ALGORITHMIC FAIRNESS

Kristina Lerman

USC Information Sciences Institute

DSCI 552 – Spring 2021

March 3, 2021

Topics

- Bias in data
  - Sources of bias in data
  - Understand the impact of bias on data analysis
  - Learn how to evaluate bias in data
  - Computational strategies to mitigate bias in data
- Algorithmic fairness
  - What is fairness in AI?
  - Bias in data and algorithmic fairness
  - Measures of algorithmic fairness; the impossibility of total fairness
  - Methods: Improving fairness by debiasing data

The promise and the perils of AI

AI eliminates human biases from the decision process

Algorithms are only as good as the data they were trained on

- Data inherits implicit and explicit **biases** persisting in society

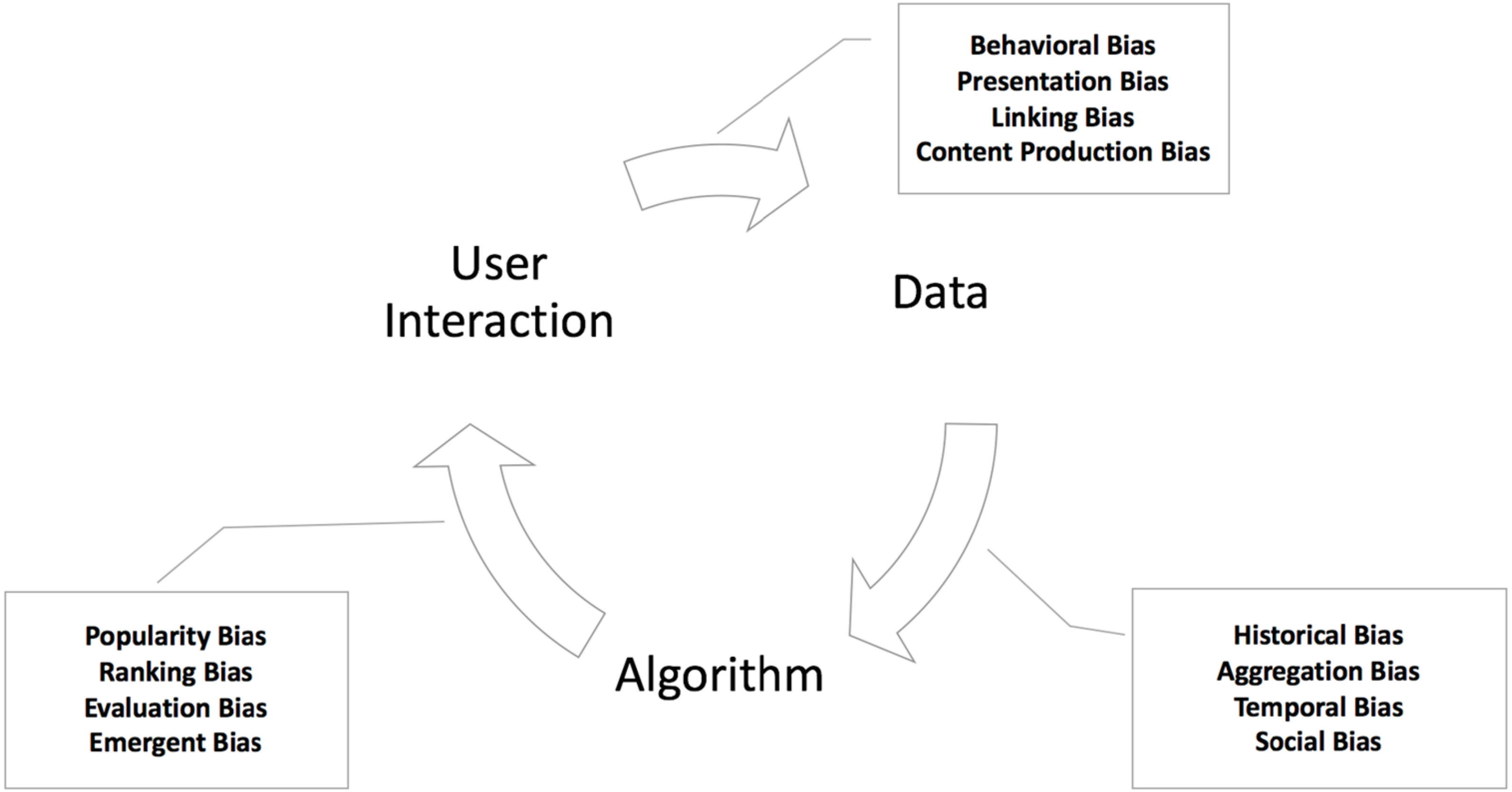
- AI trained on biased data may **discriminate** against protected classes

- AI may reproduce and **amplify** existing discrimination, exclusion and inequality

"Unthinking reliance on data mining can deny historically disadvantaged and vulnerable groups full participation in society"

[Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. *Calif. L. Rev.*, *104*, 671.]

Biases amplify each other



User Interaction

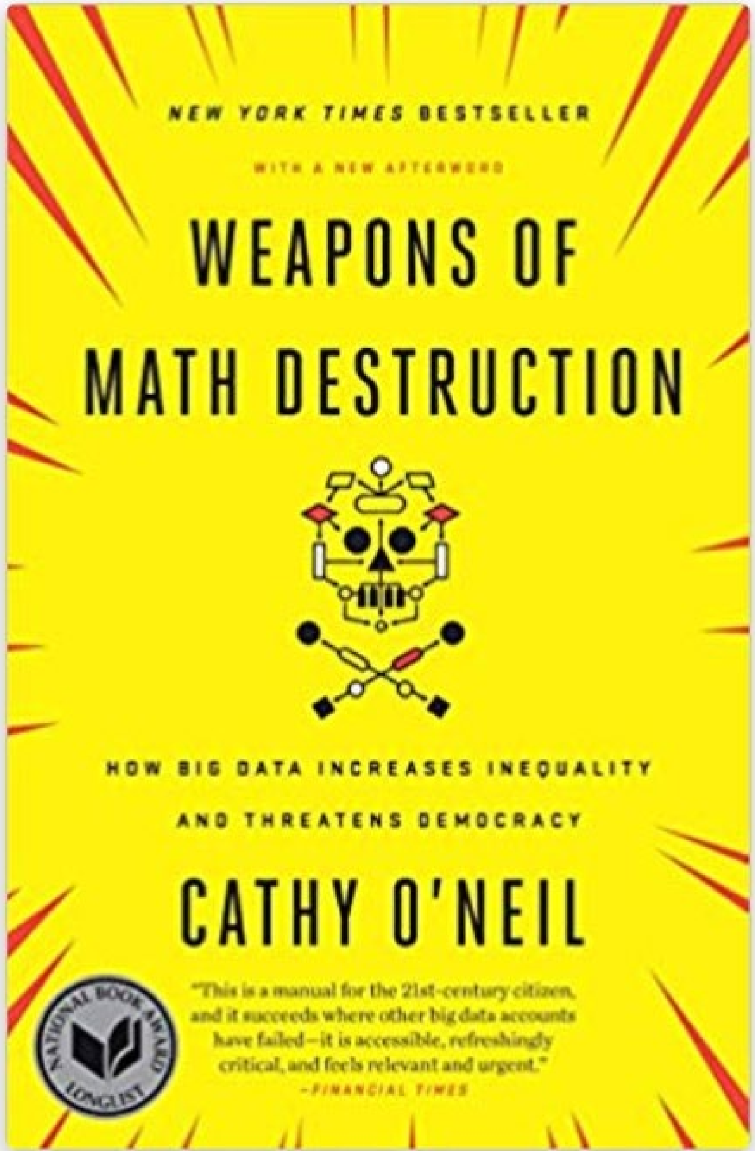
Data

Algorithm

Behavioral Bias
Presentation Bias
Linking Bias
Content Production Bias

Historical Bias
Aggregation Bias
Temporal Bias
Social Bias

Popularity Bias
Ranking Bias
Evaluation Bias
Emergent Bias



NEW YORK TIMES BESTSELLER

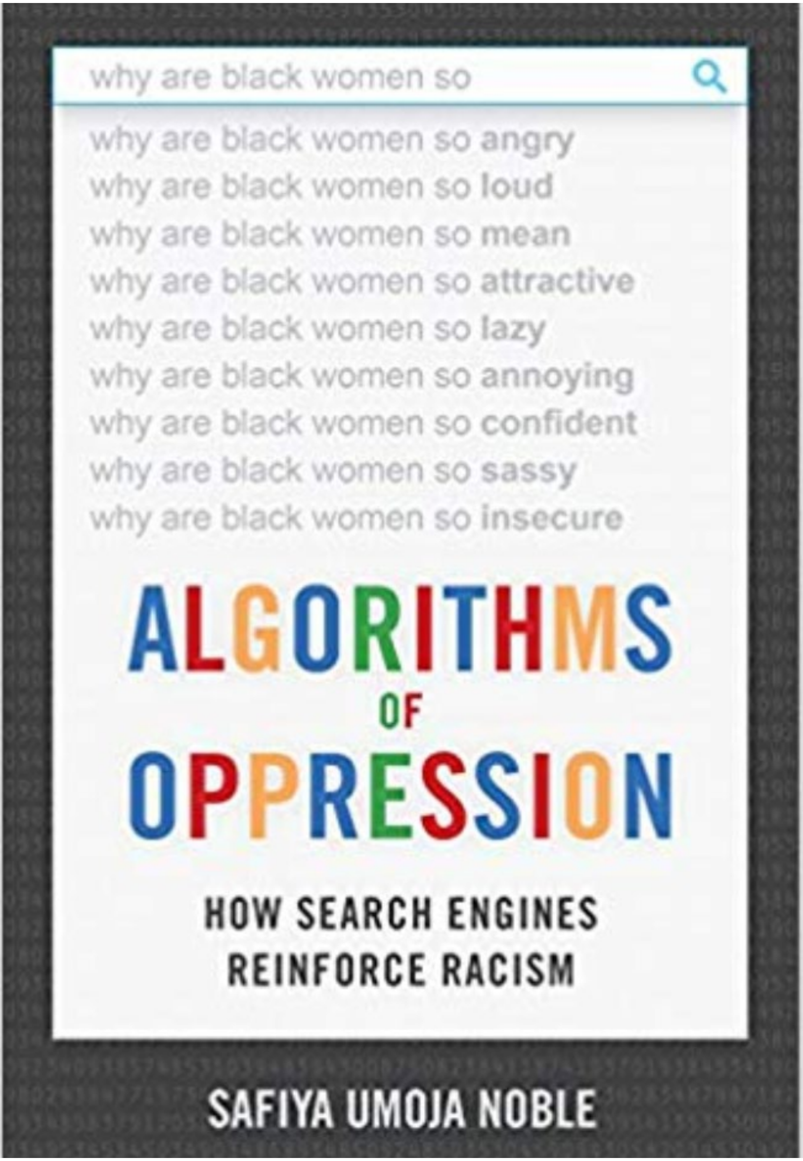
WITH A NEW AFTERWORD

WEAPONS OF MATH DESTRUCTION

HOW BIG DATA INCREASES INEQUALITY AND THREATENS DEMOCRACY

CATHY O'NEIL

"This is a manual for the 21st-century citizen, and it succeeds where other big data accounts have failed—it is accessible, refreshingly critical, and feels relevant and urgent."
—FINANCIAL TIMES



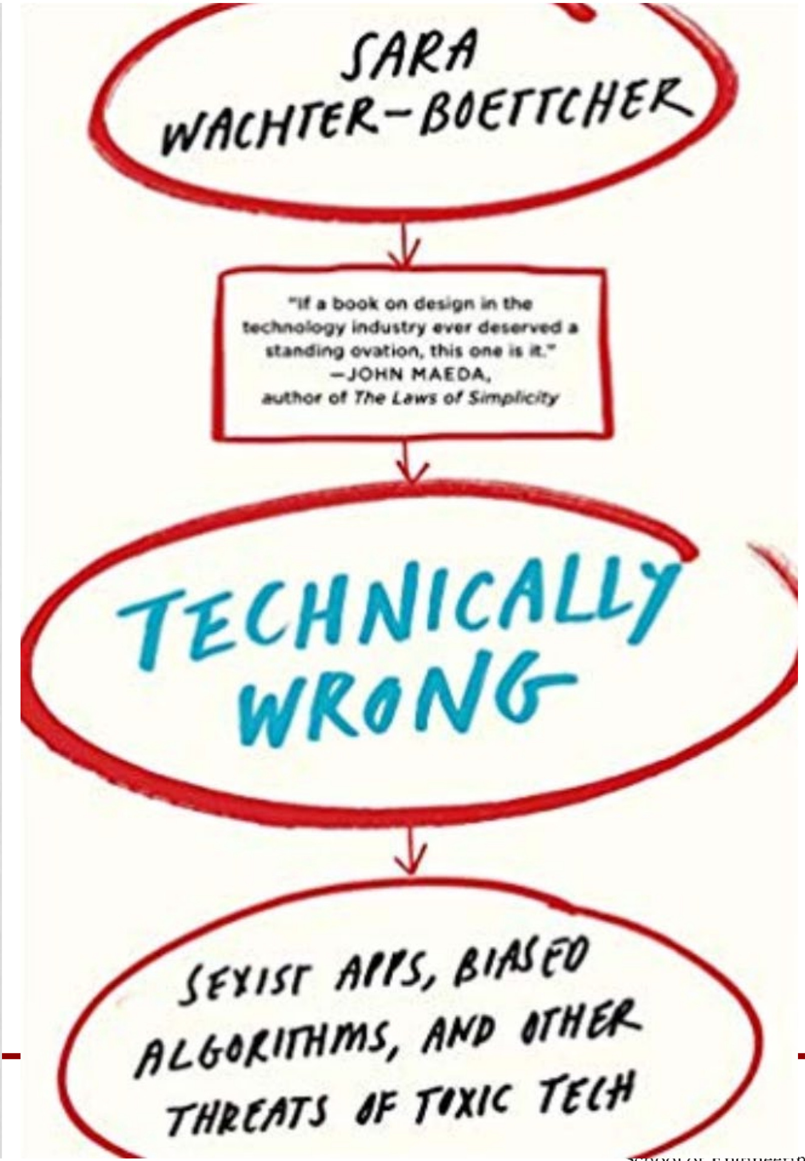
why are black women so

why are black women so angry
why are black women so loud
why are black women so mean
why are black women so attractive
why are black women so lazy
why are black women so annoying
why are black women so confident
why are black women so sassy
why are black women so insecure

ALGORITHMS OF OPPRESSION

HOW SEARCH ENGINES REINFORCE RACISM

SAFIYA UMOJA NOBLE



SARA WACHTER-BOETTCHER

"If a book on design in the technology industry ever deserved a standing ovation, this one is it."
—JOHN MAEDA,
author of *The Laws of Simplicity*

TECHNICALLY WRONG

SEXIST APPS, BIASED ALGORITHMS, AND OTHER THREATS OF TOXIC TECH

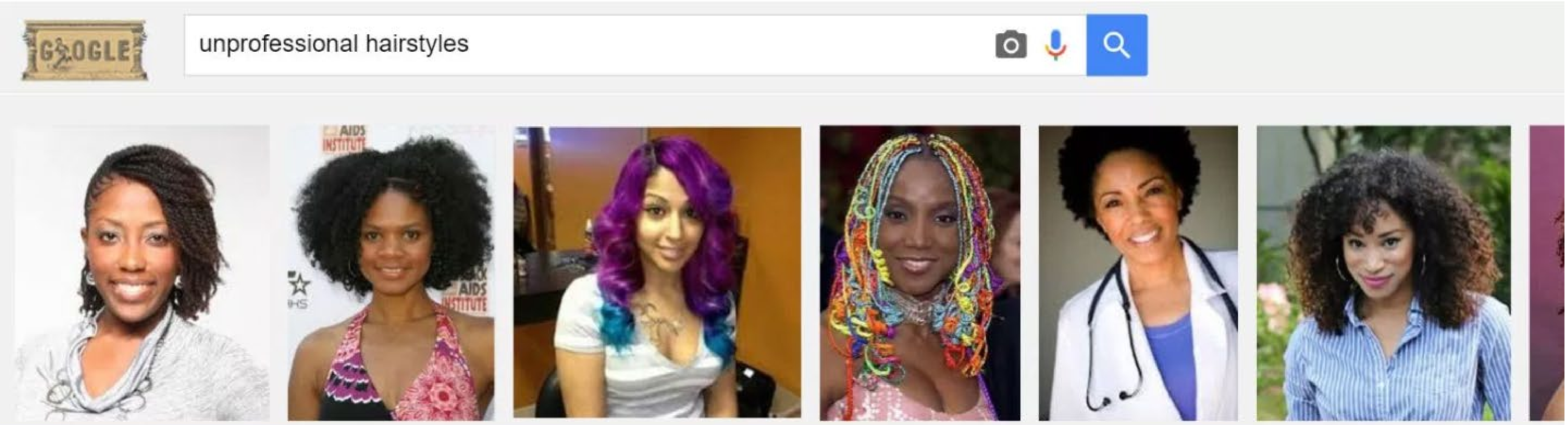
- “big data analytics have the potential to eclipse longstanding civil rights protections in how personal information is used in housing, credit, employment, health, education, and the marketplace.”
    - White House report *Big Data: Seizing Opportunities, Preserving Values*, 2014

Protected classes

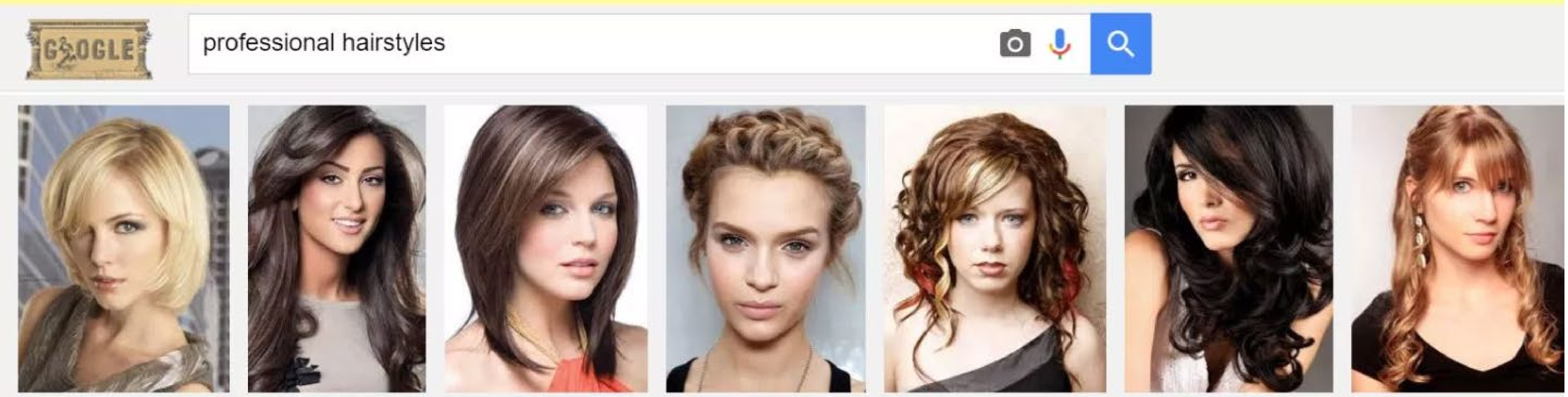
US Federal laws define protected classes to include:

- Race.
- Color.
- Religion.
- National origin or ancestry.
- Sex/Gender.
- Age.
- Physical or mental disability.
- Veteran status.
- Genetic information.
- Citizenship.

SOME EXAMPLES OF ALGORITHMIC BIAS



unprofessional hairstyles



professional hairstyles

Gender classification: error rate

der was misidentified in **up to 7 percent of lighter-skinned fema** photos.

sidentified in **up to 12 percent of darker-skinned male**

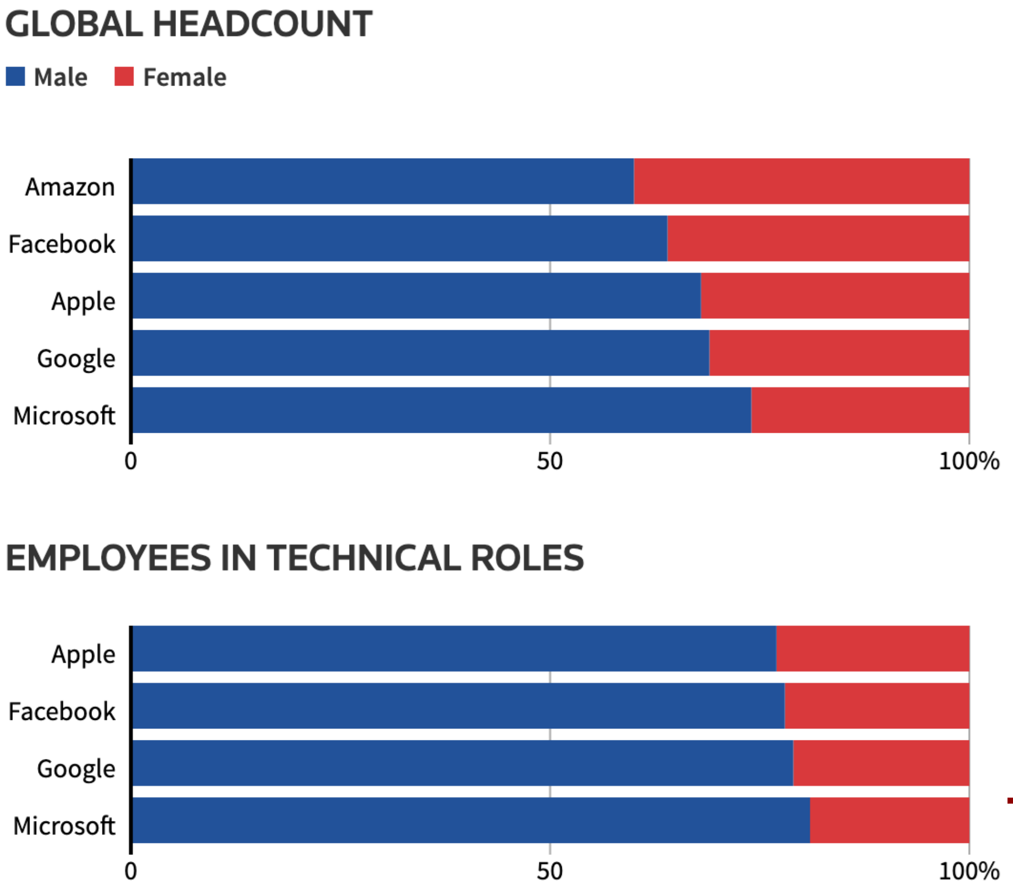
misidentified in **35 percent of darker-skinned females**

er was misidentified in **up to 1 percent of lighter-skinned males** photos.

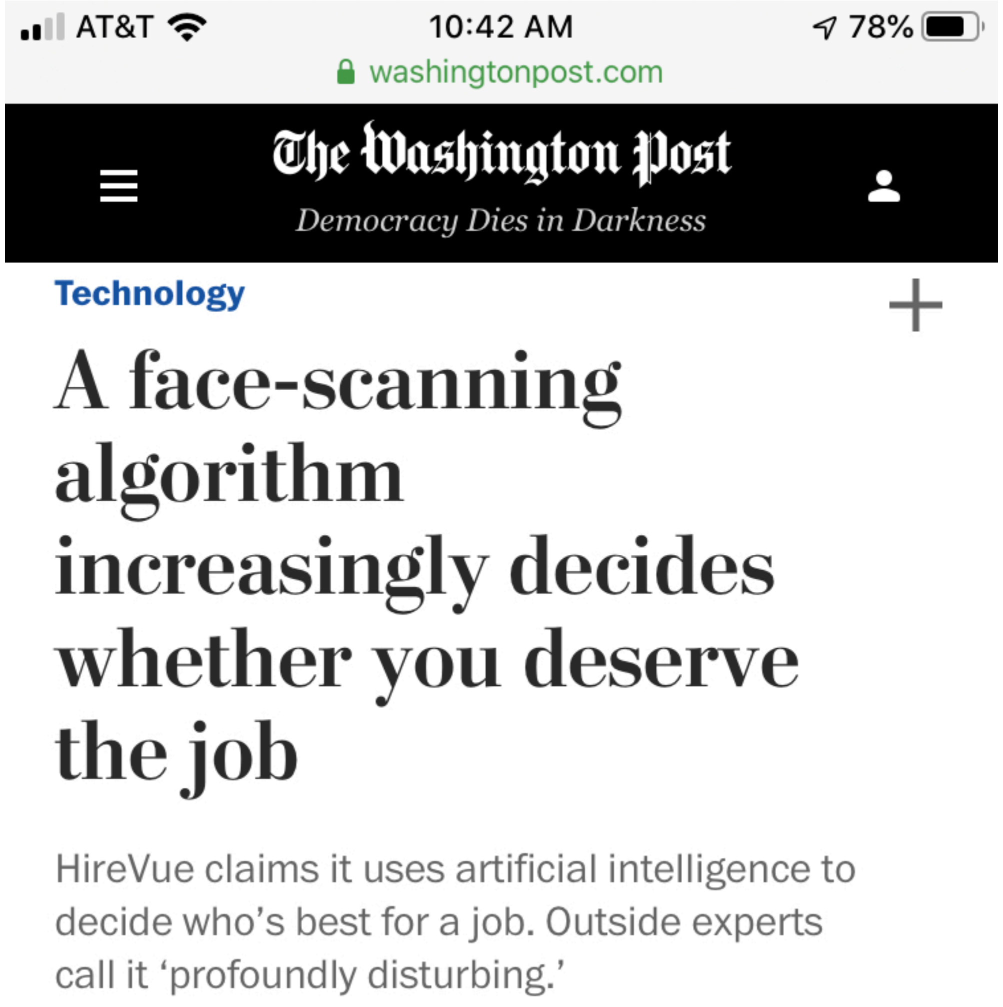
https://www.nytimes.com/2018/02/09/technology/facial-recognition-race-artificial-intelligence.html

Amazon hiring tool

- 2014 - Amazon develops AI system to screen resumes.
- One year later, realized the screening tool contains gender bias.
- Why?



AI-based job recruiting



AT&T 10:42 AM 78%
washingtonpost.com

The Washington Post
Democracy Dies in Darkness

Technology

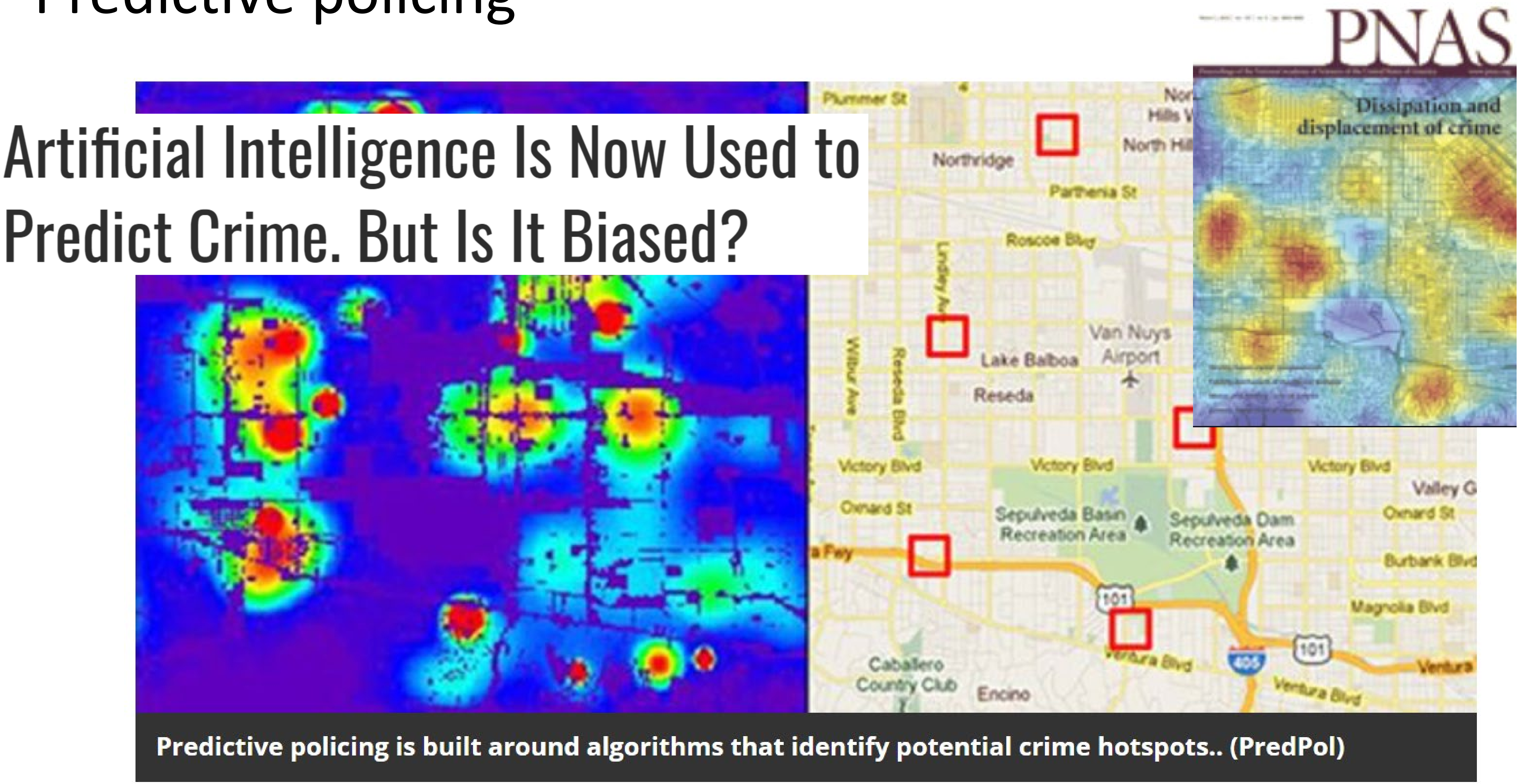
A face-scanning algorithm increasingly decides whether you deserve the job

HireVue claims it uses artificial intelligence to decide who's best for a job. Outside experts call it 'profoundly disturbing.'

An artificial intelligence hiring system has become a powerful gatekeeper for some of America's most prominent employers, reshaping how companies assess their workforce — and how prospective employees prove their worth.

Designed by the recruiting-technology firm HireVue, the system uses candidates' computer or cellphone cameras to analyze their facial movements, word choice and speaking voice before ranking them against other applicants based on an automatically generated "employability" score.

Predictive policing

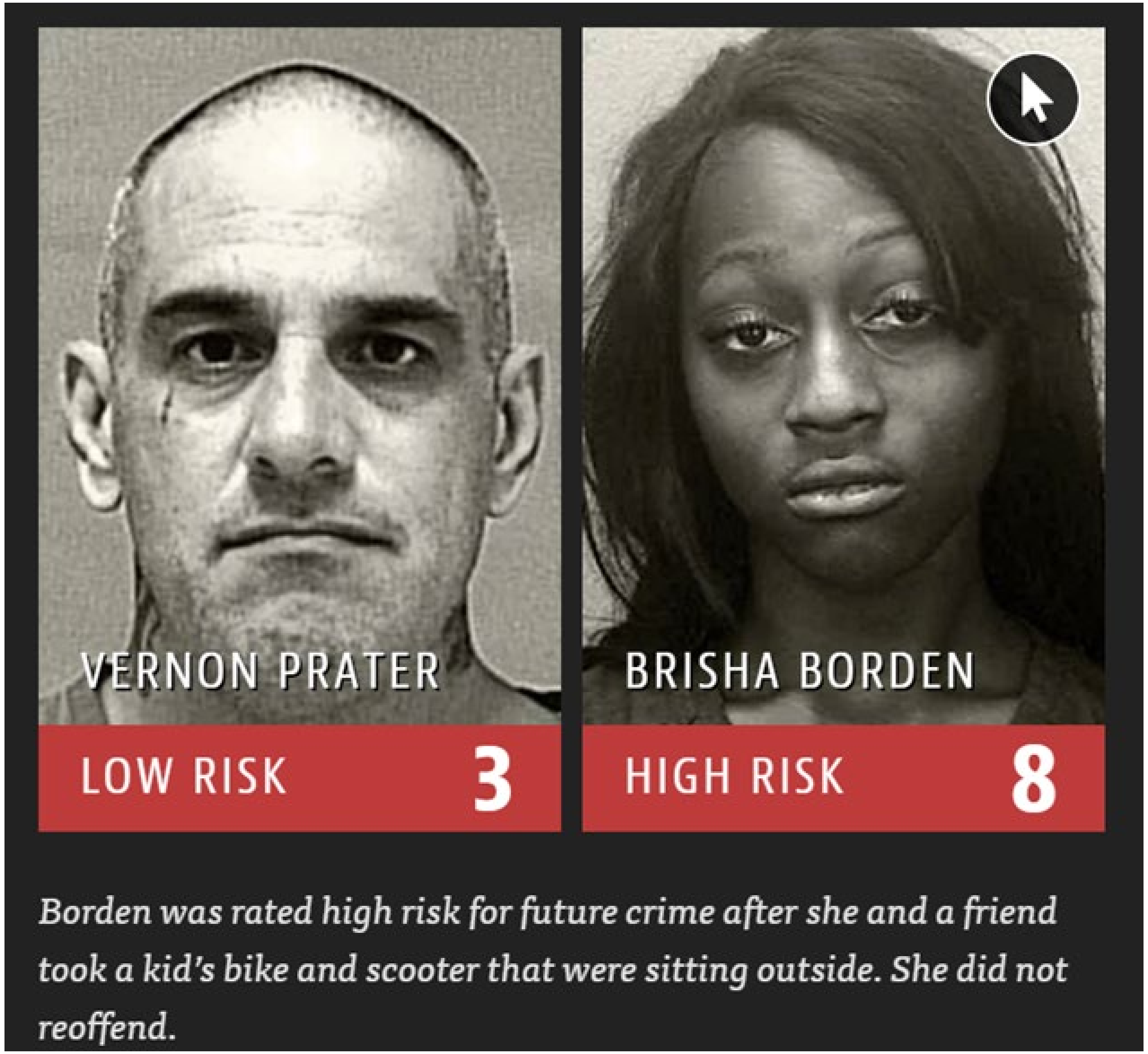


Artificial Intelligence Is Now Used to Predict Crime. But Is It Biased?

Predictive policing is built around algorithms that identify potential crime hotspots.. (PredPol)

Bias in automated criminal risk assessment

COMPAS tool systematically gives black defendants higher risk scores for future recidivism



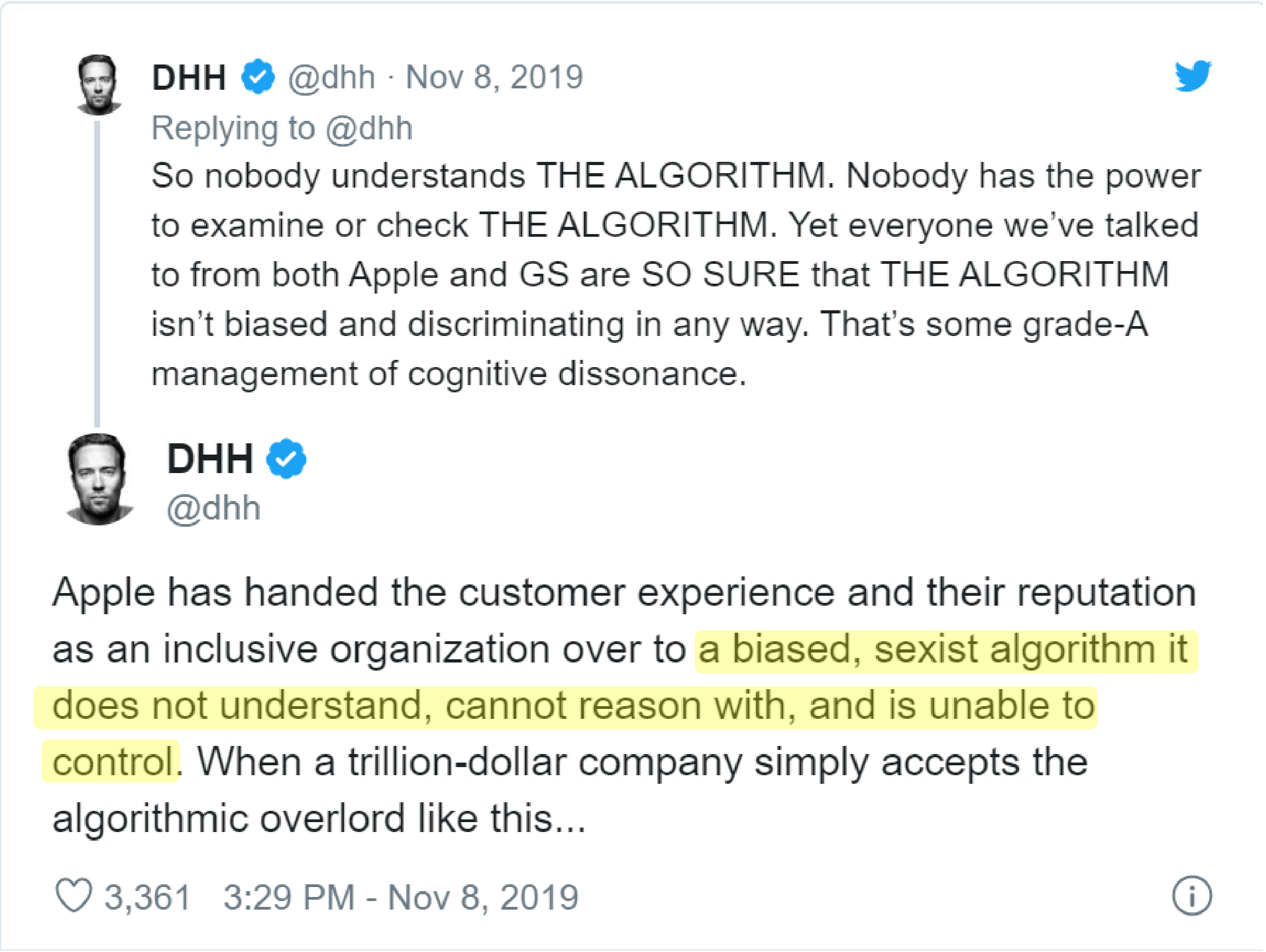
Borden was rated high risk for future crime after she and a friend took a kid’s bike and scooter that were sitting outside. She did not reoffend.

https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

The New York Times

Apple Card Investigated After Gender Discrimination Complaints

A prominent software developer said on Twitter that the credit card was “sexist” against women applying for credit.



DHH @dhh · Nov 8, 2019
Replying to @dhh
So nobody understands THE ALGORITHM. Nobody has the power to examine or check THE ALGORITHM. Yet everyone we’ve talked to from both Apple and GS are SO SURE that THE ALGORITHM isn’t biased and discriminating in any way. That’s some grade-A management of cognitive dissonance.

DHH
@dhh

Apple has handed the customer experience and their reputation as an inclusive organization over to a biased, sexist algorithm it does not understand, cannot reason with, and is unable to control. When a trillion-dollar company simply accepts the algorithmic overlord like this...

♡ 3,361 3:29 PM - Nov 8, 2019

How does data come to discriminate?

- Discrimination persists in American society in employment, housing, credit, consumer markets, academia, and entertainment
- While intentional discrimination and explicit prejudice are now less common, "institutional" discrimination through implicit biases and 'business as usual' attitudes account for disparate treatment of protected classes → encoded in patterns in data
- Absence of explicit prejudice does not guarantee impartiality
  - AI could unintentionally, rather than maliciously, discriminate by identifying and reinforcing existing patterns of discrimination
  - … and inadvertently amplify inequalities by giving disadvantaged groups less favorable treatment

How does data come to discriminate?

- Where does the unintended discrimination effects of data come from?

- Five mechanisms by which data mining may systematically disadvantage protected classes
  1. Defining the “Target Variable” and “Class Labels”
  2. Training data
  3. Feature selection
  4. Proxies
  5. Masking

1. Defining the “Target Variable” and “Class Labels”

- Data mining identifies statistical relationships in data (i.e., correlations) between target variables and features
  - Statistical patterns on which to base future decisions
  - Learned models can classify new entities, estimate unobserved features, predict future outcomes
  - How are target variables defined? –
    - What is “creditworthiness”? An abstraction developed for the purpose of evaluating loan default risk
    - what is a “good” employee? Objective measures?

2. Training data

- Biased training data leads to discriminatory models
  (1) If prejudice played a role in generating data, algorithms will happily reproduce patterns of prejudice
  (2) if data mining draws inferences from a biased sample of the population, these inferences may systematically disadvantage those who are under- or overrepresented in the data.
- Special considerations
  - *Labels* (ground truth data) are often subjective. This can skew modeling results (see example next page)
  - *Data collected* about protected classes may be systematically incorrect, or non-representative → may discriminate against protected classes.
    - Underrepresentation vs overrepresentation

Biases in hiring decisions

- St. George's Hospital (UK) developed a computer program to help sort medical school applicants based on its previous admissions decisions. Those admissions decisions systematically discriminated against racial minorities and women with similar credentials to other applicants'. By learning from prior biased decisions, St. George's Hospital devised an automated process that propagated the same prejudices.
- The computer may learn to discriminate against certain female or black applicants if trained on prior hiring decisions in which an employer has consistently rejected jobseekers with degrees from women's or historically black colleges.

3. Feature selection

- AI designers make choices about what attributes to collect and model
  - May systematically omit details to reliably resolve members of protected classes with details necessary to achieve equally accurate determinations residing at a level of granularity and coverage that the selected features fail to achieve.
  - Obtaining detailed information to resolve protected classes can be expensive.
    - See “redlining” – using a coarse proxy to make decisions , e.g., zipcode leads to less accurate decisions than fine-grained data, but is easier to collect

4. Proxies

- Features are highly correlated with membership in the protected class
  - E.g., “redlining” – zipcode used instead of race to deny loan applications

5. Masking

- Decision makers can disguise their prejudicial views by manipulating data collection, labeling, feature selection, etc.
  - E.g., decision makers could knowingly and purposefully bias the collection of data to ensure that mining suggests rules that are less favorable to members of protected classes, such as “redlining”

Measuring fairness

What is fairness?

- Not one established definition, many from many different levels.
  - Social science
  - Philosophy
  - Law
- In computer science, three macro definitions:
  - Group fairness
    - Treat different groups equally.
  - Individual fairness
    - Give similar predictions to similar individuals.
  - Subgroup fairness
    - Apply group fairness to a large collection of subgroups.

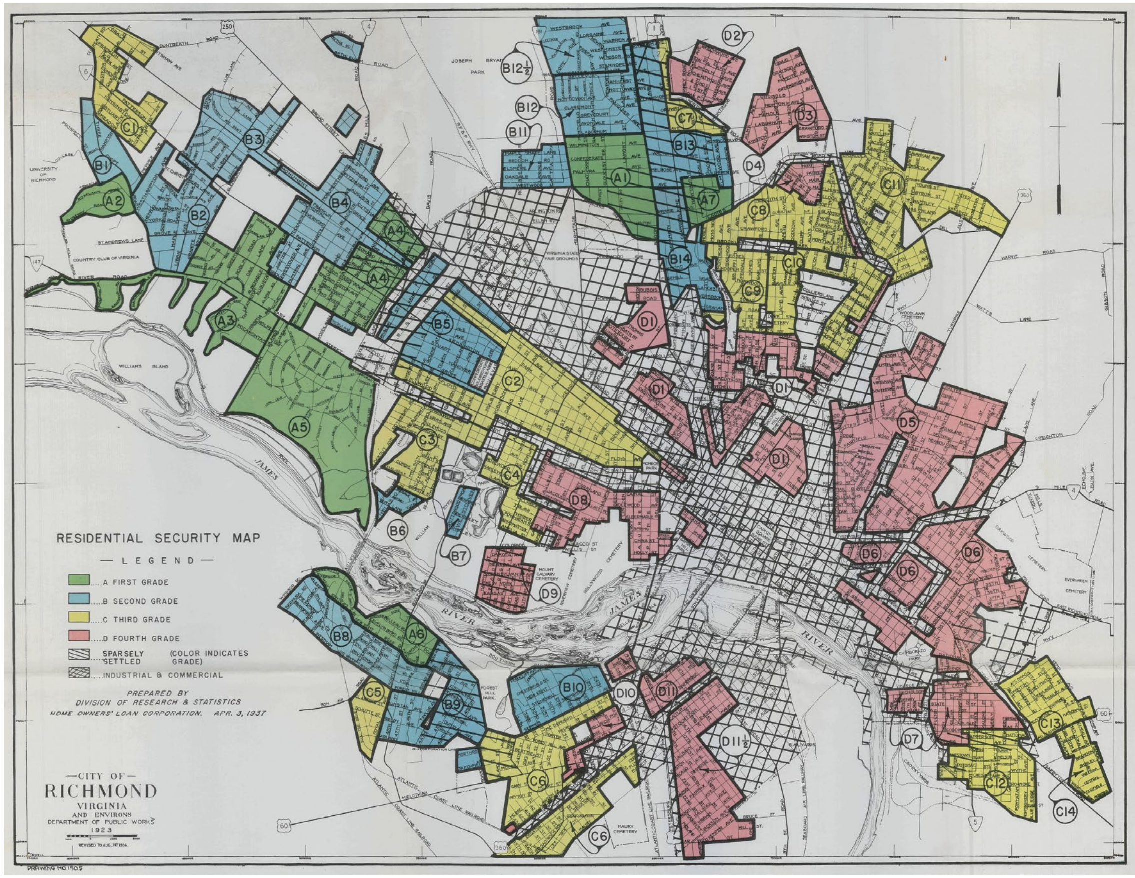
Notation

- A – set of protected attributes
- X – all other observable attributes
- U – set of latent attributes not observed
- Y – outcome to be predicted
- Ŷ – predictor, dependent on A, X, U.

Fairness through Unawareness

- *An algorithm is fair so long as any protected attributes A are not explicitly used in the decision-making process.*
  - Any mapping Ŷ : X → Y that excludes A satisfies this.
  - However, it can be biased for many reasons.
    - X still has biases from variables that correlate with *A*.

Counterexample: Redlining


RESIDENTIAL SECURITY MAP
— LEGEND —
🟩.....A FIRST GRADE
🟦.....B SECOND GRADE
🟨.....C THIRD GRADE
🟥.....D FOURTH GRADE
▨.....SPARSELY SETTLED (COLOR INDICATES GRADE)
▩.....INDUSTRIAL & COMMERCIAL
PREPARED BY
DIVISION OF RESEARCH & STATISTICS
HOME OWNERS' LOAN CORPORATION. APR. 3, 1937
—CITY OF—
RICHMOND
VIRGINIA
AND ENVIRONS
DEPARTMENT OF PUBLIC WORKS
1923
REVISED TO AUG. 1ST 1936.
DRAWING NO 1505
JAMES RIVER

- Banks marked certain neighborhoods as risky investments, because Blacks lived there
- Although race was not used explicitly in loan decisions, discrimination/unfairness is still possible
  - Zipcode as a proxy of race

Statistical Parity (Group Fairness)

Equalize two groups S, T at the level of outcomes

- E.g., *S* = minority, *T* = majority

$$\Pr[\text{credit} = 1 \mid S] = \Pr\,[\text{credit} = 1 \mid T]$$

“Fraction of people in S getting credit is the same as in T.”

Conditional Statistical Parity

- $P(d = 1 \mid L = l, A = m) = P(d = 1 \mid L = l, A = f)$
- *Difference:* Considering these factors, protected and unprotected instances should have the same probability of success.
- L must be legitimate.
- In the credit example, legitimate factors could be:
  - Credit history, employment, amount requested.

Equalized Odds

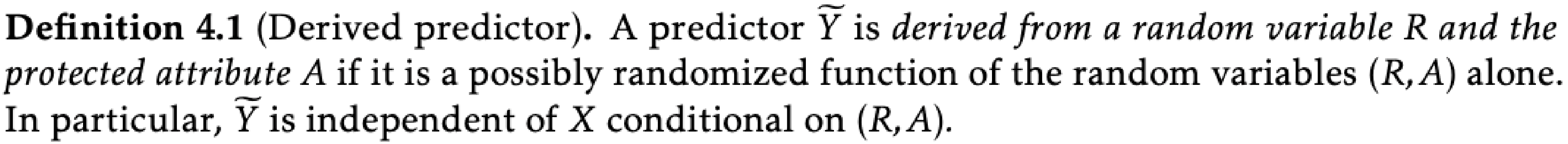
- Definition: $\hat{Y} \perp A | Y$
- Prediction does not provide information about protected attribute *A* beyond what Y already does.
- $P(\hat{Y} |\ Y = y, A = m) = P(\hat{Y}\ |\ Y = y, A = f)$
- Protects against accuracy disparity.

Equality of Opportunity

- In a binary case, we think of Y=1 as the “advantaged” outcome.
- Require non-discrimination only within this outcome.
- E.g., people who pay back their loan ought to have an equal opportunity of getting the loan in the first place.
- $P(\hat{Y} = 1 \mid A = m, Y = 1) = P(\hat{Y} = 1 \mid A = f, Y = 1)$.
- Relaxation of Equalized Odds.

Achieving Eq. Odds and Eq. of Opportunity

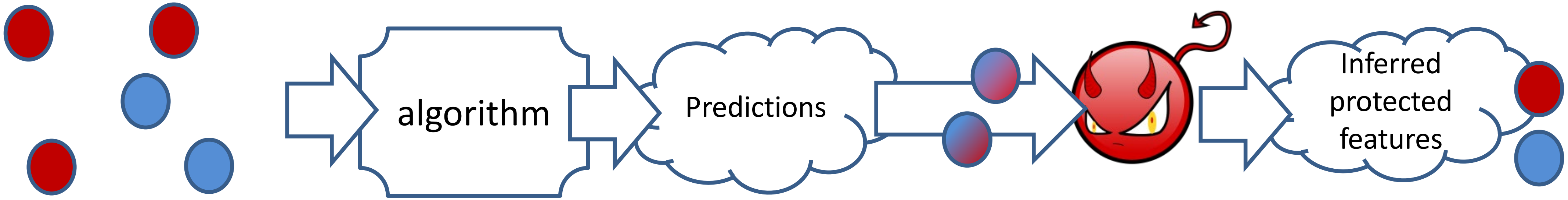
- Goal: find an Eq. odds or Eq. Opportunity predictor $\tilde{Y}$
  - Derived from a (possibly discriminatory) predictor, $\hat{Y}$



Definition 4.1 (Derived predictor). A predictor $\widetilde{Y}$ is *derived from a random variable R and the protected attribute A* if it is a possibly randomized function of the random variables (R, A) alone. In particular, $\widetilde{Y}$ is independent of X conditional on (R, A).

- The joint distribution is required at training time.
- At prediction time, we only have R, A.

Alternate ways to measure fairness



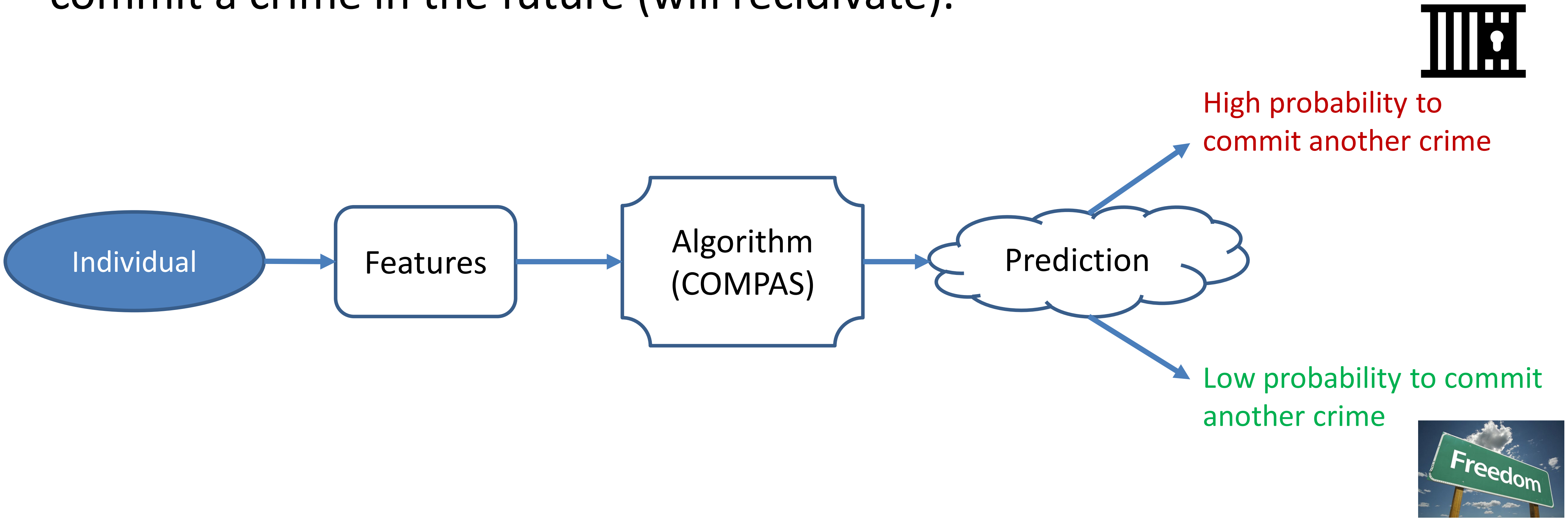
- Do predictions reveal information about the protected feature?
  - Imagine an evil actor who reverse engineers values of the protected features (A) from the predictions (Ŷ)
  - Measuring dependency between Ŷ and A

1. Pearson correlation between the predictions and the protected feature
2. Mutual information between the predictions and the protected feature
3. Accuracy of predicting the protected features from predictions

FAIRNESS IN ALGORITHMIC RISK SCORES

Automated risk assessment

- Increasingly common in the judicial system in the US
- Algorithm computes a score predicting the likelihood a defendant will commit a crime in the future (will recidivate).



Individual

Features

Algorithm (COMPAS)

Prediction

High probability to commit another crime

Low probability to commit another crime

Benefits and failings of risk scores

- Increase efficiency and reduce prejudice
  - Used by judges to inform decisions about bail, sentencing, parole, conditions for release, etc.
  - Hoped to mitigate existing sources of bias in human judgments
- In 2014 US Attorney General Eric Holder warned risk scores may be injecting bias into the courts
  - “Although these measures were crafted with the best of intentions, I am concerned that they inadvertently undermine our efforts to ensure individualized and equal justice. They may exacerbate unwarranted and unjust disparities that are already far too common in our criminal justice system and in our society.”

COMPAS

- Correctional Offender Management Profiling for Alternative Sanctions
- Assigns defendants risk scores between 1 and 10 that indicate how likely they are to commit a violent crime based on more than 100 factors, including age, sex and criminal history.
  - Defendants with scores of 7 reoffend at twice the rate as those with scores of 3.
- Help judges make decisions about whether to release a defendant or hold them in jail while waiting for trial
- Deployed before it was rigorously tested

Where does the data for COMPAS come from?

- Level of Service Inventory – Revised (LSI)
- A lengthy questionnaire for the prisoner to fill out, with questions like
  - “How many prior arrests have you had?”
  - “What part did alcohol and drugs play?”
  - “The first time you were involved with the police?”
- The questions themselves discriminate against minority classes and contribute to the feedback loop amplifying bias

ProPublica study of racial disparities in risk scores

- Records of 7000 people arrested in Broward County, FL (2013-2014)
- COMPAS predicted whether defendant will commit a crime in 2 years

		What actually happened	
		Did commit a crime	**Did not commit a crime**
COMPAS prediction	**Will commit crime**	True Positives (TP)	*False Positives (FP)*
	Will not commit a crime	*False Negatives (FN)*	True Negatives (TN)

https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Definitions of fairness

- *Statistical parity* - same proportion of defendants are detained in each race group. For example, white and black defendants are detained at equal rates.
- *Conditional statistical parity* - controlling for “legitimate” risk factors, an equal proportion of defendants are detained within each race group
  - E.g., among defendants with same number of prior convictions, black and white defendants are detained at equal rates.
- *Predictive equality* – the accuracy of decisions is equal across race groups, as measured by false positive rate (FPR). This condition means that among defendants who did not go on to commit a violent crime after release, detention rates are equal across race groups.

The tension

- Judges must balance two factors:
  - Let a guilty person go?
    - Increase in crimes committed by released defendants
  - Put an innocent person in jail?
    - social and economic costs of detaining innocent defendants.

Benefit vs cost

Benefit : expected number of crimes prevented

Cost : expected number of innocent people detained

- Define *utility* of a decision rule, that balances benefits and costs
- What is the decision rule that optimized the utility?

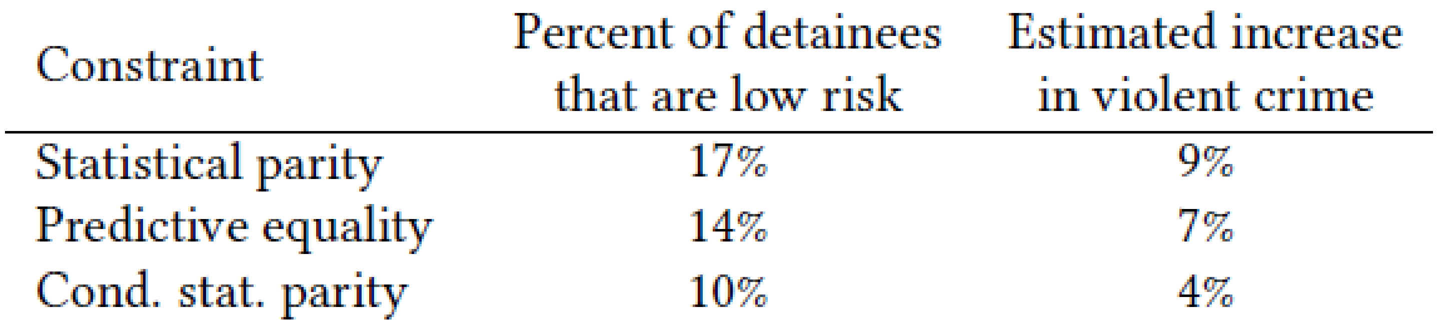
Tensions between different views of fairness

- Fair decision rule v1
  - Treat all individuals equally regardless of race
  - Uses the same threshold for risk scores across both groups
  - For example, $\phi = 3$
  - Maximizes utility by detaining defendants for p(y)>ϕ
- Fair decision rule v2 (satisfying statistical parity)
  - Each group must have different threshold, eg, $\phi = 3$ for white and $\phi = 5$ for blacks
  - Maximizes utility when detains the same proportion of defendants in each group
- 2 views of fairness in conflict

Cost of fairness

[Corbett-Davies, Pierson, Feller, Goel and Huq. (2016). Algorithmic decision making and the cost of fairness. In *KDD*.]

- Does fairness decrease public safety?
  - How much does the algorithm increase crime by mistakenly releasing violent offenders



Constraint	Percent of detainees that are low risk	Estimated increase in violent crime
Statistical parity	17%	9%
Predictive equality	14%	7%
Cond. stat. parity	10%	4%

→The cost to satisfying popular notions of algorithmic fairness is reducing public safety.

→**In general, fairness reduces accuracy (fairness-accuracy tradeoff)**

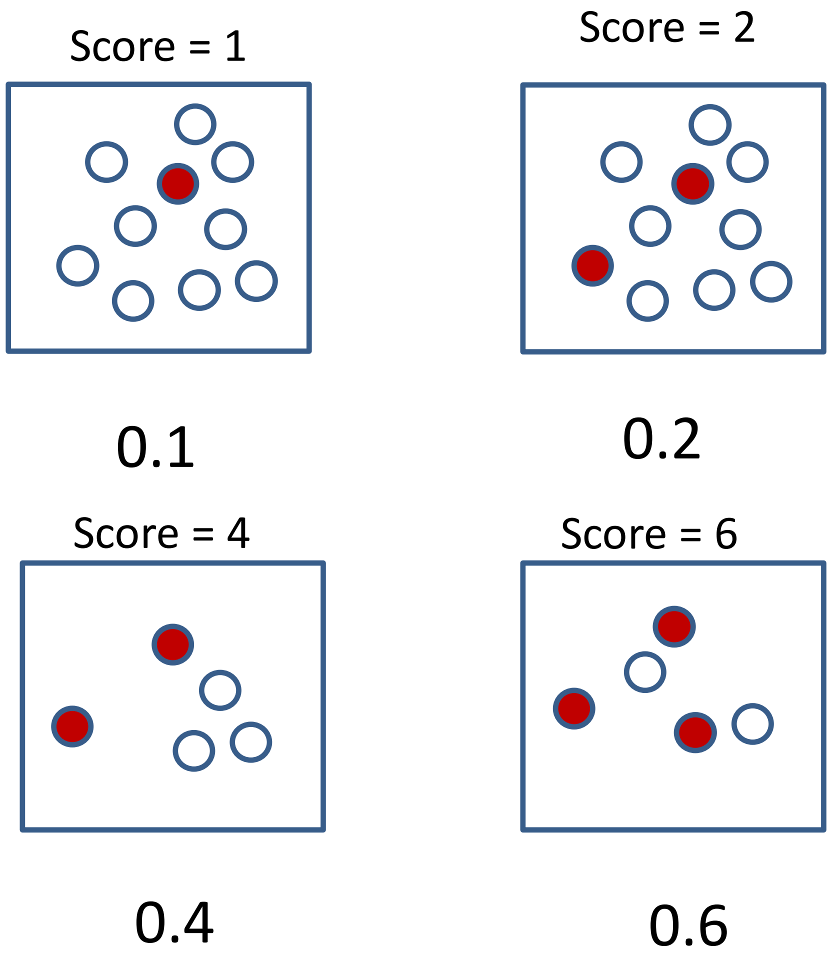
Inherent Trade-Offs in the Fair Determination of Risk Scores

[Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). Inherent trade-offs in the fair determination of risk scores.]

- Algorithmic predictions should have the same effectiveness regardless of group membership
- What kind of guarantees about fairness can we make?
  - We can define intuitive properties of fairness
  - But, they are incompatible with each other and can only be simultaneously satisfied in a few special cases.

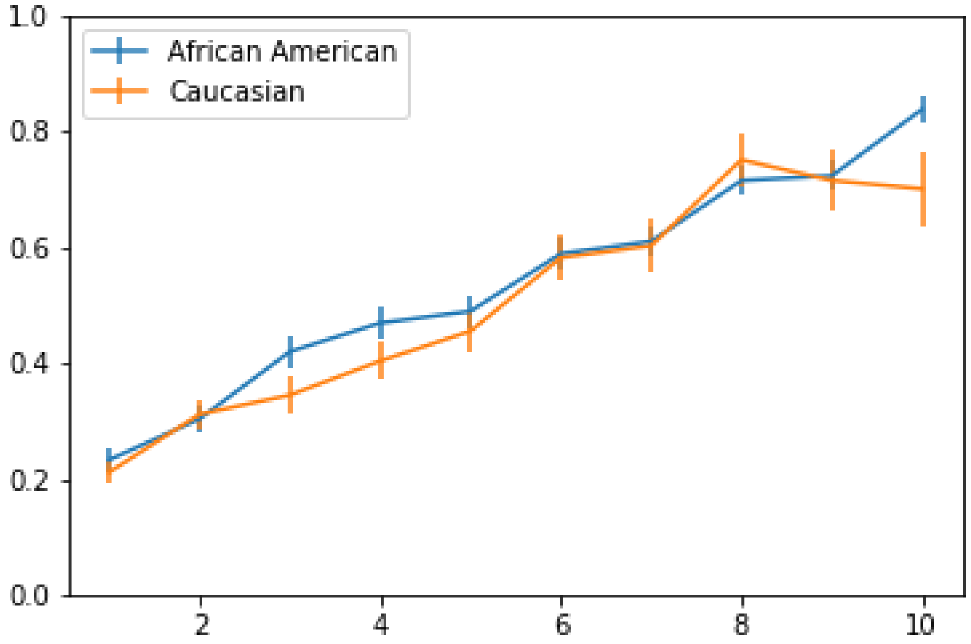
Fairness properties of risk assessments

- *Calibration within groups* requires that for each group, a v_b fraction of people in bin *b* are positive
- Societal benefits to calibration



COMPAS risk scores are well calibrated

Probability of recidivism

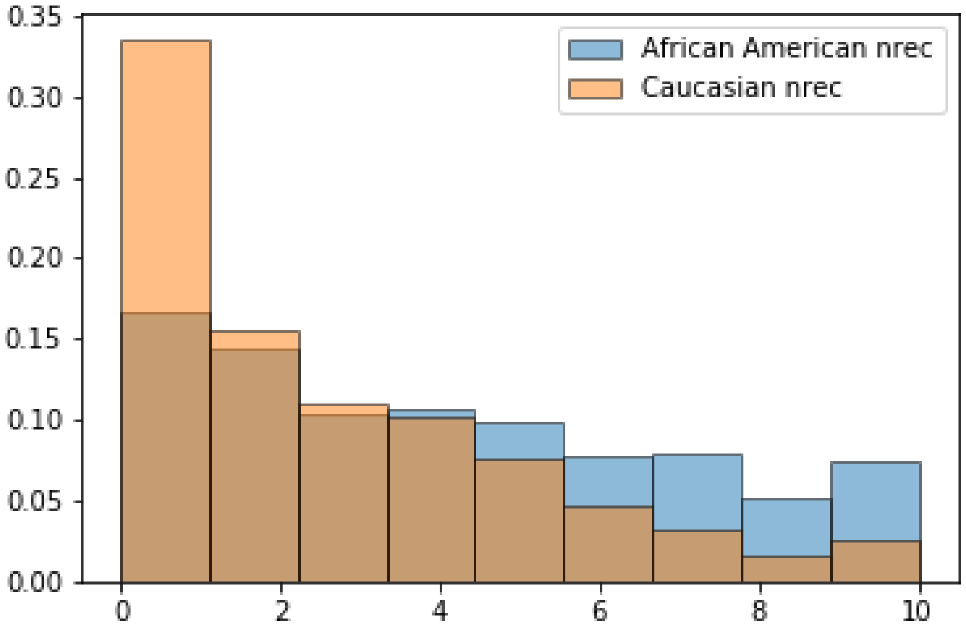


Risk score

Fairness of risk predictions

- *Balance for the negative class* requires that the average scores of positive members in group A equals average score of positive members in group B
  - i.e., average score of defendants who never commit another crime should be the same for each group
  - assignment of scores shouldn't be systematically more inaccurate for negative instances in one group than the other.

Negative instances: Defendants who do not commit another crime



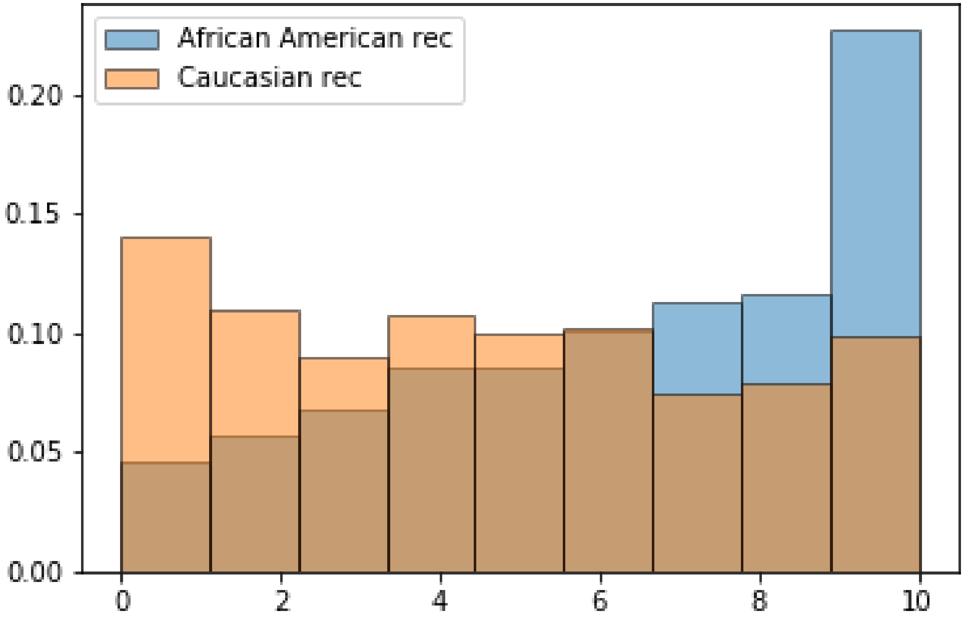
Risk score

Fairness of risk predictions

- *Balance for the positive class* symmetrically requires that average score of defendants who do commit another crime should be the same in each group

- both groups should have equal false positive and false negative rates
  - Balance is independent of *statistical parity*

Positive instances: Defendants who commit another crime

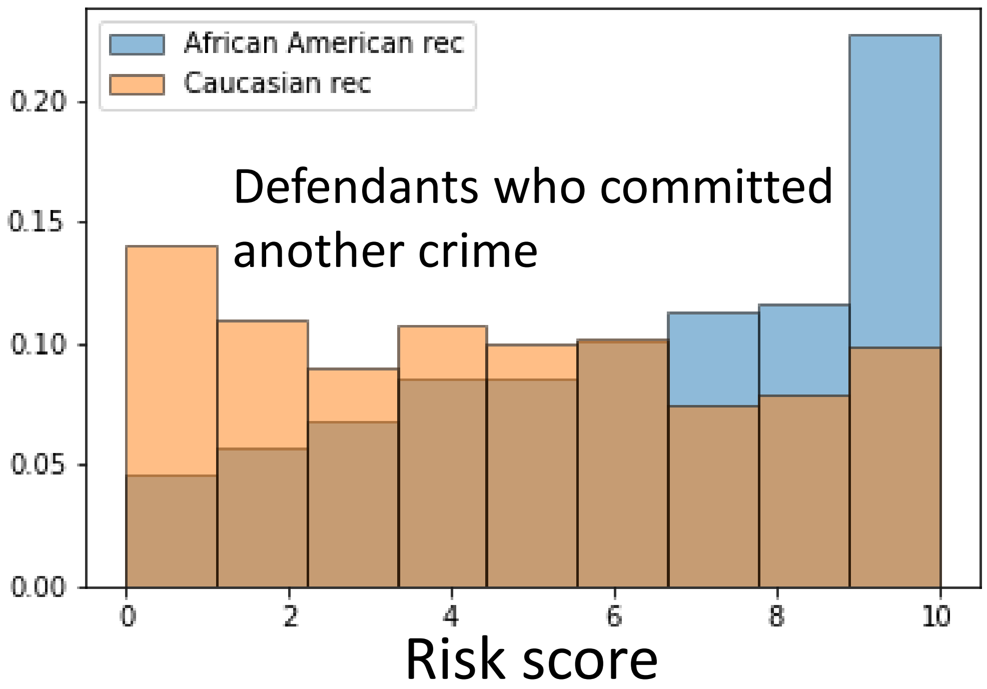


Risk score

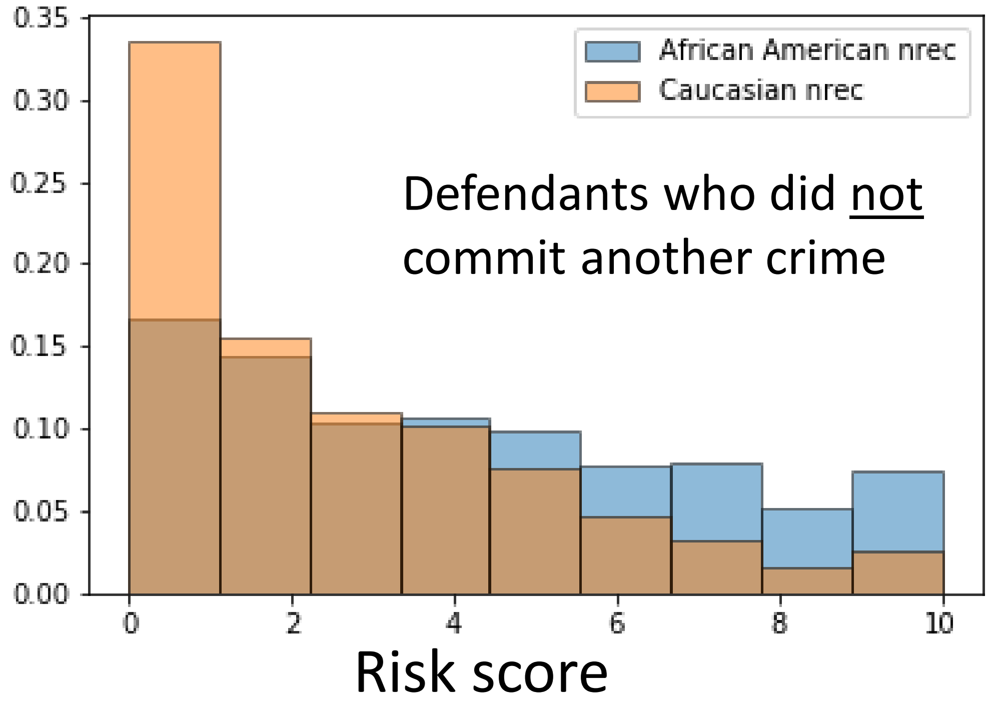
Asymmetry of COMPAS mistakes

White defendants were mislabeled as low risk more often than blacks.

Black defendants were mislabeled as high risk at twice the rate of whites



Defendants who committed another crime



Defendants who did not commit another crime

Dueling notions of fairness

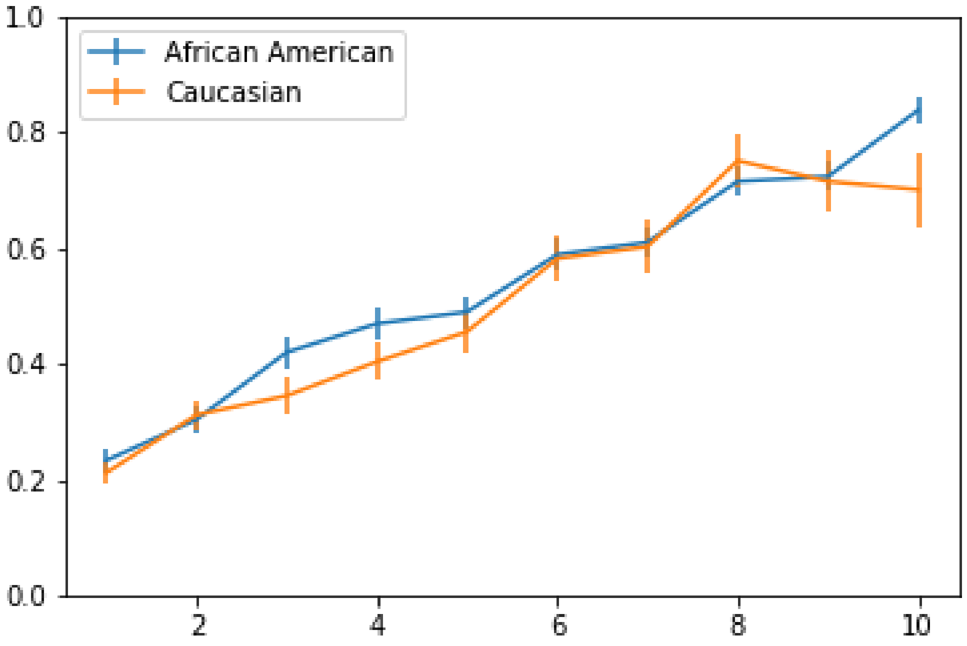
- **Calibration condition:** we treat people with the same score similarly to one other, regardless of their group (race)
  - I.e., conditioned on predictions, outcomes should not depend on race. Individuals should be given similar bails, and similar sentences. (COMPAS's fairness claims)
- **Balance conditions:** if two people in different groups have the same future behavior (recidivate or not), they should be treated the same way
  - I.e., conditioned on outcomes, predictions should not depend on race. Members of one group who never commit another crime should not have consistently higher scores than members of the other group (ProPublica's fairness claims)

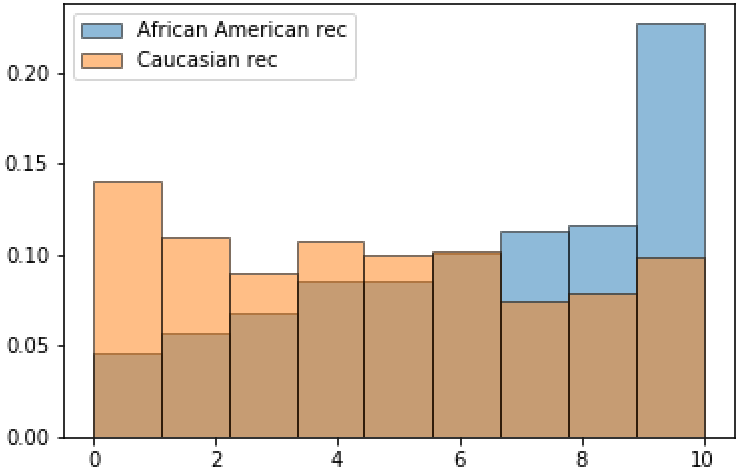
Fundamental limits of fairness

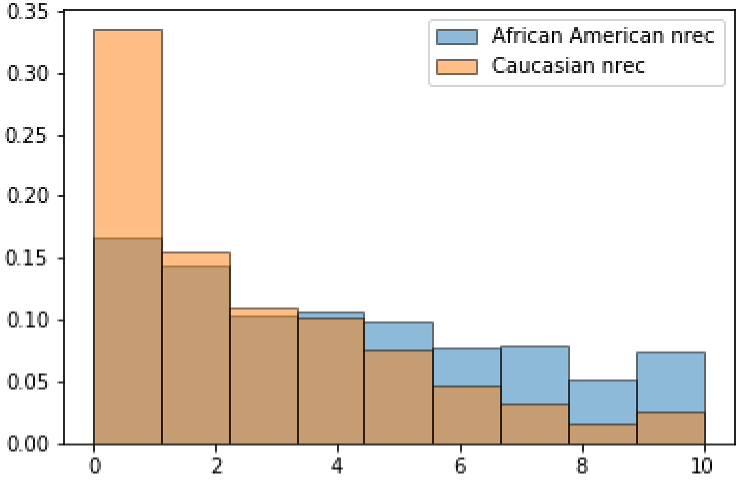
- The fairness conditions (calibration, balance) cannot be simultaneously satisfied except for special cases
  - Perfect prediction: model makes no prediction errors
  - Equal base rates: recidivism rate is the same for both groups. Is it achievable?

- In all other cases, you must sacrifice at least one fairness condition

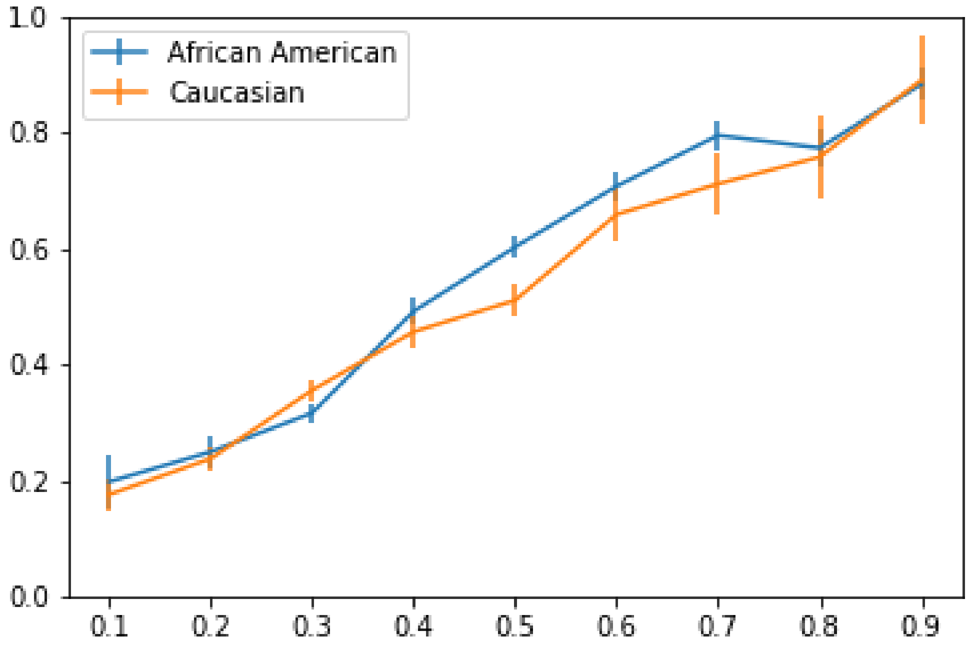
Calibration & Balance: COMPAS

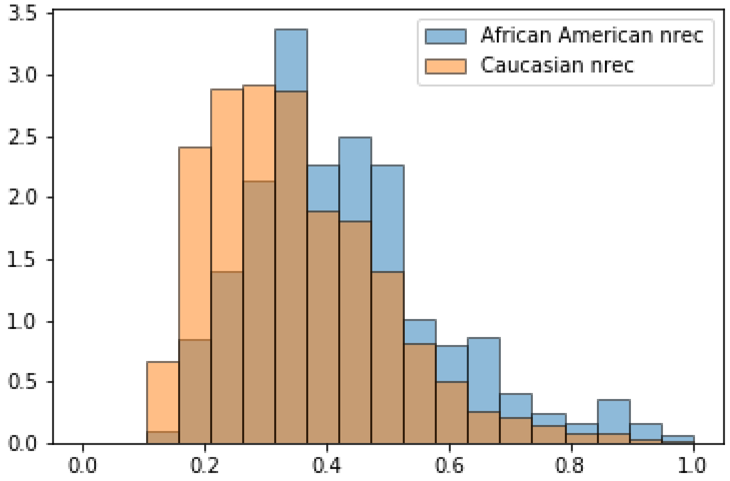


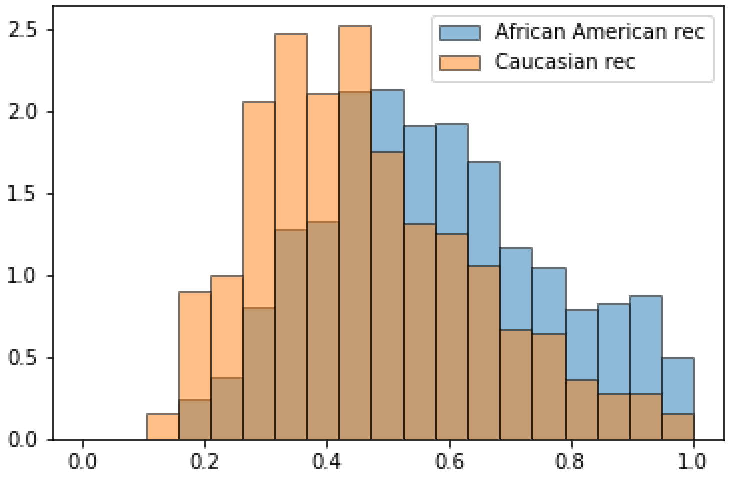




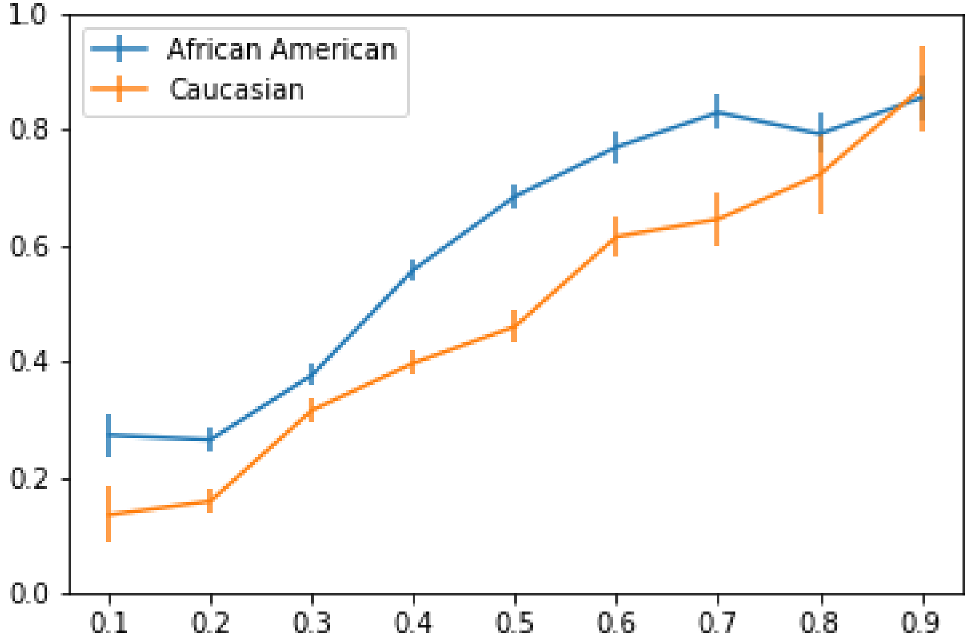
Race blind regression

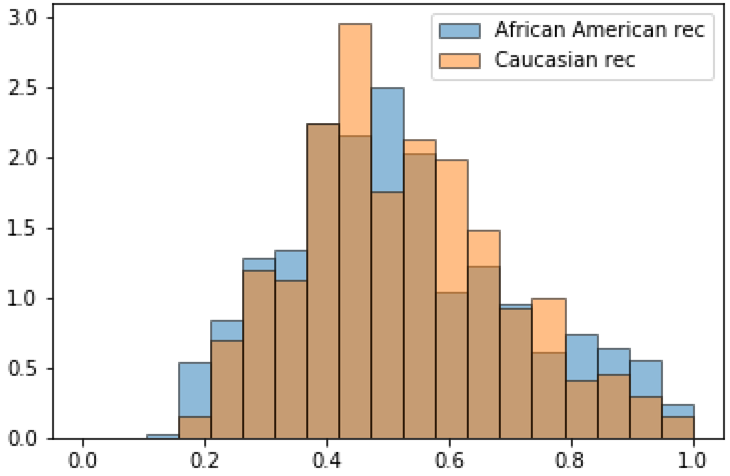


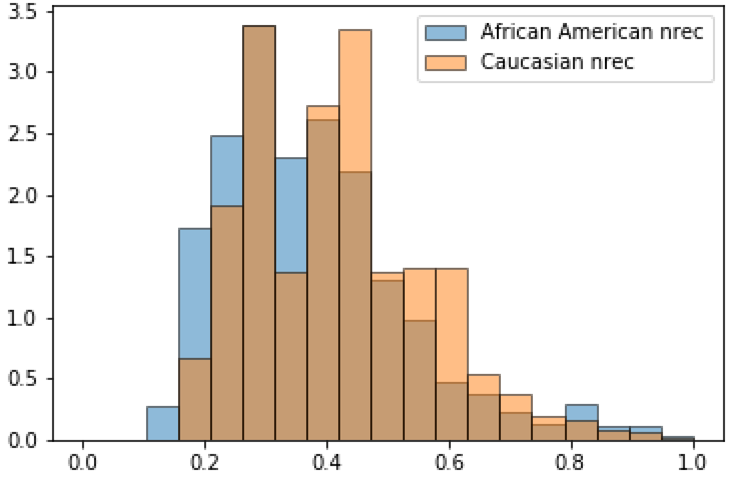




De-biased regression





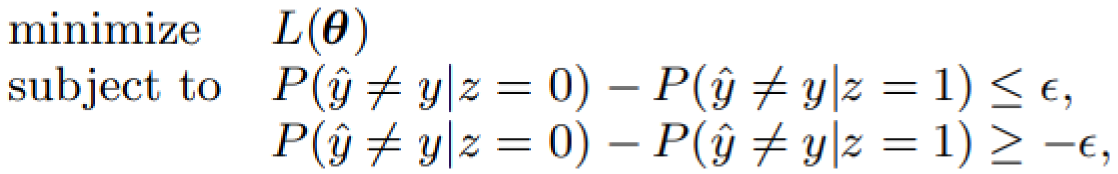


MITIGATING BIAS TO IMPROVE FAIRNESS

Fairness of misclassifications

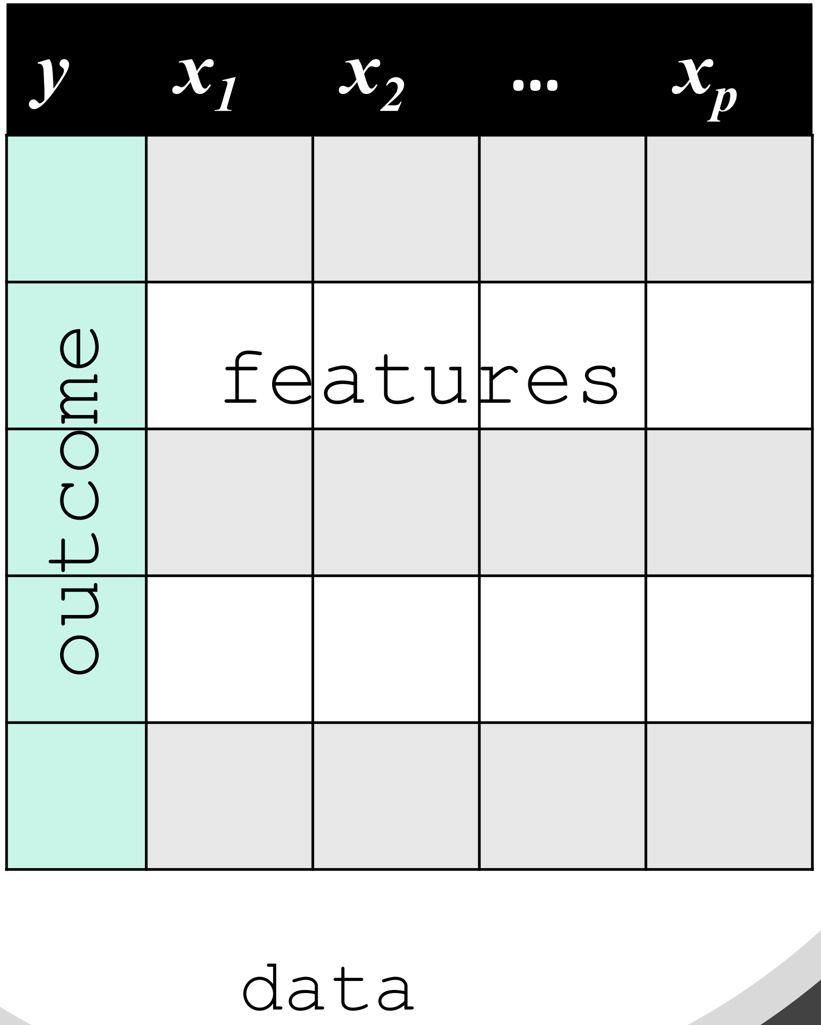
- Disparate treatment
  - Similar people should be treated the same by the algorithm, regardless of the group they belong to
- Disparate mistreatment
  - Rate of errors for each group is different
- Incorporating fairness constraints into classifiers – logistic regression without disparate mistreatment

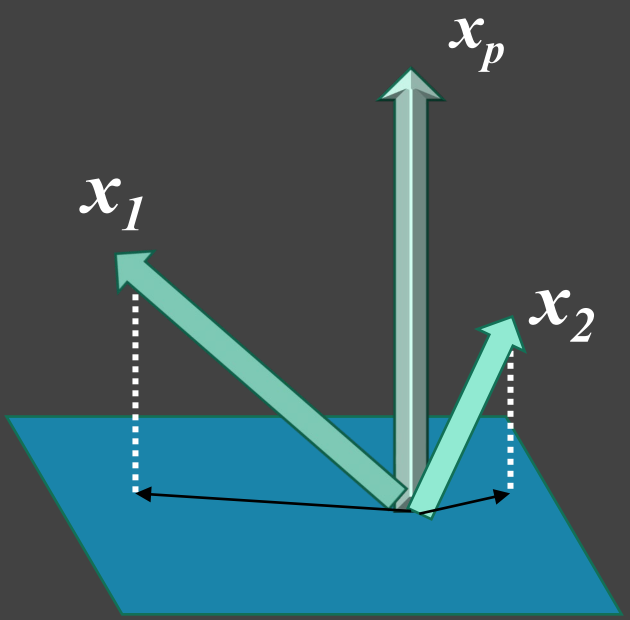
Zafar et al. (2016) “Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment”



$$\begin{array}{ll} \text{minimize} & L(\boldsymbol{\theta}) \\ \text{subject to} & P(\hat{y} \neq y | z = 0) - P(\hat{y} \neq y | z = 1) \leq \epsilon, \\ & P(\hat{y} \neq y | z = 0) - P(\hat{y} \neq y | z = 1) \geq -\epsilon, \end{array}$$

Fair representations via linear orthogonolization



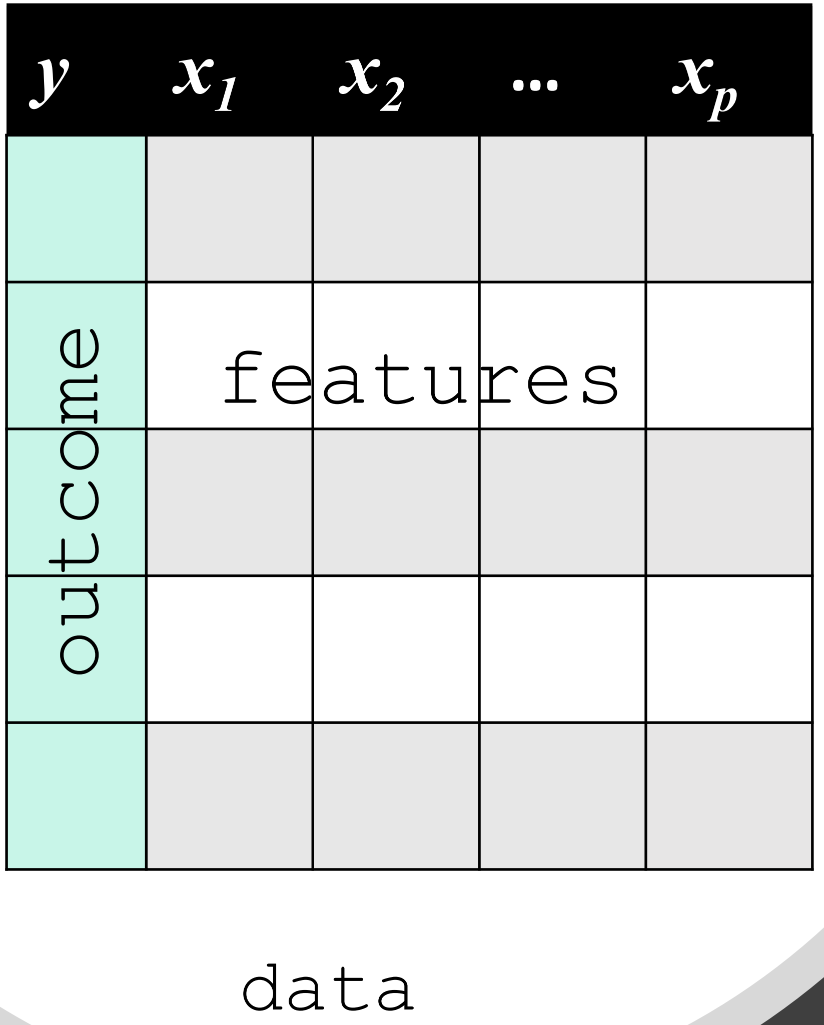


1. Create a null space orthogonal to protected feature(s) x_p

2. Features x_i projected unto the null space are independent of the protected features

3. Parameter $\lambda \in [0,1]$ adjusts the level of fairness

Fair representations via linear orthogonolization



y x_1 x_2 … x_p

outcome

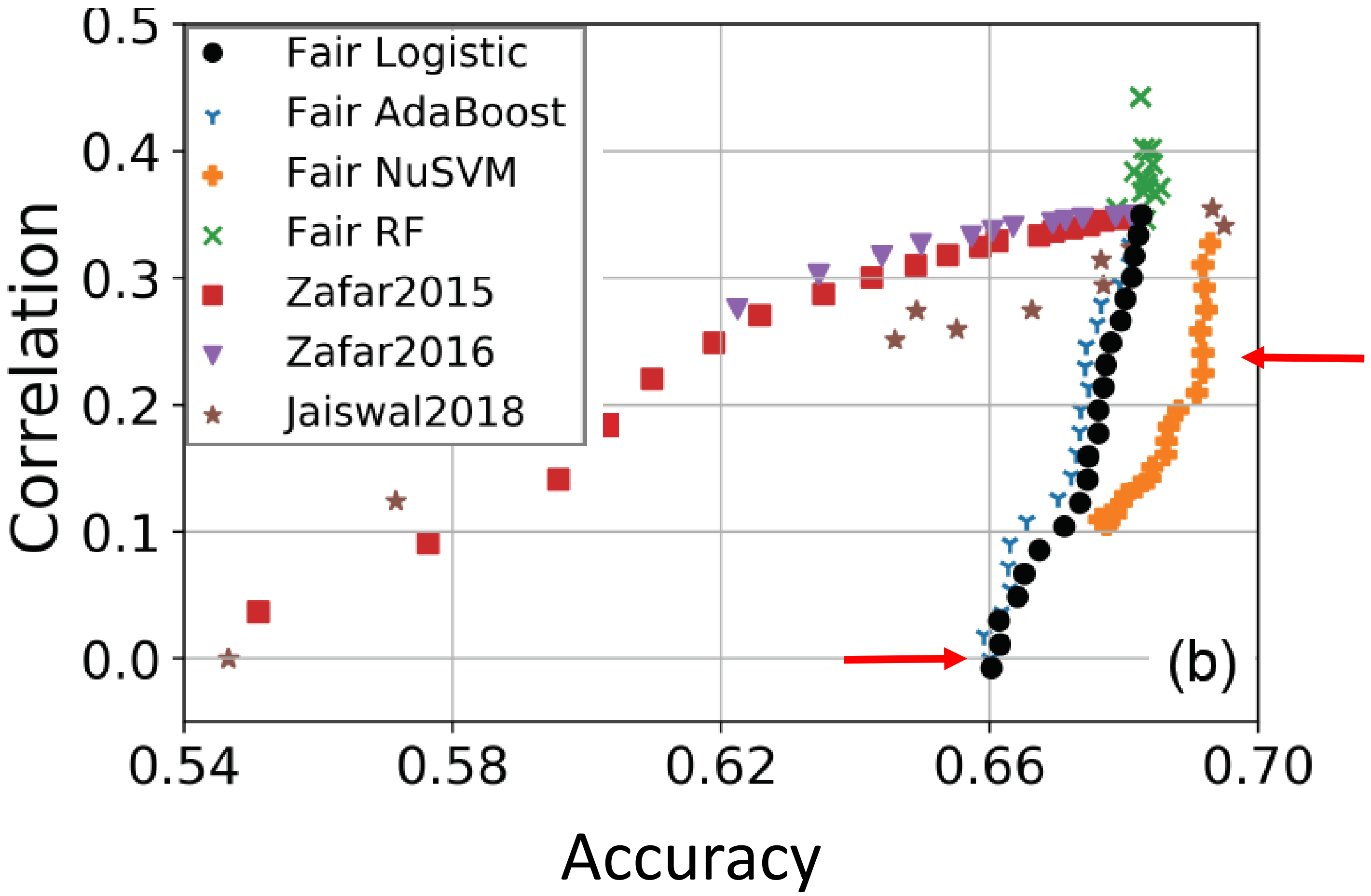
features

data

- Interpretable
- Used with different models
  - Regression
  - Decision trees
  - SVM, etc
  - Neural networks
- More fair predictions
- More accurate than state-of-the-art fairness methods

He, Y., Burghardt, K., & Lerman, K. (2020, February). A Geometric Solution to Fair Representations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* https://arxiv.org/abs/1910.12854

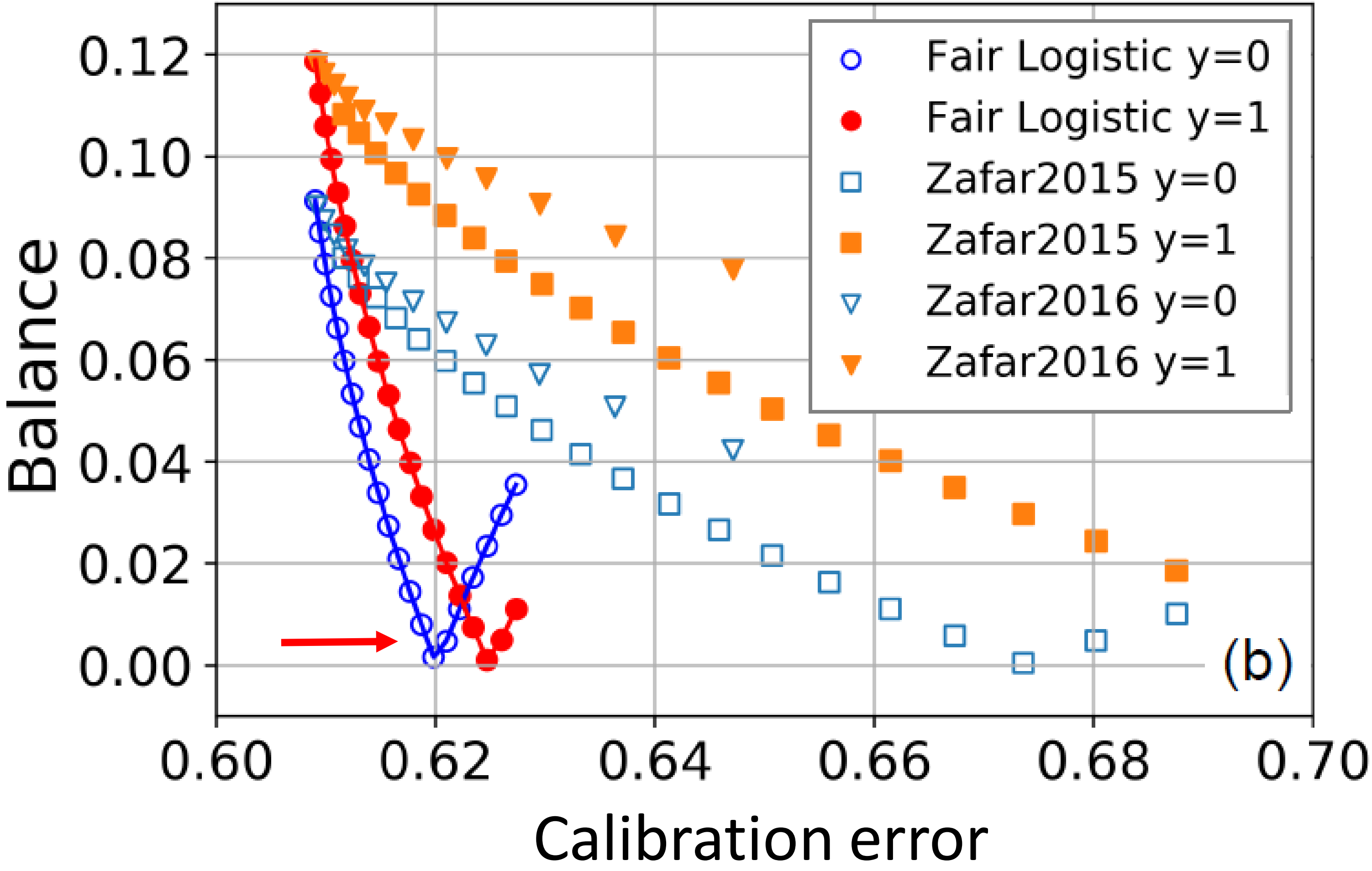
COMPAS: Fairness vs Accuracy tradeoffs



Correlation of predictions $\hat{Y}$ with protected features A

He, Y., Burghardt, K., & Lerman, K. (2020, February). A Geometric Solution to Fair Representations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* https://arxiv.org/abs/1910.12854

COMPAS: Balance vs Calibration tradeoffs



Difference between the means for the two classes

He, Y., Burghardt, K., & Lerman, K. (2020, February). A Geometric Solution to Fair Representations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* https://arxiv.org/abs/1910.12854

Lessons learned: How algorithms can discriminate

- Transparency
  - How is the data collected? Is the bias built into data collection?
    - Choices of what data to collect, what variables to model reflect ideology and blind spots
  - Are participants aware of being modeled? Do they know how their data is used?
  - Does the model rely on proxies of protected features? Are they valid?
- Asymmetry
  - Who is harmed by the mistakes? Is the damage born by all groups equally?
  - Is the harm of False Positives comparable to the harms of False Negatives?
- Feedback
  - Does the model learn from mistakes? Is it updated when ground truth changes?
  - Does it create a feedback loop that exacerbates the damage?

- Questions?

- Virtual office hour
- https://usc.zoom.us/j/95136500603?pwd=VEJhblhWK25lT2N3RC9FNWk3eTJKQT09