

# Graduate AI

## Lecture 27:

## Ethics and AI II

Teachers:  
Zico Kolter  
Ariel Procaccia (this time)

# FAIRNESS IN ML

- AI algorithms are supposedly unbiased
- But they may make use of features that interact with protected attributes
- For example, zip code is sometimes correlated with race
- There is a fast-growing body of evidence for discrimination by AI algorithms



# EXAMPLE: AD DELIVERY

## Racism is Poisoning Online Ad Delivery, Says Harvard Professor

Google searches involving black-sounding names are more likely to serve up ads suggestive of a criminal record than white-sounding names, says computer scientist

February 4, 2013

**“Have you ever been arrested? Imagine the question not appearing in the solitude of your thoughts as you read this paper, but appearing explicitly whenever someone queries your name in a search engine.”**

So begins Latanya Sweeney at Harvard University in a compelling paper arguing that racial discrimination plagues online ad delivery.

Many people will have experience Googling friends, colleagues and relatives to find out about their online presence—the websites on which they appear, their pictures, hobbies and so on.

Ad related to latanya sweeney ⓘ  
**Latanya Sweeney Truth**  
www.instantcheckmate.com/  
Looking for Latanya Sweeney? Check Latanya Sweeney's Arrests.

Ads by Google  
**Latanya Sweeney Arrested?**  
1) Enter Name and State. 2) Access Full Background Checks Instantly.  
www.instantcheckmate.com/

**Latanya Sweeney**  
Public Records Found For: Latanya Sweeney. View Now.  
www.publicrecords.com/

**La Tanya**  
Search for La Tanya Look Up Fast Results now!  
www.ask.com/La+Tanya

Screenshot of a Google ad.

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

[Datta et al. 2016]

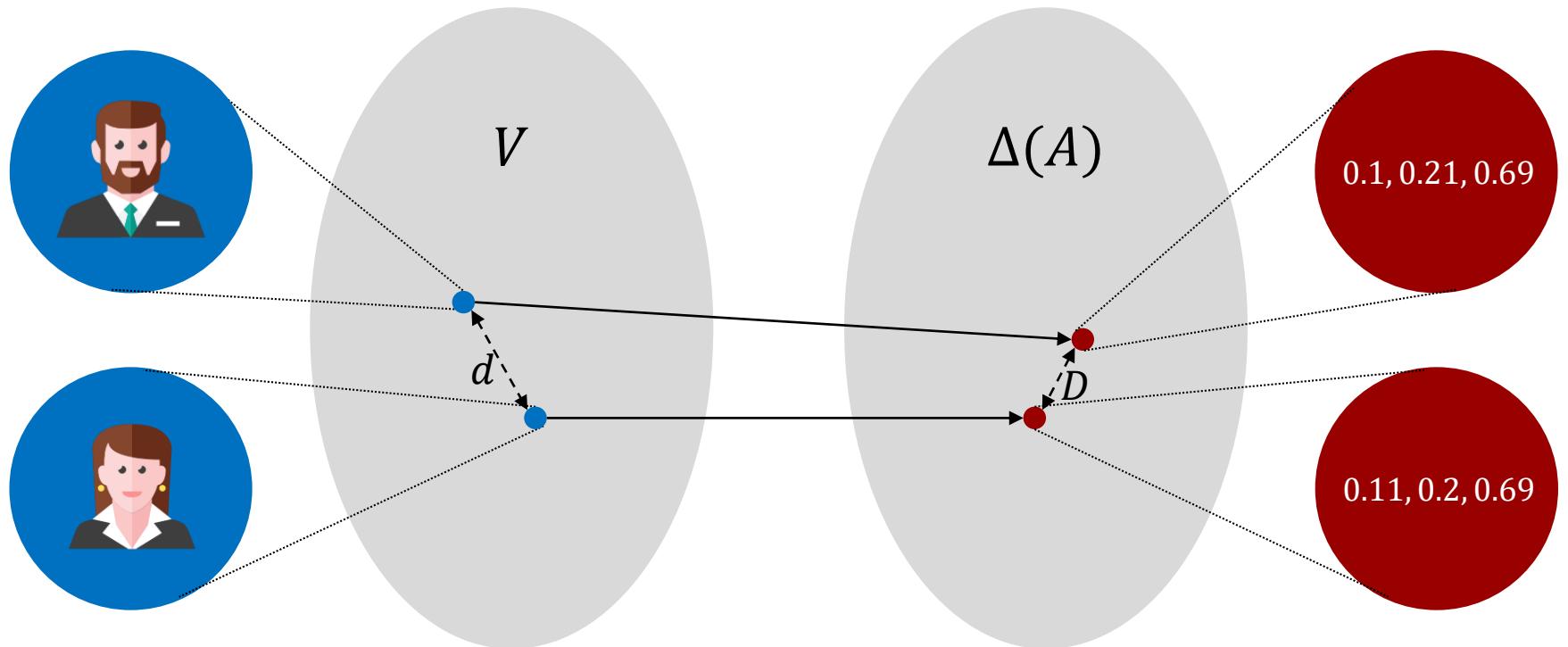
# INDIVIDUAL FAIRNESS

- Model introduced by Dwork et al. (2012)
- Set of individuals  $V$  and outcomes  $A$
- Metric on individuals  $d: V \times V \rightarrow \mathbb{R}^+$
- Metric  $D$  on distributions over outcomes
- Randomized classifier  $M: V \rightarrow \Delta(A)$
- $M$  satisfies the **Lipschitz property** if for all  $x, y \in V$ ,

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



# 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  $L: V \times A \rightarrow \mathbb{R}^+$
- This leads to the optimization problem

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

# INDIVIDUAL FAIRNESS

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

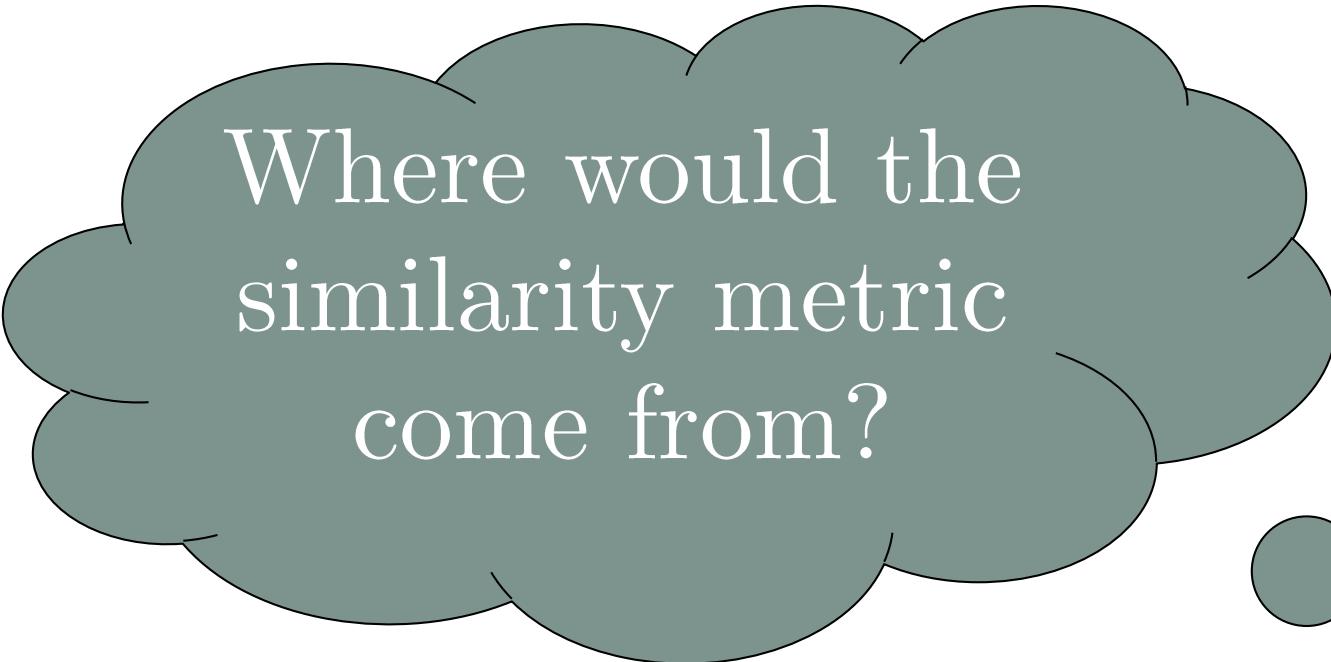
$$D_{tv}(P, Q) = \max_{E \subseteq A} |P(E) - Q(E)|$$

- Lemma: (we skip the simple proof)

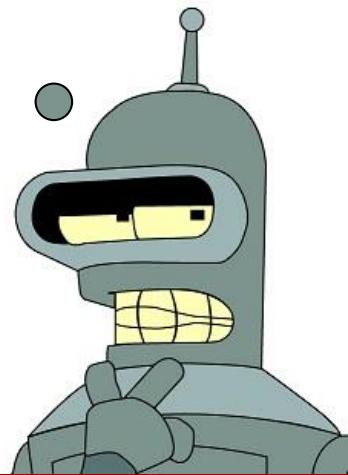
$$D_{tv}(P, Q) = \frac{1}{2} \sum_{a \in A} |P(a) - Q(a)|$$

- When  $D = D_{tv}$ , the optimization problem is a linear program!





Where would the  
similarity metric  
come from?

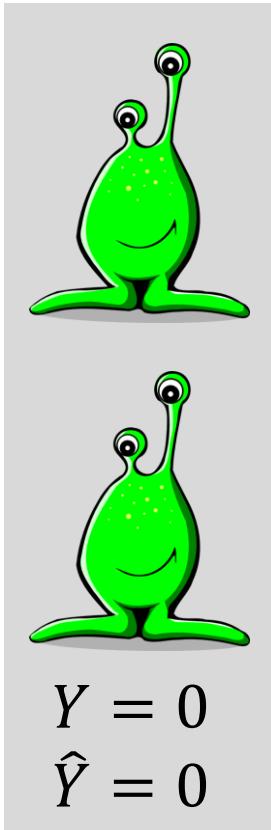


# GROUP FAIRNESS

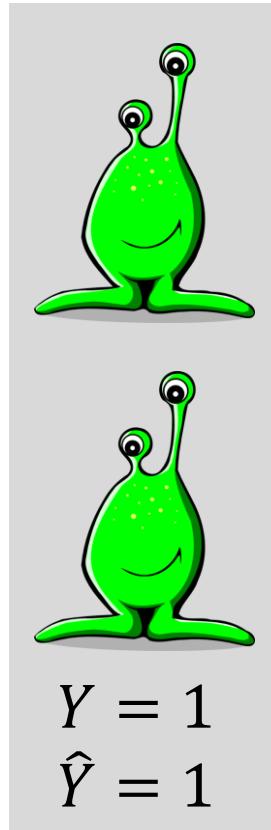
- Assume we are making a binary decision  $\hat{Y} \in \{0,1\}$ , and there is a protected attribute  $A \in \{0,1\}$
- Demographic parity:  
$$\Pr[\hat{Y} = 1 | A = 0] = \Pr[\hat{Y} = 1 | A = 1]$$
- May accept unqualified individuals when  $A = 0$ , and qualified individuals when  $A = 1$ !



# GROUP FAIRNESS

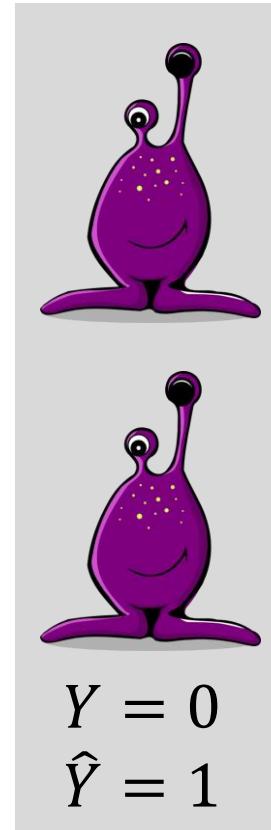


$$Y = 0$$
$$\hat{Y} = 0$$

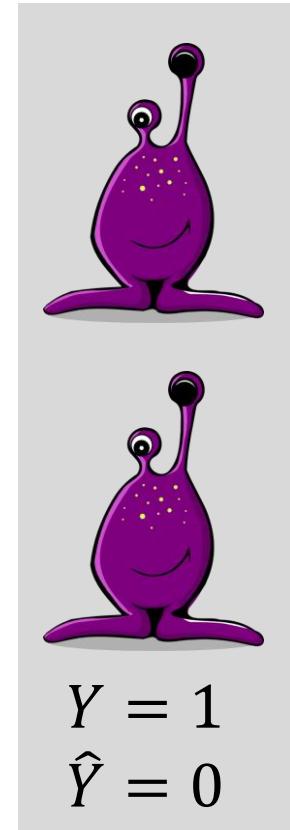


$$Y = 1$$
$$\hat{Y} = 1$$

$$A = 0$$



$$Y = 0$$
$$\hat{Y} = 1$$



$$Y = 1$$
$$\hat{Y} = 0$$

$$A = 1$$

# GROUP FAIRNESS

- We will follow the exposition of Hardt et al. [2016]
- $\hat{Y}$  satisfies **equalized odds** with respect to protected attribute  $A$  and outcome  $Y$  if  $\hat{Y}$  and  $A$  are independent conditional on  $Y$
- That is, for all  $y \in \{0,1\}$ ,

$$\begin{aligned}\Pr[\hat{Y} = 1 \mid A = 0, Y = y] \\ = \Pr[\hat{Y} = 1 \mid A = 1, Y = y]\end{aligned}$$

# RELATIONS BETWEEN PROPERTIES

- Demographic parity:

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

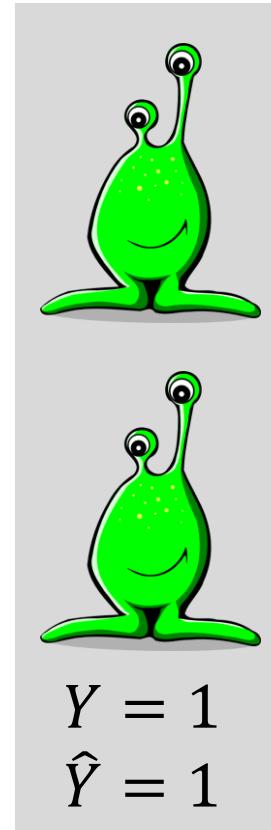
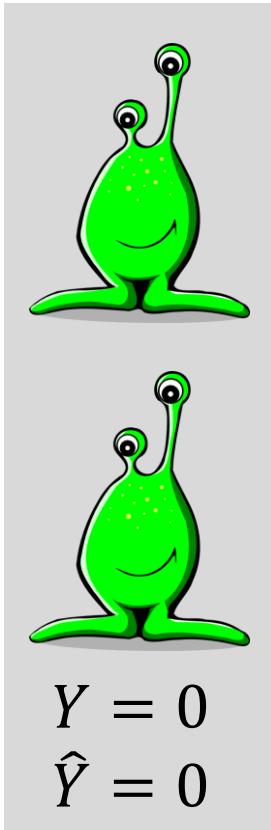
- Equalized odds: For all  $y \in \{0,1\}$ ,

$$\begin{aligned}\Pr[\hat{Y} = 1 \mid A = 0, Y = y] \\ = \Pr[\hat{Y} = 1 \mid A = 1, Y = y]\end{aligned}$$

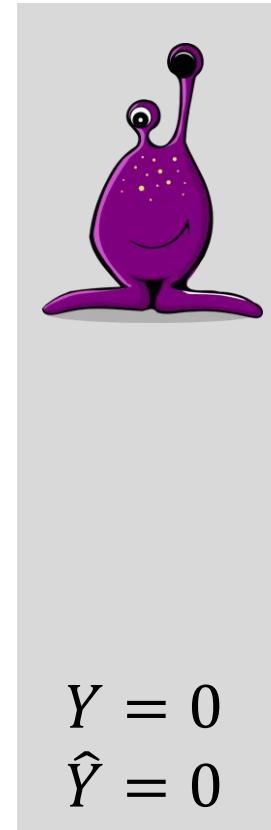
- Poll 1: Relation between demographic parity and equalized odds?
  1. Demographic parity  $\Rightarrow$  equalized odds
  2. Equalized odds  $\Rightarrow$  demographic parity
  3. Incomparable



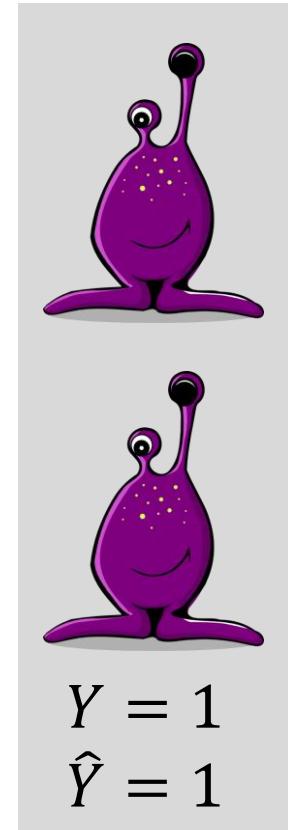
# GROUP FAIRNESS



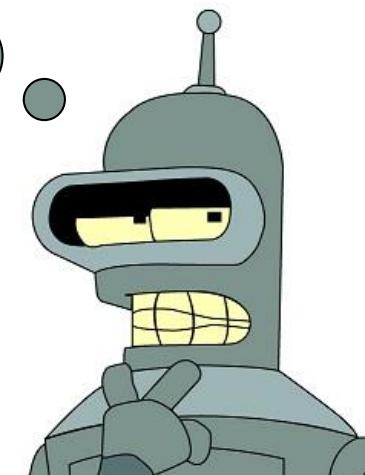
$A = 0$



$A = 1$



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

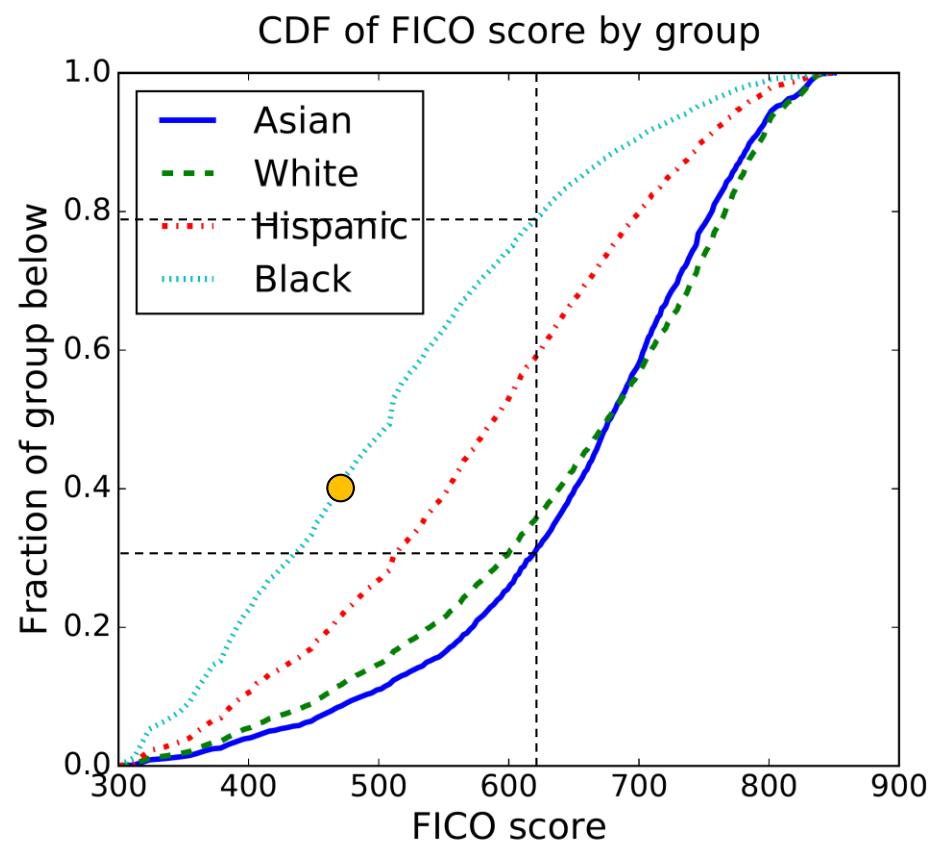
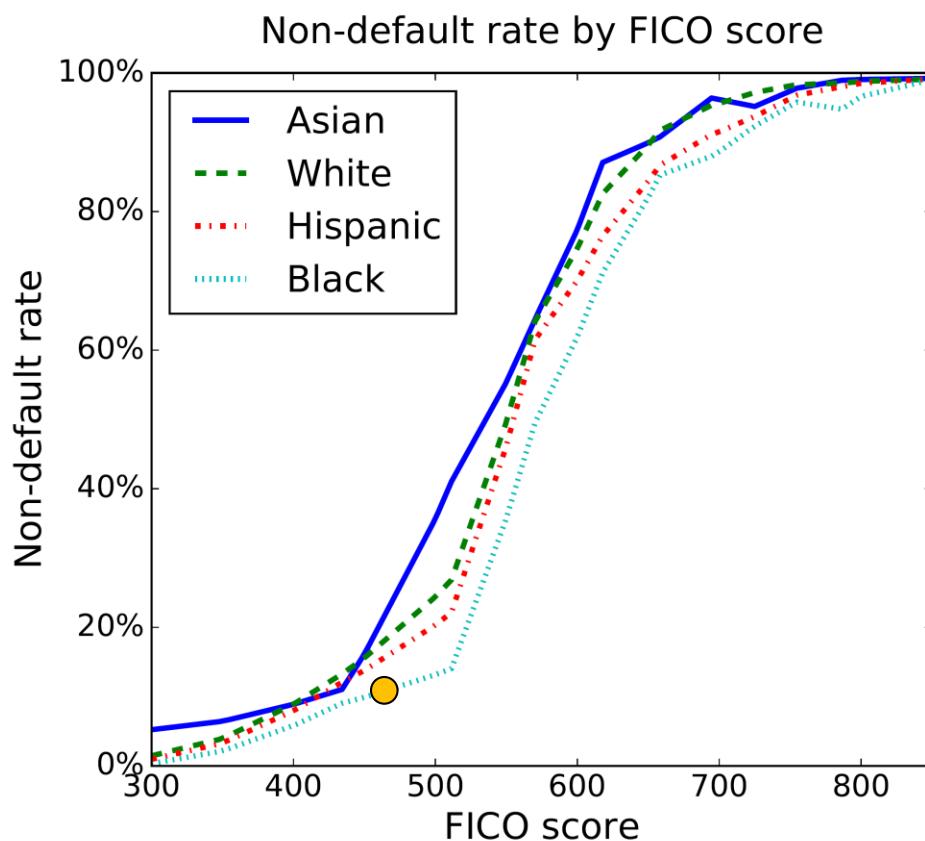


# EXAMPLE: FICO SCORES

- 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
- Based on features, such as number of bank accounts used, that may interact with race in unfair ways

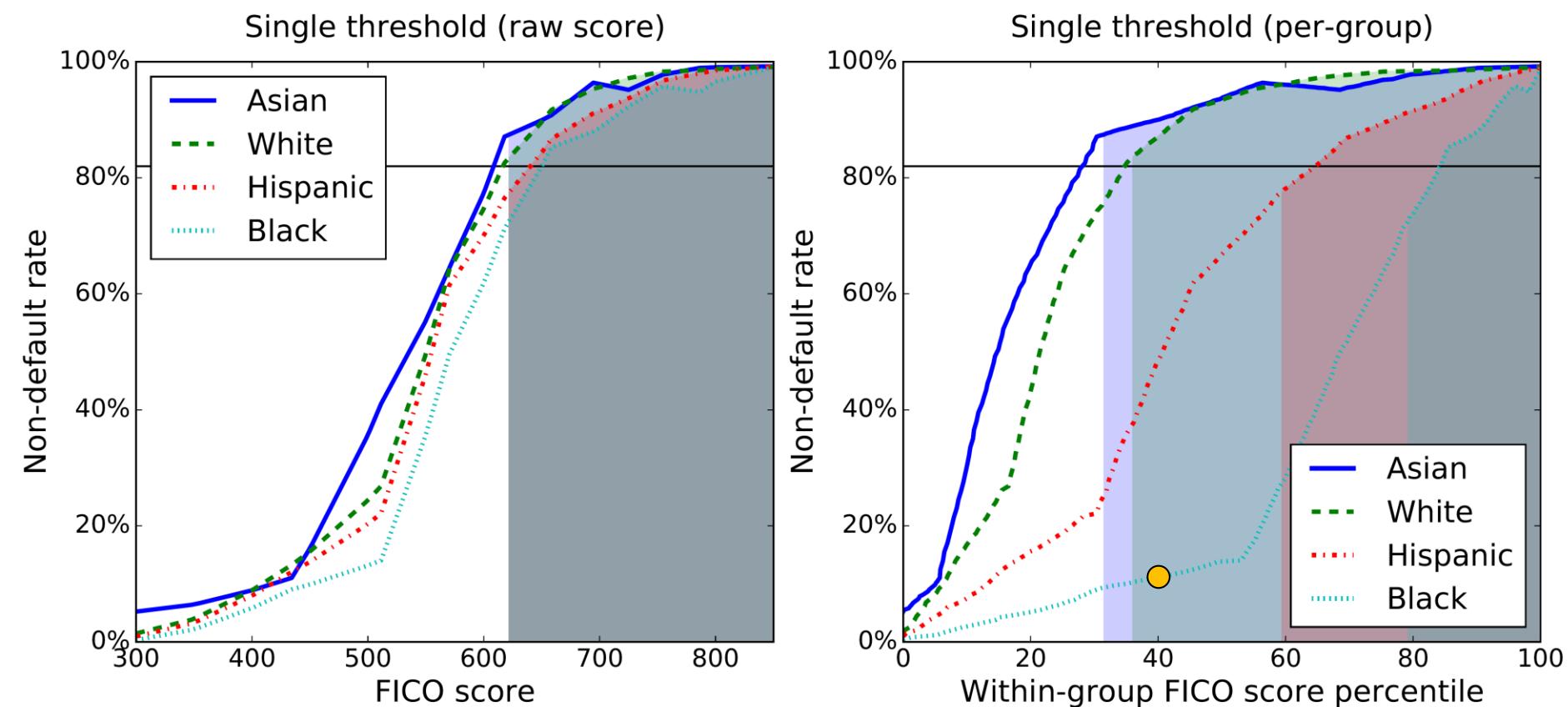


# EXAMPLE: FICO SCORES



[Hardt et al. 2016]

# EXAMPLE: FICO SCORES



[Hardt et al. 2016]

# ACHIEVING EQUALIZED ODDS

- We wish to derive a classifier  $\tilde{Y}$  from a possibly discriminatory classifier  $\hat{Y}$
- $\tilde{Y}$  is derived from  $\hat{Y}$  and  $A$  if it is a possibly randomized function of  $(\hat{Y}, A)$  alone
- $\tilde{Y}$  is completely described by four parameters in  $[0,1]$  corresponding to  $\Pr[\tilde{Y} = 1 \mid \hat{Y} = \hat{y}, A = a]$  for  $\hat{y}, a \in \{0,1\}$



# ACHIEVING EQUALIZED ODDS

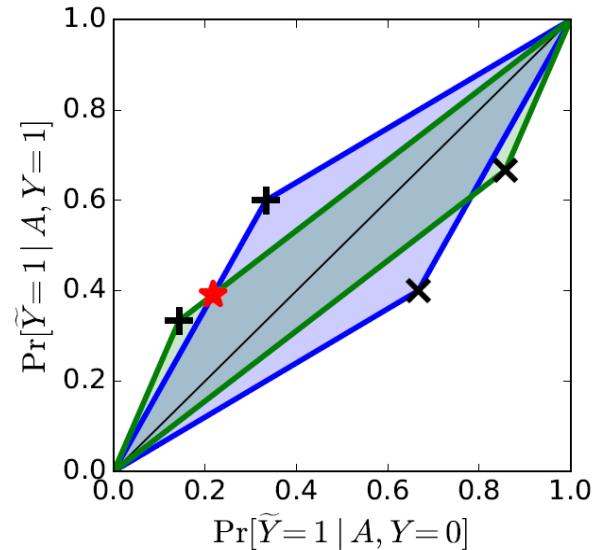
- Define  $\gamma_a(\hat{Y})$  as  
 $(\Pr[\hat{Y} = 1 | A = a, Y = 0], \Pr[\hat{Y} = 1 | A = a, Y = 1])$
- Poll 2:  $\hat{Y}$  satisfies equalized odds iff
  1.  $\gamma_0(\hat{Y}) = \gamma_1(\hat{Y})$
  2.  $\gamma_0(\hat{Y}) \leq \gamma_1(\hat{Y})$
  3.  $\gamma_0(\hat{Y}) = 1 - \gamma_1(\hat{Y})$
  4.  $\gamma_0(\hat{Y}) = 0$  and  $\gamma_1(\hat{Y}) = 1$



# ACHIEVING EQUALIZED ODDS\*

- **Lemma:**  $\tilde{Y}$  is derived iff for all  $a \in \{0,1\}$ ,  
 $\gamma_a(\tilde{Y}) \in P_a(\hat{Y})$ , where  $P_a(\hat{Y})$  is  
 $\text{conv}\{(0,0), \gamma_a(\hat{Y}), \gamma_a(1 - \hat{Y}), (1,1)\}$

- Achievable region ( $A = 0$ )
- Achievable region ( $A = 1$ )
- Overlap
- $+$  Result for  $\tilde{Y} = \hat{Y}$
- $\times$  Result for  $\tilde{Y} = 1 - \hat{Y}$
- $\star$  Equalized odds optimum



\*Just for fun

# ACHIEVING EQUALIZED ODDS\*

- Loss function  $\ell: \{0,1\}^2 \rightarrow \mathbb{R}^+$  gives the loss  $\ell(\hat{y}, y)$  of predicting  $\hat{y}$  when the label is  $y$
- The optimization problem is

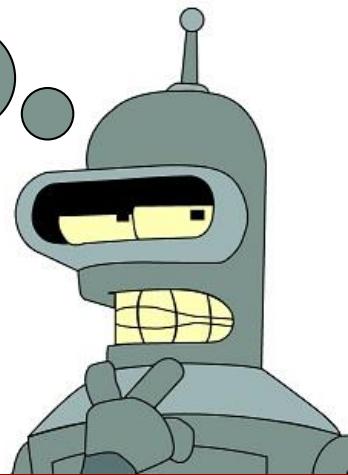
$$\begin{aligned} & \min \mathbb{E}[\ell(\tilde{Y}, Y)] \\ \text{s.t. } & \forall a \in \{0,1\}, \gamma_a(\tilde{Y}) \in P_a(\hat{Y}) \\ & \gamma_0(\tilde{Y}) = \gamma_1(\tilde{Y}) \end{aligned}$$

- Theorem: This is a linear program



\*Just for fun

The construction of  $\tilde{Y}$   
depends on the joint  
distribution of  $(\hat{Y}, A, Y)$   
at training time, but at  
prediction time we only  
have access to  $(\hat{Y}, A)$



\*Just for fun

# SUMMARY

- Definitions
  - Lipschitz property
  - Statistical parity
  - Equalized odds
- Algorithms:
  - LP for Lipschitz classifiers
  - LP for deriving a classifier satisfying equalized odds

