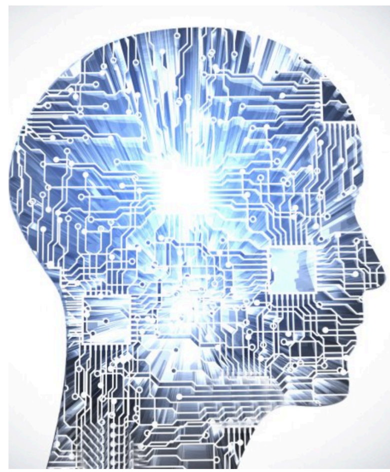
Bias and Fairness in AI/ML models

Swati Gupta
Assistant Professor
School of Industrial and Systems Engineering,
Georgia Institute of Technology

October 25, 2018

Digital Data Flows Master Class: Emerging Technologies

Machine Learning Pipeline



Data → **Machine Learning/AI** → **Data driven decisions**

What is the effect of these decisions on human well-being?

What is Bias/Fairness?

nature.com : Sitemap

nature International weekly journal of science

Home | News & Comment | Research | Careers & Jobs | Current Issue | Archive | Audio & Video | For Aut

Archive > Volume 551 > Issue 7679 > Comment > Article

NATURE | COMMENT

Four ethical priorities for neurotechnologies and AI

Rafael Yuste, Sara Goering, Blaise Agüera y Arcas, Guoqiang Bi, Jose M. Carmena, Adrian Carter, Joseph J. Fins, Phoebe Friesen, Jack Gallant, Jane E. Huggins, Judy Illes, Philipp Kellmeyer, Eran Klein, Adam Marblestone, Christine Mitchell, Erik Parens, Michelle Pham, Alan Rubel, Norihiro Sadato, Laura Specker Sullivan, Mina Teicher, David Wasserman, Anna Wexler, Meredith Whittaker & Jonathan Wolpaw

08 November 2017

What is Bias/Fairness?

“**Bias.** When scientific or **technological decisions** are based on a narrow set of systemic, structural or **social concepts and norms**, the resulting technology can **privilege certain groups** and harm others.” – Nature comment

Amazon to Bring Same-Day Delivery to Roxbury After Outcry

by **Spencer Soper**
April 26, 2016, 5:19 PM EDT *Updated on* April 26, 2016, 8:22 PM EDT

CORNELL CHRONICLE

pics | Campus & Community | All Stories | In the News | Expert Quotes | Ezra Magazine

Rating systems may discriminate against Uber drivers

By Leslie Morris | December 15, 2016

Amazon to Bring Same-Day Delivery to Roxbury After Outcry

by **Spencer Soper**
April 26, 2016, 5:19 PM EDT *Updated on* April 26, 2016, 8:22 PM EDT

CORNELL CHRONICLE

pics Campus & Community All

Rating systems may d

By

PROPUBLICA TOPICS ▾ SERIES ▾ NEWS APPS GET INVOLVED IMPACT ABOUT

MACHINE BIAS

Facebook Lets Advertisers Exclude Users by Race

Facebook's system allows advertisers to exclude black, Hispanic, and other "ethnic affinities" from seeing ads.

by **Julia Angwin** and **Terry Parris Jr.**, Oct. 28, 2016, 1 p.m. EDT

Amazon to Bring Same-Day

THE WALL STREET JOURNAL.

Subscribe No

$1 FOR 2 MO

Home World U.S. Politics Economy Business Tech Markets Opinion Life & Arts Real Estate WSJ. Magazine

Dyn Says Cyberattack Has Ended, Investigation Continues

Visa Taps Blockchain for Cross-Border Payment Plan

Airbnb Revises New York Rules Amid Possible Legislation

Russian Hacker Suspected of LinkedIn Attack Indicted in U.S.

Settl Mobi Plans

DIGITS

Google Mistakenly Tags Black People as 'Gorillas,' Showing Limits of Algorithms

The Marshall Project Nonprofit journalism about criminal justice

SEARCH ABOUT SUPPO

JUSTICE TALK

What You Need To Know About Predictive Policing

Key background reading before our discussion on predictive policing on Wednesday, February 24th.

Amazon to Bring Same-Day

THE WALL STREET JOURNAL.

Subscribe No

$1 FOR 2 MO

Home World U.S. Politics Economy Business Tech Markets Opinion Life & Arts Real Estate WSJ. Magazine

Dyn Sa
Cyberattac
Ended, Inve
Continues

DIGITS

Google M
Algorith

DE GRUYTER OPEN Proceedings on Privacy Enhancing Technologies 2015; 2015 (1):92–112

Amit Datta*, Michael Carl Tschantz, and Anupam Datta

Automated Experiments on Ad Privacy Settings

A Tale of Opacity, Choice, and Discrimination

Abstract: To partly address people's concerns over web tracking, Google has created the Ad Settings webpage to provide information about and some choice over the profiles Google creates on users. We present AdFisher, an automated tool that explores how user behaviors, Google's ads, and Ad Settings interact. AdFisher can run browser-based experiments and analyze data using machine learning and significance tests. Our tool uses a rigorous experimental design and statistical analysis to

serious privacy concern. Colossal amounts of collected data are used, sold, and resold for serving targeted content, notably advertisements, on websites (e.g., [1]). Many websites providing content, such as news, outsource their advertising operations to large third-party ad networks, such as Google's DoubleClick. These networks embed tracking code into webpages across many sites providing the network with a more global view of each user's behaviors.

The Marshall Proje

SUPPO

JUSTICE TALK

What Policing

Key background reading before our discussion on predictive policing on Wednesday, February 24th.

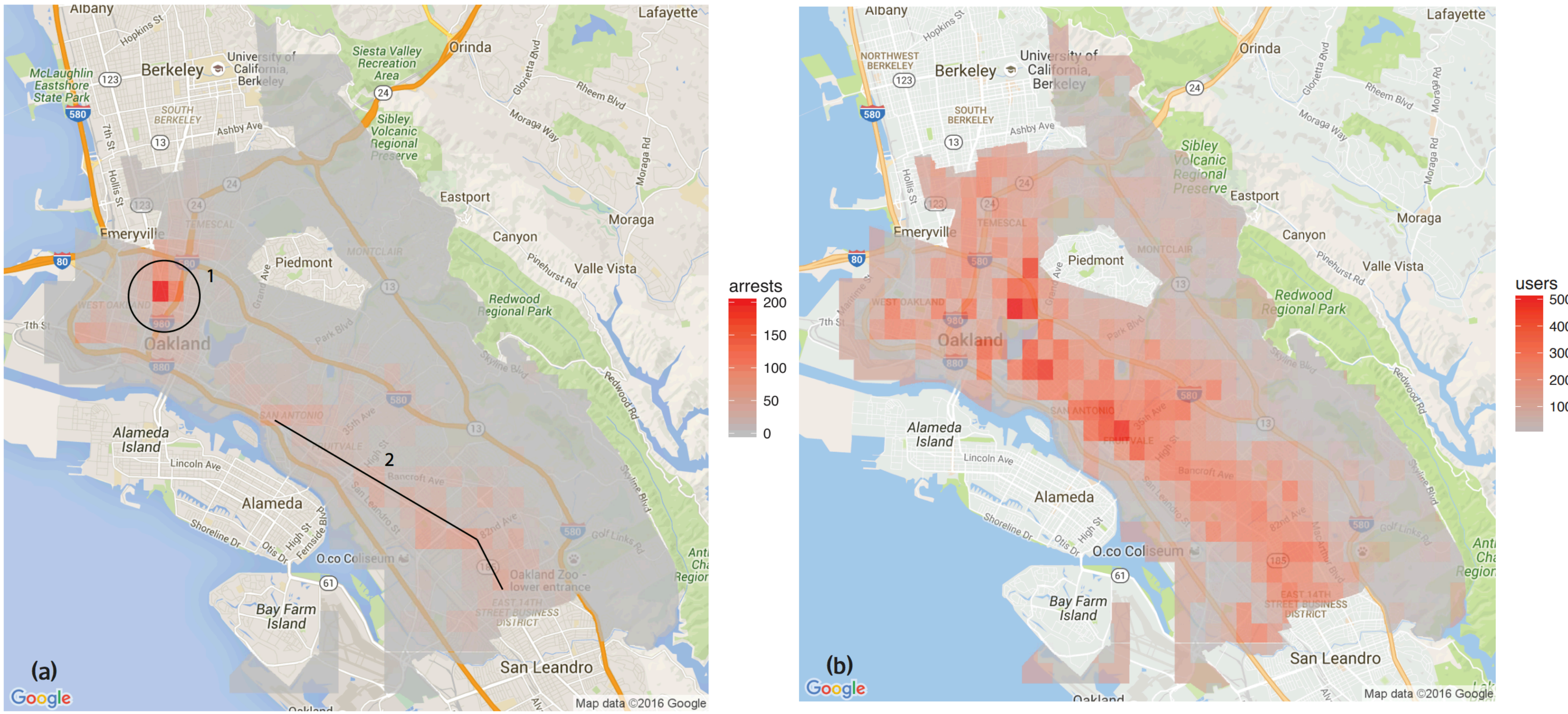
Outline of the talk

- **Bias in the data, models and variables**
- **Fairness Metrics**
  - Statistical measures
  - Equity measures
- **Trolley Problem of Choice**

Predictive Policing

[Lum, Isaac, 2016]

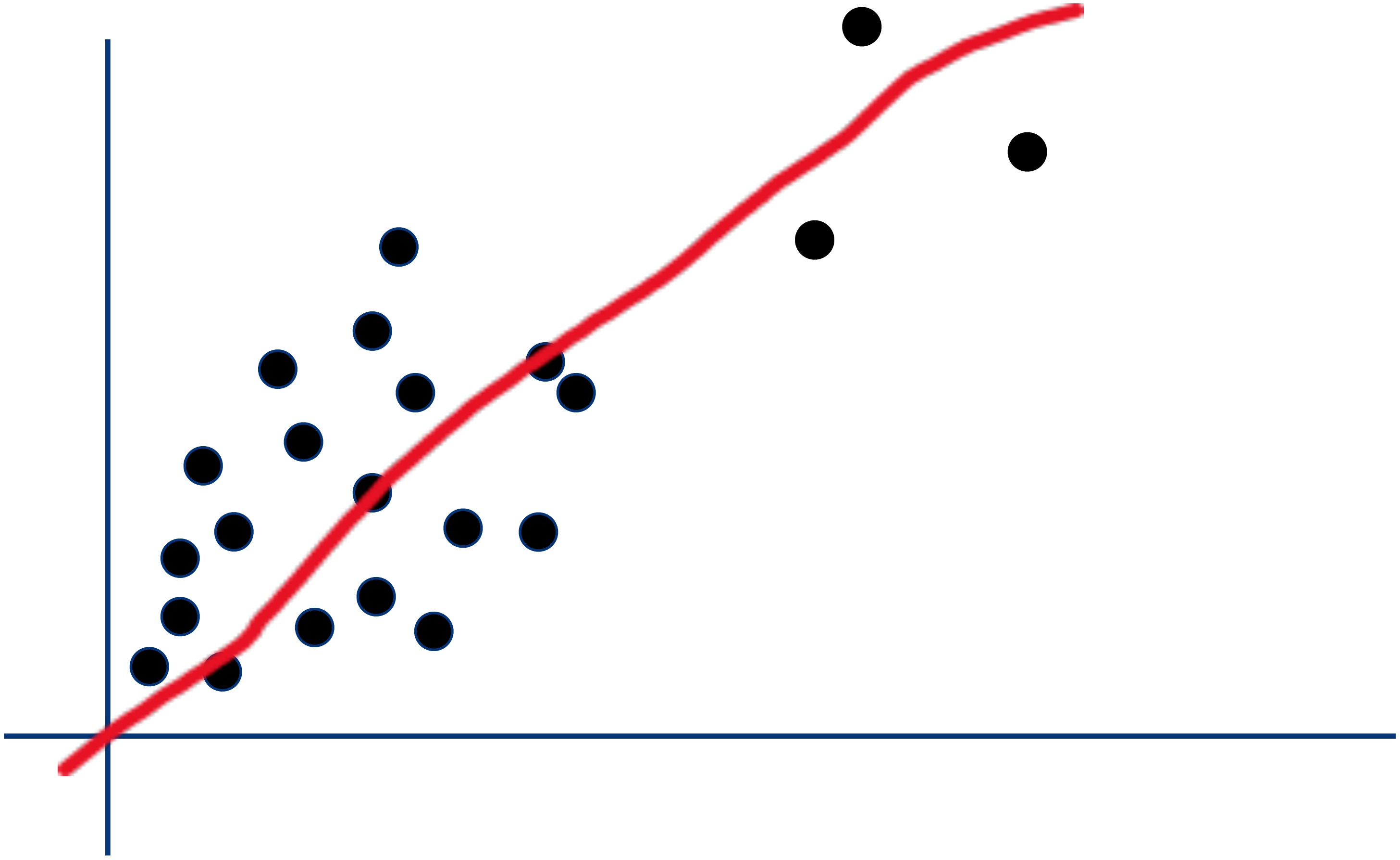
“application of analytical techniques to identify likely targets for police intervention and prevent crime or solve past crimes by making statistical predictions”



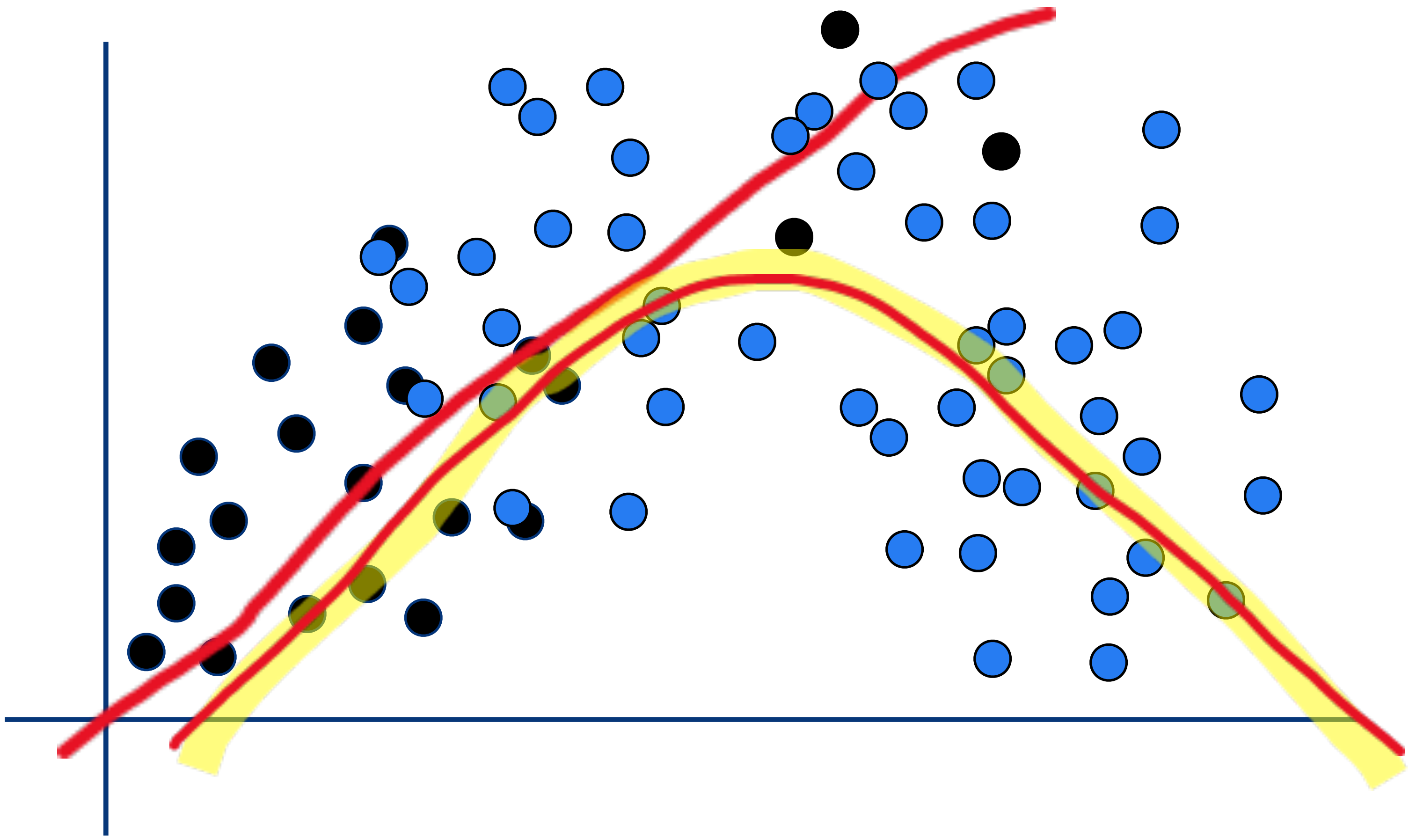
Heat map of drug arrests made

FIGURE 1 (a) Number of drug arrests made by Oakland police department, 2010. (1) West Oakland, (2) International Boulevard. (b) Estimated number of drug users, based on 2011 National Survey on Drug Use and Health

ML finds patterns in data

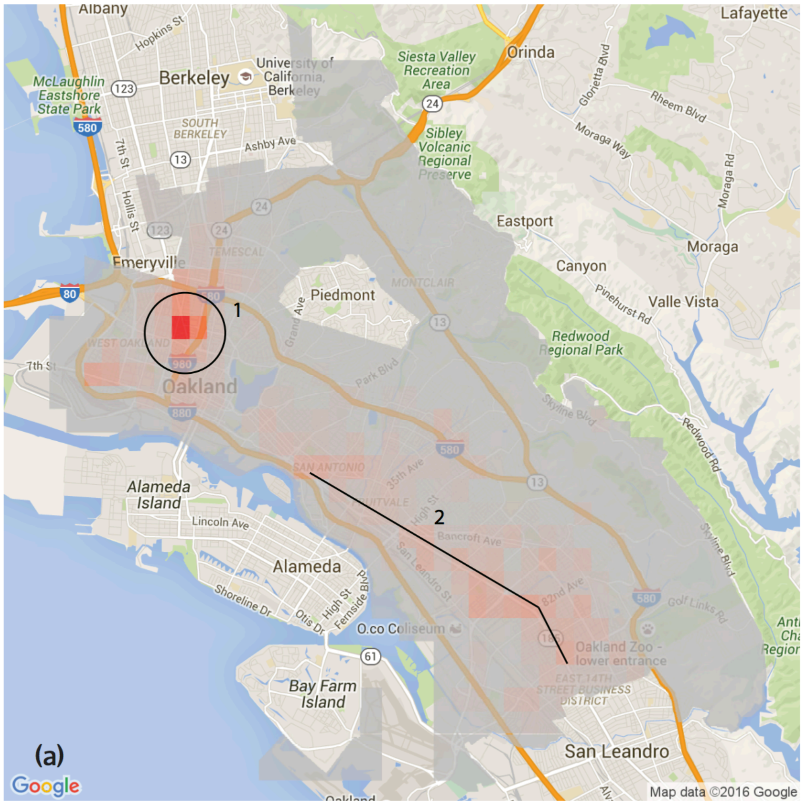


ML finds patterns in data



PredPol: crime type, time, loc

[Kristian Lum, William Isaac, 2016]



Heat map of drug arrests made

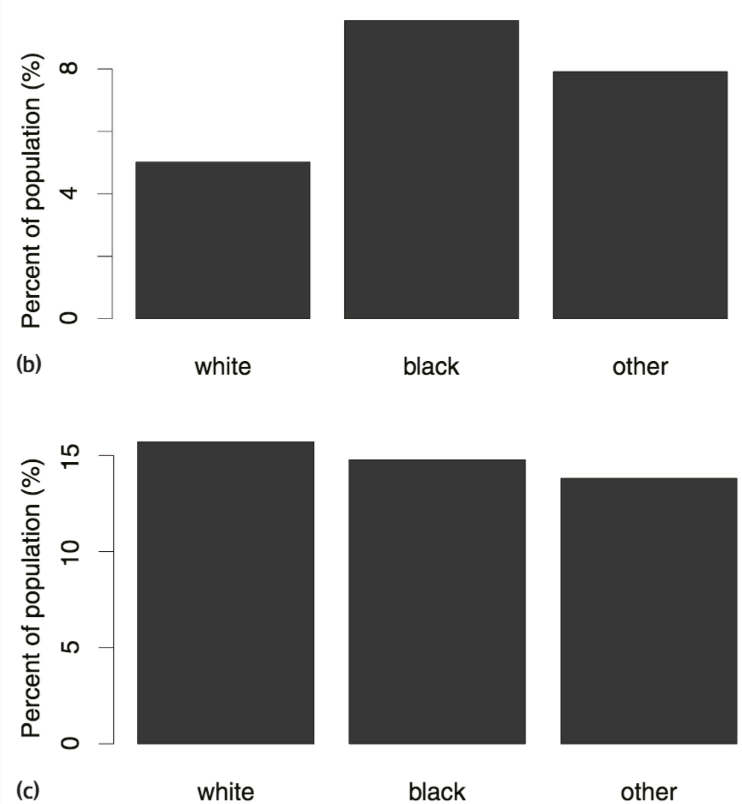
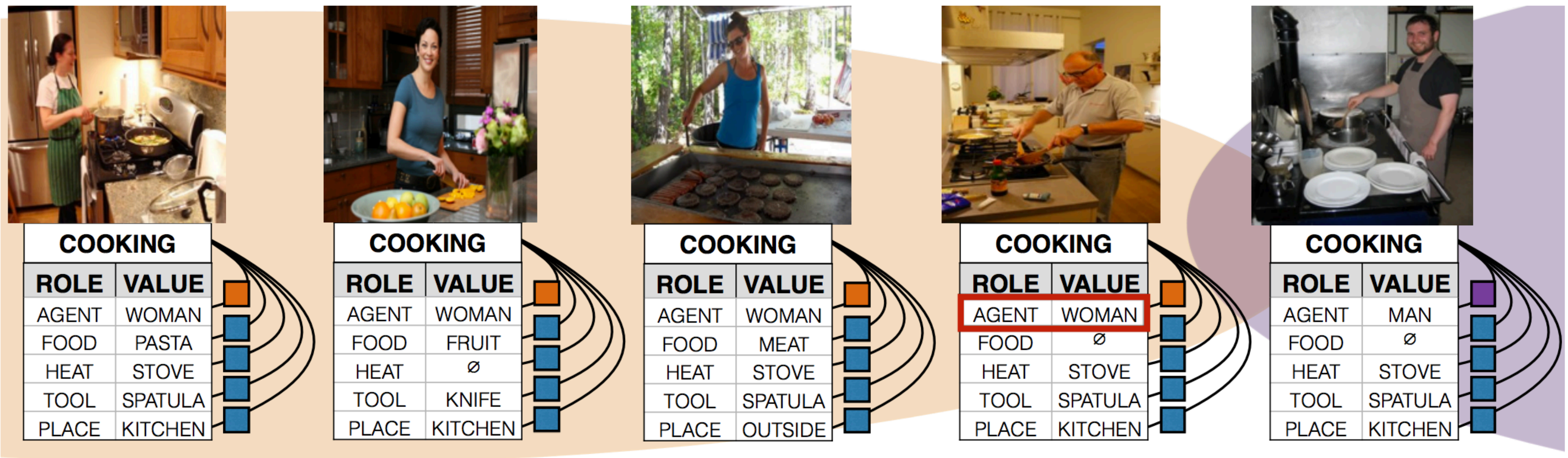


FIGURE 2 (a) Number of days with targeted policing for drug crimes in areas flagged by PredPol analysis of Oakland police data. (b) Targeted policing for drug crimes, by race. (c) Estimated drug use by race

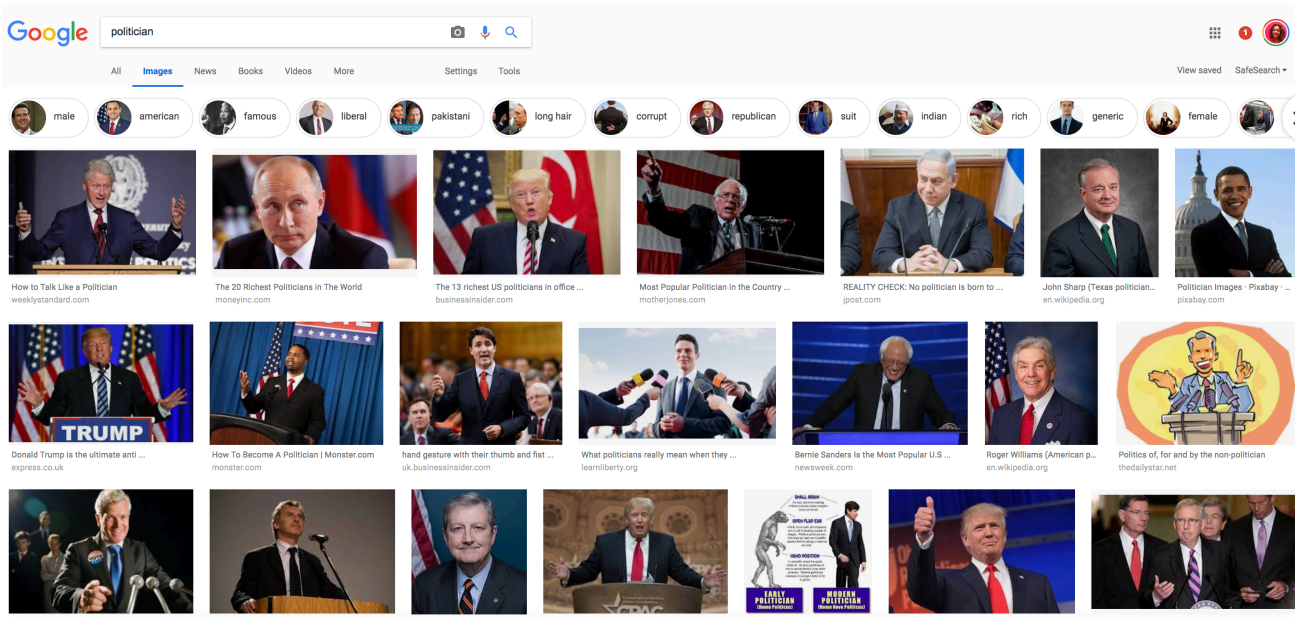
Not just about collection

We live in a biased society, so it's inevitable that data collected about that society will be biased: inherent bias, test data, feedback, proxies..



Not just about collection

We live in a biased society, so it's inevitable that data collected about that society will be biased: inherent bias, test data, feedback, proxies..



Not just about collection

We live in a biased society, so it's inevitable that data collected about that society will be biased: inherent bias, test data, feedback, proxies..

DE GRUYTER OPEN   Proceedings on Privacy Enhancing Technologies 2014; 1 (11):1–21

Amit Datta*, Michael Carl Tschantz, and Anupam Datta

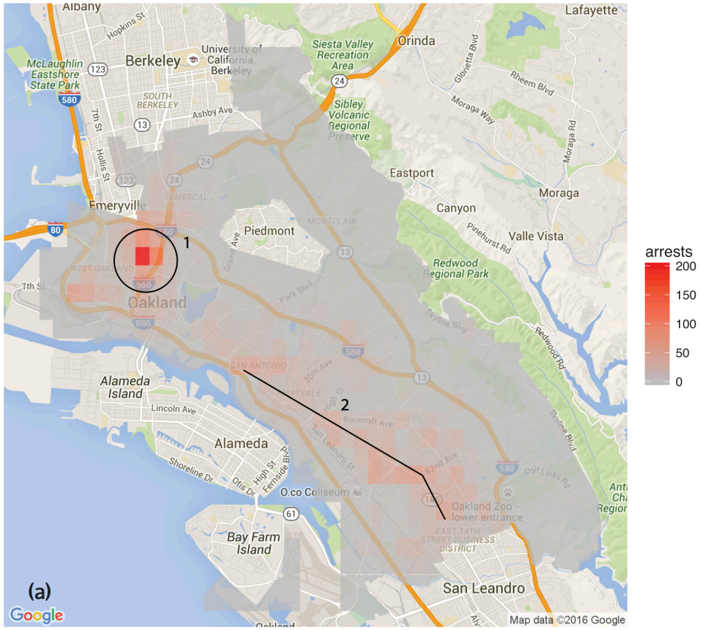
Automated Experiments on Ad Privacy Settings

A Tale of Opacity, Choice, and Discrimination

"We also found that setting the gender to female resulted in getting fewer instances of an ad related to high paying jobs than setting it to male. "

Not just about collection

We live in a biased society, so it's inevitable that data collected about that society will be biased: inherent bias, test data, feedback, proxies..



Proxies:
predicting crime using data **on arrests**,
Not on **incidence of crime**.

We do not want such biases to propagate into systems that make life-changing decisions.

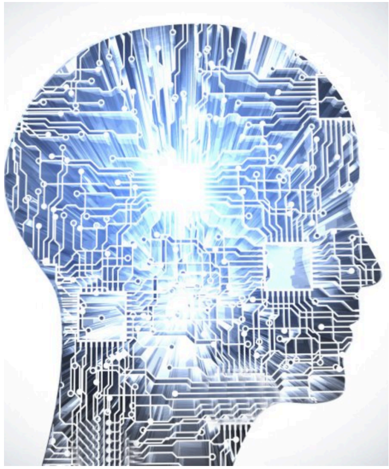
Outline of the talk

- **Bias in the data, models and variables**
- **Fairness Metrics**
  - Statistical measures
  - Equity measures
- **Trolley Problem of Choice**

Machine Learning Pipeline

Data

→



Machine Learning/AI

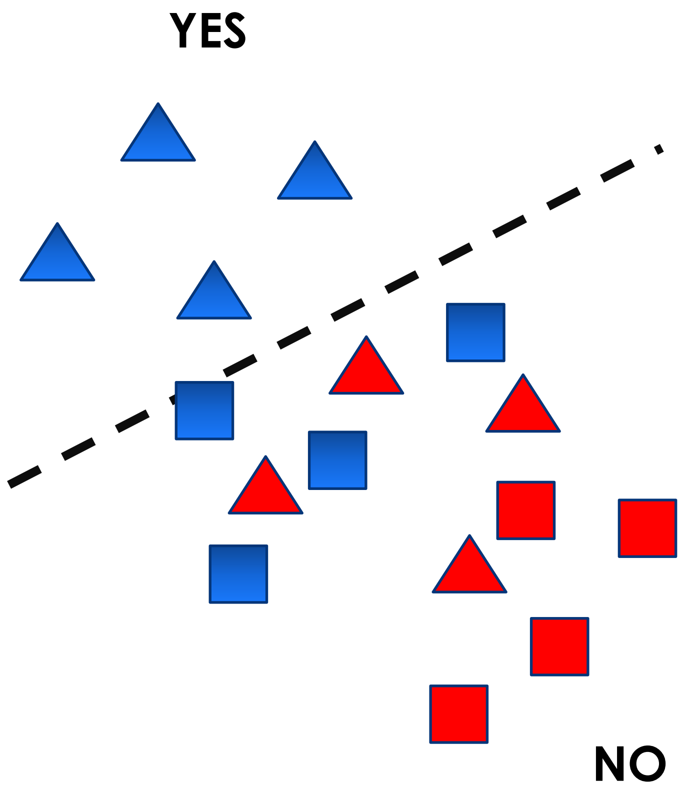
→

Data driven decisions

What is the effect of these decisions on human well-being?

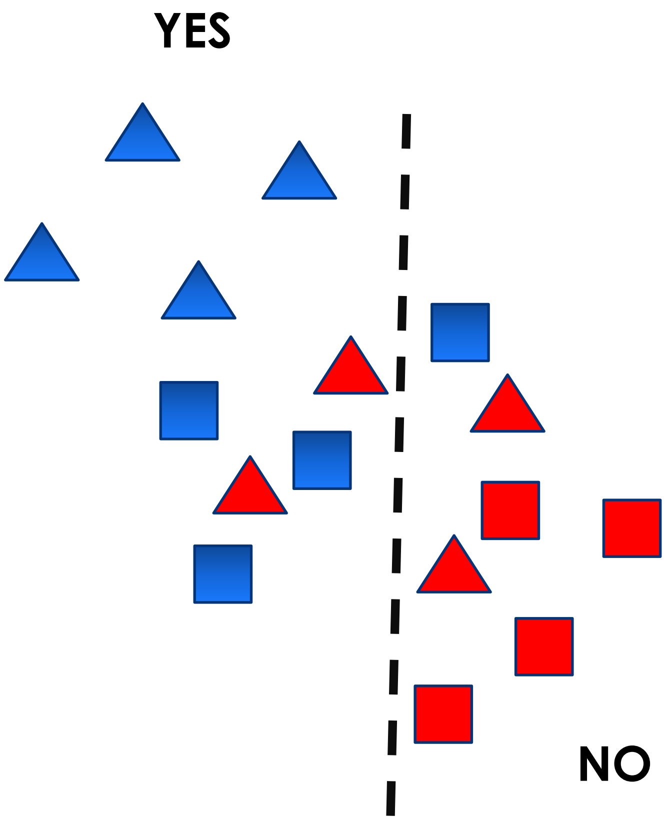
Classification

Hired for job or not, will re-offend or not (prison), given a loan or not.



Statistical Definitions of Fairness

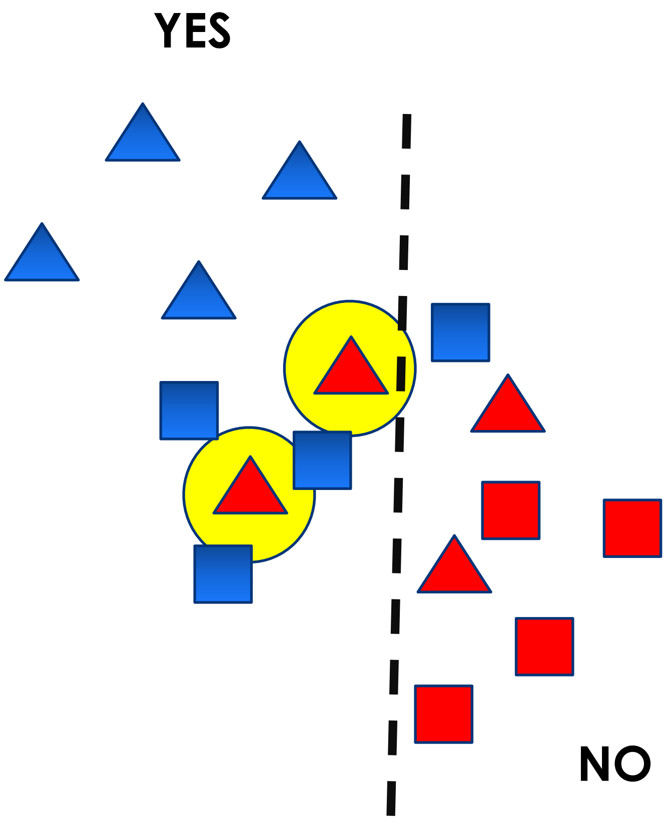
Hired for job or not, **will re-offend or not (prison)**, given a loan or not.



Is it **fair** to achieve highest accuracy in classification?

Statistical Definitions of Fairness

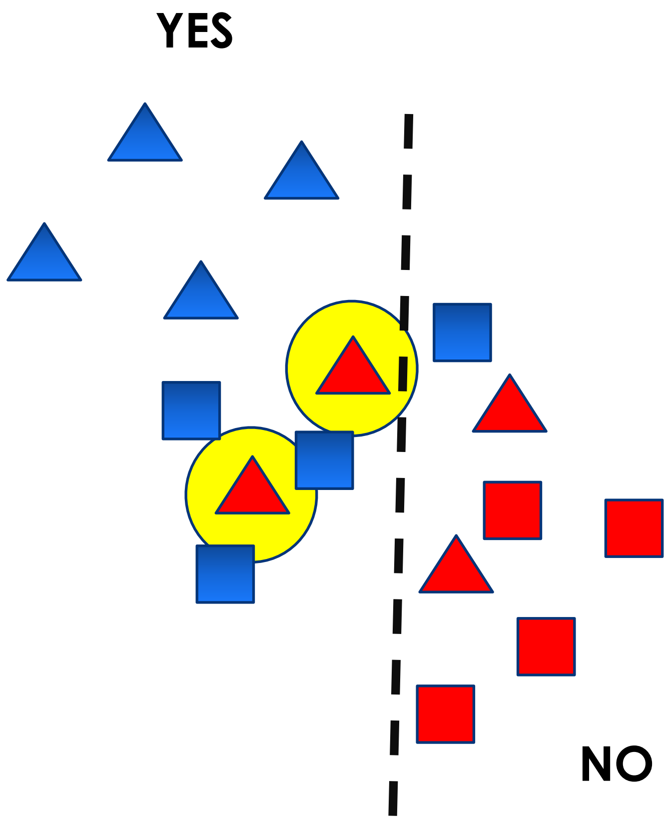
Hired for job or not, **will re-offend or not (prison),** given a loan or not.

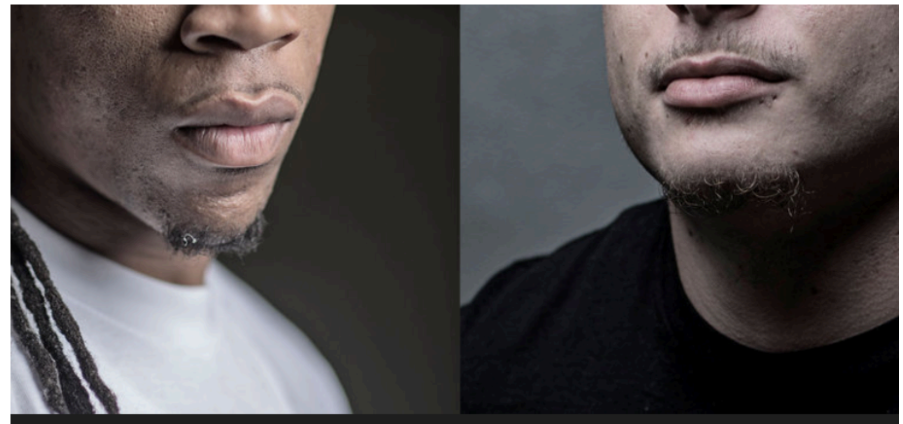


COMPAS Risk Score: ProPublica

Statistical Definitions of Fairness

Hired for job or not, **will re-offend or not (prison),** given a loan or not.





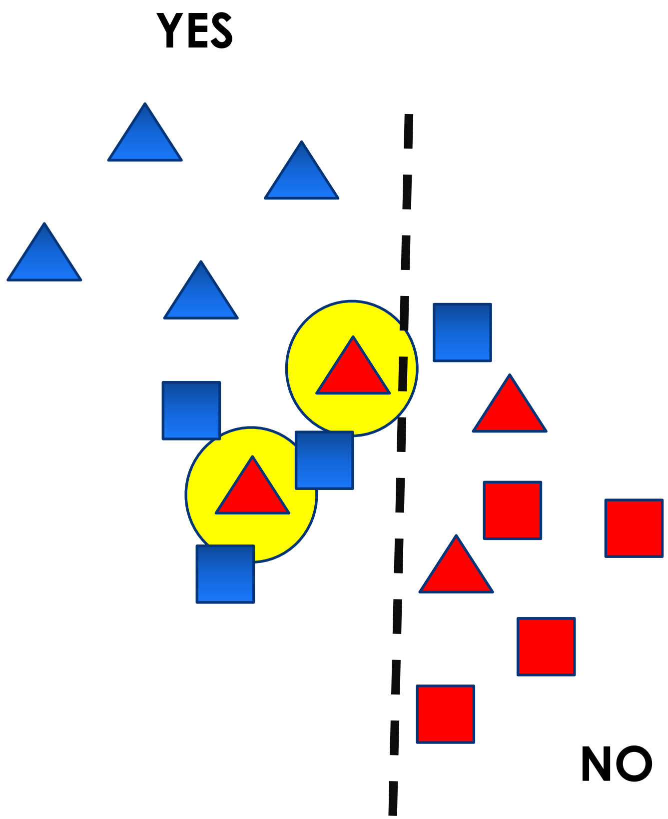
Prediction Fails Differently for Black Defendants

	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

Overall, Northpointe's assessment tool correctly predicts recidivism 61 percent of the time. But blacks are almost twice as likely as whites to be labeled a higher risk but not actually re-offend. It makes the opposite mistake among whites: They are much more likely than blacks to be labeled lower risk but go on to commit other crimes. (Source: ProPublica analysis of data from Broward County, Fla.)

Statistical Definitions of Fairness

Hired for job or not, **will re-offend or not (prison),** given a loan or not.

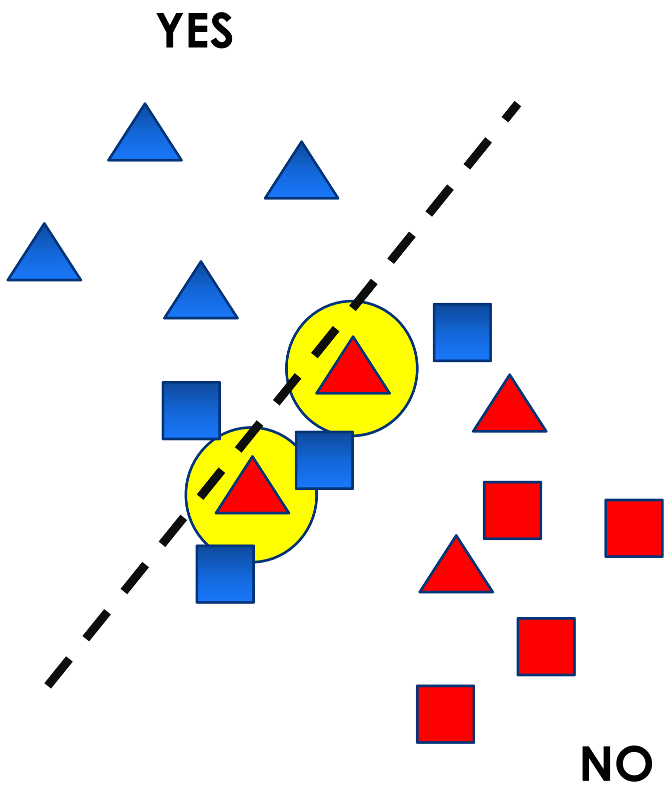


Is it **fair** to achieve highest accuracy in classification?

Or is it **fair** to balance **false positives** across the groups?

Statistical Definitions of Fairness

Hired for job or not, **will re-offend or not (prison),** given a loan or not.

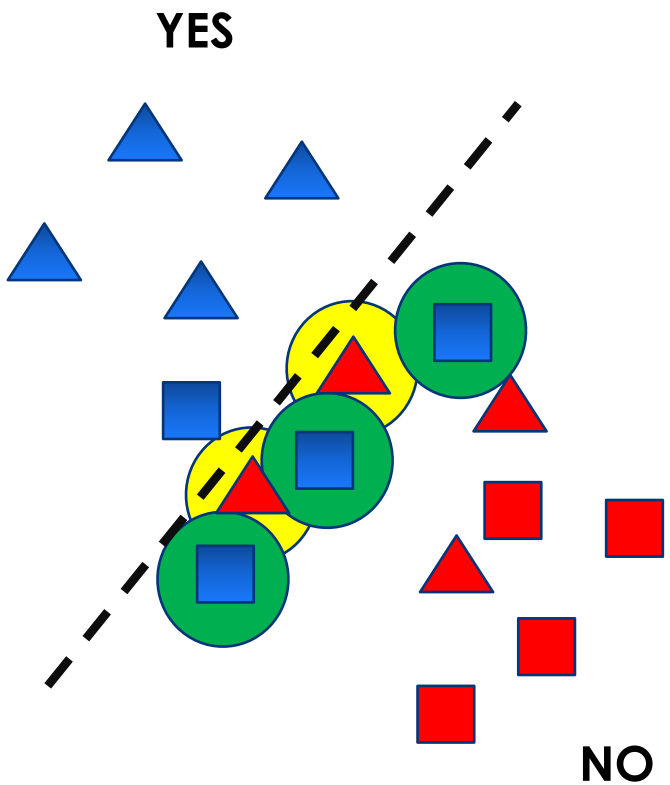


Is it **fair** to achieve highest accuracy in classification?

Or is it **fair** to balance **false positives** across the groups?

Statistical Definitions of Fairness

Hired for job or not, **will re-offend or not (prison)**, given a loan or not.



Is it **fair** to achieve highest accuracy in classification?

Or is it **fair** to balance **false positives** across the groups?

False negatives?

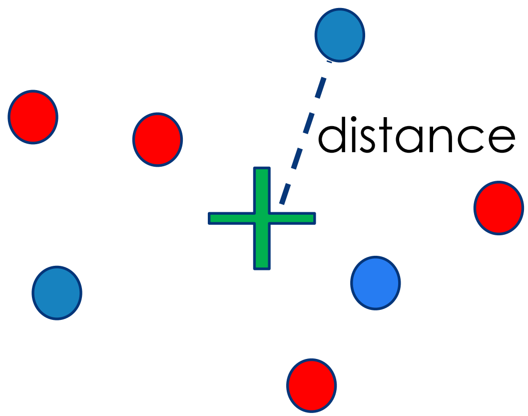
Statistical Definitions of Fairness

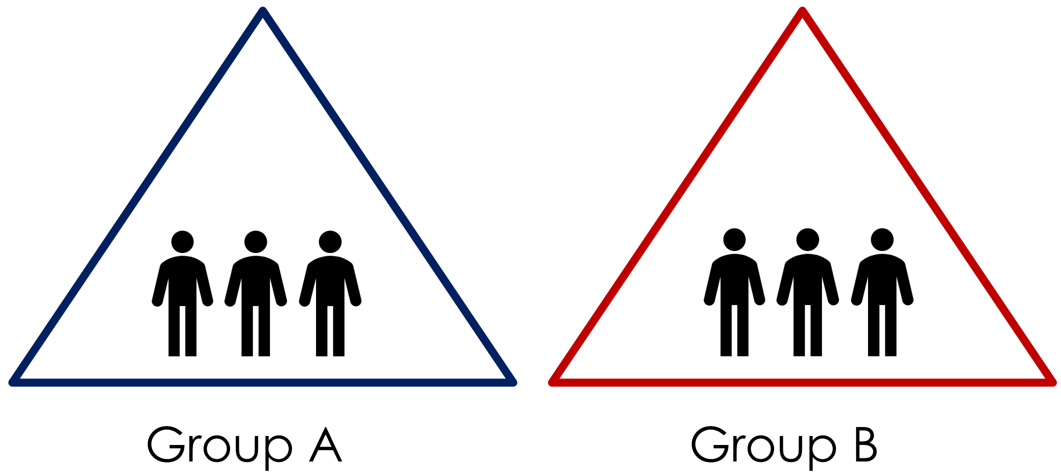
		True condition				
	Total population	Condition positive	Condition negative	Prevalence $= \frac{\Sigma \text{ Condition positive}}{\Sigma \text{ Total population}}$	Accuracy (ACC) = $\frac{\Sigma \text{ True positive} + \Sigma \text{ True negative}}{\Sigma \text{ Total population}}$	
Predicted condition	Predicted condition positive	**True positive**, Power	**False positive**, Type I error	Positive predictive value (PPV), Precision = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Predicted condition positive}}$	False discovery rate (FDR) = $\frac{\Sigma \text{ False positive}}{\Sigma \text{ Predicted condition positive}}$	
	Predicted condition negative	**False negative**, Type II error	**True negative**	False omission rate (FOR) = $\frac{\Sigma \text{ False negative}}{\Sigma \text{ Predicted condition negative}}$	Negative predictive value (NPV) = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Predicted condition negative}}$	
		True positive rate (TPR), Recall, Sensitivity, probability of detection $= \frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$	False positive rate (FPR), Fall-out, probability of false alarm $= \frac{\Sigma \text{ False positive}}{\Sigma \text{ Condition negative}}$	Positive likelihood ratio (LR+) $= \frac{\text{TPR}}{\text{FPR}}$	Diagnostic odds ratio (DOR) $= \frac{\text{LR+}}{\text{LR−}}$	F_1 score = $\frac{2}{\frac{1}{\text{Recall}} + \frac{1}{\text{Precision}}}$
		False negative rate (FNR), Miss rate $= \frac{\Sigma \text{ False negative}}{\Sigma \text{ Condition positive}}$	Specificity (SPC), Selectivity, True negative rate (TNR) $= \frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$	Negative likelihood ratio (LR−) $= \frac{\text{FNR}}{\text{TNR}}$		

In fact,
different stakeholders might have different points of view

Equity Metrics of Fairness

What about general decisions: *how much loan to give*? ***where to place an emergency room***? *Where to schedule deliveries*?

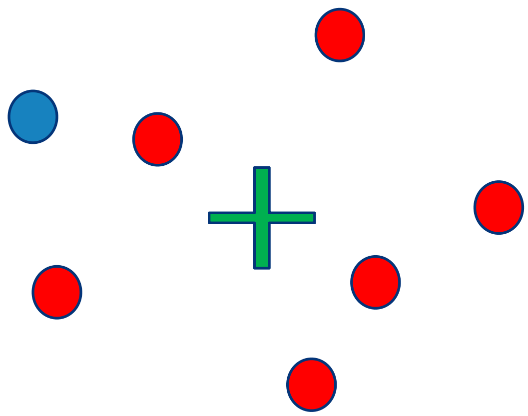


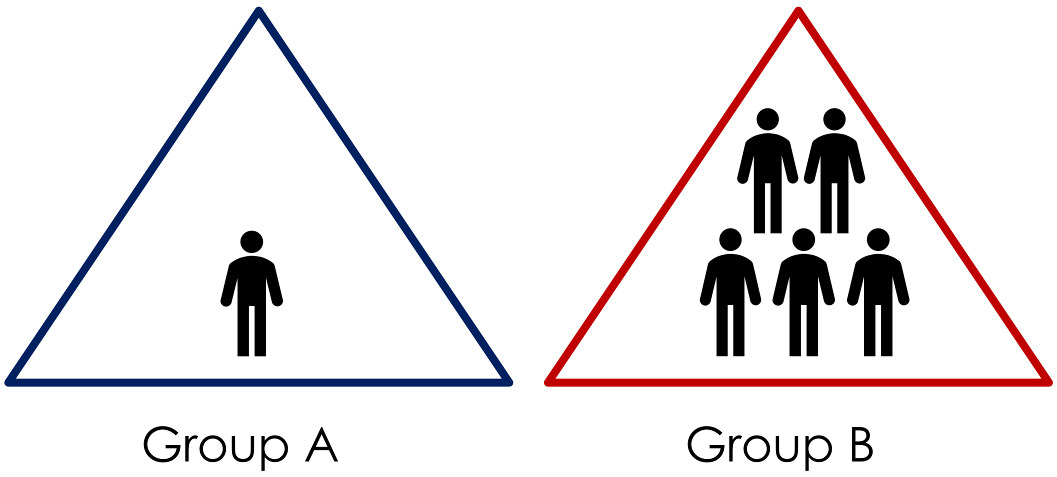


Is it **fair** to minimize **total distance** travelled by any group?

Equity Metrics of Fairness

What about general decisions: *how much loan to give*? ***where to place an emergency room***? *Where to schedule deliveries*?

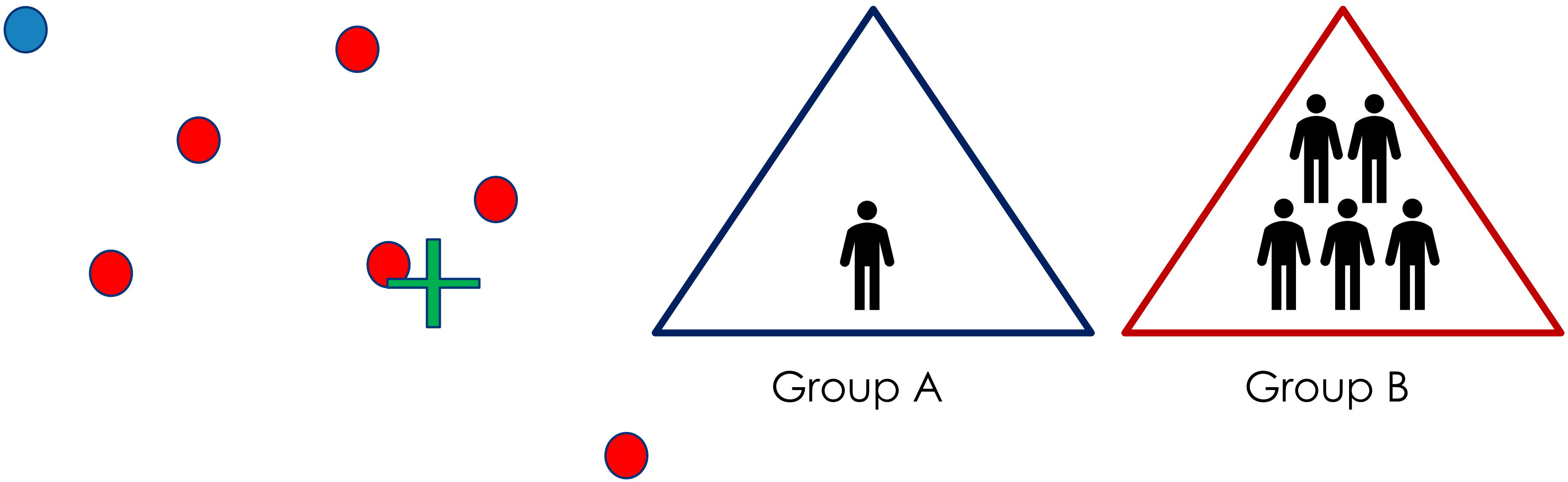




Is it **fair** to minimize **total distance** travelled by any group?

Equity Metrics of Fairness

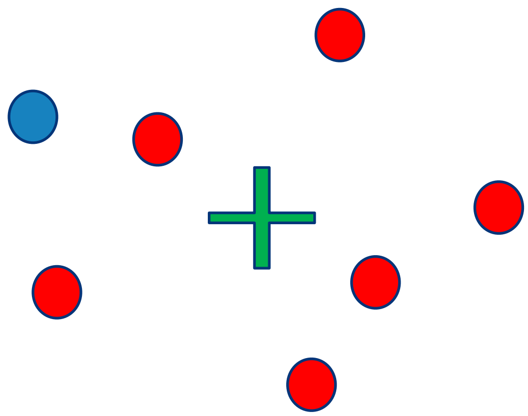
What about general decisions: *how much loan to give*? ***where to place an emergency room***? *Where to schedule deliveries*?

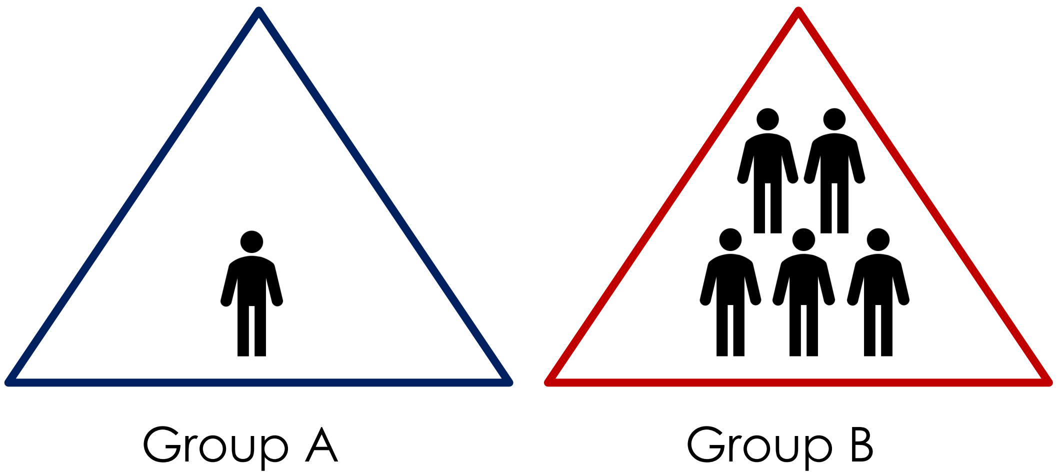


Is it **fair** to minimize **total distance** travelled by any group?

Equity Metrics of Fairness

What about general decisions: *how much loan to give*? ***where to place an emergency room***? *Where to schedule deliveries*?

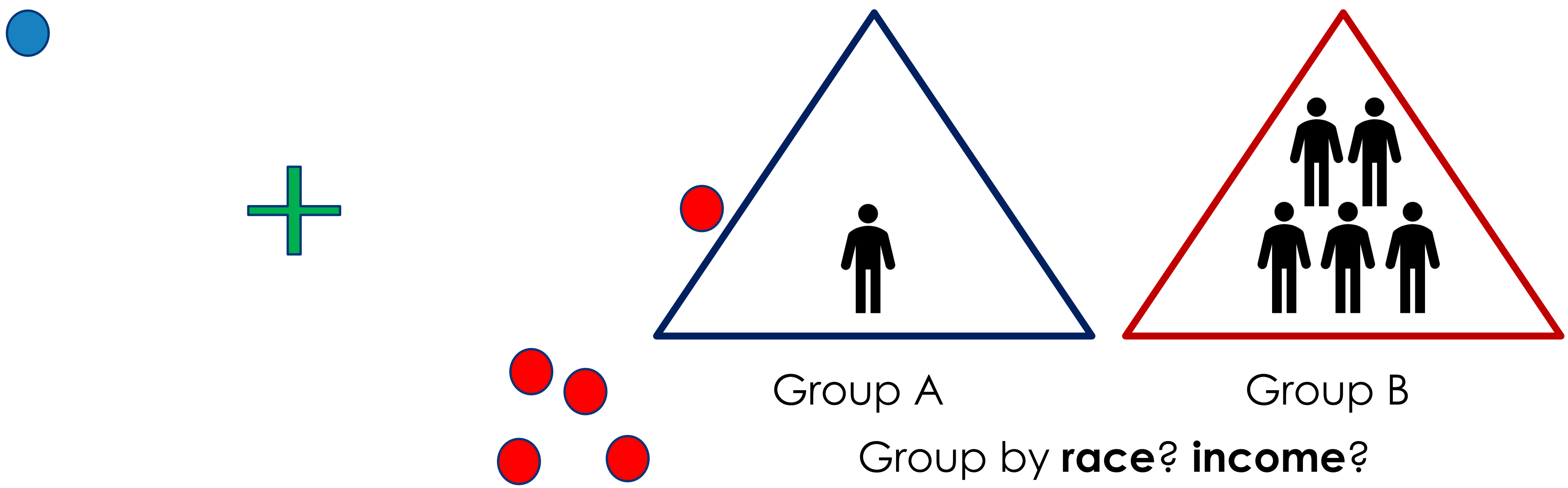




Is it **fair** to minimize **average distance** travelled by any group (per person)?

Equity Metrics of Fairness

What about general decisions: *how much loan to give*? ***where to place an emergency room***? *Where to schedule deliveries?*



Group A

Group B

Group by **race**? **income**? **insurance**?

Is it **fair** to minimize **average distance** travelled by any group (per person)?

Equity Metrics of Fairness

Table 3a
Framework for equity measures

Scaling	Reference distribution: Peer	Mean	Attribute
None	$\lvert E_i - E_h \rvert^p$	$\lvert E_i - \bar{E} \rvert^p$	$\lvert E_i - A_i \rvert^p$
	(4) $\sum_h \sum_i \lvert E_i - E_h \rvert$	(2) $\sum_i (E_i - \bar{E})^2$	(1) $\max_i E_i$
	(6) $\max_{i,h} \lvert E_i - E_h \rvert$	(3) $\sum_i \lvert E_i - \bar{E} \rvert$	(14) $\sum_i \lvert E_i - A_i \rvert$
	(7) $\max_i E_i - \min_i E_i$	(8) $\max_i \lvert E_i - \bar{E} \rvert$	
	(9) $\max_i \sum_j \lvert E_i - E_j \rvert$	(20) $\frac{1}{N} \sum_i (\log E_i - \log \bar{E})^2$	
	(10) $\sum_i \max_j \lvert E_i - E_j \rvert$		

Table 3b
Framework for equity measures

Scaling	Reference distribution: Peer	Mean	Attribute
Normalized	$\frac{\lvert E_i - E_h \rvert^p}{\bar{E}}$	$\frac{\lvert E_i - \bar{E} \rvert^p}{\bar{E}}$	$\left\lvert \frac{E_i}{\bar{E}} - \frac{A_i}{\bar{A}} \right\rvert^p$
	(5) $\frac{\sum_i \sum_h \lvert E_i - E_h \rvert}{2N^2 \bar{E}}$	(17) $\frac{\sqrt{\sum (E_i - \bar{E})^2}}{\bar{E}}$	(11) $\frac{1}{N} \sum_i \left\lvert \frac{E_i}{\bar{E}} - \frac{A_i}{\bar{A}} \right\rvert$
		(18) $\frac{\sum_i \lvert E_i - \bar{E} \rvert}{2N\bar{E}}$	(12) $\sqrt{\frac{1}{N} \sum_i \left\lvert \frac{E_i}{\bar{E}} - \frac{A_i}{\bar{A}} \right\rvert^2}$
		(19) $\frac{1}{N} \frac{\sum_i \lvert E_i \log E_i - \bar{E} \log \bar{E} \rvert}{\bar{E}}$	

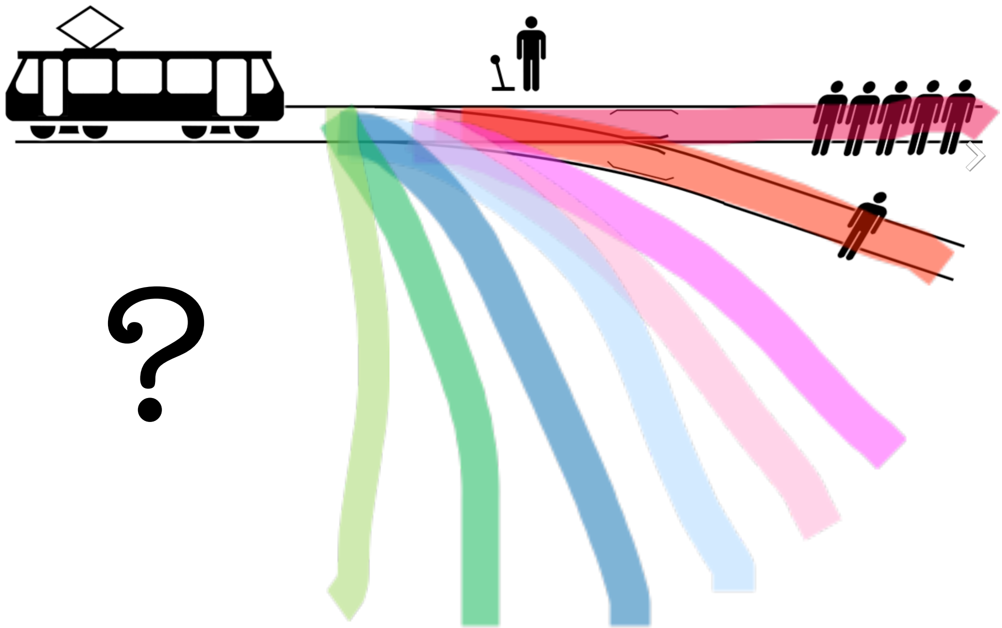
Table 3c
Framework for equity measures

Scaling	Reference distribution: Peer	Mean	Attribute
General	$\left\lvert \frac{E_i}{A_i} - \frac{E_h}{A_h} \right\rvert^p$	$\left\lvert \frac{E_i}{A_i} - \frac{\bar{E}}{\bar{A}} \right\rvert^p$	$\left\lvert \frac{E_i - A_i}{A_i} \right\rvert^p$
	(15) $\sum_i \sum_h \left\lvert \frac{E_i}{A_i} - \frac{E_h}{A_h} \right\rvert$	(13) $\sum_i \left[\frac{E_i}{A_i} - \frac{\bar{E}}{\bar{A}} \right]^2$	(16) $\sum_i \left\lvert \frac{E_i - A_i}{A_i} \right\rvert$

Outline of the talk

- **Bias in the data, models and variables**
- **Fairness Metrics**
  - Statistical measures
  - Equity measures
- **Trolley Problem of Choice**

Which **fairness** do we want?

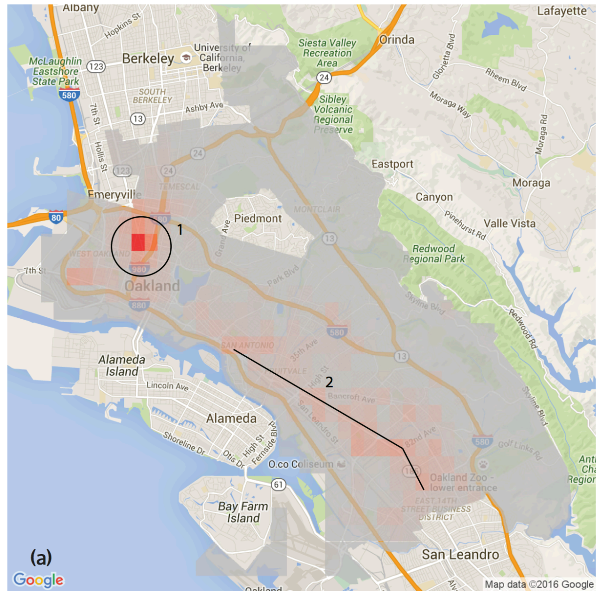


At least 50 ways to be fair

This has to be a **collective decision** we need to **consciously** reach at, after a **deeper dive** into the **application**.

Which **fairness** do we want?

Social Scientist:
Arrest data is not a good proxy for crime data



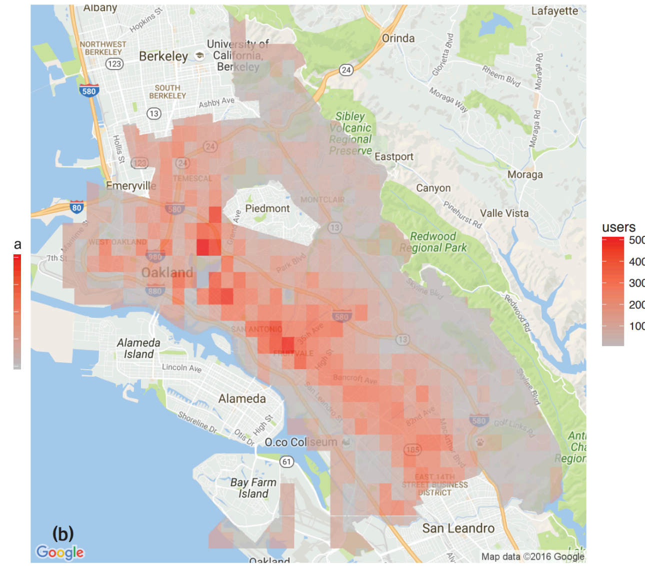


FIGURE 1 (a) Number of drug arrests made by Oakland police department, 2010. (1) West Oakland, (2) International Boulevard. (b) Estimated number of drug users, based on 2011 National Survey on Drug Use and Health

This has to be a **collective decision** we need to **consciously** reach at, after a **deeper dive** into the **application**.

Which **fairness** do we want?

Lawyer/Policy maker: *Cannot use protected classes for making decisions.*

Race (Civil Rights Act of 1964), **Color** (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)

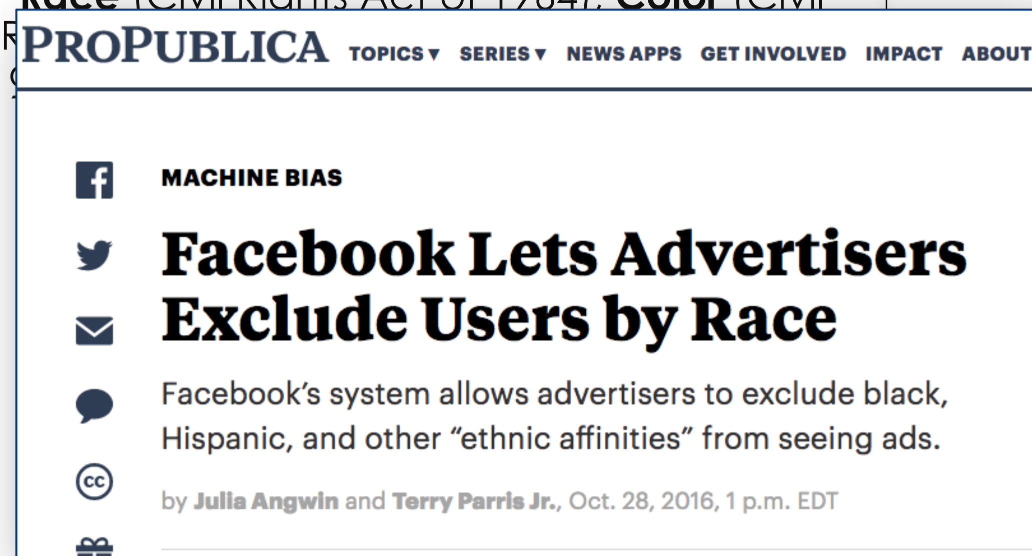
Disparate Treatment v/s Impact

This has to be a **collective decision** we need to **consciously** reach at, after a **deeper dive** into the **application**.

Which **fairness** do we want?

Lawyer/Policy maker: *Cannot use protected classes for making decisions.*

Race (Civil Rights Act of 1964), **Color** (Civil


PROPUBLICA TOPICS ▾ SERIES ▾ NEWS APPS GET INVOLVED IMPACT ABOUT

MACHINE BIAS

Facebook Lets Advertisers Exclude Users by Race

Facebook's system allows advertisers to exclude black, Hispanic, and other "ethnic affinities" from seeing ads.

by **Julia Angwin** and **Terry Parris Jr.**, Oct. 28, 2016, 1 p.m. EDT

Nondiscrimination Act)

Disparate Treatment v/s Impact

This has to be a **collective decision** we need to **consciously** reach at, after a **deeper dive** into the **application**.

Which **fairness** do we want?

Algorithm designer: *awareness of protected classes can fix bias*

This has to be a **collective decision** we need to **consciously** reach at, after a **deeper dive** into the **application**.

Which **fairness** do we want?

Statistician: *cannot have equal false positive, negative rates & calibration simultaneously*

Prediction Fails Differently for Black Defendants

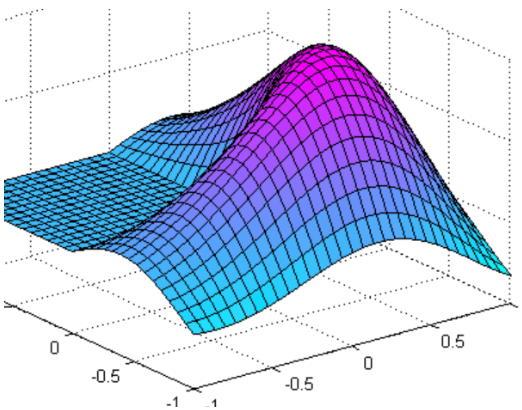
	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

Overall, Northpointe's assessment tool correctly predicts recidivism 61 percent of the time. But blacks are almost twice as likely as whites to be labeled a higher risk but not actually re-offend. It makes the opposite mistake among whites: They are much more likely than blacks to be labeled lower risk but go on to commit other crimes. (Source: ProPublica analysis of data from Broward County, Fla.)

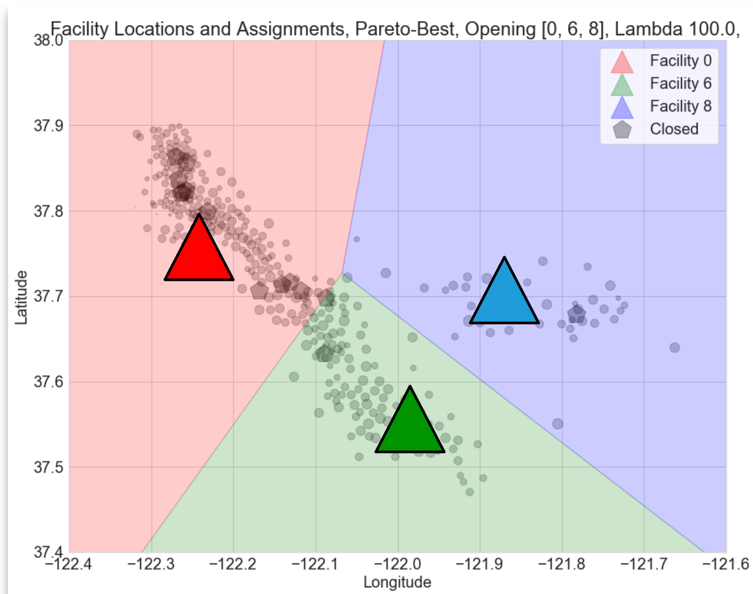
COMPAS Debate: Northpointe v/s ProPublica

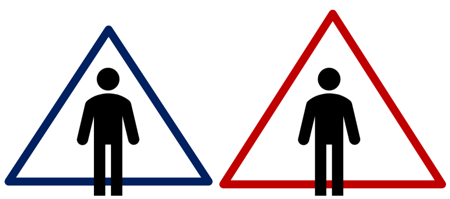
This has to be a **collective decision** we need to **consciously** reach at, after a **deeper dive** into the **application**.

Which **fairness** do we want?



Optimizer: *can at times have approximately fair solutions for multiple metrics together*

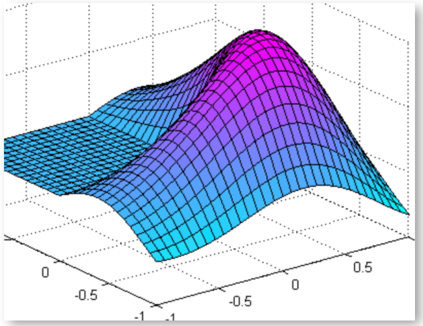




Group by **race**? **income**? **insurance**?

This has to be a **collective decision** we need to **consciously** reach at, after a **deeper dive** into the **application**.

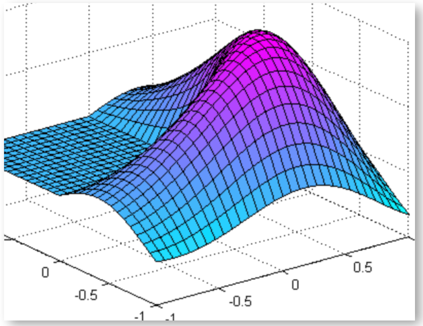
Which **fairness** do we want?



Economists, Behavioral scientists, Humans-in the loop, ..

This has to be a **collective decision** we need to **consciously** reach at, after a **deeper dive** into the **application**.

Which **fairness** do we want?



*

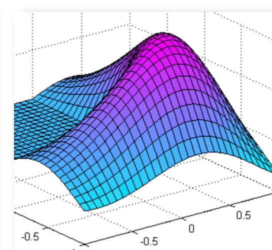
This has to be a **collective decision** we need to **consciously** reach at, after a **deeper dive** into the **application**.

Summary

Transparency, Interpretability, Gameability?

- **Bias in the data, models and variables**
  - Collection, Feedback, Proxies, Test Data, Representation..
- **Fairness Metrics**
  - Statistical measures: accuracy, false positive rate, true positive rate, calibration, ...
  - Equity measures: general decisions, average metric, total metric, group choice, …
- **Trolley Problem of Choice:** it's an inclusive story



*

Questions? swatig@gatech.edu, **www.swatigupta.tech**