Fall 2022 | Lecture 20

Fairness

Ariel Procaccia | Harvard University

UNFAIRNESS

- AI algorithms are supposedly unbiased
- But they are trained based on data that encodes societal biases, and may exacerbate those biases
- There is a large body of evidence for discrimination by AI algorithms

EXAMPLE: AD DELIVERY

Title	URL	Coefficient	appears in agents		total appearances	
			female	male	female	male
Top ads for identifying the simulated female group						
Jobs (Hiring Now)	www.jobsinyourarea.co	0.34	6	3	45	8
4Runner Parts Service	www.westernpatoyotaservice.com	0.281	6	2	36	5
Criminal Justice Program	www3.mc3.edu/Criminal+Justice	0.247	5	1	29	1
Goodwill - Hiring	goodwill.careerboutique.com	0.22	45	15	121	39
UMUC Cyber Training	www.umuc.edu/cybersecuritytraining	0.199	19	17	38	30
Top ads for identifying agents in the simulated male group						
$200k+ Jobs - Execs Only	careerchange.com	−0.704	60	402	311	1816
Find Next $200k+ Job	careerchange.com	−0.262	2	11	7	36
Become a Youth Counselor	www.youthcounseling.degreeleap.com	−0.253	0	45	0	310
CDL-A OTR Trucking Jobs	www.tadrivers.com/OTRJobs	−0.149	0	1	0	8
Free Resume Templates	resume-templates.resume-now.com	−0.149	3	1	8	10

[Datta et al. 2015]

EXAMPLE: CRIMINAL JUSTICE

Bernard Parker, left, was rated high risk; Dylan Fugett was rated low risk. (Josh Ritchie for ProPublica)

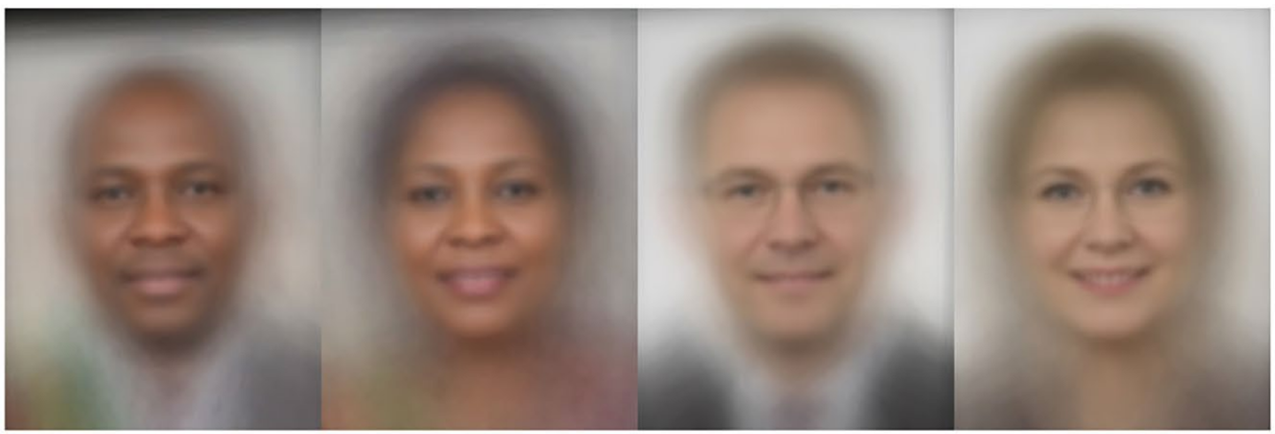
Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica
May 23, 2016

EXAMPLE: FACIAL RECOGNITION

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%



[Buolamwini, 2019]

Cynthia Dwork

1958–

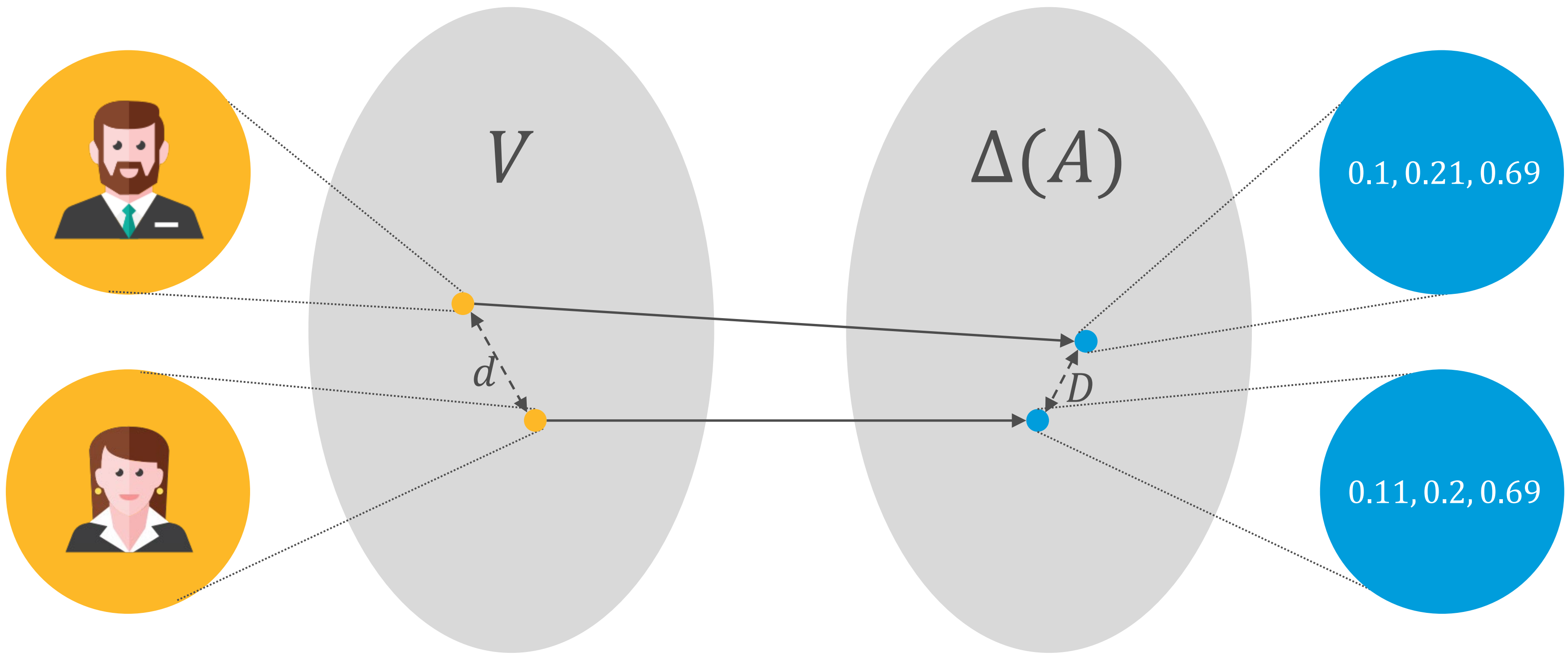
Professor of Computer Science at Harvard. In the last 15 years, played a pivotal role in the formation of differential privacy and fair AI.

INDIVIDUAL FAIRNESS

- Set of individuals V and outcomes A
- Randomized classifier $M: V \to \Delta(A)$ where $\Delta(A)$ is distributions over outcomes
- Metric on individuals $d: V \times V \to \mathbb{R}^+$
- Metric D on distributions over outcomes
- M satisfies the Lipschitz property if for all $x, y \in V$,

$$D\big(M(x), M(y)\big) \leq d(x, y)$$

INDIVIDUAL FAIRNESS



INDIVIDUAL FAIRNESS

- We can get a Lipschitz classifier by setting $M(x) = M(y)$ for all $x, y \in V$
- But we want to minimize a loss function

$$L: V \times A \to \mathbb{R}^+$$

- This leads to the optimization problem

$$\begin{aligned} \min \; & \sum_{x \in V} \sum_{a \in A} \mu_x(a) \cdot L(x, a) \\ \text{s.t.} \; & \forall x, y \in V, D\left(\mu_x, \mu_y\right) \leq d(x, y) \\ & \forall x \in V, \mu_x \in \Delta(A) \end{aligned}$$

INDIVIDUAL FAIRNESS

- Various options for the metric D
- Example: total variation, defined for distributions P and Q as

$$D_{tv}(P,Q) = \frac{1}{2}\sum_{a\in A}|P(a) - Q(a)|$$

- When $D = D_{tv}$, the optimization problem is a linear program
- Poll 1 (brainstorming): Where would the similarity metric d come from?

ENVY-FREENESS, REVISITED

- Each $x \in V$ has a utility u_{xa} for each outcome $a \in A$
- A randomized classifier M is envy free if and only if for all $x, y \in V$,

$$\mathbb{E}_{a \sim M(x)}[u_{xa}] \geq \mathbb{E}_{a \sim M(y)}[u_{xa}]$$

- This gives a completely different way of thinking about individual fairness
- But envy-freeness isn't useful in situations where there is a desirable and an undesirable outcome, like bail and loans

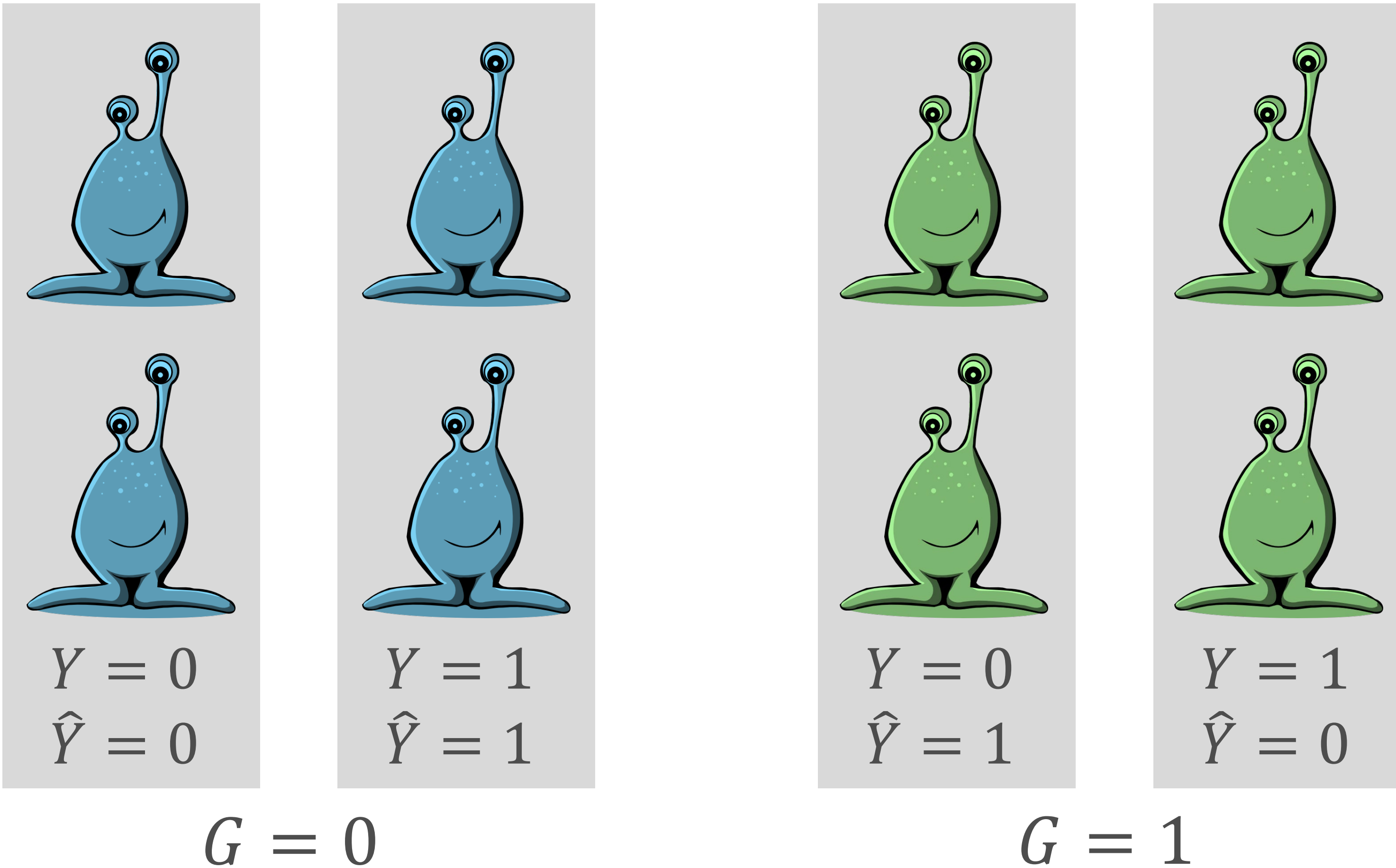
GROUP FAIRNESS

- Assume we are making a binary decision $\hat{Y} \in \{0,1\}$, and there is a legally protected attribute $G \in \{0,1\}$
- Demographic parity:

$$\Pr[\hat{Y} = 1 \mid G = 0] = \Pr[\hat{Y} = 1 \mid G = 1]$$

- May accept unqualified individuals when $G = 0$, and qualified individuals when $G = 1$!

GROUP FAIRNESS



This classifier satisfies demographic parity!

GROUP FAIRNESS

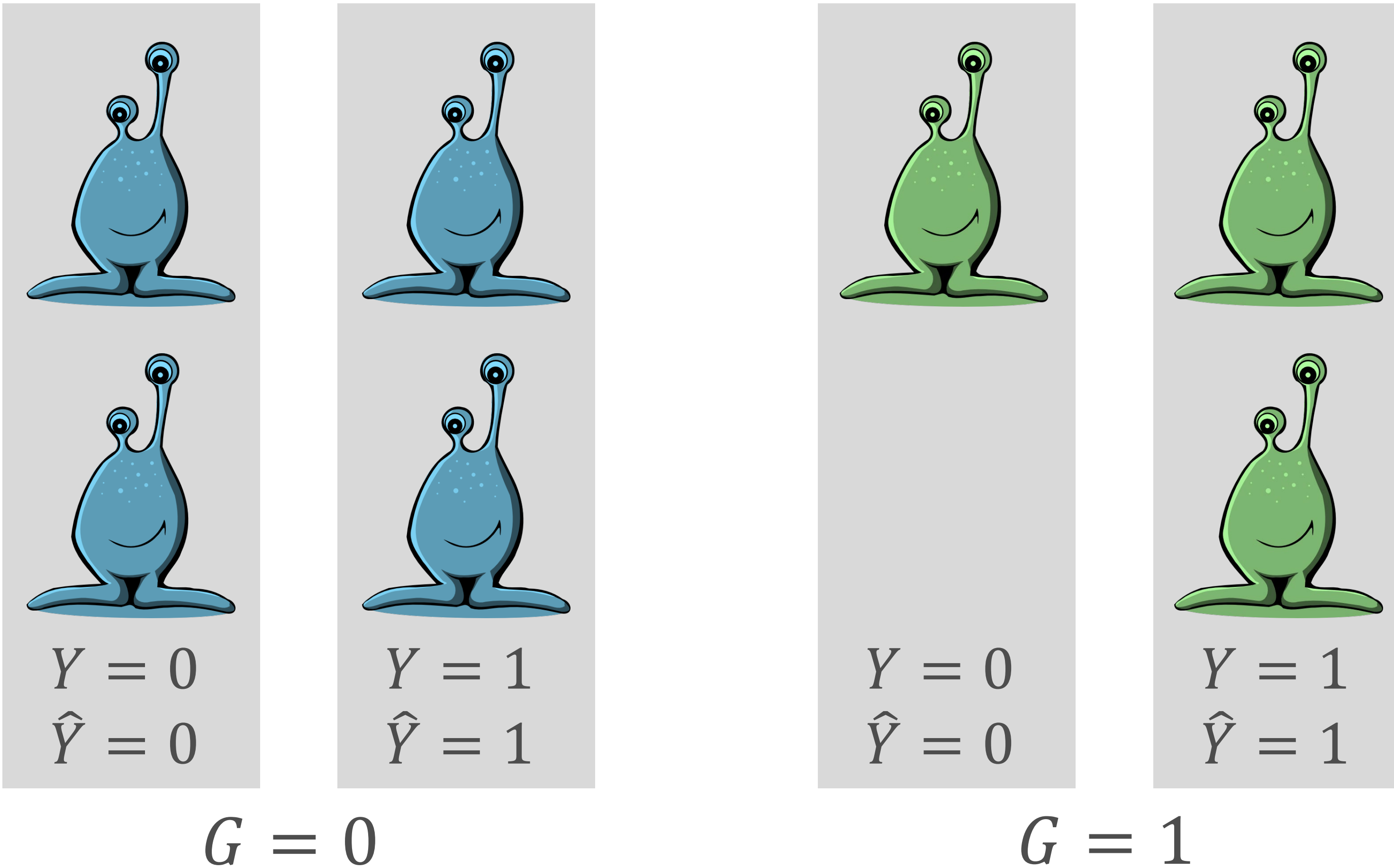
- $\hat{Y}$ satisfies equalized odds with respect to protected attribute G if the groups have equal false positive and false negative rates
- That is, for all $y, \hat{y} \in \{0,1\}$,

$$\Pr[\hat{Y} = \hat{y} \mid G = 0, Y = y] = \Pr[\hat{Y} = \hat{y} \mid G = 1, Y = y]$$

RELATIONS BETWEEN PROPERTIES

- Demographic parity:
  $$\Pr[\hat{Y} = 1 \mid G = 0] = \Pr[\hat{Y} = 1 \mid G = 1]$$
- Equalized odds: For all $y, \hat{y} \in \{0,1\}$,
  $$\Pr[\hat{Y} = \hat{y} \mid G = 0, Y = y]$$
  $$= \Pr[\hat{Y} = \hat{y} \mid G = 1, Y = y]$$
- Poll 2: Relation between demographic parity and equalized odds?
  - Demographic parity $\Rightarrow$ equalized odds
  - Equalized odds $\Rightarrow$ demographic parity
  - Incomparable

GROUP FAIRNESS

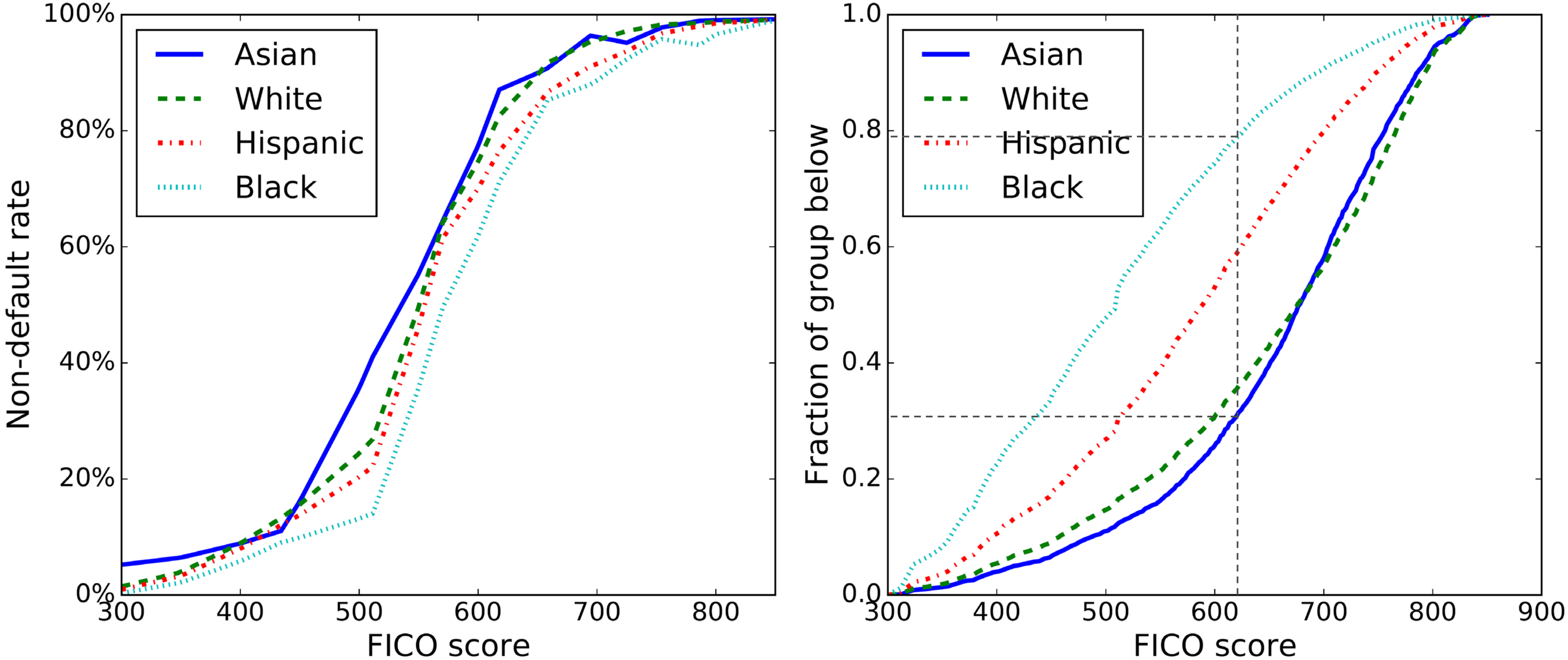


$\hat{Y} = Y$ may not satisfy demographic parity!

EXAMPLE: FICO SCORES

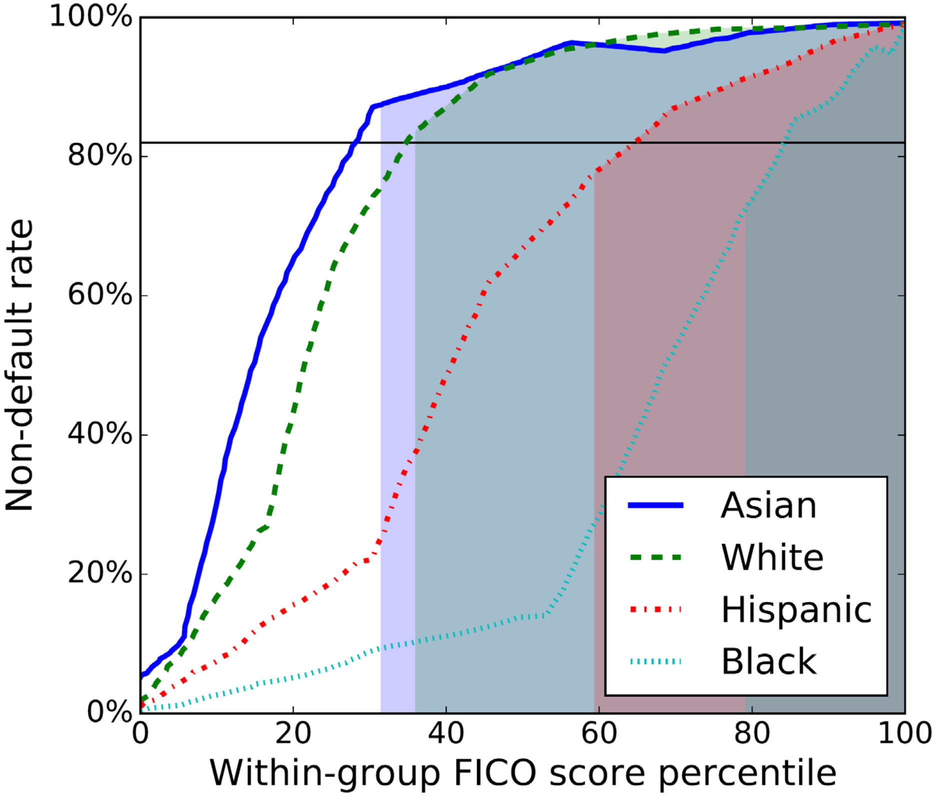
- FICO scores are a proprietary classifier widely used in the United States to predict credit worthiness
- Range from 300 to 850, where cutoff of 620 is commonly used for prime-rate loans, which corresponds to a default rate of 18%

EXAMPLE: FICO SCORES



[Hardt et al. 2016]

EXAMPLE: FICO SCORES



[Hardt et al. 2016]

IMPOSSIBILITY FOR RISK SCORES

- Each person has a feature vector $\boldsymbol{\sigma}$
- $p_{\boldsymbol{\sigma}}$ denotes the fraction of people with feature vector $\boldsymbol{\sigma}$ and a true positive label
- A person in group $G \in \{0,1\}$ has a given probability of exhibiting feature vector $\boldsymbol{\sigma}$
- A risk assignment is an assignment of people to bins, where each bin b is labeled with a score v_b seen as the probability of a positive label

IMPOSSIBILITY FOR RISK SCORES

- Calibration within groups is achieved when for each group G and each bin b, the expected number of members of group G in b who belong to the positive class is a v_b fraction of the expected number of members of group G assigned to b
- Equalized odds requires that the average score assigned to members of group 0 who belong to the negative (resp., positive) class would be the same as the average score assigned to people of group 1 who belong to the negative (resp., positive) class

IMPOSSIBILITY FOR RISK SCORES

- Can we achieve calibration together with equalized odds?
  - Perfect prediction: For each feature vector $\boldsymbol{\sigma}$, either $p_{\boldsymbol{\sigma}} = 0$ or $p_{\boldsymbol{\sigma}} = 1$
  - Equal base rates: The two groups have the same fraction of members in the positive class
- Theorem: If a risk assignment satisfies calibration and equalized odds, the instance must allow for perfect prediction or have equal base rates

FAIRNESS IN INDUSTRY