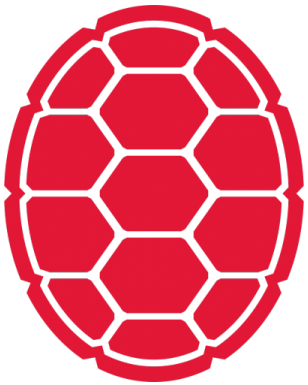


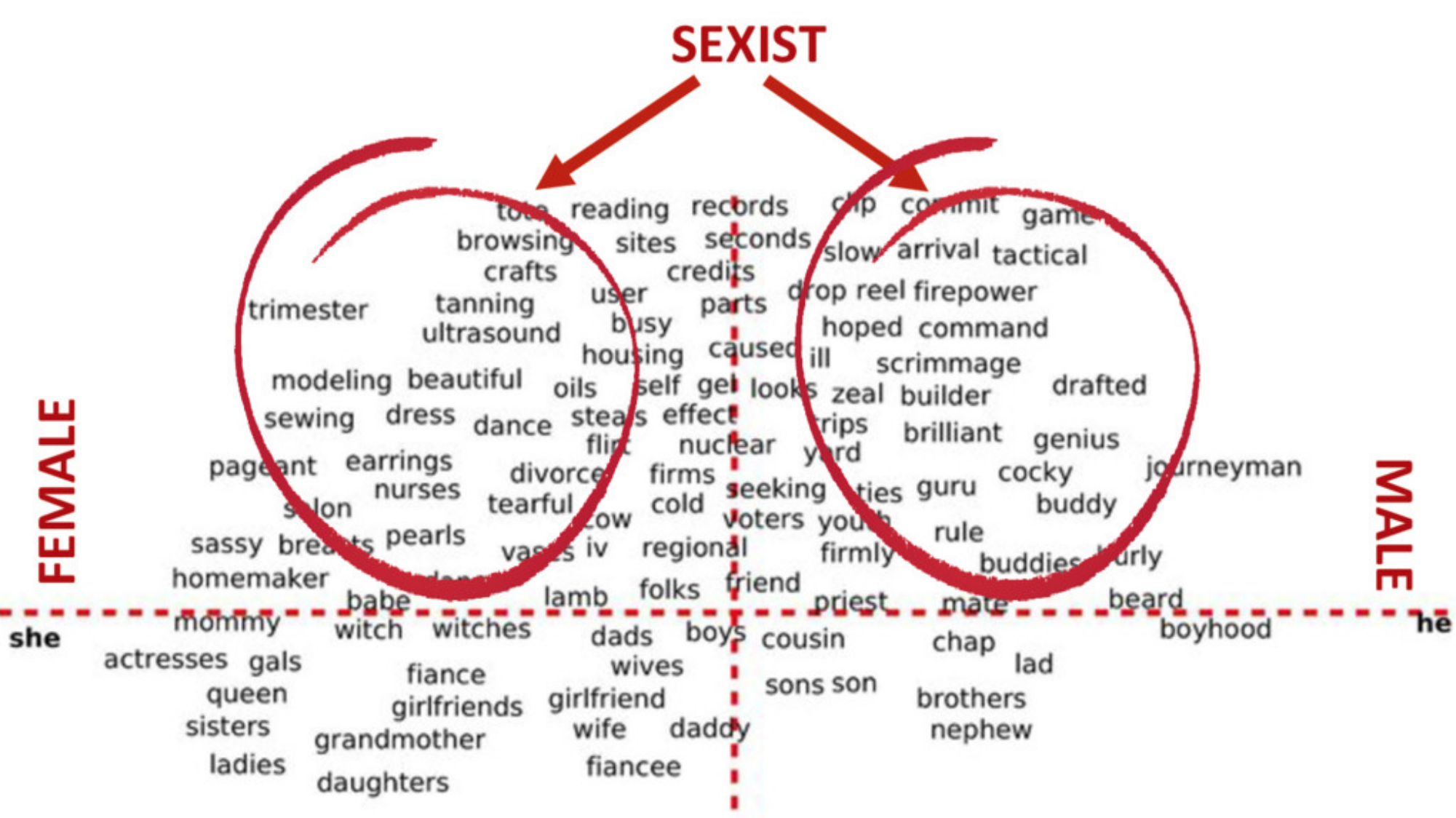
Fairness, Accountability, and Transparency

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

Slides/ideas adapted from Adam Tauman Kalai and Moritz Hardt

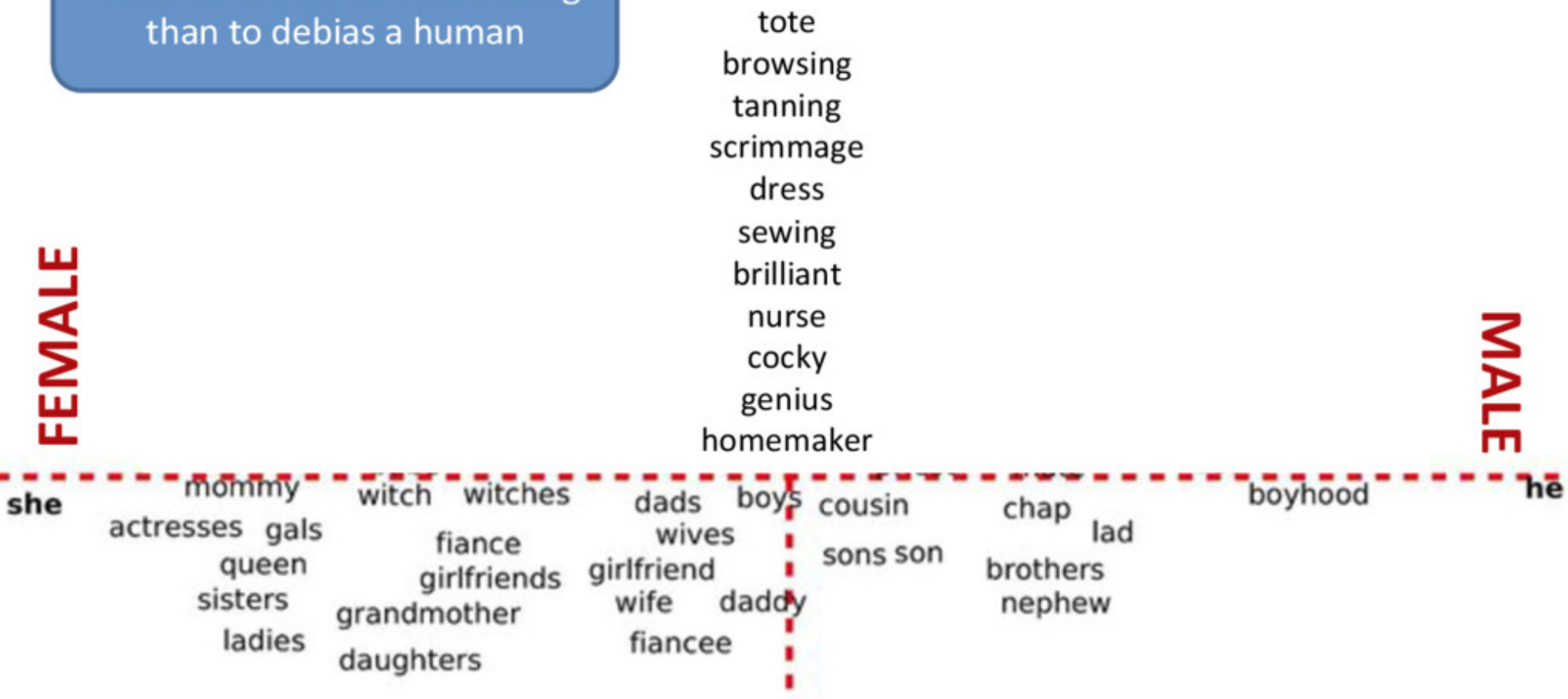
Our data reflect our world . . .

- 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


SEXIST
FEMALE
MALE
she
he
reading
records
browsing
sites
seconds
crafts
credits
trimester
tanning
user
parts
ultrasound
busy
housing
caused
modeling
beautiful
oils
self
get
looks
sewing
dress
dance
steals
effect
flirt
nuclear
pageant
earrings
divorce
firms
nurses
seeking
tearful
cold
salon
pearls
regional
sassy
homemaker
babe
lamb
folks
friend
mommy
witch
witches
dads
boys
cousin
actresses
gals
wives
fiance
queen
girlfriends
girlfriend
sisters
wife
daddy
grandmother
ladies
fiancee
daughters
arrival
tactical
drop
reel
firepower
hoped
command
ill
scrimmage
zeal
builder
drafted
brilliant
genius
cocky
guru
buddy
rule
firmly
buddies
priest
mate
beard
boyhood
chap
lad
sons
son
brothers
nephew
DEFINITIONAL
The embedding captures gender stereotypes *and* sexism.
(related [Schmidt '15])

~~SEXIST~~

Easier to debias an embedding than to debias a human

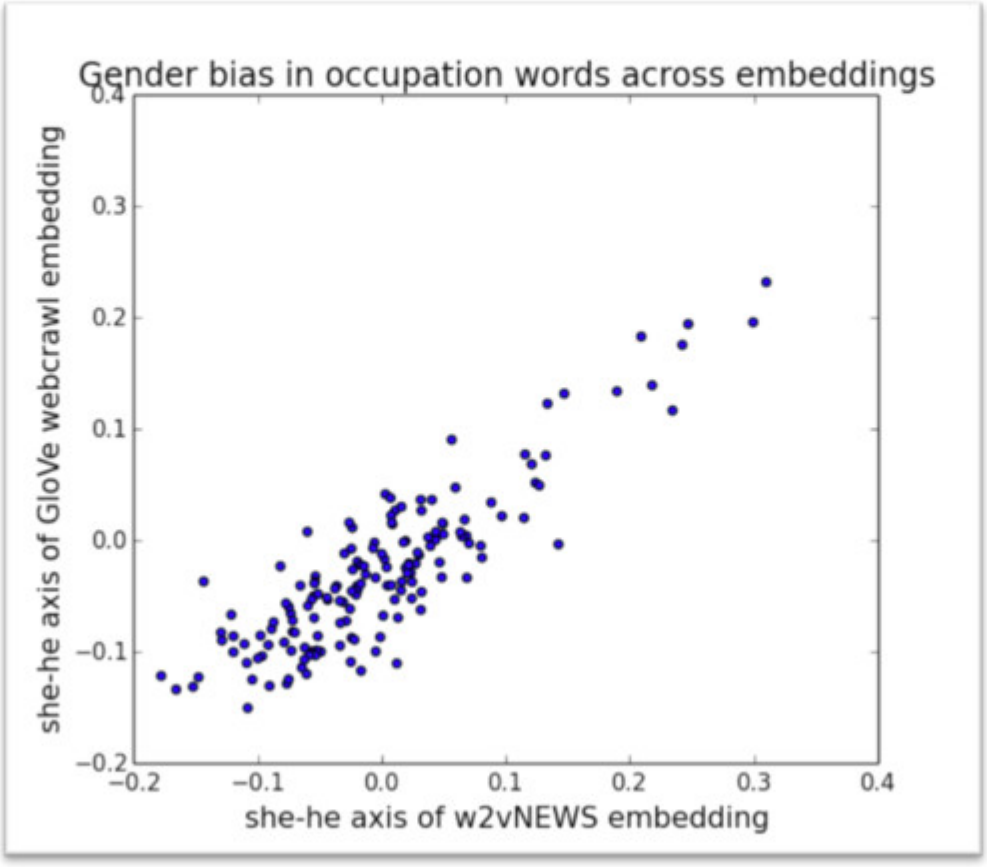


DEFINITIONAL

(related [Schmidt '15])

Consistency of embedding stereotype

GloVe trained on web crawl



word2vec trained on Google news

Each dot is an occupation; Spearman = 0.8

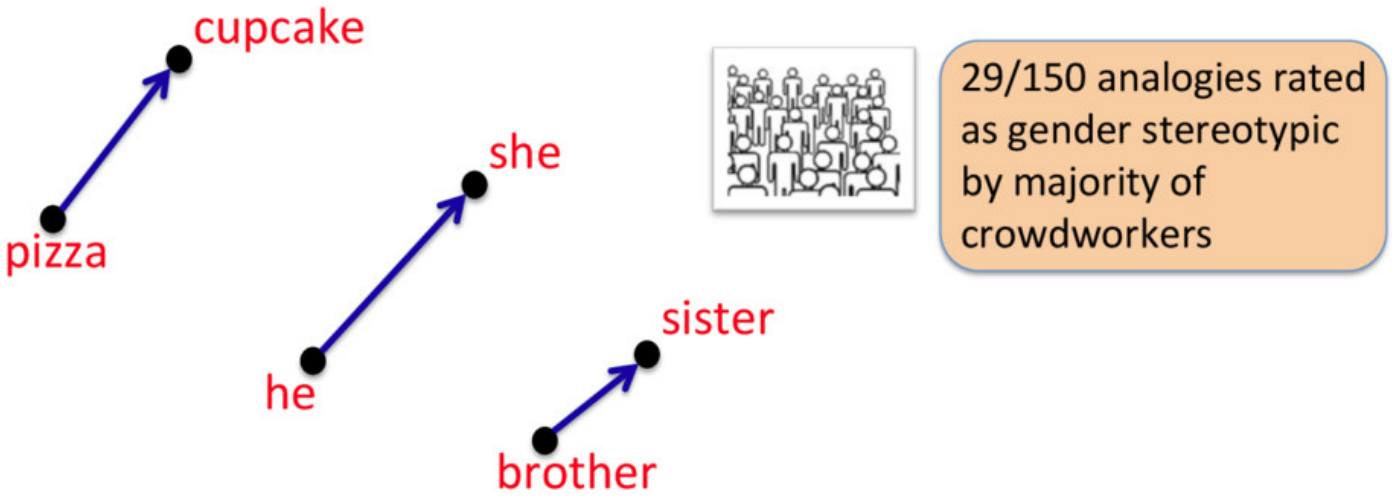
Doesn’t matter source or algorithm

Bias encoded in some dimensions

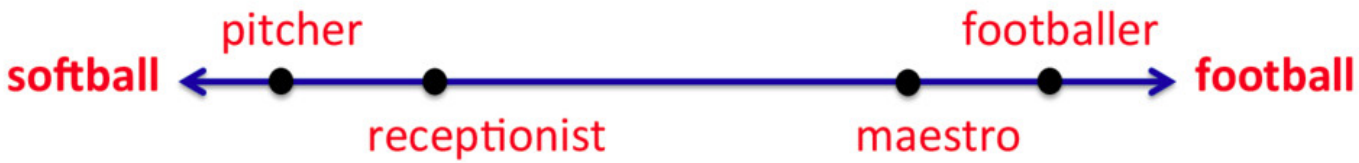
Analogies

he:x::she:y

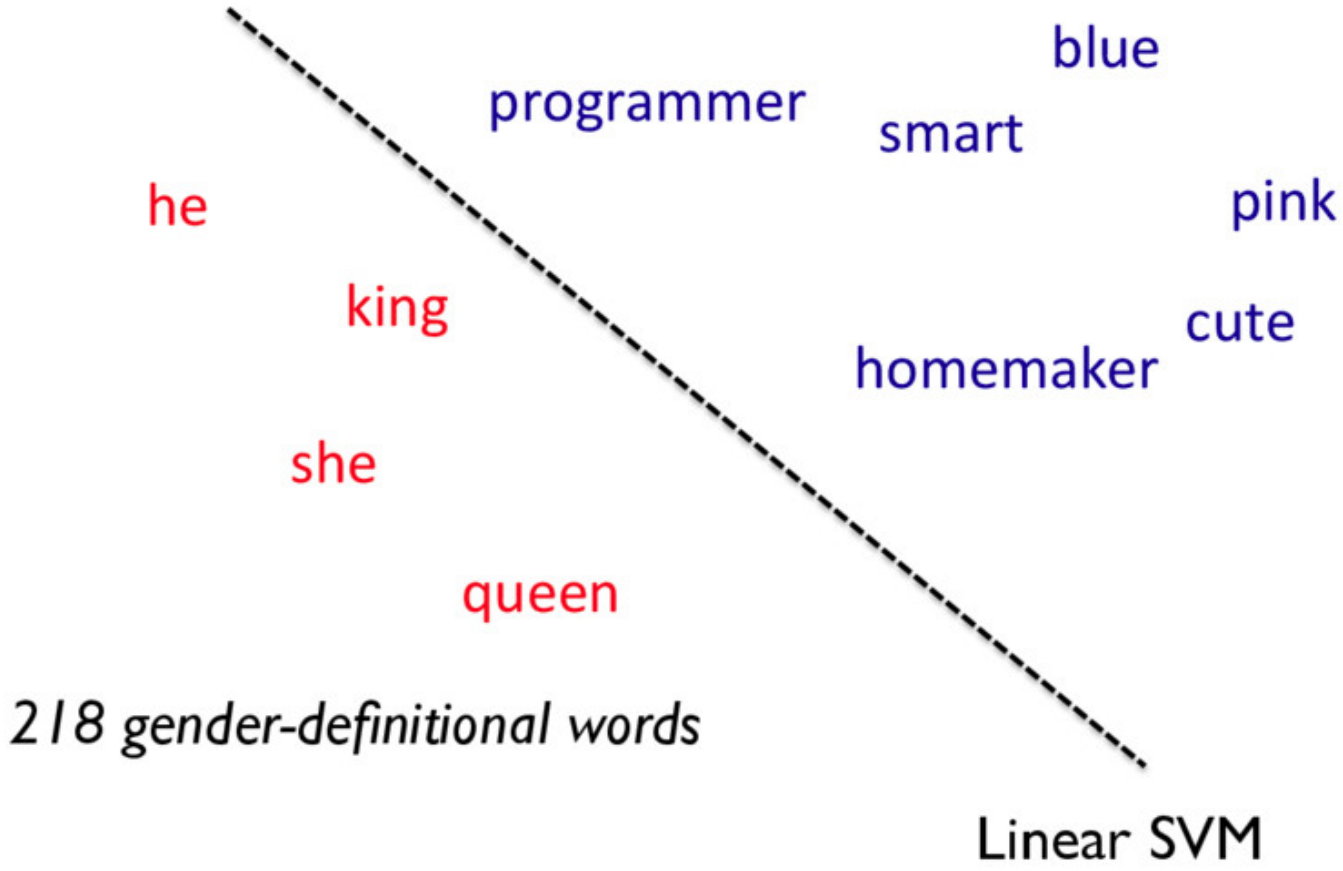
$$\min \cos(\text{he} - \text{sh}, x - y) \text{s.t.} ||x - y||_2 < \delta \quad (1)$$



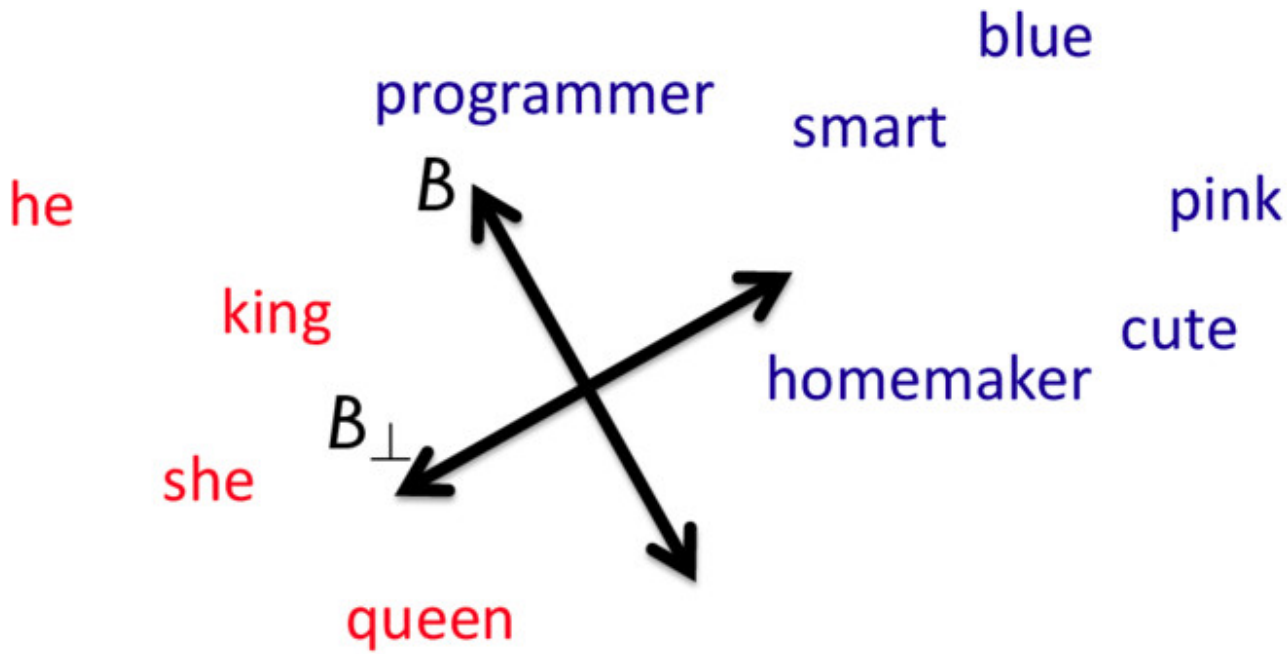
Bias Where it Shouldn't Be



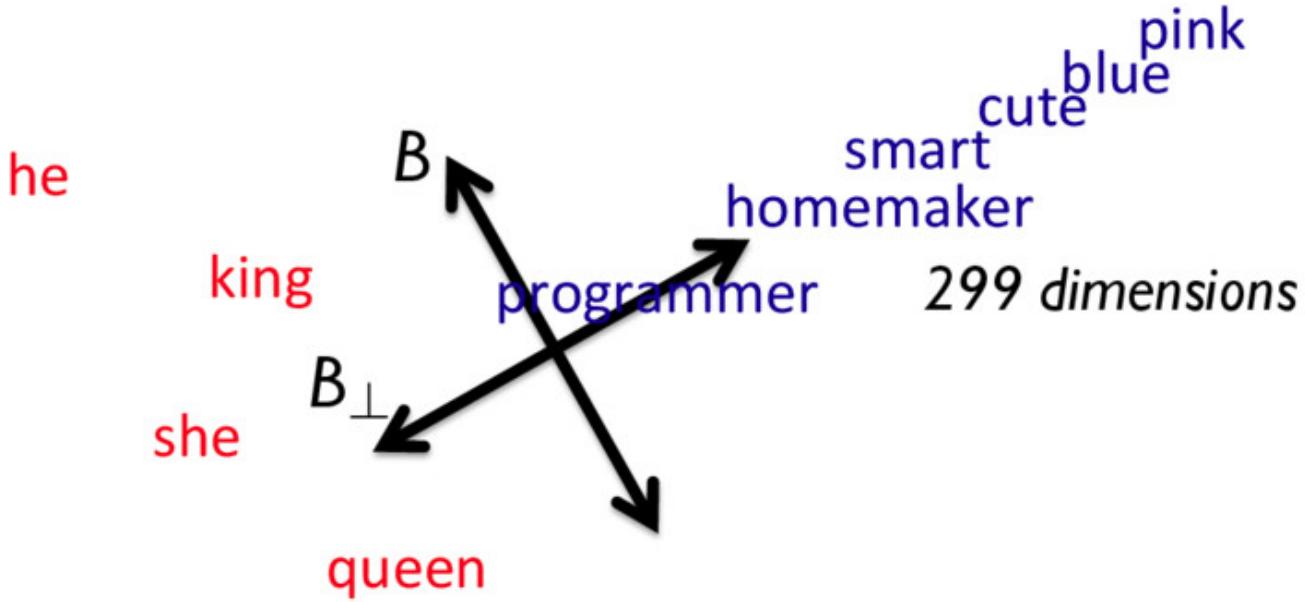
Debiasing


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

Debiasing

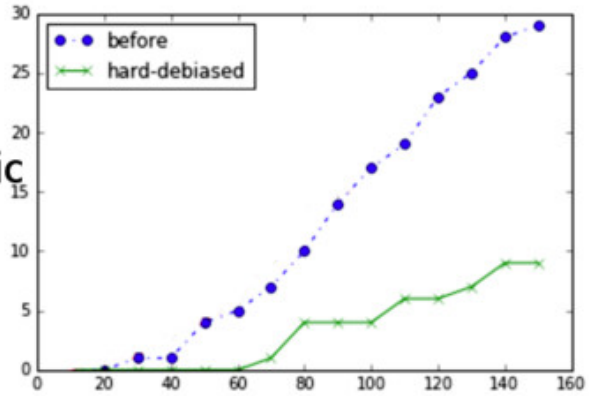

blue
programmer
smart
he
B
pink
king
cute
homemaker
B⊥
she
queen

Debiasing


pink
blue
cute
smart
homemaker
programmer
299 dimensions
B
B⊥
he
king
she
queen

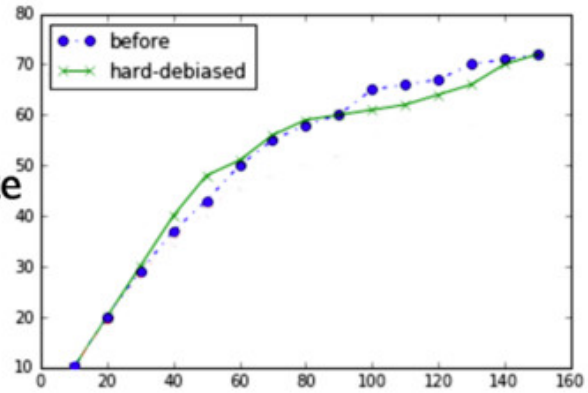
Debiasing

stereotypic analogies



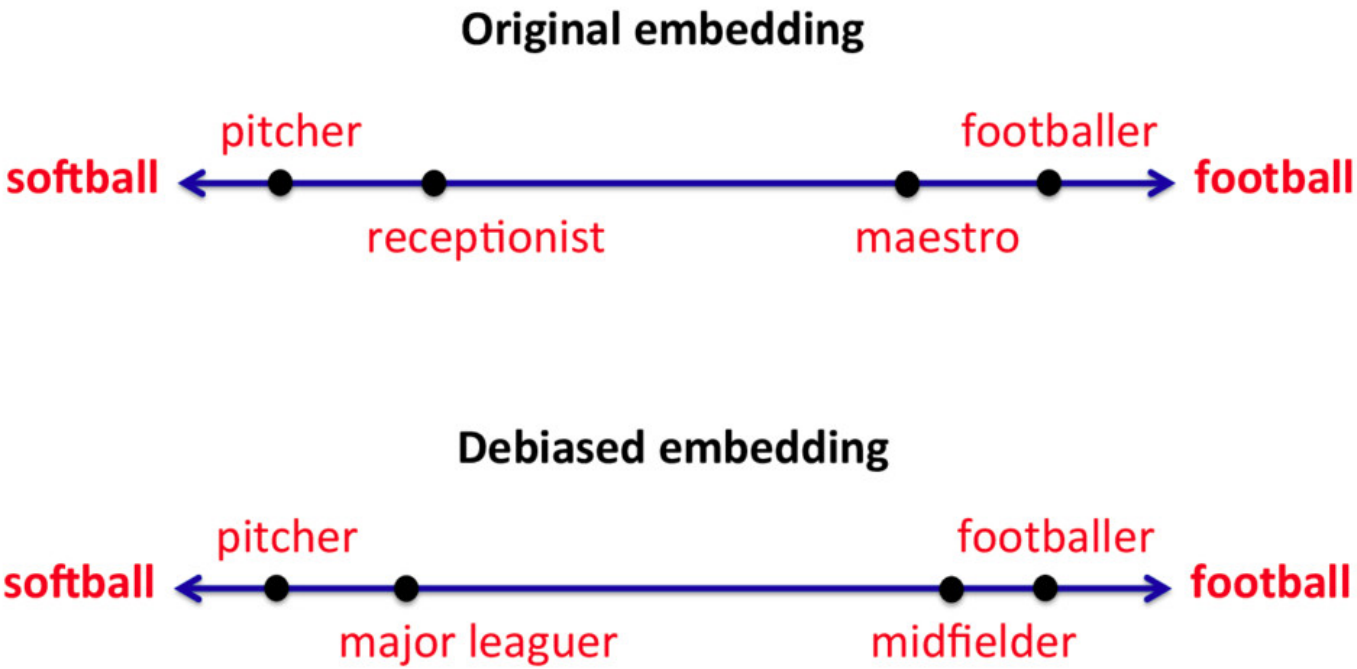
analogies generated

appropriate analogies



analogies generated

Debiasing


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

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 ($p(\hat{y} \mid A = a) = p(\hat{y} \mid A = a')$

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 ($p(\hat{y} \mid A = a) = p(\hat{y} \mid A = a')$)

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 ($p(\hat{y} \mid A = a) = p(\hat{y} \mid A = a')$ (doesn't allow perfect prediction)
- Equalized odds:

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

  - Perfect predictor always satisfies
  - Protects against accuracy disparity

Fairness, Accountability, and Transparency

- Like much of machine learning, we have problems and no clear solutions
- What I've presented here are just first steps
- The important thing is to think about data, algorithms, and employing them in a way that thinks through consequences
- Don't blindly trust algorithms / data