Machine Learning for Cities
CUSP-GX 7033 B, Spring 2025

Lecture 7: Algorithmic Fairness, Discrimination, and Bias

Much of this material is re-used with permission from:
"Data-driven discrimination and fairness-aware classification" (M. Jankowiak)
"Bias and discrimination in data-driven decision making" (A. Chouldechova)
"Identifying significant predictive bias in classifiers" (Z. Zhang and D.B. Neill)

Outline of today's lecture

- **The need for fairness in algorithms: motivation and examples.**
- Preventing disparate impact: a case study in criminal justice.
- Group fairness: tweaking ML algorithms to prevent discrimination.
- Calibration: Detecting and fixing systematic biases in risk prediction.

Why should we care about fairness?

Online algorithms can exacerbate demographic and socioeconomic disparities, e.g., through price discrimination or targeted advertising.

Sensitive decisions at the individual level: school admissions, job applications, loan/credit approval, insurance premiums…

Policing: geographic and demographic biases in targeted patrolling, "stop and frisk", assumption of guilt/innocence, citation vs. warning…

Criminal justice: biases in sentencing and parole/probation decisions.

Provision of city services: resource disparities by neighborhood.

Many other quality of life factors: food deserts, poverty, environmental risk factors (e.g., pollution), access to fresh water…

Websites Vary Prices, Deals Based on Users' Information

By **JENNIFER VALENTINO-DEVRIES**, **JEREMY SINGER-VINE** and **ASHKAN SOLTANI**
December 24, 2012

It was the same Swingline stapler, on the same Staples.com website. But for Kim Wamble, the price was $15.79, while the price on Trude Frizzell's screen, just a few miles away, was $14.29.

A key difference: where Staples seemed to think they were located.

http://www.wsj.com/articles/SB10001424127887323777204578189391813881534

What happened: lower store density in poor & ethnic minority neighborhoods → higher prices → racially disparate impact.

IMAGE: PERCENTAGE OF WOMEN IN TOP 100 GOOGLE IMAGE SEARCH RESULTS FOR CEO IS: 11 PERCENT. PERCENTAGE OF US CEOS WHO ARE WOMEN IS: 27 PERCENT. view more ›

Websites Vary Prices, Deals Based on Users' Information

By JENNIFER VALENTINO-DEVRIES, JEREMY SINGER-VINE and ASHKAN SOLTANI
December 24, 2012

It was the same Swingline stapler, on the same Staples.com website. But for Kim[...] the price was $15.79, while the price on Trude Frizzell's screen, just a few miles away, was $14.29.

A key difference: where Staples seemed to think they were located.

Unforeseen consequences!

http://www.wsj.com/articles/SB10001424127887323777204578189391813881534

<u>What happened</u>: lower store density in poor & ethnic minority neighborhoods → higher prices → racially disparate impact.

Reinforcing undesirable status quo!

IMAGE: PERCENTAGE OF WOMEN IN TOP 100 GOOGLE IMAGE SEARCH RESULTS FOR CEO IS: 11 PERCENT. PERCENTAGE OF US CEOS WHO ARE WOMEN IS: 27 PERCENT. view more ›

Google accused of racism after black names are 25% more likely to bring up adverts for criminal records checks

- Professor finds 'significant discrimination' in ad results, with black names 25 per cent more likely to be linked to arrest record check services
- She compared typically black names like 'Ebony' and 'DeShawn' with typically white ones like 'Jill' and 'Geoffrey'

Ad related to **Darnell Bacon**

Darnell Bacon, Arrested?

www.instantcheckmate.com/

1) Enter Name and State. 2) Access Full Background Checks Instantly.

http://arxiv.org/abs/1301.6822

An Analysis of the New York City Police Department's "Stop-and-Frisk" Policy in the Context of Claims of Racial Bias

Andrew GELMAN, Jeffrey FAGAN, and Alex KISS

Recent studies by police departments and researchers confirm that police stop persons of racial and ethnic minority groups more often than whites relative to their proportions in the population. However, it has been argued that stop rates more accurately reflect rates of crimes committed by each ethnic group, or that stop rates reflect elevated rates in specific social areas, such as neighborhoods or precincts. Most of the research on stop rates and police–citizen interactions has focused on traffic stops, and analyses of pedestrian stops are rare. In this article we analyze data from 125,000 pedestrian stops by the New York Police Department over a 15-month period. We disaggregate stops by police precinct and compare stop rates by racial and ethnic group, controlling for previous race-specific arrest rates. We use hierarchical multilevel models to adjust for precinct-level variability, thus directly addressing the question of geographic heterogeneity that arises in the analysis of pedestrian stops. We find that persons of African and Hispanic descent were stopped more frequently than whites, even after controlling for precinct variability and race-specific estimates of crime participation.

KEY WORDS: Criminology; Hierarchical model; Multilevel model; Overdispersed Poisson regression; Police stops; Racial bias.

How can we use machine learning to identify and reduce biases?

How can we avoid introducing new biases, or exacerbating existing biases, when we perform data-driven analyses?

Big data claims to be neutral. It isn't.

Advocates of algorithmic techniques like data mining argue that they eliminate human biases from the decision-making process. But an algorithm is only as good as the data it works with. Data mining can inherit the prejudices of prior decision-makers or reflect the widespread biases that persist in society at large. Often, the "patterns" it discovers are simply preexisting societal patterns of inequality and exclusion. Unthinking reliance on data mining can deny members of vulnerable groups full participation in society. Worse still, because the resulting discrimination is almost always an unintentional emergent property of the algorithm's use rather than a conscious choice by its programmers, it can be unusually hard to identify the source of the problem or to explain it to a court.

— "Big Data's Disparate Impact," Barocas & Selbst

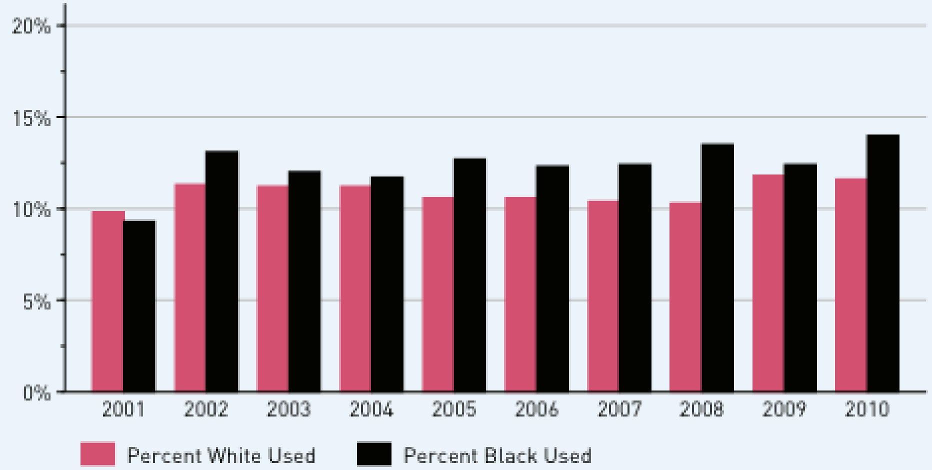
What's your Y?

- **Y** = *candidate was hired* vs. **Y** = *employee productivity*

- Target variable bias: in recidivism prediction, we:
  want **Y** = **re-offense**, have **Y** = **re-arrest**

FIGURE 21

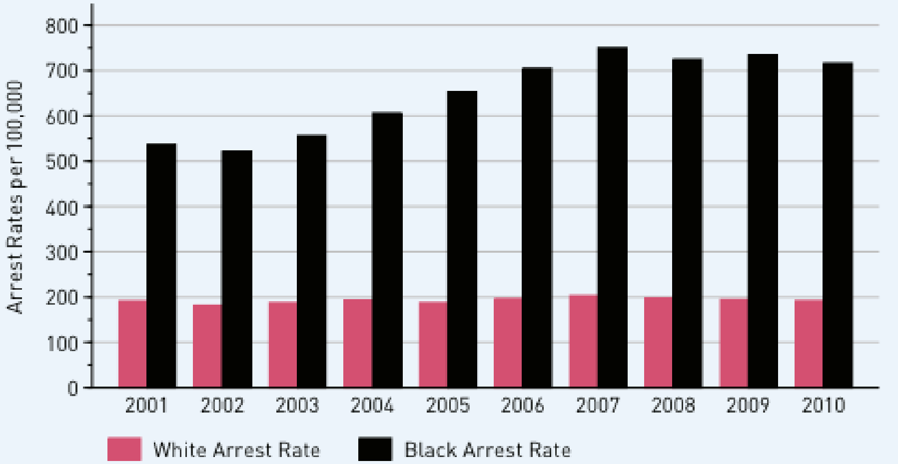
Marijuana Use by Race: Used Marijuana in Past 12 Months (2001-2010)



Source: National Household Survey on Drug Abuse and Health, 2001-2010

FIGURE 10

Arrest Rates for Marijuana Possession by Race (2001-2010)



Source: FBI/Uniform Crime Reporting Program Data and U.S. Census Data

Outline of today's lecture

- The need for fairness in algorithms: motivation and examples.
- **Preventing disparate impact: a case study in criminal justice.**
- Group fairness: tweaking ML algorithms to prevent discrimination.
- Calibration: Detecting and fixing systematic biases in risk prediction.

Material re-used with permission from:
"Bias and discrimination in data-driven decision making" (A. Chouldechova)

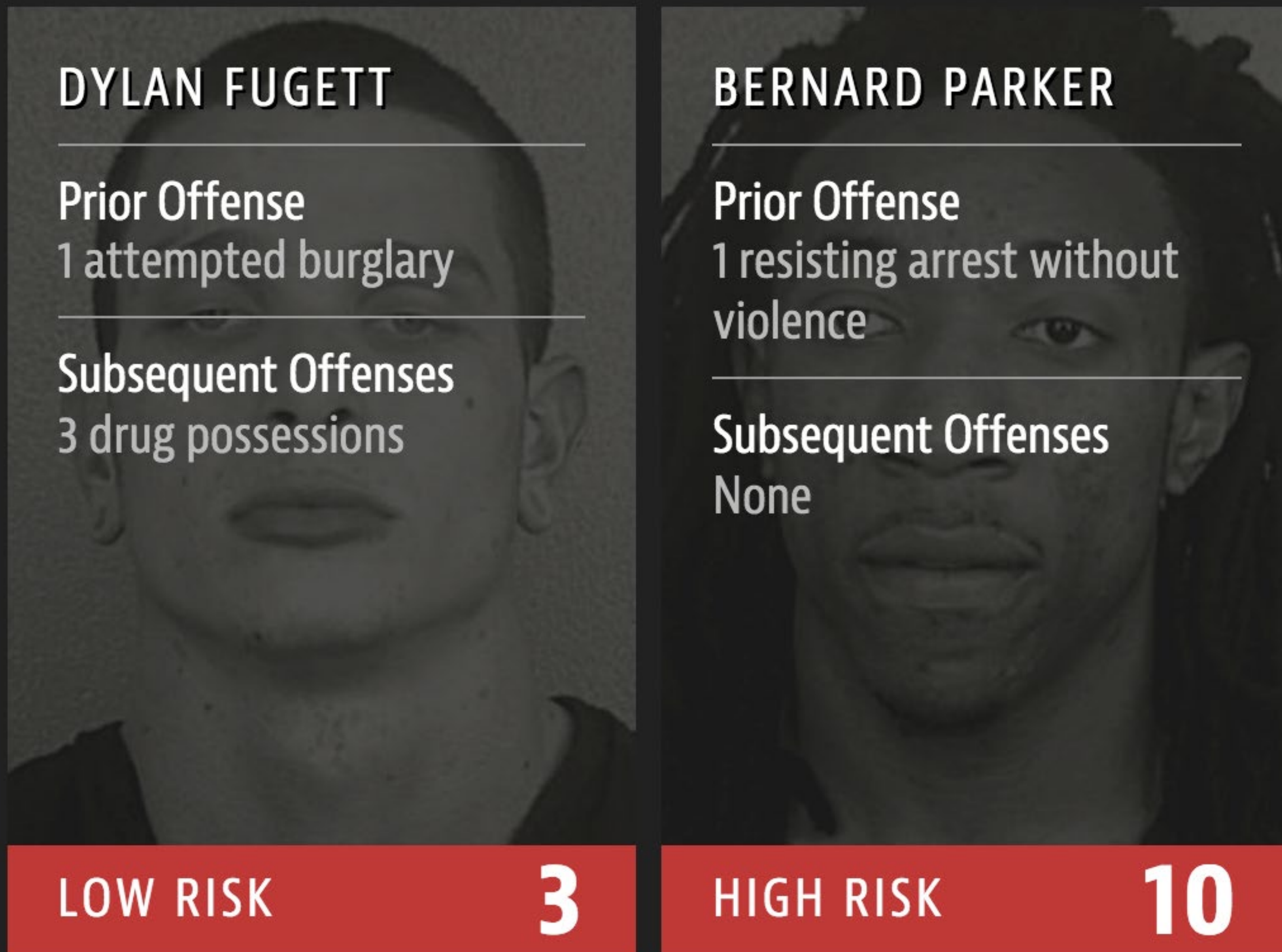
Bernard Parker, left, was rated high risk; Dylan Fugett was rated low risk. Josh Ritchie for ProPublica)

Machine Bias

There's software used across the country to predict future criminals.
And it's biased against blacks.

Source:
Julia Angwin,
Jeff Larson,
Surya Mattu and
Lauren Kirchner, *ProPublica*

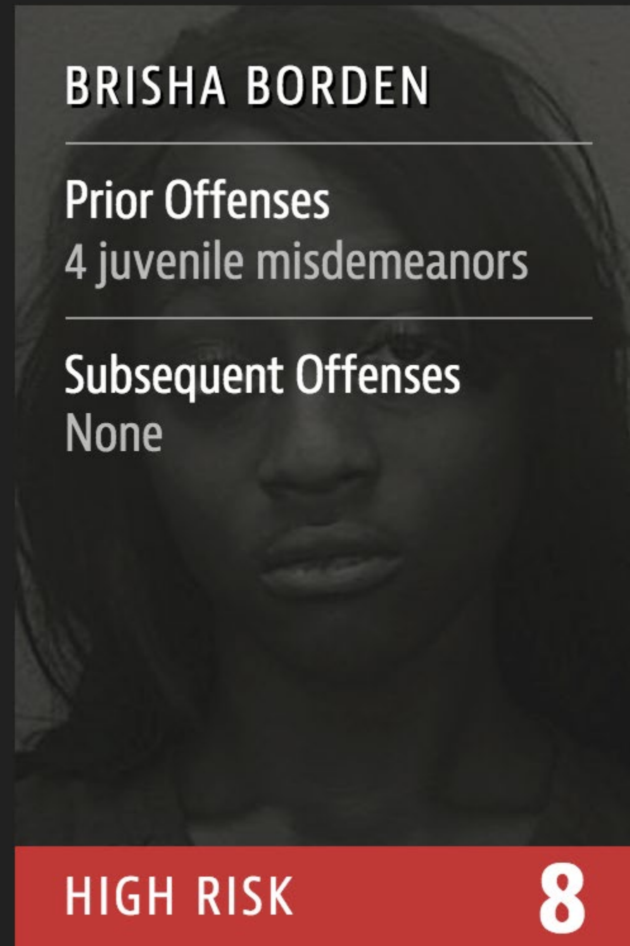
Two Drug Possession Arrests



Fugett was rated low risk after being arrested with cocaine and marijuana. He was arrested three times on drug charges after that.

Source: *ProPublica*

Two Petty Theft Arrests



Borden was rated high risk for future crime after she and a friend took a kid's bike and scooter that were sitting outside. She did not reoffend.

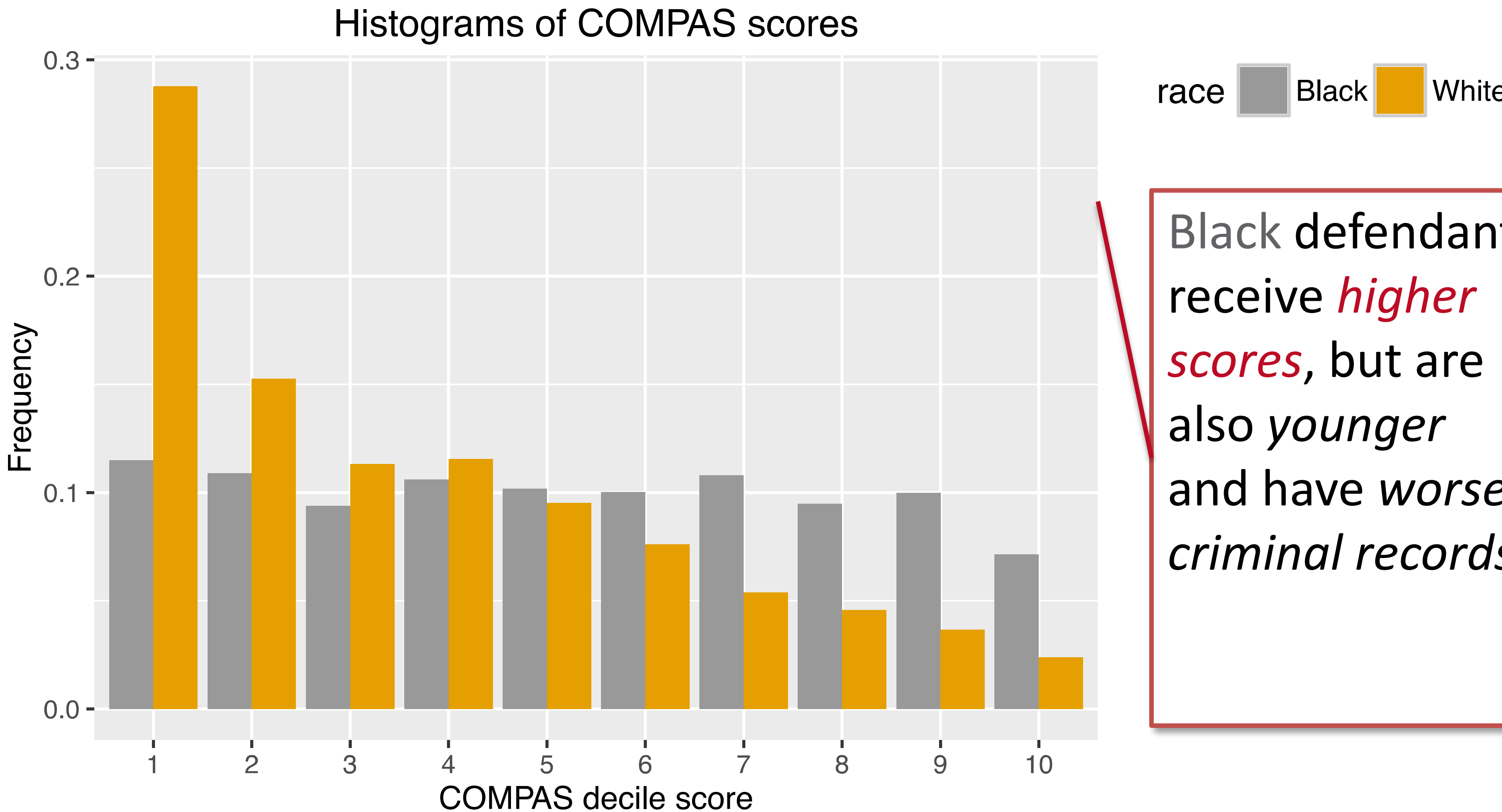
Source: *ProPublica*

Broward County data

- **Source**: ProPublica's data on criminal defendants in Broward County, Florida in 2013 – 2014, outcome assessed through April 2016
- **Score**: COMPAS score, scale 1 – 10

Background	Black $(n = 3696)$		White $(n = 2454)$
Age	32.7 (10.9)	<	37.7 (12.8)
Male (%)	82.4	>	76.9
Number of Priors	4.44 (5.58)	>	2.59 (3.8)
Any priors? (%)	76.4	>	65.9
Felony (%)	68.9	>	60.3
COMPAS Score	5.37 (2.83)	>	3.74 (2.6)

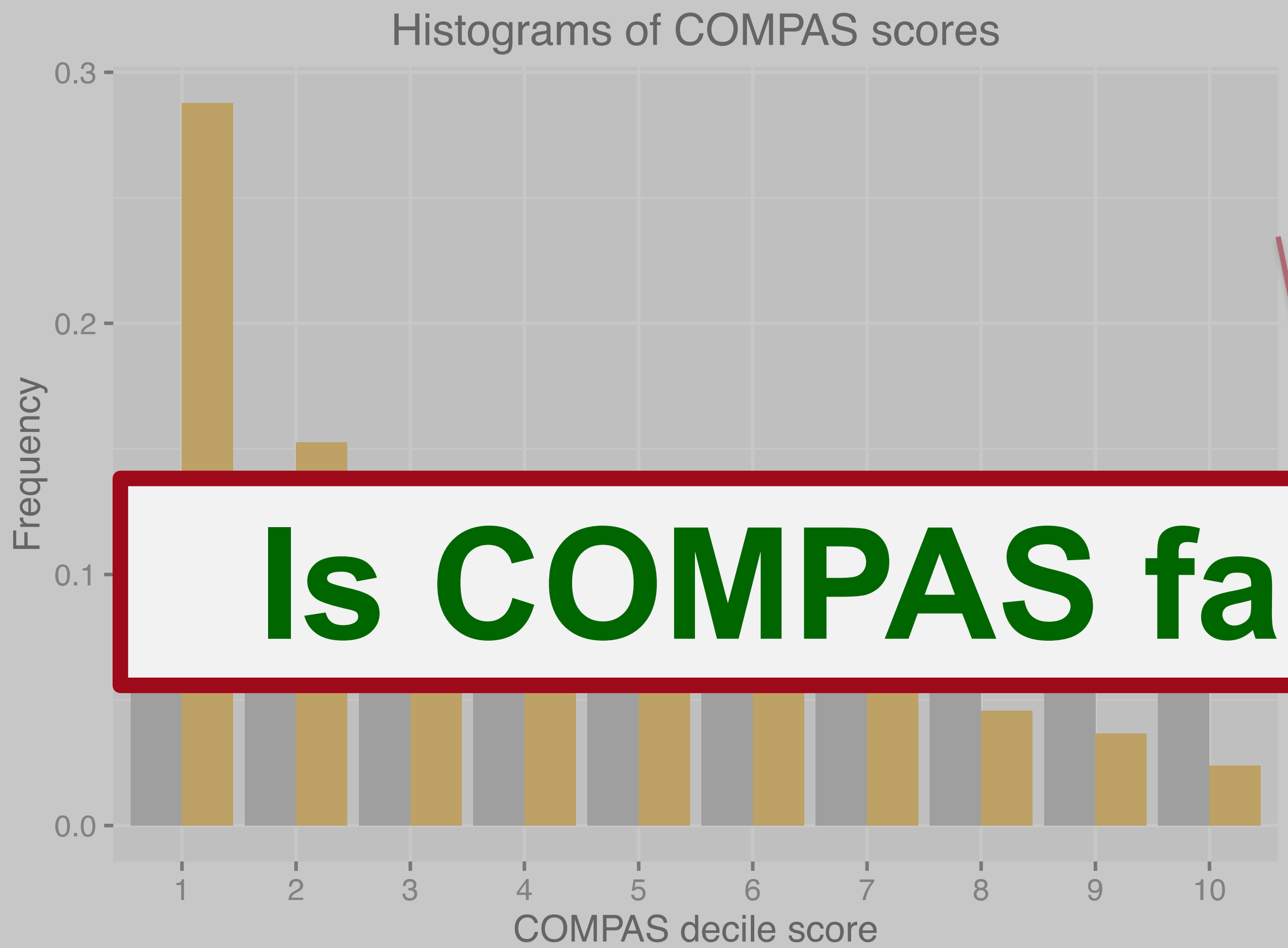
Sample averages (standard deviations)



Black defendants receive *higher scores*, but are also *younger* and have *worse criminal records*.

Outcome	Black	White
Recidivism (%)	51.4	39.4
Violent Recidivism (%)	13.40	9.05

Observed recidivism *prevalence* is *higher* among Black defendants.

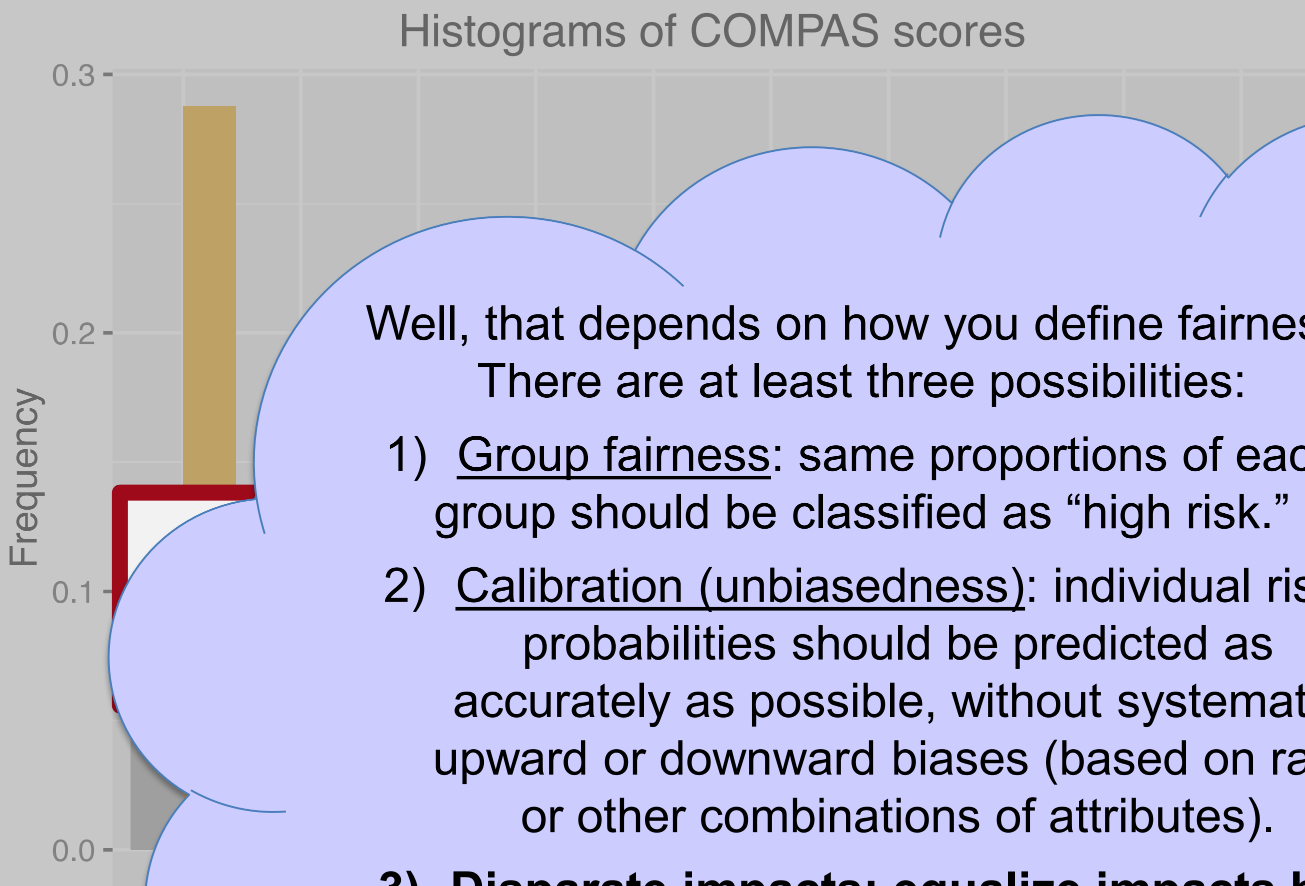


Is COMPAS fair?

Black defendants receive *higher scores*, but are ...er ...vorse ...cords.

Outcome	Black	White
Recidivism (%)	51.4	39.4
Violent Recidivism (%)	13.40	9.05

Observed recidivism *prevalence* is *higher* among Black defendants.



Well, that depends on how you define fairness!
There are at least three possibilities:

1) Group fairness: same proportions of each group should be classified as "high risk." (?)
2) Calibration (unbiasedness): individual risk probabilities should be predicted as accurately as possible, without systematic upward or downward biases (based on race or other combinations of attributes).
3) **Disparate impacts: equalize impacts by calibrating false positive and false negative rates in each group.**

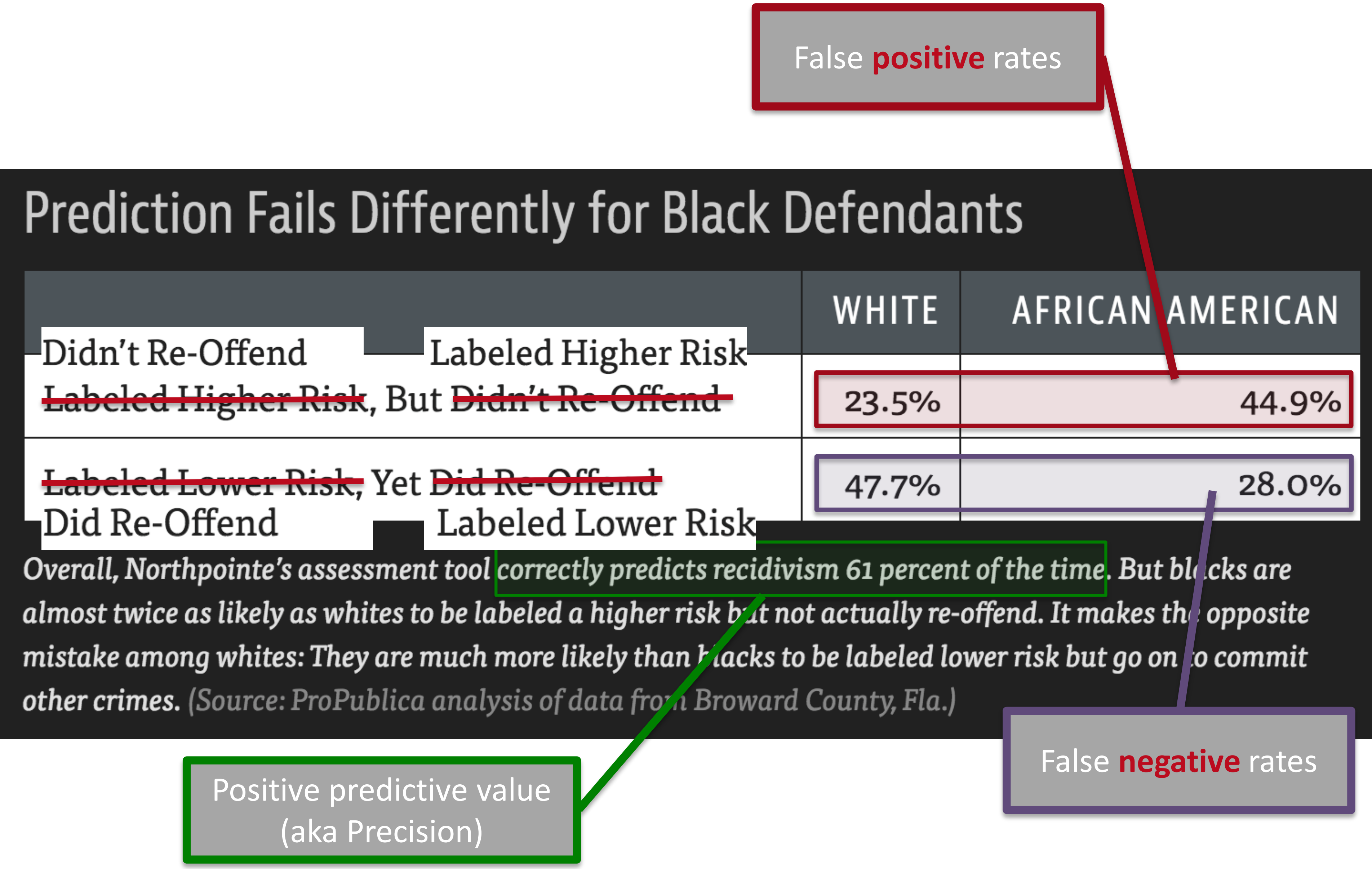
Outcome		
Recidivism (%)	51.4	
Violent Recidivism (%)	13.40	9.05

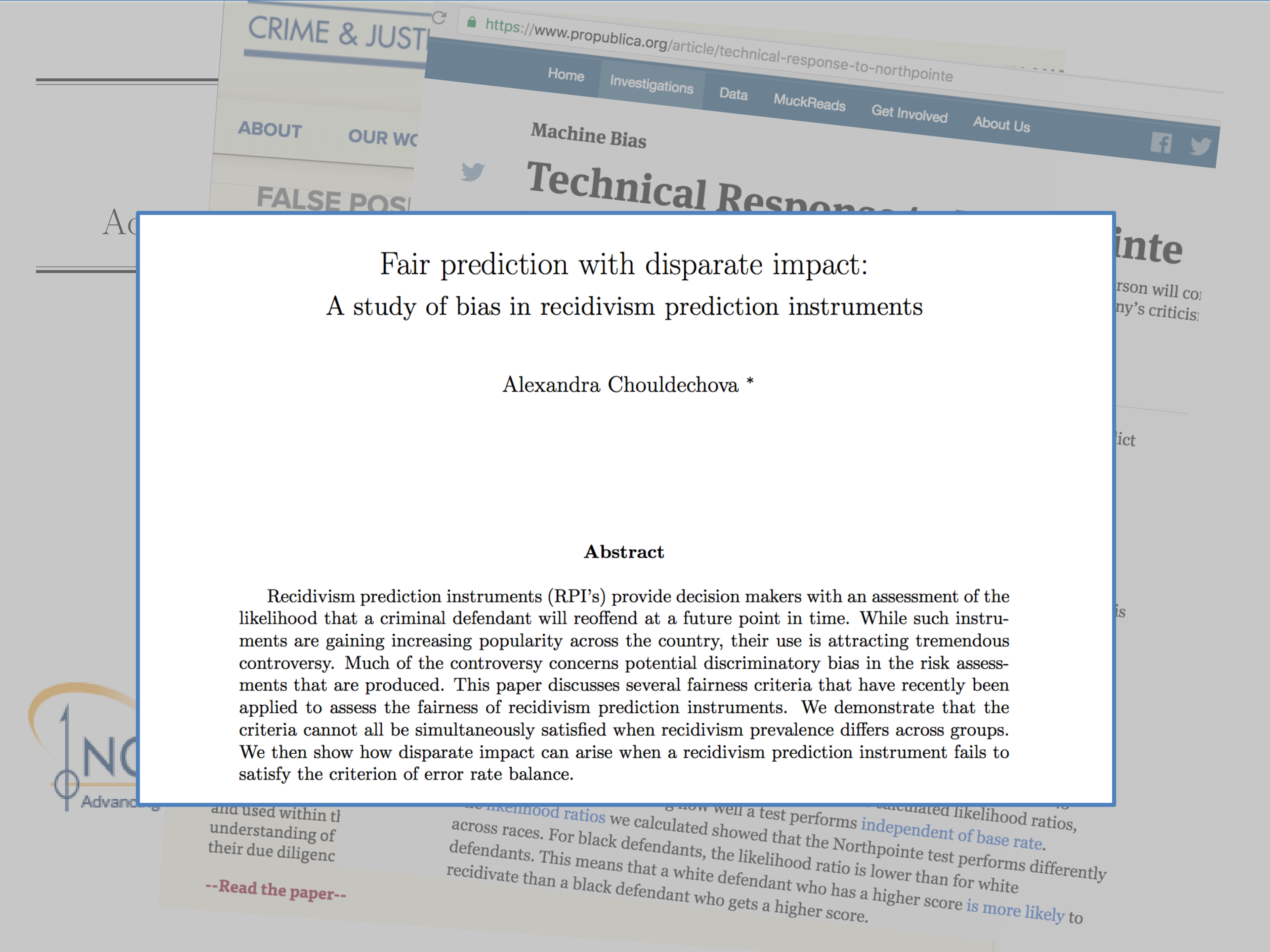
…ecidivism …prevalence is *higher* among Black defendants.

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%

Overall, Northpointe's assessment tool correctly predicts recidivism 61 percent of the time. But blacks 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. (Source: ProPublica analysis of data from Broward County, Fla.)


False positive rates
Prediction Fails Differently for Black Defendants
WHITE
AFRICAN AMERICAN
Didn't Re-Offend
Labeled Higher Risk
Labeled Higher Risk, But Didn't Re-Offend
23.5%
44.9%
Labeled Lower Risk, Yet Did Re-Offend
47.7%
28.0%
Did Re-Offend
Labeled Lower Risk
Overall, Northpointe's assessment tool correctly predicts recidivism 61 percent of the time. But blacks 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. (Source: ProPublica analysis of data from Broward County, Fla.)
False negative rates
Positive predictive value (aka Precision)



Fair prediction with disparate impact:
A study of bias in recidivism prediction instruments

Alexandra Chouldechova *

Abstract

Recidivism prediction instruments (RPI's) provide decision makers with an assessment of the likelihood that a criminal defendant will reoffend at a future point in time. While such instruments are gaining increasing popularity across the country, their use is attracting tremendous controversy. Much of the controversy concerns potential discriminatory bias in the risk assessments that are produced. This paper discusses several fairness criteria that have recently been applied to assess the fairness of recidivism prediction instruments. We demonstrate that the criteria cannot all be simultaneously satisfied when recidivism prevalence differs across groups. We then show how disparate impact can arise when a recidivism prediction instrument fails to satisfy the criterion of error rate balance.

The debate in a nutshell

ProPublica

COMPAS is biased

COMPAS has a 1.9x *higher FPR* and 1.7x *lower FNR* among Black defendants

Northpointe

COMPAS is fair

COMPAS satisfies *predictive parity**

(*has equal PPV across groups)

Among defendants classified as High Risk, 63% of Black defendants and 59% of White defendants are observed to reoffend.

i.e., COMPAS is "equally predictive of recidivism" for Black and White defendants.

It turns out:

1. When recidivism prevalence differs across groups:
   predictive parity → *or rate imbalance*
2. *Error rate imbalance* leads to disparate impact under policies that assign stricter penalties to individuals assessed as higher-risk.

Fairness metrics

- Score S. If $S > s_{HR}$, say the person is "High Risk"
- Outcome $Y = \begin{cases} 0, & \text{does not recidivate} \\ 1, & \text{recidivates} \end{cases}$
- Group membership, e.g., Race $R \in \{b, w\}$

Question

What does it mean for S to be fair with respect to R?

Typical approach

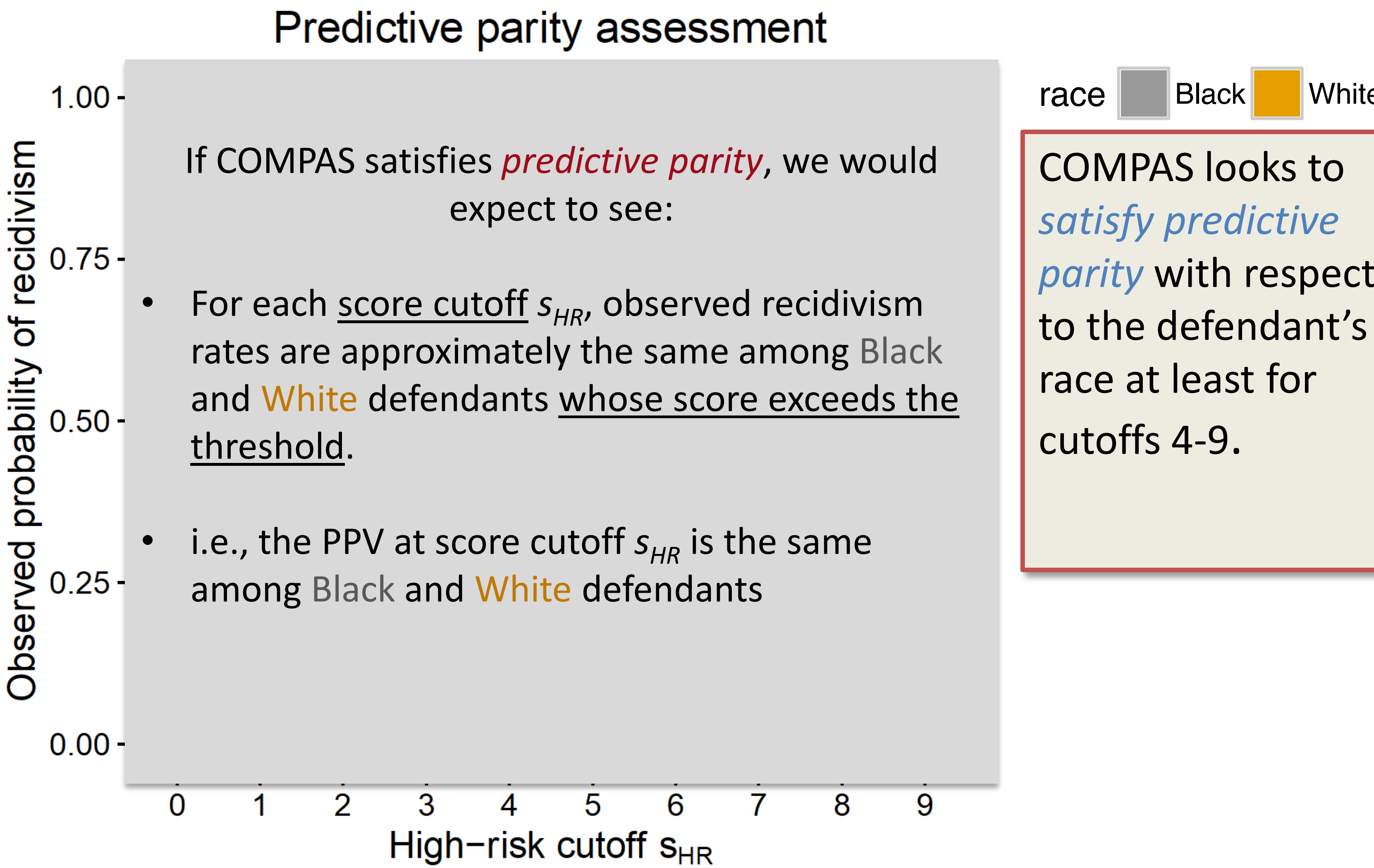
Compare various accuracy and error metrics across groups.

Building blocks: Confusion tables

- Score *S*. If $S > s_{HR}$, say the person is "High Risk"
- Outcome $Y = \begin{cases} 0, & \text{does not recidivate} \\ 1, & \text{recidivates} \end{cases}$
- Group membership, e.g., Race $R \in \{b, w\}$

Predictive parity (Northpointe's criterion)

$$\mathbb{P}(\text{reoffend} \mid \text{classified HR}, R = b) = \mathbb{P}(\text{reoffend} \mid \text{classified HR}, R = w)$$



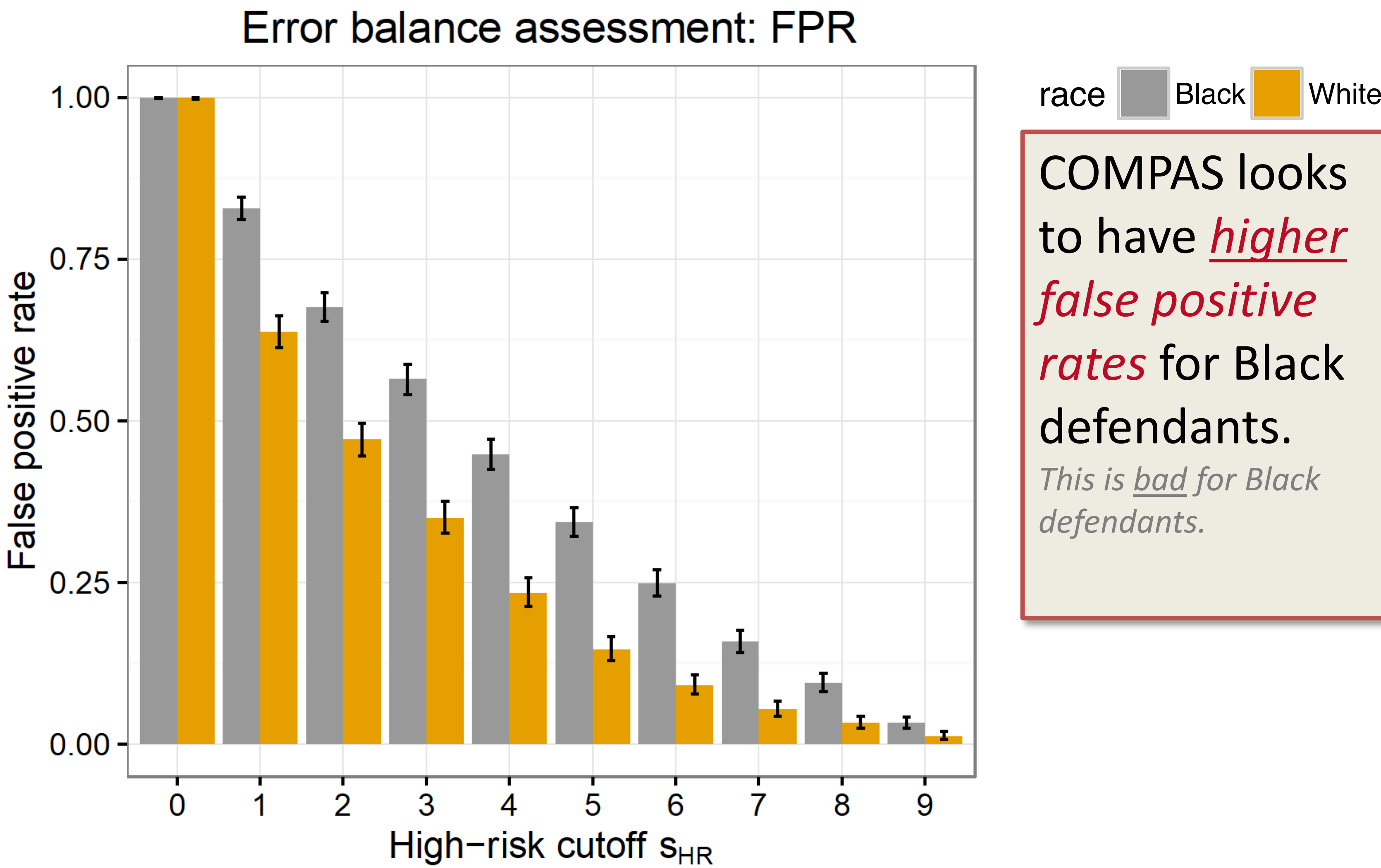
If COMPAS satisfies *predictive parity*, we would expect to see:

- For each score cutoff s_{HR}, observed recidivism rates are approximately the same among Black and White defendants whose score exceeds the threshold.
- i.e., the PPV at score cutoff s_{HR} is the same among Black and White defendants

COMPAS looks to *satisfy predictive parity* with respect to the defendant's race at least for cutoffs 4-9.

False positive rate balance (ProPublica criterion)

$$\mathbb{P}(\text{ classified HR } \mid \text{ do not reoffend }, R = b) = \mathbb{P}(\text{ classified HR } \mid \text{ do not reoffend }, R = w)$$

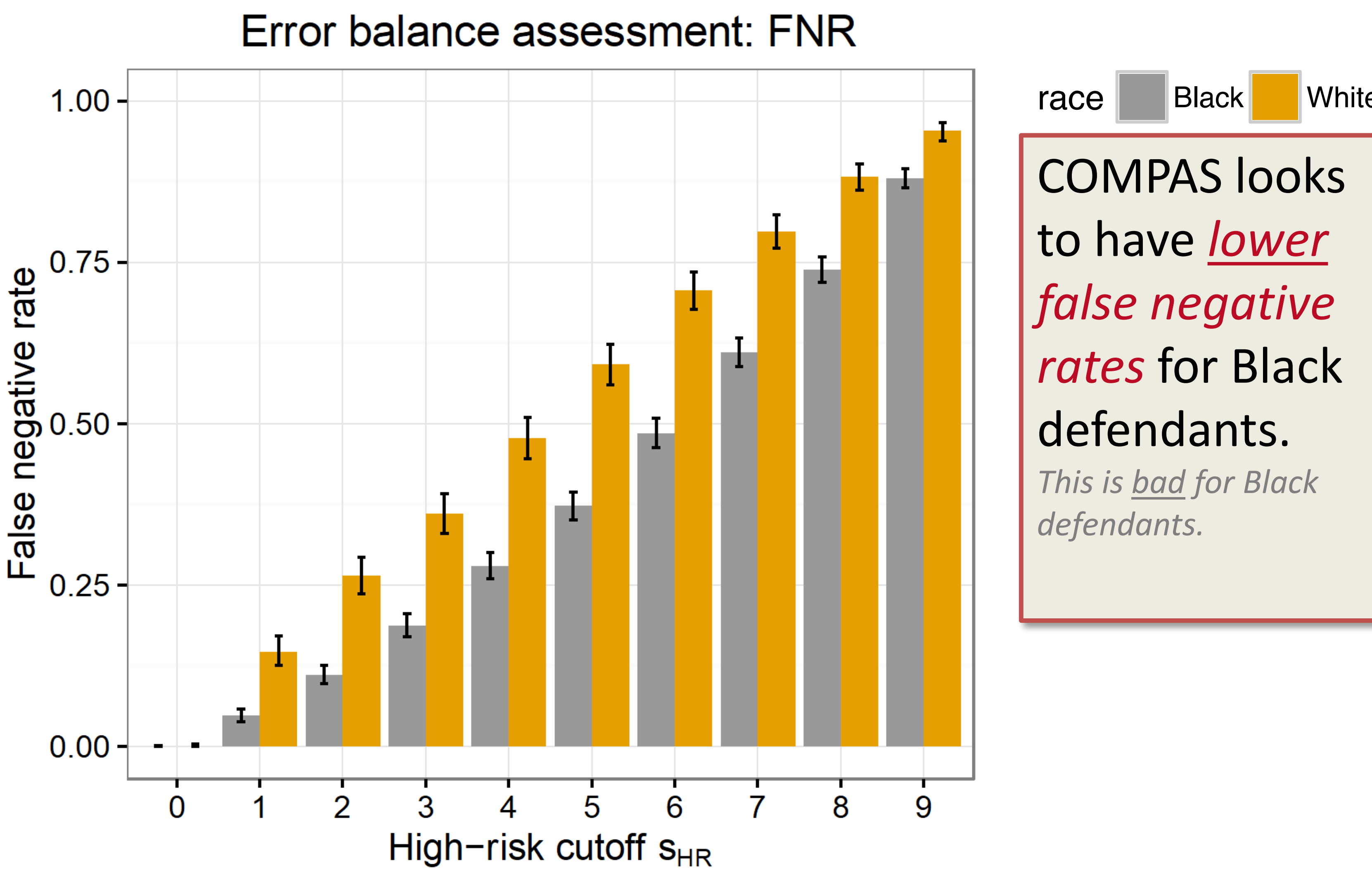


COMPAS looks to have ***higher false positive rates*** for Black defendants.

This is bad for Black defendants.

False negative rate balance (ProPublica criterion)

$$\mathbb{P}(\text{ classified LR } \mid \text{ reoffend } , R = b) = \mathbb{P}(\text{ classified LR } \mid \text{ reoffend } , R = w)$$



COMPAS looks to have *lower false negative rates* for Black defendants.

This is bad for Black defendants.

Looking back at the high-risk cutoff of $s_{HR} = 4$

Black defendants

	Low-Risk	High-Risk
Non-recid	990	805
Recid	532	1369

White defendants

	Low-Risk	High-Risk
Non-Recid	1139	349
Recid	461	505

metric	value
n	3696
prevalence	0.514
PPV	0.630
FPR	0.448
FNR	0.280

metric	value
n	2454
prevalence	0.394
PPV	0.591
FPR	0.235
FNR	0.477

predictive parity

false positive rate imbalance

false negative rate imbalance

Can we fix it?

NO WE CAN'T

Bias in Criminal Risk Scores Is Mathematically Inevitable, Researchers Say

ProPublica's analysis of bias against black defendants in criminal risk scores has prompted research showing that the disparity can be addressed — if the algorithms focus on the fairness of outcomes.

by Julia Angwin and Jeff Larson
ProPublica, Dec. 30, 2016, 4:44 p.m.

25 Comments | Print

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

This is part of an ongoing investigation

Machine Bias

We're investigating algorithmic injustice and the formulas that increasingly influence our lives.

Latest Stories in this Project

Facebook Doesn't Tell Users Everything It Really Knows About Them

Facebook Says it Will Stop Allowing Some Advertisers to Exclude Users by Race

Where Traditional DNA Testing Fails, Algorithms Take Over

Facebook Lets Advertisers Exclude Users by Race

Breaking the Black Box: How Machines Learn to Be Racist

Predictive parity implies Error rate imbalance

Predictive parity criterion requires:

The Positive Predictive Value (PPV) of S should be the same for all values of R.

Key relationship:

prevalence

$$\mathrm{FPR} = \frac{p}{1-p}\,\frac{1-\mathrm{PPV}}{\mathrm{PPV}}(1-\mathrm{FNR})$$

error rates

Takeaway Chouldechova 2016, Kleinberg *et al.* 2016

When the *prevalence* differs across groups, requiring that the PPV's be equal implies that the FNR and FPR *cannot both be equal* across those groups.

(Except in edge cases such as when PPV = 1)

Sentencing guidelines

Guidelines provide a *range of possible sentences* [t_{min}, t_{max}] based on a convicted offender's *current crime* and *criminal history.*

Pennsylvania Commission on Sentencing

§303.16. Basic Sentencing Matrix. **7th Edition (12/28/2012)**

Level	OGS	Example Offenses	Prior Record Score 0	1	2	3	4	5	RFEL	REVOC	AGG/MIT
LEVEL 3 State/ Cnty Incar RIP trade	7 (F2)	Burglary-Home/No Person Present Statutory Sexual Assault Theft (>$50,000-$100,000) Identity Theft (3rd/subq) PWID Cocaine (5-<10 g)	6-14 BC	9-16 BC	12-18 BC	15-21 BC	18-24 BC	24-30 BC	35-45 BC	NA	+/- 6
	6	Agg Assault-Cause Fear of SBI Homicide by Vehicle Burglary-Not a Home/Person Prsnt Theft (>$25,000-$50,000) Arson-Endanger Property PWID Cocaine (2<5 g)	3-12 BC	6-14 BC	9-16 BC	12-18 BC	15-21 BC	21-27 BC	27-40 BC	NA	+/- 6
LEVEL 2	5 (F3)	Burglary F2 Theft (>$2000-$25,000) Bribery PWID Marij (1-<10 lbs)	RS-9	1-12 BC	3-14 BC	6-16 BC	9-16 BC	12-18 BC	24-36 BC	NA	+/- 3
		Indecent Assault M2									

Broward County data is insufficient to calculate prior record scores for our cohort, so we'll assume a prior score of 1 for the empirical analysis.

We'll consider two *risk-based sentencing policies*:

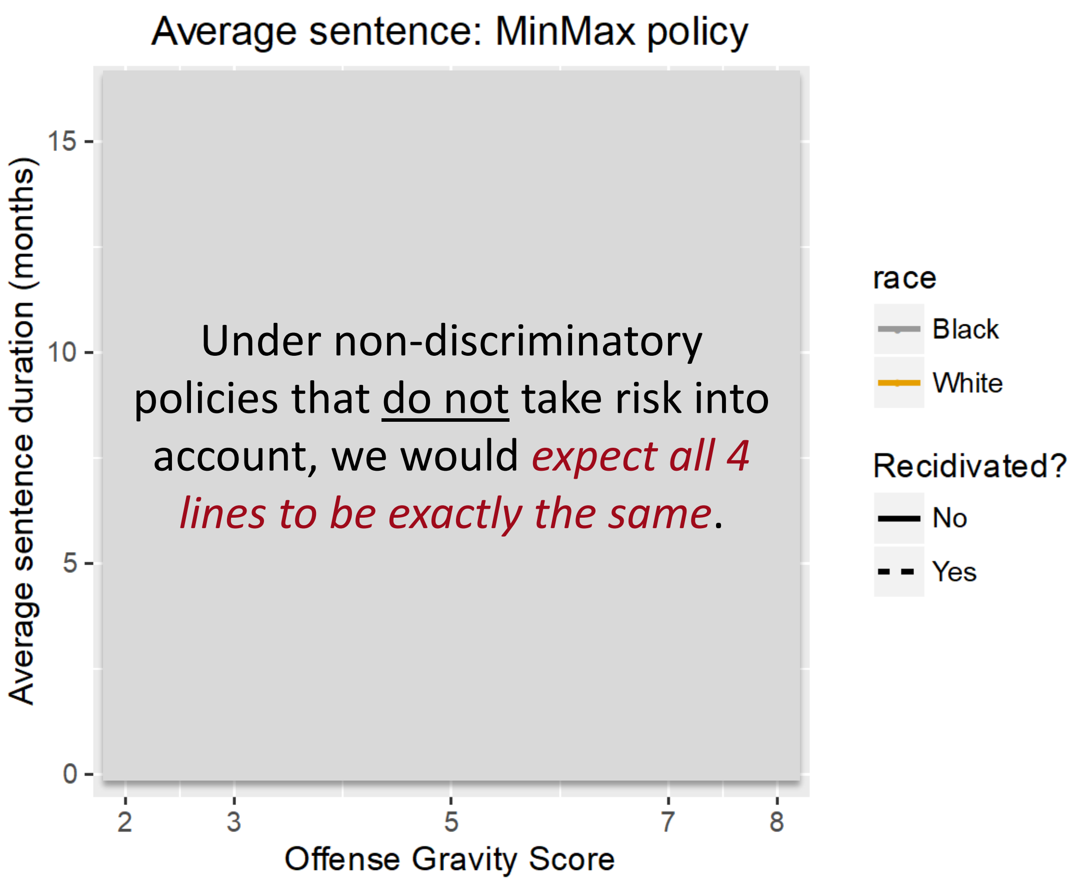
MinMax

$$\text{sentence}_{\text{MM}} = \begin{cases} t_{\min} & \text{if defendant is Low-risk} \\ t_{\max} & \text{if defendant is High-risk} \end{cases}$$

Cnty Incar RIP trade	6	Agg Assault-Cause Fear of SBI Homicide by Vehicle Burglary-Not a Home/Person Prsnt Theft (>$25,000-$50,000) Arson-Endanger Property PWID Cocaine (2<5 g)	3-12 BC	6-14 BC	9-16 BC	12-18 BC	15-21 BC	21-27 BC	27-40 BC	NA	+/- 6

Interpolation

$$\text{sentence}_{\text{INT}} = t_{\min} + \frac{s-1}{9}(t_{\max} - t_{\min})$$



OGS	2	3	5	7	8
Offense Degree	(M2)	(M1)	(F3)	(F2)	(F1)

* Except at OGS level 8, observed differences in average sentences between Black and White defendants in both the recidivating and non-recidivating groups are statistically significant at the 0.01 level.

- *Recidivists* would receive longer sentences than *non-recidivists*
- Black defendants would receive significantly *longer sentences* compared to White defendants.
- Among *non-recidivists*, this is due to the *higher* FPR among Black defendants.
- Among *recidivists*, this is due to the *higher* FNR among White defendants.

Can we mitigate disparate impact?

- **Yes. Two possible approaches:**
  a) Re-build scoring model to maximize accuracy subject to *error rate balance constraints* (see e.g., Zafar et al. (2016))
  b) Rebalance FNR and FPR by using *different score thresholds* across groups (see e.g., Hardt, Price, Srebro (2016))

- Let's try approach (b):
  - Use a COMPAS score cutoff of **6** for Black defendants, while keeping a cutoff of **4** for White defendants.

Allowing group-specific cutoffs

Before: Cutoff = 4 for both groups

Black defendants: metric	value	White defendants: metric	value
n	3696	n	2454
prevalence	0.51	prevalence	0.39
PPV	0.63	PPV	0.59
FPR	0.45	FPR	0.24
FNR	0.28	FNR	0.48

imbalanced!

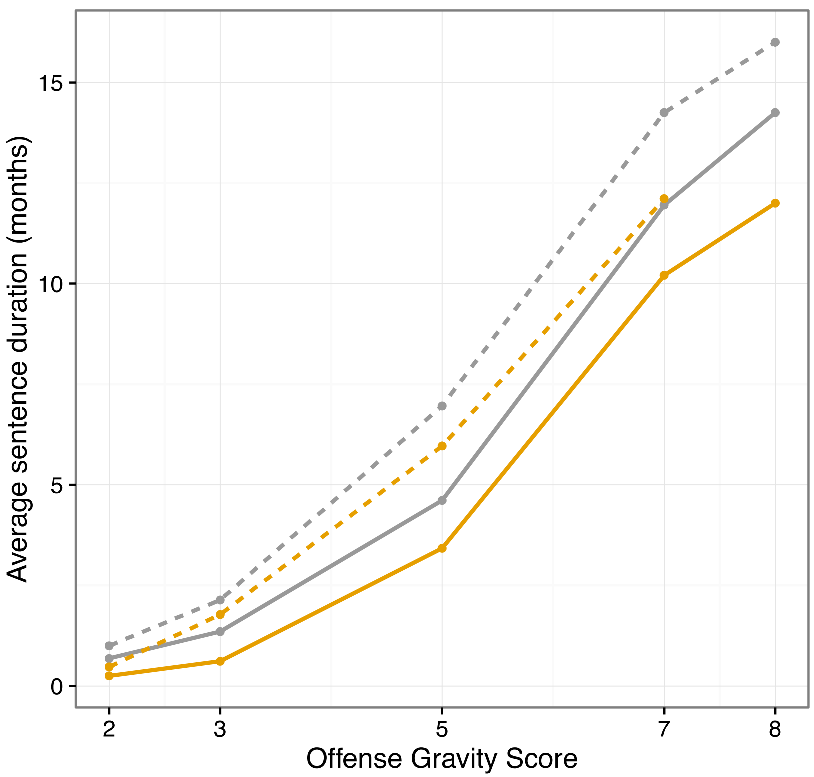
After: Cutoff = 6 for Black def.'s Cutoff = 4 for White def.

Black defendants: metric	value	White defendants: metric	value
n	3696	n	2454
prevalence	0.51	prevalence	0.39
PPV	0.69	PPV	0.59
FPR	0.25	FPR	0.24
FNR	0.49	FNR	0.48

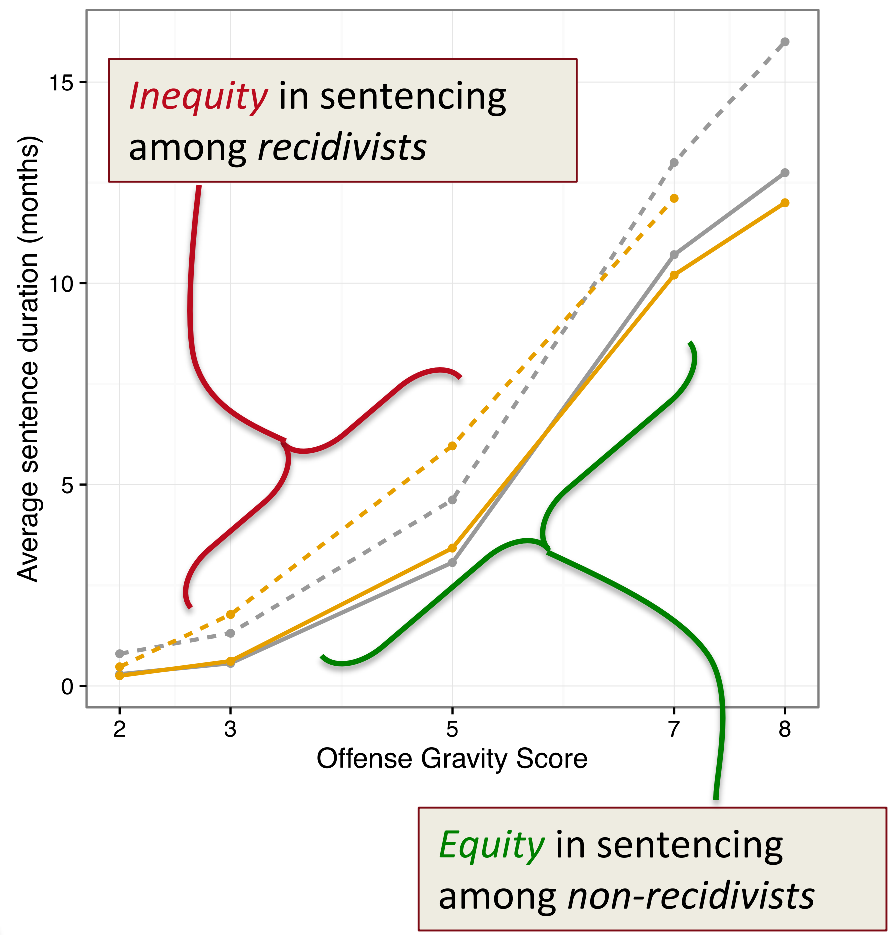
balanced!

Did we succeed?

MinMax Sentencing, *Before*



MinMax Sentencing, *After cutoff change*



race — Black — White

Recidivated? — No - - Yes

Takeaway

Balancing *overall* error rates is *insufficient*. Balance must be achieved at *sufficiently fine levels of granularity*.

Outline of today's lecture

- The need for fairness in algorithms: motivation and examples.
- Preventing disparate impact: a case study in criminal justice.
- **Group fairness: tweaking ML algorithms to prevent discrimination.**
- Calibration: Detecting and fixing systematic biases in risk prediction.

Statistical Parity

or "group fairness": an entirely different notion of fairness

$$\mathbb{P}(\textbf{hired} \mid \textbf{man}) = \mathbb{P}(\textbf{hired} \mid \textbf{woman})$$

- Reasonable fairness criterion in settings such as **employment**
  - *Not reasonable in risk assessment*

- Lots of recent work on constructing models that satisfy statistical parity

- **Caution:**
  - Statistical parity shouldn't be an end in itself
  - E.g., Could just hire top 10% of men and a random 10% of women
    - *Self-fulfilling prophecy* of discrimination (Dwork et. al.)

Discrimination

"Discrimination refers to an unjustified distinction of individuals based on their membership, or perceived membership, in a certain group or category. Justified distinctions are exceptions explicitly admitted by law, such as imposing a minimum age for voting in elections, or that are proven (sometimes in court) as being objective and legitimate, such as requiring a man for a male character in a film. Some groups, traditionally subject to discrimination, are explicitly listed as 'protected groups' by national and international human rights laws."

from "A multidisciplinary survey on discrimination analysis"

How algorithms can discriminate

- Problem specification (definition of target variable, features, etc.)

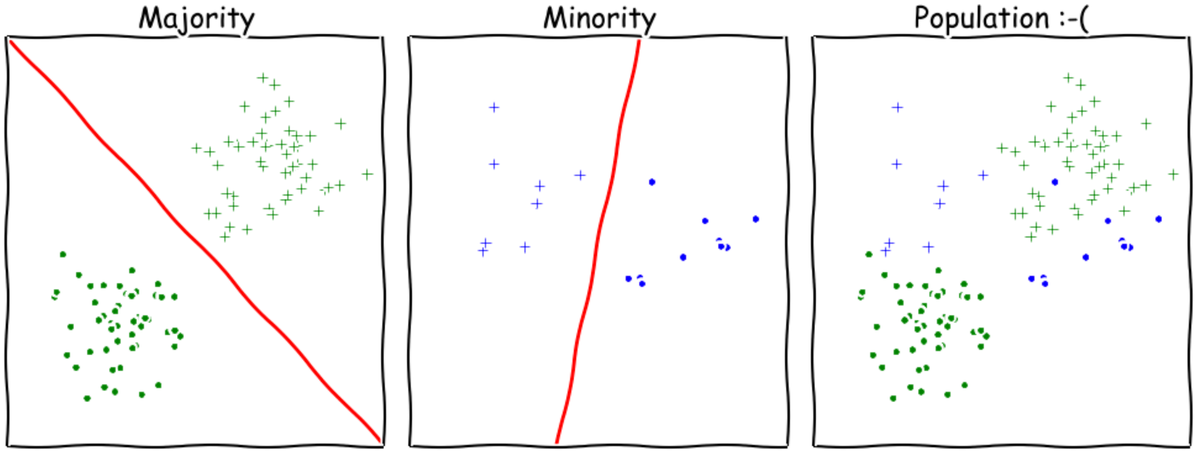
e.g. electing to hire on the basis of predicted tenure can be more likely to have a disparate impact on certain protected classes than hiring decisions that turn on some estimate of worker productivity

How algorithms can discriminate

- Problems with training data:

— the data generating process itself was inherently discriminatory

— biased/imbalanced data samples



How algorithms can discriminate

via proxies

“when the criteria that are genuinely relevant in making rational and well-informed decisions also happen to serve as reliable proxies for class membership. In other words, the very same criteria that correctly sort individuals according to their predicted likelihood of excelling at a job may also sort individuals according to class membership”

How algorithms can discriminate

via proxies

hidden attribute

Customer no.	Gender	Age	Hp	Driving style	Risk
#1	Male	30 years	High	Aggressive	+
#2	Male	35 years	Low	Aggressive	-
#3	Female	24 years	Med.	Calm	-
#4	Female	18 years	Med.	Aggressive	+
#5	Male	65 years	High	Calm	-
#6	Male	54 years	Low	Aggressive	+
#7	Female	21 years	Low	Calm	-
#8	Female	29 years	Med.	Calm	-

from "Why Unbiased Computational Processes Can Lead to Discriminative Decision Procedures," Calders & Žliobaitė

Simple fixes that don't work

removing the sensitive attribute

Customer no.	Ethnicity	Work exp.	Postal code	Loan decision
#1	European	12 years	1212	+
#2	Asian	2 years	1010	-
#3	European	5 years	1221	+
#4	Asian	10 years	1011	-
#5	European	10 years	1200	+
#6	Asian	5 years	1001	-
#7	European	12 years	1212	+
#8	Asian	2 years	1010	-

from "Why Unbiased Computational Processes Can Lead to Discriminative Decision Procedures," Calders & Žliobaitė

Simple fixes that don't work

Building separate models for each value of the sensitive attribute

Applicant no.	Gender	Test score	Level	Acceptance
#1	Male	82	A	+
#2	Female	85	A	+
#3	Male	75	B	+
#4	Female	75	B	-
#5	Male	65	A	-
#6	Female	62	A	-
#7	Male	91	B	+
#8	Female	81	B	+

from "Why Unbiased Computational Processes Can Lead to Discriminative Decision Procedures," Calders & Žliobaitė

Adverse affect and the 80% rule

Adverse effect refers to a total employment process which results in a significantly higher percentage of a protected group in the candidate population being rejected for employment, placement, or promotion. The difference between the rejection rates for a protected group and the remaining group must be statistically significant at the .05 level. In addition, if the acceptance rate of the protected group is greater than or equal to 80% of the acceptance rate of the remaining group, then adverse effect is said to be not present by definition (Section 7.1).

from Biddle

Adverse affect and the 80% rule

Definition 1.1 (Disparate Impact ("80% rule")). *Given data set $D = (X, Y, C)$, with* protected attribute *X (e.g., race, sex, religion, etc.), remaining attributes Y, and binary class to be predicted C (e.g., "will hire"), we will say that D has* disparate impact *if*

$$\frac{\Pr(C = YES|X = 0)}{\Pr(C = YES|X = 1)} \leq \tau = 0.8$$

for positive outcome class YES and majority protected attribute 1 where $\Pr(C = c|X = x)$ denotes the conditional probability (evaluated over D) that the class outcome is $c \in C$ given protected attribute $x \in X$.[1]

from "Certifying and removing disparate impact",

Adverse affect and the 80% rule

consider a classifier defined by a **decision boundary;**

to satisfy the 80% rule our classifier must satisfy

$$\frac{P(d_\theta(\mathbf{x}) > 0 | X = 0)}{P(d_\theta(\mathbf{x}) > 0 | X = 1)} \geq 0.80$$

unfortunately this isn't a very well-behaved objective function
(as a function of the parameters theta)

One quantitative learning approach to fairness: decision boundary covariance

decision boundary

sensitive class

$$
\begin{aligned}
\mathrm{Cov}(\mathbf{z}, d_{\boldsymbol{\theta}}(\mathbf{x})) &= \mathbb{E}[(\mathbf{z}-\bar{\mathbf{z}})d_{\boldsymbol{\theta}}(\mathbf{x})] - \mathbb{E}[(\mathbf{z}-\bar{\mathbf{z}})]\bar{d}_{\boldsymbol{\theta}}(\mathbf{x}) \\
&= \mathbb{E}[(\mathbf{z}-\bar{\mathbf{z}})d_{\boldsymbol{\theta}}(\mathbf{x})] \\
&\approx \frac{1}{N}\sum_{i=1}^{N}(\mathbf{z}_i-\bar{\mathbf{z}})\, d_{\boldsymbol{\theta}}(\mathbf{x}_i),
\end{aligned}
$$

"Learning Fair Classifiers." Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez-Rodriguez, Krishna P. Gummadi

http://arxiv.org/pdf/1507.05259v3.pdf

Decision boundary covariance

$$\frac{1}{N}\sum_{i=1}^{N}(\mathbf{z}_i-\bar{\mathbf{z}})\,\boldsymbol{\theta}^T\mathbf{x}_i.$$

this will serve as our measure of unfairness

making this small does NOT guarantee that the 80% rule will be satisfied; but as we will see, in practice a small value of the covariance will typically lead to a balanced ratio of

$$\frac{P(d_\theta(\mathbf{x})>0|X=0)}{P(d_\theta(\mathbf{x})>0|X=1)}$$

Maximizing accuracy under fairness constraints

$$p(y_i = 1 | \mathbf{x}_i, \boldsymbol{\theta}) = \frac{1}{1 + e^{\boldsymbol{\theta}^T \mathbf{x}_i}}$$

our constrained optimization problem with fairness constraints:

$$\begin{array}{ll} \text{minimize} & -\sum_{i=1}^{N} \log p(y_i | \mathbf{x}_i, \boldsymbol{\theta}) \\ \text{subject to} & \frac{1}{N} \sum_{i=1}^{N} (\mathbf{z}_i - \bar{\mathbf{z}}) \, \boldsymbol{\theta}^T \mathbf{x}_i \leq \mathbf{c}, \\ & \frac{1}{N} \sum_{i=1}^{N} (\mathbf{z}_i - \bar{\mathbf{z}}) \, \boldsymbol{\theta}^T \mathbf{x}_i \geq -\mathbf{c}. \end{array}$$

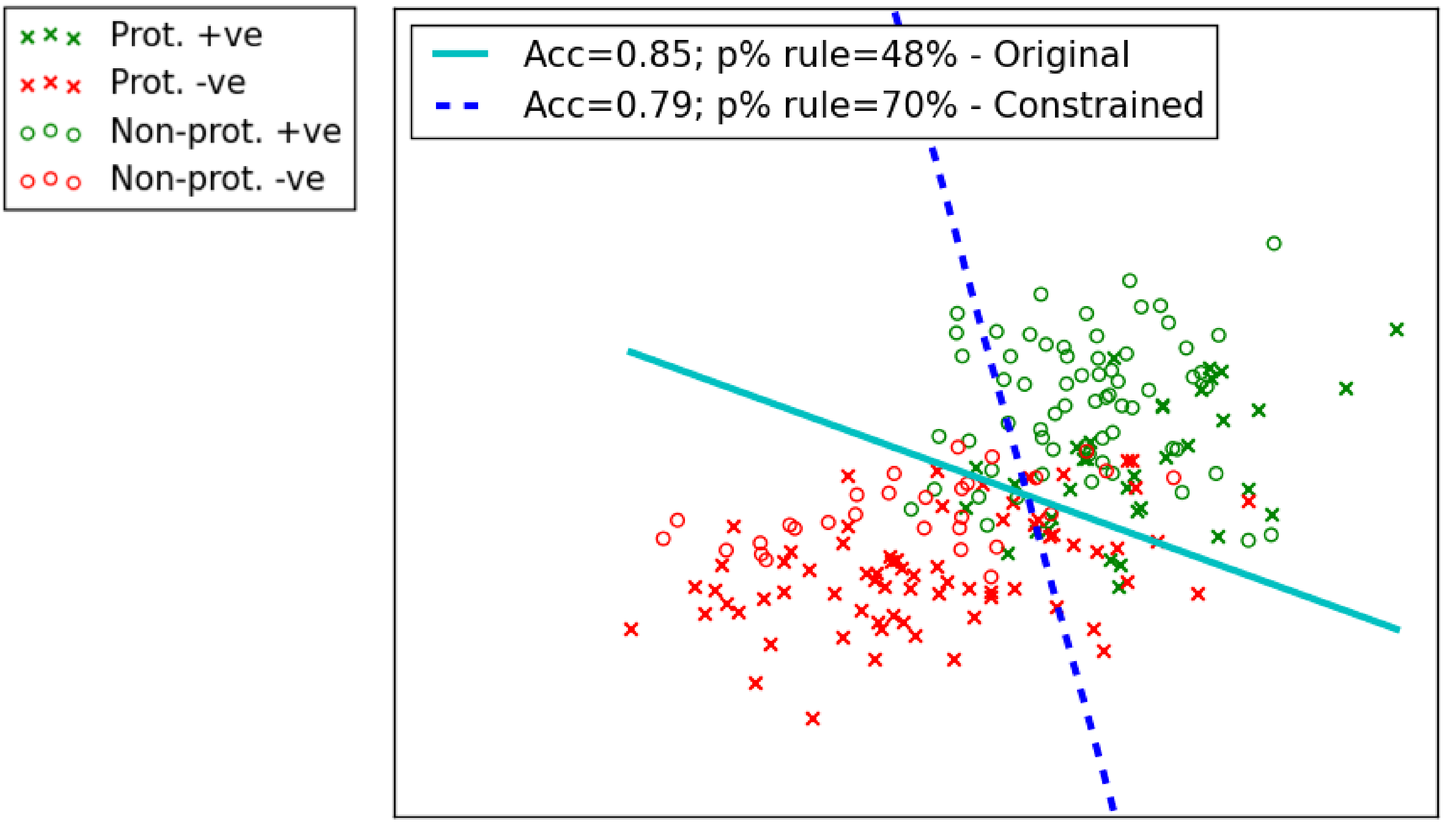
logistic regression case

Maximizing accuracy under fairness constraints

$$
\begin{array}{ll}
\text{minimize} & \|\boldsymbol{\theta}\|^2 + C\sum_{i=1}^{n}\xi_i \\
\text{subject to} & y_i\boldsymbol{\theta}^T\mathbf{x}_i \geq 1-\xi_i, \forall i \in \{1,\ldots,n\} \\
& \xi_i \geq 0, \forall i \in \{1,\ldots,n\}, \\
& \frac{1}{N}\sum_{i=1}^{N}\left(\mathbf{z}_i - \bar{\mathbf{z}}\right)\boldsymbol{\theta}^T\mathbf{x}_i \leq \mathbf{c}, \\
& \frac{1}{N}\sum_{i=1}^{N}\left(\mathbf{z}_i - \bar{\mathbf{z}}\right)\boldsymbol{\theta}^T\mathbf{x}_i \geq -\mathbf{c},
\end{array}
$$

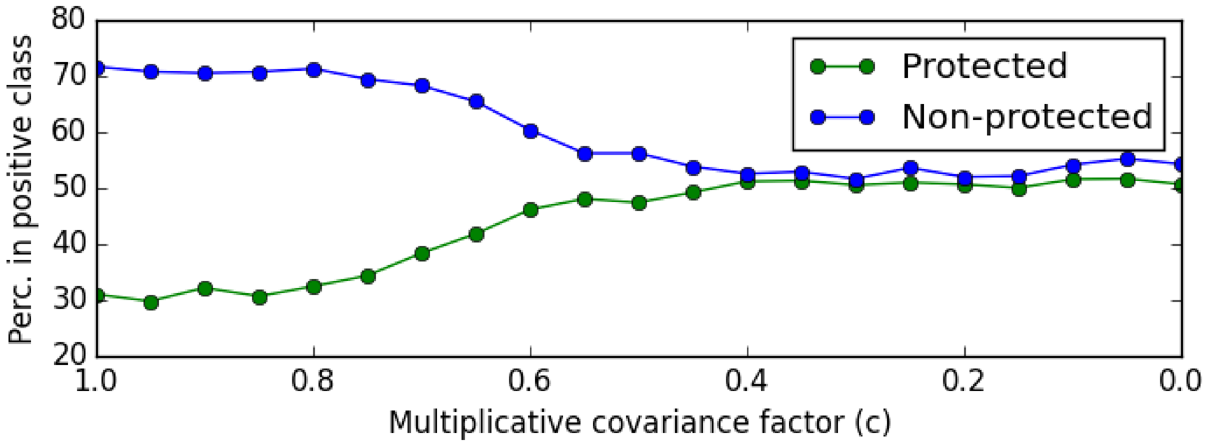
linear SVM case

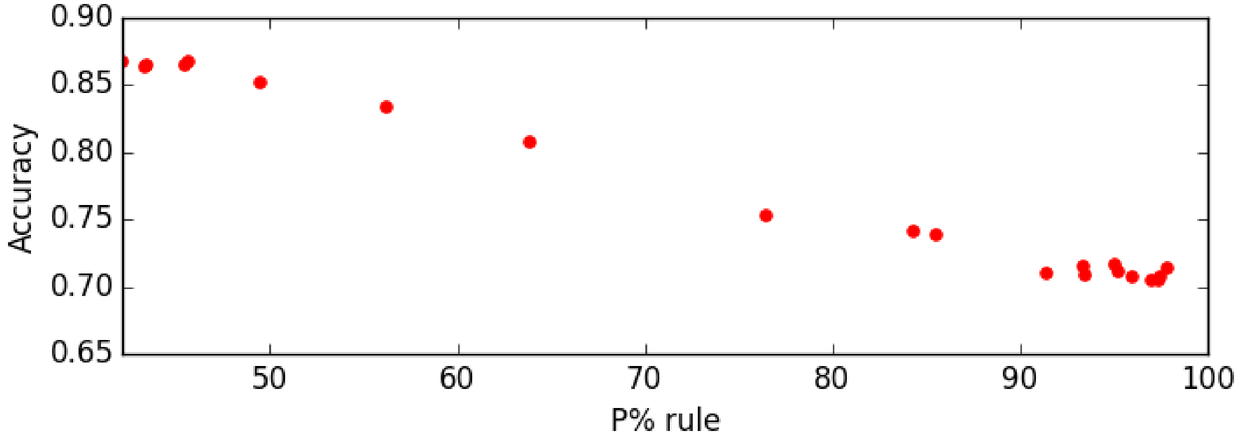
Maximizing accuracy under fairness constraints



https://github.com/mbilalzafar/fair-classification

Maximizing accuracy under fairness constraints





Outline of today's lecture

- The need for fairness in algorithms: motivation and examples.
- Preventing disparate impact: a case study in criminal justice.
- Group fairness: tweaking ML algorithms to prevent discrimination.
- **Calibration: Detecting and fixing systematic biases in risk prediction.**

Another perspective (mine…)

Whether our goal is to achieve group fairness or reduce disparate impacts, our first step should be to predict risk as accurately as possible.

In particular, we wish to **detect** and **correct** any systematic **biases** in risk prediction that a classifier may have (i.e., over-predicting or under-predicting risk for a specific attribute or combination of attributes).

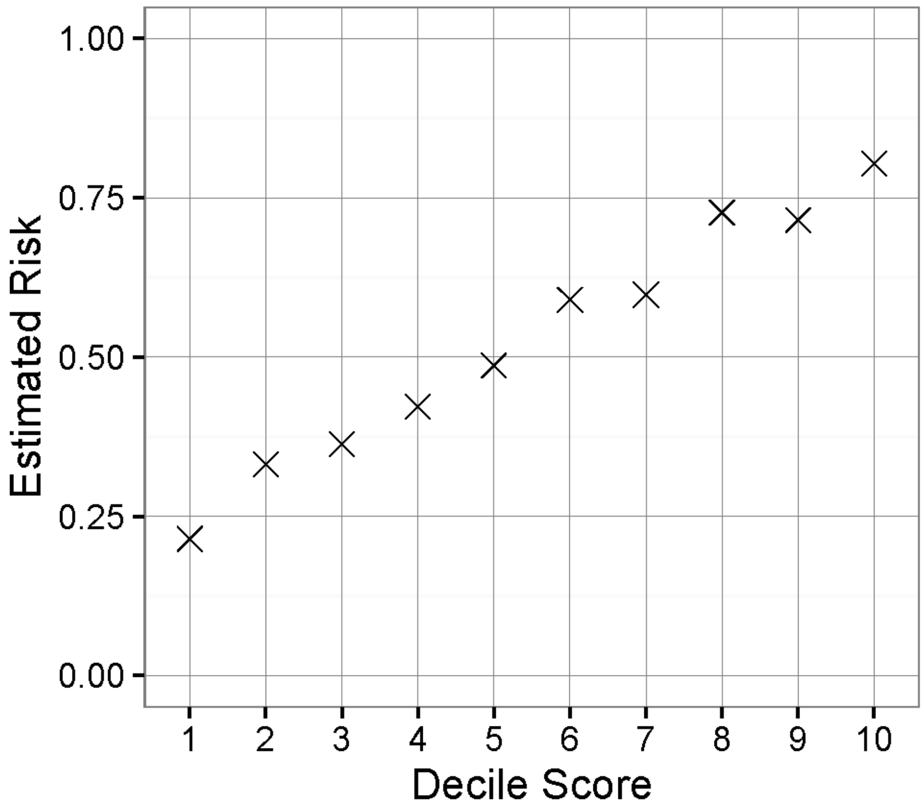
Thus we developed a new **subset scan** method to identify subgroups where classifier predictions are significantly biased (Zhang & Neill, 2016).

Assume a dataset with inputs x_i, binary labels $y_i \in \{0,1\}$, and the classifier's risk predictions $\widehat{p_i} = \Pr(y_i = 1)$.

Search space: subspaces defined by a subset of values for each attribute (e.g., "white and Asian males under 25")

Score function: a log-likelihood ratio statistic. H_0: $\widehat{p_i}$ correctly calibrated; $H_1(S)$: constant multiplicative increase or decrease in odds of $y_i = 1$ for subspace S.

Results of bias scan on COMPAS

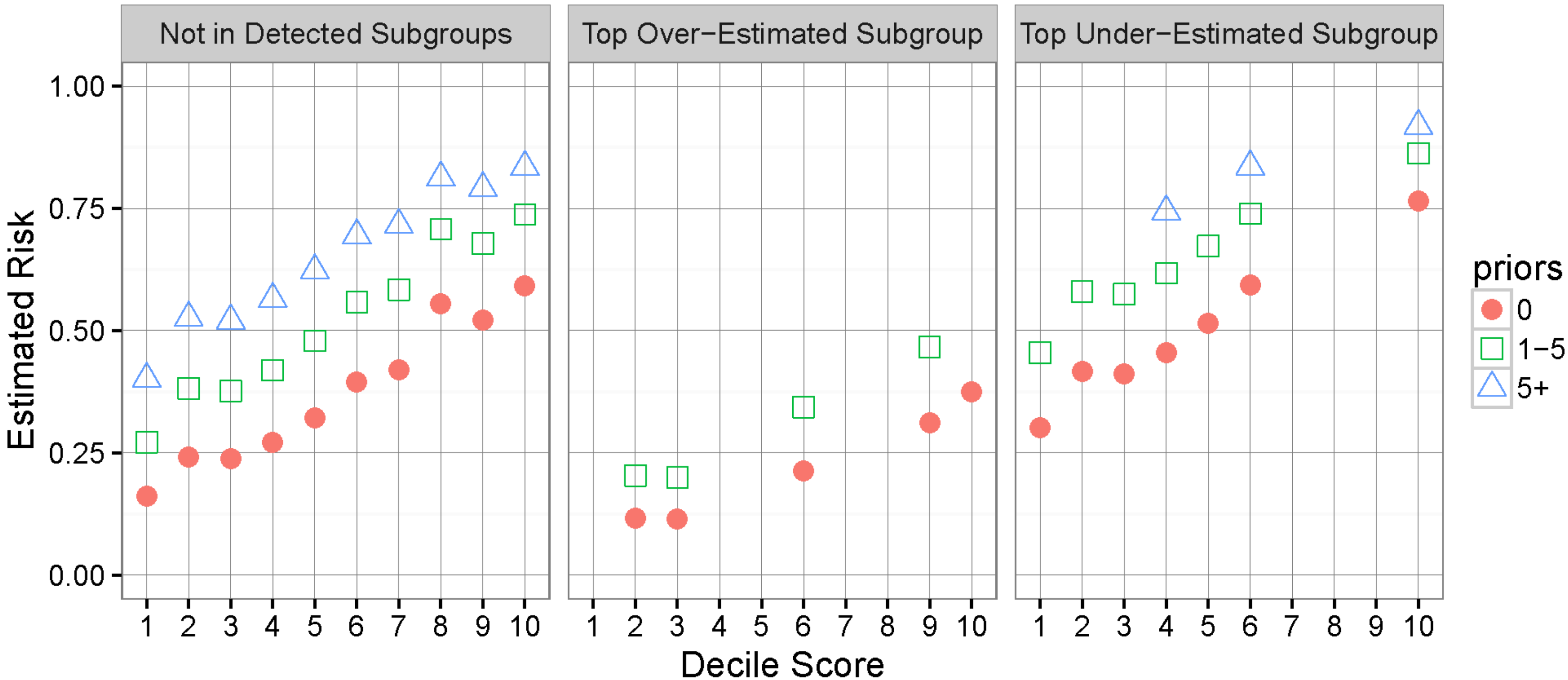


Start with maximum likelihood risk estimates for each COMPAS decile score.

Detection result 1: COMPAS underestimates the importance of prior offenses, overestimating risk for 0 priors, and underestimating risk for 5 or more priors.

Detection result 2: Even controlling for prior offenses, COMPAS still underestimates risk for males under 25, and overestimates risk for females who committed misdemeanors.

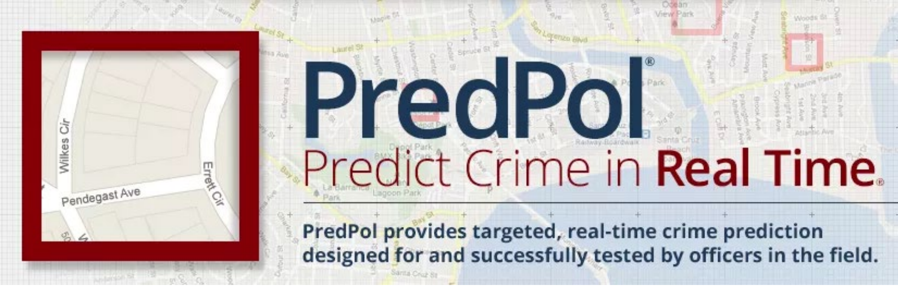
Results of bias scan on COMPAS



After controlling for prior offenses and membership in the two detected subgroups, there are no significant systematic biases in prediction.

Thorny question: given individual risk predictions, what to do with them?

The bigger picture



Big Data: A Report on Algorithmic Systems, Opportunity, and Civil Rights

Executive Office of the President

May 2016

THE WHITE HOUSE
WASHINGTON

Chronicle Of Social Change

Chronicle Webpage

www

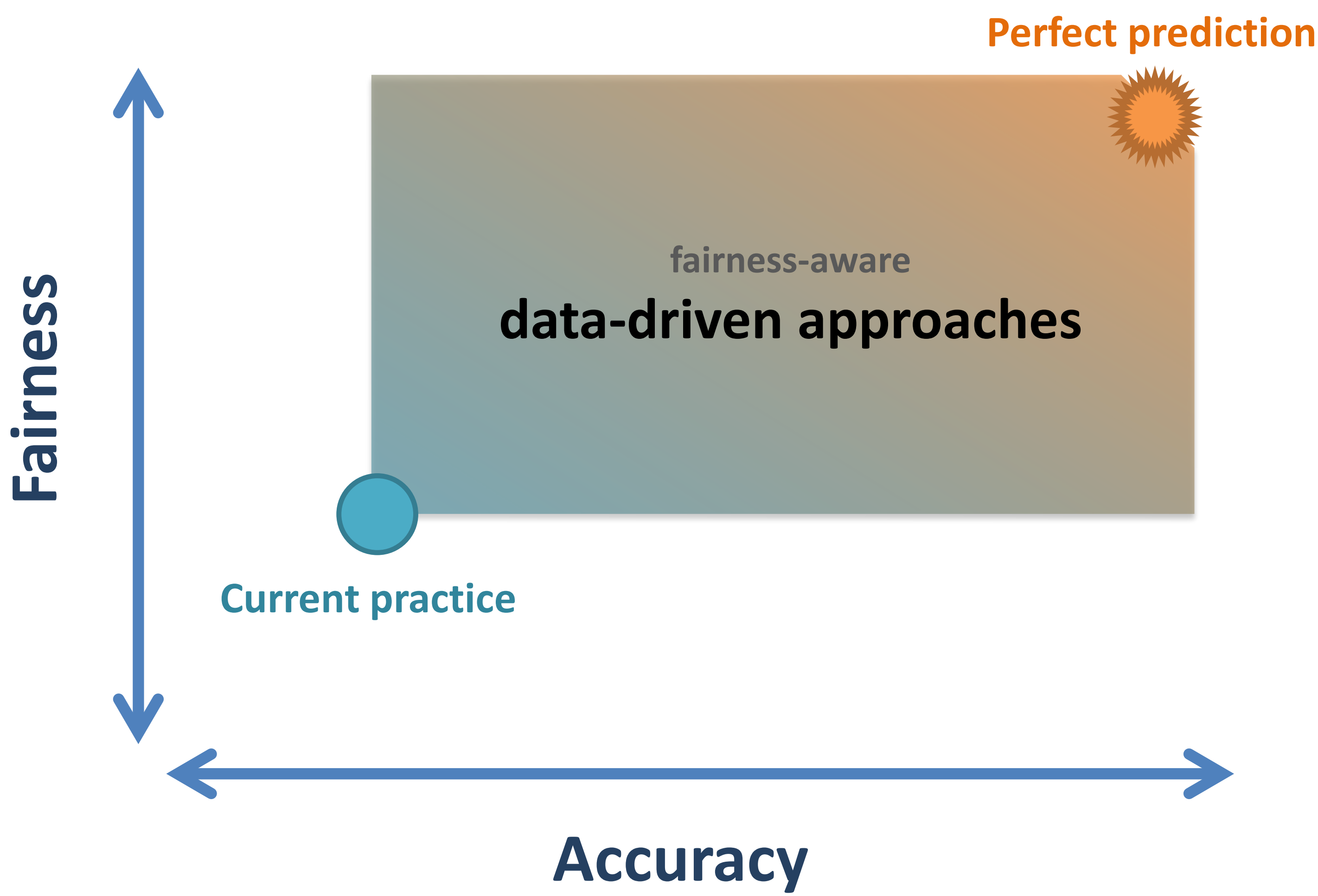
California Bets on Big Data to Predict Child Abuse

CADE METZ BUSINESS 07.11.16 7:00 AM

ARTIFICIAL INTELLIGENCE IS SETTING UP THE INTERNET FOR A HUGE CLASH WITH EUROPE

GETTY IMAGES

Should we adopt data-driven approaches?



References

- Resources for fairness, accountability, and transparency in ML: https://www.fatml.org/resources/relevant-scholarship
- A. Chouldechova. Fair prediction with disparate impact: a study of bias in recidivism prediction instruments. *Big Data*, 5(2): 153-163, 2017.
- Z. Zhang and D.B. Neill. Identifying significant predictive bias in classifiers. https://arxiv.org/pdf/1611.08292.pdf. In *NIPS Workshop on Interpretable Machine Learning*, 2016.
- A Romei & S Ruggieri. A multidisciplinary survey on discrimination analysis. http://www.di.unipi.it/~ruggieri/Papers/ker.pdf
- S. Barocas and A.D. Selbst. Big Data's Disparate Impact. In *104 California Law Review* 671, 2016.
- Žliobaitė, I. Measuring discrimination in algorithmic decision making. *Data Mining and Knowledge Discovery*, 31(4): 1060-1089, 2017. http://www.zliobaite.com/publications
- M. Bilal Zafar, *et al*. Learning Fair Classifiers. Tech. report, 2016. http://arxiv.org/pdf/1507.05259v3.pdf
- S. Feldman et al.. Certifying and removing disparate impact. In *Proc. KDD 2015*, http://sorelle.friedler.net/papers/kdd_disparate_impact.pdf