Quantifying bias in machine decisions

Sharad Goel
Stanford Computational Policy Lab

Summary

Most proposed mathematical measures of fairness are poor proxies for detecting discrimination.

Attempts to satisfy these formal measures of fairness can lead to discriminatory or otherwise perverse decisions.

Summary

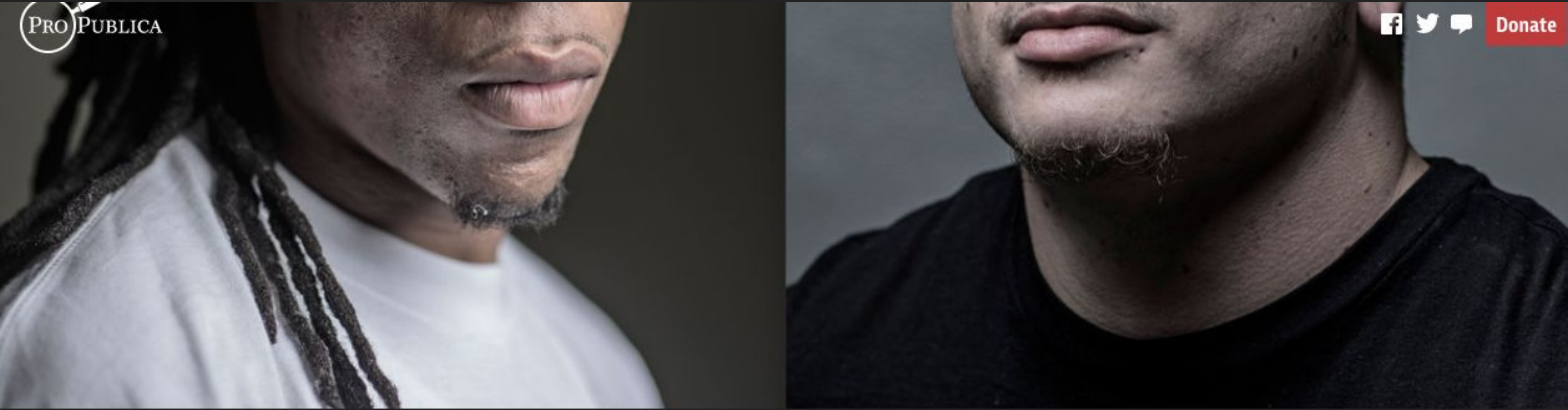
Most proposed mathematical measures of fairness are poor proxies for detecting discrimination.

Attempts to satisfy these formal measures of fairness can lead to discriminatory or otherwise perverse decisions.

This is a controversial message so please push back!

Part I

Algorithmic decision making



Bernard Parker, left, was rated high risk; Dylan Fugett was rated low risk. (Josh Ritchie for ProPublica)

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

Pretrial detention

A detailed case study

Judges must decide which arrested defendants should be released while awaiting trial and which should be detained.

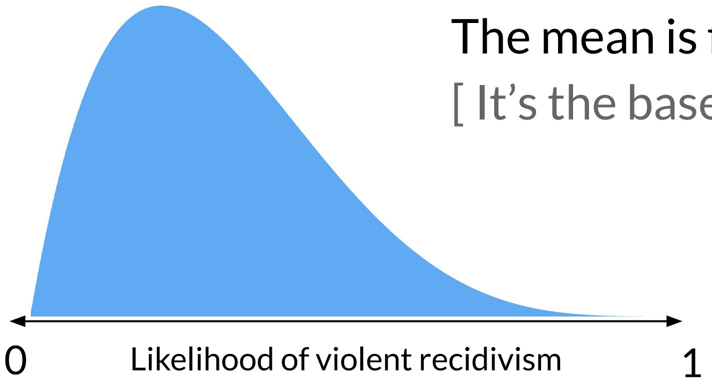
The goal is to balance the social and financial costs of incarceration with the benefits of reducing pretrial crime.

Key assumptions

1. We know the true label **Y** (i.e,. whether a defendant would have reoffended if released).
   [Y is true counterfactual, with no measurement error.]

2. We know the true risk: $r_x = P(Y=1 \mid X=x)$
   [Reasonable when we have lots of data.]

Risk distributions



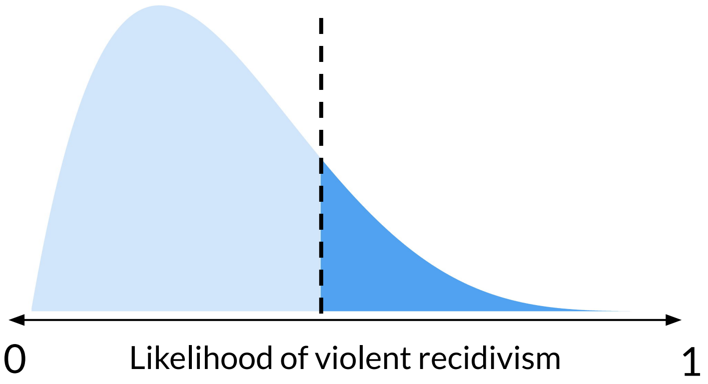
The mean is fixed for all choices of X

[It's the base rate of recidivism.]

The shape can change based on our choice of X

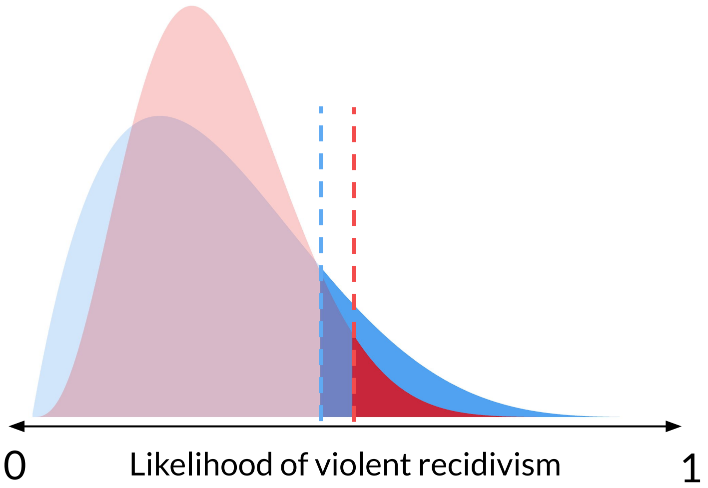
From risk to decisions

Threshold rules



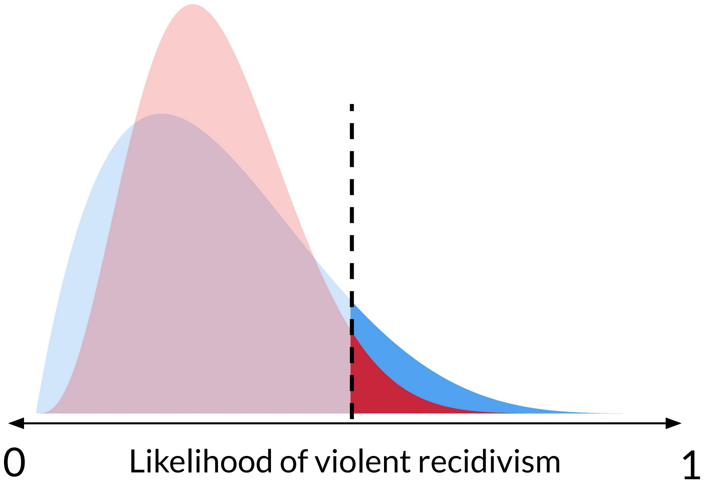
A threshold of ***t*** means that we're willing to detain at most $^{1}/_{t}$ extra defendants to prevent one extra violent crime.

A double standard



We could detain fewer members of the blue group while decreasing overall detention and crime.

Fairness of a single threshold



Equally risky people are treated equally, regardless of group membership. No “taste-based” discrimination. Inline with legal norms. This is what is done in practice.

Part II

Prevailing mathematical definitions of *fairness*

Popular mathematical definitions of *fairness*

Calibration

[Outcome is independent of group membership given risk.]

Classification parity

[e.g., false positive rates are equal across groups.]

Anti-classification

[Protected characteristics are not used by the algorithm.]

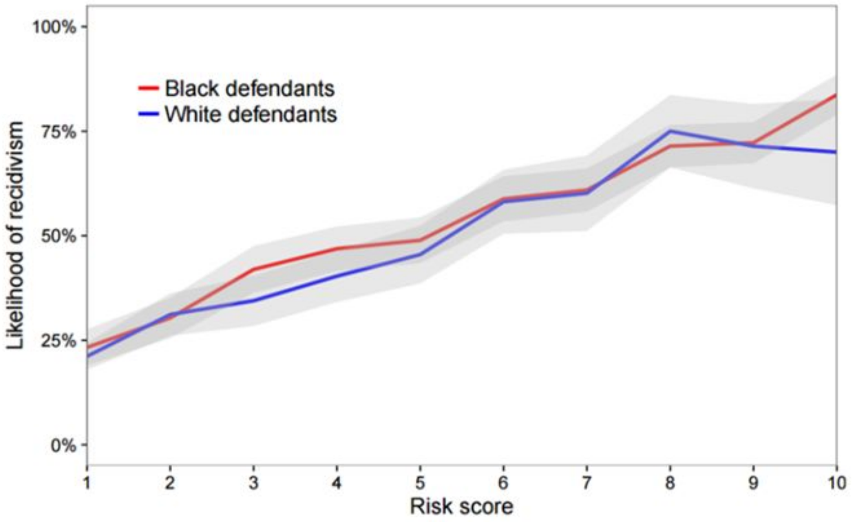
Popular mathematical definitions of *fairness*

All three definitions are problematic formalizations of long-standing legal and social norms.

1. **Calibration** does not preclude taste-based discrimination
2. **Classification parity** almost always leads to taste-based discrimination
3. **Anti-classification** often leads to taste-based discrimination

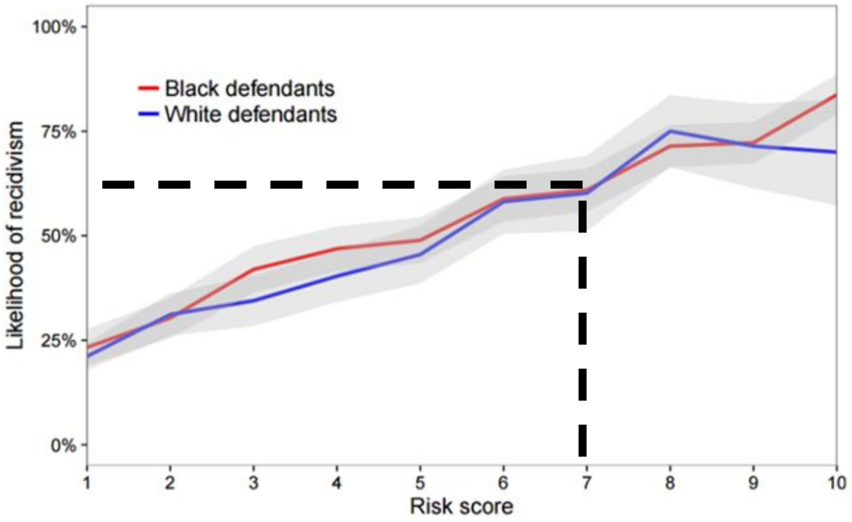
Calibration

Conditional on risk score, groups should reoffend at equal rates

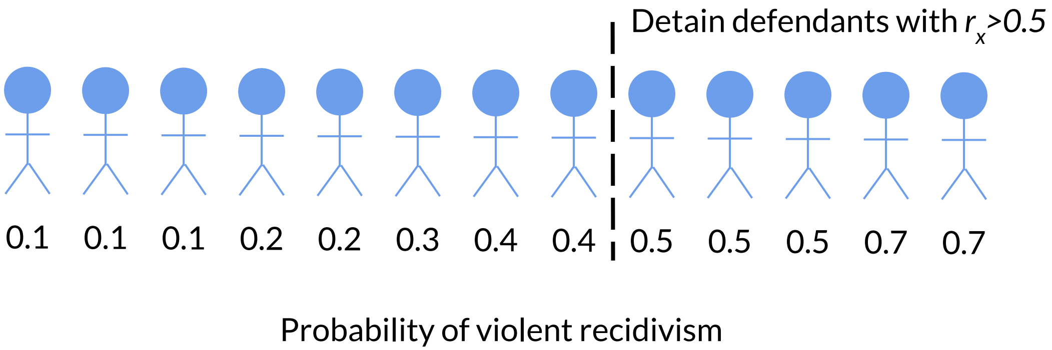


Calibration

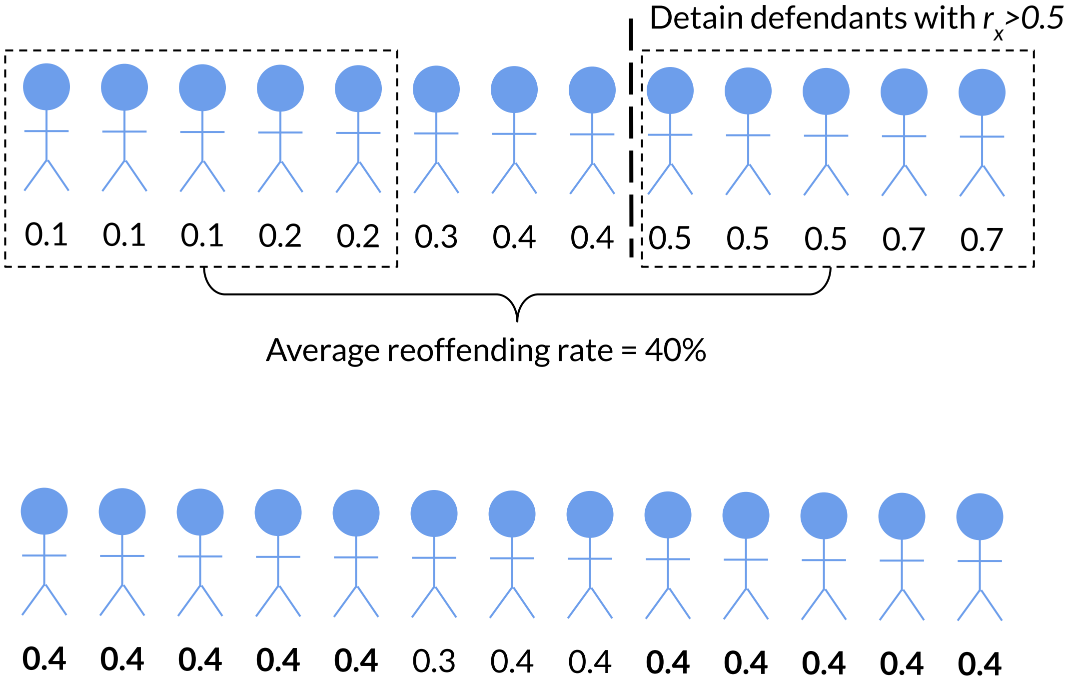
Conditional on risk score, groups should reoffend at equal rates



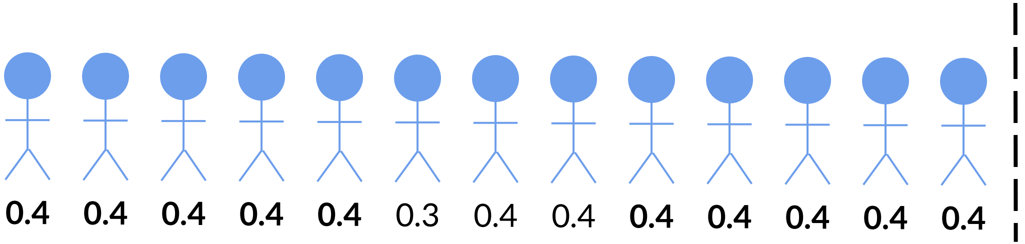
Discrimination with calibrated scores



A new set of calibrated scores



A new set of calibrated scores



The scores are still calibrated, but no blue defendants are detained.

In practice this could be achieved by choosing features that aren't predictive for the blue group.

Ensuring calibrated scores don't discriminate

We can't assess the fairness of an algorithm without seeing the features used. [Since informative features may have been ignored to discriminate; modern version of redlining.]

Algorithm designers should train the *best* risk scores possible. [Omitting features can lead to discrimination.]

Classification parity

Defines fairness as requiring equality in some aggregate statistic across groups.

False positive rate
False negative rate [1 - recall]
Positive predictive value [precision]
Negative predictive value
Proportion classified positive [e.g., detention rates]

False positive rate parity

The false positive rates are equal for all groups.

$$\text{False positive rate} = \frac{\text{Wouldn't have reoffended \& "high risk"}}{\text{Wouldn't have reoffended}}$$

ProPublica used this definition to allege bias in COMPAS.

Error rate disparities in Broward County

31% vs. **15%**

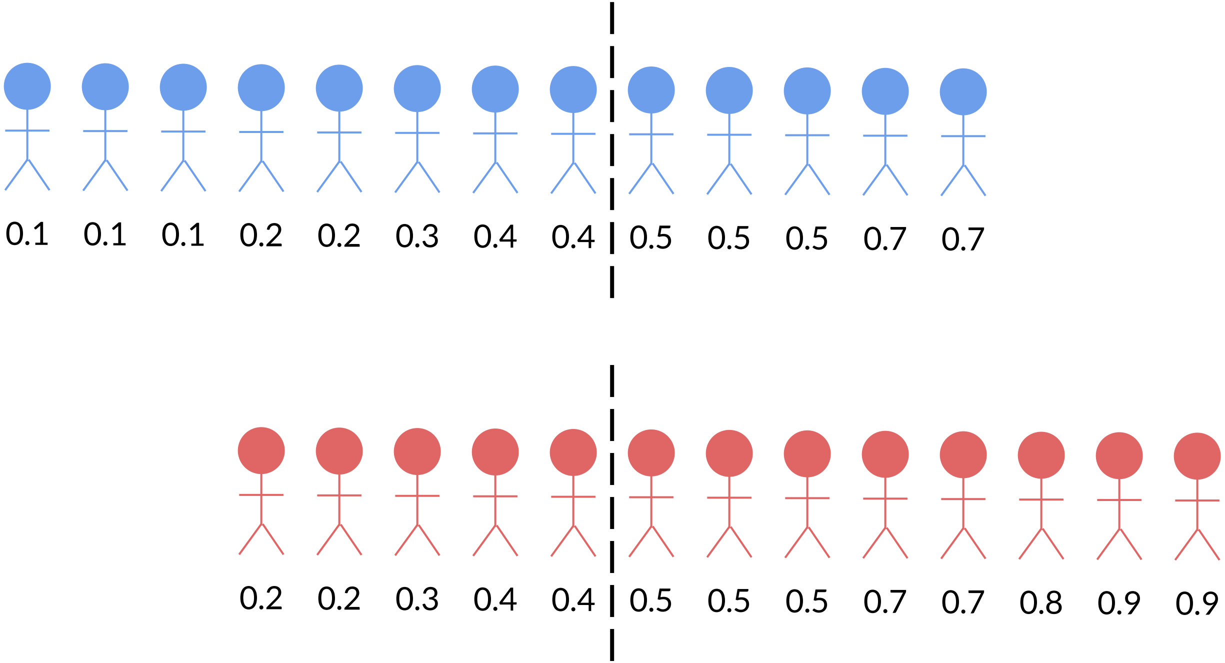
of black defendants **who did not reoffend**

of white defendants **who did not reoffend**

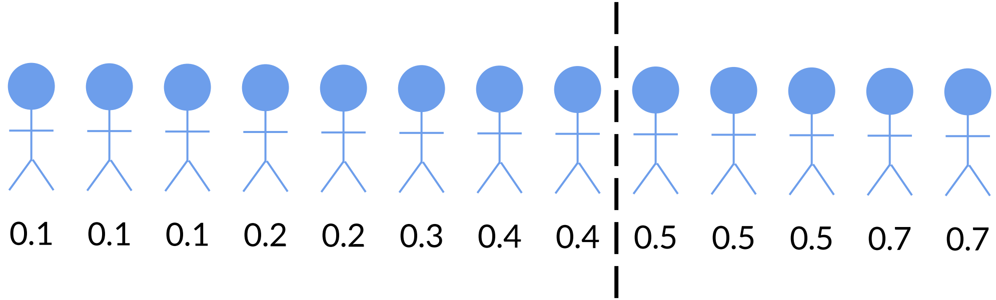
were deemed **high risk** of committing a violent crime

[Higher false positive rates for black defendants]

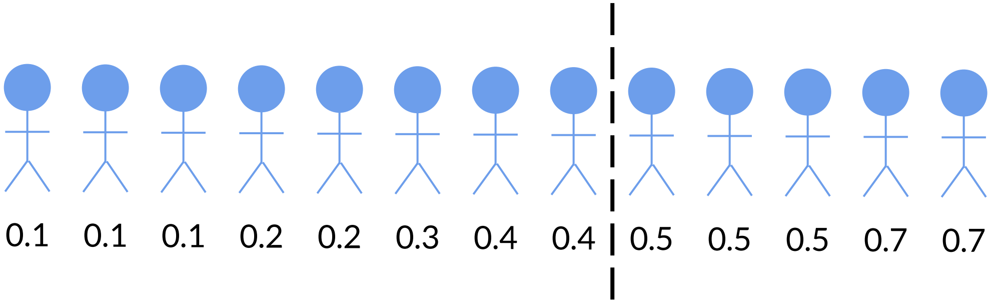
Calculating false positive rates



Calculating false positive rates

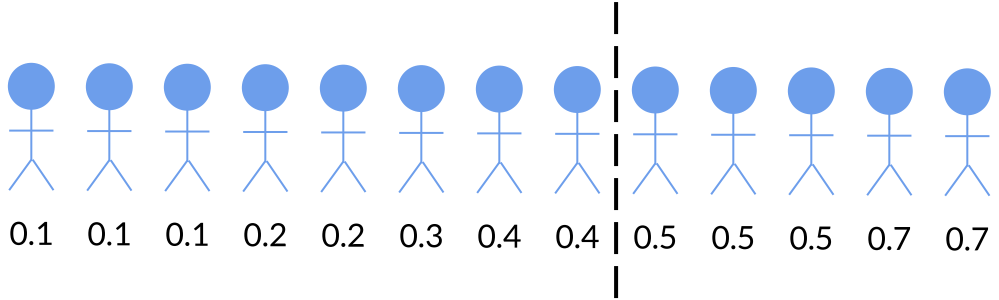


Calculating false positive rates

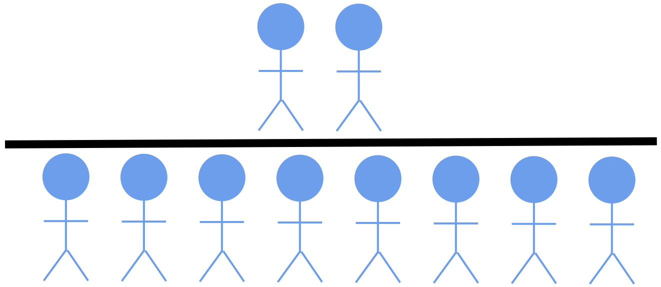


$$\frac{\text{Did not reoffend \& “high risk”}}{\text{Did not reoffend}} =$$

Calculating false positive rates

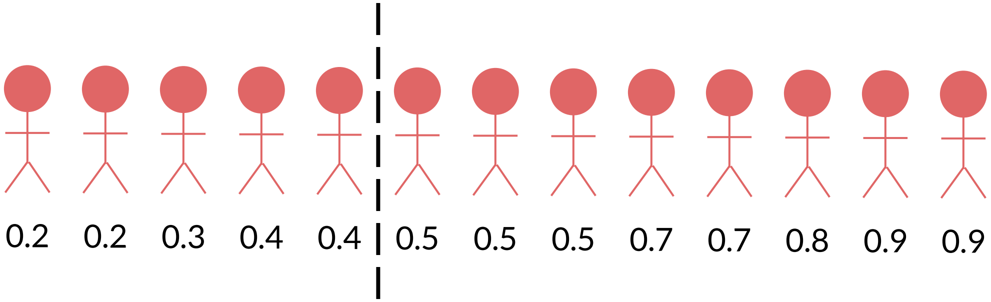


$\dfrac{\text{Did not reoffend \& "high risk"}}{\text{Did not reoffend}}$ =

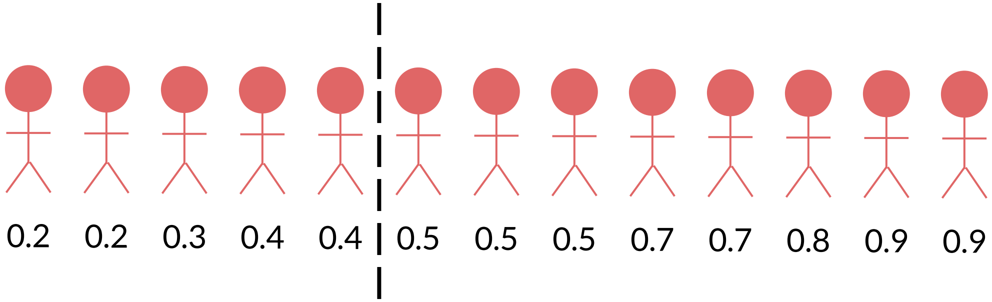


= **25%** false positive rate

Calculating false positive rates



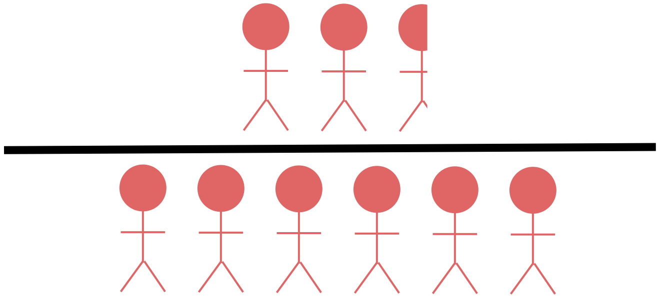
Calculating false positive rates



Did not reoffend & “high risk”

Did not reoffend

=

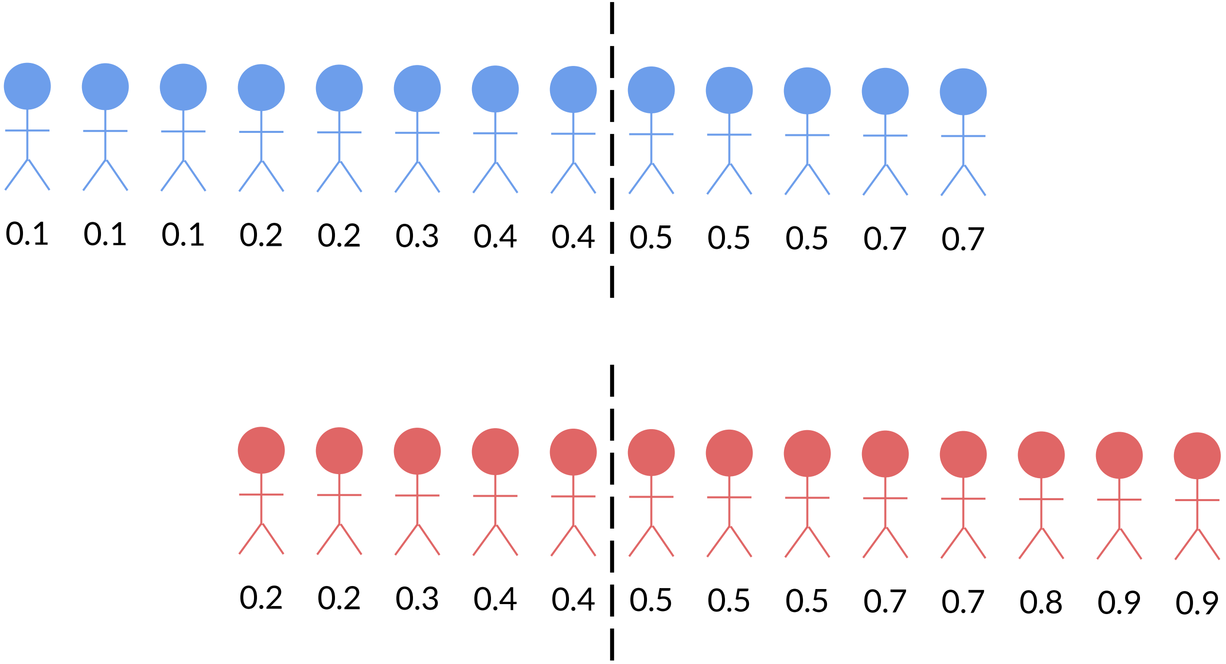


=

40%

false positive rate

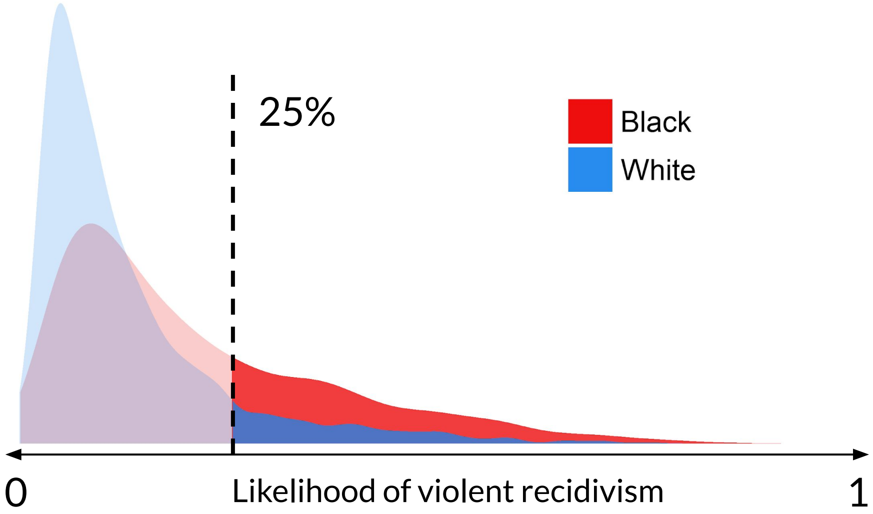
Calculating false positive rates



25%
false positive rate

40%
false positive rate

Why do false positive rates differ?



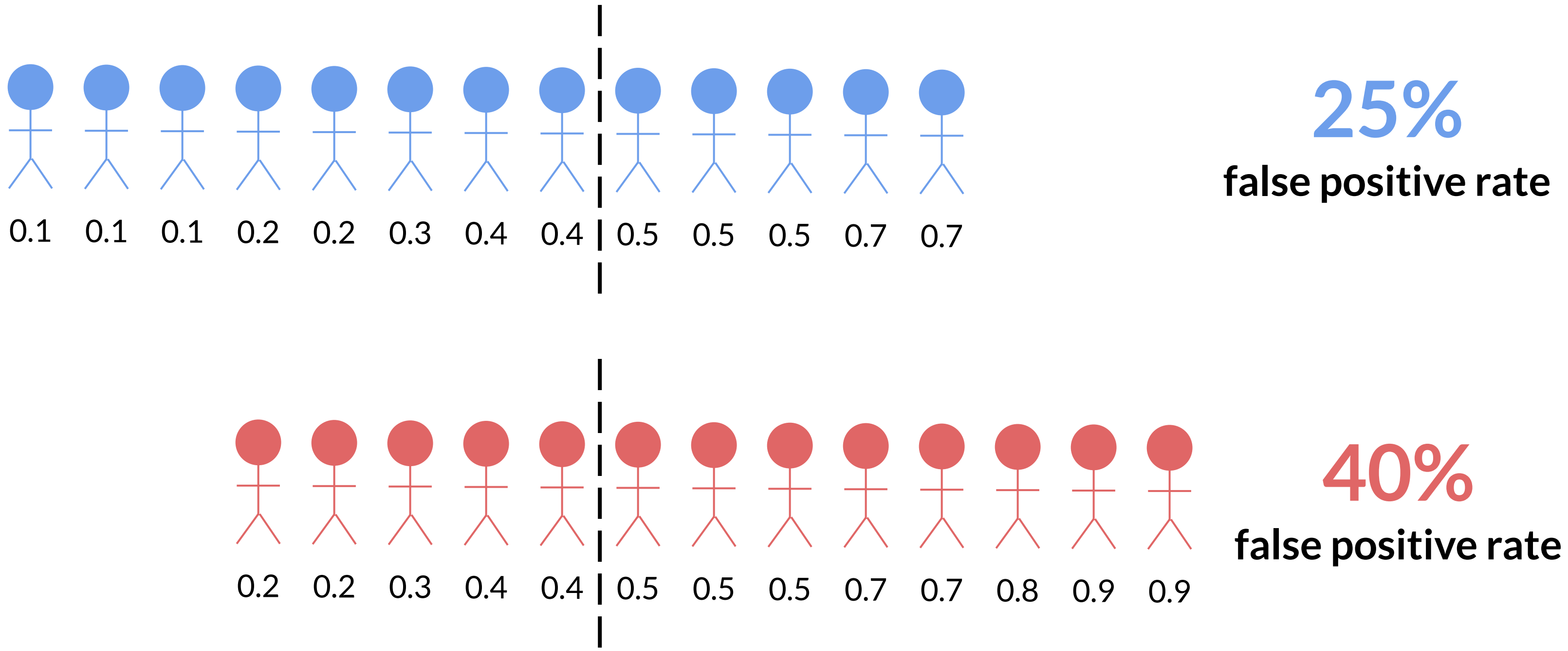
Black and white defendants have different risk distributions

Infra-marginality

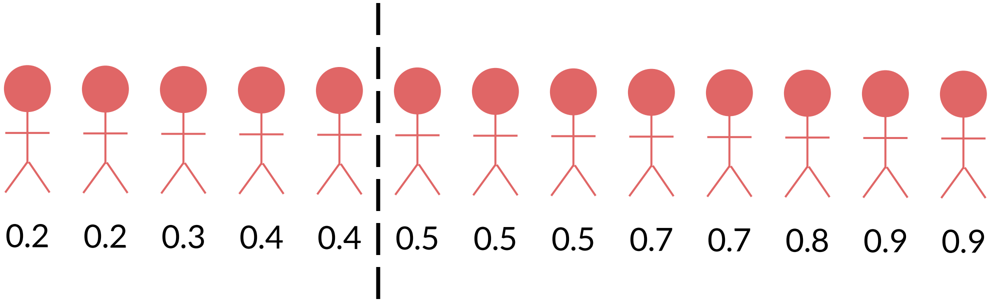
The false positive rate is an *infra-marginal* statistic—it depends not only on a group's threshold but on its distribution of risk.

Infra-marginal statistics are misleading proxies for the threshold when risk distributions differ.

The problem with false positive rates

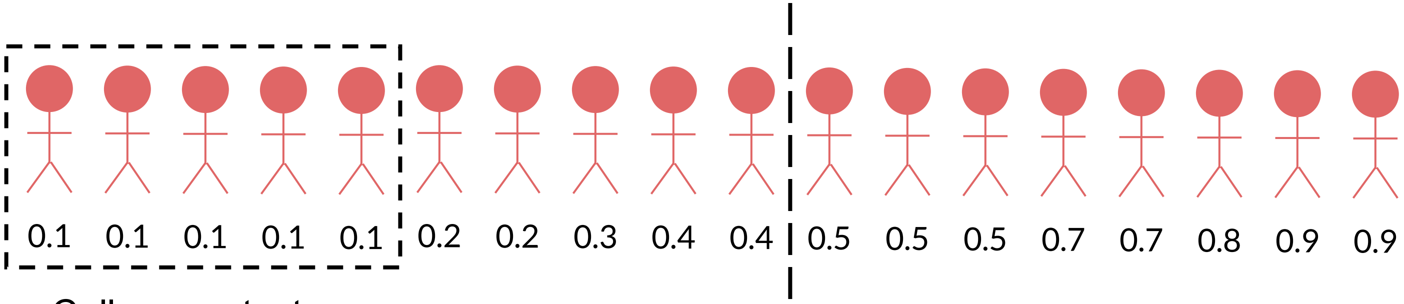


The problem with false positive rates



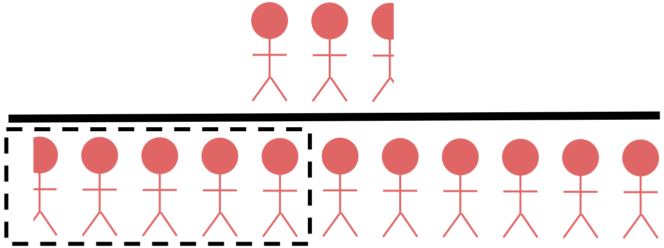
The problem with false positive rates

The problem with false positive rates



Did not reoffend & "high risk"
/
Did not reoffend

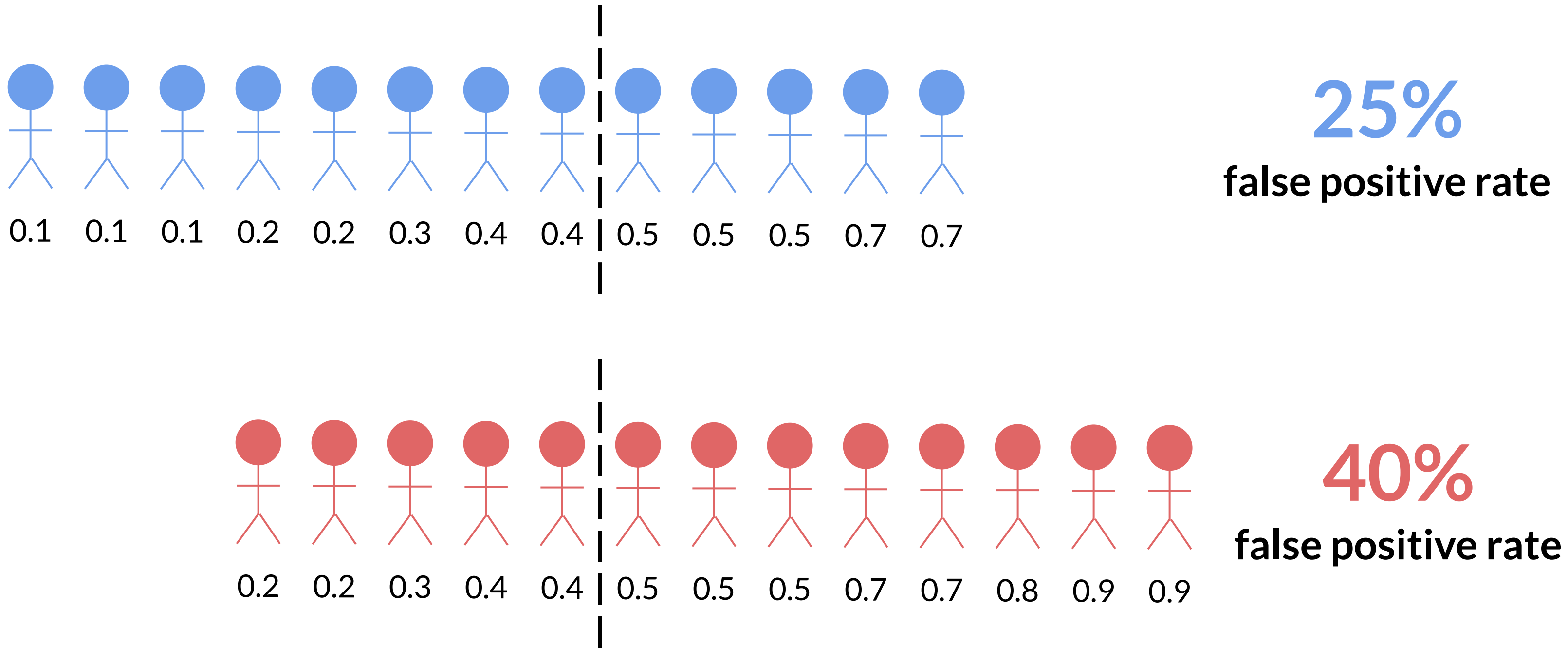
=



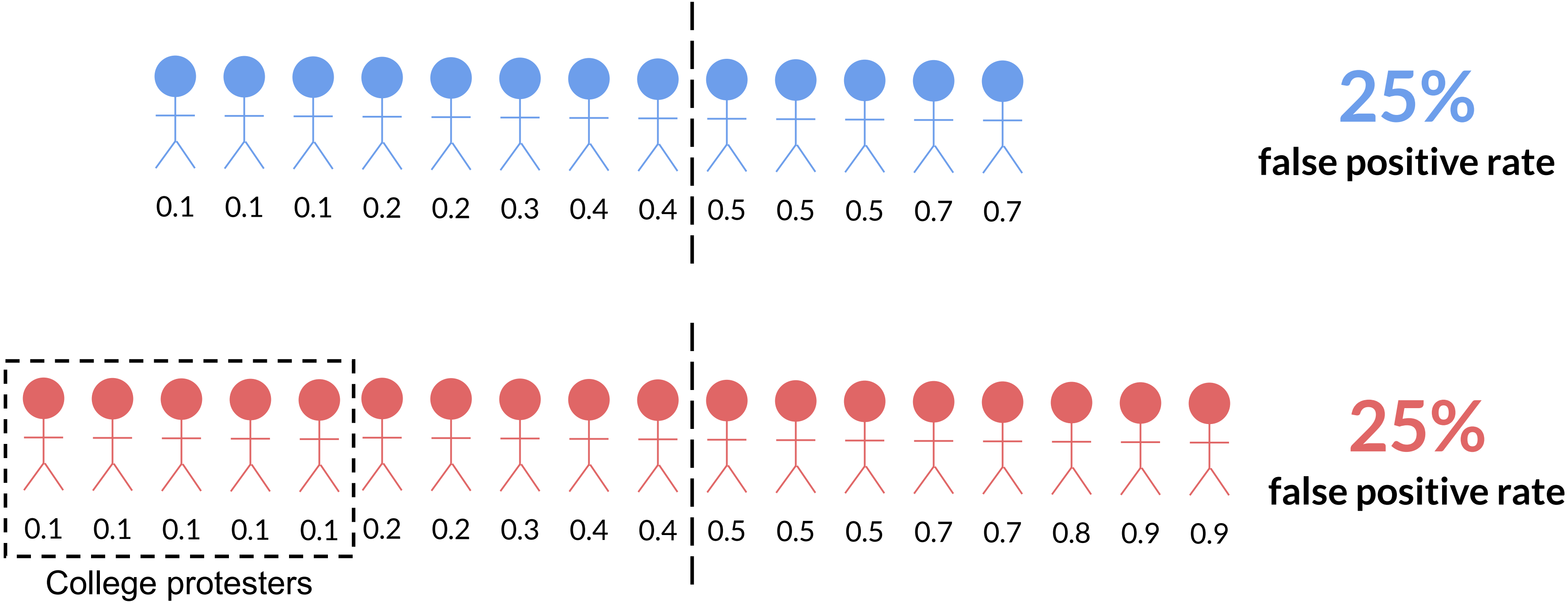
=

25%
false positive rate

The problem with false positive rates



The problem with false positive rates



Classification parity

Many proposed definitions of fairness try to equalize some aggregate statistic between groups.
[Precision parity, statistical parity, recall parity, equalized odds]

All these definitions compare **infra-marginal** statistics, so they have the same problems as false positive rates. They are all unreliable measures of discrimination.

Anti-classification

Intuitively, a *fair* algorithm shouldn't use protected classes.
[e.g., decisions shouldn't explicitly depend on race or gender.]

Many have argued a fair algorithm thus shouldn't use *proxies.*

The problem with anti-classification

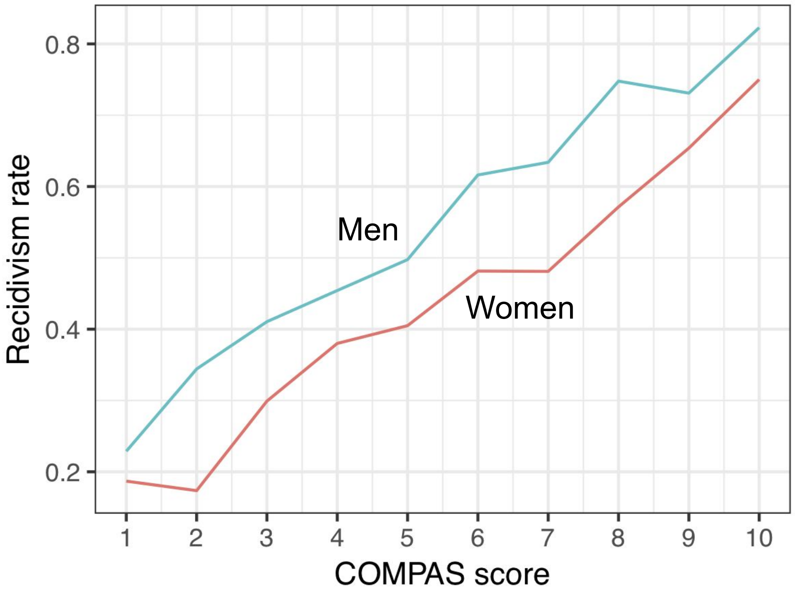
Under traditional legal and economic notions of fairness, it may be warranted to use protected class when making certain decisions.

The problem with anti-classification

In Broward County, women are less likely to reoffend than men of the same age with similar criminal histories.

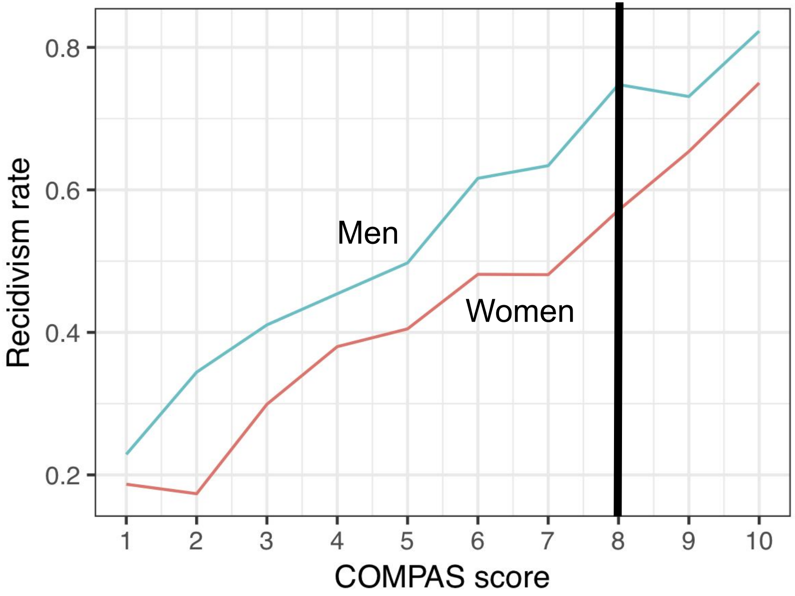
A gender-blind risk score

Broward County, Florida



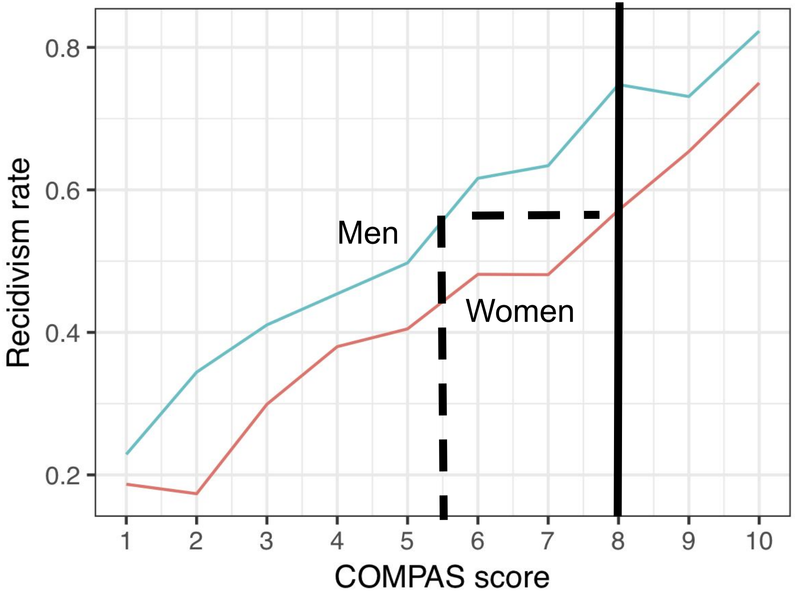
A gender-blind risk score

Broward County, Florida



A gender-blind risk score

Broward County, Florida



The problem with anti-classification

Gender-blind risk models can lead to taste-based discrimination.

One can fix this problem by using one model for men and another for women [or by including gender in the model].
[Wisconsin uses gender-specific risk assessment tools.]

Part III

Bias in the data

Are the data *biased*?

Two types of bias:

1. Biased labels
   [***Y*** doesn’t perfectly measure what we care about]

2. Biased predictors
   [Features that are differentially predictive]

Biased labels

St. George’s Hospital in the UK developed an algorithm to sort medical school applicants. Algorithm trained to mimic past admissions decisions made by humans.

Biased labels

St. George's Hospital in the UK developed an algorithm to sort medical school applicants. Algorithm trained to mimic past admissions decisions made by humans.

But past decisions were biased against women and minorities.
[The algorithm codified discrimination.]

Biased labels

In reality we measure who is *arrested* or *convicted*, not who [would have] committed a crime.

Biased labels

In reality we measure who is *arrested* or *convicted*, not who [would have] committed a crime.

Increased policing in minority areas might make certain arrest types [e.g., for drugs] a problematic measure of actual crime.

Some outcomes [e.g., violent crime] seem less prone to measurement error.

Biased predictors

Marijuana arrests are likely *biased*: minority users more likely to be arrested than white users.

Including it in the model will overstate the risk of minorities.
[Conditional on marijuana arrests, white defendants are more likely to reoffend.]

Biased predictors

Marijuana arrests are likely *biased*: minority users more likely to be arrested than white users.

Including it in the model will overstate the risk of minorities. [Conditional on marijuana arrests, white defendants are more likely to reoffend.]

If the labels are unbiased, we can fix biased predictors with appropriate interactions. [Contrary to anti-classification.]

Biased predictors

“In New Orleans, when I worked there as a public defender, the significance of arrest varied by race. If a black man had three arrests in his past, it suggested only that he had been living in New Orleans. Black men were arrested all the time for trivial things. If a white man had three past arrests, on the other hand, it suggested that he was really bad news!”
[Sandra Mayson, “Bias in, bias out”]

Part IV

Coda

Math ≠ equity

There are many formal, mathematical definitions of fairness.

Nearly none of these definitions map to established legal or social understandings of equity.

Algorithms ≠ policy

Statistical algorithms are often good at synthesizing information, but we must still set effective and equitable policy.

In the case of pretrial decisions, we might limit money bail and/or consider non-custodial interventions.

Stanford Computational Policy Lab

policylab.stanford.edu