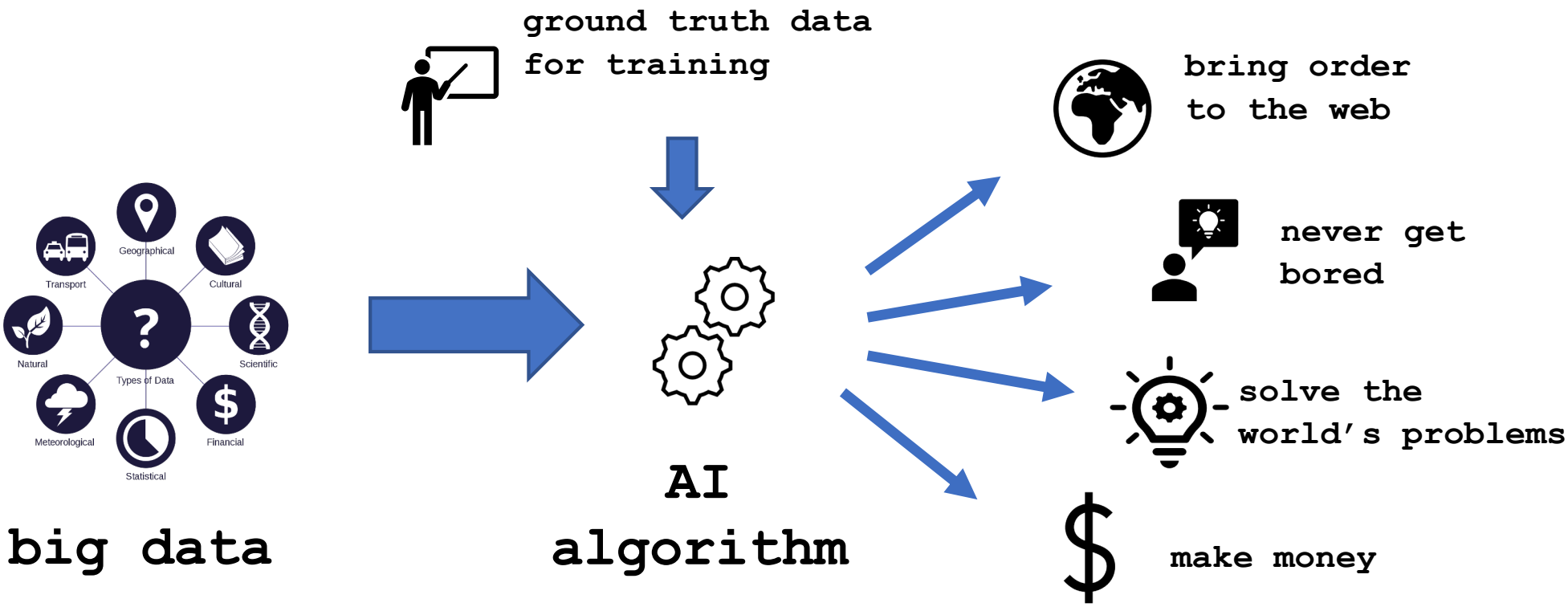
Ethics in AI

Dr. Fred Morstatter

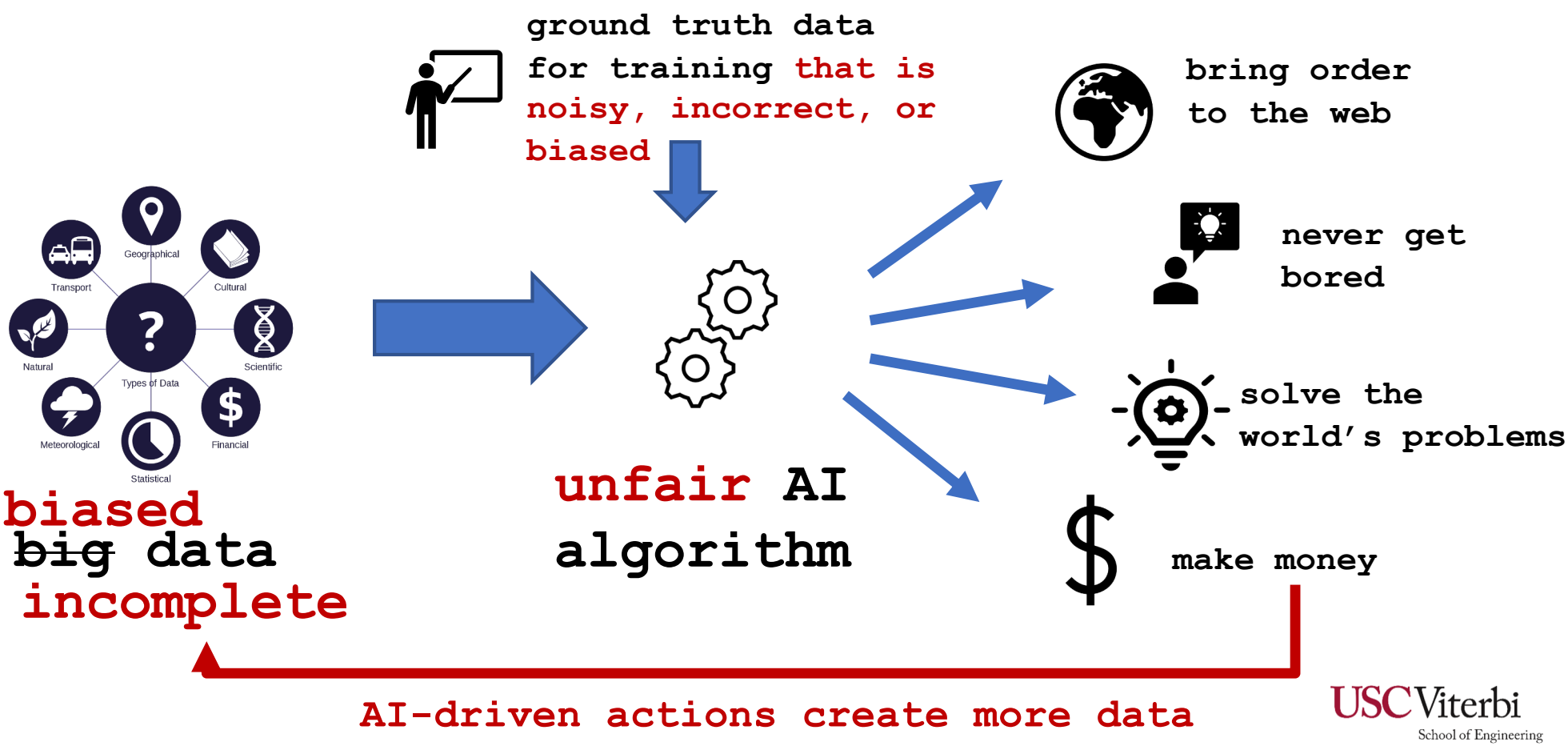
USC Viterbi
School of Engineering

AI systems



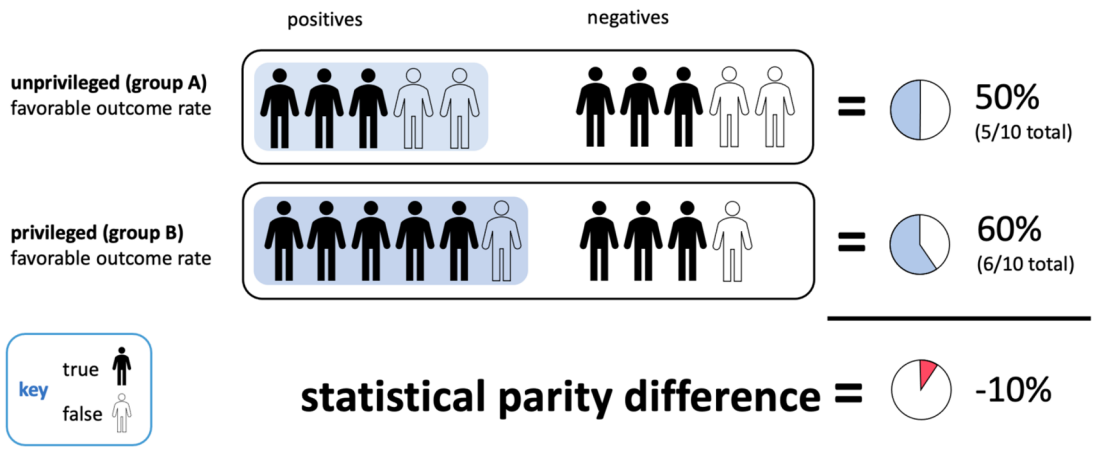
USC Viterbi
School of Engineering

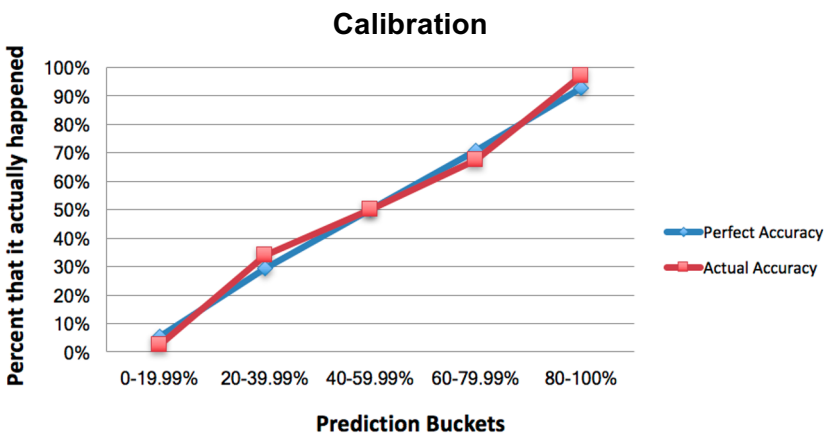
AI systems



USC Viterbi School of Engineering

Defining Fairness

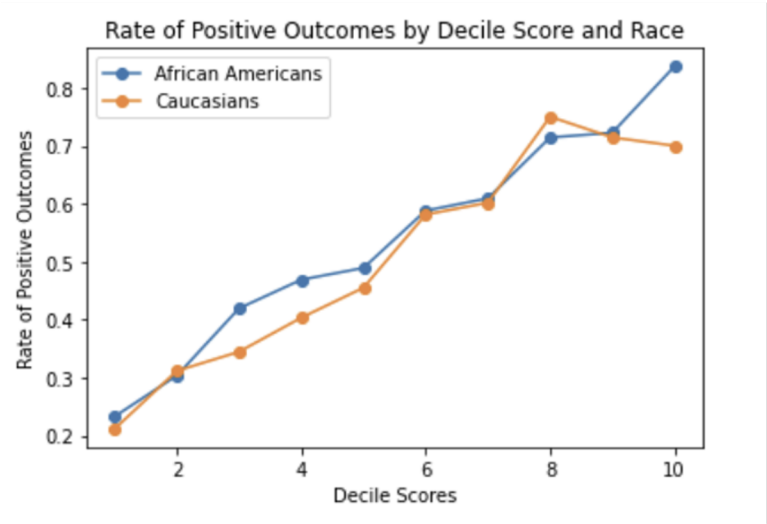


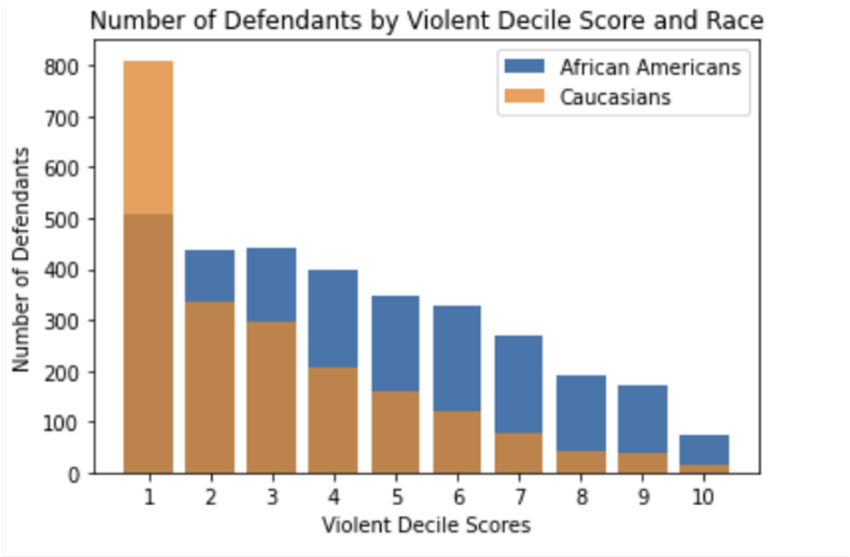


Group Level Graphic Credit: https://www.xyonix.com/blog/how-to-detect-and-mitigate-harmful-societal-bias-in-your-organizations-ai
Calibration: https://electionbettingodds.com/TrackRecord.html

USC Viterbi School of Engineering

Calibration and Eq. Odds: COMPAS

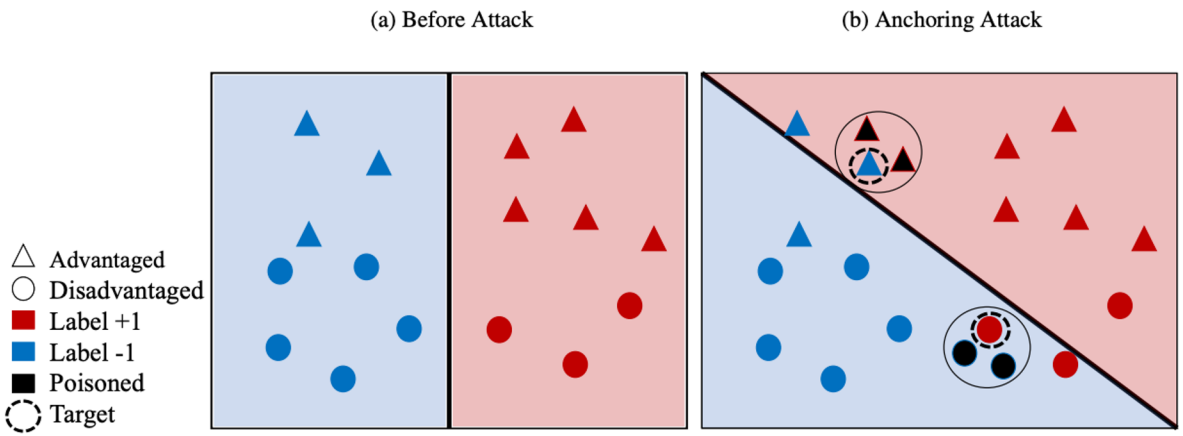


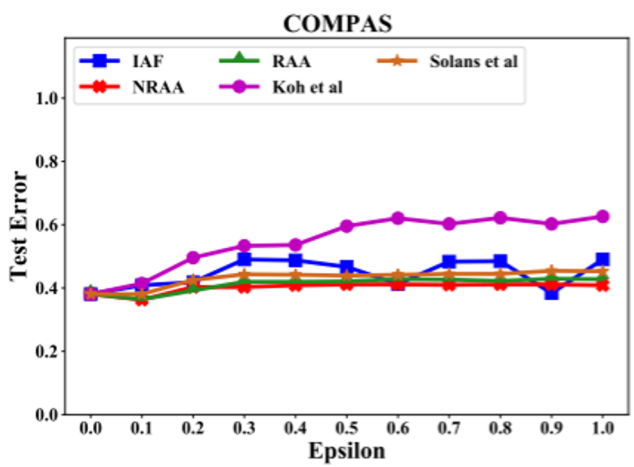


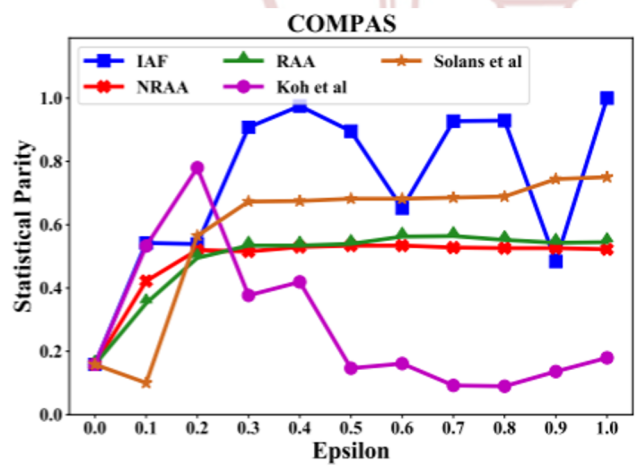
https://towardsdatascience.com/fairness-in-decision-making-for-criminal-justice-2eaaaf947142

USC Viterbi School of Engineering

Attacks Targeting Fairness







Mehrabi, N., Naveed, M., Morstatter, F., & Galstyan, A. (2021, May). Exacerbating Algorithmic Bias through Fairness Attacks. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 35, No. 10, pp. 8930-8938).

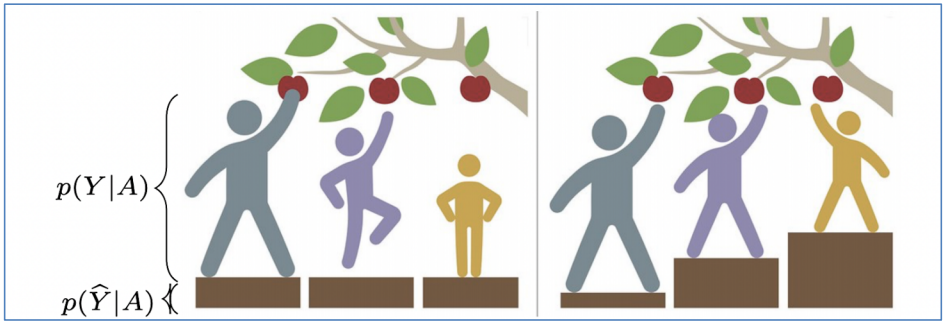
Fairness

Allocation Harms

- "A system withholds certain resources or opportunities from some groups."*
- E.g. Loan allocation systems.

Representational Harms

- Harm to the identity of certain groups.
- Regardless of allocating resources or opportunities.

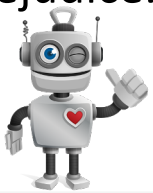


Mehrabi, N., Huang, Y., & Morstatter, F. (2020). Statistical Equity: A Fairness Classification Objective. arXiv preprint arXiv:2005.07293.

USC Viterbi School of Engineering

Representational Harms

- Investigating representational harms in commonsense knowledge resources and tasks.
- What is Commonsense Knowledge?
  - "Facts about the world that all humans are expected to know."**
  - Commonsense knowledge should be facts NOT stereotypes or biases in terms of favoritism or prejudice.



en Fish — CapableOf ⟶ en swim
Weight: 2.0

en A politician — CapableOf ⟶ en tell lies
Weight: 1.0

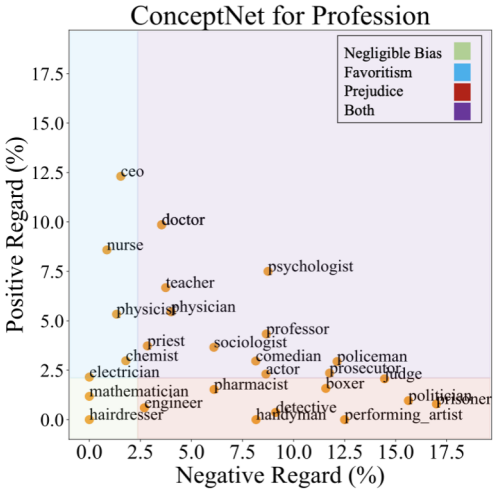
***This presentation contains content that may be offensive or upsetting.**

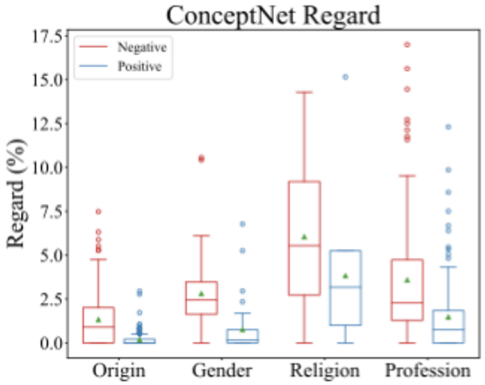
****https://en.wikipedia.org/wiki/Commonsense_knowledge_(artificial_intelligence)**

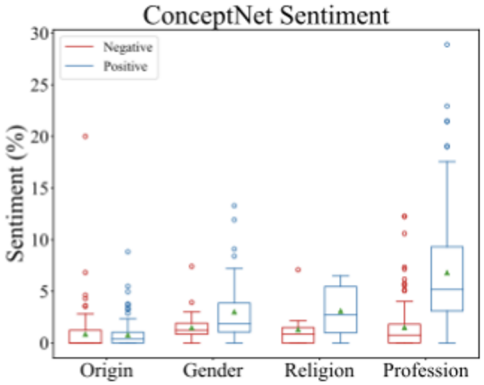
USC Viterbi School of Engineering

Finding Common Sense Bias

- Observed both overgeneralization and disparity in overgeneralization in both Knowledge resources.
- The degree of severity was higher in GenericsKB compared to ConceptNet.







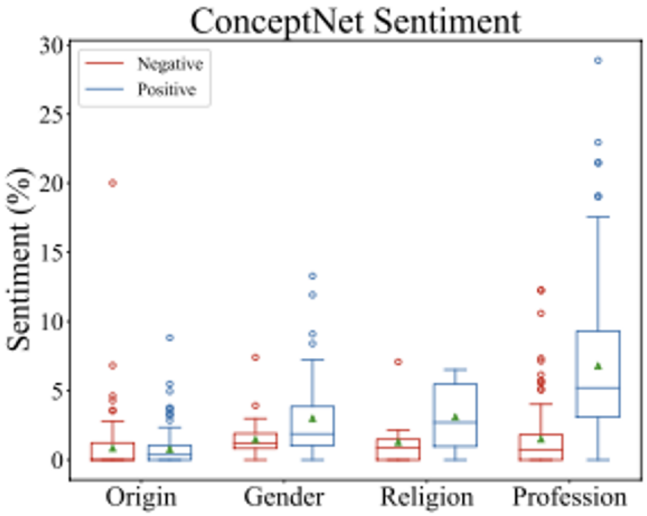
Mehrabi, Ninareh, et al. "Lawyers are Dishonest? Quantifying Representational Harms in Commonsense Knowledge Resources." *EMNLP 2021.*

USC Viterbi School of Engineering

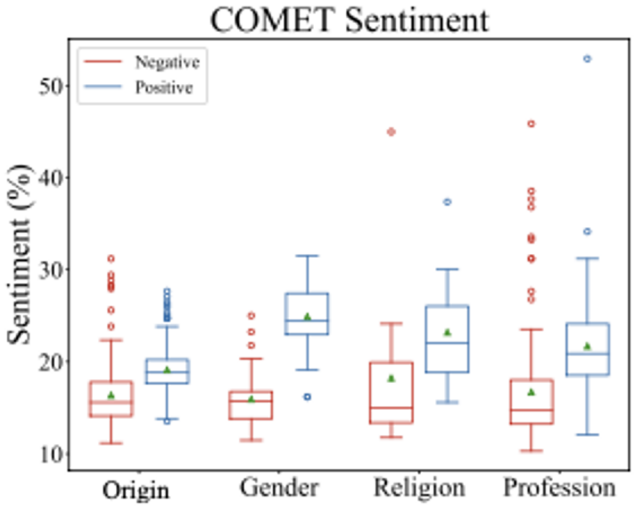
Downstream Task: COMET

- COMET is a knowledge completion/generation model.
- Trained on GPT and then ConceptNet.
- Input the same target words used in ConceptNet, collect triples
- Observed both overgeneralization and disparity in overgeneralization in both downstream tasks.

Raw Data



Predictions made by model trained on data

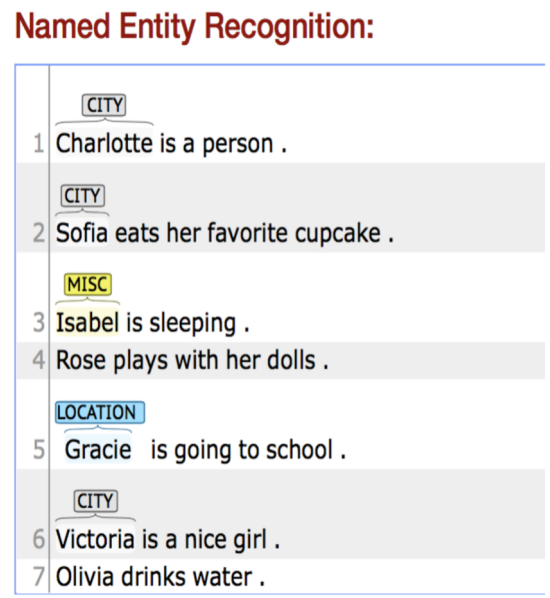


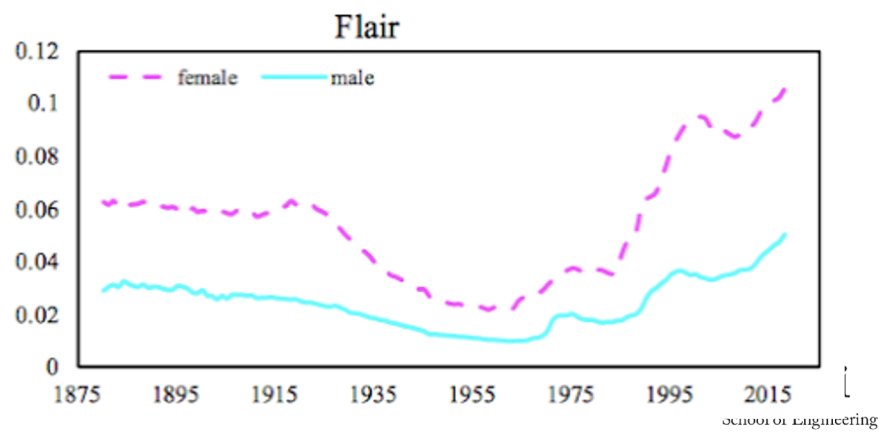
Mehrabi, Ninareh, et al. "Lawyers are Dishonest? Quantifying Representational Harms in Commonsense Knowledge Resources." *EMNLP 2021.*

USC Viterbi School of Engineering

Named Entity Recognition

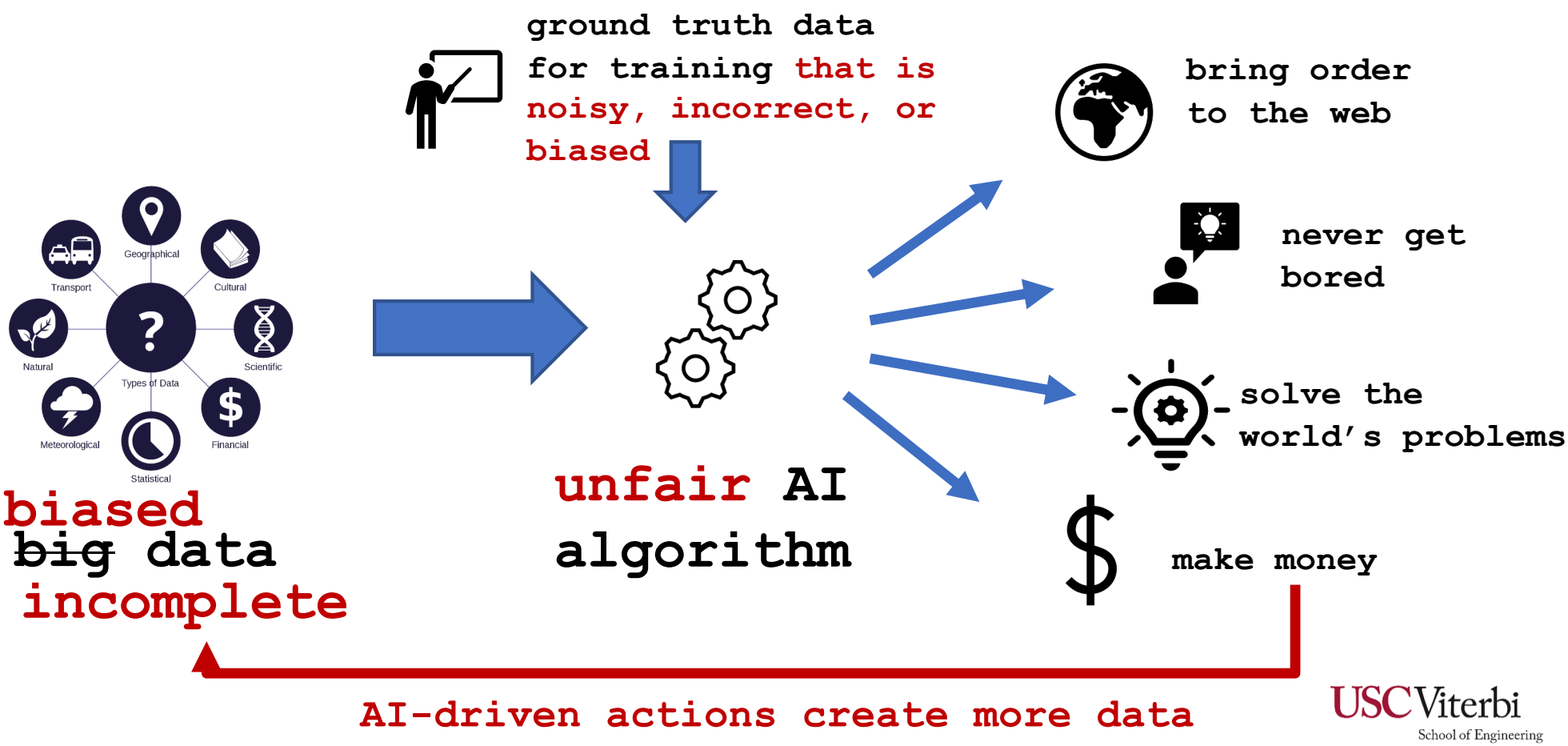
- Named Entity Recognition (NER) systems do not perform equivalently across names from different genders.
- More female names are being mis-tagged or not tagged at all compared to male names.





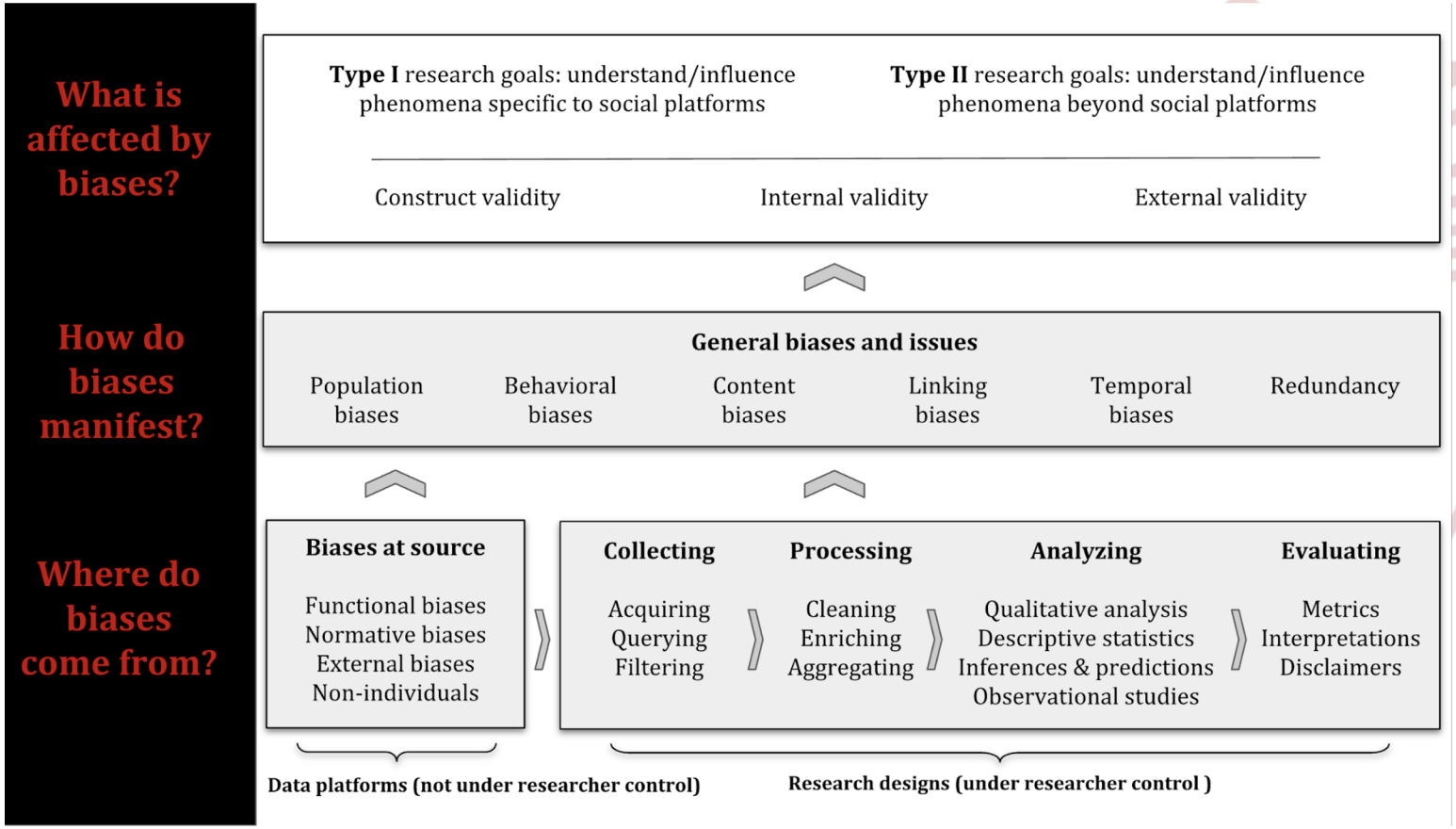
Mehrabi, N., Gowda, T., Morstatter, F., Peng, N., & Galstyan, A. (2020, July). Man is to person as woman is to location: Measuring gender bias in named entity recognition. In Proceedings of the 31st ACM Conference on Hypertext and Social Media (pp. 231-232).

AI systems



USC Viterbi School of Engineering

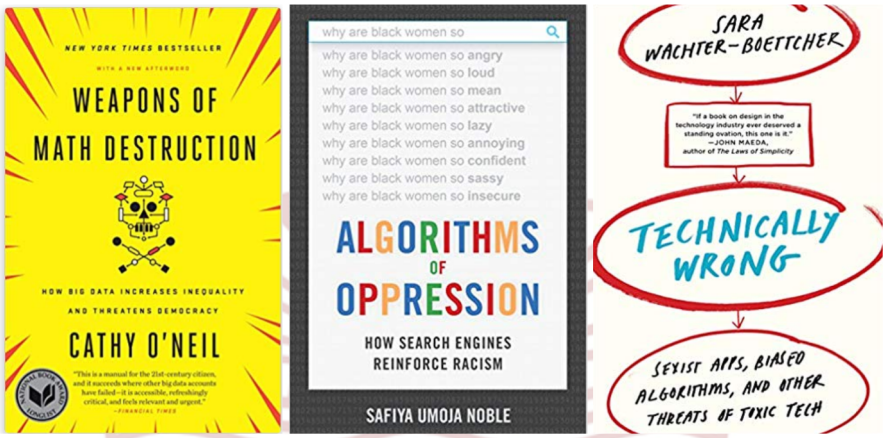
Bias in Research Pipeline



Olteanu, Alexandra, et al. "Social data: Biases, methodological pitfalls, and ethical boundaries." *Frontiers in Big Data* 2 (2019): 13.

USC Viterbi School of Engineering

Conclusions



- Much of AI bias is a reflection of human bias.
- Many definitions of fairness.
- Representational harms are nefarious; difficult to detect.
- Biases can be introduced at any step of the research process.
- Get your hands dirty: https://aif360.mybluemix.net/.

Contact: fred@isi.edu

RESEARCH-ARTICLE

A Survey on Bias and Fairness in Machine Learning

Authors: Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, Aram Galstyan

Authors Info & Claims

ACM Computing Surveys, Volume 54, Issue 6 • July 2022 • Article No.: 115, pp 1–35 • https://doi.org/10.1145/3457607

USC Viterbi School of Engineering