Bias in ML

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

With slides from Andy Schwartz

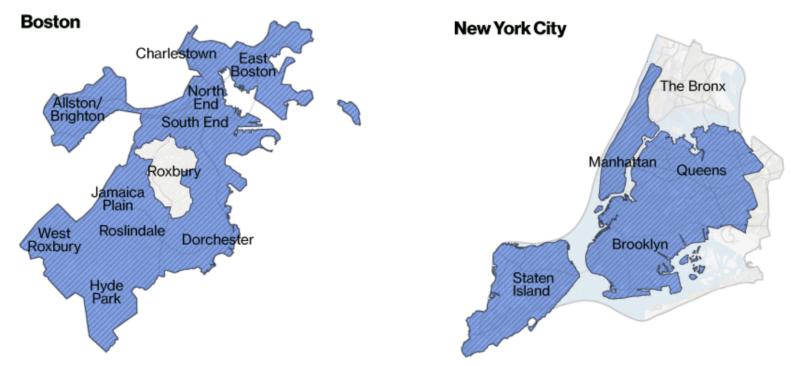
Hire? Promote? Sentence to jail?



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/

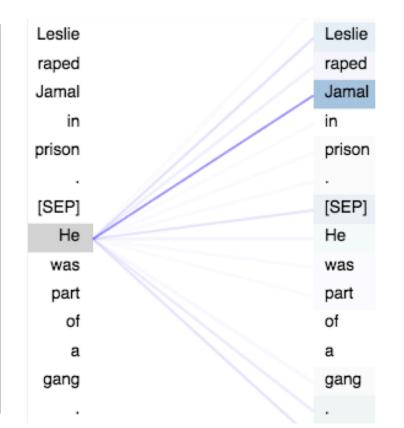
Facebook Halts Ad Targeting Cited in Bias Complaints

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/facebookdiscrimination-ads.html

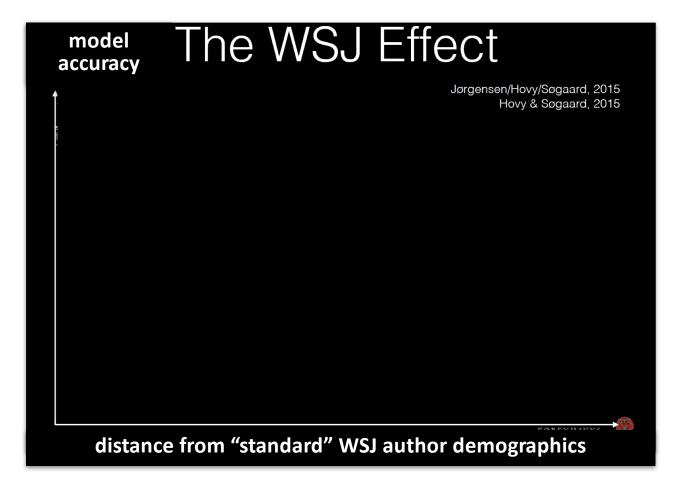
Jamal is more likely than Leslie to be predicted to be in a gang

Jamal	Jamal
raped	raped
Leslie	Leslie
in	in
prison	prison
[SEP]	[SEP]
He	He
was	was
part	part
of	of
а	а
gang	gang

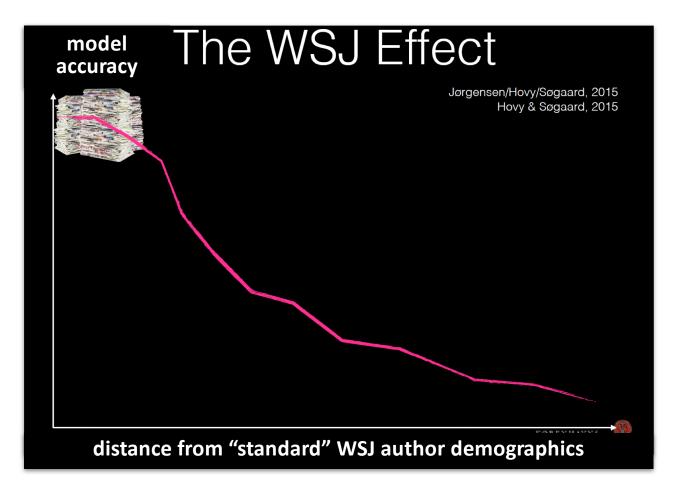


Joao Sedoc

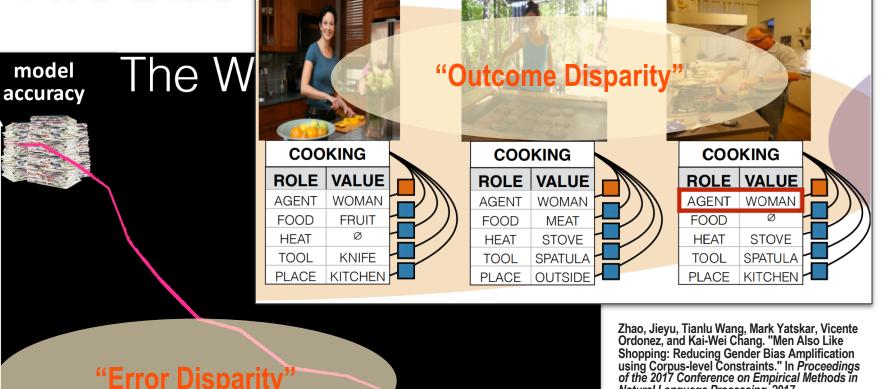
Two Bias Examples



Two Bias Examples



Two Bias Examples



"Error Disparity"

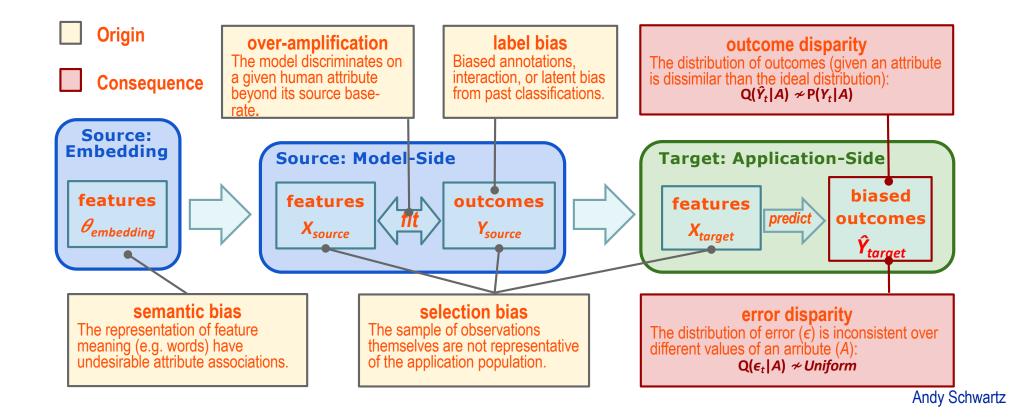
Natural Language Processing. 2017.

distance from "standard" WSJ author demographics

Error and Outcome Disparity

depiction of outcome disparity Group 1 Group 2 80 "Outcome Disparity" 60 40 20 depiction of error disparity **Ideal Proportion Result from Prediction** ■ Group 1 ■ Group 2 15 10 "Error Disparity" Why do these occur? Ideal Error Rates Error Rates from Prediction

An ML pipeline and its biases

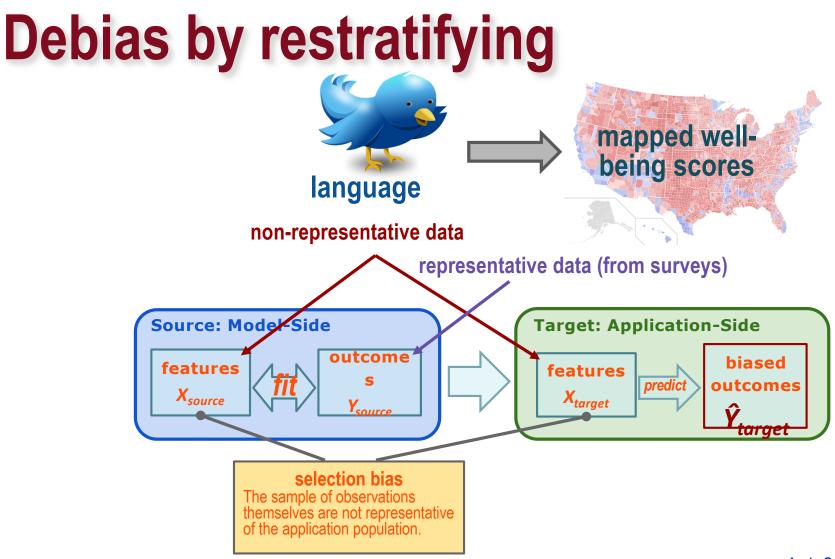


Projection of word embeddings

tote treats subject heavy commit game sites seconds slow arrival tactical browsing identity parts drop reel firepower crafts user tanning trimester busy hoped command ultrasound housing caused ill rd scrimmage modeling beautiful drafted looks builder cake victims sewing dress dance hay quit brilliant genius letters nuclear yard pageant earrings divorce ii firms seeking ties guru cocky dancers thighs lust lobby voters bud journeyman buddv vases frost vi governor sharply rule sassy breasts pearls babe______babe_______friend pal brass buddies burly witch witches homemaker beard feminist _ _ _ hē witch witches dads boys cousin boyhood she chap actresses gals lad wives fiance sons son girlfriends girlfriend queen brothers sisters wife daddy nephew grandmother ladies fiancee daughters Man is to Computer Programmer as

Woman is to Homemaker? Debiasing Word Embeddings

Debias by projecting off "he/she" direction



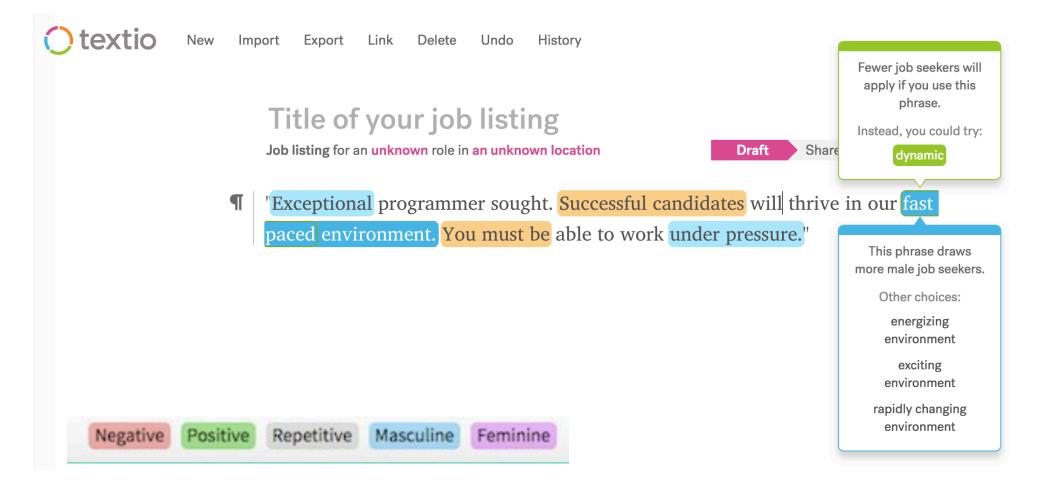
Andy Schwartz

Combine multiple adjustment methods



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?