Noname manuscript No.
(will be inserted by the editor)

Algorithmic Fairness Datasets: the Story so Far

Alessandro Fabris · Stefano Messina ·
Gianmaria Silvello · Gian Antonio Susto

Received: date / Accepted: date

Abstract Data-driven algorithms are studied and deployed in diverse domains to support critical decisions, directly impacting people's well-being. As a result, a growing community of researchers has been investigating the equity of existing algorithms and proposing novel ones, advancing the understanding of risks and opportunities of automated decision-making for historically disadvantaged populations. Progress in fair Machine Learning (ML) and equitable algorithm design hinges on data, which can be appropriately used only if adequately documented. Unfortunately, the algorithmic fairness 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 target this data documentation debt by surveying over two hundred datasets employed in algorithmic fairness research, and 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 supports multiple contributions. Firstly, 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. Secondly, we document hundreds of available alternatives, annotating their domain and supported fairness tasks, along with additional properties of interest for fairness practitioners and researchers, including their format, cardinality, and the sensitive attributes they encode. We summarize this information, zooming in on the tasks, domains, and roles of these resources. Finally, we analyze these datasets from the perspective of five important data curation topics: anonymization,

Dipartimento di Ingegneria dell'Informazione
Università di Padova
Via Giovanni Gradenigo 6B – 35131 Padova, Italy
E-mail: fabrisal@dei.unipd.it, silvello@dei.unipd.it, gianantonio.susto@unipd.it

consent, inclusivity, labeling of sensitive attributes, and transparency. 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.

Keywords Algorithmic Fairness · Datasets · Documentation Debt

1 Introduction

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, investigating algorithms through the lens of justice, equity, bias, power, and harms. A line of work gaining 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 (Gebru et al., 2018; Holland et al., 2018; Bender and Friedman, 2018; Geiger et al., 2020; Jo and Gebru, 2020; Miceli et al., 2021). Most prominently, Gebru et al. (2018) and Holland et al. (2018) proposed two complementary documentation frameworks, called *Datasheets for Datasets* and *Dataset Nutrition Labels*, 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 Machine Learning (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 (Bandy and Vincent, 2021; Garbin et al., 2021), auditing their properties (Prabhu and Birhane, 2020) and tracing their usage (Peng et al., 2021).

Data documentation is important and caters to different goals. It increases transparency, favouring an improved understanding of the data and resulting models (Hutchinson et al., 2021), it reduces chances of data misuse (Gebru et al., 2018) and supports accountability in dataset and model creation (Hutchinson et al., 2021), it helps connect the data with its context to guide scientific inquiry (Paullada et al., 2020), and it makes the values influencing dataset curation explicit (Scheuerman et al., 2021). Technical debt is a cost incurred in software development when speed of execution is prioritized over quality (Hutchinson et al., 2021). In recent work, Bender et al. (2021) propose the notion of *documentation debt*, in relation to training sets that are undocumented and too large to document retrospectively, which 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

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

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 (UCI Machine Learning Repository, 1994), 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 (He et al., 2020b; Yang et al., 2020a; Baharlouei et al., 2020; Lohaus et al., 2020; Martinez et al., 2020; Wang et al., 2021; Perrone et al., 2021; Sharma et al., 2021), while existing yet overlooked documentation shows that this feature cannot be reliably retrieved (Grömping, 2019). Moreover, the mere fact that a dataset exists and is relevant to a given task or a given domain may be unknown. The BUPT Faces datasets, for instance, were presented as the second existing resource for face analysis with race annotations (Wang and Deng, 2020). However several resources were already available at the time, including Labeled Faces in the Wild (Han and Jain, 2014), UTK Face (Zhang et al., 2017b), Racial Faces in the Wild (Wang et al., 2019e), and Diversity in Faces (Merler et al., 2019).[2]

To tackle the documentation debt of the algorithmic fairness community, we survey the datasets used in over 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 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 and a nutrition label for each of them. From these documentation efforts, we extract a summary of the merits and limitations of popular algorithmic fairness benchmarks, a categorization of domains and fairness tasks for existing datasets, and a set of best practices for the curation of novel resources.

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

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, and age.
- **Survey of existing alternatives**. We compile standardized and compact documentation for over two hundred resources used in fair ML research, annotating their domain, the tasks they support, and the roles they play 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 categorization of tasks and domains investigated in algorithmic fairness research (summarized in Tables 2 and 3).
- **Best practices for the curation of novel resources**. We analyze different approaches to anonymization, consent, inclusivity, labeling, and transparency across these datasets. By comparing existing approaches and discussing their advantages, we make the underlying concerns visible and practical, and extract best practices to inform the curation of new datasets and post-hoc remedies to existing ones.

Roadmap. Readers looking for alternative fairness datasets should prioritize Section 5, Appendix A, and take account of the web app under development (see Footnote 10). 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, namely Adult (§ 4.1), COMPAS (§ 4.2), and German Credit (§ 4.3), and provides an overall summary of their merits and limitations as fairness benchmarks (§ 4.4). Section 5 discusses alternative fairness resources from the perspective of the underlying domains (§ 5.1), the fair ML tasks they support (§ 5.2), and the roles they play (§ 5.3). Section 6 presents important topics in data curation, discussing existing approaches and best practices to avoid re-identification (§ 6.1), elicit informed consent (§ 6.2), consider inclusivity (§ 6.3), collect sensitive attributes (§ 6.4), and document datasets (§ 6.5). Section 7 summarizes the broader benefits of our documentation effort and envisioned uses for the research community. Finally, Section 8 contains concluding remarks and recommendations. Interested readers may find the data briefs in Appendix A, followed by the detailed documentation produced for Adult (Appendix B), COMPAS (Appendix C), and German Credit (Appendix D).

2 Related Work

2.1 Algorithmic fairness surveys

Multiple surveys about algorithmic fairness have been published in the literature (Mehrabi et al., 2021; Caton and Haas, 2020; Pessach and Shmueli, 2020). 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 (Mehrabi et al., 2021), software packages and projects which address fairness in ML (Caton and Haas, 2020), or describe selected sub-fields of algorithmic fairness (Pessach and Shmueli, 2020). Datasets are typically not emphasized in these works, which is also true of domain-specific surveys on algorithmic fairness, focused e.g. on ranking (Pitoura et al., 2021), Natural Language Processing (NLP) (Sun et al., 2019) and computational medicine (Sun et al., 2019). As an exception, Pessach and Shmueli (2020) and Zehlike et al. (2021) list and briefly describe 12 popular algorithmic fairness datasets, and 19 datasets employed in fair ranking research, respectively.

2.2 Data studies

The work most closely related (and concurrently carried out) to ours is Le Quy et al. (2022). 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, and (4) the analysis and distillation of best practices for data curation.Different goals of the research community, such as selection of 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 (2021) focus on resources commonly used as training sets in computer vision, with attention to associated labels and underlying taxonomies. Fabbrizzi et al. (2021) also consider computer vision datasets, describing types of bias affecting them, along with methods for discovering and measuring bias, while Scheuerman et al. (2021) analyze the values encoded in their documentation. Koch et al. (2021) 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. (2021) 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.

(2020) evaluate transparency in the documentation of labeling practices employed in over 100 datasets about Twitter. Leonelli and Tempini (2020) study practices of collection, cleaning, visualization, sharing, and analysis across a variety of research domains. Romei and Ruggieri (2014) 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 (2020) focus on Imagenet (ILSVRC 2012) which they analyze along the lines of consent, problematic content, and individual re-identification. Kizhner et al. (2020) 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 (Garbin et al., 2021) and the BookCorpus (Bandy and Vincent, 2021). Among popular fairness datasets, COMPAS has drawn scrutiny from multiple works, analysing its numerical idiosyncrasies (Barenstein, 2019) and sources of bias (Bao et al., 2021). Ding et al. (2021) study numerical idiosyncrasies in the Adult dataset, and propose a novel version, for which they provide a datasheet. Grömping (2019) discuss issues resulting from coding mistakes in German Credit.

Our work combines the breadth of multi-dataset and the depth of single-dataset studies. On one hand, we survey 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 perks and limitations. Moreover, by making our data briefs available, we hope to contribute a useful tool to the research community, favouring further data studies and analyses, as outlined in Section 7.

2.3 Documentation frameworks

Several data documentation frameworks have been proposed in the literature; three popular ones are described below. *Datasheets for Datasets* (Gebru et al., 2018) 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* (Bender and Friedman, 2018), which is tailored for NLP, requiring domain-specific information on language variety and speaker demographics. *Dataset Nutrition Labels* (Holland et al., 2018) 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 (Arnold et al., 2019; Partnership on AI, 2022).

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.[3] 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.[4]

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 (Zhao et al., 2021), compensating selection bias to improve accuracy without attention to sensitive attributes (Kato et al., 2019), enhancing desirable discriminating properties of models (Chen et al., 2018a), or generally focused on model performance (Li et al., 2018; Zhong et al., 2019). 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] `https://facctconference.org/network/`

[4] 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.

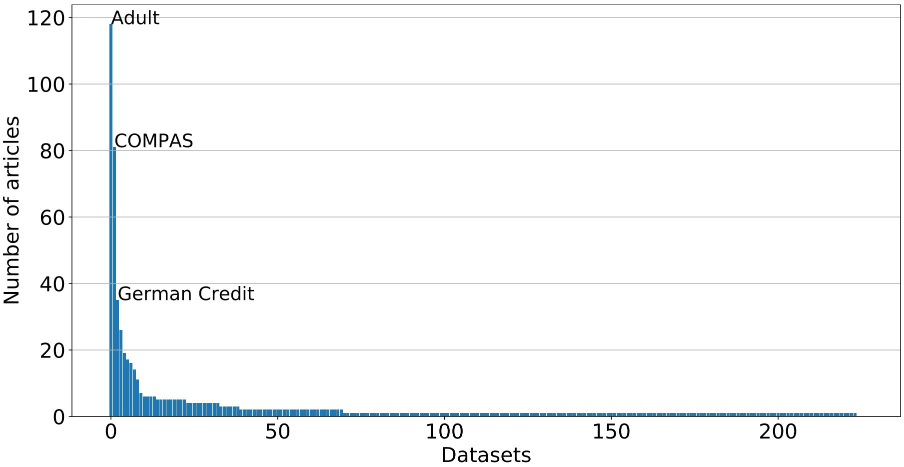

Fig. 1: Utilization of datasets in fairness research follows a long tail distribution.

(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 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;
– 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 seldom when points (1), (2), and (3) outlined above result in little to no information about 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 with a description of the underlying coding procedure. 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 (Koch et al., 2021). 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. (2017c), 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. 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.

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. Attributes in the dataset 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 (US Dept. of Commerce Bureau of the Census, 1995), 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 (Ding et al., 2021). 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 (Moore et al., 2000). Another source of noise is top-coding of the variable "capital-gain" (saturation to $99,999) to avoid the re-identification of certain individuals (US Dept. of Commerce Bureau of the Census, 1995). 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 (Ding et al., 2021).

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

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. 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.

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 (Larson et al., 2016). Moreover, in its official release (ProPublica, 2016), the COMPAS dataset features redundant variables and data leakage due to spuriously time-dependent recidivism rates (Barenstein, 2019). 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 (Mishler et al., 2021). The lack of a standard preprocessing protocol beyond the one by ProPublica (ProPublica, 2016), 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, several risk factors considered by the COMPAS algorithm are absent from the dataset (Dieterich et al., 2016). As an additional concern, race categories lack Native Hawaiian or Other Pacific Islander, while Hispanic is redefined as race instead of ethnicity (Bao et al., 2021). 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 which can be termed *data bias*. With terminology from Friedler et al. (2021), when it comes to datasets encoding complex human phenomena, there is often a disconnect between the *construct space* (what we aim 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 across individuals or groups. COMPAS, for example, is a dataset about criminal offense. Offense is central to the prediction target Y, aimed at encoding 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 (Larson et al., 2016), and are thus clearly influenced by spatially differentiated policing practices (Fogliato et al., 2021). This is also true of criminal history encoded in COMPAS covariates, again mediated by arrest and policing practices which may be racially biased (Bao et al., 2021; Mayson, 2018). As a result, the true fairness of an algorithm, just like its accuracy, may differ significantly from what is reported on biased data. For example, algorithms that achieve equality of true positive rates across sensitive groups on COMPAS are deemed fair under the *equal opportunity* measure (Hardt et al., 2016). 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 we ultimately care about (Friedler et al., 2021).

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. (2021) 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 (Grömping, 2019). 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 (UCI Machine Learning Repository, 1994). A revised version with correct variable encodings, called South German Credit, was donated to UCI Machine Learning Repository (2019) with an accompanying report (Grömping, 2019).

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 (UCI Machine Learning Repository, 1994) 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 (Le Quy et al., 2022). This coding error, along with others discussed in Grömping (2019) was corrected in a novel release of the dataset (UCI Machine Learning Repository, 2019). 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 (Wang et al., 2021; Le Quy et al., 2022). These coding mistakes are part of a documentation debt whose influence continues to affect the algorithmic fairness community.

4.4 Summary

Adult, COMPAS, and German Credit are the most used datasets in the surveyed algorithmic fairness literature, despite the limitations summarized in Table 1. Their status as de facto fairness benchmarks is probably due to their use in seminal works (Pedreshi et al., 2008; Calders et al., 2009) and influential articles (Angwin et al., 2016) 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 (Bao et al., 2021). On 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 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. (2021) 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. In section 5.3 we discuss the different roles played by these datasets in fairness research, such as supporting training and benchmarking.

5.1 Domain

Algorithmic fairness concerns arise in any domain where Automated Decision Making (ADM) systems may influence human well-being. Unsurprisingly, the datasets in our survey 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 2, we report a subdivision of the surveyed datasets in different macrodomains.

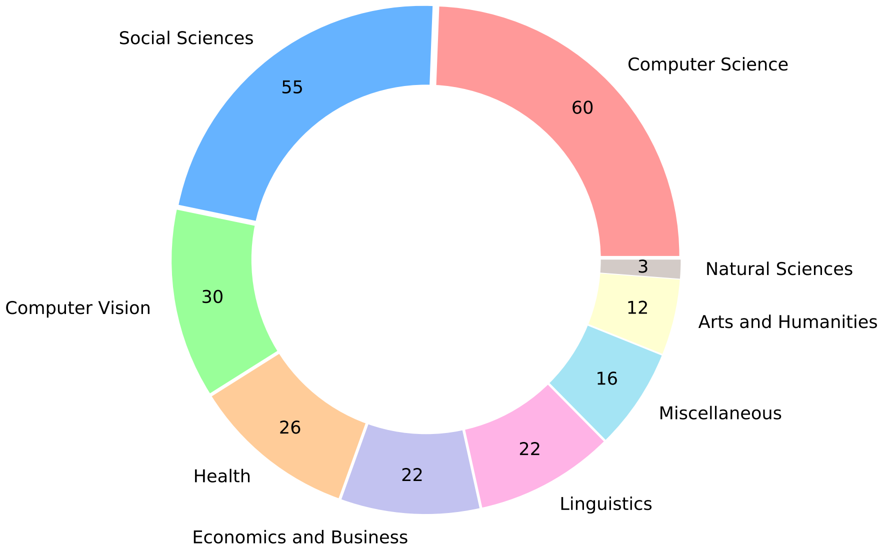

Fig. 2: Datasets employed in fairness research span diverse domains. See Table 2 for a detailed breakdown.

We mostly follow the area-category taxonomy by Scimago,[5] 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 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 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 sci-

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

ence 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, 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 (Multitask 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).

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.

As a whole, many of these datasets encode fundamental human activities where algorithms and ADM systems have been studied and deployed. Alertness and attention to equity seems 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 o ensive 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 Classi cation
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
tari s	US Harmonized Tari Schedule
insurance	Italian Car Insurance
division of goods	Spliddit Divide Goods
nance	
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 2: A selection of datasets through the lens of the domain taxonomy.

5.2 Task and setting

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 (Chouldechova and Roth, 2020), also thanks to practitioner surveys (Holstein et al., 2019) and studies of specific legal challenges (Andrus et al., 2021). 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*, describing them in the second part of this section. While our work focuses on fair ML datasets, it is cognizant of the wide variety of tasks tackled in the algorithmic fairness literature, which are captured in a specific field of our data briefs. In this section, we provide an overview of common tasks and settings studied on these datasets, showing their variety and diversity. Table 3 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.

5.2.1 Task

Fair classification (Calders and Verwer, 2010; Dwork et al., 2012) is the most common task by far. Typically, it involves equalizing some measure of interest across subpopulations, such as the 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 ones overall (§ 4), i.e., Adult, COMPAS, and German Credit.

Fair regression (Berk et al., 2017) 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 predicting the rate of violent crimes in different communities.

Fair ranking (Yang and Stoyanovich, 2017) requires ordering candidate items based on their relevance to a current need. Fairness concerns 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, Last.fm, Million Song Dataset, Amazon Recommendations, Adressa) and search engines (Yahoo! c14B Learning to Rank, Microsoft Learning to Rank, TREC Robust04).

Table 3: Most used datasets by algorithmic fairness task and setting.

Task	Datasets
Fair classi cation	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 di usion/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 O ensive 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 classi cation	CheXpert; CelebA; Civil Comments

Fair matching (Kobren et al., 2019) is similar to ranking as they are both tasks defined on two-sided markets. This task, however, 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 (Coston et al., 2020) 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 impacting 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 (Creager et al., 2019) concerns the study of features learnt 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 (Chierichetti et al., 2017) 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).

Fair anomaly detection (Zhang and Davidson, 2021), also called **outlier detection** (Davidson and Ravi, 2020), 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 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 (Schutzman, 2020) 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 precinct-level aggregated information about demographics and political leaning of voters in US districts.

Fair task assignment and **truth discovery** (Goel and Faltings, 2019; Li et al., 2020d) 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 spatio-temporal process learning (Shang et al., 2020) focuses on the estimation of models for processes which evolve in time and space. Surveyed 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 neighbourhoods and special attention to their racial composition.

Fair graph diffusion (Farnad et al., 2020) 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** (Ramachandran et al., 2021) 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 (Babaioff et al., 2019; Huang et al., 2020) 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 (Celis et al., 2018) 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 data generation (Xu et al., 2018) 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 learnt 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 (Burke et al., 2018a).

Fair graph mining (Kang et al., 2020) 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 (Kallus and Zhou, 2021) concerns learning and deploying an optimal pricing policy for revenue while maintaining equity of access to services and consumer welfare across groups. 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 (Celis et al., 2019a) 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 (Qian et al., 2015) 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 across all drivers.

Fair entity resolution (Cotter et al., 2019) is a task focused on deciding whether multiple records refer to the same entity, which is useful, for instance, 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 (Kiritchenko and Mohammad, 2018) 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) (Bolukbasi et al., 2016) 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) (Bordia and Bowman, 2019) 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) (Stanovsky et al., 2019) concerns automatic translation of text from a source language into a target one. 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 (Tatman, 2017) requires 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.

5.2.2 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 the surveyed 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 (Kearns et al., 2018) 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 (Chen et al., 2019a), encrypted (Kilbertus et al., 2018) or corrupted by noise (Lamy et al., 2019). 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 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 (Ji et al., 2020), noisy (Wang et al., 2021), or only available in aggregate form (Sabato and Yom-Tov, 2020).

Robust fairness problems arise under perturbations to the training set (Huang and Vishnoi, 2019), adversarial attacks (Nanda et al., 2021) and dataset shift (Singh et al., 2021). 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 (Liu et al., 2018; D'Amour et al., 2020) 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 (Zafar et al., 2017b) denotes work informed, explicitly or implicitly, by the preferences of stakeholders. For data subjects this is related to notions of envy-freeness and loss aversion (Ali et al., 2019b); for policy-makers it permits an indication of how to trade-off different fairness measures (Zhang et al., 2020c) or direct demonstrations of fair outcomes (Galhotra et al., 2021).

Multi-stage fairness (Madras et al., 2018b) 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 (Zhao et al., 2020b) 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** (Coston et al., 2019), 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 (Bagdasaryan et al., 2019; Jagielski et al., 2019) studies the interplay between privacy-preserving mechanisms and fairness constraints. Works in this space consider the equity of machine learning models designed to avoid 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** (Li et al., 2020b), where algorithms are trained across multiple decentralized devices, **fair incremental learning** (Zhao et al., 2020a), where novel classes may be added to the learning problem over time, **fair active learning** (Noriega-Campero et al., 2019), allowing for the acquisition of novel information during inference, and **fair selective classification** (Jones et al., 2021), 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 surveyed datasets.

5.3 Role

The 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. (2019), describes six different roles of datasets in machine learning. We adopt their framework to analyse fair ML datasets, adding to the taxonomy two roles that are specific to fairness research.

A source of real data. While synthetic datasets 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 progress. Datasets can spur 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 which 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 (Angwin et al., 2016) 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 surpassing these baselines are often considered a sign that the task is "solved" and that harder benchmarks are required (Barocas et al., 2019). 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 providing 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. (2019), 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 instance, are often used to train word embeddings and investigate their biases (Brunet et al., 2019; Liang and Acuna, 2020; Papakyriakopoulos et al., 2020). Several algorithmic fairness works aim to study and mitigate undesirable biases in learnt 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 learnt from training sets that encode relations between features and target variable in a representative fashion. One example from the fairness literature is Large Movie Review, used to train sentiment analysis models, later audited for fairness (Liang and Acuna, 2020). 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 (Brunet et al., 2019). 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. (2020) 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 instance, 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 (Yang et al., 2020b) and non-consensuality (Prabhu and Birhane, 2020). 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 (Moreland et al., 2020). To evaluate its representativeness for the overall US population, Coston et al. (2021) have studied selection biases in this dataset.

In algorithmic fairness research, datasets play similar roles to the ones they play in machine learning according to Barocas et al. (2019), 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, likely 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.

In sum, existing fairness datasets originate from a variety of domains, support diverse tasks, and play different roles in the algorithmic fairness literature. We hope our work will contribute to establishing principled data practices in the field, to guide an optimal usage of these resources. In the next section we continue our discussion on the key features of these datasets with a change of

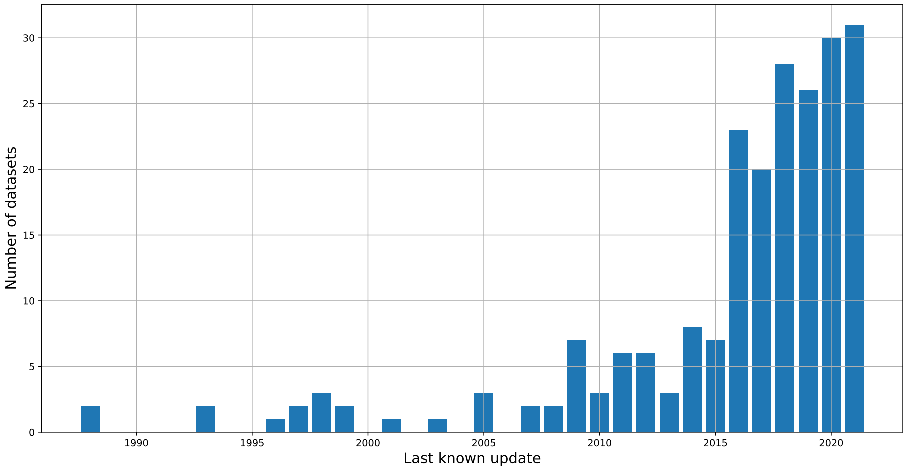

Fig. 3: Most datasets employed in algorithmic fairness were created or updated after 2015, with a clear growth in recent years.

perspective, asking which lessons can be learnt from existing resources for the curation of novel ones.

6 Best Practices for Dataset Curation

In this section, we analyze the surveyed datasets from different perspectives, typical of critical data studies, human-computer interaction, and computer-supported cooperative work. In particular, we discuss concerns of re-identification (§ 6.1), consent (§ 6.2), inclusivity (§ 6.3), sensitive attribute labeling (§ 6.4) and transparency (§ 6.5). We describe a range of approaches and consideration to these topics, ranging from negligent to conscientious. Our aim is to make these concerns and related desiderata more visible and concrete, to help inform responsible curation of novel fairness resources, whose number has been increasing in recent years (Figure 3).

6.1 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 successfully mapped back to the respective individuals. Their identity is thus discovered and associated with the information encoded in the dataset features. Examples of external re-identification attacks include de-anonymization of movie ratings from the Netflix prize dataset (Narayanan and Shmatikov, 2008), identification of profiles based on social media group membership (Wondracek et al., 2010), and identification of people depicted in verifiably pornographic categories of ImageNet (Prabhu and Birhane, 2020). These analyses were carried out as “attacks” by external teams for demonstrative purposes, but dataset curators and stakeholders may undertake similar efforts internally (McKenna, 2019b).

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 tremendous, including phishing attacks, blackmail, threat, and manipulation (Kröger et al., 2021). 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 (2020) 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. In this work, we find that 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 (El Emam et al., 2012). Indeed, health is a domain where a culture of patient record confidentiality is widely established and there is a 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. (2019) 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. may help to reduce the risk of re-identification. A first set of approaches concerns the distribution of data artefacts. 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 are mixed release strategies: NLSY has some publicly available data, while access to further information that may favour 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). While 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 (Yang et al., 2020b).

Table 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, 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 low. In this work we survey four popular ones, taken from Zafar et al. (2017c); Donini et al. (2018); Lipton et al. (2018); Singh and Joachims (2019). 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. (2019) on MS-Celeb-1M. Data synthesis may also be applied to augment datasets with artificial sensitive attributes in a principled fashion (MovieLens – (Burke et al., 2018a)). Finally, resources designed to externally probe services, algorithms, and platforms, to estimate the direct effect of a feature of interest (e.g. gender, race), may rely on hypothetical profiles (Bertrand and Mullainathan, 2004; Fabris et al., 2021). This approach can support evaluations of *fairness through unawareness* (Grgic-Hlaca et al., 2016), of which Italian Car Insurance is an example.

One 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 typically considered conflicting objectives for a dataset (Wieringa et al., 2021). 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 afford 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 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.

6.2 Consent

Motivation. In the context of data, *informed consent* is an agreement between a data processor and a subject, aimed at allowing 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 (European Union, 2016), requiring it to be freely given, specific, informed, and unambiguous. Paullada et al. (2020) 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 thus an important tool in a healthy data ecosystem that favours 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. According to the Creative Commons organization, however, their licenses are not suited to protect privacy and cover research ethics (Merkley, 2019). In computer vision, and especially in face recognition, consent and copyright are often considered and discussed jointly, and Creative Commons licenses are frequently taken as an all-inclusive permit encompassing intellectual property, consent, and ethics (Prabhu and Birhane, 2020). Merler et al. (2019), 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”. Indeed lack of consent is a widespread and far-reaching problem in face recognition datasets (Keyes et al., 2019). Prabhu and Birhane (2020) 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 (Guo et al., 2016b). As suggested by its name, the dataset should feature only celebrities, “to enable our training, testing, and re-distributing under certain licenses” (Guo et al., 2016b). 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 (Murgia, 2019).

Positive examples. FACES, an experimental psychology dataset on emotion-related stimuli, represents a positive exception in the face analysis domain. Due its small cardinality, it was possible to obtain informed consent from every participant. 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. (1992) 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,[6] 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 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 (Takac and Zabovsky, 2012). While 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 favouring 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.

This stands in contrast with Framingham and other datasets derived from medical studies, where consent may be provided or refused with fine granularity (Levy et al., 2010). In this regard, let us consider a consent form from a recent Framingham exam (Framingham Heart Study, 2021). The form comes with five different consent boxes which cover participation in examination, use of resulting data, participation in genetic studies, sharing of data with external entities, and notification of findings to 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 are the risks 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."

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

- “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 deliver a Spanish version, showing attention to linguistic minorities. Overall, this approach seems geared towards trust and truly informed consent.

In some cases, consent is made unapplicable by necessity. Allegheny Child Welfare, for instance, 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 situations 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 requesting 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 a chance to revoke it and be removed from a dataset. This is an active area of research, studying specific tools for consent management (Albanese et al., 2020) and approaches for retroactive removal of an instance from a model's training set (Ginart et al., 2019). 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. The dataset was removed by Microsoft after widespread criticism and claims of privacy infringement. Despite this fact, it remains available via academic torrents (Peng et al., 2021). 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 covered in this survey. This fact demonstrates that harms related to data artefacts are not simply remedied via retirement or redaction. Ethical considerations about consent and potential harms to people must be more than an afterthought and need to enter the discussion during design.

6.3 Inclusivity

Motivation. Issues of representation, inclusion and diversity are central to the fair ML community. Due to historical biases stemming from structural inequalities, some populations and their perspectives are underrepresented in certain domains and in related data artefacts (Jo and Gebru, 2020). For example, the person subtree of ImageNet contains images that skew toward male, young and light skin individuals (Yang et al., 2020b). Female entities were found to be underrepresented in popular datasets for coreference resolution

(Zhao et al., 2018). Even datasets that match natural group proportions may support the development of biased tools with low accuracy for minorities.

Recent works have demonstrated the disparate performance of tools on sensitive subpopulations in domains such as health care (Obermeyer and Mullainathan, 2019), speech recognition (Tatman, 2017), and computer vision (Buolamwini and Gebru, 2018). 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 (Jo and Gebru, 2020). 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, the curators of ImageNet proposed a demographic balancing solution based on a web interface that removes the images of overrepresented categories (Yang et al., 2020b). A natural alternative is the collection of novel instances, a solution adopted for Framingham. This dataset stems 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 a greater racial and ethnic diversity in the town (Tsao and Vasan, 2015).

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 (Radford et al., 2019). The authors crawled every document reachable from outbound Reddit links that collected at least 3 *karma*. While this was considered a useful heuristic to achieve size and quality, it ended up skewing this resource towards content appreciated by Reddit users, who are predominantly male, young, and enjoy good internet access. This should act as reminder that size does not guarantee diversity (Bender et al., 2021), 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* approach 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 collection of data is ideal to ensure that various cultures are represented in the manner in which they would like to be seen (Jo and Gebru, 2020). Unfortunately, we found no instance of datasets where 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 removing instances whose gender is neither female nor male (Flanigan et al., 2020).

Inclusivity does not guarantee benefits. To avoid downstream harms, inclusion by itself 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 (Prabhu and Birhane, 2020). 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 (Hendricks et al., 2018). 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 in general, 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" by simply collecting more data, or granting equal performance across different groups identified by a given sensitive attribute.

6.4 Sensitive Attribute Labelling

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. On the contrary, research in human-computer interaction, computer-supported cooperative work, and critical data studies argues that this belief is superficial, limited and potentially harmful (Muller et al., 2019; Crawford and Paglen, 2021).

Data is, quite simply, a human-influenced entity (Miceli et al., 2021), 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 cognizant of the context surrounding data than theoretical researchers, report significant awareness of how curation and annotation choices influence their data and its relation with the underlying phenomena (Muller et al., 2019). In an interview, a senior text classification researcher responsible for ground truth annotation shows consciousness of their own influence on datasets by stating “I am the ground truth.” (Muller et al., 2019).

Sensitive attributes, such as race and gender, are no exception in this regard. Inconsistencies in racial annotation are rather common within the same system (Lum et al., 2020) and, even more so, across different systems (Scheuerman et al., 2020; Khan and Fu, 2021). 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 5.

Table 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 may report their demographics during registration (Bing US Queries, MovieLens, Libimseti), or that were gathered 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 labelled 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" (Liu et al., 2015). In a similar fashion, Diversity in Faces consists of images labeled with gender and age by workers hired through the Figure Eight crowd-sourcing 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 of fair compensation of labour, 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,[7] as is the case with RtGender. In situations where text written by individuals is available, rule-based approaches exploiting 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 (Locatello et al., 2019). Alternatively, datasets can be augmented with simulated demographics, as done by Madnani et al. (2017) who randomly assigned a native language to test-takers in ASAP, or through the technique of Burke et al. (2018a), which they demonstrate on MovieLens.

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

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 labelled by a surgical dermatologist with the Fitzpatrick skin type of the subjects (Fitzpatrick, 1988). 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.[8]

Similarly to Scheuerman et al. (2020), 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" (Zhang et al., 2014).

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)" (Guo et al., 2016b). For these reasons, 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 annotated externally 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. Purposes and aims should be documented clearly, while also reflecting on other uses and potential for misuse of a dataset (Gebru et al., 2018). Dataset curators may use documentation to discuss these aspects and specify limitations for

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

the intended use of a resource (Peng et al., 2021). In the next section we focus on documentation and why it represents a key component of data curation.

6.5 Transparency

Motivation. Transparent and accurate documentation is a fundamental part of data quality. Its absence may lead to serious issues, including lack of reproducibility, concerns of scientific validity, ethical problems, and harms (Barocas et al., 2019). Clear documentation can shine a light on inevitable choices made by dataset creators and on the context surrounding the data. In the absence of this information, the curation mechanism mediating reality and data is hidden; the data becomes one with its context, to the point that interpretation of numerical results can be misleading and overarching (Bao et al., 2021).

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 measures of algorithmic fairness such as sufficiency and separation (Barocas et al., 2019) 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) (Friedler et al., 2021). 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 about who collected and annotated the data, how, and for which purpose (Gebru et al., 2018; Denton et al., 2020). This provides dataset users with information they can leverage to select appropriate datasets for their tasks and avoid unintentional misuse (Gebru et al., 2018). Other actors, such as reviewers, may also access the official documentation of a dataset to ensure that it is employed in compliance with its stated purpose, guidelines, and terms of use (Peng et al., 2021). Overall data documentation plays a fundamental role in increasing transparency and accountability (Hutchinson et al., 2021), favouring responsible and reflexive data curation (Miceli et al., 2021; Jo and Gebru, 2020), and a correct utilization of these resources (Paullada et al., 2020).

Positive examples. In this survey, we find examples of excellent documentation 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 to collect 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 which 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 no 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 it consists of 10,000 outdoor face images from the web that were labelled manually with gender. Most face datasets suffer from opaque documentation, especially concerning 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 about employee attrition, which the hosting website describes as containing "fictional data", without any additional information. Nonetheless, 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 may also be produced retrospectively (Bandy and Vincent, 2021; Garbin et al., 2021). 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 mistakes (Grömping, 2019). As mentioned in Section 4.3, 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 by Grömping (2019) shows that this is not the case, which has 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 artefact of the wrong coding with which the dataset has been, and still is, officially distributed in the UCI Machine Learning Repository (1994). Until the report and the new redacted dataset (UCI Machine Learning Repository, 2019) 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 (He et al., 2020b; Yang et al., 2020a; Baharlouei et al., 2020; Lohaus et al., 2020; Martinez et al., 2020; Wang et al., 2021). 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 the 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. Indeed, 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.

7 Broader Relevance to the Community

Along with the analyses presented in this work, through the lens of tasks supported, domains spanned, and roles played by algorithmic fairness datasets, we are releasing the underlying data briefs, as a further contribution for the research community. Data briefs are a short documentation format providing essential information on datasets used in fairness research. Data briefs are composed of ten fields, detailed in appendix A, derived from shared vocabularies such as Data Catalog Vocabulary (DCAT); to be compliant with the FAIR data principles (Wilkinson et al., 2016), we also defined a schema called `fdo` to model the relationships between the terms and to make the links to external vocabularies explicit. We leverage `fdo` to format the data briefs as a Resource Description Framework (RDF) (Miller, 1998), and to make them available as linked open data, thus supporting data reuse, interoperability, and interpretability.[9] Our final aim is to release, update, and maintain a web app, 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, such as:

- Informing the choice of datasets for experimental evaluations of fair ML methods, including domain-oriented and task-oriented search.
- Directing studies of data bias, and other quantitative and qualitative analyses, including retrospective documentation efforts, towards popular (or otherwise important) resources.
- Identifying areas and sub-problems that are understudied in the algorithmic fairness literature.
- Supporting multi-dataset studies, focused on resources united by a common characteristic, such as encoding a given sensitive attribute (Scheuerman et al., 2020), concerning computer vision (Fabbrizzi et al., 2021), or being popular in the fairness literature (Le Quy et al., 2022).

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

[10] This resource will be released at `https://fairnessdatasets.dei.unipd.it/`

8 Conclusions and Recommendations

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, improve the utilization of existing datasets, and the curation of novel ones.

With respect to existing resources, we have shown that the most popular datasets in the fairness literature (Adult, COMPAS, and German Credit) have limited merits beyond originating from human processes and encoding protected attributes. On the other hand, several negative aspects call into question their current status of general-purpose fairness benchmarks, including contrived prediction tasks, noisy data, severe coding mistakes, limitations in encoding sensitive attributes, and age.

We have documented over two hundred datasets to provide viable alternatives, annotating their domain, the tasks they support, and discussing the roles they play 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 classification*, 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 algorithmic fairness research. Tasks and domain annotations are made available in our data briefs to facilitate the work of researchers and practitioners interested in the study of algorithmic fairness applied to specific domains or tasks. By assembling sparse information on hundreds of datasets into a single document, we aim to provide a useful reference to support both domain-oriented and task-oriented dataset search.

At the same time, we have analyzed issues connected to re-identification, consent, inclusivity, labeling, and transparency running across these datasets. By describing a range of approaches and attentiveness to these topics, we aim to make them more visible and concrete. On 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 endeavours of data curation. Our recommendations complement (and do not replace) a growing body of work studying key aspects in the life cycle of datasets (Gebru et al., 2018; Jo and Gebru, 2020; Prabhu and Birhane, 2020; Crawford and Paglen, 2021; Peng et al., 2021; Hutchinson et al., 2021).

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.

Inclusivity is also important for social relevance, as it allows for a wider representation, and supports analyses that take into account important groups. However, inclusivity is insufficient in itself. 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.

Understanding and budgeting for all these aspects during early design phases, rather than after collection or release, can be invaluable for data subjects, data users, and society. While possible remedies exist, data is an extremely fluid asset allowing for easy reproduction and derivatives of all sorts; remedies applied to a dataset do not necessarily benefit its derivates. 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; as a result, we have found and described a range of weaknesses and best practices that can be adopted to reduce opacity and mitigate concerns of privacy and inclusion. 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.

Acknowledgements 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.

References

Abbasi M, Bhaskara A, Venkatasubramanian S (2021) Fair clustering via equitable group representations. In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, Association for Computing Machinery, New York, NY, USA, FAccT '21, p 504–514, DOI 10.1145/3442188.3445913, URL https://doi.org/10.1145/3442188.3445913

Adragna R, Creager E, Madras D, 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)”, 2011.06485
Agarwal A, Beygelzimer A, Dudik M, Langford J, Wallach H (2018) A reductions approach to fair classification. In: Dy J, Krause A (eds) Proceedings of the 35th International Conference on Machine Learning, PMLR, Stockholmsmässan, Stockholm Sweden, Proceedings of Machine Learning Research, vol 80, pp 60–69, URL http://proceedings.mlr.press/v80/agarwal18a.html
Agarwal A, Dudik M, Wu ZS (2019) Fair regression: Quantitative definitions and reduction-based algorithms. In: Chaudhuri K, Salakhutdinov R (eds) Proceedings of the 36th International Conference on Machine Learning, PMLR, Long Beach, California, USA, Proceedings of Machine Learning Research, vol 97, pp 120–129, URL http://proceedings.mlr.press/v97/agarwal19d.html
Agrawal M, Zitnik M, Leskovec J, et al. (2018) Large-scale analysis of disease pathways in the human interactome. In: PSB, World Scientific, pp 111–122
Ahmadian S, Epasto A, Knittel M, Kumar R, Mahdian M, Moseley B, Pham P, Vassilvitskii S, Wang Y (2020) Fair hierarchical clustering. In: Larochelle H, Ranzato M, Hadsell R, Balcan M, Lin H (eds) Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual, URL https://proceedings.neurips.cc/paper/2020/hash/f10f2da9a238b746d2bac55759915f0d-Abstract.html
Aka O, Burke K, Bauerle A, Greer C, Mitchell M (2021) Measuring Model Biases in the Absence of Ground Truth, Association for Computing Machinery, New York, NY, USA, p 327–335. URL https://doi.org/10.1145/3461702.3462557
Albanese G, Calbimonte JP, Schumacher M, Calvaresi D (2020) Dynamic consent management for clinical trials via private blockchain technology. Journal of ambient intelligence and humanized computing pp 1–18
Ali J, Babaei M, Chakraborty A, Mirzasoleiman B, Gummadi KP, Singla A (2019a) On the fairness of time-critical influence maximization in social networks. NeurIPS 2019 workshop: “Human-Centric Machine Learning”, 1905.06618
Ali J, Zafar MB, Singla A, Gummadi KP (2019b) Loss-aversively fair classification. In: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, Association for Computing Machinery, New York, NY, USA, AIES ’19, p 211–218, DOI 10.1145/3306618.3314266, URL https://doi.org/10.1145/3306618.3314266
Ali J, Lahoti P, Gummadi KP (2021) Accounting for Model Uncertainty in Algorithmic Discrimination, Association for Computing Machinery, New York, NY, USA, p 336–345. URL https://doi.org/10.1145/3461702.3462630
Amini A, Soleimany AP, Schwarting W, Bhatia SN, 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, Association for Computing Machinery, New York, NY, USA, AIES ’19, p

289–295, DOI 10.1145/3306618.3314243, URL https://doi.org/10.1145/3306618.3314243
Anderson E (1936) The species problem in iris. Annals of the Missouri Botanical Garden 23(3):457–509
Andrus M, Spitzer E, Brown J, 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, Association for Computing Machinery, New York, NY, USA, FAccT '21, p 249–260, DOI 10.1145/3442188.3445888, URL https://doi.org/10.1145/3442188.3445888
Andrzejak RG, Lehnertz K, Mormann F, Rieke C, David P, Elger CE (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
Angwin J, Larson J, Mattu S, Kirchner L (2016) Machine bias. URL https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing
Arjovsky M, Bottou L, Gulrajani I, Lopez-Paz D (2020) Invariant risk minimization. 1907.02893
Arnold M, Bellamy RK, Hind M, Houde S, Mehta S, Mojsilović A, Nair R, Ramamurthy KN, 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
Atwood J, Srinivasan H, Halpern Y, Sculley D (2019) Fair treatment allocations in social networks. NeurIPS 2019 workshop: "Fair ML for Health", 1911.05489
Awasthi P, Beutel A, Kleindessner M, Morgenstern J, Wang X (2021) Evaluating fairness of machine learning models under uncertain and incomplete information. In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, Association for Computing Machinery, New York, NY, USA, FAccT '21, p 206–214, DOI 10.1145/3442188.3445884, URL https://doi.org/10.1145/3442188.3445884
Babaeianjelodar M, Lorenz S, Gordon J, Matthews J, Freitag E (2020) Quantifying gender bias in different corpora. In: Companion Proceedings of the Web Conference 2020, Association for Computing Machinery, New York, NY, USA, WWW '20, p 752–759, DOI 10.1145/3366424.3383559, URL https://doi.org/10.1145/3366424.3383559
Babaioff M, Nisan N, Talgam-Cohen I (2019) Fair allocation through competitive equilibrium from generic incomes. In: Proceedings of the Conference on Fairness, Accountability, and Transparency, Association for Computing Machinery, New York, NY, USA, FAT* '19, p 180, DOI 10.1145/3287560.3287582, URL https://doi.org/10.1145/3287560.3287582
Backurs A, Indyk P, Onak K, Schieber B, Vakilian A, Wagner T (2019) Scalable fair clustering. In: Chaudhuri K, Salakhutdinov R (eds) Proceedings of the 36th International Conference on Machine Learning, PMLR, Long Beach, California, USA, Proceedings of Machine Learning Re-

search, vol 97, pp 405–413, URL http://proceedings.mlr.press/v97/backurs19a.html
Bagdasaryan E, Poursaeed O, Shmatikov V (2019) Differential privacy has disparate impact on model accuracy. In: Wallach H, Larochelle H, Beygelzimer A, d'Alché-Buc F, Fox E, Garnett R (eds) Advances in Neural Information Processing Systems, Curran Associates, Inc., vol 32, URL https://proceedings.neurips.cc/paper/2019/file/fc0de4e0396fff257ea362983c2dda5a-Paper.pdf
Baharlouei S, Nouiehed M, Beirami A, Razaviyayn M (2020) Rényi fair inference. In: International Conference on Learning Representations, URL https://openreview.net/forum?id=HkgsUJrtDB
Bakker MA, Tu DP, Valdés HR, Gummadi KP, Varshney KR, Weller A, Pentland A (2019) Dadi: Dynamic discovery of fair information with adversarial reinforcement learning. NeurIPS 2019 workshop: "Human-Centric Machine Learning", 1910.13983
Bakker MA, Tu DP, Gummadi KP, Pentland AS, Varshney KR, Weller A (2021) Beyond Reasonable Doubt: Improving Fairness in Budget-Constrained Decision Making Using Confidence Thresholds, Association for Computing Machinery, New York, NY, USA, p 346–356. URL https://doi.org/10.1145/3461702.3462575
Ball-Burack A, Lee MSA, Cobbe J, 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, Association for Computing Machinery, New York, NY, USA, FAccT '21, p 116–128, DOI 10.1145/3442188.3445875, URL https://doi.org/10.1145/3442188.3445875
Bandy J, Vincent N (2021) Addressing" documentation debt" in machine learning research: A retrospective datasheet for bookcorpus. arXiv preprint arXiv:210505241
Bao M, Zhou A, Zottola S, Brubach B, Desmarais S, Horowitz A, Lum K, Venkatasubramanian S (2021) It's compaslicated: The messy relationship between rai datasets and algorithmic fairness benchmarks. arXiv preprint arXiv:210605498
Barabas C, Dinakar K, Doyle C (2019) The problems with risk assessment tools. URL https://www.nytimes.com/2019/07/17/opinion/pretrial-ai.html
Barbaro M (2007) In apparel, all tariffs aren't created equal. URL https://www.nytimes.com/2007/04/28/business/28gender.html
Barenstein M (2019) Propublica's compas data revisited. arXiv preprint arXiv:190604711
Barman-Adhikari A, Begun S, Rice E, Yoshioka-Maxwell A, Perez-Portillo A (2016) Sociometric network structure and its association with methamphetamine use norms among homeless youth. Social science research 58:292–308
Barocas S, Hardt M, Narayanan A (2019) Fairness and Machine Learning. fairmlbook.org, http://www.fairmlbook.org

Baudry JP, Cardoso M, Celeux G, Amorim MJ, Ferreira AS (2015) Enhancing the selection of a model-based clustering with external categorical variables. Advances in Data Analysis and Classification 9(2):177–196

Behaghel L, Crépon B, 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, DOI 10.1257/app.6.4.142, URL https://www.aeaweb.org/articles?id=10.1257/app.6.4.142

Belitz C, Jiang L, Bosch N (2021) Automating Procedurally Fair Feature Selection in Machine Learning, Association for Computing Machinery, New York, NY, USA, p 379–389. URL https://doi.org/10.1145/3461702.3462585

Bender EM, 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, DOI 10.1162/tacl_a_00041, URL https://www.aclweb.org/anthology/Q18-1041

Bender EM, Gebru T, McMillan-Major A, Shmitchell S (2021) On the dangers of stochastic parrots: Can language models be too big? Association for Computing Machinery, New York, NY, USA, FAccT '21, p 610–623, DOI 10.1145/3442188.3445922, URL https://doi.org/10.1145/3442188.3445922

Benenson R, Popov S, Ferrari V (2019) Large-scale interactive object segmentation with human annotators. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 11700–11709

Bera S, Chakrabarty D, Flores N, Negahbani M (2019) Fair algorithms for clustering. In: Wallach H, Larochelle H, Beygelzimer A, d'Alché-Buc F, Fox E, Garnett R (eds) Advances in Neural Information Processing Systems, Curran Associates, Inc., vol 32, pp 4954–4965, URL https://proceedings.neurips.cc/paper/2019/file/fc192b0c0d270dbf41870a63a8c76c2f-Paper.pdf

Beretta E, Vetrò A, Lepri B, Martin JCD (2021) Detecting discriminatory risk through data annotation based on bayesian inferences. In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, Association for Computing Machinery, New York, NY, USA, FAccT '21, p 794–804, DOI 10.1145/3442188.3445940, URL https://doi.org/10.1145/3442188.3445940

Berk R, Heidari H, Jabbari S, Joseph M, Kearns M, Morgenstern J, Neel S, Roth A (2017) A convex framework for fair regression. KDD 2017 workshop: "Fairness, Accountability, and Transparency in Machine Learning (FAT/ML)", 1706.02409

Bertin-Mahieux T, Ellis DPW, Whitman B, Lamere P (2011) The million song dataset. In: Proceedings of the 12th International Society for Music Information Retrieval Conference, ISMIR, Miami, United States, pp 591–596, DOI 10.5281/zenodo.1415820, URL https://doi.org/10.5281/zenodo.1415820

Bertrand M, Mullainathan S (2004) Are emily and greg more employable than lakisha and jamal? a field experiment on labor market discrimination. Amer-

ican economic review 94(4):991–1013
Beutel A, Chen J, Zhao Z, Chi EH (2017) Data decisions and theoretical implications when adversarially learning fair representations. KDD 2017 workshop: "Fairness, Accountability, and Transparency in Machine Learning (FAT/ML)", 1707.00075
Biega AJ, Diaz F, Ekstrand MD, Kohlmeier S (2019) Overview of the trec 2019 fair ranking track. In: The Twenty-Eighth Text REtrieval Conference (TREC 2019) Proceedings
Biswas A, Mukherjee S (2021) Ensuring Fairness under Prior Probability Shifts, Association for Computing Machinery, New York, NY, USA, p 414–424. URL https://doi.org/10.1145/3461702.3462596
Black E, Fredrikson M (2021) Leave-one-out unfairness. In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, Association for Computing Machinery, New York, NY, USA, FAccT '21, p 285–295, DOI 10.1145/3442188.3445894, URL https://doi.org/10.1145/3442188.3445894
Black E, Yeom S, Fredrikson M (2020) Fliptest: Fairness testing via optimal transport. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, Association for Computing Machinery, New York, NY, USA, FAT* '20, p 111–121, DOI 10.1145/3351095.3372845, URL https://doi.org/10.1145/3351095.3372845
Blodget SL, 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)", 1707.00061
Blodgett SL, Green L, 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, Association for Computational Linguistics, Austin, Texas, pp 1119–1130, DOI 10.18653/v1/D16-1120, URL https://www.aclweb.org/anthology/D16-1120
Bolukbasi T, Chang KW, Zou JY, Saligrama V, Kalai AT (2016) Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In: Lee D, Sugiyama M, Luxburg U, Guyon I, Garnett R (eds) Advances in Neural Information Processing Systems, Curran Associates, Inc., vol 29, pp 4349–4357, URL https://proceedings.neurips.cc/paper/2016/file/a486cd07e4ac3d270571622f4f316ec5-Paper.pdf
Bordes A, Usunier N, Garcia-Duran A, Weston J, Yakhnenko O (2013) Translating embeddings for modeling multi-relational data. In: Burges CJC, Bottou L, Welling M, Ghahramani Z, Weinberger KQ (eds) Advances in Neural Information Processing Systems, Curran Associates, Inc., vol 26, URL https://proceedings.neurips.cc/paper/2013/file/1cecc7a77928ca8133fa24680a88d2f9-Paper.pdf
Bordia S, Bowman SR (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 Linguis-

tics: Student Research Workshop, Association for Computational Linguistics, Minneapolis, Minnesota, pp 7–15, DOI 10.18653/v1/N19-3002, URL https://aclanthology.org/N19-3002

Borkan D, Dixon L, Sorensen J, Thain N, 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, Association for Computing Machinery, New York, NY, USA, WWW '19, p 491–500, DOI 10.1145/3308560.3317593, URL https://doi.org/10.1145/3308560.3317593

Bose A, Hamilton W (2019) Compositional fairness constraints for graph embeddings. In: Chaudhuri K, Salakhutdinov R (eds) Proceedings of the 36th International Conference on Machine Learning, PMLR, Long Beach, California, USA, Proceedings of Machine Learning Research, vol 97, pp 715–724, URL http://proceedings.mlr.press/v97/bose19a.html

Bower A, Niss L, Sun Y, 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)", 1807.00461

Bower A, Eftekhari H, Yurochkin M, Sun Y (2021) Individually fair rankings. In: International Conference on Learning Representations, URL https://openreview.net/forum?id=71zCSP_HuBN

Brennan T, Dieterich W, Ehret B (2009) Evaluating the predictive validity of the compas risk and needs assessment system. Criminal Justice and Behavior 36(1):21–40, DOI 10.1177/0093854808326545, URL https://doi.org/10.1177/0093854808326545, https://doi.org/10.1177/0093854808326545

Brockman G, Cheung V, Pettersson L, Schneider J, Schulman J, Tang J, Zaremba W (2016) Openai gym. arXiv preprint arXiv:160601540

Brooks-Gunn J, Liaw Fr, Klebanov PK (1992) Effects of early intervention on cognitive function of low birth weight preterm infants. The Journal of pediatrics 120(3):350–359

Brozovsky L, Petricek V (2007) Recommender system for online dating service. ArXiv abs/cs/0703042

Brožovský L (2006) Recommender system for a dating service. Master's thesis, Charles University in Prague, Prague, Czech Republic, URL http://colfi.wz.cz/colfi.pdf

Brubach B, Chakrabarti D, Dickerson J, Khuller S, Srinivasan A, Tsepenekas L (2020) A pairwise fair and community-preserving approach to k-center clustering. In: III HD, Singh A (eds) Proceedings of the 37th International Conference on Machine Learning, PMLR, Virtual, Proceedings of Machine Learning Research, vol 119, pp 1178–1189, URL http://proceedings.mlr.press/v119/brubach20a.html

Brunet ME, Alkalay-Houlihan C, Anderson A, Zemel R (2019) Understanding the origins of bias in word embeddings. In: Chaudhuri K, Salakhutdinov R (eds) Proceedings of the 36th International Conference on Machine Learning, PMLR, Long Beach, California, USA, Proceedings of Machine Learning Research, vol 97, pp 803–811, URL http://proceedings.mlr.press/v97/

brunet19a.html
Buolamwini J, Gebru T (2018) Gender shades: Intersectional accuracy disparities in commercial gender classification. In: Friedler SA, Wilson C (eds) Proceedings of the 1st Conference on Fairness, Accountability and Transparency, PMLR, New York, NY, USA, Proceedings of Machine Learning Research, vol 81, pp 77–91, URL http://proceedings.mlr.press/v81/buolamwini18a.html
Burke R, Kontny J, Sonboli N (2018a) Synthetic attribute data for evaluating consumer-side fairness. RecSys 2018 workshop: "Workshop on Responsible Recommendation (FAT/Rec)", 1809.04199
Burke R, Sonboli N, Ordonez-Gauger A (2018b) Balanced neighborhoods for multi-sided fairness in recommendation. In: Friedler SA, Wilson C (eds) Proceedings of the 1st Conference on Fairness, Accountability and Transparency, PMLR, New York, NY, USA, Proceedings of Machine Learning Research, vol 81, pp 202–214, URL http://proceedings.mlr.press/v81/burke18a.html
Buyl M, De Bie T (2020) DeBayes: a Bayesian method for debiasing network embeddings. In: III HD, Singh A (eds) Proceedings of the 37th International Conference on Machine Learning, PMLR, Virtual, Proceedings of Machine Learning Research, vol 119, pp 1220–1229, URL http://proceedings.mlr.press/v119/buyl20a.html
Cai W, Gaebler J, Garg N, Goel S (2020) Fair allocation through selective information acquisition. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, Association for Computing Machinery, New York, NY, USA, AIES '20, p 22–28, DOI 10.1145/3375627.3375823, URL https://doi.org/10.1145/3375627.3375823
Caldas S, Duddu SMK, Wu P, Li T, Konečný J, McMahan HB, Smith V, Talwalkar A (2018) Leaf: A benchmark for federated settings. arXiv preprint arXiv:181201097
Calders T, Verwer S (2010) Three naive bayes approaches for discrimination-free classification. Data Min Knowl Discov 21(2):277–292, DOI 10.1007/s10618-010-0190-x, URL https://doi.org/10.1007/s10618-010-0190-x
Calders T, Kamiran F, Pechenizkiy M (2009) Building classifiers with independency constraints. In: 2009 IEEE International Conference on Data Mining Workshops, pp 13–18, DOI 10.1109/ICDMW.2009.83
Caliskan A, Bryson J, Narayanan A (2017) Semantics derived automatically from language corpora contain human-like biases. Science 356(6334):183–186, DOI 10.1126/science.aal4230
Calmon F, Wei D, Vinzamuri B, Natesan Ramamurthy K, Varshney KR (2017) Optimized pre-processing for discrimination prevention. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, Garnett R (eds) Advances in Neural Information Processing Systems, Curran Associates, Inc., vol 30, pp 3992–4001, URL https://proceedings.neurips.cc/paper/2017/file/9a49a25d845a483fae4be7e341368e36-Paper.pdf
Canetti R, Cohen A, Dikkala N, Ramnarayan G, Scheffler S, Smith A (2019) From soft classifiers to hard decisions: How fair can we be? In: Proceed-

ings of the Conference on Fairness, Accountability, and Transparency, Association for Computing Machinery, New York, NY, USA, FAT* '19, p 309–318, DOI 10.1145/3287560.3287561, URL https://doi.org/10.1145/3287560.3287561

Caragiannis I, Kurokawa D, Moulin H, Procaccia AD, Shah N, Wang J (2016) The unreasonable fairness of maximum nash welfare. In: Proceedings of the 2016 ACM Conference on Economics and Computation, Association for Computing Machinery, New York, NY, USA, EC '16, p 305–322, DOI 10.1145/2940716.2940726, URL https://doi.org/10.1145/2940716.2940726

Cardoso RL, Meira Jr W, Almeida V, Zaki MJ (2019) A framework for benchmarking discrimination-aware models in machine learning. In: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, Association for Computing Machinery, New York, NY, USA, AIES '19, p 437–444, DOI 10.1145/3306618.3314262, URL https://doi.org/10.1145/3306618.3314262

Carvalho M, Lodi A (2019) Game theoretical analysis of kidney exchange programs. arXiv preprint arXiv:191109207

Caton S, Haas C (2020) Fairness in machine learning: A survey. arXiv preprint arXiv:201004053

Celis E, Keswani V, Straszak D, Deshpande A, Kathuria T, Vishnoi N (2018) Fair and diverse DPP-based data summarization. In: Dy J, Krause A (eds) Proceedings of the 35th International Conference on Machine Learning, PMLR, Stockholmsmässan, Stockholm Sweden, Proceedings of Machine Learning Research, vol 80, pp 716–725, URL http://proceedings.mlr.press/v80/celis18a.html

Celis E, Mehrotra A, Vishnoi N (2019a) Toward controlling discrimination in online ad auctions. In: Chaudhuri K, Salakhutdinov R (eds) Proceedings of the 36th International Conference on Machine Learning, PMLR, Long Beach, California, USA, Proceedings of Machine Learning Research, vol 97, pp 4456–4465, URL http://proceedings.mlr.press/v97/mehrotra19a.html

Celis LE, Keswani V (2020) Implicit diversity in image summarization. Proceedings of the ACM on Human-Computer Interaction 4(CSCW2):1–28, DOI 10.1145/3415210, URL http://dx.doi.org/10.1145/3415210

Celis LE, Deshpande A, Kathuria T, Vishnoi NK (2016) How to be fair and diverse? DTL 2016 workshop: "Fairness, Accountability, and Transparency in Machine Learning (FAT/ML)", 1610.07183

Celis LE, Huang L, Keswani V, Vishnoi NK (2019b) Classification with fairness constraints: A meta-algorithm with provable guarantees. In: Proceedings of the Conference on Fairness, Accountability, and Transparency, Association for Computing Machinery, New York, NY, USA, FAT* '19, p 319–328, DOI 10.1145/3287560.3287586, URL https://doi.org/10.1145/3287560.3287586

Celis LE, Keswani V, Vishnoi N (2020a) Data preprocessing to mitigate bias: A maximum entropy based approach. In: III HD, Singh A (eds) Proceedings of the 37th International Conference on Machine Learning, PMLR, Virtual, Proceedings of Machine Learning Research, vol 119, pp 1349–1359, URL http://proceedings.mlr.press/v119/celis20a.html

Celis LE, Mehrotra A, Vishnoi NK (2020b) Interventions for ranking in the presence of implicit bias. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, Association for Computing Machinery, New York, NY, USA, FAT* '20, p 369–380, DOI 10.1145/3351095.3372858, URL https://doi.org/10.1145/3351095.3372858

Celma O (2010) Music Recommendation and Discovery in the Long Tail. Springer

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, Association for Computing Machinery, New York, NY, USA, KDD '20, p 873–881, DOI 10.1145/3394486.3403130, URL https://doi.org/10.1145/3394486.3403130

Chakraborty A, Patro GK, Ganguly N, Gummadi KP, Loiseau P (2019) Equality of voice: Towards fair representation in crowdsourced top-k recommendations. In: Proceedings of the Conference on Fairness, Accountability, and Transparency, Association for Computing Machinery, New York, NY, USA, FAT* '19, p 129–138, DOI 10.1145/3287560.3287570, URL https://doi.org/10.1145/3287560.3287570

Chapelle O, Chang Y (2010) Yahoo! learning to rank challenge overview. In: Proceedings of the 2010 International Conference on Yahoo! Learning to Rank Challenge - Volume 14, JMLR.org, YLRC'10, p 1–24

Chaudhari HA, Lin S, Linda O (2020) A general framework for fairness in multistakeholder recommendations. RecSys 2020 workshop: "3rd FAccTRec Workshop on Responsible Recommendation", 2009.02423

Chelba C, Mikolov T, Schuster M, Ge Q, Brants T, Koehn P, Robinson T (2014) One billion word benchmark for measuring progress in statistical language modeling. In: INTERSPEECH-2014

Chen B, Deng W, Shen H (2018a) Virtual class enhanced discriminative embedding learning. In: Bengio S, Wallach H, Larochelle H, Grauman K, Cesa-Bianchi N, Garnett R (eds) Advances in Neural Information Processing Systems, Curran Associates, Inc., vol 31, URL https://proceedings.neurips.cc/paper/2018/file/d79aac075930c83c2f1e369a511148fe-Paper.pdf

Chen CW, Lamere P, Schedl M, Zamani H (2018b) Recsys challenge 2018: Automatic music playlist continuation. Association for Computing Machinery, New York, NY, USA, RecSys '18, p 527–528, DOI 10.1145/3240323.3240342, URL https://doi.org/10.1145/3240323.3240342

Chen I, Johansson FD, Sontag D (2018c) Why is my classifier discriminatory? In: Bengio S, Wallach H, Larochelle H, Grauman K, Cesa-Bianchi N, Garnett R (eds) Advances in Neural Information Processing Systems, Curran Associates, Inc., vol 31, URL https://proceedings.neurips.cc/paper/2018/file/1f1baa5b8edac74eb4eaa329f14a0361-Paper.pdf

Chen J, Kallus N, Mao X, Svacha G, Udell M (2019a) Fairness under unawareness: Assessing disparity when protected class is unobserved. In: Proceedings of the Conference on Fairness, Accountability, and Transparency, Association for Computing Machinery, New York, NY, USA, FAT* '19, p 339–348, DOI 10.1145/3287560.3287594, URL https://doi.org/10.1145/3287560.3287594

Chen X, Fain B, Lyu L, Munagala K (2019b) Proportionally fair clustering. In: Chaudhuri K, Salakhutdinov R (eds) Proceedings of the 36th International Conference on Machine Learning, PMLR, Long Beach, California, USA, Proceedings of Machine Learning Research, vol 97, pp 1032–1041, URL http://proceedings.mlr.press/v97/chen19d.html

Chen Y, Mahoney C, Grasso I, Wali E, Matthews A, Middleton T, Njie M, Matthews J (2021) Gender Bias and Under-Representation in Natural Language Processing Across Human Languages, Association for Computing Machinery, New York, NY, USA, p 24–34. URL https://doi.org/10.1145/3461702.3462530

Cheng P, Hao W, Yuan S, Si S, Carin L (2021a) Fairfil: Contrastive neural debiasing method for pretrained text encoders. In: International Conference on Learning Representations, URL https://openreview.net/forum?id=N6JECD-PI5w

Cheng V, Suriyakumar VM, Dullerud N, Joshi S, 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, Association for Computing Machinery, New York, NY, USA, FAccT '21, p 149–160, DOI 10.1145/3442188.3445879, URL https://doi.org/10.1145/3442188.3445879

Chierichetti F, Kumar R, Lattanzi S, Vassilvitskii S (2017) Fair clustering through fairlets. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, Garnett R (eds) Advances in Neural Information Processing Systems, Curran Associates, Inc., vol 30, pp 5029–5037, URL https://proceedings.neurips.cc/paper/2017/file/978fce5bcc4eccc88ad48ce3914124a2-Paper.pdf

Chiplunkar A, Kale S, Ramamoorthy SN (2020) How to solve fair k-center in massive data models. In: III HD, Singh A (eds) Proceedings of the 37th International Conference on Machine Learning, PMLR, Virtual, Proceedings of Machine Learning Research, vol 119, pp 1877–1886, URL http://proceedings.mlr.press/v119/chiplunkar20a.html

Cho J, Hwang G, Suh C (2020) A fair classifier using kernel density estimation. In: Larochelle H, Ranzato M, Hadsell R, Balcan MF, Lin H (eds) Advances in Neural Information Processing Systems, Curran Associates, Inc., vol 33, pp 15088–15099, URL https://proceedings.neurips.cc/paper/2020/file/ac3870fcad1cfc367825cda0101eee62-Paper.pdf

Cho WI, Kim J, Yang J, Kim NS (2021) Towards cross-lingual generalization of translation gender bias. In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, Association for Computing Machinery, New York, NY, USA, FAccT '21, p 449–457, DOI 10.1145/3442188.3445907, URL https://doi.org/10.1145/3442188.3445907

Choi K, Grover A, Singh T, Shu R, Ermon S (2020a) Fair generative modeling via weak supervision. In: III HD, Singh A (eds) Proceedings of the 37th International Conference on Machine Learning, PMLR, Virtual, Proceedings of Machine Learning Research, vol 119, pp 1887–1898, URL http://proceedings.mlr.press/v119/choi20a.html

Choi Y, Dang M, den Broeck GV (2020b) Group fairness by probabilistic modeling with latent fair decisions. NeurIPS 2020 workshop: "Algorithmic Fairness through the Lens of Causality and Interpretability (AFCI)", 2009.09031

Chouldechova A (2017) Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. Big data 5 2:153–163

Chouldechova A, G'Sell M (2017) Fairer and more accurate, but for whom? KDD 2017 workshop: "Fairness, Accountability, and Transparency in Machine Learning (FAT/ML)", 1707.00046

Chouldechova A, Roth A (2020) A snapshot of the frontiers of fairness in machine learning. Commun ACM 63(5):82–89, DOI 10.1145/3376898, URL https://doi.org/10.1145/3376898

Chuang CY, Mroueh Y (2021) Fair mixup: Fairness via interpolation. In: International Conference on Learning Representations, URL https://openreview.net/forum?id=DNl5s5BXeBn

Chzhen E, Denis C, Hebiri M, Oneto L, 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, Garnett R (eds) Advances in Neural Information Processing Systems, Curran Associates, Inc., vol 32, pp 12760–12770, URL https://proceedings.neurips.cc/paper/2019/file/ba51e6158bcaf80fd0d834950251e693-Paper.pdf

Chzhen E, Denis C, Hebiri M, Oneto L, Pontil M (2020a) Fair regression via plug-in estimator and recalibration with statistical guarantees. In: Larochelle H, Ranzato M, Hadsell R, Balcan M, Lin H (eds) Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual, URL https://proceedings.neurips.cc/paper/2020/hash/ddd808772c035aed16d42ad3559be5f-Abstract.html

Chzhen E, Denis C, Hebiri M, Oneto L, Pontil M (2020b) Fair regression with wasserstein barycenters. In: Larochelle H, Ranzato M, Hadsell R, Balcan MF, Lin H (eds) Advances in Neural Information Processing Systems, Curran Associates, Inc., vol 33, pp 7321–7331, URL https://proceedings.neurips.cc/paper/2020/file/51cdbd2611e844ece5d80878eb770436-Paper.pdf

Cohany SR, Polivka AE, Rothgeb JM (1994) Revisions in the current population survey effective january 1994. Emp & Earnings 41:13

Corbett-Davies S, Pierson E, Feller A, Goel S, 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, Association for Computing Machinery, New York, NY, USA, KDD '17, p 797–806, DOI 10.1145/3097983.3098095, URL https://doi.org/10.1145/3097983.3098095

Cortez P, Silva AMG (2008) Using data mining to predict secondary school student performance. In: Proceedings of 5th FUture BUsiness TEChnology Conference

Coston A, Ramamurthy KN, Wei D, Varshney KR, Speakman S, Mustahsan Z, Chakraborty S (2019) Fair transfer learning with missing protected attributes. In: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, Association for Computing Machinery, New York, NY, USA, AIES '19, p 91–98, DOI 10.1145/3306618.3314236, URL https://doi.org/10.1145/3306618.3314236

Coston A, Mishler A, Kennedy EH, Chouldechova A (2020) Counterfactual risk assessments, evaluation, and fairness. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, Association for Computing Machinery, New York, NY, USA, FAT* '20, p 582–593, DOI 10.1145/3351095.3372851, URL https://doi.org/10.1145/3351095.3372851

Coston A, Guha N, Ouyang D, Lu L, Chouldechova A, Ho DE (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, Association for Computing Machinery, New York, NY, USA, FAccT '21, p 173–184, DOI 10.1145/3442188.3445881, URL https://doi.org/10.1145/3442188.3445881

Cotter A, Gupta M, Jiang H, Srebro N, Sridharan K, Wang S, Woodworth B, You S (2018) Training fairness-constrained classifiers to generalize. ICML 2018 workshop: "Fairness, Accountability, and Transparency in Machine Learning (FAT/ML)"

Cotter A, Gupta M, Jiang H, Srebro N, Sridharan K, Wang S, Woodworth B, You S (2019) Training well-generalizing classifiers for fairness metrics and other data-dependent constraints. In: Chaudhuri K, Salakhutdinov R (eds) Proceedings of the 36th International Conference on Machine Learning, PMLR, Long Beach, California, USA, Proceedings of Machine Learning Research, vol 97, pp 1397–1405, URL http://proceedings.mlr.press/v97/cotter19b.html

Crawford K, Paglen T (2021) Excavating ai: the politics of images in machine learning training sets. URL https://excavating.ai/

Creager E, Madras D, Jacobsen JH, Weis M, Swersky K, Pitassi T, Zemel R (2019) Flexibly fair representation learning by disentanglement. In: Chaudhuri K, Salakhutdinov R (eds) Proceedings of the 36th International Conference on Machine Learning, PMLR, Long Beach, California, USA, Proceedings of Machine Learning Research, vol 97, pp 1436–1445, URL http://proceedings.mlr.press/v97/creager19a.html

Creager E, Madras D, Pitassi T, Zemel R (2020) Causal modeling for fairness in dynamical systems. In: III HD, Singh A (eds) Proceedings of the 37th International Conference on Machine Learning, PMLR, Virtual, Proceedings of Machine Learning Research, vol 119, pp 2185–2195, URL http://proceedings.mlr.press/v119/creager20a.html

Creager E, Jacobsen JH, Zemel R (2021) Exchanging lessons between algorithmic fairness and domain generalization. URL https://openreview.net/forum?id=DC1Im3MkGG, neurIPS 2020 workshop: "Algorithmic Fairness through the Lens of Causality and Interpretability (AFCI)"

D'Amour A, Srinivasan H, Atwood J, Baljekar P, Sculley D, 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, Association for Computing Machinery, New York, NY, USA, FAT* '20, p 525–534, DOI 10.1145/3351095.3372878, URL https://doi.org/10.1145/3351095.3372878

Dash A, Chakraborty A, Ghosh S, Mukherjee A, Gummadi KP (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, Association for Computing Machinery, New York, NY, USA, FAccT '21, p 873–884, DOI 10.1145/3442188.3445944, URL https://doi.org/10.1145/3442188.3445944

Datta S, Posada J, Olson G, Li W, O'Reilly C, Balraj D, Mesterhazy J, Pallas J, Desai P, Shah N (2020) A new paradigm for accelerating clinical data science at stanford medicine. arXiv preprint arXiv:200310534

David KE, Liu Q, Fong R (2020) Debiasing convolutional neural networks via meta orthogonalization. NeurIPS 2020 workshop: "Algorithmic Fairness through the Lens of Causality and Interpretability (AFCI)", 2011.07453

Davidson I, Ravi SS (2020) A framework for determining the fairness of outlier detection. In: ECAI 2020, IOS Press, pp 2465–2472

Davidson T, Warmsley D, Macy MW, 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, AAAI Press, pp 512–515, URL https://aaai.org/ocs/index.php/ICWSM/ICWSM17/paper/view/15665

De-Arteaga M, Romanov A, Wallach H, Chayes J, Borgs C, Chouldechova A, Geyik S, Kenthapadi K, Kalai AT (2019) 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, Association for Computing Machinery, New York, NY, USA, FAT* '19, p 120–128, DOI 10.1145/3287560.3287572, URL https://doi.org/10.1145/3287560.3287572

Delobelle P, Temple P, Perrouin G, Frénay B, Heymans P, Berendt B (2020) Ethical adversaries: Towards mitigating unfairness with adversarial machine learning. ECMLPKDD 2020 workshop: "BIAS 2020: Bias and Fairness in AI", 2005.06852

Deng J, Dong W, Socher R, Li L, Kai Li, Li Fei-Fei (2009) Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp 248–255, DOI 10.1109/CVPR.2009.5206848

Denton E, Hanna A, Amironesei R, Smart A, Nicole H, Scheuerman MK (2020) Bringing the people back in: Contesting benchmark machine learning datasets. arXiv preprint arXiv:200707399

Deshpande KV, Pan S, Foulds JR (2020) Mitigating demographic bias in ai-based resume filtering. In: Adjunct Publication of the 28th ACM Conference on User Modeling, Adaptation and Personalization, Association for Computing Machinery, New York, NY, USA, UMAP '20 Adjunct, p 268–275, DOI 10.1145/3386392.3399569, URL https://doi.org/10.1145/3386392.3399569

Detrano R, Janosi A, Steinbrunn W, Pfisterer M, Schmid JJ, Sandhu S, Guppy KH, Lee S, 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

Devlin J, Chang MW, Lee K, 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), Association for Computational Linguistics, Minneapolis, Minnesota, pp 4171–4186

Dhamala J, Sun T, Kumar V, Krishna S, Pruksachatkun Y, Chang KW, 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, Association for Computing Machinery, New York, NY, USA, FAccT '21, p 862–872, DOI 10.1145/3442188.3445924, URL https://doi.org/10.1145/3442188.3445924

Diana E, Gill W, Kearns M, Kenthapadi K, Roth A (2021) Minimax Group Fairness: Algorithms and Experiments, Association for Computing Machinery, New York, NY, USA, p 66–76. URL https://doi.org/10.1145/3461702.3462523

DiCiccio C, Vasudevan S, Basu K, Kenthapadi K, Agarwal D (2020) Evaluating Fairness Using Permutation Tests, Association for Computing Machinery, New York, NY, USA, p 1467–1477. URL https://doi.org/10.1145/3394486.3403199

Dickens C, Singh R, Getoor L (2020) Hyperfair: A soft approach to integrating fairness criteria. RecSys 2020 workshop: "3rd FAccTRec Workshop on Responsible Recommendation", 2009.08952

Dieterich W, Mendoza C, Brennan T (2016) Compas risk scales: Demonstrating accuracy equity and predictive parity

Ding F, Hardt M, Miller J, Schmidt L (2021) Retiring adult: New datasets for fair machine learning. Advances in Neural Information Processing Systems 34

Dixon L, Li J, Sorensen J, Thain N, Vasserman L (2018) Measuring and mitigating unintended bias in text classification. In: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, Association for Computing Machinery, New York, NY, USA, AIES '18, p 67–73, DOI 10.1145/3278721.3278729, URL https://doi.org/10.1145/3278721.3278729

Donini M, Oneto L, Ben-David S, Shawe-Taylor JS, Pontil M (2018) Empirical risk minimization under fairness constraints. In: Bengio S, Wallach H, Larochelle H, Grauman K, Cesa-Bianchi N, Garnett R (eds) Advances in Neural Information Processing Systems, Curran Associates, Inc., vol 31, pp 2791–2801, URL https://proceedings.neurips.cc/paper/2018/file/83cdcec08fbf90370fcf53bdd56604ff-Paper.pdf

Dressel J, Farid H (2018) The accuracy, fairness, and limits of predicting recidivism. Science advances 4(1):eaao5580

Duarte MF, Hu YH (2004) Vehicle classification in distributed sensor networks. Journal of Parallel and Distributed Computing 64(7):826 – 838, DOI https://doi.org/10.1016/j.jpdc.2004.03.020, URL http://www.sciencedirect.com/science/article/pii/S0743731504000255, computing and Communication in Distributed Sensor Networks

Dwork C, Hardt M, Pitassi T, Reingold O, Zemel R (2012) Fairness through awareness. In: Proceedings of the 3rd Innovations in Theoretical Computer Science Conference, Association for Computing Machinery, New York, NY, USA, ITCS '12, p 214–226, DOI 10.1145/2090236.2090255, URL https://doi.org/10.1145/2090236.2090255

Dwork C, Immorlica N, Kalai AT, Leiserson M (2017) Decoupled classifiers for fair and efficient machine learning. KDD 2017 workshop: "Fairness, Accountability, and Transparency in Machine Learning (FAT/ML)", 1707.06613

Dwork C, Immorlica N, Kalai AT, Leiserson M (2018) Decoupled classifiers for group-fair and efficient machine learning. In: Friedler SA, Wilson C (eds) Proceedings of the 1st Conference on Fairness, Accountability and Transparency, PMLR, New York, NY, USA, Proceedings of Machine Learning Research, vol 81, pp 119–133, URL http://proceedings.mlr.press/v81/dwork18a.html

Ebner NC, Riediger M, 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

Eidinger E, Enbar R, Hassner T (2014) Age and gender estimation of unfiltered faces. IEEE Transactions on Information Forensics and Security 9(12):2170–2179, DOI 10.1109/TIFS.2014.2359646

Ekstrand MD, Tian M, Azpiazu IM, Ekstrand JD, Anuyah O, McNeill D, Pera MS (2018) All the cool kids, how do they fit in?: Popularity and demographic biases in recommender evaluation and effectiveness. In: Friedler SA, Wilson C (eds) Proceedings of the 1st Conference on Fairness, Accountability and Transparency, PMLR, New York, NY, USA, Proceedings of Machine Learning Research, vol 81, pp 172–186, URL http://proceedings.mlr.press/v81/ekstrand18b.html

El Emam K, Arbuckle L, Koru G, Eze B, Gaudette L, Neri E, Rose S, Howard J, 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

El Halabi M, Mitrović S, Norouzi-Fard A, Tardos J, Tarnawski JM (2020) Fairness in streaming submodular maximization: Algorithms and hardness.

In: Larochelle H, Ranzato M, Hadsell R, Balcan MF, Lin H (eds) Advances in Neural Information Processing Systems, Curran Associates, Inc., vol 33, pp 13609–13622, URL https://proceedings.neurips.cc/paper/2020/file/9d752cb08ef466fc480fba981cfa44a1-Paper.pdf

Elzayn H, Jabbari S, Jung C, Kearns M, Neel S, Roth A, Schutzman Z (2019) Fair algorithms for learning in allocation problems. In: Proceedings of the Conference on Fairness, Accountability, and Transparency, Association for Computing Machinery, New York, NY, USA, FAT* '19, p 170–179, DOI 10.1145/3287560.3287571, URL https://doi.org/10.1145/3287560.3287571

Epstein L, Landes W, Posner R (2013) The Behavior of Federal Judges: A Theoretical and Empirical Study of Rational Choice. Harvard University Press, URL https://books.google.it/books?id=RcQEBeic3ecC

Equivant (2019) Practitioner's guide to compas core. URL https://www.equivant.com/wp-content/uploads/Practitioners-Guide-to-COMPAS-Core-040419.pdf

Esmaeili S, Brubach B, Tsepenekas L, Dickerson J (2020) Probabilistic fair clustering. In: Larochelle H, Ranzato M, Hadsell R, Balcan MF, Lin H (eds) Advances in Neural Information Processing Systems, Curran Associates, Inc., vol 33, pp 12743–12755, URL https://proceedings.neurips.cc/paper/2020/file/95f2b84de5660ddf45c8a34933a2e66f-Paper.pdf

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

Fabbrizzi S, Papadopoulos S, Ntoutsi E, Kompatsiaris I (2021) A survey on bias in visual datasets. arXiv preprint arXiv:210707919

Fabris A, Mishler A, Gottardi S, Carletti M, Daicampi M, Susto GA, Silvello G (2021) Algorithmic Audit of Italian Car Insurance: Evidence of Unfairness in Access and Pricing, Association for Computing Machinery, New York, NY, USA, p 458–468. URL https://doi.org/10.1145/3461702.3462569

Farnad G, Babaki B, Gendreau M (2020) A unifying framework for fairness-aware influence maximization. In: Companion Proceedings of the Web Conference 2020, Association for Computing Machinery, New York, NY, USA, WWW '20, p 714–722, DOI 10.1145/3366424.3383555, URL https://doi.org/10.1145/3366424.3383555

Farnadi G, Kouki P, Thompson SK, Srinivasan S, Getoor L (2018) A fairness-aware hybrid recommender system. RecSys 2018 workshop: "Workshop on Responsible Recommendation (FAT/Rec)", 1809.09030

Farnadi G, Babaki B, Carvalho M (2019) Enhancing fairness in kidney exchange program by ranking solutions. NeurIPS 2019 workshop: "Fair ML for Health", 1911.05489

Fehrman E, Muhammad AK, Mirkes EM, Egan V, Gorban AN (2017) The five factor model of personality and evaluation of drug consumption risk. In: Palumbo F, Montanari A, Vichi M (eds) Data Science, Springer International Publishing, Cham, pp 231–242

Fehrman E, Egan V, Gorban AN, Levesley J, Mirkes EM, Muhammad AK (2019) Personality Traits and Drug Consumption: A story told by data. Springer

Feldman M, Friedler SA, Moeller J, Scheidegger C, Venkatasubramanian S (2015) Certifying and removing disparate impact. In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Association for Computing Machinery, New York, NY, USA, KDD '15, p 259–268, DOI 10.1145/2783258.2783311, URL https://doi.org/10.1145/2783258.2783311

Ferraro A, Bogdanov D, Serra X, Yoon J (2019) Artist and style exposure bias in collaborative filtering based music recommendations. ISMIR 2019 workshop: "Workshop on Designing Human-Centric MIR Systems", 1911.04827

Fish B, Kun J, Lelkes Á (2015) Fair boosting : a case study. ICML 2015 workshop: "Fairness, Accountability, and Transparency in Machine Learning (FAT/ML)"

Fisher J, Palfrey D, Christodoulopoulos C, Mittal A (2020) Measuring social bias in knowledge graph embeddings. AKBC 2020 workshop: "Bias in Automatic Knowledge Graph Construction", 1912.02761

Fisher RA (1936) The use of multiple measurements in taxonomic problems. Annals of eugenics 7(2):179–188

Fisman R, Iyengar S, Kamenica E, Simonson I (2006) Gender differences in mate selection: Evidence from a speed dating experiment. The Quarterly Journal of Economics 121:673–697, DOI 10.1162/qjec.2006.121.2.673

Fitzpatrick TB (1988) The validity and practicality of sun-reactive skin types i through vi. Archives of dermatology 124(6):869–871

Flanigan B, Gölz P, Gupta A, Procaccia AD (2020) Neutralizing self-selection bias in sampling for sortition. In: Larochelle H, Ranzato M, Hadsell R, Balcan M, Lin H (eds) Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual, URL https://proceedings.neurips.cc/paper/2020/hash/48237d9f2dea8c74c2a72126cf63d933-Abstract.html

Florez OU (2019) On the unintended social bias of training language generation models with data from local media. NeurIPS 2019 workshop: "Human-Centric Machine Learning", 1911.00461

Fogliato R, Xiang A, Lipton Z, Nagin D, 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), Virtual Event, pp 100–111, DOI 10.1145/3461702.3462538

Founta A, Djouvas C, Chatzakou D, Leontiadis I, Blackburn J, Stringhini G, Vakali A, Sirivianos M, 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, AAAI Press, pp 491–500, URL https://aaai.org/ocs/index.php/ICWSM/ICWSM18/paper/view/17909

Framingham Heart Study (2021) Framingham heart study. offspring exam 10, omni 1 exam 5. research consent form. URL https://framinghamheartstudy.org/files/2021/01/FHS-Offspring-Exam-10-Omni-1-Exam-5-Informed-Consent-English-Language-v21.pdf

Friedler SA, Scheidegger C, Venkatasubramanian S, Choudhary S, Hamilton EP, Roth D (2019) A comparative study of fairness-enhancing interventions in machine learning. In: Proceedings of the Conference on Fairness, Accountability, and Transparency, Association for Computing Machinery, New York, NY, USA, FAT* '19, p 329–338, DOI 10.1145/3287560.3287589, URL https://doi.org/10.1145/3287560.3287589

Friedler SA, Scheidegger C, Venkatasubramanian S (2021) The (im)possibility of fairness: Different value systems require different mechanisms for fair decision making. Commun ACM 64(4):136–143, DOI 10.1145/3433949, URL https://doi.org/10.1145/3433949

Galhotra S, Saisubramanian S, Zilberstein S (2021) Learning to Generate Fair Clusters from Demonstrations, Association for Computing Machinery, New York, NY, USA, p 491–501. URL https://doi.org/10.1145/3461702.3462558

Garbin C, Rajpurkar P, Irvin J, Lungren MP, Marques O (2021) Structured dataset documentation: a datasheet for chexpert. 2105.03020

Garg S, Perot V, Limtiaco N, Taly A, Chi EH, Beutel A (2019) Counterfactual fairness in text classification through robustness. In: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, Association for Computing Machinery, New York, NY, USA, AIES '19, p 219–226, DOI 10.1145/3306618.3317950, URL https://doi.org/10.1145/3306618.3317950

Gastwirth JL, 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

Ge H, Caverlee J, Lu H (2016) Taper: A contextual tensor-based approach for personalized expert recommendation. In: Proceedings of the 10th ACM Conference on Recommender Systems, Association for Computing Machinery, New York, NY, USA, RecSys '16, p 261–268, DOI 10.1145/2959100.2959151, URL https://doi.org/10.1145/2959100.2959151

Gebru T, Morgenstern J, Vecchione B, Vaughan JW, Wallach H, Daumé III H, Crawford K (2018) Datasheets for datasets. arXiv preprint arXiv:180309010

Geiger RS, Yu K, Yang Y, Dai M, Qiu J, Tang R, 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, Association for Computing Machinery, New York, NY, USA, FAT* '20, p 325–336, DOI 10.1145/3351095.3372862, URL https://doi.org/10.1145/3351095.3372862

Gelman A, Fagan J, 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

Gerritse EJ, de Vries AP (2020) Effect of debiasing on information retrieval. In: Boratto L, Faralli S, Marras M, Stilo G (eds) Bias and Social Aspects in Search and Recommendation, Springer International Publishing, Cham, pp 35–42

Ghadiri M, Samadi S, Vempala S (2021) Socially fair k-means clustering. In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, Association for Computing Machinery, New York, NY, USA, FAccT '21, p 438–448, DOI 10.1145/3442188.3445906, URL https://doi.org/10.1145/3442188.3445906

Ginart A, Guan M, Valiant G, Zou JY (2019) Making ai forget you: Data deletion in machine learning. In: Wallach H, Larochelle H, Beygelzimer A, d'Alché-Buc F, Fox E, Garnett R (eds) Advances in Neural Information Processing Systems, Curran Associates, Inc., vol 32, URL https://proceedings.neurips.cc/paper/2019/file/cb79f8fa58b91d3af6c9c991f63962d3-Paper.pdf

Go A, Bhayani R, Huang L (2009) Twitter sentiment classification using distant supervision. Processing 150

Goel N, Faltings B (2019) Crowdsourcing with fairness, diversity and budget constraints. In: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, Association for Computing Machinery, New York, NY, USA, AIES '19, p 297–304, DOI 10.1145/3306618.3314282, URL https://doi.org/10.1145/3306618.3314282

Goel N, Yaghini M, Faltings B (2018) Non-discriminatory machine learning through convex fairness criteria. In: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, Association for Computing Machinery, New York, NY, USA, AIES '18, p 116, DOI 10.1145/3278721.3278722, URL https://doi.org/10.1145/3278721.3278722

Goel N, Amayuelas A, Deshpande A, 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)", 2012.11448

Goel S, Rao JM, 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

Goel S, Perelman M, Shroff R, Sklansky D (2017) Combatting police discrimination in the age of big data. New Criminal Law Review 20(2):181–232, DOI 10.1525/nclr.2017.20.2.181

Goelz P, Kahng A, Procaccia AD (2019) Paradoxes in fair machine learning. In: Wallach H, Larochelle H, Beygelzimer A, d'Alché-Buc F, Fox E, Garnett R (eds) Advances in Neural Information Processing Systems, Curran Associates, Inc., vol 32, pp 8342–8352, URL https://proceedings.neurips.cc/paper/2019/file/bbc92a647199b832ec90d7cf57074e9e-Paper.pdf

Golbeck J, Ashktorab Z, Banjo RO, Berlinger A, Bhagwan S, Buntain C, Cheakalos P, Geller AA, Gergory Q, Gnanasekaran RK, Gunasekaran RR, Hoffman KM, Hottle J, Jienjitlert V, Khare S, Lau R, Martindale MJ, Naik S, Nixon HL, Ramachandran P, Rogers KM, Rogers L, Sarin MS, Shahane G, Thanki J, Vengataraman P, Wan Z, Wu DM (2017) A large labeled corpus for online harassment research. In: Proceedings of the 2017 ACM on Web Science Conference, Association for Computing Machinery, New York, NY, USA, WebSci '17, p 229–233, DOI 10.1145/3091478.3091509, URL https://doi.org/10.1145/3091478.3091509

Goldstein H (1991) Multilevel modelling of survey data. Journal of the Royal Statistical Society Series D (The Statistician) 40(2):235–244, URL http://www.jstor.org/stable/2348496

Gong S, Liu X, Jain AK (2021) Mitigating face recognition bias via group adaptive classifier. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 3414–3424

Gordaliza P, Barrio ED, Fabrice G, Loubes JM (2019) Obtaining fairness using optimal transport theory. In: Chaudhuri K, Salakhutdinov R (eds) Proceedings of the 36th International Conference on Machine Learning, PMLR, Long Beach, California, USA, Proceedings of Machine Learning Research, vol 97, pp 2357–2365, URL http://proceedings.mlr.press/v97/gordaliza19a.html

Gordon J, Babaeianjelodar M, Matthews J (2020) Studying political bias via word embeddings. In: Companion Proceedings of the Web Conference 2020, Association for Computing Machinery, New York, NY, USA, WWW '20, p 760–764, DOI 10.1145/3366424.3383560, URL https://doi.org/10.1145/3366424.3383560

Goyal Y, Khot T, Summers-Stay D, Batra D, 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, pp 6904–6913

Graffam J, Shinkfield AJ, Hardcastle L (2008) The perceived employability of ex-prisoners and offenders. International Journal of Offender Therapy and Comparative Criminology 52(6):673–685, DOI 10.1177/0306624X07307783, URL https://doi.org/10.1177/0306624X07307783, https://doi.org/10.1177/0306624X07307783

Green B, 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, Association for Computing Machinery, New York, NY, USA, FAT* '19, p 90–99, DOI 10.1145/3287560.3287563, URL https://doi.org/10.1145/3287560.3287563

Greenwald AG, McGhee DE, Schwartz JL (1998) Measuring individual differences in implicit cognition: the implicit association test. Journal of personality and social psychology 74(6):1464

Grgic-Hlaca N, Zafar M, Gummadi K, Weller A (2016) The case for process fairness in learning: Feature selection for fair decision making. NeurIPS 2016 workshop: "Machine Learning and the Law"

Grömping U (2019) South German Credit Data: Correcting a Widely Used Data Set. Report. Tech. rep., Beuth University of Applied Sciences Berlin, URL http://www1.beuth-hochschule.de/FB_II/reports/Report-2019-004.pdf

Gulla JA, Zhang L, Liu P, Özgöbek O, Su X (2017) The adressa dataset for news recommendation. In: Proceedings of the International Conference on Web Intelligence, Association for Computing Machinery, New York, NY, USA, WI '17, p 1042–1048, DOI 10.1145/3106426.3109436, URL https://doi.org/10.1145/3106426.3109436

Gungor A (2018) Benchmarking authorship attribution techniques using over a thousand books by fifty victorian era novelists. Master's thesis, Purdue University

Guo G, Zhang J, Yorke-Smith N (2016a) A novel evidence-based bayesian similarity measure for recommender systems. ACM Trans Web 10(2), DOI 10.1145/2856037, URL https://doi.org/10.1145/2856037

Guo W, Caliskan A (2021) Detecting Emergent Intersectional Biases: Contextualized Word Embeddings Contain a Distribution of Human-like Biases, Association for Computing Machinery, New York, NY, USA, p 122–133. URL https://doi.org/10.1145/3461702.3462536

Guo Y, Zhang L, Hu Y, He X, Gao J (2016b) Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In: Leibe B, Matas J, Sebe N, Welling M (eds) Computer Vision – ECCV 2016, Springer International Publishing, Cham, pp 87–102

Guvenir HA, Acar B, Demiroz G, Cekin A (1997) A supervised machine learning algorithm for arrhythmia analysis. In: Computers in Cardiology 1997, IEEE, pp 433–436

Han H, Jain AK (2014) Age, gender and race estimation from unconstrained face images. URL http://biometrics.cse.msu.edu/Publications/Face/HanJain_UnconstrainedAgeGenderRaceEstimation_MSUTechReport2014.pdf

Hannák A, Wagner C, Garcia D, Mislove A, Strohmaier M, 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, Association for Computing Machinery, New York, NY, USA, CSCW '17, p 1914–1933, DOI 10.1145/2998181.2998327, URL https://doi.org/10.1145/2998181.2998327

Har-Peled S, 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, Garnett R (eds) Advances in Neural Information Processing Systems, Curran Associates, Inc., vol 32, pp 13176–13187, URL https://proceedings.neurips.cc/paper/2019/file/742141ceda6b8f6786609d31c8ef129f-Paper.pdf

Harb E, Lam HS (2020) Kfc: A scalable approximation algorithm for k-center fair clustering. In: Larochelle H, Ranzato M, Hadsell R, Balcan MF, Lin H (eds) Advances in Neural Information Processing Systems, Curran Associates, Inc., vol 33, pp 14509–14519, URL https://proceedings.neurips.cc/paper/2020/file/a6d259bfbfa2062843ef543e21d7ec8e-Paper.pdf

Hardt M, Price E, Price E, Srebro N (2016) Equality of opportunity in supervised learning. In: Lee D, Sugiyama M, Luxburg U, Guyon I, Garnett R (eds) Advances in Neural Information Processing Systems, Curran Associates, Inc., vol 29, pp 3315–3323, URL https://proceedings.neurips.cc/paper/2016/file/9d2682367c3935defcb1f9e247a97c0d-Paper.pdf

Harper FM, Konstan JA (2015) The movielens datasets: History and context. ACM Trans Interact Intell Syst 5(4), DOI 10.1145/2827872, URL https://doi.org/10.1145/2827872

Hashimoto T, Srivastava M, Namkoong H, Liang P (2018) Fairness without demographics in repeated loss minimization. In: Dy J, Krause A (eds) Proceedings of the 35th International Conference on Machine Learning, PMLR, Stockholmsmässan, Stockholm Sweden, Proceedings of Machine Learning Research, vol 80, pp 1929–1938, URL http://proceedings.mlr.press/v80/hashimoto18a.html

He R, McAuley J (2016) Ups and downs. Proceedings of the 25th International Conference on World Wide Web DOI 10.1145/2872427.2883037, URL http://dx.doi.org/10.1145/2872427.2883037

He R, Kang WC, McAuley J (2017) Translation-based recommendation. In: Proceedings of the Eleventh ACM Conference on Recommender Systems, Association for Computing Machinery, New York, NY, USA, RecSys '17, p 161–169, DOI 10.1145/3109859.3109882, URL https://doi.org/10.1145/3109859.3109882

He Y, Burghardt K, Guo S, 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)", 2010.16409

He Y, Burghardt K, Lerman K (2020b) A geometric solution to fair representations. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, Association for Computing Machinery, New York, NY, USA, AIES '20, p 279–285, DOI 10.1145/3375627.3375864, URL https://doi.org/10.1145/3375627.3375864

Heidari H, Ferrari C, Gummadi K, 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, Garnett R (eds) Advances in Neural Information Processing Systems, Curran Associates, Inc., vol 31, pp 1265–1276, URL https://proceedings.neurips.cc/paper/2018/file/be3159ad04564bfb90db9e32851ebf9c-Paper.pdf

Heidari H, Loi M, Gummadi KP, 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, Association for Computing Machinery, New York, NY, USA, FAT* '19, p 181–190, DOI 10.1145/3287560.3287584, URL https://doi.org/10.1145/3287560.3287584

Heidari H, Nanda V, Gummadi K (2019b) On the long-term impact of algorithmic decision policies: Effort unfairness and feature segregation through social learning. In: Chaudhuri K, Salakhutdinov R (eds) Proceedings of the 36th International Conference on Machine Learning, PMLR, Long Beach, California, USA, Proceedings of Machine Learning Research, vol 97, pp 2692–2701, URL http://proceedings.mlr.press/v97/heidari19a.html

Hendricks LA, Burns K, Saenko K, Darrell T, Rohrbach A (2018) Women also snowboard: Overcoming bias in captioning models. In: Ferrari V, Hebert M, Sminchisescu C, Weiss Y (eds) Computer Vision – ECCV 2018, Springer International Publishing, Cham, pp 793–811

Higgins I, Matthey L, Pal A, Burgess C, Glorot X, Botvinick M, Mohamed S, Lerchner A (2017) beta-vae: Learning basic visual concepts with a constrained variational framework. In: ICLR

Holland S, Hosny A, Newman S, Joseph J, Chmielinski K (2018) The dataset nutrition label: A framework to drive higher data quality standards. arXiv preprint arXiv:180503677

Hollywood J, McKay K, Woods D, Agniel D (2019) Real time crime centers in chicago. URL https://www.rand.org/content/dam/rand/pubs/research_reports/RR3200/RR3242/RAND_RR3242.pdf

Holmes MD, Smith BW, Freng AB, Muñoz EA (2008) Minority threat, crime control, and police resource allocation in the southwestern united states. Crime & Delinquency 54(1):128–152, DOI 10.1177/0011128707309718, URL https://doi.org/10.1177/0011128707309718, https://doi.org/10.1177/0011128707309718

Holstein K, Wortman Vaughan J, Daumé III H, Dudik M, Wallach H (2019) 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), Glasgow, UK, pp 1–16

Houvardas J, Stamatatos E (2006) N-gram feature selection for authorship identification. In: International conference on artificial intelligence: Methodology, systems, and applications, Springer, pp 77–86

Hu L, Chen Y (2020) Fair classification and social welfare. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, Association for Computing Machinery, New York, NY, USA, FAT* '20, p 535–545, DOI 10.1145/3351095.3372857, URL https://doi.org/10.1145/3351095.3372857

Hu Y, Wu Y, Zhang L, Wu X (2020) Fair multiple decision making through soft interventions. In: Larochelle H, Ranzato M, Hadsell R, Balcan M, Lin H (eds) Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual, URL https://proceedings.neurips.cc/paper/2020/hash/d0921d442ee91b89ad95059d13df618-Abstract.html

Huan W, Wu Y, Zhang L, Wu X (2020) Fairness through equality of effort. In: Companion Proceedings of the Web Conference 2020, Association for Computing Machinery, New York, NY, USA, WWW '20, p 743–751, DOI 10.1145/3366424.3383558, URL https://doi.org/10.1145/3366424.3383558

Huang GB, Ramesh M, Berg T, Learned-Miller E (2007) Labeled faces in the wild: A database for studying face recognition in unconstrained environments

Huang L, Vishnoi N (2019) Stable and fair classification. In: Chaudhuri K, Salakhutdinov R (eds) Proceedings of the 36th International Conference on Machine Learning, PMLR, Long Beach, California, USA, Proceedings of Machine Learning Research, vol 97, pp 2879–2890, URL http://proceedings.mlr.press/v97/huang19e.html

Huang L, Jiang S, Vishnoi N (2019) Coresets for clustering with fairness constraints. In: Advances in Neural Information Processing Systems, pp 7589–7600

Huang L, Wei J, 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, Association for Computing Machinery, New York, NY, USA, AIES '20, p 293–299, DOI 10.1145/3375627.3375848, URL https://doi.org/10.1145/3375627.3375848

Hull J (1994) A database for handwritten text recognition research. IEEE Transactions on Pattern Analysis and Machine Intelligence 16(5):550–554, DOI 10.1109/34.291440

Hussain S, Dahan NA, Ba-Alwib FM, 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

Hutchinson B, Prabhakaran V, Denton E, Webster K, Zhong Y, Denuyl S (2020) Unintended machine learning biases as social barriers for persons with disabilities. SIGACCESS Access Comput (125), DOI 10.1145/3386296.3386305, URL https://doi.org/10.1145/3386296.3386305

Hutchinson B, Smart A, Hanna A, Denton E, Greer C, Kjartansson O, Barnes P, 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, Association for Computing Machinery, New York, NY, USA, FAccT '21, p 560–575

Häußler M Walter (1979) Empirische ergebnisse zu diskriminationsverfahren bei kreditscoringsystemen. URL https://link.springer.com/article/10.1007/BF01917956

International Warfarin Pharmacogenetics Consortium (2009) Estimation of the warfarin dose with clinical and pharmacogenetic data. New England Journal of Medicine 360(8):753–764

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, 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, AAAI Press, 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, pp 590–597, 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

Islam R, Pan S, Foulds JR (2021) Can We Obtain Fairness For Free?, Association for Computing Machinery, New York, NY, USA, p 586–596. URL https://doi.org/10.1145/3461702.3462614

Jabbari S, Ou HC, Lakkaraju H, 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"

Jagielski M, Kearns M, Mao J, Oprea A, Roth A, Malvajerdi SS, Ullman J (2019) Differentially private fair learning. In: Chaudhuri K, Salakhutdinov R (eds) Proceedings of the 36th International Conference on Machine Learning, PMLR, Long Beach, California, USA, Proceedings of Machine Learning Research, vol 97, pp 3000–3008, URL http://proceedings.mlr.press/v97/jagielski19a.html

Ji D, Smyth P, 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, Lin H (eds) Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual, URL https://proceedings.neurips.cc/paper/2020/hash/d83de59e10227072a9c034ce10029c39-Abstract.html

Jiang W, Pardos ZA (2021) Towards Equity and Algorithmic Fairness in Student Grade Prediction, Association for Computing Machinery, New York, NY, USA, p 608–617. URL https://doi.org/10.1145/3461702.3462623

Jo ES, 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, Association for Computing Machinery, New York, NY, USA, FAT* '20, p 306–316, DOI 10.1145/3351095.3372829, URL https://doi.org/10.1145/3351095.3372829

Johnson AE, Pollard TJ, Shen L, Lehman LH, Feng M, Ghassemi M, Moody B, Szolovits P, Celi LA, Mark RG (2016) Mimic-iii, a freely accessible critical care database. Scientific data 3:160035

Johnson AE, Pollard TJ, Greenbaum NR, Lungren MP, Deng Cy, Peng Y, Lu Z, Mark RG, Berkowitz SJ, Horng S (2019) Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. arXiv preprint arXiv:190107042

Jones E, Sagawa S, Koh PW, Kumar A, Liang P (2021) Selective classification can magnify disparities across groups. In: International Conference on Learning Representations, URL https://openreview.net/forum?id=N0M_4BkQ05i

Jones M, Nguyen H, Nguyen T (2020) Fair k-centers via maximum matching. In: III HD, Singh A (eds) Proceedings of the 37th International Conference on Machine Learning, PMLR, Virtual, Proceedings of Machine Learning Research, vol 119, pp 4940–4949, URL http://proceedings.mlr.press/v119/jones20a.html

Jung S, Lee D, Park T, Moon T (2021) Fair feature distillation for visual recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp 12115–12124

Kallus N, Zhou A (2018) Residual unfairness in fair machine learning from prejudiced data. In: Dy J, Krause A (eds) Proceedings of the 35th International Conference on Machine Learning, PMLR, Stockholmsmässan, Stockholm Sweden, Proceedings of Machine Learning Research, vol 80, pp 2439–2448, URL http://proceedings.mlr.press/v80/kallus18a.html

Kallus N, 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, Garnett R (eds) Advances in Neural Information Processing Systems, Curran Associates, Inc., vol 32, pp 3426–3437, URL https://proceedings.neurips.cc/paper/2019/file/d54e99a6c03704e95e6965532dec148b-Paper.pdf

Kallus N, 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, Garnett R (eds) Advances in Neural Information Processing Systems, Curran Associates, Inc., vol 32, pp 3438–3448, URL https://proceedings.neurips.cc/paper/2019/file/73e0f7487b8e5297182c5a711d20bf26-Paper.pdf

Kallus N, Zhou A (2021) Fairness, welfare, and equity in personalized pricing. In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, Association for Computing Machinery, New York, NY, USA, FAccT '21, p 296–314, DOI 10.1145/3442188.3445895, URL https://doi.org/10.1145/3442188.3445895

Kallus N, Mao X, Zhou A (2020) Assessing algorithmic fairness with unobserved protected class using data combination. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, Association for Computing Machinery, New York, NY, USA, FAT* '20, p 110, DOI 10.1145/3351095.3373154, URL https://doi.org/10.1145/3351095.3373154

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, Association for Computing Machinery, New York, NY, USA, KDD '03, p 583–588, DOI 10.1145/956750.956823, URL https://doi.org/10.1145/956750.956823

Kang J, He J, Maciejewski R, Tong H (2020) InFoRM: Individual Fairness on Graph Mining, Association for Computing Machinery, New York, NY, USA, p 379–389. URL https://doi.org/10.1145/3394486.3403080

Kannel WB, McGee DL (1979) Diabetes and cardiovascular disease: the framingham study. Jama 241(19):2035–2038

Karako C, 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)”, 1809.03577

Karkkainen K, 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, pp 1548–1558

Karlan DS, Zinman J (2008) Credit elasticities in less-developed economies: Implications for microfinance. American Economic Review 98(3):1040–68

Kasy M, Abebe R (2021) Fairness, equality, and power in algorithmic decision-making. In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, Association for Computing Machinery, New York, NY, USA, FAccT ’21, p 576–586, DOI 10.1145/3442188.3445919, URL https://doi.org/10.1145/3442188.3445919

Kato M, Teshima T, Honda J (2019) Learning from positive and unlabeled data with a selection bias. In: International Conference on Learning Representations, URL https://openreview.net/forum?id=rJzLciCqKm

Kearns M, Neel S, Roth A, Wu ZS (2018) Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In: Dy J, Krause A (eds) Proceedings of the 35th International Conference on Machine Learning, PMLR, Stockholmsmässan, Stockholm Sweden, Proceedings of Machine Learning Research, vol 80, pp 2564–2572, URL http://proceedings.mlr.press/v80/kearns18a.html

Kearns M, Neel S, Roth A, Wu ZS (2019) An empirical study of rich subgroup fairness for machine learning. In: Proceedings of the Conference on Fairness, Accountability, and Transparency, Association for Computing Machinery, New York, NY, USA, FAT* ’19, p 100–109, DOI 10.1145/3287560.3287592, URL https://doi.org/10.1145/3287560.3287592

Keswani V, Lease M, Kenthapadi K (2021) Towards Unbiased and Accurate Deferral to Multiple Experts, Association for Computing Machinery, New York, NY, USA, p 154–165. URL https://doi.org/10.1145/3461702.3462516

Keyes O, Stevens N, Wernimont J (2019) The government is using the most vulnerable people to test facial recognition software

Khan Z, 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, Association for Computing Machinery, New York, NY, USA, FAccT ’21, p 587–597, DOI 10.1145/3442188.3445920, URL https://doi.org/10.1145/3442188.3445920

Kilbertus N, Gascon A, Kusner M, Veale M, Gummadi K, Weller A (2018) Blind justice: Fairness with encrypted sensitive attributes. In: Dy J, Krause A (eds) Proceedings of the 35th International Conference on Machine Learning, PMLR, Stockholmsmässan, Stockholm Sweden, Proceedings of Machine Learning Research, vol 80, pp 2630–2639, URL http://proceedings.mlr.

press/v80/kilbertus18a.html
Kim E, Bryant D, Srikanth D, Howard A (2021) Age Bias in Emotion Detection: An Analysis of Facial Emotion Recognition Performance on Young, Middle-Aged, and Older Adults, Association for Computing Machinery, New York, NY, USA, p 638–644. URL https://doi.org/10.1145/3461702.3462609
Kim H, Mnih A (2018) Disentangling by factorising. In: Dy J, Krause A (eds) Proceedings of the 35th International Conference on Machine Learning, PMLR, Proceedings of Machine Learning Research, vol 80, pp 2649–2658, URL http://proceedings.mlr.press/v80/kim18b.html
Kim JS, Chen J, Talwalkar A (2020) FACT: A diagnostic for group fairness trade-offs. In: III HD, Singh A (eds) Proceedings of the 37th International Conference on Machine Learning, PMLR, Virtual, Proceedings of Machine Learning Research, vol 119, pp 5264–5274, URL http://proceedings.mlr.press/v119/kim20a.html
Kim MP, Ghorbani A, 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, Association for Computing Machinery, New York, NY, USA, AIES '19, p 247–254, DOI 10.1145/3306618.3314287, URL https://doi.org/10.1145/3306618.3314287
Kiritchenko S, 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, Association for Computational Linguistics, New Orleans, Louisiana, pp 43–53, DOI 10.18653/v1/S18-2005, URL https://aclanthology.org/S18-2005
Kizhner I, Terras M, Rumyantsev M, Khokhlova V, Demeshkova E, Rudov I, 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, DOI 10.1093/llc/fqaa055, URL https://doi.org/10.1093/llc/fqaa055, https://academic.oup.com/dsh/article-pdf/36/3/607/40873280/fqaa055.pdf
Klare BF, Klein B, Taborsky E, Blanton A, Cheney J, Allen K, Grother P, Mah A, Jain AK (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)
Kleindessner M, Awasthi P, Morgenstern J (2019a) Fair k-center clustering for data summarization. In: Chaudhuri K, Salakhutdinov R (eds) Proceedings of the 36th International Conference on Machine Learning, PMLR, Long Beach, California, USA, Proceedings of Machine Learning Research, vol 97, pp 3448–3457, URL http://proceedings.mlr.press/v97/kleindessner19a.html
Kleindessner M, Samadi S, Awasthi P, Morgenstern J (2019b) Guarantees for spectral clustering with fairness constraints. In: Chaudhuri K, Salakhutdinov R (eds) Proceedings of the 36th International Conference on Machine Learning, PMLR, Long Beach, California, USA, Proceedings of Machine Learning Research, vol 97, pp 3458–3467, URL http://proceedings.mlr.

press/v97/kleindessner19b.html
Knees P, Hübler M (2019) Towards uncovering dataset biases: Investigating record label diversity in music playlists. ISMIR 2019 workshop: "Workshop on Designing Human-Centric MIR Systems"
Kobren A, Saha B, McCallum A (2019) Paper matching with local fairness constraints. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Association for Computing Machinery, New York, NY, USA, KDD '19, p 1247–1257, DOI 10.1145/3292500.3330899, URL https://doi.org/10.1145/3292500.3330899
Koch B, Denton E, Hanna A, Foster JG (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), URL https://openreview.net/forum?id=zNQBIBKJRkd
Kocijan V, Camburu OM, 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)", 2011.01837
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, AAAI Press, KDD'96, p 202–207
Komiyama J, Takeda A, Honda J, Shimao H (2018) Nonconvex optimization for regression with fairness constraints. In: Dy J, Krause A (eds) Proceedings of the 35th International Conference on Machine Learning, PMLR, Stockholmsmässan, Stockholm Sweden, Proceedings of Machine Learning Research, vol 80, pp 2737–2746, URL http://proceedings.mlr.press/v80/komiyama18a.html
Konstantakis G, Promponas G, Dretakis M, Papadakos P (2020) Bias goggles: Exploring the bias of web domains through the eyes of users. In: Boratto L, Faralli S, Marras M, Stilo G (eds) Bias and Social Aspects in Search and Recommendation, Springer International Publishing, Cham, pp 66–71
Koolen C (2018) Reading beyond the female the relationship between perception of author gender and literary quality. PhD thesis, University of Amsterdam
Koolen C, 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, Association for Computational Linguistics, Valencia, Spain, pp 12–22, DOI 10.18653/v1/W17-1602, URL https://www.aclweb.org/anthology/W17-1602
Krizhevsky A (2009) Learning multiple layers of features from tiny images
Kröger JL, Miceli M, Müller F (2021) How data can be used against people: A classification of personal data misuses. Available at SSRN 3887097
Kuhlman C, Rundensteiner E (2020) Rank aggregation algorithms for fair consensus. Proc VLDB Endow 13(12):2706–2719, DOI 10.14778/3407790.3407855, URL https://doi.org/10.14778/3407790.3407855

Kuhlman C, Gerych W, 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, Association for Computing Machinery, New York, NY, USA, AIES '21, p 674–682, DOI 10.1145/3461702.3462588, URL https://doi.org/10.1145/3461702.3462588

Kulshrestha J, Eslami M, Messias J, Zafar MB, Ghosh S, Gummadi KP, Karahalios K (2017) 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, Association for Computing Machinery, New York, NY, USA, CSCW '17, p 417–432, DOI 10.1145/2998181.2998321, URL https://doi.org/10.1145/2998181.2998321

Kushmerick N (1999) Learning to remove internet advertisements. In: Proceedings of the Third Annual Conference on Autonomous Agents, Association for Computing Machinery, New York, NY, USA, AGENTS '99, p 175–181, DOI 10.1145/301136.301186, URL https://doi.org/10.1145/301136.301186

Kusner MJ, Loftus J, Russell C, Silva R (2017) Counterfactual fairness. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, Garnett R (eds) Advances in Neural Information Processing Systems, Curran Associates, Inc., vol 30, pp 4066–4076, URL https://proceedings.neurips.cc/paper/2017/file/a486cd07e4ac3d270571622f4f316ec5-Paper.pdf

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

von Kügelgen J, Karimi AH, Bhatt U, Valera I, Weller A, Schölkopf B (2021) On the fairness of causal algorithmic recourse. NeurIPS 2020 workshop: "Algorithmic Fairness through the Lens of Causality and Interpretability (AFCI)", 2010.06529

Lahoti P, Beutel A, Chen J, Lee K, Prost F, Thain N, Wang X, Chi E (2020) Fairness without demographics through adversarially reweighted learning. In: Larochelle H, Ranzato M, Hadsell R, Balcan MF, Lin H (eds) Advances in Neural Information Processing Systems, Curran Associates, Inc., vol 33, pp 728–740, URL https://proceedings.neurips.cc/paper/2020/file/07fc15c9d169ee48573edd749d25945d-Paper.pdf

Lake BM, Salakhutdinov R, Tenenbaum JB (2015) Human-level concept learning through probabilistic program induction. Science 350(6266):1332–1338, DOI 10.1126/science.aab3050, URL https://science.sciencemag.org/content/350/6266/1332, https://science.sciencemag.org/content/350/6266/1332.full.pdf

Lamy A, Zhong Z, Menon AK, Verma N (2019) Noise-tolerant fair classification. In: Wallach H, Larochelle H, Beygelzimer A, d'Alché-Buc F, Fox E, Garnett R (eds) Advances in Neural Information Processing Systems, Curran Associates, Inc., vol 32, pp 294–306, URL https://proceedings.neurips.cc/paper/2019/file/8d5e957f297893487bd98fa830fa6413-Paper.pdf

Lan C, Huan J (2017) Discriminatory transfer. KDD 2017 workshop: "Fairness, Accountability, and Transparency in Machine Learning (FAT/ML)", 1707.00780

Larson J, Mattu S, Kirchner L, Angwin J (2016) How we analyzed the compas recidivism algorithm. URL https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

Le Quy T, Roy A, Iosifidis V, Zhang W, Ntoutsi E (2022) A survey on datasets for fairness-aware machine learning. WIREs Data Mining and Knowledge Discovery n/a(n/a):e1452, DOI https://doi.org/10.1002/widm.1452, URL https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/widm.1452, https://wires.onlinelibrary.wiley.com/doi/pdf/10.1002/widm.1452

Leavy S, Meaney G, Wade K, Greene D (2019) Curatr: A Platform for Semantic Analysis and Curation of Historical Literary Texts, pp 354–366. DOI 10.1007/978-3-030-36599-8_31

Leavy S, Meaney G, Wade K, Greene D (2020) Mitigating gender bias in machine learning data sets. In: Boratto L, Faralli S, Marras M, Stilo G (eds) Bias and Social Aspects in Search and Recommendation, Springer International Publishing, Cham, pp 12–26

Lecun Y, Bottou L, Bengio Y, Haffner P (1998) Gradient-based learning applied to document recognition. Proceedings of the IEEE 86(11):2278–2324, DOI 10.1109/5.726791

LeCun Y, Fu Jie Huang, 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., vol 2, pp II–104 Vol.2, DOI 10.1109/CVPR.2004.1315150

Lee H, Kizilcec RF (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)", 2007.00088

Leonelli S, Tempini N (2020) Data journeys in the sciences. Springer Nature

Leskovec J, Mcauley J (2012) Learning to discover social circles in ego networks. In: Pereira F, Burges CJC, Bottou L, Weinberger KQ (eds) Advances in Neural Information Processing Systems, Curran Associates, Inc., vol 25, URL https://proceedings.neurips.cc/paper/2012/file/7a614fd06c325499f1680b9896beedeb-Paper.pdf

Leskovec J, Kleinberg J, Faloutsos C (2007) Graph evolution: Densification and shrinking diameters. ACM transactions on Knowledge Discovery from Data (TKDD) 1(1):2–es

Lesmana NS, Zhang X, Bei X (2019) Balancing efficiency and fairness in on-demand ridesourcing. In: Wallach H, Larochelle H, Beygelzimer A, d'Alché-Buc F, Fox E, Garnett R (eds) Advances in Neural Information Processing Systems, Curran Associates, Inc., vol 32, pp 5309–5319, URL https://proceedings.neurips.cc/paper/2019/file/3070e6addcd702cb58de5d7897bfdae1-Paper.pdf

Levy D, Splansky GL, Strand NK, Atwood LD, Benjamin EJ, Blease S, Cupples LA, D'Agostino Sr RB, Fox CS, 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

Li P, Zhao H, Liu H (2020a) Deep fair clustering for visual learning. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)

Li P, Wang Y, Zhao H, Hong P, Liu H (2021) On dyadic fairness: Exploring and mitigating bias in graph connections. In: International Conference on Learning Representations, URL https://openreview.net/forum?id=xgGS6PmzNq6

Li T, Sanjabi M, Beirami A, Smith V (2020b) Fair resource allocation in federated learning. In: International Conference on Learning Representations, URL https://openreview.net/forum?id=ByexElSYDr

Li Y, Ning Y, Liu R, Wu Y, 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, Association for Computing Machinery, New York, NY, USA, WWW '20, p 733–742, DOI 10.1145/3366424.3383557, URL https://doi.org/10.1145/3366424.3383557

Li Y, Sun H, Wang WH (2020d) Towards Fair Truth Discovery from Biased Crowdsourced Answers, Association for Computing Machinery, New York, NY, USA, p 599–607. URL https://doi.org/10.1145/3394486.3403102

Li Z, Zhao H, Liu Q, Huang Z, Mei T, 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, Association for Computing Machinery, New York, NY, USA, KDD '18, p 1734–1743, DOI 10.1145/3219819.3220014, URL https://doi.org/10.1145/3219819.3220014

Liang L, Acuna DE (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, Association for Computing Machinery, New York, NY, USA, FAT* '20, p 403–412, DOI 10.1145/3351095.3375623, URL https://doi.org/10.1145/3351095.3375623

Lin TY, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P, Zitnick CL (2014) Microsoft coco: Common objects in context. In: Fleet D, Pajdla T, Schiele B, Tuytelaars T (eds) Computer Vision – ECCV 2014, Springer International Publishing, Cham, pp 740–755

Lipton Z, McAuley J, Chouldechova A (2018) Does mitigating ml's impact disparity require treatment disparity? In: Bengio S, Wallach H, Larochelle H, Grauman K, Cesa-Bianchi N, Garnett R (eds) Advances in Neural Information Processing Systems, Curran Associates, Inc., vol 31, URL https://proceedings.neurips.cc/paper/2018/file/8e0384779e58ce2af40eb365b318cc32-Paper.pdf

Liu D, Shafi Z, Fleisher W, Eliassi-Rad T, Alfeld S (2021) RAWLSNET: Altering Bayesian Networks to Encode Rawlsian Fair Equality of Opportunity, Association for Computing Machinery, New York, NY, USA, p 745–755. URL https://doi.org/10.1145/3461702.3462618

Liu LT, Dean S, Rolf E, Simchowitz M, Hardt M (2018) Delayed impact of fair machine learning. In: Dy J, Krause A (eds) Proceedings of the 35th International Conference on Machine Learning, PMLR, Stockholmsmässan, Stockholm Sweden, Proceedings of Machine Learning Research, vol 80, pp 3150–3158, URL http://proceedings.mlr.press/v80/liu18c.html

Liu LT, Simchowitz M, Hardt M (2019) The implicit fairness criterion of unconstrained learning. In: Chaudhuri K, Salakhutdinov R (eds) Proceedings of the 36th International Conference on Machine Learning, PMLR, Long Beach, California, USA, Proceedings of Machine Learning Research, vol 97, pp 4051–4060, URL http://proceedings.mlr.press/v97/liu19f.html

Liu LT, Wilson A, Haghtalab N, Kalai AT, Borgs C, 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, Association for Computing Machinery, New York, NY, USA, FAT* '20, p 381–391, DOI 10.1145/3351095.3372861, URL https://doi.org/10.1145/3351095.3372861

Liu W, Burke R (2018) Personalizing fairness-aware re-ranking. RecSys 2018 workshop: "Workshop on Responsible Recommendation (FAT/Rec)", 1809.02921

Liu Z, Luo P, Wang X, Tang X (2015) Deep learning face attributes in the wild. 1411.7766

Locatello F, Abbati G, Rainforth T, Bauer S, Schölkopf B, Bachem O (2019) On the fairness of disentangled representations. In: Wallach H, Larochelle H, Beygelzimer A, d'Alché-Buc F, Fox E, Garnett R (eds) Advances in Neural Information Processing Systems, Curran Associates, Inc., vol 32, pp 14611–14624, URL https://proceedings.neurips.cc/paper/2019/file/1b486d7a5189ebe8d8c46afc64b0d1b4-Paper.pdf

Lohaus M, Perrot M, Luxburg UV (2020) Too relaxed to be fair. In: III HD, Singh A (eds) Proceedings of the 37th International Conference on Machine Learning, PMLR, Virtual, Proceedings of Machine Learning Research, vol 119, pp 6360–6369, URL http://proceedings.mlr.press/v119/lohaus20a.html

Louizos C, Swersky K, Li Y, Welling M, Zemel RS (2016) The variational fair autoencoder. In: Bengio Y, LeCun Y (eds) 4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings, URL http://arxiv.org/abs/1511.00830

Lowe H, Ferris TA, Hernandez P, Weber S (2009) Stride - an integrated standards-based translational research informatics platform. AMIA Annual Symposium proceedings AMIA Symposium 2009:391–5

Lu Q, Getoor L (2003) Link-based classification. In: Proceedings of the Twentieth International Conference on International Conference on Machine Learning, AAAI Press, ICML'03, p 496–503

Lum K, Johndrow J (2016) A statistical framework for fair predictive algorithms. DTL 2016 workshop: "Fairness, Accountability, and Transparency in Machine Learning (FAT/ML)", 1610.08077

Lum K, Boudin C, 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, Association for Computing Machinery, New York, NY, USA, FAT* '20, p 482–491, DOI 10.1145/3351095.3372846, URL https://doi.org/10.1145/3351095.3372846

Luong BT, Ruggieri S, Turini F (2016) Classification rule mining supported by ontology for discrimination discovery. In: 2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW), pp 868–875, DOI 10.1109/ICDMW.2016.0128

Maas AL, Daly RE, Pham PT, Huang D, Ng AY, 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, Association for Computational Linguistics, Portland, Oregon, USA, pp 142–150, URL https://www.aclweb.org/anthology/P11-1015

Madnani N, Loukina A, von Davier A, Burstein J, 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, Association for Computational Linguistics, Valencia, Spain, pp 41–52, DOI 10.18653/v1/W17-1605, URL https://www.aclweb.org/anthology/W17-1605

Madras D, Creager E, Pitassi T, Zemel R (2018a) Learning adversarially fair and transferable representations. In: Dy J, Krause A (eds) Proceedings of the 35th International Conference on Machine Learning, PMLR, Stockholmsmässan, Stockholm Sweden, Proceedings of Machine Learning Research, vol 80, pp 3384–3393, URL http://proceedings.mlr.press/v80/madras18a.html

Madras D, Pitassi T, 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, Garnett R (eds) Advances in Neural Information Processing Systems, Curran Associates, Inc., vol 31, pp 6147–6157, URL https://proceedings.neurips.cc/paper/2018/file/09d37c08f7b129e96277388757530c72-Paper.pdf

Madras D, Creager E, Pitassi T, 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, Association for Computing Machinery, New York, NY, USA, FAT* '19, p 349–358, DOI 10.1145/3287560.3287564, URL https://doi.org/10.1145/3287560.3287564

Mahabadi S, Vakilian A (2020) Individual fairness for k-clustering. In: III HD, Singh A (eds) Proceedings of the 37th International Conference on Machine Learning, PMLR, Virtual, Proceedings of Machine Learning Research, vol 119, pp 6586–6596, URL http://proceedings.mlr.press/v119/mahabadi20a.html

Maity S, Xue S, Yurochkin M, Sun Y (2021) Statistical inference for individual fairness. In: International Conference on Learning Representations, URL https://openreview.net/forum?id=z9k8BWL-_2u

Mandal D, Deng S, Jana S, Wing JM, Hsu DJ (2020) Ensuring fairness beyond the training data. In: Larochelle H, Ranzato M, Hadsell R, Balcan M, Lin H (eds) Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual, URL https://proceedings.neurips.cc/paper/2020/hash/d6539d3b57159bab6a72e106beb45bd-Abstract.html

Manjunatha V, Saini N, Davis LS (2019) Explicit bias discovery in visual question answering models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)

Martinez N, Bertran M, Sapiro G (2020) Minimax pareto fairness: A multi objective perspective. In: III HD, Singh A (eds) Proceedings of the 37th International Conference on Machine Learning, PMLR, Virtual, Proceedings of Machine Learning Research, vol 119, pp 6755–6764, URL http://proceedings.mlr.press/v119/martinez20a.html

Mary J, Calauzènes C, Karoui NE (2019) Fairness-aware learning for continuous attributes and treatments. In: Chaudhuri K, Salakhutdinov R (eds) Proceedings of the 36th International Conference on Machine Learning, PMLR, Long Beach, California, USA, Proceedings of Machine Learning Research, vol 97, pp 4382–4391, URL http://proceedings.mlr.press/v97/mary19a.html

Mastrandrea R, Fournet J, 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, DOI 10.1371/journal.pone.0136497, URL http://dx.doi.org/10.1371/journal.pone.0136497

Mattei N, Saffidine A, 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)

Mattei N, Saffidine A, Walsh T (2018b) Fairness in deceased organ matching. In: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, Association for Computing Machinery, New York, NY, USA, AIES '18, p 236–242, DOI 10.1145/3278721.3278749, URL https://doi.org/10.1145/3278721.3278749

Mayson SG (2018) Bias in, bias out. YAle lJ 128:2218

McAuley J, Targett C, Shi Q, 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, Association for Computing Machinery, New York, NY, USA, SIGIR '15, p 43–52, DOI 10.1145/2766462.2767755, URL https://doi.org/10.1145/2766462.2767755

McCallum AK, Nigam K, Rennie J, Seymore K (2000) Automating the construction of internet portals with machine learning. Information Retrieval 3(2):127–163

McDuff D, Ma S, Song Y, Kapoor A (2019) Characterizing bias in classifiers using generative models. In: Wallach H, Larochelle H, Beygelzimer A, d'Alché-Buc F, Fox E, Garnett R (eds) Advances in Neural Information Processing Systems, Curran Associates, Inc., vol 32, pp 5403–5414, URL https://proceedings.neurips.cc/paper/2019/file/7f018eb7b301a66658931cb8a93fd6e8-Paper.pdf

McFee B, Bertin-Mahieux T, Ellis DP, Lanckriet GR (2012) The million song dataset challenge. In: Proceedings of the 21st International Conference on World Wide Web, Association for Computing Machinery, New York, NY, USA, WWW '12 Companion, p 909–916, DOI 10.1145/2187980.2188222, URL https://doi.org/10.1145/2187980.2188222

McKenna L (2019a) A history of the current population survey and disclosure avoidance. URL https://www2.census.gov/adrm/CED/Papers/FY20/2019-04-McKenna-cps%20and%20da.pdf

McKenna L (2019b) A history of the us census bureau's disclosure review board. URL https://www2.census.gov/adrm/CED/Papers/FY20/2019-04-McKenna-DRB.pdf

McMahan B, Moore E, Ramage D, Hampson S, y Arcas BA (2017) Communication-Efficient Learning of Deep Networks from Decentralized Data. In: Singh A, Zhu J (eds) Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, PMLR, Fort Lauderdale, FL, USA, Proceedings of Machine Learning Research, vol 54, pp 1273–1282, URL http://proceedings.mlr.press/v54/mcmahan17a.html

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, Association for Computing Machinery, New York, NY, USA, AIES '19, p 313–320, DOI 10.1145/3306618.3314290, URL https://doi.org/10.1145/3306618.3314290

Meek C, Thiesson B, Heckerman D (2002) The learning-curve sampling method applied to model-based clustering. Journal of Machine Learning Research 2(Feb):397–418

Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A (2021) A survey on bias and fairness in machine learning. ACM Comput Surv 54(6), DOI 10.1145/3457607, URL https://doi.org/10.1145/3457607

Mehrotra A, Celis LE (2021) Mitigating bias in set selection with noisy protected attributes. In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, Association for Computing Machinery, New York, NY, USA, FAccT '21, p 237–248, DOI 10.1145/3442188.3445887, URL https://doi.org/10.1145/3442188.3445887

Mehrotra R, Anderson A, Diaz F, Sharma A, Wallach H, Yilmaz E (2017) Auditing search engines for differential satisfaction across demographics. In: Proceedings of the 26th International Conference on World Wide Web Companion, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, WWW '17 Companion, p 626–633, DOI 10.1145/3041021.3054197, URL https://doi.org/10.1145/3041021.3054197

Merkley R (2019) Use and fair use: Statement on shared images in facial recognition ai

Merler M, Ratha N, Feris RS, Smith JR (2019) Diversity in faces. 1901.10436

Metevier B, Giguere S, Brockman S, Kobren A, Brun Y, Brunskill E, Thomas PS (2019) Offline contextual bandits with high probability fairness guarantees. In: Wallach H, Larochelle H, Beygelzimer A, d'Alché-Buc F, Fox E, Garnett R (eds) Advances in Neural Information Processing Systems, Curran Associates, Inc., vol 32, pp 14922–14933, URL https://proceedings.neurips.cc/paper/2019/file/d69768b3da745b77e82cdbddcc8bac98-Paper.pdf

Mhasawade V, Chunara R (2021) Causal Multi-Level Fairness, Association for Computing Machinery, New York, NY, USA, p 784–794. URL https://doi.org/10.1145/3461702.3462587

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)

Miceli M, Yang T, Naudts L, Schuessler M, Serbanescu D, 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, Association for Computing Machinery, New York, NY, USA, FAccT '21, p 161–172

Miller E (1998) An introduction to the resource description framework. D-lib Magazine

Mirkin S, Nowson S, Brun C, Perez J (2015) Motivating personality-aware machine translation. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Lisbon, Portugal, pp 1102–1108, DOI 10.18653/v1/D15-1130, URL https://www.aclweb.org/anthology/D15-1130

Mishler A, Kennedy EH, 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, Association for Computing Machinery, New York, NY, USA, FAccT '21, p 386–400, DOI 10.1145/3442188.3445902, URL https://doi.org/10.1145/3442188.3445902

Mishra S, He S, Belli L (2020) Assessing demographic bias in named entity recognition. AKBC 2020 workshop: "Bias in Automatic Knowledge Graph Construction", 2008.03415

Mislove A, Viswanath B, Gummadi KP, 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, Association for Computing Machinery, New York, NY, USA, WSDM '10, p 251–260, DOI 10.1145/1718487.1718519, URL https://doi.org/10.1145/1718487.1718519

Moore JC, Stinson LL, Welniak EJ (2000) Income measurement error in surveys: A review. Journal of Official Statistics-Stockholm- 16(4):331–362

Moreland A, Herlihy C, Tynan MA, Sunshine G, McCord RF, Hilton C, Poovey J, Werner AK, Jones CD, Fulmer EB, 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

Moro S, Cortez P, Rita P (2014) A data-driven approach to predict the success of bank telemarketing. Decis Support Syst 62:22–31

Mozannar H, Ohannessian M, Srebro N (2020) Fair learning with private demographic data. In: III HD, Singh A (eds) Proceedings of the 37th International Conference on Machine Learning, PMLR, Virtual, Proceedings of Machine Learning Research, vol 119, pp 7066–7075, URL http://proceedings.mlr.press/v119/mozannar20a.html

Mukherjee D, Yurochkin M, Banerjee M, Sun Y (2020) Two simple ways to learn individual fairness metrics from data. In: III HD, Singh A (eds) Proceedings of the 37th International Conference on Machine Learning, PMLR, Virtual, Proceedings of Machine Learning Research, vol 119, pp 7097–7107, URL http://proceedings.mlr.press/v119/mukherjee20a.html

Muller M, Lange I, Wang D, Piorkowski D, Tsay J, Liao QV, Dugan C, Erickson T (2019) How Data Science Workers Work with Data: Discovery, Capture, Curation, Design, Creation, Association for Computing Machinery, New York, NY, USA, p 1–15. URL https://doi.org/10.1145/3290605.3300356

Murgia M (2019) Microsoft quietly deletes largest public face recognition data set. URL https://www.ft.com/content/7d3e0d6a-87a0-11e9-a028-86cea8523dc2

Nabi R, Malinsky D, Shpitser I (2019) Learning optimal fair policies. In: Chaudhuri K, Salakhutdinov R (eds) Proceedings of the 36th International Conference on Machine Learning, PMLR, Long Beach, California, USA, Proceedings of Machine Learning Research, vol 97, pp 4674–4682, URL http://proceedings.mlr.press/v97/nabi19a.html

Namata G, London B, Getoor L, Huang B, EDU U (2012) Query-driven active surveying for collective classification. In: 10th International Workshop on Mining and Learning with Graphs, vol 8

Nanda V, Xu P, Sankararaman KA, Dickerson JP, 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, Association for Computing Machinery, New York, NY, USA, AIES '20, p 131, DOI 10.1145/3375627.3375818, URL https://doi.org/10.1145/3375627.3375818

Nanda V, Dooley S, Singla S, Feizi S, Dickerson JP (2021) Fairness through robustness: Investigating robustness disparity in deep learning. In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, Association for Computing Machinery, New York, NY, USA, FAccT '21, p 466–477, DOI 10.1145/3442188.3445910, URL https://doi.org/10.1145/3442188.3445910

Narayanan A, Shmatikov V (2008) Robust de-anonymization of large sparse datasets. In: 2008 IEEE Symposium on Security and Privacy (sp 2008), IEEE, pp 111–125

Nasr M, Tschantz MC (2020) Bidding strategies with gender nondiscrimination constraints for online ad auctions. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, Association for Computing Machinery, New York, NY, USA, FAT* '20, p 337–347, DOI 10.1145/3351095.3375783, URL https://doi.org/10.1145/3351095.3375783

Ngong IC, Maughan K, Near JP (2020) Towards auditability for fairness in deep learning. NeurIPS 2020 workshop: "Algorithmic Fairness through the Lens of Causality and Interpretability (AFCI)", 2012.00106

NLST Trial Research Team (2011) The national lung screening trial: overview and study design. Radiology 258(1):243–253

Noriega-Campero A, Bakker MA, Garcia-Bulle B, Pentland AS (2019) Active fairness in algorithmic decision making. In: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, Association for Computing Machinery, New York, NY, USA, AIES '19, p 77–83, DOI 10.1145/3306618.3314277, URL https://doi.org/10.1145/3306618.3314277

Noriega-Campero A, Garcia-Bulle B, Cantu LF, Bakker MA, Tejerina L, Pentland A (2020) Algorithmic targeting of social policies: Fairness, accuracy, and distributed governance. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, Association for Computing Machinery, New York, NY, USA, FAT* '20, p 241–251, DOI 10.1145/3351095.3375784, URL https://dl.acm.org/doi/pdf/10.1145/3351095.3375784

Nuttall DL, Goldstein H, Prosser R, Rasbash J (1989) Differential school effectiveness. International Journal of Educational Research 13(7):769–776, DOI https://doi.org/10.1016/0883-0355(89)90027-X, URL https://www.sciencedirect.com/science/article/pii/088303558990027X

Obermeyer Z, 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, Association for Computing Machinery, New York, NY, USA, FAT* '19, p 89, DOI 10.1145/3287560.3287593, URL https://doi.org/10.1145/3287560.3287593

Ogura H, Takeda A (2020) Convex fairness constrained model using causal effect estimators. In: Companion Proceedings of the Web Conference 2020, Association for Computing Machinery, New York, NY, USA, WWW '20, p 723–732, DOI 10.1145/3366424.3383556, URL https://doi.org/10.1145/3366424.3383556

Olave M, Rajkovic V, Bohanec M (1989) An application for admission in public school systems. Expert Systems in Public Administration 1:145–160

Oneto L, Siri A, Luria G, Anguita D (2017) Dropout prediction at university of genoa: a privacy preserving data driven approach. In: ESANN

Oneto L, Donini M, Elders A, Pontil M (2019a) Taking advantage of multitask learning for fair classification. In: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, Association for Computing Machinery, New York, NY, USA, AIES '19, p 227–237, DOI 10.1145/3306618.3314255, URL https://doi.org/10.1145/3306618.3314255

Oneto L, Donini M, Maurer A, Pontil M (2019b) Learning fair and transferable representations. NeurIPS 2019 workshop: “Human-Centric Machine Learning”

Oneto L, Donini M, Luise G, Ciliberto C, Maurer A, Pontil M (2020) Exploiting MMD and sinkhorn divergences for fair and transferable representation learning. In: Larochelle H, Ranzato M, Hadsell R, Balcan M, Lin H (eds) Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual, URL https://proceedings.neurips.cc/paper/2020/hash/af9c0e0c1dee63e5cad8b7ed1a5be96-Abstract.html

Pandey A, Caliskan A (2021) Disparate Impact of Artificial Intelligence Bias in Ridehailing Economy’s Price Discrimination Algorithms, Association for Computing Machinery, New York, NY, USA, p 822–833. URL https://doi.org/10.1145/3461702.3462561

Papakyriakopoulos O, Hegelich S, Serrano JCM, Marco F (2020) Bias in word embeddings. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, Association for Computing Machinery, New York, NY, USA, FAT* ’20, p 446–457, DOI 10.1145/3351095.3372843, URL https://dl.acm.org/doi/pdf/10.1145/3351095.3372843

Paraschakis D, 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)”

Partnership on AI (2022) About ML. Tech. rep., URL https://partnershiponai.org/workstream/about-ml/

Patro GK, Chakraborty A, Ganguly N, Gummadi KP (2019) Incremental fairness in two-sided market platforms: On smoothly updating recommendations. NeurIPS 2019 workshop: “Human-Centric Machine Learning”, 1909.10005

Paullada A, Raji ID, Bender EM, Denton E, Hanna A (2020) Data and its (dis) contents: A survey of dataset development and use in machine learning research. arXiv preprint arXiv:201205345

Pedreshi D, Ruggieri S, Turini F (2008) Discrimination-aware data mining. In: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Association for Computing Machinery, New York, NY, USA, KDD ’08, p 560–568, DOI 10.1145/1401890.1401959, URL https://doi.org/10.1145/1401890.1401959

Peng K, Mathur A, Narayanan A (2021) Mitigating dataset harms requires stewardship: Lessons from 1000 papers. arXiv preprint arXiv:210802922

Perrone V, Donini M, Zafar MB, Schmucker R, Kenthapadi K, Archambeau C (2021) Fair Bayesian Optimization, Association for Computing Machinery, New York, NY, USA, p 854–863. URL https://doi.org/10.1145/3461702.3462629

Pessach D, Shmueli E (2020) Algorithmic fairness. arXiv preprint arXiv:200109784

Peters M, Neumann M, Iyyer M, Gardner M, Clark C, Lee K, 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), Association for Computational Linguistics, New Orleans, Louisiana, pp 2227–2237

Peters ME, Lecocq D (2013) Content extraction using diverse feature sets. In: Proceedings of the 22Nd International Conference on World Wide Web, pp 89–90

Pfohl S, Marafino B, Coulet A, Rodriguez F, Palaniappan L, Shah NH (2019) Creating fair models of atherosclerotic cardiovascular disease risk. In: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, Association for Computing Machinery, New York, NY, USA, AIES '19, p 271–278, DOI 10.1145/3306618.3314278, URL https://doi.org/10.1145/3306618.3314278

Pinard M (2010) Collateral consequences of criminal convictions: Confronting issues of race and dignity. NYUL Rev 85:457

Pitoura E, Stefanidis K, Koutrika G (2021) Fairness in rankings and recommendations: An overview. The VLDB Journal pp 1–28

Pleiss G, Raghavan M, Wu F, Kleinberg J, Weinberger KQ (2017) On fairness and calibration. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, Garnett R (eds) Advances in Neural Information Processing Systems, Curran Associates, Inc., vol 30, pp 5680–5689, URL https://proceedings.neurips.cc/paper/2017/file/b8b9c74ac526fffbeb2d39ab038d1cd7-Paper.pdf

Pont-Tuset J, Uijlings J, Changpinyo S, Soricut R, Ferrari V (2020) Connecting vision and language with localized narratives. In: European Conference on Computer Vision, Springer, pp 647–664

Prabhu VU, Birhane A (2020) Large image datasets: A pyrrhic win for computer vision? arXiv preprint arXiv:200616923

Preoţiuc-Pietro D, Ungar L (2018) User-level race and ethnicity predictors from Twitter text. In: Proceedings of the 27th International Conference on Computational Linguistics, Association for Computational Linguistics, Santa Fe, New Mexico, USA, pp 1534–1545, URL https://www.aclweb.org/anthology/C18-1130

ProPublica (2016) Compas analysis github repository. URL https://github.com/propublica/compas-analysis

ProPublica (2021) Propublica data store terms. URL https://www.propublica.org/datastore/terms

Pujol D, McKenna R, Kuppam S, Hay M, Machanavajjhala A, Miklau G (2020) Fair decision making using privacy-protected data. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, Association for Computing Machinery, New York, NY, USA, FAT* '20, p 189–199, DOI 10.1145/3351095.3372872, URL https://doi.org/10.1145/3351095.3372872

Qian S, Cao J, Mouël FL, Sahel I, 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, Association for Computing Machinery, New York, NY, USA, KDD '15, p 955–964, DOI 10.1145/2783258.2783261, URL https://doi.org/10.1145/2783258.2783261

Qin T, Liu TY (2013) Introducing letor 4.0 datasets. 1306.2597

Quadrianto N, Sharmanska V (2017) Recycling privileged learning and distribution matching for fairness. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, Garnett R (eds) Advances in Neural Information Processing Systems, Curran Associates, Inc., vol 30, pp 677–688, URL https://proceedings.neurips.cc/paper/2017/file/250cf8b51c773f3f8dc8b4be867a9a02-Paper.pdf

Quadrianto N, Sharmanska V, Thomas O (2019) Discovering fair representations in the data domain. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)

Radford A, Narasimhan K, Salimans T, Sutskever I (2018) Improving language understanding by generative pre-training

Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I (2019) Language models are unsupervised multitask learners

Radin J (2017) "digital natives": How medical and indigenous histories matter for big data. Osiris 32(1):43–64

Raff E, Sylvester J (2018) Gradient reversal against discrimination. ICML 2018 workshop: "Fairness, Accountability, and Transparency in Machine Learning (FAT/ML)", 1807.00392

Raff E, Sylvester J, 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, Association for Computing Machinery, New York, NY, USA, AIES '18, p 243–250, DOI 10.1145/3278721.3278742, URL https://doi.org/10.1145/3278721.3278742

Rahmattalabi A, Vayanos P, Fulginiti A, Rice E, Wilder B, Yadav A, Tambe M (2019) Exploring algorithmic fairness in robust graph covering problems. In: Wallach H, Larochelle H, Beygelzimer A, d'Alché-Buc F, Fox E, Garnett R (eds) Advances in Neural Information Processing Systems, Curran Associates, Inc., vol 32, pp 15776–15787, URL https://proceedings.neurips.cc/paper/2019/file/1d7c2aae840867027b7edd17b6aaa0e9-Paper.pdf

Raj A, Wood C, Montoly A, Ekstrand MD (2020) Comparing fair ranking metrics. RecSys 2020 workshop: "3rd FAccTRec Workshop on Responsible Recommendation", 2009.01311

Raji ID, 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, Association for Computing Machinery, New York, NY, USA, AIES '19, p 429–435, DOI 10.1145/3306618.3314244, URL https://doi.org/10.1145/3306618.3314244

Ramachandran GS, Brugere I, Varshney LR, Xiong C (2021) GAEA: Graph Augmentation for Equitable Access via Reinforcement Learning, Association for Computing Machinery, New York, NY, USA, p 884–894. URL https://doi.org/10.1145/3461702.3462615

Ramaswamy VV, Kim SSY, 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), pp 9301–9310

Red V, Kelsic ED, Mucha PJ, Porter MA (2011) Comparing community structure to characteristics in online collegiate social networks. SIAM review 53(3):526–543

Redmond M, Baveja A (2002) A data-driven software tool for enabling cooperative information sharing among police departments. European Journal of Operational Research 141:660–678, DOI 10.1016/S0377-2217(01)00264-8

Redmond U, Cunningham P (2013) A temporal network analysis reveals the unprofitability of arbitrage in the prosper marketplace. Expert Systems with Applications 40(9):3715–3721, DOI https://doi.org/10.1016/j.eswa.2012.12.077, URL https://www.sciencedirect.com/science/article/pii/S0957417412013188

Reed SE, Zhang Y, Zhang Y, Lee H (2015) Deep visual analogy-making. In: Cortes C, Lawrence N, Lee D, Sugiyama M, Garnett R (eds) Advances in Neural Information Processing Systems, Curran Associates, Inc., vol 28, pp 1252–1260, URL https://proceedings.neurips.cc/paper/2015/file/e07413354875be01a996dc560274708e-Paper.pdf

Rezaei A, Liu A, Memarrast O, Ziebart B (2021) Robust fairness under covariate shift. NeurIPS 2020 workshop: “Algorithmic Fairness through the Lens of Causality and Interpretability (AFCI)”, 2010.05166

Riederer C, Chaintreau A (2017) The price of fairness in location based advertising. DOI 10.18122/B2MD8C, URL https://doi.org/10.18122/B2MD8C, recSys 2017 workshop: “Workshop on Responsible Recommendation (FAT/Rec)”

Rocher L, Hendrickx JM, De Montjoye YA (2019) Estimating the success of re-identifications in incomplete datasets using generative models. Nature communications 10(1):1–9

Rodolfa KT, Salomon E, Haynes L, Mendieta IH, Larson J, 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, Association for Computing Machinery, New York, NY, USA, FAT* ’20, p 142–153, DOI 10.1145/3351095.3372863, URL https://doi.org/10.1145/3351095.3372863

Roh Y, Lee K, Whang S, Suh C (2020) FR-train: A mutual information-based approach to fair and robust training. In: III HD, Singh A (eds) Proceedings of the 37th International Conference on Machine Learning, PMLR, Virtual, Proceedings of Machine Learning Research, vol 119, pp 8147–8157, URL http://proceedings.mlr.press/v119/roh20a.html

Roh Y, Lee K, Whang SE, Suh C (2021) Fairbatch: Batch selection for model fairness. In: International Conference on Learning Representations, URL https://openreview.net/forum?id=YNnpaAKeCfx

Romano Y, Bates S, Candes E (2020) Achieving equalized odds by resampling sensitive attributes. In: Larochelle H, Ranzato M, Hadsell R, Balcan MF, Lin H (eds) Advances in Neural Information Processing Systems, Curran Associates, Inc., vol 33, pp 361–371, URL https://proceedings.neurips.cc/paper/2020/file/03593ce517feac573fdaafa6dcedef61-Paper.pdf

Romei A, Ruggieri S (2014) A multidisciplinary survey on discrimination analysis. The Knowledge Engineering Review 29(5):582–638, DOI 10.1017/S0269888913000039

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

Rozemberczki B, Allen C, Sarkar R (2021) Multi-scale attributed node embedding. Journal of Complex Networks 9(2):cnab014

Rudinger R, May C, Van Durme B (2017) Social bias in elicited natural language inferences. In: Proceedings of the First ACL Workshop on Ethics in Natural Language Processing, Association for Computational Linguistics, Valencia, Spain, pp 74–79, DOI 10.18653/v1/W17-1609, URL https://www.aclweb.org/anthology/W17-1609

Rudinger R, Naradowsky J, Leonard B, 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), Association for Computational Linguistics, New Orleans, Louisiana, pp 8–14, DOI 10.18653/v1/N18-2002, URL https://www.aclweb.org/anthology/N18-2002

Ruoss A, Balunovic M, Fischer M, Vechev M (2020) Learning certified individually fair representations. In: Larochelle H, Ranzato M, Hadsell R, Balcan MF, Lin H (eds) Advances in Neural Information Processing Systems, Curran Associates, Inc., vol 33, pp 7584–7596, URL https://proceedings.neurips.cc/paper/2020/file/55d491cf951b1b920900684d71419282-Paper.pdf

Russell C, Kusner MJ, Loftus J, Silva R (2017) When worlds collide: Integrating different counterfactual assumptions in fairness. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, Garnett R (eds) Advances in Neural Information Processing Systems, Curran Associates, Inc., vol 30, pp 6414–6423, URL https://proceedings.neurips.cc/paper/2017/file/1271a7029c9df08643b631b02cf9e116-Paper.pdf

Sabato S, Yom-Tov E (2020) Bounding the fairness and accuracy of classifiers from population statistics. In: III HD, Singh A (eds) Proceedings of the 37th International Conference on Machine Learning, PMLR, Virtual, Proceedings of Machine Learning Research, vol 119, pp 8316–8325, URL http://proceedings.mlr.press/v119/sabato20a.html

Saenko K, Kulis B, Fritz M, Darrell T (2010) Adapting visual category models to new domains. In: Proceedings of the 11th European Conference on Computer Vision: Part IV, Springer-Verlag, Berlin, Heidelberg, ECCV'10,

p 213–226
Sagawa S, Koh PW, Hashimoto TB, Liang P (2020) Distributionally robust neural networks. In: International Conference on Learning Representations, URL https://openreview.net/forum?id=ryxGuJrFvS
Samadi S, Tantipongpipat U, Morgenstern JH, Singh M, Vempala S (2018) The price of fair pca: One extra dimension. In: Bengio S, Wallach H, Larochelle H, Grauman K, Cesa-Bianchi N, Garnett R (eds) Advances in Neural Information Processing Systems, Curran Associates, Inc., vol 31, pp 10976–10987, URL https://proceedings.neurips.cc/paper/2018/file/cc4af25fa9d2d5c953496579b75f6f6c-Paper.pdf
Savani Y, White C, Govindarajulu NS (2020) Intra-processing methods for debiasing neural networks. In: Larochelle H, Ranzato M, Hadsell R, Balcan M, Lin H (eds) Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual, URL https://proceedings.neurips.cc/paper/2020/hash/1d8d70dddf147d2d92a634817f01b239-Abstract.html
Scheuerman MK, Wade K, Lustig C, Brubaker JR (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), DOI 10.1145/3392866, URL https://doi.org/10.1145/3392866
Scheuerman MK, Hanna A, 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
Schumann C, Ricco S, Prabhu U, Ferrari V, Pantofaru C (2021) A Step Toward More Inclusive People Annotations for Fairness, Association for Computing Machinery, New York, NY, USA, p 916–925. URL https://doi.org/10.1145/3461702.3462594
Schutzman Z (2020) Trade-offs in fair redistricting. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, Association for Computing Machinery, New York, NY, USA, AIES '20, p 159–165, DOI 10.1145/3375627.3375802, URL https://doi.org/10.1145/3375627.3375802
Segal S, Adi Y, Pinkas B, Baum C, Ganesh C, Keshet J (2021) Fairness in the Eyes of the Data: Certifying Machine-Learning Models, Association for Computing Machinery, New York, NY, USA, p 926–935. URL https://doi.org/10.1145/3461702.3462554
Sen P, Namata G, Bilgic M, Getoor L, Galligher B, Eliassi-Rad T (2008) Collective classification in network data. AI magazine 29(3):93–93
Shah K, Gupta P, Deshpande A, Bhattacharyya C (2021) Rawlsian Fair Adaptation of Deep Learning Classifiers, Association for Computing Machinery, New York, NY, USA, p 936–945. URL https://doi.org/10.1145/3461702.3462592
Shang J, Sun M, Lam NS (2020) List-Wise Fairness Criterion for Point Processes, Association for Computing Machinery, New York, NY, USA, p 1948–1958. URL https://doi.org/10.1145/3394486.3403246
Sharifi-Malvajerdi S, Kearns M, Roth A (2019) Average individual fairness: Algorithms, generalization and experiments. In: Wallach H, Larochelle H,

Beygelzimer A, d'Alché-Buc F, Fox E, Garnett R (eds) Advances in Neural Information Processing Systems, Curran Associates, Inc., vol 32, pp 8242–8251, URL https://proceedings.neurips.cc/paper/2019/file/0e1feae55e360ff05fef58199b3fa521-Paper.pdf

Sharma S, Henderson J, 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, Association for Computing Machinery, New York, NY, USA, AIES '20, p 166–172, DOI 10.1145/3375627.3375812, URL https://doi.org/10.1145/3375627.3375812

Sharma S, Zhang Y, Ríos Aliaga JM, Bouneffouf D, Muthusamy V, Varshney KR (2020b) Data augmentation for discrimination prevention and bias disambiguation. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, Association for Computing Machinery, New York, NY, USA, AIES '20, p 358–364, DOI 10.1145/3375627.3375865, URL https://doi.org/10.1145/3375627.3375865

Sharma S, Gee AH, Paydarfar D, Ghosh J (2021) FaiR-N: Fair and Robust Neural Networks for Structured Data, Association for Computing Machinery, New York, NY, USA, p 946–955. URL https://doi.org/10.1145/3461702.3462559

Shekhar S, Shah N, Akoglu L (2021) Fairod: Fairness-aware outlier detection. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, Association for Computing Machinery, New York, NY, USA, AIES '21, p 210–220, DOI 10.1145/3461702.3462517, URL https://doi.org/10.1145/3461702.3462517

Shen JH, Fratamico L, Rahwan I, Rush AM (2018) Darling or babygirl? investigating stylistic bias in sentiment analysis. KDD 2018 workshop: "Fairness, Accountability, and Transparency in Machine Learning (FAT/ML)"

Shermis MD (2014) State-of-the-art automated essay scoring: Competition, results, and future directions from a united states demonstration. Assessing Writing 20:53–76, DOI https://doi.org/10.1016/j.asw.2013.04.001, URL https://www.sciencedirect.com/science/article/pii/S1075293513000196

Singh A, Joachims T (2018) Fairness of exposure in rankings. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Association for Computing Machinery, New York, NY, USA, KDD '18, p 2219–2228, DOI 10.1145/3219819.3220088, URL https://doi.org/10.1145/3219819.3220088

Singh A, Joachims T (2019) Policy learning for fairness in ranking. In: Wallach H, Larochelle H, Beygelzimer A, d'Alché-Buc F, Fox E, Garnett R (eds) Advances in Neural Information Processing Systems, Curran Associates, Inc., vol 32, pp 5426–5436, URL https://proceedings.neurips.cc/paper/2019/file/9e82757e9a1c12cb710ad680db11f6f1-Paper.pdf

Singh H, Singh R, Mhasawade V, Chunara R (2021) Fairness violations and mitigation under covariate shift. In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, Association for Computing Machinery, New York, NY, USA, FAccT '21, p 3–13, DOI 10.1145/3442188.3445865, URL https://doi.org/10.1145/3442188.3445865

Singh M, Ramamurthy KN (2019) Understanding racial bias in health using the medical expenditure panel survey data. NeurIPS 2019 workshop: "Fair ML for Health", 1911.01509

Slack D, Friedler S, Givental E (2019a) Fair meta-learning: Learning how to learn fairly. URL https://drive.google.com/file/d/1F5YF1Ar1hJ7l2H7zIsC35SzXOWqUylVW/view, neurIPS 2019 workshop: "Human-Centric Machine Learning"

Slack D, Friedler S, Givental E (2019b) Fairness warnings. URL https://drive.google.com/file/d/1eeu703ulWkehk0WEepYDwXg2KXSwOzc2/view, neurIPS 2019 workshop: "Human-Centric Machine Learning"

Slack D, Friedler SA, Givental E (2020) Fairness warnings and fair-maml: Learning fairly with minimal data. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, Association for Computing Machinery, New York, NY, USA, FAT* '20, p 200–209, DOI 10.1145/3351095.3372839, URL https://doi.org/10.1145/3351095.3372839

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

Smith JW, Everhart J, Dickson W, Knowler W, Johannes R (1988) Using the adap learning algorithm to forecast the onset of diabetes mellitus. Proceedings Symposium on Computer Applications in Medical Care p 261—265, URL https://europepmc.org/articles/PMC2245318

Solans D, Fabbri F, Calsamiglia C, Castillo C, Bonchi F (2021) Comparing Equity and Effectiveness of Different Algorithms in an Application for the Room Rental Market, Association for Computing Machinery, New York, NY, USA, p 978–988. URL https://doi.org/10.1145/3461702.3462600

Sonboli N, Burke R (2019) Localized fairness in recommender systems. In: Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization, Association for Computing Machinery, New York, NY, USA, UMAP'19 Adjunct, p 295–300, DOI 10.1145/3314183.3323845, URL https://doi.org/10.1145/3314183.3323845

Sonboli N, Burke R, Mattei N, Eskandanian F, Gao T (2020) "and the winner is...": Dynamic lotteries for multi-group fairness-aware recommendation. RecSys 2020 workshop: "3rd FAccTRec Workshop on Responsible Recommendation", 2009.02590

Speakman S, Sridharan S, 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, Association for Computing Machinery, New York, NY, USA, COMPASS '18

Speicher T, Ali M, Venkatadri G, Ribeiro FN, Arvanitakis G, Benevenuto F, Gummadi KP, Loiseau P, Mislove A (2018a) Potential for discrimination in online targeted advertising. In: Friedler SA, Wilson C (eds) Proceedings of the 1st Conference on Fairness, Accountability and Transparency, PMLR, New York, NY, USA, Proceedings of Machine Learning Research, vol 81,

pp 5–19, URL http://proceedings.mlr.press/v81/speicher18a.html
Speicher T, Heidari H, Grgic-Hlaca N, Gummadi KP, Singla A, Weller A, Zafar MB (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, Association for Computing Machinery, New York, NY, USA, KDD '18, p 2239–2248, DOI 10.1145/3219819.3220046, URL https://doi.org/10.1145/3219819.3220046
Squire RF (2019) Measuring and correcting sampling bias in safegraph patterns for more accurate demographic analysis. URL https://www.safegraph.com/blog/measuring-and-correcting-sampling-bias-for-accurate-demographic-analysis/?utm_source=content&utm_medium=referral&utm_campaign=colabnotebook&utm_content=panel_bias
Stanovsky G, Smith NA, Zettlemoyer L (2019) Evaluating gender bias in machine translation. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, pp 1679–1684, DOI 10.18653/v1/P19-1164, URL https://aclanthology.org/P19-1164
Steed R, 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, Association for Computing Machinery, New York, NY, USA, FAccT '21, p 701–713, DOI 10.1145/3442188.3445932, URL https://doi.org/10.1145/3442188.3445932
Strack B, Deshazo J, Gennings C, Olmo Ortiz JL, Ventura S, Cios K, Clore J (2014) Impact of hba1c measurement on hospital readmission rates: Analysis of 70,000 clinical database patient records. BioMed research international 2014:781670, DOI 10.1155/2014/781670
Sühr T, Biega AJ, Zehlike M, Gummadi KP, 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, Association for Computing Machinery, New York, NY, USA, KDD '19, p 3082–3092, DOI 10.1145/3292500.3330793, URL https://doi.org/10.1145/3292500.3330793
Sühr T, Hilgard S, Lakkaraju H (2021) Does Fair Ranking Improve Minority Outcomes? Understanding the Interplay of Human and Algorithmic Biases in Online Hiring, Association for Computing Machinery, New York, NY, USA, p 989–999. URL https://doi.org/10.1145/3461702.3462602
Sun T, Gaut A, Tang S, Huang Y, ElSherief M, Zhao J, Mirza D, Belding E, Chang KW, Wang WY (2019) Mitigating gender bias in natural language processing: Literature review. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, pp 1630–1640, DOI 10.18653/v1/P19-1159, URL https://aclanthology.org/P19-1159

Sun Y, Han J, Gao J, Yu Y (2009) itopicmodel: Information network-integrated topic modeling. In: 2009 Ninth IEEE International Conference on Data Mining, pp 493–502, DOI 10.1109/ICDM.2009.43

Swinger N, De-Arteaga M, Heffernan IV NT, Leiserson MD, Kalai AT (2019) What are the biases in my word embedding? In: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, Association for Computing Machinery, New York, NY, USA, AIES '19, p 305–311, DOI 10.1145/3306618.3314270, URL https://doi.org/10.1145/3306618.3314270

Takac L, Zabovsky M (2012) Data analysis in public social networks. In: International scientific conference and international workshop present day trends of innovations, vol 1

Tan YC, Celis LE (2019) Assessing social and intersectional biases in contextualized word representations. In: Wallach H, Larochelle H, Beygelzimer A, d'Alché-Buc F, Fox E, Garnett R (eds) Advances in Neural Information Processing Systems, Curran Associates, Inc., vol 32, pp 13230–13241, URL https://proceedings.neurips.cc/paper/2019/file/201d546992726352471cfea6b0df0a48-Paper.pdf

Tang J, Zhang J, Yao L, Li J, Zhang l, Su Z (2008) Arnetminer: Extraction and mining of academic social networks. pp 990–998, DOI 10.1145/1401890.1402008

Tantipongpipat U, Samadi S, Singh M, Morgenstern JH, Vempala S (2019) Multi-criteria dimensionality reduction with applications to fairness. In: Wallach H, Larochelle H, Beygelzimer A, d'Alché-Buc F, Fox E, Garnett R (eds) Advances in Neural Information Processing Systems, Curran Associates, Inc., vol 32, pp 15161–15171, URL https://proceedings.neurips.cc/paper/2019/file/2201611d7a08ffda97e3e8c6b667a1bc-Paper.pdf

Taskesen B, Blanchet J, Kuhn D, Nguyen VA (2021) A statistical test for probabilistic fairness. In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, Association for Computing Machinery, New York, NY, USA, FAccT '21, p 648–665, DOI 10.1145/3442188.3445927, URL https://doi.org/10.1145/3442188.3445927

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, Association for Computational Linguistics, Valencia, Spain, pp 53–59, DOI 10.18653/v1/W17-1606, URL https://www.aclweb.org/anthology/W17-1606

Tavallaee M, Bagheri E, Lu W, Ghorbani AA (2009) A detailed analysis of the kdd cup 99 data set. In: 2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications, pp 1–6, DOI 10.1109/CISDA.2009.5356528

Team Conduent Public Safety Solutions (2018) Real time crime forecasting challenge: Post-mortem analysis challenge performance

Tjong Kim Sang EF, 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, pp 142–147, URL https://www.aclweb.org/anthology/

W03-0419
Tong S, Kagal L (2020) Investigating bias in image classification using model explanations. ICML 2020 workshop: “Workshop on Human Interpretability in Machine Learning (WHI)”, 2012.05463
Toutanova K, Chen D, Pantel P, Poon H, Choudhury P, Gamon M (2015) Representing text for joint embedding of text and knowledge bases. DOI 10.18653/v1/D15-1174
Tsang A, Wilder B, Rice E, Tambe M, Zick Y (2019) Group-fairness in influence maximization. In: International Joint Conference on Artificial Intelligence
Tsao CW, Vasan RS (2015) Cohort profile: The framingham heart study (fhs): overview of milestones in cardiovascular epidemiology. International journal of epidemiology 44(6):1800–1813
Tschandl P, Rosendahl C, 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
Tziavelis N, Giannakopoulos I, Doka K, Koziris N, Karras P (2019) Equitable stable matchings in quadratic time. In: Wallach H, Larochelle H, Beygelzimer A, d'Alché-Buc F, Fox E, Garnett R (eds) Advances in Neural Information Processing Systems, Curran Associates, Inc., vol 32, pp 457–467, URL https://proceedings.neurips.cc/paper/2019/file/cb70ab375662576bd1ac5aaf16b3fca4-Paper.pdf
UCI Machine Learning Repository (1994) Statlog (german credit data) data set. URL https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)
UCI Machine Learning Repository (1996) Adult data set. URL https://archive.ics.uci.edu/ml/datasets/adult
UCI Machine Learning Repository (2019) South german credit data set. URL https://archive.ics.uci.edu/ml/datasets/South+German+Credit
US Dept of Commerce Bureau of the Census (1978) The current population survey: Design and methodology
US Dept of Commerce Bureau of the Census (1995) Current population survey: Annual demographic file, 1994
US Federal Reserve (2007) Report to the congress on credit scoring and its effects on the availability and affordability of credi
Ustun B, Westover MB, Rudin C, Bianchi MT (2016) Clinical prediction models for sleep apnea: The importance of medical history over symptoms. Journal of Clinical Sleep Medicine 12(02):161–168, DOI 10.5664/jcsm.5476, URL https://jcsm.aasm.org/doi/abs/10.5664/jcsm.5476, https://jcsm.aasm.org/doi/pdf/10.5664/jcsm.5476
Ustun B, Liu Y, Parkes D (2019) Fairness without harm: Decoupled classifiers with preference guarantees. In: Chaudhuri K, Salakhutdinov R (eds) Proceedings of the 36th International Conference on Machine Learning, PMLR, Long Beach, California, USA, Proceedings of Machine Learning Research, vol 97, pp 6373–6382, URL http://proceedings.mlr.press/v97/ustun19a.html

V E S, 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, DOI 10.1080/22797254.2020.1725789, URL https://doi.org/10.1080/22797254.2020.1725789, https://doi.org/10.1080/22797254.2020.1725789
V E S, Park J, Cho Y (2020) Using data mining techniques for bike sharing demand prediction in metropolitan city. Computer Communications 153:353–366, DOI https://doi.org/10.1016/j.comcom.2020.02.007, URL https://www.sciencedirect.com/science/article/pii/S0140366419318997
Vaithianathan R, Putnam-Hornstein E, Jiang N, Nand P, Maloney T (2017) Developing predictive models to support child maltreatment hotline screening decisions: Allegheny county methodology and implementation. URL https://www.alleghenycountyanalytics.us/wp-content/uploads/2019/05/16-ACDHS-26_PredictiveRisk_Package_050119_FINAL-2.pdf
Valera I, Singla A, 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, Garnett R (eds) Advances in Neural Information Processing Systems, Curran Associates, Inc., vol 31, pp 1769–1778, URL https://proceedings.neurips.cc/paper/2018/file/0a113ef6b61820daa5611c870ed8d5ee-Paper.pdf
Van Horn G, Mac Aodha O, Song Y, Cui Y, Sun C, Shepard A, Adam H, Perona P, Belongie S (2018) The inaturalist species classification and detection dataset. 1707.06642
Van Horn G, Cole E, Beery S, Wilber K, Belongie S, 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, pp 12884–12893
Vargo A, Zhang F, Yurochkin M, Sun Y (2021) Individually fair gradient boosting. In: International Conference on Learning Representations, URL https://openreview.net/forum?id=JBAa9we1AL
Vig J, Gehrmann S, Belinkov Y, Qian S, Nevo D, Singer Y, Shieber SM (2020) Investigating gender bias in language models using causal mediation analysis. In: Larochelle H, Ranzato M, Hadsell R, Balcan M, Lin H (eds) Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual, URL https://proceedings.neurips.cc/paper/2020/hash/92650b2e92217715fe312e6fa7b90d82-Abstract.html
Vijayaraghavan P, Vosoughi S, 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), Association for Computational Linguistics, Vancouver, Canada, pp 478–483, DOI 10.18653/v1/P17-2076, URL https://www.aclweb.org/anthology/P17-2076
Voigt R, Jurgens D, Prabhakaran V, Jurafsky D, Tsvetkov Y (2018) RtGender: A corpus for studying differential responses to gender. In: Proceedings of the Eleventh International Conference on Language Resources and Eval-

uation (LREC 2018), European Language Resources Association (ELRA), Miyazaki, Japan, URL https://www.aclweb.org/anthology/L18-1445

Voorhees E (2005) Overview of the trec 2005 robust retrieval track. URL https://trec.nist.gov/pubs/trec13/papers/ROBUST.OVERVIEW.pdf

Wadsworth C, Vera F, 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)”, 1807.00199

Wah C, Branson S, Welinder P, Perona P, Belongie S (2011) The caltech-ucsd birds200-2011 dataset. Advances in Water Resources - ADV WATER RESOUR

Wan M, McAuley J (2018) Item recommendation on monotonic behavior chains. In: Proceedings of the 12th ACM Conference on Recommender Systems, Association for Computing Machinery, New York, NY, USA, RecSys ’18, p 86–94, DOI 10.1145/3240323.3240369, URL https://doi.org/10.1145/3240323.3240369

Wang A, Pruksachatkun Y, Nangia N, Singh A, Michael J, Hill F, Levy O, 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, Garnett R (eds) Advances in Neural Information Processing Systems, Curran Associates, Inc., vol 32, URL https://proceedings.neurips.cc/paper/2019/file/4496bf24afe7fab6f046bf4923da8de6-Paper.pdf

Wang A, Singh A, Michael J, Hill F, Levy O, Bowman SR (2019b) GLUE: A multi-task benchmark and analysis platform for natural language understanding. In: International Conference on Learning Representations, URL https://openreview.net/forum?id=rJ4km2R5t7

Wang H, Grgic-Hlaca N, Lahoti P, Gummadi KP, Weller A (2019c) An empirical study on learning fairness metrics for compas data with human supervision. NeurIPS 2019 workshop: “Human-Centric Machine Learning”, 1910.10255

Wang H, Ustun B, Calmon F (2019d) Repairing without retraining: Avoiding disparate impact with counterfactual distributions. In: Chaudhuri K, Salakhutdinov R (eds) Proceedings of the 36th International Conference on Machine Learning, PMLR, Long Beach, California, USA, Proceedings of Machine Learning Research, vol 97, pp 6618–6627, URL http://proceedings.mlr.press/v97/wang19l.html

Wang J, Liu Y, Levy C (2021) Fair classification with group-dependent label noise. In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, Association for Computing Machinery, New York, NY, USA, FAccT ’21, p 526–536, DOI 10.1145/3442188.3445915, URL https://doi.org/10.1145/3442188.3445915

Wang M, Deng W (2020) Mitigating bias in face recognition using skewness-aware reinforcement learning. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)

Wang M, Deng W, Hu J, Tao X, 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, pp 692–702

Wang S, Guo W, Narasimhan H, Cotter A, Gupta M, Jordan M (2020a) Robust optimization for fairness with noisy protected groups. In: Larochelle H, Ranzato M, Hadsell R, Balcan MF, Lin H (eds) Advances in Neural Information Processing Systems, Curran Associates, Inc., vol 33, pp 5190–5203, URL https://proceedings.neurips.cc/paper/2020/file/37d097caf1299d9aa79c2c2b843d2d78-Paper.pdf

Wang T, 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)", 2011.08398

Wang Z, Qinami K, Karakozis IC, Genova K, Nair P, Hata K, Russakovsky O (2020b) Towards fairness in visual recognition: Effective strategies for bias mitigation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)

Waseem Z, Hovy D (2016) Hateful symbols or hateful people? predictive features for hate speech detection on Twitter. In: Proceedings of the NAACL Student Research Workshop, Association for Computational Linguistics, San Diego, California, pp 88–93, DOI 10.18653/v1/N16-2013, URL https://www.aclweb.org/anthology/N16-2013

Webster K, Recasens M, Axelrod V, Baldridge J (2018) Mind the gap: A balanced corpus of gendered ambiguous pronouns. 1810.05201

Weeks M, Clair S, Borgatti S, Radda K, Schensul J (2002) Social networks of drug users in high-risk sites: Finding the connections. AIDS and Behavior 6:193–206, DOI 10.1023/A:1015457400897

Wick M, panda s, Tristan JB (2019) Unlocking fairness: a trade-off revisited. In: Wallach H, Larochelle H, Beygelzimer A, d'Alché-Buc F, Fox E, Garnett R (eds) Advances in Neural Information Processing Systems, Curran Associates, Inc., vol 32, pp 8783–8792, URL https://proceedings.neurips.cc/paper/2019/file/373e4c5d8edfa8b74fd4b6791d0cf6dc-Paper.pdf

Wieringa J, Kannan P, Ma X, Reutterer T, Risselada H, Skiera B (2021) Data analytics in a privacy-concerned world. Journal of Business Research 122:915–925, DOI https://doi.org/10.1016/j.jbusres.2019.05.005, URL https://www.sciencedirect.com/science/article/pii/S0148296319303078

Wightman L, Ramsey H, Council LSA (1998) LSAC National Longitudinal Bar Passage Study. LSAC research report series, Law School Admission Council, URL https://books.google.it/books?id=WdA7AQAAIAAJ

Wilder B, Ou HC, de la Haye K, Tambe M (2018) Optimizing network structure for preventative health. In: Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems, International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, AAMAS '18, p 841–849

Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten JW, da Silva Santos LB, Bourne PE, et al. (2016) The fair guiding principles for scientific data management and stewardship. Scientific data 3(1):1–9

Williams JV, Razavian N (2019) Quantification of bias in machine learning for healthcare: A case study of renal failure prediction. URL https://drive.google.com/file/d/1dvJfvVLIQVeeKaLrMlXfX6lcVTzhkDQ0/view, neurIPS 2019 workshop: “Fair ML for Health”

Williamson R, Menon A (2019) Fairness risk measures. In: Chaudhuri K, Salakhutdinov R (eds) Proceedings of the 36th International Conference on Machine Learning, PMLR, Long Beach, California, USA, Proceedings of Machine Learning Research, vol 97, pp 6786–6797, URL http://proceedings.mlr.press/v97/williamson19a.html

Wilson C, Ghosh A, Jiang S, Mislove A, Baker L, Szary J, Trindel K, 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, Association for Computing Machinery, New York, NY, USA, FAccT ’21, p 666–677, DOI 10.1145/3442188.3445928, URL https://doi.org/10.1145/3442188.3445928

Wondracek G, Holz T, Kirda E, Kruegel C (2010) A practical attack to de-anonymize social network users. In: 2010 IEEE Symposium on Security and Privacy, pp 223–238, DOI 10.1109/SP.2010.21

Wu Y, Zhang L, Wu X (2018) 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, Association for Computing Machinery, New York, NY, USA, KDD ’18, p 2536–2544, DOI 10.1145/3219819.3220087, URL https://doi.org/10.1145/3219819.3220087

Wu Y, Zhang L, Wu X, 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, Garnett R (eds) Advances in Neural Information Processing Systems, Curran Associates, Inc., vol 32, pp 3404–3414, URL https://proceedings.neurips.cc/paper/2019/file/44a2e0804995faf8d2e3b084a1e2db1d-Paper.pdf

Xiao H, Rasul K, Vollgraf R (2017) Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. 1708.07747

Xiao W, Zhao H, Pan H, Song Y, Zheng VW, 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, Association for Computing Machinery, New York, NY, USA, KDD ’19, p 235–245, DOI 10.1145/3292500.3330965, URL https://doi.org/10.1145/3292500.3330965

Xie M, Lauritsen JL (2012) Racial context and crime reporting: A test of Black’s stratification hypothesis. Journal of Quantitative Criminology 28(2):265–293

Xu D, Yuan S, Zhang L, Wu X (2018) Fairgan: Fairness-aware generative adversarial networks. In: 2018 IEEE International Conference on Big Data (Big Data), IEEE, pp 570–575

Xu R, Cui P, Kuang K, Li B, Zhou L, Shen Z, Cui W (2020) Algorithmic Decision Making with Conditional Fairness, Association for Computing Machinery, New York, NY, USA, p 2125–2135. URL https://doi.org/10.1145/3394486.3403263

Xu X, Huang Y, Shen P, Li S, Li J, Huang F, Li Y, 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), pp 578–586

Yang F, Cisse M, Koyejo S (2020a) Fairness with overlapping groups; a probabilistic perspective. In: Larochelle H, Ranzato M, Hadsell R, Balcan MF, Lin H (eds) Advances in Neural Information Processing Systems, Curran Associates, Inc., vol 33, pp 4067–4078, URL https://proceedings.neurips.cc/paper/2020/file/29c0605a3bab4229e46723f89cf59d83-Paper.pdf

Yang K, Stoyanovich J (2017) Measuring fairness in ranked outputs. In: Proceedings of the 29th International Conference on Scientific and Statistical Database Management, Association for Computing Machinery, New York, NY, USA, SSDBM '17, DOI 10.1145/3085504.3085526, URL https://doi.org/10.1145/3085504.3085526

Yang K, Qinami K, Fei-Fei L, Deng J, 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, Association for Computing Machinery, New York, NY, USA, FAT* '20, p 547–558, DOI 10.1145/3351095.3375709, URL https://dl.acm.org/doi/pdf/10.1145/3351095.3375709

Yang M, Kim B (2019) Benchmarking attribution methods with relative feature importance. 1907.09701

Yao S, Huang B (2017a) Beyond parity: Fairness objectives for collaborative filtering. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, Garnett R (eds) Advances in Neural Information Processing Systems, Curran Associates, Inc., vol 30, pp 2921–2930, URL https://proceedings.neurips.cc/paper/2017/file/e6384711491713d29bc63fc5eeb5ba4f-Paper.pdf

Yao S, Huang B (2017b) New fairness metrics for recommendation that embrace differences. KDD 2017 workshop: "Fairness, Accountability, and Transparency in Machine Learning (FAT/ML)", 1706.09838

Yeh IC, 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, DOI https://doi.org/10.1016/j.eswa.2007.12.020, URL http://www.sciencedirect.com/science/article/pii/S0957417407006719

Yi S, Xiaogang W, Xiaoou T (2013) Deep convolutional network cascade for facial point detection. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition, pp 3476–3483, DOI 10.1109/CVPR.2013.446

Yi S, Wang S, Joshi S, Ghassemi M (2019) Fair and robust treatment effect estimates: Estimation under treatment and outcome disparity with deep neural models. URL https://drive.google.com/file/d/1hUHRovnfzxnPaselTczzuQfvGU9jbTI1/view, neurIPS 2019 workshop: "Fair ML for Health"

Yurochkin M, Sun Y (2021) Sensei: Sensitive set invariance for enforcing individual fairness. In: International Conference on Learning Representations, URL https://openreview.net/forum?id=DktZb97_Fx

Yurochkin M, Bower A, Sun Y (2020) Training individually fair ml models with sensitive subspace robustness. In: International Conference on Learning Representations, URL https://openreview.net/forum?id=B1gdkxHFDH

Zafar MB, Valera I, Gomez Rodriguez M, Gummadi KP (2017a) Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In: Proceedings of the 26th International Conference on World Wide Web, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, WWW '17, p 1171–1180, DOI 10.1145/3038912.3052660, URL https://doi.org/10.1145/3038912.3052660

Zafar MB, Valera I, Rodriguez M, Gummadi K, Weller A (2017b) From parity to preference-based notions of fairness in classification. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, Garnett R (eds) Advances in Neural Information Processing Systems, Curran Associates, Inc., vol 30, pp 229–239, URL https://proceedings.neurips.cc/paper/2017/file/82161242827b703e6acf9c726942a1e4-Paper.pdf

Zafar MB, Valera I, Rogriguez MG, Gummadi KP (2017c) Fairness constraints: Mechanisms for fair classification. In: Artificial Intelligence and Statistics, PMLR, pp 962–970

Zehlike M, Yang K, Stoyanovich J (2021) Fairness in ranking: A survey. arXiv preprint arXiv:210314000

Zhang BH, Lemoine B, Mitchell M (2018) Mitigating unwanted biases with adversarial learning. In: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, Association for Computing Machinery, New York, NY, USA, AIES '18, p 335–340, DOI 10.1145/3278721.3278779, URL https://doi.org/10.1145/3278721.3278779

Zhang H, Davidson I (2021) Towards fair deep anomaly detection. In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, Association for Computing Machinery, New York, NY, USA, FAccT '21, p 138–148, DOI 10.1145/3442188.3445878, URL https://doi.org/10.1145/3442188.3445878

Zhang H, Lu AX, Abdalla M, McDermott M, Ghassemi M (2020a) Hurtful words: Quantifying biases in clinical contextual word embeddings. In: Proceedings of the ACM Conference on Health, Inference, and Learning, Association for Computing Machinery, New York, NY, USA, CHIL '20, p 110–120, DOI 10.1145/3368555.3384448, URL https://doi.org/10.1145/3368555.3384448

Zhang J, Bareinboim E (2018) Equality of opportunity in classification: A causal approach. In: Bengio S, Wallach H, Larochelle H, Grauman K, Cesa-Bianchi N, Garnett R (eds) Advances in Neural Information Processing Systems, Curran Associates, Inc., vol 31, pp 3671–3681, URL https://proceedings.neurips.cc/paper/2018/file/ff1418e8cc993fe8abcfe3ce2003e5c5-Paper.pdf
Zhang L, Wu Y, Wu X (2017a) Achieving non-discrimination in data release. In: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Association for Computing Machinery, New York, NY, USA, KDD '17, p 1335–1344, DOI 10.1145/3097983.3098167, URL https://doi.org/10.1145/3097983.3098167
Zhang X, Khaliligarekani M, Tekin C, 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, Garnett R (eds) Advances in Neural Information Processing Systems, Curran Associates, Inc., vol 32, pp 15269–15278, URL https://proceedings.neurips.cc/paper/2019/file/7690dd4db7a92524c684e3191919eb6b-Paper.pdf
Zhang X, Tu R, Liu Y, Liu M, Kjellström H, Zhang K, Zhang C (2020b) How do fair decisions fare in long-term qualification? In: Larochelle H, Ranzato M, Hadsell R, Balcan M, Lin H (eds) Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual, URL https://proceedings.neurips.cc/paper/2020/hash/d6d231705f96d5a3aeb3a76402e49a3-Abstract.html
Zhang Y (2005) Bayesian graphical model for adaptive information filtering. PhD thesis, Carnegie Mellon University.
Zhang Y, Bellamy R, 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, Association for Computing Machinery, New York, NY, USA, AIES '20, p 400–406, DOI 10.1145/3375627.3375862, URL https://doi.org/10.1145/3375627.3375862
Zhang Z, Neill DB (2017) Identifying significant predictive bias in classifiers. KDD 2017 workshop: "Fairness, Accountability, and Transparency in Machine Learning (FAT/ML)", 1611.08292
Zhang Z, Luo P, Loy CC, Tang X (2014) Facial landmark detection by deep multi-task learning. DOI 10.1007/978-3-319-10599-4_7
Zhang Z, Luo P, Loy CC, Tang X (2015) Learning deep representation for face alignment with auxiliary attributes. IEEE transactions on pattern analysis and machine intelligence 38(5):918–930
Zhang Z, Song Y, Qi H (2017b) Age progression/regression by conditional adversarial autoencoder. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)
Zhao B, Xiao X, Gan G, Zhang B, Xia ST (2020a) Maintaining discrimination and fairness in class incremental learning. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)

Zhao C, Li C, Li J, Chen F (2020b) Fair meta-learning for few-shot classification. In: 2020 IEEE International Conference on Knowledge Graph (ICKG), pp 275–282, DOI 10.1109/ICBK50248.2020.00047

Zhao H, Gordon G (2019) Inherent tradeoffs in learning fair representations. In: Wallach H, Larochelle H, Beygelzimer A, d'Alché-Buc F, Fox E, Garnett R (eds) Advances in Neural Information Processing Systems, Curran Associates, Inc., vol 32, pp 15675–15685, URL https://proceedings.neurips.cc/paper/2019/file/b4189d9de0fb2b9cce090bd1a15e3420-Paper.pdf

Zhao H, Coston A, Adel T, Gordon GJ (2020c) Conditional learning of fair representations. In: International Conference on Learning Representations, URL https://openreview.net/forum?id=Hkekl0NFPr

Zhao J, Wang T, Yatskar M, Ordonez V, Chang KW (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, Association for Computational Linguistics, Copenhagen, Denmark, pp 2979–2989, DOI 10.18653/v1/D17-1323, URL https://www.aclweb.org/anthology/D17-1323

Zhao J, Wang T, Yatskar M, Ordonez V, Chang KW (2018) 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), Association for Computational Linguistics, New Orleans, Louisiana, pp 15–20, DOI 10.18653/v1/N18-2003, URL https://www.aclweb.org/anthology/N18-2003

Zhao Y, Kong S, 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), pp 15759–15768

Zheng Y, Dave T, Mishra N, Kumar H (2018) Fairness in reciprocal recommendations: A speed-dating study. In: Adjunct Publication of the 26th Conference on User Modeling, Adaptation and Personalization, Association for Computing Machinery, New York, NY, USA, UMAP '18, p 29–34, DOI 10.1145/3213586.3226207, URL https://doi.org/10.1145/3213586.3226207

Zhong Y, Deng W, Wang M, Hu J, Peng J, Tao X, 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)

Zhou B, Lapedriza A, Khosla A, Oliva A, Torralba A (2018) Places: A 10 million image database for scene recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 40(6):1452–1464, DOI 10.1109/TPAMI.2017.2723009

Zhu Y, Kiros R, Zemel R, Salakhutdinov R, Urtasun R, Torralba A, Fidler S (2015) Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. 1506.06724

Zhu Z, Wang J, Zhang Y, Caverlee J (2018) Fairness-aware recommendation of information curators. RecSys 2018 workshop: "Workshop on Responsible Recommendation (FAT/Rec)", 1809.03040

Žliobaitė I (2015) On the relation between accuracy and fairness in binary classification. ICML 2015 workshop: "Fairness, Accountability, and Transparency in Machine Learning (FAT/ML)", 1505.05723

Žliobaitė I, Kamiran F, Calders T (2011) Handling conditional discrimination. In: 2011 IEEE 11th International Conference on Data Mining, pp 992–1001, DOI 10.1109/ICDM.2011.72

Appendix A 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. 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)[11]; to be compliant with the FAIR data principles (Wilkinson et al., 2016), 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.[12] 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 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

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

[12] 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`.

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,[13] with the addition of "news", "social media", "social networks", "sports" and "food". Table 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 5.2. 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 (Pleiss et al., 2017). 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".

A.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).
{ **A liation of creators**: US Census Bureau.
{ **Domain**: linguistics.
{ **Tasks in fairness literature**: fair subset selection under unawareness (Mehrotra and Celis, 2021).
{ **Data spec**: tabular data.
{ **Sample size**: 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.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.
{ **A liation of creators**: FiveThirtyEight.
{ **Domain**: political science.
{ **Tasks in fairness literature**: limited-label fairness evaluation (Sabato and Yom-Tov, 2020).

[13] `https://www.scimagojr.com/journalrank.php`

- **Data spec**: tabular data.
- **Sample size**: 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.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 (Harb and Lam, 2020).
- **Data spec**: author-author pairs.
- **Sample size**: 30K nodes (authors) connected by 200K edges (co-author relations).
- **Year**: 2009.
- **Sensitive features**: author.
- **Link**: not available
- **Further info**: Sun et al. (2009)

A.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 (Kang et al., 2020).
- **Data spec**: author-author pairs.
- **Sample size**: 19K nodes (authors) connected by 200K edges (indications of co-authorship) (AstroPh). 23K nodes connected by 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. (2007)

A.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 (2018) labeled each image in Adience with Fitzpatrick skin type.
- **A liation of creators**: Adience; Open University of Israel.
- **Domain**: computer vision.
- **Tasks in fairness literature**: data bias evaluation (Buolamwini and Gebru, 2018), robust fairness evaluation (Nanda et al., 2021).
- **Data spec**: image.
- **Sample size**: 30K images of 2K subjects.
- **Year**: 2014.
- **Sensitive features**: age, gender, skin type.
- **Link**: `https://talhassner.github.io/home/projects/Adience/Adience-data.html`
- **Further info**: Eidinger et al. (2014); Buolamwini and Gebru (2018)

A.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.
- **A liation of creators**: Norwegian University of Science and Technology; Adresseavisen.
- **Domain**: news, information systems.
- **Tasks in fairness literature**: fair ranking (Chakraborty et al., 2019).
- **Data spec**: user-article pairs.
- **Sample size**: 3M ratings by 15M readers over $1K$ articles.
- **Year**: 2018.
- **Sensitive features**: geography.
- **Link**: `http://reclab.idi.ntnu.no/dataset/`
- **Further info**: (Gulla et al., 2017)

A.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 B for extensive documentation.
- **A liation of creators**: Silicon Graphics Inc.
- **Domain**: economics.
- **Tasks in fairness literature**: fairness evaluation (Sharma et al., 2020a; Cardoso et al., 2019; Oneto et al., 2019a; Friedler et al., 2019; Chen et al., 2018c; Lipton et al., 2018; Pleiss et al., 2017; DiCiccio et al., 2020; Speicher et al., 2018b; Feldman et al., 2015; Maity et al., 2021; Kim et al., 2020; Liu et al., 2019; Williamson and Menon, 2019; von Kügelgen et al., 2021; Ngong et al., 2020; Jabbari et al., 2020; Huan et al., 2020; Žliobaitė,

2015; Islam et al., 2021; Segal et al., 2021), fair classification (He et al., 2020b; Sharma et al., 2020b; Goel et al., 2018; Raff et al., 2018; Zhang et al., 2018; Hu and Chen, 2020; Celis et al., 2019b; Yang et al., 2020a; Cho et al., 2020; Savani et al., 2020; Wu et al., 2019; Donini et al., 2018; Quadrianto and Sharmanska, 2017; Calmon et al., 2017; Xu et al., 2020; Zhang et al., 2017a; Yurochkin and Sun, 2021; Vargo et al., 2021; Chuang and Mroueh, 2021; Roh et al., 2021; Yurochkin et al., 2020; Baharlouei et al., 2020; Lohaus et al., 2020; Martinez et al., 2020; Roh et al., 2020; Celis et al., 2020a; Cotter et al., 2019; Gordaliza et al., 2019; Wang et al., 2019d; Agarwal et al., 2018; Creager et al., 2021; Delobelle et al., 2020; Ogura and Takeda, 2020; Feldman et al., 2015; Zafar et al., 2017c; Fish et al., 2015; Raff and Sylvester, 2018; Mhasawade and Chunara, 2021; Perrone et al., 2021; Shah et al., 2021; Sharma et al., 2021), fair clustering (Abbasi et al., 2021; Ghadiri et al., 2021; Harb and Lam, 2020; Ahmadian et al., 2020; Huang et al., 2019; Bera et al., 2019; Chierichetti et al., 2017; Brubach et al., 2020; Mahabadi and Vakilian, 2020; Backurs et al., 2019; Mary et al., 2019; Wang and Saar-Tsechansky, 2020; Berk et al., 2017; Beutel et al., 2017), fair clustering under unawareness (Esmaeili et al., 2020), fair active classification (Noriega-Campero et al., 2019; Bakker et al., 2019, 2021), fair preference-based classification (Ali et al., 2019b; Ustun et al., 2019; Mukherjee et al., 2020), fair classification under unawareness (Lahoti et al., 2020; Wang et al., 2020a; Mozannar et al., 2020; Kilbertus et al., 2018), fair anomaly detection (Zhang and Davidson, 2021; Shekhar et al., 2021), fairness evaluation under unawareness (Awasthi et al., 2021), robust fairness evaluation (Black and Fredrikson, 2021), data bias evaluation (Beretta et al., 2021), rich-subgroup fairness evaluation (Kearns et al., 2019; Chouldechova and G'Sell, 2017), fair representation learning (Ruoss et al., 2020; Zhao and Gordon, 2019; Zhao et al., 2020c; Louizos et al., 2016; Quadrianto et al., 2019; Madras et al., 2018a), fair multi-stage classification (Hu et al., 2020; Goel et al., 2020), robust fair classification (Mandal et al., 2020; Huang and Vishnoi, 2019; Rezaei et al., 2021), dynamical fair classification (Zhang et al., 2019), fair ranking evaluation (Kallus and Zhou, 2019b), fair data summarization (Chiplunkar et al., 2020; Jones et al., 2020; Kleindessner et al., 2019a; Celis et al., 2018; El Halabi et al., 2020; Belitz et al., 2021), fair regression (Agarwal et al., 2019), limited-label fair classification (Chzhen et al., 2019; Wang et al., 2021; Choi et al., 2020b), limited-label fairness evaluation (Ji et al., 2020), preference-based fair clustering (Galhotra et al., 2021).

{ **Data spec**: tabular data.
{ **Sample size**: 50K instances.
{ **Year**: 1996.
{ **Sensitive features**: age, sex, race.
{ **Link**: `https://archive.ics.uci.edu/ml/datasets/adult`
{ **Further info**: Kohavi (1996); UCI Machine Learning Repository (1996); US Dept. of Commerce Bureau of the Census (1995); Ding et al. (2021); McKenna (2019a,b)

A.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.

{ **A liation 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 (Coston et al., 2020), fair risk assessment (Mishler et al., 2021).
{ **Data spec**: tabular data.
{ **Sample size**: 80K calls.
{ **Year**: 2019.
{ **Sensitive features**: age, race, gender of child.
{ **Link**: not available
{ **Further info**: Vaithianathan et al. (2017)

A.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.
{ **A liation of creators**: Indian Institute of Technology; Max Planck Institute for Software Systems.
{ **Domain**: information systems.
{ **Tasks in fairness literature**: fair ranking evaluation (Dash et al., 2021).
{ **Data spec**: item-recommendation pairs.
{ **Sample size**: 1M recommendations associated with 20K items.
{ **Year**: 2021.
{ **Sensitive features**: brand ownership.
{ **Link**: not available
{ **Further info**: Dash et al. (2021)

A.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.
{ **A liation of creators**: University of California, San Diego.
{ **Domain**: information systems.
{ **Tasks in fairness literature**: fair ranking (Patro et al., 2019).
{ **Data spec**: user-product pairs (reviews).
{ **Sample size**: 200M reviews of products.
{ **Year**: 2018.
{ **Sensitive features**: none.
{ **Link**: `https://nijianmo.github.io/amazon/index.html`
{ **Further info**: McAuley et al. (2015); He and McAuley (2016)

A.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.
{ **A liation 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 (Kallus and Zhou, 2019a).
{ **Data spec**: tabular data.
{ **Sample size**: 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. (2014)

A.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.
{ **A liation of creators**: National University of Singapore; National University of Southern California.
{ **Domain**: public health.
{ **Tasks in fairness literature**: fair graph diffusion (Farnad et al., 2020).
{ **Data spec**: agent-agent pairs.
{ **Sample size**: 20 synthetic networks, containing 500 individuals each.
{ **Year**: 2019.
{ **Sensitive features**: ethnicity, gender, age, geography.
{ **Link**: `https://github.com/bwilder0/fair_influmax_code_release`
{ **Further info**: Wilder et al. (2018); Tsang et al. (2019)

A.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.
{ **A liation of creators**: Massachusetts Institute of Technology; Massachusetts General Hospital; Harvard Medical School.
{ **Domain**: sleep medicine.

{ **Tasks in fairness literature**: fair preference-based classification (Ustun et al., 2019).
{ **Data spec**: mixture (time series and tabular data).
{ **Sample size**: 2K patients.
{ **Year**: 2016.
{ **Sensitive features**: age, sex.
{ **Link**: not available
{ **Further info**: Ustun et al. (2016)

A.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.
{ **A liation of creators**: Tsinghua University; IBM.
{ **Domain**: library and information sciences.
{ **Tasks in fairness literature**: fair graph mining (Buyl and De Bie, 2020).
{ **Data spec**: article-article pairs.
{ **Sample size**: 5M papers connected by 50M citations.
{ **Year**: 2021.
{ **Sensitive features**: author.
{ **Link**: `http://www.arnetminer.org/citation`
{ **Further info**: Tang et al. (2008); `https://www.aminer.org/`

A.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.
{ **A liation of creators**: Bilkent University; Baskent University.
{ **Domain**: cardiology.
{ **Tasks in fairness literature**: fair classification (Donini et al., 2018; Mary et al., 2019), robust fair classification (Rezaei et al., 2021), limited-label fair classification (Chzhen et al., 2019).
{ **Data spec**: tabular data.
{ **Sample size**: 500 patients.
{ **Year**: 1997.
{ **Sensitive features**: age, sex.
{ **Link**: `https://archive.ics.uci.edu/ml/datasets/arrhythmia`
{ **Further info**: Guvenir et al. (1997)

A.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 (Tong and Kagal, 2020).
- **Data spec**: image.
- **Sample size**: 800 images of athletes and 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 (2020)

A.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 (Madnani et al., 2017).
- **Data spec**: text.
- **Sample size**: 20K student essays.
- **Year**: 2012.
- **Sensitive features**: none.
- **Link**: `https://www.kaggle.com/c/asap-aes/data/`
- **Further info**: Shermis (2014)

A.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 (Savani et al., 2020; Baharlouei et al., 2020; Zafar et al., 2017c; Diana et al., 2021; Shah et al., 2021), fair clustering (Abbasi et al., 2021; Harb and Lam, 2020; Ahmadian et al., 2020; Huang et al., 2019; Bera et al., 2019; Mahabadi and Vakilian, 2020; Backurs et al., 2019; Chierichetti et al., 2017), fair data summarization (El Halabi et al., 2020), fair classification under unawareness (Kilbertus et al., 2018), fairness evaluation (Lipton et al., 2018; Islam et al.,

2021), limited-label fairness evaluation (Ji et al., 2020), preference-based fair clustering (Galhotra et al., 2021).
- **Data spec**: tabular data.
- **Sample size**: 40K phone contacts.
- **Year**: 2012.
- **Sensitive features**: age.
- **Link**: `https://archive.ics.uci.edu/ml/datasets/Bank+Marketing`
- **Further info**: Moro et al. (2014)

A.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 (Solans et al., 2021).
- **Data spec**: lister-seeker pairs.
- **Sample size**: 4M pairs.
- **Year**: 2021.
- **Sensitive features**: gender, age, spoken language, "gay-friendliness".
- **Link**: not available
- **Further info**: Solans et al. (2021)

A.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 (Lin et al., 2014) into scene images from MiniPlaces (Zhou et al., 2018). 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 (David et al., 2020).
- **Data spec**: image.
- **Sample size**: 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 (2019)

A.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 (Jiang and Pardos, 2021).
- **Data spec**: tabular data.
- **Sample size**: 2M enrollments across 80K students.
- **Year**: 2021.
- **Sensitive features**: gender, race.
- **Link**: not available
- **Further info**: Jiang and Pardos (2021)

A.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 (De-Arteaga et al., 2019), fair classification (Yurochkin and Sun, 2021).
- **Data spec**: text.
- **Sample size**: 400K biographies.
- **Year**: 2018.
- **Sensitive features**: gender.
- **Link**: `https://github.com/Microsoft/biosbias`
- **Further info**: De-Arteaga et al. (2019)

A.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 (Cho et al., 2021).
- **Data spec**: text.
- **Sample size**: 1K templates.
- **Year**: 2021.
- **Sensitive features**: gender.
- **Link**: `https://github.com/nolongerprejudice/tgbi-x`
- **Further info**: Cho et al. (2021)

A.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 (Mehrotra et al., 2017).
- **Data spec**: query-result pairs.
- **Sample size**: 30M (non-unique) queries issued by 4M distinct users.
- **Year**: 2017.
- **Sensitive features**: age, gender.
- **Link**: not available
- **Further info**: Mehrotra et al. (2017)

A.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 (Dhamala et al., 2021).
- **Data spec**: text.
- **Sample size**: 20K prompts.
- **Year**: 2021.
- **Sensitive features**: gender, race, religion, profession, political leaning.
- **Link**: `https://github.com/amazon-research/bold`
- **Further info**: Dhamala et al. (2021)

A.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 (Radford et al., 2018) and BERT (Devlin et al., 2019) language models were trained on this dataset.
- **A liation of creators**: University of Toronto; Massachusetts Institute of Technology.
- **Domain**: linguistics.
- **Tasks in fairness literature**: data bias evaluation (Tan and Celis, 2019).
- **Data spec**: text.
- **Sample size**: 1B words in 74M sentences from 11K books.
- **Year**: unknown.
- **Sensitive features**: textual references to people and their demographics.
- **Link**: not available
- **Further info**: Zhu et al. (2015); Bandy and Vincent (2021)

A.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 (§ A.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 (§ A.124), on which the BUPT Faces datasets are heavily based.
- **A liation of creators**: Beijing University of Posts and Telecommunications.
- **Domain**: computer vision.
- **Tasks in fairness literature**: fair reinforcement learning (Wang and Deng, 2020), fair classification (Xu et al., 2021), fair representation learning (Gong et al., 2021).
- **Data spec**: image.
- **Sample size**: 2M images of 40K celebrities (BUPT-Globalface); 1M images of 30K celebrities (BUPT-Balancedface).
- **Year**: 2019.
- **Sensitive features**: race.
- **Link**: `http://www.whdeng.cn/RFW/Trainingdataste.html`
- **Further info**: Wang and Deng (2020)

A.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.
- **A liation of creators**: Shopify.
- **Domain**: information systems.
- **Tasks in fairness literature**: fair ranking (Karako and Manggala, 2018).
- **Data spec**: image.
- **Sample size**: 3K images.
- **Year**: present.
- **Sensitive features**: gender.
- **Link**: not available
- **Further info**: Karako and Manggala (2018); `https://burst.shopify.com/`

A.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.
- **A liation of creators**: Google.
- **Domain**: linguistics.
- **Tasks in fairness literature**: fair entity resolution (Cotter et al., 2019).
- **Data spec**: text.
- **Sample size**: 15K samples.
- **Year**: 2019.
- **Sensitive features**: geography, business size.
- **Link**: not available
- **Further info**: Cotter et al. (2019)

A.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.
- **A liation of creators**: Jain University Bangalore.
- **Domain**: education.
- **Tasks in fairness literature**: fair data generation (Liu et al., 2021).
- **Data spec**: tabular data.
- **Sample size**: 200 students.
- **Year**: 2020.
- **Sensitive features**: gender.
- **Link**: `https://www.kaggle.com/datasets/benroshan/factors-affecting-campus-placement`
- **Further info**:

A.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.
- **A liation of creators**: University of Michigan.
- **Domain**: computer vision.
- **Tasks in fairness literature**: fair representation learning (Locatello et al., 2019).
- **Data spec**: image.
- **Sample size**: 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. (2015)

A.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".

- **A liation of creators**: Chinese University of Hong Kong.
- **Domain**: computer vision.
- **Tasks in fairness literature**: fair classification (Savani et al., 2020; Kim et al., 2019; Chuang and Mroueh, 2021; Lohaus et al., 2020; Creager et al., 2019; Jung et al., 2021), fair anomaly detection (Zhang and Davidson, 2021), bias discovery (Amini et al., 2019) fair anomaly detection (Zhang and Davidson, 2021), fairness evaluation of private classification (Cheng et al., 2021b), fairness evaluation of selective classification (Jones et al., 2021), fairness evaluation (Wang et al., 2020b; Segal et al., 2021), fair representation learning (Quadrianto et al., 2019), fair data summarization (Chiplunkar et al., 2020), fair data generation (Choi et al., 2020a; Ramaswamy et al., 2021).
- **Data spec**: image.
- **Sample size**: 200K face images of over 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. (2015)

A.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.
- **A liation of creators**: Stanford University.
- **Domain**: radiology.
- **Tasks in fairness literature**: fairness evaluation of selective classification (Jones et al., 2021), fairness evaluation of private classification (Cheng et al., 2021b).
- **Data spec**: image.
- **Sample size**: 200K chest radiographs from 60K patients.
- **Year**: 2019.
- **Sensitive features**: sex, age (of patient).
- **Link**: `https://stanfordmlgroup.github.io/competitions/chexpert/`
- **Further info**: Irvin et al. (2019); Garbin et al. (2021)

A.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).
- **A liation of creators**: City of Chicago.
- **Domain**: transportation.
- **Tasks in fairness literature**: fair pricing evaluation (Pandey and Caliskan, 2021).
- **Data spec**: tabular data.
- **Sample size**: 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-additional.html`; `http://dev.cityofchicago.org/open%20data/data%20portal/2019/04/12/tnp-taxi-privacy.html`

A.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 (Wang et al., 2020b; Jung et al., 2021), fair incremental learning (Zhao et al., 2020a), robust fairness evaluation (Nanda et al., 2021).
- **Data spec**: image.
- **Sample size**: $6K$ 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 (2009)
- **Variants**: CIFAR-10S (Wang et al., 2020b) is a modified version specifically aimed at studying biases in image classification across an artificial sensitive attribute (color/grayscale).

A.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 (Li et al., 2021).
- **Data spec**: paper-paper pairs.
- **Sample size**: 3K articles connected by 5K citations.
- **Year**: 2016.
- **Sensitive features**: none.
- **Link**: `http://networkrepository.com/citeseer.php`
- **Further info**: Lu and Getoor (2003)

A.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 (Adragna et al., 2020; Yurochkin and Sun, 2021; Chuang and Mroueh, 2021), fairness evaluation of selective classification (Jones et al., 2021), fair robust toxicity classification (Adragna et al., 2020), fairness evaluation of toxicity classification (Hutchinson et al., 2020), fairness evaluation (Babaeianjelodar et al., 2020).
- **Data spec**: text.
- **Sample size**: 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. (2019)

A.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 $60K$ 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 (Flanigan et al., 2020).
- **Data spec**: tabular data.
- **Sample size**: 2K pool participants.
- **Year**: 2020.
- **Sensitive features**: gender, age, education, urban/rural, geography, ethnicity.
- **Link**: not available
- **Further info**: Flanigan et al. (2020); `https://www.climateassembly.uk/`

A.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 (Zheng et al., 2018), preference-based fair ranking (Paraschakis and Nilsson, 2020).
- **Data spec**: person-person pairs.
- **Sample size**: 10K dating records involving 400 people.

- **Year**: 2016.
- **Sensitive features**: gender, age, race, geography.
- **Link**: `https://data.world/annavmontoya/speed-dating-experiment`
- **Further info**: Fisman et al. (2006)

A.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 (Yang et al., 2020a; Sharifi-Malvajerdi et al., 2019; Heidari et al., 2018; Lohaus et al., 2020; Cotter et al., 2019; Creager et al., 2019; Cotter et al., 2018), fair regression evaluation (Heidari et al., 2019a), fair few-shot learning (Slack et al., 2020, 2019a), rich-subgroup fairness evaluation (Kearns et al., 2019), rich-subgroup fair classification (Kearns et al., 2018), fair regression (Chzhen et al., 2020a,b; Romano et al., 2020; Agarwal et al., 2019; Mary et al., 2019; Komiyama et al., 2018; Ogura and Takeda, 2020; Berk et al., 2017; Diana et al., 2021), fair representation learning (Ruoss et al., 2020), robust fair classification (Mandal et al., 2020), fair private classification (Jagielski et al., 2019), fairness evaluation of transfer learning (Lan and Huan, 2017), preference-based fair clustering (Galhotra et al., 2021).
- **Data spec**: tabular data.
- **Sample size**: $2K$ 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 (2002)

A.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 C for extensive documentation.

- **Affiliation of creators**: ProPublica.
- **Domain**: law.
- **Tasks in fairness literature**: fair classification (He et al., 2020b; Sharma et al., 2020b; Goel et al., 2018; Oneto et al., 2019a; Celis et al., 2019b; Canetti et al., 2019; Cho et al., 2020; Savani et al., 2020; Donini et al., 2018; Heidari et al., 2018; Russell et al., 2017; Quadrianto and Sharmanska, 2017; Calmon et al., 2017; DiCiccio et al., 2020; Xu et al., 2020; Vargo et al., 2021; Roh et al., 2021; Maity et al., 2021; Lohaus et al., 2020; Roh et al., 2020; Celis et al., 2020a; Cotter et al., 2019; Mary et al., 2019; Wang et al., 2019d; Delobelle et al., 2020; Ogura and Takeda, 2020; Lum and Johndrow, 2016; Zafar et al., 2017a; Berk et al., 2017; Wadsworth et al., 2018; Diana et al., 2021; Perrone et al., 2021; Ali et al., 2021), fairness evaluation (Cardoso et al., 2019; McNamara, 2019; Kasy and Abebe, 2021; Taskesen et al., 2021; Friedler et al., 2019; Wick et al., 2019; Zhang and Bareinboim, 2018; Pleiss et al., 2017; Chaibub Neto, 2020; Speicher et al., 2018b; Corbett-Davies et al., 2017; Liu et al., 2019; Agarwal et al., 2018; Ngong et al., 2020; Jabbari et al., 2020; Chouldechova, 2017; Grgic-Hlaca et al., 2016; Islam et al., 2021), fair risk assessment (Coston et al., 2020; Mishler et al., 2021; Nabi et al., 2019), fair *task assignment* (Goel and Faltings, 2019), fair classification under unawareness (Lahoti et al., 2020; Lamy et al., 2019; Chzhen et al., 2019; Kilbertus et al., 2018), data bias evaluation (Beretta et al., 2021), fair representation learning (Ruoss et al., 2020; Zhao et al., 2020c; Bower et al., 2018), robust fair classification (Mandal et al., 2020; Rezaei et al., 2021; Biswas and Mukherjee, 2021), dynamical fairness evaluation (Zhang et al., 2020b), fair reinforcement learning (Metevier et al., 2019), fair ranking evaluation (Kallus and Zhou, 2019b; Yang and Stoyanovich, 2017), fair multi-stage classification (Madras et al., 2018b), dynamical fair classification (Valera et al., 2018), preference-based fair classification (Zafar et al., 2017b; Ustun et al., 2019), fair regression (Komiyama et al., 2018), fair multi-stage classification (Goel et al., 2020), limited-label fair classification (Chzhen et al., 2019; Wang et al., 2021; Choi et al., 2020b), robust fairness evaluation (Slack et al., 2020, 2019b), rich subgroup fairness evaluation (Chouldechova and G'Sell, 2017; Zhang and Neill, 2017).
- **Data spec**: tabular data.
- **Sample size**: 12K defendants.
- **Year**: 2016.
- **Sensitive features**: sex, age, race.
- **Link**: `https://github.com/propublica/compas-analysis`
- **Further info**: Angwin et al. (2016); Larson et al. (2016)

A.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**: 3K articles connected by 5K citations.
- **Year**: 2019.
- **Sensitive features**: none.
- **Link**: `https://relational.fit.cvut.cz/dataset/CORA`
- **Further info**: McCallum et al. (2000); Sen et al. (2008)

A.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.
- **A liation of creators**: Instituto Nacional de Estadística y Censos.
- **Domain**: economics.
- **Tasks in fairness literature**: fair classification (Noriega-Campero et al., 2020).
- **Data spec**: tabular data.
- **Sample size**: 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.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.
- **A liation of creators**: Chung-Hua University; Thompson Rivers University.
- **Domain**: finance.
- **Tasks in fairness literature**: fair classification (Cho et al., 2020; Berk et al., 2017), fair clustering (Harb and Lam, 2020; Ghadiri et al., 2021; Harb and Lam, 2020; Bera et al., 2019), fair clustering under unawareness (Esmaeili et al., 2020), fair classification under unawareness (Wang et al., 2020a), fair data summarization (Tantipongpipat et al., 2019; Samadi et al., 2018), fairness evaluation (Lipton et al., 2018), fair anomaly detection (Shekhar et al., 2021).
- **Data spec**: tabular data.
- **Sample size**: 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 (2009)

A.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.

- **A liation of creators**: Yale University; Dartmouth College.
- **Domain**: finance.
- **Tasks in fairness literature**: fair pricing evaluation (Kallus and Zhou, 2021).
- **Data spec**: tabular data.
- **Sample size**: 50K clients.
- **Year**: 2008.
- **Sensitive features**: gender, age, geography.
- **Link**: `http://doi.org/10.3886/E113240V1`
- **Further info**: Karlan and Zinman (2008)

A.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 (Dressel and Farid, 2018). 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.
- **A liation of creators**: Dartmouth College.
- **Domain**: law.
- **Tasks in fairness literature**: fair *truth discovery* (Li et al., 2020d), fair *task assignment* (Li et al., 2020d; Goel and Faltings, 2019)
- **Data spec**: judge-defendant pair.
- **Sample size**: 1K defendants from COMPAS and 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**: (Dressel and Farid, 2018)
- **Variants**: a similar dataset was collected by Wang et al. (2019c).

A.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.
- **A liation of creators**: University College Dublin.
- **Domain**: literature.
- **Tasks in fairness literature**: data bias evaluation (Leavy et al., 2020).
- **Data spec**: text.
- **Sample size**: 20K books.
- **Year**: 2020.
- **Sensitive features**: textual references to people and their demographics.
- **Link**: `http://curatr.ucd.ie/`
- **Further info**: Leavy et al. (2019)

A.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.
- **A liation of creators**: University of Maryland.
- **Domain**: information systems, management information systems.
- **Tasks in fairness literature**: fair ranking (Deshpande et al., 2020).
- **Data spec**: text.
- **Sample size**: 100 resumes.
- **Year**: 2020.
- **Sensitive features**: ethnic group.
- **Link**: not available
- **Further info**: Deshpande et al. (2020)

A.49 Dallas Police Incidents

- **Description**: this dataset is due to the Dallas OpenData initiative[14] 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[15].
- **A liation of creators**: Dallas Police Department.
- **Domain**: law.
- **Tasks in fairness literature**: fair spatio-temporal process learning (Shang et al., 2020).
- **Data spec**: tabular.
- **Sample size**: 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.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.
- **A liation of creators**: Massachusetts Institute of Technology.
- **Domain**: social media.
- **Tasks in fairness literature**: fairness evaluation of sentiment analysis (Shen et al., 2018).
- **Data spec**: mixture.
- **Sample size**: 80K profiles.
- **Year**: 2017.
- **Sensitive features**: gender, age, political orientation, geography.
- **Link**: not available
- **Further info**: Vijayaraghavan et al. (2017)

[14] `https://www.dallasopendata.com/`

[15] `https://www.kaggle.com/carrie1/dallaspolicereportedincidents`

A.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.
{ **A liation of creators**: Virginia Commonwealth University; University of Cordoba; Polish Academy of Sciences.
{ **Domain**: endocrinology.
{ **Tasks in fairness literature**: fair clustering (Chierichetti et al., 2017; Bera et al., 2019; Backurs et al., 2019; Mahabadi and Vakilian, 2020; Huang et al., 2019; Bera et al., 2019).
{ **Data spec**: tabular data.
{ **Sample size**: 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. (2014)

A.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.
{ **A liation of creators**: IBM.
{ **Domain**: computer vision.
{ **Tasks in fairness literature**: fair representation learning (Quadrianto et al., 2019), fairness evaluation of private classification (Bagdasaryan et al., 2019).
{ **Data spec**: image.
{ **Sample size**: 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. (2019)

A.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.

{ **A liation 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 (Donini et al., 2018; Mary et al., 2019), evaluation of data bias (Beretta et al., 2021), limited-label fair classification (Chzhen et al., 2019), robust fair classification (Rezaei et al., 2021).
{ **Data spec**: tabular data.
{ **Sample size**: 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. (2017, 2019)

A.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.
{ **A liation 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 (Kleindessner et al., 2019b).
{ **Data spec**: person-person pairs.
{ **Sample size**: 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. (2002)

A.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 (Locatello et al., 2019; Creager et al., 2019).
- **Data spec**: image.
- **Sample size**: 700K images.
- **Year**: 2017.
- **Sensitive features**: none.
- **Link**: https://github.com/deepmind/dsprites-dataset
- **Further info**: Higgins et al. (2017)

A.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 (Agarwal et al., 2018; Xu et al., 2020; Zhang et al., 2017a; Lohaus et al., 2020), fairness evaluation (Cardoso et al., 2019).
- **Data spec**: tabular data.
- **Sample size**: 60K respondents.
- **Year**: 2001.
- **Sensitive features**: sex, age, citizenship.
- **Link**: https://sites.google.com/site/conditionaldiscrimination/
- **Further info**: Žliobaitė et al. (2011); 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.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 (He et al., 2020a).
- **Data spec**: tabular data.
- **Sample size**: 2K counties.
- **Year**: 2019.
- **Sensitive features**: geography.
- **Link**: https://www.edgap.org/
- **Further info**:

A.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.
- **A liation of creators**: University of Bonn.
- **Domain**: neurology.
- **Tasks in fairness literature**: robust fairness evaluation (Black and Fredrikson, 2021).
- **Data spec**: time series.
- **Sample size**: 500 individuals, each summarized by 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. (2001)

A.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.
- **A liation of creators**: Salesforce.
- **Domain**: transportation.
- **Tasks in fairness literature**: fair graph augmentation (Ramachandran et al., 2021).
- **Data spec**: location-location pairs.
- **Sample size**: $2K$ nodes (locations), connected by 8K edges (bus lines).
- **Year**: 2020.
- **Sensitive features**: race.
- **Link**: `https://github.com/salesforce/GAEA`
- **Further info**: Ramachandran et al. (2021)

A.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.
- **A liation of creators**: National Research Council Canada.
- **Domain**: linguistics.
- **Tasks in fairness literature**: fair sentiment analysis evaluation (Liang and Acuna, 2020).
- **Data spec**: text.
- **Sample size**: 9K sentences.
- **Year**: 2018.
- **Sensitive features**: race, gender.
- **Link**: `https://saifmohammad.com/WebPages/Biases-SA.html`
- **Further info**: Kiritchenko and Mohammad (2018)

A.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.
- **A liation of creators**: Stanford University.
- **Domain**: social networks.
- **Tasks in fairness literature**: fair graph mining (Li et al., 2021).
- **Data spec**: user-user pairs.
- **Sample size**: 4K people connected by 90K friend relations.
- **Year**: 2012.
- **Sensitive features**: geography, gender.
- **Link**: `https://snap.stanford.edu/data/egonets-Facebook.html`
- **Further info**: Leskovec and Mcauley (2012)

A.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.
- **A liation of creators**: University of Edinburgh.
- **Domain**: social networks.
- **Tasks in fairness literature**: fair graph mining evaluation (Kang et al., 2020).
- **Data spec**: page-page pairs.
- **Sample size**: 20K nodes (pages) connected by 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. (2021)

A.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.
- **A liation of creators**: Max Planck Institute for Human Development.
- **Domain**: computer vision, experimental psychology.
- **Tasks in fairness literature**: fairness evaluation (Kim et al., 2021).
- **Data spec**: image.
- **Sample size**: 2K images of 200 people.
- **Year**: 2010.
- **Sensitive features**: age, gender.
- **Link**: `https://faces.mpib-berlin.mpg.de/imeji/`
- **Further info**: Ebner et al. (2010)

A.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 (Cheng et al., 2021b).
{ **A liation of creators**: University of California, Los Angeles.
{ **Domain**: computer vision.
{ **Tasks in fairness literature**: fairness evaluation of private classification (Cheng et al., 2021b).
{ **Data spec**: image.
{ **Sample size**: 100K images.
{ **Year**: 2019.
{ **Sensitive features**: race, age, gender, skin tone.
{ **Link**: `https://github.com/joojs/fairface`
{ **Further info**: Karkkainen and Joo (2021)

A.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.
{ **A liation of creators**: Worcester Polytechnic Institute.
{ **Domain**: sports.
{ **Tasks in fairness literature**: fair ranking evaluation (Kuhlman et al., 2021).
{ **Data spec**: player-expert pairs.
{ **Sample size**: 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 (2020)

A.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.
{ **A liation of creators**: Zalando.
{ **Domain**: computer vision.
{ **Tasks in fairness literature**: robust fairness evaluation (Black and Fredrikson, 2021).
{ **Data spec**: image.
{ **Sample size**: 70K images across 10 product categories.
{ **Year**: 2017.
{ **Sensitive features**: none.
{ **Link**: `https://github.com/zalandoresearch/fashion-mnist`
{ **Further info**: Xiao et al. (2017)

A.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.
{ **A liation of creators**: Google; University of Texas at Austin; Toyota Technological Institute at Chicago.
{ **Domain**: finance.
{ **Tasks in fairness literature**: fairness evaluation (Hardt et al., 2016), dynamical fair classification (Liu et al., 2020), dynamical fairness evaluation (Zhang et al., 2020b; Liu et al., 2018; Creager et al., 2020), fair resource allocation (Goelz et al., 2019).
{ **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**: US Federal Reserve (2007); Hardt et al. (2016); Barocas et al. (2019)

A.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.
{ **A liation of creators**: unknown.
{ **Domain**: sports.
{ **Tasks in fairness literature**: fairness evaluation under unawareness (Awasthi et al., 2021).
{ **Data spec**: tabular data.
{ **Sample size**: 20K players.
{ **Year**: 2019.
{ **Sensitive features**: geography.
{ **Link**: `https://www.kaggle.com/stefanoleone992/fifa-20-complete-player-dataset`
{ **Further info**:

A.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.
{ **A liation of creators**: Northeastern University; Nanyang Technological University; American University of Beirut; University of Cambridge.
{ **Domain**: information systems, movies.
{ **Tasks in fairness literature**: fair ranking (Liu and Burke, 2018).
{ **Data spec**: user-movie pairs and user-user pairs.

- **Sample size**: 40K ratings by 2K users over 2K movies.
- **Year**: 2011.
- **Sensitive features**: none.
- **Link**: `https://guoguibing.github.io/librec/datasets.html`
- **Further info**: Guo et al. (2016a)

A.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.
- **A liation of creators**: National Heart, Lung, and Blood Institute (NHLBI); Boston University.
- **Domain**: cardiology.
- **Tasks in fairness literature**: fair ranking evaluation (Kallus and Zhou, 2019b).
- **Data spec**: mixture.
- **Sample size**: 15K respondents.
- **Year**: present.
- **Sensitive features**: age, sex, race.
- **Link**: `https://framinghamheartstudy.org/`
- **Further info**: Kannel and McGee (1979); Tsao and Vasan (2015)

A.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 (Bordes et al., 2013), 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.
- **A liation of creators**: Microsoft; Stanford University.
- **Domain**: information systems.
- **Tasks in fairness literature**: fair graph mining (Bose and Hamilton, 2019), fairness evaluation in graph mining (Fisher et al., 2020).
- **Data spec**: entity-relation-entity triples.
- **Sample size**: 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. (2015)

A.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 (Kocijan et al., 2020).
- **Data spec**: text.
- **Sample size**: 9K sentences.
- **Year**: 2018.
- **Sensitive features**: gender.
- **Link**: `https://github.com/google-research-datasets/gap-coreference`
- **Further info**: Webster et al. (2018)

A.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 (UCI Machine Learning Repository, 1994). 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 (2019) with an accompanying report (Grömping, 2019). See Appendix D 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 (He et al., 2020b; Sharma et al., 2020b; Raff et al., 2018; Celis et al., 2019b; Yang et al., 2020a; Donini et al., 2018; Vargo et al., 2021; Baharlouei et al., 2020; Lohaus et al., 2020; Martinez et al., 2020; Mary et al., 2019; Delobelle et al., 2020; Raff and Sylvester, 2018; Perrone et al., 2021; Sharma et al., 2021), fairness evaluation (Friedler et al., 2019; Feldman et al., 2015), fair active resource allocation (Cai et al., 2020), preference-based fair classification (Zhang et al., 2020c), fair active classification (Noriega-Campero et al., 2019), fair classification under unawareness (Kilbertus et al., 2018), robust fairness evaluation (Black and Fredrikson, 2021), fair representation learning (Ruoss et al., 2020; Louizos et al., 2016), fair reinforcement learning (Metevier et al., 2019), fair ranking evaluation (Kallus and Zhou, 2019b; Wu et al., 2018; Yang and Stoyanovich, 2017), fair ranking (Singh and Joachims, 2019; Bower et al., 2021), fair multi-stage classification (Goel et al., 2020), limited-label fair classification (Chzhen et al., 2019; Wang et al., 2021; Choi et al., 2020b), limited-label fairness evaluation (Ji et al., 2020).
- **Data spec**: tabular data.
- **Sample size**: 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 (2019)

A.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 (Papakyriakopoulos et al., 2020).
- **Data spec**: text.
- **Sample size**: 20M posts comments and tweets.
- **Year**: 2020.
- **Sensitive features**: textual references to people and their demographics.
- **Link**: not available
- **Further info**: Papakyriakopoulos et al. (2020)

A.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 (Wang et al., 2019a).
- **Affiliation of creators**: New York University; University of Washington; DeepMind.
- **Domain**: linguistics.
- **Tasks in fairness literature**: fairness evaluation (Babaeianjelodar et al., 2020; Rudinger et al., 2017), bias evaluation in language models (Cheng et al., 2021a), fairness evaluation of selective classification (Jones et al., 2021).
- **Data spec**: text.
- **Sample size**: 100 400K samples. Datasets have variable sizes spanning three orders of magnitude.
- **Year**: 2018.
- **Sensitive features**: none.
- **Link**: `https://gluebenchmark.com/`
- **Further info**: Wang et al. (2019b)

A.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 (2018). 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.
{ **A liation of creators**: University of California, San Diego.
{ **Domain**: literature, information systems.
{ **Tasks in fairness literature**: fair ranking evaluation (Raj et al., 2020), fairness evaluation (Chen et al., 2018c).
{ **Data spec**: user-book pairs.
{ **Sample size**: 200M records from 900K users over 2M books.
{ **Year**: 2019.
{ **Sensitive features**: author.
{ **Link**: `https://sites.google.com/eng.ucsd.edu/ucsdbookgraph/`
{ **Further info**: Wan and McAuley (2018)

A.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).
{ **A liation of creators**: University of California, San Diego.
{ **Domain**: information systems.
{ **Tasks in fairness literature**: fair ranking (Patro et al., 2019).
{ **Data spec**: user-business pairs.
{ **Sample size**: 10M reviews and ratings from 5M users on 3M local businesses.
{ **Year**: 2018.
{ **Sensitive features**: geography.
{ **Link**: `https://cseweb.ucsd.edu/~jmcauley/datasets.html#google_local`
{ **Further info**: He et al. (2017)

A.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.
{ **A liation of creators**: FORTH-ICS, University of Crete.
{ **Domain**: .
{ **Tasks in fairness literature**: bias discovery(Konstantakis et al., 2020).
{ **Data spec**: page-page pairs.
{ **Sample size**: 900k pages from 90k domains.
{ **Year**: 2020.
{ **Sensitive features**: none.
{ **Link**: `https://pangaia.ics.forth.gr/bias-goggles/about.html#Dataset`
{ **Further info**: Konstantakis et al. (2020)

A.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.
{ **A liation of creators**: University College Dublin.
{ **Domain**: news.
{ **Tasks in fairness literature**: data bias evaluation (Leavy et al., 2020).
{ **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. (2020)

A.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.
{ **A liation of creators**: Medical University of Vienna; University of Queensland.
{ **Domain**: dermatology.
{ **Tasks in fairness literature**: fair classification (Martinez et al., 2020).
{ **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. (2018)

A.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.
{ **A liation of creators**: Harvey Relief Handiworks; Harvey Relief Coalition.
{ **Domain**: social work.
{ **Tasks in fairness literature**: fair spatio-temporal process learning (Shang et al., 2020).
{ **Data spec**: tabular data.
{ **Sample size**: 1K help requests.
{ **Year**: 2017.
{ **Sensitive features**: geography.
{ **Link**: not available
{ **Further info**: `http://harveyrelief.handiworks.co/`

A.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.
{ **A liation 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 (Pleiss et al., 2017), fair active classification (Noriega-Campero et al., 2019).
{ **Data spec**: tabular data.
{ **Sample size**: 1K patients.
{ **Year**: 1988.
{ **Sensitive features**: age, sex.
{ **Link**: `https://archive.ics.uci.edu/ml/datasets/heart+disease`
{ **Further info**: Detrano et al. (1989)

A.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.
{ **A liation 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 (Madras et al., 2018b), fair representation learning (Louizos et al., 2016), fair classification (Raff et al., 2018; Raff and Sylvester, 2018), fair transfer learning (Madras et al., 2018a), fairness evaluation (Islam et al., 2021).
{ **Data spec**: tabular data.
{ **Sample size**: 150K patients.
{ **Year**: 2011.
{ **Sensitive features**: age, sex.
{ **Link**: `https://www.kaggle.com/c/hhp/data`
{ **Further info**: El Emam et al. (2012)

A.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 (Kleindessner et al., 2019b).
- **Data spec**: student-student pairs.
- **Sample size**: 300 students.
- **Year**: 2015.
- **Sensitive features**: gender.
- **Link**: `http://www.sociopatterns.org/datasets/high-school-contact-and-friendship-networks/`
- **Further info**: Mastrandrea et al. (2015)

A.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 (Chen et al., 2019a; Kallus et al., 2020).
- **Data spec**: tabular data.
- **Sample size**: 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.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 (Rahmattalabi et al., 2019).
- **Data spec**: person-person pairs.
- **Sample size**: 300 youth.
- **Year**: 2015.
- **Sensitive features**: age, gender, sexual orientation, race and ethnicity.
- **Link**: not available
- **Further info**: Barman-Adhikari et al. (2016)

A.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 (Liu et al., 2021).
- **Data spec**: tabular data.
- **Sample size**: $1K$ 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.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 (Celis et al., 2020b).
- **Data spec**: tabular data.
- **Sample size**: 400K students.
- **Year**: 2009.
- **Sensitive features**: gender, birth category.
- **Link**: not available
- **Further info**: Celis et al. (2020b)

A.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 (2018).
- **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 (Buolamwini and Gebru, 2018).
- **Data spec**: image.
- **Sample size**: 6K images of 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. (2015)

A.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 (Oneto et al., 2019b, 2020).
- **Data spec**: unknown.
- **Sample size**: 30K students from 140 secondary schools.
- **Year**: unknown.
- **Sensitive features**: age, sex, ethnicity.
- **Link**: not available
- **Further info**: (Nuttall et al., 1989; Goldstein, 1991)

A.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 (Caliskan et al., 2017), following the Implicit Association Test (IAT) from social psychology (Greenwald et al., 1998). 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 (Steed and Caliskan, 2021).
- **Data spec**: image.
- **Sample size**: 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 (2021)

A.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 (Prabhu and Birhane, 2020; Yang et al., 2020b; Crawford and Paglen, 2021).
- **Affiliation of creators**: Princeton University.

- **Domain**: computer vision.
- **Tasks in fairness literature**: fair classification (Dwork et al., 2018), bias discovery (Amini et al., 2019), data bias evaluation (Yang et al., 2020b), fair incremental learning (Zhao et al., 2020a), fairness evaluation (Dwork et al., 2017).
- **Data spec**: image.
- **Sample size**: 14M images depicting 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**: Deng et al. (2009); Barocas et al. (2019); Prabhu and Birhane (2020); Yang et al. (2020b); Crawford and Paglen (2021)

A.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 (Tjong Kim Sang and De Meulder, 2003), 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.
- **A liation of creators**: Twitter.
- **Domain**: linguistics.
- **Tasks in fairness literature**: fairness evaluation in entity recognition (Mishra et al., 2020).
- **Data spec**: text.
- **Sample size**: 50K sentences.
- **Year**: 2020.
- **Sensitive features**: gender, race and religion.
- **Link**: `https://github.com/napsternxg/NER_bias`
- **Further info**: Mishra et al. (2020)

A.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.
- **A liation of creators**: California Institute of Technology; University of Edinburgh; Google; Cornell University; iNaturalist.
- **Domain**: biology.
- **Tasks in fairness literature**: fairness evaluation of private classification (Bagdasaryan et al., 2019).
- **Data spec**: image.
- **Sample size**: 3M images from 10K different species of plants and animals.
- **Year**: 2021.
- **Sensitive features**: none.
- **Link**: `https://github.com/visipedia/inat_comp`
- **Further info**: Van Horn et al. (2018); Van Horn et al. (2021)

A.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 (Pujol et al., 2020).
- **Data spec**: tabular data.
- **Sample size**: 30 state.
- **Year**: unknown.
- **Sensitive features**: geography.
- **Link**: https://www.indiabudget.gov.in/budget_archive/es2006-07/chapt2007/tab97.pdf
- **Further info**:

A.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 (Belitz et al., 2021).
- **Data spec**: tabular data.
- **Sample size**: 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. (2018)

A.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 (Madras et al., 2019; Yi et al., 2019).
- **Data spec**: mixture.
- **Sample size**: 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. (1992)

A.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.
{ **A liation of creators**: Columbia University.
{ **Domain**: social media.
{ **Tasks in fairness literature**: fair advertising (Riederer and Chaintreau, 2017).
{ **Data spec**: unknown.
{ **Sample size**: 1M photos from 40K users.
{ **Year**: 2017.
{ **Sensitive features**: race, gender, geography.
{ **Link**: not available
{ **Further info**: Riederer and Chaintreau (2017)

A.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.
{ **A liation of creators**: University College Dublin.
{ **Domain**: pattern recognition.
{ **Tasks in fairness literature**: fair anomaly detection (Shekhar et al., 2021).
{ **Data spec**: tabular data.
{ **Sample size**: 3K image encodings.
{ **Year**: 1998.
{ **Sensitive features**: none.
{ **Link**: `https://archive.ics.uci.edu/ml/datasets/internet+advertisements`
{ **Further info**: Kushmerick (1999)

A.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.
{ **A liation of creators**: Missouri Botanical Garden; Washington University.
{ **Domain**: plant science.

- **Tasks in fairness literature**: fair clustering (Chen et al., 2019b; Abbasi et al., 2021).
- **Data spec**: tabular data.
- **Sample size**: 100 samples from three species of Iris.
- **Year**: 1988.
- **Sensitive features**: none.
- **Link**: `https://archive.ics.uci.edu/ml/datasets/iris`
- **Further info**: (Anderson, 1936; Fisher, 1936)

A.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.
- **A liation of creators**: University of Padua; Carnegie Mellon University; University of Udine.
- **Domain**: economics.
- **Tasks in fairness literature**: fair pricing evaluation (Fabris et al., 2021).
- **Data spec**: tabular data.
- **Sample size**: 2K driver profiles.
- **Year**: 2021.
- **Sensitive features**: gender, birthplace.
- **Link**: not available
- **Further info**: Fabris et al. (2021)

A.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.
- **A liation of creators**: Massachusetts Institute of Technology.
- **Domain**: computer networks.
- **Tasks in fairness literature**: fair clustering (Chen et al., 2019b).
- **Data spec**: tabular data.
- **Sample size**: 7M connections.
- **Year**: 1999.
- **Sensitive features**: none.
- **Link**: `http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html`
- **Further info**: Tavallaee et al. (2009)

A.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.
{ **A liation of creators**: Université de Montréal; Polytechnique de Montréal.
{ **Domain**: public health.
{ **Tasks in fairness literature**: fair matching evaluation (Farnadi et al., 2019).
{ **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 (2019)

A.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.
{ **A liation of creators**: unknown.
{ **Domain**: public health.
{ **Tasks in fairness literature**: fairness matching evaluation (Mattei et al., 2018b).
{ **Data spec**: tabular data.
{ **Sample size**: unknown.
{ **Year**: 2018.
{ **Sensitive features**: age, geography, blood type.
{ **Link**: not available
{ **Further info**: Mattei et al. (2018a)

A.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.
{ **A liation of creators**: Kiva; DePaul University.
{ **Domain**: finance.
{ **Tasks in fairness literature**: fair ranking (Burke et al., 2018b; Liu and Burke, 2018; Sonboli et al., 2020), bias discovery (Sonboli and Burke, 2019).
{ **Data spec**: tabular data.
{ **Sample size**: 1M transactions involving 100K loans and 200K users.
{ **Year**: 2018.
{ **Sensitive features**: gender, geography, activity.
{ **Link**: not available
{ **Further info**: Sonboli and Burke (2019)

A.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.
- **A liation of creators**: University of Massachussets, Amherst; Stony Brook University.
- **Domain**: computer vision.
- **Tasks in fairness literature**: fair data summarization (Samadi et al., 2018), fair clustering (Ghadiri et al., 2021), robust fairness evaluation (Black and Fredrikson, 2021), fairness evaluation (Segal et al., 2021).
- **Data spec**: image.
- **Sample size**: 13K face images of 6K individuals.
- **Year**: 2007.
- **Sensitive features**: gender, age, race.
- **Link**: `http://vis-www.cs.umass.edu/lfw/`
- **Further info**: Huang et al. (2007); Han and Jain (2014); Gebru et al. (2018)

A.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.
- **A liation of creators**: Stanford University.
- **Domain**: linguistics.
- **Tasks in fairness literature**: fair sentiment analysis evaluation (Liang and Acuna, 2020).
- **Data spec**: text.
- **Sample size**: 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. (2011)

A.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).
- **A liation of creators**: Barcelona Music and Audio Technologies; Universitat Pompeu Fabra.
- **Domain**: music, information systems.
- **Tasks in fairness literature**: fair ranking evaluation (Ekstrand et al., 2018).
- **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 (2010)

A.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.
{ **A liation of creators**: Capital One.
{ **Domain**: news.
{ **Tasks in fairness literature**: data bias evaluation (Florez, 2019).
{ **Data spec**: text.
{ **Sample size**: 60K articles.
{ **Year**: 2019.
{ **Sensitive features**: textual references to people and their demographics.
{ **Link**: not available
{ **Further info**: Florez (2019)

A.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.
{ **A liation of creators**: Law School Admission Council (LSAC).
{ **Domain**: education.
{ **Tasks in fairness literature**: fair classification (Yang et al., 2020a; Cho et al., 2020; Russell et al., 2017; Agarwal et al., 2018; Berk et al., 2017), rich-subgroup fairness evaluation (Kearns et al., 2019), fair classification under unawareness (Lahoti et al., 2020; Lamy et al., 2019), fairness evaluation (Black et al., 2020; Kusner et al., 2017), fair regression (Chzhen et al., 2020a,b; Agarwal et al., 2019; Komiyama et al., 2018), fair representation learning (Ruoss et al., 2020), robust fair classification (Mandal et al., 2020), limited-label fair classification (Wang et al., 2021).
{ **Data spec**: tabular data.
{ **Sample size**: 20K examinees.
{ **Year**: 1998.
{ **Sensitive features**: sex, race, age.
{ **Link**: not available
{ **Further info**: Wightman et al. (1998)

A.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.
{ **A liation of creators**: Charles University in Prague; Libimseti.
{ **Domain**: sociology, information systems.

- **Tasks in fairness literature**: fair matching (Tziavelis et al., 2019).
- **Data spec**: user-user pairs.
- **Sample size**: 10M ratings over 200K users.
- **Year**: 2007.
- **Sensitive features**: gender.
- **Link**: `http://colfi.wz.cz/`
- **Further info**: Brožovský (2006); Brozovsky and Petricek (2007)

A.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.
- **A liation of creators**: Los Angeles City Attorney's Office; University of Chicago.
- **Domain**: law.
- **Tasks in fairness literature**: fair classification (Rodolfa et al., 2020).
- **Data spec**: tabular data.
- **Sample size**: 1M unique individuals associated with 2M cases.
- **Year**: 2020.
- **Sensitive features**: race, ethnicity.
- **Link**: not available
- **Further info**: (Rodolfa et al., 2020)

A.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.
- **A liation of creators**: Agency for Healthcare Research and Quality.
- **Domain**: health policy.
- **Tasks in fairness literature**: fair transfer learning (Coston et al., 2019), fair regression (Romano et al., 2020), fairness evaluation (Singh and Ramamurthy, 2019), robust fair classification (Biswas and Mukherjee, 2021), fair classification (Sharma et al., 2021).
- **Data spec**: tabular data.
- **Sample size**: 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.114 MGGG States

{ **Description**: developed by the Metric Geometry and Gerrymandering Group[16], 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[17], demographic data from IPUMS[18] and election data from MIT Election and Data Science [19]. Source and precise data format vary by state.
{ **A liation of creators**: Tufts University.
{ **Domain**: political science.
{ **Tasks in fairness literature**: fair districting for electoral precincts (Schutzman, 2020).
{ **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.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 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.
{ **A liation of creators**: Microsoft.
{ **Domain**: information systems.
{ **Tasks in fairness literature**: fair ranking (Bower et al., 2021).
{ **Data spec**: query document pairs.
{ **Sample size**: 30K queries.
{ **Year**: 2013.
{ **Sensitive features**: none.
{ **Link**: https://www.microsoft.com/en-us/research/project/mslr/
{ **Further info**: (Qin and Liu, 2013)

A.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 (Knees and Hübler, 2019).
{ **A liation of creators**: Spotify; Johannes Kepler University; University of Massachusetts.
{ **Domain**: music, information systems.
{ **Tasks in fairness literature**: data bias evaluation (Knees and Hübler, 2019).
{ **Data spec**: tabular data.

[16] https://mggg.org/
[17] https://www.census.gov/geographies/mapping-files.html
[18] https://www.nhgis.org/
[19] https://electionlab.mit.edu/

- **Sample size**: 1M playlists containing 2M unique tracks by 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. (2018b)

A.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.
- **A liation of creators**: Columbia University; The Echo Nest.
- **Domain**: music, information systems.
- **Tasks in fairness literature**: dynamical evaluation of fair ranking (Ferraro et al., 2019).
- **Data spec**: user-song pairs.
- **Sample size**: 50M play counts over 1M users and 400K songs.
- **Year**: 2012.
- **Sensitive features**: artist; geography.
- **Link**: `http://millionsongdataset.com/`; `https://www.kaggle.com/c/msdchallenge`
- **Further info**: Bertin-Mahieux et al. (2011); McFee et al. (2012)

A.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.
- **A liation 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 (Cheng et al., 2021b).
- **Data spec**: images.
- **Sample size**: 400K images of 70K patients.
- **Year**: 2019.
- **Sensitive features**: sex.
- **Link**: `https://physionet.org/content/mimic-cxr-jpg/2.0.0/`
- **Further info**: (Johnson et al., 2019)

A.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.

{ **A liation of creators**: Massachusetts Institute of Technology; Beth Israel Deaconess Medical Center; A*STAR.
{ **Domain**: critical care medicine.
{ **Tasks in fairness literature**: fair classification (Martinez et al., 2020), fairness evaluation (Chen et al., 2018c; Zhang et al., 2020a), robust fair classification (Singh et al., 2021).
{ **Data spec**: mixture.
{ **Sample size**: 60K patients.
{ **Year**: 2016.
{ **Sensitive features**: age, ethnicity, gender.
{ **Link**: `https://mimic.mit.edu/`
{ **Further info**: Johnson et al. (2016)

A.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 (Brockman et al., 2016), 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.
{ **A liation of creators**: Google.
{ **Domain**: N/A.
{ **Tasks in fairness literature**: dynamical fair resource allocation (Atwood et al., 2019; D'Amour et al., 2020), dynamical fair classification (D'Amour et al., 2020).
{ **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. (2020)

A.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.
{ **A liation of creators**: AT&T Labs.
{ **Domain**: computer vision.
{ **Tasks in fairness literature**: fair clustering (Har-Peled and Mahabadi, 2019; Li et al., 2020a), fair anomaly detection (Zhang and Davidson, 2021), fair classification (Creager et al., 2021), fairness evaluation (Segal et al., 2021).
{ **Data spec**: image.
{ **Sample size**: 70K images across 10 digits.
{ **Year**: 1998.
{ **Sensitive features**: none.
{ **Link**: `http://yann.lecun.com/exdb/mnist/`

- **Further info**: Lecun et al. (1998); Barocas et al. (2019)
- **Variants**:
  - MNIST-USPS (Li et al., 2020a): merge with USPS dataset of handwritten digits (Hull, 1994).
  - Color-reverse MNIST (Li et al., 2020a) or MNIST-Invert (Zhang and Davidson, 2021): images from MNIST, reversed via $p = 255 \quad p$ for each pixel p.
  - Color MNIST (Arjovsky et al., 2020): 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.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.
- **A liation of creators**: unknown.
- **Domain**: finance.
- **Tasks in fairness literature**: fair transfer learning (Coston et al., 2019).
- **Data spec**: tabular data.
- **Sample size**: 200K people.
- **Year**: unknown.
- **Sensitive features**: age, gender.
- **Link**: not available
- **Further info**: Speakman et al. (2018)

A.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.
- **A liation of creators**: University of Minnesota.
- **Domain**: information systems, movies.
- **Tasks in fairness literature**: fair ranking (Burke et al., 2018b; Sonboli et al., 2020; Dickens et al., 2020; Farnadi et al., 2018; Liu and Burke, 2018), fair ranking evaluation (Ekstrand et al., 2018; Yao and Huang, 2017a,b), fair data summarization (El Halabi et al., 2020), fair representation learning (Oneto et al., 2020, 2019b), fair graph mining (Buyl and De Bie, 2020; Bose and Hamilton, 2019), fair data generation (Burke et al., 2018a).
- **Data spec**: user-movie pairs.
- **Sample size**: 1M reviews by 6K users over 4K movies.
- **Year**: 2003.
- **Sensitive features**: gender, age.
- **Link**: `https://grouplens.org/datasets/movielens/1m/`
- **Further info**: Harper and Konstan (2015)

A.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)" (Guo et al., 2016b). 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 (§ A.153) and BUPT Faces (§ A.27). The dataset was recently augmented with gender and nationality data automatically inferred from biographies of people (McDuff et al., 2019). From nationality, a race-related attribute was also annotated on a subset of 20,000 images.
- **A liation of creators**: Microsoft.
- **Domain**: computer vision.
- **Tasks in fairness literature**: fairness evaluation through artificial data generation (McDuff et al., 2019).
- **Data spec**: image.
- **Sample size**: 10M images representing 100K people.
- **Year**: 2016.
- **Sensitive features**: gender, race, geography.
- **Link**: not available
- **Further info**: Guo et al. (2016b); McDuff et al. (2019); Murgia (2019)

A.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 (Zhao et al., 2017; Hendricks et al., 2018).
- **A liation 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 (David et al., 2020), fair classification (Hendricks et al., 2018).
- **Data spec**: image.
- **Sample size**: 300K images.
- **Year**: 2014.
- **Sensitive features**: gender.
- **Link**: `https://cocodataset.org/`
- **Further info**: Lin et al. (2014)

A.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 (Yi et al., 2013), 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.
{ **A liation of creators**: The Chinese University of Hong Kong.
{ **Domain**: computer vision.
{ **Tasks in fairness literature**: fair clustering (Li et al., 2020a).
{ **Data spec**: image.
{ **Sample size**: 10K images.
{ **Year**: 2014.
{ **Sensitive features**: gender.
{ **Link**: `http://mmlab.ie.cuhk.edu.hk/projects/TCDCN.html`
{ **Further info**: Zhang et al. (2014, 2015)

A.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.
{ **A liation of creators**: US Bureau of Labor Statistics.
{ **Domain**: demography.
{ **Tasks in fairness literature**: fair regression (Komiyama et al., 2018; Chzhen et al., 2020b,a).
{ **Data spec**: tabular data.
{ **Sample size**: 10K respondents (NLSY79); 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.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.
{ **A liation 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 (Ustun et al., 2019).
{ **Data spec**: image.
{ **Sample size**: 50K participants.
{ **Year**: 2020.
{ **Sensitive features**: age, ethnicity, race, sex.
{ **Link**: `https://cdas.cancer.gov/nlst/`
{ **Further info**: NLST Trial Research Team (2011); `https://www.cancer.gov/types/lung/research/nlst`

A.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).
{ **A liation of creators**: The New York Times.
{ **Domain**: news.
{ **Tasks in fairness literature**: bias evaluation in WEs (Brunet et al., 2019).
{ **Data spec**: text.
{ **Sample size**: 2M articles.
{ **Year**: 2008.
{ **Sensitive features**: textual references to people and their demographics.
{ **Link**: `https://catalog.ldc.upenn.edu/LDC2008T19`
{ **Further info**:

A.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`.
{ **A liation of creators**: University of Amsterdam.
{ **Domain**: literature.
{ **Tasks in fairness literature**: fairness evaluation (Koolen and van Cranenburgh, 2017).
{ **Data spec**: text.
{ **Sample size**: 50 novels.
{ **Year**: 2017.
{ **Sensitive features**: gender, geography (of author).
{ **Link**: not available
{ **Further info**: Koolen and van Cranenburgh (2017); Koolen (2018)

A.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.
{ **A liation of creators**: North Carolina State Board of Elections.
{ **Domain**: political science.
{ **Tasks in fairness literature**: data bias evaluation (Coston et al., 2021), fair clustering (Abbasi et al., 2021), fairness evaluation of advertisement (Speicher et al., 2018a).
{ **Data spec**: tabular data.
{ **Sample size**: 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.[20].

[20] `https://l2-data.com/states/north-carolina/`

A.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.
{ **A liation 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 (Romano et al., 2020).
{ **Data spec**: tabular data.
{ **Sample size**: 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. (1989)

A.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.
{ **A liation of creators**: University of Illinois.
{ **Domain**: transportation.
{ **Tasks in fairness literature**: fair matching (Lesmana et al., 2019; Nanda et al., 2020).
{ **Data spec**: tabular data.
{ **Sample size**: 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.[21].

A.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

[21] `http://www.andresmh.com/nyctaxitrips/`

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.

{ **A liation of creators**: Yale Universiy.
{ **Domain**: information systems.
{ **Tasks in fairness literature**: fair subset selection under unawareness (Mehrotra and Celis, 2021).
{ **Data spec**: image.
{ **Sample size**: 10K images of 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 (2020)

A.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.
{ **A liation of creators**: University of California, Berkeley.
{ **Domain**: computer vision.
{ **Tasks in fairness literature**: fair clustering (Li et al., 2020a).
{ **Data spec**: image.
{ **Sample size**: 4K images.
{ **Year**: 2011.
{ **Sensitive features**: none.
{ **Link**: `https://paperswithcode.com/dataset/office-31`
{ **Further info**: Saenko et al. (2010)

A.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.
{ **A liation of creators**: unknown.
{ **Domain**: sports.
{ **Tasks in fairness literature**: fair clustering (Huang et al., 2019).
{ **Data spec**: tabular data.
{ **Sample size**: 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.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 (Li et al., 2020b).
- **Data spec**: image.
- **Sample size**: 2K images from 50 different alphabets.
- **Year**: 2019.
- **Sensitive features**: none.
- **Link**: `https://github.com/brendenlake/omniglot`
- **Further info**: Lake et al. (2015)

A.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[22], 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 (Peters et al., 2018) were trained on this benchmark.
- **Affiliation of creators**: Google; University of Edinburgh; Cantab Research Ltd.
- **Domain**: linguistics.
- **Tasks in fairness literature**: data bias evaluation (Tan and Celis, 2019).
- **Data spec**: text.
- **Sample size**: 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. (2014)

A.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.
- **Domain**: information systems.
- **Tasks in fairness literature**: fairness evaluation (Hannák et al., 2017).
- **Data spec**: query-result pairs.
- **Sample size**: 10K workers (Fiverr); 4K (TaskRabbit).
- **Year**: 2017.
- **Sensitive features**: gender, race.
- **Link**: not available
- **Further info**: Hannák et al. (2017)

[22] `http://statmt.org/wmt11/training-monolingual.tgz`

A.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 (Kuznetsova et al., 2020), visual relationships (Kuznetsova et al., 2020), and validation+test segmentations are drawn fully manually (Benenson et al., 2019); while segmentations in train are drawn using an interactive algorithm (Benenson et al., 2019). 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 (Pont-Tuset et al., 2020)).
{ **A liation of creators**: Google.
{ **Domain**: computer vision.
{ **Tasks in fairness literature**: data bias evaluation (Schumann et al., 2021), fairness evaluation (Aka et al., 2021).
{ **Data spec**: image.
{ **Sample size**: 9M images.
{ **Year**: 2020.
{ **Sensitive features**: gender, age.
{ **Link**: `https://storage.googleapis.com/openimages/web/index.html`
{ **Further info**: (Kuznetsova et al., 2020; Schumann et al., 2021; Benenson et al., 2019; Pont-Tuset et al., 2020)

A.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.
{ **A liation of creators**: unknown.
{ **Domain**: library and information sciences.
{ **Tasks in fairness literature**: fair matching (Kobren et al., 2019).
{ **Data spec**: paper-reviewer pairs.
{ **Sample size**: 200 reviewers for 100 papers (MIDL); 1K reviewers for 3K papers (CVPR). 3K reviewers for 5K papers (CVPR2018).
{ **Year**: 2019.
{ **Sensitive features**: none.
{ **Link**: not available
{ **Further info**: Kobren et al. (2019)

A.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.
- **A liation of creators**: Philadelphia Police Department.
- **Domain**: law.
- **Tasks in fairness literature**: fair resource allocation (Elzayn et al., 2019).
- **Data spec**: tabular data.
- **Sample size**: 1M crime incidents.
- **Year**: present.
- **Sensitive features**: geography.
- **Link**: `https://www.opendataphilly.org/dataset/crime-incidents`
- **Further info**:

A.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.
- **A liation of creators**: Massachusetts Institute of Technology; Microsoft.
- **Domain**: computer vision.
- **Tasks in fairness literature**: fair classification (Kim et al., 2019; Amini et al., 2019), fairness evaluation (Buolamwini and Gebru, 2018; Raji and Buolamwini, 2019), bias discovery (Kim et al., 2019; Amini et al., 2019).
- **Data spec**: image.
- **Sample size**: 1K images of 1K individuals.
- **Year**: 2018.
- **Sensitive features**: gender, skin type.
- **Link**: `http://gendershades.org/`
- **Further info**: Buolamwini and Gebru (2018)

A.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.
- **A liation of creators**: Logistics Management Institute; National Institute of Diabetes Digestive and Kidney Diseases; John Hopkins University.
- **Domain**: endocrinology.
- **Tasks in fairness literature**: fairness evaluation (Sharma et al., 2020a), fair clustering (Chen et al., 2019b).
- **Data spec**: tabular data.
- **Sample size**: 800 subjects.

- **Year**: 2016.
- **Sensitive features**: age.
- **Link**: `https://www.kaggle.com/uciml/pima-indians-diabetes-database`
- **Further info**: Smith et al. (1988); Radin (2017)

A.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 (El Halabi et al., 2020).
- **Data spec**: user-user pairs.
- **Sample size**: 2M nodes (profiles) connected by 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 (2012)

A.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 (Yurochkin et al., 2020; Mukherjee et al., 2020), bias discovery in WEs (Swinger et al., 2019).
- **Data spec**: tabular data.
- **Sample size**: 3K unique names (NYC names); 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.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 (Noriega-Campero et al., 2020).
- **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.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.
{ **A liation of creators**: Stanford University; Chan Zuckerberg Biohub.
{ **Domain**: biology.
{ **Tasks in fairness literature**: fair graph mining (Kang et al., 2020).
{ **Data spec**: protein-protein pairs.
{ **Sample size**: 20K proteins (nodes) linked by 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. (2018)

A.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.
{ **A liation of creators**: Prosper; University College Dublin.
{ **Domain**: finance.
{ **Tasks in fairness literature**: fair classification (Li et al., 2020c).
{ **Data spec**: lender-borrower pairs.
{ **Sample size**: 3M loan records involving 100K people.
{ **Year**: 2015.
{ **Sensitive features**: none.
{ **Link**: `http://mlg.ucd.ie/datasets/prosper.html`
{ **Further info**: Redmond and Cunningham (2013)

A.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.
{ **A liation of creators**: University of Maryland.
{ **Domain**: library and information sciences.
{ **Tasks in fairness literature**: fair graph mining (Li et al., 2021).

- **Data spec**: article-article pairs.
- **Sample size**: 20K articles connected by 40K citations.
- **Year**: 2020.
- **Sensitive features**: none.
- **Link**: `https://linqs.soe.ucsc.edu/data`
- **Further info**: Namata et al. (2012)

A.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 (Wilson et al., 2021).
- **Data spec**: tabular data.
- **Sample size**: 10K users.
- **Year**: 2021.
- **Sensitive features**: gender, race.
- **Link**: not available
- **Further info**: Wilson et al. (2021)

A.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 (Ball-Burack et al., 2021).
- **Data spec**: text.
- **Sample size**: 5M tweets from 4K users.
- **Year**: 2018.
- **Sensitive features**: race, gender, age.
- **Link**: `http://www.preotiuc.ro/`
- **Further info**: Preoţiuc-Pietro and Ungar (2018)

A.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 (§ A.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.

- **A liation of creators**: Beijing University of Posts; Telecommunications and Canon Information Technology (Beijing).
- **Domain**: computer vision.
- **Tasks in fairness literature**: fair reinforcement learning (Wang and Deng, 2020), fair representation learning (Gong et al., 2021).
- **Data spec**: image.
- **Sample size**: 50K images of 10K individuals.
- **Year**: 2019.
- **Sensitive features**: race (inferrred).
- **Link**: `http://www.whdeng.cn/RFW/testing.html`
- **Further info**: Wang et al. (2019e)

A.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$.
- **A liation of creators**: National Institute of Justice.
- **Domain**: law.
- **Tasks in fairness literature**: fair spatio-temporal process learning (Shang et al., 2020).
- **Data spec**: tabular data.
- **Sample size**: 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 (2018)

A.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 10K probation cases "representative of 79,043 probationers".
- **A liation of creators**: US Department of Justice; National Association of Criminal Justice Planners.
- **Domain**: law.
- **Tasks in fairness literature**: limited-label fair classification (Wang and Saar-Tsechansky, 2020).
- **Data spec**: tabular data.
- **Sample size**: 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.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.
{ **A liation of creators**: Pushshift data.
{ **Domain**: social media, linguistics.
{ **Tasks in fairness literature**: bias evaluation in language models (Guo and Caliskan, 2021).
{ **Data spec**: text.
{ **Sample size**: 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 (2021)

A.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.
{ **A liation of creators**: New York University; New York University Langone Medical Center.
{ **Domain**: nephrology.
{ **Tasks in fairness literature**: fairness evaluation (Williams and Razavian, 2019).
{ **Data spec**: tabular data.
{ **Sample size**: 2M patients.
{ **Year**: 2019.
{ **Sensitive features**: age, gender, race.
{ **Link**: not available
{ **Further info**: Williams and Razavian (2019)

A.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.
{ **A liation of creators**: University of the Aegean.
{ **Domain**: news.
{ **Tasks in fairness literature**: fair clustering (Harb and Lam, 2020).
{ **Data spec**: text.
{ **Sample size**: 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 (2006)

A.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.
{ **A liation of creators**: Haverford College.
{ **Domain**: law.
{ **Tasks in fairness literature**: fairness evaluation (Feldman et al., 2015; Friedler et al., 2019), limited-label fairness evaluation (Ji et al., 2020).
{ **Data spec**: tabular data.
{ **Sample size**: 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 (2009); Miao (2010)

A.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.
{ **A liation of creators**: MPI-SWS; Rice University; Northeastern University.
{ **Domain**: social networks.
{ **Tasks in fairness literature**: fair graph diffusion (Ali et al., 2019a).
{ **Data spec**: user-user pairs.
{ **Sample size**: 1K profiles connected by 40K edges.
{ **Year**: 2010.
{ **Sensitive features**: none.
{ **Link**: not available
{ **Further info**: Mislove et al. (2010)

A.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.
{ **A liation of creators**: Huygens ING – KNAW; University of Amsterdam; Fryske Akademy.
{ **Domain**: literature.
{ **Tasks in fairness literature**: fairness evaluation (Koolen and van Cranenburgh, 2017).
{ **Data spec**: text.
{ **Sample size**: 400 novels.

- **Year**: 2017.
- **Sensitive features**: gender (of author).
- **Link**: not available
- **Further info**: Koolen and van Cranenburgh (2017); `https://literaryquality.huygens.knaw.nl/`

A.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 (Sühr et al., 2019).
- **Data spec**: driver-job pairs.
- **Sample size**: 1K drivers handling 200K job requests.
- **Year**: 2019.
- **Sensitive features**: geography.
- **Link**: not available
- **Further info**: Sühr et al. (2019)

A.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.[23] 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 (Babaeianjelodar et al., 2020).
- **Data spec**: text.
- **Sample size**: 2M posts with 25M comments.
- **Year**: 2018.
- **Sensitive features**: gender.[24].
- **Link**: `https://nlp.stanford.edu/robvoigt/rtgender/`
- **Further info**: (Voigt et al., 2018)

[23] e.g. `https://en.wikipedia.org/wiki/Category:American_female_tennis_players`

[24] Annotations for Facebook and TED come from Wikipedia and Mirkin et al. (2015) respectively. Reddit and Fitocracy rely on self-reported labels.

A.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 (Squire, 2019) and externally (Coston et al., 2021).
{ **A liation of creators**: Safegraph.
{ **Domain**: urban studies.
{ **Tasks in fairness literature**: data bias evaluation (Coston et al., 2021).
{ **Data spec**: mixture.
{ **Sample size**: 7M POI.
{ **Year**: present.
{ **Sensitive features**: geography.
{ **Link**: `https://www.safegraph.com/academics`
{ **Further info**: `https://docs.safegraph.com/v4.0/docs`

A.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.
{ **A liation of creators**: École Polytechnique Fédérale de Lausanne (EPFL); Microsoft; University of California, Berkeley.
{ **Domain**: information systems.
{ **Tasks in fairness literature**: fair data summarization (Celis et al., 2016, 2018).
{ **Data spec**: image.
{ **Sample size**: 800 images.
{ **Year**: 2016.
{ **Sensitive features**: male/female.
{ **Link**: `goo.gl/hNukfP`
{ **Further info**: Celis et al. (2016)

A.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.
{ **A liation of creators**: US Census Bureau.
{ **Domain**: demography.
{ **Tasks in fairness literature**: fairness evaluation of private resource allocation (Pujol et al., 2020).

{ **Data spec**: tabular data.
{ **Sample size**: 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.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.
{ **A liation of creators**: Stanford University.
{ **Domain**: social media.
{ **Tasks in fairness literature**: fair federated learning (Li et al., 2020b).
{ **Data spec**: text.
{ **Sample size**: 2M tweets by 600K accounts.
{ **Year**: 2012.
{ **Sensitive features**: textual references to people and their demographics.
{ **Link**: `http://help.sentiment140.com/home`
{ **Further info**: Go et al. (2009)

A.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.
{ **A liation of creators**: Sunchon National University.
{ **Domain**: transportation.
{ **Tasks in fairness literature**: fair regression (Diana et al., 2021).
{ **Data spec**: time series.
{ **Sample size**: 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 (2020); V E et al. (2020), `https://data.seoul.go.kr/index.do`

A.169 Shakespeare

{ **Description**: this dataset is available as part of the LEAF benchmark for federated learning (Caldas et al., 2018). 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.

{ **A liation of creators**: Google; Carnegie Mellon University; Determined AI.
{ **Domain**: literature.
{ **Tasks in fairness literature**: fair federated learning (Li et al., 2020b).
{ **Data spec**: text.
{ **Sample size**: 4M tokens over 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**: McMahan et al. (2017); Caldas et al. (2018)

A.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.
{ **A liation of creators**: Shanghai Jiao Tong University; CITI-INRIA Lab.
{ **Domain**: transportation.
{ **Tasks in fairness literature**: fair routing (Qian et al., 2015).
{ **Data spec**: unknown.
{ **Sample size**: unknown.
{ **Year**: 2015.
{ **Sensitive features**: geography.
{ **Link**: not available
{ **Further info**: Qian et al. (2015)

A.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.
{ **A liation of creators**: DeepMind; Wayve.
{ **Domain**: computer vision.
{ **Tasks in fairness literature**: fair representation learning (Locatello et al., 2019), fair data generation (Choi et al., 2020a).
{ **Data spec**: image.
{ **Sample size**: 500K images.
{ **Year**: 2018.
{ **Sensitive features**: none.
{ **Link**: `https://github.com/deepmind/3d-shapes`
{ **Further info**: Kim and Mnih (2018)

A.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.[25] This dataset was annotated automatically with a binary Fitzpatrick skin tone label (Cheng et al., 2021b).

[25] `https://www.kaggle.com/c/siim-isic-melanoma-classification`

{ **A liation 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 (Cheng et al., 2021b).
{ **Data spec**: image.
{ **Sample size**: 30K images of 2K patients.
{ **Year**: 2020.
{ **Sensitive features**: skin type.
{ **Link**: urlhttps://doi.org/10.34970/2020-ds01
{ **Further info**: Rotemberg et al. (2021)

A.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.
{ **A liation of creators**: New York University; NEC Labs America.
{ **Domain**: computer vision.
{ **Tasks in fairness literature**: fair representation learning (Locatello et al., 2019).
{ **Data spec**: image.
{ **Sample size**: 100K images.
{ **Year**: 2005.
{ **Sensitive features**: none.
{ **Link**: `https://cs.nyu.edu/~ylclab/data/norb-v1.0-small/`
{ **Further info**: LeCun et al. (2004)

A.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.
{ **A liation of creators**: Spliddit.
{ **Domain**: economics.
{ **Tasks in fairness literature**: fair preferece-based resource allocation (Babaioff et al., 2019).
{ **Data spec**: tabular data.
{ **Sample size**: 1K division instances.
{ **Year**: 2016.
{ **Sensitive features**: none.
{ **Link**: not available
{ **Further info**: Caragiannis et al. (2016); `http://www.spliddit.org/apps/goods`

A.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.
- **A liation of creators**: Stanford University.
- **Domain**: medicine.
- **Tasks in fairness literature**: fair risk assessment (Pfohl et al., 2019).
- **Data spec**: mixture.
- **Sample size**: 3M individuals.
- **Year**: present.
- **Sensitive features**: race, ethnicity, gender, age.
- **Link**: `https://starr.stanford.edu/`
- **Further info**: Lowe et al. (2009); Datta et al. (2020)

A.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.
- **A liation of creators**: US Department of Justice.
- **Domain**: law.
- **Tasks in fairness literature**: fairness evaluation of multi-stage classification (Green and Chen, 2019).
- **Data spec**: tabular data.
- **Sample size**: 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.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.
- **A liation of creators**: Hong Kong University of Science and Technology; WeBank.
- **Domain**: social media.
- **Tasks in fairness literature**: fairness evaluation (Xiao et al., 2019).
- **Data spec**: user-post pairs.
- **Sample size**: 50K users interacting over 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. (2019)

A.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.
- **A liation of creators**: New York Police Department.
- **Domain**: law.
- **Tasks in fairness literature**: preference-based fair classification (Zafar et al., 2017b), robust fair classification (Kallus and Zhou, 2018), fair classification under unawareness (Kilbertus et al., 2018), fairness evaluation (Goel et al., 2017), fair classification (Ali et al., 2021).
- **Data spec**: tabular data.
- **Sample size**: 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. (2007); Goel et al. (2016)

A.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.
- **A liation of creators**: Chicago Police Department; Illinois Institute of Technology.
- **Domain**: law.
- **Tasks in fairness literature**: fairness evaluation (Black et al., 2020).
- **Data spec**: tabular data.
- **Sample size**: 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. (2019)

A.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, socioecononomical, 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 (Chzhen et al., 2020b,a; Heidari et al., 2019b), rich-subgroup fairness evaluation (Kearns et al., 2019), fair data summarization (Jones et al., 2020; Belitz et al., 2021).
- **Data spec**: tabular data.
- **Sample size**: 600 students.
- **Year**: 2014.
- **Sensitive features**: sex, age.
- **Link**: `https://archive.ics.uci.edu/ml/datasets/student+performance`
- **Further info**: Cortez and Silva (2008)

A.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 (Chiplunkar et al., 2020).
- **Data spec**: user-sushi pairs.
- **Sample size**: 5K respondents.
- **Year**: 2016.
- **Sensitive features**: gender, age, geography.
- **Link**: `https://www.kamishima.net/%20sushi/`
- **Further info**: Kamishima (2003)

A.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,[26] 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 (Sabato and Yom-Tov, 2020).
- **Data spec**: tabular data.
- **Sample size**: statistics for 20 cancer types across 50 US states.
- **Year**: 2020.
- **Sensitive features**: geography.
- **Link**: `https://github.com/sivansabato/bfa/blob/master/cancer_data.m`
- **Further info**: Sabato and Yom-Tov (2020)

[26] `https://gis.cdc.gov/Cancer/USCS/DataViz.html`

A.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++ (Zhu et al., 2018).
- **A liation of creators**: Texas A&M University.
- **Domain**: social media.
- **Tasks in fairness literature**: fair ranking (Zhu et al., 2018).
- **Data spec**: user-topic pairs.
- **Sample size**: 10K Twitter lists featuring 8K list members.
- **Year**: 2016.
- **Sensitive features**: race.
- **Link**: not available
- **Further info**: Ge et al. (2016)

A.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.
- **A liation of creators**: Technische Universität Berlin; Harvard University.
- **Domain**: information systems.
- **Tasks in fairness literature**: fair ranking evaluation (Sühr et al., 2021), multi-stage fairness evaluation (Sühr et al., 2021).
- **Data spec**: query-worker pairs.
- **Sample size**: 3 rankings (one per query) of 10 workers.
- **Year**: 2021.
- **Sensitive features**: gender.
- **Link**: not available
- **Further info**: Sühr et al. (2021)

A.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.
- **A liation 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 (Segal et al., 2021).

- **Data spec**: time series.
- **Sample size**: 600 speakers, each uttering 10 sentences.
- **Year**: 1993.
- **Sensitive features**: dialect, gender.
- **Link**: `https://catalog.ldc.upenn.edu/LDC93S1`
- **Further info**: `https://en.wikipedia.org/wiki/TIMIT`

A.186 Toy Dataset 1

- **Description**: this dataset consists of 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 $p(x|y=1) = N([2;2],[5,1;1,5])$; $p(x|y= \ 1) = N([\ 2;\ 2],[10,1;1,3])$. 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: $x' = [\cos(\ ),\ \sin(\ );\sin(\ ),\cos(\ )]x$. Parameter 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 (Zafar et al., 2017c; Roh et al., 2020), fair preference-based classification (Zafar et al., 2017b; Ali et al., 2019b), fair few-shot learning (Slack et al., 2019a, 2020), fair classification under unawareness (Kilbertus et al., 2018).
- **Data spec**: tabular data.
- **Sample size**: 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. (2017c)

A.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 (Singh and Joachims, 2019; Bower et al., 2021).
- **Data spec**: query-document pairs.
- **Sample size**: 1K relevance judgements overs 100 queries with 10 candidate documents.
- **Year**: 2019.
- **Sensitive features**: N/A.
- **Link**: `https://github.com/ashudeep/Fair-PGRank`
- **Further info**: Singh and Joachims (2019)

A.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:

$$
\begin{aligned}
z_i & \quad \text{Bernoulli}(0.5) \\
\text{hair\_length}_i | z_i = 1 & \quad 35 \quad \text{Beta}(2, 2) \\
\text{hair\_length}_i | z_i = 0 & \quad 35 \quad \text{Beta}(2, 7) \\
\text{work\_exp}_i | z_i & \quad \text{Poisson}(25 + 6 z_i) \quad \text{Normal}(20, 0.2) \\
y_i | \text{work\_exp}_i & \quad 2 \quad \text{Bernoulli}(p_i) \quad 1, \\
\text{where} p_i = & \; 1/(1 + \exp[\;(\;25.5 + 2.5 \text{work\_exp})])
\end{aligned}
$$

.

- **Affiliation of creators**: Carnegie Mellon University; University of California, San Diego.
- **Domain**: N/A.
- **Tasks in fairness literature**: fairness evaluation (Lipton et al., 2018; Black et al., 2020).
- **Data spec**: tabular data.
- **Sample size**: 2K points.
- **Year**: 2018.
- **Sensitive features**: N/A.
- **Link**: not available
- **Further info**: Lipton et al. (2018)

A.189 Toy Dataset 4

- **Description**: in this toy example, features are generated according to four 2-dimensional isotropic Gaussian distributions with different mean and variance ^{2}. Each of the four distributions corresponds to a different combination of binary label y and protected attribute s as follows: (1) $s = a, y = +1$: $= (\;1, \;1), \;^2 = 0.8$; (2) $s = a, y = \;1$: $= (1, 1), \;^2 = 0.8$; (3) $s = b, y = +1$: $= (0.5, \;0.5), \;^2 = 0.5$; (4) $s = b, y = \;1$: $= (0.5, 0.5), \;^2 = 0.5$.
- **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 (Donini et al., 2018), fairness evaluation (Williamson and Menon, 2019).
- **Data spec**: tabular data.
- **Sample size**: 6K points.
- **Year**: 2018.
- **Sensitive features**: N/A.
- **Link**: `https://github.com/jmikko/fair_ERM`
- **Further info**: Donini et al. (2018)

A.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.
{ **A liation of creators**: National Institute of Standards and Technology.
{ **Domain**: news, information systems.
{ **Tasks in fairness literature**: fair ranking evaluation (Gerritse and de Vries, 2020).
{ **Data spec**: query-document pairs.
{ **Sample size**: 300K relevance judgements over 200 queries and 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 (2005)

A.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.
{ **A liation of creators**: University of Edinburgh.
{ **Domain**: social networks.
{ **Tasks in fairness literature**: fair graph mining (Kang et al., 2020).
{ **Data spec**: user-user pairs.
{ **Sample size**: 30K nodes (users) connected by 400K edges (mutual friendship).
{ **Year**: 2019.
{ **Sensitive features**: none.
{ **Link**: `http://snap.stanford.edu/data/twitch-social-networks.html`
{ **Further info**: Rozemberczki et al. (2021)

A.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.
{ **A liation 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 (Ball-Burack et al., 2021).
- **Data spec**: text.
- **Sample size**: 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. (2018)

A.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 (Ball-Burack et al., 2021).
- **Data spec**: text.
- **Sample size**: 20K tweets.
- **Year**: 2016.
- **Sensitive features**: textual references to people and their demographics.
- **Link**:
- **Further info**: Waseem and Hovy (2016)

A.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 (Ball-Burack et al., 2021), fair multi-stage classification (Keswani et al., 2021).
- **Data spec**: text.
- **Sample size**: 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. (2017)

A.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.
- **A liation of creators**: University of Maryland.
- **Domain**: social media.
- **Tasks in fairness literature**: fairness evaluation (Ball-Burack et al., 2021).
- **Data spec**: text.
- **Sample size**: 40K tweets.
- **Year**: 2017.
- **Sensitive features**: textual references to people and their demographics.
- **Link**: not available
- **Further info**: Golbeck et al. (2017)

A.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.
- **A liation 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**: 30K search results containing 30K distinct tweets from 20K users.
- **Year**: 2016.
- **Sensitive features**: political leaning.
- **Link**: not available
- **Further info**: Kulshrestha et al. (2017)

A.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.
- **A liation of creators**: Clarkson University.
- **Domain**: social media.
- **Tasks in fairness literature**: bias audit (Gordon et al., 2020).

- **Data spec**: text.
- **Sample size**: 1M tweets from 500 accounts.
- **Year**: 2020.
- **Sensitive features**: political leaning.
- **Link**: not available
- **Further info**: Gordon et al. (2020)

A.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++.
- **A liation of creators**: Indian Institute of Technology Kharagpur; Max Planck Institute for Software Systems; Grenoble INP.
- **Domain**: social media.
- **Tasks in fairness literature**: fair ranking (Chakraborty et al., 2019).
- **Data spec**: text.
- **Sample size**: 200M tweets by 10M users and 10K trending topics.
- **Year**: 2018.
- **Sensitive features**: race, age, and gender.
- **Link**: not available
- **Further info**: Chakraborty et al. (2019)

A.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.
- **A liation of creators**: University of Massachusetts Amherst.
- **Domain**: social media, linguistics.
- **Tasks in fairness literature**: fairness evaluation of sentiment analysis (Shen et al., 2018), fairness evaluation of private classification (Bagdasaryan et al., 2019), fairness evaluation (Ball-Burack et al., 2021), robust fair language model (Hashimoto et al., 2018), fairness evaluation of language identification (Blodget and O'Connor, 2017).
- **Data spec**: text.
- **Sample size**: 8M tweets.
- **Year**: 2016.
- **Sensitive features**: dialect (related to race).
- **Link**: `http://slanglab.cs.umass.edu/TwitterAAE/`
- **Further info**: Blodgett et al. (2016)

A.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.
- **A liation of creators**: US International Trade Commission.
- **Domain**: economics.
- **Tasks in fairness literature**: fairness evaluation (Luong et al., 2016).
- **Data spec**: tabular data.
- **Sample size**: unknown.
- **Year**: present.
- **Sensitive features**: gender.
- **Link**: `https://hts.usitc.gov/current`
- **Further info**: Barbaro (2007)

A.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.
- **A liation of creators**: University of Genoa.
- **Domain**: education.
- **Tasks in fairness literature**: fair regression (Chzhen et al., 2020b,a), fair representation learning (Oneto et al., 2019b, 2020).
- **Data spec**: tabular data.
- **Sample size**: 5K students.
- **Year**: unknown.
- **Sensitive features**: ethnicity, gender, financial status.
- **Link**: not availalbe
- **Further info**: Oneto et al. (2017)

A.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.
- **A liation of creators**: Facebook; University of North Carolina; Harvard University; University of Oxford.
- **Domain**: social networks.
- **Tasks in fairness literature**: fair graph mining (Li et al., 2021), fair graph augmentation (Ramachandran et al., 2021).
- **Data spec**: user-user pairs.
- **Sample size**: 20K people connected by 1M friend relations (Oklahoma97); 20K people connected by 1M friend relations (UNC28); 30K people connected by 1M friend relations (Michigan23); 1K people connected by 20K friend relations (Reed98); 1K people connected by 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. (2011)

A.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.
{ **A liation of creators**: Microsoft.
{ **Domain**: demography.
{ **Tasks in fairness literature**: fair clustering (Backurs et al., 2019; Huang et al., 2019; Bera et al., 2019), fair clustering under unawareness (Esmaeili et al., 2020), limited-label fairness evaluation (Sabato and Yom-Tov, 2020).
{ **Data spec**: tabular data.
{ **Sample size**: 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. (2002)

A.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).
{ **A liation of creators**: US Bureau of Labor Statistics; US Census Bureau.
{ **Domain**: economics.
{ **Tasks in fairness literature**: fair subset selection under unawareness (Mehrotra and Celis, 2021).
{ **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.205 US Federal Judges

{ **Description**: this dataset was extracted from Epstein et al. (2013) 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. (2013), 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.

{ **A liation of creators**: Yale University.
{ **Domain**: law.
{ **Tasks in fairness literature**: fair subset selection (Huang et al., 2020).
{ **Data spec**: judge-case pairs.
{ **Sample size**: 300 judges selected for 2K cases.
{ **Year**: 2020.
{ **Sensitive features**: gender.
{ **Link**: not available
{ **Further info**: Huang et al. (2020)

A.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.
{ **A liation of creators**: Cornell University.
{ **Domain**: education.
{ **Tasks in fairness literature**: fairness evaluation (Lee and Kizilcec, 2020).
{ **Data spec**: tabular data.
{ **Sample size**: unknown.
{ **Year**: 2020.
{ **Sensitive features**: gender, racial-ethnic group.
{ **Link**: not available
{ **Further info**: Lee and Kizilcec (2020)

A.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.
{ **A liation of creators**: University of Tennessee.
{ **Domain**: computer vision.
{ **Tasks in fairness literature**: robust fairness evaluation (Nanda et al., 2021), fairness evaluation of private classification (Bagdasaryan et al., 2019), fairness evaluation (Segal et al., 2021), fair classification (Jung et al., 2021).
{ **Data spec**: image.
{ **Sample size**: 20K face images.
{ **Year**: 2017.
{ **Sensitive features**: age, gender, race (inferred).
{ **Link**: `https://susanqq.github.io/UTKFace/`
{ **Further info**: Zhang et al. (2017b)

A.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.
- **A liation of creators**: University of Wisconsin-Madison.
- **Domain**: signal processing.
- **Tasks in fairness literature**: fair federated learning (Li et al., 2020b).
- **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 (2004)

A.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.
- **A liation of creators**: Purdue University.
- **Domain**: literature.
- **Tasks in fairness literature**: fair clustering (Harb and Lam, 2020).
- **Data spec**: text.
- **Sample size**: 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 (2018)

A.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 (Lin et al., 2014) 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.
- **A liation of creators**: Georgia Institute of Technology; Carnegie Mellon University; Army Research Lab; Facebook AI Research.
- **Domain**: computer vision.
- **Tasks in fairness literature**: bias discovery (Manjunatha et al., 2019).
- **Data spec**: mixture (image, text).
- **Sample size**: 1M questions over 300K images.
- **Year**: 2017.
- **Sensitive features**: visual and textual references to gender.
- **Link**: `https://visualqa.org/`
- **Further info**: Goyal et al. (2017)

A.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.
{ **A liation of creators**: PharmGKB; International Warfarin Pharmacogenetics Consortium.
{ **Domain**: pharmacology.
{ **Tasks in fairness literature**: fairness evaluation under unawareness (Kallus et al., 2020).
{ **Data spec**: tabular data.
{ **Sample size**: 6*K* patients.
{ **Year**: 2009.
{ **Sensitive features**: sex, ethnicity, age.
{ **Link**: `https://www.pharmgkb.org/downloads`
{ **Further info**: International Warfarin Pharmacogenetics Consortium (2009)

A.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 (Wah et al., 2011), 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 (Zhou et al., 2018). 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.
{ **A liation of creators**: Stanford University; Microsoft.
{ **Domain**: computer vision.
{ **Tasks in fairness literature**: fairness evaluation of selective classification (Jones et al., 2021).
{ **Data spec**: image.
{ **Sample size**: 10K images.
{ **Year**: 2021.
{ **Sensitive features**: none.
{ **Link**: `https://github.com/ejones313/worst-group-sc/tree/main/src/data`
{ **Further info**: Sagawa et al. (2020)

A.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 (Peters and Lecocq, 2013) and Newspaper[27] extractors was exploited. The curators performed deduplication and removed all Wikipedia pages to reduce text overlap with Wikipedia-based datasets.

[27] `https://github.com/codelucas/newspaper`

- **Affiliation of creators**: OpenAI.
- **Domain**: linguistics.
- **Tasks in fairness literature**: data bias evaluation (Tan and Celis, 2019).
- **Data spec**: text.
- **Sample size**: 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. (2019)

A.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 (Jones et al., 2020).
- **Data spec**: tabular data.
- **Sample size**: 400 businesses.
- **Year**: 2014.
- **Sensitive features**: geography.
- **Link**: `http://archive.ics.uci.edu/ml/datasets/wholesale+customers`
- **Further info**: Baudry et al. (2015)

A.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 (Fisher et al., 2020).
- **Data spec**: item-property-value triples.
- **Sample size**: 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.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 (Liang and Acuna, 2020; Papakyriakopoulos et al., 2020; Brunet et al., 2019; Chen et al., 2021).
- **Data spec**: text.
- **Sample size**: 6M articles (EN), 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.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, (Garg et al., 2019; Shah et al., 2021), fairness evaluation (Dixon et al., 2018).
- **Data spec**: text.
- **Sample size**: 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.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 (Kallus and Zhou, 2021).
- **Data spec**: tabular data.
- **Sample size**: 2K respondents.
- **Year**: 2015.
- **Sensitive features**: age, gender, geography.
- **Link**: `https://snd.gu.se/sv/catalogue/study/snd0987/1#dataset`
- **Further info**: Slunge (2015)

A.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. (Vig et al., 2020).
- **Data spec**: text.
- **Sample size**: 3K sentences.
- **Year**: 2020.
- **Sensitive features**: gender.
- **Link**: `https://github.com/uclanlp/corefBias/tree/master/WinoBias/wino`
- **Further info**: Zhao et al. (2018)

A.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 (Vig et al., 2020), fairness evaluation in entity recognition (Mishra et al., 2020).
- **Data spec**: text.
- **Sample size**: 700 sentences.
- **Year**: 2018.
- **Sensitive features**: gender.
- **Link**: `https://github.com/rudinger/winogender-schemas`
- **Further info**: Rudinger et al. (2018)
- **Variants**: Winogender-NER (Mishra et al., 2020) is a modified version of the template appropriate for named entity recognition.

A.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 (Greenwald et al., 1998), 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 (Brunet et al., 2019; Guo and Caliskan, 2021).
- **Data spec**: text.
- **Sample size**: 10 groups of words, with 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. (2017)

A.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.
{ **A liation of creators**: Yahoo! Labs.
{ **Domain**: marketing.
{ **Tasks in fairness literature**: fair advertising (Celis et al., 2019a; Nasr and Tschantz, 2020).
{ **Data spec**: advertiser-keyword pairs.
{ **Sample size**: 20M bids by 10K advertisers over 1K search queries.
{ **Year**: after 2003.
{ **Sensitive features**: none.
{ **Link**: `https://webscope.sandbox.yahoo.com/catalog.php?datatype=a`
{ **Further info**:

A.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.
{ **A liation of creators**: Yahoo! Labs.
{ **Domain**: information systems.
{ **Tasks in fairness literature**: fair ranking (Singh and Joachims, 2019).
{ **Data spec**: query-document pairs.
{ **Sample size**: 40K queries, 900K documents.
{ **Year**: 2011.
{ **Sensitive features**: none.
{ **Link**: `https://webscope.sandbox.yahoo.com/catalog.php?datatype=c`
{ **Further info**: Chapelle and Chang (2010)

A.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.

- **A liation of creators**: University of Washington.
- **Domain**: social media.
- **Tasks in fairness literature**: fairness evaluation of speech recognition (Tatman, 2017).
- **Data spec**: tabular data.
- **Sample size**: 100 speakers.
- **Year**: 2016.
- **Sensitive features**: gender, geography.
- **Link**: `https://github.com/rctatman/youtubeDialectAccuracy`
- **Further info**: Tatman (2017)

A.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.
- **A liation of creators**: Carnegie Mellon University.
- **Domain**: news, information systems.
- **Tasks in fairness literature**: fair ranking (Singh and Joachims, 2018).
- **Data spec**: user-story pairs.
- **Sample size**: 10K interaction logs.
- **Year**: 2009.
- **Sensitive features**: news provider.
- **Link**: `https://users.soe.ucsc.edu/~yiz/papers/data/YOWStudy/`
- **Further info**: Zhang (2005); `https://users.soe.ucsc.edu/~yiz/piir/`

A.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.
- **A liation of creators**: Boston University; Zillow Group.
- **Domain**: information systems.
- **Tasks in fairness literature**: fair ranking (Chaudhari et al., 2020).
- **Data spec**: unknown.
- **Sample size**: 10K search sessions featuring 40K property listings.
- **Year**: 2020.
- **Sensitive features**: ownership, tour availability.
- **Link**: not available
- **Further info**: Chaudhari et al. (2020)

Appendix B Adult

Key references include Cohany et al. (1994); Kohavi (1996); UCI Machine Learning Repository (1996); US Dept. of Commerce Bureau of the Census (1995); Ding et al. (2021); McKenna (2019a,b).

B.1 Datasheet

B.1.1 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 (Kohavi, 1996).
- **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 (UCI Machine Learning Repository, 1996)
- **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.

B.1.2 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 (UCI Machine Learning Repository, 1996), 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 9 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 (US Dept. of Commerce Bureau of the Census, 1995) but this information is not reported in the dataset.

- Are there recommended data splits?
  Yes . The dataset comes with a speci ed train/test split made using MLC++ GenCV-Files, resulting in a 2/3{1/3 random split (UCI Machine Learning Repository, 1996). 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 de nitional di culties, di erences 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-identi cation of certain individuals (US Dept. of Commerce Bureau of the Census, 1995).
- 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 con dential?
  Yes . The data is protected by Title 13 of the United States Code, protecting individuals against identi cation from Census data. [28]
- Does the dataset contain data that, if viewed directly, might be o ensive, 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-identi cation was bounded by envisioned categories. These are (female, male) for sex and (White, Black, American Indian/Aleut Eskimo, Asian or Paci c Islander, Other) for race. Table 6 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-identi cation, 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 (Ding et al., 2021). Moreover, re-identi cation studies internal to the Census Bureau pointed to combinations of variables that could potentially be used to re-identify respondents from Census microdata (McKenna, 2019b).
- 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 nancial situation of respondents.
- Any other comments?
  A precise de nition 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 classi cation tasks would be unusual.

B.1.3 Collection process

- How was the data associated with each instance acquired?
  Trained interviewers asked questions directly to respondents (US Dept. of Commerce Bureau of the Census, 1995). 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 (US Dept. of Commerce Bureau of the Census, 1995).

[28] 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 6: Demographic Characteristics of the Adult dataset.

{ 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 strati ed 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 (US Dept. of Commerce Bureau of the Census, 1978). 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 a liated 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 (McKenna, 2019b) 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 noti ed 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 noti ed 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. [29]
- 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-identi cation studies have been conducted both internally (McKenna, 2019b) and externally (Rocher et al., 2019) on Census Bureau data. McKenna (2019b) mention nding combinations of variables on Census les that can lead to successful reidenti cation, which were subsequently removed or protected with noise injection. Rocher et al. (2019) demonstrate on the Adult dataset that the likelihood of a speci c individual to have been correctly re-identi ed can be estimated with high accuracy. We are unaware of studies about the potential impact of successful re-identi cation on respondents.

B.1.4 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.

B.1.5 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 speci c applications, Adult is used as a benchmark for classi ers in many elds 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 (1996). A more recent list is available under the beta version of the UCI ML Repository. [30] See Appendix A.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 signi cance 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 classi cation 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 (Ding et al., 2021). Some numerical results on Adult may be an artifact of this threshold choice.

[29] https://www2.census.gov/programs-surveys/cps/advance_letter.pdf

[30] https://archive-beta.ics.uci.edu/ml/datasets/2

– 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.

B.1.6 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 (UCI Machine Learning Repository, 1996).
– How is the dataset distributed?
  The dataset is available as a csv le .
– 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 .

B.1.7 Maintenance

– Who is supporting/hosting/maintaining the dataset?
  The dataset is hosted and maintained by the UCI Machine Learning Repository (UCI Machine Learning Repository, 1996).
– 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 speci c 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 (Ding et al., 2021).
– 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 .

B.2 Data Nutrition Label

B.2.1 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 classi ers tasked with predicting whether individual income exceeds $50K/yr based on demographic and socio-economic information. Also known as \Census Income" dataset.

Table 7: Metadata of the Adult dataset

B.2.2 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 8: Provenance of the Adult dataset

B.2.3 Variables

VARIABLES	
age	Respondent's age.
workclass	Broad classi cation 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 9: 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- shing (Farming, forestry, and shing) Transport-moving (Transportation and material moving) Priv-house-serv (Private household service, e.g. cooks, cleaners) Protective-serv (Protective service, e.g. re ghters, police) Armed-Forces
relationship	Familial role wihtin household. Wife Own-child Husband Not-in-family Other-relative Unmarried

Table 10: 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	Pro ts 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 11: Variables of the Adult dataset (3/3).

B.2.4 Statistics

STATISTICS						
Ordinal						
name	type	count	uniqueEntries	mostFrequent	leastFrequent	missing
education-num	int	48842	16	9	1	0

Table 12: Ordinal variables statistics of the Adult dataset

Categorical						
name	type	count	uniqueEntries	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 13: 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	178144.5	1490400	189664.13	105604.03	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 14: Quantitative variables statistics of the Adult dataset.

Appendix C COMPAS

Key references include Angwin et al. (2016); Larson et al. (2016); Dieterich et al. (2016); ProPublica (2016); Equivant (2019); Brennan et al. (2009); Bao et al. (2021); Barenstein (2019).

C.1 Datasheet

C.1.1 Motivation

{ For what purpose was the dataset created?
This dataset was created for an external audit of racial biases in the Correctional Offender Management Pro ling 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), Je Larson (data editor), Surya Mattu (contributing researcher), Lauren Kirchner (senior reporting fellow). All four contributors were a liated with ProPublica at the time.

{ Who funded the creation of the dataset?
The dataset curation work was likely remunerated by ProPublica.

C.1.2 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 (ProPublica, 2016) 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-years-violent.csv).

{ 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 (Larson et al., 2016). 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 (Barenstein, 2019).

{ 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 o ense (potentially violent) that took place after their COMPAS screening but within two years. The definition of recidivism and the two-year cuto were selected by ProPublica sta to align their audit with de nitions by Northpointe (Brennan et al., 2009; Angwin et al., 2016).

- 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 o ense (c _o ense_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 Sheri 's O ce, Broward County Clerk's Of- ce 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 rst adjudication, arrest rate and vocational educational scale (Dieterich et al., 2016). 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% (Larson et al., 2016). 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 (Barenstein, 2019). 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 con dential?
  No . However it does contains rst names and last names of defendants, connecting them to their criminal history.
- Does the dataset contain data that, if viewed directly, might be o ensive, 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 identi es population by age, sex and race. The curators of the COMPAS dataset maintained the race classi cations used by the Broward County Sheri 's O ce, identifying individuals as Asian, Black, Hispanic, Native American and White (Larson et al., 2016). Age is reported as an integer, sex as either Male or Female. A distribution along these dimensions is reported in Table 15 which summarizes data in compas-scores-two-years.csv . Distributions in remaining les are similar.
- Is it possible to identify individuals , either directly or indirectly from the dataset?
  Yes . The dataset reports defendants' rst 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.

C.1.3 Collection process

- How was the data associated with each instance acquired?
  The data was obtained cross-referencing three sources. From the Broward County Sheri 's O ce 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 O ce website matching them based on date of birth, rst

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 15: Demographic Characteristics of compas-scores-two-years.

and last names. The dataset was augmented with jail records provided by the Broward County Sheri 's O ce. 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 Sheri 's O ce, Broward County Clerk's O ce, and Florida Department of Corrections. The curators of the COMPAS dataset obtained records from the County Sheri 's O ce through a public records request, while data from the County Clerk's O ce and the Florida Department of Correction was downloaded from their o cial website, matching the methodology of a COMPAS validation study (Larson et al., 2016).

{ 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 conve-nience 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 de nition of pretrial defendants (Dieterich et al., 2016)) 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 speci c sampling e ects, inevitably mediated by law enforcement practices Xie and Lauritsen (2012); Holmes et al. (2008).

{ 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 (Angwin et al., 2016) and accompanying technical report (Larson et al., 2016), namely Julia Angwin, Je 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 rst released by ProPublica in May 2016 (ProPublica, 2016).

{ 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 Sheri 's O ce in Florida through a public records request, from the Broward County Clerk's O ce through the o cial website and through the Florida Department of Corrections through the o cial website. Collection from interested individuals would not have been viable.

{ Were the individuals in question noti ed about the data collection?
Likely no . Most of the COMPAS data was publicly available and downloaded from the o cial websites of Broward County Clerk's O ce 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 nancially (Pinard, 2010; Angwin et al., 2016).

{ 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 speci cally focused on the COMPAS dataset. More broadly, public availability of criminal records is related to studies on the employability of o enders (Gra am et al., 2008).

C.1.4 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 (ProPublica, 2016). 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 de ned as misdemeanor or felony, (4) the COMPAS score is not clearly de ned. 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 o cial ProPublica github repository (ProPublica, 2016). 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 o cial ProPublica github repository (ProPublica, 2016). The software used to cross-reference data from separate sources is not publicly available.

C.1.5 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 di erent algorithms and, more broadly, to study de nitions of algorithmic fairness.

– Is there a repository that links to any or all papers or systems that use the dataset?
  See Appendix A.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 arti cially in ated recidivism base rates (Barenstein, 2019), 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 (Mishler et al., 2021). From a broader perspective, the data is likely in uenced by historical biases in criminal justice, with di erential impact on di erent communities (Holmes et al., 2008; Xie and Lauritsen, 2012; Angwin et al., 2016). Zooming out further, the use of automated risk assessment tools in pretrial decisions is the subject of controversial debate (Barabas et al., 2019) 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. (2021) 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.

C.1.6 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 o cial github repository (ProPublica, 2016).
– 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 (ProPublica, 2021). 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 .

C.1.7 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 o cial erratum. An external report highlighting anomalies in the data is available (Barenstein, 2019).

{ Will the dataset be updated?
Likely no . In the event of an update, ProPublica's data store ToU speci es users are solely responsible for checking their sites for updates (ProPublica, 2021)

{ 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 .

C.2 Data Nutrition Label

The following analysis refers to compas-scores-two-years.csv after applying the standard COMPAS preprocessing (ProPublica, 2016).

C.2.1 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 (ProPublica, 2021)
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 16: Metadata of COMPAS dataset.

C.2.2 Provenance

PROVENANCE	
Source	
Name	Broward County Sheri 's O ce
Url	http://www.sheriff.org/
email	//
Name	Broward County Clerk's O ce
Url	https://www.browardclerk.org
email	Eclerk@browardclerk.org
Name	Florida Department of Corrections
Url	http://www.dc.state.fl.us/
email	FDCCitizenServices@fdc.my orida.com
Authors	
Names	Julia Angwin, Je 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 17: Provenance of COMPAS dataset.

C.2.3 Variables

VARIABLES	
id	Unique identi er assigned by the authors
name	Defendant's rst and last name
rst	Defendant's rst name
last	Defendant's last name
compas_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 greater than 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 18: 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 identi er
c_o ense_date	Date on which the o ense was committed
c_arrest_date	Date on which defendant was arrested
c_days_from_compas	Days elapsed between o ense/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 identi er for recidivist o ense
r_charge_degree	Degree of recidivist charge
r_days_from_arrest	Days elapsed between date of recidivist o ense (r_o ense_date) and date of recidivist incarceration (r_jail_in)
r_o ense_date	Date of recidivist o ense
r_charge_desc	Textual description of recidivist charge
r_jail_in	Date of incarceration for recidivist o ense
r_jail_out	Date of release for recidivist o ense

Table 19: 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 identi er for violent recidivist o ense
vr_charge_degree	Degree of violent recidivist o ense
vr_o ense_date	Date of violent recidivist o ense
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 20: Variables of COMPAS dataset (3/3).

C.2.4 Statistics

STATISTICS						
Ordinal						
name	type	count	uniqueEntries	mostFrequent	leastFrequent	missing
id	int	6,172	6,172	multiple	multiple	0
compas_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	Greater than 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_o ense_date	date	6,172	737	multiple	multiple	1388
c_arrest_date	date	6,172	417	2013-02-06	multiple	8425
r_o ense_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_o ense_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 21: Ordinal variables statistics of COMPAS dataset

Categorical						
name	type	count	uniqueEntries	mostFrequent	leastFrequent	missing
name	string	6,172	9,128	mutiple	multiple	0
rst	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 22: 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 23: Quantitative variables statistics of COMPAS dataset.

Appendix D German Credit

Key references include Häußler (1979); UCI Machine Learning Repository (1994); Grömping (2019); UCI Machine Learning Repository (2019).

D.1 Datasheet

D.1.1 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 (UCI Machine Learning Repository, 1994). Presumably, he donated the dataset to the European Statlog project and a representative of Strathclyde University donated it to UCI (Grömping, 2019).
- 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 (Häußler, 1979). 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.

D.1.2 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 (Häußler, 1979). 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 27-29.
- 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 (1994). 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 (Grömping, 2019). 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 (Grömping, 2019). 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.73). More mistakes in the documentation of eight variables and the relative errata are outlined in Grömping (2019). A clean version of the dataset is available at UCI Machine Learning Repository (2019).
- 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 (UCI Machine Learning Repository, 1994). A summary based on amended documentation (Grömping, 2019) is presented in Table 24.

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 24: Demographic characteristics of the German credit dataset.

{ **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.

D.1.3 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).

{ **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 (Häußler, 1979). It became publicly available in November 1994 (UCI Machine Learning Repository, 1994).

{ **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 noti ed 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.

D.1.4 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.

D.1.5 Uses

– **For what tasks has the dataset been used?**
  The dataset was originally used to study the problem of automated credit scoring (Häußler, 1979). 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 (1994). A more recent list is available under the beta version of the UCI ML Repository.[31] See Appendix A.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 (UCI Machine Learning Repository, 1994), 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 (UCI Machine Learning Repository, 1994) 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 (UCI Machine Learning Repository, 2019) associated with the erratum (Grömping, 2019) is recommended.

D.1.6 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 (UCI Machine Learning Repository, 1994)
– **How is the dataset distributed?**
  The dataset is available as a **csv le**.
– **When was the dataset distributed?**
  The dataset was released to the UCI ML Repository in **November 1994**.
– **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.

[31] `https://archive-beta.ics.uci.edu/ml/datasets/144`

D.1.7 Maintenance

- **Who is supporting/hosting/maintaining the dataset?**
  The dataset is hosted and maintained by the **UCI Machine Learning Repository** (UCI Machine Learning Repository, 1994). A clean and well-documented version of the same dataset donated by Ulrike Gromping (UCI Machine Learning Repository, 2019) 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 (UCI Machine Learning Repository, 2019) and accompanying report (Grömping, 2019) are available online.
- **Will the dataset be updated?**
  Likely no. The recently released South German Credit Data Set (UCI Machine Learning Repository, 2019) 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 (UCI Machine Learning Repository, 2019) and the old version (UCI Machine Learning Repository, 1994) 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.

D.2 Data Nutrition Label

For the sake of correctness, we report redacted information based on the new South German Credit Data Set (UCI Machine Learning Repository, 2019) and accompanying documentation (Grömping, 2019).

D.2.1 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 25: Metadata of South German Credit dataset.

D.2.2 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 26: Provenance of South German Credit dataset

D.2.3 Variables

VARIABLES	
status	Checking account balance (in Deutsche Mark) 1 (no checking account) 2 (< 0 DM) 3 (0 ... < 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 27: 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 ... < 500 DM) 4 (500 ... < 1000 DM) 5 (1000 DM)
employment_duration	Duration of applicant's current employment 1 (unemployed) 2 (< 1 year) 3 (1 ... < 4 years) 4 (4 ... < 7 years) 5 (7 years)
installment_rate	Installment amount to disposable income ratio [%] 1 (35) 2 (25 ... < 35) 3 (20 ... < 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 ... < 4 years) 3 (4 ... < 7 years) 4 (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 28: 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(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 29: Variables of South German Credit dataset (3/3).

D.2.4 Statistics

STATISTICS

Ordinal

name	type	count	unique	mostFrequent	leastFrequent	missing
status	string	1000	4	4 (200)	3 (0 ...<200)	0
savings	string	1000	5	1 (unknown/no savings)	4 (500 ...<1000)	0
employment_duration	string	1000	5	3 (1 ...<4)	1 (unemployed)	0
installment_rate	string	1000	4	4 (<20)	1 35	0
present_residence	string	1000	4	4 (7 yrs)	1 (< 1 yr)	0
number_credits	string	1000	4	1 (1)	4 (6)	0
people liable	string	1000	2	2 (0 to 2)	1 (3 or more)	0

Table 30: Ordinal variables statistics of South German Credit dataset

Categorical

name	type	count	uniqueEntries	mostFrequent	leastFrequent	missing
credit_history	string	1000	5	2 (no credits taken)	0 (delay in paying off)	0
purpose	string	1000	11	3 (furniture/equipment)	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 (building 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/unsk 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 31: 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 32: Quantitative variables statistics of South German Credit dataset.