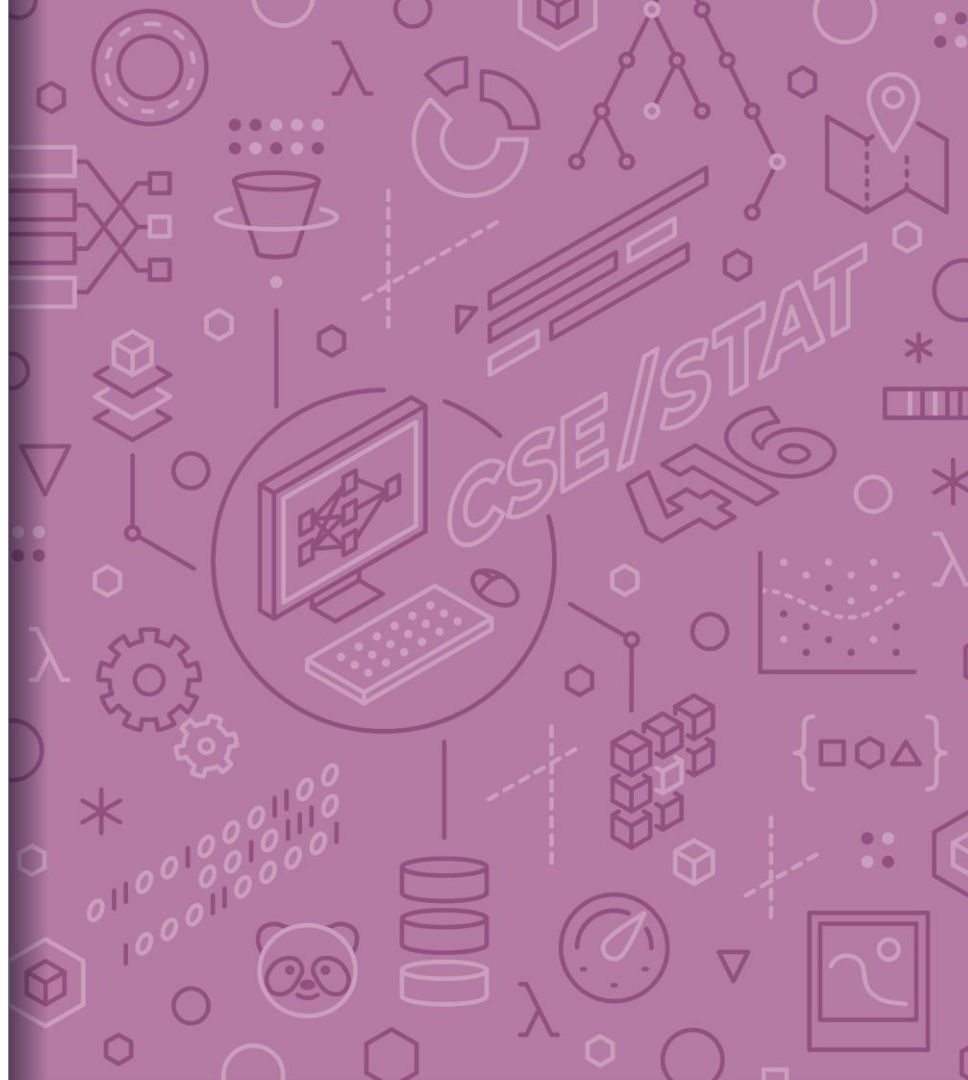


CSE/STAT 416

Bias and Fairness in ML

Tanmay Shah
Paul G. Allen School of Computer Science & Engineering
University of Washington

April 17, 2024



Administrivia


- HW2 due Tomorrow
- LR3 out Friday
- Midterm
 - Released on gradescope upcoming Monday, due Wednesday
 - Timed
 - Make Ed post for any questions



ML and Society

ML Systems Gone Wrong



INDEPENDENT
**GOOGLE'S ALGORITHM SHOWS
PRESTIGIOUS JOB ADS TO MEN,
BUT NOT TO WOMEN**



REUTERS Business Markets World Politics TV More

BUSINESS NEWS OCTOBER 9, 2018 / 8:12 PM / 5 MONTHS AGO

Amazon scraps secret AI recruiting tool that showed bias against women

Jeffrey Dastin 8 MIN READ  

The New York Times

Facebook Engages in Housing Discrimination With Its Ad Practices, U.S. Says

By Katie Benner, Glenn Thrush and Mike Isaac

March 28, 2019       168

MIT Technology Review

Intelligent Machines

How to Fix Silicon Valley's Sexist Algorithms

Computers are inheriting gender bias implanted in language data sets—and not everyone thinks we should correct it.

PRO PUBLICA

Machine Bias

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

COMPAS

An ML model created by NorthPointe used to predict likelihood of inmates to “recidivate”. Eventually started use in Florida in judges’ decision for parole

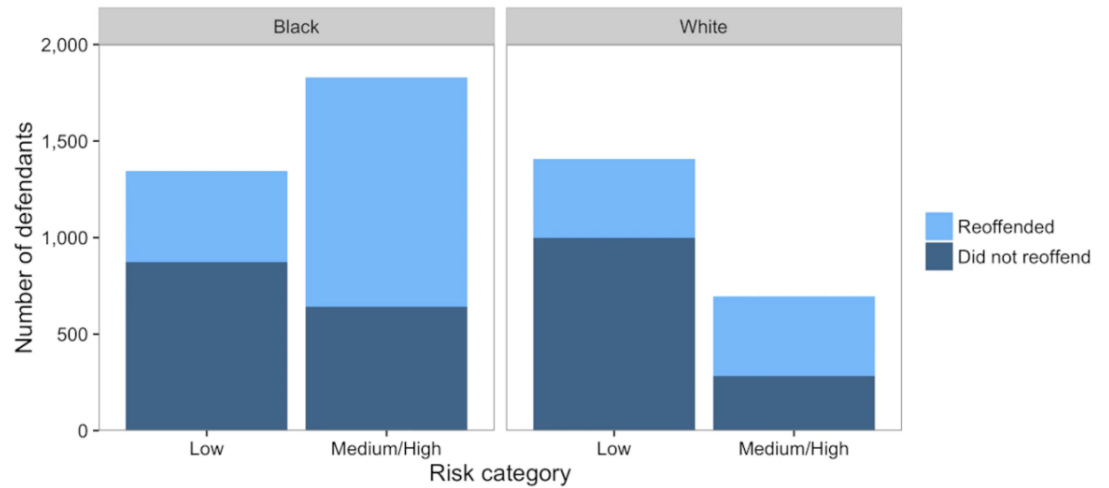
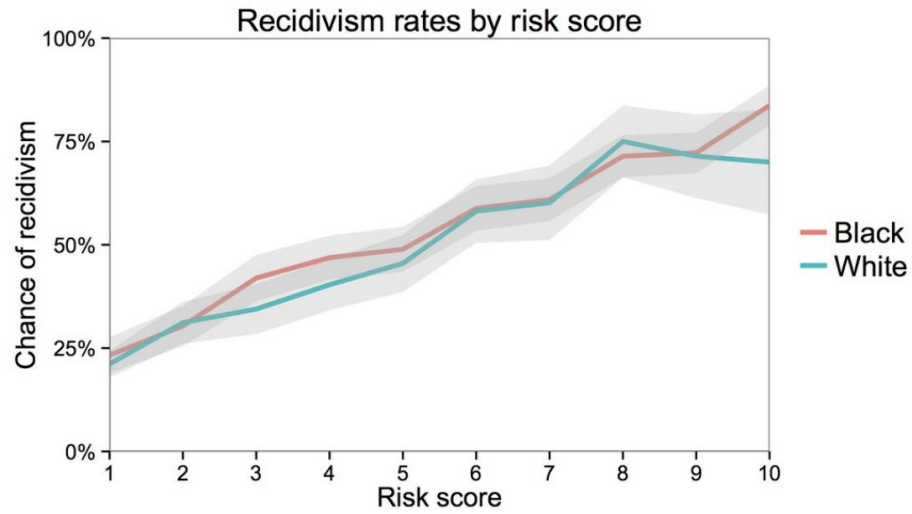
ProPublica (a news org) investigated the model and [wrote](#) that the model exhibited biased behavior against people of color. Particularly, they found that the model would predict higher risk scores for black people.

Northpointe [countered](#) and claimed that their scores were well **calibrated** (e.g., when the predict score of 9/10 that person recidivates about 90% of the time).

- Interesting [follow up](#) from ProPublica

So the question is: Who is right? Is it right to use this model?

COMPAS



Why Biased Outcomes?

Probably not the case that someone explicitly coded the model to be biased against a particular race. In fact, race was not even a question that was on the survey inmates took!

More often than not, biased outcomes from a model come from **the data it learns from** rather than some explicit choice from the modeler.

- “Garbage in → Garbage out”
- “Bias in → Bias out”



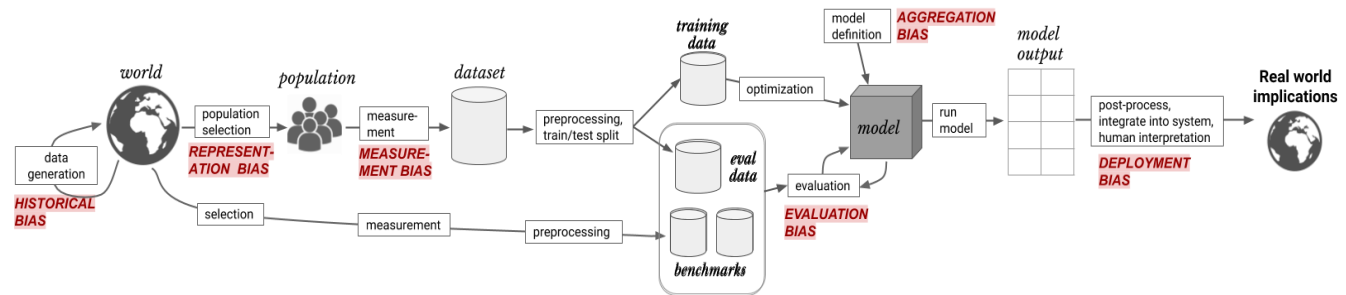
Sources of Bias

Sources of Bias

Discussion heavily based on Suresh and Guttag (2020)

Six common sources of bias:

- Historical bias
- Representation Bias
- Measurement Bias
- Aggregation Bias
- Evaluation Bias
- Deployment Bias



[A FRAMEWORK FOR UNDERSTANDING UNINTENDED CONSEQUENCES OF MACHINE LEARNING](#), BY HARINI SURESH AND JOHN V. GUTTAG, 2020

Historical Bias

The world we lived in is one that contains biases for/against certain demographics. Even 'accurate' data could still be harmful.

- Historical bias exists even with perfect sampling or feature measurement (other sources of bias are possible)!

Examples:

- In 2018, 5% of Fortune 500 CEOs were women. Should search results for "CEO" match this statistic? Could reflecting the world (even if accurately) perpetuate more harm?



Representation Bias

When the *training data* we collect does not contain representative samples of the true distribution.

Examples:

- If we use data gathered from smart phones, we would likely be underestimating poorer and older populations.
- ImageNet (a very popular image dataset) with 1.2 million images. About 45% of these images were taken in the US and the majority of the rest in North America and Western Europe. Only about 1% and 2.1% of the images come from China and India respectively.



Measurement Bias

Often we are gathering data that contains (noisy) proxies of characteristics of interest. Some examples:

- Financial responsibility → Credit Score
- Crime Rate → Arrest Rate
- Intelligence → SAT Score

If these measurements are not measured equally across groups or places (or aren't relevant to the task at hand), this can be another source of bias.



Measurement Bias (cont.)

Examples:

- If factory workers are monitored more often, more errors are spotted. This can result in a **feedback loop** to encourage more monitoring in the future.
 - Same principles at play with predictive policing. Minoritized communities were more heavily policed in the past, which causes more instances of documented crime, which then leads to more policing in the future.
- Women are more likely to be misdiagnosed (or not diagnosed) for conditions where self-reported pain is a symptom. In this case aspect of our data “diagnosed with X” is a biased proxy for “has condition X”.
- The feature we measure is a poor representation of the quality of interest (e.g., SAT score doesn’t actually measure intelligence)

Aggregation Bias

When we use a “one-sized fits all” model that does not accurately serve every group equally.

Examples:

- HbA1c levels (used to monitor and diagnose diabetes) differ in very complex ways across ethnicities and sexes. One model for everyone might not be the right choice, even if everyone is represented well in the training data.



Evaluation Bias

Similar to representation bias, but focused more on the data we evaluate or test ourselves against. If the evaluation dataset or benchmark doesn't represent the world well, we have evaluation bias.

- **Benchmarks** are common datasets used to evaluate models from different researchers.

Examples:

- If it is common to report accuracy on a benchmark, this might hide disparate performance on subgroups.
- Drastically worse performance for facial recognition software when used on faces of darker-skinned females. Common evaluation datasets for facial recognition only had 5-7% had faces of darker-skinned women.

Deployment Bias

When how a model was intended to be used and how it is actually used when deployed in the real-world.

Examples:

- Crime risk prediction models might be evaluated to achieve good calibration, but the model designers might not have evaluated the model's use in the context of determining prison sentence lengths.
- People are complex and when using models to aid their decisions, might make incorrect assumptions about what a model says.

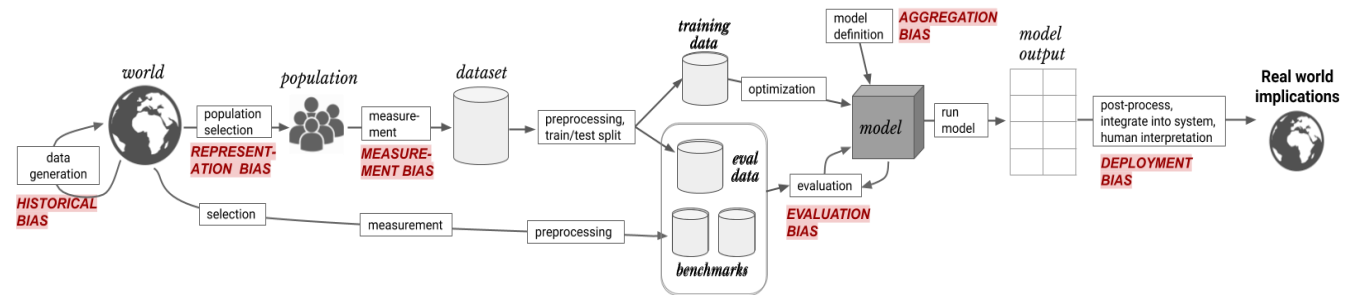


Sources of Bias

Discussion heavily based on Suresh and Guttag (2020)

Six common sources of bias:

- Historical bias
- Representation Bias
- Measurement Bias
- Aggregation Bias
- Evaluation Bias
- Deployment Bias



[A FRAMEWORK FOR UNDERSTANDING UNINTENDED CONSEQUENCES OF MACHINE LEARNING](#), BY HARINI SURESH AND JOHN V. GUTTAG, 2020

Brain Break



Fairness in ML

Fairness

What does it mean for a model to be fair or unfair? Can we come up with a numeric way of measuring fairness?

Lots of work in the field of ML and fairness is looking into mathematical definitions of fairness to help us spot when something might be unfair.

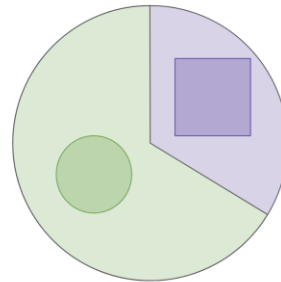
- There is not going to be one central definition of fairness, as each definition is a mathematical statement of which behaviors are/aren't allowed.
- Different definitions of fairness can be contradictory!

Today, we will focus on notions of **group fairness** in an attempt to prevent discriminatory outcomes.

Example: College Admissions

Will use a very simplified example of college admissions. This is **not** an endorsement of such a system or a statement of how we think the world does/should work. Will make MANY simplifying assumptions (which are unrealistic).

- There is a single definition of “success” for college applicants, and the goal of an admissions decision is to predict “success”
- The only thing we will use as part of our decision is SAT Score
- To talk about group fairness, will assume everyone belongs to exactly one of two races: Circles (66%) or Squares (33%).



Notation

Example: College admission only using SAT Score

X input about a person for prediction

- Example: $X = \text{SAT Score}$

A variable indicating which group X belongs in

- Example: $A = \square$ or $A = \bigcirc$

Y the “true label”

- Example: $Y = +$ if truly successful in college, $Y = -$ if not

$\hat{Y} = \hat{f}(X)$ is our prediction for Y using a learned model \hat{f}

- Example: $\hat{Y} = +$ if predicted successful, $\hat{Y} = -$ otherwise



Fairness

Definition 1: “Shape Blind”

To avoid unfair decisions, prevent the model from every looking at protected attribute (e.g., if the applicant is Circle/Square).

Often called “**Fairness through unawareness**”

Doesn’t work in practice. This does not prevent historical or measurement bias. Protected attributes can be unintentionally inferred from other, related attributes (e.g., in some cities, zip code can be deeply correlated with race).



Confusion Matrix

For binary classification, there are only two types of mistakes

- $\hat{y} = +1, y = -1$
- $\hat{y} = -1, y = +1$

Generally we make a **confusion matrix** to understand mistakes.

		Predicted Label	
		+	-
True Label	+	True Positive (TP)	False Negative (FN)
	-	False Positive (FP)	True Negative (TN)

Binary Classification Measures

Notation

- $C_{TP} = \#TP$, $C_{FP} = \#FP$, $C_{TN} = \#TN$, $C_{FN} = \#FN$
- $N = C_{TP} + C_{FP} + C_{TN} + C_{FN}$
- $N_P = C_{TP} + C_{FN}$, $N_N = C_{FP} + C_{TN}$

Error Rate

$$\frac{C_{FP} + C_{FN}}{N}$$

Accuracy Rate

$$\frac{C_{TP} + C_{TN}}{N}$$

False Positive rate (FPR)

$$\frac{C_{FP}}{N_N}$$

False Negative Rate (FNR)

$$\frac{C_{FN}}{N_P}$$

True Positive Rate or Recall

$$\frac{C_{TP}}{N_P}$$

Precision

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

F1-Score

$$2 \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

[See more!](#)

Fairness

Definition 2: Statistical Parity

Idea: “Admit decisions are equivalent across groups”

$$\Pr(\hat{Y} = + | A = \square) = \Pr(\hat{Y} = + | A = \circ)$$

Also phrased as matching demographic statistics (e.g., if 33% of population are Squares, 33% of those admitted should be Square).

Pros:

- Aligns with certain legal definitions of equity.

Cons:

- A rather weak in fairness requirements. Allows for strategies that might not be desirable (e.g., random selection, self-fulfilling prophecy)

Fairness

Definition 3: Equal Opportunity

Idea: True positive rate should be equivalent across groups

$$\Pr(\hat{Y} = + | A = \square, Y = +) = \Pr(\hat{Y} = + | A = \circ, Y = +)$$

Pros:

- Better controls for true outcome

Cons:

- More complex to explain to non-experts
- Only protects for the positive outcome

Note: Equality of true positives is the same as equality of false negatives

Fairness

Definition 4: Predictive equality

Idea: True negative rate should be equivalent across groups

$$\Pr(\hat{Y} = - | A = \square, Y = -) = \Pr(\hat{Y} = - | A = \circ, Y = -)$$

Same idea as equal opportunity, but controlling for different statistic. Might be favorable in situations you care more about false positives than a false negative.

Note: Equality of true negatives is the same as equality of false positives

And many,
many more

List of demographic fairness criteria			
Name	Closest relative	Note	Reference
Statistical parity	Independence	Equivalent	Dwork et al. (2011)
Group fairness	Independence	Equivalent	
Demographic parity	Independence	Equivalent	
Conditional statistical parity	Independence	Relaxation	Corbett-Davies et al. (2017)
Darlington criterion (4)	Independence	Equivalent	Darlington (1971)
Equal opportunity	Separation	Relaxation	Hardt, Price, Srebro (2016)
Equalized odds	Separation	Equivalent	Hardt, Price, Srebro (2016)
Conditional procedure accuracy	Separation	Equivalent	Berk et al. (2017)
Avoiding disparate mistreatment	Separation	Equivalent	Zafar et al. (2017)
Balance for the negative class	Separation	Relaxation	Kleinberg, Mullainathan, Raghavan (2016)
Balance for the positive class	Separation	Relaxation	Kleinberg, Mullainathan, Raghavan (2016)
Predictive equality	Separation	Relaxation	Chouldechova (2016)
Equalized correlations	Separation	Relaxation	Woodworth (2017)
Darlington criterion (3)	Separation	Relaxation	Darlington (1971)
Cleary model	Sufficiency	Equivalent	Cleary (1966)
Conditional use accuracy	Sufficiency	Equivalent	Berk et al. (2017)
Predictive parity	Sufficiency	Relaxation	Chouldechova (2016)
Calibration within groups	Sufficiency	Equivalent	Chouldechova (2016)
Darlington criterion (1), (2)	Sufficiency	Relaxation	Darlington (1971)

Which one to use?

We can't tell you! Each definition makes its own statement on what fairness means. Choosing a fairness measure is an explicit statement of what values we hold when thinking about fairness.

Takeaway: Discrimination in ML models is a crucial problem we need to work on. It's not a problem that will only be solved algorithmically. We need people (e.g., policymakers, regulators, philosophers, developers) to be in the loop to determine the values we want to encode into our systems.



Next time

Will discuss some limitations in these definitions (particularly how they contradict) and how we can think about fairness as a philosophy (or worldview).



Recap

Theme: It's important to give terms to abstract notions like bias and fairness so we can have concrete things to look out for. There is not one right perspective though!

Ideas:

- Calibration
- Impacts of ML Systems on society
- Sources of bias
 - Historical bias
 - Representation Bias
 - Measurement Bias
 - Aggregation Bias
 - Evaluation Bias
 - Deployment Bias
- Definitions of fairness
 - Fairness through unawareness
 - Statistical parity
 - Equal opportunity
 - Predictive equality