Designing Equitable Risk Models for Lending and Beyond

Sharad Goel
Stanford Computational Policy Lab

Summary

Part I. Many common mathematical definitions of algorithmic fairness are at odd with important understandings of equity.

Summary

Part I. Many common mathematical definitions of algorithmic fairness are at odd with important understandings of equity.

Part II. We can often design more equitable systems by explicitly separating prediction from decision making.

Part I

Assessing bias in risk models

Are risk models *fair*?

Statistical models of risk are now used by experts in finance, medicine, criminal justice, and beyond to guide high-stakes decisions.

Are risk models *fair*?

Statistical models of risk are now used by experts in finance, medicine, **criminal justice,** and beyond to guide high-stakes decisions.

Pretrial release decisions

“Release on recognizance” or set bail

Shortly after arrest, judges must decide whether to release or detain defendants while they await trial.

Goal is to balance flight risk and public safety against the financial and social burdens of bail.

Risk assessment tools

In jurisdictions across the United States, judges are now incorporating the results of risk assessment tools when making pretrial decisions.

These statistical tools typically assess the likelihood a defendant will **fail to appear** at trial or **commit future crimes**.
[We call this the defendant's *risk* of FTA or criminal activity.]

Algorithmic risk assessment

An example: the Public Safety Assessment (PSA)

Failure to Appear (FTA)	
Risk Factor	**Points**
Pending charge at the time of offense	No = 0 Yes = 1
Prior conviction (misdemeanor or felony)	No = 0 Yes = 1
Prior failure to appear in past 2 years	0 = 0 1 = 2 2 or more = 4
Prior failure to appear older than 2 years	No = 0 Yes = 1

Algorithmic risk assessment

An example: the Public Safety Assessment (PSA)

A hypothetical defendant:

- No pending charges
- 2 prior convictions
- 2 prior FTA's in last 2 years
- No prior FTA's before that

Failure to Appear (FTA)	
Risk Factor	**Points**
Pending charge at the time of offense	No = 0 Yes = 1
Prior conviction (misdemeanor or felony)	No = 0 Yes = 1
Prior failure to appear in past 2 years	0 = 0 1 = 2 2 or more = 4
Prior failure to appear older than 2 years	No = 0 Yes = 1

Algorithmic risk assessment

An example: the Public Safety Assessment (PSA)

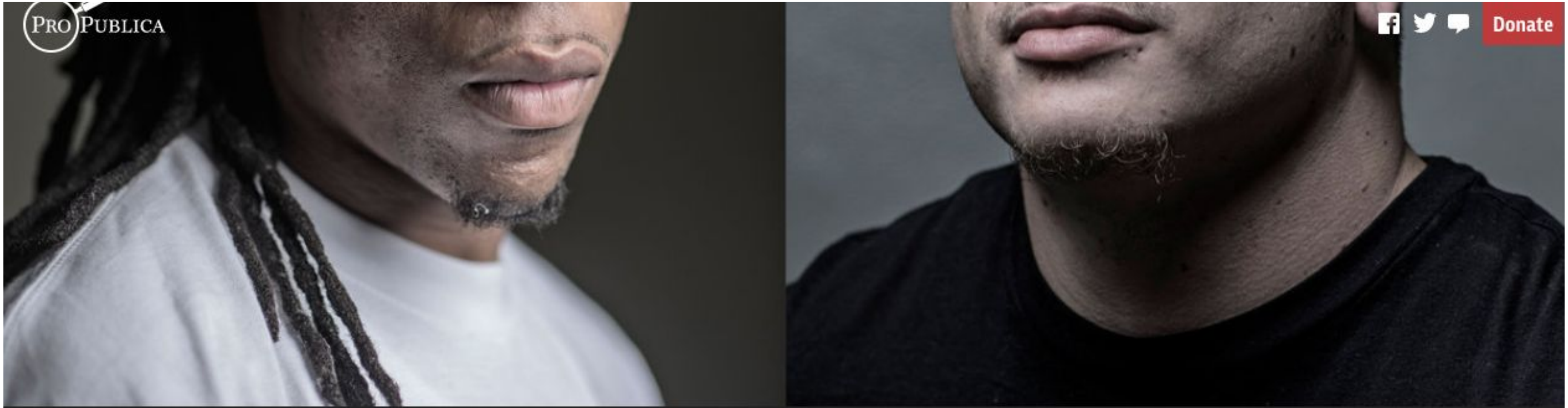
A hypothetical defendant:

- No pending charges
- 2 prior convictions
- 2 prior FTA's in last 2 years
- No prior FTA's before that

5/7

"High risk"

Failure to Appear (FTA)	
Risk Factor	**Points**
Pending charge at the time of offense	No = 0 Yes = 1
Prior conviction (misdemeanor or felony)	No = 0 Yes = 1
Prior failure to appear in past 2 years	0 = 0 1 = 2 2 or more = 4
Prior failure to appear older than 2 years	No = 0 Yes = 1



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

A critique of *fair* machine learning

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.

Corbett-Davies & Goel, Science Advances [R&R]
Corbett-Davies et al., KDD [2017]

A mathematical definition of fairness

Classification parity

An algorithm is considered to be *fair* if error rates are [approximately] equal for white and Black defendants.

A mathematical definition of fairness

Proposed legislation in Idaho [2019]

"Pretrial risk assessment algorithms shall not be used … by the state until first shown to be **free of bias**, ...[meaning] that an algorithm has been formally tested and...the **rate of error is balanced** as between protected classes and those not in protected classes."

[This requirement was removed from the final bill.]

A mathematical definition of fairness

False positive rate

A common mathematical definition of fairness is demanding equal false positive rates [used by ProPublica].

$$\text{False positive rate} = \frac{\text{Did not reoffend \& ``high risk''}}{\text{Did not reoffend}}$$

Error rate disparities in Broward County

31% vs. **15%**

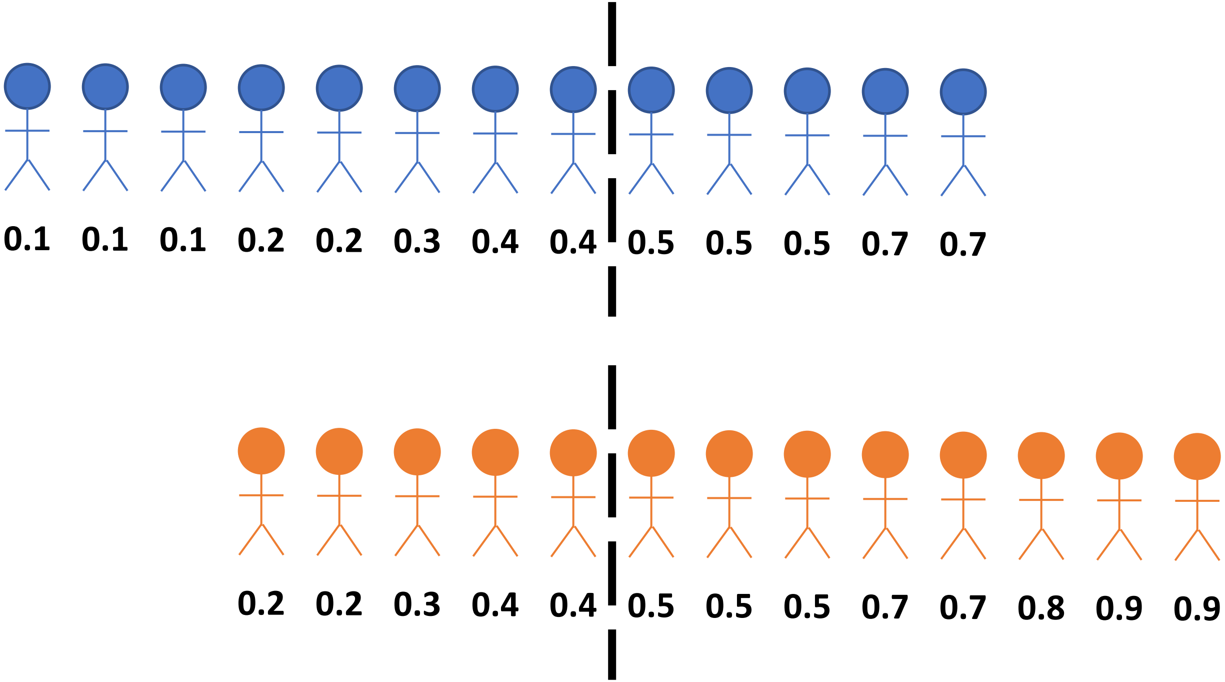
of Black defendants
who did not reoffend

of white defendants
who did not reoffend

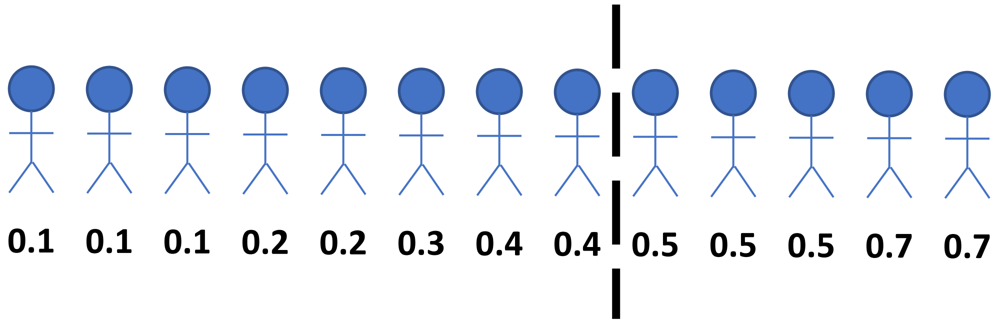
were deemed **high risk** of committing a violent crime

[Higher false positive rates for black defendants]

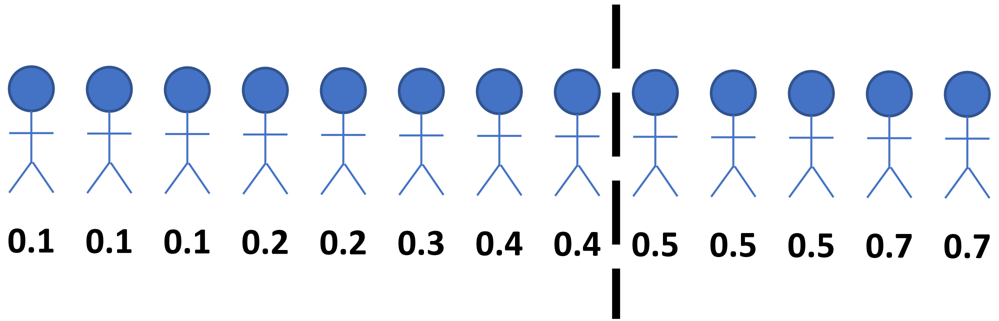
False positive rates



False positive rates



False positive rates

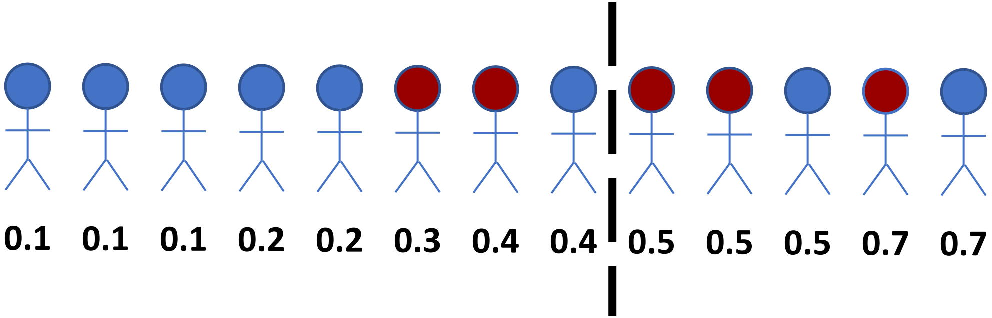


Did not reoffend & “high risk”

Did not reoffend

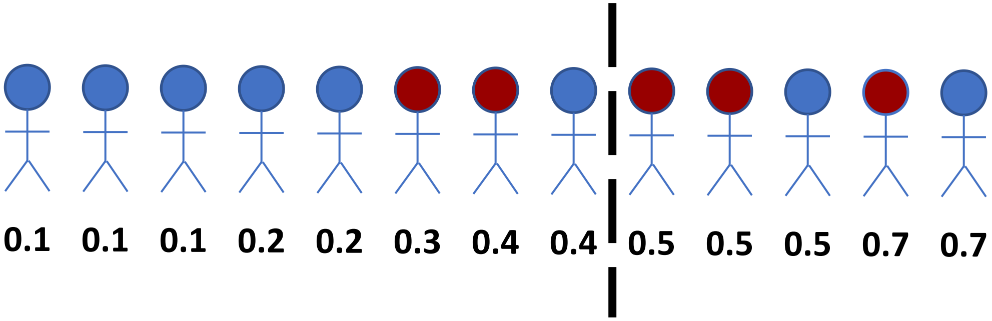
=

False positive rates

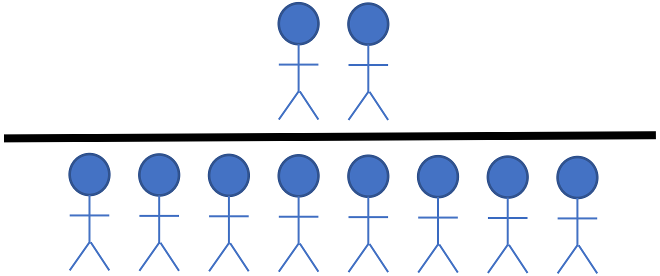


$$\frac{\text{Did not reoffend } \& \text{ “high risk”}}{\text{Did not reoffend}} =$$

False positive rates



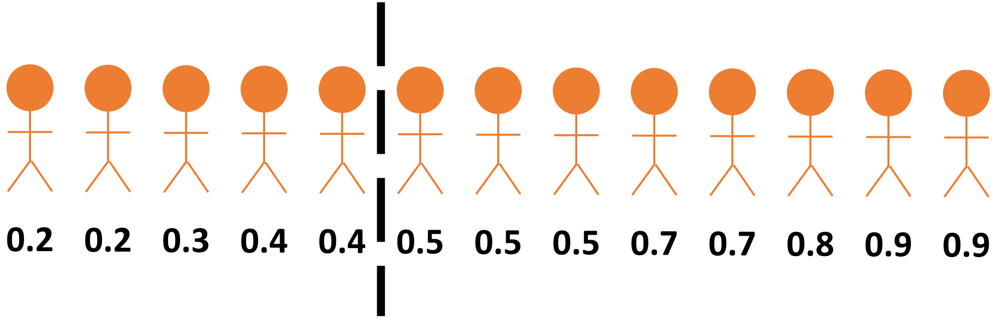
$$\frac{\text{Did not reoffend \& “high risk”}}{\text{Did not reoffend}} =$$



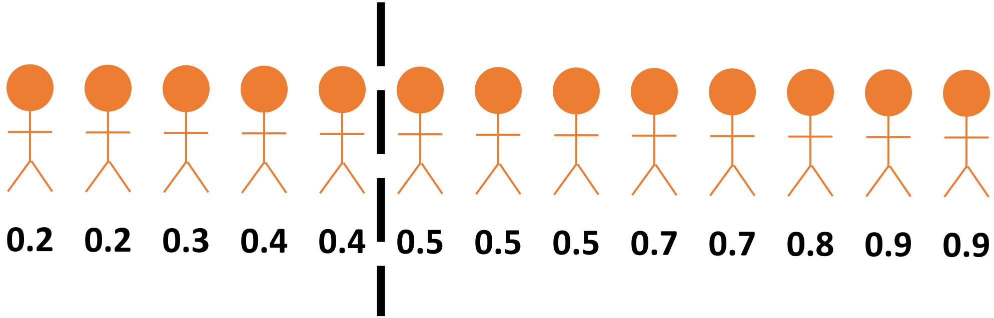
= **25%**

false positive rate

False positive rates



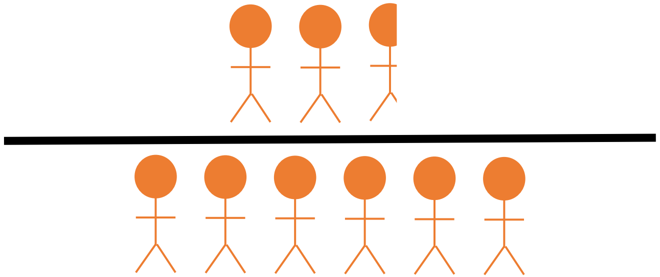
False positive rates



Did not reoffend & “high risk”

Did not reoffend

=

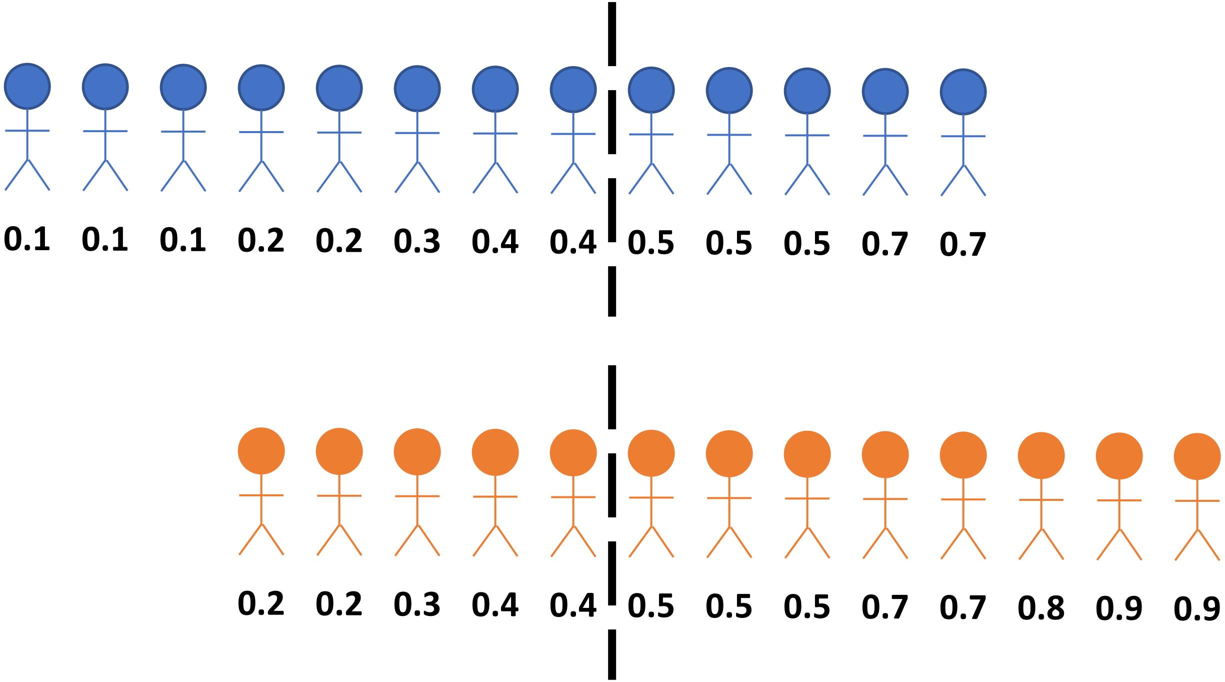


=

42%

false positive rate

False positive rates



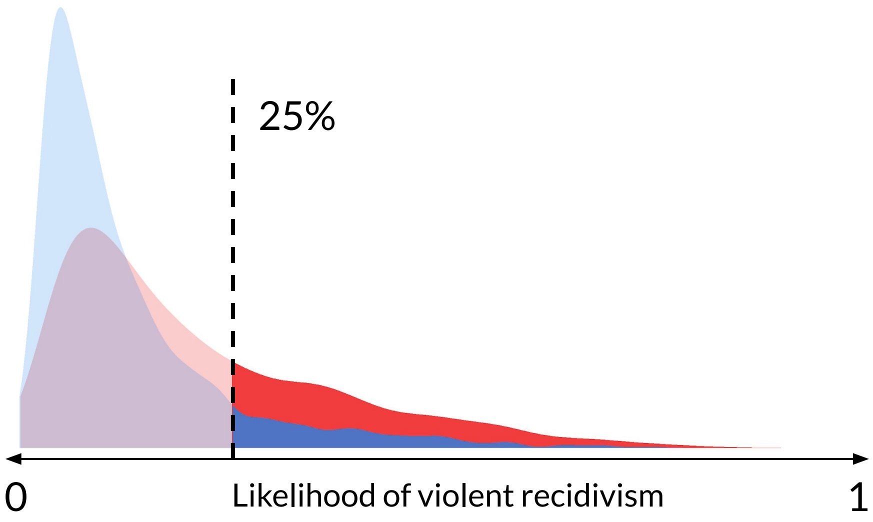
25%
false positive rate

42%
false positive rate

The problem of 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.

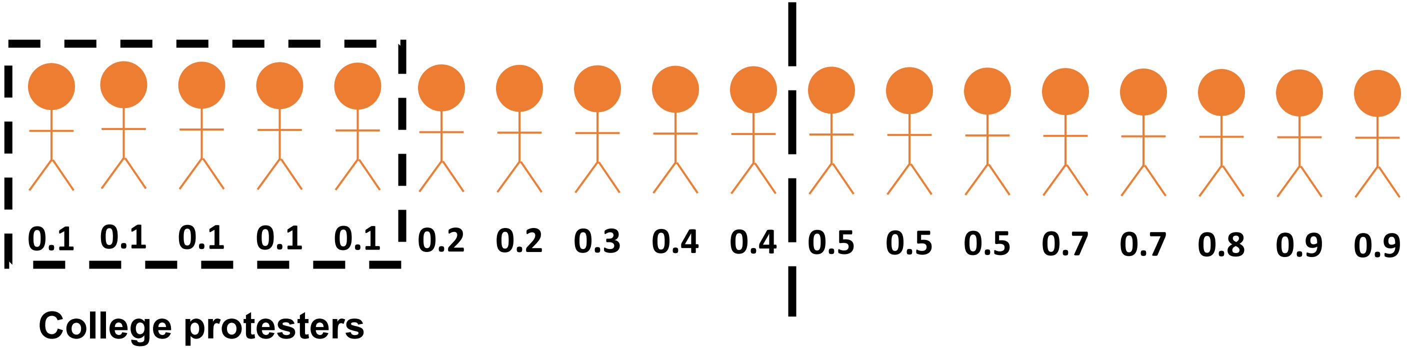
Broward County risk distributions



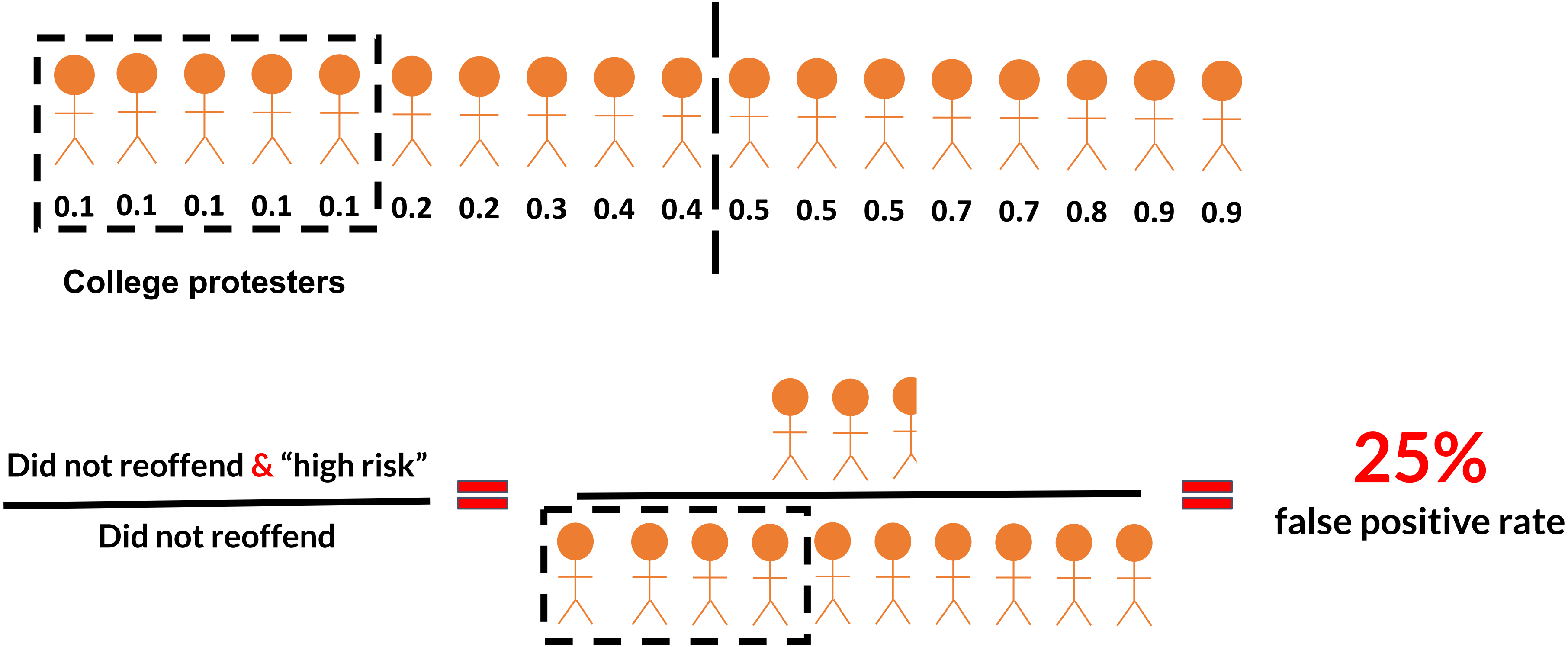
Black and white defendants have different risk distributions

The problem with false positive rates

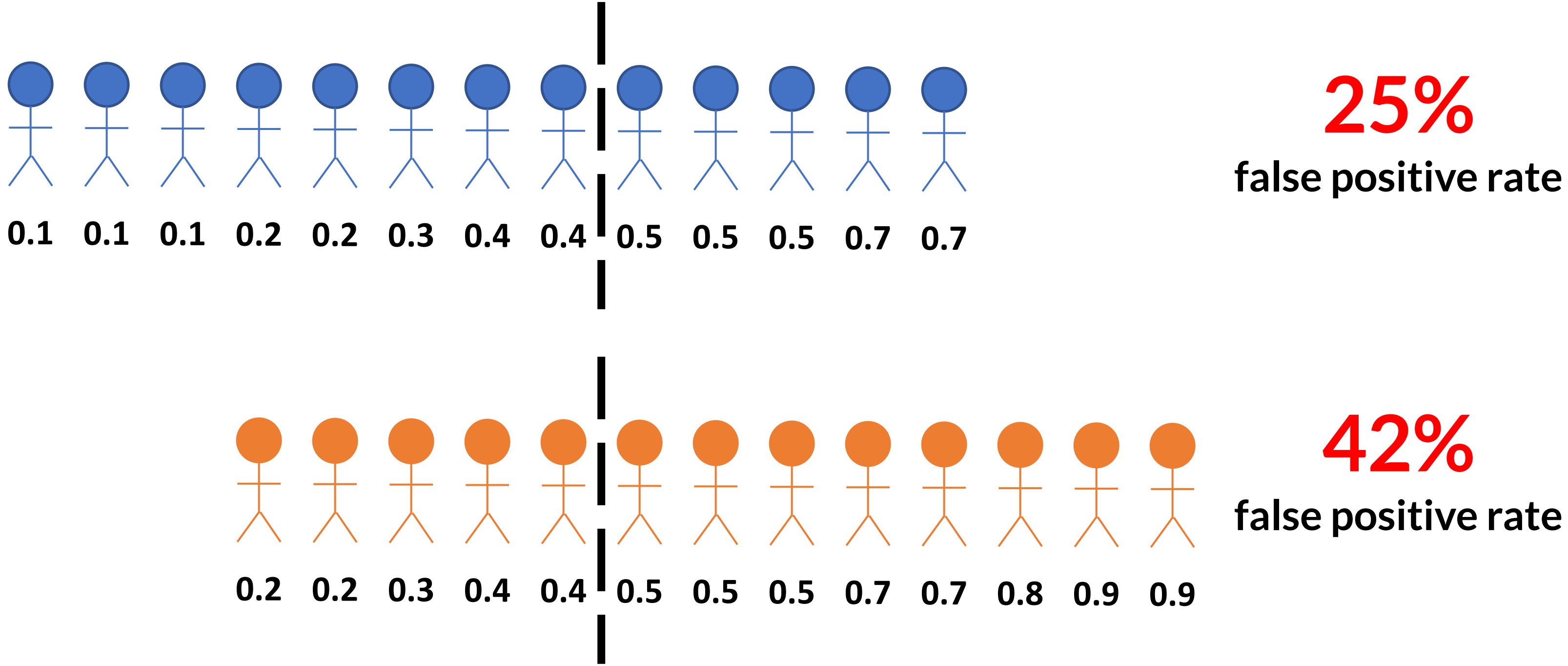
The problem with false positive rates



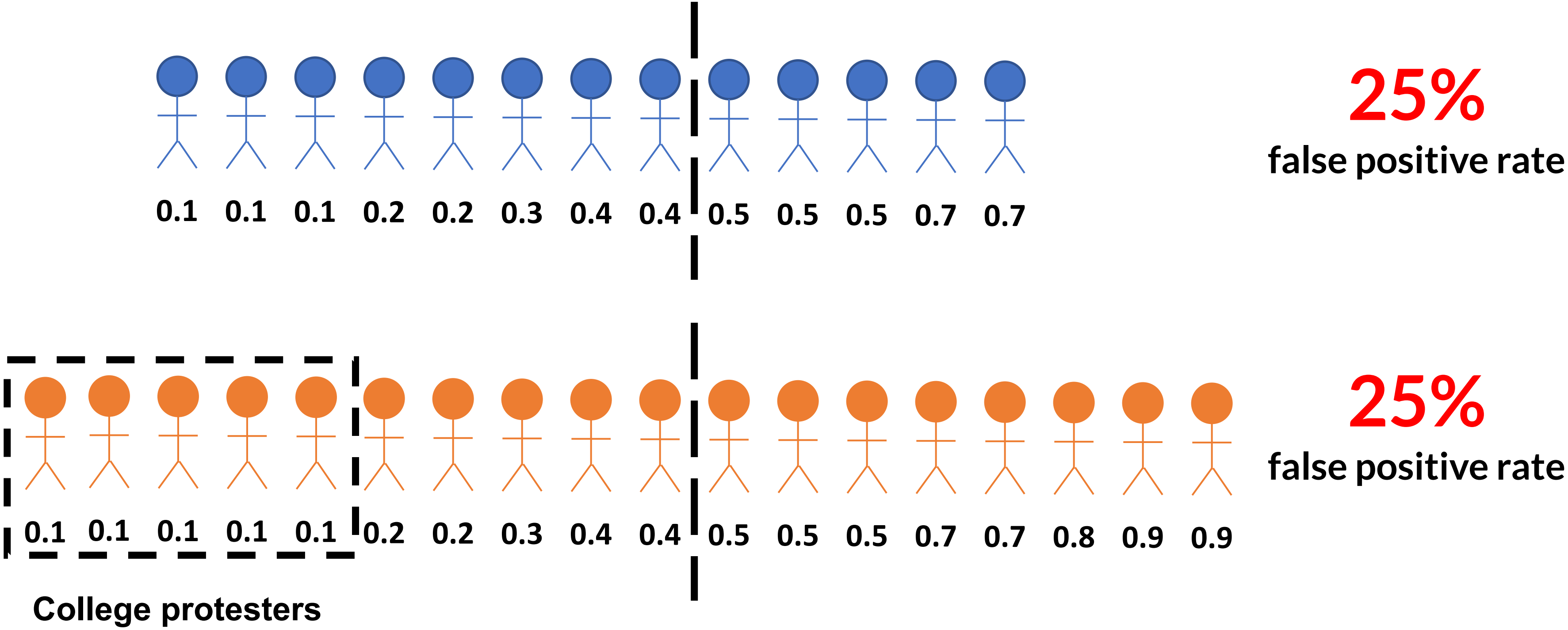
The problem with false positive rates



The problem with false positive rates



The problem with false positive rates



Anti-classification

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

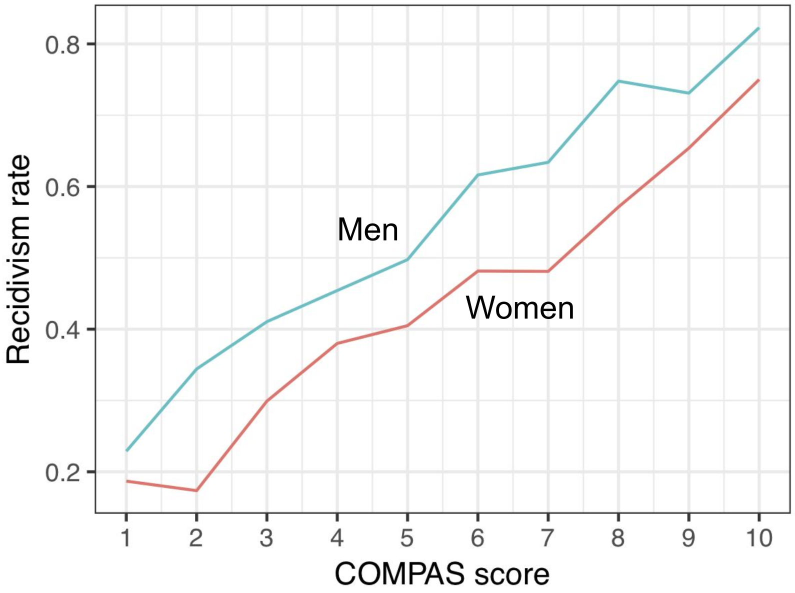
But discrimination is still possible using “blind” policies.
[e.g., redlining in financial services]

The problem with anti-classification

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

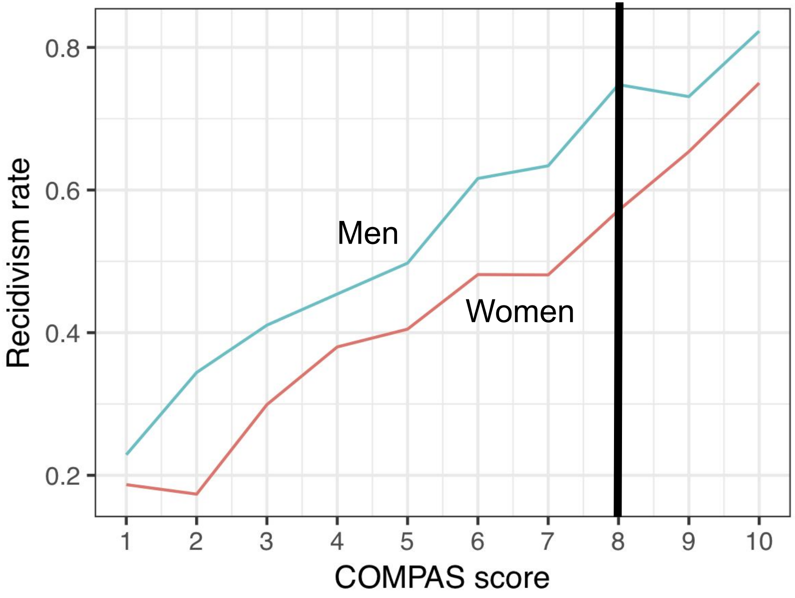
A gender-blind risk score

Broward County, Florida



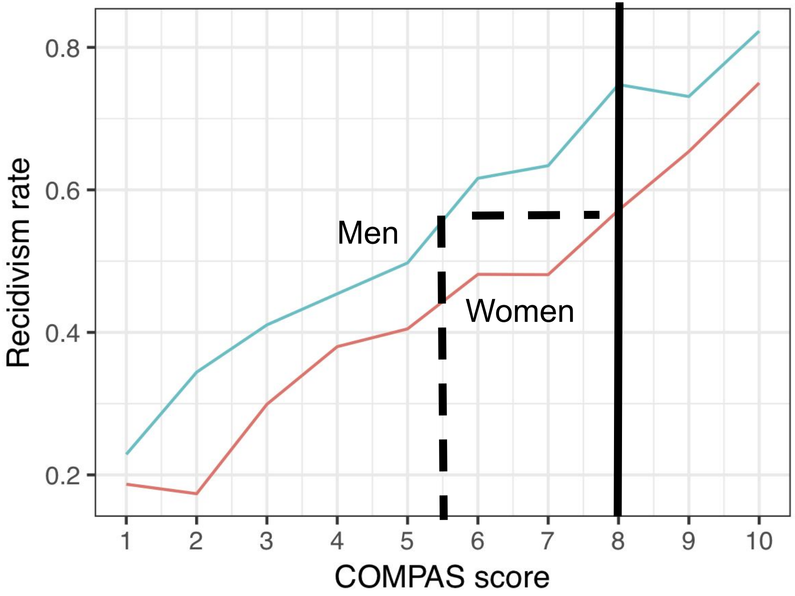
A gender-blind risk score

Broward County, Florida



A gender-blind risk score

Broward County, Florida



The problem with anti-classification

Gender-neutral risk models can lead to 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.]

Are the data *biased*?

Biased labels
[Measurement error]

Algorithm estimates the probability a defendant will be *observed / reported* committing a future violent crime.

Since reported crime is only a proxy for actual crime, estimates might be biased.

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.]

Part II

Designing equitable algorithmic policies

Algorithms ≠ policy

Separate risk estimation from policy decisions.

Statistical algorithms are often good at synthesizing information to estimate risk. But we must still set equitable policy.

In the case of pretrial decisions, we might limit money bail and/or consider non-custodial interventions. In the financial sector, we might offer support services to change one's risk profile.

Inequities in lending

Motivation

20% of U.S. households have no mainstream credit

[Not eligible for small-dollar loans]

Inequities in lending
Motivation

20% of U.S. households have no mainstream credit

[Not eligible for small-dollar loans]

“About three in four ... households with no mainstream credit stayed current on bills in the past 12 months”

[Apaam et al. 2017]

Inequities in lending

Motivation

20% of U.S. households have no mainstream credit

[Not eligible for small-dollar loans]

“About three in four ... households with no mainstream credit stayed current on bills in the past 12 months”

[Apaam et al. 2017]

These households are disproportionately Black & Hispanic.

How can we design a more inclusive lending policy?

Inequities in lending

The challenge

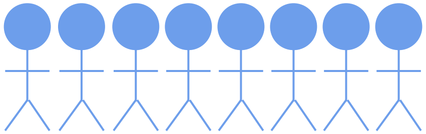
We want to:

- Allocate resources to underserved groups
  [Individuals without mainstream credit]

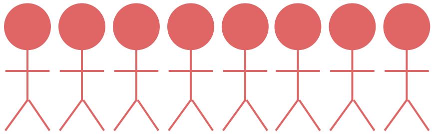
- while remaining relatively efficient.
  [Giving loans to those who are most likely to repay]

Equity in loans

Illustrative example



Unbanked



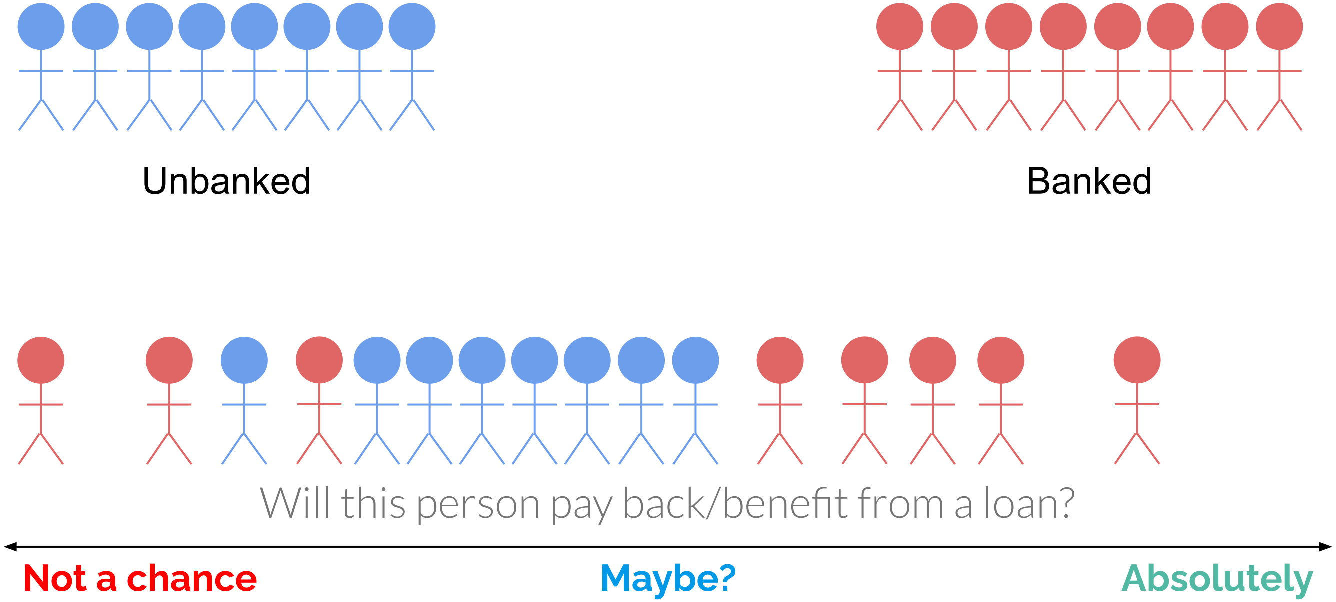
Banked

Will this person pay back/benefit from a loan?

Not a chance | **Maybe?** | **Absolutely**

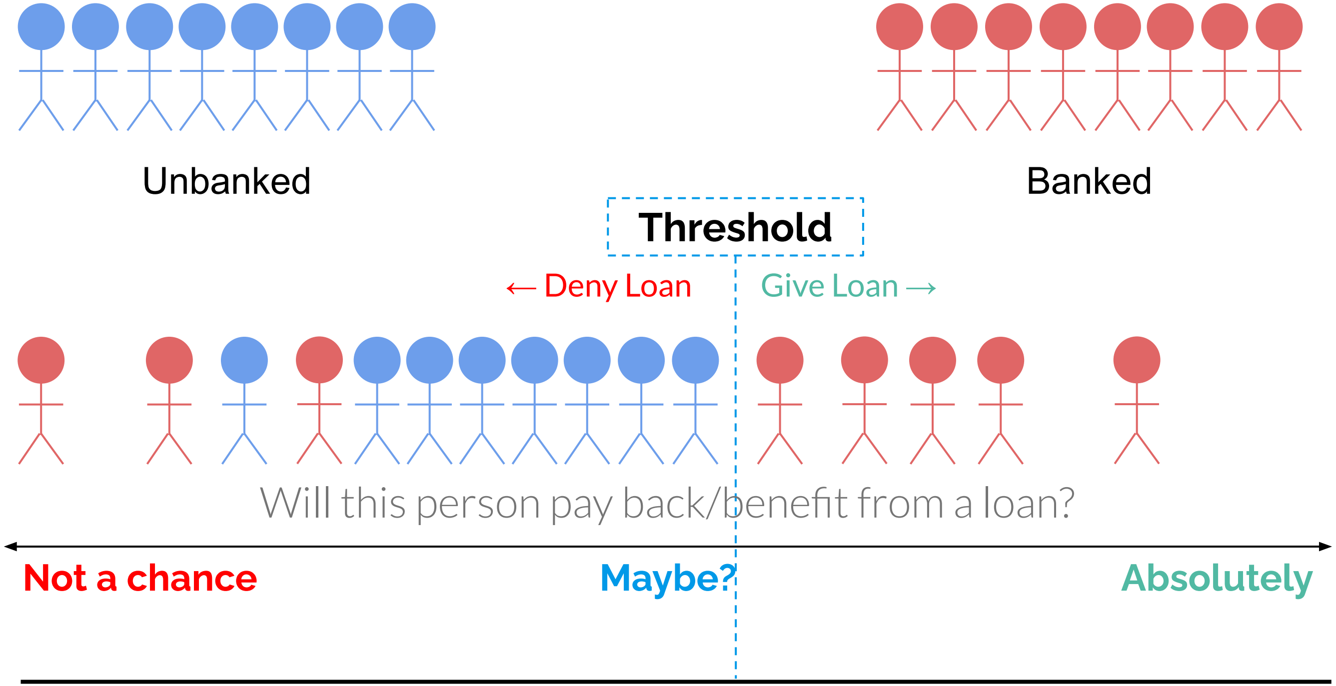
Equity in loans

Illustrative example

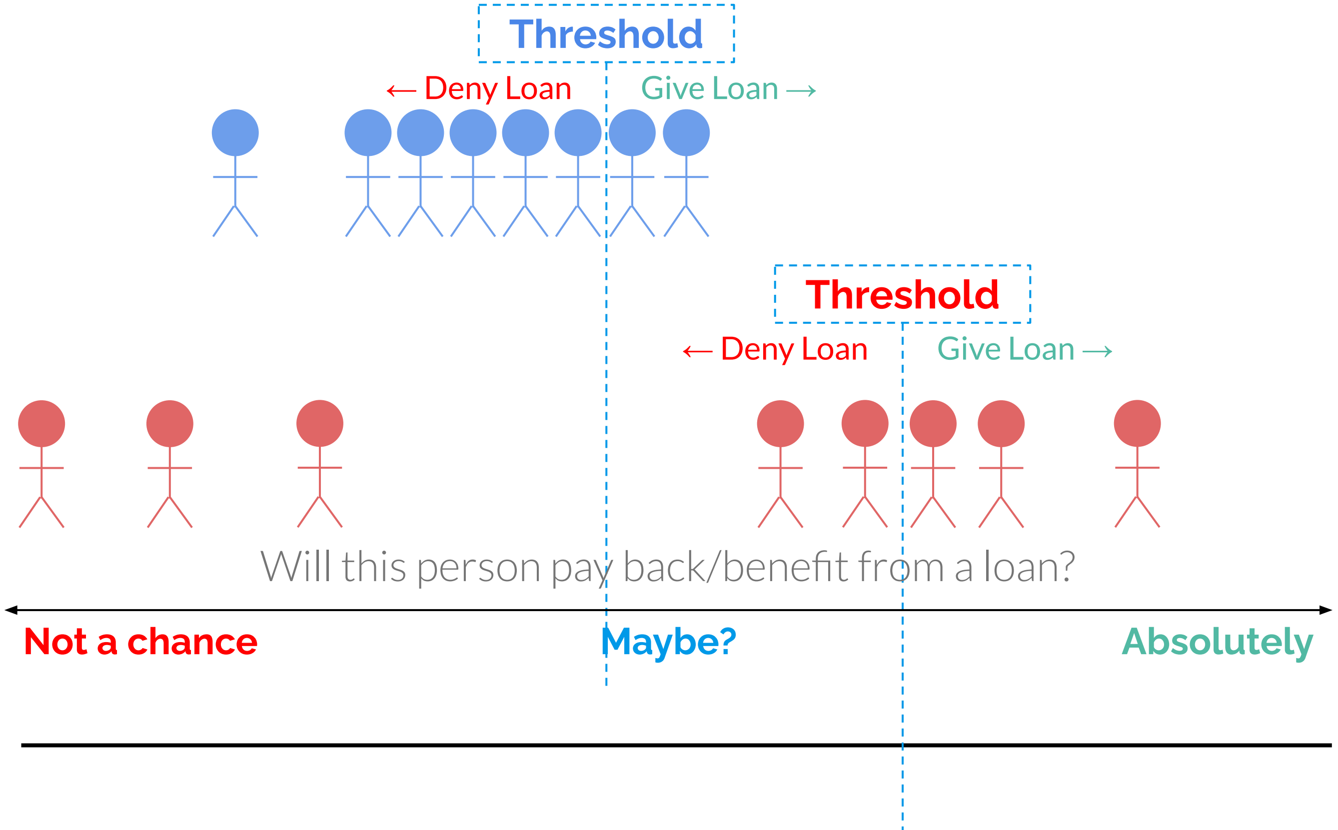


Equity in loans

Illustrative example



Equity in loans

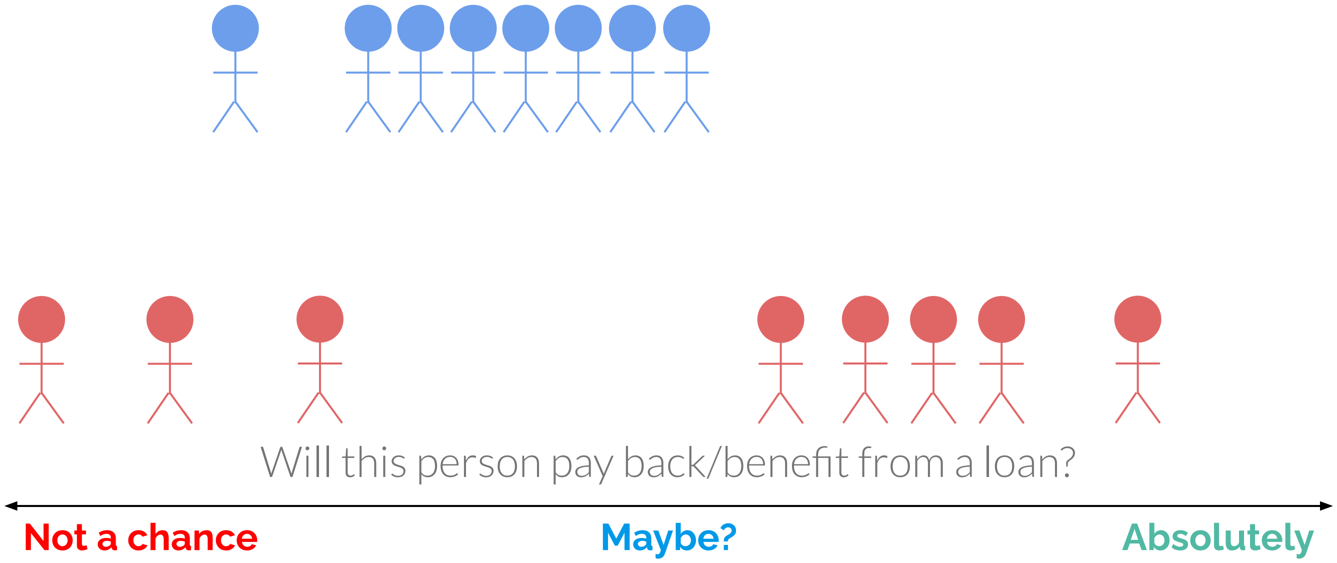


Selective screening

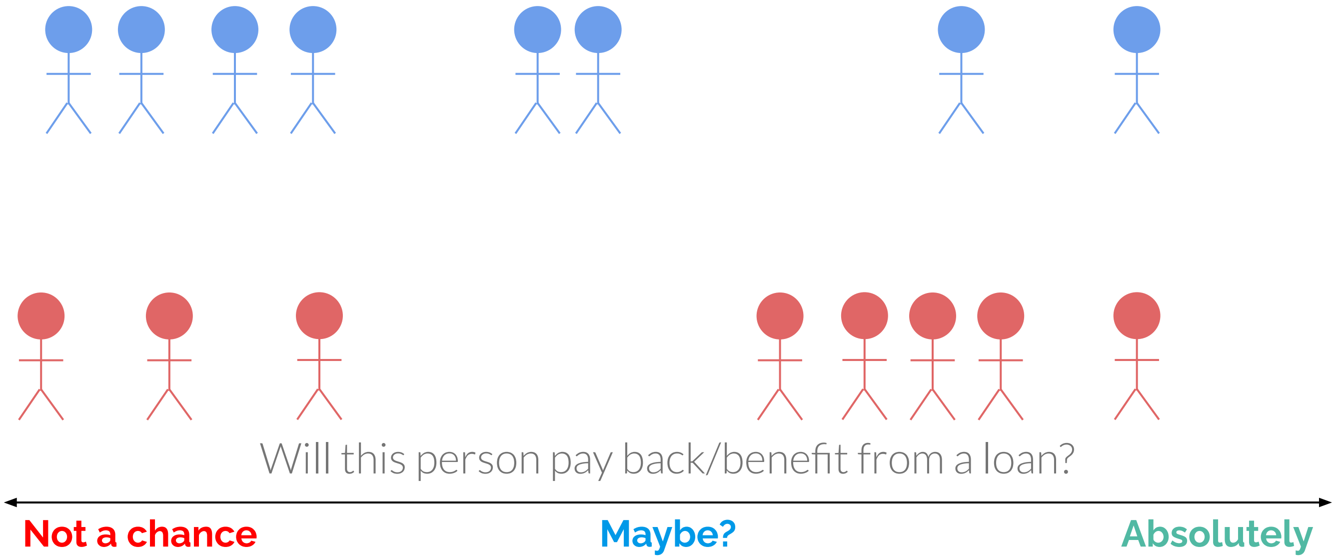
A strategy for reducing inequities

Get more information on *some* individuals without mainstream credit who may in fact be creditworthy.
[e.g., examine household bills — requires time and money]

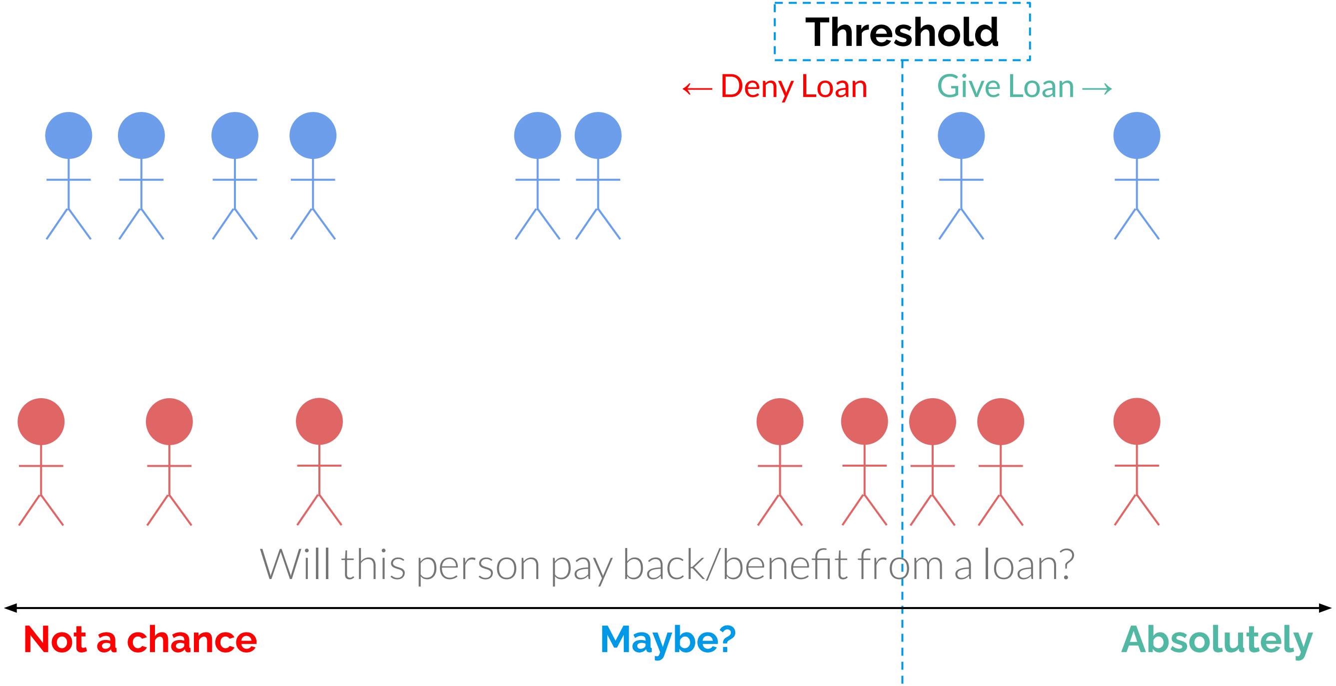
Equity in loans: screening



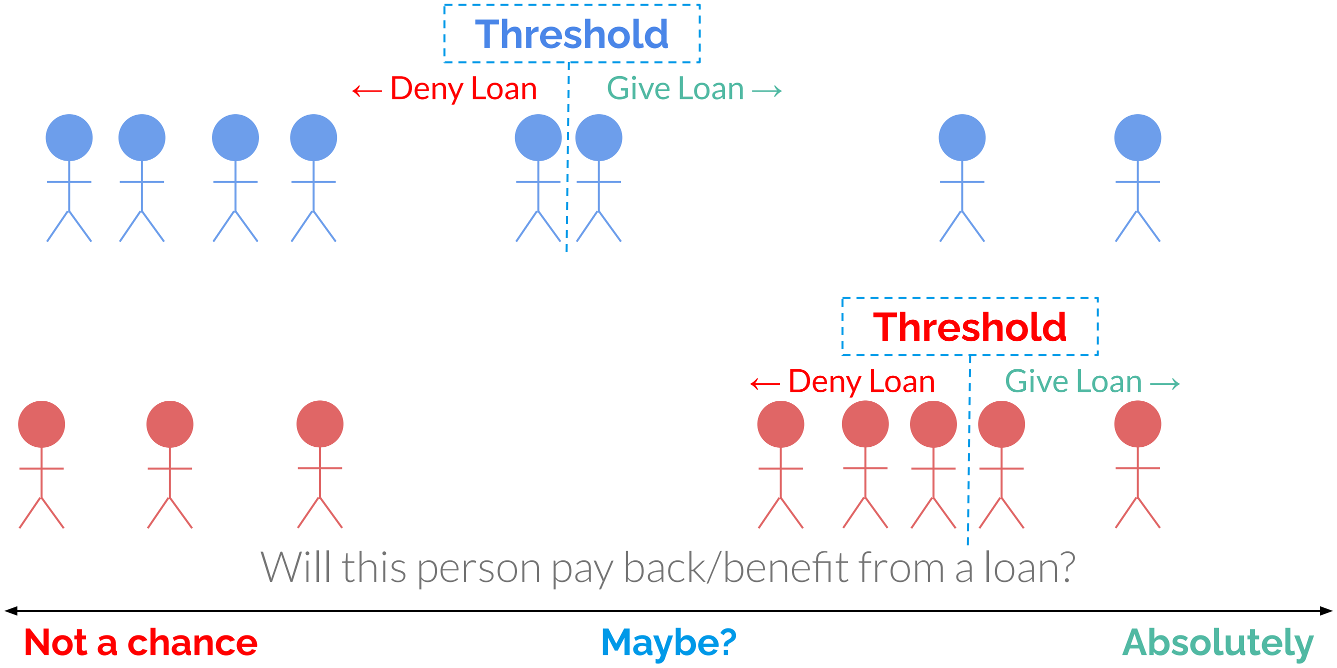
Equity in loans: screening



Equity in loans: screening



Equity in loans: screening



Selective screening

A strategy for reducing inequities

We developed a simple, statistical method for selecting a subset of individuals to screen.

Intuitively, we screen people "close" to the threshold, for whom the added information may plausibly make a difference in the lending decision.

[We formulate the problem as a constrained optimization.]

German credit experiment

Simulation

We conduct a stylized simulation exercise to examine the efficacy of this approach.

German credit experiment

1,000 individuals, 70% of whom are creditworthy.

German credit experiment

1,000 individuals, 70% of whom are creditworthy.

We consider two groups:

1. Those who own a residence [28%]
2. Those who do not [72%]

Greater proportion of homeowners are creditworthy.
[74% vs. 60%]

German credit experiment

1,000 individuals, 70% of whom are creditworthy.

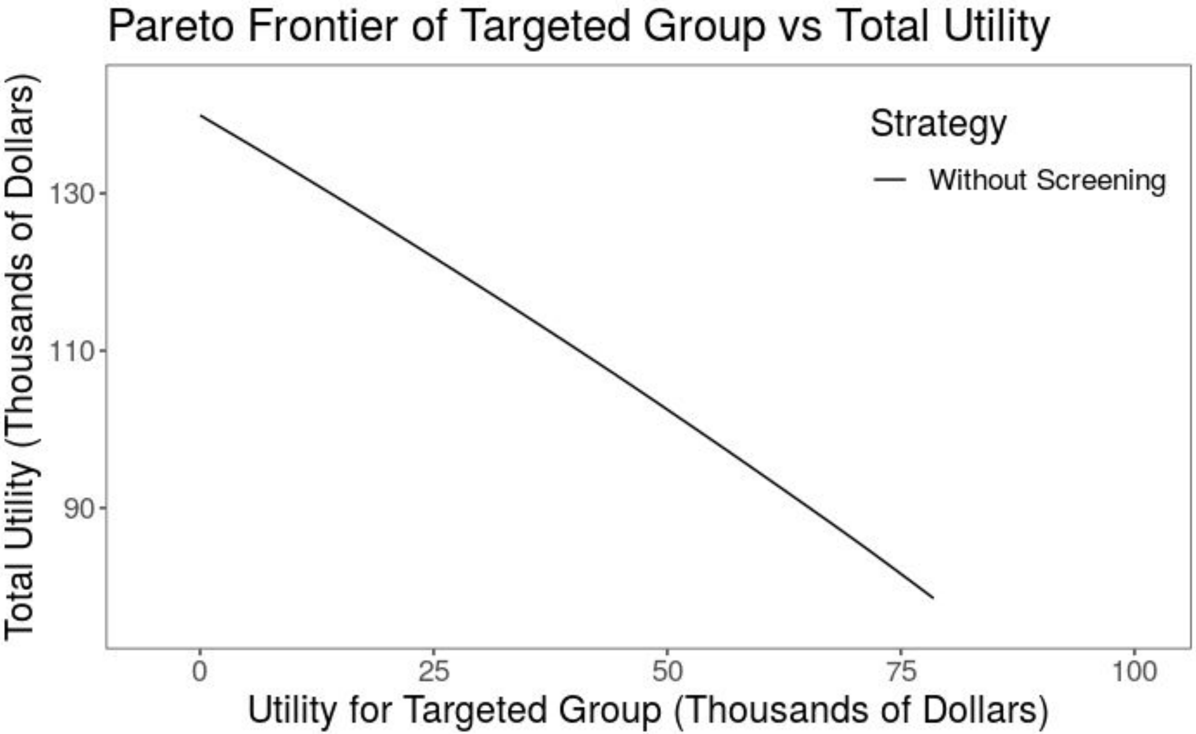
We consider two groups:

1. Those who own a residence [28%]
2. Those who do not [72%]

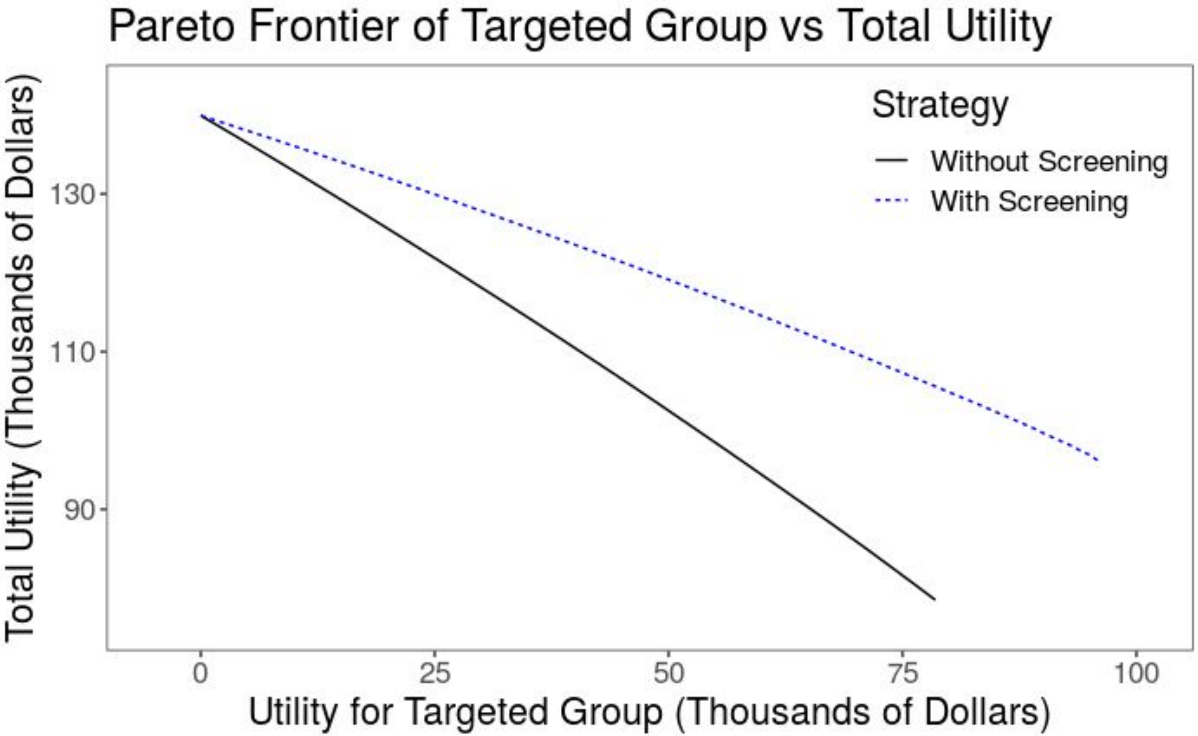
Greater proportion of homeowners are creditworthy.
[74% vs. 60%]

We assume the cost of screening is 10% the loan amount.
[Imagine $1,000 loans with $100 for additional screening.]

German credit experiment



German credit experiment



Summary

Equitable decision making generally requires examining the trade-off between competing concerns.
[Traditional fairness definitions are often overly rigid.]

Important to understand the value of acquiring information and, more broadly, the value of interventions.
[Traditional fairness work treats information as static.]

References

The Measure and Mismeasure of Fairness:
A Critical Review of Fair Machine Learning
Sam Corbett-Davies and Sharad Goel

Fair Allocation through Selective Information Acquisition
William Cai, Johann Gaebler, Nikhil Garg, and Sharad Goel

Stanford Computational Policy Lab

policylab.stanford.edu