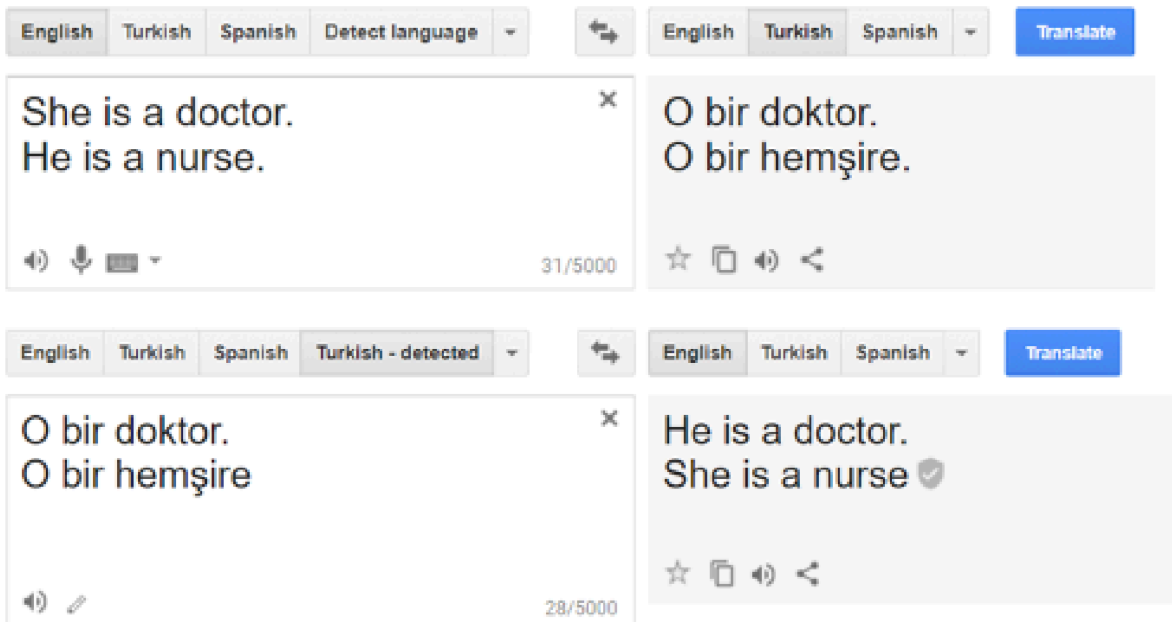
CS37300
PURDUE UNIVERSITY

DEC 1, 2023

DATA MINING

FAIRNESS



From the book “Fairness and Machine Learning”

EXAMPLE: AMAZON SAME-DAY DELIVERY

- 2016 study found dissimilarities in demographic racial make up of the places where it was offered
- Amazon insists the system is data-driven, no explicit use of race as a predictor
- Looking beyond intent: looking at discrimination and impact
- Impact analogy: click driven optimization can lead to echo chambers

MACHINE LEARNING LOOP

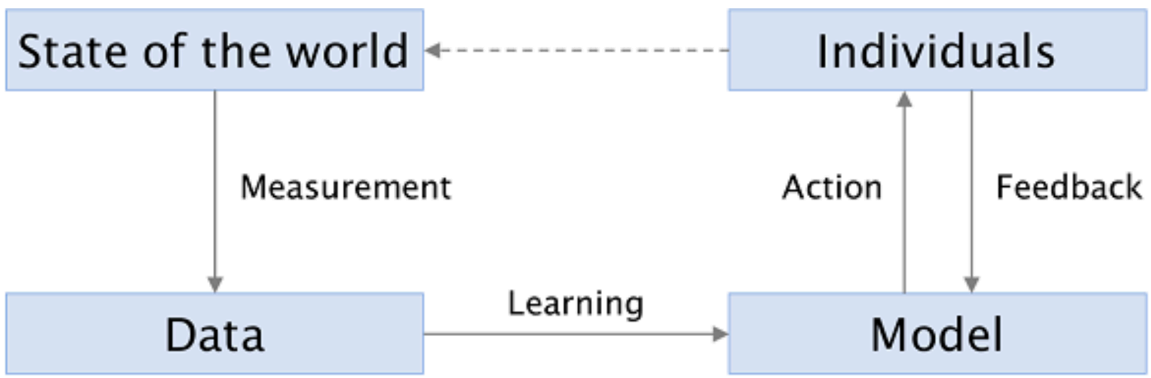


Image from "fairness and ML"

STATE OF SOCIETY: INHERENT DISPARITIES EXISTING IN THE SYSTEM

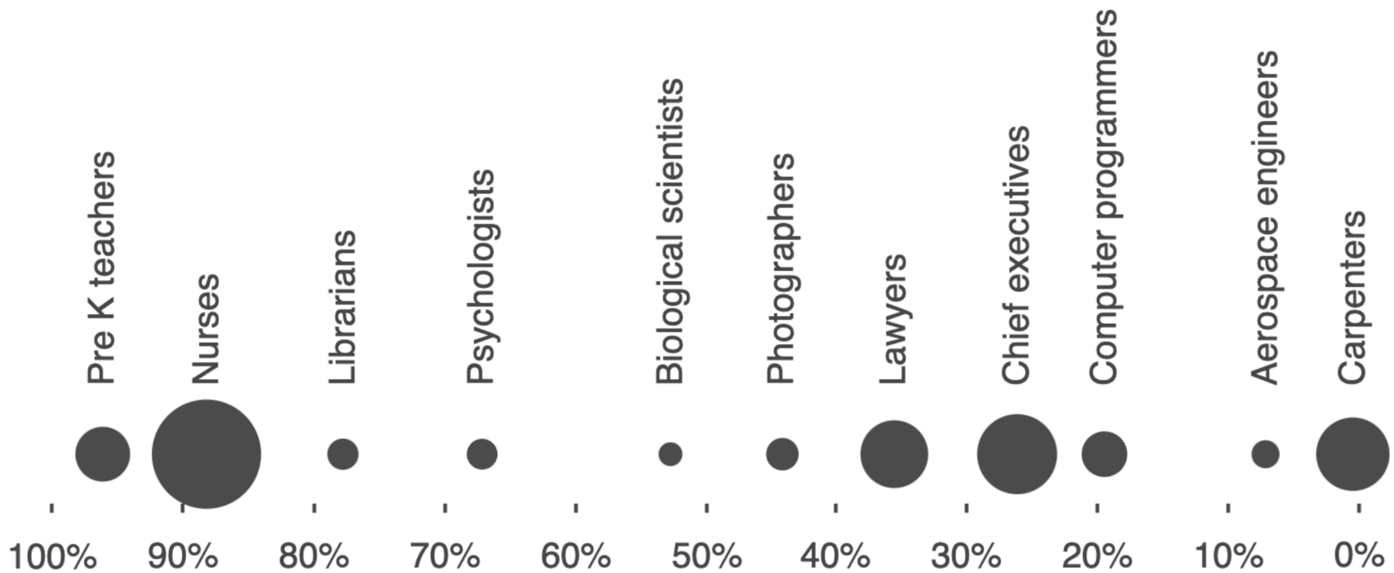


Image from “fairness and ML”

EXAMPLE: “STREET BUMP”

- Project by city of Boston to read smart phone sensors to pinpoint location of potholes
- Still biased towards people that own costlier smart phones
- ML systems rely on data collection, and the people component is almost impossible to overlook

CHALLENGES WITH MEASUREMENT

- 2017 NYT article about “Blacks and Hispanics are more underrepresented in colleges than in 1980s”
- Based on self-reporting of race
- “Multiracial” was introduced as a category only in 2008
- The study ignored this category, and its impact on reporting

MEASUREMENT: DEFINING THE TARGET VARIABLE

- Some cases are easy: click or no click
- Some cases are not that hard: movie ratings
- Some cases can be really hard: how do you define a "good employee", or a "successful student"
  - What's a good metric for a sales manager or a professor's promotion?
    - Every such decision impacts the employee's behavior

FROM DATA TO MODELS

- Stereotyped Correlation
- Causation
  - Statistical tools to test for causation using “interventions”
    - Would this applicant be still rejected if everything else remaining the same, the race was different?

A CASE AGAINST AUTOMATED DECISION MAKING

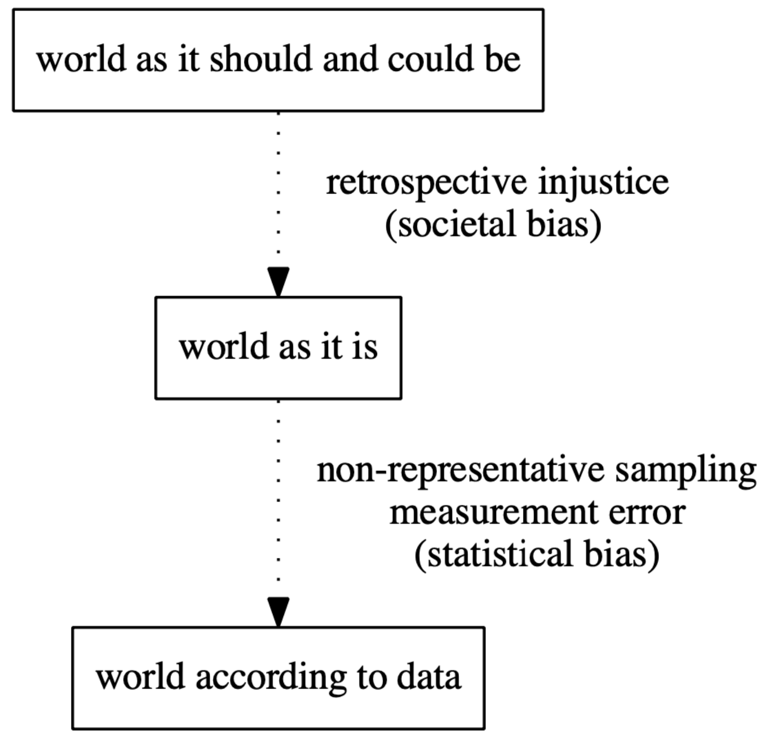
- In short: AI is not “there” yet
- Limited interpretability: spurious correlations
- Lack of inductive reasoning on handling unseen cases
- Un-representative data
- Definitions of targets
- Incentives propagate pre-existing “biases”
- Liability issues ? If a self-driving truck crashes due to a software issue, who is responsible?

GOALS OF A PREDICTIVE SYSTEM

- Examples:
  - Criminal Justice : predict probability of a repeat crime, or set policies to reduce the number of people to be detained without endangering public safety, or probability of failure to show up at a hearing
  - University admissions: probability of “success” of an admitted candidate
  - Social planner: Benevolent social welfare or reform
  - Omitted pay off bias

FAIRNESS BUCKETS

- Human biases in decision making
- Data biases: predictive models will predict/reproduce unfair patterns
  - What is the population? e.g. for loan applications, it is inherently people who are short on money. There may be unfairness within the process that exist for such a biased sample of the population



From: “Prediction-Based Decisions and Fairness: A Catalogue of Choices, Assumptions, and Definitions

DATA BIASES

- Selective labels problem
- Differential systematic measurement errors
  - Error is greater for some groups because of lack of proper “documented” measurements.
    - Immigrant lending example
    - Re-arrest probability not accounting the bias in decision to arrest in the first place
- Societal non-statistical biases

FIXING BIASES

- Preprocess the data to remove biases
- Regularize the learning mechanism
  - Modelling choices e.g. choose more interpretable models
- Post-process the learnt model
  - Interpretability is important

PROTECTED GROUPS

- “Sensitive” features with groupings into advatanged vs disadvantaged groups
- Goal: Uniformity in outcome across the protected features
- Example: Posterior prediction probabilities are equal across groups.

CLASSIFICATION FAIRNESS: EQUAL PREDICTION OUTCOME

- But score may include protected attributes
  - Solution: “protect” or remove these attributes from score calculation

- Equal accuracy check: For a protected binary attribute X_p with D as the prediction: $P(D = y \mid X_p = 0) = P(D = y \mid X_p = 1)$

CONFUSION MATRIX

	Y=1	Y=0	P(Y=1|D)	P(Y=0|D)
D=1	True Positive (**TP**)	False Positive (**FP**)	P(Y=1|D=1): Positive Predictive Value (**PPV**)	P(Y=0|D=0): False Discovery Rate (**FDR**)
D=0	False Negative (**FN**)	True Negative (**TN**)	P(Y=1|D=0): False Omission Rate (**FOR**)	P(Y=0|D=0): Negative Predictive Value (**NPR**)
P(D=1|Y)	P(D=1 | Y=1): True Positive Rate (**TPR**)	P(D=1|Y=0): False Positive Rate (**FPR**)		
P(D=0|Y)	P(D=0|Y=1): False Negative Rate (**FNR**)	P(D=0|Y=0): True Negative rate (**TNR**)		P(D=Y): Accuracy

Example Checks: TPR=FNR => $D \perp X_p \mid Y = 1$ (Equal opportunity).

$D \perp X_p \mid Y$ Equalized odds. "Similar people treated similarly".

$Y \perp X_p \mid D$ Sufficiency

From: "Prediction-Based Decisions and Fairness:
A Catalogue of Choices, Assumptions, and Definitions

PARITY ACROSS GROUPS

- AUC parity
- Balance for classes (negative vs positive)
- Calibration within groups

LACK OF HARMONY AMONG MEASURES

- The various definitions or constraints may not agree with each other
- COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) too deployed in criminal justice settings
  - Satisfies equal PPV. Those who were classified as high risk, the proportion of defendants re-arrested is approx equal regardless of race
  - Fails FPR: Defendants who did not get re-arrested, black defendants were twice as likely to be classified as high-risk
- Theoretical analyses have shown provable disagreement between some of these measures

EXAMPLE: FAIR FEATURE SELECTION

- Partition the features into non-overlapping groups.
- Can select only upto ki features group i.
    - Contrast with “can select k features out of d”
- Can employ greedy-like algorithms

PROCEDURAL FAIRNESS

- Based on perceived human notions of fairness rather than that of outcome
- Example: Feature selection based on what features are perceived fair
- In some cases can be enumerated as constraints, resulting in a constrained optimization problem.
- Example: Lets U_S be set of users that think feature S is fair.

$$h(\mathsf{T}) := 1 - \frac{\left|\bigcap_{s \in \mathsf{T}} \mathcal{U}_s\right|}{|\mathcal{U}|}$$