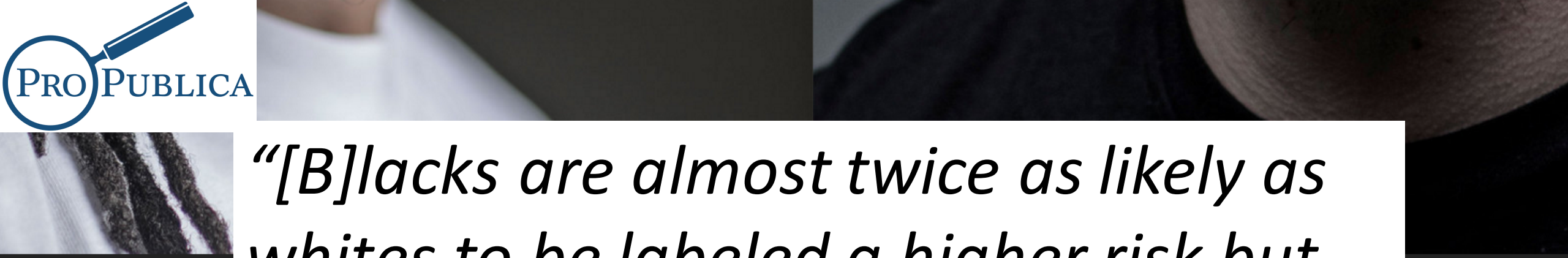
Bias In, Bias Out

Adventures in Algorithmic Fairness

Sandy Mayson ✦ November 3, 2016



“[B]lacks are almost twice as likely as whites to be labeled a higher risk but not actually re-offend. It makes the opposite mistake among whites: They are much more likely than blacks to be labeled lower risk but go on to commit other crimes.”

COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity

Performance of the COMPAS Risk Scales in Broward County

Northpointe Inc.
Research Department

William Dieterich, Ph.D.
Christina Mendoza, M.S.
Tim Brennan, Ph.D.

July 8, 2016

Alternate metrics of fairness

Prediction Fails Differently for Black Defendants

	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

Of people who didn't reoffend, % labeled HR

Of people who did reoffend, % labeled LR

	White	African American
Labeled Higher Risk, But Didn't Re-Offend	41%	37%
Labeled Lower Risk, Yet Did Re-Offend	29%	35%

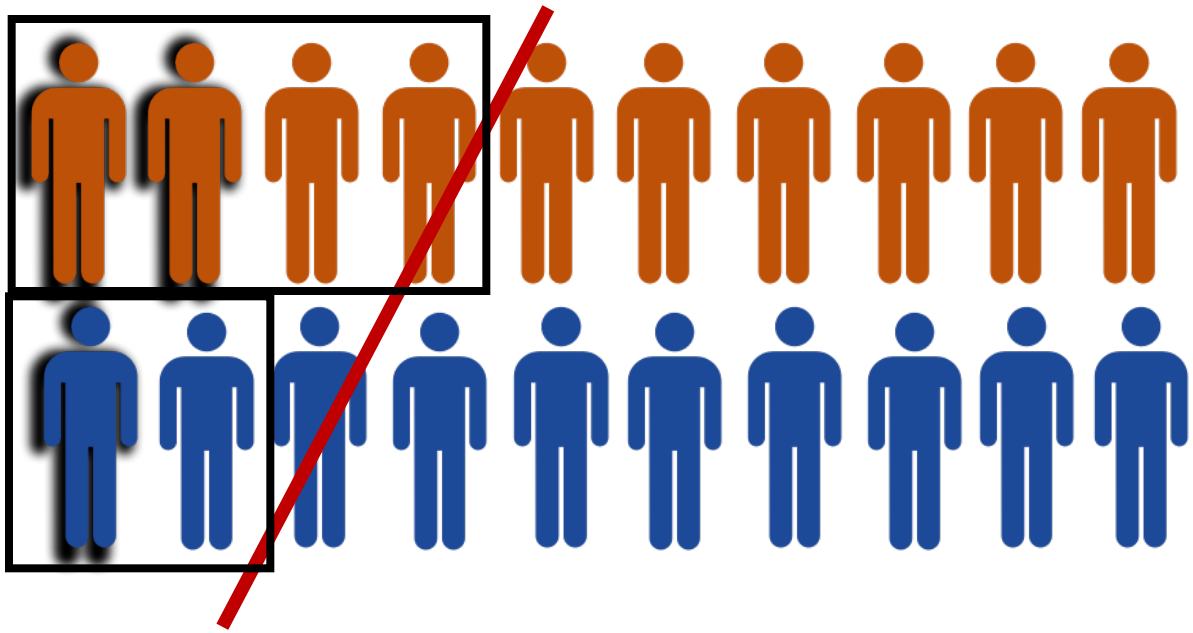
Of people labeled HR, % who didn't reoffend

Of people labeled LR, % who did reoffend

Inequality in the World

Two subgroups with different base rates

Equal predictive accuracy for each group

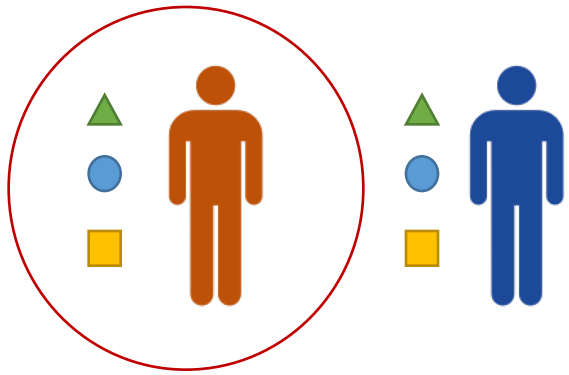


% of group forecast to recidivate	% of non-recidivists forecast to recidivate
40%	25%
20%	11%

What is “Bias”?

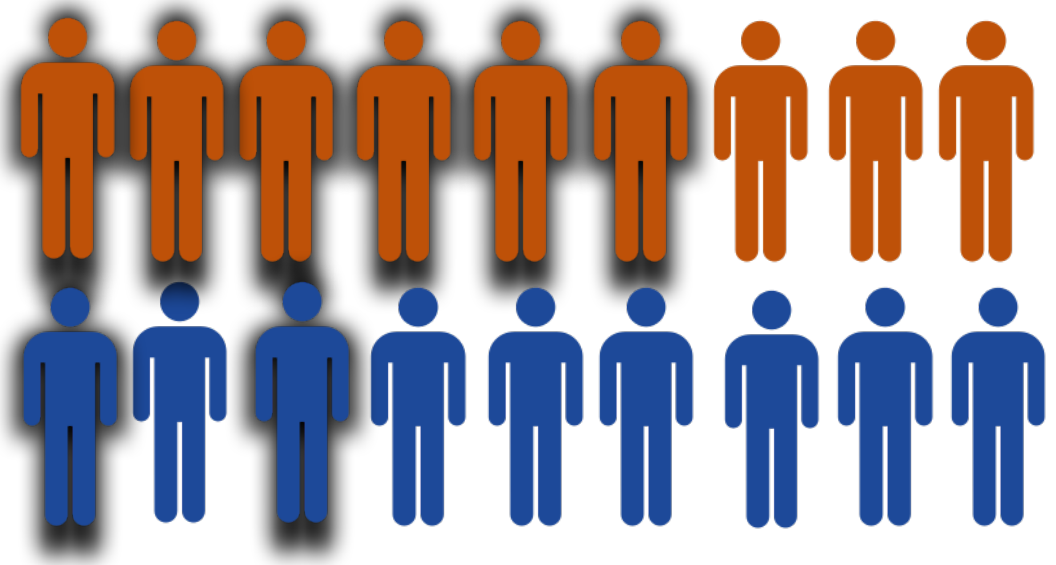
Two Concepts

Disparate Treatment



Formal or intentional differential treatment of “similarly situated” people

Disparate Impact



No intentional discrimination, but disparate outcomes for groups defined by the trait of concern

Disparate Treatment

- Formal / intentional classification on basis of trait of concern
- Concerns individual outcomes
- Fairness metric: *Equal Treatment (with respect to trait of concern)*
- Value: Anti-classification

Relevant Legal Frameworks

- **Federal Constitution: Equal Protection Doctrine**
  - Laws / state actions that explicitly classify or intentionally discriminate on the basis of certain protected characteristics (incl. race, alienage, national origin, classifications that burden "fundamental rights," sex, illegitimacy, maybe sexual orientation) are subject to heightened scrutiny
- **Federal statutory law, e.g. Title VII of the Civil Rights Act of 1964**
  - Prohibits disparate treatment in employment on basis of race, color, religion, sex, or national origin
- **State constitutional and statutory law**

Disparate Impact

- Practices "that are fair in form, but discriminatory in operation." Griggs v. Duke Power Co., 401 U.S. 424, 431 (U.S. 1971)
- Concerns group outcomes
- Fairness metric: *Equal Outcomes (with respect to group trait of concern)*
- Value: Anti-subordination

Relevant Legal Frameworks

- **Federal statutory law, e.g. Title VII of the Civil Rights Act of 1964**
  - Prohibits disparate impact in employment on basis of race, color, religion, sex, or national origin, if the discriminatory practice is not job-related or if employer could have used less discriminatory means
  - Fuzzy guidance on what counts as "disparate impact"

- **Some state statutory law**

Forms of Disparate Impact

And responsive fairness metrics

1. Differential % of subgroups forecast for Outcome X
   - Fairness metric = ***Demographic Parity*** (Hardt et al, 2016), ***Statistical Parity*** (Berk et al, 2016)

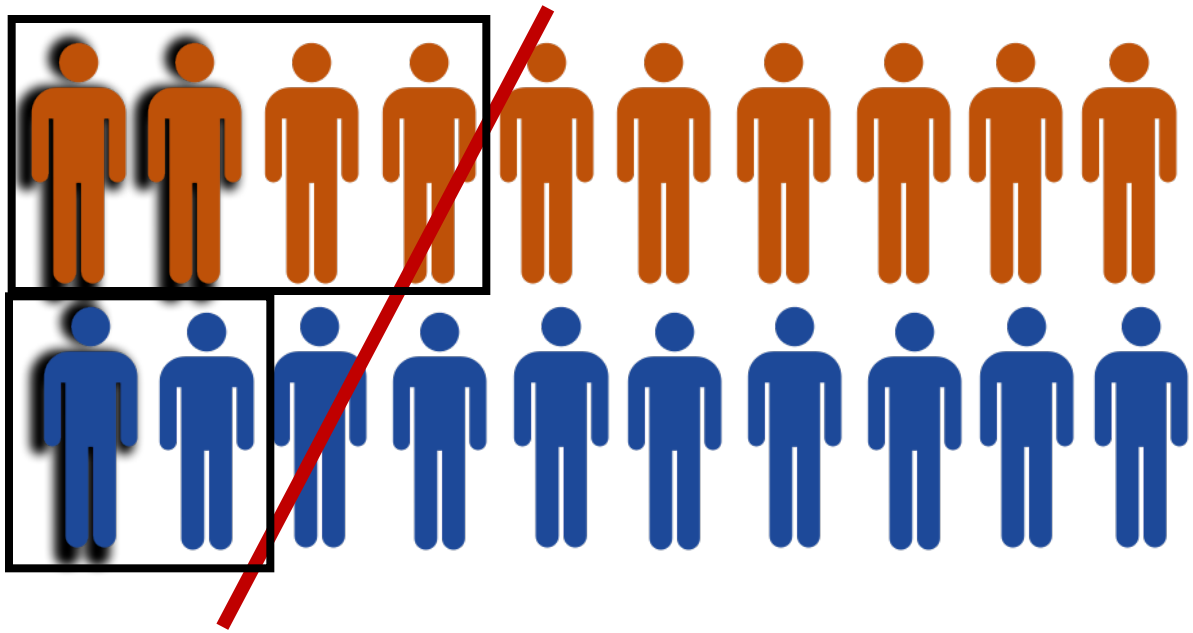
2. Differential predictive accuracy
   - Fairness metric = ***Predictive Parity*** (Northpointe); ***Equality of Outcome*** (Berk et al, 2016)

3. Differential true positive / negative or false positive / negative rates
   - Fairness metrics =
     - ***Procedural parity*** *(with respect to tpr / tnr / fpr / fnr)*
     - ***Equality of opportunity*** *(with respect to a positive or negative classification)* (Hardt et al, 2016; Berk et al, 2016; Joseph et al, 2016)

Inequality in the World

Two subgroups with different base rates

Equal predictive accuracy for each group



	% of group forecast to recidivate	% of non-recidivists forecast to recidivate
	40%	25%
	20%	11%

Other Group Fairness Metrics…

Suggested by Berk et al., 2016

1. ***Overall parity*** – overall procedural accuracy is the same across subgroups ("estimated probability of a correct classification: either a true positive or a true negative")

2. ***Cost ratio equality*** – ratio of false negatives to false positives is the same across subgroups

3. ***Total fairness*** – all metrics of fairness are satisfied

Tradeoffs!

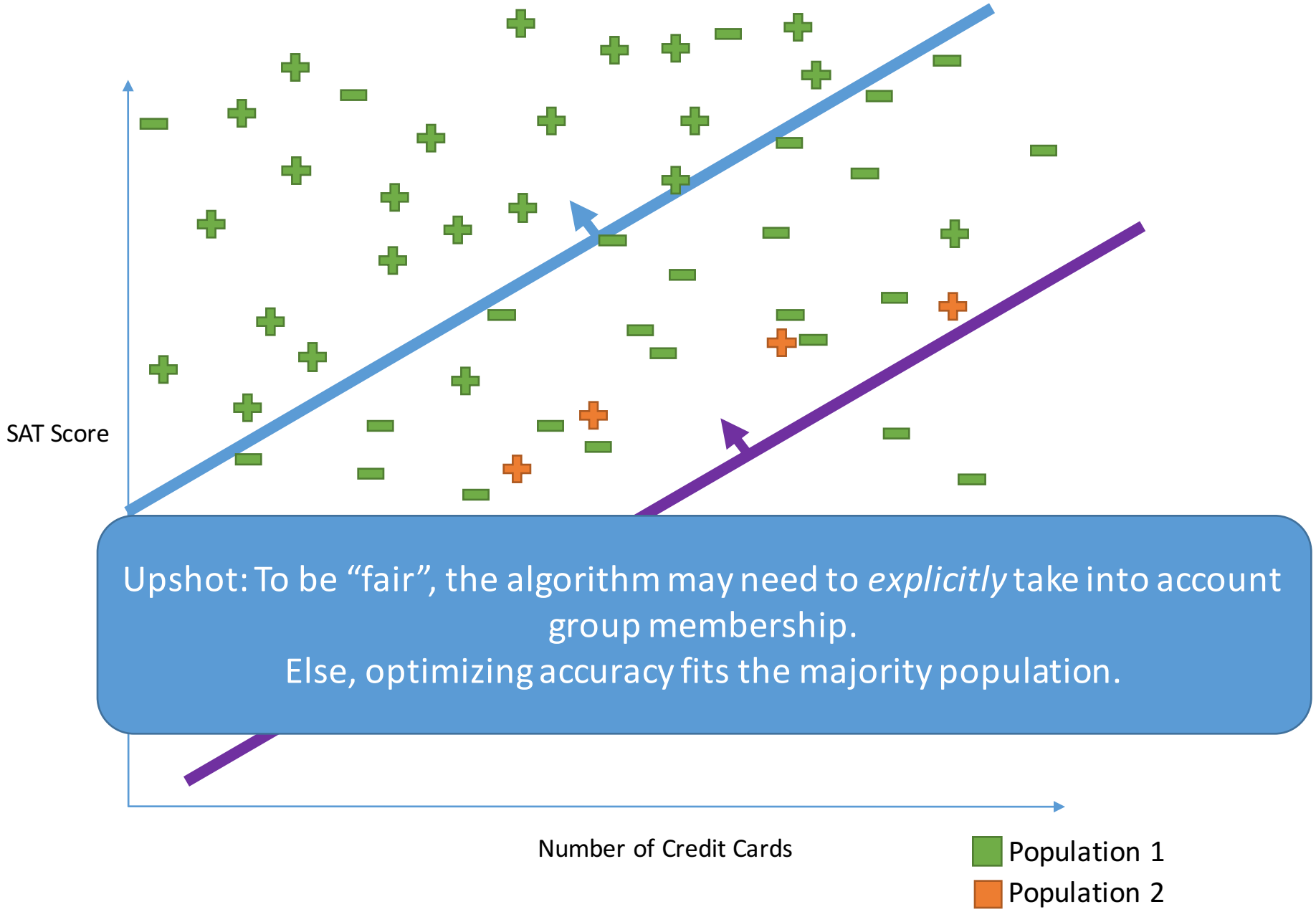
1. Fairness v. Accuracy

2. Fairness v. Fairness

   - *Disparate treatment v. disparate impact*
     - Most obvious way to mitigate disparate impact is to treat subgroups differently (e.g. affirmative action)
     - Obstacle = Equal protection, anti-discrimination statutes

   - *Disparate impact (predictive parity) v. disparate impact (demographic parity, or equality of opportunity)*

(From Aaron Roth's presentation for Optimizing Government series, September 22, 2016.)

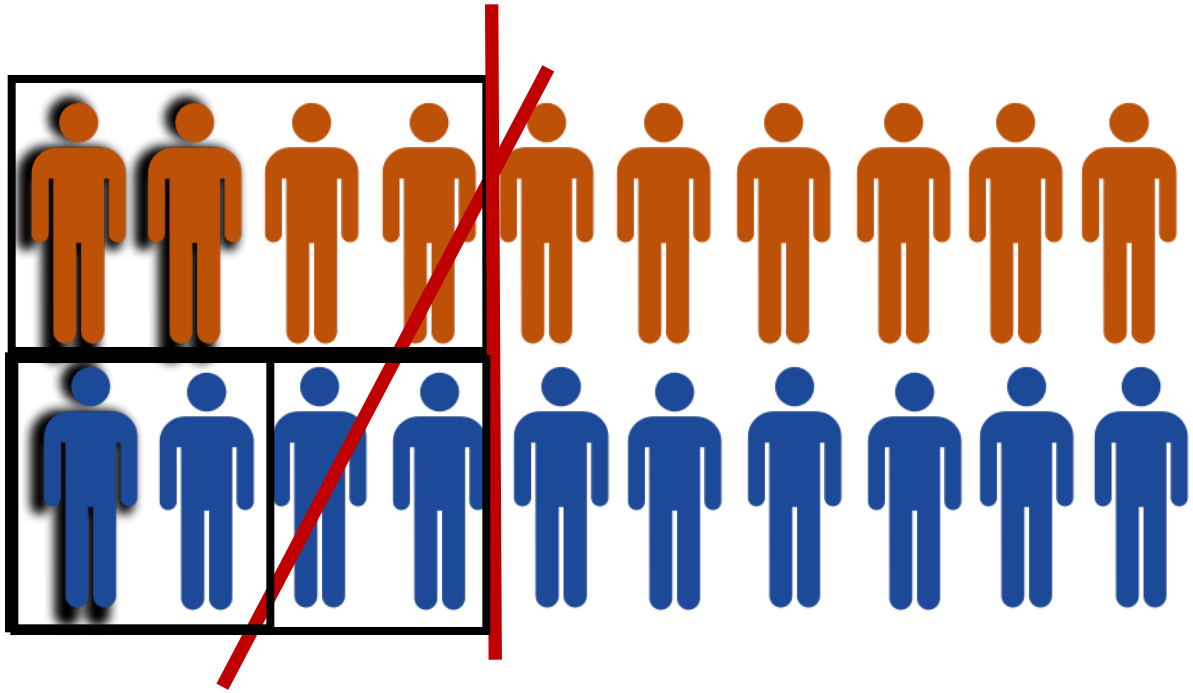


Upshot: To be "fair", the algorithm may need to *explicitly* take into account group membership.
Else, optimizing accuracy fits the majority population.

Inequality in the World

Two subgroups with different base rates

UN Equal predictive accuracy for each group



% of group forecast to recidivate	% of non-recidivists forecast to recidivate
40%	25%
20%	11%
40%	33%

Situating "Fairness in Learning"

Joseph, Kearns, Morgenstern & Roth

Fairness metric: *Within a given pool of applicants, "a worse applicant is never favored over a better one."*

- Aim to guarantee "fairness at the individual level"
- Responsive to concern about disparate treatment of two people similarly situated (equally qualified for Outcome X, insofar as that is knowable on the basis of the data)

Situating "Fairness in Learning"

In the Disparate Treatment / Impact Framework

Disparate Treatment

✓ On basis of non-merit-relevant factor

? On basis of race, sex, national origin, religion, etc.

Disparate Impact

X Predictive Parity

X Demographic / Statistical Parity

- Procedural Parity, w/r/t…
  - ✓ True positive rate
  - ? True negative rate
  - False positive rate
  - False negate rate

X Overall parity

X Cost ratio equality

X Total fairness

Additional Notes

On "*Fairness in Learning*"

1. **Relies on randomization**
   - Ensures that equally qualified people have an equal *chance* of Outcome X, but may choose randomly among a group of equally qualified people.
   - Equality of opportunity in a roll of the dice; not equality of outcome.
   - Is randomness arbitrariness?

2. **Model is a sequential decisionmaker, learns quickly under relatively stable conditions, can learn through exploration**
   - Vs. some machine-learning applications, like criminal justice risk assessment (algorithms applied in static form for some amount of time; can't engage in randomized exploration)

3. **Doesn't address problems with proxy outcome measures**
   - *E.g.* in criminal justice we can't accurately measure *commission* of new crime, so resort to a proxy measure (*arrest* for new crime) that may embed a degree of irrational discrimination.