The University of Texas at Austin
Department of Computer Science
College of Natural Sciences

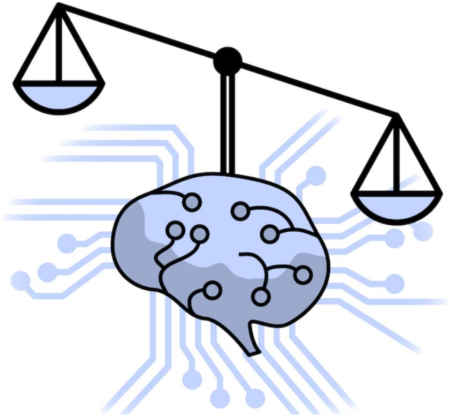
Algorithmic Fairness

SDS 384 Scientific Machine Learning, Spring 2023

Justin Lerma, Chris Lawson, Jeffrey Gordon

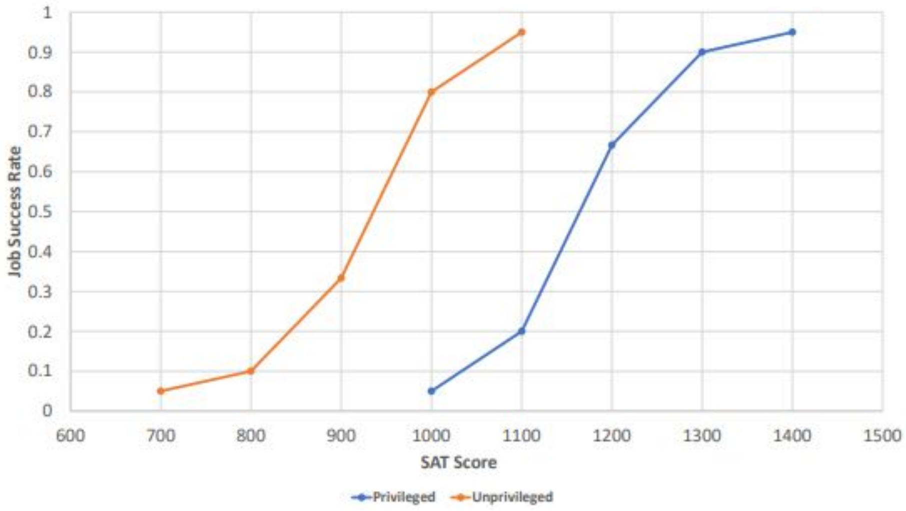
Introduction / Motivation

- Increasing number of decisions being controlled by ML
    - Hiring Managers
    - Influencing Sentencing
    - Suggested Ads
- Misconception when machines always being “objective”



Potential Causes

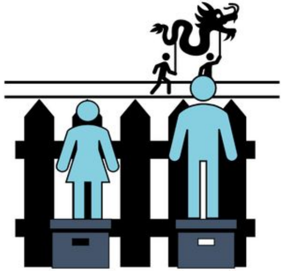
- Bias in existing datasets
- Missing Data
- Disproportionate data among groups
- Caused by proxy through sensitive attributes

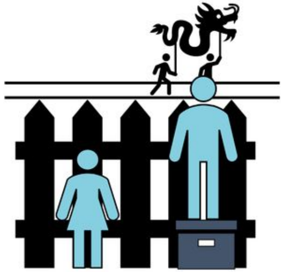


Current Legal Definitions

- **Disparate Impact:**
    - Negatively affecting members of a protected class more than others even if by a seemingly neutral policy

- **Disparate Treatment:**
    - Intentionally treating an individual differently based on his/her membership in a protected class


Disparate Impact


Disparate Treatment

Measures of Algorithmic Bias

Disparate Impact

- Was designed to mathematically represent the legal notion
- Approaches fairness through True Positive Rates
- Tries to show that the TPR across attributes is the same within some error ε

$$\frac{P[\hat{Y} = 1 | S \neq 1]}{P[\hat{Y} = 1 | S = 1]} \geq 1 - \varepsilon$$

Measures of Algorithmic Bias

Demographic Parity

- Similar to Disparate Impact, but a difference is taken rather than a ratio.

- A smaller value indicates better fairness

$$\left|P[\hat{Y} = 1 | S = 1] - P[\hat{Y} = 1 | S \neq 1]\right| \leq \varepsilon$$

Measures of Algorithmic Bias

Equalized Odds

- Similar to Demographic Parity, but now computes it for both TPR and FPR
- A classifier needs to satisfy both equalized odds constraints to be fully accurate.
- Requires the ground truth (Y)

$$\left|P[\hat{Y} = 1|S = 1, Y = \underline{0]} - P[\hat{Y} = 1|S \neq 1, Y = \underline{0]}\right| \leq \varepsilon$$

$$\left|P[\hat{Y} = 1|S = 1, Y = \underline{1]} - P[\hat{Y} = 1|S \neq 1, Y = \underline{1]}\right| \leq \varepsilon$$

Measures of Algorithmic Bias

Equal Opportunity

- Requires only the TPRs to be similar across groups

$$\left|P[\hat{Y} = 1 | S \neq 1, Y = 1] - P[\hat{Y} = 1 | S = 1, Y = 1]\right| \leq \varepsilon$$

- Downside is that it may increase disparity in terms of the other error.

- Issues arise when base rates differ between people.

Measures of Algorithmic Bias

Individual Fairness

- Addresses the lack of individual notions of fairness as all previous measures were for groups

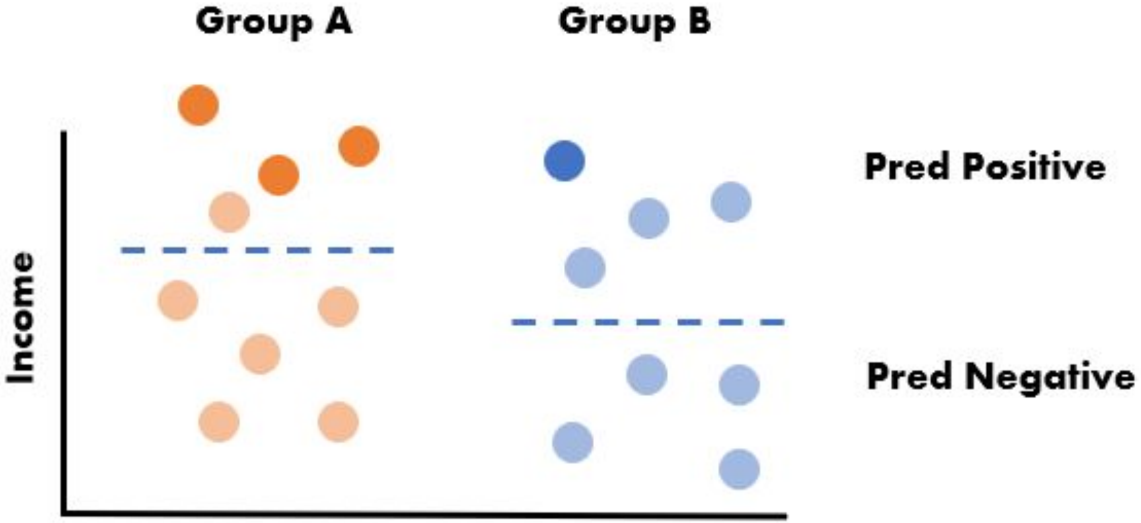
- A distance metric d(i,j) is required to compare similar individuals i and j

- Downside is that defining d(i,j) is incredibly difficult

$$\left|P(\hat{Y}^{(i)} = y|X^{(i)}, S^{(i)}) - P(\hat{Y}^{(j)} = y|X^{(j)}, S^{(j)})\right| \leq \varepsilon;\ if\ d(i,j) \approx 0$$

Visualization of Algorithms

Problem: Deciding who gets a loan



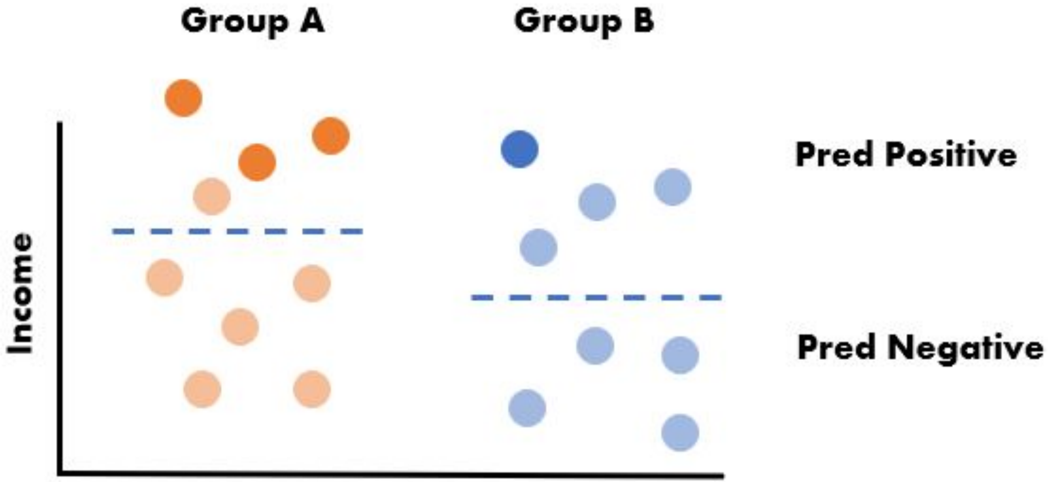
Visualization of Algorithms

Demographic Parity / Disparate Impact

PR for A: 4 / 9 = 40%

PR for B: 4 / 8 = 50%

Is this a good model?



$$\left|P[\hat{Y} = 1|S = 1] - P[\hat{Y} = 1|S \neq 1]\right| \leq \varepsilon$$

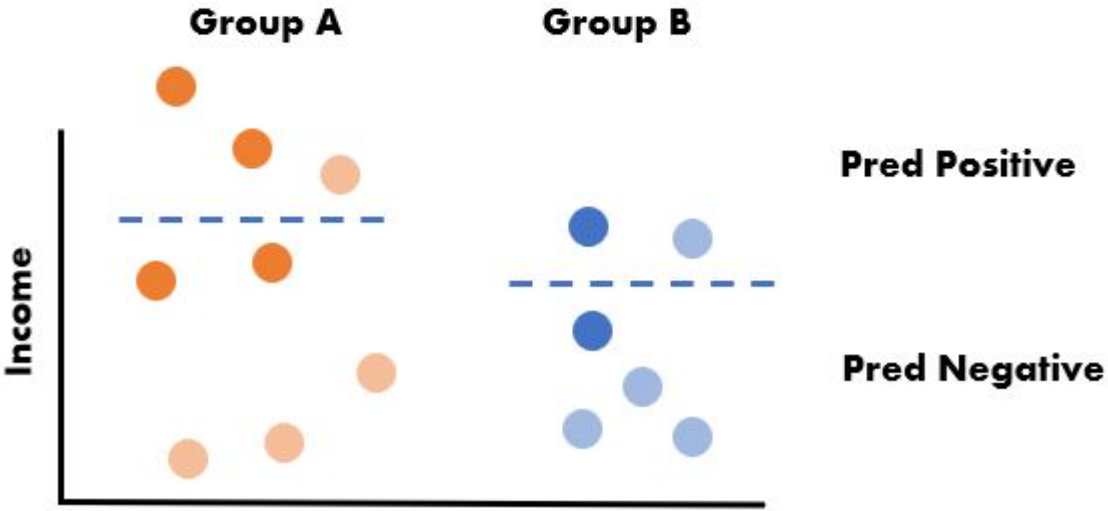
Visualization of Algorithms

Equalized Odds

TPR for A: 2 / 4 = 50%
FPR for A: 1 / 4 = 25%

TPR for B: 1 / 2 = 50%
FPR for B: 1 / 4 = 25%

Is this a good model?



$$\left|P[\hat{Y} = 1|S = 1, Y = 0] - P[\hat{Y} = 1|S \neq 1, Y = 0]\right| \leq \varepsilon$$

$$\left|P[\hat{Y} = 1|S = 1, Y = 1] - P[\hat{Y} = 1|S \neq 1, Y = 1]\right| \leq \varepsilon$$

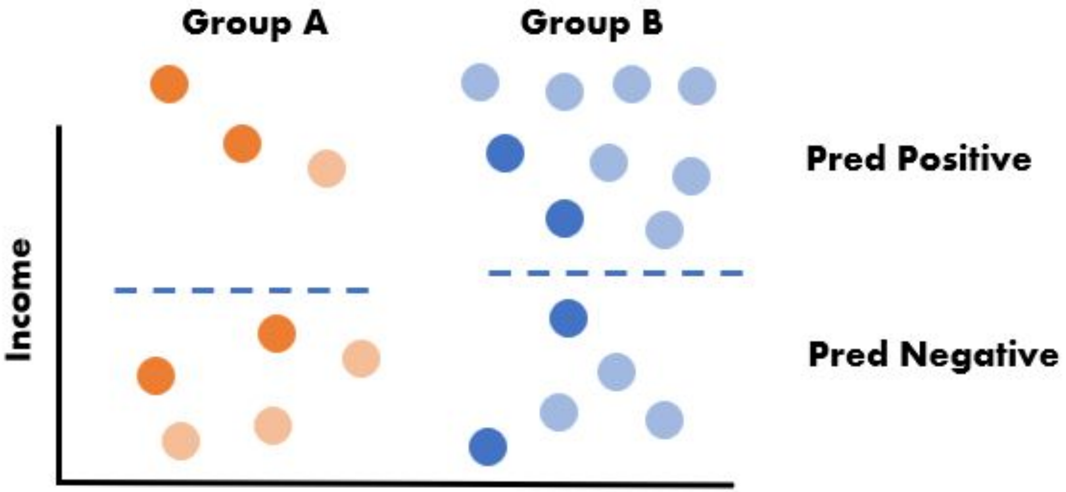
Visualization of Algorithms

Equal Opportunity

TPR for A: 2 / 4 = 50%

TPR for B: 2 / 4 = 50%

Is this a good model?



$$\left|P[\hat{Y} = 1 | S \neq 1, Y = 1] - P[\hat{Y} = 1 | S = 1, Y = 1]\right| \leq \varepsilon$$

What should we choose?

- Take into account the proper:
    - Legal
    - Ethical
    - Social Context

- Consider Fairness - Accuracy Tradeoff

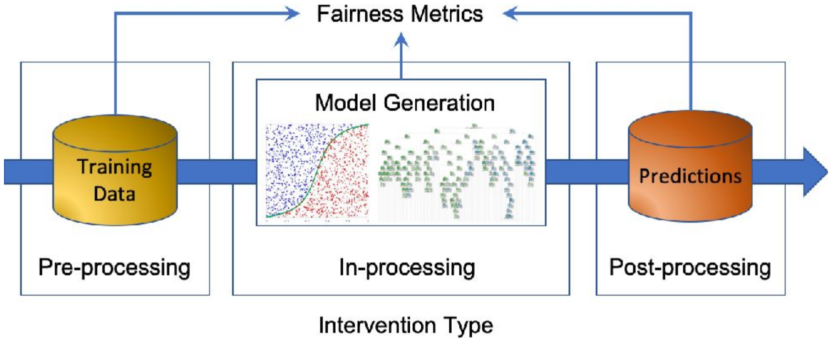
Fairness Enhancing Mechanisms

Pre-Process Mechanisms: Manipulation of data before feeding it into an algorithm.

In-Process Mechanisms: Modifying algorithms to account for fairness.

Post-Process Mechanisms: Post processing

of the output scores of the classifier.



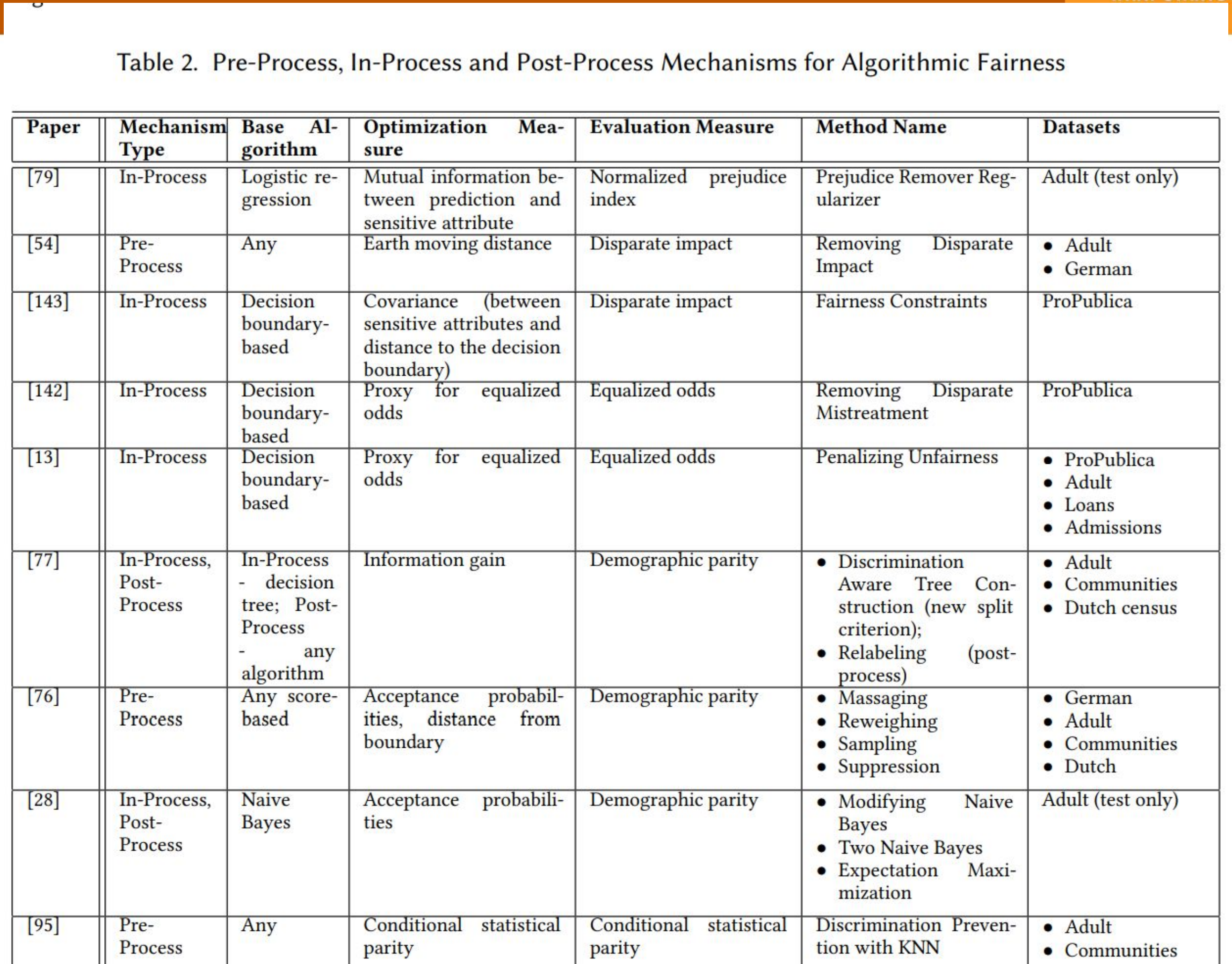


Table 2. Pre-Process, In-Process and Post-Process Mechanisms for Algorithmic Fairness

Paper	Mechanism Type	Base Algorithm	Optimization Measure	Evaluation Measure	Method Name	Datasets
[79]	In-Process	Logistic regression	Mutual information between prediction and sensitive attribute	Normalized prejudice index	Prejudice Remover Regularizer	Adult (test only)
[54]	Pre-Process	Any	Earth moving distance	Disparate impact	Removing Disparate Impact	• Adult • German
[143]	In-Process	Decision boundary-based	Covariance (between sensitive attributes and distance to the decision boundary)	Disparate impact	Fairness Constraints	ProPublica
[142]	In-Process	Decision boundary-based	Proxy for equalized odds	Equalized odds	Removing Disparate Mistreatment	ProPublica
[13]	In-Process	Decision boundary-based	Proxy for equalized odds	Equalized odds	Penalizing Unfairness	• ProPublica • Adult • Loans • Admissions
[77]	In-Process, Post-Process	In-Process - decision tree; Post-Process - any algorithm	Information gain	Demographic parity	• Discrimination Aware Tree Construction (new split criterion); • Relabeling (post-process)	• Adult • Communities • Dutch census
[76]	Pre-Process	Any score-based	Acceptance probabilities, distance from boundary	Demographic parity	• Massaging • Reweighing • Sampling • Suppression	• German • Adult • Communities • Dutch
[28]	In-Process, Post-Process	Naive Bayes	Acceptance probabilities	Demographic parity	• Modifying Naive Bayes • Two Naive Bayes • Expectation Maximization	Adult (test only)
[95]	Pre-Process	Any	Conditional statistical parity	Conditional statistical parity	Discrimination Prevention with KNN	• Adult • Communities

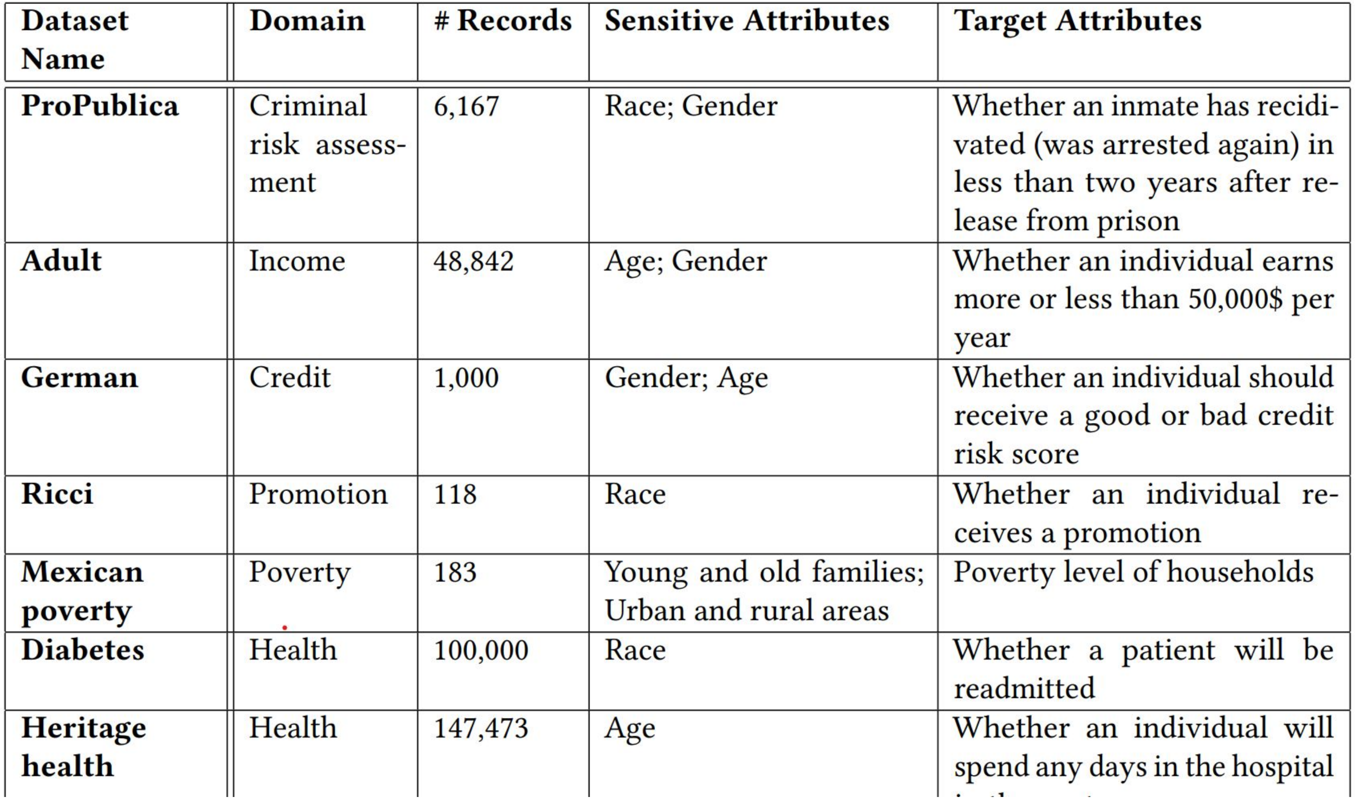
Which mechanisms to use

Depends on the situation:

- Pre-processing: nonspecific
- Post-processing: nonspecific
- Inprocessing: coupled with the algorithm

Performances of each mechanism method vary across datasets with no conclusive dominate method, however more extensive experiments are needed

Table 3. Common Benchmark Datasets for Algorithmic Fairness



Dataset Name	Domain	# Records	Sensitive Attributes	Target Attributes
ProPublica	Criminal risk assessment	6,167	Race; Gender	Whether an inmate has recidivated (was arrested again) in less than two years after release from prison
Adult	Income	48,842	Age; Gender	Whether an individual earns more or less than 50,000$ per year
German	Credit	1,000	Gender; Age	Whether an individual should receive a good or bad credit risk score
Ricci	Promotion	118	Race	Whether an individual receives a promotion
Mexican poverty	Poverty	183	Young and old families; Urban and rural areas	Poverty level of households
Diabetes	Health	100,000	Race	Whether a patient will be readmitted
Heritage health	Health	147,473	Age	Whether an individual will spend any days in the hospital

Emerging Research on Algorithmic Fairness

Fairness in Sequential Learning

- Sequential Learning deals with models that receive data in an online fashion
- Feedback loops allow current decision to have influence over future ones
- Due to this, must consider fairness at each time step & balance exploitation vs. exploration to make decisions

Markov Decision Process

- RL Sequential Model that defines fairness as one action will not be preferred over another if its discounted long term reward is lower

Time - Dependent Models

- Post Process mechanism that says that two people who arrive at the same time and have similar features should have similar labels

Challenges

- Feedback loops can amplify bias if not handled from the beginning
- Exploration might be considered unethical and favoring one group over another

Fairness in Adversarial Learning

- GANs generate fake data that simulates real data
- A network G sends fake data to a network A, which sends a guess of if the data is real or fake to G. G uses response to update network
- Fairness issues in applications such as filters, which could favor a skin color for making someone more attractive

Minimax

- To apply fairness, model is structured as a minimax problem, trying to maximize the predictors accuracy while minimizing the adversaries capability to predict the sensitive feature

Feedback, Fair Representations, Synthetic Data

- Use a feedback loop to see if a classifier is fair or not, update if needed
- Use GAN to create fair features that a model will have a harder time distinguishing groups
- Use GAN to generate unbiased data form the original data, and train classifier

Fair Word Embedding

- Word embeddings transform words into vector representations
- Words that occur in the same context will be closer in vector space
- Bias is created when gender defining words are in context with gender neutral words (like “he” and “CEO”, or “male” and “programmer”)

Hard-Debiasing

- Post process mechanism to remove gender bias
- Identifies the gender dimension, defines neutral words, then negates the projection of the neutral words w.r.t. The gender direction, so the neutral word bias goes to 0

Soft-Debiasing

- Reduces bias while maintaining some similarity to original vector
- More control over accuracy-bias tradeoff

Gender-Neutral Vectors

- In process mechanism that encourages pairs of gender indication words to have more distance, and encourages gender neutral words to be orthogonal to gender direction

- Overall, more extensive research is needed in order to actually remove the bias and not just hide them

Fair Visual Description

- Bias is apparent in computer vision problems (underrepresentation of dark skinned faces, or labeling images incorrectly due to naive appearance similarities)
- Hard to mitigate because you need to add fairness to the NLP model and the CV model
- Datasets are inherently unbalanced as well
- Some methods to add fairness is to solve a constrained optimization problem using Lagrange relaxation, and directly using a person's appearance information in the image. This makes the model more cautious when there's no gender info in the image.

Fairness in Recommender Systems

- Apparent bias, such as recommending higher paying jobs to men and not women
- Still need to take personalization into account when reducing bias
- **C-Fairness:** Fairness considered recommendations are independent of group
- **P-Fairness:** Items from different providers are recommended equally
- Some studies used a constraint based integer program to have both C and P Fairness
- Post Process Mechanism that modifies entries in the recommendation matrix so it's hard to predict sensitive attributes (E-Fairness is the measure of this)
- Notion of Recommendation Independence requires recommendation results are independent from sensitive attribute

Fairness in Causal Learning

- Causal learning relies on data with a cause and effect structure
- Enhances fairness if you understand the cause and effect, since you can decide which bias should be allowed or not
- Gives you insight in the source of bias, which increases model transparency
- Counterfactual fairness: Measures the extent to which you can build the same prediction model from a privileged group vs. an unprivileged group
- High score means predictions aren't affected by sensitive attributes
- In practice, hard to obtain the correct causal model

Discussion & Conclusion

- Presented overview of algorithmic fairness
- Described the causes, measures, trade offs, and solutions
- Described ongoing research in the field
- There are still several open challenges
  - Dealing with dataset bias
  - Fairness-Accuracy tradeoff
  - Transparency on how model addresses fairness
- AI is becoming integrated with our daily lives, so dealing with the question of fairness is inevitable