STEVENS
INSTITUTE OF TECHNOLOGY
1870

Fairness in Machine Learning

Reference to: Tutorial at the 19th ACM Conference on Economics and Computation (EC'18), Ithaca, NY, June 18, 2018

STEVENS
INSTITUTE OF TECHNOLOGY
1870

Measuring Discrimination in Human Decisions

Reference to: Tutorial at the 19th ACM Conference on Economics and Computation (EC'18), Ithaca, NY, June 18, 2018

Discrimination in the Law

- **Disparate treatment**
  - Individuals are treated differently due to prejudice against racial, gender, and protected trait groups.

    [Equal Protection Clause of the 14th Amendment.]
- **Unjustified disparate impact**
  - A policy that appears neutral delivers differential results that cannot be justified by a valid, nondiscriminatory interest.

    [Civil Rights Act. Fair Housing Act. And various state statutes.]

Discrimination in Economics

- **Taste-based Discrimination**
  - The decision maker shows a willingness to discriminate at the expense of utility.

- **Statistical discrimination**
  - To maximize profit, the decision-maker draws logical conclusions based on group membership.

Reference to: Tutorial at the 19th ACM Conference on Economics and Computation (EC'18), Ithaca, NY, June 18, 2018

Taste-based Discrimination

- Decision makers derive utility from discriminating and thus act sub-optimally [relative to a profit-maximizing agent].

  - **An example**:

    A mother hires a less-qualified female nanny over a more-qualified male nanny to satisfy the employer's gender bias.

Reference to: Tutorial at the 19th ACM Conference on Economics and Computation (EC'18), Ithaca, NY, June 18, 2018
Figure from: https://creazilla.com/nodes/36077-childcare-worker-man-clipart

Testing for Discrimination

- Applying the same standard to all individuals is optimal [e.g., hiring everyone above a certain threshold, regardless of group membership]

- Statistical discrimination tests in human decisions often aim to determine whether decision-makers apply different standards to different groups.

A motivating Example

Vehicle searches

- Police need probable cause to search a vehicle for contraband.
  - Do officers apply the probable cause standard equally to drivers of all races? [If not, they could find more contraband while conducting the same number of searches.]

Reference to: Tutorial at the 19th ACM Conference on Economics and Computation (EC'18), Ithaca, NY, June 18, 2018
Figure from: https://www.istockphoto.com/illustrations/police-searching-car

Benchmark Test

A simple test for discrimination

- Are white and black drivers searched at similar rates?
  - A higher search rate for black drivers might indicate a lower (and discriminatory) bar for searching them.
  - But without more information, it could also be the case that whites and blacks are held to the same standard, but that more black drivers are above the search threshold.

Reference to: Tutorial at the 19th ACM Conference on Economics and Computation (EC'18), Ithaca, NY, June 18, 2018

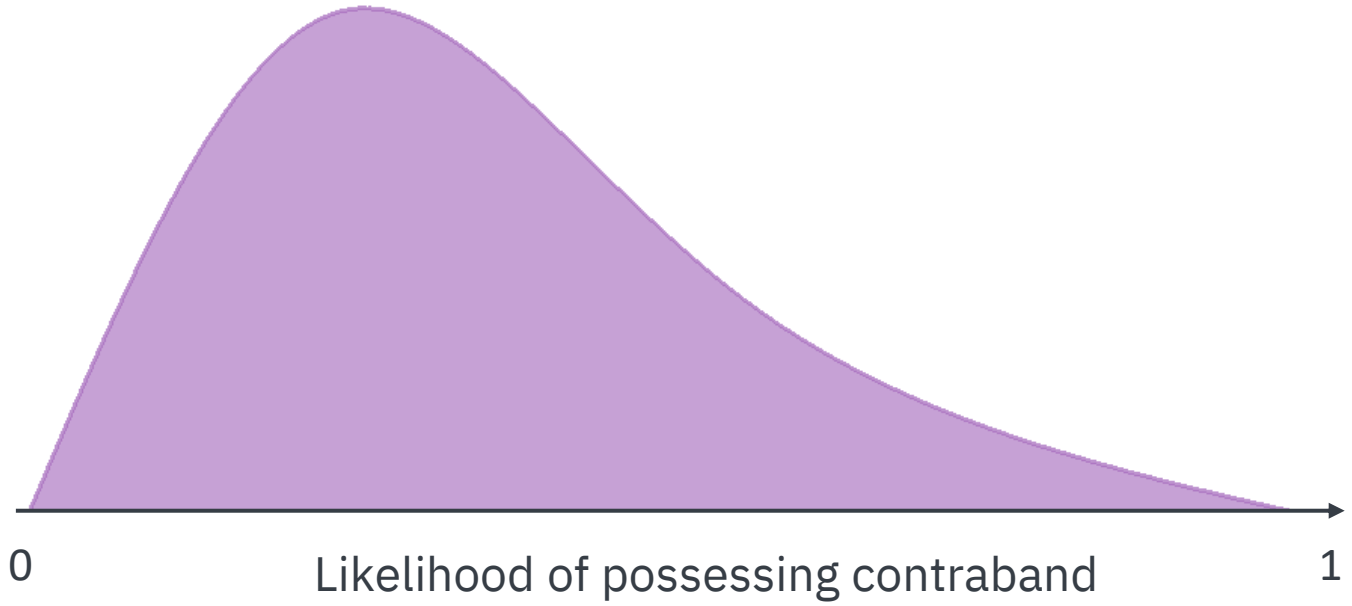
Outcome Test

A “better” test for discrimination

- Are the search outcomes the same for drivers of different races? [i.e., are the hit rates equal?]
  - A lower hit rate for black drivers might mean they are searched on the basis of less evidence.

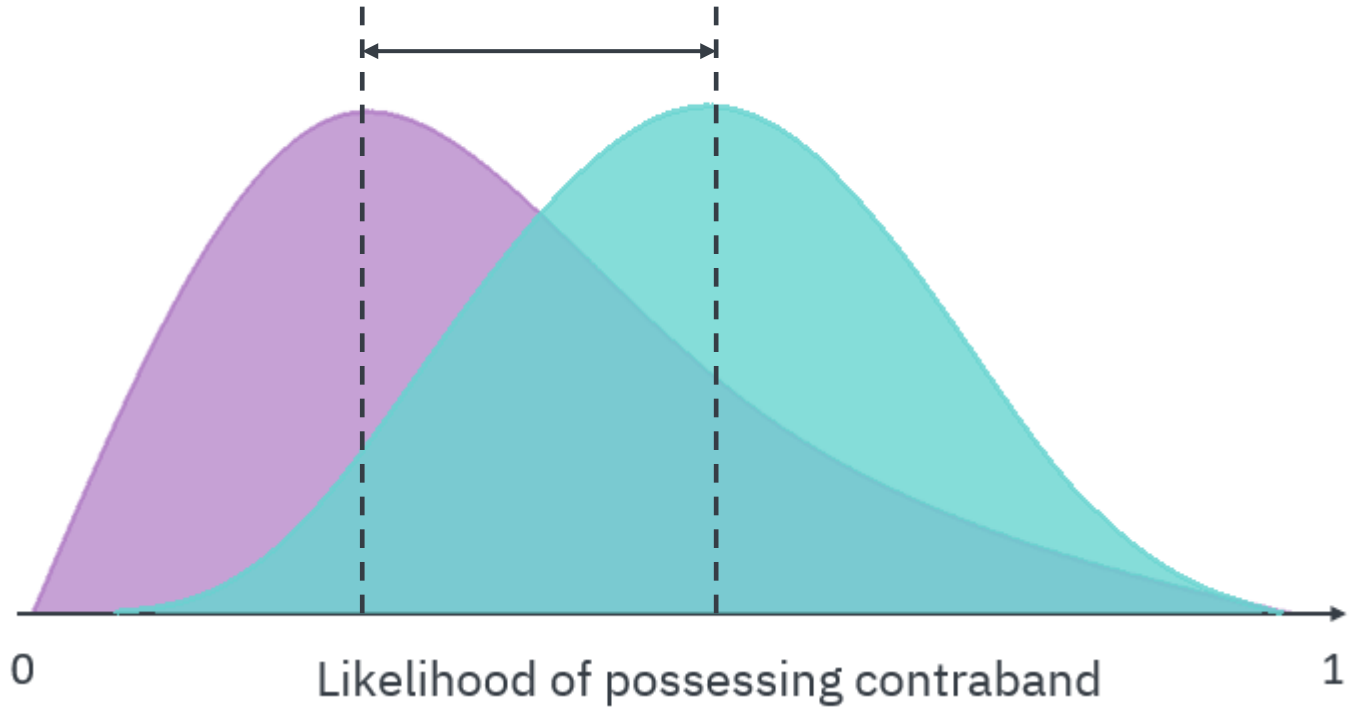
Reference to: Tutorial at the 19th ACM Conference on Economics and Computation (EC'18), Ithaca, NY, June 18, 2018
Figure from: https://www.istockphoto.com/illustrations/police-search-car

Risk Distributions



Reference to: Tutorial at the 19th ACM Conference on Economics and Computation (EC'18), Ithaca, NY, June 18, 2018

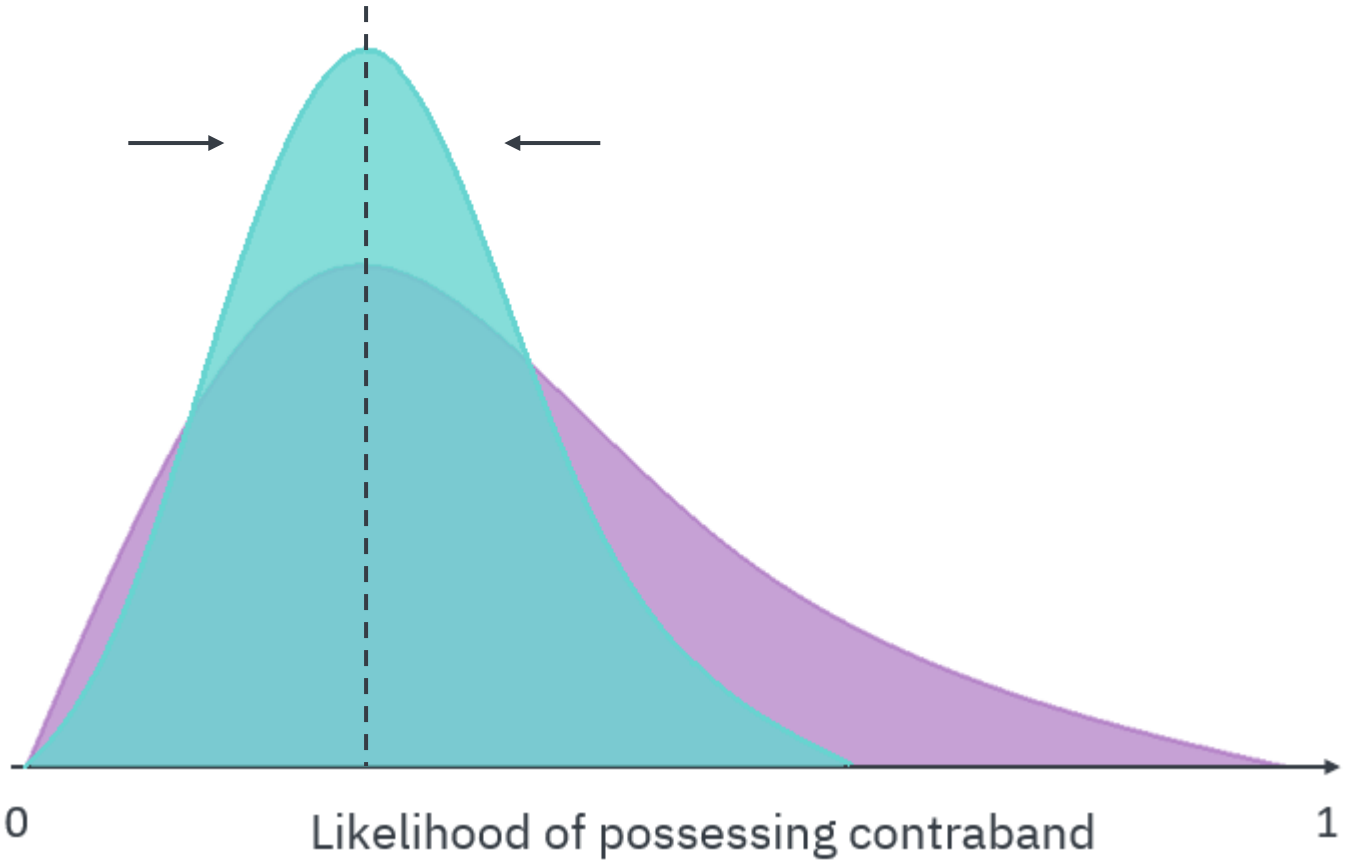
Risk Distributions



Different average risk = different rates of carrying contraband

Reference to: Tutorial at the 19th ACM Conference on Economics and Computation (EC'18), Ithaca, NY, June 18, 2018

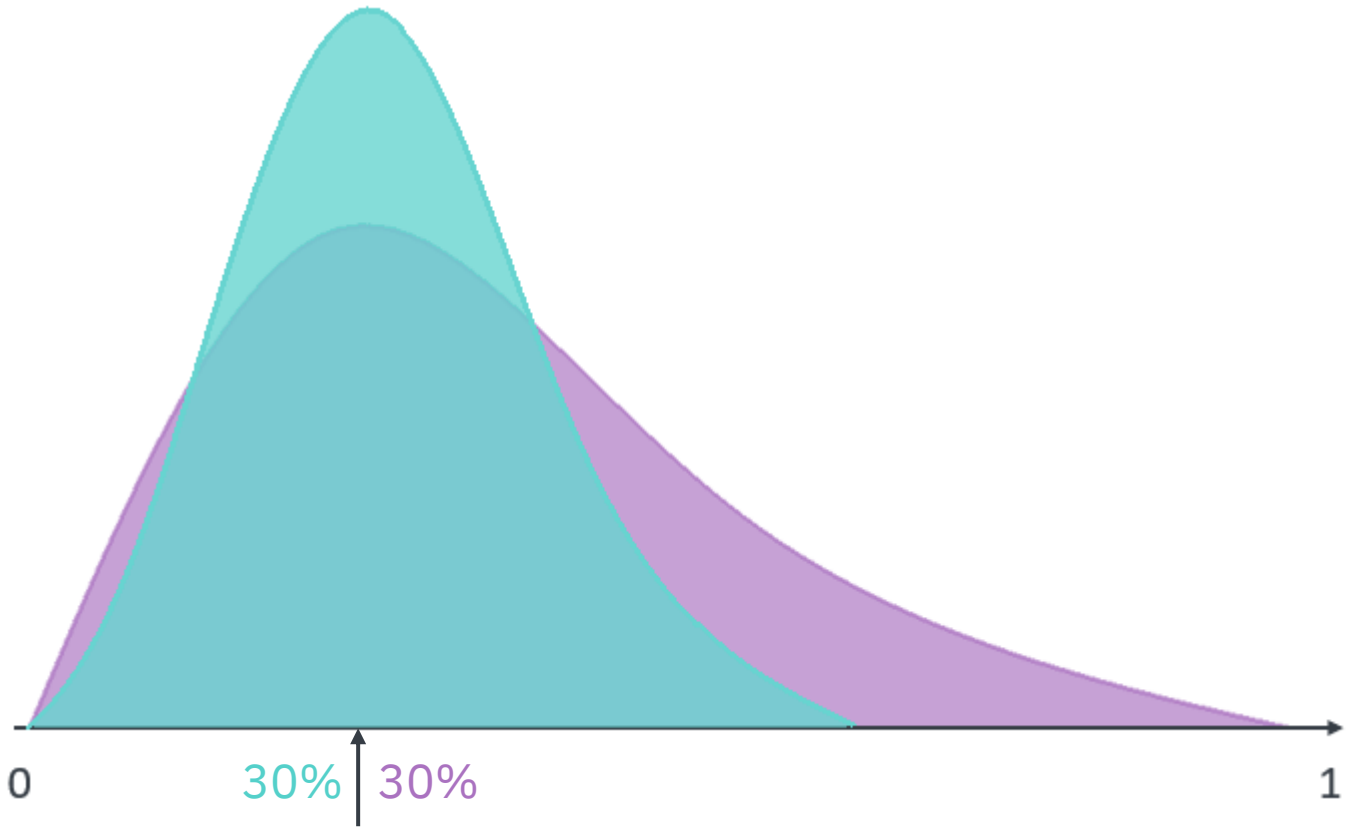
Risk Distributions



Lower variance = harder to determine who's carrying contraband

Reference to: Tutorial at the 19th ACM Conference on Economics and Computation (EC'18), Ithaca, NY, June 18, 2018

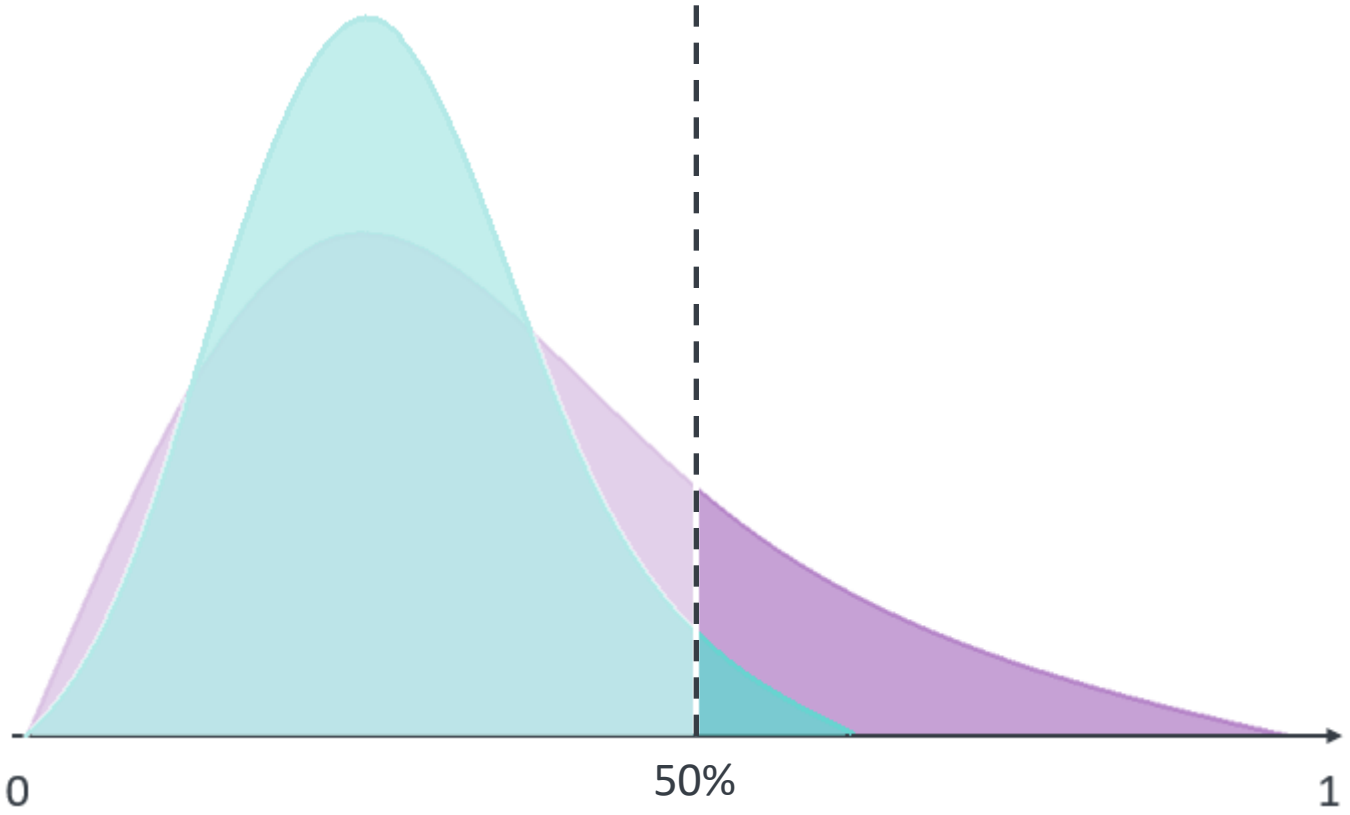
The Problem with the Outcome Test



Two groups who both carry contraband 30% of the time.

Reference to: Tutorial at the 19th ACM Conference on Economics and Computation (EC'18), Ithaca, NY, June 18, 2018

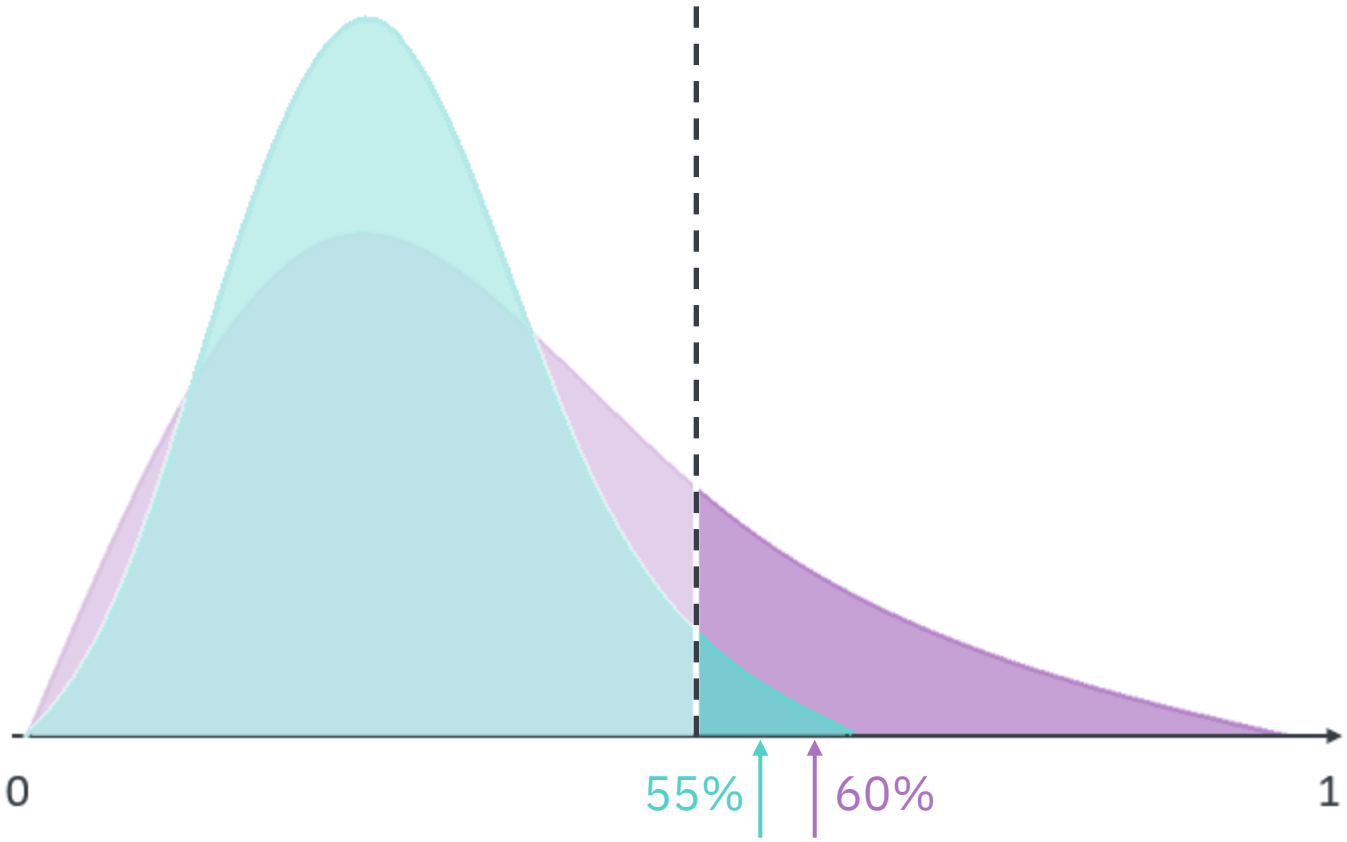
The Problem with the Outcome Test



Police search if there's a greater than 50% chance they'll find contraband. [A facially neutral, non-discriminatory search policy.]

Reference to: Tutorial at the 19th ACM Conference on Economics and Computation (EC'18), Ithaca, NY, June 18, 2018

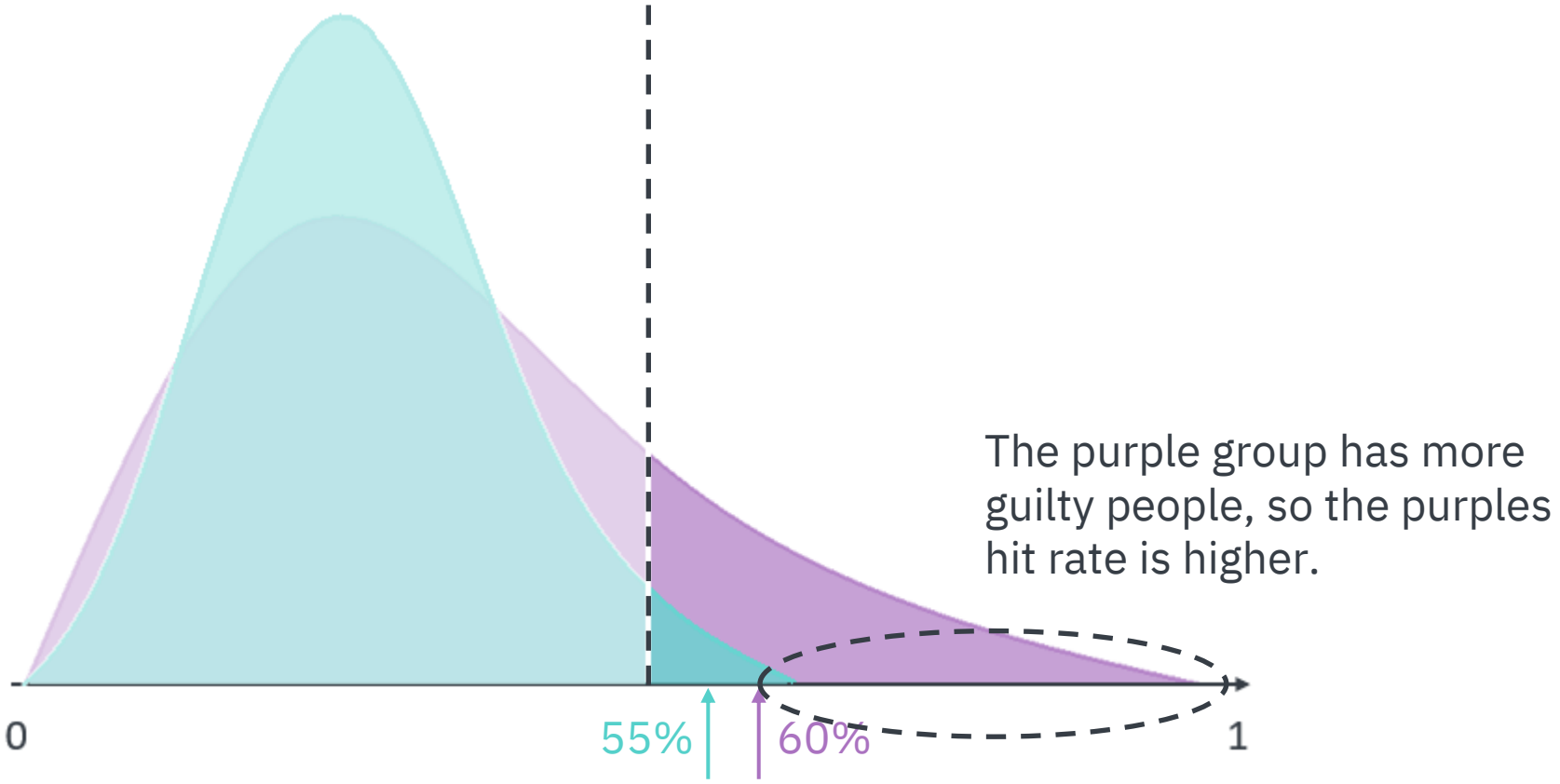
The Problem with the Outcome Test



Searches of the purple group are more successful.
[The outcome test incorrectly suggests bias against the green group.]

Reference to: Tutorial at the 19th ACM Conference on Economics and Computation (EC'18), Ithaca, NY, June 18, 2018

The Problem with the Outcome Test

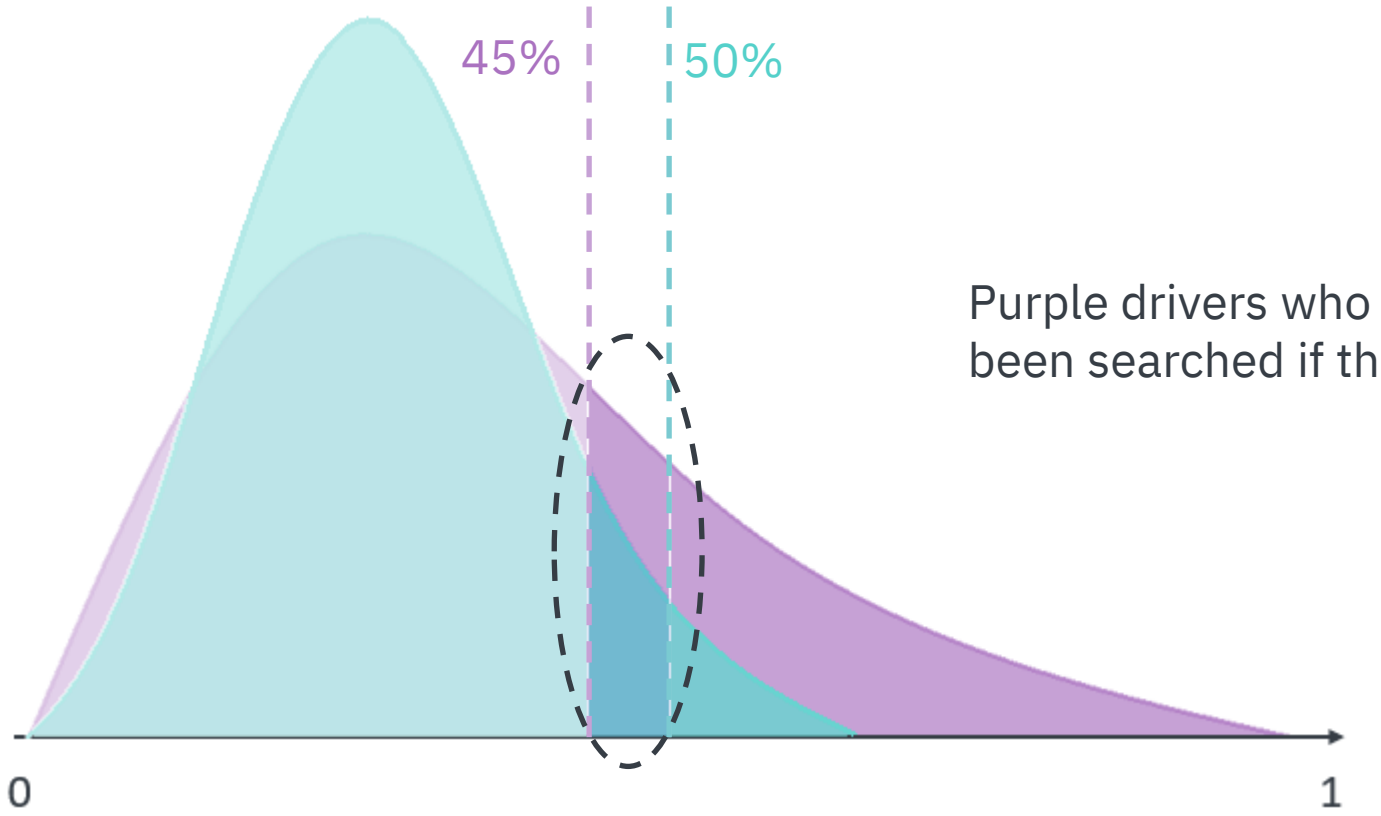


The purple group has more guilty people, so the purples hit rate is higher.

Searches of the purple group are more successful.
[The outcome test incorrectly suggests bias against the green group.]

Reference to: Tutorial at the 19th ACM Conference on Economics and Computation (EC'18), Ithaca, NY, June 18, 2018

The Problem with the Outcome Test

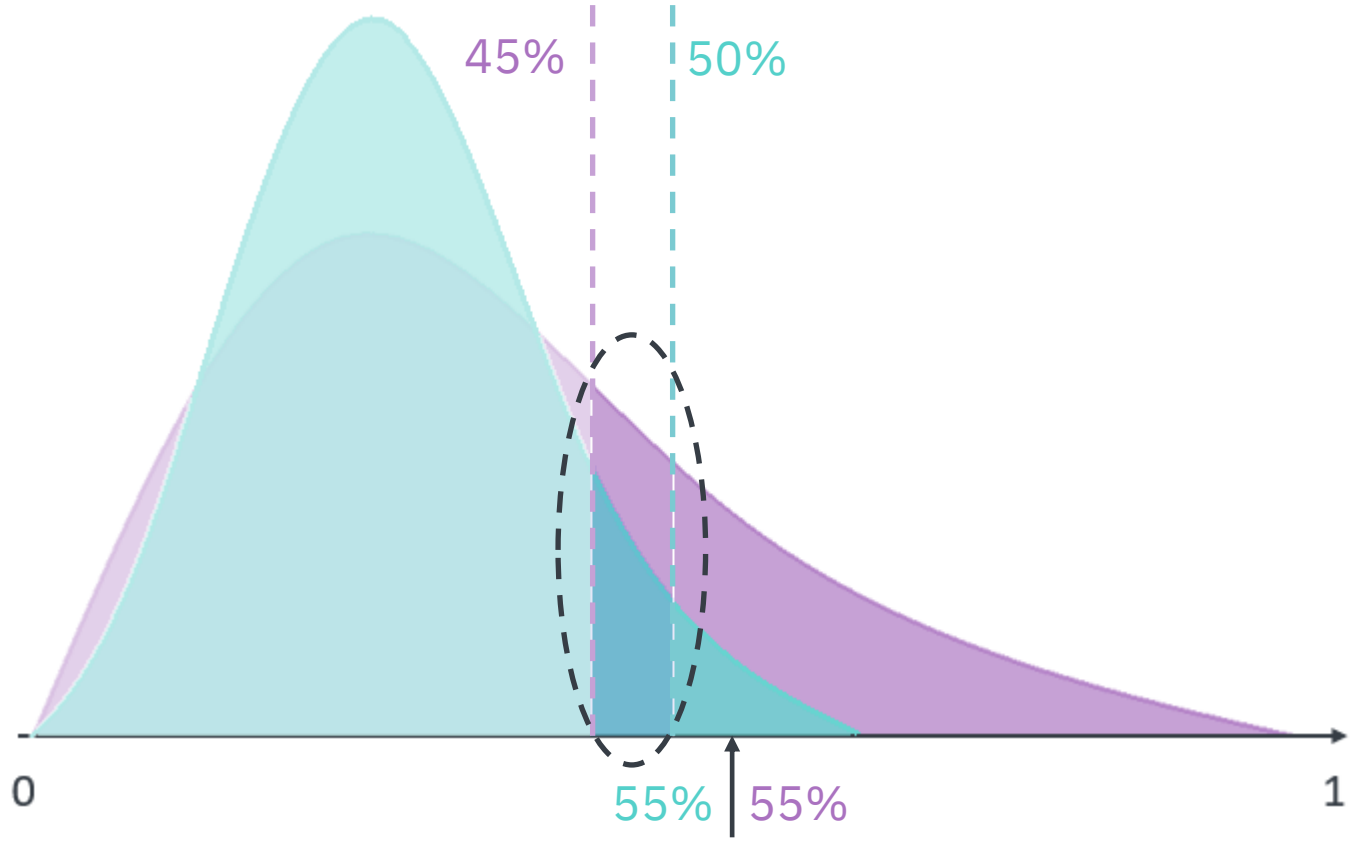


Purple drivers who wouldn't have been searched if they were green.

Police are acting sub-optimally to discriminate against purple drivers.

Reference to: Tutorial at the 19th ACM Conference on Economics and Computation (EC'18), Ithaca, NY, June 18, 2018

The Problem with the Outcome Test



But the hit rates are equal.
[The outcome test incorrectly finds no discrimination.]

The Problem of Infra-Marginality

- Infra marginal statistics-those that average over individuals away from the margin-depend on both the threshold applied and the distribution of risk.

- These statistics are imperfect proxies for the threshold, and hence problematic measures of taste-based discrimination.

Reference to: Tutorial at the 19th ACM Conference on Economics and Computation (EC'18), Ithaca, NY, June 18, 2018

Infra-Marginality in Raleigh, NC

- 16% vs. 13%

  Black hit rate — White hit rate

- Searches of black drivers are *more* successful than searches of white drivers, and so the outcome test suggests bias against *white* drivers.

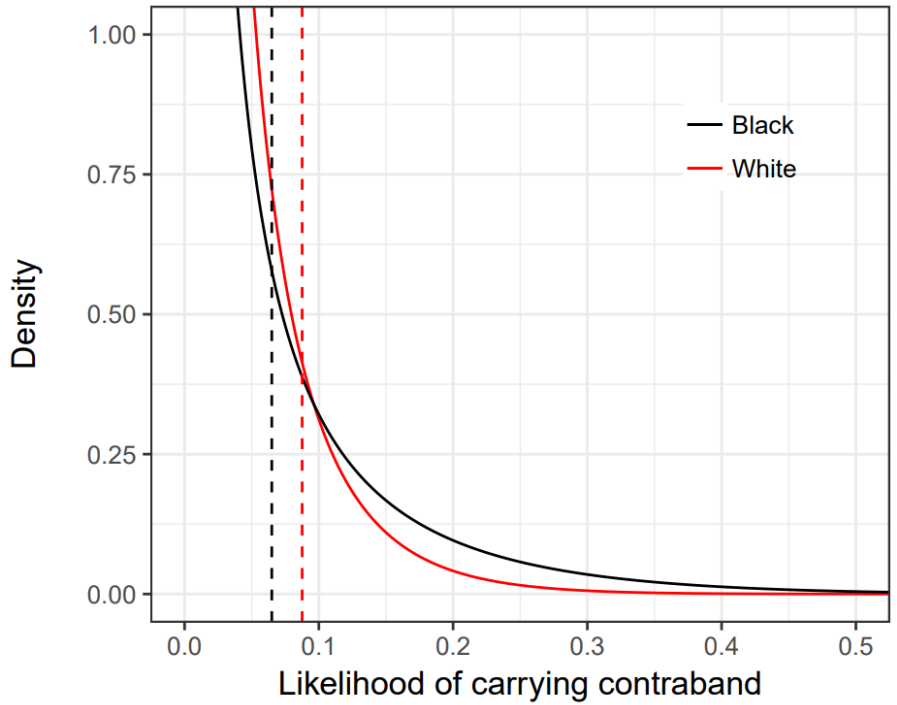
Reference to: Tutorial at the 19th ACM Conference on Economics and Computation (EC'18), Ithaca, NY, June 18, 2018

Infra-Marginality in Raleigh, NC

- Black drivers in Raleigh are three times more likely to carry contraband in plain view, and so their risk distribution has a very heavy tail.

Infra-Marginality in Raleigh, NC

- Black drivers in Raleigh are three times more likely to carry contraband in plain view, and so their risk distribution has a very heavy tail.



- Tests for discrimination that account for the shape of the risk distributions find that officers apply a lower standard when searching blacks. [Simoiu et al., 2017]
- Infra-marginality is real. In this case, the outcome test failed to detect bias against black drivers.

Reference to: Tutorial at the 19th ACM Conference on Economics and Computation (EC'18), Ithaca, NY, June 18, 2018

Connection to ML Predictions

- We can think of the hit rate as the precision of human decision makers. It is the fraction of those classified positive (i.e., searched) who had contraband.

- As the Raleigh example shows, precision can be a misleading proxy for the threshold is applied, and thus is a problematic measure of discrimination.

Reference to: Tutorial at the 19th ACM Conference on Economics and Computation (EC'18), Ithaca, NY, June 18, 2018

STEVENS
INSTITUTE OF TECHNOLOGY
1870

Measuring discrimination in algorithmic decisions

Algorithmic Risk Assessments

- Many high-stakes decisions are made by first estimating the risk of an individual based on the available information.
- Lending is based on the risk of default; pretrial detention is based on the risk of pretrial recidivism.
- Decisions guided by statistical risk assessments can be more effective and fair than those made by intuition alone.

Reference to: Tutorial at the 19th ACM Conference on Economics and Computation (EC'18), Ithaca, NY, June 18, 2018

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.

Reference to: Tutorial at the 19th ACM Conference on Economics and Computation (EC'18), Ithaca, NY, June 18, 2018

Risk Assessments in Broward County, FL

ProPublica analyzed 3,000 white and black defendants assigned COMPAS scores in Broward County, Florida. [Also determined whether these defendants recidivated.]

Reference to: Tutorial at the 19th ACM Conference on Economics and Computation (EC'18), Ithaca, NY, June 18, 2018

A risk assessment tool

The Public Safety Assessment [New Jersey, San Francisco, elsewhere]

New Violent Criminal Activity (maximum total weight = 7 points)		
Risk Factor	**Weights**	
Current violent offense	No Yes	0 2
Current violent offense & 20 years old or younger	No Yes	0 1
Pending charge at the time of the offense	No Yes	0 1
Prior conviction	No Yes	0 1
Prior violent conviction	No convictions 1 or 2 convictions 3+ convictions	0 1 2

The data

- ProPublica analyzed 3,000 white and black defendants assigned COMPAS scores in Broward County, Florida.

  [Also determine whether these defendants recidivated.]

Key Assumptions

Common to most papers on algorithmic fairness

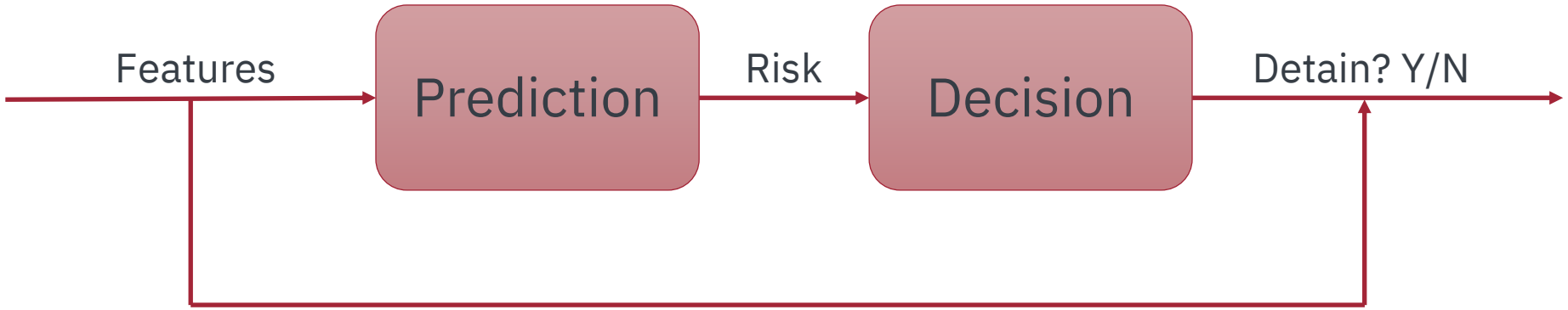
1. We know the true label $\boldsymbol{Y}$ (i.e.,. whether a defendant would have reoffended if released).

   [Y is true counterfactual, with no measurement error.]

2. We know the true risk: $\boldsymbol{r_x = P(Y = 1 \mid X = x)}$

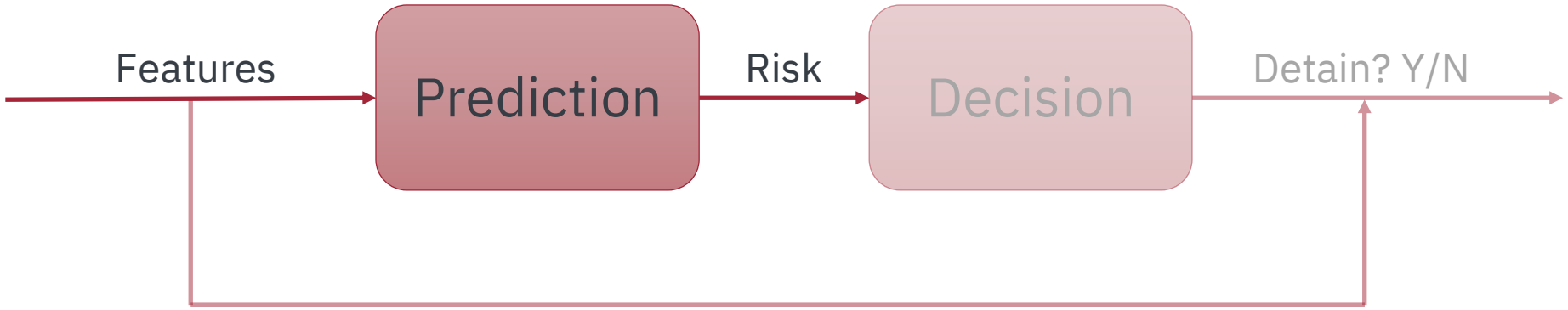
   [Reasonable when we have lots of data.]

From Features to Decisions



How should we go from feature to decisions?

From Features to Decisions



The risk $r_x = P(Y = 1 \mid X = x)$ is fixed once we choose the features X.

Reference to: Tutorial at the 19th ACM Conference on Economics and Computation (EC'18), Ithaca, NY, June 18, 2018

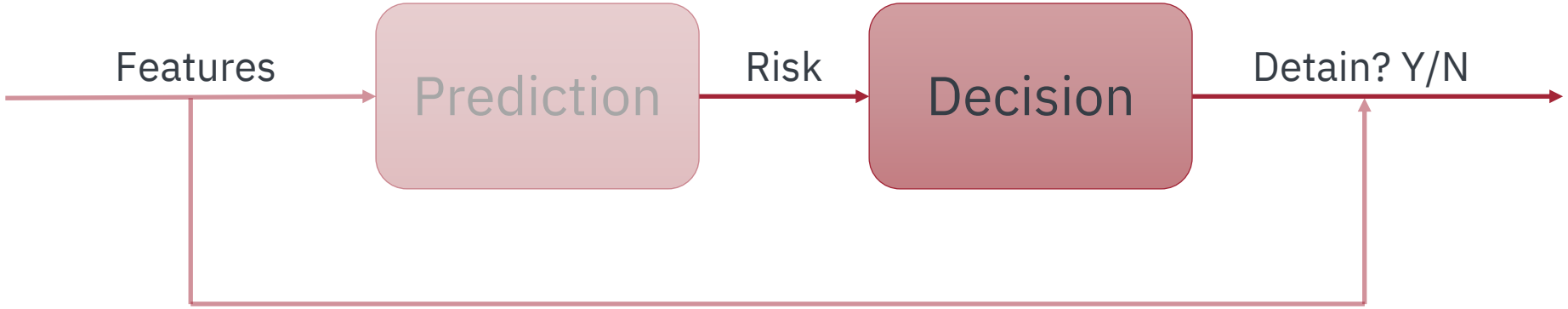
Risk distributions

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.

Reference to: Tutorial at the 19th ACM Conference on Economics and Computation (EC'18), Ithaca, NY, June 18, 2018

From Risk to Decisions



It’s common to use a threshold on risk
[The decision is independent of the features given risk.]

Reference to: Tutorial at the 19th ACM Conference on Economics and Computation (EC'18), Ithaca, NY, June 18, 2018

Choosing a Threshold

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

Reference to: Tutorial at the 19th ACM Conference on Economics and Computation (EC'18), Ithaca, NY, June 18, 2018

Applying a Threshold

A threshold rule optimally trades off between detention and recidivism.

Reference to: Tutorial at the 19th ACM Conference on Economics and Computation (EC'18), Ithaca, NY, June 18, 2018

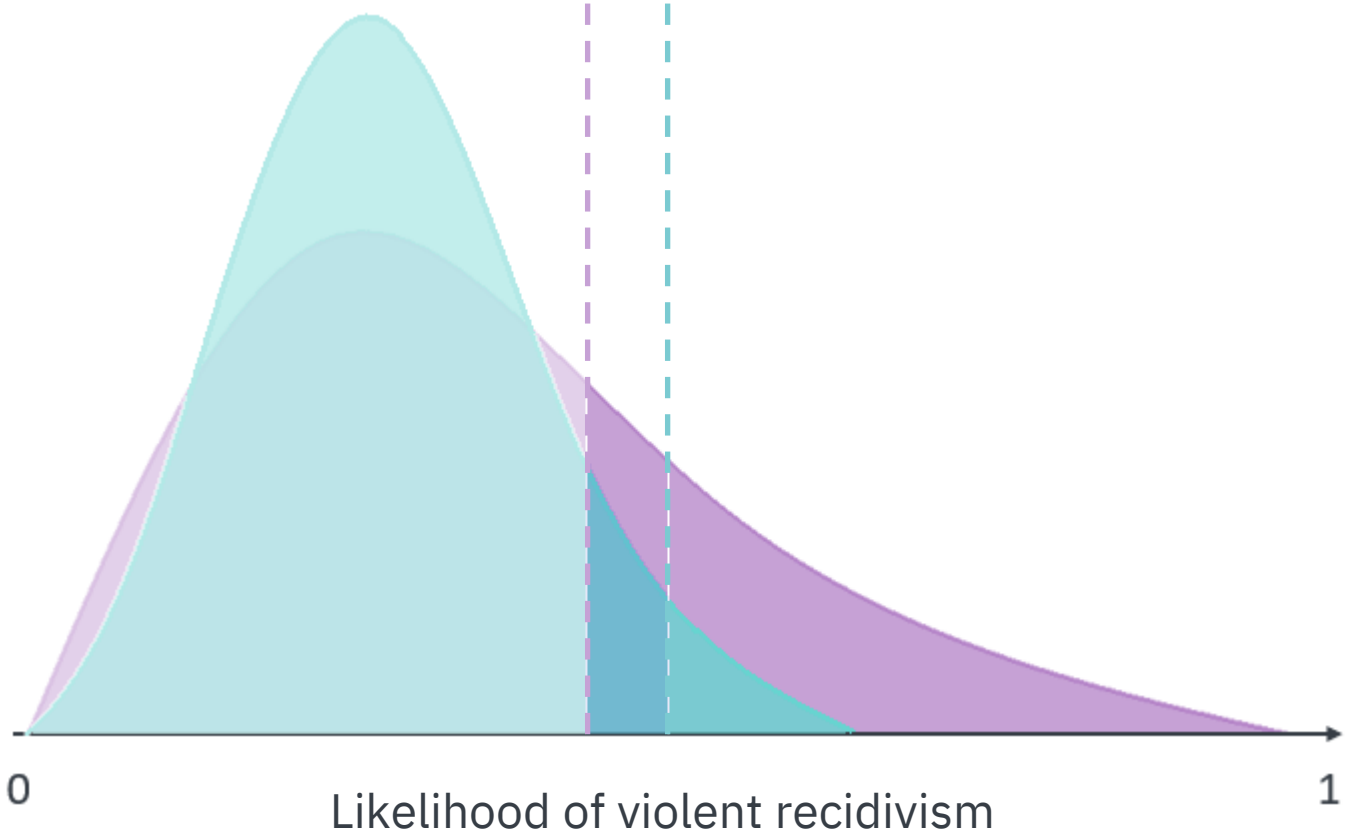
Taste-based discrimination

- Taste-based discrimination means one is willing to sacrifice utility to satisfy a preference to detain members of one group.

- If one group faces a lower threshold, we could detain fewer defendants from that group while also reducing overall detention and crime.

Reference to: Tutorial at the 19th ACM Conference on Economics and Computation (EC'18), Ithaca, NY, June 18, 2018

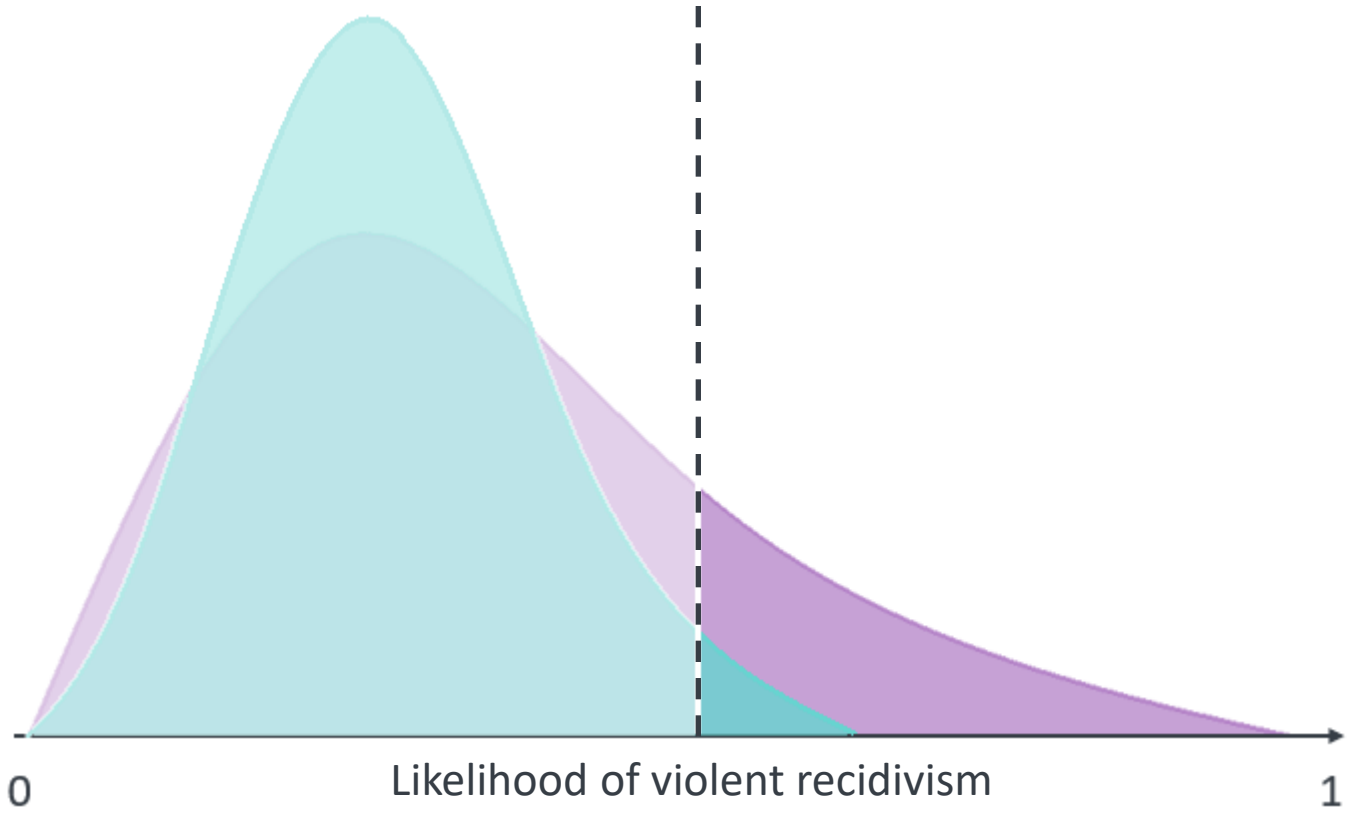
Taste-Based Discrimination



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

Reference to: Tutorial at the 19th ACM Conference on Economics and Computation (EC'18), Ithaca, NY, June 18, 2018

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.

Reference to: Tutorial at the 19th ACM Conference on Economics and Computation (EC'18), Ithaca, NY, June 18, 2018

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.]

Reference to: Tutorial at the 19th ACM Conference on Economics and Computation (EC'18), Ithaca, NY, June 18, 2018

Popular Mathematical Definitions of Fairness

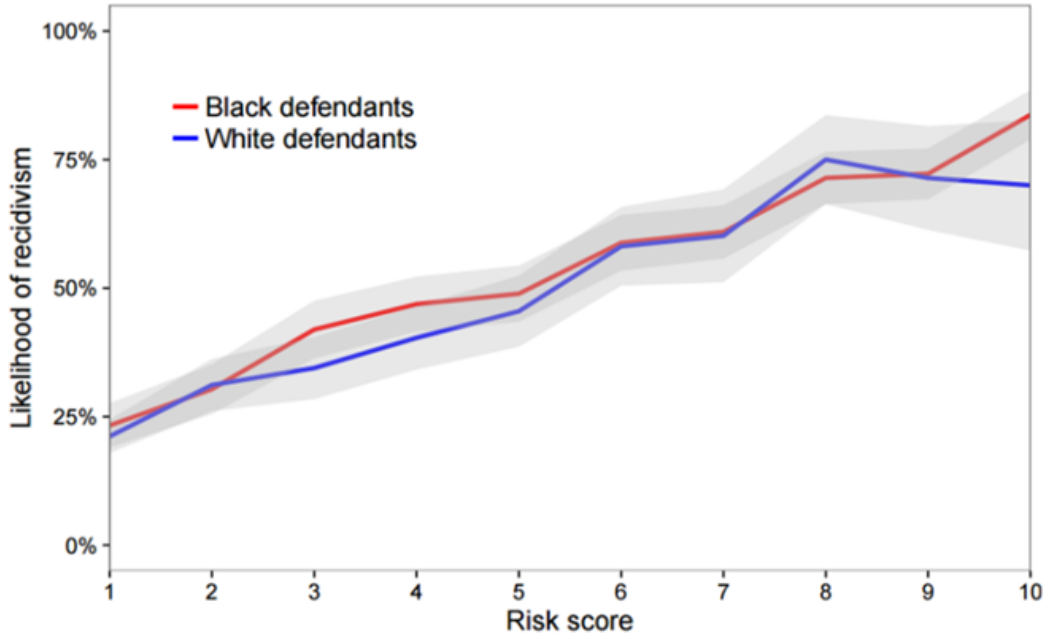
- All three definitions are problematic formalizations of long-standing legal and social norms.

  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

Reference to: Tutorial at the 19th ACM Conference on Economics and Computation (EC'18), Ithaca, NY, June 18, 2018

Calibration

- Conditional on risk score, groups should reoffend at equal rates.
- Calibration does not preclude taste-based discrimination.



Reference to: Tutorial at the 19th ACM Conference on Economics and Computation (EC'18), Ithaca, NY, June 18, 2018

Discrimination with Calibrated Scores

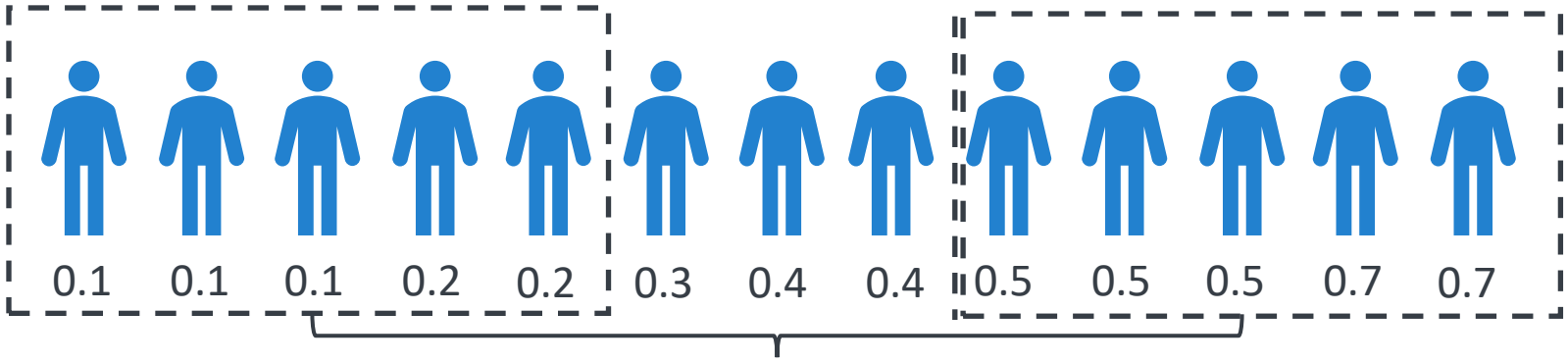
Detain defendants with $r_x > 0.5$

Probability of violent recidivism

Reference to: Tutorial at the 19th ACM Conference on Economics and Computation (EC'18), Ithaca, NY, June 18, 2018

A New Set of Calibrated Scores

Detain defendants with $r_x > 0.5$



Reference to: Tutorial at the 19th ACM Conference on Economics and Computation (EC'18), Ithaca, NY, June 18, 2018

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.

Reference to: Tutorial at the 19th ACM Conference on Economics and Computation (EC'18), Ithaca, NY, June 18, 2018

Ensuring Calibrated Scores don’t Discriminate

- Algorithm designers should train the best risk scores possible.
  [They should use all predictive features available.]

- 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.]

Reference to: Tutorial at the 19th ACM Conference on Economics and Computation (EC'18), Ithaca, NY, June 18, 2018

False Positive Rate Parity

- The false positive rates are equal for all groups.
- $$\text{False positive rate} = \frac{\text{Wouldn't have reoffened \& detained}}{\text{Wouldn't have reoffended}}$$
- ProPublica used this definition to allege bias in COMPAS.

Reference to: Tutorial at the 19th ACM Conference on Economics and Computation (EC'18), Ithaca, NY, June 18, 2018

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]

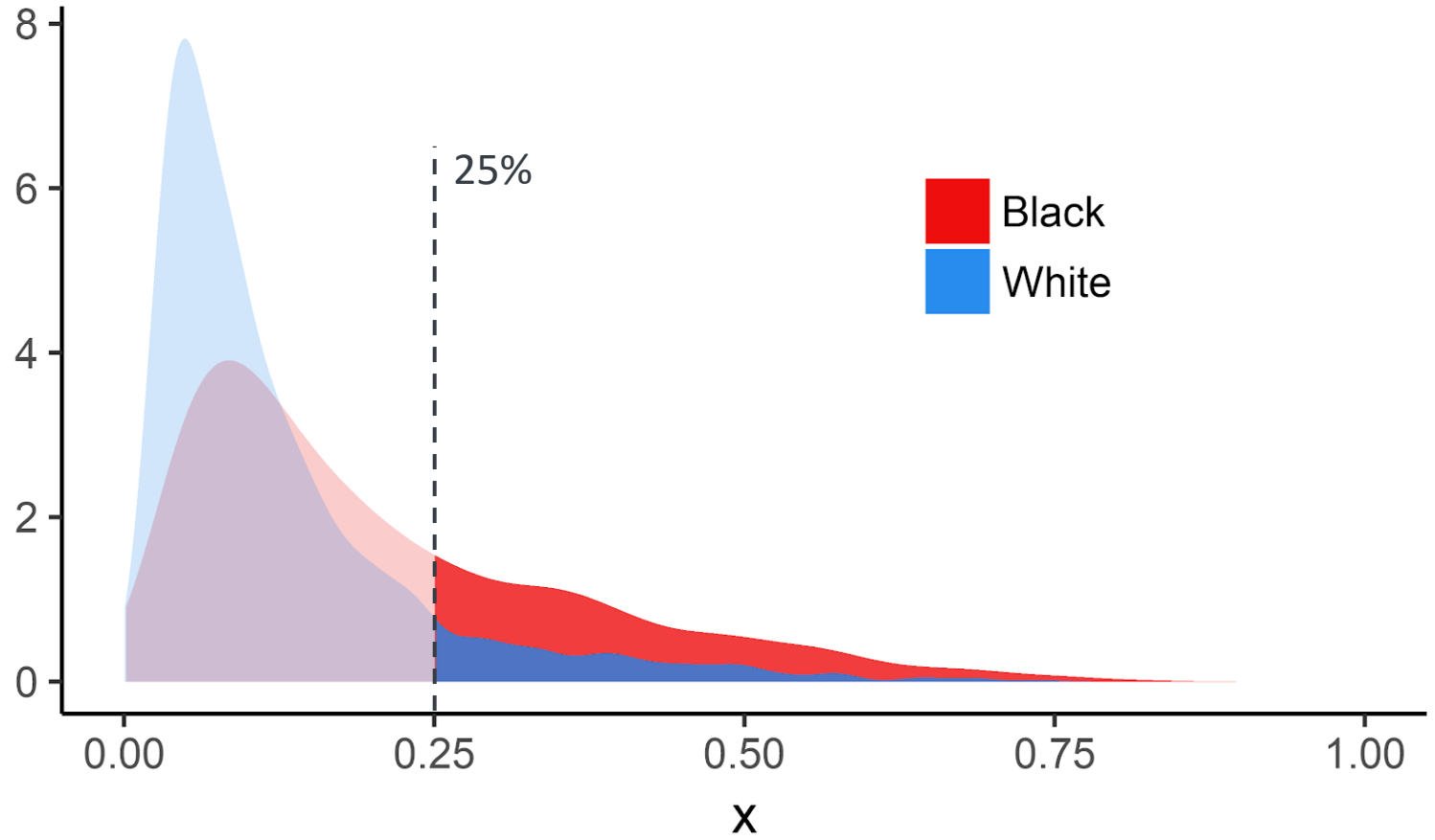
Reference to: Tutorial at the 19th ACM Conference on Economics and Computation (EC'18), Ithaca, NY, June 18, 2018

False Positive Rate Misconceptions

1. A higher false positive rate for some group implies discrimination (i.e., a lower bar for detaining that group)

Reference to: Tutorial at the 19th ACM Conference on Economics and Computation (EC'18), Ithaca, NY, June 18, 2018

Why do false positive rates differ?



Black and white defendants have different risk distributions.

Reference to: Tutorial at the 19th ACM Conference on Economics and Computation (EC'18), Ithaca, NY, June 18, 2018

Calculating False Positive Rates

Reference to: Tutorial at the 19th ACM Conference on Economics and Computation (EC'18), Ithaca, NY, June 18, 2018

Calculating False Positive Rates

$$\frac{\text{Did not reoffened \& detained}}{\text{Wouldn't have reoffended}} = \frac{\text{👥}}{\text{👥👥👥👥👥👥}} = 25\%$$

false positive rate

Reference to: Tutorial at the 19th ACM Conference on Economics and Computation (EC'18), Ithaca, NY, June 18, 2018

Calculating False Positive Rates

$$\frac{\text{Did not reoffened \& detained}}{\text{Wouldn't have reoffended}} = \frac{\text{👥}}{\text{👥👥👥👥👥👥}} = 40\%$$

false positive rate

Reference to: Tutorial at the 19th ACM Conference on Economics and Computation (EC'18), Ithaca, NY, June 18, 2018

Calculating False Positive Rates

25%
false positive rate

40%
false positive rate

Reference to: Tutorial at the 19th ACM Conference on Economics and Computation (EC'18), Ithaca, NY, June 18, 2018

Infra-Marginality Returns

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

- As we saw with the Raleigh example, infra-marginal statistics are misleading proxies for the threshold when risk distributions differ.

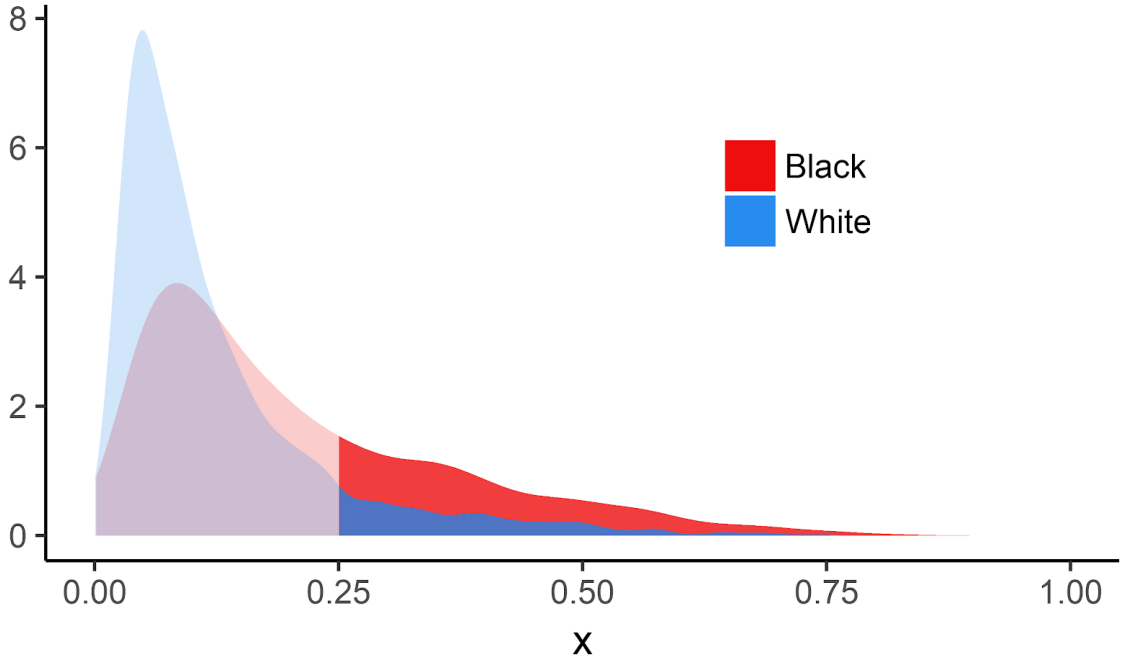
False Positive Rate Misconceptions

1. A higher false positive rate for some group implies discrimination (i.e., a lower bar for detaining that group)
2. A higher false positive rate for a minority group is due to a lack of data, either:
   a) a lack of training examples [rows]
   b) a lack of predictive features [columns]

Reference to: Tutorial at the 19th ACM Conference on Economics and Computation (EC'18), Ithaca, NY, June 18, 2018

A Lack of Training Data?

- If the base rates differ, the risk distributions will always differ, regardless of how many data points we have. And even if base rates are similar, the distributions may still differ.



Likelihood of violent recidivism

Reference to: Tutorial at the 19th ACM Conference on Economics and Computation (EC'18), Ithaca, NY, June 18, 2018

Not Enough Predictors?

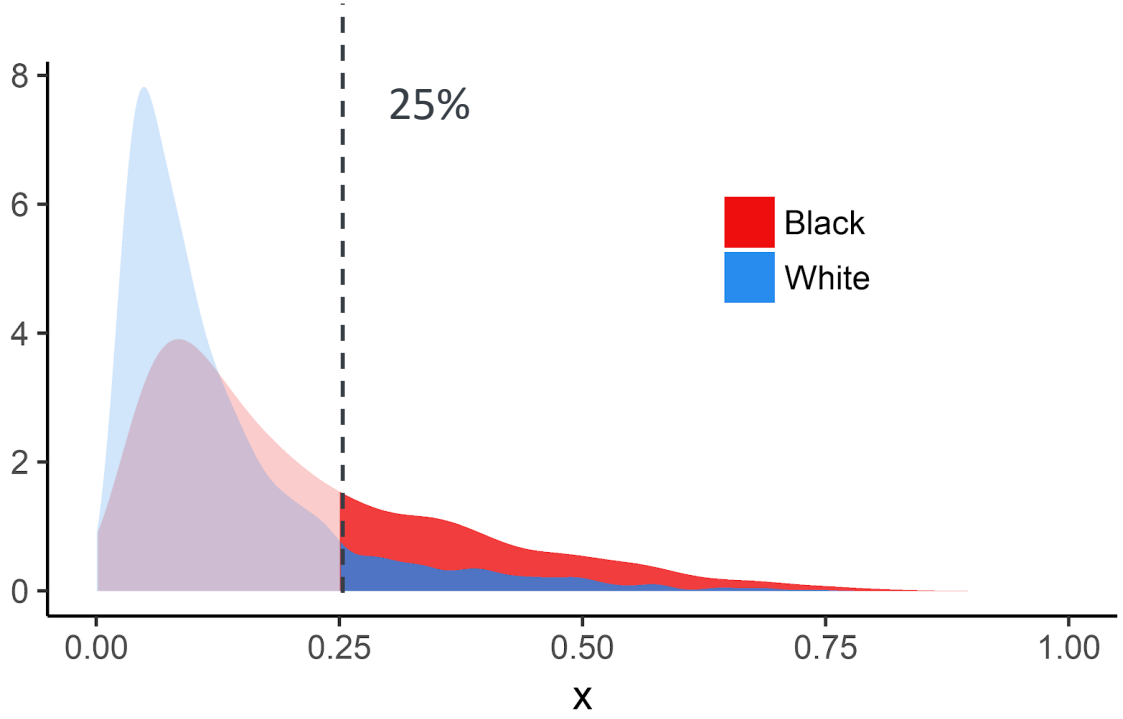
- If we acquire new predictive features the risk distribution shifts outwards, since we're better able to distinguish between recidivists and non-recidivists.

- This can actually increase the false positive rate, since it might result in more defendants lying above the threshold.

Reference to: Tutorial at the 19th ACM Conference on Economics and Computation (EC'18), Ithaca, NY, June 18, 2018

Equalizing FPRs by Ignoring Information

- The average black defendant is below the threshold



Black and white defendants have different risk distributions.

Reference to: Tutorial at the 19th ACM Conference on Economics and Computation (EC'18), Ithaca, NY, June 18, 2018

Equalizing FPRs by Ignoring Information

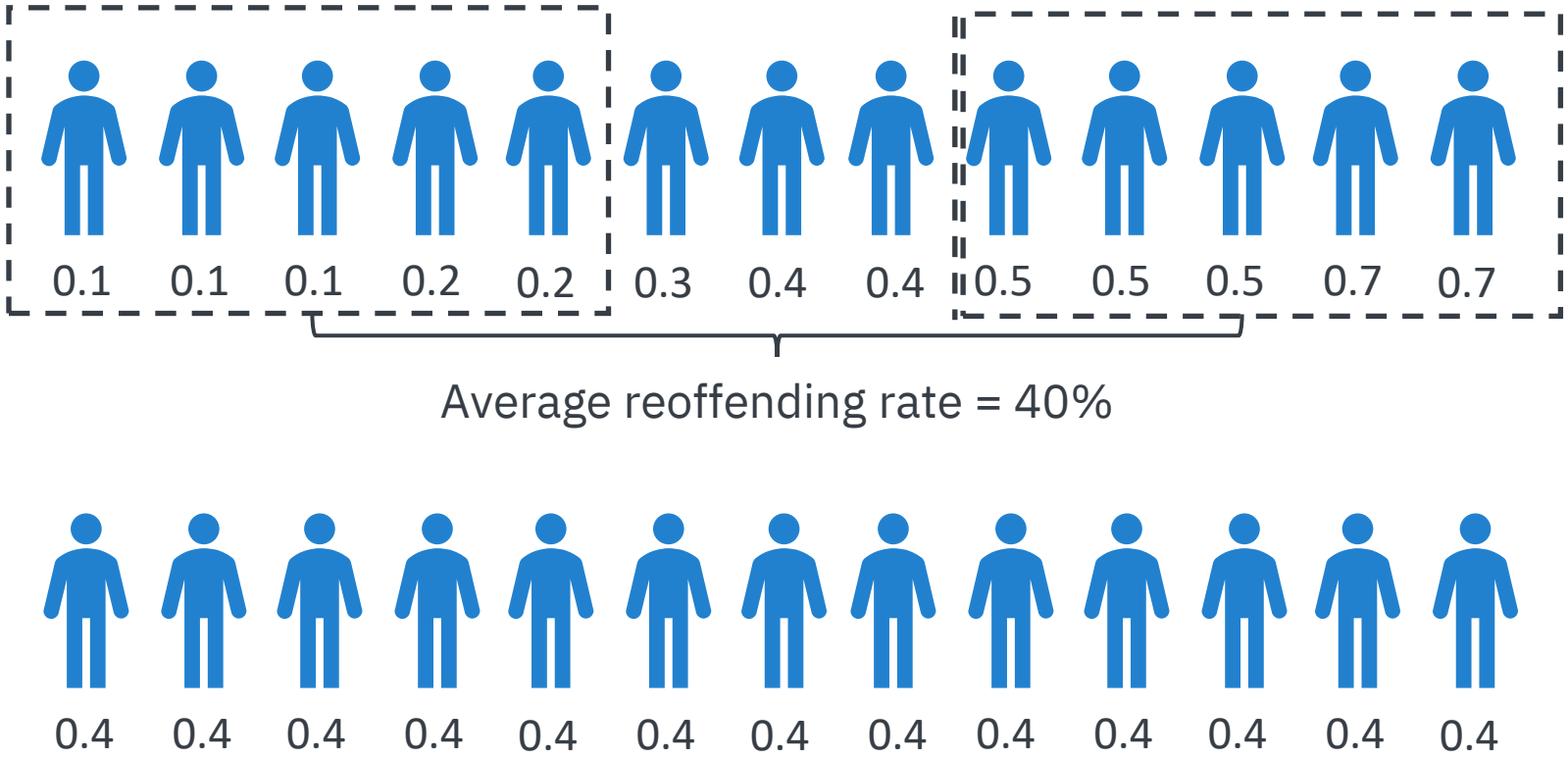
- The average black defendant is below the threshold.
- As the risk scores get worse we lose the ability to distinguish between high and low risk defendants.

  [Everyone starts to look like the average defendant]
- Thus, we can lower the black false positive rate by making the black risk scores worse.

Reference to: Tutorial at the 19th ACM Conference on Economics and Computation (EC'18), Ithaca, NY, June 18, 2018

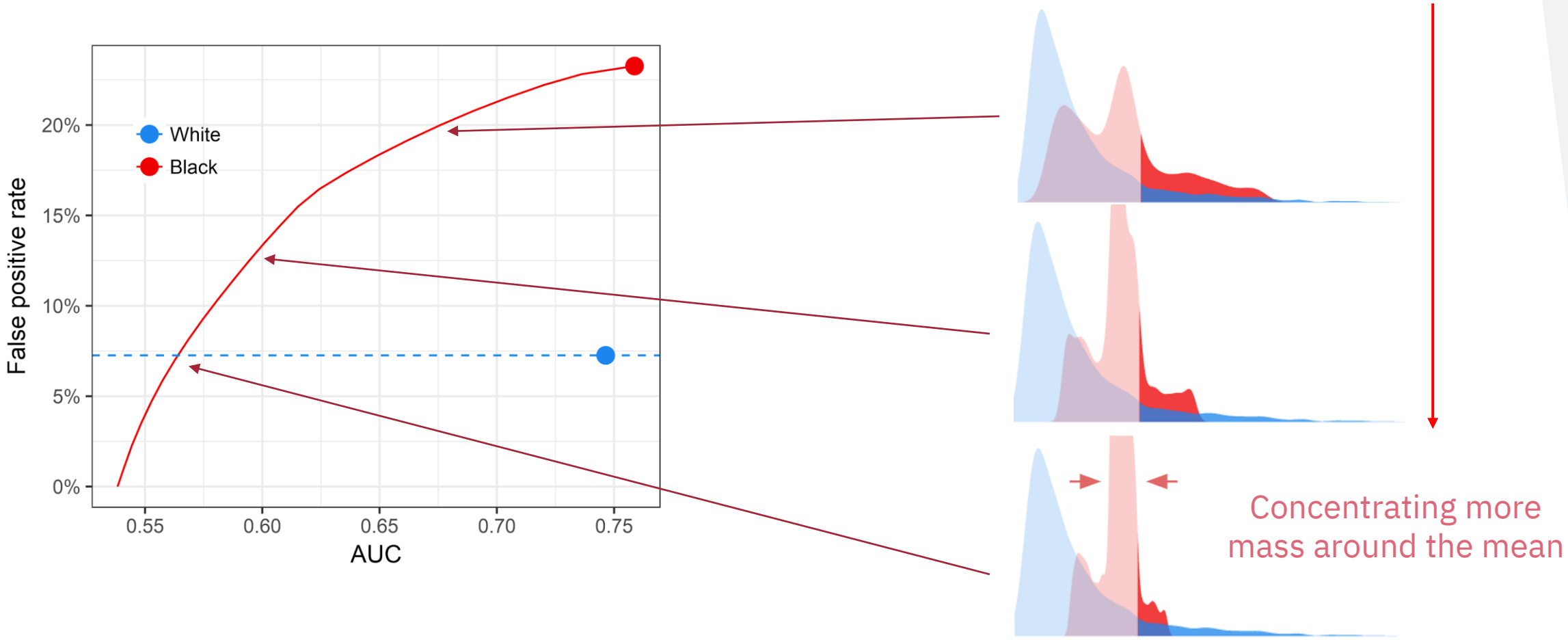
Equalizing FPRs by Ignoring Information

Detain defendants with $r_x > 0.5$



Reference to: Tutorial at the 19th ACM Conference on Economics and Computation (EC'18), Ithaca, NY, June 18, 2018

Equalizing FPRs by Ignoring Information



False positive rates are equalized when the black risk scores have almost no predictive validity (AUC = 0.56)

Reference to: Tutorial at the 19th ACM Conference on Economics and Computation (EC'18), Ithaca, NY, June 18, 2018

False Positive Rate Misconceptions

1. A higher false positive rate for some group implies discrimination (i.e., a lower bar for detaining that group)
2. A higher false positive rate for a minority group is due to a lack of data, either:
   a) a lack of training examples [rows]
   b) a lack of predictive features [columns]
3. False positive rates are a proxy for group well-being

Reference to: Tutorial at the 19th ACM Conference on Economics and Computation (EC'18), Ithaca, NY, June 18, 2018

The Problem with False Positive Rates

25%
false positive rate

40%
false positive rate

Reference to: Tutorial at the 19th ACM Conference on Economics and Computation (EC'18), Ithaca, NY, June 18, 2018

The Problem with False Positive Rates

$$\frac{\text{Did not reoffened \& detained}}{\text{Wouldn't have reoffended}} = \frac{\text{▲}}{\text{▲▲▲▲▲▲▲▲▲}} = 25\%$$

false positive rate

Reference to: Tutorial at the 19th ACM Conference on Economics and Computation (EC'18), Ithaca, NY, June 18, 2018

The Problem with False Positive Rates

25%
false positive rate

40%
false positive rate

Reference to: Tutorial at the 19th ACM Conference on Economics and Computation (EC'18), Ithaca, NY, June 18, 2018

The Problem with False Positive Rates

25%
false positive rate

25%
false positive rate

Reference to: Tutorial at the 19th ACM Conference on Economics and Computation (EC'18), Ithaca, NY, June 18, 2018

Fairness using Confusion Matrices

	Non-recidivist	Recidivist
Released	*TN*	*FN*
Detained	*FP*	*TP*

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

Reference to: Tutorial at the 19th ACM Conference on Economics and Computation (EC'18), Ithaca, NY, June 18, 2018

Fairness using Confusion Matrices

	Non-recidivist	Recidivist
Released	*TN*	*FN*
Detained	*FP*	*TP*

All these definitions compare infra-marginal statistics, so they have the same problems as false positive rates.

They are all unreliable measures of taste-based discrimination.

Reference to: Tutorial at the 19th ACM Conference on Economics and Computation (EC'18), Ithaca, NY, June 18, 2018

Are the data biased?

- Two types of bias:

  1. Biased predictors

     [Features that are differentially predictive]

  2. Biased labels

     [***Y*** doesn't perfectly measure what we care about]

Reference to: Tutorial at the 19th ACM Conference on Economics and Computation (EC'18), Ithaca, NY, June 18, 2018

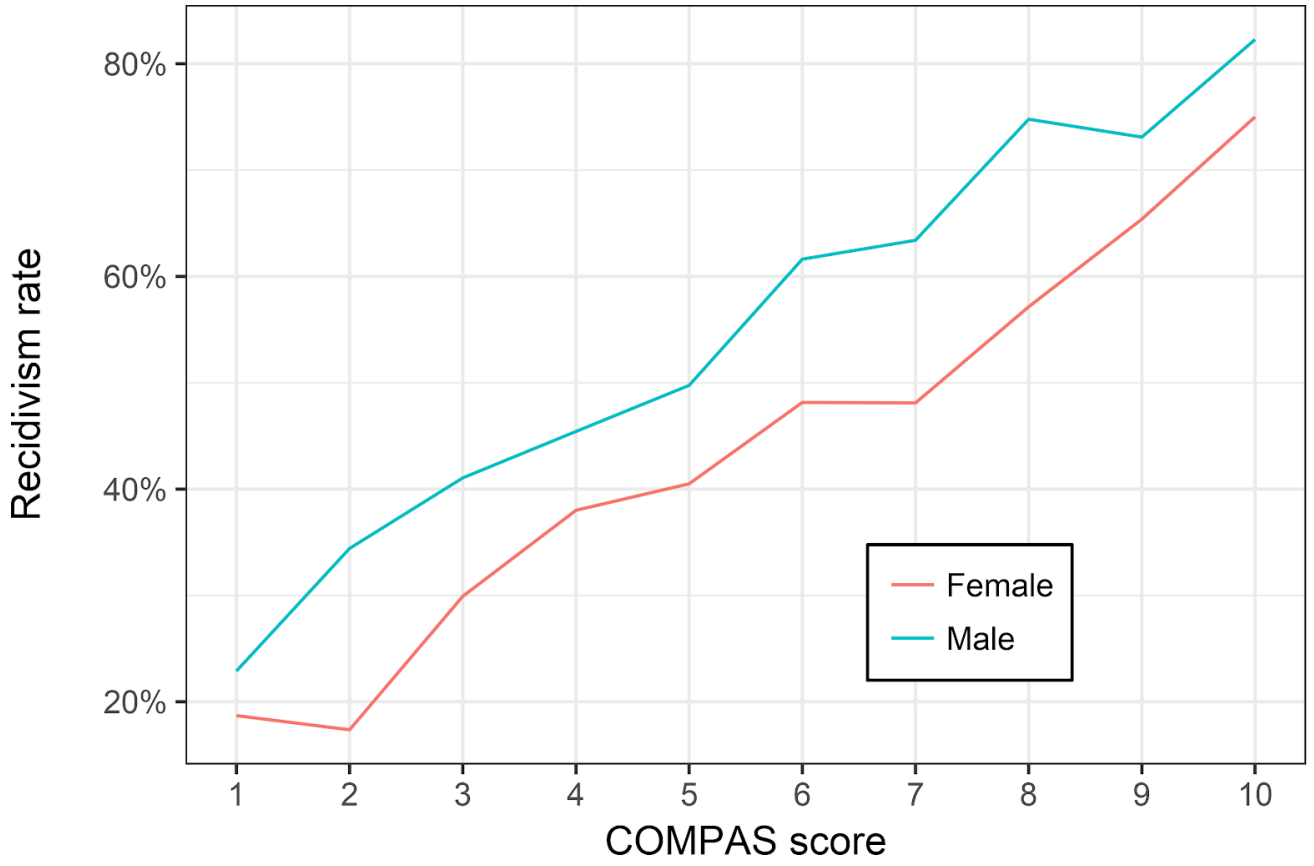
Biased predictors

- 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.]

Reference to: Tutorial at the 19th ACM Conference on Economics and Computation (EC'18), Ithaca, NY, June 18, 2018

Gender bis in Broward County



Reference to: Tutorial at the 19th ACM Conference on Economics and Computation (EC'18), Ithaca, NY, June 18, 2018

The problem with anti-classification

- Gender-neutral 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.]

Reference to: Tutorial at the 19th ACM Conference on Economics and Computation (EC'18), Ithaca, NY, June 18, 2018

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.

Reference to: Tutorial at the 19th ACM Conference on Economics and Computation (EC'18), Ithaca, NY, June 18, 2018

STEVENS
INSTITUTE OF TECHNOLOGY
1870

Coda

Mathematical definitions

- There are many formal, mathematical definitions of fairness, most of which cannot be simultaneously satisfied.

Reference to: Tutorial at the 19th ACM Conference on Economics and Computation (EC'18), Ithaca, NY, June 18, 2018

Math ≠ equity

- There are many formal, mathematical definitions of fairness, most of which cannot be simultaneously satisfied.
- Nearly none of these definitions map to established legal or social understandings of equity.

Reference to: Tutorial at the 19th ACM Conference on Economics and Computation (EC'18), Ithaca, NY, June 18, 2018

Math ≠ equity

Three important consequences

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

2. Attempts to equalize these measures can itself lead to discriminatory or otherwise perverse decisions.

3. The idea that there are trade-offs between different measures is largely illusory.

Reference to: Tutorial at the 19th ACM Conference on Economics and Computation (EC'18), Ithaca, NY, June 18, 2018

Threshold rules

- In many decision-making settings, applying a single threshold rule to risk scores is both efficient and equitable.
  [We assume lots of data + accurate outcome measures.]

Reference to: Tutorial at the 19th ACM Conference on Economics and Computation (EC'18), Ithaca, NY, June 18, 2018

Important caveats

1. We have focused on the immediate costs and benefits of decisions. Long-term equilibrium effects might justify multiple thresholds.

2. Multiple thresholds might also be justified by different individual-level costs. [Detaining a single parent more costly than detaining a defendant without children.]

3. Some decisions [e.g., college admissions] apply to groups rather than individuals. [There are externalities.]

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.

Reference to: Tutorial at the 19th ACM Conference on Economics and Computation (EC'18), Ithaca, NY, June 18, 2018

STEVENS
INSTITUTE OF TECHNOLOGY
1870

THANK **YOU**

Stevens Institute of Technology
1 Castle Point Terrace, Hoboken, NJ 07030