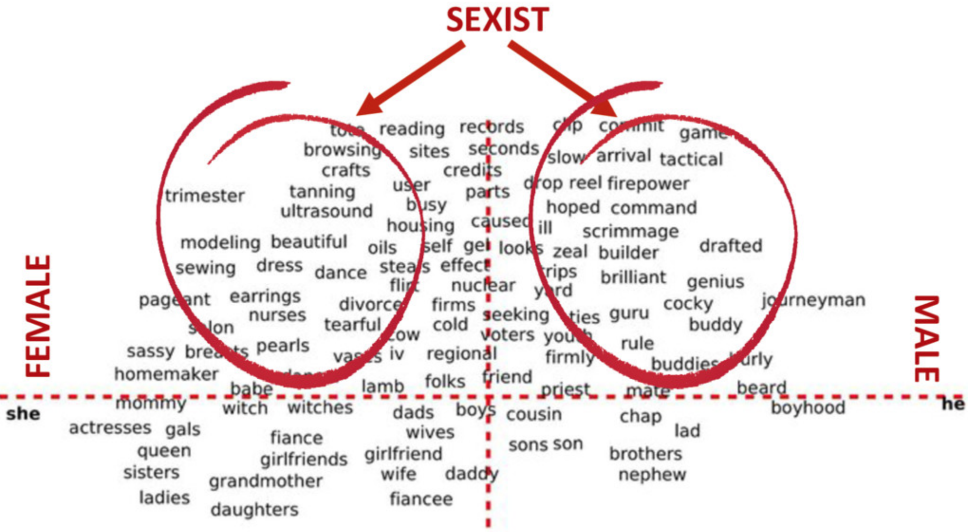
Fairness, Accountability, and Transparency

Machine Learning: Jordan Boyd-Graber
University of Maryland
BIASED REPRESENTATIONS

Slides/ideas adapted from Adam Tauman Kalai and Moritz Hardt

Our data reflect our world . . .

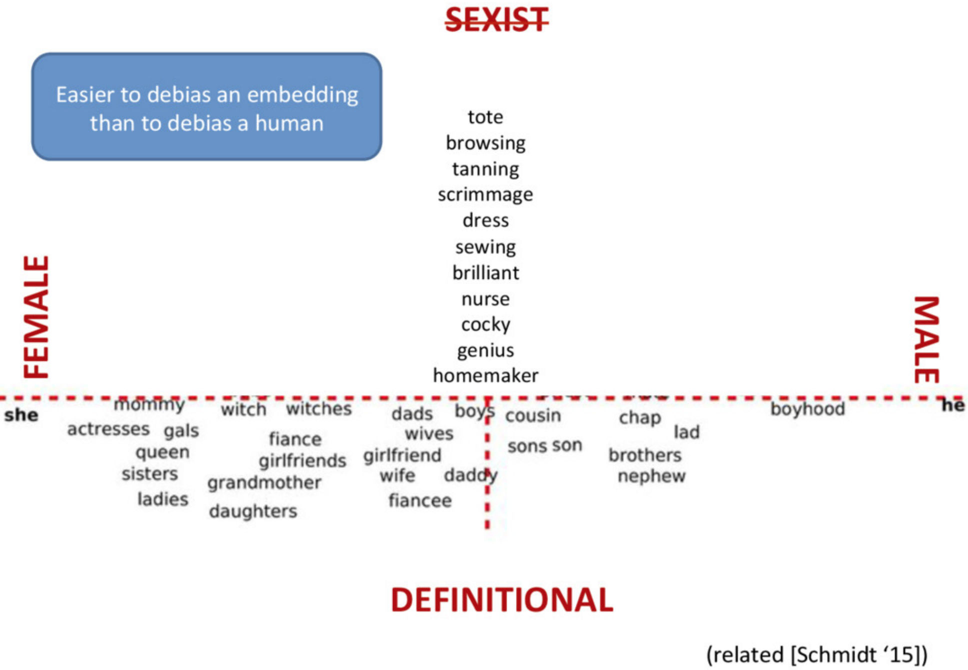
- Word representations learned from massive amounts of data
- Reflect prejudices and messiness of our world
- But learned representations used for many tasks
  - Detecting “bad” behavior online
  - Matching resumes to jobs
  - Recommendations



DEFINITIONAL

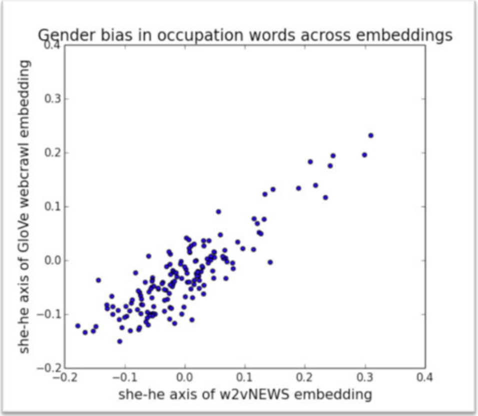
The embedding captures gender stereotypes *and* sexism.

(related [Schmidt '15])


~~SEXIST~~
Easier to debias an embedding than to debias a human
tote
browsing
tanning
scrimmage
dress
sewing
brilliant
nurse
cocky
genius
homemaker
FEMALE
MALE
she
mommy
actresses
gals
queen
sisters
ladies
witch
witches
fiance
girlfriends
grandmother
daughters
dads
wives
girlfriend
wife
fiancee
boys
daddy
cousin
sons son
chap
lad
brothers
nephew
boyhood
he
DEFINITIONAL
(related [Schmidt '15])

Consistency of embedding stereotype

GloVe trained on web crawl



Each dot is an occupation; Spearman = 0.8

word2vec trained on Google news

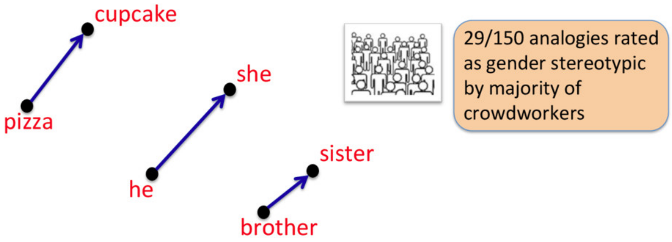
Doesn't matter source or algorithm

Bias encoded in some dimensions

Analogies

he:x::she:y

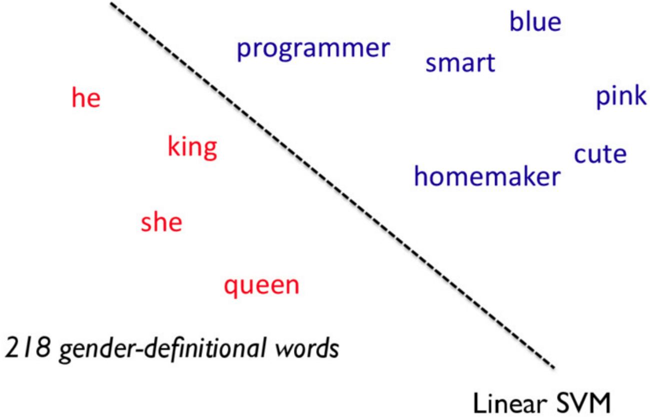
$$\min \cos(\text{he} - \text{sh}, x - y) \text{s.t.} ||x - y||_2 < \delta \tag{1}$$



$$\min \cos(\text{he} - \text{she}, \text{x} - \text{y}) \text{ such that } ||\text{x} - \text{y}||_2 < \delta$$

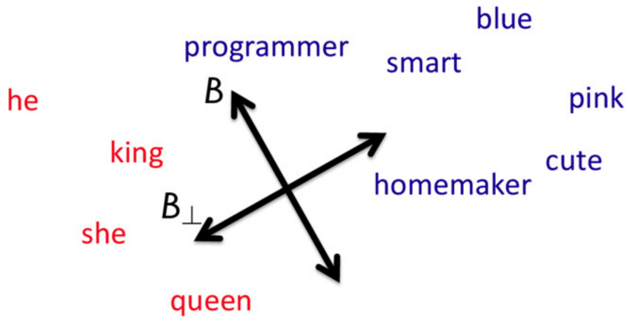
Bias Where it Shouldn't Be

Debiasing



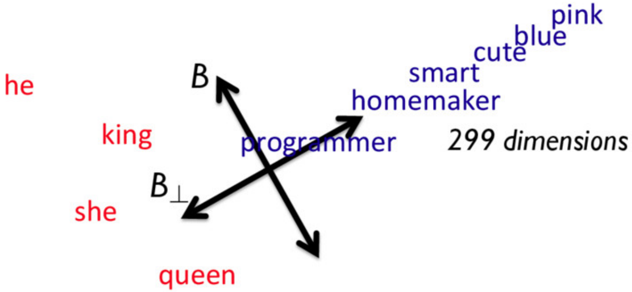
blue
programmer
smart
he
pink
king
cute
homemaker
she
queen
218 gender-definitional words
Linear SVM

Debiasing



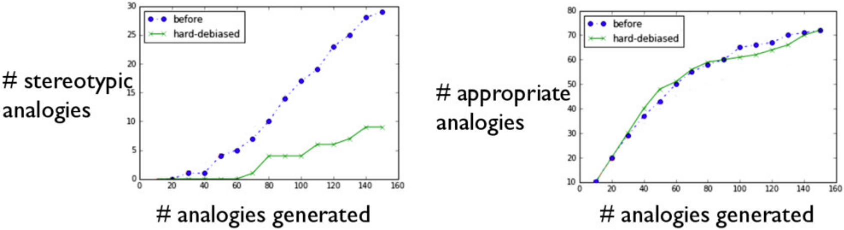
blue
programmer
smart
B
he
pink
king
cute
homemaker
$B_\perp$
she
queen

Debiasing



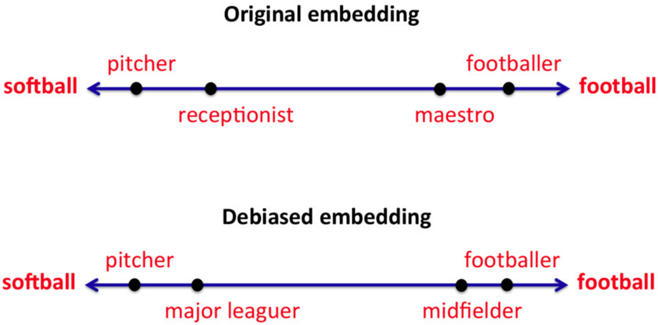
pink
blue
cute
smart
homemaker
he
king
programmer
299 dimensions
she
queen
B
$B_\perp$

Debiasing



stereotypic analogies
before
hard-debiased
analogies generated
appropriate analogies
before
hard-debiased
analogies generated

Debiasing


Original embedding
pitcher
footballer
softball
football
receptionist
maestro
Debiased embedding
pitcher
footballer
softball
football
major leaguer
midfielder

Data are biased . . .

- Our data (societies) are biased
- Can we make algorithms better than the data?
- Can we define fairness for tasks like sentencing, loan approval, etc.

Defining Fairness

What does non-discriminatory mean?

Target y, predictor $\hat{y}$ from features x and protected attribute a.

- Don't want to remove a
- Don't want parity $(p(\hat{y} \,|\, A = a) = p(\hat{y} \,|\, A = a')$

Defining Fairness

What does non-discriminatory mean?

Target y, predictor $\hat{y}$ from features x and protected attribute a.

- Don't want to remove a (correlations, accuracy disparity)
- Don't want parity $(p(\hat{y} \,|\, A=a) = p(\hat{y} \,|\, A=a')$

Defining Fairness

What does non-discriminatory mean?

Target y, predictor $\hat{y}$ from features x and protected attribute a.

- Don't want to remove a (correlations, accuracy disparity)
- Don't want parity $(p(\hat{y} \mid A=a) = p(\hat{y} \mid A=a')$ (doesn't allow perfect prediction)

  Also, can have accuracy disparity: give loans to qualified $A=0$ and random $A=1$

Defining Fairness

What does non-discriminatory mean?

Target y, predictor $\hat{y}$ from features x and protected attribute a.

- Don't want to remove a (correlations, accuracy disparity)
- Don't want parity $(p(\hat{y} \,|\, A = a) = p(\hat{y} \,|\, A = a')$ (doesn't allow perfect prediction)
- Equalized odds:

$$p(\hat{y} \,|\, Y = y, A = a) = P(\hat{y} \,|\, Y = y, A = a') \tag{2}$$

  - Perfect predictor always satisfies
  - Protects against accuracy disparity