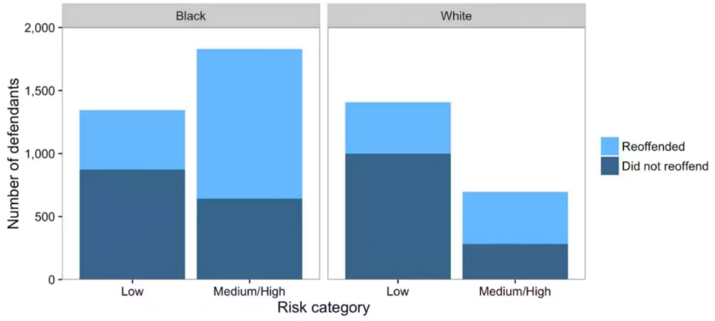
Tufts COMP 135: Introduction to Machine Learning

https://www.cs.tufts.edu/comp/135/2019s/

Fairness, Ethics, and Machine Learning



Many ideas/slides attributable to:
Alexandra Chouldechova
Moritz Hardt

Prof. Mike Hughes

Fairness:
Unit Objectives

- How to think systematically about end-to-end ML
  - Where does data come from? What might be missing?
  - What features am I measuring? What protected information can leak in unintentionally?
  - Who will be impacted?
- How to define and measure notions fairness
  - Use concepts: accuracy, TPR, FPR, PPV, NPV
  - What is achievable? What is impossible?

Example Concerns about Fairness

Unfair Hiring?

TheUpshot

ROBO RECRUITING

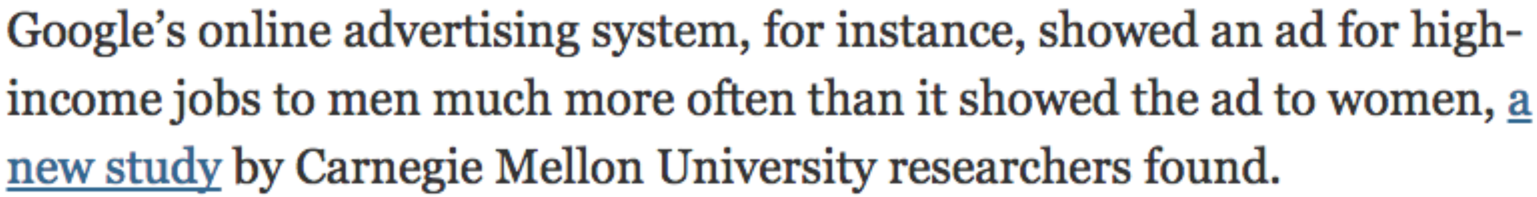
Can an Algorithm Hire Better Than a Human?

Claire Cain Miller @clairecm JUNE 25, 2015

"[H]iring could become faster and less expensive, and [...] lead recruiters to more highly skilled people who are better matches for their companies. Another potential result: a more diverse workplace. The software relies on data to surface candidates from a wide variety of places and match their skills to the job requirements, free of human biases."

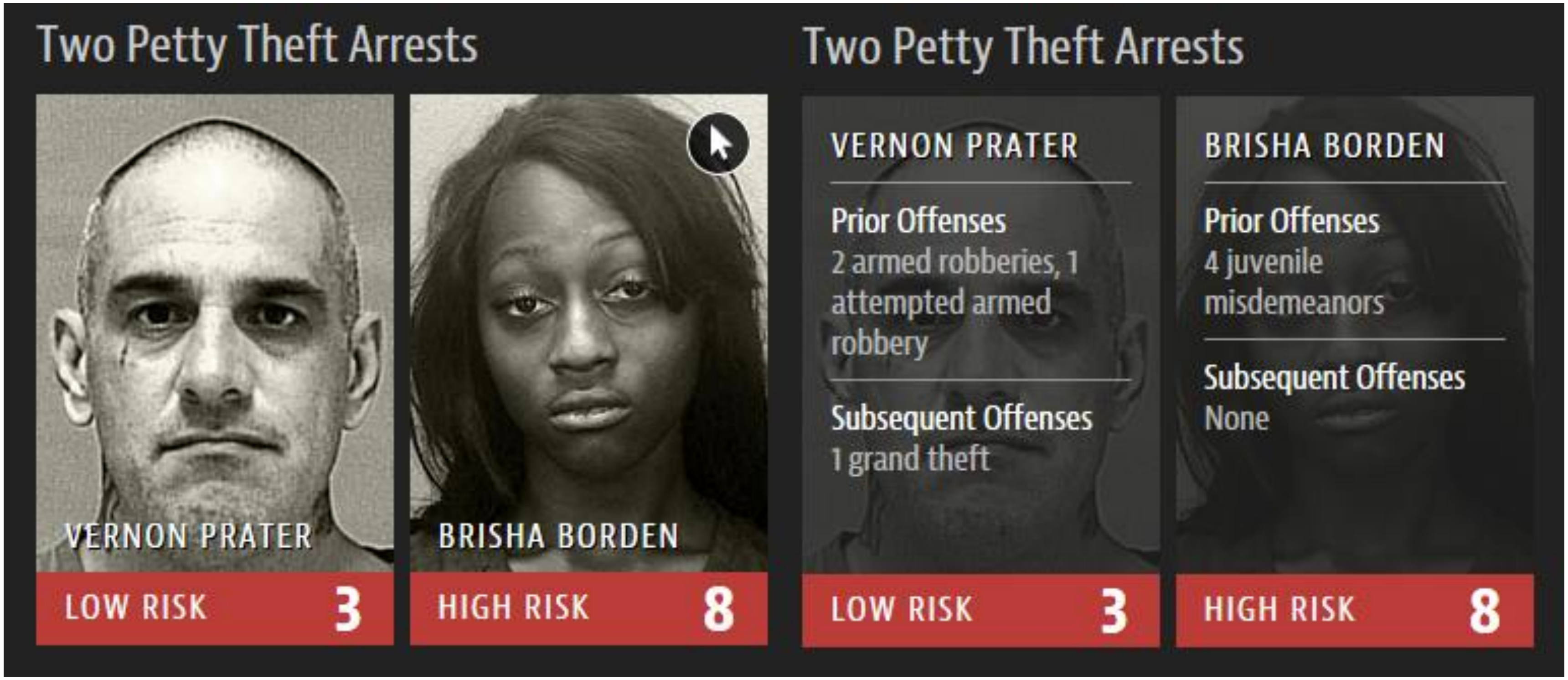
Miller (2015)

Unfair job recommendations?



Google's online advertising system, for instance, showed an ad for high-income jobs to men much more often than it showed the ad to women, a new study by Carnegie Mellon University researchers found.

Unfair Future Crime Prediction?



Two Petty Theft Arrests

VERNON PRATER

LOW RISK 3

BRISHA BORDEN

HIGH RISK 8

Two Petty Theft Arrests

VERNON PRATER

Prior Offenses
2 armed robberies, 1 attempted armed robbery

Subsequent Offenses
1 grand theft

LOW RISK 3

BRISHA BORDEN

Prior Offenses
4 juvenile misdemeanors

Subsequent Offenses
None

HIGH RISK 8

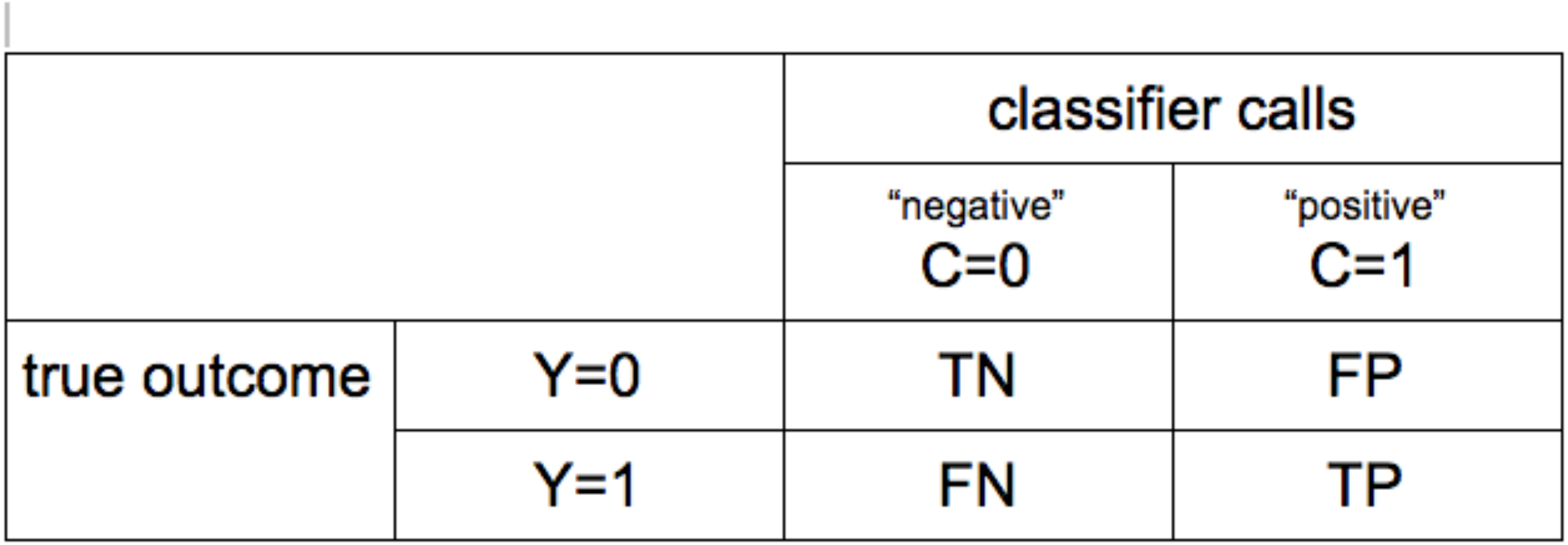
“Recidivism” means likelihood of being a repeat offender

Focus Today: Binary Classifier

- Let’s say we have two groups, A and B
  - Could be any protected group (race / gender / age)

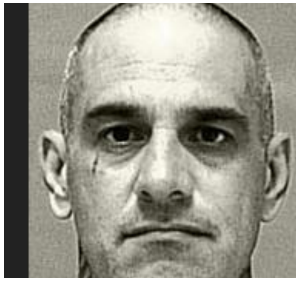
- We’re trying to build a binary classifier that will predict individuals as HIGH or LOW risk
  - Likelihood of recidivism
  - Ability to pay back a loan

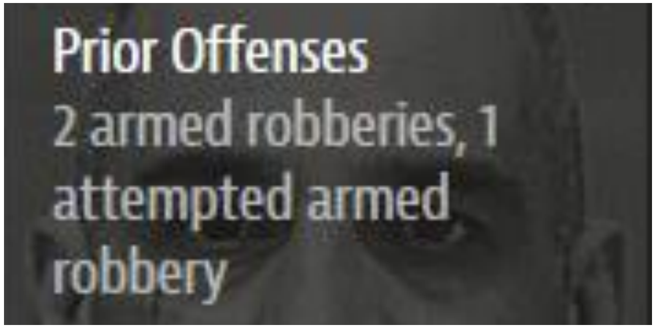
Notation for Binary Classifier



		classifier calls	
		“negative” C=0	“positive” C=1
true outcome	Y=0	TN	FP
	Y=1	FN	TP

Case Study: The COMPAS future crime prediction algorithm

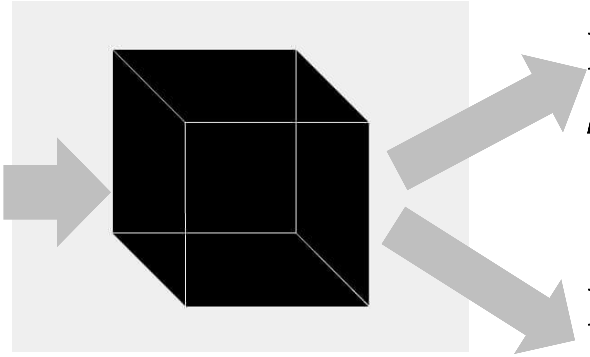




Prior Offenses
2 armed robberies, 1 attempted armed robbery

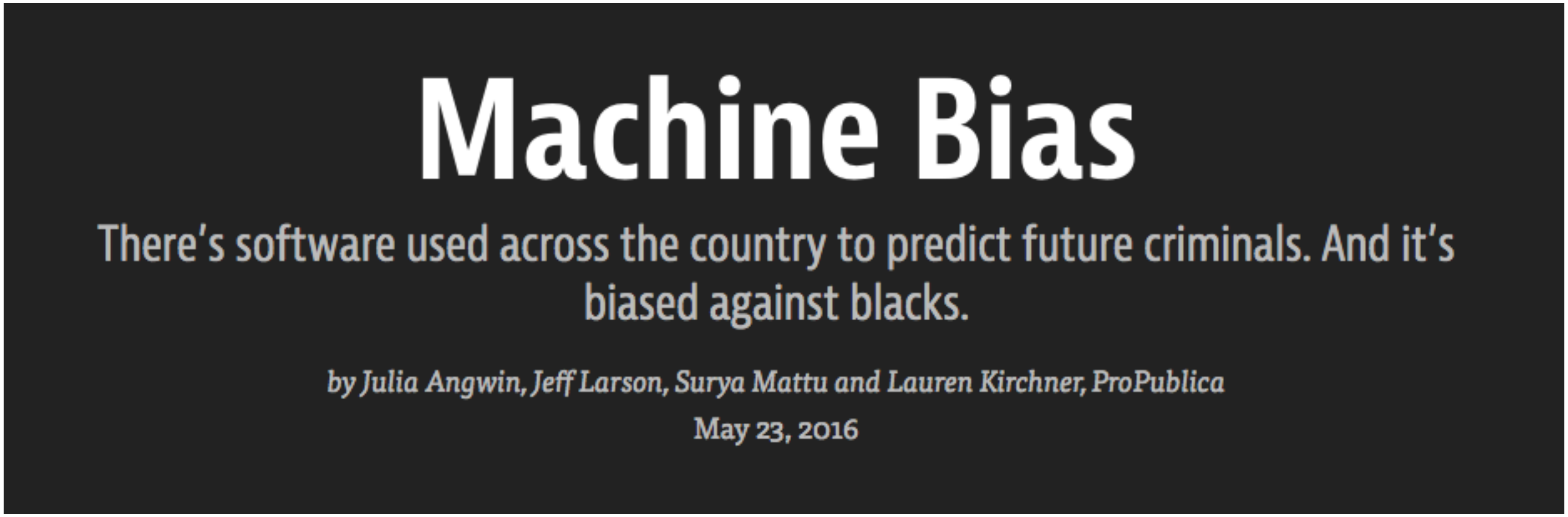
other features (e.g. demographics, questionnaire answers, family history)

COMPAS classifier



HIGH RISK of future crime
hold in jail before trial

LOW RISK of future crime
release before trial

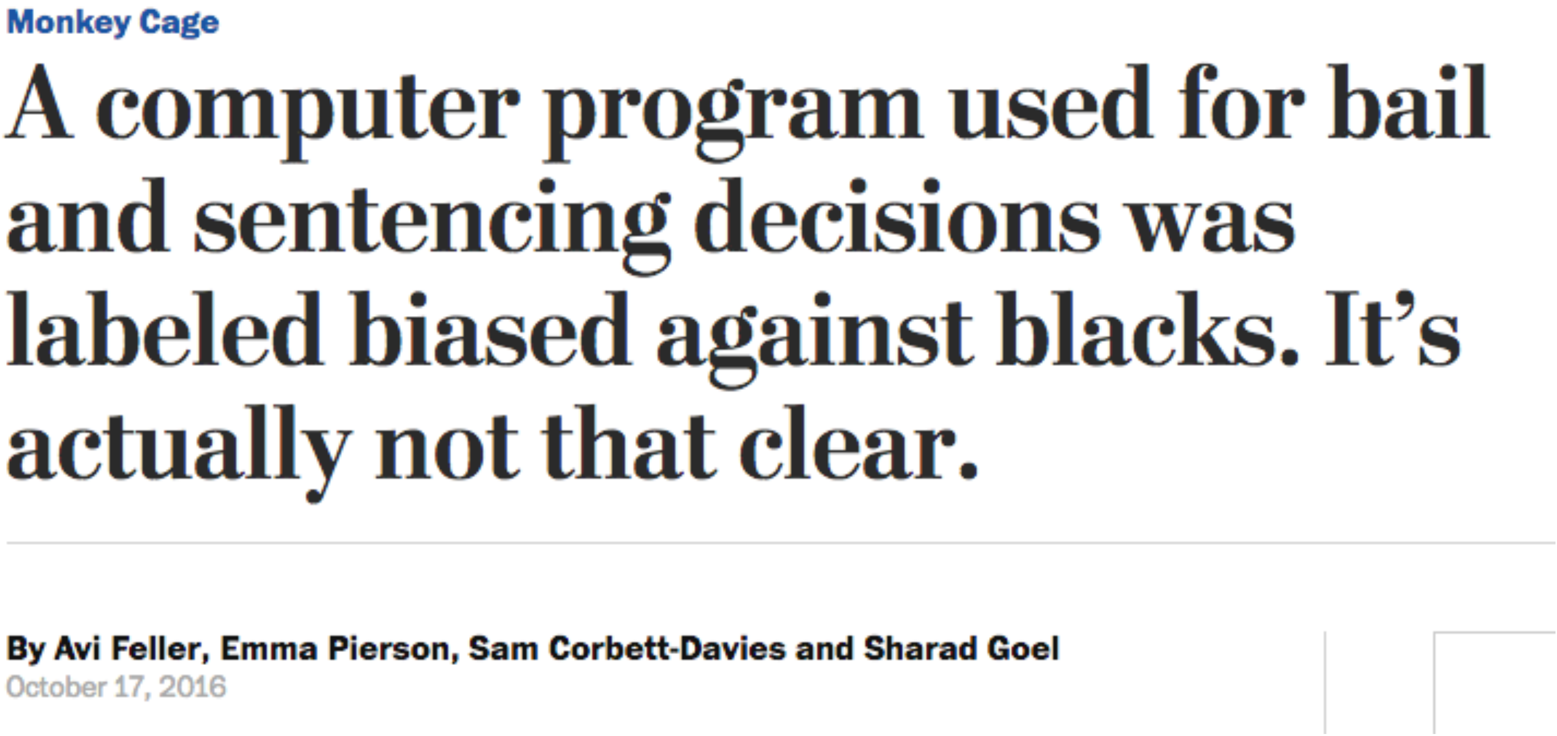


Machine Bias

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



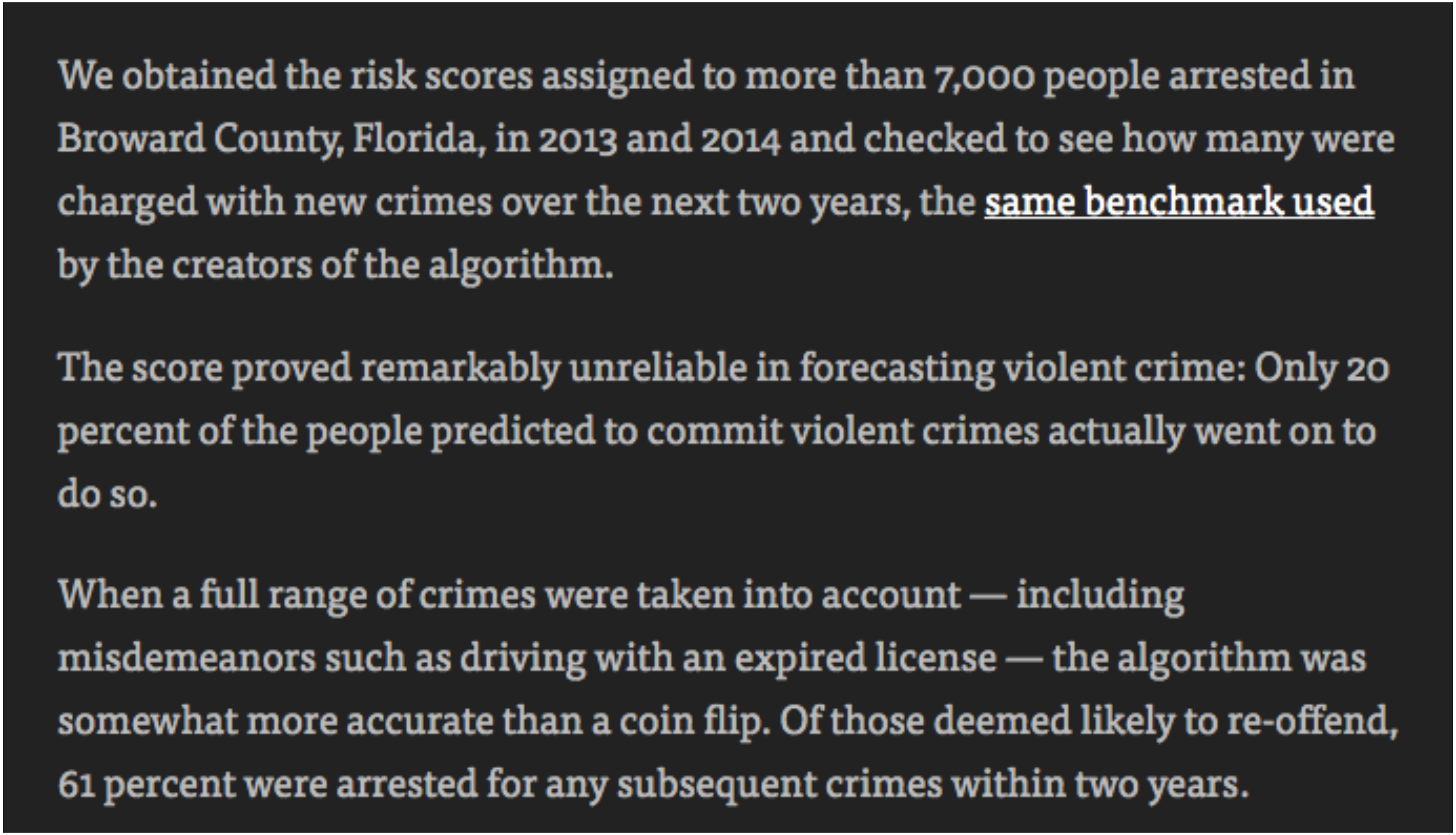
Monkey Cage

A computer program used for bail and sentencing decisions was labeled biased against blacks. It’s actually not that clear.

By Avi Feller, Emma Pierson, Sam Corbett-Davies and Sharad Goel

October 17, 2016

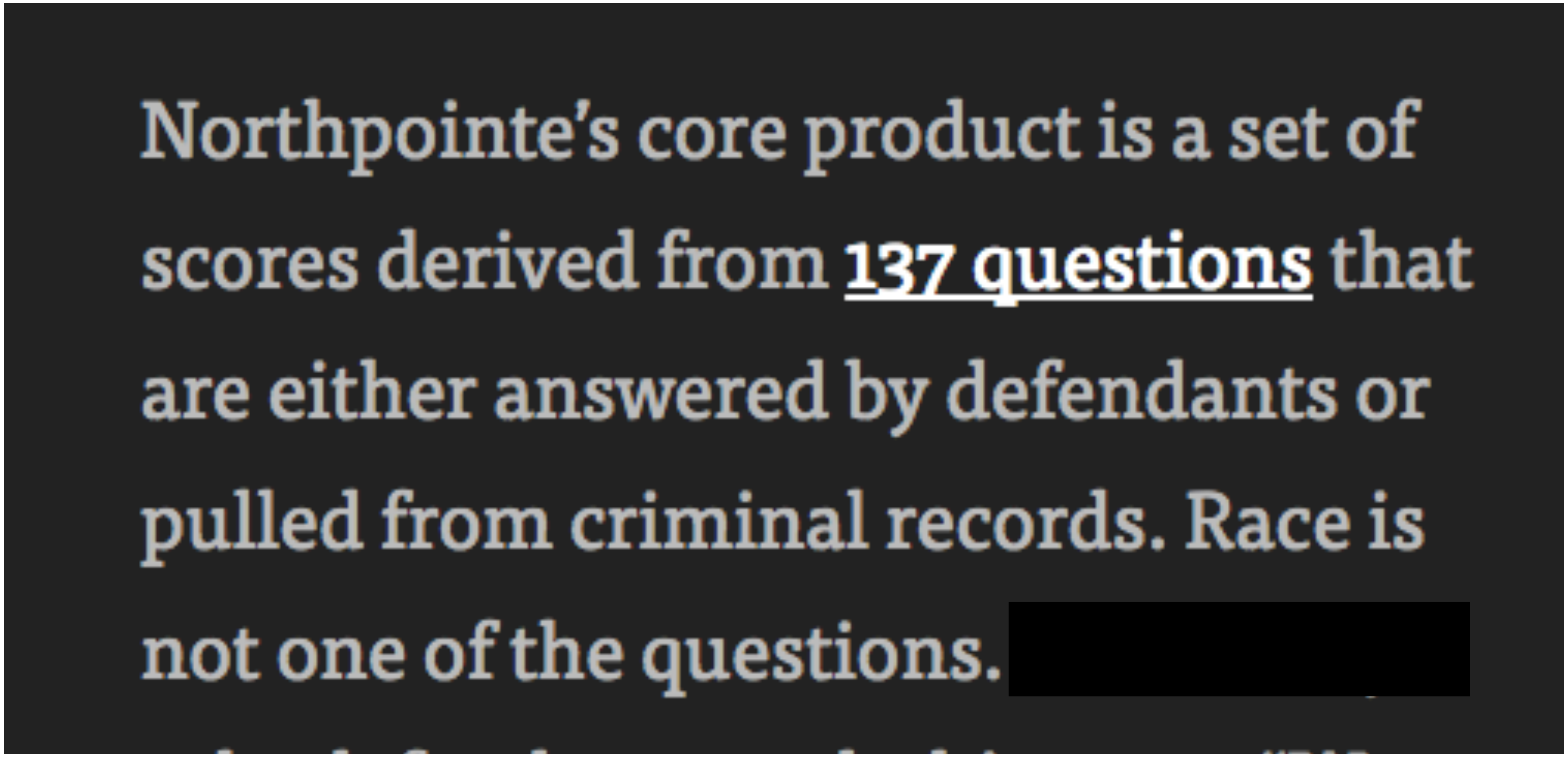
The COMPAS tool assigns defendants scores from 1 to 10 that indicate how likely they are to reoffend based on more than 100 factors, including age, sex and criminal history. Notably, race is not used. These scores profoundly affect defendants’ lives: defendants who are defined as medium or high risk, with scores of 5-10, are more likely to be detained while awaiting trial than are low-risk defendants, with scores of 1-4.



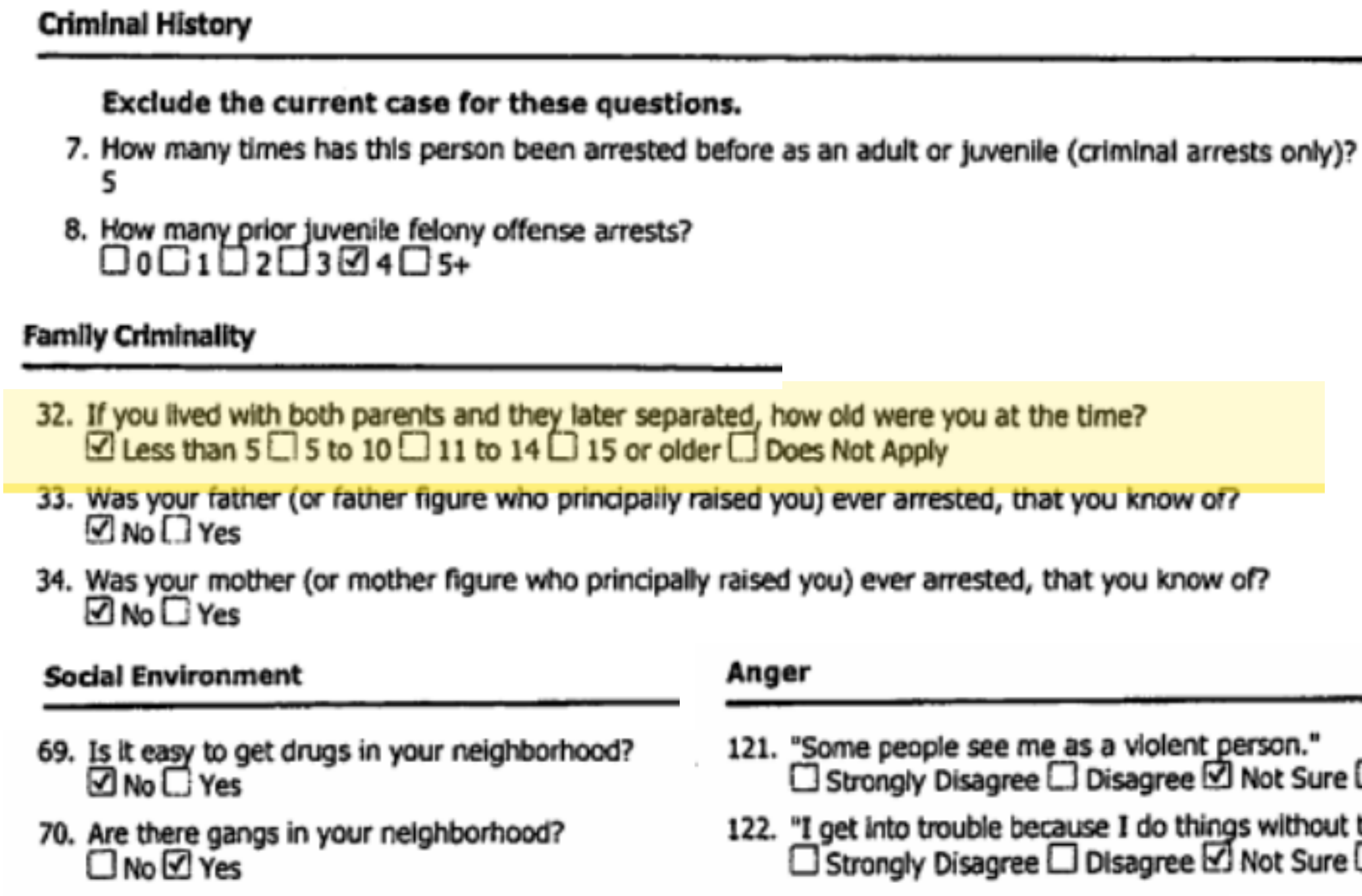
We obtained the risk scores assigned to more than 7,000 people arrested in Broward County, Florida, in 2013 and 2014 and checked to see how many were charged with new crimes over the next two years, the same benchmark used by the creators of the algorithm.

The score proved remarkably unreliable in forecasting violent crime: Only 20 percent of the people predicted to commit violent crimes actually went on to do so.

When a full range of crimes were taken into account — including misdemeanors such as driving with an expired license — the algorithm was somewhat more accurate than a coin flip. Of those deemed likely to re-offend, 61 percent were arrested for any subsequent crimes within two years.



Northpointe's core product is a set of scores derived from 137 questions that are either answered by defendants or pulled from criminal records. Race is not one of the questions. [redacted]



Criminal History

Exclude the current case for these questions.

7. How many times has this person been arrested before as an adult or juvenile (criminal arrests only)?
5

8. How many prior juvenile felony offense arrests?
☐ 0 ☐ 1 ☐ 2 ☐ 3 ☑ 4 ☐ 5+

Family Criminality

32. If you lived with both parents and they later separated, how old were you at the time?
☑ Less than 5 ☐ 5 to 10 ☐ 11 to 14 ☐ 15 or older ☐ Does Not Apply

33. Was your father (or father figure who principally raised you) ever arrested, that you know of?
☑ No ☐ Yes

34. Was your mother (or mother figure who principally raised you) ever arrested, that you know of?
☑ No ☐ Yes

Social Environment

69. Is it easy to get drugs in your neighborhood?
☑ No ☐ Yes

70. Are there gangs in your neighborhood?
☐ No ☑ Yes

Anger

121. "Some people see me as a violent person."
☐ Strongly Disagree ☐ Disagree ☑ Not Sure

122. "I get into trouble because I do things without
☐ Strongly Disagree ☐ Disagree ☑ Not Sure

Full Document: https://www.documentcloud.org/documents/2702103-Sample-Risk-Assessment-COMPAS-CORE.html

Breakout

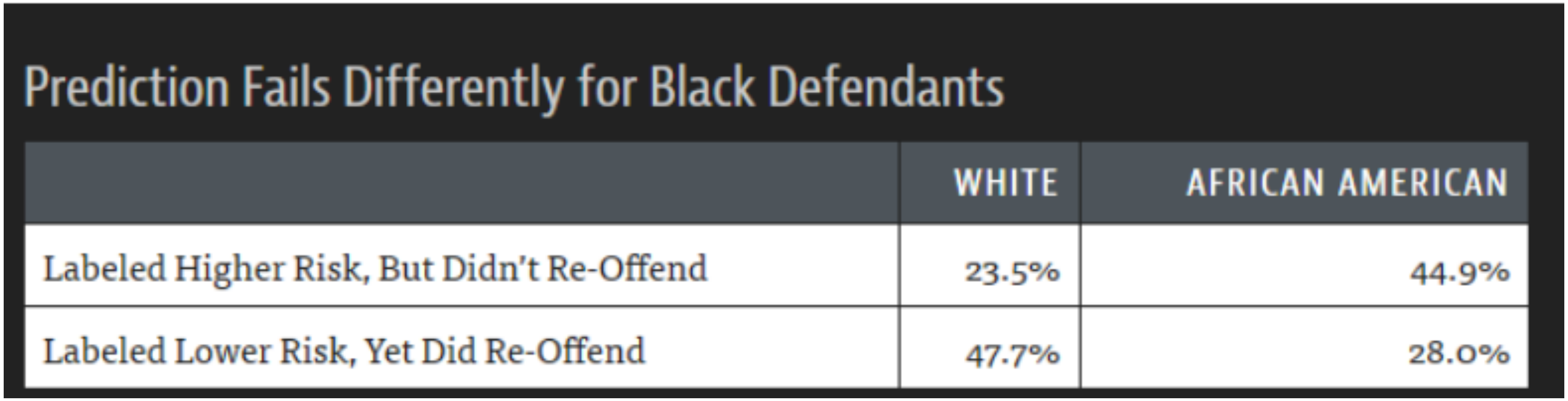
Open in your browser: Discussion Guide GDoc

(link on the schedule page for today)

Three parts, in order

1) Discussion of readings
    1) Follow prompts in the Guide
2) Interactive exploration: “Loan Classifier”
    1) Follow link in the Guide
3) Worksheet: When can a classifier be fair to two groups?

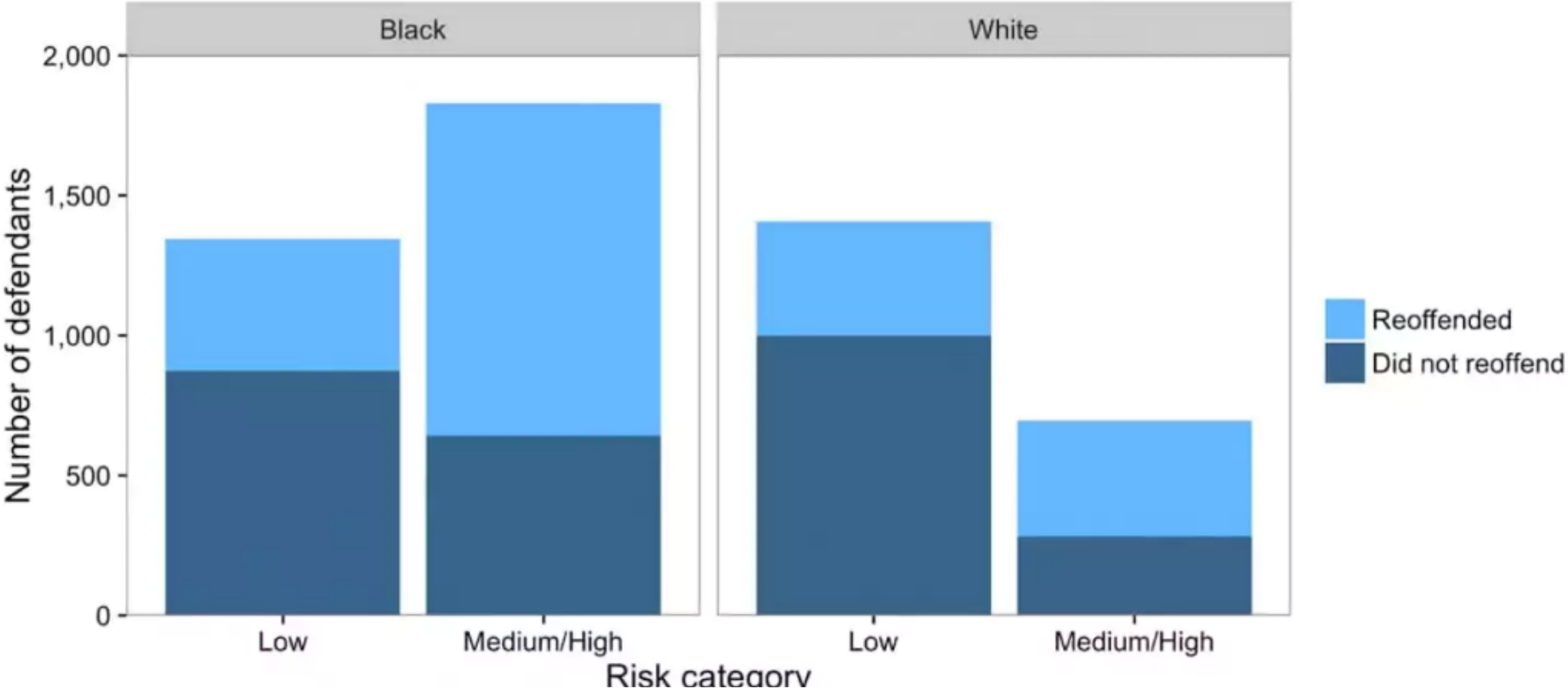
ProPublica says:
“Groups have different **False Pos. Rates**”



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%

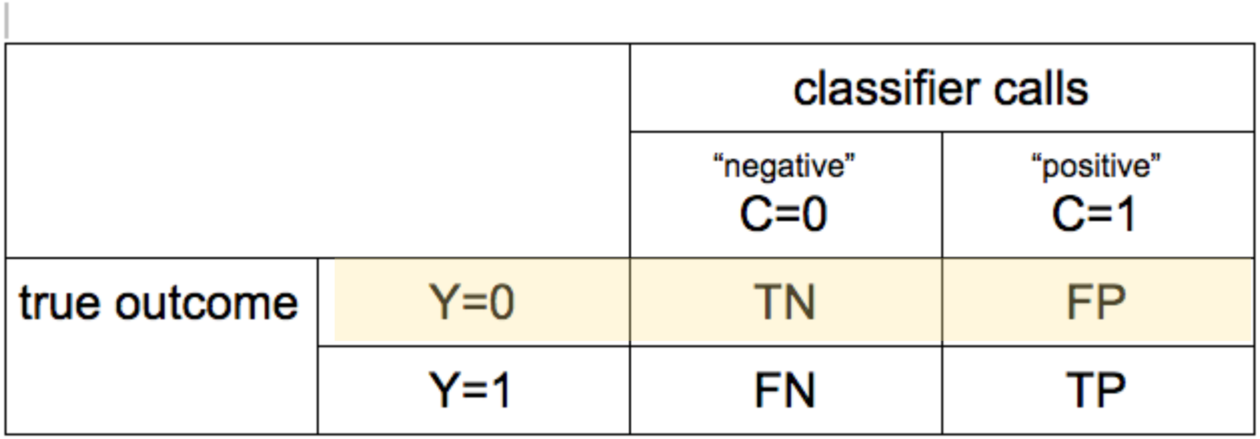
Compas Team Says:
"Groups have same **predictive value**"



- Within each risk category, the proportion of defendants who reoffend is approximately the same regardless of race; this is Northpointe's definition of fairness.

False Positive Rate $= \dfrac{FP}{FP + TN}$

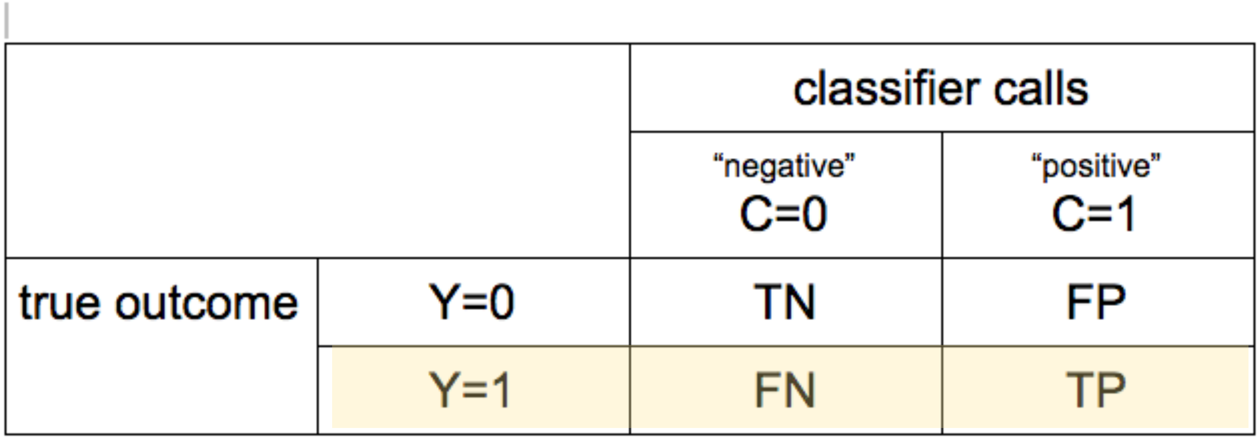
- When true outcome is 0, how often does classifier say “1”.



		classifier calls	
		“negative” C=0	“positive” C=1
true outcome	Y=0	TN	FP
	Y=1	FN	TP

True Positive Rate $= \dfrac{TP}{TP + FN}$

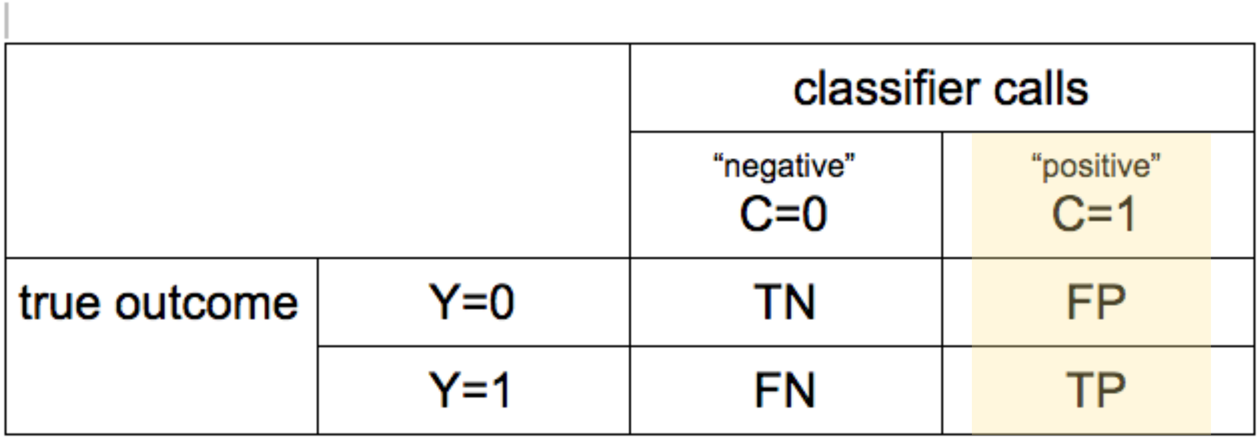
- When true outcome is 1, how often does classifier say “1”.



		classifier calls	
		“negative” C=0	“positive” C=1
true outcome	Y=0	TN	FP
	Y=1	FN	TP

Positive Predictive Value = $\frac{\text{TP}}{\text{TP + FP}}$

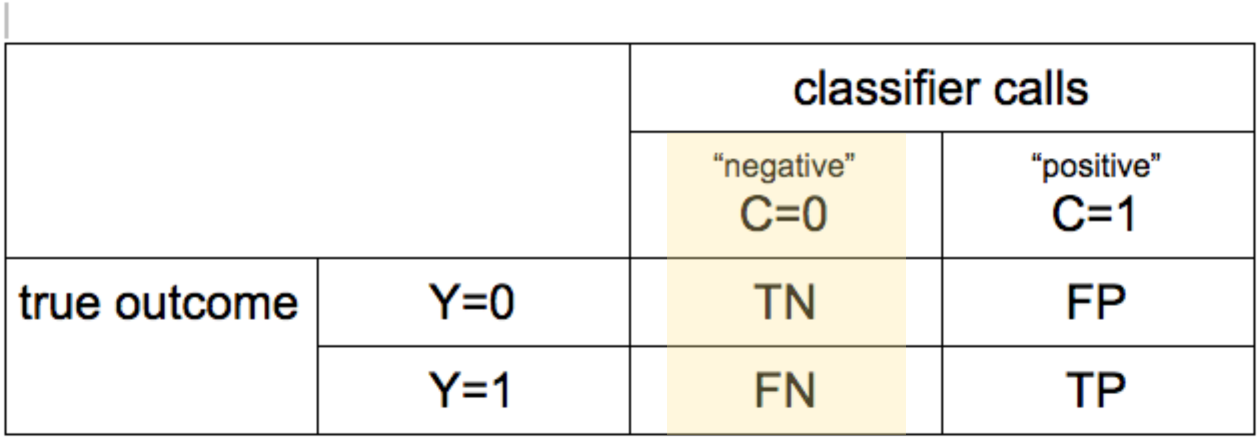
When classifier says “1”, how often is true label 1.



		classifier calls	
		“negative” C=0	“positive” C=1
true outcome	Y=0	TN	FP
	Y=1	FN	TP

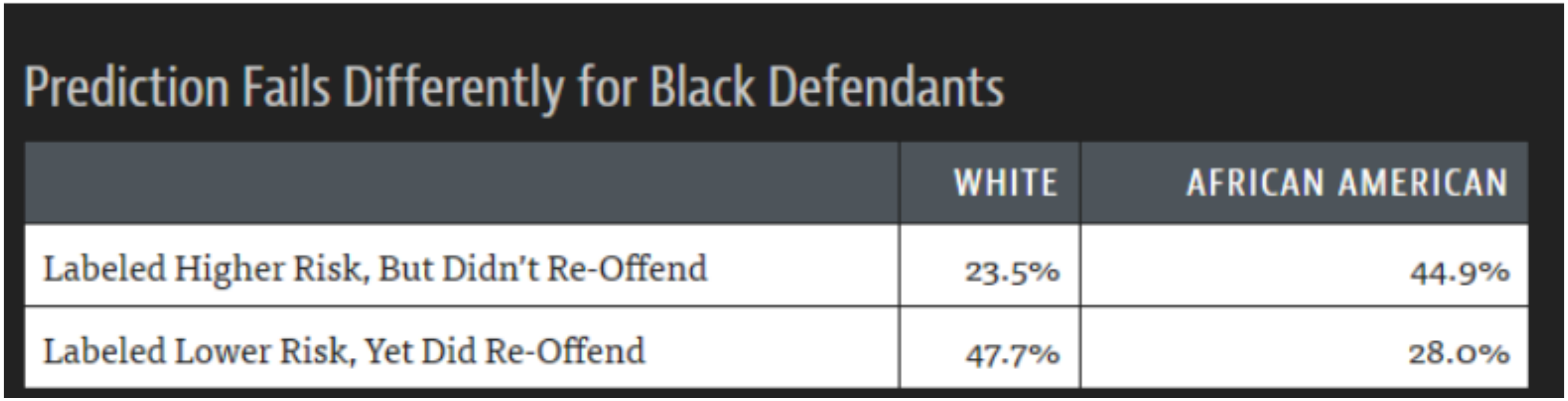
Negative Predictive Value = $\frac{\text{TN}}{\text{TN + FN}}$

When classifier says “0”, how often is true label 0.



		classifier calls	
		“negative” C=0	“positive” C=1
true outcome	Y=0	TN	FP
	Y=1	FN	TP

ProPublica says:
“Groups have different **False Pos. Rates**”



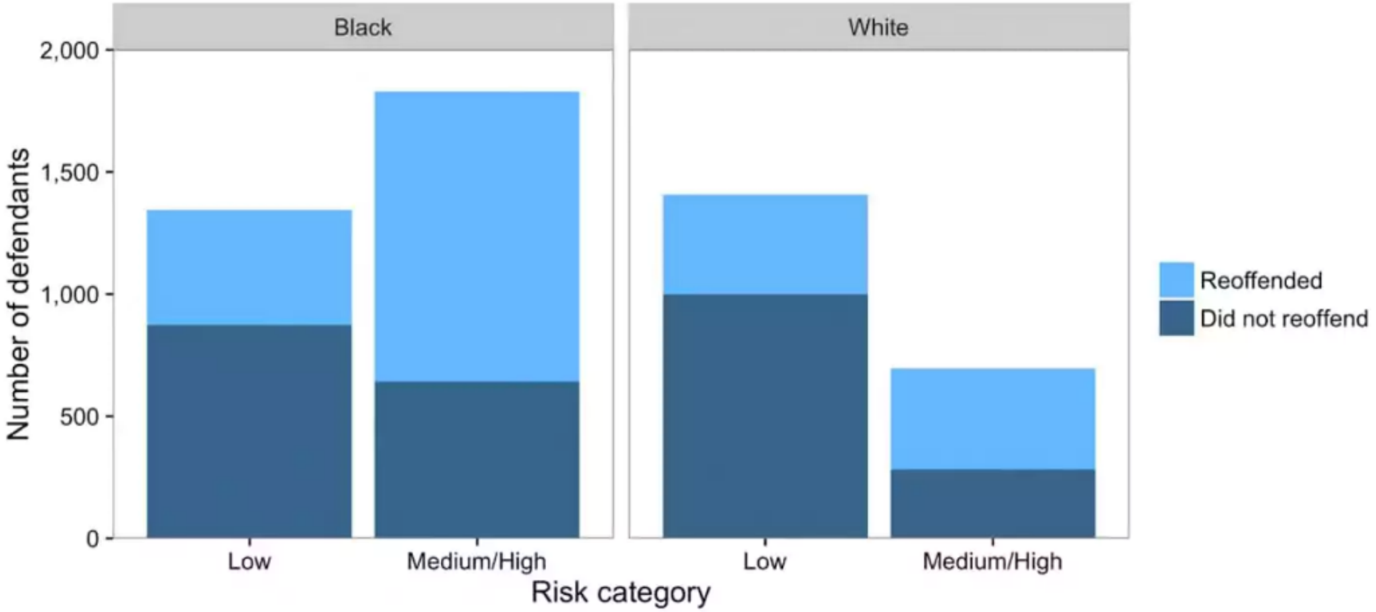
Prediction Fails Differently for Black Defendants

	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%

		classifier calls	
		“negative” C=0	“positive” C=1
true outcome	Y=0	TN	FP
	Y=1	FN	TP

$$\frac{FP}{FP + TN}$$

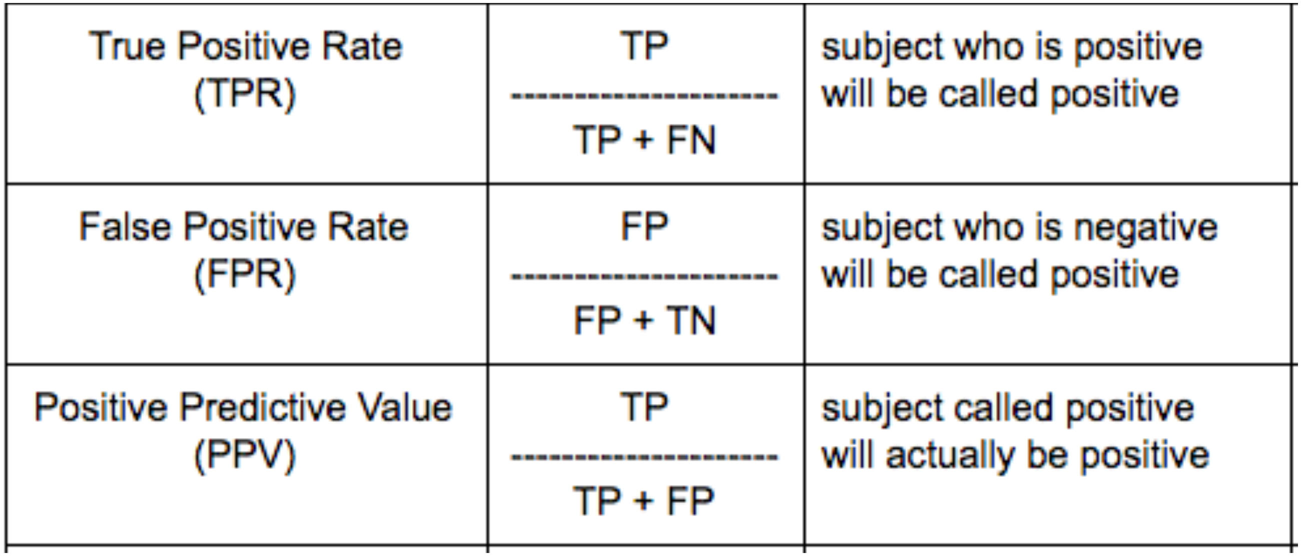
Compas Team Says:
“Groups have same **predictive value**”



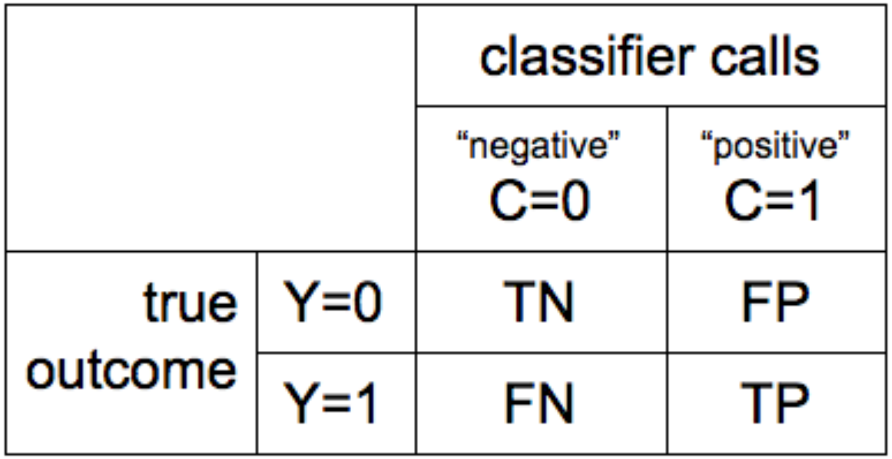
		classifier calls	
		“negative” C=0	“positive” C=1
true outcome	Y=0	TN	FP
	Y=1	FN	TP

$$\frac{TP}{TP + FP}$$

Worksheet



True Positive Rate (TPR)	$\frac{TP}{TP + FN}$	subject who is positive will be called positive
False Positive Rate (FPR)	$\frac{FP}{FP + TN}$	subject who is negative will be called positive
Positive Predictive Value (PPV)	$\frac{TP}{TP + FP}$	subject called positive will actually be positive



		classifier calls	
		"negative" C=0	"positive" C=1
true outcome	Y=0	TN	FP
	Y=1	FN	TP

Equation of the Day

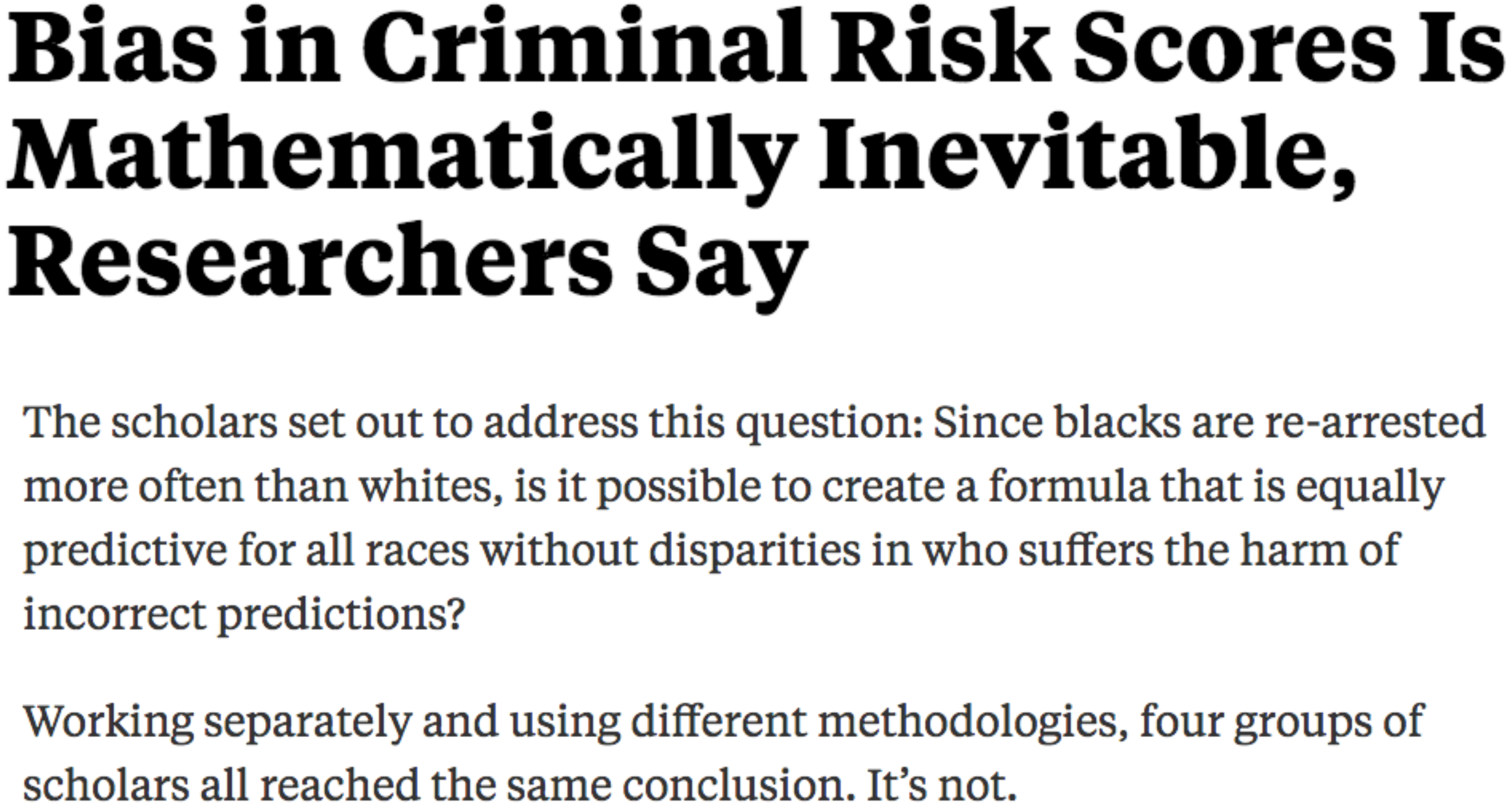
$$\mathrm{FPR} = \frac{p}{1-p} \frac{1 - \mathrm{PPV}}{\mathrm{PPV}} \mathrm{TPR}$$

where prevalence p = Pr(Y = 1)

If two groups have different p values, can we simultaneously have TPR parity AND FPR parity AND PPV parity AND NPV parity?

Unless one of these is true:
* classifier is perfect
* prevalence is same (same crime rate for A and B)

We **must** choose one over the other:
    Disparate Treatment (PPV or NPV not equal)
or Disparate Impact (FPR or TPR not equal)



Bias in Criminal Risk Scores Is Mathematically Inevitable, Researchers Say

The scholars set out to address this question: Since blacks are re-arrested more often than whites, is it possible to create a formula that is equally predictive for all races without disparities in who suffers the harm of incorrect predictions?

Working separately and using different methodologies, four groups of scholars all reached the same conclusion. It's not.

https://www.propublica.org/article/bias-in-criminal-risk-scores-is-mathematically-inevitable-researchers-say

Fairness: Unit Objectives

- How to think systematically about end-to-end ML
  - Where does data come from?
  - What features am I measuring? What protected information can leak in unintentionally?
  - Who will be impacted?
- How to define and measure notions fairness
  - Use concepts: accuracy, TPR, FPR, PPV, NPV
  - What is achievable? What is impossible?