Ethics and Fairness in ML

CS5785 Fall 2019

- Ethics in private/public sector ML
  Preview benefits of tech for public services

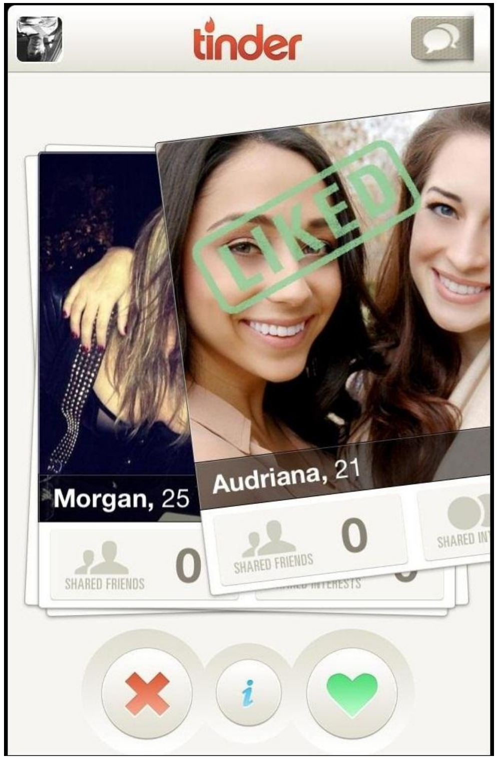
- Fairness in machine learning: 2016+

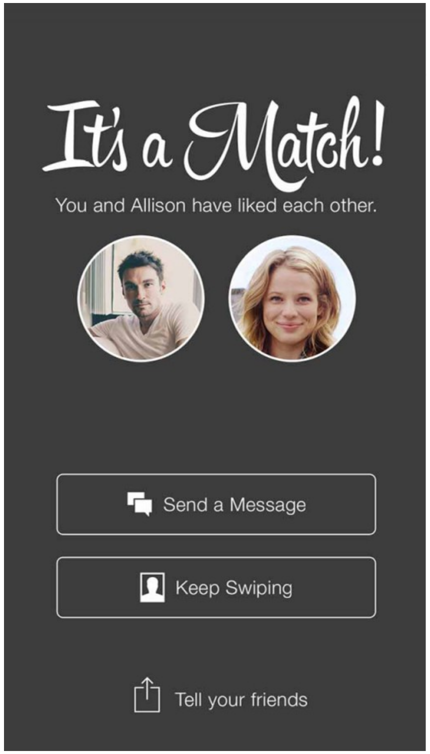
Example borrowed from Delip Rao

Today's business is metric-driven!

May want to increase:
- % right swipes
- % matches

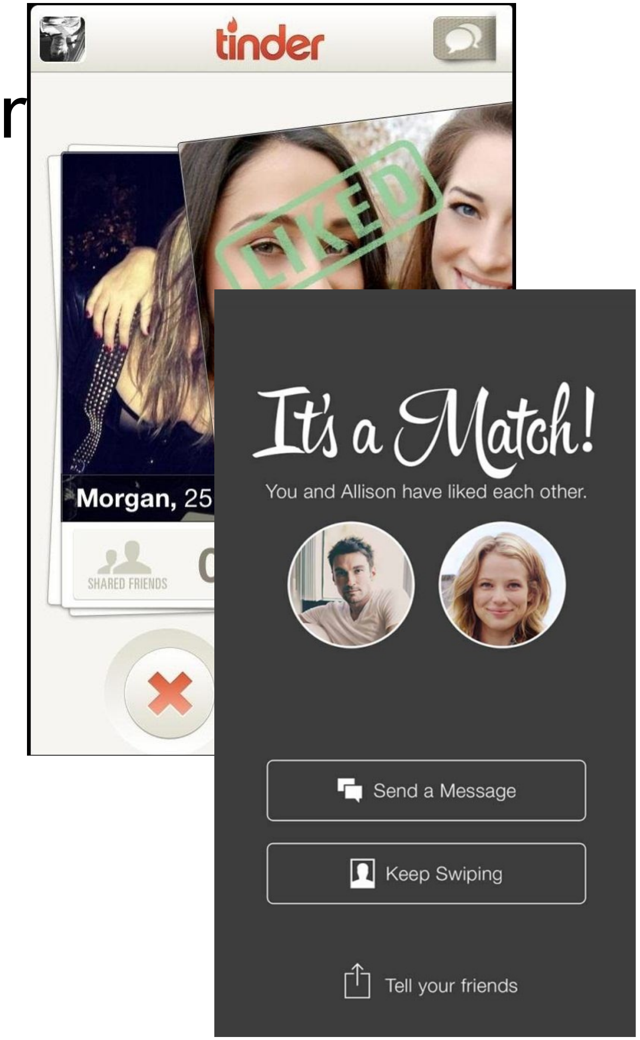
Opportunities:
Tons and tons of data,
mostly clean data,
many rich features





Example borrowed from Delip Rao

- Say we can improve metrics by including as a feature skin color (extracted using computer vision) for the ranking algorithm
- Should we?
- What about self-identified ethnicity (e.g. in profile)?
- Are recommendations restricted based on gender/ sex/orientation ok?



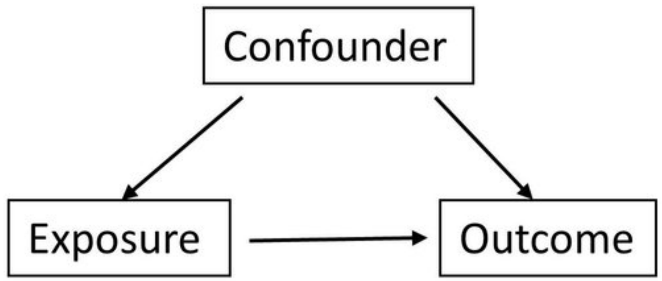
Real world data is confounded

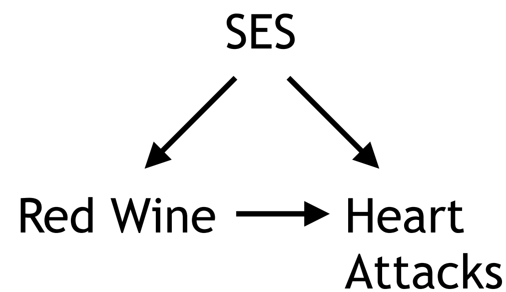
- Sometimes the confounding can lead to clear error and harm

- Sometimes the confounding is due to history

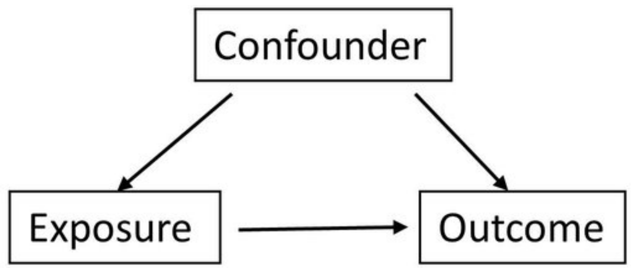
- Sometimes both

(A) Confounding

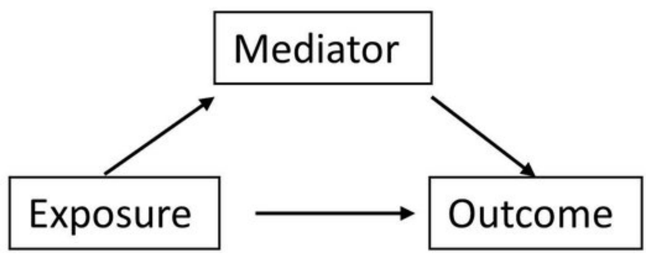


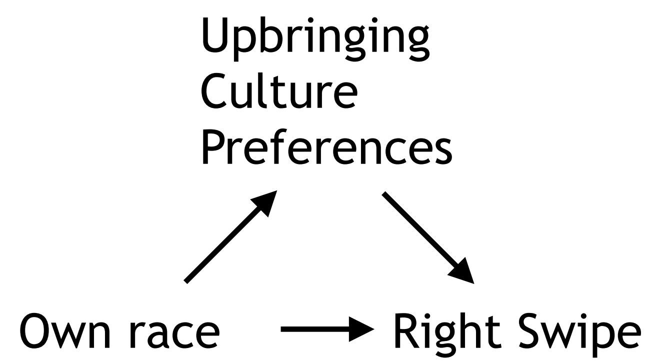


(A) Confounding



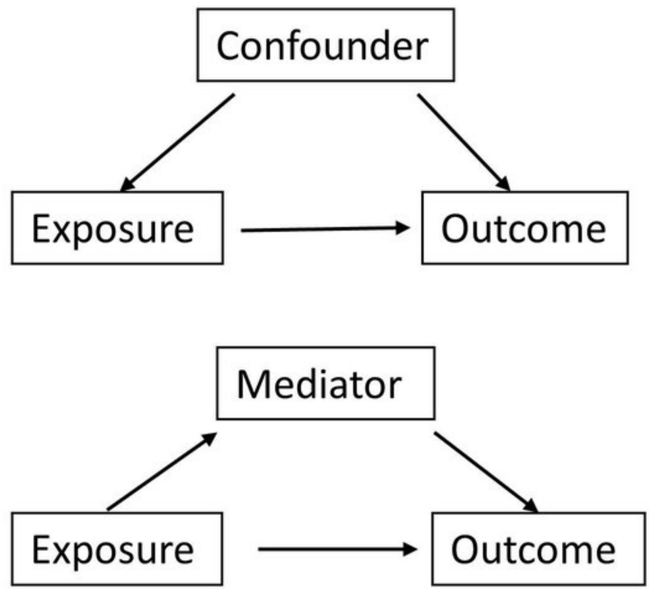
(B) Mediation

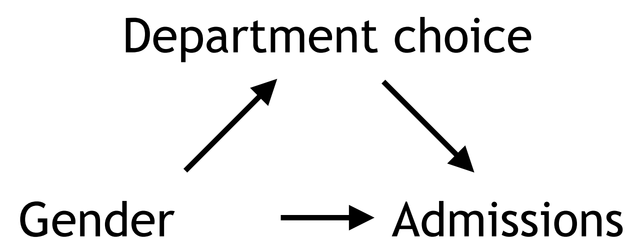


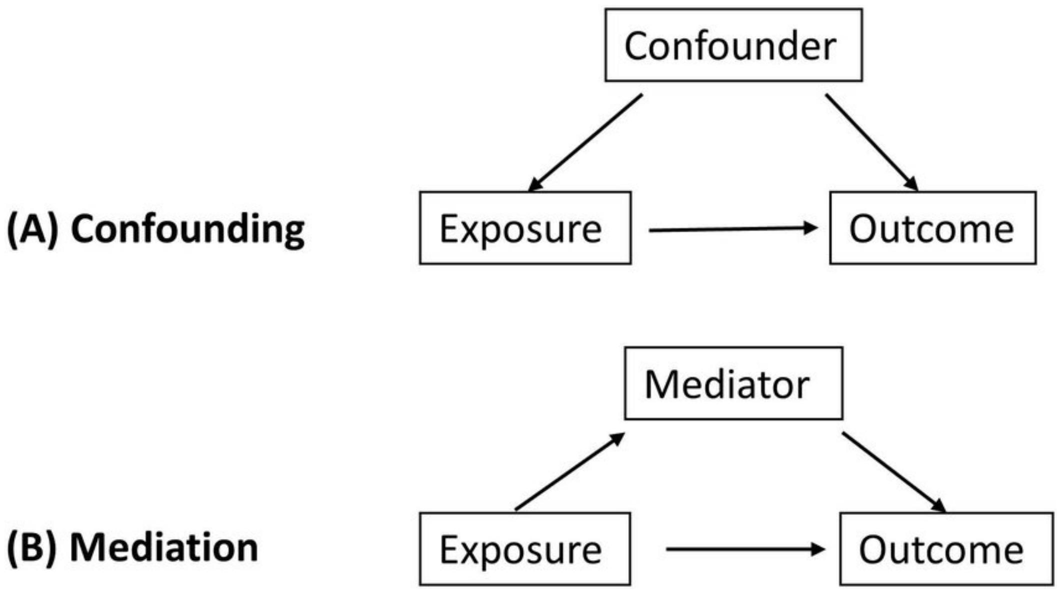


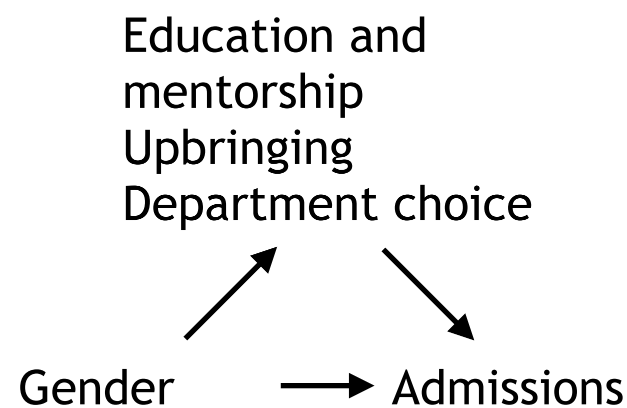
(A) Confounding

(B) Mediation






Confounder
(A) Confounding
Exposure
Outcome
Mediator
(B) Mediation
Exposure
Outcome


Education and
mentorship
Upbringing
Department choice
Gender
Admissions

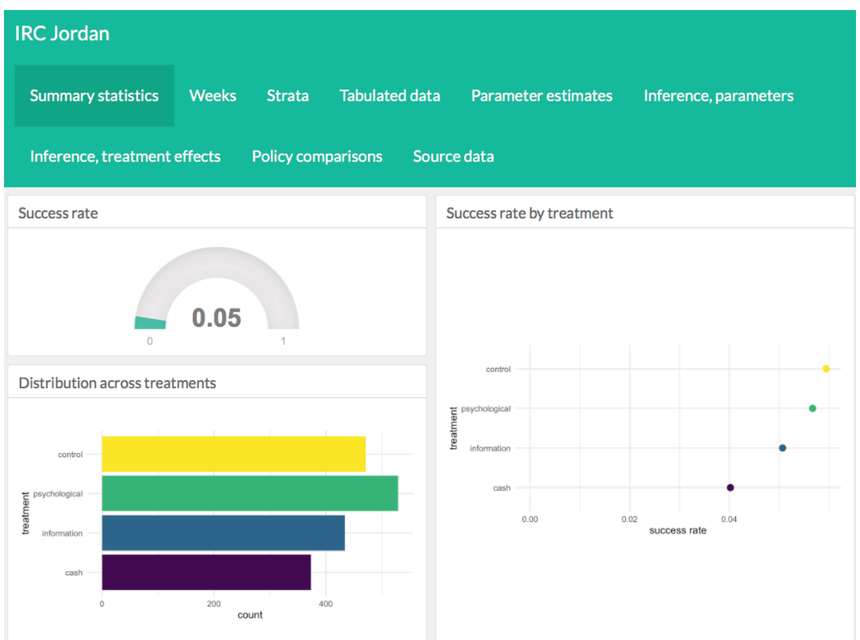
Relationships between causality and fairness

- Causal modeling lets us be precise about sources of bias; or “problematic” causal pathways of effects
- On the flipside: we care about *unfairness* if we do not want to perpetuate injustice:
  But the language of *improvement* of welfare is causal inference and policy

Predictive analytics and allocation of resources in the public sector

Bandits to allocate labor market interventions (cash, psychological, information interventions)

Caria, Stefano et al. 2019. "Job Search Assistance for Refugees in Jordan." AEA RCT Registry. September 06. https://doi.org/10.1257/rct.3870-2.0.



Service Type	Number Assigned	Percent Reentered
Emergency Shelter	2897	56.20
Transitional Housing	1927	40.22
Rapid Rehousing	589	53.48
Homelessness Prevention	2061	24.16
Total	7474	43.03

The Optimization Problem

Let x_{ij} be a binary variable representing whether or not household i is placed in intervention j. Then, the Integer Programming problem is given by

$$\min_{x_{ij}} \sum_i \sum_j p_{ij} x_{ij}$$

$$\text{subject to} \quad \sum_j x_{ij} = 1 \quad \forall i$$

$$\sum_i x_{ij} \leq C_j \quad \forall j$$

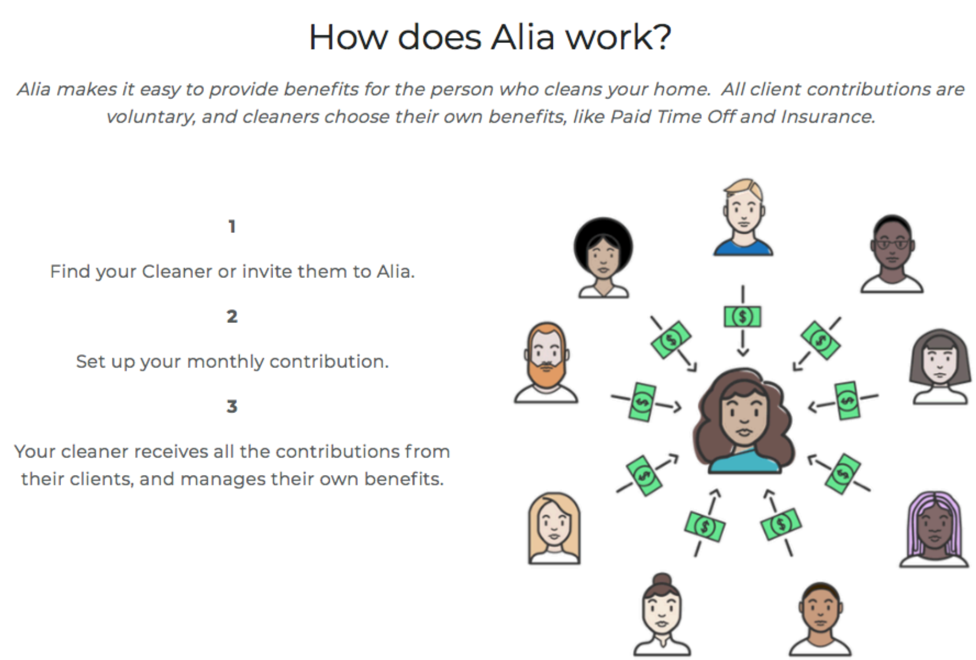
Using causal ML to allocate households to homelessness interventions (shelters, rapid rehousing, interventional resources)

Kube, Amanda, Sanmay Das, and Patrick J. Fowler. "Allocating interventions based on predicted outcomes: A case study on homelessness services." *Proceedings of the AAAI Conference on Artificial Intelligence*. 2019.

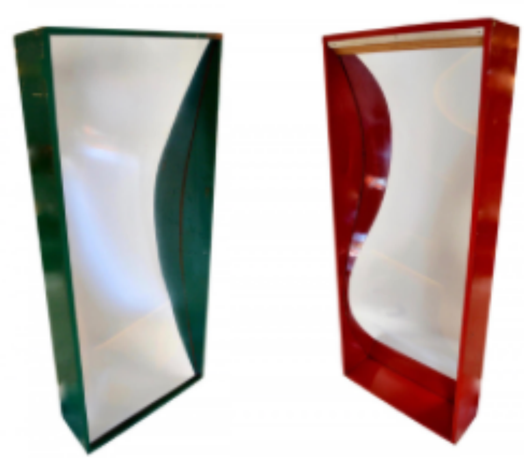
Tech that addresses market failures in social services

NowPow: personalized referrals for social services
(Medicare/Medicaid research spin-out)

Alia: Portable benefits for home cleaners (National Domestic Workers Alliance (NDWA))



ML needs to learn from data from the real world



Is it supposed to be a transparent interface?

Does it introduce distortions?

On the flipside: what are useful distortions?

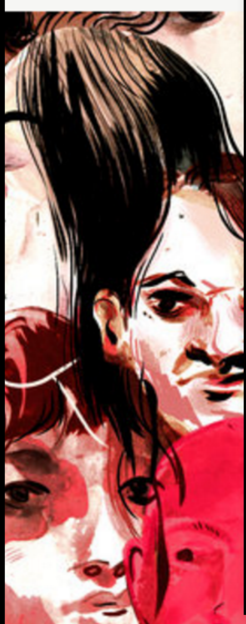
The Marsha

UT DONATE

The cing

Should prison

PREDICTIVE POLICING: USING MACHINE LEARNING TO DETECT PATTERNS OF CRIME



Errol Damelin from his office overlooking London's Regent's Park. "They're

Impartiality of learning machines

- Is it enough to just use colorblind/genderblind/X-blind data?
- Is justice blind? Do algorithms help?
- Do they hurt?
- Can an algorithm be racist if its inputs are colorblind?
- What is algorithmic bias?
- What bias is allowed? What bias isn't allowed?

HIDDEN BIAS

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

ON A SPRING AFTERNOON IN 2014, Brisha Borden was running late to pick up her god-sister from school when she spotted an unlocked kid's blue Huffy bicycle and a silver Razor scooter. Borden and a friend grabbed the bike and scooter and tried to ride them down the street in the Fort Lauderdale suburb of Coral Springs.

Just as the 18-year-old girls were realizing they were too big for the tiny conveyances — which belonged to a 6-year-old boy — a woman came running after them saying, "That's my kid's stuff." Borden and her friend immediately dropped the bike and scooter and walked away.

But it was too late — a neighbor who witnessed the heist had already called the police. Borden and her friend were arrested and charged with burglary and petty theft for the

termining the
d and the
cisions about

at rely on data
lgorithms are
gorithms adjust
earchers in
human

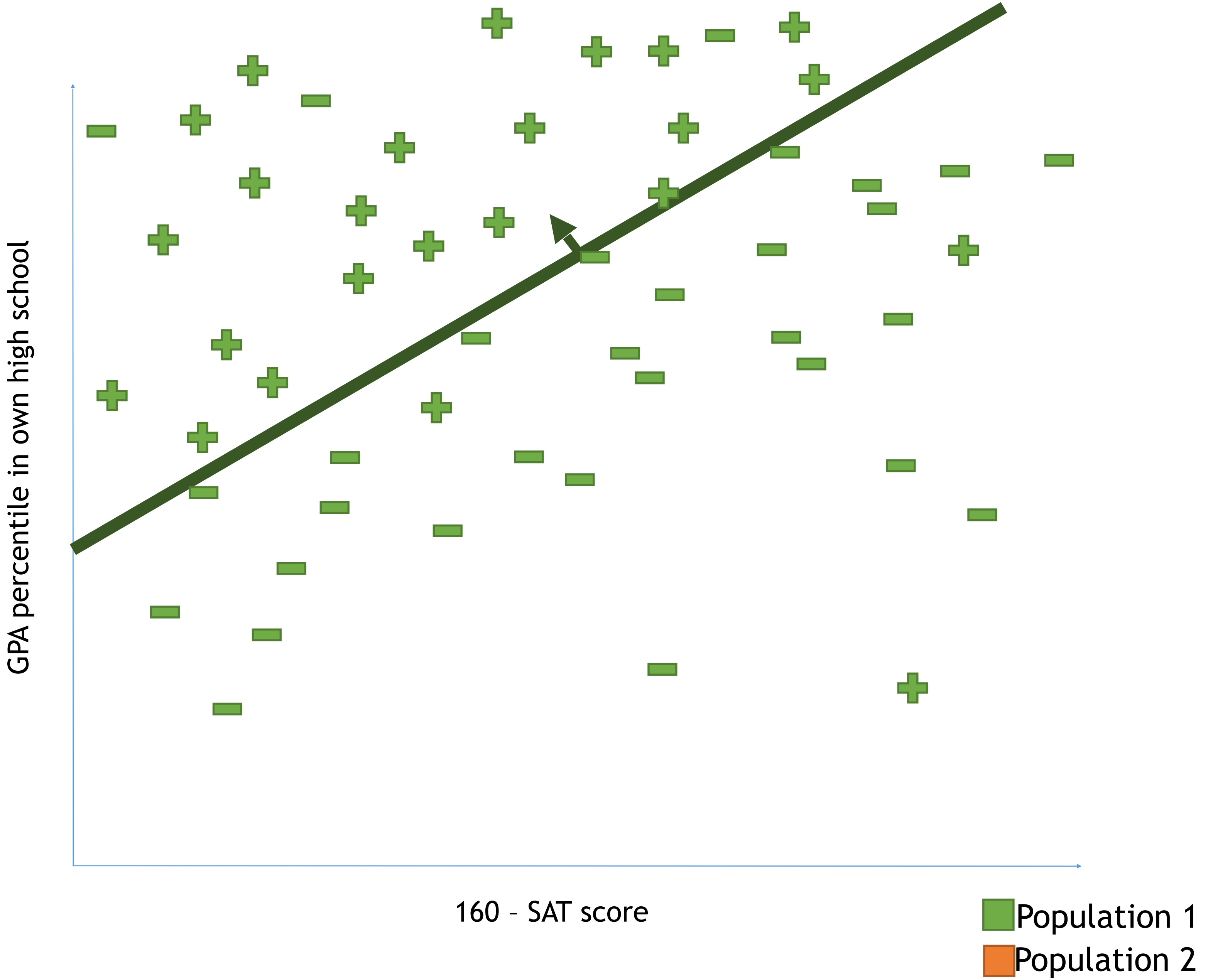
ad for high-
to women, a

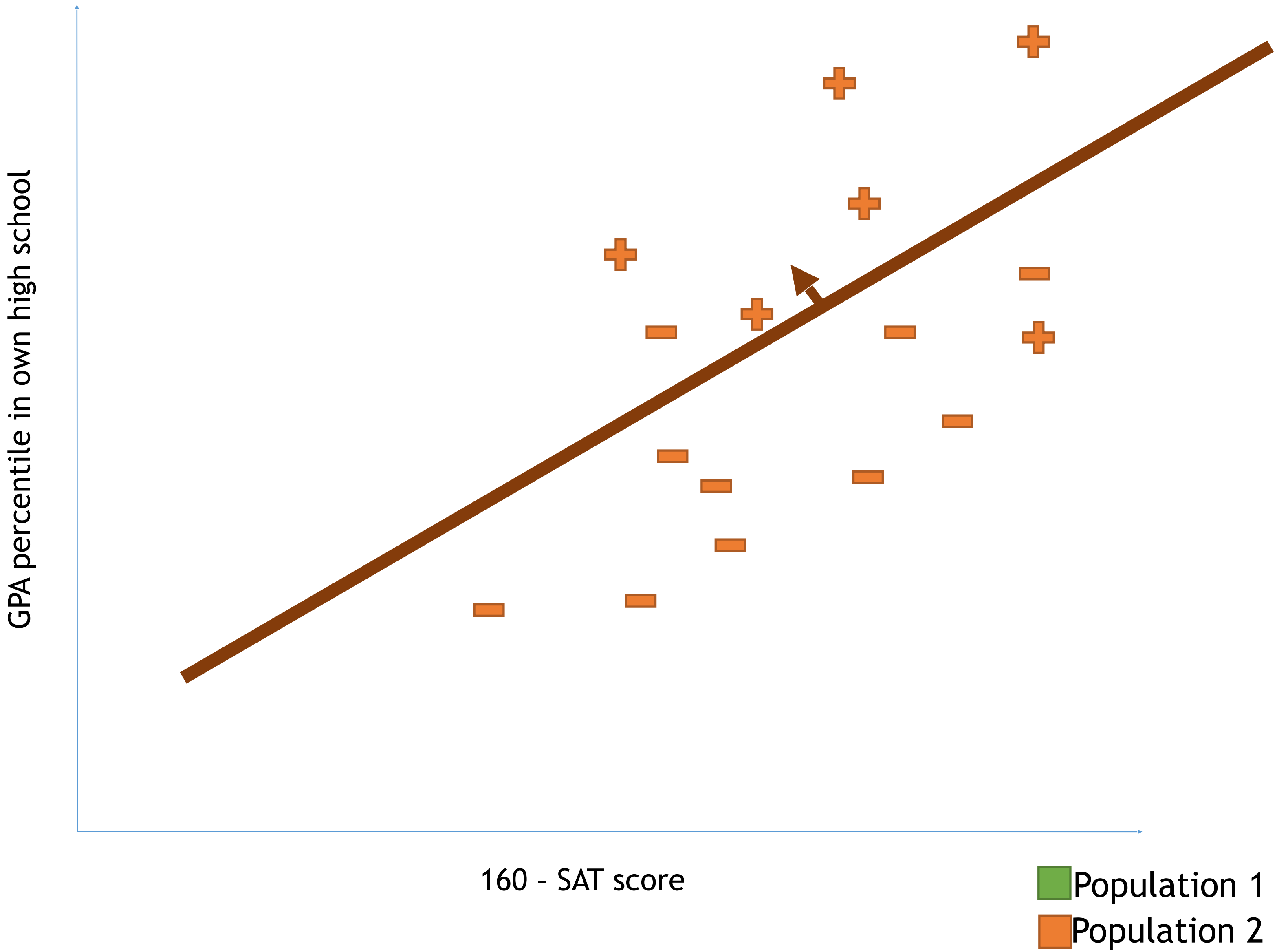
records were
vely black
Commission said
neighborhoods

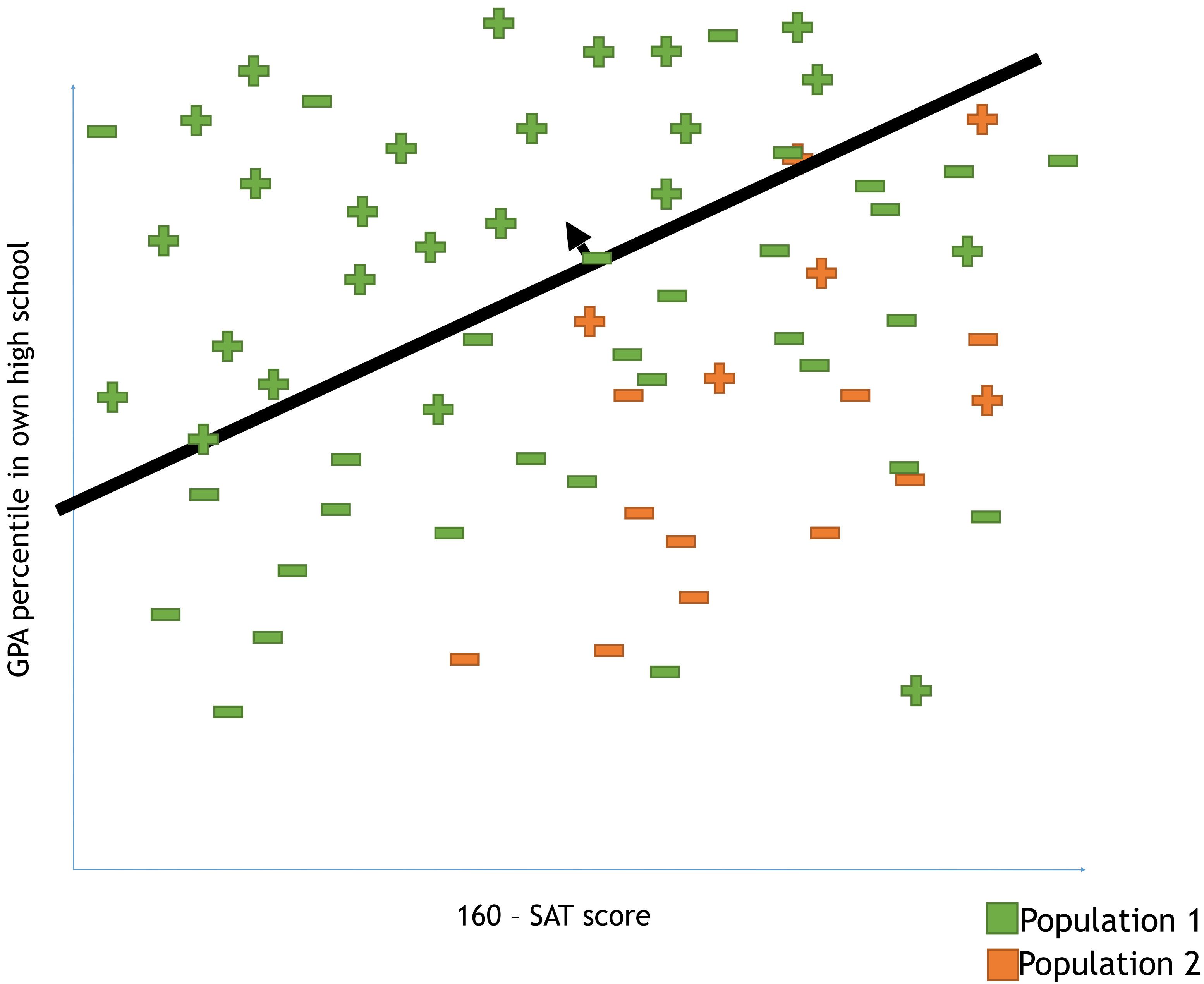
Adapted from Aaron Roth

Why might machines be "unfair"?

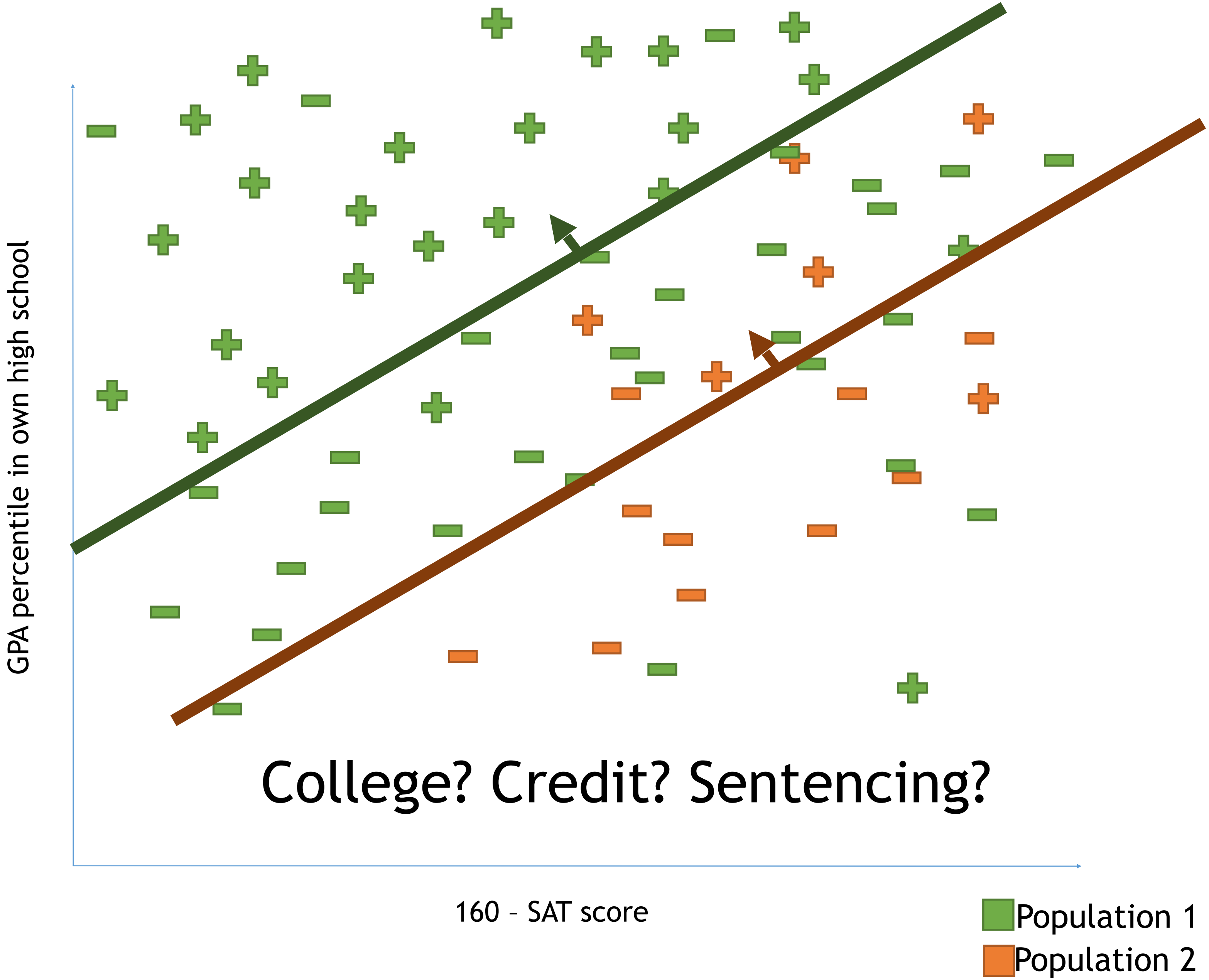
- Many reasons:
  - Data might encode existing biases
    - E.g. Y labels are "arrested" rather than "committed crime"
  - Data collection feedback loops
    - E.g. only observe paid back vs defaulted if the loan was approved and credited.
  - Different populations with different life-courses.
    - E.g. "SAT score" might correlate with eventual academic success differently in populations that employ SAT tutors.
    - E.g. "# accounts opened" reflects *both* creditworthiness and ethno-culturally determined factors
  - Less data (by definition) about minority populations.


GPA percentile in own high school
160 – SAT score
Population 1
Population 2


GPA percentile in own high school
160 – SAT score
Population 1
Population 2

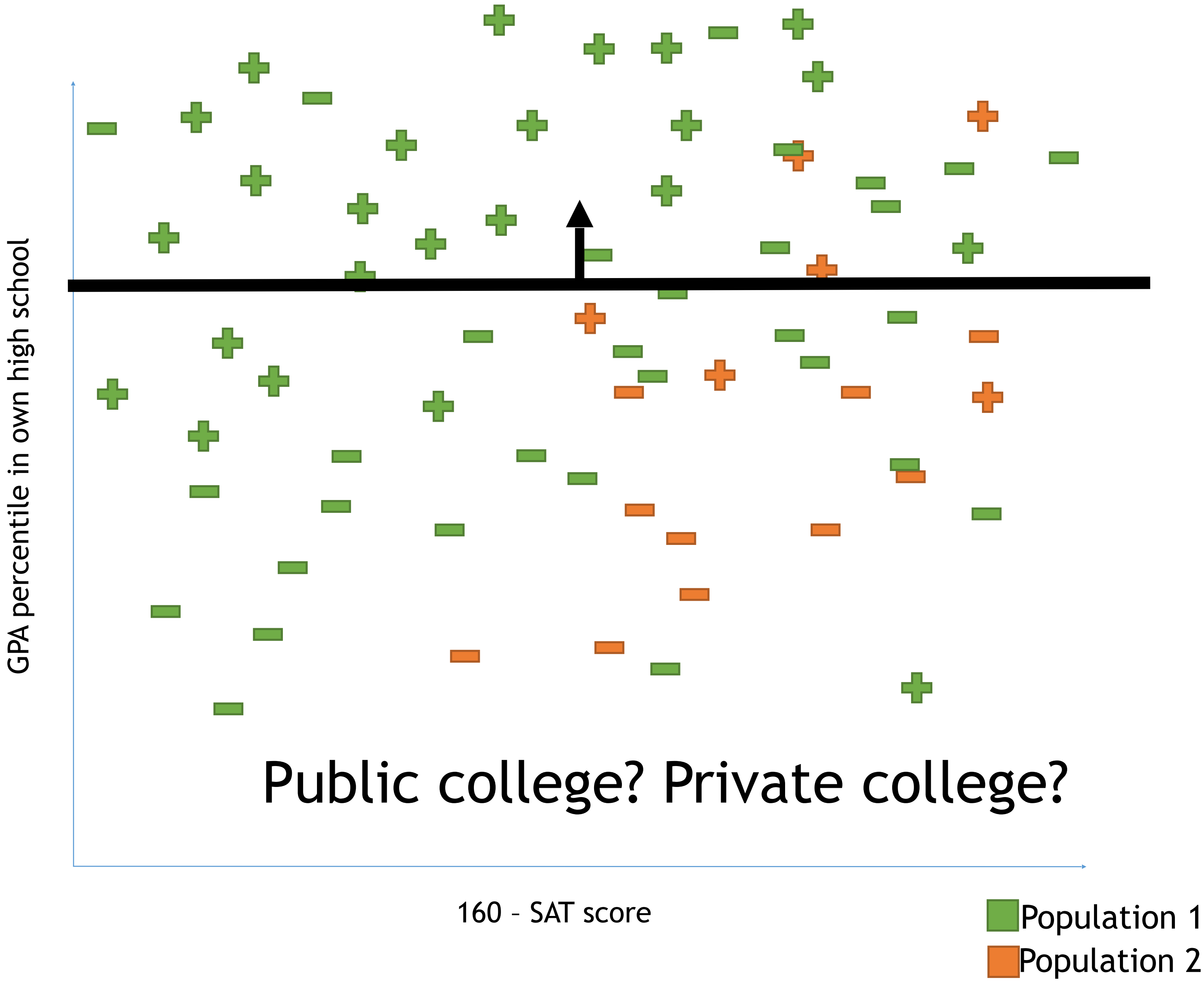

GPA percentile in own high school
160 – SAT score
Population 1
Population 2

When is affirmative action ok?

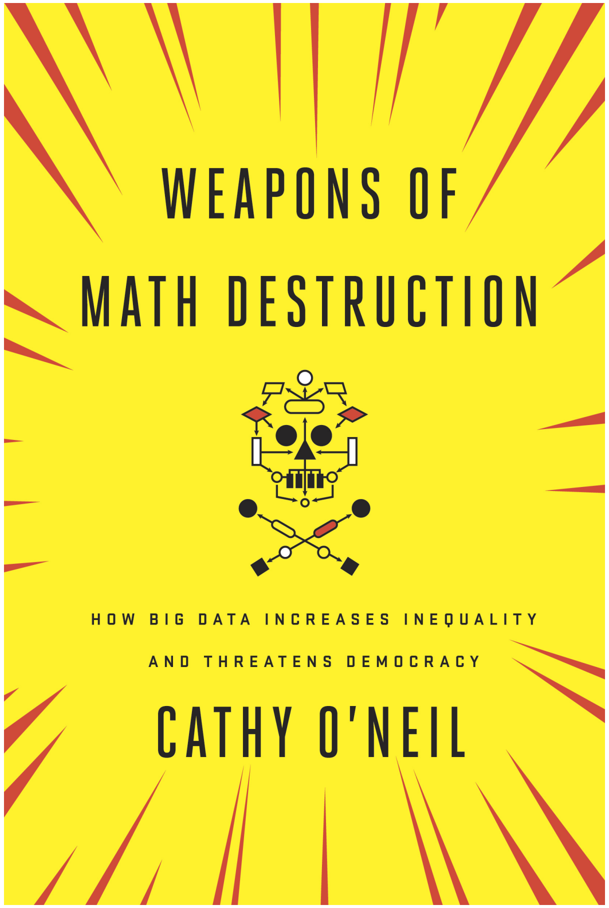


College? Credit? Sentencing?

What kind of affirmative action is ok?



Affirmative action beyond the data:
Societal values and aspirations



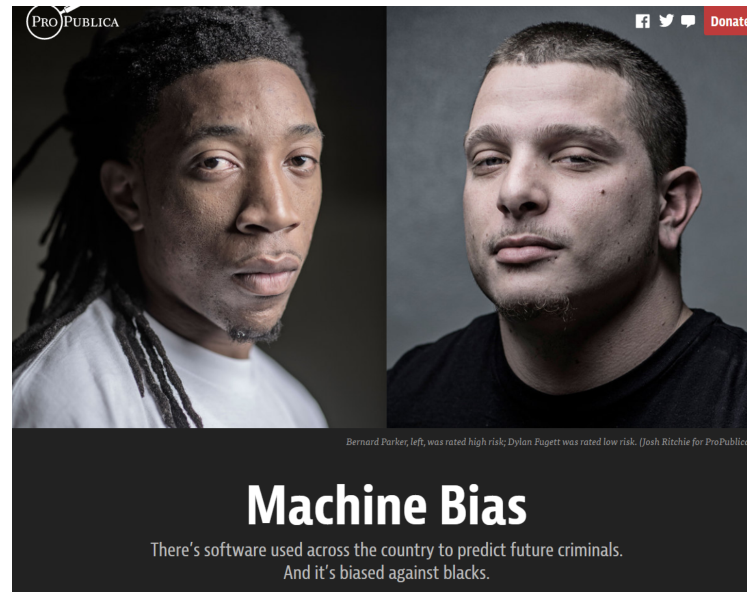
"If we allowed a model to be used for college admissions in 1870, we'd still have 0.7% of women going to college."
(on her blog mathbabe.org)

What does discrimination law aim to achieve?

Disparate Treatment	**Disparate Impact**
Procedural fairness	Distributive justice
Equality of opportunity	Minimized inequality of outcome

Defining Fairness:
The case of Northpointe COMPAS

- ML model to provide a risk score that predicts: “will this defendant commit a crime within their next two years of freedom?”
- Race is not an input feature
- Used for bail and sentencing
- Famed investigation by ProPublica on use in FL: biased against black offenders

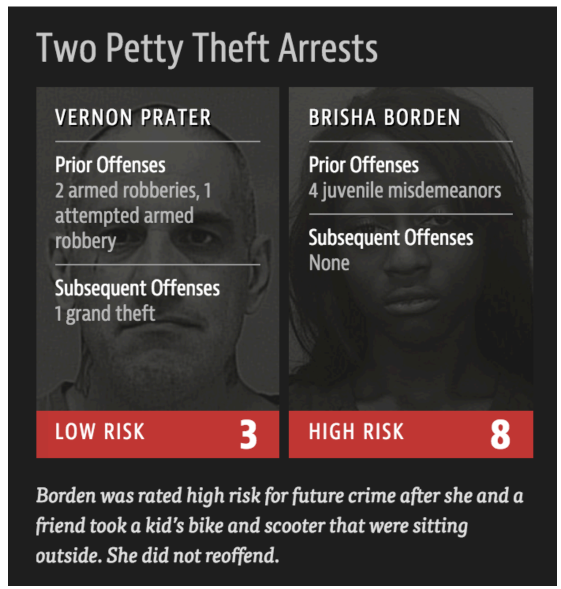


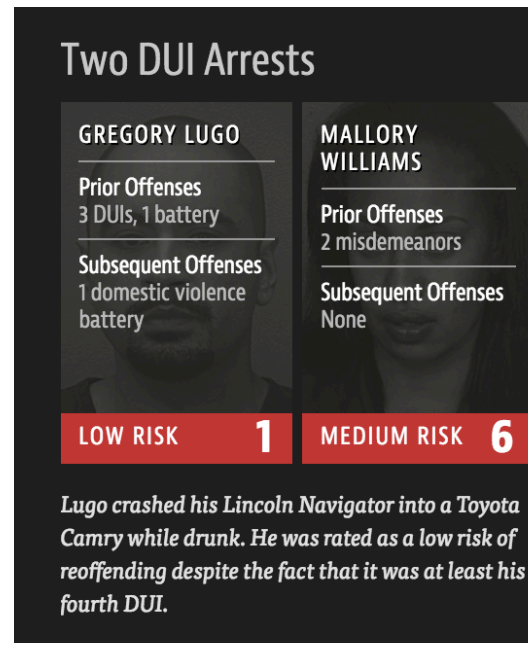
Defining Fairness:
The case of Northpointe COMPAS

The algorithm, called COMPAS, is used nationwide to decide whether defendants awaiting trial are too dangerous to be released on bail. In May, the investigative news organization ProPublica claimed that COMPAS is biased against black defendants. Northpointe, the Michigan-based company that created the tool, released its own report questioning ProPublica's analysis. ProPublica rebutted the rebuttal, academic researchers entered the fray, this newspaper's Wonkblog weighed in, and even the Wisconsin Supreme Court cited the controversy in its recent ruling that upheld the use of COMPAS in sentencing.

Defining Fairness:
The case of Northpointe COMPAS

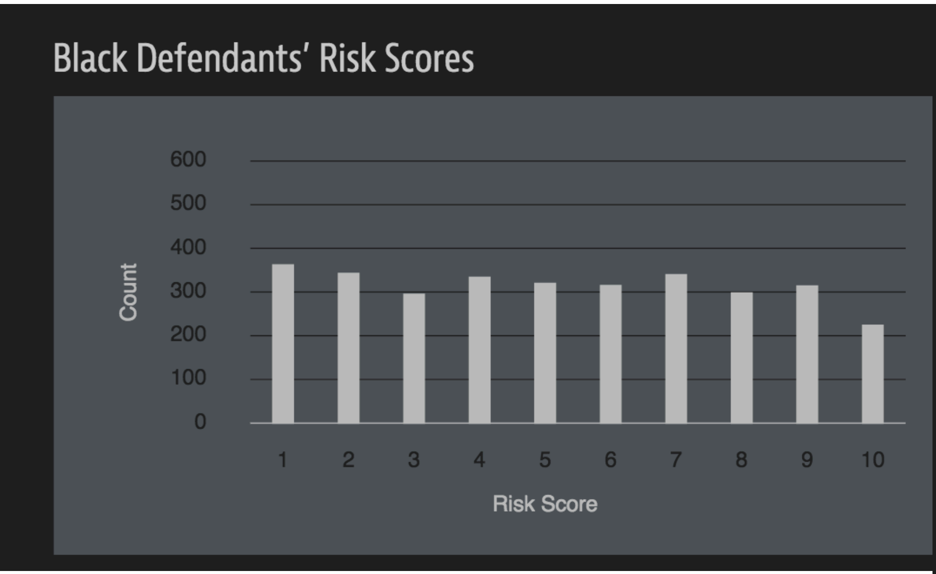
ProPublica report





Defining Fairness:
The case of Northpointe COMPAS

ProPublica report



These charts show that scores for white defendants were skewed toward lower-risk categories. Scores for black defendants were not. *(Source: ProPublica analysis of data from Broward County, Fla.)*

Defining Fairness:
The case of Northpointe COMPAS

ProPublica report

Prediction Fails Differently for Black Defendants

		WHITE	AFRICAN AMERICAN
FPR	Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
FNR	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.)*

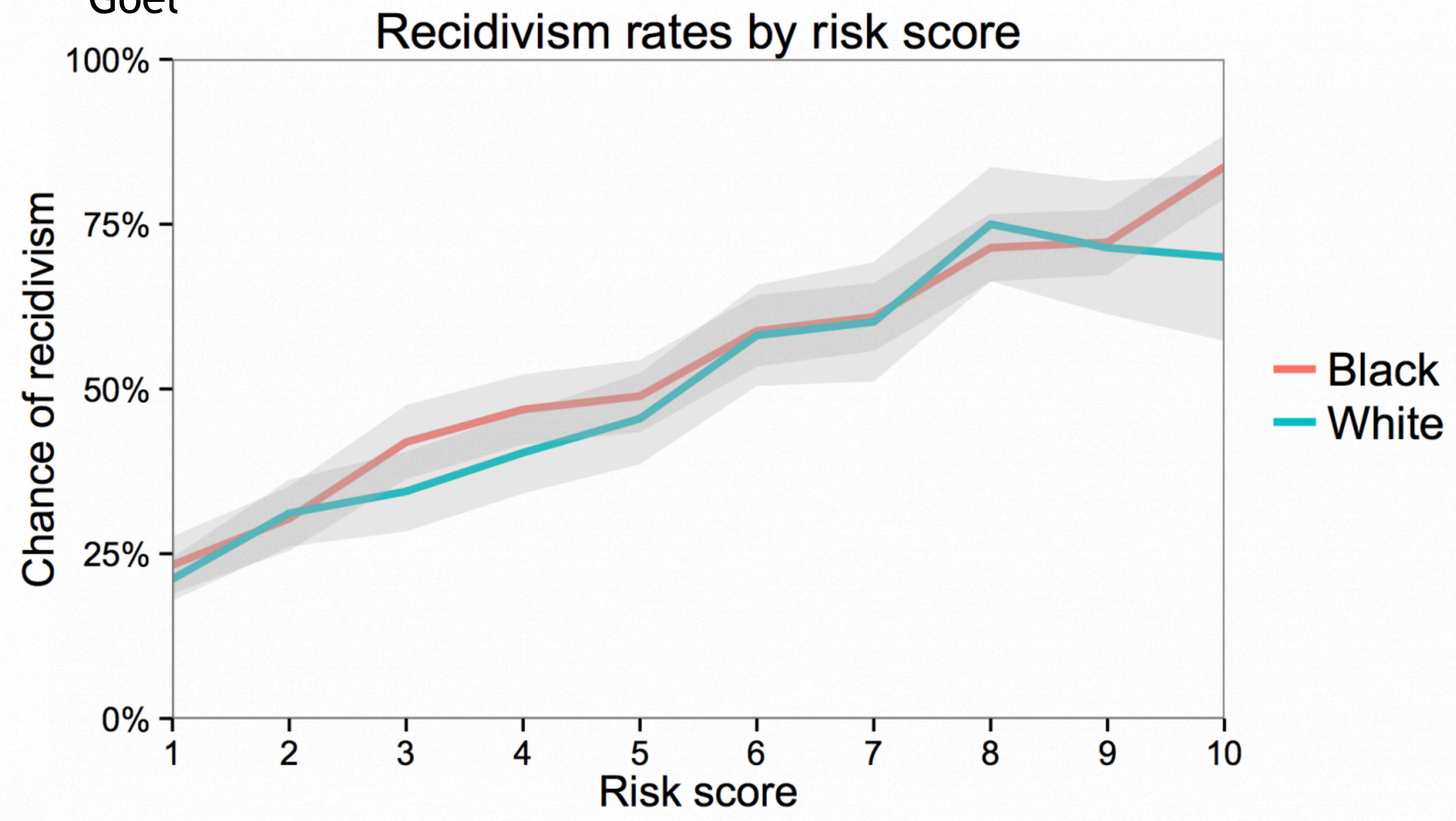
Defining Fairness:
The case of Northpointe COMPAS

- Algorithms are racist! Down with algorithms!
- Maybe so… but not so fast
- Maybe ML indeed has no place in justice system
- But was COMPAS really “unfair”?
- If so, can it be made “fair”?

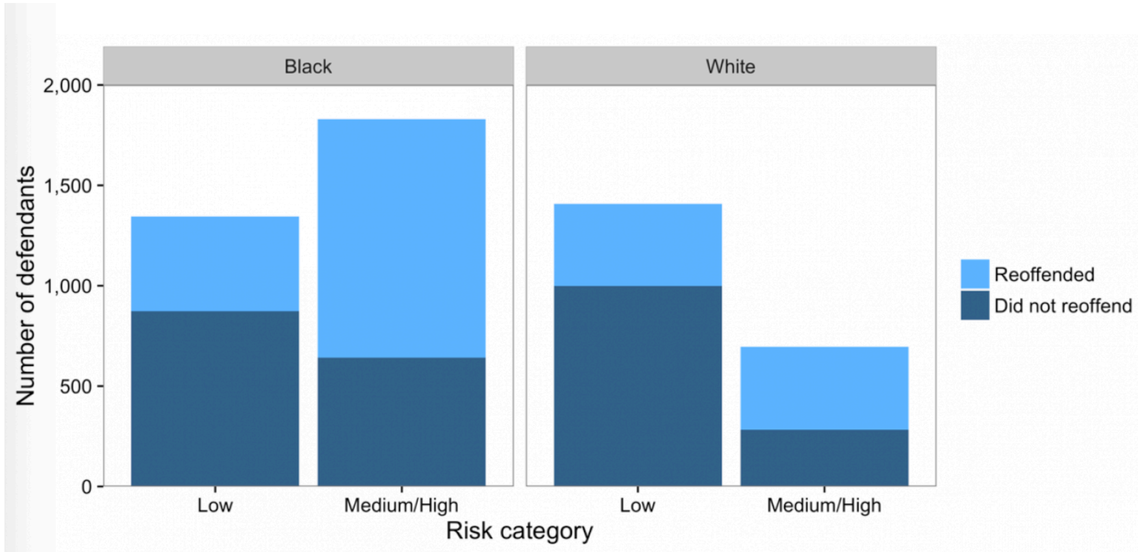
Washington Post

"A computer program used for bail and sentencing decisions was labeled biased against blacks. It's actually not that clear."

By Sam Corbett-Davies, Emma Pierson, Avi Feller and Sharad Goel



“A computer program used for bail and sentencing decisions was labeled biased against blacks. It’s actually not that clear.”
By Sam Corbett-Davies, Emma Pierson, Avi Feller and Sharad Goel



- Within each risk category, the proportion of defendants who reoffend is approximately the same regardless of race (Northpointe’s definition of fairness)
- The overall recidivism rate for black defendants is higher than for white defendants (52% vs. 39%)
- Black defendants are more likely to be classified as med/high risk (58% vs. 33%)
- Black defendants who don’t reoffend are predicted to be riskier than white defendants who don’t reoffend (ProPublica’s criticism of the algorithm)

ProPublica's evidence of bias

	White Defendants	Black Defendants
Proportion of those who **didn't** reoffend labeled as **med/high risk**	24%	45%
Proportion of those who **did** reoffend labeled as **low risk**	48%	28%

Sam Corbett-Davies, Emma Pierson, Avi Feller and Sharad Goel

Northpointe's evidence of fairness

	White Defendants	Black Defendants
Proportion of those labeled as **med/high risk** who **did** reoffend	59%	63%
Proportion of those labeled as **low risk** who **didn't** reoffend	71%	65%

Sam Corbett-Davies, Emma Pierson, Avi Feller and Sharad

Can’t have it all! – How unfair!

- Northpointe says fair would be
  1. Positive precision is the same across groups
  2. Negative precision is the same across groups
- ProPublica says fair would be
  3. True positive rate is the same across groups
  4. False positive rate is the same across groups
- **Fact of life**: Can never have all of 1-4 *unless* either we can make *perfect* predictions or the groups have the *same proportion* of positive instances
- See Kleinberg, Mullainathan and Raghavan ‘16 fairmlbook.org

Can't have it all! – How unfair!

- Northpointe says fair would be
  1. Positive precision is the same across groups
  2. Negative precision is the same across groups
- ProPublica says fair would be
  3. True positive rate is the same across groups
  4. False positive rate is the same across groups
- **Fact of life**: If we enforce 3 (and give up 1-2) by having different risk score thresholds by group, we will end up with 7% more freely roaming reoffenders
  - Anyway, race-based threshold won't hold up in a case brought by lower-threshold person using 14th Amendment
  - See Corbett-Davies et al 17

What fairness do we want? At what price?

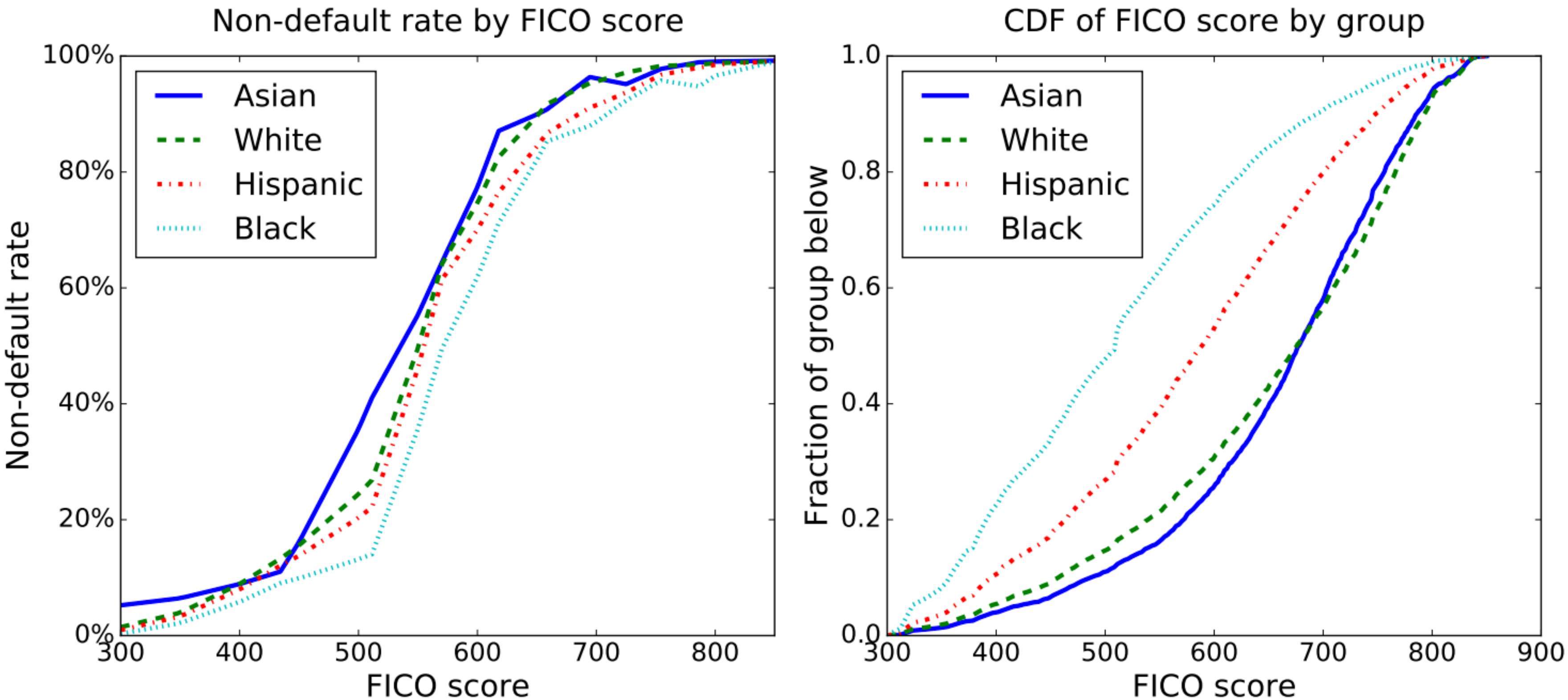
- In many cases, a good form of fairness is:
- True positive rate is the same across groups
  = equality of opportunities for qualified individuals
- FICO score *should* be independent of race given creditworthiness
  - Treat African-American creditworthy person the same as Asian-American creditworthy person
  - Don’t use variables like # bank accounts as proxies for race (or rather as proxies for creditworthiness via race)
- A *qualified* non-cis-male *should* be treated the same as a *qualified* cis-male when hiring

Adjusting for fairness

- "In-processing":
  Constrained optimization to learn a model that satisfies "fairness" constraints

- "Post-processing"
  Adjust a given black-box model to satisfy "fairness" constraints

- "Data pre-processing"
  Learn a representation of the data that satisfies independence properties



Hardt, Price & Srebro ‘16

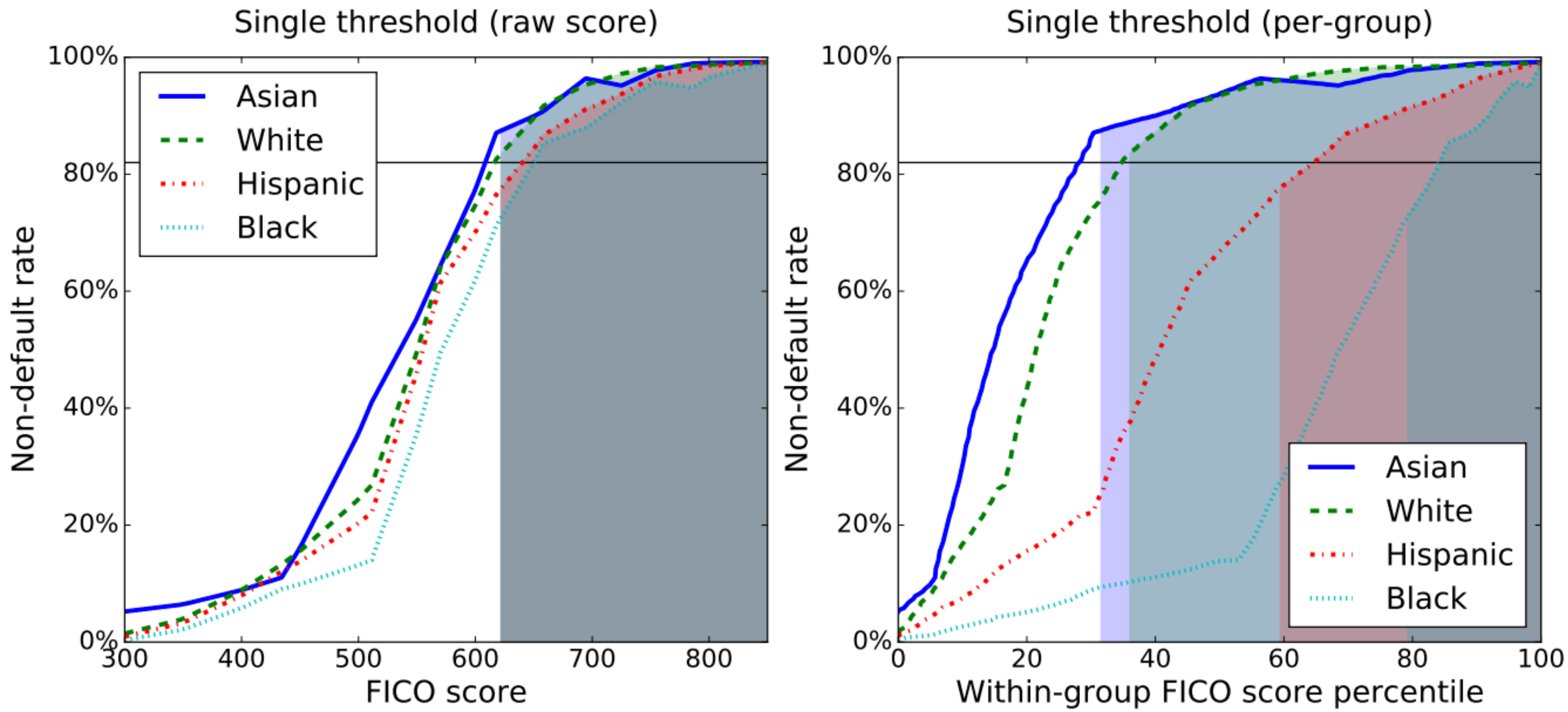


Figure 8: The common FICO threshold of 620 corresponds to a non-default rate of 82%. Rescaling the x axis to represent the within-group thresholds (right), $\Pr[\widehat{Y} = 1 \mid Y = 1, A]$ is the fraction of the area under the curve that is shaded. This means black non-defaulters are much less likely to qualify for loans than white or Asian ones, so a race blind score threshold violates our fairness definitions.

Hardt, Price & Srebro '16

These examples (hiring, lending, crime) are high-stakes & controversial
(which you might not end up working in)

Black-box ML has no guarantee of being aligned with human, societal values

Can product design and development that leverages ML, aligned with human values, be a value proposition?

Other concerns: ethics in data collection

The New York Times

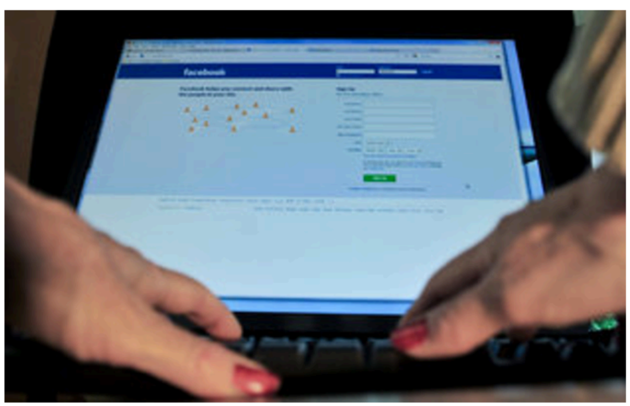
LOG IN

TECHNOLOGY

Facebook Tinkers With Users' Emotions in News Feed Experiment, Stirring Outcry

By **VINDU GOEL** JUNE 29, 2014

216



Facebook revealed that it had altered the news feeds of over half a million users in its study.
Karen Bleier/Agence France-Presse — Getty Images

To Facebook, we are all lab rats.

Other concerns: privacy
avoidable vs unavoidable

Fredrikson, Jha, Ristenpart '15

Figure 1: An image recovered using a new model inversion attack (left) and a training set image of the victim (right). The attacker is given only the person's name and access to a facial recognition system that returns a class confidence score.

Gaydar: Facebook friendships expose sexual orientation

by Carter Jernigan and Behram F.T. Mistree

You Can't Keep Your Secrets From Twitter

On the Internet, no one knows you're secretly a man (or woman), right? Think again. Just by examining patterns in tweets, you can infer a Twitter user's gender. A look at the words (Etsy, Jeep, redneck...) that make men and women give themselves away.