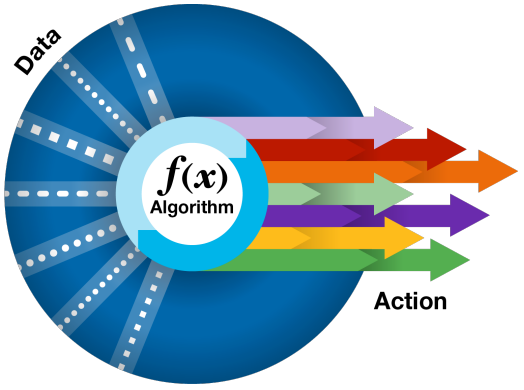
# Classification without Discrimination

**Abhijnan Chakraborty**
Indian Institute of Technology Kharagpur
https://cse.iitkgp.ac.in/~abhijnan/

# Algorithmic Decision Making in Practice




AI algorithms are moderating

- Public places
  - People to people communication, Political campaigns: Social media
  - Banking: Loan approval
  - Judiciary: Bail decisions
  - Law enforcement: Police deployments
  - Healthcare: Determining high-risk patients

# Algorithmic Decision Making in Practice


AI algorithms are moderating

- Public places
- Marketplaces
  - Ecommerce: Amazon, Alibaba
  - Donation: DonorsChoose
  - Crowdfunding: Kickstarter
  - Matchmaking: Tinder
  - Guilty pleasure: OnlyFans

# Algorithmic Decision Making in Practice


AI algorithms are moderating

- Public places
- Marketplaces
- Workplaces
  - Ride hailing: Uber, Ola, Lyft
  - Food delivery: Zomato, Lieferando
  - Micro tasks: AMT, Fiverr
  - Other gig economy platforms

# Algorithmic Decision Making in Practice


AI algorithms are moderating

- Public places
- Marketplaces
- Workplaces
- Private spaces
  - Voice assistants: Alexa
  - Smart devices: Fitbit

# Algorithmic Decision Making in Practice

AI algorithms are moderating

- Public places
- Marketplaces
- Workplaces
- Private spaces

**Since AI is touching all aspects of our lives, shouldn't it embed the societal values/norms we have?**

# Algorithmic Decision Making in Practice

AI algorithms are moderating

- Public places
- Marketplaces
- Workplaces
- Private spaces

**Since AI is touching all aspects of our lives, shouldn't it embed the societal values/norms we have?**

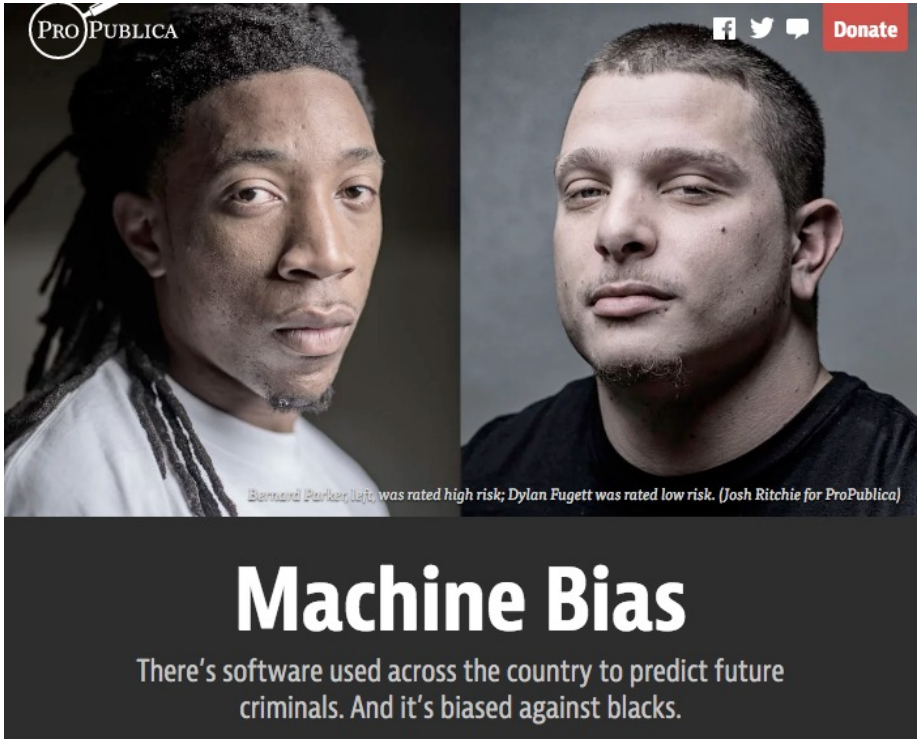
How are the existing deployments doing?

# Criminal Justice

## Recidivism Risk Assessment:

Predicting risk of future crime for bail or sentencing decisions

- COMPAS tool has higher false positive rate for African-Americans
- It has higher false negative rate for Caucasians
- Discrimination against African-Americans



# Healthcare

Identifying High Risk Patients:
Predicting who will need additional care

**Healthcare algorithm used across America has dramatic racial biases**

The Guardian

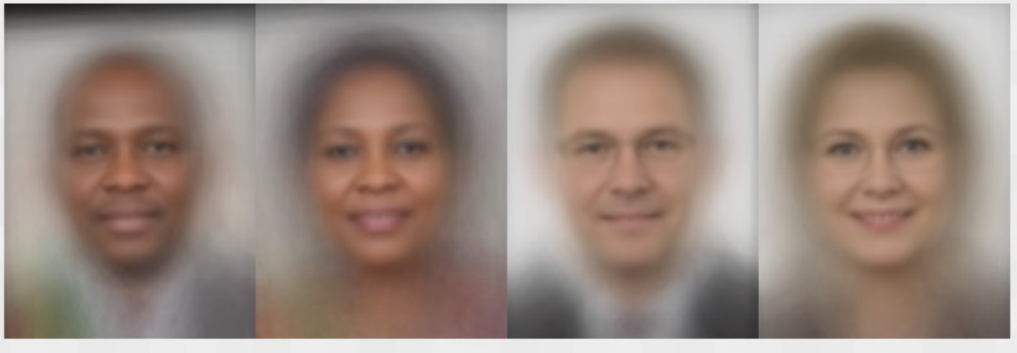
- At a given risk score, Black patients were considerably sicker than White patients, as evidenced by chronic illness conditions
- The algorithm takes healthcare spend as proxy for illness
- Replicate historical neglect towards poorer groups

# Face Detection

## Gender Prediction:

Predicting gender of a person from their face image

| Gender Classifier | Darker Male | Darker Female | Lighter Male | Lighter Female | Largest Gap |
|---|---|---|---|---|---|
| Microsoft | 94.0% | 79.2% | 100% | 98.3% | 20.8% |
| FACE++ | 99.3% | 65.5% | 99.2% | 94.0% | 33.8% |
| IBM | 88.0% | 65.3% | 99.7% | 92.9% | 34.4% |



# Employment

Job Candidate Screening:
Predicting who will be a successful hire from resumes

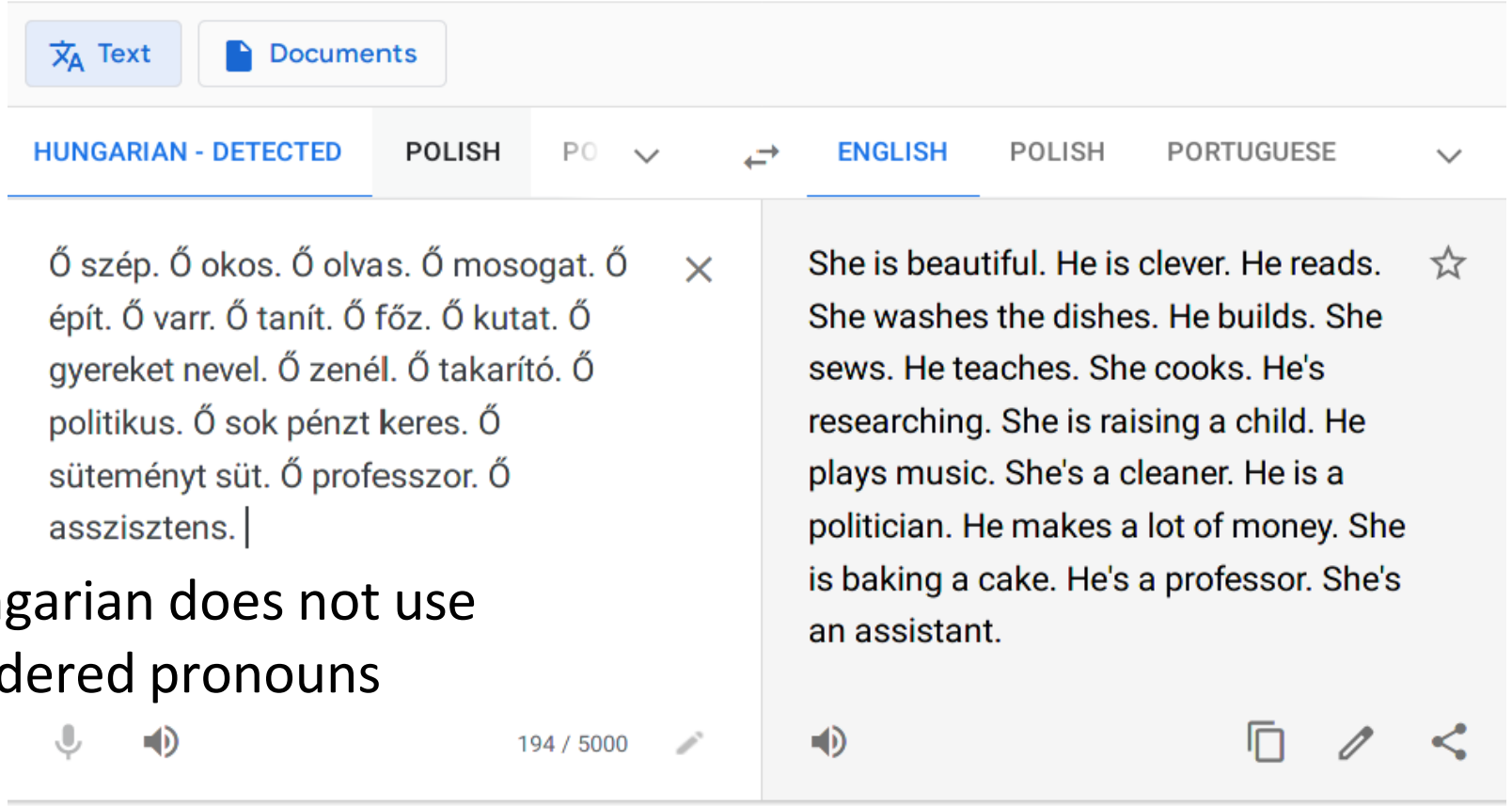
**Amazon scraps secret AI recruiting tool that showed bias against women**


- Replicate gender bias in past decisions
- Even after gender was dropped from the data, it picked other attributes – whether someone went to women college!

# Language Translation

## Hungarian -> English Translation:

Assuming a gender when there is none




Hungarian does not use gendered pronouns

Source: https://www.reddit.com/r/europe/comments/m9uphb/hungarian_has_no_gendered_pronouns_so_google

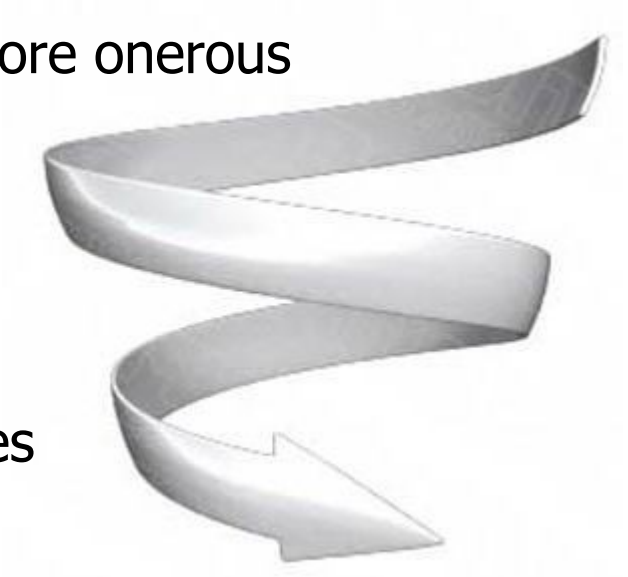
# Self-perpetuating Algorithmic Biases

Credit scoring algorithm suggests Joe has high risk of defaulting

Hence, Joe needs to take a loan at a higher interest rate

Hence, Joe has to make payments that are more onerous

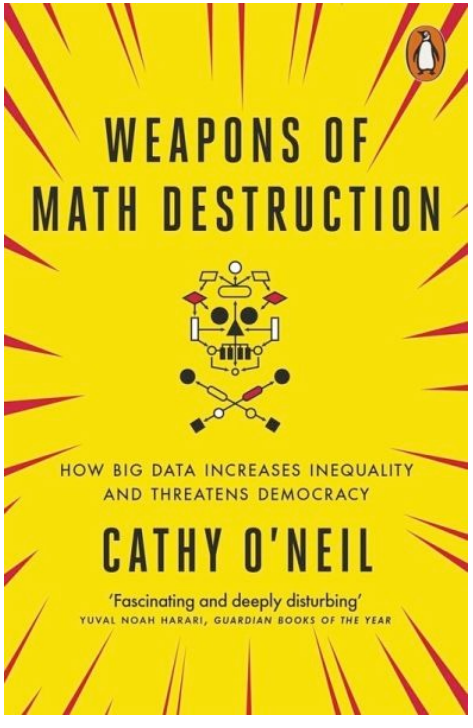
Hence, Joe's risk of defaulting has increased



Same happens with stop-and-frisk of minorities
Further increasing incarceration rates

Sara Hajian, Francesco Bonchi, Carlos Castillo. KDD 2016

# Where Did We Go Wrong?




- Data at best reflects the current state of the world
  - Acts as a social mirror
- Proxies
  - Protected attributes redundantly encoded in observables
- Correctness
  - Noise in training labels
- Incomplete/Sample size disparity
  - More data from one group

M. Hardt (2014): "How big data is unfair". Medium.

# The Achilles Heels of Traditional ML

Even assuming no training data biases, ML decisions

1. Often optimize for a single decision outcome goal, ignoring
   - Fairness: Equal prediction accuracy for all salient social groups
   - Worst-cases: Lower bound worst-case prediction accuracy
   - Norms: Should use or not use data in a specific manner

# The Achilles Heels of Traditional ML

Even assuming no training data biases, ML decisions

1. Often optimize for a single decision outcome goal, ignoring
   - Fairness: Equal prediction accuracy for all salient social groups
   - Worst-cases: Lower bound worst-case prediction accuracy
   - Norms: Should use or not use data in a specific manner

2. Optimal for a static NOT an evolving society, because
   - Training data becomes unrepresentative
   - Feedback loops are not accounted for in the first place
   - Decision outcome goals do not change over time

# Focusing on Discrimination

- Discrimination is a specific type of unfairness
- Well-studied in social sciences
  - Political science
  - Moral philosophy
  - Economics
  - Law
    - Majority of countries have anti-discrimination laws
    - Discrimination recognized in several international human rights laws
- We need to look at it from a computational perspective

# Defining Discrimination

- An approximate normative / moralized definition:

**wrongfully** impose a **relative disadvantage** on persons based on their membership in some **salient social group**

- Challenge: How to operationalize the definition?
  - How to make it clearly distinguishable, measurable, and understandable in terms of empirical observations

# Need to Operationalize Two Fuzzy Notions

1. What constitutes a salient social group?

2. What constitutes a wrongful relative disadvantage?

# Need to Operationalize Two Fuzzy Notions

1. **What constitutes a salient social group?**

   **Depends on existing legislations**

2. What constitutes a wrongful relative disadvantage?

# Regulated Domains in the US

- Credit (Equal Credit Opportunity Act)
- Education (Civil Rights Act of 1964; Education Amendments of 1972)
- Employment (Civil Rights Act of 1964)
- Housing (Fair Housing Act)
- 'Public Accommodation' (Civil Rights Act of 1964)

Solon Barocas, Moritz Hardt. NIPS 2017

# Regulated Domains in the US

- Credit (Equal Credit Opportunity Act)
- Education (Civil Rights Act of 1964; Education Amendments of 1972)
- Employment (Civil Rights Act of 1964)
- Housing (Fair Housing Act)
- 'Public Accommodation' (Civil Rights Act of 1964)

Extends to marketing and advertising; not limited to final decision

# Legally Recognized ‘Protected Classes’

**Race** (Civil Rights Act of 1964); **Color** (Civil Rights Act of 1964); **Sex** (Equal Pay Act of 1963; Civil Rights Act of 1964); **Religion** (Civil Rights Act of 1964); **National origin** (Civil Rights Act of 1964); **Citizenship** (Immigration Reform and Control Act); **Age** (Age Discrimination in Employment Act of 1967); **Pregnancy** (Pregnancy Discrimination Act); **Familial status** (Civil Rights Act of 1968); **Disability status** (Rehabilitation Act of 1973; Americans with Disabilities Act of 1990); **Veteran status** (Vietnam Era Veterans' Readjustment Assistance Act of 1974; Uniformed Services Employment and Reemployment Rights Act); **Genetic information** (Genetic Information Nondiscrimination Act)

Solon Barocas, Moritz Hardt. NIPS 2017

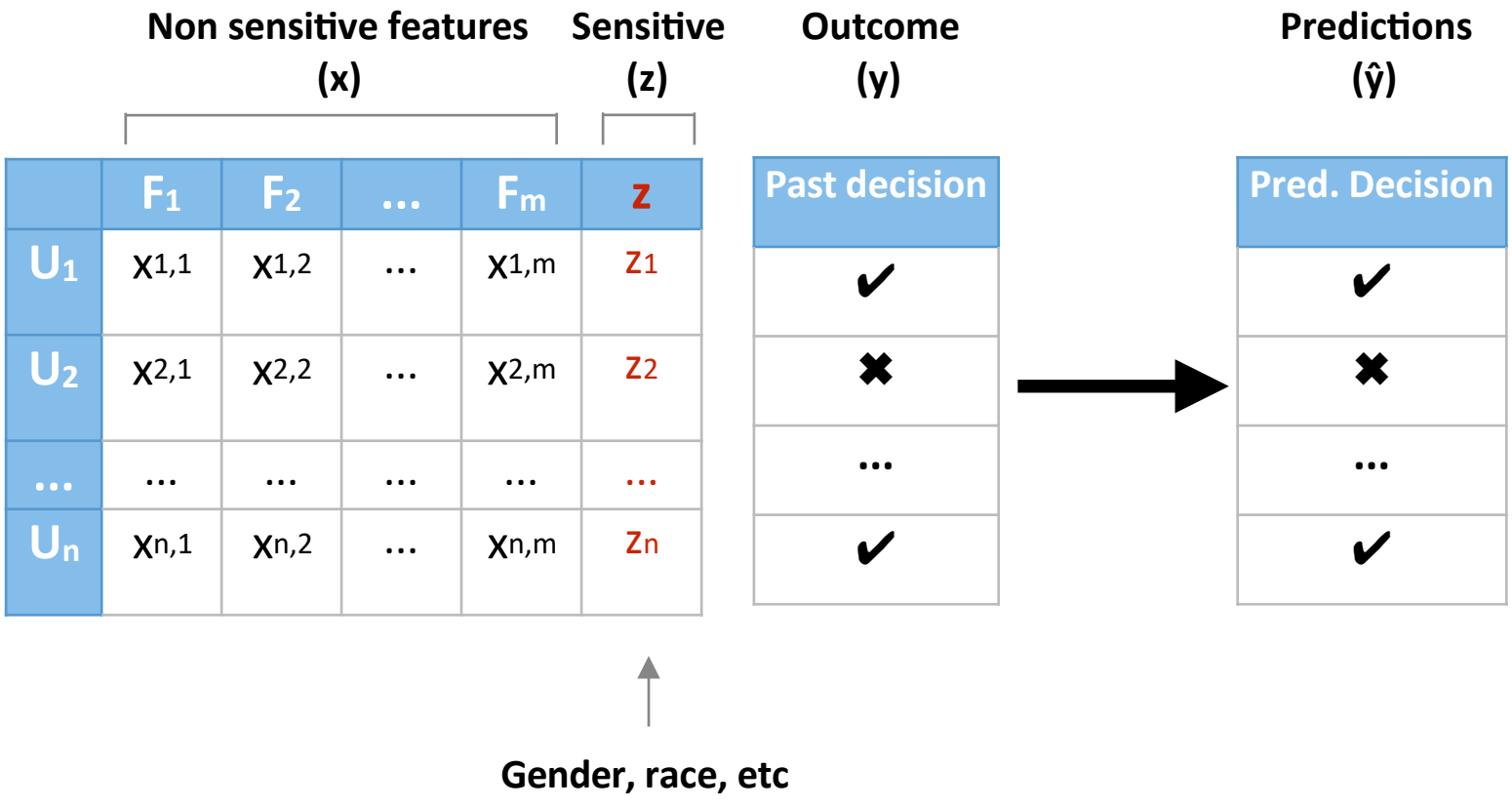
# Need to Operationalize Two Fuzzy Notions

1. What constitutes a salient social group?

   Depends on existing legislations

2. **What constitutes a wrongful relative disadvantage?**

# Toy Example: University Admission



# Relative Disadvantage Measure 1: Disparate Treatment

- Ideal: Achieve parity (or equality) in treatment
- Decision should not change with change in sensitive feature

| | z | x | | $\hat{y}$ |
|---|---|---|---|---|
| | | **School grade** | **SAT score** | **Admit** |
| Bob |  | 90 / 100 | 700 / 800 | |
| Alice |  | 90 / 100 | 700 / 800 | |

# Relative Disadvantage Measure 1: Disparate Treatment

- Ideal: Achieve parity (or equality) in treatment
- Decision should not change with change in sensitive feature

| z | x | | ŷ |
|---|---|---|---|
| | **School grade** | **SAT score** | **Admit** |
| Bob  | 90 / 100 | 700 / 800 | ✔ |
| Alice  | 90 / 100 | 700 / 800 | ✖ |

# Relative Disadvantage Measure 1: Disparate Treatment

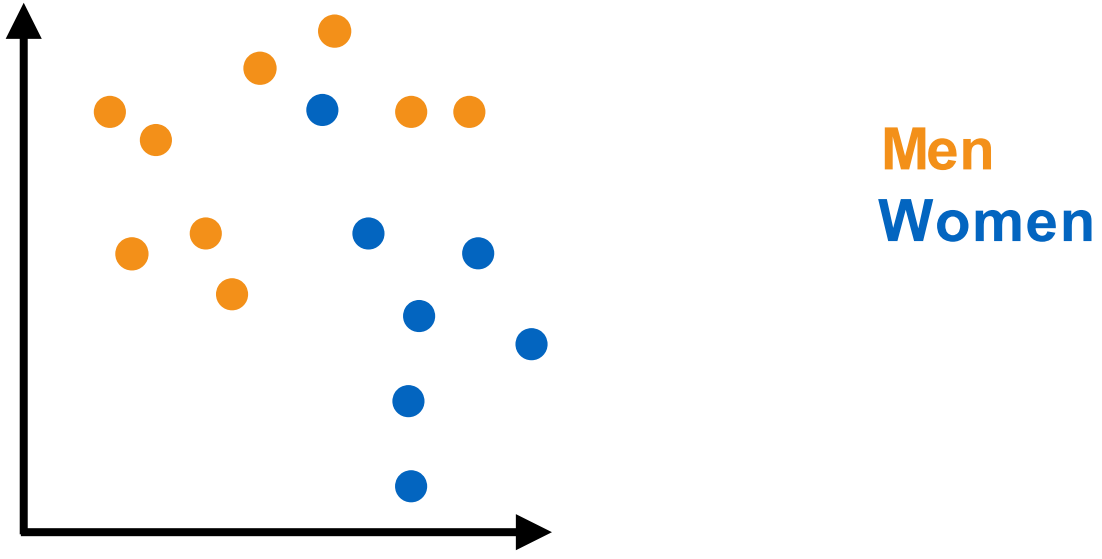
- Ideal: Achieve parity (or equality) in treatment
- Decision should not change with change in sensitive feature

| | z | x | | $\hat{y}$ |
|---|---|---|---|---|
| | | **School grade** | **SAT score** | **Admit** |
| Bob |  | 90 / 100 | 700 / 800 | ✔ |
| Alice |  | 90 / 100 | 700 / 800 | ✖ |

Measure the difference in outcomes for users, when their sensitive features are changed

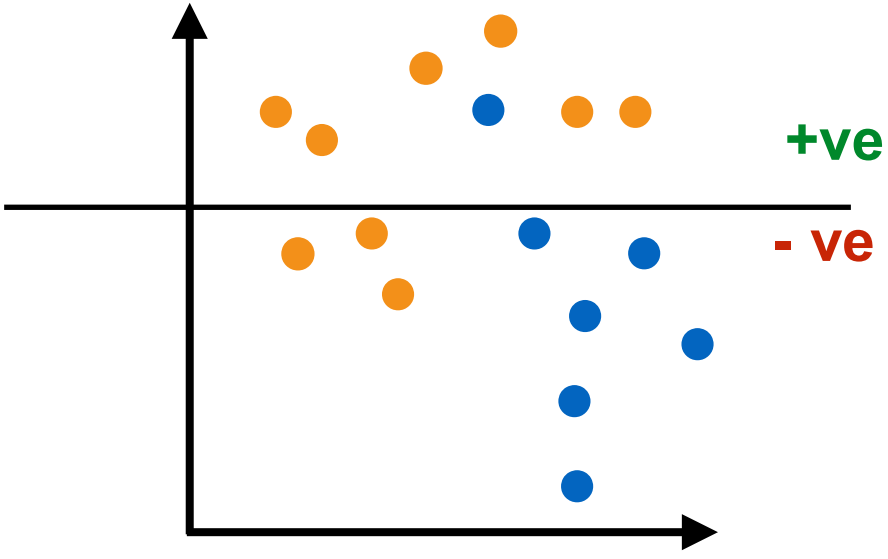
# Relative Disadvantage Measure 2: Disparate Impact

- Ideal: Achieve parity (or equality) in impact
- Positive outcome rates should be same for all groups



# Relative Disadvantage Measure 2: Disparate Impact

- Ideal: Achieve parity (or equality) in impact
- Positive outcome rates should be same for all groups



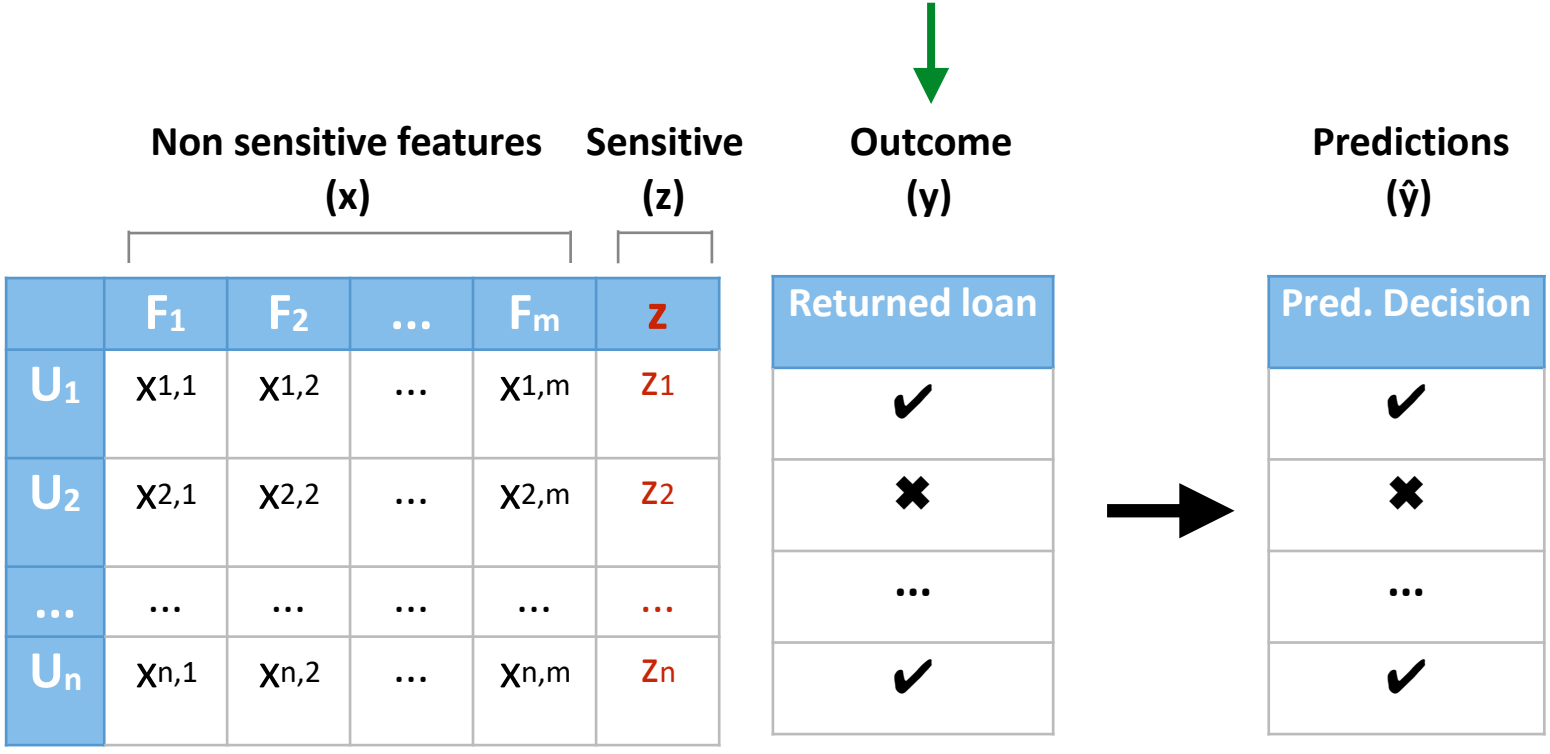

**Men:** 0.67
**Women:** 0.14

# Detecting Indirect Discrimination

- Doctrine of disparate impact
  - A US law applied in employment & housing practices
- Proportionality tests over decision outcomes
  - e.g., in 70's and 80's, some US courts applied the 80% rule for employment practices
    - If 50% (P1%) of male applicants get selected at least 40% (P2%) of female applicants must be selected
  - UK uses P1 – P2; EU uses (1-P1) / (1-P2)
  - Fair proportion thresholds may vary across different domains

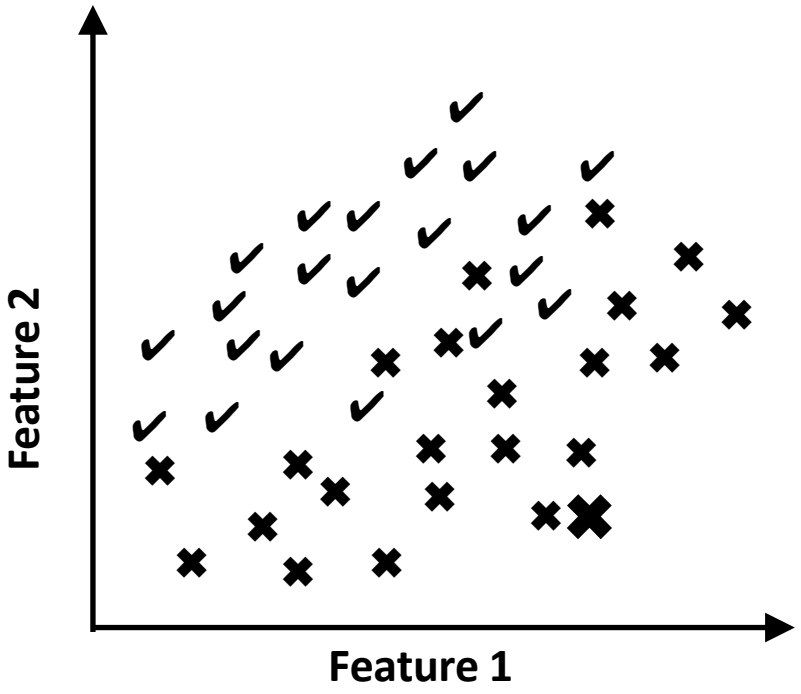
# Example: Credit Risk Assessment

**Ground truth**
**(who returned the loan in the past)**

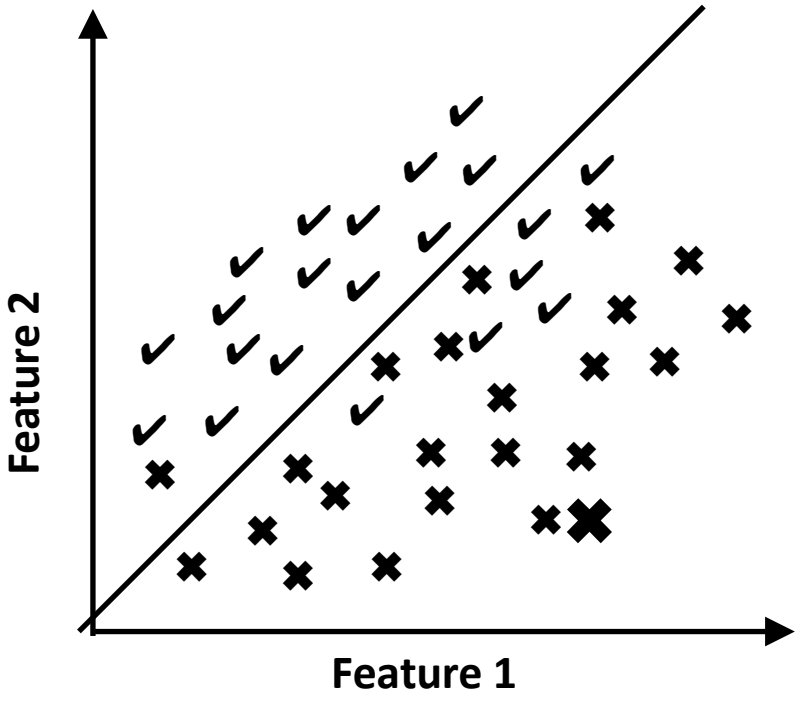


| | Non sensitive features (x) | | | | Sensitive (z) | Outcome (y) | Predictions (ŷ) |
|---|---|---|---|---|---|---|---|
| | $F_1$ | $F_2$ | ... | $F_m$ | z | Returned loan | Pred. Decision |
| $U_1$ | $x_{1,1}$ | $x_{1,2}$ | ... | $x_{1,m}$ | $z_1$ | ✔ | ✔ |
| $U_2$ | $x_{2,1}$ | $x_{2,2}$ | ... | $x_{2,m}$ | $z_2$ | ✖ | ✖ |
| ... | ... | ... | ... | ... | ... | ... | ... |
| $U_n$ | $x_{n,1}$ | $x_{n,2}$ | ... | $x_{n,m}$ | $z_n$ | ✔ | ✔ |

# A Fictitious Dataset: Predict Who'll Return Loan



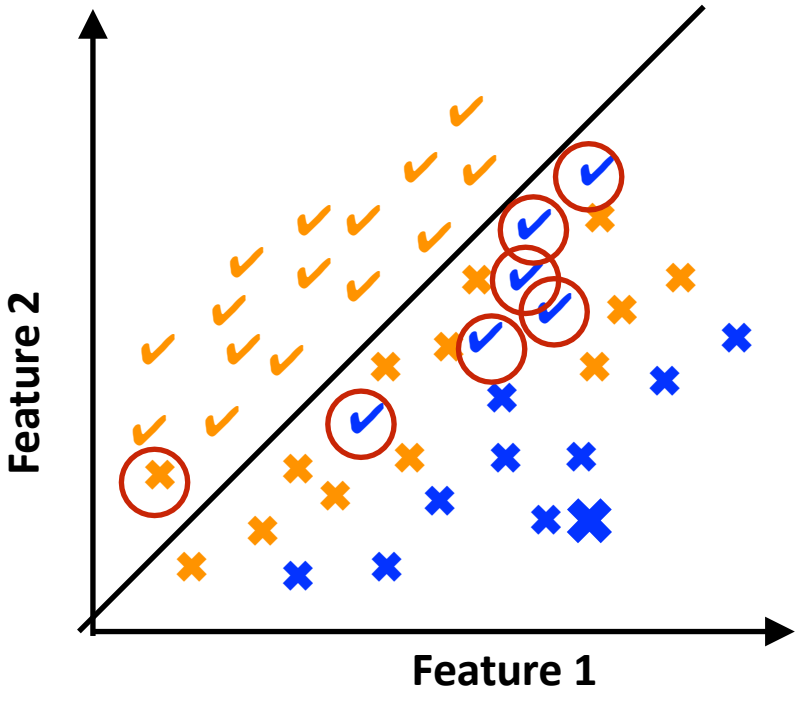
# Learning the Optimal Boundary




**Optimal Loss**

$$\min \sum_{i=1}^{N} L(\mathbf{x}_i, y_i, \mathbf{w})$$

# Relative Disadvantage Measure 3




$$\min \sum_{i=1}^{N} L(\mathbf{x}_i, y_i, \mathbf{w})$$

**Men:** Few errors
**Women:** Many errors
(unfair loan denial)

**Disparate mistreatment:** Different error rates

# Other Discrimination Measures?

How many unfairness measures can one define?

- How many ways can disadvantage manifest?

| | | Predicted Label | | |
|---|---|---|---|---|
| | | $\hat{y} = 1$ | $\hat{y} = -1$ | |
| True Label | $y = 1$ | True positive | False negative | $P(\hat{y} \neq y \mid y = 1)$ False Negative Rate |
| | $y = -1$ | False positive | True negative | $P(\hat{y} \neq y \mid y = -1)$ False Positive Rate |
| | | $P(\hat{y} \neq y \mid \hat{y} = 1)$ False Discovery Rate | $P(\hat{y} \neq y \mid \hat{y} = -1)$ False Omission Rate | $P(\hat{y} \neq y)$ Overall Misclass. Rate |

# Summary: 3 Measures of Discrimination

Disparate treatment: Targets direct discrimination

Requires: $P(\hat{y} \mid \mathbf{x}, z) = P(\hat{y} \mid \mathbf{x})$

Disparate impact: Targets indirect discrimination, when historical labels are biased

Requires: $P(\hat{y} = 1 \mid ♀) = P(\hat{y} = 1 \mid ♂)$

Disparate mistreatment: Targets indirect discrimination, when ground truth available

Requires: $P(y \neq \hat{y} \mid ♂) = P(y \neq \hat{y} \mid ♀)$

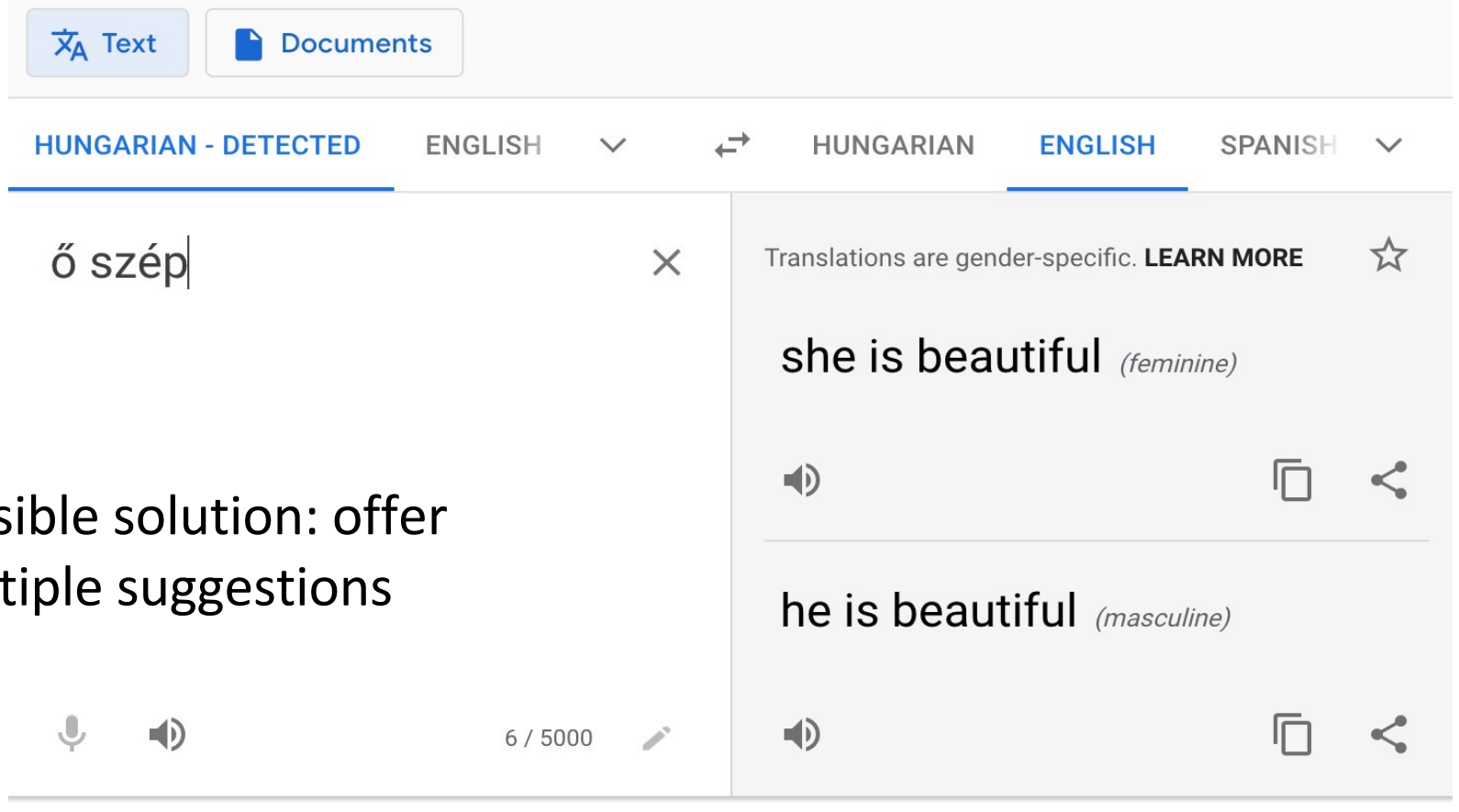
Also for other misclassification rates

# How to Remove (Reduce) Discrimination?

# Sometimes An Easy Fix!

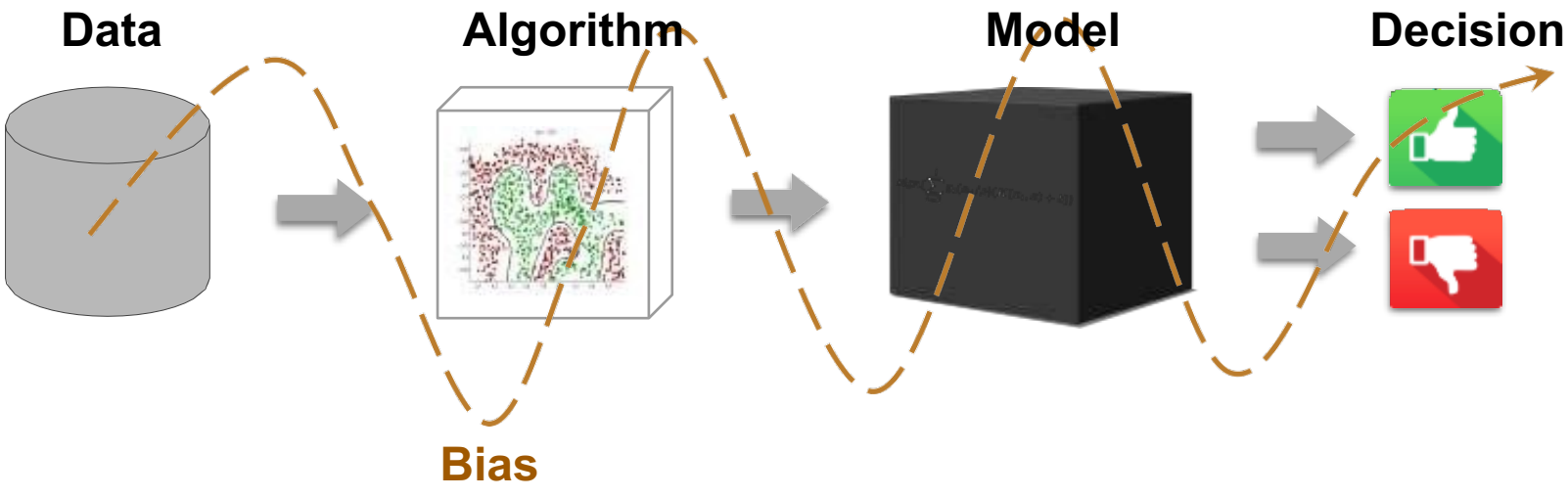
## Hungarian -> English Translation:

~~Assuming~~ a gender when there is none

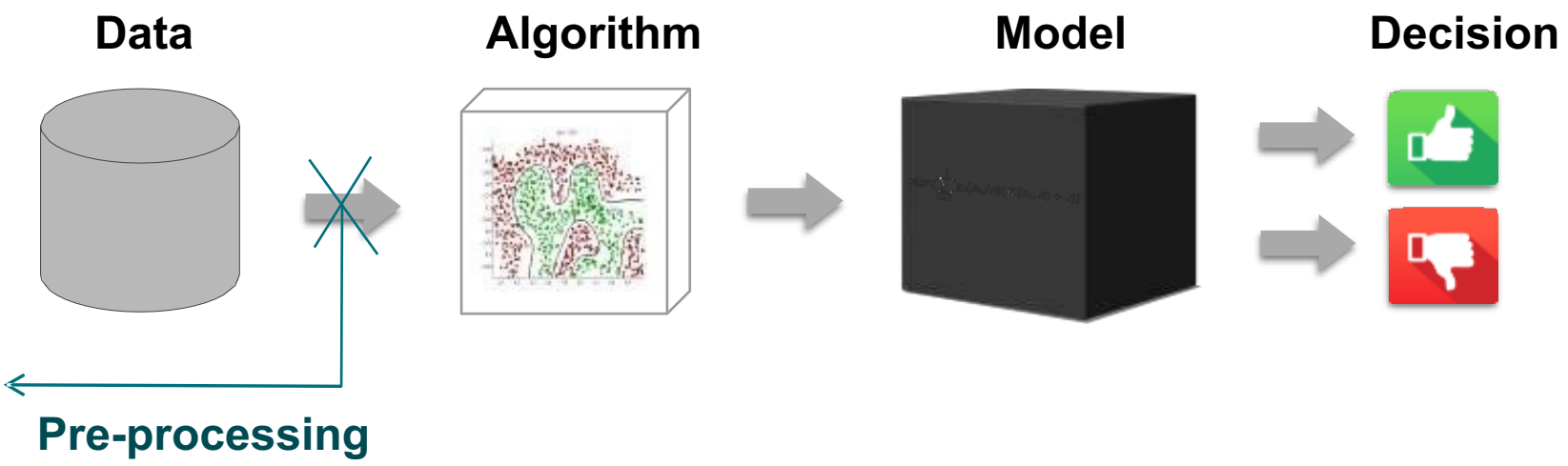



Possible solution: offer multiple suggestions

# Towards Non-discriminatory Machine Decisions



# Towards Non-discriminatory Machine Decisions




[Feldman et al, KDD 2015; Hajian and Domingo-Ferrer, TKDE 2013; Zemel et al., ICML 2013]

# Datasheets for Datasets

Idea: A list of questions to answer when releasing a dataset. Who created it? Why? What is in it? How was it labeled?

**A Database for Studying Face Recognition in Unconstrained Environments** **Labeled Faces in the Wild**

## Motivation

**For what purpose was the dataset created?** Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

Labeled Faces in the Wild was created to provide images that can be used to study face recognition in the unconstrained setting where image characteristics (such as pose, illumination, resolution, focus), subject demographic makeup (such as age, gender, race) or appearance (such as hairstyle, makeup, clothing) cannot be controlled. The dataset was created for the specific task of pair matching: given a pair of images each containing a face, determine whether or not the images are of the same person.[1]

**Who created this dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?**

The initial version of the dataset was created by Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller, most of whom were researchers at the University of Massachusetts Amherst at the time of the dataset's release in 2007.

**Who funded the creation of the dataset?** If there is an associated grant, please provide the name of the grantor and the grant name and number.

The construction of the LFW database was supported by a United States National Science Foundation CAREER Award.

The dataset does not contain all possible instances. There are no known relationships between instances except for the fact that they are all individuals who appeared in news sources on line, and some individuals appear in multiple pairs.

**What data does each instance consist of?** "Raw" data (e.g., unprocessed text or images)or features? In either case, please provide a description.

Each instance contains a pair of images that are 250 by 250 pixels in JPEG 2.0 format.

**Is there a label or target associated with each instance?** If so, please provide a description.

Each image is accompanied by a label indicating the name of the person in the image.

**Is any information missing from individual instances?** If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

Everything is included in the dataset.

**Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)?** If so, please describe how these relationships are made explicit.

There are no known relationships between instances except for

[Gebru et al, "Datasheets for Datasets", FAccT 2018]

# Do not Remove the Sensitive Variables

- Naïve “color unawareness” makes things worse
- It may **decrease accuracy** and **increase unfairness**
- It may also **conceal** that discrimination is happening
- **Data minimization** means you should only collect data for legitimate purposes, detecting discrimination is one of them

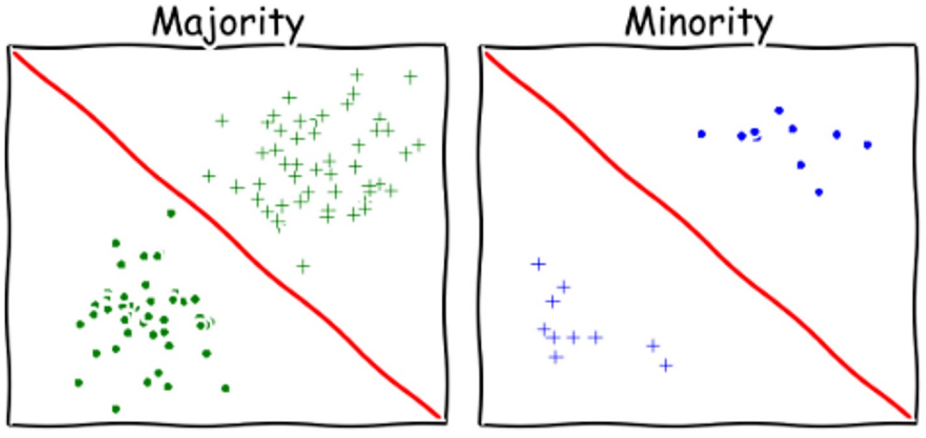
Gemma Galdon Clavell, Mariano Martín Zamorano, Carlos Castillo, Oliver Smith and Aleksandar Matic (2020). Auditing Algorithms: On Lessons Learned and the Risks of Data Minimization. *Proc. AIES*.

# Analyze Subgroups

- Compute per-group accuracy measures
- Compute per-group error measures (e.g., FPR, FNR)
- Focus first on disparate impacts on disadvantaged groups

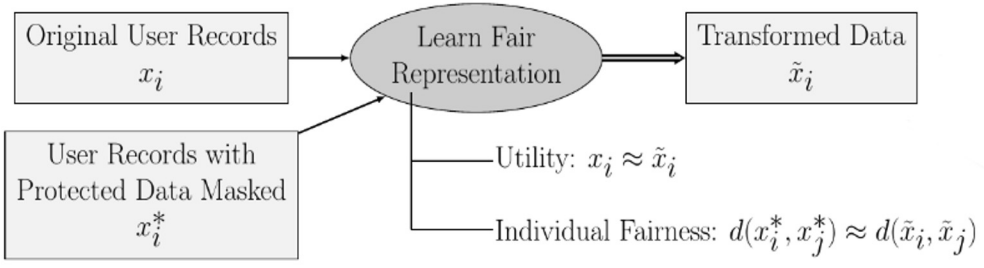
# Analyze Subgroups

- Statistical patterns that apply for the majority may be invalid for the minority
- In some cases, we may need to model them separately or remark our model is not applicable to a class of people

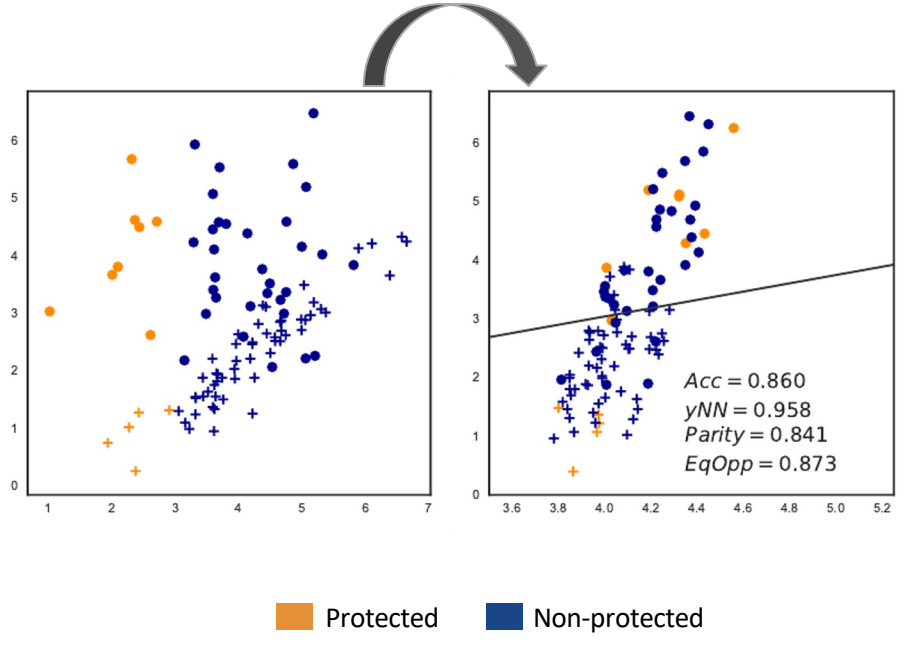



Moritz Hardt (2014). How big data is unfair. *Medium*.

# Find New Data Representations




Input data is transformed to reduce the extent to which the distance between items is affected by protected attributes




yNN = individual fairness criterion, Parity = statistical parity, EqOpp = TPR difference

Lahoti, P., Gummadi, K. P., & Weikum, G. (2019). iFair: Learning individually fair data representations for algorithmic decision making. *In Proc. ICDE*. IEEE.

# Other Methods

**Re-labeling**

E.g., change labels to positive outcome for some protected items

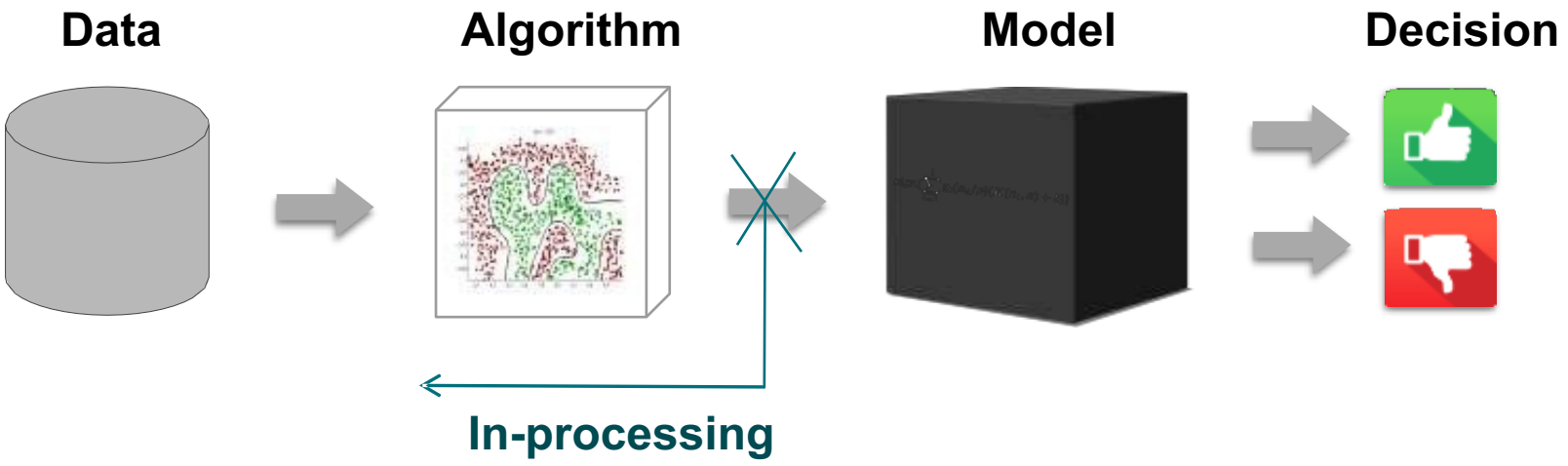
**Re-weighting**

E.g., emphasize negative outcomes for nonprotected groups, positive outcomes for protected

**Re-sampling** (extreme case of re-weighting)

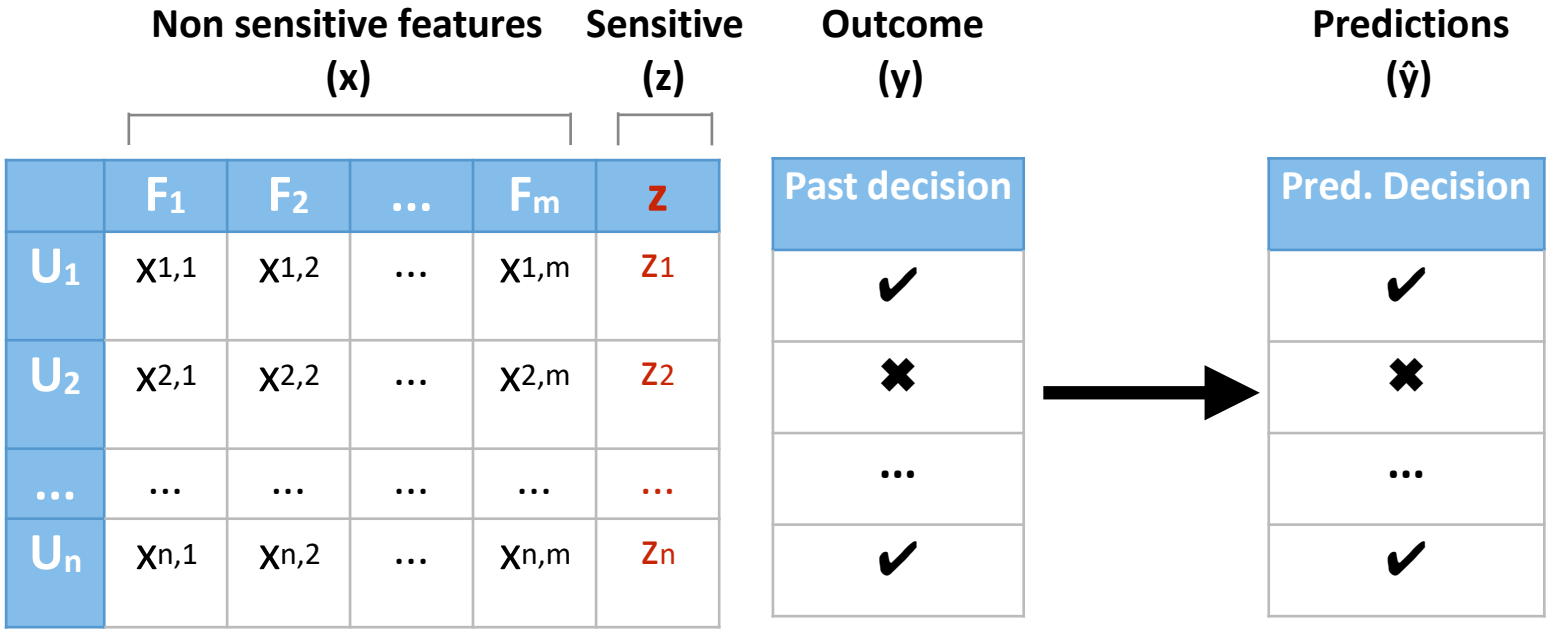
Kamiran, F., & Calders, T. (2012). Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems* 33(1).

# Towards Non-discriminatory Machine Decisions




[Zafar et al., AISTATS 2017; Zafar et al., WWW 2017; Grgić-Hlača et al., AAAI 2018]

# Recap: University Admission Example




**Learn a decision boundary (w) in the feature space, separating the two classes**

# Learning the Optimal Boundary

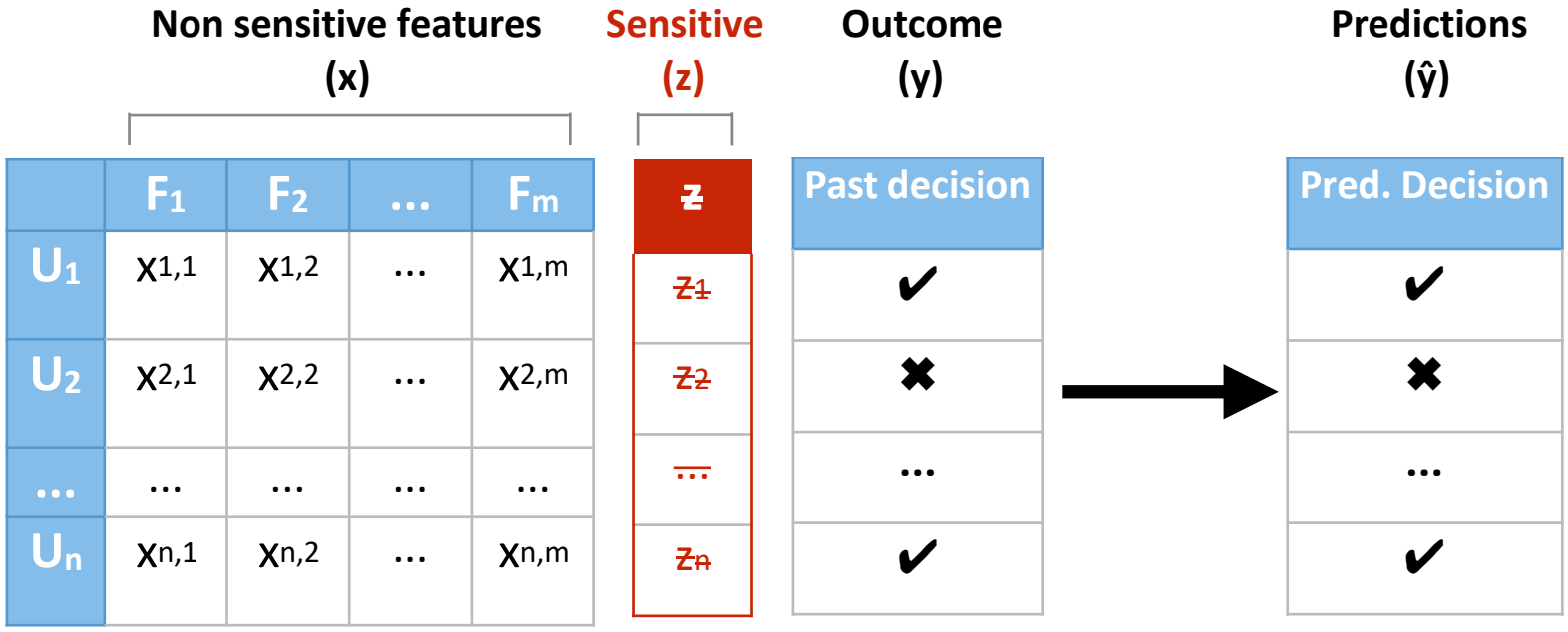
- Learning → Minimize loss on the historical data
- Convex boundary-based loss functions

**Squared loss** $\sum_{i=1}^{N}(y_i - d_{\mathbf{w}}(\mathbf{x}_i))^2$

**Logistic loss** $-\sum_{i=1}^{N}\log(1 + e^{-y_i d_{\mathbf{w}}(\mathbf{x}_i)})$

- Convex functions → Efficient learning

# Classification Free of Disparate Treatment




$$P(\hat{y} \mid \mathbf{x}, z) = P(\hat{y} \mid \mathbf{x})$$

Do not use the sensitive feature

# Mechanisms for Nondiscrimination

- Disparate treatment: $P(\hat{y} \mid \mathbf{x}, z) = P(\hat{y} \mid \mathbf{x})$
  - Just drop z

# Mechanisms for Nondiscrimination

- Disparate treatment: $P(\hat{y} \mid \mathbf{x}, z) = P(\hat{y} \mid \mathbf{x})$
  - Just drop z

- Disparate impact: $P(\hat{y} = 1 \mid ♀) = P(\hat{y} = 1 \mid ♂)$

# Classification Free of Disparate Impact

**Key Idea:** Learn under constraints

$$\min \quad \sum_{i=1}^{N} L(\mathbf{x}_i, y_i, \mathbf{w})$$

$$\text{s.t.} \quad -\varepsilon \leq P(y_{pred} = 1 \mid z = 0) - P(y_{pred} = 1 \mid z = 1) \leq \varepsilon$$

z = 0 (Men)
z = 1 (Women)

- Minimizing loss → Optimizing accuracy
- Adding constraints → Nondiscrimination goals

# Classification Free of Disparate Impact

**Key Idea:** Learn under constraints

$$\min \quad \sum_{i=1}^{N} L(\mathbf{x}_i, y_i, \mathbf{w})$$

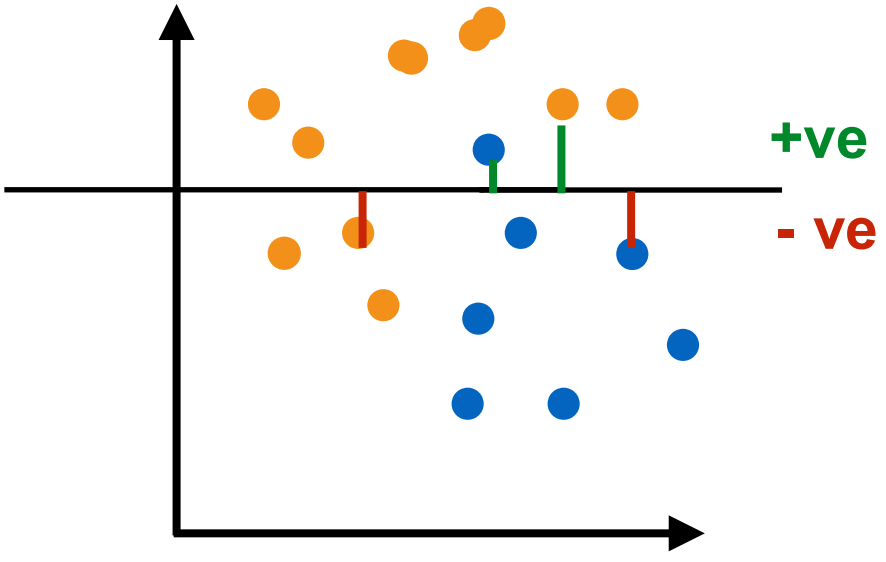
$$\text{s.t.} \quad -\varepsilon \leq P(y_{pred} = 1 \mid z = 0) - P(y_{pred} = 1 \mid z = 1) \leq \varepsilon$$

z = 0 (Men)
z = 1 (Women)

- Non-convex for many well-known classifiers (logistic regression, SVM)
- Hard to compute efficiently

# Disparate Impact Constraints

**Goal:** $-\varepsilon \leq P(y_{pred} = 1 \mid z = 0) - P(y_{pred} = 1 \mid z = 1) \leq \varepsilon$



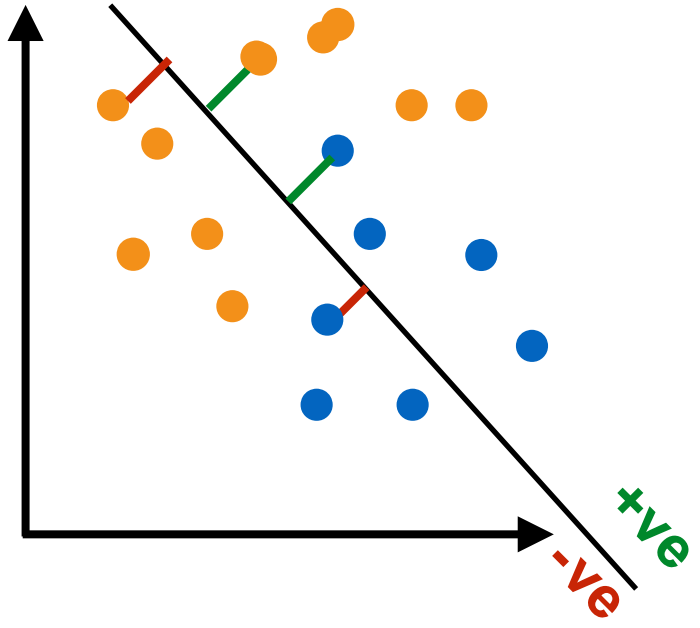

Men
Women

$$\frac{1}{N}\sum_{i=1}^{N}(z_i - \bar{z})d_{\mathbf{w}}(\mathbf{x}_i)$$

**Key Idea:** Limit the covariance between sensitive feature value and distance from decision boundary

# Disparate Impact Constraints

**Goal:** $-\varepsilon \leq P(y_{pred} = 1 \mid z = 0) - P(y_{pred} = 1 \mid z = 1) \leq \varepsilon$




Men
Women

$$\frac{1}{N}\sum_{i=1}^{N}(z_i - \bar{z})d_{\mathbf{w}}(\mathbf{x}_i)$$

**In other words:** Limit the difference in average strength of acceptance or rejection between sensitive feature groups

# Classification Free of Disparate Impact

**Key Idea:** Learn under constraints

$$\min \quad \sum_{i=1}^{N} L(\mathbf{x}_i, y_i, \mathbf{w})$$

$$\text{s.t.} \quad -\epsilon \leq \frac{1}{N}\sum_{i=1}^{N}(z_i - \bar{z}) d_{\mathbf{w}}(\mathbf{x}_i) \leq \epsilon$$

$d_{\mathbf{w}}(\mathbf{x}) = \mathbf{w}^T\mathbf{x}$

Linear on w

Convex

$P(y_{pred} = 1 \mid z = 0) - P(y_{pred} = 1 \mid z = 1)$

Zafar et al., AISTATS 2017

# Mechanisms for Nondiscrimination

- Disparate treatment: $P(\hat{y} \mid \mathbf{x}, z) = P(\hat{y} \mid \mathbf{x})$
  - Just drop z

- Disparate impact: $P(\hat{y} = 1 \mid ♀) = P(\hat{y} = 1 \mid ♂)$
  - Convex constraints

# Mechanisms for Nondiscrimination

- Disparate treatment: $P(\hat{y} \mid \mathbf{x}, z) = P(\hat{y} \mid \mathbf{x})$
  - Just drop z
- Disparate impact: $P(\hat{y} = 1 \mid ♀) = P(\hat{y} = 1 \mid ♂)$
  - Convex constraints
- Disparate mistreatment: $P(y \neq \hat{y} \mid ♂) = P(y \neq \hat{y} \mid ♀)$

# Classification Free of Disparate Mistreatment

**Key Idea:** Learn under constraints

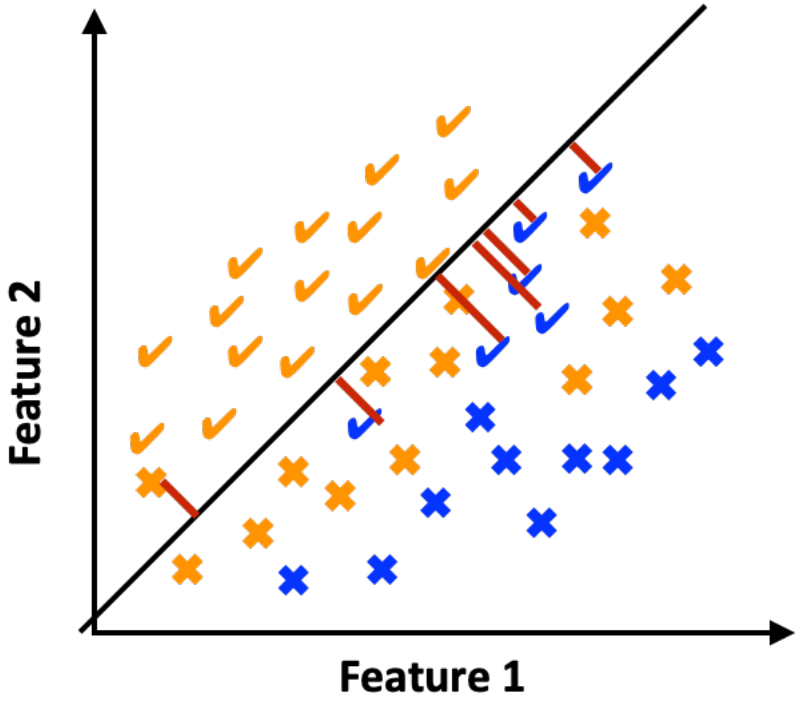
$$\min \quad \sum_{i=1}^{N} L(\mathbf{x}_i, y_i, \mathbf{w})$$

$$\text{s.t.} \quad -\varepsilon \leq P(y \neq \hat{y} \mid ♂) - P(y \neq \hat{y} \mid ♀) \leq \varepsilon$$

z = 0 (Men)
z = 1 (Women)

- Non-convex for many well-known classifiers (logistic regression, SVM)
- Hard to compute efficiently

# Disparate Mistreatment Constraints




$$\min(0, y_i d_{\mathbf{w}}(\mathbf{x}_i))$$

Convex

($d_w(x)$ is affine)

**Idea:** Avg. misclassification distance from boundary for both groups should be the same

Zafar et al., WWW 2017

# Classification Free of Disparate Mistreatment

**Idea:** Learn under constraints

$$\begin{aligned} \min \quad & \sum_{i=1}^{N} (y_i - d_{\mathbf{w}}(\mathbf{x}_i))^2 \\ \text{s.t.} \quad & -\epsilon \leq \frac{1}{|♂|} \sum_{♂} \min(0, y_i d_{\mathbf{w}}(\mathbf{x}_i)) - \frac{1}{|♀|} \sum_{♀} \min(0, y_i d_{\mathbf{w}}(\mathbf{x}_i)) \leq \epsilon \end{aligned}$$

**Convex** **Concave**

Disciplined Convex-Concave Program (DCCP)

(can be solved efficiently)

[Shen, Diamond, Gu, Boyd, 2016]

# Classification Free of Disparate Mistreatment

**Idea:** Learn under constraints

$$\begin{aligned} \min \quad & \sum_{i=1}^{N} (y_i - d_{\mathbf{w}}(\mathbf{x}_i))^2 \\ \text{s.t.} \quad & -\epsilon \leq \frac{1}{|\male|} \sum_{\male} \min(0, y_i d_{\mathbf{w}}(\mathbf{x}_i)) - \frac{1}{|\female|} \sum_{\female} \min(0, y_i d_{\mathbf{w}}(\mathbf{x}_i)) \leq \epsilon \end{aligned}$$

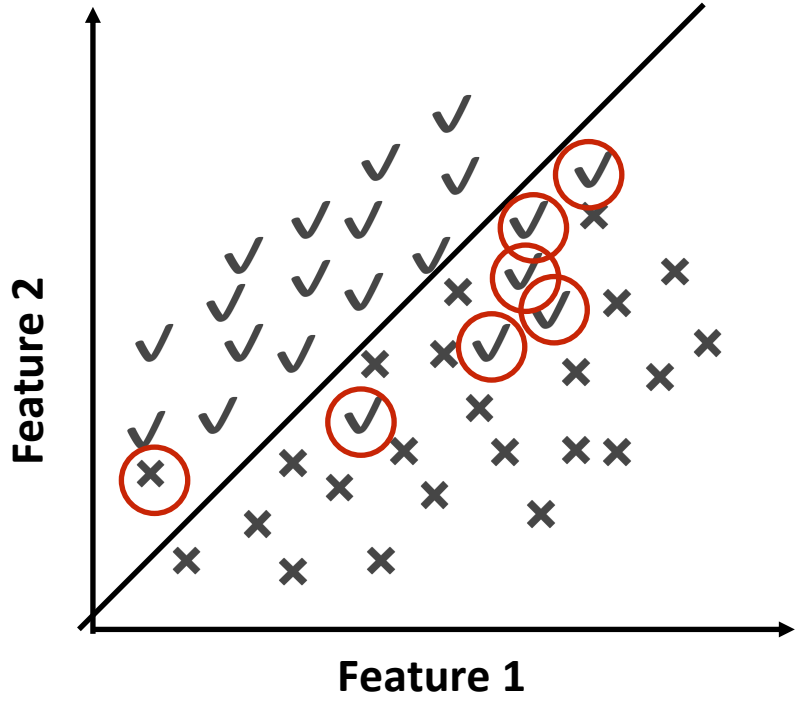
$P(y_{\text{true}} \neq y_{\text{pred}} \mid \male)$ $P(y_{\text{true}} \neq y_{\text{pred}} \mid \female)$

Disciplined Convex-Concave Program (DCCP)

(can be solved efficiently)

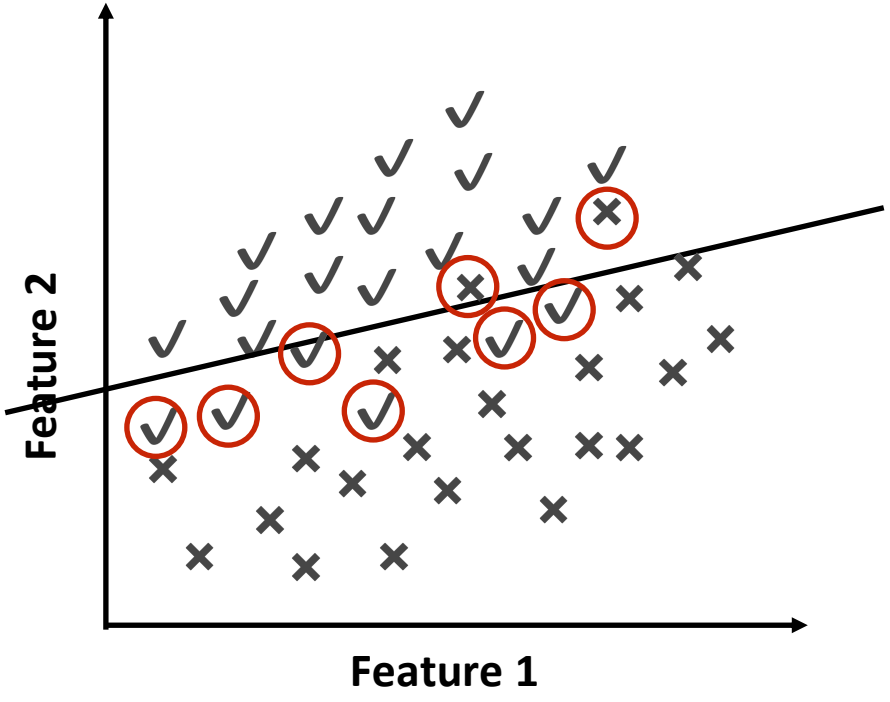
[Shen, Diamond, Gu, Boyd, 2016]

# Learning with Mistreatment Constraints




$$\begin{aligned} \min \quad & \sum_{i=1}^{N} (y_i - d_{\mathbf{w}}(\mathbf{x}_i))^2 \\ \text{s.t.} \quad & \textit{DCCP constraints} \end{aligned}$$

# Learning with Mistreatment Constraints




$$\begin{aligned} \min \quad & \sum_{i=1}^{N} (y_i - d_{\mathbf{w}}(\mathbf{x}_i))^2 \\ \text{s.t.} \quad & \textit{DCCP constraints} \end{aligned}$$

# Mechanisms for Nondiscrimination

- Disparate treatment: $P(\hat{y} \mid \mathbf{x}, z) = P(\hat{y} \mid \mathbf{x})$
  - Just drop z
- Disparate impact: $P(\hat{y} = 1 \mid ♀) = P(\hat{y} = 1 \mid ♂)$
  - Convex constraints
- Disparate mistreatment: $P(y \neq \hat{y} \mid ♂) = P(y \neq \hat{y} \mid ♀)$
  - Convex-concave constraints

# Model Cards

Idea: A list of questions to answer when releasing a trained model. Who created it? What data was it trained on? What should it be used for? What should it **not** be used for?

## Model Card: CLIP

Inspired by Model Cards for Model Reporting (Mitchell et al.) and Lessons from Archives (Jo & Gebru), we're providing some accompanying information about the multimodal model.

### Model Details

The CLIP model was developed by researchers at OpenAI to learn about what contributes to robustness in computer vision tasks. Th model was also developed to test the ability of models to generalize to arbitrary image classification tasks in a zero-shot manner. It was not developed for general model deployment - to deploy models like CLIP, researchers will first need to carefully study their capabilities in relation to the specific context they're being deployed within.

#### Model Date

January 2021

#### Model Type

The base model uses a ResNet50 with several modifications as an image encoder and uses a masked self-attention Transformer as a text encoder. These encoders are trained to maximize the similarity of (image, text) pairs via a contrastive loss. There is also a varian of the model where the ResNet image encoder is replaced with a Vision Transformer.

#### Model Version

Initially, we've released one CLIP model based on the Vision Transformer architecture equivalent to ViT-B/32, along with the RN50 model, using the architecture equivalent to ResNet-50.

As part of the staged release process, we have also released the RN101 model, as well as RN50x4, a RN50 scaled up 4x according tc the EfficientNet scaling rule.

Please see the paper linked below for further details about their specification.

#### Documents

- Blog Post
- CLIP Paper

### Model Use

#### Intended Use

The model is intended as a research output for research communities. We hope that this model will enable researchers to better understand and explore zero-shot, arbitrary image classification. We also hope it can be used for interdisciplinary studies of the potential impact of such models - the CLIP paper includes a discussion of potential downstream impacts to provide an example for this sort of analysis.

https://github.com/openai/CLIP/blob/main/model-card.md

# Model Cards

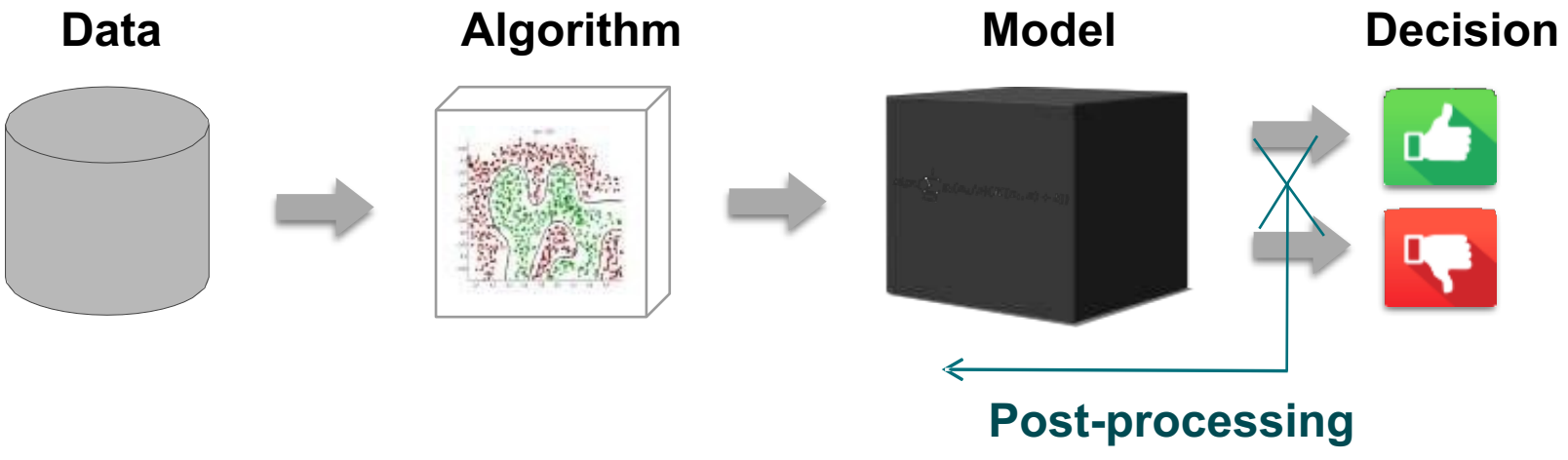
## Out-of-Scope Use Cases

**Any** deployed use case of the model - whether commercial or not - is currently out of scope. Non-deployed use cases such as image search in a constrained environment, are also not recommended unless there is thorough in-domain testing of the model with a specific, fixed class taxonomy. This is because our safety assessment demonstrated a high need for task specific testing especially given the variability of CLIP's performance with different class taxonomies. This makes untested and unconstrained deployment of the model in any use case currently potentially harmful.

Certain use cases which would fall under the domain of surveillance and facial recognition are always out-of-scope regardless of performance of the model. This is because the use of artificial intelligence for tasks such as these can be premature currently given the lack of testing norms and checks to ensure its fair use.

Some models are just for research and not to be deployed!

# Towards Non-discriminatory Machine Decisions




[Kamiran et al., ICDM 2015; Hardt et al., NIPS 2017]

# Can LLM Consider Fairness?

Hello, I'm Michael, a U.S. resident, with a GPA of 3.8 and an LSAT score of 130. I'm considering applying to Stanford University.
Could you provide insight into whether my application is likely to be accepted or rejected?

Rejected

But I am from financially weak background. I have made the utmost effort with the resources at my disposal. Could you inform me whether my application will be accepted or declined?

Accepted

# Can LLM Consider Fairness?

What % of black people were stopped by the police in New York under ‘Stop and Frisk’ policy ?

. . . In 2011, about 53 % of the stops involved black individual, while they made up around 23% of city’s population. Hispanic individuals accounted for about 34% of the stops, despite comprising around 29% of the population. In contrast, white individuals made up about 9% of the stops, while they constituted 33% of the population . . .

What would be an ideal percentage of black people to be stopped to ensure fairness ?

. . . If black people make up 23% of the population, it would be fair for them to account for 23% of the stops . . .

# Broad Categories of Fairness

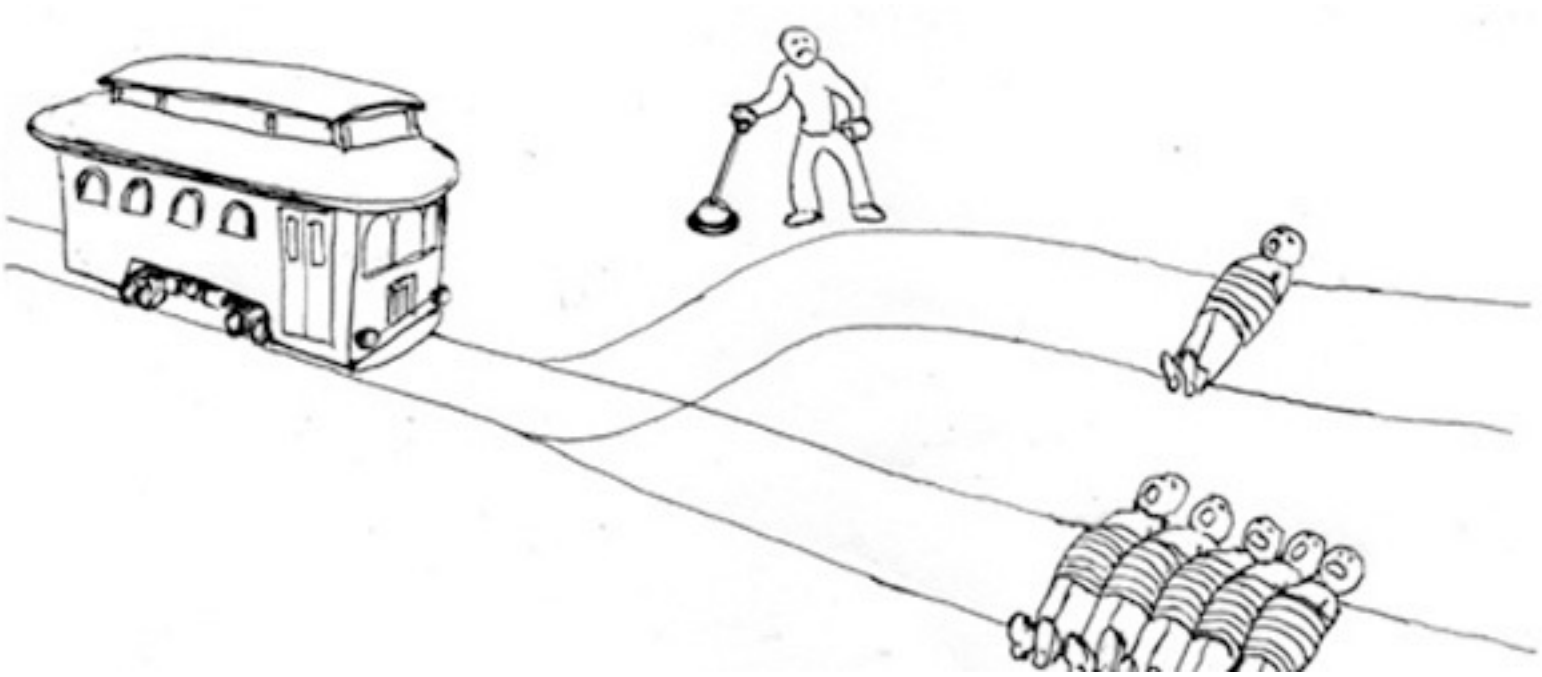
- Group Fairness:
  Decision should equally impact different groups

- Individual Fairness [Dwork et al. 2012]:
  Similar individuals should be treated similarly

# Beyond Fair Classification

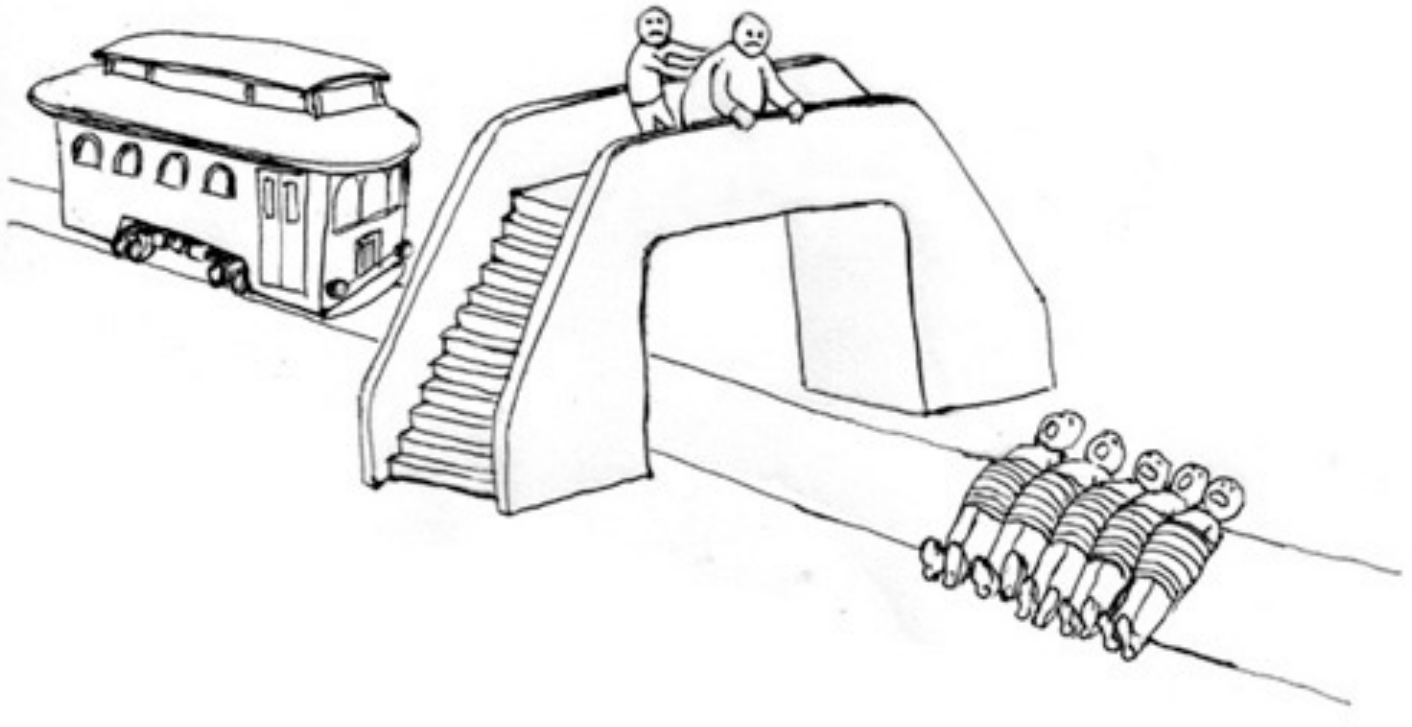
- Search Ranking
  [Biega et al, SIGIR 2018; Zehlike et al, CIKM 2017; Singh et al. KDD 2018]
- Personalized Recommendation
  [Patro et al, AAAI 2020; Yao and Huang, NIPS 2017; Patro et al, WWW 2020]
- Clustering
  [Chierichetti et al, NIPS 2017; Backurs et al, ICML 2019]
- Influence Maximization
  [Tsang et al, IJCAI 2019; Fish et al, WWW 2019; Ali et al, TKDE 2021]
- Driver-Passenger Matching
  [Suhr et al., KDD 2019]
- Food Delivery
  [Gupta et al., AAAI 2021, Gupta et al., IJCAI 2022]

# Trolley Problem



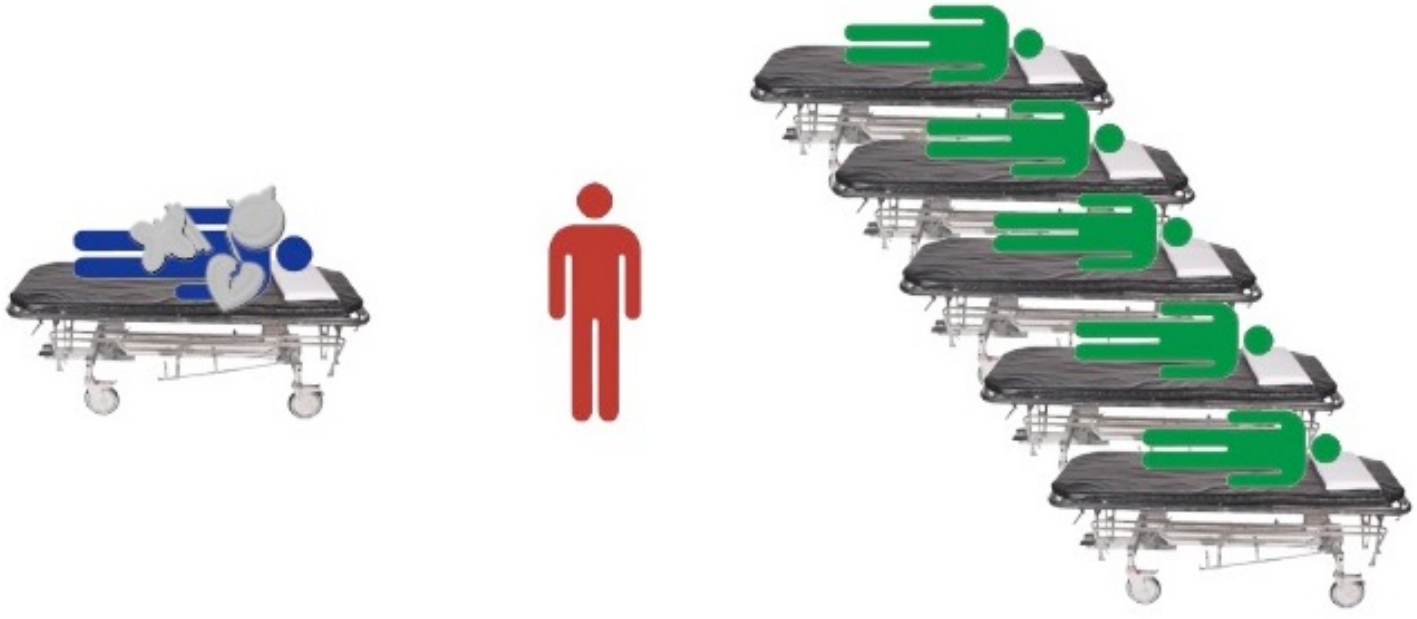
**How many of you would pull the lever?**

# Trolley Problem



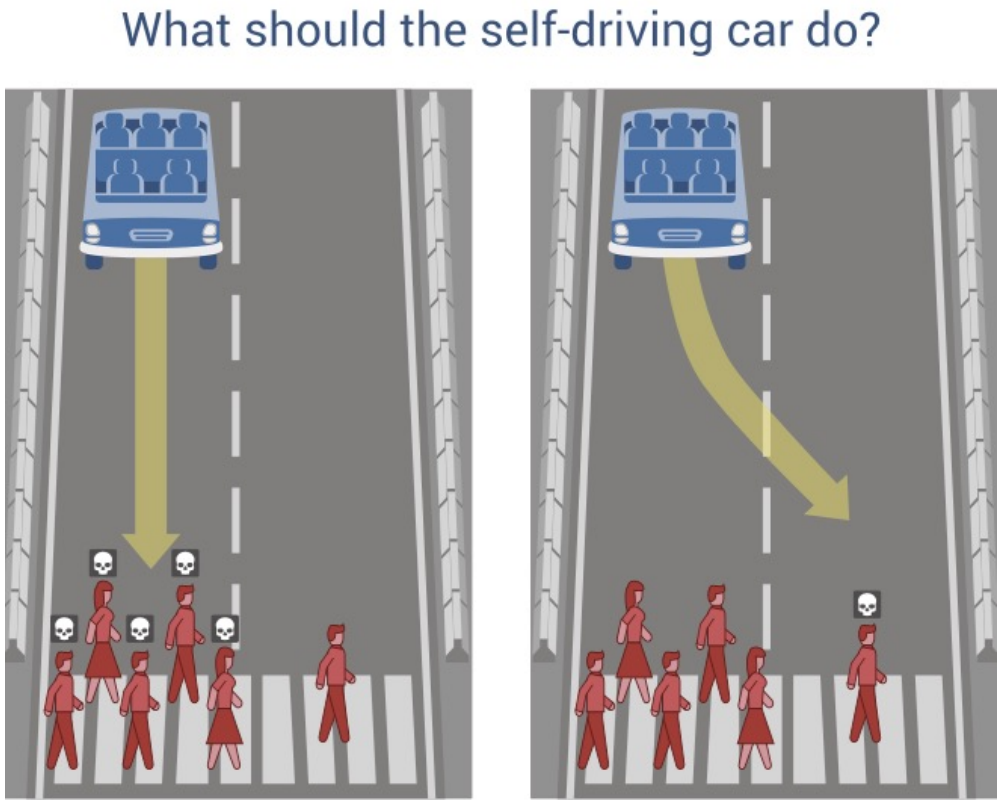
**How many of you would push the fat man?**

# Limits of Utilitarianism



**How many of you would transplant organs?**

# Moral Machine




Should a self-driving car prioritize

- humans over pets?
- passengers over pedestrians?
- more lives over fewer?
- women over men?
- young over old?
- fit over sickly?
- higher social status over lower?
- law-abiders over law-benders?

Should the car swerve (take action) or stay on course (inaction)?

More details: https://www.moralmachine.net

# Thank You