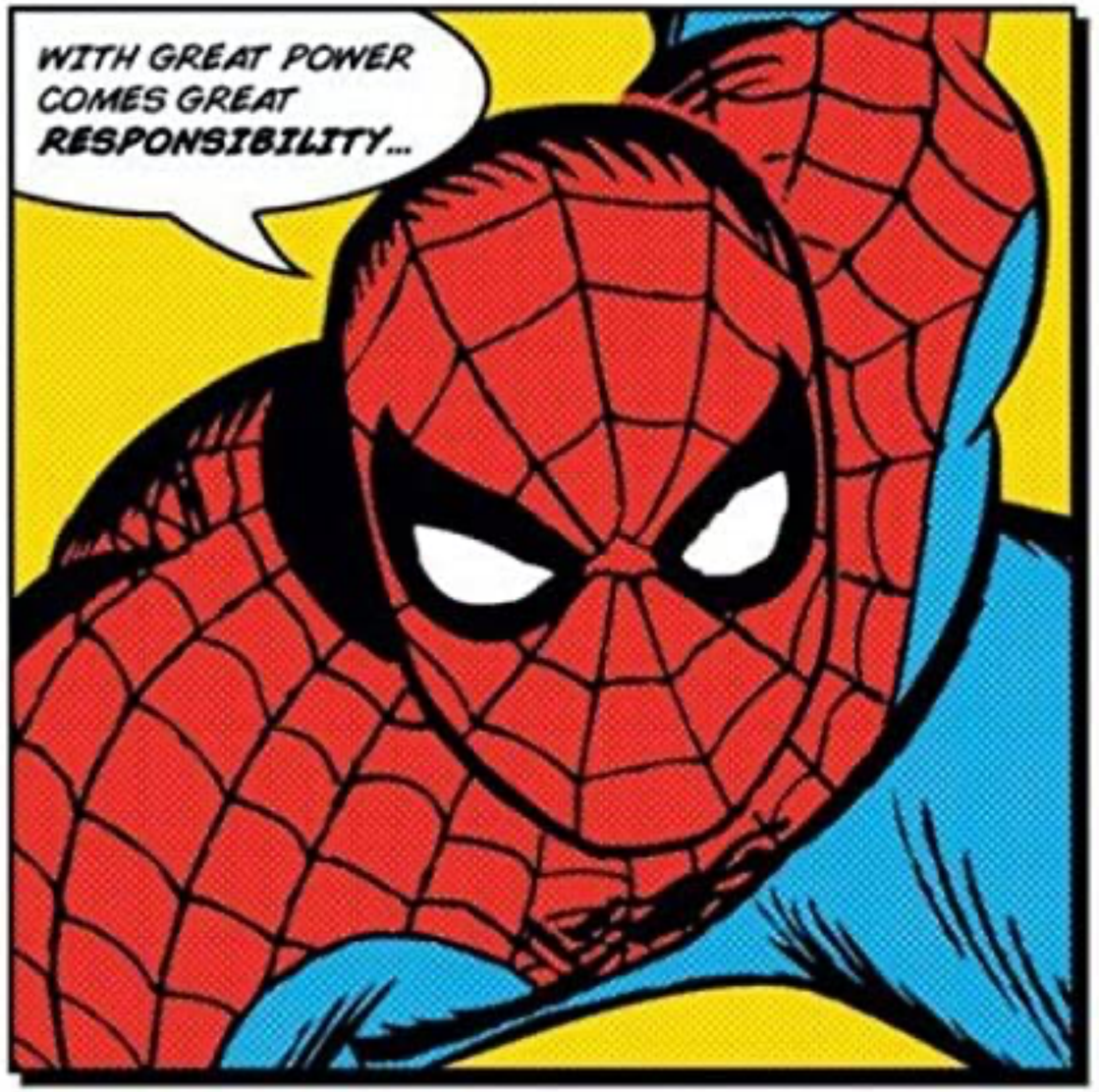
Ethics and Fairness in ML

Machine Learning Course - CS-433
18 Nov 2025
Robert West
(Slide credits: Nicolas Flammarion)

EPFL


WITH GREAT POWER
COMES GREAT
RESPONSIBILITY...

Special thanks - disclaimer - ©

These slides are essentially based on:

- The MLSS and NeurIPS tutorials of Moritz Hardt mrtz.org/
- The book *Fairness and Machine Learning* of Solon Barocas, Moritz Hardt, and Arvind Narayanan fairmlbook.org

And also:

- The lecture of Nathan Kallus at Cornell

Moritz Hardt

Failure of fairness through unawareness

Amazon uses data-driven techniques to decide the neighborhoods it will offer free same-day delivery

Disparities in the demographic makeup of these neighborhoods

→ White residents were more than twice as likely as Black residents to have access to this service

Article from Bloomberg, 2016

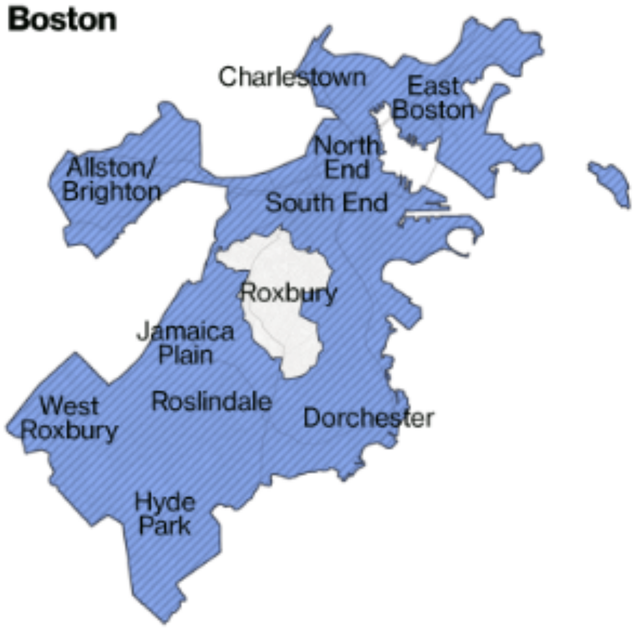
Failure of fairness through unawareness

Certainly, Amazon was just predicting a number of purchases, which correlates with wealth which correlates with race in the US.

They did not look at their customers' race when building their product

Example of just using ML without concern for fairness issues which leads to ethical issues

Discarding "sensitive attributes" does not solve the fairness problem and can aggravate them



Article from Bloomberg, 2016

Discrimination in ML

Discrimination: didn't we actually learn how to discriminate in the previous lectures?

We will be concerned with **unjustified bases for differentiation**:

- Practical irrelevance
  - Sexual orientation in employment decisions
- Moral irrelevance
  - Disability status in hiring decisions

Discrimination: didn't we actually learn how to discriminate in the previous lectures?

We will be concerned with **unjustified bases for differentiation**:

- Practical irrelevance

  Sexual orientation in employment decisions

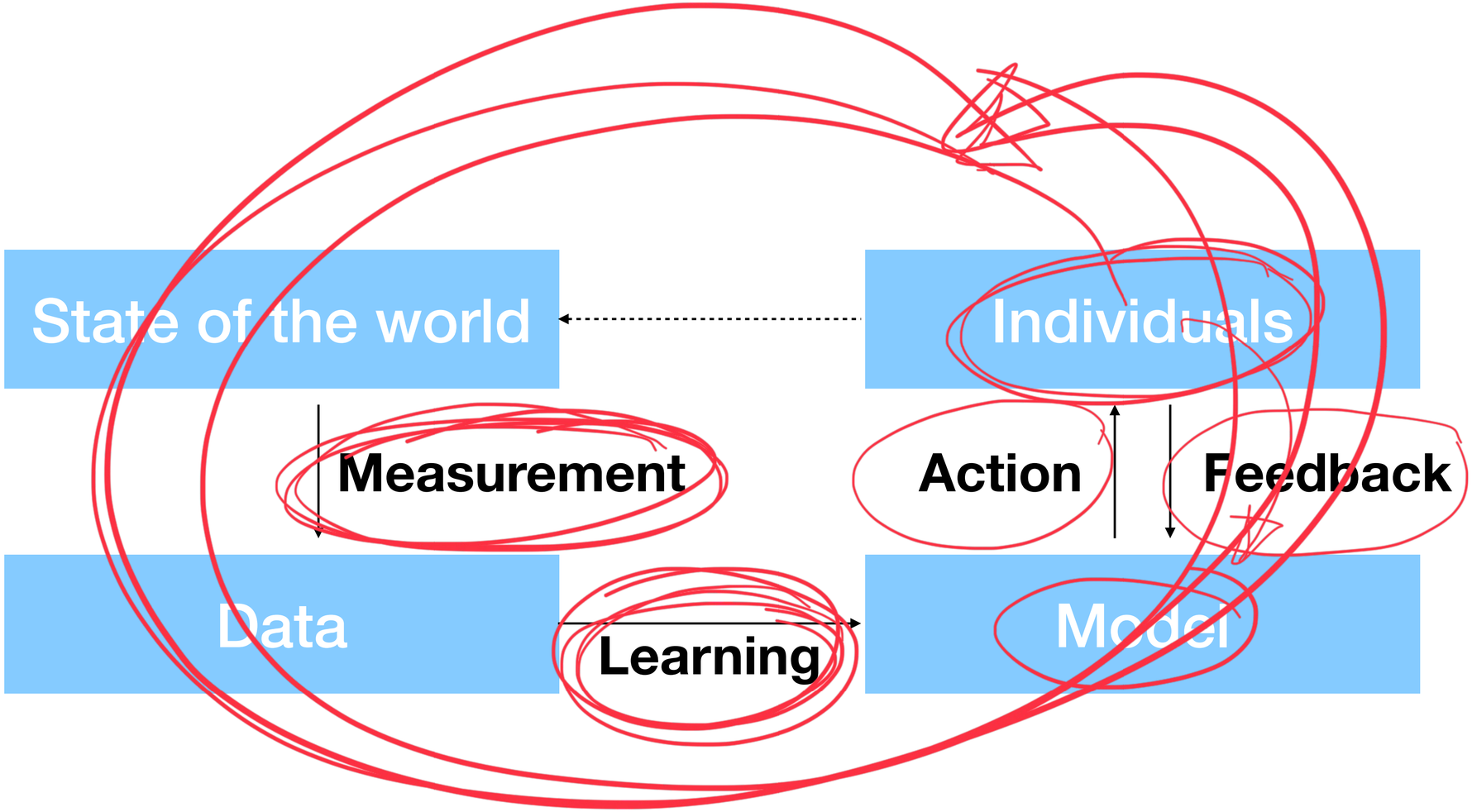
- Moral irrelevance

  Disability status in hiring decisions

Discrimination is **domain-specific**: concerned with opportunities that affect people's lives

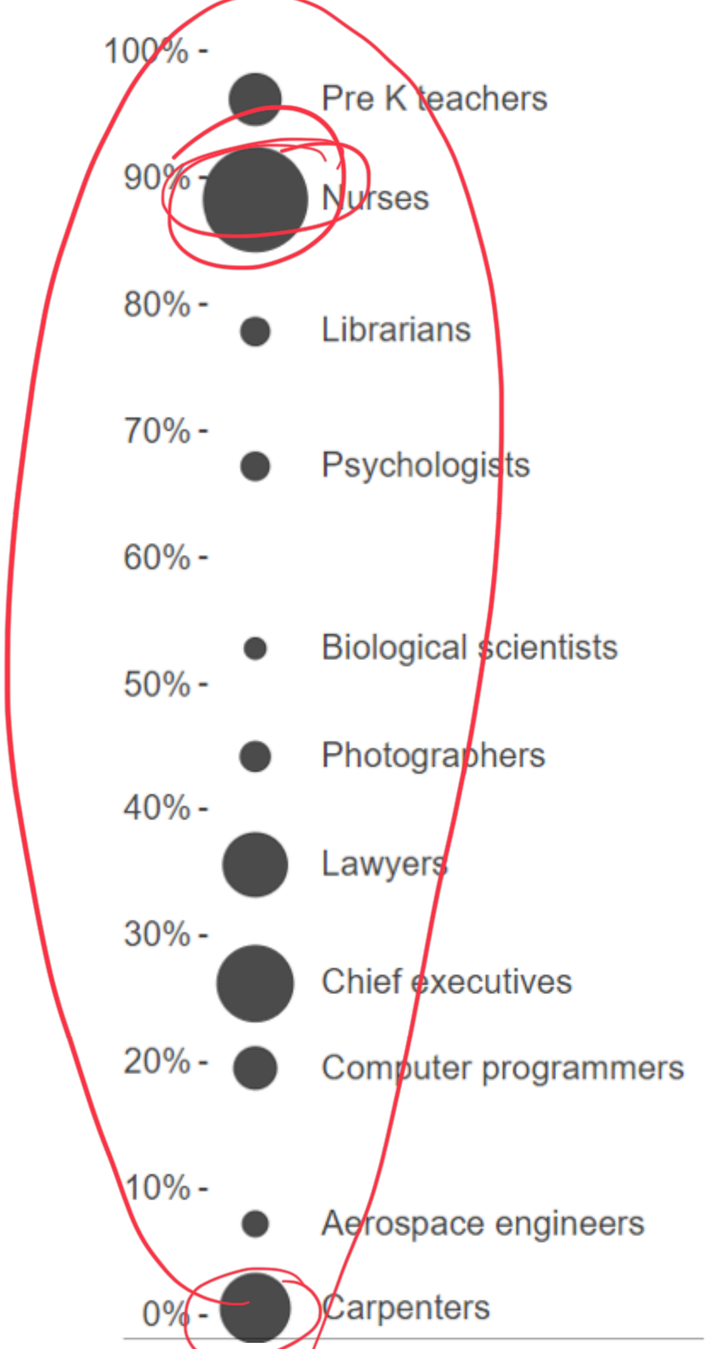
Discrimination is **group-specific**: concerned with social categories that have served as the basis for unjustified and systematically adverse treatment in the past

The machine learning loop



Applications about people

- Most ML applications are about people: 14 out of the top 30 Kaggle competitions concern tasks where decisions are made about individuals
- Training data often encode existing demographic disparities
- Social stereotypes may be perpetuated by applications of ML algorithms to these tasks
  Ex: Automated essay scoring: train data come from human graders with possible stereotype



Percentage of women in a sample of occupations in USA

Applications that are not about people?

Example borrowed from Delip Rao

Measurements are not without problems

Measurement involves

- Defining your variable of interest (e.g., how to measure intelligence, credit score, crime rate, etc.?)
- Defining the process for interacting with the real world (e.g., questionnaires, images, interviews, etc.?)
- Turning observations into numbers. i.e., collecting the data (e.g., camera color balance)

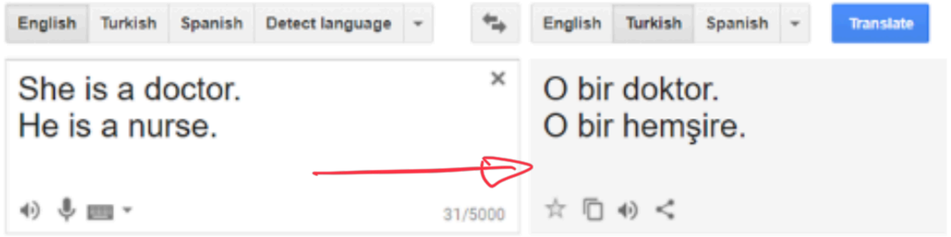
Measuring attributes of people can be subjective and challenging

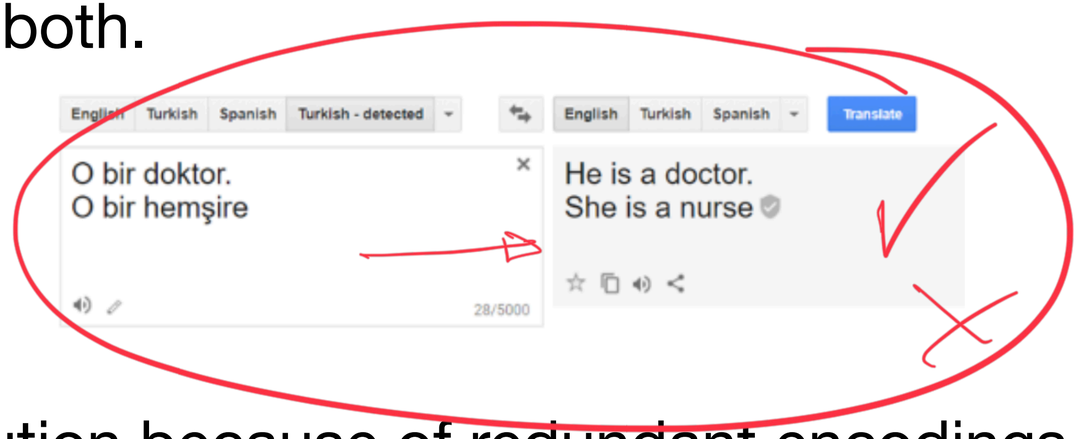
It is crucial to understand the provenance of the data as a practitioner

From data to models: disparities can be preserved

Some patterns in the training data represent **knowledge** - we want to learn, while other patterns represent **stereotypes** - we want to avoid learning.

→ML algorithms cannot distinguish between these two. Without specific intervention, ML algorithms will extract both.

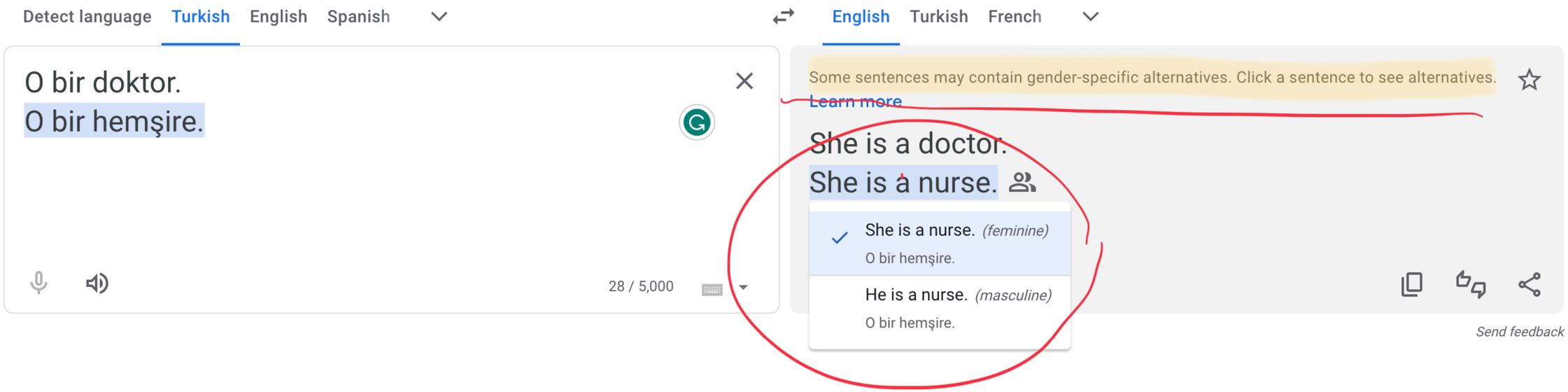




Removing, e.g., the gender is not a solution because of redundant encodings, i.e., other attributes (e.g., profession) that may correlate with, e.g., the gender.

These redundant encodings may be relevant to the problem at hand.

From data to models: disparities can be preserved but should be fixed



From data to models: disparities can be introduced

Sample size disparity:

- Uniform subsampling from population leads to fewer data about minorities
- If minority groups are in addition underrepresented, then even fewer data

ML works best with a lot of data → ML may work less well for minorities

True error is an average criterion → low true error may hide terrible performance for a minority group

Pseudonym

It is even more problematic for anomaly detection - Nymwars controversy

Conclusion: learning algorithms generalize based on the majority culture, leading to higher error rate for minority groups. This is because of our goal to avoid overfitting.

Toy example

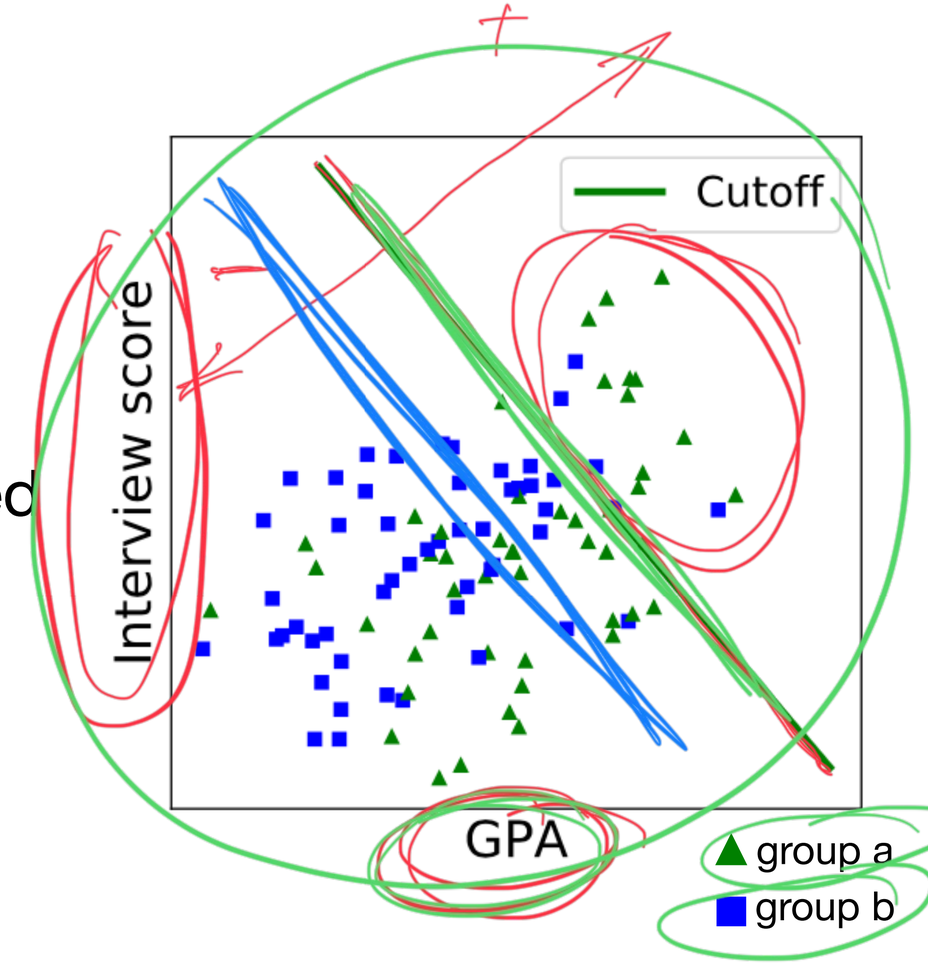
A hiring committee makes decisions based on applicants' college GPA and interview score

The classifier does not take into account which group a candidate belongs to

However, the triangles are more likely to be selected than the squares: the ground truth we predict is systematically lower for the squares than for the triangles

Why:

- Post-hiring: bias inside the company
- Pre-hiring: disparities in education
- Combinations of different factors



Data from past candidates

How to decrease the disparity of our selection procedure?

1. GPA is correlated with the demographic attributes - proxy
   → But simply omitting it would decrease the accuracy of our model

2. Pick different cutoffs so that candidates from both groups have the same probability of being hired
   → but two candidates with the same attributes may receive different decisions depending on their groups

3. Change the model to weigh less the GPA and increase diversity between selected candidates

ONE SIZE
DOESN'T FIT ALL

Fairness criteria in classification

Formal setting: classification

Data are described by covariates X and outcomes variable $Y \in \{0,1\}$

Goal: given a new X you want to predict its label Y

How:

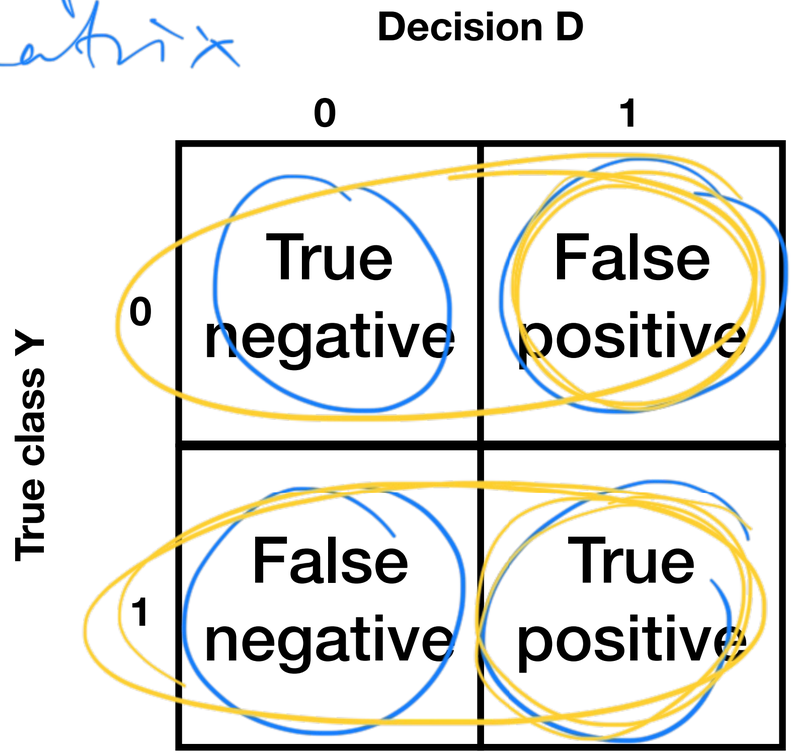
1. Use an algorithm to produce a score function $R = r(X)$
   - Bayes-optimal score: $\mathbb{P}_{\mathcal{D}}\{Y = 1 | X\}$
   - Learned from labeled data, e.g., in logistic regression
2. Make binary decisions according to the threshold rule $D = 1_{R>t}$

Today: we assume R given and are interested in the decision process

Statistical classification criteria

confusion matrix

$(D=1 \wedge Y=0) \vee (D=0 \wedge Y=1) \equiv D \neq Y$



True positive rate: $\mathbb{P}(D = 1 \mid Y = 1)$

False positive rate: $\mathbb{P}(D = 1 \mid Y = 0)$

True negative rate: $\mathbb{P}(D = 0 \mid Y = 0)$

False negative rate: $\mathbb{P}(D = 0 \mid Y = 1)$

The choice of the threshold t in the decision rule D will depend on the classification criteria we pick

Sensitive attributes

In many tasks, X can encode sensitive attributes of an individual

We introduce additional random variable A encoding membership status in a protected class

No fairness through unawareness: removing/ignoring sensitive attributes is not solving the problem

- Many features slightly correlated with the sensitive attribute can be used to recover the attribute
- If we remove the attribute, the classifier may still find a redundant encoding in terms of other features and we'd learn an equivalent classifier

Three fundamental fairness criteria

Idea: equalize different statistical quantities involving group membership A

$\rightarrow$ it dates back to the 1960s with the seminal work of Anne Cleary

Most of the fairness criteria are properties of (A, Y, R):

- Independence: R independent of A
- Separation: R independent of A, conditional on Y
- Sufficiency: Y independent of A conditional on R

Independence: equalizing acceptance rate

It requires the random variables A and R to be independent, denoted by $A \perp R$

Implies, for any two groups a, b:

$$\mathbb{P}(D = 1 \mid A = a) = \mathbb{P}(D = 1 \mid A = b)$$

→ The acceptance rate is the same in all groups: equal positive rate

Limitations of independence

This criterion does not rule out unfair practice. Let's imagine a company which

- hires with care (i.e., makes good decisions) in a group a at some rate $p > 0$
- hires without care (i.e., makes poor decisions) in a group b with the same rate p

→ acceptance in both groups is identical

→ unqualified applicants are more likely to be selected in the group b

→ members of the group b will appear to perform less well than those of a

It can happen on its own if there is less data in one group

A positive output can either be a false positive or a true positive

→ we shouldn't be able to match true positives in one group with false positives in another

Separation: equalizing error rates

It requires the random variables A and R to be independent conditional on the target variable Y, denoted by $A \perp R \mid Y$

It implies for all groups a, b:

$$\mathbb{P}(D = 1 \mid Y = 0, A = a) = \mathbb{P}(D = 1 \mid Y = 0, A = b) \quad \text{(equal false positive rate)}$$

$$\mathbb{P}(D = 0 \mid Y = 1, A = a) = \mathbb{P}(D = 0 \mid Y = 1, A = b) \quad \text{(equal false negative rate)}$$

This is a **post-hoc criterion**: at decision time, we do not know who is a positive/negative instance

It can be computed in retrospect, by collecting groups of positive and negative instances

Sufficiency:

separation: $A \perp R \mid Y$

It requires the random variables A and Y to be independent conditional on R, denoted by $A \perp Y \mid R$ $\equiv \mathbb{P}(Y \mid R, A) = \mathbb{P}(Y \mid R)$ $\equiv \mathbb{P}(A \mid R, Y) = \mathbb{P}(A \mid R)$

For all groups a, b and values r we have:

$$\mathbb{P}(Y = 1 \mid R = r, A = a) = \mathbb{P}(Y = 1 \mid R = r, A = b) = \mathbb{P}(Y = 1 \mid R = r)$$

Meaning: for predicting Y we do not need to know A if we have R

Calibration and sufficiency

Def: A score R is calibrated if

$$\mathbb{P}(Y = 1 \,|\, R = r) = r$$

→ you can interpret your score as a probability

→ a priori guarantee: score value r corresponds to positive outcome rate r

Calibration by group:

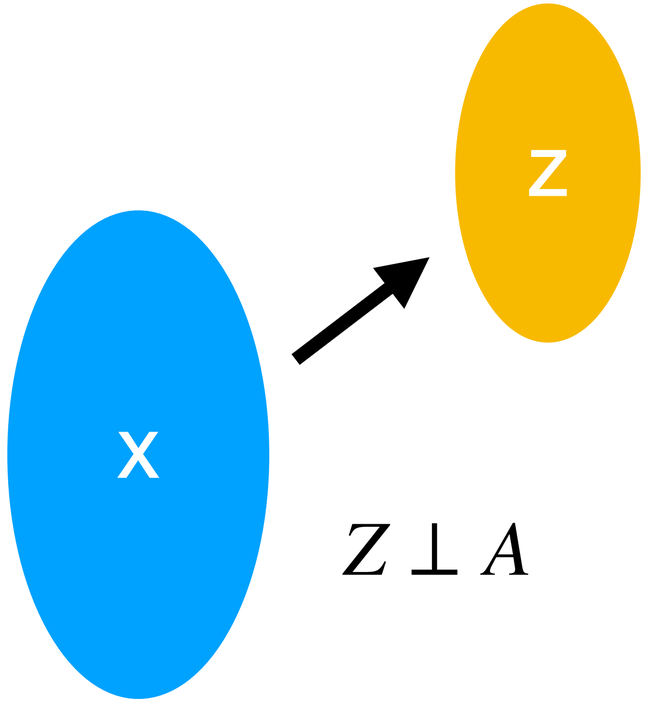
$$\mathbb{P}(Y = 1 \,|\, R = r, A = a) = r$$

Fact: Calibration by group implies sufficiency

Remark: it is also possible to go from sufficiency to calibration

How to achieve fairness criteria

- Post-processing: adjust your learned classifier so that it becomes uncorrelated with the sensitive attribute A
- At training time: work the constraint into the optimization process
- Pre-processing: adjust your features so that they become uncorrelated with the sensitive attribute A: e.g., use deep learning to learn a representation of the data independent of A, while representing original data as well as possible - Zemel et al., 2015



Can we satisfy them simultaneously?

Three criteria:

- Independence: R independent of A
  $\Longrightarrow$ equal acceptance rate
- Separation: R independent of A, conditional on Y
  $\Longrightarrow$ equal error rate
- Sufficiency: Y independent of A conditional on R
  $\Longrightarrow$ calibration by group

Informal theorem: any of these criteria are mutually exclusive - except in degenerate cases!

Recap

- ML models ultimately interact with the world, and their design should account for their impact. It's not only about the training.
- There is no fairness through unawareness. Naive data selection and ML techniques can perpetuate or introduce unwanted disparities. Careful pre-processing and post-processing are often necessary.
- We have examined statistical tools to formally reason about fairness criteria.

Bonus - Incompatibility results: trade-offs are necessary

1. Independence vs sufficiency: If A and Y are not independent, then sufficiency and independence cannot both hold

   Proof: $A \perp R$ and $A \perp Y \mid R \implies A \perp (Y, R) \implies A \perp Y$

2. Independence vs separation: if A is not independent of Y and R is not independent of Y, then independence and separation cannot both hold

   Proof: $A \perp R$ and $A \perp R \mid Y \implies A \perp Y$ or $R \perp Y$

Bonus - Proof of the second implication

Claim: $A \perp R$ and $A \perp R \mid Y \implies A \perp Y$ or $R \perp Y$

Proof: $\mathbb{P}(R = r \mid A = a) = \sum_y \mathbb{P}(R = r \mid A = a,\ Y = y)\mathbb{P}(Y = y \mid A = a)$

Since $A \perp R$ and $A \perp R \mid Y$:

$$\mathbb{P}(R = r) = \mathbb{P}(R = r \mid A = a) = \sum_y \mathbb{P}(R = r \mid Y = y)\mathbb{P}(Y = y \mid A = a)$$

We also have

$$\mathbb{P}(R = r) = \sum_y \mathbb{P}(R = r \mid Y = y)\mathbb{P}(Y = y)$$

Thus

$$\sum_y \mathbb{P}(R = r \mid Y = y)\mathbb{P}(Y = y \mid A = a) = \sum_y \mathbb{P}(R = r \mid Y = y)\mathbb{P}(Y = y)$$

Bonus - Proof of the second implication

Since $Y \in \{0,1\}$ it implies

$$\begin{aligned}\mathbb{P}(R = r \,|\, Y = 0)\mathbb{P}(Y = 0 \,|\, A = a) + \mathbb{P}(R = r \,|\, Y = 1)\mathbb{P}(Y = 1 \,|\, A = a) \\ = \mathbb{P}(R = r \,|\, Y = 0)\mathbb{P}(Y = 0) + \mathbb{P}(R = r \,|\, Y = 1)\mathbb{P}(Y = 1)\end{aligned}$$

It directly implies

$$\begin{aligned}\mathbb{P}(Y = 0)(\mathbb{P}(R = r \,|\, Y = 0) - \mathbb{P}(R = r \,|\, Y = 1)) \\ = \mathbb{P}(Y = 0 \,|\, A = a)(\mathbb{P}(R = r \,|\, Y = 0) - \mathbb{P}(R = r \,|\, Y = 1))\end{aligned}$$

Therefore either $\mathbb{P}(Y = 0) = \mathbb{P}(Y = 0 \,|\, A = a)$ and $A \perp Y$

Or $\mathbb{P}(R = r \,|\, Y = 0) = \mathbb{P}(R = r \,|\, Y = 1)$ and $Y \perp R$

Bonus - Incompatibility results: trade-offs are necessary

3. Separation vs sufficiency: Assume all events in the joint distribution of (A, R, Y) have positive probability and assume $A \not\perp\!\!\!\perp Y$. Then, separation and sufficiency cannot both hold
   Proof:
   $A \perp R \,|\, Y$ and $A \perp Y \,|\, R \implies A \perp (R, Y) \implies A \perp R$ and $A \perp Y$