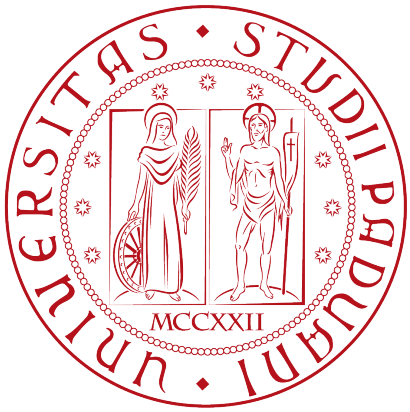
Algorithmic Fairness Datasets: Curation, Selection, and Applications



Alessandro Fabris

Advisor: Prof. Gian Antonio Susto

Co-advisor: Prof. Gianmaria Silvello

Coordinator: Prof. Andrea Neviani

Department of Information Engineering
University of Padua

This dissertation is submitted for the degree of
Doctor of Philosophy

December 2022

To the mentors I have been lucky to find.
To those who make home a fantastic place to be.

Abstract

This thesis supports measurements of algorithmic fairness from a data-centric perspective. First, we tackle the problem of dataset selection, highlighting misguided practices prevalent in the field, and providing solutions for more principled approaches. Second, we turn to dataset curation. We design and collect datasets for a fairness audit of algorithms deployed nation-wide in Italy and zoom out to study dataset curation more broadly. We distill a set of best practices for data curation based on hundreds of datasets used in algorithmic fairness research. Third, we tackle the problem of measuring fairness in practical settings where information on sensitive attributes is not available. Finally, we target the gap between fairness definitions and their mathematical formulation, proposing and validating novel measures of equity in information access. Overall, this thesis navigates the tension between algorithmic equity and the complexity of data acquisition, supporting fairness, accountability, and transparency for technology developed by and for a responsible society.

Table of contents

List of figures

List of tables

Chapter 1

Introduction

Algorithms are used to assist humans with perception, forecasting, and decision-making in sensitive contexts, including education [201], healthcare [596], hiring [828], lending [496], and law enforcement [21]. Compared to human processes, automated procedures offer several advantages. First, algorithms are typically capable of quick predictions, making them faster than humans in many applications of interest. Second, they can take into account more variables and more complex patterns. Third, they can be deployed in parallel across multiple instances, making them highly scalable. Fourth, algorithms can deliver repeatable outcomes, that is, they are designed to only consider the features of interest, to process them consistently, and to neglect irrelevant exogenous factors which, on the other hand, may influence human decisions [188, 666]. Alongside these technical desiderata, important ethical requirements for algorithms have been articulated in recent years [408]. The efforts of civil society, regulatory agencies, policy makers, researchers, and practitioners have contributed to defining diverse requirements for trustworthy algorithms [364], such as the respect for human autonomy, harm avoidance, explainability, and fairness. *Algorithmic fairness*, closely related to equity and non-discrimination, is the focus of this thesis. To simplify, an algorithm used for decisions about human subjects is fair if it does not wrongly impose on subjects a relative disadvantage according to social characteristics like race or gender [552]. Algorithms designed and trained without explicit attention to fairness are unlikely to satisfy this desideratum [101, 483, 607], with suboptimal performance for entire categories, such as women and African-American subjects, who are directly harmed by their deployment.

Discussions on fairness and non-discrimination may seem strange in the context of AI and machine learning, where discriminatory power is typically a desired property of models. *Discrimination*, in this case, is an overloaded word. Algorithmic fairness aims to develop algorithms that can quantitatively distinguish individuals in support of a given task (desirable discrimination) without systematically imposing a disadvantage on a social

group (undesirable discrimination). Throughout this thesis, we let the word *discrimination* denote this inequitable imposition, which is problematic for different reasons [50]. From a technical perspective, discriminatory algorithms can be challenged for lack of relevance and granularity. Discrimination relies on membership in a social group, which should often be irrelevant to the problem at hand, and implies a coarse approach where individuals from a given group receive similar treatment. In addition, discriminatory algorithms can be accused of prejudice and disrespect, as they treat entire groups as inferior to others. Furthermore, discrimination typically refers to attributes over which people have no control, such as their birthplace or ethnicity, giving subjects little power to improve their situation. Finally, a very compelling reason to foster fairness and non-discrimination in decision making is to avoid compounding existing injustice. Indeed, some categories may appear different from others at decision time due to past injustice; as an example, a group may consist of individuals who seem high-risk or low-potential, due to structural societal biases including wealth disparity, less time for personal development, and past discrimination.

To address these concerns, algorithmic fairness is emerging as a field focused on uncovering, studying, and countering undesirable algorithmic discrimination. Similarly to other quantitative fields, algorithmic fairness is based on data. In fact, the quality of datasets employed in research and practice are central to the validity of experiments and generalization of results in this field. Noisy, inaccurate, or otherwise non-representative data inevitably restrain the validity and utility of associated findings [443]. The approaches adopted in this scholarly field to select data for experiments bear consequence on the reliability and accuracy of their findings [382]. Downstream effects of data issues triggered by poor practice that undervalues data quality are both common and avoidable [694]. To date, an analysis of the data practices employed in algorithmic fairness is lacking in the literature, therefore hindering understanding, reflection, and potential improvements in this key area.

Clearly, the right data for a given experiment is not always readily available. Curating a new dataset becomes necessary in this case. This occurrence is very common in algorithmic audits [559, 793], which are data-driven quantitative analyses aimed at uncovering potential discrimination in a given algorithm. Most audits in the literature focus on US-centric applications [41], while algorithms and use cases from European countries have received less attention in comparison [148]. This lack of scrutiny is concerning as it can lead to an undetected violation of the values encoded in legislation and normative frameworks, including fairness and equity.

In addition, it should be noted that, during dataset design and curation, there are some important data choices impacting data subjects and dataset users, including documentation and consent elicitation. These recurrent aspects of data curation are often neglected or treated

separately. Therefore, dataset curators are left with insufficient and sparse guidance on curatorial best practice, often without practical examples. Poor data curation can have serious consequences ranging from poor data usage [382] to potential harm to data subjects [468].

After their curation, datasets can be repurposed to support goals that were not originally envisioned. This is often the case for algorithmic fairness, which is becoming a common desideratum for products. A specific data requirement to measure algorithmic fairness is the knowledge of sensitive attributes, which, however, is often unavailable due to specific legislation, privacy-by-design standards, and a data minimization ethos [17, 82]. Therefore, it is frequently impossible to measure, let alone improve, algorithmic fairness in these applications. Posterior annotation of sensitive attributes is possible but can also have a negative impact on the privacy of individuals and runs the risk of favoring profiling along sensitive attributes.

Along with datasets, *measures* are a key component of the fairness measurement process. The right data alone does not guarantee an accurate fairness measurement, unless coupled with the right fairness measure, i.e., with a careful mathematical formulation of what it means for an algorithm to be fair in a given context. A plethora of fairness measures are available to researchers and practitioners and their simultaneous satisfaction is often impossible [50, 153]. This fact suggests that a careful and contextual definition of fairness is crucial. However, fairness measures are often introduced as self-evident requirements, without appropriate contextualization and comparison with other available measures. In fair ranking alone, tens if not hundreds of metrics have been proposed [794]. This abundance, often coupled with a lack of discussion on which measures are more appropriate for a given context, can lead to adoption of suboptimal measures, which do not suitably capture the property of interest, and may fail to catch unjust outcomes and harms for individuals.

1.1 Objectives and Contributions.

Note that these algorithmic fairness challenges can be mapped to different stages of model development, similarly to *data cascades* [694], which are data issues affecting algorithmic development downstream at various stages. Figure 1.1, adapted from Sambasivan et al. [694], introduces a schematic representation of data-driven algorithmic development, mapping its stages to the central chapters of this thesis. Fairness measurements are key to model evaluation, and stand to benefit from progress in data practices at upstream stages proposed in this work. To support improved, contextualized, and responsible fairness measurements, the main objectives of this thesis are as follows.

- Enable principled approaches for dataset selection in algorithmic fairness.

Fig. 1.1 Stages of data-driven algorithmic development.

- Audit an algorithm deployed nation-wide in Italy and distill a set of best practices for dataset curation.
- Support fairness measurements under unawareness of sensitive attributes while respecting individual privacy.
- Study and develop fair ranking measures with a contextualized approach.

Towards our first goal, we undertake a thorough documentation effort, by tracking dataset usage across algorithmic fairness articles published in proceedings of important conferences over seven years. We find over two hundred different datasets and rigorously show that their utilization follows a power law, with most datasets used in only one article and few datasets used in ten articles or more. We demonstrate a disproportionate popularity for three datasets, documenting them in depth, and calling into question their suitability as fairness benchmarks. For each of the (200+) datasets found, we produce standardized documentation enabling practitioners and researchers to find datasets according to fundamental properties such as their domain, the tasks supported in the fairness literature, or the sensitive attributes encoded.

Next, we identify the Italian car insurance market as a high-stakes domain where algorithms play a central role to mediate access and pricing. We study their regulatory and social context, and turn to dataset curation to perform an audit of these algorithms informed by relevant normative frameworks. In addition, we study important topics in dataset curation, namely re-identification, consent, inclusivity, labeling, and documentation. We analyze a large set of fairness datasets from these perspectives, finding different approaches and levels of attention to these topics, distilling them into a set of best practices to reduce harm for data subjects and improve communication in future endeavors of data curation.

Moreover, we study the problem of posterior annotation to measure fairness when sensitive attributes are unknown. We demonstrate that quantification learning [316] is broadly applicable in this setting and substantially outperforms previously proposed approaches due to

its intrinsic robustness to distribution drift. We show this fact over five experimental protocols, proposed by us to mirror different challenges encountered when estimating algorithmic fairness. We also show that quantification methods can produce accurate estimates at the sample level, while preventing precise inference at the individual level.

Finally, we turn to fairness measures for ranking algorithms. We study *pairwise fairness*, a family of measures proposed in the fair ranking literature without proper justification. We develop a normative and behavioral grounding for pairwise fairness. Leveraging measurement theory and user browsing models, we derive an interpretation of pairwise fairness centered on the construct of producer dissatisfaction, tying pairwise fairness to perceptions of ranking quality. Highlighting the key limitations of prior pairwise measures, we introduce a set of reformulations that allow us to capture behavioral and practical aspects of ranking systems. Additionally, we focus on text search and target the construct of *gender stereotype reinforcement* introduced by ranking algorithms. We propose the first measure for this construct, study its validity, and apply it, on carefully chosen datasets, to quantify the extent to which different information retrieval algorithms can reinforce or counter gender stereotypes.

1.2 Outline

This thesis is organized as follows. Chapter 2 contextualizes this work, introducing relevant background from algorithmic fairness, information retrieval evaluation, and data studies. In Chapter 3, we focus on dataset selection. We identify the most popular resources for algorithmic fairness research and describe their limitations. Next, we present our dataset documentation framework, supporting task-oriented and domain-oriented search of alternative resources. In Chapter 4, we turn to dataset curation. First, we present *Italian Car Insurance*, a dataset we designed and collected to audit important algorithms, responsible for access and pricing in the Italian car insurance industry. Our analyses highlight problematic practices which go against anti-discrimination law at the national and international level. We then zoom out, considering a large selection of algorithmic fairness datasets, and their position with respect to important topics of data curation. We distill our findings into a set of best practices for data anonymization, consent, inclusivity, labeling, and transparency. In Chapter 5, we consider the problem of measuring algorithmic fairness on datasets where sensitive features are missing. We demonstrate the suitability of quantification learning algorithms in this common setting, with superior performance under different types of dataset shift, and their potential to discourage misuse such as profiling and other undesirable inferences against individuals. In Chapter 6, we zoom in on a specific task, focusing on measures for fair ranking. First, we analyze pairwise fairness, describing the underlying construct,

its relationship to other measures of fair ranking, and its limitations, proposing targeted improvements to overcome them. Second, we develop the first measure of gender stereotype reinforcement in search engines, and extensively analyze its validity on carefully selected datasets. Finally, in Chapter 7, we report our conclusions and directions for future work. It is worth noting that each chapter presents additional specific background, allowing interested readers to consult individual chapters independently.

1.3 Publications

Most results in this thesis have been published (or accepted for publication) at conferences and journals on algorithmic fairness, information retrieval, and data mining [70, 207, 240, 247–253]. They are presented below in chronological order.

Journal Papers

- A. Fabris, A. Esuli, A. Moreo Fernández, F. Sebastiani: Measuring Fairness under Unawareness of Sensitive Attributes: A Quantification-Based Approach. *Journal of Artificial Intelligence Research*. 2023. Accepted with minor revisions.
- A. Fabris, S. Messina, G. Silvello, G.A. Susto: Algorithmic Fairness Datasets: the Story so Far. *Data Mining and Knowledge Discovery,* special issue on Bias and Fairness in AI. 2022.
- A. Fabris, A. Purpura, G. Silvello, G.A. Susto: Gender stereotype reinforcement: Measuring the gender bias conveyed by ranking algorithms. *Information Processing & Management*. 2021. **Best PhD paper award**.

Conference Papers

- A. Fabris, G. Silvello, G.A. Susto, A. Biega: Pairwise Fairness in Ranking as a Dissatisfaction Measure. *Proceedings of the 16th ACM conference on Web Search and Data Mining*. WSDM 2023, 27 February–3 March, Singapore. In press.
- A. Fabris, S. Messina, G. Silvello, G.A. Susto: Tackling Documentation Debt: A Survey on Algorithmic Fairness Datasets. *Proceedings of the second ACM conference on Equity and Access in Algorithms, Mechanisms, and Optimization*. EAAMO 2022, 6–9 October, Arlington, USA.
- A. Fabris, A. Mishler, S. Gottardi, M. Carletti, M. Daicampi, G. A. Susto, G. Silvello: Algorithmic Audit of Italian Car Insurance: Evidence of Unfairness in Access and Pricing. *Proceedings of the 4th AAAI/ACM Conference on Artificial Intelligence, Ethics and Society*. AIES 2021, 19–21 May, Virtual Event.
- D. Biasion, A. Fabris, G. Silvello, G. A. Susto: Gender Bias in Italian Word Embeddings. *Proceedings of the Seventh Italian Conference on Computational Linguistics*. CLIC-IT 2020, 1–3 March, Virtual Event.

Workshop Papers

- A. Fabris, A. Purpura, G. Silvello, G.A. Susto: Measuring Gender Stereotype Reinforcement in Information Retrieval Systems. *Proceedings of the 12th Italian Information Retrieval Workshop* . IIR 2021, 13-15 Septeber, Bari, Italy.
- G.M. Di Nunzio, A. Fabris, G. Silvello, G.A. Susto: Incentives for Item Duplication under Fair Ranking Policies. *Proceedings of the 2nd International Workshop on Algorithmic Bias in Search and Recommendation*. BIAS@ECIR2021, Virtual Event.

Books

- A. Esuli, A. Fabris, A. Moreo Fernández, F. Sebastiani: Learning to Quantify. *The Information Retrieval Series. Springer*. 2023. In press.

Chapter 2

Background

2.1 Algorithmic Fairness

As a result of an increasing adoption of automated decision making, algorithmic fairness has recently received increasing attention from academia, industry, media, and government. In academia, many articles have surveyed algorithmic fairness across domains [119, 552, 634] and in specific areas [639, 860], with an in-progress textbook [50] and entire conferences dedicated to the topic, including the ACM Conference on Fairness, Accountability, and Transparency (FAccT), the AAAI/ACM Conference on Artificial Intelligence, Ethics and Society (AIES) and the ACM conference on Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO). This work has inspired numerous software packages [465, 734, 820] and industry efforts, including consulting services for technology ethics,[1] dedicated corporate teams,[2] and public bounty challenges.[3] Popular books [179, 599, 610] and mass media [427] have contributed to a growing awareness of algorithmic fairness problems outside technical circles. Consequently, several regulatory initiatives have been undertaken to articulate uniform guidelines and norms for the responsible development of artificial intelligence and algorithmic decision-making [243, 258].

Algorithmic fairness is a multifaceted construct that can be defined in multiple ways [398]. Narrow views concentrate on granting equal opportunities to individuals who appear to be similar at decision time [219], while broader conceptualizations take into account past social structures holding back some social groups, and require their restructuring [667]. Even when restricting the space of possible measures to a specific notion of fairness, there are

[1] https://www.eticasconsulting.com/

[2] https://www.microsoft.com/en-us/research/theme/fate/

[3] https://blog.twitter.com/engineering/en_us/topics/insights/2021/algorithmic-bias-bounty-challenge

always multiple ways to *operationalize*, i.e., translate into mathematical form, said notion. As an example, it is reasonable to request of a model equal performance across different social groups, e.g., men and women. Depending on the task and point of view considered, this can amount to a requirement of parity in measures derived from the groupwise confusion matrices, such as precision or recall. Mathematically, let Y denote the ground truth variable, with values in $\{\oplus, \ominus\}$, $\hat{Y}$ its algorithmic estimate, and S binary gender, a simplistic sensitive attribute with values $\{\text{man}, \text{woman}\}$. Depending on whether precision (Prec) or recall (Rec) is more important for the problem at hand, disparity measures can be defined as

$$\delta^{\text{Prec}} = \Pr(Y = \oplus | \hat{Y} = \oplus, S = \text{man}) - \Pr(Y = \oplus | \hat{Y} = \oplus, S = \text{woman}) \tag{2.1}$$

$$\delta^{\text{Rec}} = \Pr(\hat{Y} = \oplus | Y = \oplus, S = \text{man}) - \Pr(\hat{Y} = \oplus | Y = \oplus, S = \text{woman}), \tag{2.2}$$

where a value of $\delta = 0$ indicates maximum fairness, while extreme values $\delta \in \{-1, 1\}$ indicate maximum unfairness. The choice of Equation (2.1) or (2.2) is not merely philosophical but bears immediate consequences for the associated evaluation and model selection process. In fact, equality of precision and recall for both the positive and negative class has been shown to be incompatible under realistic conditions [50, 153, 452]. Therefore, researchers and practitioners must carefully select how they measure and optimize fairness among hundreds of definitions proposed in the literature [118, 552, 572]. In Chapter 6 we propose two novel measures of fair ranking, analyzing in depth their appropriateness, novelty, and validity [558]. The choice of a measure is further complicated by different axes of unfairness which may be taken into account, i.e. gender, race, or other attributes that are relevant to the context at hand. Crucially, most fairness definitions require knowledge of these sensitive attributes (denoted by S in equations 2.1 and 2.2), which may be unavailable following privacy concerns and data minimization principles [17, 82]. In fact, *fairness under unawareness* of sensitive attributes is a setting of high practical interest, as analyzed in Chapter 5.

It is worth noting that the very same measures of algorithmic fairness can be used to evaluate human decisions. The study of mechanisms that lead humans to mental shortcuts such as gender stereotypes and their harmful application in different contexts has a long tradition [53, 81, 186, 224, 234, 466, 640]. Chapter 6 engages with this literature to present a measure of gender stereotype reinforcement in text search engines. The study of discrimination in algorithmic decision-making and its sources, on the other hand, is more recent [552, 572]. We can distinguish different causes of algorithmic discrimination, associated with different phases of algorithmic development. Model evaluation, for example, can play a critical role. Classical aggregative measures such as accuracy are dominated by majority groups and, thus, may fail to catch malfunctions affecting minorities. Also the choice of model class ahead of

its training may influence its fairness. An important source of algorithmic discrimination for data-driven models is often the training data itself. This is due to a combination of three types of bias typically present, to a different extent, in every dataset. *Sample bias* refers to datasets that are not fully representative of the population to which the model will be deployed. For example, loan decision models are trained on data on individuals who were granted credit in the past. Moreover, datasets that are scraped online are unlikely to represent the general population offline. *Measurement bias* refers to measurement errors and disconnects between what a feature is intended to capture and what it ends up measuring. For example, judicial risk assessment models consider prior arrests as a proxy for crime. Arrest data is mediated by policing patterns, typically concentrated in poorer neighborhoods. This leads to measurement errors that differ according to the social group of an individual: crime data for individuals may be underestimated or overestimated depending on their wealth, leading to systematic underestimation or overestimation of risk. Finally, *societal bias* will inevitably affect any dataset. For example, educational data on student performance can accurately record race-specific differences in grades. At the same time, it would miss precious information on the potential and ambition of each student, which, unlike grades, is unlikely to be affected by the same differences. Society in this case represents a mediator between student potential and attainment, with race-dependent effects.

Clearly, datasets play a central role in algorithmic fairness, as in other data-driven fields. To assess algorithmic fairness, we must choose, along with an appropriate fairness measure, the data on which that measurement should take place. This choice, although somewhat neglected, is fundamental in both theoretical studies and in real-world applications. Fairness testing in practice requires the curation of ad-hoc datasets, including numerous choices of data selection and labeling, which may significantly affect fairness [607]. We discuss data curation in Chapter 4. Furthermore, theoretical contributions to algorithmic fairness, including novel algorithms and approaches, are typically demonstrated on selected datasets. While some results have general validity, other studies, especially empirical ones, may yield different conclusions depending on the dataset at hand [213, 282]. Crucially, as shown in Chapter 3, the data practices of algorithmic fairness researchers have crystallized around few datasets of dubious merit, which have a disproportionate influence on the results obtained by the community. In Chapter 3 we perform a critical analysis on hundreds of alternative resources with the goal of supporting principled approaches to dataset selection in algorithmic fairness.

2.2 Data Studies and Documentation

Data is not objective or neutral; rather it is the result of different choices by its creators, including which instances and features should be included, excluded, or preprocessed, and how should inevitable measurement noise be treated, if at all. Critical Data Studies [92, 387] elaborate on this fact from a sociological, political, economical, and philosophical perspective, engaging with the broader context around data artifacts and their influence on society [54, 180, 274]. This attention to data also covers its procurement and its ethics, including concerns about privacy and consent for data subjects [560]. Chapter 4 analyzes the datasets employed in the algorithmic fairness literature from this perspective. Overall, datasets are useful abstractions that simplify the underlying phenomena, making them computationally treatable. The fact that they can be repurposed and play multiple roles, as shown in Chapter 3 (Section 3.5), can cause a disconnect between the data and the contextual information required to handle it properly, due to inadequate abstraction and insufficient documentation [387].

In fact, public datasets are typically used by dataset consumers without direct knowledge of the data curation process, except for the documentation available with the dataset. In recent years, several frameworks have been proposed to improve the reproducibility of results, increase the equity of data-driven models, favor reflection on part of dataset creators, and increase accountability and awareness in the machine learning community [58, 292, 366]. This line of research has contributed to the transparency of data curation initiatives, reducing the associated *documentation debt* [42, 59] i.e., the suboptimal use of datasets due to poor documentation. Formal documentation, critical analyses, and quantitative audits of datasets can also be published retrospectively [213, 329, 644], drawing the attention of the research community to important aspects of a dataset that were previously understudied. While these initiatives reduce the opacity of specific data artifacts, the problem of documentation sparsity remains: information on datasets is scattered across repositories, websites, and scientific articles, and there is a high chance that it will not reach interested parties. Chapter 3 is especially related to this line of work, describing a centralized documentation initiative aimed at reducing documentation sparsity, pointing out poor data practices, and allowing their improvement on part of the algorithmic fairness community.

2.3 Information Retrieval Evaluation

Well-grounded evaluation, i.e. a clear operationalization of success in line with desiderata, is a key factor for the advancement of disciplines in computer science [400, 604, 829].

Information Retrieval (IR) is a a field with a long tradition of rigorous evaluation [161, 162, 347] contributing to the development of highly successful applications such as modern search engines. IR systems must operate well in a variety of domains, users, contexts, and tasks. To exemplify, different IR systems support patient or rushed users searching items such as web pages, legal, or medical documents, in settings where the consequences of missing a relevant item can range from unimportant to catastrophic. To develop reliable systems in each of these settings, an appropriate and contextualized measurement of retrieval performance is fundamental.

From the early initiatives in library and information science [161, 162] to more recent evaluation campaigns such as the Text REtrieval Evaluation Conference (TREC – [801]) and its sister conferences, including the Forum for Information Retrieval Evaluation (FIRE), the NII Testbeds and Community for Information access Research project (NCTIR), and the Conference and Labs of the Evaluation Forum (CLEF), the IR research community has greatly benefited from constant attention to datasets and measures. In a similar fashion, this thesis argues that datasets and measures are the two key components of algorithmic fairness.

Specific datasets are often curated in IR to address challenges such as web search [352], civil litigation discovery [51], question answering [799], and spam filtering [166]. More recently, new datasets and measures have been developed to study the problem of fairness in ranking [71, 73], to which we contribute in Chapter 6. IR datasets, termed *collections*, consist of a set of queries that mirror typical information needs, a set of items that may be relevant to an information need, and a set of relevance judgements, typically issued by human assessors for a limited set of query-item pairs, encoding whether the item is indeed relevant to the query and should be retrieved by a system in response to it. Collections are curated to develop, test, and compare IR systems. The fact that one system performs better than another can be a general fact or a specific result valid only on a given collection. Therefore, the properties of collection are studied to understand their advantages and limitations in terms of generalization [100, 412]. Ideally, superior performance in a collection should guarantee higher user satisfaction in practical applications. A disconnect between the two may arise from disparate factors, including the subjective nature of relevance judgements and the choice of a final performance measurement, which should summarize system performance to mirror the point of view of users as closely as possible. The properties of collections are studied precisely for this reason: to help the IR community steer away from results that are an artefact of a given dataset [9, 362, 695]. In a similar vein, Chapter 3 of this work discusses the limitations of common datasets used in algorithmic fairness research and provides a critical analysis of existing alternatives.

Chapter 3

Dataset Selection

Following the widespread study and application of data-driven algorithms in contexts that are central to people's well-being, a large community of researchers has coalesced around the growing field of algorithmic fairness and equity, investigating algorithms through the lens of justice, bias, power, harms, and equality [21, 50, 101, 153, 282, 628]. A line of work that has gained traction in the field, intersecting with critical data studies, human-computer interaction, and computer-supported cooperative work, focuses on data transparency and standardized documentation processes to describe key characteristics of datasets [58, 292, 293, 366, 405, 564]. Most prominently, Gebru et al. [292] and Holland et al. [366] proposed two complementary documentation frameworks, called *Datasheets for Datasets* and *Dataset Nutrition Labels*, respectively, to improve data curation practices and favour more informed data selection and utilization for dataset users. Overall, this line of work has contributed to an unprecedented attention to dataset documentation in ML, including a novel track focused on datasets at the Conference on Neural Information Processing Systems (NeurIPS), an initiative to support dataset tracking in repositories for scholarly articles,[1] and dedicated works producing retrospective documentation for existing datasets [42, 287], auditing their properties [644] and tracing their usage [630].

Data documentation is important and caters to different goals. It increases transparency, favoring improved understanding of the data and resulting models [385], reduces the chances of data misuse [292] and supports accountability in dataset and model creation [385], it helps connect the data with its context to guide scientific inquiry [627], and makes the values influencing the curation of datasets explicit [700]. Technical debt is a cost incurred in software development when speed of execution is prioritized over quality [385]. In recent work, Bender et al. [59] propose the notion of *documentation debt*, in relation to training sets that are undocumented and too large to document retrospectively, which compounds

[1] https://medium.com/paperswithcode/datasets-on-arxiv-1a5a8f7bd104

over time with serious consequences on dataset understanding and use. We extend this definition to the collection of datasets employed in a given field of research. We see two components at work contributing to the documentation debt of a research community. On the one hand, *opacity* is the result of poor documentation affecting single datasets, contributing to misunderstandings and misuse of specific resources. On the other hand, when relevant information exists but does not reach interested parties, there is a problem of documentation *sparsity*. An example that is particularly relevant for the algorithmic fairness community is represented by the German Credit dataset [774], a popular resource in this field. Many works of algorithmic fairness, including recent ones, carry out experiments on this dataset using sex as a protected attribute [34, 356, 514, 534, 633, 713, 810, 841], while existing, yet overlooked, documentation shows that this feature cannot be reliably recovered [329].

Indeed, German Credit is one of the most frequently used datasets in algorithmic fairness research. Together with Adult [459] and COMPAS [21], they represent *de-facto* benchmarks for algorithmic fairness research. These resources have become established after being used in seminal works [107, 629] and influential articles [21] on algorithmic fairness. Although some of their limitations have been studied individually [43, 213, 329], their overall status as reference resources in the fairness literature remains unquestioned, due to a combination of documentation opacity and sparsity. In fact, a thorough and standardized evaluation of their merits and limitations has not been carried out, thus favoring the proliferation of this default data utilization practice. Our first goal is to challenge this practice and evaluate its merits through a centralized, in-depth documentation effort.

O1: Analyze the suitability of the most popular datasets as algorithmic fairness benchmarks.

This objective is supported by (1) a preliminary investigation to rigorously identify the most frequently used datasets in the fairness literature, (2) a standardized documentation effort, compiling *Datasheets for Datasets* and *Dataset Nutrition Labels* for these datasets, and (3) a summary of key results.

Another consequence of the data documentation debt is that the mere existence of a dataset and its relevance to a given task or a given domain may be unknown. For example, the BUPT Faces datasets were presented as the second existing resource for face analysis with race annotations [812]. However several resources were already available at the time, including Labeled Faces in the Wild [338], UTK Face [875], Racial Faces in the Wild [813], and Diversity in Faces [556].[2] Unawareness of resources and lack of information about them is one of the key factors reinforcing suboptimal data practices in the algorithmic

[2]Hereafter, for brevity, we only report dataset names. The relevant references and additional information can be found in Appendix A.

fairness community, and can be tackled with a wide documentation initiative. For example, it would be important to enable domain-oriented and task-oriented search across a wide array of resources to support practitioners and researchers looking for datasets in their field of interest.

O2: Enable principled approaches for fairness dataset selection.

Towards this goal, we survey the datasets used in more than 500 articles on fair ML and equitable algorithmic design, presented at seven major conferences, considering each edition in the period 2014–2021, and more than twenty domain-specific workshops in the same period. We find over 200 datasets employed in studies of algorithmic fairness, for which we produce compact and standardized documentation, called *data briefs*. Data briefs are intended as a lightweight format to document fundamental properties of data artifacts used in algorithmic fairness, including their purpose, their features, with particular attention to sensitive ones, the underlying labeling procedure, and the envisioned ML task, if any. To favor domain-based and task-based search from dataset users, data briefs also indicate the domain of the processes that produced the data (e.g., radiology) and list the fairness tasks studied on a given dataset (e.g. fair ranking). For this endeavor, we have contacted creators and knowledgeable practitioners identified as primary points of contact for the datasets. We received feedback (incorporated into the final version of the data briefs) from 79 curators and practitioners, whose contribution is acknowledged at the end of this work.

This chapter is organized as follows. First, in Section 3.1, we identify and describe the three most popular datasets in algorithmic fairness. We add to and unify recent scholarship on these datasets, calling into question their suitability as general-purpose fairness benchmarks. Next, we present alternative fairness resources and approaches to select them, enabling a domain-oriented, task-oriented, and role-oriented search. In Section 3.2, we describe our documentation initiative, presenting a novel framework to document datasets called *data briefs*, the inclusion criteria we adopted, and the related work. In Section 3.3 we describe the different domains spanned by fairness datasets. In Section 3.4 we provide a fine-grained categorization of fairness tasks, and algorithmic fairness datasets on which they have been studied. In Section 3.5 we discuss the different roles these datasets play in fairness research. These sections serve the dual purpose of introducing the relevant categories to practitioners interested in dataset selection for a specific task or domain, and providing a critical analysis of algorithmic fairness research so far. Finally, Section 3.6 summarizes the results of this chapter and discusses the immediate benefits and broader relevance of our initiative to the research community.

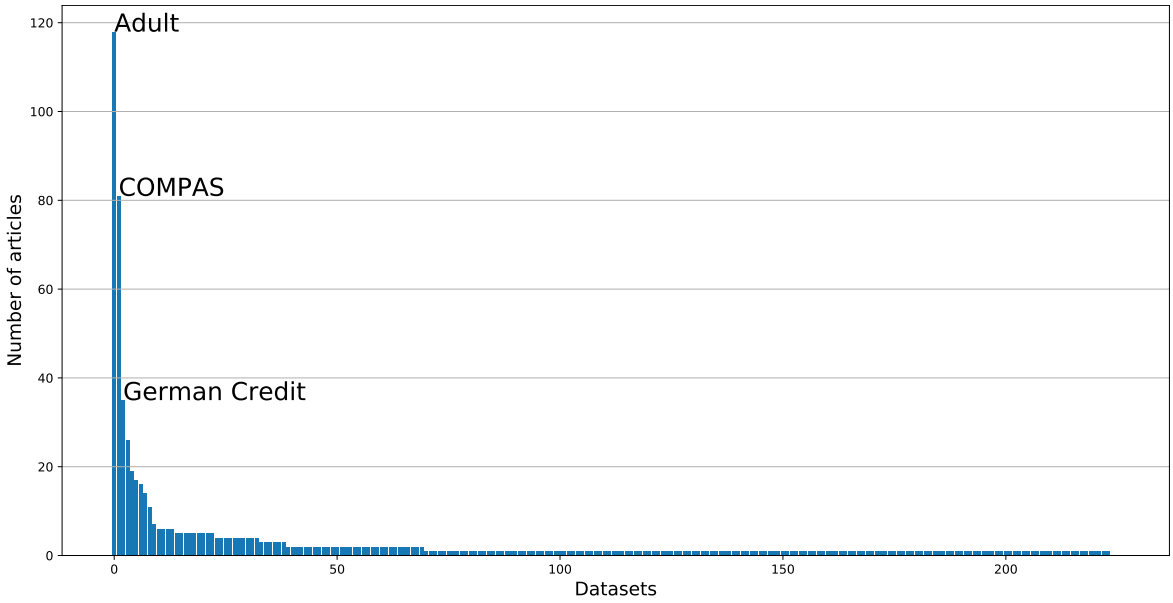


Fig. 3.1 Utilization of datasets in fairness research follows a long-tail distribution.

3.1 Limitations of Algorithmic Fairness Benchmarks

Dataset utilization in scholarly works on algorithmic fairness follows a long-tail distribution, reflecting findings of data use in computer vision [457]. Figure 3.1 depicts useage counts for datasets in fairness articles recently published at selected venues.[3] Over 100 datasets are only used once, also because some of these resources are not publicly available. Complementing this long tail is a short head of nine resources used in ten or more articles. These datasets are Adult (118 usages), COMPAS (81), German Credit (35), Communities and Crime (26), Bank Marketing (19), Law School (17), CelebA (16), MovieLens (14), and Credit Card Default (11). The tenth most used resource is the toy dataset from Zafar et al. [856], used in 7 articles. In this section, we summarize the positive and negative aspects of the three most popular datasets, namely Adult, COMPAS, and German Credit, informed by the extensive documentation in Appendix A.

3.1.1 Adult

The Adult dataset was created as a resource to benchmark the performance of machine learning algorithms on socially relevant data. Each instance is a person who responded to the March 1994 US Current Population Survey, represented along demographic and socio-economic dimensions, with features describing their profession, education, age, sex,

[3] The inclusion criteria are described in Section 3.2.

race, personal, and financial condition. The dataset was extracted from the census database, preprocessed, and donated to UCI Machine Learning Repository in 1996 by Ronny Kohavi and Barry Becker. A binary variable encoding whether respondents' income is above $50,000 was chosen as the target of the prediction task associated with this resource.

Adult inherits some positive sides from the best practices employed by the US Census Bureau. Although later filtered somewhat arbitrarily, the original sample was designed to be representative of the US population. Trained and compensated interviewers collected the data. The attributes in the data set are self-reported and provided by consensual respondents. Finally, the original data from the US Census Bureau is well documented, and its variables can be mapped to Adult by consulting the original documentation [781], except for a variable denominated `fnlwgt`, whose precise meaning is unclear.

A negative aspect of this dataset is the contrived prediction task associated with it. Income prediction from socio-economic factors is a task whose social utility appears rather limited. Even discounting this aspect, the arbitrary $50,000 threshold for the binary prediction task is high, and model properties such as accuracy and fairness are very sensitive to it [213]. Furthermore, there are several sources of noise affecting the data. Approximately 7% of the data points have missing values, plausibly due to problems with data recording and coding, or the inability of the respondents to recall information. Moreover, the tendency of respondents in household surveys to underreport their income is a common concern of the Census Bureau [576]. Another source of noise is top-coding of the variable "capital-gain" (saturation to $99,999) to avoid the re-identification of certain individuals [781]. Finally, the dataset is rather old; sensitive attribute "race" contains the outdated "Asian Pacific Islander" class. It should be noted that a set of similar resources was recently made available, allowing more current socio-economic studies of the US population [213].

3.1.2 COMPAS

This dataset was created for an external audit of racial biases in the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) risk assessment tool developed by Northpointe (now Equivant), which estimates the likelihood of a defendant becoming a recidivist. Instances represent defendants scored by COMPAS in Broward County, Florida, between 2013–2014, reporting their demographics, criminal record, custody and COMPAS scores. Defendants' public criminal records were obtained from the Broward County Clerk's Office website matching them according to date of birth, first and last names. The dataset was augmented with jail records and COMPAS scores provided by the Broward County Sheriff's Office. Finally, public incarceration records were downloaded from the Florida Department of Corrections website. The instances are associated with two target variables (is_recid and

Table 3.1 Limitations of popular algorithmic fairness datasets.

	Adult	COMPAS	German Credit
Age	Old (1994)	Recent (2013–2016)	Very old (1973–1975)
Prediction task	Contrived (income > 50K$)	Realistic (recidivism)	Realistic (creditworthiness)
Sensitive attributes	Outdated racial categories	Outdated racial categories	Sex cannot be retrieved
Sources of noise	Top-coding; tendency to under-report income	Data leakage; label bias; clerical errors	Incorrect code table
Sample representativeness	US working population	Convenience sample (Broward County)	Artificial sample (credit granted, negative class oversampled)
Preprocessing needed	Handling missing values (7%)	Handling missing values (80%); removing redundant features; ground truth on detainment	None
Additional concerns	Accuracy and fairness are sensitive to arbitrary 50K$ threshold	Potential for misguided discussion on criminal justice	Interpretability and exploratory analyses are invalid

is_violent_recid), indicating whether defendants were booked in jail for a criminal offense (potentially violent) that occurred after their COMPAS screening and within two years.

On the upside, this dataset is recent and captures some relevant aspects of the COMPAS risk assessment tool and the criminal justice system in Broward County. On the downside, it was compiled from disparate sources, hence there are clerical errors and mismatches [483]. Furthermore, in its official release [646], the COMPAS dataset features redundant variables and data leakage due to spuriously time-dependent recidivism rates [46]. For these reasons, researchers must perform further preprocessing in addition to the standard one from ProPublica. More subjective choices are required of researchers interested in counterfactual evaluation of risk-assessment tools, due to the absence of a clear indication of whether defendants were detained or released pre-trial [568]. The lack of a standard preprocessing protocol beyond the one by ProPublica [646], which is insufficient to handle these factors, may cause reproducibility issues and difficulty in comparing methods. Furthermore, according to Northpointe's response to ProPublica's study, several risk factors considered by the COMPAS algorithm are absent from the dataset [212]. As an additional concern, race categories lack Native Hawaiian or Other Pacific Islander, while Hispanic is redefined as race instead of ethnicity [43]. Finally, defendants' personal information (e.g. race and criminal history) is available in conjunction with obvious identifiers, making re-identification of defendants trivial.

COMPAS also represents a case of a broad phenomenon that can be termed *data bias*. With the terminology of Friedler et al. [281], when it comes to datasets that encode complex human phenomena, there is often a disconnect between the *construct space* (what we want to measure) and the *observed space* (what we end up observing). This may be especially problematic if the difference between construct and observation is uneven between individuals or groups. COMPAS, for example, is a dataset on criminal offenses. Offense is central to the prediction target Y, which aims to encode recidivism, and to the available covariates X, summarizing criminal history. However, the COMPAS dataset (observed space) is an imperfect proxy for the criminal patterns it should summarize (construct space). The prediction labels Y actually encode re-arrest, instead of re-offense [483], and are thus clearly influenced by spatially differentiated policing practices [276]. This is also true for the criminal history encoded in the COMPAS covariates, again mediated by arrest and policing practices that may be racially biased [43, 539]. As a result, the true fairness of an algorithm, like its accuracy, may differ significantly from what is reported on biased data. For example, algorithms that achieve equality of true positive rates between sensitive groups in COMPAS are deemed fair under the *equal opportunity* measure [344]. However, if both the training set on which this objective is enforced and the test set on which it is measured are affected by race-dependent noise described above, those algorithms are only "fair" in an abstract observed space, but not in the real construct space that we ultimately care about [281].

Overall, these considerations paint a mixed picture for a dataset of high social relevance that was extremely useful to catalyze attention on algorithmic fairness issues, displaying at the same time several limitations in terms of its continued use as a flexible benchmark for fairness studies of all sorts. In this regard, Bao et al. [43] suggest avoiding the use of COMPAS to demonstrate novel approaches in algorithmic fairness, as considering data without a proper context may lead to misleading conclusions, which could misguide the broader debate on criminal justice and risk assessment.

3.1.3 German Credit

The German Credit dataset was created to study the problem of computer-assisted credit decisions at a regional Bank in southern Germany. Instances represent loan applicants from 1973 to 1975, who were deemed creditworthy and were granted a loan, causing a natural selection bias. Within this sample, bad credits are oversampled to favor a balance in target classes [329]. The data summarizes the financial situation of the applicants, their credit history, and their personal situation, including housing and number of liable people. A binary variable encoding whether each loan recipient punctually paid each installment is the target of a classification task. Among the covariates, marital status and sex are jointly encoded in a

single variable. Many documentation errors are present in the UCI entry associated with this resource [774]. A revised version with the correct variable encodings, called South German Credit, was donated to UCI Machine Learning Repository [776] with an accompanying report [329].

The greatest upside of this data set is that it captures a real-world application of credit scoring in a bank. On the downside, the data is half a century old, significantly limiting the societally useful insights that can be gleaned from it. Most importantly, the popular release of this dataset [774] comes with highly inaccurate documentation that contains incorrect variable codings. For example, the variable reporting whether loan recipients are foreign workers has its coding reversed, so that, apparently, fewer than 5% of the loan recipients in the dataset would be German. Luckily, this error has no impact on numerical results obtained from this dataset, as it is irrelevant at the level of abstraction provided by raw features, with the exception of potentially counterintuitive explanations in works of interpretability and exploratory analysis [485]. This coding error, along with others discussed in Grömping [329] was corrected in a novel release of the dataset [776]. Unfortunately and most importantly for the fair ML community, retrieving the sex of loan applicants is simply not possible, unlike the original documentation suggested. This is due to the fact that one value of this feature was used to indicate both women who are divorced, separated, or married, and men who are single, while the original documentation reported each feature value to correspond to same-sex applicants (either male-only or female-only). This particular coding error ended up having a non-negligible impact on the fair ML community, where many works studying group fairness extract sex from the joint variable and use it as a sensitive attribute, even years after the redacted documentation was published [485, 810]. These coding errors are part of a documentation debt whose influence continues to affect the algorithmic fairness community.

3.2 Beyond Benchmarks: Addressing Documentation Debt

3.2.1 Data Briefs: a Novel Documentation Framework

Several data documentation frameworks have been proposed in the literature, including *datasheets for datasets* [292], *data statements* [58], and *dataset nutrition labels* [366]. These are thorough yet cumbersome frameworks, which do not scale to a wider documentation effort with limited resources. For this reason, we propose and produce *data briefs*, a lightweight documentation format designed for algorithmic fairness datasets. Data briefs include fields specific to fair ML, such as sensitive attributes and tasks for which the dataset has been used in the algorithmic fairness literature.

More in detail, data briefs are composed of ten fields derived from shared vocabularies such as Data Catalog Vocabulary (DCAT)[4]; to be compliant with the FAIR data principles [825], we also defined a schema (with namespace `fdo`) to model the relationships between the terms, to make the links to external vocabularies explicit, and map the data briefs to a machine-readable RDF graph.[5] The `fdo` schema has been defined by reusing, as much as possible, existing terminology from established vocabularies. Below we detail the fields of the data briefs and present their correspondence to DCAT and `fdo` properties:

Description. This is a free-text field reporting (1) the aim/purpose of a data artifact (i.e., why it was developed/collected), as stated by curators or inferred from context; (2) a high-level description of the available features; (3) the labeling procedure for annotated attributes, with special attention to sensitive ones, if any; (4) the envisioned ML task, if any. Corresponds to `dct:description` in DCAT.

Affiliation of creators. Typically derived from reports, articles, or official web pages that present a dataset. Datasets can be derivatives of other datasets (e.g., Adult). We typically refer to the final resource while providing the prior context where appropriate. In the DCAT vocabulary, it is the affiliation of a `dct:publisher` (for published resources) or a `dct:creator`.

Domain. The main field where the data is used (e.g., computer vision for ImageNet) or the field studying the processes and phenomena that produced the dataset (e.g., radiology for CheXpert). Corresponds to `fdo:Domain` in the `fdo` schema.

Tasks in fairness literature. An indication of the task performed on the dataset in each article that uses the current resource. Corresponds to `fdo:Task`.

Data spec. The main format of the data. The categories envisaged are text, image, time-series, tabular data, and pairs. The latter denotes a special type of tabular data where rows and columns correspond to entities and cells to a relation between them, such as relevance for query-document pairs, ratings for user-item pairs, co-authorship relation for author-author pairs. A "mixture" category was added for multimodal data resources. Corresponds to `dct:type` in DCAT.

Sample size. Dataset cardinality. Corresponds to `fdo:sampleSize` in `fdo`.

[4] http://www.w3.org/ns/dcat, with namespace `dct`

[5] Schema publicly available at https://fairnessdatasets.dei.unipd.it/schema/; RDF graph publicly available at https://zenodo.org/record/6518370#.YnOSKFTMJhF.

Year. Last known update to the dataset. For resources whose collection and curation are ongoing (e.g., Framingham) we write "present". Corresponds to `dct:modified`.

Sensitive features. Sensitive attributes in the dataset. These are typically explicitly annotated, but may include implicit ones, such as textual references to people and their demographics in text datasets. References to gender, for instance, can easily be retrieved from English-language text corpora based on intrinsically gendered words, such as she, man, aunt. Corresponds to `fdo:sensitiveFeature`.

Link. A link to the website where the resource can be downloaded or requested. Corresponds to `dcat:landingPage`.

Further information. Reference to works and web pages describing the dataset.

Following the algorithmic fairness literature, we define sensitive features as encoding membership to groups that are salient for society and have some special protection based on the law, including race, ethnicity, sex, gender, and age. We may occasionally stretch this definition and report features considered sensitive in some works, such as political leaning or education, so long as they reflect essential divisions in society. We also report domain-specific attributes considered sensitive in a given context, such as language for Section 203 determinations or brand ownership for Amazon Recommendations. We follow the language of the available documentation for the names and values of sensitive features, including distinctions between race and ethnicity. For datasets that report geographic information at any granularity (GPS coordinates, neighbourhoods, countries) we report "geography" among the sensitive attributes. If an article considers features to be sensitive in an arbitrary fashion (e.g., sepal width in the Iris dataset), we do not report it in the respective field.

For the dataset domain, we follow the area-category taxonomy defined by Scimago,[6] with the addition of "news", "social media", "social networks", "sports" and "food". Table 3.2 contains a summary of the included datasets through this domain-based taxonomy. Tasks in the fairness literature were labeled via open coding. The final taxonomy is detailed in Section 3.4.

3.2.2 Inclusion Criteria

Our aim is to compile and publish standardized documentation on the datasets employed in the most important research venues for algorithmic fairness research. To this end, we consider (1) every article published in the proceedings of domain-specific conferences such

[6]https://www.scimagojr.com/journalrank.php

as the ACM Conference on Fairness, Accountability, and Transparency (FAccT), and the AAAI/ACM Conference on Artificial Intelligence, Ethics and Society (AIES); (2) every article published in proceedings of well-known machine learning and data mining conferences, including the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), the Conference on Neural Information Processing Systems (NeurIPS), the International Conference on Machine Learning (ICML), the International Conference on Learning Representations (ICLR), the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD); (3) every article available from Past Network Events and Older Workshops and Events of the FAccT network.[7] We consider the period from 2014, the year of the first workshop on Fairness, Accountability, and Transparency in Machine Learning, to June 2021, thus including works presented at FAccT, ICLR, AIES, and CVPR in 2021.[8]

To target works of algorithmic fairness, we select a subsample of these articles whose titles contain either of the following strings, where the star symbol represents the wildcard character: `*fair*` (targeting e.g. fairness, unfair), `*bias*` (biased, debiasing), `discriminat*` (discrimination, discriminatory), `*equal*` (equality, unequal), `*equit*` (equity, equitable), disparate (disparate impact), `*parit*` (parity, disparities). These selection criteria are centered on equity-based notions of fairness, typically operationalized by measuring disparity in some algorithmic property across individuals or groups of individuals. Through manual inspection by two authors, we discard articles where these keywords are used with a different meaning. Discarded works, for example, include articles on handling pose distribution bias [883], compensating selection bias to improve accuracy without paying attention to sensitive attributes [432], enhancing desirable discriminating properties of models [136], or generally focused on model performance [503, 885]. This leaves us with 558 articles.

From the articles that pass this initial screening, we select datasets treated as important data artifacts, either being used to train/test an algorithm or undergoing a data audit, i.e., an in-depth analysis of different properties. We produce a data brief for these datasets by (1) reading the information provided in the surveyed articles, (2) consulting the references provided, and (3) reviewing scholarly articles or official websites found by querying popular search engines with the dataset name. We discard the following:

- Word Embeddings (WEs). We only consider the corpora they are trained on, provided that WEs are trained as part of a given work and not taken off the shelf;
- toy datasets, i.e., simulations with no connection to real-world processes, unless they are used in more than one article, which we take as a sign of importance in the field;

[7] https://facctconference.org/network/

[8] We are working on an update covering more recent work, including articles presented at the ACM conference on Equity and Access in Algorithms, Mechanisms, and Optimization.

- auxiliary resources that are only used as a minor source of ancillary information, such as the percentage of US residents in each state;
- datasets for which the available information is insufficient. This happens very rarely when points (1), (2), and (3) outlined above result in little to no information on the curators, purposes, features, and format of a dataset. For popular datasets, this is never the case.

For each of the 226 datasets satisfying the above criteria, we produce a data brief, available in Appendix A.

3.2.3 Related Work

Algorithmic fairness surveys

Multiple surveys on algorithmic fairness have been published in the literature [119, 552, 634]. These works typically focus on describing and classifying important measures of algorithmic fairness and methods to enhance it. Some articles also discuss sources of bias [552], software packages and projects that address fairness in ML [119], or describe selected subfields of algorithmic fairness [634]. Datasets are typically not emphasized in these works, which is also true of domain-specific surveys on algorithmic fairness, focused e.g. on ranking [639], Natural Language Processing (NLP) [749] and computational medicine [749]. As an exception, Pessach and Shmueli [634] and Zehlike et al. [860] list and briefly describe 12 popular algorithmic fairness datasets, and 19 datasets employed in fair ranking research, respectively.

Data studies

The work most closely related to our critical review is Le Quy et al. [485]. The authors perform a detailed analysis of 15 tabular datasets used in works of algorithmic fairness, listing important metadata (e.g. domain, protected attributes, collection period, and location), and carrying out an exploratory analysis of the probabilistic relationship between features. Our work complements it by placing more emphasis on (1) a rigorous methodology for the inclusion of resources, (2) a wider selection of (more than 200) datasets spanning different data types, including text, image, timeseries, and tabular data, (3) a fine-grained evaluation of domains and tasks associated with each dataset. Different goals of the research community, such as selecting appropriate resources for experimentation and data studies, can benefit from the breadth and depth of both works.

Other works analyzing multiple datasets along specific lines have been published in recent years. Crawford and Paglen [181] focus on resources commonly used as training sets in computer vision, with attention to associated labels and underlying taxonomies. Fabbrizzi et al. [246] also consider computer vision datasets, describing the types of bias that affect them, along with methods to discover and measure bias, while Scheuerman et al. [700] analyze the values encoded in their documentation. Koch et al. [457] study the data employed in machine learning research and show a concentration of work on a small number of benchmark datasets curated at few well-resourced institutions. Peng et al. [630] analyze ethical concerns in three popular face and person recognition datasets, arising from derivative datasets and models, lack of clarity of licenses, and dataset management practices. Geiger et al. [293] evaluate transparency in the documentation of labeling practices employed in over 100 datasets about Twitter. Leonelli and Tempini [492] study practices of collection, cleaning, visualization, sharing, and analysis in a variety of research domains. Romei and Ruggieri [681] survey techniques and data for discrimination analysis, focused on measuring, rather than enforcing, equity in human processes.

A different, yet related, family of articles provides deeper analyses of single datasets. Prabhu and Birhane [644] focus on Imagenet (ILSVRC 2012) which they analyze along the lines of consent, problematic content, and individual re-identification. Kizhner et al. [450] study issues of representation in the Google Arts and Culture project across countries, cities and institutions. Some works provide datasheets for a given resource, such as CheXpert [287] and the BookCorpus [42]. Among the popular fairness datasets, COMPAS has drawn scrutiny from multiple works, analyzing its numerical idiosyncrasies [46] and sources of bias [43]. Ding et al. [213] study numerical idiosyncrasies in the Adult dataset, and propose a novel version, for which they provide a datasheet. Grömping [329] discuss issues resulting from coding errors in German Credit.

Our work combines the breadth of multi-dataset and the depth of single-dataset studies. On one hand, we include numerous resources used in works of algorithmic fairness, analyzing them across multiple dimensions. On the other hand, we identify the most popular resources, compiling their *datasheet* and *nutrition label*, and summarize their benefits and limitations. Moreover, by making our data briefs available, we hope to contribute a useful tool to the research community, favoring further data studies and analyses, as outlined in Section 3.6.

Documentation frameworks

Several data documentation frameworks have been proposed in the literature; three popular ones are described below. *Datasheets for Datasets* [292] are a general-purpose qualitative framework with over 50 questions covering key aspects of datasets, such as motivation, com-

position, collection, preprocessing, uses, distribution, and maintenance. Another qualitative framework is represented by *Data statements* [58], which is tailored for NLP, requiring domain-specific information on language variety and speaker demographics. *Dataset Nutrition Labels* [366] describe a complementary, quantitative framework, focused on numerical aspects such as the marginal and joint distribution of variables. More broadly, recent initiatives focused on ML and AI documentation strongly emphasize data documentation [23, 624].

Popular datasets require close scrutiny; for this reason, we adopt these frameworks, producing three datasheets and nutrition labels for Adult, German Credit, and COMPAS. However, this approach does not scale to a larger documentation effort with limited resources. For this reason, we propose and produce *data briefs*, a lightweight documentation format designed for algorithmic fairness datasets. Data briefs include fields specific to fair ML, such as sensitive attributes and tasks for which the dataset has been used in the algorithmic fairness literature.

3.3 Fairness Domains

Algorithmic fairness concerns arise in any domain where Automated Decision Making (ADM) systems may influence human well-being. Unsurprisingly, the datasets in our critical review reflect a variety of areas where ADM systems are studied or deployed, including criminal justice, education, search engines, online marketplaces, emergency response, social media, medicine, and hiring. In Figure 3.2, we report a subdivision of selected datasets in different macrodomains. We mostly follow the area-category taxonomy of Scimago,[9] departing from it where appropriate. For example, we consider computer vision and linguistics macrodomains of their own, for the purposes of algorithmic fairness, as much fair ML work has been published in both disciplines. Below, we present a description of each macrodomain and its main subdomains, summarized in detail in Table 3.2.

Computer Science. Datasets from this macrodomain are very well represented, comprising *information systems, social media, library and information sciences, computer networks, and signal processing*. *Information systems* heavily feature datasets on search engines for various items such as text, images, worker profiles, and real estate, retrieved in response to queries issued by users (Occupations in Google Images, Scientist+Painter, Zillow Searches, Barcelona Room Rental, Burst, TaskRabbit, Online Freelance Marketplaces, Bing US Queries, Symptoms in Queries). Other datasets represent problems of item recommen-

[9] See the "subject area" and "subject category" drop down menus from https://www.scimagojr.com/journalrank.php, accessed on March 15, 2022

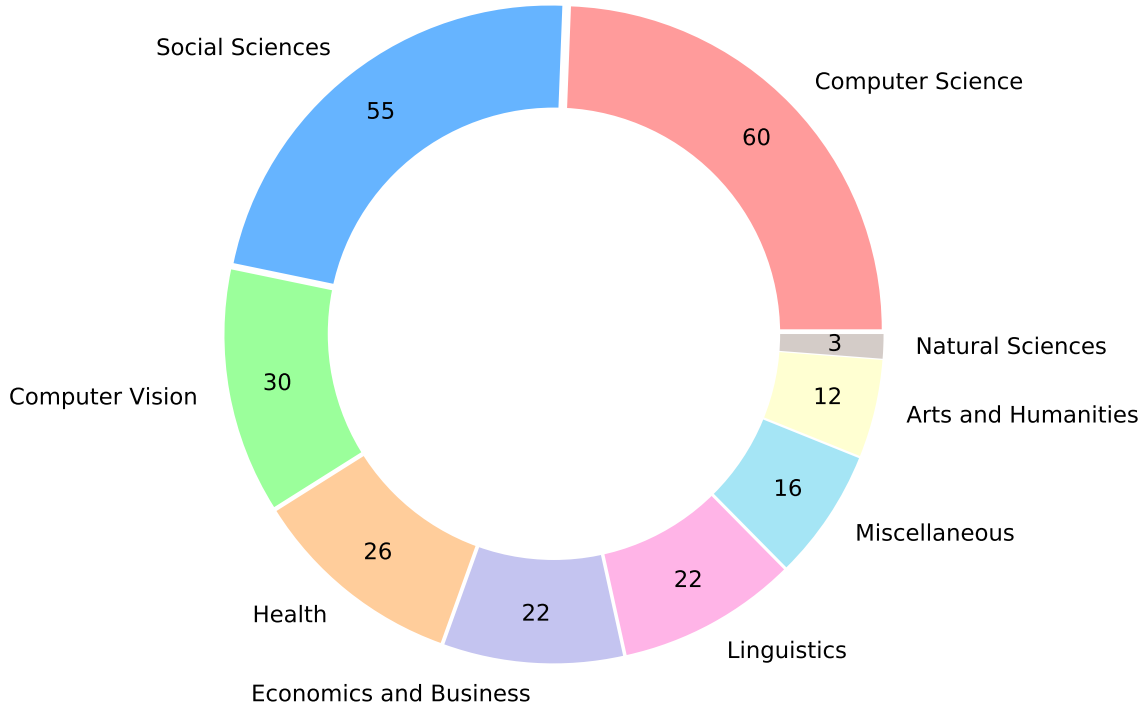


Fig. 3.2 Datasets employed in fairness research span various domains. See Table 3.2 for a detailed breakdown.

dation, covering products, businesses, and movies (Amazon Recommendations, Amazon Reviews, Google Local, MovieLens, FilmTrust). The remaining datasets in this subdomain represent knowledge bases (Freebase15k-237, Wikidata) and automated screening systems (CVs from Singapore, Pymetrics Bias Group). Datasets from *social media* that are not focused on links and relationships between people are also considered part of computer science in this critical review. These resources often focus on text, powering tools, and analyses of hate speech and toxicity (Civil Comments, Twitter Abusive Behavior, Twitter Offensive Language, Twitter Hate Speech Detection, Twitter Online Harassment), dialect (TwitterAAE), and political leaning (Twitter Presidential Politics). Twitter is by far the most represented platform, while datasets from Facebook (German Political Posts), Steeemit (Steemit), Instagram (Instagram Photos), Reddit (RtGender, Reddit Comments), Fitocracy (RtGender), and YouTube (YouTube Dialect Accuracy) are also present. Datasets from *library and information sciences* are mainly focused on academic collaboration networks (Cora Papers, CiteSeer Papers, PubMed Diabetes Papers, ArnetMiner Citation Network, 4area, Academic Collaboration Networks), except for a data set on peer review of scholarly manuscripts (Paper-Reviewer Matching).

Social Sciences. Datasets from social sciences are also plentiful, spanning *law, education, social networks, demography, social work, political science, transportation, sociology* and *urban studies*. *Law* datasets are mostly focused on recidivism (Crowd Judgement, COMPAS, Recidivism of Felons on Probation, State Court Processing Statistics, Los Angeles City Attorney's Office Records) and crime prediction (Strategic Subject List, Philadelphia Crime Incidents, Stop, Question and Frisk, Real-Time Crime Forecasting Challenge, Dallas Police Incidents, Communities and Crime), with granularity that spans the range from individuals to communities. In the area of *education* we find datasets that encode application processes (Nursery, IIT-JEE), student performance (Student, Law School, UniGe, ILEA, US Student Performance, Indian Student Performance, EdGap, Berkeley Students), including attempts at automated grading (Automated Student Assessment Prize), and placement information after school (Campus Recruitment). Some datasets on student performance support studies of differences across schools and educational systems, for which they report useful features (Law School, ILEA, EdGap), while the remaining datasets are more focused on differences in the individual condition and outcome for students, typically within the same institution. Datasets about *social networks* mostly concern online social networks (Facebook Ego-networks, Facebook Large Network, Pokec Social Network, Rice Facebook Network, Twitch Social Networks, University Facebook Networks), except for High School Contact and Friendship Network, which also features offline relations. *Demography* datasets comprise census data from different countries (Dutch Census, Indian Census, National Longitudinal Survey of Youth, Section 203 determinations, US Census Data (1990)). Datasets from *social work* cover complex personal and social problems, including child maltreatment prevention (Allegheny Child Welfare), emergency response (Harvey Rescue), and drug abuse prevention (Homeless Youths' Social Networks, DrugNet). Resources from *political science* describe registered voters (North Carolina Voters), electoral precincts (MGGG States), polling (2016 US Presidential Poll), and sortition (Climate Assembly UK). *Transportation* data summarizes trips and fares from taxis (NYC Taxi Trips, Shanghai Taxi Trajectories), ride-hailing (Chicago Ridesharing, Ride-hailing App), and bike sharing services (Seoul Bike Sharing), along with public transport coverage (Equitable School Access in Chicago). *Sociology* resources summarize online (Libimseti) and offline dating (Columbia University Speed Dating). Finally, we assign SafeGraph Research Release to *urban studies*.

Computer Vision. This is an area of early success for artificial intelligence, where fairness typically concerns learned representations and equality of performance across classes. The selected articles feature several popular datasets on image classification (ImageNet, MNIST, Fashion MNIST, CIFAR), visual question answering (Visual Question Answering), segmentation, and captioning (MS-COCO, Open Images Dataset). We find more than ten

face analysis datasets (Labeled Faces in the Wild, UTK Face, Adience, FairFace, IJB-A, CelebA, Pilot Parliaments Benchmark, MS-Celeb-1M, Diversity in Faces, Multi-task Facial Landmark, Racial Faces in the Wild, BUPT Faces), including one from experimental psychology (FACES), for which fairness is most often intended as the robustness of classifiers across different subpopulations, without much regard for downstream benefits or harms to these populations. Synthetic images are popular to study the relationship between fairness and disentangled representations (dSprites, Cars3D, shapes3D). Similar studies can be conducted on datasets with spurious correlations between subjects and backgrounds (Waterbirds, Benchmarking Attribution Methods) or gender and occupation (Athletes and health professionals). Finally, the Image Embedding Association Test dataset is a fairness benchmark to study biases in image embeddings across religion, gender, age, race, sexual orientation, disability, skin tone, and weight. It is worth noting that this significant proportion of computer vision datasets is not an artifact of including CVPR in the list of candidate conferences, which contributed just five additional datasets (Multi-task Facial Landmark, Office31, Racial Faces in the Wild, BUPT Faces, Visual Question Answering).

Health. This macrodomain, comprising medicine, psychology, and pharmacology displays a notable diversity of subdomains interested by fairness concerns. The specialties represented in the datasets are maily medical, including *public health* (Antelope Valley Networks, Willingness-to-Pay for Vaccine, Kidney Matching, Kidney Exchange Program), *cardiology* (Heart Disease, Arrhythmia, Framingham), *endocrinology* (Diabetes 130-US Hospitals, Pima Indians Diabetes Dataset), *health policy* (Heritage Health, MEPS-HC). Specialties such as *radiology* (National Lung Screening Trial, MIMIC-CXR-JPG, CheXpert) and *dermatology* (SIIM-ISIC Melanoma Classification, HAM10000) feature several image datasets for their strong connections to medical imaging. Other specialties include *critical care medicine* (MIMIC-III), *neurology* (Epileptic Seizures), *pediatrics* (Infant Health and Development Program), *sleep medicine* (Apnea), *nephrology* (Renal Failure), *pharmacology* (Warfarin) and *psychology* (Drug Consumption, FACES). These datasets are often extracted from care data from multiple medical centers to study automated diagnosis. Resources derived from longitudinal studies, including Framingham and Infant Health and Development Program are also present. Works of algorithmic fairness in this domain are typically concerned with obtaining models with similar performance for patients across race and sex.

Linguistics. In addition to the textual resources we already described, such as those derived from social media, several datasets employed in algorithmic fairness literature can be assigned to the domain of linguistics and Natural Language Processing (NLP). There are many examples of resources curated to be fairness benchmarks for different tasks, including machine translation (Bias in Translation Templates), sentiment analysis (Equity

Evaluation Corpus), coreference resolution (Winogender, Winobias, GAP Coreference), named entity recognition (In-Situ), language models (BOLD) and word embeddings (WEAT). Other datasets have been considered for their size and importance for pretraining text representations (Wikipedia dumps, One billion word benchmark, BookCorpus, WebText) or their utility as NLP benchmarks (GLUE, Business Entity Resolution). Speech recognition resources have also been considered (TIMIT).

Economics and Business. This macrodomain comprises datasets from *economics*, *finance*, *marketing*, and *management information systems*. *Economics* datasets mostly consist of census data focused on wealth (Adult, US Family Income, Poverty in Colombia, Costarica Household Survey) and other resources that summarize employment (ANPE), tariffs (US Harmonized Tariff Schedules), insurance (Italian Car Insurance), and division of goods (Spliddit Divide Goods). *Finance* resources feature data on microcredit and peer-to-peer lending (Mobile Money Loans, Kiva, Prosper Loans Network), mortgages (HMDA), loans (German Credit, Credit Elasticities), credit scoring (FICO) and default prediction (Credit Card Default). *Marketing* datasets describe marketing campaigns (Bank Marketing), customer data (Wholesale) and advertising bids (Yahoo! A1 Search Marketing). Finally, datasets from *management information systems* summarize information on automated hiring (CVs from Singapore, Pymetrics Bias Group) and employee retention (IBM HR Analytics).

Miscellaneous. This macrodomain contains several datasets originating from the *news* domain (Yow news, Guardian Articles, Latin Newspapers, Adressa, Reuters 50 50, New York Times Annotated Corpus, TREC Robust04). Other resources include datasets on food (Sushi), sports (Fantasy Football, FIFA 20 Players, Olympic Athletes), and toy datasets (Toy Dataset 1–4).

Arts and Humanities. In this area, we mostly find *literature* datasets, which contain text from literary works (Shakespeare, Curatr British Library Digital Corpus, Victorian Era Authorship Attribution, Nominees Corpus, Riddle of Literary Quality), which are typically studied with NLP tools. Other datasets in this domain include domain-specific information systems about books (Goodreads Reviews), *movies* (MovieLens) and *music* (Last.fm, Million Song Dataset, Million Playlist Dataset).

Natural Sciences. This domain is represented with three datasets from *biology* (iNaturalist), *biochemestry* (PP-Pathways) and *plant science*, with the classic Iris dataset.

In general, many of these datasets encode fundamental human activities where algorithms and ADM systems have been studied and deployed. Alertness and attention to equity appear especially important in specific domains, including social sciences, computer science, medicine, and economics. Here, the potential for impact may result in large benefits, but also great harm, particularly for vulnerable populations and minorities, more likely to be neglected

during the design, training, and testing of an ADM. After concentrating on domains, in the next section we analyze the variety of tasks studied in works of algorithmic fairness and supported by these datasets.

Domain	Sample datasets
Computer Science	
social media	
toxicity and hate speech	Civil Comments, Wikipedia Toxic Comments, Twitter offensive language
political leaning	Twitter Presidential Politics
dialect	TwitterAAE
library and information sciences	
collaboration networks	Paper-Reviewer Matching, 4area, ArnetMiner Citation Network
peer review	Paper-Reviewer Matching
information systems	
search engines	Online Freelance Marketplaces, Bing US Queries, Symptoms in Queries
recommender systems	Amazon Recommendations, Amazon Reviews, MovieLens
knowledge bases	Freebase15k-237, Wikidata
computer networks	KDD Cup 99
pattern recognition	Internet Ads
signal processing	Vehicle
Social Sciences	
urban studies	SafeGraph Research Release
social networks	University Facebook Networks, Pokec Social Network, Rice Facebook Network
demography	US Census Data (1990), Dutch Census, National Longitudinal Survey of Youth
sociology	Columbia University Speed Dating, Libimseti
law	
recidivism prediction	COMPAS, Recidivism of Felons on Probation, State Court Processing Statistics
crime prediction	Communities and Crime, Stop, Question and Frisk, Strategic Subject List
political science	
registered voters	North Carolina Voters
electoral precincts	MGGG States
polling	2016 US Presidential Poll
sortition	Climate Assembly UK
education	
application processes	Nursery, IIT-JEE
student performance	Student, Law School, UniGe
post-education placement	Campus Recruitment
social work	
child maltreatment prevention	Allegheny Child Welfare
emergency response	Harvey Rescue

drug abuse prevention	Homeless Youths' Social Networks, DrugNet
transportation	
taxi trips	NYC Taxi Trips, Shanghai Taxi Trajectories
ride hailing	Chicago Ridesharing, Ride-hailing App
bike sharing	Seoul Bike Sharing
public transport	Equitable School Access in Chicago
Computer Vision	
general purpose	ImageNet, MNIST, CIFAR
face analysis	CelebA, Pilot Parliaments Benchmar, FairFace
synthetic	dSprites, Cars3D, shapes3D
Health	
sleep medicine	Apnea
critical care medicine	MIMIC-III
public health	Kidney Exchange Program, Willingness-to-Pay for Vaccine, Kidney Matching
cardiology	Arrhythmia, Heart Disease, Framingham
neurology	Epileptic Seizures
pediatrics	Infant Health and Development Program (IHDP)
dermatology	HAM10000, SIIM-ISIC Melanoma Classification
medicine	Stanford Medicine Research Data Repository
pharmacology	Warfarin
endocrinology	Diabetes 130-US Hospitals, Pima Indians Diabetes Dataset (PIDD)
nephrology	Renal Failure
radiology	CheXpert, MIMIC-CXR-JPG, National Lung Screening Trial (NLST)
health policy	Heritage Health, MEPS-HC
applied psychology	Drug Consumption
experimental psychology	FACES
Economics and Business	
economics	
census	Adult, US Family Income, Poverty in Colombia
employment	ANPE
tariffs	US Harmonized Tariff Schedule
insurance	Italian Car Insurance
division of goods	Spliddit Divide Goods
finance	
peer-to-peer lending	Mobile Money Loans, Kiva, Prosper Loans Network
mortgages	HMDA
credit scoring	FICO
other credit	German Credit, Credit Card Default, Credit Elasticities
marketing	
marketing campaigns	Bank Marketing
advertising bids	Yahoo! A1 Search Marketing, Wholesale

management information systems	
automated hiring	Pymetrics Bias Group, CVs from Singapore
employee retention	IBM HR Analytics
Linguistics	
general purpose	Wikipedia dumps, One billion word benchmark, BookCorpus
fairness benchmarks	Bias in Translation Templates, Equity Evaluation Corpus, Winogender
Arts and Humanities	
music	Million Playlist Dataset (MPD), Million Song Dataset (MSD), Last.fm
literature	Goodreads Reviews, Riddle of Literary Quality, Nominees Corpus
movies	MovieLens, FilmTrust
Natural Sciences	
biology	iNaturalist Datasets
biochemestry	PP-Pathways
plant science	Iris
Miscellaneous	
news	TREC Robust04, New York Times Annotated Corpus, Reuters 50 50
sports	Fantasy Football, FIFA 20 Players, Olympic Athletes
food	Sushi

Table 3.2 A selection of datasets through the lens of the domain taxonomy.

3.4 Fairness Tasks and Settings

Researchers and practitioners are showing an increasing interest in algorithmic fairness, proposing solutions for many different *tasks*, including fair classification, regression, and ranking. At the same time, the academic community is developing an improved understanding of important challenges that run across different tasks in the algorithmic fairness space [155], also thanks to practitioner surveys [369] and studies of specific legal challenges [17]. To exemplify, the presence of noise corrupting labels for sensitive attributes represents a challenge that may apply across different tasks, including fair classification, regression, and ranking. We refer to these challenges as *settings*, and we describe them in the second part of this section. In this section, we provide an overview of common tasks and settings studied on these datasets, showing their variety and diversity. Table 3.3 summarizes these tasks, listing the three most used datasets for each task. When describing a task, we explicitly highlight datasets that are particularly relevant to it, even when outside of the top three.

3.4.1 Task

Fair classification [108, 220] is by far the most common task. Typically, it involves equalizing some measure of interest across subpopulations, such as recall, precision, or accuracy for different racial groups. On the other hand, individually fair classification focuses on the idea that similar individuals (low distance in the covariate space) should be treated similarly (low distance in the outcome space), often formalized as a Lipschitz condition. Unsurprisingly, the most common datasets for fair classification are the most popular overall (section 3.1), i.e., Adult, COMPAS, and German Credit.

Fair regression [64] concentrates on models that predict a real-valued target, requiring the average loss to be balanced across groups. Individual fairness in this context may require losses to be as uniform as possible across all individuals. Fair regression is a less popular task, often studied on the Communities and Crime dataset, where the task is to predict the rate of violent crimes in different communities.

Fair ranking [843] requires the ordering of candidate items based on their relevance to a current need. Fairness in this context may concern both the people producing the items that are being ranked (e.g. artists) and those consuming the items (users of a music streaming platform). It is typically studied in applications of recommendation (MovieLens, Amazon Recommendations, Last.fm, Million Song Dataset, Adressa) and search engines (Yahoo! c14B Learning to Rank, Microsoft Learning to Rank, TREC Robust04).

Fair matching [456] is similar to ranking as they are both tasks defined on two-sided markets. However, this task is focused on highlighting and matching pairs of items on both

sides of the market, without emphasis on the ranking component. Datasets for this task are from diverse domains, including dating (Libimseti, Columbia University Speed Dating) transportation (NYC Taxi Trips, Ride-hailing App) and organ donation (Kidney Matching, Kidney Exchange Program).

Fair risk assessment [171] studies algorithms that score instances in a dataset according to a predefined type of risk. Relevant domains include healthcare and criminal justice. Key differences with respect to classification are an emphasis on real-valued scores rather than labels, and awareness that the risk assessment process can lead to interventions that impact the target variable. For this reason, fairness concerns are often defined in a counterfactual fashion. The most popular dataset for this task is COMPAS, followed by datasets from medicine (IHDP, Stanford Medicine Research Data Repository), social work (Allegheny Child Welfare), Economics (ANPE) and Education (EdGap).

Fair representation learning [183] concerns the study of features learned by models as intermediate representations for inference tasks. A popular line of work in this space, called *disentaglement*, aims to learn representations where a single factor of import corresponds to a single feature. Ideally, this approach should select representations where sensitive attributes cannot be used as proxies for target variables. Cars3D and dSprites are popular datasets for this task, consisting of synthetic images depicting controlled shape types under a controlled set of rotations. Post-processing approaches are also applicable to obtain fair representations from biased ones via debiasing.

Fair clustering [146] is an unsupervised task that deals with the division of a sample into homogeneous groups. Fairness may be intended as an equitable representation of protected subpopulations in each cluster, or in terms of average distance from the cluster center. While Adult is the most common dataset for problems of fair clustering, other resources often used for this task include Bank Marketing, Diabetes 130-US Hospitals, Credit Card Default and US Census Data (1990).

Fair anomaly detection [864], also called **outlier detection** [192], aims to identify surprising or anomalous points in a dataset. Fairness requirements involve equalizing salient quantities (e.g. acceptance rate, recall, precision, distribution of anomaly scores) across populations of interest. This problem is particularly relevant for members of minority groups, who, in the absence of specific attention to dataset inclusivity, are less likely to fit the norm in the feature space.

Fair districting [705] is the division of a territory into electoral districts for political elections. Fairness notions brought forth in this space are either outcome-based, requiring that seats earned by a party roughly match their share of the popular vote, or procedure-based, ignoring outcomes and requiring that counties or municipalities are split as little as possible.

MGGG States is a reference resource for this task, providing aggregated precinct-level information about demographics and political leaning of voters in US districts.

Fair task assignment and **truth discovery** [307, 502] are different subproblems in the same area, focused on the subdivision of work and the aggregation of answers in crowdsourcing. Here fairness may be intended concerning errors in the aggregated answer, requiring errors to be balanced across subpopulations of interest, or in terms of the work load imposed to workers. A dataset suitable for this task is Crowd Judgement, which contains crowd-sourced recidivism predictions.

Fair spatio-temporal process learning [711] focuses on the estimation of models for processes that evolve in time and space. Selected applications include crime forecasting (Real-Time Crime Forecasting Challenge, Dallas Police Incidents) and disaster relief (Harvey Rescue), with fairness requirements focused on equalization of performance across different neighborhoods and special attention to their racial composition.

Fair graph diffusion [254] models and optimizes the propagation of information and influence over networks and its probability of reaching individuals of different sensitive groups. Applications include obesity prevention (Antelope Valley Networks) and drug use prevention (Homeless Youths' Social Networks). **Fair graph augmentation** [664] is a similar task, defined in graphs that define access to resources based on existing infrastructure (e.g. transportation), which can be augmented under a budget to increase equity. This task has been proposed to improve school access (Equitable School Access in Chicago) and information availability in social networks (Facebook100).

Fair resource allocation/subset selection [30, 380] can often be formalized as a classification problem with constraints on the number of positives. Fairness requirements are similar to those of classification. Subset selection may be employed to choose a group of people from a wider set for a given task (US Federal Judges, Climate Assembly UK). Resource allocation concerns the division of goods (Spliddit Divide Goods) and resources (ML Fairness Gym, German Credit).

Fair data summarization [120] refers to presenting a summary of datasets that is equitable to subpopulations of interest. It may involve finding a small subset representative of a larger dataset (strongly linked to subset selection) or selecting the most important features (dimensionality reduction). Approaches for this task have been applied to select a subset of images (Scientist+Painter) or customers (Bank Marketing), that represent the underlying population across sensitive demographics.

Fair data generation [837] deals with generating "fair" data points and labels, which can be used as training or test sets. Approaches in this space may be used to ensure an equitable representation of protected categories in data generation processes learned from

biased datasets (CelebA, IBM HR Analytics), and to evaluate biases in existing classifiers (MS-Celeb-1M). Data generation may also be limited to synthesizing artificial sensitive attributes [102].

Fair graph mining [425] focuses on representations and prediction tasks on graph structures. Fairness may be defined either as a lack of bias in representations, or with respect to a final inference task defined on the graph. Fair graph mining approaches have been applied to knowledge bases (Freebase15k-237, Wikidata), collaboration networks (CiteSeer Paper, Academic Collaboration Networks) and social network datasets (Facebook Large Network, Twitch Social Networks).

Fair pricing [422] concerns learning and deploying an optimal pricing policy for revenue while maintaining equity in access to services and consumer welfare across sensitive groups. Datasets employed in fair pricing are from the economics (Credit Elasticities, Italian Car Insurance), transportation (Chicago Ridesharing), and public health domains (Willingness-to-Pay for Vaccine).

Fair advertising [121] is also concerned with access to goods and services. It comprises both bidding strategies and auction mechanisms which may be modified to reduce discrimination with respect to the gender or race composition of the audience that sees an ad. One publicly available dataset for this subtask is Yahoo! A1 Search Marketing.

Fair routing [650] is the task of suggesting an optimal path from a starting location to a destination. For this task, experimentation has been carried out on a semi-synthetic traffic dataset (Shanghai Taxi Trajectories). The proposed fairness measure requires equalizing the driving cost per customer between all drivers.

Fair entity resolution [175] is a task that focuses on deciding whether multiple records refer to the same entity, which is useful, for example, for the construction and maintenance of knowledge bases. Business Entity Resolution is a proprietary dataset for fair entity resolution, where constraints of performance equality across chain and non-chain businesses can be tested. Winogender and Winobias are publicly available datasets developed to study gender biases in pronoun resolution.

Fair sentiment analysis [448] is a well-established instance of fair classification, where text snippets are typically classified as positive, negative, or neutral, depending on the sentiment they express. Fairness is intended with respect to the entities mentioned in the text (e.g. men and women). The central idea is that the estimated sentiment for a sentence should not change if female entities (e.g. “her”, “woman”, “Mary”) are substituted with their male counterparts (“him”, “man”, “James”). The Equity Evaluation Corpus is a benchmark developed to assess gender and race bias in sentiment analysis models.

Bias in Word Embeddings (WEs) [83] is the study of undesired semantics and stereotypes captured by vectorial representations of words. WEs are typically trained on large text corpora (Wikipedia dumps) and audited for associations between gendered words (or other words connected to sensitive attributes) and stereotypical or harmful concepts, such as the ones encoded in WEAT.

Bias in Language Models (LMs) [87] is, quite similarly, the study of biases in LMs, which are flexible models of human language based on contextualized word representations, which can be employed in a variety of linguistics and NLP tasks. LMs are trained on large text corpora from which they may learn spurious correlations and stereotypes. The BOLD dataset is an evaluation benchmark for LMs, based on prompts that mention different socio-demographic groups. LMs complete these prompts into full sentences, which can be tested along different dimensions (sentiment, regard, toxicity, emotion, and gender polarity).

Fair Machine Translation (MT) [742] concerns the automatic translation of text from a source language into a target language. MT systems can exhibit gender biases, such as a tendency to translate gender-neutral pronouns from the source language into gendered pronouns of the target language in accordance with gender stereotypes. For example, a "nurse" mentioned in a gender-neutral context in the source sentence may be rendered with feminine grammar in the target language. Bias in Translation Templates is a set of short templates to test such biases.

Fair speech recognition [760] requires an accurate annotation of spoken language into text across different demographics. YouTube Dialect Accuracy is a dataset developed to audit the accuracy of YouTube's automatic captions across two genders and five dialects of English. Similarly, TIMIT is a classic speech recognition dataset annotated with American English dialect and gender of speaker.

3.4.2 Setting

As noted at the beginning of this section, there are several *settings* (or challenges) that run across different tasks described above. Some of these settings are specific to fair ML, such as ensuring fairness across an exponential number of groups, or in the presence of noisy labels for sensitive attributes. Other settings are connected with common ML challenges, including few-shot and privacy-preserving learning. Below, we describe common settings encountered in selected articles. Most of these settings are tested on fairness datasets which are popular overall, i.e. Adult, COMPAS, and German Credit. We highlight situations where this is not the case, potentially due to a given challenge arising naturally in some other dataset.

Rich-subgroup fairness [434] is a setting where fairness properties are required to hold not only for a limited number of protected groups but also in an exponentially large number

of subpopulations. This line of work represents an attempt to bridge the normative reasoning underlying individual and group fairness.

Fairness under unawareness is a general expression to indicate problems where sensitive attributes are missing [139], encrypted [441] or corrupted by noise [480]. These problems respond to real-world challenges related to the confidential nature of protected attributes that individuals may wish to hide, encrypt, or obfuscate. This setting is most commonly studied on highly popular fairness datasets (Adult, COMPAS), moderately popular ones (Law School and Credit Card Default), and a dataset about home mortgage applications in the US (HMDA).

Limited-label fairness comprises settings with limited information on the target variable, including situations where labeled instances are few [403], noisy [810], or only available in aggregate form [688].

Robust fairness problems arise under perturbations to the training set [379], adversarial attacks [589] and dataset shift [724]. This line of research is often connected with work on robust machine learning, extending the stability requirements beyond accuracy-related metrics to fairness-related ones.

Dynamical fairness [187, 508] entails repeated decisions in changing environments, potentially affected by the very algorithm that is being studied. Works in this space study the co-evolution of algorithms and populations on which they act over time. For example, an algorithm that achieves equality of acceptance rates across protected groups in a static setting may generate further incentives for the next generation of individuals from historically disadvantaged groups. Popular resources for this setting are FICO and the ML Fairness GYM.

Preference-based fairness [855] denotes work informed, explicitly or implicitly, by the preferences of stakeholders. For people subjected to a decision, this is related to notions of envy-freeness and loss aversion [13]; alternatively, policy-makers can express indications on how to trade-off different fairness measures [871], or experts can provide demonstrations of fair outcomes [285].

Multi-stage fairness [525] refers to settings where several decision makers coexist in a compound decision-making process. Decision makers, both humans and algorithms, can act with different levels of coordination. A fundamental question in this setting is how to ensure fairness under the composition of different decision mechanisms.

Fair few-shot learning [877] aims to develop fair ML solutions in the presence of few data samples. The problem is closely related to, and possibly solved by, **fair transfer learning** [172] where the goal is to exploit the knowledge gained on a problem to solve a different but related one. Datasets where this setting arises naturally are Communities and

Crime, where one may restrict the training set to a subset of US states, and Mobile Money Loans, which consists of data from different African countries.

Fair private learning [33, 399] studies the interplay between privacy-preserving mechanisms and fairness constraints. Work in this space considers the equity of machine learning models designed to avoid the leakage of information about individuals in the training set. Common domains for datasets employed in this setting are face analysis (UTK Face, FairFace, Diversity in Face) and medicine (CheXpert, SIIM-ISIC Melanoma Classification, MIMIC-CXR-JPG).

Additional settings that are less common include **fair federated learning** [500], where algorithms are trained across multiple decentralized devices, **fair incremental learning** [876], where novel classes can be added to the learning problem over time, **fair active learning** [600], allowing for the acquisition of novel information during inference and **fair selective classification** [411], where predictions are issued only if model confidence is above a certain threshold.

Overall, we found a variety of tasks defined on fairness datasets, ranging from generic, such as *fair classification*, to narrow and specifically defined on certain datasets, such as *fair districting* on MGGG States and *fair truth discovery* on Crowd Judgement. Orthogonally to this dimension, many settings or challenges may arise to complicate these tasks, including noisy labels, system dynamics, and privacy concerns. Quite clearly, algorithmic fairness research has been expanding in both directions, by studying a variety of tasks under diverse and challenging settings. In the next section, we analyze the roles played in scholarly works by the datasets we consider.

Table 3.3 Most used datasets by algorithmic fairness task and setting.

Task	Datasets
Fair classification	Adult; COMPAS; German Credit
Fair regression	Communities and Crime; Law School; Student
Fair ranking	MovieLens; German Credit; Kiva
Fair matching	NYC Taxi Trips; Libimseti; Columbia University Speed Dating
Fair risk assessment	COMPAS; Allegheny Child Welfare; Infant Health and Development Program (IHDP)
Fair representation learning	Adult; COMPAS; dSprites
Fair clustering	Adult; Bank Marketing; Diabetes 130-US Hospitals
Fair anomaly detection	Adult; MNIST; Credit Card Default
Fair districting	MGGG States
Fair task assignment	Crowd Judgement; COMPAS
Fair spatio-temporal process learning	Real-Time Crime Forecasting Challenge; Dallas Police Incidents; Harvey Rescue
Fair graph diffusion/augmentation	University Facebook Networks; Antelope Valley Networks; Rice Facebook Network
Fair resource allocation/subset selection	ML Fairness Gym; US Federal Judges; Climate Assembly UK
Fair data summarization	Adult; Student; Credit Card Default
Fair data generation	CelebA; MovieLens; shapes3D
Fair graph mining	MovieLens; Freebase15k-237; PP-Pathways
Fair pricing	Willingness-to-Pay for Vaccine; Credit Elasticities; Italian Car Insurance
Fair advertising	Yahoo! A1 Search Marketing; North Carolina Voters; Instagram Photos
Fair routing	Shanghai Taxi Trajectories
Fair entity resolution	Winogender; Winobias; Business Entity Resolution
Fair sentiment analysis	Popular Baby Names; Equity Evaluation Corpus (EEC); TwitterAAE
Bias in word embeddings	Wikipedia dumps; Word Embedding Association Test (WEAT); Popular Baby Names
Bias in language models	TwitterAAE; BOLD; GLUE
Fair machine translation	Bias in Translation Templates
Fair speech recognition	YouTube Dialect Accuracy; TIMIT
Setting	**Datasets**
Rich-subgroup fairness	Adult; COMPAS; Communities and Crime
Fairness under unawareness	Adult; COMPAS; HMDA
Limited-label fairness	Adult; German Credit; COMPAS
Robust fairness	COMPAS; Adult; MEPS-HC
Dynamical fairness	FICO; ML Fairness Gym; COMPAS
Preference-based fairness	Adult; COMPAS; Toy Dataset 1
Multi-stage fairness	Adult; Heritage Health; Twitter Offensive Language
Fair few-shot learning	Communities and Crime; Toy Dataset 1; Mobile Money Loans
Fair private learning	UTK Face; CheXpert; FairFace
Fair federated learning	Vehicle; Sentiment140; Shakespeare
Fair incremental learning	ImageNet; CIFAR
Fair active learning	Adult; German Credit; Heart Disease
Fair selective classification	CheXpert; CelebA; Civil Comments

3.5 Fairness Roles

Datasets used in algorithmic fairness research can play different roles. For example, some may be used to train novel algorithms, while others are suited to test existing algorithms from a specific point of view. Chapter 7 of Barocas et al. [50], describes six different roles of datasets in machine learning. We adopt their framework to analyze fair ML datasets, adding to the taxonomy two roles that are specific to fairness research.

A source of real data. Although synthetic data sets and simulations may be suited to demonstrate specific properties of a novel method, the usefulness of an algorithm is typically established on data from the real world. More than a sign of immediate applicability to important challenges, good performance on real-world sources of data signals that the researchers did not make up the data to suit the algorithm. This is likely the most common role for fairness datasets, especially common for the ones hosted on the UCI ML repository, including Adult, German Credit, Communities and Crime, Diabetes 130-US Hospitals, Bank Marketing, Credit Card Default, US Census Data (1990). These resources owe their popularity in fair ML research to being a product of human processes and to encoding protected attributes. Quite simply, they are sources of real human data.

A catalyst of domain-specific or task-specific progress. Data sets can stimulate algorithmic insight and bring about domain-specific progress. Civil Comments is a great example of this role, powering the Jigsaw Unintended Bias in Toxicity Classification challenge. The challenge responds to a specific need in the space of automated moderation against toxic comments in online discussion. Early attempts at toxicity detection resulted in models that associate mentions of frequently attacked identities (e.g. gay) with toxicity, due to spurious correlations in training sets. The dataset and associated challenge tackle this issue by providing toxicity ratings for comments, along with labels encoding whether members of a certain group are mentioned, favouring measurement of undesired bias. Many other datasets can play a similar role, including, Winogender, Winobias and the Equity Evaluation Corpus. In a broader sense, COMPAS and the accompanying study [21] have been an important catalyst, not for a specific task, but for fairness research overall.

A way to numerically track progress on a problem. This role is common for machine learning benchmarks that also provide human performance baselines. Algorithmic methods approaching or exceeding these baselines are often considered a sign that the task is "solved" and that harder benchmarks are required [50]. Algorithmic fairness is a complicated, context-dependent, contested construct whose correct measurement is continuously debated. Due to this reason, we are unaware of any dataset with a similar role in the algorithmic fairness literature.

A resource to compare models. Practitioners interested in solving a specific problem may take a large set of algorithms and test them on a group of datasets that are representative of their problem, in order to select the most promising ones. For well-established ML challenges, there are often leaderboards that provide a concise comparison between algorithms for a given task, which may be used for model selection. This setting is rare in the fairness literature, also due to inherent difficulties in establishing a single measure of interest in the field. One notable exception is represented by Friedler et al. [282], who employed a suite of four datasets (Adult, COMPAS, German Credit, Ricci) to compare the performance of four different approaches to fair classification.

A source of pre-training data. Flexible, general-purpose models are often pre-trained to encode useful representations, which are later fine-tuned for specific tasks in the same domain. For example, large text corpora are often employed to train language models and word embeddings, which are later specialized to support a variety of downstream NLP applications. Wikipedia dumps, for example, are often used to train word embeddings and investigate their biases [99, 504, 620]. Several algorithmic fairness works aim to study and mitigate undesirable biases in learned representations. Corpora like Wikipedia dumps are used to obtain representations via realistic pretraining procedures that mimic common machine learning practice as closely as possible.

A source of training data. Models for a specific task are typically learned from training sets that encode relations between features and target variable in a representative fashion. An example from the fairness literature is Large Movie Review, used to train sentiment analysis models, later audited for fairness [504]. For fairness audits, one alternative would be resorting to publicly available models, but sometimes a close control on the training corpus and procedure is necessary. Indeed, it is interesting to study issues of model fairness in relation to biases present in the respective training corpora, which can help explain the causes of bias [99]. Some works measure biases in internal model representations before and after fine-tuning on a training set, and regard the difference as a measure of bias in the training set. Babaeianjelodar et al. [29] employ this approach to measure biases in RtGender, Civil Comments, and datasets from GLUE.

A representative summary of a service. Much important work in the fairness literature is focused on measuring fairness and harms in the real world. This line of work includes audits of products and services, which rely on datasets extracted from the application of interest. Datasets created for this purpose include Amazon Recommendations, Pymetrics Bias Group, Occupations in Google Images, Zillow Searches, Online Freelance Marketplaces, Bing US Queries, YouTube Dialect Accuracy. Several other datasets were originally created

for this purpose and later repurposed in the fairness literature as sources of real data, including Stop Question and Frisk, HMDA, Law School, and COMPAS.

An important source of data. Some datasets acquire a pivotal role in research and industry, to the point of being considered a de-facto standard for a given purpose. This status warrants closer scrutiny of the dataset, through which researchers aim to uncover potential biases and problematic aspects that may impact models and insights derived from the dataset. ImageNet, for example, is a dataset with millions of images across thousands of categories. Since its release in 2011, this resource has been used to train, benchmark, and compare hundreds of computer vision models. Given its status in machine learning research, ImageNet has been the subject of two quantitative investigations analyzing its biases and other problematic aspects in the person subtree, uncovering issues of representation [842] and non-consensuality [644]. A different data bias audit was carried out on SafeGraph Research Release. SafeGraph data captures mobility patterns in the US, with data from nearly 50 million mobile devices obtained and maintained by Safegraph, a private data company. Their recent academic release has become a fundamental resource for pandemic research, to the point of being used by the Centers for Disease Control and Prevention to measure the effectiveness of social distancing measures [577]. To evaluate its representativeness for the general population of the United States, Coston et al. [170] have studied selection biases in this dataset.

In algorithmic fairness research, datasets play a similar role to the one they play in machine learning according to Barocas et al. [50], including training, catalyzing attention, and signalling awareness of common data practices. One notable exception is that fairness datasets are not used to track algorithmic progress on a problem over time, probably due to the fact that there is no consensus on a single measure to be reported. On the other hand, two roles peculiar to fairness research are summarizing a service or product that is being audited, and representing an important resource whose biases and ethical aspects are particularly worthy of attention. We note that these roles are not mutually exclusive and that datasets can play multiple roles. COMPAS, for example, was originally curated to perform an audit of pretrial risk assessment tools and was later used extensively in fair ML research as a source of real human data, becoming, overall, a catalyst for fairness research and debate.

3.6 Chapter Outcomes

In this chapter, we have surveyed algorithmic fairness datasets, analyzing important limitations of three benchmarks, surveying alternative resources, and presenting a critical analysis of the field and its impact across several domains and tasks.

Regarding the first objective of this chapter (**O1**), we rigorously identified Adult, COMPAS, and German Credit as the most used datasets in algorithmic fairness literature and we studied their limitations, summarized in Table 3.1. Their status as de facto fairness benchmarks is probably due to their use in seminal works [107, 629] and influential articles [21] on algorithmic fairness. Once this fame was created, researchers had clear incentives to study novel problems and approaches on these datasets, which, as a result, have become even more established benchmarks in the algorithmic fairness literature [43]. Under close scrutiny, the fundamental merit of these datasets lies in originating from human processes, encoding protected attributes, and having different base rates for the target variable across sensitive groups. Their use in recent works on algorithmic fairness can be interpreted as a signal that the authors have a basic awareness of default data practices in the field and that the data was not made up to fit the algorithm. Overarching claims of significance in real-world scenarios stemming from experiments on these datasets should be met with skepticism. Experiments that claim to extract a sex variable from the German Credit dataset should be considered noisy at best.

As for alternatives, Bao et al. [43] suggest employing well-designed simulations. A complementary avenue, targeted in this chapter, is to enable an informed approach to dataset selection and assist researchers and practitioners in seeking alternative datasets that are relevant to the problem at hand (**O2**). To achieve this goal, we have presented a lightweight format for the documentation of fairness datasets and applied it to a wide range of datasets featured in the top venues for fair ML research (Section 3.2). We make available to the research community the resulting data briefs, providing key information about over 200 datasets employed in algorithmic fairness research. In addition, we contribute a critical analysis of algorithmic fairness research over the past seven years. Fairness datasets originate from a variety of domains, support various tasks, and play different roles in the algorithmic fairness literature, discussed in Sections 3.3–3.5. Overall, we hope that this work will contribute to establishing principled data selection practices in the field, and guide an optimal usage of these resources.

The analyses presented in this chapter are just one contribution enabled by the underlying documentation initiative. Our final aim is to release, update, and maintain a web app for the data briefs, which can be queried by researchers and practitioners to find the most relevant datasets, according to their specific needs.[10] We envision several benefits for the algorithmic fairness and data studies communities, including: (1) informing the choice of datasets for experimental evaluations of fair ML methods, including domain-oriented, task-oriented, and role-oriented search; (2) directing studies of data bias, and other quantitative

[10]This resource is available at https://fairnessdatasets.dei.unipd.it/

and qualitative analyses, including retrospective documentation efforts, toward popular (or otherwise important) resources; (3) identifying areas and sub-problems that are understudied in the algorithmic fairness literature; and (4) supporting multi-dataset studies, focused on resources united by a common characteristic, such as encoding a given sensitive attribute [701], concerning computer vision [246], or being popular in the fairness literature [485].

Chapter 4

Dataset Curation

A principled approach to resource selection is key to an informed utilization of existing datasets. However, certain endeavors of research and industry are highly specific and cannot be pursued with available data alone. Novel resources must be curated as a result. Data curation consists of a set of choices and processes that impact the collection, annotation, handling, and distribution of data. In the algorithmic fairness literature, there are several examples of datasets that were created to favor progress in a specific task, including Winogender [685], Winobias [881], and the Equity Evaluation Corpus [448], designed to study gender biases in algorithms for coreference resolution and sentiment analysis. Novel datasets are also necessary when studying a service or a product from a fresh perspective, and new data is collected as a representative summary. This is a common need for algorithmic audits and analyses of human decision making. Prominent examples include COMPAS [21, 483] and Stop, Question and Frisk,[1] two datasets curated to study decision making in policing and the judicial system in the US.

An algorithmic audit can be characterized as a study of algorithms, products, and services aimed at uncovering meaningful relationships between inputs and outputs. As automation becomes increasingly embedded in society, processes designed to reverse engineer and uncover key aspects of algorithms and automated decision systems are fundamental. Auditing is a central part of fairness, accountability, and transparency, allowing communities to keep technology and decision systems in check and ensure that they are aligned with specific values and requirements. Among many notable works in this area, researchers have audited personalization in search engines [339], price discrimination on e-commerce platforms [340], racial bias in judicial risk assessment [727], sources of bias in political queries on Twitter [472], gender- and race-based differences in the accuracy of face analysis technology [101],

[1] https://www1.nyc.gov/site/nypd/stats/reports-analysis/stopfrisk.page

and radicalization on YouTube [673]. To each of these studies corresponds a data curation effort.

Most of these audits are US-centric in several ways: they study American systems or products (e.g. judicial risk assessment in the US), from cultural and linguistic perspectives (e.g. Democratic/Republic political dichotomy), and through racial lenses (e.g. White/African-American) that are especially relevant in the US. Indeed, other regions can be neglected by algorithmic fairness research, and run the risk of maintaining biased systems and algorithms which systematically harm disadvantaged groups. The car insurance system in Italy, for example, is an important socio-technical system of sizeable and widespread impact within the country. With nearly 40 million vehicles insured annually [223], this is a critical system, which determines large and recurring expenses, affecting the majority of the Italian population. Both access to market and pricing are increasingly determined by algorithms and automated systems, such as actuarial models and comparison websites [416], which require scrutiny through the lens of fairness and equity. The first objective of this chapter is to fill this gap.

O1: Design and collect a dataset to audit key algorithms regulating access and pricing in the Italian car insurance industry through the lens of algorithmic fairness.

It should be noted that algorithmic fairness datasets, even those curated with the best intentions and motivated by the pursuit of equity and justice, may end up harming some populations and hindering progress. Invisibility and exclusion, for example, can be a form of harm: due to historical inequality, some populations and their perspectives are underrepresented in certain domains and datasets [405]. Artifacts, systems, and models built and validated with these datasets can perform suboptimally for under-represented populations, as shown in domains such as health care [607], speech recognition [760], and computer vision [101]. Inclusion, in itself, does not guarantee benefits: personal data can be used against people in several ways [468], from which they can be protected if the data is properly anonymized and consent mechanisms are in place to favor tighter control over one's digital trails [627]. Furthermore, datasets can be used correctly only if properly documented. This is a fundamental aspect of data quality, which favors reproducibility, scientific validity, and harm avoidance [50]. One key feature of algorithmic fairness datasets, requiring accurate documentation, are sensitive attributes and their procurement, for which several approaches are available with trade-offs between scalability, accuracy, and respect of data subject rights [701]. Overall, this is a set of important themes in data curation that are often overlooked, especially during the early stages of dataset design, due to a lack of consideration and clear guidance. In this chapter, we aim to make these concerns visible and practical and extract a

set of best practices. To do so, we compare existing approaches across hundreds of fairness datasets considered in the previous chapter and discuss their advantages and limitations.

O2: Distill a set of best practices for dataset curation with respect to anonymization, consent, inclusivity, labeling, and transparency.

The remainder of this chapter is organized as follows. First, Section 4.1 addresses **O1** and describes the Italian Car Insurance dataset, supporting analyses of access and pricing discrimination in the Italian car insurance market. We provide the motivation and background for this resource, discuss its design and collection, illustrate the main results supported by the data, and discuss the underlying curatorial choices. Next, we broaden our analysis and study curatorial practices across hundreds of algorithmic fairness datasets. In Section 4.2, we tackle **O2** and discuss concerns of re-identification, consent, inclusivity, annotation, and transparency. We describe a range of approaches and consideration to these topics, ranging from negligent to conscientious, and distill a set of best practices for dataset curation. We make these concerns and related desiderata more visible and concrete, to help inform responsible curation of novel fairness resources. Finally, Section 4.3 concludes the chapter with a discussion of its main outcomes and lessons learned.

4.1 The Italian Car Insurance Dataset

This section centers around a dataset we curated to investigate differential pricing and access in the Italian car insurance market. With reference to the domains, tasks, and roles of algorithmic fairness datasets, introduced in Chapter 3, the key characteristics of this resource are the following:

Domain. This dataset, investigating the importance of selected factors in car insurance, squarely belongs to *insurance economics*.

Task. The main supported task is *fair pricing evaluation*, where the relevant notion of "fairness" is derived from the normative framework described in Section 4.1.2.

Role. This dataset represents a *summary of a service*. More in detail, Italian Car Insurance supports a driver-centric analysis of pricing and access to car insurance in Italy, as mediated by popular comparison websites.

Figure 4.1 reports the data brief for this dataset.

4.1.1 Motivation

Car ownership is an important factor for employment and, more broadly, for participation in the economic, social, and political organization of many societies [615, 732]. This may

A.1.101 Italian Car Insurance

- **Description**: this resource was curated to study discriminatory practices in the Italian car insurance market. More specifically, the data was collected to estimate the direct effect of gender and birthplace on yearly quoted premiums. It was collected in 2020 from a popular Italian car insurance comparison website, where the curators tried different hypothetical driver profiles and collected the quotes provided by nine companies. Along with gender and birthplace, additional driver features include age, city of residence, insured vehicle, mileage, and a summary of claim history.
- **Affiliation of creators**: University of Padua; Carnegie Mellon University; University of Udine.
- **Domain**: economics.
- **Tasks in fairness literature**: fair pricing evaluation [245].
- **Data spec**: tabular data.
- **Sample size**: $\sim$ 2K driver profiles.
- **Year**: 2021.
- **Sensitive features**: gender, birthplace.
- **Link**: not available
- **Further info**: Fabris et al. [245]

Fig. 4.1 Data brief of the Italian Car Insurance dataset.

be especially true in Italy, second only to Luxembourg for car ownership among EU states [205]. Auto insurance makes vehicle ownership and usage less hazardous from an individual financial perspective [280]. It acts as a risk-pooling device, covering drivers against liability for bodily injury and property damage in exchange for a premium. Companies are developing increasingly complex machine learning-based models for car insurance pricing [132, 840].

The legal liability connected to driving a vehicle in Italy, and the car insurance system itself, are known as the Motor Vehicle Liability system (*Responsabilità Civile Autoveicoli* - RCA). It is mandatory to purchase RCA coverage before using or keeping a motor vehicle on Italian public roads. RCA is regulated by the national Institute for the Supervision of Insurance (*Istituto Per la Vigilanza Sulle Assicurazioni* - IVASS), which oversees the industry, protecting customers and ensuring transparency, while also promoting market stability and financial viability of businesses.

In the last decade, the use of sensitive features, such as gender and nationality, has been regulated in the Italian car insurance industry. The European Union has adopted legislation that prohibits the direct use of gender to set insurance premiums [226, 227]. After finding evidence of discrimination in pricing on the basis of nationality, IVASS and the National Anti-Racial Discrimination Office (*Ufficio Nazionale Antidiscriminazioni Razziali* - UNAR) issued a soft regulation that encourages companies to avoid using nationality-related factors, such as birthplace and citizenship, as inputs to risk models [395, 777].

Concurrently, comparison websites (also called aggregators) have become a primary point of access to RCA subscription, claiming half of the total gross written premiums in the Italian vehicle insurance market in 2017 [416]. Due to their growing importance, comparison websites have come under increased scrutiny and regulation. In a previous investigation on

RCA aggregators [394], IVASS found anecdotal evidence of *access differences* in connection with risk profile: result pages from comparison websites seemed to display fewer quotes for driver profiles associated with higher risk. Although the evidence, based on a limited sample of 7 driver profiles, was not conclusive, this was highlighted as a potential problem of differential treatment, providing uneven opportunities to different driver segments.

To the best of our knowledge, to date no study has analyzed the direct influence of gender on car insurance pricing in Italy, despite the laws and regulations that limit its use. Furthermore, we are unaware of any study that has examined the influence of nationality-related features on pricing since regulations were issued by IVASS and UNAR. Finally, we are unaware of a systematic study of access differences in RCA comparison websites.

In this section, we describe Italian Car Insurance (§ A.1.101) a dataset we curated to close this gap. It consists of 20,000 quotes collected on a popular comparison website to audit access differences in the aggregator, along with the RCA pricing practices of nine companies, representing a significant share of the market. The dataset caters to **O1** and addresses the following research questions (1) What are the factors that play a major role in setting RCA premiums? (2) Do gender and birthplace directly influence quoted premiums? (3) Do riskier driver profiles see fewer quotes on comparison websites?

4.1.2 Background and Related Work

Protected attributes and fairness criteria

We focus on gender and birthplace as sensitive features both (1) because there exists legislation regulating their use for insurance pricing in Italy, and (2) because RCA websites require users to input these features before generating quotes. Other features that are commonly invoked in studies of fairness and discrimination are either not collected by insurance websites (e.g. race and ethnicity) or are currently permitted under the law as inputs to risk models (e.g. age).

Gender is often conflated with sex in European insurance legislation [226, 227, 242]. The forms on the websites we crawled prompt “The driver is”, providing the options “female” and “male”. We refer to this feature as *gender* throughout the manuscript and follow the binary framing common in the industry and current legislation.

The principle of gender equality is enshrined in Articles 21 and 23 of the Charter of Fundamental Rights of the European Union [241, 242]. Gender equality has been explicitly operationalized in the context of insurance, with Article 5(1) of Council Directive 2004/113/EC [176], which states that no difference in individuals’ premiums can result from the use of gender as an explicit factor, and is fully confirmed in a 2011 judgement by the

European Court of Justice [227]. Official guidelines on the application of the ruling [226] explicitly mention motor insurance, clarifying that indirect discrimination remains possible where justifiable: "For example, price differentiation based on the size of a car engine in the field of motor insurance should remain possible, even if statistically men drive cars with more powerful engines". Moreover, information about gender may still be collected, stored and used, e.g. to monitor portfolio mix or for the purposes of reinsurance.

Nationality-related features were freely used as input to actuarial models in the Italian industry until 2010, when a Tunisian citizen residing in Italy since 1992 sued his car insurance company after being quoted a 30% surcharge due to his citizenship. The lawsuit was later extended to other companies found to engage in similar practices. Following extensive press coverage and further evidence presented by non-Italian citizens, the matter came to the attention of UNAR, who opened an investigation in concert with IVASS and the National Association of Insurance Companies (*Associazione Nazionale fra le Imprese Assicuratrici*). IVASS reported that 25% of companies in their sample took nationality into account as a risk factor. UNAR contacted companies found to charge foreign-born drivers more than Italian-born drivers; one company clarified that birthplace is intended as a proxy for the country where drivers obtain their license, and that learning to drive under different traffic rules and road signs represents an important risk factor. Based on these circumstances, in light of extensive analysis of national and European anti-discrimination law, UNAR issued a general recommendation to the industry, requesting that companies charge the same premiums to Italian and non-Italian citizens, all else being equal [777].

Shortly thereafter, two companies under lawsuit issued a press release, stressing the absence of discriminatory intentions in their practices and committing to remove citizenship from the parameters explicitly used in their risk models [24]. The lawsuit was thereby settled and dropped. Finally, in 2014, IVASS issued a soft regulation, recalling and echoing the recommendation from UNAR, with wording explicitly focused on *birthplace* [395]. IVASS invited all insurance companies operating in Italy to "reconsider this criterion and put in place any activity deemed necessary in order for car insurance quotes and contracts not to take country of birth into account". This regulation clarifies unambiguously that birthplace – not only citizenship – is a sensitive factor, and that its direct utilization in actuarial models is considered discriminatory by IVASS. Henceforth, we refer to a single nationality-related variable, i.e. *birthplace*, given that this is the information currently queried on the websites and distinguish it from citizenship where relevant for the discussion.

In summary, the regulatory framework against discrimination in car insurance described above, comprising EU legislation on gender and Italian soft regulation on birthplace, permits

the collection of protected attributes while forbidding their direct utilization, thus aligning with the criterion of *Fairness Through Unawareness* [328, 475] defined below.

Definition (Fairness Through Unawareness - FTU). Consider a function (equivalently, "algorithm") $f : \mathscr{S} \times \mathscr{X} \mapsto \mathscr{Y}$, where $\mathscr{S}$ represents a sensitive feature, $\mathscr{X}$ represents additional covariates, and $\mathscr{Y}$ represents an output space. The algorithm satisfies FTU with respect to S as long as $f(s,x) = f(s',x)$ for all $s,s' \in \mathscr{S}$ and $x \in \mathscr{X}$. In other words, the algorithm essentially does not utilize the sensitive feature. If the sensitive feature does not form part of the input to the algorithm, then the algorithm trivially satisfies FTU with respect to that feature.

Given its alignment with current regulation, we adopt FTU as the relevant criterion for the purposes of our algorithmic audit, while recognizing that other notions of fairness may be salient in different contexts.

Comparison websites

Comparison websites act as digital intermediaries between customers and insurance providers, typically charging the latter a commission while providing a free service to the former [394, 416]. Their penetration in the European car insurance market has increased dramatically in the last decade. Focusing on the Italian market, in 2017 aggregators reached a 48% share of the total motor gross written premiums [416]. Beyond their importance for direct sales as insurance brokers, comparison websites provide a useful information service for drivers, who can efficiently compare different car insurance options from a single point of access and benefit from increased market transparency.

In a 2014 investigation on comparison websites, IVASS highlighted some critical aspects [394]. They tested 7 different driver profiles on 6 competing comparison websites, finding anecdotal evidence of *access differences*: The result pages appeared to display fewer quotes on driver profiles associated with higher risk. This was stressed by IVASS as a potential problem of differential treatment, providing uneven opportunities to different driver segments. Responding to concerns outlined in the report, some comparison websites provided a technical explanation related to timeouts. The IVASS report concluded that it was impossible to determine whether the variability in the number of quotes was due to technical or strategic reasons.

It should be noted that Italian law imposes a dual *duty to contract*, which applies to both drivers and insurers. According to Article 132 of the Private Insurance Code [392], insurers are required to offer RCA coverage to all drivers, regardless of their risk profile. Article 132-bis, introduced in 2017, recognizes the growing importance of intermediaries in car insurance, such as brokers and agents, who are required to inform customers exhaustively

and transparently with respect to premiums offered by all companies with which they have a broker agreement.

Our reasons for resorting to a comparison website to acquire car insurance quotes are twofold. On the one hand, our objective is to analyze the direct impact of gender and birthplace on quoted prices in a driver-centric fashion, using this prevalent modality of market access. On the other hand, we are interested in auditing patterns of unequal treatment for different users anecdotally highlighted by IVASS.

Car insurance audit

Other researchers have investigated discrimination in the price of auto insurance. Harrington and Niehaus [349] used data from Missouri to examine whether insurance profits were higher in ZIP codes with a higher percentage of minorities. They found no evidence of redlining or other racial discrimination. In subsequent work, Ong and Stoll [616] arrived at a different conclusion. They gathered 836 quotes, varying only the ZIP code while holding all other inputs constant. They found that, after accounting for risk factors, socioeconomic factors in a neighborhood, such as the percentage of poor residents and black residents, correlated with higher premiums. This work is closest to our study, as it is based on quotes gathered with full control of the inputs, some of which are fixed while others are varied according to an experimental design. Most recently, ProPublica analyzed car insurance payouts and premiums in California, Illinois, Texas, and Missouri, coming to similar conclusions that redlining practices affect minority neighborhoods unfavourably [20, 482].

To our knowledge, no such audit has been conducted for the Italian market; our aim is to close a transparency gap between industry practice and current regulation on the equity of RCA pricing and access.

4.1.3 Design of Experiment

Our Design Of Experiment (DOE) and data collection procedure are motivated by the three research questions described in Section 4.1. We choose a common strategy to study access and pricing in RCA: gathering quotes from several companies on a popular RCA comparison website as we vary the drivers' profiles across features that are known *a-priori* to generate sizeable variations in RCA premiums, as detailed in technical reports, trade magazines, and domain-specific websites [169, 595].

We tried to balance this principle of sizeable output variability with that of sample representativeness. For example, when deciding which vehicles to consider, we restricted our options to the best-selling cars in the Italian market, thus neglecting pricey luxury vehicles

Table 4.1 DOE for data collection.

Feature	Values tested	Brief description
gender	F, M	driver's gender
birthplace	Milan, Rome, Naples, Romania, Ghana, Laos	driver's place of birth
age	18, 25, 32	driver's age
city	Milan, Rome, Naples	driver's residence
car	OLED, NSEP	insured vehicle type
km_driven	10,000, 30,000	kms driven yearly
class	0, 4, 9, 14, 18, None (-1)	claim history summary

which are likely associated with the most expensive RCA quotes, but are also far from a representative choice for the average Italian driver.

We define a full factorial experiment, based on protected features (gender and birthplace) and features which are widely recognized as significant for pricing such as driver age, municipality of residence, car, annual mileage, and claim history [168, 595, 709]. Table 4.1 summarizes our DOE.

We let **gender** take the two values allowed on the comparison website: male (M) and female (F). For **birthplace**, we consider Romania, an EU member state with over 1.1 million citizens residing in Italy, along with Ghana and Laos, two countries in completely different geographical areas that also differ greatly for the number of citizens residing in Italy, estimated at 49,543 and 69, respectively [391].[2] It is worth noting that most companies are unlikely to have more than a few tens of Laos-born drivers available as data points to infer the "effect" of this factor level. For this reason, pricing policies connected with this factor level plausibly stem from subjective (potentially inadvertent) choices rather than statistically significant inference. Along with these countries, we also consider the three largest Italian cities in northern (Milan), central (Rome) and southern Italy (Naples).

According to data on recently underwritten RCA contracts [169, 709], most of the **age**-related premium variability is concentrated in the youngest age groups. The mean price for the youngest bracket (18-24) is nearly double the national average; premiums decrease with age up to the bracket (35-44), where they align with the national average. For this reason, we focus on a young segment of the population, aged 18, 25 and 32, who, as is typical of Italians at their age, have been driving for 0, 7 and 14 years, respectively.

[2]While the quoted source reports the number of people with foreign citizenship residing in Italy, the forms in websites we utilized query for their birthplace. Given that Italy has a naturalization rate close to 2% [245], it seems unlikely that the number of Laos citizens and Laos-born people residing in Italy will differ by orders of magnitude.

For **city** of residence, we consider (again) Milan, Rome, and Naples. These are the three largest cities in Italy and represent cultural and economic hubs in northern, central, and southern Italy, respectively. Among the ten most populous Italian cities, residents of Naples and Milan pay, on average, the highest and the lowest RCA premiums, respectively [169, 709].

The type of insured **car** is reported to have a significant impact on the quoted price, with age, engine displacement, and fuel system cited by trade magazines as key factors. Among the best-selling vehicles from 2008 to 2020 [778, 779], the most favorable combination for insurance price is achieved by a 2020 Fiat Panda with a 1.2 litre petrol engine (new, small engine, petrol - abbreviated as NSEP), while the least favorable is a 2008 Fiat Bravo fitted with a 2.0 litre diesel engine (old, large engine, diesel - OLED).

Yearly mileage, or **kilometers driven**, is often cited as an important factor, due to the longer time on the road and consequent risk of causing an accident. We let this feature take values 10,000 (a common default setting in aggregators) and 30,000.

Finally, the Bonus-Malus System (BMS) **class** [762] summarizes the driver claim history, which is updated yearly. Classes 1 and 18 are the best and worst, respectively. New drivers start in class 14, but can alternatively choose to acquire the BMS class of a relative from the same household when purchasing their first auto insurance [393]. Every year, their class improves by 1 if they had no at-fault accidents and increases by 2 otherwise. We investigate the full range available for this feature, from class 1 to class 18, including classes 4, 9 and 14 as intermediate values. The aggregator distinguishes between class 1 and "class 1 for one year or more"; we select the latter value and label it "class 0". Finally, we also test a profile without a driving record (class None), which should be equivalent to class 14.

The full factorial design results in 2,592 unique factor level combinations (*profiles*), of which 1/6 are excluded due to inadmissibility: 18-year-olds are not allowed to drive powerful cars (OLED), reducing the size of the experiment to 2,160 profiles. In setting the (constant) values of the remaining features that are not factors in our study, we aimed for plausible values that are compatible with our chosen factor levels. Our subject is employed, single, and has no children. They are the only driver of the insured vehicle, which is used for both work and leisure.

Overall, this DOE is cognizant of important themes for data curation introduced at the beginning of this chapter. On the privacy side, we opt for *hypothetical profiles*, ensuring extremely low chances of harms related to **re-identification** (studied later in Section 4.2.1), since no driver contributed their personal data to our collection. For the same reason, an analysis of (and planning for) individual **consent** (Section 4.2.2) does not apply and is therefore not addressed. **Inclusivity** (Section 4.2.3) is a strong point of this dataset, as

is typical of datasets created to assess group biases in services, products, and algorithms. Italian Car Insurance represents an example of the *equal* approach to inclusivity, where protected and non-protected groups are represented with the same proportion to support statistically significant statements across sensitive populations. The values for **sensitive attributes** (Section 4.2.4) are imposed by design with the choice of hypothetical profiles, and do not require annotation. Finally, **transparency** and accurate documentation (Section 4.2.5) are a primary concern in this endeavor of data curation and analysis. Our assumptions, choices, and mistakes with respect to design and collection are detailed in the current and subsequent section.

4.1.4 Data Collection

We gather data from a famous comparison website, consistently present in the top two search engine results for the query "comparatore RCA" (RCA comparison website) and meaningful variations thereof. The insurance groups represented in the search results cover more than 60% of the RCA market. To avoid disrupting the service of the website, we collect fewer than 200 quotes per day during July 2020, over a period of 17 days. We envision 3 plausible sources of disturbance in the pricing signal: (1) the evolution of actuarial models and pricing schemes over time, (2) session duration, with time spent on the website potentially factored into the pricing scheme, and (3) A/B testing on behalf of insurance companies, the comparison website or both. To compensate for these effects, we design a doubly-nested randomization with a control group, summarized in Figure 4.2 and described hereafter.

Protected features, likely to cause small fluctuations, which we aim to measure carefully, are bundled and rotated. While keeping all other factors constant, we sequentially execute 12 queries, one for each combination of gender and birthplace, normally over a single session, occasionally over two. We call this sequence of 12 queries, identical for every factor except gender and birthplace, a *block*. This is the inner loop, which is randomized so that each combination of gender and birthplace has an equal chance of occurring at any of the 12 slots in the block. Two profiles that differ only for birthplace (gender) make up a birthplace-(gender-) *protected pair*, as exemplified in Figure 4.2. According to the normative reasoning introduced in Section 4.1.2, profiles in a protected pair should obtain identical quotes. The remaining features are combined via cartesian product and permuted, thus randomizing the order of blocks. This is the outer loop, comprising $B = 180$ blocks in total.

The above procedure should disentangle the features of interest, in particular the protected features, from slow price fluctuations deriving from the evolution of actuarial models and pricing schemes. We further control for unaccounted factors, such as A/B testing or session,

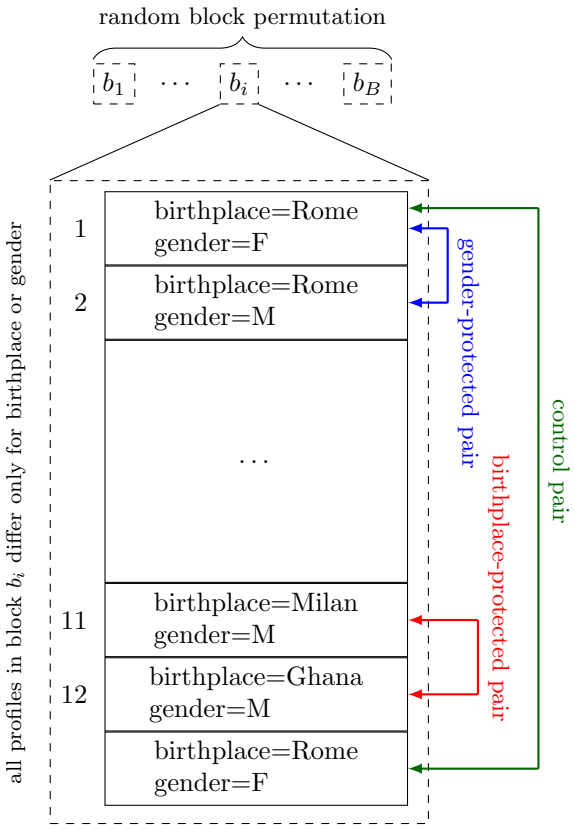


Fig. 4.2 Schematic for Italian Car Insurance data collection procedure. Quotes for profiles that differ only for birthplace or gender are grouped into B *blocks* of 12 ($B = 180$) and collected sequentially in random order. A control quote, identical to the first one, is collected at the end of each block. The blocks are randomly permuted.

by repeating every 12th query. We call these *control* queries, each of them gathered at the end of a block, identical to a regular query gathered at the beginning of the block, with which they form a *control pair*. Overall, the data collection procedure requires the execution of 2,340 queries – one for each of the 2,160 unique profiles and 180 control queries.[3] Each query returns between 5 and 12 price quotes, depending on which companies appear in the search results.

In total, we gather 19,608 yearly quotes from 9 companies (not including control queries), which are summarized in Table 4.2. Companies are labeled `c1` to `c9`, with arbitrary numbering. Depending on product portfolio and agreements with the comparison website, each company offers up to three different insurance products (labeled '/a', '/b' and '/c'). Products from the same company differ in whether they require a tracking device and whether they include premium services, such as road assistance and coverage of legal expenses. Only one company

[3]Due to a design flaw, we only executed control queries for the final 71 blocks.

Table 4.2 Summary of collected insurance quotes.

Company	Num. Quotes	Frequency	Track
c1/a	1728	80%	
c1/b	1728	80%	YES
c1/c	1152	53%	
c2	1787	83%	
c3/a	1477	68%	
c3/b	388	18%	YES
c3/c	690	32%	YES
c4	1628	75%	
c5	2148	99%	
c6/a	717	33%	
c6/b	1428	66%	
c6/c	360	17%	YES
c7	102	4%	
c8	2115	98%	
c9	2160	100%	

(c9) provides a quote for every tested profile; two more companies (c8 and c4) appear very frequently (in 98% and 99% of the query results, respectively). The remaining companies appear 4-83% of the time. This is a first hint of access differences, which will be analyzed in the next section.

4.1.5 Main Results

Most Important Factors

Methods. In this section, we address the first question by analyzing the average impact each factor has on yearly quoted prices. The comparison website orders quotes for a given profile from cheapest (at the top) to most expensive (at the bottom); hence, we refer to an analysis focused on k cheapest quotes as `top-k`. We perform the following analyses:

- `top1`: examines the cheapest quote obtained for each profile. This analysis adopts the perspective of a driver driven solely by expense minimization.
- `top5`: average of the five cheapest quotes obtained for each profile. Average prices correspond to a dual point of view: (1) a driver who is not necessarily seeking the cheapest product; (2) a driver who is "shopping around" on the website, comparing several insurance options with their current contract. At least five quotes were returned for each profile.

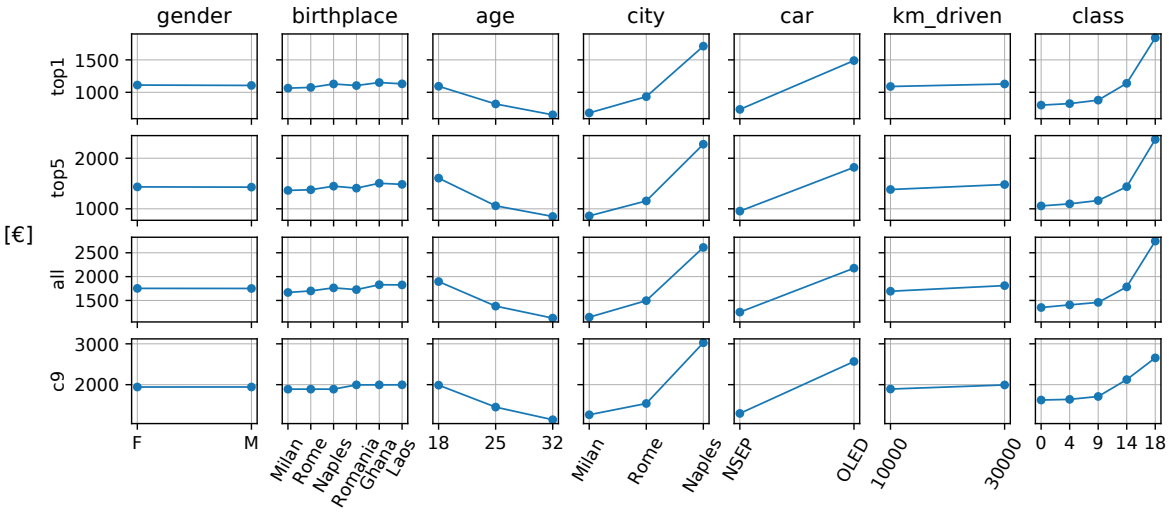


Fig. 4.3 Overview of factor influence on insurance price. Each column represents a different factor, with all its tested levels on the *x* axis. The *y* axis depicts the influence of each factor, as a mean price for all profiles with a given factor level, in different analyses: cheapest quote (`top1` - row 1), average of 5 cheapest quotes (`top5` - row 2), average of all quotes (`all` - row 3) and quote by company `c9` (row 4), the only company present in every result page. Age, city, car and class are confirmed to strongly influence average price, while mileage and protected factors appear to be less important in comparison.

- `all`: average of all quotes obtained for each profile.
- `c9`: quotes provided by the only company that appeared on the result pages for each profile tested, i.e., `c9`.

For each of the four analyses above, we first reduce all the quotes on the result page to a single price, either by selecting the relevant quote (in the `top1` and `c9` analyses) or by averaging the selected quotes (in the `top5` and `all` analyses). Thus, each profile corresponds to a single price within each analysis.

Results. Figure 4.3 summarizes these analyses (one per row), with each column representing a different feature. Each panel plots the mean quoted price in euros for all profiles with a given feature value, represented on the *y* axis, versus feature values along the *x* axis. For example, the top left panel reports the results of the `top1` analysis, depicting the average of each "female" profile and the average of each "male" profile.

Notice that the results are robust across all analyses (rows) in Figure 4.3.[4] Age, city, car, and class have a strong effect, confirming our DOE considerations for including them. Mileage (km_driven), on the other hand, has a weak effect, despite being reported as a powerful predictor of the number of claims at-fault [265, 491]. We hypothesize that this is

[4] An analysis focused on median values, omitted to save space, yields equivalent results.

due to the low verifiability of this feature, which is self-reported and difficult to verify for insurance companies in the absence of a tracking device fitted on the vehicle.

Among protected features, birthplace seems to be utilized to differentiate not only between different countries, but also Italian cities, though the effects are smaller than for the previously mentioned factors. Gender, on the other hand, seems to play a negligible role. The absence of a clear effect for this feature should not be interpreted as a guarantee that it does not directly influence actuarial models. Rather, it means that, if gender-based differences are present, they do not on average favor men or women. The next section provides in-depth analysis of the role of gender and birthplace.

Discrimination Analysis

Methods. In this section, we focus on the direct influence of protected attributes, i.e. birthplace and gender, on price. While the previous section considered average prices across feature levels, here we examine the distribution of price differences δ for pairs of profiles that differ only in one protected attribute (e.g. F-M for gender, Ghana-Milan for birthplace), which we refer to as *protected pairs*. For FTU to hold rigorously, the result pages presented to protected pairs of profiles should be identical. To compensate for the effect of external factors (such as A/B testing), modest differences are deemed acceptable as long as they remain comparable to differences between two identical queries (*control pairs*). Recall that protected pairs are always gathered within the same block (Figure 4.2), so that the effect of time or browser session on any given protected pair should be minimal and smaller than the effect it has on control pairs, which are gathered by definition at the very beginning and end of a block.

We conduct `top1` and `top5` analyses, collapsing each set of query results into a single price as described above. Again, these analyses adopt the perspective of a driver who is only interested in the cheapest possible quote (`top1`) or a driver who is shopping around and comparing policies (`top5`). Within each analysis, we consider the vector that contains price differences δ for all protected pairs with two given factor levels (e.g. female and male for gender). We compute its median value $m(\delta)$ and report the p-value from a sign test, which tests the null hypothesis that the median difference for each pair of profiles is 0, meaning e.g. that we are as likely to observe a difference in favor of men as a difference in favor of women. If we reject the null hypothesis, then we are compelled to conclude that FTU does not hold, though, of course, failure to reject the null hypothesis does not guarantee that FTU does hold. In particular, while the condition $m(\delta) = 0$ ensures that no protected group is *systematically* at a disadvantage, it does not provide any guarantee about price differences directly determined by a protected attribute in a pair and compensated for by a

Table 4.3 Summary of discrimination analysis. For protected pairs, we consider the vector of differences (δ) in `top1` and `top5` values. We report the percentage of ties (within a 5€ tolerance threshold - Ties_5), the 5th and 95th percentiles ($\eta_{.05}(\delta)$, $\eta_{.95}(\delta)$), the median difference $m(\delta)$, and the p-value from a sign test described in the main text. Both birthplace and gender can have a sizeable direct influence on the quotes that drivers see; the influence of birthplace is more frequent, substantial, and systematic, with drivers who are not born in Milan suffering financial disadvantages relative to Milan.

			`top1`					`top5`			
Attribute	Pairs	Ties_5	$\eta_{.05}(\delta)$	$\eta_{.95}(\delta)$	$m(\delta)$	p	Ties_5	$\eta_{.05}(\delta)$	$\eta_{.95}(\delta)$	$m(\delta)$	p
birthplace	Rome vs Milan	23%	-238 €	207 €	10 €	7.6e−08	5%	-202 €	240 €	7 €	3.0e−04
birthplace	Naples vs Milan	27%	-60 €	274 €	27 €	3.2e−16	6%	-50 €	331 €	53 €	7.9e−31
birthplace	Romania vs Milan	37%	-81 €	253 €	17 €	3.2e−16	9%	-86 €	225 €	39 €	6.8e−27
birthplace	Ghana vs Milan	30%	-90 €	553 €	57 €	7.9e−31	5%	-48 €	521 €	84 €	2.6e−61
birthplace	Laos vs Milan	30%	-46 €	312 €	56 €	7.9e−31	6%	-60 €	437 €	78 €	2.0e−58
gender	F vs M	78%	-48 €	127 €	0 €	5.3e−02	39%	-173 €	187 €	0 €	2.1e−01
noise control		93%	-33 €	0 €	0 €	2.3e−01	89%	-6 €	11 €	0 €	5.0e−01

difference of opposite sign in another protected pair. To this end, we also compute the 5th and 95th percentiles (labeled $\eta_{.05}(\delta)$ and $\eta_{.95}(\delta)$, respectively), along with the percentage of protected pairs for which the quote difference δ is within a tolerance threshold of 5€ (Ties_5). We compare these values with those computed for control pairs. To satisfy FTU, we would expect protected pairs and control pairs to exhibit non-zero differences with comparable frequency (summarized by Ties_5) and magnitude (summarized by $\eta_{.05}(\delta)$ and $\eta_{.95}(\delta)$).

Results. Our numerical analysis is presented in Table 4.3. Rows 1-5 relate to birthplace, where Milan acts as a baseline, and positive values represent a surcharge incurred by drivers born in Rome, Naples, Romania, Ghana, and Laos, respectively. Row 6 shows analogous results where the protected attribute is gender and positive differences are unfavorable for female drivers. A final row is added to summarize the effect of noise by reporting results for control pairs.

Focusing on the median difference $m(\delta)$, we find no systematic gender bias: the median is zero for both `top1` and `top5` analyses, with insignificant p-values, even before correcting for multiple hypotheses testing. However, we find some sizeable price differences for gender-protected pairs, which are centered around zero, thus placing no gender at a systematic disadvantage. This finding will be discussed in the next paragraph. On the other hand, birthplace is used predominantly in one direction, to the advantage of drivers born in Milan. Their `top1` and `top5` average quotes are consistently lower than those of foreign-born drivers from Laos, Ghana, and Romania, with median `top5` differences of 78€, 84€ and 39€, respectively. Changing birthplace from Milan to Naples also results in significantly higher quotes ($m(\delta)$ equal to 27€ for `top1` and 53€ for `top5`). Although less sizeable, drivers born in Rome also find a significant median difference compared to their Milan-born counterparts

($m(\delta)$ equal to 10€ for `top1` and 7€ for `top5`). This is the first result we are aware of demonstrating that pricing algorithms return different quotes for drivers born in different Italian cities, even when all remaining factors are held equal. All p-values associated with birthplace are significant.

Considering the magnitude of differences directly induced by protected attributes, we find that the gender- and birthplace-based differences $\eta_{.95}(\delta)$-$\eta_{.05}(\delta)$ in the `top5` results range from 311€ to 569€, compared to a value of 17€ for control pairs. The frequency of Ties_5 for `top5` is below 10% for all birthplace-protected pairs and below 40% for gender-protected pairs, compared against a value of 89% for control pairs. Similar if somewhat weaker patterns obtain in the `top1` results. We interpret these findings as evidence that gender and birthplace have a direct and substantial influence in the result pages of this comparison website. Histograms for these differences are reported in Figure 4.4.

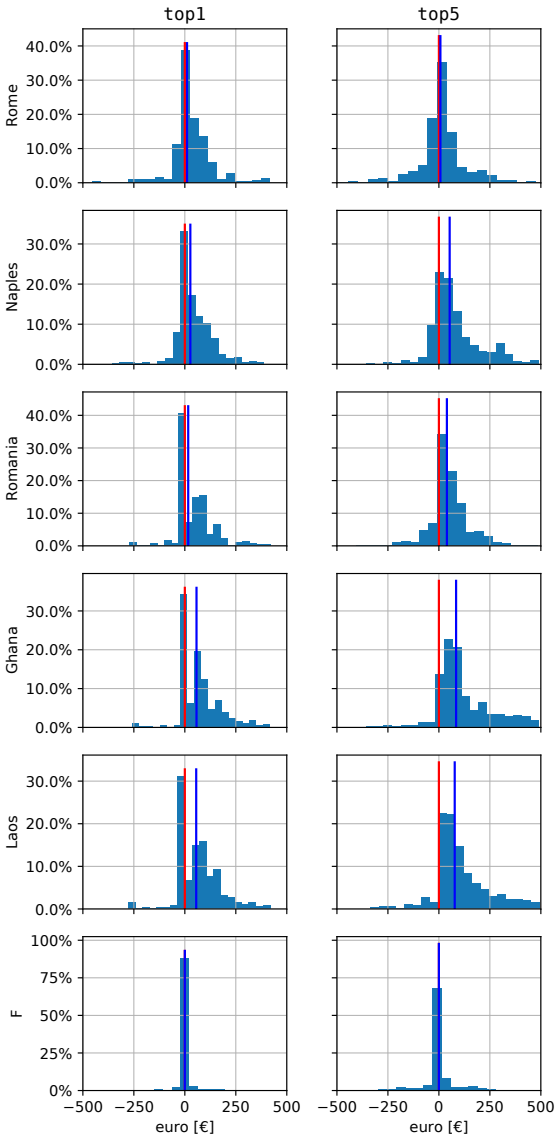


Fig. 4.4 Birthplace- and gender-based discrimination. Histogram of paired differences in cheapest quote (col. 1) and average of 5 cheapest quotes (col. 2). Rows 1-5 focus on birthplace, with Milan as a baseline, depicting paired differences with respect to Rome (row 1), Naples (row 2), Romania (row 3), Ghana (row 4), Laos (row 5). Positive means Milan is cheaper. Row 6 depicts gender-based differences (F-M). Vertical blue lines represent the median difference, while red lines are the median difference between regular and control quotes (zero). The x axis is clipped between -500 and 500 €. Both birthplace and gender can have a sizeble influence, the former being more frequent, strong and systematic in one direction.

In summary, the pricing algorithms generating the RCA quotes that drivers obtain through this popular aggregator violate FTU: when all else is held constant, both gender and birthplace

have sizable effects on the quoted prices, even though, in the case of gender, the direction of this effect is not systematic, i.e. the median effect is 0. Given that aggregators have become a primary point of access to RCA subscription, these results point to potentially non-trivial violations of existing laws and regulations.

It is not immediately clear to what extent these results arise from the pricing algorithms of individual companies versus the behavior of the aggregator. In this regard, we note that 4 out of 9 of the companies in the results do not appear to use gender or birthplace directly for pricing insurance. This suggests (1) that these results are probably not due to the aggregator alone and (2) that the use of gender and birthplace does not qualify as a fundamental business need, which could otherwise partially explain violations of FTU. While the aggregator may in theory offer different prices than those offered on the companies' own websites, studies of prevalent business models for aggregators suggest the contrary [394, 416]. To investigate whether the pricing patterns we find are independent of the aggregator, in Appendix B we analyze a dataset gathered from the website of a single insurance company, comparing these quotes with those obtained in the aggregator. Overall, we find the effect of the aggregator to be negligible, if any.

Access Differences

Methods. In this section, we analyze the effect of each factor included in our DOE on the frequency f_q with which insurance companies appear in quotes for specific profiles. We report the results of four companies listed in Table 4.2, for which f_q displays a clear dependence on one or more factors. We also aggregate these results from the perspective of users, detailing how different features affect the average number of quotes they see.

Results. Figure 4.5 contains a summary of our results, where each column represents a factor, with all its possible values on the x axis. Rows 1-4 depict f_q for c1, c2, c4, and c7, respectively. Interesting patterns emerge from the 2160 profiles that were tested. Company c1 is never present on the result pages for 18-year-old drivers. Company c2 is absent from the result pages for drivers in the worst BMS class (18). Both these results are very strong, since c1 and c2 are otherwise present 100% of the time. Company c4 is always present on the results pages for residents of Rome and Milan, but its frequency of appearance drops to 26% for Naples. Company c7, appearing only in 4% of the result pages, seems to focus on Italian-born drivers of non-OLED cars with no claim history.

Row 5 of Figure 4.5 aggregates these results from the user's perspective, detailing how different characteristics affect the average number of quotes they see, reported on the y axis. Age plays a major role, with 18- and 32-year-olds seeing on average 7.1 and 9.8 quotes, respectively. Municipality is also important: more quotes are available for Milan residents

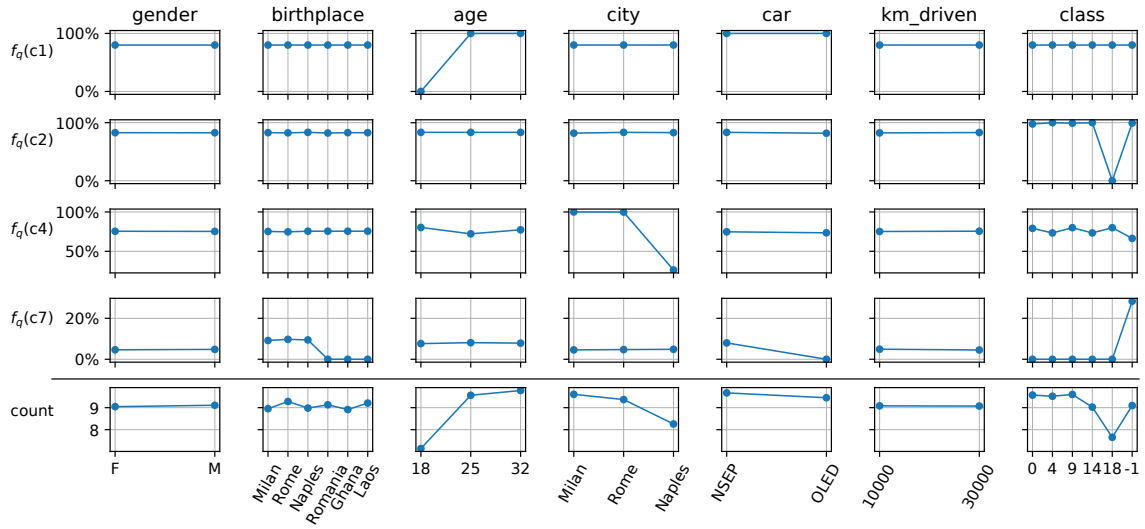


Fig. 4.5 Influence of each factor on frequency of appearance in result pages for company `c1` (row 1), `c2` (row 2), `c4` (row 3), `c7` (row 4) and total number of quotes (row 5). Strong patterns are visible for factors age, city and class.

than for Naples residents. Claim history is another important factor, with drivers in BMS classes above 9 seeing fewer quotes than drivers with more favorable classes. In general, these are also factors that have a strong influence on insurance premiums, as depicted in Figure 4.3. Profiles perceived as risky see fewer, more expensive quotes.

Like price, the number of quotes may be subject to noise, due e.g. to A/B testing or technical issues. We quantify this effect by considering control pairs. We notice that 17% of the result page pairs differ by 1 in the number of quotes returned, resulting in an average absolute difference of 0.17 quotes for identical profiles. We regard this figure as an estimate of noise affecting the number of quotes returned by the aggregator in its result pages. As shown in the bottom row of Figure 4.5, age, city, and class induce systematic differences, one order of magnitude larger than this threshold.

To illustrate the potential impact of these findings on drivers, let us consider matching profiles that differ only for age, and let us pair 18-year-olds with their 32-year-old counterparts. In 26% of the resulting pairs, the company providing the cheapest quote to the 32-year-old driver is absent from the result page of the matching 18-year-old. This clearly reduces the opportunities available to some younger drivers, hiding potentially favorable premiums from them, which in turn can contribute to an increase in their expenses. This problem is also relevant for factors that are not associated with systematic access differences. Focusing on gender, for example, if we consider gender-protected pairs such that $\delta_{\texttt{top1}} > \eta_{.95}(\delta_{\texttt{top1}})$, i.e. the pairs with most extreme `top1` differences in favor of men (top 5 percentiles), we find

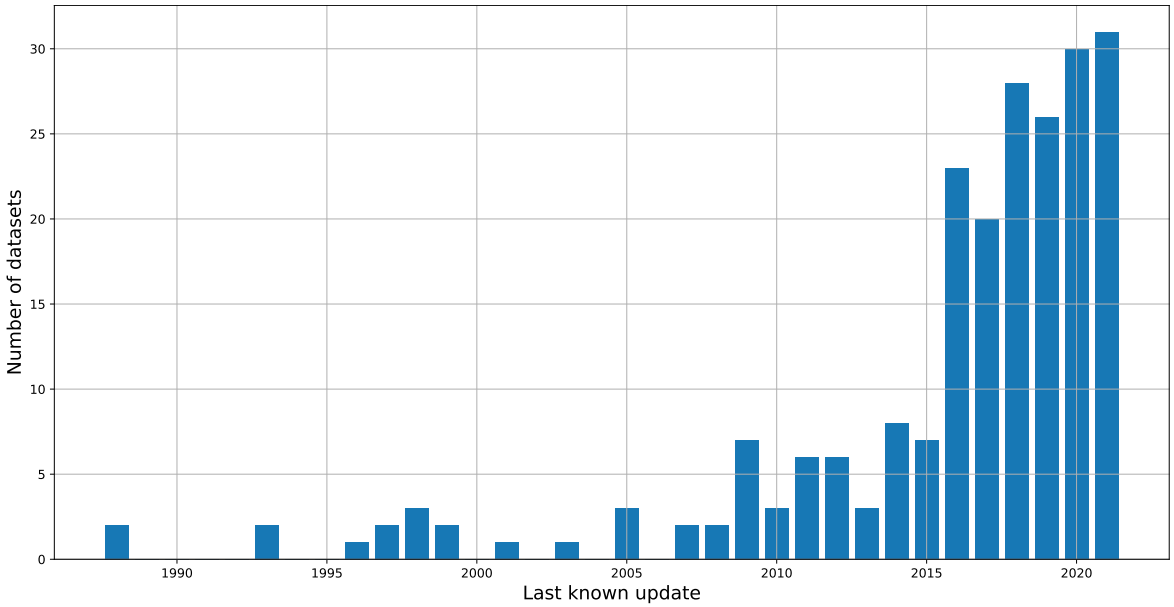


Fig. 4.6 Most datasets used in algorithmic fairness were created or updated after 2015, with a clear growth in recent years.

that the company providing the cheapest quote to the male profile is absent from the result page of his female counterpart 84% of the time. In other words, the most extreme differences in `top1` price for gender-protected pairs appear to be caused by output variability on the aggregator result pages.

4.2 Curatorial Best Practices

As more algorithmic fairness datasets are created each year (Figure 4.6), it becomes increasingly important for curators to understand the importance of ethical data curation for harm avoidance and, subsequently, to embrace curatorial best practices. In this section, we analyze algorithmic fairness datasets from the perspective of five important data curation topics: anonymization, consent, inclusivity, labeling of sensitive attributes, and transparency. Consider the datasets that meet the inclusion criteria outlined in Section 3.2.2, we discuss different approaches and levels of attention to these topics, making them tangible, and distill them into a set of best practices for the curation of novel resources.

4.2.1 Re-identification

Motivation. Data re-identification (or de-anonymization) is a practice through which instances in a dataset, theoretically representing people in an anonymized fashion, are success-

fully mapped back to the respective individuals. Therefore, their identity is discovered and associated with the information encoded in the features of the dataset. Examples of external re-identification attacks include de-anonymization of movie ratings from the Netflix prize dataset [593], identification of profiles based on social media group membership [830], and identification of people depicted in verifiably pornographic categories of ImageNet [644]. These analyses were carried out as "attacks" by external teams for demonstration purposes, but curators and stakeholders may undertake similar efforts internally [546].

There are multiple harms connected to data re-identification, especially the ones featured in algorithmic fairness research, due to their social significance. Depending on the domain and breadth of information provided by a dataset, malicious actors may acquire information about mobility patterns, consumer habits, political leaning, psychological traits, and medical conditions of individuals, just to name a few. The potential for misuse is enormous, including phishing attacks, blackmail, threat, and manipulation [468]. Face recognition datasets are especially prone to successful re-identification as, by definition, they contain information strongly connected with a person's identity. The problem also extends to general-purpose computer vision datasets. In a recent dataset audit, Prabhu and Birhane [644] found images of beach voyeurism and other non-consensual depictions in ImageNet, and were able to identify the victims using reverse image search engines, highlighting downstream risks of blackmail and other forms of abuse.

Disparate consideration. Fairness datasets are proofed against re-identification with a full range of measures and care. Perhaps surprisingly, some datasets allow for straightforward re-identification of individuals, providing their full names. We do not discuss these resources here to avoid amplifying the harms discussed above. Other datasets afford plausible re-identification, providing social media handles and aliases, such as Twitter Abusive Behavior, Sentiment140, Facebook Large Network, and Google Local. Columbia University Speed Dating may also fall in this category due to a restricted population from which the sample is drawn, and provision of age, field of study, and ZIP code where participants grew up in addition. In contrast, many datasets come with strong guarantees against de-anonymization, which is especially typical of health data, such as MIMIC-III and Heritage Health [232]. Indeed, health is a domain where a culture of confidentiality of patient records is widely established and there is strong attention to harm avoidance. Also datasets describing scholarly works and academic collaboration networks (Academic Collaboration Networks, PubMed Diabetes Papers, Cora, CiteSeer) are typically de-identified, with numerical IDs substituting names. This is possibly a sign of attention to anonymization from curators when the data represents potential colleagues. As a consequence, researchers are protected from related harms, but posterior annotation of sensitive attributes similarly to Biega et al. [72] becomes

difficult or impossible. One notable exception is ArnetMiner Citation Network, derived from an online platform which is especially focused on data mining from academic social networks and profiling of researchers.

Mitigating factors. A wide range of factors, summarized in Table 4.4. may help to reduce the risk of re-identification. A first set of approaches concerns the distribution of data artifacts. Some datasets are simply kept private, minimizing risks in this regard. These include UniGe, US Student Performance, Apnea, Symptoms in Queries and Pymetrics Bias Group, the last two being proprietary datasets that are not disclosed to preserve intellectual property. Twitter Online Harrassment is available upon request to protect the identities of Twitter users that were included. Another interesting approach is mixed release strategies: NLSY has some publicly available data, while access to further information that may favor re-identification (e.g. ZIP code and census tract) is restricted. For crawl-based datasets, it is possible to keep a resource private while providing code to recreate it (Bias in Bios). Although this may alleviate some concerns, it will not deter motivated actors. As a post-hoc remedy, proactive removal of problematic instances is also a possibility, as shown by recent work on ImageNet [842].

Table 4.4 Mitigating factors against re-identification.

Mitigating factor	Example datasets
Controlled distribution	
Private dataset	UniGe, Pymetrics Bias Group
Availability upon request	Twitter Online Harrassment
Mixed release strategy	NLSY
Code-based reconstruction	Bias in Bios
Data perturbation	
Obfuscation	Yahoo! c14B Learn to Rank, Microsoft Learning to Rank
Top-coding	Adult
Blurring	Chicago Ridesharing
Targeted scrubbing	ASAP
Aggregation	FICO
Synthesis	
Synthetic data	Toy Dataset 1–4
Semi-synthetic data	Antelope Valley Networks, Kidney Matching
Hypothetical profiles	Italian Car Insurance
Age	German Credit

Another family of approaches is based on redaction, aggregation, and injection of noise. Obfuscation typically involves the distribution of proprietary company data at a level of abstraction which maintains utility to a company while hindering reconstruction of the

underlying human-readable data, which also makes re-identification highly unlikely (Yahoo! c14B Learn to Rank, Microsoft Learning to Rank). Noise injection can take many forms, such as top-coding (Adult), i.e., saturation of certain variables, and blurring (Chicago Ridesharing), i.e., disclosure at coarse granularity. Targeted scrubbing of identifiable information is also rather common, with ad-hoc techniques applied in different domains. For example, the curators of ASAP, a dataset featuring student essays, removed personally identifying information from the essays using named entity recognition and several heuristics. Finally, the aggregation of data into subpopulations of interest also supports the anonymity of the underlying individuals (FICO).

So far, we have covered datasets that feature human data derived from real-world processes. Toy datasets, on the other hand, are perfectly safe from this point of view; however, their social relevance is inevitably lower. In this work we survey four popular ones, taken from Donini et al. [216], Lipton et al. [506], Singh and Joachims [723], Zafar et al. [856]. Semi-synthetic datasets aim for the best of both worlds by generating artificial data from models that emulate the key characteristics of the underlying processes, as is the case with Antelope Valley Networks, Kidney Matching, and the generative adversarial network trained by McDuff et al. [543] on MS-Celeb-1M. Data synthesis may also be applied to augment datasets with artificial sensitive attributes in a principled fashion (MovieLens – [102]). Finally, resources designed to externally investigate services, algorithms, and platforms, to estimate the direct effect of a feature of interest (e.g. gender, race), may rely on hypothetical profiles [66, 250]. This approach can support evaluations of *fairness through unawareness* [327], of which Italian Car Insurance (Section 4.1) is an example.

The last important factor is the *age* of a dataset. Re-identification of old information about individuals requires matching with auxiliary resources from the same period, which are less likely to be maintained than comparable resources from recent years. Moreover, even if successful, the consequences of re-identification are likely mitigated by dataset age, as old information about individuals is less likely to support harm against them. The German Credit dataset, for example, represents loan applicants from 1973–1975, whose re-identification and subsequent harm appears less likely than re-identification for more recent datasets in the same domain.

Anonymization vs social relevance. Utility and privacy are generally considered conflicting objectives for a dataset [822]. If we define social relevance as the breadth and depth of societally useful insights that can be derived from a dataset, a similar conflict with privacy becomes clear. Old datasets hardly provide any insight that is actionable and relevant to current applications. Insight derived from synthetic datasets is inevitably questionable. Noise injection increases uncertainty and reduces the precision of claims. Obfuscation hinders

subsequent annotation of sensitive attributes. Conservative release strategies increase friction and deter people from obtaining and analyzing the data. The most socially relevant fairness datasets typically feature confidential information (e.g. criminal history and financial situation) in conjunction with sensitive attributes of individuals (e.g. race and sex). For these reasons, the social impact afforded by a dataset and the safety against re-identification of included individuals are potentially conflicting objectives that require careful balancing. In the next section, we discuss informed consent, another important aspect for the privacy of data subjects.

4.2.2 Consent

Motivation. In the context of data, *informed consent* is an agreement between a data processor and a subject, intended to allow the collection and use of personal information while guaranteeing some control to the subject. It is emphasized in Article 7 and Recitals (42) and (43) of the General Data Protection Regulation [244], which require consent to be freely given, specific, informed, and unambiguous. Paullada et al. [627] note that in the absence of individual control on personal information, anyone with access to the data can process it with little oversight, possibly against the interest and well-being of data subjects. Consent is therefore an important tool in a healthy data ecosystem that favors development, trust, and dignity.

Negative examples. A separate framework, often conflated with consent, is copyright. Licenses such as Creative Commons discipline how academic and creative works can be shared and built upon, with proper credit attribution. However, according to the Creative Commons organization, their licenses are not suited to protect privacy and cover research ethics [555]. In computer vision, and especially in face recognition, consent and copyright are often considered and discussed jointly, and Creative Commons licenses are often taken as an all-inclusive permit encompassing intellectual property, consent, and ethics [644]. Merler et al. [556], for example, mention privacy and copyright concerns in the construction of Diversity in Faces. These concerns are apparently jointly solved by obtaining images from YFCC-100M, due to the fact that "a large portion of the photos have Creative Commons license". In fact, lack of consent is a widespread and far-reaching problem in face recognition datasets [439]. Prabhu and Birhane [644] find several examples of non-consensual images in large-scale computer vision datasets. A particularly egregious example covered in this survey is MS-Celeb-1M, released in 2016 as the largest publicly available training set for face recognition in the world [335]. As suggested by its name, the dataset should feature only celebrities, "to enable our training, testing, and re-distributing under certain licenses" [335].

However, the dataset was later found to feature several people who are in no way celebrities and must simply maintain an online presence. The dataset was retracted for this reason [586].

Positive examples. FACES, an experimental psychology dataset on emotion-related stimuli, represents a positive exception in the face analysis domain. Due to its small cardinality, it was possible to obtain informed consent from all participants. One domain where informed consent doctrine has been well established for decades is medicine; fairness datasets from this space are typically sensitive to the topic. Experiments such as randomized controlled trials always require consent elicitation and often discuss the process in the respective articles. Infant Health and Development Program (IHDP), for instance, is a dataset used to study fair risk assessment. It was collected through the IHDP program, carried out between 1985 and 1988 in the US to evaluate the effectiveness of comprehensive early intervention in reducing developmental and health problems in low birth weight premature infants. Brooks-Gunn et al. [95] clearly state that "of the 1302 infants who met enrollment criteria, 274 (21%) had parents who refused consent and 43 were withdrawn before entry into the assigned group". Longitudinal studies require trust and continued participation. They typically produce insights and data thanks to participants who have read and signed an informed consent form. Examples of such datasets include Framingham, stemming from a study on cardiovascular disease, and the National Longitudinal Survey of Youth, following the lives of representative samples of US citizens, focusing on their labor market activities and other significant life events. Field studies and derived datasets (DrugNet, Homeless Youths' Social Networks) are also attentive to informed consent.

The FRIES framework. According to the Consentful Tech Project,[5] consent should be *Freely given*, *Reversible*, *Informed*, *Enthusiastic*, and *Specific* (FRIES). Below, we expand on these points and discuss some fairness datasets through the FRIES lens. Pokec Social Network summarizes the networks of Pokec users, a popular social network in Slovakia and the Czech Republic. Due to default privacy settings being predefined as public, a wealth of information for each profile was collected by curators, including information on demographics, politics, education, marital status, and children [753]. Although privacy settings are a useful tool to control personal data, default public settings are arguably misleading and do not amount to *freely given* consent. In the presence of more conservative predefined settings, a user can explicitly choose to publicly share their information. This may be interpreted as consent to share one's information here and now with other users; more loose interpretations favoring data collection and distribution are also possible, but they seem rather lacking in *specificity*. It is far from clear that choosing public profile settings entails consent to become part of a study and a publicly available dataset for years to come.

[5] https://www.consentfultech.io/

This contrasts with Framingham and other datasets derived from medical studies, where consent may be provided or refused with fine granularity [497]. In this regard, consider a consent form from a recent Framingham exam [279]. The form comes with five different consent boxes that cover participation in examination, use of the resulting data, participation in genetic studies, sharing of data with external entities, and notification of findings to the subject. Before the consent boxes, a well-structured document informs participants on the reasons for this study, clarifies that they can choose to drop out without penalties at any point, provides a point of contact, explains what will happen in the study, and what the risks are to the subject. Some examples of accessible language and open explanations include the following.

- "You have the right to refuse to allow your data and samples to be used or shared for further research. Please check the appropriate box in the selection below."
- "There is a potential risk that your genetic information could be used to your disadvantage. For example, if genetic research findings suggest a serious health problem, that could be used to make it harder for you to get or keep a job or insurance."
- "However, we cannot guarantee total privacy. [...] Once information is given to outside parties, we cannot promise that it will be kept private."

Moreover, the consent form is accessible from a website that promises to provide a Spanish version, showing attention to linguistic minorities. Overall, this approach is geared towards trust and truly informed consent.

In some cases, consent is made unapplicable by necessity. Allegheny Child Welfare, for example, stems from an initiative by the Allegheny County's Department of Human Services to develop assistive tools to support child maltreatment hotline screening decisions. Individuals who resort to this service are in a situation of need and emergency that makes *enthusiastic* consent highly unlikely. Similar considerations arise in any situation where data subjects are in a state of need and can only access a service by providing their data. A clear example is Harvey Rescue, the result of crowdsourced efforts to connect rescue parties with people seeking help in the Houston area. Moreover, the provision of data is mandatory in some cases, such as the US census, which conflicts with meaningful, let alone enthusiastic, consent.

Finally, consent should be *reversible*, giving individuals the opportunity to revoke it and be removed from a dataset. This is an active area of research, studying specific tools for consent management [10] and approaches for retroactive removal of an instance from a model's training set [304]. Unfortunately, even when discontinued or redacted, some

datasets remain available through backchannels and derivatives. MS-Celeb-1M is, again, a negative example in this regard. Microsoft removed the dataset after widespread criticism and claims of privacy infringement. Despite this fact, it remains available through academic torrents [630]. Moreover, MS-Celeb-1M was used as a source of images for several datasets derived from it, including the BUPT Faces and Racial Faces in the Wild datasets. This fact demonstrates that harms related to data artifacts are not simply remedied through retirement or redaction. Ethical considerations about consent and potential harms to people must be more than an afterthought and must be included in the discussion during design.

4.2.3 Inclusivity

Motivation. Issues of representation, inclusion and diversity are central to the fair ML community. Due to historical biases arising from structural inequalities, some populations and their perspectives are underrepresented in certain domains and related data artifacts [405]. For example, the person subtree of ImageNet contains images that skew towards male, young, and light skin individuals [842]. Female entities were found to be underrepresented in popular datasets for coreference resolution [881]. Even datasets that match natural group proportions may support the development of biased tools with low accuracy for minorities.

Recent work has demonstrated the disparate performance of tools in sensitive subpopulations in domains such as health care [607], speech recognition [760], and computer vision [101]. Inclusivity and diversity are often considered a primary solution in this regard, both in training sets, which support the development of better models, and test sets, capable of flagging such issues.

Positive examples. Ideally, inclusivity should begin with a clear definition of data collection objectives [405]. Indeed, we find that diversity and representation are strong points of datasets that were created to assess biases in services, products, and algorithms (BOLD, HMDA, FICO, Law School, Scientist+Painter, CVs from Singapore, YouTube Dialect Accuracy, Pilot Parliaments Benchmark), which were designed and curated with special attention to sensitive groups. We also find instances of ex-post remedies to issues of diversity. As an example, ImageNet curators proposed a demographic balancing solution based on a web interface that removes images from overrepresented categories [842]. A natural alternative is the collection of novel instances, a solution adopted for Framingham. This dataset comes from a study of key factors that contribute to cardiovascular disease, with participants recruited in Framingham, Massachusetts over multiple decades. Recent cohorts were especially designed to reflect greater racial and ethnic diversity in the city [769].

Negative examples. Among the datasets we surveyed, we highlight one whose low inclusivity is rather obvious. WebText is a 40 GB text dataset that supported training of the

GPT-2 language model [655]. The authors crawled every document accessible from outbound Reddit links that collected at least 3 *karma*. Although this was considered a useful heuristic to achieve size and quality, it ended up skewing this resource toward content appreciated by Reddit users, who are predominantly male, young, and enjoy good internet access. This should serve as a reminder that size does not guarantee diversity [59], and that sampling biases are almost inevitable.

Inclusivity is nuanced. While inclusivity surely requires an attention to subpopulations, a more precise definition may depend on context and application. Based on the task at hand, an ideal sample may feature all subpopulations with equal presence, or proportionally to their share in the overall population. Let us call these the *equal* and *proportional* approaches to diversity. The equal approach is typical of datasets that are meant to be evaluation benchmarks (Pilot Parliaments Benchmark, Winobias) and allow for statistically significant statements on performance differences across groups. On the other hand, the proportional approach is rather common in datasets collected by census offices, such as US Census Data (1990), and in resources aimed precisely at studying issues of representation in services and products (Occupations in Google Images).

Open-ended data collection is ideal to ensure that various cultures are represented in the manner in which they would like to be seen [405]. Unfortunately, we found no instance of datasets in which sensitive labels were self-reported according to open-ended responses. On the contrary, individuals with non-conforming gender identities were excluded from some datasets and analyses. Bing US Queries is a proprietary dataset used to study differential user satisfaction with the Bing search engine across different demographic groups. It consists of a subset of Bing users who provided their gender at registration according to a binary categorization, which misrepresents or simply excludes non-binary users from the subset. Moreover, a dataset may be inclusive and encode gender in a non-binary gender fashion (Climate Assembly UK), but, if used in conjunction with an auxiliary dataset where gender has binary encoding, a common solution is to remove instances whose gender is neither female nor male [272].

Inclusivity does not guarantee benefits. To avoid downstream harm, inclusion alone is insufficient. The context in which people and sensitive groups are represented should always be taken into account. Despite its overall skew towards male subjects, ImageNet has a high female-to-male ratio in classes such as `bra`, `bikini` and `maillot`, which often feature images that are voyeuristic, pornographic, and non-consensual [644]. Similarly, in MS-COCO, a famous dataset for object recognition, there is roughly a 1:3 female-to-male ratio, increasing to 0.95 for images of kitchens [360]. This sort of representation is unlikely

to benefit women in any way and, on the contrary, may contribute to reinforce stereotypes and support harmful biases.

Another clear (but often ignored) disconnect between the inclusion of a group and benefits to it is represented by the task at hand and, more generally, by possible uses afforded by a dataset. In this regard, we find many datasets from the face recognition domain, which are presented as resources geared towards inclusion (Diversity in Faces, BUPT Faces, UTK Face, FairFace, Racial Faces in the Wild). Attention to subpopulations in this context is still called "diversity" (Diversity in Faces, FairFace, Racial Faces in the Wild) or "social awareness" (BUPT Faces), but is driven by business imperatives and goals of robustness for a technology that can very easily be employed for surveillance purposes, and become detrimental to vulnerable populations included in datasets. In a similar vein, the FACES dataset has been used to measure age bias in emotion detection, a task whose applications and benefits for individuals remain dubious.

Overall, attention to subpopulations is an upside of many datasets we surveyed. However, inclusion, representation, and diversity can be defined in different ways according to the problem at hand. Individuals would rather be included on their own terms, and decide whether and how they should be represented. The problems of diversity and robustness have some clear commonalities, as the former can be seen as a means towards the latter, but it seems advisable to maintain a clear separation between the two and to avoid equating either one with fairness. Algorithmic fairness will not be "solved" simply by collecting more data, or granting equal performance across different groups identified by a given sensitive attribute.

4.2.4 Sensitive Attribute Labeling

Motivation. Datasets are often taken as factual information that supports objective computation and pattern extraction. The etymology of the word "data", meaning "given", is rather revealing in this sense. In contrast, research in human-computer interaction, computer-supported cooperative work, and critical data studies argues that this belief is superficial, limited, and potentially harmful [181, 584].

Data is, quite simply, a human-influenced entity [564], determined by a chain of discretionary decisions on measurement, sampling and categorization, which shape how and by whom data will be collected and annotated, according to which taxonomy and based on which guidelines. Data science professionals, often more aware of the context surrounding data than theoretical researchers, report a significant awareness of how curation and annotation choices influence their data and its relation with the underlying phenomena [584]. In an interview, a senior text classification researcher responsible for ground truth annotation shows awareness of their own influence on datasets by stating "I am the ground truth." [584].

Sensitive attributes, such as race and gender, are no exception in this regard. Inconsistencies in racial annotation are quite common within the same system [518] and, even more so, across different systems [440, 701]. External annotation (either human or algorithmic) is essentially based on co-occurrence of specific traits with membership in a group, thus running the risk of encoding and reinforcing stereotypes. Self-reported labels overcome this issue, although they are still based on an imposed taxonomy, unless provided in an open-ended fashion. In this section, we discuss the practices through which sensitive attributes are annotated in datasets used in algorithmic fairness research, which are summarized in Table 4.5.

Table 4.5 Approaches to demographic data procurement.

Approach	Example datasets
Self-reported labels	Bing US Queries, MovieLens, Libimset, Adult, HMDA, Law School, Sushi, Willingness-to-Pay for Vaccine
Expert labels	Pilot Parliaments Benchmark
Non-expert labels	CelebFaces Attributes, Diversity in Faces, FairFace, Occupations in Google Images
ML algorithm	Racial Faces in the Wild, Instagram Photos, BUPT Faces, UTK Face
ML algorithm + annotators	FairFace, Open Images Dataset
Rule- / knowledge-based algorithm	RtGender, Bias in Bios, Demographics on Twitter, TwitterAAE

Procurement of sensitive attributes. Self-reported labels for sensitive attributes are typical of datasets that represent users of a service, who can report their demographics during registration (Bing US Queries, MovieLens, Libimseti), or were collected through surveys (HMDA, Adult, Law School, Sushi, Willingness-to-Pay for Vaccine). These are all resources for which collection of protected attributes was envisioned at design, potentially as an optional step. However, when sensitive attributes are not available, their annotation may be possible through different mechanisms.

A common approach is having sensitive attributes labeled by non-experts, often workers hired on crowdsourcing platforms. CelebFaces Attributes Dataset (CelebA) features images of celebrities from the CelebFaces dataset, augmented with annotations of landmark location and categorical attributes, including gender, skin tone and age, which were annotated by a "professional labeling company" [512]. In a similar fashion, Diversity in Faces consists of images labeled with gender and age by workers hired through the Figure Eight crowdsourcing platform, while the creators of FairFace hired workers on Amazon Mechanical Turk

to annotate gender, race, and age in a public image dataset. This practice also raises concerns about fair compensation of labor, which are not discussed in this work.

Some creators employ algorithms to obtain sensitive labels. Face datasets curators often resort to the Face++ API (Racial Faces in the Wild, Instagram Photos, BUPT Faces) or other algorithms (UTK Face, FairFace). In essence labeling is classifying; hence measuring and reporting accuracy for this procedure would be in order, but rarely happens. Creators occasionally note that automated labels were validated (FairFace) or substantially enhanced (Open Images Dataset) by human annotators, and very seldom report inter-annotator agreement (Occupations in Google Images).

Other examples of external labels include the geographic origin of candidates in resumes (CVs from Singapore), political leaning of US Twitter profiles (Twitter Political Searches), English dialect of tweets (TwitterAAE), and gender of subjects featured in image search results for professions (Occupations in Google Images). Annotation may also rely on external knowledge bases such as Wikipedia,[6] as is the case with RtGender. In situations where text written by individuals is available, rule-based approaches that exploit gendered nouns ("woman") or pronouns ("she") are also applicable (Bias in Bios, Demographics on Twitter).

Some datasets may simply have no sensitive attribute. These are often used in works of individual fairness, but may occasionally support studies of group fairness. For example, dSprites is a synthetic computer vision dataset where regular covariates may play the role of sensitive variables [513]. Alternatively, datasets can be augmented with simulated demographics, as done by Madnani et al. [522] who randomly assigned a native language to test participants in ASAP, or through the technique of Burke et al. [102], which they demonstrate on MovieLens.

Face datasets. Posterior annotation is especially common in computer vision datasets. The Pilot Parliaments Benchmark, for instance, was devised as a testbed for face analysis algorithms. It consists of images of parliamentary representatives from three African and three European countries that were labeled by a surgical dermatologist with the Fitzpatrick skin type of the subjects [271]. This is a dermatological scale for skin color, which can be retrieved from people's appearance. On the contrary, annotations of race or ethnicity from a photo are simplistic at best, and it should be clear that they actually capture *perceived race* from the perspective of the annotator (FairFace, BUPT Faces). Careful nomenclature is an important first step to improve the transparency of a dataset and make the underlying context more visible.[7]

[6] https://en.wikipedia.org/wiki/Category:American_female_tennis_players

[7] In this thesis, we discuss sensitive attributes following the naming convention in the accompanying documentation of a dataset, avoiding a critical terminology discussion .

Similarly to Scheuerman et al. [701], we find that documentation accompanying face recognition datasets hardly ever describes how specific taxonomies for gender and race were chosen, conveying a false impression of objectivity. A description of the annotation process is typically present but minimal. For Multi-task Facial Landmark, for instance, we only know that "The ground truths of the related tasks are labeled manually" [872].

Annotation trade-offs. It is worth re-emphasizing that sensitive label assignment is a classification task that rests on assumptions. Annotation of race and gender in images, for example, is based on the idea that they can be accurately ascertained from pictures, which is an oversimplification of these constructs. The envisioned classes (e.g. binary gender) are another subjective choice stemming from the point of view of dataset curators and may reflect narrow or outdated conceptions and potentially harm the data subjects. In this regard, a quote from the curators of MS-Celeb-1M, who do not annotate race, but consider it for their sampling strategy, is particularly striking: "We cover all the major races in the world (Caucasian, Mongoloid, and Negroid)" [335]. For these reasons, the external annotation of sensitive attributes is controversial and inevitably influenced by dataset curators.

On the other hand, external annotation may be the only way to test specific biases. Occupations in Google Images, for instance, is an image dataset collected to study gender and skin tone diversity in image search results for various professions. The creators hired workers on Amazon Mechanical Turk to label the gender (male, female) and Fitzpatrick skin tone (Type 1–6) of the primary person in each image. The Pilot Parliaments Benchmark was also externally annotated to obtain a benchmark for the evaluation of face analysis technology, with a balanced representation of gender and skin type. Different purposes can motivate data collection and annotation of sensitive attributes. The purpose and objective should be clearly documented, while also reflecting on other uses and the potential for misuse of a dataset [292]. Dataset curators may use documentation to discuss these aspects and specify limitations for the intended use of a resource [630]. In the next section, we focus on documentation and why it represents a key component of data curation.

4.2.5 Transparency

Motivation. Transparent and accurate documentation is a fundamental part of data quality. Its absence may cause serious problems, including lack of reproducibility, concerns about scientific validity, ethical problems, and harms [50]. Clear documentation can shine a light on the inevitable choices made by dataset creators and on the context surrounding the data. In the absence of this information, the curation mechanism that mediates reality and data is hidden; the data becomes one with its context, to the point that the interpretation of numerical results can be misleading and overarching [43].

The "ground truth" labels (typically indicated with the letter y), which are the target of prediction tasks in some datasets, such as indications of recidivism in COMPAS, are especially sensitive in this regard. Indeed, not only accuracy and related quality metrics, but also algorithmic fairness measures such as sufficiency and separation [50] are based on y labels and the ability of ML algorithms to replicate them, implicitly granting them a special status of truthfulness. In reality, however, these labels may be biased and incorrect due to multiple causes, including, very frequently, a disconnect between what we aim to measure in an ideal construct space (e.g., offense in the case of COMPAS) and what we can actually measure in the observed space (e.g., arrest) [281]. Fair ML algorithms (measures) can only partly overcome (catch) these biases, and actually run the risk of further reifying them. Proper documentation does not solve this issue, but equips practitioners and researchers with the necessary awareness to handle these biases.

More broadly, good documentation should discuss and explain features, providing context on who collected and annotated the data, how, and for which purpose [199, 292]. This provides dataset users with information they can leverage to select appropriate datasets for their tasks and avoid unintentional misuse [292]. Other actors, such as reviewers, can also access the official documentation of a dataset to ensure that it is used in accordance with its stated purpose, guidelines, and terms of use [630]. Overall data documentation plays a fundamental role in increasing transparency and accountability [385], favoring responsible and reflexive data curation [405, 564], and the correct use of these resources [627].

Positive examples. In this survey, we find excellent documentation examples in datasets related to studies and experiments, including CheXpert, Framingham, and NLSY. Indeed, datasets curated by medical institutions and census offices are often well documented. The ideal source of good documentation are descriptor articles published in conjunction with a dataset (e.g. MIMIC-III), typically offering stronger guarantees than web pages in terms of quality and permanence. Official websites hosting and distributing datasets are also important for collecting updates, errata, and additional information that may not be available at the time of release. The Million Song Dataset and Goodreads Reviews, for instance, are available on websites that contain a useful overview of the respective dataset, a list of updates, code samples, pointers to documentation, and contacts for further questions.

Negative examples. On the other hand, some datasets are opaque and poorly documented. Among publicly available ones, Arrhythmia is distributed with a description of the features, but without context about the purposes, actors, and subjects involved in the data collection. Similarly, the whole curation process and composition of Multi-task Facial Landmark is described in a short paragraph, explaining that it consists of 10,000 outdoor face images from the web that were labeled manually with gender. Most face datasets suffer from opaque

documentation, especially regarding the choice of sensitive labels and their annotation. For semi-synthetic resources, proper documentation is especially important, to let users understand the broader applicability and implications of numerical analyses performed on a dataset. IBM HR Analytics is a resource on employee attrition, which the hosting website describes as containing "fictional data", without any additional information. Nevertheless, this data was plausibly generated in a principled fashion, and (even partial) disclosure of the underlying data generation mechanism would benefit dataset users.

Retrospective documentation. Good documentation can also be produced retrospectively [42, 287]. German Credit is an interesting example of a dataset that was poorly documented for decades, until the recent publication of a report correcting severe coding errors [329]. From the old documentation, it seemed possible to retrieve the sex of data subjects from a feature jointly encoding sex and marital status. The *dataset archaeology* work of Grömping [329] shows that this is not the case, which is of particular relevance for many algorithmic fairness works using this dataset with sex as a protected feature, as this feature is simply not available. Numerical results obtained in this setting may be an artifact of the wrong coding with which the dataset has been, and still is, officially distributed in the UCI Machine Learning Repository [774]. Until the report and the new redacted dataset [776] become well-known, the old version will remain prevalent and more mistakes will be made. In other words, while the documentation debt for this particular dataset has been retrospectively addressed (*opacity*), many algorithmic fairness works published after the report continue to use the German Credit dataset with sex as a protected attribute [34, 356, 514, 534, 810, 841]. This is an issue of documentation *sparsity*, where the right information exists but does not reach interested parties, including researchers and reviewers.

Documentation is a fundamental part of data curation, with most responsibility resting on creators. However, dataset users can also play a role in mitigating documentation debt by proactively looking for information about the resources they plan to use. Brief summaries discussing and motivating the chosen datasets can be included in scholarly articles, at least in supplementary materials when conflicting with page limitations. In fact, documentation debt is a problem for the whole research community, which can be addressed collectively with retrospective contributions and clarifications. We argue that it is also up to individual researchers to seek contextual information for situating the data they want to use.

In this section, we have analyzed issues connected to re-identification, consent, inclusivity, and transparency running across algorithmic fairness datasets. By describing a range of approaches and attention to these topics, we aim to make them more visible and concrete. On the one hand, this may prove valuable to inform post-hoc data interventions aimed at mitigating potential harms caused by existing datasets. On the other hand, as novel datasets

are increasingly curated, published, and adopted in fairness research, it is important to motivate these concerns, make them tangible, and distill existing approaches into best practices, which we summarize below, for future endeavors of data curation. Our recommendations complement (and do not replace) a growing body of work that studies key aspects in the life cycle of datasets [181, 292, 385, 405, 630, 644].

The social relevance of data, intended as the breadth and depth of societally useful insights afforded by datasets, is a central requirement in fairness research. Unfortunately, this may conflict with user privacy, favouring re-identification or leaving consideration of consent in the background. Consent should be considered during the initial design of a dataset, in accordance with existing frameworks, such as the FRIES framework outlined in the Consentful Tech project. Moreover, different strategies are available to alleviate concerns of re-identification, including noise injection, conservative release, and (semi)synthetic data generation. Algorithmic fairness is motivated by aims of justice and harm avoidance for people, which should be extended to data subjects with careful curatorial choices.

Inclusivity allows for a wider representation and supports analyses that take into account important groups. However, inclusivity is insufficient in itself. Ideally, inclusivity should begin with a clear definition of the objectives of a data collection effort [405]. This step is important to clarify the benefits of including a given population. Possible uses afforded by a dataset should always be considered, evaluating costs and benefits for the data subjects and the wider population. In the absence of these considerations, acritical inclusivity runs the risk of simply supporting system robustness across sensitive attributes, such as race and gender, rebranded as fairness.

Sensitive attributes are a key ingredient to measure inclusion and increase the social relevance of a dataset. Although often impractical, it is typically preferable for sensitive attributes to be self-reported by data subjects. Externally assigned labels and taxonomies can harm individuals by erasing their needs and points of view. Sensitive attribute labelling is thus a shortcut whose advantages and disadvantages should be carefully weighted and, if chosen, it should be properly documented. Possible approaches based on human labour include expert and non-expert annotation, while automated approaches range from simple rule-based systems to complex and opaque algorithms. To label is to classify, hence measuring and reporting per-group accuracy is in order. Some labeling endeavours are more sensible than others: while skin tone can arguably be retrieved from pictures, annotations of race from an image actually capture *perceived race* from the perspective of the annotator. Rigorous nomenclature favours better understanding and clarifies the subjectivity of certain labels.

Reliable documentation shines a light on inevitable choices made by dataset creators and on the context surrounding the data. This provides dataset users with information they can

leverage to select appropriate datasets for their tasks and avoid unintentional misuse. Datasets for which some curation choices are poorly documented may appear more objective at first sight. However, it should be clear that objective data and turbid data are very different things. Proper documentation increases transparency, trust, and understanding. At a minimum, it should include the purpose of a data artifact, a description of the sample, the features and related annotation procedures, along with an explicit discussion of the associated task, if any. It should also clarify who was involved in the different stages of the data development procedure, with special attention to annotation. Data documentation also supports reviewers and readers of academic research in assessing whether a dataset was selected with good reason and utilized in compliance with creators' guidelines.

4.3 Chapter Outcomes

Italian car insurance

The Italian Car Insurance dataset was collected to achieve objective **O1** and audit important algorithms that mediate access to and pricing of insurance for drivers in Italy. It was gathered from the most popular comparison website for RCA in Italy, following a driver-centric perspective and guaranteeing a reasonable market coverage. It was designed as a full-factorial experiment with noise control, allowing robust estimates of the effect of selected variables on access and pricing. More in detail, this resource has been used to answer the following research questions.

What are the factors that play a major role in setting RCA premiums?
We examined the prices stratified on each feature, averaged across each of the remaining factors. We found that driver age, city, vehicle, and claim history are important factors for RCA pricing, at least for the sample we considered. Contrary to our expectations, the levels we tested for mileage led to small average price differences, probably due to the low verifiability of this feature. Birthplace and gender also induced smaller average fluctuations, which we analyzed more in detail in light of their sensitive nature and existing legislation against their direct use.

Do gender and birthplace directly influence quoted premiums?
Both factors have a direct influence on the quotes offered to users: we paired driver profiles so that they only differ for gender or birthplace, and found that quotes provided to them vary frequently and substantially. These differences are greater than those present in control (identical) pairs.

More in detail, we analyzed the distribution of paired differences, finding that gender-related differences are centered around zero, confirming the finding that no gender is systematically at a disadvantage. However, some sizeable differences were measured in both directions, showing that gender can have a direct non-negligible influence on quoted price. Birthplace-related differences, on the other hand, exhibit patterns of systematic discrimination. Foreign-born drivers and natives of Naples are consistently charged more expensive premiums compared to drivers born in Milan, *ceteribus paribus*. We interpret these findings as a violation of Fairness Through Unawareness (FTU), which is the fairness principle that (most closely) aligns with European legislation on gender equality in insurance [226, 227] and Italian soft regulation against nationality-based discrimination [395, 777]. We repeated our data collection procedure on a single company's website, focusing on the most representative subset of our sample. Comparative analysis supports the key trends discussed above, confirming that the influence of the aggregator on quoted prices is moderate, if any.

Do riskier driver profiles see fewer quotes on comparison websites? We analyzed the frequency with which insurers appear on the result pages for different profiles, finding that some companies are systematically absent from result pages for certain driver segments. In summary, 18-year-olds, drivers with a bad claim history, and Naples residents appear to be the least desirable categories in our dataset: when they query the comparison website, they end up receiving, on average, fewer quotes. These results are consistent with anecdotal findings from IVASS on aggregator output variability, associating riskier profiles with fewer RCA quotes [394]. The evidence we found on our medium-size sample represents a confirmation that strategic choices seem to be in place, providing users of comparison websites with unequal opportunity and access to products based on their risk profile.

It should be noted that Italian Car Insurance consists of quotes for 2,160 driver profiles, a dataset of limited size and not fully representative of the Italian driving population at large. Moreover, we were only able to examine a subset of the relevant features, which does not fully characterize the behavior of the pricing algorithm. Although our experiments show a violation of FTU, we did not attempt to quantify the impact of the discrimination that we uncovered on Italian society at large. This would be a large and complex endeavor and an interesting target for future work. Despite its moderate size, Italian Car Insurance features several protected groups, and is attentive to inclusivity. The related chances of re-identification are small, if any, and it is not problematic with respect to individual consent. Moreover, it is thoroughly documented and no controversial practice was employed for the annotation of sensitive attributes. In general, this resource is aware of important desiderata for data curation summarized below.

Curatorial best practices

To study curation topics more broadly, we have analyzed issues connected to re-identification, consent, inclusivity, and transparency running across over 200 algorithmic fairness datasets. By describing a range of approaches and attention to these topics, we made them more visible and concrete. On the one hand, this may prove valuable to inform post-hoc data interventions aimed at mitigating potential harms caused by existing datasets. On the other hand, as novel datasets are increasingly curated, published, and adopted in fairness research, it is important to motivate these concerns, make them tangible, and distill existing approaches into best practices, which we summarize below, for future data curation efforts (**O2**).

The social relevance of data, intended as the breadth and depth of societally useful insights afforded by datasets, is a central requirement in fairness research. Unfortunately, this may conflict with user privacy, favouring re-identification or leaving consideration of consent in the background. Consent should be considered during the initial design of a dataset, in accordance with existing frameworks, such as the FRIES framework outlined in the Consentful Tech project. Moreover, different strategies are available to alleviate concerns of re-identification, including noise injection, conservative release, and (semi)synthetic data generation. Algorithmic fairness is motivated by aims of justice and harm avoidance for people, which should be extended to data subjects with careful curatorial choices.

Inclusivity allows for a wider representation and supports analyses that take into account important groups. It is nuanced and may require equal representation across all groups or proportional to their prevalence in a wider population. Importantly, inclusivity is insufficient in itself. Ideally, inclusivity should begin with a clear definition of the objectives of a data collection effort [405]. This step is fundamental to clarify the benefits of including a given population. Possible uses afforded by a dataset should always be considered, evaluating costs and benefits for the data subjects and the wider population. In the absence of these considerations, acritical inclusivity runs the risk of simply supporting system robustness across sensitive attributes, such as race and gender, rebranded as fairness.

Sensitive attributes are a key ingredient to measure inclusion and increase the social relevance of a dataset. Although often impractical, it is typically preferable for sensitive attributes to be self-reported by data subjects. Externally assigned labels and taxonomies can harm individuals by erasing their needs and points of view. Sensitive attribute labelling is thus a shortcut whose advantages and disadvantages should be carefully weighted and, if chosen, it should be properly documented. Possible approaches based on human labour include expert and non-expert annotation, while automated approaches range from simple rule-based systems to complex and opaque algorithms. To label is to classify, hence measuring and reporting per-group accuracy is in order. Some labeling endeavours are more sensible than

others: while skin tone can arguably be retrieved from pictures, annotations of race from an image actually capture *perceived race* from the perspective of the annotator. Rigorous nomenclature favours better understanding and clarifies the subjectivity of certain labels.

Reliable documentation shines a light on inevitable choices made by dataset creators and on the context surrounding the data. This provides dataset users with information they can leverage to select appropriate datasets for their tasks and avoid unintentional misuse. Datasets for which some curation choices are poorly documented may appear more objective at first sight. However, it should be clear that objective data and turbid data are very different things. Proper documentation increases transparency, trust, and understanding. At a minimum, it should include the purpose of a data artifact, a description of the sample, the features and related annotation procedures, along with an explicit discussion of the associated task, if any. It should also clarify who was involved in the different stages of the data development procedure, with special attention to annotation. Data documentation also supports reviewers and readers of academic research in assessing whether a dataset was selected with good reason and utilized in compliance with creators' guidelines.

Our recommendations complement (and do not replace) a growing body of work studying key aspects in the life cycle of datasets [181, 292, 405, 630, 644]. Indeed, these data curation topics can be further investigated and elaborated upon; in the next chapter we focus on a specific aspect of dataset annotation.

Chapter 5

Dataset Annotation

Sensitive attributes, such as race and gender, are of primary importance in algorithmic fairness datasets. Their availability enables fairness audits [101], measures of diversity [571], and disaggregated analyses [48], which can highlight sizeable differences in opportunities and outcomes across populations of interest [250]. Unfortunately, sensitive demographic data, such as race or sex of subjects, are often not available, since practitioners find several barriers to obtaining these data, both during model development and after deployment. Among these barriers, legislation plays an important role, prohibiting the collection of sensitive attributes in some domains [82]. Even in the absence of explicit prohibition, privacy-by-design standards and a data minimization ethos often push companies in the direction of avoiding the collection of sensitive data from their customers. Similarly, the prospect of negative media coverage is a clear concern, so companies often err on the side of caution and inaction [17]. The unavailability of these data makes the measurement of fairness and diversity nontrivial. For these reasons, in a recent survey of industry professionals, most of the respondents stated that the availability of tools that support fairness auditing in the absence of individual-level demographics would be very useful [370]. In other words, the problem of measuring group fairness when the values of the sensitive attributes are unknown, called *fairness under unawareness*, is pressing and requires ad hoc solutions.

O1: Develop annotation methods tailored for measuring classifier fairness under unawareness of sensitive attributes.

In the this chapter, we tackle the problem of annotating a dataset to estimate fairness under unawareness by using techniques from *quantification* [316], a supervised learning task concerned with estimating, rather than the class labels of individual data points, the class prevalence values for samples of such data points, i.e., group-level quantities, such as the percentage of women in a given sample. Quantification methods address two pressing facets

of the fairness under unawareness problem: (1) their estimates are robust to *distribution shift* (i.e., to the fact that the distribution of the labels in the unlabeled data may significantly differ from the analogous distribution in the training data), which is often inevitable since populations evolve, and demographic data are unlikely to be representative of every condition encountered at deployment time; (2) they allow the estimation of sample-level quantities but do not allow the inference of sensitive attributes at the individual level, which is beneficial since the latter might lead to the inappropriate and nonconsensual utilization of this sensitive information, reducing individuals' agency over data [18]. Quantification methods achieve these goals by *directly* targeting group-level prevalence estimates. They do so through a variety of approaches, including, e.g., dedicated loss functions, task-specific adjustments, and *ad hoc* model selection procedures.

More in detail, we target **O1** through the following contributions. (1) **Quantifying fairness under unawareness**. We show that measuring fairness under unawareness can be cast as a problem of prevalence estimation at the sample level, and solved with approaches of proven consistency established in the quantification literature. This fact suggests that individual label annotation is not necessarily required to perform fairness audits and disaggregated analyses. (2) **Experimental protocols for five major challenges**. Drawing from the algorithmic fairness literature, we identify five important challenges that arise in estimating fairness under unawareness. These challenges are encountered in real-world applications, and include the non-stationarity of the processes generating the data and the variable cardinality of the available samples. For each such challenge, we define and formalise a precise experimental protocol, through which we compare the performance of quantifiers (i.e., sample-level prevalence estimators) generated by six different quantification methods. (3) **Decoupling sample-level and individual-level inferences.** We consider the problem of potential model misuse to maliciously infer demographic characteristics at the individual level, which is a concern for *proxy methods*. Through a set of experiments, we demonstrate two methods that yield precise estimates of demographic disparity but poor classification performance, thus decoupling the (desirable) objective of sample-level prevalence estimation from the (undesirable) objective of individual-level class label prediction.

The outline of this chapter is as follows. Section 5.1 summarizes the notation and background. Section 5.2 introduces related works. After giving a primer on quantification, with an emphasis on the approaches we consider in this work, Section 5.3 shows how these approaches can be leveraged to augment datasets with sample-level annotations and measure fairness under unawareness of sensitive attributes. Section 5.4 presents our experiments, in which we tackle, one by one, each of the five major challenges mentioned above. We then summarize and

Table 5.1 Main notational conventions used in this chapter.

symbol	meaning
$\mathbf{x} \in \mathscr{X}$	a data point, i.e., a vector of non-sensitive attribute values
$s \in \mathscr{S}$	a value for the sensitive attribute , with $\mathscr{S} = \{0,1\}$
$y \in \mathscr{Y}$	a class from the target domain $\mathscr{Y} = \{\ominus, \oplus\}$
$X, S, Y, \hat{Y}$	random variables for data points, non-sensitive attributes, classes, and class predictions
$h(\mathbf{x})$	a classifier $h : \mathscr{X} \to \mathscr{Y}$ issuing predictions in $\mathscr{Y}$ for data points in $\mathscr{X}$
$k(\mathbf{x})$	a classifier $k : \mathscr{X} \to \mathscr{S}$ issuing predictions in $\mathscr{S}$ for data points in $\mathscr{X}$
σ	a sample, i.e., a non-empty set of data points drawn from $\mathscr{X}$
$p_\sigma(s)$	true prevalence of sensitive attribute value s in sample σ
$\hat{p}_\sigma(s)$	estimate of the prevalence of sensitive attribute value s in sample σ
$\hat{p}^q_\sigma(s)$	estimate $\hat{p}_\sigma(s)$ obtained via quantifier q
$q(\sigma)$	a quantifier $q : 2^{\mathscr{X}} \to [0,1]$ estimating the prevalence of the positive class of sensitive attribute S in a sample
$\mathscr{D}_1$	set of pairs $(\mathbf{x}_i, y_i) \in (\mathscr{X}, \mathscr{Y})$ for training classifier $h(\mathbf{x})$
$\mathscr{D}_2$	set of pairs $(\mathbf{x}_i, s_i) \in (\mathscr{X}, \mathscr{S})$ for training quantifier $q(\sigma)$
$\mathscr{D}_3$	set of points $\mathbf{x}_i \in \mathscr{X}$ to which $h(\mathbf{x})$ and $q(\sigma)$ are to be applied
$\mathscr{D}_2^y$	short for $\mathscr{D}_2^{\hat{Y}=y} = \{(\mathbf{x}_i, s_i) \in \mathscr{D}_2 \mid h(\mathbf{x}_i) = y\}$
$\mathscr{D}_3^y$	short for $\mathscr{D}_3^{\hat{Y}=y} = \{\mathbf{x}_i \in \mathscr{D}_3 \mid h(\mathbf{x}_i) = y\}$
$\check{\mathscr{D}}$	a set derived from $\mathscr{D}$ according to an experimental protocol among those detailed in Sections 5.4.3–5.4.7

discuss these results (Section 5.5) and present the chapter outcomes (Section 5.6), describing key limitations and directions for future work.

The code to reproduce the experiments presented in this chapter is available at https://github.com/alessandro-fabris/ql4facct.

5.1 Preliminaries

5.1.1 Notation

In this chapter, we use the following notation, summarized in Table 5.1. By $\mathbf{x}$ we indicate a data point drawn from a domain $\mathscr{X}$, represented through a set X of nonsensitive attributes

(i.e., features). We use S to denote a sensitive attribute that takes values in $\mathscr{S} = \{0,1\}$, and by $s \in \mathscr{S}$ a value that S can take. By Y we indicate a class (representing the target of a prediction task) taking values in a binary domain $\mathscr{Y} = \{\ominus, \oplus\}$, and by $y \in \mathscr{Y}$ a value that Y can take. The symbol σ denotes a *sample*, i.e., a non-empty set of data points drawn from $\mathscr{X}$. By $p_\sigma(s)$ we indicate the true prevalence of an attribute value s in the sample σ, while by $\hat{p}^q_\sigma(s)$ we indicate the estimate of this prevalence obtained by means of a quantifier q, which we define as a function $q : 2^{\mathscr{X}} \to [0,1]$. Since $0 \leq p_\sigma(s) \leq 1$ and $0 \leq \hat{p}^q_\sigma(s) \leq 1$ for all $s \in \mathscr{S}$, and since $\sum_{s\in\mathscr{S}} p_\sigma(s) = \sum_{s\in\mathscr{S}} \hat{p}^q_\sigma(s) = 1$, the $p_\sigma(s)$'s and the $\hat{p}^q_\sigma(s)$'s form two probability distributions in $\mathscr{S}$. We also introduce the random variable $\hat{Y}$, which denotes a predicted label. By $\Pr(V = v)$ we indicate, as usual, the probability that a random variable V takes the value v, which we shorten as $\Pr(v)$ when V is clear from the context, since X, S, Y can also be seen as random variables. By $h : \mathscr{X} \to \mathscr{Y}$ we indicate a binary classifier that assigns classes in $\mathscr{Y}$ to data points in $\mathscr{X}$; by $k : \mathscr{X} \to \mathscr{S}$ we instead indicate a binary classifier that assigns sensitive attribute values in $\mathscr{S}$ to data points (e.g., that predicts the sensitive attribute value of a certain data item $\mathbf{x}$). It is worth re-emphasizing that both h and k only use nonsensitive attributes X as input variables, For ease of use, we will interchangeably write $h(\mathbf{x}) = y$ or $h_y(\mathbf{x}) = 1$, and $k(\mathbf{x}) = s$ or $k_s(\mathbf{x}) = 1$.

5.1.2 Background

Several criteria for group fairness have been proposed in the machine learning literature, typically requiring equalization of some conditional or marginal property of the distribution of sensitive variable S, ground truth Y, and classifier estimate $\hat{Y}$ [219, 345, 592]. The main criteria of observational group fairness [50], i.e., the ones computed directly from groupwise confusion matrices, are defined as follows.

Definition 1. Given a classifier $h : \mathscr{X} \to \mathscr{Y}$ issuing predictions $\hat{y} = h(\mathbf{x})$, and given the respective ground truth labels y, the following groupwise disparities with respect to the attribute S can be defined.

$$\text{Demographic Disparity: } \delta_h^{S,\text{DD}} = \Pr(\hat{Y} = \oplus | S = 1) - \Pr(\hat{Y} = \oplus | S = 0)$$
$$\text{True Positive Rate Disparity: } \delta_h^{S,\text{TPRD}} = \Pr(\hat{Y} = \oplus | S = 1, Y = \oplus) - \Pr(\hat{Y} = \oplus | S = 0, Y = \oplus)$$
$$\text{True Negative Rate Disparity: } \delta_h^{S,\text{TNRD}} = \Pr(\hat{Y} = \ominus | S = 1, Y = \ominus) - \Pr(\hat{Y} = \ominus | S = 0, Y = \ominus)$$
$$\text{Positive Predicted Value Disparity: } \delta_h^{S,\text{PPVD}} = \Pr(Y = \oplus | S = 1, \hat{Y} = \oplus) - \Pr(Y = \oplus | S = 0, \hat{Y} = \oplus)$$
$$\text{Negative Predicted Value Disparity: } \delta_h^{S,\text{NPVD}} = \Pr(Y = \ominus | S = 1, \hat{Y} = \ominus) - \Pr(Y = \ominus | S = 0, \hat{Y} = \ominus)$$

□

Demographic disparity, for example, measures whether the prevalence of the positive class is the same between subpopulations identified by the sensitive attribute S; a value $\delta_h^{S,\mathrm{DD}} = 0$ indicates maximum fairness, while values of $\delta_h^{S,\mathrm{DD}} = -1$ or $\delta_h^{S,\mathrm{DD}} = +1$ indicate minimum fairness, i.e., maximum advantage for $S = 0$ over $S = 1$ or vice versa. We illustrate the problem of measuring fairness under unawareness using an example focused on demographic disparity.

Example 1. Assume that S stands for "race", $S = 1$ for "African-American" and $S = 0$ for "White",[1], and that the classifier, deployed by a bank, is responsible for recommending loan applicants for acceptance, classifying them as "grant" ($\oplus$) or "deny" ($\ominus$). For simplicity, let us assume that the outcome of the classifier will be translated directly into a decision without human supervision. The bank might want to check that the fraction of loan recipients out of the total number of applicants is approximately the same in the African-American and White subpopulations. In other words, the bank might want $\delta_h^{S,\mathrm{DD}}$ to be close to 0. Of course, if the bank is aware of the race of each applicant, this constraint is very easy to check and, potentially, enforce. If the bank is unaware of the applicants' race, the problem is not trivial, and can be addressed by the method we propose in this chapter.

5.2 Related Work

5.2.1 Fairness under Unawareness

Unavailability of sensitive attribute values poses a major challenge for internal and external fairness audits. When these values are unknown, it is sometimes possible to seek expert advice to annotate them (Section 4.2.4). Alternatively, disclosure procedures have been proposed for subjects to provide their sensitive attributes to a trusted third party [792] or to share them encrypted [442]. Another line of research studies the problem of reliably estimating measures of group fairness, in classification [27, 140, 417] and ranking [299, 449], without access to sensitive attributes, via proxy variables.

Chen et al. [140] is the work most closely related to ours. The authors study the problem of estimating the demographic disparity of a classifier, exploiting the values of non-sensitive attributes X as proxies to infer the value of the sensitive variable S. Starting from a naïve

[1] While acknowledging its limitations [746] we follow the race categorization adopted by the US Census Bureau wherever possible.

approach, dubbed *threshold estimator* (**TE**), which estimates $\mu(s) = \Pr(\hat{Y} = \oplus | S = s)$ as

$$\hat{\mu}^{\mathrm{TE}(s)} = \frac{\sum_{\mathbf{x}_i} k_s(\mathbf{x}_i) h_{\oplus}(\mathbf{x}_i)}{\sum_{\mathbf{x}_i} k_s(\mathbf{x}_i)} \tag{5.1}$$

i.e., by using a hard classifier $k_s : \mathscr{X} \to \{0,1\}$ (which outputs Boolean decisions regarding membership in a sensitive group $S = s$), they propose a *weighted estimator* (**WE**) with better convergence properties.

$$\hat{\mu}^{\mathrm{WE}}(s) = \frac{\sum_{\mathbf{x}_i} \pi_s(\mathbf{x}_i) h_{\oplus}(\mathbf{x}_i)}{\sum_{\mathbf{x}_i} \pi_s(\mathbf{x}_i)} \tag{5.2}$$

WE exploits a soft classifier $\pi_s : \mathscr{X} \to [0,1]$ that outputs posterior probabilities $\Pr(s|\mathbf{x}_i)$. The posteriors represent the probability that the classifier attributes to the fact that $\mathbf{x}_i$ belongs to the subpopulation with sensitive attribute $S = s$. The authors argue that the naïve estimator of Equation (5.1) has a tendency to exaggerate disparities, and show that WE mitigates this problem under the hypothesis that $\pi_s(\mathbf{x}_i)$ produces well-calibrated posterior probabilities. A contribution of this chapter is to show that TE and WE are just instances of a broad family of estimators (Proposition 2). Moreover, we consider alternative methods from the same family and show them to outperform both TE and WE on an extensive suite of experiments (Section 5.4).

Kallus et al. [417] study the problem of measuring a classifier's demographic disparity, true positive rate disparity, and true negative rate disparity in a setting with access to a primary dataset involving $(\hat{Y}, Z)$ and an auxiliary dataset involving (S, Z), where Z is a generic set of proxy variables, potentially disjoint from X. They show that reliably estimating the demographic disparity of a classifier issuing predictions $\hat{Y}$ when Z is not highly informative with respect to $\hat{Y}$ or S is infeasible. Moreover, they provide upper and lower bounds for the true value of the estimand in a setting where the primary and auxiliary datasets are drawn from marginalisations of a common joint distribution. This chapter departs from this setting in two important ways, focusing on realistic conditions for internal fairness audits. First, we take into account the non-stationarity of the processes generating the data and do not assume the primary and auxiliary dataset to be marginalisations of the same joint distribution. Rather, we identify different sources of distribution shift and formalize them into protocols to test the performance of different estimators in a more realistic setting (Sections 5.4.3–5.4.7). Second, we hypothesize that, from within the company deploying a classifier $h(\mathbf{x})$, the available proxy variables Z comprise X, and thus are highly informative with respect to $\hat{Y}$.

Awasthi et al. [27] characterize the structure of the best estimator for sensitive attributes when the final estimand is a classifier's disparity in true positive rates across protected groups.

They show that the test accuracy of the attribute classifier and its performance as an estimator of the true positive rate disparity are not necessarily correlated. We contribute to this line of research, demonstrating the possibility to decouple the *classification* performance of a model when deployed for sensitive attribute inference at the individual level, which constitutes a privacy infringement, from its *quantification* performance in applications where it is used for group-level estimates (Section 5.4.8). This line of work opens the possibility of developing estimators that reliably measure group fairness under unawareness of sensitive attributes, while guaranteeing privacy at the individual level.

5.2.2 Quantification and Fairness

The application of quantification methods in algorithmic fairness research is not entirely new. Biswas and Mukherjee [76] study the problem of enforcing fair classification under distribution shift, which potentially affects different demographic groups at different rates. They define a notion of fairness based on the proportionality between the prevalence of positives in a protected group $S = s$ and the group-specific acceptance rate of a classifier issuing predictions $\hat{Y}$. This notion, called *proportional equality*, is defined by the quantity

$$\mathrm{PE} = \left| \frac{\Pr(Y = \oplus | S = 1)}{\Pr(Y = \oplus | S = 0)} - \frac{\Pr(\hat{Y} = \oplus | S = 1)}{\Pr(\hat{Y} = \oplus | S = 0)} \right|$$

calculated on a test set $\mathscr{D}$, where low values of PE correspond to fairer predictions $\hat{Y}$. In the presence of distribution shift between training and testing conditions, the true group-specific prevalences $\Pr(Y = \oplus | S = 1)$ and $\Pr(Y = \oplus | S = 0)$ are unknown. The authors use an approach from the quantification literature to estimate these prevalence values, integrating them into a wider system aimed at optimizing PE.

In other words, previous work applying quantification to problems of algorithmic fairness concentrates on *enforcing* classifier fairness under unawareness of *target labels*. Our work, on the other hand, aims at *measuring* classifier fairness under unawareness of *sensitive attributes*.

5.3 Measuring Fairness under Unawareness: A Quantification-Based Method

In this section, we first present a primer on quantification (Section 5.3.1), and then show how to measure fairness under unawareness with quantification (Section 5.3.2), discussing the properties of the resulting estimators.

5.3.1 Learning to Quantify

Quantification (also known as *supervised prevalence estimation*, or *learning to quantify*) is the task of training, by means of supervised learning, a predictor that estimates the relative frequency (also known as *prevalence*, or *prior probability*) of the classes of interest in a sample of unlabeled data points, where the data used to train the predictor are a set of labeled data points; see González et al. [316] for a survey of quantification research.

Definition 2. Given a sample σ of data points $\mathbf{x} \in \mathscr{X}$, with unknown target labels in domain $\mathscr{S}$, a *quantifier* $q(\sigma)$ is an estimator $q : 2^{\mathscr{X}} \rightarrow [0,1]$ that predicts the prevalence of class s in the sample σ as $\hat{p}^q_\sigma(s) = q(\sigma)$.

Remark 1. The above definition is deliberately broad to include the trivial *classify and count* baseline introduced below. In practice, a method is *truly* quantification based when explicitly targeting prevalence estimates, rather than simply treating them as a byproduct of classification. This includes methods that use dedicated loss functions, task-specific adjustments, and ad hoc model selection procedures. Typically, the prevalence estimates issued by these methods display desirable properties of unbiasedness and convergence.

Quantification can be trivially solved via classification, i.e., by classifying all the unlabelled data points using a standard classifier, counting, for each class, the data points that have been assigned to the class and normalizing. However, it has unequivocally been shown (see, among many others, Fernandes Vaz et al. [262], Forman [277], González et al. [316], González-Castro et al. [317], Moreo and Sebastiani [580]) that solving quantification by means of this *classify and count* (CC) method is suboptimal and that there are more accurate quantification methods. The key reason behind this is the fact that many applicative scenarios suffer from *distribution shift*, therefore, the class prevalence values in the training set may substantially differ from the class prevalence values in the unlabeled data that the classifier issues predictions for [578]. The presence of distribution shift means that the well-known IID assumption, on which most learning algorithms for training classifiers are based, does not hold; in turn, this means that CC will perform suboptimally on scenarios that exhibit distribution shift, and that the higher the amount of shift, the worse we can expect CC to perform.

A wide variety of quantification methods have been defined in the literature. In the experiments presented in this chapter, we compare six such methods, which we briefly present in this section. One of them is the trivial CC baseline; we have chosen the other five methods over other contenders because they are simple and proven, and because some of them (especially the ACC, PACC, SLD and HDy methods; see below) have shown top-notch performance in recent comparative tests run in other domains [579, 580]. We briefly describe

them here, with direct reference to the application we are interested in, i.e., estimating the prevalence of a protected subgroup.

As mentioned above, an obvious way to solve quantification (used, among others, in Equation (5.1)) is by aggregating the predictions of a "hard" classifier, i.e., a classifier $k_s : \mathscr{X} \to \{0,1\}$ that outputs Boolean decisions regarding membership in a sensitive group (defined by constraint $S = s$). The (trivial) *classify and count* (**CC**) quantifier then comes down to computing

$$\hat{p}_\sigma^{\mathrm{CC}}(s) = \frac{\sum_{\mathbf{x}_i \in \sigma} k_s(\mathbf{x}_i)}{|\sigma|}. \tag{5.3}$$

Alternatively, quantification methods can use a "soft" classifier $\pi_s : \mathscr{X} \to [0,1]$ that produces posterior probabilities $\Pr(s|\mathbf{x}_i)$. The resulting *probabilistic classify and count* quantifier (**PCC**) [57] is defined by the equation

$$\hat{p}_\sigma^{\mathrm{PCC}}(s) = \frac{\sum_{\mathbf{x}_i \in \sigma} \pi_s(\mathbf{x}_i)}{|\sigma|}. \tag{5.4}$$

It should be noted that PCC and CC are clearly related to WE and TE, summarized by Equations (5.1) and (5.2), as shown later in Proposition 2.

A different and popular quantification method consists of applying an *adjustment* to the prevalence $\hat{p}_\sigma^{\mathrm{CC}}(s)$ estimated through "classify and count". It is easy to check that, in the binary case, the true prevalence $p_\sigma(s)$ and the estimated prevalence $\hat{p}_\sigma^{\mathrm{CC}}(s)$ are such that

$$p_\sigma(s) = \frac{\hat{p}_\sigma^{\mathrm{CC}}(s) - \mathrm{fpr}_{k_s}}{\mathrm{tpr}_{k_s} - \mathrm{fpr}_{k_s}} \tag{5.5}$$

where tpr_{k_s} and fpr_{k_s} stand for *true positive rate* and *false positive rate* of the classifier k_s used to obtain $\hat{p}_\sigma^{\mathrm{CC}}(s)$. The values of tpr_{k_s} and fpr_{k_s} are unknown, but can be estimated through a k-fold cross-validation on the training data. This boils down to using the results $k_s(\mathbf{x}_i)$ obtained in the k-fold cross-validation (i.e., $\mathbf{x}_i$ ranges on the training items) in equations

$$\hat{\mathrm{tpr}}_{k_s} = \frac{\sum_{\{(\mathbf{x}_i,s_i)|s_i=s\}} k_s(\mathbf{x}_i)}{|\{(\mathbf{x}_i,s_i)|s_i = s\}|} \qquad \hat{\mathrm{fpr}}_{k_s} = \frac{\sum_{\{(\mathbf{x}_i,s_i)|s_i\neq s\}} k_s(\mathbf{x}_i)}{|\{(\mathbf{x}_i,s_i)|s_i \neq s\}|}. \tag{5.6}$$

We obtain estimates of $p_\sigma^{\mathrm{ACC}}(s)$, which define the *adjusted classify and count* method (**ACC**) [277], by replacing tpr_{k_s} and fpr_{k_s} in Equation 5.5 with estimates of Equation (5.6), i.e.,

$$\hat{p}_\sigma^{\mathrm{ACC}}(s) = \frac{\hat{p}_\sigma^{\mathrm{CC}}(s) - \hat{\mathrm{fpr}}_{k_s}}{\hat{\mathrm{tpr}}_{k_s} - \hat{\mathrm{fpr}}_{k_s}}. \tag{5.7}$$

If the soft classifier $\pi_s(\mathbf{x}_i)$ is used in place of $k_s(\mathbf{x}_i)$, analogues of $\hat{\text{tpr}}_{k_s}$ and $\hat{\text{fpr}}_{k_s}$ from Equation (5.6) can be defined as

$$\hat{\text{tpr}}_{\pi} = \frac{\sum_{\{(\mathbf{x}_i,s_i)|s_i=s\}} \pi_s(\mathbf{x}_i)}{|\{(\mathbf{x}_i,s_i)|s_i=s\}|} \qquad \hat{\text{fpr}}_{\pi} = \frac{\sum_{\{(\mathbf{x}_i,s_i)|s_i\neq s\}} \pi_s(\mathbf{x}_i)}{|\{(\mathbf{x}_i,s_i)|s_i\neq s\}|}. \tag{5.8}$$

We obtain $p_\sigma^{\text{PACC}}(s)$ estimates, which define the *probabilistic adjusted classify and count* method (**PACC**) [57], by replacing all factors on the right-hand side of Equation (5.7) with their "soft" counterparts from Equations (5.4) and (5.8), i.e.,

$$\hat{p}_\sigma^{\text{PACC}}(s) = \frac{\hat{p}_\sigma^{\text{PCC}}(s) - \hat{\text{fpr}}_{\pi}}{\hat{\text{tpr}}_{\pi} - \hat{\text{fpr}}_{\pi}}. \tag{5.9}$$

A further method is the one proposed in [690] (which we here call **SLD**, from the names of its proposers), which consists of training a probabilistic classifier and then using the Expectation–Maximization (EM) algorithm (i) to update (in an iterative, mutually recursive way) the posterior probabilities that the classifier returns, and (ii) to re-estimate the class prevalence values of the test set until convergence. This makes the method robust to distribution shift, since the iterative process allows the estimates of the prevalence values to become increasingly attuned to the changed conditions found in the unlabeled set. The pseudocode describing the SLD algorithm can be found in Appendix C.1.

Lastly, we consider **HDy** [317], a probabilistic binary quantification method that views quantification as the problem of minimizing the divergence (measured in terms of the Hellinger Distance) between two cumulative distributions of posterior probabilities returned by the classifier, one from the unlabeled examples and the other from a validation set. HDy looks for the mixture parameter α that best fits the validation distribution (consisting of a mixture of a "positive" and a "negative" distribution) to the unlabelled distribution, and returns α as the estimated prevalence of the positive class. Here, robustness to distribution shift is achieved by the analysis of the distribution of the posterior probabilities in the unlabeled set, which reveals how conditions have changed with respect to the training data. A more detailed description of HDy can be found in Appendix C.2.

5.3.2 Using Quantification to Measure Fairness under Unawareness

We assume the existence, in the operational setup, of three separate sets of data points:

- A *training set* $\mathscr{D}_1$ for h, $\mathscr{D}_1 = \{(\mathbf{x}_i, y_i) \mid \mathbf{x}_i \in \mathscr{X}, y_i \in \mathscr{Y}\}$, typically of large size, where h is the classifier whose fairness we want to measure. Given the difficulties inherent

in demographic data procurement mentioned in the introduction, we assume that the sensitive attribute S is not part of the vectorial representation X.

- A small *auxiliary set* $\mathscr{D}_2 = \{(\mathbf{x}_i, s_i) \mid \mathbf{x}_i \in \mathscr{X}, s_i \in \mathscr{S}\}$, containing demographic data, employed to train quantifiers for the sensitive attribute.
- A set $\mathscr{D}_3 = \{\mathbf{x}_i \mid \mathbf{x}_i \in \mathscr{X}\}$ of unlabelled data points, which are the data to which the classifier h is to be applied, representing the deployment conditions, i.e., the dataset we aim to augment. Alternatively, $\mathscr{D}_3$ could also be a labelled held-out test set available at a company, if it has acted proactively rather than reactively, for pre-deployment audits [663]. In our experiments we will use labelled data and call $\mathscr{D}_3$ the *test set*, on which the fairness of the classifier h should be measured.

It is worth re-emphasizing that, from the perspective of the estimation task at hand, i.e., estimating the fairness of the classifier h, $\mathscr{D}_2$ represents the quantifier's training set, while $\mathscr{D}_3$ is its test set.

Proposition 1. Observational measures of algorithmic fairness, such as the ones introduced in Definition 1, can be computed, under unawareness of sensitive attributes, by estimating the prevalence of the sensitive attribute in specific subsets of the test set.

Proof. We prove this statement for TPRD in Definition 1, which we recall below:

$$\text{True Positive Rate Disparity: } \delta_h^{S,\text{TPRD}} = \Pr(\hat{Y} = \oplus | S = 1, Y = \oplus) - \Pr(\hat{Y} = \oplus | S = 0, Y = \oplus)$$

Both terms in the above equation can be written as

$$\begin{aligned}\Pr(\hat{Y} = \oplus | S = s, Y = \oplus) &= \frac{\Pr(Y = \oplus, \hat{Y} = \oplus, S = s)}{\Pr(Y = \oplus, S = s)} \\ &= \underbrace{\frac{\Pr(S = s | Y = \oplus, \hat{Y} = \oplus)}{\Pr(S = s | Y = \oplus)}}_{\text{obtained from prevalence estimator}} \cdot \underbrace{\frac{\Pr(Y = \oplus, \hat{Y} = \oplus)}{\Pr(Y = \oplus)}}_{\text{known quantity}}\end{aligned}$$

In other words, TPRD can be calculated by estimating the prevalence of the sensitive attribute among the positives and the true positives in $\mathscr{D}_3$. Analogous results can be proven for other measures of observational fairness, under the assumption that Y and $\hat{Y}$ are known. □

Remark 2. This proposition is important for two reasons. First, it shows that inference of sensitive attributes at the individual level is not necessary to measure fairness under unawareness; rather, prevalence estimates in given subsets are sufficient. Second, it suggests that methods directly targeting prevalence estimates (i.e., *quantifiers*) are especially suited in this setting.

Notice that, for the purposes of a fairness audit, it is common to assume that the ground truth variable Y is available in $\mathscr{D}_3$. In the banking scenario of Example 1, this is only partially realistic, as the outcomes for the accepted applicants are eventually observed, but the outcomes for the rejected applicants remain unknown, leaving us with a problem of sample selection bias [40]. This is an instance of a general estimation problem, common to all fairness criteria that require knowledge of the ground truth variable Y, such as TPRD, TNRD, PPVD, and NPVD in Definition 1. This represents an open research problem [688, 810] that is beyond the scope of this work.

In the remainder of this chapter, we avoid this issue by focusing on a detailed study of demographic disparity (DD), which does not require information on ground truth Y, leaving additional measures of observational fairness for future work. Following [140], we write DD as

$$\delta_h^S = \Pr(\hat{Y} = \oplus | S = 1) - \Pr(\hat{Y} = \oplus | S = 0) = \mu(1) - \mu(0), \tag{5.10}$$

where

$$\mu(s) = \Pr(\hat{Y} = \oplus | S = s) \tag{5.11}$$

is the acceptance rate of the individuals in the group $S = s$. To estimate the demographic disparity of a classifier $h(\mathbf{x})$ in the test set $\mathscr{D}_3$, we can use any quantification approach from Section 5.3.1. Applying Bayes' theorem to Equation (5.11), we obtain

$$\begin{aligned} \mu(s) &= p_{\mathscr{D}_3}(\oplus|s) \\ &= p_{\mathscr{D}_3^{\oplus}}(s)\frac{p_{\mathscr{D}_3}(\oplus)}{p_{\mathscr{D}_3}(s)}, \end{aligned} \tag{5.12}$$

where we use $p_{\mathscr{D}_3}(\oplus)$ as a shorthand of $p_{\mathscr{D}_3}(h(\mathbf{x}) = \oplus)$, and where we have defined

$$\begin{aligned} \mathscr{D}_3^{\oplus} &= \{\mathbf{x} \in \mathscr{D}_3 \mid h(\mathbf{x}) = \oplus\} \\ \mathscr{D}_3^{\ominus} &= \{\mathbf{x} \in \mathscr{D}_3 \mid h(\mathbf{x}) = \ominus\}. \end{aligned}$$

Since $p_{\mathscr{D}_3}(\oplus)$ is known (it is the fraction of items in $\mathscr{D}_3$ that have been assigned class $\oplus$ by the classifier h), in order to compute $\mu(s)$ through Equation (5.12), for $s \in \{0,1\}$, we only need to estimate the prevalence values $\hat{p}_{\mathscr{D}_3^{\oplus}}(s)$ and $\hat{p}_{\mathscr{D}_3^{\ominus}}(s)$; the latter is needed to estimate the denominator of Equation (5.12), i.e., the prevalence $p_{\mathscr{D}_3}(s)$ of the sensitive attribute value

s in the entire test set $\mathscr{D}_3$, since

$$p_{\mathscr{D}_3}(s) = p_{\mathscr{D}_3^{\oplus}}(s) \cdot p_{\mathscr{D}_3}(\oplus) + p_{\mathscr{D}_3^{\ominus}}(s) \cdot p_{\mathscr{D}_3}(\ominus). \tag{5.13}$$

In order to compute $p_{\mathscr{D}_3^{\oplus}}(s)$ and $p_{\mathscr{D}_3^{\ominus}}(s)$ we can use a quantification-based approach, which can be easily integrated into existing machine learning workflows, as summarized by the method below.

Method. Quantification-Based Estimate of Demographic Disparity.

1. The classifier $h : \mathscr{X} \to \mathscr{Y}$ is trained on $\mathscr{D}_1$ and ready for deployment, e.g., to estimate the creditworthiness of individuals. The assumption that, at this training stage, we are unaware of the sensitive attribute S is due to the inherent difficulties in demographic data procurement already mentioned at the beginning of this chapter.

2. We use the classifier h to classify the auxiliary set $\mathscr{D}_2$, thus inducing a partition of $\mathscr{D}_2$ into $\mathscr{D}_2^{\oplus} = \{(\mathbf{x}_i, s_i) \in \mathscr{D}_2 \mid h(\mathbf{x}) = \oplus\}$ and $\mathscr{D}_2^{\ominus} = \{(\mathbf{x}_i, s_i) \in \mathscr{D}_2 \mid h(\mathbf{x}) = \ominus\}$.

3. We use $\mathscr{D}_2^{\oplus}$ as the training set for the quantifier $q_{\oplus}(s)$, whose task will be to estimate the prevalence of value s (e.g., African-American applicants) on sets of data points labeled with class $\oplus$ (e.g., creditworthy applicants). Likewise, we use $\mathscr{D}_2^{\ominus}$ as the training set for a quantifier $q_{\ominus}(s)$ whose task will be to estimate the prevalence of s on sets of data points labeled with $\ominus$. Intuitively, separate quantifiers specialized on different subpopulations (of positively and negatively classified individuals) should perform better than a single quantifier. The ablation study in Section 5.4.9 supports this hypothesis.

4. The classifier h is deployed, classifying the test set $\mathscr{D}_3$, thus inducing a partition of $\mathscr{D}_3$ into creditworthy $\mathscr{D}_3^{\oplus} = \{\mathbf{x} \in \mathscr{D}_3 \mid h(\mathbf{x}) = \oplus\}$ and non-creditworthy $\mathscr{D}_3^{\ominus} = \{\mathbf{x} \in \mathscr{D}_3 \mid h(\mathbf{x}) = \ominus\}$.

5. We apply the quantifier $q_{\oplus}$ to $\mathscr{D}_3^{\oplus}$ to obtain an estimate $\hat{p}^{q_{\oplus}}_{\mathscr{D}_3^{\oplus}}(s)$ of the prevalence of s in $\mathscr{D}_3^{\oplus}$, and we apply $q_{\ominus}$ to $\mathscr{D}_3^{\ominus}$ to obtain an estimate $\hat{p}^{q_{\ominus}}_{\mathscr{D}_3^{\ominus}}(s)$ of the prevalence of s in $\mathscr{D}_3^{\ominus}$. Recall from Section 5.1.1 that $\hat{p}^{q}_{\sigma}(s)$ denotes the prevalence of an attribute value s in a set σ as estimated via quantification method q.

6. To avoid numerical instability in the denominator of Equation (5.15) below, we apply Laplace smoothing to the estimated prevalence values $\hat{p}^{q_{\oplus}}_{\mathscr{D}_3^{\oplus}}(s)$ and $\hat{p}^{q_{\ominus}}_{\mathscr{D}_3^{\ominus}}(s)$. We use the variant that uses known incidence rates, using $\mathscr{D}_2^{\ominus}$ and $\mathscr{D}_2^{\oplus}$ as the control populations,

and assume a pseudocount $\alpha = 1/2$. We thus compute the smoothed estimator

$$\tilde{p}^{q_\oplus}_{\mathscr{D}_3^\oplus}(s) = \frac{\hat{p}^{q_\oplus}_{\mathscr{D}_3^\oplus}(s) \cdot |\mathscr{D}_3^\oplus| + p_{\mathscr{D}_2^\oplus}(s) \cdot \alpha \cdot |\mathscr{Y}|}{|\mathscr{D}_3^\oplus| + \alpha \cdot |\mathscr{Y}|} = \frac{\hat{p}^{q_\oplus}_{\mathscr{D}_3^\oplus}(s) \cdot |\mathscr{D}_3^\oplus| + p_{\mathscr{D}_2^\oplus}(s)}{|\mathscr{D}_3^\oplus| + 1}$$

and analogously for $\tilde{p}^{q_\ominus}_{\mathscr{D}_3^\ominus}(s)$.

7. Finally, we estimate the demographic disparity of h, defined in Equation (5.10), as

$$\hat{\delta}^S_h = \hat{\mu}(1) - \hat{\mu}(0) \tag{5.14}$$

where, as from Equations (5.12) and (5.13),

$$\hat{\mu}(s) = \tilde{p}^{q_\oplus}_{\mathscr{D}_3^\oplus}(s) \cdot \frac{p_{\mathscr{D}_3}(\oplus)}{\tilde{p}^{q_\oplus}_{\mathscr{D}_3^\oplus}(s) \cdot p_{\mathscr{D}_3}(\oplus) + \tilde{p}^{q_\ominus}_{\mathscr{D}_3^\ominus}(s) \cdot p_{\mathscr{D}_3}(\ominus)} \tag{5.15}$$

Remark 3. Therefore, prevalence estimates $\hat{p}^{q_\oplus}_{\mathscr{D}_3^\oplus}(s)$ and $\hat{p}^{q_\ominus}_{\mathscr{D}_3^\ominus}(s)$, obtained with a quantification method of the type introduced in Section 5.3.1, can be translated into estimates of the demographic disparity of a classifier using Equations (5.14) and (5.15). Importantly, the bias and variance of that estimate depend on the properties of the underlying quantification method, which have been characterized in the quantification literature. For example, SLD, ACC, and PACC have been shown to be *Fisher-consistent*, that is, unbiased, under prior probability shift [262, 758]. In other words, we expect Equation (5.14) instantiated with SLD, PCC, and PACC to provide unbiased estimates when $\mathscr{D}_2$ and $\mathscr{D}_3$ are linked by prior probability shift. We verify this property in Sections 5.4.3 and 5.4.4.

It is worth noting that the weighted estimator (WE) introduced in [140], summarized by Equation (5.2), can be viewed as a special case of this approach, as shown by the following proposition.

Proposition 2. The weighted estimator of Equation (5.2) is a special case of quantification-based estimation of demographic disparity, instantiated with the PCC quantification method. Moreover, the threshold estimator of Equation (5.1) corresponds to CC.

Proof. See Appendix C.3. □

Remark 4. This proposition shows that PCC and WE are equivalent, and that the trivial CC quantifier is equivalent to TE. We treat these methods as prior art and refer to them as CC and PCC for consistency of exposition.

This quantification-based method of addressing demographic disparity is suitable for internal fairness audits, since it allows for unawareness of the sensitive attribute S (i) in the set $\mathscr{D}_1$ used to train the classifier h to be audited, and (ii) in the set $\mathscr{D}_3$ on which this classifier is going to be deployed; it only requires the availability of an auxiliary data set $\mathscr{D}_2$ where the attribute S is labeled. Dataset $\mathscr{D}_2$ may originate from a targeted effort, such as interviews [35], surveys sent to customers asking for voluntary disclosure of sensitive attributes [17], or other optional means of sharing demographic information [67, 68]. Alternatively, it could be derived from data acquisitions carried out for other purposes [284].

Finally, note that in this chapter, we assume the existence of a single binary sensitive attribute S only for ease of exposition. However, our approach can be used straightforwardly in the case in which *multiple* sensitive attributes are present at the same time; in this case, one can simply measure demographic disparity with respect to each sensitive attribute separately (if interested in independent evaluations) or jointly (if emphasizing intersectionality [301]). Our approach can also be extended to deal with *categorical, non-binary* attributes. In this case, one needs (1) to extend the notion of demographic disparity to the case of non-binary attributes. This can be done, e.g., by considering, instead of the simple difference between two acceptance rates $\mu(s)$ as in Equation (5.10), the variance of the acceptance rates across the possible values of S, or the difference between the highest and lowest acceptance rate $\max_{s\in\mathscr{S}}\mu(s) - \min_{s\in\mathscr{S}}\mu(s)$; and (2) to use a single-label multiclass (rather than a binary) quantification system. Concerning this, note that all the methods discussed in Section 5.3.1 except HDy admit straightforward extensions from the binary case to the single-label multiclass case (see [580] for details). HDy is a method for binary quantification only, but it can be adapted to the single-label multiclass scenario by training a binary quantifier for each class in one-vs-all fashion, estimating the prevalence of each class independently of the others, and normalising the obtained prevalence values so that they sum to 1.

5.4 Experiments

5.4.1 General Setup

In this section, we perform an evaluation of different estimators of demographic disparity. We propose five experimental protocols (Sections 5.4.3–5.4.7) summarized in Table 5.2. Each protocol addresses a major challenge that may arise in estimating fairness under unawareness, and does so by varying the size and the mutual distribution shift of the training, auxiliary, and test sets. Protocol names are in the form `action-characteristic-dataset`, as they act on datasets ($\mathscr{D}_1$, $\mathscr{D}_2$ or $\mathscr{D}_3$), modifying their characteristics (size or class prevalence) through

Table 5.2 Summary of experimental protocols.

Protocol name	Variable	See section
`sample-prev-`$\mathscr{D}_3$	joint distribution of $(S,\hat{Y})$ in $\mathscr{D}_3$, via sampling	§ 5.4.3
`sample-prev-`$\mathscr{D}_2$	joint distribution of $(S,\hat{Y})$ in $\mathscr{D}_2$, via sampling	§ 5.4.4
`sample-size-`$\mathscr{D}_2$	size of $\mathscr{D}_2$, via sampling	§ 5.4.5
`sample-prev-`$\mathscr{D}_1$	joint distribution of (S,Y) in $\mathscr{D}_1$, via sampling	§ 5.4.6
`flip-prev-`$\mathscr{D}_1$	joint distribution of (S,Y) in $\mathscr{D}_1$, via label flipping	§ 5.4.7

one of two actions (sampling or flipping of labels). We investigate the performance of six estimators of demographic disparity in each of the five challenges/protocols, keeping the remaining factors constant. For every protocol, we perform an extensive empirical evaluation as follows:

- We compare the performance of each estimation technique on three datasets (Adult, COMPAS, and CreditCard). The datasets and respective preprocessing are described in detail in Section 5.4.2. We focus our discussion (and we present plots – see Figures 5.1–5.7) on the experiments carried out on the Adult dataset, while we summarise numerically the results on COMPAS and CreditCard (Tables 5.4–5.8), discussing them only when significant differences from Adult arise.

- We divide a given data set into three subsets $\mathscr{D}_A, \mathscr{D}_B, \mathscr{D}_C$ of identical sizes and identical joint distribution over (S,Y). We perform five random such splits; in order to test each estimator under the same conditions, these splits are the same for every method. For each split, we permute the role of the stratified subsets $\mathscr{D}_A, \mathscr{D}_B, \mathscr{D}_C$, so that each subset alternatively serves as the training set ($\mathscr{D}_1$), or auxiliary set ($\mathscr{D}_2$), or the test set ($\mathscr{D}_3$). We test all (six) such permutations.

- Whenever an experimental protocol requires sampling from a set, for instance when artificially altering a class prevalence value, we perform 10 different samplings. To perform extensive experiments at a reasonable computational cost, every time an experimental protocol requires changing a dataset $\mathscr{D}$ into a version $\breve{\mathscr{D}}$ characterized by distribution shift, we also reduce its cardinality to $|\breve{\mathscr{D}}| = 500$. Further details and implications of this choice on each experimental protocol are provided in the context of the protocol's setup (e.g., Section 5.4.6).

- Different learning approaches can be used to train the sensitive attribute classifier k_s underlying the quantification methods. We test Logistic Regression (LR) and

Support Vector Machines (SVMs).[2] Sections 5.4.3–5.4.7 report results of quantification algorithms wrapped around a classifier trained via LR. Analogous results obtained with SVMs are reported in Appendix C.4.

- We train the classifier h, whose demographic disparity we aim to estimate, using LR with balanced class weights (i.e., loss weights inversely proportional to class frequencies).

- To measure the performance of different quantifiers, we report the signed estimation error, derived from Equations (5.10) and (5.14) as

$$e = \hat{\delta}_h^S - \delta_h^S = [\hat{\mu}(1) - \hat{\mu}(0)] - [\mu(1) - \mu(0)] \tag{5.16}$$

  We refer to $|e|$ as the Absolute Error (AE), and evaluate the results of our experiments by Mean Absolute Error (MAE) and Mean Squared Error (MSE), defined as

$$\text{MAE}(E) = \frac{1}{|E|} \sum_{e_i \in E} |e_i| \tag{5.17}$$

$$\text{MSE}(E) = \frac{1}{|E|} \sum_{e_i \in E} e_i^2 \tag{5.18}$$

  where the mean of the signed estimation errors e_i is computed over multiple experiments E. Overall, our experiments consist of more than 700,000 separate estimations of demographic disparity.

The remainder of this section is organized as follows. Section 5.4.2 presents the datasets that we have chosen and the preprocessing steps that we apply. Sections 5.4.3–5.4.7 motivate and detail each of the five experimental protocols, reporting the performance of different demographic disparity estimators. Section 5.4.8 shows that reliable fairness auditing may be decoupled from undesirable misuse aimed at inferring the values of the sensitive attribute at an individual level. Finally, Section 5.4.9 describes an ablation study, aimed at investigating the benefits of training and maintaining multiple class-specific quantifiers.

[2]Some among the quantification methods we test in this study require the classifier to output posterior probabilities (as is the case for classifiers trained via LR). If a classifier natively outputs classification scores that are not probabilities (as is the case for classifiers trained via SVM), we convert the former into the latter via the Platt [641] probability calibration method.

5.4.2 Datasets

We perform our experiments on three datasets. We choose Adult and COMPAS, the two most common datasets in the algorithmic fairness community (see Section 3.1), and Credit Card Default (hereafter: CreditCard), which serves as a representative use case for a bank performing a fairness audit of a prediction tool used internally.[3] For each dataset, we standardize the selected features by subtracting the mean and scaling to unit variance. Below, we summarize the key contextual information about these datasets and link to their data briefs; the respective statistics are reported in Table 5.3.

Adult (§ A.1.7).[4] One of the most popular resources in the UCI Machine Learning Repository, the Adult dataset was curated to benchmark the performance of machine learning algorithms. It was extracted from the March 1994 US Current Population Survey and represents respondents along demographic and socioeconomic dimensions, reporting, e.g., their sex, race, educational attainment, and occupation. Each instance comes with a binary label, encoding whether its income exceeds $50,000, which is the target of the associated classification task. We consider "sex" the sensitive attribute S, with a binary categorization of respondents as "Female" or "Male". From the non-sensitive attributes X, we remove "education-num" (a redundant feature), "relationship" (where the values "husband" and "wife" are near-perfect predictors of "sex"), and "fnlwgt" (a variable released by the US Census Bureau to encode how representative each instance is of the overall population). Categorical variables are dummy-encoded and instances with missing values (7%) are removed.

COMPAS (§ A.1.41).[5] This dataset was curated to audit racial biases in the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) risk assessment tool, which estimates the probability that a defendant will become a recidivist [21, 483]. The dataset represents defendants who were scored for recidivism risk by COMPAS in Broward County, Florida between 2013 and 2014, summarizing their demographics, criminal history, custody, and COMPAS scores. We consider the `compas-scores-two-years` subset published by ProPublica on github, consisting of defendants who were observed for two years after screening, for whom a binary recidivism ground truth is available. We follow standard pre-processing to remove noisy instances [646]. We focus on "race" as a protected attribute S, restricting the data to defendants labeled "African-American" or "Caucasian". Our attributes X are the age of the defendant ("age", an integer), the number of juvenile felonies, misdemeanours, and other convictions ("juv_fel_count", "juv_misd_count", "juv_other_count", all

[3] There are two reasons for presenting our experiments on Adult and COMPAS, despite criticizing them in Section 3.1, i.e., (1) the work in this chapter precedes our critique and (2) we are not immune to the incentive structure favoring experimentation with well-known datasets.

[4] https://archive.ics.uci.edu/ml/datasets/adult

[5] https://github.com/propublica/compas-analysis

integers), the number of prior crimes ("priors_count", an integer) and the degree of current charge ("c_charge_degree", felony or misdemeanour, dummy-encoded).

CreditCard (§ A.1.44).[6] This resource was curated to study automated credit card default prediction, following a wave of defaults in Taiwan. The dataset summarizes the payment history of customers of an important Taiwanese bank from April to October 2005. Demographics, marital status, and education of customers are also provided, along with the amount of credit given and a binary variable encoding the default on payment within the next month, which is the associated prediction task. We consider "sex" (binarily encoded) as the sensitive attribute S and keep every other variable in X, preprocessing categorical ones via dummy-encoding ("education", "marriage", "pay_0", "pay_2", "pay_3", "pay_4", "pay_5", "pay_6"). Differently from Adult, we keep marital status as its values are not trivial predictors of the sensitive attribute.

Table 5.3 Dataset statistics after preprocessing.

Dataset	Adult	COMPAS	CreditCard
# data points	45,222	5,278	30,000
# non-sensitive features	84	6	81
sensitive attribute	sex	race	sex
$S=1$	Male	Caucasian	Male
$\Pr(S=1)$	0.675	0.398	0.396
target variable	income	recidivist	default
$Y=\oplus$	>\$50,000	no	no
$\Pr(Y=\oplus)$	0.248	0.498	0.779

5.4.3 Distribution Shift in the Test Set

Motivation and setup

The first experimental protocol models a setting in which the test set $\mathscr{D}_3$ shows a significant distribution shift with respect to the sets $\mathscr{D}_1$ and $\mathscr{D}_2$ available during training of h and k. In other words, in this protocol, $\mathscr{D}_1$ and $\mathscr{D}_2$ are marginalisations of the same joint distribution, while $\mathscr{D}_3$ (more precisely $\breve{\mathscr{D}}_3$) is drawn from a different joint distribution. We consider two sub-protocols (`sample-prev-`$\mathscr{D}_3^{\ominus}$ and `sample-prev-`$\mathscr{D}_3^{\oplus}$) that model changes in the distribution of a sensitive variable S in $\mathscr{D}_3^{\ominus}$ and $\mathscr{D}_3^{\oplus}$, the test subsets of negatively or positively

[6] https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients. Note that we discuss variables with the names they are given in the tabular data (.xls file), which do not match those in the documentation.

predicted instances. More in detail, we let $\Pr(s|\ominus)$ (or its dual $\Pr(s|\oplus)$) in $\breve{\mathscr{D}}_3$ range on eleven evenly spaced values between 0 and 1. For example, under sub-protocol `sample-prev-`$\mathscr{D}_3^{\ominus}$, we vary the distribution of sensitive attribute S in $\breve{\mathscr{D}}_3^{\ominus}$, so that $\Pr(s|\ominus) \in \{0.0, 0.1 \ldots, 0.9, 1.0\}$, while keeping the distribution in $\breve{\mathscr{D}}_3^{\oplus}$ fixed. For both sub-protocols, in each repetition we sample subsets of the test set $\mathscr{D}_3$ such that $|\breve{\mathscr{D}}_3^{\ominus}| = |\breve{\mathscr{D}}_3^{\oplus}| = 500$. Pseudocode 1 describes the protocol when acting on $\mathscr{D}_3^{\ominus}$; the case for $\mathscr{D}_3^{\oplus}$ is analogous and consists of swapping the roles of $\mathscr{D}_3^{\ominus}$ and $\mathscr{D}_3^{\oplus}$ in Lines 18 and 19.

This protocol accounts for the inevitable evolution of phenomena, especially those related to human behaviour. Indeed, it is common in real-world scenarios for data generation processes to be nonstationary and change across development and deployment, due, e.g., to seasonality, changes in the spatiotemporal application context, or any sort of unmodeled novelty and difference in populations [214, 528, 578]. Given that most work on algorithmic fairness focuses on decisions or predictions about people, and given inevitable changes in human lives, values, and behavior, the above considerations about non-stationarity seem particularly relevant. For example, data available from one population is often repurposed to train algorithms that will be deployed on a different population, requiring ad hoc fair learning approaches [173] and evoking the *portability trap* of fair machine learning [707]. In addition, agents can respond to new technology in their social context and adapt their behaviour accordingly [374, 771], causing *ripple effects* [707] and *feedback loops* [531]. Furthermore, as a concrete (although spurious) example of shift in a popular fairness-related dataset, the repeated offence rate for defendants in the COMPAS dataset [483] increases sharply between 2013 and 2014 [46, 76]. Finally, personalized pricing constitutes an increasingly possible practice with non-trivial fairness concerns [423] and inevitable shifts due to changing habits and environments [720].

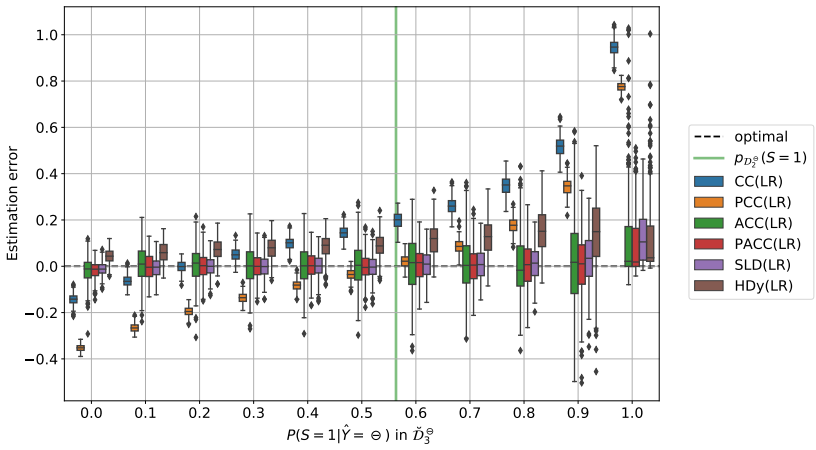
In this protocol, quantifiers are tested on subsets $\breve{\mathscr{D}}_3^{\ominus}$, $\breve{\mathscr{D}}_3^{\oplus}$ that exhibit a different prevalence of sensitive attribute s compared to their counterparts $\mathscr{D}_2^{\ominus}$, $\mathscr{D}_2^{\oplus}$ in the auxiliary set. More specifically, with this protocol we vary the joint distribution of $(S, \hat{Y})$ to directly influence the demographic disparity of the classifier h in the test set $\mathscr{D}_3$, and move it away from the value δ_h^S of the same measure that we would obtain on the set $\mathscr{D}_2$. This is a fundamental evaluation protocol, as it makes our estimand different across $\mathscr{D}_2$ and $\mathscr{D}_3$ (or, more precisely, its modified version $\breve{\mathscr{D}}_3$), which is typically expected in practice. If this was not the case, a practitioner could simply resort to an explicit calculation of the demographic disparity in the auxiliary set $\mathscr{D}_2$ and consider it representative of any deployment condition. Given this reasoning, this protocol imposes large variations in the demographic disparity of h between $\mathscr{D}_2$ and $\mathscr{D}_3$, which act as the training set and the test set, respectively, for our quantifiers. For example, on Adult, δ_h^S is approximately equal to 0.3 in $\mathscr{D}_2$, while in $\mathscr{D}_3$ we let it vary in the

Input : • Dataset $\mathscr{D}$;
• Classifier learner CLS;
• Quantification method Q;
Output : • MAE of the demographic disparity estimates ;
• MSE of the demographic disparity estimates ;

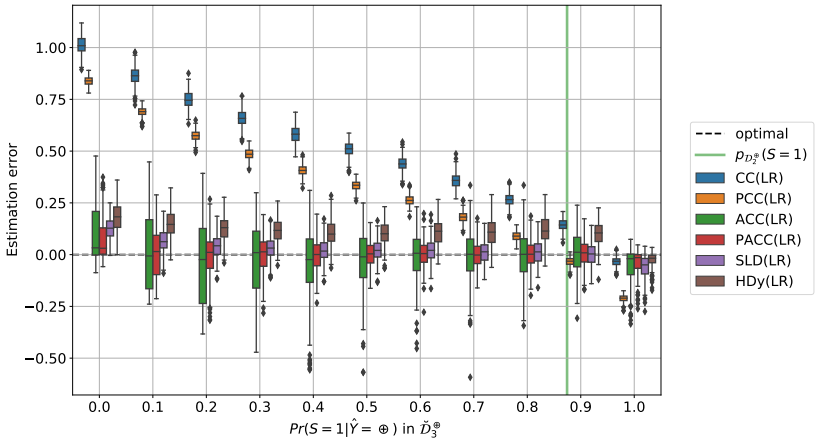
$E \leftarrow \emptyset$;
for *5 random splits* **do**
  $\mathscr{D}_A, \mathscr{D}_B, \mathscr{D}_C \leftarrow$ split_stratify($\mathscr{D}$) ;
  for $\mathscr{D}_1, \mathscr{D}_2, \mathscr{D}_3 \in$ permutations($\mathscr{D}_A, \mathscr{D}_B, \mathscr{D}_C$) **do**
    `/* Learn a classifier` $h : \mathscr{X} \to \mathscr{Y}$ `*/`
    $h \leftarrow$ CLS.fit($\mathscr{D}_1$) ;
    $\mathscr{D}_2^{\ominus} \leftarrow \{(\mathbf{x}_i, s_i) \in \mathscr{D}_2 \mid h(\mathbf{x}_i) = \ominus\}$;
    $\mathscr{D}_2^{\oplus} \leftarrow \{(\mathbf{x}_i, s_i) \in \mathscr{D}_2 \mid h(\mathbf{x}_i) = \oplus\}$;
    `/* Learn quantifiers` $q_y : 2^{\mathscr{X}} \to [0,1]$ `*/`
    $q_{\ominus} \leftarrow$ Q.fit($\mathscr{D}_2^{\ominus}$) ;
    $q_{\oplus} \leftarrow$ Q.fit($\mathscr{D}_2^{\oplus}$) ;
    `/* Split instances in` $\mathscr{D}_3$ `based on predicted labels from` h `*/`
    $\mathscr{D}_3^{\ominus} \leftarrow \{\mathbf{x}_i \in \mathscr{D}_3 \mid h(\mathbf{x}_i) = \ominus\}$;
    $\mathscr{D}_3^{\oplus} \leftarrow \{\mathbf{x}_i \in \mathscr{D}_3 \mid h(\mathbf{x}_i) = \oplus\}$;
    for *10 repeats* **do**
      for $p \in \{0.1, 0.2, \ldots, 0.9\}$ **do**
        `/* Generate samples from` $\mathscr{D}_3^{\ominus}$ `at desired prevalence and size, and uniform samples from` $\mathscr{D}_3^{\oplus}$ `at desired size */`
        $\breve{\mathscr{D}}_3^{\ominus} \sim \mathscr{D}_3^{\ominus}$ with $p_{\breve{\mathscr{D}}_3^{\ominus}}(s) = p$ and $|\breve{\mathscr{D}}_3^{\ominus}| = 500$;
        $\breve{\mathscr{D}}_3^{\oplus} \sim \mathscr{D}_3^{\oplus}$ with $|\breve{\mathscr{D}}_3^{\oplus}| = 500$;
        `/* Use quantifiers to estimate demographic prevalence */`
        $\hat{p}^{q_{\ominus}}_{\breve{\mathscr{D}}_3^{\ominus}}(s) \leftarrow q_{\ominus}(\breve{\mathscr{D}}_3^{\ominus})$;
        $\hat{p}^{q_{\oplus}}_{\breve{\mathscr{D}}_3^{\oplus}}(s) \leftarrow q_{\oplus}(\breve{\mathscr{D}}_3^{\oplus})$;
        `/* Compute the signed error of the demographic disparity estimate */`
        $e \leftarrow$ compute error using $\hat{p}^{q_{\ominus}}_{\breve{\mathscr{D}}_3^{\ominus}}(s)$, $\hat{p}^{q_{\oplus}}_{\breve{\mathscr{D}}_3^{\oplus}}(s)$ and Equation (5.16)
        $E \leftarrow E \cup \{e\}$
      end
    end
  end
end
mae $\leftarrow$ MAE(E) ;
mse $\leftarrow$ MSE(E) ;
return mae, mse

Pseudocode 1: Protocol `sample-prev-`$\mathscr{D}_3$, shown for variations of prevalence values in class $y = \ominus$.

range $[-0.7, 0.9]$. Despite these sizeable variations, we expect that methods such as SLD, ACC, and PACC perform well, due to their proven unbiasedness in this setting (Remark 3).



(a) Protocol `sample-prev-`$\mathscr{D}_3^{\ominus}$



(b) Protocol `sample-prev-`$\mathscr{D}_3^{\oplus}$

Fig. 5.1 Experiments conducted according to protocol `sample-prev-`$\mathscr{D}_3$ on the Adult dataset. The figure shows the distribution of the estimation error (on the y axis) as $\breve{\mathscr{D}}_3$ is sampled with a given $\Pr(S=1|Y=\ominus)$ value (a) or with a given $\Pr(S=1|Y=\oplus)$ value (b), which are shown on the x axis. The green line indicates the value of $\Pr(S=1)$ as observed in $\mathscr{D}_2^{\ominus}$ (a) or in $\mathscr{D}_2^{\oplus}$ (b).

Results

In Figure 5.1 we report the performance of CC, PCC, ACC, PACC, SLD, and HDy on the Adult dataset under the `sample-prev-`$\mathscr{D}_3$ experimental protocol. The estimation error (Equation 5.16) is reported on the y axis, as we vary the prevalence of the protected group in the test set, which is displayed on the x axis. Figure 5.1a concentrates on prevalence variations in $\mathscr{D}_3^{\ominus}$, while Figure 5.1b considers variations of the prevalence of the protected group in $\mathscr{D}_3^{\oplus}$. Each boxplot summarizes the results of 5 random splits, 6 role permutations, and 10 samplings of $\breve{\mathscr{D}}_3$, for a total of 300 repetitions for each combination of 6 methods

Table 5.4 Results obtained in the experiments run according to protocol `sample-prev-`$\mathscr{D}_3$.

		↓ MAE	↓ MSE	↑ $P(\text{AE} < 0.1)$	↑ $P(\text{AE} < 0.2)$
Adult	CC(LR)	0.382 ±0.304	0.239 ±0.305	0.207	0.386
	PCC(LR)	0.299 ±0.237	0.146 ±0.199	0.235	0.427
	ACC(LR)	0.103 ±0.097	0.020 ±0.047	0.595	0.870
	PACC(LR)	0.061 ±0.059	0.007 ±0.016	0.813	0.970
	SLD(LR)	**0.055** ±0.052	**0.006** ±0.012	**0.846**	**0.980**
	HDy(LR)	0.110 ±0.079	0.018 ±0.032	0.500	0.893
COMPAS	CC(LR)	0.541 ±0.369	0.429 ±0.472	0.118	0.237
	PCC(LR)	0.337 ±0.242	0.172 ±0.214	0.181	0.344
	ACC(LR)	0.495 ±0.363	0.377 ±0.471	0.143	0.252
	PACC(LR)	0.252 ±0.213	0.109 ±0.184	0.287	0.492
	SLD(LR)	**0.169** ±0.139	**0.048** ±0.077	**0.385**	**0.669**
	HDy(LR)	0.267 ±0.213	0.116 ±0.176	0.250	0.472
CreditCard	CC(LR)	0.345 ±0.241	0.177 ±0.212	0.172	0.339
	PCC(LR)	0.325 ±0.213	0.151 ±0.157	0.176	0.340
	ACC(LR)	0.341 ±0.259	0.183 ±0.256	0.189	0.367
	PACC(LR)	0.259 ±0.211	0.111 ±0.173	0.269	0.480
	SLD(LR)	**0.190** ±0.148	**0.058** ±0.086	**0.334**	**0.609**
	HDy(LR)	0.251 ±0.190	0.099 ±0.142	0.248	0.478

and 11 values that vary on the x axis. Boxes enclose the two central quartiles (separated by a median horizontal line), while whiskers surround points in the outer quartiles, except for outliers marked with diamonds.

Similar trends emerge under both sub-protocols. CC and PCC show a clear trend along the x axis, vastly over- or underestimating the demographic disparity of h, and prove to be unreliable in settings where the prevalence values in the unlabeled (test) set shift away from the prevalence values of the training set. In sub-protocol `sample-prev-`$\mathscr{D}_3^{\oplus}$, summarised in Figure 5.1b, the prevalence of men ($S = 1$) in $\breve{\mathscr{D}}_3^{\oplus}$, used to test one of the quantifiers, is almost always lower than the prevalence in the respective training set $\mathscr{D}_2^{\oplus}$, reported with a vertical green line. As a result, quantifiers trained on $\mathscr{D}_2^{\oplus}$ tend to systematically overestimate the prevalence of males in $\mathscr{D}_3^{\oplus}$, thus also overestimating $\mu(1)$ and δ_h^S, according to Equations (5.14) and (5.15). Similar considerations hold for sub-protocol `sample-prev-`$\mathscr{D}_3^{\ominus}$, with a sign flip.

ACC, PACC, SLD and HDy, on the other hand, display low bias, even under sizeable prevalence shift. Their variance is higher than those of CC and PCC, but their estimation error is moderate overall. The condition $\Pr(S = 1|\hat{Y} = \ominus) = 1$ (right-most point in Figure

5.1a) is particularly critical for every method due to $p_{\mathscr{D}_3}(s=0)$ dropping below 0.1, thus making small estimation errors for the denominator of Equation 5.15 especially impactful on $\hat{\mu}(0)$.

The results of the COMPAS and CreditCard datasets are reported in Table 5.4, together with a summary of the results of the Adult dataset we have just discussed. The first and second columns indicate the MAE and MSE values (lower is better), while the third and fourth columns indicate the probability that the Absolute Error (AE) falls below 0.1 and 0.2 across the entire experimental protocol (higher is better). **Boldface** indicates the best method for a given dataset and metric. The superscripts † and ‡ denote the methods (if any) whose error scores (MAE, MSE) are *not* statistically significantly different from the best according to a paired sample, two-tailed t-test at different confidence levels. Symbol † indicates $0.001 < p\text{-value} < 0.05$ while symbol ‡ indicates $0.05 \leq p\text{-value}$; the absence of any such symbol indicates $p\text{-value} \leq 0.001$ (i.e., that the performance of the method is statistically significantly different from that of the best method). Overall, SLD strikes the best balance between bias and variance. PACC is the second-best approach, outperforming ACC and PCC, demonstrating the utility of combining posterior probabilities and adjustments when the latter can reliably be estimated. The trends we discussed also hold for COMPAS and CreditCard. Note that both datasets appear to provide a setting harder than Adult for the inference of the sensitive attribute S from the non-sensitive attributes X.

5.4.4 Distribution Shift in the Auxiliary Set

Motivation and setup

This protocol is analogous to protocol `sample-prev-`$\mathscr{D}_3$ (Section 5.4.3), but for the fact that it focuses on shifts in the auxiliary set $\mathscr{D}_2$, while $\mathscr{D}_1$ and $\mathscr{D}_3$ remain at their natural prevalence. Similarly to Section 5.4.3, we assess the signed estimation error under shifts that affect $\mathscr{D}_2^{\ominus}$ or $\mathscr{D}_2^{\oplus}$, that is, the subsets of $\mathscr{D}_2$ labeled positively or negatively by the classifier h. Here too, we consider two experimental sub-protocols, describing variations in the prevalence of sensitive attribute s in either subset. More specifically, we let $\Pr(s|\ominus)$ (or its dual $\Pr(s|\oplus)$) take 9 evenly spaced values between 0.1 and 0.9. We avoid extreme values of 0 and 1, which would make either demographic group $S=0$ or $S=1$ absent from the training set of one quantifier. For example, in sub-protocol `sample-prev-`$\mathscr{D}_2^{\ominus}$ we let the prevalence $\Pr(s|\ominus)$ in $\breve{\mathscr{D}}_2^{\ominus}$ take values in $\{0.1, 0.2\ldots, 0.8, 0.9\}$, while the remaining subset $\breve{\mathscr{D}}_2^{\oplus}$ remains at its natural prevalence $\Pr(s|\oplus)$. For each repetition, we set $|\breve{\mathscr{D}}_2^{\ominus}| = |\breve{\mathscr{D}}_2^{\oplus}| = 500$. This makes for a challenging quantification setting and allows for fast training of multiple quantifiers across

many repetitions. Pseudocode 3 describes the protocol when acting on $\mathscr{D}_2^{\ominus}$; the case for $\mathscr{D}_2^{\oplus}$ is analogous and comes down to swapping the roles of $\mathscr{D}_2^{\ominus}$ and $\mathscr{D}_2^{\oplus}$ in Lines 12 and 13.

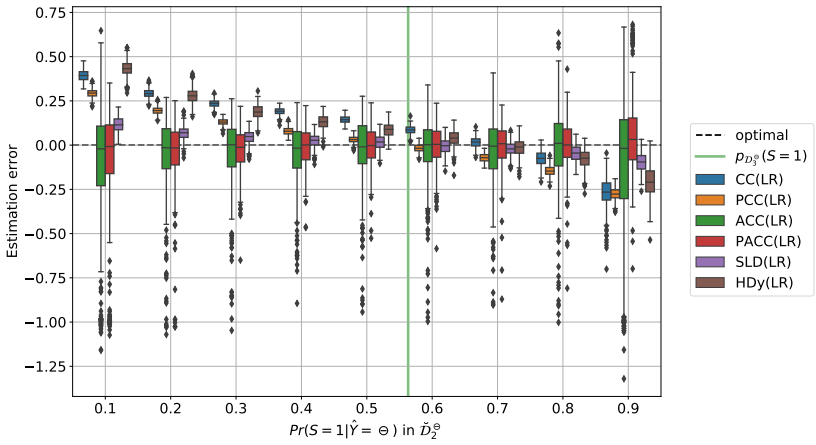
This protocol captures issues of representativity in demographic data, e.g., due to nonuniform response rates across subpopulations [702, 703]. Given the importance of trust for the provision of one's sensitive attributes, in some domains this provision is considered akin to a *data donation* [17]. Individuals from groups that were historically served with worse quality or had lower acceptance rates for a service can be reluctant to disclose their membership in those groups, fearing that it may be used against them as grounds for rejection or discrimination [351]. This may be especially true for people who perceive themselves to be at high risk of rejection, and this can cause complex selection biases, jointly dependent on S and Y, or S and $\hat{Y}$ if individuals have some knowledge of the classification procedure. For example, health care providers may be advised to collect information about the race of patients to monitor the quality of services across subpopulations. In a field study, 28% of patients reported discomfort in revealing their own race to a clerk, with African-American patients significantly less comfortable than white patients on average [35].

Through this protocol, we may expect to find patterns similar to those highlighted in Section 5.4.3, with the roles of the auxiliary set $\mathscr{D}_2$ and the test set $\mathscr{D}_3$ now switched. Under this protocol, $\mathscr{D}_2$ has a lower cardinality and variable prevalence (and is noted by $\breve{\mathscr{D}}_2$ for this reason), while $\mathscr{D}_3$ is left to its original cardinality and prevalence of the sensitive attribute s.

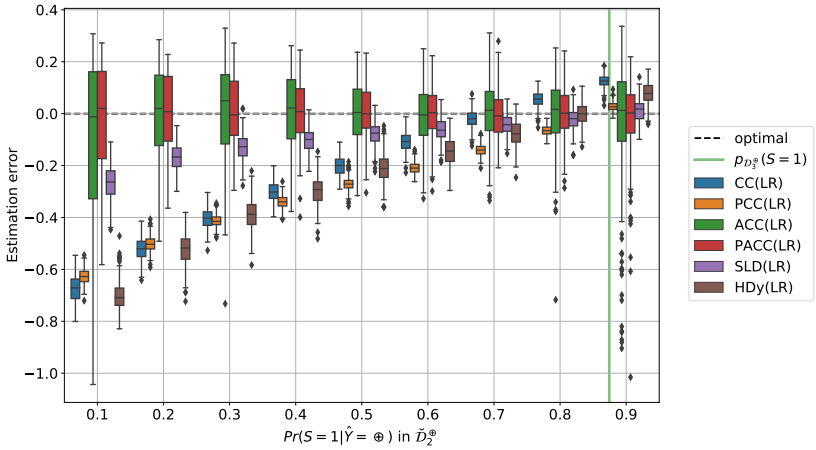
Results

Figure 5.2 shows the signed estimation error on the y axis, as we vary, on the x axis, the prevalence of the sensitive attribute in $\mathscr{D}_2^{\ominus}$ (Figure 5.2a) and $\mathscr{D}_2^{\oplus}$ (Figure 5.2b). CC, PCC, and HDy are fairly sensitive to shifts in their training set. In sub-protocol `sample-prev-`$\mathscr{D}_2^{\oplus}$, symmetrically to the sub-protocol `sample-prev-`$\mathscr{D}_3^{\oplus}$ discussed in the previous section, the prevalence of males ($S = 1$) in subset $\mathscr{D}_2^{\oplus}$, used to train one of the quantifiers, is almost always lower than the prevalence in the respective test subset $\mathscr{D}_3^{\oplus}$, indicated with a vertical green line. As a result, quantifiers trained on $\mathscr{D}_2^{\oplus}$ tend to systematically underestimate the prevalence of males in $\mathscr{D}_3^{\oplus}$ and underestimate the (signed) demographic disparity of the classifier h.

ACC and PACC require splitting their training set to estimate the respective adjustments (Equations 5.6–5.9), and suffer from a reduced cardinality $|\breve{\mathscr{D}}_2| = 1{,}000$. Their performance worsens substantially with respect to protocol `sample-prev-`$\mathscr{D}_3$, where $|\mathscr{D}_2| > 15{,}000$. Indeed, these methods have been shown to be *Fisher-consistent* under prior probability shift [262, 758], that is, they are guaranteed to be accurate, thanks to the respective adjustments, if $\mathscr{D}_2$ is large enough and linked to $\mathscr{D}_3$ by prior probability shift. While the latter condition



(a) Protocol `sample-prev-`$\mathscr{D}_2^{\ominus}$



(b) Protocol `sample-prev-`$\mathscr{D}_2^{\oplus}$

Fig. 5.2 Protocol `sample-prev-`$\mathscr{D}_2$ on the Adult dataset. Distribution of the estimation error (y axis) as $\check{\mathscr{D}}_2$ is sampled with a given $\Pr(S=1|Y=\ominus)$ value, plot (a), or $\Pr(S=1|Y=\oplus)$ value, plot (b) (x axis). The green line indicates the value of $\Pr(S=1)$ as observed in $\mathscr{D}_3^{\ominus}$, plot (a), or $\mathscr{D}_3^{\oplus}$, plot (b).

holds, the former is violated under this protocol, hence ACC and PACC are unbiased (in expectation), but display a large variance, due to unstable adjustments. SLD, on the other hand, shows moderate variance and bias. These effects are especially evident at the extremes of the x axis, which correspond to settings where few instances with $S=0$ or $S=1$ are available for quantifier training. In turn, the few positives (negatives) make it particularly difficult to reliably estimate tpr_{k_s} (tnr_{k_s}), as required by Equations 5.7 and 5.9. For example, in Figure 5.2a we see that the error of ACC ranges between -1.3 and 0.7. Given that the true demographic disparity of the classifier h is $\delta_h^S = 0.3$, these are the worst possible errors, corresponding to extreme estimates $\hat{\delta}_h^S = -1$ and $\hat{\delta}_h^S = 1$, respectively. Finally, it should be

Table 5.5 Results obtained in the experiments run according to protocol `sample-prev-`$\mathscr{D}_2$.

		↓ MAE	↓ MSE	↑ $P(\text{AE} < 0.1)$	↑ $P(\text{AE} < 0.2)$
Adult	CC(LR)	0.230 ±0.177	0.084 ±0.118	0.274	0.523
	PCC(LR)	0.213 ±0.169	0.074 ±0.103	0.323	0.551
	ACC(LR)	0.159 ±0.178	0.057 ±0.159	0.439	0.789
	PACC(LR)	0.112 ±0.118	0.026 ±0.093	0.559	0.889
	SLD(LR)	**0.081** ±0.070	**0.011** ±0.020	**0.705**	**0.929**
	HDy(LR)	0.219 ±0.188	0.084 ±0.128	0.345	0.573
COMPAS	CC(LR)	0.498 ±0.253	0.312 ±0.260	0.044	0.128
	PCC(LR)	0.264 ±0.186	0.104 ±0.126	0.227	0.431
	ACC(LR)	0.469 ±0.276	0.296 ±0.303	0.080	0.184
	PACC(LR)	0.338 ±0.254	0.179 ±0.250	0.185	0.356
	SLD(LR)	**0.160** ±0.123	**0.041** ±0.060	**0.386**	**0.678**
	HDy(LR)	0.255 ±0.189	0.101 ±0.135	0.246	0.463
CreditCard	CC(LR)	0.429 ±0.252	0.248 ±0.236	0.103	0.225
	PCC(LR)	0.204 ±0.140	0.061 ±0.073	0.287	0.551
	ACC(LR)	0.535 ±0.316	0.387 ±0.353	0.085	0.165
	PACC(LR)	0.512 ±0.311	0.359 ±0.343	0.094	0.171
	SLD(LR)	**0.171** ±0.123	**0.044** ±0.058	**0.348**	**0.645**
	HDy(LR)	0.222 ±0.159	0.074 ±0.101	0.260	0.508

noted that PACC outperforms ACC, thanks to efficient use of posteriors $\pi_s(\mathbf{x}_i)$ in place of binary decisions $k_s(\mathbf{x}_i)$.

These trends also hold for COMPAS and CreditCard, as summarized in Table 5.5. Similarly to Table 5.4, we find that, under large shifts between the auxiliary and the test set, the estimation of demographic disparity is more difficult in COMPAS and CreditCard than in Adult. Overall, these experiments show that CC and PCC fare poorly under prior probability shift, and are outperformed by estimators with better theoretical guarantees.

5.4.5 Reduced Cardinality of the Auxiliary Set

Motivation and setup

In this experimental protocol, we focus on the size of the auxiliary set $\mathscr{D}_2$, studying its influence on the estimation problem. Our goal is to understand how small this set can be before degrading the performance of our estimation techniques. We use subsets $\breve{\mathscr{D}}_2$ of the auxiliary set, obtained by sampling from it instances uniformly without replacement. We let their cardinality $|\breve{\mathscr{D}}_2|$ take five values evenly spaced on a logarithmic scale, between a

minimum size $|\breve{\mathscr{D}}_2|$=1,000 and a maximum size $|\breve{\mathscr{D}}_2| = |\mathscr{D}_2|$. In other words, we let the cardinality of the auxiliary set take five different values between 1,000 and $|\mathscr{D}_2|$ in a geometric progression. As described in Section 5.4.1, for each cardinality of the auxiliary set we wish to test, we perform ten samplings over five splits and six permutations, for a total of 300 repetitions per approach per dataset. Pseudocode 4 describes Protocol `sample-size-`$\mathscr{D}_2$.

This protocol is justified by the well-documented difficulties in the acquisition of demographic data for industry professionals, which vary depending on the domain, the company and other factors of disparate nature [17, 68, 82, 284, 370]. As an example, Galdon Clavell et al. [284] perform an internal fairness audit of a personalized wellness recommendation app, for which sensitive features are not collected during production, following the principles of data minimization. However, sensitive features were available from an initial development phase in an auxiliary set, whose size was determined by prior considerations. Furthermore, in the US, the collection of sensitive attributes is highly industry dependent, ranging from mandatory to forbidden, depending on the fragmented regulation applicable in each domain [82]. High-quality auxiliary sets can be obtained through optional surveys [828], for which response rates are highly dependent on trust, and can be improved by making the intended use of the data clearer [17], directly impacting the cardinality of $\mathscr{D}_2$.

Therefore, the cardinality of the auxiliary set $\mathscr{D}_2$ is an interesting variable in the context of fairness audits. The estimation methods that we consider have peculiar data requirements, such as the need to estimate true/false positive rates. For this reason, interesting patterns should emerge from this protocol. We expect the key trends for the estimation error to vary monotonically with $|\breve{\mathscr{D}}_2|$, which is why we let it vary according to a geometric progression.

Results

The signed estimation error on the Adult dataset under this experimental protocol is illustrated in Figure 5.3, as we vary the cardinality $|\breve{\mathscr{D}}_2|$ along the x axis. Clearly, the variance for each approach decreases as the size of $\breve{\mathscr{D}}_2$ increases. Additionally, slight biases may improve, as is the case with HDy, whose median error approaches zero as $|\breve{\mathscr{D}}_2|$ increases. These trends are a direct confirmation of hints already obtained from the protocols discussed above. The most striking trend is the unreliability of ACC and PACC (and especially the former) in the small-data regime.

Similar results are obtained for COMPAS and CreditCard, as reported in Table 5.6. Across the three datasets, PACC and ACC perform quite poorly due to the difficulty in estimating tpr_{k_s} and fpr_{k_s} with the few labeled data points available from $\breve{\mathscr{D}}_2$. On the other hand, both SLD and HDy are fairly reliable. PCC stands out as a strong performer, with low bias and low variance. This is due to the fact that, under this experimental protocol, there

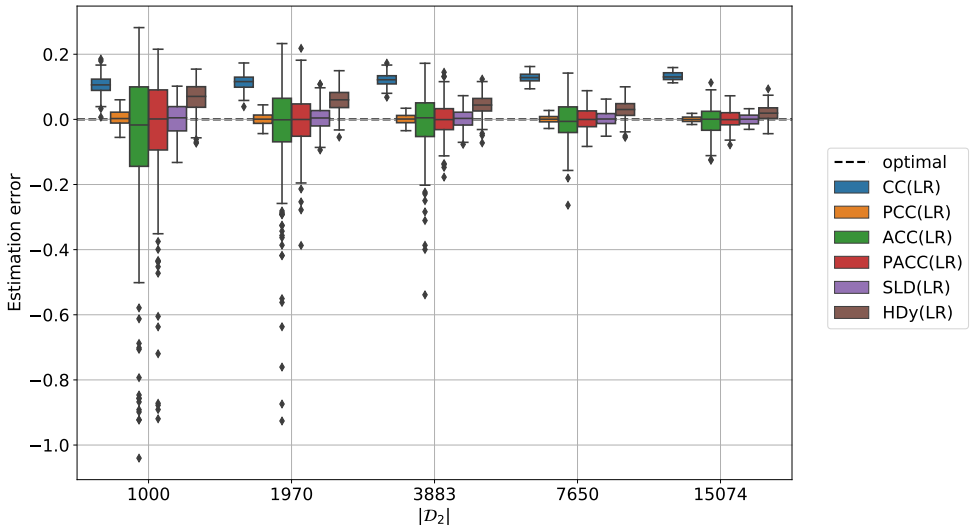


Fig. 5.3 Protocol `sample-size-`$\mathscr{D}_2$ on the Adult dataset. Distribution of the estimation error (y axis) as the cardinality $|\breve{\mathscr{D}}_2|$ is varied (x axis).

is no shift between the auxiliary set $\mathscr{D}_2$, on which the quantifiers are trained, and the test set $\mathscr{D}_3$, on which they are tested. Since the current protocol focuses on the cardinality of the auxiliary set, $\mathscr{D}_2$ and $\mathscr{D}_3$ remain stratified subsets of the Adult dataset, with identical distributions over (S,Y). In turn, this favours PCC, which relies on the fact that the posterior probabilities of its underlying classifier k are well-calibrated on $\mathscr{D}_3$.[7]

5.4.6 Distribution Shift in the Training Set via Sampling

Motivation and setup

With this protocol we evaluate the impact of shifts in the training set $\mathscr{D}_1$, by drawing different subsets $\breve{\mathscr{D}}_1$ as we vary $\Pr(Y=S)$.[8] More specifically, we vary $\Pr(Y=S)$ between 0 and 1 with a step of 0.1. In other words, we sample at random from $\mathscr{D}_1$ a proportion p of instances $(\mathbf{x}_i,s_i,y_i)$ such that $Y=S$ and a proportion $(1-p)$ such that $Y\neq S$, with $p\in\{0.0,0.1,\ldots,0.9,1.0\}$. We choose a limited cardinality $|\breve{\mathscr{D}}_1|=500$, which allows us to

[7] Posterior probabilities $\Pr(s|\mathbf{x})$ are said to be *well-calibrated* when, given a sample σ drawn from $\mathscr{X}$

$$\lim_{|\sigma|\to\infty}\frac{|\{\mathbf{x}\in s|\Pr(s|\mathbf{x})=\alpha\}|}{|\{\mathbf{x}\in\sigma|\Pr(s|\mathbf{x})=\alpha\}|}=\alpha.$$

i.e., when for big enough samples, α approximates the true proportion of data points belonging to class s among all data points for which $\Pr(s|\mathbf{x})=\alpha$.

[8] While Y and S take values from different domains, by $Y=S$ we mean $(Y=\oplus\wedge S=1)\vee(Y=\ominus\wedge S=0)$, i.e. a situation where positive outcomes are associated with group $S=1$ and negative outcomes with group $S=0$.

Table 5.6 Results obtained in the experiments run according to protocol `sample-size-`$\mathscr{D}_2$

		↓ MAE	↓ MSE	↑ $P(\text{AE} < 0.1)$	↑ $P(\text{AE} < 0.2)$
Adult	CC(LR)	0.120 ±0.022	0.015 ±0.005	0.159	**1.000**
	PCC(LR)	**0.012** ±0.010	**0.000** ±0.000	**1.000**	**1.000**
	ACC(LR)	0.083 ±0.113	0.020 ±0.082	0.747	0.928
	PACC(LR)	0.055 ±0.079	0.009 ±0.048	0.856	0.969
	SLD(LR)	0.025 ±0.020	0.001 ±0.002	0.996	**1.000**
	HDy(LR)	0.047 ±0.033	0.003 ±0.004	0.922	**1.000**
COMPAS	CC(LR)	0.353 ±0.047	0.127 ±0.032	0.000	0.005
	PCC(LR)	**0.030** ±0.020	**0.001** ±0.001	**0.999**	**1.000**
	ACC(LR)	0.381 ±0.213	0.190 ±0.214	0.097	0.186
	PACC(LR)	0.265 ±0.212	0.115 ±0.183	0.247	0.467
	SLD(LR)	0.135 ±0.098	0.028 ±0.038	0.441	0.765
	HDy(LR)	0.108 ±0.082	0.018 ±0.027	0.549	0.858
CreditCard	CC(LR)	0.177 ±0.078	0.037 ±0.030	0.177	0.629
	PCC(LR)	**0.016** ±0.013	**0.000** ±0.001	**1.000**	**1.000**
	ACC(LR)	0.337 ±0.266	0.184 ±0.259	0.203	0.368
	PACC(LR)	0.299 ±0.255	0.154 ±0.240	0.261	0.445
	SLD(LR)	0.053 ±0.043	0.005 ±0.008	0.871	0.985
	HDy(LR)	0.057 ±0.046	0.005 ±0.009	0.831	0.991

perform multiple repetitions at reasonable computational costs, as described in Section 5.4.1. Although this may affect the quality of the classifier h, this aspect is not the central focus of the present work.

This experimental protocol aligns with biased data collection procedures, sometimes referred to as *censored data* [419]. Indeed, it is common for the ground truth variable to represent a mere proxy for the actual quantity of interest, with nontrivial sampling effects between the two. For example, the validity of arrest data as a proxy for offence has been brought into question [276]. In fact, in this domain, different sources of sampling bias can be in action, such as uneven allocation of police resources between jurisdictions and neighborhoods [368] and lower levels of cooperation in populations who feel oppressed by law enforcement [836].

By varying $\Pr(Y = S)$ we impose a spurious correlation between Y and S, which may be picked up by the classifier h. In extreme situations, such as when $\Pr(Y = S) \simeq 1$, a classifier h can confound the concepts behind S and Y. In turn, we expect this to unevenly affect the acceptance rates for the two demographic groups, effectively changing the demographic

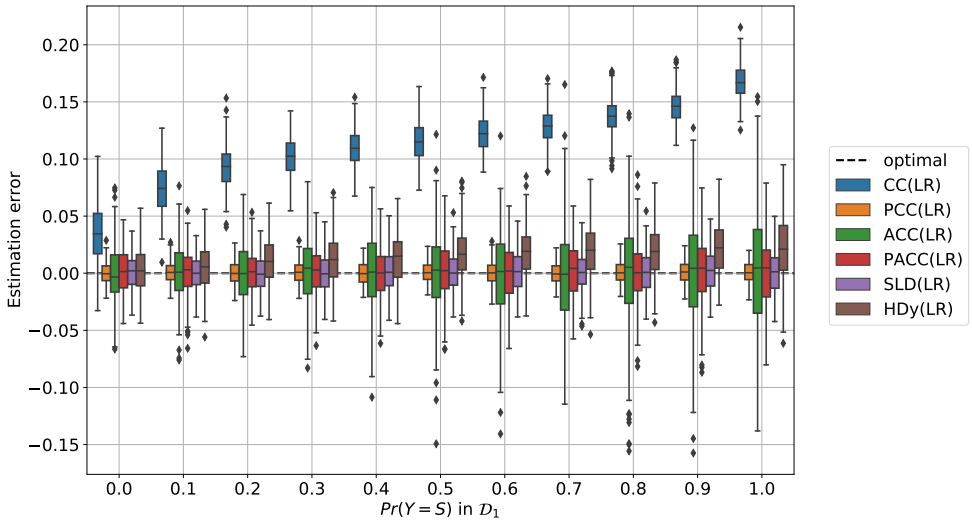


Fig. 5.4 Protocol `sample-prev-`$\mathscr{D}_1$ on the Adult dataset. Distribution of the estimation error (y axis) as $\breve{\mathscr{D}}_1$ is sampled with a given $\Pr(Y = S)$ (x axis). Each boxplot summarizes the results of 5 random splits, 6 role permutations and 10 samplings of $\breve{\mathscr{D}}_1$.

disparity of h, i.e., our estimand δ_h^S. Pseudocode 5 describes the main steps to implement Protocol `sample-prev-`$\mathscr{D}_1$.

Results

In Figure 5.4, the y axis depicts the estimation error (Equation 5.16), as we vary $\Pr(Y = S)$ along the x axis. Each quantification approach outperforms vanilla CC, which overestimates the demographic disparity of the classifier h, i.e., its estimate is larger than the ground truth value, so $\hat{\delta}_h^{S,\mathrm{CC}} > \delta_h^S$. ACC, PCC, PACC, SLD and HDy display a negligible bias and a reliable estimate of demographic disparity. The absolute error for these techniques is always below 0.1, except for a few outliers.

Results for the COMPAS and CreditCard datasets are reported in Table 5.7. Confirming the results of previous protocols, these datasets provide a harder setting for the estimate of demographic disparity, as shown by higher MAE and MSE, which, for instance, increase by one order of magnitude for SLD and PACC moving from Adult to COMPAS. PCC is the best performer, for the same reasons discussed in Section 5.4.3, i.e., the absence of shift between $\mathscr{D}_2$ and $\mathscr{D}_3$.

Table 5.7 Results obtained in the experiments run according to protocol `sample-prev-`$\mathscr{D}_1$.

		↓ MAE	↓ MSE	↑ $P(\text{AE} < 0.1)$	↑ $P(\text{AE} < 0.2)$
Adult	CC(LR)	0.112 ±0.038	0.014 ±0.008	0.321	0.998
	PCC(LR)	**0.008** ±0.005	**0.000** ±0.000	**1.000**	**1.000**
	ACC(LR)	0.029 ±0.024	0.001 ±0.003	0.983	**1.000**
	PACC(LR)	0.019 ±0.014	0.001 ±0.001	**1.000**	**1.000**
	SLD(LR)	0.013 ±0.010	0.000 ±0.000	**1.000**	**1.000**
	HDy(LR)	0.022 ±0.016	0.001 ±0.001	**1.000**	**1.000**
COMPAS	CC(LR)	0.328 ±0.091	0.116 ±0.056	0.022	0.081
	PCC(LR)	**0.026** ±0.019	**0.001** ±0.001	**1.000**	**1.000**
	ACC(LR)	0.349 ±0.211	0.166 ±0.192	0.130	0.252
	PACC(LR)	0.194 ±0.164	0.065 ±0.115	0.345	0.607
	SLD(LR)	0.114 ±0.083	0.020 ±0.027	0.512	0.849
	HDy(LR)	0.096 ±0.076	0.015 ±0.023	0.605	0.897
CreditCard	CC(LR)	0.152 ±0.095	0.032 ±0.036	0.338	0.711
	PCC(LR)	**0.010** ±0.007	**0.000** ±0.000	**1.000**	**1.000**
	ACC(LR)	0.187 ±0.152	0.058 ±0.094	0.347	0.626
	PACC(LR)	0.130 ±0.106	0.028 ±0.046	0.487	0.777
	SLD(LR)	0.047 ±0.037	0.004 ±0.005	0.902	0.998
	HDy(LR)	0.061 ±0.047	0.006 ±0.009	0.814	0.989

5.4.7 Distribution Shift in the Training Set via Label Flipping

Motivation and setup

Certain biases in the training set resulting from domain-specific practices, such as the use of arrest as a substitute for the offense, can be modeled as either a selection bias [276] or a label bias that distorts the ground truth variable Y [275]. With this experimental protocol, we impose the latter bias by actively flipping some ground truth labels Y in $\mathscr{D}_1$ based on their sensitive attribute. Similarly to `sample-prev-`$\mathscr{D}_1$, this protocol achieves a given association between the target Y and the sensitive variable S in the training set $\mathscr{D}_1$. However, instead of sampling, it does so by flipping the Y label of some data points. More specifically, we impose $\Pr(Y = \ominus | S = 0) = \Pr(Y = \oplus | S = 1) = p$ and let p take values across 11 evenly spaced values between 0 and 1. For every value of p, we first sample a random subset $\breve{\mathscr{D}}_1$ of the training set with cardinality 500. Next, we actively flip some Y labels in both demographic groups, until both $\Pr(Y = \ominus | S = 0)$ and $\Pr(Y = \oplus | S = 1)$ reach the desired value of $p \in \{0.0, 0.1, \ldots, 0.9, 1.0\}$. Finally, we train a classifier h on the attributes X and the modified ground truth Y of $\breve{\mathscr{D}}_1$.

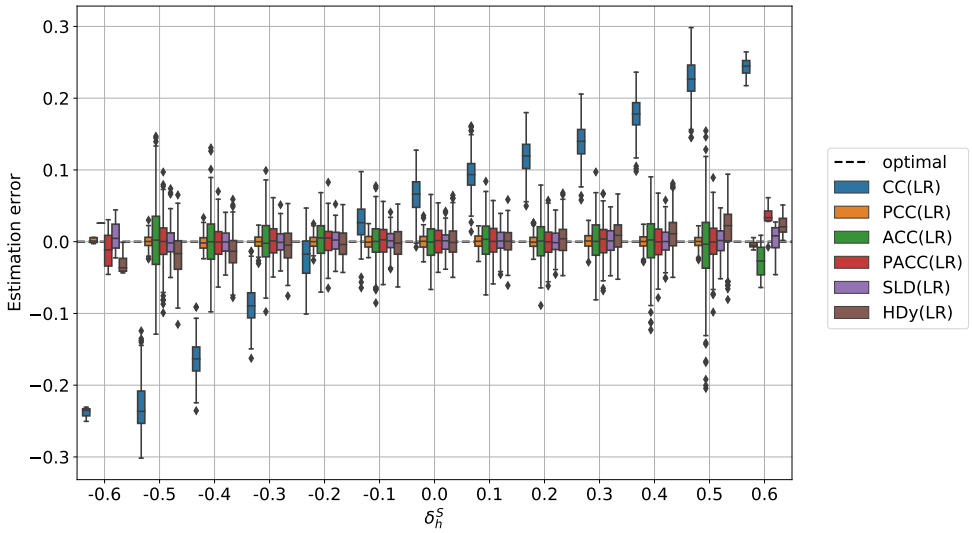


Fig. 5.5 Protocol `flip-prev-`$\mathscr{D}_1$ on the Adult dataset. Distribution of the estimation error (y axis) as δ_h^S varies (x axis).

This experimental protocol is compatible with settings where training data capture a distorted ground truth due to systematic biases and group-dependent annotation accuracy [811]. As an example, the quality of medical diagnoses can depend on race, sex, and socioeconomic status [303]. Furthermore, health care expenditures have been used as a proxy to train an algorithm deployed nationwide in the US to estimate patients' health care needs, resulting in a systematic underestimation of the needs of African-American patients [607]. In the hiring domain, employer response rates to resumes have been found to vary depending on the perceived ethnic origin of the applicant's name [66]. These are all examples where the "ground truth" associated with a dataset is distorted to the disadvantage of a sensitive demographic group.

Similarly to Section 5.4.6, we expect that this experimental protocol will cause significant variations in the demographic disparity of the classifier h due to the strong correlation we impose between S and Y by label flipping. The pseudocode that describes this protocol is essentially the same as in Pseudocode 5, simply replacing the sampling in line 8 with the label flipping procedure described above; therefore, we omit it.

Results

Figure 5.5 illustrates the key trends caused by this experimental protocol on the Adult dataset. A clear trend is visible along the x axis, reporting the true demographic disparity δ_h^S for the classifier h (Equation 5.10), quantized with a step of 0.1. We choose to depict the true demographic disparity on the x axis as it is the estimand, hence a quantity of

Table 5.8 Results obtained in the experiments run according to protocol `flip-prev-`$\mathscr{D}_1$.

		↓ MAE	↓ MSE	↑ $P(\text{AE} < 0.1)$	↑ $P(\text{AE} < 0.2)$
Adult	CC(LR)	0.151 ±0.072	0.028 ±0.021	0.274	0.706
	PCC(LR)	**0.008** ±0.006	**0.000** ±0.000	**1.000**	**1.000**
	ACC(LR)	0.030 ±0.025	0.002 ±0.003	0.982	0.999
	PACC(LR)	0.020 ±0.015	0.001 ±0.001	**1.000**	**1.000**
	SLD(LR)	0.014 ±0.011	0.000 ±0.000	**1.000**	**1.000**
	HDy(LR)	0.022 ±0.017	0.001 ±0.001	1.000	**1.000**
COMPAS	CC(LR)	0.388 ±0.116	0.164 ±0.083	0.027	0.068
	PCC(LR)	**0.027** ±0.020	**0.001** ±0.001	**0.998**	**1.000**
	ACC(LR)	0.392 ±0.211	0.198 ±0.199	0.105	0.194
	PACC(LR)	0.195 ±0.160	0.063 ±0.106	0.337	0.611
	SLD(LR)	0.115 ±0.084	0.020 ±0.027	0.513	0.836
	HDy(LR)	0.094 ±0.075	0.015 ±0.023	0.612	0.906
CreditCard	CC(LR)	0.159 ±0.101	0.036 ±0.037	0.345	0.640
	PCC(LR)	**0.011** ±0.009	**0.000** ±0.000	**1.000**	**1.000**
	ACC(LR)	0.223 ±0.185	0.084 ±0.130	0.307	0.565
	PACC(LR)	0.147 ±0.117	0.035 ±0.056	0.420	0.725
	SLD(LR)	0.056 ±0.043	0.005 ±0.007	0.843	0.995
	HDy(LR)	0.071 ±0.055	0.008 ±0.012	0.732	0.973

interest by definition. The error incurred by CC shows a linear trend that goes from severe underestimation (for low values of the x axis) to severe overestimation (for large values of the x axis). In other words, the (signed) estimation error increases with the true demographic disparity of the classifier h, a phenomenon also noticed by Chen et al. [140]. All remaining approaches compensate for this weakness and display a good estimation error: PCC, ACC, PACC, SLD, and HDy have low variance and a median estimation close to zero across different values of the estimand. Table 5.8 summarizes similar results on COMPASS and CreditCard; PCC remains well-calibrated and very effective, while SLD and HDy also have good performance.

5.4.8 Quantifying without Classifying

Motivation and setup

The motivating use case for this chapter is dataset augmentation for internal audits of group fairness, to characterize a model and its potential to harm sensitive categories of users. Following Awasthi et al. [27], we envision this as an important first step in empowering

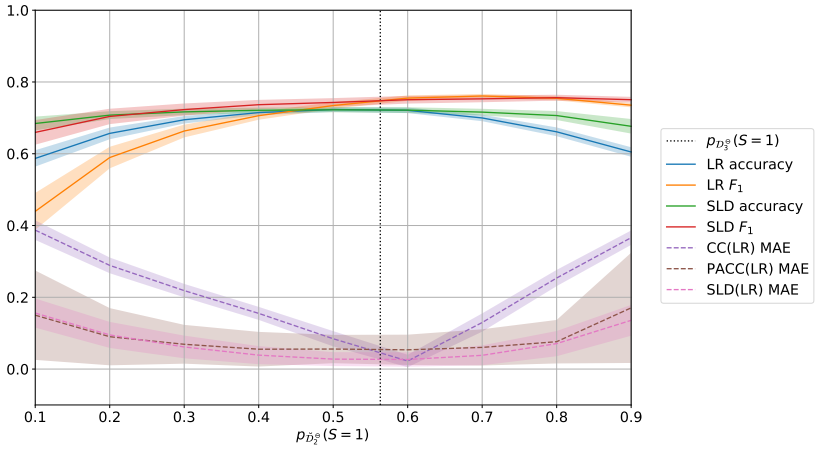
practitioners to argue for resources and, more broadly, to advocate for a deeper understanding and careful evaluation of models. Unfortunately, developing a tool to infer demographic information, even if motivated by careful intentions and good faith, leaves open the possibility for misuse, especially at an individual level. Once a predictive tool, also capable of instance-level classification, is available, it will be tempting for some actors to exploit it precisely for this purpose.

For example, the *Bayesian Improved Surname Geocoding* (BISG) method was designed to estimate population-level disparities in health care [235], but later used to identify individuals potentially eligible for settlements related to discriminatory practices of auto lenders [16, 464]. Automatic inference of sensitive attributes of individuals is problematic for several reasons. Such procedure exploits the co-occurrence of membership in a group and display of a given trait, running the risk of learning, encoding, and reinforcing stereotypical associations. Although also true for group-level estimates, this practice is particularly troublesome at the individual level, where it is likely to cause harm to people who do not fit the norm, resulting, for example, in misgendering and the associated negative effects [547]. Even when "accurate", the mere act of externally assigning sensitive labels can be problematic. For example, gender assignment can be forceful and cause psychological harm for individuals [438].

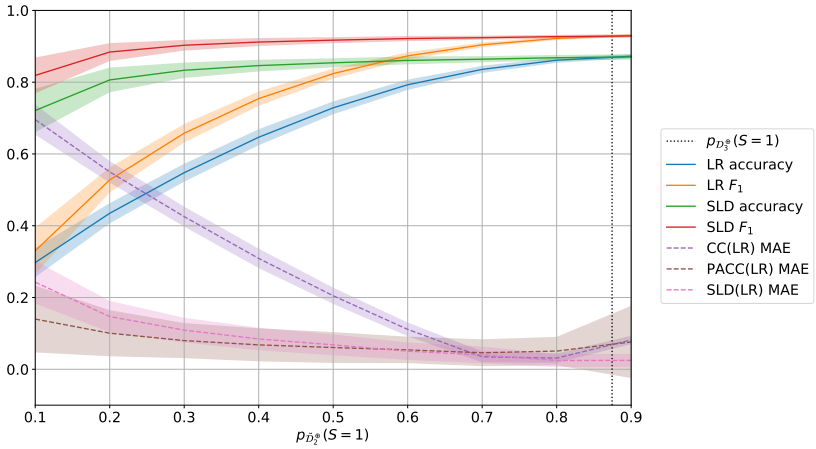
In this section, our aim is to demonstrate that it is possible to decouple the objective of (group-level) quantification of sensitive attributes from that of (individual-level) classification. For each protocol in the previous sections, we compute the accuracy and F_1 score (defined below) of the sensitive attribute classifier k underlying the quantifiers tested, comparing it with their estimation error for class prevalence of the same sensitive attribute (Equation 5.16). Accuracy is the proportion of correctly classified instances over the total (Equation 5.19) while F_1 is the harmonic mean of precision and recall (Equation 5.20). Both measures can be computed from the counters of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN), as follows:

$$\text{accuracy} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}} \tag{5.19}$$

$$F_1 = \begin{cases} \dfrac{2\text{TP}}{2\text{TP}+\text{FP}+\text{FN}} & \text{if } \text{TP}+\text{FP}+\text{FN} > 0 \\ 1 & \text{if } \text{TP}=\text{FP}=\text{FN}=0 \end{cases} \tag{5.20}$$



(a) Protocol `sample-prev-`$\mathscr{D}_2^{\ominus}$

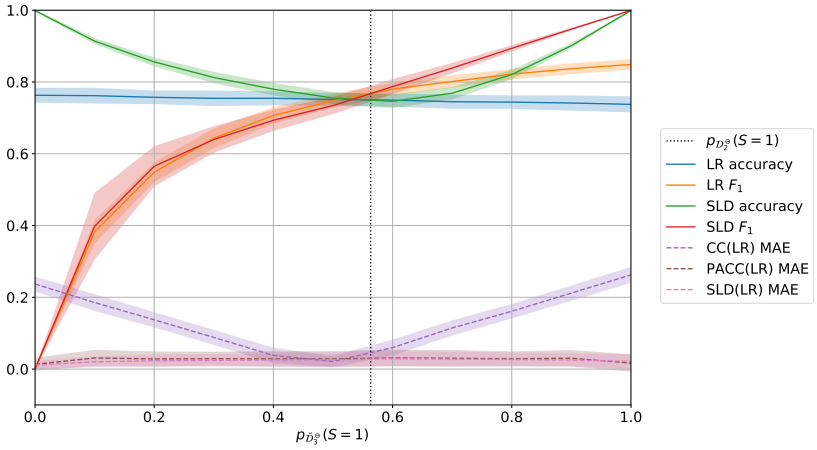


(b) Protocol `sample-prev-`$\mathscr{D}_2^{\oplus}$

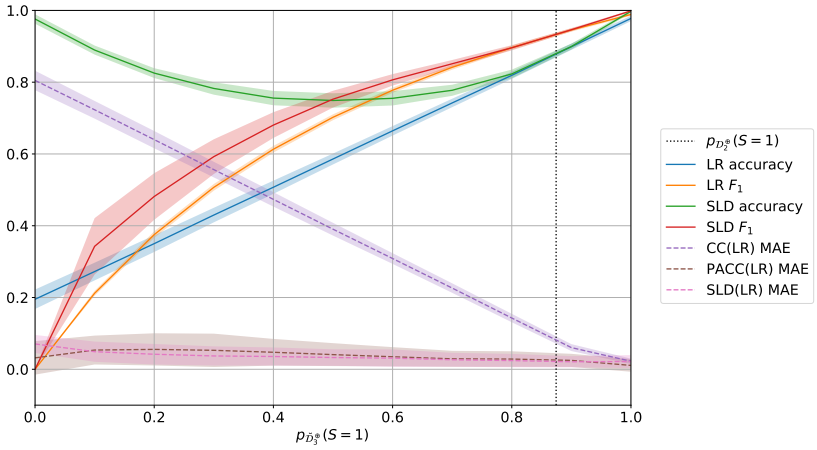
Fig. 5.6 Performance of CC, SLD and PACC on the Adult dataset when used for quantification (MAE – lower is better, dashed) and classification (F_1, accuracy – higher is better, solid) under protocol `sample-prev-`$\mathscr{D}_2$. The classification performance of PACC is equivalent to that of CC (both equal to the performance of the underlying LR), and we thus omit it for readability.

Results

Figures 5.6 and 5.7 displays the quantification performance (MAE – dashed) and classification performance (F_1, accuracy – solid) of CC, SLD and PACC on the Adult dataset under protocols `sample-prev-`$\mathscr{D}_2$ and `sample-prev-`$\mathscr{D}_3$, respectively. As usual, we describe the results for LR-based learners and report their SVM-based duals in the appendix (Figures C.6 and C.7). To evaluate the quantification performance of each approach, we simply report their MAE in estimating the prevalence $p_{\mathscr{D}_3^{\ominus}}(S=1)$, $p_{\mathscr{D}_3^{\oplus}}(S=1)$ in either test subset, depending on the protocol at hand. To assess the performance of the sensitive attribute classifier k



(a) Protocol `sample-prev-`$\mathscr{D}_3^{\ominus}$



(b) Protocol `sample-prev-`$\mathscr{D}_3^{\oplus}$

Fig. 5.7 Performance of CC, SLD and PACC on the Adult dataset when used for quantification (MAE – lower is better, dashed) and classification (F_1, accuracy – higher is better, solid) under protocol `sample-prev-`$\mathscr{D}_3$. The classification performance of PACC is equivalent to that of CC (both equal to the performance of the underlying LR), and we thus omit it for readability.

underlying each quantifier, we proceed as follows. For CC and PACC, we simply run k (LR) on either $\mathscr{D}_3^{\ominus}$ or $\mathscr{D}_3^{\oplus}$, reporting its accuracy and F_1 score in inferring the sensitive attribute of individual instances. The classification performance scores of the classifiers underlying CC and PACC are equivalent, so we omit the latter from Figures 5.6 and 5.7 for readability. For SLD, we take the novel posteriors obtained by applying the EM algorithm to either test subset, and use them for classification with a threshold of 0.5.

Clearly, SLD improves both the quantification and classification performance of the classifier k. In terms of quantification, its MAE is consistently below that of CC, and in

terms of classification, it displays better F_1 and accuracy. However, under large prevalence shifts across the auxiliary set $\mathscr{D}_2$ and the test set $\mathscr{D}_3$, its classification performance becomes unreliable. In particular, under protocol `sample-prev-`$\mathscr{D}_3^{\ominus}$ (resp. `sample-prev-`$\mathscr{D}_3^{\oplus}$) in Figure 5.7a (resp. Figure 5.7b), for low values of the x axis, i.e., when the true prevalence values $p_{\mathscr{D}_3^{\ominus}}(S=1)$ (resp. $p_{\mathscr{D}_3^{\oplus}}(S=1)$) becomes small, the SLD-based classifier starts to act as a trivial rejector with low recall, and hence low F_1 score. On the other hand, the quantification performance of SLD does not degrade in the same way, since its MAE is low and flat across the entire x axis in Figures 5.7a and 5.7b. This is a first hint of the fact that classification and quantification performance may be decoupled.

PACC is another method that significantly outperforms CC in estimating the prevalence of sensitive attributes in both test subsets $\mathscr{D}_3^{\ominus}$, $\mathscr{D}_3^{\oplus}$. In fact, its MAE is well aligned with that of SLD, displaying low quantification error under all protocols (Figures 5.6–5.7). On the other hand, its classification performance is aligned with the accuracy and F_1 score of CC, which is unsatisfactory and can even become worse than random. This fact shows that it is possible to build models which yield good prevalence estimates for the sensitive attribute within a sample, without providing reliable demographic estimates for single instances. Indeed, quantification methods of type *aggregative* (that is, based on the output of a classifier – like all methods we use in this study) are suited to repair the initial prevalence estimate (computed by classifying and counting) without precise knowledge of which specific data points have been misclassified. In the context of models to augment data and measure fairness under unawareness of sensitive attributes, we highlight this as a positive result, decoupling a desirable ability to estimate sample-level prevalence from the potential for undesirable misuse at the individual level.

5.4.9 Ablation Study

Motivation and setup

In the previous sections, we tested six approaches to estimate demographic disparity. For each approach, we used multiple quantifiers for the sensitive attribute S, namely one for each class in the codomain of the classifier h, as described in Step 3 of the method for quantification-based estimate of demographic disparity. In the binary setting adopted in this work, where $\mathscr{Y} = \{\ominus, \oplus\}$, we trained two quantifiers. A quantifier was trained on the set of positively-classified instances of the auxiliary set $\mathscr{D}_2^{\oplus} = \{(\mathbf{x}_i, s_i) \in \mathscr{D}_2 \mid h(\mathbf{x}) = \oplus\}$ and deployed to quantify the prevalence of sensitive instances (such that $S = s$) within the test subset $\mathscr{D}_3^{\oplus}$. The remaining quantifier was trained on $\mathscr{D}_2^{\ominus}$ and deployed on $\mathscr{D}_3^{\ominus}$.

Training and maintaining multiple quantifiers is more expensive and cumbersome than having a single one. Firstly, quantifiers that depend on the classification outcome $\hat{y} = h(\mathbf{x})$ require retraining every time h is modified, e.g., due to a model update being rolled out. Second, the maintenance cost is multiplied by the number of classes $|\mathscr{Y}|$ that are possible for the outcome variable. To ensure that these downsides are compensated by performance improvements, we perform an ablation study and evaluate the performance of different estimators of demographic disparity supported by a single quantifier.

In this section, we concentrate on three estimation approaches, namely PCC, SLD, and PACC. SLD and PACC are among the best overall performers, displaying low bias or variance across all protocols. PCC shows great performance in situations where its posteriors are well-calibrated on $\mathscr{D}_3$. We compare their performance in two settings. In the first setting, adopted so far, two separate quantifiers $q_\ominus$ and $q_\oplus$ are trained on $\mathscr{D}_2^\ominus$, $\mathscr{D}_2^\oplus$ and deployed on $\mathscr{D}_3^\ominus$, $\mathscr{D}_3^\oplus$, respectively. In the second setting, we train a single quantifier q on $\mathscr{D}_2$ and deploy it separately on $\mathscr{D}_3^\ominus$ and $\mathscr{D}_3^\oplus$ to estimate $\hat{\boldsymbol{\delta}}_h^S$ using Equations (5.14) and (5.15), specialized so that $q_\ominus$ and $q_\oplus$ are the same quantifier.

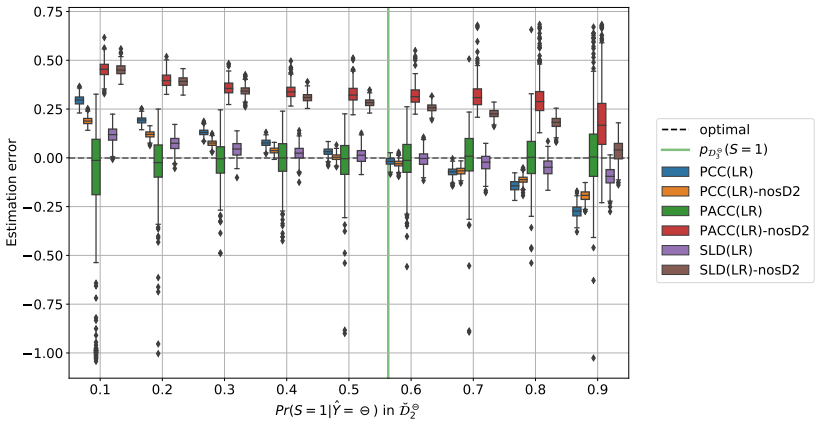
Results

Figure 5.8 summarizes results for the Adult dataset under two protocols that are representative of the overall trends, namely `sample-prev-`$\mathscr{D}_2$ (Figure 5.8a) and `sample-prev-`$\mathscr{D}_3$ (Figure 5.8b).[9] The y axis depicts the estimation error of PCC, SLD, PACC, and their single-quantifier counterparts, denoted by the suffix "nosD2" to indicate that the auxiliary set $\mathscr{D}_2$ is not split into $\mathscr{D}_2^\ominus$, $\mathscr{D}_2^\oplus$ during training. The x axis depicts the quantity of interest varied under each protocol.

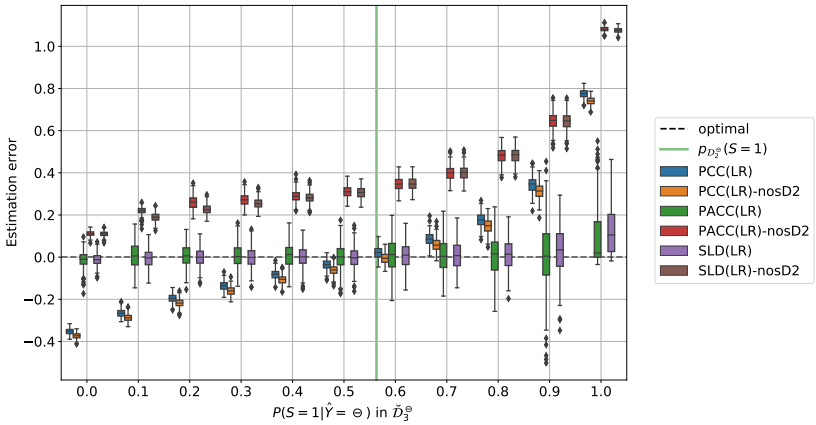
Interestingly, PCC appears to be rather insensitive to the ablation study, so that the estimation errors of PCC and PCC-nosD2 are well-aligned. PCC-nosD2 performs slightly better under the protocol `sample-prev-`$\mathscr{D}_2$, where the auxiliary set is small, and splitting it to learn separate quantifiers may result in poor performance. The opposite is true for PACC-nosD2, showing a clear decline in performance in the single-quantifier setting. This is due to the fact that the estimates of tpr (and fpr) in $\mathscr{D}_3^\oplus$ and $\mathscr{D}_3^\ominus$ for the adjustment (Equation 5.9) are more precise when issued by dedicated estimators rather than a single one computed without splitting $\mathscr{D}_2$. SLD-nosD2 also shows a sizeable performance decay.

Under all protocols, the performance of SLD and PACC is compromised in the absence of class-specific quantifiers $q_\ominus$ and $q_\oplus$. If a single quantifier is trained on the full auxiliary

[9] In the interest of brevity, the figures in this section refer to LR-based quantification on the Adult dataset under two protocols. The results for SVM-based quantifiers under each protocol are depicted in the Appendix (Figures C.1 – C.5). Analogous results hold on CreditCard and COMPAS.



(a) Protocol `sample-prev-`$\mathscr{D}_2^{\ominus}$



(b) Protocol `sample-prev-`$\mathscr{D}_3^{\ominus}$

Fig. 5.8 Ablation study on the Adult dataset. Distribution of the estimation error (y axis) for CC, PACC, SLD, and their single-quantifier counterparts, as $\Pr(S=1|\hat{Y}=\ominus)$ vary in $\mathscr{D}_2$, plot (a), and $\mathscr{D}_3$, plot (b).

set $\mathscr{D}_2$, the corrections brought about by SLD and PACC can end up worsening, rather than improving, the prevalence estimates of vanilla CC. PCC is less sensitive to the ablation, showing small performance differences in both directions under the single quantifier setting. In general, it seems beneficial to partition the auxiliary set into subsets $\mathscr{D}_2^{\ominus}$ and $\mathscr{D}_2^{\oplus}$ according to the method in Section 5.3.2.

5.5 Discussion

Overall, this chapter shows that quantification approaches are suited to annotate datasets at the sample level with sensitive attributes and measure demographic parity under unawareness if a small auxiliary dataset, containing sensitive and non-sensitive attributes, is available. This

is a common setting in real-world scenarios, where such datasets may originate from targeted efforts or voluntary disclosure. Despite an inevitable selection bias, these datasets still represent a valuable asset for fairness audits, if coupled with robust estimation approaches. Indeed, several quantification methods tested in this work provide precise estimates of demographic disparity despite the distribution shift across training and testing caused by selection bias and other distribution shifts that arise in the context of human processes. This is an important improvement over CC and PCC, previously studied in the algorithmic fairness literature as the *threshold estimator* and *weighted estimator* [140]. SLD strikes the best balance in performance across all protocols; we suggest its adoption, especially when the distribution shift between development and deployment conditions has not been carefully characterized. Moreover, while the development of proxy methods typically comes with a potential for misuse on individuals (e.g., profiling), quantification approaches demonstrate the potential to circumvent this issue. More in detail, from the above experimental section, we summarize the following trends concerning different approaches to measure demographic parity under unawareness.

Datasets can be annotated with sensitive information using quantification. For example, the problem of measuring group fairness under unawareness can be cast as a prevalence estimation problem and effectively solved using methods of proven consistency from the quantification literature. We demonstrate several estimators that outperform previously proposed methods [140], corresponding to CC and PCC, i.e., two weak baselines in the quantification literature.

CC is suboptimal. Naïve Classify-and-Count represents the default approach for practitioners unaware of quantification. Ad hoc quantification methods outperform CC in most combinations of 5 protocols, 3 datasets, and 2 underlying learners.

PCC suffers under distribution shift. As long as the underlying posteriors are well-calibrated, PCC is a strong performer. However, when its training set and test set have different prevalence values for the sensitive attribute S, a common situation in practice, PCC displays a systematic estimation bias, which increases sharply with the prior probability shift between training and test.

HDy, ACC and PACC deteriorate in the small data regime. These methods require splitting their training set (that is, the auxiliary set $\mathscr{D}_2$), so their performance drops faster when its cardinality is small. PACC and ACC display good median performance but a large variance; the former method always outperforms the latter.

SLD strikes a good balance. This method was shown to be the best performer under (the inevitable) distribution shift between the auxiliary set $\mathscr{D}_2$ and the test set $\mathscr{D}_3$, with a moderate performance decrease when $|\mathscr{D}_2|$ becomes small. However, in situations where it is

not possible to maintain separate quantifiers for positively and negatively predicted instances, its performance drops substantially.

Decoupling is possible. Methods such as SLD and PACC fare much better than CC in estimating group-level quantities (such as demographic parity), while if misused for individual classification of sensitive attributes, the improvement is minor (SLD) or zero (PACC).

5.6 Chapter Outcomes

Measuring the differential impact of models on groups of individuals is important to understand their effects in the real world and their tendency to encode and reinforce divisions and privilege across sensitive attributes. Unfortunately, in practice, demographic attributes are often not available. In this work, we have taken the perspective of responsible practitioners, interested in internal fairness audits of production models. We have proposed a novel approach to annotate data with sensitive attributes and measure group fairness under unawareness, utilizing methods from the quantification literature. These methods are specifically designed for group-level prevalence estimation rather than individual-level classification. Since practitioners who try to measure fairness under unawareness are precisely interested in prevalence estimates of sensitive attributes (Proposition 1), it is useful for the fairness and quantification communities to exchange lessons.

More in detail, we have studied the problem of estimating a classifier's fairness under unawareness of sensitive attributes, with access to an auxiliary set of data for which demographic information is available, a highly relevant setting for real-world fairness audits [697]. We have shown how this can be cast as a quantification problem, and solved with established approaches of proven consistency. We have conducted a detailed empirical evaluation of different methods and their properties focused on demographic parity. Drawing from the algorithmic fairness literature, we have identified five important factors for this problem, associating each of them with a formal evaluation protocol. We have tested several quantification-based approaches, which, under realistic assumptions for an internal fairness audit, outperform previously proposed estimators in the fairness literature. We have discussed their benefits, including the unbiasedness guarantees of some methods, and mitigation of misuse at an individual level.

Future work may require a deeper study of the relation between classification and quantification performance and the extent to which these two objectives can be decoupled. It would be interesting to explicitly target decoupling through learners aimed at maximizing quantification performance subject to a low classification performance constraint. Ideally,

decoupling should provide precise privacy guarantees to individuals while allowing for precise sample-level estimates. Another important avenue for future work is the study of confidence intervals for fairness estimates provided by quantification methods. A reliable indication of confidence for estimates of group fairness may be invaluable for a practitioner measuring diversity or arguing for resources and attention to the disparate effects of a model on different populations. Finally, the estimators presented in this work may be plugged into optimization procedures aimed at improving, rather than measuring, algorithmic fairness. It will be interesting to evaluate fairness estimators in this broader context and extend them, e.g., to ranking problems and counterfactual settings.

Chapter 6

Measures and Datasets

Ranking is a central component in Search Engines (SE), two-sided markets, recommender and match-making systems. These platforms, collectively referred to as Information Access Systems (IAS), act as intermediaries between providers and consumers of items of diverse nature, facilitating access to information, entertainment, accommodation, products, services, jobs, and workers.

Fairness in ranking is a central aim of industry and academia, with entire tracks,[1] workshops,[2] and corporate teams devoted to such a complex topic. Research and development in this space are focused on ensuring that stakeholders of ranking systems do not unjustly suffer discrimination or harm. Alongside datasets, measures play a central role in the fairness measurement process. We borrow from Ekstrand et al. [229] to present a fair-ranking taxonomy along two key dimensions: the type of harm that is being prevented and the people who benefit from this effort. Based on their position in the information pipeline, fairness can help the following stakeholders.

1. Consumers of item (e.g., SE users) are concerned by various problems in this space, including privacy, content diversity, and targeted advertisement, which can result in unequal opportunity for different segments of the population [122, 283, 286].

2. Producers of items (e.g., marketplace sellers) are interested in being read, viewed, and clicked by consumers, whose attention should be directed to items based on equitable criteria [721, 831, 844, 857].

3. Information subjects (e.g., people mentioned in items or SE queries) may be present in contexts where they would rather not appear or, conversely, neglected or censored out against their will [39, 433, 599, 617].

[1] https://fair-trec.github.io/
[2] http://bias.disim.univaq.it/

These stakeholders can be affected by harms of different nature.

1. Distributional harms are inequities in access to a resource of interest, such as education, jobs, credit, possibility of parole, and exposure [74, 721, 857].

2. Representational harms typically occur when individuals and groups are unable to self-determine their image, which can end up being stereotyped, inadequate, or offensive [2, 599, 617].

In recent years, a deluge of fair ranking measures have been proposed [230, 639, 794]. Tens, if not hundreds, of fair ranking measures coexist, often capturing similar properties [660]. Part of this proliferation and redundancy is due to a lack of normative reasoning: fairness measures are often introduced as self-evident prerequisites for equity, resulting in downstream uncertainty as to what exactly is being measured or optimized. In contrast, it is fundamental to distinguish between the *construct* of a measure, that is, the theoretical property targeted by it (e.g. fame), and its *operationalization*, that is, the mathematical formulation adopted to capture this property (e.g. number of followers) [398]. This distinction enables a clear understanding of measures and their applicability individually, along with a more informed study and consideration of the available measures as a whole.

Pairwise fairness [67, 471, 591] is a family of group measures for item providers, where groups descend from sensitive attributes, such as race and sex. Their underlying construct has not been characterized in prior literature, preventing an understanding and informed adoption on part of the research community. In summary, pairwise fairness counts how frequently a ranking arranges two items from different groups according to their merit, and whether wrong pairings benefit one group more than another. Although this notion of equity seems valid at face value, it does not directly measure producer benefits and, for this reason, deserves further scrutiny.

O1: Study pairwise fairness with a focus on the underlying construct, its relation to other fair ranking measures, its limitations, and possible improvements.

Despite a multitude of fair ranking measures [230, 639, 794], some aspects of equitable ranking have been neglected in the literature. Most importantly, very few measures study fairness towards information subjects and related issues [230], despite substantial evidence of representational harms caused by IAS, reinforcing damaging stereotypes along gendered and racial lines [433, 599, 617]. Perhaps surprisingly, no quantitative approach has been developed to estimate these trends.

In particular, SE have been shown to perpetuate well-known gender stereotypes identified in psychology literature and to influence users accordingly. For example, Google Image

search results reflect current gender differences in profession, with a tendency to slightly exaggerate [433]. In addition, manipulating the representation of women and men in job search results can significantly affect people's perceptions of gender ratios in those jobs [433]. Moreover, a study on Bing photos found more women in depictions of warm traits (e.g. sensitive), while competence traits (e.g. intelligent) yielded more photos of men [617]. Problematic and stereotypical results can also be found in autocomplete suggestions for gendered queries [599]. These results highlight the importance of measuring and counteracting gender biases in SE.

O2: Develop a measure to quantify the tendency of a SE to reinforce gender stereotypes in its users.

This chapter is organized as follows. In Section 6.1, we target **O1** with a careful study and improvement of pairwise fairness. Leveraging browsing models and measurement theory, we propose an interpretation centered around the construct of producer dissatisfaction, explicitly connecting pairwise fairness to perceptions of injustice and platform quality. After highlighting the key limitations of prior measures, we introduce a set of targeted enhancements and combine them into a novel measure, analyzing its connection with popular measures of fair ranking. Next, in Section 6.2 we focus on **O2**, proposing the Gender Stereotype Reinforcement (GSR) measure, which quantifies the tendency of a SE to support gender stereotypes. Through the critical lens of construct validity, we validate the proposed measure on synthetic and real collections. Subsequently, we use GSR to compare widely-used information retrieval algorithms, including lexical, semantic, and neural models. Finally, Section 6.3 summarizes the results of the chapter, contextualizing their relevance for the fairness and ranking communities. The notational conventions common to both sections are summarized in Table 6.1.

6.1 Dissatisfaction Induced by Pairwise Swaps

Information Access Systems (IAS) facilitate user interactions with content by ranking and presenting items to their users according to estimated merit or relevance [32, 406]. *Content producers* in Information Access Systems (IAS) are increasingly recognized as stakeholders whose economic and societal needs must be taken into account, along with those of *consumers*, to foster a fruitful and equitable information ecosystem [230, 639, 860, 861]. Their needs can be considered individually [75, 90, 207] or based on group membership [71, 696, 722] determined by sensitive attributes such as gender or race. To these ends,

symbol	meaning
$\mathscr{I}$	set of available items or documents
$\mathscr{Q}$	set of queries in search history
$N = \lvert\mathscr{Q}\rvert$	number of queries in search history
$i \in \mathscr{I}$	a specific item
$q_j \in \mathscr{Q}$	a specific query
r_i^j	relevance of item i to query q_j
σ_j	a ranking (permutation of items) issued in response to query q_j
σ_{j*}	ideal ranking for q_j, i.e., $\sigma_{j*} = \text{argsort}(r_i^j)$
$\sigma_j(k)$	item at rank k in σ_j
$\sigma_j^{-1}(i)$	rank of item i in σ_j

Table 6.1 Notation for fair ranking measures.

several measures of *fairness* in ranking have been proposed, capturing notions of equity of *exposure* [209, 722], *representation* [25, 843], or *pairwise accuracy* [67, 471].

As mentioned in the introduction of this chapter, when considering a measure, it is important to distinguish between its construct, that is, the theoretical property targeted by the measure (e.g. fame), and its operationalization, that is, the mathematical formulation adopted to capture this property (e.g. number of followers) [398]. In this regard, exposure- and representation-based measures in fair ranking operationalize well-defined constructs, clearly connected to the desiderata of producers. They measure the presence of salient groups of providers in the most visible positions of a ranking, increasing their chance of being viewed by IAS users and consequently gain benefits, such as clicks, purchases, or downloads. On the contrary, in the prior literature, pairwise fairness has not been clearly associated with a quantity of practical interest for producers [67, 471, 591]. In a nutshell, measures of pairwise fairness quantify how often the rank of two items from different groups reflects their merit and whether mismatched pairs are systematically in favor of one group. This notion of equity is less clearly connected with immediate producer benefits and thus deserves further scrutiny.

In this section, we perform an in-depth study of pairwise fairness. First, we provide an interpretation of pairwise fairness grounded in browsing models [116], developing a rigorous distinction between the construct and its operationalization [398]. We show that pairwise fairness can capture perceived injustice on part of item producers, and thus operationalize their *dissatisfaction* with the output of an IAS. Second, we highlight several limitations of existing pairwise fairness metrics and derive a novel metric that overcomes the issues. Our measure improves on previous proposals by modeling realistic browsing behaviors, individual user perspectives, and relevance ties. It captures key aspects of observed injustice

symbol	meaning
$\mathscr{S} = \{A, B\}$	alphabet for sensitive attribute of item producer
$s \in \mathscr{S}$	value for sensitive attribute
$i \in s$	membership of item i in sensitive group s
$d(i, i')$	indicator function for discordant pair

Table 6.2 Notation for DIPS.

and dissatisfaction, related to perceived quality of IAS and likelihood of withdrawal on the part of producers, which is a central concern for platform owners. Finally, we characterize the relationship between pairwise and exposure-based measures analytically and empirically. We show their key similarities, inherited from browsing models, and highlight their differences due to the underlying normative constructs.

6.1.1 Background and Related Work

Fair ranking. Typically supported by the objectives of bias mitigation [85, 127], equity [75, 722], and diversity [542, 744], fair ranking is concerned with accurately ordering items without unjust discrimination and is complicated by factors such as multiple protected groups [859], outliers [696], and duplicates [207]. Recent surveys and comparative analyses of fair ranking omit measures of pairwise fairness [660, 859] or simply frame them as accuracy-based [230, 639]. A clear discussion of the construct underlying pairwise fairness is lacking in the literature [67, 471, 591], hindering an informed adoption of these measures and understanding of how they relate to item producers and equity towards them.

Notation. The key notation for this section is reported in Table 6.2; it is worth recalling that the common notation for this chapter is summarized in Table 6.1. In this section, we drop the subscript j, that is, $\sigma_j = \sigma$, since there is no need to discuss multiple queries. Let $\mathscr{I}$ denote a set of items that need to be ranked, and let i be an item from this set. Let r_i denote the relevance of item i in a given ranking. Moreover, let s denote a sensitive attribute, taking binary values in $s \in \mathscr{S} = \{A, B\}$ for ease of exposition.[3] Let $i \in s$ denote the membership of i in group s. Let σ_* denote an "ideal" ranking, i.e., a permutation which orders items decreasingly by relevance: $\sigma_* = \text{argsort}(r_i)$. Finally, let σ denote a ranking returned by the IAS in response to a query, and $\sigma(k)$ indicate the item ranked at position k by σ.

[3] We follow the literature on pairwise fairness and consider binary sensitive attributes. Extensions to settings with more than two groups can be defined in multiple ways starting from individual measures (§6.1.3) and are left to future work.

Discordant pairs. Central to pairwise fairness is the definition of a *discordant pair*. Two items $i, i' \in \mathscr{I}$ represent a discordant pair if their relative ordering in σ and σ_* differs. More formally, let $\sigma^{-1}(i)$ be the position of item i in ranking σ, i.e., $\sigma^{-1}(i) = k \iff \sigma(k) = i$. Given two rankings, σ and σ_*, the indicator function for a discordant pair is defined as

$$d(i,i') = \underbrace{\mathbb{1}(\sigma^{-1}(i) < \sigma^{-1}(i'), \sigma_*^{-1}(i) > \sigma_*^{-1}(i'))}_{d_F(i,i')} + \underbrace{\mathbb{1}(\sigma^{-1}(i) > \sigma^{-1}(i'), \sigma_*^{-1}(i) < \sigma_*^{-1}(i'))}_{d_U(i,i')}$$

In other words, i can be part of a discordant pair when ranking σ unfairly places it at an advantage (d_F) or a disadvantage (d_U) over another item i'. We explicitly remark this by writing $d(i,i') = d_F(i,i') + d_U(i,i')$, where subscripts F and U indicate that the first item is part of a Favorable Discordant Pair (FDP) or an Unfavorable Discordant Pair (UDP).

Pairwise Fairness. Inter-Group Inaccuracy (IGI) [67] and Rank Equality Error (REE) [471], the most popular measures of pairwise fairness, are defined as

$$M_{AB} = \frac{1}{C_{AB}} \cdot \sum_{i \in A} \sum_{i' \in B} d_U(i,i'). \tag{6.1}$$

The key difference between IGI and REE is the normalizing constant C_{AB}. We defer a detailed analysis of this aspect to Section 6.1.3. M_{AB} measures how often items $i \in A$ are part of a UDP with items $i' \in B$. On the contrary, M_{BA} measures the frequency of cross-group UDPs where items from B are at a disadvantage. Following [67], fairness requirements can be expressed as

$$M_{AB} - M_{BA} = 0, \tag{6.2}$$

i.e., prescribing equality in the frequency of discordant pairs between sensitive groups. Unfortunately, a clear discussion of the normative reasoning behind this measure, along with its merits and limitations, is lacking in the literature, thus hindering an informed adoption of pairwise fairness and its proper contextualization in the fair ranking landscape.

6.1.2 A Critical Review of Pairwise Fairness

Following Jacobs and Wallach [398], we discuss fairness measures distinguishing between their construct, i.e., the theoretical property targeted by the measure, and their operationalization, i.e., the particular mathematical formulation meant to capture that property. Ideally,

fairness measures should follow clear normative reasoning, explicitly discussing what it means for an algorithm to be equitable and from which perspective. In other words, the value-laden requirements for an equitable algorithm (*construct*) should be enunciated and discussed, before finding a quantitative way to measure their fulfillment (*operationalization*). However, often fairness measures are introduced as self-evident prerequisites for equity, resulting in downstream uncertainty as to what exactly is being measured or optimized.

The above considerations are especially applicable to measures of pairwise fairness in ranking. For example, REE is based on the "postulate that there is value in considering error-based fairness criteria for rankings" [471]. Similarly, for IGI, Beutel et al. [67] "draw on the intuition of Hardt et al. [344] for equality of odds, where the fairness of a classifier is quantified by comparing either its false positive rate and/or false negative rate." Although somehow related to fairness, intended as equalization of some property between groups, according to Equation (6.2), a clear explanation of the construct behind these measures is lacking in the literature. In this section, we target this gap by analyzing pairwise fairness measures in depth and retrospectively mapping them to their underlying construct(s).

Implicit browsing models

First, it is useful to observe that REE and IGI are closely related to Kendall's Tau [436], a rank correlation measure defined as $\tau(\sigma,\sigma_*) = 1 - \frac{2}{C}\cdot\sum_i\sum_{i'\neq i} d(i,i')$, with $C = n(n-1)/2$. In essence, computing Kendall's Tau requires enumerating every pair of items and counting discordant ones. Following Equation 6.1, let us define *inaccuracy* as the frequency of discordant pairs in σ and σ_*

$$M = \frac{1}{C}\cdot\sum_i\sum_{i'\neq i} d(i,i'), \tag{6.3}$$

from which Kendall's Tau is computed via the linear transformation $\tau = 1 - 2\cdot M$. For the sake of simplicity, we will temporarily concentrate on Kendall's Tau and its interpretation(s), and subsequently reintroduce the complexity of sensitive attributes to specifically study REE and IGI.

Second, note that item pairs can be enumerated by browsing the ranking σ according to a cascade model [178]. To enumerate every pair, we can browse σ from top ($k = 0$) to bottom ($k = n-1$) and compare the current item $\sigma(k)$ (the item at rank k in σ) with items further up in the ranking, to determine whether they constitute a discordant pair.

$$M = \frac{1}{C}\cdot\sum_{k=1}^{n-1}\sum_{k'=0}^{k-1} d(\sigma(k),\sigma(k'))$$

With shorthand notation, we write the indicator function for a discordant pair of items ranked by σ in positions (k,k') as $d(k,k') = d(\sigma(k),\sigma(k'))$.

Moreover, let us define a trivial browsing model, according to which users visit the positions in a ranking with uniform (unit) probability across all ranks. More formally, $F(k) = 1 \ \forall k$, where $F(k)$ denotes the probability that users will visit the item $\sigma(k)$. With this notation, we can write the following alternative formulas for M:

$$\textbf{Item-centric } M = \frac{1}{C} \cdot \sum_{k=0}^{n-1} \sum_{k'=0}^{k-1} F(k') d_U(k,k')$$
$$\textbf{User-centric: } M = \frac{1}{C} \cdot \sum_{k=0}^{n-1} F(k) \sum_{k'=0}^{k-1} d_U(k,k') \tag{6.4}$$

As shown in the next section, these alternative formulations capture the perspectives and desiderata of item producers (item-centric) or item consumers (user-centric). They are equivalent under the trivial browsing model defined above, but in general they yield different results for M. Both provide a way to count and weigh each pair of items by sequentially traversing a ranking according to a specified browsing model $F(k)$.

Interpretations

Below we provide two alternative interpretations of Kendall's Tau based on the above formulations:

- **Item-centric.** Producers of items at each rank k evaluate ranking σ by focusing on the most visible injustices against their item $\sigma(k)$. Their dissatisfaction with σ grows every time they encounter a UDP for $\sigma(k)$, that is, an item of lesser relevance ranked better than their own. The inner summation $\sum_{k'=0}^{k-1} F(k') d_U(k,k')$ is a weighted counter of UDPs, with weight proportional to the visibility of the unjustly favored item. According to this interpretation, Kendall's Tau operationalizes aggregate producer dissatisfaction with σ for unjustly favoring other items.

- **User-centric.** Users browse the ranking σ sequentially, visiting items in rank k with probability $F(k)$. Every time they visit an item $\sigma(k)$, if an item of lesser relevance was unduly positioned above it, users add 1 to a counter measuring wasted effort in arriving at the item in position k. According to this interpretation, Kendall's Tau operationalizes user dissatisfaction due to wasted browsing effort.

These interpretations are also applicable to group-based measures of pairwise fairness, such as IGI and REE (Equation 6.1), with the caveat of focusing on cross-group comparisons.

To exemplify, let us focus on the item-centric formulation and consider

$$M_{AB} = \frac{1}{C_{AB}} \cdot \sum_{k=0}^{n-1} \sum_{k'=0}^{k-1} F(k') d_U(k,k') \cdot \mathbb{1}(\sigma(k) \in A, \sigma(k') \in B)$$

Item-centric interpretations for IGI and REE convey the dissatisfaction of items and producers from one group for being unjustly ranked worse than items of lesser relevance from a different group. More specifically, suppose that an item in position k' belongs to group B; the producers of group A evaluate whether this item is unjustly ranked above their items despite having lower merit. They contribute to an inter-group dissatisfaction counter, which is weighted according to the probability of a visit at rank k', i.e., to the visibility of the unjustly favored item. In other words, if an item i' is unjustly ranked better than another item i, but in a position with low visibility (such as $\sigma^{-1}(i') = 900$ under a top-heavy browsing model), the producer of i is unlikely to notice, while they are more likely to observe the UDP and increase their dissatisfaction if i' is very visible. According to this interpretation, M_{AB} represents the dissatisfaction of group A with the ranking σ, due to their items being unjustly ranked below the items of group B (in expectation over the browsing model $F(k)$ and after normalization). Pairwise fairness is thus connected to observed injustice, perceived quality of platform service, and likelihood of withdrawal on part of item producers.

User-centric interpretations, on the other hand, center around wasted effort due to user attention being diverted to items of lower interest from a different group. Users visit an item with probability $F(k)$, taking into account its group (say, $s = A$). They evaluate how much effort they wasted to reach this item because they examined items of inferior relevance from different groups. According to this interpretation, the counter measures wasted effort to reach items in group A that are unduly ranked below items in group B, and M_{AB} represents a normalized expectation of cross-group wasted effort over the browsing model.

6.1.3 Proposed Measure

The fact that multiple interpretations are possible speaks for the flexibility of pairwise measures in operationalizing multiple constructs, related to user and item satisfaction. Clearly, these measures can be adapted in different ways towards capturing these concepts, but in their current, non-specialized form they fail to realize their full potential. Overall, this is due to several limitations of IGI and REE. In this section, we describe these limitations and propose specific refinements to overcome them.

Individual pairwise fairness

Limitation. Just like aggregate performance measures can obscure poor performance for groups of people [49, 318], group measures can obscure poor performance for individuals. IGI and REE focus solely on groups, completely neglecting the individual perspective. Capturing individual perspectives is desirable for more precise analyses and requires a version of these measures with finer granularity.

Refinement. To overcome this limitation, we define an individual version of pairwise fairness that captures the dissatisfaction of each item. This is a straightforward extension that allows analyses at a finer granularity.

$$M_i = \sum_{i'=0}^{n-1} d_U(i,i') \tag{6.5}$$

Moreover, if sensitive groups are salient, item producers may take this information into account. We can account separately for the envy towards items of different groups by defining

$$M_{iA} = \sum_{i'=0}^{n-1} d_U(i,i') \cdot \mathbb{1}(i' \in A);\ \ M_{iB} = \sum_{i'=0}^{n-1} d_U(i,i') \cdot \mathbb{1}(i' \in B),$$

from which Equation (6.1) emerges additively as

$$M_{AB} = \frac{1}{C_{AB}} \sum_{i \in A} M_{iB};\ \ M_{BA} = \frac{1}{C_{BA}} \sum_{i \in B} M_{iA}$$

In other words, we can define individual dissatisfaction measures (Equation 6.5), from which groupwise dissatisfaction (Equation 6.1) emerges naturally as a sum of individual components. This is a useful property, providing an intuitive connection between individual and group perspectives, and guaranteeing that interventions at the individual level, making M_i smaller $\forall i \in \{0,\dots,n-1\}$, will be beneficial at the group level, resulting in reductions in M_{AB} and M_{BA}.

Top-heaviness

Limitation. The lack of assumptions about browsing behavior in pairwise fairness is only apparent. In fact, REE and IGI are compatible with the trivial browsing model described in Equation (6.4), i.e., $F(k) = 1\ \ \forall k$, but can be specialized with more realistic ones. Indeed, UDPs towards the top of a ranking are likely to be more visible and cause greater dissatisfaction. Equation (6.1) fails to capture this aspect.

Refinement. As previously shown, we can explicitly plug user browsing models into pairwise fairness and define

$$M_i = \sum_{k=0}^{n-1} F(k) d_U(i, \sigma(k))$$

In other words, dissatisfaction M_i of the item i is a weighted counter of UDPs, with weights proportional to the probability of visiting the item unjustly ranked better than i.

Like in Section 6.1.3, we can incorporate group membership into the individual measure

$$M_{iB} = \sum_{k=0}^{n-1} F(k) d_U(i, \sigma(k)) \cdot \mathbb{1}(\sigma(k) \in B)$$

and aggregate it to summarize cross-group dissatisfaction

$$M_{AB} = \frac{1}{C_{AB}} \sum_{i \in A} \sum_{k=0}^{n-1} F(k) d_U(i, \sigma(k)) \cdot \mathbb{1}(\sigma(k) \in B). \tag{6.6}$$

In summary, pairwise fairness measures can be flexibly modified, both at the individual and group level, to incorporate a desired position bias, by choosing a suitable user browsing model $F(k)$.

Tie handling

Limitation. Pairwise measures are of limited utility in situations where relevance r_i admits ties. Indeed, this is a frequent situation: in recommender systems, for example, user ratings are often quantized [348], while, in information retrieval, relevance judgements are typically discrete, either binary or graded [346].

Refinement. Recalling that $\sigma_* = \text{argsort}(r_i)$, we rewrite the indicator function for UDPs as

$$d_U(i, i') = \mathbb{1}(\sigma^{-1}(i) > \sigma^{-1}(i'), r_i > r_{i'})$$

explicitly showing that relevance ties are simply neglected. As a remedy, we propose to extend the notion of UDP to handle ties as

$$\begin{aligned} d_U(i, i') = &\mathbb{1}(\sigma^{-1}(i) > \sigma^{-1}(i'), r_i > r_{i'}) \\ &+ c_t \mathbb{1}(\sigma^{-1}(i) > \sigma^{-1}(i'), r_i = r_{i'}) \end{aligned} \tag{6.7}$$

where c_t models the dissatisfaction of an item ranked below another item of the same relevance. If i and i' are tied (such that $r_i = r_{i'}$) the item with a worse rank considers this pair a *partial UDP*. Sensible values for c_t range in $(0,1)$, where $c_t = 1$ corresponds to equating partial UDPs to proper UDPs in terms of dissatisfaction, while $c_t = 0$ indicates indifference to comparisons with items of the same relevance.

Normalization

Limitation. Normalization is a seemingly intuitive, often overlooked procedure. Its main benefit lies in making the range of a measure independent of the context, bounding it below a given value, typically equal to one, in every operational context. In turn, this fact makes the values of a measure simpler to understand and, potentially, plug into a multi-term objective function. However, these benefits are accompanied by downsides, which are not always understood, and represent a potentially dangerous instance of *unknown unknowns*. To illustrate them in the context of fair ranking, let us recall that IGI and REE can be written as

$$M_{AB}^{\text{IGI,REE}} = \frac{1}{C_{AB}^{\text{IGI,REE}}} \cdot \sum_{i \in A} \sum_{i' \in B} d_U(i, i')$$

with different normalizing constants

$$C_{AB}^{\text{IGI}} = \sum_{i \in A} \sum_{i' \in B} \mathbb{1}(r_i > r_{i'}); \;\; C_{AB}^{\text{REE}} = N_A \cdot N_B, \tag{6.8}$$

where N_A (N_B) denotes the number of items in $\mathscr{I}$ that belong to the sensitive group A (B). In other words, IGI is normalized with respect to a worst-case scenario which takes into account the ground truth relevance r_i, and its distribution across sensitive groups, while REE is normalized with respect to the *a-priori* worst case, which does not take into account r_i. This means that, for REE, the normalizing constant is the same for M_{AB} and M_{BA} ($C_{AB}^{\text{REE}} = C_{BA}^{\text{REE}}$), while for IGI they usually differ ($C_{AB}^{\text{IGI}} \neq C_{BA}^{\text{IGI}}$).

At first sight, the normalization mechanism underlying IGI seems preferable as, in a way, it makes the respective measures M_{AB}^{IGI} and M_{BA}^{IGI} less context dependent. In fact, regardless of the distribution of relevance scores r_i across groups, M_{AB}^{IGI} and M_{BA}^{IGI} can always reach the theoretical maximum ($M = 1$), in the presence of a particularly unfortunate ranking that systematically puts either group at total disadvantage. This is not true for their REE counterparts.

However, this property comes with a large downside, that is, it becomes unclear how M_{AB}^{IGI} and M_{BA}^{IGI} should be compared. This fact is best explained with a toy example where the ideal ranking is $\sigma_* = [i_0^A, i_1^B, i_2^A, i_3^A]$, and where the superscript s in i^s denotes membership of i to

sensitive group s. In this situation, as is typical, we have different constants for IGI ($C_{AB}^{\text{IGI}} = 1$, $C_{BA}^{\text{IGI}} = 2$) and equal constants for REE ($C_{AB}^{\text{REE}} = C_{BA}^{\text{REE}} = 3$). A ranking $\sigma = [i_2^A, i_1^B, i_0^A, i_3^A]$, obtained by simply exchanging i_0^A and i_2^A in σ_*, produces two UDPs, one (i_0^A, i_1^B) in favor of group B and another one (i_1^B, i_2^A) in favor of group A. The resulting measures for IGI are $M_{AB}^{\text{IGI}} = 1 \gg M_{BA}^{\text{IGI}} = 0.5$. Taken at face value, this would suggest that group B is largely favored over group A, and that the latter should be more dissatisfied with σ than with the former. We argue that this is not necessarily true, as, from a groupwise perspective, σ and σ_* are equivalent. In fact, under the IGI normalization scheme, comparing M_{AB}^{IGI} and M_{BA}^{IGI} is not trivial. This is a very practical problem, since fairness, according to Equation (6.2), is defined precisely as the difference between these quantities.

Refinement. We choose a REE inspired normalization scheme, using the same constant for M_{AB} and M_{BA}, independently of the relevance scores. In Equation (6.6), where we have to take the browsing model into account, we define

$$C_{AB} = C_{BA} = \max\left(N_A \cdot \sum_{k=0}^{N_B-1} F(k), N_B \cdot \sum_{k=0}^{N_A-1} F(k)\right) \tag{6.9}$$

The first term represents a worst-case scenario in which every item in group B is unduly ranked above every item in group A (hence the multiplying factor N_A) and occupies the most visible positions (hence the summation). The second term represents its dual, and the $\max(\cdot)$ function takes care of selecting the largest between these two. This definition allows two desirable properties to hold simultaneously: (1) the difference $M_{AB} - M_{BA}$ is bounded between $(-1, 1)$ and (2) it retains the property of clearly identifying with its sign the advantaged group, since positive (negative) values correspond to rankings σ with more UDPs against group A (B).

Dissatisfaction Induced by Pairwise Swaps

We propose a measure of dissatisfaction that incorporates these refinements. We call this measure Dissatisfaction Induced by Pairwise Swaps (DIPS), and define it as

$$M_{AB}^{\text{DIPS}} = \frac{1}{C_{AB}^{\text{DIPS}}} \sum_{i=0}^{n-1} \sum_{k=0}^{n-1} F(k) d_U(i, \sigma(k)) \cdot \mathbb{1}(i \in A, \sigma(k) \in B), \tag{6.10}$$

DIPS (i) handles ties by setting $c_t = 1$ in the definition of $d_U(\cdot)$ in Equation (6.7), (ii) is normalized with a symmetric constant (independent of relevance scores) according to Equation (6.9), and (iii) inherits a top-heavy behavior from suitable browsing models $F(k)$. Browsing models capture the fact that dissatisfaction is more likely to increase when injustice

is visible. The tunable parameters for DIPS are the browsing model $F(k)$ and the tie-handling constant c_t. For the latter, we recommend an intermediate value $c_t = 0.5$, while the former depends on the application and should be tuned to context-specific browsing behaviour.

Exposure-based measures are a very popular family of fairness measures, which are also based on browsing models [73, 75, 722]. In the remainder of this section, we study their connection with DIPS. We concentrate on a groupwise perspective to allow for a direct comparison with REE and IGI, and leave an analysis centered on individual perspectives to future work.

A Critical Review of Exposure-Based Fairness. Exposure-based measures, in their groupwise version, define an ideal target exposure (T_A, T_B) for each group and measure the distance between this target and the actual exposure (E_A, E_B) assigned by ranking σ to each group. We define the normalized misallocation vector as

$$\delta^\sigma = [\delta_A^\sigma, \delta_B^\sigma] = \left[\frac{T_A}{T_A + T_B} - \frac{E_A}{E_A + E_B}, \frac{T_B}{T_A + T_B} - \frac{E_B}{E_A + E_B} \right] \tag{6.11}$$

where, for a given group s, E_s is the sum of individual exposure values granted by σ to items in the sensitive group s: $E_s = \sum_{i \in s} F(\sigma^{-1}(i))$. To summarize the overall unfairness of a ranking σ in aggregate, we follow Biega et al. [75], and report the ℓ_1 norm of δ^σ. In the following, we consider three measures that differ for the normative reasoning according to which the target exposure quotas are established.

According to Equity of Attention (EA – Biega et al. [75]), the target exposure for a group should be proportional to the sum of the relevance of the items in the said group:

$$T_s^{\text{EA}} = \sum_{i \in s} r_i \tag{6.12}$$

Following a different normative reasoning, we can define a version of EA inspired by demographic parity [50, 108], which requires, for each group, a share of attention proportional to its representation in the overall population

$$T_s^{\text{EA-dp}} = N_s / N. \tag{6.13}$$

Expected Exposure (EE – Diaz et al. [209]) also relies on relevance scores to specify its target exposure; however, different from EA, it assigns to relevance judgements an *ordinal* validity: if item i is more relevant than (as relevant as) i', it should get more (as much) exposure. *How much* more exposure is not explicitly specified by the normative reasoning underlying EE, and remains related to the browsing model $F(k)$. This is contrary to EA, which assigns to relevance judgements a *scale ratio* validity: if item i is twice as relevant as

i', it should get twice as much exposure. Therefore, the target quota for EE is dependent on $F(k)$, and can be expressed numerically as

$$t_i = \text{mean}_{\{i'|r_{i'}=r_i\}}(F(\sigma_*^{-1}(i'))) \tag{6.14}$$

$$T_s^{\text{EE}} = \sum_{i \in s} t_i \tag{6.15}$$

where t_i is the exposure target quota for item i. In a simple setting without relevance ties, t_i is equal to the exposure granted to i by the ideal ranking σ_* under $F(k)$. If ties are present, it is the average exposure granted to items of the same relevance as i by σ_*.

DIPS and Exposure-Based Fairness. According to exposure-based measures, individual misallocation is the difference between the target quota of an item and its actual exposure $F(\sigma^{-1}(i))$, i.e., its probability of a visit given ranking σ. For example, EA defines the target quota of an item as its share of overall relevance $c_i = r_i / \sum_{i'} r_{i'}$. Under EA, individual misallocation M_i can be written as

$$\begin{aligned} M_i^{\text{EA}} &= c_i \sum_{i'=0}^{n-1} F(\sigma^{-1}(i')) - F(\sigma^{-1}(i)) \\ &= \sum_{k=0}^{n-1} p_s(\sigma(k)) \left[c_i(k+1) - \Pr(\sigma^{-1}(i) \leq k) \right], \end{aligned}$$

where $p_s(\sigma(k))$ denotes the probability of a user stopping browsing at position k, and $F(k)$ is the resulting probability of a visit.[4] Moreover, recall that DIPS can be expressed at the individual level as

$$\begin{aligned} M_i^{\text{DIPS}} &= \sum_{k=0}^{n-1} F(k) d_U(i, \sigma(k)) \\ &= \sum_{k=0}^{n-1} p_s(\sigma(k)) \sum_{k'=0}^{k} d_U(i, \sigma(k')) \end{aligned}$$

These formulas show that EA and DIPS can both be expressed as a weighted sum, over stopping probabilities $p_s(\sigma(k))$, of two quantities that are directly related. DIPS counts the number of UDPs for the item i up to rank k, while EA computes the (negative) probability $\Pr(\sigma^{-1}(i) \leq k)$ that item i is among the top k. Clearly, the probability of being in the top ranks tends to decrease with the number of UDPs. For this reason, we expect DIPS and exposure-based measures to behave similarly. At the same time, these measures operationalize different

[4] Under cascade (sequential) browsing models, the probability of receiving a visit at rank k is equal to the sum of the probability of stopping at any rank greater than or equal to k [116]: $F(k) = \sum_{k'=k}^{n-1} p_s(\sigma(k'))$.

constructs; hence, we expect them to capture different aspects of a ranking. For example, a ranking can assign to an item i its ideal exposure quota ($M_i^{EA} = 0$), while reserving the most visible positions for items of lesser relevance, thus causing substantial dissatisfaction in i ($M_i^{DIPS} \gg 0$), induced by highly visible UDPs.

6.1.4 Datasets

We exploit synthetic and real-world data to study the behaviour of DIPS and analyze its relation with exposure-based fairness measures.

Synthetic data

To compare fair ranking measures in a principled fashion, we build a synthetic dataset with full control on group representation, merit, and ranking policies. We consider a controlled setting with a binary sensitive attribute $s \in \{A, B\}$, where groups have equal representation over a total of $N = 1,000$ items, and sizeable differences in relevance scores. More in detail, we set $N_A = N_B = 500$, and draw relevance scores from group-specific, uniform distributions $f_A(r_i) = \text{unif}(0.5, 1)$ and $f_B(r_i) = \text{unif}(0.2, 0.7)$. In other words, all items of high relevance ($0.7 < r_i \leq 1$) belong to group A, items of intermediate relevance ($0.5 \leq r_i \leq 0.7$) belong to both groups with the same probability, and items of low relevance ($0 \leq r_i < 0.5$) are entirely from group B. The distribution of relevance scores between groups is depicted in panel (1) of Figure 6.1a. We choose the browsing model underlying rank biased precision [574] for DIPS, EA, and EE, modeling a top-heavy probability of visit with exponential decay: $F(k) = \gamma^k, \ \gamma = 0.9$.

Real-world data

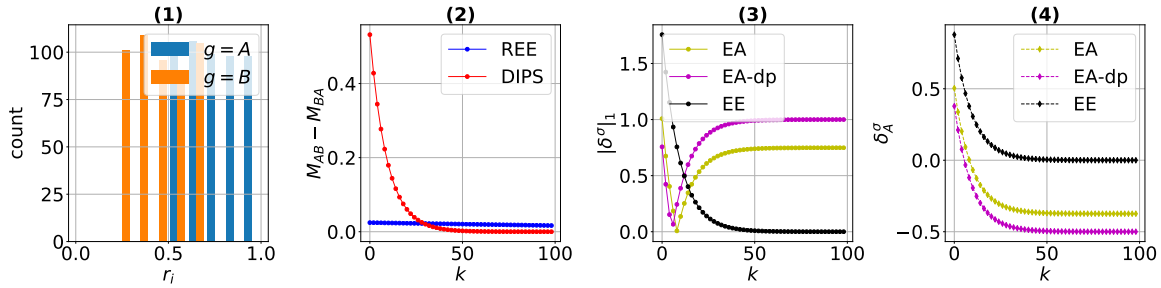
We complement our discussion on the similarities and differences between pairwise and exposure-based fairness measures by experimenting with a real-world dataset and a popular manipulation mechanism in fair ranking. We use the Entrepreneurs dataset [300], which consists of a list of US startup founders who received Series A funding over the period 2016-2021, obtained from Crunchbase.[5] The curators summed the Series A funds of founders who were part of multiple funding rounds during this time frame. They collected the name, image, and self-identified binary gender of the founders. Entrepreneurs are ranked by inflation adjusted funding, which is considered the merit parameter r_i, reported in panel (1) of Figure 6.1c. The sensitive attribute is binary gender, with a representation ratio of 9:1 in favor of

[5] https://crunchbase.com/

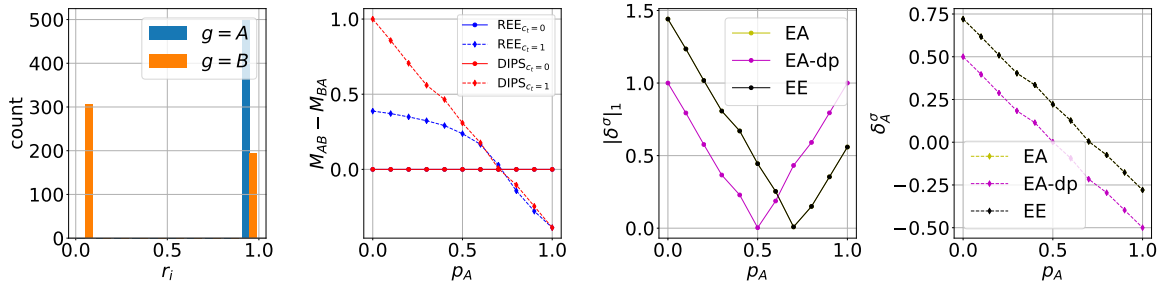
men (group A). Notice that the y axis is broken, to highlight the prevalence of low-relevance items from group A while favouring readability at higher values of r_i.

6.1.5 Experiments

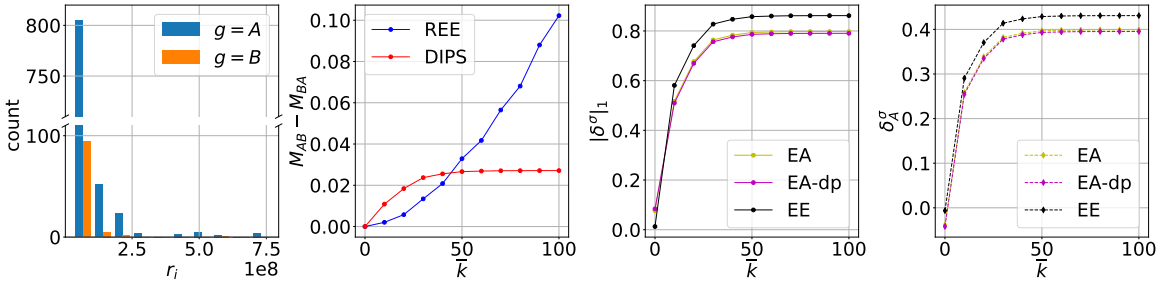
DIPS is a measure of pairwise fairness, yet it is grounded in browsing models like exposure-based fairness. In this section, we study the similarities and differences between DIPS, pairwise measures, and exposure-based measures empirically, by analyzing the behaviour of REE, EA, EA-dp, EE, and DIPS on synthetic and real-world datasets.



(a) Synthetic data: systematic group advantage.



(b) Synthetic data: relevance ties.



(c) Real-world data: fairness intervention to ensure minimum representation.

Fig. 6.1 Distribution of relevance r_i (1) and comparison of pairwise fairness measures REE and DIPS (2) with exposure-based measures EE, EA, EA-dp: $|\delta^\sigma|_1$ (3) and δ^σ_A (4).

Systematic group advantage

Setup. In this experiment on synthetic data, we consider a promotion mechanism that advances the 20 most relevant items from group B. This mechanism mimics an intervention to favor group B in a non-purely meritocratic way, where merit is summarized by relevance r_i. This advantage may be the result of (1) a fairness intervention aimed at giving more visibility to this group [75, 722], or (2) a profit-driven intervention by the platform owner, e.g., aimed at increasing visibility for a group of highly profitable items [402]. In either case, such an intervention can cause dissatisfaction for the remaining items.

We vary the destination rank k for the promoted items in the top position, allowing k to take values in $(0,99)$. For example, setting $k=0$, we promote the 20 most relevant items from $g=B$ to the ranks $\{0,1,\ldots,19\}$, while the relative positions of the remaining items remain unchanged, that is, their rank increases according to $\sigma^{-1}(i)=\sigma_*^{-1}(i)+20$.

Results. The results of this experiment are reported in Figure 6.1a. The values of $M_{AB}-M_{BA}$ (Equation 6.2) for REE and DIPS are shown in panel (2). Notice that no promotion takes place to the advantage of $s=A$, hence $M_{BA}=0$. DIPS is very sensitive to the promotion rank for items in group B, exhibiting an exponential decay, while REE is mostly flat. Moreover, the value $M_{AB}^{\text{DIPS}}>0.5$ for $k=0$ captures a strong dissatisfaction, while $M_{AB}^{\text{REE}}\ll 0.1$ is much smaller in comparison.

The remaining panels concentrate on three exposure-based measures (EE, EA, EA-dp). Panel (3) of Figure 6.1a reports the aggregate measure $|\delta^\sigma|_1$, i.e., the ℓ_1 norm of the misallocation vector in Equation (6.2), while Panel (4) reports the groupwise measure δ_A^σ for group A, i.e., the first component of Equation (6.2). The groupwise misallocation in panel (4) clearly shows a monotonic trend with exponential decay, as expected from the browsing model $F(k)$. It is worth recalling that positive values indicate underexposure for group A. Promoting items from group B to the most visible positions reduces the exposure E_A available for group A, and therefore δ_A^σ increases as items from group B are promoted to better positions, corresponding to lower values of k on the x axis. It should be noted that the aggregate measure ($|\delta^\sigma|_1$) in panel (3) derives directly from the groupwise measure δ_A^σ in panel (4). In the binary case considered in this example, it is equal to twice its absolute value, since $|\delta^\sigma|_1=2\cdot\text{abs}(\delta_A^\sigma)$.

Interpretation. The large value $M_{AB}^{\text{DIPS}}>0.5$ for $k=0$ captures the strong dissatisfaction that is likely to arise in group A if many items in another group were unjustly promoted to the top ranks. These highly visible promotions are unjust from the perspective of $s=A$ in the sense that they do not reflect the merit values encoded in r_i and σ_*. A large value for M_{AB}^{DIPS} adequately summarizes a situation where the items in group A are highly dissatisfied,

as the promoted items form visible UDPs with most of the items in group A. The same is not true for $M_{AB}^{\text{REE}} \ll 0.1$, suggesting that, under the (implicit) normative reasoning of REE, the dissatisfaction of group A would be very far from its theoretical maximum.

Turning to exposure-based measures, the disaggregated measures δ_A^{σ}, depicted in panel (4), are all equal up to a constant, which depends on the differences in the underlying normative reasoning presented in Section 6.1.3. Furthermore, these measures have the same profile as DIPS in the left panel. As discussed in Section 6.1.3, UDPs (in the absence of FDPs) directly result in missed exposure and higher values of EA, EA-dp, and EE. Since the same top-heavy browsing model $F(k)$ is assumed across these measures, they end up having a similar profile with exponential decay. This fact suggests that, if items producers have a notion of merit r_i, any intervention that assigns exposure to a group beyond its merit, as encoded by r_i, may generate a proportional amount of dissatisfaction in the remaining groups.

Relevance ties

Setup. Relevance ties are common in ranking problems and datasets [346, 348, 550, 887]. To study the behavior of DIPS and related measures in the presence of ties, we round the relevance scores in the synthetic data to the nearest integer, leaving us with binary values $r_i^q = \text{round}(r_i)$, depicted in panel (1) of Figure 6.1b. We consider the rankings of maximum utility $\sigma = \text{argsort}(r_i^q)$, where we vary the policy for tie-breaking. At each position of the ranking σ, each policy places the item of maximum relevance among those that are not already placed in better positions; if items of the same relevance are available from both groups, we draw the best available item from $s = A$ with probability $p_A \in \{0, 0.1, \ldots, 1\}$, or from $s = B$ with probability $p_B = 1 - p_A$. We consider a tie-aware and a tie-indifferent variant of REE and DIPS, obtained by selecting $c_t = 1$ and $c_t = 0$, respectively, in Equation (6.7).

Results. Figure 6.1b shows the values for each measure, averaged over 100 repetitions. Panel (2) shows both versions of REE and DIPS. As expected, the tie-indifferent variant of both measures is flat at zero. In fact, $\sigma = \text{argsort}(r_i^q)$ is a meritocratic ranking; therefore, no proper UDPs are present. For $c_t = 1$, both DIPS and REE span a wide range of values, capturing the sizeable cross-group dissatisfaction that is likely to arise in this setting with ranking policies that systematically favor one group over another in the case of ties.

EE, EA, and EA-dp are depicted in panels (3)-(4), as aggregate ($|\delta_g|_1$) and individual component (δ_A^{σ}), respectively. Interestingly, the difference between EE and EA becomes negligible; hence, they are indistinguishable in the plots. This is expected in situations such

as this, with binary relevance judgements and enough relevant items ($r_i^q = 1$) to make the exposure available for irrelevant items ($r_i^q = 0$), under a meritocratic ranking, negligible, according to the browsing model $F(k)$. Note that EA-dp is optimized when $p_A = 0.5$, i.e., when both groups get the same exposure in expectation. EE and EA are optimized when the tie breaking mechanism draws more often from $s = A$ ($p_A \simeq 0.7$), as this is the prevalence of group A among the relevant items. Overall, EA and EE have the same profile and zero-crossing as the tie-aware version of DIPS.

Interpretation. Measures of pairwise fairness can aptly model dissatisfaction in contexts where relevance ties are present, a situation that is fairly common in ranking problems. This is achieved by extending the concept of UDP to account for relevance ties. If instead we stick to the regular definition of UDP, any systematic advantage for one group will go unnoticed, as testified by the (constant and null) values of REE and DIPS instantiated with $c_t = 0$. Furthermore, this experiment confirms a close connection between DIPS and exposure-based measures.

Enforcing minimum representation

Setup. We deploy the fairness intervention of Zehlike et al. [858], imposing a minimum representation for women in the Entrepreneurs dataset. More specifically, we require a minimum percentage $p_{\min} = 0.5$ of women in every prefix of the final ranking, up to a given ranking $\bar{k}$. As a motivating example for such an intervention, consider a trade magazine that compiles a chart of successful entrepreneurs with attention to gender representation. The goals of relevance and gender representation can be achieved with a ranking σ that is aware of the raised funding while featuring a minimum percentage of women in every prefix up to a given rank $\bar{k}$. Low values of $\bar{k}$ correspond to mild gender parity requirements, only enforced at the top positions of the ranking (up to $\bar{k}$). High values of $\bar{k}$, on the other hand, correspond to more strict requirements, where minimum representation must also be maintained further down the ranking. In this experiment, we vary $\bar{k} \in \{0, \ldots, 100\}$.

Results. The results for this experiment are reported in Figure 6.1c. As usual, panel (2) focuses on REE and DIPS. The latter increases sharply for small values ($\bar{k} < 20$), where an increased representation corresponds to highly visible UDPs under a top-heavy browsing model. Around rank $\bar{k} = 40$ DIPS becomes flat, as these ranks have low visibility. REE also increases with $\bar{k}$, but, unlike DIPS, the increase accelerates with $\bar{k}$. This is due to the fact that, to satisfy the minimum representation requirement, the number of UPDs increases superlinearly with $\bar{k}$.

EE, EA, and EA-dp are represented in panels (3)-(4). The groupwise measure displays a concave profile, similar to DIPS, since promotions after rank $k = 40$ have a negligible impact on exposure. As usual, EE is minimized by the null manipulation $\bar{k} = 0$; EA and EA-dp are very close to it as, in this particular setting, women entrepreneurs have a low overall representation ($T_B^{\text{EA-dp}} = N_B/N \simeq 0.1$) and, subsequently, a low share of the overall relevance ($T_B^{\text{EA}} = \sum_{i \in B} r_i / \sum_i r_i \simeq 0.1$). The sizeable values of EE, for $\bar{k} \geq 40$, suggest that group B (women) gains a significant exposure from this intervention, clearly at the expense of group A (men).[6]

DIPS, on the other hand, has low values $|M_{AB}^{\text{DIPS}} - M_{BA}^{\text{DIPS}}| \ll 0.1$. This is due to the fact that the female entrepreneurs occupying these highly visible positions in the final ranking σ have greater relevance (r_i) than most of the other entrepreneurs. In other words, despite a substantial visibility gain for female entrepreneurs, the most visible positions occupied by them do not represent a UDP for most male entrepreneurs. For example, when $\bar{k} \geq 20$, among the 20 most visible positions, accounting for more than 80% of the overall exposure, we find ten female entrepreneurs who are in the top decile for raised funding, i.e., their relative merit with respect to most other candidates is unquestionable. This follows from the fact that the fairness manipulation employed is aware of relevance, so female entrepreneurs with higher r_i are promoted first. Different ranking policies, naïvely enforcing representation without paying attention to relevance, would yield high values of DIPS.

Interpretation. On the one hand, this experiment shows that, when DIPS and exposure-based measures are instantiated with the same top-heavy browsing model $F(k)$, they are similarly influenced by fairness interventions toward the top of a ranking, while ignoring swaps at less visible positions; they display similar "saturated" profiles as a result. On the other hand, the absolute values of these measures can differ substantially. In essence, exposure-based measures are based on a comparison between *groupwise* merit and groupwise representation among the most visible items in the final ranking. Although DIPS is focused similarly on the most visible items, it takes into account their *individual* merits. For instance, an item whose relevance is in the highest decile can be promoted to the most visibile position, i.e, with a sizeable impact on exposure, without increasing the dissatisfaction counter of most items, i.e., with a small impact on the aggregate DIPS measure. Overall, while showing some clear similarities, DIPS and exposure-based measures operationalize different constructs and capture different properties. Overall, our analyses show that fairness-enhancing interventions in ranking may cause dissatisfaction for non-protected groups, but merit-based policies will mitigate this downside.

[6]Recall that δ_A^σ is a normalized quantity, i.e., $0 \leq \delta_A^\sigma \leq 1$

6.1.6 Discussion

In this section, we have targeted an important gap concerning measures of pairwise fairness in ranking. First, we have shown that the construct and normative reasoning behind the most popular measures of pairwise fairness, i.e., Inter-Group Inaccuracy (IGI) [67] and Rank Equality Error (REE) [471], are not discussed in the literature, hindering a clear understanding and informed adoption of these measures. To address this problem, we have retrospectively mapped these measures to the item-centric construct of producer *dissatisfaction* induced by a non-meritocratic ranking.

Second, we have highlighted some clear limitations of REE and IGI, proposing a new measure called *Dissatisfaction Induced by Pairwise Swaps (DIPS)* to overcome them. DIPS can suitably operationalize a perception of injustice by ranked items for being positioned below less worthy items from different groups, thus quantifying cross-group dissatisfaction, perceived quality of IAS, and the likelihood of withdrawal on part of producers.

Third, we have studied the relationship between DIPS, pairwise measures, and exposure-based measures such as Equity of Attention and Expected Exposure. We have shown that pairwise fairness can be grounded in browsing models, highlighting the similarities between exposure-based measures and DIPS, which can be instantiated with the same top-heavy model, while also stressing their differences, which follow from operationalizing fundamentally different constructs.

Overall, this section improves measures of pairwise fairness, situates them more precisely in the broader context of fair ranking, and contributes to the debate on the normative reasoning behind algorithmic fairness. As a result, we hope to incentivize a more informed adoption of pairwise fairness on part of practitioners and researchers and, more broadly, of fair ranking measures and approaches.

6.2 Gender Stereotype Reinforcement

As a further contribution to the fair ranking literature, in this section, we focus on subject fairness, and develop the first measure of gender stereotype reinforcement in SE. From the outset, we stress the division between the construct we aim to capture and its operationalization. In general, stereotypes can be modeled as associative networks of concepts [699]. They often arise from co-occurrence of features [484], such as membership in a group and the display of certain traits and roles, which become linked in a Bayesian fashion based on culture and direct observation [365]. In this regard, women and men are particularly salient categories, recognizable since an early age, and available for stereotypical association with traits, behaviors, and events [533]. In turn, even when outspokenly rejected, gender

stereotypes influence the lives of women and men both descriptively and prescriptively, shaping the qualities, priorities, and needs that members of each gender are expected to possess [234]. During their lives, individuals are frequently exposed to information about gender, through direct experience and indirect information derived from social interactions and cultural representations [224], often portrayed by the media.

Cultivation theory [295], historically focused on television, posits that increasing exposure to a medium and its contents leads to a progressive alignment to the beliefs, culture, and reality depicted in the televised world. Within this framework, the way women are depicted on primetime television has been studied; recent analysis highlights persistent representational stereotypes related to physical appearance and warmth [726], confirmed by public opinion [224, 757]. According to cultivation theory, heavy viewers are likely to be influenced in their perception of the real world, due to the availability heuristic [719]: in judging frequency and normality (e.g. of women being affectionate), they resort to the examples that come to their mind, the media being a potential source of information to recall. The availability heuristic has been proposed and verified as a general shortcut in human cognitive processes [772], and recently studied as a bias that arises while exploring SE result pages [605].

Inevitably, SE influence users, helping them to link topics, concepts, and people as they read, browse, and acquire knowledge. Therefore, they can play an important role in countering or reinforcing stereotypes. For example, Google image search results were found to reflect current gender differences in occupation, with a tendency to slight exaggeration [433]; at the time of the study, searching images of a job with a female-to-male ratio of 1:4 in the employed population, such as software engineer, would yield pictures depicting women in less than 20% of the results. Moreover, manipulation of female-to-male representation in job search results, artificially increasing the presence of one gender in images, significantly affected people's perception of gender ratios in that occupation [433]. A study on Bing photos found a higher frequency of women in depictions of warm traits (e.g. sensitive), while men are more common in searches for competence traits (e.g., intelligent) [617]. These results highlight the importance of measuring and counteracting bias in SE, as recently pointed out by critical race and gender studies scholarship [599].

Gender stereotypes held by people are commonly measured in two ways: directly, on the basis of an individual agreeing with statements about gender and specific traits [224]; indirectly, via Implicit Association Tests (IAT) between mental representations of objects [325] or assessment of attitude through priming [257]. Indirect tests allow for an unobtrusive assessment of attitudes towards groups (determined e.g., by gender and ethnicity) and can measure association of categories, such as `women`, with words from a specific domain, such as `family`, even when subconscious. Large text corpora sourced from the web, such as

Wikipedia, have been found to echo some of the above biases: as an example, Wikipedia entries related to women are more likely to mention marriage- and sex-related contents and events [322]. Interestingly, Word Embeddings can be used to detect gender-related biases in the corpus on which they have been trained [196, 288].

Word Embeddings (WE) are vectorial representations of words computed automatically using different supervised and unsupervised machine learning approaches [414, 566]. Most frequently, they are learned from large text corpora available online (such as Wikipedia, Google News, and Common Crawl, capturing semantic relationships of words based on their usage. Recent work [110] shows that WE retain the stereotypical associations from their training corpora, encoding a full spectrum of biases from the IAT, including gender-related ones about career and family, science, and arts. Additional problematic depictions of men and women have been identified in these WE, including sexist analogies (such as $\texttt{woman} - \texttt{man} \simeq \texttt{midwife} - \texttt{doctor} \simeq \texttt{whore} - \texttt{coward}$ [84]) and representation of jobs skewed with respect to gender, in ways that reflect current gender gaps in the US workforce [194, 288, 648]. For this reason, WE have been proposed as an unobtrusive measurement tool of the average bias of the many contributors to these corpora and, generalizing, from the society in which they live [288] or the language they speak [196]. Based on co-occurrence with intrinsically gendered terms within the text corpora (such as `woman` and `man`), a *genderedness* score can be derived for each word in the embedding space. Among the words with a high score, some are duly gendered (`hers`, `his`), while others reflect an accidental status quo aligned with stereotype (`hygienist`, `electrician` - stereotypically female and male, respectively). This bias, undesirable when WE are part of a socio-technical system, is an interesting property we can leverage to measure gender stereotypes in SE.

In this section, we leverage gender bias encoded in WE to detect and quantify the extent to which a SE responds to stereotypically gendered queries with documents containing stereotypical language of the same polarity. We propose the *Gender Stereotype Reinforcement* (GSR) measure that is specifically tailored to quantify the tendency of SE results to support gender stereotypes. Contributions of this section include (1) the GSR measure tailored for SE and its evaluation within the *construct validity* framework; (2) an audit, in terms of GSR, of several widely-known and used ranking algorithms; and (3) the estimation of the impact of different WE debiasing approaches, both on ranking effectiveness and countering gender bias.

6.2.1 Background and Related Work

WE and neural models in Information Retrieval

Word2Vec [566] was the first widely used WE model. Word2Vec can learn similar representations for terms used in similar contexts in the training data, typically corpora of millions of documents in natural language. In addition, as the embedded word representations learned with Word2Vec reflect the usage distribution of respective terms, they have been used as a proxy for the semantic similarity of terms in many NLP applications. The popularity of Word2Vec also paved the way for other machine learning approaches to obtain embedded word representations such as GloVe [631] and FastText [414]. WE models were soon adopted in the Information Retrieval (IR) domain, promoting the exploration of deep learning approaches for document retrieval [573]. Hereafter, we let Information Retrieval indicate the field concerned with building the infrastructure and algorithms required to retrieve relevant items in response to user queries, of which web SE are the most visible application.

Lexical approaches such as `tf-idf` [692], `QLM` [862] and `BM25` [675] were the first and most popular techniques adopted for document retrieval. Nevertheless, these retrieval models do not take into account terms which are not contained in the user query nor their semantics. For this reason, the potential offered by embedded word representations – that is, the possibility to represent the meaning of a term and compare it with others in a measurable way – was soon put to use by newly proposed retrieval models.

The simplest WE-based document retrieval approach in our experiments is named `w2v-add` [802]. In this case, we compute a query and a document representation averaging the WE of the terms they contain, and then rank documents according to their cosine similarity to the query vector. This approach, however, reduces the query/document representation problem to the core. For example, it does not take into account the relative importance of each term. `w2v-si` solves this problem: queries and documents representations are obtained by computing a weighted average of the word vectors of their terms, and then the documents are ranked as in the previous case. Each term weight corresponds to its self-information (si) which is a term specificity measure similar to IDF [177].

Among the first most successful deep learning models for IR, there is Deep Relevance Matching Model (`DRMM` - [333]). `DRMM` uses embedded representation of words to compute the similarity between every pair of terms in a user query and each document in a ranked list. Another paradigmatic approach in the neural IR field is MatchPyramid (`MP` - [619]). This approach, originally proposed as a document classification model, was also successfully applied to the ranked task. For our study, we select these two approaches, which are popular

in IR and easy to use. Moreover, they allow us to evaluate the impact of diverse WEs in different neural IR architectures.

Gender stereotype in WE and SE

A convincing body of research shows that WE learned on large corpora of text available online encode cultural aspects, some of which undesirable. Among them, worth noting at the core of this work are gender-related biases which comprise: sexist analogies [84], stereotypical association of gender with science and arts [110], representation of occupations correlated with differences in female and male employment [194, 288, 648], gender roles in career and within the family [110]. Communion (also called warmth) and agency are two other dimensions that are consistently associated with gender [224]; in line with this stereotype, we find that warm traits (e.g. "emotional") have female polarity in the embedding space, while agentic traits (e.g. "aggressive") are more commonly associated with men (Section 6.2.2). A wealth of studies in the literature on psychology and labor economics confirms the presence of the aforementioned biases in society [129, 186, 224, 269, 361, 602, 603], making their presence in WE particularly interesting.

Several of these biases, found in SE, potentially reinforce gender stereotypes through powerful and widespread search tools available to the public. Kay et al. [433] show that gender bias in image search results is exaggerated: the gender distribution for Google image results on jobs is correlated with and amplifies differences in female and male employment. Bing images associate agentic traits with men and warm traits with women [617]. Monster and CareerBuilder were audited, displaying group unfairness against female candidates in 1/3 of the job titles surveyed [141]. This does not imply that these SEs are likely to have the same biased WE as part of their algorithmic machinery. Rather, finding that known gender biases in SE are also encoded in vectorial representations of words suggests that WE can be used as a tool to measure gender bias in SE.

In this respect, Bolukbasi et al. [84] find that gender-related information for each word in the embedding space is mostly confined within a single dimension:

1. They propose ten word pairs to define gender: `she-he`, `her-his`, `woman-man`, `Mary-John`, `herself-himself`, `daughter-son`, `mother-father`, `gal-guy`, `girl-boy`, `female-male`.

2. For each pair, they compute the difference between the two word vectors, obtaining ten candidate vectors (dimensions) to encode gender.

3. They stack the ten vectors into a single matrix, on which they perform a principal component analysis, finding 60% of the variance explained by the first principal

component, subsequently treated as the gender subspace w_g. We dub the *genderedness* score of a word w, its scalar projection along the gender subspace

$$g(w) = \frac{w \cdot w_g}{|w||w_g|}, \tag{6.16}$$

and use it as a building block to operationalize GSR.

The sign and magnitude of $g(w)$ determine the polarity and strength of gender-association for word w - e.g. $g(\texttt{sister}) = 0.31$, $g(\texttt{brother}) = -0.22$. After identifying a gender subspace (or direction w_g), Bolukbasi et al. [84] remove gender-related information from most words via orthogonal projection. Only intrinsically gendered word pairs (such as `she`, `he`) retain a nonzero component in the gender subspace. Prost et al. [648] propose a *strong* variant of this approach where the procedure applies to the whole vocabulary. This family of debiasing techniques seem limited and imperfect [314], with gender information redundantly encoded along multiple dimensions and, thus, hard to eradicate. The confirmation of this statement is given in the context of SE and gender stereotype in Section 6.2.4, where we evaluate the impact of regular and strong debiasing with respect to the performance and GSR of IR models based on WE.

Fairness and diversity in IR

As mentioned in the introduction of this chapter, fair ranking is a fundamental aim of IAS and, more specifically, IR systems. In this space, Gerritse [296] study the impact of debiasing WE [84] on query reformulation algorithms based on Word2Vec embeddings. This work is closest to our evaluation of the effects of debiasing in Section 6.2.4, where we perform a complementary analysis on different IR algorithms that are purely based on WE. Metrics and approaches from fair ranking can also be employed to measure and favor diversified topical coverage [286], where political leaning or sentiment take on the role of a protected attribute that should have reasonable diversification across search results. This flavor of fairness overlaps with diversity and novelty research from the IR community [115, 160, 851].

In some areas, such as political search in social media, it is interesting to evaluate how diversity and bias in search results can be influenced by (implicit) bias in queries. For instance, Kulshrestha et al. [472] find that Twitter's response to queries about US political candidates tends to give better ranking to tweets from sources with the same political leaning as the candidate. Although different in methods and objective, our work is conceptually similar as we are interested in evaluating how a construct measured on queries (stereotypical genderedness) relates to the same construct measured on search results.

Construct validity and reliability

Construct validity, in its modern connotation, is a unified view on the desired properties of a measure aimed at quantifying a given construct that enables an overall judgment about adequacy and appropriateness based on empirical evidence and theoretical arguments [557]. Embedded in this definition is a clear distinction between, on the one hand, the unobservable theoretical attribute we are trying to evaluate (the *construct*, e.g. "teacher quality"), with its context and underlying theme and, on the other, the way the construct becomes operational through a measurement model (the *operationalization*).

We follow Jacobs and Wallach [397], who describe seven components of construct validity, which we summarize below:

1. *Face validity*. How plausible does the measurement model look compared to the construct? Answers to this question are highly subjective and little more than a preliminary step.

2. *Content validity*. Is there a coherent understanding of the theoretical construct? Is the selected operationalization in accordance with it?

3. *Convergent validity*. Does our measurement agree with other measurements of the same construct?

4. *Discriminant validity*. What else is the measurement capturing? Are there other constructs which are justifiably or unexpectedly correlated with the proposed measurement?

5. *Predictive validity*. Are any other properties likely to be influenced by our construct? Is our operationalization of the construct related to those properties as expected?

6. *Hypothesis validity*. Are the construct and its operationalization meaningful and useful, so that they can be used to test hypotheses and raise new questions?

7. *Consequential validity*. Should our measure be used? In what context can it be employed and what would the consequences be?

The next section describes in detail Gender Stereotype Reinforcement (GSR) as a construct, referring to the supporting literature from social psychology, which deals with the common understanding of GSR and its *content validity* as a construct. Taking into account the key properties of GSR, Section 6.2.2 also discusses its *discriminant validity*, tied to domain-specificity of language, along with its *convergent validity* in a wider context of fairness metrics. *Consequential validity* and *hypothesis validity* are related to current limitations

and future work, discussed in Section 6.2.5. In the absence of a user study, *predictive validity* cannot be properly discussed. Within the context of gender stereotypes in SE, the only user study that we are aware of focuses on image retrieval [433], while our proposed measure deals with textual data. Due to its subjective nature, we do not specifically address *face validity*.

Finally, we also discuss GSR *reliability*, a more familiar concept to computer scientists. This dimension of measurement theory summarizes how robust, repeatable, and reliable a measure is. In summary, reliability depends on the stability of the measured quantity, the precision of measurement tools, and process noise; Section 6.2.4 is devoted to this property.

6.2.2 Proposed Measure

We articulate our approach, untangling the definition of a *construct*, i.e. the phenomenon we want to study, from its subsequent *operationalization*, which details how the phenomenon can be measured from observed data [397, 558].

Construct

Our aim is to quantify to what extent a SE can reinforce gender stereotypes in users. We call this construct *Gender Stereotype Reinforcement* (GSR), resorting to supporting concepts from the psychology literature before giving a formal definition. This incremental process is important to establish the *content validity* of GSR as a construct.

Definition 1. Stereotype.
Stereotypes are beliefs about groups of individuals with a common trait, widely held by a population of interest. Their appearance is likely influenced by the strength of an observational link, i.e. how often one position along a dimension (such as gender) co-occurs with another (such as warmth) [484].

Stereotypical associations picked up by individuals can be attributed to culture and socialization [365]. Bayesian principles are believed to be at play in the acquisition of culture, which is often screened and mediated by search technology, whose trustworthiness is generally taken for granted [337]. In other words, our cognition is receptive to repeated co-occurrence of topics and entities. It may therefore end up forming links between them, also thanks to the media and technology we interact with on a daily basis.

Definition 2. Gender Stereotype.

A gender stereotype is a generalized view or preconception about attributes or characteristics, or the roles that are or ought to be possessed by, or performed by, women and men.[7]

Stereotypes about gender have been studied in a variety of contexts, including school [186], workplace [81], parenthood [185] and search for romantic partners [623], with respect to several aspects such as depiction, perception (of self and others) and outcomes. Common themes have been identified through decades of scholarship, including *agency* and propensity to science, *communion* and importance of appearance [234].

As a well-researched example, historical meta-analysis over seven decades confirms agency and communion as consistently and increasingly salient in US opinion polls about gender differences [224]. Agency, perceived as predominantly male, refers to the drive for achievement, while communion is related to caring for others and is increasingly associated with women.

Definition 3. Direct gender stereotype.
Association of a stereotypically gendered concept with people of the respective gender.

This applies to any sentence where preconceptions about one gender are directly associated to a member of that gender, mentioned through a noun (`man`), adjective (`his`), pronoun (`he`) or name (`John`).
Example: *She is affectionate*.

Definition 4. Indirect gender stereotype.
The link of a stereotypically gendered concept with another stereotypically gendered concept, commonly associated with the same gender.

This definition is based on a view of culture, social constructs, and stereotypes as networks of concepts [302, 626] and implicit associations [63, 257, 325]. Co-occurrence of stereotypical characteristics and traits, commonly associated to one gender, may reinforce a link in a network of stereotypes about women and men. To exemplify, we argue that beliefs about stereotypically female (male) jobs are likely to fall on women (men). Research in social cognition and political science highlights that networks of stereotypes associated with protected attributes, such as gender and ethnicity, can play a role in a person's perception, without them being aware of it [63]. This may happen to a person, even if they sincerely dislike that stereotype [257].
Example: *The nurse is affectionate*

Characterization of the GSR construct. Given the above terminology, we characterize GSR in the context of IR as the SE's tendency to respond to stereotypically gendered queries with

[7] https://www.ohchr.org/en/issues/women/wrgs/pages/genderstereotypes.aspx

documents containing stereotypical language with the same polarity. We defer a thorough definition, complete with mathematical formalization, to Definition 8.
In societal systems, GSR is measured by the agreement of human constituents with gender stereotype descriptors [640, 767]. In operationalizing this construct, our objective is to quantify the impact of SE on the perception of gender: more specifically, its alignment to existing direct and indirect stereotypes encoded in culture and language. Search results may end up reinforcing gender stereotypes if, when responding to potentially stereotypical queries, their language is skewed along gendered lines with matching polarity.

Intuitively, the influence that people around us may exert can be regarded as the societal counterpart of documents and their language in the context of SE. An example is a SE which, responding to a query about nursing, displays documents with a strong representation of women (*direct* stereotype), or emphasis on attributes related to communion (*indirect* stereotype).

Operationalization

After defining our construct, we show how it can be made operational. This involves illustrating our assumptions and their interaction with the building blocks of our measurement model [397]. We begin by defining the basic concepts in the context of search.

Definition 5. Ranked list.
The response of a SE to a query, i.e. a permutation σ of documents, decreasingly ordered by (estimated) relevance with respect to a given user query. A ranked list may not include all the items available in $\mathscr{I}$, in which case it is represented as a length-K prefix of the whole permutation vector, denoted $\sigma(1:K)$.

Definition 6. Search history.
A set of (query, ranked list) pairs representing the interactions of one or more users with a SE.

Stereotype formation may be conceptualized as an acquisition of culture and associations of a particular kind, taking place through *repeated* interaction. The response to a single query, though anecdotally interesting, is less informative than a set of responses to different queries. Hence, we refer to a *search history*, a somewhat overloaded expression, which potentially encompasses every past user interaction with a SE, including the pages they visited along with very detailed logs of click behavior, browsing, and permanence.

Our usage of the expression is different in two ways. (1) It applies to any subset of user interactions with an SE, including, for instance, only recent ones. We do not require a

symbol	meaning
w	a word
$g(w)$	genderedness of word w
$g(q_j)$	genderedness of query q_j
$g_{q_j}(i)$	genderedness of document i retrieved for q_j
$g_{q_j}(\sigma_j)$	genderedness of ranked list σ_j retrieved for q_j
μ_q	average genderedness of queries from $\mathscr{Q}$
$\mathrm{std}_{g(q)}$	standard deviation in genderedness of queries
$\mu_{q,\sigma}$	average genderedness of ranked lists of documents
$m_s(\mathscr{Q},\mathscr{I})$	GSR for system s on collection $(\mathscr{Q},\mathscr{I})$

Table 6.3 Notation for GSR.

complete list of queries issued and results shown. (2) The level of granularity and depth of logging entailed by our definition is minimal. This work aims to audit and model SEs rather than users. For this reason we do not require click logs, which are user-dependent and thus accidental with respect to our analysis. More in general, Definition 6 adapts to data coming from multiple users in a bundled and anonymized fashion, as well as data collected and curated by a practitioner. These differences are important to correctly assess the applicability, practicality, and ethics of our operationalization.

To summarize the following sections, we assume that a strong correlation between *genderedness* of queries and of the ranked document list in a search history reinforces gender stereotypes. In the following, we gradually introduce related quantities; the adopted notation is summarized in Table 6.3.

Measuring gender stereotype. Stereotypes about gender are plentiful and pervasive, likely due to the fact that the underlying categories (especially the classical female-male dichotomy) are available to our cognition from an early age on a daily basis. Preferential association of a concept or topic to men or women is measured by surveying a population of individuals. The study of gender-based associations thus depends on resources, time available, and research agendas.

Increasing evidence from the field of NLP shows that, among the powerful results and interesting properties of WEs, their geometry captures well-known stereotypes related to gender [84, 110, 203, 288, 621, 648]. Techniques have been proposed to isolate the genderedness of a word along a single direction [84]. Based on this approach, each word is associated with a "gender score" consisting of a signed scalar value. In a convention employed hereafter, a strongly positive (negative) score will be a proxy for a strong association to female (male) gender. The upper part of Figure 6.2 shows, as a simplified example, the

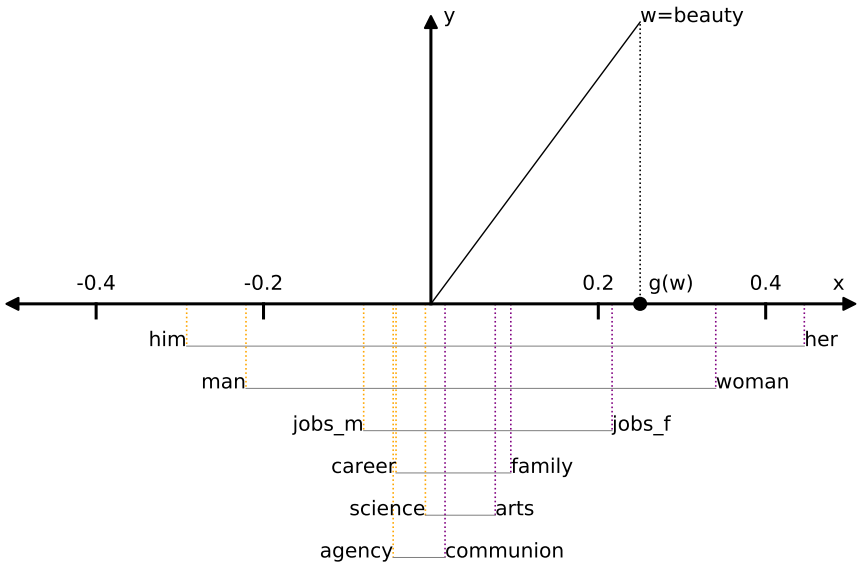


Fig. 6.2 Computation of genderedness for $w =$ `beauty`. Gender direction is on x axis, while y axis represents subspace orthogonal to gender. w displays a significant component along the gender direction, hence we consider it stereotypically female in the embedding space and the underlying text corpus. See, in Appendix D, Table D.1 for *agency* vs *communion* ($p = 2.2\text{x}10^{-2}$), Table D.2 for *science* vs *arts* ($p = 1.8\text{x}10^{-3}$), Table D.3 for *career* vs *family* ($p = 5.5\text{x}10^{-4}$), Table D.4 for *jobs_m* vs *jobs_f* ($p = 0$). P-values are computed with four one-tailed permutation tests on the genderedness of the words in each table.

projection of the word `beauty`, which is strongly positive and thus associated with the female gender.[8]

To validate genderedness, encoded by Equation (6.16), as a score of perceived masculinity/femininity, we test it against known gender stereotypes. Two commonly studied constructs in psychology literature are *agency* and *communion* [224, 361], alternatively dichotomized as *competence* and *warmth* [269]. Agency, stereotypically associated to men, is related to ability and drive to pursue one's goals, while displaying leadership and assertiveness. Communion, prevalent in female stereotypes, relates to a person's orientation towards others and their well-being, suggesting propensity for caring, nurturing, compassion, and emotion.

Attitude towards mathematics and science has been measured implicitly [186, 603] and explicitly [186]. Studies provide evidence of the cognitive link between math and male gender at an early age. This association is often studied in opposition to arts (and language) which are found to be predominantly associated with female gender [602, 603].

[8]For obvious reasons, a figure can only represent 2 out of the 300 dimensions in which the `w2v` embedding is encoded.

Career orientation, in opposition to family, is another dimension related to gender [602]. Career can also be broken down into sectors. Some professions have a very high male representation, while other work is carried out overwhelmingly by women [129].

In considering research on gender stereotypes, four opposing associations emerged, which are described above. We compute their genderedness as follows: for *agency* vs *communion* we summarize the genderedness of either construct with the average genderedness of adjectives in Table D.1, taken from [224]. With the same averaging procedure, we follow [602] for terms related to *science* vs *arts* (Table D.2) and [603] for *career* vs *family* (Table D.3). Finally, we sample the 20 most gendered single-word jobs of [129], shown in Table D.4, and perform the same computation, dubbing this comparison *jobs_m* vs *jobs_f*.

The results are summarized in the lower part of Figure 6.2, where we also report the projections of `woman`, `man`, `her`, `his` for comparison. All four stereotypes are confirmed, with male cluster projections (orange) falling to the left of their female counterparts (purple). According to one-tailed permutation tests, the dichotomy *agency* versus *communion* is the least gendered, significant at $p = 2.2\text{x}10^{-2}$. Interestingly, the strongest association with gender is *jobs_m* versus *jobs_f* ($p = 0$), originating from census data and representing occupations with extreme skew in gender distribution.

We conclude that projection along the gender subspace (although potentially noisy for single terms) is, on average, a suitable proxy for stereotypical association with gender.

Modeling stereotype in query-document pairs. Semantic memory is a specific aspect of human memory that holds general knowledge of concepts. It is regarded as a widely distributed neural network [626]. Associative network structures, often referred to as schemas, are commonly used in neuroscience as models that represent complex constructs that guide behavior [302]. This suggests that any acquisition of knowledge and culture is in part based on the formation of rich networks of concepts. The acquisition and articulation of stereotypes are not conceptually different: a recent line of work employs network analysis to study stereotypical associations as clusters and subclusters of concepts [699].

We are interested in modeling the potential association of concepts, with a tendency to cluster along a gendered dimension. Search technologies play an important role in helping users build links between concepts. When issuing a query, SE users are likely to be receptive to the formation of new links between concepts of their query and information found in ranked lists [433]. If a document i retrieved for a query q_j (e.g. `nurse`) contains terms most aligned with the genderedness of q_j (e.g. `care`, `woman`, `Mary`) it may end up reinforcing gender stereotype through an association of such concepts.

To evaluate the stereotypical gender agreement between query q_j and document i, we calculate their mean genderedness $g(q_j)$ and $g_{q_j}(i)$ as schematized in Figure 6.3. Both

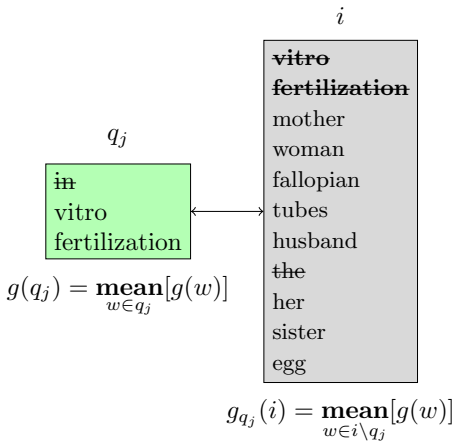


Fig. 6.3 Concepts from query associated with concepts from document along gender dimension. Before computing the average genderedness of query and documents, stop words are removed (struck through font), and query terms which explicitly appear in a document are neglected (bold and struck through).

queries and documents are represented as bag-of-words (stop words are removed). $g(q_j)$ is subsequently computed as the average genderedness of remaining query terms. For $g_{q_j}(i)$, query terms are removed (bold in Figure 6.3) before performing the same averaging procedure. Our goal is to model the alignment of query-document concepts along stereotypically gendered lines. When computing $g_{q_j}(i)$, we therefore neglect document terms which also appear in the query, to remove the spurious bias due to redundant self-linking. For this reason, $g_{q_j}(i)$ depends on the query, as illustrated by subscript q_j.

Computing the genderedness of a ranked list. As is generally known, users seldom dabble in result pages beyond the first, and the likelihood of view decreases with document rank [263, 401, 407]. Performance metrics in IR have taken this aspect into account, assigning more importance to top-ranked rather than low-ranked documents [401, 575]. A widely adopted evaluation measure, based on this user model, is the Discounted Cumulative Gain [401], which weighs documents according to a coefficient that decreases with rank in a logarithmic fashion. This weighing scheme is applied to measure the effectiveness of a ranked list based on the relevance and position of the documents within it. Our approach is identical, except for our focus on genderedness rather than relevance.

Figure 6.4 shows a ranked list σ_j of documents, retrieved for a query q_j. A weight vector $\overline{w}$ is computed with rank-based logarithmic discount and normalized. The genderedness of a ranked list $g_{q_j}(\sigma_j)$ is calculated as the weighted average of the genderedness of documents in σ_j with weight vector $\overline{w}$.

Definition 7. Genderedness of a ranked list.

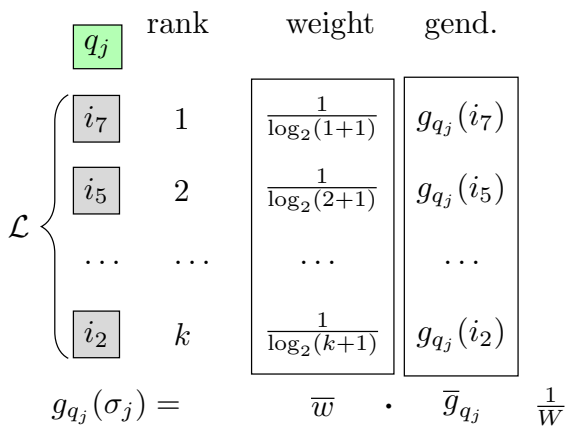


Fig. 6.4 Genderedness of ranked list is computed as a weighted average of the genderedness of each document retrieved, with weight computed according to a rank-based logarithmic discount. Note that in the calculation of $g_{q_j}(\sigma_j)$, without any loss of generality, we opt for a base-2 logarithm. W is a normalizing constant, i.e. the sum of elements in $\overline{w}$.

Let $\overline{w} = [w_1, \ldots w_K]$ be a weight vector such that K is the length of the ranked list, $W = \sum_{k=1}^{K} w_k$ and $w_k = 1/\log_2(k+1), k \in [1, K]$. Then, the genderedness of σ_j is defined as

$$g_{q_j}(\sigma_j) = \frac{1}{W} \sum_{k=1}^{K} w_k \cdot g_{q_j}(\sigma_j(k)), \tag{6.17}$$

where $\sigma_j(k)$ is the item at rank k in σ_j and $g_{q_j}(\sigma_j(k))$ is its genderedness.

As a toy example, which will be expanded and further discussed in Example 1, suppose that we have the following setting with a single-term query and two retrieved documents:
$q_j =$ `electrician`
$i_1 =$ `The man is an electrician.`
$i_2 =$ `The woman is an electrician.`
$\sigma_j = [i_1, i_2]$.
Then, according to Definition 7, the genderedness of ranked list σ_j is computed as follows:

$$g_{q_j}(\sigma_j) = \frac{1}{W}\left[\frac{1}{log(2)} g_{q_j}(i_1) + \frac{1}{log(3)} g_{q_j}(i_2)\right] = -3.8\text{x}10^{-3},$$

where i_2 is less important in this weighted average, being the last document in σ_j. Its genderednesss $g_{q_j}(i_2)$ is thus discounted accordingly, and the negative value of $g_{q_j}(i_1)$, albeit smaller in modulo than that of $g_{q_j}(i_2)$, prevails.

From ranked list to search history. Multiple search results constitute a search history which may reinforce gender stereotypes. If the language of documents in ranked lists (more specifically their genderedness) consistently agrees with that of user's queries, it is reasonable to assume that the search history supports concept clustering along a gender-stereotypical dimension.

More precisely, given a set of queries $\mathscr{Q}$ and a set of ranked lists (one per query) returned by a system s, we compute a linear fit between query genderedness $g(q_j)$ and ranked list genderedness $g_{q_j}(\sigma_j)$, considering it as a summary of the GSR carried out by s on $\mathscr{Q}$. Below is a summary of the steps to measure GSR:

- Genderedness of a word w is measured as its projection along the gender direction (Equation 6.16).
- Genderedness $g(q_j)$ of a query is defined as average genderedness of the terms in the query, after removing the stop words.
- Genderedness $g_{q_j}(i)$ of a document i retrieved for a query q_j, is computed as average genderedness of its terms, neglecting the stop words and query terms.
- Genderedness $g_{q_j}(\sigma_j)$ of a ranked list $sigma_j$ is computed as a weighted average of documents' genderedness. An inverse logarithmic function of rank determines the weight of each document.
- Given a set of queries $\mathscr{Q}$, a set of ranked lists (one per query) retrieved by a system s from a collection $\mathscr{I}$, and the linear fit between query genderedness and ranked list genderedness, the GSR is the slope $m_s(\mathscr{Q},\mathscr{I})$ of the linear fit.

Hence, GSR is formally defined as follows:

Definition 8. Gender Stereotype Reinforcement (GSR).
Let $\mathscr{Q}$ with cardinality N be a set of queries, $g(q_j)$, $j \in [1,N]$ the genderedness of $q_j \in \mathscr{Q}$; let $\mathscr{I}$ be a corpus of documents and σ_j the ranked list (permutation of documents) provided by a system s for the query q_j. Then, GSR of s on collection $(\mathscr{Q},\mathscr{I})$ is defined as:

$$m_s(\mathscr{Q},\mathscr{I}) = \frac{1}{\text{std}^2_{g(q)}} \frac{1}{N} \sum_{j=1}^{N} (g(q_j) - \mu_q)(g_{q_j}(\sigma_j) - \mu_{q,\sigma}). \tag{6.18}$$

GSR weighs the extent to which a SE responds to stereotypically gendered queries with documents containing stereotypical language with the same polarity. In the above equation,

$\mu_q, \text{std}^2_{g(q)}$ are query genderedness mean and variance over $\mathscr{Q}$, and $\mu_{q,\sigma}$ is the average genderedness of ranked lists of documents.

We chose slope instead of correlation, since the latter quantifies the predictability of the genderedness of a ranked list, given that of the query. The former also captures the extent to which highly "female" and "male" queries are answered with completely different language along the gender dimension.

Example 1. We build a toy document collection to show how GSR captures gender stereotypes. From [129] we sample the single-word jobs with the widest gender gap (Table D.4).

- $\mathscr{Q}$ is the set of (single-word) queries of occupations considered hereafter.

  High female representation: hygienist, secretary, hairdresser, dietician, paralegal, receptionist, phlebotomist, maid, nurse, typist.

  High male representation: stonemason, roofer, electrician, plumber, carpenter, firefighter, millwright, welder, machinist, driver.

- $\mathscr{I}$ is the set of all documents derived from permutations of "The ⟨person⟩ is a ⟨job⟩", with ⟨person⟩ $\in \{\texttt{man}, \texttt{woman}\}$ and ⟨job⟩ from all occupation entries in Table D.4.

- `N` is a neutral retrieval system that returns, for each query, both documents (female and male) in which the query term appears (Figure 6.5, center, $m_{\texttt{N}}(\mathscr{Q}, \mathscr{I}) = 0$).

- `S` is a retrieval system returning, for each query, only the stereotypical document (Figure 6.5, left, $m_{\texttt{S}}(\mathscr{Q}, \mathscr{I}) = 1.61$). For example, given a query about a job with high male representation, `S` would only provide documents mentioning men.

- `CS` is a retrieval system returning, for each query, only the counter-stereotypical document (Figure 6.5, right, $m_{\texttt{CS}}(\mathscr{Q}, \mathscr{I}) = -1.61$). Contrary to `S`, given a query about a job with a high representation of men, `CS` would only provide documents that mention women.

Figure 6.5 shows the behavior of three synthetic search engines measured by GSR. Each point represents a query q_j, with its genderedness $g(q_j)$ on the x axis and the genderedeness of documents retrieved for q_j ($g_{q_j}(\sigma_j)$) on the y axis. GSR is the slope of the linear fit. `CS` and `S` are quite extreme, as they only return documents that challenge the gender gap in occupations or fully reinforce it, while N is neutral. GSR successfully captures this aspect with zero slope for `N` and significantly nonzero slopes for `S` and `CS`, equal in magnitude and opposite in sign. The magnitude of GSR for `S` and `CS` is very large compared, for example,

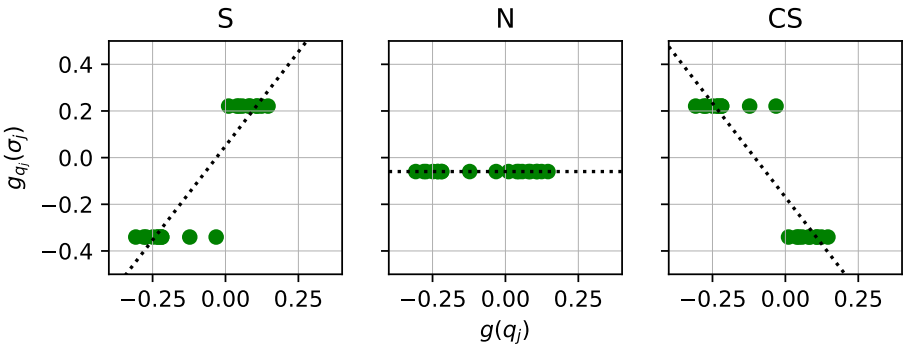


Fig. 6.5 GSR on toy dataset $(\mathscr{Q}, \mathscr{I})$ for different retrieval systems: stereotypical (S), neutral (N), counter-stereotypical (CS). GSR is the slope of linear fit, taking values $m_{\mathtt{S}}(\mathscr{Q}, \mathscr{I}) = 1.61$, $m_{\mathtt{N}}(\mathscr{Q}, \mathscr{I}) = 0$, $m_{\mathtt{CS}}(\mathscr{Q}, \mathscr{I}) = -1.61$.

against GSR values of real IR algorithms in a news collection, of the order of magnitude of 10^{-2} (Table 6.4). This depends on (1) the collection $(\mathscr{Q}, \mathscr{I})$ itself, especially conceived for *direct gender stereotype*, (2) the systems S and CS which are extreme as they respond to job-related queries with documents mentioning women or men in accordance with, or in opposition to, stereotypes related to gender gaps in the occupations mentioned within a query. Overall, this experiment shows that GSR is suited to capture direct gender stereotypes. This is further confirmed by experiments on a shared IR collection (Section 6.2.4).

Key properties. The toy example presented above defines a controlled setting where we can test the *convergent validity* of our construct with metrics of algorithmic fairness and diversity in IR [160, 286]. The IR task can be framed as a binary classification problem – i.e. classifying documents as relevant or non-relevant – with a binary protected attribute encoding whether a document is stereotypical or not. A document is deemed stereotypical for a query if it displays genderedness of same polarity. We assume a search history (in this context: solution to the classification problem) to be reasonable if, for every query q_j, the documents included in the ranked list σ_j contain the query term.

Therefore, in our controlled setting, for each query, such as `driver`, a maximum of two documents can be retrieved, namely `The woman is a driver` (counter-stereotypical) and `The man is a driver` (stereotypical), i.e. the ones which contain the term `driver` from the complete permutation described above. This is a sensible assumption and makes enumeration feasible.

We enumerate every reasonable solution and, for each, compute GSR along with the percentage of stereotypical documents among the retrieved ones, equivalent to *statistical parity fairness* from Gao and Shah [286], already employed in the context of fair ranking to enforce equal exposure of SE users to different topics. The results of Figure 6.6 show a

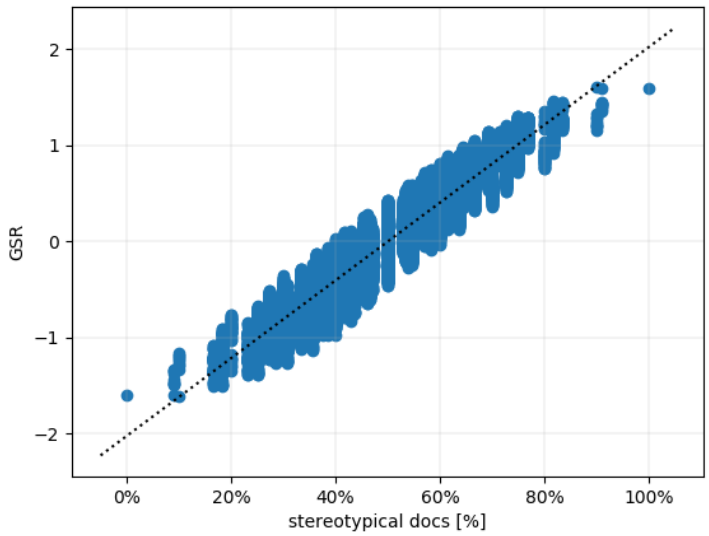


Fig. 6.6 Agreement between GSR and percentage of stereotypical documents among retrieved ones, equivalent to *statistical parity fairness* [286]. Pearson's $r = 0.92$, $p < 1\text{e}{-40}$.

strong agreement between these quantities, which stems from the very definition of the slope coefficient.

$$m_{x,y} = \frac{1}{\text{std}_x^2 N}\sum_{i=1}^{N}(x-\mu_x)(y-\mu_y) = \frac{1}{\text{std}_x^2 N}\left[\sum_{S}(x-\mu_x)(y-\mu_y)+\sum_{CS}(x-\mu_x)(y-\mu_y)\right], \tag{6.19}$$

where we explicitly partitioned the retrieved documents into stereotypical (S) and counter-stereotypical (CS). This partition is equivalent to topical group assignment for which statistical parity fairness enforces equal user exposure [286]. Specializing Equation (6.19) for GSR, we get

$$m_s(\mathscr{Q},\mathscr{I}) = \frac{1}{\text{std}_{g(q)}^2 NW}\left[\sum_{(q_j,i)\in S}\frac{(g(q_j)-\mu_q)(g_{q_j}(i)-\mu_{q,\sigma})}{log_2(\sigma_j^{-1}(i)+1)}+\sum_{(q_j,i)\in CS}\frac{(g(q_j)-\mu_q)(g_{q_j}(i)-\mu_{q,\sigma})}{log_2(\sigma_j^{-1}(i)+1)}\right], \tag{6.20}$$

where we have used Equations (6.17) and (6.18). Documents i that are stereotypical for a query q_j (first summation in Equation (6.20)) bring a positive contribution to the slope coefficient $m_s(\mathscr{Q},\mathscr{I})$, while counter-stereotypical documents bring a negative one. Equal

exposure (50% stereotypical documents) does not entail neutrality (GSR=0) and vice versa, however, the two measurements are clearly correlated (Pearson's $r = 0.92$, significant at $p < 1e{-}40$). Indeed, GSR is a measure of *weighted* statistical parity between stereotypical and counter-stereotypical documents, with weight proportional to genderedness of query times genderedness of document. Although not central in this toy example, document position in ranked list $\sigma_j^{-1}(i)$ is a further weighing factor through logarithmic discount.

A discussion focused on *discriminant validity* (Section 6.2.1) of GSR is due. In any sensible text corpus, words from the same domain (e.g. medicine) and *a fortiori* subdomain (e.g. gynecology) are likely to co-occur and their word vectors will end up close to one another, duly capturing their semantic proximity. At the same time, to satisfy an information need (e.g. query "in vitro fertilization"), it will be necessary to employ the specific language of relevant fields to which the query refers (such as gynecology). For this reason, some query-document agreement should be expected in language and, more specifically, in genderedness. Thus, in non-trivial settings, any reasonable SE is expected to have positive GSR. Figure 6.3 is a real example of this aspect, depicting a query and a relevant document (both represented as bags-of-words, the latter subsampled for brevity) taken from the Robust04 collection [346]. The document surely contains domain-specific language; however, it is also centered around female entities, echoing old gender roles in the framing of involuntary childlessness [585].

In other words, if a kernel of truth is present in some stereotypes [632], positive GSR captures a kernel of relevance, and its value is fundamentally influenced by documents available to respond to a query. Taking into account this aspect, it is fundamental to provide a baseline GSR for relevant documents. System N is an example of such a baseline in the toy setting of Example 1. Upward deviation of GSR from this baseline (as computed differentially or through a ratio) is regarded as the SE's contribution towards reinforcement of gender stereotype. The baseline GSR for relevant documents captures a mixture of *historical bias* [751] and inevitable domain-specificity of language. For this reason, when measuring GSR, it is important to have a list of relevant documents as a baseline, which is to be externally validated and reasonably considered a ground truth.

6.2.3 Datasets

Synthetic data

To test GSR against indirect gender stereotypes (Definition 4), we build a synthetic dataset of queries $\mathscr{Q}$ and documents $\mathscr{I}$ where a SE might promote gender stereotypes, or counter them. We simulate three SEs, designed to be stereotypical, counter-stereotypical or neutral. In the following, we summarize the dataset $(\mathscr{Q}, \mathscr{I})$ and the simulated SE working on the dataset.

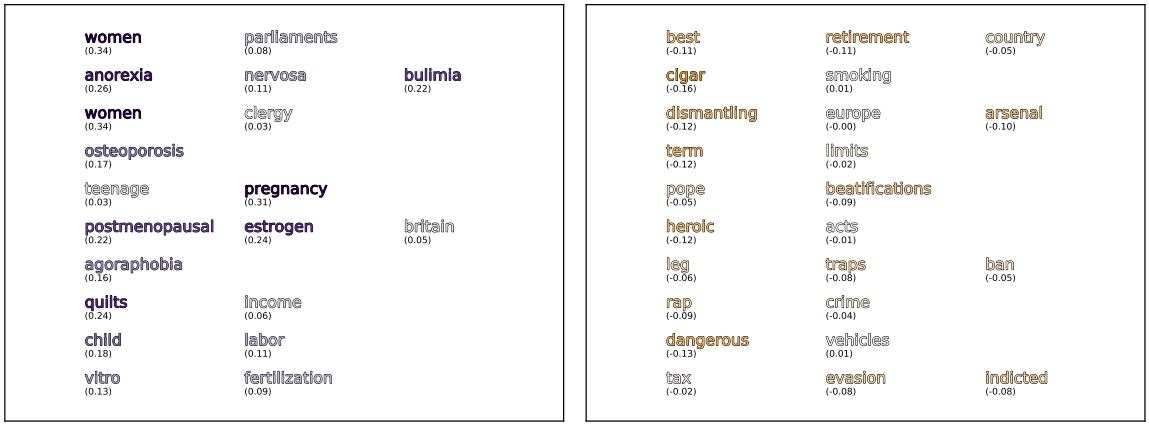
- $\mathscr{Q}$ is the set of (single-word) queries consisting of occupations with a large gap in gender representation from Table D.4.
- $\mathscr{I}$ is the set of all documents derived from permutations of "The ⟨job⟩ is ⟨adjective⟩", with ⟨job⟩ from Table D.4 and ⟨adjective⟩ from Table D.1. The adjectives considered are commonly used to assess gender stereotypes held by a population, and are descriptive of *communion* (commonly considered a female trait) and *agency* (often associated with males).
- N is a neutral retrieval system (search engine) that returns, for each query, all documents in which the query term appears.
- S is a retrieval system that returns only the stereotypical document in which the query term appears. Predominantly female (male) jobs are therefore associated to communion (agency) adjectives - e.g. `The plumber is hardworking`.
- CS is a retrieval system returning only the counter-stereotypical document. Predominantly female (male) jobs are associated with agency (communion) adjectives.

Real-world data

We analyze GSR on a widely used TREC evaluation collection: TREC 2004 Robust Track [346], hereafter called Robust04. This collection consists of about 528K news documents and 249 queries. We have selected this candidate collection for three main reasons: (1) the large number of queries ($N = 249$) for which relevance judgments from human annotators are available; (2) the domain, news, where the importance of SE in mediating user access is wide and well established [371], and (3) its relevance within the IR community. The data brief for this dataset is available in Appendix A (§ A.1.190).

To qualitatively evaluate whether Robust04 is an interesting collection for GSR analysis, we focus on the most gendered queries according to Word2Vec (`w2v`), contributing the most to GSR in Equation (6.18), and assess whether they contain recognizable gender stereotypes. We restrict our analysis to topic titles, inspecting the 10 most "female" and "male" queries according to `w2v`. These are the queries q_j whose title has the highest and lowest *genderedness* $g(q_j)$, calculated as the average projection of query terms onto the gender direction of `w2v` [84]. The most gendered queries are depicted in Figures 6.7a and 6.7b.

Among the most "female" queries, few are *intrinsically gendered*, such as `women in parliaments` (topic 321) and `women clergy` (topic 445). Some more queries are *biologically gendered* (such as `postmenopausal estrogen britain` and `osteoporosis` – topics



(a) Female (b) Male

Fig. 6.7 Most gendered queries from Robust04 under `w2v`. The text is printed with color-coded gradient where strongly male words are orange, strongly female words are purple, neutral words are white. Terms' projection along gender direction can be read below each word. Stop words are removed from queries.

356 and 403 respectively), describing topics biologically associated with women. The remaining queries can be described as *culturally gendered*. Some are associated with disorders with apparently higher incidence on the female population (`agoraphobia` and `anorexia nervosa bulimia` – topics 348 and 369). The final three (`quilts income`, `child labor`, `in vitro fertilization` – topics 418, 440, 368) seem to capture unnecessary or even harmful stereotypes related to communion [224] and gender roles [585]. Topic 440 (`child labor`) highlights a limit of GSR and the underlying word representations. In the presence of a polysemous word, its embedding encodes a mixed representation of the different uses this word. In this example, the embedding of `labor` has been influenced by the meaning related to giving birth, while the query has a different intent. Contextualized approaches may be useful to mitigate this issue.

All "male" queries seem to be *culturally gendered*, containing terms loosely related to agency (such as `dismantling`, `heroic`, `evasion`), and occupation (`retirement`, `term`), contrary to communion (`dangerous`, `arsenal`, `crime`, `traps`), and more frequently associated with men (`cigar`, `rap`).

Overall, most of these queries associate gender (as encoded by `w2v`) with undesirable and harmful concepts. For this reason, Robust04 is a reasonable collection for studying GSR.

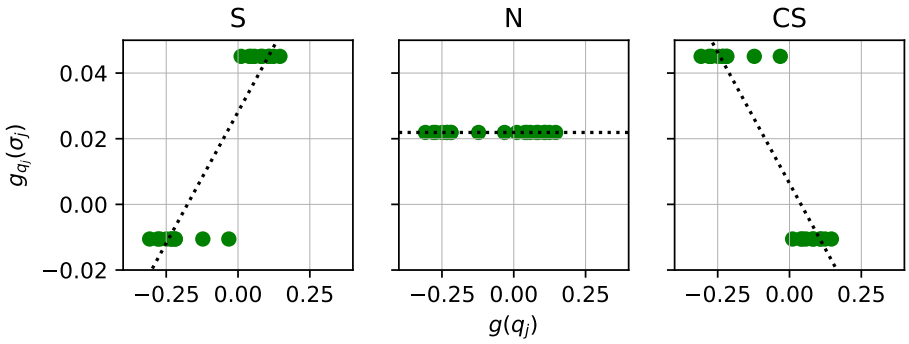


Fig. 6.8 GSR on synthetic dataset for different retrieval systems: stereotypical (S), neutral (N), counter-stereotypical (CS). GSR is the slope of linear fit, taking values $m_{\mathtt{S}}(\mathscr{Q},\mathscr{I}) = 0.16$, $m_{\mathtt{N}}(\mathscr{Q},\mathscr{I}) = 0$, $m_{\mathtt{CS}}(\mathscr{Q},\mathscr{I}) = -0.16$.

6.2.4 Experiments

Recall that GSR captures both problematic query-document associations by a SE, and domain-specificity of language embedded in a collection of documents and queries. Hence, a perfect SE, which retrieves all and only the relevant documents for each query, is expected to have a positive GSR. This stems from the fact that a document is more likely to be relevant for a query if it contains specific language from the query domain, and words from the same domain tend to cluster together in the WE space, and consequently in the gender subspace.

Therefore, we use the GSR of the perfect SE as a baseline against which to compare the real SE. In other words, a SE can be said to counter gender stereotypes, even if it displays a positive GSR, as long as its GSR is smaller than that of the perfect SE. On the contrary, a SE that reinforces gender stereotypes will have a larger GSR. For this reason, we perform tests on shared test collections based on the Cranfield paradigm [163], for which relevance judgments have been provided by qualified human assessors.

Indirect gender stereotypes

Setup. We hypothesize that GSR can measure *indirect* gender stereotypes (Definition 4) arising from the clustering of concepts and the segregation of language along the gender direction, such as the association of stereotypically gendered occupations and traits. For example, GSR should highlight situations where SE respond to queries about jobs with strong male representation with documents that focus on traits related to agency. To test this hypothesis, we employ the synthetic dataset described in section 6.2.3.

Results. The results are summarized in Figure 6.8. System S reinforces indirect stereotypes, since it links occupation and personality roles along gender-stereotypical lines, strengthening

gender clusters. The proposed measure successfully captures this aspect, along with the neutral nature of N and the counter-stereotypical nature of CS.
Interpretation. GSR can detect indirect gender stereotype reinforcement for gender stereotypes that are encoded in the underlying WE.

GSR of popular IR systems

Setup. SE based on WE, such as `w2v` and FastText (`ftt` - [565]), are expected to have a higher GSR than purely lexical ones, i.e. they should be more prone to support gender stereotypes due to the problematic gender information embedded in word representations. On the Robust04 collection (Section 6.2.3), we evaluate GSR for three families of well-known IR systems, which could serve as a basis for a SE:

- Lexical: these algorithms are based on matching query terms to document terms, without any information on semantics. In this group, we include three models inspired by different key paradigms: the widely used probabilistic model `BM25` [675], a popular language model [862] (i.e. Language Modelling with Bayesian smoothing and a Dirichlet Prior) called `QLM`, and the classic vector space model `tf-idf` [692].

- Semantic: IR systems based on WE have been proposed [802], with the idea of exploiting the latent relationship between words encoded by embeddings. We test `w2v_add`, `w2v_si` [802] and `ftt_add` (`w2v_add`'s counterpart based on `ftt` WE) for this family.

- Neural: WE can be fed as input to neural networks, which in turn learn to match the signals of user queries with that of relevant documents. We consider Deep Relevance Matching Model (DRMM - [333]), and Match Pyramid (MP - [619]). The embeddings used as input to these systems are `w2v` trained on Google News [565].

If a query q_j has K_j relevant documents, according to the assessors' judgments, we compute the GSR for each system on the top K_j documents retrieved by it, denoted as $\sigma_j(1:K_j)$. This makes the GSR of the perfect SE (retrieving all and only the K relevant documents) directly comparable to that of the systems at hand.
Results. As a preliminary illustration, Figure 6.9 shows the GSR for three systems.

- On the left, a search engine retrieving random documents.

- In the middle, a perfect search engine which retrieves all and only relevant documents, also ranking them perfectly according to relevance judgements.

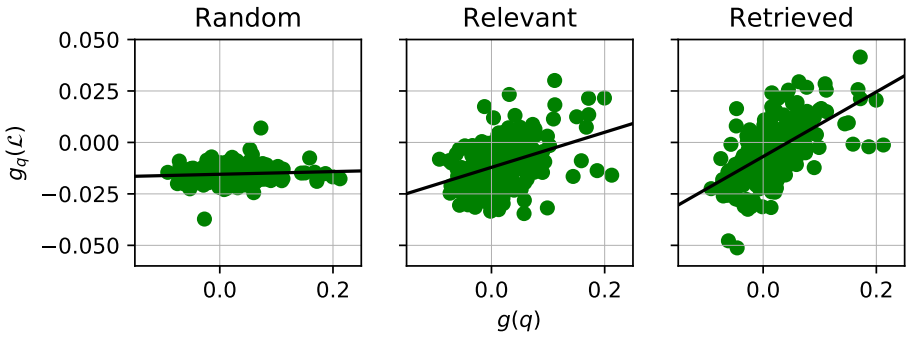


Fig. 6.9 GSR for relevant, random and retrieved docs. The x axis represents genderedness of queries $g(q_j)$, while y axis represents the genderedness of ranked document list $g_{q_j}(\sigma_j)$. GSR is the slope of the linear fit of the scatter plot.

- On the right, a search engine based on `w2v_add` [802].

As mentioned in Section 6.2.2, while discussing *discriminant validity*, positive GSR can be associated with relevance. This is shown by the positive GSR of the perfect SE in the middle panel, compared with the near-zero GSR of random retrieval. This is not surprising; it is due to a combination of language specificity and *historical bias* [751] potentially present in news coverage, as discussed in Section 6.2.2. For this reason, hereafter we will report for comparison the GSR of the perfect SE.

Figure 6.10 shows the GSR for the search results of eight different retrieval systems. Each panel contains a scatter plot of 249 different points (one for each query in Robust04), along with their linear fit (solid) and the linear fit of the perfect SE for comparison (dashed). Panels (4)-(6), depicting `w2v_add`, `w2v_si`, `ftt_add`, confirm that semantic SE have higher GSR than lexical ones, namely `QLM`, `tf-idf` and `BM25`, depicted in panels (1)-(3). This fact is easy to explain: SE based on gender-biased WE inherit the bias and tend to reinforce it.

Interestingly, neural systems based on the same word representation (`MP`, `DRMM`), shown in panels (7) and (8) respectively, seem to dampen this effect, thanks to successful weight adjustment during training, which reduces the importance of the (biased) gender direction in `w2v`.

Interpretation. As expected, semantic models based on biased WE are likely to reinforce gender stereotypes, even when based on an IDF-inspired weighting scheme (as in the case of `w2v_si`), aimed at assigning greater importance to terms that contain more information. Neural systems seem capable to dampen this effect. Lexical models have a low GSR, comparable to that of the ideal SE.

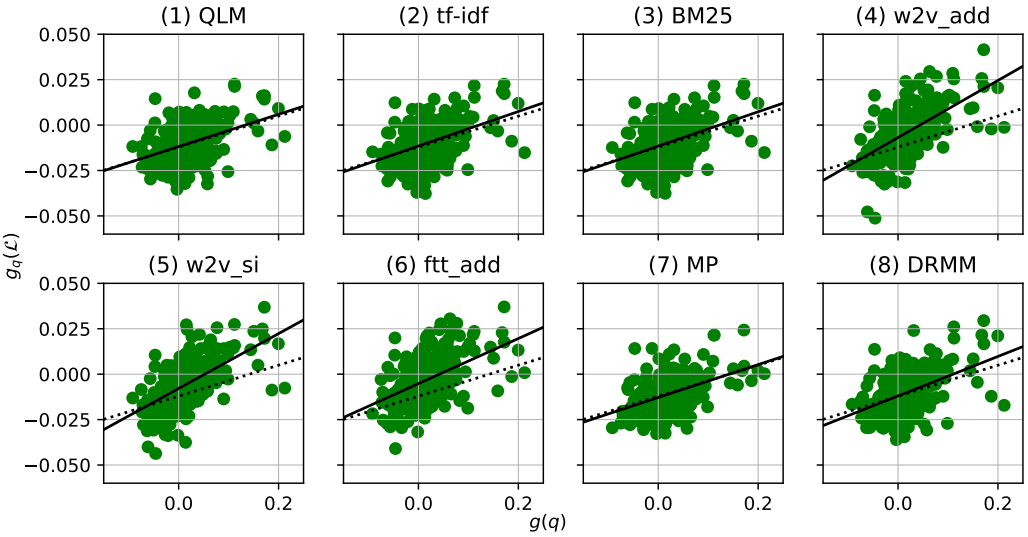


Fig. 6.10 GSR for different systems on Robust04 according to gender direction of `w2v`. The x axis represents the genderedness of queries $g(q_j)$, while the y axis represents the genderedness of ranked document lists $g_{q_j}(\sigma_j)$. GSR is the slope of the linear fit through the scatter plot (solid). The dashed line is the linear fit of the perfect SE, reported for comparison.

Effect of debiasing

Setup. The gender direction along which we measure GSR can be removed from the embeddings, by means of orthogonal projection [84]. We evaluate the impact of this operation on the performance and GSR of semantic and neural SE based on WE. For each system that relies on WE, we repeat the previous retrieval task with three different versions of `w2v` embeddings. *Regular* embeddings are the original version trained on Google News. *Debiased* embeddings are obtained by eliminating the gender direction from neutral words while maintaining it for gendered words such as `woman` [84]. *Strong debiased* embeddings take this procedure a step further, eliminating the gender component from each word [648]. The same debiasing procedures are applied to FastText embeddings fed to `ftt_add`.

Results. GSR values are reported in Table 6.4, under the header **GSR (`w2v`)**. Debiasing is effective in reducing GSR for systems where it is particularly high, namely the semantic ones, purely based on WE (`w2v_add`, `w2v_si`, `ftt_add`). However, even for these systems the reduction is quite weak, ranging between 10%-25%, as gender information leaks along different directions, orthogonal to the one that is eliminated through debiasing. This aspect has been previously studied [314], to conclude that "the gender-direction provides a way to measure the gender-association of a word but does not determine it". Our results confirm that

System	GSR (w2v)		GSR (ftt)	
	absolute	relative	absolute	relative
perfect	$8.5\text{x}10^{-2}$	0%	$9.3\text{x}10^{-2}$	0%
w2v_add				
regular	$16\text{x}10^{-2}$	84%	$15\text{x}10^{-2}$	62%
debiased	$14\text{x}10^{-2}$	62%	$14\text{x}10^{-2}$	53%
strong debiased	$14\text{x}10^{-2}$	62%	$14\text{x}10^{-2}$	53%
w2v_si				
regular	$15\text{x}10^{-2}$	77%	$15\text{x}10^{-2}$	64%
debiased	$13\text{x}10^{-2}$	55%	$14\text{x}10^{-2}$	53%
strong debiased	$13\text{x}10^{-2}$	54%	$14\text{x}10^{-2}$	53%
ftt_add				
regular	$12\text{x}10^{-2}$	46%	$15\text{x}10^{-2}$	58%
debiased	$11\text{x}10^{-2}$	35%	$13\text{x}10^{-2}$	44%
strong debiased	$11\text{x}10^{-2}$	35%	$13\text{x}10^{-2}$	43%
MP (w2v)				
regular	$9.0\text{x}10^{-2}$	6%	$9.6\text{x}10^{-2}$	4%
debiased	$9.0\text{x}10^{-2}$	6%	$9.7\text{x}10^{-2}$	5%
strong debiased	$9.0\text{x}10^{-2}$	6%	$9.7\text{x}10^{-2}$	5%
DRMM (w2v)				
regular	$11\text{x}10^{-2}$	28%	$12\text{x}10^{-2}$	24%
debiased	$11\text{x}10^{-2}$	25%	$11\text{x}10^{-2}$	21%
strong debiased	$11\text{x}10^{-2}$	26%	$11\text{x}10^{-2}$	21%
lexical				
QLM	$8.9\text{x}10^{-2}$	4%	$10\text{x}10^{-2}$	12%
tf-idf	$9.5\text{x}10^{-2}$	11%	$10\text{x}10^{-2}$	14%
BM25	$9.4\text{x}10^{-2}$	11%	$10\text{x}10^{-2}$	14%

Table 6.4 GSR measured according to w2v and ftt with 2 significant figures. Raw GSR values are shown (dubbed absolute), along with relative values, obtained from the former, as a percentage of the GSR value for the perfect search engine. Agreement betweeen w2v and ftt: Spearman's $\rho = 0.96$, $p < 1\text{e}{-10}$.

this is true for a measure based on the gender direction such as GSR. Furthermore, *strong debiasing* brings no major advantage compared to simple debiasing.

The impact on performance is very limited, as shown in Table 6.5, reporting Mean Average Precision (MAP), precision for the top-10 ranked documents (P@10) and Normalized Discounted Cumulative Gain for the top-100 ranked documents (nDCG@100). We focus on systems based on WE, leaving aside lexical ones, since our interest is to evaluate the impact of debiasing on classical performance measures. Our results, which are in line with the prior art [532], show that debiasing (both regular and strong) produces negligible changes to average performance.

How does such a minor impact on performance coexist with a moderate yet sizeable impact on GSR (shown in Table 6.4)? Figure 6.11 answers this question, depicting the

System	MAP	nDCG@100	P@10
`w2v_add`			
regular	0.067	0.170	0.174
debiased	0.068 (+2%)	0.171 (+0%)	0.176 (+1%)
str. deb.	0.068 (+1%)	0.171 (+0%)	0.175 (+0%)
`w2v_si`			
regular	0.093	0.213	0.216
debiased	0.094 (+1%)‡	0.213 (+0%)	0.217 (+0%)
str. deb.	0.094 (+1%)†	0.213 (+0%)	0.217 (+0%)
`ftt_add`			
regular	0.056	0.144	0.150
debiased	0.056 (+0%)	0.144 (+0%)	0.148 (-1%)
str. deb.	0.056 (+0%)	0.144 (+0%)	0.147 (-2%)
`MP (w2v)`			
regular	0.151	0.283	0.287
debiased	0.148 (-2%)	0.279 (-1%)	0.285 (-1%)
str. deb.	0.148 (-2%)	0.279 (-1%)	0.285 (-1%)
`DRMM (w2v)`			
regular	0.260	0.423	0.456
debiased	0.259 (−1%)‡	0.422 (+0%)	0.454 (+0%)
str. deb.	0.259 (−0%)†	0.421 (−1%)†	0.457 (+0%)

Table 6.5 Impact of regular debiasing [84] and strong debiasing [648] on performance of models based on WE. A Student's t test is computed between regular and debiased versions of the same algorithm, with significance at $p = 0.05$ and $p = 0.01$ denoted by $\dagger, \ddagger$ respectively.

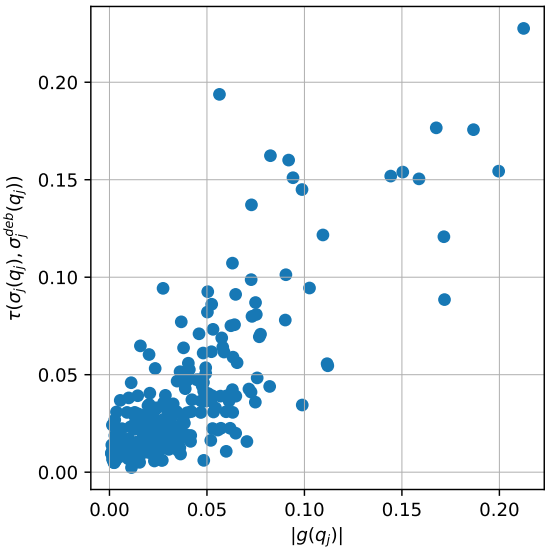


Fig. 6.11 Impact of regular debiasing on `w2v_add`: absolute value of query genderedness, on the x axis, and difference between top-100 documents retrieved by `w2v_add` before and after debiasing, on the y axis, as measured by Kendall τ distance. Pearson's $r = 0.81$, $p < 1e-50$.

Kendall τ distance for ranked lists of documents (top-100) retrieved by `w2v_add` before and after debiasing (on the y axis), against the genderedness of the respective query in absolute value (on the x axis). From the plot an expected property of debiasing WE emerges in the context of IR algorithms: the most impacted queries are the ones with high genderedness, which are also the most important ones for GSR. If most queries have a low gender score, then the impact on aggregated performance will be insignificant.

Interpretation. Debiasing moderately reduces GSR and has a negligible impact on performance.

Reliability

Setup. Word representations learned with different techniques and corpora such as `w2v` (Google News) and `GloVe` (Wikipedia) have already been shown to exhibit a similar bias along the gender direction [84, 110, 288]. To test the reliability of GSR, we check how dependent it is on a specific WE implementation. We do so by computing GSR based on FastText (`ftt`) WE, trained on Common Crawl and check its agreement with `w2v`-based GSR. The procedure to isolate a gender direction and project word vectors from `w2v` onto it

[84] is perfectly applicable to different WE. We compute GSR according to `ftt` embeddings, and compare it against results from previous sections obtained with `w2v`.

Results. As a preliminary check, we compute the correlation between query genderedness measured by `w2v` and `ftt`. Figure 6.12 is a scatter plot of the genderedness of 249 queries from Robust04 under `w2v` and `ftt`, which shows a strong correlation between the two (Pearson's $r = 0.78$, $p < 1e{-}40$). This preliminary check confirms that `w2v` and `ftt` are likely to encode stereotypically gendered concepts in similar ways.

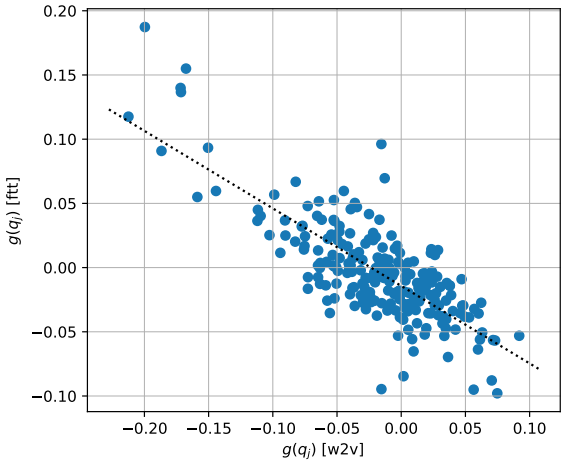


Fig. 6.12 Genderedness of Robust04 queries, according to `w2v` (x axis) and `ftt` (y axis). Correlation: Pearson's $r = 0.78$, $p < 1e{-}40$.

Table 6.4 shows the values of `ftt`-based GSR, in columns 3 and 4. SE can be ranked according to the GSR scores computed with `ftt` and `w2v`. We regard the correlation of these scores as a measure of the reliability of GSR across different WE. In other words, we would like the ranking determined by `w2v`-based GSR and `ftt`-based GSR to agree as much as possible, as measured by Spearman's rank coefficient (ρ). The values in Table 6.4 (either absolute or relative) yield Spearman's $\rho = 0.96$, with a p-value $p < 1e{-}10$. We conclude that GSR and the underlying gender direction is fairly reliable across `w2v` and `ftt` embeddings, despite the different text corpora from which they were learned (Google News and Common Crawl, respectively).

We anticipated that SE based on `w2v` (namely `w2v_add`, `w2v_si`) would have a higher score when GSR is measured according to the same `w2v` WE, than when GSR is `ftt`-based. Similarly, `ftt_add` has higher GSR when computed according to `ftt` than according to `w2v`. This is not surprising, given each SE based on word representations inherits the peculiar biases of its underlying WE, which are the same biases that GSR captures. Despite this, the `w2v`-based and `ftt`-based GSRs show solid overall agreement.

Interpretation. GSR is reliable and stable across WE that differ in architecture and choice of (large) training text corpus from the web domain.

GSR and direct stereotypes

Setup. To interpret the results from GSR, we investigate its relationship with explicit mentions of female and male entities. For every document, we compute a binary measure of "intrinsic genderedness". For the sake of simplicity, a document is considered:

- Intrinsically male, if it contains more male mentions than female ones.
- Intrinsically female, if it contains more female mentions than male ones.
- Neutral, otherwise.

We hypothesize that a high GSR will result in associating stereotypically gendered queries (such as those depicted in Figures 6.7a, 6.7b) with intrinsically gendered documents of the same polarity. In other words, GSR should capture direct gender stereotypes (Definition 3), taking large values for SE which associate stereotypically female (male) concepts with female (male) entities. This hypothesis relates to the *content validity* of GSR, as we would expect our measure to capture this form of direct bias.

To assess intrinsic genderedness of documents, male and female names are sourced from `nltk`'s names corpus.[9] Gendered nouns, adjectives, and titles are obtained starting from definitional pairs and gender-specific words in Appendix C of Bolukbasi et al. [84], of which we only keep words that specifically refer to a person. Under this criterion, `aunt` is considered a female entity, while `pregnancy` is not. The resulting word list referring to gendered entities is reported in D.2.

We compare the "perfect" search engine (dubbed `P`) (retrieving all and only the relevant documents for each query) against one based on `w2v_add`, which has the highest GSR among the tested systems. As a comparison, we also include `QLM` and `MP` (low GSR) and `ftt_add` (medium GSR). Each system is compared to `P` as follows:

- For each query q_j, with K_j relevant documents, we consider the top K items in a ranking $\sigma_j(1:K_j)$ returned by a SE. We compute the number of intrinsically female and male documents (dubbed $f(q_j)$ and $m(q_j)$ respectively). We use their ratio as a summary of the representation gap in the search results ($\text{gap}(q_j) = \frac{m(q_j)}{f(q_j)}$). For a given list of search results $\sigma_j(1:K_j)$, $\text{gap}(q_j)$ quantifies the extent to which top-ranked documents in σ_j tend to mention more male entities than female ones.

[9] https://www.nltk.org/book/ch02.html

- For each query q_j, we compute gap(q_j) under `P`, the perfect SE, and `sys`, the system at hand (`w2v_add`, `ftt_add`, `QLM` and `MP`).
- Their difference, $\Delta_{\text{gap}}(q_j) = gap(q_j)^{\texttt{sys}} - gap(q_j)^{\texttt{P}}$ summarizes the over- or under-exposure of user to documents with male entities, compared with the ground truth of system `P`.
- Based on the sign of $\Delta_{\text{gap}}(q_j)$, we determine whether `sys` favors male documents (if positive) or female documents (if negative).
- To test our hypothesis, we compute sgn$(\Delta_{\text{gap}}(q_j))$ for each query q_j and compare it with the genderedness $g(q_j)$ of the query.

We expect low $g(q_j)$ (stereotypically male queries) to be associated with over-representation of male entities, high $g(q_j)$ with under-representation. Furthermore, this relationship should be strong for `w2v_add`, weaker for `ftt_add`, and absent from `QLM` and `MP`

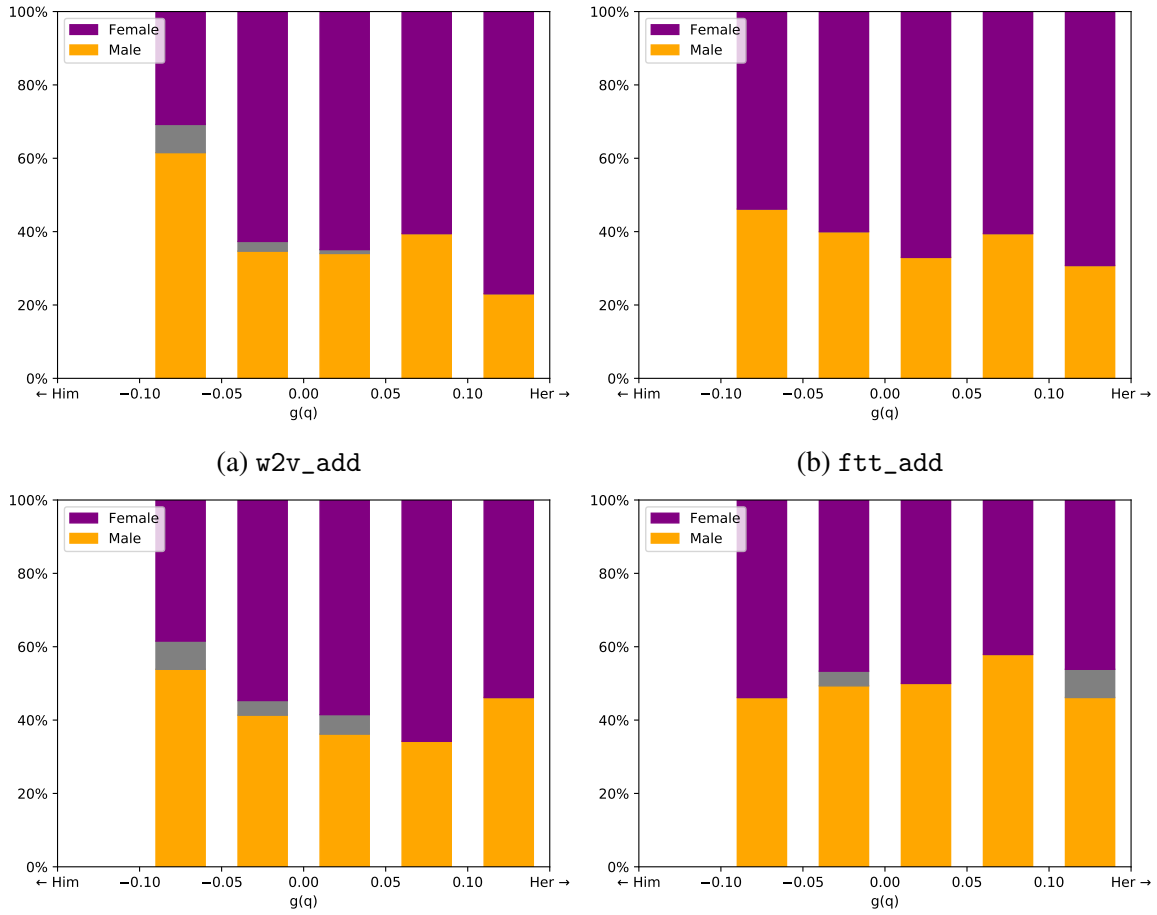
Results. Figure 6.13a confirms our expectation for `w2v_add`, with a clear trend along the x axis. We expect the trend to be less evident for `ftt_add`, given its lower GSR, which is confirmed by Figure 6.13b.

`QLM` and `MP` (low GSR) are represented in Figures 6.13c and 6.13d. The former seems to have a weak trend similar to that of `ftt_add`, disconfirmed however by the last bin, which contains the most gendered queries ($|g(q_j)| > 0.1$) but does not significantly favor "intrinsically female" documents. The latter displays no trend along the x axis. In summary, as anticipated, no consistent trend is visible for systems with low GSR.

Interpretation. We conclude that GSR captures this form of direct gender stereotype: SE with a high GSR associate stereotypically gendered queries with documents mentioning people of the same gender.

6.2.5 Discussion

We defined Gender Stereotype Reinforcement (GSR) in SE, a construct describing the tendency of a SE to reinforce direct and indirect biases about gender, which we made operational employing WE as a measurement tool. We validated our approach against well-studied gender stereotypes from the psychology literature, and exploited the framework of construct validity [397] to critically evaluate our novel measure. We found that GSR captures gender stereotypes, while also being influenced by the relevance of documents retrieved for each query. This is due to the domain-specificity of language: queries and relevant documents are likely to share some specific vocabulary, whose words cluster in the embedding space and,



(a) `w2v_add` (b) `ftt_add`

(c) `QLM` (d) `MP`

Fig. 6.13 On the y axis, in orange (purple) percentage of queries where intrinsically male (female) documents are over-represented among retrieved ones. The complementary set, depicted in gray, is the percentage of queries for which neither gender is over-represented ($\Delta_{gap}(q_j) = 0$). The x axis is a quantization on query genderedness $g(q_j)$. Top panes depict semantic systems `w2v_add` (high GSR) and `ftt_add` (medium GSR), bottom panes show systems from the lexical and neural family (`QLM` and `MP` - low GSR). No query in Robust04 has $g(q_j) < -0.1$, the bin is therefore empty.

subsequently, along the gender subspace. This aspect can be compensated for when assessor judgments are available. In this regard, TREC collection Robust04 [346] has proven to be a suitable dataset to measure the extent to which different IR algorithms reinforce gender stereotypes. This is due to availability of relevance judgments, the large number of queries, and the interesting content of some queries from a gender stereotype perspective.

Subsequently, we studied how lexical, semantic, and neural IR models reinforce gender stereotypes. We found that semantic models, based on biased WE, are more prone to reinforcement of gender stereotypes, while neural models based on the same word representations

can mitigate this effect; neural models exhibit low GSR, comparable to that of lexical systems. The reliability of these conclusions was tested with two different sets of WE (Word2Vec and FastText), identifying strong agreement between the two measurements.

Finally, we evaluated the impact of debiasing WE on downstream IR tasks. Regular debiasing [84] and strong debiasing [648] have a similar effect, reducing GSR to a significant, yet moderate extent. We conclude that the gender direction encoded by WE is a useful proxy for the gender-related biases contained in the large online corpora on which they have been trained. However, debiasing techniques based on projecting WE orthogonally to the same gender direction are superficial and insufficient, due to redundant encoding of stereotypical information. This also explains the minimum impact debiasing has on model performance.

In summary, GSR can measure associations of documents and queries along gendered lines, detecting and quantifying polarization in the language used to respond to stereotypically female and male queries. We showed that GSR captures the difference in the number of stereotypical and counter-stereotypical documents within a search history, drawing a parallel with existing statistical parity metrics [286].

A limitation of our measurement is the compositional model employed to assemble word scores into document scores, which does not account for syntactic structure, thus neglecting important information, such as negation. A second drawback is the noisy nature of the gender information encoded in WE, which should discourage the deployment of GSR on small collections, unless supported by human supervision. These observations are crucial to discuss the *consequential validity* of the proposed measure. If GSR were to be integrated as part of the ranking function of a SE, it would likely favor documents which appear to be gender-neutral or counter-stereotypical for the queries issued by users. Indeed, it would be possible for document providers to target our measure, ensuring that their documents are not flagged as stereotypical for some queries of interest. Moreover, intrinsically gendered queries, such as `women in parliament`, require special care; low GSR may contradict user preferences. For these reasons, we consider our operationalization of GSR a preliminary attempt to measure gender stereotype reinforcement in SE, with limited consequential validity in fully automated contexts. Future work should include an exploration of different compositional models, based, for instance, on dependency parsers, and novel approaches to compute a gender score for words and phrases, including *ad-hoc* training [882]. Finally, it will be interesting to measure GSR in cross-lingual scenarios; grammatical gender may pose an additional challenge in some languages, especially for the isolation of gender information along a single direction.

To our knowledge, GSR is the first measure in the domain of IR capable of quantifying a specific type of representational harm, namely gender stereotypes. This opens the possibility to quantitatively study the interplay between distributional and representational harms,

making GSR very promising in terms of *hypothesis validity* and its future uses. In the context of job search, it would be meaningful to study this interaction, due to the high stakes, the proven existence of biased tools [141], and the availability of datasets [194]. As noticed by Chen et al. [141], search results in resume SE, which are biased with respect to gender, may lead to a dual harm: an immediate one, for the providers of CVs, competing to appear in the current search, and a long-term one, for the perception and future decisions of recruiters.

6.3 Chapter Outcomes

In this chapter, we have targeted important gaps in the fair ranking literature. Toward **O1**, we have performed a critical analysis of pairwise fairness, a family of measures lacking clear contextualization in the prior literature. After retrospectively mapping these measures to a plausible construct, we have proposed several enhancements grounded in user models and combined them into a novel pairwise measure, studying its connection with popular fairness measures. To address **O2**, we have drawn from social psychology to rigorously define the GSR construct, which we have operationalized and verified extensively, leveraging validity theory. We have audited IR algorithms from different families through GSR to quantify their tendency to reinforce gender stereotypes in their users. The individual contributions of each study are summarized in Sections 6.1.6 and 6.2.5. In this section, we present a comprehensive discussion of the outcomes of these works and their joint significance for fair ranking research.

Primarily, our work demonstrates the importance of a precise division between construct and operationalization for fair ranking. Untangling these components from each other brings several benefits to the underlying measures, their proponents, and their users. First and foremost, this approach leads to a reflection about the construct itself and whether there is a clear understanding of the property that should be captured by a given measure. Negative answers should lead to domain-specific research and refinement of the target. After a construct is made explicit, measurement proponents can employ domain-specific knowledge, such as user models, for the operationalization. This is especially important in fair ranking, where researchers can benefit from a long tradition of studies about user browsing models. In addition, a clear statement of the target construct serves the purpose of circumscribing the intended use and applicability of a measure, assisting researchers and practitioners in choosing the right measure for their goals. This is especially important in a field such as algorithmic fairness, where the number of new measures proposed in the literature is growing steadily. The benefit of this framework for measurement users is twofold: on the one hand, they can focus on measurements whose underlying construct most closely aligns with their

goals; on the other hand, they can understand the operationalization choices made by the proponents and deviate from them to tailor a measure according to their needs. More broadly, distinguishing the construct from its operationalization favors understanding and progress for the research community as a whole, enabling more informed comparisons between fairness measures and studies of their similarities and differences.

This chapter also ties back to the relevance of dataset selection and curation discussed in Chapters 3 and 4. Just like other data-driven fields, the quality and reliability of fair ranking research depends on the data supporting it. Firstly, datasets should be selected based on the task and domain at hand. This preliminary requirement is not always met, as testified by fair ranking research leveraging COMPAS and German Credit [421, 843]. Additionally, it is important to perform preliminary analyses of datasets, and evaluate whether they are adequate to study a given property, i.e., they can generate sizeable variations in the construct at hand. For example, GSR is related to ranking tasks and to the domain of information systems. However, not many datasets from these fields are likely to contain queries of interest for GSR analysis, i.e., with potential to reinforce or counter gender stereotypes in search engine users. Our choice of the Robust04 collection for GSR experimentation was informed by such analyses. As a substitute or supplement to real-world datasets, synthetic ones may be employed. If the construct at hand is well understood, synthetic resources generated in a principled fashion can produce valuable insight with tight control over variables of interest. Overall, accurate fairness measurements require contextually relevant measures computed on carefully selected datasets.

Chapter 7

Conclusion

This thesis supports principled and contextualized approaches to measure algorithmic fairness, enabling improvements for two key components of the fairness problem space, namely data and measures. With respect to datasets, we have rigorously shown that the scholarly field of algorithmic fairness has converged to suboptimal data practices, designating few datasets with serious limitations as de-facto fairness benchmarks. To overcome this major limitation, we have developed approaches and guidelines to select datasets in a principled fashion, to augment them with sensitive information, and to curate novel resources with attention to data ethics. Regarding measures, we have shown the importance of a rigorous distinction between fairness constructs and their operationalization, drawing from normative contexts specific of *information access* to firstly define a target property, and subsequently formulate it in quantitative terms. Finally, we have combined situated fairness measures with careful data practices for effective algorithmic audits.

Our research on data focused on selection, curation, and augmentation. Firstly, we have surveyed and analyzed the prevalent data practices in algorithmic fairness research, considering all works published in the proceedings of seven important conferences from the early years of the field to the present day. We have identified the most popular resources, namely Adult, COMPAS, and German Credit, and studied their limitations, including age, contrived prediction tasks, wrong labels, noisy values, low representativeness and reproducibility. To overcome this limitation, we have contributed improvements to the field's data practices in multiple directions. Firstly, we have undertaken a comprehensive dataset documentation initiative, proposing a lightweight documentation format tailored to the needs of fairness researchers, enabling search and analysis along multiple dimensions. Our study of the field through this lens demonstrates the breadth of domains impacted by algorithmic decision-making, the variety of tasks that can benefit from algorithmic fairness, and the common challenges encountered in the field. Secondly, we have analyzed important topics

in data ethics such as re-identification, consent, inclusivity, labeling, and transparency. By demonstrating a wide spectrum of attention to these topics, ranging from conscientious to neglectful, we have made these concerns tangible and distilled a set of best practices for the responsible curation of novel fairness resources. Third, we have considered the problem of augmenting existing datasets with sensitive attributes to support fairness measurements. We have developed quantification-based algorithms to measure fairness under unawareness of sensitive attributes. We have theoretically proved that they generalize the prior art, and empirically shown that they reliably outperform it. We have also explored the potential of these approaches to decouple estimates of sample-level quantities, such as fairness and diversity, from inferences at the individual level, opening the way to proxy methods that protect individual agency and privacy from information misuse.

On the measurement side, we have targeted important gaps in the fair ranking literature, caused by a lack of context and normative reasoning. Firstly, we have studied pairwise fairness, showing that the construct behind it is underspecified, and we have filled this gap, tying pairwise fairness with producer dissatisfaction. We have highlighted some limitations in the current operationalization of pairwise fairness, and proposed a new measure to overcome them, studying its connection with exposure-based measures of fair ranking. Secondly, we have defined gender stereotype reinforcement in information access, a novel construct measuring the tendency to reinforce gender biases in search engine users, which we made operational employing word embeddings as a measurement tool. After validating our measure, engaging with the relevant literature from social psychology and validity theory, we have tested different families of information retrieval methods, showing that semantic methods are more prone to reinforcing gender stereotypes than lexical and neural methods. We have performed our experiments on a carefully selected dataset where gender stereotypes can arise due to the presence of stereotypically gendered queries. Finally, we have performed an audit of key algorithms responsible for access and pricing in the Italian car insurance industry. This study is based on a precise analysis of the relevant normative requirements and their subsequent operationalization in a mathematical formulation. Our measurement, carried out on a carefully curated dataset, shows that birthplace is often used as a pricing factor, putting foreign-born drivers at a systematic financial disadvantage, and that other factors, such as age, claim history, and city of residence, can play an important role in accessing different insurance products through comparison websites.

It is worth noting some limitations of this thesis to guide future work. The documentation initiative described in Chapter 3 represents a static snapshot of datasets used in algorithmic fairness research during the period 2014–2021. To ensure that this effort results in an accessible resource supporting up-to-date search along relevant axes, we believe the data briefs

should be made accessible through a web app, which has recently been released.[1] Updating and maintaining this resource will be a challenge, and potentially require engagement and collaboration with the wider research community. Our recommendations and best practices for ethical data curation, in Chapter 4, cover a limited set of topics, without attention to the availability of resources and the dataset development workflow. Future work in this direction should consider additional topics of interest, including worker rights and data licensing, bearing in mind power differentials and resource availability for different curators, and situating suggestions within the dataset development life cycle. Chapter 5 presents novel estimators to measure fairness under unawareness of sensitive attributes based on quantification methods. Although quantification is geared toward group-level, rather than individual-level estimates, it is still based on proxy methods, which can be misused for inferences about individuals, compromising their agency and privacy. The potential of these estimators to decouple group-level and individual-level estimates has only been explored in this work and should be studied more thoroughly. Finally, our work on fair ranking measures presented in Chapter 6, while informed by measurement theory, requires further validation. Future work should include dedicated experiments to measure the response of human subjects to different information access systems with respect to satisfaction and gender stereotypes.

Overall, this thesis is informed by and contributes to open challenges in algorithmic fairness. Data cascades, that is, data issues leading to suboptimal model evaluation and deployment, hinder both research and practice in the field. As we have shown, the algorithmic fairness community converged towards inadequate datasets employed as de-facto benchmarks; this thesis calls for and enables a more principled utilization of plentiful datasets available in this space. While algorithmic fairness aims at ensuring shared benefits and reduced iniquity of algorithmic decision-making, it should also protect the people included in datasets. Our work provides guidelines for responsible data curation supporting harm avoidance for data subjects. Furthermore, we have highlighted a disconnect between fairness research and practice, exacerbated by a misalignment between fairness measures and constructs. Our work bridges this gap by incorporating normative reasoning and relevant legislation into mathematical formulations of fairness. Finally, we have combined principled measures with adequate data to support situated and robust fairness audits under noisy or missing personal information. This thesis navigates the tension between algorithmic equity and the complexities of data acquisition, supporting fairness, accountability, and transparency in a society pursuing responsible automation.

[1] www.fairnessdatasets.dei.unipd.it

References

[1] Abbasi, M., Bhaskara, A., and Venkatasubramanian, S. (2021). Fair clustering via equitable group representations. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 504–514, New York, NY, USA. Association for Computing Machinery.

[2] Abbasi, M., Friedler, S. A., Scheidegger, C., and Venkatasubramanian, S. (2019). Fairness in representation: quantifying stereotyping as a representational harm. In *Proc. of SIAM 2019*, pages 801–809.

[3] Adragna, R., Creager, E., Madras, D., and Zemel, R. (2020). Fairness and robustness in invariant learning: A case study in toxicity classification. NeurIPS 2020 workshop: "Algorithmic Fairness through the Lens of Causality and Interpretability (AFCI)".

[4] Agarwal, A., Beygelzimer, A., Dudik, M., Langford, J., and Wallach, H. (2018). A reductions approach to fair classification. In Dy, J. and Krause, A., editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 60–69, Stockholmsmässan, Stockholm Sweden. PMLR.

[5] Agarwal, A., Dudik, M., and Wu, Z. S. (2019). Fair regression: Quantitative definitions and reduction-based algorithms. In Chaudhuri, K. and Salakhutdinov, R., editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 120–129, Long Beach, California, USA. PMLR.

[6] Agrawal, M., Zitnik, M., Leskovec, J., et al. (2018). Large-scale analysis of disease pathways in the human interactome. In *PSB*, pages 111–122. World Scientific.

[7] Ahmadian, S., Epasto, A., Knittel, M., Kumar, R., Mahdian, M., Moseley, B., Pham, P., Vassilvitskii, S., and Wang, Y. (2020). Fair hierarchical clustering. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

[8] Aka, O., Burke, K., Bauerle, A., Greer, C., and Mitchell, M. (2021). *Measuring Model Biases in the Absence of Ground Truth*, page 327–335. Association for Computing Machinery, New York, NY, USA.

[9] Al-Maskari, A., Sanderson, M., Clough, P., and Airio, E. (2008). The good and the bad system: does the test collection predict users' effectiveness? In *Proceedings of the 31st*

annual international ACM SIGIR conference on Research and development in information retrieval, pages 59–66.

[10] Albanese, G., Calbimonte, J.-P., Schumacher, M., and Calvaresi, D. (2020). Dynamic consent management for clinical trials via private blockchain technology. *Journal of ambient intelligence and humanized computing*, pages 1–18.

[11] Ali, J., Babaei, M., Chakraborty, A., Mirzasoleiman, B., Gummadi, K. P., and Singla, A. (2019a). On the fairness of time-critical influence maximization in social networks. NeurIPS 2019 workshop: "Human-Centric Machine Learning".

[12] Ali, J., Lahoti, P., and Gummadi, K. P. (2021). *Accounting for Model Uncertainty in Algorithmic Discrimination*, page 336–345. Association for Computing Machinery, New York, NY, USA.

[13] Ali, J., Zafar, M. B., Singla, A., and Gummadi, K. P. (2019b). Loss-aversively fair classification. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '19, page 211–218, New York, NY, USA. Association for Computing Machinery.

[14] Amini, A., Soleimany, A. P., Schwarting, W., Bhatia, S. N., and Rus, D. (2019). Uncovering and mitigating algorithmic bias through learned latent structure. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '19, page 289–295, New York, NY, USA. Association for Computing Machinery.

[15] Anderson, E. (1936). The species problem in iris. *Annals of the Missouri Botanical Garden*, 23(3):457–509.

[16] Andriotis, A. and Ensign, R. L. (2015). US Government uses race test for $80 million in payments. The Wall Street Journal, October 29, 2015.

[17] Andrus, M., Spitzer, E., Brown, J., and Xiang, A. (2021). What we can't measure, we can't understand: Challenges to demographic data procurement in the pursuit of fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 249–260, New York, NY, USA. Association for Computing Machinery.

[18] Andrus, M. and Villeneuve, S. (2022). Demographic-reliant algorithmic fairness: Characterizing the risks of demographic data collection in the pursuit of fairness. In *Proceedings of the 5th ACM Conference on Fairness, Accountability, and Transparency (FAccT 2022)*, Seoul, Republic of Korea.

[19] Andrzejak, R. G., Lehnertz, K., Mormann, F., Rieke, C., David, P., and Elger, C. E. (2001). Indications of nonlinear deterministic and finite-dimensional structures in time series of brain electrical activity: Dependence on recording region and brain state. *Physical Review E*, 64(6):061907.

[20] Angwin, J., Larson, J., Kirchner, L., and Mattu, S. (2017). Minority neighborhoods pay higher car insurance premiums than white areas with the same risk. Machine bias, ProPublica, New York, NY, USA.

[21] Angwin, J., Larson, J., Mattu, S., and Kirchner, L. (2016). Machine bias.

[22] Arjovsky, M., Bottou, L., Gulrajani, I., and Lopez-Paz, D. (2020). Invariant risk minimization.

[23] Arnold, M., Bellamy, R. K., Hind, M., Houde, S., Mehta, S., Mojsilović, A., Nair, R., Ramamurthy, K. N., Olteanu, A., Piorkowski, D., et al. (2019). Factsheets: Increasing trust in ai services through supplier's declarations of conformity. *IBM Journal of Research and Development*, 63(4/5):6–1.

[24] ASGI - Association for Judicial Studies on Immigration (2012). Polizze rc auto: le assicurazioni quixa e zurich non applicheranno più il parametro della cittadinanza che rendeva più care le tariffe applicate ai cittadini stranieri. http://old.asgi.it/home_asgi.php%3Fn=2057&l=it.html.

[25] Asudeh, A., Jagadish, H. V., Stoyanovich, J., and Das, G. (2019). Designing fair ranking schemes. In *Proceedings of the 2019 International Conference on Management of Data*, SIGMOD '19, page 1259–1276, New York, NY, USA. Association for Computing Machinery.

[26] Atwood, J., Srinivasan, H., Halpern, Y., and Sculley, D. (2019). Fair treatment allocations in social networks. NeurIPS 2019 workshop: "Fair ML for Health".

[27] Awasthi, P., Beutel, A., Kleindessner, M., Morgenstern, J., and Wang, X. (2021a). Evaluating fairness of machine learning models under uncertain and incomplete information. In *Proceedings of the 4th ACM Conference on Fairness, Accountability, and Transparency (FAccT 2021)*, pages 206–214, Toronto, CA.

[28] Awasthi, P., Beutel, A., Kleindessner, M., Morgenstern, J., and Wang, X. (2021b). Evaluating fairness of machine learning models under uncertain and incomplete information. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 206–214, New York, NY, USA. Association for Computing Machinery.

[29] Babaeianjelodar, M., Lorenz, S., Gordon, J., Matthews, J., and Freitag, E. (2020). Quantifying gender bias in different corpora. In *Companion Proceedings of the Web Conference 2020*, WWW '20, page 752–759, New York, NY, USA. Association for Computing Machinery.

[30] Babaioff, M., Nisan, N., and Talgam-Cohen, I. (2019). Fair allocation through competitive equilibrium from generic incomes. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 180, New York, NY, USA. Association for Computing Machinery.

[31] Backurs, A., Indyk, P., Onak, K., Schieber, B., Vakilian, A., and Wagner, T. (2019). Scalable fair clustering. In Chaudhuri, K. and Salakhutdinov, R., editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 405–413, Long Beach, California, USA. PMLR.

[32] Baeza-Yates, R., Ribeiro-Neto, B., et al. (1999). *Modern information retrieval*, volume 463. ACM press New York.

[33] Bagdasaryan, E., Poursaeed, O., and Shmatikov, V. (2019). Differential privacy has disparate impact on model accuracy. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

[34] Baharlouei, S., Nouiehed, M., Beirami, A., and Razaviyayn, M. (2020). Rényi fair inference. In *International Conference on Learning Representations*.

[35] Baker, D. W., Cameron, K. A., Feinglass, J., Georgas, P., Foster, S., Pierce, D., Thompson, J. A., and Hasnain-Wynia, R. (2005). Patients' attitudes toward health care providers collecting information about their race and ethnicity. *Journal of General Internal Medicine*, 20(10):895–900.

[36] Bakker, M. A., Tu, D. P., Gummadi, K. P., Pentland, A. S., Varshney, K. R., and Weller, A. (2021). *Beyond Reasonable Doubt: Improving Fairness in Budget-Constrained Decision Making Using Confidence Thresholds*, page 346–356. Association for Computing Machinery, New York, NY, USA.

[37] Bakker, M. A., Tu, D. P., Valdés, H. R., Gummadi, K. P., Varshney, K. R., Weller, A., and Pentland, A. (2019). Dadi: Dynamic discovery of fair information with adversarial reinforcement learning. NeurIPS 2019 workshop: "Human-Centric Machine Learning".

[38] Ball-Burack, A., Lee, M. S. A., Cobbe, J., and Singh, J. (2021). Differential tweetment: Mitigating racial dialect bias in harmful tweet detection. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 116–128, New York, NY, USA. Association for Computing Machinery.

[39] Bamman, D., O'Connor, B., and Smith, N. (2012). Censorship and deletion practices in chinese social media. *First Monday*, 17(3).

[40] Banasik, J., Crook, J., and Thomas, L. (2003). Sample selection bias in credit scoring models. *Journal of the Operational Research Society*, 54(8):822–832.

[41] Bandy, J. (2021). Problematic machine behavior: A systematic literature review of algorithm audits. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW1).

[42] Bandy, J. and Vincent, N. (2021). Addressing "documentation debt" in machine learning: A retrospective datasheet for bookcorpus. In Vanschoren, J. and Yeung, S., editors, *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1.

[43] Bao, M., Zhou, A., Zottola, S., Brubach, B., Desmarais, S., Horowitz, A., Lum, K., and Venkatasubramanian, S. (2021). It's compaslicated: The messy relationship between rai datasets and algorithmic fairness benchmarks. *arXiv preprint arXiv:2106.05498*.

[44] Barabas, C., Dinakar, K., and Doyle, C. (2019). The problems with risk assessment tools.

[45] Barbaro, M. (2007). In apparel, all tariffs aren't created equal.

[46] Barenstein, M. (2019). Propublica's compas data revisited. *arXiv preprint arXiv:1906.04711*.

[47] Barman-Adhikari, A., Begun, S., Rice, E., Yoshioka-Maxwell, A., and Perez-Portillo, A. (2016). Sociometric network structure and its association with methamphetamine use norms among homeless youth. *Social science research*, 58:292–308.

[48] Barocas, S., Guo, A., Kamar, E., Krones, J., Morris, M. R., Vaughan, J. W., Wadsworth, W. D., and Wallach, H. (2021a). *Designing Disaggregated Evaluations of AI Systems: Choices, Considerations, and Tradeoffs*, page 368–378. Association for Computing Machinery, New York, NY, USA.

[49] Barocas, S., Guo, A., Kamar, E., Krones, J., Morris, M. R., Vaughan, J. W., Wadsworth, W. D., and Wallach, H. (2021b). *Designing Disaggregated Evaluations of AI Systems: Choices, Considerations, and Tradeoffs*, page 368–378. Association for Computing Machinery, New York, NY, USA.

[50] Barocas, S., Hardt, M., and Narayanan, A. (2019). *Fairness and Machine Learning*. fairmlbook.org. http://www.fairmlbook.org.

[51] Baron, J. R., Lewis, D. D., and Oard, D. W. (2006). Trec 2006 legal track overview. In *Proceedings of the Fifteenth Text REtrieval Conference*.

[52] Baudry, J.-P., Cardoso, M., Celeux, G., Amorim, M. J., and Ferreira, A. S. (2015). Enhancing the selection of a model-based clustering with external categorical variables. *Advances in Data Analysis and Classification*, 9(2):177–196.

[53] Bauer, N. M. (2015). Emotional, sensitive, and unfit for office? gender stereotype activation and support female candidates. *Political Psychology*, 36(6):691–708.

[54] Beer, D. (2016). How should we do the history of big data? *Big Data & Society*, 3(1):2053951716646135.

[55] Behaghel, L., Crépon, B., and Gurgand, M. (2014). Private and public provision of counseling to job seekers: Evidence from a large controlled experiment. *American Economic Journal: Applied Economics*, 6(4):142–74.

[56] Belitz, C., Jiang, L., and Bosch, N. (2021). *Automating Procedurally Fair Feature Selection in Machine Learning*, page 379–389. Association for Computing Machinery, New York, NY, USA.

[57] Bella, A., Ferri, C., Hernández-Orallo, J., and Ramírez-Quintana, M. J. (2010). Quantification via probability estimators. In *Proceedings of the 11th IEEE International Conference on Data Mining (ICDM 2010)*, pages 737–742, Sydney, AU.

[58] Bender, E. M. and Friedman, B. (2018). Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

[59] Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

[60] Benenson, R., Popov, S., and Ferrari, V. (2019). Large-scale interactive object segmentation with human annotators. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11700–11709.

[61] Bera, S., Chakrabarty, D., Flores, N., and Negahbani, M. (2019). Fair algorithms for clustering. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 32, pages 4954–4965. Curran Associates, Inc.

[62] Beretta, E., Vetrò, A., Lepri, B., and Martin, J. C. D. (2021). Detecting discriminatory risk through data annotation based on bayesian inferences. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 794–804, New York, NY, USA. Association for Computing Machinery.

[63] Berinsky, A. J. and Mendelberg, T. (2005). The indirect effects of discredited stereotypes in judgments of jewish leaders. *American Journal of Political Science*, 49(4):845–864.

[64] Berk, R., Heidari, H., Jabbari, S., Joseph, M., Kearns, M., Morgenstern, J., Neel, S., and Roth, A. (2017). A convex framework for fair regression. KDD 2017 workshop: "Fairness, Accountability, and Transparency in Machine Learning (FAT/ML)".

[65] Bertin-Mahieux, T., Ellis, D. P. W., Whitman, B., and Lamere, P. (2011). The million song dataset. In *Proceedings of the 12th International Society for Music Information Retrieval Conference*, pages 591–596, Miami, United States. ISMIR.

[66] Bertrand, M. and Mullainathan, S. (2004). Are emily and greg more employable than lakisha and jamal? a field experiment on labor market discrimination. *American economic review*, 94(4):991–1013.

[67] Beutel, A., Chen, J., Doshi, T., Qian, H., Wei, L., Wu, Y., Heldt, L., Zhao, Z., Hong, L., Chi, E. H., and Goodrow, C. (2019a). Fairness in recommendation ranking through pairwise comparisons. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, page 2212–2220, New York, NY, USA. Association for Computing Machinery.

[68] Beutel, A., Chen, J., Doshi, T., Qian, H., Woodruff, A., Luu, C., Kreitmann, P., Bischof, J., and Chi, E. H. (2019b). Putting fairness principles into practice: Challenges, metrics, and improvements. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '19, page 453–459, New York, NY, USA. Association for Computing Machinery.

[69] Beutel, A., Chen, J., Zhao, Z., and Chi, E. H. (2017). Data decisions and theoretical implications when adversarially learning fair representations. KDD 2017 workshop: "Fairness, Accountability, and Transparency in Machine Learning (FAT/ML)".

[70] Biasion, D., Fabris, A., Silvello, G., and Susto, G. A. (2020). Gender bias in italian word embeddings. In *Proc. of the Seventh Italian Conference on Computational Linguistics*.

[71] Biega, A. J., Diaz, F., Ekstrand, M. D., Feldman, S., and Kohlmeier, S. (2021). Overview of the trec 2020 fair ranking track. *arXiv preprint arXiv:2108.05135*.

[72] Biega, A. J., Diaz, F., Ekstrand, M. D., and Kohlmeier, S. (2019). Overview of the trec 2019 fair ranking track. In *The Twenty-Eighth Text REtrieval Conference (TREC 2019) Proceedings*.

[73] Biega, A. J., Diaz, F., Ekstrand, M. D., and Kohlmeier, S. (2020). Overview of the trec 2019 fair ranking track. *arXiv preprint arXiv:2003.11650*.

[74] Biega, A. J., Gummadi, K. P., and Weikum, G. (2018a). Equity of attention: Amortizing individual fairness in rankings. In *Proc of 41st ACM SIGIR*, SIGIR '18, page 405–414, New York. Association for Computing Machinery.

[75] Biega, A. J., Gummadi, K. P., and Weikum, G. (2018b). Equity of attention: Amortizing individual fairness in rankings. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '18, page 405–414, New York, NY, USA. Association for Computing Machinery.

[76] Biswas, A. and Mukherjee, S. (2021). *Ensuring Fairness under Prior Probability Shifts*, page 414–424. Association for Computing Machinery, New York, NY, USA.

[77] Black, E. and Fredrikson, M. (2021). Leave-one-out unfairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 285–295, New York, NY, USA. Association for Computing Machinery.

[78] Black, E., Yeom, S., and Fredrikson, M. (2020). Fliptest: Fairness testing via optimal transport. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 111–121, New York, NY, USA. Association for Computing Machinery.

[79] Blodget, S. L. and O'Connor, B. (2017). Racial disparity in natural language processing: A case study of social media african-american english. KDD 2017 workshop: "Fairness, Accountability, and Transparency in Machine Learning (FAT/ML)".

[80] Blodgett, S. L., Green, L., and O'Connor, B. (2016). Demographic dialectal variation in social media: A case study of African-American English. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1119–1130, Austin, Texas. Association for Computational Linguistics.

[81] Bobbitt-Zeher, D. (2011). Gender discrimination at work: Connecting gender stereotypes, institutional policies, and gender composition of workplace. *Gender & Society*, 25(6):764–786.

[82] Bogen, M., Rieke, A., and Ahmed, S. (2020). Awareness in practice: Tensions in access to sensitive attribute data for antidiscrimination. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 492–500, New York, NY, USA. Association for Computing Machinery.

[83] Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., and Kalai, A. T. (2016a). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 29, pages 4349–4357. Curran Associates, Inc.

[84] Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., and Kalai, A. T. (2016b). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in neural information processing systems*, pages 4349–4357.

[85] Boratto, L., Fenu, G., and Marras, M. (2021). Connecting user and item perspectives in popularity debiasing for collaborative recommendation. *Information Processing & Management*, 58(1):102387.

[86] Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J., and Yakhnenko, O. (2013). Translating embeddings for modeling multi-relational data. In Burges, C. J. C., Bottou, L., Welling, M., Ghahramani, Z., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

[87] Bordia, S. and Bowman, S. R. (2019). Identifying and reducing gender bias in word-level language models. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 7–15, Minneapolis, Minnesota. Association for Computational Linguistics.

[88] Borkan, D., Dixon, L., Sorensen, J., Thain, N., and Vasserman, L. (2019). Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion Proceedings of The 2019 World Wide Web Conference*, WWW '19, page 491–500, New York, NY, USA. Association for Computing Machinery.

[89] Bose, A. and Hamilton, W. (2019). Compositional fairness constraints for graph embeddings. In Chaudhuri, K. and Salakhutdinov, R., editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 715–724, Long Beach, California, USA. PMLR.

[90] Bower, A., Eftekhari, H., Yurochkin, M., and Sun, Y. (2021). Individually fair rankings. In *International Conference on Learning Representations*.

[91] Bower, A., Niss, L., Sun, Y., and Vargo, A. (2018). Debiasing representations by removing unwanted variation due to protected attributes. ICML 2018 workshop: "Fairness, Accountability, and Transparency in Machine Learning (FAT/ML)".

[92] Boyd, D. and Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, communication & society*, 15(5):662–679.

[93] Brennan, T., Dieterich, W., and Ehret, B. (2009). Evaluating the predictive validity of the compas risk and needs assessment system. *Criminal Justice and Behavior*, 36(1):21–40.

[94] Brockman, G., Cheung, V., Pettersson, L., Schneider, J., Schulman, J., Tang, J., and Zaremba, W. (2016). Openai gym. *arXiv preprint arXiv:1606.01540*.

[95] Brooks-Gunn, J., Liaw, F.-r., and Klebanov, P. K. (1992). Effects of early intervention on cognitive function of low birth weight preterm infants. *The Journal of pediatrics*, 120(3):350–359.

[96] Brozovsky, L. and Petricek, V. (2007). Recommender system for online dating service. *ArXiv*, abs/cs/0703042.

[97] Brožovský, L. (2006). Recommender system for a dating service. Master's thesis, Charles University in Prague, Prague, Czech Republic.

[98] Brubach, B., Chakrabarti, D., Dickerson, J., Khuller, S., Srinivasan, A., and Tsepenekas, L. (2020). A pairwise fair and community-preserving approach to k-center clustering. In III, H. D. and Singh, A., editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1178–1189, Virtual. PMLR.

[99] Brunet, M.-E., Alkalay-Houlihan, C., Anderson, A., and Zemel, R. (2019). Understanding the origins of bias in word embeddings. In Chaudhuri, K. and Salakhutdinov, R., editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 803–811, Long Beach, California, USA. PMLR.

[100] Buckley, C., Dimmick, D., Soboroff, I., and Voorhees, E. (2006). Bias and the limits of pooling. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 619–620.

[101] Buolamwini, J. and Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proceedings of the 1st ACM Conference on Fairness, Accountability and Transparency (FAT* 2018)*, pages 77–91, New York, US.

[102] Burke, R., Kontny, J., and Sonboli, N. (2018a). Synthetic attribute data for evaluating consumer-side fairness. RecSys 2018 workshop: "Workshop on Responsible Recommendation (FAT/Rec)".

[103] Burke, R., Sonboli, N., and Ordonez-Gauger, A. (2018b). Balanced neighborhoods for multi-sided fairness in recommendation. In Friedler, S. A. and Wilson, C., editors, *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 202–214, New York, NY, USA. PMLR.

[104] Buyl, M. and De Bie, T. (2020). DeBayes: a Bayesian method for debiasing network embeddings. In III, H. D. and Singh, A., editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1220–1229, Virtual. PMLR.

[105] Cai, W., Gaebler, J., Garg, N., and Goel, S. (2020). Fair allocation through selective information acquisition. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, AIES '20, page 22–28, New York, NY, USA. Association for Computing Machinery.

[106] Caldas, S., Duddu, S. M. K., Wu, P., Li, T., Konečnỳ, J., McMahan, H. B., Smith, V., and Talwalkar, A. (2018). Leaf: A benchmark for federated settings. *arXiv preprint arXiv:1812.01097*.

[107] Calders, T., Kamiran, F., and Pechenizkiy, M. (2009). Building classifiers with independency constraints. In *2009 IEEE International Conference on Data Mining Workshops*, pages 13–18.

[108] Calders, T. and Verwer, S. (2010). Three naive bayes approaches for discrimination-free classification. *Data Min. Knowl. Discov.*, 21(2):277–292.

[109] Caliskan, A., Bryson, J., and Narayanan, A. (2017a). Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

[110] Caliskan, A., Bryson, J. J., and Narayanan, A. (2017b). Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

[111] Calmon, F., Wei, D., Vinzamuri, B., Natesan Ramamurthy, K., and Varshney, K. R. (2017). Optimized pre-processing for discrimination prevention. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30, pages 3992–4001. Curran Associates, Inc.

[112] Canetti, R., Cohen, A., Dikkala, N., Ramnarayan, G., Scheffler, S., and Smith, A. (2019). From soft classifiers to hard decisions: How fair can we be? In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 309–318, New York, NY, USA. Association for Computing Machinery.

[113] Caragiannis, I., Kurokawa, D., Moulin, H., Procaccia, A. D., Shah, N., and Wang, J. (2016). The unreasonable fairness of maximum nash welfare. In *Proceedings of the 2016 ACM Conference on Economics and Computation*, EC '16, page 305–322, New York, NY, USA. Association for Computing Machinery.

[114] Cardoso, R. L., Meira Jr., W., Almeida, V., and Zaki, M. J. (2019). A framework for benchmarking discrimination-aware models in machine learning. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '19, page 437–444, New York, NY, USA. Association for Computing Machinery.

[115] Carpineto, C., D'Amico, M., and Romano, G. (2012). Evaluating subtopic retrieval methods: Clustering versus diversification of search results. *Information Processing & Management*, 48(2):358–373.

[116] Carterette, B. (2011). System effectiveness, user models, and user utility: A conceptual framework for investigation. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '11, page 903–912, New York, NY, USA. Association for Computing Machinery.

[117] Carvalho, M. and Lodi, A. (2019). Game theoretical analysis of kidney exchange programs. *arXiv preprint arXiv:1911.09207*.

[118] Castelnovo, A., Crupi, R., Greco, G., Regoli, D., Penco, I. G., and Cosentini, A. C. (2022). A clarification of the nuances in the fairness metrics landscape. *Scientific Reports*, 12(1):1–21.

[119] Caton, S. and Haas, C. (2020). Fairness in machine learning: A survey. *arXiv preprint arXiv:2010.04053*.

[120] Celis, E., Keswani, V., Straszak, D., Deshpande, A., Kathuria, T., and Vishnoi, N. (2018). Fair and diverse DPP-based data summarization. In Dy, J. and Krause, A., editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 716–725, Stockholmsmässan, Stockholm Sweden. PMLR.

[121] Celis, E., Mehrotra, A., and Vishnoi, N. (2019a). Toward controlling discrimination in online ad auctions. In Chaudhuri, K. and Salakhutdinov, R., editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 4456–4465, Long Beach, California, USA. PMLR.

[122] Celis, E., Mehrotra, A., and Vishnoi, N. (2019b). Toward controlling discrimination in online ad auctions. In *Proc. of ICML 2019*, pages 4456–4465.

[123] Celis, L. E., Deshpande, A., Kathuria, T., and Vishnoi, N. K. (2016). How to be fair and diverse? DTL 2016 workshop: "Fairness, Accountability, and Transparency in Machine Learning (FAT/ML)".

[124] Celis, L. E., Huang, L., Keswani, V., and Vishnoi, N. K. (2019c). Classification with fairness constraints: A meta-algorithm with provable guarantees. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 319–328, New York, NY, USA. Association for Computing Machinery.

[125] Celis, L. E. and Keswani, V. (2020). Implicit diversity in image summarization. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2):1–28.

[126] Celis, L. E., Keswani, V., and Vishnoi, N. (2020a). Data preprocessing to mitigate bias: A maximum entropy based approach. In III, H. D. and Singh, A., editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1349–1359, Virtual. PMLR.

[127] Celis, L. E., Mehrotra, A., and Vishnoi, N. K. (2020b). Interventions for ranking in the presence of implicit bias. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 369–380, New York, NY, USA. Association for Computing Machinery.

[128] Celma, O. (2010). *Music Recommendation and Discovery in the Long Tail*. Springer.

[129] Census Bureau (2019). Current population survey. Accessed = 2020-02-12.

[130] Chaibub Neto, E. (2020). A causal look at statistical definitions of discrimination. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '20, page 873–881, New York, NY, USA. Association for Computing Machinery.

[131] Chakraborty, A., Patro, G. K., Ganguly, N., Gummadi, K. P., and Loiseau, P. (2019). Equality of voice: Towards fair representation in crowdsourced top-k recommendations. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 129–138, New York, NY, USA. Association for Computing Machinery.

[132] Chapados, N., Bengio, Y., Vincent, P., Ghosn, J., Dugas, C., Takeuchi, I., and Meng, L. (2001). Estimating car insurance premia: A case study in high-dimensional data inference. In *Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic*, NIPS'01, page 1369–1376, Cambridge, MA, USA. MIT Press.

[133] Chapelle, O. and Chang, Y. (2010). Yahoo! learning to rank challenge overview. In *Proceedings of the 2010 International Conference on Yahoo! Learning to Rank Challenge - Volume 14*, YLRC'10, page 1–24. JMLR.org.

[134] Chaudhari, H. A., Lin, S., and Linda, O. (2020). A general framework for fairness in multistakeholder recommendations. RecSys 2020 workshop: "3rd FAccTRec Workshop on Responsible Recommendation".

[135] Chelba, C., Mikolov, T., Schuster, M., Ge, Q., Brants, T., Koehn, P., and Robinson, T. (2014). One billion word benchmark for measuring progress in statistical language modeling. In *INTERSPEECH-2014*.

[136] Chen, B., Deng, W., and Shen, H. (2018a). Virtual class enhanced discriminative embedding learning. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

[137] Chen, C.-W., Lamere, P., Schedl, M., and Zamani, H. (2018b). Recsys challenge 2018: Automatic music playlist continuation. RecSys '18, page 527–528, New York, NY, USA. Association for Computing Machinery.

[138] Chen, I., Johansson, F. D., and Sontag, D. (2018c). Why is my classifier discriminatory? In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

[139] Chen, J., Kallus, N., Mao, X., Svacha, G., and Udell, M. (2019a). Fairness under unawareness: Assessing disparity when protected class is unobserved. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 339–348, New York, NY, USA. Association for Computing Machinery.

[140] Chen, J., Kallus, N., Mao, X., Svacha, G., and Udell, M. (2019b). Fairness under unawareness: Assessing disparity when protected class is unobserved. In *Proceedings of the 2nd ACM Conference on Fairness, Accountability, and Transparency (FAT* 2019)*, page 339–348, Atlanta, US.

[141] Chen, L., Ma, R., Hannák, A., and Wilson, C. (2018d). Investigating the impact of gender on rank in resume search engines. In *Proc. of the 2018 chi conference on human factors in computing systems*, pages 1–14.

[142] Chen, X., Fain, B., Lyu, L., and Munagala, K. (2019c). Proportionally fair clustering. In Chaudhuri, K. and Salakhutdinov, R., editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 1032–1041, Long Beach, California, USA. PMLR.

[143] Chen, Y., Mahoney, C., Grasso, I., Wali, E., Matthews, A., Middleton, T., Njie, M., and Matthews, J. (2021). *Gender Bias and Under-Representation in Natural Language Processing Across Human Languages*, page 24–34. Association for Computing Machinery, New York, NY, USA.

[144] Cheng, P., Hao, W., Yuan, S., Si, S., and Carin, L. (2021a). Fairfil: Contrastive neural debiasing method for pretrained text encoders. In *International Conference on Learning Representations*.

[145] Cheng, V., Suriyakumar, V. M., Dullerud, N., Joshi, S., and Ghassemi, M. (2021b). Can you fake it until you make it? impacts of differentially private synthetic data on downstream classification fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 149–160, New York, NY, USA. Association for Computing Machinery.

[146] Chierichetti, F., Kumar, R., Lattanzi, S., and Vassilvitskii, S. (2017). Fair clustering through fairlets. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30, pages 5029–5037. Curran Associates, Inc.

[147] Chiplunkar, A., Kale, S., and Ramamoorthy, S. N. (2020). How to solve fair k-center in massive data models. In III, H. D. and Singh, A., editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1877–1886, Virtual. PMLR.

[148] Chiusi, F., Fischer, S., Kayser-Bril, N., and Spielkamp, M. (2020). Automating society.

[149] Cho, J., Hwang, G., and Suh, C. (2020). A fair classifier using kernel density estimation. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 15088–15099. Curran Associates, Inc.

[150] Cho, W. I., Kim, J., Yang, J., and Kim, N. S. (2021). Towards cross-lingual generalization of translation gender bias. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 449–457, New York, NY, USA. Association for Computing Machinery.

[151] Choi, K., Grover, A., Singh, T., Shu, R., and Ermon, S. (2020a). Fair generative modeling via weak supervision. In III, H. D. and Singh, A., editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1887–1898, Virtual. PMLR.

[152] Choi, Y., Dang, M., and den Broeck, G. V. (2020b). Group fairness by probabilistic modeling with latent fair decisions. NeurIPS 2020 workshop: "Algorithmic Fairness through the Lens of Causality and Interpretability (AFCI)".

[153] Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5 2:153–163.

[154] Chouldechova, A. and G'Sell, M. (2017). Fairer and more accurate, but for whom? KDD 2017 workshop: "Fairness, Accountability, and Transparency in Machine Learning (FAT/ML)".

[155] Chouldechova, A. and Roth, A. (2020). A snapshot of the frontiers of fairness in machine learning. *Commun. ACM*, 63(5):82–89.

[156] Chuang, C.-Y. and Mroueh, Y. (2021). Fair mixup: Fairness via interpolation. In *International Conference on Learning Representations*.

[157] Chzhen, E., Denis, C., Hebiri, M., Oneto, L., and Pontil, M. (2019). Leveraging labeled and unlabeled data for consistent fair binary classification. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 32, pages 12760–12770. Curran Associates, Inc.

[158] Chzhen, E., Denis, C., Hebiri, M., Oneto, L., and Pontil, M. (2020a). Fair regression via plug-in estimator and recalibration with statistical guarantees. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

[159] Chzhen, E., Denis, C., Hebiri, M., Oneto, L., and Pontil, M. (2020b). Fair regression with wasserstein barycenters. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 7321–7331. Curran Associates, Inc.

[160] Clarke, C. L., Kolla, M., Cormack, G. V., Vechtomova, O., Ashkan, A., Büttcher, S., and MacKinnon, I. (2008). Novelty and diversity in information retrieval evaluation. In *Proc. of SIGIR 2008*, pages 659–666.

[161] Cleverdon, C. (1960). Report on the first stage of an investigation into the comparative efficiency of indexing systems. Technical report.

[162] Cleverdon, C. (1962). Report on the testing and analysis of an investigation into the comparative efficiency of indexing systems. Technical report.

[163] Cleverdon, C. (1997). *The Cranfield Tests on Index Language Devices*, page 47–59. Morgan Kaufmann Publishers Inc.

[164] Cohany, S. R., Polivka, A. E., and Rothgeb, J. M. (1994). Revisions in the current population survey effective january 1994. *Emp. & Earnings*, 41:13.

[165] Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., and Huq, A. (2017). Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '17, page 797–806, New York, NY, USA. Association for Computing Machinery.

[166] Cormack, G. (2007). Trec 2007 spam track overview. In *Proceedings of the Sixteenth Text REtrieval Conference (TREC 2007)*.

[167] Cortez, P. and Silva, A. M. G. (2008). Using data mining to predict secondary school student performance. In *Proceedings of 5th FUture BUsiness TEChnology Conference*.

[168] Cosconati, M. (2018). No news is good news: moral hazard in oligopolistic insurance markets. Quaderno IVASS 10, IVASS, Rome.

[169] Cosconati, M., Medori, V., Serafini, D., Scialanga, G. L., Visani, C., Ianni, A., and Matarazzo, L. (2020). Iper: L'andamento dei prezzi effettivi per la garanzia r.c.auto nel secondo trimestre 2020. Bollettino Statistico 8, IVASS, Rome.

[170] Coston, A., Guha, N., Ouyang, D., Lu, L., Chouldechova, A., and Ho, D. E. (2021). Leveraging administrative data for bias audits: Assessing disparate coverage with mobility data for covid-19 policy. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 173–184, New York, NY, USA. Association for Computing Machinery.

[171] Coston, A., Mishler, A., Kennedy, E. H., and Chouldechova, A. (2020). Counterfactual risk assessments, evaluation, and fairness. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 582–593, New York, NY, USA. Association for Computing Machinery.

[172] Coston, A., Ramamurthy, K. N., Wei, D., Varshney, K. R., Speakman, S., Mustahsan, Z., and Chakraborty, S. (2019a). Fair transfer learning with missing protected attributes. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '19, page 91–98, New York, NY, USA. Association for Computing Machinery.

[173] Coston, A., Ramamurthy, K. N., Wei, D., Varshney, K. R., Speakman, S., Mustahsan, Z., and Chakraborty, S. (2019b). Fair transfer learning with missing protected attributes. In *Proceedings of the 2nd AAAI/ACM Conference on AI, Ethics, and Society (AIES 2019)*, pages 91–98, Honolulu, US.

[174] Cotter, A., Gupta, M., Jiang, H., Srebro, N., Sridharan, K., Wang, S., Woodworth, B., and You, S. (2018). Training fairness-constrained classifiers to generalize. ICML 2018 workshop: "Fairness, Accountability, and Transparency in Machine Learning (FAT/ML)".

[175] Cotter, A., Gupta, M., Jiang, H., Srebro, N., Sridharan, K., Wang, S., Woodworth, B., and You, S. (2019). Training well-generalizing classifiers for fairness metrics and other data-dependent constraints. In Chaudhuri, K. and Salakhutdinov, R., editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 1397–1405, Long Beach, California, USA. PMLR.

[176] Council of the EU (2004). Implementing the principle of equal treatment between men and women in the access to and supply of goods and services l-373-37. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:32004L0113.

[177] Cover, T. and Thomas, J. (2012). *Elements of information theory*. John Wiley & Sons.

[178] Craswell, N., Zoeter, O., Taylor, M., and Ramsey, B. (2008). An experimental comparison of click position-bias models. In *Proceedings of the 2008 International Conference on Web Search and Data Mining*, WSDM '08, page 87–94, New York, NY, USA. Association for Computing Machinery.

[179] Crawford, K. (2021). *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*. Yale University Press.

[180] Crawford, K., Gray, M. L., and Miltner, K. (2014). Big datal critiquing big data: Politics, ethics, epistemologyl special section introduction. *International Journal of Communication*, 8:10.

[181] Crawford, K. and Paglen, T. (2021). Excavating ai: the politics of images in machine learning training sets.

[182] Creager, E., Jacobsen, J.-H., and Zemel, R. (2021). Exchanging lessons between algorithmic fairness and domain generalization. NeurIPS 2020 workshop: “Algorithmic Fairness through the Lens of Causality and Interpretability (AFCI)".

[183] Creager, E., Madras, D., Jacobsen, J.-H., Weis, M., Swersky, K., Pitassi, T., and Zemel, R. (2019). Flexibly fair representation learning by disentanglement. In Chaudhuri, K. and Salakhutdinov, R., editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 1436–1445, Long Beach, California, USA. PMLR.

[184] Creager, E., Madras, D., Pitassi, T., and Zemel, R. (2020). Causal modeling for fairness in dynamical systems. In III, H. D. and Singh, A., editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 2185–2195, Virtual. PMLR.

[185] Cuddy, A. J., Fiske, S. T., and Glick, P. (2004). When professionals become mothers, warmth doesn’t cut the ice. *Journal of Social issues*, 60(4):701–718.

[186] Cvencek, D., Meltzoff, A. N., and Greenwald, A. G. (2011). Math-gender stereotypes in elementary school children. *Child Development*, 82(3):766–779.

[187] D’Amour, A., Srinivasan, H., Atwood, J., Baljekar, P., Sculley, D., and Halpern, Y. (2020). Fairness is not static: Deeper understanding of long term fairness via simulation studies. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* ’20, page 525–534, New York, NY, USA. Association for Computing Machinery.

[188] Danziger, S., Levav, J., and Avnaim-Pesso, L. (2011). Extraneous factors in judicial decisions. *Proceedings of the National Academy of Sciences*, 108(17):6889–6892.

[189] Dash, A., Chakraborty, A., Ghosh, S., Mukherjee, A., and Gummadi, K. P. (2021). When the umpire is also a player: Bias in private label product recommendations on e-commerce marketplaces. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT ’21, page 873–884, New York, NY, USA. Association for Computing Machinery.

[190] Datta, S., Posada, J., Olson, G., Li, W., O’Reilly, C., Balraj, D., Mesterhazy, J., Pallas, J., Desai, P., and Shah, N. (2020). A new paradigm for accelerating clinical data science at stanford medicine. *arXiv preprint arXiv:2003.10534*.

[191] David, K. E., Liu, Q., and Fong, R. (2020). Debiasing convolutional neural networks via meta orthogonalization. NeurIPS 2020 workshop: “Algorithmic Fairness through the Lens of Causality and Interpretability (AFCI)".

[192] Davidson, I. and Ravi, S. S. (2020). A framework for determining the fairness of outlier detection. In *ECAI 2020*, pages 2465–2472. IOS Press.

[193] Davidson, T., Warmsley, D., Macy, M. W., and Weber, I. (2017). Automated hate speech detection and the problem of offensive language. In *Proceedings of the Eleventh International Conference on Web and Social Media, ICWSM 2017, Montréal, Québec, Canada, May 15-18, 2017*, pages 512–515. AAAI Press.

[194] De-Arteaga, M., Romanov, A., Wallach, H., Chayes, J., Borgs, C., Chouldechova, A., Geyik, S., Kenthapadi, K., and Kalai, A. T. (2019a). Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *Proc. of FAT* 2019*, pages 120–128. ACM.

[195] De-Arteaga, M., Romanov, A., Wallach, H., Chayes, J., Borgs, C., Chouldechova, A., Geyik, S., Kenthapadi, K., and Kalai, A. T. (2019b). Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 120–128, New York, NY, USA. Association for Computing Machinery.

[196] DeFranza, D., Mishra, H., and Mishra, A. (2020). How language shapes prejudice against women: An examination across 45 world languages. *Journal of Personality and Social Psychology*.

[197] Delobelle, P., Temple, P., Perrouin, G., Frénay, B., Heymans, P., and Berendt, B. (2020). Ethical adversaries: Towards mitigating unfairness with adversarial machine learning. ECMLPKDD 2020 workshop: "BIAS 2020: Bias and Fairness in AI".

[198] Deng, J., Dong, W., Socher, R., Li, L., Kai Li, and Li Fei-Fei (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255.

[199] Denton, E., Hanna, A., Amironesei, R., Smart, A., Nicole, H., and Scheuerman, M. K. (2020). Bringing the people back in: Contesting benchmark machine learning datasets. *arXiv preprint arXiv:2007.07399*.

[200] Deshpande, K. V., Pan, S., and Foulds, J. R. (2020). Mitigating demographic bias in ai-based resume filtering. In *Adjunct Publication of the 28th ACM Conference on User Modeling, Adaptation and Personalization*, UMAP '20 Adjunct, page 268–275, New York, NY, USA. Association for Computing Machinery.

[201] Dessì, D., Fenu, G., Marras, M., and Reforgiato Recupero, D. (2018). Coco: Semantic-enriched collection of online courses at scale with experimental use cases. In *World Conference on Information Systems and Technologies*, pages 1386–1396. Springer.

[202] Detrano, R., Janosi, A., Steinbrunn, W., Pfisterer, M., Schmid, J.-J., Sandhu, S., Guppy, K. H., Lee, S., and Froelicher, V. (1989). International application of a new probability algorithm for the diagnosis of coronary artery disease. *The American Journal of Cardiology*, 64(5):304–310.

[203] Dev, S. and Phillips, J. (2019). Attenuating bias in word vectors. In Chaudhuri, K. and Sugiyama, M., editors, *Proc. of Machine Learning Research*, volume 89, pages 879–887. PMLR.

[204] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

[205] DG MOVE - EU Directorate-General for Mobility and Transport (2019). Statistical pocketbook 2019. https://ec.europa.eu/transport/facts-fundings/statistics/pocketbook-2019_en.

[206] Dhamala, J., Sun, T., Kumar, V., Krishna, S., Pruksachatkun, Y., Chang, K.-W., and Gupta, R. (2021). Bold: Dataset and metrics for measuring biases in open-ended language generation. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 862–872, New York, NY, USA. Association for Computing Machinery.

[207] Di Nunzio, G. M., Fabris, A., Silvello, G., and Susto, G. A. (2021). Incentives for item duplication under fair ranking policies. In Boratto, L., Faralli, S., Marras, M., and Stilo, G., editors, *Advances in Bias and Fairness in Information Retrieval*, pages 64–77, Cham. Springer International Publishing.

[208] Diana, E., Gill, W., Kearns, M., Kenthapadi, K., and Roth, A. (2021). *Minimax Group Fairness: Algorithms and Experiments*, page 66–76. Association for Computing Machinery, New York, NY, USA.

[209] Diaz, F., Mitra, B., Ekstrand, M. D., Biega, A. J., and Carterette, B. (2020). Evaluating stochastic rankings with expected exposure. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, CIKM '20, page 275–284, New York, NY, USA. Association for Computing Machinery.

[210] DiCiccio, C., Vasudevan, S., Basu, K., Kenthapadi, K., and Agarwal, D. (2020). *Evaluating Fairness Using Permutation Tests*, page 1467–1477. Association for Computing Machinery, New York, NY, USA.

[211] Dickens, C., Singh, R., and Getoor, L. (2020). Hyperfair: A soft approach to integrating fairness criteria. RecSys 2020 workshop: "3rd FAccTRec Workshop on Responsible Recommendation".

[212] Dieterich, W., Mendoza, C., and Brennan, T. (2016). Compas risk scales: Demonstrating accuracy equity and predictive parity.

[213] Ding, F., Hardt, M., Miller, J., and Schmidt, L. (2021). Retiring adult: New datasets for fair machine learning. *Advances in Neural Information Processing Systems*, 34.

[214] Ditzler, G., Roveri, M., Alippi, C., and Polikar, R. (2015). Learning in nonstationary environments: A survey. *IEEE Computational Intelligence*, 10(4):12–25.

[215] Dixon, L., Li, J., Sorensen, J., Thain, N., and Vasserman, L. (2018). Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18, page 67–73, New York, NY, USA. Association for Computing Machinery.

[216] Donini, M., Oneto, L., Ben-David, S., Shawe-Taylor, J. S., and Pontil, M. (2018). Empirical risk minimization under fairness constraints. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 31, pages 2791–2801. Curran Associates, Inc.

[217] Dressel, J. and Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. *Science advances*, 4(1):eaao5580.

[218] Duarte, M. F. and Hu, Y. H. (2004). Vehicle classification in distributed sensor networks. *Journal of Parallel and Distributed Computing*, 64(7):826 – 838. Computing and Communication in Distributed Sensor Networks.

[219] Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. (2012a). Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference (ITCS 2012)*, pages 214–226, Cambridge, US.

[220] Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. (2012b). Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, ITCS '12, page 214–226, New York, NY, USA. Association for Computing Machinery.

[221] Dwork, C., Immorlica, N., Kalai, A. T., and Leiserson, M. (2017). Decoupled classifiers for fair and efficient machine learning. KDD 2017 workshop: "Fairness, Accountability, and Transparency in Machine Learning (FAT/ML)".

[222] Dwork, C., Immorlica, N., Kalai, A. T., and Leiserson, M. (2018). Decoupled classifiers for group-fair and efficient machine learning. In Friedler, S. A. and Wilson, C., editors, *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 119–133, New York, NY, USA. PMLR.

[223] D'Aurizio, L., Mattei, P., and Mosco, V. (2021). L'attività assicurativa nel comparto auto (2015 – 2020). Bollettino Statistico 16, IVASS, Rome.

[224] Eagly, A. H., Nater, C., Miller, D. I., Kaufmann, M., and Sczesny, S. (2019). Gender stereotypes have changed: A cross-temporal meta-analysis of us public opinion polls from 1946 to 2018. *American psychologist*, 75(3):301–315.

[225] Ebner, N. C., Riediger, M., and Lindenberger, U. (2010). Faces—a database of facial expressions in young, middle-aged, and older women and men: Development and validation. *Behavior research methods*, 42(1):351–362.

[226] EC - European Commission (2012). Guidelines on the application of council directive 2004/113/ec to insurance, in the light of the judgment of the court of justice of the european union in case c-236/09 (test-achats) c-11/1. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52012XC0113(01).

[227] ECJ - European Court of Justice (2011). Association belge des consommateurs test-achats asbl v conseil des ministres (2011) c-236/09. https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX:62009CJ0236.

[228] Eidinger, E., Enbar, R., and Hassner, T. (2014). Age and gender estimation of unfiltered faces. *IEEE Transactions on Information Forensics and Security*, 9(12):2170–2179.

[229] Ekstrand, M. D., Burke, R., and Diaz, F. (2019). Fairness and discrimination in retrieval and recommendation. In *Proc. of SIGIR 2019*, page 1403–1404. ACM.

[230] Ekstrand, M. D., Das, A., Burke, R., and Diaz, F. (2021). Fairness and discrimination in information access systems. *arXiv preprint arXiv:2105.05779*.

[231] Ekstrand, M. D., Tian, M., Azpiazu, I. M., Ekstrand, J. D., Anuyah, O., McNeill, D., and Pera, M. S. (2018). All the cool kids, how do they fit in?: Popularity and demographic biases in recommender evaluation and effectiveness. In Friedler, S. A. and Wilson, C., editors, *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 172–186, New York, NY, USA. PMLR.

[232] El Emam, K., Arbuckle, L., Koru, G., Eze, B., Gaudette, L., Neri, E., Rose, S., Howard, J., and Gluck, J. (2012). De-identification methods for open health data: the case of the heritage health prize claims dataset. *Journal of medical Internet research*, 14(1):e33.

[233] El Halabi, M., Mitrović, S., Norouzi-Fard, A., Tardos, J., and Tarnawski, J. M. (2020). Fairness in streaming submodular maximization: Algorithms and hardness. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 13609–13622. Curran Associates, Inc.

[234] Ellemers, N. (2018). Gender stereotypes. *Annual Review of Psychology*, 69(1):275–298.

[235] Elliott, M. N., Morrison, P. A., Fremont, A., McCaffrey, D. F., Pantoja, P., and Lurie, N. (2009). Using the Census Bureau's surname list to improve estimates of race/ethnicity and associated disparities. *Health Services and Outcomes Research Methodology*, 9(2):69–83.

[236] Elzayn, H., Jabbari, S., Jung, C., Kearns, M., Neel, S., Roth, A., and Schutzman, Z. (2019). Fair algorithms for learning in allocation problems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 170–179, New York, NY, USA. Association for Computing Machinery.

[237] Epstein, L., Landes, W., and Posner, R. (2013). *The Behavior of Federal Judges: A Theoretical and Empirical Study of Rational Choice*. Harvard University Press.

[238] Equivant (2019). Practitioner's guide to compas core.

[239] Esmaeili, S., Brubach, B., Tsepenekas, L., and Dickerson, J. (2020). Probabilistic fair clustering. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 12743–12755. Curran Associates, Inc.

[240] Esuli, A., Fabris, A., Moreo, A., and Sebastiani, F. (2023). *Learning to Quantify*. Springer. In press.

[241] EU - European Union (2000). Charter of fundamental rights of the european union c-364/01. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32000X1218%2801%29.

[242] EU - European Union (2012). Charter of fundamental rights of the european union c-326/391. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:12012P/TXT.

[243] European Commission (2021). Proposal for a regulation of the european parliament and of the council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts.

[244] European Union (2016). Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/ec (general data protection regulation). https://eur-lex.europa.eu/eli/reg/2016/679/2016-05-04.

[245] Eurostat (2020). Migration and migrant population statistics. https://ec.europa.eu/eurostat/statistics-explained/index.php/Migration_and_migrant_population_statistics#Acquisitions_of_citizenship:_EU-27_Member_States_granted_citizenship_to_672_thousand_persons_in_2018.

[246] Fabbrizzi, S., Papadopoulos, S., Ntoutsi, E., and Kompatsiaris, I. (2021). A survey on bias in visual datasets. *arXiv preprint arXiv:2107.07919*.

[247] Fabris, A., Esuli, A., Moreo, A., and Sebastiani, F. (2021a). Measuring fairness under unawareness via quantification. *arXiv preprint arXiv:2109.08549*.

[248] Fabris, A., Messina, S., Silvello, G., and Susto, G. A. (2022a). Algorithmic fairness datasets: the story so far. *Data Mining and Knowledge Discovery*.

[249] Fabris, A., Messina, S., Silvello, G., and Susto, G. A. (2022b). Tackling documentation debt: A survey on algorithmic fairness datasets. In *Equity and Access in Algorithms, Mechanisms, and Optimization*, EAAMO '22, New York, NY, USA. Association for Computing Machinery.

[250] Fabris, A., Mishler, A., Gottardi, S., Carletti, M., Daicampi, M., Susto, G. A., and Silvello, G. (2021b). *Algorithmic Audit of Italian Car Insurance: Evidence of Unfairness in Access and Pricing*, page 458–468. Association for Computing Machinery, New York, NY, USA.

[251] Fabris, A., Purpura, A., Silvello, G., and Susto, G. A. (2020). Gender stereotype reinforcement: Measuring the gender bias conveyed by ranking algorithms. *Information Processing & Management*, 57(6):102377.

[252] Fabris, A., Purpura, A., Silvello, G., and Susto, G. A. (2021c). Measuring gender stereotype reinforcement in information retrieval systems. In *Proc. of the 12th Italian Information Retrieval Workshop*.

[253] Fabris, A., Silvello, G., Susto, G. A., and Biega, A. (2023). Pairwise fairness in ranking as a dissatisfaction measure. In *Proc. of the Sixteenth ACM International Conference on Web Search and Data Mining*, New York, NY, USA. ACM. In press.

[254] Farnad, G., Babaki, B., and Gendreau, M. (2020). A unifying framework for fairness-aware influence maximization. In *Companion Proceedings of the Web Conference 2020*, WWW '20, page 714–722, New York, NY, USA. Association for Computing Machinery.

[255] Farnadi, G., Babaki, B., and Carvalho, M. (2019). Enhancing fairness in kidney exchange program by ranking solutions. NeurIPS 2019 workshop: "Fair ML for Health".

[256] Farnadi, G., Kouki, P., Thompson, S. K., Srinivasan, S., and Getoor, L. (2018). A fairness-aware hybrid recommender system. RecSys 2018 workshop: "Workshop on Responsible Recommendation (FAT/Rec)".

[257] Fazio, R. H., Jackson, J. R., Dunton, B. C., and Williams, C. J. (1995). Variability in automatic activation as an unobtrusive measure of racial attitudes: A bona fide pipeline? *Journal of personality and social psychology*, 69(6):1013–1027.

[258] Federal Trade Commission (2021). Trade regulation rule on commercial surveillance.

[259] Fehrman, E., Egan, V., Gorban, A. N., Levesley, J., Mirkes, E. M., and Muhammad, A. K. (2019). *Personality Traits and Drug Consumption: A story told by data*. Springer.

[260] Fehrman, E., Muhammad, A. K., Mirkes, E. M., Egan, V., and Gorban, A. N. (2017). The five factor model of personality and evaluation of drug consumption risk. In Palumbo, F., Montanari, A., and Vichi, M., editors, *Data Science*, pages 231–242, Cham. Springer International Publishing.

[261] Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., and Venkatasubramanian, S. (2015). Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, page 259–268, New York, NY, USA. Association for Computing Machinery.

[262] Fernandes Vaz, A., Izbicki, R., and Bassi Stern, R. (2019). Quantification under prior probability shift: The ratio estimator and its extensions. *Journal of Machine Learning Research*, 20:79:1–79:33.

[263] Ferrante, M., Ferro, N., and Maistro, M. (2014). Injecting user models and time into precision via markov chains. In *Proc. of SIGIR 2014*, page 597–606. ACM.

[264] Ferraro, A., Bogdanov, D., Serra, X., and Yoon, J. (2019). Artist and style exposure bias in collaborative filtering based music recommendations. ISMIR 2019 workshop: "Workshop on Designing Human-Centric MIR Systems".

[265] Ferreira Jr, J. and Minikel, E. (2012). Measuring per mile risk for pay-as-you-drive automobile insurance. *Transportation research record*, 2297(1):97–103.

[266] Fish, B., Kun, J., and Lelkes, Á. (2015). Fair boosting : a case study. ICML 2015 workshop: "Fairness, Accountability, and Transparency in Machine Learning (FAT/ML)".

[267] Fisher, J., Palfrey, D., Christodoulopoulos, C., and Mittal, A. (2020). Measuring social bias in knowledge graph embeddings. AKBC 2020 workshop: "Bias in Automatic Knowledge Graph Construction".

[268] Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of eugenics*, 7(2):179–188.

[269] Fiske, S. T., Cuddy, A. J., Glick, P., and Xu, J. (2002). A model of (often mixed) stereotype content: Competence and warmth respectively follow from perceived status and competition. In *Journal of Personality and Social Psychology*, pages 878–902.

[270] Fisman, R., Iyengar, S., Kamenica, E., and Simonson, I. (2006). Gender differences in mate selection: Evidence from a speed dating experiment. *The Quarterly Journal of Economics*, 121:673–697.

[271] Fitzpatrick, T. B. (1988). The validity and practicality of sun-reactive skin types i through vi. *Archives of dermatology*, 124(6):869–871.

[272] Flanigan, B., Gölz, P., Gupta, A., and Procaccia, A. D. (2020). Neutralizing self-selection bias in sampling for sortition. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

[273] Florez, O. U. (2019). On the unintended social bias of training language generation models with data from local media. NeurIPS 2019 workshop: "Human-Centric Machine Learning".

[274] Floridi, L. (2011). *The philosophy of information*. Oxford University Press.

[275] Fogliato, R., Chouldechova, A., and G'Sell, M. (2020). Fairness evaluation in presence of biased noisy labels. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS 2020)*, pages 2325–2336, Virtual Event.

[276] Fogliato, R., Xiang, A., Lipton, Z., Nagin, D., and Chouldechova, A. (2021). On the validity of arrest as a proxy for offense: Race and the likelihood of arrest for violent crimes. In *Proceedings of the 4th AAAI/ACM Conference on AI, Ethics, and Society (AIES 2021)*, pages 100–111, Virtual Event.

[277] Forman, G. (2008). Quantifying counts and costs via classification. *Data Mining and Knowledge Discovery*, 17(2):164–206.

[278] Founta, A., Djouvas, C., Chatzakou, D., Leontiadis, I., Blackburn, J., Stringhini, G., Vakali, A., Sirivianos, M., and Kourtellis, N. (2018). Large scale crowdsourcing and characterization of twitter abusive behavior. In *Proceedings of the Twelfth International Conference on Web and Social Media, ICWSM 2018, Stanford, California, USA, June 25-28, 2018*, pages 491–500. AAAI Press.

[279] Framingham Heart Study (2021). Framingham heart study. offspring exam 10, omni 1 exam 5. research consent form.

[280] Frezal, S. and Barry, L. (2019). Fairness in uncertainty: Some limits and misinterpretations of actuarial fairness. *Journal of Business Ethics*, pages 1–10.

[281] Friedler, S. A., Scheidegger, C., and Venkatasubramanian, S. (2021). The (im)possibility of fairness: Different value systems require different mechanisms for fair decision making. *Commun. ACM*, 64(4):136–143.

[282] Friedler, S. A., Scheidegger, C., Venkatasubramanian, S., Choudhary, S., Hamilton, E. P., and Roth, D. (2019). A comparative study of fairness-enhancing interventions in machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 329–338, New York, NY, USA. Association for Computing Machinery.

[283] Fu, Z., Ren, K., Shu, J., Sun, X., and Huang, F. (2015). Enabling personalized search over encrypted outsourced data with efficiency improvement. *IEEE transactions on parallel and distributed systems*, 27(9):2546–2559.

[284] Galdon Clavell, G., Martín Zamorano, M., Castillo, C., Smith, O., and Matic, A. (2020). Auditing algorithms: On lessons learned and the risks of data minimization. In *Proceedings of the 3rd AAAI/ACM Conference on AI, Ethics, and Society (AIES 2020)*, pages 265–271, New York, US.

[285] Galhotra, S., Saisubramanian, S., and Zilberstein, S. (2021). *Learning to Generate Fair Clusters from Demonstrations*, page 491–501. Association for Computing Machinery, New York, NY, USA.

[286] Gao, R. and Shah, C. (2020). Toward creating a fairer ranking in search engine results. *Information Processing & Management*, 57(1):to appear.

[287] Garbin, C., Rajpurkar, P., Irvin, J., Lungren, M. P., and Marques, O. (2021). Structured dataset documentation: a datasheet for chexpert.

[288] Garg, N., Schiebinger, L., Jurafsky, D., and Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):3635–3644.

[289] Garg, S., Perot, V., Limtiaco, N., Taly, A., Chi, E. H., and Beutel, A. (2019). Counterfactual fairness in text classification through robustness. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '19, page 219–226, New York, NY, USA. Association for Computing Machinery.

[290] Gastwirth, J. L. and Miao, W. (2009). Formal statistical analysis of the data in disparate impact cases provides sounder inferences than the us government's 'four-fifths' rule: an examination of the statistical evidence in ricci v. destefano. *Law, Probability & Risk*, 8(2):171–191.

[291] Ge, H., Caverlee, J., and Lu, H. (2016). Taper: A contextual tensor-based approach for personalized expert recommendation. In *Proceedings of the 10th ACM Conference on Recommender Systems*, RecSys '16, page 261–268, New York, NY, USA. Association for Computing Machinery.

[292] Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., and Crawford, K. (2018). Datasheets for datasets. *arXiv preprint arXiv:1803.09010*.

[293] Geiger, R. S., Yu, K., Yang, Y., Dai, M., Qiu, J., Tang, R., and Huang, J. (2020). Garbage in, garbage out? do machine learning application papers in social computing report where human-labeled training data comes from? In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 325–336, New York, NY, USA. Association for Computing Machinery.

[294] Gelman, A., Fagan, J., and Kiss, A. (2007). An analysis of the new york city police department's "stop-and-frisk" policy in the context of claims of racial bias. *Journal of the American statistical association*, 102(479):813–823.

[295] Gerbner, G., Gross, L., Morgan, M., and Signorielli, N. (1986). Living with television: The dynamics of the cultivation process. *Perspectives on media effects*, 1986:17–40.

[296] Gerritse, E. (2019). Impact of debiasing word embeddings on information retrieval. In *Proc. of FDIA 2019*, page 54–59. CEUR-ws.

[297] Gerritse, E. J. and de Vries, A. P. (2020). Effect of debiasing on information retrieval. In Boratto, L., Faralli, S., Marras, M., and Stilo, G., editors, *Bias and Social Aspects in Search and Recommendation*, pages 35–42, Cham. Springer International Publishing.

[298] Ghadiri, M., Samadi, S., and Vempala, S. (2021). Socially fair k-means clustering. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 438–448, New York, NY, USA. Association for Computing Machinery.

[299] Ghazimatin, A., Kleindessner, M., Russel, C., Abedjan, Z., and Golebiowski, J. (2022). Measuring fairness of rankings under noisy sensitive information. In *Proceedings of the 5th ACM Conference on Fairness, Accountability, and Transparency (FAccT 2022)*, Seoul, Republic of Korea.

[300] Ghosh, A., Dutt, R., and Wilson, C. (2021a). When fair ranking meets uncertain inference. In *Proceedings of the 44th International ACM Conference on Research and Development in Information Retrieval (SIGIR 2021)*, pages 1033–1043, (Virtual Event).

[301] Ghosh, A., Genuit, L., and Reagan, M. (2021b). Characterizing intersectional group fairness with worst-case comparisons. In *Proceedings of the 2nd AAAI Workshop on Artificial Intelligence Diversity, Belonging, Equity, and Inclusion (AIDBEI 2021)*, pages 22–34, [Virtual Event].

[302] Ghosh, V. E. and Gilboa, A. (2014). What is a memory schema? a historical perspective on current neuroscience literature. *Neuropsychologia*, 53:104–114.

[303] Gianfrancesco, M. A., Tamang, S., Yazdany, J., and Schmajuk, G. (2018). Potential biases in machine learning algorithms using electronic health record data. *JAMA Internal Medicine*, 178(11):1544–1547.

[304] Ginart, A., Guan, M., Valiant, G., and Zou, J. Y. (2019). Making ai forget you: Data deletion in machine learning. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

[305] Go, A., Bhayani, R., and Huang, L. (2009). Twitter sentiment classification using distant supervision. *Processing*, 150.

[306] Goel, N., Amayuelas, A., Deshpande, A., and Sharma, A. (2020). The importance of modeling data missingness in algorithmic fairness: A causal perspective. NeurIPS 2020 workshop: “Algorithmic Fairness through the Lens of Causality and Interpretability (AFCI)".

[307] Goel, N. and Faltings, B. (2019). Crowdsourcing with fairness, diversity and budget constraints. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, AIES ’19, page 297–304, New York, NY, USA. Association for Computing Machinery.

[308] Goel, N., Yaghini, M., and Faltings, B. (2018). Non-discriminatory machine learning through convex fairness criteria. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES ’18, page 116, New York, NY, USA. Association for Computing Machinery.

[309] Goel, S., Perelman, M., Shroff, R., and Sklansky, D. (2017). Combatting police discrimination in the age of big data. *New Criminal Law Review*, 20(2):181–232.

[310] Goel, S., Rao, J. M., Shroff, R., et al. (2016). Precinct or prejudice? understanding racial disparities in new york city’s stop-and-frisk policy. *Annals of Applied Statistics*, 10(1):365–394.

[311] Goelz, P., Kahng, A., and Procaccia, A. D. (2019). Paradoxes in fair machine learning. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 32, pages 8342–8352. Curran Associates, Inc.

[312] Golbeck, J., Ashktorab, Z., Banjo, R. O., Berlinger, A., Bhagwan, S., Buntain, C., Cheakalos, P., Geller, A. A., Gergory, Q., Gnanasekaran, R. K., Gunasekaran, R. R., Hoffman, K. M., Hottle, J., Jienjitlert, V., Khare, S., Lau, R., Martindale, M. J., Naik, S., Nixon, H. L., Ramachandran, P., Rogers, K. M., Rogers, L., Sarin, M. S., Shahane, G., Thanki, J., Vengataraman, P., Wan, Z., and Wu, D. M. (2017). A large labeled corpus for online harassment research. In *Proceedings of the 2017 ACM on Web Science Conference*, WebSci ’17, page 229–233, New York, NY, USA. Association for Computing Machinery.

[313] Goldstein, H. (1991). Multilevel modelling of survey data. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 40(2):235–244.

[314] Gonen, H. and Goldberg, Y. (2019). Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. In *Proc. of NAACL 2019*, pages 609–614. ACL.

[315] Gong, S., Liu, X., and Jain, A. K. (2021). Mitigating face recognition bias via group adaptive classifier. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3414–3424.

[316] González, P., Castaño, A., Chawla, N. V., and del Coz, J. J. (2017). A review on quantification learning. *ACM Computing Surveys*, 50(5):74:1–74:40.

[317] González-Castro, V., Alaiz-Rodríguez, R., and Alegre, E. (2013). Class distribution estimation based on the Hellinger distance. *Information Sciences*, 218:146–164.

[318] Gorantla, S., Deshpande, A., and Louis, A. (2021). On the problem of underranking in group-fair ranking. In Meila, M. and Zhang, T., editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 3777–3787. PMLR.

[319] Gordaliza, P., Barrio, E. D., Fabrice, G., and Loubes, J.-M. (2019). Obtaining fairness using optimal transport theory. In Chaudhuri, K. and Salakhutdinov, R., editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2357–2365, Long Beach, California, USA. PMLR.

[320] Gordon, J., Babaeianjelodar, M., and Matthews, J. (2020). Studying political bias via word embeddings. In *Companion Proceedings of the Web Conference 2020*, WWW '20, page 760–764, New York, NY, USA. Association for Computing Machinery.

[321] Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., and Parikh, D. (2017). Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6904–6913.

[322] Graells-Garrido, E., Lalmas, M., and Menczer, F. (2015). First women, second sex: Gender bias in wikipedia. In *Proc. of HT 2015*, pages 165–174.

[323] Graffam, J., Shinkfield, A. J., and Hardcastle, L. (2008). The perceived employability of ex-prisoners and offenders. *International Journal of Offender Therapy and Comparative Criminology*, 52(6):673–685.

[324] Green, B. and Chen, Y. (2019). Disparate interactions: An algorithm-in-the-loop analysis of fairness in risk assessments. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 90–99, New York, NY, USA. Association for Computing Machinery.

[325] Greenwald, A. G., McGhee, D. E., and Schwartz, J. L. (1998a). Measuring individual differences in implicit cognition: the implicit association test. *Journal of personality and social psychology*, 74(6):1464–1480.

[326] Greenwald, A. G., McGhee, D. E., and Schwartz, J. L. (1998b). Measuring individual differences in implicit cognition: the implicit association test. *Journal of personality and social psychology*, 74(6):1464.

[327] Grgic-Hlaca, N., Zafar, M., Gummadi, K., and Weller, A. (2016a). The case for process fairness in learning: Feature selection for fair decision making. NeurIPS 2016 workshop: "Machine Learning and the Law".

[328] Grgic-Hlaca, N., Zafar, M. B., Gummadi, K. P., and Weller, A. (2016b). The case for process fairness in learning: Feature selection for fair decision making. In *NIPS Symposium on Machine Learning and the Law*, volume 1, page 2.

[329] Grömping, U. (2019). South German Credit Data: Correcting a Widely Used Data Set. Report. Technical report, Beuth University of Applied Sciences Berlin.

[330] Gulla, J. A., Zhang, L., Liu, P., Özgöbek, O., and Su, X. (2017). The adressa dataset for news recommendation. In *Proceedings of the International Conference on Web Intelligence*, WI '17, page 1042–1048, New York, NY, USA. Association for Computing Machinery.

[331] Gungor, A. (2018). Benchmarking authorship attribution techniques using over a thousand books by fifty victorian era novelists. Master's thesis, Purdue University.

[332] Guo, G., Zhang, J., and Yorke-Smith, N. (2016a). A novel evidence-based bayesian similarity measure for recommender systems. *ACM Trans. Web*, 10(2).

[333] Guo, J., Fan, Y., Ai, Q., and Croft, W. B. (2016b). A deep relevance matching model for ad-hoc retrieval. pages 55–64. ACM Press.

[334] Guo, W. and Caliskan, A. (2021). *Detecting Emergent Intersectional Biases: Contextualized Word Embeddings Contain a Distribution of Human-like Biases*, page 122–133. Association for Computing Machinery, New York, NY, USA.

[335] Guo, Y., Zhang, L., Hu, Y., He, X., and Gao, J. (2016c). Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In Leibe, B., Matas, J., Sebe, N., and Welling, M., editors, *Computer Vision – ECCV 2016*, pages 87–102, Cham. Springer International Publishing.

[336] Guvenir, H. A., Acar, B., Demiroz, G., and Cekin, A. (1997). A supervised machine learning algorithm for arrhythmia analysis. In *Computers in Cardiology 1997*, pages 433–436. IEEE.

[337] Halavais, A. (2008). *Search Engine Society*. Polity Press.

[338] Han, H. and Jain, A. K. (2014). Age, gender and race estimation from unconstrained face images.

[339] Hannak, A., Sapiezynski, P., Molavi Kakhki, A., Krishnamurthy, B., Lazer, D., Mislove, A., and Wilson, C. (2013). Measuring personalization of web search. In *Proceedings of the 22nd International Conference on World Wide Web*, WWW '13, page 527–538, New York, NY, USA. Association for Computing Machinery.

[340] Hannak, A., Soeller, G., Lazer, D., Mislove, A., and Wilson, C. (2014). Measuring price discrimination and steering on e-commerce web sites. In *Proceedings of the 2014 Conference on Internet Measurement Conference*, IMC '14, page 305–318, New York, NY, USA. Association for Computing Machinery.

[341] Hannák, A., Wagner, C., Garcia, D., Mislove, A., Strohmaier, M., and Wilson, C. (2017). Bias in online freelance marketplaces: Evidence from taskrabbit and fiverr. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, CSCW '17, page 1914–1933, New York, NY, USA. Association for Computing Machinery.

[342] Har-Peled, S. and Mahabadi, S. (2019). Near neighbor: Who is the fairest of them all? In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 32, pages 13176–13187. Curran Associates, Inc.

[343] Harb, E. and Lam, H. S. (2020). Kfc: A scalable approximation algorithm for k-center fair clustering. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 14509–14519. Curran Associates, Inc.

[344] Hardt, M., Price, E., Price, E., and Srebro, N. (2016a). Equality of opportunity in supervised learning. In Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 29, pages 3315–3323. Curran Associates, Inc.

[345] Hardt, M., Price, E., and Srebro, N. (2016b). Equality of opportunity in supervised learning. In *Proceedings of the 29th Annual Conference on Neural Information Processing Systems (NIPS 2016)*, pages 3323–3331, Barcelona, ES.

[346] Harman, D. (1992). The darpa tipster project. *SIGIR Forum*, 26(2):26–28.

[347] Harman, D. (2011). Information retrieval evaluation. *Synthesis Lectures on Information Concepts, Retrieval, and Services*, 3(2):1–119.

[348] Harper, F. M. and Konstan, J. A. (2015). The movielens datasets: History and context. *ACM Trans. Interact. Intell. Syst.*, 5(4).

[349] Harrington, S. E. and Niehaus, G. (1998). Race, redlining, and automobile insurance prices. *The Journal of Business*, 71(3):439–469.

[350] Hashimoto, T., Srivastava, M., Namkoong, H., and Liang, P. (2018). Fairness without demographics in repeated loss minimization. In Dy, J. and Krause, A., editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1929–1938, Stockholmsmässan, Stockholm Sweden. PMLR.

[351] Hasnain-Wynia, R. and Baker, D. W. (2006). Obtaining data on patient race, ethnicity, and primary language in health care organizations: Current challenges and proposed solutions. *Health Services Research*, 41(4p1):1501–1518.

[352] Hawking, D. and Craswell, N. (2005). The very large collection and web tracks. *TREC: Experiment and evaluation in information retrieval*.

[353] He, R., Kang, W.-C., and McAuley, J. (2017). Translation-based recommendation. In *Proceedings of the Eleventh ACM Conference on Recommender Systems*, RecSys '17, page 161–169, New York, NY, USA. Association for Computing Machinery.

[354] He, R. and McAuley, J. (2016). Ups and downs. *Proceedings of the 25th International Conference on World Wide Web*.

[355] He, Y., Burghardt, K., Guo, S., and Lerman, K. (2020a). Inherent trade-offs in the fair allocation of treatments. NeurIPS 2020 workshop: "Algorithmic Fairness through the Lens of Causality and Interpretability (AFCI)".

[356] He, Y., Burghardt, K., and Lerman, K. (2020b). A geometric solution to fair representations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, AIES '20, page 279–285, New York, NY, USA. Association for Computing Machinery.

[357] Heidari, H., Ferrari, C., Gummadi, K., and Krause, A. (2018). Fairness behind a veil of ignorance: A welfare analysis for automated decision making. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 31, pages 1265–1276. Curran Associates, Inc.

[358] Heidari, H., Loi, M., Gummadi, K. P., and Krause, A. (2019a). A moral framework for understanding fair ml through economic models of equality of opportunity. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 181–190, New York, NY, USA. Association for Computing Machinery.

[359] Heidari, H., Nanda, V., and Gummadi, K. (2019b). On the long-term impact of algorithmic decision policies: Effort unfairness and feature segregation through social learning. In Chaudhuri, K. and Salakhutdinov, R., editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2692–2701, Long Beach, California, USA. PMLR.

[360] Hendricks, L. A., Burns, K., Saenko, K., Darrell, T., and Rohrbach, A. (2018). Women also snowboard: Overcoming bias in captioning models. In Ferrari, V., Hebert, M., Sminchisescu, C., and Weiss, Y., editors, *Computer Vision – ECCV 2018*, pages 793–811, Cham. Springer International Publishing.

[361] Hentschel, T., Heilman, M. E., and Peus, C. V. (2019). The multiple dimensions of gender stereotypes: a current look at men's and women's characterizations of others and themselves. *Frontiers in psychology*, 10(11).

[362] Hersh, W., Turpin, A., Price, S., Chan, B., Kramer, D., Sacherek, L., and Olson, D. (2000). Do batch and user evaluations give the same results? In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 17–24.

[363] Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., and Lerchner, A. (2017). beta-vae: Learning basic visual concepts with a constrained variational framework. In *ICLR*.

[364] High-Level Expert Group on AI (2019). Ethics guidelines for trustworthy ai.

[365] Hinton, P. (2017). Implicit stereotypes and the predictive brain: cognition and culture in "biased" person perception. *Palgrave Communications*, 3(1):1–9.

[366] Holland, S., Hosny, A., Newman, S., Joseph, J., and Chmielinski, K. (2018). The dataset nutrition label: A framework to drive higher data quality standards. *arXiv preprint arXiv:1805.03677*.

[367] Hollywood, J., McKay, K., Woods, D., and Agniel, D. (2019). Real time crime centers in chicago.

[368] Holmes, M. D., Smith, B. W., Freng, A. B., and Muñoz, E. A. (2008). Minority threat, crime control, and police resource allocation in the southwestern united states. *Crime & Delinquency*, 54(1):128–152.

[369] Holstein, K., Wortman Vaughan, J., Daumé III, H., Dudik, M., and Wallach, H. (2019a). Improving fairness in machine learning systems: What do industry practitioners need? In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI 2019)*, pages 1–16, Glasgow, UK.

[370] Holstein, K., Wortman Vaughan, J., Daumé III, H., Dudik, M., and Wallach, H. (2019b). Improving fairness in machine learning systems: What do industry practitioners need? In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI 2019)*, pages 1–16, Glasgow, UK.

[371] Hong, S. and Kim, N. (2018). Will the internet promote democracy? search engines, concentration of online news readership, and e-democracy. *Journal of Information Technology & Politics*, 15(4):388–399.

[372] Houvardas, J. and Stamatatos, E. (2006). N-gram feature selection for authorship identification. In *International conference on artificial intelligence: Methodology, systems, and applications*, pages 77–86. Springer.

[373] Hu, L. and Chen, Y. (2020). Fair classification and social welfare. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 535–545, New York, NY, USA. Association for Computing Machinery.

[374] Hu, L., Immorlica, N., and Vaughan, J. W. (2019). The disparate effects of strategic manipulation. In *Proceedings of the 2nd ACM Conference on Fairness, Accountability, and Transparency (FAT* 2019)*, pages 259–268, Atlanta, US.

[375] Hu, Y., Wu, Y., Zhang, L., and Wu, X. (2020). Fair multiple decision making through soft interventions. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

[376] Huan, W., Wu, Y., Zhang, L., and Wu, X. (2020). Fairness through equality of effort. In *Companion Proceedings of the Web Conference 2020*, WWW '20, page 743–751, New York, NY, USA. Association for Computing Machinery.

[377] Huang, G. B., Ramesh, M., Berg, T., and Learned-Miller, E. (2007). Labeled faces in the wild: A database for studying face recognition in unconstrained environments.

[378] Huang, L., Jiang, S., and Vishnoi, N. (2019). Coresets for clustering with fairness constraints. In *Advances in Neural Information Processing Systems*, pages 7589–7600.

[379] Huang, L. and Vishnoi, N. (2019). Stable and fair classification. In Chaudhuri, K. and Salakhutdinov, R., editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2879–2890, Long Beach, California, USA. PMLR.

[380] Huang, L., Wei, J., and Celis, E. (2020). Towards just, fair and interpretable methods for judicial subset selection. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, AIES '20, page 293–299, New York, NY, USA. Association for Computing Machinery.

[381] Hull, J. (1994). A database for handwritten text recognition research. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(5):550–554.

[382] Hullman, J., Kapoor, S., Nanayakkara, P., Gelman, A., and Narayanan, A. (2022). The worst of both worlds: A comparative analysis of errors in learning from data in psychology and machine learning. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '22, page 335–348, New York, NY, USA. Association for Computing Machinery.

[383] Hussain, S., Dahan, N. A., Ba-Alwib, F. M., and Ribata, N. (2018). Educational data mining and analysis of students' academic performance using weka. *Indonesian Journal of Electrical Engineering and Computer Science*, 9(2):447–459.

[384] Hutchinson, B., Prabhakaran, V., Denton, E., Webster, K., Zhong, Y., and Denuyl, S. (2020). Unintended machine learning biases as social barriers for persons with disabilities. *SIGACCESS Access. Comput.*, (125).

[385] Hutchinson, B., Smart, A., Hanna, A., Denton, E., Greer, C., Kjartansson, O., Barnes, P., and Mitchell, M. (2021). Towards accountability for machine learning datasets: Practices from software engineering and infrastructure. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 560–575, New York, NY, USA. Association for Computing Machinery.

[386] Häußler, Walter, M. (1979). Empirische ergebnisse zu diskriminationsverfahren bei kreditscoringsystemen.

[387] Iliadis, A. and Russo, F. (2016). Critical data studies: An introduction. *Big Data & Society*, 3(2):2053951716674238.

[388] International Warfarin Pharmacogenetics Consortium (2009). Estimation of the warfarin dose with clinical and pharmacogenetic data. *New England Journal of Medicine*, 360(8):753–764.

[389] Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R., Shpanskaya, K., Seekins, J., Mong, D., Halabi, S., Sandberg, J., Jones, R., Larson, D., Langlotz, C., Patel, B., Lungren, M., and Ng, A. (2019). Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *33rd AAAI Conference on Artificial Intelligence, AAAI 2019, 31st Innovative Applications of Artificial Intelligence Conference, IAAI 2019 and the 9th AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019*, 33rd AAAI Conference on Artificial Intelligence, AAAI 2019, 31st Innovative Applications of Artificial Intelligence Conference, IAAI 2019 and the 9th AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, pages 590–597. AAAI Press. Publisher Copyright: © 2019, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.; 33rd AAAI Conference on Artificial Intelligence, AAAI 2019, 31st Annual Conference on Innovative Applications of Artificial Intelligence, IAAI 2019 and the 9th AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019 ; Conference date: 27-01-2019 Through 01-02-2019.

[390] Islam, R., Pan, S., and Foulds, J. R. (2021). *Can We Obtain Fairness For Free?*, page 586–596. Association for Computing Machinery, New York, NY, USA.

[391] Istat (2020). Stranieri residenti al 1° gennaio - cittadinanza. http://dati.istat.it/Index.aspx?DataSetCode=DCIS_POPSTRCIT1.

[392] Italian govt. (2007a). 2007 decreto legislativo 7 settembre 2005, n. 209. https://www.normattiva.it/uri-res/N2Ls?urn:nir:stato:decreto.legislativo:2005-09-07;209.

[393] Italian govt. (2007b). Legge 2 aprile 2007, n. 40. http://www.normattiva.it/uri-res/N2Ls?urn:nir:stato:legge:2007-04-02;40!vig=2020-09-18.

[394] IVASS - Institute for the Supervision of Insurance (2014a). Investigation into comparison websites in the italian insurance market. https://www.ivass.it/consumatori/azioni-tutela/indagini-tematiche/documenti/INVESTIGATION_INTO_COMPARISON_WEBSITES_IN_THE_ITALIAN_INSURANCE_MARKET.pdf?language_id=3.

[395] IVASS - Institute for the Supervision of Insurance (2014b). Lettera al mercato prot. n. 45-14-007503, 26 novembre 2014. https://www.ivass.it/consumatori/azioni-tutela/lettere-mercato/documenti/Tariffazione_del_rischio_r.c.auto_Fattore_tariffario_nazionalita_di_nascita.pdf.

[396] Jabbari, S., Ou, H.-C., Lakkaraju, H., and Tambe, M. (2020). An empirical study of the trade-offs between interpretability and fairness. In *ICML 2020 Workshop on Human Interpretability in Machine Learning, preliminary version*. ICML 2020 workshop: "Workshop on Human Interpretability in Machine Learning (WHI".

[397] Jacobs, A. Z. and Wallach, H. (2019). Measurement and fairness.

[398] Jacobs, A. Z. and Wallach, H. (2021). Measurement and fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 375–385, New York, NY, USA. Association for Computing Machinery.

[399] Jagielski, M., Kearns, M., Mao, J., Oprea, A., Roth, A., Malvajerdi, S. S., and Ullman, J. (2019). Differentially private fair learning. In Chaudhuri, K. and Salakhutdinov, R., editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 3000–3008, Long Beach, California, USA. PMLR.

[400] Japkowicz, N. (2006). Why question machine learning evaluation methods. In *AAAI workshop on evaluation methods for machine learning*, pages 6–11.

[401] Järvelin, K. and Kekäläinen, J. (2002). Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems*, 20(4):422–446.

[402] Jeffries, A. and Yin, L. (2021). Amazon puts itw own "brands" abve better rated products. https://themarkup.org/amazons-advantage/2021/10/14/amazon-puts-its-own-brands-first-above-better-rated-products.

[403] Ji, D., Smyth, P., and Steyvers, M. (2020). Can I trust my fairness metric? assessing fairness with unlabeled data and bayesian inference. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

[404] Jiang, W. and Pardos, Z. A. (2021). *Towards Equity and Algorithmic Fairness in Student Grade Prediction*, page 608–617. Association for Computing Machinery, New York, NY, USA.

[405] Jo, E. S. and Gebru, T. (2020). Lessons from archives: Strategies for collecting sociocultural data in machine learning. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 306–316, New York, NY, USA. Association for Computing Machinery.

[406] Joachims, T. (2002). Optimizing search engines using clickthrough data. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '02, page 133–142, New York, NY, USA. Association for Computing Machinery.

[407] Joachims, T., Granka, L., Pan, B., Hembrooke, H., and Gay, G. (2017). Accurately interpreting clickthrough data as implicit feedback. *SIGIR Forum*, 51(1):4–11.

[408] Jobin, A., Ienca, M., and Vayena, E. (2019). The global landscape of ai ethics guidelines. *Nature Machine Intelligence*, 1(9):389–399.

[409] Johnson, A. E., Pollard, T. J., Greenbaum, N. R., Lungren, M. P., Deng, C.-y., Peng, Y., Lu, Z., Mark, R. G., Berkowitz, S. J., and Horng, S. (2019). Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*.

[410] Johnson, A. E., Pollard, T. J., Shen, L., Lehman, L. H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L. A., and Mark, R. G. (2016). Mimic-iii, a freely accessible critical care database. *Scientific data*, 3:160035.

[411] Jones, E., Sagawa, S., Koh, P. W., Kumar, A., and Liang, P. (2021). Selective classification can magnify disparities across groups. In *International Conference on Learning Representations*.

[412] Jones, K. S. (1973). Collection properties influencing automatic term classification performance. *Information Storage and Retrieval*, 9(9):499–513.

[413] Jones, M., Nguyen, H., and Nguyen, T. (2020). Fair k-centers via maximum matching. In III, H. D. and Singh, A., editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4940–4949, Virtual. PMLR.

[414] Joulin, A., Grave, E., Bojanowski, P., and Mikolov, T. (2017). Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431. ACL.

[415] Jung, S., Lee, D., Park, T., and Moon, T. (2021). Fair feature distillation for visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12115–12124.

[416] Kaesler, S., Lorenz, J.-T., and Schollmeier, F. (2018). Friends or foes: The rise of european aggregators and their impact on traditional insurers. https://tinyurl.com/y26lghy9.

[417] Kallus, N., Mao, X., and Zhou, A. (2020a). Assessing algorithmic fairness with unobserved protected class using data combination. In *Proceedings of the 3rd Conference on Fairness, Accountability, and Transparency (FAT* 2020)*, page 110, Barcelona, ES.

[418] Kallus, N., Mao, X., and Zhou, A. (2020b). Assessing algorithmic fairness with unobserved protected class using data combination. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 110, New York, NY, USA. Association for Computing Machinery.

[419] Kallus, N. and Zhou, A. (2018). Residual unfairness in fair machine learning from prejudiced data. In *Proceedings of the 35th International Conference on Machine Learning (ICML 2018)*, pages 2439–2448, Stockholm, SE.

[420] Kallus, N. and Zhou, A. (2019a). Assessing disparate impact of personalized interventions: Identifiability and bounds. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 32, pages 3426–3437. Curran Associates, Inc.

[421] Kallus, N. and Zhou, A. (2019b). The fairness of risk scores beyond classification: Bipartite ranking and the xauc metric. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 32, pages 3438–3448. Curran Associates, Inc.

[422] Kallus, N. and Zhou, A. (2021a). Fairness, welfare, and equity in personalized pricing. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 296–314, New York, NY, USA. Association for Computing Machinery.

[423] Kallus, N. and Zhou, A. (2021b). Fairness, welfare, and equity in personalized pricing. In *Proceedings of the 4th ACM Conference on Fairness, Accountability, and Transparency (FAccT 2021)*, pages 296–314, Toronto, CA.

[424] Kamishima, T. (2003). Nantonac collaborative filtering: Recommendation based on order responses. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '03, page 583–588, New York, NY, USA. Association for Computing Machinery.

[425] Kang, J., He, J., Maciejewski, R., and Tong, H. (2020). *InFoRM: Individual Fairness on Graph Mining*, page 379–389. Association for Computing Machinery, New York, NY, USA.

[426] Kannel, W. B. and McGee, D. L. (1979). Diabetes and cardiovascular disease: the framingham study. *Jama*, 241(19):2035–2038.

[427] Kantayya, S. (2020). Coded bias.

[428] Karako, C. and Manggala, P. (2018). Using image fairness representations in diversity-based re-ranking for recommendations. UMAP 2018 workshop: "Fairness in User Modeling, Adaptation and Personalization (FairUMAP)".

[429] Karkkainen, K. and Joo, J. (2021). Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1548–1558.

[430] Karlan, D. S. and Zinman, J. (2008). Credit elasticities in less-developed economies: Implications for microfinance. *American Economic Review*, 98(3):1040–68.

[431] Kasy, M. and Abebe, R. (2021). Fairness, equality, and power in algorithmic decision-making. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 576–586, New York, NY, USA. Association for Computing Machinery.

[432] Kato, M., Teshima, T., and Honda, J. (2019). Learning from positive and unlabeled data with a selection bias. In *International Conference on Learning Representations*.

[433] Kay, M., Matuszek, C., and Munson, S. A. (2015). Unequal representation and gender stereotypes in image search results for occupations. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 3819–3828. ACM.

[434] Kearns, M., Neel, S., Roth, A., and Wu, Z. S. (2018). Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In Dy, J. and Krause, A., editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2564–2572, Stockholmsmässan, Stockholm Sweden. PMLR.

[435] Kearns, M., Neel, S., Roth, A., and Wu, Z. S. (2019). An empirical study of rich subgroup fairness for machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 100–109, New York, NY, USA. Association for Computing Machinery.

[436] Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika*, 30(1/2):81–93.

[437] Keswani, V., Lease, M., and Kenthapadi, K. (2021). *Towards Unbiased and Accurate Deferral to Multiple Experts*, page 154–165. Association for Computing Machinery, New York, NY, USA.

[438] Keyes, O. (2018). The misgendering machines: Trans/HCI implications of automatic gender recognition. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):88:1–88:22.

[439] Keyes, O., Stevens, N., and Wernimont, J. (2019). The government is using the most vulnerable people to test facial recognition software.

[440] Khan, Z. and Fu, Y. (2021). One label, one billion faces: Usage and consistency of racial categories in computer vision. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 587–597, New York, NY, USA. Association for Computing Machinery.

[441] Kilbertus, N., Gascon, A., Kusner, M., Veale, M., Gummadi, K., and Weller, A. (2018a). Blind justice: Fairness with encrypted sensitive attributes. In Dy, J. and Krause, A., editors, *Proceedings of the 35th International Conference on Machine Learning*,

volume 80 of *Proceedings of Machine Learning Research*, pages 2630–2639, Stockholmsmässan, Stockholm Sweden. PMLR.

[442] Kilbertus, N., Gascon, A., Kusner, M., Veale, M., Gummadi, K., and Weller, A. (2018b). Blind justice: Fairness with encrypted sensitive attributes. In *Proceedings of the 35th International Conference on Machine Learning (ICML 2018)*, pages 2630–2639, Stockholm, SE.

[443] Kilkenny, M. F. and Robinson, K. M. (2018). Data quality:"garbage in–garbage out".

[444] Kim, E., Bryant, D., Srikanth, D., and Howard, A. (2021). *Age Bias in Emotion Detection: An Analysis of Facial Emotion Recognition Performance on Young, Middle-Aged, and Older Adults*, page 638–644. Association for Computing Machinery, New York, NY, USA.

[445] Kim, H. and Mnih, A. (2018). Disentangling by factorising. In Dy, J. and Krause, A., editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2649–2658. PMLR.

[446] Kim, J. S., Chen, J., and Talwalkar, A. (2020). FACT: A diagnostic for group fairness trade-offs. In III, H. D. and Singh, A., editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 5264–5274, Virtual. PMLR.

[447] Kim, M. P., Ghorbani, A., and Zou, J. (2019). Multiaccuracy: Black-box post-processing for fairness in classification. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '19, page 247–254, New York, NY, USA. Association for Computing Machinery.

[448] Kiritchenko, S. and Mohammad, S. (2018). Examining gender and race bias in two hundred sentiment analysis systems. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 43–53, New Orleans, Louisiana. Association for Computational Linguistics.

[449] Kırnap, O., Diaz, F., Biega, A., Ekstrand, M., Carterette, B., and Yilmaz, E. (2021). Estimation of fair ranking metrics with incomplete judgments. In *Proceedings of the 2021 Web Conference (WWW 2021)*, pages 1065—-1075, Ljubljana, SL.

[450] Kizhner, I., Terras, M., Rumyantsev, M., Khokhlova, V., Demeshkova, E., Rudov, I., and Afanasieva, J. (2020). Digital cultural colonialism: measuring bias in aggregated digitized content held in Google Arts and Culture. *Digital Scholarship in the Humanities*, 36(3):607–640.

[451] Klare, B. F., Klein, B., Taborsky, E., Blanton, A., Cheney, J., Allen, K., Grother, P., Mah, A., and Jain, A. K. (2015). Pushing the frontiers of unconstrained face detection and recognition: Iarpa janus benchmark a. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[452] Kleinberg, J., Mullainathan, S., and Raghavan, M. (2016). Inherent trade-offs in the fair determination of risk scores. DTL 2016 workshop: "Fairness, Accountability, and Transparency in Machine Learning (FAT/ML)".

[453] Kleindessner, M., Awasthi, P., and Morgenstern, J. (2019a). Fair k-center clustering for data summarization. In Chaudhuri, K. and Salakhutdinov, R., editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 3448–3457, Long Beach, California, USA. PMLR.

[454] Kleindessner, M., Samadi, S., Awasthi, P., and Morgenstern, J. (2019b). Guarantees for spectral clustering with fairness constraints. In Chaudhuri, K. and Salakhutdinov, R., editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 3458–3467, Long Beach, California, USA. PMLR.

[455] Knees, P. and Hübler, M. (2019). Towards uncovering dataset biases: Investigating record label diversity in music playlists. ISMIR 2019 workshop: "Workshop on Designing Human-Centric MIR Systems".

[456] Kobren, A., Saha, B., and McCallum, A. (2019). Paper matching with local fairness constraints. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, page 1247–1257, New York, NY, USA. Association for Computing Machinery.

[457] Koch, B., Denton, E., Hanna, A., and Foster, J. G. (2021). Reduced, reused and recycled: The life of a dataset in machine learning research. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

[458] Kocijan, V., Camburu, O.-M., and Lukasiewicz, T. (2020). The gap on gap: Tackling the problem of differing data distributions in bias-measuring datasets. NeurIPS 2020 workshop: "Algorithmic Fairness through the Lens of Causality and Interpretability (AFCI)".

[459] Kohavi, R. (1996). Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD'96, page 202–207. AAAI Press.

[460] Komiyama, J., Takeda, A., Honda, J., and Shimao, H. (2018). Nonconvex optimization for regression with fairness constraints. In Dy, J. and Krause, A., editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2737–2746, Stockholmsmässan, Stockholm Sweden. PMLR.

[461] Konstantakis, G., Promponas, G., Dretakis, M., and Papadakos, P. (2020). Bias goggles: Exploring the bias of web domains through the eyes of users. In Boratto, L., Faralli, S., Marras, M., and Stilo, G., editors, *Bias and Social Aspects in Search and Recommendation*, pages 66–71, Cham. Springer International Publishing.

[462] Koolen, C. (2018). *Reading beyond the female The relationship between perception of author gender and literary quality*. PhD thesis, University of Amsterdam.

[463] Koolen, C. and van Cranenburgh, A. (2017). These are not the stereotypes you are looking for: Bias and fairness in authorial gender attribution. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 12–22, Valencia, Spain. Association for Computational Linguistics.

[464] Koren, J. R. (2016). Feds use Rand formula to spot discrimination. the GOP calls it junk science. Los Angeles Times, August 23, 2016.

[465] Kozodoi, N. and Varga, T. (2021). The fairness r package.

[466] Kray, L. J., Thompson, L., and Galinsky, A. (2001). Battle of the sexes: gender stereotype confirmation and reactance in negotiations. *Journal of personality and social psychology*, 80(6):942.

[467] Krizhevsky, A. (2009). Learning multiple layers of features from tiny images.

[468] Kröger, J. L., Miceli, M., and Müller, F. (2021). How data can be used against people: A classification of personal data misuses. *Available at SSRN 3887097*.

[469] Kuhlman, C., Gerych, W., and Rundensteiner, E. (2021). Measuring group advantage: A comparative study of fair ranking metrics. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '21, page 674–682, New York, NY, USA. Association for Computing Machinery.

[470] Kuhlman, C. and Rundensteiner, E. (2020). Rank aggregation algorithms for fair consensus. *Proc. VLDB Endow.*, 13(12):2706–2719.

[471] Kuhlman, C., VanValkenburg, M., and Rundensteiner, E. (2019). Fare: Diagnostics for fair ranking using pairwise error metrics. In *The World Wide Web Conference*, WWW '19, page 2936–2942, New York, NY, USA. Association for Computing Machinery.

[472] Kulshrestha, J., Eslami, M., Messias, J., Zafar, M. B., Ghosh, S., Gummadi, K. P., and Karahalios, K. (2017a). Quantifying search bias: Investigating sources of bias for political searches in social media. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, CSCW '17, page 417–432, New York, NY, USA. Association for Computing Machinery.

[473] Kulshrestha, J., Eslami, M., Messias, J., Zafar, M. B., Ghosh, S., Gummadi, K. P., and Karahalios, K. (2017b). Quantifying search bias: Investigating sources of bias for political searches in social media. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, CSCW '17, page 417–432, New York, NY, USA. Association for Computing Machinery.

[474] Kushmerick, N. (1999). Learning to remove internet advertisements. In *Proceedings of the Third Annual Conference on Autonomous Agents*, AGENTS '99, page 175–181, New York, NY, USA. Association for Computing Machinery.

[475] Kusner, M. J., Loftus, J., Russell, C., and Silva, R. (2017a). Counterfactual fairness. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30, pages 4066–4076. Curran Associates, Inc.

[476] Kusner, M. J., Loftus, J., Russell, C., and Silva, R. (2017b). Counterfactual fairness. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30, pages 4066–4076. Curran Associates, Inc.

[477] Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Malloci, M., Kolesnikov, A., et al. (2020). The open images dataset v4. *International Journal of Computer Vision*, 128(7):1956–1981.

[478] Lahoti, P., Beutel, A., Chen, J., Lee, K., Prost, F., Thain, N., Wang, X., and Chi, E. (2020). Fairness without demographics through adversarially reweighted learning. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 728–740. Curran Associates, Inc.

[479] Lake, B. M., Salakhutdinov, R., and Tenenbaum, J. B. (2015). Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338.

[480] Lamy, A., Zhong, Z., Menon, A. K., and Verma, N. (2019). Noise-tolerant fair classification. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 32, pages 294–306. Curran Associates, Inc.

[481] Lan, C. and Huan, J. (2017). Discriminatory transfer. KDD 2017 workshop: "Fairness, Accountability, and Transparency in Machine Learning (FAT/ML)".

[482] Larson, J., Angwin, J., Kirchner, L., Mattu, S., Haner, D., Saccucci, M., Newsom-Stewart, K., Cohen, A., and Romm, M. (2017). How we examined racial discrimination in auto insurance prices. Machine bias, ProPublica, New York, NY, USA.

[483] Larson, J., Mattu, S., Kirchner, L., and Angwin, J. (2016). How we analyzed the COMPAS recidivism algorithm. ProPublica, May 23, 2016.

[484] Le Pelley, M. E., Reimers, S. J., Calvini, G., Spears, R., Beesley, T., and Murphy, R. A. (2010). Stereotype formation: Biased by association. *Journal of Experimental Psychology: General*, 139(1):138–161.

[485] Le Quy, T., Roy, A., Iosifidis, V., Zhang, W., and Ntoutsi, E. (2022). A survey on datasets for fairness-aware machine learning. *WIREs Data Mining and Knowledge Discovery*, n/a(n/a):e1452.

[486] Leavy, S., Meaney, G., Wade, K., and Greene, D. (2019). *Curatr: A Platform for Semantic Analysis and Curation of Historical Literary Texts*, pages 354–366.

[487] Leavy, S., Meaney, G., Wade, K., and Greene, D. (2020). Mitigating gender bias in machine learning data sets. In Boratto, L., Faralli, S., Marras, M., and Stilo, G., editors, *Bias and Social Aspects in Search and Recommendation*, pages 12–26, Cham. Springer International Publishing.

[488] Lecun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

[489] LeCun, Y., Fu Jie Huang, and Bottou, L. (2004). Learning methods for generic object recognition with invariance to pose and lighting. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, volume 2, pages II–104 Vol.2.

[490] Lee, H. and Kizilcec, R. F. (2020). Evaluation of fairness trade-offs in predicting student success. International Conference on Educational Data Mining workshop: “Fairness, Accountability, and Transparency, in Educational Data (Mining)".

[491] Lemaire, J., Park, S. C., and Wang, K. (2015). The use of annual mileage as a rating variable. *Astin Bulletin*, 46(1):39.

[492] Leonelli, S. and Tempini, N. (2020). *Data journeys in the sciences*. Springer Nature.

[493] Leskovec, J., Kleinberg, J., and Faloutsos, C. (2007). Graph evolution: Densification and shrinking diameters. *ACM transactions on Knowledge Discovery from Data (TKDD)*, 1(1):2–es.

[494] Leskovec, J. and Mcauley, J. (2012). Learning to discover social circles in ego networks. In Pereira, F., Burges, C. J. C., Bottou, L., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc.

[495] Lesmana, N. S., Zhang, X., and Bei, X. (2019). Balancing efficiency and fairness in on-demand ridesourcing. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 32, pages 5309–5319. Curran Associates, Inc.

[496] Lessmann, S., Baesens, B., Seow, H.-V., and Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research*, 247(1):124–136.

[497] Levy, D., Splansky, G. L., Strand, N. K., Atwood, L. D., Benjamin, E. J., Blease, S., Cupples, L. A., D’Agostino Sr, R. B., Fox, C. S., Kelly-Hayes, M., et al. (2010). Consent for genetic research in the framingham heart study. *American Journal of Medical Genetics Part A*, 152(5):1250–1256.

[498] Li, P., Wang, Y., Zhao, H., Hong, P., and Liu, H. (2021). On dyadic fairness: Exploring and mitigating bias in graph connections. In *International Conference on Learning Representations*.

[499] Li, P., Zhao, H., and Liu, H. (2020a). Deep fair clustering for visual learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[500] Li, T., Sanjabi, M., Beirami, A., and Smith, V. (2020b). Fair resource allocation in federated learning. In *International Conference on Learning Representations*.

[501] Li, Y., Ning, Y., Liu, R., Wu, Y., and Hui Wang, W. (2020c). Fairness of classification using users’ social relationships in online peer-to-peer lending. In *Companion Proceedings of the Web Conference 2020*, WWW ’20, page 733–742, New York, NY, USA. Association for Computing Machinery.

[502] Li, Y., Sun, H., and Wang, W. H. (2020d). *Towards Fair Truth Discovery from Biased Crowdsourced Answers*, page 599–607. Association for Computing Machinery, New York, NY, USA.

[503] Li, Z., Zhao, H., Liu, Q., Huang, Z., Mei, T., and Chen, E. (2018). Learning from history and present: Next-item recommendation via discriminatively exploiting user behaviors. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '18, page 1734–1743, New York, NY, USA. Association for Computing Machinery.

[504] Liang, L. and Acuna, D. E. (2020). Artificial mental phenomena: Psychophysics as a framework to detect perception biases in ai models. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 403–412, New York, NY, USA. Association for Computing Machinery.

[505] Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In Fleet, D., Pajdla, T., Schiele, B., and Tuytelaars, T., editors, *Computer Vision – ECCV 2014*, pages 740–755, Cham. Springer International Publishing.

[506] Lipton, Z., McAuley, J., and Chouldechova, A. (2018). Does mitigating ml's impact disparity require treatment disparity? In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

[507] Liu, D., Shafi, Z., Fleisher, W., Eliassi-Rad, T., and Alfeld, S. (2021). *RAWLSNET: Altering Bayesian Networks to Encode Rawlsian Fair Equality of Opportunity*, page 745–755. Association for Computing Machinery, New York, NY, USA.

[508] Liu, L. T., Dean, S., Rolf, E., Simchowitz, M., and Hardt, M. (2018). Delayed impact of fair machine learning. In Dy, J. and Krause, A., editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 3150–3158, Stockholmsmässan, Stockholm Sweden. PMLR.

[509] Liu, L. T., Simchowitz, M., and Hardt, M. (2019). The implicit fairness criterion of unconstrained learning. In Chaudhuri, K. and Salakhutdinov, R., editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 4051–4060, Long Beach, California, USA. PMLR.

[510] Liu, L. T., Wilson, A., Haghtalab, N., Kalai, A. T., Borgs, C., and Chayes, J. (2020). The disparate equilibria of algorithmic decision making when individuals invest rationally. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 381–391, New York, NY, USA. Association for Computing Machinery.

[511] Liu, W. and Burke, R. (2018). Personalizing fairness-aware re-ranking. RecSys 2018 workshop: "Workshop on Responsible Recommendation (FAT/Rec)".

[512] Liu, Z., Luo, P., Wang, X., and Tang, X. (2015). Deep learning face attributes in the wild.

[513] Locatello, F., Abbati, G., Rainforth, T., Bauer, S., Schölkopf, B., and Bachem, O. (2019). On the fairness of disentangled representations. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 32, pages 14611–14624. Curran Associates, Inc.

[514] Lohaus, M., Perrot, M., and Luxburg, U. V. (2020). Too relaxed to be fair. In III, H. D. and Singh, A., editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 6360–6369, Virtual. PMLR.

[515] Louizos, C., Swersky, K., Li, Y., Welling, M., and Zemel, R. S. (2016). The variational fair autoencoder. In Bengio, Y. and LeCun, Y., editors, *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*.

[516] Lowe, H., Ferris, T. A., Hernandez, P., and Weber, S. (2009). Stride - an integrated standards-based translational research informatics platform. *AMIA ... Annual Symposium proceedings. AMIA Symposium*, 2009:391–5.

[517] Lu, Q. and Getoor, L. (2003). Link-based classification. In *Proceedings of the Twentieth International Conference on International Conference on Machine Learning*, ICML'03, page 496–503. AAAI Press.

[518] Lum, K., Boudin, C., and Price, M. (2020). The impact of overbooking on a pre-trial risk assessment tool. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 482–491, New York, NY, USA. Association for Computing Machinery.

[519] Lum, K. and Johndrow, J. (2016). A statistical framework for fair predictive algorithms. DTL 2016 workshop: "Fairness, Accountability, and Transparency in Machine Learning (FAT/ML)".

[520] Luong, B. T., Ruggieri, S., and Turini, F. (2016). Classification rule mining supported by ontology for discrimination discovery. In *2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW)*, pages 868–875.

[521] Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., and Potts, C. (2011). Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.

[522] Madnani, N., Loukina, A., von Davier, A., Burstein, J., and Cahill, A. (2017). Building better open-source tools to support fairness in automated scoring. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 41–52, Valencia, Spain. Association for Computational Linguistics.

[523] Madras, D., Creager, E., Pitassi, T., and Zemel, R. (2018a). Learning adversarially fair and transferable representations. In Dy, J. and Krause, A., editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 3384–3393, Stockholmsmässan, Stockholm Sweden. PMLR.

[524] Madras, D., Creager, E., Pitassi, T., and Zemel, R. (2019). Fairness through causal awareness: Learning causal latent-variable models for biased data. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 349–358, New York, NY, USA. Association for Computing Machinery.

[525] Madras, D., Pitassi, T., and Zemel, R. (2018b). Predict responsibly: Improving fairness and accuracy by learning to defer. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 31, pages 6147–6157. Curran Associates, Inc.

[526] Mahabadi, S. and Vakilian, A. (2020). Individual fairness for k-clustering. In III, H. D. and Singh, A., editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 6586–6596, Virtual. PMLR.

[527] Maity, S., Xue, S., Yurochkin, M., and Sun, Y. (2021). Statistical inference for individual fairness. In *International Conference on Learning Representations*.

[528] Malinin, A., Band, N., Chesnokov, G., Gal, Y., Gales, M. J., Noskov, A., Ploskonosov, A., Prokhorenkova, L., Provilkov, I., Raina, V., et al. (2021). Shifts: A dataset of real distributional shift across multiple large-scale tasks. *arXiv preprint arXiv:2107.07455*.

[529] Mandal, D., Deng, S., Jana, S., Wing, J. M., and Hsu, D. J. (2020). Ensuring fairness beyond the training data. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

[530] Manjunatha, V., Saini, N., and Davis, L. S. (2019). Explicit bias discovery in visual question answering models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[531] Mansoury, M., Abdollahpouri, H., Pechenizkiy, M., Mobasher, B., and Burke, R. (2020). Feedback loop and bias amplification in recommender systems. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management (CIKM 2020)*, pages 2145–2148, Virtual Event.

[532] Marchesin, S., Purpura, A., and Silvello, G. (2019). Focal elements of neural information retrieval models. an outlook through a reproducibility study. *Information Processing & Management*, page to appear.

[533] Martin, C. L. and Ruble, D. N. (2010). Patterns of gender development. *Annual review of psychology*, 61:353–381.

[534] Martinez, N., Bertran, M., and Sapiro, G. (2020). Minimax pareto fairness: A multi objective perspective. In III, H. D. and Singh, A., editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 6755–6764, Virtual. PMLR.

[535] Mary, J., Calauzènes, C., and Karoui, N. E. (2019). Fairness-aware learning for continuous attributes and treatments. In Chaudhuri, K. and Salakhutdinov, R., editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 4382–4391, Long Beach, California, USA. PMLR.

[536] Mastrandrea, R., Fournet, J., and Barrat, A. (2015). Contact patterns in a high school: A comparison between data collected using wearable sensors, contact diaries and friendship surveys. *PLOS ONE*, 10(9):e0136497.

[537] Mattei, N., Saffidine, A., and Walsh, T. (2018a). An axiomatic and empirical analysis of mechanisms for online organ matching. In *Proceedings of the 7th International Workshop on Computational Social Choice (COMSOC)*.

[538] Mattei, N., Saffidine, A., and Walsh, T. (2018b). Fairness in deceased organ matching. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18, page 236–242, New York, NY, USA. Association for Computing Machinery.

[539] Mayson, S. G. (2018). Bias in, bias out. *YAle lJ*, 128:2218.

[540] McAuley, J., Targett, C., Shi, Q., and van den Hengel, A. (2015). Image-based recommendations on styles and substitutes. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '15, page 43–52, New York, NY, USA. Association for Computing Machinery.

[541] McCallum, A. K., Nigam, K., Rennie, J., and Seymore, K. (2000). Automating the construction of internet portals with machine learning. *Information Retrieval*, 3(2):127–163.

[542] McDonald, G., Macdonald, C., and Ounis, I. (2022). Search results diversification for effective fair ranking in academic search. *Information Retrieval Journal*, 25(1):1–26.

[543] McDuff, D., Ma, S., Song, Y., and Kapoor, A. (2019). Characterizing bias in classifiers using generative models. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 32, pages 5403–5414. Curran Associates, Inc.

[544] McFee, B., Bertin-Mahieux, T., Ellis, D. P., and Lanckriet, G. R. (2012). The million song dataset challenge. In *Proceedings of the 21st International Conference on World Wide Web*, WWW '12 Companion, page 909–916, New York, NY, USA. Association for Computing Machinery.

[545] McKenna, L. (2019a). A history of the current population survey and disclosure avoidance.

[546] McKenna, L. (2019b). A history of the us census bureau's disclosure review board.

[547] McLemore, K. A. (2015). Experiences with misgendering: Identity misclassification of transgender spectrum individuals. *Self and Identity*, 14(1):51–74.

[548] McMahan, B., Moore, E., Ramage, D., Hampson, S., and y Arcas, B. A. (2017). Communication-Efficient Learning of Deep Networks from Decentralized Data. In Singh, A. and Zhu, J., editors, *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 1273–1282, Fort Lauderdale, FL, USA. PMLR.

[549] McNamara, D. (2019). Equalized odds implies partially equalized outcomes under realistic assumptions. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '19, page 313–320, New York, NY, USA. Association for Computing Machinery.

[550] McSherry, F. and Najork, M. (2008). Computing information retrieval performance measures efficiently in the presence of tied scores. In Macdonald, C., Ounis, I., Plachouras, V., Ruthven, I., and White, R. W., editors, *Advances in Information Retrieval*, pages 414–421, Berlin, Heidelberg. Springer Berlin Heidelberg.

[551] Meek, C., Thiesson, B., and Heckerman, D. (2002). The learning-curve sampling method applied to model-based clustering. *Journal of Machine Learning Research*, 2(Feb):397–418.

[552] Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Comput. Surv.*, 54(6).

[553] Mehrotra, A. and Celis, L. E. (2021). Mitigating bias in set selection with noisy protected attributes. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 237–248, New York, NY, USA. Association for Computing Machinery.

[554] Mehrotra, R., Anderson, A., Diaz, F., Sharma, A., Wallach, H., and Yilmaz, E. (2017). Auditing search engines for differential satisfaction across demographics. In *Proceedings of the 26th International Conference on World Wide Web Companion*, WWW '17 Companion, page 626–633, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

[555] Merkley, R. (2019). Use and fair use: Statement on shared images in facial recognition ai.

[556] Merler, M., Ratha, N., Feris, R. S., and Smith, J. R. (2019). Diversity in faces.

[557] Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American psychologist*, 50(9):741–749.

[558] Messick, S. (1998). Test validity: A matter of consequence. *Social Indicators Research*, 45(1-3):35–44.

[559] Metaxa, D., Park, J. S., Robertson, R. E., Karahalios, K., Wilson, C., Hancock, J., Sandvig, C., et al. (2021). Auditing algorithms: Understanding algorithmic systems from the outside in. *Foundations and Trends® in Human–Computer Interaction*, 14(4):272–344.

[560] Metcalf, J. and Crawford, K. (2016). Where are human subjects in big data research? the emerging ethics divide. *Big Data & Society*, 3(1):2053951716650211.

[561] Metevier, B., Giguere, S., Brockman, S., Kobren, A., Brun, Y., Brunskill, E., and Thomas, P. S. (2019). Offline contextual bandits with high probability fairness guarantees. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 32, pages 14922–14933. Curran Associates, Inc.

[562] Mhasawade, V. and Chunara, R. (2021). *Causal Multi-Level Fairness*, page 784–794. Association for Computing Machinery, New York, NY, USA.

[563] Miao, W. (2010). Did the results of promotion exams have a disparate impact on minorities? using statistical evidence in ricci v. destefano. *Journal of Statistics Education*, 18(3).

[564] Miceli, M., Yang, T., Naudts, L., Schuessler, M., Serbanescu, D., and Hanna, A. (2021). Documenting computer vision datasets: An invitation to reflexive data practices. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 161–172, New York, NY, USA. Association for Computing Machinery.

[565] Mikolov, T., Grave, E., Bojanowski, P., Puhrsch, C., and Joulin, A. (2018). Advances in pre-training distributed word representations. In *Proc. of LREC 2018*.

[566] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

[567] Mirkin, S., Nowson, S., Brun, C., and Perez, J. (2015). Motivating personality-aware machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1102–1108, Lisbon, Portugal. Association for Computational Linguistics.

[568] Mishler, A., Kennedy, E. H., and Chouldechova, A. (2021). Fairness in risk assessment instruments: Post-processing to achieve counterfactual equalized odds. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 386–400, New York, NY, USA. Association for Computing Machinery.

[569] Mishra, S., He, S., and Belli, L. (2020). Assessing demographic bias in named entity recognition. AKBC 2020 workshop: "Bias in Automatic Knowledge Graph Construction".

[570] Mislove, A., Viswanath, B., Gummadi, K. P., and Druschel, P. (2010). You are who you know: Inferring user profiles in online social networks. In *Proceedings of the Third ACM International Conference on Web Search and Data Mining*, WSDM '10, page 251–260, New York, NY, USA. Association for Computing Machinery.

[571] Mitchell, M., Baker, D., Moorosi, N., Denton, E., Hutchinson, B., Hanna, A., Gebru, T., and Morgenstern, J. (2020). *Diversity and Inclusion Metrics in Subset Selection*, page 117–123. Association for Computing Machinery, New York, NY, USA.

[572] Mitchell, S., Potash, E., Barocas, S., D'Amour, A., and Lum, K. (2021). Algorithmic fairness: Choices, assumptions, and definitions. *Annual Review of Statistics and Its Application*, 8(1):141–163.

[573] Mitra, B. and Craswell, N. (2018). An introduction to neural information retrieval. *Foundations and Trends® in Information Retrieval*, 13(1):1–126.

[574] Moffat, A. and Zobel, J. (2008a). Rank-biased precision for measurement of retrieval effectiveness. *ACM Trans. Inf. Syst.*, 27(1).

[575] Moffat, A. and Zobel, J. (2008b). Rank-biased precision for measurement of retrieval effectiveness. *ACM Transactions on Information Systems*, 27(1):1–27.

[576] Moore, J. C., Stinson, L. L., and Welniak, E. J. (2000). Income measurement error in surveys: A review. *Journal of Official Statistics-Stockholm-*, 16(4):331–362.

[577] Moreland, A., Herlihy, C., Tynan, M. A., Sunshine, G., McCord, R. F., Hilton, C., Poovey, J., Werner, A. K., Jones, C. D., Fulmer, E. B., et al. (2020). Timing of state and territorial covid-19 stay-at-home orders and changes in population movement—united states, march 1–may 31, 2020. *Morbidity and Mortality Weekly Report*, 69(35):1198.

[578] Moreno-Torres, J. G., Raeder, T., Alaiz-Rodríguez, R., Chawla, N. V., and Herrera, F. (2012). A unifying view on dataset shift in classification. *Pattern Recognition*, 45(1):521–530.

[579] Moreo, A. and Sebastiani, F. (2021). Re-assessing the "classify and count" quantification method. In *Proceedings of the 43rd European Conference on Information Retrieval (ECIR 2021)*, volume II, pages 75–91, Lucca, IT.

[580] Moreo, A. and Sebastiani, F. (2022). Tweet sentiment quantification: An experimental re-evaluation. *PLoS ONE*. Forthcoming.

[581] Moro, S., Cortez, P., and Rita, P. (2014). A data-driven approach to predict the success of bank telemarketing. *Decis. Support Syst.*, 62:22–31.

[582] Mozannar, H., Ohannessian, M., and Srebro, N. (2020). Fair learning with private demographic data. In III, H. D. and Singh, A., editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 7066–7075, Virtual. PMLR.

[583] Mukherjee, D., Yurochkin, M., Banerjee, M., and Sun, Y. (2020). Two simple ways to learn individual fairness metrics from data. In III, H. D. and Singh, A., editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 7097–7107, Virtual. PMLR.

[584] Muller, M., Lange, I., Wang, D., Piorkowski, D., Tsay, J., Liao, Q. V., Dugan, C., and Erickson, T. (2019). *How Data Science Workers Work with Data: Discovery, Capture, Curation, Design, Creation*, page 1–15. Association for Computing Machinery, New York, NY, USA.

[585] Mumtaz, Z., Shahid, U., and Levay, A. (2013). Understanding the impact of gendered roles on the experiences of infertility amongst men and women in punjab. *Reproductive health*, 10(3):1–10.

[586] Murgia, M. (2019). Microsoft quietly deletes largest public face recognition data set.

[587] Nabi, R., Malinsky, D., and Shpitser, I. (2019). Learning optimal fair policies. In Chaudhuri, K. and Salakhutdinov, R., editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 4674–4682, Long Beach, California, USA. PMLR.

[588] Namata, G., London, B., Getoor, L., Huang, B., and EDU, U. (2012). Query-driven active surveying for collective classification. In *10th International Workshop on Mining and Learning with Graphs*, volume 8.

[589] Nanda, V., Dooley, S., Singla, S., Feizi, S., and Dickerson, J. P. (2021). Fairness through robustness: Investigating robustness disparity in deep learning. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 466–477, New York, NY, USA. Association for Computing Machinery.

[590] Nanda, V., Xu, P., Sankararaman, K. A., Dickerson, J. P., and Srinivasan, A. (2020). Balancing the tradeoff between profit and fairness in rideshare platforms during high-demand hours. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, AIES '20, page 131, New York, NY, USA. Association for Computing Machinery.

[591] Narasimhan, H., Cotter, A., Gupta, M., and Wang, S. (2020). *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(04):5248–5255.

[592] Narayanan, A. (2018). 21 fairness definitions and their politics. In *Tutorial presented at the 1st ACM Conference on Fairness, Accountability and Transparency (FAT* 2018)*, New York, US.

[593] Narayanan, A. and Shmatikov, V. (2008). Robust de-anonymization of large sparse datasets. In *2008 IEEE Symposium on Security and Privacy (sp 2008)*, pages 111–125. IEEE.

[594] Nasr, M. and Tschantz, M. C. (2020). Bidding strategies with gender nondiscrimination constraints for online ad auctions. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 337–347, New York, NY, USA. Association for Computing Machinery.

[595] National Consumer Union (2019). Costo assicurazione auto: da cosa dipende? https://www.consumatori.it/auto-moto/costo-assicurazione-auto/.

[596] Ngiam, K. Y. and Khor, W. (2019). Big data and machine learning algorithms for health-care delivery. *The Lancet Oncology*, 20(5):e262–e273.

[597] Ngong, I. C., Maughan, K., and Near, J. P. (2020). Towards auditability for fairness in deep learning. NeurIPS 2020 workshop: "Algorithmic Fairness through the Lens of Causality and Interpretability (AFCI)".

[598] NLST Trial Research Team (2011). The national lung screening trial: overview and study design. *Radiology*, 258(1):243–253.

[599] Noble, S. U. (2018). *Algorithms of oppression: How search engines reinforce racism*. NYU Press.

[600] Noriega-Campero, A., Bakker, M. A., Garcia-Bulle, B., and Pentland, A. S. (2019). Active fairness in algorithmic decision making. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '19, page 77–83, New York, NY, USA. Association for Computing Machinery.

[601] Noriega-Campero, A., Garcia-Bulle, B., Cantu, L. F., Bakker, M. A., Tejerina, L., and Pentland, A. (2020). Algorithmic targeting of social policies: Fairness, accuracy, and distributed governance. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 241–251, New York, NY, USA. Association for Computing Machinery.

[602] Nosek, B. A., Banaji, M. R., and Greenwald, A. G. (2002a). Harvesting implicit group attitudes and beliefs from a demonstration web site. *Group Dynamics: Theory, Research, and Practice*, 6(1):101–115.

[603] Nosek, B. A., Banaji, M. R., and Greenwald, A. G. (2002b). Math=male, me=female, therefore math≠me. *Journal of Personality and Social Psychology*, 83(1):44–59.

[604] Novaković, J. D., Veljović, A., Ilić, S. S., Papić, Ž., and Milica, T. (2017). Evaluation of classification models in machine learning. *Theory and Applications of Mathematics & Computer Science*, 7(1):39–46.

[605] Novin, A. and Meyers, E. (2017). Making sense of conflicting science information: Exploring bias in the search engine result page. In *Proc. of CHIIR 2017*, page 175–184. ACM.

[606] Nuttall, D. L., Goldstein, H., Prosser, R., and Rasbash, J. (1989). Differential school effectiveness. *International Journal of Educational Research*, 13(7):769–776.

[607] Obermeyer, Z. and Mullainathan, S. (2019). Dissecting racial bias in an algorithm that guides health decisions for 70 million people. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 89, New York, NY, USA. Association for Computing Machinery.

[608] Ogura, H. and Takeda, A. (2020). Convex fairness constrained model using causal effect estimators. In *Companion Proceedings of the Web Conference 2020*, WWW '20, page 723–732, New York, NY, USA. Association for Computing Machinery.

[609] Olave, M., Rajkovic, V., and Bohanec, M. (1989). An application for admission in public school systems. *Expert Systems in Public Administration*, 1:145–160.

[610] O'Neil, C. (2016). *Weapons of Math Destruction*. Crown Books.

[611] Oneto, L., Donini, M., Elders, A., and Pontil, M. (2019a). Taking advantage of multitask learning for fair classification. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '19, page 227–237, New York, NY, USA. Association for Computing Machinery.

[612] Oneto, L., Donini, M., Luise, G., Ciliberto, C., Maurer, A., and Pontil, M. (2020). Exploiting MMD and sinkhorn divergences for fair and transferable representation learning. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

[613] Oneto, L., Donini, M., Maurer, A., and Pontil, M. (2019b). Learning fair and transferable representations. NeurIPS 2019 workshop: "Human-Centric Machine Learning".

[614] Oneto, L., Siri, A., Luria, G., and Anguita, D. (2017). Dropout prediction at university of genoa: a privacy preserving data driven approach. In *ESANN*.

[615] Ong, P. M. (2002). Car ownership and welfare-to-work. *Journal of Policy Analysis and Management*, 21(2):239–252.

[616] Ong, P. M. and Stoll, M. A. (2007). Redlining or risk? a spatial analysis of auto insurance rates in los angeles. *Journal of Policy Analysis and Management*, 26(4):811–830.

[617] Otterbacher, J., Bates, J., and Clough, P. (2017). Competent men and warm women: Gender stereotypes and backlash in image search results. In *Proc. of CHI 2017*, page 6620–6631. ACM.

[618] Pandey, A. and Caliskan, A. (2021). *Disparate Impact of Artificial Intelligence Bias in Ridehailing Economy's Price Discrimination Algorithms*, page 822–833. Association for Computing Machinery, New York, NY, USA.

[619] Pang, L., Lan, Y., Guo, J., Xu, J., Wan, S., and Cheng, X. (2016). Text matching as image recognition. In *Proc. of AAAI 2016*, page 2793–2799. AAAI Press.

[620] Papakyriakopoulos, O., Hegelich, S., Serrano, J. C. M., and Marco, F. (2020a). Bias in word embeddings. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 446–457, New York, NY, USA. Association for Computing Machinery.

[621] Papakyriakopoulos, O., Hegelich, S., Serrano, J. C. M., and Marco, F. (2020b). Bias in word embeddings. In *Proc. of FAT* 2020*, page 446–457. ACM.

[622] Paraschakis, D. and Nilsson, B. (2020). Matchmaking under fairness constraints: a speed dating case study. ECIR 2020 workshop: "International Workshop on Algorithmic Bias in Search and Recommendation (BIAS 2020)".

[623] Park, L. E., Young, A. F., and Eastwick, P. W. (2015). (psychological) distance makes the heart grow fonder: Effects of psychological distance and relative intelligence on men's attraction to women. *Personality and Social Psychology Bulletin*, 41(11):1459–1473.

[624] Partnership on AI (2022). About ML. Technical report.

[625] Patro, G. K., Chakraborty, A., Ganguly, N., and Gummadi, K. P. (2019). Incremental fairness in two-sided market platforms: On smoothly updating recommendations. NeurIPS 2019 workshop: "Human-Centric Machine Learning".

[626] Patterson, K., Nestor, P. J., and Rogers, T. T. (2007). Where do you know what you know? the representation of semantic knowledge in the human brain. *Nature Reviews Neuroscience*, 8(12):976–987.

[627] Paullada, A., Raji, I. D., Bender, E. M., Denton, E., and Hanna, A. (2020). Data and its (dis) contents: A survey of dataset development and use in machine learning research. *arXiv preprint arXiv:2012.05345*.

[628] Pedreshi, D., Ruggieri, S., and Turini, F. (2008a). Discrimination-aware data mining. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, page 560–568, New York, NY, USA. Association for Computing Machinery.

[629] Pedreshi, D., Ruggieri, S., and Turini, F. (2008b). Discrimination-aware data mining. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, page 560–568, New York, NY, USA. Association for Computing Machinery.

[630] Peng, K., Mathur, A., and Narayanan, A. (2021). Mitigating dataset harms requires stewardship: Lessons from 1000 papers. *arXiv preprint arXiv:2108.02922*.

[631] Pennington, J., Socher, R., and Manning, C. (2014). Glove: Global vectors for word representation. volume 14, pages 1532–1543.

[632] Penton-Voak, I. S., Pound, N., Little, A. C., and Perrett, D. I. (2006). Personality judgments from natural and composite facial images: More evidence for a "kernel of truth" in social perception. *Social cognition*, 24(5):607–640.

[633] Perrone, V., Donini, M., Zafar, M. B., Schmucker, R., Kenthapadi, K., and Archambeau, C. (2021). *Fair Bayesian Optimization*, page 854–863. Association for Computing Machinery, New York, NY, USA.

[634] Pessach, D. and Shmueli, E. (2020). Algorithmic fairness. *arXiv preprint arXiv:2001.09784*.

[635] Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

[636] Peters, M. E. and Lecocq, D. (2013). Content extraction using diverse feature sets. In *Proceedings of the 22Nd International Conference on World Wide Web*, pages 89–90.

[637] Pfohl, S., Marafino, B., Coulet, A., Rodriguez, F., Palaniappan, L., and Shah, N. H. (2019). Creating fair models of atherosclerotic cardiovascular disease risk. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '19, page 271–278, New York, NY, USA. Association for Computing Machinery.

[638] Pinard, M. (2010). Collateral consequences of criminal convictions: Confronting issues of race and dignity. *NYUL Rev.*, 85:457.

[639] Pitoura, E., Stefanidis, K., and Koutrika, G. (2021). Fairness in rankings and recommendations: An overview. *The VLDB Journal*, pages 1–28.

[640] Plante, I., De la Sablonnière, R., Aronson, J. M., and Théorêt, M. (2013). Gender stereotype endorsement and achievement-related outcomes: The role of competence beliefs and task values. *Contemporary Educational Psychology*, 38(3):225–235.

[641] Platt, J. C. (2000). Probabilistic outputs for support vector machines and comparison to regularized likelihood methods. In Smola, A., Bartlett, P., Schölkopf, B., and Schuurmans, D., editors, *Advances in Large Margin Classifiers*, pages 61–74. The MIT Press, Cambridge, MA.

[642] Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., and Weinberger, K. Q. (2017). On fairness and calibration. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30, pages 5680–5689. Curran Associates, Inc.

[643] Pont-Tuset, J., Uijlings, J., Changpinyo, S., Soricut, R., and Ferrari, V. (2020). Connecting vision and language with localized narratives. In *European Conference on Computer Vision*, pages 647–664. Springer.

[644] Prabhu, V. U. and Birhane, A. (2020). Large image datasets: A pyrrhic win for computer vision? *arXiv preprint arXiv:2006.16923*.

[645] Preoţiuc-Pietro, D. and Ungar, L. (2018). User-level race and ethnicity predictors from Twitter text. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1534–1545, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

[646] ProPublica (2016). Compas analysis github repository.

[647] ProPublica (2021). Propublica data store terms.

[648] Prost, F., Thain, N., and Bolukbasi, T. (2019). Debiasing embeddings for reduced gender bias in text classification. In *Proc. of the 1st ACL Workshop on Gender Bias for Natural Language Processing*, pages 69–75. ACL.

[649] Pujol, D., McKenna, R., Kuppam, S., Hay, M., Machanavajjhala, A., and Miklau, G. (2020). Fair decision making using privacy-protected data. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 189–199, New York, NY, USA. Association for Computing Machinery.

[650] Qian, S., Cao, J., Mouël, F. L., Sahel, I., and Li, M. (2015). Scram: A sharing considered route assignment mechanism for fair taxi route recommendations. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, page 955–964, New York, NY, USA. Association for Computing Machinery.

[651] Qin, T. and Liu, T.-Y. (2013). Introducing letor 4.0 datasets.

[652] Quadrianto, N. and Sharmanska, V. (2017). Recycling privileged learning and distribution matching for fairness. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30, pages 677–688. Curran Associates, Inc.

[653] Quadrianto, N., Sharmanska, V., and Thomas, O. (2019). Discovering fair representations in the data domain. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[654] Radford, A., Narasimhan, K., Salimans, T., and Sutskever, I. (2018). Improving language understanding by generative pre-training.

[655] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language models are unsupervised multitask learners.

[656] Radin, J. (2017). "digital natives": How medical and indigenous histories matter for big data. *Osiris*, 32(1):43–64.

[657] Raff, E. and Sylvester, J. (2018). Gradient reversal against discrimination. ICML 2018 workshop: "Fairness, Accountability, and Transparency in Machine Learning (FAT/ML)".

[658] Raff, E., Sylvester, J., and Mills, S. (2018). Fair forests: Regularized tree induction to minimize model bias. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18, page 243–250, New York, NY, USA. Association for Computing Machinery.

[659] Rahmattalabi, A., Vayanos, P., Fulginiti, A., Rice, E., Wilder, B., Yadav, A., and Tambe, M. (2019). Exploring algorithmic fairness in robust graph covering problems. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 32, pages 15776–15787. Curran Associates, Inc.

[660] Raj, A. and Ekstrand, M. D. (2022). Measuring fairness in ranked results: An analytical and empirical comparison. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

[661] Raj, A., Wood, C., Montoly, A., and Ekstrand, M. D. (2020). Comparing fair ranking metrics. RecSys 2020 workshop: "3rd FAccTRec Workshop on Responsible Recommendation".

[662] Raji, I. D. and Buolamwini, J. (2019). Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial ai products. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '19, page 429–435, New York, NY, USA. Association for Computing Machinery.

[663] Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., and Barnes, P. (2020). Closing the ai accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 33–44, New York, NY, USA. Association for Computing Machinery.

[664] Ramachandran, G. S., Brugere, I., Varshney, L. R., and Xiong, C. (2021). *GAEA: Graph Augmentation for Equitable Access via Reinforcement Learning*, page 884–894. Association for Computing Machinery, New York, NY, USA.

[665] Ramaswamy, V. V., Kim, S. S. Y., and Russakovsky, O. (2021). Fair attribute classification through latent space de-biasing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9301–9310.

[666] Ramji-Nogales, J., Schoenholtz, A. I., and Schrag, P. G. (2007). Refugee roulette: Disparities in asylum adjudication. *Stan. L. Rev.*, 60:295.

[667] Rawls, J. (1971). *A Theory of Justice*. Belknap Press.

[668] Red, V., Kelsic, E. D., Mucha, P. J., and Porter, M. A. (2011). Comparing community structure to characteristics in online collegiate social networks. *SIAM review*, 53(3):526–543.

[669] Redmond, M. and Baveja, A. (2002). A data-driven software tool for enabling cooperative information sharing among police departments. *European Journal of Operational Research*, 141:660–678.

[670] Redmond, U. and Cunningham, P. (2013). A temporal network analysis reveals the unprofitability of arbitrage in the prosper marketplace. *Expert Systems with Applications*, 40(9):3715–3721.

[671] Reed, S. E., Zhang, Y., Zhang, Y., and Lee, H. (2015). Deep visual analogy-making. In Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 28, pages 1252–1260. Curran Associates, Inc.

[672] Rezaei, A., Liu, A., Memarrast, O., and Ziebart, B. (2021). Robust fairness under covariate shift. NeurIPS 2020 workshop: "Algorithmic Fairness through the Lens of Causality and Interpretability (AFCI)".

[673] Ribeiro, M. H., Ottoni, R., West, R., Almeida, V. A. F., and Meira, W. (2020). Auditing radicalization pathways on youtube. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 131–141, New York, NY, USA. Association for Computing Machinery.

[674] Riederer, C. and Chaintreau, A. (2017). The price of fairness in location based advertising. RecSys 2017 workshop: "Workshop on Responsible Recommendation (FAT/Rec)".

[675] Robertson, S. E. and Zaragoza, U. (2009). The Probabilistic Relevance Framework: BM25 and Beyond. *Foundations and Trends in Information Retrieval (FnTIR)*, 3(4):333–389.

[676] Rocher, L., Hendrickx, J. M., and De Montjoye, Y.-A. (2019). Estimating the success of re-identifications in incomplete datasets using generative models. *Nature communications*, 10(1):1–9.

[677] Rodolfa, K. T., Salomon, E., Haynes, L., Mendieta, I. H., Larson, J., and Ghani, R. (2020). Case study: Predictive fairness to reduce misdemeanor recidivism through social service interventions. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 142–153, New York, NY, USA. Association for Computing Machinery.

[678] Roh, Y., Lee, K., Whang, S., and Suh, C. (2020). FR-train: A mutual information-based approach to fair and robust training. In III, H. D. and Singh, A., editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 8147–8157, Virtual. PMLR.

[679] Roh, Y., Lee, K., Whang, S. E., and Suh, C. (2021). Fairbatch: Batch selection for model fairness. In *International Conference on Learning Representations*.

[680] Romano, Y., Bates, S., and Candes, E. (2020). Achieving equalized odds by resampling sensitive attributes. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 361–371. Curran Associates, Inc.

[681] Romei, A. and Ruggieri, S. (2014). A multidisciplinary survey on discrimination analysis. *The Knowledge Engineering Review*, 29(5):582–638.

[682] Rotemberg, V., Kurtansky, N., Betz-Stablein, B., Caffery, L., Chousakos, E., Codella, N., Combalia, M., Dusza, S., Guitera, P., Gutman, D., et al. (2021). A patient-centric dataset of images and metadata for identifying melanomas using clinical context. *Scientific data*, 8(1):1–8.

[683] Rozemberczki, B., Allen, C., and Sarkar, R. (2021). Multi-scale attributed node embedding. *Journal of Complex Networks*, 9(2):cnab014.

[684] Rudinger, R., May, C., and Van Durme, B. (2017). Social bias in elicited natural language inferences. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 74–79, Valencia, Spain. Association for Computational Linguistics.

[685] Rudinger, R., Naradowsky, J., Leonard, B., and Van Durme, B. (2018). Gender bias in coreference resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 8–14, New Orleans, Louisiana. Association for Computational Linguistics.

[686] Ruoss, A., Balunovic, M., Fischer, M., and Vechev, M. (2020). Learning certified individually fair representations. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 7584–7596. Curran Associates, Inc.

[687] Russell, C., Kusner, M. J., Loftus, J., and Silva, R. (2017). When worlds collide: Integrating different counterfactual assumptions in fairness. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30, pages 6414–6423. Curran Associates, Inc.

[688] Sabato, S. and Yom-Tov, E. (2020). Bounding the fairness and accuracy of classifiers from population statistics. In III, H. D. and Singh, A., editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 8316–8325, Virtual. PMLR.

[689] Saenko, K., Kulis, B., Fritz, M., and Darrell, T. (2010). Adapting visual category models to new domains. In *Proceedings of the 11th European Conference on Computer Vision: Part IV*, ECCV'10, page 213–226, Berlin, Heidelberg. Springer-Verlag.

[690] Saerens, M., Latinne, P., and Decaestecker, C. (2002). Adjusting the outputs of a classifier to new a priori probabilities: A simple procedure. *Neural Computation*, 14(1):21–41.

[691] Sagawa, S., Koh, P. W., Hashimoto, T. B., and Liang, P. (2020). Distributionally robust neural networks. In *International Conference on Learning Representations*.

[692] Salton, G. and McGill, M. J. (1983). *Introduction to Modern Information Retrieval*. McGraw-Hill, New York, USA.

[693] Samadi, S., Tantipongpipat, U., Morgenstern, J. H., Singh, M., and Vempala, S. (2018). The price of fair pca: One extra dimension. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 31, pages 10976–10987. Curran Associates, Inc.

[694] Sambasivan, N., Kapania, S., Highfill, H., Akrong, D., Paritosh, P., and Aroyo, L. M. (2021). "everyone wants to do the model work, not the data work": Data cascades in high-stakes ai. In *proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–15.

[695] Sanderson, M., Paramita, M. L., Clough, P., and Kanoulas, E. (2010). Do user preferences and evaluation measures line up? In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 555–562.

[696] Sarvi, F., Heuss, M., Aliannejadi, M., Schelter, S., and de Rijke, M. (2022). Understanding and mitigating the effect of outliers in fair ranking. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, WSDM '22, page 861–869, New York, NY, USA. Association for Computing Machinery.

[697] Sato, M. (2022). Instagram will start asking some users for their race and ethnicity.

[698] Savani, Y., White, C., and Govindarajulu, N. S. (2020). Intra-processing methods for debiasing neural networks. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

[699] Sayans-Jiménez, P., van Harreveld, F., Dalege, J., and Rojas Tejada, A. J. (2019). Investigating stereotype structure with empirical network models. *European Journal of Social Psychology*, 49(3):604–621.

[700] Scheuerman, M. K., Hanna, A., and Denton, E. (2021). Do datasets have politics? disciplinary values in computer vision dataset development. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–37.

[701] Scheuerman, M. K., Wade, K., Lustig, C., and Brubaker, J. R. (2020). How we've taught algorithms to see identity: Constructing race and gender in image databases for facial analysis. *Proc. ACM Hum.-Comput. Interact.*, 4(CSCW1).

[702] Schouten, B., Bethlehem, J., Beullens, K., Kleven, O., Loosveldt, G., Luiten, A., Rutar, K., Shlomo, N., and Skinner, C. (2012). Evaluating, comparing, monitoring, and improving representativeness of survey response through R-indicators and partial R-indicators. *International Statistical Review*, 80(3):382–399.

[703] Schouten, B., Cobben, F., and Bethlehem, J. (2009). Indicators for the representativeness of survey response. *Survey Methodology*, 35(1):101–113.

[704] Schumann, C., Ricco, S., Prabhu, U., Ferrari, V., and Pantofaru, C. (2021). *A Step Toward More Inclusive People Annotations for Fairness*, page 916–925. Association for Computing Machinery, New York, NY, USA.

[705] Schutzman, Z. (2020). Trade-offs in fair redistricting. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, AIES '20, page 159–165, New York, NY, USA. Association for Computing Machinery.

[706] Segal, S., Adi, Y., Pinkas, B., Baum, C., Ganesh, C., and Keshet, J. (2021). *Fairness in the Eyes of the Data: Certifying Machine-Learning Models*, page 926–935. Association for Computing Machinery, New York, NY, USA.

[707] Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., and Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems. In *Proceedings of the 2nd ACM Conference on Fairness, Accountability, and Transparency (FAT* 2019)*, pages 59–68, Atlanta, US.

[708] Sen, P., Namata, G., Bilgic, M., Getoor, L., Galligher, B., and Eliassi-Rad, T. (2008). Collective classification in network data. *AI magazine*, 29(3):93–93.

[709] Serafini, D., Medori, V., Cosconati, M., Scialanga, G. L., Visani, C., Ianni, A., and Matarazzo, L. (2020). Iper: L'andamento dei prezzi effettivi per la garanzia r.c.auto nel primo trimestre 2020. Bollettino Statistico 6, IVASS, Rome.

[710] Shah, K., Gupta, P., Deshpande, A., and Bhattacharyya, C. (2021). *Rawlsian Fair Adaptation of Deep Learning Classifiers*, page 936–945. Association for Computing Machinery, New York, NY, USA.

[711] Shang, J., Sun, M., and Lam, N. S. (2020). *List-Wise Fairness Criterion for Point Processes*, page 1948–1958. Association for Computing Machinery, New York, NY, USA.

[712] Sharifi-Malvajerdi, S., Kearns, M., and Roth, A. (2019). Average individual fairness: Algorithms, generalization and experiments. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 32, pages 8242–8251. Curran Associates, Inc.

[713] Sharma, S., Gee, A. H., Paydarfar, D., and Ghosh, J. (2021). *FaiR-N: Fair and Robust Neural Networks for Structured Data*, page 946–955. Association for Computing Machinery, New York, NY, USA.

[714] Sharma, S., Henderson, J., and Ghosh, J. (2020a). Certifai: A common framework to provide explanations and analyse the fairness and robustness of black-box models. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, AIES '20, page 166–172, New York, NY, USA. Association for Computing Machinery.

[715] Sharma, S., Zhang, Y., Ríos Aliaga, J. M., Bouneffouf, D., Muthusamy, V., and Varshney, K. R. (2020b). Data augmentation for discrimination prevention and bias disambiguation. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, AIES '20, page 358–364, New York, NY, USA. Association for Computing Machinery.

[716] Shekhar, S., Shah, N., and Akoglu, L. (2021). Fairod: Fairness-aware outlier detection. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '21, page 210–220, New York, NY, USA. Association for Computing Machinery.

[717] Shen, J. H., Fratamico, L., Rahwan, I., and Rush, A. M. (2018). Darling or babygirl? investigating stylistic bias in sentiment analysis. KDD 2018 workshop: "Fairness, Accountability, and Transparency in Machine Learning (FAT/ML)".

[718] Shermis, M. D. (2014). State-of-the-art automated essay scoring: Competition, results, and future directions from a united states demonstration. *Assessing Writing*, 20:53–76.

[719] Shrum, L. J. (1995). Assessing the social influence of television: A social cognition perspective on cultivation effects. *Communication Research*, 22(4):402–429.

[720] Sindreu, J. (2021). Covid-19 wrecked the algorithms that set airfares, but they won't stay dumb. The Wall Street Journal, May 17, 2021.

[721] Singh, A. and Joachims, T. (2018a). Fairness of exposure in rankings. In *Proc. of KDD 2018*, page 2219–2228. ACM.

[722] Singh, A. and Joachims, T. (2018b). Fairness of exposure in rankings. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '18, page 2219–2228, New York, NY, USA. Association for Computing Machinery.

[723] Singh, A. and Joachims, T. (2019). Policy learning for fairness in ranking. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 32, pages 5426–5436. Curran Associates, Inc.

[724] Singh, H., Singh, R., Mhasawade, V., and Chunara, R. (2021). Fairness violations and mitigation under covariate shift. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 3–13, New York, NY, USA. Association for Computing Machinery.

[725] Singh, M. and Ramamurthy, K. N. (2019). Understanding racial bias in health using the medical expenditure panel survey data. NeurIPS 2019 workshop: "Fair ML for Health".

[726] Sink, A. and Mastro, D. (2017). Depictions of gender on primetime television: A quantitative content analysis. *Mass Communication and Society*, 20(1):3–22.

[727] Skeem, J. L. and Lowenkamp, C. T. (2016). Risk, race, and recidivism: Predictive bias and disparate impact. *Criminology*, 54(4):680–712.

[728] Slack, D., Friedler, S., and Givental, E. (2019a). Fair meta-learning: Learning how to learn fairly. NeurIPS 2019 workshop: "Human-Centric Machine Learning".

[729] Slack, D., Friedler, S., and Givental, E. (2019b). Fairness warnings. NeurIPS 2019 workshop: “Human-Centric Machine Learning".

[730] Slack, D., Friedler, S. A., and Givental, E. (2020). Fairness warnings and fair-maml: Learning fairly with minimal data. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* ’20, page 200–209, New York, NY, USA. Association for Computing Machinery.

[731] Slunge, D. (2015). The willingness to pay for vaccination against tick-borne encephalitis and implications for public health policy: evidence from sweden. *PloS one*, 10(12):e0143875.

[732] Smart, M. J. and Klein, N. J. (2020). Disentangling the role of cars and transit in employment and labor earnings. *Transportation*, 47(3):1275–1309.

[733] Smith, J. W., Everhart, J., Dickson, W., Knowler, W., and Johannes, R. (1988). Using the adap learning algorithm to forecast the onset of diabetes mellitus. *Proceedings. Symposium on Computer Applications in Medical Care*, page 261—265.

[734] Sokol, K., Santos-Rodriguez, R., and Flach, P. (2022). Fat forensics: A python toolbox for algorithmic fairness, accountability and transparency. *Software Impacts*, page 100406.

[735] Solans, D., Fabbri, F., Calsamiglia, C., Castillo, C., and Bonchi, F. (2021). *Comparing Equity and Effectiveness of Different Algorithms in an Application for the Room Rental Market*, page 978–988. Association for Computing Machinery, New York, NY, USA.

[736] Sonboli, N. and Burke, R. (2019). Localized fairness in recommender systems. In *Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization*, UMAP’19 Adjunct, page 295–300, New York, NY, USA. Association for Computing Machinery.

[737] Sonboli, N., Burke, R., Mattei, N., Eskandanian, F., and Gao, T. (2020). "and the winner is...": Dynamic lotteries for multi-group fairness-aware recommendation. RecSys 2020 workshop: “3rd FAccTRec Workshop on Responsible Recommendation".

[738] Speakman, S., Sridharan, S., and Markus, I. (2018). Three population covariate shift for mobile phone-based credit scoring. In *Proceedings of the 1st ACM SIGCAS Conference on Computing and Sustainable Societies*, COMPASS ’18, New York, NY, USA. Association for Computing Machinery.

[739] Speicher, T., Ali, M., Venkatadri, G., Ribeiro, F. N., Arvanitakis, G., Benevenuto, F., Gummadi, K. P., Loiseau, P., and Mislove, A. (2018a). Potential for discrimination in online targeted advertising. In Friedler, S. A. and Wilson, C., editors, *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 5–19, New York, NY, USA. PMLR.

[740] Speicher, T., Heidari, H., Grgic-Hlaca, N., Gummadi, K. P., Singla, A., Weller, A., and Zafar, M. B. (2018b). A unified approach to quantifying algorithmic unfairness: Measuring individual & group unfairness via inequality indices. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD ’18, page 2239–2248, New York, NY, USA. Association for Computing Machinery.

[741] Squire, R. F. (2019). Measuring and correcting sampling bias in safegraph patterns for more accurate demographic analysis.

[742] Stanovsky, G., Smith, N. A., and Zettlemoyer, L. (2019). Evaluating gender bias in machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1679–1684, Florence, Italy. Association for Computational Linguistics.

[743] Steed, R. and Caliskan, A. (2021). Image representations learned with unsupervised pre-training contain human-like biases. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 701–713, New York, NY, USA. Association for Computing Machinery.

[744] Stoyanovich, J., Yang, K., and Jagadish, H. (2018). Online set selection with fairness and diversity constraints. In *Proceedings of the EDBT Conference*.

[745] Strack, B., Deshazo, J., Gennings, C., Olmo Ortiz, J. L., Ventura, S., Cios, K., and Clore, J. (2014). Impact of hba1c measurement on hospital readmission rates: Analysis of 70,000 clinical database patient records. *BioMed research international*, 2014:781670.

[746] Strmic-Pawl, H. V., Jackson, B. A., and Garner, S. (2018). Race counts: Racial and ethnic data on the u.s. census and the implications for tracking inequality. *Sociology of Race and Ethnicity*, 4(1):1–13.

[747] Sühr, T., Biega, A. J., Zehlike, M., Gummadi, K. P., and Chakraborty, A. (2019). Two-sided fairness for repeated matchings in two-sided markets: A case study of a ride-hailing platform. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, page 3082–3092, New York, NY, USA. Association for Computing Machinery.

[748] Sühr, T., Hilgard, S., and Lakkaraju, H. (2021). *Does Fair Ranking Improve Minority Outcomes? Understanding the Interplay of Human and Algorithmic Biases in Online Hiring*, page 989–999. Association for Computing Machinery, New York, NY, USA.

[749] Sun, T., Gaut, A., Tang, S., Huang, Y., ElSherief, M., Zhao, J., Mirza, D., Belding, E., Chang, K.-W., and Wang, W. Y. (2019). Mitigating gender bias in natural language processing: Literature review. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1630–1640, Florence, Italy. Association for Computational Linguistics.

[750] Sun, Y., Han, J., Gao, J., and Yu, Y. (2009). itopicmodel: Information network-integrated topic modeling. In *2009 Ninth IEEE International Conference on Data Mining*, pages 493–502.

[751] Suresh, H. and Guttag, J. V. (2019). A framework for understanding unintended consequences of machine learning. *arXiv preprint arXiv:1901.10002*.

[752] Swinger, N., De-Arteaga, M., Heffernan IV, N. T., Leiserson, M. D., and Kalai, A. T. (2019). What are the biases in my word embedding? In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '19, page 305–311, New York, NY, USA. Association for Computing Machinery.

[753] Takac, L. and Zabovsky, M. (2012). Data analysis in public social networks. In *International scientific conference and international workshop present day trends of innovations*, volume 1.

[754] Tan, Y. C. and Celis, L. E. (2019). Assessing social and intersectional biases in contextualized word representations. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 32, pages 13230–13241. Curran Associates, Inc.

[755] Tang, J., Zhang, J., Yao, L., Li, J., Zhang, l., and Su, Z. (2008). Arnetminer: Extraction and mining of academic social networks. pages 990–998.

[756] Tantipongpipat, U., Samadi, S., Singh, M., Morgenstern, J. H., and Vempala, S. (2019). Multi-criteria dimensionality reduction with applications to fairness. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 32, pages 15161–15171. Curran Associates, Inc.

[757] Tantleff-Dunn, S., Barnes, R. D., and Larose, J. G. (2011). It's not just a "woman thing:" the current state of normative discontent. *Eating disorders*, 19(5):392–402.

[758] Tasche, D. (2017). Fisher consistency for prior probability shift. *Journal of Machine Learning Research*, 18:95:1–95:32.

[759] Taskesen, B., Blanchet, J., Kuhn, D., and Nguyen, V. A. (2021). A statistical test for probabilistic fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 648–665, New York, NY, USA. Association for Computing Machinery.

[760] Tatman, R. (2017). Gender and dialect bias in YouTube's automatic captions. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 53–59, Valencia, Spain. Association for Computational Linguistics.

[761] Tavallaee, M., Bagheri, E., Lu, W., and Ghorbani, A. A. (2009). A detailed analysis of the kdd cup 99 data set. In *2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications*, pages 1–6.

[762] Taylor, G. (1997). Setting a bonus-malus scale in the presence of other rating factors. *ASTIN Bulletin*, 27(2):319–327.

[763] Team Conduent Public Safety Solutions (2018). Real time crime forecasting challenge: Post-mortem analysis challenge performance.

[764] Tjong Kim Sang, E. F. and De Meulder, F. (2003). Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

[765] Tong, S. and Kagal, L. (2020). Investigating bias in image classification using model explanations. ICML 2020 workshop: "Workshop on Human Interpretability in Machine Learning (WHI)".

[766] Toutanova, K., Chen, D., Pantel, P., Poon, H., Choudhury, P., and Gamon, M. (2015). Representing text for joint embedding of text and knowledge bases.

[767] Tresh, F., Steeden, B., de Moura, G. R., Leite, A. C., Swift, H. J., and Player, A. (2019). Endorsing and reinforcing gender and age stereotypes: The negative effect on self-rated leadership potential for women and older workers. *Frontiers in psychology*, 10.

[768] Tsang, A., Wilder, B., Rice, E., Tambe, M., and Zick, Y. (2019). Group-fairness in influence maximization. In *International Joint Conference on Artificial Intelligence*.

[769] Tsao, C. W. and Vasan, R. S. (2015). Cohort profile: The framingham heart study (fhs): overview of milestones in cardiovascular epidemiology. *International journal of epidemiology*, 44(6):1800–1813.

[770] Tschandl, P., Rosendahl, C., and Kittler, H. (2018). The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data*, 5(1):1–9.

[771] Tsirtsis, S., Tabibian, B., Khajehnejad, M., Singla, A., Schölkopf, B., and Gomez-Rodriguez, M. (2019). Optimal decision making under strategic behavior. *arXiv preprint arXiv:1905.09239*.

[772] Tversky, A. and Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive psychology*, 5(2):207–232.

[773] Tziavelis, N., Giannakopoulos, I., Doka, K., Koziris, N., and Karras, P. (2019). Equitable stable matchings in quadratic time. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 32, pages 457–467. Curran Associates, Inc.

[774] UCI Machine Learning Repository (1994). Statlog (german credit data) data set.

[775] UCI Machine Learning Repository (1996). Adult data set.

[776] UCI Machine Learning Repository (2019). South german credit data set.

[777] UNAR - National Anti-Racial Discrimination Office (2012). Repertorio n.16 del 31 gennaio 2012. http://www.prefettura.it/FILES/AllegatiPag/1247/raccomandazione%20generale%20Tariffe%20polizze%20RCA.pdf.

[778] UNRAE - Unione Nazionale Rappresentanti Autoveicoli Esteri (2017). Il mercato italiano negli ultimi 10 anni. https://www.federmetano.it/wp-content/uploads/2018/06/AnnualReportUNRAE_2017_web_5b28b32173ff0.pdf.

[779] UNRAE - Unione Nazionale Rappresentanti Autoveicoli Esteri (2020). Immatricolazione in italia di autovetture e fuoristrada. top 10 per alimentazione - maggio 2020. http://www.unrae.it/files/maggio%20Top%2010%20alimentazione_5ed5061f56ab3.pdf.

[780] US Dept. of Commerce Bureau of the Census (1978). The current population survey: Design and methodology.

[781] US Dept. of Commerce Bureau of the Census (1995). Current population survey: Annual demographic file, 1994.

[782] US Federal Reserve (2007). Report to the congress on credit scoring and its effects on the availability and affordability of credi.

[783] Ustun, B., Liu, Y., and Parkes, D. (2019). Fairness without harm: Decoupled classifiers with preference guarantees. In Chaudhuri, K. and Salakhutdinov, R., editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6373–6382, Long Beach, California, USA. PMLR.

[784] Ustun, B., Westover, M. B., Rudin, C., and Bianchi, M. T. (2016). Clinical prediction models for sleep apnea: The importance of medical history over symptoms. *Journal of Clinical Sleep Medicine*, 12(02):161–168.

[785] V E, S. and Cho, Y. (2020). A rule-based model for seoul bike sharing demand prediction using weather data. *European Journal of Remote Sensing*, 53(sup1):166–183.

[786] V E, S., Park, J., and Cho, Y. (2020). Using data mining techniques for bike sharing demand prediction in metropolitan city. *Computer Communications*, 153:353–366.

[787] Vaithianathan, R., Putnam-Hornstein, E., Jiang, N., Nand, P., and Maloney, T. (2017). Developing predictive models to support child maltreatment hotline screening decisions: Allegheny county methodology and implementation.

[788] Valera, I., Singla, A., and Gomez Rodriguez, M. (2018). Enhancing the accuracy and fairness of human decision making. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 31, pages 1769–1778. Curran Associates, Inc.

[789] Van Horn, G., Cole, E., Beery, S., Wilber, K., Belongie, S., and Mac Aodha, O. (2021). Benchmarking representation learning for natural world image collections. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12884–12893.

[790] Van Horn, G., Mac Aodha, O., Song, Y., Cui, Y., Sun, C., Shepard, A., Adam, H., Perona, P., and Belongie, S. (2018). The inaturalist species classification and detection dataset.

[791] Vargo, A., Zhang, F., Yurochkin, M., and Sun, Y. (2021). Individually fair gradient boosting. In *International Conference on Learning Representations*.

[792] Veale, M. and Binns, R. (2017). Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data. *Big Data and Society*, 4(2):1–17.

[793] Vecchione, B., Levy, K., and Barocas, S. (2021). Algorithmic auditing and social justice: Lessons from the history of audit studies. In *Equity and Access in Algorithms, Mechanisms, and Optimization*, EAAMO '21, New York, NY, USA. Association for Computing Machinery.

[794] Verma, S., Gao, R., and Shah, C. (2020). Facets of fairness in search and recommendation. ECIR 2020 workshop: "International Workshop on Algorithmic Bias in Search and Recommendation (BIAS 2020)".

[795] Vig, J., Gehrmann, S., Belinkov, Y., Qian, S., Nevo, D., Singer, Y., and Shieber, S. M. (2020). Investigating gender bias in language models using causal mediation analysis. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

[796] Vijayaraghavan, P., Vosoughi, S., and Roy, D. (2017). Twitter demographic classification using deep multi-modal multi-task learning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 478–483, Vancouver, Canada. Association for Computational Linguistics.

[797] Voigt, R., Jurgens, D., Prabhakaran, V., Jurafsky, D., and Tsvetkov, Y. (2018). RtGender: A corpus for studying differential responses to gender. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

[798] von Kügelgen, J., Karimi, A.-H., Bhatt, U., Valera, I., Weller, A., and Schölkopf, B. (2021). On the fairness of causal algorithmic recourse. NeurIPS 2020 workshop: "Algorithmic Fairness through the Lens of Causality and Interpretability (AFCI)".

[799] Voorhees, E. (2002). Overview of thetrec 2002 question answering track. In *Proceedings of the Eleventh Text REtrieval Conference (TREC 2002)*.

[800] Voorhees, E. (2005). Overview of the trec 2005 robust retrieval track.

[801] Voorhees, E. and Harman, D. (2005). *TREC: Experiment and evaluation in information retrieval*. MIT Press.

[802] Vulić, I. and Moens, M.-F. (2015). Monolingual and cross-lingual information retrieval models based on (bilingual) word embeddings. In *Proc. of SIGIR 2015*, page 363–372. ACM.

[803] Wadsworth, C., Vera, F., and Piech, C. (2018). Achieving fairness through adversarial learning: an application to recidivism prediction. ICML 2018 workshop: "Fairness, Accountability, and Transparency in Machine Learning (FAT/ML)".

[804] Wah, C., Branson, S., Welinder, P., Perona, P., and Belongie, S. (2011). The caltech-ucsd birds200-2011 dataset. *Advances in Water Resources - ADV WATER RESOUR*.

[805] Wan, M. and McAuley, J. (2018). Item recommendation on monotonic behavior chains. In *Proceedings of the 12th ACM Conference on Recommender Systems*, RecSys '18, page 86–94, New York, NY, USA. Association for Computing Machinery.

[806] Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. (2019a). Superglue: A stickier benchmark for general-purpose language understanding systems. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

[807] Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. R. (2019b). GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *International Conference on Learning Representations*.

[808] Wang, H., Grgic-Hlaca, N., Lahoti, P., Gummadi, K. P., and Weller, A. (2019c). An empirical study on learning fairness metrics for compas data with human supervision. NeurIPS 2019 workshop: "Human-Centric Machine Learning".

[809] Wang, H., Ustun, B., and Calmon, F. (2019d). Repairing without retraining: Avoiding disparate impact with counterfactual distributions. In Chaudhuri, K. and Salakhutdinov, R., editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6618–6627, Long Beach, California, USA. PMLR.

[810] Wang, J., Liu, Y., and Levy, C. (2021a). Fair classification with group-dependent label noise. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 526–536, New York, NY, USA. Association for Computing Machinery.

[811] Wang, J., Liu, Y., and Levy, C. (2021b). Fair classification with group-dependent label noise. In *Proceedings of the 4th ACM Conference on Fairness, Accountability, and Transparency (FAccT 2021)*, pages 526–536, Toronto, CA.

[812] Wang, M. and Deng, W. (2020). Mitigating bias in face recognition using skewness-aware reinforcement learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[813] Wang, M., Deng, W., Hu, J., Tao, X., and Huang, Y. (2019e). Racial faces in the wild: Reducing racial bias by information maximization adaptation network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 692–702.

[814] Wang, S., Guo, W., Narasimhan, H., Cotter, A., Gupta, M., and Jordan, M. (2020a). Robust optimization for fairness with noisy protected groups. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 5190–5203. Curran Associates, Inc.

[815] Wang, T. and Saar-Tsechansky, M. (2020). Augmented fairness: An interpretable model augmenting decision-makers' fairness. NeurIPS 2020 workshop: "Algorithmic Fairness through the Lens of Causality and Interpretability (AFCI)".

[816] Wang, Z., Qinami, K., Karakozis, I. C., Genova, K., Nair, P., Hata, K., and Russakovsky, O. (2020b). Towards fairness in visual recognition: Effective strategies for bias mitigation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[817] Waseem, Z. and Hovy, D. (2016). Hateful symbols or hateful people? predictive features for hate speech detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93, San Diego, California. Association for Computational Linguistics.

[818] Webster, K., Recasens, M., Axelrod, V., and Baldridge, J. (2018). Mind the gap: A balanced corpus of gendered ambiguous pronouns.

[819] Weeks, M., Clair, S., Borgatti, S., Radda, K., and Schensul, J. (2002). Social networks of drug users in high-risk sites: Finding the connections. *AIDS and Behavior*, 6:193–206.

[820] Wexler, J. (2018). The what-if tool: Code-free probing of machine learning models.

[821] Wick, M., panda, s., and Tristan, J.-B. (2019). Unlocking fairness: a trade-off revisited. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 32, pages 8783–8792. Curran Associates, Inc.

[822] Wieringa, J., Kannan, P., Ma, X., Reutterer, T., Risselada, H., and Skiera, B. (2021). Data analytics in a privacy-concerned world. *Journal of Business Research*, 122:915–925.

[823] Wightman, L., Ramsey, H., and Council, L. S. A. (1998). *LSAC National Longitudinal Bar Passage Study*. LSAC research report series. Law School Admission Council.

[824] Wilder, B., Ou, H. C., de la Haye, K., and Tambe, M. (2018). Optimizing network structure for preventative health. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '18, page 841–849, Richland, SC. International Foundation for Autonomous Agents and Multiagent Systems.

[825] Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., et al. (2016). The fair guiding principles for scientific data management and stewardship. *Scientific data*, 3(1):1–9.

[826] Williams, J. V. and Razavian, N. (2019). Quantification of bias in machine learning for healthcare: A case study of renal failure prediction. NeurIPS 2019 workshop: "Fair ML for Health".

[827] Williamson, R. and Menon, A. (2019). Fairness risk measures. In Chaudhuri, K. and Salakhutdinov, R., editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6786–6797, Long Beach, California, USA. PMLR.

[828] Wilson, C., Ghosh, A., Jiang, S., Mislove, A., Baker, L., Szary, J., Trindel, K., and Polli, F. (2021). Building and auditing fair algorithms: A case study in candidate screening. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 666–677, New York, NY, USA. Association for Computing Machinery.

[829] Wohlin, C., Runeson, P., Höst, M., Ohlsson, M. C., Regnell, B., and Wesslén, A. (2012). *Experimentation in software engineering*. Springer Science & Business Media.

[830] Wondracek, G., Holz, T., Kirda, E., and Kruegel, C. (2010). A practical attack to de-anonymize social network users. In *2010 IEEE Symposium on Security and Privacy*, pages 223–238.

[831] Wu, Y., Zhang, L., and Wu, X. (2018a). On discrimination discovery and removal in ranked data using causal graph. In *Proc. of KDD 2018*, pages 2536–2544.

[832] Wu, Y., Zhang, L., and Wu, X. (2018b). On discrimination discovery and removal in ranked data using causal graph. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '18, page 2536–2544, New York, NY, USA. Association for Computing Machinery.

[833] Wu, Y., Zhang, L., Wu, X., and Tong, H. (2019). Pc-fairness: A unified framework for measuring causality-based fairness. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 32, pages 3404–3414. Curran Associates, Inc.

[834] Xiao, H., Rasul, K., and Vollgraf, R. (2017). Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms.

[835] Xiao, W., Zhao, H., Pan, H., Song, Y., Zheng, V. W., and Yang, Q. (2019). Beyond personalization: Social content recommendation for creator equality and consumer satisfaction. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, page 235–245, New York, NY, USA. Association for Computing Machinery.

[836] Xie, M. and Lauritsen, J. L. (2012). Racial context and crime reporting: A test of Black's stratification hypothesis. *Journal of Quantitative Criminology*, 28(2):265–293.

[837] Xu, D., Yuan, S., Zhang, L., and Wu, X. (2018). Fairgan: Fairness-aware generative adversarial networks. In *2018 IEEE International Conference on Big Data (Big Data)*, pages 570–575. IEEE.

[838] Xu, R., Cui, P., Kuang, K., Li, B., Zhou, L., Shen, Z., and Cui, W. (2020). *Algorithmic Decision Making with Conditional Fairness*, page 2125–2135. Association for Computing Machinery, New York, NY, USA.

[839] Xu, X., Huang, Y., Shen, P., Li, S., Li, J., Huang, F., Li, Y., and Cui, Z. (2021). Consistent instance false positive improves fairness in face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 578–586.

[840] Yan, C., Wang, X., Liu, X., Liu, W., and Liu, J. (2020). Research on the ubi car insurance rate determination model based on the cnn-hvsvm algorithm. *IEEE Access*, 8:160762–160773.

[841] Yang, F., Cisse, M., and Koyejo, S. (2020a). Fairness with overlapping groups; a probabilistic perspective. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 4067–4078. Curran Associates, Inc.

[842] Yang, K., Qinami, K., Fei-Fei, L., Deng, J., and Russakovsky, O. (2020b). Towards fairer datasets: Filtering and balancing the distribution of the people subtree in the imagenet hierarchy. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 547–558, New York, NY, USA. Association for Computing Machinery.

[843] Yang, K. and Stoyanovich, J. (2017a). Measuring fairness in ranked outputs. In *Proceedings of the 29th International Conference on Scientific and Statistical Database Management*, SSDBM '17, New York, NY, USA. Association for Computing Machinery.

[844] Yang, K. and Stoyanovich, J. (2017b). Measuring fairness in ranked outputs. In *Proc. of the 29th International Conference on Scientific and Statistical Database Management*, pages 1–6.

[845] Yang, M. and Kim, B. (2019). Benchmarking attribution methods with relative feature importance.

[846] Yao, S. and Huang, B. (2017a). Beyond parity: Fairness objectives for collaborative filtering. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30, pages 2921–2930. Curran Associates, Inc.

[847] Yao, S. and Huang, B. (2017b). New fairness metrics for recommendation that embrace differences. KDD 2017 workshop: "Fairness, Accountability, and Transparency in Machine Learning (FAT/ML)".

[848] Yeh, I.-C. and hui Lien, C. (2009). The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Systems with Applications*, 36(2, Part 1):2473 – 2480.

[849] Yi, S., Wang, S., Joshi, S., and Ghassemi, M. (2019). Fair and robust treatment effect estimates: Estimation under treatment and outcome disparity with deep neural models. NeurIPS 2019 workshop: "Fair ML for Health".

[850] Yi, S., Xiaogang, W., and Xiaoou, T. (2013). Deep convolutional network cascade for facial point detection. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3476–3483.

[851] Yu, H.-T., Jatowt, A., Blanco, R., Joho, H., Jose, J. M., Chen, L., and Yuan, F. (2018). Revisiting the cluster-based paradigm for implicit search result diversification. *Information Processing & Management*, 54(4):507–528.

[852] Yurochkin, M., Bower, A., and Sun, Y. (2020). Training individually fair ml models with sensitive subspace robustness. In *International Conference on Learning Representations*.

[853] Yurochkin, M. and Sun, Y. (2021). Sensei: Sensitive set invariance for enforcing individual fairness. In *International Conference on Learning Representations*.

[854] Zafar, M. B., Valera, I., Gomez Rodriguez, M., and Gummadi, K. P. (2017a). Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th International Conference on World Wide Web*, WWW '17, page 1171–1180, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

[855] Zafar, M. B., Valera, I., Rodriguez, M., Gummadi, K., and Weller, A. (2017b). From parity to preference-based notions of fairness in classification. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30, pages 229–239. Curran Associates, Inc.

[856] Zafar, M. B., Valera, I., Rogriguez, M. G., and Gummadi, K. P. (2017c). Fairness constraints: Mechanisms for fair classification. In *Artificial Intelligence and Statistics*, pages 962–970. PMLR.

[857] Zehlike, M., Bonchi, F., Castillo, C., Hajian, S., Megahed, M., and Baeza-Yates, R. (2017a). Fa* ir: A fair top-k ranking algorithm. In *Proc. of CIKM 2017*, pages 1569–1578.

[858] Zehlike, M., Bonchi, F., Castillo, C., Hajian, S., Megahed, M., and Baeza-Yates, R. (2017b). Fa* ir: A fair top-k ranking algorithm. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 1569–1578.

[859] Zehlike, M., Sühr, T., Baeza-Yates, R., Bonchi, F., Castillo, C., and Hajian, S. (2022a). Fair top-k ranking with multiple protected groups. *Information Processing & Management*, 59(1):102707.

[860] Zehlike, M., Yang, K., and Stoyanovich, J. (2022b). Fairness in ranking, part i: Score-based ranking. *ACM Comput. Surv.* Just Accepted.

[861] Zehlike, M., Yang, K., and Stoyanovich, J. (2022c). Fairness in ranking, part ii: Learning-to-rank and recommender systems. *ACM Comput. Surv.* Just Accepted.

[862] Zhai, C. (2008). Statistical Language Models for Information Retrieval. A Critical Review. *Foundations and Trends in Information Retrieval (FnTIR)*, 2(3):137–213.

[863] Zhang, B. H., Lemoine, B., and Mitchell, M. (2018). Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18, page 335–340, New York, NY, USA. Association for Computing Machinery.

[864] Zhang, H. and Davidson, I. (2021). Towards fair deep anomaly detection. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 138–148, New York, NY, USA. Association for Computing Machinery.

[865] Zhang, H., Lu, A. X., Abdalla, M., McDermott, M., and Ghassemi, M. (2020a). Hurtful words: Quantifying biases in clinical contextual word embeddings. In *Proceedings of the ACM Conference on Health, Inference, and Learning*, CHIL '20, page 110–120, New York, NY, USA. Association for Computing Machinery.

[866] Zhang, J. and Bareinboim, E. (2018). Equality of opportunity in classification: A causal approach. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 31, pages 3671–3681. Curran Associates, Inc.

[867] Zhang, L., Wu, Y., and Wu, X. (2017a). Achieving non-discrimination in data release. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '17, page 1335–1344, New York, NY, USA. Association for Computing Machinery.

[868] Zhang, X., Khaliligarekani, M., Tekin, C., and liu, m. (2019). Group retention when using machine learning in sequential decision making: the interplay between user dynamics and fairness. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 32, pages 15269–15278. Curran Associates, Inc.

[869] Zhang, X., Tu, R., Liu, Y., Liu, M., Kjellström, H., Zhang, K., and Zhang, C. (2020b). How do fair decisions fare in long-term qualification? In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

[870] Zhang, Y. (2005). *Bayesian Graphical Model for Adaptive Information Filtering*. PhD thesis, Carnegie Mellon University.

[871] Zhang, Y., Bellamy, R., and Varshney, K. (2020c). Joint optimization of ai fairness and utility: A human-centered approach. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, AIES '20, page 400–406, New York, NY, USA. Association for Computing Machinery.

[872] Zhang, Z., Luo, P., Loy, C. C., and Tang, X. (2014). Facial landmark detection by deep multi-task learning.

[873] Zhang, Z., Luo, P., Loy, C. C., and Tang, X. (2015). Learning deep representation for face alignment with auxiliary attributes. *IEEE transactions on pattern analysis and machine intelligence*, 38(5):918–930.

[874] Zhang, Z. and Neill, D. B. (2017). Identifying significant predictive bias in classifiers. KDD 2017 workshop: "Fairness, Accountability, and Transparency in Machine Learning (FAT/ML)".

[875] Zhang, Z., Song, Y., and Qi, H. (2017b). Age progression/regression by conditional adversarial autoencoder. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[876] Zhao, B., Xiao, X., Gan, G., Zhang, B., and Xia, S.-T. (2020a). Maintaining discrimination and fairness in class incremental learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[877] Zhao, C., Li, C., Li, J., and Chen, F. (2020b). Fair meta-learning for few-shot classification. In *2020 IEEE International Conference on Knowledge Graph (ICKG)*, pages 275–282.

[878] Zhao, H., Coston, A., Adel, T., and Gordon, G. J. (2020c). Conditional learning of fair representations. In *International Conference on Learning Representations*.

[879] Zhao, H. and Gordon, G. (2019). Inherent tradeoffs in learning fair representations. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 32, pages 15675–15685. Curran Associates, Inc.

[880] Zhao, J., Wang, T., Yatskar, M., Ordonez, V., and Chang, K.-W. (2017). Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2979–2989, Copenhagen, Denmark. Association for Computational Linguistics.

[881] Zhao, J., Wang, T., Yatskar, M., Ordonez, V., and Chang, K.-W. (2018a). Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20, New Orleans, Louisiana. Association for Computational Linguistics.

[882] Zhao, J., Zhou, Y., Li, Z., Wang, W., and Chang, K.-W. (2018b). Learning gender-neutral word embeddings. In *Proc. of EMNLP 2018*, pages 4847–4853. ACL.

[883] Zhao, Y., Kong, S., and Fowlkes, C. (2021). Camera pose matters: Improving depth prediction by mitigating pose distribution bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15759–15768.

[884] Zheng, Y., Dave, T., Mishra, N., and Kumar, H. (2018). Fairness in reciprocal recommendations: A speed-dating study. In *Adjunct Publication of the 26th Conference on User Modeling, Adaptation and Personalization*, UMAP '18, page 29–34, New York, NY, USA. Association for Computing Machinery.

[885] Zhong, Y., Deng, W., Wang, M., Hu, J., Peng, J., Tao, X., and Huang, Y. (2019). Unequal-training for deep face recognition with long-tailed noisy data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[886] Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., and Torralba, A. (2018). Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(6):1452–1464.

[887] Zhou, K., Xue, G.-R., Zha, H., and Yu, Y. (2008). Learning to rank with ties. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '08, page 275–282, New York, NY, USA. Association for Computing Machinery.

[888] Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A., and Fidler, S. (2015). Aligning books and movies: Towards story-like visual explanations by watching movies and reading books.

[889] Zhu, Z., Wang, J., Zhang, Y., and Caverlee, J. (2018). Fairness-aware recommendation of information curators. RecSys 2018 workshop: "Workshop on Responsible Recommendation (FAT/Rec)".

[890] Žliobaité, I. (2015). On the relation between accuracy and fairness in binary classification. ICML 2015 workshop: "Fairness, Accountability, and Transparency in Machine Learning (FAT/ML)".

[891] Žliobaité, I., Kamiran, F., and Calders, T. (2011). Handling conditional discrimination. In *2011 IEEE 11th International Conference on Data Mining*, pages 992–1001.

Appendix A

Supplementary Materials to Chapter 3

In this appendix, we report the data briefs and benchmark documentation informing Chapter 3.

A.1 Data briefs

Data briefs were drafted by the first author and reviewed by the remaining authors. For over 95% of the surveyed datasets, we identified at least one contact involved in the data curation process or familiar with the dataset, who received a preliminary version of the respective data brief and a request for corrections and additions. Their contributions are acknowledged at the end of this section. Data briefs are meant as short documentation providing essential information on datasets used in fairness research. Data briefs are composed of ten fields derived from shared vocabularies such as Data Catalog Vocabulary (DCAT)[1]; to be compliant with the FAIR data principles [825], we also defined a schema (with namespace `fdo`) to model the relationships between the terms, to make the links to external vocabularies explicit, and map the data briefs to a machine-readable RDF graph.[2] The `fdo` schema has been defined by reusing, as much as possible, existing terminology from established vocabularies. In the following we detail the fields of the data briefs and present their correspondence to DCAT and `fdo` properties:

Description. This is a free-text field reporting (1) the aim/purpose of a data artifact (i.e., why it was developed/collected), as stated by curators or inferred from context; (2) a high-level description of the available features; (3) the labeling procedure for annotated

[1] http://www.w3.org/ns/dcat, with namespace `dct`

[2] Schema publicly available at https://fairnessdatasets.dei.unipd.it/schema/; RDF graph publicly available at https://zenodo.org/record/6518370#.YnOSKFTMJhF. To favour consultation and dynamical querying of the data briefs, we are working to release a web app at https://fairnessdatasets.dei.unipd.it.

attributes, with special attention to sensitive ones, if any; (4) the envisioned ML task, if any. Corresponds to `dct:description` in DCAT.

Affiliation of creators. Typically derived from reports, articles, or official web pages presenting a dataset. Datasets can be derivatives of other datasets (e.g., Adult). We typically refer to the final resource while providing the prior context where appropriate. In DCAT vocabulary, it is the affiliation of a `dct:publisher` (for published resources) or a `dct:creator`.

Domain. The main field where the data is used (e.g., computer vision for ImageNet) or the field studying the processes and phenomena that produced the dataset (e.g., radiology for CheXpert). Corresponds to `fdo:Domain` in the `fdo` schema.

Tasks in fairness literature. An indication of the task performed on the dataset in each surveyed article that uses the current resource. Corresponds to `fdo:Task`.

Data spec. The main format of the data. The envisioned categories are text, image, time-series, tabular data, and pairs. The latter denotes a special type of tabular data where rows and columns correspond to entities and cells to a relation between them, such as relevance for query-document pairs, ratings for user-item pairs, co-authorship relation for author-author pairs. A "mixture" category was added for resources with multimodal data. Corresponds to `dct:type` in DCAT.

Sample size. Dataset cardinality. Corresponds to `fdo:sampleSize` in `fdo`.

Year. Last known update to the dataset. For resources whose collection and curation are ongoing (e.g., Framingham) we write "present". Corresponds to `dct:modified`.

Sensitive features. Sensitive attributes in the dataset. These are typically explicitly annotated, but may include implicit ones, such as textual references to people and their demographics in text datasets. References to gender, for instance, can easily be retrieved from English-language text corpora based on intrinsically gendered words, such as she, man, aunt. Corresponds to `fdo:sensitiveFeature`.

Link. A link to the website where the resource can be downloaded or requested. Corresponds to `dcat:landingPage`.

Further information. Reference to works and web pages describing the dataset.

Following the algorithmic fairness literature, we define sensitive features as encoding membership to groups that are salient for society and have some special protection based

on the law, including race, ethnicity, sex, gender, and age. We may occasionally stretch this definition and report features considered sensitive in some works, such as political leaning or education, so long as they reflect essential divisions in society. We also report domain-specific attributes considered sensitive in a given context, such as language for Section 203 determinations or brand ownership for Amazon Recommendations. We follow the language of the available documentation for the names and values of sensitive features, including distinctions between race and ethnicity. For datasets that report geographical information at any granularity (GPS coordinates, neighbourhoods, countries) we report "geography" among the sensitive attributes. If an article considers features to be sensitive in an arbitrary fashion (e.g., sepal width in the Iris dataset), we do not report it in the respective field.

For the dataset domain, we follow the area-category taxonomy defined by Scimago,[3] with the addition of "news", "social media", "social networks", "sports" and "food". Table 3.2 contains a summary of the surveyed datasets through this domain-based taxonomy. Tasks in the fairness literature were labeled via open coding. The final taxonomy is detailed in Section 3.4. We distinguish between works that are more focused on evaluation rather than a proposal of novel solutions by writing, e.g. "fair ranking evaluation" instead of "fair ranking". We use "evaluation" as a broad term for works focusing on analyses of algorithms, products, platforms, or datasets and their properties from multiple fairness and accuracy perspectives. With some abuse of nomenclature, we also use this label for works that focus on properties of fairness metrics [642]. Unless otherwise specified, "fairness evaluation" is about fair classification, which is the most common task. Exploratory approaches focused on discovering biases that are not fully specified ex-ante are indicated with the label "bias discovery".

We would like to thank the following researchers and dataset creators for the useful feedback on the data briefs: Alain Barrat, Luc Behaghel, Asia Biega, Marko Bohanec, Chris Burgess, Robin Burke, Alejandro Noriega Campero, Margarida Carvalho, Abhijnan Chakraborty, Robert Cheetham, Won Ik Cho, Paulo Cortez, Thomas Davidson, Maria De-Arteaga, Lucas Dixon, Danijela Djordjević, Michele Donini, Marco Duarte, Natalie Ebner, Elaine Fehrman, H. Altay Guvenir, Moritz Hardt, Irina Higgins, Yu Hen Hu, Rachel Huddart, Lalana Kagal, Dean Karlan, Vijay Keswani, Been Kim, Hyunjik Kim, Jiwon Kim, Svetlana Kiritchenko, Pang Wei Koh, Joseph A. Konstan, Varun Kumar, Jeremy Andrew Irvin, Jamie N. Larson, Jure Leskovec, Jonathan Levy, Andrea Lodi, Oisin Mac Aodha, Loic Matthey, Julian McAuley, Brendan McMahan, Sergio Moro, Luca Oneto, Orestis Papakyriakopoulos, Stephen Robert Pfohl, Christopher G. Potts, Mike Redmond, Kit Rodolfa, Ben Roshan, Veronica Rotemberg, Rachel Rudinger, Sivan Sabato, Kate Saenko, Mark D. Shermis, Daniel

[3]https://www.scimagojr.com/journalrank.php

Slunge, David Solans, Luca Soldaini, Efstathios Stamatatos, Ryan Steed, Rachael Tatman, Schrasing Tong, Alan Tsang, Sathishkumar V E, Andreas van Cranenburgh, Lucy Vasserman, Roland Vollgraf, Alex Wang, Zeerak Waseem, Kellie Webster, Bryan Wilder, Nick Wilson, I-Cheng Yeh, Elad Yom-Tov, Neil Yorke-Smith, Michal Zabovsky, Yukun Zhu.

A.1.1 2010 Frequently Occurring Surnames

- **Description**: this dataset reports all surnames occurring 100 or more times in the 2010 US Census, broken down by race (White, Black, Asian and Pacific Islander (API), American Indian and Alaskan Native only (AIAN), multiracial, or Hispanic).
- **Affiliation of creators**: US Census Bureau.
- **Domain**: linguistics.
- **Tasks in fairness literature**: fair subset selection under unawareness [553].
- **Data spec**: tabular data.
- **Sample size**: $\sim$ 200K surnames.
- **Year**: 2016.
- **Sensitive features**: race.
- **Link**: https://www.census.gov/topics/population/genealogy/data/2010_surnames.html
- **Further info**: https://www2.census.gov/topics/genealogy/2010surnames/surnames.pdf

A.1.2 2016 US Presidential Poll

- **Description**: this dataset was collected and maintained by FiveThirtyEight, a website specialized in opinion poll analysis. This resource was developed with the goal of providing an aggregated estimate based on multiple polls, weighting each input according to sample size, recency, and historical accuracy of the polling organization. For each poll, the dataset provides the period of data collection, its sample size, the pollster conducting it, their rating, and a url linking to the source data.
- **Affiliation of creators**: FiveThirtyEight.
- **Domain**: political science.
- **Tasks in fairness literature**: limited-label fairness evaluation [688].
- **Data spec**: tabular data.
- **Sample size**: $\sim$ 13K poll results.
- **Year**: 2016.
- **Sensitive features**: geography.
- **Link**: http://projects.fivethirtyeight.com/general-model/president_general_polls_2016.csv
- **Further info**: https://projects.fivethirtyeight.com/2016-election-forecast/

A.1.3 4area

- **Description**: this dataset was extracted from DBLP to study the problem of topic modeling on documents connected by links in a graph structure. The creators extracted from DBLP articles published at 20 major conferences from four related areas, i.e., database, data mining, machine learning, and information retrieval. Each author is associated with four continuous variables based on the fraction of research papers published in these areas. The associated task is the prediction of these attributes.
- **Affiliation of creators**: University of Illinois at Urbana-Champaign.

- **Domain**: library and information sciences.
- **Tasks in fairness literature**: fair clustering [343].
- **Data spec**: author-author pairs.
- **Sample size**: $\sim$ 30K nodes (authors) connected by $\sim$ 200K edges (co-author relations).
- **Year**: 2009.
- **Sensitive features**: author.
- **Link**: not available
- **Further info**: Sun et al. [750]

A.1.4 Academic Collaboration Networks

- **Description**: these dataset represent two collaboration networks from the preprint server arXiv, covering scientific papers submitted to the astrophysics (AstroPh) and condensed matter (CondMat) physics categories. Each node in the network is an author, with links indicating co-authorship of one or more articles. Nodes are indicated with ids, hence information about the researchers in the graph is not immediately available. These datasets were developed to study the evolution of graphs over time.
- **Affiliation of creators**: Carnegie Mellon University; Cornell University.
- **Domain**: library and information sciences.
- **Tasks in fairness literature**: fair graph mining [425].
- **Data spec**: author-author pairs.
- **Sample size**: $\sim$19K nodes (authors) connected by $\sim$ 200K edges (indications of co-authorship) (AstroPh). $\sim$23K nodes connected by $\sim$ 93K edges (CondMat).
- **Year**: 2009.
- **Sensitive features**: none.
- **Link**: http://snap.stanford.edu/data/ca-AstroPh.html (AstroPh) and http://snap.stanford.edu/data/ca-CondMat.html (CondMat)
- **Further info**: Leskovec et al. [493]

A.1.5 Adience

- **Description**: this resource was developed to favour the study of automated age and gender identification from images of faces. Photos were sourced from Flickr albums, among the ones automatically uploaded from iPhone and made available under Creative Commons license. All images were manually labeled for age, gender and identity "using both the images themselves and any available contextual information". These annotations are fundamental for the tasks associated with this dataset, i.e. age and gender estimation. One author of Buolamwini and Gebru [101] labeled each image in Adience with Fitzpatrick skin type.
- **Affiliation of creators**: Adience; Open University of Israel.
- **Domain**: computer vision.
- **Tasks in fairness literature**: data bias evaluation [101], robust fairness evaluation [589].
- **Data spec**: image.
- **Sample size**: $\sim$ 30K images of $\sim$ 2K subjects.
- **Year**: 2014.
- **Sensitive features**: age, gender, skin type.
- **Link**: https://talhassner.github.io/home/projects/Adience/Adience-data.html
- **Further info**: Buolamwini and Gebru [101], Eidinger et al. [228]

A.1.6 Adressa

- **Description**: this dataset was curated as part of the RecTech project on recommendation technology owned by Adresseavisen (shortened to Adressa) a large Norwegian newspaper. It summarizes one week of traffic to the newspaper website by both subscribers and non-subscribers, during February 2017. The dataset describes reading events, i.e. a reader accessing an article, providing access timestamps and user information inferred from their IP. Specific information about the articles is also available, including author, keywords, body, and mentioned entities. The dataset curators also worked on an extended version of the dataset (Adressa 20M), ten times larger than the one described here.
- **Affiliation of creators**: Norwegian University of Science and Technology; Adresseavisen.
- **Domain**: news, information systems.
- **Tasks in fairness literature**: fair ranking [131].
- **Data spec**: user-article pairs.
- **Sample size**: $\sim$ 3M ratings by $\sim$ 15M readers over $\sim 1K$ articles.
- **Year**: 2018.
- **Sensitive features**: geography.
- **Link**: http://reclab.idi.ntnu.no/dataset/
- **Further info**: [330]

A.1.7 Adult

- **Description**: this dataset was created as a resource to benchmark the performance of machine learning algorithms on socially relevant data. Each instance is a person who responded to the March 1994 US Current Population Survey, represented along demographic and socio-economic dimensions, with features describing their profession, education, age, sex, race, personal and financial condition. The dataset was extracted from the census database, preprocessed, and donated to UCI Machine Learning Repository in 1996 by Ronny Kohavi and Barry Becker. A binary variable encoding whether respondents' income is above $50,000 was chosen as the target of the prediction task associated with this resource. See Appendix A.2 for extensive documentation.
- **Affiliation of creators**: Silicon Graphics Inc.
- **Domain**: economics.
- **Tasks in fairness literature**: fairness evaluation [114, 138, 210, 261, 282, 376, 390, 396, 446, 506, 509, 527, 597, 611, 642, 706, 714, 740, 798, 827, 890], fair classification [4, 34, 111, 124, 126, 149, 156, 175, 182, 197, 216, 261, 266, 308, 319, 356, 373, 514, 534, 562, 608, 633, 652, 657, 658, 678, 679, 698, 710, 713, 715, 791, 809, 833, 838, 841, 852, 853, 856, 863, 867], fair clustering [1, 7, 31, 61, 64, 69, 98, 146, 298, 343, 378, 526, 535, 815], fair clustering under unawareness [239], fair active classification [36, 37, 600], fair preference-based classification [13, 583, 783], fair classification under unawareness [441, 478, 582, 814], fair anomaly detection [716, 864], fairness evaluation under unawareness [28], robust fairness evaluation [77], data bias evaluation [62], rich-subgroup fairness evaluation [154, 435], fair representation learning [515, 523, 653, 686, 878, 879], fair multi-stage classification [306, 375], robust fair classification [379, 529, 672], dynamical fair classification [868], fair ranking evaluation [421], fair data summarization [56, 120, 147, 233, 413, 453], fair regression [5], limited-label fair classification [152, 157, 810], limited-label fairness evaluation [403], preference-based fair clustering [285].
- **Data spec**: tabular data.
- **Sample size**: $\sim$ 50K instances.
- **Year**: 1996.

- **Sensitive features**: age, sex, race.
- **Link**: https://archive.ics.uci.edu/ml/datasets/adult
- **Further info**: Ding et al. [213], Kohavi [459], McKenna [545, 546], UCI Machine Learning Repository [775], US Dept. of Commerce Bureau of the Census [781]

A.1.8 Allegheny Child Welfare

- **Description**: this dataset stems from an initiative by the Allegheny County's Department of Human Services to develop assistive tools to support child maltreatment hotline screening decisions. Referrals received by Allegheny County via a hotline between September 2008 and April 2016 were assembled into a dataset. To obtain a relevant history and follow-up time for each referral, a subset of samples spanning the period from April 2010 to April 2014 is considered. Each data point pertains to a referral for suspected child abuse or neglect and contains a wealth of information from the integrated data management systems of Allegheny County. This data includes cross-sector administrative information for individuals associated with a report of child abuse or neglect, including data from child protective services, mental health services, drug, and alcohol services. The target to be estimated by risk models is future child harm, as measured e.g. by re-referrals, which complements the role of the screening staff who are focused on the information currently available about the referral.
- **Affiliation of creators**: Allegheny County Department of Human Services; Auckland University of Technology; University of Southern California; University of Auckland; University of California.
- **Domain**: social work.
- **Tasks in fairness literature**: fairness evaluation of risk assessment [171], fair risk assessment [568].
- **Data spec**: tabular data.
- **Sample size**: $\sim$ 80K calls.
- **Year**: 2019.
- **Sensitive features**: age, race, gender of child.
- **Link**: not available
- **Further info**: Vaithianathan et al. [787]

A.1.9 Amazon Recommendations

- **Description**: this dataset was crawled to study anti-competitive behaviour on Amazon, and the extent to which Amazon's private label products are recommended on the platform. Considering the categories *backpack* and *battery*, where Amazon is known to have a strong private label presence, the creators gathered a set of organic and sponsored recommendations from Amazon.in, exploiting snowball sampling. Metadata for each product was also collected, including user rating, number of reviews, brand, seller.
- **Affiliation of creators**: Indian Institute of Technology; Max Planck Institute for Software Systems.
- **Domain**: information systems.
- **Tasks in fairness literature**: fair ranking evaluation [189].
- **Data spec**: item-recommendation pairs.
- **Sample size**: $\sim$ 1M recommendations associated with $\sim$ 20K items.
- **Year**: 2021.
- **Sensitive features**: brand ownership.
- **Link**: not available
- **Further info**: Dash et al. [189]

A.1.10 Amazon Reviews

- **Description**: this is large-scale dataset of over ten million products and respective reviews on Amazon, spanning more than two decades. It was created to study the problem of image-based recommendation and its dynamics. Rich metadata are available for both products and reviews. Reviews consist of ratings, text, reviewer name, and review ID, while products include title, price, image, and sales rank of product.
- **Affiliation of creators**: University of California, San Diego.
- **Domain**: information systems.
- **Tasks in fairness literature**: fair ranking [625].
- **Data spec**: user-product pairs (reviews).
- **Sample size**: $\sim$ 200M reviews of products.
- **Year**: 2018.
- **Sensitive features**: none.
- **Link**: https://nijianmo.github.io/amazon/index.html
- **Further info**: He and McAuley [354], McAuley et al. [540]

A.1.11 ANPE

- **Description**: this dataset represents a large randomized controlled trial, assigning job seekers in France to a program run by the Public employment agency (ANPE), or to a program outsourced to private providers by the Unemployment insurance organization (Unédic). The data involves 400 public employment branches and over 200,000 job-seekers. Data about job seekers includes their demographics, their placement program and the subsequent duration of unemployment spells.
- **Affiliation of creators**: Paris School of Economics; Institute of Labor Economics; CREST; ANPE; Unédic; Direction de l'Animation de la Recherche et des Études Statistiques.
- **Domain**: economics.
- **Tasks in fairness literature**: fairness evaluation of risk assessment [420].
- **Data spec**: tabular data.
- **Sample size**: $\sim$ 200K job seekers.
- **Year**: 2012.
- **Sensitive features**: age, gender, nationality.
- **Link**: https://www.openicpsr.org/openicpsr/project/113904/version/V1/view?path=/openicpsr/113904/fcr:versions/V1/Archive&type=folder
- **Further info**: Behaghel et al. [55]

A.1.12 Antelope Valley Networks

- **Description**: this a set of synthetic datasets generated to study the problem of influence maximization for obesity prevention. Samples of agents are generated to emulate the demographic and obesity distribution across regions in the Antelope Valley in California, exploiting data from the US Census, the Los Angeles County Department of Public Health, and Los Angeles Times Mapping L.A. project. Each agent in the network has a geographic region, gender, ethnicity, age, and connections to other agents, which are more frequent for agents with similar attributes. Agents are also assigned a weight status, which may change based on interactions with other agents in their ego-network, emulating social learning.
- **Affiliation of creators**: National University of Singapore; National University of Southern California.
- **Domain**: public health.
- **Tasks in fairness literature**: fair graph diffusion [254].
- **Data spec**: agent-agent pairs.

- **Sample size**: $\sim$ 20 synthetic networks, containing $\sim$ 500 individuals each.
- **Year**: 2019.
- **Sensitive features**: ethnicity, gender, age, geography.
- **Link**: https://github.com/bwilder0/fair_influmax_code_release
- **Further info**: Tsang et al. [768], Wilder et al. [824]

A.1.13 Apnea

- **Description**: this dataset results from a sleep medicine study focused on establishing important factors for the automated diagnosis of Obstructive Sleep Apnea (OSA). The task associated with this dataset is the prediction of medical condition (OSA/no OSA) from available patient features, which include demographics, medical history, and symptoms.
- **Affiliation of creators**: Massachusetts Institute of Technology; Massachusetts General Hospital; Harvard Medical School.
- **Domain**: sleep medicine.
- **Tasks in fairness literature**: fair preference-based classification [783].
- **Data spec**: mixture (time series and tabular data).
- **Sample size**: $\sim$ 2K patients.
- **Year**: 2016.
- **Sensitive features**: age, sex.
- **Link**: not available
- **Further info**: Ustun et al. [784]

A.1.14 ArnetMiner Citation Network

- **Description**: this dataset is one of the many resources made available by the ArnetMiner online service. The ArnetMiner system was developed for the extraction and mining of data from academic social networks, with a focus on profiling of researchers. The DBLP Citation Network is extracted from academic resources, such as DBLP, ACM and MAG (Microsoft Academic Graph). The dataset captures the relationships between scientific articles and their authors in a connected graph structure. It can be used for tasks such as community discovery, topic modeling, centrality and influence analysis. In its latest versions, the dataset comprises over 20 fields, including paper title, keywords, abstract, venue, year, along with authors, and their affiliations. The ArnetMiner project was partially funded by the Chinese National High-tech R&D Program, the National Science Foundation of China, IBM China Research Lab, the Chinese Young Faculty Research Funding program and Minnesota China Collaborative Research Program.
- **Affiliation of creators**: Tsinghua University; IBM.
- **Domain**: library and information sciences.
- **Tasks in fairness literature**: fair graph mining [104].
- **Data spec**: article-article pairs.
- **Sample size**: $\sim$ 5M papers connected by $\sim$ 50M citations.
- **Year**: 2021.
- **Sensitive features**: author.
- **Link**: http://www.arnetminer.org/citation
- **Further info**: Tang et al. [755]; https://www.aminer.org/

A.1.15 Arrhythmia

- **Description**: data provenance for this set of patient records seems uncertain. The first work referencing this dataset dates to 1997 and details a machine learning approach for the diagnosis of arrhythmia, which presumably motivated its collection. Each data point describes a different patient; features include demographics, weight and height and clinical measurements from ECG signals, along with the diagnosis of a cardiologist into 16 different classes of arrhythmia (including none), which represents the target variable.
- **Affiliation of creators**: Bilkent University; Baskent University.
- **Domain**: cardiology.
- **Tasks in fairness literature**: fair classification [216, 535], robust fair classification [672], limited-label fair classification [157].
- **Data spec**: tabular data.
- **Sample size**: $\sim$ 500 patients.
- **Year**: 1997.
- **Sensitive features**: age, sex.
- **Link**: https://archive.ics.uci.edu/ml/datasets/arrhythmia
- **Further info**: Guvenir et al. [336]

A.1.16 Athletes and health professionals

- **Description**: the datasets were developed to study the effects of bias in image classification. The health professional dataset (doctors and nurses) contains race and gender as sensitive features and the athlete dataset (basketball and volleyball players) contains gender and jersey color as sensitive features. Each subgroup, separated by combinations of sensitive features, is roughly balanced at 200 images. The collected data was manually examined by the curators to remove stylized images and images containing both females and males.
- **Affiliation of creators**: Massachusetts Institute of Technology.
- **Domain**: computer vision.
- **Tasks in fairness literature**: bias discovery [765].
- **Data spec**: image.
- **Sample size**: $\sim$ 800 images of athletes and $\sim$ 500 images of health professionals.
- **Year**: 2020.
- **Sensitive features**: Gender (both), race (health professionals), jersey color (athletes).
- **Link**: https://github.com/ghayat2/Datasets
- **Further info**: Tong and Kagal [765]

A.1.17 Automated Student Assessment Prize (ASAP)

- **Description**: this dataset was collected to evaluate the feasibility of automated essay scoring. It consists of a collection of essays by US students in grade levels 7–10, rated by at least two human raters. The dataset comes with a predefined training/validation/test split and powers the Hewlett Foundation Automated Essay Scoring competition on Kaggle. The curators tried to remove personally identifying information from the essays using Named Entity Recognizer (NER) and several heuristics.
- **Affiliation of creators**: University of Akron; The Common Pool; OpenEd Solutions.
- **Domain**: education.
- **Tasks in fairness literature**: fair regression evaluation [522].
- **Data spec**: text.

- **Sample size**: $\sim$ 20K student essays.
- **Year**: 2012.
- **Sensitive features**: none.
- **Link**: https://www.kaggle.com/c/asap-aes/data/
- **Further info**: Shermis [718]

A.1.18 Bank Marketing

- **Description**: often simply called *Bank* dataset in the fairness literature, this resource was produced to support a study of success factors in telemarketing of long-term deposits within a Portuguese bank, with data collected over the period 2008–2010. Each data point represents a telemarketing phone call and includes client-specific features (e.g. job, education), features about the marketing phone call (e.g. day of the week and duration) and meaningful environmental features (e.g. euribor). The classification target is a binary variable indicating client subscription to a term deposit.
- **Affiliation of creators**: Instituto Universitário de Lisboa (ISCTE-IUL), ISTAR, Lisboa; University of Minho.
- **Domain**: marketing.
- **Tasks in fairness literature**: fair classification [34, 208, 698, 710, 856], fair clustering [1, 7, 31, 61, 146, 343, 378, 526], fair data summarization [233], fair classification under unawareness [441], fairness evaluation [390, 506], limited-label fairness evaluation [403], preference-based fair clustering [285].
- **Data spec**: tabular data.
- **Sample size**: $\sim$ 40K phone contacts.
- **Year**: 2012.
- **Sensitive features**: age.
- **Link**: https://archive.ics.uci.edu/ml/datasets/Bank+Marketing
- **Further info**: Moro et al. [581]

A.1.19 Barcelona Room Rental

- **Description**: this dataset summarizes the operations of a room rental platform in Barcelona over 30 months, from January 2017 through June 2019. It contains information about over $60,000$ users, divided into those seeking (seeker) and those listing (lister) a room. The data consists of lister-seeker pairs, such that a seeker is recommended for a room and lister. Recommendations are provided by a set of different recommender systems (recsys). For each pair, the data reports the rank in which each seeker was listed, the recsys providing the recommendation, and the post-recommendation interaction, if any, along with demographic information on both users. Textual indications of "gay-friendliness" in user profiles is treated as a sensitive feature (among others), as sexual orientation was previously found to be a discriminating factor in access to housing.
- **Affiliation of creators**: University Pompeu Fabra; Eurecat; Institute for Political Economy and Governance; ISI Foundation.
- **Domain**: information systems.
- **Tasks in fairness literature**: fair ranking evaluation [735].
- **Data spec**: lister-seeker pairs.
- **Sample size**: $\sim$ 4M pairs.
- **Year**: 2021.
- **Sensitive features**: gender, age, spoken language, "gay-friendliness".
- **Link**: not available
- **Further info**: Solans et al. [735]

A.1.20 Benchmarking Attribution Methods (BAM)

- **Description**: this dataset was developed to evaluate different explainability methods in computer vision. It was constructed by pasting object pixels from MS-COCO [505] into scene images from MiniPlaces [886]. Objects are rescaled to a variable proportion between one third and one half of the scene images onto which they are pasted. Both scene images and object images belong to ten different classes, for a total of 100 possible combinations. Scene images were chosen between the ones that do not contain the objects from the ten MS-COCO classes. This dataset enables users to freely control how each object is correlated with scenes, from which ground truth explanations can be formed. The creators also propose a few quantitative metrics to evaluate interpretability methods by either contrasting different inputs in the same dataset or contrasting two models with the same input.
- **Affiliation of creators**: Google.
- **Domain**: computer vision.
- **Tasks in fairness literature**: fair representation learning [191].
- **Data spec**: image.
- **Sample size**: $\sim$ 100K images over 10 object classes and 10 image classes.
- **Year**: 2020.
- **Sensitive features**: none.
- **Link**: https://github.com/google-research-datasets/bam
- **Further info**: Yang and Kim [845]

A.1.21 Berkeley Students

- **Description**: this dataset holds anonymized student records at UC Berkeley from Spring 2012 through Fall 2019. It consistst of enrollment information on a per-semester basis for tens of thousands of students. For each enrollment, student course scores are provided, along with student demographic information, including gender, race, entry status and parental income. The dataset supports evaluations of equity in educational outcome as well as grade predictions for academic support interventions. It is maintained by the University's Enterprise Data and Analytics unit.
- **Affiliation of creators**: University of California, Berkeley.
- **Domain**: education.
- **Tasks in fairness literature**: fair classification [404].
- **Data spec**: tabular data.
- **Sample size**: $\sim$ 2M enrollments across $\sim$ 80K students.
- **Year**: 2021.
- **Sensitive features**: gender, race.
- **Link**: not available
- **Further info**: Jiang and Pardos [404]

A.1.22 Bias in Bios

- **Description**: this dataset was developed as a large-scale study of gender bias in occupation classification. It consists of online biographies of professionals scraped from the Common Crawl. Biographies are detected in crawls when they match the regular expression "<name> is a(n) <title>", with <title> being one of twenty-eight common occupations. The gender of each person in the dataset is identified via the third person gendered pronoun, typically used in professional biographies. The envisioned task mirrors that of a job search automated system in a two-sided labor marketplace, i.e. automated

occupation classification. The dataset curators provide python code to recreate the dataset from old Common Crawls.
- **Affiliation of creators**: Carnegie Mellon University; University of Massachusetts Lowell; Microsoft; LinkedIn.
- **Domain**: linguistics, information systems.
- **Tasks in fairness literature**: fairness evaluation [195], fair classification [853].
- **Data spec**: text.
- **Sample size**: $\sim$ 400K biographies.
- **Year**: 2018.
- **Sensitive features**: gender.
- **Link**: https://github.com/Microsoft/biosbias
- **Further info**: De-Arteaga et al. [195]

A.1.23 Bias in Translation Templates

- **Description**: this resource was developed to study the problem of gender biases in machine translation. It consists of a set of short templates of the form `One thing about the man/woman, [he/she] is [a ##]`, where `[he/she]` can be a gender-neutral or gender-specific pronoun, and `[a ##]` refers to a profession or conveys sentiment. Templates are built so that the part before the comma acts as a gender-specific clue, and the part after the comma contains information about gender and sentiment/profession. Accurate translations should correctly match the grammatical gender before and after the comma, in every word where it is required by the target language. The curators identify a set of languages to which this template is easily applicable, namely German, Korean, Portuguese, and Tagalog, which are chosen for their different properties with respect to grammatical gender. Depending on which language pair is being considered for translation, the curators identify a set of criteria for the evaluation of translation quality, with special emphasis on the correctness of grammatical gender.
- **Affiliation of creators**: Seoul National University.
- **Domain**: linguistics.
- **Tasks in fairness literature**: bias evaluation of machine translation [150].
- **Data spec**: text.
- **Sample size**: $\sim$ 1K templates.
- **Year**: 2021.
- **Sensitive features**: gender.
- **Link**: https://github.com/nolongerprejudice/tgbi-x
- **Further info**: Cho et al. [150]

A.1.24 Bing US Queries

- **Description**: this dataset was created to investigate differential user satisfaction with the Bing search engine across different demographic groups. The authors selected log data of a random subset of Bing's desktop and laptop users from the English-speaking US market over a two week period. The data was preprocessed by cleaning spam and bot queries, and it was enriched with user demographics, namely age (bucketed) and gender (binary), which were self-reported by users during account registration and automatically validated by the dataset curators. Moreover, queries were labeled with topic information. Finally, four different signals were extracted from search logs, namely graded utility, reformulation rate, page click count, and successful click count.
- **Affiliation of creators**: Microsoft.

- **Domain**: information systems.
- **Tasks in fairness literature**: fair ranking evaluation [554].
- **Data spec**: query-result pairs.
- **Sample size**: $\sim$ 30M (non-unique) queries issued by $\sim$ 4M distinct users.
- **Year**: 2017.
- **Sensitive features**: age, gender.
- **Link**: not available
- **Further info**: Mehrotra et al. [554]

A.1.25 BOLD

- **Description**: this resource is a benchmark to measure biases of language models with respect to sensitive demographic attributes. The creators identified six attributes (e.g. race, profession) and values of said attribute (e.g. African American, flight nurse) for which they gather prompts from English Language Wikipedia, either from pages about the group (e.g. "A flight nurse is a registered") or people representing it (e.g. "Over the years, Isaac Hayes was able"). Prompts are fed to different language models, whose outputs are automatically labelled for sentiment, regard, toxicity, emotion and gender polarity. These labels are also validated by human annotators hired on Amazon Mechanical Turk.
- **Affiliation of creators**: Amazon; University of California, Santa Barbara.
- **Domain**: linguistics.
- **Tasks in fairness literature**: bias evaluation in language models [206].
- **Data spec**: text.
- **Sample size**: $\sim$ 20K prompts.
- **Year**: 2021.
- **Sensitive features**: gender, race, religion, profession, political leaning.
- **Link**: https://github.com/amazon-research/bold
- **Further info**: Dhamala et al. [206]

A.1.26 BookCorpus

- **Description**: this dataset was developed for the problem of learning general representations of text useful for different downstream tasks. It consist of text from 11,038 books from the web by unpublished authors available on https://www.smashwords.com/ in 2015. The BookCorpus contains thousands of duplicate books (only 7,185 are unique) and many contain copyright restrictions. The GPT [654] and BERT [204] language models were trained on this dataset.
- **Affiliation of creators**: University of Toronto; Massachusetts Institute of Technology.
- **Domain**: linguistics.
- **Tasks in fairness literature**: data bias evaluation [754].
- **Data spec**: text.
- **Sample size**: $\sim$ 1B words in $\sim$74M sentences from $\sim$11K books.
- **Year**: unknown.
- **Sensitive features**: textual references to people and their demographics.
- **Link**: not available
- **Further info**: Bandy and Vincent [42], Zhu et al. [888]

A.1.27 BUPT Faces

- **Description**: this resource consists of two datasets, developed as a large scale collection, suitable for training face verification algorithms operating on diverse populations. The underlying data collection procedure mirrors the one from RFW (subsection A.1.153), including sourcing from MS-Celeb-1M and automated annotation of so-called *race* into one of four categories: Caucasian, Indian, Asian and African. For categories where not enough images were readily available, the authors resort to the FreeBase celebrity list, downloading images of people from Google and cleaning them "both automatically and manually". The remaining images were obtained from MS-Celeb-1M (subsection A.1.124), on which the BUPT Faces datasets are heavily based.
- **Affiliation of creators**: Beijing University of Posts and Telecommunications.
- **Domain**: computer vision.
- **Tasks in fairness literature**: fair reinforcement learning [812], fair classification [839], fair representation learning [315].
- **Data spec**: image.
- **Sample size**: $\sim$ 2M images of $\sim$ 40K celebrities (BUPT-Globalface); $\sim$ 1M images of $\sim$ 30K celebrities (BUPT-Balancedface).
- **Year**: 2019.
- **Sensitive features**: race.
- **Link**: http://www.whdeng.cn/RFW/Trainingdataste.html
- **Further info**: Wang and Deng [812]

A.1.28 Burst

- **Description**: Burst is a free provider of stock photography powered by Shopify. This dataset features a subset of Burst images used as a resource to test algorithms for fair image retrieval and ranking, aimed at providing, in response to a query, a collection of photos that is balanced across demographics. Images come with human-curated tags annotated internally by the Burst team.
- **Affiliation of creators**: Shopify.
- **Domain**: information systems.
- **Tasks in fairness literature**: fair ranking [428].
- **Data spec**: image.
- **Sample size**: $\sim$ 3K images.
- **Year**: present.
- **Sensitive features**: gender.
- **Link**: not available
- **Further info**: Karako and Manggala [428]; https://burst.shopify.com/

A.1.29 Business Entity Resolution

- **Description**: A proprietary Google dataset, where the task is to predict whether a pair of business descriptions describe the same real business.
- **Affiliation of creators**: Google.
- **Domain**: linguistics.
- **Tasks in fairness literature**: fair entity resolution [175].
- **Data spec**: text.
- **Sample size**: $\sim$15K samples.
- **Year**: 2019.

- **Sensitive features**: geography, business size.
- **Link**: not available
- **Further info**: Cotter et al. [175]

A.1.30 Campus Recruitment

- **Description**: this dataset was published to Kaggle in 2020 by Ben Roshan, who was then enrolled in an MBA in Business Analytics at Jain University Bangalore. The provenance of this dataset is not clear. It was provided by a Jain University professor as a class resource to study and experiment with data analysis. It encodes information about students at an Indian institution, including their degree, their performance in school and placement information at the end of school, including salary.
- **Affiliation of creators**: Jain University Bangalore.
- **Domain**: education.
- **Tasks in fairness literature**: fair data generation [507].
- **Data spec**: tabular data.
- **Sample size**: $\sim$ 200 students.
- **Year**: 2020.
- **Sensitive features**: gender.
- **Link**: https://www.kaggle.com/datasets/benroshan/factors-affecting-campus-placement
- **Further info**:

A.1.31 Cars3D

- **Description**: this dataset consists of CAD-generated models of 199 cars rendered from from 24 rotation angles. Originally devised for visual analogy making, it is also used for more general research on learning disentangled representation.
- **Affiliation of creators**: University of Michigan.
- **Domain**: computer vision.
- **Tasks in fairness literature**: fair representation learning [513].
- **Data spec**: image.
- **Sample size**: $\sim$ 5K images.
- **Year**: 2020.
- **Sensitive features**: none.
- **Link**: https://github.com/google-research/disentanglement_lib/tree/master/disentanglement_lib/data/ground_truth
- **Further info**: Reed et al. [671]

A.1.32 CelebA

- **Description**: CelebFaces Attributes Dataset (CelebA) features images of celebrities from the CelebFaces dataset, augmented with annotations of landmark location and binary attributes. The attributes, ranging from highly subjective features (e.g. attractive, big nose) and potentially offensive (e.g. double chin) to more objective ones (e.g. black hair) were annotated by a "professional labeling company".
- **Affiliation of creators**: Chinese University of Hong Kong.
- **Domain**: computer vision.

- **Tasks in fairness literature**: fair classification [156, 183, 415, 447, 514, 698], fair anomaly detection [864], bias discovery [14] fair anomaly detection [864], fairness evaluation of private classification [145], fairness evaluation of selective classification [411], fairness evaluation [706, 816], fair representation learning [653], fair data summarization [147], fair data generation [151, 665].
- **Data spec**: image.
- **Sample size**: $\sim$ 200K face images of over $\sim$ 10K unique individuals.
- **Year**: 2015.
- **Sensitive features**: gender, age, skin tone.
- **Link**: http://mmlab.ie.cuhk.edu.hk/projects/CelebA.html
- **Further info**: Liu et al. [512]

A.1.33 CheXpert

- **Description**: this dataset consists of chest X-ray images from patients that have been treated at the Stanford Hospital between October 2002 and July 2017. Each radiograph, either frontal or lateral, is annotated for the presence of 14 observations related to medical conditions. Most annotations were automatically extracted from free text radiology reports and validated against a set of 1,000 held-out reports, manually reviewed by a radiologist. For a subset of the X-ray images, high-quality labels are provided by a group of 3 radiologists. The task associated with this dataset is the automated diagnosis of medical conditions from radiographs.
- **Affiliation of creators**: Stanford University.
- **Domain**: radiology.
- **Tasks in fairness literature**: fairness evaluation of selective classification [411], fairness evaluation of private classification [145].
- **Data spec**: image.
- **Sample size**: $\sim$ 200K chest radiographs from 60K patients.
- **Year**: 2019.
- **Sensitive features**: sex, age (of patient).
- **Link**: https://stanfordmlgroup.github.io/competitions/chexpert/
- **Further info**: Garbin et al. [287], Irvin et al. [389]

A.1.34 Chicago Ridesharing

- **Description**: this resource describes all trips reported by ridesharing companies to the City of Chicago, starting November 2018. It is the result of an ongoing transparency effort, following the introduction of a city-wide ordinance requiring the disclosure of trips and and fares on part of transportation network providers. For each trip, this dataset reports geographical information (pickup and dropoff), duration and cost. To avoid individual re-identification, the granularity of times and locations is reduced to the nearest 15-minutes interval and census tract. Moreover, for rare combinations of census tract an interval, location data is provided at coarser granularity (community area).
- **Affiliation of creators**: City of Chicago.
- **Domain**: transportation.
- **Tasks in fairness literature**: fair pricing evaluation [618].
- **Data spec**: tabular data.
- **Sample size**: $\sim$ 200M trips.
- **Year**: present.
- **Sensitive features**: geography.

- **Link**: https://data.cityofchicago.org/Transportation/Transportation-Network-Providers-Trips/m6dm-c72p
- **Further info**: http://dev.cityofchicago.org/open%20data/data%20portal/2020/04/28/tnp-trips-2019-addition html; http://dev.cityofchicago.org/open%20data/data%20portal/2019/04/12/tnp-taxi-privacy.html

A.1.35 CIFAR

- **Description**: CIFAR-10 and CIFAR-100 are a labelled subset of the 80 million tiny images database. CIFAR consists of 32x32 colour images that students were paid to annotate. The project, aimed at advancing the effectiveness of supervised learning techniques in computer vision, was funded by the the Canadian Institute for Advanced Research, after which the dataset is named.
- **Affiliation of creators**: University of Toronto.
- **Domain**: computer vision.
- **Tasks in fairness literature**: fair classification [415, 816], fair incremental learning [876], robust fairness evaluation [589].
- **Data spec**: image.
- **Sample size**: $\sim 6K$ images x 10 classes (CIFAR-10) or 600 images x 100 classes (CIFAR-100).
- **Year**: 2009.
- **Sensitive features**: none.
- **Link**: https://www.cs.toronto.edu/~kriz/cifar.html
- **Further info**: Krizhevsky [467]
- **Variants**: CIFAR-10S [816] is a modified version specifically aimed at studying biases in image classification across an artificial sensitive attribute (color/grayscale).

A.1.36 CiteSeer Papers

- **Description**: this dataset was created to study the problem of link-based classification of connected entities. The creators extracted a network of papers from CiteSeer, belonging to one of six categories: Agents, Artificial Intelligence, Database, Human Computer Interaction, Machine Learning and Information Retrieval. Each article is associated with a bag-of-word representation, and the associated task is classification into one of six topics.
- **Affiliation of creators**: University of Maryland.
- **Domain**: library and information sciences.
- **Tasks in fairness literature**: fair graph mining [498].
- **Data spec**: paper-paper pairs.
- **Sample size**: $\sim$ 3K articles connected by $\sim$ 5K citations.
- **Year**: 2016.
- **Sensitive features**: none.
- **Link**: http://networkrepository.com/citeseer.php
- **Further info**: Lu and Getoor [517]

A.1.37 Civil Comments

- **Description**: this dataset derives from an archive of the Civil Comments platform, a browser plugin for independent news sites, whose users peer-reviewed each other's comments with civility ratings. When the plugin shut down, they decided to make comments and metadata available, including the

crowd-sourced toxicity ratings. A subset of this dataset was later annotated with a variety of sensitive attributes, capturing whether members of a certain group are mentioned in comments. This dataset powers the Jigsaw Unintended Bias in Toxicity Classification challenge.
- **Affiliation of creators**: Jigsaw; Civil Comments.
- **Domain**: social media.
- **Tasks in fairness literature**: fair toxicity classification [3, 156, 853], fairness evaluation of selective classification [411], fair robust toxicity classification [3], fairness evaluation of toxicity classification [384], fairness evaluation [29].
- **Data spec**: text.
- **Sample size**: $\sim$ 2M comments.
- **Year**: 2019.
- **Sensitive features**: race/ethnicity, gender, sexual orientation, religion, disability.
- **Link**: https://www.kaggle.com/c/jigsaw-unintended-bias-in-toxicity-classification
- **Further info**: Borkan et al. [88]

A.1.38 Climate Assembly UK

- **Description**: this resource was curated to study the problem of subset selection for *sortition*, a political system where decisions are taken by a subset of the whole voting population selected at random. The data describes participants to Climate Assembly UK, a panel organized by the Sortition Foundation in 2020. With the goal of understanding public opinion on how the UK can meet greenhouse gas emission targets. The panel consisted of 110 UK residents selected from a pool of 1,715 who responded to an invitation from the Sortition Foundation reaching $\sim 60K$ citizens. Features for each subject in the pool describe their demographics and climate concern level.
- **Affiliation of creators**: Carnegie Mellon University; Harvard University; Sortition Foundation.
- **Domain**: political science.
- **Tasks in fairness literature**: fair subset selection [272].
- **Data spec**: tabular data.
- **Sample size**: $\sim$ 2K pool participants.
- **Year**: 2020.
- **Sensitive features**: gender, age, education, urban/rural, geography, ethnicity.
- **Link**: not available
- **Further info**: Flanigan et al. [272]; https://www.climateassembly.uk/

A.1.39 Columbia University Speed Dating

- **Description**: this dataset is a result of a speed dating experiment aimed at understanding preferences in mate selection in men and women. Subjects were recruited from students at Columbia University. Fourteen rounds were conducted with different proportions of male and female subjects, over the period 2002–2004, with participants meeting each potential mate for four minutes and rating them thereafter on six attributes. They also provide an overall evaluation of each potential mate and a binary decision indicating interest in meeting again. Before an event, each participant filled in a survey disclosing their preferences, expectations, and demographics. The inference task associated with this dataset is optimal recommendation in symmetrical two-sided markets.
- **Affiliation of creators**: Columbia University; Harvard University; Stanford University.
- **Domain**: sociology.
- **Tasks in fairness literature**: fair matching [884], preference-based fair ranking [622].

- **Data spec**: person-person pairs.
- **Sample size**: $\sim$ 10K dating records involving $\sim$ 400 people.
- **Year**: 2016.
- **Sensitive features**: gender, age, race, geography.
- **Link**: https://data.world/annavmontoya/speed-dating-experiment
- **Further info**: Fisman et al. [270]

A.1.40 Communities and Crime

- **Description**: this dataset was curated to develop a software tool supporting the work of US police departments. It was especially aimed at identifying similar precincts to exchange best practices and share experiences among departments. The creators were supported by the police departments of Camden (NJ) and Philadelphia (PA). The factors included in the dataset were the ones deemed most important to define similarity of communities from the perspective of law enforcement; they were chosen with the help of law enforcement officials from partner institutions and academics of criminal justice, geography and public policy. The dataset includes socio-economic factors (aggregate data on age, income, immigration, and racial composition) obtained from the 1990 US census, along with information about policing (e.g. number of police cars available) based on the 1990 Law Enforcement Management and Administrative Statistics survey, and crime data derived from the 1995 FBI Uniform Crime Reports. In its released version on UCI, the task associated with the dataset is predicting the total number of violent crimes per 100K population in each community. The most referenced version of this dataset was preprocessed with a normalization step; after receiving multiple requests, the creators also published an unnormalized version.
- **Affiliation of creators**: La Salle University; Rutgers University.
- **Domain**: law.
- **Tasks in fairness literature**: fair classification [174, 175, 183, 357, 514, 712, 841], fair regression evaluation [358], fair few-shot learning [728, 730], rich-subgroup fairness evaluation [435], rich-subgroup fair classification [434], fair regression [5, 64, 158, 159, 208, 460, 535, 608, 680], fair representation learning [686], robust fair classification [529], fair private classification [399], fairness evaluation of transfer learning [481], preference-based fair clustering [285].
- **Data spec**: tabular data.
- **Sample size**: $\sim 2K$ communities.
- **Year**: 2009.
- **Sensitive features**: race, geography.
- **Link**: https://archive.ics.uci.edu/ml/datasets/communities+and+crime and http://archive.ics.uci.edu/ml/datasets/communities+and+crime+unnormalized
- **Further info**: Redmond and Baveja [669]

A.1.41 COMPAS

- **Description**: this dataset was created for an external audit of racial biases in the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) risk assessment tool developed by Northpointe (now Equivant), which estimates the likelihood of a defendant becoming a recidivist. Instances represent defendants scored by COMPAS in Broward County, Florida, between 2013–2014, reporting their demographics, criminal record, custody and COMPAS scores. Defendants' public criminal records were obtained from the Broward County Clerk's Office website matching them based on date of birth,

first and last names. The dataset was augmented with jail records and COMPAS scores provided by the Broward County Sheriff's Office. Finally, public incarceration records were downloaded from the Florida Department of Corrections website. Instances are associated with two target variables (is_recid and is_violent_recid), indicating whether defendants were booked in jail for a criminal offense (potentially violent) that occurred after their COMPAS screening but within two years. See Appendix A.3 for extensive documentation.

- **Affiliation of creators**: ProPublica.
- **Domain**: law.
- **Tasks in fairness literature**: fair classification [12, 64, 111, 112, 124, 126, 149, 175, 197, 208, 210, 216, 308, 356, 357, 514, 519, 527, 535, 608, 611, 633, 652, 678, 679, 687, 698, 715, 791, 803, 809, 838, 854], fairness evaluation [4, 114, 130, 153, 165, 282, 327, 390, 396, 431, 509, 549, 597, 642, 740, 759, 821, 866], fair risk assessment [171, 568, 587], fair *task assignment* [307], fair classification under unawareness [157, 441, 478, 480], data bias evaluation [62], fair representation learning [91, 686, 878], robust fair classification [76, 529, 672], dynamical fairness evaluation [869], fair reinforcement learning [561], fair ranking evaluation [421, 843], fair multi-stage classification [525], dynamical fair classification [788], preference-based fair classification [783, 855], fair regression [460], fair multi-stage classification [306], limited-label fair classification [152, 157, 810], robust fairness evaluation [729, 730], rich subgroup fairness evaluation [154, 874].
- **Data spec**: tabular data.
- **Sample size**: $\sim$ 12K defendants.
- **Year**: 2016.
- **Sensitive features**: sex, age, race.
- **Link**: https://github.com/propublica/compas-analysis
- **Further info**: Angwin et al. [21], Larson et al. [483]

A.1.42 Cora Papers

- **Description**: this resource was produced within the wider development effort for *Cora*, an Internet portal for computer science research papers available in the early 2000s. The portal supported keyword search, topical categorization of articles, and citation mapping. This dataset consists of articles and citation links between them. It contains bag-of-word representations for the text of each article, and the associated task is classification into one of seven topics.
- **Affiliation of creators**: Just Research Carnegie Mellon University; Massachusetts Institute of Technology; Univeristy of Maryland; Lawrence Livermore National Laboratory.
- **Domain**: library and information sciences.
- **Tasks in fairness literature**: .
- **Data spec**: article-article pairs.
- **Sample size**: $\sim$ 3K articles connected by $\sim$ 5K citations.
- **Year**: 2019.
- **Sensitive features**: none.
- **Link**: https://relational.fit.cvut.cz/dataset/CORA
- **Further info**: McCallum et al. [541], Sen et al. [708]

A.1.43 Costarica Household Survey

- **Description**: this data comes from the national household survey of Costa Rica, performed by the national institute of statistics and census (Instituto Nacional de Estadística y Censos). The survey is

aimed at measuring the socio-economical situation in the country and informing public policy. The data collection procedure is specially designed to allow for precise conclusions with respect to six different regions of the country and about differences in urban vs rural areas; stratification along these variables is deemed suitable. The 2018 survey contains a special section on the crimes suffered by respondents.
- **Affiliation of creators**: Instituto Nacional de Estadística y Censos.
- **Domain**: economics.
- **Tasks in fairness literature**: fair classification [601].
- **Data spec**: tabular data.
- **Sample size**: $\sim$ 13K households.
- **Year**: 2018.
- **Sensitive features**: sex, age, birthplace, disability, geography, family size.
- **Link**: https://www.inec.cr/encuestas/encuesta-nacional-de-hogares
- **Further info**: https://www.inec.cr/sites/default/files/documetos-biblioteca-virtual/enaho-2018.pdf

A.1.44 Credit Card Default

- **Description**: this dataset was built to investigate automated mechanisms for credit card default prediction following a wave of defaults in Taiwan connected to patters of card over-issuing and over-usage. The dataset contains payment history of customers of an important Taiwanese bank, from April to October 2005. Demographics, marital status, and education of customers are also provided, along with the amount of credit and a binary variable encoding default on payment, which is the target variable of the associated task.
- **Affiliation of creators**: Chung-Hua University; Thompson Rivers University.
- **Domain**: finance.
- **Tasks in fairness literature**: fair classification [64, 149], fair clustering [61, 298, 343, 343], fair clustering under unawareness [239], fair classification under unawareness [814], fair data summarization [693, 756], fairness evaluation [506], fair anomaly detection [716].
- **Data spec**: tabular data.
- **Sample size**: $\sim$ 30K credit card holders.
- **Year**: 2016.
- **Sensitive features**: gender, age.
- **Link**: https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients
- **Further info**: Yeh and hui Lien [848]

A.1.45 Credit Elasticities

- **Description**: this dataset stems from a randomized trial conducted by a consumer lender in South Africa to study loan price elasticity. Prior customers were contacted by mail with limited-time loan offers at variable and randomized interest rates. The aim of the study was understanding the relationship between interest rate and customer acceptance rates, along with the benefits for the lender. Customers who accepted and received formal approval, filled in a short survey with factors of interest for the study, including demographics, education, and prior borrowing history.
- **Affiliation of creators**: Yale University; Dartmouth College.
- **Domain**: finance.
- **Tasks in fairness literature**: fair pricing evaluation [422].
- **Data spec**: tabular data.

- **Sample size**: $\sim$ 50K clients.
- **Year**: 2008.
- **Sensitive features**: gender, age, geography.
- **Link**: http://doi.org/10.3886/E113240V1
- **Further info**: Karlan and Zinman [430]

A.1.46 Crowd Judgement

- **Description**: this dataset was assembled to compare the performance of the COMPAS recidivism risk prediction system against that of non-expert human assessors [217]. A subset of 1,000 defendants were selected from the COMPAS dataset. Crowd-sourced assessors were recruited through Amazon Mechanical Turk. They were presented with a summary of each defendant, including demographics and previous criminal history, and asked to predict whether they would recidivate within 2 years of their most recent crime. These judgements, assembled via plain majority voting, ended up exhibiting accuracy and fairness levels comparable to that displayed by the COMPAS system. While this dataset was assembled for an experiment, it was later used to study the problem of fairness in crowdsourced judgements.
- **Affiliation of creators**: Dartmouth College.
- **Domain**: law.
- **Tasks in fairness literature**: fair *truth discovery* [502], fair *task assignment* [307, 502]
- **Data spec**: judge-defendant pair.
- **Sample size**: $\sim$ 1K defendants from COMPAS and $\sim$ 400 crowd-sourced labellers. Each defendant is judged by 20 different labellers.
- **Year**: 2018.
- **Sensitive features**: sex, age and race of defendants and crowd-sourced judges.
- **Link**: https://farid.berkeley.edu/downloads/publications/scienceadvances17/
- **Further info**: [217]
- **Variants**: a similar dataset was collected by Wang et al. [808].

A.1.47 Curatr British Library Digital Corpus

- **Description**: this dataset is a subset of English language digital texts from the British Library focused on volumes of 19th-century fiction, obtained through the Curatr platform. It was selected for the well-researched presence of stereotypical and binary concepts of gender in this literary production. The goal of the creators was studying gender biases in large text corpora and their relationship with biases in word embeddings trained on those corpora.
- **Affiliation of creators**: University College Dublin.
- **Domain**: literature.
- **Tasks in fairness literature**: data bias evaluation [487].
- **Data spec**: text.
- **Sample size**: $\sim$ 20K books.
- **Year**: 2020.
- **Sensitive features**: textual references to people and their demographics.
- **Link**: http://curatr.ucd.ie/
- **Further info**: Leavy et al. [486]

A.1.48 CVs from Singapore

- **Description**: this dataset was developed to test demographic biases in resume filtering. In particular, the authors studied nationality bias in automated resume filtering in Singapore, across the three major ethnic groups of the city state: Chinese, Malaysian and Indian. The dataset consists of 135 resumes (45 per ethnic group) used for application to finance jobs in Singapore, collected by Jai Janyani. The dataset only includes resumes for which the origin of the candidates can be reliably inferred to be either Chinese, Malaysian, or Indian from education and initial employment. The dataset also comprises 9 finance job postings from China, Malaysia, and India (3 per country). All job-resume pairs are rated for relevance/suitability by three annotators.
- **Affiliation of creators**: University of Maryland.
- **Domain**: information systems, management information systems.
- **Tasks in fairness literature**: fair ranking [200].
- **Data spec**: text.
- **Sample size**: $\sim$ 100 resumes.
- **Year**: 2020.
- **Sensitive features**: ethnic group.
- **Link**: not available
- **Further info**: Deshpande et al. [200]

A.1.49 Dallas Police Incidents

- **Description**: this dataset is due to the Dallas OpenData initiative[4] and "reflects crimes as reported to the Dallas Police Department" beginning June 1, 2014. Each incident comes with rich spatio-temporal data, information about the victim, the officers involved and the type of crime. A subset of the dataset is available on Kaggle[5].
- **Affiliation of creators**: Dallas Police Department.
- **Domain**: law.
- **Tasks in fairness literature**: fair spatio-temporal process learning [711].
- **Data spec**: tabular.
- **Sample size**: $\sim$ 800K incidents.
- **Year**: present.
- **Sensitive features**: age, race, and gender (of victim), geography.
- **Link**: https://www.dallasopendata.com/Public-Safety/Police-Incidents/qv6i-rri7
- **Further info**:

A.1.50 Demographics on Twitter

- **Description**: this dataset was developed to test demographic classifiers on Twitter data. In particular, the tasks associated with this resource are the automatic inference of gender, age, location and political orientation of users. The true values for these attributes, which act as a ground truth for learning algorithms, were inferred from tweets and user bios, such as the ones containing the regexp "I'm a <gendered noun>", with gendered nouns including mother, woman, father, man.
- **Affiliation of creators**: Massachusetts Institute of Technology.
- **Domain**: social media.

[4]https://www.dallasopendata.com/
[5]https://www.kaggle.com/carrie1/dallaspolicereportedincidents

- **Tasks in fairness literature**: fairness evaluation of sentiment analysis [717].
- **Data spec**: mixture.
- **Sample size**: $\sim$ 80K profiles.
- **Year**: 2017.
- **Sensitive features**: gender, age, political orientation, geography.
- **Link**: not available
- **Further info**: Vijayaraghavan et al. [796]

A.1.51 Diabetes 130-US Hospitals

- **Description**: this dataset contains 10 years of care data from 130 US hospitals extracted from Health Facts, a clinical database associated with a multi-institution data collection program. The dataset was extracted to study the association between the measurement of HbA1c (glycated hemoglobin) in human bloodstream and early hospital readmission, and was donated to UCI in 2014. The dataset includes patient demographics, in-hospital procedures, and diagnoses, along with information about subsequent readmissions.
- **Affiliation of creators**: Virginia Commonwealth University; University of Cordoba; Polish Academy of Sciences.
- **Domain**: endocrinology.
- **Tasks in fairness literature**: fair clustering [31, 61, 61, 146, 378, 526].
- **Data spec**: tabular data.
- **Sample size**: $\sim$ 100K patients.
- **Year**: 2014.
- **Sensitive features**: age, race, gender.
- **Link**: https://archive.ics.uci.edu/ml/datasets/diabetes+130-us+hospitals+for+years+1999-2008
- **Further info**: Strack et al. [745]

A.1.52 Diversity in Faces (DiF)

- **Description**: this large dataset was created to favour the development and evaluation of robust face analysis algorithms across diverse demographics and domain-specific features, such as craniofacial distances and facial contrast). One million images of people's faces from Flickr were labelled, mostly automatically, according to 10 different coding schemes, comprising, e.g., cranio-facial measurements, pose, and demographics. Age and gender were inferred both automatically and by human workers. Statistics about the diversity of this dataset along these coded measures are available in the accompanying report.
- **Affiliation of creators**: IBM.
- **Domain**: computer vision.
- **Tasks in fairness literature**: fair representation learning [653], fairness evaluation of private classification [33].
- **Data spec**: image.
- **Sample size**: $\sim$ 1M images.
- **Year**: 2019.
- **Sensitive features**: skin color, age, and gender.
- **Link**: https://www.ibm.com/blogs/research/2019/01/diversity-in-faces/
- **Further info**: Merler et al. [556]

A.1.53 Drug Consumption

- **Description**: this dataset was collected by Elaine Fehrman between March 2011 and March 2012 after receiving approval from relevant ethics boards from the University of Leicester. The goal of this dataset is to seek patterns connecting an individual's risk of drug consumption with demographics and psychometric measurements of the Big Five personality traits (NEO-FFI-R), impulsivity (BIS-11), and sensation seeking (ImpSS). The study employed an online survey tool from Survey Gizmo to recruit participants world-wide; over 93% of the final usable sample reported living in an English-speaking country. Target variables summarize the consumption of 18 psychoactive substances on an ordinal scale ranging from never using the drug to using it over a decade ago, or in the last decade, year, month, week, or day. The 18 substances considered in the study are classified as central nervous system depressants, stimulants, or hallucinogens and comprise the following: alcohol, amphetamines, amyl nitrite, benzodiazepines, cannabis, chocolate, cocaine, caffeine, crack, ecstasy, heroin, ketamine, legal highs, LSD, methadone, magic mushrooms, nicotine, and Volatile Substance Abuse (VSA), along with one fictitious drug (Semeron) introduced to identify over-claimers. A version of the dataset donated to the UCI Machine Learning Repository is associated with 18 prediction tasks, i.e. one per substance.
- **Affiliation of creators**: Rampton Hospital; Nottinghamshire Healthcare NHS Foundation Trust; University of Leicester; University of Nottingham; University of Salahaddin.
- **Domain**: applied psychology.
- **Tasks in fairness literature**: fair classification [216, 535], evaluation of data bias [62], limited-label fair classification [157], robust fair classification [672].
- **Data spec**: tabular data.
- **Sample size**: $\sim$ 2K respondents.
- **Year**: 2016.
- **Sensitive features**: age, gender, ethnicity, geography.
- **Link**: https://archive.ics.uci.edu/ml/datasets/Drug+consumption+%28quantified%29
- **Further info**: Fehrman et al. [259, 260]

A.1.54 DrugNet

- **Description**: this dataset was collected to study drug consumption patterns in connection with social ties and behaviour of drug users. This work puts particular emphasis on situations at risk of disease transmission and to assess the opportunity for prevention via recruitment of peer educators to demonstrate, disseminate and support HIV prevention practices among their connections. Participants were recruited in Hartford neighbourhoods of high drug-use activity, mostly via street outreach and recruitment by early participants. Eligibility criteria included being at least 18 years old, using an illicit drug, and signing an informed consent form. Each participant provided data about their drug use, most common sites of usage, HIV risk practices associated with drug use and sexual behavior, and social ties deemed important by the respondent and their demographics.
- **Affiliation of creators**: Institute for Community Research of Hartford; Hispanic Health Council, Hartford; Boston College.
- **Domain**: social work, social networks.
- **Tasks in fairness literature**: fair graph clustering [454].
- **Data spec**: person-person pairs.
- **Sample size**: $\sim$ 300 people.
- **Year**: 2016.
- **Sensitive features**: ethnicity, sex, age.

- **Link**: https://sites.google.com/site/ucinetsoftware/datasets/covert-networks/drugnet
- **Further info**: Weeks et al. [819]

A.1.55 dSprites

- **Description**: this dataset was assembled by researchers affiliated with Google DeepMind as an artificial benchmark for unsupervised methods aimed at learning disentangled data representations. Each image in the dataset consists of a black-and-white sprite with variable shape, scale, orientation and position. Together these are the *generative factors* underlying each image. Ideally, systems trained on this data should learn disentangled representations, such that latent image representations are clearly associated with changes in a single generative factor.
- **Affiliation of creators**: Google.
- **Domain**: computer vision.
- **Tasks in fairness literature**: fair representation learning [183, 513].
- **Data spec**: image.
- **Sample size**: $\sim$ 700K images.
- **Year**: 2017.
- **Sensitive features**: none.
- **Link**: https://github.com/deepmind/dsprites-dataset
- **Further info**: Higgins et al. [363]

A.1.56 Dutch Census

- **Description**: this dataset was derived from the 2001 census carried out by the Dutch Central Bureau for Statistics to gather data about family composition, economic activities, levels of education, and occupation of Dutch citizens and foreigners from various countries of origin. A version of the dataset commonly employed in the fairness research literature has been preprocessed and made available online. The associated task is the classification of individuals into high-income and low-income professions.
- **Affiliation of creators**: Bournemouth University; TU Eindhoven.
- **Domain**: demography.
- **Tasks in fairness literature**: fair classification [4, 514, 838, 867], fairness evaluation [114].
- **Data spec**: tabular data.
- **Sample size**: $\sim$ 60K respondents.
- **Year**: 2001.
- **Sensitive features**: sex, age, citizenship.
- **Link**: https://sites.google.com/site/conditionaldiscrimination/
- **Further info**: Žliobaité et al. [891]; https://microdata.worldbank.org/index.php/catalog/2102/data-dictionary/F2?file_name=NLD2001-P-H; https://www.cbs.nl/nl-nl/publicatie/2004/31/the-dutch-virtual-census-of-2001

A.1.57 EdGap

- **Description**: this dataset focuses on education performance in different US counties, with a focus on inequality of opportunity and its connection to socioeconomic factors. Along with average SAT and ACT test scores by county, this dataset reports socioeconomic data from the American Community Survey by the Bureau of Census, including household income, unemployment, adult educational attainment, and family structure. Importantly, some states require all students to take ACT or SAT tests, while others do

not. As a result, average test scores are inherently higher in states that do not require all students to test, and they are not directly comparable to average scores in states where testing is mandatory.
- **Affiliation of creators**: Memphis Teacher Residency.
- **Domain**: education.
- **Tasks in fairness literature**: fair risk assessment [355].
- **Data spec**: tabular data.
- **Sample size**: $\sim$ 2K counties.
- **Year**: 2019.
- **Sensitive features**: geography.
- **Link**: https://www.edgap.org/
- **Further info**:

A.1.58 Epileptic Seizures

- **Description**: this dataset was curated to study electroencephalographic (EEG) time series in relation to epilepsy. The dataset consists of EEG recordings from healthy volunteers with eyes closed and eyes open, and from epilepsy patients during seizure-free intervals and during epileptic seizures. Volunteers and patients are recorded for 23.6-sec. A version of this dataset, used in fairness research, was donated to UCI Machine Learning Repository by researchers affiliated with Rochester Institute of Technology in 2017, with a classification task based on the patients' condition and state at the time of recording. The data was later removed from UCI at the original curators' request.
- **Affiliation of creators**: University of Bonn.
- **Domain**: neurology.
- **Tasks in fairness literature**: robust fairness evaluation [77].
- **Data spec**: time series.
- **Sample size**: $\sim$ 500 individuals, each summarized by $\sim$ 4K-points time series.
- **Year**: 2017.
- **Sensitive features**: none.
- **Link**: https://archive.ics.uci.edu/ml/datasets/Epileptic+Seizure+Recognition; http://epileptologie-bonn.de/cms/upload/workgroup/lehnertz/eegdata.html
- **Further info**: Andrzejak et al. [19]

A.1.59 Equitable School Access in Chicago

- **Description**: this resource was assembled from disparate sources to evaluate school access in Chicago for different race groups. A transportation network was inferred from data on public bus lines available on the Chicago Transit Authority website. Data on school location and quality evaluation was obtained from the Chicago Public School data portal. Finally, demographic information on race representation in different tracts was retrieved from the 2010 US census.
- **Affiliation of creators**: Salesforce.
- **Domain**: transportation.
- **Tasks in fairness literature**: fair graph augmentation [664].
- **Data spec**: location-location pairs.
- **Sample size**: $\sim 2K$ nodes (locations), connected by $\sim$ 8K edges (bus lines).
- **Year**: 2020.
- **Sensitive features**: race.
- **Link**: https://github.com/salesforce/GAEA
- **Further info**: Ramachandran et al. [664]

A.1.60 Equity Evaluation Corpus (EEC)

- **Description**: this dataset was compiled to audit sentiment analysis systems for gender and race bias. It is based on 11 short sentence templates; 7 templates include emotion words, while the remaining 4 do not. Moreover, each sentence includes one gender- or race-associated word, such as names predominantly associated with African American or European American people. Gender-related words consist of names, nouns, and pronouns.
- **Affiliation of creators**: National Research Council Canada.
- **Domain**: linguistics.
- **Tasks in fairness literature**: fair sentiment analysis evaluation [504].
- **Data spec**: text.
- **Sample size**: $\sim$ 9K sentences.
- **Year**: 2018.
- **Sensitive features**: race, gender.
- **Link**: https://saifmohammad.com/WebPages/Biases-SA.html
- **Further info**: Kiritchenko and Mohammad [448]

A.1.61 Facebook Ego-networks

- **Description**: this dataset was collected to study the problem of identifying users' social circles, i.e. categorizing links between nodes in a social network. The data represents ten ego-networks whose central user was asked to fill in a survey and manually identify the circles to which their friends belonged. Features from each profile, including education, work and location are anonymized.
- **Affiliation of creators**: Stanford University.
- **Domain**: social networks.
- **Tasks in fairness literature**: fair graph mining [498].
- **Data spec**: user-user pairs.
- **Sample size**: $\sim$ 4K people connected by $\sim$ 90K friend relations.
- **Year**: 2012.
- **Sensitive features**: geography, gender.
- **Link**: https://snap.stanford.edu/data/egonets-Facebook.html
- **Further info**: Leskovec and Mcauley [494]

A.1.62 Facebook Large Network

- **Description**: this dataset was developed to study the effectiveness of node embeddings for learning tasks defined on graphs. The dataset concentrates on verified Facebook pages of politicians, governmental organizations, television shows, and companies, represented as nodes, while edges represent mutual likes. In addition, each page comes with node embeddings which are extracted from the textual description of each page. The original task on this dataset is page category classification.
- **Affiliation of creators**: University of Edinburgh.
- **Domain**: social networks.
- **Tasks in fairness literature**: fair graph mining evaluation [425].
- **Data spec**: page-page pairs.
- **Sample size**: $\sim$20K nodes (pages) connected by $\sim$ 200K edges (mutual likes).
- **Year**: 2019.
- **Sensitive features**: none.
- **Link**: http://snap.stanford.edu/data/facebook-large-page-page-network.html
- **Further info**: Rozemberczki et al. [683]

A.1.63 FACES

- **Description**: this resource contains images of Caucasian individuals of variable age and gender under six predefined facial expressions (neutrality, sadness, disgust, fear, anger, and happiness). This dataset is described as a database of emotion-related stimuli for scientific research. Subjects were hired through a model agency in Berlin, and suitably informed about the purpose of the photo-shooting session, thereafter signing an informed consent document. Each model reported their own age and gender. The necessary facial expressions were carefully explained with the help of a manual, with attention to the position of muscles. Photographs were obtained and post-processed in a standardized fashion, and later validated by raters of different ages with respect to the perceived expression and age of subjects. At a later stage, images were also annotated for attractiveness and distinctiveness. Currently, a small subset of the images is publicly available, while the full dataset is available after registration.
- **Affiliation of creators**: Max Planck Institute for Human Development.
- **Domain**: computer vision, experimental psychology.
- **Tasks in fairness literature**: fairness evaluation [444].
- **Data spec**: image.
- **Sample size**: $\sim$ 2K images of $\sim$ 200 people.
- **Year**: 2010.
- **Sensitive features**: age, gender.
- **Link**: https://faces.mpib-berlin.mpg.de/imeji/
- **Further info**: Ebner et al. [225]

A.1.64 FairFace

- **Description**: this dataset was developed as a balanced resource for face analysis with diverse race, gender and age composition. The associated task is race, gender and age classification. Starting from a large public image dataset (Yahoo YFCC100M), the authors sampled images incrementally to ensure diversity with respect to race, for which they considered seven categories: White, Black, Indian, East Asian, Southeast Asian, Middle East, and Latino. Sensitive attributes were annotated by workers on Amazon Mechanical Turk, and also through a model based on these annotations. Faces with low agreement between model and annotators were manually re-verified by the dataset curators. This dataset was annotated automatically with a binary Fitzpatrick skin tone label [145].
- **Affiliation of creators**: University of California, Los Angeles.
- **Domain**: computer vision.
- **Tasks in fairness literature**: fairness evaluation of private classification [145].
- **Data spec**: image.
- **Sample size**: $\sim$ 100K images.
- **Year**: 2019.
- **Sensitive features**: race, age, gender, skin tone.
- **Link**: https://github.com/joojs/fairface
- **Further info**: Karkkainen and Joo [429]

A.1.65 Fantasy Football

- **Description**: this resource was curated to study the problem of fair ranking aggregation. The creators collected rankings of National Football League players from the top 25 experts on the popular fantasy sports website FantasyPros. The data covers 16 weeks during the 2019 football season. Players

are assigned to different sensitive groups based on the conference of their team (American Football Conference or National Football Conference). The data available online concentrates on wide receivers.
- **Affiliation of creators**: Worcester Polytechnic Institute.
- **Domain**: sports.
- **Tasks in fairness literature**: fair ranking evaluation [469].
- **Data spec**: player-expert pairs.
- **Sample size**: $\sim$ 50 players, ranked by 25 experts (on a weekly basis), over 16 weeks.
- **Year**: 2020.
- **Sensitive features**: football conference.
- **Link**: https://arcgit.wpi.edu/cakuhlman/VLDB2020/tree/master/charts/data
- **Further info**: Kuhlman and Rundensteiner [470]

A.1.66 Fashion MNIST

- **Description**: this dataset is based on product assortement from the Zalando website. It contains gray-scale resized versions of thumbnail images of unique clothing products, labeled by in-house fashion experts according to their category, including e.g. trousers, coat and shirt. The envisioned task is object classification. The dataset, sharing the same size and structure as MNIST, was developed to provide a harder and more representative task, and to replace MNIST as a popular computer vision benchmark.
- **Affiliation of creators**: Zalando.
- **Domain**: computer vision.
- **Tasks in fairness literature**: robust fairness evaluation [77].
- **Data spec**: image.
- **Sample size**: $\sim$ 70K images across 10 product categories.
- **Year**: 2017.
- **Sensitive features**: none.
- **Link**: https://github.com/zalandoresearch/fashion-mnist
- **Further info**: Xiao et al. [834]

A.1.67 FICO

- **Description**: based on a sample of 301,536 TransUnion TransRisk scores from 2003, this dataset was created to study the problem of adjusting predictors for compliance with the equality of opportunity fairness metric. The TransUnion data was preprocessed and aggregated to summarize the CDF of risk scores by race (Non-Hispanic white, Black, Hispanic, Asian). The original data comes from a 2007 report to the US Congress on credit scoring and its effects on the availability and affordability of credit carried out by a dedicated Federal Reserve working group. The collection, creation, processing, and aggregation was carried out by the working group; the data was later scraped by the creators, who made it available without any modification.
- **Affiliation of creators**: Google; University of Texas at Austin; Toyota Technological Institute at Chicago.
- **Domain**: finance.
- **Tasks in fairness literature**: fairness evaluation [344], dynamical fair classification [510], dynamical fairness evaluation [184, 508, 869], fair resource allocation [311].
- **Data spec**: tabular data.
- **Sample size**: N/As. CDFs are provided over risk scores which are normalized (0-100%) and quantized with step 0.5%.

- **Year**: 2016.
- **Sensitive features**: race.
- **Link**: https://github.com/fairmlbook/fairmlbook.github.io/tree/master/code/creditscore/data
- **Further info**: Barocas et al. [50], Hardt et al. [344], US Federal Reserve [782]

A.1.68 FIFA 20 Players

- **Description**: this dataset was scraped by Stefano Leone and made available on Kaggle. It includes the players' data for the Career Mode from FIFA 15 to FIFA 20, a popular football game. Several tasks are envisioned for this dataset, including a historical comparison of players.
- **Affiliation of creators**: unknown.
- **Domain**: sports.
- **Tasks in fairness literature**: fairness evaluation under unawareness [28].
- **Data spec**: tabular data.
- **Sample size**: $\sim$ 20K players.
- **Year**: 2019.
- **Sensitive features**: geography.
- **Link**: https://www.kaggle.com/stefanoleone992/fifa-20-complete-player-dataset
- **Further info**:

A.1.69 FilmTrust

- **Description**: this dataset was crawled from the entire FilmTrust website, a movie recommendation service with a social network component. The dataset comprises user-movie ratings on a 5-star scale and user-user indications of trust about movie taste. This resource can be used to train and evaluate recommender systems.
- **Affiliation of creators**: Northeastern University; Nanyang Technological University; American University of Beirut; University of Cambridge.
- **Domain**: information systems, movies.
- **Tasks in fairness literature**: fair ranking [511].
- **Data spec**: user-movie pairs and user-user pairs.
- **Sample size**: $\sim$ 40K ratings by $\sim$ 2K users over $\sim$ 2K movies.
- **Year**: 2011.
- **Sensitive features**: none.
- **Link**: https://guoguibing.github.io/librec/datasets.html
- **Further info**: Guo et al. [332]

A.1.70 Framingham

- **Description**: the Framingham Heart Study began in 1948 under the direction of the National Heart, Lung, and Blood Institute (NHLBI), with the goal of identifying key factors that contribute to cardiovascular disease, given a mounting epidemic of cardiovascular disease whose etiology was mostly unknown at the time. Six different cohorts have been recruited over the years among citizens of Framingham, Massachusetts, without symptoms of cardiovascular disease. After the original cohort, two more were enrolled from the children and grandchildren of the first one. Additional cohorts were also started to reflect the increased racial and ethnic diversity in the town of Framingham. Participants in the study

report on their habits (e.g. physical activity, smoking) and undergo regular physical examination and laboratory tests.
- **Affiliation of creators**: National Heart, Lung, and Blood Institute (NHLBI); Boston University.
- **Domain**: cardiology.
- **Tasks in fairness literature**: fair ranking evaluation [421].
- **Data spec**: mixture.
- **Sample size**: $\sim$ 15K respondents.
- **Year**: present.
- **Sensitive features**: age, sex, race.
- **Link**: https://framinghamheartstudy.org/
- **Further info**: Kannel and McGee [426], Tsao and Vasan [769]

A.1.71 Freebase15k-237

- **Description**: Freebase was a collaborative knowledge base which allowed its community members to fill in structured data about diverse entities and relations between them. This database was developed from a prior Freebase dataset [86], pruning it from redundant relations and augmenting it with textual relationships from the ClueWeb12 corpus. The creators of this dataset worked on the joint optimization of entity knowledge base and representations of the entities' textual relations, with the goal of providing representations of entities suited for knowledge base completion.
- **Affiliation of creators**: Microsoft; Stanford University.
- **Domain**: information systems.
- **Tasks in fairness literature**: fair graph mining [89], fairness evaluation in graph mining [267].
- **Data spec**: entity-relation-entity triples.
- **Sample size**: $\sim$ 15K entities connected by 170K edges (relations).
- **Year**: 2016.
- **Sensitive features**: demographics of people featured in entities and their relations.
- **Link**: https://www.microsoft.com/en-us/download/details.aspx?id=52312
- **Further info**: Toutanova et al. [766]

A.1.72 GAP Coreference

- **Description**: this resource was developed as a gender-balanced coreference resolution dataset, useful for auditing gender-dependent differences in the accuracy of existing pronoun resolution algorithms and for training new algorithms that are less gender-biased. The dataset consists of thousands of ambiguous pronoun-name pairs in sentences extracted from Wikipedia. Several measures are taken to avoid the success of naïve heuristics and to favour diversity. Most notably, while the initial (automated) stage of the data collection pipeline extracts contexts with a female:male ratio of 1:9, feminine pronouns are oversampled to achieve a 1:1 ratio. Each example is presented to and annotated for coreference by three in-house workers.
- **Affiliation of creators**: Google.
- **Domain**: linguistics.
- **Tasks in fairness literature**: data bias evaluation [458].
- **Data spec**: text.
- **Sample size**: $\sim$ 9K sentences.
- **Year**: 2018.
- **Sensitive features**: gender.
- **Link**: https://github.com/google-research-datasets/gap-coreference
- **Further info**: Webster et al. [818]

A.1.73 German Credit

- **Description**: the German Credit dataset was created to study the problem of automated credit decisions at a regional Bank in southern Germany. Instances represent loan applicants from 1973 to 1975, who were deemed creditworthy and were granted a loan, bringing about a natural selection bias. The data summarizes their financial situation, credit history and personal situation, including housing and number of liable people. A binary variable encoding whether each loan recipient punctually payed every installment is the target of a classification task. Among covariates, marital status and sex are jointly encoded in a single variable. Many documentation mistakes are present in the UCI entry associated with this resource [774]. Due to one of these mistakes, users of this dataset are led to believe that the variable sex can be retrieved from the joint marital_status-sex variable, however this is false. A revised version with correct variable encodings, called South German Credit, was donated to UCI Machine Learning Repository [776] with an accompanying report [329]. See Appendix A.4 for extensive documentation.
- **Affiliation of creators**: Hypo Bank (OP/EDV-VP); Universität Hamburg; Strathclyde University (German Credit); Beuth University of Applied Sciences Berlin (South German Credit).
- **Domain**: finance.
- **Tasks in fairness literature**: fair classification [34, 124, 197, 216, 356, 514, 534, 535, 633, 657, 658, 713, 715, 791, 841], fairness evaluation [261, 282], fair active resource allocation [105], preference-based fair classification [871], fair active classification [600], fair classification under unawareness [441], robust fairness evaluation [77], fair representation learning [515, 686], fair reinforcement learning [561], fair ranking evaluation [421, 832, 843], fair ranking [90, 723], fair multi-stage classification [306], limited-label fair classification [152, 157, 810], limited-label fairness evaluation [403].
- **Data spec**: tabular data.
- **Sample size**: $\sim$ 1K.
- **Year**: 1994 (German Credit); 2020 (South German Credit).
- **Sensitive features**: age, geography.
- **Link**: https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data) (German Credit); https://archive.ics.uci.edu/ml/datasets/South+German+Credit+%28UPDATE%29 (South German Credit)
- **Further info**: Grömping [329]

A.1.74 German Political Posts

- **Description**: this dataset was used as a training set for German word embeddings, with the goal of investigating biases in word representations. The authors used the Facebook and Twitter APIs to collect posts and comments from the social media channels of six main political parties in Germany (CDU/CSU, SPD, Bundnis90/Die Grünen, FDP, Die Linke, AfD). Facebook posts are from the period 2015–2018, while tweets were collected between January and October 2018. Overall, the dataset consists of millions of posts, for a total of half a billion tokens. A subset of the Facebook comments (100,000) were labeled by human annotators based on whether they contain sexist content, with four sub-labels indicating sexist comments, sexist buzzwords, gender-related compliments, statements against gender equality and assignment of gender stereotypical roles to people.
- **Affiliation of creators**: Technical University of Munich.
- **Domain**: social media.
- **Tasks in fairness literature**: bias evaluation in WEs [620].
- **Data spec**: text.

- **Sample size**: $\sim$ 20M posts comments and tweets.
- **Year**: 2020.
- **Sensitive features**: textual references to people and their demographics.
- **Link**: not available
- **Further info**: Papakyriakopoulos et al. [620]

A.1.75 GLUE

- **Description**: this benchmark was assembled to reliably evaluate the progress of natural language processing models. It consists of multiple datasets and associated tasks from the natural language processing domain, including paraphrase detection, textual entailment, sentiment analysis and question answering. Given the quick progress registered by language models on GLUE, a similar benchmark called SuperGLUE was subsequently released comprising more challenging and diverse tasks [806].
- **Affiliation of creators**: New York University; University of Washington; DeepMind.
- **Domain**: linguistics.
- **Tasks in fairness literature**: fairness evaluation [29, 684], bias evaluation in language models [144], fairness evaluation of selective classification [411].
- **Data spec**: text.
- **Sample size**: $\sim 100 - 400$K samples. Datasets have variable sizes spanning three orders of magnitude.
- **Year**: 2018.
- **Sensitive features**: none.
- **Link**: https://gluebenchmark.com/
- **Further info**: Wang et al. [807]

A.1.76 Goodreads Reviews

- **Description**: there are several versions of this dataset, corresponding to different crawls. Here we refer to the most well documented one by Wan and McAuley [805]. This resource consists of anonymized reviews collected from public user *book shelves*. Rich metadata is available for books and reviews, including. authors, country code, publisher, userid, rating, timestamp, and text. A few medium-size subsamples focused on specific book genres are available. The task typically associated with this resource is book recommendation.
- **Affiliation of creators**: University of California, San Diego.
- **Domain**: literature, information systems.
- **Tasks in fairness literature**: fair ranking evaluation [661], fairness evaluation [138].
- **Data spec**: user-book pairs.
- **Sample size**: $\sim$ 200M records from $\sim$ 900K users over $\sim$ 2M books.
- **Year**: 2019.
- **Sensitive features**: author.
- **Link**: https://sites.google.com/eng.ucsd.edu/ucsdbookgraph/
- **Further info**: Wan and McAuley [805]

A.1.77 Google Local

- **Description**: this dataset contains reviews and ratings from millions of users on local businesses from five different continents. Businesses are labelled with nearly 50 thousand categories. This resource was collected as a real world example of interactions between users and ratable items, with the goal

of testing novel recommendation approaches. The dataset comprises data that is specific to users (e.g. places lived), businesses (e.g. GPS coordinates), and reviews (e.g. timestamps).
- **Affiliation of creators**: University of California, San Diego.
- **Domain**: information systems.
- **Tasks in fairness literature**: fair ranking [625].
- **Data spec**: user-business pairs.
- **Sample size**: $\sim$ 10M reviews and ratings from $\sim$ 5M users on $\sim$ 3M local businesses.
- **Year**: 2018.
- **Sensitive features**: geography.
- **Link**: https://cseweb.ucsd.edu/~jmcauley/datasets.html#google_local
- **Further info**: He et al. [353]

A.1.78 Greek Websites

- **Description**: this dataset was created to demonstrate the *bias goggles* tools, which enables users to explore diverse bias aspects connected with popular Greek web domains. The dataset is a subset of the Greek web, crawled from Greek websites that cover politics and sports, represent big industries, or are generally popular. Starting from a seed of hundreds of websites, crawlers followed the links up to depth 7, avoiding popular sites such as Facebook and Twitter. The final dataset has a graph structure, comprising pages and links between them.
- **Affiliation of creators**: FORTH-ICS, University of Crete.
- **Domain**: .
- **Tasks in fairness literature**: bias discovery[461].
- **Data spec**: page-page pairs.
- **Sample size**: $\sim$ 900k pages from $\sim$ 90k domains.
- **Year**: 2020.
- **Sensitive features**: none.
- **Link**: https://pangaia.ics.forth.gr/bias-goggles/about.html#Dataset
- **Further info**: Konstantakis et al. [461]

A.1.79 Guardian Articles

- **Description**: this dataset consists of articles from *The Guardian*, retrieved from The Guardian Open Platform API. In particular, the authors crawled every article that appeared on the website between 2009 and 2018. They created this dataset to demonstrate a framework for the identification of gender biases in training data for machine learning.
- **Affiliation of creators**: University College Dublin.
- **Domain**: news.
- **Tasks in fairness literature**: data bias evaluation [487].
- **Data spec**: text.
- **Sample size**: unknown.
- **Year**: 2020.
- **Sensitive features**: textual references to people and their demographics.
- **Link**: not available
- **Further info**: Leavy et al. [487]

A.1.80 HAM10000

- **Description**: the dataset comprises 10,015 dermatoscopic images collected over a period of 20 years the Department of Dermatology at the Medical University of Vienna, Austria and the skin cancer practice of Cliff Rosendahl in Queensland, Australia. Images were acquired and stored through different modalities; each image depicts a lesion and comes with metadata detailing the region of skin lesion, patient demographics, and diagnosis, which is the target variable. The dataset was employed for the lesion disease classification of the ISIC 2018 challenge.
- **Affiliation of creators**: Medical University of Vienna; University of Queensland.
- **Domain**: dermatology.
- **Tasks in fairness literature**: fair classification [534].
- **Data spec**: image.
- **Sample size**: ∼10K images.
- **Year**: 2018.
- **Sensitive features**: age, sex.
- **Link**: https://doi.org/10.7910/DVN/DBW86T
- **Further info**: Tschandl et al. [770]

A.1.81 Harvey Rescue

- **Description**: this dataset is the result of crowdsourced efforts to connect rescue parties with people requesting help in the Houston area, mostly due to the flooding caused by Hurricane Harvey. Most requests are from August 28, 2017, and were sent via social media; they are timestamped and associated with the location of the people seeking help.
- **Affiliation of creators**: Harvey Relief Handiworks; Harvey Relief Coalition.
- **Domain**: social work.
- **Tasks in fairness literature**: fair spatio-temporal process learning [711].
- **Data spec**: tabular data.
- **Sample size**: ∼1K help requests.
- **Year**: 2017.
- **Sensitive features**: geography.
- **Link**: not available
- **Further info**: http://harveyrelief.handiworks.co/

A.1.82 Heart Disease

- **Description**: this dataset is a collection of medical data from separate groups of patients referred for cardiac catheterisation and coronary angiography at 5 different medical centers, namely the Cleveland Clinic (data from 1981–1984), the Hungarian Institute of Cardiology in Budapest (1983–1987), the Long Beach Veterans Administration Medical Center (1984–1987) and the University Hospitals of Basel and Zurich (1985). The binary target variable in this dataset encodes a diagnosis of Coronary artery disease. Covariates relate to patient demographics, exercise data (e.g. maximum heart rate) and routine test data (e.g. resting blood pressure). Overall, 76 covariates are available but 14 are recommended. Names and social security numbers of the patients were initially available, but have been removed from the publicly available dataset.
- **Affiliation of creators**: Veterans Administration Medical Center, Long Beach; Hungarian Institute of Cardiology, Budapest; University Hospital, Zurich; University Hospital, Basel; Studer Corporation; Stanford University.

- **Domain**: cardiology.
- **Tasks in fairness literature**: fairness evaluation [642], fair active classification [600].
- **Data spec**: tabular data.
- **Sample size**: $\sim$ 1K patients.
- **Year**: 1988.
- **Sensitive features**: age, sex.
- **Link**: https://archive.ics.uci.edu/ml/datasets/heart+disease
- **Further info**: Detrano et al. [202]

A.1.83 Heritage Health

- **Description**: this dataset was developed as part of the Heritage Health Prize competition with the goal of reducing the cost of health care by decreasing the number of avoidable hospitalizations. The competition requires predicting the number of days a patient will spend in hospital during the 12 months following a cutoff date. The dataset features basic demographic information about patients, along with data about prior hospitalizations (e.g. length of stay and diagnosis), laboratory tests and prescriptions.
- **Affiliation of creators**: CHEO Research Institute, Inc; University of Ottawa; University of Maryland; Privacy Analytics, Inc; Kaggle; Heritage Provider Network.
- **Domain**: health policy.
- **Tasks in fairness literature**: fair multi-stage classification [525], fair representation learning [515], fair classification [657, 658], fair transfer learning [523], fairness evaluation [390].
- **Data spec**: tabular data.
- **Sample size**: $\sim$ 150K patients.
- **Year**: 2011.
- **Sensitive features**: age, sex.
- **Link**: https://www.kaggle.com/c/hhp/data
- **Further info**: El Emam et al. [232]

A.1.84 High School Contact and Friendship Network

- **Description**: this dataset was developed to compare and contrast different methods commonly employed to measure human interaction and build the underlying social network. Data corresponds to interactions and friendship relations between students of a French high school in Marseilles. The authors consider four different methods of network data collection, namely face-to-face contacts measured by two concurrent methods (sensors and diaries), self-reported friendship surveys, and Facebook links.
- **Affiliation of creators**: Aix Marseille Université; Université de Toulon; Centre national de la recherche scientifique; ISI Foundation.
- **Domain**: social networks.
- **Tasks in fairness literature**: fair graph clustering [454].
- **Data spec**: student-student pairs.
- **Sample size**: $\sim$ 300 students.
- **Year**: 2015.
- **Sensitive features**: gender.
- **Link**: http://www.sociopatterns.org/datasets/high-school-contact-and-friendship-networks/
- **Further info**: Mastrandrea et al. [536]

A.1.85 HMDA

- **Description**: The Home Mortgage Disclosure Act (HMDA) is a US federal law from 1975 mandating that financial institutions maintain and disclose information about mortgages to the public. Companies submit a Loan Application Register (LAR) to the Federal Financial Institutions Examination Council FFIEC who maintain and disclose the data. The LAR format is subject to changes, such as the one which happened in 2017. From 2018 onward, entries to the LAR comprise information about the financial institution (e.g. geography, id), the applicants (e.g. demographics, income), the house (e.g. value, construction method), the mortgage conditions (type, interest rate, amount) and the outcome. Ethnicity, race, and sex of applicants are self-reported.
- **Affiliation of creators**: Federal Financial Institutions Examination Council.
- **Domain**: finance.
- **Tasks in fairness literature**: fairness evaluation under unawareness [139, 418].
- **Data spec**: tabular data.
- **Sample size**: $\sim$ 200M records.
- **Year**: present.
- **Sensitive features**: sex, geography, race, ethnicity.
- **Link**: https://ffiec.cfpb.gov/data-browser/
- **Further info**: https://ffiec.cfpb.gov/; https://www.consumerfinance.gov/data-research/hmda/

A.1.86 Homeless Youths' Social Networks

- **Description**: this dataset was collected to study methamphetamine use norms among homeless youth in association with their social networks. A sample of homeless youth aged 13–25 years was recruited between 2011—2012 from two drop-in centers in California. After obtaining informed consent/assent, participants filled in a survey and answered questions from an interview. The survey included questions on demographics, migratory status, educational status and housing. To reconstruct the social network between them, each participant provided information for up to 50 people with whom they had interacted during the previous 30 days.
- **Affiliation of creators**: University of Denver; University of Southern California.
- **Domain**: social work.
- **Tasks in fairness literature**: fair graph diffusion [659].
- **Data spec**: person-person pairs.
- **Sample size**: $\sim$ 300 youth.
- **Year**: 2015.
- **Sensitive features**: age, gender, sexual orientation, race and ethnicity.
- **Link**: not available
- **Further info**: Barman-Adhikari et al. [47]

A.1.87 IBM HR Analytics

- **Description**: based on the information available on Kaggle, this is a fictional dataset created by IBM data scientists. It describes employees along dimensions that may be relevant for attrition, the target variable encoding employee departure. Available covariates include information on employee background (education, number of prior companies), work satisfaction (recent promotions, environment and job satisfaction) and seniority (years at the company, years in current role, job level).

- **Affiliation of creators**: IBM.
- **Domain**: information systems, management information systems.
- **Tasks in fairness literature**: fair data generation [507].
- **Data spec**: tabular data.
- **Sample size**: $\sim 1K$ employees.
- **Year**: 2019.
- **Sensitive features**: gender.
- **Link**: https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset
- **Further info**: https://github.com/IBM/employee-attrition-aif360

A.1.88 IIT-JEE

- **Description**: this dataset was released in response to a Right to Information application filed in June 2009, and contains country-wide results for the Joint Entrance Exam (EET) to Indian Institutes of Technology (IITs), a group of prestigious engineering schools in India. The dataset contains the marks obtained by every candidate who took the test in 2009, divided according to the specific Math, Physics, and Chemistry sections of the test. Demographics such as ZIP code, gender, and birth categories (ethnic categories relating to the caste system) are also included.
- **Affiliation of creators**: Indian Institute of Technology, Kharagpur.
- **Domain**: education.
- **Tasks in fairness literature**: fair ranking [127].
- **Data spec**: tabular data.
- **Sample size**: $\sim$ 400K students.
- **Year**: 2009.
- **Sensitive features**: gender, birth category.
- **Link**: not available
- **Further info**: Celis et al. [127]

A.1.89 IJB-A

- **Description**: the IARPA Janus Benchmark A (IJB-A) dataset was proposed as a face recognition benchmark with wide geographic representation and pose variation for subjects. It consists of *in-the-wild* images and videos of 500 subjects, obtained through internet searches over Creative Commons licensed content. The subjects were manually specified by the creators of the dataset to ensure broad geographic representation. The tasks associated with the dataset are face identification and verification. The dataset curators also collected the subjects' skin color and gender, through an unspecified annotation procedure. Similar protected attributes (gender and Fitzpatrick skin type) were labelled by one author of Buolamwini and Gebru [101].
- **Affiliation of creators**: Noblis; National Institute of Standards and Technology (NIST); Intelligence Advanced Research Projects Activity (IARPA); Michigan State University.
- **Domain**: computer vision.
- **Tasks in fairness literature**: data bias evaluation [101].
- **Data spec**: image.
- **Sample size**: $\sim$ 6K images of $\sim$ 500 subjects.
- **Year**: 2015.
- **Sensitive features**: gender, skin color.
- **Link**: https://www.nist.gov/itl/iad/image-group/ijb-dataset-request-form
- **Further info**: Klare et al. [451]

A.1.90 ILEA

- **Description**: this dataset was created by the Inner London Education Authority (ILEA) considering data from 140 British schools. It comprises the results of public examinations taken by students of age 16 over the period 1985–1987. These values are used as a measurement of school effectiveness, with emphasis on quality of education and equality of opportunity for students of different backgrounds and ethnicities. Student-level records report their sex and ethnicity, while school-level factors include the percentage of students eligible for free meals and the percentage of girls in each institute.
- **Affiliation of creators**: Inner London Education Authority (ILEA).
- **Domain**: education.
- **Tasks in fairness literature**: fair representation learning [612, 613].
- **Data spec**: unknown.
- **Sample size**: $\sim$ 30K students from 140 secondary schools.
- **Year**: unknown.
- **Sensitive features**: age, sex, ethnicity.
- **Link**: not available
- **Further info**: [313, 606]

A.1.91 Image Embedding Association Test (iEAT)

- **Description**: the Image Embedding Association Test (iEAT) is a resource for quantifying biased associations between representations of social concepts and attributes in images. It mimics seminal work on biases in WEs [109], following the Implicit Association Test (IAT) from social psychology [326]. The curators identified several combinations of target concepts (e.g. young) and attributes (e.g. pleasant), testing similarities between representations of these concepts learnt by unsupervised computer vision models. For each attribute/concept they obtained a set of images from the IAT, the CIFAR-100 dataset or Google Image Search, which act as the source of images and the associated sensitive attribute labels.
- **Affiliation of creators**: Carnegie Mellon University; George Washington University.
- **Domain**: computer vision.
- **Tasks in fairness literature**: fairness evaluation of learnt representations [743].
- **Data spec**: image.
- **Sample size**: $\sim$ 200 image for 15 iEATs.
- **Year**: 2021.
- **Sensitive features**: religion, gender, age, race, sexual orientation, disability, skin tone, weight.
- **Link**: https://github.com/ryansteed/ieat/tree/master/data
- **Further info**: Steed and Caliskan [743]

A.1.92 ImageNet

- **Description**: Imagenet is one of the most influential machine learning dataset of the 2010s. Much important work on computer vision, including early breakthroughs in deep learning has been sparked by ImageNet Large Scale Visual Recognition Challenge (ILSVRC), a competition held yearly from 2010 to 2017. The most used portion of ImageNet is indeed the data powering the classification task in ILSVRC 2012, featuring 1,000 classes, over 100 of which represent different dog breeds. Recently, several problematic biases were found in the `person` subtree of ImageNet, tracing their causes and proposing approaches to remove them [181, 644, 842].

- **Affiliation of creators**: Princeton University.
- **Domain**: computer vision.
- **Tasks in fairness literature**: fair classification [222], bias discovery [14], data bias evaluation [842], fair incremental learning [876], fairness evaluation [221].
- **Data spec**: image.
- **Sample size**: $\sim$ 14M images depicting $\sim$ 20K categories (synsets).
- **Year**: 2021.
- **Sensitive features**: people's gender and other sensitive annotations may be present in synsets from the person subtree.
- **Link**: https://image-net.org/
- **Further info**: Barocas et al. [50], Crawford and Paglen [181], Deng et al. [198], Prabhu and Birhane [644], Yang et al. [842]

A.1.93 In-Situ

- **Description**: this dataset was curated to measure biases in named entity recognition algorithms, based on gender, race and religion of people represented by entities. The authors exploit census data to build a list of 123 names typical of men and women of different race and religion. Next, they extract 289 sentences mentioning people from the CoNLL 2003 NER test data [764], itself derived from Reuters 1990s news stories. Finally, they substitute the unigram person entity from the CoNLL 2003 shared task with each of names obtained previously as specific to a demographic group.
- **Affiliation of creators**: Twitter.
- **Domain**: linguistics.
- **Tasks in fairness literature**: fairness evaluation in entity recognition [569].
- **Data spec**: text.
- **Sample size**: $\sim$ 50K sentences.
- **Year**: 2020.
- **Sensitive features**: gender, race and religion.
- **Link**: https://github.com/napsternxg/NER_bias
- **Further info**: Mishra et al. [569]

A.1.94 iNaturalist Datasets

- **Description**: these datasets were curated as challenging real-world benchmarks for large-scale fine-grained visual classification and feature visually similar classes with large class imbalance. They consist of images of plants and animals from iNaturalist, a social network where nature enthusiasts share information and observations about biodiversity. There are four different releases of the dataset: 2017, 2018, 2019, and 2021. A subset of the images are also annotated with bounding boxes and have additional metadata such as where and when the images were captured.
- **Affiliation of creators**: California Institute of Technology; University of Edinburgh; Google; Cornell University; iNaturalist.
- **Domain**: biology.
- **Tasks in fairness literature**: fairness evaluation of private classification [33].
- **Data spec**: image.
- **Sample size**: $\sim$ 3M images from $\sim$ 10K different species of plants and animals.
- **Year**: 2021.
- **Sensitive features**: none.
- **Link**: https://github.com/visipedia/inat_comp
- **Further info**: [789, 790]

A.1.95 Indian Census

- **Description**: very little information seems to be available on this dataset. It represents a count of residents of 35 Indian states, repeated every ten years between 1951 and 2001.
- **Affiliation of creators**: Office of the Registrar General of India.
- **Domain**: demography.
- **Tasks in fairness literature**: fairness evaluation of private resource allocation [649].
- **Data spec**: tabular data.
- **Sample size**: $\sim$ 30 state.
- **Year**: unknown.
- **Sensitive features**: geography.
- **Link**: https://www.indiabudget.gov.in/budget_archive/es2006-07/chapt2007/tab97.pdf
- **Further info**:

A.1.96 Indian Student Performance

- **Description**: this dataset was curated to support educational data mining algorithms. The creators collected data from three colleges of Assam, India (Duliajan College, Doomdooma College, and Digboi College). Each data point represents a student, summarizing information on their demographics (gender, caste), family (occupation and qualification of parents), and school fruition (study hours, attendance, home-to-school travel). Among the latter there are four variables summarizing student performance in different classes and examinations, which represent the response variable of a prediction task.
- **Affiliation of creators**: Dibrugarh University; Sana'a University; Abdelmalek Essaâdi University.
- **Domain**: education.
- **Tasks in fairness literature**: fair data summarization [56].
- **Data spec**: tabular data.
- **Sample size**: $\sim$ 300 students.
- **Year**: 2018.
- **Sensitive features**: gender, caste, geography.
- **Link**: https://archive.ics.uci.edu/ml/datasets/Student+Academics+Performance
- **Further info**: Hussain et al. [383]

A.1.97 Infant Health and Development Program (IHDP)

- **Description**: this dataset is the result of the IHDP program carried out between 1985 and 1988 in the US. A longitudinal randomized trial was conducted to evaluate the effectiveness of comprehensive early intervention in reducing developmental and health problems in low birth weight premature infants. Families in the experimental group received an intervention based on an educational program delivered through home visits, a daily center-based program and a parent supporting group. Children in the study were assessed across multiple cognitive, behavioral, and health dimensions longitudinally in four phases at ages 3, 5, 8, and 18. The dataset also contains information on household composition, source of health care, parents' demographics and employment.
- **Affiliation of creators**: unknown.
- **Domain**: pediatrics.
- **Tasks in fairness literature**: fair risk assessment [524, 849].
- **Data spec**: mixture.
- **Sample size**: $\sim$ 1K infants.
- **Year**: 1993.

- **Sensitive features**: race and ethnicity (of parents), age (maternal), gender (of infant).
- **Link**: https://www.icpsr.umich.edu/web/HMCA/studies/9795
- **Further info**: Brooks-Gunn et al. [95]

A.1.98 Instagram Photos

- **Description**: this dataset was crawled from Instagram to explore trade-offs between fairness and revenue in platforms that serve ads to their users. The authors crawled metadata from photos (location and tags) and users (names), using Kevin Systrom as a seed user and cascading into profiles that like or comment photos. The curators concentrated on cities with enough geotagged data, namely New York and Los Angeles. Moreover, they labeled the users with gender and race. Gender was labeled via US social security data, using the proportion of babies with a given name registered with either gender. Gender was only assigned to users with a first name for which there were both at least 50 births and 95% of recorded births were one gender. Race were labeled using the Face++ API on a subset of photos. Photos were not downloaded, rather they were fed to Face++ via their publicly available URL. Finally, the ground truth labels were validated by two research assistants. To emulate a location-based advertisement model, the creators devised a task aimed at predicting what topics a user will be interested in, given their locations from previous check-ins.
- **Affiliation of creators**: Columbia University.
- **Domain**: social media.
- **Tasks in fairness literature**: fair advertising [674].
- **Data spec**: unknown.
- **Sample size**: $\sim$ 1M photos from $\sim$ 40K users.
- **Year**: 2017.
- **Sensitive features**: race, gender, geography.
- **Link**: not available
- **Further info**: Riederer and Chaintreau [674]

A.1.99 Internet Ads

- **Description**: this dataset was assembled to study the problem of automated advertisement removal in browsers. It consists of images crawled from randomly generated urls, manually classified as ad/no-ad. Image encodings are derived from raw html, thus containing no information about pixel values, but rather encoding width, height, anchor text and image source. The associated task is classifying each image encoding as an ad or a no-ad image.
- **Affiliation of creators**: University College Dublin.
- **Domain**: pattern recognition.
- **Tasks in fairness literature**: fair anomaly detection [716].
- **Data spec**: tabular data.
- **Sample size**: $\sim$ 3K image encodings.
- **Year**: 1998.
- **Sensitive features**: none.
- **Link**: https://archive.ics.uci.edu/ml/datasets/internet+advertisements
- **Further info**: Kushmerick [474]

A.1.100 Iris

- **Description**: the most popular dataset on the UCI Machine Learning Repository was created by E. Anderson and popularized by R.A. Fisher in the pattern recognition community in the 1930s. The measurements in this collection represent the length and width of sepal and petals of different Iris flowers, collected to evaluate the morphological variation of different Iris species. The typical learning task associated with this dataset is labelling the species based on the available measurements.
- **Affiliation of creators**: Missouri Botanical Garden; Washington University.
- **Domain**: plant science.
- **Tasks in fairness literature**: fair clustering [1, 142].
- **Data spec**: tabular data.
- **Sample size**: $\sim$ 100 samples from three species of Iris.
- **Year**: 1988.
- **Sensitive features**: none.
- **Link**: https://archive.ics.uci.edu/ml/datasets/iris
- **Further info**: [15, 268]

A.1.101 Italian Car Insurance

- **Description**: this resource was curated to study discriminatory practices in the Italian car insurance market. More specifically, the data was collected to estimate the direct effect of gender and birthplace on yearly quoted premiums. It was collected in 2020 from a popular Italian car insurance comparison website, where the curators tried different hypothetical driver profiles and collected the quotes provided by nine companies. Along with gender and birthplace, additional driver features include age, city of residence, insured vehicle, mileage, and a summary of claim history.
- **Affiliation of creators**: University of Padua; Carnegie Mellon University; University of Udine.
- **Domain**: economics.
- **Tasks in fairness literature**: fair pricing evaluation [250].
- **Data spec**: tabular data.
- **Sample size**: $\sim$ 2K driver profiles.
- **Year**: 2021.
- **Sensitive features**: gender, birthplace.
- **Link**: not available
- **Further info**: Fabris et al. [250]

A.1.102 KDD Cup 99

- **Description**: this dataset was developed for a data mining competition on cybersecurity, focused on building an automated network intrusion detector based on TCP dump data. The task is predicting whether a connection is legitimate and inoffensive or symptomatic of an attack, such as denial-of-service or user-to-root; tens of attack classes have been simulated and annotated within this dataset. The available features include basic TCP/IP information, network traffic and contextual features, such as number of failed login attempts.
- **Affiliation of creators**: Massachusetts Institute of Technology.
- **Domain**: computer networks.
- **Tasks in fairness literature**: fair clustering [142].
- **Data spec**: tabular data.
- **Sample size**: $\sim$ 7M connections.

- **Year**: 1999.
- **Sensitive features**: none.
- **Link**: http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html
- **Further info**: Tavallaee et al. [761]

A.1.103 Kidney Exchange Program

- **Description**: this dataset is based on data of the Canadian Kidney Paired Donation Program (KPD) to study strategic behavior among entities controlling part of the incompatible patient-donor pairs. Based on data from the Canadian Blood Services on the KPD and census, these instances were generated. The random instance generator is available upon request. The instances are weighted graphs. The incompatible patient-donor pairs represent the vertices of the graph, an arc means that the donor of a vertex is compatible with the patient of another vertex, and weights represent the benefit of the donation. Compatibility is encoded based on true blood type distribution and risk of transplant rejection.
- **Affiliation of creators**: Université de Montréal; Polytechnique de Montréal.
- **Domain**: public health.
- **Tasks in fairness literature**: fair matching evaluation [255].
- **Data spec**: patient-donor pairs.
- **Sample size**: 180.
- **Year**: 2020.
- **Sensitive features**: blood type, geography.
- **Link**: https://github.com/mxmmargarida/KEG
- **Further info**: Carvalho and Lodi [117]

A.1.104 Kidney Matching

- **Description**: this dataset was created via a simulator based on real data provided by the Organ and Tissue Authority of Australia. The data was validated against additional information from the Australian Bureau of Statistics, the Public and Research sets, and Wikipedia. The simulator models the probability distribution over the Blood Type and State of donors and patients, along with the quality of a donated organ (summarized by Kidney Donor Patient Index) and of a patient (quantified by the Expected Post-Transplant Survival). The envisioned task for this data is optimal matching of organs and patients.
- **Affiliation of creators**: unknown.
- **Domain**: public health.
- **Tasks in fairness literature**: fairness matching evaluation [538].
- **Data spec**: tabular data.
- **Sample size**: unknown.
- **Year**: 2018.
- **Sensitive features**: age, geography, blood type.
- **Link**: not available
- **Further info**: Mattei et al. [537]

A.1.105 Kiva

- **Description**: this dataset was obtained from kiva.org, a non-profit organization allowing low-income entrepreneurs and students to borrow money through loan crowdfunding. The data summarizes all transactions occurred in 2017. Transactions are typically between 25$ to 50$ and range from 5$ to

10,000$. Features include information about the loan, such as its purpose, sector and amount, and data specific to the borrower and their demographics. Women are prevalent in this dataset, probably due to the priorities of partner organizations and the easier access to capital enjoyed by men in many countries.
- **Affiliation of creators**: Kiva; DePaul University.
- **Domain**: finance.
- **Tasks in fairness literature**: fair ranking [103, 511, 737], bias discovery [736].
- **Data spec**: tabular data.
- **Sample size**: $\sim$ 1M transactions involving $\sim$ 100K loans and $\sim$ 200K users.
- **Year**: 2018.
- **Sensitive features**: gender, geography, activity.
- **Link**: not available
- **Further info**: Sonboli and Burke [736]

A.1.106 Labeled Faces in the Wild (LFW)

- **Description**: LFW is a public benchmark for face verification, maintained by researchers affiliated with the University of Massachusetts. It was built to measure the progress of face verification systems in unconstrained settings (e.g. variable pose, illumination, resolution). The dataset consists of images of people who appeared in the news, labelled with the name of the respective individual. According to perception of human coders who were later asked to annotate this dataset, images mostly skew white, male and below 60.
- **Affiliation of creators**: University of Massachussets, Amherst; Stony Brook University.
- **Domain**: computer vision.
- **Tasks in fairness literature**: fair data summarization [693], fair clustering [298], robust fairness evaluation [77], fairness evaluation [706].
- **Data spec**: image.
- **Sample size**: $\sim$ 13K face images of $\sim$ 6K individuals.
- **Year**: 2007.
- **Sensitive features**: gender, age, race.
- **Link**: http://vis-www.cs.umass.edu/lfw/
- **Further info**: Gebru et al. [292], Han and Jain [338], Huang et al. [377]

A.1.107 Large Movie Review

- **Description**: a set of reviews from IMDB, collected, filtered and preprocessed by researchers affiliated with Stanford University. Polarity judgements are balanced in terms of positive and negative reviews and automatically inferred from star-based ratings, so that 7 or more is positive, while 4 or less is considered negative. The dataset was collected to provide a large benchmark for sentiment analysis algorithms.
- **Affiliation of creators**: Stanford University.
- **Domain**: linguistics.
- **Tasks in fairness literature**: fair sentiment analysis evaluation [504].
- **Data spec**: text.
- **Sample size**: $\sim$ 50K reviews.
- **Year**: 2011.
- **Sensitive features**: textual references to people and their demographics.
- **Link**: https://ai.stanford.edu/~amaas/data/sentiment/
- **Further info**: Maas et al. [521]

A.1.108 Last.fm

- **Description**: the Last.fm datasets were collected via the Last.fm API with the purpose of studying music consumption, discovery and recommendation on the web. Two datasets are provided: LFM1K, comprising timestamped listening habits of a limited user sample (∼1K) at song granularity, and LFM360K, containing the top 50 most played artists of a wider user population (∼360K).
- **Affiliation of creators**: Barcelona Music and Audio Technologies; Universitat Pompeu Fabra.
- **Domain**: music, information systems.
- **Tasks in fairness literature**: fair ranking evaluation [231].
- **Data spec**: user-song pairs (LFM1K); user-artist pairs (LFM360K).
- **Sample size**: ∼19M timestamped records of ∼1K users playing songs from ∼170K artists (LFM1K); ∼ 20M play counts (user-artist pairs) for ∼400K users over ∼300K artists (LFM360K).
- **Year**: 2010.
- **Sensitive features**: user age, gender, geography; artist.
- **Link**: http://ocelma.net/MusicRecommendationDataset/
- **Further info**: Celma [128]

A.1.109 Latin Newspapers

- **Description**: this dataset was built to study gender bias in language models and their connection with the corpora they have been trained on. It was built crawling articles from the websites of three newspapers from Chile, Peru, and Mexico. More detailed information about this resource seems to be missing.
- **Affiliation of creators**: Capital One.
- **Domain**: news.
- **Tasks in fairness literature**: data bias evaluation [273].
- **Data spec**: text.
- **Sample size**: ∼ 60K articles.
- **Year**: 2019.
- **Sensitive features**: textual references to people and their demographics.
- **Link**: not available
- **Further info**: Florez [273]

A.1.110 Law School

- **Description**: This dataset was collected to study performance in law school and bar examination of minority examinees in connection with affirmative action programs established after 1967 and subsequent anecdotal reports suggesting low bar passage rates for black examinees. Students, law schools, and state boards of bar examiners contributed to this dataset. The study tracks students who entered law school in fall 1991 through three or more years of law school and up to five administrations of the bar examination. Variables include demographics of candidates (e.g. age, race, sex), their academic performance (undergraduate GPA, law school admission test, and GPA), personal condition (e.g. financial responsibility for others during law school) along with information about law schools and bar exams (e.g. geographical area where it was taken). The associated task in machine learning is prediction of passage of the bar exam.
- **Affiliation of creators**: Law School Admission Council (LSAC).
- **Domain**: education.

- **Tasks in fairness literature**: fair classification [4, 64, 149, 687, 841], rich-subgroup fairness evaluation [435], fair classification under unawareness [478, 480], fairness evaluation [78, 476], fair regression [5, 158, 159, 460], fair representation learning [686], robust fair classification [529], limited-label fair classification [810].
- **Data spec**: tabular data.
- **Sample size**: $\sim$ 20K examinees.
- **Year**: 1998.
- **Sensitive features**: sex, race, age.
- **Link**: not available
- **Further info**: Wightman et al. [823]

A.1.111 Libimseti

- **Description**: this dataset was collected to explore the effectiveness of recommendations in online dating services based on collaborative filtering. It was collected in collaboration with employees of the dating platform libimseti.cz, one of the largest Czech dating websites at the time. The data consists of anonymous ratings provided by (and to) users of the web service on a 10-point scale.
- **Affiliation of creators**: Charles University in Prague; Libimseti.
- **Domain**: sociology, information systems.
- **Tasks in fairness literature**: fair matching [773].
- **Data spec**: user-user pairs.
- **Sample size**: $\sim$10M ratings over $\sim$200K users.
- **Year**: 2007.
- **Sensitive features**: gender.
- **Link**: http://colfi.wz.cz/
- **Further info**: Brozovsky and Petricek [96], Brožovský [97]

A.1.112 Los Angeles City Attorney's Office Records

- **Description**: this dataset was extracted from the Los Angeles City Attorney's case management system. It consists of a collection of records aimed at powering data-driven approaches to decision making and resource allocation for misdemeanour recidivism reduction via individually tailored social service interventions. Focusing on cases handled by the office between 1995–2017, the data includes information about jail bookings, charges, court appearances, outcomes, and demographics.
- **Affiliation of creators**: Los Angeles City Attorney's Office; University of Chicago.
- **Domain**: law.
- **Tasks in fairness literature**: fair classification [677].
- **Data spec**: tabular data.
- **Sample size**: $\sim$ 1M unique individuals associated with $\sim$ 2M cases.
- **Year**: 2020.
- **Sensitive features**: race, ethnicity.
- **Link**: not available
- **Further info**: [677]

A.1.113 MEPS-HC

- **Description**: the Medical Expenditure Panel Survey (MEPS) data is collected by the US Department of Health and Human Services, to survey healthcare spending and utilization by US citizens. Overall, this is a set of large-scale surveys of families and individuals, their employers, and medical providers (e.g. doctors, hospitals, pharmacies). The Household Component (HC) focuses on households and individuals, who provide information about their demographics, medical conditions and expenses, health insurance coverage, and access to care. Individuals included in a panel undergo five rounds of interviews over two years. Healthcare expenditure is often regarded as a target variable in machine learning applications, where it has been used as a proxy for healthcare utilization, with the goal of identifying patients in need.
- **Affiliation of creators**: Agency for Healthcare Research and Quality.
- **Domain**: health policy.
- **Tasks in fairness literature**: fair transfer learning [172], fair regression [680], fairness evaluation [725], robust fair classification [76], fair classification [713].
- **Data spec**: tabular data.
- **Sample size**: $\sim$ 30K, variable on a yearly basis.
- **Year**: present.
- **Sensitive features**: gender, ethnicity, age.
- **Link**: https://meps.ahrq.gov/mepsweb/data_stats/download_data_files.jsp
- **Further info**: https://www.ahrq.gov/data/meps.html

A.1.114 MGGG States

- **Description**: developed by the Metric Geometry and Gerrymandering Group[6], this dataset contains precinct-level aggregated information about demographics and political leaning of voters in each district. The data hinges on several distinct sources of data, including GIS mapping files from the US Census. Bureau[7], demographic data from IPUMS[8] and election data from MIT Election and Data Science [9]. Source and precise data format vary by state.
- **Affiliation of creators**: Tufts University.
- **Domain**: political science.
- **Tasks in fairness literature**: fair districting for electoral precincts [705].
- **Data spec**: mixture.
- **Sample size**: variable number of precincts (thousands) per state.
- **Year**: 2021.
- **Sensitive features**: race, political affiliation (representation in different precincts).
- **Link**: https://github.com/mggg-states
- **Further info**: https://mggg.org/

A.1.115 Microsoft Learning to Rank

- **Description**: this dataset was released to spur advances in learning to rank algorithms, capable of producing a list of documents in response to a text query, ranked according to their relevance for the

[6]https://mggg.org/
[7]https://www.census.gov/geographies/mapping-files.html
[8]https://www.nhgis.org/
[9]https://electionlab.mit.edu/

query. The dataset contains relevance judgements for query-document pairs, obtained "from a retired labeling set" of the Bing search engine. Over 100 numerical features are provided for each query-document pair, summarizing the salient lexical properties of the pair and the quality of the webpage, including its page rank.

- **Affiliation of creators**: Microsoft.
- **Domain**: information systems.
- **Tasks in fairness literature**: fair ranking [90].
- **Data spec**: query document pairs.
- **Sample size**: $\sim$ 30K queries.
- **Year**: 2013.
- **Sensitive features**: none.
- **Link**: https://www.microsoft.com/en-us/research/project/mslr/
- **Further info**: [651]

A.1.116 Million Playlist Dataset (MPD)

- **Description**: this dataset powered the 2018 RecSys Challenge on automatic playlist continuation. It consists of a sample of public Spotify playlists created by US Spotify users between 2010–2017. Each playlist consists of a title, track list and additional metadata. For each track, MPD provides the title, artist, album, duration and Spotify pointers. User data is anonymized. The dataset was augmented with record label information crawled from the web [455].
- **Affiliation of creators**: Spotify; Johannes Kepler University; University of Massachusetts.
- **Domain**: music, information systems.
- **Tasks in fairness literature**: data bias evaluation [455].
- **Data spec**: tabular data.
- **Sample size**: $\sim$ 1M playlists containing $\sim$ 2M unique tracks by $\sim$ 300K artists.
- **Year**: 2018.
- **Sensitive features**: artist, record label.
- **Link**: https://www.aicrowd.com/challenges/spotify-million-playlist-dataset-challenge
- **Further info**: Chen et al. [137]

A.1.117 Million Song Dataset (MSD)

- **Description**: this dataset was created as a large-scale benchmark for algorithms in the musical domain. Song data was acquired through The Echo Nest API, capturing a wide array of information about the song (duration, loudness, key, tempo, etc.) and the artist (name, id, location, etc.). In total the dataset creators retrieved one million songs, and for each song 55 fields are provided as metadata. This dataset also powers the Million Song Dataset Challenge, integrating the MSD with implicit feedback from taste profiles gather from an undisclosed set of applications.
- **Affiliation of creators**: Columbia University; The Echo Nest.
- **Domain**: music, information systems.
- **Tasks in fairness literature**: dynamical evaluation of fair ranking [264].
- **Data spec**: user-song pairs.
- **Sample size**: $\sim$ 50M play counts over $\sim$ 1M users and $\sim$ 400K songs.
- **Year**: 2012.
- **Sensitive features**: artist; geography.
- **Link**: http://millionsongdataset.com/; https://www.kaggle.com/c/msdchallenge
- **Further info**: Bertin-Mahieux et al. [65], McFee et al. [544]

A.1.118 MIMIC-CXR-JPG

- **Description**: this dataset was curated to encourage research in medical computer vision. It consists of chest x-rays sourced from the Beth Israel Deaconess Medical Center between 2011–2016. Each image is tagged with one or more of fourteen labels, derived from the corresponding free-text radiology reports via natural language processing tools. A subset of 687 report-label pairs have been validated by a board of certified radiologists with 8 years of experience.
- **Affiliation of creators**: Massachusetts Institute of Technology; Beth Israel Deaconess Medical Center; Stanford University; Harvard Medical School; National Library of Medicine.
- **Domain**: radiology.
- **Tasks in fairness literature**: fairness evaluation of private classification [145].
- **Data spec**: images.
- **Sample size**: $\sim$ 400K images of $\sim$ 70K patients.
- **Year**: 2019.
- **Sensitive features**: sex.
- **Link**: https://physionet.org/content/mimic-cxr-jpg/2.0.0/
- **Further info**: [409]

A.1.119 MIMIC-III

- **Description**: this dataset was extracted from a database of patients admitted to critical care units at the Beth Israel Deaconess Medical Center in Boston (MA), following the widespread adoption of digital health records in US hospitals. Data comprises vital signs, medications, laboratory measurements, notes and observations by care providers, fluid balance, procedure codes, diagnostic codes, imaging reports, length of stay, survival data, and demographics. The dataset spans over a decade of intensive care unit stays for adult and neonatal patients.
- **Affiliation of creators**: Massachusetts Institute of Technology; Beth Israel Deaconess Medical Center; A*STAR.
- **Domain**: critical care medicine.
- **Tasks in fairness literature**: fair classification [534], fairness evaluation [138, 865], robust fair classification [724].
- **Data spec**: mixture.
- **Sample size**: $\sim$ 60K patients.
- **Year**: 2016.
- **Sensitive features**: age, ethnicity, gender.
- **Link**: https://mimic.mit.edu/
- **Further info**: Johnson et al. [410]

A.1.120 ML Fairness Gym

- **Description**: this resource was developed to study the long-term behaviour and emergent properties of fair ML systems. It is an extension of OpenAI Gym [94], simulating the actions of agents within environments as Markov Decision Processes. As of 2021, four environments have been released. (1) *Lending* emulates the decisions of a bank, based on perceived credit-worthiness of individuals, which is distributed according to an artificial sensitive feature. (2) *Attention allocation* concentrates on agents tasked with monitoring sites for incidents. (3) *College admission* relates to sequential game theory, where agents represent colleges and environments contain students capable of strategically manipulating

their features at different costs, for instance through preparation courses. (4) *Infectious disease* models the problem of vaccine allocation and its long-term consequences on people in different demographic groups.
- **Affiliation of creators**: Google.
- **Domain**: N/A.
- **Tasks in fairness literature**: dynamical fair resource allocation [26, 187], dynamical fair classification [187].
- **Data spec**: time series.
- **Sample size**: variable.
- **Year**: 2020.
- **Sensitive features**: synthetic.
- **Link**: https://github.com/google/ml-fairness-gym
- **Further info**: D'Amour et al. [187]

A.1.121 MNIST

- **Description**: one of the most famous resources in computer vision, this dataset was created from an earlier database released by the National Institute of Standards and Technology (NIST). It consists of hand-written digits collected among high-school students and Census Bureau employees, which have to be correctly labelled by image processing systems. Several augmentations have also been used in the fairness literature, discussed at the end of this section.
- **Affiliation of creators**: AT&T Labs.
- **Domain**: computer vision.
- **Tasks in fairness literature**: fair clustering [342, 499], fair anomaly detection [864], fair classification [182], fairness evaluation [706].
- **Data spec**: image.
- **Sample size**: $\sim$ 70K images across 10 digits.
- **Year**: 1998.
- **Sensitive features**: none.
- **Link**: http://yann.lecun.com/exdb/mnist/
- **Further info**: Barocas et al. [50], Lecun et al. [488]
- **Variants**:
  - MNIST-USPS [499]: merge with USPS dataset of handwritten digits [381].
  - Color-reverse MNIST [499] or MNIST-Invert [864]: images from MNIST, reversed via $p = 255 - p$ for each pixel p.
  - Color MNIST [22]: images from MNIST colored red or green based on class label.
  - C-MNIST: images from MNIST, such that both digits and background are colored.

.

A.1.122 Mobile Money Loans

- **Description**: this dataset captures the ongoing collaboration between some banks and mobile network operators in East Africa. Phone data, including mobile money transactions, is used as "soft" financial data to create a credit score. Mobile money (bank-less) transactions represent a low-barrier tool for the financial inclusion of the poor and are fairly popular in some African countries.

- **Affiliation of creators**: unknown.
- **Domain**: finance.
- **Tasks in fairness literature**: fair transfer learning [172].
- **Data spec**: tabular data.
- **Sample size**: $\sim$ 200K people.
- **Year**: unknown.
- **Sensitive features**: age, gender.
- **Link**: not available
- **Further info**: Speakman et al. [738]

A.1.123 MovieLens

- **Description**: first released in 1998, MovieLens datasets represent user ratings from the movie recommender platform run by the GroupLens research group from the University of Minnesota. While different datasets have been released by GroupLens, in this section we concentrate on MovieLens 1M, the one predominantly used in fairness research. User-system interactions take the form of a quadruple (UserID, MovieID, Rating, Timestamp), with ratings expressed on a 1-5 star scale. The dataset also reports user demographics such as age and gender, which is voluntarily provided by the users.
- **Affiliation of creators**: University of Minnesota.
- **Domain**: information systems, movies.
- **Tasks in fairness literature**: fair ranking [103, 211, 256, 511, 737], fair ranking evaluation [231, 846, 847], fair data summarization [233], fair representation learning [612, 613], fair graph mining [89, 104], fair data generation [102].
- **Data spec**: user-movie pairs.
- **Sample size**: $\sim$ 1M reviews by $\sim$ 6K users over $\sim$ 4K movies.
- **Year**: 2003.
- **Sensitive features**: gender, age.
- **Link**: https://grouplens.org/datasets/movielens/1m/
- **Further info**: Harper and Konstan [348]

A.1.124 MS-Celeb-1M

- **Description**: this dataset was created as a large scale public benchmark for face recognition. The creators cover a wide range of countries and emphasizes diversity echoing outdated notions of race: "We cover all the major races in the world (Caucasian, Mongoloid, and Negroid)" [335]. While (in theory) containing only images of celebrities, the dataset was found to feature people who simply must maintain an online presence, and was retracted for this reason. Despite termination of the hosting website, the dataset is still searched for, available and used to build new fairness datasets, such as RFW (subsection A.1.153) and BUPT Faces (subsection A.1.27). The dataset was recently augmented with gender and nationality data automatically inferred from biographies of people [543]. From nationality, a race-related attribute was also annotated on a subset of 20,000 images.
- **Affiliation of creators**: Microsoft.
- **Domain**: computer vision.
- **Tasks in fairness literature**: fairness evaluation through artificial data generation [543].
- **Data spec**: image.
- **Sample size**: $\sim$ 10M images representing $\sim$ 100K people.
- **Year**: 2016.

- **Sensitive features**: gender, race, geography.
- **Link**: not available
- **Further info**: Guo et al. [335], McDuff et al. [543], Murgia [586]

A.1.125 MS-COCO

- **Description**: this dataset was created with the goal of improving the state of the art in object recognition. The dataset consists of over 300,000 labeled images collected from Flickr. Each image was annotated based on whether it contains one or more of the 91 object types proposed by the authors. Segmentations are also provided to indicate the region where objects are located in each image. Finally, five human-generated captions are provided for each image. Annotation, segmentation and captioning were performed by human annotators hired on Amazon Mechanical Turk. A subset of the images depicting people have been augmented with gender labels "man" and "woman" based on whether captions mention one word but not the other [360, 880].
- **Affiliation of creators**: Cornell University; Toyota Technological Institute; Facebook; Microsoft; Brown University; California Institute of Technology; University of California at Irvine.
- **Domain**: computer vision.
- **Tasks in fairness literature**: fair representation learning [191], fair classification [360].
- **Data spec**: image.
- **Sample size**: $\sim$ 300K images.
- **Year**: 2014.
- **Sensitive features**: gender.
- **Link**: https://cocodataset.org/
- **Further info**: Lin et al. [505]

A.1.126 Multi-task Facial Landmark (MTFL)

- **Description**: this dataset was developed to evaluate the effectiveness of multi-task learning in problems of facial landmark detection. The dataset builds upon an existing collection of outdoor face images sourced from the web already labelled with bounding boxes and landmarks [850], by annotating whether subjects are smiling or wearing glasses, along with their gender and pose. These annotations, whose provenance is not documented, allow researchers to define additional classification tasks for their multi-task learning pipeline.
- **Affiliation of creators**: The Chinese University of Hong Kong.
- **Domain**: computer vision.
- **Tasks in fairness literature**: fair clustering [499].
- **Data spec**: image.
- **Sample size**: $\sim$ 10K images.
- **Year**: 2014.
- **Sensitive features**: gender.
- **Link**: http://mmlab.ie.cuhk.edu.hk/projects/TCDCN.html
- **Further info**: Zhang et al. [872, 873]

A.1.127 National Longitudinal Survey of Youth

- **Description**: the National Longitudinal Surveys from the US Bureau of Labor Statistics follow the lives of representative samples of US citizens, focusing on their labor market activities and other significant life events. Subjects periodically provide responses to questions about their education, employment, housing, income, health, and more. Two different cohorts were started in 1979 (NLSY79) and (NLSY97), which have been associated with machine learning tasks of income prediction and GPA prediction respectively.
- **Affiliation of creators**: US Bureau of Labor Statistics.
- **Domain**: demography.
- **Tasks in fairness literature**: fair regression [158, 159, 460].
- **Data spec**: tabular data.
- **Sample size**: $\sim$ 10K respondents (NLSY79); $\sim$ 9K respondents (NLSY97).
- **Year**: present.
- **Sensitive features**: age, race, sex.
- **Link**: https://www.bls.gov/nls/nlsy79.htm (NLSY79); https://www.bls.gov/nls/nlsy97.htm (NLSY97)
- **Further info**:

A.1.128 National Lung Screening Trial (NLST)

- **Description**: the NLST was a randomized controlled trial aimed at understanding whether imaging through low-dose helical computed tomography reduces lung cancer mortality relative to chest radiography. Participants were recruited at 33 screening centers across the US, among subjects deemed at risk of lung cancer based on age and smoking history, and were made aware of the trial. A breadth of features about participants is available, including demographics, disease history, smoking history, family history of lung cancer, type, and results of screening exams.
- **Affiliation of creators**: National Cancer Institute's Division of Cancer Prevention, Division of Cancer Treatment and Diagnosis.
- **Domain**: radiology.
- **Tasks in fairness literature**: fair preference-based classification [783].
- **Data spec**: image.
- **Sample size**: $\sim$ 50K participants.
- **Year**: 2020.
- **Sensitive features**: age, ethnicity, race, sex.
- **Link**: https://cdas.cancer.gov/nlst/
- **Further info**: NLST Trial Research Team [598]; https://www.cancer.gov/types/lung/research/nlst

A.1.129 New York Times Annotated Corpus

- **Description**: this corpus contains nearly two million articles published in The New York Times over the period 1987–2007. For some articles, annotations by library scientists are available, including topics, mentioned entities, and summaries. The data is provided in News Industry Text Format (NITF).
- **Affiliation of creators**: The New York Times.
- **Domain**: news.
- **Tasks in fairness literature**: bias evaluation in WEs [99].
- **Data spec**: text.

- **Sample size**: $\sim$ 2M articles.
- **Year**: 2008.
- **Sensitive features**: textual references to people and their demographics.
- **Link**: https://catalog.ldc.upenn.edu/LDC2008T19
- **Further info**:

A.1.130 Nominees Corpus

- **Description**: this corpus was curated to study gender-related differences in literary production, with attention to perception of quality. It consists of fifty Dutch-language fiction novels nominated for either the AKO Literatuurprijs(shortlist) or the Libris Literatuur Prijs (longlist) in the period 2007–2012. The corpus was curated to control for nominee gender and country of origin. Word counts, LIWC counts, and metadata for this dataset are available at http://dx.doi.org/10.17632/tmp32v54ss.2.
- **Affiliation of creators**: University of Amsterdam.
- **Domain**: literature.
- **Tasks in fairness literature**: fairness evaluation [463].
- **Data spec**: text.
- **Sample size**: $\sim$ 50 novels.
- **Year**: 2017.
- **Sensitive features**: gender, geography (of author).
- **Link**: not available
- **Further info**: Koolen [462], Koolen and van Cranenburgh [463]

A.1.131 North Carolina Voters

- **Description**: US voter data is collected, curated, and maintained for multiple reasons. Data about voters in North Carolina is collected publicly as part of voter registration requirements and also privately. Private companies curating these datasets sell voter data as part of products, which include outreach lists and analytics. These datasets include voters' full names, address, demographics, and party affiliation.
- **Affiliation of creators**: North Carolina State Board of Elections.
- **Domain**: political science.
- **Tasks in fairness literature**: data bias evaluation [170], fair clustering [1], fairness evaluation of advertisement [739].
- **Data spec**: tabular data.
- **Sample size**: $\sim$ 8M voters.
- **Year**: present.
- **Sensitive features**: race, ethnicity, age, geography.
- **Link**: https://www.ncsbe.gov/results-data/voter-registration-data
- **Further info**:
- **Variants**: a privately curated version of this dataset is maintained by L2.[10].

[10]https://l2-data.com/states/north-carolina/

A.1.132 Nursery

- **Description**: this dataset encodes applications for a nursery school in Ljubljana, Slovenia. To favour transparent and objective decision-making, a computer-based decision support system was developed for the selection and ranking of applications. The target variable reported is thus the output of an expert systems based on a set of rules, taking as an input information about the family, including housing, occupation and financial status, included in the dataset. The variables were reportedly constructed in a careful manner, taking into account laws that were in force at that time and following advice given by leading experts in that field. However, the variables also appear to be coded rather subjectively. For example, the variable *social condition* admits as a value *Slightly problematic*, allegedly reserved for "When education ability of parents is low (unequal, inconsistent education, exaggerated pretentiousness or indulgence, neurotic reactions of parents), or there are improper relations in family (easier forms of parental personality disturbances, privileged or ignored children, conflicts in the family)". Given that the true map between inputs and outputs is known, this resource is mostly useful to evaluate methods of structure discovery.
- **Affiliation of creators**: University of Maribor; Jožef Stefan Institute; University of Ljubljana; Center for Public Enterprises in Developing Countries.
- **Domain**: education.
- **Tasks in fairness literature**: fair classification [680].
- **Data spec**: tabular data.
- **Sample size**: $\sim$ 10K combinations of input data (hypothetical applicants).
- **Year**: 1997.
- **Sensitive features**: family wealth.
- **Link**: https://archive.ics.uci.edu/ml/datasets/nursery
- **Further info**: Olave et al. [609]

A.1.133 NYC Taxi Trips

- **Description**: this dataset was collected through a Freedom of Information Law request from the NYC Taxi and Limousine Commission. Data points represent New York taxi trips over 4 years (2010–2013), complete with spatio-temporal data, trip duration, number of passengers, and cost. Reportedly, the dataset contains a large number of errors, including misreported trip distance, duration, and GPS coordinates. Overall, these errors account for 7% of all trips in the dataset.
- **Affiliation of creators**: University of Illinois.
- **Domain**: transportation.
- **Tasks in fairness literature**: fair matching [495, 590].
- **Data spec**: tabular data.
- **Sample size**: $\sim$ 700M taxi trips.
- **Year**: 2016.
- **Sensitive features**: none.
- **Link**: https://experts.illinois.edu/en/datasets/new-york-city-taxi-trip-data-2010-2013-2
- **Further info**: https://bit.ly/3yrT8jt
- **Variants**: a similar, smaller dataset was obtained by Chris Whong from the NYC Taxi and Limousine Commission under the Freedom of Information Law.[11].

[11] http://www.andresmh.com/nyctaxitrips/

A.1.134 Occupations in Google Images

- **Description**: this dataset was collected to study gender and skin tone diversity in image search results for jobs, and its relation with gender and race conentration in different professions. The dataset consists of the top 100 results for 96 occupations from Google Image Search, collected in December 2019. The creators hired workers on Amazon Mechanical Turk to label the gender (male, female) and Fitzpatrick skin tone (Type 1–6) of the primary person in each image, adding "Not applicable" and "Cannot determine" as possible options. Three labels were collected for each image, to which the majority label was assigned where possible.
- **Affiliation of creators**: Yale Universiy.
- **Domain**: information systems.
- **Tasks in fairness literature**: fair subset selection under unawareness [553].
- **Data spec**: image.
- **Sample size**: $\sim$ 10K images of $\sim$100 occupations.
- **Year**: 2019.
- **Sensitive features**: gender, skin tone (inferred).
- **Link**: https://drive.google.com/drive/u/0/folders/1j9I5ESc-7NRCZ-zSD0C6LHjeNp42RjkJ
- **Further info**: Celis and Keswani [125]

A.1.135 Office31

- **Description**: this dataset was curated to support domain adaptation algorithms for computer vision systems. It features images of 31 different office tools (e.g. chair, keyboard, printer) from 3 different domains: listings on Amazon, high quality camera images, low quality webcam shots.
- **Affiliation of creators**: University of California, Berkeley.
- **Domain**: computer vision.
- **Tasks in fairness literature**: fair clustering [499].
- **Data spec**: image.
- **Sample size**: $\sim$ 4K images.
- **Year**: 2011.
- **Sensitive features**: none.
- **Link**: https://paperswithcode.com/dataset/office-31
- **Further info**: Saenko et al. [689]

A.1.136 Olympic Athletes

- **Description**: this is a historical sports-related dataset on the modern Olympic Games from their first edition in 1896 to the 2016 Rio Games. The dataset was consolidated by Randi H Griffin utilizing SportsReference as the primary source of information. For each athlete, the dataset comprises demographics, height, weight, competition, and medal.
- **Affiliation of creators**: unknown.
- **Domain**: sports.
- **Tasks in fairness literature**: fair clustering [378].
- **Data spec**: tabular data.
- **Sample size**: $\sim$ 300K athletes.
- **Year**: 2018.
- **Sensitive features**: sex, age.
- **Link**: https://www.kaggle.com/heesoo37/120-years-of-olympic-history-athletes-and-results
- **Further info**: https://www.sports-reference.com/

A.1.137 Omniglot

- **Description**: this dataset was designed to study the problem of automatically learning basic visual concepts. It consists of handwritten characters from different alphabets drawn online via Amazon Mechanical Turk by 20 different people.
- **Affiliation of creators**: New York University; University of Toronto; Massachusetts Institute of Technology.
- **Domain**: computer vision.
- **Tasks in fairness literature**: fair few-shot learning [500].
- **Data spec**: image.
- **Sample size**: $\sim$ 2K images from 50 different alphabets.
- **Year**: 2019.
- **Sensitive features**: none.
- **Link**: https://github.com/brendenlake/omniglot
- **Further info**: Lake et al. [479]

A.1.138 One billion word benchmark

- **Description**: this dataset was proposed in 2014 as a benchmark for language models. The authors sourced English textual data from the EMNLP 6th workshop on Statistical Machine Translation[12], more specifically the Monolingual language model training data, comprising a news crawl from 2007–2011 and data from the European Parliament website. Preprocessing includes removal of duplicate sentences, rare words (appearing less than 3 times) and mapping out-of-vocabulary words to the <UNK> token. The ELMo contextualized WEs [635] were trained on this benchmark.
- **Affiliation of creators**: Google; University of Edinburgh; Cantab Research Ltd.
- **Domain**: linguistics.
- **Tasks in fairness literature**: data bias evaluation [754].
- **Data spec**: text.
- **Sample size**: $\sim$ 800M words.
- **Year**: 2014.
- **Sensitive features**: textual references to people and their demographics.
- **Link**: https://opensource.google/projects/lm-benchmark
- **Further info**: Chelba et al. [135]

A.1.139 Online Freelance Marketplaces

- **Description**: this dataset was created to audit racial and gender biases on TaskRabbit and Fiverr, two popular online freelancing marketplaces. The dataset was built by crawling workers' profiles from both websites, including metadata, activities, and past job reviews. Profiles were later annotated with perceived demographics (gender and race) by Amazon Mechanical Turk based on profile images. On TaskRabbit, the authors executed search queries for all task categories in the 10 largest cities where the service is available, logging workers' ranking in search results. On Fiverr, they concentrated on 9 tasks of diverse nature. The total number of queries that were issued on each platform, resulting in as many search result pages, is not explicitly stated.
- **Affiliation of creators**: Northeastern University, GESIS Leibniz Institute for the Social Sciences, University of Koblenz-Landau, ETH Zürich.

[12] http://statmt.org/wmt11/training-monolingual.tgz

- **Domain**: information systems.
- **Tasks in fairness literature**: fairness evaluation [341].
- **Data spec**: query-result pairs.
- **Sample size**: $\sim$ 10K workers (Fiverr); $\sim$ 4K (TaskRabbit).
- **Year**: 2017.
- **Sensitive features**: gender, race.
- **Link**: not available
- **Further info**: Hannák et al. [341]

A.1.140 Open Images Dataset

- **Description**: this dataset was curated to improve and measure the performance of computer vision algorithms. Images with CC-BY license were downloaded from Flickr, and further filtered to remove near-duplicates, inappropriate content, and images appearing elsewhere in the internet. Different versions of this dataset were released, progressively adding a wealth of information on these images, including labels, bounding boxes, segmentation masks, visual relationships, and localized narratives. Bounding boxes relate to 600 classes, including "person", which admits "girl", "boy", "woman", and "man" as a subclass. Image-level labels are generated automatically and verified by humans, resulting in annotations for a subset of present and absent classes (positive and negative image-level labels). Based on the positive image-level labels, spatial annotations are produced by human annotators: bounding boxes [477], visual relationships [477], and validation+test segmentations are drawn fully manually [60]; while segmentations in train are drawn using an interactive algorithm [60]. Further, independent of any other annotations, rich localized dense image captions are collected by asking humans to provide detailed free-form image descriptions while they hover the mouse over the regions they describe (Localized Narratives [643]).
- **Affiliation of creators**: Google.
- **Domain**: computer vision.
- **Tasks in fairness literature**: data bias evaluation [704], fairness evaluation [8].
- **Data spec**: image.
- **Sample size**: $\sim$ 9M images.
- **Year**: 2020.
- **Sensitive features**: gender, age.
- **Link**: https://storage.googleapis.com/openimages/web/index.html
- **Further info**: [60, 477, 643, 704]

A.1.141 Paper-Reviewer Matching

- **Description**: this dataset summarizes the peer review assignment process of 3 different conferences, namely one edition of Medical Imaging and Deep Learning (MIDL) and two editions of the Conference on Computer Vision and Pattern Recognition (called CVPR and CVPR2018). The data, provided by OpenReview and the Computer Vision Foundation, consist of a matrix of paper-reviewer affinities, a set of coverage constraints to ensure each paper is properly reviewed, and a set of upper bound constraints to avoid imposing an excessive burden on reviewers.
- **Affiliation of creators**: unknown.
- **Domain**: library and information sciences.
- **Tasks in fairness literature**: fair matching [456].
- **Data spec**: paper-reviewer pairs.

- **Sample size**: $\sim$ 200 reviewers for $\sim$ 100 papers (MIDL); $\sim$ 1K reviewers for $\sim$ 3K papers (CVPR). $\sim$ 3K reviewers for $\sim$ 5K papers (CVPR2018).
- **Year**: 2019.
- **Sensitive features**: none.
- **Link**: not available
- **Further info**: Kobren et al. [456]

A.1.142 Philadelphia Crime Incidents

- **Description**: this dataset is provided as part of OpenDataPhilly initiative. It summarizes hundreds of thousands of crime incidents handled by the Philadelphia Police Department over a period of ten years (2006–2016). The dataset comes with fine spatial and temporal granularity and has been used to monitor seasonal and historical trends and measure the effect of police strategies.
- **Affiliation of creators**: Philadelphia Police Department.
- **Domain**: law.
- **Tasks in fairness literature**: fair resource allocation [236].
- **Data spec**: tabular data.
- **Sample size**: $\sim$ 1M crime incidents.
- **Year**: present.
- **Sensitive features**: geography.
- **Link**: https://www.opendataphilly.org/dataset/crime-incidents
- **Further info**:

A.1.143 Pilot Parliaments Benchmark (PPB)

- **Description**: this dataset was developed as a benchmark with a balanced representation of gender and skin type to evaluate the performance of face analysis technology. The dataset features images of parliamentary representatives from three African countries (Rwanda, Senegal, South Africa) and three European countries (Iceland, Finland, Sweden) to achieve a good balance between skin type and gender while reducing potential harms connected with lack of consent from the people involved. Three annotators provided gender and Fitzpatrick labels. A certified surgical dermatologist provided the definitive Fitzpatrick skin type labels. Gender was annotated based on name, gendered title, and photo appearance.
- **Affiliation of creators**: Massachusetts Institute of Technology; Microsoft.
- **Domain**: computer vision.
- **Tasks in fairness literature**: fair classification [14, 447], fairness evaluation [101, 662], bias discovery [14, 447].
- **Data spec**: image.
- **Sample size**: $\sim$ 1K images of $\sim$ 1K individuals.
- **Year**: 2018.
- **Sensitive features**: gender, skin type.
- **Link**: http://gendershades.org/
- **Further info**: Buolamwini and Gebru [101]

A.1.144 Pima Indians Diabetes Dataset (PIDD)

- **Description**: this resource owes its name to the respective entry on the UCI repository (now unavailable), and was derived from a medical study of Native Americans from the Gila River Community, often called Pima. The study was initiated in the 1960s by the National Institute of Diabetes and Digestive and Kidney Diseases and found a large prevalence of *diabetes mellitus* in this population. The dataset commonly available nowadays represents a subset of the original study, focusing on women of age 21 or older. It reports whether they tested positive for diabetes, along with eight covariates that were found to be significant risk factors for this population. These include the number of pregnancies, skin thickness, and body mass index, based on which algorithms should predict the test results.
- **Affiliation of creators**: Logistics Management Institute; National Institute of Diabetes Digestive and Kidney Diseases; John Hopkins University.
- **Domain**: endocrinology.
- **Tasks in fairness literature**: fairness evaluation [714], fair clustering [142].
- **Data spec**: tabular data.
- **Sample size**: $\sim$ 800 subjects.
- **Year**: 2016.
- **Sensitive features**: age.
- **Link**: https://www.kaggle.com/uciml/pima-indians-diabetes-database
- **Further info**: Radin [656], Smith et al. [733]

A.1.145 Pokec Social Network

- **Description**: this graph dataset summarizes the networks of Pokec users, a social network service popular in Slovakia and Czech Republic. Due to default privacy settings being predefined as public, a wealth of information for each profile was collected by curators including information on demographics, politics, education, marital status, and children wherever available. This resource was collected to perform data analysis in social networks.
- **Affiliation of creators**: University of Zilina.
- **Domain**: social networks.
- **Tasks in fairness literature**: fair data summarization [233].
- **Data spec**: user-user pairs.
- **Sample size**: $\sim$ 2M nodes (profiles) connected by $\sim$ 30M edges (friendship relations).
- **Year**: 2013.
- **Sensitive features**: gender, geography, age.
- **Link**: https://snap.stanford.edu/data/soc-pokec.html
- **Further info**: Takac and Zabovsky [753]

A.1.146 Popular Baby Names

- **Description**: this dataset summarizes birth registration in New York City, focusing on names sex and race of newborns, providing a reliable source of data to assess naming trends in New York. A similar nation-wide database is maintained by the US Social Security Administration.
- **Affiliation of creators**: City of New York, Department of Health and Mental Hygiene (NYC names); United States Social Security Administration (US names).
- **Domain**: linguistics.
- **Tasks in fairness literature**: fair sentiment analysis [583, 852], bias discovery in WEs [752].
- **Data spec**: tabular data.

- **Sample size**: $\sim$ 3K unique names (NYC names); $\sim$ 30K unique names (US names).
- **Year**: 2021.
- **Sensitive features**: sex, race.
- **Link**: https://catalog.data.gov/dataset/popular-baby-names (NYC names); https://www.ssa.gov/oact/babynames/limits.html (US names)
- **Further info**:

A.1.147 Poverty in Colombia

- **Description**: this dataset stems from an official survey of households performed yearly by the Colombian national statistics department (Departamento Administrativo Nacional de Estadística). The survey is aimed at soliciting information about employment, income, and demographics. The data serves as an input for studies on poverty in Colombia.
- **Affiliation of creators**: Departamento Administrativo Nacional de Estadística.
- **Domain**: economics.
- **Tasks in fairness literature**: fair classification [601].
- **Data spec**: tabular data.
- **Sample size**: unknown.
- **Year**: 2018.
- **Sensitive features**: age, sex, geography.
- **Link**: https://www.dane.gov.co/index.php/estadisticas-por-tema/pobreza-y-condiciones-de-vida/pobreza-y-desigualdad/pobreza-monetaria-y-multidimensional-en-colombia-2018
- **Further info**: https://www.dane.gov.co/files/investigaciones/condiciones_vida/pobreza/2018/bt_pobreza_monetaria_18.pdf

A.1.148 PP-Pathways

- **Description**: this dataset represents a network of physical interactions between proteins that are experimentally documented in humans. The dataset was assembled to study the problem of automated discovery of the proteins (nodes) associated with a given disease. Starting from a few known disease-associated proteins and a a map of protein-protein interactions (edges), the task is to find the full list of proteins associated with said disease.
- **Affiliation of creators**: Stanford University; Chan Zuckerberg Biohub.
- **Domain**: biology.
- **Tasks in fairness literature**: fair graph mining [425].
- **Data spec**: protein-protein pairs.
- **Sample size**: $\sim$ 20K proteins (nodes) linked by $\sim$ 300K physical interactions.
- **Year**: 2018.
- **Sensitive features**: none.
- **Link**: http://snap.stanford.edu/biodata/datasets/10000/10000-PP-Pathways.html
- **Further info**: Agrawal et al. [6]

A.1.149 Prosper Loans Network

- **Description**: this dataset represents transactions on the Prosper marketplace, a famous peer-to-peer lending service where US-based users can register as lenders or borrowers. This resource has a graph structure and covers the period 2005–2011. Loan records include user ids, timestamps, loan amount,

and rate. The dataset was first associated with a study of arbitrage and its profitability in a peer-to-peer lending system.
- **Affiliation of creators**: Prosper; University College Dublin.
- **Domain**: finance.
- **Tasks in fairness literature**: fair classification [501].
- **Data spec**: lender-borrower pairs.
- **Sample size**: $\sim$ 3M loan records involving $\sim$ 100K people.
- **Year**: 2015.
- **Sensitive features**: none.
- **Link**: http://mlg.ucd.ie/datasets/prosper.html
- **Further info**: Redmond and Cunningham [670]

A.1.150 PubMed Diabetes Papers

- **Description**: this dataset was created to study the problem of classification of connected entities via active learning. The creators extracted a set of articles related to diabetes from PubMed, along with their citation network. The task associated with the dataset is inferring a label specifying the type of diabetes addressed in each publication. For this task, TF/IDF-weighted term frequencies of every article are available.
- **Affiliation of creators**: University of Maryland.
- **Domain**: library and information sciences.
- **Tasks in fairness literature**: fair graph mining [498].
- **Data spec**: article-article pairs.
- **Sample size**: $\sim$ 20K articles connected by $\sim$ 40K citations.
- **Year**: 2020.
- **Sensitive features**: none.
- **Link**: https://linqs.soe.ucsc.edu/data
- **Further info**: Namata et al. [588]

A.1.151 Pymetrics Bias Group

- **Description**: Pymetrics is a company that offers a candidate screening tool to employers. Candidates play a core set of twelve games, derived from psychological studies. The resulting gamified psychological measurements are exploited to build predictive models for hiring, where positive examples are provided by high-performing employees from the employer. Pymetrics staff maintain a *Pymetrics Bias Group* dataset for internal fairness audits by asking players to fill in an optional demographic survey after they complete the games.
- **Affiliation of creators**: Pymetrics.
- **Domain**: information systems, management information systems.
- **Tasks in fairness literature**: fairness evaluation [828].
- **Data spec**: tabular data.
- **Sample size**: $\sim$ 10K users.
- **Year**: 2021.
- **Sensitive features**: gender, race.
- **Link**: not available
- **Further info**: Wilson et al. [828]

A.1.152 Race on Twitter

- **Description**: this dataset was collected to power applications of user-level race prediction on Twitter. Twitter users were hired through Qualtrics, were they filled in a survey providing their Twitter handle and demographics, including race, gender, age, education, and income. The dataset creators downloaded the most recent 3,200 tweets by the users who provided their handle. The data, allegedly released in an anonymized and aggregated format, appears to be unavailable.
- **Affiliation of creators**: University of Pennsylvania.
- **Domain**: social media.
- **Tasks in fairness literature**: fairness evaluation [38].
- **Data spec**: text.
- **Sample size**: $\sim$ 5M tweets from $\sim$ 4K users.
- **Year**: 2018.
- **Sensitive features**: race, gender, age.
- **Link**: http://www.preotiuc.ro/
- **Further info**: Preoţiuc-Pietro and Ungar [645]

A.1.153 Racial Faces in the Wild (RFW)

- **Description**: this dataset was developed as a benchmark for face verification algorithms operating on diverse populations. The dataset comprises 4 clusters of images extracted from MS-Celeb-1M (subsection A.1.124), a dataset that was discontinued by Microsoft due to privacy violations. Clusters are of similar size and contain individuals labelled Caucasian, Asian, Indian and African. Half of the labels (Asian, Indian) are derived from the "Nationality attribute of FreeBase celebrities"; the remaining half (Caucasian, African) is automatically estimated via the Face++ API. This attribute is referred to as "race" by the authors, who also assert "carefully and manually" cleaning every image. Clusters feature multiple images of each individual to allow for face verification applications.
- **Affiliation of creators**: Beijing University of Posts; Telecommunications and Canon Information Technology (Beijing).
- **Domain**: computer vision.
- **Tasks in fairness literature**: fair reinforcement learning [812], fair representation learning [315].
- **Data spec**: image.
- **Sample size**: $\sim$ 50K images of $\sim$ 10K individuals.
- **Year**: 2019.
- **Sensitive features**: race (inferrred).
- **Link**: http://www.whdeng.cn/RFW/testing.html
- **Further info**: Wang et al. [813]

A.1.154 Real-Time Crime Forecasting Challenge

- **Description**: this dataset was assembled and released by the US National Institute of Justice in 2017 with the goal of advancing the state of automated crime forecasting. It consists of calls-for-service (CFS) records provided by the Portland Police Bureau for the period 2012–2017. Each CFS record contains spatio-temporal data and crime-related categories. The dataset was released as part of a challenge with a toal prize of 1,200,000$.
- **Affiliation of creators**: National Institute of Justice.
- **Domain**: law.

- **Tasks in fairness literature**: fair spatio-temporal process learning [711].
- **Data spec**: tabular data.
- **Sample size**: $\sim$ 700K CFS records.
- **Year**: 2017.
- **Sensitive features**: geography.
- **Link**: https://nij.ojp.gov/funding/real-time-crime-forecasting-challenge-posting#data
- **Further info**: Team Conduent Public Safety Solutions [763]

A.1.155 Recidivism of Felons on Probation

- **Description**: this dataset covers probation cases of persons who were sentenced in 1986 in 32 urban and suburban US jurisdictions. It was assembled to study the behaviour of individuals on probation and their compliance with court orders across states. Possible outcomes include successful discharge, new felony rearrest, and absconding. The information on probation cases was frequently obtained through manual reviews and transcription of probation files, mostly by college students. Variables include probationer's demographics, educational level, wage, history of convictions, disciplinary hearings and probation sentences. The final dataset consists of $\sim$ 10K probation cases "representative of 79,043 probationers".
- **Affiliation of creators**: US Department of Justice; National Association of Criminal Justice Planners.
- **Domain**: law.
- **Tasks in fairness literature**: limited-label fair classification [815].
- **Data spec**: tabular data.
- **Sample size**: $\sim$ 10K probation cases.
- **Year**: 2005.
- **Sensitive features**: sex, race, ethnicity, age.
- **Link**: https://www.icpsr.umich.edu/web/NACJD/studies/9574
- **Further info**: https://bjs.ojp.gov/data-collection/recidivism-survey-felons-probation

A.1.156 Reddit Comments

- **Description**: this resource consists of Reddit comments and relative metadata, crawled and made available online for research purposes. While the available dumps cover the period 2006-2021, below the "sample size" field refers to comments from 2014 used in one surveyed work.
- **Affiliation of creators**: Pushshift data.
- **Domain**: social media, linguistics.
- **Tasks in fairness literature**: bias evaluation in language models [334].
- **Data spec**: text.
- **Sample size**: $\sim$ 500M comments.
- **Year**: 2021.
- **Sensitive features**: textual references to people and their demographics.
- **Link**: https://files.pushshift.io/reddit/comments/
- **Further info**: Guo and Caliskan [334]

A.1.157 Renal Failure

- **Description**: the dataset was created to compare the performance of two different algorithms for automated renal failure risk assessment. Considering patients who received care at NYU Langone Medical Center, each entry encodes their health records, demographics, disease history, and lab results.

The final version of the dataset has a cutoff date, considering only patients who did not have kidney failure by that time, and reporting, as a target ground truth, whether they proceeded to have kidney failure within the next year.

- **Affiliation of creators**: New York University; New York University Langone Medical Center.
- **Domain**: nephrology.
- **Tasks in fairness literature**: fairness evaluation [826].
- **Data spec**: tabular data.
- **Sample size**: $\sim$ 2M patients.
- **Year**: 2019.
- **Sensitive features**: age, gender, race.
- **Link**: not available
- **Further info**: Williams and Razavian [826]

A.1.158 Reuters 50 50

- **Description**: this dataset was extracted from the Reuters Corpus Volume 1 (RCV1), a large corpus of newswire stories, to study the problem of authorship attribution. The 50 most prolific authors were selected from RCV1, considering only texts labeled corporate/industrial. The dataset consists of short news stories from these authors, labelled with the name of the author.
- **Affiliation of creators**: University of the Aegean.
- **Domain**: news.
- **Tasks in fairness literature**: fair clustering [343].
- **Data spec**: text.
- **Sample size**: $\sim$ 5K articles.
- **Year**: 2011.
- **Sensitive features**: author, textual references to people and their demographics.
- **Link**: http://archive.ics.uci.edu/ml/datasets/Reuter_50_50
- **Further info**: Houvardas and Stamatatos [372]

A.1.159 Ricci

- **Description**: this dataset relates to the US supreme court labor case on discrimination *Ricci vs DeStefano* (2009), connected with the disparate impact doctrine. It represents 118 firefighter promotion tests, providing the scores and race of each test taker. Eighteen firefighters from the New Haven Fire Department claimed "reverse discrimination" after the city refused to certify a promotion examination where they had obtained high scores. The reasons why city officials avoided certifying the examination included concerns of potential violation of the 'four-fiths' rule, as, given the vacancies at the time, no black firefighter would be promoted. The dataset was published and popularized by Weiwen Miao for pedagogical use.
- **Affiliation of creators**: Haverford College.
- **Domain**: law.
- **Tasks in fairness literature**: fairness evaluation [261, 282], limited-label fairness evaluation [403].
- **Data spec**: tabular data.
- **Sample size**: $\sim$ 100 test takers.
- **Year**: 2018.
- **Sensitive features**: race.
- **Link**: http://jse.amstat.org/jse_data_archive.htm; https://github.com/algofairness/fairness-comparison/tree/master/fairness/data/raw
- **Further info**: Gastwirth and Miao [290], Miao [563]

A.1.160 Rice Facebook Network

- **Description**: this dataset repesents the Facebook sub-network of students and alumni of Rice University. It consists of a crawl of reachable profiles in the Rice Facebook network, augmented with academic information obtained from Rice University directories. This collection was created to study the problem of inferring unknown attributes in a social network based on the network graph and attributes that are available for a fraction of users.
- **Affiliation of creators**: MPI-SWS; Rice University; Northeastern University.
- **Domain**: social networks.
- **Tasks in fairness literature**: fair graph diffusion [11].
- **Data spec**: user-user pairs.
- **Sample size**: $\sim$ 1K profiles connected by 40K edges.
- **Year**: 2010.
- **Sensitive features**: none.
- **Link**: not available
- **Further info**: Mislove et al. [570]

A.1.161 Riddle of Literary Quality

- **Description**: this text corpus was assembled to study the factors that correlate with the acceptance of a text as literary (or non-literary) and good (or bad). It consists of 401 Dutch-language novels published between 2007–2012. These works were selected for being bestsellers or often lent from libraries in the period 2009–2012. Due to copyright reasons, the data is not publicly available.
- **Affiliation of creators**: Huygens ING – KNAW; University of Amsterdam; Fryske Akademy.
- **Domain**: literature.
- **Tasks in fairness literature**: fairness evaluation [463].
- **Data spec**: text.
- **Sample size**: $\sim$ 400 novels.
- **Year**: 2017.
- **Sensitive features**: gender (of author).
- **Link**: not available
- **Further info**: Koolen and van Cranenburgh [463]; https://literaryquality.huygens.knaw.nl/

A.1.162 Ride-hailing App

- **Description**: this dataset was gathered from a ride-hailing app operating in an undisclosed major Asian city. It summarizes spatio-temporal data about ride requests (jobs) and assignments to drivers during 29 consecutive days. The data tracks the position and status of taxis logging data every 30-90 seconds.
- **Affiliation of creators**: Max Planck Institute for Software Systems; Max Planck Institute for Informatics.
- **Domain**: transportation.
- **Tasks in fairness literature**: fair matching [747].
- **Data spec**: driver-job pairs.
- **Sample size**: $\sim$ 1K drivers handling $\sim$ 200K job requests.
- **Year**: 2019.
- **Sensitive features**: geography.
- **Link**: not available
- **Further info**: Sühr et al. [747]

A.1.163 RtGender

- **Description**: this dataset captures differences in online commenting behaviour to posts and videos of female and male users. It was created by collecting posts and top-level comments from four platforms: Facebook, Reddit, Fitocracy, TED talks. For each of the four sources, the possibility to reliably report the gender of the poster or presenter shaped the data collection procedure. Authors of posts and videos were selected among users self-reporting their gender or public figures for which gender annotations were available. For instance, the authors created two Facebook-based datasets: one containing all posts and associated top-level comments for all 412 members of US parliament who have public Facebook pages, and a similar one for 105 American public figures (journalists, novelists, actors, actresses, etc.). The gender of these figures was derived based on their presence on Wikipedia category pages relevant for gender.[13] The gender of commenters and a reliable ID to identify them across comments may be useful for some analyses. The authors report commenters' first names and a randomized ID, which should support these goals, while reducing chances of re-identification based on last name and Facebook ID.
- **Affiliation of creators**: Stanford University; University of Michigan; Carnegie Mellon University.
- **Domain**: social media, linguistics.
- **Tasks in fairness literature**: fairness evaluation [29].
- **Data spec**: text.
- **Sample size**: $\sim$ 2M posts with $\sim$ 25M comments.
- **Year**: 2018.
- **Sensitive features**: gender.[14].
- **Link**: https://nlp.stanford.edu/robvoigt/rtgender/
- **Further info**: [797]

A.1.164 SafeGraph Research Release

- **Description**: this dataset captures mobility patterns in the US and Canada. It is maintained by SafeGraph, a data company powering analytics about access to Points-of-Interest (POI) and mobility, including pandemic research. SafeGraph data is sourced from millions of mobile devices, whose users allow location tracking by some apps. The *Research Release* dataset consists of aggregated estimates of hourly visit counts to over 6 million POI. Given the increasing importance of SafeGraph data, directly influencing not only private initiative but also public policy, audits of data representativeness are being carried out both internally [741] and externally [170].
- **Affiliation of creators**: Safegraph.
- **Domain**: urban studies.
- **Tasks in fairness literature**: data bias evaluation [170].
- **Data spec**: mixture.
- **Sample size**: $\sim$ 7M POI.
- **Year**: present.
- **Sensitive features**: geography.
- **Link**: https://www.safegraph.com/academics
- **Further info**: https://docs.safegraph.com/v4.0/docs

[13] e.g. https://en.wikipedia.org/wiki/Category:American_female_tennis_players

[14] Annotations for Facebook and TED come from Wikipedia and Mirkin et al. [567] respectively. Reddit and Fitocracy rely on self-reported labels.

A.1.165 Scientist+Painter

- **Description**: this resource was crawled to study the problem of fair and diverse representation in subsets of instances selected from a large dataset, with a focus on gender concentration in professions. The dataset consists of approximately 800 images that equally represent male scientists, female scientists, male painters, and female painters. These images were gathered from Google image search, selecting the top 200 medium sized JPEG files that passed the strictest level of Safe Search filtering. Then, each image was processed to obtain sets of 128-dimensional SIFT descriptors. The descriptors are combined, subsampled and then clustered using k-means into 256 clusters.
- **Affiliation of creators**: École Polytechnique Fédérale de Lausanne (EPFL); Microsoft; University of California, Berkeley.
- **Domain**: information systems.
- **Tasks in fairness literature**: fair data summarization [120, 123].
- **Data spec**: image.
- **Sample size**: $\sim$ 800 images.
- **Year**: 2016.
- **Sensitive features**: male/female.
- **Link**: goo.gl/hNukfP
- **Further info**: Celis et al. [123]

A.1.166 Section 203 determinations

- **Description**: this dataset is created in support of the language minority provisions of the Voting Rights Act, Section 203. The data contains information about limited-English proficient voting population by jurisdiction, which is used to determine whether election materials must be printed in minority languages. For each combination of language protected by Section 203 and US jurisdiction, the dataset provides information about total population, population of voting age, US citizen population of voting age, combining this information with language spoken at home and overall English proficiency.
- **Affiliation of creators**: US Census Bureau.
- **Domain**: demography.
- **Tasks in fairness literature**: fairness evaluation of private resource allocation [649].
- **Data spec**: tabular data.
- **Sample size**: $\sim$ 600K combinations of jurisdictions and languages potentially spoken therein.
- **Year**: 2017.
- **Sensitive features**: geography, language.
- **Link**: https://www.census.gov/data/datasets/2016/dec/rdo/section-203-determinations.html
- **Further info**: https://www.census.gov/programs-surveys/decennial-census/about/voting-rights/voting-rights-determination-file.2016.html

A.1.167 Sentiment140

- **Description**: this dataset was created to study the problem of sentiment analysis in social media, envisioning applications of product quality and brand reputation analysis via Twitter monitoring. The sentiment of tweets, retrieved via Twitter API, is automatically inferred based on the presence of emoticons conveying joy or sadness. This dataset is part of the LEAF benchmark for federated learning. In federated learning settings, devices correspond to accounts.

- **Affiliation of creators**: Stanford University.
- **Domain**: social media.
- **Tasks in fairness literature**: fair federated learning [500].
- **Data spec**: text.
- **Sample size**: $\sim$ 2M tweets by $\sim$ 600K accounts.
- **Year**: 2012.
- **Sensitive features**: textual references to people and their demographics.
- **Link**: http://help.sentiment140.com/home
- **Further info**: Go et al. [305]

A.1.168 Seoul Bike Sharing

- **Description**: this resource, summarizing hourly public rental history of *Seoul Bikes*, was curated to study the problem of bike sharing demand prediction. The data was downloaded from the Seoul Public Data Park website of South Korea and spans one year of utilization (December 2017 to November 2018) of Seoul Bikes, a bike sharing system that started in 2015. This dataset consists of hourly information about weather (e.g. temperature, solar radiation, rainfall) and time (date, time, season, holiday), along with the number of bikes rented at each hour, which is the target of a prediction task.
- **Affiliation of creators**: Sunchon National University.
- **Domain**: transportation.
- **Tasks in fairness literature**: fair regression [208].
- **Data spec**: time series.
- **Sample size**: $\sim$ 9K hourly points.
- **Year**: 2020.
- **Sensitive features**: none.
- **Link**: https://archive.ics.uci.edu/ml/datasets/Seoul+Bike+Sharing+Demand
- **Further info**: V E and Cho [785], V E et al. [786], https://data.seoul.go.kr/index.do

A.1.169 Shakespeare

- **Description**: this dataset is available as part of the LEAF benchmark for federated learning [106]. It is built from "The Complete Works of William Shakespeare", where each speaking role represents a different device. The task envisioned for this dataset is next character prediction.
- **Affiliation of creators**: Google; Carnegie Mellon University; Determined AI.
- **Domain**: literature.
- **Tasks in fairness literature**: fair federated learning [500].
- **Data spec**: text.
- **Sample size**: $\sim$ 4M tokens over $\sim$ 1K speaking roles.
- **Year**: 2020.
- **Sensitive features**: textual references to people and their demographics.
- **Link**: https://www.tensorflow.org/federated/api_docs/python/tff/simulation/datasets/shakespeare
- **Further info**: Caldas et al. [106], McMahan et al. [548]

A.1.170 Shanghai Taxi Trajectories

- **Description**: this semi-synthetic dataset represents the road network and traffic patterns of Shanghai. Trajectories were collected from thousands of taxis operating in Shanghai. Spatio-temporal traffic patterns were extracted from these trajectories and used to build the dataset.
- **Affiliation of creators**: Shanghai Jiao Tong University; CITI-INRIA Lab.
- **Domain**: transportation.
- **Tasks in fairness literature**: fair routing [650].
- **Data spec**: unknown.
- **Sample size**: unknown.
- **Year**: 2015.
- **Sensitive features**: geography.
- **Link**: not available
- **Further info**: Qian et al. [650]

A.1.171 shapes3D

- **Description**: this dataset is an artificial benchmark for unsupervised methods aimed at learning disentangled data representations. It consists of imagesof 3D shapes in a walled environment, with variable floor colour, wall colour, object colour, scale, shape and orientation.
- **Affiliation of creators**: DeepMind; Wayve.
- **Domain**: computer vision.
- **Tasks in fairness literature**: fair representation learning [513], fair data generation [151].
- **Data spec**: image.
- **Sample size**: $\sim$ 500K images.
- **Year**: 2018.
- **Sensitive features**: none.
- **Link**: https://github.com/deepmind/3d-shapes
- **Further info**: Kim and Mnih [445]

A.1.172 SIIM-ISIC Melanoma Classification

- **Description**: this dataset was developed to advance the study of automated melanoma classification. The resource consists of dermoscopy images from six medical centers. Images in the dataset are tagged with a patient identifier, allowing lesions from the same patient to be mapped to one another. Images were queried from medical databases among patients with dermoscopy imaging from 1998 to 2019, ranging in quality from 307,200 to 24,000,000 pixels. A curated subset is employed for the 2020 ISIC Grand Challenge.[15] This dataset was annotated automatically with a binary Fitzpatrick skin tone label [145].
- **Affiliation of creators**: Memorial Sloan Kettering Cancer Center; University of Queensland; University of Athens; IBM; Universitat de Barcelona; Melanoma Institute Australia; Sydney Melanoma Diagnostic Center; Emory University; Medical University of Vienna; Mayo Clinic; SUNY Downstate Medical School; Stony brook Medical School; Rabin Medical Center; Weill Cornell Medical College.
- **Domain**: dermatology.
- **Tasks in fairness literature**: fairness evaluation of private classification [145].
- **Data spec**: image.

[15] https://www.kaggle.com/c/siim-isic-melanoma-classification

- **Sample size**: $\sim$ 30K images of $\sim$ 2K patients.
- **Year**: 2020.
- **Sensitive features**: skin type.
- **Link**: urlhttps://doi.org/10.34970/2020-ds01
- **Further info**: Rotemberg et al. [682]

A.1.173 SmallNORB

- **Description**: this dataset was assembled by researchers affiliated with New York University as a benchmark for robust object recognition under variable pose and lighting conditions. It consists of images of 50 different toys belonging to 5 categories (four-legged animals, human figures, airplanes, trucks, and cars) obtained by 2 different cameras.
- **Affiliation of creators**: New York University; NEC Labs America.
- **Domain**: computer vision.
- **Tasks in fairness literature**: fair representation learning [513].
- **Data spec**: image.
- **Sample size**: $\sim$ 100K images.
- **Year**: 2005.
- **Sensitive features**: none.
- **Link**: https://cs.nyu.edu/~ylclab/data/norb-v1.0-small/
- **Further info**: LeCun et al. [489]

A.1.174 Spliddit Divide Goods

- **Description**: this dataset summarizes instances of usage of the *divide goods* feature of Spliddit, a not-for-profit academic endeavor providing easy access to fair division methods. A typical use case for the service is inheritance division. Participants express their preferences by dividing 1,000 points between the available goods. In response, the service provides suggestions that are meant to maximize the overall satisfaction of all stakeholders.
- **Affiliation of creators**: Spliddit.
- **Domain**: economics.
- **Tasks in fairness literature**: fair preferece-based resource allocation [30].
- **Data spec**: tabular data.
- **Sample size**: $\sim$ 1K division instances.
- **Year**: 2016.
- **Sensitive features**: none.
- **Link**: not available
- **Further info**: Caragiannis et al. [113]; http://www.spliddit.org/apps/goods

A.1.175 Stanford Medicine Research Data Repository

- **Description**: this is a data lake/repository developed at Stanford University, supporting a number of data sources and access pipelines. The aim of the underlying project is favouring access to clinical data for research purposes through flexible and robust management of medical data. The data comes from Stanford Health Care, the Stanford Children's Hospital, the University Healthcare Alliance and Packard Children's Health Alliance clinics.
- **Affiliation of creators**: Stanford University.

- **Domain**: medicine.
- **Tasks in fairness literature**: fair risk assessment [637].
- **Data spec**: mixture.
- **Sample size**: ∼3M individuals.
- **Year**: present.
- **Sensitive features**: race, ethnicity, gender, age.
- **Link**: https://starr.stanford.edu/
- **Further info**: Datta et al. [190], Lowe et al. [516]

A.1.176 State Court Processing Statistics (SCPS)

- **Description**: this resource was curated as part of the SCPS program. The program tracked felony defendants from charging by the prosecutor until disposition of their cases for a maximum of 12 months (24 months for murder cases). The data represents felony cases filed in approximately 40 populous US counties in the period 1990-2009. Defendants are summarized by 106 variables summarizing demographics, arrest charges, criminal history, pretrial release and detention, adjudication, and sentencing.
- **Affiliation of creators**: US Department of Justice.
- **Domain**: law.
- **Tasks in fairness literature**: fairness evaluation of multi-stage classification [324].
- **Data spec**: tabular data.
- **Sample size**: $\sim$ 200K defendants.
- **Year**: 2014.
- **Sensitive features**: gender, race, age, geography.
- **Link**: https://www.icpsr.umich.edu/web/NACJD/studies/2038/datadocumentation
- **Further info**: https://bjs.ojp.gov/data-collection/state-court-processing-statistics-scps

A.1.177 Steemit

- **Description**: this resource was collected to test novel approaches for personalized content recommendation in social networks. It consists of two separate datasets summarizing interactions in the Spanish subnetwork and the English subsnetwork of Steemit, a blockchain-based social media website. The datasets summarize user-post interactions in a binary fashion, using comments as a proxy for positive engagement. The datasets cover a whole year of commenting activities over the period 2017–2018 and comprise the text of posts.
- **Affiliation of creators**: Hong Kong University of Science and Technology; WeBank.
- **Domain**: social media.
- **Tasks in fairness literature**: fairness evaluation [835].
- **Data spec**: user-post pairs.
- **Sample size**: $\sim$ 50K users interacting over $\sim$ 200K posts.
- **Year**: 2019.
- **Sensitive features**: textual references to people and their demographics.
- **Link**: https://github.com/HKUST-KnowComp/Social-Explorative-Attention-Networks
- **Further info**: Xiao et al. [835]

A.1.178 Stop, Question and Frisk

- **Description**: Stop, Question and Frisk (SQF) is an expression that commonly refers to a New York City policing program under which officers can briefly detain, question, and search a citizen if the officer has a reasonable suspicion of criminal activity. Concerns about race-based disparities in this practice have been expressed multiple times, especially in connection with the subjective nature of "reasonable suspicion" and the fact that being in a "high-crime area" lawfully lowers the bar of want may constitute reasonable suspicion. The NYPD has a policy of keeping track of most stops, recording them in UF-250 forms which are maintained centrally and distributed by the NYPD. The form includes several information such as place and time of a stop, the duration of the stop and its outcome along with data on demographics and physical appearance of the suspect. Currently available data pertains to years 2003–2020.
- **Affiliation of creators**: New York Police Department.
- **Domain**: law.
- **Tasks in fairness literature**: preference-based fair classification [855], robust fair classification [419], fair classification under unawareness [441], fairness evaluation [309], fair classification [12].
- **Data spec**: tabular data.
- **Sample size**: $\sim$ 1M records.
- **Year**: 2021.
- **Sensitive features**: race, age, sex, geography.
- **Link**: https://www1.nyc.gov/site/nypd/stats/reports-analysis/stopfrisk.page
- **Further info**: Gelman et al. [294], Goel et al. [310]

A.1.179 Strategic Subject List

- **Description**: this dataset was funded through a Bureau of Justice Assistance grant and leveraged by the Illinois Institute of Technology to develop the Chicago Police Department's Strategic Subject Algorithm. The algorithm provides a risk score which reflects an individual's probability of being involved in a shooting incident either as a victim or an offender. For each individual, the dataset provides information about the circumstances of their arrest, their demographics and criminal history. The dataset covers arrest data from the period 2012–2016; the associated program was discontinued in 2019.
- **Affiliation of creators**: Chicago Police Department; Illinois Institute of Technology.
- **Domain**: law.
- **Tasks in fairness literature**: fairness evaluation [78].
- **Data spec**: tabular data.
- **Sample size**: $\sim$ 400K individuals.
- **Year**: 2020.
- **Sensitive features**: ace, sex, age.
- **Link**: https://data.cityofchicago.org/Public-Safety/Strategic-Subject-List-Historical/4aki-r3np
- **Further info**: Hollywood et al. [367]

A.1.180 Student

- **Description**: the data was collected from two Portuguese public secondary schools in the Alentejo region, to investigate student achievement prediction and identify decisive factors in student success. The data tracks student performance in Mathematics and Portuguese through school year 2005-2006

and is complemented by demographic, socio-econonomical, and personal data obtained through a questionnaire. Numerical grades (20-point scale) collected by students over three terms are typically the target of the associated prediction task.
- **Affiliation of creators**: University of Minho.
- **Domain**: education.
- **Tasks in fairness literature**: fair regression [158, 159, 359], rich-subgroup fairness evaluation [435], fair data summarization [56, 413].
- **Data spec**: tabular data.
- **Sample size**: $\sim$ 600 students.
- **Year**: 2014.
- **Sensitive features**: sex, age.
- **Link**: https://archive.ics.uci.edu/ml/datasets/student+performance
- **Further info**: Cortez and Silva [167]

A.1.181 Sushi

- **Description**: this dataset was sourced online via a commercial survey service to evaluate rank-based approaches to solicit preferences and provide recommendations. The dataset captures the preferences for different types of sushi held by people in different areas of Japan. These are encoded both as ratings in a 5-point scale and ordered lists of preferences, which recommenders should learn via collaborative filtering. Demographic data was also collected to study geographical preference patterns.
- **Affiliation of creators**: Japanese National Institute of Advanced Industrial Science and Technology (AIST).
- **Domain**: .
- **Tasks in fairness literature**: fair data summarization [147].
- **Data spec**: user-sushi pairs.
- **Sample size**: $\sim$ 5K respondents.
- **Year**: 2016.
- **Sensitive features**: gender, age, geography.
- **Link**: https://www.kamishima.net/%20sushi/
- **Further info**: Kamishima [424]

A.1.182 Symptoms in Queries

- **Description**: the purpose of this dataset is to study, using only aggregate statistics, the fairness and accuracy of a classifier that predicts whether an individual has a certain type of cancer based on their Bing search queries. The dataset does not include individual data points. It provides, for each US state, and for 18 types of cancer, the proportion of individuals who have this cancer in the state according to CDC 2019 data,[16] and the proportion of individuals who are predicted to have this cancer according to the classifier that was calculated using Bing queries.
- **Affiliation of creators**: Microsoft; Ben-Gurion University of the Negev.
- **Domain**: information systems, public health.
- **Tasks in fairness literature**: limited-label fairness evaluation [688].
- **Data spec**: tabular data.
- **Sample size**: statistics for $\sim$ 20 cancer types across $\sim$ 50 US states.
- **Year**: 2020.

[16] https://gis.cdc.gov/Cancer/USCS/DataViz.html

- **Sensitive features**: geography.
- **Link**: https://github.com/sivansabato/bfa/blob/master/cancer_data.m
- **Further info**: Sabato and Yom-Tov [688]

A.1.183 TAPER Twitter Lists

- **Description**: this resource was collected to study the problem of personalized expert recommendation, leveraging Twitter lists where users labelled other users as relevant for (or expert in) a given topic. The creators started from a seed dataset of over 12 million geo-tagged Twitter lists, which they filtered to only keep US-based users in topics: news, music, technology, celebrities, sports, business, politics, food, fashion, art, science, education, marketing, movie, photography, and health. A subset of this dataset was annotated with user race (whites and non-whites) via Face++ [889].
- **Affiliation of creators**: Texas A&M University.
- **Domain**: social media.
- **Tasks in fairness literature**: fair ranking [889].
- **Data spec**: user-topic pairs.
- **Sample size**: $\sim$ 10K Twitter lists featuring $\sim$ 8K list members.
- **Year**: 2016.
- **Sensitive features**: race.
- **Link**: not available
- **Further info**: Ge et al. [291]

A.1.184 TaskRabbit

- **Description**: this resource was assembled to study the effectiveness of fair ranking approaches in improving outcomes for protected groups in online hiring. It consists of the top 10 results returned by the online freelance marketplace TaskRabbit for three queries: "Shopping", "Event staffing", and "Moving Assistace". The geographic location for a query was especially selected to yield a ranking with 3 female candidates among the top 10, with most of them appearing in the bottom 5, which may be a motivating condition for a fairness intervention. Candidates' gender was manually labelled by creators based on pronoun usage and profile pictures. For each profile, the authors extracted information on job suitability, including TaskRabbit relevance scores, number of completed tasks and positive reviews.
- **Affiliation of creators**: Technische Universität Berlin; Harvard University.
- **Domain**: information systems.
- **Tasks in fairness literature**: fair ranking evaluation [748], multi-stage fairness evaluation [748].
- **Data spec**: query-worker pairs.
- **Sample size**: 3 rankings (one per query) of $\sim$ 10 workers.
- **Year**: 2021.
- **Sensitive features**: gender.
- **Link**: not available
- **Further info**: Sühr et al. [748]

A.1.185 TIMIT

- **Description**: this resource was curated to power studies of phonetics and to evaulate systems of automated speech recognition. The dataset features speakers of different American English dialects, and includes time-aligned orthographic, phonetic and word transcriptions. Utterances are sampled at a 16kHz frequency.

- **Affiliation of creators**: University of Pennsylvania; National Institute of Standards and Technology; Massachusetts Institute of Technology; SRI International; Texas Instruments.
- **Domain**: linguistics.
- **Tasks in fairness literature**: fairness evaluation of speech recognition [706].
- **Data spec**: time series.
- **Sample size**: $\sim$ 600 speakers, each uttering $\sim$ 10 sentences.
- **Year**: 1993.
- **Sensitive features**: dialect, gender.
- **Link**: https://catalog.ldc.upenn.edu/LDC93S1
- **Further info**: https://en.wikipedia.org/wiki/TIMIT

A.1.186 Toy Dataset 1

- **Description**: this dataset consists of $\sim$ 4K points generated as follows. Binary class labels y are generated at random for each point. Next, two-dimensional features x are assigned to each point, sampling from gaussian distributions whose mean and variance depend on y, so that $p(x|y=1) = \mathcal{N}([2;2],[5,1;1,5])$; $p(x|y=-1) = \mathcal{N}([-2;-2],[10,1;1,3])$. Finally, each point's sensitive attribute z is sampled from a Bernoulli distribution so that $p(z=1) = p(x'|y=1)/(p(x'|y=1)+p(x'|y=1))$, where x' is a rotated version of x: $x' = [\cos(\phi),-\sin(\phi);\sin(\phi),\cos(\phi)]x$. Parameter ϕ controls the correlation between class label y and sensitive attribute z.
- **Affiliation of creators**: Max Planck Institute for Software Systems.
- **Domain**: N/A.
- **Tasks in fairness literature**: fair classification [678, 856], fair preference-based classification [13, 855], fair few-shot learning [728, 730], fair classification under unawareness [441].
- **Data spec**: tabular data.
- **Sample size**: $\sim$ 4K points.
- **Year**: 2017.
- **Sensitive features**: N/A.
- **Link**: https://github.com/mbilalzafar/fair-classification/tree/master/disparate_impact/synthetic_data_demo
- **Further info**: Zafar et al. [856]

A.1.187 Toy Dataset 2

- **Description**: this dataset contains synthetic relevance judgements over pairs of queries and documents that are biased against a minority group. For each query, there are 10 candidate documents, 8 from group G_0 and 2 from minority group G_1. Each document is associated with a feature vector (x_1,x_2), with both components sampled uniformly at random from the interval $(0,3)$. The relevance of documents is set to $y = x_1 + x_2$ and clipped between 0 and 5. Feature x_2 is then corrupted and replaced by zero for group G_1, leading to a biased representation between groups, such that any use of x_2 should lead to unfair rankings.
- **Affiliation of creators**: Cornell University.
- **Domain**: N/A.
- **Tasks in fairness literature**: fair ranking [90, 723].
- **Data spec**: query-document pairs.
- **Sample size**: $\sim$ 1K relevance judgements overs $\sim$ 100 queries with $\sim$ 10 candidate documents.
- **Year**: 2019.

- **Sensitive features**: N/A.
- **Link**: https://github.com/ashudeep/Fair-PGRank
- **Further info**: Singh and Joachims [723]

A.1.188 Toy Dataset 3

- **Description**: this dataset was created to demonstrate undesirable properties of a family of fair classification approaches. Each instance in the dataset is associated with a sensitive attribute z, a target variable y encoding employability, one feature that is important for the problem at hand and correlated with z (work_experience) and a second feature which is unimportant yet also correlated with z (hair_length). The data generating process is the following:

$$
\begin{aligned}
z_i &\sim \text{Bernoulli}(0.5) \\
\text{hair\_length}_i | z_i = 1 &\sim 35 \cdot \text{Beta}(2,2) \\
\text{hair\_length}_i | z_i = 0 &\sim 35 \cdot \text{Beta}(2,7) \\
\text{work\_exp}_i | z_i &\sim \text{Poisson}(25 + 6z_i) - \text{Normal}(20, 0.2) \\
y_i | \text{work\_exp}_i &\sim 2 \cdot \text{Bernoulli}(p_i) - 1, \\
\text{where} p_i &= 1/(1 + \exp[-(-25.5 + 2.5\text{work\_exp})])
\end{aligned}
$$

.

- **Affiliation of creators**: Carnegie Mellon University; University of California, San Diego.
- **Domain**: N/A.
- **Tasks in fairness literature**: fairness evaluation [78, 506].
- **Data spec**: tabular data.
- **Sample size**: $\sim$ 2K points.
- **Year**: 2018.
- **Sensitive features**: N/A.
- **Link**: not available
- **Further info**: Lipton et al. [506]

A.1.189 Toy Dataset 4

- **Description**: in this toy example, features are generated according to four 2-dimensional isotropic Gaussian distributions with different mean μ and variance σ^2. Each of the four distributions corresponds to a different combination of binary label y and protected attribute s as follows: (1) $s = a, y = +1 : \mu = (-1,-1), \sigma^2 = 0.8$; (2) $s = a, y = -1 : \mu = (1,1), \sigma^2 = 0.8$; (3) $s = b; y = +1 : \mu = (0.5,-0.5), \sigma^2 = 0.5$; (4) $s = b, y = -1 : \mu = (0.5,0.5), \sigma^2 = 0.5$.
- **Affiliation of creators**: Istituto Italiano di Tecnologia; University of Genoa; University of Waterloo; University College London.
- **Domain**: N/A.
- **Tasks in fairness literature**: fair classification [216], fairness evaluation [827].
- **Data spec**: tabular data.
- **Sample size**: $\sim$ 6K points.

- **Year**: 2018.
- **Sensitive features**: N/A.
- **Link**: https://github.com/jmikko/fair_ERM
- **Further info**: Donini et al. [216]

A.1.190 TREC Robust04

- **Description**: this classic information retrieval collection is a set of topics, documents and relevance judgements collected as part of the Text REtrieval Conference (TREC) 2004 Robust Retrieval Track to catalyze research improving the consistency of information retrieval technology. Documents are taken from articles published during the 1990s in the Financial Times Limited, the Federal Register, the Foreign Broadcast Information Service, and the Los Angeles Times. Graded relevance (not relevant, relevant, highly relevant) was judged by human assessors for a subset of all possible topic-document combinations, which were selected as "promising" by the automated systems that entered the TREC initiative. The associated task is predicting the relevance of documents for various textual queries.
- **Affiliation of creators**: National Institute of Standards and Technology.
- **Domain**: news, information systems.
- **Tasks in fairness literature**: fair ranking evaluation [297].
- **Data spec**: query-document pairs.
- **Sample size**: $\sim$ 300K relevance judgements over $\sim$ 200 queries and $\sim$ 500K documents.
- **Year**: 2005.
- **Sensitive features**: textual references to people and their demographics.
- **Link**: https://trec.nist.gov/data/t13_robust.html
- **Further info**: Voorhees [800]

A.1.191 Twitch Social Networks

- **Description**: this dataset was developed to study the effectiveness of node embeddings for learning tasks defined on graphs. This resource concentrates on Twitch content creators streaming in 6 different languages. The dataset has users as nodes, mutual friendships as edges, and node embedddings summarizing games liked, location and streaming habits. The original task on this dataset is predicting whether a streamer uses explicit language.
- **Affiliation of creators**: University of Edinburgh.
- **Domain**: social networks.
- **Tasks in fairness literature**: fair graph mining [425].
- **Data spec**: user-user pairs.
- **Sample size**: $\sim$30K nodes (users) connected by $\sim$ 400K edges (mutual friendship).
- **Year**: 2019.
- **Sensitive features**: none.
- **Link**: http://snap.stanford.edu/data/twitch-social-networks.html
- **Further info**: Rozemberczki et al. [683]

A.1.192 Twitter Abusive Behavior

- **Description**: this dataset is the result of an eight-month crowdsourced study of various forms of abusive behavior on Twitter. The authors began by considering a wide variety of inappropriate speech categories, analyzing how they are used by amateur annotators hired on CrowdFlower. After two exploratory

rounds, they merged some labels and eliminated others, converging to a final four-class categorization into (normal, spam, abusive, hateful), requiring five crowdsourced judgements per tweet. Tweets were sampled according to a boosted random sampling technique. A large part of the dataset is randomly sampled, with the addition of tweets that are likely to belong to one or more of the minority (non-normal) classes. The dataset is available as a table mapping tweet IDs to behavior category, making it possible to identify Twitter users in this dataset.

- **Affiliation of creators**: Aristotle University of Thessaloniki; Cyprus University of Technology; Telefonica; University of Alabama at Birmingham; University College London.
- **Domain**: social media.
- **Tasks in fairness literature**: fairness evaluation of harmful content detection [38].
- **Data spec**: text.
- **Sample size**: $\sim$ 100K tweets.
- **Year**: 2018.
- **Sensitive features**: textual references to people and their demographics.
- **Link**: https://github.com/ENCASEH2020/hatespeech-twitter
- **Further info**: Founta et al. [278]

A.1.193 Twitter Hate Speech Detection

- **Description**: this dataset was developed to study the problem of automated hate speech detection. The creators used the Twitter API to search for tweets containing racist and sexist terms and hashtags. The annotation was carried out by the authors, with an external review by a 25-year-old woman studying gender studies. After identifying a list of eleven criteria to identify hate speech against a minority, each tweet was labelled as sexism, racism or none. The task associated with this resource is hate speech detection. The dataset is available as a table mapping tweet IDs to hate speech category, making it possible to identify Twitter users in this dataset.
- **Affiliation of creators**: University of Copenhagen.
- **Domain**: social media.
- **Tasks in fairness literature**: fairness evaluation [38].
- **Data spec**: text.
- **Sample size**: $\sim$ 20K tweets.
- **Year**: 2016.
- **Sensitive features**: textual references to people and their demographics.
- **Link**:
- **Further info**: Waseem and Hovy [817]

A.1.194 Twitter Offensive Language

- **Description**: this dataset was developed to study the problem of automated hate speech detection, and to distinguish between hate speech and other kinds of offensive language. The creators used the Twitter API to search for tweets containing terms from a hate speech lexicon compiled by *Hatebase.org*. Workers on CrowdFlower annotated a random subset of these tweets as hate speech, offensive but not hate speech, or neither offensive nor hate speech. Workers were explicitly told that the mere presence of a slur word does not amount to hate speech. Three of more workers annotated each tweet.
- **Affiliation of creators**: Cornell University; Qatar Computing Research Institute.
- **Domain**: social media.
- **Tasks in fairness literature**: fairness evaluation [38], fair multi-stage classification [437].

- **Data spec**: text.
- **Sample size**: $\sim$ 20K tweets.
- **Year**: 2017.
- **Sensitive features**: textual references to people and their demographics.
- **Link**: https://github.com/t-davidson/hate-speech-and-offensive-language/tree/master/data
- **Further info**: Davidson et al. [193]

A.1.195 Twitter Online Harrassment

- **Description**: this dataset was developed as multidisciplinary resource to study online harrassment. The authors searched a stream of tweets for keywords likely to denote violent, offensive, threatening or hateful content based on race, gender, religion and sexual orientation. They developed coding guidelines to label a tweet as harrassing or non/harrassing and spent three weeks reviewing and refining it, annotating sample tweets as a group, and discussing the results. The curators are not publicly sharing the dataset due to Twitter terms of service restrictions and privacy concerns about individuals whose tweets are included; researchers can request access.
- **Affiliation of creators**: University of Maryland.
- **Domain**: social media.
- **Tasks in fairness literature**: fairness evaluation [38].
- **Data spec**: text.
- **Sample size**: $\sim$ 40K tweets.
- **Year**: 2017.
- **Sensitive features**: textual references to people and their demographics.
- **Link**: not available
- **Further info**: Golbeck et al. [312]

A.1.196 Twitter Political Searches

- **Description**: this dataset was collected to study political biases in Twitter search results, due to political leaning of tweets and biases in the Twitter ranking algorithm. The authors identified 25 popular political queries in December 2015, and collected relevant tweets during a week in which two presidential debates occurred, via the Twitter streaming API. Tweets were annotated based on users' political leaning. Users' leaning was automatically inferred from their topics of interest, via a classifier trained on representative sets of democratic and republican users. Both the accuracy of classifiers and the validity of user leaning as a proxy for tweet leaning was validated by workers recruited on Amazon Mechanical Turk.
- **Affiliation of creators**: Max Planck Institute for Software Systems; University of Illinois at Urbana-Champaign; Indian Institute of Engineering Science and Technology, Shibpur; Adobe Research.
- **Domain**: social media.
- **Tasks in fairness literature**: social media.
- **Data spec**: query-result pairs.
- **Sample size**: $\sim$ 30K search results containing $\sim$ 30K distinct tweets from $\sim$ 20K users.
- **Year**: 2016.
- **Sensitive features**: political leaning.
- **Link**: not available
- **Further info**: Kulshrestha et al. [473]

A.1.197 Twitter Presidential Politics

- **Description**: this dataset was created by collecting tweets, through the Twitter API, from 576 accounts linked to presidential candidates and members of congress, from the entire account history until December 2019. Out of all the accounts considered, 258 accounts were classified as Republican and 318 as Democratic. The dataset was collected to build a political bias subspace from word embeddings, which could be a flexible tool to quantitatively investigate political leaning in text-based media.
- **Affiliation of creators**: Clarkson University.
- **Domain**: social media.
- **Tasks in fairness literature**: bias audit [320].
- **Data spec**: text.
- **Sample size**: $\sim$ 1M tweets from $\sim$ 500 accounts.
- **Year**: 2020.
- **Sensitive features**: political leaning.
- **Link**: not available
- **Further info**: Gordon et al. [320]

A.1.198 Twitter Trending Topics

- **Description**: this dataset was used to study the problem of fair recommendation. It comprises a random sample (1%) of all tweets posted in the US between February and July 2017, obtained through the Twitter Streaming API. This sample is paired with a collection of trending Twitter topics queried every 15-minutes through the Twitter REST API in July 2017. User interest in each topic was inferred using Twitter lists and follower-followee graphs. Finally, user demographics were also annotated to evaluate how user interest in different topics skews with respect to race, age, and gender. These attributes were obtained feeding user profile images to Face++.
- **Affiliation of creators**: Indian Institute of Technology Kharagpur; Max Planck Institute for Software Systems; Grenoble INP.
- **Domain**: social media.
- **Tasks in fairness literature**: fair ranking [131].
- **Data spec**: text.
- **Sample size**: $\sim$ 200M tweets by $\sim$ 10M users and $\sim$ 10K trending topics.
- **Year**: 2018.
- **Sensitive features**: race, age, and gender.
- **Link**: not available
- **Further info**: Chakraborty et al. [131]

A.1.199 TwitterAAE

- **Description**: this resource was developed to study the use of dialect language on social media. The authors used Twitter APIs to collect public tweets sent on mobile phones from US users in 2013. They devise a distant supervision approach based on geolocation to annotate the probable language/dialect of the tweet, distinguishing between African American English (AAE) and Standard American English (SAE). To validate their approach, the creators studied the phonological and syntactic divergence of AAE tweets vs. SAE tweets, ensuring they align with linguistic phenomena that typically distinguish these variants of English.
- **Affiliation of creators**: University of Massachusetts Amherst.

- **Domain**: social media, linguistics.
- **Tasks in fairness literature**: fairness evaluation of sentiment analysis [717], fairness evaluation of private classification [33], fairness evaluation [38], robust fair language model [350], fairness evaluation of language identification [79].
- **Data spec**: text.
- **Sample size**: $\sim$ 8M tweets.
- **Year**: 2016.
- **Sensitive features**: dialect (related to race).
- **Link**: http://slanglab.cs.umass.edu/TwitterAAE/
- **Further info**: Blodgett et al. [80]

A.1.200 US Harmonized Tariff Schedules (HTS)

- **Description**: this resource represents a comprehensive classification system for goods imported in the US, which defines the applicable tariffs. It defines a fine-grained categorization for goods, based e.g. on their material and shape. The chapter on apparel was explicitly criticized for its differential treatment of men's and women's clothing, effectively resulting in discriminatory tariffs for consumers.
- **Affiliation of creators**: US International Trade Commission.
- **Domain**: economics.
- **Tasks in fairness literature**: fairness evaluation [520].
- **Data spec**: tabular data.
- **Sample size**: unknown.
- **Year**: present.
- **Sensitive features**: gender.
- **Link**: https://hts.usitc.gov/current
- **Further info**: Barbaro [45]

A.1.201 UniGe

- **Description**: this dataset is connected with the *DROP@UNIGE* project, aimed at studying the dynamics of university dropout, focusing on the University of Genoa as a case study. In ML fairness literature, the most common version of the dataset focuses on students who enrolled in 2017. Students are associated with attributes describing their ethnicity, gender, financial status, and prior school experience. The target variable encodes early academic success, as summarized by students' grades at the end of the first semester.
- **Affiliation of creators**: University of Genoa.
- **Domain**: education.
- **Tasks in fairness literature**: fair regression [158, 159], fair representation learning [612, 613].
- **Data spec**: tabular data.
- **Sample size**: $\sim$ 5K students.
- **Year**: unknown.
- **Sensitive features**: ethnicity, gender, financial status.
- **Link**: not availalbe
- **Further info**: Oneto et al. [614]

A.1.202 University Facebook Networks

- **Description**: a collection of 100 datasets shared with researchers in anonymized format by Adam D'Angelo of Facebook. The datasets used in the fairness literature consist of a 2005 snapshot from the Facebook network of the Universities of Oklahoma (Oklahoma97), North Carolina (UNC28), Caltech (Caltech36), Reed College (Reed98), and Michigan State (Michigan23), and links between them. User data comprises gender, class year, and anonymized data fields representing high school, major, and dormitory residences.
- **Affiliation of creators**: Facebook; University of North Carolina; Harvard University; University of Oxford.
- **Domain**: social networks.
- **Tasks in fairness literature**: fair graph mining [498], fair graph augmentation [664].
- **Data spec**: user-user pairs.
- **Sample size**: $\sim$ 20K people connected by $\sim$ 1M friend relations (Oklahoma97); $\sim$ 20K people connected by $\sim$ 1M friend relations (UNC28); $\sim$ 30K people connected by $\sim$ 1M friend relations (Michigan23); $\sim$ 1K people connected by $\sim$ 20K friend relations (Reed98); $\sim$ 1K people connected by $\sim$ 20K friend relations (Caltech36).
- **Year**: 2017.
- **Sensitive features**: gender.
- **Link**: http://networkrepository.com/socfb-Oklahoma97.php (Oklahoma97); http://networkrepository.com/socfb-UNC28.php (UNC28); https://networkrepository.com/socfb-Michigan23.php (Michigan23); https://networkrepository.com/socfb-Reed98.php (Reed98); https://networkrepository.com/socfb-Caltech36.php (Caltech36)
- **Further info**: Red et al. [668]

A.1.203 US Census Data (1990)

- **Description**: this resource is a one percent sample extracted from the 1990 US census data as a benchmark for clustering algorithms on large datasets. It contains a variety of features about different aspects of participants' lives, including demographics, wealth, and military service.
- **Affiliation of creators**: Microsoft.
- **Domain**: demography.
- **Tasks in fairness literature**: fair clustering [31, 61, 378], fair clustering under unawareness [239], limited-label fairness evaluation [688].
- **Data spec**: tabular data.
- **Sample size**: $\sim$ 2M respondents.
- **Year**: 1999.
- **Sensitive features**: age, sex.
- **Link**: https://archive.ics.uci.edu/ml/datasets/US+Census+Data+(1990)
- **Further info**: Meek et al. [551]

A.1.204 US Family Income

- **Description**: this resource was compiled from the Current Population Survey (CPS) Annual Social and Economic (ASEC) Supplement. It contains income data for over 80,000 thousand US families, broken down by age and race (White, Black, Asian, and Hispanic).
- **Affiliation of creators**: US Bureau of Labor Statistics; US Census Bureau.

- **Domain**: economics.
- **Tasks in fairness literature**: fair subset selection under unawareness [553].
- **Data spec**: tabular data.
- **Sample size**: 4 races x 12 age categories x 41 income categories.
- **Year**: 2020.
- **Sensitive features**: age, race.
- **Link**: https://www.census.gov/data/tables/time-series/demo/income-poverty/cps-finc/finc-02.html
- **Further info**: https://www2.census.gov/programs-surveys/cps/techdocs/cpsmar20.pdf

A.1.205 US Federal Judges

- **Description**: this dataset was extracted from Epstein et al. [237] to study the problem of judicial subset selection from the point of view of justice, fairness and interpretability. Given the fact that in several judicial systems a subset of judges is selected from the whole judicial body to decide the outcome of appeals, the creators extract cases were three judges are required from Epstein et al. [237], covering the period 2000–2004. They emulate prior probabilities of affirmance/reversal for specific judges based on their past decisions. The task associated with this dataset is the optimal selection of a subset of judges, so that the procedure is interpretable, the subset contains at least one female (junior) judge and the decision of the subset coincides with the decision of the whole judicial body.
- **Affiliation of creators**: Yale University.
- **Domain**: law.
- **Tasks in fairness literature**: fair subset selection [380].
- **Data spec**: judge-case pairs.
- **Sample size**: $\sim$300 judges selected for $\sim$ 2K cases.
- **Year**: 2020.
- **Sensitive features**: gender.
- **Link**: not available
- **Further info**: Huang et al. [380]

A.1.206 US Student Performance

- **Description**: this resource represents students at an undisclosed US research university, spanning the Fall 2014 to Spring 2019 terms. The associated task is predicting student success based on university administrative records. Student features include demographics and academic information on prior achievement and standardized test scores.
- **Affiliation of creators**: Cornell University.
- **Domain**: education.
- **Tasks in fairness literature**: fairness evaluation [490].
- **Data spec**: tabular data.
- **Sample size**: unknown.
- **Year**: 2020.
- **Sensitive features**: gender, racial-ethnic group.
- **Link**: not available
- **Further info**: Lee and Kizilcec [490]

A.1.207 UTK Face

- **Description**: the dataset was developed as a diverse resource for face regression and progression (models of aging), where diversity is intended with respect to age, gender and race. The creators sourced part of the images from two existing datasets (Morph and CACD datasets). To increase the representation of some age groups, additional images were crawled from major search engines based on specific keywords (e.g., baby). Age, gender, and race were estimated through an algorithm and validated by a human annotator.
- **Affiliation of creators**: University of Tennessee.
- **Domain**: computer vision.
- **Tasks in fairness literature**: robust fairness evaluation [589], fairness evaluation of private classification [33], fairness evaluation [706], fair classification [415].
- **Data spec**: image.
- **Sample size**: $\sim$ 20K face images.
- **Year**: 2017.
- **Sensitive features**: age, gender, race (inferred).
- **Link**: https://susanqq.github.io/UTKFace/
- **Further info**: Zhang et al. [875]

A.1.208 Vehicle

- **Description**: this dataset comprises measurements from a distributed network of acoustic, seismic, and infrared sensors, as different types of military vehicles are driven in their proximity. This dataset was developed as part of a project supported by DARPA for the task of vehicle detection and type classification.
- **Affiliation of creators**: University of Wisconsin-Madison.
- **Domain**: signal processing.
- **Tasks in fairness literature**: fair federated learning [500].
- **Data spec**: time series.
- **Sample size**: unknown.
- **Year**: 2013.
- **Sensitive features**: none.
- **Link**: http://www.ecs.umass.edu/mduarte/Software.html
- **Further info**: Duarte and Hu [218]

A.1.209 Victorian Era Authorship Attribution

- **Description**: this resource was developed to benchmark different authorship attribution techniques. Querying the Gdelt database, the creators focus on English language authors from the 19th century with at least five books available. The corpus was split into text fragments of 1,000 words each. Only the most frequent 10,000 words were kept, while the remaining ones were removed.
- **Affiliation of creators**: Purdue University.
- **Domain**: literature.
- **Tasks in fairness literature**: fair clustering [343].
- **Data spec**: text.
- **Sample size**: $\sim$ 100K text fragments.
- **Year**: 2018.
- **Sensitive features**: textual references to people and their demographics.
- **Link**: http://archive.ics.uci.edu/ml/datasets/Victorian+Era+Authorship+Attribution
- **Further info**: Gungor [331]

A.1.210 Visual Question Answering (VQA)

- **Description**: this dataset is curated as a benchmark for open-ended visual question answering. The collection features both real images from MS-COCO [505] and abstract scenes with human figures. Questions and answers were compiled by workers on Mechanical Turk who were instructed to formulate questions that require seeing the associated image for a correct answer.
- **Affiliation of creators**: Georgia Institute of Technology; Carnegie Mellon University; Army Research Lab; Facebook AI Research.
- **Domain**: computer vision.
- **Tasks in fairness literature**: bias discovery [530].
- **Data spec**: mixture (image, text).
- **Sample size**: $\sim$ 1M questions over $\sim$ 300K images.
- **Year**: 2017.
- **Sensitive features**: visual and textual references to gender.
- **Link**: https://visualqa.org/
- **Further info**: Goyal et al. [321]

A.1.211 Warfarin

- **Description**: this dataset was collected as part of a study about algorithmic estimation of optimal warfarin dosage as an oral anticoagulation treatment. The study was carried out by the International Warfarin Pharmacogenetics Consortium, comprising 21 research groups from 9 countries and 4 continents. The dataset was co-curated by staff at the Pharmacogenomics Knowledge Base (PharmGKB) including, for thousands of patients at centers around the world, their demographics, comorbities, other medications and genetic factors, along with the steady-state dose of warfarin that led to stable levels of anticoagulation without adverse events.
- **Affiliation of creators**: PharmGKB; International Warfarin Pharmacogenetics Consortium.
- **Domain**: pharmacology.
- **Tasks in fairness literature**: fairness evaluation under unawareness [418].
- **Data spec**: tabular data.
- **Sample size**: $\sim 6K$ patients.
- **Year**: 2009.
- **Sensitive features**: sex, ethnicity, age.
- **Link**: https://www.pharmgkb.org/downloads
- **Further info**: International Warfarin Pharmacogenetics Consortium [388]

A.1.212 Waterbirds

- **Description**: this computer vision dataset consists of photos where subjects and backgrounds are carefully paired to induce spurious correlations. Subjects are birds, taken from the CUB dataset [804], divided into waterbirds and landbirds. Pixel-level segmentation masks are exploited to cut out subjects and paste them onto land or water backgrounds from the Places dataset [886]. While in the provided validation and test splits both landbirds and waterbirds appear with the same frequency on either background, the training split is imbalanced so that 95% of all waterbirds are placed against a water background and 95% of all landbirds are depicted against a land background.
- **Affiliation of creators**: Stanford University; Microsoft.
- **Domain**: computer vision.

- **Tasks in fairness literature**: fairness evaluation of selective classification [411].
- **Data spec**: image.
- **Sample size**: $\sim$ 10K images.
- **Year**: 2021.
- **Sensitive features**: none.
- **Link**: https://github.com/ejones313/worst-group-sc/tree/main/src/data
- **Further info**: Sagawa et al. [691]

A.1.213 WebText

- **Description**: this resource is a web scrape collected to train the GPT-2 language model. The authors considered all outbound links from Reddit which collected at least 3 *karma*. This inclusion criterion signals that the link received some upvotes by redditors and is treated as a quality heuristic for the webpage. To extract text data from each link, a combination of Dragnet [636] and Newspaper[17] extractors was exploited. The curators performed deduplication and removed all Wikipedia pages to reduce text overlap with Wikipedia-based datasets.
- **Affiliation of creators**: OpenAI.
- **Domain**: linguistics.
- **Tasks in fairness literature**: data bias evaluation [754].
- **Data spec**: text.
- **Sample size**: $\sim$ 8M documents.
- **Year**: 2019.
- **Sensitive features**: textual references to people and their demographics.
- **Link**: https://github.com/openai/gpt-2-output-dataset (partial)
- **Further info**: Radford et al. [655]

A.1.214 Wholesale

- **Description**: this dataset represents Portuguese businesses from the catering industry purchasing goods from the same wholesaler. The businesses are located in Lisbon, Oporto, and a third undisclosed area; 298 are from the Horeca (Hotel/Restaurant/Café) channel and 142 from the Retail channel. Each data point comprises this information along with yearly expenditures on different categories of products (e.g. milk, frozen goods, delicatessen). Collection of this data was presumably carried out by the wholesaler in a business intelligence initiative primarily aimed at customer segmentation and targeted marketing.
- **Affiliation of creators**: Université Pierre et Marie Curie; University Institute of Lisbon; INRIA.
- **Domain**: marketing.
- **Tasks in fairness literature**: fair data summarization [413].
- **Data spec**: tabular data.
- **Sample size**: $\sim$ 400 businesses.
- **Year**: 2014.
- **Sensitive features**: geography.
- **Link**: http://archive.ics.uci.edu/ml/datasets/wholesale+customers
- **Further info**: Baudry et al. [52]

[17] https://github.com/codelucas/newspaper

A.1.215 Wikidata

- **Description**: founded in 2012, Wikidata is a free, collaborative, multilingual knowledge base, maintained by editors and partly automated. It consists of items linked by properties. The most common items include humans, administrative territorial entities, architectural structures, chemical compounds, films, and scholarly articles.
- **Affiliation of creators**: Wikimedia Foundation.
- **Domain**: information systems.
- **Tasks in fairness literature**: fairness evaluation in graph mining [267].
- **Data spec**: item-property-value triples.
- **Sample size**: $\sim$ 90M items.
- **Year**: present.
- **Sensitive features**: demographics of people featured in entities (age, sex, geography) and their relations.
- **Link**: https://www.wikidata.org/wiki/Wikidata:Data_access
- **Further info**: https://www.wikidata.org/wiki/Wikidata:Main_Page

A.1.216 Wikipedia dumps

- **Description**: Wikipedia dumps are maintained and updated regularly by the Wikimedia Foundation. Typically, they contain every article available in a language at a given time. As a large source of curated text, they have often been used by the natural language processing and computational linguistics communities to extract models of human language. We find usage of German, English, Mandarin Chinese, Spanish, Arabic, French, Farsi, Urdu, and Wolof dumps in the surveyed articles.
- **Affiliation of creators**: Wikimedia Foundation.
- **Domain**: linguistics.
- **Tasks in fairness literature**: bias evaluation in WEs [99, 143, 504, 620].
- **Data spec**: text.
- **Sample size**: $\sim$ 6M articles (EN), $\sim$ 3M articles (DE) as of May 2021.
- **Year**: present.
- **Sensitive features**: textual references to people and their demographics.
- **Link**: https://dumps.wikimedia.org/enwiki/; https://dumps.wikimedia.org/dewiki/
- **Further info**: https://meta.wikimedia.org/wiki/Data_dumps

A.1.217 Wikipedia Toxic Comments

- **Description**: this dataset was developed as a resource to analyze discourse and personal attacks on Wikipedia talk pages, which are used by editors to discuss improvements. It is aimed at using ML for better online conversations and flag posts that are likely to make other participants leave. The data consists of Wikipedia comments labelled by 5,000 crowd-workers according to their toxicity level (toxic, severe_toxic) and type (obscene, threat, insult, identity_hate). This resource powers a public Kaggle competition.
- **Affiliation of creators**: Wikimedia foundation; Google.
- **Domain**: social media.
- **Tasks in fairness literature**: fair classification, [289, 710], fairness evaluation [215].
- **Data spec**: text.
- **Sample size**: $\sim$ 160K comments.
- **Year**: 2017.
- **Sensitive features**: textual reference to people and their demographics.
- **Link**: https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge
- **Further info**: https://www.perspectiveapi.com/research/

A.1.218 Willingness-to-Pay for Vaccine

- **Description**: this dataset resulted from a study of willingness to pay for a vaccine against tick-borne encephalitis in Sweden. Thousands of citizens from different areas of the country filled in a survey about exposure, risk perception, knowledge, and protective behavior related to ticks and tick-borne diseases, along with socioeconomic information. The central question of the survey asks how much respondents would be willing to pay for a vaccine that provides a three-year protection against tick-borne encephalitis.
- **Affiliation of creators**: University of Gothenburg.
- **Domain**: public health.
- **Tasks in fairness literature**: fair pricing evaluation [422].
- **Data spec**: tabular data.
- **Sample size**: $\sim$ 2K respondents.
- **Year**: 2015.
- **Sensitive features**: age, gender, geography.
- **Link**: https://snd.gu.se/sv/catalogue/study/snd0987/1#dataset
- **Further info**: Slunge [731]

A.1.219 Winobias

- **Description**: similarly to Winogender, this benchmark was built to study coreference resolution and gender bias, focusing on words that relate to professions with diverse gender representation. Example: "The physician hired the secretary because he (she) was overwhelmed with clients". The correct pronoun resolution is clear from the syntax or semantics of the sentence and can be either stereotypical or counter-stereotypical. The accuracy of biased coreference resolution systems will vary accordingly.
- **Affiliation of creators**: University of California Los Angeles; University of Virginia; Allen Institute for Artificial Intelligence.
- **Domain**: linguistics.
- **Tasks in fairness literature**: fair entity resolution evaluation. [795].
- **Data spec**: text.
- **Sample size**: $\sim$ 3K sentences.
- **Year**: 2020.
- **Sensitive features**: gender.
- **Link**: https://github.com/uclanlp/corefBias/tree/master/WinoBias/wino
- **Further info**: Zhao et al. [881]

A.1.220 Winogender

- **Description**: this dataset was crafted to systematically study gender bias in systems for coreference resolution, the task of resolving whom pronouns refer to in a sentence. This resource follows the Winograd schemas, with sentence templates mentioning a profession (nurse), a participant (patient), and a pronoun referring to either one of them: "The nurse notified the patient that her/his/their shift would be ending in an hour." Sentence templates have been crafted so that the pronoun resolution can be done unambiguously based on contextual information, hence unbiased systems should display similar error rates, regardless of gender concentrations in different professions. The ground truth for each sentence has been validated by workers on Mechanical Turk with accuracy over 99%.
- **Affiliation of creators**: Johns Hopkins University.

- **Domain**: linguistics.
- **Tasks in fairness literature**: fair entity resolution evaluation [795], fairness evaluation in entity recognition [569].
- **Data spec**: text.
- **Sample size**: $\sim$ 700 sentences.
- **Year**: 2018.
- **Sensitive features**: gender.
- **Link**: https://github.com/rudinger/winogender-schemas
- **Further info**: Rudinger et al. [685]
- **Variants**: Winogender-NER [569] is a modified version of the template appropriate for named entity recognition.

A.1.221 Word Embedding Association Test (WEAT)

- **Description**: this resource was created to audit biases in English WEs. Following the Implicit Association Test (IAT) from social psychology [326], this dataset defines two groups of target words, relating e.g. to flowers and insects, and two groups of attribute words, relating e.g. to pleasantness and unpleasantness. The dataset can be used to measure biased associations between the target words and the attribute words represented by a set of WEs. WEAT comprises ten tests across different word categories. The most salient for the purposes of algorithmic fairness support tests of associations between race and pleasantness, age and pleasantness, gender and career (vs family), gender and propensity to math (vs arts). Race-related words are first names predominantly associated with African American or European American individuals. Gender is encoded in a similar fashion, or with intrinsically gendered words (e.g. mother).
- **Affiliation of creators**: Princeton University; University of Bath.
- **Domain**: linguistics.
- **Tasks in fairness literature**: bias evaluation in WEs [99, 334].
- **Data spec**: text.
- **Sample size**: $\sim$10 groups of words, with $\sim$10-60 words in each group.
- **Year**: 2017.
- **Sensitive features**: race, gender.
- **Link**: https://arxiv.org/pdf/1608.07187.pdf
- **Further info**: Caliskan et al. [109]

A.1.222 Yahoo! A1 Search Marketing

- **Description**: this dataset contains bids from all advertisers who participated in Yahoo! Search Marketing auctions for the top 1000 search queries from June 15, 2002, to June 14, 2003. The identities of advertisers and the queries they target are anonymized for confidentiality reasons.
- **Affiliation of creators**: Yahoo! Labs.
- **Domain**: marketing.
- **Tasks in fairness literature**: fair advertising [121, 594].
- **Data spec**: advertiser-keyword pairs.
- **Sample size**: $\sim$ 20M bids by $\sim$ 10K advertisers over $\sim$ 1K search queries.
- **Year**: after 2003.
- **Sensitive features**: none.
- **Link**: https://webscope.sandbox.yahoo.com/catalog.php?datatype=a
- **Further info**:

A.1.223 Yahoo! c14B Learning to Rank

- **Description**: this resource consists of 2 datasets which encode the interactions of Yahoo! users with the search engine in the US and an unknown Asian country. This data is a subset of the entire training set used internally to train the ranking functions of the Yahoo! search engine. Textual features are deliberately obfuscated and the final data consists of numerical features which encode query-document pairs. Query-document pairs are assigned multigraded relevance judgements by a professional editor.
- **Affiliation of creators**: Yahoo! Labs.
- **Domain**: information systems.
- **Tasks in fairness literature**: fair ranking [723].
- **Data spec**: query-document pairs.
- **Sample size**: $\sim$ 40K queries, $\sim$ 900K documents.
- **Year**: 2011.
- **Sensitive features**: none.
- **Link**: https://webscope.sandbox.yahoo.com/catalog.php?datatype=c
- **Further info**: Chapelle and Chang [133]

A.1.224 YouTube Dialect Accuracy

- **Description**: this dataset was curated to audit the accuracy of YouTube's automated captioning system across two genders and five dialects of English. Eighty speakers were sampled from videos matching the query "accent challenge <region>" or "accent tag <region>", where <region> is one of five areas selected for geographic separation and distinct local dialects: California, Georgia, New England, New Zealand and Scotland. This curation choice targets a popular internet phenomenon (called "accent tag", "dialect meme" or "accent challenge") consisting of videos of people from different areas presenting themselves and their linguistic background, subsequently reading a list of words designed to elicit pronounciation differences dependent on dialect. This resource focuses only on the word portion of these videos, with a "phonetically-trained listener familiar with the dialects" performing the annotation for word caption accuracy.
- **Affiliation of creators**: University of Washington.
- **Domain**: social media.
- **Tasks in fairness literature**: fairness evaluation of speech recognition [760].
- **Data spec**: tabular data.
- **Sample size**: $\sim$ 100 speakers.
- **Year**: 2016.
- **Sensitive features**: gender, geography.
- **Link**: https://github.com/rctatman/youtubeDialectAccuracy
- **Further info**: Tatman [760]

A.1.225 Yow news

- **Description**: this dataset was collected to support research on personalized information integration and retrieval. The data, consisting of implicit and explicit user feedback stored in interaction logs, was gathered in a user study via a special browser accessing a web-based news story filtering system. The task associated with this resource is personalized news recommendation.
- **Affiliation of creators**: Carnegie Mellon University.
- **Domain**: news, information systems.

- **Tasks in fairness literature**: fair ranking [722].
- **Data spec**: user-story pairs.
- **Sample size**: $\sim$ 10K interaction logs.
- **Year**: 2009.
- **Sensitive features**: news provider.
- **Link**: https://users.soe.ucsc.edu/~yiz/papers/data/YOWStudy/
- **Further info**: Zhang [870]; https://users.soe.ucsc.edu/~yiz/piir/

A.1.226 Zillow Searches

- **Description**: this is a proprietary dataset from Zillow, a famous real estate marketplace. It consists of a random sample of over 13,000 search sessions covering more than 36,000 property listings. Each listing consists of several features, some of which are considered salient by the creators and a sensible target for fair ranking algorithms. Among these are the ownership of the house (Zillow, independent realtor, new construction listed by builders) and the availability of 3D/video tours of the property. This dataset was collected internally to study the problem of fair recommendation and ranking on Zillow data.
- **Affiliation of creators**: Boston University; Zillow Group.
- **Domain**: information systems.
- **Tasks in fairness literature**: fair ranking [134].
- **Data spec**: unknown.
- **Sample size**: $\sim$ 10K search sessions featuring $\sim$ 40K property listings.
- **Year**: 2020.
- **Sensitive features**: ownership, tour availability.
- **Link**: not available
- **Further info**: Chaudhari et al. [134]

A.2 Adult

Key references include Cohany et al. [164], Ding et al. [213], Kohavi [459], McKenna [545, 546], UCI Machine Learning Repository [775], US Dept. of Commerce Bureau of the Census [781].

A.2.1 Datasheet

Motivation

- **For what purpose was the dataset created?**

  The Adult dataset was created as a resource to benchmark the performance of machine learning algorithms. Rather than powering a specific task or application, the dataset was likely chosen as a real-world source of socially relevant data [459].

- **Who created the dataset?**

  Barry Becker extracted this dataset from the 1994 Census database. Ronny Kohavi and Barry Becker donated it to UCI Machine Learning Repository in 1996. At that time, both were working for Silicon Graphics Inc [775]

- **Who funded the creation of the dataset?**

  The underlying database is a product of the Current Population Survey (CPS) of March 1994, a joint effort by the US Census Bureau and the US Bureau of Labor Statistics (BLS), funded by the US federal government. The extraction of Adult from the larger database was plausibly part of work remunerated by Silicon Graphics.

Composition

- **What do the instances that comprise the dataset represent?**

  Each instance is a **March 1994 CPS respondent**, represented along demographic and socio-economic dimensions.

- **How many instances are there in total?**

  The dataset consists of **48,842 instances**.

- **Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?**

  Adult contains individuals from a **sample** of US households, extracted from the 1994 Annual Social and Economic Supplement (ASEC) of the CPS with the following query:

  $$(AAGE > 16)\&\&(AGI > 100)\&\&(AFNLWGT > 1)\&\&(HRSWK > 0).$$

  This means Adult focuses on a subset of ASEC respondents aged 17 or older, whose income is above $100, working at least 1 hour per week. While these were conceived as conditions to filter out noisy records [775], they may introduce sampling effects. Moreover, the 1994 CPS data was itself a sample, selected according to Census Bureau best practices, reaching over 70,000 households in nearly 2,000 US counties. The March 1994 CPS sample aimed at obtaining more reliable information on the Hispanic population, and was hence extended to an additional 2,500 eligible housing units.

- **What data does each instance consist of?**

  Each instance consists of a combination of nominal, ordinal and continuous attributes, denominated age, workclass, fnlwgt, education, education-num, marital-status, occupation, relationship, race, sex, capital-gain, capital-loss, hours-per-week, native-country. See Table A.4 for a detailed explanation of features and their values.

- **Is there a label or target associated with each instance?**

  Yes. Each person instance comes with a binary label encoding whether their income is above a $50,000$ threhsold.

- **Is any information missing from individual instances?**

  Yes. Over 7% of the instances have missing values. This is likely due to issues with data recording and coding or respondents' inability to recall information.

- **Are relationships between individual instances made explicit e.g., users' movie ratings, social network links)?**

  No. Some instances are related persons from the same household [781] but this information is not reported in the dataset.

- **Are there recommended data splits?**

  Yes. The dataset comes with a specified train/test split made using MLC++ GenCVFiles, resulting in a 2/3–1/3 random split [775]. The training set consists of 32561 instances, the test set of 16281 instances.

- **Are there any errors, sources of noise, or redundancies in the dataset?**

  Yes. Sources of error include definitional difficulties, differences in interpretation of questions, respondents inability or unwillingness to provide correct information, errors made during data collection, data processing or missing value imputation. The tendency in household surveys for respondents to under-report their income was an explicit concern. Finally, noise infusion such as topcoding (saturation to $99,999) was applied to avoid re-identification of certain individuals [781].

- **Is the dataset self-contained, or does it link to or otherwise rely on external resources?**

  The dataset is **self-contained**.

- **Does the dataset contain data that might be considered confidential?**

  Yes. The data is protected by Title 13 of the United States Code, protecting individuals against identification from Census data.[18]

- **Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?**

  No, not strictly. Interpreting the question more broadly, however, the envisioned racial and sexual categories may be deemed inadequate.

- **Does the dataset identify any subpopulations (e.g., by age, gender)?**

  Yes. The dataset provides information on sex, age and race of respondents. These were self-reported, although self-identification was bounded by envisioned categories. These are (female, male) for sex and (White, Black, American Indian/Aleut Eskimo, Asian or Pacific Islander, Other) for race. Table A.1 summarizes the marginal distribution of the Adult dataset across these subpopulations.

- **Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset?**

  Unknown. Important variables for data re-identification, such as birth date or ZIP code, are absent from the Adult dataset. However, instances in this dataset may be linked to the original CPS 1994 data [213]. Moreover, re-identification studies internal to the Census Bureau pointed to combinations of variables that could potentially be used to re-identify respondents from Census microdata [546].

- **Does the dataset contain data that might be considered sensitive in any way?**

  Yes. This dataset contains sensitive data, such as sex, race, native country and financial situation of respondents.

- **Any other comments?**

  A precise definition for the variable called fnlwgt is unknown. It was used by Census Bureau statisticians to obtain population-level estimates from the CPS sample. For this reason, its use in classification tasks would be unusual.

[18] https://www.census.gov/about/policies/privacy/data_stewardship/title_13_-_protection_of_confidential_information.html

Demographic Caracteristic	Values
Percentage of male subjects	66.85%
Percentage of female subjects	33.15%
Percentage of White subjects	85.50%
Percentage of Black subjects	9.60%
Percentage of Asian-Pac-Islander subjects	3.11%
Percentage of Amer-Indian-Eskimo subjects	0.96%
Percentage of people belonging to other races	0.83%
Percentage of people between 16-19 years old	5.14%
Percentage of people between 20-29 years old	24.58%
Percentage of people between 30-39 years old	26.47%
Percentage of people between 40-49 years old	21.95%
Percentage of people between 50-59 years old	13.55%
Percentage of people between 60-69 years old	6.25%
Percentage of people between 70-79 years old	1.67%
Percentage of people between 80-89 years old	0.27%
Percentage of people between 90-99 years old	0.11%

Table A.1 Demographic Characteristics of the Adult dataset.

Collection process

- **How was the data associated with each instance acquired?**

  Trained interviewers asked questions directly to respondents [781]. The data was made available through US Census data products which were used by Barry Becker to extract the Adult dataset.

- **What mechanisms or procedures were used to collect the data?**

  Interviewers conducted the survey either in person at the respondent's home or by phone. They used laptop computers with ad-hoc software to prompt questions and record answers. At the end of each day, interviewers transmitted the collected data via modem to the Bureau headquarters [781].

- **If the dataset is a sample from a larger set, what was the sampling strategy?**

  A probabilistic sample was selected according to US Census Bureau best practice, with a muti-stage stratified design. The US territory was divided into strata, from which one county (or group of counties) was selected. From each selected county a sample of addresses was later obtained and added to the

sample [780]. Barry Becker extracted a "set of reasonably clean records" using the following conditions:

$$(AAGE > 16)\&\&(AGI > 100)\&\&(AFNLWGT > 1)\&\&(HRSWK > 0).$$

- **Who was involved in the data collection process and how were they compensated?**

  Interviewers trained by the US Census Bureau were involved in the data collection process. Data extraction was later performed by Barry Becker while affiliated with Silicon Graphics. Their compensation is unknown.

- **Over what timeframe was the data collected?**

  Respondents were interviewed in March 1994, while the Adult dataset was donated to UCI ML Repository in May 1996.

- **Were any ethical review processes conducted?**

  The Microdata Review Panel likely reviewed this data for compliance with Title 13 [546] and authorized its publication.

- **Was the data collected from the individuals in question directly, or obtain it via third parties or other sources?**

  Directly. US Census Bureau interviewers collected the data through interviews, conducted in person or over the phone. Danny Kohavi and Barry Becker later processed this data, obtaining it from the Census Bureau website.

- **Were the individuals in question notified about the data collection?**

  Yes. Individuals knew they were part of a sample chosen by the Census Bureau chosen for statistical analysis. They were not notified about their data being included in the Adult dataset.

- **Did the individuals in question consent to the collection and use of their data?**

  Yes. For the CPS, participation is voluntary. A recent version of the information provided to respondents before interviews is available on the US Census Website.[19]

- **If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses?**

  Unknown.

- **Has an analysis of the potential impact of the dataset and its use on data subjects been conducted?**

  Yes. Re-identification studies have been conducted both internally [546] and externally [676] on Census Bureau data. McKenna [546] mention finding combinations of variables on Census files that can lead to successful re-identification, which were subsequently removed or protected with noise injection. Rocher et al. [676] demonstrate on the Adult dataset that the likelihood of a specific individual to have been correctly re-identified can be estimated with high accuracy. We are unaware of studies about the potential impact of successful re-identification on respondents.

[19] https://www2.census.gov/programs-surveys/cps/advance_letter.pdf

Preprocessing/cleaning/labelling

- **Was any preprocessing/cleaning/labeling of the data done?**

  Yes. Preprocessing operations by the Census Bureau include missing value imputation and topcoding. Furthermore, Barry Becker and Ron Kohavi binarized the income variable ($>$ \$50K) and discarded several CPS respondents who are not included in the Adult dataset.

- **Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data?**

  Unknown.

- **Is the software used to preprocess/clean/label the instances available?**

  Likely no. It seems unlikely for the code to be available 25 years after its last known use.

Uses

- **For what tasks has the dataset been used?**

  This dataset probably owes its status in the ML community to an early position of publicly-available and interesting resource based on real-world data. For this reason, rather than powering specific applications, Adult is used as a benchmark for classifiers in many fields of machine learning. Due to its encoding of sensitive attributes, it has also become the most used dataset in the fair ML literature.

- **Is there a repository that links to any or all papers or systems that use the dataset?**

  Yes. A selection of early works (pre-2005) using this dataset can be found in UCI Machine Learning Repository [775]. A more recent list is available under the beta version of the UCI ML Repository.[20] See Appendix A.1.7 for a (non-exhaustive) list of algorithmic fairness works using this resource.

- **What (other) tasks could the dataset be used for?**

  The Adult dataset is used in tasks where data of social significance is deemed important, for example privacy-preserving ML.

- **Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?**

  Yes. The threshold used to quantize income for a binary classification task is very high (\$50K). As a result a trivial rejector achieves very large accuracy on the black subpopulation (93%). For the same reason, models are often more accurate for the female subpopulation than for the male one [213]. Some numerical results on Adult may be an artifact of this threshold choice.

- **Are there tasks for which the dataset should not be used?**

  Based on the previous answer, we caution against drawing overarching conclusions based on experimental results obtained on this dataset alone.

Distribution

- **Is the dataset distributed to third parties outside of the entity on behalf of which the dataset was created?**

  Yes. The dataset is publicly available [775].

[20] https://archive-beta.ics.uci.edu/ml/datasets/2

- **How is the dataset distributed?**

  The dataset is available as a **csv file**.

- **When was the dataset distributed?**

  The dataset was released on the UCI ML Repository in **May 1996**.

- **Is the dataset distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?**

  Yes. The UCI ML repository has a citation policy. Terms of Use concerning the privacy of CPS respondents are likely to apply.

- **Have any third parties imposed IP-based or other restrictions on the data associated with the instances?**

  Likely no. We are unaware of any IP-based restrictions.

- **Do any export controls or other regulatory restrictions apply to the dataset or to individual instances?**

  Likely no.

Maintenance

- **Who is supporting/hosting/maintaining the dataset?**

  The dataset is hosted and maintained by the **UCI Machine Learning Repository** [775].

- **How can the owner/curator/manager of the dataset be contacted?**

  Comments and inquiries may be directed at ml-repository@ics.uci.edu. Ronny Kohavi is the primary contact for this specific resource, available at ronnyk@live.com.

- **Is there an erratum?**

  Likely no. We are unaware of any erratum.

- **Will the dataset be updated?**

  A superset of the dataset without quantization of the target income variable is available [213].

- **If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances?**

  Unknown.

- **Will older versions of the dataset continue to be supported/hosted/maintained?**

  Unless otherwise indicated, the Adult dataset will remain hosted on the UCI ML Repository in its current version.

- **If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?**

  Unknown.

A.2.2 Data Nutrition Label

Metadata

METADATA	
Filenames	adult
Format	csv
Url	https://archive.ics.uci.edu/ml/datasets/adult
Domain	Economics
Keywords	US census, income
Type	Tabular
Rows	48842
Columns	14
% of missing cells	0.9%
Rows with missing cells	7%
License	UCI Repository citation policy
Released	May 1996
Range	1994
Description	A benchmark for classifiers tasked with predicting whether individual income exceeds $50K/yr based on demographic and socio-economic information. Also known as “Census Income” dataset.

Table A.2 Metadata of the Adult dataset

Provenance

PROVENANCE	
Source	
Name	U. S. Census Bureau
Url	https://www.census.gov/en.html
email	//
Authors	
Names	Ronny Kohavi and Barry Becker
Url	https://archive.ics.uci.edu/ml/datasets/
email	ronnyk@live.com

Table A.3 Provenance of the Adult dataset

Variables

VARIABLES	
age	Respondent's age.
workclass	Broad classification of employment, with following envisioned classes. Private Self-emp-not-inc (Self employed not-incorporated) Self-emp-inc (Self employed incorporated) Federal-gov Local-gov State-gov Without-pay (Without pay in family business) Never-worked
fnlwgt	Variable used to produce population estimates from the CPS sample.
education	Educational attainment of respondent. Preschool 1st-4th 5th-6th 7th-8th 9th 10th 11th 12th (no diploma) HS-grad (High school graduation) Some-college (no degree) Assoc-voc (associate degree in college, vocation program) Assoc-acdm (associate degree in college, academic program) Bachelors Masters Prof-school (professional school) Doctorate
education-num	Ordinal encoding of previous variable.

Table A.4 Variables of the Adult dataset (1/3).

VARIABLES	
marital-status	Respondent's marital status, with following envisioned classes. Married-civ-spouse (married, civilian spouse present) Divorced Never-married Separated Widowed Married-spouse-absent Married-AF-spouse (married, armed force spouse)
occupation	Job of respondent. Tech-support (Technical, sales, and administrative support) Craft-repair (Precision production, craft, and repair) Other-service Sales Exec-managerial (Managerial and professional speciality) Prof-specialty (Professional speciality) Handlers-cleaners (Handlers, equipment cleaners, helpers, and laborers) Machine-op-inspct (Operators, fabricators, and laborers) Adm-clerical (Administrative support occupations, including clerical) Farming-fishing (Farming, forestry, and fishing) Transport-moving (Transportation and material moving) Priv-house-serv (Private household service, e.g. cooks, cleaners) Protective-serv (Protective service, e.g. firefighters, police) Armed-Forces
relationship	Familial role wihtin household. Wife Own-child Husband Not-in-family Other-relative Unmarried

Table A.5 Variables of the Adult dataset (2/3).

VARIABLES	
race	Respondent's race. Amer-Indian-Eskimo Asian-Pac-Islander Black White Other
sex	Respondent's sex. Female Male
capital-gain	Profits from sale of assets.
capital-loss	Losses from sale of assets.
hours-per-week	Average hours of work per week.
native-country	Native Country of respondent
target variable	Does respondent's income exceed $50,000?

Table A.6 Variables of the Adult dataset (3/3).

Statistics

STATISTICS						
Ordinal						
name	type	count	unique	mostFrequent	leastFrequent	missing
education-num	int	48842	16	9	1	0

Table A.7 Ordinal variables statistics of the Adult dataset

Categorical						
name	type	count	unique	mostFrequent	leastFrequent	missing
workclass	string	48842	8	Private	Never-worked	2799
education	string	48842	16	HS-grad	Preschool	0
marital-status	string	48842	7	Married-civ-spouse	Married-AF-spouse	0
occupation	string	48842	14	Prof-specialty	Armed-Forces	2809
relationship	string	48842	6	Husband	Other-relative	0
race	string	48842	5	White	Other	0
sex	string	48842	2	Male	Female	0
native-country	string	48842	41	United-States	Holand-Netherlands	857
target variable	string	48842	2	$<=$ 50K	$>$ 50K	0

Table A.8 Categorical variables statistics of the Adult dataset

Quantitative									
name	type	count	min	median	max	mean	stdDev	miss	zeros
age	int	48842	17	37	90	38.64	13.71	0	0
fnlwgt	int	48842	12285	178144	1490400	189664	105604	0	0
capital-gain	int	48842	0	0	99999	1079.07	7452.02	0	44807
capital-loss	int	48842	0	0	4356	87.50	403	0	46560
hours-per-week	int	48842	1	40	99	40.42	12.39	0	0

Table A.9 Quantitative variables statistics of the Adult dataset.

A.3 COMPAS

Key references include Angwin et al. [21], Bao et al. [43], Barenstein [46], Brennan et al. [93], Dieterich et al. [212], Equivant [238], Larson et al. [483], ProPublica [646].

A.3.1 Datasheet

Motivation

- **For what purpose was the dataset created?**

  This dataset was created for an external audit of racial biases in the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) risk assessment tool developed by Northpointe (now Equivant), which estimates the likelihood of a defendant becoming a recidivist.

- **Who created the dataset and on behalf of which entity?**

  The dataset was created by Julia Angwin (senior reporter), Jeff Larson (data editor), Surya Mattu (contributing researcher), Lauren Kirchner (senior reporting fellow). All four contributors were affiliated with ProPublica at the time.

- **Who funded the creation of the dataset?**

  The dataset curation work was likely remunerated by ProPublica.

Composition

- **What do the instances that comprise the dataset represent?**

  Each instance is a person that was scored for risk of recidivism by the COMPAS system in Broward County, Florida, between 2013–2014. In other words, instances are **defendants**.

- **How many instances are there in total?**

  The COMPAS dataset [646] consists of **11,757** defendants assessed at the pretrial stage (`compas-scores.csv`). A separate dataset is released for a subset of 7,214 defendants that were observed for two years after screening (`compas-scores-two-years.csv`). Finally a smaller subset of 4,743 defendants focuses on violent recidivism (`compas-scores-two-year`

- **Does the dataset contain all possible instances or is it a sample of instances from a larger set?**

The dataset represents a **convenience sample** of all individuals that were scored by the COMPAS tool. It concentrates on defendants in Broward County, as it is a large jurisdiction in a state with strong open-records laws [483]. Moreover, due to Broward County using COMPAS primarily in release/detain decisions prior to a defendant's trial, scores assessed at parole, probation or other stages were discarded. A notable anomaly in the sample is the low amount of defendants screened between June and July 2013 compared to the remaining time span of the COMPAS dataset [46].

- **What data does each instance consist of?**

  Instances represent Broward County defendants scored with COMPAS for risk of recidivism. For each defendant the data provided by ProPublica includes tens of variables (~ 50) summarizing their demographics, criminal record, custody and COMPAS scores.

- **Is there a label or target associated with each instance?**

  Yes. Instances are associated with two target variables (is_recid and is_violent_recid), indicating whether defendants were booked in jail with a criminal offense (potentially violent) that took place after their COMPAS screening but within two years. The definition of recidivism and the two-year cutoff were selected by ProPublica staff to align their audit with definitions by Northpointe [21, 93].

- **Is any information missing from individual instances?**

  Yes. There are several columns where data is missing for one or more instances, including dates when defendants committed the offense (c_offense_date) were incarcerated (c_jail_in) or released (c_jail_out). Missingness in this dataset is not surprising as its curation was a complex endeavour that required cross-referencing information from three separate sources, namely Broward County Sheriff's Office, Broward County Clerk's Office and Florida Department of Corrections. Moreover, Northpointe's response to the ProPublica's study points out important risk factors considered by the COMPAS algorithm that are not present in the dataset, among which the criminal involvement scale, drug problems sub-scale, age at first adjudication, arrest rate and vocational educational scale [212]. Finally, a clear indication of whether defendants were released or detained pretrial seems to be missing.

- **Are relationships between individual instances made explicit?**

  No. While it is plausile for some Broward County defendants to be connected, this information is not available.

- **Are there recommended data splits?**

No.

- **Are there any errors, sources of noise, or redundancies in the dataset?**

  Yes. Clerical errors in records caused incorrect matches between individuals' COMPAS scores and their criminal records, leading to an error rate close to 4% [483]. Moreover, an important temporal trend was spuriously introduced by ProPublica's preprocessing in `compas-scores-two-years.csv` and `compas-scores-two-years-violent.csv`, due to which defendants with a screening date after April 2014 are all recidivists [46]. In terms of redundancies, `compas-scores.csv` contains two identical columns (called decile_score and decile_score.1).

- **Is the dataset self-contained, or does it link to or otherwise rely on external resources?**

  The dataset is **self-contained**.

- **Does the dataset contain data that might be considered confidential?**

  No. However it does contains first names and last names of defendants, connecting them to their criminal history.

- **Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?**

  Yes. The column vr_charge_desc describing violent recidivism charges is one such example.

- **Does the dataset identify any subpopulations (e.g., by age, gender)?**

  Yes. The dataset identifies population by age, sex and race. The curators of the COMPAS dataset maintained the race classifications used by the Broward County Sheriff's Office, identifying individuals as Asian, Black, Hispanic, Native American and White [483]. Age is reported as an integer, sex as either Male or Female. A distribution along these dimensions is reported in Table A.10 which summarizes data in `compas-scores-two-years.csv`. Distributions in remaining files are similar.

- **Is it possible to identify individuals , either directly or indirectly from the dataset?**

  Yes. The dataset reports defendants' first name, last name and date of birth.

- **Does the dataset contain data that might be considered sensitive in any way?**

  Yes. The COMPAS dataset reports individuals' race, criminal history, full name and date of birth.

compas-scores-two-years	
Demographic Caracteristic	**Values**
Percentage of male subjects	80.83%
Percentage of female subjects	19.17%
Percentage of African-American subjects	51.46%
Percentage of Caucasian subjects	33.63%
Percentage of Hispanic subjects	8.67%
Percentage of Asian subjects	0.48%
Percentage of Native American subjects	0.20%
Percentage of people belonging to other races	5.56%
Percentage of people under-19 years old	0.42%
Percentage of people between 20-29 years old	42.41%
Percentage of people between 30-39 years old	28.04%
Percentage of people between 40-49 years old	14.60%
Percentage of people between 50-59 years old	11.00%
Percentage of people between 60-69 years old	3.01%
Percentage of people over-70 years old	0.51%

Table A.10 Demographic Characteristics of compas-scores-two-years.

Collection process

- **How was the data associated with each instance acquired?**

  The data was obtained cross-referencing three sources. From the Broward County Sheriff's Office in Florida, ProPublica obtained COMPAS scores associated with all 18,610 people scored in 2013 and 2014. Defendants' public criminal records were obtained from the Broward County Clerk's Office website matching them based on date of birth, first and last names. The dataset was augmented with jail records provided by the Broward County Sheriff's Office. Finally public incarceration records were downloaded from the Florida Department of Corrections website.

- **What mechanisms or procedures were used to collect the data?**

The original data was plausibly recorded by employees of the Broward County Sheriff's Office, Broward County Clerk's Office, and Florida Department of Corrections. The curators of the COMPAS dataset obtained records from the County Sheriff's Office through a public records request, while data from the County Clerk's Office and the Florida Department of Correction was downloaded from their official website, matching the methodology of a COMPAS validation study [483].

- **If the dataset is a sample from a larger set, what was the sampling strategy?**

  In terms of auditing the COMPAS risk assessment tool, this dataset represents a **convenience sample**, focused on a single county and scoring period 2013–2014. Considering a single county in a state with strong open-records laws reduced the data cross-referencing overhead. Concentrating on recent scores predating the study by 2–3 years kept the study timely and permitted a measurement of recidivism aligned with the one by Northpointe. The fact that Northpointe's response to the ProPublica study only contains minor criticism of the sample (concerning the definition of pretrial defendants [212]) may be interpreted as testimony to its overall quality. More broadly and beyond the COMPAS audit, arrest data as a proxy for crime brings about specific sampling effects, inevitably mediated by law enforcement practices [368, 836].

- **Who was involved in the data collection process and how were they compensated?**

  The original data was plausibly recorded by Broward County and Florida Department of Corrections employees. On ProPublica's side, we assume that key curation choices were made and implemented by four employees credited in the article [21] and accompanying technical report [483], namely Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner. Given the focus on arrest data, the Broward County law enforcement community is also important in the data sampling process.

- **Over what timeframe was the data collected?**

  COMPAS scores are from 2013 and 2014, while jail records cover the period from January 2013 to April 2016. The dataset was first released by ProPublica in May 2016 [646].

- **Were any ethical review processes conducted?**

  Unknown.

- **Was the data collected from the individuals in question directly, or obtained via third parties or other sources?**

  The data was obtained **via third parties**, namely the Broward County Sheriff's Office in Florida through a public records request, from the Broward County Clerk's Office

through the official website and through the Florida Department of Corrections through the official website. Collection from interested individuals would not have been viable.

- **Were the individuals in question notified about the data collection?**

  Likely no. Most of the COMPAS data was publicly available and downloaded from the official websites of Broward County Clerk's Office and the Florida Department of Corrections.

- **Did the individuals in question consent to the collection and use of their data?**

  Likely no. Public availability of arrest/conviction records is associated with collateral consequences that typically damage subjects socially and financially [21, 638].

- **If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses?**

  Likely no.

- **Has an analysis of the potential impact of the dataset and its use on data subjects been conducted?**

  Likely no. We are unaware of analyses specifically focused on the COMPAS dataset. More broadly, public availability of criminal records is related to studies on the employability of offenders [323].

Preprocessing/cleaning/labelling

- **Was any preprocessing/cleaning/labeling of the data done?**

  Yes. Instances were discarded if assessed with COMPAS at parole, probation or other stages in the criminal justice system. This data is unavailable. Moreover, ProPublica published its datasets with accompanying preprocessing code which has become standard [646]. The standard preprocessing removes instances for which (1) arrest dates or charge dates are not within 30 days of the COMPAS assessment, (2) true recidivism cannot be decided, (3) charge degree is not defined as misdemeanor or felony, (4) the COMPAS score is not clearly defined. The remaining COMPAS scores were bucketed into low, medium and high risk.

- **Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data?**

  Yes. The data is available in the official ProPublica github repository [646]. This is an intermediate data artifact, already cross-referenced by ProPublica across three separate sources.

- **Is the software used to preprocess/clean/label the instances available?**

  Yes. The standard preprocessing software can be found in the official ProPublica github repository [646]. The software used to cross-reference data from separate sources is not publicly available.

Uses

- **For what tasks has the dataset been used?**

  The creators used this dataset to audit the COMPAS tool for racial bias. In the literature it has also been used to evaluate the fairness and accuracy of different algorithms and, more broadly, to study definitions of algorithmic fairness.

- **Is there a repository that links to any or all papers or systems that use the dataset?**

  See Appendix A.1.41 for a (non-exhaustive) list of algorithmic fairness works using this resource.

- **What (other) tasks could the dataset be used for?**

  In terms of immediate applications, the dataset could be used to train novel recidivism risk assessment tools. From a methodological perspective, COMPAS may be used in high-stakes domains connected with decision-making about human subjects, including explainable and privacy-preserving ML.

- **Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?**

  From a very narrow perspective, the fact that all defendants with a screening date after April 2014 are recidivists introduces artificially inflated recidivism base rates [46], which would likely be inherited by tools trained on the COMPAS dataset. Moreover, the dataset contains no clear indication concerning pretrial detention or release of defendants. Therefore, researchers must come up with subjective criteria to label individuals as detained or released if they are interested in studying pretrial detention as an intervention deviating from a default course of action [568]. From a broader perspective, the data is likely influenced by historical biases in criminal justice, with differential impact on different communities [21, 368, 836]. Zooming out further, the use of automated risk assessment tools in pretrial decisions is the subject of controversial debate [44] which cannot be overlooked.

- **Are there tasks for which the dataset should not be used?**

Given the above considerations and the narrow geographical scope of the dataset, COMPAS should not be used to train and deploy risk assessment tools for the judicial system. In research settings, users should exercise care in selecting both rows and columns. Bao et al. [43] suggest avoiding the use of COMPAS to demonstrate novel approaches in algorithmic fairness, as considering data without proper context may bring to misleading conclusions which could misguidedly enter the broader debate on criminal justice.

Distribution

- **Is the dataset distributed to third parties outside of the entity on behalf of which the dataset was created?**

  Yes. The COMPAS dataset is publicly available.

- **How is the dataset distributed?**

  The dataset is hosted on ProPublica's official github repository [646].

- **When was the dataset distributed?**

  Since **May 2016**.

- **Is the dataset distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?**

  As of June 2021 the COMPAS dataset is freely distributed under ProPublica's standard ToU [647]. The dataset cannot be republished in its entirety, it cannot be sold, and can only be used for publication if ProPublica's work is properly referenced.

- **Have any third parties imposed IP-based or other restrictions on the data associated with the instances?**

  Likely no.

- **Do any export controls or other regulatory restrictions apply to the dataset or to individual instances?**

  Unknown.

Maintenance

- **Who is supporting/hosting/maintaining the dataset?**

  The dataset is currently hosted and maintained by **ProPublica** on github.

- **How can the owner/curator/manager of the dataset be contacted?**

  The contact for ProPublica's data store is data.store@propublica.org.

- **Is there an erratum?**

  No. There is no official erratum. An external report highlighting anomalies in the data is available [46].

- **Will the dataset be updated?**

  Likely no. In the event of an update, ProPublica's data store ToU specifies users are solely responsible for checking their sites for updates [647]

- **If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances?**

  Unknown.

- **Will older versions of the dataset continue to be supported/hosted/maintained?**

  Unknown.

- **If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?**

  Likely no.

A.3.2 Data Nutrition Label

The following analysis refers to compas-scores-two-years.csv after applying the standard COMPAS preprocessing [646].

Metadata

METADATA	
Filenames	compas-scores-two-years
Format	csv
Url	https://www.propublica.org/datastore/dataset/compas-recidivism-risk-score-data-and-analysis
Domain	Law
Keywords	risk assessment, pretrial, recidivism
Type	Tabular
Rows	6,172
Columns	57
% missing cells	5%
Rows with missing cells	100%
License	ProPublica's ToU [647]
Released	May 2016
Range	2013-2014 for COMPAS scores, 2013-2016 for arrest and detention history.
Description	Dataset curated by ProPublica to audit COMPAS software for racial biases, focusing on Broward County 2013–2014.

Table A.11 Metadata of COMPAS dataset.

Provenance

PROVENANCE	
Source	
Name	Broward County Sheriff's Office
Url	http://www.sheriff.org/
email	//
Name	Broward County Clerk's Office
Url	https://www.browardclerk.org
email	Eclerk@browardclerk.org
Name	Florida Department of Corrections
Url	http://www.dc.state.fl.us/
email	FDCCitizenServices@fdc.myflorida.com
Authors	
Names	Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner
Url	https://www.propublica.org/datastore/dataset/ compas-recidivism-risk-score-data-and-analysis
email	data.store@propublica.org

Table A.12 Provenance of COMPAS dataset.

Variables

VARIABLES	
id	Unique identifier assigned by the authors
name	Defendant's first and last name
first	Defendant's first name
last	Defendant's last name
screening_date	Day defendant was scored by COMPAS
sex	Defendant's sex
dob	Defendant's date of birth
age	Defendant's age
age_cat	Age quantization: less than 25 25-45 > 45
race	Defendant's race: African-American Asian Caucasian Hispanic Native American Other
juv_fel_count	Number of juvenile felonies
decile_score	COMPAS recidivism score (10-point scale)
juv_misd_count	Number of juvenile misdemeanors
juv_other_count	Number of other juvenile convictions (not considering misdemeanor and felonies)
priors_count	Number of prior crimes
days_b_screening_arrest	Days between imprisonment (c_jail_in) and COMPAS screening (compas_screening_date)

Table A.13 Variables of COMPAS dataset (1/3).

VARIABLES	
c_jail_in	Date of imprisonment
c_jail_out	Date of release
c_case_number	Alpha-numeric case identifier
c_offense_date	Date on which the offense was committed
c_arrest_date	Date on which defendant was arrested
c_days_from_compas	Days elapsed between offense/arrest and the date of COMPAS screening
c_charge_degree	Degree of charge: F (felony) M (misdemeanor)
c_charge_desc	Textual description of charge
is_recid	Binary indication of recidivism.
r_case_number	Alpha-numeric case identifier for recidivist offense
r_charge_degree	Degree of recidivist charge
r_days_from_arrest	Days elapsed between date of recidivist offense (r_offense_date) and date of recidivist incarceration (r_jail_in)
r_offense_date	Date of recidivist offense
r_charge_desc	Textual description of recidivist charge
r_jail_in	Date of incarceration for recidivist offense
r_jail_out	Date of release for recidivist offense

Table A.14 Variables of COMPAS dataset (2/3).

VARIABLES	
violent_recid	Unknown; all nan
is_violent_recid	Binary indication of violent recidivism. If true, then is_recid is true.
vr_case_number	Alpha-numeric case identifier for violent recidivist offense
vr_charge_degree	Degree of violent recidivist offense
vr_offense_date	Date of violent recidivist offense
vr_charge_desc	Textual description of the violent recidivist charge
type_of_assessment	Type of COMPAS assessment - all 'Risk of Recidivism'.
decile_score_1	Identical to decile_score
score_text	Quantization of decile_score: LOW (1-4) MEDIUM (5-7) HIGH (8-10).
screening_date	Identical to compas_screening_date
v_type_of_assessment	Type of COMPAS violent assessment - all 'Risk of Violence'.
v_decile_score	COMPAS violent recidivism score (10-point scale)
v_score_text	Quantization of v_decile_score: LOW (1-4) MEDIUM (5-7) HIGH (8-10).
v_screening_date	Identical to compas_screening_date.
in_custody	Unknown
out_custody	Unknown
priors_count.1	Identical to priors_count.
start	Unknown
end	Unknown
event	Unknown
two_year_recid	Unknown

Table A.15 Variables of COMPAS dataset (3/3).

Statistics

STATISTICS						
Ordinal						
name	type	count	unique	mostFrequent	leastFrequent	missing
id	int	6,172	6,172	multiple	multiple	0
screening_date	date	6,172	685	2013-04-20	multiple	0
dob	date	6,172	4,830	multiple	multiple	0
age_cat	string	6,172	3	25 - 45	>45	0
c_jail_in	date	6,172	6,172	multiple	multiple	433
c_jail_out	date	6,172	6,161	2013-09-14 05:58:00	multiple	433
c_offense_date	date	6,172	737	multiple	multiple	1388
c_arrest_date	date	6,172	417	2013-02-06	multiple	8425
r_offense_date	date	6,172	1,041	2014-12-08	multiple	3,182
r_jail_in	date	6,172	928	multiple	multiple	4,175
r_jail_out	date	6,172	893	multiple	multiple	4,175
vr_offense_date	date	6,172	505	2015-08-15	multiple	5,480
v_score_text	string	6,172	3	Low	High	0
v_screening_date	date	6,172	685	2013-04-20	multiple	0
score_text	string	6,172	3	Low	High	0
screening_date	date	6,172	685	2013-04-20	multiple	0
in_custody	date	6,172	1,087	multiple	multiple	0
out_custody	date	6,172	1,097	2020-01-01	multiple	0

Table A.16 Ordinal variables statistics of COMPAS dataset

Categorical

name	type	count	unique	mostFrequent	leastFrequent	missing
name	string	6,172	9,128	mutiple	multiple	0
first	string	6,172	2,493	michael	multiple	0
last	string	6,172	3,465	williams	multiple	0
sex	string	6,172	2	Male	Female	0
race	string	6,172	6	African-American	Native American	0
c_case_number	string	6,172	6,172	multiple	multiple	0
c_charge_desc	string	6,172	390	Battery	multiple	5
c_charge_degree	string	6,172	2	F	M	0
r_case_number	string	6,172	2,991	multiple	multiple	3,182
r_charge_desc	string	6,172	319	Possess Cannabis/ 20 Grams Or Less	multiple	3,228
r_charge_degree	string	6,172	11	(M1)	(F5)	0
vr_case_number	string	6,172	693	multiple	multiple	5,480
vr_charge_desc	string	6,172	82	Battery	multiple	5,480
vr_charge_degree	string	6,172	10	(M1)	(F5)	5,480
type_of_assessment	string	6,172	1	Risk of Recidivism	Risk of Recidivism	0
v_type_of_assessment	string	6,172	1	Risk of Violence	Risk of Violence	0
is_recid	binary	6,172	2	0	1	0
is_violent_recid	binary	6,172	2	0	1	0
event	binary	6,172	2	0	1	0
two_year_recid	binary	6,172	2	0	1	0

Table A.17 Categorical variables statistics of COMPAS dataset

Quantitative									
name	type	count	min	median	max	mean	stdDev	miss	zeros
age	int	6,172	18	31	96	34.53	11.73	0	0
juv_fel_count	int	6,172	0	0	20	0.06	0.46	0	5,964
juv_misd_count	int	6,172	0	0	13	0.09	0.50	0	5,820
juv_other_count	int	6,172	0	0	9	0.11	0.47	0	5,711
priors_count	int	6,172	0	1	38	3.25	4.74	0	2,085
days_b_screening_arrest	int	6,172	-30.0	-1	30.0	-1.74	5.08	0	1,379
c_days_from_compas	int	6,172	0	1	9,485	24.90	276.81	0	869
r_days_from_arrest	int	6,172	-1	0	993	20.10	76.54	4,175	1,452
decile_score	int	6,172	1	4	10	4.42	2.84	0	0
v_decile_score	int	6,172	1	3	10	3.64	2.49	0	0
start	int	6,172	0	0	937	13.32	50.14	0	3,485
end	int	6,172	0	539	1,186	555.05	400.26	0	1

Table A.18 Quantitative variables statistics of COMPAS dataset.

A.4 German Credit

Key references include Grömping [329], Häußler [386], UCI Machine Learning Repository [774, 776].

A.4.1 Datasheet

Motivation

- **For what purpose was the dataset created?**

  This dataset was created to study the problem of automated credit decisions at a regional Bank in southern Germany.

- **Who created the dataset and on behalf of which entity?**

  The dataset was created at a regional Bank of southern Germany (most likely Hypo Bank) and first used by Walter Häußler in the late 1970s as part of his PhD thesis. Hans Hofmann, affiliated with Universität Hamburg at the time, is credited as dataset source [774]. Presumably, he donated the dataset to the European Statlog project and a representative of Strathclyde University donated it to UCI [329].

- **Who funded the creation of the dataset?**

  The first known work using the dataset describes it as originating from a regional Bank of southern Germany [386]. Given the affiliation of the author is Hypo Bank, which fit the description at the time, we assume the dataset was collected, curated and funded at Hypo Bank.

Composition

- **What do the instances that comprise the dataset represent?**

  Instances represent Hypo bank **loan recipients** from 1973–1975.

- **How many instances are there in total?**

  The dataset consists of **1,000** instances.

- **Does the dataset contain all possible instances or is it a sample of instances from a larger set?**

  In principle this is a **convenience sample**, consisting of people who were deemed creditworthy by a bank clerk. A representative sample stemming from indiscriminate credit grants would not have been viable [386]. However, if the envisioned application was *post-screening* credit decisions, the influence of this selection bias would be

reduced. Finally loan recipients associated with delayed payment or loan default ("bad credit") are oversampled (30%).

- **What data does each instance consist of?**

  For each instance, 13 categorical and 7 quantitative variables are provided, summarizing their financial situation, credit history, and personal situation, including housing, number of liable people, and a mixed variable encoding marital status and sex. A more through description is deferred to Tables A.22-A.24.

- **Is there a label or target associated with each instance?**

  Yes. A binary label encodes whether loan recipients punctually payed each installment ("good credit") or not ("bad credit"). The latter label includes a range of situations from delayed payment up to loan default.

- **Is any information missing from individual instances?**

  No. No cell is missing, however the variable "property" has a level jointly encoding the conditions "no property" and "unknown". A similar joint encoding exists for "savings", so some values may actually be deemed missing for these variables.

- **Are relationships between individual instances made explicit?**

  No. There are no known relationships between instances.

- **Are there recommended data splits?**

  No.

- **Are there any errors, sources of noise, or redundancies in the dataset?**

  Yes. The dataset documentation is filled with errors, so that several levels of categorical variables do not correspond to what they should according to the official documentation from UCI Machine Learning Repository [774]. This is not necessarily an issue if one is purely interested in the evaluation of a method. For example, according to the official documentation, a majority of loan recipients are foreign workers, while in reality this should appear rather strange and indeed is not true [329]. Computationally, this will make no difference, as the input to a machine learning method will remain the same. However if one is interested to the context surrounding the data, as should be the case with fairness research, the wrong encoding poses several problems. The most significant problem is the impression that one can retrieve people's sex from the joint sex-marital-status encoding, which is simply false as a single level corresponds to both single males and divorced/separated/married females [329]. Despite this information being available since 2019, the fairness community does not seem to have taken notice.

Several experiments of algorithmic fairness on this dataset consider the protected attribute "sex" (sometimes even called "gender"). These experiments are part of work recently published in the most reputable venues for fairness research (Appendix A.1.73). More mistakes in the documentation of eight variables and the relative errata are outlined in Grömping [329]. A clean version of the dataset is available at UCI Machine Learning Repository [776].

- **Is the dataset self-contained, or does it link to or otherwise rely on external resources?**

  The dataset is **self-contained**.

- **Does the dataset contain data that might be considered confidential?**

  Yes. The dataset summarizes customers' financial and personal situation, including past credit history.

- **Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?**

  No.

- **Does the dataset identify any subpopulations?**

  Yes. The dataset identifies subpopulation by age and sex. Sex is jointly encoded with marital status and cannot be retrieved, contrary to documentation accompanying the dataset [774]. A summary based on amended documentation [329] is presented in Table A.19.

- **Is it possible to identify individuals, either directly or indirectly, from the dataset?**

  Likely no, especially given the fact that these records data back to almost 50 years ago. Also, important variables for re-identification, such as ZIP cose and date of birth are missing and many other variables are bucketed.

- **Does the dataset contain data that might be considered sensitive in any way?**

  Yes. For each instance, the dataset encodes sex, marital status and financial situation.

Collection process

- **How was the data associated with each instance acquired?**

  The data was collected by Hypo bank clerks. Some variables were observable (e.g. credit history with the bank), other variables were reported by subjects (e.g. loan purpose).

Demographic Caracteristic	Values
Percentage of people under-19 years old	0.20%
Percentage of people between 20-29 years old	36.70%
Percentage of people between 30-39 years old	33.20%
Percentage of people between 40-49 years old	17.60%
Percentage of people between 50-59 years old	7.20%
Percentage of people between 60-69 years old	4.40%
Percentage of people over-70 years old	0.70%
Percentage of people who are male : divorced/separated	5.00%
Percentage of people who are female : non-single or male : single	31.00%
Percentage of people who are male : married/widowed	54.80%
Percentage of people who are female : single	9.20%

Table A.19 Demographic characteristics of the German credit dataset.

- **What mechanisms or procedures were used to collect the data?**

  Unknown.

- **If the dataset is a sample from a larger set, what was the sampling strategy?**

  The so-called "bad credits" are heavily oversampled to make the classification problem more balanced. A natural selection bias is present in the data, as it only consist of applicants who were deemed creditworthy and were thus granted a loan.

- **Who was involved in the data collection process and how were they compensated?**

  The data was likely collected by Hypo bank clerks. Walter Häußler was likely involved in sample selection.

- **Over what timeframe was the data collected?**

  The dataset covers loans granted in the period **1973–1975**. Its first publicly-known use dates back to 1979 [386]. It became publicly available in November 1994 [774].

- **Were any ethical review processes conducted?**

  Unknown.

- **Was the data collected from the individuals in question directly, or obtained via third parties or other sources?**

Likely both. Some variables were necessarily collected from loan applicants (e.g. loan purpose), while other variables were likely available from bank records (e.g. credit history with the bank).

- **Were the individuals in question notified about the data collection?**

  Individuals provided some of this data as part of a loan application. Collection and notification practices for variables like credit history are unclear.

- **Did the individuals in question consent to the collection and use of their data?**

  Likely yes, for the purposes of the immediate credit decision. However it seems implausible they agreed to their data becoming publicly available in an anonymized fashion.

- **If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses?**

  Likely no.

- **Has an analysis of the potential impact of the dataset and its use on data subjects been conducted?**

  Unknown.

Preprocessing/cleaning/labelling

- **Was any preprocessing/cleaning/labeling of the data done?**

  Yes. Some instances were discarded. Remaining instances were associated with a binary label according to compliance with the contract. Bucketing took place on several variables, including balance on checking and savings account (A1, A6) and duration of current employment (A7). Sex and marital status were jointly coded (A9).

- **Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data?**

  Unknown.

- **Is the software used to preprocess/clean/label the instances available?**

  Likely no.

Uses

- **For what tasks has the dataset been used?**

The dataset was originally used to study the problem of automated credit scoring [386]. Similarly to the Adult dataset, since becoming publicly available it has been used as a benchmark in various machine learning fields.

- **Is there a repository that links to any or all papers or systems that use the dataset?**

  Yes. A selection of early works (pre-2005) using this dataset can be found in UCI Machine Learning Repository [774]. A more recent list is available under the beta version of the UCI ML Repository.[21] See Appendix A.1.73 for a (non-exhaustive) list of algorithmic fairness works using this resource.

- **What (other) tasks could the dataset be used for?**

  The German Credit could be used in fields that concentrate on socially relevant goals and require socially relevant data, such as privacy and explainability. The task at hand is always credit scoring.

- **Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?**

  Contrary to documentation accompanying the dataset [774], the sex of loan recipients cannot be reliably retrieved. Works of algorithmic fairness should not use this feature.

- **Are there tasks for which the dataset should not be used?**

  In its most common version [774] the German Credit dataset should not be used in works of explainability/interpretability as the incorrect documentation would result in counter-intuitive explanations. The 2019 version [776] associated with the erratum [329] is recommended.

Distribution

- **Is the dataset distributed to third parties outside of the entity on behalf of which the dataset was created?**

  Yes. The dataset is publicly available [774]

- **How is the dataset distributed?**

  The dataset is available as a **csv file**.

- **When was the dataset distributed?**

  The dataset was released to the UCI ML Repository in **November 1994**.

[21] https://archive-beta.ics.uci.edu/ml/datasets/144

- **Is the dataset distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?**

  Yes. The UCI ML repository has a citation policy.

- **Have any third parties imposed IP-based or other restrictions on the data associated with the instances?**

  Likely no. We are unaware of any IP-based restrictions.

- **Do any export controls or other regulatory restrictions apply to the dataset or to individual instances?**

  Unknown.

Maintenance

- **Who is supporting/hosting/maintaining the dataset?**

  The dataset is hosted and maintained by the **UCI Machine Learning Repository** [774]. A clean and well-documented version of the same dataset donated by Ulrike Gromping [776] is also available on the same repository.

- **How can the owner/curator/manager of the dataset be contacted?**

  The dataset donor, Hans Hofmann retired in 2008. Comments and inquiries for UCI may be sent to ml-repository@ics.uci.edu.

- **Is there an erratum?**

  Yes. A clean data release [776] and accompanying report [329] are available online.

- **Will the dataset be updated?**

  Likely no. The recently released South German Credit Data Set [776] may be considered an update.

- **If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances?**

  Unknown.

- **Will older versions of the dataset continue to be supported/hosted/maintained?**

  Unless otherwise indicated, both the new [776] and the old version [774] of the German Credit dataset will remain hosted on the UCI ML Repository in its current version.

- **If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?**

Unknown.

A.4.2 Data Nutrition Label

For the sake of correctness, we report redacted information based on the new South German Credit Data Set [776] and accompanying documentation [329].

Metadata

METADATA	
Filenames	SouthGermanCredit
Format	.asc
Url	https://archive.ics.uci.edu/ml/datasets/South+German+Credit
Domain	Economics
Keywords	credit scoring, Germany, loan, classification
Type	Tabular
Rows	1000
Columns	21
% missing cells	0%
Rows with missing cells	0%
License	UCI Repository citation policy
Released	November 2019
Range	1973-1975
Description	This dataset encodes socio-economical features of loan recipients from a bank in southern Germany, along with binary variable encoding whether they punctually payed every installment, which is he target of a classification task.

Table A.20 Metadata of South German Credit dataset.

Provenance

PROVENANCE	
Source	
Name	Walter Häußler
Url	https://archive.ics.uci.edu/ml/datasets/Statlog+%28German+Credit+Data%29
email	//
Authors	
Names	Ulrike Grömping
Url	https://archive.ics.uci.edu/ml/datasets/South+German+Credit
email	groemping@bht-berlin.de

Table A.21 Provenance of South German Credit dataset

Variables

VARIABLES	
status	Checking account balance (in Deutsche Mark) 1 (no checking account) 2 (< 0 DM) 3 ($0 \leq ... < 200$ DM) 4 (≥ 200 DM)
duration	Credit duration (in months)
credit_history	Applicant's credit history 0 (delay in past payments) 1 (critical account/other credits elsewhere) 2 (no credits taken/all credits paid back duly) 3 (existing credits paid back duly till now) 4 (all credits at this bank paid back duly)
purpose	Purpose of loan 0 (other) 1 (new car) 2 (used car) 3 (furniture/equipment) 4 (radio/television) 5 (domestic appliances) 6 (repairs) 7 (education) 8 (vacation) 9 (retraining) 10 (business)
amount	Credit amount (result of unknown monotonic transformation)

Table A.22 Variables of South German Credit dataset (1/3).

VARIABLES	
savings	Savings account balance (in Deutsche Mark) 1 (unknown/ no savings account) 2($<$ 100 DM) 3 ($100 \leq ... < 500$ DM) 4 ($500 \geq ... < 1000$ DM) 5 ($\geq$ 1000 DM)
employment_duration	Duration of applicant's current employment 1 (unemployed) 2 ($<$ 1 year) 3 ($1 \leq ... < 4$ years) 4 ($4 \leq ... < 7$ years) 5 ($\geq$ 7 years)
installment_rate	Installment amount to disposable income ratio $[\%]$ 1 ($\geq$35) 2 ($25 \leq ... < 35$) 3 ($20 \leq ... < 25$) 4 ($<$ 20)
personal_status_sex	Joint encoding of sex and marital status of applicant 1 (male - divorced/separated) 2 (female - non single or male - single) 3 (male - married/widowed) 4 (female - single)
other_debtors	Presence of co-debtor or guarantor 1 (none) 2 (co-applicant) 3 (guarantor)
present_residence	Years living at current address 1 ($<$ 1 year) 2 ($1 \leq ... < 4$ years) 3 ($4 \leq ... < 7$ years) 4 ($\geq$ 7 years)
property	Applicant's most valuable property 1 (unknown / no property) 2 (car or other) 3 (building soc. savings agr / life insurance) 4 (real estate)

Table A.23 Variables of South German Credit dataset (2/3).

VARIABLES	
age	Applicant's age (years)
other_installment_plans	Installment plans with other banks 1 (bank) 2 (stores) 3 (none)
housing	Type of housing 1 (for free) 2 (rent) 3 (own)
number_credits	Number of credits (ongoing or past, including current) with this bank 1 (1) 2 (2-3) 3 (4-5) 4($\geq$ 6)
job	Applicant's job and emplyability 1 (unemployed/ unskilled - non-resident) 2 (unskilled - resident) 3 (skilled employee / official) 4 (manager / self-empl. / highly qualif. employee)
people_liable	Number of people who financially depend on the applicant 1 (3 or more) 2 (0 to 2)
telephone	Presence of telephone landline registered under applicant's name (2) or not (1)
foreign_worker	Foreign worker (1) or not (2)
credit_risk	Punctually payed back every installment (1) or not (2)

Table A.24 Variables of South German Credit dataset (3/3).

Statistics

STATISTICS						
Ordinal						
name	type	count	unique	mostFrequent	leastFrequent	missing
status	string	1000	4	4 ($\geq$200)	3 (0$\leq$...$<$200)	0
savings	string	1000	5	1 (unk./no sav.)	4 (500$\leq$...$<$1000)	0
employment_duration	string	1000	5	3 (1$\leq$...$<$4)	1 (unemployed)	0
installment_rate	string	1000	4	4 ($<$20)	1 $\geq$35	0
present_residence	string	1000	4	4 ($\geq$7 yrs)	1 ($<$ 1 yr)	0
number_credits	string	1000	4	1 (1)	4 ($\geq$6)	0
people liable	string	1000	2	2 (0 to 2)	1 (3 or more)	0

Table A.25 Ordinal variables statistics of South German Credit dataset

Categorical						
name	type	count	unique	mostFrequent	leastFrequent	missing
credit_history	string	1000	5	2 (no credits taken)	0 (delay)	0
purpose	string	1000	11	3 (furnit/equip)	8 (vacation)	0
status_sex	string	1000	4	3 (male-marr/widow)	1 (male-divorc/separ)	0
other_debtors	string	1000	3	1 (none)	2 (co-appliant)	0
property	string	1000	4	3 (build. soc. savings)	4 (real estate)	0
other_plans	string	1000	3	3 (none)	2 (stores)	0
housing	string	1000	3	2 (rent)	3 (own)	0
job	string	1000	4	3 (skilled empl/offic)	1 (unempl/ non-res)	0
telephone	string	1000	2	1 (no)	2 (yes)	0
foreign_worker	string	1000	2	2 (no)	1 (yes)	0
credit_risk	string	1000	2	1 (good)	0 (bad)	0

Table A.26 Categorical variables statistics of South German Credit dataset

Quantitative									
name	type	count	min	median	max	mean	stdDev	miss	zeros
duration	number	1000	4	18	72	20.90	12.06	0	0
amount	number	1000	250	2319.50	18424	3271.25	2822.75	0	0
age	number	1000	19	33	75	35.54	11.35	0	0

Table A.27 Quantitative variables statistics of South German Credit dataset.

Appendix B

Supplementary Materials to Chapter 4

In this appendix, we report additional visualizations and analyses for Section 4.1.

B.1 Aggregator Influence on Premiums

B.1.1 Methods

Based on reports on aggregators and their typical business model, we expect their influence on quoted prices to be null or negligible [394, 416]. In this section, we verify that the key pricing trends obtained on the comparison website are also present on an individual company website. Considering a single company and a single product (`c1/a`), we repeat our data collection procedure, with doubly-nested randomization and control (summarized in Figure 4.2), directly on the company website. We concentrate on a subset of our dataset, comprising 32-year-old drivers with BMS classes 0 and 4. We choose this subset since (1) `c1/a` is always present in the respective aggregator result pages, allowing for a direct comparison; and (2) this is the most representative subset in our sample, as very young drivers and BMS classes above 4 are quite rare among Italian RCA subscribers [169, 709]. The resulting dataset consists of 288 regular quotes and 24 control quotes, gathered in the second half of December, 2020.

Concentrating on `c1/a`, we mimic the analyses from Section 4.1.5, i.e. an overview of the most important factors and a discrimination analysis focused on protected pairs. As the dataset collected from the aggregator predates this one by six months, we do not attempt a rigorous characterization of the comparison website effect in terms of fees and discounts. Instead, we are mainly interested in evaluating whether the key trends from Section 4.1.5 are confirmed.

Table B.1 Consistency in birthplace- and gender-related trends in `c1/a` pricing across the two datasets. Differences in price for protected pairs (δ). Rows are consistent with Table 4.3. Columns report the percentage of ties within a 5€ tolerance threshold (Ties_5), 5th, 95th percentile and median difference ($\eta_{.05}(\delta)$, $\eta_{.05}(\delta)$ and $m(\delta)$), as computed from the dataset gathered on the aggregator (Aggr.) and directly on the company website six months later (Comp.). Trends are stable across both datasets.

		Ties_5		$\eta_{.05}(\delta)$		$m(\delta)$		$\eta_{.95}(\delta)$	
Attribute	Pairs	Comp.	Aggr.	Comp.	Aggr.	Comp.	Aggr.	Comp.	Aggr.
birthplace	Rome vs Milan	29%	17%	-1 €	-1 €	16 €	23 €	121 €	197 €
birthplace	Naples vs Milan	20%	17%	-1 €	-1 €	144 €	143 €	294 €	435 €
birthplace	Romania vs Milan	42%	38%	-6 €	-10 €	0 €	6 €	124 €	211 €
birthplace	Ghana vs Milan	8%	0%	0 €	92 €	213 €	243 €	1116 €	867 €
birthplace	Laos vs Milan	8%	0%	0 €	87 €	213 €	243 €	1116 €	914 €
gender	F vs M	80%	76%	0 €	-5 €	0 €	0 €	14 €	19 €
noise control		96%	100%	0 €	0 €	0 €	0 €	0 €	0 €

B.1.2 Results

Figure B.1 depicts the effect of each factor, as an average price for profiles sharing a factor level, across each of the remaining factors, similarly to Figure 4.3. Overall trends are

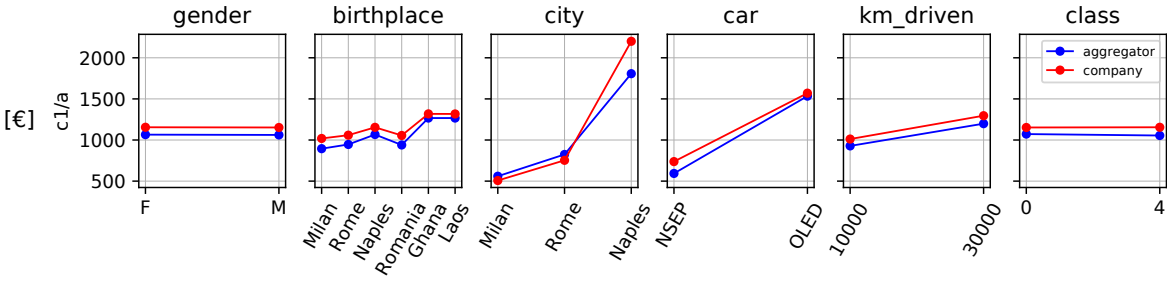


Fig. B.1 Consistency of general trends in `c1/a` pricing across the two datasets. Each panel represents a different factor, with its levels on the x axis. The y axis depicts the mean price for all profiles with a given factor level, gathered on the aggregator (blue) and company website (red). Trends are stable across both datasets, despite having been collected six months apart.

confirmed for every factor across all levels.

Table B.1 reports summary statistics for protected pairs on both datasets; corresponding histograms are available in Figure B.2. Similarly to Table 4.3, we consider both birthplace- and gender-protected pairs in rows 1-6, with a final row focused on control pairs. In each row, we report the frequency of ties within a 5€ tolerance threshold (Ties_5), along with the median, 5th and 95th percentiles, labelled $m(\delta)$, $\eta_{.05}(\delta)$ and $\eta_{.95}(\delta)$ respectively.

Overall we find stable trends across both datasets, as summarized in Table B.1.

- About 80% of gender-protected pairs are tied. Ties are less frequent between birthplace-protected pairs within the EU (17%-42%) and very rare when comparing drivers born in Milan with their counterparts born in Ghana or Laos (0%-8%).
- $\eta_{.05}(\delta)$ is weakly (if at all) negative, showing that the baseline factor level (Milan for birthplace, male for gender) is rarely at a disadvantage.
- $m(\delta)$ is similar in both datasets, confirming a systematic and sizeable financial disadvantage for drivers born in Naples, Ghana and Laos ($m(\delta) > 100$€).
- $\eta_{.95}(\delta)$ is always larger than 100€ for birthplace-protected pairs, reaching a 1,000€ surcharge for Ghana and Laos.
- noise control shows minimal differences for identical queries.

In sum, these results show that the effect of the comparison website on the prices quoted in its result pages (if any) is modest in comparison with the effect of pricing algorithms employed by company `c1`. As a final remark, it is worth highlighting the strong financial disadvantage measured for Laos-born drivers despite the small number of Laos citizens residing in Italy [391] and available to company `c1` to infer the "effect" of this feature in risk models.[1]

Figure B.2 depicts histograms for price differences quoted to protected pairs of profiles for the insurance product labelled `c1/a`, reporting both the price obtained on the aggregator (blue) and the price obtained on the company website (orange). Rows are consistent with Table B.1. These results pertain to a subset of the full sample, as described above. Despite the fact that the aggregator dataset predates the company dataset by six months, key trends are stable.

[1] In this case, we can likely rule out that the feature is being used as a proxy for the country where drivers learned to drive, since the company website explicitly queries the year of arrival in Italy, and our input, 2004, predates by 2 years the driver's license issue date.

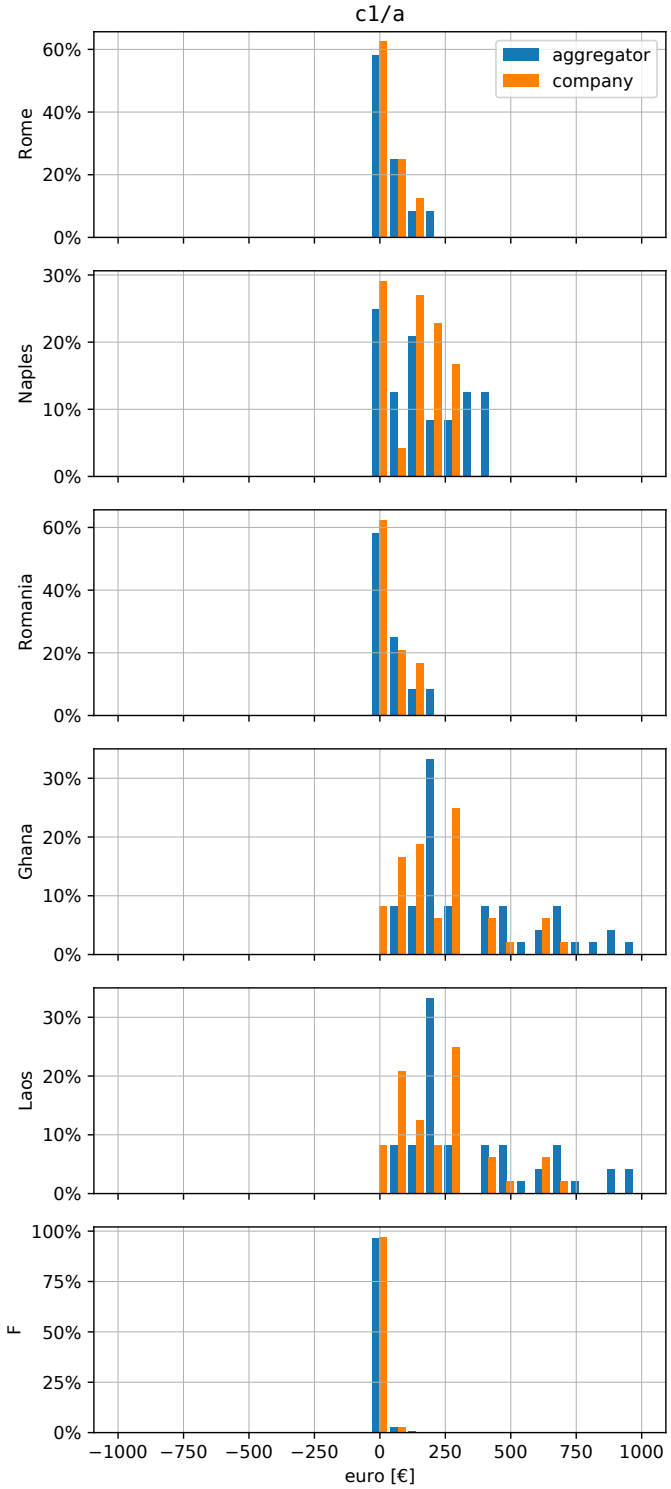


Fig. B.2 Consistency of trends for birthplace- and gender-based discrimination on company website and aggregator. Histogram of differences in `c1/a` quote provided on aggregator (blue) and on company website (orange) to different protected pairs. Rows are consistent with Figure 4.4. The x axis is clipped between -1000 and 1000 €. Key trends are stable, despite the fact that datasets have been collected six months apart.

Appendix C

Supplementary Materials to Chapter 5

In this appendix, we describe the SLD and HDy quantification methods, we report the full proof of Proposition 2, and we present the results of fairness measurements based on SVM rather than LR, along with the pseudocode for three protocols from Chapter 5.

C.1 The SLD Method

SLD [690] produces prevalence estimates $\hat{p}_{\sigma}^{\mathrm{SLD}}(s)$ iteratively, using EM algorithms. In detail, given two sets, L and U, where the former represents the *labelled* one (training set) and the latter represents the *unlabelled* one (test set). The method iterates until convergence (i.e., the difference between the prevalence estimated across two consecutive iterations is less than a tolerance factor ε –we use $\varepsilon = 1e-4$) or until a maximum number of iterations is reached. The pseudocode describing SLD is as follows:

Input : Class prevalence values $p_L(s)$ on L;
Posterior probabilities $\pi_s(\mathbf{x}_i)$, for all $\mathbf{x}_i \in U$;
Output : Estimates $\hat{p}_U(s)$ of class prevalence values on U;

`/* Initialisation */`
$t \leftarrow 0$;
for $s \in S$ **do**
  $\hat{p}_U^{(t)}(s) \leftarrow p_L(s)$;
  for $\mathbf{x}_i \in U$ **do**
    $\Pr^{(t)}(s|\mathbf{x}_i) \leftarrow \pi_s(\mathbf{x}_i)$;
  end
end
`/* Main Iteration Cycle */`
while *stopping condition* = *false* **do**
  $t \leftarrow t+1$;
  for $s \in S$ **do**
    for $\mathbf{x}_i \in U$ **do**

$$\Pr^{(t)}(s|\mathbf{x}_i) \leftarrow \frac{\frac{\hat{p}_U^{(t-1)}(s)}{\hat{p}_U^{(0)}(s)} \cdot \overset{(0)}{\Pr}(s|\mathbf{x}_i)}{\sum_{s \in S} \frac{\hat{p}_U^{(t-1)}(s)}{\hat{p}_U^{(0)}(s)} \cdot \overset{(0)}{\Pr}(s|\mathbf{x}_i)}$$

    end
    $\hat{p}_U^{(t)}(s) \leftarrow \frac{1}{|U|} \sum_{\mathbf{x}_i \in U} \overset{(t)}{\Pr}(s|\mathbf{x}_i)$
  end
end
`/* Generate output */`
for $s \in S$ **do**
  $\hat{p}_U^{\mathrm{SLD}}(s) \leftarrow \hat{p}_U^{(t)}(s)$
end

Pseudocode 2: The SLD algorithm [690].

C.2 The HDy Method

HDy [317] measures the divergence between two distributions of posterior probabilities (i.e., as returned by a calibrated classifier) v and u in terms of the Hellinger Distance (HD), defined as

$$\text{HD}(v,u) = \sqrt{\int \left(\sqrt{v(x)} - \sqrt{u(x)}\right)^2 dx}$$

The HD between two continuous distributions v and u is typically approximated by discretizing v and u across bins and then integrating

$$\hat{\text{HD}}(V,U) = \sqrt{\sum_{i=1}^{b} \left(\sqrt{\frac{|V_i|}{|V|}} - \sqrt{\frac{|U_i|}{|U|}}\right)^2}$$

with V and U the discrete distributions, b the number of bins and V_i, U_i representing the frequency in the ith bin for each distribution, respectively.

The method seeks the α parameter that yields the smallest distance between the validation distribution V (typically, a held-out split of the training set that has not been used to train the classifier) and the unlabelled distribution U, i.e.,

$$\alpha^* = \arg\min_{\alpha \in [0,1]} \hat{\text{HD}}(V^\alpha, U)$$

where V^α is the mixture of the positive distribution ($V^{S=1}$) and the negative distribution ($V^{S=0}$) defined by

$$V^\alpha(x) = (1-\alpha) \cdot V^{S=0}(x) + \alpha \cdot V^{S=1}(x)$$

HDy returns α^* as the sought positive class prevalence

$$\hat{p}_\sigma^{\text{HDy}}(1) = \alpha^*$$

Since the number of bins b could have a significant impact on the calculation, one typically returns the median of the distribution of the best α's found for a range of b's (in our case, we explore $b \in [10, 20, 30, \ldots, 110]$).

C.3 Proof of Proposition 2

We show that Equation 5.2 and Equation 5.15 are equivalent when the latter is instantiated by prevalence estimates given by PCC:

$$\hat{\mu}^{\text{PCC}}(s) = \hat{p}^{\text{PCC}}_{\mathscr{D}_3^{\oplus}}(s) \frac{p_{\mathscr{D}_3}(\oplus)}{\hat{p}^{\text{PCC}}_{\mathscr{D}_3^{\oplus}}(s) p_{\mathscr{D}_3}(\oplus) + \hat{p}^{\text{PCC}}_{\mathscr{D}_3^{\ominus}}(s) p_{\mathscr{D}_3}(\ominus)}$$

The terms in the denominator can be written as

$$\begin{aligned}\hat{p}^{\text{PCC}}_{\mathscr{D}_3^{\oplus}}(s) &= \frac{\sum_{\mathbf{x}_i \in \mathscr{D}_3^{\oplus}} \pi_s(\mathbf{x}_i)}{|\mathscr{D}_3^{\oplus}|} \\ &= \frac{\sum_{\mathbf{x}_i} \pi_s(\mathbf{x}_i) h_{\oplus}(\mathbf{x}_i)}{\sum_{\mathbf{x}_i} h_{\oplus}(\mathbf{x}_i)}\end{aligned}$$

$$\hat{p}^{\text{PCC}}_{\mathscr{D}_3^{\ominus}}(s) = \frac{\sum_{\mathbf{x}_i} \pi_s(\mathbf{x}_i)(1 - h_{\oplus}(\mathbf{x}_i))}{\sum_{\mathbf{x}_i}(1 - h_{\oplus}(\mathbf{x}_i))}$$

$$p_{\mathscr{D}_3}(\oplus) = \frac{\sum_{\mathbf{x}_i} h_{\oplus}(\mathbf{x}_i)}{|\mathscr{D}_3|}$$

$$p_{\mathscr{D}_3}(\ominus) = \frac{\sum_{\mathbf{x}_i}(1 - h_{\oplus}(\mathbf{x}_i))}{|\mathscr{D}_3|}$$

Plugging them into the denominator yields

$$\begin{aligned}\hat{\mu}^{\text{PCC}}(s) &= \hat{p}^{\text{PCC}}_{\mathscr{D}_3^{\oplus}}(s) \frac{p_{\mathscr{D}_3}(\oplus)}{\frac{\sum_{\mathbf{x}_i} \pi_s(\mathbf{x}_i)}{|\mathscr{D}_3|}} \\ &= \frac{\sum_{\mathbf{x}_i} \pi_s(\mathbf{x}_i) h_{\oplus}(\mathbf{x}_i)}{\sum_{\mathbf{x}_i} h_{\oplus}(\mathbf{x}_i)} \cdot \frac{\sum_{\mathbf{x}_i} h_{\oplus}(\mathbf{x}_i)}{|\mathscr{D}_3|} \cdot \frac{|\mathscr{D}_3|}{\sum_{\mathbf{x}_i} \pi_s(\mathbf{x}_i)} \\ &= \hat{\mu}^{\text{WE}}(s)\end{aligned}$$

The equivalence between CC and TE is straightforward. □

C.4 SVM-based Quantification

In this appendix we report the results of experiments, analogous to the ones in Sections 5.4.6-5.4.8, where quantifiers are wrapped around an SVM classifier rather than an LR classifier. The experimental protocols are summarized in Tables C.1-C.5. The ablation study is depicted in Figures C.1-C.5. Experiments on decoupling the quantification performance of a model from its classification performance are reported in Figures C.6 and C.7.

Table C.1 Results obtained in the experiments run according to protocol `sample-prev-`$\mathscr{D}_3$ with the SVM-based classifier.

		↓ MAE	↓ MSE	↑ $P(\text{AE} < 0.1)$	↑ $P(\text{AE} < 0.2)$
Adult	CC(SVM)	0.410 ±0.323	0.273 ±0.341	0.193	0.365
	PCC(SVM)	0.308 ±0.244	0.154 ±0.210	0.230	0.412
	ACC(SVM)	0.107 ±0.105	0.022 ±0.053	0.606	0.857
	PACC(SVM)	0.059 ±0.057	0.007 ±0.016	0.824	0.971
	SLD(SVM)	**0.056** ±0.050	**0.006** ±0.011	**0.836**	**0.983**
	HDy(SVM)	0.104 ±0.078	0.017 ±0.028	0.546	0.895
COMPAS	CC(SVM)	0.543 ±0.370	0.432 ±0.474	0.115	0.235
	PCC(SVM)	0.339 ±0.243	0.174 ±0.216	0.179	0.343
	ACC(SVM)	0.497 ±0.346	0.367 ±0.448	0.127	0.224
	PACC(SVM)	0.269 ±0.207	0.115 ±0.165	0.250	0.445
	SLD(SVM)	**0.227** ±0.202	**0.092** ±0.154	**0.335**	**0.566**
	HDy(SVM)	0.265 ±0.204	0.112 ±0.162	0.238	0.459
CreditCard	CC(SVM)	0.346 ±0.241	0.178 ±0.213	0.171	0.335
	PCC(SVM)	0.329 ±0.215	0.155 ±0.161	0.173	0.335
	ACC(SVM)	0.358 ±0.270	0.201 ±0.276	0.175	0.348
	PACC(SVM)	0.267 ±0.215	0.118 ±0.180	0.252	0.473
	SLD(SVM)	$0.243^{\ddagger}$ ±0.191	$0.096^{\dagger}$ ±0.143	0.268	0.496
	HDy(SVM)	**0.237** ±0.186	**0.090** ±0.137	**0.271**	**0.507**

Table C.2 Results obtained in the experiments run according to protocol `sample-prev-`$\mathscr{D}_2$ with the SVM-based classifier.

		↓ MAE	↓ MSE	↑ $P(\text{AE} < 0.1)$	↑ $P(\text{AE} < 0.2)$
Adult	CC(SVM)	0.217 ±0.168	0.075 ±0.106	0.286	0.554
	PCC(SVM)	0.242 ±0.190	0.095 ±0.129	0.303	0.507
	ACC(SVM)	0.150 ±0.169	0.051 ±0.147	0.458	0.799
	PACC(SVM)	0.111 ±0.114	0.025 ±0.085	0.555	0.888
	SLD(SVM)	**0.095** ±0.100	**0.019** ±0.067	**0.634**	**0.929**
	HDy(SVM)	0.182 ±0.151	0.056 ±0.084	0.381	0.634
COMPAS	CC(SVM)	0.506 ±0.255	0.321 ±0.267	0.036	0.116
	PCC(SVM)	0.266 ±0.187	0.106‡ ±0.128	0.226	0.430
	ACC(SVM)	0.479 ±0.276	0.306 ±0.305	0.076	0.175
	PACC(SVM)	0.356 ±0.260	0.194 ±0.261	0.167	0.324
	SLD(SVM)	0.297 ±0.244	0.148 ±0.228	0.231	0.424
	HDy(SVM)	**0.255** ±0.192	**0.102** ±0.141	**0.240**	**0.479**
CreditCard	CC(SVM)	0.428 ±0.253	0.247 ±0.237	0.106	0.229
	PCC(SVM)	**0.209** ±0.142	**0.064** ±0.076	**0.285**	**0.537**
	ACC(SVM)	0.531 ±0.316	0.382 ±0.352	0.090	0.164
	PACC(SVM)	0.542 ±0.313	0.391 ±0.352	0.078	0.145
	SLD(SVM)	0.445 ±0.284	0.279 ±0.288	0.118	0.237
	HDy(SVM)	0.246 ±0.193	0.098 ±0.150	0.256	0.487

Table C.3 Results obtained in the experiments run according to protocol `sample-size-`$\mathscr{D}_2$ with the SVM-based classifier

		↓ MAE	↓ MSE	↑ $P(\text{AE} < 0.1)$	↑ $P(\text{AE} < 0.2)$
Adult	CC(SVM)	0.131 ±0.028	0.018 ±0.007	0.133	**1.000**
	PCC(SVM)	**0.012** ±0.011	**0.000** ±0.000	**1.000**	**1.000**
	ACC(SVM)	0.081 ±0.107	0.018 ±0.076	0.759	0.935
	PACC(SVM)	0.051 ±0.066	0.007 ±0.036	0.873	0.977
	SLD(SVM)	0.043 ±0.062	0.006 ±0.030	0.907	0.971
	HDy(SVM)	0.045 ±0.034	0.003 ±0.005	0.918	0.999
COMPAS	CC(SVM)	0.355 ±0.044	0.128 ±0.031	0.000	0.003
	PCC(SVM)	**0.029** ±0.019	**0.001** ±0.001	**0.999**	**1.000**
	ACC(SVM)	0.389 ±0.212	0.196 ±0.212	0.090	0.171
	PACC(SVM)	0.284 ±0.231	0.134 ±0.210	0.233	0.444
	SLD(SVM)	0.228 ±0.199	0.092 ±0.158	0.305	0.555
	HDy(SVM)	0.130 ±0.102	0.027 ±0.041	0.461	0.778
CreditCard	CC(SVM)	0.189 ±0.079	0.042 ±0.032	0.132	0.552
	PCC(SVM)	**0.016** ±0.013	**0.000** ±0.001	**1.000**	**1.000**
	ACC(SVM)	0.358 ±0.269	0.201 ±0.267	0.181	0.328
	PACC(SVM)	0.322 ±0.257	0.169 ±0.248	0.213	0.385
	SLD(SVM)	0.243 ±0.192	0.096 ±0.142	0.284	0.497
	HDy(SVM)	0.105 ±0.096	0.020 ±0.051	0.583	0.876

Table C.4 Results obtained in the experiments run according to protocol `sample-prev-`$\mathscr{D}_1$ with the SVM-based classifier.

		↓ MAE	↓ MSE	↑ $P(\text{AE} < 0.1)$	↑ $P(\text{AE} < 0.2)$
Adult	CC(SVM)	0.126 ±0.047	0.018 ±0.011	0.268	0.959
	PCC(SVM)	**0.007** ±0.005	**0.000** ±0.000	**1.000**	**1.000**
	ACC(SVM)	0.032 ±0.032	0.002 ±0.014	0.968	0.998
	PACC(SVM)	0.018 ±0.014	0.001 ±0.001	**1.000**	**1.000**
	SLD(SVM)	0.013 ±0.010	0.000 ±0.000	**1.000**	**1.000**
	HDy(SVM)	0.022 ±0.016	0.001 ±0.001	**1.000**	**1.000**
COMPAS	CC(SVM)	0.334 ±0.087	0.119 ±0.055	0.018	0.063
	PCC(SVM)	**0.026** ±0.018	**0.001** ±0.001	**1.000**	**1.000**
	ACC(SVM)	0.349 ±0.196	0.160 ±0.174	0.123	0.221
	PACC(SVM)	0.208 ±0.179	0.075 ±0.134	0.332	0.578
	SLD(SVM)	0.170 ±0.166	0.057 ±0.117	0.422	0.707
	HDy(SVM)	0.113 ±0.089	0.021 ±0.031	0.528	0.839
CreditCard	CC(SVM)	0.152 ±0.100	0.033 ±0.038	0.360	0.708
	PCC(SVM)	**0.010** ±0.007	**0.000** ±0.000	**1.000**	**1.000**
	ACC(SVM)	0.194 ±0.160	0.063 ±0.104	0.342	0.618
	PACC(SVM)	0.132 ±0.108	0.029 ±0.046	0.482	0.778
	SLD(SVM)	0.110 ±0.091	0.020 ±0.032	0.560	0.845
	HDy(SVM)	0.080 ±0.061	0.010 ±0.014	0.683	0.953

Table C.5 Results obtained in the experiments run according to protocol `flip-prev-`$\mathscr{D}_1$ with the SVM-based classifier.

		↓ MAE	↓ MSE	↑ $P(\text{AE} < 0.1)$	↑ $P(\text{AE} < 0.2)$
Adult	CC(SVM)	0.175 ±0.085	0.038 ±0.028	0.231	0.544
	PCC(SVM)	**0.007** ±0.006	**0.000** ±0.000	**1.000**	**1.000**
	ACC(SVM)	0.032 ±0.028	0.002 ±0.004	0.969	0.999
	PACC(SVM)	0.020 ±0.015	0.001 ±0.001	**1.000**	**1.000**
	SLD(SVM)	0.015 ±0.012	0.000 ±0.001	**1.000**	**1.000**
	HDy(SVM)	0.022 ±0.018	0.001 ±0.001	1.000	**1.000**
COMPAS	CC(SVM)	0.395 ±0.113	0.169 ±0.083	0.021	0.055
	PCC(SVM)	**0.027** ±0.019	**0.001** ±0.001	**0.998**	**1.000**
	ACC(SVM)	0.399 ±0.204	0.201 ±0.193	0.094	0.174
	PACC(SVM)	0.207 ±0.176	0.074 ±0.131	0.325	0.587
	SLD(SVM)	0.160 ±0.146	0.047 ±0.095	0.418	0.722
	HDy(SVM)	0.112 ±0.084	0.020 ±0.028	0.528	0.842
CreditCard	CC(SVM)	0.165 ±0.105	0.038 ±0.039	0.328	0.627
	PCC(SVM)	**0.012** ±0.009	**0.000** ±0.000	**1.000**	**1.000**
	ACC(SVM)	0.227 ±0.186	0.086 ±0.130	0.303	0.542
	PACC(SVM)	0.144 ±0.120	0.035 ±0.059	0.442	0.742
	SLD(SVM)	0.118 ±0.095	0.023 ±0.036	0.512	0.819
	HDy(SVM)	0.092 ±0.070	0.013 ±0.019	0.621	0.913

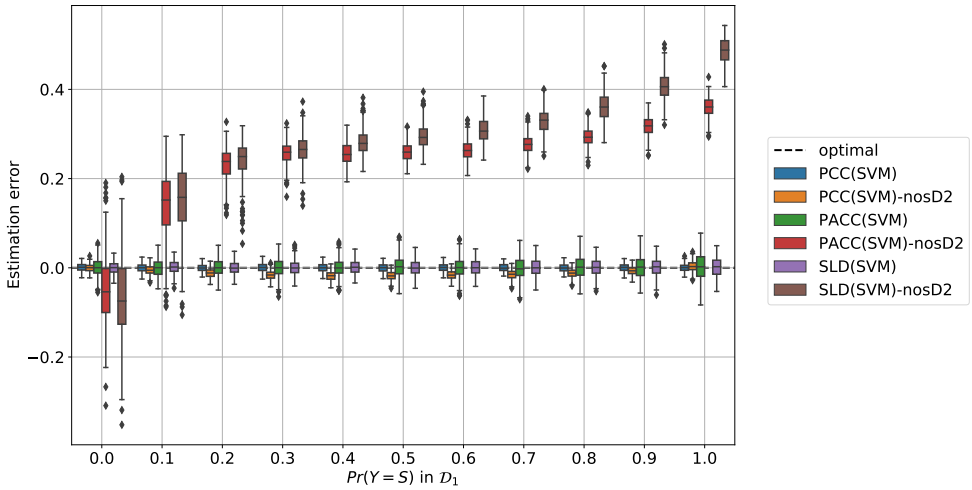


Fig. C.1 Results obtained in the ablation study on the Adult dataset with SVM-based quantification for protocol `sample-prev-`$\mathscr{D}_1$.

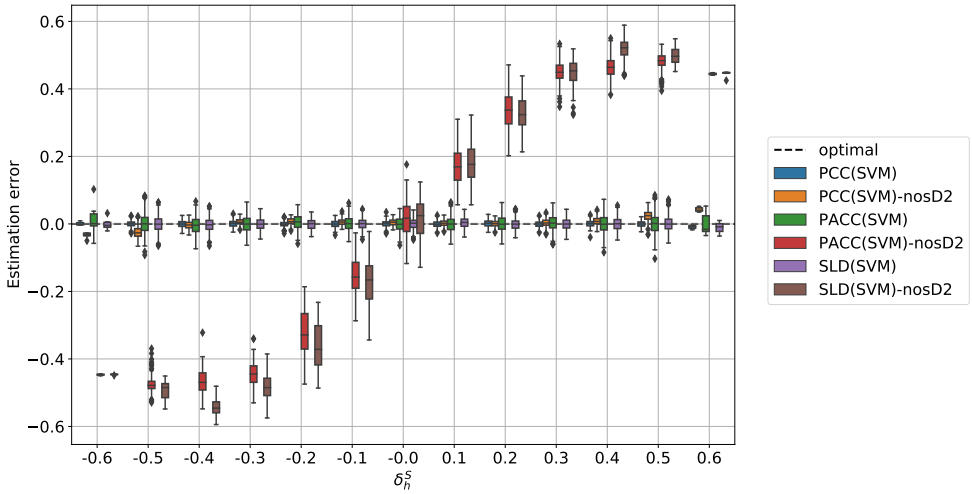


Fig. C.2 Results obtained in the ablation study on the Adult dataset with SVM-based quantification for protocol `flip-prev-`$\mathscr{D}_1$.

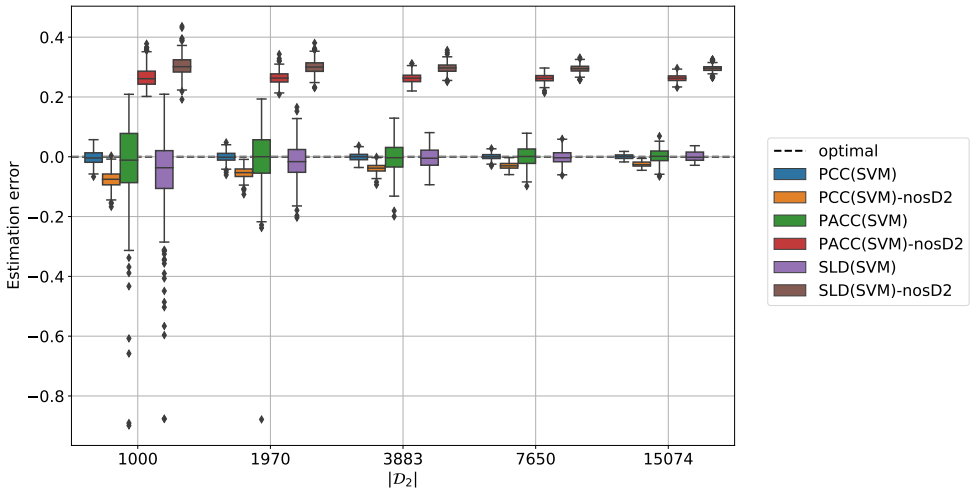
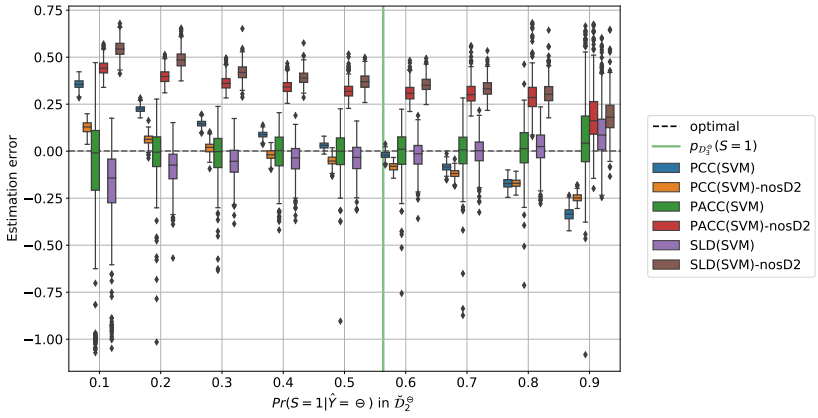
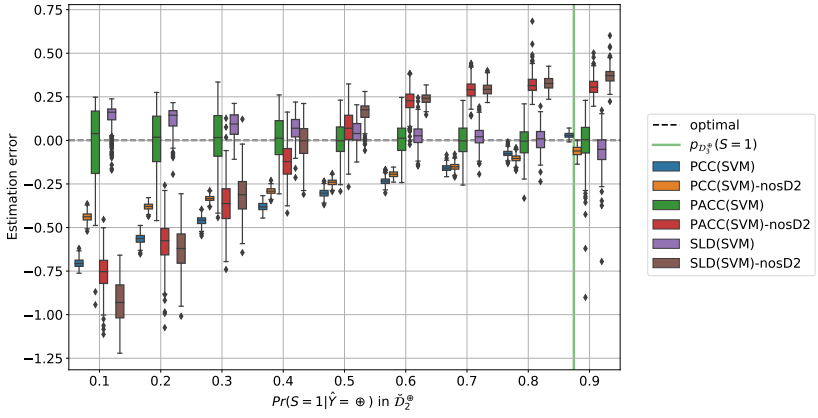


Fig. C.3 Results obtained in the ablation study on the Adult dataset with SVM-based quantification for protocol `sample-size-`$\mathscr{D}_2$.

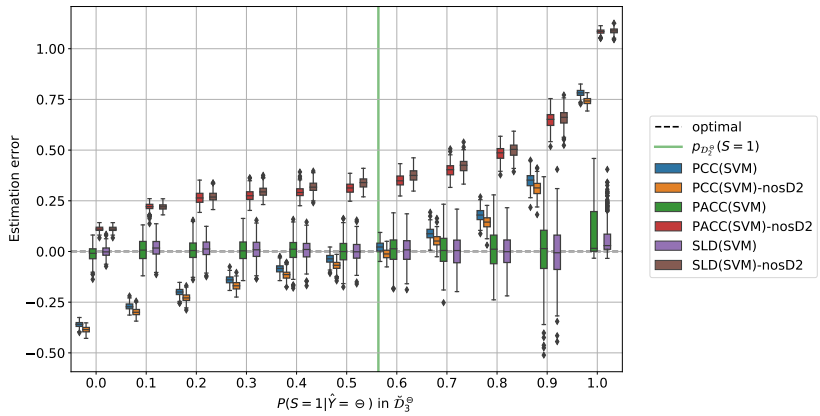


(a) Protocol `sample-prev-`$\mathscr{D}_2^{\ominus}$

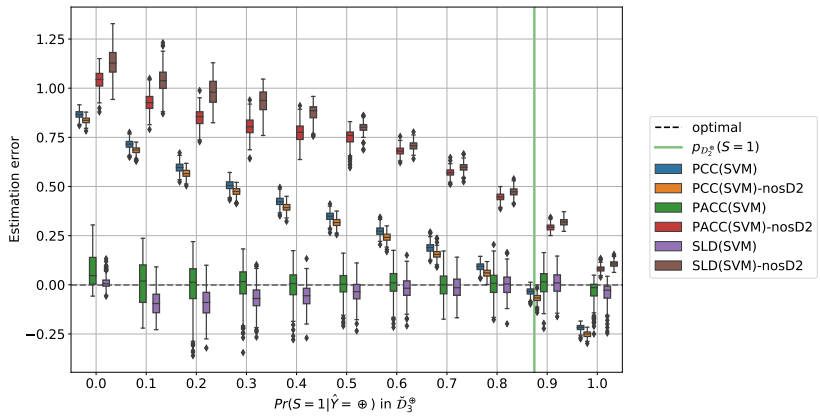


(b) Protocol `sample-prev-`$\mathscr{D}_2^{\oplus}$

Fig. C.4 Results obtained in the ablation study on the Adult dataset with SVM-based quantification for protocol `sample-prev-`$\mathscr{D}_2$.

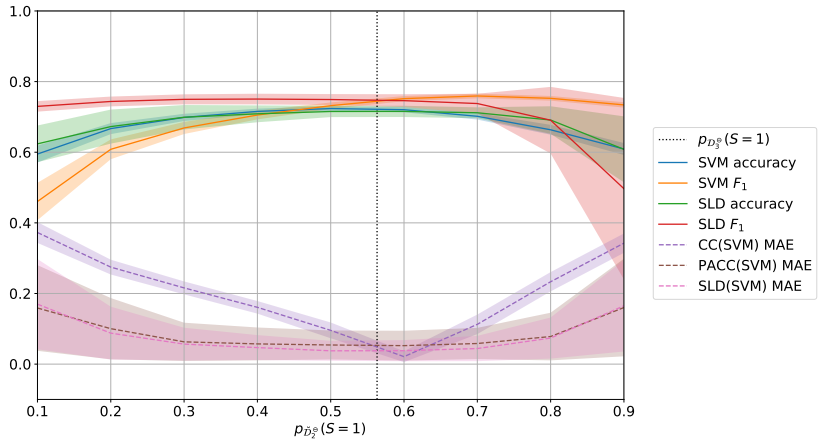


(a) Protocol $\texttt{sample-prev-}\mathscr{D}_3^{\ominus}$

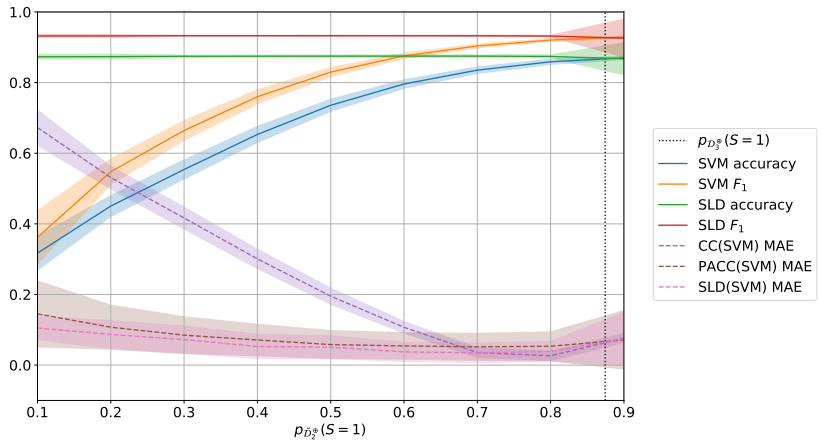


(b) Protocol $\texttt{sample-prev-}\mathscr{D}_3^{\oplus}$

Fig. C.5 Results obtained in the ablation study on the Adult dataset with SVM-based quantification for protocol $\texttt{sample-prev-}\mathscr{D}_3$.

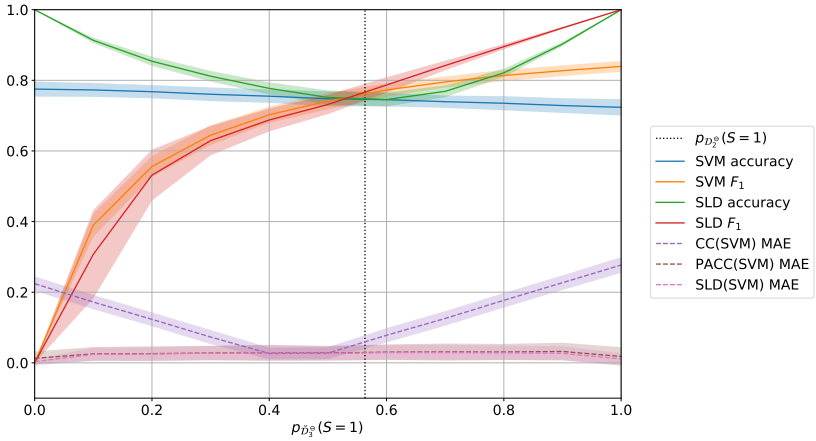


(a) Protocol `sample-prev-`$\mathscr{D}_2^{\ominus}$

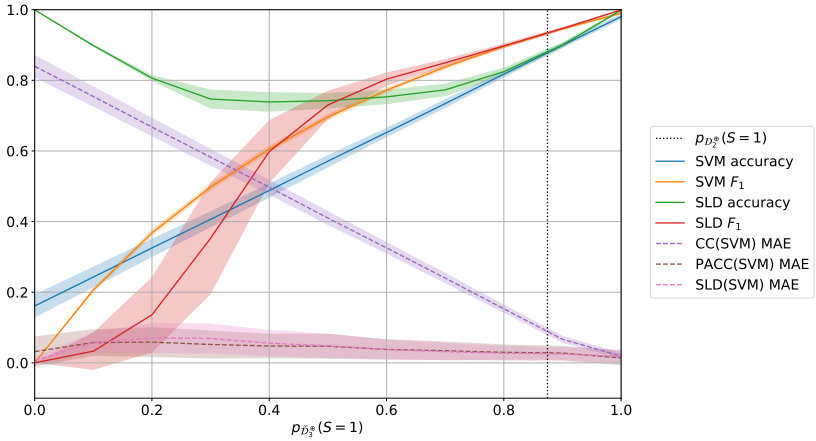


(b) Protocol `sample-prev-`$\mathscr{D}_2^{\oplus}$

Fig. C.6 Performance of SVM-based methods CC, SLD and PACC on the Adult dataset when used for quantification (MAE – lower is better) and classification (F_1, accuracy – higher is better) under protocol `sample-prev-`$\mathscr{D}_2$. The classification performance of PACC is equivalent to that of CC (both equal to the performance of the underlying SVM), and we thus omit it for readability.



(a) Protocol `sample-prev-`$\mathscr{D}_3^{\ominus}$



(b) Protocol `sample-prev-`$\mathscr{D}_3^{\oplus}$

Fig. C.7 Performance of SVM-based methods CC, SLD and PACC on the Adult dataset when used for quantification (MAE – lower is better) and classification (F_1, accuracy – higher is better) under protocol `sample-prev-`$\mathscr{D}_3$. The classification performance of PACC is equivalent to that of CC (both equal to the performance of the underlying SVM), and we thus omit it for readability.

C.5 Pseudocode

This section reports pseudocode for protocols `sample-prev-`$\mathscr{D}_2$ (Pseudocode 3), `sample-size-`$\mathscr{D}_2$ (Pseudocode 4), and `sample-prev-`$\mathscr{D}_1$ (Pseudocode 5).

Input : • Dataset $\mathscr{D}$;
• Classifier learner CLS;
• Quantification method Q;
Output : • MAE of the demographic disparity estimates ;
• MSE of the demographic disparity estimates ;

$E \leftarrow \emptyset$;
for *5 random splits* **do**
  $\mathscr{D}_A, \mathscr{D}_B, \mathscr{D}_C \leftarrow$ split_stratify($\mathscr{D}$) ;
  for $\mathscr{D}_1, \mathscr{D}_2, \mathscr{D}_3 \in$ permutations($\mathscr{D}_A, \mathscr{D}_B, \mathscr{D}_C$) **do**
    `/* Learn a classifier` $h : \mathscr{X} \to \mathscr{Y}$ `*/`
    $h \leftarrow$ CLS.fit($\mathscr{D}_1$) ;
    $\mathscr{D}_2^{\ominus} \leftarrow \{(\mathbf{x}_i, s_i) \in \mathscr{D}_2 \mid h(\mathbf{x}_i) = \ominus\}$;
    $\mathscr{D}_2^{\oplus} \leftarrow \{(\mathbf{x}_i, s_i) \in \mathscr{D}_2 \mid h(\mathbf{x}_i) = \oplus\}$;
    for *10 repeats* **do**
      for $p \in \{0.1, 0.2, \ldots, 0.9\}$ **do**
        `/* Generate samples from` $\mathscr{D}_2^{\ominus}$ `at desired prevalence and size, and uniform samples from` $\mathscr{D}_2^{\oplus}$ `at desired size */`
        $\check{\mathscr{D}}_2^{\ominus} \sim \mathscr{D}_2^{\ominus}$ with $p_{\check{\mathscr{D}}_2^{\ominus}}(s) = p$ and $|\check{\mathscr{D}}_2^{\ominus}| = 500$;
        $\check{\mathscr{D}}_2^{\oplus} \sim \mathscr{D}_2^{\oplus}$ with $|\check{\mathscr{D}}_2^{\oplus}| = 500$;
        `/* Learn quantifiers` $q_y : 2^{\mathscr{X}} \to [0,1]$ `*/`
        $q_{\ominus} \leftarrow$ Q.fit($\check{\mathscr{D}}_2^{\ominus}$) ;
        $q_{\oplus} \leftarrow$ Q.fit($\check{\mathscr{D}}_2^{\oplus}$) ;
        `/* Use quantifiers to estimate demographic prevalence */`
        $\mathscr{D}_3^{\ominus} \leftarrow \{\mathbf{x}_i \in \mathscr{D}_3 \mid h(\mathbf{x}_i) = \ominus\}$;
        $\mathscr{D}_3^{\oplus} \leftarrow \{\mathbf{x}_i \in \mathscr{D}_3 \mid h(\mathbf{x}_i) = \oplus\}$;
        $\hat{p}^{q_{\ominus}}_{\mathscr{D}_3^{\ominus}}(s) \leftarrow q_{\ominus}(\mathscr{D}_3^{\ominus})$;
        $\hat{p}^{q_{\oplus}}_{\mathscr{D}_3^{\oplus}}(s) \leftarrow q_{\oplus}(\mathscr{D}_3^{\oplus})$;
        `/* Compute the signed error of the demographic disparity estimate */`
        $e \leftarrow$ compute error using $\hat{p}^{q_{\ominus}}_{\mathscr{D}_3^{\ominus}}(s)$, $\hat{p}^{q_{\oplus}}_{\mathscr{D}_3^{\oplus}}(s)$ and Equation 5.16
        $E \leftarrow E \cup \{e\}$
      end
    end
  end
end
mae $\leftarrow$ MAE(E) ;
mse $\leftarrow$ MSE(E) ;
return mae, mse

Pseudocode 3: Protocol `sample-prev-`$\mathscr{D}_2$, shown for variations of prevalence values in class $y = \ominus$.

Input : • Dataset $\mathscr{D}$;
• Classifier learner CLS;
• Quantification method Q;
Output : • MAE of the demographic disparity estimates ;
• MSE of the demographic disparity estimates ;

$E \leftarrow \emptyset$;
for *5 random splits* **do**
  $\mathscr{D}_A, \mathscr{D}_B, \mathscr{D}_C \leftarrow$ split_stratify($\mathscr{D}$) ;
  for $\mathscr{D}_1, \mathscr{D}_2, \mathscr{D}_3 \in$ permutations($\mathscr{D}_A, \mathscr{D}_B, \mathscr{D}_C$) **do**
    `/* Learn a classifier` $h : \mathscr{X} \to \mathscr{Y}$ `*/`
    $h \leftarrow$ CLS.fit($\mathscr{D}_1$) ;
    for *10 repeats* **do**
      for *size* $s \in$ *logspace(**from:** 1000, **to:** $|\mathscr{D}_2|$, **steps:** 5)* **do**
        `/* Generate samples from` $\mathscr{D}_2$ `at desired size */`
        $\breve{\mathscr{D}}_2 \sim \mathscr{D}_2$ with $|\breve{\mathscr{D}}_2| = s$;
        `/* Learn quantifiers` $q_y : 2^{\mathscr{X}} \to [0,1]$ `*/`
        $\breve{\mathscr{D}}_2^{\ominus} \leftarrow \{(\mathbf{x}_i, s_i) \in \breve{\mathscr{D}}_2 \mid h(\mathbf{x}_i) = \ominus\}$;
        $\breve{\mathscr{D}}_2^{\oplus} \leftarrow \{(\mathbf{x}_i, s_i) \in \breve{\mathscr{D}}_2 \mid h(\mathbf{x}_i) = \oplus\}$;
        $q_{\ominus} \leftarrow$ Q.fit($\breve{\mathscr{D}}_2^{\ominus}$) ;
        $q_{\oplus} \leftarrow$ Q.fit($\breve{\mathscr{D}}_2^{\oplus}$) ;
        `/* Use quantifiers to estimate demographic prevalence */`
        $\mathscr{D}_3^{\ominus} \leftarrow \{\mathbf{x}_i \in \mathscr{D}_3 \mid h(\mathbf{x}_i) = \ominus\}$;
        $\mathscr{D}_3^{\oplus} \leftarrow \{\mathbf{x}_i \in \mathscr{D}_3 \mid h(\mathbf{x}_i) = \oplus\}$;
        $\hat{p}^{q_{\ominus}}_{\mathscr{D}_3^{\ominus}}(s) \leftarrow q_{\ominus}(\mathscr{D}_3^{\ominus})$;
        $\hat{p}^{q_{\oplus}}_{\mathscr{D}_3^{\oplus}}(s) \leftarrow q_{\oplus}(\mathscr{D}_3^{\oplus})$;
        `/* Compute the signed error of the demographic disparity estimate */`
        $e \leftarrow$ compute error using $\hat{p}^{q_{\ominus}}_{\mathscr{D}_3^{\ominus}}(s)$, $\hat{p}^{q_{\oplus}}_{\mathscr{D}_3^{\oplus}}(s)$ and Equation 5.16
        $E \leftarrow E \cup \{e\}$
      end
    end
  end
end
mae $\leftarrow$ MAE(E) ;
mse $\leftarrow$ MSE(E) ;
return mae, mse

Pseudocode 4: Protocol `sample-size-`$\mathscr{D}_2$.

Input : • Dataset $\mathscr{D}$;
• Classifier learner CLS;
• Quantification method Q;
Output : • MAE of the demographic disparity estimates ;
• MSE of the demographic disparity estimates ;

$E \leftarrow \emptyset$;
for *5 random splits* **do**
  $\mathscr{D}_A, \mathscr{D}_B, \mathscr{D}_C \leftarrow$ split_stratify($\mathscr{D}$) ;
  for $\mathscr{D}_1, \mathscr{D}_2, \mathscr{D}_3 \in$ permutations($\mathscr{D}_A, \mathscr{D}_B, \mathscr{D}_C$) **do**
    for *10 repeats* **do**
      for $p \in \{0.0, 0.1, \ldots, 0.9, 1.0\}$ **do**
        `/* Generate samples from` $\mathscr{D}_1$ `at desired prevalence */`
        $\breve{\mathscr{D}}_1 \sim \mathscr{D}_1$ with $P(Y = S) = p$ and $|\breve{\mathscr{D}}_1| = 500$;
        `/* Learn a classifier` $h : \mathscr{X} \to \mathscr{Y}$ `*/`
        $h \leftarrow$ CLS.fit($\breve{\mathscr{D}}_1$) ;
        `/* Learn quantifiers` $q_y : 2^{\mathscr{X}} \to [0,1]$ `*/`
        $\mathscr{D}_2^{\ominus} \leftarrow \{(\mathbf{x}_i, s_i) \in \mathscr{D}_2 \mid h(\mathbf{x}_i) = \ominus\}$;
        $\mathscr{D}_2^{\oplus} \leftarrow \{(\mathbf{x}_i, s_i) \in \mathscr{D}_2 \mid h(\mathbf{x}_i) = \oplus\}$;
        $q_{\ominus} \leftarrow$ Q.fit($\mathscr{D}_2^{\ominus}$) ;
        $q_{\oplus} \leftarrow$ Q.fit($\mathscr{D}_2^{\oplus}$) ;
        `/* Use quantifiers to estimate demographic prevalence */`
        $\mathscr{D}_3^{\ominus} \leftarrow \{\mathbf{x}_i \in \mathscr{D}_3 \mid h(\mathbf{x}_i) = \ominus\}$;
        $\mathscr{D}_3^{\oplus} \leftarrow \{\mathbf{x}_i \in \mathscr{D}_3 \mid h(\mathbf{x}_i) = \oplus\}$;
        $\hat{p}^{q_{\ominus}}_{\mathscr{D}_3^{\ominus}}(s) \leftarrow q_{\ominus}(\mathscr{D}_3^{\ominus})$;
        $\hat{p}^{q_{\oplus}}_{\mathscr{D}_3^{\oplus}}(s) \leftarrow q_{\oplus}(\mathscr{D}_3^{\oplus})$;
        `/* Compute the signed error of the demographic disparity estimate */`
        $e \leftarrow$ compute error using $\hat{p}^{q_{\ominus}}_{\mathscr{D}_3^{\ominus}}(s)$, $\hat{p}^{q_{\oplus}}_{\mathscr{D}_3^{\oplus}}(s)$ and Equation 5.16
        $E \leftarrow E \cup \{e\}$
      end
    end
  end
end
mae $\leftarrow$ MAE(E) ;
mse $\leftarrow$ MSE(E) ;
return mae, mse

Pseudocode 5: Protocol `sample-prev-`$\mathscr{D}_1$.

Appendix D

Supplementary Materials to Chapter 6

In this appendix, we present the controlled vocabulary we adopted to study gender stereotypes in Section 6.2.

D.1 Traits and Terms for Stereotypical Associations

agency	**communion**
aggressive	affectionate
ambitious	compassionate
arrogant	emotional
confident	generous
corageous	honest
critical	nurturing
decisive	outgoing
demanding	patient
hardworking	polite
independent	romantic
possessive	sensitive
proud	unselfish
selfish	
strong	
stubborn	

Table D.1 *agency* vs *communion*: adjectives associated to each construct [224].

science	**arts**
astronomy	art
chemistry	dance
Einstein	drama
experiment	literature
NASA	novel
physics	poetry
science	Shakespeare
technology	symphony

Table D.2 *science* vs *arts*: associated attributes [602].

career	**family**
business	children
career	cousin
corporation	family
executive	home
management	marriage
office	parents
professional	relatives
salary	wedding

Table D.3 *career* vs *family*: associated attributes [603].

Predominantly male			Predominantly female		
occupation	**%F**	**%M**	**occupation**	**%F**	**%M**
stonemason	0.7	99.3	hygienist	96.0	4.0
roofer	1.9	98.1	secretary	93.2	6.8
electrician	2.2	97.8	hairdresser	92.3	7.7
plumber	2.7	97.3	dietician	92.1	7.9
carpenter	2.8	97.2	paralegal	89.6	10.4
firefighter	3.3	96.7	receptionist	89.3	10.7
millwright	5.0	95.0	phlebotomist	89.3	10.7
welder	5.3	94.7	maid	89.0	11.0
machinist	5.6	94.4	nurse	88.9	11.1
driver	6.7	93.3	typist	86.0	14.0

Table D.4 *jobs_m* vs *jobs_f*: occupations with highest gender gap in representation [129].

D.2 Gendered Entities

The following are used in section 6.2.4 to detect mentions of intrinsically gendered entities.

Words associated with male entities:

actor, actors, bachelor, bachelors, bloke, blokes, boy, boys, boyfriend, boyfriends, brother, brothers, brethren, businessman, businessmen, chairman, chairmen, chap, chaps, congressman, congressmen, councilman, councilmen, dad, daddy, dads, dude, dudes, ex-boyfriend, ex-boyfriends, exboyfriend, exboyfriends, father, fathers, fella, fellas, gentleman, gentlemen, godfather, godfathers, grandfather, grandfathers, grandpa, grandson, grandsons, guy, guys, handyman, handymen, he, him, himself, his, husband, husbands, king, kings, lad, lads, male, males, man, men, monk, monks, mr, nephew, nephews, pa, prince, princes, salesman, salesmen, schoolboy, schoolboys, son, sons, spokesman, spokesmen, statesman, statesmen, stepfather, stepfathers, stepson, stepsons, uncle, uncles, waiter, waiters.

Words associated with female entities:

actress, actresses, aunt, aunts, ballerina, ballerinas, bride, brides, businesswoman, businesswomen, chairwoman, chairwomen, congresswoman, congresswomen, councilwoman, councilwomen, daughter, daughters, exgirlfriend, exgirlfriends, ex-girlfriend, ex-girlfriends, female, females, gal, gals, girl, girls, girlfriend, girlfriends, godmother, godmothers, granddaughter, granddaughters, grandma, grandmas, grandmother, grandmothers, her, hers, herself, hostess, hostesses, housewife, housewives, lady, ladies, ma, maid, maiden, maids, mama, mom, mommy, moms, mother, mothers, ms, mrs, niece, nieces, nun, nuns, princess, princesses, queen, queens, schoolgirl, schoolgirls, she, sister, sisters, spokeswoman, spokeswomen, stepdaughter, stepmother, waitress, waitresses, wife, wives, woman, women.