Tackling Documentation Debt: A Survey on Algorithmic Fairness Datasets

Alessandro Fabris
University of Padua
Padua, Italy
fabrisal@dei.unipd.it

Stefano Messina
University of Padua
Padua, Italy
stefano.messina@studenti.unipd.it

Gianmaria Silvello
University of Padua
Padua, Italy
silvello@dei.unipd.it

Gian Antonio Susto
University of Padua
Padua, Italy
susto@dei.unipd.it

ABSTRACT

A growing community of researchers has been investigating the equity of algorithms, advancing the understanding of risks and opportunities of automated decision-making for historically disadvantaged populations. Progress in fair Machine Learning (ML) hinges on data, which can be appropriately used only if adequately documented. Unfortunately, the research community, as a whole, suffers from a collective data documentation debt caused by a lack of information on specific resources (*opacity*) and scatteredness of available information (*sparsity*). In this work, we survey over two hundred datasets employed in algorithmic fairness research, producing standardized and searchable documentation for each of them. Moreover we rigorously identify the three most popular fairness datasets, namely Adult, COMPAS, and German Credit, for which we compile in-depth documentation. This unifying documentation effort targets *documentation sparsity* and supports multiple contributions. In the first part of this work, we summarize the merits and limitations of Adult, COMPAS, and German Credit, adding to and unifying recent scholarship, calling into question their suitability as general-purpose fairness benchmarks. To overcome this limitation, we document hundreds of available alternatives, annotating their domain and the algorithmic fairness tasks they support, along with additional properties of interest for fairness practitioners and researchers, including their format, cardinality, and the sensitive attributes they encode. In the second part, we summarize this information, zooming in on the domains and tasks supported by these resources. Overall, we assemble and summarize sparse information on hundreds of datasets into a single resource, which we make available to the community, with the aim of tackling the data documentation debt.

This is the conference version of Fabris et al. [27], which presents a more comprehensive and detailed analysis of fairness datasets.

EAAMO '22, October 6–9, 2022, Arlington, VA, USA

ACM ISBN 978-1-4503-9477-2/22/10…$15.00
https://doi.org/10.1145/3551624.3555286

KEYWORDS

Algorithmic fairness, Data studies, Documentation debt.

ACM Reference Format:
Alessandro Fabris, Stefano Messina, Gianmaria Silvello, and Gian Antonio Susto. 2022. Tackling Documentation Debt: A Survey on Algorithmic Fairness Datasets. In *Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO '22), October 6–9, 2022, Arlington, VA, USA.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3551624.3555286

1 INTRODUCTION

Data documentation is important and caters to different goals. It increases transparency, favouring an improved understanding of the data and resulting models [39], it reduces chances of data misuse [31] and supports accountability in dataset and model creation [39], it helps connect the data with its context to guide scientific inquiry [65], and it makes the values influencing dataset curation explicit [71]. In the field of software development, technical debt is a cost incurred when speed of execution is prioritized over quality [39]. In recent work, Bender et al. [9] propose the notion of *documentation debt*, in relation to training sets that are undocumented and too large to document retrospectively. This debt compounds 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 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*. One example that is particularly relevant for the algorithmic fairness community is represented by the German Credit dataset [75], 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 [5, 35, 57, 59, 67, 73, 78, 79], while existing yet overlooked documentation shows that this feature cannot be reliably retrieved [34].[1]

To tackle the documentation debt of the algorithmic fairness community, we survey the datasets used in over 500 articles on fair ML and equitable algorithms, presented at seven major conferences, considering each edition in the period 2014–2021, and more than

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

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 endeavour, 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 article. Moreover, we identify and carefully analyze the three datasets most often utilized in the surveyed articles (Adult, COMPAS, and German Credit), retrospectively producing a *datasheet* [31] and a *nutrition label* [36] for each of them. From these documentation efforts, we extract a summary of the merits and limitations of popular algorithmic fairness benchmarks, and a categorization of alternative resources with respect to domains and tasks in works of algorithmic fairness. Overall, we make the following contributions.

- **Unified analysis of popular fairness benchmarks**. We produce datasheets and nutrition labels for Adult, COMPAS, and German Credit, from which we extract a summary of their merits and limitations. We add to and unify recent scholarship on these datasets, calling into question their suitability as general-purpose fairness benchmarks due to contrived prediction tasks, noisy data, severe coding mistakes, limitations in encoding sensitive attributes, and age. Table 1 summarizes this contribution.
- **Survey of existing alternatives**. We compile standardized and compact documentation for over two hundred resources used in fair ML research, annotating their domain and the tasks they support in works of algorithmic fairness. By assembling sparse information on hundreds of datasets into a single document, we aim to support multiple goals by researchers and practitioners, including domain-oriented and task-oriented search by dataset users. Contextually, we provide a novel taxonomy of tasks and domains investigated in algorithmic fairness research (summarized in Tables 2 and 3).

Roadmap. Readers looking for alternative fairness datasets should prioritize Section 5, Appendix A, and take account of the web app under development (see Footnote 5). Overall, this work is organized as follows. Section 2 introduces related works. Section 3 presents the methodology and inclusion criteria of this survey. Section 4 analyzes the perks and limitations of the most popular datasets. Section 5 discusses alternative fairness resources from the perspective of the underlying domains and supported tasks. Finally, Section 6 contains concluding remarks and details the broader importance of this work for the research community. Interested readers may find the data briefs in Appendix A, followed by the detailed documentation produced for Adult, COMPAS, and German Credit.

2 RELATED WORK

2.1 Data studies

In recent years, several works analyzing multiple datasets along specific lines have been published. Crawford and Paglen [19] focus on resources commonly used as training sets in computer vision, with attention to associated labels and underlying taxonomies. Fabbrizzi et al. [26] also consider computer vision datasets, describing types of bias affecting them, along with methods for discovering and measuring bias, while Scheuerman et al. [71] analyze the values encoded in their documentation. Koch et al. [49] 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. [66] analyze ethical concerns in three popular face and person recognition datasets, stemming from derivative datasets and models, lack of clarity of licenses, and dataset management practices. Geiger et al. [32] evaluate transparency in the documentation of labeling practices employed in over 100 datasets about Twitter.

The work most closely related (and concurrently carried out) to ours is Le Quy et al. [52]. 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 (over 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.

2.2 Documentation frameworks

Several data documentation frameworks have been proposed in the literature; three popular ones are described below. *Datasheets for Datasets* [31] are a general-purpose qualitative framework with over fifty questions covering key aspects of datasets, such as motivation, composition, collection, preprocessing, uses, distribution, and maintenance. Another qualitative framework is represented by *Data statements* [8], which is tailored for NLP, requiring domain-specific information on language variety and speaker demographics. *Dataset Nutrition Labels* [36] 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 [2, 64].

Popular datasets require close scrutiny; for this reason we adopt these frameworks, producing three datasheets and nutrition labels for Adult, German Credit, and COMPAS. This approach, however, does 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, described in Appendix A, include fields specific to fair ML, such sensitive attributes and tasks for which the dataset has been used in the algorithmic fairness literature.

3 METHODOLOGY

In this work, we consider (1) every article published in the proceedings of domain-specific conferences such 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.[2] 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.[3]

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 around 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 instance, include articles on handling pose distribution bias [87], compensating selection bias to improve accuracy without attention to sensitive attributes [45], enhancing desirable discriminating properties of models [14], or generally focused on model performance [55, 88]. 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 provided references, and (3) reviewing scholarly articles or official websites found by querying popular search engines with the dataset name. From this effort, we rigorously identify the three most popular resources, whose perks and limitations are summarized in the next section.

4 MOST POPULAR DATASETS

Figure 1 depicts the number of articles using each dataset, showing that dataset utilization in surveyed scholarly works follows a long tail distribution, reflecting findings of data use in computer vision [49]. 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. [82], used in 7 articles. In this section, we summarize positive and negative aspects of the three most popular datasets, namely Adult, COMPAS, and German Credit, informed by extensive documentation in Appendices B, C, and D.

4.1 Adult

The Adult dataset was created as a resource to benchmark the performance of machine learning algorithms on socially relevant data. Adult inherits some positive sides from the best practices employed by the US Census Bureau, including sample representativeness and fair compensation of labor. 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 [24]. Furthermore, there are several sources of noise affecting the data. Roughly 7% of the data points have missing values, plausibly due to issues with data recording and coding, or respondents' inability to recall information. Moreover, the tendency in household surveys for respondents to under-report their income is a common concern of the Census Bureau [61]. Another source of noise is top-coding of the variable "capital-gain" (saturation to $99,999) to avoid the re-identification of certain individuals [77]. Finally, the dataset is rather old; sensitive attribute "race" contains the outdated "Asian Pacific Islander" class. It is worth noting that a set of similar resources was recently made available, allowing more current socio-economic studies of the US population [24].

4.2 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. 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 clerical errors and mismatches are present [51]. Moreover, in its official release [68], the COMPAS dataset features redundant variables and data leakage due to spuriously time-dependent recidivism rates [7]. For these reasons, researchers must perform further preprocessing in addition to the standard one by 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 [60]. The lack of a standard preprocessing protocol beyond the one by ProPublica [68], which is insufficient to handle these factors, may cause issues of reproducibility and difficulty in comparing methods. Moreover, according to Northpointe's response to the ProPublica's study,

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

[3] 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.

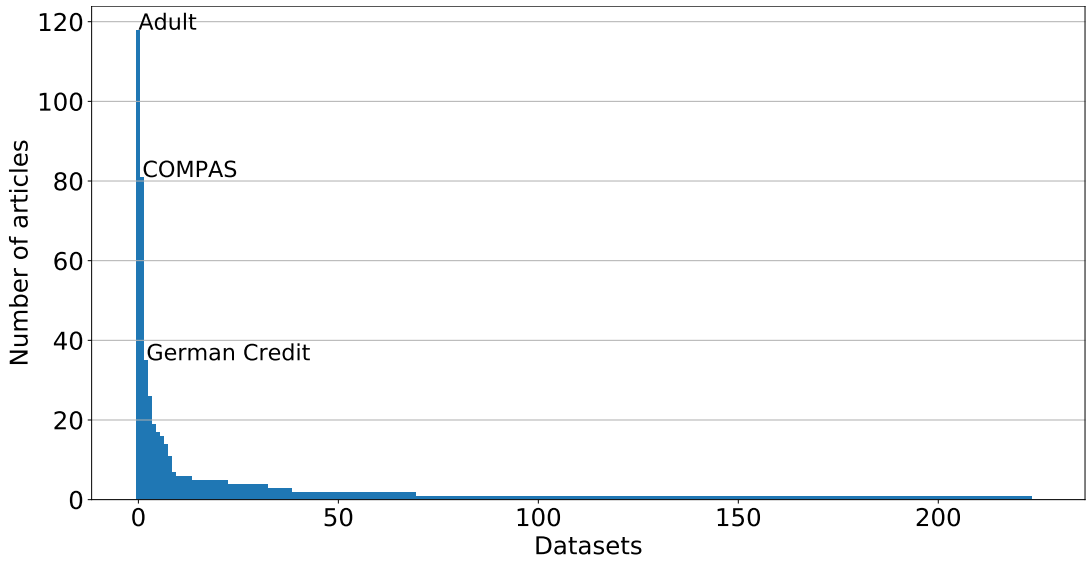

Figure 1: Utilization of datasets in fairness research follows a long tail distribution.

several risk factors considered by the COMPAS algorithm are absent from the dataset [23]. As an additional concern, race categories lack Native Hawaiian or Other Pacific Islander, while Hispanic is redefined as race instead of ethnicity [6]. Finally, defendants' personal information (e.g. race and criminal history) is available in conjunction with obvious identifiers, making re-identification of defendants trivial.

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. [6] suggest avoiding the use of COMPAS to demonstrate novel approaches in algorithmic fairness, as considering the data without proper context may lead to misleading conclusions, which could misguidedly enter the broader debate on criminal justice and risk assessment.

4.3 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, bringing about a natural selection bias. Within this sample, bad credits are oversampled to favour a balance in target classes [34]. The data summarizes applicants' 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 the 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 [75]. A revised version with correct variable encodings, called South German Credit, was donated to UCI Machine Learning Repository [76] with an accompanying report [34].

The greatest upside of this dataset is the fact that it captures a real-world application of credit scoring at 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 [75] comes with highly inaccurate documentation which contains wrong 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 afforded by raw features, with the exception of potentially counterintuitive explanations in works of interpretability and exploratory analysis [52]. This coding error, along with others discussed in Grömping [34] was corrected in a novel release of the dataset [76]. 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 [52, 78]. These coding mistakes are part of a documentation debt whose influence continues to affect the algorithmic fairness community.

4.4 Summary

On close scrutiny, the fundamental merit of these datasets lies in originating from human processes, encoding protected attributes,

Table 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

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 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 extracting a sex variable from the German Credit dataset should be considered noisy at best. As for alternatives, Bao et al. [6] suggest employing well-designed simulations. A complementary avenue is to seek different datasets that are relevant for the problem at hand. We hope that the two hundred data briefs accompanying this work will prove useful in this regard, favouring both domain-oriented and task-oriented searches, according to the classification discussed in the next section.

5 EXISTING ALTERNATIVES

In this section, we discuss existing fairness resources from different perspectives. In section 5.1 we describe the different domains spanned by fairness datasets. In section 5.2 we provide a categorization of fairness tasks supported by the same resources.

5.1 Domain

In Figure 2, we report a subdivision of the surveyed datasets in different macrodomains. We mostly follow the area-category taxonomy by Scimago,[4] 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 selection of macrodomains and subdomains, summarized in detail in Table 3 (Appendix A)

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 recommendation, 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 survey. These resources are often focused 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 dataset about 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 a granularity spanning 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,

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

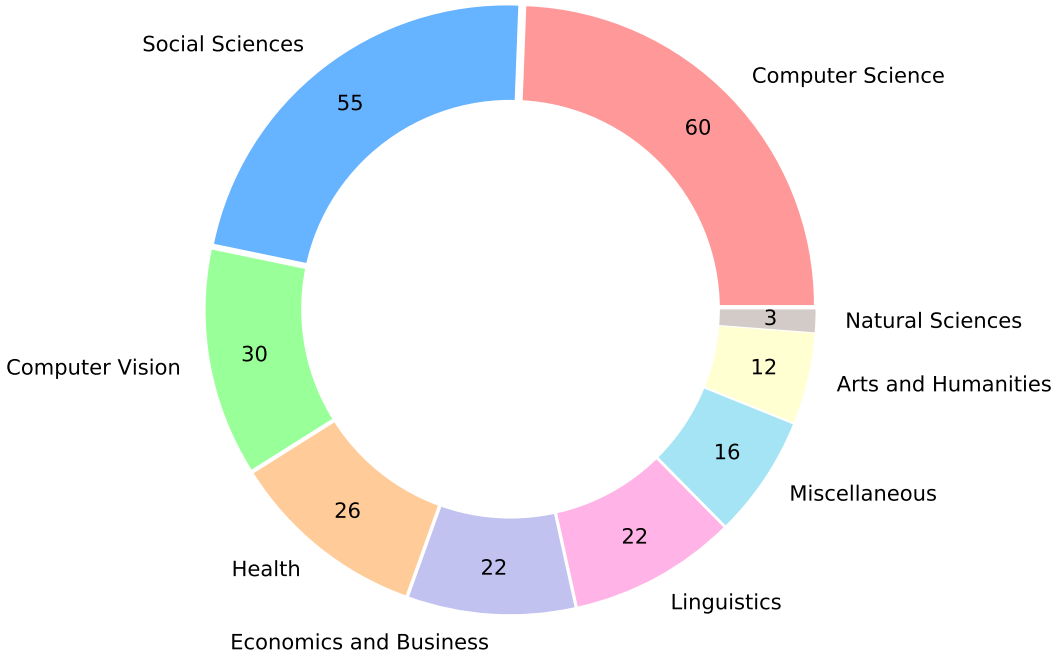

Figure 2: Datasets employed in fairness research span diverse domains. See Table 3 (Appendix A) for a detailed breakdown.

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, also featuring 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 surveyed 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 over 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. Specialties represented in the surveyed datasets are mostly 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 with 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 of multiple medical centers to study problems of 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 the ones 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 which 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 about automated hiring (CVs from Singapore, Pymetrics Bias Group) and employee retention (IBM HR Analytics).

5.2 Task and setting

In this section, we provide an overview of common tasks and settings studied on these datasets, showing their variety and diversity. We use the word *task* to indicate ML problems, such as classification or regression, and *setting* to denote a challenge that runs across different tasks, such as the presence of noise corrupting labels for sensitive attributes. Table 2 summarizes the tasks and settings, listing, for each, the three most used datasets. When describing tasks and settings, we explicitly highlight datasets that are particularly relevant, even when outside of the top three. For brevity, we present a selection of tasks and settings; a thorough treatment is presented in Fabris et al. [27]

5.2.1 Task.

Fair classification [11, 25] is the most common task by far. Group fairness involves equalizing some measure of interest across subpopulations, while individual fairness focuses on ensuring similar treatment for similar individuals.Unsurprisingly, the most common datasets for fair classification are the most popular ones overall (§ 4), i.e., Adult, COMPAS, and German Credit.

Fair regression [10] concentrates on models that predict a real-valued target, requiring the average loss to be balanced across groups. Fair regression is a less popular task, often studied on the Communities and Crime dataset, where the task is predicting the rate of violent crimes in different communities.

Fair ranking [80] requires ordering candidate items based on their relevance to a current need. Fairness concerns both the people producing the items that are being ranked and those consuming the items. It is typically studied in applications of recommendation and search (MovieLens, Last.fm, Million Song Dataset, TREC Robust04).

Fair matching [48] focuses on highlighting and matching pairs of items on both sides of a two-sided 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 [17] studies algorithms that score instances in a dataset according to a predefined type of risk. 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 [20] concerns the study of features learnt by models as intermediate representations for inference tasks. 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 [16] is an unsupervised task concerned with the division of a sample into homogenous 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, other resources often used for this task include Bank Marketing, Diabetes 130-US Hospitals, Credit Card Default and US Census Data (1990).

Table 2: 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

Fair anomaly detection [83], also called **outlier detection** [22], is aimed at identifying surprising or anomalous points in a dataset. Fairness requirements involve equalizing key measures (e.g. acceptance rate, recall, distribution of anomaly scores) across populations of interest. This problem is particularly relevant for minority groups, who, in the absence of specific attention to dataset inclusivity, are less likely to fit the norm in the feature space.

Fair task assignment and **truth discovery** [33, 54] are different subproblems in the same area, focused on the subdivision of work and the aggregation of answers in crowdsourcing. Fairness may be intended concerning errors in the aggregated answer, requiring error rates to be balanced across groups, or in terms of the work load imposed to workers. A dataset suitable for this task is Crowd Judgement, containing crowd-sourced recidivism predictions.

Fair graph diffusion [29] 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** [69] is a similar task, defined on graphs which model 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 [3, 38] can 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 [12] refers to equity in data reduction. 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 groups.

Fair graph mining [44] focuses on representations and prediction on graph structures. Fairness is defined 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 [43] concerns learning and deploying an optimal pricing policy for revenue while maintaining equity of access to services and consumer welfare across groups [28]. Employed datasets are from the economics (Credit Elasticities, Italian Car Insurance), transportation (Chicago Ridesharing), and public health domains (Willingness-to-Pay for Vaccine).

Fair advertising [13] 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.

5.2.2 Setting. Most 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 [46] is a setting where fairness properties are required to hold not only for a limited number of protected groups, but across 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 [15], encrypted [47] or corrupted by noise [50]. This setting is most commonly studied on highly popular fairness dataset (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 labelled instances are few [41], noisy [78], or only available in aggregate form [70].

Robust fairness problems arise under perturbations to the training set [37], adversarial attacks [62] and dataset shift [74]. This line of research is often connected with work in robust machine learning, extending the stability requirements beyond accuracy-related metrics to fairness-related ones.

Dynamical fairness [21, 56] 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. Popular resources for this setting are FICO and the ML Fairness GYM.

Preference-based fairness [81] denotes work informedby the preferences of stakeholders. For data subjects this is related to notions of envy-freeness and loss aversion [1]; for policy-makers it permits an indication of how to trade-off different fairness measures [84] or direct demonstrations of fair outcomes [30].

Multi-stage fairness [58] refers to settings where several decision makers coexist in a compound decision-making process. Decision makers, both humans and algorithmic, may act with different levels of coordination. A fundamental question in this setting is how to ensure fairness under composition of different decision mechanisms.

Fair few-shot learning [86] aims at developing fair ML solutions in the presence of a small amount of data samples. The problem is closely related to, and possibly solved by, **fair transfer learning** [18]. 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 [4, 40] studies the interplay between privacy-preserving mechanisms and fairness constraints. 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** [53], where algorithms are trained across multiple decentralized devices, **fair incremental learning** [85], where novel classes may be added to the learning problem over time, **fair active learning** [63], allowing for the acquisition of novel information during inference, and **fair selective classification** [42], where predictions are issued only if model confidence is above a certain threshold.

6 CONCLUSIONS

Algorithmic fairness is a young research area, undergoing a fast expansion, with diverse contributions in terms of methodology and applications. Progress in the field hinges on different resources, including, very prominently, datasets. In this work, we have surveyed hundreds of datasets used in the fair ML and algorithmic equity literature to help the research community reduce its documentation debt, identify gaps, and improve the utilization of existing resources.

We have rigorously identified the most popular datasets in the literature, and carried out a thorough documentation effort for Adult, COMPAS and German Credit. Our work unifies and adds to

recent literature on data studies, calling into question their current status of general-purpose fairness benchmarks, due to contrived prediction tasks, noisy data, severe coding mistakes, limitations in encoding sensitive attributes, and age. In a practical demonstration of documentation debt and its consequences, we find several works of algorithmic fairness using German Credit with sex as a protected attribute, while careful analysis of recent documentation shows that this feature cannot be reliably retrieved from the data.

We have documented over two hundred datasets to provide viable alternatives, annotating their domain and the tasks they support in works of algorithmic fairness. We have shown that the processes generating the data belong to many different domains, including, for instance, criminal justice, education, search engines, online marketplaces, emergency response, social media, medicine, hiring, and finance. At the same time, we have described a variety of tasks studied on these resources, ranging from generic, such as *fair regression*, to narrow such as *fair districting* and *fair truth discovery*. Overall, such diversity of domains and tasks provides a glimpse into the variety of human activities and applications that can be impacted by automated decision making, and that can benefit from fair ML and algorithmic equity research.

Dataset tasks, domains, and the whole metadata are made available in our data briefs (Appendix A), which we plan to update on a yearly basis.[5] We envision several benefits for the algorithmic equity and data studies research communities, including (1) informing the choice of datasets for experimental evaluations of fair algorithms, including domain-oriented and task-oriented search, (2) directing studies of data bias, and other quantitative and qualitative analyses, including retrospective documentation efforts, towards 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 [72], concerning computer vision [26], or being popular in the fairness literature [52].

In this work, we have targeted the collective documentation debt of the algorithmic fairness community, resulting from the opacity surrounding certain resources and the sparsity of existing documentation. We have mainly targeted sparsity in a centralized documentation effort. Similarly to other types of data interventions, useful documentation can be produced after release, but, as shown in this work, the documentation debt may propagate nonetheless. In a mature research community, curators, users and reviewers can all contribute to cultivating a data documentation culture and keep the overall documentation debt in check.

ACKNOWLEDGMENTS

The authors 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 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.

[5] We aim to release a web app for dataset search at https://fairnessdatasets.dei.unipd.it/.

REFERENCES

[1] Junaid Ali, Muhammad Bilal Zafar, Adish Singla, and Krishna P. Gummadi. 2019. Loss-Aversively Fair Classification. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (AIES '19)*. Association for Computing Machinery, New York, NY, USA, 211–218. https://doi.org/10.1145/3306618.3314266

[2] Matthew Arnold, Rachel KE Bellamy, Michael Hind, Stephanie Houde, Sameep Mehta, Aleksandra Mojsilović, Ravi Nair, K Natesan Ramamurthy, Alexandra Olteanu, David Piorkowski, et al. 2019. FactSheets: Increasing trust in AI services through supplier's declarations of conformity. *IBM Journal of Research and Development* 63, 4/5 (2019), 6–1.

[3] Moshe Babaioff, Noam Nisan, and Inbal Talgam-Cohen. 2019. Fair Allocation through Competitive Equilibrium from Generic Incomes. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19)*. Association for Computing Machinery, New York, NY, USA, 180. https://doi.org/10.1145/3287560.3287582

[4] Eugene Bagdasaryan, Omid Poursaeed, and Vitaly Shmatikov. 2019. Differential Privacy Has Disparate Impact on Model Accuracy. In *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Eds.), Vol. 32. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2019/file/fc0de4e0396fff257ea362983c2dda5a-Paper.pdf

[5] Sina Baharlouei, Maher Nouiehed, Ahmad Beirami, and Meisam Razaviyayn. 2020. Rényi Fair Inference. In *International Conference on Learning Representations*. https://openreview.net/forum?id=HkgsUJrtDB

[6] Michelle Bao, Angela Zhou, Samantha Zottola, Brian Brubach, Sarah Desmarais, Aaron Horowitz, Kristian Lum, and Suresh Venkatasubramanian. 2021. It's COMPASlicated: The Messy Relationship between RAI Datasets and Algorithmic Fairness Benchmarks. *arXiv preprint arXiv:2106.05498* (2021).

[7] Matias Barenstein. 2019. ProPublica's COMPAS Data Revisited. *arXiv preprint arXiv:1906.04711* (2019).

[8] Emily M. Bender and Batya Friedman. 2018. Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. *Transactions of the Association for Computational Linguistics* 6 (2018), 587–604. https://doi.org/10.1162/tacl_a_00041

[9] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? *(FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 610–623. https://doi.org/10.1145/3442188.3445922

[10] Richard Berk, Hoda Heidari, Shahin Jabbari, Matthew Joseph, Michael Kearns, Jamie Morgenstern, Seth Neel, and Aaron Roth. 2017. A Convex Framework for Fair Regression. arXiv:cs.LG/1706.02409 KDD 2017 workshop: "Fairness, Accountability, and Transparency in Machine Learning (FAT/ML)".

[11] Toon Calders and Sicco Verwer. 2010. Three Naive Bayes Approaches for Discrimination-Free Classification. *Data Min. Knowl. Discov.* 21, 2 (Sept. 2010), 277–292. https://doi.org/10.1007/s10618-010-0190-x

[12] Elisa Celis, Vijay Keswani, Damian Straszak, Amit Deshpande, Tarun Kathuria, and Nisheeth Vishnoi. 2018. Fair and Diverse DPP-Based Data Summarization. In *Proceedings of the 35th International Conference on Machine Learning (Proceedings of Machine Learning Research)*, Jennifer Dy and Andreas Krause (Eds.), Vol. 80. PMLR, Stockholmsmässan, Stockholm Sweden, 716–725. http://proceedings.mlr.press/v80/celis18a.html

[13] Elisa Celis, Anay Mehrotra, and Nisheeth Vishnoi. 2019. Toward Controlling Discrimination in Online Ad Auctions. In *Proceedings of the 36th International Conference on Machine Learning (Proceedings of Machine Learning Research)*, Kamalika Chaudhuri and Ruslan Salakhutdinov (Eds.), Vol. 97. PMLR, Long Beach, California, USA, 4456–4465. http://proceedings.mlr.press/v97/mehrotra19a.html

[14] Binghui Chen, Weihong Deng, and Haifeng Shen. 2018. Virtual Class Enhanced Discriminative Embedding Learning. In *Advances in Neural Information Processing Systems*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (Eds.), Vol. 31. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2018/file/d79aac075930c83c2f1e369a511148fe-Paper.pdf

[15] Jiahao Chen, Nathan Kallus, Xiaojie Mao, Geoffry Svacha, and Madeleine Udell. 2019. Fairness Under Unawareness: Assessing Disparity When Protected Class Is Unobserved. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19)*. Association for Computing Machinery, New York, NY, USA, 339–348. https://doi.org/10.1145/3287560.3287594

[16] Flavio Chierichetti, Ravi Kumar, Silvio Lattanzi, and Sergei Vassilvitskii. 2017. Fair Clustering Through Fairlets. In *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc., 5029–5037. https://proceedings.neurips.cc/paper/2017/file/978fce5bcc4eccc88ad48ce3914124a2-Paper.pdf

[17] Amanda Coston, Alan Mishler, Edward H. Kennedy, and Alexandra Chouldechova. 2020. Counterfactual Risk Assessments, Evaluation, and Fairness. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 582–593. https://doi.org/10.1145/3351095.3372851

[18] Amanda Coston, Karthikeyan Natesan Ramamurthy, Dennis Wei, Kush R. Varshney, Skyler Speakman, Zairah Mustahsan, and Supriyo Chakraborty. 2019. Fair Transfer Learning with Missing Protected Attributes. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (AIES '19)*. Association for Computing Machinery, New York, NY, USA, 91–98. https://doi.org/10.1145/3306618.3314236

[19] Kate Crawford and Trevor Paglen. 2021. Excavating AI: the Politics of Images in Machine Learning Training Sets. https://excavating.ai/

[20] Elliot Creager, David Madras, Joern-Henrik Jacobsen, Marissa Weis, Kevin Swersky, Toniann Pitassi, and Richard Zemel. 2019. Flexibly Fair Representation Learning by Disentanglement. In *Proceedings of the 36th International Conference on Machine Learning (Proceedings of Machine Learning Research)*, Kamalika Chaudhuri and Ruslan Salakhutdinov (Eds.), Vol. 97. PMLR, Long Beach, California, USA, 1436–1445. http://proceedings.mlr.press/v97/creager19a.html

[21] Alexander D'Amour, Hansa Srinivasan, James Atwood, Pallavi Baljekar, D. Sculley, and Yoni Halpern. 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)*. Association for Computing Machinery, New York, NY, USA, 525–534. https://doi.org/10.1145/3351095.3372878

[22] Ian Davidson and Selvan Suntiha Ravi. 2020. A framework for determining the fairness of outlier detection. In *ECAI 2020*. IOS Press, 2465–2472.

[23] William Dieterich, Christina Mendoza, and Tim Brennan. 2016. COMPAS risk scales: Demonstrating accuracy equity and predictive parity.

[24] Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. 2021. Retiring adult: New datasets for fair machine learning. *Advances in Neural Information Processing Systems* 34 (2021).

[25] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through Awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference (ITCS '12)*. Association for Computing Machinery, New York, NY, USA, 214–226. https://doi.org/10.1145/2090236.2090255

[26] Simone Fabbrizzi, Symeon Papadopoulos, Eirini Ntoutsi, and Ioannis Kompatsiaris. 2021. A survey on bias in visual datasets. *arXiv preprint arXiv:2107.07919* (2021).

[27] Alessandro Fabris, Stefano Messina, Gianmaria Silvello, and Gian Antonio Susto. 2022. Algorithmic Fairness Datasets: the Story so Far. *Data Mining and Knowledge Discovery* (2022). https://doi.org/10.1007/s10618-022-00854-z to appear.

[28] Alessandro Fabris, Alan Mishler, Stefano Gottardi, Mattia Carletti, Matteo Daicampi, Gian Antonio Susto, and Gianmaria Silvello. 2021. *Algorithmic Audit of Italian Car Insurance: Evidence of Unfairness in Access and Pricing*. Association for Computing Machinery, New York, NY, USA, 458–468. https://doi.org/10.1145/3461702.3462569

[29] Golnoosh Farnad, Behrouz Babaki, and Michel Gendreau. 2020. A Unifying Framework for Fairness-Aware Influence Maximization. In *Companion Proceedings of the Web Conference 2020 (WWW '20)*. Association for Computing Machinery, New York, NY, USA, 714–722. https://doi.org/10.1145/3366424.3383555

[30] Sainyam Galhotra, Sandhya Saisubramanian, and Shlomo Zilberstein. 2021. *Learning to Generate Fair Clusters from Demonstrations*. Association for Computing Machinery, New York, NY, USA, 491–501. https://doi.org/10.1145/3461702.3462558

[31] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2018. Datasheets for datasets. *arXiv preprint arXiv:1803.09010* (2018).

[32] R. Stuart Geiger, Kevin Yu, Yanlai Yang, Mindy Dai, Jie Qiu, Rebekah Tang, and Jenny Huang. 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)*. Association for Computing Machinery, New York, NY, USA, 325–336. https://doi.org/10.1145/3351095.3372862

[33] Naman Goel and Boi Faltings. 2019. Crowdsourcing with Fairness, Diversity and Budget Constraints. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (AIES '19)*. Association for Computing Machinery, New York, NY, USA, 297–304. https://doi.org/10.1145/3306618.3314282

[34] Ulrike Grömping. 2019. *South German Credit Data: Correcting a Widely Used Data Set. Report*. Technical Report. Beuth University of Applied Sciences Berlin. http://www1.beuth-hochschule.de/FB_II/reports/Report-2019-004.pdf

[35] Yuzi He, Keith Burghardt, and Kristina Lerman. 2020. A Geometric Solution to Fair Representations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES '20)*. Association for Computing Machinery, New York, NY, USA, 279–285. https://doi.org/10.1145/3375627.3375864

[36] Sarah Holland, Ahmed Hosny, Sarah Newman, Joshua Joseph, and Kasia Chmielinski. 2018. The dataset nutrition label: A framework to drive higher data quality standards. *arXiv preprint arXiv:1805.03677* (2018).

[37] Lingxiao Huang and Nisheeth Vishnoi. 2019. Stable and Fair Classification. In *Proceedings of the 36th International Conference on Machine Learning (Proceedings of Machine Learning Research)*, Kamalika Chaudhuri and Ruslan Salakhutdinov (Eds.), Vol. 97. PMLR, Long Beach, California, USA, 2879–2890. http://proceedings.mlr.press/v97/huang19e.html

[38] Lingxiao Huang, Julia Wei, and Elisa Celis. 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)*. Association for Computing Machinery, New York, NY, USA, 293–299. https://doi.org/10.1145/3375627.3375848

[39] Ben Hutchinson, Andrew Smart, Alex Hanna, Emily Denton, Christina Greer, Oddur Kjartansson, Parker Barnes, and Margaret Mitchell. 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)*. Association for Computing Machinery, New York, NY, USA, 560–575.

[40] Matthew Jagielski, Michael Kearns, Jieming Mao, Alina Oprea, Aaron Roth, Saeed Sharifi Malvajerdi, and Jonathan Ullman. 2019. Differentially Private Fair Learning. In *Proceedings of the 36th International Conference on Machine Learning (Proceedings of Machine Learning Research)*, Kamalika Chaudhuri and Ruslan Salakhutdinov (Eds.), Vol. 97. PMLR, Long Beach, California, USA, 3000–3008. http://proceedings.mlr.press/v97/jagielski19a.html

[41] Disi Ji, Padhraic Smyth, and Mark Steyvers. 2020. Can I Trust My Fairness Metric? Assessing Fairness with Unlabeled Data and Bayesian Inference. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (Eds.). https://proceedings.neurips.cc/paper/2020/hash/d83de59e10227072a9c034ce10029c39-Abstract.html

[42] Erik Jones, Shiori Sagawa, Pang Wei Koh, Ananya Kumar, and Percy Liang. 2021. Selective Classification Can Magnify Disparities Across Groups. In *International Conference on Learning Representations*. https://openreview.net/forum?id=N0M_4BkQ05i

[43] Nathan Kallus and Angela Zhou. 2021. Fairness, Welfare, and Equity in Personalized Pricing. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 296–314. https://doi.org/10.1145/3442188.3445895

[44] Jian Kang, Jingrui He, Ross Maciejewski, and Hanghang Tong. 2020. *InFoRM: Individual Fairness on Graph Mining*. Association for Computing Machinery, New York, NY, USA, 379–389. https://doi.org/10.1145/3394486.3403080

[45] Masahiro Kato, Takeshi Teshima, and Junya Honda. 2019. Learning from Positive and Unlabeled Data with a Selection Bias. In *International Conference on Learning Representations*. https://openreview.net/forum?id=rJzLciCqKm

[46] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. 2018. Preventing Fairness Gerrymandering: Auditing and Learning for Subgroup Fairness. In *Proceedings of the 35th International Conference on Machine Learning (Proceedings of Machine Learning Research)*, Jennifer Dy and Andreas Krause (Eds.), Vol. 80. PMLR, Stockholmsmässan, Stockholm Sweden, 2564–2572. http://proceedings.mlr.press/v80/kearns18a.html

[47] Niki Kilbertus, Adria Gascon, Matt Kusner, Michael Veale, Krishna Gummadi, and Adrian Weller. 2018. Blind Justice: Fairness with Encrypted Sensitive Attributes. In *Proceedings of the 35th International Conference on Machine Learning (Proceedings of Machine Learning Research)*, Jennifer Dy and Andreas Krause (Eds.), Vol. 80. PMLR, Stockholmsmässan, Stockholm Sweden, 2630–2639. http://proceedings.mlr.press/v80/kilbertus18a.html

[48] Ari Kobren, Barna Saha, and Andrew McCallum. 2019. Paper Matching with Local Fairness Constraints. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '19)*. Association for Computing Machinery, New York, NY, USA, 1247–1257. https://doi.org/10.1145/3292500.3330899

[49] Bernard Koch, Emily Denton, Alex Hanna, and Jacob Gates Foster. 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)*. https://openreview.net/forum?id=zNQBIBKJRkd

[50] Alex Lamy, Ziyuan Zhong, Aditya K Menon, and Nakul Verma. 2019. Noise-tolerant fair classification. In *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Eds.), Vol. 32. Curran Associates, Inc., 294–306. https://proceedings.neurips.cc/paper/2019/file/8d5e957f297893487bd98fa830fa6413-Paper.pdf

[51] Jeff Larson, Surya Mattu, Lauren Kirchner, and Julia Angwin. 2016. How We Analyzed the COMPAS Recidivism Algorithm. https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

[52] Tai Le Quy, Arjun Roy, Vasileios Iosifidis, Wenbin Zhang, and Eirini Ntoutsi. 2022. A survey on datasets for fairness-aware machine learning. *WIREs Data Mining and Knowledge Discovery* n/a, n/a (2022), e1452. https://doi.org/10.1002/widm.1452 arXiv:https://wires.onlinelibrary.wiley.com/doi/pdf/10.1002/widm.1452

[53] Tian Li, Maziar Sanjabi, Ahmad Beirami, and Virginia Smith. 2020. Fair Resource Allocation in Federated Learning. In *International Conference on Learning Representations*. https://openreview.net/forum?id=ByexElSYDr

[54] Yanying Li, Haipei Sun, and Wendy Hui Wang. 2020. *Towards Fair Truth Discovery from Biased Crowdsourced Answers*. Association for Computing Machinery, New York, NY, USA, 599–607. https://doi.org/10.1145/3394486.3403102

[55] Zhi Li, Hongke Zhao, Qi Liu, Zhenya Huang, Tao Mei, and Enhong Chen. 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)*. Association for Computing Machinery, New York, NY, USA, 1734–1743. https://doi.org/10.1145/3219819.3220014

[56] Lydia T. Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. 2018. Delayed Impact of Fair Machine Learning. In *Proceedings of the 35th International Conference on Machine Learning (Proceedings of Machine Learning Research)*, Jennifer Dy and Andreas Krause (Eds.), Vol. 80. PMLR, Stockholmsmässan, Stockholm Sweden, 3150–3158. http://proceedings.mlr.press/v80/liu18c.html

[57] Michael Lohaus, Michael Perrot, and Ulrike Von Luxburg. 2020. Too Relaxed to Be Fair. In *Proceedings of the 37th International Conference on Machine Learning (Proceedings of Machine Learning Research)*, Hal Daumé III and Aarti Singh (Eds.), Vol. 119. PMLR, Virtual, 6360–6369. http://proceedings.mlr.press/v119/lohaus20a.html

[58] David Madras, Toni Pitassi, and Richard Zemel. 2018. Predict Responsibly: Improving Fairness and Accuracy by Learning to Defer. In *Advances in Neural Information Processing Systems*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (Eds.), Vol. 31. Curran Associates, Inc., 6147–6157. https://proceedings.neurips.cc/paper/2018/file/09d37c08f7b129e96277388757530c72-Paper.pdf

[59] Natalia Martinez, Martin Bertran, and Guillermo Sapiro. 2020. Minimax Pareto Fairness: A Multi Objective Perspective. In *Proceedings of the 37th International Conference on Machine Learning (Proceedings of Machine Learning Research)*, Hal Daumé III and Aarti Singh (Eds.), Vol. 119. PMLR, Virtual, 6755–6764. http://proceedings.mlr.press/v119/martinez20a.html

[60] Alan Mishler, Edward H. Kennedy, and Alexandra Chouldechova. 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)*. Association for Computing Machinery, New York, NY, USA, 386–400. https://doi.org/10.1145/3442188.3445902

[61] Jeffrey C Moore, Linda L Stinson, and Edward J Welniak. 2000. Income measurement error in surveys: A review. *Journal of Official Statistics-Stockholm-* 16, 4 (2000), 331–362.

[62] Vedant Nanda, Samuel Dooley, Sahil Singla, Soheil Feizi, and John P. Dickerson. 2021. Fairness Through Robustness: Investigating Robustness Disparity in Deep Learning. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 466–477. https://doi.org/10.1145/3442188.3445910

[63] Alejandro Noriega-Campero, Michiel A. Bakker, Bernardo Garcia-Bulle, and Alex 'Sandy' Pentland. 2019. Active Fairness in Algorithmic Decision Making. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (AIES '19)*. Association for Computing Machinery, New York, NY, USA, 77–83. https://doi.org/10.1145/3306618.3314277

[64] Partnership on AI. 2022. *About ML*. Technical Report. https://partnershiponai.org/workstream/about-ml/

[65] Amandalynne Paullada, Inioluwa Deborah Raji, Emily M Bender, Emily Denton, and Alex Hanna. 2020. Data and its (dis) contents: A survey of dataset development and use in machine learning research. *arXiv preprint arXiv:2012.05345* (2020).

[66] Kenny Peng, Arunesh Mathur, and Arvind Narayanan. 2021. Mitigating dataset harms requires stewardship: Lessons from 1000 papers. *arXiv preprint arXiv:2108.02922* (2021).

[67] Valerio Perrone, Michele Donini, Muhammad Bilal Zafar, Robin Schmucker, Krishnaram Kenthapadi, and Cédric Archambeau. 2021. *Fair Bayesian Optimization*. Association for Computing Machinery, New York, NY, USA, 854–863. https://doi.org/10.1145/3461702.3462629

[68] ProPublica. 2016. COMPAS analysis github repository. https://github.com/propublica/compas-analysis

[69] Govardana Sachithanandam Ramachandran, Ivan Brugere, Lav R. Varshney, and Caiming Xiong. 2021. *GAEA: Graph Augmentation for Equitable Access via Reinforcement Learning*. Association for Computing Machinery, New York, NY, USA, 884–894. https://doi.org/10.1145/3461702.3462615

[70] Sivan Sabato and Elad Yom-Tov. 2020. Bounding the fairness and accuracy of classifiers from population statistics. In *Proceedings of the 37th International Conference on Machine Learning (Proceedings of Machine Learning Research)*, Hal Daumé III and Aarti Singh (Eds.), Vol. 119. PMLR, Virtual, 8316–8325. http://proceedings.mlr.press/v119/sabato20a.html

[71] Morgan Klaus Scheuerman, Alex Hanna, and Emily Denton. 2021. Do datasets have politics? Disciplinary values in computer vision dataset development. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–37.

[72] Morgan Klaus Scheuerman, Kandrea Wade, Caitlin Lustig, and Jed R. Brubaker. 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, Article 058 (May 2020), 35 pages. https://doi.org/10.1145/3392866

[73] Shubham Sharma, Alan H. Gee, David Paydarfar, and Joydeep Ghosh. 2021. *FaiR-N: Fair and Robust Neural Networks for Structured Data*. Association for Computing Machinery, New York, NY, USA, 946–955. https://doi.org/10.1145/3461702.3462559

[74] Harvineet Singh, Rina Singh, Vishwali Mhasawade, and Rumi Chunara. 2021. Fairness Violations and Mitigation under Covariate Shift. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 3–13. https://doi.org/10.1145/3442188.3445865

[75] UCI Machine Learning Repository. 1994. Statlog (German Credit Data) Data Set. https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)

[76] UCI Machine Learning Repository. 2019. South German Credit Data Set. https://archive.ics.uci.edu/ml/datasets/South+German+Credit

[77] US Dept. of Commerce Bureau of the Census. 1995. Current Population Survey: Annual Demographic File, 1994.

[78] Jialu Wang, Yang Liu, and Caleb Levy. 2021. Fair Classification with Group-Dependent Label Noise. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 526–536. https://doi.org/10.1145/3442188.3445915

[79] Forest Yang, Mouhamadou Cisse, and Sanmi Koyejo. 2020. Fairness with Overlapping Groups; a Probabilistic Perspective. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 4067–4078. https://proceedings.neurips.cc/paper/2020/file/29c0605a3bab4229e46723f89cf59d83-Paper.pdf

[80] Ke Yang and Julia Stoyanovich. 2017. Measuring Fairness in Ranked Outputs. In *Proceedings of the 29th International Conference on Scientific and Statistical Database Management (SSDBM '17)*. Association for Computing Machinery, New York, NY, USA, Article 22, 6 pages. https://doi.org/10.1145/3085504.3085526

[81] Muhammad Bilal Zafar, Isabel Valera, Manuel Rodriguez, Krishna Gummadi, and Adrian Weller. 2017. From Parity to Preference-based Notions of Fairness in Classification. In *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc., 229–239. https://proceedings.neurips.cc/paper/2017/file/82161242827b703e6acf9c726942a1e4-Paper.pdf

[82] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P Gummadi. 2017. Fairness Constraints: Mechanisms for Fair classification. In *Artificial Intelligence and Statistics*. PMLR, 962–970.

[83] Hongjing Zhang and Ian Davidson. 2021. Towards Fair Deep Anomaly Detection. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 138–148. https://doi.org/10.1145/3442188.3445878

[84] Yunfeng Zhang, Rachel Bellamy, and Kush Varshney. 2020. 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)*. Association for Computing Machinery, New York, NY, USA, 400–406. https://doi.org/10.1145/3375627.3375862

[85] Bowen Zhao, Xi Xiao, Guojun Gan, Bin Zhang, and Shu-Tao Xia. 2020. Maintaining Discrimination and Fairness in Class Incremental Learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[86] Chen Zhao, Changbin Li, Jincheng Li, and Feng Chen. 2020. Fair Meta-Learning For Few-Shot Classification. In *2020 IEEE International Conference on Knowledge Graph (ICKG)*. 275–282. https://doi.org/10.1109/ICBK50248.2020.00047

[87] Yunhan Zhao, Shu Kong, and Charless Fowlkes. 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)*. 15759–15768.

[88] Yaoyao Zhong, Weihong Deng, Mei Wang, Jiani Hu, Jianteng Peng, Xunqiang Tao, and Yaohai Huang. 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)*.