FlipTest: Fairness Auditing via Optimal Transport

Emily Black[*]
Carnegie Mellon University
emilybla@andrew.cmu.edu

Samuel Yeom[*]
Carnegie Mellon University
syeom@cs.cmu.edu

Matt Fredrikson
Carnegie Mellon University
mfredrik@cs.cmu.edu

Abstract

We present FlipTest, a black-box auditing technique for uncovering subgroup discrimination in predictive models. Combining the concepts of individual and group fairness, we search for discrimination by matching individuals in different protected groups to each other, and comparing their classifier outcomes. Specifically, we formulate a GAN-based approximation of the optimal transport mapping, and use it to translate the distribution of one protected group to that of another, returning pairs of *in-distribution samples* that *statistically correspond to one another*. We then define the *flipset*: the set of individuals whose classifier output changes post-translation, which intuitively corresponds to the set of people who were harmed because of their protected group membership. To shed light on why the model treats a given subgroup differently, we introduce the *transparency report*: a ranking of features that are most associated with the model's behavior on the flipset. We show that this provides a computationally inexpensive way to identify subgroups that are harmed by model discrimination, including in cases where the model satisfies population-level group fairness criteria.

1 Introduction

With the recent use of machine learning in sensitive applications such as predictive policing [13] and child welfare [33], the question of whether this leads to unfair outcomes has gained widespread attention. These concerns are not merely hypothetical. Racial bias in the COMPAS recidivism prediction model [2] and gender bias in Amazons hiring model [8] show that discriminatory behaviors in models can have wide-reaching, and often irreversible, effects.

We present FlipTest, a *black-box, efficient, and interpretable* fairness auditing approach. Black-box audits assume the ability to query a model and record its outputs—thus they do not require the release of proprietary information from potentially reluctant companies, and are applicable to a wide range of model types. In contrast with prior work [22, 31] that performs a worst-case exponential search over possible subgroups, FlipTest only requires a pair of targeted queries per data point. Finally, the result of the audit is a corresponding pair of input examples that potentially cause the model to predict different outcomes, and thus offers an easily understood form of transparency into how models interact with notions of fairness.

Group fairness [1, 14, 17, 18, 21] describes a set of criteria which ask whether the model treats different protected groups similarly in aggregate. However, there are several downsides to group fairness metrics [6, 24**?**], among which is that models can "pass" group fairness audits while still behaving unfairly towards individuals, or even targeted subgroups [26]. Instead, FlipTest reasons about the model's behavior on individuals to look for evidence of discrimination based on membership in a protected group. To achieve this, we apply the optimal transport mapping [34] (Section 3) to map individuals from one group (e.g., men) to their counterpart in another (e.g., women). Using the optimal transport mapping to compare group outcomes is advantageous in that it generates

[*] Equal contribution

Preprint. Under review.

statistically correspondent pairs of individuals, both within the distribution used to train the model. This allows for a meaningful comparison without relying on a causal model of the data, and can be sufficient to establish discrimination according to law in many countries (e.g., *disparate impact* in the US [29] and *indirect discrimination* in the UK [28]).

We then examine the *flipset*, or set of points whose prediction changes under the optimal transport mapping, to identify the presence of discrimination, as well as the characteristics of the subgroup that the model discriminates against (Section 4). Membership in a flipset does not necessarily imply discrimination against a group or individual. For example, an inter-group disparity in the model's output may be permitted if there is a sufficient justification such as a "business necessity" [4, §II.B]. To shed light on such matters, we show how to construct *transparency reports* from the information contained in a flipset (Section 4.1), which identify the features that are most associated with the model's differential treatment. These results can be further refined by invoking complementary methods that establish causal relationships between the model's inputs and outputs [9].

The optimal transport mapping can be computed by solving a linear program (Section 3.1). Nonetheless, this approach does not scale to large datasets, and we show that the resulting mapping suffers from instability with respect to small changes in the data. To address both concerns, we formulate an approximation based on Generative Adversarial Networks (GANs) [15], and show that it is feasible to construct good, stable approximations of the precise mapping (Section 3.2). However, because this mapping is an approximation that may introduce noise at the individual level, we pose audits for subgroup discrimination rather than for single individuals.

To summarize, our main contributions are: *(1)* FlipTest, a black-box, efficient, interpretable auditing technique to detect discrimination in machine learning models; *(2)* a GAN-based approximation of the optimal transport mapping (Section 3.2), which we demonstrate empirically is well-suited to the needs of FlipTest (Section 3.3); and *(3)* the application of FlipTest to two case studies involving predictive policing (Section 4.2) and hiring (Section 4.4), which demonstrate that our approach can efficiently identify concrete examples of unfair model behavior even in cases where the model satisfies group fairness criteria.

2 Related Work

Counterfactual Fairness and Audit FlipTest is a counterfactual notion of fairness in the sense that we compare the model's behavior on a real input and a counterfactual, generated input. The work of Kusner et al. [25] also take this approach, but their counterfactual point is generated through a causal graph. Similarly, Datta et al. [9] perform causal interventions on input features to study which features are influential in changing in the output of the model, Wachter et al. [35] generate simple L_1-nearest counterfactuals as a form of explanations for model outputs, and Ustun et al. [32] develop a method that outlines what actions individuals can take to change their classification outcome in linear models. However, these methods generate potentially unrealistic, out-of-distribution points, which can jeopardize the conclusions drawn from these approaches. By contrast, the counterfactual points that we generate conform to the data distribution.

Optimal Transport Others have proposed using the optimal transport map in the context of fairness, but to the best of our knowledge, it has not yet been used as an auditing or transparency mechanism. Del Barrio et al. [10] use the optimal transport mapping to extend to the multivariate setting a previous method [14] of pre-processing univariate data to make the data independent of the protected attribute. Concurrent work from Yang et al. [37] also develop a method of approximating an (unbalanced) optimal transport mapping using GANs. Their formulation is closely related to ours, but they do not consider its application to auditing fairness.

Individual Fairness Individual fairness criteria [11, 12, 20, 40] bind guarantees about the fairness of a model's behavior to every individual, as opposed to an aggregated statistic. Dwork et al. [12] note that, in some cases, the model must essentially be a constant function to satisfy individual fairness and group fairness at the same time. They propose an alternative that applies an optimal transport mapping to one of the groups, obtaining a transformed dataset on which they solve individual fairness. Our audit is motivated by this approach, but we specifically look for potentially discriminatory differences between pairs of individuals *that belong to different groups*, and use optimal transport to construct pairs of individuals that exemplify these differences.

Subgroup Fairness One application of this work is uncovering *subgroup unfairness* [19**?**], i.e., identifying subgroups that are possibly harmed as a result of their group membership. Previously, the authors of FairTest [31] used a decision tree to find subgroups with high discriminatory association, while taking care to ensure that the association is statistically significant. Kearns et al. [22] prove that checking for subgroup fairness is equivalent to weak agnostic learning, which is computationally hard in the worst case. Our audit differs from these works in that we do not require the auditor to specify the subgroups of interest before the audit. Zhang et al. [41] find a computationally faster way to search the exponentially large number of subgroups, but their method also relies solely on group fairness, which we demonstrate can leave some forms of bias undetected.

3 Optimal Transport Mapping

In this section, we describe the optimal transport problem and formulate a linear program that solves for an exact optimal transport mapping (Section 3.1). Because this formulation does not scale to large datasets, and additionally suffers from instability, we show how to approximate the mapping by solving a GAN objective (Section 3.2). We conclude by showing that the approximation is suitably stable, while still emulating the precise formulation well (Section 3.3).

We first introduce the notation we will use throughout this section and the next. Let $\mathcal{S}$ and $\mathcal{S}'$ be two distributions defined over the feature space $\mathcal{X}$. In practice, we do not know these distributions, so we usually deal with observations of points drawn from these distributions instead. We will use the sets $S = \{\boldsymbol{x}_1, \dots, \boldsymbol{x}_n\}$ and $S' = \{\boldsymbol{x}'_1, \dots, \boldsymbol{x}'_n\}$ to denote the observed points, where $n = |S| = |S'|$.[2]

Let $c : \mathcal{X} \times \mathcal{X} \to [0, \infty)$ be a cost function that describes the cost of moving between two points in the feature space $\mathcal{X}$. Intuitively, an optimal transport mapping from S to S' is a minimum-cost way to move the points in S such that the end result is S'. Thus, if more similar pairs of points have a lower cost, an optimal transport mapping describes how to match points in S with their similar counterparts in S'.‘‘ Formally, an optimal transport mapping can be defined as a bijection $f : S \to S'$ that minimizes the expected cost $\mathbb{E}[c(\boldsymbol{x}, f(\boldsymbol{x}))] = \frac{1}{n}\sum_{i=1}^{n} c(\boldsymbol{x}_i, f(\boldsymbol{x}_i))$.

3.1 Linear Programming Formulation

Let $c_{ij} = c(\boldsymbol{x}_i, \boldsymbol{x}'_j)$, and let t_{ij} be an indicator variable representing whether $f(\boldsymbol{x}_i) = \boldsymbol{x}'_j$. Then, we can optimize the following linear program over the t_{ij}'s to compute an optimal transport mapping:

$$\begin{array}{lll} \text{minimize} & \sum_{i=1}^{n}\sum_{j=1}^{n} c_{ij}t_{ij} & \\ \text{subject to} & \sum_{j=1}^{n} t_{ij} = 1, & \forall i \\ & \sum_{i=1}^{n} t_{ij} = 1, & \forall j \\ & t_{ij} \geq 0, & \forall i, j \end{array} \tag{1}$$

Since the feasible region in (1) is bounded, the minimum is attained at one of the vertices of the feasible region. In this problem, the vertices are the points such that $t_{ij} = 0$ or 1 for all i and j, so we have a 0-1 solution. Then, we can define an optimal transport mapping as $f(\boldsymbol{x}_i) = \sum_{j=1}^{n} t^*_{ij}\boldsymbol{x}'_j$, where t^*_{ij} is the value of t_{ij} in the 0-1 solution to (1).

3.2 Approximation via Generative Adversarial Networks

While the linear program in the previous section gives us an exact optimal transport mapping, solving it does not scale well to large values of n. In addition, the mapping given by (1) is not defined for points outside of S. In this section, we show how to use a generative adversarial network (GAN) [15] to approximate an optimal transport mapping in a way that avoids both of these issues.

In a typical GAN, the inputs to the generator are drawn from a simple distribution, such as a Gaussian, in order to provide the generator with a source of entropy for its output distribution. However, in this case we want to use G as an optimal transport mapping, so we draw the inputs from $\mathcal{S}$.

[2] Here, we assume that $|S| = |S'|$ for ease of exposition, as this assumption allows the resulting exact optimal transport mapping to be deterministic. The general case where $|S| \neq |S'|$ can be handled through the use of randomized optimal transport mappings, and our approximation method does not require that the two sets have equal size.

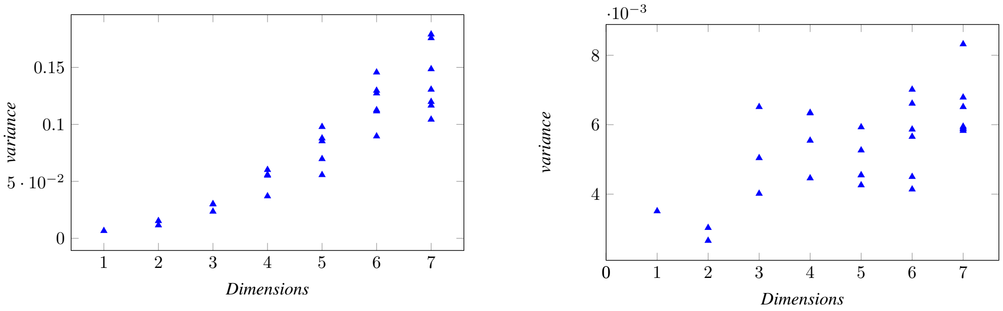


Figure 1: Variance of the exact optimal transport mapping (left) and the GAN approximation (right) over the different random draws of the observed points S and S'. Note the difference in the vertical scales: the GAN approximation has much smaller variance, order of 1×10^{-3}. The horizontal axis represents the number of dimensions in the feature space $\mathcal{X}$, and each plotted point represents the variance of one of the features. More details are given in Section 3.3.

For concreteness, we now use a Wasserstein GAN [3] as an example, but we note that Theorem 1 can easily be extended to other types of GANs as well. When training a typical Wasserstein GAN with the sets of observed points S and S', the generator's loss function is $\frac{1}{n}\sum_{\boldsymbol{x}\in S} D(G(\boldsymbol{x}))$, and the discriminator's loss function is $\frac{1}{n}\sum_{\boldsymbol{x}'\in S'} D(\boldsymbol{x}') - \frac{1}{n}\sum_{\boldsymbol{x}\in S} D(G(\boldsymbol{x}))$. For the purpose of finding an optimal transport mapping, we modify the generator's loss function to take into account the cost of moving from a point in S to a point in S':

$$L_G = \frac{1}{n}\sum_{\boldsymbol{x}\in S} D(G(\boldsymbol{x})) + \frac{\lambda}{n}\sum_{\boldsymbol{x}\in S} c(\boldsymbol{x}, G(\boldsymbol{x})) \tag{2}$$

Our modified generator has two objectives, with the parameter λ controlling their relative importance: generating the correct output distribution $\mathcal{S}'$, and minimizing the expected cost $c(\boldsymbol{x}, G(\boldsymbol{x}))$. Theorem 1 formalizes the intuition that these objectives are also those of an optimal transport mapping.

Theorem 1. *Suppose that G^* is a minimizer of L_G among all G such that $G(S) = S'$. Then, G^* is an exact optimal transport mapping from S to S'.*

Proof. If $G(S) = S'$, G induces a bijection between S and S', so we have $\frac{1}{n}\sum_{\boldsymbol{x}\in S} D(G(\boldsymbol{x})) = \frac{1}{n}\sum_{\boldsymbol{x}'\in S'} D(\boldsymbol{x}')$. Then, (2) becomes

$$L_G = \frac{1}{n}\sum_{\boldsymbol{x}'\in S'} D(\boldsymbol{x}') + \frac{\lambda}{n}\sum_{\boldsymbol{x}\in S} c(\boldsymbol{x}, G(\boldsymbol{x})).$$

The first term does not depend on G, so if G^* minimizes L_G, it also minimizes $\frac{\lambda}{n}\sum_{\boldsymbol{x}\in S} c(\boldsymbol{x}, G(\boldsymbol{x}))$. This is exactly the definition of an optimal transport mapping. □

Although the generator G will not satisfy $G(S) = S'$ in practice, Theorem 1 motivates the use of this generator to approximate an optimal transport mapping. This GAN approximation has several advantages over the linear program (1). First, the linear program creates n^2 variables, so it takes at least $\Omega(n^2)$ time to solve the linear program. By contrast, adding more training points to GANs does not increase the training time as much. Second, unlike the linear program mapping, the GAN generator mapping is defined for all points in the distribution $\mathcal{S}$, thus it can be used to map previously unseen points as well. Finally, as we will show in Section 3.3, an exact optimal transport mapping has a tendency to "regress toward the mean" of the distribution $\mathcal{S}'$. Furthermore, the mapping is not very stable and changes drastically depending on which sets S and S' were drawn from $\mathcal{S}$ and $\mathcal{S}'$. Our experimental results show that the GAN mapping is more stable and resistant to regression to the mean.

3.3 Stability

First, we compare the behavior of an exact optimal transport mapping to that of our GAN approximation. We fix a point $\boldsymbol{x} \in S$ and then draw multiple different samples of the other $n-1$ points from $\mathcal{S}$ and n points from $\mathcal{S}'$. Thus, we have different sampled sets S and S' each time, and we can observe how the random drawing affects the point $f(\boldsymbol{x})$.

In all of the experiments in this paper, we used the L_1 distance as the cost function. The exact optimal transport mapping (Section 3.1) was computed with Gurobi [16] on Python 3, and for the GAN approximation (Section 3.2), we trained a Wasserstein GAN [3] using Keras [5] with a Theano backend [30] on Python 3. For these experiments, we mapped the standard multivariate normal distribution to itself. Because the size of the linear program in Section 3.1 increases at least quadratically with the size n of the dataset, we used $n = 500$ in these experiments to make the runtime reasonably low. Each experiment was repeated with 100 different random draws of the dataset.

In the first set of experiments, we set $\boldsymbol{x}$ to the zero vector and noted the variance of $f(\boldsymbol{x})$ under both types of mappings. The result is plotted in Figure 1 and shows that the GAN approximation is much more stable than the exact mapping. In the second set of experiments, we set $\boldsymbol{x}$ to be the one vector and observed the mean of $f(\boldsymbol{x})$. Since we map a distribution to itself, f should roughly be the identity function, and the mean of $f(\boldsymbol{x})$ should be similar to $\boldsymbol{x}$. While this was the case for the GAN approximation, the exact mapping displayed a significant "regression-to-the-mean" effect that increased with the number of dimensions in the feature space. The differences in both mean and variance persisted when we changed the data distribution by making the features correlated with each other.

These differences can be explained by the fact that the exact mapping tends to overfit to the observed points, since it has to map every point to an observed point. On the other hand, GANs provide some implicit regularization [27] that makes the mapping generalize better. As a result, the GAN approximation is better suited for evaluating the fairness of a model that is trained to generalize, and we will exclusively use GANs for the rest of the experiments.

4 Application to Detecting Discrimination

We leverage the optimal transport mapping to gather two main pieces of information from a machine learning model: who may experience discrimination, and which features are associated with this effect. In Section 4.1, we describe *flipsets*, which we use to answer the first question, and show how they are used to construct *transparency reports* that answer the second. We then illustrate the application of these ideas to a biased predictive policing model (Section 4.2), as well as a hiring model (Section 4.4) that contains a subtle form of subgroup discrimination while maintaining group fairness. FlipTest uncovers discrimination in both classifiers by pointing out unusually large flipsets, and identifies the correct source of bias in each model through the transparency report.

4.1 Flipsets and Transparency Reports

We begin by introducing the *flipset* (Definition 1), which is the set of points whose image under a transport mapping is assigned a different label by a binary classifier.

Definition 1 (Flipset)**.** *Let $h : \mathcal{X} \to \{0, 1\}$ be a classifier and $G : \mathcal{S} \to \mathcal{S}'$ be a generator defined by the loss in Equation 2. The flipset $F(h, G)$ is the set of points in S whose mapping into $\mathcal{S}'$ under G changes classification.*

$$F(h, G) = \{\boldsymbol{x} \in S | h(\boldsymbol{x}) \neq h(G(\boldsymbol{x}))\} \tag{3}$$

The positive *and* negative *partitions of $F(h, G)$ are denoted by $F^+(h, G)$ and $F^-(h, G)$.*

$$F^+(h, G) = \{\boldsymbol{x} \in S | h(\boldsymbol{x}) > h(G(\boldsymbol{x}))\}, \quad \textit{and} \quad F^-(h, G) = \{\boldsymbol{x} \in S | h(\boldsymbol{x}) < h(G(\boldsymbol{x}))\} \tag{4}$$

In our experiments, $\mathcal{S}$ and $\mathcal{S}'$ will correspond to two groups with differing values for a protected attribute, and h will be a classifier with the potential to be unfair. For example, suppose that $\mathcal{S}$ and $\mathcal{S}'$ respectively correspond to female and male job applicants and that h is used to determine which applicants should proceed to further rounds of interview. Then $F^+(h, G)$ is the set of female applicants who proceed to the next round but whose male counterparts under G do not, and $F^-(h, G)$ is the women who do not proceed but whose male counterparts do.

Intuitively, if the distributions $\mathcal{S}$ and $\mathcal{S}'$ are equal, then the model's input features are independent of the protected attribute. Thus, the model cannot possibly discriminate on the basis of the protected attribute. Theorem 2 formalizes this intuition, stating the conditions under which the flipsets will be empty.

Theorem 2. *Let h be a predictive model, and let $G : \mathcal{S} \to \mathcal{S}'$ be a generator. If G converges to the optimal transport mapping between samples S and S' and the cost function c is a distance metric, then $S = S'$ implies $|F^+(h, G)| = |F^-(h, G)| = 0$.*

Proof. Since $S = S'$, the identity mapping id is a valid transport mapping. Moreover, c is a distance metric, so $c(\boldsymbol{x}, \boldsymbol{x}') = 0$ if and only if $\boldsymbol{x} = \boldsymbol{x}'$, which means that id is the unique optimal transport mapping. Then, we have $h(\boldsymbol{x}) = h(\mathrm{id}(\boldsymbol{x})) = h(G(\boldsymbol{x}))$ for all $\boldsymbol{x}$, so the flipsets are empty. □

In practice, the flipsets will not be empty for two reasons: (1) the GAN does not in general converge to the optimal transport mapping, and (2) $\mathcal{S} = \mathcal{S}'$ does not imply $S = S'$.

In order to verify that the first factor does not introduce enough error to invalidate the approach, we apply our method to a fair model as a control. We trained a fair model by ensuring that, from the perspective of the model, there are no distributional differences between the two protected groups. To introduce more error into the GAN mapping, we added other random features that are *dependent* on the protected attribute. More specifically, we set $\mathcal{S}$ and $\mathcal{S}'$ respectively to the normal distributions $\mathcal{N}(\boldsymbol{\mu}_{\mathcal{S}}, \mathbf{I}_6)$ and $\mathcal{N}(\boldsymbol{\mu}_{\mathcal{S}'}, \mathbf{I}_6)$, where $\boldsymbol{\mu}_{\mathcal{S}} = (0, 0, 0, 1, 1, 1)^T$, $\boldsymbol{\mu}_{\mathcal{S}'} = (0, 0, 0, -1, -1, -1)^T$, and $\mathbf{I}_6$ is the 6x6 identity matrix. Thus, the first three features had the same joint distribution for both groups, and the model was allowed to use only these three features.

We constructed the training set by drawing 10,000 points from each of $\mathcal{S}$ and $\mathcal{S}'$. In order to amplify any errors in the GAN mapping, we make the model arbitrary and complicated: we gave each point a binary class label uniformly at random and then trained with this data an SVM classifier with a radial basis function kernel. Finally, we trained a GAN ($\lambda = 10^{-4}$) using a test set, which was drawn from $\mathcal{S}$ and $\mathcal{S}'$ in the same way as the training set, and observed the size of the flipset on the test set.

Our GAN correctly mapped the distribution $\mathcal{S}$ to $\mathcal{S}'$, changing each of the first three features by less than 0.01 on average and the other three by close to 2 on average. As a result, even though the model had a very irregular decision boundary, with approximately 63% training accuracy (the test accuracy was, of course, close to 50%) and 53% positive classification rate, the flipsets were small. Only 167 of the 10,000 points flipped from positive to negative under the GAN mapping, and 148 of the 10,000 points flipped from negative to positive. These numbers stayed similar when we altered the distributions $\mathcal{S}$ and $\mathcal{S}'$ so that the features were not necessarily normal and uncorrelated.

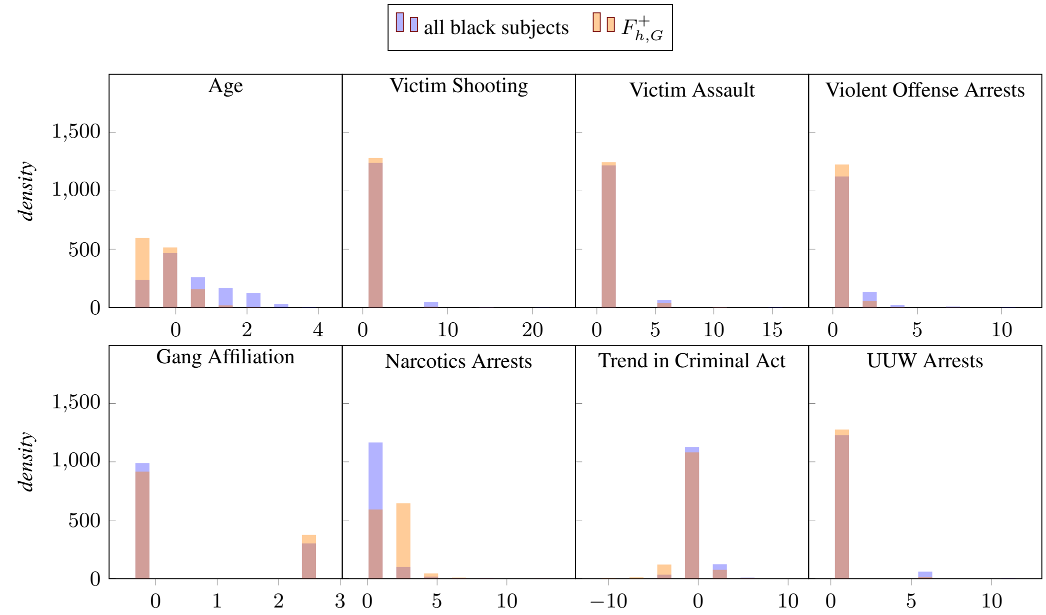
When h may be biased, two quantities derived from the flipset can provide useful information for audit: the relative magnitudes of $F^-(h, G)$ and $F^+(h, G)$, and the difference between marginal distributions of the flipset and $\mathcal{S}$. The relative sizes of $F^-(h, G)$ and $F^+(h, G)$ give a first indication that there may be subgroup discrimination. In addition, by examining the difference between the marginals of $F^+(h, G)$, $F^-(h, G)$, and $\mathcal{S}$, we gain more information about which specific subgroups may be discriminated against. In particular, when the flipset marginal differs from the general population, the locations receiving more mass in the flipset correspond to subgroups for which the model might be discriminatory; we develop this insight further in Sections 4.2 and 4.4. A *transparency report* (Definition 2) supplements this information by identifying features that change the most, and most consistently, under G. These features are likely candidates for the underlying "cause" of the apparently discriminatory behavior, and can be examined further using causal methods [9] to make a final determination about their causal influence.

Definition 2 (Transparency Report)**.** *Let $h : \mathcal{X} \to \{0, 1\}$ be a classifier, $G : \mathcal{S} \to \mathcal{S}'$ be a generator defined by the loss in Equation 2, and $F(h, G)$ be the corresponding flipset. If $\mathcal{X} \subseteq \mathbb{R}^d$, we can compute the following vectors, each of whose coordinate corresponds to a feature in $\mathcal{X}$:*

$$\frac{1}{|F^+(h,G)|} \sum_{\boldsymbol{x} \in F^+(h,G)} \boldsymbol{x} - G(\boldsymbol{x}) \quad \textit{and} \quad \frac{1}{|F^+(h,G)|} \sum_{\boldsymbol{x} \in F^+(h,G)} \mathrm{sign}(\boldsymbol{x} - G(\boldsymbol{x})) \tag{5}$$

Together, these vectors define a transparency report*, which consists of two rankings of the features in $\mathcal{X}$, each sorted by the absolute value of each coordinate.*

Intuitively, these features ranked highest by the transparency report are those that are most statistically associated with the model's differences in behavior on the flipset. As we show in Sections 4.2 and 4.4, these often align precisely with the features used by the model to discriminate.



Feature	Mean Sign
Narcotic Arr	0.99
Trend	−0.78
Gang Aff	0.26
Vio Off	0.04
Vic Assault	0.03
Age	−0.01
UUW Arr	0.00
Vic Shooting	0.00

Feature	Mean Diff
Narcotic Arr	1.32
Gang Aff	0.71
Trend	−0.20
Vic Assault	0.15
Vio Off	0.11
UUW Arr	0.20
Age	−0.01
Vic Shooting	0.01

Figure 2: Distribution of the high risk to low risk flipset for a black to white mapping (orange) and the overall black subject population (purple). On the x axis we have (normalized) feature value; y axis is number of individuals. The distribution of $F^+(h, G)$ is disproportionately young when compared to that of the full distribution of black individuals, and similarly, disproportionately consists of those with prior drug arrests. Meanwhile, other features included for comparison, more or less share the same distribution as the overall black population arrested. We did not plot the flipset in the other direction, as it only has 4 individuals. The transparency report, as noted in Section 4.2, identifies the feature that the model relies on to bring about bias; narcotic arrests. Recall that all features are scaled to zero mean and unit variance.

4.2 Auditing a Biased Model: Chicago SSL Dataset

The Chicago Strategic Subject List (SSL) dataset [7] consists of arrest data collected in Chicago for the purpose of identifying which individuals are likely to be involved in a violent crime, either as a victim or a perpetrator. We used the following eight features that are also used by the SSL model [7]: number of times as victim of a shooting incident, age during last arrest, number of times as victim of aggravated battery or assault, number of prior arrests for violent offenses, gang affiliation, number of prior narcotic arrests, trend in recent criminal activity and number of prior unlawful use of weapon arrests. The target feature corresponds to the risk of being involved in a shooting, ranging in value from 0 to 500 (low to high risk). A standard least-squares regression model for this data primarily relies on age ($w_{age} \approx -50$), giving the remaining features coefficients of magnitude less than 10.

We create a classification task from this dataset by setting a threshold on score (345); 10% of the dataset has a score above this threshold. As prior work [38] has shown that models trained on this data do not appear to exhibit significant bias, we create a biased classifier that discriminates against black subjects by intervening on the coefficients of features correlated with race.

The first biased classifier only relies on age and prior narcotic arrests, assigning $w_{age} = 50$, $w_{narc} = 25$. In the SSL data, narcotic arrests are positively correlated with race, ($cov = 0.12$), but not many features that would suggest high crime risk, such as previous violent crime count or weapons arrests. By relying heavily on narcotics arrests, we make a biased classifier that is not necessarily accurate. Thus, we expect this model to assign black subjects higher average scores. We calibrate the threshold so that the model labels approximately 10% of subjects as high-risk using holdout data. On the test data, this results in 8% of all people labeled high risk, with 5% (out of 16,465) of white subjects and 9% (out of 41,560) of black subjects. We construct an optimal transport GAN to map black SSL subjects to white ones after scaling the features to zero mean and unit variance.

Specifically, we use a Wasserstein GAN [3] using Keras [5] with a Theano backend [30] on Python 3. Our Wasserstein GAN has a generator and a critic both with two dense layers of size 128; the critic has an output layer of size 1. We use relu activations on both discriminator and generator. We use the RMSProp optimizer with a learning rate of 5×10^{-5}. We train for 20,000 epochs, have a weight clip value of 0.01, and train the critic 5 times more than we train the generator. The generator output has an additional L1-loss of $\lambda = 5 \times 10^{-4}$. We use a batch size of 4. We train the GAN on 1/4 of the SSL data [7], separated into white and black groups, with only the 8 features mentioned above. This corresponds to 41,560 black subjects and 16,465 white subjects. Throughout all experiments in this section, we use random seed 100.

We calculate the flipsets based on this mapping, and find that 1290 black subjects that are marked as high risk by the model are marked as low risk when sent through the black to white mapping (out of 3,683 black subjects marked as high risk). However, only 4 black subjects marked as low risk are mapped to high-risk white subjects (out of 37,877 marked as low risk). The size of $F^+(h, G)$ aand extreme asymmetry of the flipsets suggests that there is a relationship between being black and being labeled as high risk. This reflects the reality of the model, which relies on a feature correlated with race to make its decision. From the flipset alone, however, we cannot tell if this relationship between race and outcome is a necessity or a true source of bias. To see this, we look at the marginal distribution of the flipset, and the transparency report.

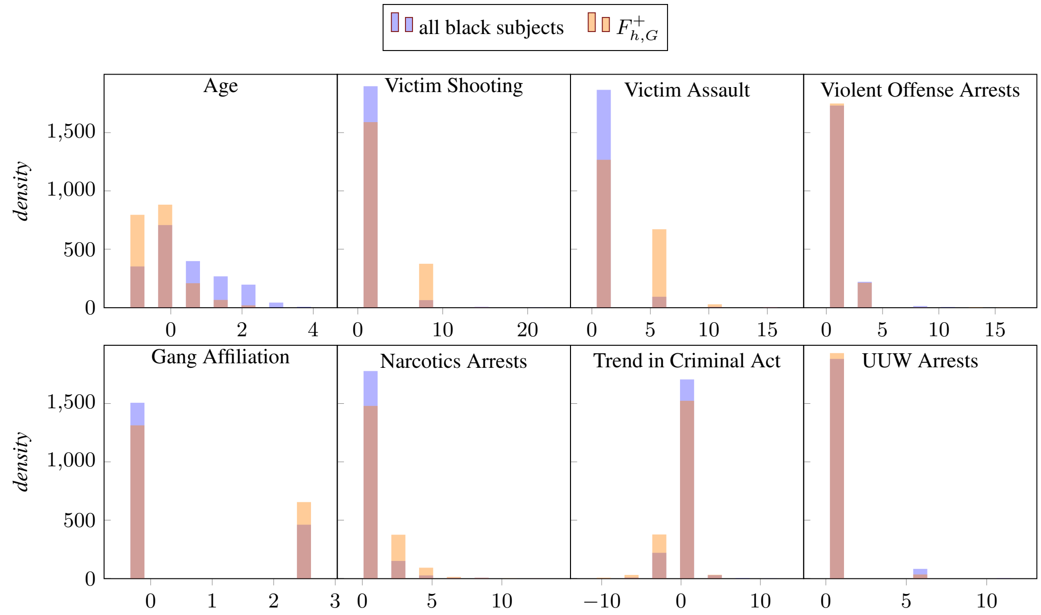
The marginal distributions of each feature, measured on $F^+(h, G)$, is shown in Figure 2. For age and narcotic arrests, the flipset subjects skew away from the full population, towards younger people with more narcotic arrests. This suggests the bias of the model affects younger people with more narcotic arrests most. Notice that the marginals largely overlap for gang affiliation, indicating that black subjects whose white optimal transport correspondents are assigned low risk do not differ from the full population of black subjects in this feature.

We can look at the transparency report (Figure 2) to examine this more closely. The feature with the highest mean difference and most consistent sign of difference is narcotic arrests, which is consistent with the type of bias that we introduced into this model. Thus, although the transparency reports only provide direct insight into statistical correlations between features and model behavior, in this case they identified precisely the feature used by the model to cause discrimination. In general, based on the information from the transparency report, an auditor can decide whether the possible reason for the model's differing treatment of the flipset—in this case narcotic arrests—is an justified.

4.3 Auditing a Biased Model with Several Biased Features: Chicago SSL

We also provide an audit of a biased classifier that relies on a combination of features to be discriminatory. We choose the three features that are the most positively correlated with race and least correlated with a history of violent crime: victim of shooting incident, victim of assault, and narcotic arrests. We set $w_{age} = 50$, $w_{narc} = 20$, $w_{vs} = 20$, $w_{va} = 20$. We give all other features weight 0. We use the same GAN and classifier calibration procedure as in the first biased classifier experiment. We also scale the features in the same way, to zero mean and unit variance. We find that there are 1,968 high-risk black subjects that get mapped to low-risk white counterparts (out of 3,683 total high risk black subjects), and 0 low-risk black subjects that get mapped to high-risk white counterparts. Again, the extreme asymmetry of $F^+(h, G)$ and $F^-(h, G)$ along with the size of $F^+(h, G)$ suggest there is some discrimination in the model.

The marginal distributions can be seen in 3: young, black individuals with nonzero accounts of being a victim of a shooting and assault, with nonzero narcotic arrests, are discriminated against by this



Feature	Mean Sign
Trend	−0.54
Narcotic Arr	0.43
Vic Assault	0.39
Vic Shooting	0.23
Gang Aff	0.10
Vio Offense	0.09
UUW Arr	0.01
Age	−0.01

Feature	Mean Diff
Vic Shooting	1.86
Vic Assault	1.50
Narcotic Arr	0.62
Gang Aff	0.26
Vio Offense	0.22
Trend	−0.09
UUW Arr	0.05
Age	−0.01

Figure 3: Distribution of the high risk to low risk flipset for a black to white mapping. We did not plot the flipset in the other direction, as it was empty. We see that this model adversely affects young black individuals who have been the victim of shooting and assault incidents, who largely have nonzero narcotic arrests and have a slightly higher chance of being in a gang. Note that the distribution over violent offenses, and several other features, is relatively unchanged between the regular distribution and the flipset, suggesting that these features do not define the flipset distribution. The transparency report gives a strong case to investigate narcotic arrests, being a victim of a shooting, or assault, and possibly trend in recent criminal activity as features to investigate for causal influence. Three of these features were the ones we chose to overweight to make the model unfair.

model. They are also more likely to be in a gang and to not have a record of unwarranted weapons arrests.

Upon viewing the transparency report, to determine which features may be a worth further investigation into the model's workings, we see that being a victim of assault, of a shooting, and having a history of narcotic arrests are the features with the highest mean difference in the flipset. Trend in recent criminal activity is the most consistent, negative change, but it is closely followed by narcotic arrests and then victim of assault and shooting. Thus, again, the model's use of the features align with the correlations shown in the transparency report.

4.4 Auditing a Group-Fair Model: Lipton et al. Dataset

Previously, Lipton et al. [26] argued that some fair learning algorithms employ a "problematic within-class discrimination mechanism". To support this argument, they create a synthetic data distribution targeted towards hiring that consists of two features: work experience and hair length. They use this distribution to train a learning algorithm by Zafar et al. [39] that seeks to equalize hiring rates across genders. They find that the resulting linear model uses hair length as a proxy for gender, thereby unfairly benefiting long-haired men and harming short-haired women. In this sec-

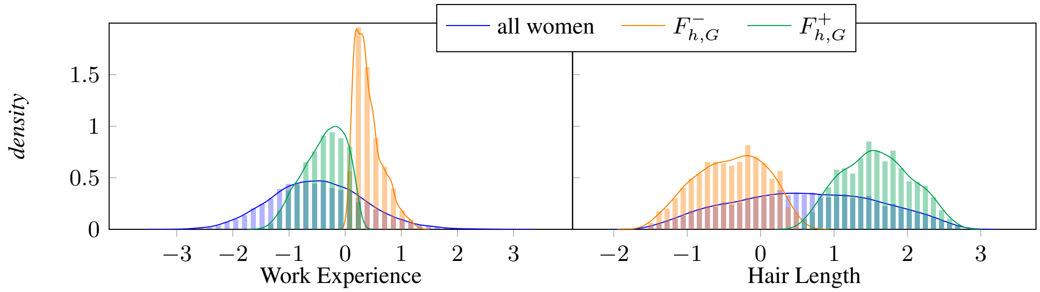


Figure 4: Distribution of women in the hiring data generated by Lipton et al. [26] (Section 4.4). The green distribution is that of the women who were hired while their male statistical counterparts were not, and the orange distribution represents those who were not hired while their male statistical counterparts were, according to a model of Zafar et al. [39] that seeks to equalize hiring rates across genders.

tion, we replicate this experimental setting to demonstrate that our method can detect unfair behavior in a model that appears to be fair at the population level.

We trained a linear model on 10,000 men and 10,000 women drawn from the male and female distributions, respectively, after scaling each feature to zero mean and unit variance. Then, we trained an optimal transport GAN with the same experimental setup as in Section 4.2, with the exceptions of $\lambda = 10^{-4}$ and the batch size $b = 64$. All experiments in this section use random seed 0. We evaluated the fairness of the model on a test set of 10,000 men and 10,000 women drawn from the same distributions. At the population level, the model treated the two groups similarly, hiring 27% of the men and 30% of the women. However, when we mapped the women to the men 715 women were disadvantaged by the model (rejected with hired male counterpart) whereas 1215 were advantaged (hired with rejected male counterpart).

These flipsets comprise a much larger portion of the population than those encountered in Section 4.1, suggesting some discriminatory behavior at the subgroup level. Looking more closely at distributions of these flipsets (Figure 4) provides insight into the subgroups experiencing discrimination: the disadvantaged women tend to have shorter hair and longer work experience than the average woman, and the advantaged women have the opposite characteristics. This is consistent with the observation of Lipton et al. [26] that the model penalizes people with a masculine characteristic (short hair) in order to equalize hiring rates.

In addition, the transparency report shows that the disadvantaged women have much less (1.3 standard deviations) work experience and slightly longer (0.5 SD) hair than their counterparts, while the advantaged women have slightly less (0.6 SD) work experience and much longer (1.6 SD) hair than their counterparts. This suggests that the disadvantaged women are disadvantaged due to their short work experience and that the advantaged women are advantaged due to their hair length. We note that in general, this is not sufficient to establish *causal* claims about why the model behaves this way; for example, the difference in hair length may be due to the model's use of some other feature that is correlated with hair length. In this case, however, the model's weights—which define its causal behavior—and the result of the transparency report do agree. The model, which is a linear regressor, weights the two features almost equally ($w_{hair} = 1.4$, $w_{work} = 1.2$). Since work experience, but not hair length, is a legitimate factor in most hiring decisions, after this deeper investigation into the model we may conclude that this apparently fair model in fact discriminates against some men.

5 Conclusion

We presented FlipTest, a fairness auditing framework that identifies statistical discrimination in machine learning models. FlipTest combines notions of group and individual fairness by constructing a flipset. As we showed in Section 4.4, the flipset sheds light on examples of unfairness in a groupwise fair model, leading to the identification of specific subgroups that are treated unfairly by the model. FlipTest is efficient, leveraging a GAN-based approximation of the optimal transport map to identify likely subgroups from a moderate number ($O(n)$) of queries to the model.

The primary limitations of our approach follow from its use of GANs to construct the flipset. Namely, it inherits the the well-known problems of training and tuning GANs on new data, and without special considerations in the generator architecture and loss, may give inadequate results on datasets with many categorical features. However, it will also inherit the benefits of advances in this area. As future work, extending the framework beyond binary classifiers, which are the most commonly studied case in the fairness literature, and exploring the application of FlipTest to uncovering additional types of discrimination are both promising directions.

Acknowledgment

This material is based upon work supported by the National Science Foundation under Grant No. CNS-1704845.

References

[1] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna M. Wallach. A reductions approach to fair classification. *CoRR*, abs/1803.02453, 2018. URL http://arxiv.org/abs/1803.02453.

[2] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. *ProPublica*, 2016.

[3] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein GAN. *arXiv preprint arXiv:1701.07875*, 2017.

[4] Solon Barocas and Andrew D Selbst. Big data's disparate impact. *California Law Review*, 104:671–732, 2016.

[5] François Chollet et al. Keras. https://keras.io, 2015.

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

[7] City of Chicago. Strategic Subject List. https://data.cityofchicago.org/Public-Safety/Strategic-Subject-List/4aki-r3np, 2017.

[8] Jeffrey Dastin. Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters*, 2018.

[9] Anupam Datta, Shayak Sen, and Yair Zick. Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. In *IEEE Symposium on Security and Privacy*, pages 598–617, 2016.

[10] Eustasio del Barrio, Fabrice Gamboa, Paula Gordaliza, and Jean-Michel Loubes. Obtaining fairness using optimal transport theory. *arXiv preprint arXiv:1806.03195*, 2018.

[11] Cynthia Dwork and Christina Ilvento. Fairness under composition. *CoRR*, abs/1806.06122, 2018. URL http://arxiv.org/abs/1806.06122.

[12] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Innovations in Theoretical Computer Science*, pages 214–226, 2012.

[13] Equivant. Practitioner's guide to COMPAS core. http://www.equivant.com/wp-content/uploads/Practitioners-Guide-to-COMPAS-Core-040419.pdf, 2019.

[14] Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 259–268, 2015.

[15] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2672–2680, 2014.

[16] Gurobi Optimization, LLC. Gurobi optimizer reference manual. https://www.gurobi.com/documentation/8.1/refman.pdf, 2019.

[17] Sara Hajian and Josep Domingo-Ferrer. A methodology for direct and indirect discrimination prevention in data mining. *IEEE transactions on knowledge and data engineering*, 25(7): 1445–1459, 2012.

[18] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, pages 3315–3323, 2016.

[19] Úrsula Hébert-Johnson, Michael Kim, Omer Reingold, and Guy Rothblum. Multicalibration: Calibration for the (computationally-identifiable) masses. In *International Conference on Machine Learning*, pages 1944–1953, 2018.

[20] Matthew Joseph, Michael J. Kearns, Jamie Morgenstern, and Aaron Roth. Fairness in learning: Classic and contextual bandits. *CoRR*, abs/1605.07139, 2016. URL http://arxiv.org/abs/1605.07139.

[21] Faisal Kamiran and Toon Calders. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33(1):1–33, 2012.

[22] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *International Conference on Machine Learning*, pages 2564–2572, 2018.

[23] Michael J. Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. An empirical study of rich subgroup fairness for machine learning. *CoRR*, abs/1808.08166, 2018. URL http://arxiv.org/abs/1808.08166.

[24] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. In *Innovations in Theoretical Computer Science*, pages 43:1–43:23, 2017.

[25] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4066–4076. Curran Associates, Inc., 2017. URL http://papers.nips.cc/paper/6995-counterfactual-fairness.pdf.

[26] Zachary Lipton, Julian McAuley, and Alexandra Chouldechova. Does mitigating ML's impact disparity require treatment disparity? In *Advances in Neural Information Processing Systems*, pages 8125–8135, 2018.

[27] Cong Ma, Kaizheng Wang, Yuejie Chi, and Yuxin Chen. Implicit regularization in nonconvex statistical estimation: Gradient descent converges linearly for phase retrieval and matrix completion. In *International Conference on Machine Learning*, pages 3351–3360, 2018.

[28] Parliament of the United Kingdom. Equality Act 2010. https://www.legislation.gov.uk/ukpga/2010/15/contents, 2010.

[29] Supreme Court of the United States. *Griggs v. Duke Power Co.* 401 U.S. 424, 1971.

[30] Theano Development Team. Theano: A Python framework for fast computation of mathematical expressions. *arXiv preprint arXiv:1605.02688*, 2016.

[31] Florian Tramèr, Vaggelis Atlidakis, Roxana Geambasu, Daniel Hsu, Jean-Pierre Hubaux, Mathias Humbert, Ari Juels, and Huang Lin. Fairtest: Discovering unwarranted associations in data-driven applications. In *IEEE European Symposium on Security and Privacy*, pages 401–416, 2017.

[32] Berk Ustun, Alexander Spangher, and Yang Liu. Actionable recourse in linear classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, pages 10–19, New York, NY, USA, 2019. ACM. ISBN 978-1-4503-6125-5. doi: 10.1145/3287560.3287566. URL http://doi.acm.org/10.1145/3287560.3287566.

[33] Rhema Vaithianathan, Emily Putnam-Hornstein, Nan Jiang, Parma Nand, and Tim Maloney. Developing predictive models to support child maltreatment hotline screening decisions: Allegheny County methodology and implementation. https://www.alleghenycountyanalytics.us/wp-content/uploads/2018/02/DevelopingPredictiveRiskModels-package_011618.pdf, 2017.

[34] Cédric Villani. *Optimal transport: Old and new*, volume 338. Springer Science & Business Media, 2008.

[35] Sandra Wachter, Brent D. Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *CoRR*, abs/1711.00399, 2017. URL http://arxiv.org/abs/1711.00399.

[36] Blake E. Woodworth, Suriya Gunasekar, Mesrob I. Ohannessian, and Nathan Srebro. Learning non-discriminatory predictors. *CoRR*, abs/1702.06081, 2017. URL http://arxiv.org/abs/1702.06081.

[37] Karren D. Yang and Caroline Uhler. Scalable unbalanced optimal transport using generative adversarial networks. *CoRR*, abs/1810.11447, 2018. URL http://arxiv.org/abs/1810.11447.

[38] Samuel Yeom, Anupam Datta, and Matt Fredrikson. Hunting for discriminatory proxies in linear regression models. In *Advances in Neural Information Processing Systems*, pages 4568–4578, 2018.

[39] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P Gummadi. Fairness constraints: Mechanisms for fair classification. In *Artificial Intelligence and Statistics*, pages 962–970, 2017.

[40] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In Sanjoy Dasgupta and David McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, 2013.

[41] Zhe Zhang and Daniel B Neill. Identifying significant predictive bias in classifiers. *arXiv preprint arXiv:1611.08292*, 2016.