MACHINE LEARNING AND DISCRIMINATION

CS181

Spring 2018

ME!

- Kate Vredenburgh
  - kvredenburgh@fas.harvard.edu
  - Office hours: Wednesday 12-2pm, Emerson 306

THE CREDIT PROBLEM

- Phoebe Robinson and Jessica Williams come to you asking for a loan to make a pilot to pitch a series to HBO.
- You have to make a prediction about whether they'll pay back their loan.

- Income
- Income to debt ratio
- ...Location?
- …Race?

IS ALL INFORMATION FAIR GAME?

Is there information that banks and other private companies *shouldn't* be allowed to use in making judgments about creditworthiness? Why not?

No	Yes
The assessment of the risk that will pay back a loan (their creditworthiness) involves a certain level of uncertainty.	People making these decisions are fallible human agents with implicit biases.
In a society where loan repayment differs between groups, effective credit assessment may require using facts about someone's race as the basis for a loan decision.	Facts about who can repay a loan are partially determined by facts about past and current injustice.

WHAT WAS BEHIND THE PROS AND CONS

- A conception of the social good.
  - The principles and values by which social institutions establish basic rights and liberties (freedom of occupation; free speech), distribute scarce resources (university places; political offices), organize work (profit maximizing; meaningful work for a wide range of employees), etc.

CURRENT USES OF ML

- Generating credit scores
- Predicting recidivism in prospective parolees
- Evaluating job candidates

Social goods!

MACHINE LEARNING AND THE SOCIAL GOOD

How can machine learning help us to realize socially good outcomes?

- What does it mean to do these tasks accurately?
- What does it mean to do these tasks ethically?

FOCUS: DISCRIMINATION

What is discrimination?

- Disparate treatment
- Disparate impact

DISPARATE TREATMENT

- Involves classifying someone in an impermissible way.
  - Some types of classification are morally neutral, whereas other types are morally problematic.
- *Intent* to discriminate, either by explicitly referring to class membership or not.

DISPARATE IMPACT

Looks at the *consequences* of classification on certain groups.

- No intent required.
- Practices with a disproportionate effect on one group do not cause disparate impact if they are "grounded in sound business considerations." (http://www.scotusblog.com/2015/06/paul-hancock-fha/)

PROTECTED ATTRIBUTES

- Age
- Disability
- National origin
- Race/color
- Religion
- Sex

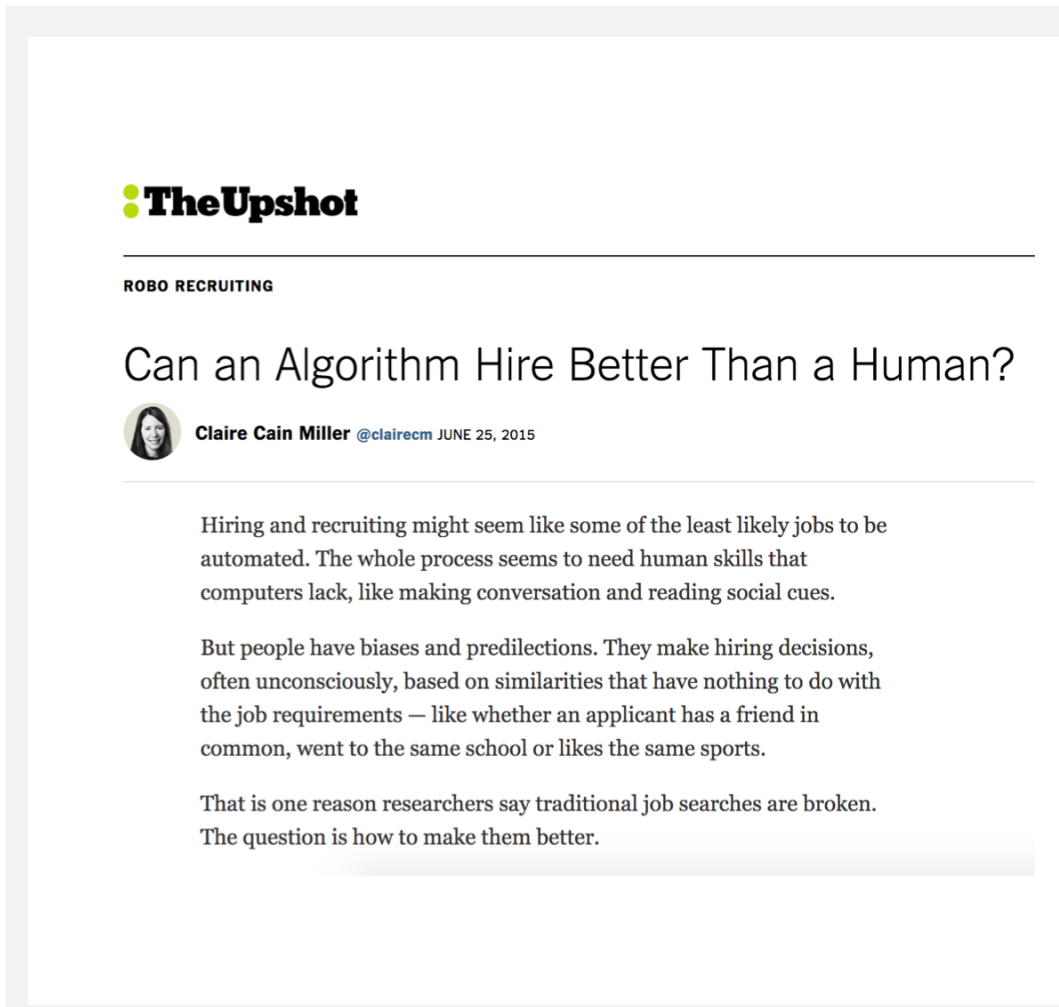
(from the US Equal Opportunity Employment Commission)

MACHINE LEARNING, ACCURACY, AND DISCRIMINATION

- Motivating idea: Discriminatory bias makes classification *less accurate*.
  - So, avoiding discrimination is a matter of building more accurate models.

DATA AS ANTIDOTE TO DISCRIMINATION?

- Objective metrics
- Controlled inputs
- Alternative to fuzzy, biased human reasoning



The Upshot

ROBO RECRUITING

Can an Algorithm Hire Better Than a Human?

Claire Cain Miller @clairecm JUNE 25, 2015

Hiring and recruiting might seem like some of the least likely jobs to be automated. The whole process seems to need human skills that computers lack, like making conversation and reading social cues.

But people have biases and predilections. They make hiring decisions, often unconsciously, based on similarities that have nothing to do with the job requirements — like whether an applicant has a friend in common, went to the same school or likes the same sports.

That is one reason researchers say traditional job searches are broken. The question is how to make them better.

"If they succeed, they say, hiring could become faster and less expensive, and their data could lead recruiters to more highly skilled people who are better matches for their companies. *Another potential result: a more diverse workplace*. The software relies on data to surface candidates from a wide variety of places and match their skills to the job requirements, free of human biases."

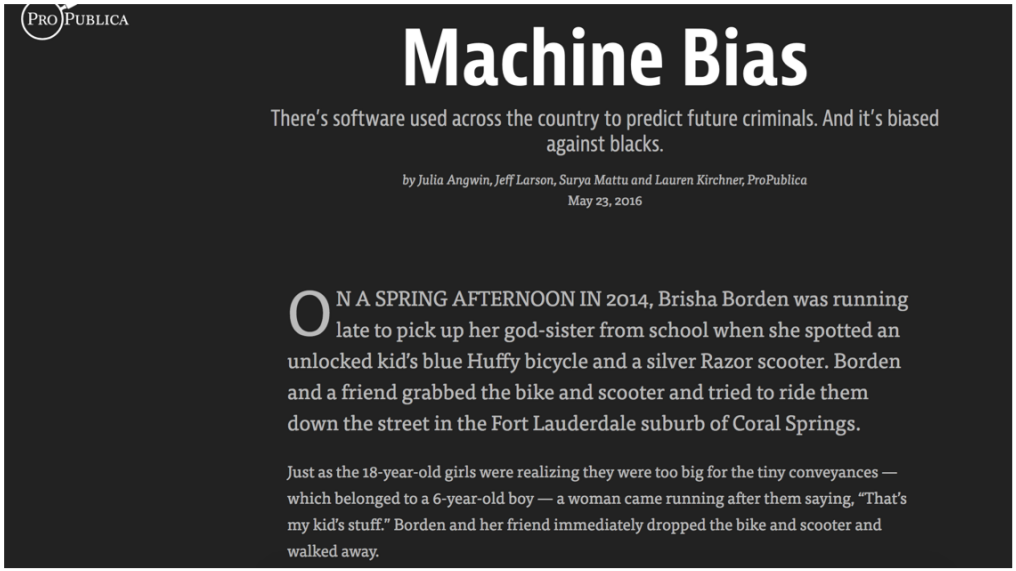
AUTOMATIC LOAN UNDERWRITING

"Compared with traditional manual underwriting, AU [automated underwriting] more accurately predicts default, and AU's greater accuracy results in higher borrower approval rates, especially for underserved applicants." (Gates, Perry, and Zorn 2002: 370)

BIAS IN TRAINING DATA

- A machine learning system may be trained on data infused with human bias
  - Recidivism scores based on prior arrests, age of first police contact, parents' incarceration record

O'Neil, Cathy. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown, 2016.



PRO PUBLICA

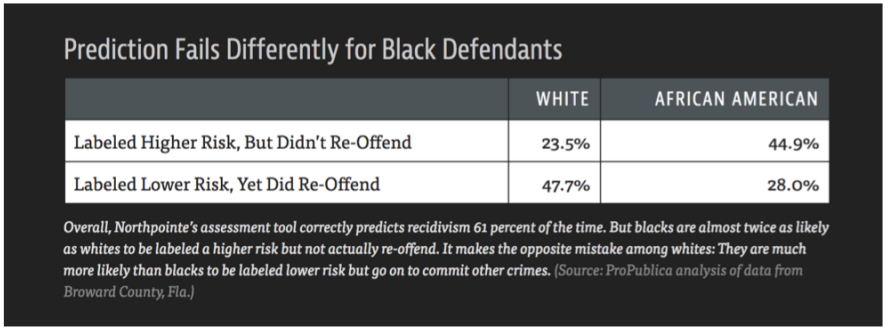
Machine Bias

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.

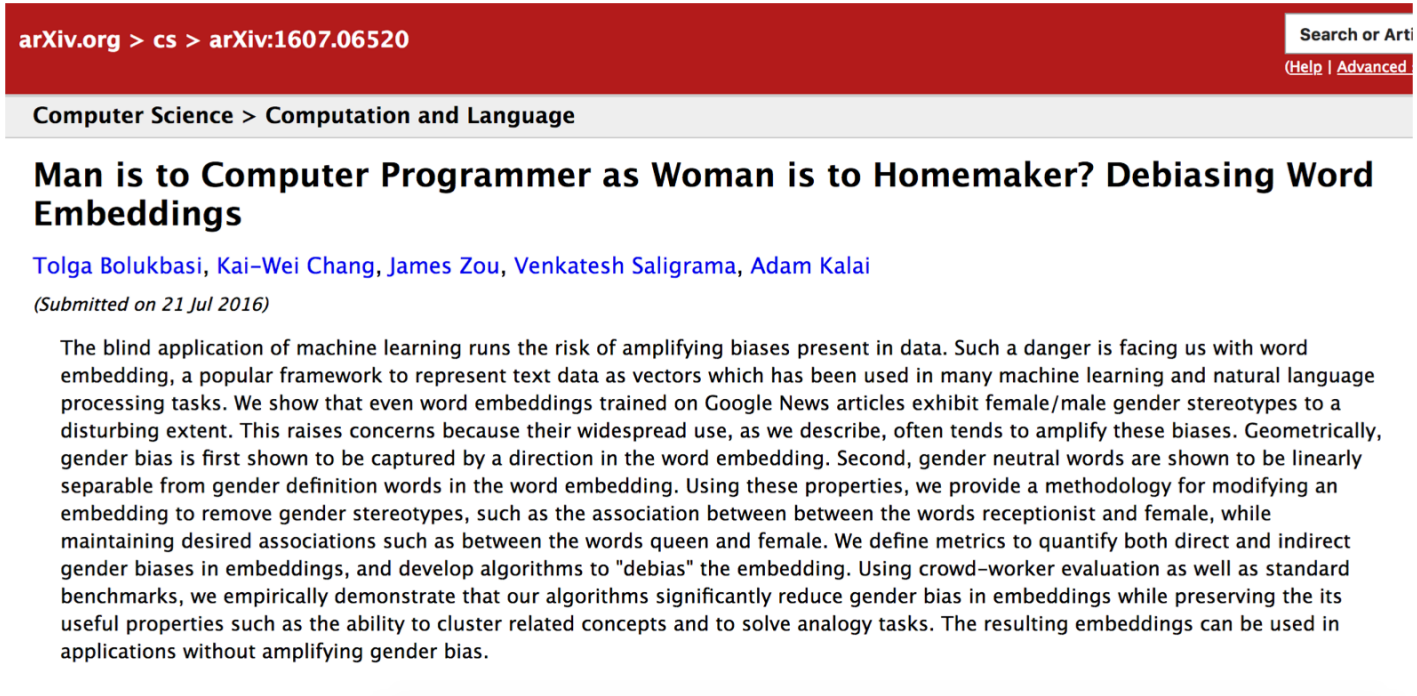


Prediction Fails Differently for Black Defendants

	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.)*

PROPUBLICA STUDY OF NORTHPOINTE SOFTWARE

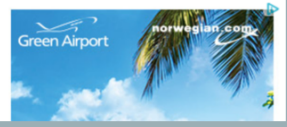


GENDER BIAS IN WORD EMBEDDINGS

Buolamwini and Gebru (2018) found an 8.1% vs 20.6% difference in error rate for male vs female faces and a 11.8% vs 19.2% difference in error rate for lighter vs darker faces in Microsoft, IBM, and Face++ classifiers.

Google Photos Tags Two African-Americans As Gorillas Through Facial Recognition Software

Maggie Zhang, FORBES STAFF
I write about technology, innovation, and startups. FULL BIO



Facial recognition software is biased towards white men, researcher finds

Biases are seeping into software

By Lauren Goode | @LaurenGoode | Feb 11, 2018, 2:00pm EST

Facial-Recognition Software Might Have a Racial Bias Problem

Depending on how algorithms are trained, they could be significantly more accurate when identifying white faces than African American ones.

CLARE GARVIE AND JONATHAN FRANKLE | APR 7, 2016 | TECHNOLOGY

SOURCES OF BIAS

- Over- and under-sampling.
- Skewed sample.
- Feature choice/limited features.
- Proxies/redundant encodings.

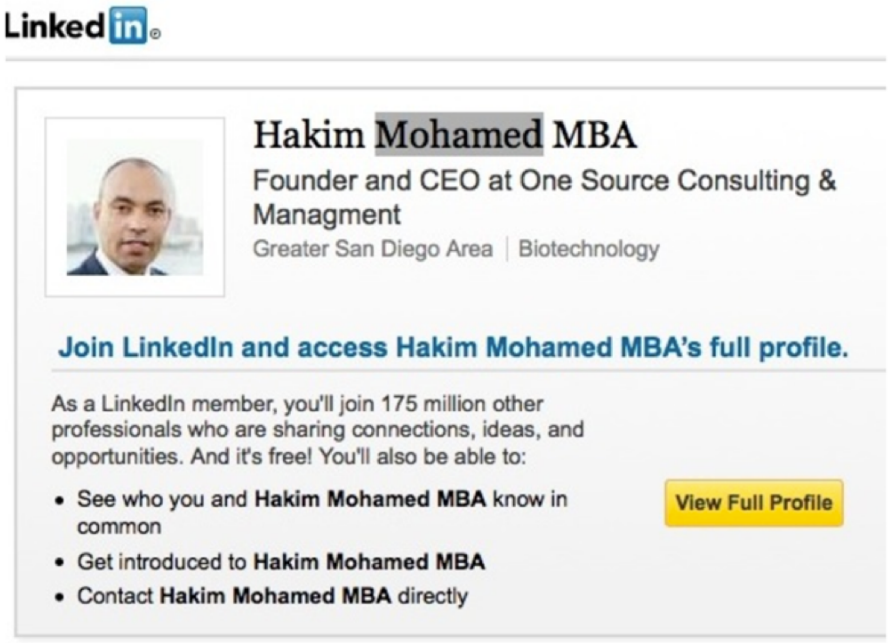
BIAS IN TRAINING DATA

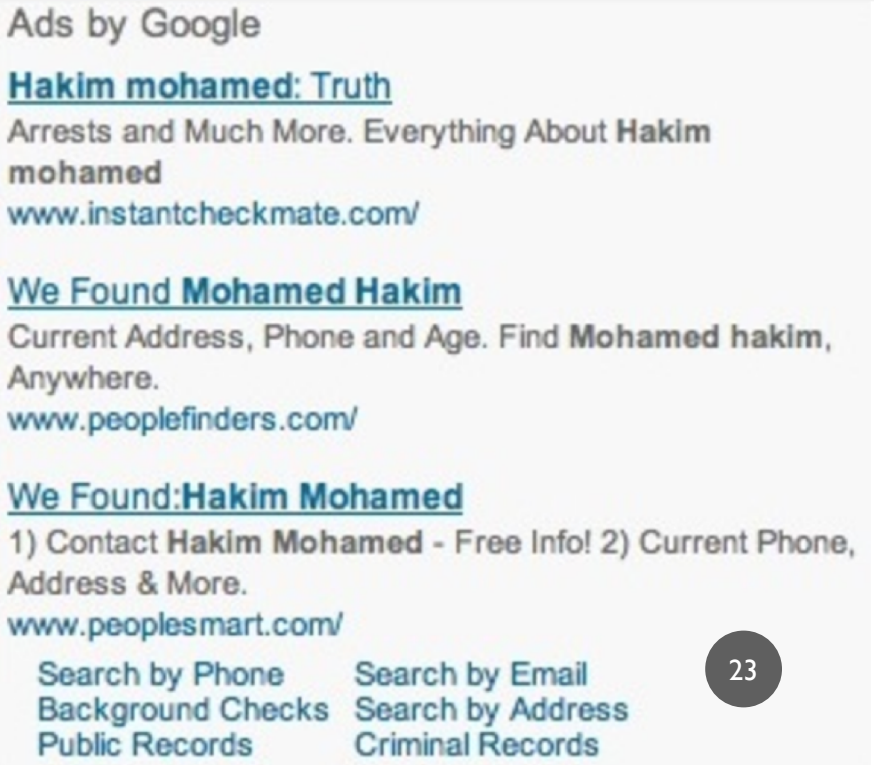
Problem: identifying (or generating) an ***unbiased*** dataset from which to draw generalizations.

Goal: make **accurate** generalizations, in order to do **ethical** machine learning.

A COUNTER-EXAMPLE

Latanya Sweeney. "Discrimination in Online Ad Delivery." *Communications of the ACM* (2013).





A COUNTER-EXAMPLE

Latanya Sweeney. "Discrimination in Online Ad Delivery." *Communications of the ACM* (2013):

LATANYA N. BROWN-ROBERTSON, PHD

Phone 301-860-3661 / Email: lnbrown@bowiestate.edu

Dr. LaTanya N. Brown-

Ads by Google

Latanya Brown, Arrested?
1) Enter Name and State. 2) Access Full Background Checks Instantly.
www.instantcheckmate.com/

Latonya Brown
Get **Latonya Brown** Search for **Latonya Brown**
www.ask.com/**Latonya+Brown**

We Found:**Latanya Brown**
1) Contact **Latanya Brown** - Free Info! 2) Current Phone, Address & More.

A COUNTER-EXAMPLE

Latanya Sweeney. "Discrimination in Online Ad Delivery." *Communications of the ACM* (2013):

GOAL: ACCURATE GENERALIZATIONS

Accuracy about what?

Is accuracy enough?

GOAL: ACCURATE GENERALIZATIONS?

LATANYA N. BROWN-ROBERTSON, PHD

Phone 301-860-3661 / Email: lnbrown@bowiestate.edu

Dr. LaTanya N. Brown-

Ads by Google

Latanya Brown, Arrested?
1) Enter Name and State. 2) Access Full Background Checks Instantly.
www.instantcheckmate.com/

Latonya Brown
Get **Latonya Brown** Search for **Latonya Brown**
www.ask.com/**Latonya+Brown**

We Found:**Latanya Brown**
1) Contact **Latanya Brown** - Free Info! 2) Current Phone, Address & More.

DIAGNOSIS

Accuracy is not enough.

How can we do *ethical* machine learning?

LOOK AT THE PERFORMANCE TASK

- Is the task to be optimized one that contributes to the social good?
  - Is the social good of advertising in terms of generating the most number of clicks?
- Why haven't people pursued this strategy?

OKAY, WE'RE DONE, RIGHT?

1. Make sure we have a performance task that achieve some social good.
2. Make sure we have an unbiased data set.

Hiring at Abercrombie and Fitch. Abercrombie and Fitch have hired a new computer science team to design an algorithm to classify various job applicants. You notice that African-American sales representatives have significantly fewer average sales than white sales representatives. The algorithm's output recommends hiring far fewer African-Americans than white applicants, when the percentage of applications from people of various races are adjusted for.

FIRST QUESTION: IS THIS DISCRIMINATION?

- Recall the disparate impact standard:

- Is this discrimination according to disparate impact?
- Is this discrimination?

ACTIVITY CONTINUED

You have to communicate the results to your employers, and make a recommendation about what to do. What do you say?

LESSON

Characteristics such as race, gender, socio-economic class, etc. determine other features about us that are relevant for to the outcome.

- Average wealth for white families is seven times higher than average wealth for black families.
- Wealth is relevant for whether you can pay back a loan.
- Differences in wealth are determined by historical and present injustice.

COULD MACHINE LEARNING BE PART OF THE PROBLEM?

- Machine learning is, by nature, historical.
- To effectively combat discrimination, we need to *change* these patterns.
- Machine learning, however, reinforces these patterns.
- "Even if history is an arc that bends towards justice, machine learning doesn't bend."

SOLUTIONS

- Sequential learning
- More theory
- Causal modeling
- Optimize for fairness

OPTIMIZING FOR FAIRNESS

Search for a formalized non-discrimination criterion to optimize, along with expected task performance.

FORMALIZING FAIRNESS

- A number of *observational* criteria of fairness have been proposed.
  - These criteria are *oblivious*, i.e., only depend on the joint distribution of the predictor, protected attribute, and outcome.

DEMOGRAPHIC PARITY

- Idea: the decision should be independent of the protected attribute.
  - Race, gender etc. are *irrelevant* to the decision.

- For a binary decision $Y \in \{0,1\}$ and protected attribute $A \in \{0,1\}$:

$$P\{Y=1|A=0\}=P\{Y=1|A=1\}$$

COMMON OBJECTION

Demographic parity rules out using the *perfect predictor* C=Y, where C is the predictor and Y the target variable.

- Say that we want to predict whether an individual will purchase organic shampoo.
- Whether members of certain groups purchase organic shampoo is not independent of their membership of that group.
- So, demographic parity would rule out the perfect predictor.

OTHER OBJECTIONS?

$$P\{Y=1|A=0\}=P\{Y=1|A=1\}$$

Two kinds of objections:

- In principle
- In practice

EQUALIZED ODDS

- The prediction and attribute should be independent, conditional on the outcome.
- For the predictor R, outcome Y, and protected attribute A, where all three are binary variables:

$$P(R=1 \mid A=0, Y=1) = P(R=1 \mid A=1, Y=1).$$

WELL-CALIBRATED

The outcome and protected attribute are independent, conditional on the predictor.

For the predictor R, outcome Y, and protected attribute A, where all three are binary variables:

$$P(Y=1 \mid A=0, R=1) = P(Y=1 \mid A=1, R=1).$$

"Group unaware": Hold everyone to the same standard.

MUTUALLY INCOMPATIBLE STANDARDS

Any two of the three criteria are incompatible (Kleinberg, Mullainathan, Raghavan (2016)).

Let's look at this fact in the context of the debate between ProPublica and Northpointe about whether COMPAS is biased against black defendents.

WELL-CALIBRATED BUT UNEQUAL ODDS

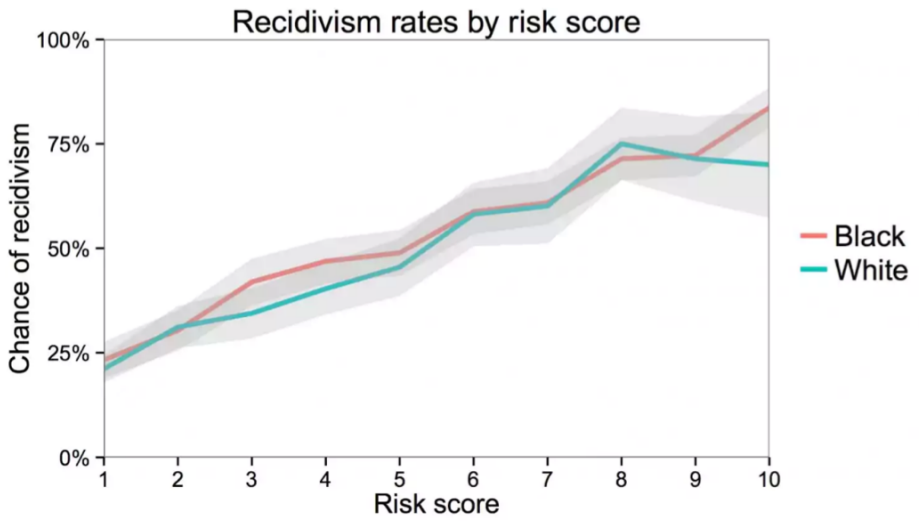
Northpointe's defense: COMPAS is well-calibrated, i.e.,

$$P(Y=1 \mid A=0, R=1) = P(Y=1 \mid A=1, R=1).$$

ProPublica's rejoinder: COMPAS has a higher false positive rate for black defendants and a higher false negative rate for white defendants, i.e., does not satisfy equalized odds:

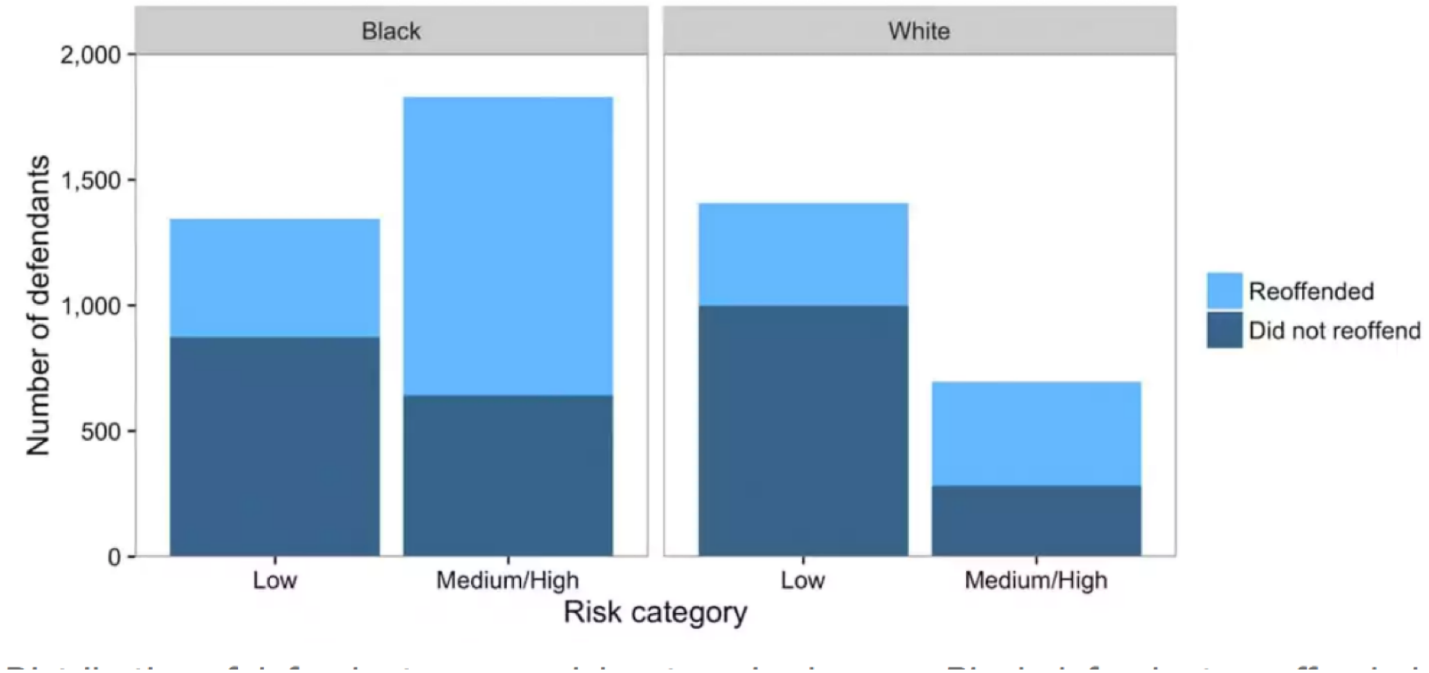
$$P(R=1 \mid A=0, Y=1) \neq P(R=1 \mid A=1, Y=1).$$

WELL-CALIBRATED



Recidivism rate by risk score and race. White and black defendants with the same risk score are roughly equally likely to reoffend. The gray bands show 95 percent confidence intervals.

From https://www.washingtonpost.com/news/monkey-cage/wp/2016/10/17/can-an-algorithm-be-racist-our-analysis-is-more-cautious-than-propublicas/?utm_term=.17f77de3ab45



From https://www.washingtonpost.com/news/monkey-cage/wp/2016/10/17/can-an-algorithm-be-racist-our-analysis-is-more-cautious-than-propublicas/?utm_term=.17f77de3ab45

WHY THIS MATTERS

- It's hard to figure out when certain fairness criteria should apply.
- If some criterion *didn't* come at a cost to the others, then you would worry less about applying one when you're uncertain (at least you aren't incurring unknown costs)!
- But, since this isn't the case, we need to understand the impact of failing to meet at some criteria.

IS THAT SURPRISING?

It seems that what's generating the problem is something we discussed earlier: background facts created by injustice (such as higher rates of being caught re-offending due to higher police scrutiny).

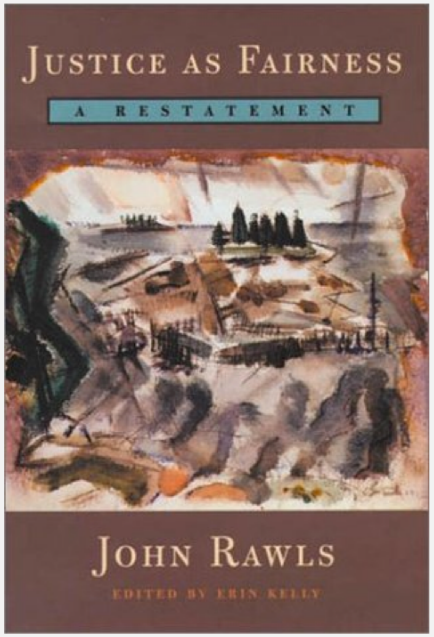
LESSON

- Can't completely formalize fairness in a set of compatible, desirable criteria.
- We need *interpretable systems*.
- We need "traditional" approaches to guaranteeing fairness as well, which focus on institutional design.

FINAL THOUGHT

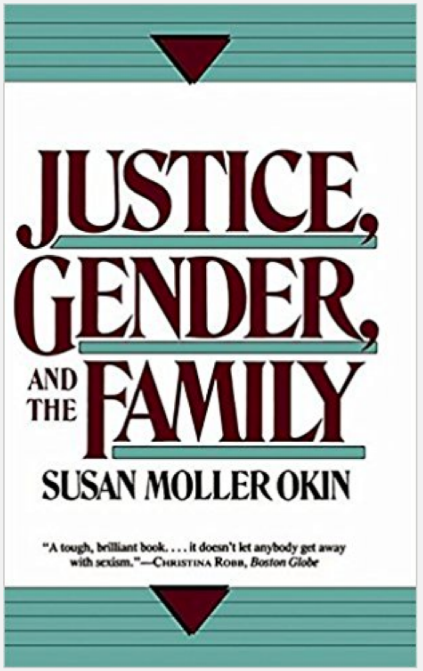
"Optimizing for equal opportunity is just one of many tools that can be used to improve machine learning systems—and mathematics alone is unlikely to lead to the best solutions. Attacking discrimination in machine learning will ultimately require a careful, multidisciplinary approach."

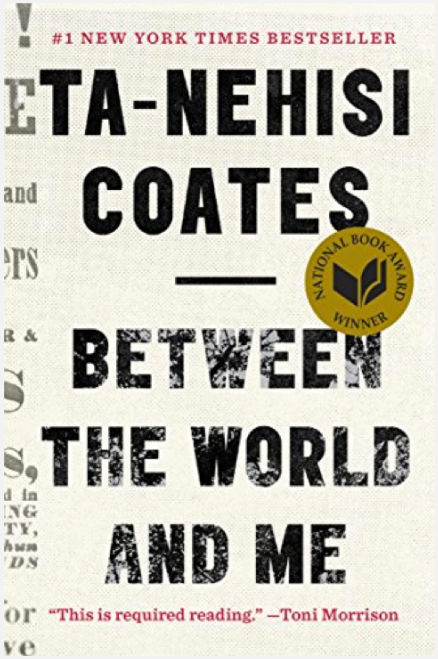
(from https://research.google.com/bigpicture/attacking-discrimination-in-ml/)

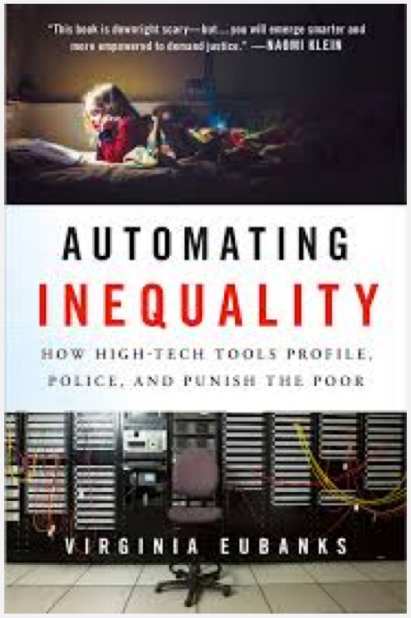


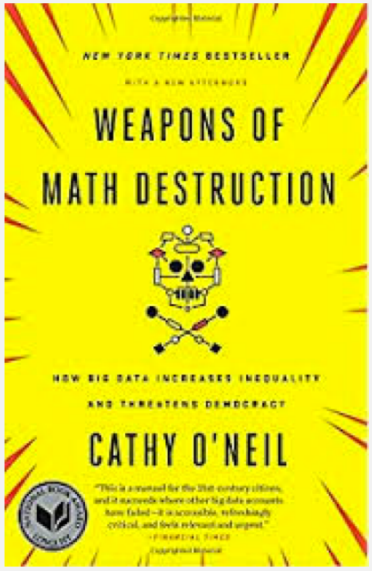
Dissatisfied because only 1/4 (25%) of neighbors are X

X	X	O	X	O
	O	O	O	O
X	X			
X	O	X	X	X
X	O	O		O









RESOURCES

Discussion of threshold classifiers: https://research.google.com/bigpicture/attacking-discrimination-in-ml/.

General resources: https://fairmlclass.github.io/