The Geometry of Gender and Ethnic Stereotypes in Word Embeddings

James Zou

Stanford University

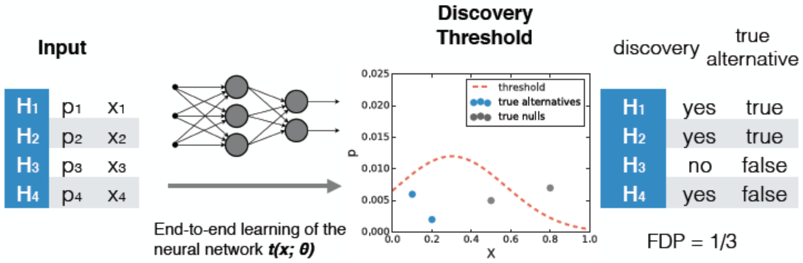
Chan-Zuckerberg BioHub

Joint work with T. Bolukbasi, K. Chang, V. Saligrama, A. Kalai, N. Garg, L. Schiebinger, D. Jurafsky

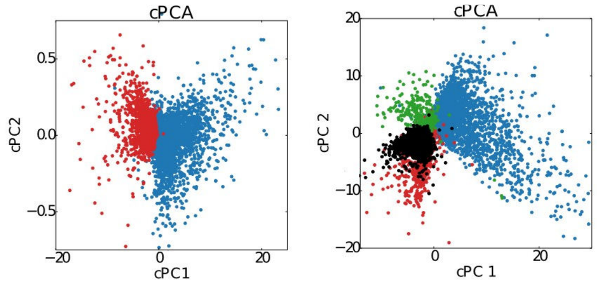
Stanford Machine Learning and CompBio Group

- New Algorithms and Theory:

NeuralFDR



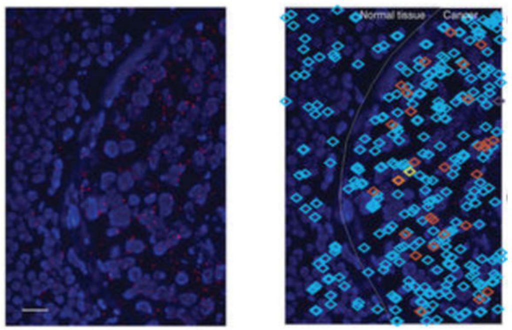
Contrastive PCA



• Patient 1 (pre) • Patient 1 (post) • Patient 2 (pre) • Patient 2 (post)

- Applications: AI for enabling new genomic technology

Spatial transcriptomics/Human cell atlas



Genome editing

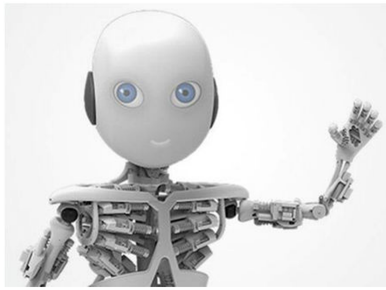
Risk prediction

ClinGen
Clinical Genome Resource

Dictionary for machine learning

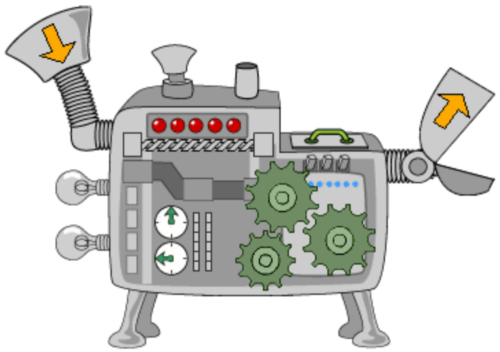
Raw data

ML algorithm



Dictionary for machine learning

Raw data

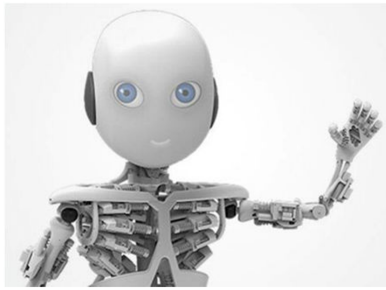


“dictionary”
word embedding

Vectors

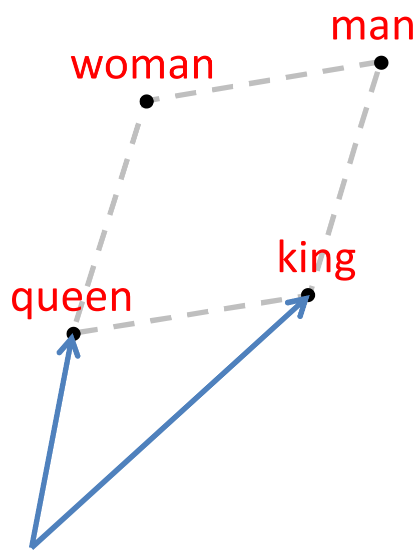
	A	B	C	D	E	F
1	the	0.056	0.043	0.051	0.08	0.006
2	cat	0.072	0.076	0.1	0.085	0.055
3	dog	0.088	0.099	0.028	0.059	0.06
4	nurse	0.03	0.018	0.058	0.074	0.055
5	doctor	0.097	0.093	0.035	0.057	0.044
6	king	0.013	0.059	0.024	0.032	0.038
7	queen	0.087	0.072	0.029	0.042	0.05
8	bird	0.042	0.044	0.006	0.003	0.003

ML algorithm



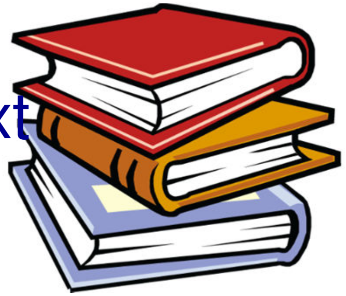
Word embedding is a dictionary

- word embedding is a dictionary: word → vector
- Related words are nearby vectors
- Geometry captures semantics



Word embedding is a dictionary

Training data: corpus of text



Context window

The dog is chasing a cat.

v_{dog} $v_{chasing}$ v_{cat}

find v's to $\max \log P(chasing|dog) + \log P(cat|dog)$

where $P(cat|dog) \propto \exp(v_{cat} \cdot v_{dog})$

Word embedding is a dictionary

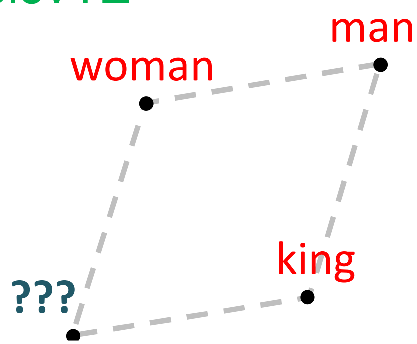
- word embedding is a dictionary: word → vector
- Related words are nearby vectors
- Geometry captures semantics
- Word2vec, GloVe and variants in other languages.

Beyond bilingual: multi-sense word embedding using multi-lingual context. 2017
Learning covariate-specific embeddings with tensor decomposition. 2017

Analogies generated by embedding

Parallelograms capture semantics: [MikolovYZ 13]

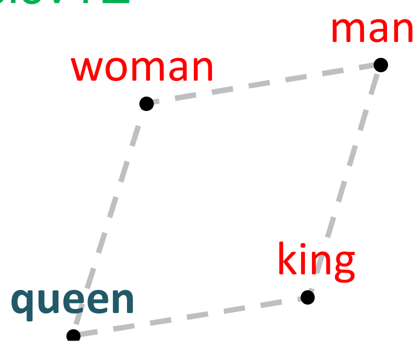
- Man:King :: Woman:***???***



Analogies generated by embedding

Parallelograms capture semantics: [MikolovYZ 13]

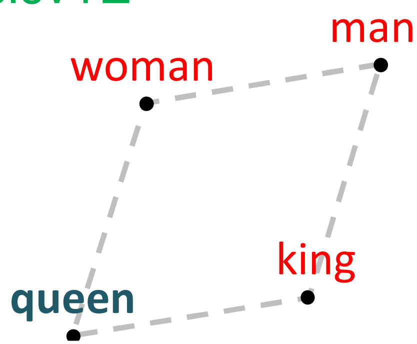
- Man:King :: Woman:*Queen*



Analogies generated by embedding

Parallelograms capture semantics: [MikolovYZ 13]

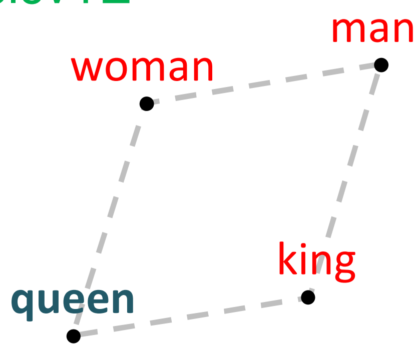
- Man:King :: Woman:*Queen*
- Paris:France :: Tokyo:*Japan*



Analogies generated by embedding

Parallelograms capture semantics: [MikolovYZ 13]

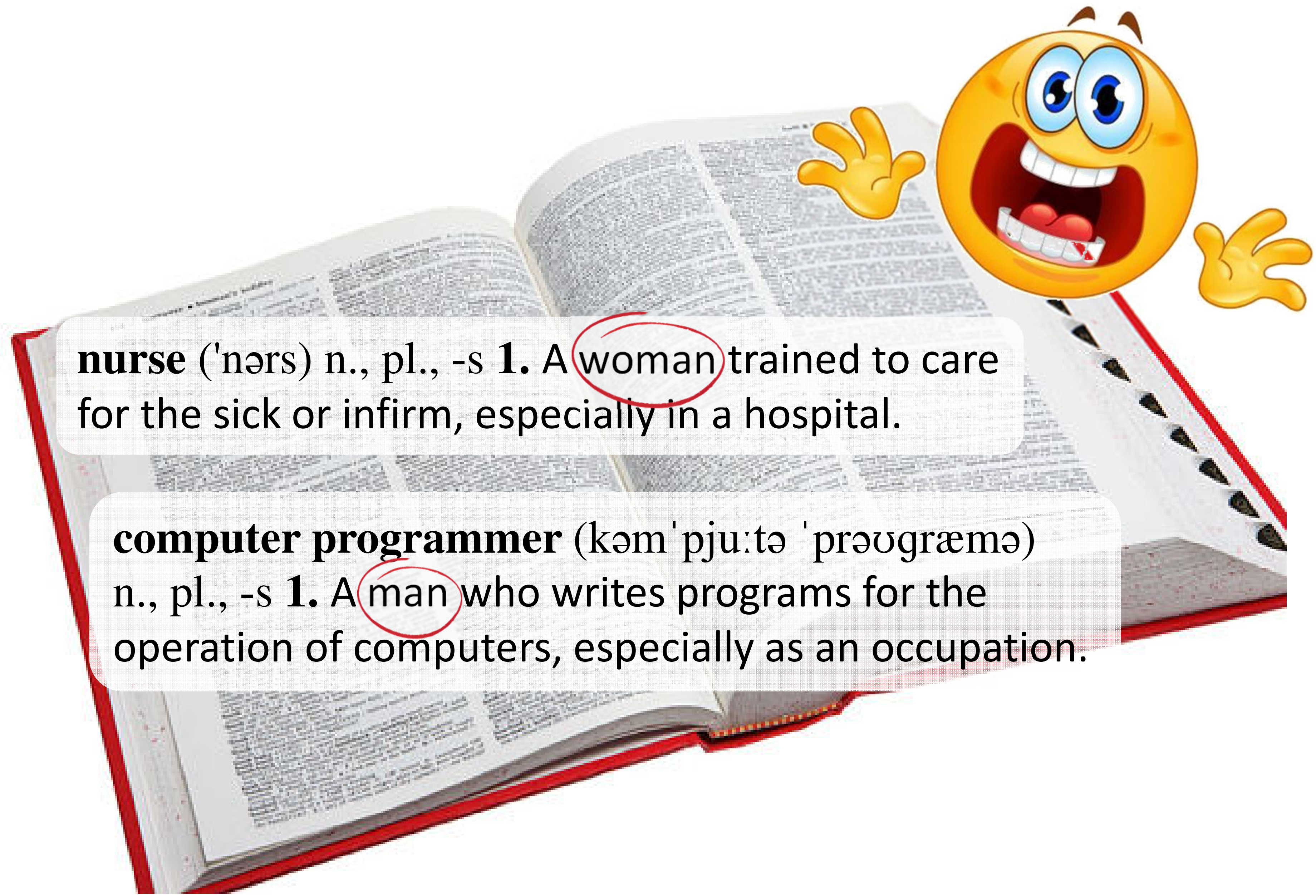
- Man:King :: Woman:*Queen*
- Paris:France :: Tokyo:*Japan*



- He:*Brother* :: She:*Sister*
- He:*Blue* :: Sh
- He:*Doctor* :: Sh
- He:*Architect* :: S *designer*
- He:*Realist* :: S
- She:*Pregnancy* :: *stone*
- He:*Computer programmer* ::

Based on word2vec trained on Google News corpus

NIPS 2016

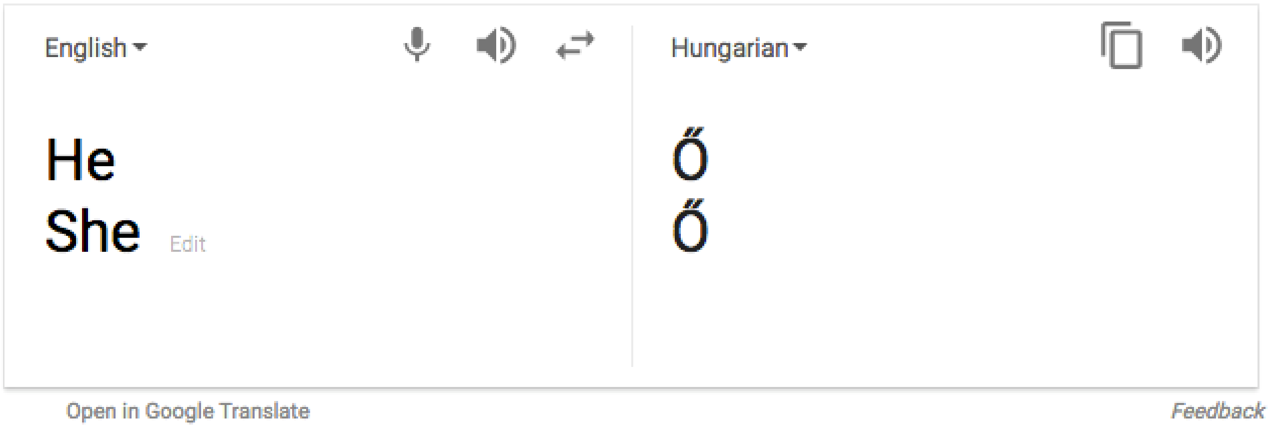


nurse ('nərs) n., pl., -s **1.** A woman trained to care for the sick or infirm, especially in a hospital.

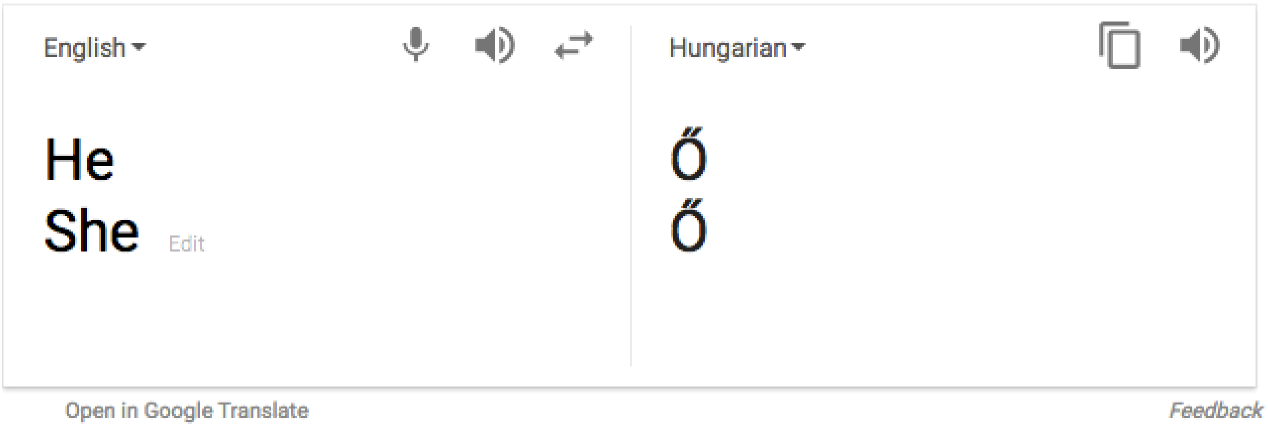
computer programmer (kəmˈpjuːtə ˈprəʊgræmə) n., pl., -s **1.** A man who writes programs for the operation of computers, especially as an occupation.

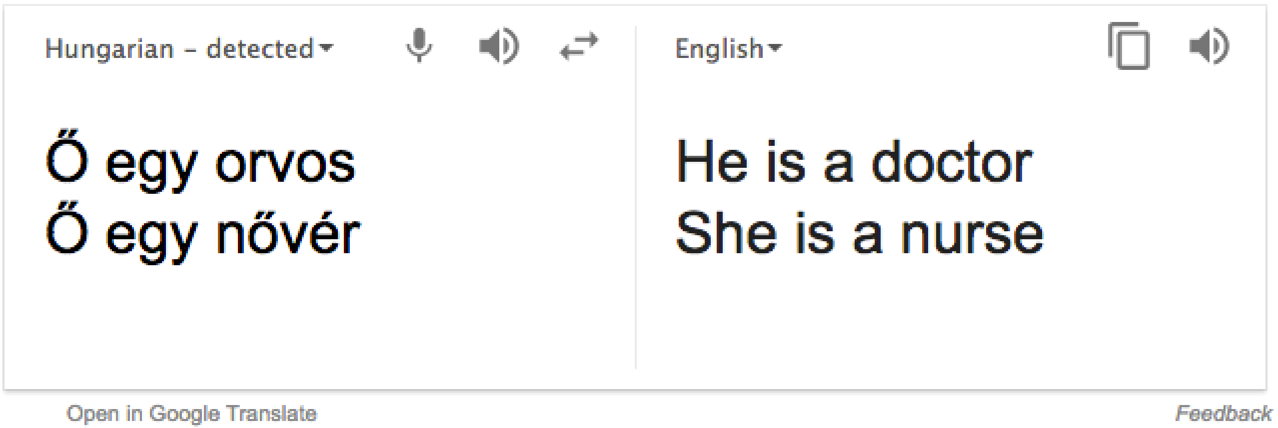
Hypothetical dictionary

Mishaps in Google Translate



Mishaps in Google Translate





Londa Schiebinger (2012) Caliskan-Islam et al. (2016)

Talk outline

1. 3 metrics for quantifying embedding stereotypes.

2. debiasing algorithm.

3. embedding as a lens to study 100 years of stereotypes.

Metric1: occupations.

327 gender neutral occupations. Project on to *she*—*he* directio

Metric1: occupations.

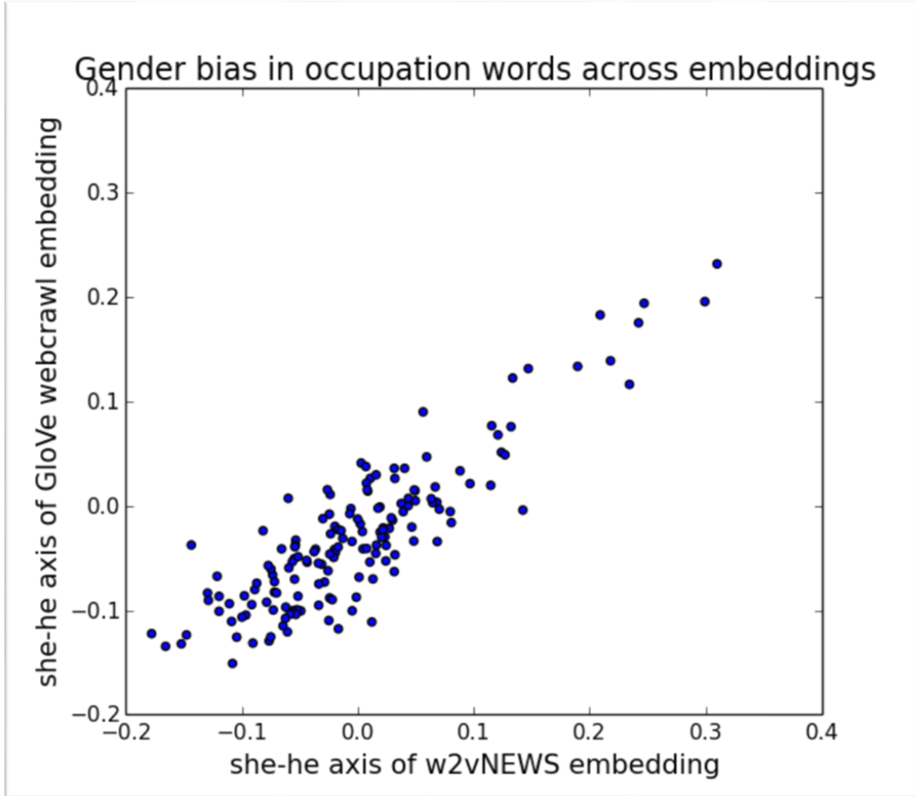
327 gender neutral occupations. Project on to *she*—*he* directio

Crowdworkers rate each occup. for gender stereotype

$$\text{Corr}(\text{projection}_{she-he}, \text{crowd rating}) = 0.51$$

Consistency of embedding stereotype

GloVe trained on web crawl

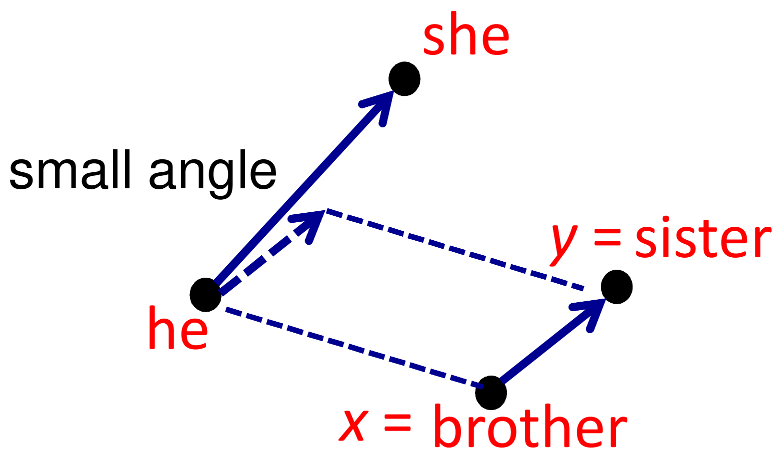


word2vec trained on Google news

Each dot is an occupation; Spearman = 0.8

Metric 2: analogies.

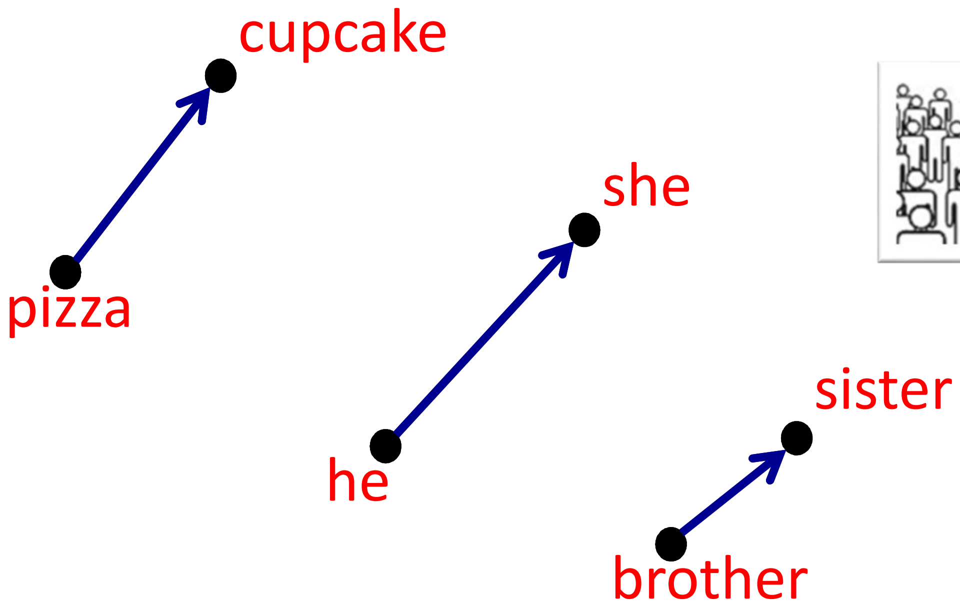
Automatically generate he : x :: she : y analogies.



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

Metric 2: analogies.

Automatically generate he : *x* :: she : *y* analogies.



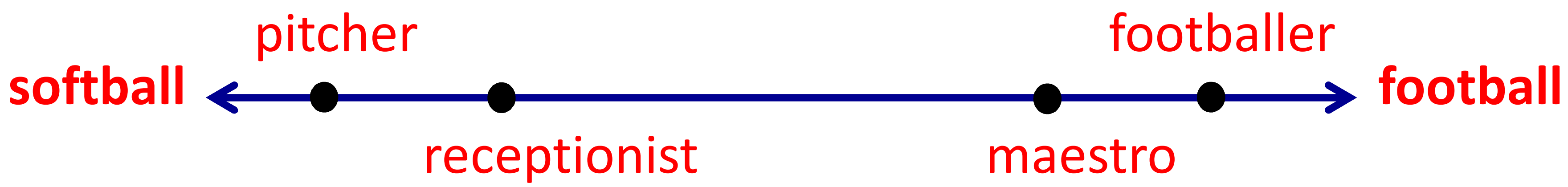
29/150 analogies rated as gender stereotypic by majority of crowdworkers

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

Metric 3: indirect bias.

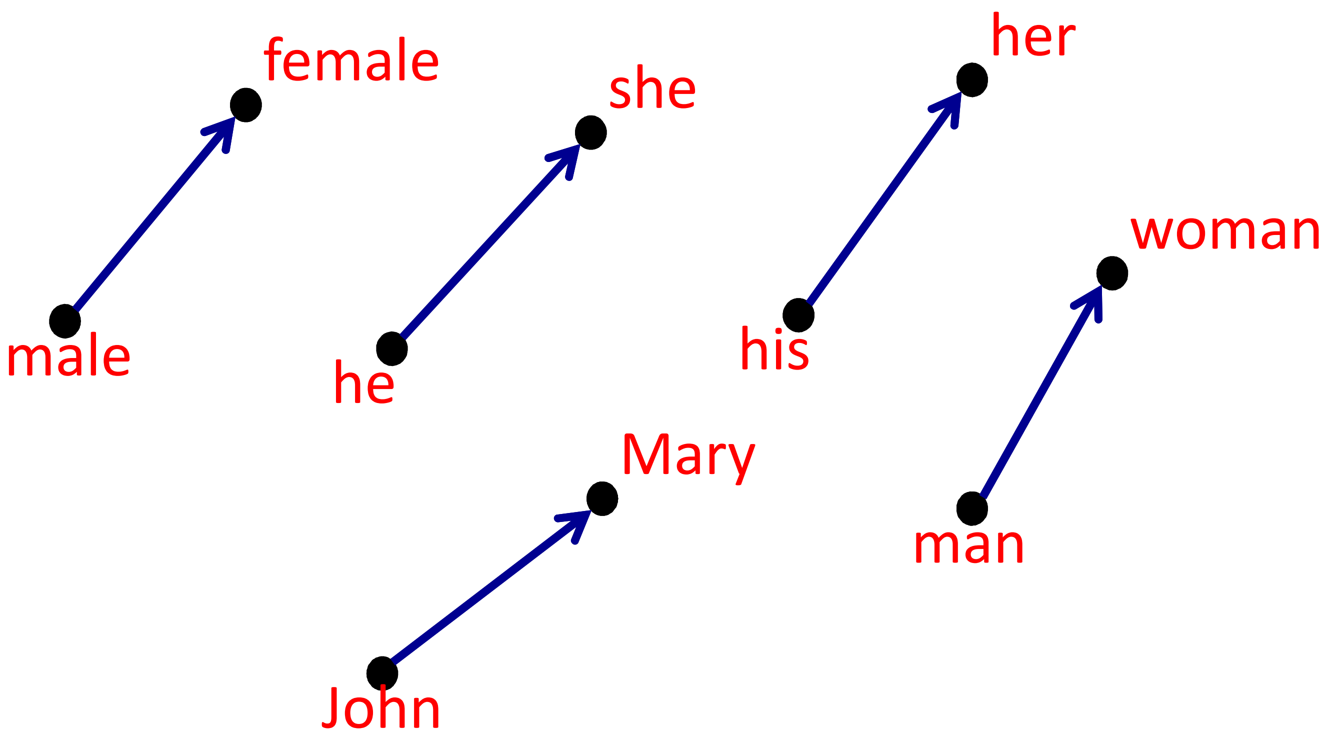
- Gender stereotype could affect the geometry between words that should be gender-neutral.

- Project occupations onto softball—football axis.



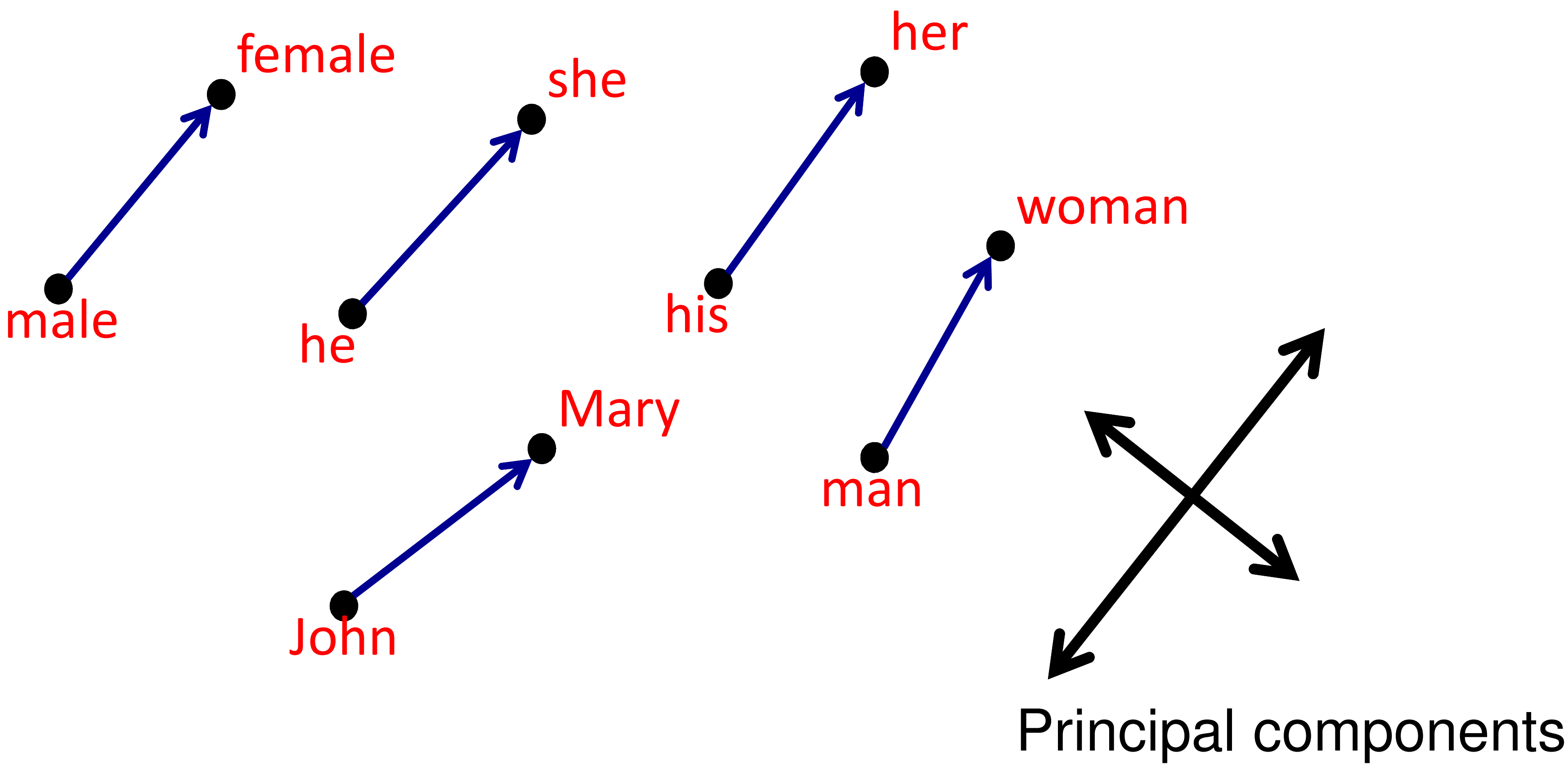
The geometry of gender

Select pairs of words that reflect gender opposites.

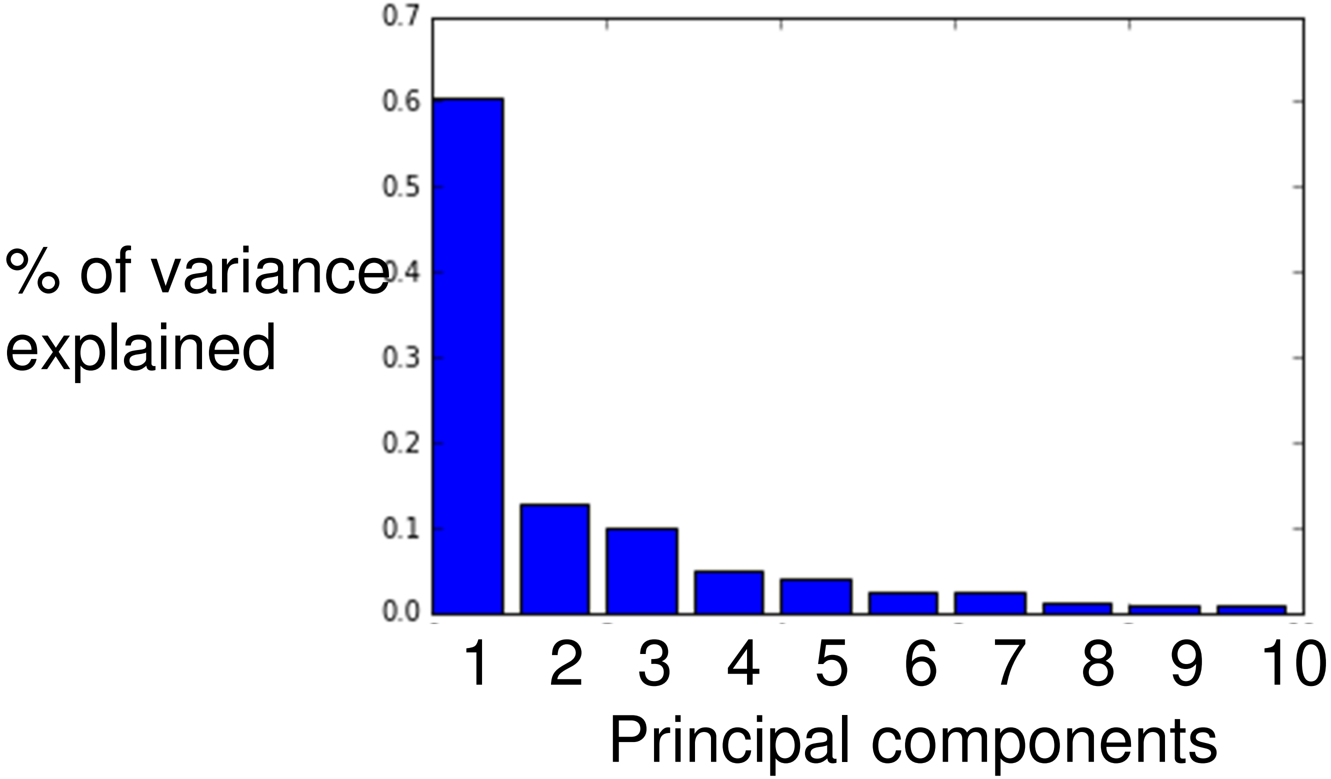


The geometry of gender

Select pairs of words that reflect gender opposites.



Geometry of gender



The top PC seems to capture the gender subspace *B*.

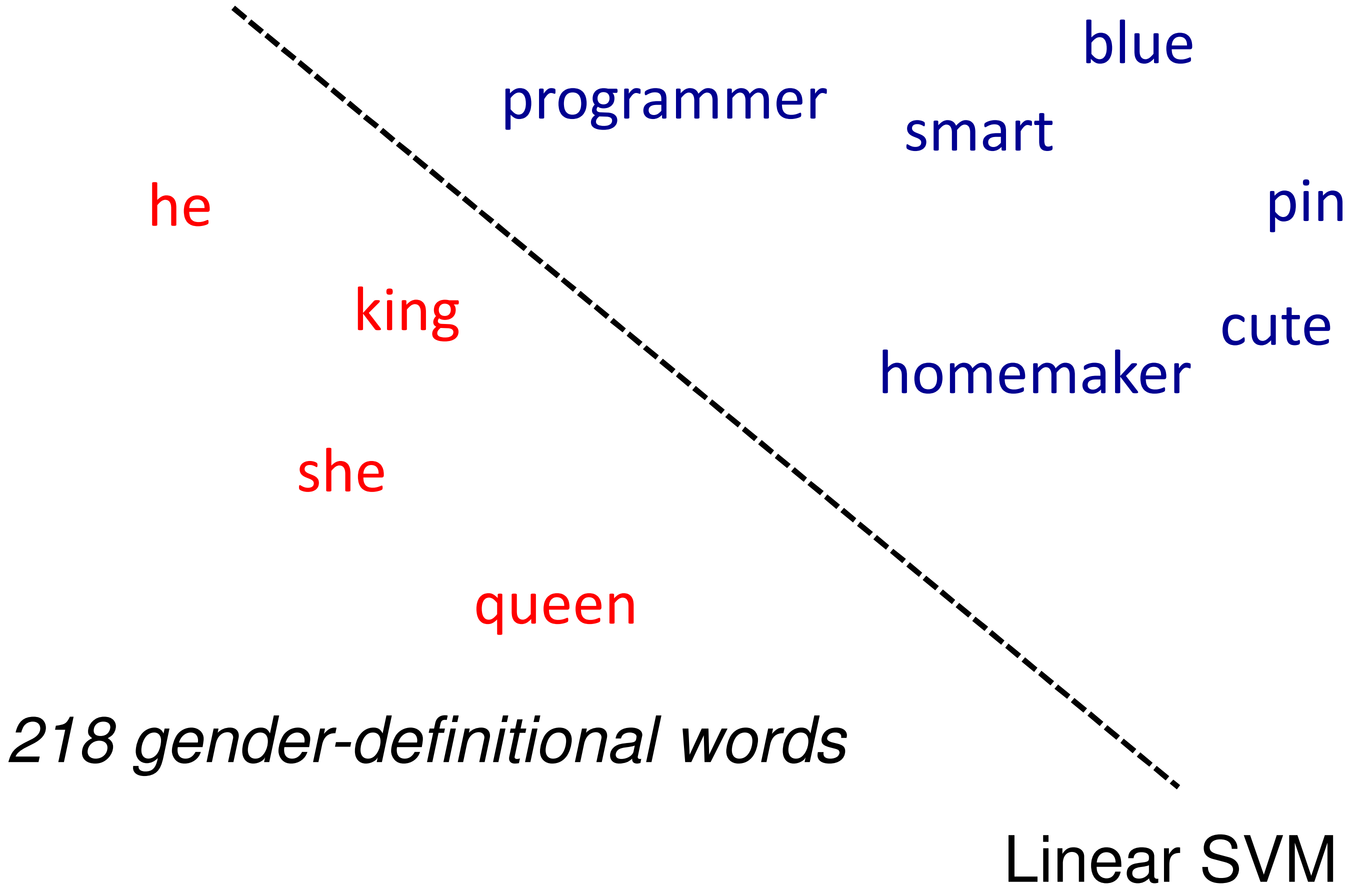
Debiasing algorithm (ver.1)

1. Identify words that are gender-neutral N and gender-definitional S.

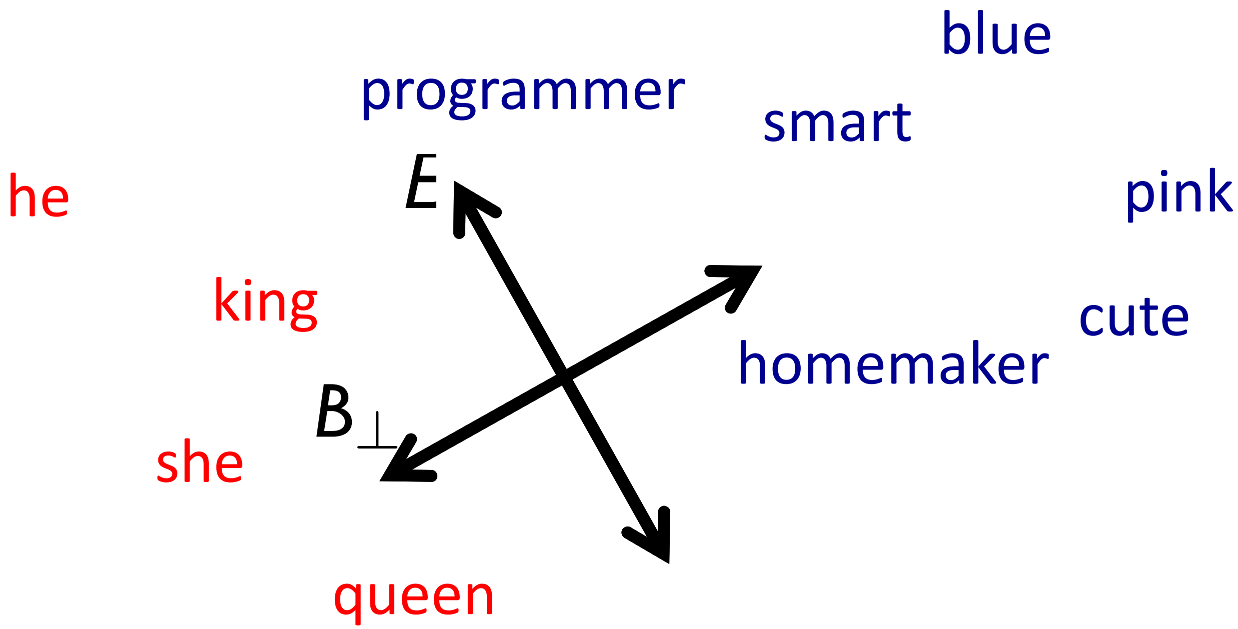
2. Project away the gender subspace from the gender-neutral words.
   $w := w - w \cdot B$ for $w \in N$ $\quad$ B is the gender subspace.

3. Normalize vectors.

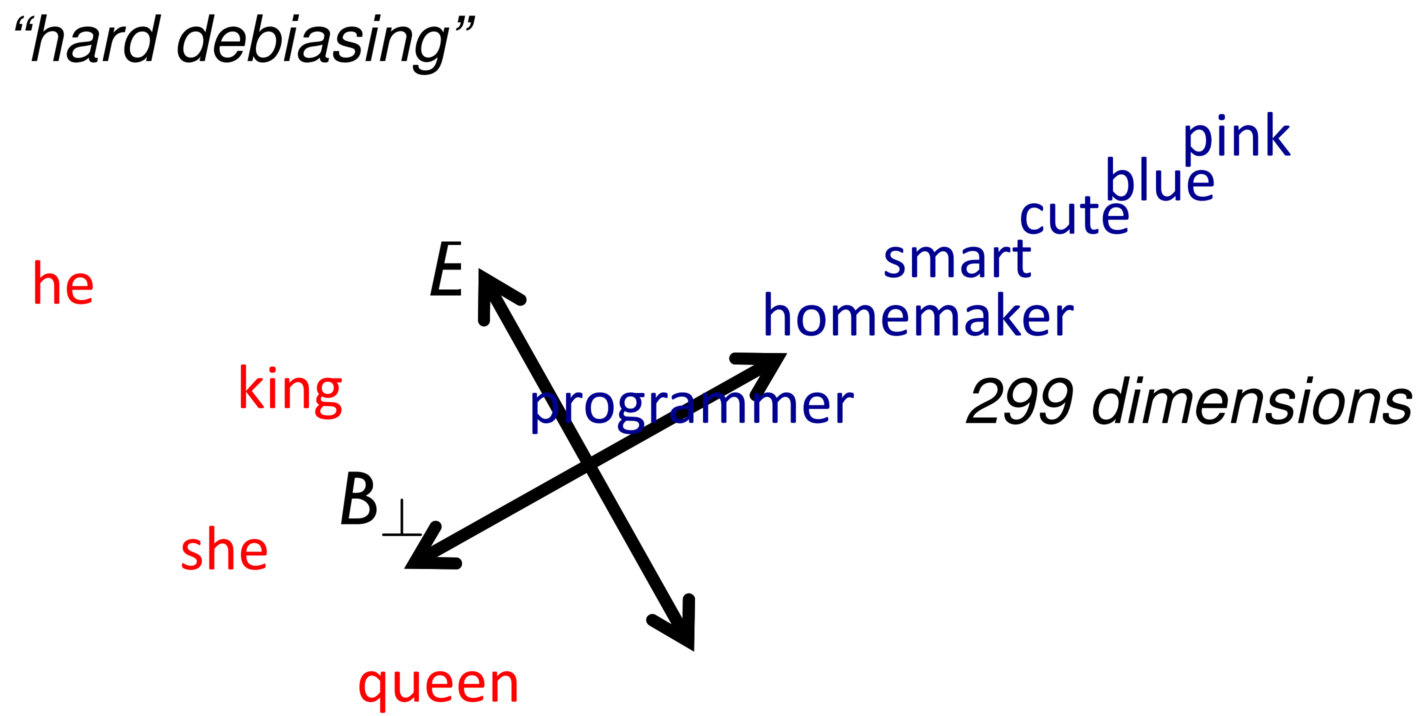
Identify gender-definitional words



Projecting away gender component



Projecting away gender component



Advanced debiasing

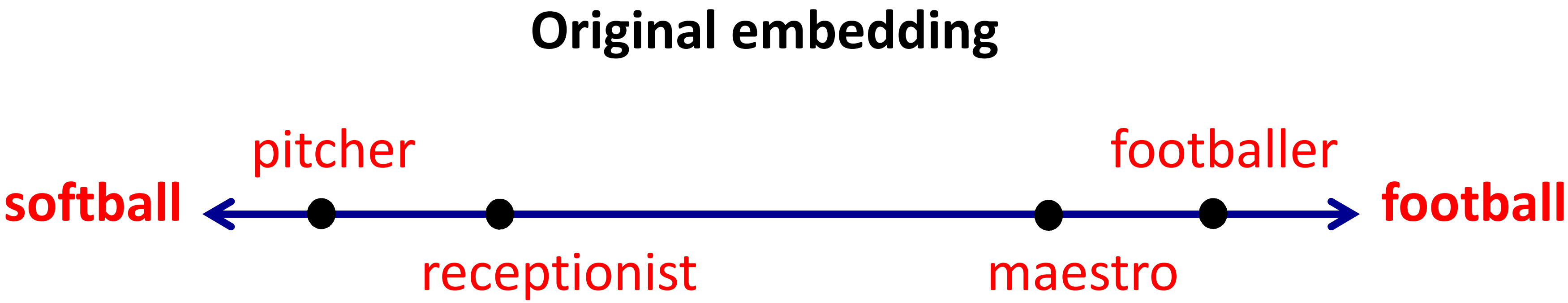
Find a linear transformation T of the gender-neutral words to reduce the gender component while not moving the words too much.

$W =$ matrix of all word vectors.

$N =$ matrix of neutral word vectors.

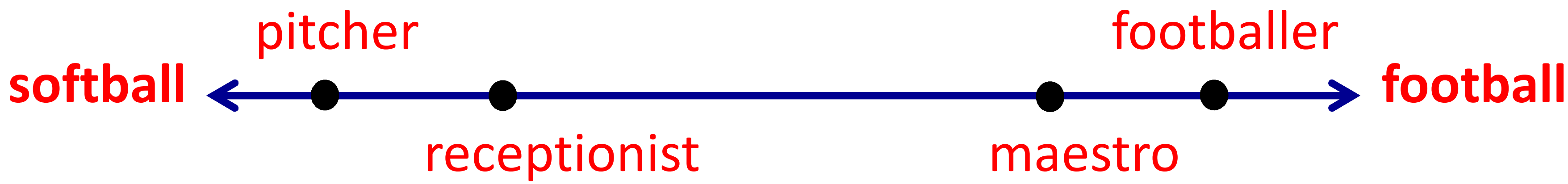
$$\min_T \underbrace{||(TW)^T(TW) - W^TW||_F^2}_{\text{don't move too much}} + \lambda \underbrace{||(TN)^T(TB)||_F^2}_{\text{minimize gender component}}$$

Debiasing results: indirect bias

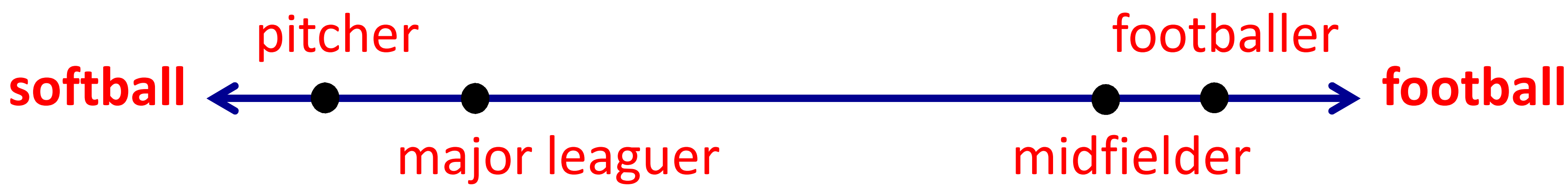


Debiasing results: indirect bias

Original embedding

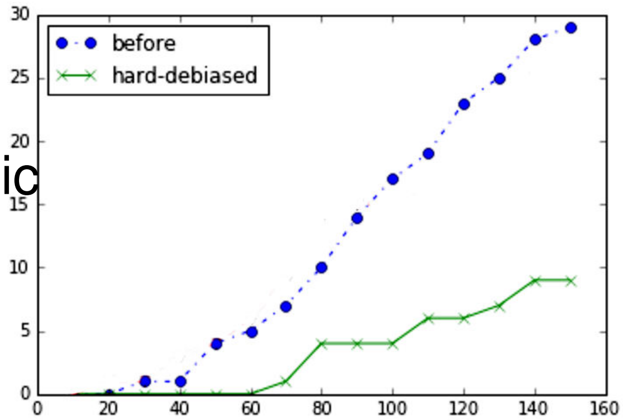


Debiased embedding



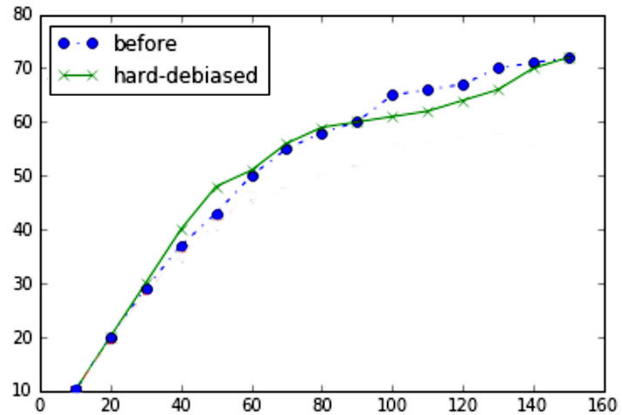
Debiasing results: analogies

stereotypic analogies



analogies generated

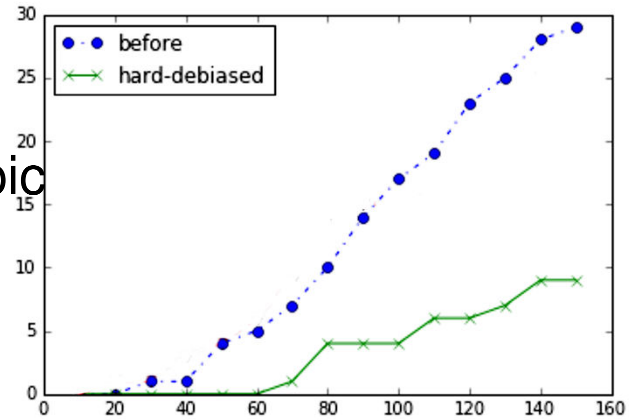
appropriate analogies



analogies generated

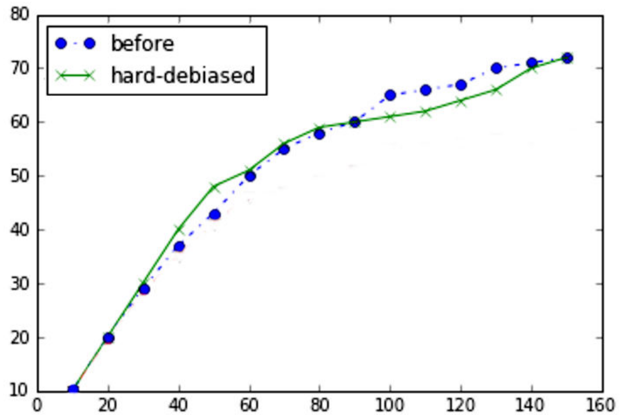
Debiasing results: analogies

stereotypic analogies



analogies generated

appropriate analogies



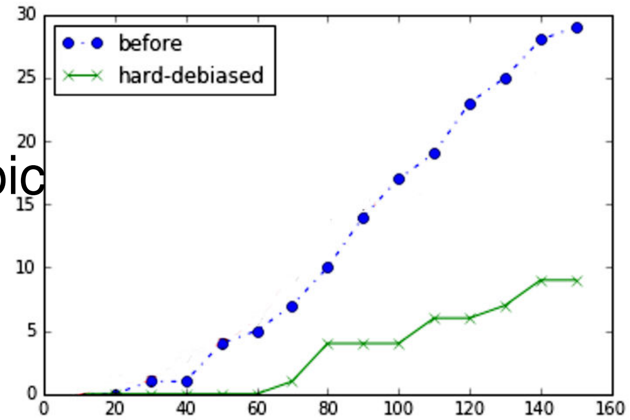
analogies generated

	RG	WS	analogy
Before	62.3	54.5	57.0
Hard-debiased	62.4	54.1	57.0

Debiasing reduced stereotypic analogies while preserving the utilities of the embedding.

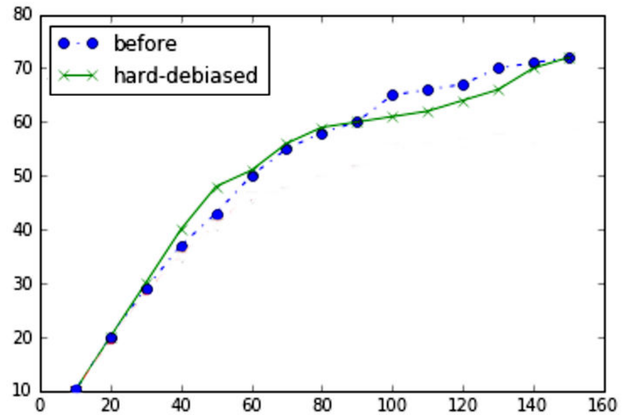
Debiasing results: analogies

stereotypic analogies



analogies generated

appropriate analogies



analogies generated

He : *King* :: She : *Queen*

He : *Doctor* :: She : *Doctor*

Debias embedding for sensitive applications

Paper: *Man is to computer programmer as woman is to homemaker? Debiasing word embeddings.* NIPS'16.

npr

He's Brilliant, She's Lovely: Teaching Computers To Be Less Sexist

August 12, 2016 · 8:01 AM ET

MIT Technology Review

Intelligent Machines

How to Fix Silicon Valley's Sexist Algorithms

MOTHERBOARD

RACISM

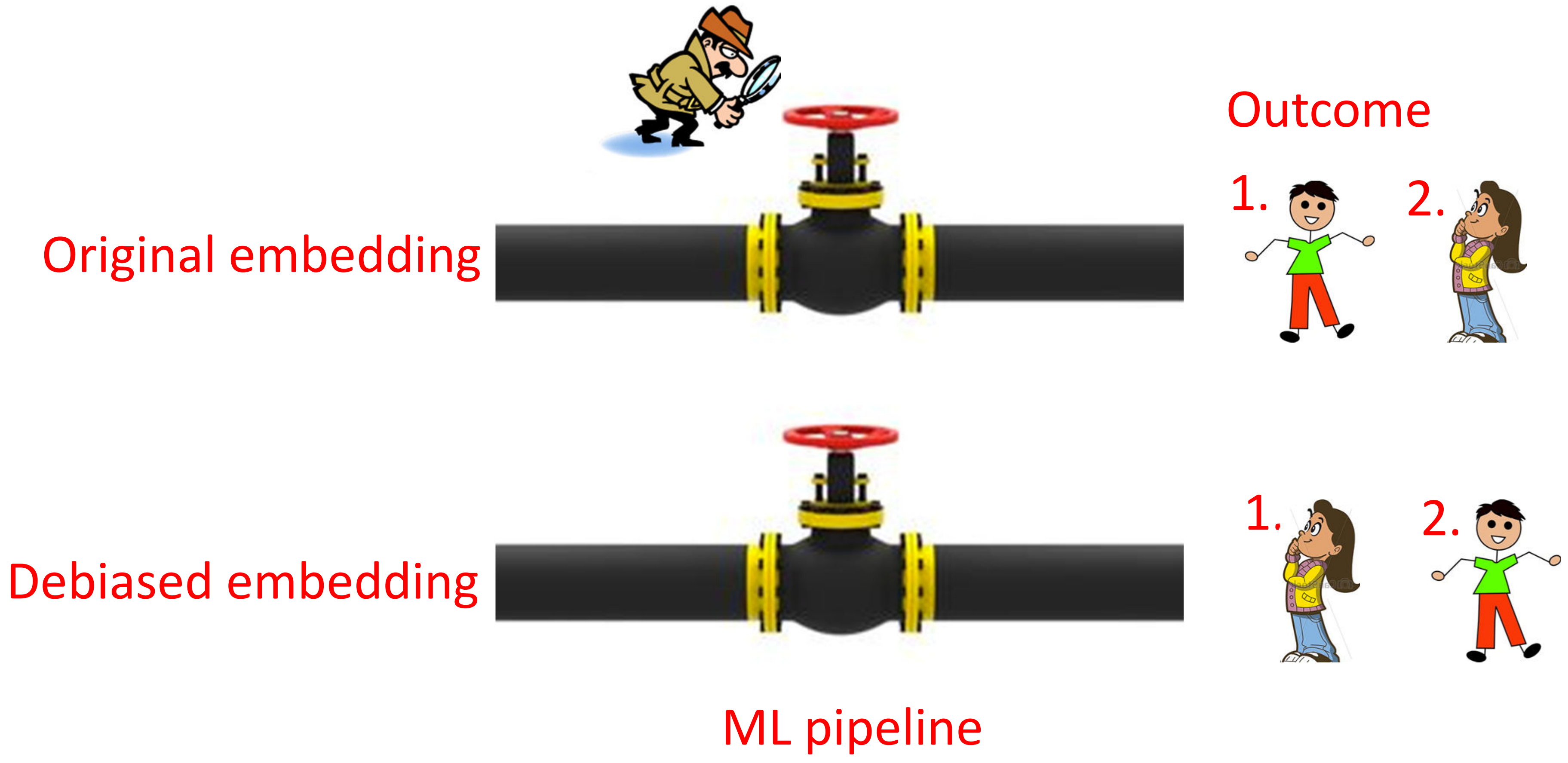
Machines Are Learning to Be Sexist Like Humans. Luckily, They're Easier to Fix.

硅谷的 AI 算法带有性别偏见，该如何修复它?

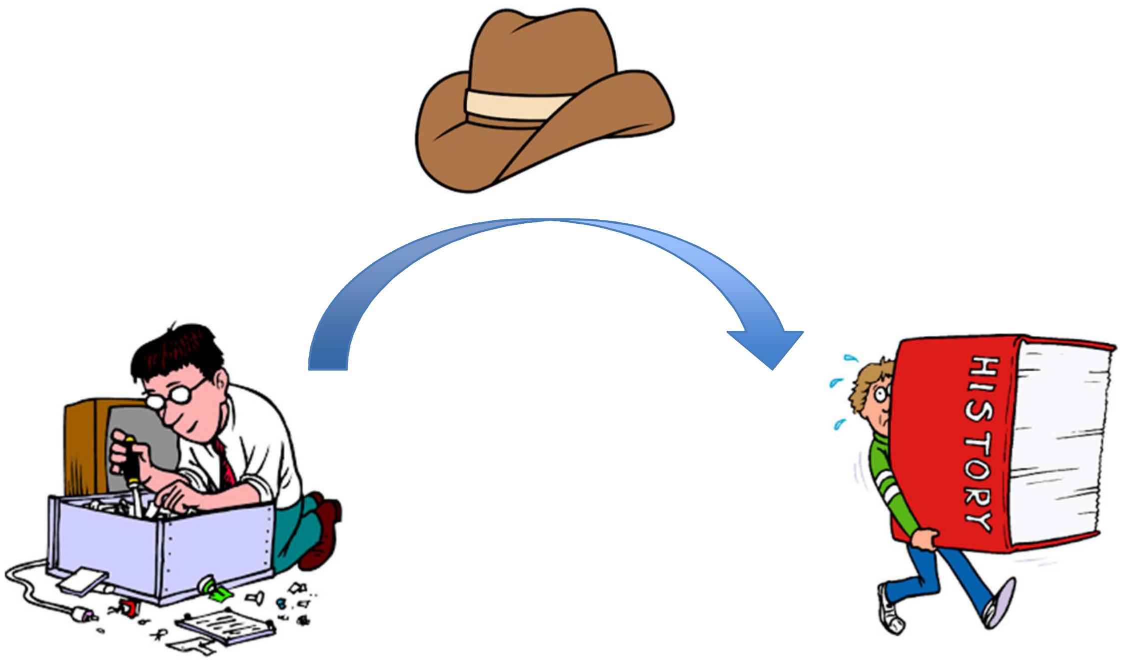
粹客网

Lazy coders are training artificial intelligences to be sexist

Use the debiased embedding to understand bias



Embedding as a lens to study history



Word embedding captures common stereotypes;
can we use this to study history?

100 years of word embeddings

Separate word embedding learned from each decade*

1910	1920	1930	1940	1950	1960	1970	1980	1990





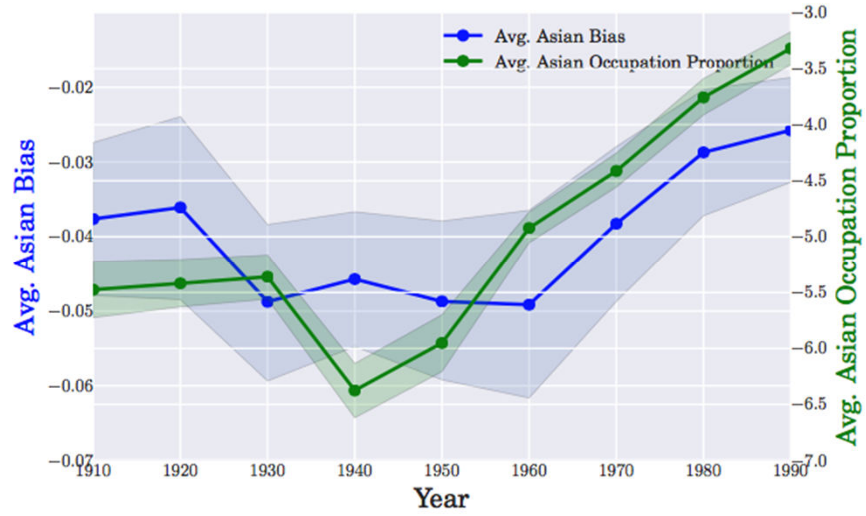
Integrate with U.S. Census and historical records

*Trained on Google books and Corpus of Historical American English.

Embedding captures Asian stereotypes

1910	1950	1990
irresponsible	disorganized	inhibited
envious	outrageous	passive
barbaric	pompous	dissolute
aggressive	unstable	haughty
transparent	effeminate	complacent
monstrous	unprincipled	forceful
hateful	venomous	fixed
cruel	disobedient	active
greedy	predatory	sensitive
bizarre	boisterous	hearty

Most Asian adjectives



Embedding Asian bias vs. census occupation

* Used U.S. census to quantify the average Asian participation in occupations.

Nikhil Garg

Embedding captures Asian stereotypes

1910	1950	1990
irresponsible	disorganized	inhibited
envious	outrageous	passive
barbaric	pompous	dissolute
aggressive	unstable	haughty
transparent	effeminate	complacent
monstrous	unprincipled	forceful
hateful	venomous	fixed
cruel	disobedient	active
greedy	predatory	sensitive
bizarre	boisterous	hearty

Most Asian adjectives

	1910	1920	1930	1940	1950	1960	1970	1980	1990
1910	1.00	0.69	0.61	0.59	0.57	0.48	0.45	0.37	0.33
1920	0.69	1.00	0.63	0.65	0.61	0.52	0.48	0.36	0.38
1930	0.61	0.63	1.00	0.65	0.58	0.48	0.51	0.40	0.40
1940	0.59	0.65	0.65	1.00	0.62	0.51	0.56	0.42	0.43
1950	0.57	0.61	0.58	0.62	1.00	0.58	0.52	0.45	0.39
1960	0.48	0.52	0.48	0.51	0.58	1.00	0.49	0.49	0.48
1970	0.45	0.48	0.51	0.56	0.52	0.49	1.00	0.48	0.43
1980	0.37	0.36	0.40	0.42	0.45	0.49	0.48	1.00	0.58
1990	0.33	0.38	0.40	0.43	0.39	0.48	0.43	0.58	1.00

1965 Immigration & Nationality Act; Asian immigration wave

Immigration growth slows; 2nd generation Asian Americans increase

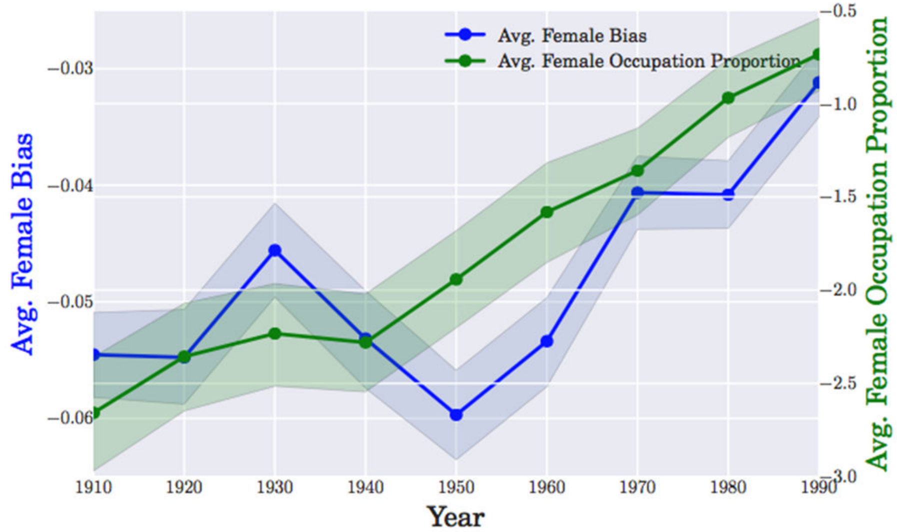
Correlation of embedding bias across decades

Garg, Schiebinger, Jurafsky, Zou *PNAS* 2018

Embedding as a lens to study history

1910	1950	1990
charming	delicate	maternal
placid	sweet	morbid
delicate	charming	artificial
passionate	transparent	physical
sweet	placid	caring
dreamy	childish	emotional
indulgent	soft	protective
playful	colorless	attractive
mellow	tasteless	soft
sentimental	agreeable	tidy

Most female adjectives



Embedding bias vs. census occupation*

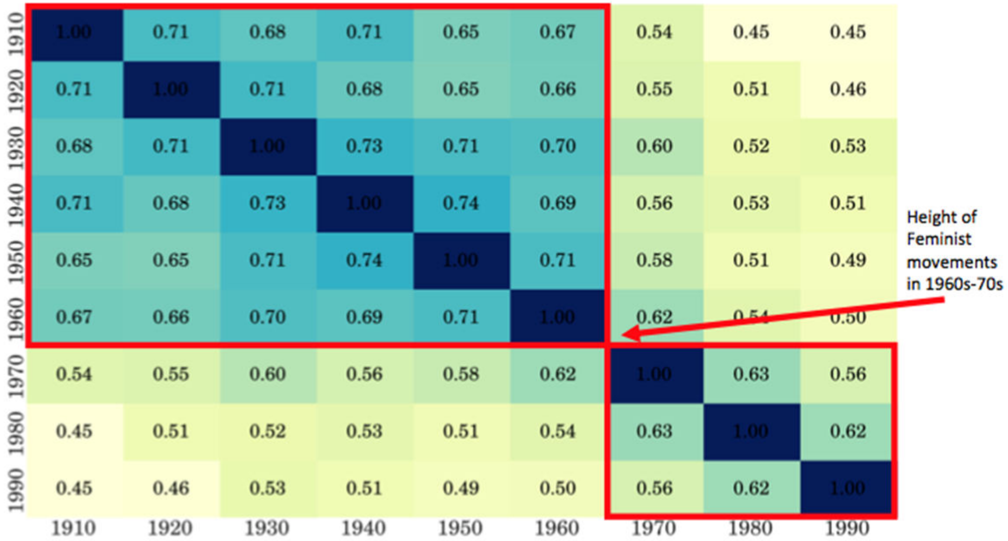
* Used U.S. census to quantify the average female participation in occupations.

Garg, Schiebinger, Jurafsky, Zou *PNAS* 2018

Embedding as a lens to study history

1910	1950	1990
charming	delicate	maternal
placid	sweet	morbid
delicate	charming	artificial
passionate	transparent	physical
sweet	placid	caring
dreamy	childish	emotional
indulgent	soft	protective
playful	colorless	attractive
mellow	tasteless	soft
sentimental	agreeable	tidy

Most female adjectives



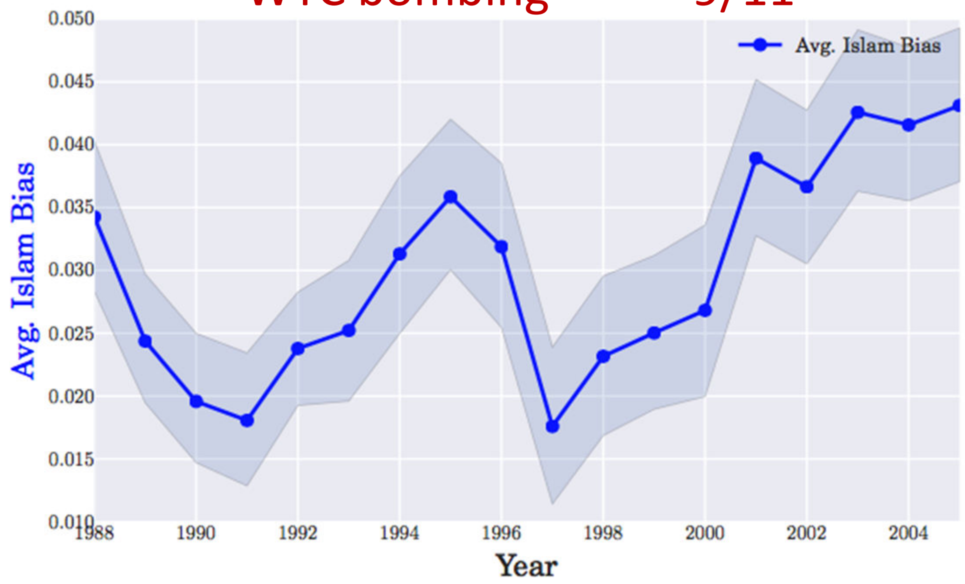
Correlation of embedding bias across decades

* Used U.S. census to quantify the average female participation in occupations.

Garg, Schiebinger, Jurafsky, Zou *PNAS* 2018

Embedding as a lens to study history

WTC bombing 9/11



Embedding trained on NY Times

Islam bias measures how close *Islam*, *mosque*, etc. are to words such as *terror*, *bomb*, *violence*.

Garg, Schiebinger, Jurafsky, Zou *PNAS* 2018

Discussion

- Geometry captures bias.
- Who's responsible: data, algorithm or user?
- Using debiased embedding for sensitive applications.
- Word embedding as a lens to study historical trends.

Papers:
Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. NIPS'16

Word embeddings as a lens to quantify 100 years of gender and ethnic stereotypes. PNAS'18

Thanks!

- Geometry captures bias.
- Who’s responsible: data, algorithm or user?
- Using debiased embedding for sensitive applications.
- Word embedding as a lens to study historical trends.

Collaborators: T. Bolukbasi, K. Chang, V. Saligrama, A. Kalai and N. Gar