Data science ethics

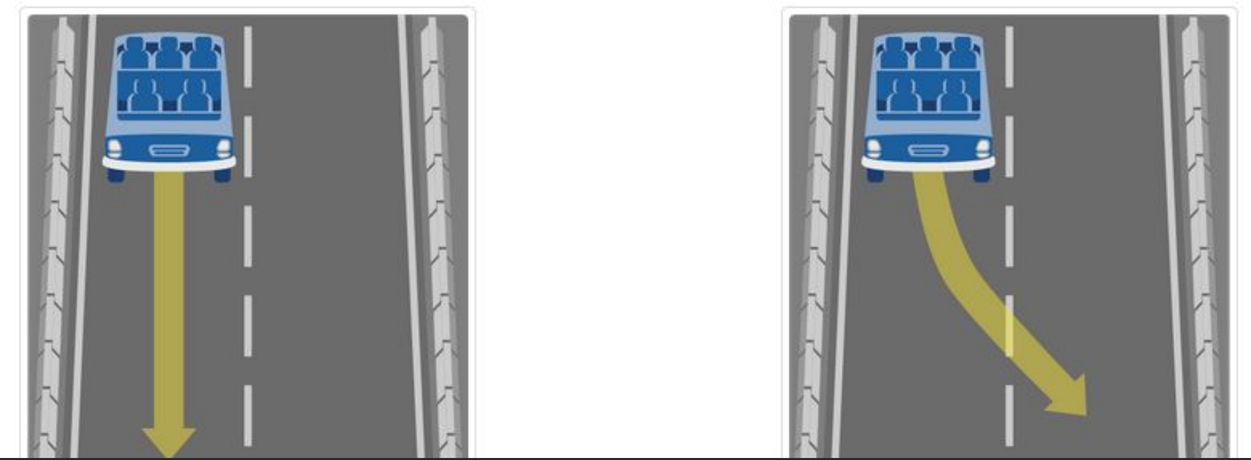
Data science ethics

Data science **ethics**

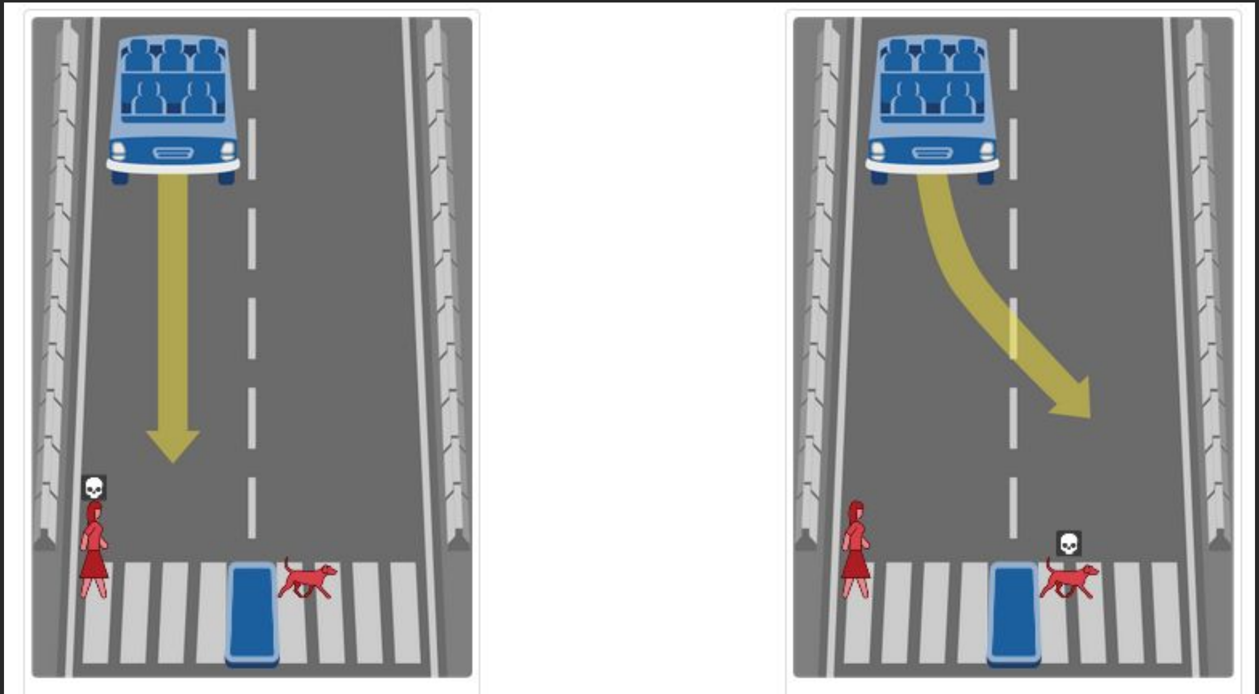
Topics

- Intro what data science ethics
- Bias & fairness

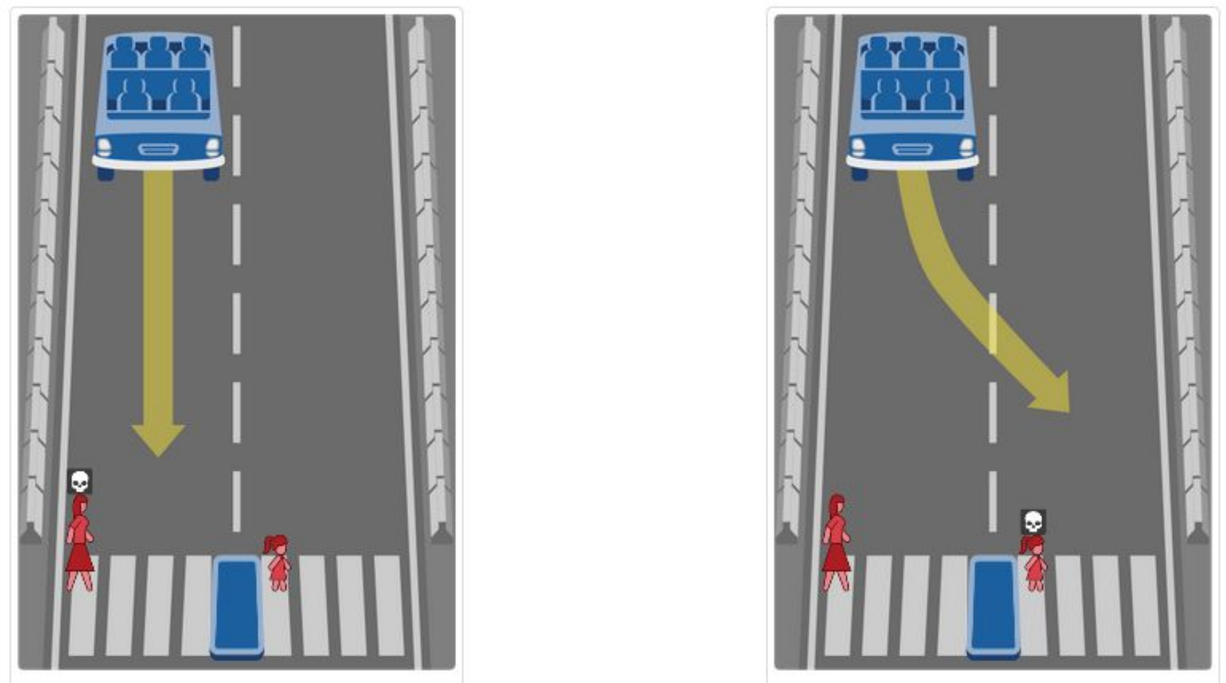
Ethical dilemma



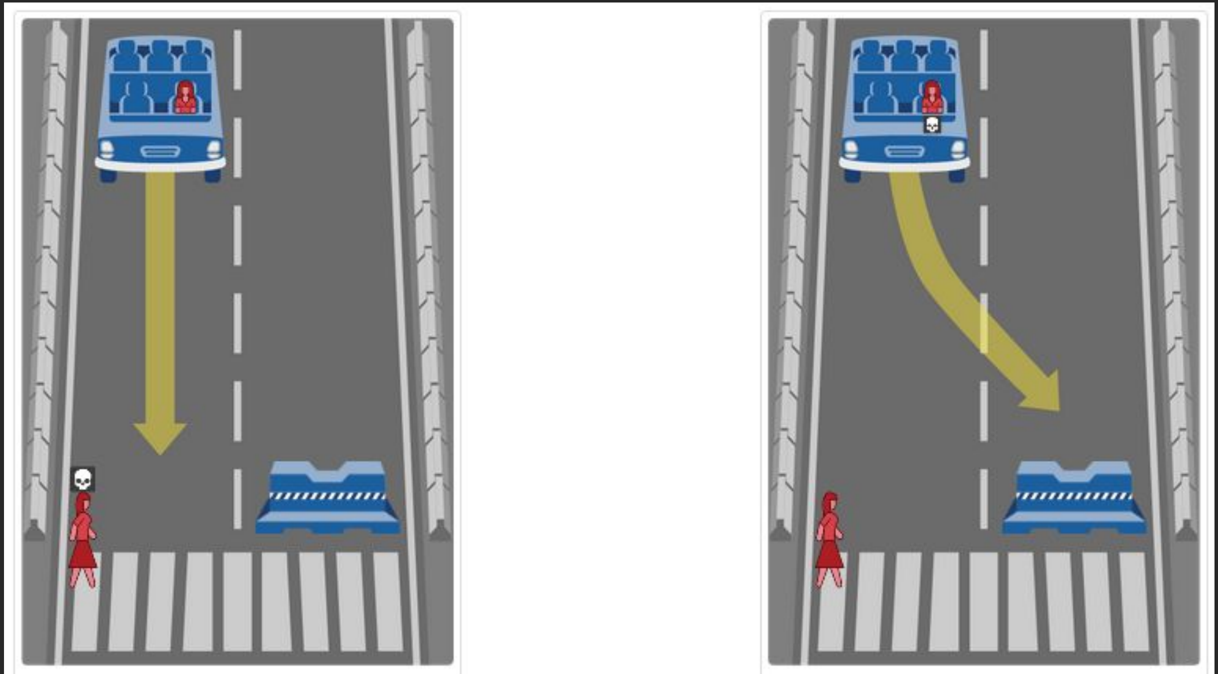
Ethical dilemma



The self-driving car dilemma



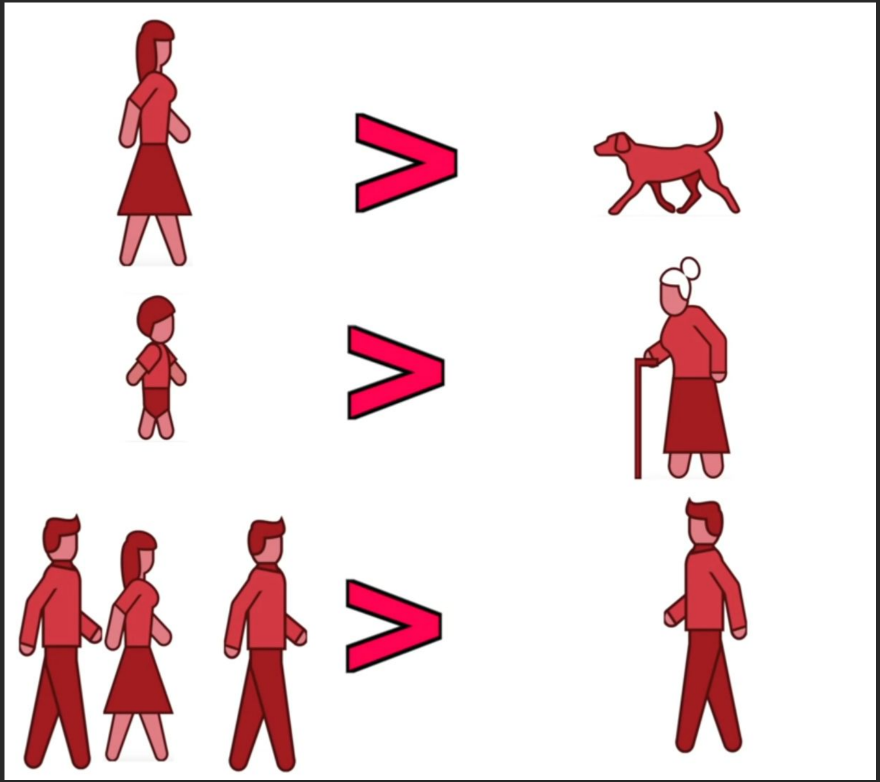
The self-driving car dilemma



MIT self-driving car survey

- Online survey
- Millions of participants
- 233 countries

MIT self-driving car survey



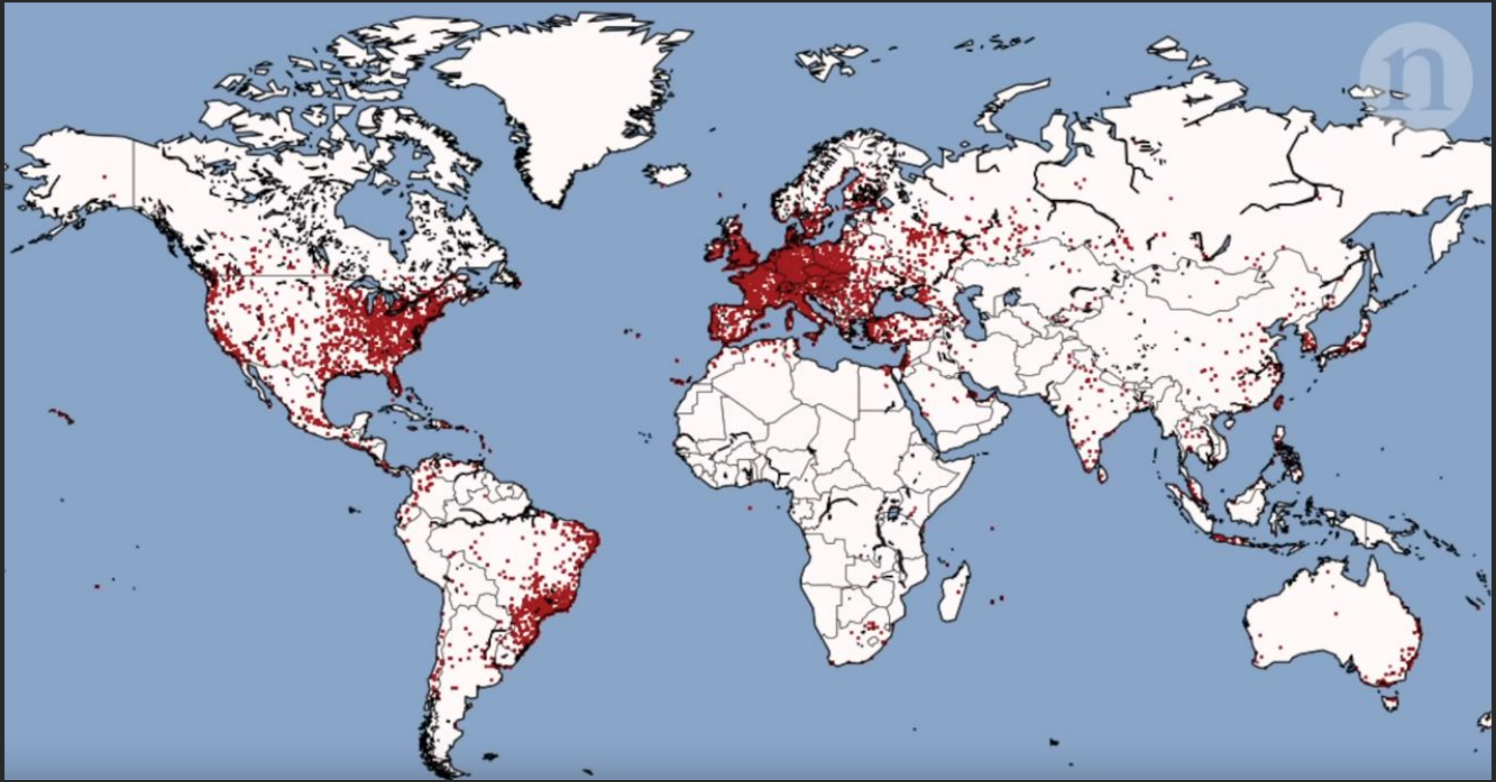
MIT self-driving car survey

- Many countries in east Asia, put a similar value on the elderly and younger lives.
- Many french speaking countries put a higher value on female lives.
- Countries with high income disparity generally valued the lives of “business people” higher than others.

Bias with the car survey data?

- Online survey
- Millions of participants
- 233 countries

Bias with the data?



Bias and fairness

Human bias

Published online 11 April 2011 | Nature | doi:10.1038/news.2011.227

News

Hungry judges dispense rough justice

When they need a break, decision-makers gravitate towards the easy option.

Zoë Corbyn

Journal of Economic Perspectives—Volume 12, Number 2—Spring 1998—Pages 41–62

Evidence on Discrimination in Mortgage Lending

Helen F. Ladd

Science faculty's subtle gender biases favor male students

Corinne A. Moss-Racusin[a,b], John F. Dovidio[b], Victoria L. Brescoll[c], Mark J. Graham[a,d], and Jo Handelsman[a,1]

[a]Department of Molecular, Cellular and Developmental Biology, [b]Department of Psychology, [c]School of Management, and [d]Department of Psychiatry, Yale University, New Haven, CT 06520

Edited* by Shirley Tilghman, Princeton University, Princeton, NJ, and approved August 21, 2012 (received for review July 2, 2012)

The Observer
Stop and search

Racial bias in police stop and search getting worse, report reveals

Despite reforms, black people are nine times more likely than white people to be checked for drugs

Mark Townsend
Home Affairs Editor

1714

Is it easier to get a job if you're Adam or Mohamed?

By Zack Adesina and Oana Marocico
BBC Inside Out

6 February 2017

Share

The two CVs sent out detailed the same level of qualifications and experience

A job seeker with an English-sounding name was offered three times the number of interviews than an applicant with a Muslim name, a BBC test found.

CNN Health » Food | Fitness | Wellness | Parenting | Live Longer   Live TV   U.S. Edition +

Drowsy driving is a factor in almost 10% of crashes, study finds

By Erin Gabriel, CNN
Updated 2:55 AM ET, Thu February 8, 2018

Sina Fazelpour
Carnegie Mellon University

Can technologies have bias?

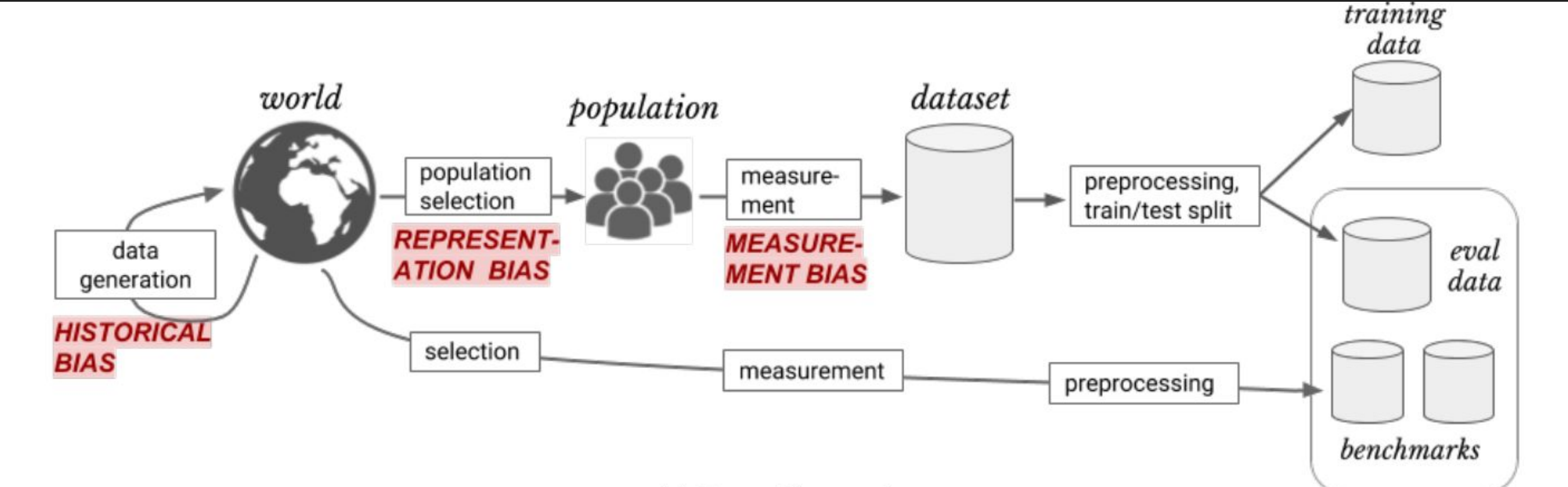
Technological biases

Humans have biases, so why does algorithmic bias matter?

- Cheap
- Scalable
- Automated
- Self-reinforcing
- Seemingly objective
- Often lacking appeals processes
- Does not just predict but also cause the future

Types of machine learning biases

Data & measurement biases



(a) Data Generation

Historical bias

Historical bias arises when there is a misalignment between world as it is and the values or objectives to be encoded and propagated in a model. It is a normative concern with the state of the world, and exists even given perfect sampling and feature selection.

Example: image search In 2018, 5% of Fortune 500 CEOs were women (Zarya, 2018). Should image search results for "CEO" reflect that number? Ultimately, a variety of stakeholders, including affected members of society, should evaluate the particular harms that this result could cause and make a judgment. This decision may be at odds with the available data even if that data is a perfect reflection of the world. Indeed, Google has recently changed their Image Search results for "CEO" to display a higher proportion of women.

Representation bias

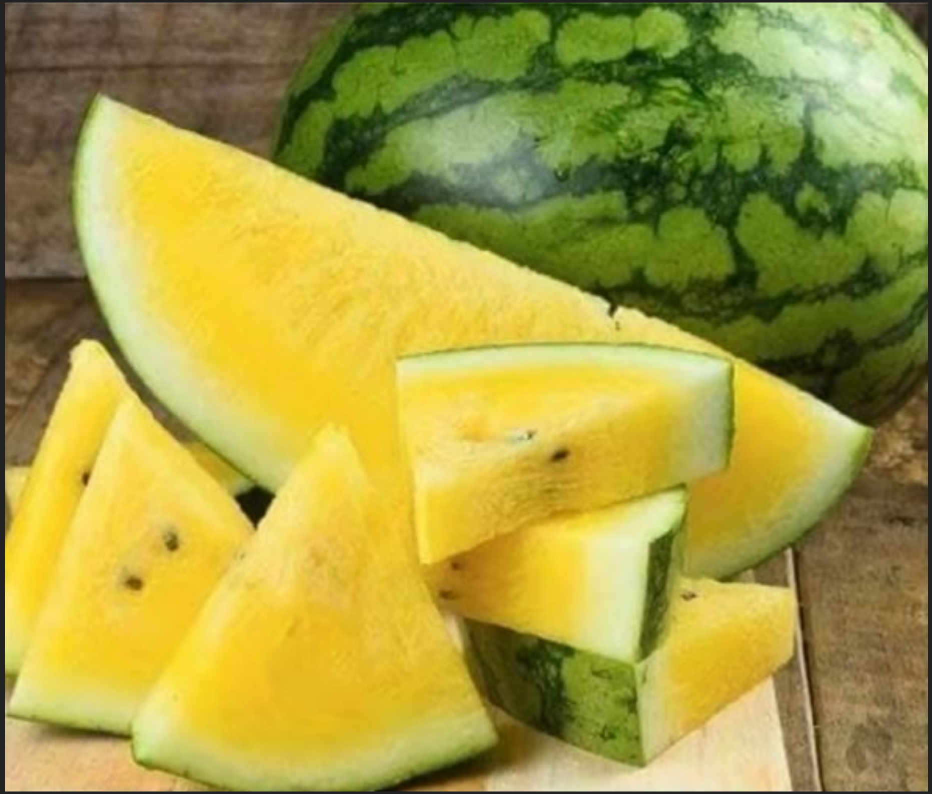
Representation bias arises while defining and sampling a development population. It occurs when the development population under-represents, and subsequently fails to generalize well, for some part of the use population.

1. **The sampling methods only reach a portion of the population.** For example, datasets collected through smartphone apps can under-represent lower-income or older groups, who are less likely to own smartphones. Similarly, medical data for a particular condition may be available only for the population of patients who were considered serious enough to bring in for further screening.

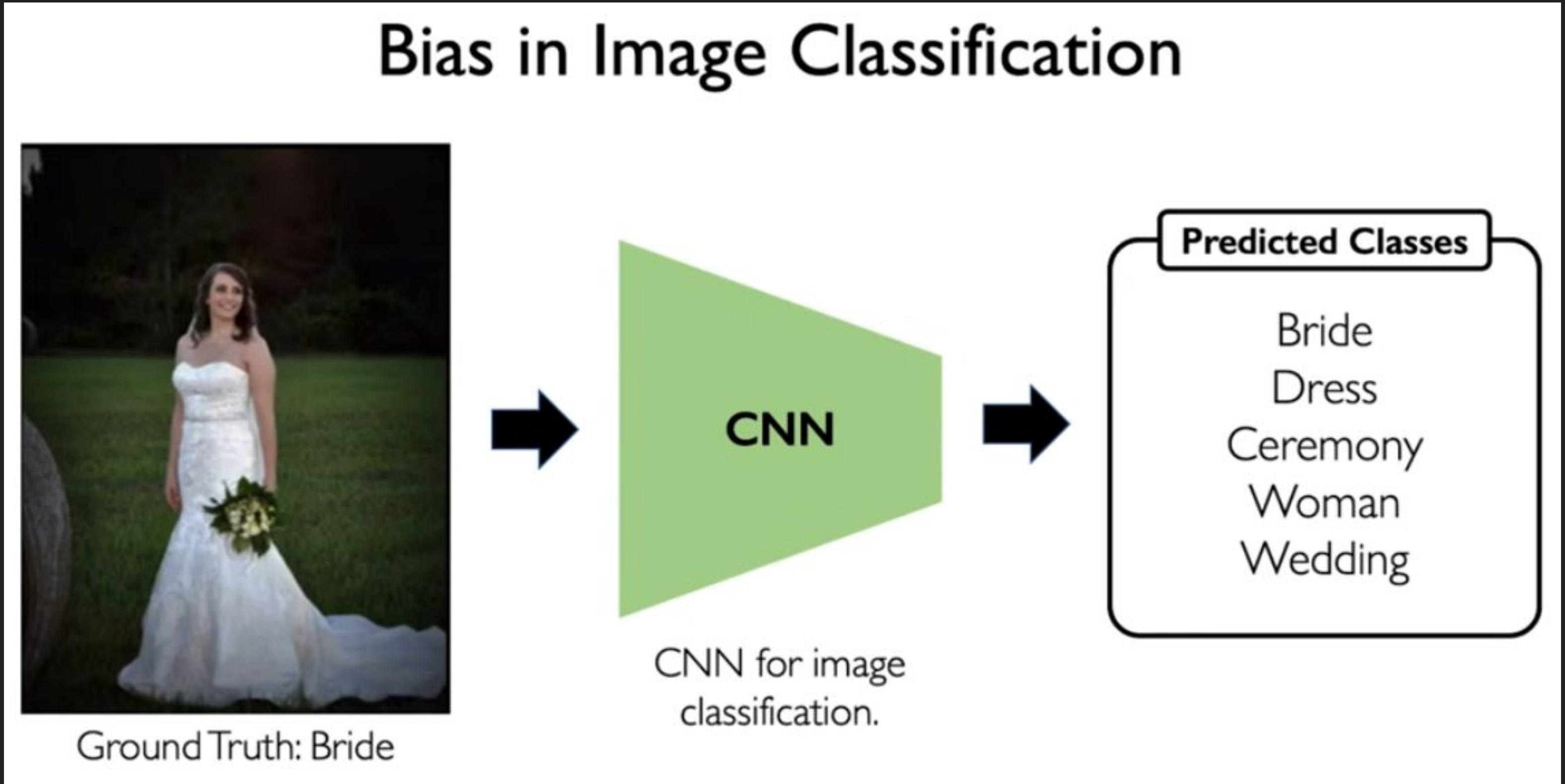
2. **The population of interest has changed or is distinct from the population used during model training.** Data that is representative of Boston, for example, may not be representative if used to analyze the population of Indianapolis. Similarly, data representative of Boston 30 years ago will likely not reflect today's population.

Representation bias

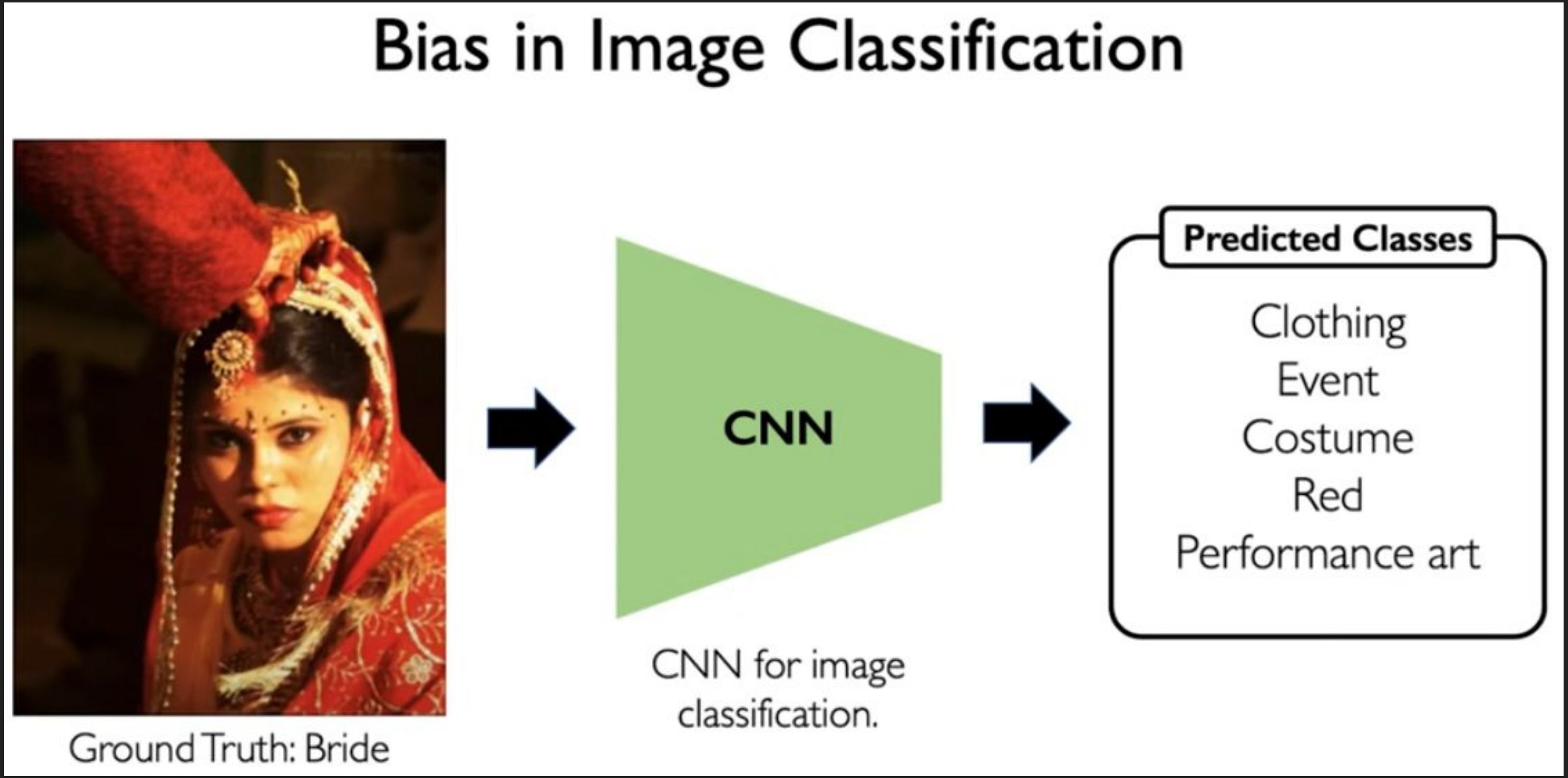
Representation bias



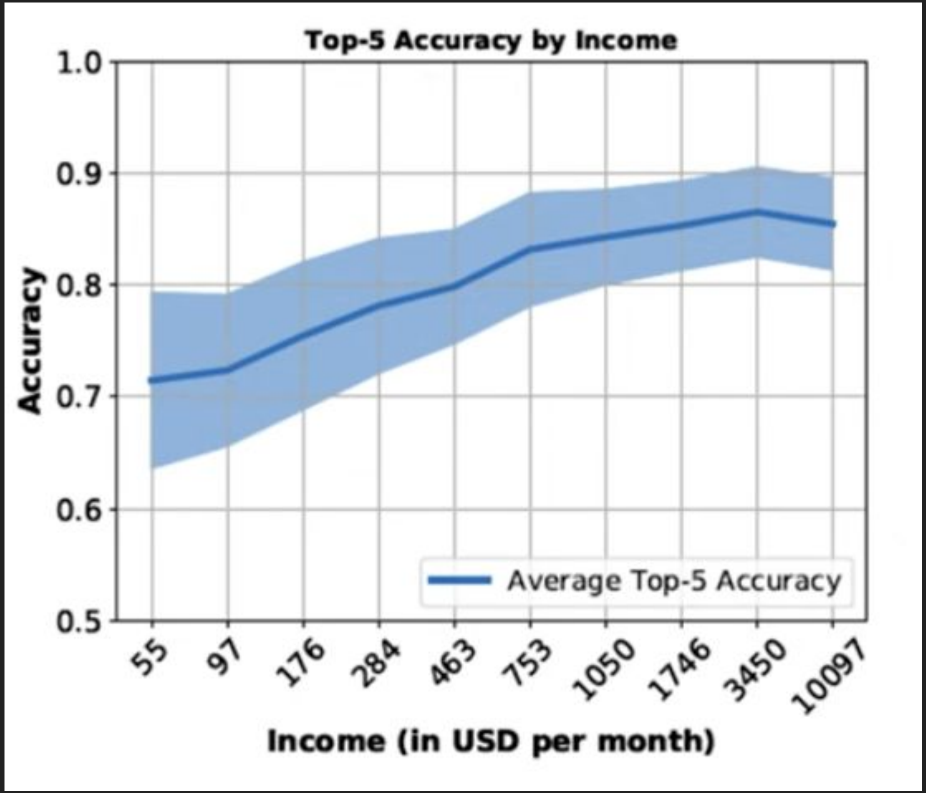
Representation bias

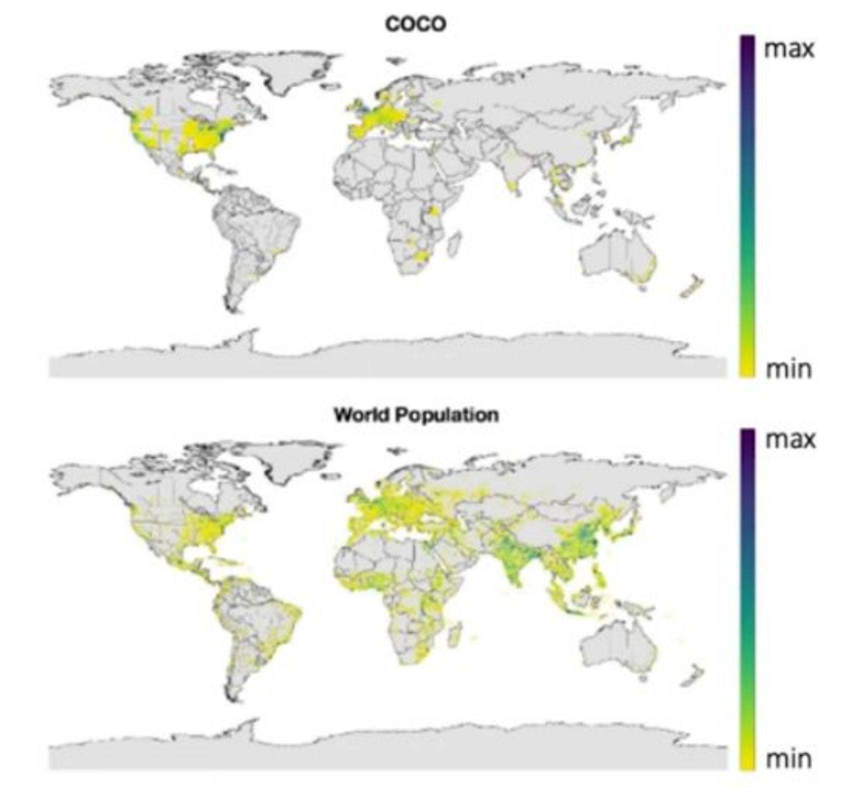


Representation bias



Representation bias





Measurement bias

Measurement Bias arises when choosing and measuring features and labels to use; these are often proxies for the desired quantities. The chosen set of features and labels may leave out important factors or introduce group- or input-dependent noise that leads to differential performance.

1. **The measurement process varies across groups.** For example, if a group of factory workers is more stringently or frequently monitored, more errors will be observed in that group. This can also lead to a feedback loop wherein the group is subject to further monitoring because of the apparent higher rate of mistakes (Barocas and Selbst, 2016;

2. **The quality of data varies across groups.** Structural discrimination can lead to systematically higher error rates in a certain group. For example, women are more likely to be misdiagnosed or not diagnosed for conditions where self-reported pain is a symptom (Calderone, 1990). In this case, "*diagnosed* with condition X" is a biased proxy for "has condition X."

3. **The defined classification task is an oversimplification.** In order to build a supervised ML model, some label to predict must be chosen. Reducing a decision to a single attribute can create a biased proxy label because it only captures a particular aspect of what we really want to measure. Consider the prediction problem of deciding whether a student will be successful (e.g., in a college admissions context). Fully capturing the outcome of 'successful student' in terms of a single measurable attribute is impossible because of its complexity. In cases such as these, algorithm designers resort to some available label such as 'GPA' (Kleinberg et al., 2018), which ignores different indicators of success achieved by parts of the population.

Measurement bias (proxy metrics)

How to determine:

- A job candidate’s fit?
- A student’s potential?
- A patient’s healthcare needs?
- A post’s engagement?

Metrics as proxies

- Prior stroke
- Cardiovascular disease
- Accidental injury
- Benign breast lump
- Colonoscopy
- Sinusitis

Does Machine Learning Automate Moral Hazard and Error?†

By SENDHIL MULLAINATHAN AND ZIAD OBERMEYER*

Gaming/Short-term

- Cancelled scheduled operations to draft extra staff to ER
- Required patients to wait outside the ER, e.g. in ambulances
- Put stretchers in hallways and classified them as "beds"
- Hospital and patients reported different wait times.

WHAT'S MEASURED IS WHAT MATTERS: TARGETS AND GAMING IN THE ENGLISH PUBLIC HEALTH CARE SYSTEM

GWYN BEVAN, CHRISTOPHER HOOD

Manipulation

TECHNOLOGY

How Facebook's Chaotic Push Into Video Cost Hundreds of Journalists Their Jobs

As media companies t
platform, they fired w

ALEXIS C. MADRIGAL AND ROBI

AIs Are Designed to Maximize Watch Time

At YouTube, we used a complex AI to pursue a simple goal: maximize watch time. Google explains this focus in the following statement:

If viewers are watching more YouTube, it signals to us that they're happier with the content they've found. It means that creators are attracting more engaged audiences. It also opens up more opportunities to generate revenue for our partners.

Goodhart's/Campbell's law

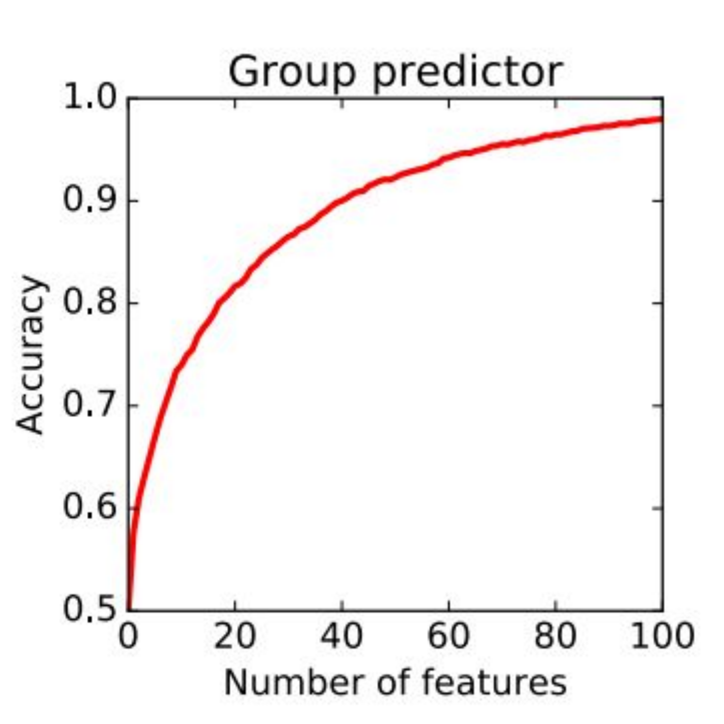
When a measure becomes a target
it ceases to be a good measure

When metrics are useful

- Use multiple metrics instead of a single score
- Listening to qualitative first-person experiences
- Involve domain experts
- Consider biases and what can go wrong beforehand
- Transparent algorithms
- Assess and review your metrics regularly
- Legislative incentives for companies to adhere

Quantifying fairness

No fairness through unawareness



What is fair?

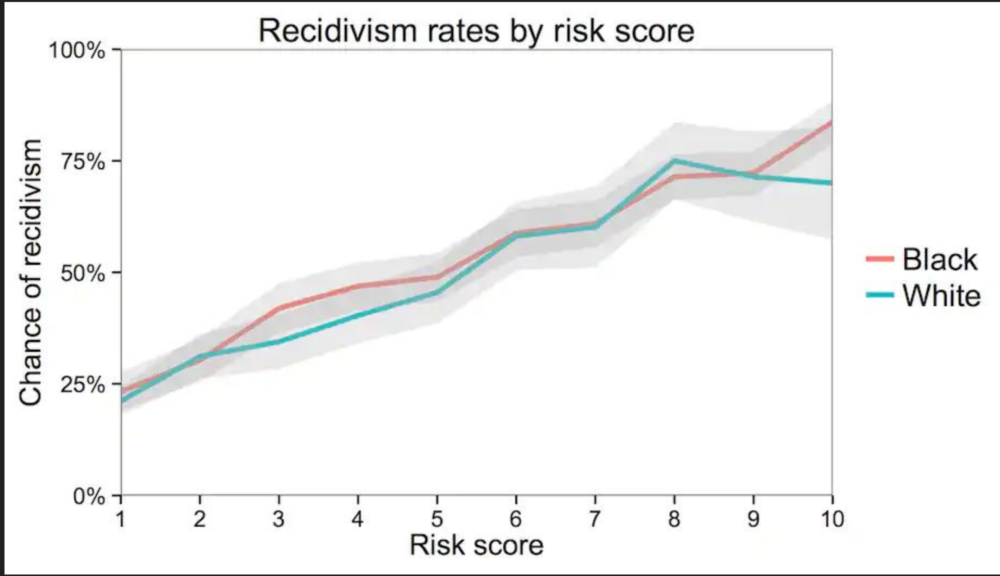
Propublica

COMPAS

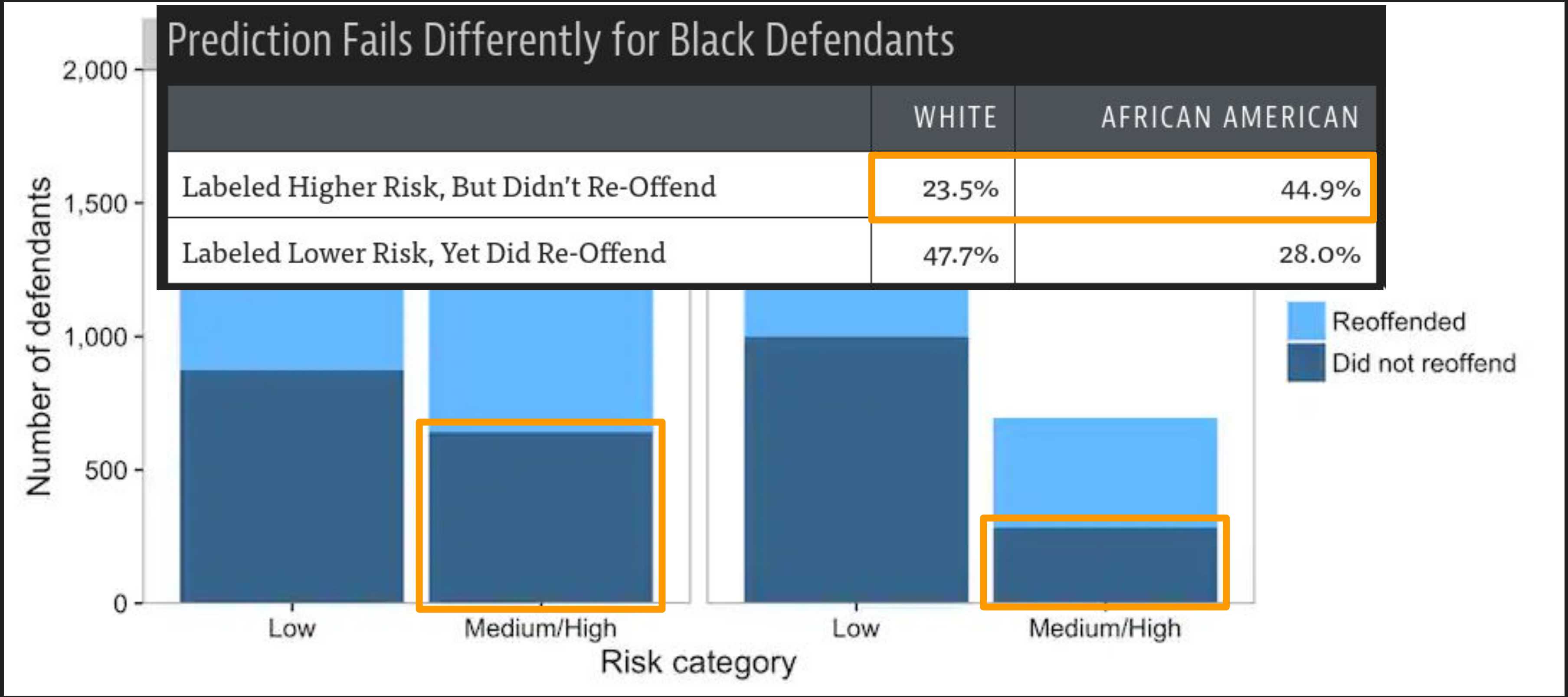
Northpointe

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%



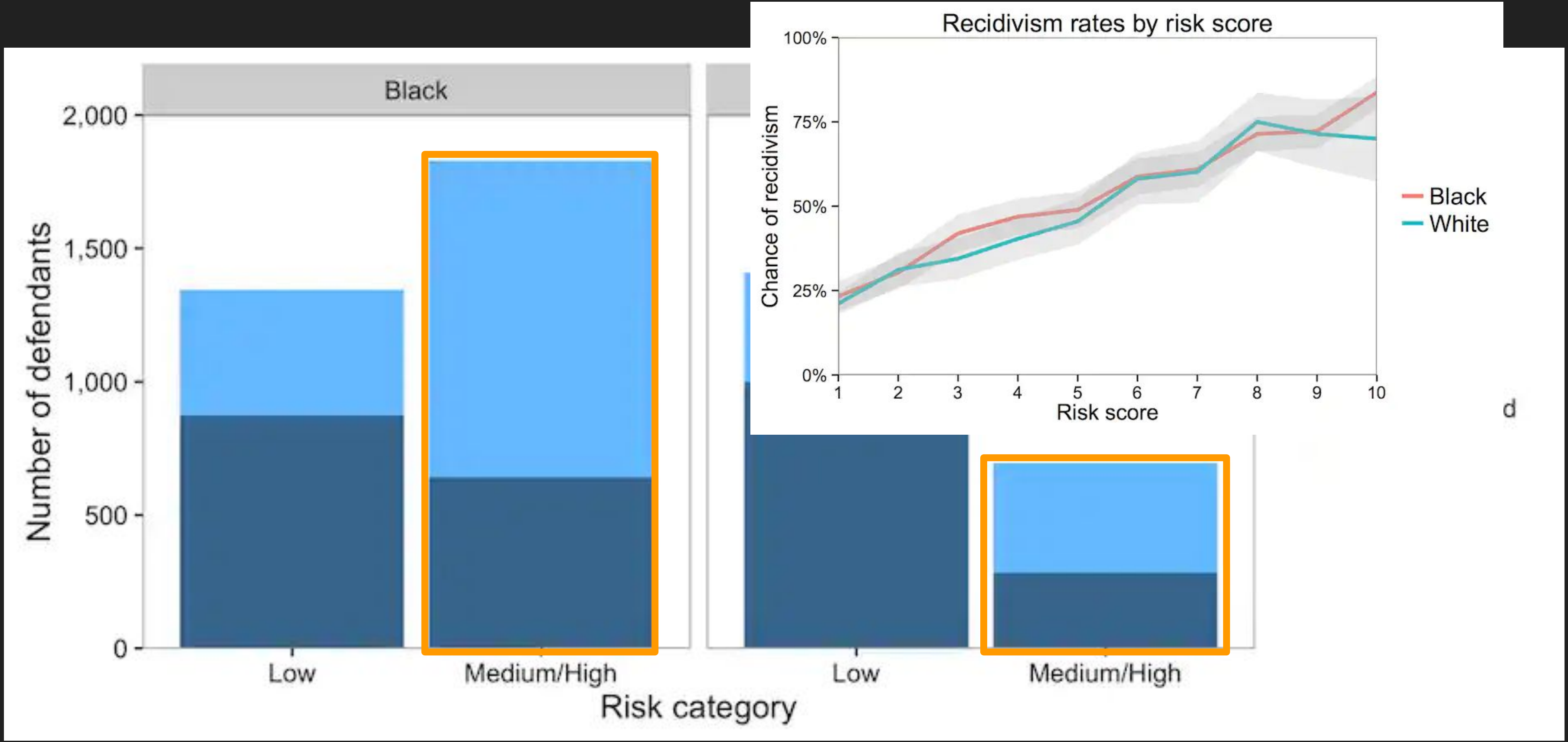
Propublica's fairness metric



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%

Northpoint's fairness metric

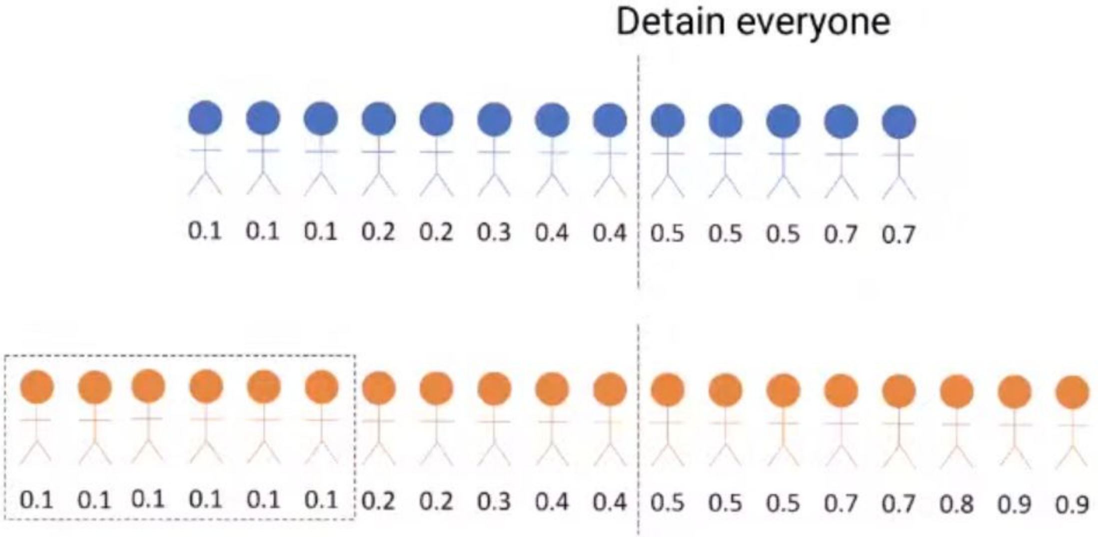


Which fairness metric is correct?

Confusion matrix & fairness metrics

<table>
<tr><td></td><td></td><td colspan="2">True condition</td><td></td><td colspan="2"></td></tr>
<tr><td></td><td>Total population</td><td>Condition positive</td><td>Condition negative</td><td>Prevalence $= \frac{\Sigma \text{ Condition positive}}{\Sigma \text{ Total population}}$</td><td colspan="2">Accuracy (ACC) = $\frac{\Sigma \text{ True positive} + \Sigma \text{ True negative}}{\Sigma \text{ Total population}}$</td></tr>
<tr><td rowspan="2">Predicted condition</td><td>Predicted condition positive</td><td>True positive</td><td>False positive,
Type I error</td><td>Positive predictive value (PPV), Precision = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Predicted condition positive}}$</td><td colspan="2">False discovery rate (FDR) = $\frac{\Sigma \text{ False positive}}{\Sigma \text{ Predicted condition positive}}$</td></tr>
<tr><td>Predicted condition negative</td><td>False negative,
Type II error</td><td>True negative</td><td>False omission rate (FOR) = $\frac{\Sigma \text{ False negative}}{\Sigma \text{ Predicted condition negative}}$</td><td colspan="2">Negative predictive value (NPV) = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Predicted condition negative}}$</td></tr>
<tr><td></td><td></td><td>True positive rate (TPR), Recall, Sensitivity, probability of detection, Power = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$</td><td>False positive rate (FPR), Fall-out, probability of false alarm $= \frac{\Sigma \text{ False positive}}{\Sigma \text{ Condition negative}}$</td><td>Positive likelihood ratio (LR+) $= \frac{\text{TPR}}{\text{FPR}}$</td><td rowspan="2">Diagnostic odds ratio (DOR) $= \frac{\text{LR+}}{\text{LR−}}$</td><td rowspan="2">$F_1$ score = $2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$</td></tr>
<tr><td></td><td></td><td>False negative rate (FNR), Miss rate = $\frac{\Sigma \text{ False negative}}{\Sigma \text{ Condition positive}}$</td><td>Specificity (SPC), Selectivity, True negative rate (TNR) $= \frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$</td><td>Negative likelihood ratio (LR−) $= \frac{\text{FNR}}{\text{TNR}}$</td></tr>
</table>

Gaming fairness metrics



Detention rate	False pos. rate
38%	25%
~~61%~~ 42%	~~42%~~ 26%

Arrest more low risk individuals in orange group!

Find the underlying problem

Failure to appear in court

One approach: Predict failure to appear, jail if risk is high.

Alternative: Recognize that people fail to appear in court due to lack of child care and transportation, work schedules, or too many court appointments. Implement steps to mitigate these issues.

Alternative is part of the Harris County Lawsuit settlement: *"require Harris County to provide free child care at courthouses, develop a two-way communication system between courts and defendants, give cell phones to poor defendants and pay for public transit or ride share services for defendants without access to transportation to court."* (Source: Houston Chronicle, April 2019)

Take-home messages

- Rather than trying to understand **IF** your model is fair, try to understand **HOW** it is unfair.
- Look at multiple fairness metrics to diagnose potential issues among diverse stakeholders.
- Be careful when optimizing on fairness criteria.
- Use domain expertise to try to understand causal relationships underlying the observed results.