Bias in ML

Learning objectives
What is bias?
Sources of bias
Types of bias
Ways to reduce bias
Bias = problems with transfer

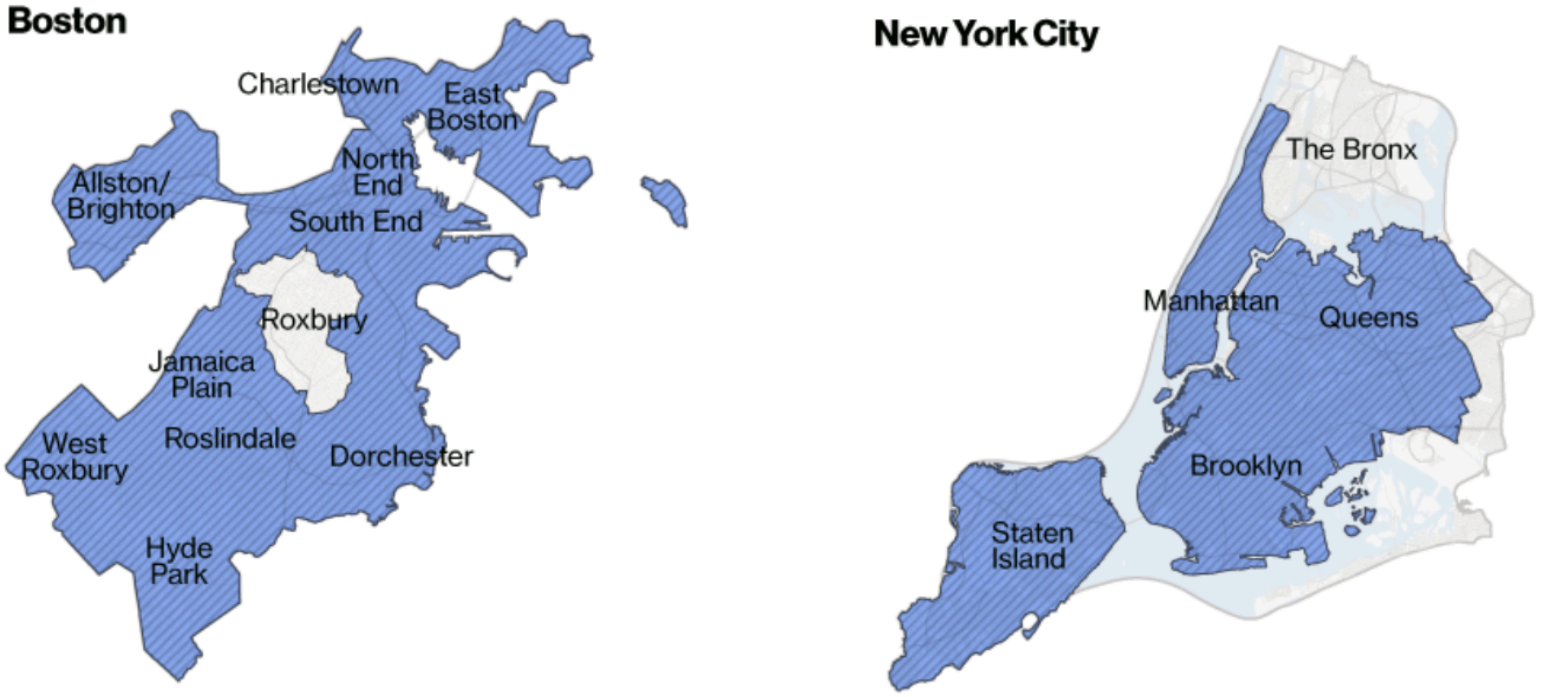
With slides from Andy Schwartz

Hire? Promote? Sentence to jail?

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

ML models often have unintended biases

“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

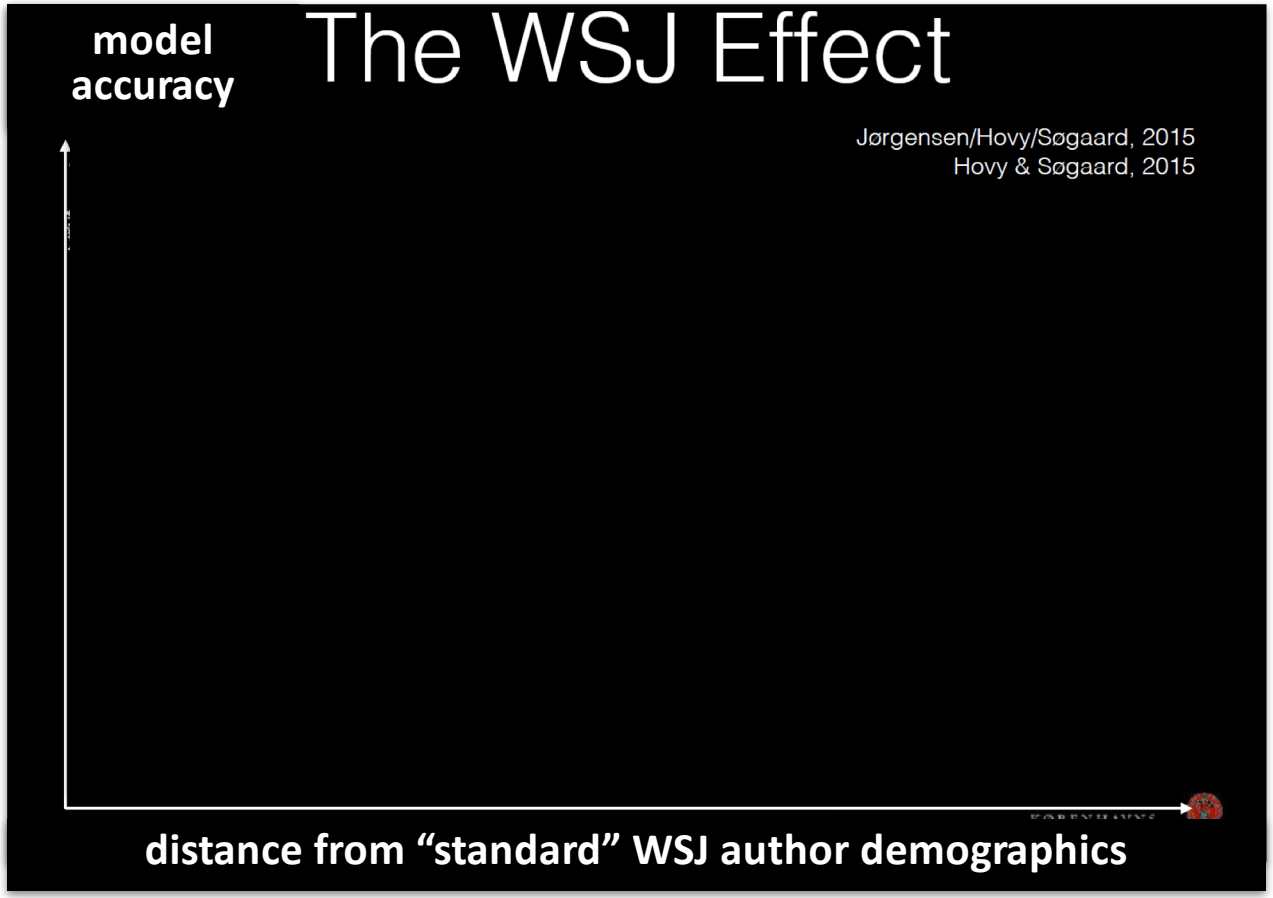
March 2019: Facebook stops allowing use of race, gender or age when targeting ads for housing, employment and credit.

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

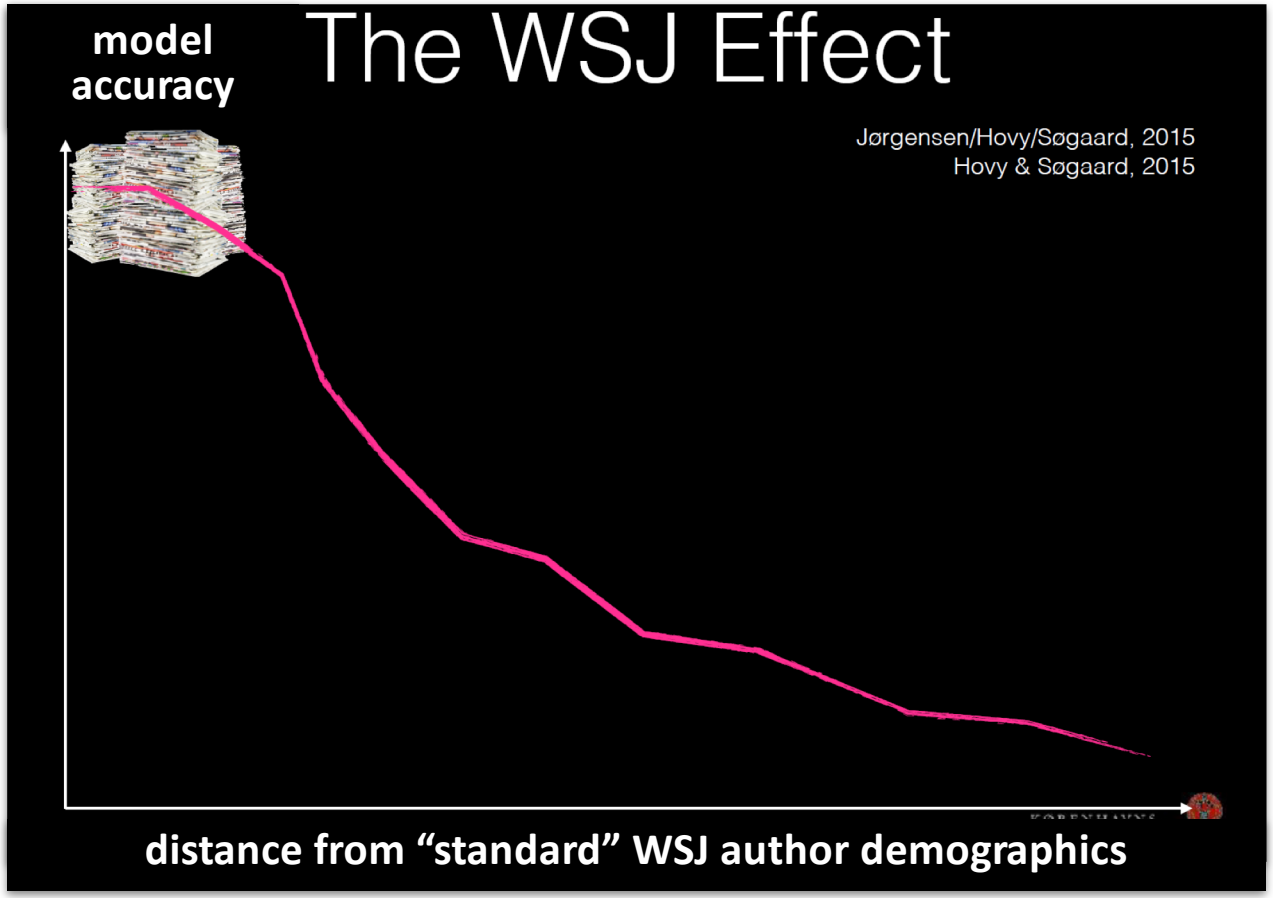
Jamal is more likely than *Leslie* to be predicted to be in a gang

Joao Sedoc

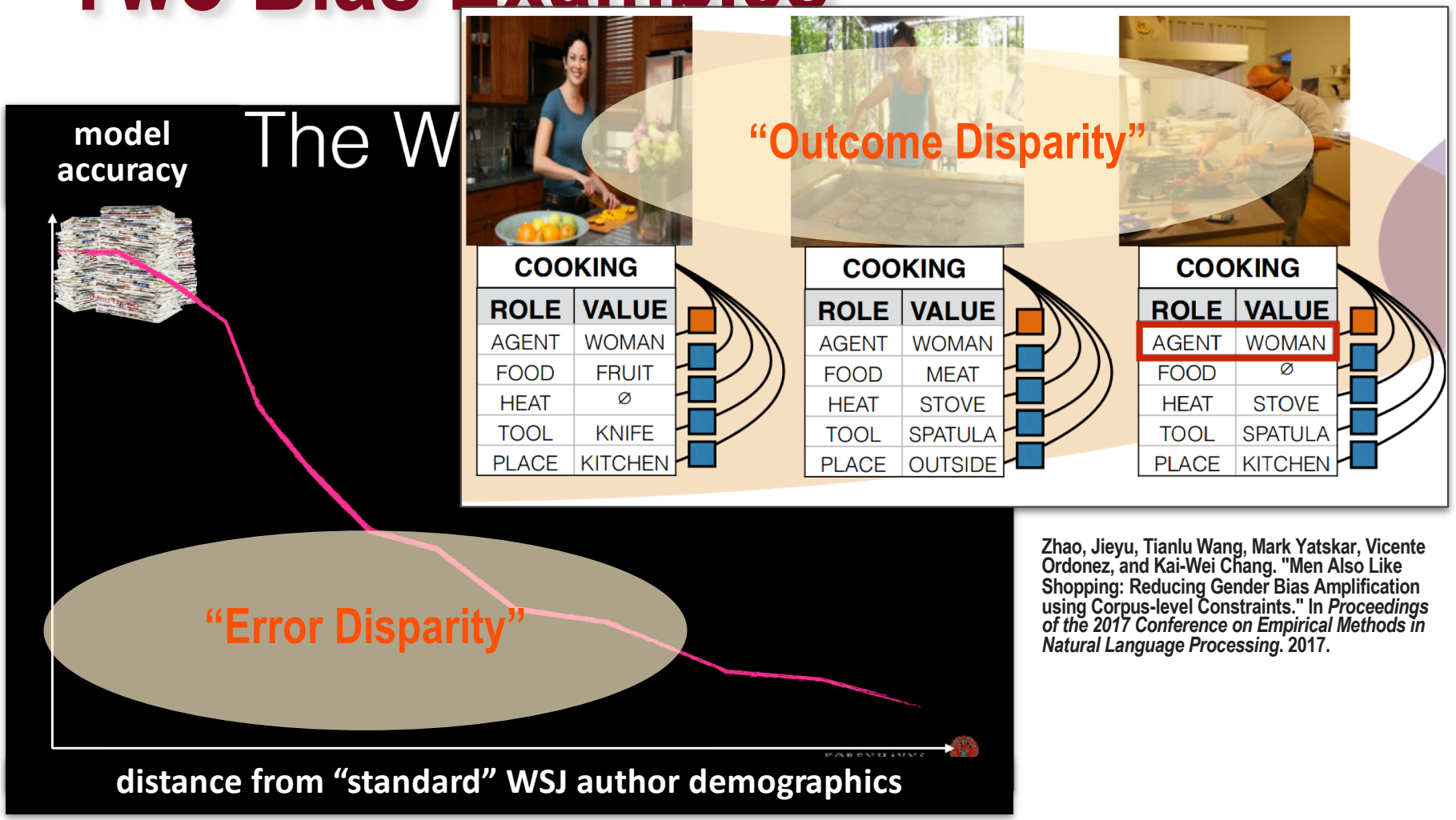
Two Bias Examples



Two Bias Examples

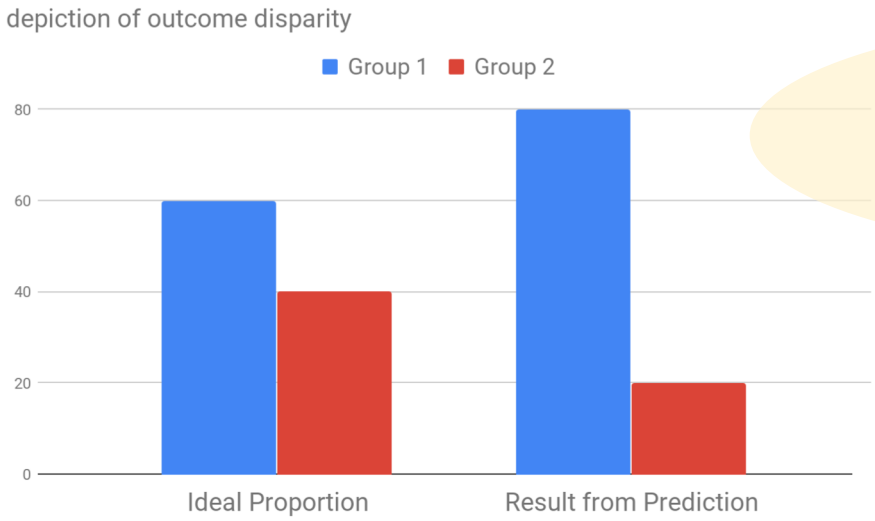


Two Bias Examples

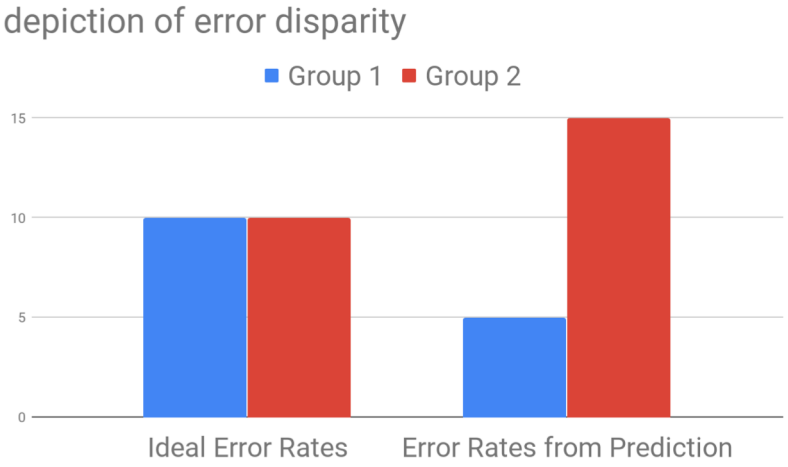


Zhao, Jieyu, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. "Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints." In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. 2017.

Error and Outcome Disparity



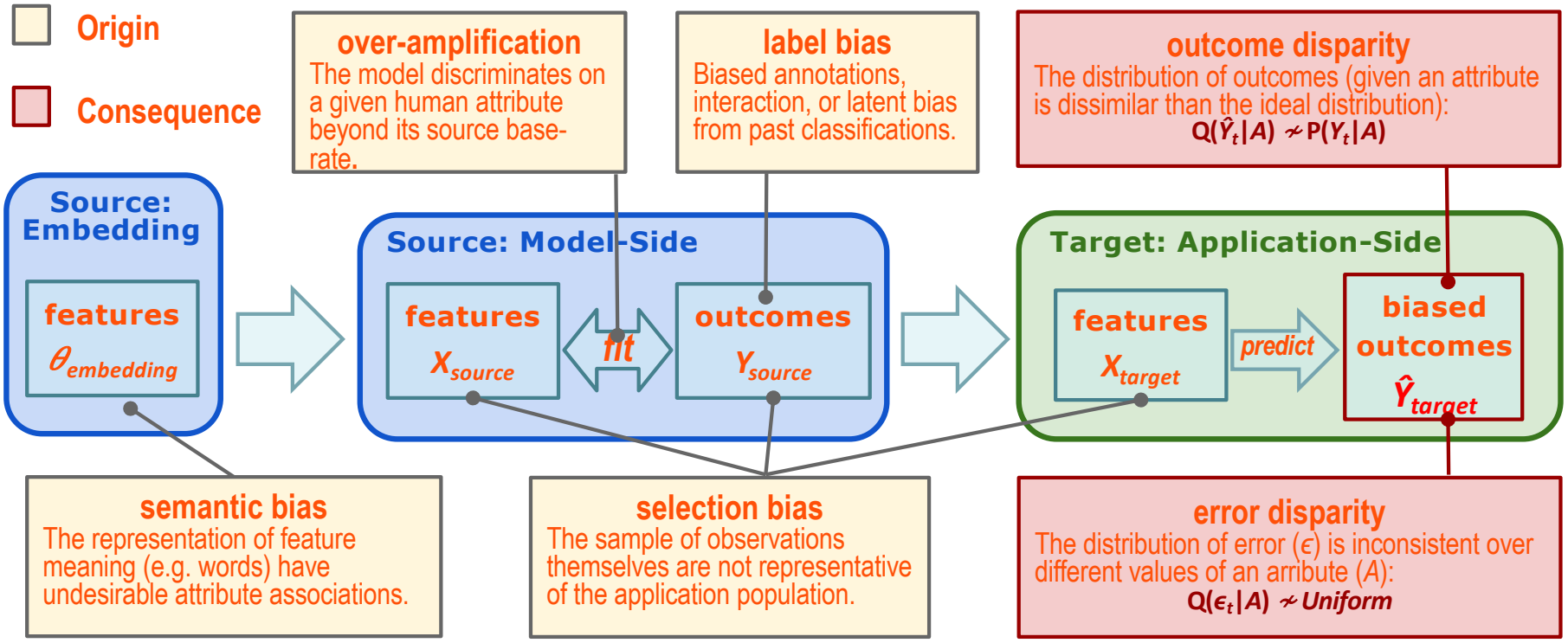
"Outcome Disparity"



"Error Disparity"

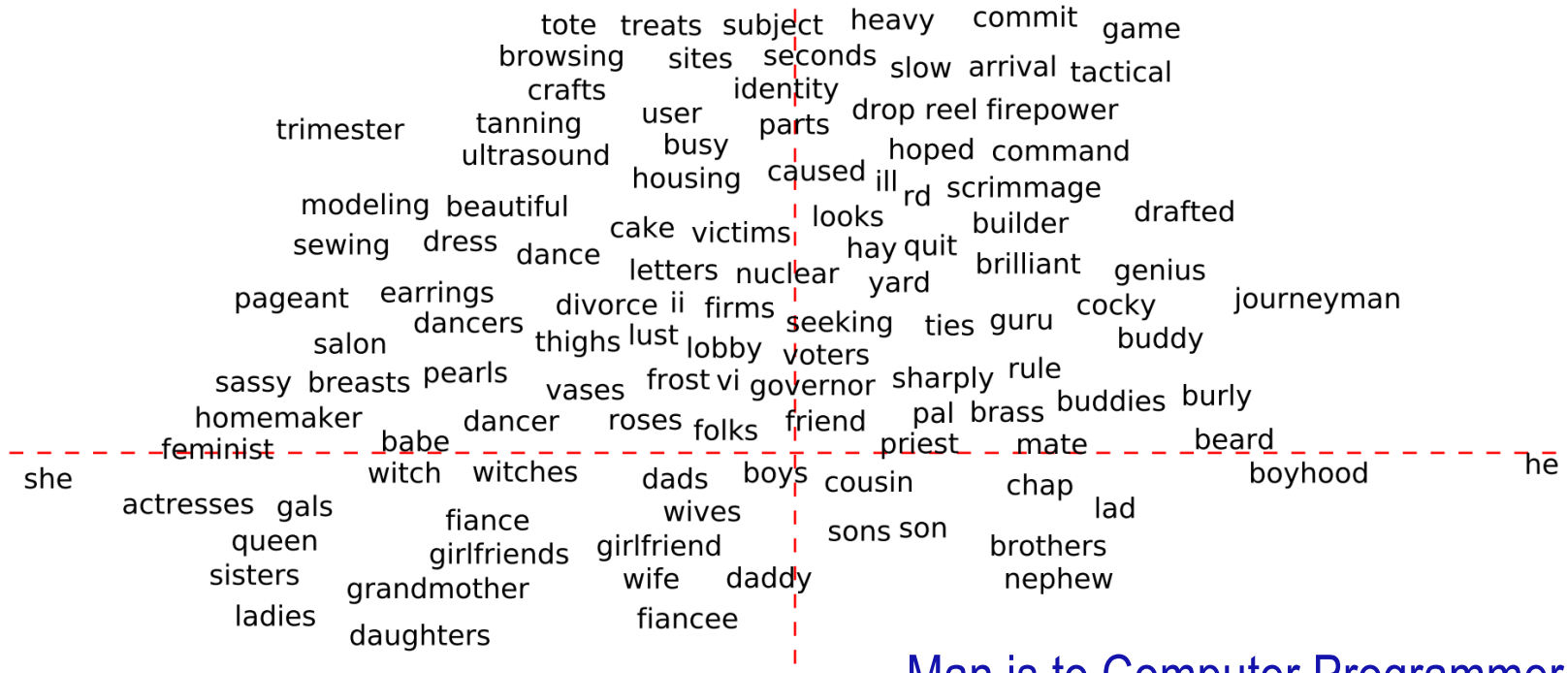
Why do these occur?

An ML pipeline and its biases



Andy Schwartz

Projection of word embeddings

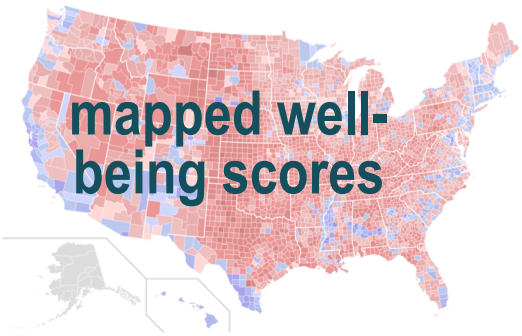


Man is to Computer Programmer as Woman is to Homemaker?
Debiasing Word Embeddings

Debias by projecting off “he/she” direction

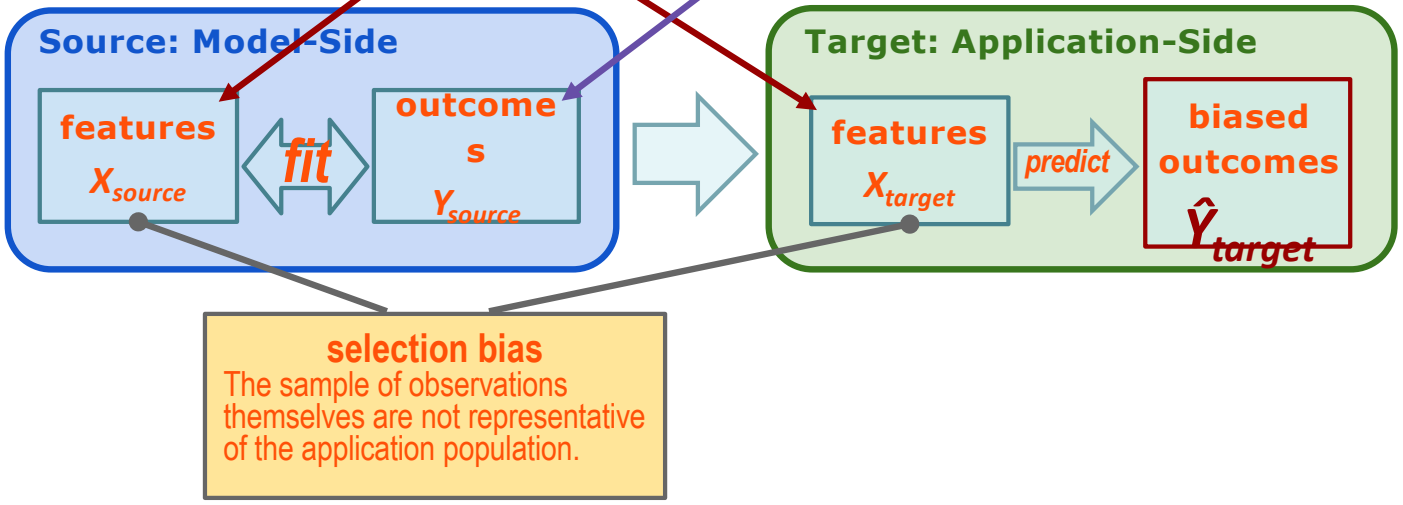
Debias by restratifying

language



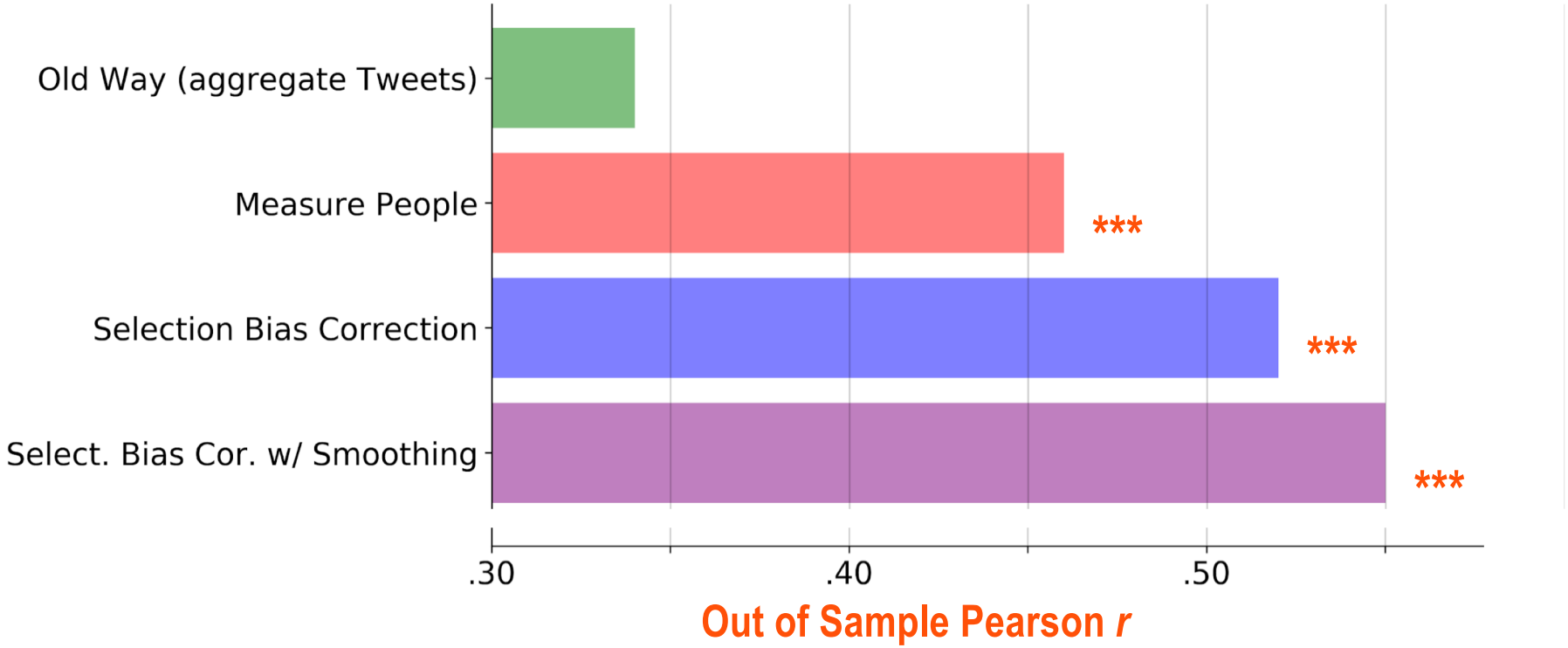
non-representative data

representative data (from surveys)



Andy Schwartz

Combine multiple adjustment methods

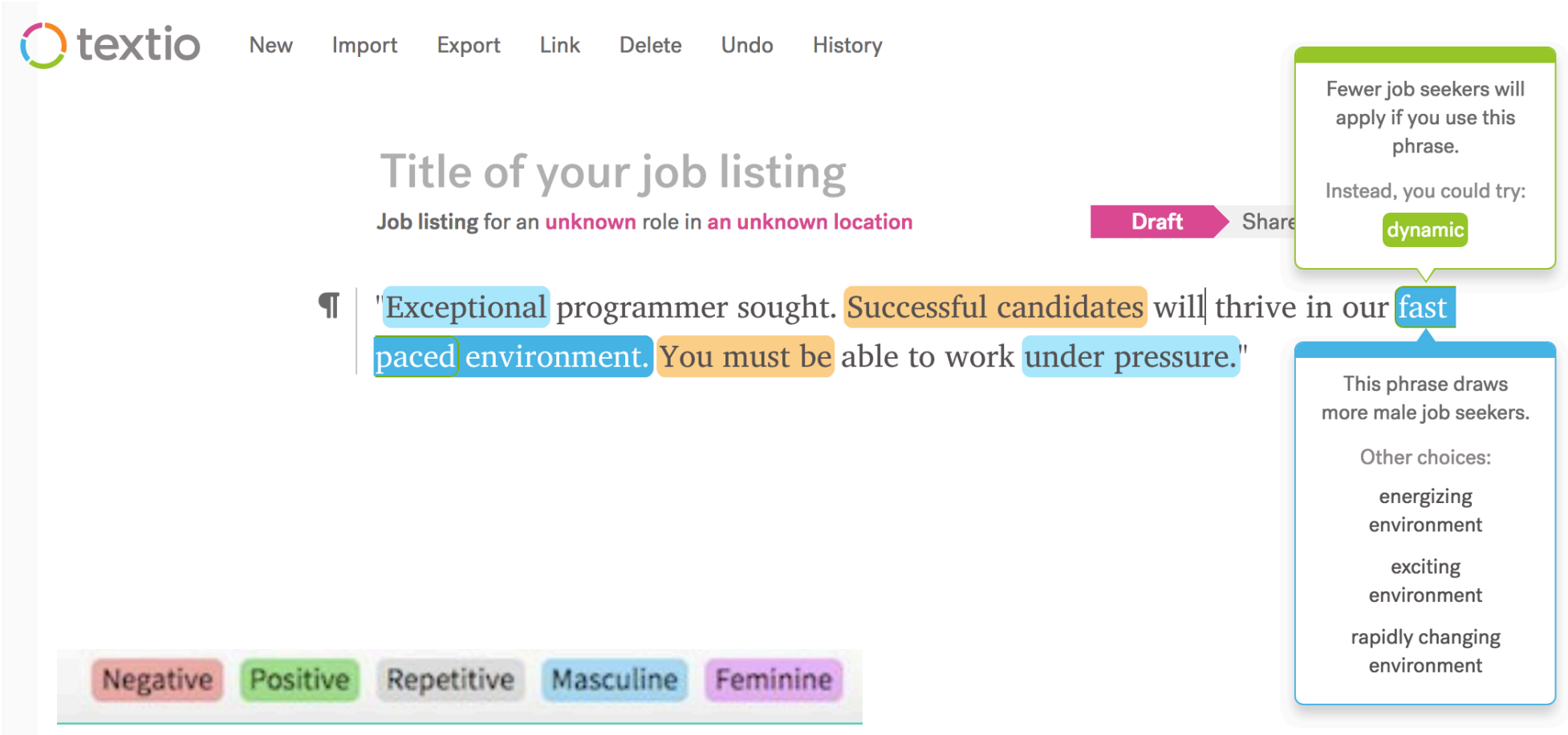


N = 2040 US counties

*** significant p < .005 improvement

Giorgi, et al. 2018, 2019

Analytics can reduce bias



Forms of ML Bias

- **Bias perpetuation**
  - Historic labels or correlations (affecting embeddings)
- **Sampling bias**
  - Non-representative training data
- **Bias amplification**
  - Under ignorance, predict the most frequently seen label
- **Majority class bias**
  - 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

Transfer Learning Questions

- **Is the correlation between features stable?**
  - If so, transfer feature transformations $z_k = g_k(\mathbf{x})$
- **Are the label frequencies stable?**
  - If not, recalibrate or adjust the threshold or restratify
- **Are the 'distant' labels representative?**
  - If not, can one adjust them?