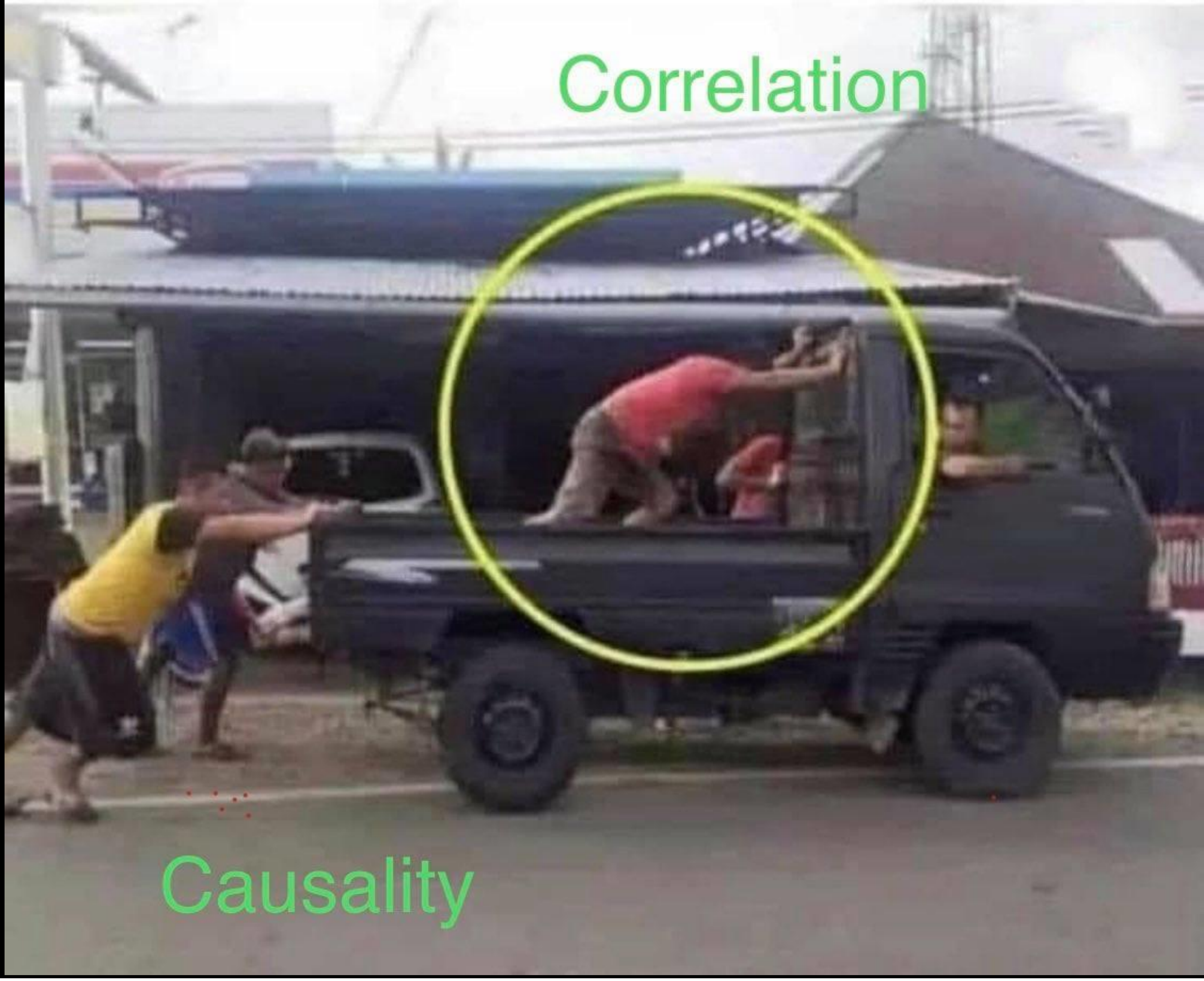
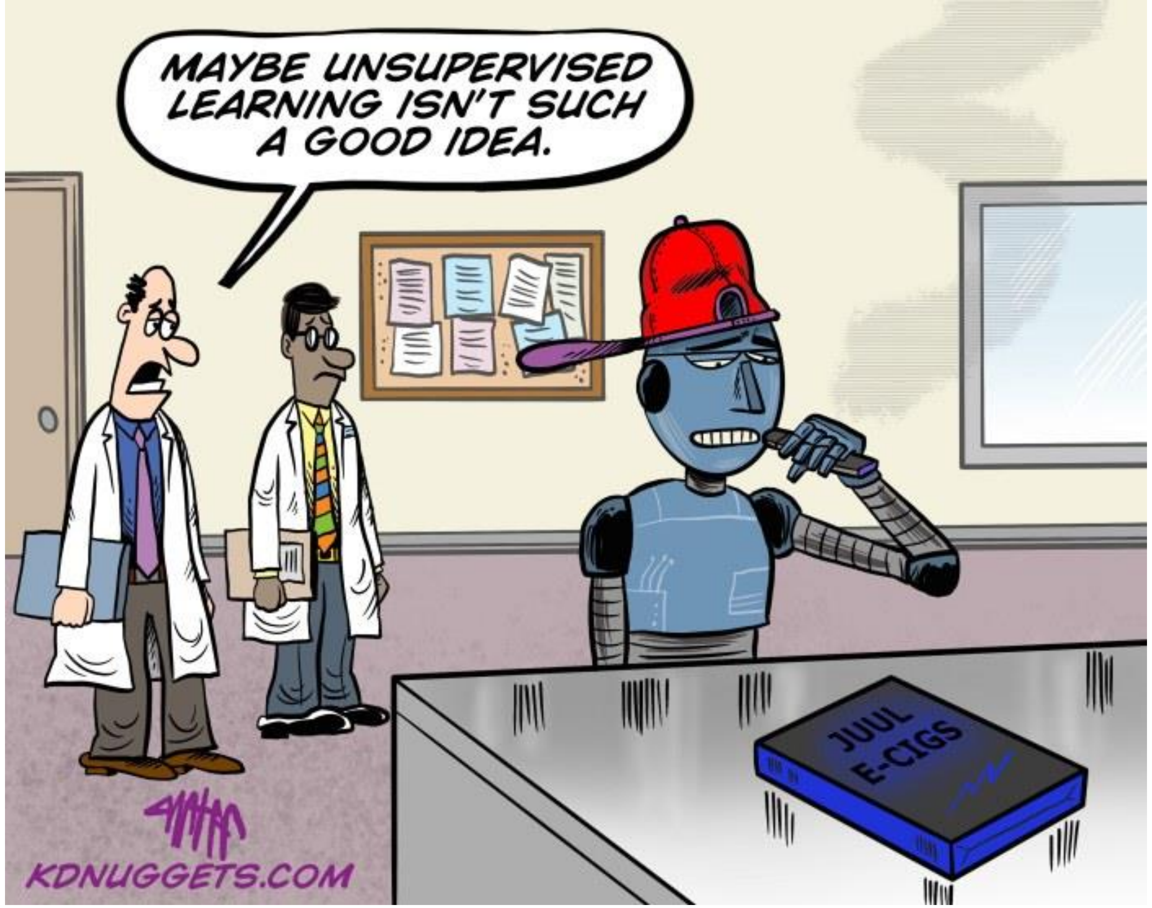
INFO 251: Applied Machine Learning



ML Harms

Today's outline

- Applied ML, start to finish
- 5 mins for course evals
- **ML harms and ethics**



Harms throughout the ML Life Cycle

- At this point in the semester, you have a strong foundation to understand how to apply ML to real-world problems
  - But this doesn't necessarily mean you *should* by using ML to address those real-world problems
- Several tensions were visible in the Togo case study
  - Exclusion and Bias
  - Data privacy and access
  - Technocracy
  - Control and authority

ML in Society

- More broadly, ML is now routinely used to make incredibly consequential decisions in all sectors of our society
  - Medical decisions
  - Parole / bail / policing / military / security decisions
  - Hiring / firing / recruiting / admissions
  - Content and product recommendations
  - Facial recognition / scene recognition / autonomous vehicles
  - ...
- The stakes are incredibly high for ML mistakes!

For a more systematic discussion:

A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle

Harini Suresh
John Guttag
hsuresh@mit.edu
guttag@mit.edu

- EAAMO '21: Equity and Access in Algorithms, Mechanisms, and Optimization

- *"To anticipate, prevent, and mitigate undesirable downstream consequences, it is critical that we understand when and how harm might be introduced throughout the ML life cycle"*

Kudos: Lauren Chambers

"Seven sources of harm in ML"

1. **Historical Bias:** Can arise even if data are perfectly measured and sampled -- if the world as it is (or was) leads to a model that produces harmful outcomes
   - Example: word embeddings reflect gender biases, for instances that "nurse" is associated with women and "engineer" with men

2. **Representation bias**: "when the development sample underrepresents some part of the population"
   - Not just about representativity: model may not have enough data to model under-represented groups, even if proportionally represented
   - Example: Lack of minority images in ImageNet (45% of images from the US!)

"Seven sources of harm in ML"

3. **Measurement bias:** when features or labels don't accurately represent the phenomenon being modeled
   - Examples: "poverty", credit scores, risk assessments

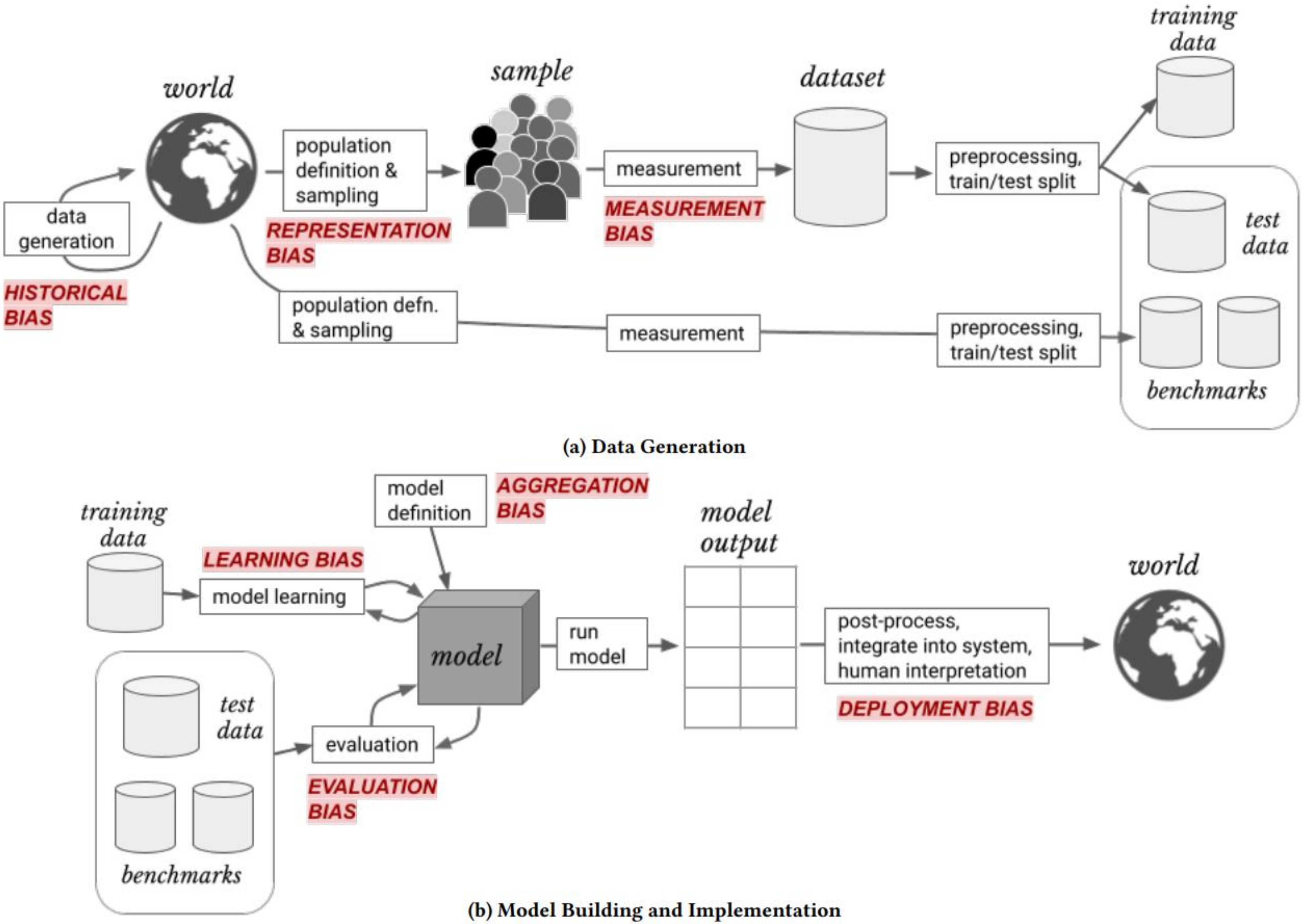
4. **Aggregation bias**: "when a one-size-fits-all model is used for data in which there are underlying groups or types of examples that should be considered differently"
   - Examples: Gender-sensitive credit scoring, quoting of rappers in in social media

“Seven sources of harm in ML”

5. **Learning bias:** “when modeling choices amplify performance disparities across different samples in the data”
   - Examples: Issues can arise when prioritizing one objective (e.g., overall accuracy) damages another (e.g., disparate impact) – see Kleinberg et al. 2017; selecting for “compact” models (e.g., pruning) can amplify performance disparities on data with underrepresented attributes
6. **Evaluation bias**: “when the benchmark data used for a particular task does not represent the use population”
   - Examples: (pre-)training on ImageNet for a different downstream task; the desire in ML circles to use standardized benchmarks and performance metrics

"Seven sources of harm in ML"

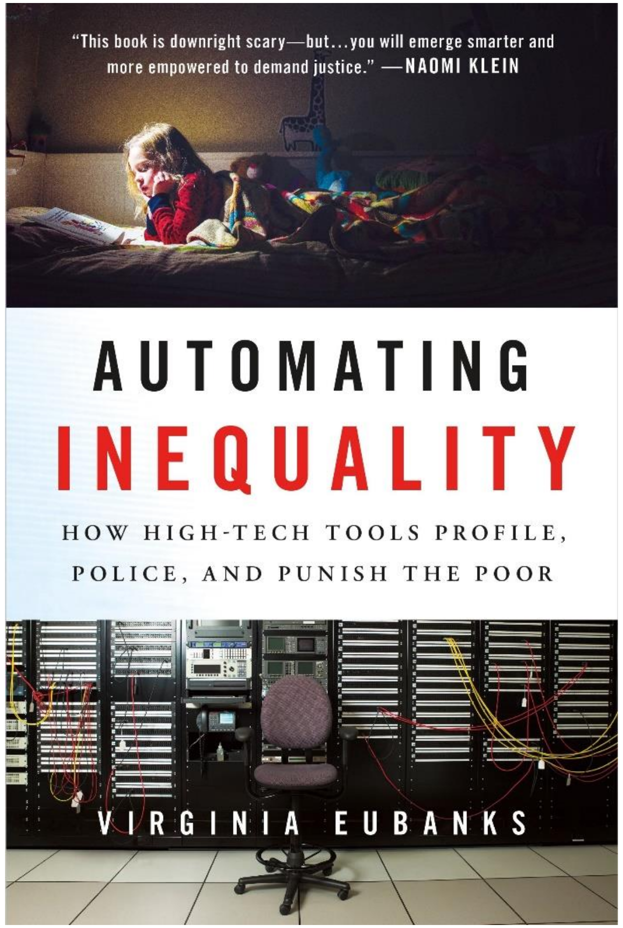
7. **Deployment bias:** when there is a mismatch between the problem a model is intended to solve and the way in which it is actually used.
   - "This often occurs when a system is built and evaluated as if it were fully autonomous, while in reality, it operates in a complicated sociotechnical system moderated by institutional structures and human decision-makers
   - Examples: When ML output is then interpreted by a human. "Despite good performance in isolation, they may end up causing harmful consequences because of phenomena such as automation or confirmation bias."

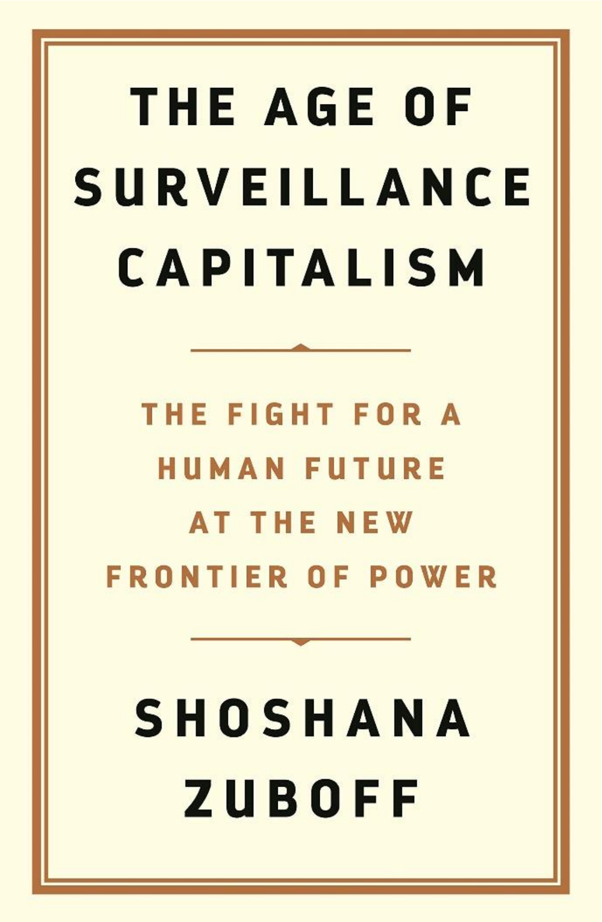


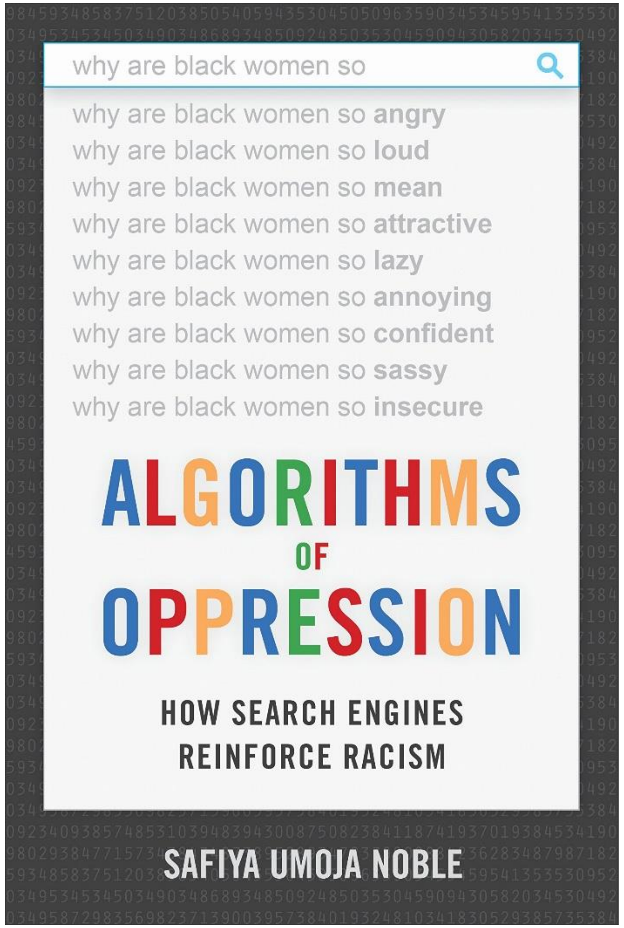
(a) Data Generation

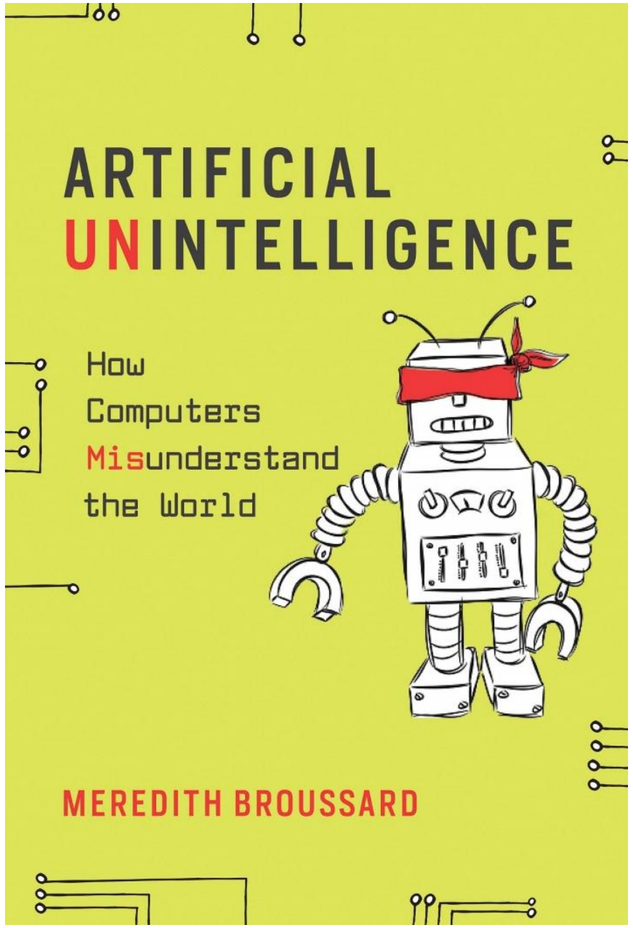
(b) Model Building and Implementation

This is just the tip of the iceberg!









Additional resources at Berkeley

- INFO 188/288: Behind the Data: Humans and Values
- CS 294-186: Algorithms & Inequality
- AFOG: Algorithmic Fairness and Opacity Working Group