



Fairness, Accountability, and Transparency

Machine Learning: Jordan Boyd-Graber
University of Maryland

BIASED REPRESENTATIONS

SEXIST

FEMALE

MALE

she

he

total reading records clip commit game
browsing sites seconds slow arrival tactical
crafts user credits drop reel firepower
trimester tanning busy parts hoped command
ultrasound housing caused ill scrimmage
modeling beautiful oils self gel looks zeal builder drafted
sewing dress dance steals effect trips brilliant genius
pageant earrings divorce flint nuclear yard cocks journeyman
salon nurses tearful low cold voters youth buddy
sassy breasts pearls vases iv regional firmly rule buddies curly
homemaker babe lamb folks friend priest mate beard
mommy witch witches dads boys cousin chap lad
actresses gals wives sons son brothers
queen girlfriends girlfriend wife daddy nephew
sisters grandmother fiancée
ladies daughters

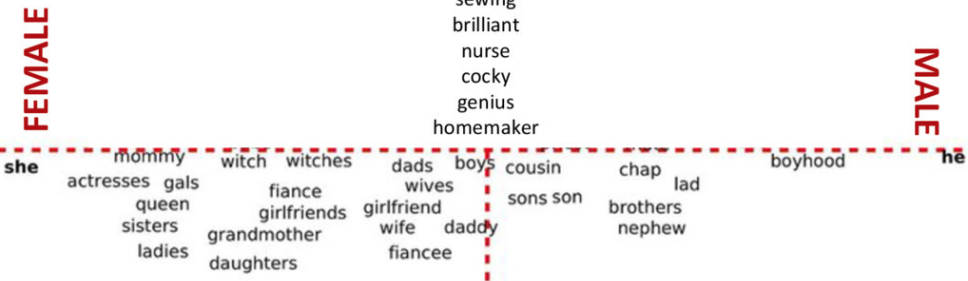
The embedding captures
gender stereotypes *and* sexism.

DEFINITIONAL

(related [Schmidt '15])

SEXIST

Easier to debias an embedding
than to debias a human

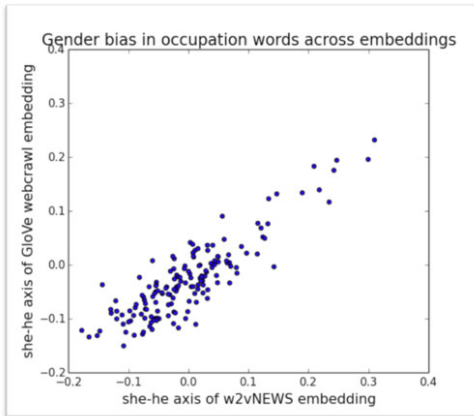


DEFINITIONAL

(related [Schmidt '15])

Consistency of embedding stereotype

GloVe trained
on web crawl



Each dot is an
occupation;
Spearman = 0.8

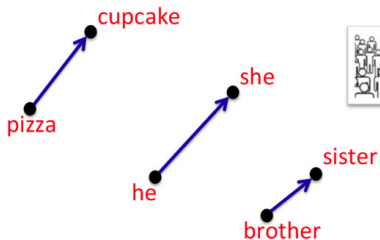
word2vec trained on Google news

Doesn't matter source or algorithm

Bias encoded in some dimensions



Analogies



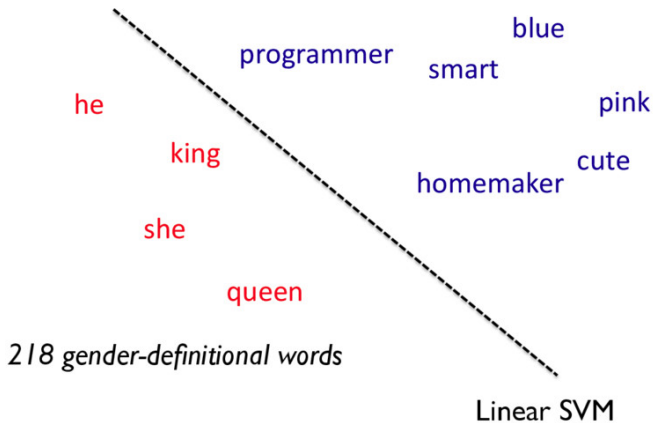
29/150 analogies rated
as gender stereotypic
by majority of
crowdworkers

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

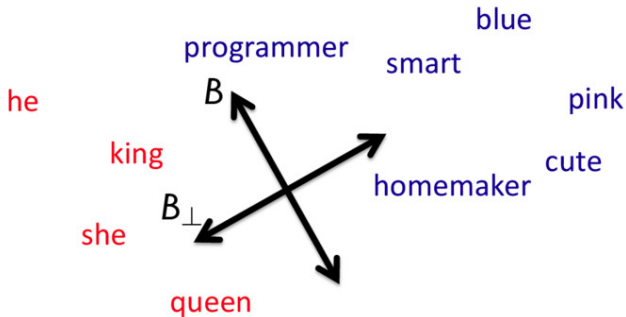
Bias Where it Shouldn't Be



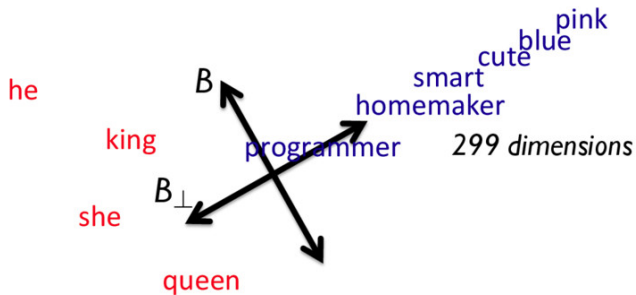
Debiasing



Debiasing

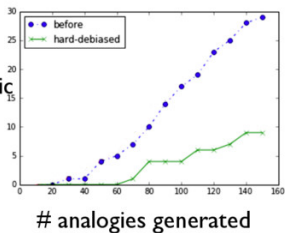


Debiasing

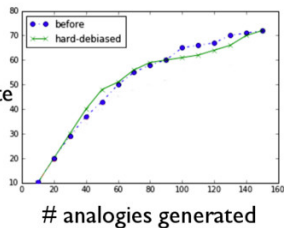


Debiasing

stereotypic
analogies



appropriate
analogies



Debiasing

