Attenuating Bias in Word Embeddings

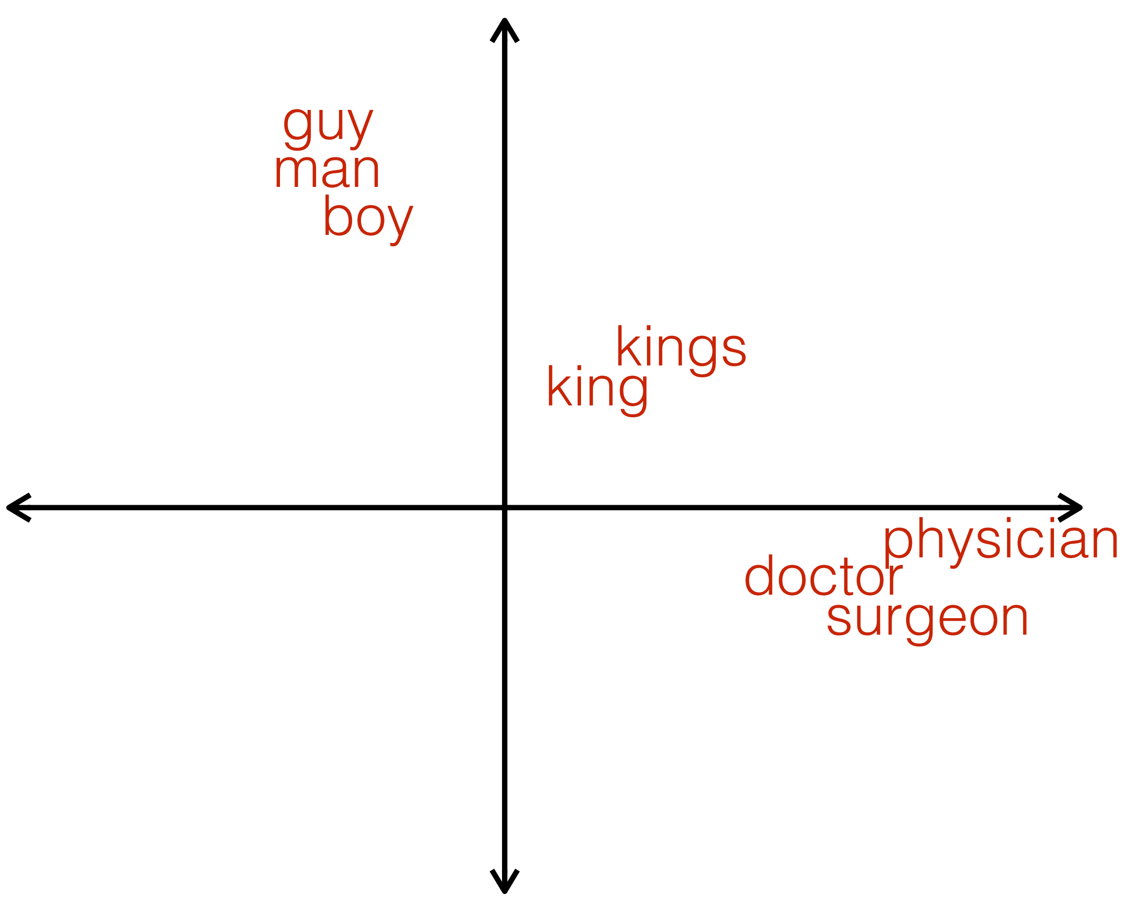
Sunipa Dev (sunipad@cs.utah.edu)
Jeff M Phillips (jeffp@cs.utah.edu)
University of Utah

April 17th
AISTATS 2019

Word Embeddings

One hot vectors or bag of words embeddings

- 100K (sometimes more) dimensional vectors
- sparse but inefficient
- contain useful semantic and syntactic information



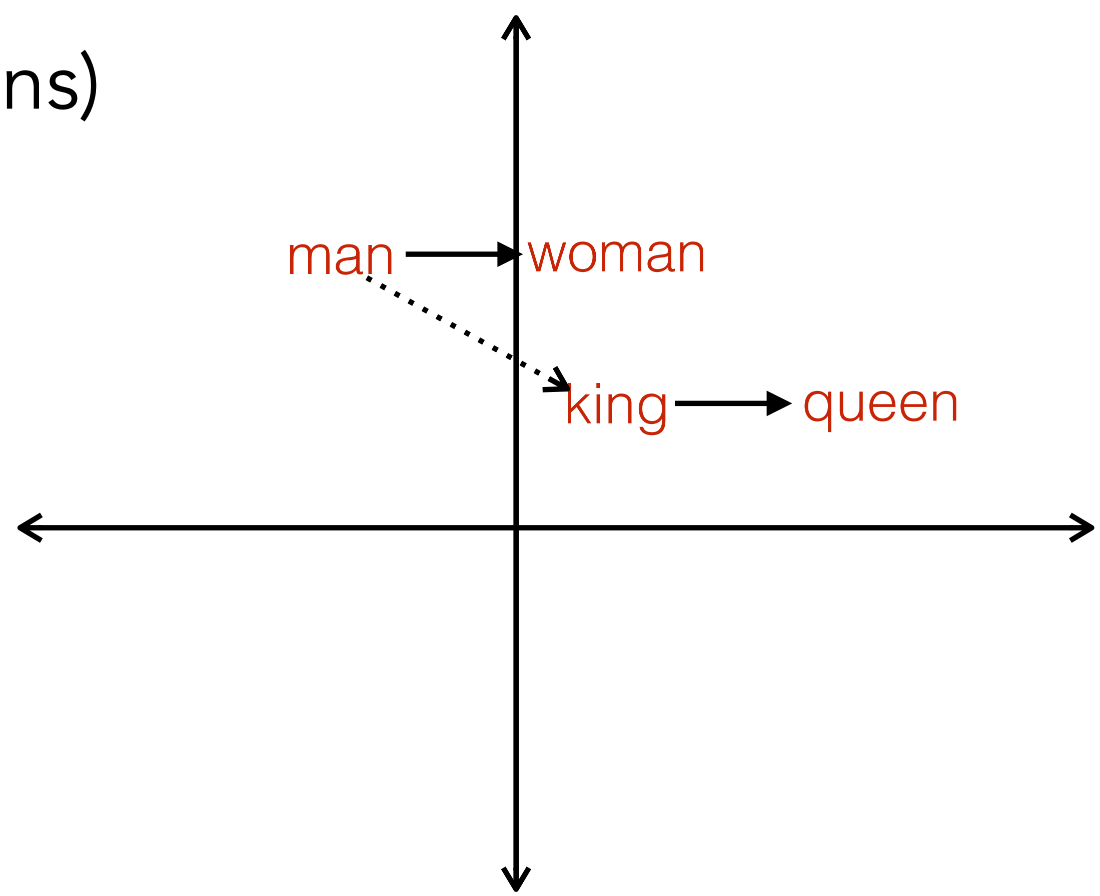
Word Embeddings

One hot vectors or bag of words embeddings

- 100K (sometimes more) dimensional vectors
- sparse but inefficient
- contain useful semantic and syntactic information

Word2Vec, GloVe, FastText

- distributed representations
- low dimensional (about ~300 dimensions)
- additional useful linear relationships



Word Embeddings

One hot vectors or bag of words embeddings

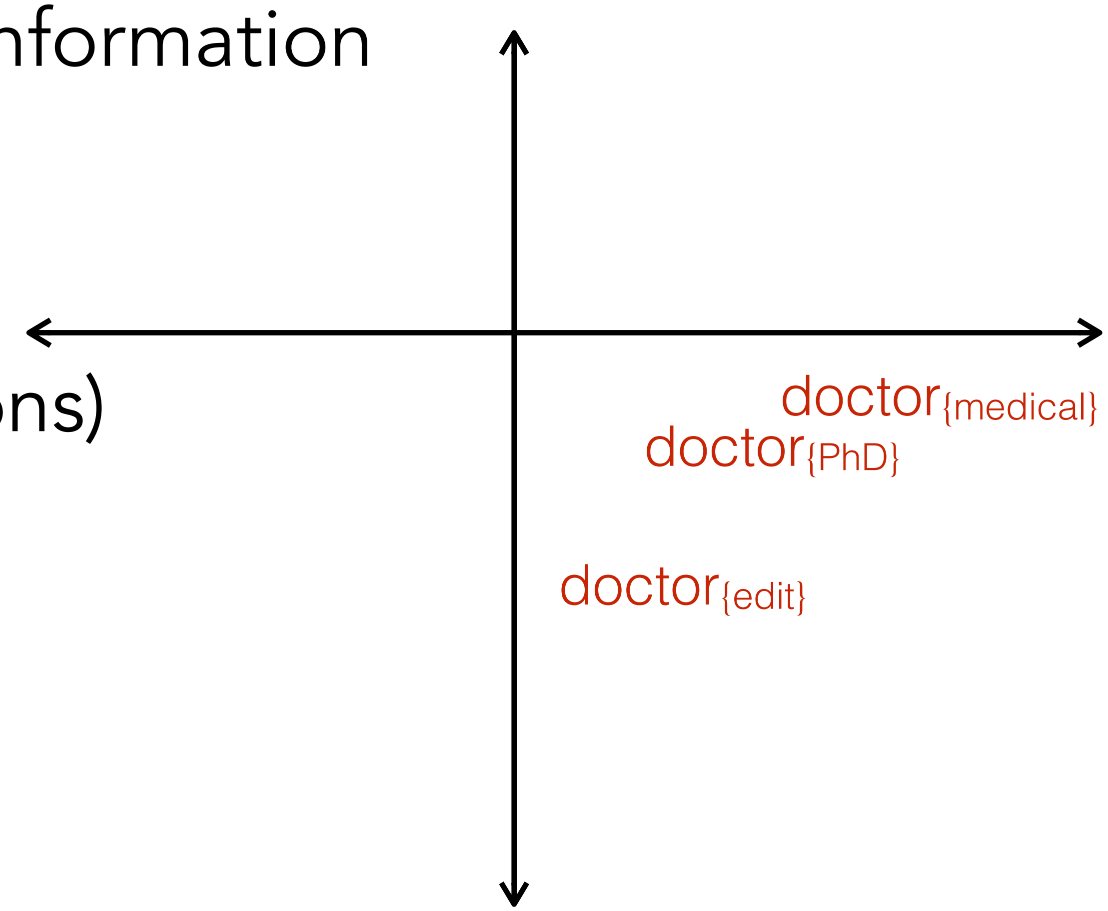
- 100K (sometimes more) dimensional vectors
- sparse but inefficient
- contain useful semantic and syntactic information

Word2Vec, GloVe, FastText

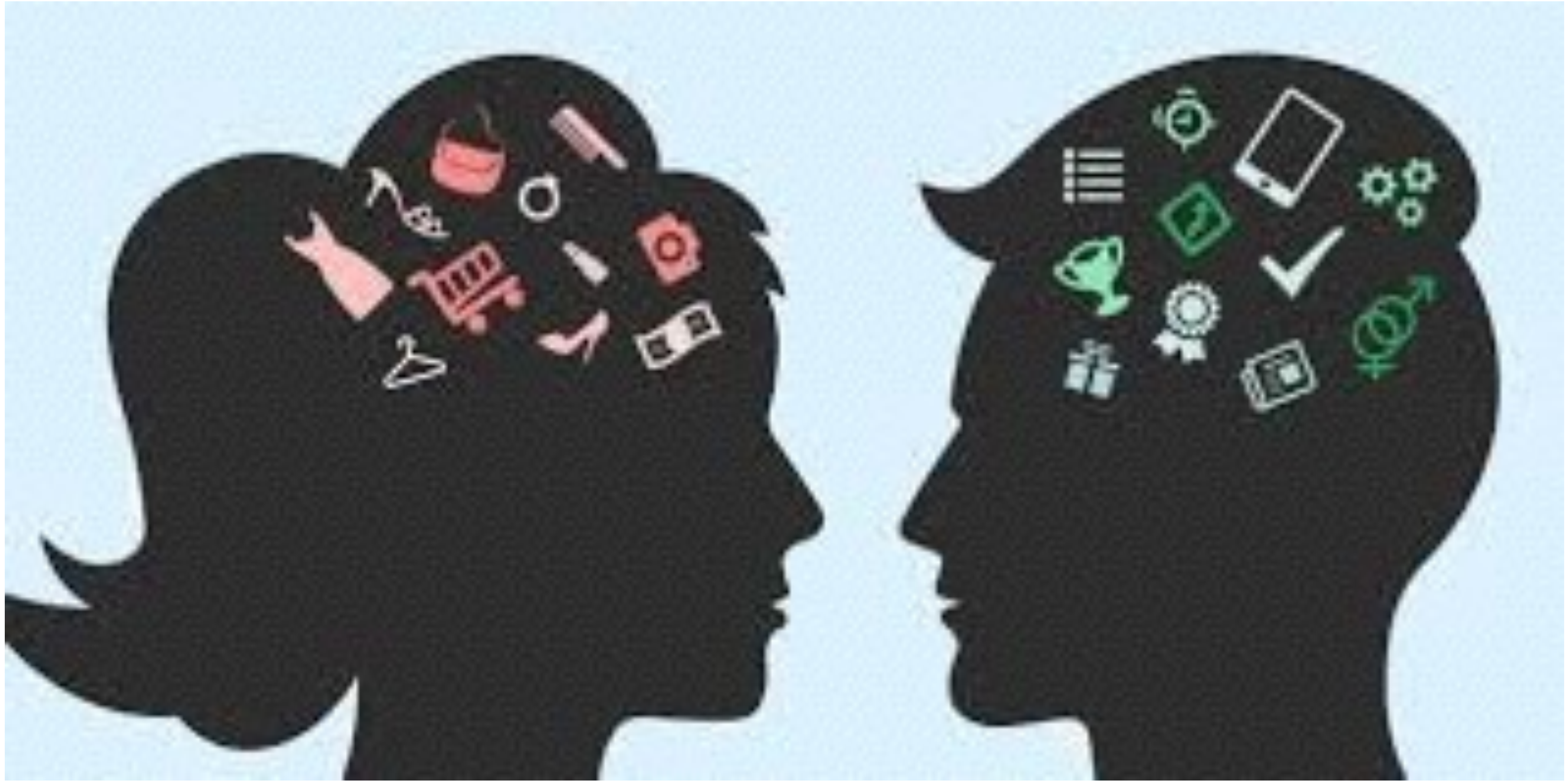
- distributed representations
- low dimensional (about ~300 dimensions)
- additional useful linear relationships

ELMo, BERT

- context sensitive embeddings
- distinguishable word senses
- dimensionality still low at about 3000 (3*1024) dimensions



Bias



Preferential association of words, topics with stereotypical connotations to word groups or names representing protected population characteristics such as gender, race, age or sexuality.

Gendered Words

Occupations

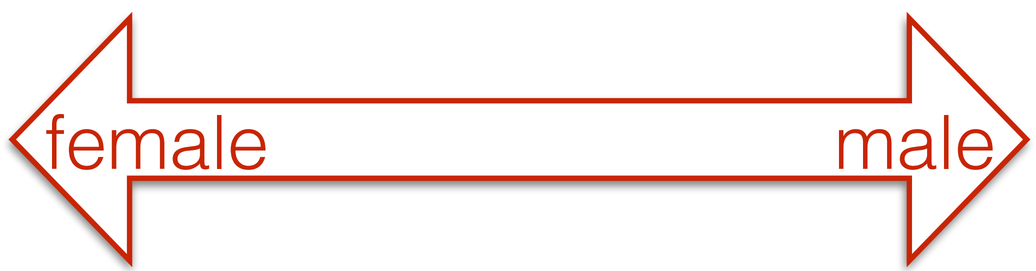
nurse
maid
housewife
prostitute

Adjectives

glamorous
diva
shimmery
beautiful

Other words

miss, maid,
motherhood,
herself, seductive,
heroine, herself



Occupations

soldier
captain
officer
footballer

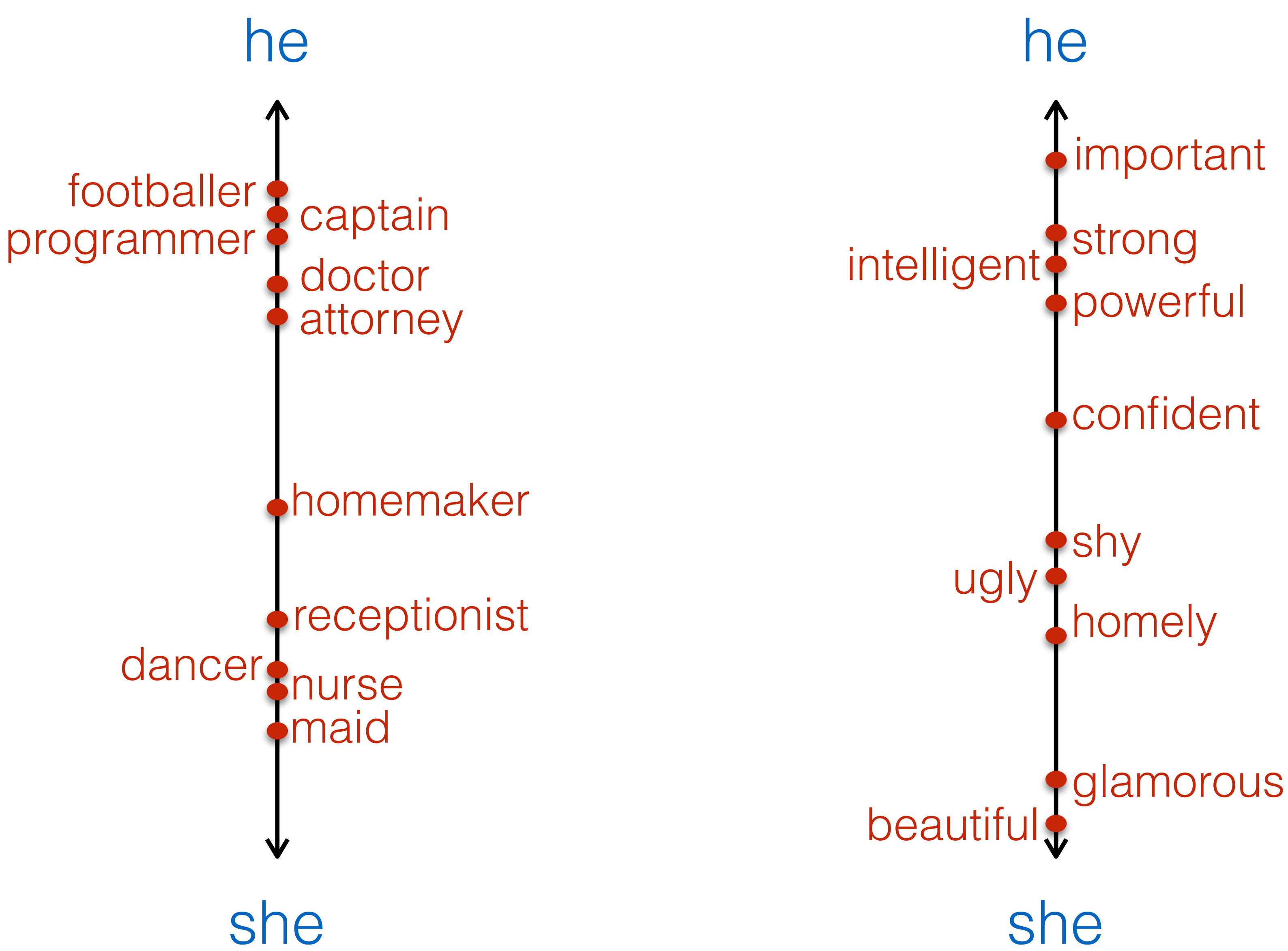
Adjectives

strong
muscular
powerful
fast

Other words

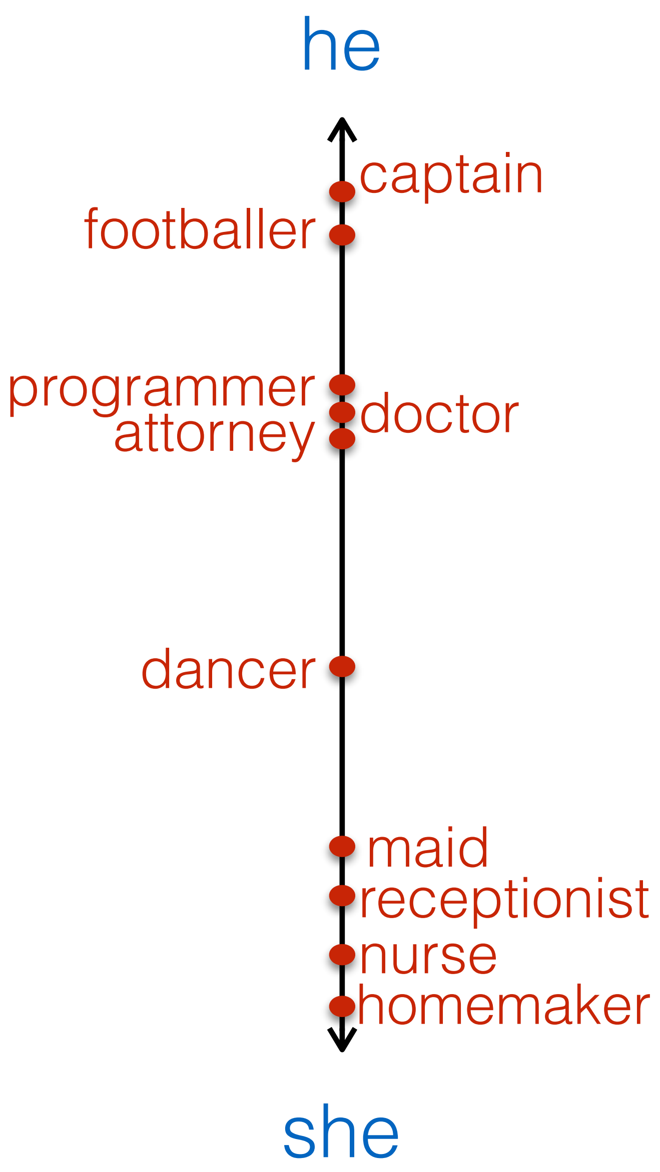
himself, sir,
congressman,
succeeded, him,
forefather, nephew

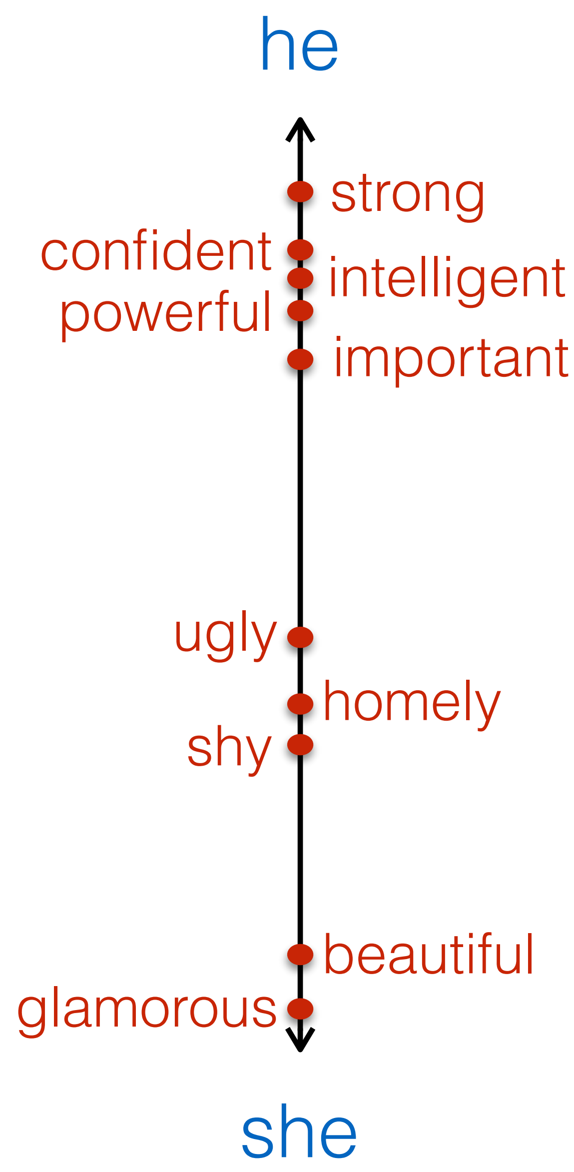
Gender Bias



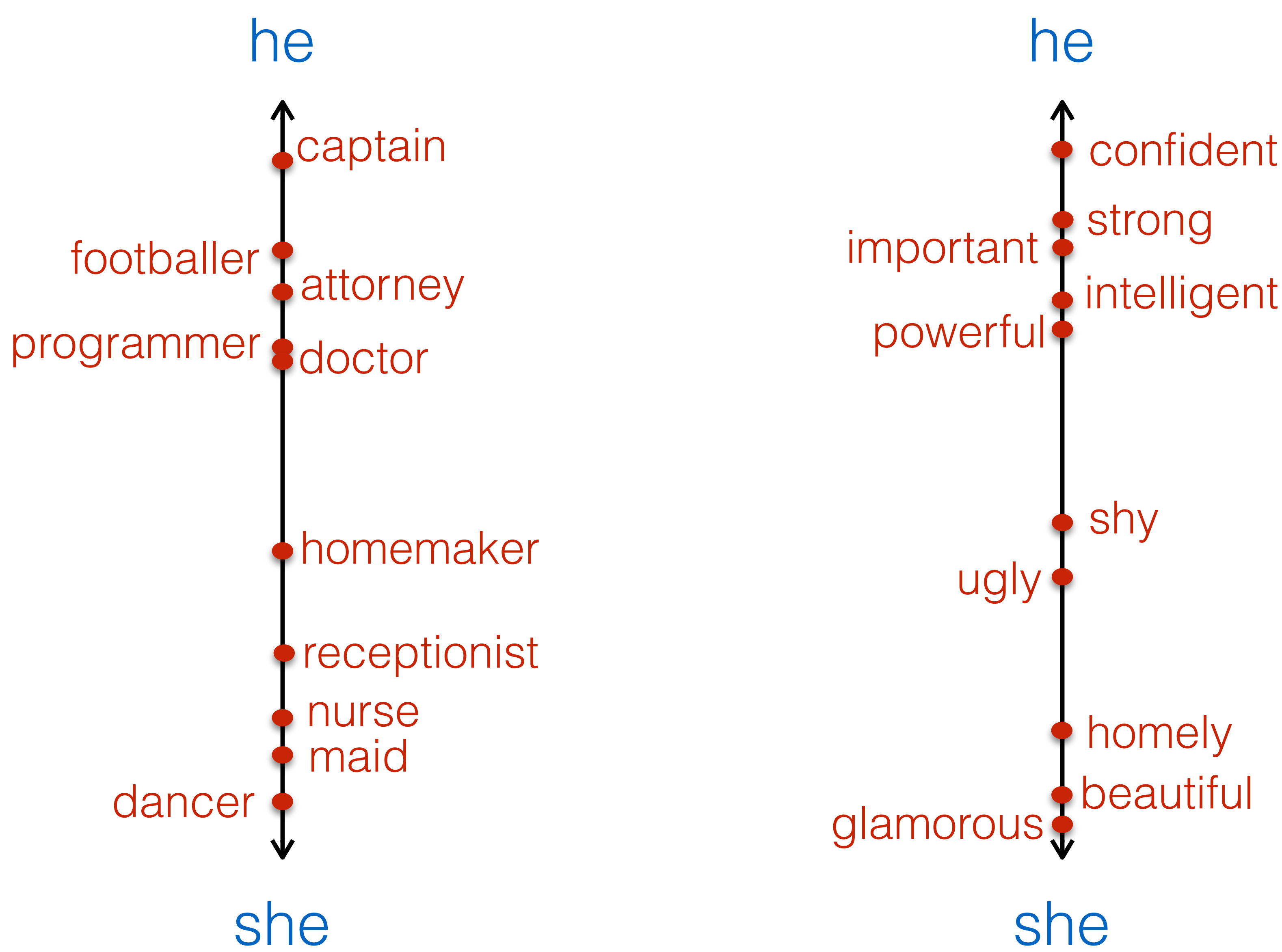
he
footballer
captain
programmer
doctor
attorney
homemaker
receptionist
dancer
nurse
maid
she
he
important
strong
intelligent
powerful
confident
shy
ugly
homely
glamorous
beautiful
she

... REPEATED OVER DATASETS ...

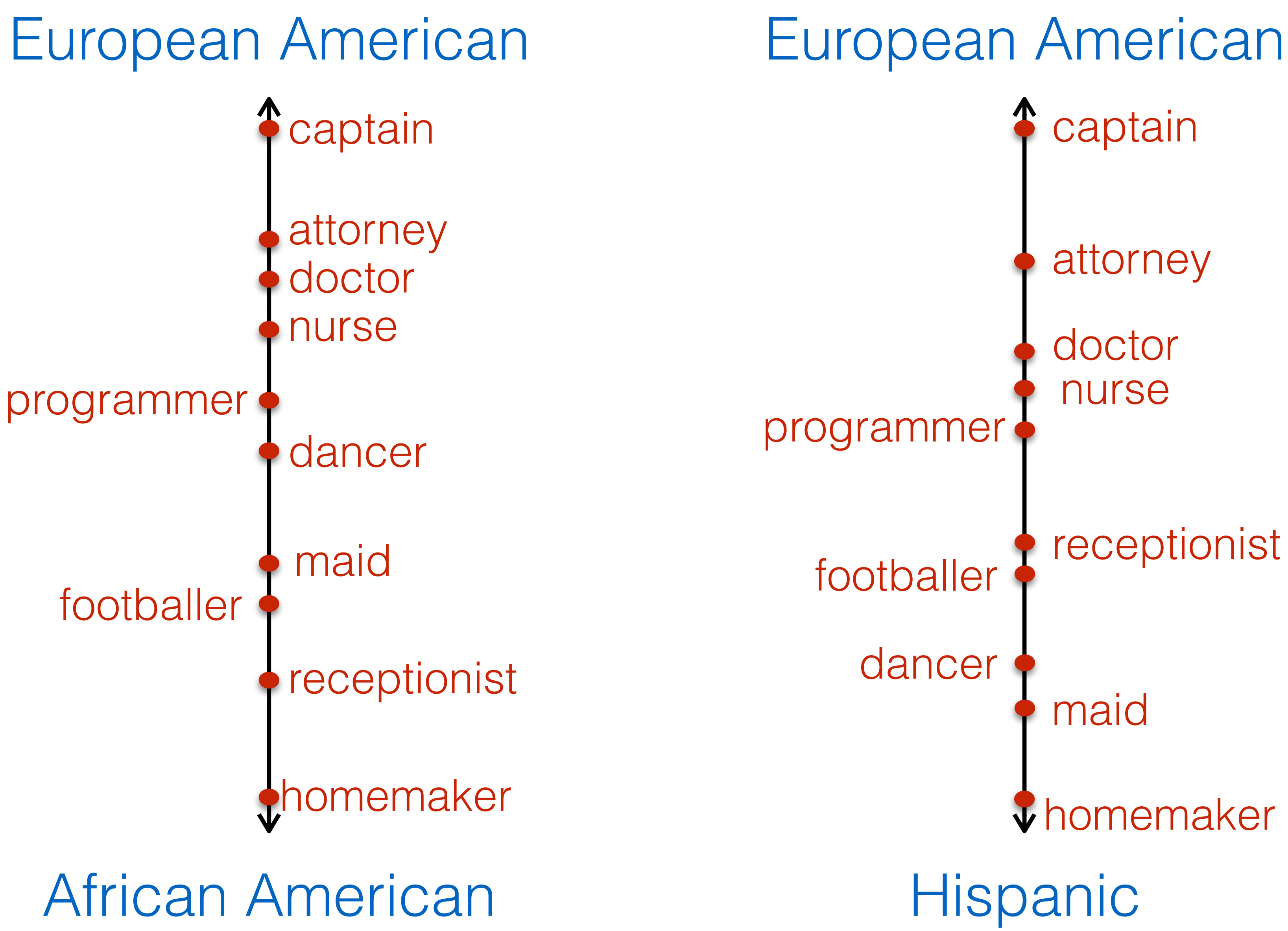




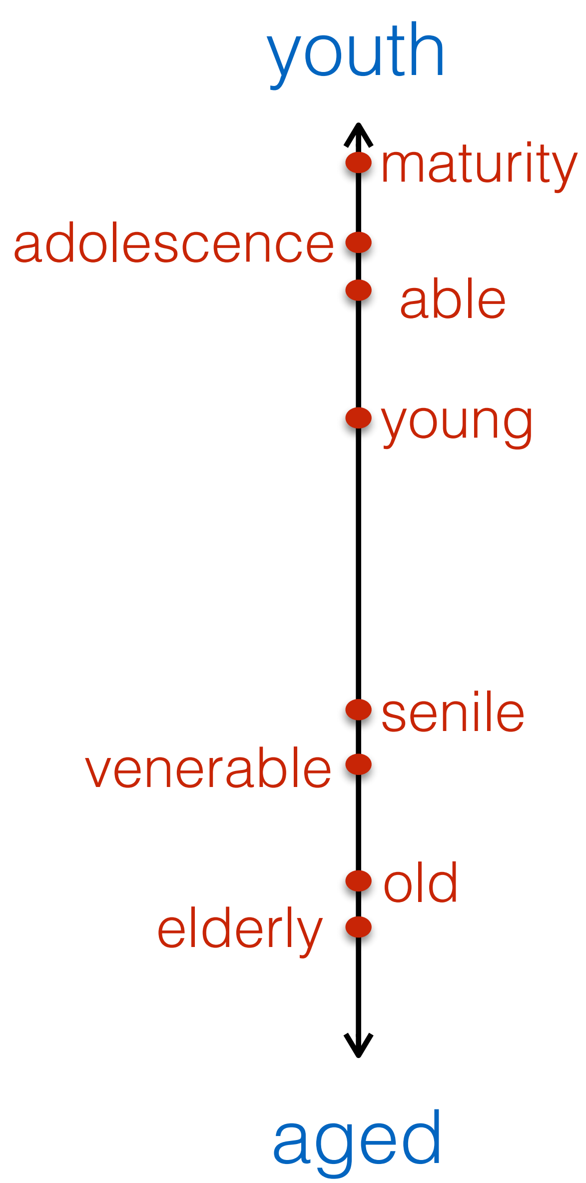
... AND EMBEDDING MECHANISMS



Racial Bias


European American
captain
attorney
doctor
nurse
programmer
dancer
maid
footballer
receptionist
homemaker
African American
European American
captain
attorney
doctor
nurse
programmer
receptionist
footballer
dancer
maid
homemaker
Hispanic

Age Based Bias

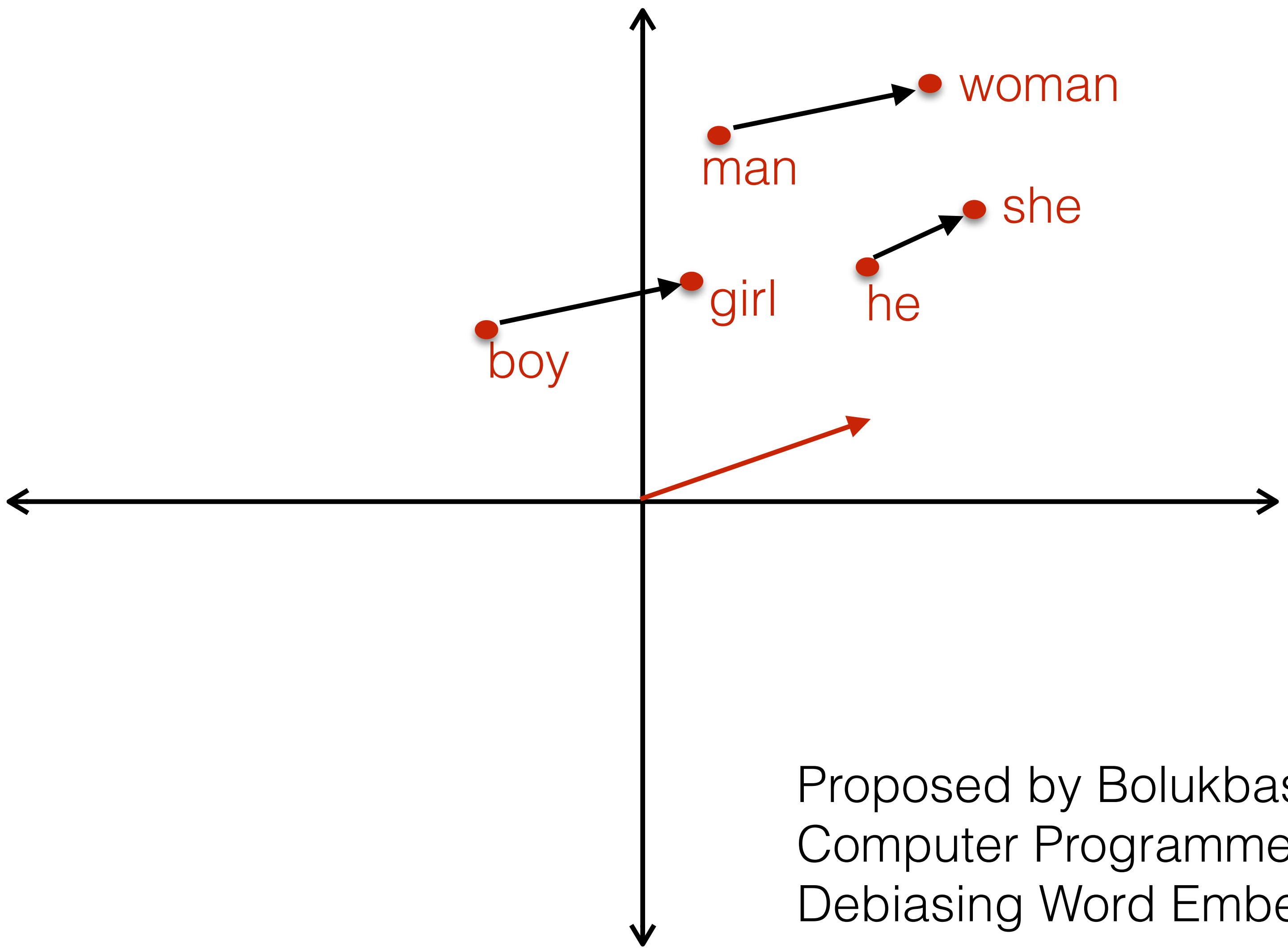


We propose new simple ways to :

- detect bias
- dampen or attenuate bias
- quantify bias

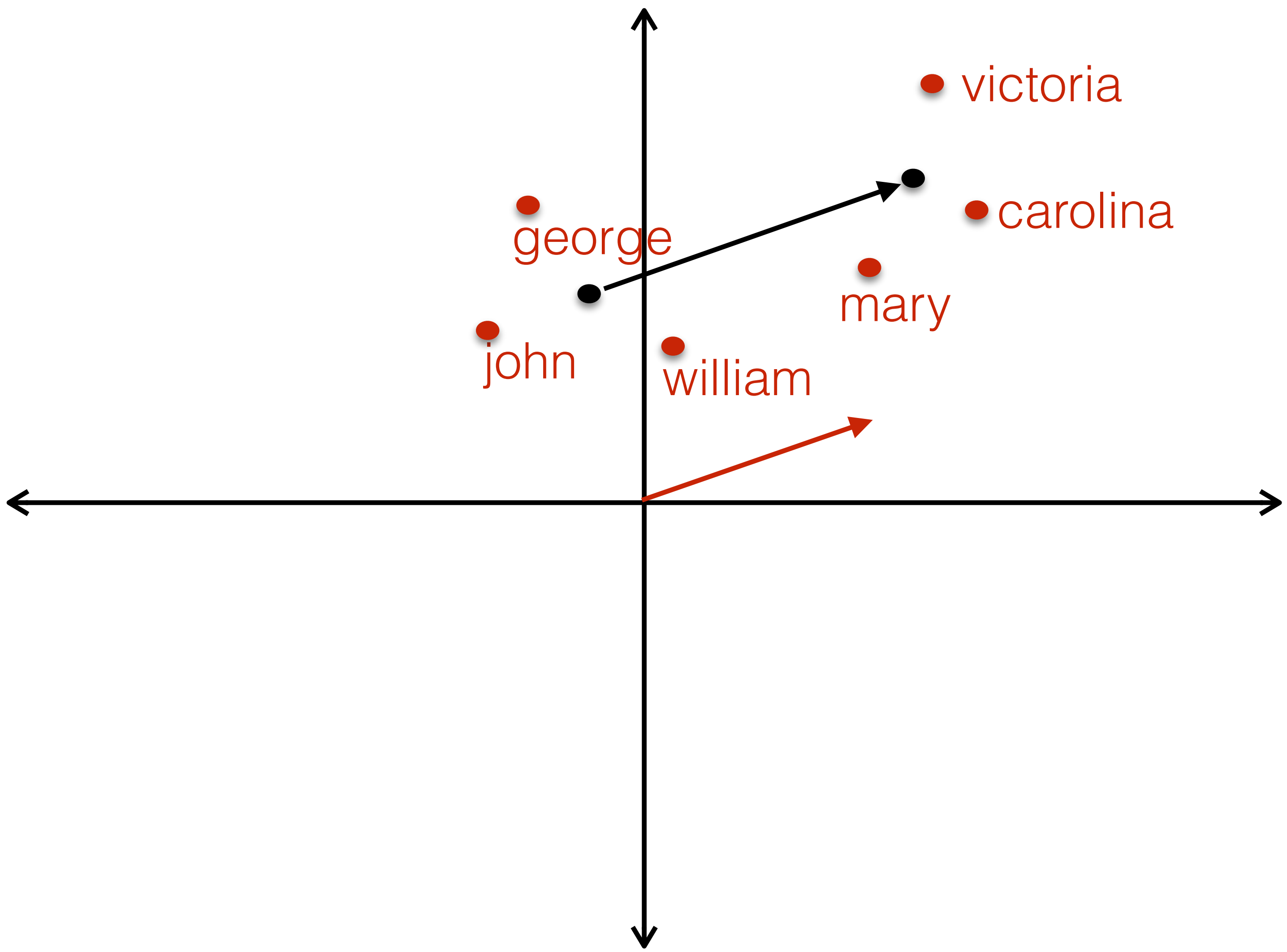
Bias Detection

Gendered Word Pairs

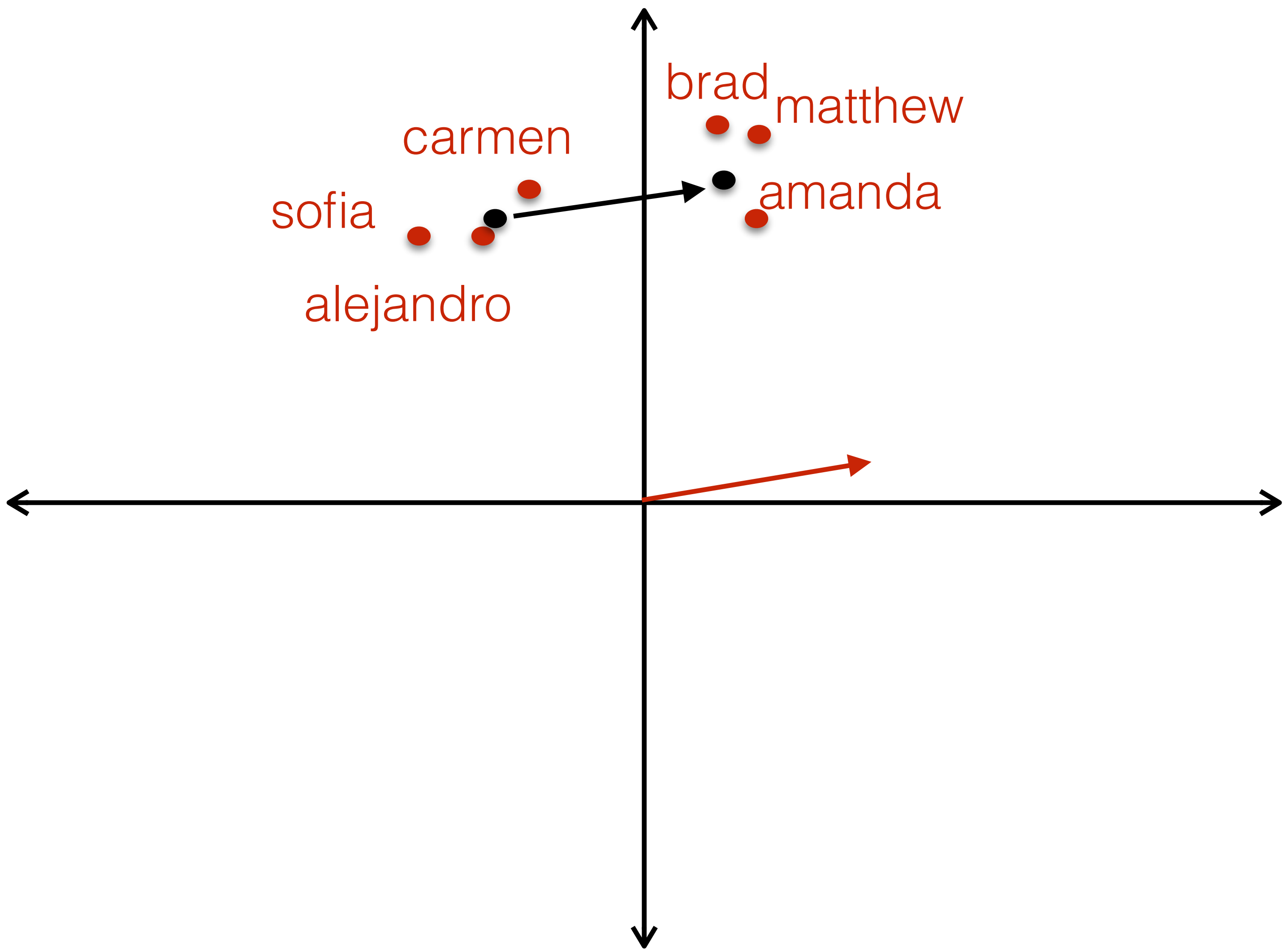


Proposed by Bolukbasi et al, 2016 : Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings

2 means method : using names



2 MEANS METHOD : USING NAMES



Dampening Bias

Flipping Raw Text

With probabilities {0.0, 0.5, 0.75, 1.0}, flip corresponding gendered words in a word pair :

- man - woman
- he - she
- boy - girl
- … and 75 such pairs

He was talking to the girl.

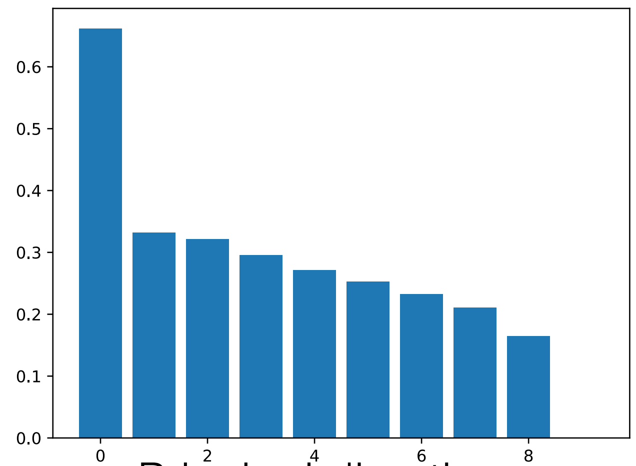
She was talking to the girl.

She was talking to the boy.

He was talking to the girl.

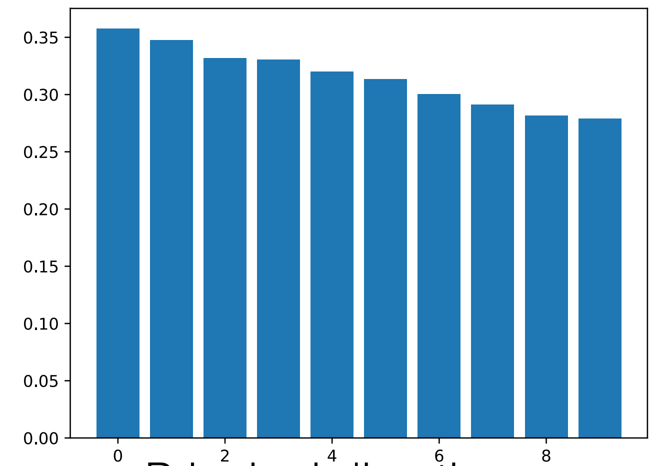
Analogy head	Original	p = 0.5	p = 0.75	p = 1.0
man : woman :: doctor	nurse	dr	dr	medicine
man : woman :: footballer	politician	midfielder	goalkeeper	Striker
he : she :: strong	weak	weak	strongly	many
he : she :: captain	mrs	lieutenant	colonel	colonel
john : mary :: doctor	nurse	medicine	surgeon	nurse

Principal Values



Principal directions

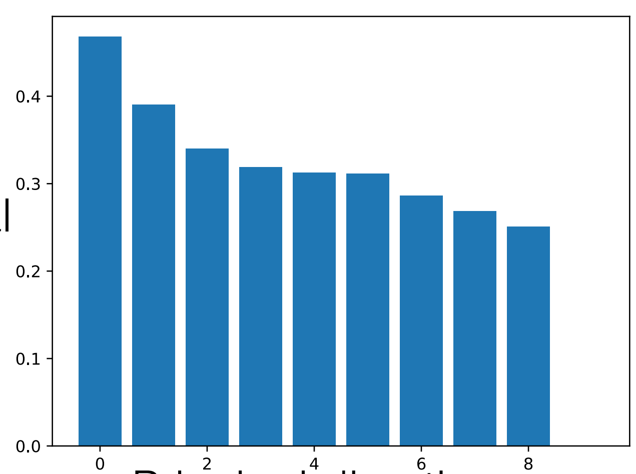
p = 0.0



Principal directions

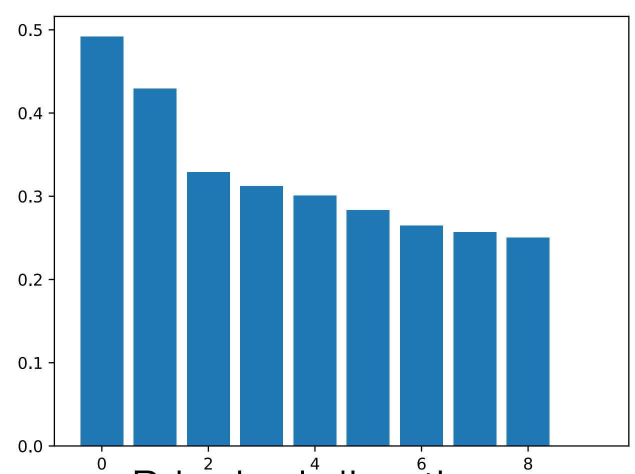
Random gaussian vectors

Principal Values



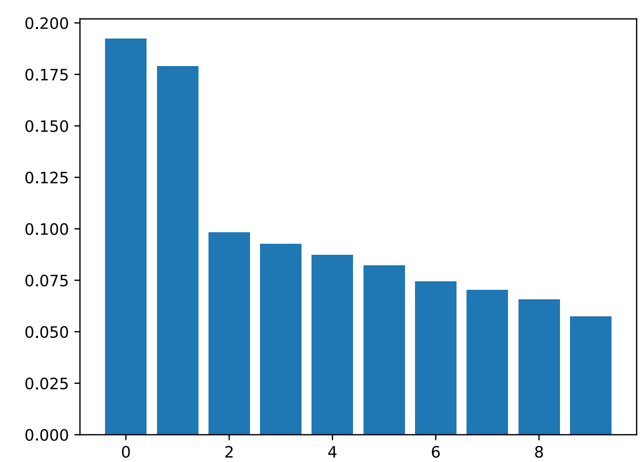
Principal directions

p = 0.50



Principal directions

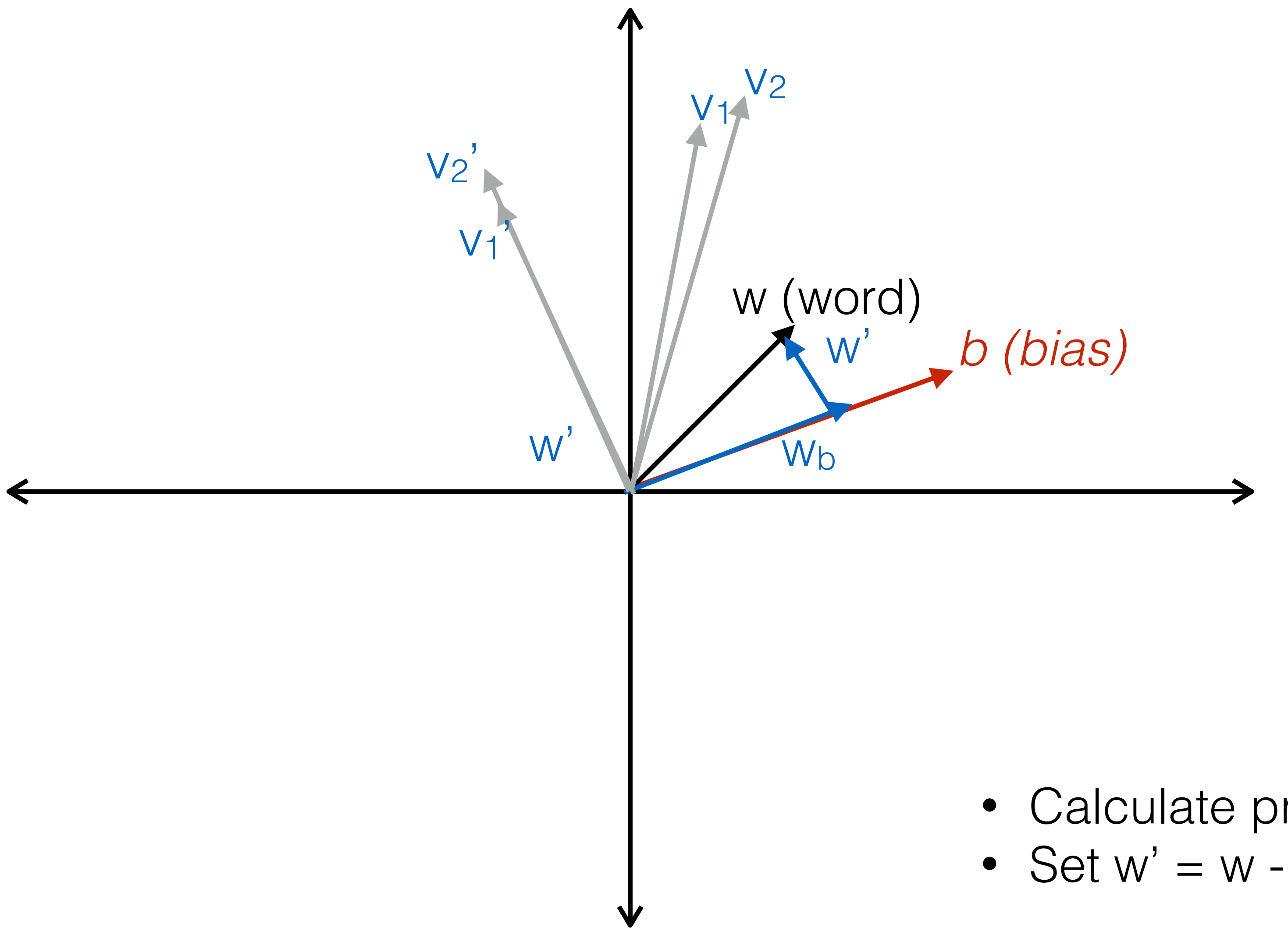
p = 0.75



Principal directions

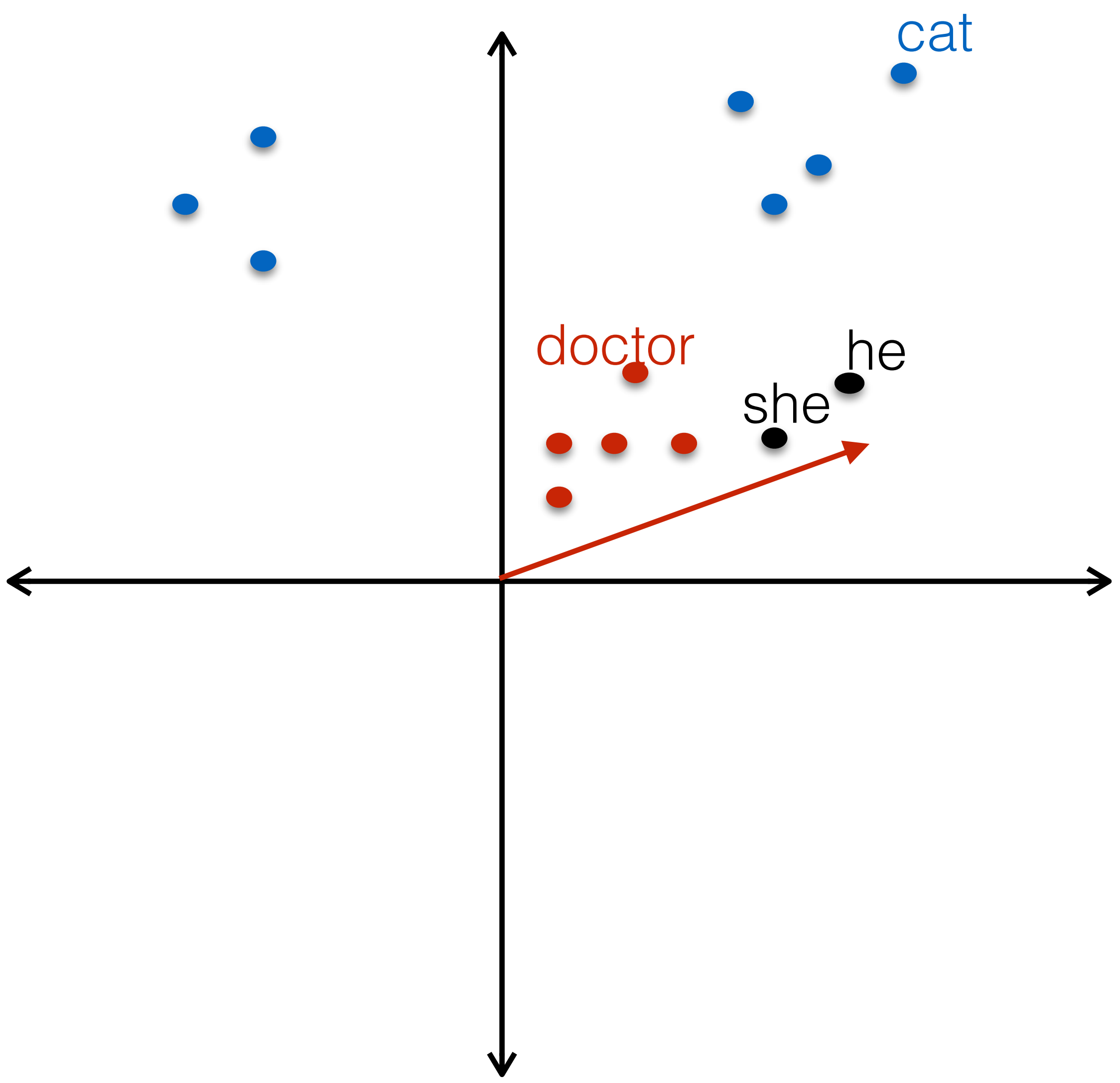
p = 1.0

Linear Projection

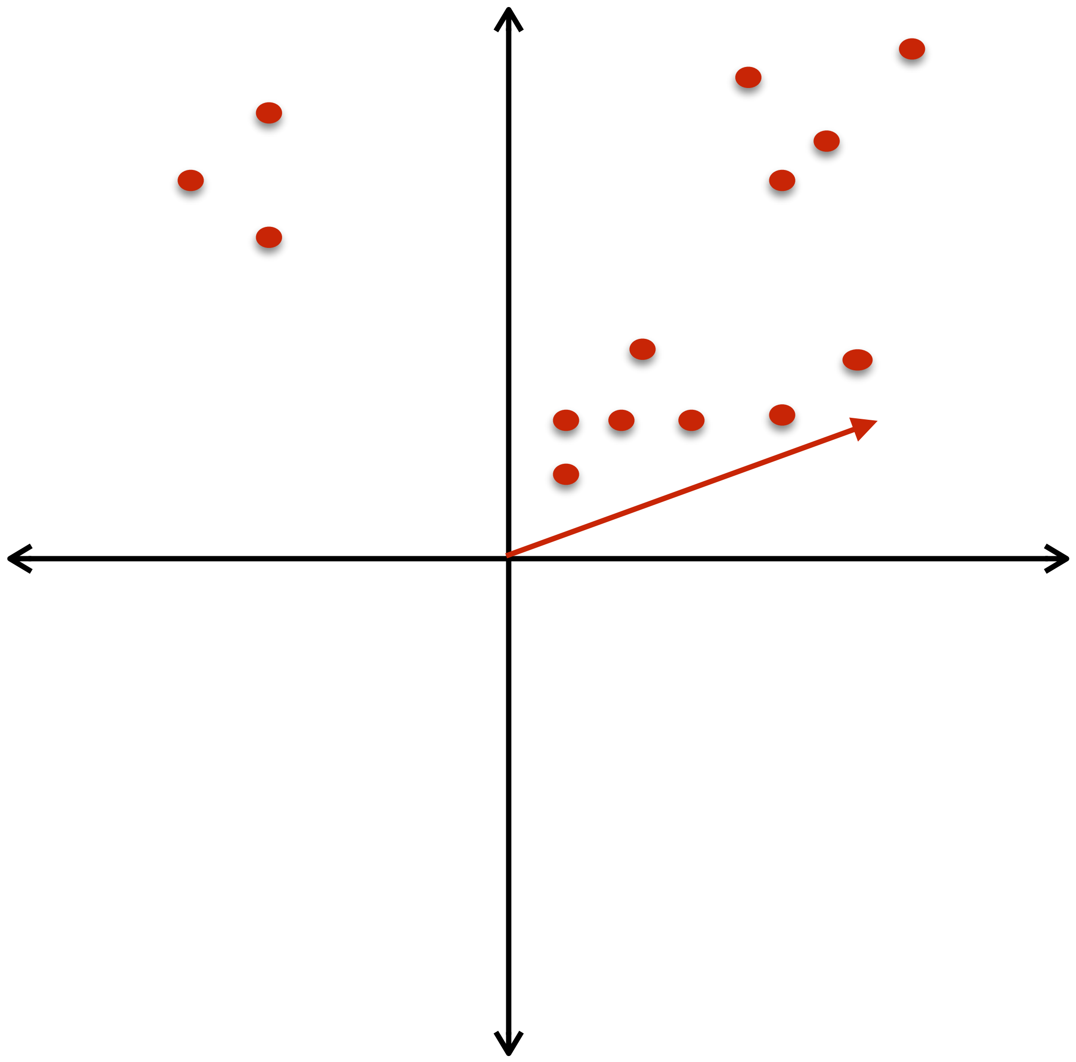


- Calculate projection, $w_b = < w,b > b$
- Set $w' = w - w_b$

Analogy head	Original	Projection
MAN : WOMAN :: DOCTOR	NURSE	PHYSICIAN
MAN : WOMAN :: FOOTBALLER	POLITICIAN	MIDFIELDER
HE : SHE :: STRONG	WEAK	STRONGER
HE : SHE :: CAPTAIN	MRS	LIEUTENANT
JOHN : MARY :: DOCTOR	NURSE	PHYSICIAN



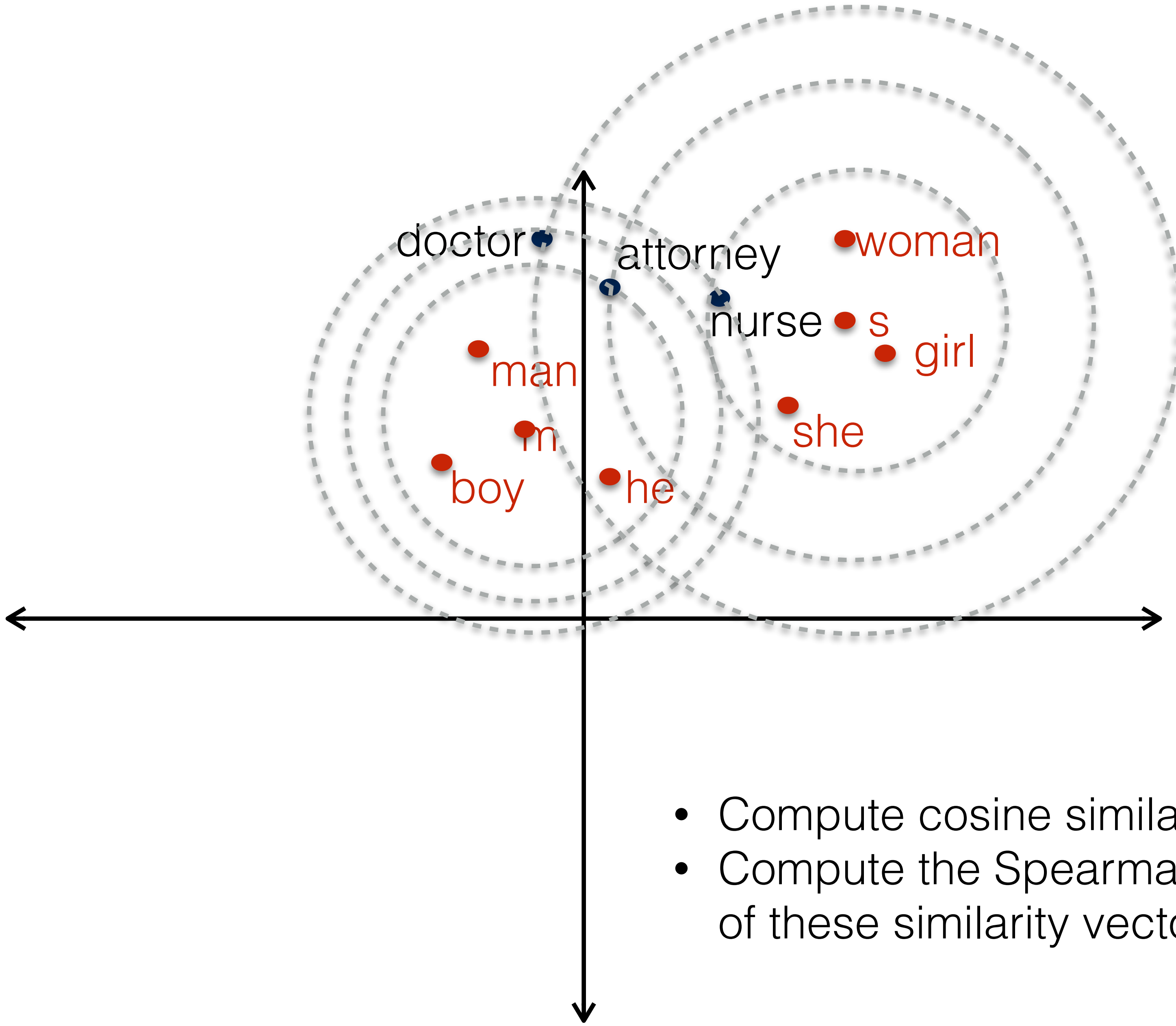
HARD DEBIASING



LINEAR PROJECTION

Quantifying Bias

Embedding Coherence Test

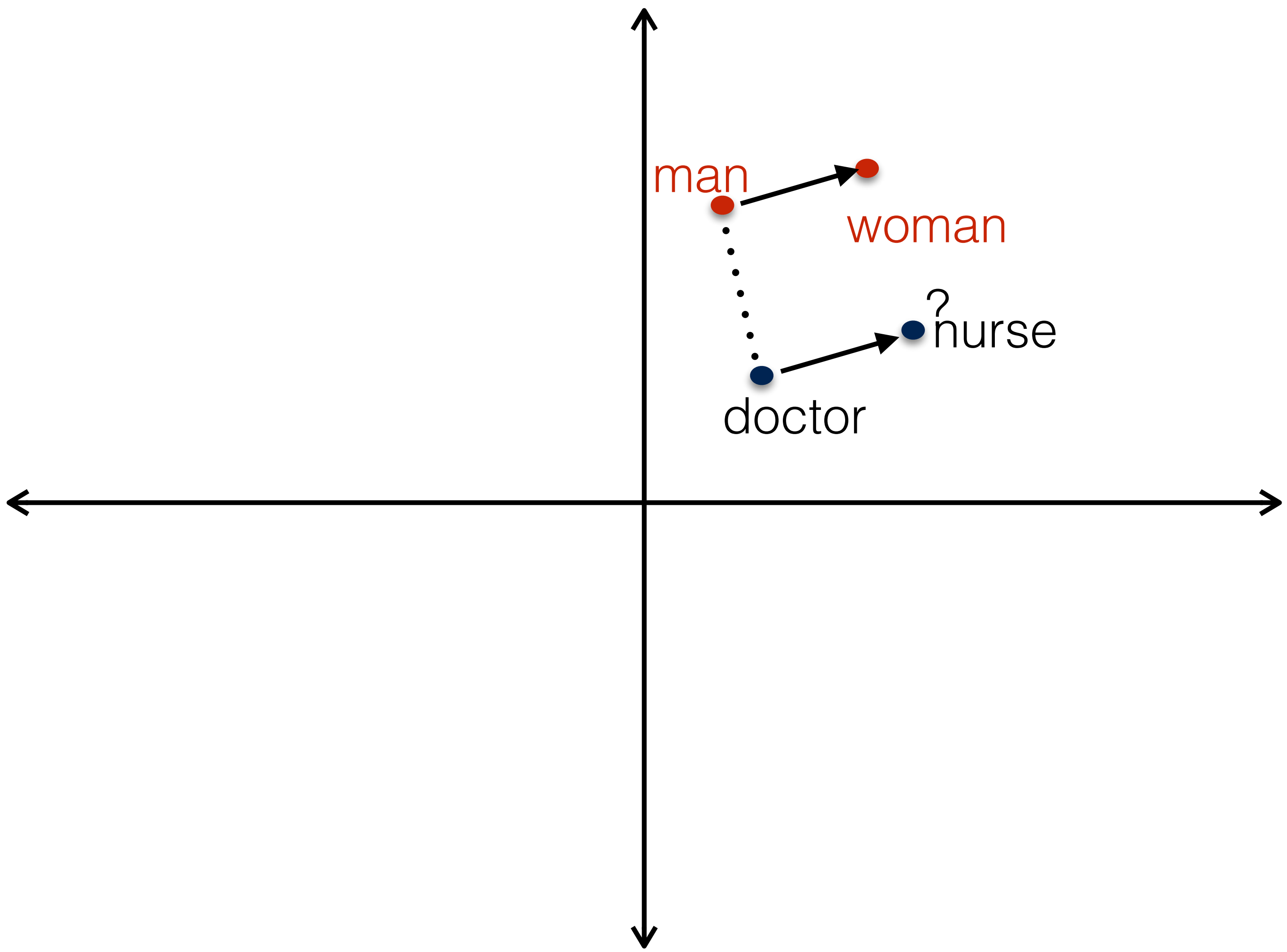


m	s
attorney	nurse
doctor	attorney
nurse	doctor

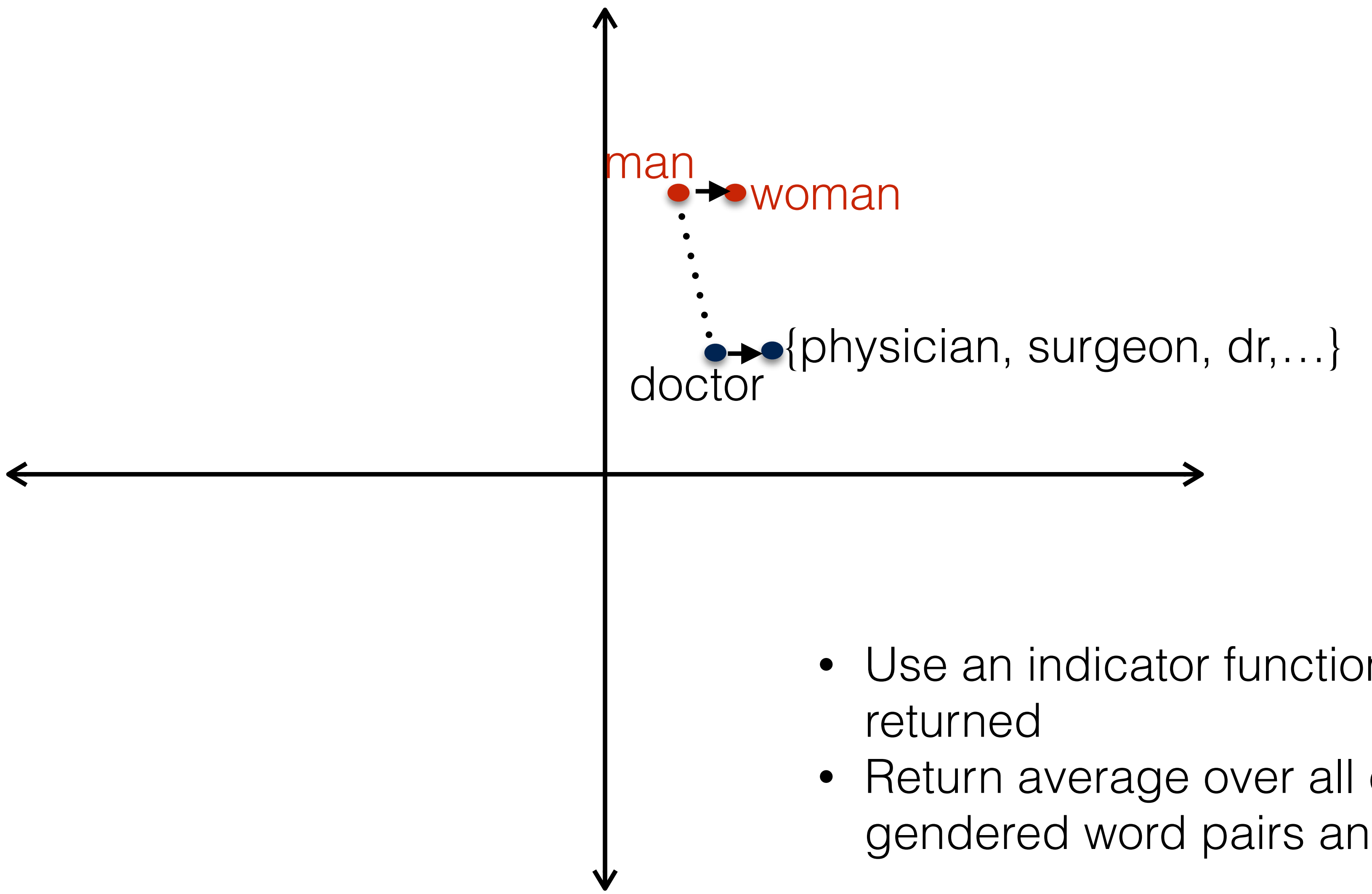
- Compute cosine similarity of m and s to target words
- Compute the Spearman Coefficient of the rank order of these similarity vectors of m and s

Test	Original	Hard Debiasing	Projection (word pairs)	Projection (names)	Flipping with p = 0.5	Flipping with p = 0.75	Flipping with p = 1.0
ECT (word pairs)	0.798	0.917	0.996	0.943	0.983	0.984	0.683
ECT (Names)	0.832	0.968	0.935	0.999	0.714	0.662	0.587

Embedding Quality Test



Embedding Quality Test



- Use an indicator function with 1 if a synonym returned
- Return average over all combinations of gendered word pairs and professions

Test	Original	Hard Debiasing	Projection (word pairs)	Projection (names)	Flipping with P = 0.5	Flipping with P = 0.75	Flipping with P = 1.0
EQT	0.128	0.145	0.283	0.291	0.131	0.098	0.085

WEAT :
(Word Embedding Association Test)

Proposed by Caliskan *et al*, for two sets of target words X and Y and attribute words A and B, the WEAT test statistic is :

$$s(X, Y, A, B) = \sum_{x \in X} s(x, A, B) - \sum_{y \in Y} s(y, A, B)$$

where,

$$s(w, A, B) = \text{mean}_{a \in A} \cos(a, w) - \text{mean}_{b \in B} \cos(b, w)$$

Test	Original	Hard Debiasing	Projection (word pairs)	Projection (names)	Flipping with P = 0.5	Flipping with P = 0.75	Flipping with P = 1.0
WEAT	1.623	1.221	1.233	1.219	1.164	1.09	1.03

Standardized Tests for Word Embedding Quality

Test	Original	Hard Debiasing	Projection (word pairs)	Projection (names)	Flipping with p = 0.5	Flipping with p = 0.75	Flipping with p = 1.0
Wsim	0.637	0.537 △ = 0.1	0.627 △ = 0.01	0.629 △ = 0.008	0.567 △ = 0.07	0.537 △= 0.01	0.536 △= 0.101
Simlex	0.324	0.314 △ = 0.01	0.321 △= 0.003	0.321 △ = 0.003	0.317 △= 0.007	0.314 △ = 0.01	0.264 △= 0.060
Google Analogy	0.623	0.561 △ = 0.062	0.565 △= 0.058	0.584 △ = 0.039	0.565 △= 0.058	0.561 △= 0.062	0.321 △= 0.302

Larger better (ECT (word pairs), ECT (names), EQT); Smaller better (WEAT, WSim, SimLex, Google Analogy)

Test	Original	Hard Debiasing	Projection (word pairs)	Projection (names)	Flipping with P = 0.5	Flipping with P = 0.75	Flipping with P = 1.0
ECT (word pairs)	0.798	0.917	0.996	0.943	0.983	0.984	0.683
ECT (Names)	0.832	0.968	0.935	0.999	0.714	0.662	0.587
EQT	0.1280	0.145	0.283	0.291	0.131	0.098	0.085
WEAT	1.623	1.221	1.233	1.219	1.164	1.09	1.03
WSim	-	0.1	0.01	0.008	0.07	0.01	0.101
SimLex	-	0.01	0.003	0.003	0.007	0.01	0.060
Google Analogy	-	0.062	0.058	0.039	0.058	0.062	0.302

WEAT Scores for Other Biases

Bias Type	Before Debiasing	After Debiasing
European American - African American	1.803	0.425
European American - Hispanic	1.461	0.480
Youth - Aged	0.915	0.704

Re-training or Post processing?

	Full Re-Training	Post Processing
Performance	?	?
Cost	$$$	¢

Summary

- Bias of different types can be detected in textual data; amplified in word embeddings.
- Names are a powerful tool for bias direction detection
- Mostly, the simple step of linear projection of all words in data away from bias direction helps debias the embedding
- Gender bias corrected GloVe embedding for Common Crawl (840B) can be found at

http://saphira.cs.utah.edu:8000/glove.cc.840b.unbiased.zip

Please contact sunipad@cs.utah.edu (or sunipa.github.io) for more details