Bias in ML

With slides from Andy Schwartz

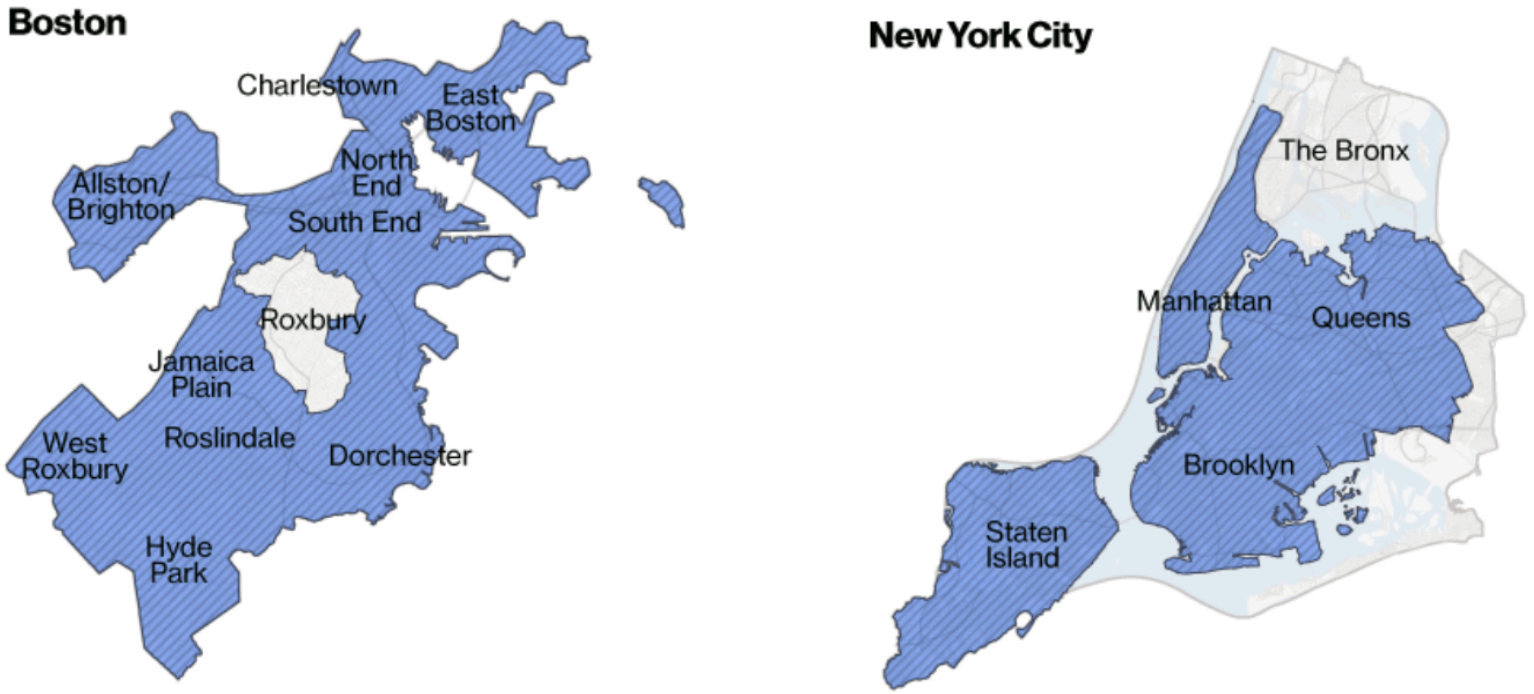
What is bias?
Sources of bias
Types of bias
Ways to reduce bias

ML models often have unintended biases

Hire? Promote? Sentence to jail?

https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

“Demographics play no role in it. Zero” - amazon



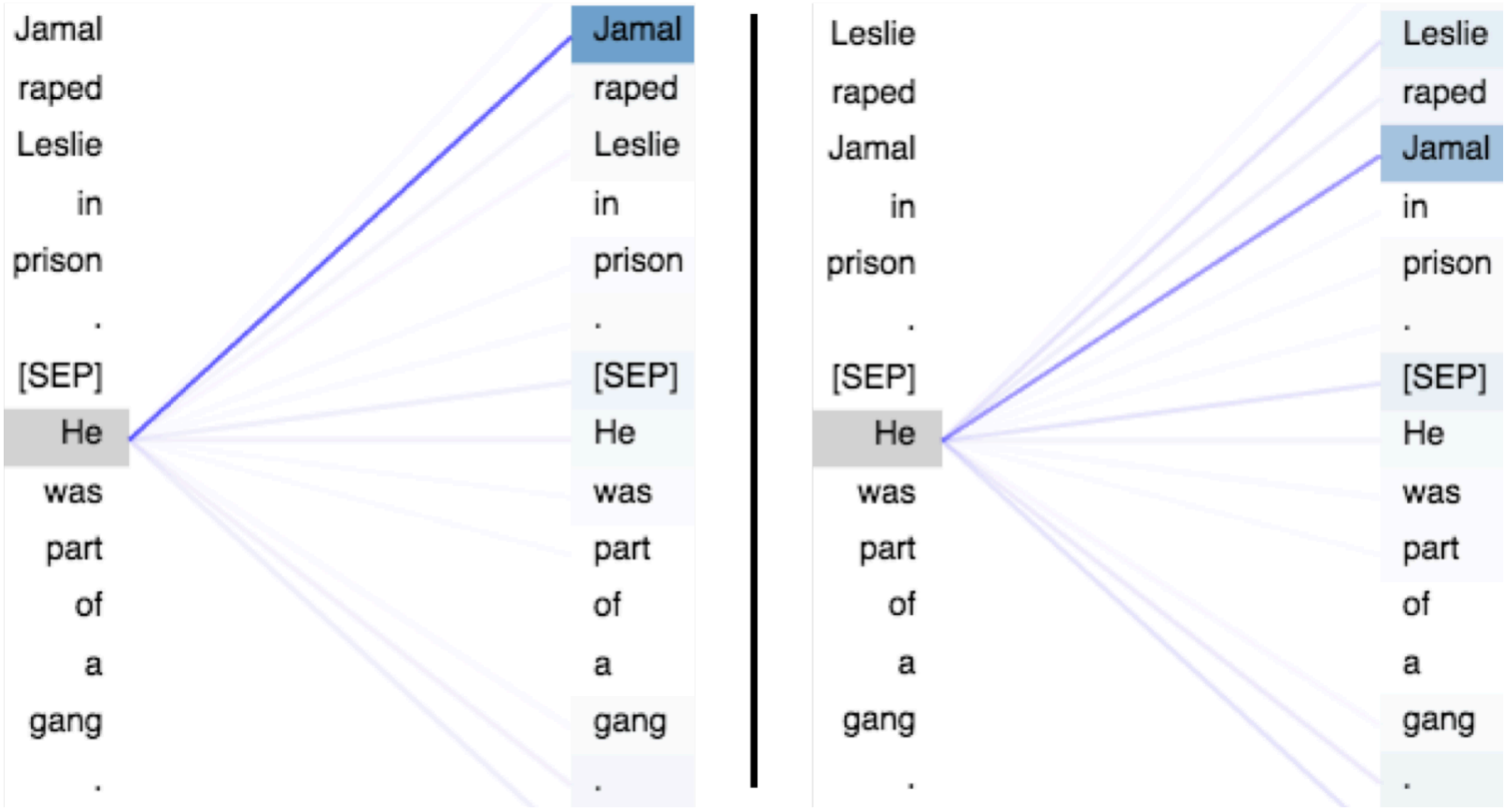
https://www.bloomberg.com/graphics/2016-amazon-same-day/

Facebook Halts Ad Targeting Cited in Bias Complaints

March 2019: Facebook stops allowing use of race, gender or age when targeting ads for housing, employment and credit.

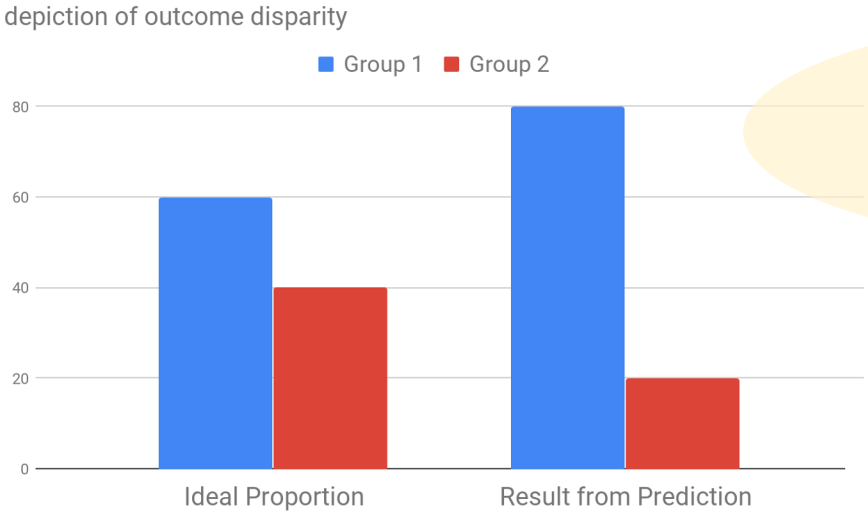
https://www.nytimes.com/2019/03/19/technology/facebook-discrimination-ads.html

Jamal is more likely than *Leslie* to be predicted to be in a gang

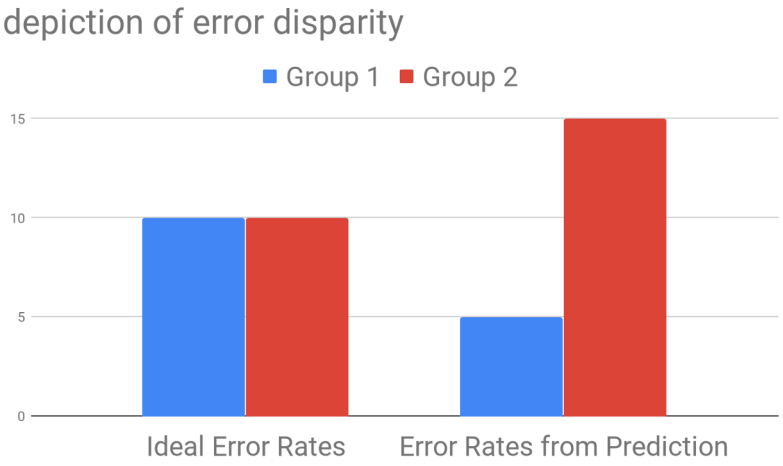


Joao Sedoc

Error and Outcome Disparity



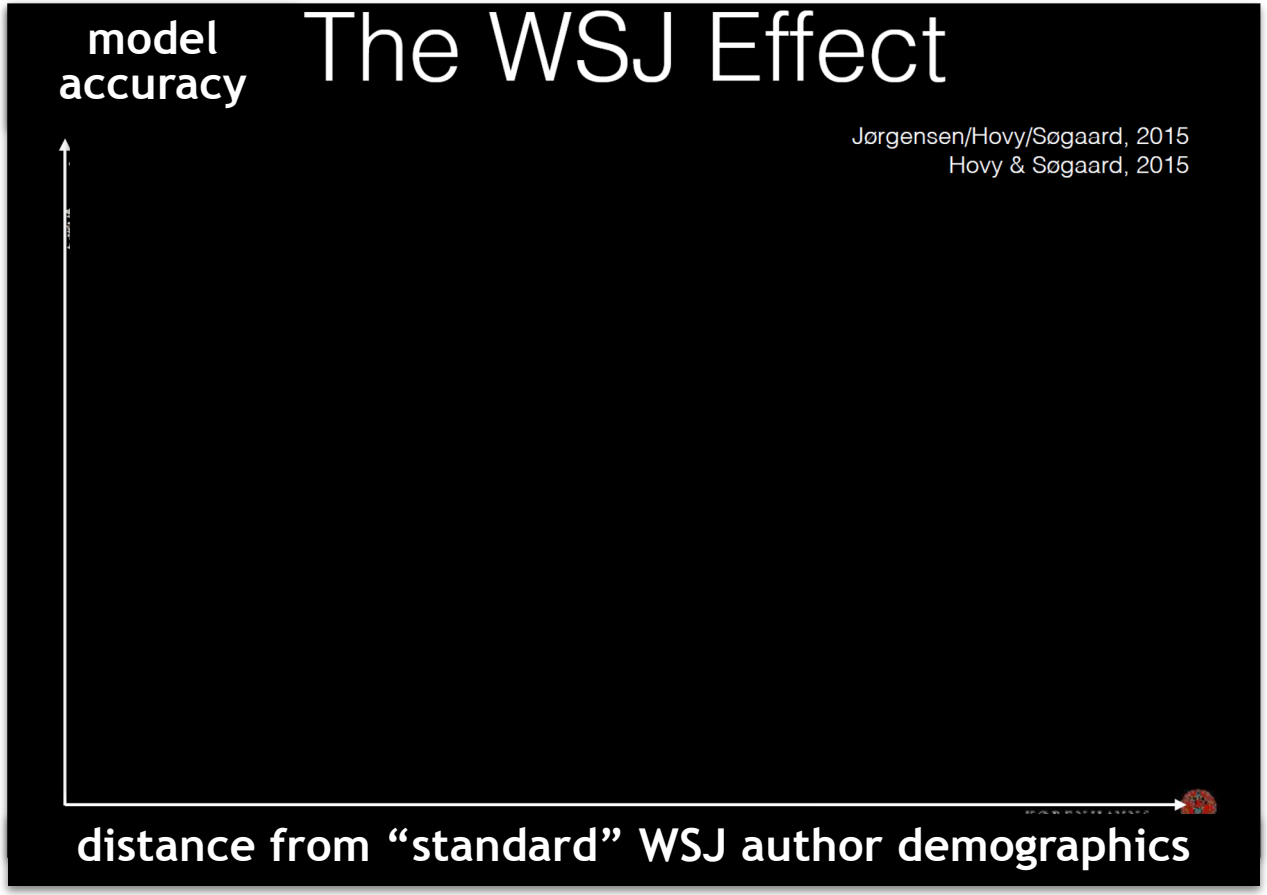
"Outcome Disparity"



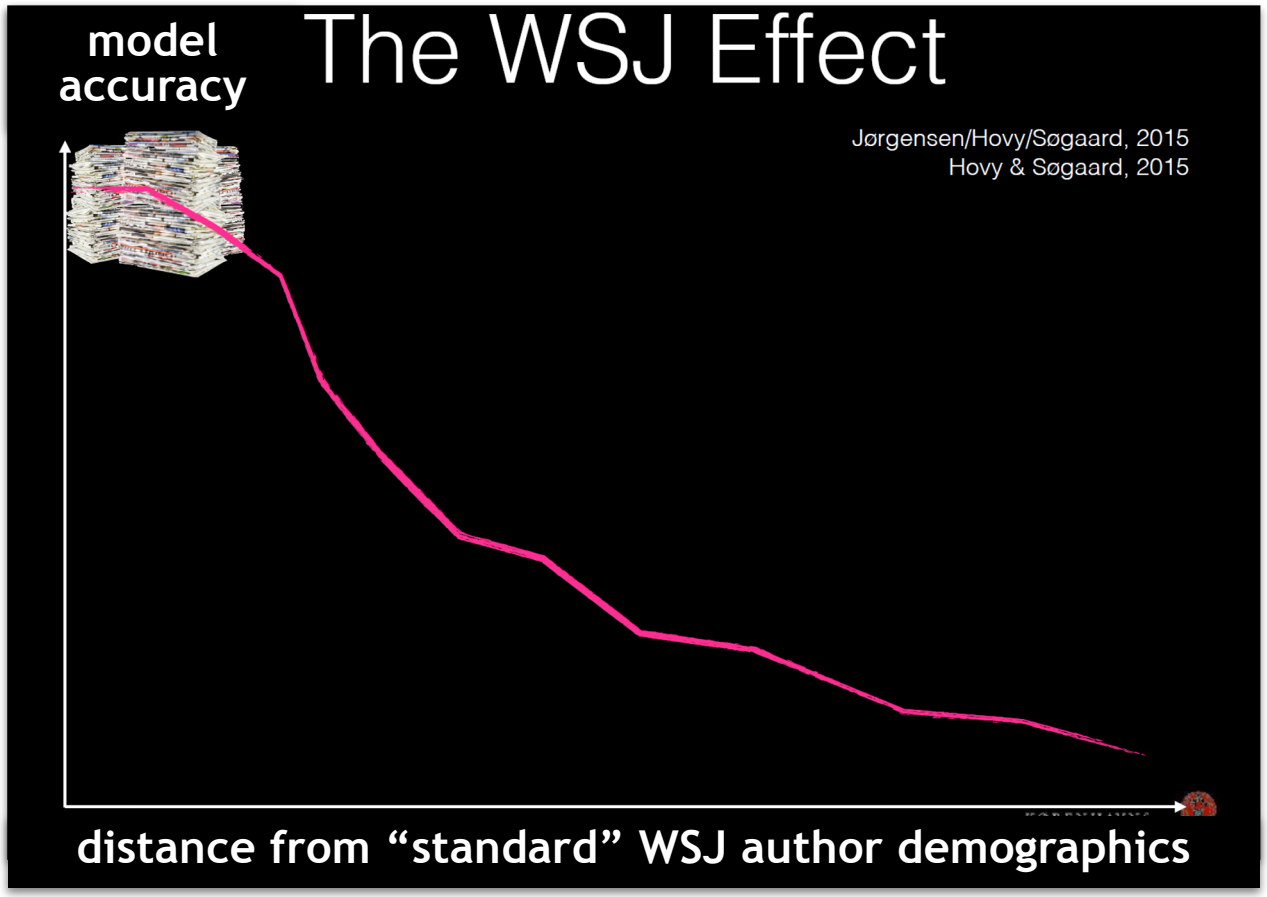
"Error Disparity"

Why do these occur?

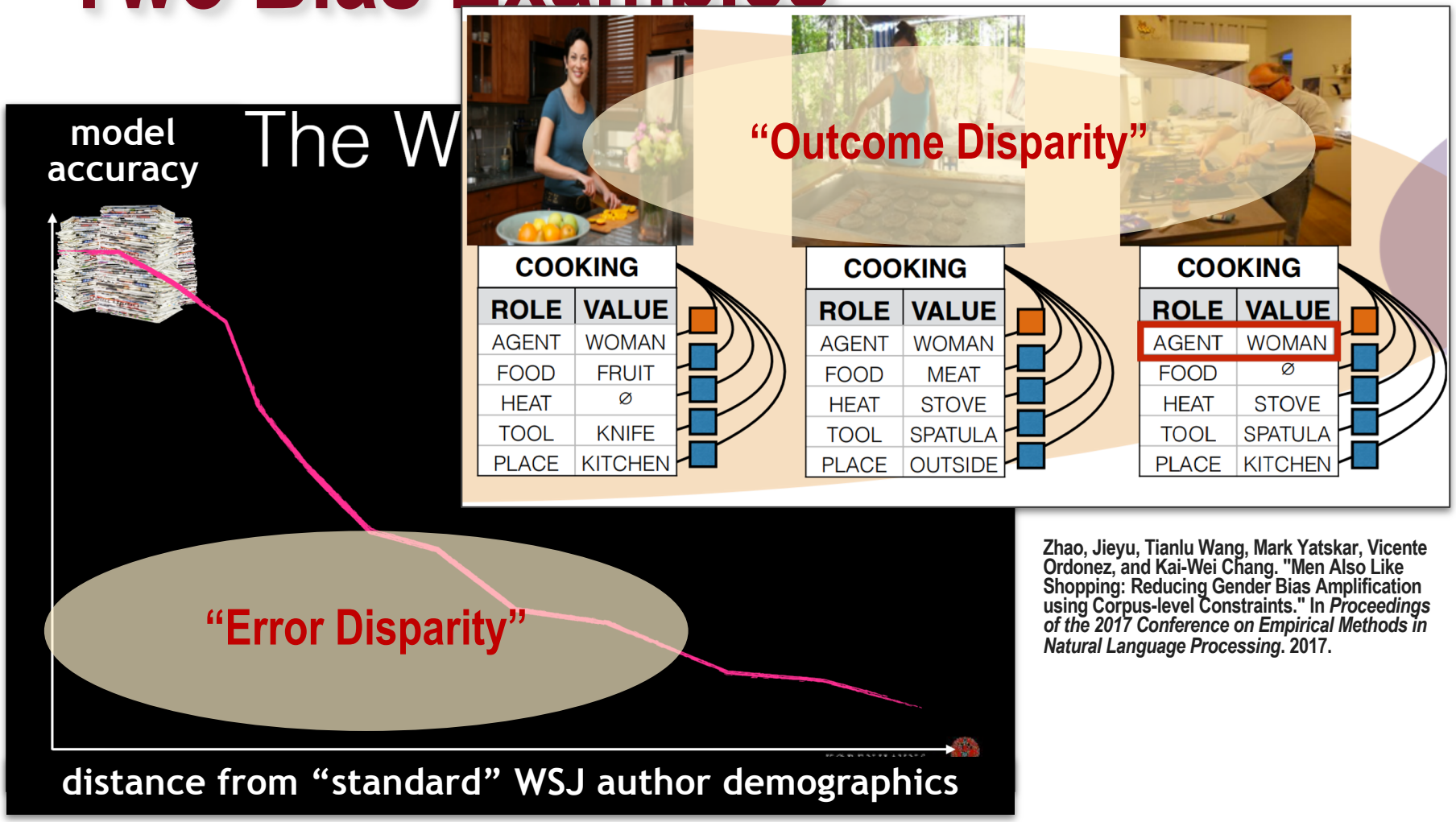
Two Bias Examples



Two Bias Examples

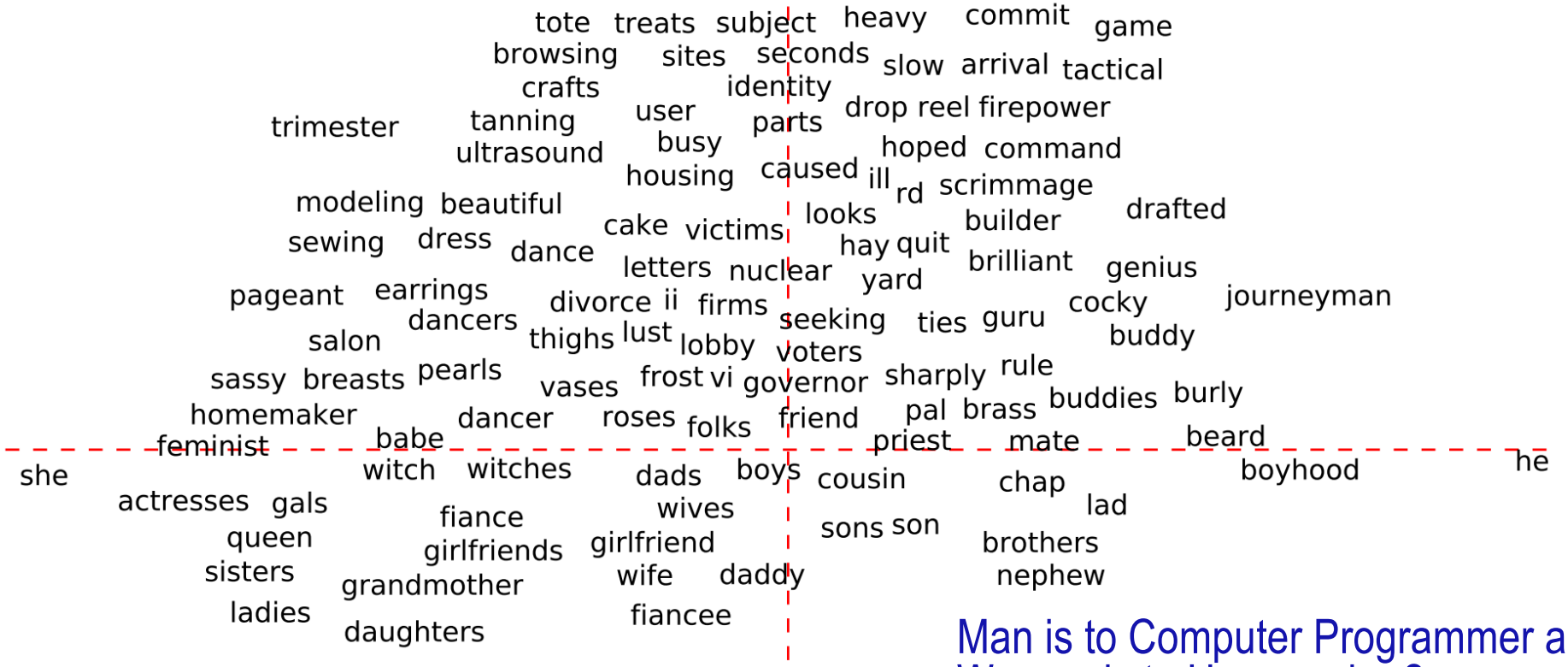


Two Bias Examples



Zhao, Jieyu, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. "Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints." In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. 2017.

Projection of word embeddings



Debias by projecting off “he/she” direction

Man is to Computer Programmer as Woman is to Homemaker?
Debiasing Word Embeddings

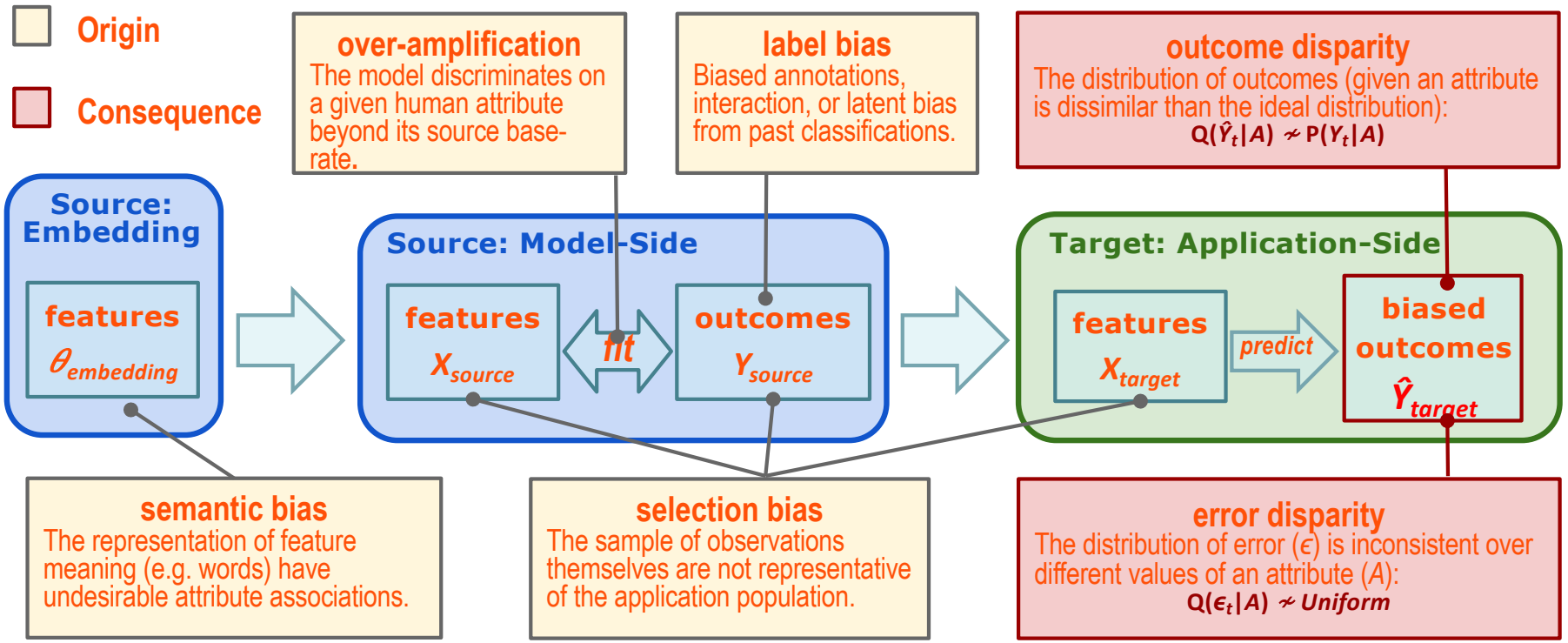
Forms of ML Bias

- **Bias perpetuation**
  - Historic labels or correlations (affecting embeddings)
- **Sampling bias**
  - Non-representative training data
- **Bias amplification ("Outcome disparity")**
  - Under ignorance, predict the most frequently seen label
- **Majority class bias ("Error disparity")**
  - Higher accuracy on more frequent classes

Bias Correction

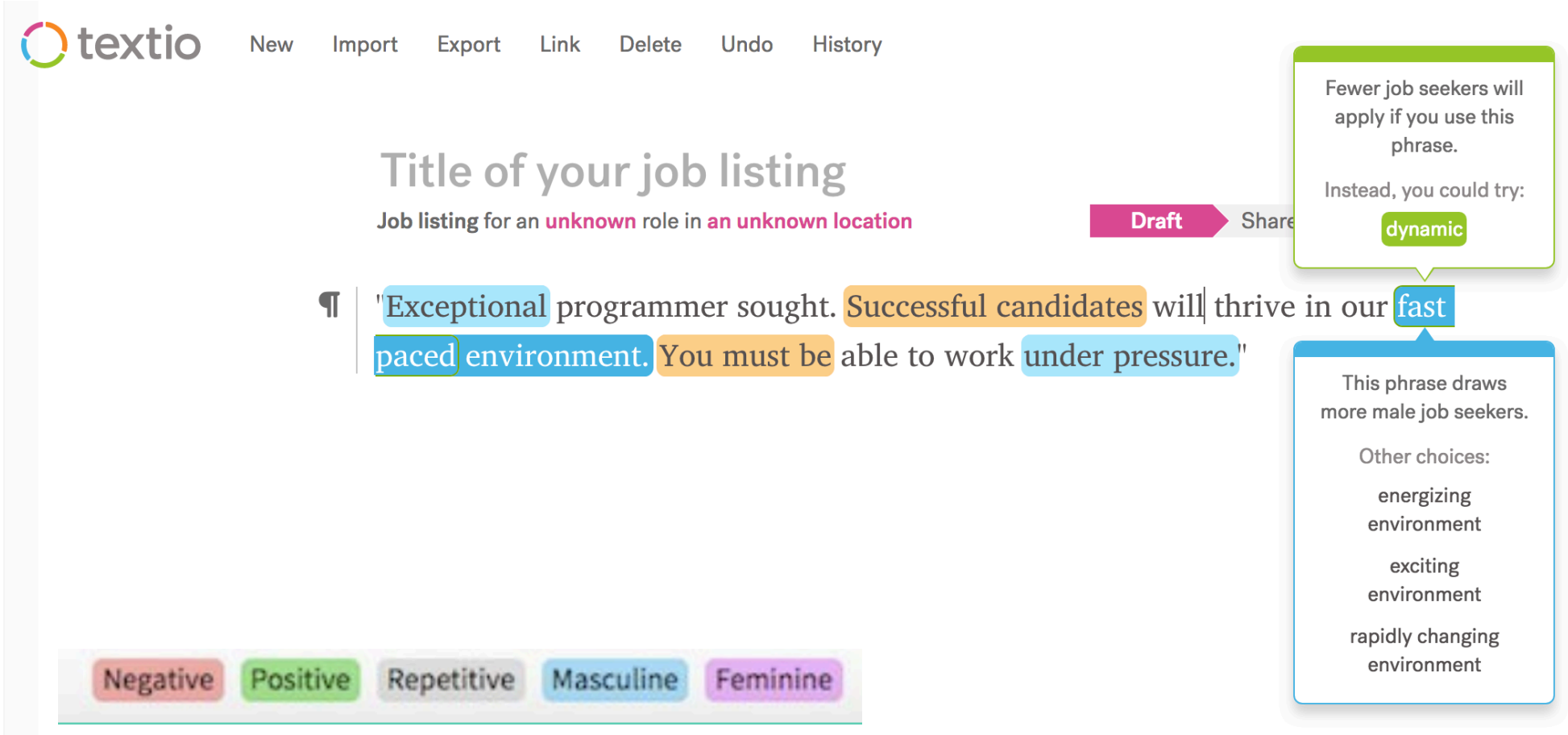
- **Bias perpetuation**
  - Adjust labels, embeddings
- **Sampling bias**
  - Re-weighting – or get more data
- **Bias amplification**
  - Recalibrate
- **Majority class bias**
  - Use loss function that treats every class equally rather than every instance

An ML pipeline and its biases



Andy Schwartz

Analytics can reduce bias



What you should know

- **Many forms of bias**
  - Perpetuation, sampling, amplification, majority class
- **Addressing bias requires picking the right training data and loss function**