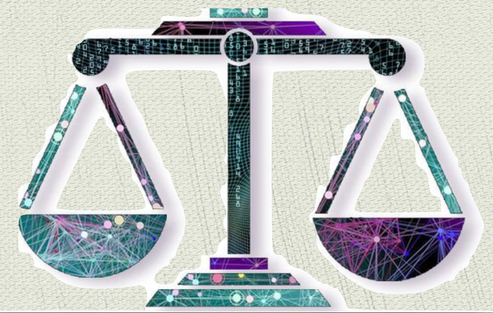
GEORGETOWN UNIVERSITY
MASSIVE DATA INSTITUTE
McCourt School of Public Policy

When is AI biased? Measuring fairness in machine learning models

Agoritsa Polyzou

Day 2



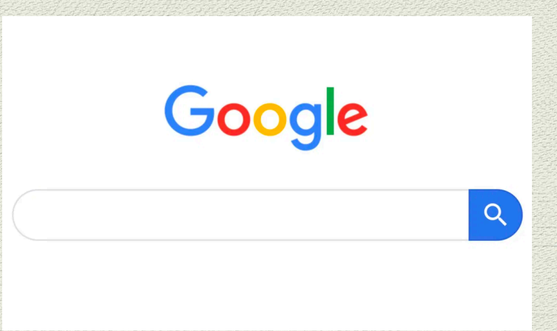
Recap!

Machine learning is …

About algorithms that give computers the tools and technology we can utilize to analyze massive amounts of data, find patterns, make predictions, and answer questions.

- Computers are good at storing, organizing, and processing **data**.
- Data: numbers, words, images, clicks, sounds, etc…

Applications



Search engines
(Understand text)

Transportation
(Self-driving cars)

Social Media
(Sentiment Analysis)

Criminal justice
(Recidivism risk)

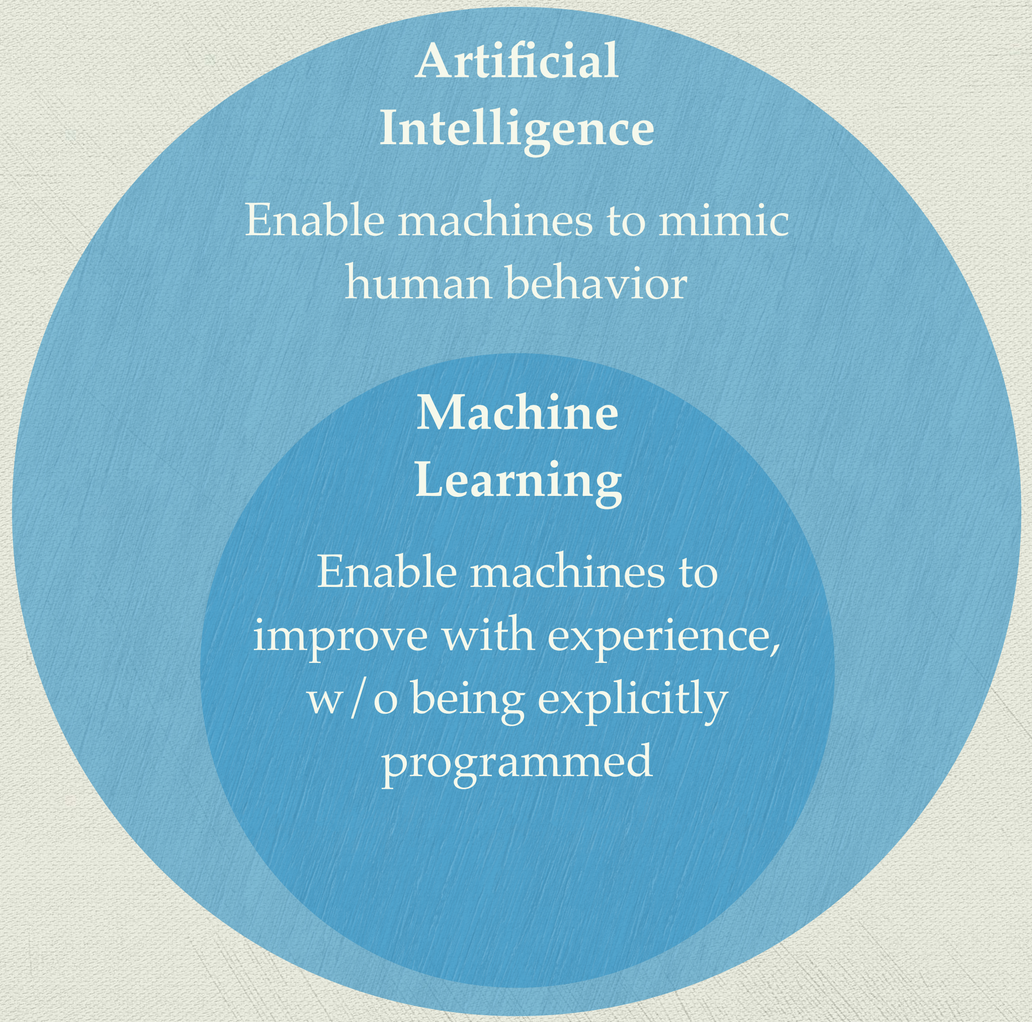
Finance
(Fraud Detection)

Government
(Handwriting recognition)

Healthcare
(Predict diagnosis)

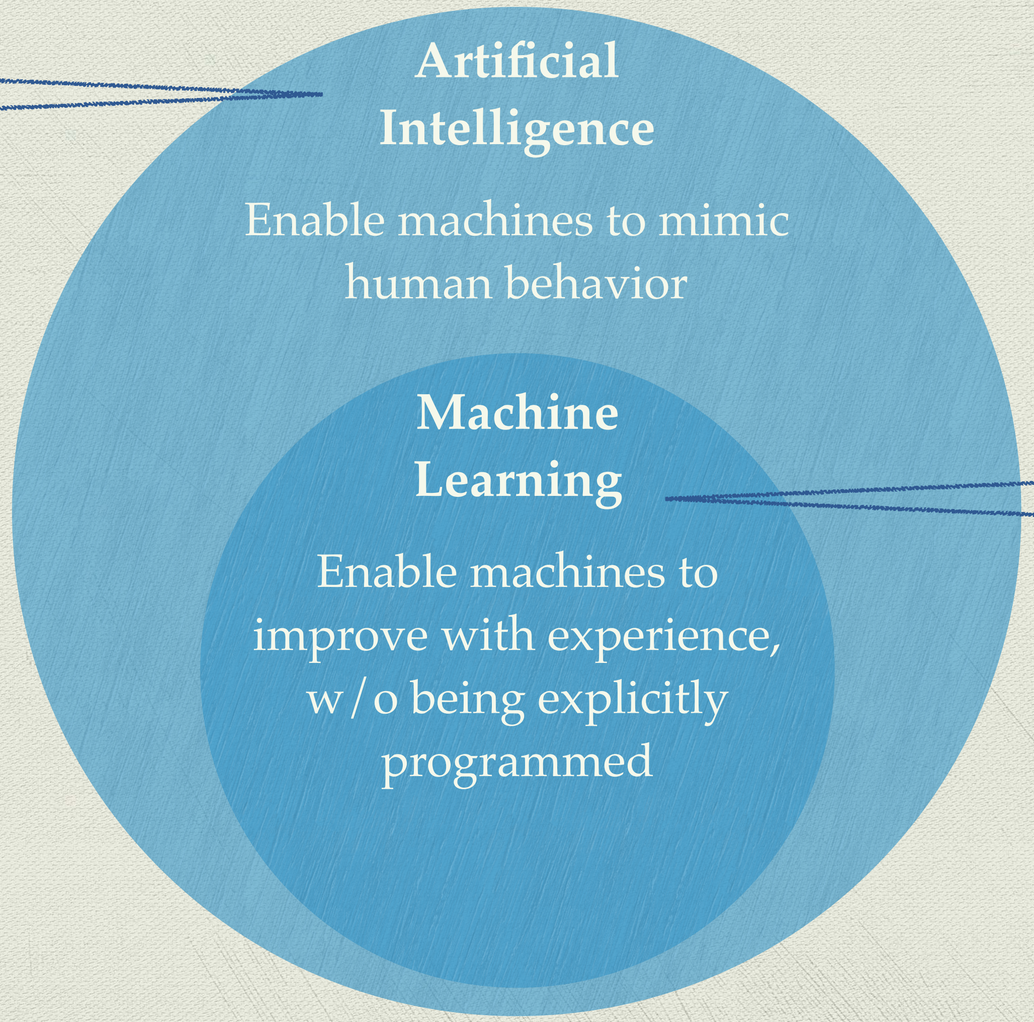
Shopping
(Product recom.)

AI and ML


Artificial Intelligence
Enable machines to mimic human behavior
Machine Learning
Enable machines to improve with experience, w/o being explicitly programmed

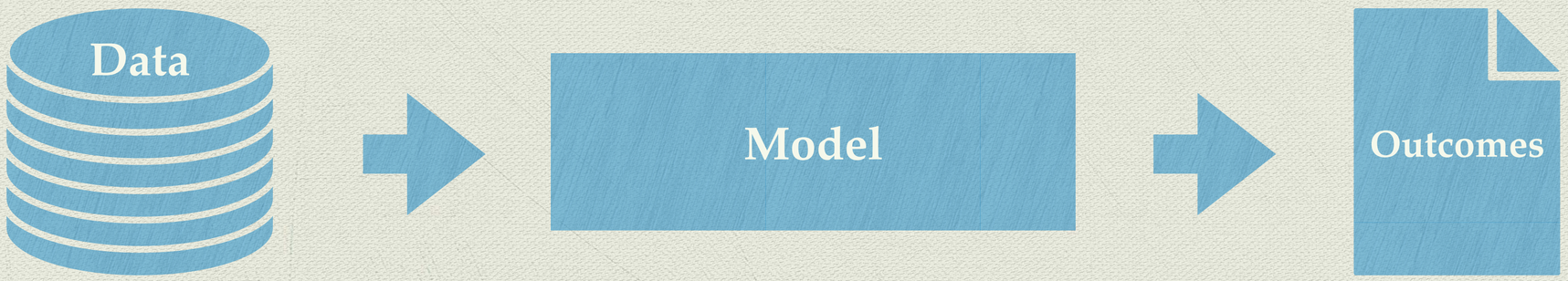
AI and ML

learning from rules or logical induction

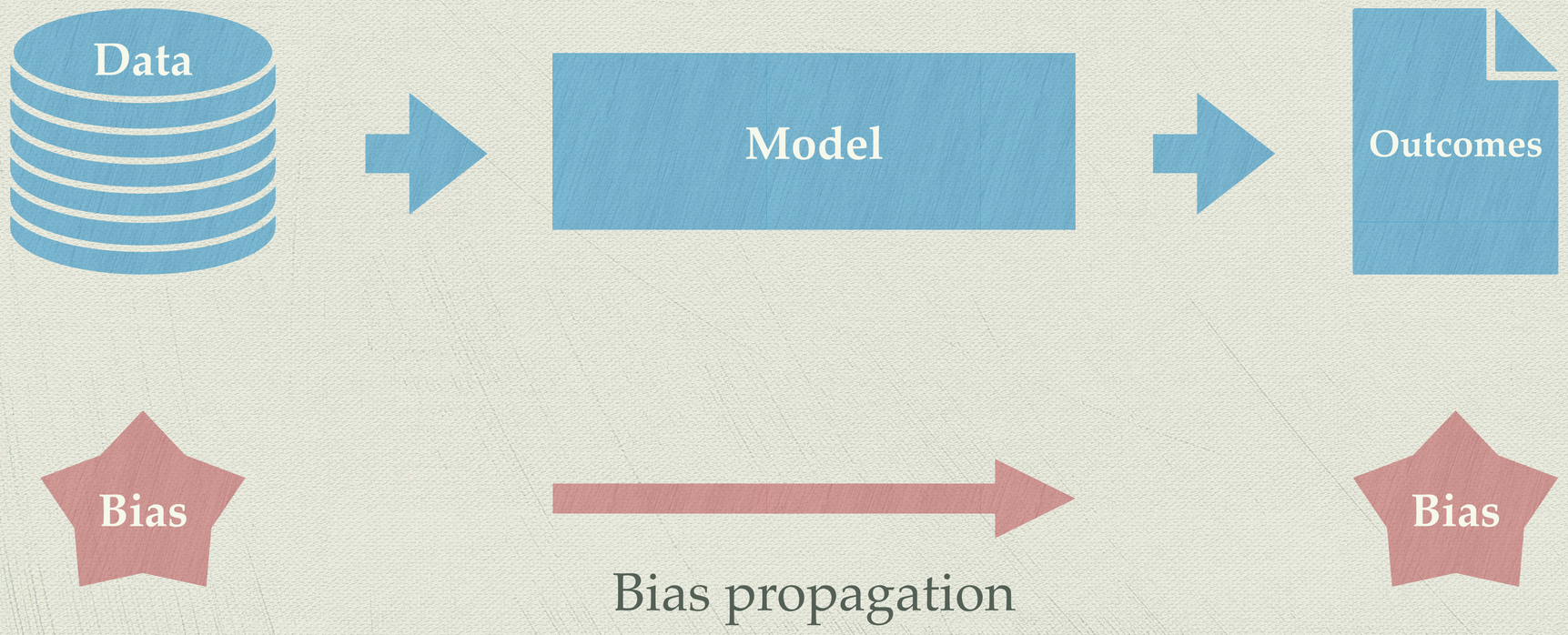


just give them the data and let them learn

Biased Data ⇒ Biased ML models



Biased Data ⇒ Biased ML models



It is not enough to know that bias exists..

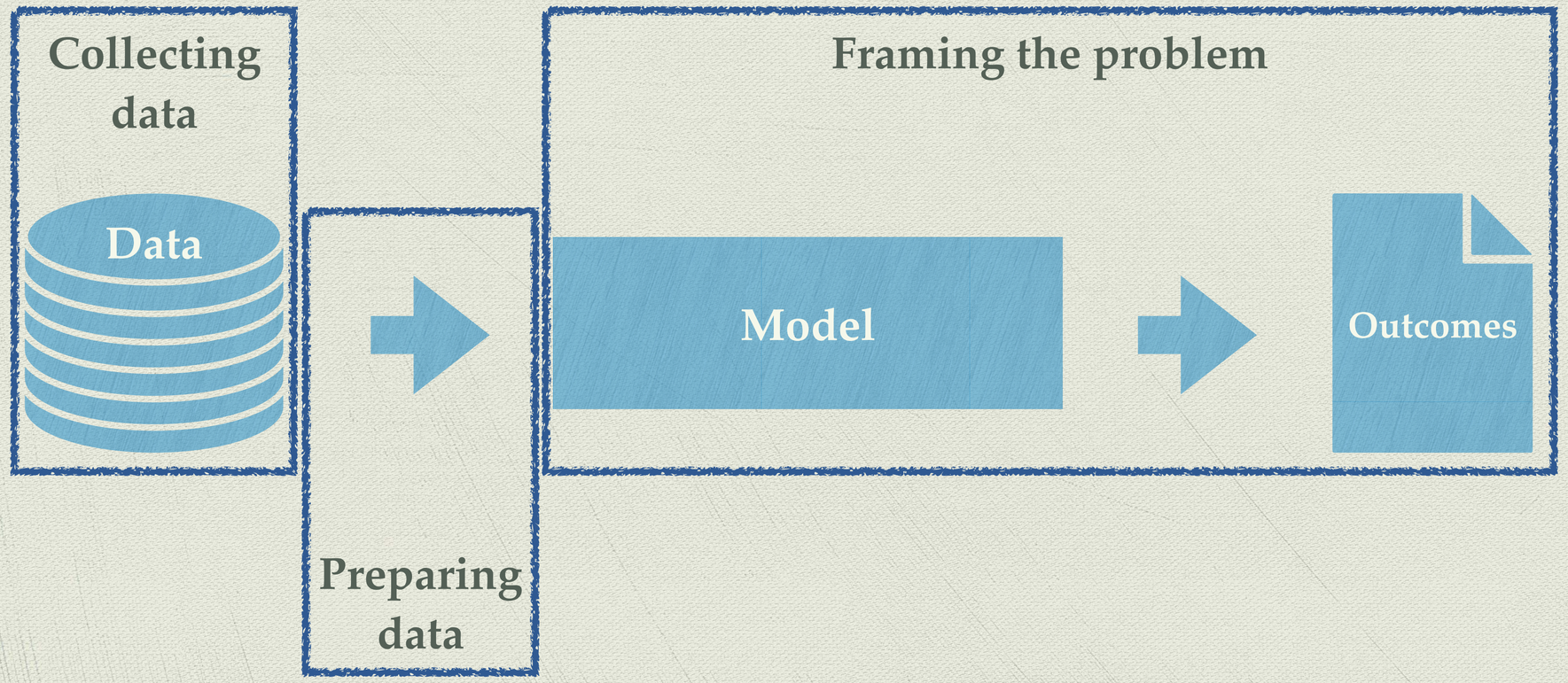
Day 1

We need to understand the mechanics of how it arises in the first place,

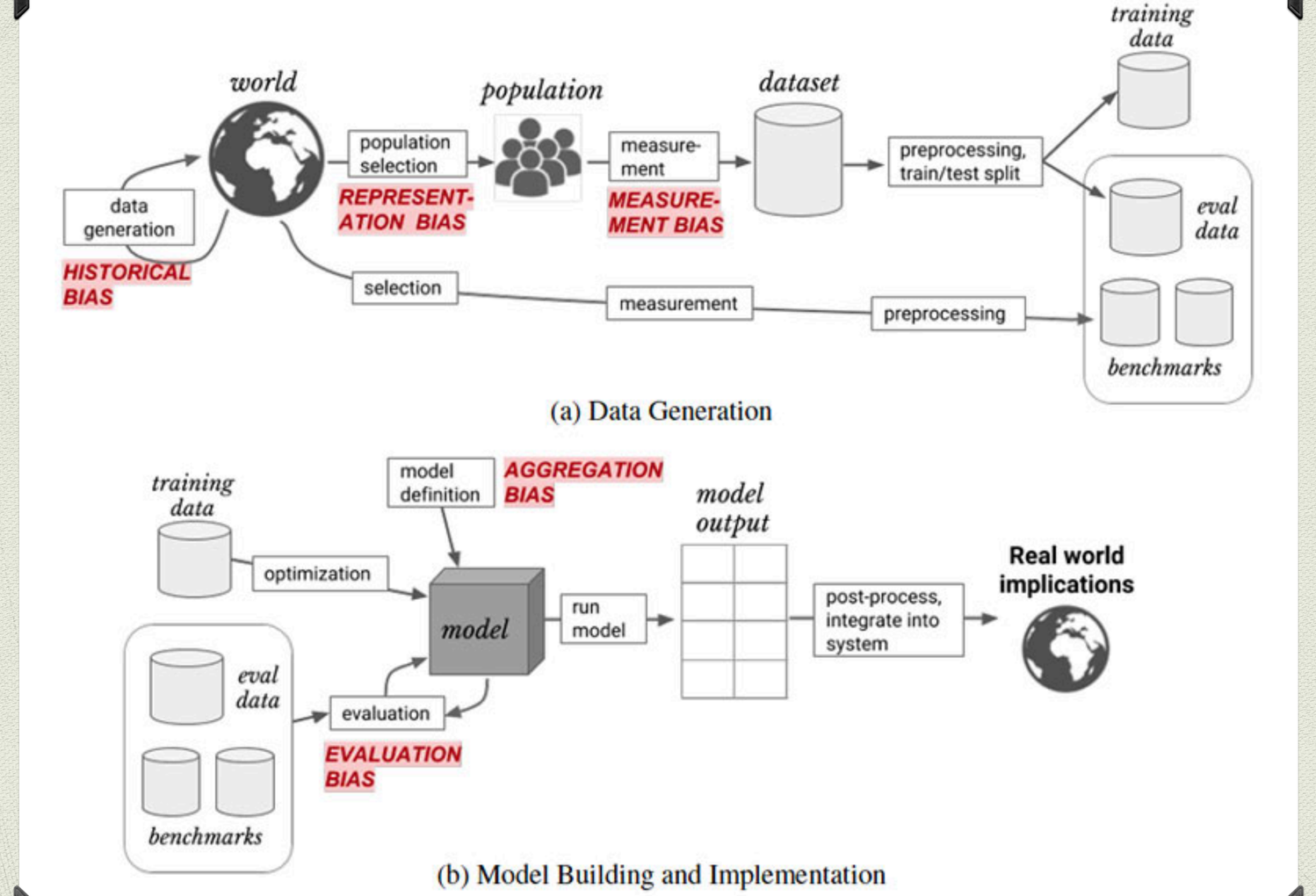
we need to be able to fix it.

Day 2

How is bias introduced in ML?



The different forms of bias



A Framework for Understanding Unintended Consequences of Machine Learning, Harini Suresh, John V. Guttag, 2020.

Next Step:

Define & measure fairness

Problem set up (1)

Let's assume that we have a classification problem.

- $x \in \mathbb{R}^p$: visible attributes of an individual, $x = (x_p, x_u)$
- a_0, a_1 : two possible outcomes to predict
- $d : \mathbb{R}^p \to \{0,1\}$: Model/decision algorithm
- $d(x_i) = k$: predict a_k for an individual

Goal: fairly predict 1 or 0 (a_0 or a_1)

Problem set up (2)

- $y \in \{0,1\}$: target (ground truth) prediction for an individual
- X, Y : random variables that take values $X = x, Y = y$ for an individual drawn randomly from the population of interest
- X_p, X_u : protected & unprotected components of X
- $s(x) \in [0,1]$: prediction score, instead of a_0, a_1
- t : threshold to go from scores to decisions, if $s(x) \geq t, d(x) = 1$

Sensitive/Protected Attributes/Features

The following are some of the attributes/features which could result in bias:

- Race
- Gender
- Color
- Religion
- National origin
- Marital status
- Sexual orientation
- Education background
- Source of income
- Age

Fairness notions

- Anti-classification
- Statistical Parity
  - Demographic parity, Equal opportunity, etc
- Individual Fairness

Anti-classification

We have algorithmic fairness when:

decisions do not consider protected attributes.

$$d(x) = d(x') \quad \text{for all} \quad x, x' \quad \text{such that} \quad x_u = x'_u$$

Anti-classification

We have algorithmic fairness when:

decisions do not consider protected attributes.

$d(x) = d(x')$ for all x, x' such that $x_u = x'_u$

Predictions

Pairs of individuals

Same unprotected attributes

Classification parity

We have algorithmic fairness when:

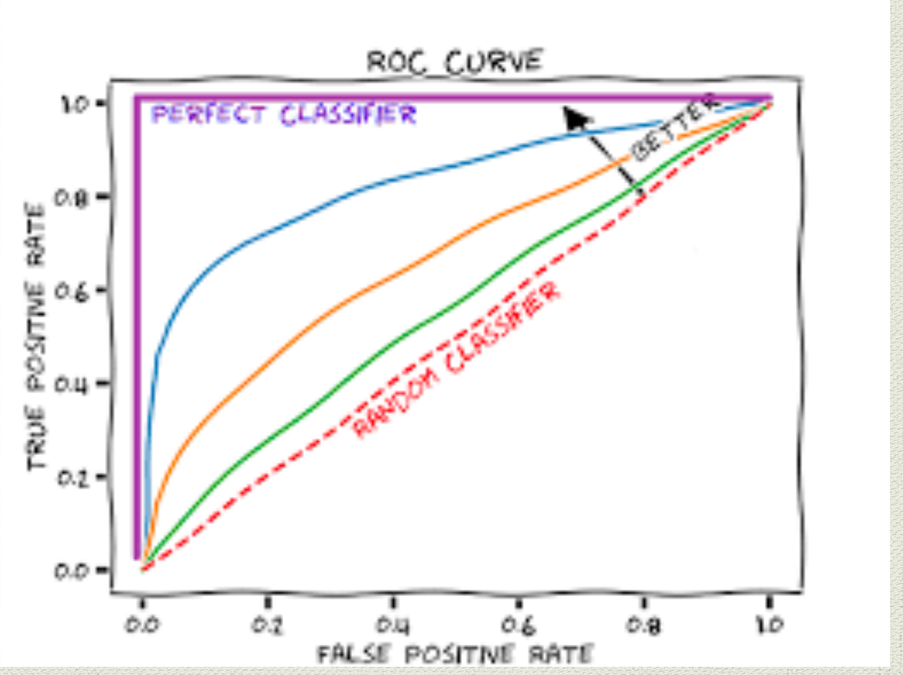
some given measure of classification error is equal across groups defined by the protected attributes.

(Binary) Classification error

Confusion matrix

		Ground truth		
		+	-	
Predicted	+	True positive (TP)	False positive (FP)	Precision = TP / (TP + FP)
	-	False negative (FN)	True negative (TN)	False negative rate FN / (FN + TN)
		Recall = TP / (TP + FN)	False positive rate FP / (FP + TN)	Accuracy = (TP + TN) / (TP + FP + TN + FN)

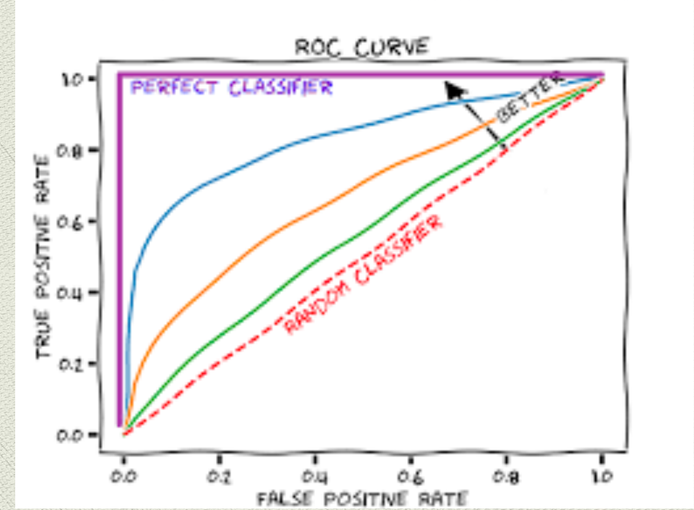
ROC curve



Classification parity - metrics

- false positive rate
- false negative rate
- precision
- recall (*Equality of opportunity*)
- accuracy
- the proportion of decisions that are positive (*demographic parity*)*
- Area under the curve (AUC)

		Ground truth		
		+	-	
Predicted	+	True positive (TP)	False positive (FP)	Precision = TP / (TP + FP)
	-	False negative (FN)	True negative (TN)	False negative rate FN / (FN + TN)
		Recall = TP / (TP + FN)	False positive rate FP / (FP + TN)	Accuracy = (TP + TN) / (TP + FP + TN + FN)



* strictly, not a measure of error, but captures an important model characteristic; derived from the confusion matrix.

Demographic parity

We have algorithmic fairness when:

positive decisions are equally likely to occur across groups defined by the protected attributes.

$$Pr(d(X) = 1 \mid X_p) \;=\; Pr(d(X) = 1)$$

Demographic parity

We have algorithmic fairness when:

positive decisions are equally likely to occur across groups defined by the protected attributes.

$$Pr(d(X) = 1 \mid X_p) = Pr(d(X) = 1)$$

Positive decisions, given the protected attributes

Positive decisions

Similar to anti-classification

Parity in false positive rates

We have algorithmic fairness when:

false positive decisions are equally likely to occur across groups.

$$Pr(d(X) = 1 \mid Y = 0,\ X_p) \ = \ Pr(d(X) = 1 \mid Y = 0)$$

Parity in false positive rates

We have algorithmic fairness when:

false positive decisions are equally likely to occur across groups.

$$\boxed{Pr(d(X) = 1 \mid Y = 0,\ X_p)} = \boxed{Pr(d(X) = 1 \mid Y = 0)}$$

False positive decisions, given the protected attributes

False positive decisions

Equalized odds

		Ground truth		
		+	-	
Predicted	+	True positive (TP)	False positive (FP)	Precision = TP / (TP + FP)
	-	False negative (FN)	True negative (TN)	False negative rate FN / (FN + TN)
		Recall = TP / (TP + FN)	False positive rate FP / (FP + TN)	Accuracy = (TP + TN) / (TP + FP + TN + FN)

We have algorithmic fairness when:

protected and unprotected groups have equal

recall & false positive rate

$$Pr(d(X) = 1 \mid Y = y, X_p) \;=\; Pr(d(X) = 1 \mid Y = y), \text{ for } y = \{0,1\}$$

Equalized odds

		Ground truth		
		+	-	
Predicted	+	True positive (TP)	False positive (FP)	Precision = TP / (TP + FP)
	-	False negative (FN)	True negative (TN)	False negative rate FN / (FN + TN)
		Recall = TP / (TP + FN)	False positive rate FP / (FP + TN)	Accuracy = (TP + TN) / (TP + FP + TN + FN)

We have algorithmic fairness when:

protected and unprotected groups have equal

recall & false positive rate

$$Pr(d(X) = 1 \mid Y = y,\ X_p) = Pr(d(X) = 1 \mid Y = y), \text{ for } y = \{0,1\}$$

True/False positive decisions, given the protected attributes

True/False positive decisions

Which metric to use?

Depends on the problem!

- What is the most important of those quantities?
- Which error will have the most significant disparate impact across protected groups?

Which metric to use?

Which metric to use?

Example: Predict if a person is diabetic ($y = 1$) or not ($y = 0$).

- True positives (TP) and True negatives (TN) are always good.
- False positives (FP) are false alarms.
- **False negatives** (FN) are worse, giving us a wrong sense of 'security'.

Which metric to use?

Example: Predict if a person is diabetic ($y = 1$) or not ($y = 0$).

- True positives (TP) and True negatives (TN) are always good.
- False positives (FP) are false alarms.
- **False negatives** (FN) are worse, giving us a wrong sense of 'security'.

Example: Predict if an email is spam ($y = 1$) or not ($y = 0$).

- True positives (TP) and True negatives (TN) are still good.
- False negatives (FN) result in some spam emails in our mailbox.
- **False positives** (FP) are worse, placing non-spam emails to spam.

Disparate Impact

For a protected attribute $S \in X_p, S = \{0,1\}$, the Disparate Impact of the classifier $d()$ is computed as:

$$DI(d, X, S) = \frac{Pr(d(X) = 1 \mid S = 1)}{Pr(d(X) = 1 \mid S = 0)}$$

The probability to positively predict individuals in group 0/1

For an ideally fair classifier : $DI(d, X, S) = 1$

Classification (i.e., statistical) parity

Demographics of those receiving positive (or negative) classifications are identical to the demographics of the population as a whole.

- Statistical parity speaks to **group** fairness.
- It equalizes outcomes across protected and non-protected groups.

Fairness Gerrymandering

- Imagine a setting with two binary features, corresponding to race (eg. black and white) and gender (eg. male and female), both randomly distributed in a population.
- A classifier labels an example positive iff it corresponds to a black man, or a white woman.
- The classifier labels 50% men and 50% women as positive, and 50% black and 50% white individuals as positive. => Fair w.r.t. demographic parity if we only consider <u>one</u> protected attribute at a time!

Individual Fairness

Individual Fairness

Assuming we have a metric *dist*() describing the similarity between pairs of individuals for the classification task at hand,

similar individuals are treated similarly.

$$D(M(x), M(x')) \leq dist(x, x')$$

Individual Fairness

Assuming we have a metric *dist*() describing the similarity between pairs of individuals for the classification task at hand,

similar individuals are treated similarly.

$$\boxed{D(M(x), M(x'))} \leq \boxed{dist(x, x')}$$

Distance between distributions over outcomes

Distance between individuals w.r.t. the task

Individual Fairness

Assuming we have a metric *dist*() describing the similarity between pairs of individuals for the classification task at hand,

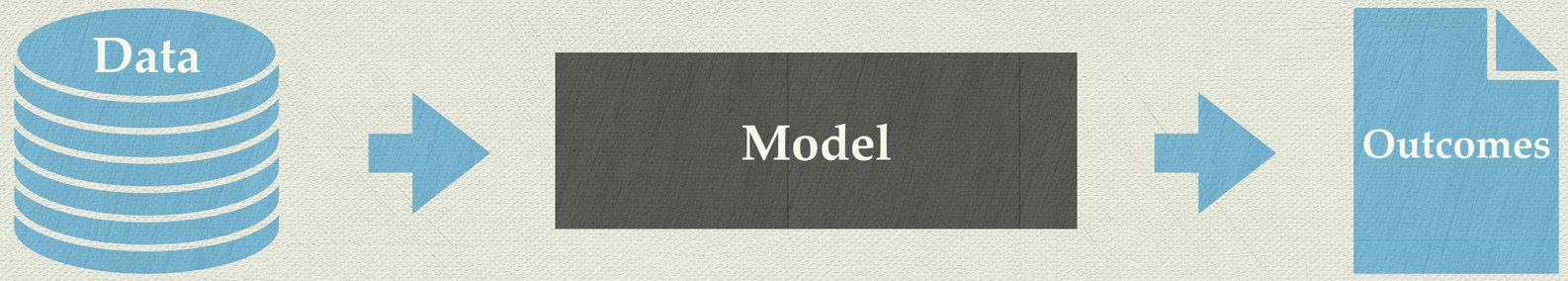
similar individuals are treated similarly.

$$\boxed{D(M(x), M(x'))} \leq \boxed{dist(x, x')}$$

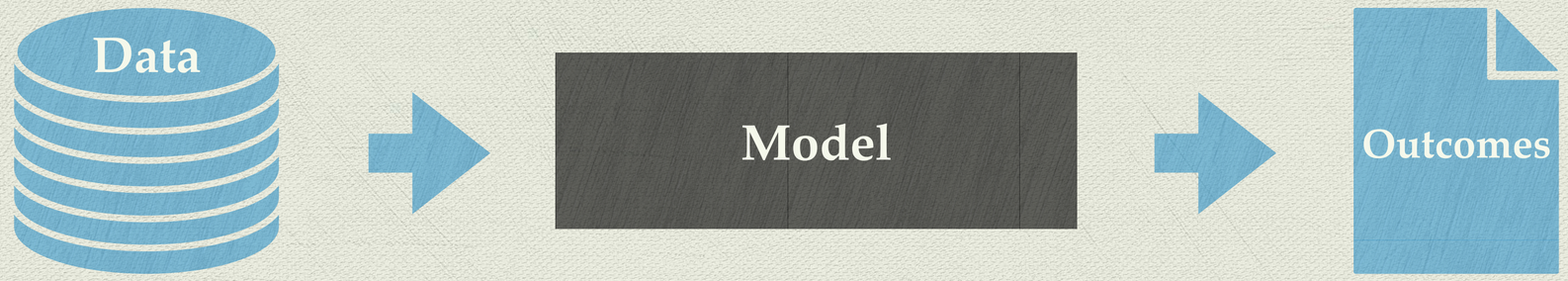
Distance between distributions over outcomes | **Distance between individuals w.r.t. the task**

Unfortunately, the need for a task-specific similarity metric has prevented its use in practice.

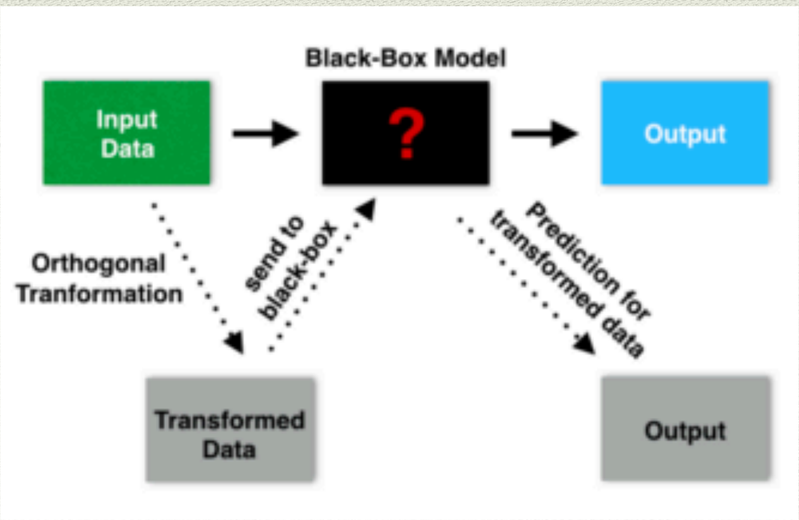
Techniques for making algorithms fair



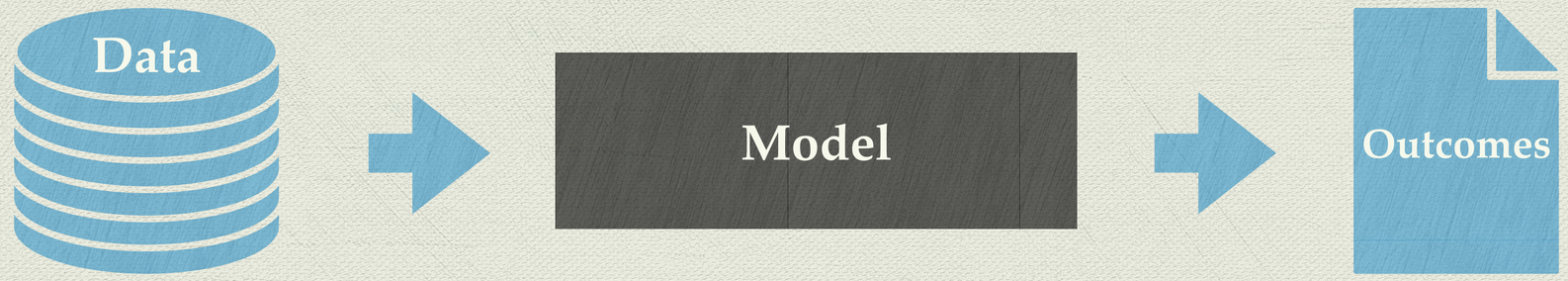
Techniques for making algorithms fair



- Data-based methods (during preprocessing)
  - Adjust the feature space to be uncorrelated w/ the sensitive attribute.

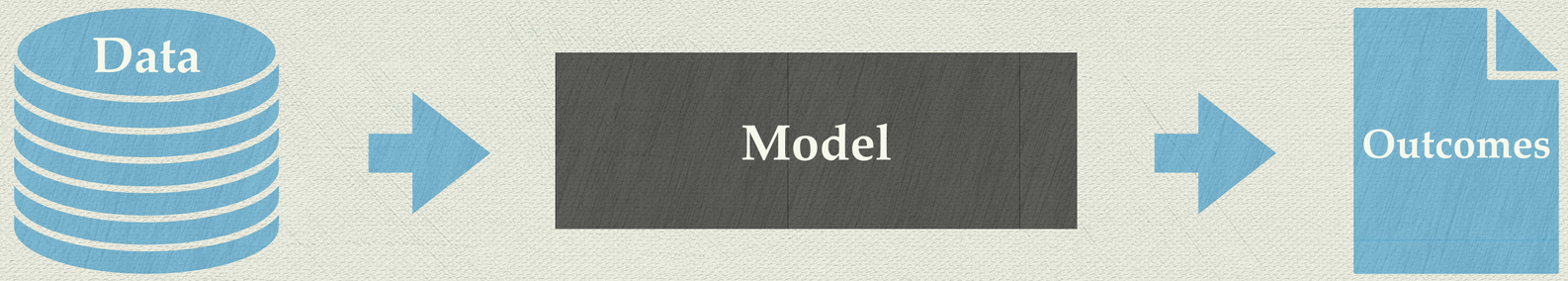


Techniques for making algorithms fair



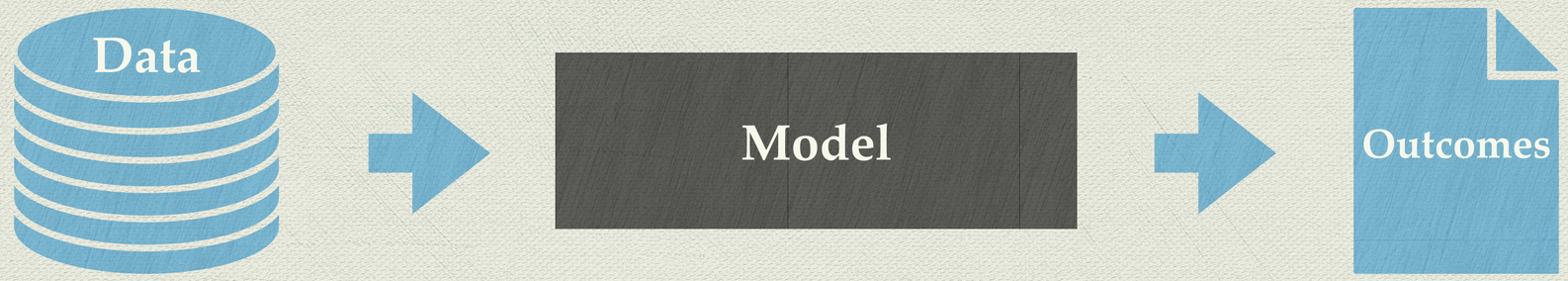
- Data-based methods (during preprocessing)
  - Adjust the feature space to be uncorrelated w/ the sensitive attribute.
- Model-based methods (at training time)
  - Include fairness constrains in the optimization process/loss function.

Techniques for making algorithms fair



- Data-based methods (during preprocessing)
  - Adjust the feature space to be uncorrelated w/ the sensitive attribute.
- Model-based methods (at training time)
  - Include fairness constrains in the optimization process/loss function.
- Post-hoc methods (during post-processing)
  - Adjust the output to be uncorrelated with the sensitive attribute.

Techniques for making algorithms fair



- Data-based methods (during preprocessing)
  - Adjust the feature space to be uncorrelated w/ the sensitive attribute.
- Model-based methods (at training time)
  - Include fairness constrains in the optimization process/loss function.
- Post-hoc methods (during post-processing)
  - Adjust the output to be uncorrelated with the sensitive attribute.

Fairness guarantees

Trade-off

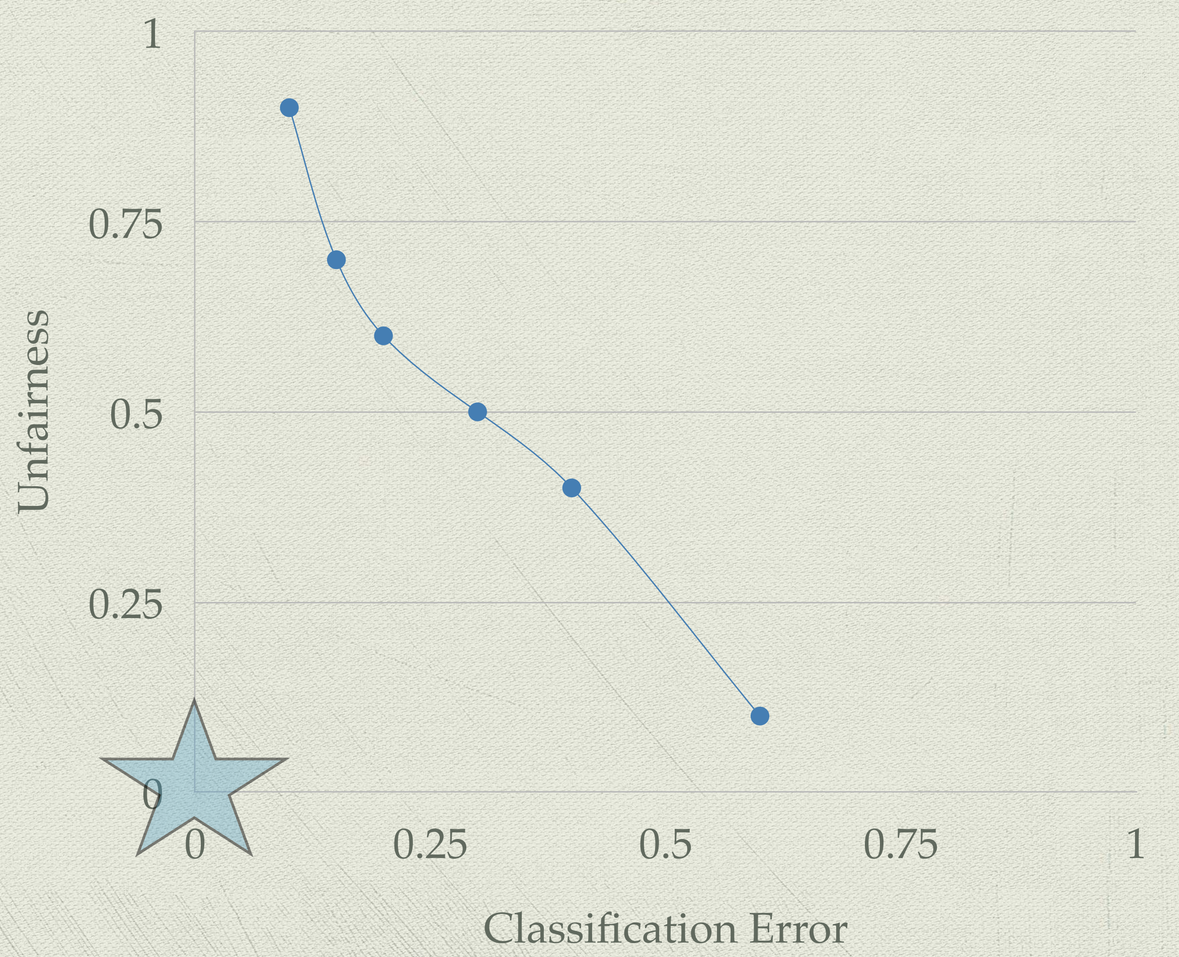
For the majority of the cases, considering fairness introduces some degradation in the quality of the predictions in terms of accuracy.

Pareto curve

Unfairness =
Recall($S = 1$) - Recall($S = 0$)

Classification Error =
1 - Accuracy

Pareto frontier:
There is no model that can achieve less unfairness with less classification error.

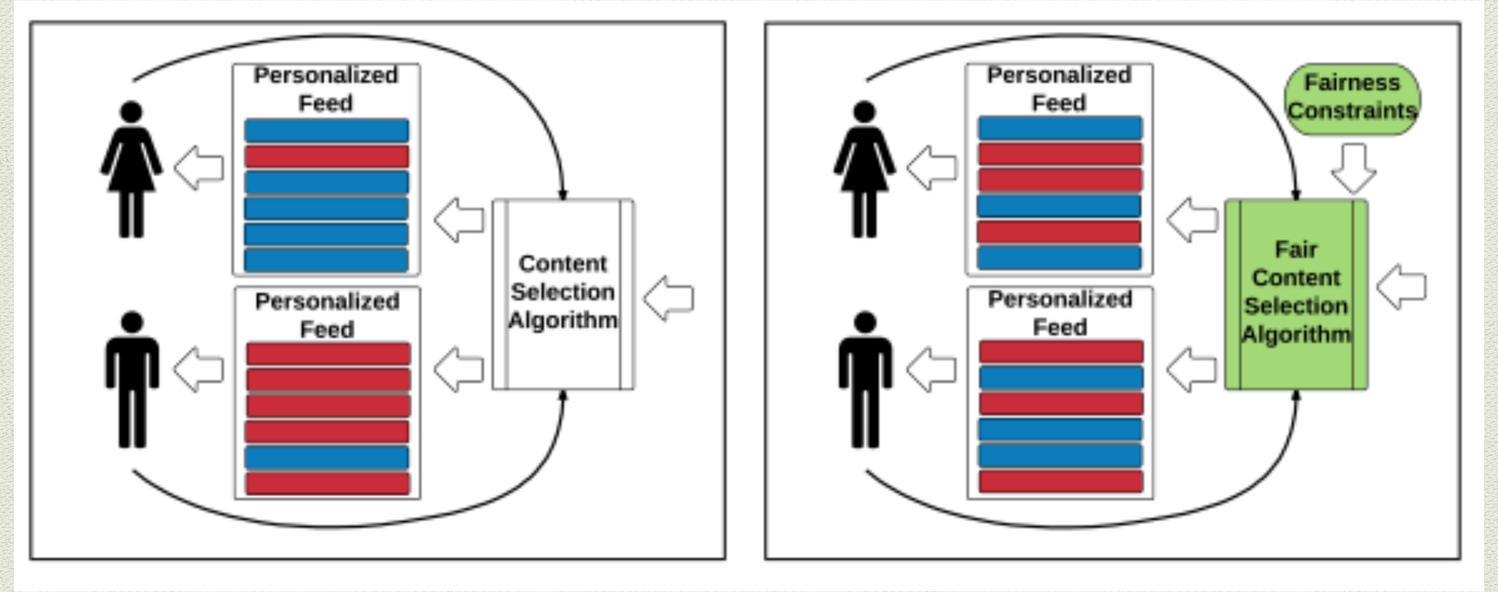


Fairness in Recommender systems

- Since biases can be inherent in the input data,
- a recommender is regarded **fair** when it does not amplify existing biases, i.e., avoids bias disparity.
- Specifically, the ratio of objects from various groups (output bias) is the same as the ratio present in the subject's input preferences (input bias).

diversity in recommendations

Recommendation (eg. news articles)



Celis, L. Elisa, Sayash Kapoor, Farnood Salehi, and Nisheeth Vishnoi. "Controlling Polarization in Personalization: An Algorithmic Framework." In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 160–69, 2019.

Thank you all!!

Questions? Concerns?

ap2744@georgetown.edu