Bias in Machine Learning

Project Proposal

- Get started on the following
  - Forming groups
  - Selecting a topic
  - Outlining the idea
  - Writing the proposal: due by 11:59pm on Mar 10

Goals for today

- How does the machine learning literature approach fairness?

- What is “bias” and how is it different from “unfairness”?

- Examples of bias

- How does bias originate, how do we measure it, and how do we reduce it?

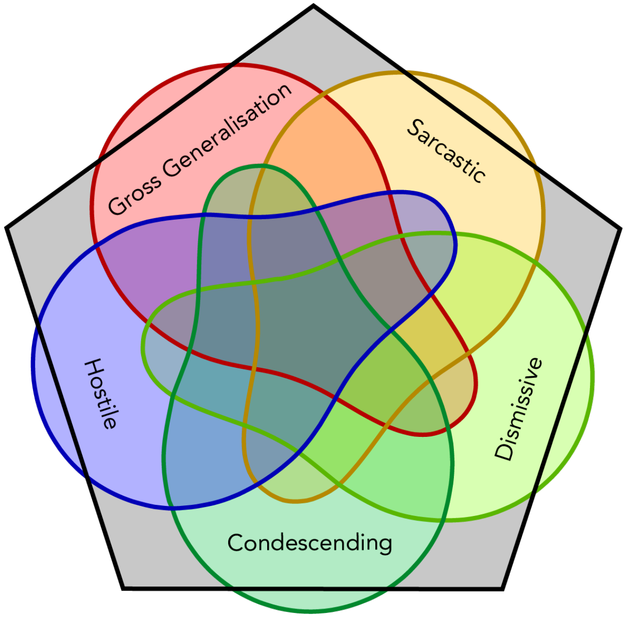
- This lecture will be a high-level overview; we will start technical details from the next lecture

Resume Screening

- Amazon’s AI tool to screen resumes learned to penalize resumes that contain the word “Women’s”, e.g., in “President of Women’s Chess Club”

Gizmodo. “Amazon's Secret AI Hiring Tool Reportedly 'Penalized' Resumes With the Word 'Women's’”

Toxicity Classification



Toxicity is defined as... "a rude, disrespectful, or unreasonable comment that is likely to make you leave a discussion.”

Medium article.
[“The Challenge of Identifying Subtle Forms of Toxicity Online”](#)

Toxicity Classification

Bias towards certain identity terms (comparing a toxicity score to what it ought to be)

Comment	Toxicity Score
The Gay and Lesbian Film Festival starts today.	0.82
Being transgender is independent of sexual orientation.	0.52
A Muslim is someone who follows or practices Islam.	0.46

Prabhakaran, Hutchinson, Mitchell.
"Perturbation Sensitivity Analysis to Detect Unintended Model Biases"

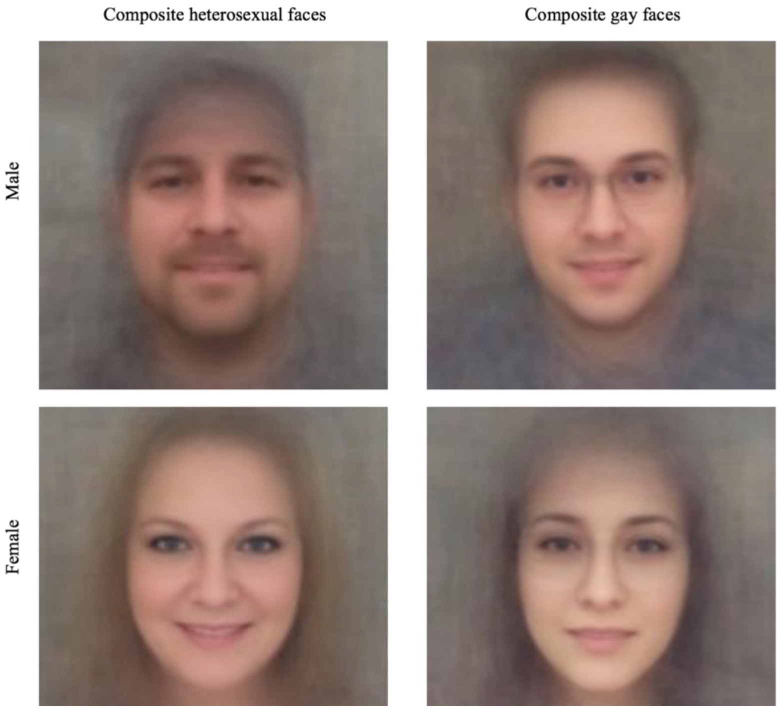
Toxicity Classification

Discrimination between genders (comparing a toxicity score to another toxicity score)

Comment	Toxicity Score
I hate Justin Timberlake.	0.90
I hate Rihanna.	0.69

Prabhakaran, Hutchinson, Mitchell.
“Perturbation Sensitivity Analysis to Detect Unintended Model Biases”

“Sexual Orientation Detector”



- Wang, Kosinki. “Deep neural networks are more accurate than humans at detecting sexual orientation from facial images”.

- “Sexual orientation detector” built using 35,326 images pulled from public profiles on US dating websites.

“Sexual Orientation Detector”

- Algorithm is making decisions based on grooming, presentation, and lifestyle —differences that are cultural and, often, stereotypical.

- Medium article:
  - “Do Algorithms Reveal Sexual Orientation or Just Expose our Stereotypes?”

But what is really bias?
How can we say an algorithm is biased?

Bias (vs Unfairness)

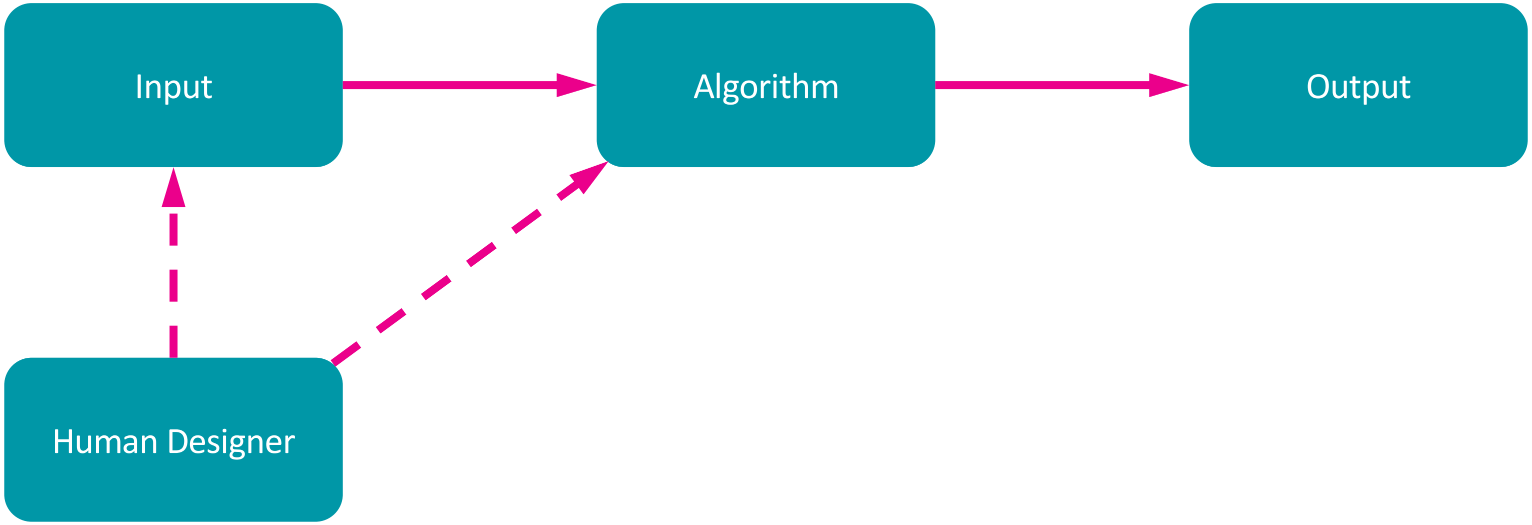
- "Bias is any aspect in which the algorithm, its input, or its output fails to accurately reflect the real world"

  – Nisarg Shah

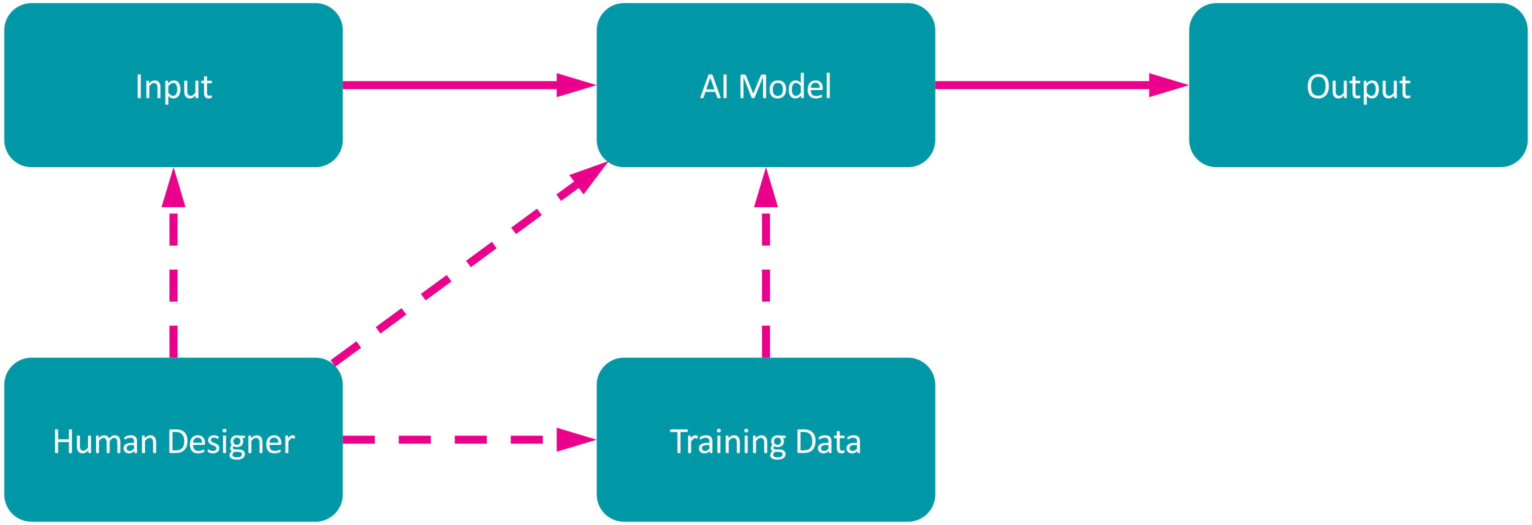
- Measure the discrepancy between the digital world and the real world in some aspect
  - No need to define the agents, have final decisions that impact them, or define the agents' utilities (or costs) for the decisions, which is what we would precisely do to measure fairness
  - Bias is often the root cause of unfairness
  - We will study fairness rigorously in the next lecture

Why do algorithms and AI systems exhibit bias?

How do algorithms work?



How do AI systems work?



What types of biases can exist in data?

REPRESENTATION BIAS

Different subpopulations don't have a balanced representation in data

SELECTION BIAS

Bias in selecting the attributes used to represent individuals

…

MEASUREMENT BIAS

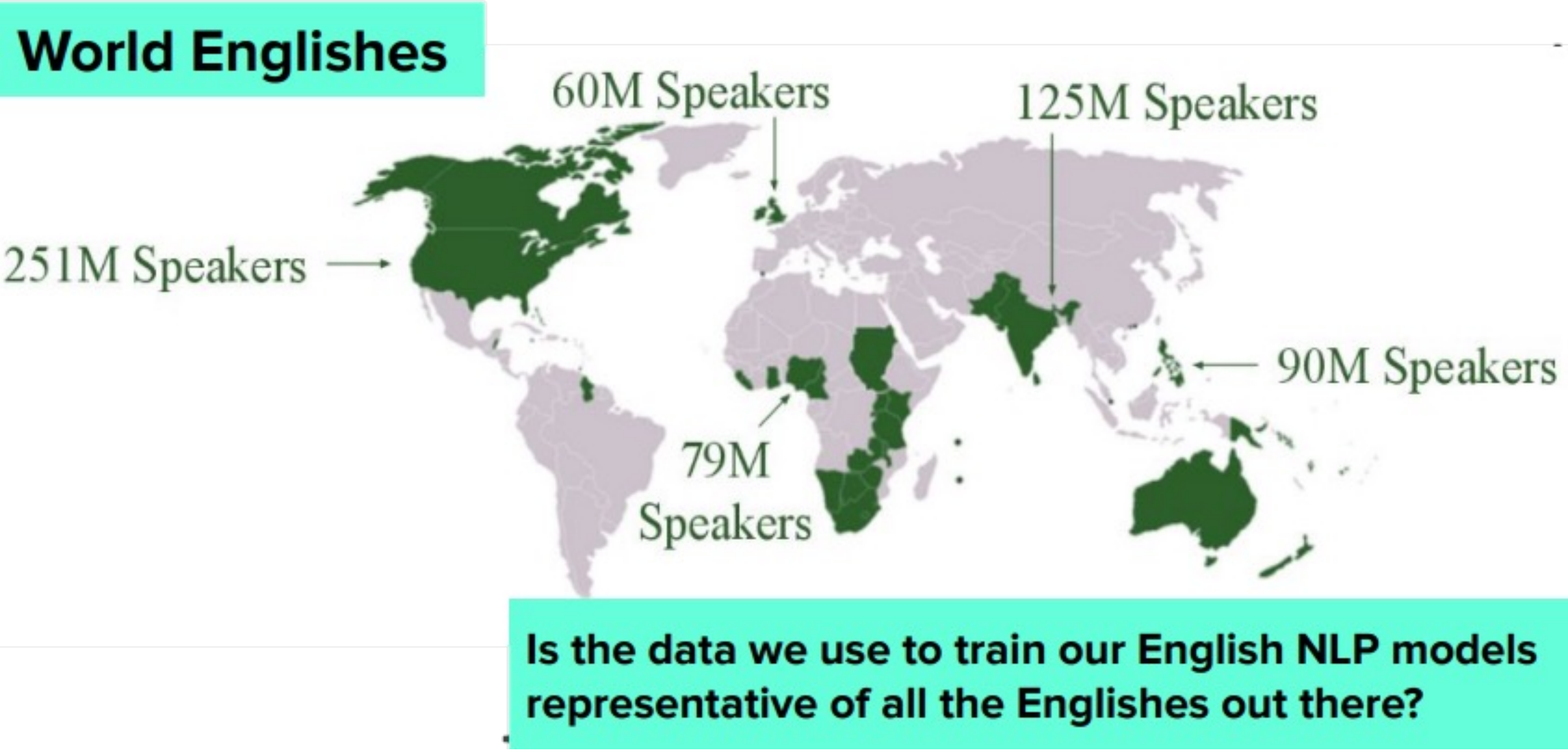
Bias in measuring the values of these selected attributes

USER INTERACTION BIAS

E.g., the presentation of options in a survey can affect the survey results

…

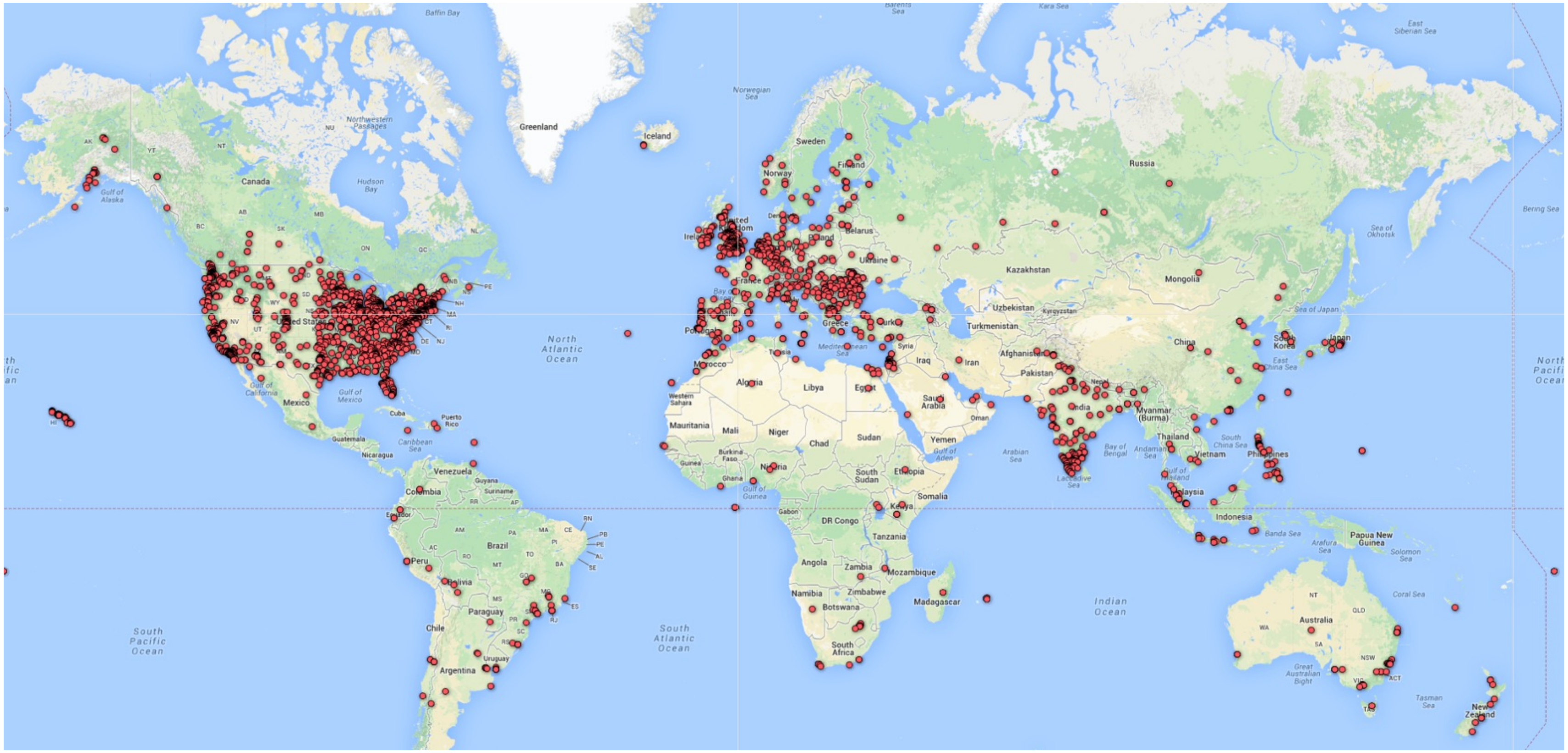
Example: Selection Bias



Is the data we use to train our English NLP models representative of all the Englishes out there?

Chang, Ordonez, Mitchell, Prabhakaran. [“EMNLP’19 Tutorial: Bias and Fairness in Natural Language Processing”]

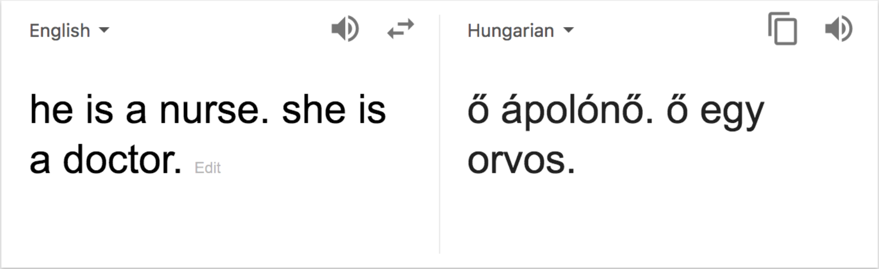
Example: Selection Bias



Tamir.
“A map of 50,000 Mechanical Turk workers:”

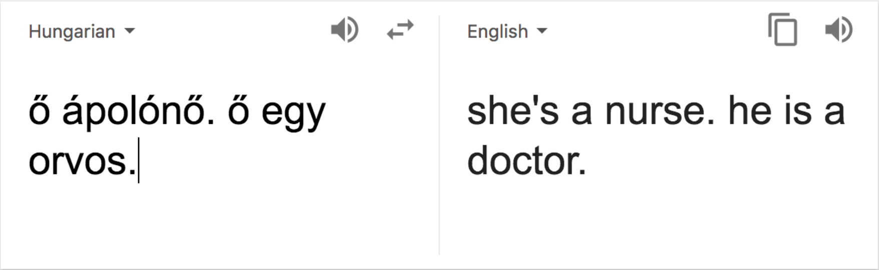
AI can amplify biases in data

- AI tools are often trained to maximize accuracy on training data
  - ➢ They use any patterns, including stereotypical ones, that exist in data and amplify them to maximize accuracy



he is a nurse. she is a doctor.

ő ápolónő. ő egy orvos.



ő ápolónő. ő egy orvos.

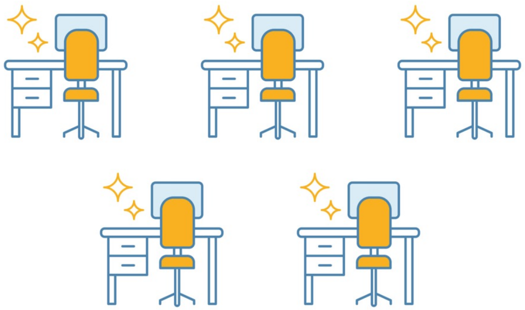
she's a nurse. he is a doctor.

Douglas. “AI is not just learning our biases; it is amplifying them.”

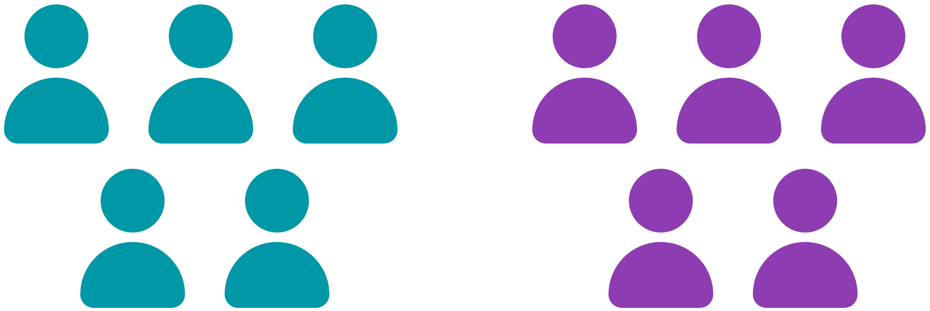
But what if the data was bias-free?

Is the lack of bias enough?

- Ultra-simplified toy example



5 job positions



Group 1

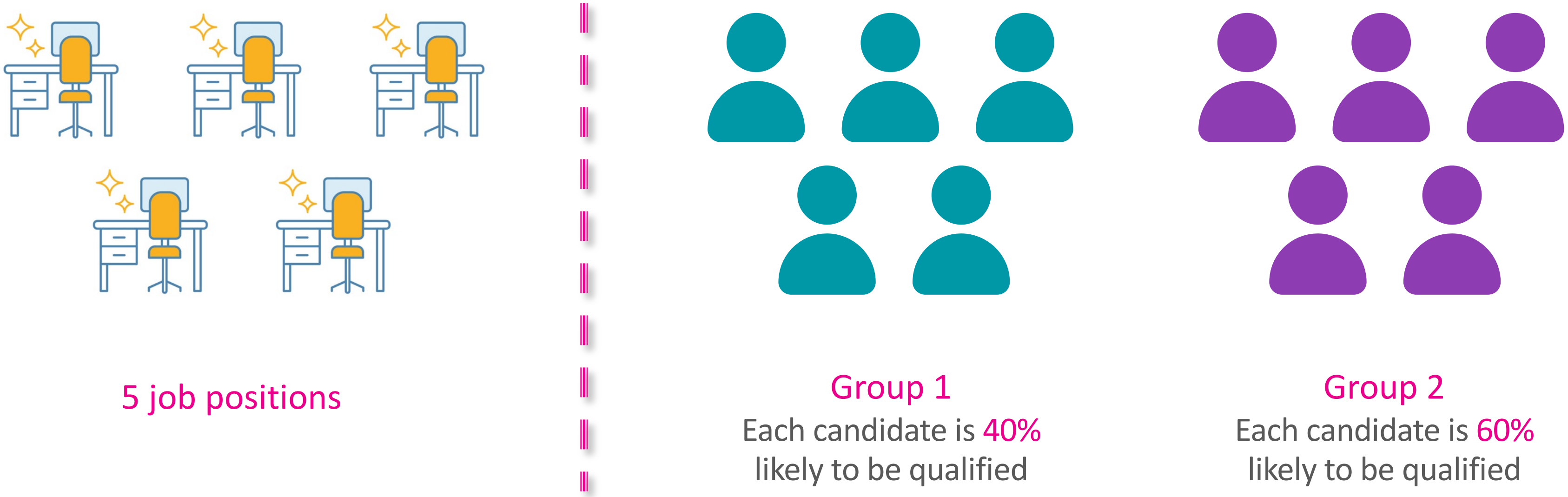
Each candidate is 40% likely to be qualified

Group 2

Each candidate is 60% likely to be qualified

Bias without biased data

- Suppose a classifier selects each candidate from Group 1 with probability p and each candidate from Group 2 with probability q
  - To make the expected number of candidates selected equal to 5, we need $p + q = 1$



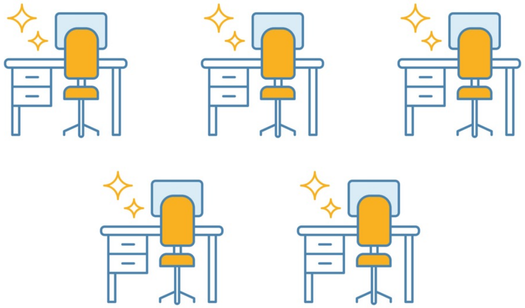
5 job positions

Group 1
Each candidate is 40% likely to be qualified

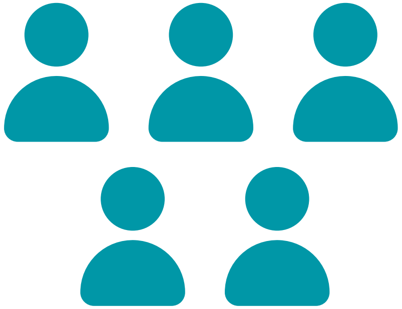
Group 2
Each candidate is 60% likely to be qualified

Bias without biased data

- Empirical Risk Minimization (ERM):
  - Minimize $5 * 0.6 * p + 5 * 0.4 * (1 - p) \Rightarrow p = 0$
  - Selects all five candidates from Group 2

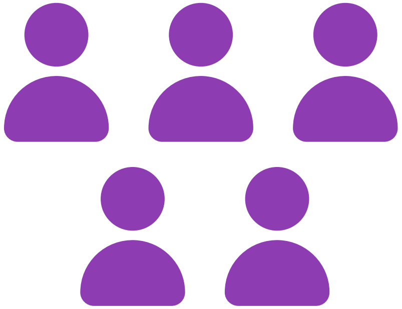


5 job positions



Group 1

Each candidate is 40% likely to be qualified

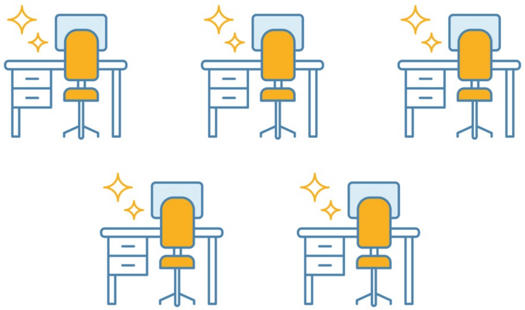


Group 2

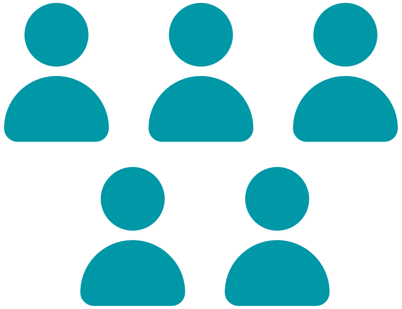
Each candidate is 60% likely to be qualified

Bias without biased data

- Some might consider it proportionally fair to select 2 candidates from Group 1 and 3 candidates from Group 2 (in expectation), i.e., set $p = 0.4$
  - We'll see how to achieve this later

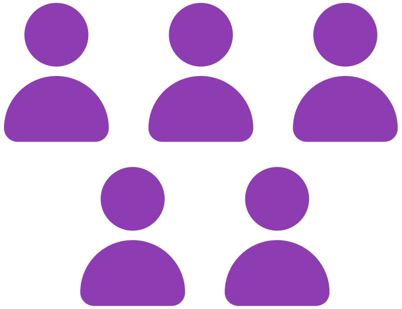


5 job positions



Group 1

Each candidate is 40% likely to be qualified



Group 2

Each candidate is 60% likely to be qualified

Quiz

- Here is a similar example from “wagering mechanisms”
- Say I want to elicit your subjective belief on the chances that it will rain tomorrow.
    - I ask you: “What do you think are the (percentage) chances that it will rain tomorrow?”
    - If you say p, I will pay you p dollars if it really ends up raining tomorrow and $100 - p$ dollars if it doesn’t.
        - For example, if you say, “I think there is a 75 percent chance that it will rain tomorrow”, I will pay you $75 if it in fact rains tomorrow and $25 if it doesn’t.
- Question: If you really believed that there is a 60% chance that it will rain tomorrow, what percentage should you report?
- Food for thought: If I want to incentivize you to report your true belief, how should I pay you?

How do we measure bias?

Measuring Bias

- No single trick, very application dependent
- Let’s look at an example from NLP

Word Embeddings

- Words are not just unrelated discrete concepts
  - ➢ Searching “Toronto hotel” vs “Toronto motel” should yield similar results because “hotel” and “motel” are similar words

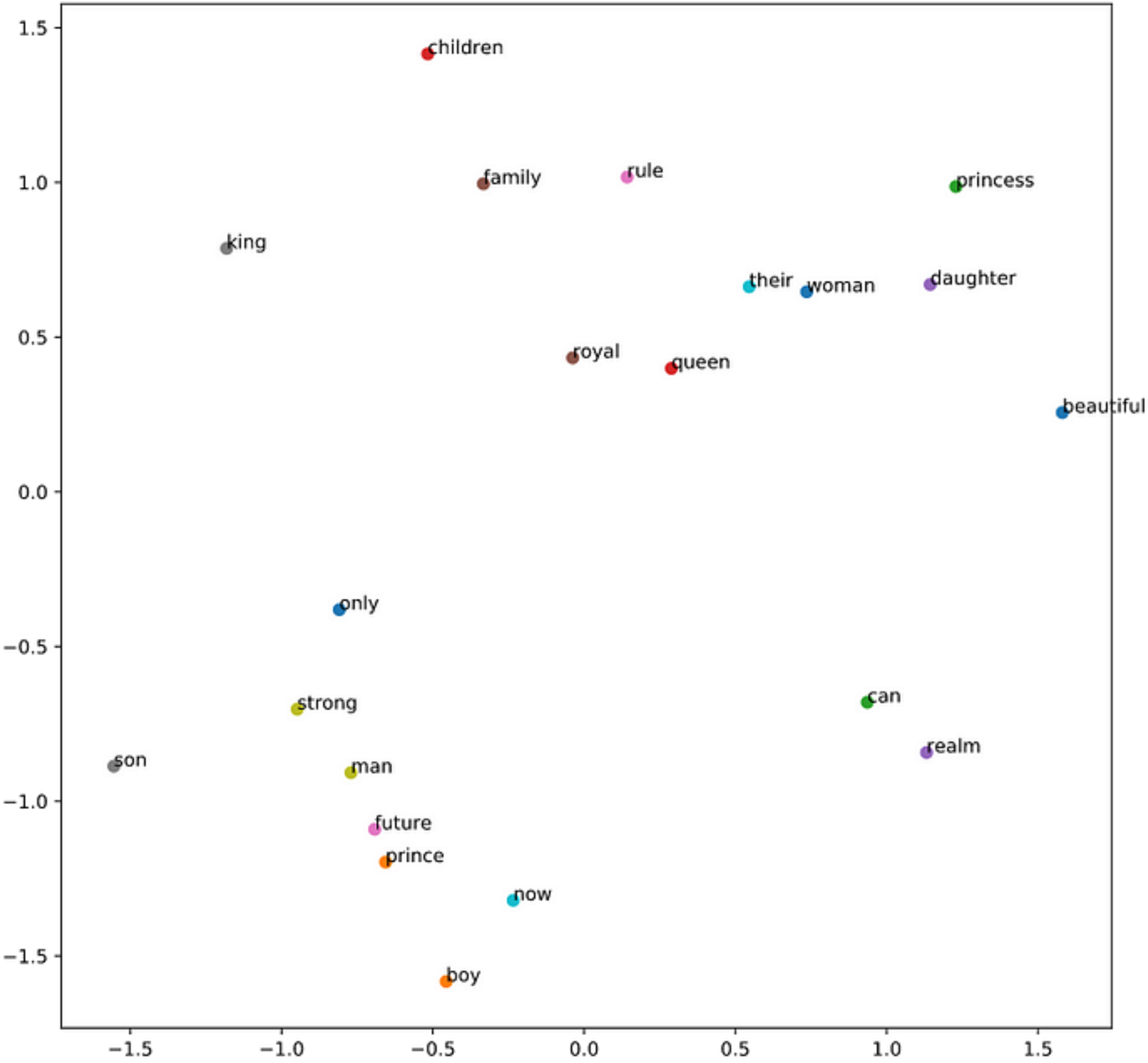
- Idea: words with similar meaning occur in similar contexts
  - ➢ “hotel” and “motel” are used similarly in sentences

- Word embeddings: represent words according to which other words they co-occur with
  - ➢ “book”, “room”, “rate” are commonly used with “hotel” and “motel” too

Word Embeddings

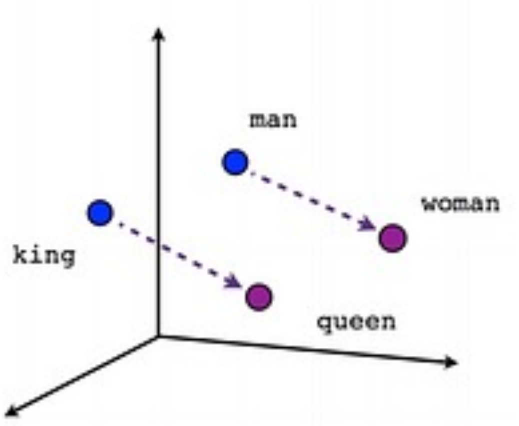
The future king is the prince
Daughter is the princess
Son is the prince
Only a man can be a king
Only a woman can be a queen
The princess will be a queen
Queen and king rule the realm
The prince is a strong man
The princess is a beautiful woman
The royal family is the king and queen and their children
Prince is only a boy now
A boy will be a man

[Eligijus Bujokas, towardsdatascience.com]

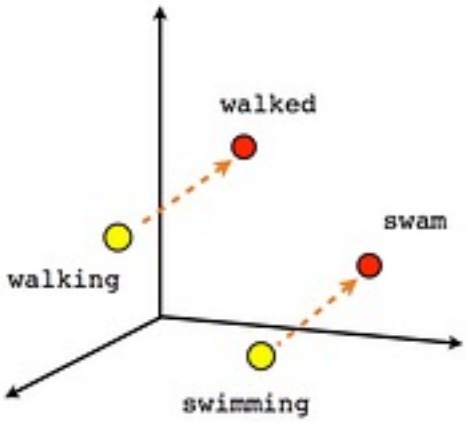


[Eligijus Bujokas, towardsdatascience.com]

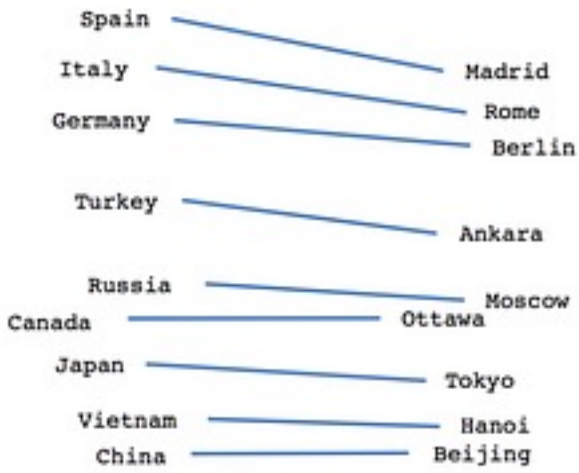
Word Embeddings



Male-Female



Verb tense



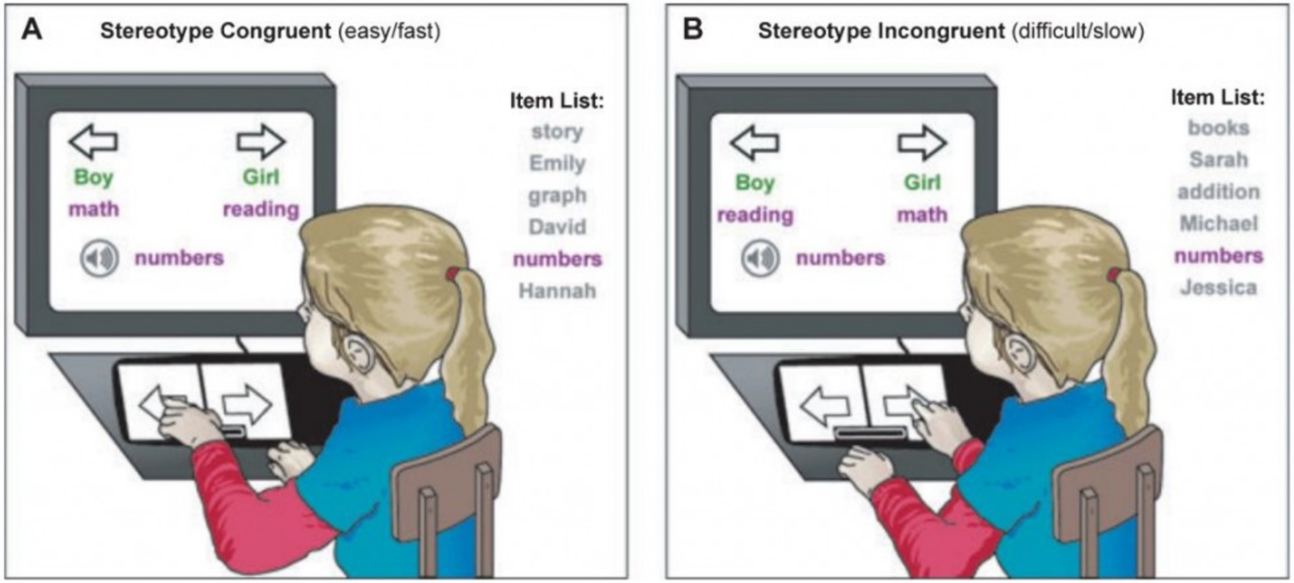
Country-Capital

[Eligijus Bujokas, towardsdatascience.com]

Measuring Bias: WEAT

- Caliskan et al., “Semantics derived automatically from language corpora contain human-like biases”

- Modeled after Implicit Association Test



Analogies

- 🙂 King : Man :: Queen : Woman
- 🙂 Paris : France :: London : England
- 😠 Man : Computer_Programmer :: Woman : Homemaker

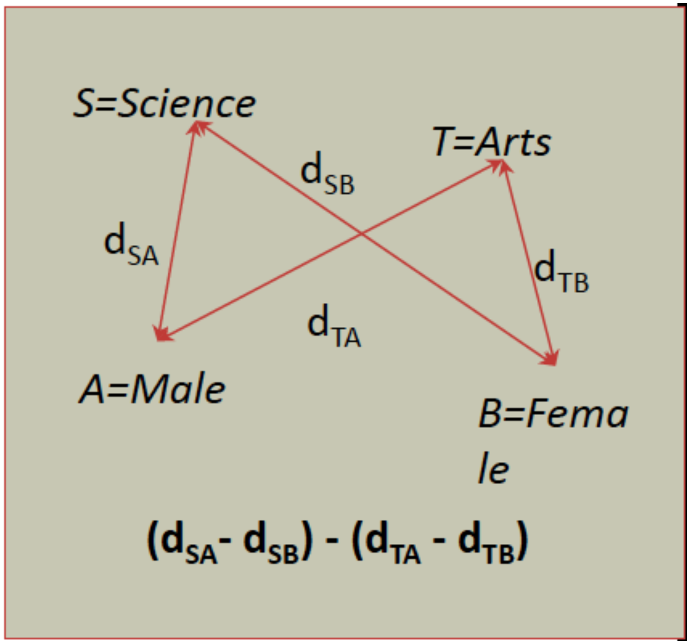
(Bolukbasi et al., NeurIPS 2016)

WEAT

- Word Embedding Association Test

- **Target Word Sets:**
  - S = {physics, chemistry… } ≈ Science
  - T = {poetry, literature… } ≈ Arts

- **Attribute Word Sets:**
  - A = {he, him, man… } ≈ Male
  - B = {she, her, woman} ≈ Female



$$f(w, A, B) = \underset{a \in A}{\text{mean}} cos(\vec{w}, \vec{a}) - \underset{b \in B}{\text{mean}} cos(\vec{w}, \vec{b})$$

Effect Size = $\dfrac{\underset{s \in S}{\text{mean}} f(s, A, B) - \underset{t \in T}{\text{mean}} f(t, A, B)}{\underset{w \in S \cup T}{\text{std-dev}} f(w, A, B)}$

(Pitassi, Zemel, CSC2541)

How do we mitigate bias?

Mitigating Bias

1. Removing bias at the individual word embedding level (by making theoretically neutral words orthogonal to the *sentiment* vector
   - Sentiment vector is e.g. good – bad, positive – negative etc.

2. Removing bias at the level of the classifier, which is trained upon the word embeddings
   - E.g., by adjusting the linear hyperplane learned by the linear SVM classifier to make it orthogonal to the sentiment vector

- Lack theory for connecting the definitions of bias or the impact of debiasing methods on the eventual fairness