NYU

Algorithmic Fairness, Discrimination, and Bias

Week 9

March 25, 2024

Spring 24 | CUSP-GX 7033 – Machine Learning for Cities | Dr. Anton Rozhkov

Today's Outline

- Definitions of fairness and bias
- The need for fairness in algorithms: motivation and examples
- Fairness-aware algorithms.
  - Next class:
    - 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
- Final Project Discussion: Groups and Proposal

DEFINITIONS OF FAIRNESS AND BIAS

THE NEED FOR FAIRNESS IN ALGORITHMS: MOTIVATION AND EXAMPLES

NYU

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

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.

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 ›

NYU

Websites Vary Prices, Deals Based on Users' Information

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

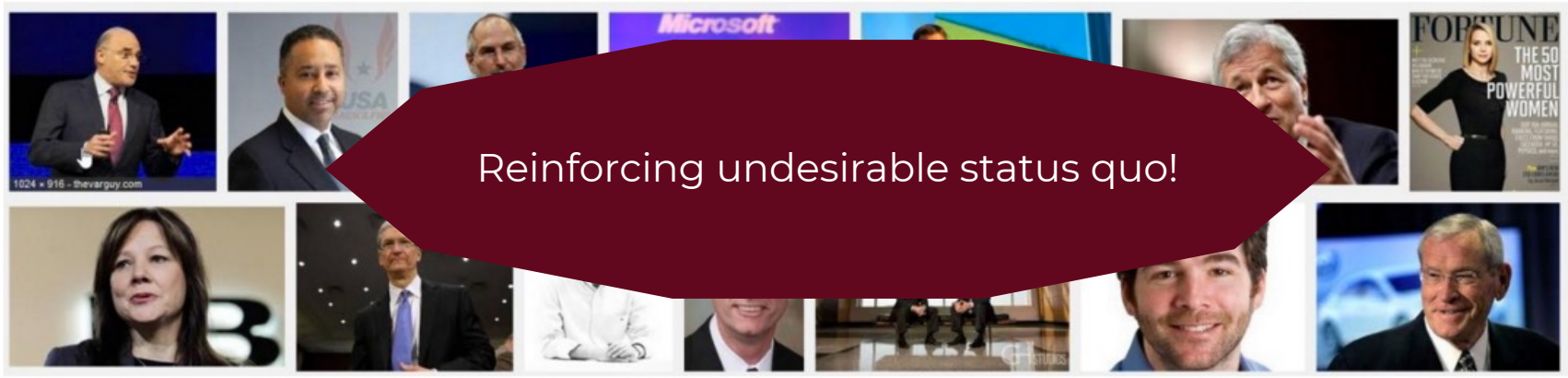
It was the same Swingline stapler, on the same Staples.com website. But for Kim Wamble, the price was $15.79, while the price on Trude Frizzell's screen, just a few miles away, was $14.29.

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

Unforeseen consequences!

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

What happened: 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 ›

NYU

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'

NYU

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?

NYU

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

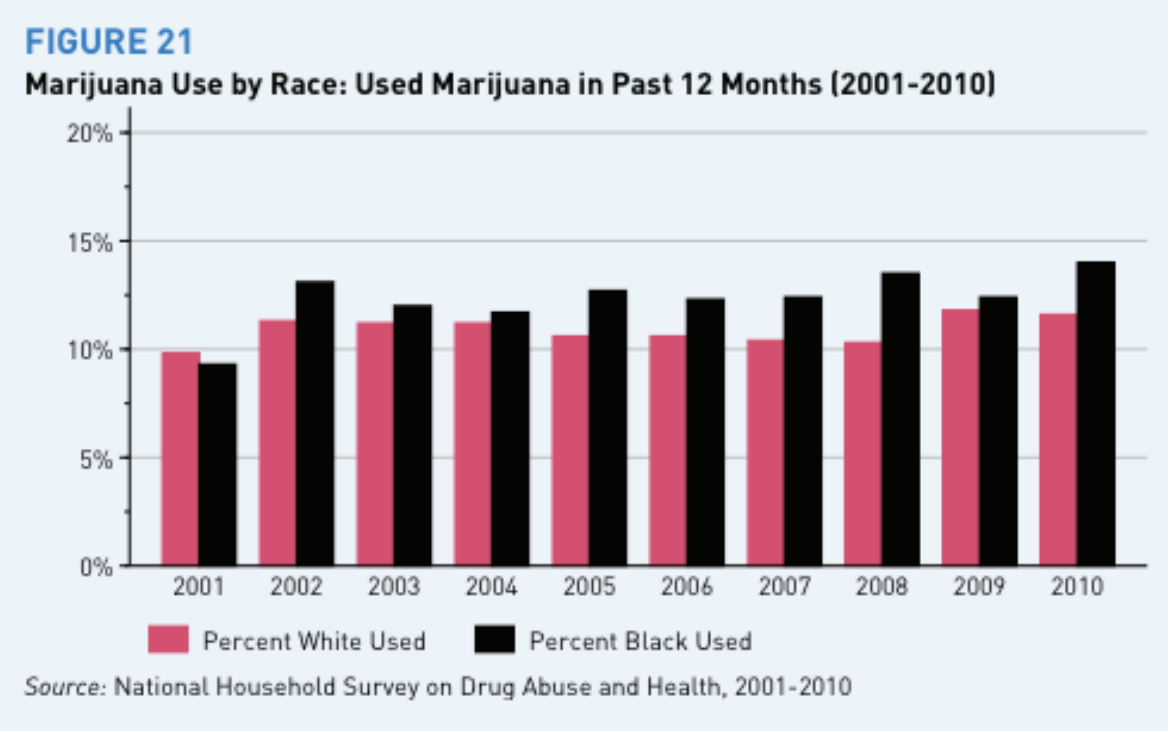
What's your Y?

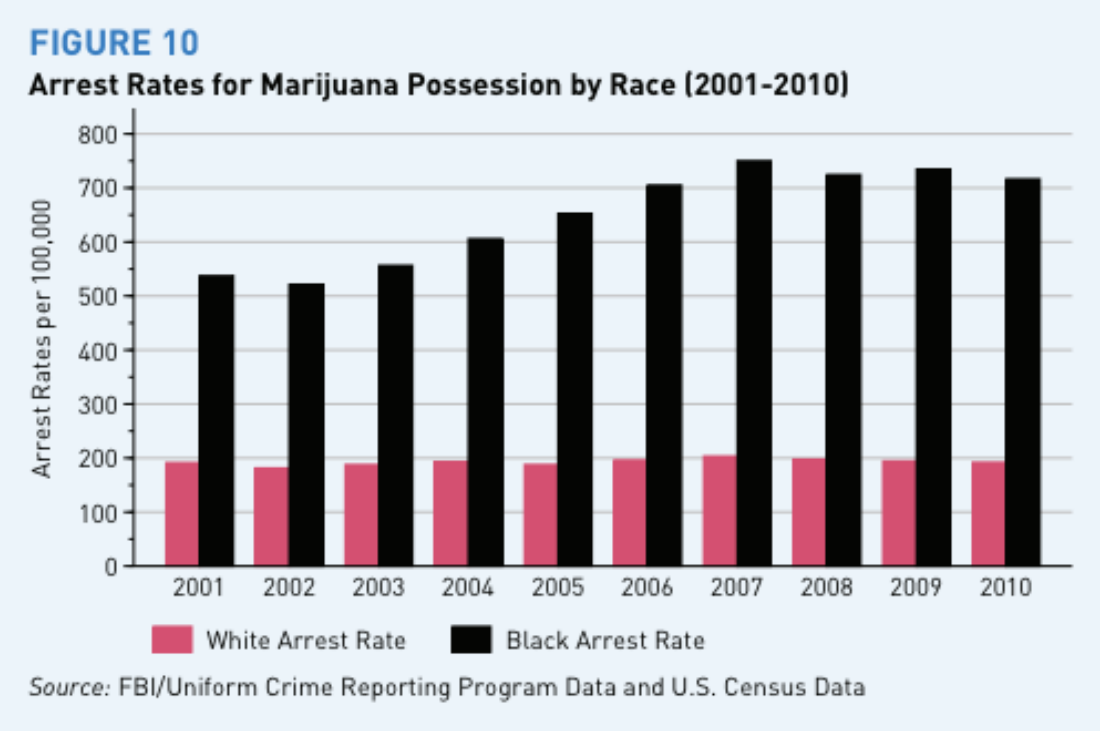
- **Y** = *candidate was hired* vs. **Y** = *employee productivity*
- Target variable bias: in recidivism prediction, we:
  want **Y** = **re-offense**, have **Y** = **re-arrest**

NYU

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**





PREVENTING DISPARATE IMPACT: A CASE STUDY IN CRIMINAL JUSTICE

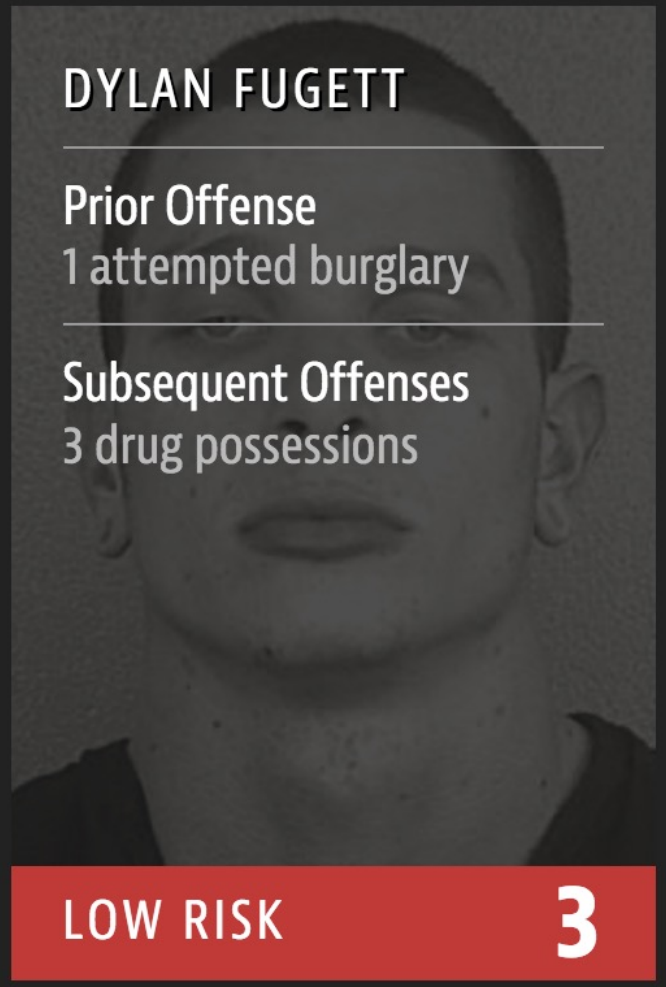
Bernard Parker, left, was rated high risk; Dylan Fugett was rated low risk. (Josh Ritchie for ProPublica)

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

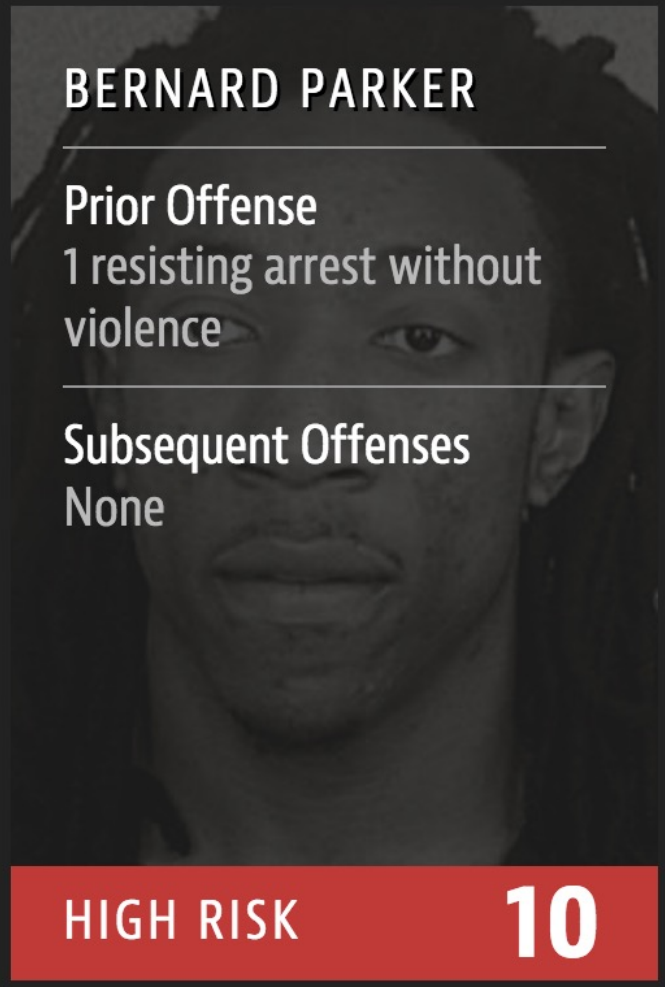


DYLAN FUGETT

Prior Offense
1 attempted burglary

Subsequent Offenses
3 drug possessions

LOW RISK 3



BERNARD PARKER

Prior Offense
1 resisting arrest without violence

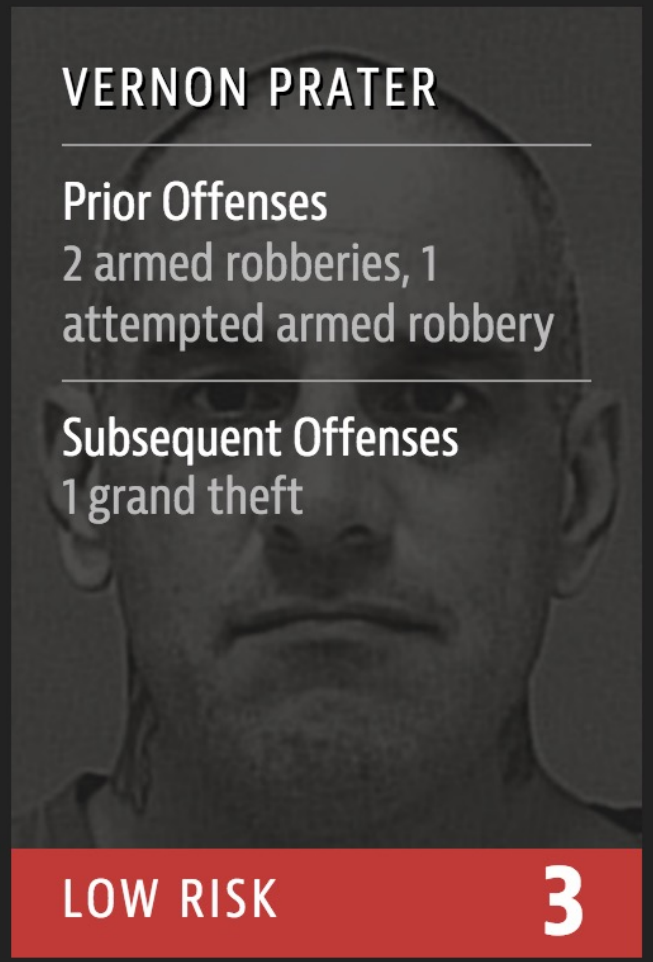
Subsequent Offenses
None

HIGH RISK 10

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

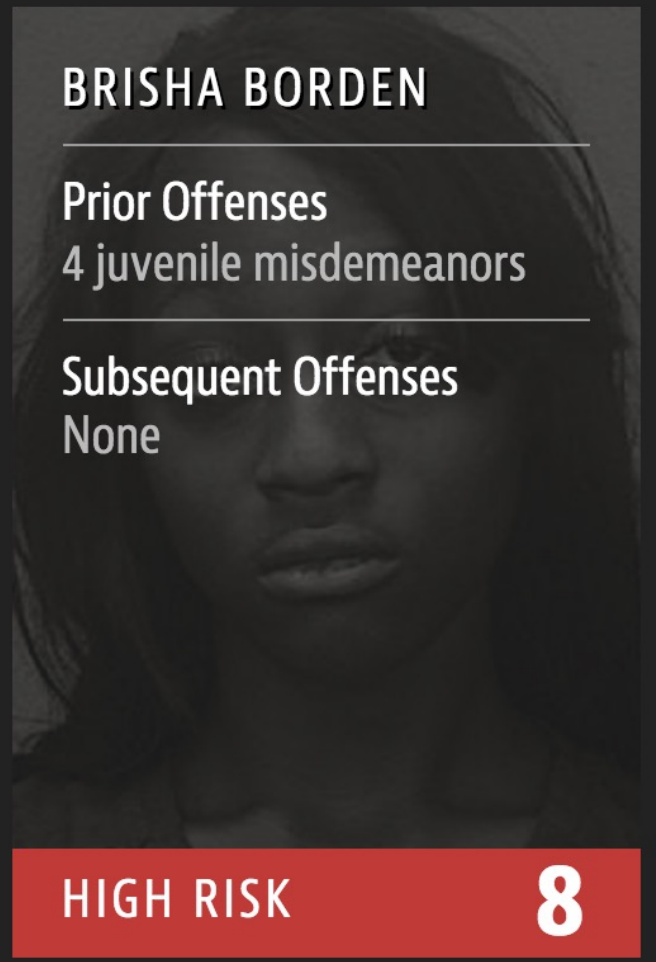


VERNON PRATER

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**

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

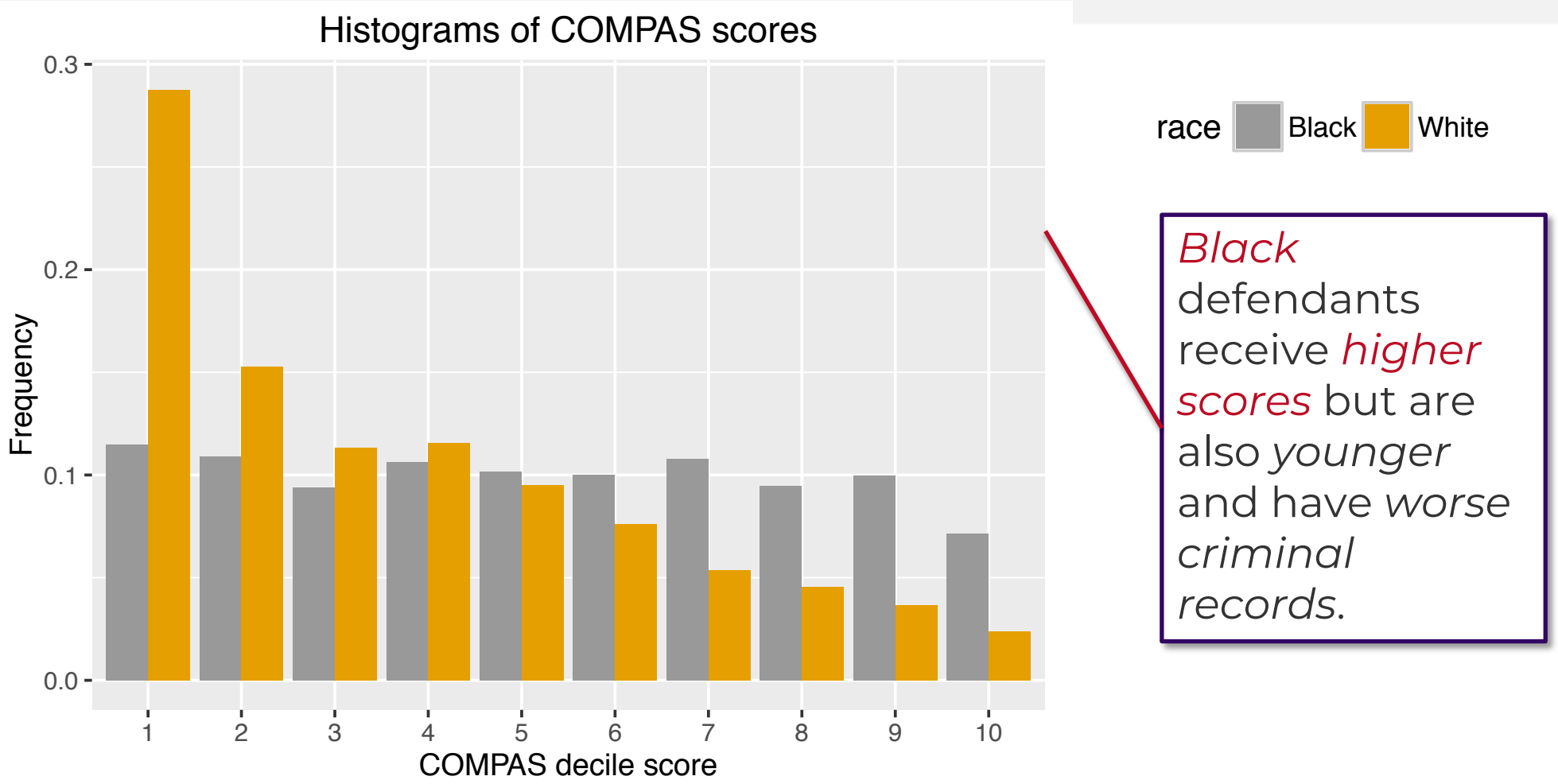
Source: ProPublica's data on criminal defendants in Broward County, Florida in 2013 – 2014, outcome assessed through April 2016

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)

NYU

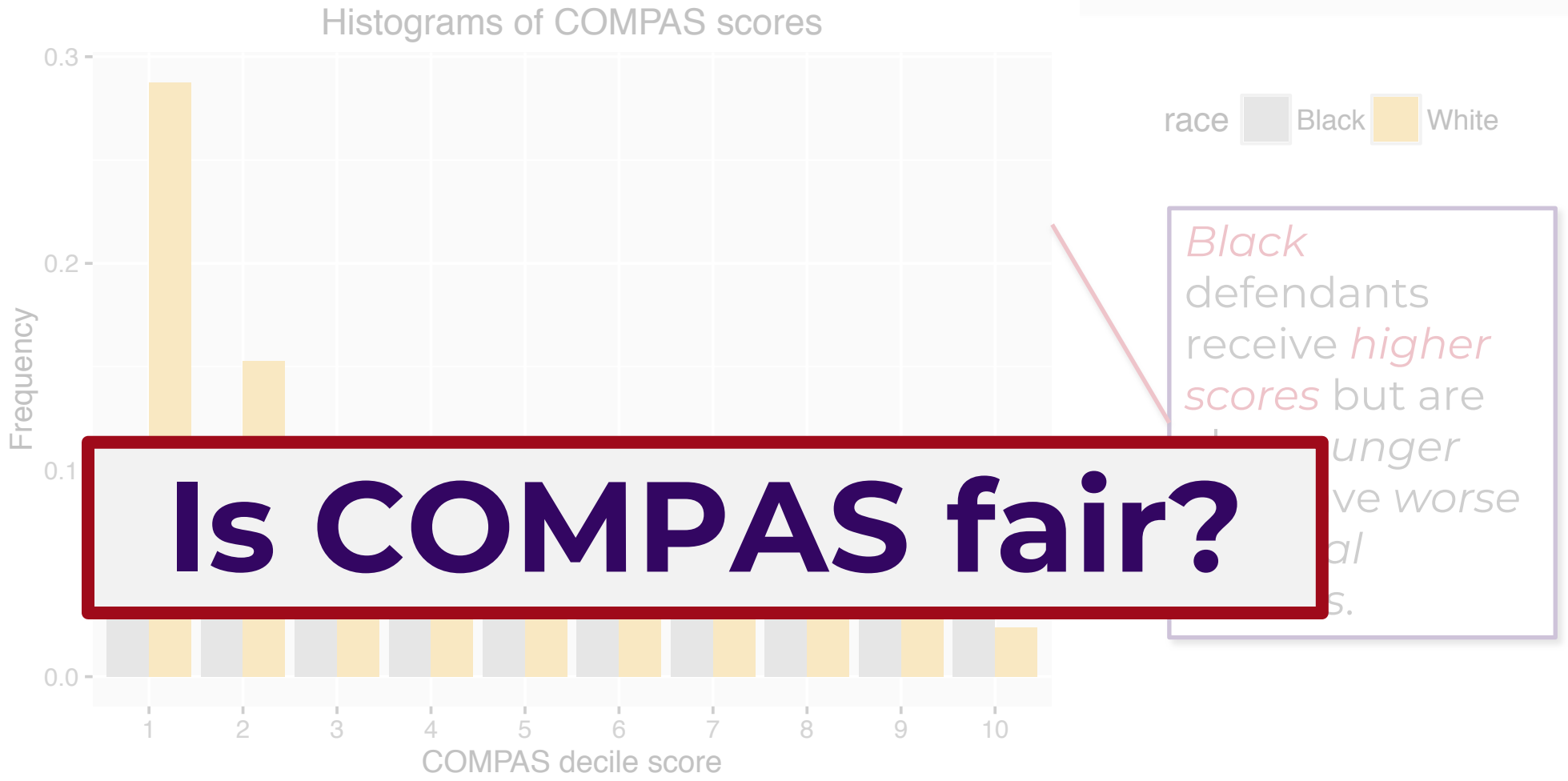


Black defendants receive *higher scores* but are also *younger* and have *worse criminal records*.

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

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

NYU



Black defendants receive *higher scores* but are *unger* *ve worse* *al* *s.*

Is COMPAS fair?

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

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

NYU

Well, that depends on how we 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.**

NYU

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~~ Did Re-Offend Labeled Lower Risk	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.)

Positive predictive value (aka Precision)

False **negative** rates

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

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 ⇨ *error rate imbalance*
2. *Error rate imbalance* leads to disparate impact under policies that assign stricter penalties to individuals assessed as higher-risk.

NYU

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.

NYU

Building Blocks: Confusion Tables

- Score S. If $S > s_{HR}$, say the person is "High Risk"
- Outcome
- Group membership, e.g., Race

Black defendants

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

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

White defendants

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

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

NYU

Predictive Parity (Northpointe's criterion)

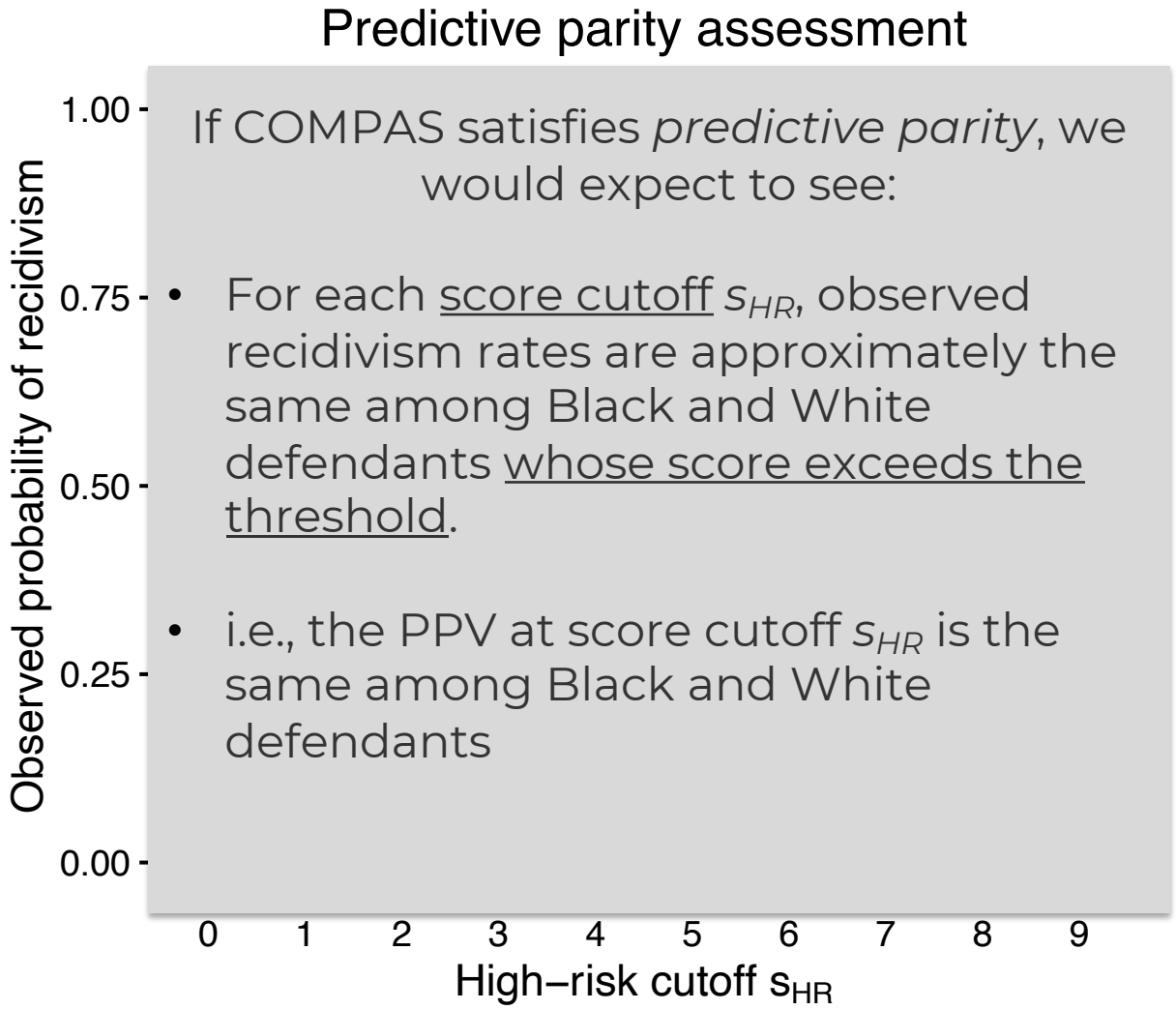
$$\mathbb{P}(Y = 1 \mid S > s_{\mathrm{HR}}, R = b) = \mathbb{P}(Y = 1 \mid S > s_{\mathrm{HR}}, R = w)$$

Predictive Parity (Northpointe's criterion)

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

Predictive Parity (Northpointe's criterion)

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

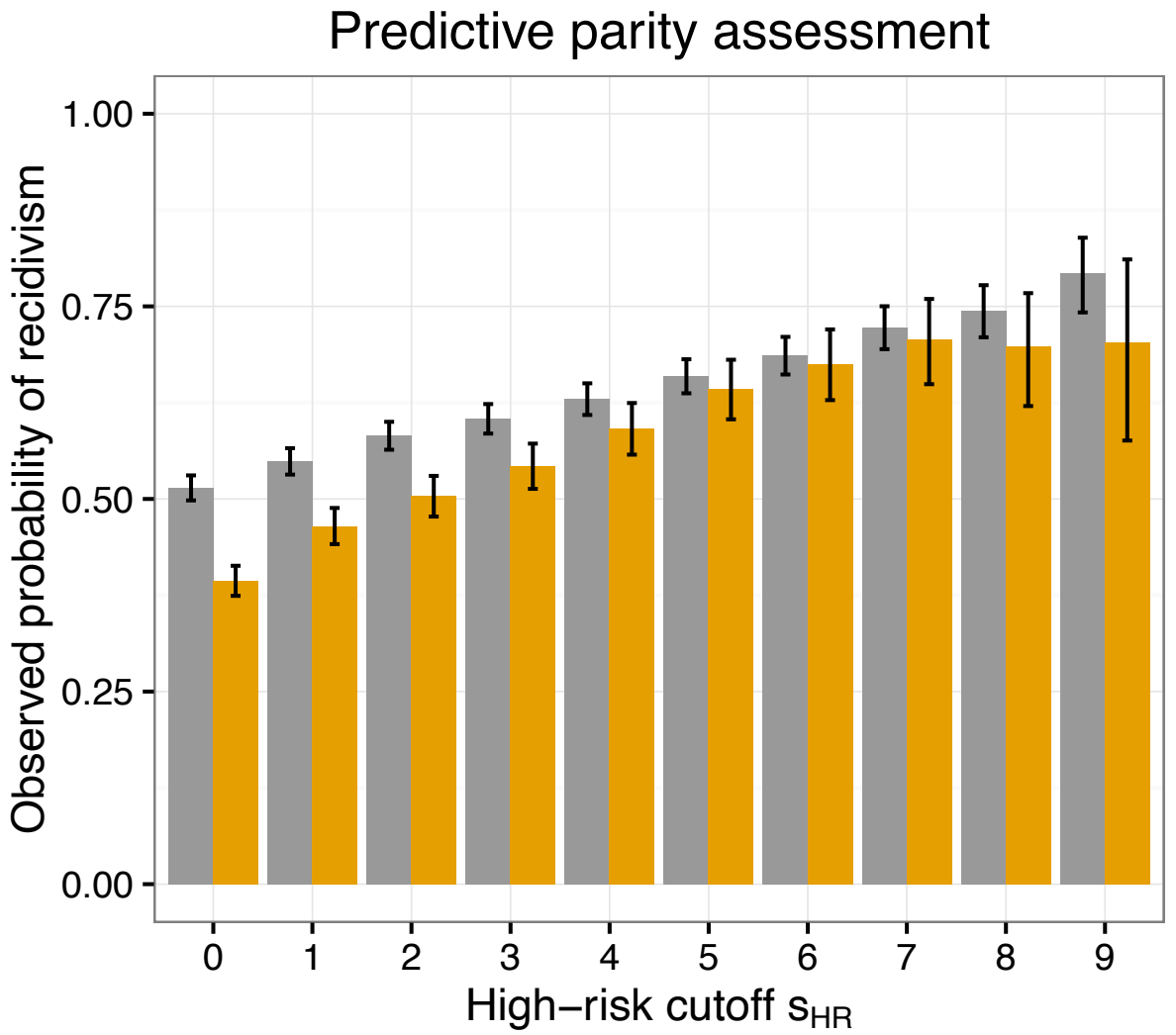


If COMPAS satisfies *predictive parity*, we would expect to see:

- 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

Predictive Parity (Northpointe's criterion)

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



race Black White

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

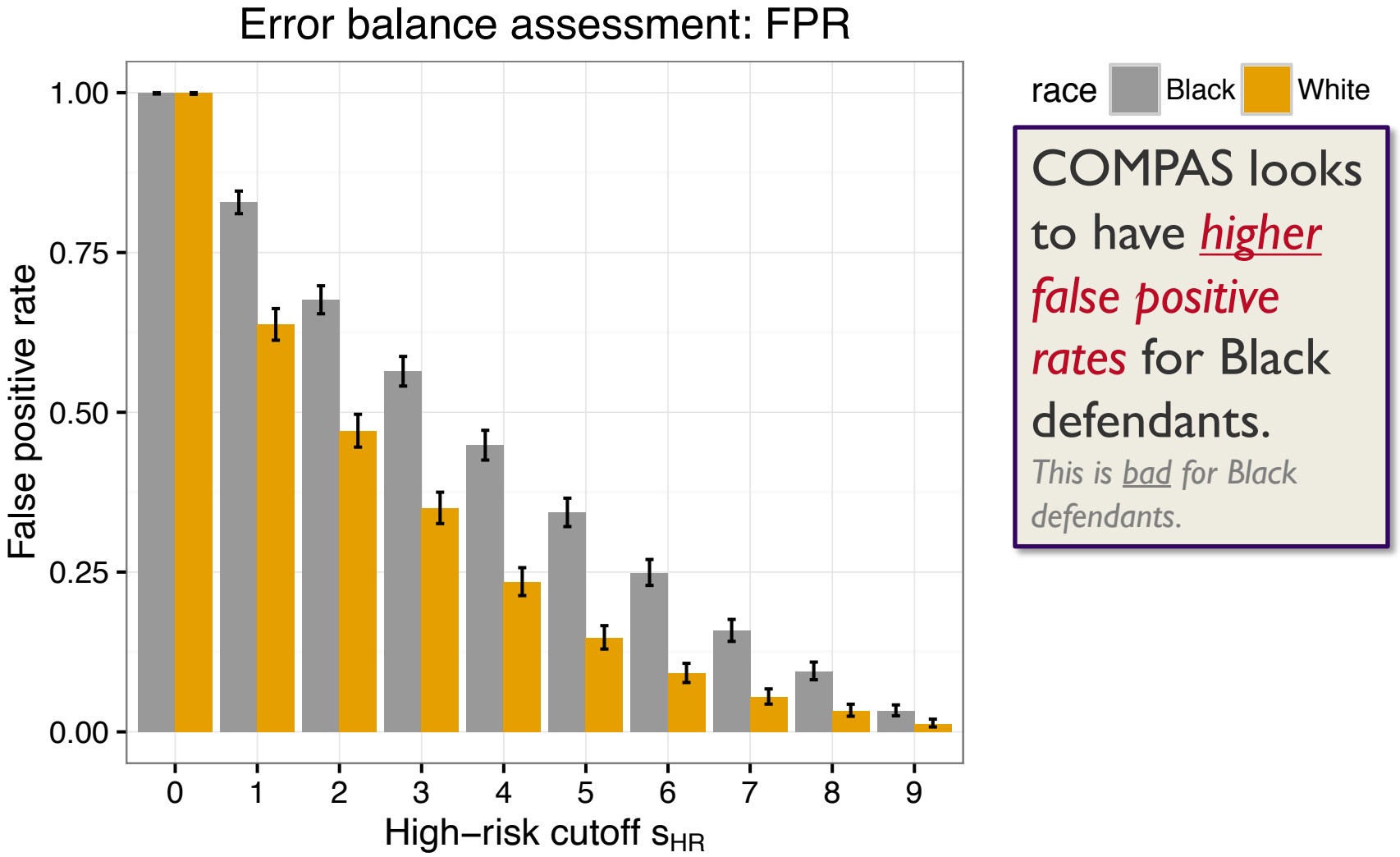
NYU

False Positive Rate Balance (ProPublica criterion)

$$\mathbb{P}(S > s_{\mathrm{HR}} \mid Y = 0, R = b) = \mathbb{P}(S > s_{\mathrm{HR}} \mid Y = 0, R = w)$$

False Positive Rate Balance (ProPublica criterion)

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

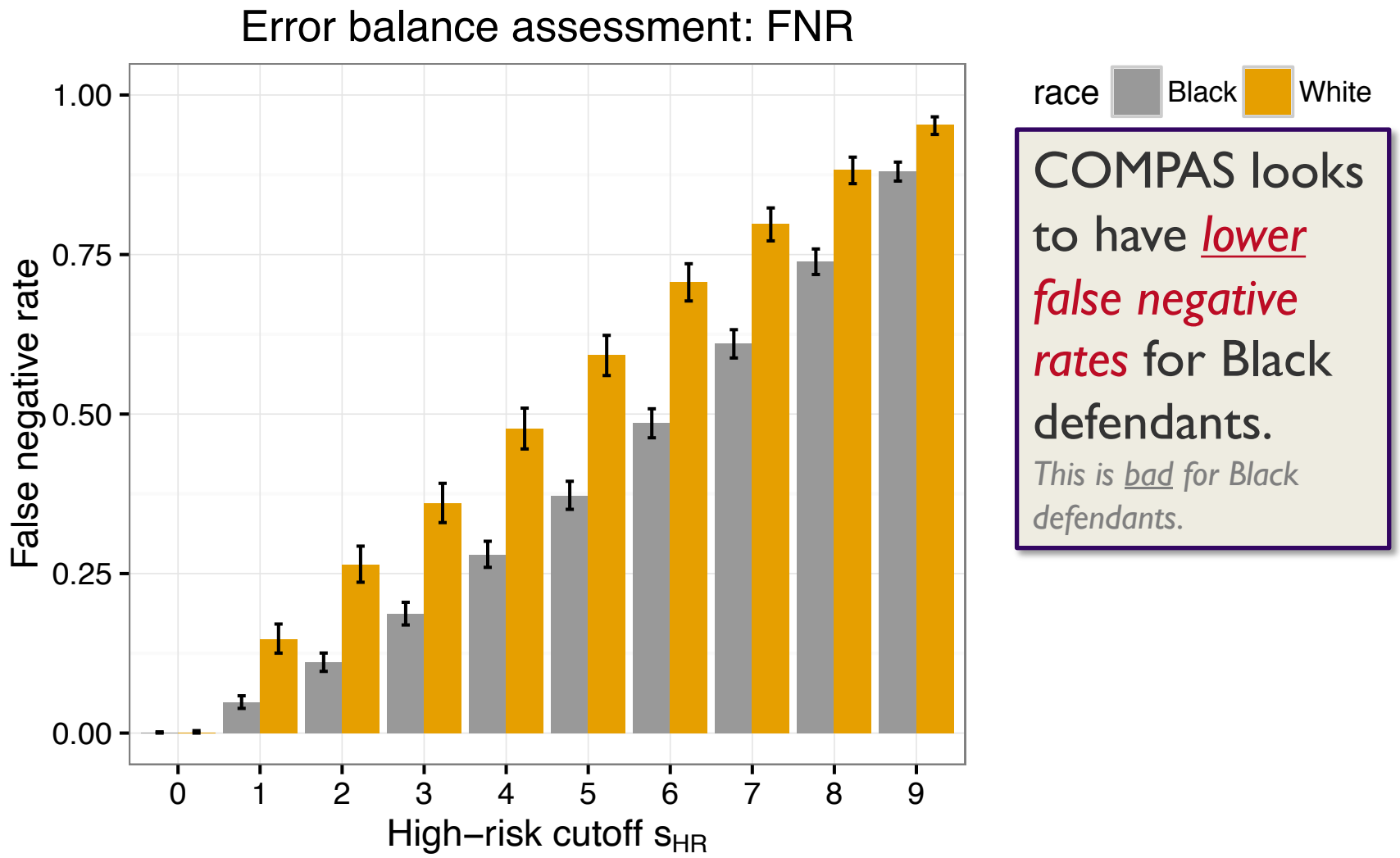


COMPAS looks to have *higher false positive rates* for Black defendants.

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 ($s_{HR} = 4$)

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.

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

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)

NYU

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									

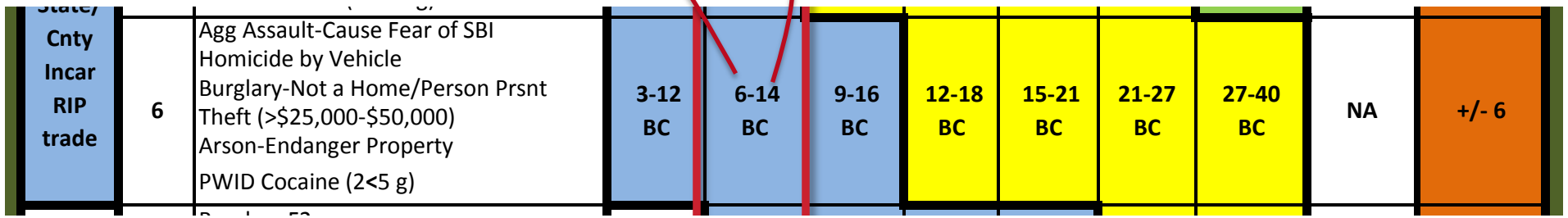
NYU

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*:

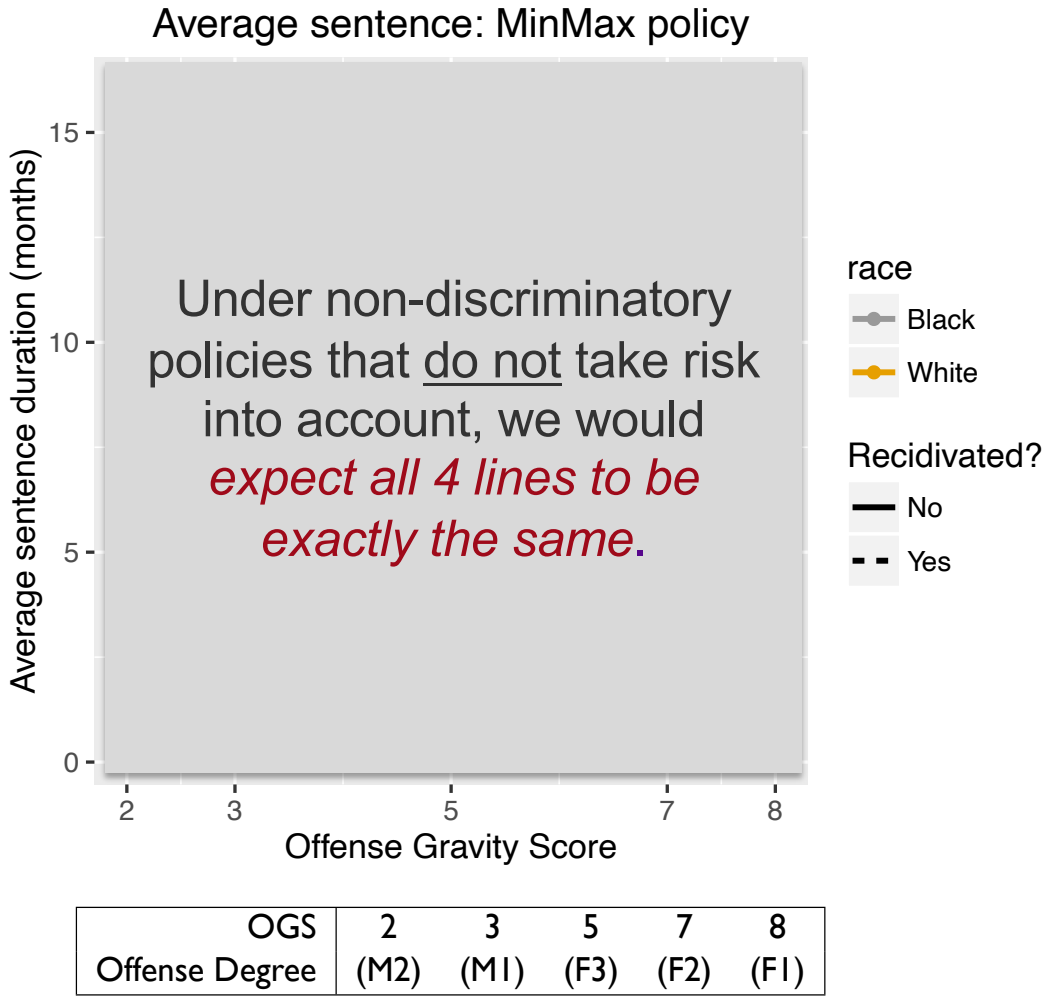
MinMax

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



Interpolation

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

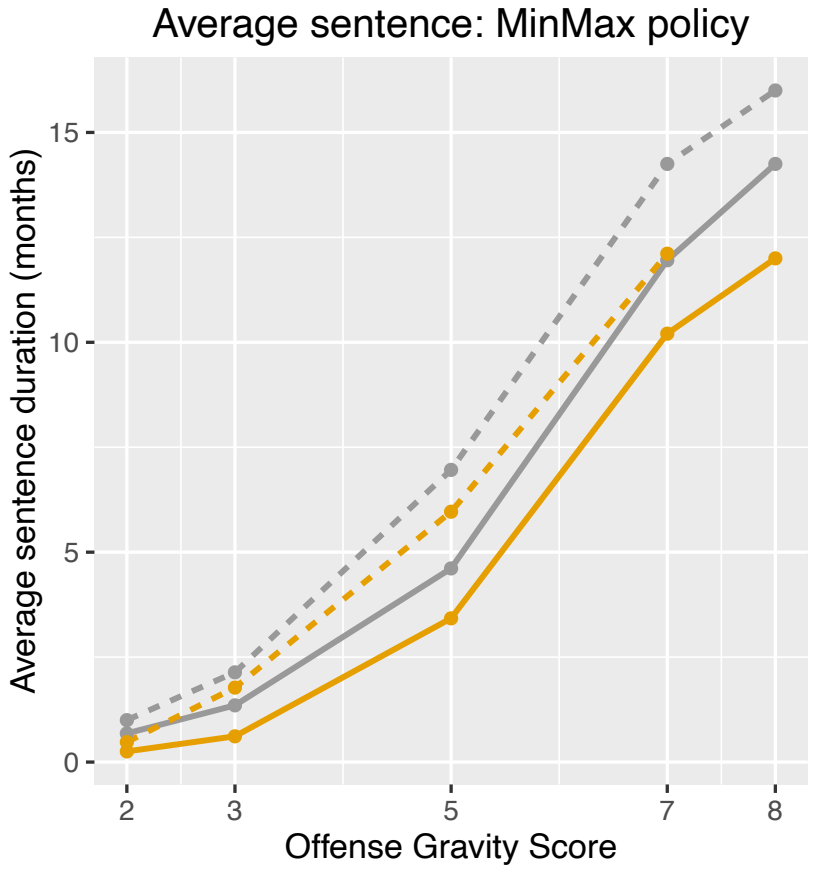


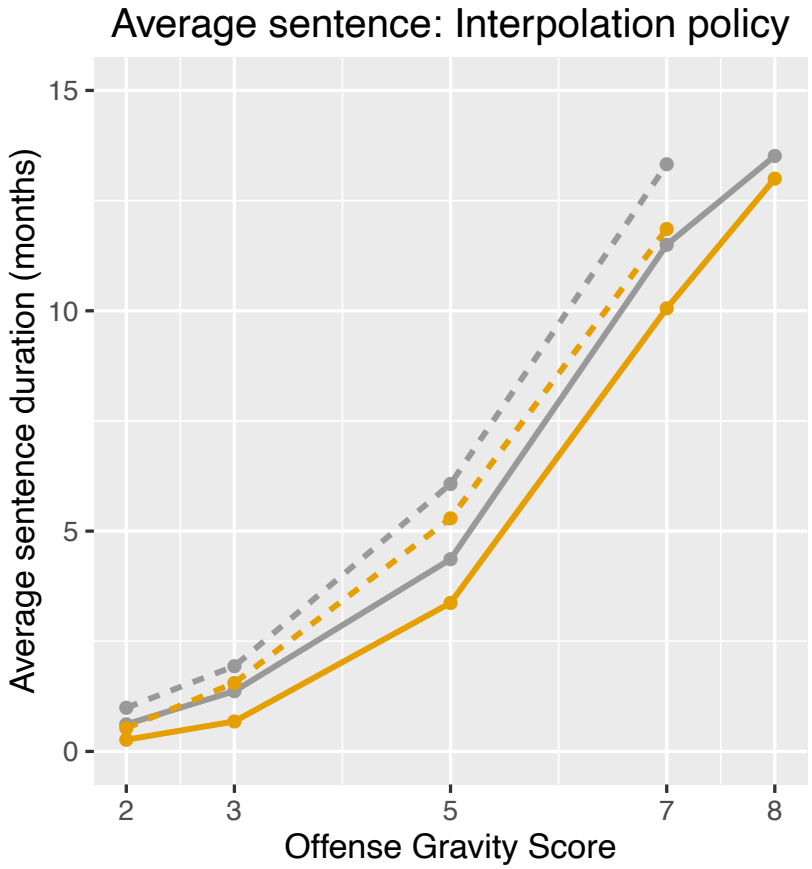
OGS	2	3	5	7	8
Offense Degree	(M2)	(M1)	(F3)	(F2)	(F1)

* 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.

NYU





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

NYU

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 Specific Cutoffs

Before:
Cutoff = 4 for both groups

Black defendants

metric	value
n	3696
prevalence	0.51
PPV	0.63
FPR	0.45
FNR	0.28

White defendants

metric	value
n	2454
prevalence	0.39
PPV	0.59
FPR	0.24
FNR	0.48

imbalanced!

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

metric	value
n	3696
prevalence	0.51
PPV	0.69
FPR	0.25
FNR	0.49

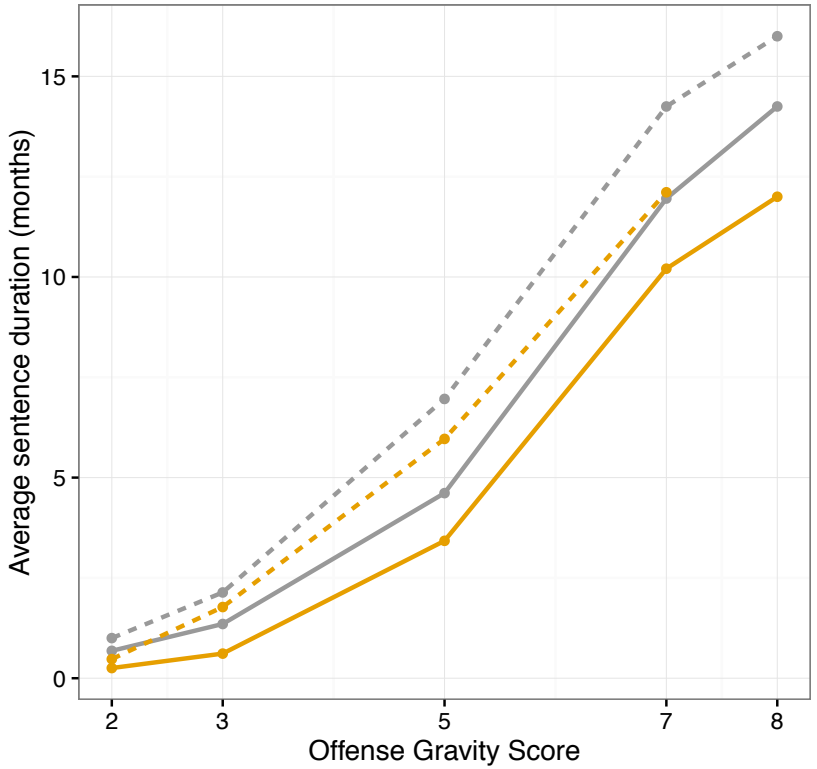
metric	value
n	2454
prevalence	0.39
PPV	0.59
FPR	0.24
FNR	0.48

balanced!

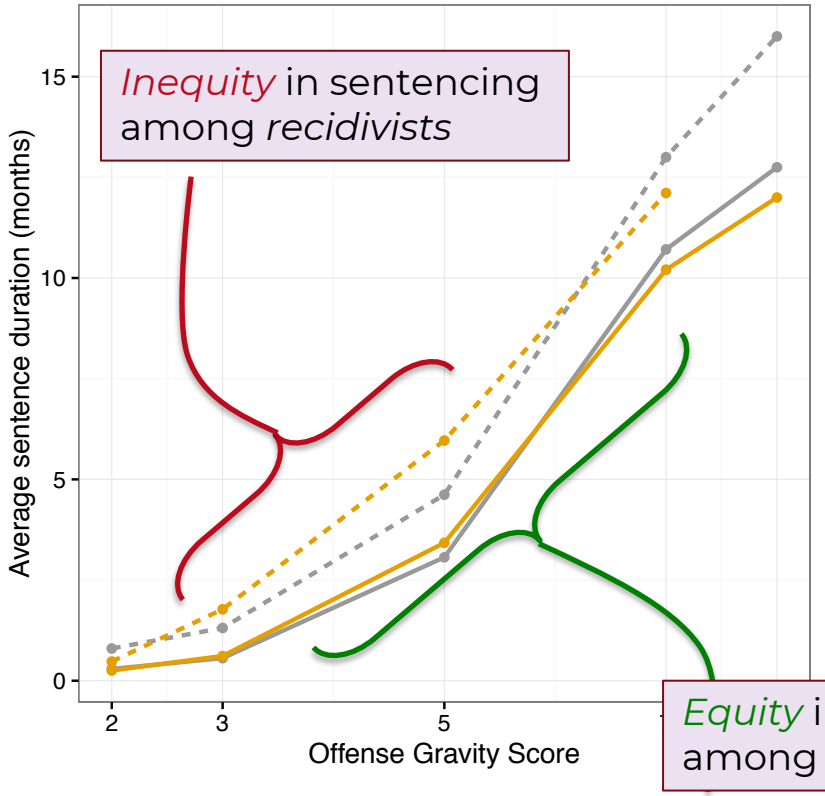
NYU

Did We Succeed?

MinMax Sentencing, *Before*



MinMax Sentencing, *After cutoff change*



race Black White

Recidivated? — No - - Yes

NYU

Takeaway

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

Lab Time

Final Project Discussion

NYU

Final Project Proposal

The proposal/abstract should be ***no more than 1 page long*** (shorter is fine), and should provide the following information:

1. Working project title
2. Names and e-mail addresses of all group members
3. Brief description of the goal of the project.
4. What datasets will you be analyzing, what methods do you think might be useful, and what do you hope to achieve?

NYU

For the Next Week (Week 10)

1. Final Project Proposal

   Due: March 31, 2024 (11:59pm)

References

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

NYU