2. Sources of Bias

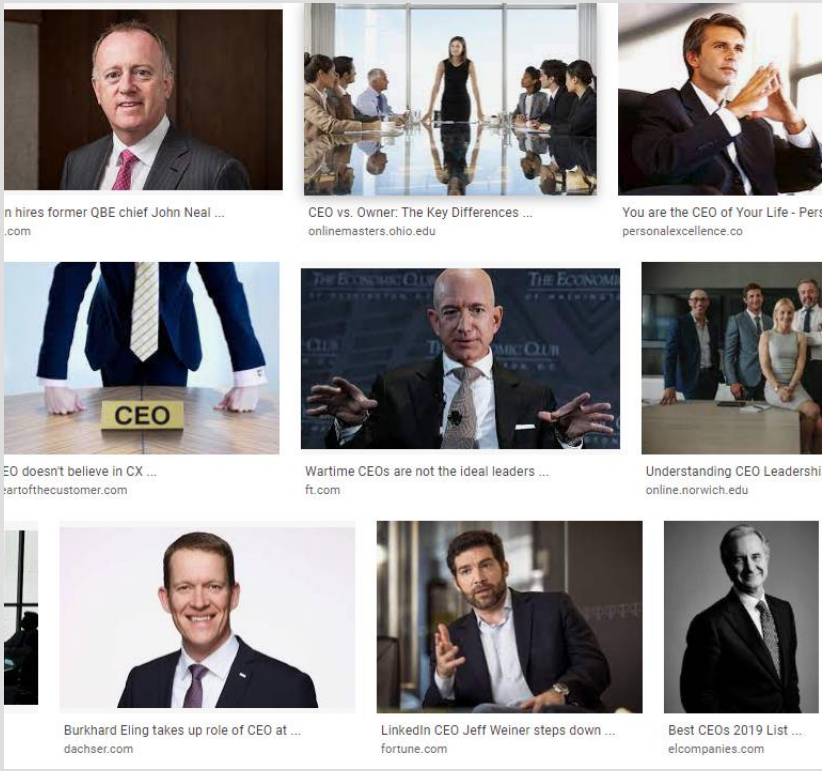
What is a bias ?

Inclination or prejudice of a decision made by an AI system which is for or against one person or group, especially in a way considered to be unfair.

From *Bias in Data-driven AI Systems - An Introductory Survey* (Ntoutsi et al., 2020)

Historical bias

- Pre-existing bias reflected in the data
- Example: What should be the results of an image search for ‘CEO’? In 2018, 5% of Fortune 500 CEOs were women.
- Google recently changed image search result to include higher proportion of women



Representation bias

- Occurs when certain parts of the input space are underrepresented
  - sampling methods only reach a portion of the population
  - population of interest ≠ training data

- Example: ImageNet. 45% from US. 1% from China.

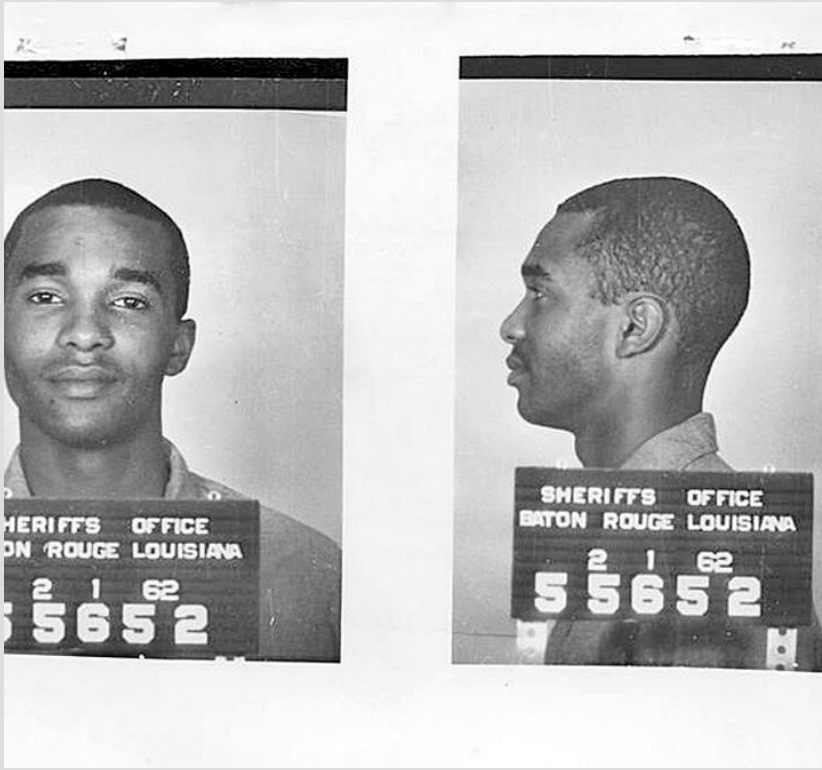
- Classifier for ‘bride’ performs worse in under-represented countries

Measurement Bias

- Occurs when selecting or collecting features.

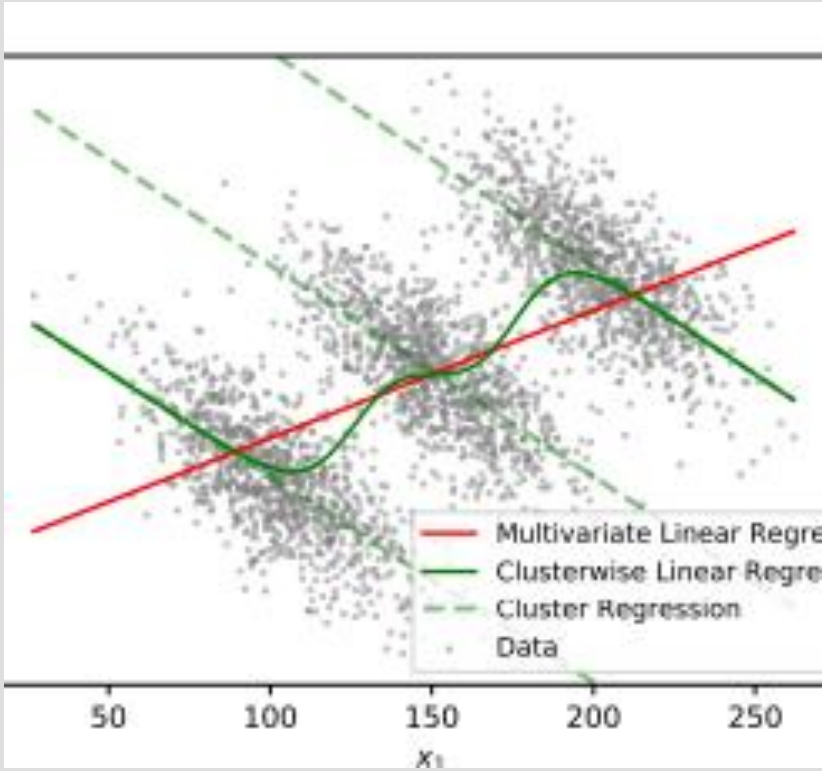
- Measurement process/ data quality may vary across groups

- Example:
  - Arrest rates used as a measure for crime rate.
  - COMPAS: prior arrests + friend/family arrests to measure recidivism



Aggregation Bias

- One-size fit all model
- Difference across groups might require several models
- Example: HbA1c levels (used to monitor diabetes) differ across gender and ethnicity



Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2019). A survey on bias and fairness in machine learning. *arXiv preprint arXiv:1908.09635*.

Evaluation bias

- When testing on benchmarks that are unbalanced compared to target population

- Example: Adience and IJB-A in Gender Shades

The Algorithmic Justice League by Joy Buolamwini

Deployment Bias

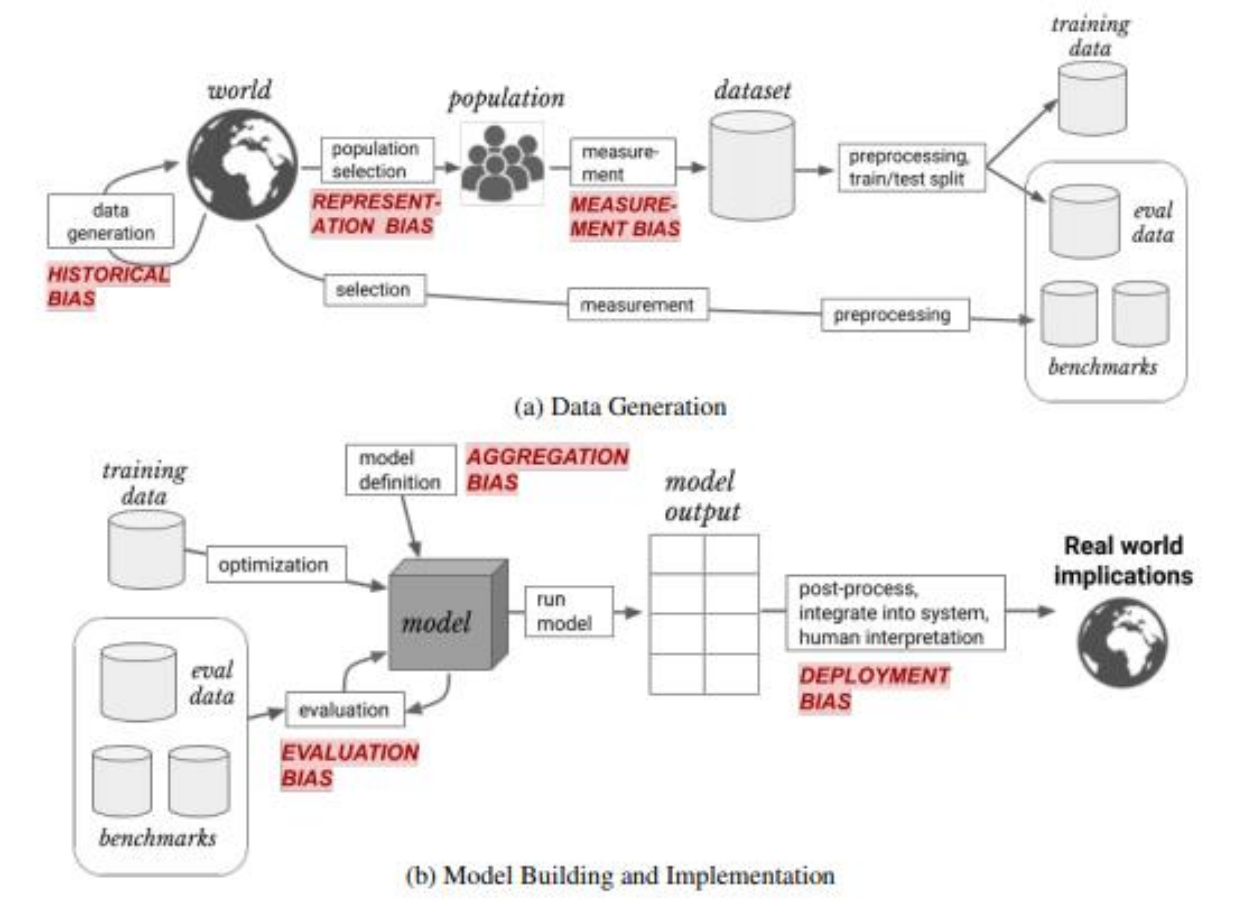
- Mismatch between design and use

- Example: person's likelihood of committing a future crime → high-stake use of determining the length of a sentence

An overview

Taken from :

Suresh, H., & Guttag, J. V. (2019). A framework for understanding unintended consequences of machine learning. *arXiv preprint arXiv:1901.10002*.



Reading/References

(Mehrabi et al., 2021). A survey on bias and fairness in machine learning

(Suresh & Guttag, 2019). A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle