

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



New York City



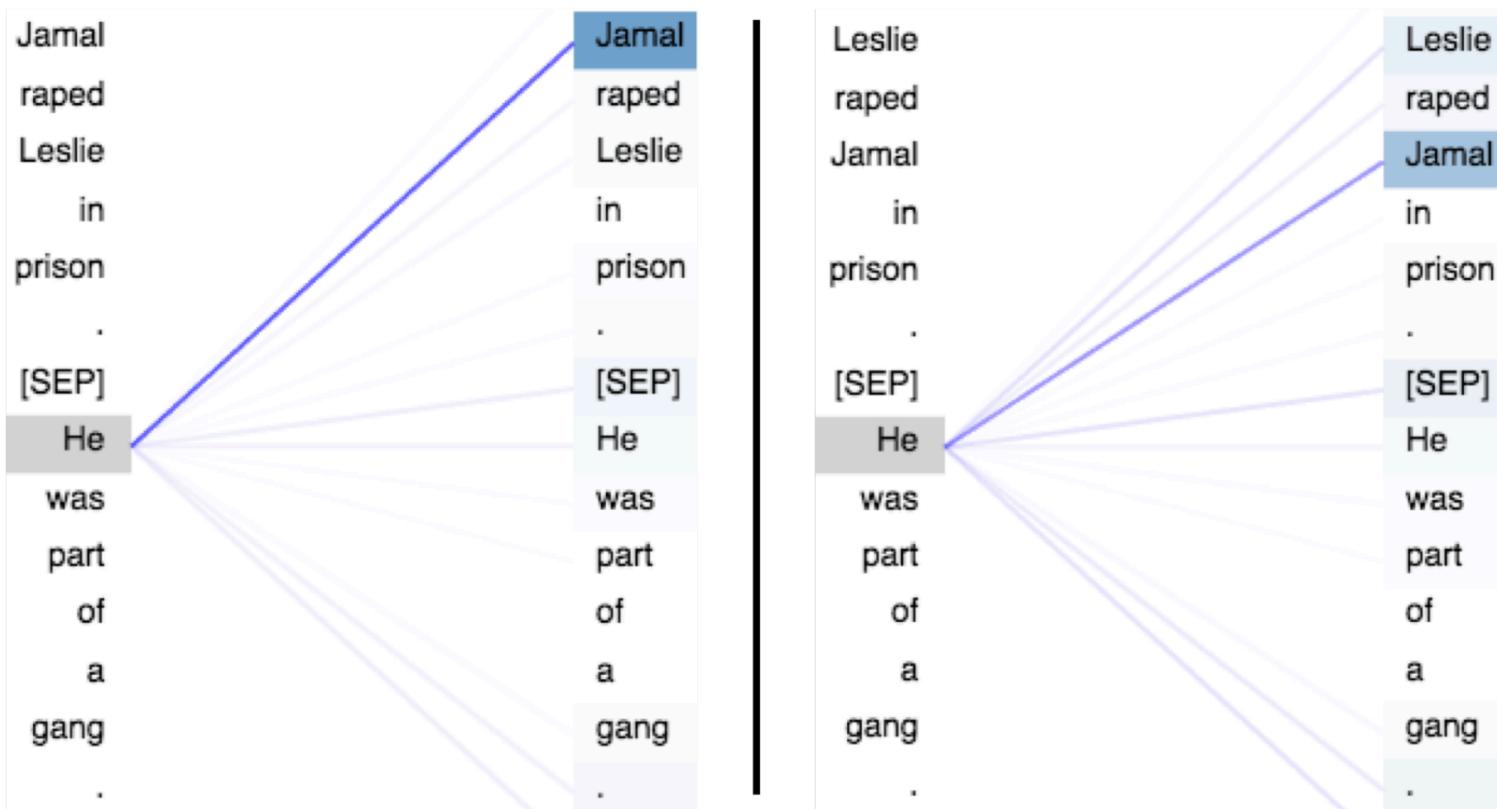
<https://www.bloomberg.com/graphics/2016-amazon-same-day/>

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>

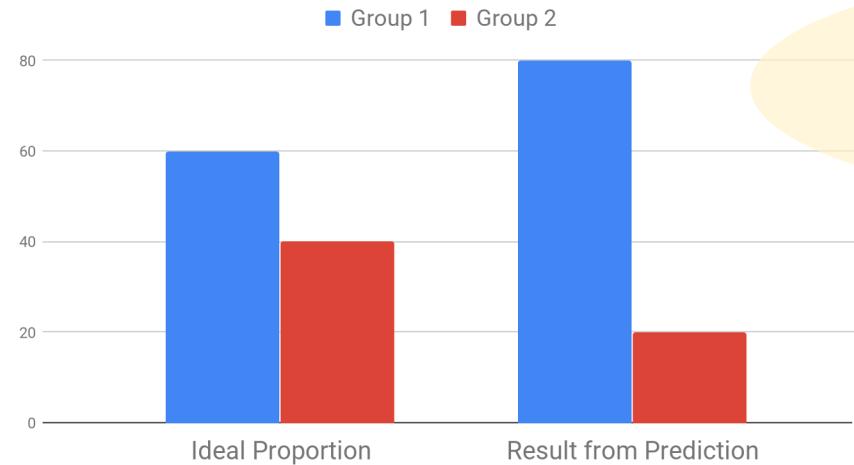
Jamal is more likely than *Leslie* to be predicted to be in a gang



Joao
Sedoc

Error and Outcome Disparity

depiction of outcome disparity



“Outcome Disparity”

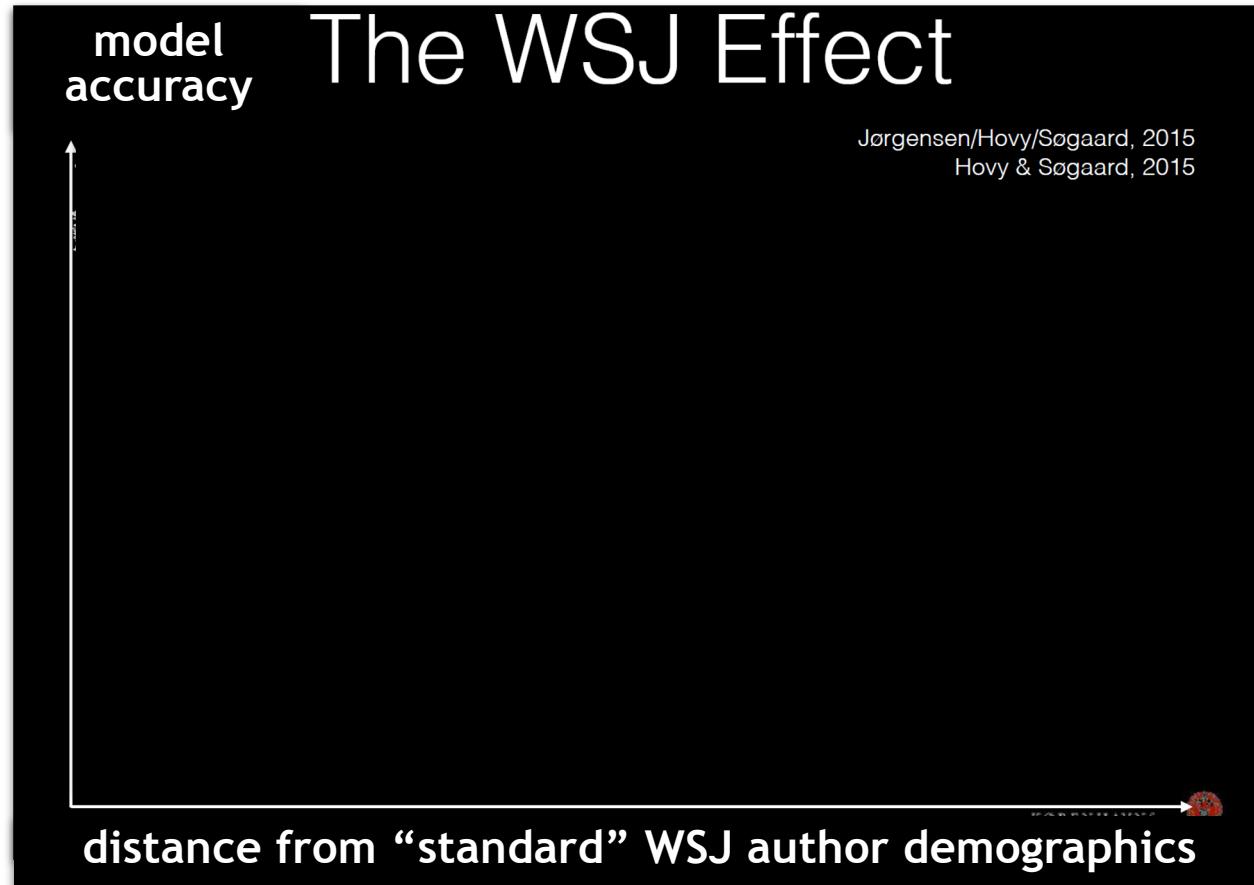
depiction of error disparity



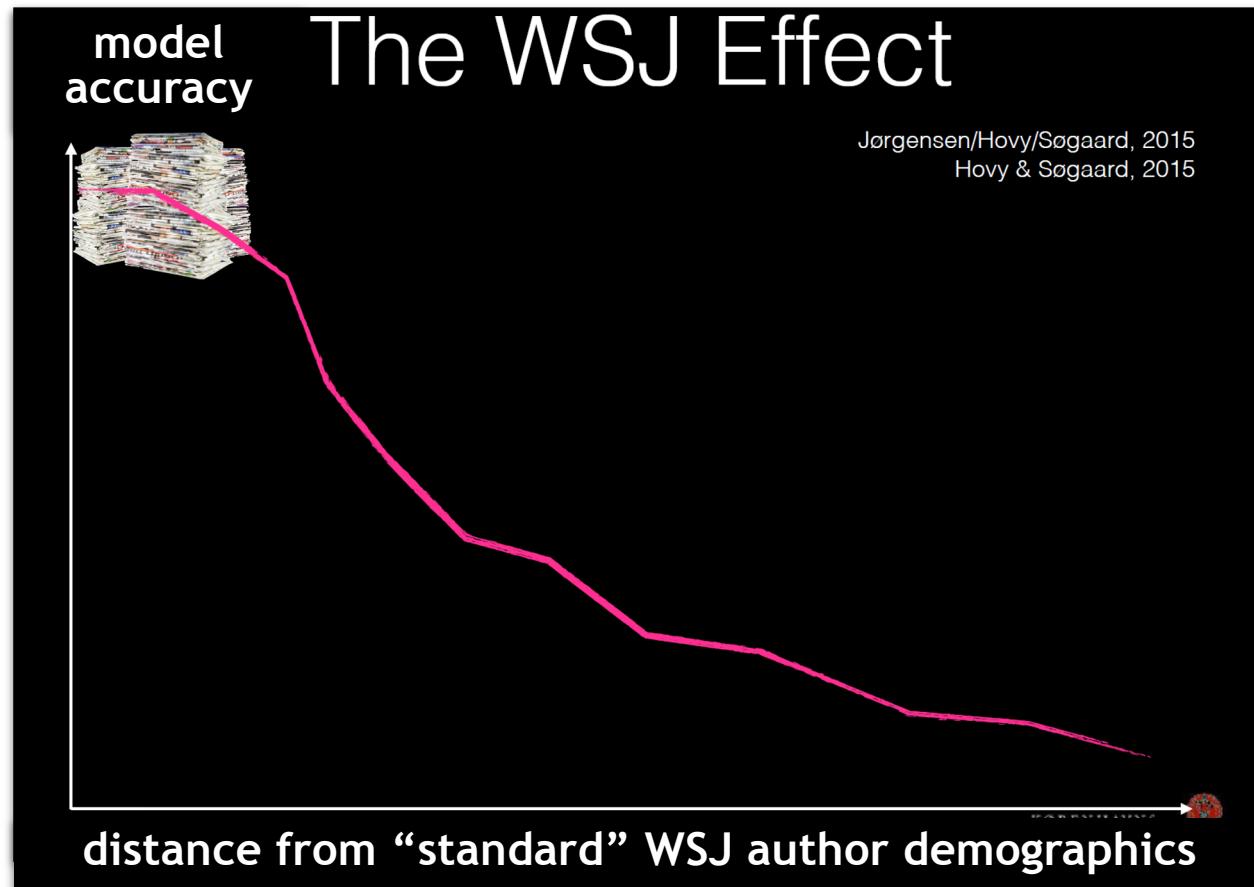
“Error Disparity”

Why do these occur?

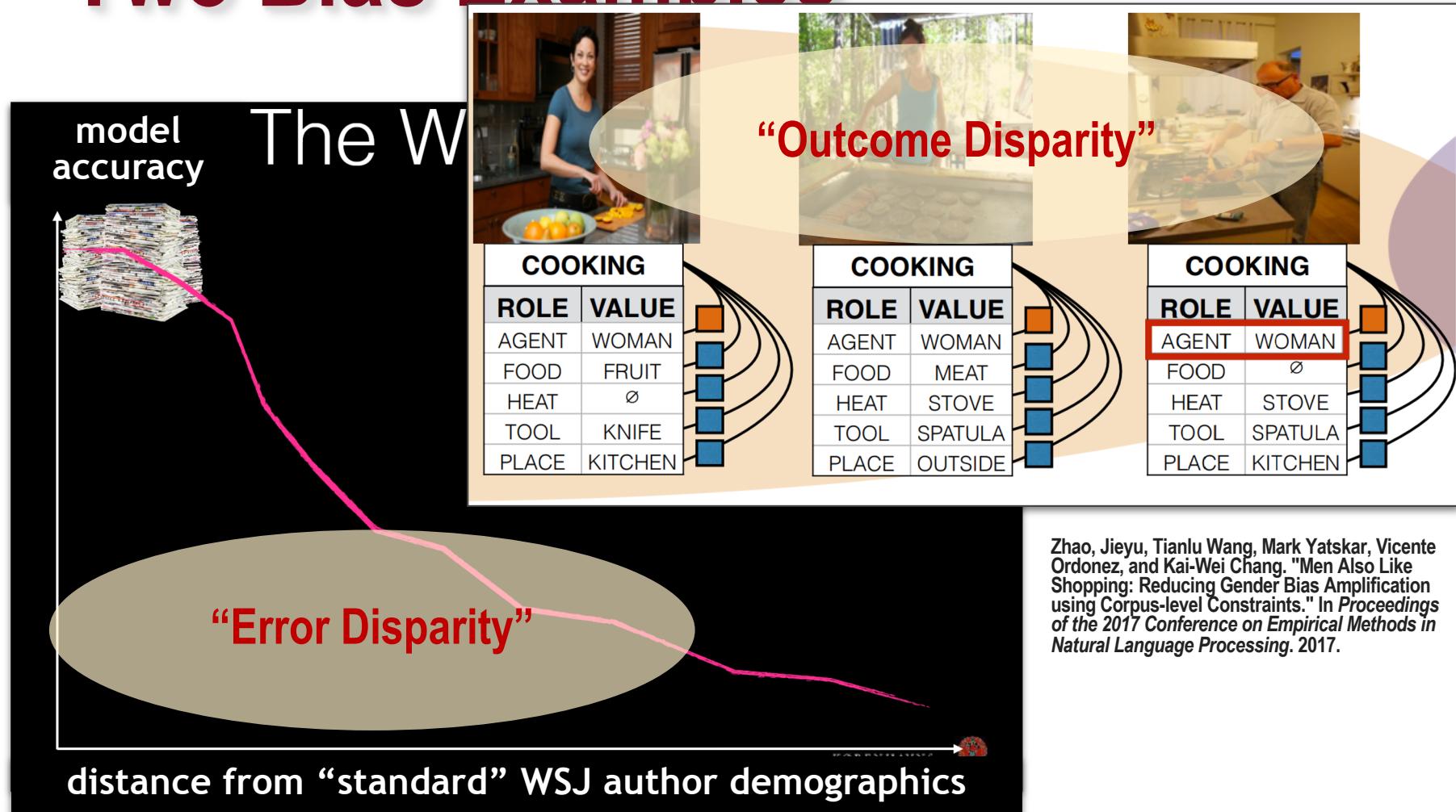
Two Bias Examples



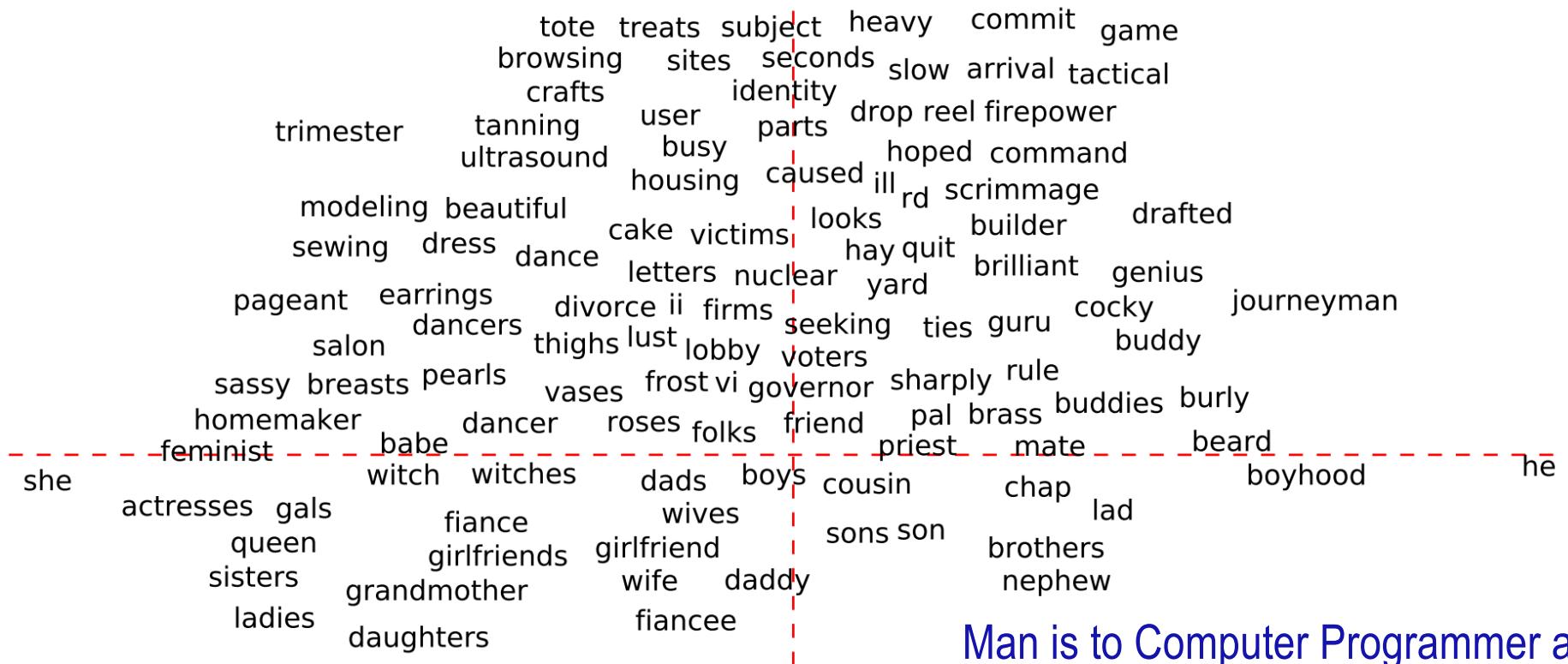
Two Bias Examples



Two Bias Examples



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

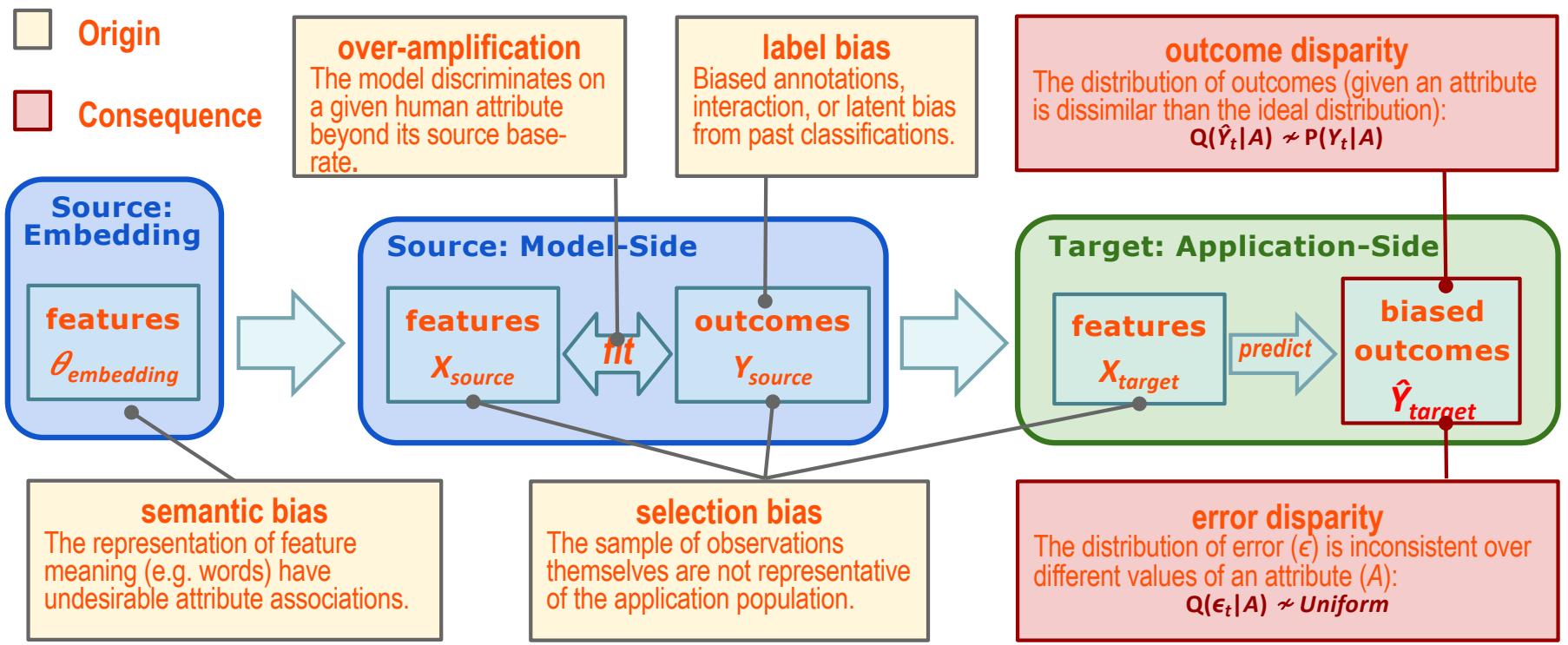
◆ Bias amplification

- 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

The screenshot shows the textio AI writing tool interface. At the top, there's a navigation bar with 'textio' logo, 'New', 'Import', 'Export', 'Link', 'Delete', 'Undo', and 'History' buttons. Below the navigation is a title 'Title of your job listing' and a subtitle 'Job listing for an **unknown** role in **an unknown location**'. To the right of the text area are 'Draft' and 'Share' buttons. The main text area contains a quote: "'Exceptional programmer sought. **Successful candidates will** thrive in our **fast** paced environment. **You must be able to work under pressure.**'". A green callout box to the right of the word 'fast' says 'Fewer job seekers will apply if you use this phrase. Instead, you could try: **dynamic**'. Another callout box below it says 'This phrase draws more male job seekers. Other choices: energizing environment, exciting environment, rapidly changing environment'. At the bottom, there are five colored buttons: 'Negative' (red), 'Positive' (green), 'Repetitive' (grey), 'Masculine' (blue), and 'Feminine' (purple).

Fewer job seekers will apply if you use this phrase.
Instead, you could try:
dynamic

This phrase draws more male job seekers.
Other choices:
energizing environment
exciting environment
rapidly changing environment

Negative Positive Repetitive Masculine Feminine

What you should know

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