Machine Learning without Biases

Learning to create "fair" models

Yordan Darakchiev
iordan93@gmail.com

Cognitive heuristics

- "Mental shortcuts" that help us solve problems
  - ✓Reduce the number of variables (feature selection)
  - ✓Reduce the "sample space" (data filtering)
  - ✓Reduce the time to reach a solution (lazy evaluation)
  - ✓Rely on past experience (greedy approach)

- Can lead to errors (**cognitive biases**)
  - ✘Confirmation bias
  - ✘Status-quo bias
  - ✘Authority bias, bandwagon bias
  - ✘Loss-aversion bias

What is bias in machine learning?

- Machine learning algorithms make decisions every day
  - Assessing employee satisfaction
  - Predicting credit defaults
  - Spotting criminals
  - Treating deadly diseases
- Our algorithms need tons of data to learn
- **When faced with radically different data, their behavior is undefined**
- We want AI to make better decisions than us
  - But it ends up amplifying our own unconscious biases

What can go wrong?

- Google image recognition, 2015
  - Recognizes pictures of African Americans as gorillas
- Microsoft Tay, 2016
  - Learns from Twitter posts
- IBM Watson, 2017
  - Memorizes the entire Urban Dictionary
- US court risk assessment, recidivism assessment, 2016-2017
  - Predicts non-Caucasians will re-offend up to 2x more
- iPhone facial recognition, 2017-2018
  - Can't recognize dark-skinned people
- Amazon recruitment, 2018
  - Strongly prefers male resumes

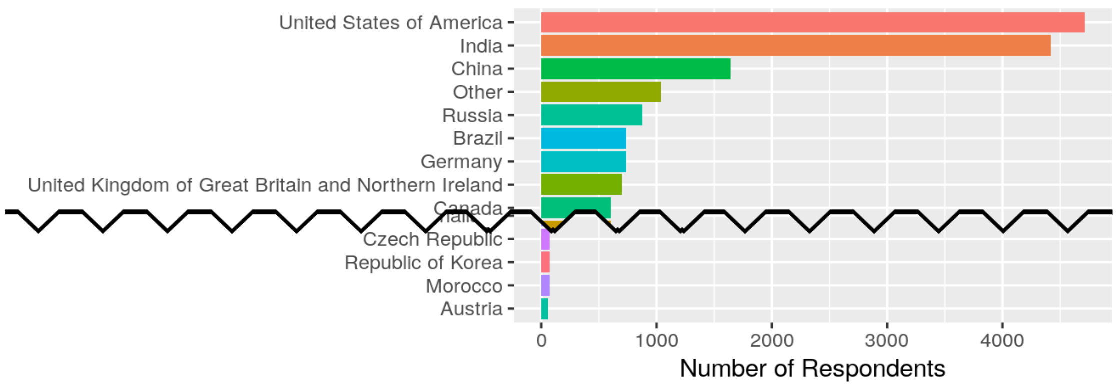
What can we do?

- Two main points
  - Mathematical algorithms to reduce bias
  - A lot of manual work
- **Gathering diverse samples is key!**
- Most machine learning algorithms act as "black boxes"
  - We don't really see how an algorithm might be biased unless we get data to prove it
- Is real-world testing safe?
  - Can we afford having two, three, or ten iterations of our algorithm run before de-biasing them?
  - Is de-biasing adding additional layers of bias?
  - What other prejudices remain?

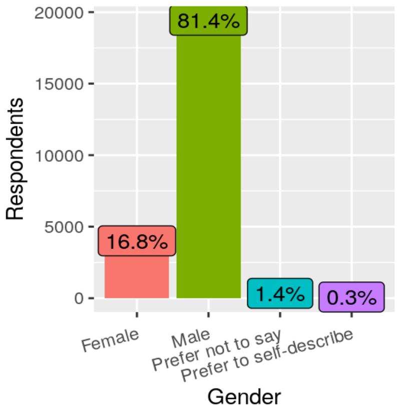
But who are we?

- Kaggle, 2018 demographics survey results (source)

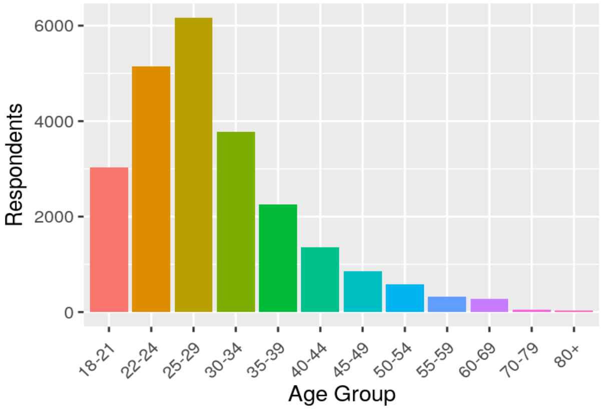
Country of Residence: US & India dominate



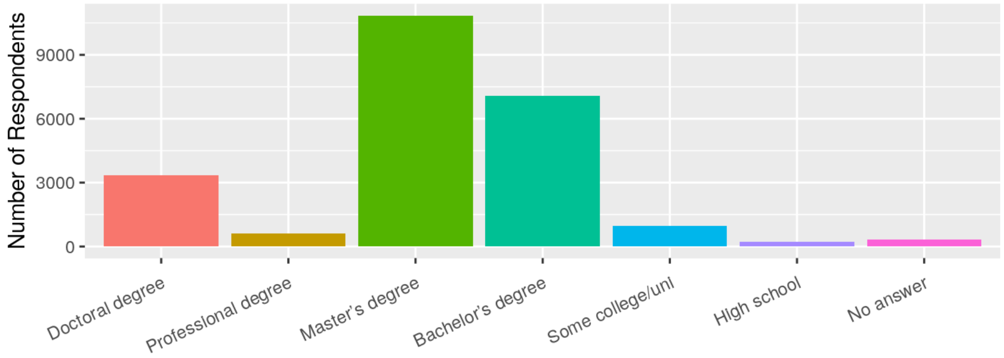
Gender imbalance



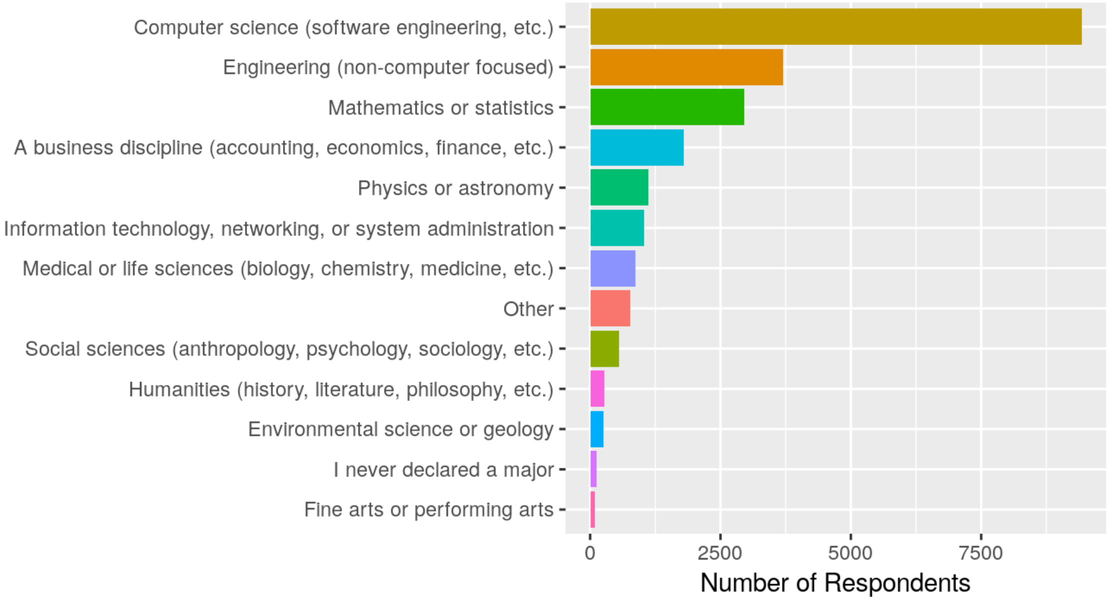
Age Groups: Here be 20-somethings



Highest Education Level: Mostly Master's & Bachelor's



Undergrad Major



How can the business help?

- We're good at maths
  - and we know how to work with data
- You know your priorities, criteria, targets, and KPIs
  - Gather **diverse** data (e.g. user backgrounds)
  - Gather **diverse** user feedback
  - Work with us to create fair, unbiased KPIs
  - Help us identify potential bias
    - E.g. gender, race, marital status, location
  - Help us create a common language
- Understand that everyone is biased...
  - ... but that doesn't mean our algorithms should be :)

Thanks!