Data Management for Causal Algorithmic Fairness*

Babak Salimi*, Bill Howe†, Dan Suciu*
University of Washington
*{bsalimi,suciu}@cs.washington.edu, †billhowe@uw.edu

Abstract

Fairness is increasingly recognized as a critical component of machine learning systems. However, it is the underlying data on which these systems are trained that often reflects discrimination, suggesting a data management problem. In this paper, we first make a distinction between associational and causal definitions of fairness in the literature and argue that the concept of fairness requires causal reasoning. We then review existing works and identify future opportunities for applying data management techniques to causal algorithmic fairness.

1 Introduction

Fairness is increasingly recognized as a critical component of machine learning (ML) systems. These systems are now routinely used to make decisions that affect people's lives [11], with the aim of reducing costs, reducing errors, and improving objectivity. However, there is enormous potential for harm: The data on which we train algorithms reflects societal inequities and historical biases, and, as a consequence, the models trained on such data will therefore reinforce and legitimize discrimination and opacity. The goal of research on algorithmic fairness is to remove bias from machine learning algorithms.

We recently argued that the algorithmic fairness problem is fundamentally a data management problem [43]. The selection of sources, the transformations applied during pre-processing, and the assumptions made during training are all sensitive to bias that can exacerbate fairness effects. The goal of this paper is to discuss the application of data management techniques in algorithmic fairness. In Sec 2 we make a distinction between associational and causal definitions of fairness in the literature and argue that the concept of fairness requires causal reasoning to capture natural situations, and that the popular associational definitions in ML can produce misleading results. In Sec 3 we review existing work and identify future opportunities for applying data management techniques to ensure causally fair ML algorithms.

2 Fairness Definitions

Algorithmic fairness considers a set of variables $\mathbf{V}$ that include a set of *protected attributes* $\mathbf{S}$ and a *response variable* Y, and a classification algorithm $\mathcal{A} : Dom(\mathbf{X}) \rightarrow Dom(O)$, where $\mathbf{X} \subseteq \mathbf{V}$, and the result is denoted

Bulletin of the IEEE Computer Society Technical Committee on Data Engineering

*This work is supported by the National Science Foundation under grants NSF III-1703281, NSF III-1614738, NSF AITF 1535565 and NSF award #1740996.

Fairness Metric	Description
Demographic Parity (DP) [7] a.k.a. Statistical Parity [12] or Benchmarking [44]	$S \perp\!\!\!\perp O$
Conditional Statistical Parity [10]	$S \perp\!\!\!\perp O \mid \mathbf{A}$
Equalized Odds (EO) [15] [2] a.k.a. Disparate Mistreatment [47]	$S \perp\!\!\!\perp O \mid Y$
Predictive Parity (PP)[9] [3] a.k.a. Outcome Test [44] or Test-fairness [9] or Calibration [9], or Matching Conditional Frequencies [15]	$S \perp\!\!\!\perp Y \mid O$

Figure 1: Common associational definitions of fairness.

O and called *outcome*. To simplify the exposition, we assume a sensitive attribute $S \in \mathbf{S}$ that classifies the population into protected $S = 1$ and privileged $S = 0$, for example, female and male, or minority and non-minority (see [48] for a survey). The first task is to define formally when an algorithm $\mathcal{A}$ is fair w.r.t. the protected attribute S; such a definition is, as we shall see, not obvious. Fairness definitions can be classified as associational or causal, which we illustrate using the following running example (see [45] for a survey on fairness definitions).

Example 1: *In 1973, UC Berkeley was sued for discrimination against females in graduate school admissions. Admission figures for the fall of 1973 showed that men applying were more likely than women to be admitted, and the difference was so large that it was unlikely to be due to chance. However, it turned out that the observed correlation was due to the indirect effect of gender on admission results through applicant's choice of department. It was shown that females tended to apply to departments with lower overall acceptance rates [41]. When broken down by department, a slight bias toward female applicants was observed, a result that did not constitute evidence for gender-based discrimination. Extending this case, suppose college admissions decisions are made independently by each department and are based on a rich collection of information about the candidates, such as test scores, grades, resumes, statement of purpose, etc. These characteristics affect not only admission decisions, but also the department to which the candidate chooses to apply. The goal is to establish conditions that guarantee fairness of admission decisions.*

2.1 Associational Fairness

A simple and appealing approach to defining fairness is by correlating the sensitive attribute S and the outcome O. This leads to several possible definitions (Fig. 1). *Demographic Parity* (DP) [12] requires an algorithm to classify both protected and privileged groups with the same probability, i.e., $\Pr(O = 1|S = 1) = \Pr(O = 1|S = 0)$. However, doing so fails to correctly model our Example 1 since it requires equal probability for males and females to be admitted, and, as we saw, failure of DP cannot be considered evidence for gender-based discrimination. This motivates *Conditional Statistical Parity* (CSP) [10], which controls for a set of admissible factors $\mathbf{A}$, i.e., $\Pr(O = 1|S = 1, \mathbf{A} = \mathbf{a}) = \Pr(O = 1|S = 0, \mathbf{A} = \mathbf{a})$. The definition is satisfied if subjects in both protected and privileged groups have equal probability of being assigned to the positive class, controlling for a set of admissible variables. In the UC Berkeley case, CSP is approximately satisfied by assuming that department is an admissible variable.

Another popular measure used for predictive classification algorithms is *Equalized Odds* (EO), which requires both protected and privileged groups to have the same false positive (FP) rate, $\Pr(O = 1|S = 1, Y = 0) = \Pr(O = 1|S = 0, Y = 0)$, and the same false negative (FN) rate, $\Pr(O = 0|S = 1, Y = 1) = \Pr(O = 0|S = 0, Y = 1)$, or, equivalently, $(O \perp\!\!\!\perp S|Y)$. In our example, assuming a classifier is trained to predict if an applicant will be admitted, then the false positive rate is the fraction of rejected applicants for

which the classifier predicted that they should be admitted, and similarly for the false negative rate: EO requires that the rates of these false predictions be the same for male and female applicants. Finally, *Predictive Parity* (PP) requires that both protected and privileged groups have the same predicted positive value (PPV), $\Pr(Y = 1|O = i, S = 0) = \Pr(Y = 1|O = i, S = 1)$ for $i, = \{1, 0\}$ or, equivalently, $Y \perp\!\!\!\perp S|O$. In our example, this implies that the probability of an applicant that actually got admitted to be correctly classified as admitted and the probability of an applicant that actually got rejected to be incorrectly classified as accepted should both be the same for male and female applicants.

An Associational Debate. Much of the literature in algorithmic fairness is motivated by controversies over a widely used commercial risk assessment system for recidivism — COMPAS by Northpointe [18]. In 2016, a team of journalists from ProPublica constructed a dataset of more than 7000 individuals arrested in Broward County, Florida between 2013 and 2014 in order to analyze the efficacy of COMPAS. In addition, they collected data on arrests for these defendants through the end of March 2016. Their assessment suggested that COMPAS scores were biased against African-Americans based on the fact that the FP rate for African-Americans (44.9%) was twice that for Caucasians (23.5%). However, the FN rate for Caucasians (47.7%) was twice as large as for African-Americans (28.0%). In other words, COMPAS scores were shown to violate EO. In response to ProPublica, Northpointe showed COMPAS scores satisfy PP, i.e., the likelihood of recidivism among high-risk offenders is the same regardless of race.

This example illustrates that associational definitions are context-specific and can be mutually exclusive; they lack universality. Indeed, it has been shown that EO and PP are incompatible. In particular, Chouldechova [9] proves the following impossibility result. Suppose that prevalence of the two populations differs, $\Pr(Y = 1|S = 0) \neq \Pr(Y = 1|S = 1)$, for example, the true rate of recidivism differs for African-Americans and Caucasians; in this case, Equalized Odds and Predictive Parity cannot hold both simultaneously. Indeed, EO implies that $FP_i/(1 - FN_i)$ is the same for both populations $S = i$, $i = 0, 1$, while PP implies that $(1 - PPV_i)/PPV_i$ must be the same. Then, the identity

$$\frac{FP_i}{1 - FN_i} = \frac{\Pr(O = 1|S = i, Y = 0)}{\Pr(O = 1|S = i, Y = 1)} = \frac{\Pr(Y = 1|S = i)}{\Pr(Y = 0|S = i)} \frac{\Pr(Y = 0|O = 1, S = i)}{\Pr(Y = 1|O = 1, S = i)} = \frac{\Pr(Y = 1|S = i)}{\Pr(Y = 0|S = i)} \frac{1 - PPV_i}{PPV_i}$$

for $i = 0, 1$, implies $\Pr(Y = 1|S = 0) = \Pr(Y = 1|S = 1)$. We revisit the impossibility result in Sec 2.3.

2.2 Causal Fairness

The lack of universality and the impossibility result for fairness definitions based on associational definitions have motivated definitions based on causality [17, 16, 25, 37, 13]. The intuition is simple: fairness holds when there is no causal relationship from the protected attribute S to the outcome O. We start with a short background on causality.

Causal DAG. A *causal DAG* G over a set of variables $\mathbf{V}$ is a directed acyclic graph that models the functional interaction between variables in $\mathbf{V}$. Each node X represents a variable in $\mathbf{V}$ that is functionally determined by: (1) its parents $\mathbf{Pa}(X)$ in the DAG, and (2) some set of *exogenous* factors that need not appear in the DAG as long as they are mutually independent. This functional interpretation leads to the same decomposition of the joint probability distribution of $\mathbf{V}$ that characterizes Bayesian networks [27]:

$$\Pr(\mathbf{V}) = \prod_{X \in \mathbf{V}} \Pr(X|\mathbf{Pa}(X)) \tag{1}$$

d-Separation. A common inference question in a causal DAG is how to determine whether a CI ($\mathbf{X} \perp\!\!\!\perp \mathbf{Y}|\mathbf{Z}$) holds. A sufficient criterion is given by the notion of d-separation, a syntactic condition ($\mathbf{X} \perp\!\!\!\perp \mathbf{Y}|_d \mathbf{Z}$) that can be checked directly on the graph (we refer the reader to [26] for details).

Counterfactuals and do Operator. A *counterfactual* is an intervention where we actively modify the state of a set of variables $\mathbf{X}$ in the real world to some value $\mathbf{X} = \mathbf{x}$ and observe the effect on some output Y. Pearl [27] described the do operator, which allows this effect to be computed on a causal DAG, denoted $\Pr(Y|do(\mathbf{X} = \mathbf{x}))$. To compute this value, we assume that X is determined by a constant function $\mathbf{X} = \mathbf{x}$ instead of a function provided by the causal DAG. This assumption corresponds to a modified graph with all edges into $\mathbf{X}$ removed, and values of the incoming variables are set to $\mathbf{x}$. For a simple example, consider three random variables $X, Y, Z \in \{0, 1\}$. We randomly flip a coin and set $Z = 0$ or $Z = 1$ with probability $1/2$; next, we set $X = Z$, and finally we set $Y = X$. The resulting causal DAG is $Z \to X \to Y$, whose equation is $\Pr(X, Y, Z) = \Pr(Z)\Pr(X|Z)\Pr(Y|X)$. The do operator lets us observe what happens in the system when we intervene by setting $X = 0$. The result is defined by removing the edge $Z \to X$, whose equation is $\Pr(Y = y, Z = z|do(X) = 0) = \Pr(Z = z)\Pr(Y = y|X = 0)$ (notice that $\Pr(X|Z)$ is missing), leading to the marginals $\Pr(Y = 0|do(X) = 0) = 1, \Pr(Y = 1|do(X) = 0) = 0$. It is important to know the casual DAG since the probability distribution is insufficient to compute the do operator; for example, if we reverse the arrows to $Y \to X \to Z$ (flip Y first, then set $X = Y$, then set $Z = X$), then $\Pr(Y = 0|do(X) = 0) = \Pr(Y = 1|do(X) = 0) = 1/2$ in other words, intervening on X has no effect on Y.

Counterfactual Fairness. Given a set of features $\mathbf{X}$, a protected attribute S, an outcome variable Y, and a set of unobserved exogenous background variables $\mathbf{U}$, Kusner et al. [17] defined a predictor O to be *counterfactually fair* if for any $\mathbf{x} \in Dom(\mathbf{X})$:

$$P(O_{S\leftarrow 0}(\mathbf{U}) = 1|\mathbf{X} = \mathbf{x}, S = 1) = P(O_{S\leftarrow 1}(\mathbf{U}) = 1|\mathbf{X} = \mathbf{x}; S = 1) \tag{2}$$

where $O_{S\leftarrow s}(\mathbf{U})$ means intervening on the protected attribute in an unspecified configuration of the exogenous factors. The definition is meant to capture the requirement that the protected attribute S should not be a cause of O at the individual level. However, this definition captures individual-level fairness only under certain strong assumptions (see [43]). Indeed, it is known in statistics that individual-level counterfactuals cannot be estimated from data [34, 35, 36].

Proxy Fairness. To avoid individual-level counterfactuals, a common approach is to study population-level counterfactuals or interventional distributions that capture the effect of interventions at population rather than individual level [28, 34, 35]. Kilbertus et al. [16] defined proxy fairness as follows:

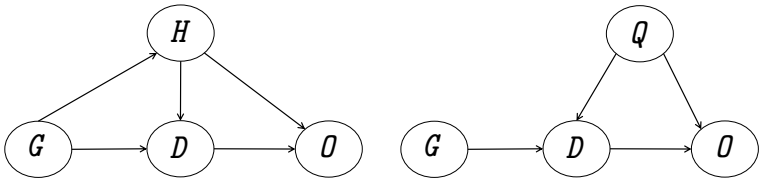
$$P(O = 1|do(\mathbf{P} = \mathbf{p})) = P(O = 1|do(\mathbf{P} = \mathbf{p}')) \tag{3}$$

for any $\mathbf{p}, \mathbf{p}' \in Dom(\mathbf{P})$, where $\mathbf{P}$ consists of proxies to a sensitive variable S (and might include S). Intuitively, a classifier satisfies proxy fairness in Eq 3 if the distribution of O under two interventional regimes in which $\mathbf{P}$ set to $\mathbf{p}$ and $\mathbf{p}'$ is the same. Thus, proxy fairness is not an individual-level notion. It has been shown that proxy fairness fails to capture group-level discrimination in general [43].

Path-Specific Fairness. These definitions are based on graph properties of the causal graph, *e.g.*, prohibiting specific paths from the sensitive attribute to the outcome [25, 22]; however, identifying path-specific causality from data requires very strong assumptions and is often impractical [4].

Interventional Fairness. To avoid issues with the aforementioned causal definitions, Salimi et al. [43] defined interventional fairness as follows: an algorithm $\mathcal{A} : Dom(\mathbf{X}) \to Dom(O)$ is $\mathbf{K}$-fair for a set of attributes $\mathbf{K} \subseteq \mathbf{V} - \{S, O\}$ w.r.t. a protected attribute S if, for any context $\mathbf{K} = \mathbf{k}$ and every outcome $O = o$, the following holds:

$$\Pr(O = o|do(S = 0), do(\mathbf{K} = \mathbf{k})) = \Pr(O = o|do(S = 1), do(\mathbf{K} = \mathbf{k})) \tag{4}$$



(a) College I (b) College II

College I	Dept. A		Dept. B		Total	
	Admitted	Applied	Admitted	Applied	Admitted	Applied
Male	16	20	16	80	32	100
Female	16	80	16	20	32	100

College II	Dept. A		Dept. B		Total	
	Admitted	Applied	Admitted	Applied	Admitted	Applied
Male	10	10	40	90	50	100
Female	40	50	10	50	50	100

Figure 2: Admission process representation in two colleges where associational notions of fairness fail (see Ex.2).

An algorithm is called *interventionally fair* if it is $\mathbf{K}$-fair for every set $\mathbf{K}$. Unlike proxy fairness, this notion correctly captures group-level fairness because it ensures that S does not affect O in *any configuration* of the system obtained by fixing other variables at some arbitrary values. Unlike counterfactual fairness, it does not attempt to capture fairness at the individual level, and therefore it uses the standard definition of intervention (the do-operator). In practice, interventional fairness is too restrictive. For example, in the UC Berkeley case, admission decisions were not interventionally fair since gender affected the admission result via applicant's choice of department. To make it practical, Salimi et al. [43] defined a notion of fairness that relies on partitioning variables into *admissible* and *inadmissible*. The former are variables through which it is permissible for the protected attribute to influence the outcome. This partitioning expresses fairness social norms and values and comes from the users. In Example 1, the user would label department as admissible since it is considered a fair use in admissions decisions and would (implicitly) label all other variables as inadmissible, for example, hobby. Then, an algorithm is called *justifiably fair* if it is $\mathbf{K}$-fair w.r.t. all supersets $\mathbf{K} \supseteq \mathbf{A}$. We illustrate with an example.

Example 2: Fig 2 shows how fair or unfair situations may be hidden by coincidences but exposed through causal analysis. In both examples, the protected attribute is gender G, and the admissible attribute is department D. Suppose both departments in College I are admitting only on the basis of their applicants' hobbies. Clearly, the admission process is discriminatory in this college because department A admits 80% of its male applicants and 20% of the female applicants, while department B admits 20% of male and 80% of female applicants. On the other hand, the admission rate for the entire college is the same 32% for both male and female applicants, falsely suggesting that the college is fair. Suppose H is a proxy to G such that $H = G$ (G and H are the same); proxy fairness then classifies this example as fair: indeed, since Gender has no parents in the causal graph, intervention is the same as conditioning; hence, $\Pr(O = 1|do(G = \text{i})) = \Pr(O = 1|G = \text{i})$ for $i = 0, 1$. Of the previous methods, only conditional statistical parity correctly indicates discrimination. We illustrate how our definition correctly classifies this examples as unfair. Indeed, assuming the user labels the department D as admissible, $\{D\}$-fairness fails because $\Pr(O = 1|do(G = 1), do(D = \text{'A'})) = \sum_h \Pr(O = 1|G = 1, D = \text{'A'}, H = h)\Pr(H = h|G = 1) = \Pr(O = 1|G = 1, D = \text{'A'}) = 0.8$, and, similarly

$\Pr(O = 1|do(G = 0), do(D = \text{'A'})) = 0.2$. Therefore, the admission process is not justifiably fair.

Now, consider the second table for College II, where both departments A and B admit only on the basis of student qualifications Q. A superficial examination of the data suggests that the admission is unfair: department A admits 80% of all females and 100% of all male applicants; department B admits 20% and 44.4%, respectively. Upon deeper examination of the causal DAG, we can see that the admission process is justifiably fair because the only path from Gender to Outcome goes through Department, which is an admissible attribute. To understand how the data could have resulted from this causal graph, suppose 50% of each gender have high qualifications and are admitted, while others are rejected, and that 50% of females apply to each department, but more qualified females apply to department A than to B (80% vs 20%). Further, suppose fewer males apply to department A, but all of them are qualified. The algorithm satisfies demographic parity and proxy fairness but fails to satisfy conditional statistical parity since $\Pr(A = 1|G = 1, D = \text{A}) = 0.8$ but $\Pr(A = 1|G = 0, D = \text{A}) = 0.2)$. Thus, conditioning on D falsely indicates discrimination in College II. One can check that the algorithm is justifiably fair, and thus our definition also correctly classifies this example; for example, $\{D\}$-fairness follows from $\Pr(O = 1|do(G = i), do(D = d)) = \sum_q \Pr(O = 1|G = i, D = d, Q = q))\Pr(Q = q|G = i) = \frac{1}{2}$. To summarize, unlike previous definitions of fairness, justifiable fairness correctly identifies College I as discriminatory and College II as fair.

2.3 Impossibility Theorem from the Causality Perspective

From the point of view of causal DAGs, EO requires that the training label Y d-separates the sensitive attribute S and the outcome of the classifier O. Intuitively, this implies that S can affect classification results only when the information comes through the training label Y. On the other hand, PP requires that the classifier outcome O d-separates the sensitive attribute S and the training labels Y. Intuitively, this implies S can affect the training labels only when the information comes thorough the outcome of classifier O. These interpretations clearly reveal the inconsistent nature of EO and PP. It is easy to show for strictly positive distributions that the CIs $(S \perp\!\!\!\perp O|Y)$ and $(S \perp\!\!\!\perp Y|O)$ imply $(S \perp\!\!\!\perp Y)$ or, equivalently, $\Pr(Y = 1|S = 0) = \Pr(Y = 1|S = 1)$ (see [43]). Indeed, from the causality perspective, EO and PP are neither sufficient nor necessary for fairness. In the causal DAG in Fig 3(b), suppose a classifier is trained on an applicant's qualifications Q to approximate admission committee decisions $\hat{O}$. It is clear that the classifier is not discriminative, yet it violates both EO and PP. The reader can verify that the causal DAG obtained by further adding an edge from Q to $\hat{O}$ (to account for the classifier outcome) does not imply the CIs $(G \perp\!\!\!\perp O|\hat{O})$ and $(G \perp\!\!\!\perp \hat{O}|O)$.

3 Data Management Techniques for Causal Fairness

3.1 Causal Fairness as Integrity Constraints

In causal DAGs, the missing arrow between two variables X and Y represents the assumption of no causal effect between them, which corresponds to the CI statement $(X \perp\!\!\!\perp Y|\mathbf{Z})$, where $\mathbf{Z}$ is a set of variables that d-separates X and Y. For example, the missing arrow between O and G in the causal DAG in Fig. 2(a) encodes the CI $(O \perp\!\!\!\perp G|H, D)$. On the other hand, the lack of certain arrows in the underling causal DAG is sufficient to satisfy different causal notions of fairness (cf. Sec 2.2). For instance, a sufficient condition for justifiable fairness in the causal DAG in Fig. 2(a) is the lack of the edge from H to O, which corresponds to the CI $(O \perp\!\!\!\perp G, H|D)$. Thus, fairness can be captured as a set of CI statements. Now to enforce fairness, instead of intervening on the causal DAG over which we have no control, we can intervene on data to enforce the corresponding CI statements.

Consequently, social causal fairness constraints can be seen as a set of integrity constraints in the form of CIs that must be preserved and enforced thorough the data science pipeline, from data gathering through the deployment of a machine learning model. The connection between CIs and well-studied integrity constraints in data management – such as Multi Valued Dependencies (MVDs) and Embedded Multi Valued Dependencies

SQL Query:

```
SELECT avg(Income)
FROM AdultData
GROUP BY Gender
```

Gender	SQL Query	Rewritten Query
Female	*0.11*	0.10
Male	0.30	0.11

Coarse-grained Explanation:

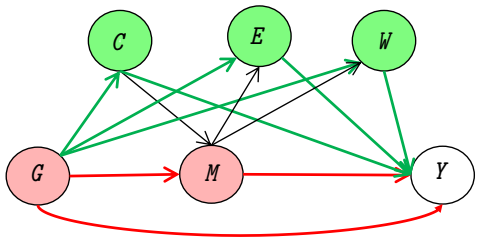
Attribute	*Res.*
MaritalStatus	*0.58*
Education	*0.13*
HoursPerwWeek	*0.04*
Age	*0.04*

Fine-grained Explanation:

Rank	MarialStatus	Gender	Income
1	Married	Male	1
2	Single	Female	0

Rank	Education	Gender	Income
1	Bachelors	Male	1
2	SomeCollage	Female	0

(a)



(b)

Figure 3: (a) HYPDB's report on the effect of gender on income (cf. Ex. 1). (b) A compact causal DAG with O = income, G = gender, M = marital status, C = age and nationality, E = education and W = work class, occupation and hours per week (cf. Ex. 3).

(EMVDs) [1] – opens the opportunity to leverage existing work in data management to detect and avoid bias in data.

3.2 Query Rewriting

In data management, *query rewriting* refers to a set of techniques to automatically modify one query into another that satisfies certain desired properties. These techniques are used to rewrite queries with views [19], in chase and backchase for complex optimizations [29], and for many other applications. This section discusses query rewriting techniques for detecting and enforcing fairness.

3.2.1 Detecting Discrimination

As argued in Sec 2.2, detecting discrimination should rely on performing a hypothesis test on the causal effect of membership in minority $S = 1$ or privileged group $S = 0$ on an outcome of an algorithm O. The gold standard for such causal hypothesis testing is a *randomized experiment* (or an *A/B test*), called such because treatments are randomly assigned to subjects. In contrast, in the context of fairness, sensitive attributes are typically imputable; hence, randomization is not even conceivable. Therefore, such queries must be answered using *observational data*, defined as data recorded from the environment with no randomization or other controls. Although causal inference in observational data has been studied in statistics for decades, causal analysis is not supported in existing online analytical processing (OLAP) tools [41]. Indeed, today, most data analysts still reach for the simplest query that computes the average of O Group By S to answer such questions, which, as shown in Ex 1, can lead to incorrect conclusions. Salimi et al. [41] took the first step toward extending existing OLAP tools to support causal analysis. Specifically, they introduced the HYPDB system, which brings together techniques from data management and causal inference to automatically rewrite SQL group-by queries into complex causal queries that support decision making. We illustrate HYPDB by applying it to a fairness question (see [40] for additional examples).

Example 3: Using UCI adult Census data [20], several prior works in algorithmic fairness have reported gender discrimination based on the fact that 11% of women have high income compared to 30% of men, which suggests a huge disparity against women. To decide whether the observed strong correlation between gender and high

income is due to discrimination, we need to understand its causes. To perform this analysis using HYPDB, one can start with the simple group-by query (Fig. 3(a)) that computes the average of Income (1 iff Income$>$ 50k) Group By Gender, which indeed suggests a strong disparity with respect to females' income. While the group-by query tells us gender and high income are highly correlated, it does not tell us why. To answer this question, HYPDB automatically infers from data that gender can potentially influence income indirectly via MaritalStatus, Education, Occupation, etc. (the indirect causal paths from G to O in Fig. 3(b)). Then, HYPDB automatically rewrites the group-by query to quantify the direct and indirect effect of gender on income. Answers to the rewritten queries suggest that the direct effect of gender on income is not significant (the effect through the arrow from G to O in Fig. 3(b)). Hence, gender essentially influences income indirectly through mediating variables. To understand the nature of this influences, HYPDB provides the user with several explanations. These show that MaritalStatus accounts for most of the indirect influence, followed by Education. However, the top fine-grained explanations for MaritalStatus reveal surprising facts: there are more married males in the data than married females, and marriage has a strong positive association with high income. It turns out that the income attribute in US census data reports the adjusted gross income as indicated in the individual's tax forms; these depend on filing status (jointly and separately), could be household income. HYPDB explanations also show that males tend to have higher levels of education than females, and higher levels of education is associated with higher incomes. The explanations generated by HYPDB illuminate crucial factors for investigating gender discrimination.

Future Extensions. Incorporating the type of analyses supported by HYPDB into data-driven decision support systems is not only crucial for sound decision making in general, but it is also important for detecting, explaining and avoiding bias and discrimination in data and analytics. Further research is required on extending HYPDB to support more complex types of queries and data, such as multi-relational and unstructured.

3.2.2 Enforcing Fairness

Raw data often goes through a series of transformations to enhance the clarity and relevance of the signal used for a particular machine learning application [3]. Filter transformations are perhaps most common, in which a subset of training data is removed based on predicates. Even if the raw data is unbiased, filtering can introduce bias [3, 41]: It is known that causal DAGs are not closed under conditioning because CIs may not hold in some subset. Hence, filtering transformations can lead to violation of causal fairness integrity constraints. It is also known that conditioning on common effects can further introduce bias even when the sensitive attribute and training labels are marginally independent [26]. This motivates the study of *fairness-aware data transformations*, where the idea is to minimally rewrite the transformation query so certain fairness constraints are guaranteed to be satisfied in the result of the transformation. This problem is closely related to that of constraint-based data transformations studied in [3]. However, fairness constraints go beyond the types of constraints considered in [3] and are more challenging to address. Note that a solution to the aforementioned problem can be used to enforce fairness-constraints for raw data by applying a fair-transformation that selects all the data.

3.3 Database Repair

Given a set of integrity constraints Γ and a database instance D that is inconsistent with Γ, the problem of repairing D is to find an instance D' that is close to D and consistent with Γ. Repair of a database can be obtained by deletions and insertions of whole tuples as well as by updating attributes. The closeness between D and D' can be interpreted in many different ways, such as the minimal number of changes or the minimal set of changes under set inclusion (refer to [6] for a survey). The problem has been studied extensively in database theory for various classes of constraints. It is NP-hard even when D consists of a single relation and Γ consists of functional dependencies [21].

Given a training data D that consists of a training label Y, a set of admissible variables $\mathbf{A}$, and a set of inadmissible variables $\mathbf{I}$, Salimi et al [43] showed that a sufficient condition for a classifier to be justifiably fair

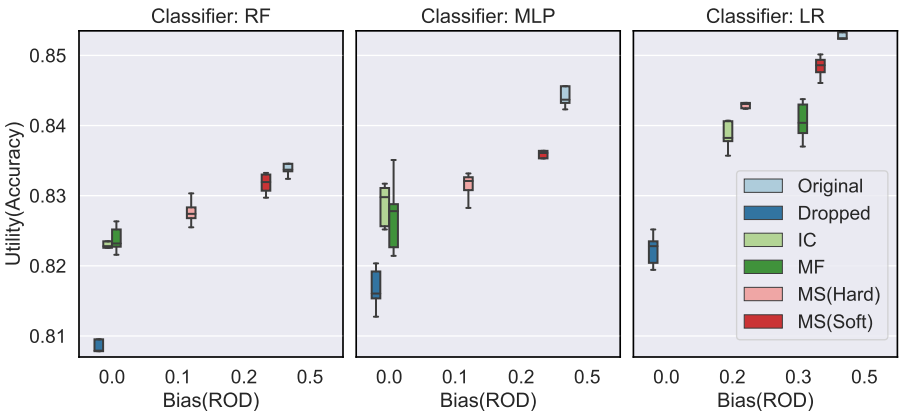


Figure 4: Performance of CAPUCHIN on Adult data.

is that the empirical distribution Pr over D satisfies the CI $(Y \perp\!\!\!\perp \mathbf{I}|\mathbf{A})$. Further, they introduced the CAPUCHIN system, which minimally repairs D by performing a sequence of database updates (viz., insertions and deletions of tuples) to obtain another training database D' that satisfies $(Y \perp\!\!\!\perp \mathbf{I}|\mathbf{A})$. Specifically, they reduced the problem to a minimal repair problem w.r.t. an MVD and developed a set of techniques, including reduction to the MaxSAT and Matrix Factorization, to address the corresponding optimization problem. We illustrate CAPUCHIN with an example.

Example 4: Suppose financial organisations use the Adult data described in Ex 1 to train an ML model to assist them in verifying the reliability of their customers. The use of raw data for training an ML model leads to a model that is discriminative against females simply because the model picks up existing bias in data, as described in Ex 3. To remove direct and indirect effects of gender on income (the red paths from G to Y in Fig. 4(b)) using the CAPUCHIN system, it is sufficient to enforce the CI $(O \perp\!\!\!\perp \mathbf{S}, \mathbf{M}|\mathbf{C}, \mathbf{E}, \mathbf{W})$ in data. Then, any model trained on the repaired data can be shown to be justifiably fair even on unseen test data under some mild assumptions [43]. To empirically assess the efficacy of the CAPUCHIN system, we repaired Adult data using the following CAPUCHINalgorithms: Matrix Factorization (MF), Independent Coupling (IC), and two versions of the MaxSAT approach: MS(Hard), which strictly enforces a CI, and MS(Soft), which approximately enforces a CI. Then, three classifiers – Linear Regression (LR), Multi-layer Perceptron (MLP), and Random Forest (RF) – were trained on both original and repaired training datasets using the set of variables $\mathbf{A} \cup \mathbf{N} \cup \mathbf{S}$. The classifier also trained on raw data using only $\mathbf{A}$, i.e., we dropped the sensitive and inadmissible variables. The utility and bias metrics for each repair method were measured using five-fold cross validation. Utility was measured by the classifiers' accuracy, and bias measured by the Ratio of Observational discrimination introduced in [43], which quantifies the effect of gender on outcome of the classifier by controlling for admissible variables (see [42] for details). Fig. 4 compares the utility and bias of CAPUCHIN repair methods on Adult data. As shown, all repair methods successfully reduced the ROD for all classifiers. The CAPUCHIN repair methods had an effect similar to dropping the sensitive and inadmissible variables completely, but they delivered much higher accuracy (because the CI was enforced approximately).

Future Extensions. The problem of repairing data w.r.t a set of CI constraints was studied in [43] for a single saturated CI constraint problem.[1] In the presence of multiple training labels and sensitive attributes, one needs to enforce multiple potentially interacting or inconsistent CIs; this is more challenging and requires further investigation. In addition, further research is required on developing approximate repair methods to be able to trade the fairness and accuracy of different ML applications.

[1] A CI statement is saturated if it contains all attributes.

3.4 Fairness-Aware Weak Supervision Methods

ML pipelines rely on massive labeled training sets. In most practical settings, such training datasets either do not exist or are very small. Constructing large labeled training datasets can be expensive, tedious, time-consuming or even impractical. This has motivated a line of work on developing techniques for addressing the data labeling bottleneck, referred to as *weak supervision methods*. The core idea is to programmatically label training data using, e.g., domain heuristics [31], crowdsourcing [32] and distant supervision [24]. In this context, the main challenges are handling noisy and unreliable sources that can potentially generate labels that are in conflict and highly correlated. State-of-the-art frameworks for weak supervision, such as Snorkel [30], handle these challenges by training label models that take advantage of conflicts between all different labeling sources to estimate their accuracy. The final training labels are obtained by combining the result of different labeling sources weighted by their estimated accuracy. While the focus of existing work is on collecting quality training labels to maximize the accuracy of ML models, the nuances of fairness cannot be captured by the exiting machinery to assess the reliability of the labeling sources. In particular, a new set of techniques is required to detect and explain whether certain labeling sources are biased and to combine their votes fairly.

3.5 Provenance for Explanation

Data provenance refers to the origin, lineage, and source of data. Various data provenance techniques have been proposed to assist researchers in understanding the origins of data [14]. Recently, data provenance techniques has been used to explain why integrity constraints fail [46]. These techniques are not immediately applicable to fairness integrity constraints, which are probabilistic. This motivates us to extend provenance to fairness or probabilistic integrity constraints in general. This extension is particularly crucial for reasoning about the fairness of training data collected from different sources by data integration and fusion, and it opens the opportunity to leverage existing techniques, such as provenance summarization [2], why-not provenance [8], and query-answers causality and responsibility [23, 38, 39, 5], explanations for database queries queries [33] to generate fine- and coarse-grained explanations for bias and discrimination.

4 Conclusions

This paper initiated a discussion on applying data management techniques in the embedding areas of algorithmic fairness in ML. We showed that fairness requires causal reasoning to capture natural situations, and that popular associational definitions in ML can produce incorrect or misleading results.

References

[1] Serge Abiteboul, Richard Hull, and Victor Vianu. *Foundations of Databases*. Addison-Wesley, 1995.

[2] Eleanor Ainy, Pierre Bourhis, Susan B Davidson, Daniel Deutch, and Tova Milo. Approximated summarization of data provenance. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, pages 483–492. ACM, 2015.

[3] Dolan Antenucci and Michael Cafarella. Constraint-based explanation and repair of filter-based transformations. *Proceedings of the VLDB Endowment*, 11(9):947–960, 2018.

[4] Chen Avin, Ilya Shpitser, and Judea Pearl. Identifiability of path-specific effects. 2005.

[5] Leopoldo Bertossi and Babak Salimi. Causes for query answers from databases: Datalog abduction, view-updates, and integrity constraints. *International Journal of Approximate Reasoning*, 90:226–252, 2017.

[6] Leopoldo E. Bertossi. *Database Repairing and Consistent Query Answering*. Synthesis Lectures on Data Management. Morgan & Claypool Publishers, 2011.

[7] Toon Calders, Faisal Kamiran, and Mykola Pechenizkiy. Building classifiers with independency constraints. In *Data mining workshops, 2009. ICDMW'09. IEEE international conference on*, pages 13–18. IEEE, 2009.

[8] Adriane Chapman and HV Jagadish. Why not? In *Proceedings of the 2009 ACM SIGMOD International Conference on Management of data*, pages 523–534. ACM, 2009.

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

[10] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 797–806. ACM, 2017.

[11] Rachel Courtland. Bias detectives: the researchers striving to make algorithms fair. *Nature*, 558, 2018.

[12] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226. ACM, 2012.

[13] Sainyam Galhotra, Yuriy Brun, and Alexandra Meliou. Fairness testing: testing software for discrimination. In *Proceedings of the 2017 11th Joint Meeting on Foundations of Software Engineering*, pages 498–510. ACM, 2017.

[14] Boris Glavic and Klaus Dittrich. Data provenance: A categorization of existing approaches. *Datenbanksysteme in Business, Technologie und Web (BTW 2007)–12. Fachtagung des GI-Fachbereichs" Datenbanken und Informationssysteme"(DBIS)*, 2007.

[15] Moritz Hardt, Eric Price, Nati Srebro, et al. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, pages 3315–3323, 2016.

[16] Niki Kilbertus, Mateo Rojas Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. Avoiding discrimination through causal reasoning. In *Advances in Neural Information Processing Systems*, pages 656–666, 2017.

[17] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In *Advances in Neural Information Processing Systems*, pages 4069–4079, 2017.

[18] Jeff Larson, Surya Mattu, Lauren Kirchner, and Julia Angwin. How we analyzed the compas recidivism algorithm. *ProPublica (5 2016)*, 9, 2016.

[19] Alon Y. Levy, Alberto O. Mendelzon, Yehoshua Sagiv, and Divesh Srivastava. Answering queries using views. In *Proceedings of the Fourteenth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, May 22-25, 1995, San Jose, California, USA*, pages 95–104, 1995.

[20] M. Lichman. Uci machine learning repository, 2013.

[21] Ester Livshits, Benny Kimelfeld, and Sudeepa Roy. Computing optimal repairs for functional dependencies. In *Proceedings of the 37th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems, Houston, TX, USA, June 10-15, 2018*, pages 225–237, 2018.

[22] Joshua R Loftus, Chris Russell, Matt J Kusner, and Ricardo Silva. Causal reasoning for algorithmic fairness. *arXiv preprint arXiv:1805.05859*, 2018.

[23] Alexandra Meliou, Wolfgang Gatterbauer, Katherine F Moore, and Dan Suciu. The complexity of causality and responsibility for query answers and non-answers. *Proceedings of the VLDB Endowment*, 4(1):34–45, 2010.

[24] Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2*, pages 1003–1011. Association for Computational Linguistics, 2009.

[25] Razieh Nabi and Ilya Shpitser. Fair inference on outcomes. In *Proceedings of the... AAAI Conference on Artificial Intelligence. AAAI Conference on Artificial Intelligence*, volume 2018, page 1931. NIH Public Access, 2018.

[26] Judea Pearl. Causality: models, reasoning, and inference. *Econometric Theory*, 19(675-685):46, 2003.

[27] Judea Pearl. *Causality*. Cambridge university press, 2009.

[28] Judea Pearl et al. Causal inference in statistics: An overview. *Statistics Surveys*, 3:96–146, 2009.

[29] Lucian Popa, Alin Deutsch, Arnaud Sahuguet, and Val Tannen. A chase too far? In *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, May 16-18, 2000, Dallas, Texas, USA.*, pages 273–284, 2000.

[30] Alexander Ratner, Stephen H Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Christopher Ré. Snorkel: Rapid training data creation with weak supervision. *Proceedings of the VLDB Endowment*, 11(3):269–282, 2017.

[31] Alexander J Ratner, Christopher M De Sa, Sen Wu, Daniel Selsam, and Christopher Ré. Data programming: Creating large training sets, quickly. In *Advances in neural information processing systems*, pages 3567–3575, 2016.

[32] Vikas C Raykar, Shipeng Yu, Linda H Zhao, Gerardo Hermosillo Valadez, Charles Florin, Luca Bogoni, and Linda Moy. Learning from crowds. *Journal of Machine Learning Research*, 11(Apr):1297–1322, 2010.

[33] Sudeepa Roy and Dan Suciu. A formal approach to finding explanations for database queries. In *Proceedings of the 2014 ACM SIGMOD international conference on Management of data*, pages 1579–1590. ACM, 2014.

[34] Donald B Rubin. *The Use of Matched Sampling and Regression Adjustment in Observational Studies*. Ph.D. Thesis, Department of Statistics, Harvard University, Cambridge, MA, 1970.

[35] Donald B Rubin. Statistics and causal inference: Comment: Which ifs have causal answers. *Journal of the American Statistical Association*, 81(396):961–962, 1986.

[36] Donald B Rubin. Comment: The design and analysis of gold standard randomized experiments. *Journal of the American Statistical Association*, 103(484):1350–1353, 2008.

[37] Chris Russell, Matt J Kusner, Joshua Loftus, and Ricardo Silva. When worlds collide: integrating different counterfactual assumptions in fairness. In *Advances in Neural Information Processing Systems*, pages 6414–6423, 2017.

[38] Babak Salimi and Leopoldo E. Bertossi. From causes for database queries to repairs and model-based diagnosis and back. In *ICDT*, pages 342–362, 2015.

[39] Babak Salimi, Leopoldo E Bertossi, Dan Suciu, and Guy Van den Broeck. Quantifying causal effects on query answering in databases. In *TaPP*, 2016.

[40] Babak Salimi, Corey Cole, Peter Li, Johannes Gehrke, and Dan Suciu. Hypdb: a demonstration of detecting, explaining and resolving bias in olap queries. *Proceedings of the VLDB Endowment*, 11(12):2062–2065, 2018.

[41] Babak Salimi, Johannes Gehrke, and Dan Suciu. Bias in olap queries: Detection, explanation, and removal. In *Proceedings of the 2018 International Conference on Management of Data*, pages 1021–1035. ACM, 2018.

[42] Babak Salimi, Luke Rodriguez, Bill Howe, and Dan Suciu. Capuchin: Causal database repair for algorithmic fairness. *arXiv preprint arXiv:1902.08283*, 2019.

[43] Babak Salimi, Luke Rodriguez, Bill Howe, and Dan Suciu. Interventional fairness: Causal database repair for algorithmic fairness. In *Proceedings of the 2019 International Conference on Management of Data*, pages 793–810. ACM, 2019.

[44] Camelia Simoiu, Sam Corbett-Davies, Sharad Goel, et al. The problem of infra-marginality in outcome tests for discrimination. *The Annals of Applied Statistics*, 11(3):1193–1216, 2017.

[45] Sahil Verma and Julia Rubin. Fairness definitions explained. In *2018 IEEE/ACM International Workshop on Software Fairness (FairWare)*, pages 1–7. IEEE, 2018.

[46] Jane Xu, Waley Zhang, Abdussalam Alawini, and Val Tannen. Provenance analysis for missing answers and integrity repairs. *IEEE Data Eng. Bull.*, 41(1):39–50, 2018.

[47] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th International Conference on World Wide Web*, pages 1171–1180. International World Wide Web Conferences Steering Committee, 2017.

[48] Indre Zliobaite. A survey on measuring indirect discrimination in machine learning. *arXiv preprint arXiv:1511.00148*, 2015.