MACHINE LEARNING AND DISCRIMINATION

Embedded EthiCS module: CS181

Spring 2019

Who am I?

- Cat Wade, philosophy graduate student
- Contact: cmcdonaldwade@g.harvard.edu
- Office hours: by appointment!

DISCRIMINATION, INFORMATION AND DECISION MAKING

The Credit Problem: a case study

THE CREDIT PROBLEM

- Phoebe Robinson and Jessica Williams come to ask you 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
- What should you take into consideration?
  - *Income*
  - *Income to debt ratio*
  - *…Location?*
  - *…Race?*

Is all information fair game?

- Credit is allocated on the basis of information about us
- Is there information that banks and other private companies *shouldn't* be allowed to use in making judgments about creditworthiness?
  - *Why/why not?*
  - *Are there specific kinds of information you think should be **off limits**?*

Is all information fair game?

YES	NO
The assessment of the risk that individuals will play back a loan (their credit worthiness) involves a certain level of uncertainty – *the more information, the more certainty*	People making these decisions are fallible human agents with implicit biases – *removing information of certain types will reduce the implicit bias risk*
In a society where loan repayment *in fact* differs between groups, effective credit assessment may require using facts about someone's race as the basis for a loan decision – *using all information will yield more accurate predictions*	Facts about who can repay a loan are partially determined by facts about past and current injustice – *certain information that might yield an accurate prediction are in fact impermissible to appeal to because that connection in itself is unjust*

What concepts are motivating these responses?

What was behind these responses?

- A conception of a the social good. This will include:
- 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)*

Current uses of Machine Learning

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

- These are all 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?

THE PLAN FOR TODAY

Learning goals:

- Understand the **concept of discrimination** and its variations
- Be able to identify the ways in which machine learning can both **enable** and **prevent discrimination**
- Practice **communicating** about discriminatory algorithms
- **Evaluate** the **trade offs** associated with different ways of **optimizing for fairness** in machine learning

- Class outline:

0. Social goods and uses of machine learning
1. Discrimination
2. Machine learning, accuracy and discrimination
   - *Understanding discrimination as inaccuracy*
   - *Discrimination despite accuracy*
3. Discrimination beyond accuracy
   - *Evaluating machine learning in terms of performance tasks*
4. ACTIVITY
5. Optimizing machine learning for fairness
   I. *Formalizing a non-discrimination criterion*
   II. *Demographic parity*
   III. *Equalizing odds*
   IV. *Well-calibrated systems*
6. Concluding discussion

DISCRIMINATION

What is it? Who does it? Against whom? When and why is it morally problematic?

What is discrimination?

1. Disparate treatment
2. Disparate impact

What is discrimination?

1. Disparate treatment
2. Disparate impact

- Involves classifying someone in an impermissible way
  - *Some types of classification are morally neutral, whereas other types are morally problematic*
- Involves the *intent* to discriminate
  - *Evidenced by explicit reference to group membership*

What is discrimination?

1. Disparate treatment
2. Disparate impact

- Looks at the *consequences* of classification/decision making on certain groups
  - *No intent required*
  - *Facially neutral*
- Practices with a disproportionate impact on a particular group **do not** cause disparate impact if they are "grounded in sound business considerations." (http://www.scotusblog.com/2015/06/paul-hancock-fha/)

What is discrimination?

1. Disparate treatment
2. Disparate impact

PROTECTED ATTRIBUTES:
- Age
- Disability
- National Origin
- Race/color
- Religion
- Sex
- (From the US Equal Opportunity Employment Commission)

- Looks at the *consequences* of classification/decision making on certain groups
  - *No intent required*
  - *Facially neutral*
- Practices with a disproportionate impact on a particular group **do not** cause disparate impact if they are ”grounded in sound business considerations.” (http://www.scotusblog.com/2015/06/paul-hancock-fha/)

What does this mean for Machine Learning?

- Sometimes Machine Learning algorithms discriminate
  - *What can we do to help combat this?*
  - *E.g., facial recognition*

- Sometimes Machine Learning algorithms can help resolve discrimination in other domains
  - *E.g., hiring*

MACHINE LEARNING, ACCURACY, AND DISCRIMINATION

Discrimination as a lack of accuracy?

Data as an antidote to discrimination?

- Motivating idea: discriminatory bias makes classification *less accurate*
  - *Discrimination impacts social goods when classification and decision making is based on **inaccurate information***
    - E.g. thinking that everyone over 7ft is a bad babysitter
  - *We can remedy this discrimination by building more accurate models*
- The greater the accuracy the less we risk discrimination
  - *How can we increase the accuracy of the model?*
    - Alternatives to fuzzy, biased human reasoning

DATA AS AN ANTIDOTE?

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."

DATA AS AN ANTIDOTE?

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)

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)

Upshot: sometimes machine learning algorithms do a better job than we would at making the most **accurate** classifications.

Further: sometimes this will **combat** discrimination in domains like hiring and credit approval.

BIAS IN TRAINING DATA (1)

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.

PROPUBLICA STUDY OF NORTHPOINTE SOFTWARE

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.)*

"Overall, Northpointe's assessment tool correctly predicts recidivism 61% 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."

BIAS IN TRAINING DATA (1)

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.

PROPUBLICA STUDY OF NORTHPOINTE SOFTWARE

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.)*

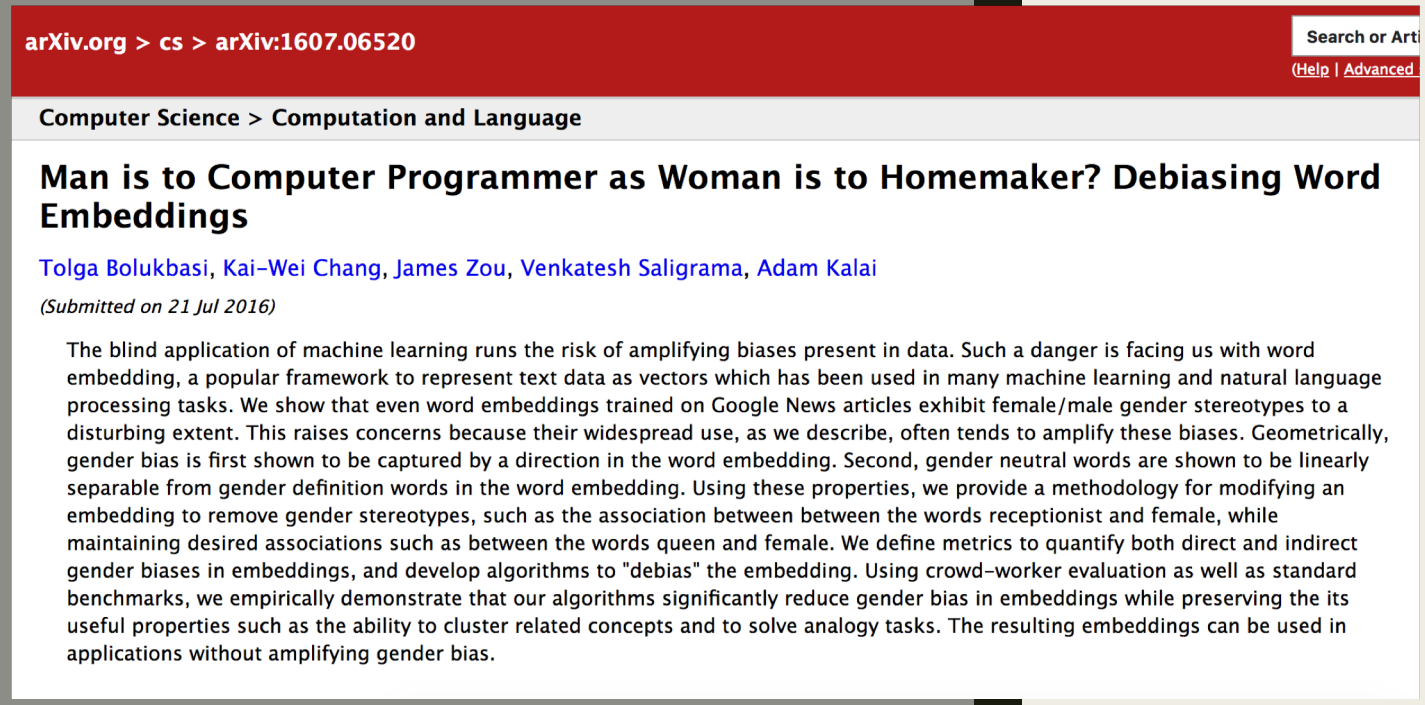
"Overall, Northpointe's assessment tool correctly predicts recidivism 61% of the time. But blacks are almost twice as likely as whites to be labeled a higher risk but not actually re-offend. It makes the opposite mistake among whites: They are much more likely than blacks to be labeled lower risk but go on to commit other crimes."

Discriminatory machine learning algorithms: bias in training data

- In this case, the biases of humans are **not** mitigated by the machine learning algorithm
  - *In fact, they are **reproduced** in the classifications that are made*
- Why does this happen?
  - *A machine learning system may be trained on **data infused with human bias***
  - *Recidivism scores such as those made by the Northpointe software are based on prior arrests, age of first police contact, parents' incarceration record*
    - This information is shaped by biases in the world and injustices more generally
- O'Neil, Cathy. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown, 2016.

BIAS IN TRAINING DATA (2)

- Automatically generated analogies from the software's vectors, such as man:woman::computer-programmer:home maker
- These reflect sexism in the original texts



arXiv.org > cs > arXiv:1607.06520

Search or Arti

(Help | Advanced

Computer Science > Computation and Language

Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, Adam Kalai

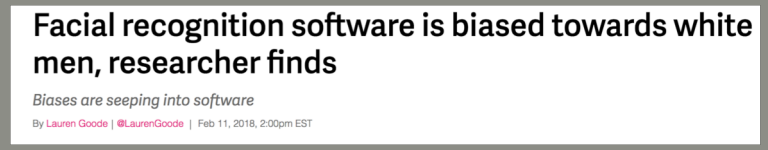
(Submitted on 21 Jul 2016)

The blind application of machine learning runs the risk of amplifying biases present in data. Such a danger is facing us with word embedding, a popular framework to represent text data as vectors which has been used in many machine learning and natural language processing tasks. We show that even word embeddings trained on Google News articles exhibit female/male gender stereotypes to a disturbing extent. This raises concerns because their widespread use, as we describe, often tends to amplify these biases. Geometrically, gender bias is first shown to be captured by a direction in the word embedding. Second, gender neutral words are shown to be linearly separable from gender definition words in the word embedding. Using these properties, we provide a methodology for modifying an embedding to remove gender stereotypes, such as the association between between the words receptionist and female, while maintaining desired associations such as between the words queen and female. We define metrics to quantify both direct and indirect gender biases in embeddings, and develop algorithms to "debias" the embedding. Using crowd-worker evaluation as well as standard benchmarks, we empirically demonstrate that our algorithms significantly reduce gender bias in embeddings while preserving the its useful properties such as the ability to cluster related concepts and to solve analogy tasks. The resulting embeddings can be used in applications without amplifying gender bias.

BIAS IN TRAINING DATA (3)

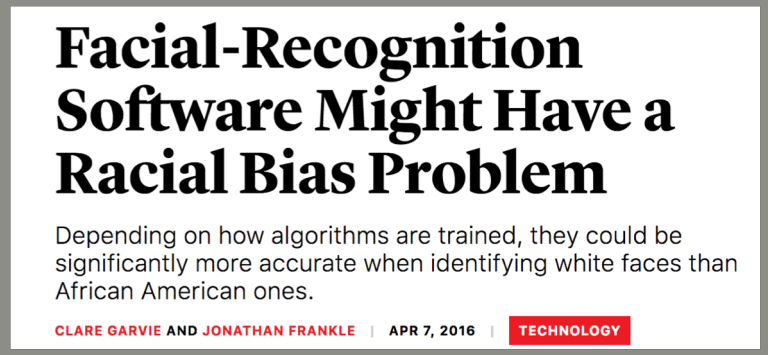

Google Photos Tags Two African-Americans As Gorillas Through Facial Recognition Software

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

- 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.
- In all these cases bias is in the data set

SOURCES OF BIAS:

- Over- and under-sampling.
- Skewed sample.
- Feature choice/limited features.
- Proxies/redundant encodings.
- Biases and injustices in the world

Bias in training data

- From these cases (1)-(3) we get the following:
  - *Machine learning algorithms can perpetuate discrimination because they are trained on biased data*
- Solution?
  - *Identify or generate an **unbiased** dataset from which to draw **accurate** generalizations*

- Upshot of this section: our overall goal should be to make accurate generalizations in order to do ethical machine learning

DISCRIMINATION DESPITE ACCURACY

Instances of discrimination in the absence of inaccuracy

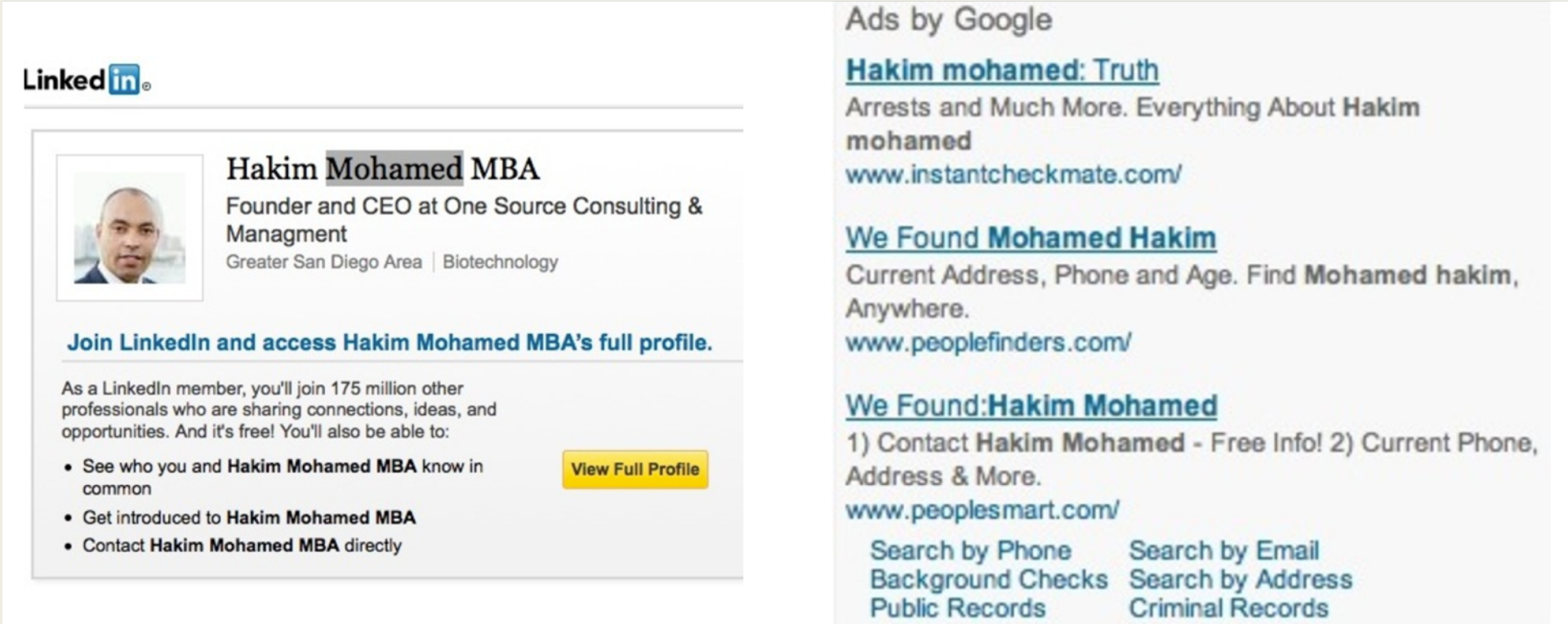
Latanya Sweeney. "Discrimination in Online Ad Delivery." *Communications of the ACM* (2013).

- What is the aim of online advertising?
  - *Get clicks*
- How do algorithms figure out what advertisements will get the most clicks?
  - *"At first, all possible ad texts are weighted the same and are presumed equally likely to produce a click. Over time, as people click one version of an ad more often than others, the weights change, so the ad text getting the most clicks eventually displays more frequently. This approach aligns the financial interests of Google, as the ad deliverer, with the advertiser"*

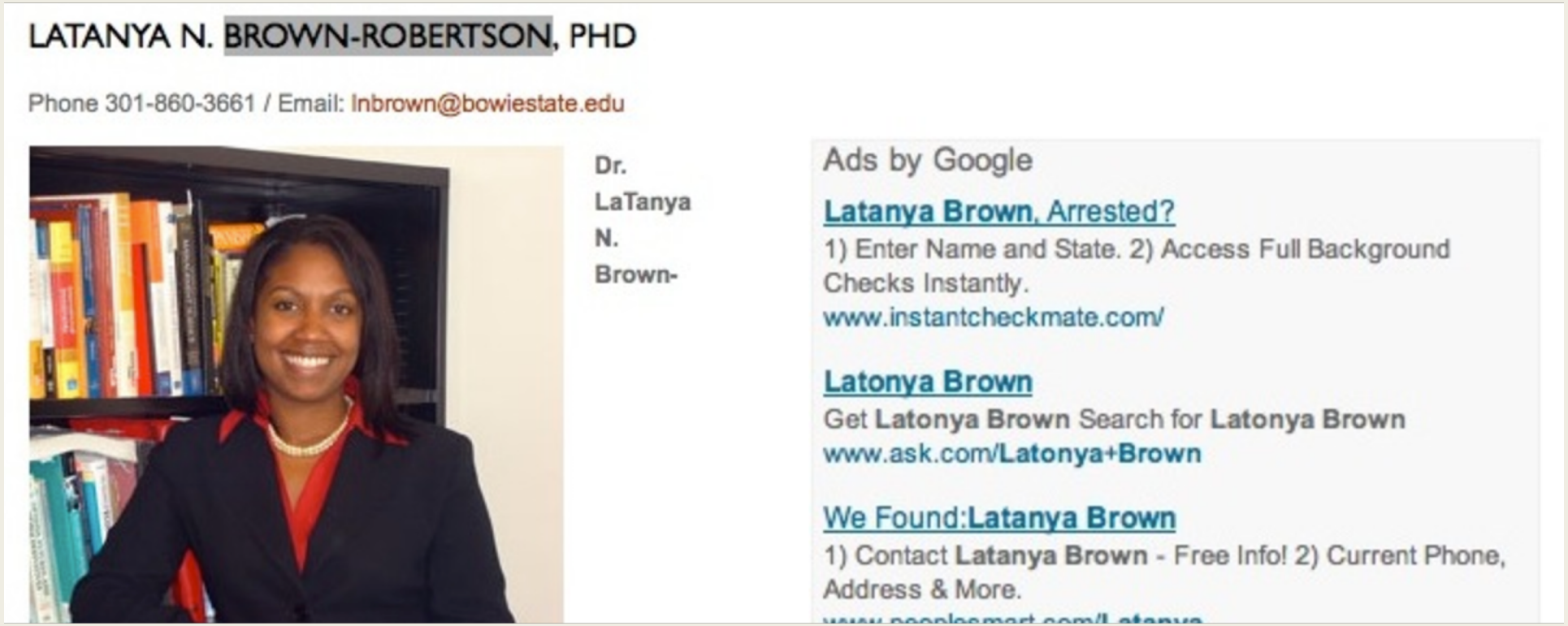
Latanya Sweeney. "Discrimination in Online Ad Delivery." *Communications of the ACM* (2013).

- The dataset is being generated from actual user behavior - unbiased
  - *Compare this to our previously biased data sets:*
    - Unbalanced arrest records in part due to over-policing
    - Sexism in written texts
    - Lack of racial diversity of faces in facial recognition training data
- Any classifications made on the basis of this data set will be accurate, relative to the generated data set
  - *Given the search query, show the advertisement that has in the past generate the most clicks*

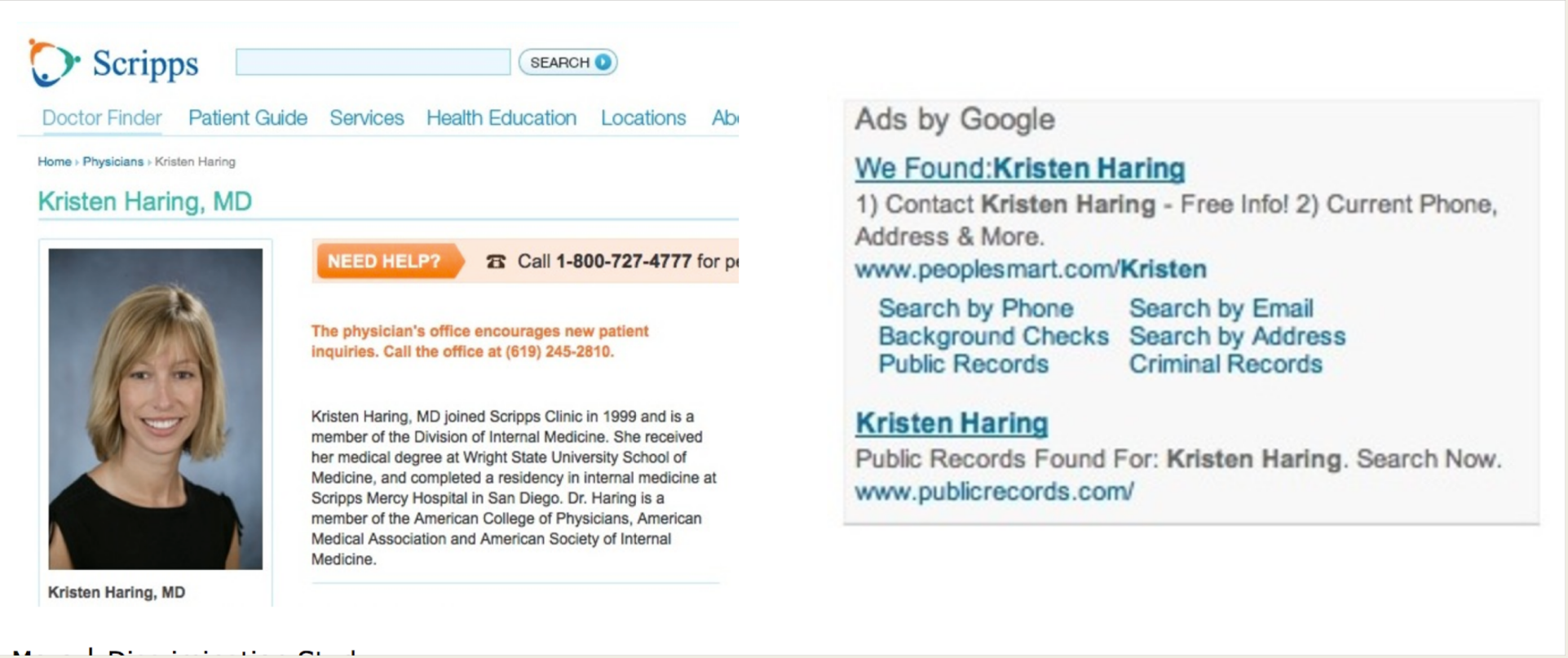
Latanya Sweeney. “Discrimination in Online Ad Delivery.” *Communications of the ACM* (2013).



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



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



What should the goal be?

- Our previous suggestion: make accurate generalizations in order to do ethical machine learning
  - *What does this case show us about accuracy and discrimination?*

- Some questions:
  - *Accuracy about what?*
  - *Is accuracy enough?*
  - *What should the performance task be?*

DISCRIMINATION BEYOND ACCURACY

How *else* can we understand *ethical* machine learning?

Look at the performance task

- Step (1): is the task to be optimized one that contributes to the social good?
  - *Directly*
  - *Indirectly*
- Step (2): can the task to be optimized by appropriated to contribute to social harms?
  - *Directly*
  - *Indirectly*
- Step (3): if the answer to step two is “yes”, what steps can be taken to mitigate those harmful effects?
  - *Question: if the performance task for advertisements isn’t just to get clicks, what could it be?*

So we’re done right?

1. Make sure we have a performance task that achieves some social good and meets the criteria we mapped out in steps (1)-(3)
2. Make sure we have an unbiased data set.

= we get A+ ethical machine learning systems?

ACTIVITY

Hiring at Forever 28. Forever 28 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.

ACTIVITY

Q1: is this discrimination?

- *Would it be disparate impact or disparate treatment?*
- *Recall the disparate impact standard:*
  - No intent required
  - Facially neutral
  - NOT disparate impact if differential treatment is ”grounded in sound business considerations”

Q2: Does this meet our other two criteria?

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

Q3: You have to communicate the results to your employers, and make a recommendation about what to do.

- *What do you say?*

ACTIVITY

Hiring at Forever 28. Forever 28 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.

Q1: is this discrimination?

Would it be disparate impact or disparate treatment?

Recall the disparate impact standard:

- No intent required
- Facially neutral
- NOT disparate impact if differential treatment is "grounded in sound business considerations"

Q2: Does this meet our other two criteria?

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

Q3: You have to communicate the results to your employers, and make a recommendation about what to do.

What do you say?

5 MINUTES TO DISCUSS

What are the lessons here?

- Characteristics such as race, gender, socio-economic class, etc. determine other features about us that are relevant to the outcome of some performance tasks.
- These *are* protected attributes, but they're *still* relevant to certain performance tasks – and performance tasks that are putative social goods at that.

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

What are the lessons here?

- 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."
- Machine learning may therefore be part of the problem

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

What now?

- Even when we optimize for accuracy, machine learning algorithms may perpetuate discrimination *even when we:*
  - *Work from an unbiased data set*
  - *Have a performance task that has social goods in mind*

- What else could we do?
  1. *Sequential learning*
  2. *More theory*
  3. *Causal modelling*
  4. *Optimizing for fairness*

What now?

- Even when we optimize for accuracy, machine learning algorithms may perpetuate discrimination *even when we:*
  - *Work from an unbiased data set*
  - *Have a performance task that has social goods in mind*

- What else could we do?
  1. *Sequential learning*
  2. *More theory*
  3. *Causal modelling*
  4. *Optimizing for fairness*

OPTIMIZING FOR FAIRNESS

Building machine learning algorithms that optimize for non-discrimination

Optimizing for fairness: 4 approaches

1. Formalizing a non-discrimination criterion
2. Demographic parity
3. Equalized odds
4. Well-calibrated systems

Optimizing for fairness (1): formalizing a non-discrimination criterion

- Identify a formalized non-discrimination criterion to optimize for, along with expected task performance
  - *What sorts of criteria might these be?*

- OBJECTION: these criteria might themselves be biased/culturally relative/difficult to quantify

Optimizing for fairness (2): demographic parity

- Idea: the decision should be **independent of protected attributes**.
  - *Race, gender etc. are **irrelevant** to the decision.*
- For a binary decision **Y** and protected attribute **A**:

$$P(Y=1|A=0) = P(Y=1|A=1)$$

The probability of some decision being made (Y=1) should be the same, regardless of the protected attribute (whether (A=1) or (A=0))

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

Optimizing for fairness (2): demographic parity

- 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.*
  - *But, demographic parity would rule out using the perfect predictor.*
- QUESTION: this is usually taken to be a devastating objection to demographic parity – but what do you think, especially in light of the *Forever 28* example?
  - *Hint: are there some cases where perfect predictors should be omitted on other grounds?*

Optimizing for fairness (3): equalized odds

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

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

The attribute (whether (A=1) or (A=0)) should not change your estimation (P) of how likely it is that some relevant predictor (R=1) holds true of the candidate. Instead, the outcome (of some decision) (Y=1) should.

ADVANTAGE: this *is* compatible with the ideal predictor, R=Y

Optimizing for fairness (3): equalized odds

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

Predictor R = Whether you were high school valedictorian (1) or not (0)
Outcome Y = Getting into Yale (1) or not (0)
Attribute A = Being gay (1), being straight (0)

<u>Buffy runs into Willow at *The Bronze*</u>

Willow: hey Buffy I got into Yale (Y=1)

Buffy: that's sweet Willow, I bet you were your high school's valedictorian (R=1)

Willow: hang on a sec, I have to tell you something – I'm gay (A=1)

Buffy: okay cool, well that doesn't make a difference to my estimation of whether you were your high school's valedictorian. (whether A=1 or A=0 doesn't make a difference to P(R=1) for Buffy, it's the outcome (Y=1) that does)

Optimizing for fairness (4): well-calibrated systems

- 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:

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

The probability of some outcome occurring (Y=1) should be unaffected by some protected attribute (whether (A=0) or (A=1)), and instead should be conditional on the relevant predictor (R=1)

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

Optimizing for fairness (4): well-calibrated systems

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

Predictor R = Whether you were high school valedictorian (1) or not (0)
Outcome Y = Getting into Yale (1) or not (0)
Attribute A = Being gay (1), being straight (0)

Buffy runs into Willow at *The Magic Box*

Buffy: hey Will, do you think I'll get into Yale? (P(Y=1))

Willow: Well, were you valedictorian at high school? (does (R=1))

Buffy: heck yeah I was, I hope that relevant fact about me and no facts about my sexuality (whether (A=0) or (A=1)) influence that outcome!

Willow: let's hope Yale uses a well-calibrated admissions algorithm

Mutually incompatible standards: well-calibrated but unequal odds

- PROBLEM: sometimes, given certain empirical circumstances, we cannot have a system be BOTH well-calibrated AND equalize the odds
- Let's look at this fact in the context of the debate between ProPublica and Northpointe about whether COMPAS is biased against black defendants:

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:

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

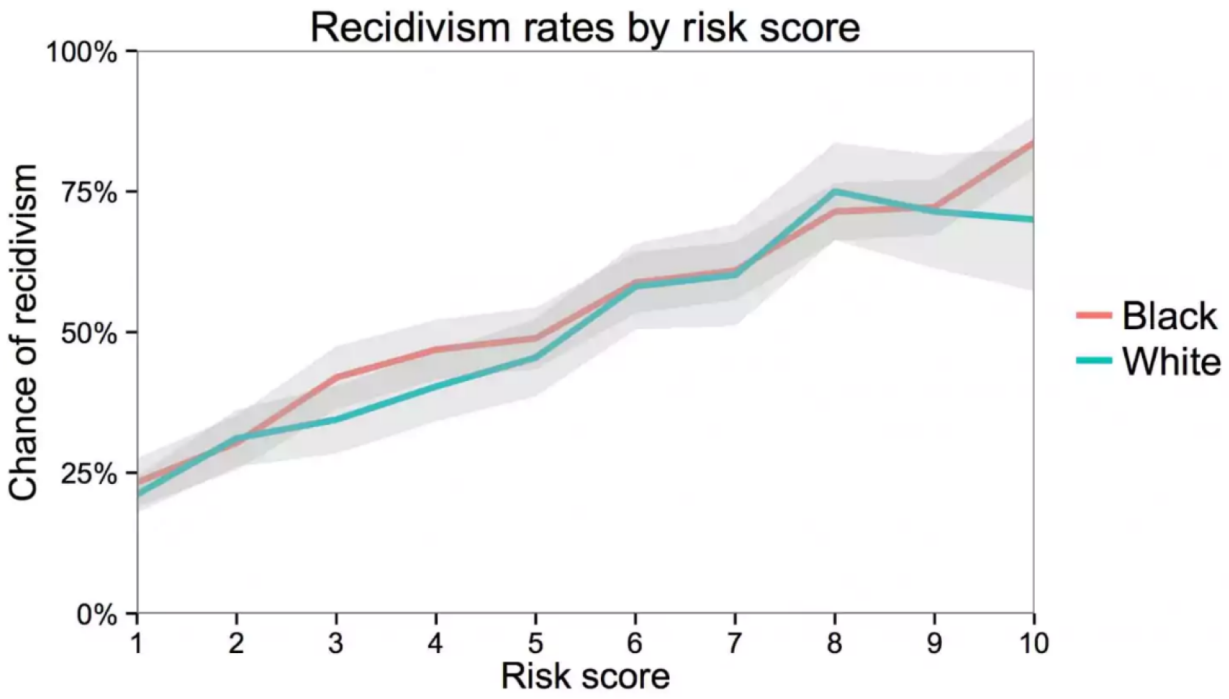
(Kleinberg, Mullainathan, Raghavan (2016)).

WELL-CALIBRATED

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

Y = whether the defendant will reoffend
A = race of the defendant
R = recidivism predictor used by COMPAS

The COMPAS system makes roughly similar recidivism predictions for defendants, regardless of their race.



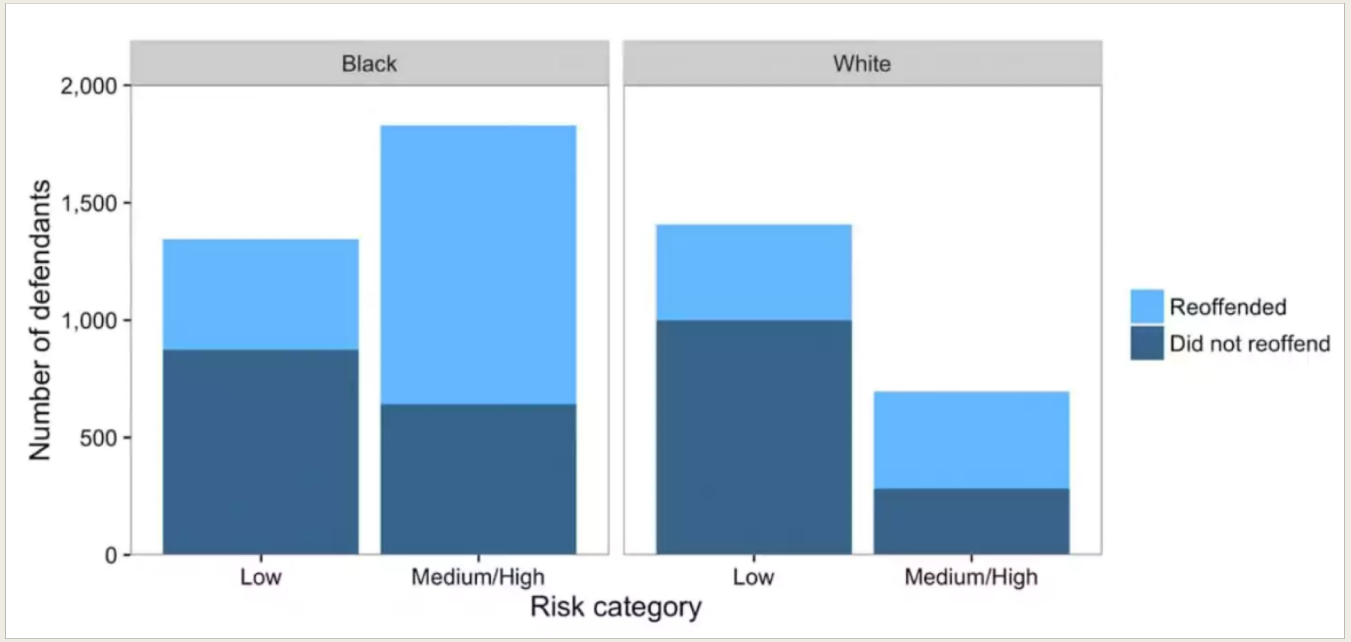
Recidivism rate by risk score and race. White and black defendants with the same risk score are roughly equally likely to reoffend. The gray bands show 95 percent confidence intervals.

EQUALIZED ODDS?

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

Y = whether the defendant will reoffend
A = race of the defendant
R = recidivism predictor used by COMPAS

<u>Michael meets Chad, who is blind, outside the courthouse</u>
Michael: Chad, COMPAS has determined that I am very likely to reoffend (Y=1)
Chad: that sucks, you must have some feature that the COMPAS recidivism algorithm uses to predict high likelihood of reoffending (R=1)
Michael: have I mentioned that I'm black?
Chad: oh, well, that's extremely likely to change the likelihood COMPAS puts on your reoffending



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

The race of the defendant makes a difference to whether the individual is placed in the low or the medium/high risk category.

Whether (A=0) or (A=1) makes a difference to the probability that COMPAS has identified some recidivism risk predictor will hold of the defendant (P(R=1)), and not just whether the defendant will/won't reoffend (Y=1)

Why did this happen?

- When certain empirical facts hold, our ability to have a well-calibrated and an odds-equalizing system breaks down
- It seems that what's generating the problem is something we discussed earlier: background facts created by injustice
  - *For example, higher rates of being caught re-offending due to higher police scrutiny*
- 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.

CONCLUDING THOUGHTS

Where does this leave optimizing for fairness?

So what now?

- Recall our 4 approaches to optimizing for fairness:

1. Formalizing a non-discrimination criterion
2. Demographic parity
3. Equalized odds
4. Well-calibrated systems

All of these approaches have promising features, but all have their drawbacks

Where does this leave us re: discrimination?

- **Conclusion 1:** we can't section off fairness in one little corner without fighting to change injustices in the world and discrimination that happens outside of machine learning systems
- This doesn't mean we can't do anything!
  - *Set some standards for fairness in certain domains, while at the same time striving to change base rates*

So what now?

- Recall our 4 approaches to optimizing for fairness:

1. Formalizing a non-discrimination criterion
2. Demographic parity
3. Equalized odds
4. Well-calibrated systems

All of these approaches have promising features, but all have their drawbacks

Where does this leave us re: discrimination?

- **Conclusion 2:** the approach we opt for is going to depend on the kind of discrimination we're hoping to counter

QUESTION: which of the approaches (1)-(4) do you think would be best to counteract:

- DISPARATE TREATMENT
- DISPARATE IMPACT

And why?

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/)

Thank you!

- Feedback: http://bit.ly/cs181ethics
- cmcdonaldwade@g.harvard.edu
  - *Reading recommendations*
  - *Philosophy class suggestions*
  - *Buffy fandom*