Bias and Fairness in NLP

Ricardo Muñoz Sánchez

The Role of AI in the World

- AI has evolved at neckbreaking speeds this past decade
- These advances mean that it is being deployed in high-stakes situations
- The idea is that machines are better, faster, and/or more accurate than humans

How Deep Learning Works

- Gather insane amounts of data
- Feed it to the computer
- Hope for the best

How Deep Learning Works

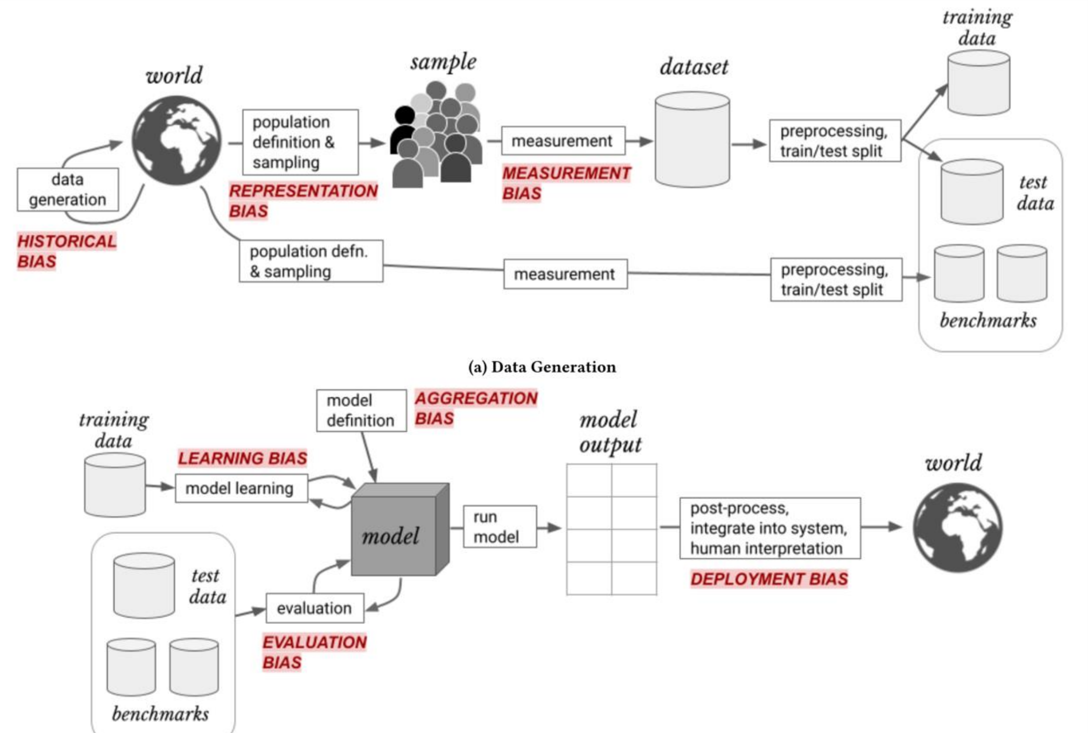
- Gather insane amounts of data
- Feed it to the computer
- Hope ~~for the best~~ it finds and exploits patterns in the data

Data Comes from Somewhere

- Humans
  - Generate data
  - Aggregate data
  - Annotate data

- But humans are biased/prejudiced!

- These biases find their way into our data and our models



(a) Data Generation

(b) Model Building and Implementation

From “A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle” by Suresh et al. (2021)

What Has Been Usually Done

- Languages
  - Mostly with English
  - Some work with gendered languages
    - Think German or the romance languages
    - Semantic vs grammatical gender

- Biases studied
  - Gender as a binary (male/female)
  - Race in the US as a binary (black/white)

Opportunities for Expansion

- Gender as a more complex phenomenon
- LGBTQ+ identities and communities
- In-group vs. out-group interactions
- Nationality, ethnicity, and/or race
- Biases based on names

GÖTEBORGS
UNIVERSITET

SPRÅKBANKENTEXT

Ricardo Muñoz Sánchez
ricardo.munoz.sanchez@svenska.gu.se
rimusa.github.io