

CSE/STAT 416

Bias and Fairness in ML

Pemi Nguyen

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

Slides by Hunter Schafer

Images were from MIT course 6.S191: AI Bias and Fairness lecture (2021)

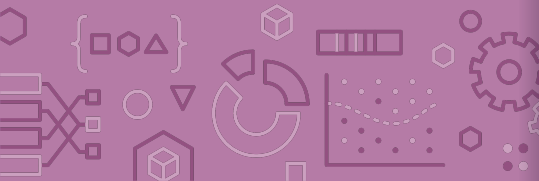
April 25, 2022



Class Logistics

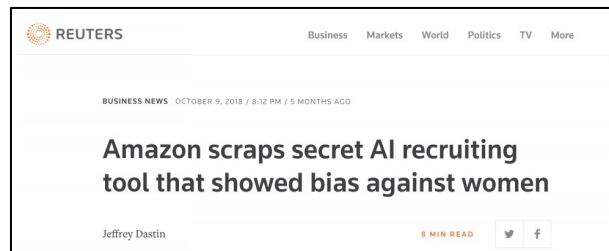
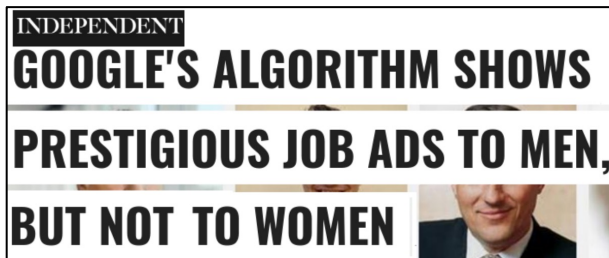
Class logistics

Information regarding week 5 and week 6 logistics has been posted on Ed. Please make sure you check it out



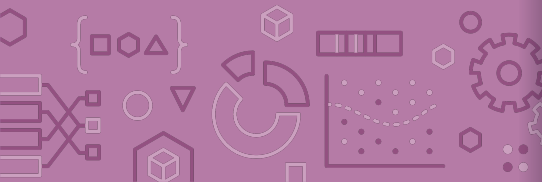
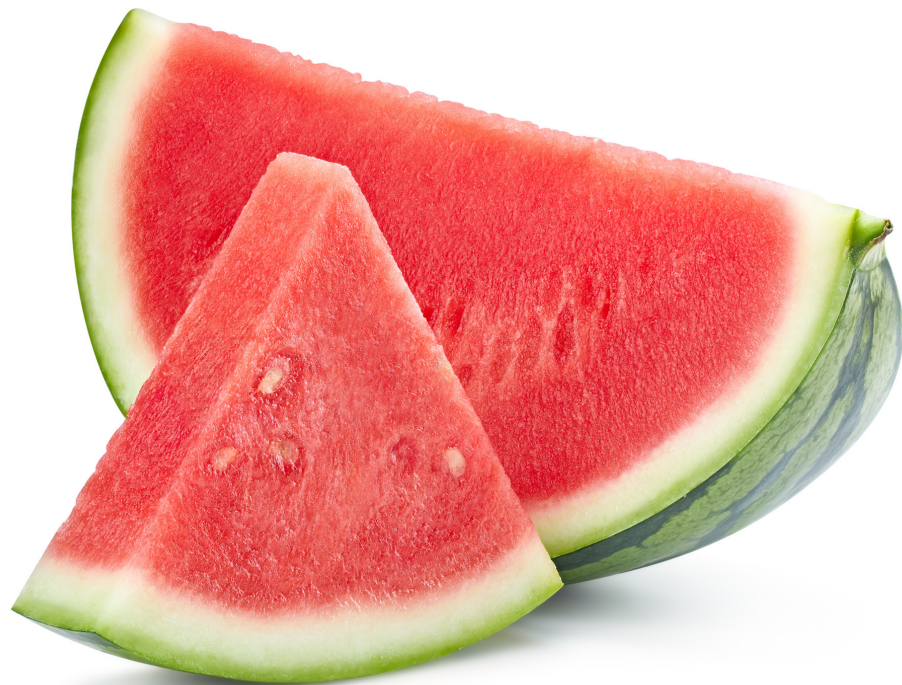
ML and Society

ML Systems Gone Wrong

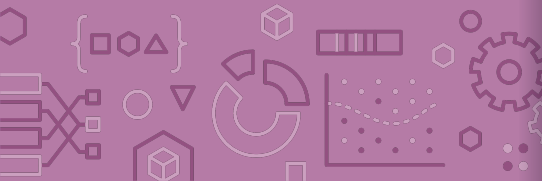
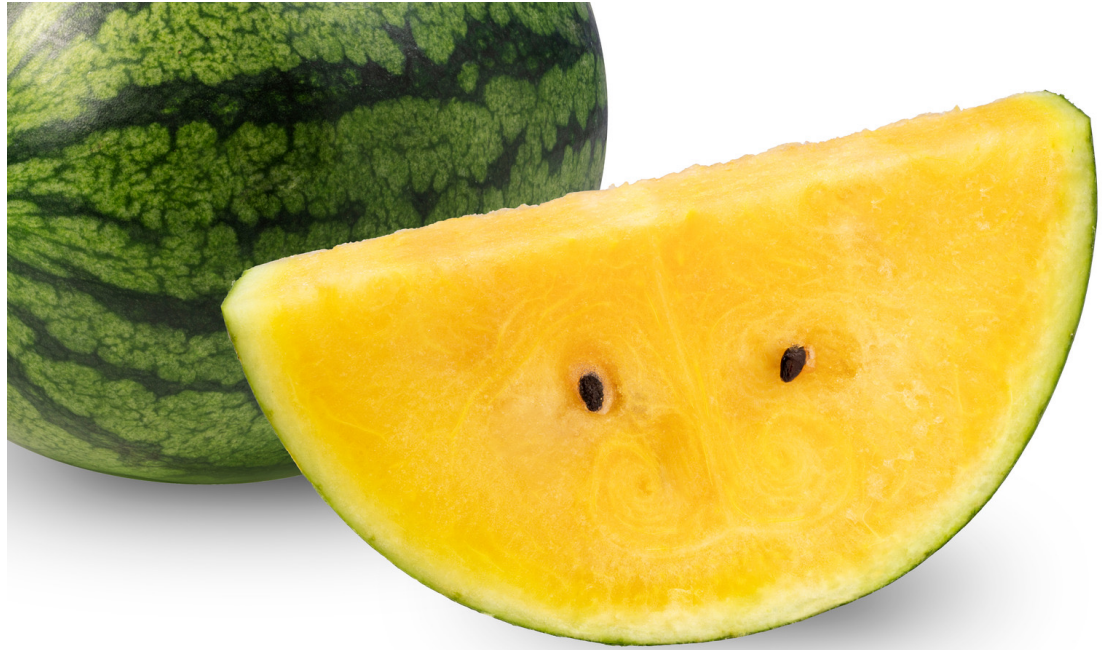


Examples of Biases

What is in
this image?



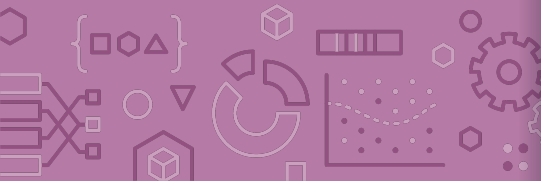
What about
this one?



Humans have biases

We don't call a red watermelon "red watermelon", because we're exposed to it frequently, and it's a typical representation ingrained in our mind. However, "yellow watermelon" is not as common.

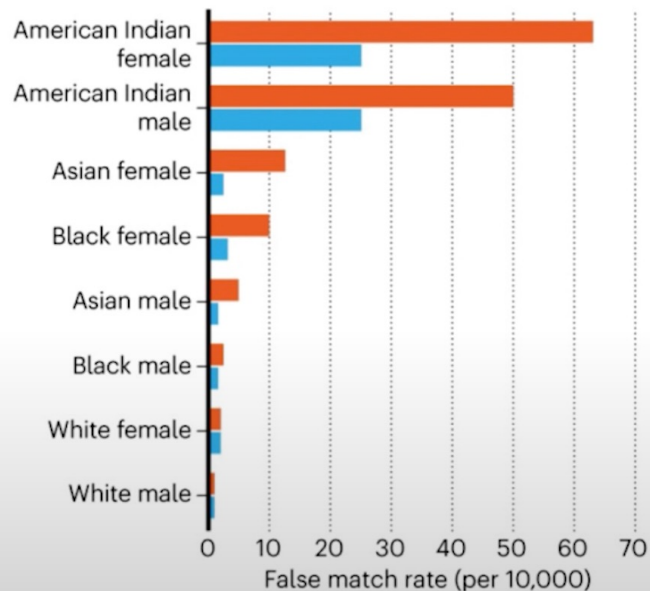
Humans have the tendency to label and categorize complex sensory inputs into simplified groups, and pay attention to atypical things. **Biases** and **stereotypes** arise when we are confronted by confounding choices.





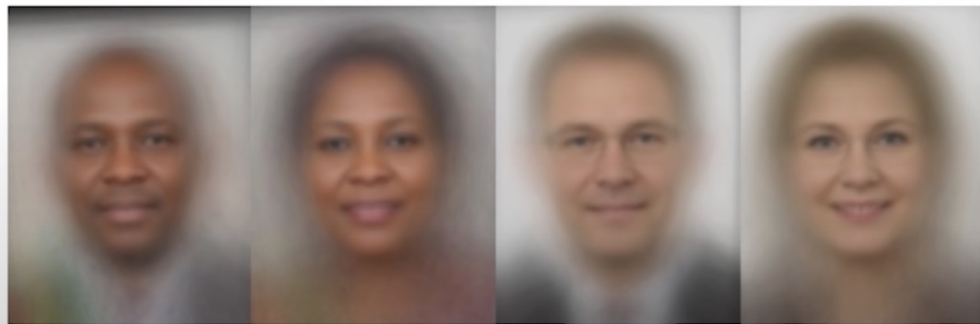
Example: Bias in Facial detection

Independent Study II

■ UK academic algorithm
■ Chinese commercial algorithm

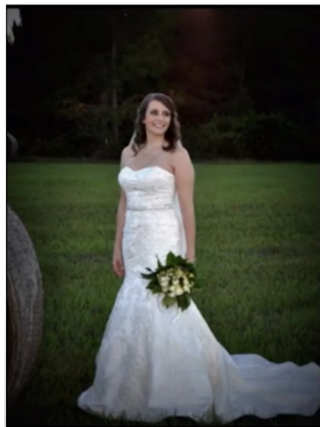


Gender Classifier	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
 Microsoft	94.0% <div><div></div></div>	79.2% <div><div></div></div>	100% <div><div></div></div>	98.3% <div><div></div></div>	20.8% <div><div></div></div>
 FACE++	99.3% <div><div></div></div>	65.5% <div><div></div></div>	99.2% <div><div></div></div>	94.0% <div><div></div></div>	33.8% <div><div></div></div>
 IBM	88.0% <div><div></div></div>	65.3% <div><div></div></div>	99.7% <div><div></div></div>	92.9% <div><div></div></div>	34.4% <div><div></div></div>

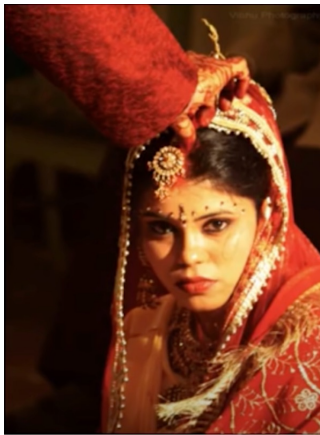


Pictures of a bride from a state-of-the-art image classifier

Example: Bias in Image Classification

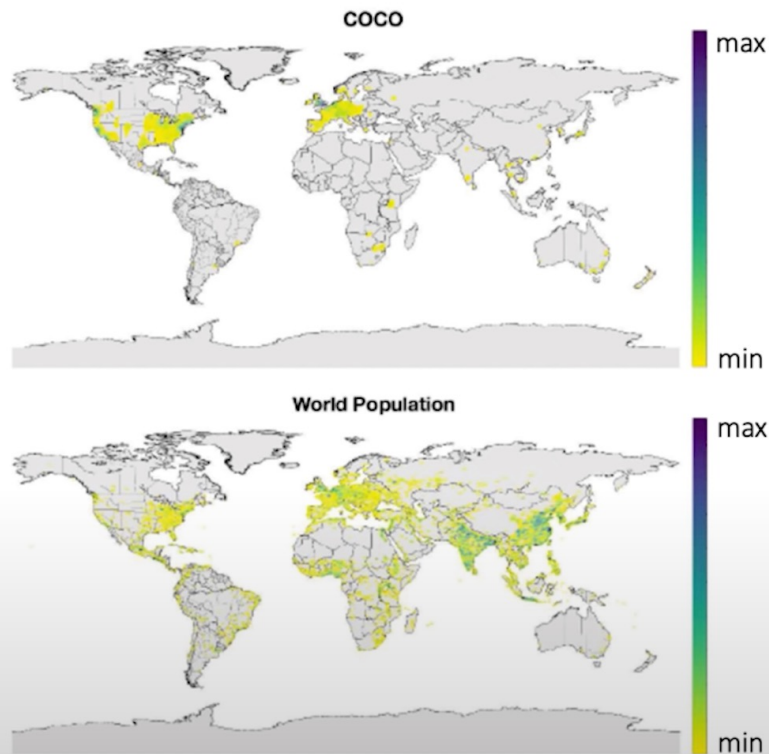
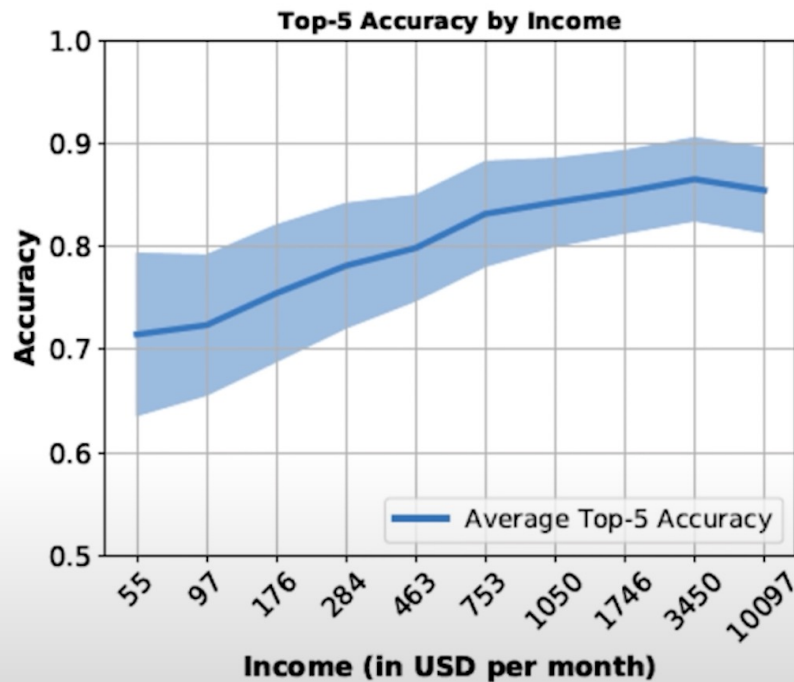


Predicted classes: Brides, dress, ceremony



Predicted classes: Clothing, event, costume

Bias correlation with income and geography



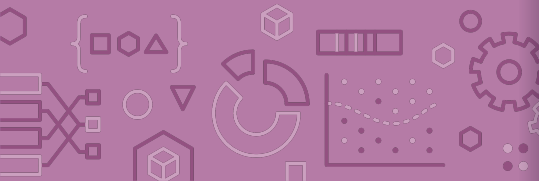
Bias from historical data

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)!

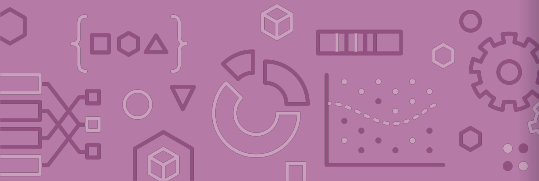
Examples:

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?



Bias in all stages in the ML pipeline

1. Data: Class imbalance from training datasets, data doesn't reflect true distribution
2. Model: Biased, unclear quality metrics, lack of interpretability
3. Training and deployment: Feedback loops that further perpetuate biases
4. Evaluation: Lack of careful analysis of the subgroups
5. Interpretation: Human errors and biased perspectives that can affect results



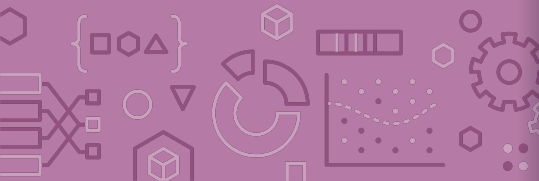
Common Biases

Data-driven biases

- Selection Bias: Data doesn't reflect randomization
- Sampling Bias: Examples of certain features are sampled more than others

Interpretation-driven biases

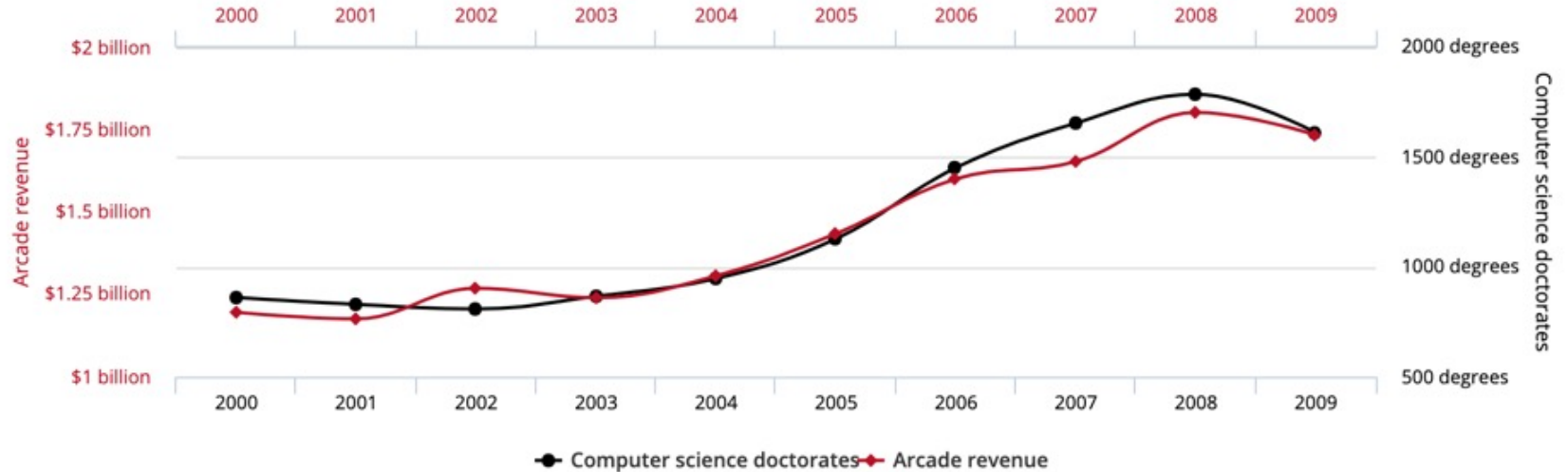
- Correlation Fallacy (Correlation \neq Causation)
- Evaluation bias (Data in the real world is much different from limited training data)
- Measurement bias



Correlation Fallacy

Total revenue generated by arcades
correlates with
Computer science doctorates awarded in the US

Correlation: 98.51% ($r=0.985065$)

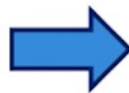


Data sources: U.S. Census Bureau and National Science Foundation

tylervigen.com

Evaluation bias

Expectation:
Cups in my dataset



Reality:
Cups from many angles



Evaluation Bias

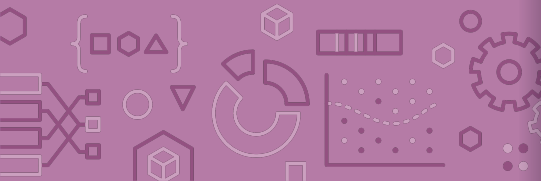
If the evaluation dataset or benchmark doesn't represent the world well, we have evaluation bias.

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


















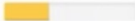
Examples:

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.



Evaluation Bias

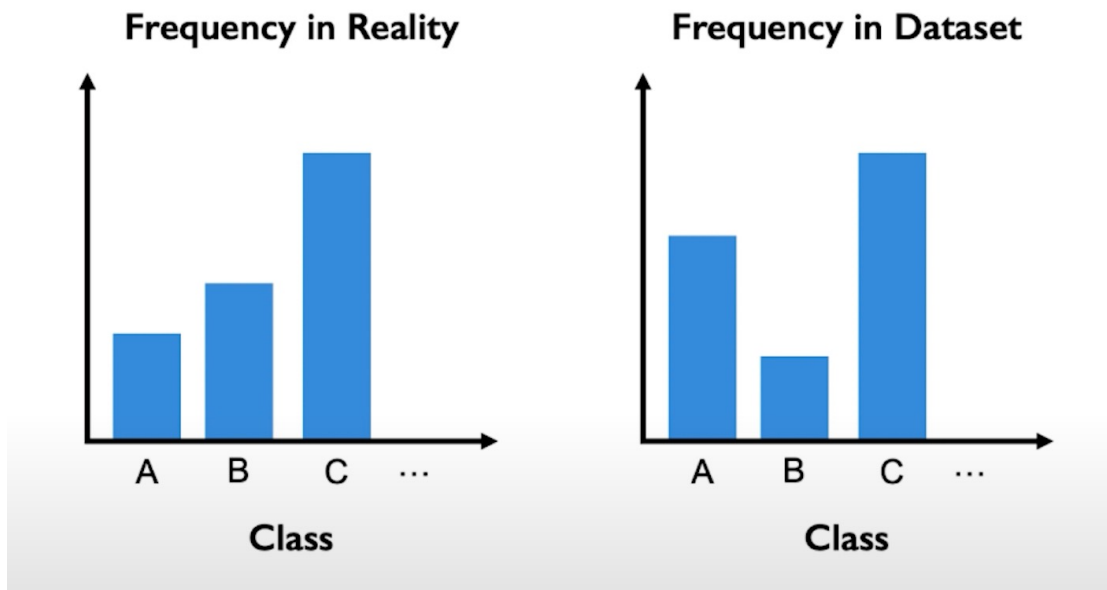
Gender Classifier	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
 Microsoft	94.0% 	79.2% 	100% 	98.3% 	20.8% 
 FACE++	99.3% 	65.5% 	99.2% 	94.0% 	33.8% 
 IBM	88.0% 	65.3% 	99.7% 	92.9% 	34.4% 



Bias due to class imbalance

Class imbalance will lead proportions of test accuracies in each class in the test set close to these of train accuracies

However, we want proportions of test accuracies in each class in the test set to be **equal**

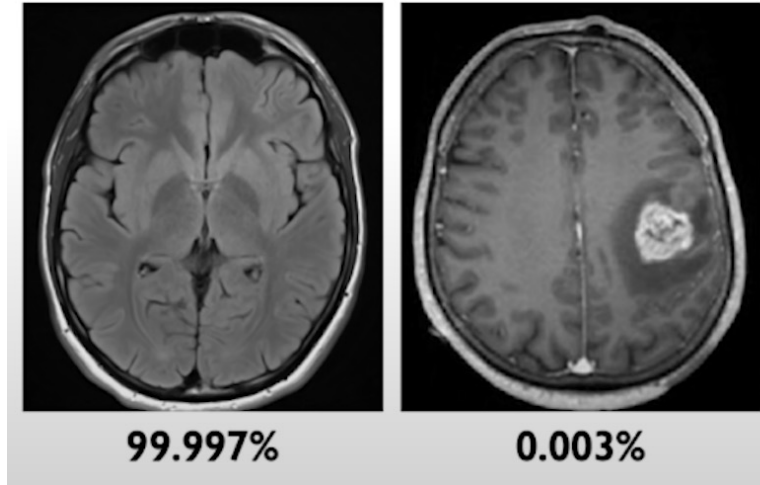
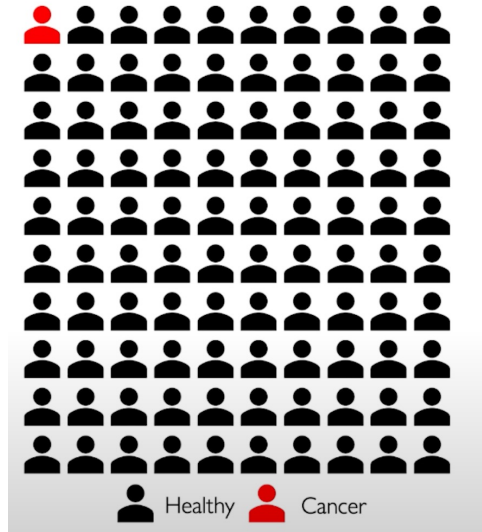


Danger of class imbalance in medical diagnosis

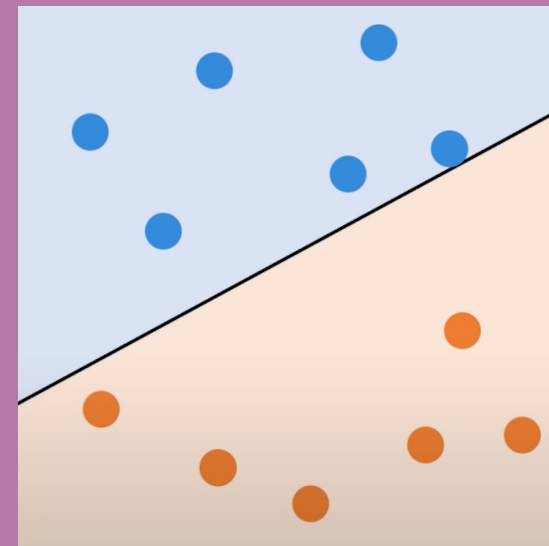
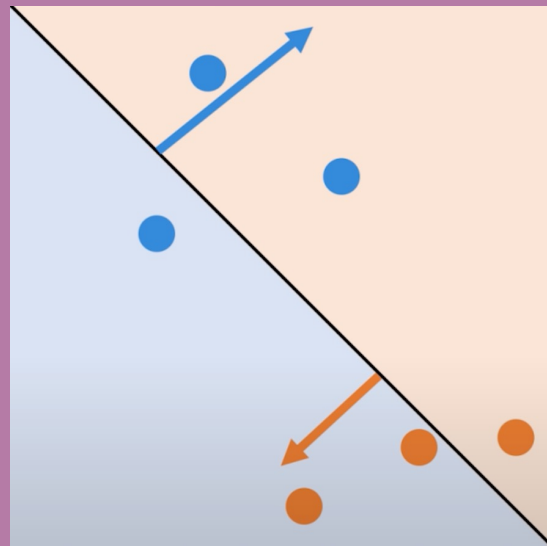
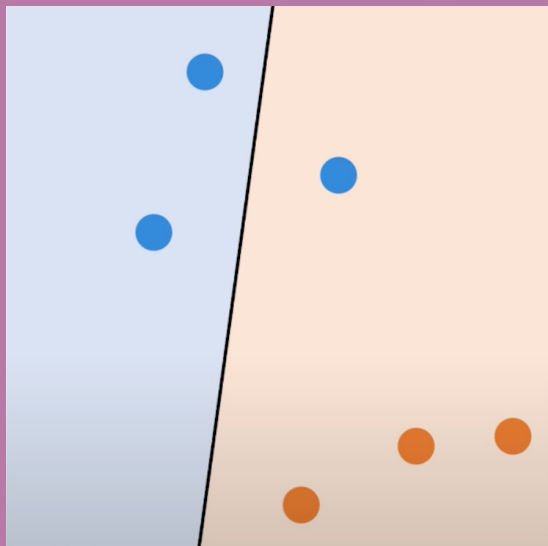
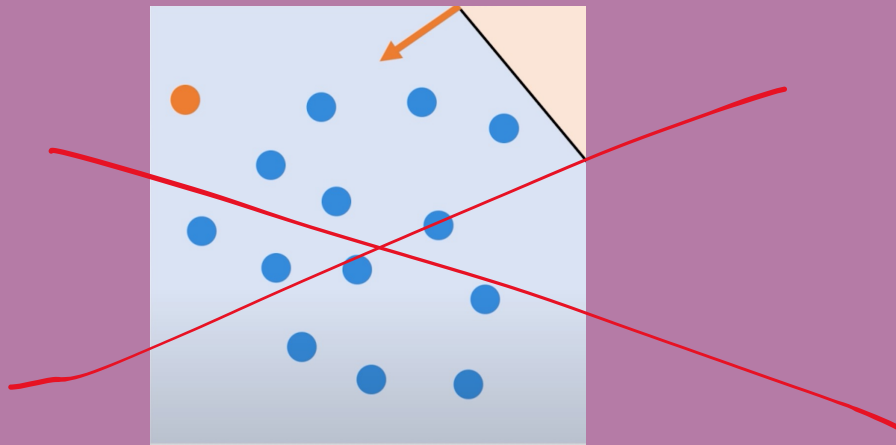
Class imbalance can affect medical diagnosis of diseases, even though the accuracy can be very very high.

Idea: Cancer is rare, but we might not want training dataset to be reflective of real-world distribution.

GBM is one of the most deadly brain tumor, which occurs to 3.19 per 100,000 people.



Learning in
dataset with
balanced classes



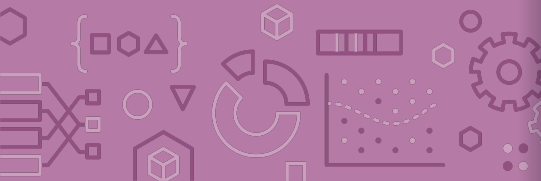
Sampling Bias

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.



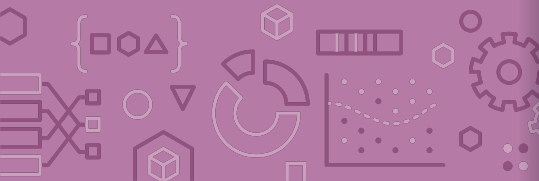
Measurement Bias

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

Crime Rate \rightarrow Arrest Rate

Intelligence \rightarrow 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.)

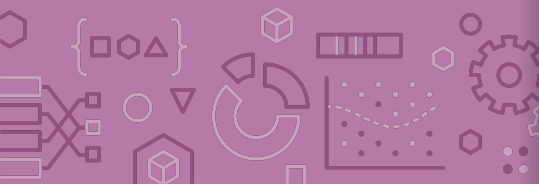
Examples:

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.

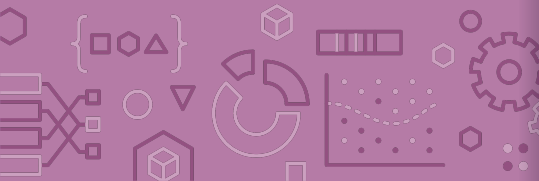
Women are more likely to be misdiagnosed (or not diagnosed) for conditions where self-reported pain is a symptom. In this case aspect of our data “diagnosed with X” is a biased proxy for “has condition X”.

The feature we measure is a poor representation of the quality of interest (e.g., SAT score doesn’t actually measure intelligence)

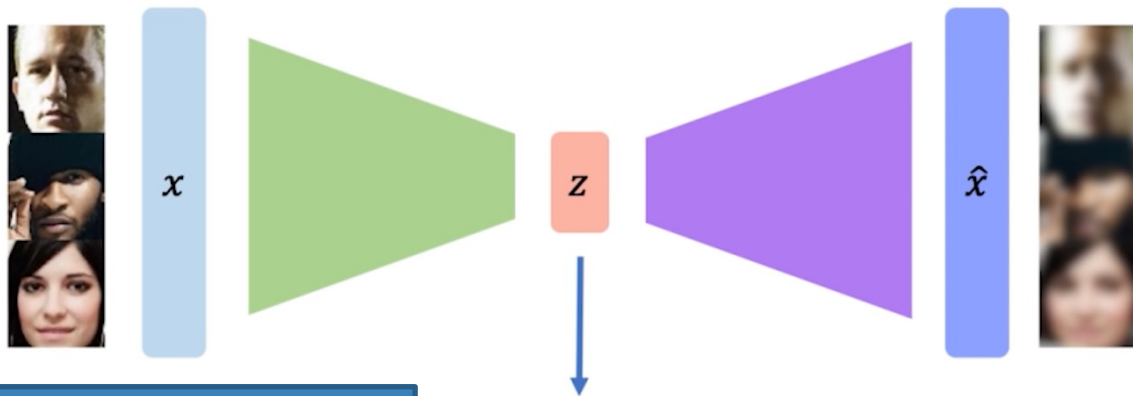


Mitigating bias

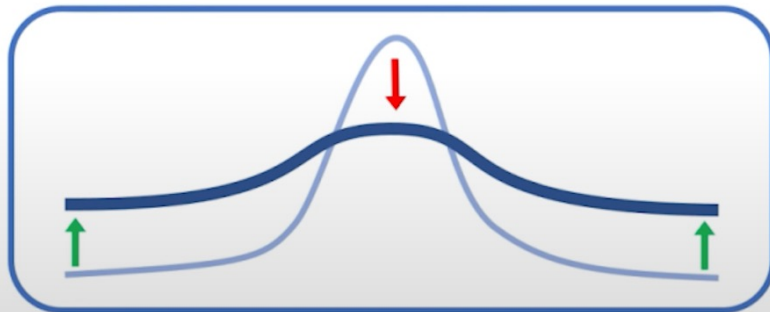
- Remove problematic signals from datasets, add signals for desired features and re-weight signals
- Constantly improve our bias evaluation metrics
- Research methods to detect and mitigate biases during training



Learning latent distributions from generative models



Use generative models to learn latent distributions from datasets and readjust the sampling distributions



Latent distributions can be used to create more fair distributions in the datasets

Fairness in ML

Fairness

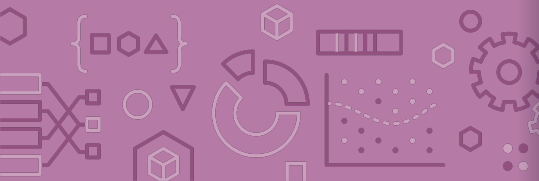
What does it mean for a model to be fair or unfair? Can we come up with a numeric way of measuring fairness?

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.



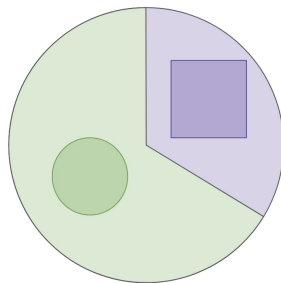
Example: College Admissions

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 = SAT Score

A variable indicating which group X belongs in

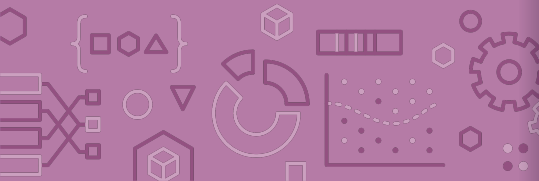
Example: $A = \blacksquare$ or $A = \bigcirc$

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



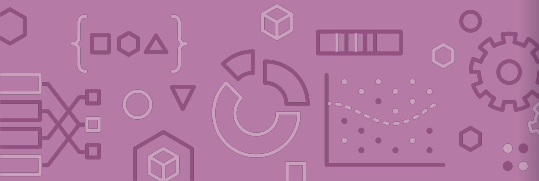
Fairness

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

$$C_{TP} = \#TP, \quad C_{FP} = \#FP, \quad C_{TN} = \#TN, \quad C_{FN} = \#FN$$

$$N = C_{TP} + C_{FP} + C_{TN} + C_{FN}$$

$$N_P = C_{TP} + C_{FN}, \quad N_N = C_{FP} + C_{TN}$$

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”

$$\Pr(\hat{Y} = + | A = \blacksquare) = \Pr(\hat{Y} = + | A = \bigcirc)$$

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

$$\Pr(\hat{Y} = + | A = \blacksquare, Y = +) = \Pr(\hat{Y} = + | A = \bigcirc, Y = +)$$

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

$$\Pr(\hat{Y} = - | A = \blacksquare, Y = -) = \Pr(\hat{Y} = - | A = \bigcirc, Y = -)$$

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?

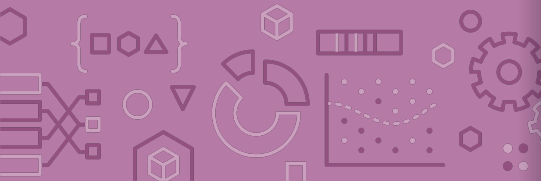
We can't tell you! Each definition makes its own statement on what fairness means. Choosing a fairness measure is an explicit statement of what values we hold when thinking about fairness.

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

On Wednesday 5/27, we'll have a special lecture by my TA **Sahil Verma, a PhD researcher in explainability and fairness in AI**. Sahil will talk about the real world example of bias issues that were uncovered by different regulatory/advisory/journalistic bodies and some examples of how the bias has been effectively tackled.



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

Bias

How to mitigate biases

Definitions of fairness

