


# Fairness

Machine Learning for Behavioral Data

May 8, 2023

# Today's Topic

Week	Lecture/Lab
8	Spring Break
9	Time Series Prediction
10	Unsupervised Learning
11	Unsupervised Learning
<b>12</b>	<b>Fairness</b>
13	Explainability
14	Project Presentations
15	Whit Monday

- 
- What is fairness?
  - Fairness metrics
  - Interpreting neural networks

# Getting ready for today's lecture...

- **If not done yet:** clone the repository containing the Jupyter notebook and data for today's lecture into your Noto workspace.
- SpeakUp room for today's lecture:

<https://go.epfl.ch/speakup-mlbd>



# Agenda

- 1) **Introduction to fairness – Cécile Hardebolle**
  - 2) Fairness in machine learning:
    - Sources of unfairness
    - Fairness metrics – evaluating model predictions
  - 3) Example on real world data
-

# Agenda

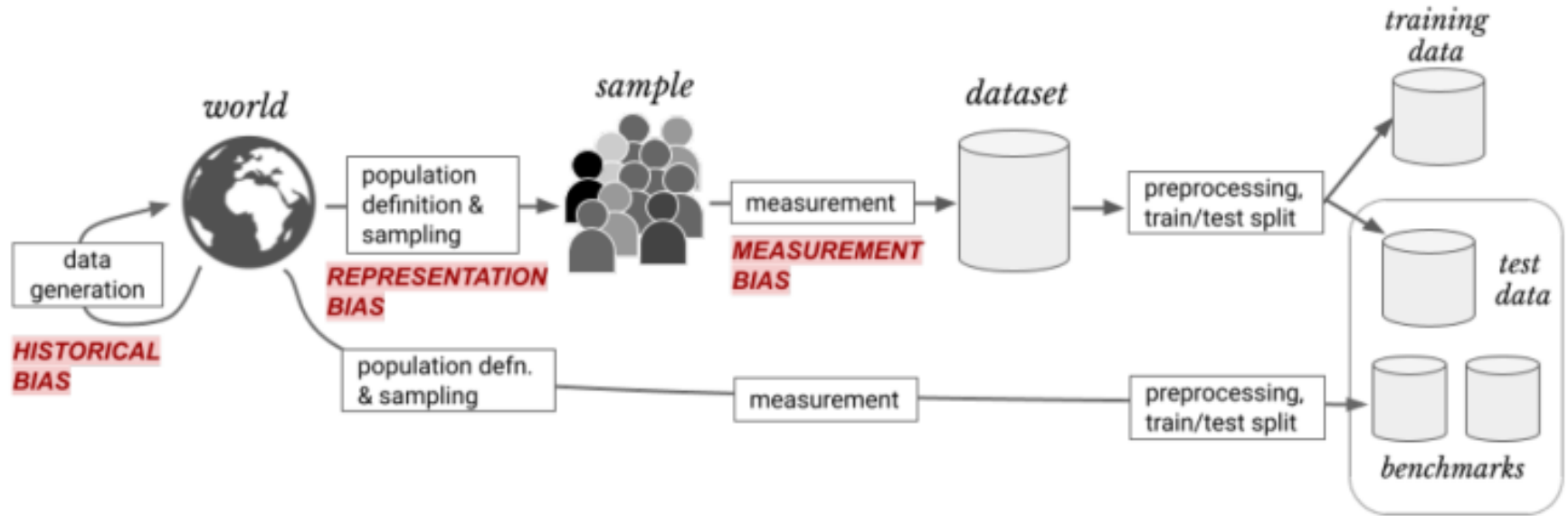
- 1) Introduction to fairness – Cécile Hardebolle
  - 2) **Fairness in machine learning:**
    - Sources of unfairness
    - Fairness metrics – evaluating model predictions
  - 3) Example on real world data
-

# Learning Objectives

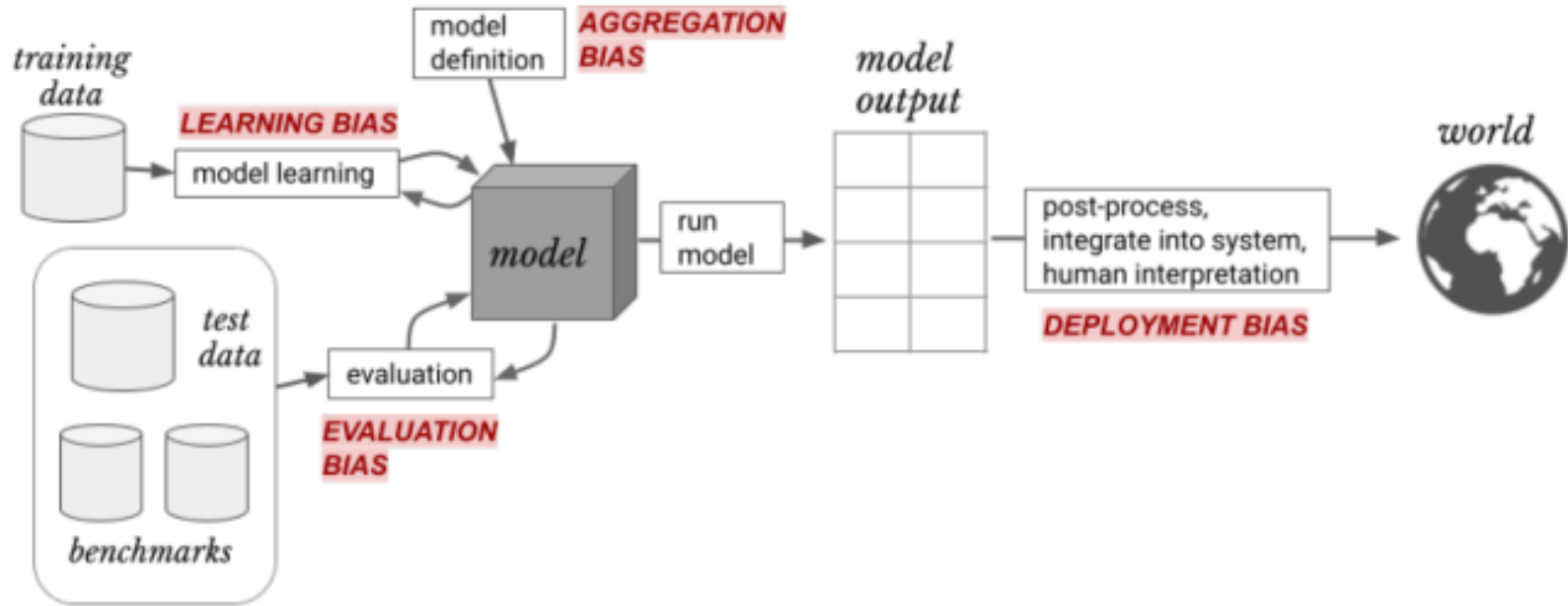
You should be able to:

- Name and explain the sources of unfairness in a machine learning pipeline
  - Explain and implement the most popular metrics for fairness
  - Perform a fairness evaluation of a machine learning model using an appropriate fairness metric
-

# Sources of Unfairness – Data Generation



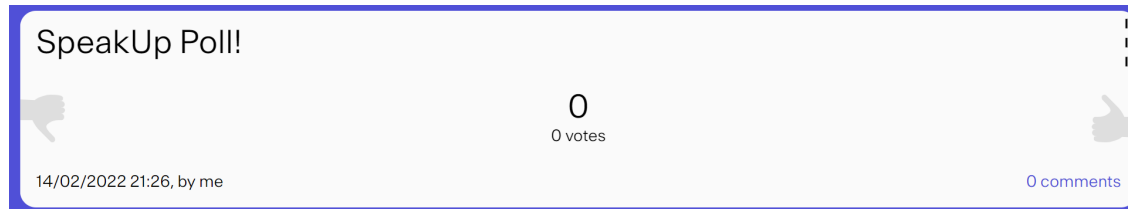
# Sources of Unfairness – Model Building





# Fairness Through Blindness

- Idea: we ignore all protected attributes in our model (e.g., we do not use protected attributes such as gender, race, etc. as features)



**Will this idea lead to a fair model?**

a) Yes

b) No

# Fairness Through Awareness

- There is not one mathematically agreed definition of fairness
  - Popular fairness metrics are
    - model-agnostic
    - defined for classification problems
-

# Problem Formalization

Notation:

- $X$  is the input to the model
  - $\hat{Y}$  is the prediction of the model
  - $T$  is the true label
  - $A$  is the protected attribute
-

# Confusion Matrix

		True Label	
		$T = 1$	$T = 0$
Predicted Label	$Y = 1$	True Positive (TP)	False Positive (FP)
	$Y = 0$	False Negative (FN)	True Negative (TN)

# Demographic Parity

- Requires equal proportion of positive predictions in each group

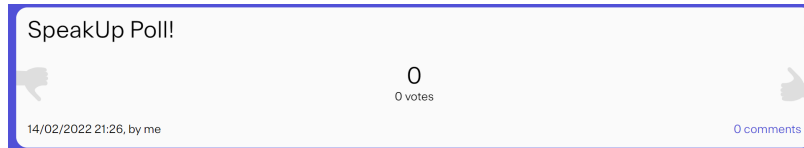
$$p(\hat{Y} = 1 | A = 1) = p(\hat{Y} = 1 | A = 0)$$

# Demographic Parity - Example

- Admittance to *Fruits* University
- Students from two schools apply: *Apple* and *Peach*
- School *Peach* has a better program, resulting in more qualified students

<i>Peach</i>	Qualified	Unqualified
Admitted	45	2
Rejected	45	8

<i>Apple</i>	Qualified	Unqualified
Admitted	5	18
Rejected	5	72



Is demographic parity fulfilled?

a) Yes

b) No

# Equalized Odds

- For any label and attribute, a classifier predicts the label equally well for all values of that attribute

$$p(\hat{Y} = 1 | A = 1, T = 1) = p(\hat{Y} = 1 | A = 0, T = 1)$$

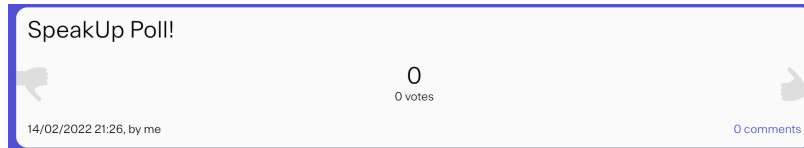
$$p(\hat{Y} = 1 | A = 1, T = 0) = p(\hat{Y} = 1 | A = 0, T = 0)$$

# Equalized Odds - Example

- Admittance to *Fruits* University
- Students from two schools apply: *Apple* and *Peach*
- School *Peach* has a better program, resulting in more qualified students

<i>Peach</i>	Qualified	Unqualified
Admitted	45	2
Rejected	45	8

<i>Apple</i>	Qualified	Unqualified
Admitted	5	18
Rejected	5	72



Are equalized odds fulfilled?

a) Yes

b) No



# Predictive Value Parity

- Probability of a sample with positive (negative) predictive value to truly belong to the positive (negative) class should be the same across attributes

$$p(T = 1|A = 1, \hat{Y} = 1) = p(T = 1|A = 0, \hat{Y} = 1)$$

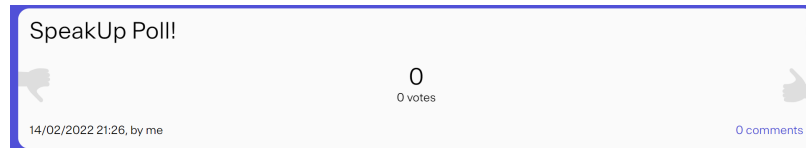
$$p(T = 0|A = 1, \hat{Y} = 0) = p(T = 0|A = 0, \hat{Y} = 0)$$

# Predictive Value Parity - Example

- Admittance to *Fruits* University
- Students from two schools apply: *Apple* and *Peach*
- School *Peach* has a better program, resulting in more qualified students

<i>Peach</i>	Qualified	Unqualified
Admitted	45	2
Rejected	45	8

<i>Apple</i>	Qualified	Unqualified
Admitted	5	18
Rejected	5	72



Is predictive value parity fulfilled?

a) Yes

b) No

# Impossibility Result

- Any two of the three criteria are mutually exclusive
    1. If  $A$  and  $T$  are not independent, then *demographic parity* and *predictive value parity* cannot simultaneously hold
    2. If  $A$  and  $\hat{Y}$  are not independent of  $T$ , then *demographic parity* and *equalized odds* cannot simultaneously hold
    3. If  $A$  and  $T$  are not independent, then *equalized odds* and *predictive value parity* cannot simultaneously hold
-

# Impossibility Result

Note that these requirements hold for *most* classifiers in real contexts:

- Base-rates of outcomes rarely are equal across groups
  - $A$  and  $T$  are usually associated when issues of fairness are relevant for the group in question
  - $\hat{Y}$  and  $T$  are usually associated, if your classifier is any good
-

# Agenda

- 1) Introduction to fairness – Cécile Hardebolle
  - 2) Fairness in machine learning:
    - Sources of unfairness
    - Fairness metrics – evaluating model predictions
  - 3) **Example on real world data**
-

# Flipped Classroom – Your Turn

- Participants: 214 EPFL students of a course taught in *flipped classroom* mode with a duration of 10 weeks
  - We have trained a classifier to predict whether a student will pass or fail the course based on their clickstream data
  - Your task:
    1. Choose one of the fairness metrics introduced in class and compute the metric for the flipped classroom classifier
    2. Tell us: is the classifier fair according to the selected metric? Why did you choose this metric?
-

# Summary

- There are multiple sources of unfairness in a machine learning pipeline
  - There is no consensus on the mathematical definition of fairness metrics
  - Different metrics assess different aspects of the classifier
  - Often, fairness metrics are mutually exclusive
  - Fairness evaluation of a classifier includes exploration of relevant characteristics of our data
-