Responsible Machine Learning

Lecture 3: Algorithmic Fairness Basics

CS 4973-05

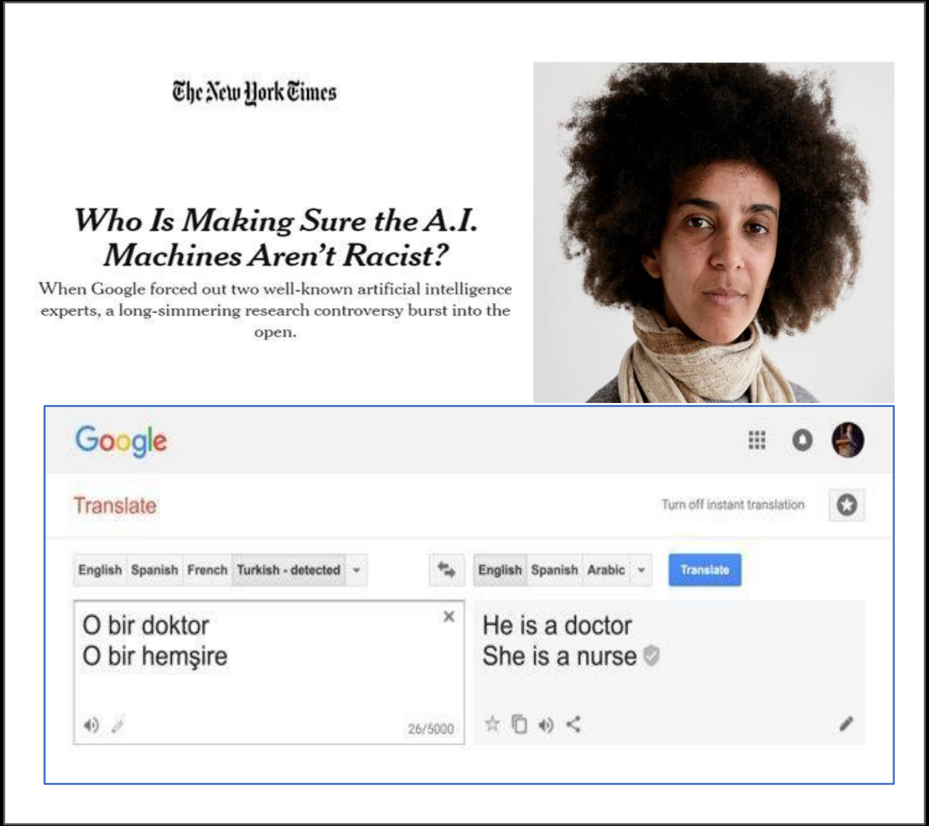
Fall 2023

Instructor: Avijit Ghosh
ghosh.a@northeastern.edu
Northeastern University, Boston, MA

Today's lecture credits: Anupam Datta+John Mitchell, Stanford CS 329T, Spring 2022 and Surya Datta|Slideshare

Fairness

Fairness Matters



- Facial Recognition
- Natural Language Processing
- Online Advertising
- Application for Credit
- College Admissions
- Judicial decisions

What is “Fair”?

What is Fair

Moral principle
Treat similar people similarly

If Avijit is similar to Jeffrey on *relevant input criteria*, then pred(Avijit) should be similar to pred(Jeffrey)

Legal requirement
Illegal to discriminate on the basis of protected characteristics

Cannot favor credit card applicants on the basis of race, gender, ….

Example Qualitative Definitions of Fairness

- **Procedural Fairness / Disparate Treatment (not very strict)**

  Models should not use protected class information as part of a decision-making process, or use other features as proxies to learn and use class membership

- **Equality of Opportunity (a little stricter)**

  Models should not give unfair (dis)advantages to one protected class over another

- **Minimized Inequality of Outcome / Demographic Parity (pretty strict)**

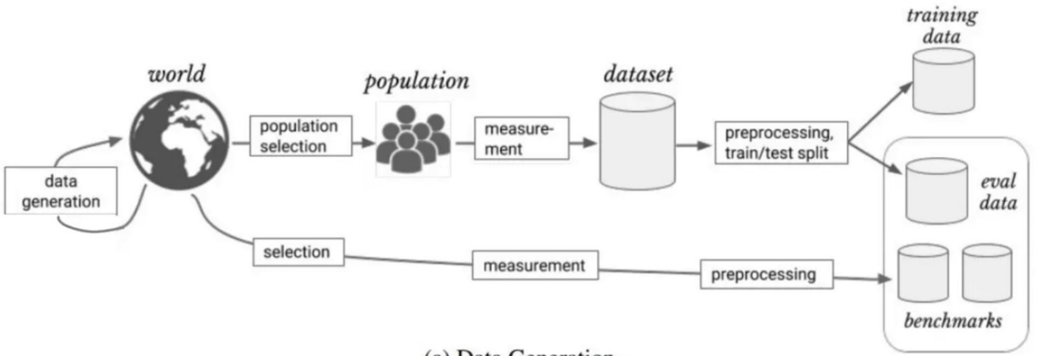
  Subject to achieving the goal the model was designed for, models should allocate resources/opportunities in a way that is as close to the demographic breakdown of the subject population across protected classes as possible

Why is Fairness a Complicated Topic?

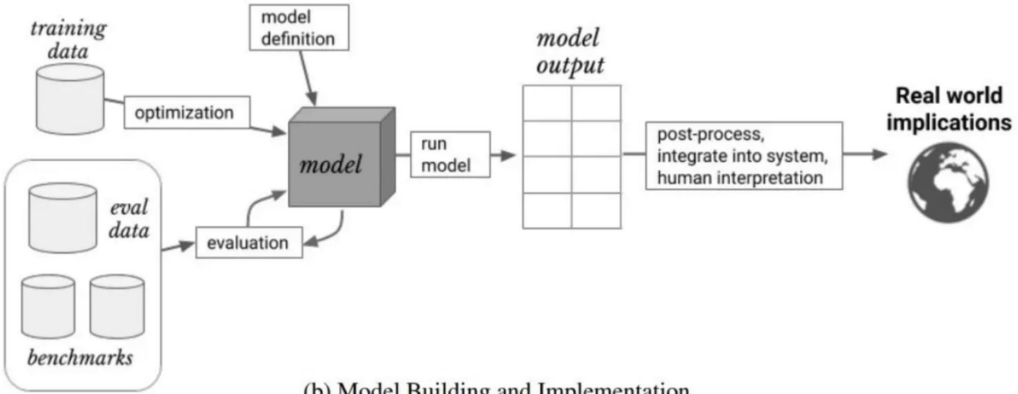
- Precise statements of compelling metrics may be mutually inconsistent
- There may be correlations between relevant and protected characteristics
- Bias in data => bias in training => bias in model

What types of ML Bias Exist?

ML Pipeline



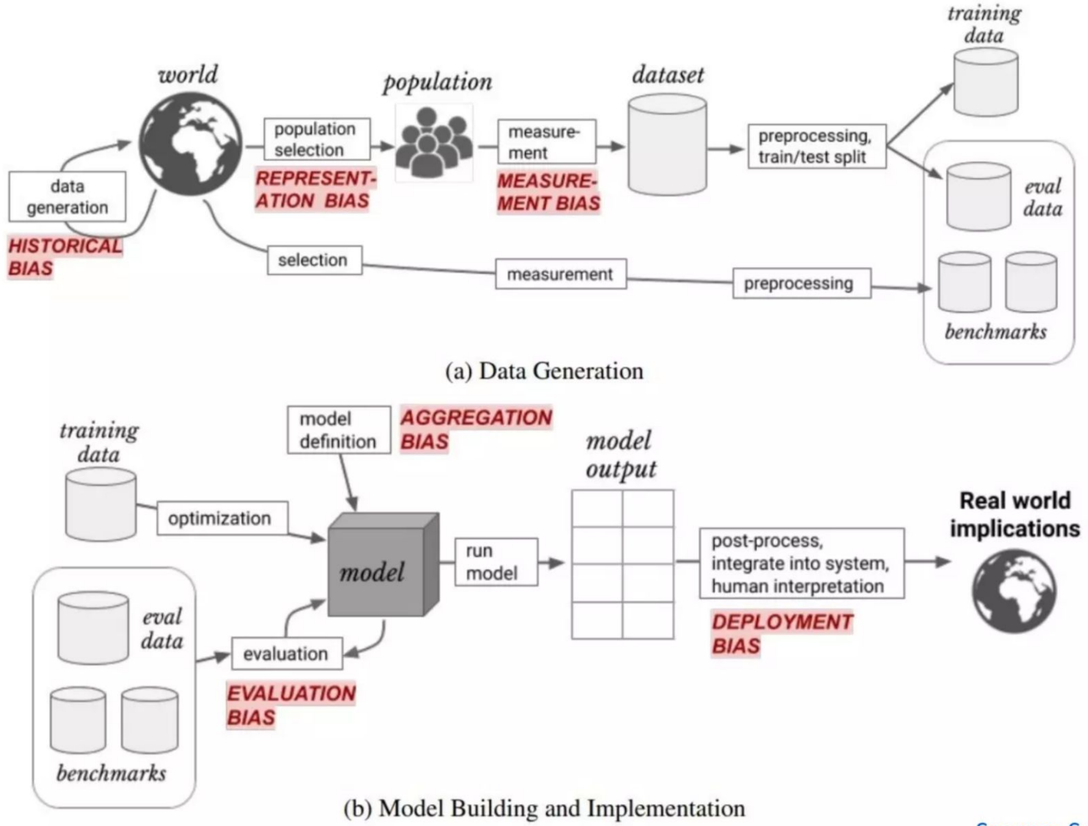
(a) Data Generation



(b) Model Building and Implementation

[Source: Suresh, Guttag 2020]

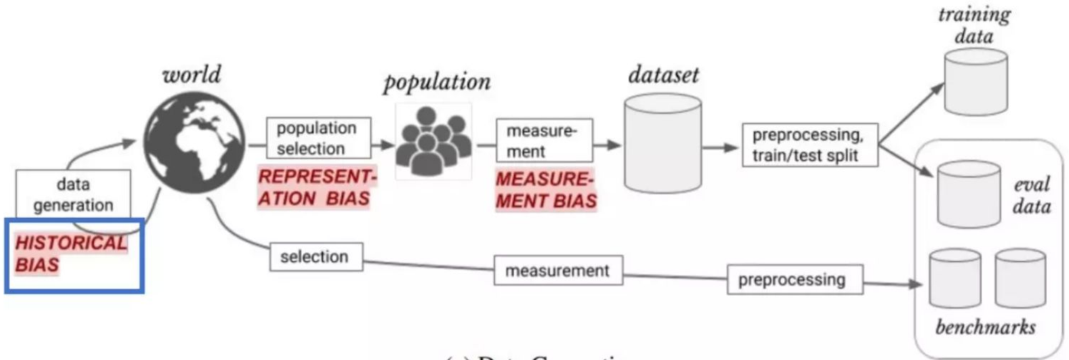
ML Pipeline



Source: Suresh, Guttag 2020

Northeastern University

Historical Bias



(a) Data Generation

LAPD ditches predictive policing program accused of racial bias

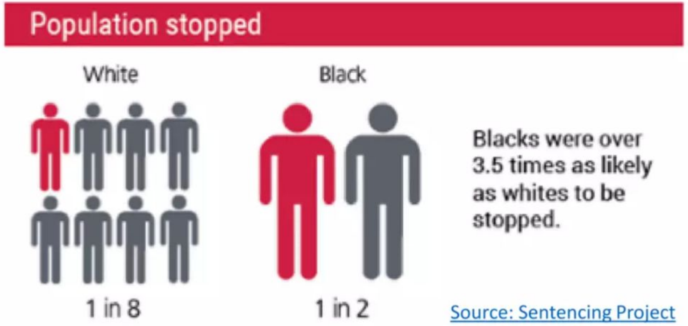
Source: The Next Web

Chicago's predictive policing tool just failed a major test

A RAND report shows that the 'Strategic Subject List' doesn't reduce homicides

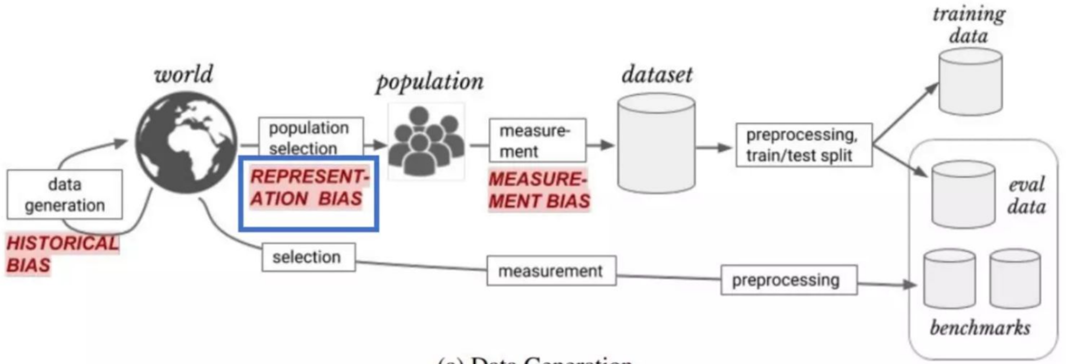
Source: The Verge

Ferguson, Missouri 2013



Source: Sentencing Project

Representation Bias



(a) Data Generation

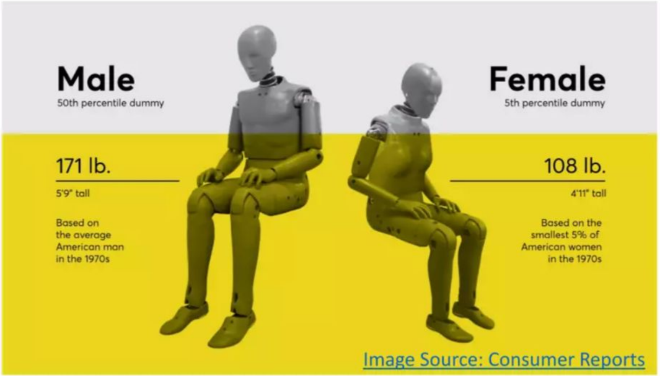
Crash Test Dummies Based on Men Pose Risks for Female Drivers

Source: Invisible Women

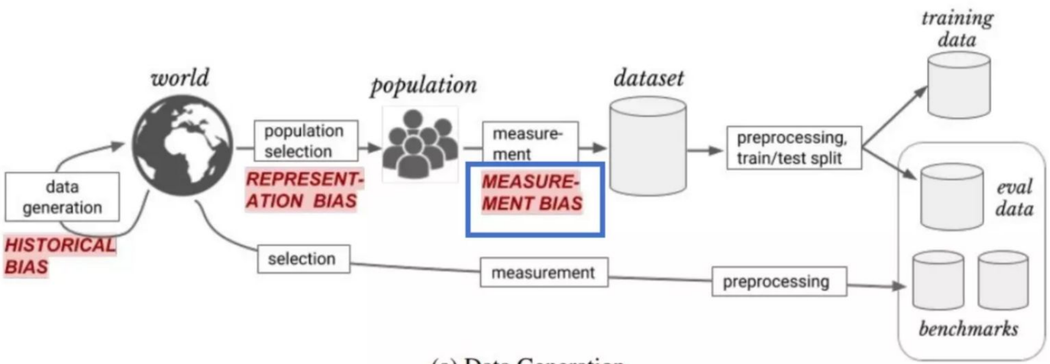
71% more likely to be **moderately injured**

47% more likely to be **seriously injured**

17% more likely to **die**



Measurement Bias



(a) Data Generation

Predicting Recidivism

Source: [“Machine Bias” by ProPublica, 2016]

Prediction Fails Differently for Black Defendants

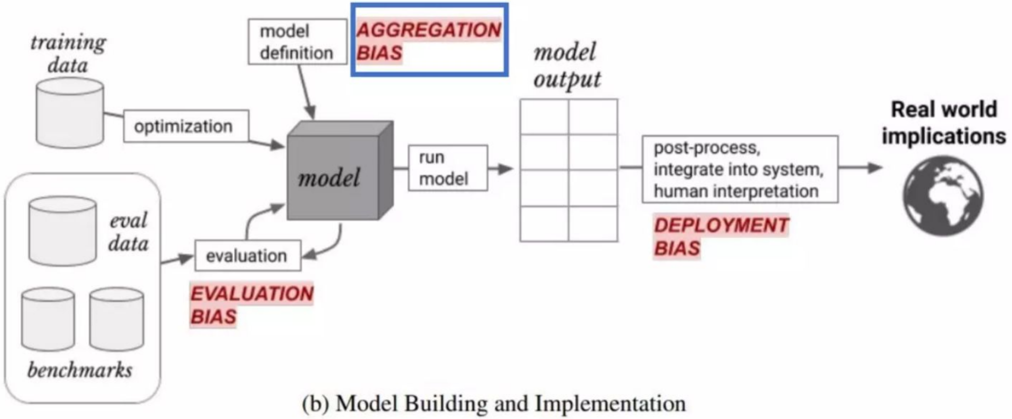
	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

Aggregation Bias

Amazon scraps secret AI recruiting tool that showed bias against women

[Source: Reuters 2018]

"In effect, Amazon's system taught itself that male candidates were preferable. It penalized resumes that included the word "women's," as in "women's chess club captain." And it downgraded graduates of two all-women's colleges, according to people familiar with the matter."



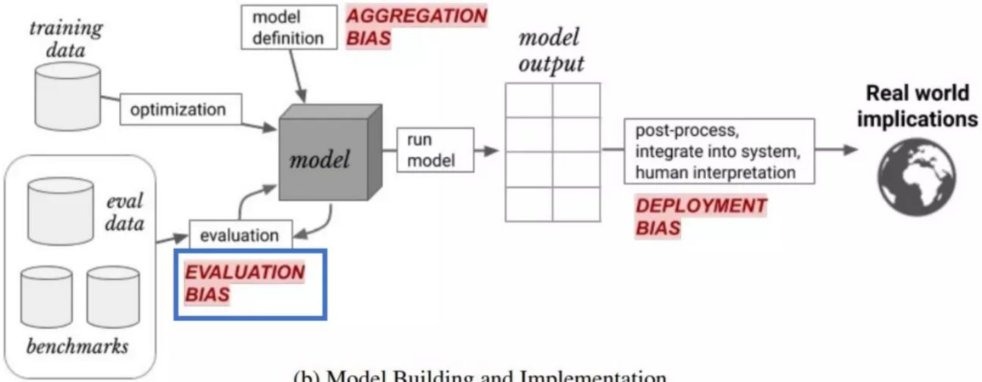
(b) Model Building and Implementation

Evaluation Bias

Gender Classifier	Overall Accuracy on all Subjects in Pilot Parlaiments Benchmark (2017)
Microsoft	93.7%
FACE++	90.0%
IBM	87.9%

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%

[Source: gendershades.org](gendershades.org)



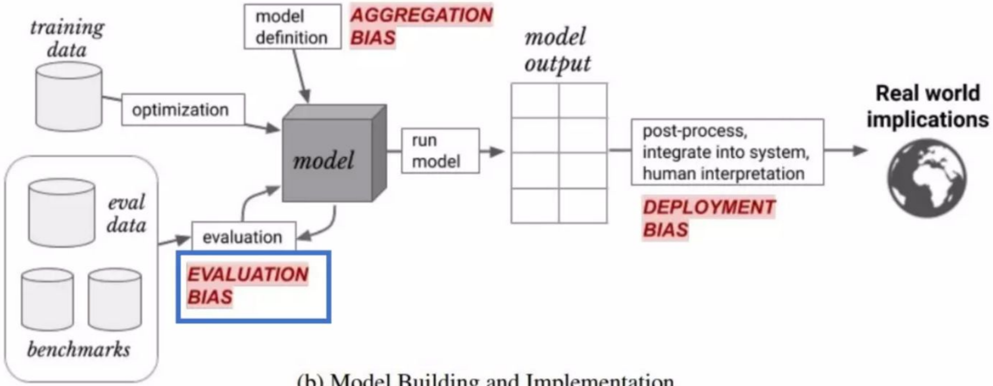
(b) Model Building and Implementation

Northeastern University

Evaluation Bias

Amazon's Face Recognition Falsely Matched 28 Members of Congress With Mugshots

Source: ACLU



(b) Model Building and Implementation

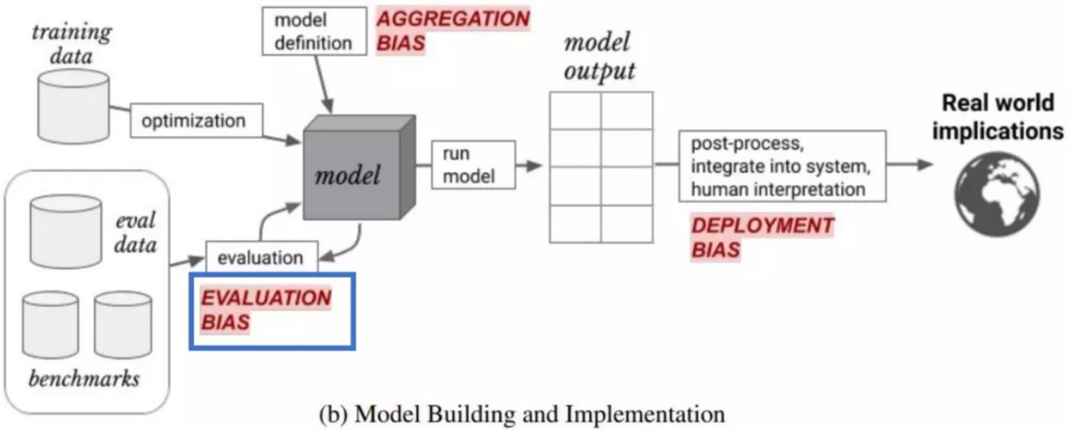
Northeastern University

Evaluation Bias

A black man was wrongfully arrested because of facial recognition

'The computer must have gotten it wrong'

[Source: The Verge]



(b) Model Building and Implementation

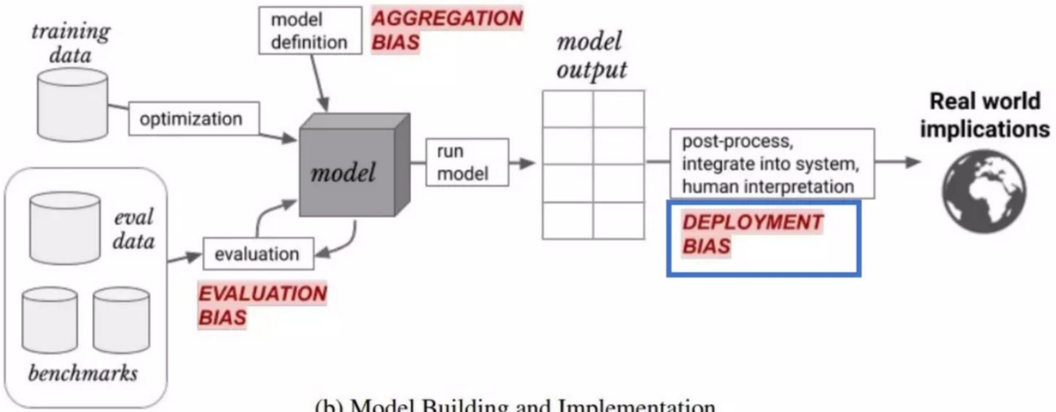
Deployment Bias

A Child Abuse Prediction Model Fails Poor Families

Why Pittsburgh's predictive analytics misdiagnoses child maltreatment and prescribes the wrong solutions

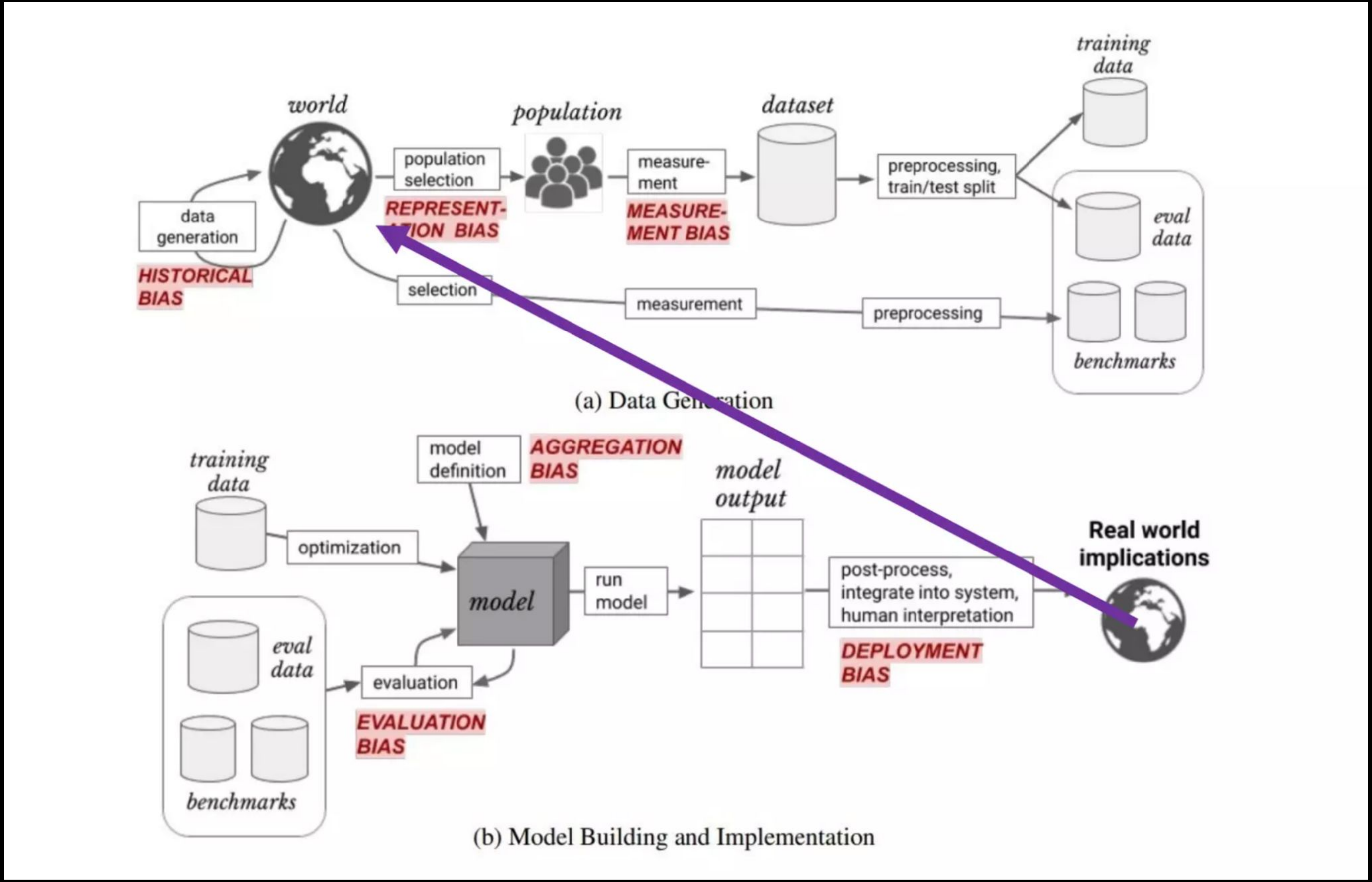
The screen that displays the AFST risk score states clearly that the system **"is not intended to make investigative or other child welfare decisions."**

[Source: Automating Inequality]



(b) Model Building and Implementation

Everything affects everything else



Let's Talk about Fairness Metrics

Defining Fairness

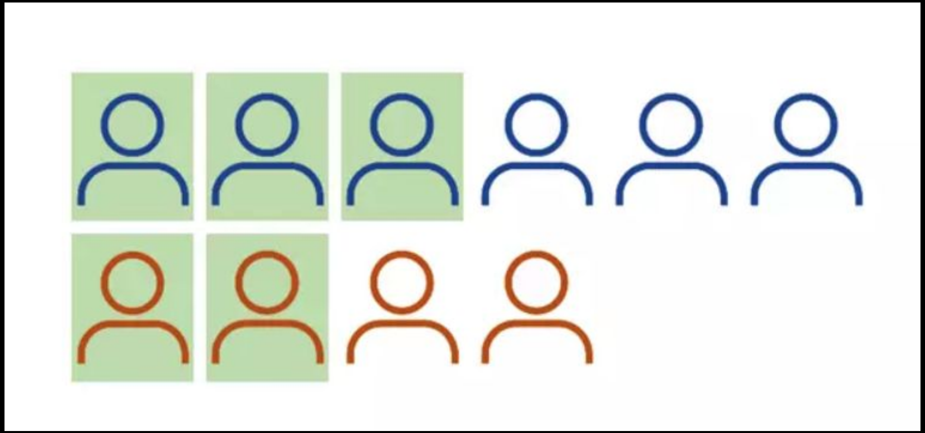
Goal: Create a metric that machine learning algorithm can use to generate fair outcomes

Definitions:

- Y is the true value (0 or 1 for binary classification)
- C is the algorithm's predicted value
- A is the protected attribute (gender, race, etc.)
  - A=1 refers to the unprivileged group, A=0 refers to privileged

Demographic Parity

“A predictor satisfies demographic parity if the likelihood of a positive outcome is the same, regardless of whether the person is in the protected group or not"



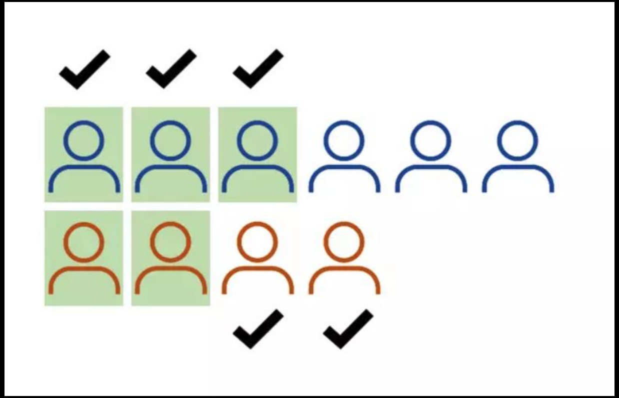
Demographic Parity

"A predictor satisfies demographic parity if the likelihood of a positive outcome is the same, regardless of whether the person is in the protected group or not"

Pros: Proportional representation of groups

Cons: Accuracy may be less in disadvantaged group

Greatly reduces effectiveness of predictor if true labels have any correlation with protected attribute



Equal Odds

"A predictor *C* satisfies equalized odds with respect to a protected attribute *A* and the true outcome *Y* if C and A are independent conditional on Y"

In a binary classification:

- C has equal true positive rates if Y=1 for both A=0 and A=1

		Y=1	Y=0
A=0	C=1	TP	FP
	C=0	FN	TN

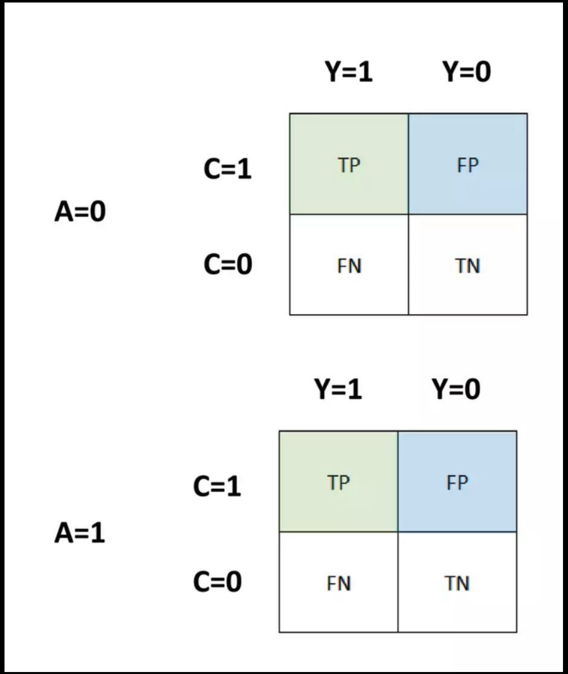
		Y=1	Y=0
A=1	C=1	TP	FP
	C=0	FN	TN

Equal Odds

"A predictor *C* satisfies equalized odds with respect to a protected attribute *A* and the true outcome *Y* if C and A are independent conditional on Y"

In a binary classification:

- C has **equal true positive rates** if Y=1 for both A=0 and A=1
- C has **equal false positive rates** if Y=0 for both A=0 and A=1



Equal Odds

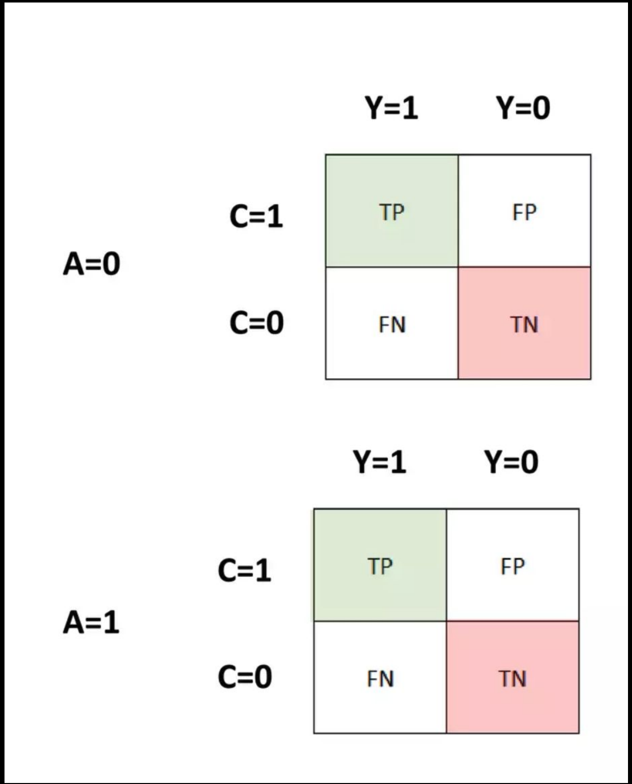
#	Qualified?	Hired?	Classification
2	Yes	Yes	True Positive
3	Yes	No	False Negative
4	No	Yes	False Positive
5	No	No	True Negative
1	Yes	Yes	True Positive
1	Yes	No	False Negative
2	No	Yes	False Positive
3	No	No	True Negative

Equal Odds

	#	Qualified?	Hired?	Classification	In-Group Rate
	2	Yes	Yes	True Positive	**2/14**
	3	Yes	No	False Negative	3/14
	4	No	Yes	False Positive	**4/14**
	5	No	No	True Negative	5/14
	1	Yes	Yes	True Positive	**1/7**
	1	Yes	No	False Negative	1/7
	2	No	Yes	False Positive	**2/7**
	3	No	No	True Negative	3/7

Equal Odds

Why don't we measure just accuracy? (TP+TN)

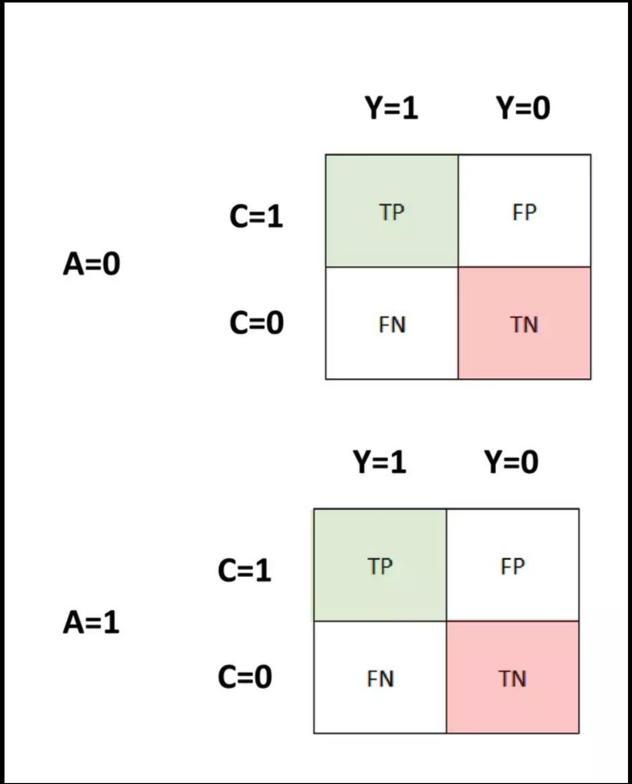


Equal Odds

Why don’t we measure just accuracy? (TP+TN)

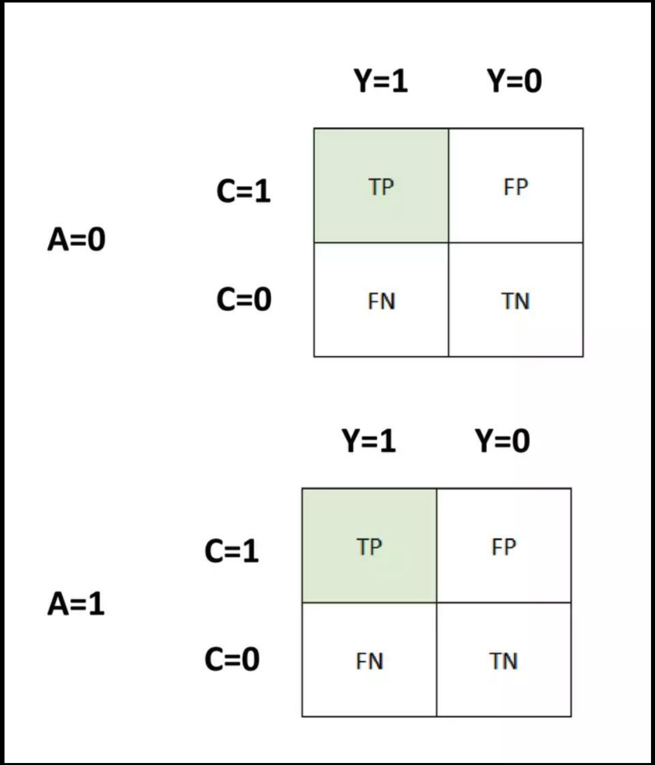
Weakness: We can "trade" the false positive rate of one group for the false negative rate for another group

Ex. Hiring from two groups. We can achieve accuracy parity by exchanging qualified applicants from privileged group for unqualified applicants from unprivileged group



Equal Opportunity

- Relaxed version of Equal Odds
- Equal true positive rates for Y=1 for both A=0 and A=1
- Useful when only care about positive outcome



Northeastern University

Other Metrics

CLASSIFICATION	Disparate impact ratio
	Statistical parity difference
	True/false positive/negative rates
	Treatment equality difference
	Equality of opportunity ratio/difference
	Conditional acceptance/rejection difference
	Predictive parity ratio/difference
REGRESSION	L1 error difference
	Mean score difference
	L2 error difference

Which Metric to use When?

These three notions of fairness are *incompatible*

- **Independence:** Predictions should be independent of membership in a protected class

$E_a[\,Y = 1\,]$

Demographic parity measures

- **Separation:** Predictions should be independent of membership in a protected class, given the true outcome (performance is the same across classes)

$E_a[\,C=1|Y=0\,]$

False Positive Rate

- **Sufficiency:** True outcomes should be independent of membership in a protected class, given the predictions (no extra information encoded in the protected class)

$E_a[C=1|Y=1]$

Predictive Parity

Fun video to watch: [21 Fairness Definitions and Their Politics]

College admissions

- **Procedural Fairness / Disparate Treatment:** "The model isn't given access to gender, so it is procedurally fair and does not treat women differently."
- **Equality of Opportunity:** "Let's equalize false negative rate so that the chance of an qualified man getting rejected is the same as the chance of a qualified woman getting rejected. If there's a correlation between gender and qualification, that's okay, so long as it's through a relevant feature such as extracurricular activity."
- **Demographic Parity:** "Gender and college qualification are completely uncorrelated and we want a class that reflects the population prevalence of men and women, so we should make sure that men and women are accepted at equal rates."

Facial recognition, Gender misclass of dark-skinned women

- **Procedural Fairness / Disparate Treatment:** "The model is just given access to a sequence of pixels, so it contains no explicit encoding of race."
- **Equality of Opportunity:** "Let's equalize false negative and/or positive rates so that the chance of someone getting misclassified does not depend on their skin color."
- **Demographic Parity:** "We want to make sure that the probability that someone is classified as male/female does not depend on their skin color."

How do we pick a fairness metric?

What kind of impact does the action the model informs have on an individual?
When the model assigns a label of 1 to someone, what kind of impact does the subsequent action have?

Are fair and accurate labels available?
Do we have access to ground-truth labels for each person which reflect the outcome that ideally should have been assigned, and is this reflective of the underlying population?

	Beneficial *Individual qualifies for a loan*	**Mixed** *Individual is selected for an experimental medical treatment*	**Harmful** *Individual is chosen for a search by police*
Yes, for all instances *Example: Prediction of credit card fraud over a time period*	**FNR (or TPR)** Ensure the proportion of people unfairly missing out on a benefit is balanced	**Treatment Equality** Ensure the ratio of false positives to false negatives is balanced	**FPR (or TNR)** Ensure the proportion of people unfairly being harmed is balanced
Only for instances with a label of 1 *Example: Qualifying for a loan*	**Predictive Parity** Ensure the number of people undeservedly helped (or harmed) as a fraction of the number of people intervened upon is balanced		
No *Example: Decision to admit a student to a school*	**Demographic Parity (Disparate Impact, Statistical Parity)** Without fair labels, we want to ensure that outcomes are equal across protected classes		

Fairness vs. Accuracy Tradeoffs

What kind of impact does the action the model informs have on an individual?
When the model assigns a label of 1 to someone, what kind of impact does the subsequent action have?

Are fair and accurate labels available? Do we have access to ground-truth labels for each person which reflect the outcome that ideally should have been assigned, and is this reflective of the underlying population?		**Beneficial** *Individual qualifies for a loan*	**Mixed** *Individual is selected for an experimental medical treatment*	**Harmful** *Individual is chosen for a search by police*
	Yes, for all instances *Prediction of credit card fraud over a time period*	**Accuracy and Fairness are aligned** Any strategy taken to improve model accuracy should also improve fairness, since the closer the model is to the training data, the more fair we are		
	Only for instances with a label of 1 *Qualifying for a loan*	**Accuracy and Fairness are partially aligned** While both accuracy and fairness benefit from correctly giving labels of 1 to people, closing the gap in precision may necessitate an accuracy drop		
	No *Decision to admit a student to a school*	**"Accuracy" and Fairness are not aligned** Satisfying demographic parity may ostensibly lower accuracy, this accuracy is measured with respect to labels that are not fair/accurate. A model that better captures the actual prevalence rate may do better in practice		

Final Thoughts

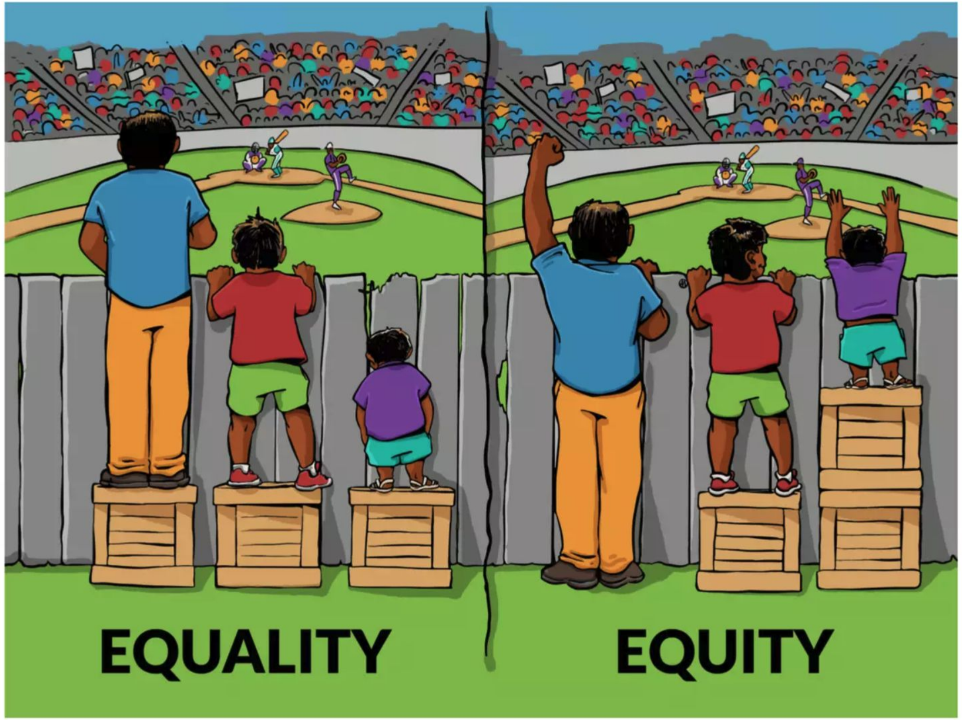


Image Source: Interaction Institute for Social Change

Thank You!

Readings for Next Class:

- Machine Bias - Propublica
- Image Cropping on Twitter: Fairness Metrics, their Limitations, and the Importance of Representation, Design, and Agency