A Visual Tour of Bias Mitigation Techniques for Word Representations

Archit Rathore, Sunipa Dev
Jeff M. Phillips, Vivek Srikumar, Bei Wang

Trigger warning
This tutorial contains examples of stereotypes seen in society and in language representations that could be potentially triggering.

Tutorial Website & Tool Download

https://www.sci.utah.edu/~beiwang/aaaibias2021

https://github.com/tdavislab/visualizing-bias





Archit Rathore, Sunipa Dev, Jeff M. Phillips, Vivek Srikumar, Bei Wang

Introduction

Representing meaning of words

What do words mean? How do they get their meaning?

cat

dog

tiger

table

Perhaps more pertinent for language technology

How can we represent the meaning of words in a form that is computationally flexible?

The company words keep

The Distributional Hypothesis: Words that occur in the same contexts have similar meanings (e.g. Zellig Harris, J.R. Firth)

Firth (1957): “*You shall know a word by the company it keeps*”

The key idea: To characterize the meaning of a word, we need to characterize the distribution of its context

What context? Commonly interpreted as neighboring words in text, but could be syntactic, semantic, discourse, pragmatic,...

Symbolic vs. Distributed representations

The strings `cat`, `tiger`, `dog` and `table` are symbols

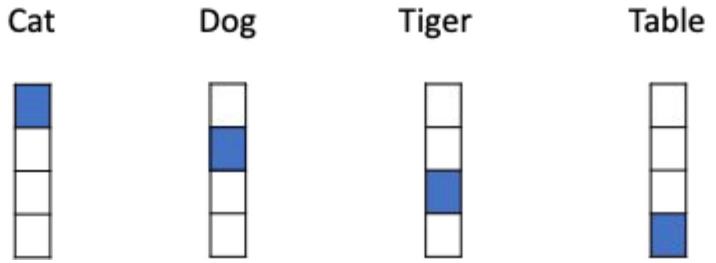
Just knowing the symbols does not tell us anything about what they mean.

1. `cat` and `tiger` are conceptually closer to each other than to `dog` or `table`
2. `cat`, `tiger` and `dog` are closer to each other than `table`

We need a representation that captures similarities between similar objects

Symbolic vs. Distributed representations

Think about feature representations



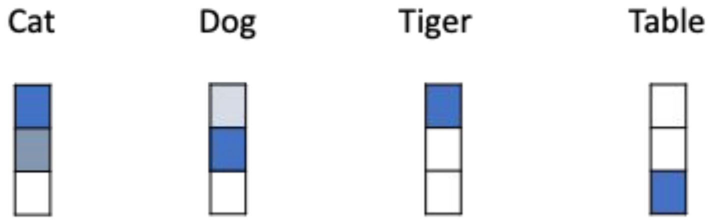
These ***one-hot vectors*** do not capture inherent similarities

Distances or dot products are all equal

Symbolic vs. Distributed representations

Distributed representations capture concept similarities better

Vector valued representations that coalesce superficially distinct concepts

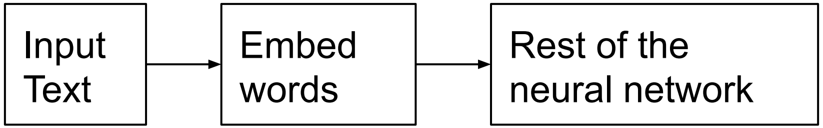


Word embeddings (or word vectors)

A mapping from words to a vector space could be:

- A fixed mapping, context independent vectors
  - Word2vec [Mikolov et al 2013], Glove [Pennington et al 2014], fastText [Joulin et al 2016]
- A parameterized mapping that produces context dependent vectors
  - ELMo [Peters et al 2018], BERT [Devlin et al 2019], RoBERTa [Chen et al 2019], etc

The first step in any neural network model for textual inputs today



Perspectives on word embeddings

1. They capture distributional semantics

   Embeddings are low dimensional vectors that are constructed by appealing to the distributional hypothesis

2. They are distributed representations of words

   The embedding dimensions represent underlying aspects of meaning, and words are characterized by membership to these latent dimensions

3. They provide features

   Word embeddings are a widely-used, convenient *learned* feature representations.

How are word embeddings trained?

Various approaches, but the common themes include:

1. Using massive unlabeled text corpora
2. Setting up a surrogate learning task that (a) does not require labeled data, and (b) produces embeddings *as a side effect*

Example: For the text

`"It was a dark and ______ night and …"`

1. Define a neural network of the form

$$P(\_\_\_\_ = \mathbf{x}) = f(Embedding[\mathbf{x}], Embedding[\text{context}])$$

2. Find embeddings that the probability for the hidden word being `stormy`

Evaluating word embeddings: Two broad approaches

1. Intrinsic evaluation: Evaluate the representation directly without training another model

2. Extrinsic evaluation: Evaluate the impact of the representation on another task

Evaluating word embeddings: Two broad approaches

1. Intrinsic evaluation: Evaluate the representation directly without training another model
   a. Typically simple tasks where success or failure is (almost) entirely a function of the representation
   b. Easy to compute, but doesn't say much about the embeddings as features

2. Extrinsic evaluation: Evaluate the impact of the representation on another task

Evaluating word embeddings: Two broad approaches

1. Intrinsic evaluation: Evaluate the representation directly without training another model
    a. Typically simple tasks where success or failure is (almost) entirely a function of the representation
    b. Easy to compute, but doesn’t say much about the embeddings as features

2. Extrinsic evaluation: Evaluate the impact of the representation on another task
    a. Typically, a neural network
    b. Can be more practically useful, but slow and depends on the quality of the model for the task being tested

Example intrinsic evaluation: Word Analogies

Complete a word analogy puzzle using the embeddings

`Queen : King :: Tigress : ?`

Example intrinsic evaluation: Word Analogies

Complete a word analogy puzzle using the embeddings

`Queen` : `King` :: `Tigress` : ?

Given word embeddings, one way to answer the question "`a` : `b` :: `c` : ?" is

$$\arg\max_d \frac{(x_a - x_b + x_c)^T x_d}{\|x_a - x_b + x_c\|}$$

Effectively finds the word such that

$$x_a - x_b \approx x_c - x_d$$

Word embeddings are great, but...

Societal biases in word embeddings

If word embeddings capture distributional information from corpora…

… and corpora possess societal stereotypes, then

the trained word embeddings may encode these stereotypes

“Feeding AI systems on the world’s beauty, ugliness, and cruelty, but expecting it to reflect only the beauty is a fantasy.”

Birhane and Prabhu (2021). "Large Image Datasets: A Pyrrhic Win for Computer Vision?", paraphrasing Ruha Benjamin (2019)

“Bias” in language technology

A fast moving field, with new techniques and perspectives being introduced almost every month

Two related lines of work:

1. New methods for quantifying biases encoded in embeddings
2. Methods for removing biases from embeddings

This tutorial: A visual exploration of debiasing

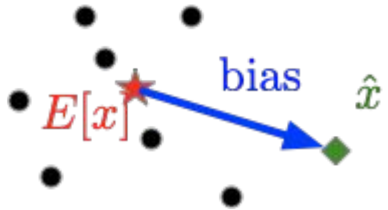
1. Biases and debiasing
   a. The various notions of bias in embeddings
   b. Measuring bias in embeddings (intrinsic and extrinsic methods)
   c. How can we attenuate bias in word embeddings? An overview of methods
2. A hands on exploration of bias
   a. A new tool for visualizing word embedding biases
   b. A visual exploration of the debiasing methods: Worked examples
3. Critiques of debiasing methods
4. Discussion

Notions of Bias

What is “bias”?

Def: difference between an estimator and its expected value

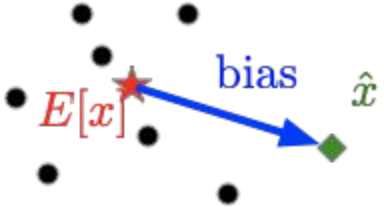
$$\hat{x} - E[x]$$



What is “bias”?

Def: difference between an estimator and its expected value

$$\hat{x} - E[x]$$

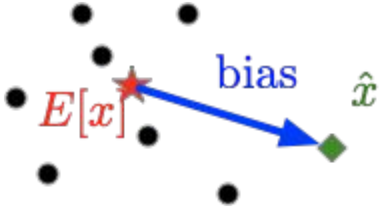


Def: an instance of prejudice, especially a personal and sometimes unreasonable outlook

What is “bias”?

Def: difference between an estimator and its expected value

$$\hat{x} - E[x]$$



Def: an instance of prejudice, especially a personal and sometimes unreasonable outlook

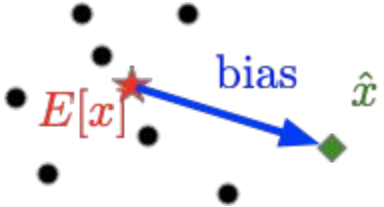
→ In machine learning .. a stereotype

Def: an oversimplified view or prejudiced attitude of a particular type of person or thing

What is “bias”?

Def: difference between an estimator and its expected value

$$\hat{x} - E[x]$$



Def: an instance of prejudice, especially a personal and sometimes unreasonable outlook

→ In machine learning .. a stereotype

Def: an oversimplified view or prejudiced attitude of a particular type of person or thing
an **oversimplification** of a **concept**

What is bias and a stereotype

An oversimplification of a concept

Ex: children are curious

Ex: dogs are friendly

Ex: nurses are women and doctors are men

Often a **negative** connotation

Harms

Kate Crawford’s NeurIPS 2017 Keynote (https://www.youtube.com/watch?v=fMym_BKWQzk)

Harms

Kate Crawford’s NeurIPS 2017 Keynote (https://www.youtube.com/watch?v=fMym_BKWQzk)

- Allocational Harms

Harms

Kate Crawford's NeurIPS 2017 Keynote (https://www.youtube.com/watch?v=fMym_BKWQzk)

- Allocational Harms
  - College acceptance
  - Bank loan applications
  - Recidivism prediction and parole

Harms

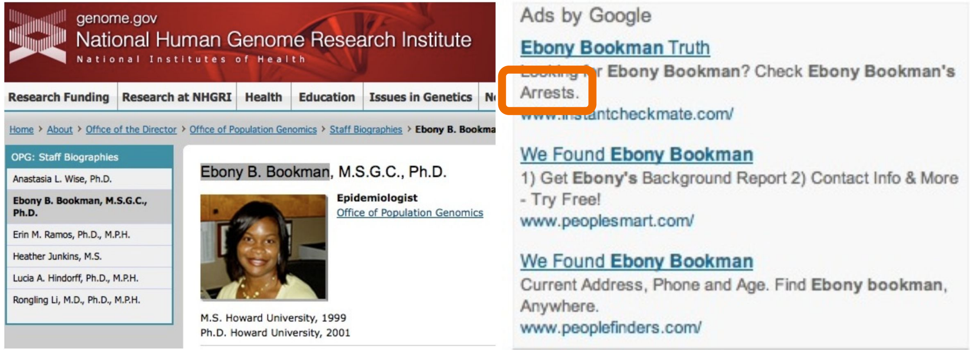
Kate Crawford's NeurIPS 2017 Keynote (https://www.youtube.com/watch?v=fMym_BKWQzk)

- Allocational Harms
  - College acceptance
  - Bank loan applications
  - Recidivism prediction and parole
  - Did your paper get into AAAI?

Harms

Kate Crawford’s NeurIPS 2017 Keynote (https://www.youtube.com/watch?v=fMym_BKWQzk)

- Allocational Harms
  - College acceptance
  - Bank loan applications
  - Recidivism prediction and parole
  - Did your paper get into AAAI?



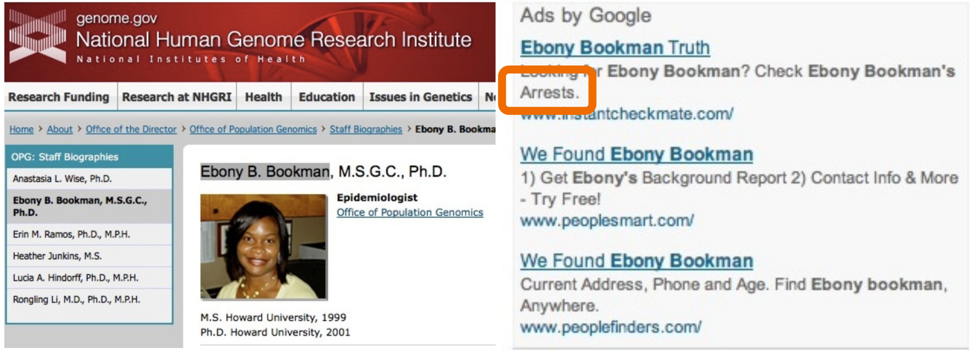
- Representational Harms
  - More subtle. How data is represented which leads to negative stereotypes / bias

Sweeney; Discreimination in Online Ad Delivery. CACM 2013

Harms

Kate Crawford’s NeurIPS 2017 Keynote (https://www.youtube.com/watch?v=fMym_BKWQzk)

- Allocational Harms
  - College acceptance
  - Bank loan applications
  - Recidivism prediction and parole
  - Did your paper get into AAAI?
- Representational Harms
  - More subtle. How data is represented which leads to negative stereotypes / bias
  - … but knowledge representation is a big part of AI



Sweeney; Discreimination in Online Ad Delivery. CACM 2013

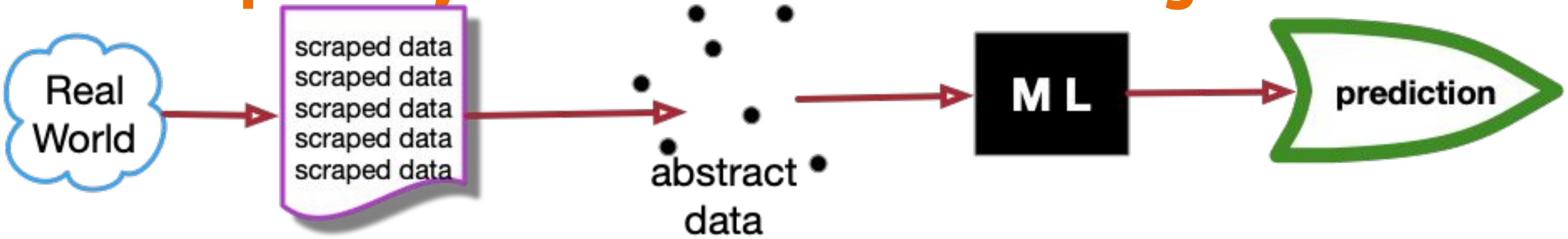
Bias + Machine Learning

Given bias

- Choice of data
- Mechanism to represent data
- Choice of learning model / algorithm

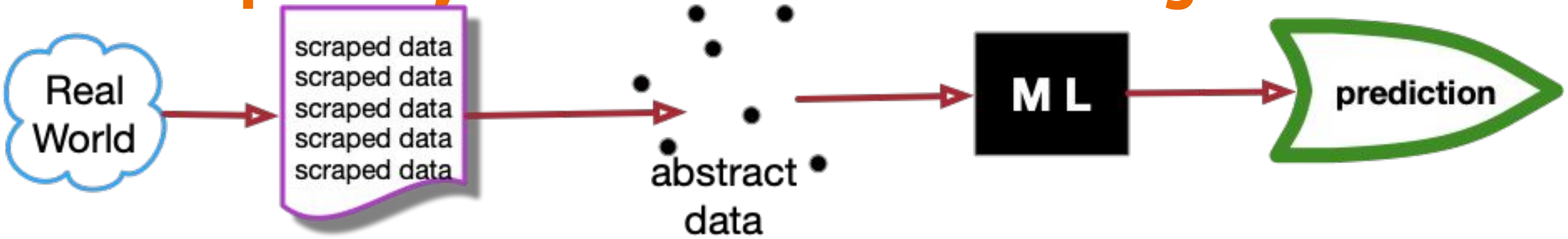
… can translate into representational or allocational harm

How to quantify bias in machine learning

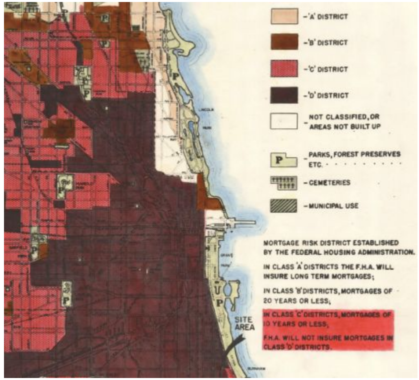


→ hard to quantify it exists (but has been done, it does exist)

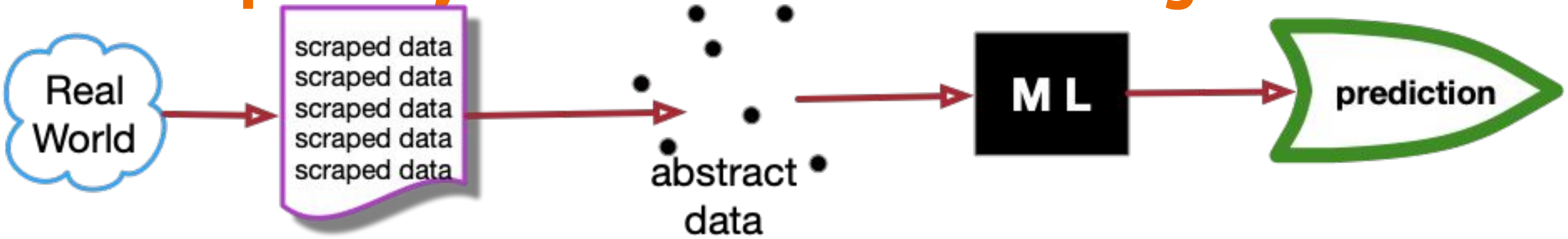
How to quantify bias in machine learning



→ hard to quantify it exists (but has been done, it does exist)
Documented examples (pro-publica, red-lining, ...)



How to quantify bias in machine learning

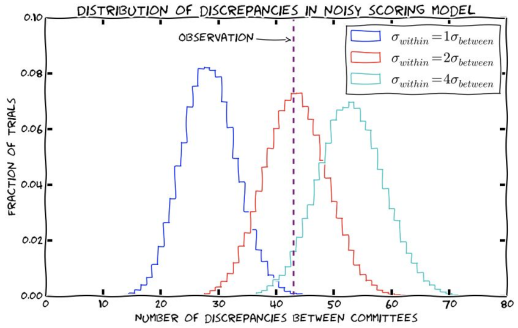


→ hard to quantify it exists (but has been done, it does exist)

Documented examples (pro-publica, red-lining, ...)

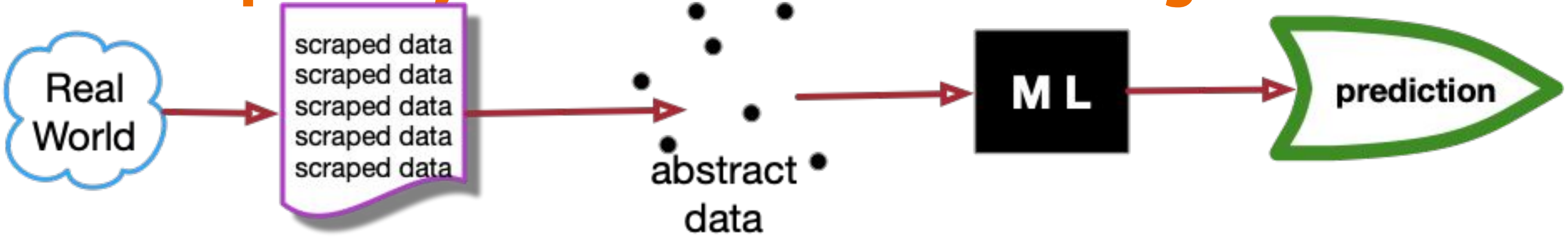
Nebulous examples (non-blind paper acceptance, policing, ...)

... harder because of potential confounding factors



Ensign et al; Runaway Feedback Loops in Predictive Policing. FAT* 2018

How to quantify bias in machine learning



→ hard to quantify it exists (but has been done, it does exist)

Documented examples (pro-publica, red-lining, ...)

Nebulous examples (non-blind paper acceptance, policing, ...)

... harder because of potential confounding factors

→ can quantify allocational harms **exist**,

but hard to quantify its true source

How to quantify bias in machine learning

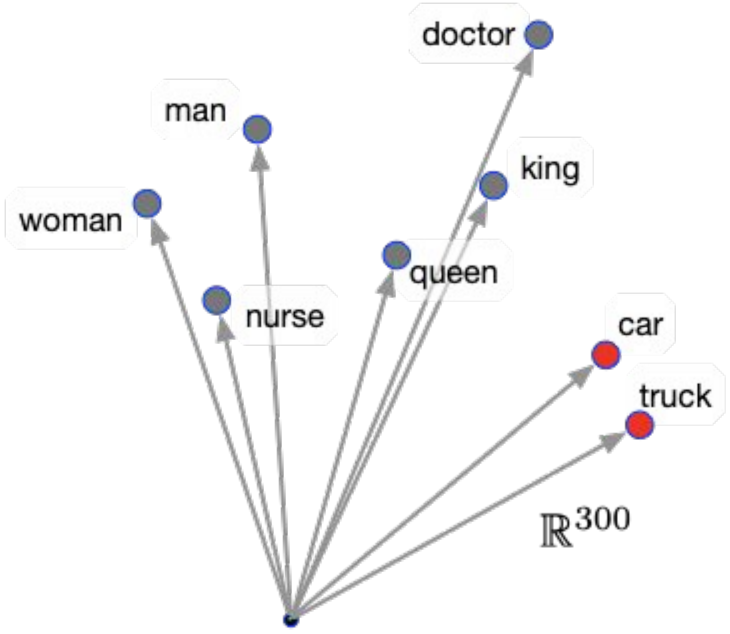
- Proxy downstream tasks
    - Simple and controlled
    - Millions of evaluations

How to quantify bias in machine learning

- Proxy downstream tasks
  - Simple and controlled
  - Millions of evaluations
- Inspecting representations
  - Direct representation harms
  - Specifically word vector embeddings

Inspecting Representations

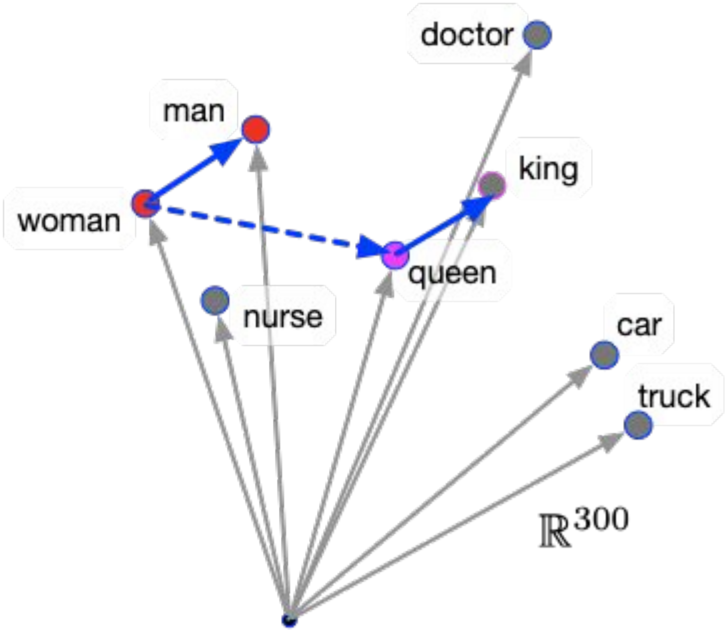
Similarity Tests



Inspecting Representations

Similarity Tests

Analogies



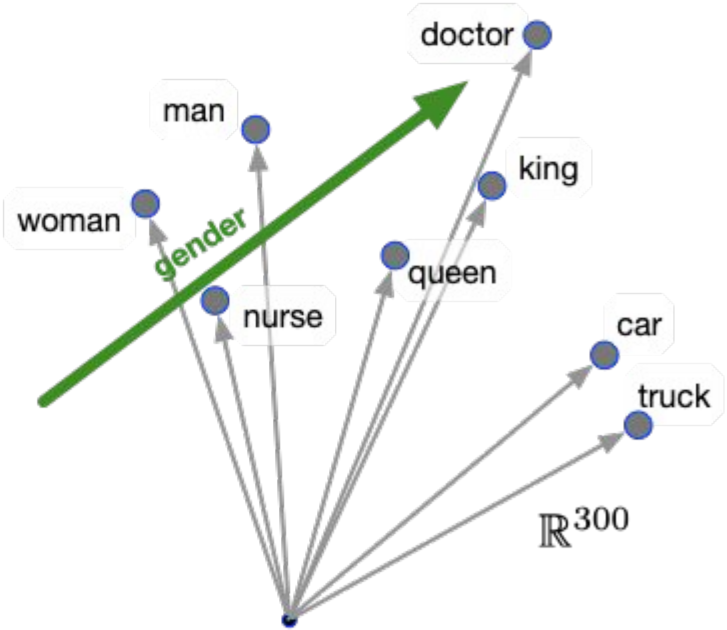
Bolukbasi et al; Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. NeurIPS 2016

Inspecting Representations

Similarity Tests

Analogies

Concept Subspace



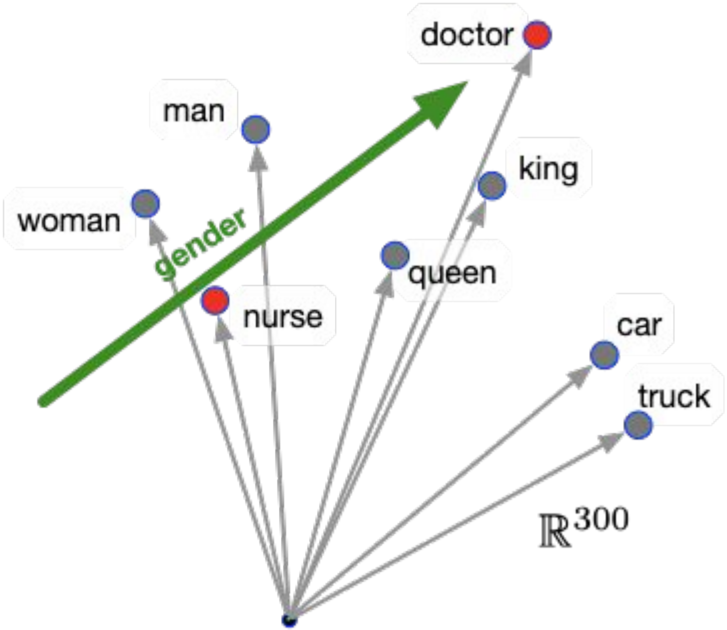
Bolukbasi et al; Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. NeurIPS 2016

Inspecting Representations

Similarity Tests

Analogies

Concept Subspace



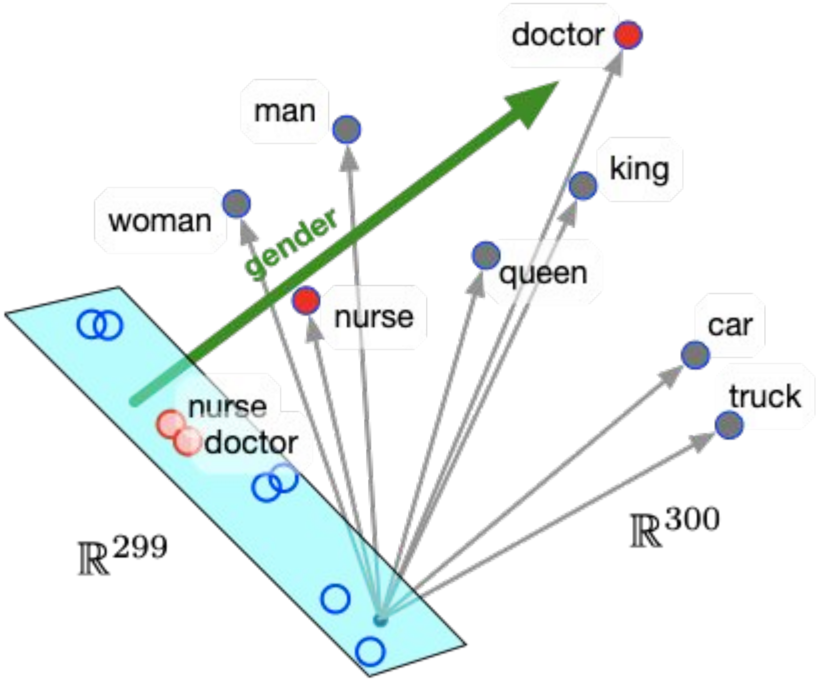
Bolukbasi et al; Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. NeurIPS 2016

Inspecting Representations

Similarity Tests

Analogies

Concept Subspace



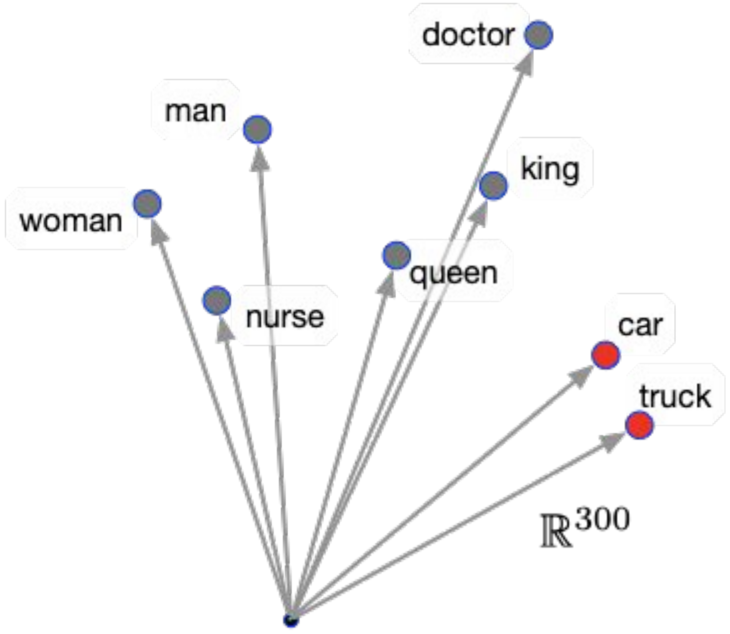
Inspecting Representations

Similarity Tests

Analogies

Concept Subspace

WEAT (implicit gender association stereotypes)



Caliskan et al; Semantics derived automatically from language corpora contain human like biases. Science 2017

Inspecting Representations

Similarity Tests

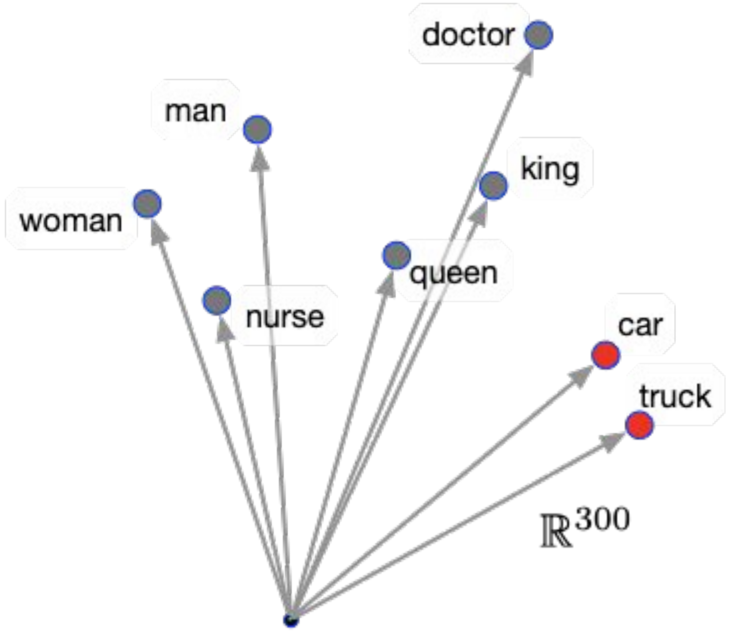
Analogies

Concept Subspace

WEAT (implicit gender association stereotypes)

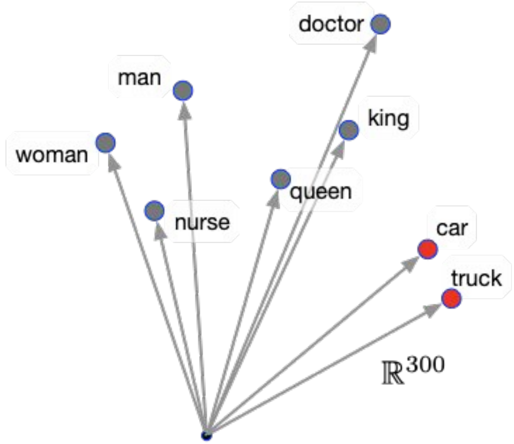
ECT, others

[aggregate results on full data]



WEAT Implicit Association Test

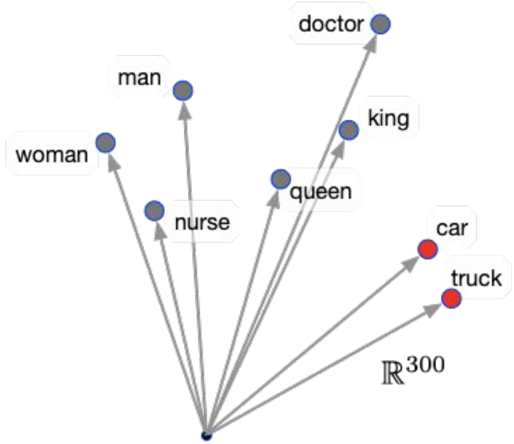
X = {man, male, ...} (definitionally male words)
Y = {woman, female, ...} (definitionally female words)



Caliskan et al; Semantics derived automatically from language corpora contain human like biases. Science 2017

WEAT Implicit Association Test

X = {man, male, ...} (definitionally male words)
Y = {woman, female, ...} (definitionally female words)
A = {programmer, engineer, scientist, ...} (stereotypical male professions)
B = {nurse, teacher, librarian, ...} (stereotypical female professions)



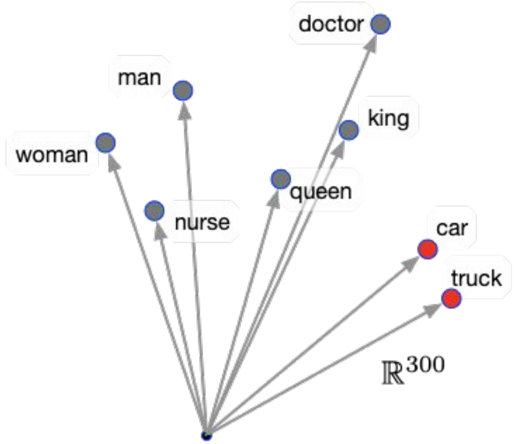
Caliskan et al; Semantics derived automatically from language corpora contain human like biases. Science 2017

WEAT Implicit Association Test

X = {man, male, ...} (definitionally male words)
Y = {woman, female, ...} (definitionally female words)
A = {programmer, engineer, scientist, ...} (stereotypical male professions)
B = {nurse, teacher, librarian, ...} (stereotypical female professions)

$$s(w, A, B) = \frac{1}{|A|} \sum_{a \in A} \cos(a, w) - \frac{1}{|B|} \sum_{b \in B} \cos(b, w)$$

association of gendered word w with sets A,B



Caliskan et al; Semantics derived automatically from language corpora contain human like biases. Science 2017

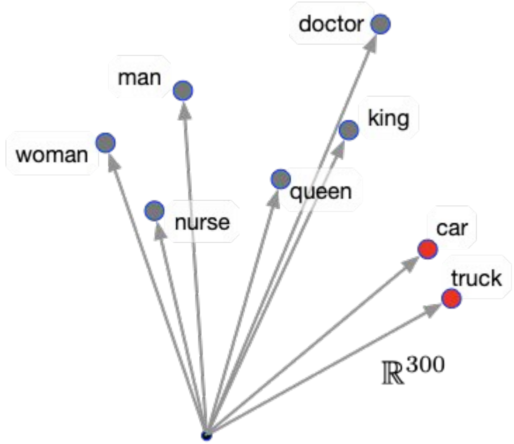
WEAT Implicit Association Test

X = {man, male, ...} (definitionally male words)
Y = {woman, female, ...} (definitionally female words)
A = {programmer, engineer, scientist, ...} (stereotypical male professions)
B = {nurse, teacher, librarian, ...} (stereotypical female professions)

$$s(w, A, B) = \frac{1}{|A|} \sum_{a \in A} \cos(a, w) - \frac{1}{|B|} \sum_{b \in B} \cos(b, w)$$

association of gendered word w with sets A,B

$$S(X, Y, A, B) = \frac{1}{|X|} \sum_{x \in X} s(x, A, B) - \frac{1}{|Y|} \sum_{y \in Y} s(y, A, B)$$



Caliskan et al; Semantics derived automatically from language corpora contain human like biases. Science 2017

WEAT Implicit Association Test

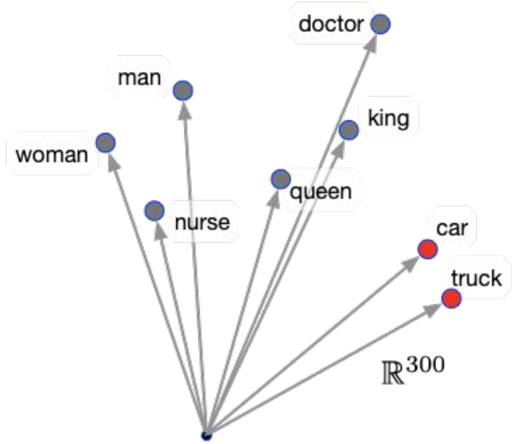
X = {man, male, ...} (definitionally male words)
Y = {woman, female, ...} (definitionally female words)
A = {programmer, engineer, scientist, ...} (stereotypical male professions)
B = {nurse, teacher, librarian, ...} (stereotypical female professions)

$$s(w, A, B) = \frac{1}{|A|} \sum_{a \in A} \cos(a, w) - \frac{1}{|B|} \sum_{b \in B} \cos(b, w)$$

association of gendered word w with sets A,B

$$S(X, Y, A, B) = \frac{1}{|X|} \sum_{x \in X} s(x, A, B) - \frac{1}{|Y|} \sum_{y \in Y} s(y, A, B)$$

S in [-2,2]. Neutral *should* be **0**. Word2Vec = **1.89**; GloVe **1.81**



ECT : Embedding Coherence Test

X = {man, male, ...} (definitionally male words)
Y = {woman, female, ...} (definitionally female words)
A = {programmer, engineer, scientist, ...} (stereotypical male professions)
B = {nurse, teacher, librarian, ...} (stereotypical female professions)

Create $\bar{x} = \frac{1}{|X|} \sum_{x \in X} x$ and $\bar{y} = \frac{1}{|Y|} \sum_{y \in Y} y$.

Dev and Phillips; Attenuating Bias in Word Vectors. AIStats 2019

ECT : Embedding Coherence Test

X = {man, male, ...} (definitionally male words)
Y = {woman, female, ...} (definitionally female words)
A = {programmer, engineer, scientist, ...} (stereotypical male professions)
B = {nurse, teacher, librarian, ...} (stereotypical female professions)

Create $\bar{x} = \frac{1}{|X|} \sum_{x \in X} x$ and $\bar{y} = \frac{1}{|Y|} \sum_{y \in Y} y$.

Determine rank order $O_X = \cos(\bar{x}, p_i) \geq \cos(\bar{x}, p_j) \geq \ldots$ for all $p \in A \cup B$ and

$$O_Y = \cos(\bar{y}, p_{i'}) \geq \cos(\bar{y}, p_{j'}) \geq \ldots$$

Dev and Phillips; Attenuating Bias in Word Vectors. AIStats 2019

ECT : Embedding Coherence Test

X = {man, male, ...} (definitionally male words)
Y = {woman, female, ...} (definitionally female words)
A = {programmer, engineer, scientist, ...} (stereotypical male professions)
B = {nurse, teacher, librarian, ...} (stereotypical female professions)

Create $\bar{x} = \frac{1}{|X|}\sum_{x \in X} x$ and $\bar{y} = \frac{1}{|Y|}\sum_{y \in Y} y$.

Determine rank order $O_X = \cos(\bar{x}, p_i) \geq \cos(\bar{x}, p_j) \geq \ldots$ for all $p \in A \cup B$ and
$O_Y = \cos(\bar{y}, p_{i'}) \geq \cos(\bar{y}, p_{j'}) \geq \ldots$

Return Spearman-Coefficient between O_X and O_Y
in $[-1, 1]$ with larger more correlated.

Glove: **0.798**

Dev and Phillips; Attenuating Bias in Word Vectors. AIStats 2019

Proxy Downstream tasks

From Natural Language Processing

- Coreference resolution (map pronoun “she” to “doctor”)?
  - Standard tasks are messy, involve many aspects

Proxy Downstream tasks

From Natural Language Processing

- Coreference resolution (map pronoun “she” to “doctor”)?
  - Standard tasks are messy, involve many aspects
- Natural Language Inference
  - MultiNLI (big, long sentences, but noisy)
  - SNLI (shorter sentences, concise)

Premise : a **doctor** bought a bagel

Hypothesis : a **woman** bought a bagel

Entailment	Neutral	Contradiction
0.87	0.11	0.02

Parikh et al; A decomposable attention model for natural language inference. EMNLP 2016

Proxy Downstream tasks

From Natural Language Processing

- Coreference resolution (map pronoun “she” to “doctor”)?
    - Standard tasks are messy, involve many aspects
- Natural Language Inference
    - MultiNLI (big, long sentences, but noisy)
    - SNLI (shorter sentences, concise)

Entailment	Neutral	Contradiction
0.87	0.11	0.02

Parikh et al; A decomposable attention model for natural language inference. EMNLP 2016

Premise : a **doctor** bought a bagel

Hypothesis 1: a **woman** bought a bagel — contradict w/p **0.91**

Hypothesis 2: a **man** bought a bagel — entails w/p **0.84**

NLI Templates

Premise : a **doctor** **bought** a **bagel**
Hypothesis 1: a **woman** **bought** a **bagel**
Hypothesis 2: a **man** **bought** a **bagel**

164 Occupations (e.g. **doctor**)
27 Verbs (e.g., **bought**)
184 Objects (e.g., **bagel**)
3 gendered word pairs (e.g., **man-woman**)

Dev *et.al.*; On Measuring and Mitigating Biased Inferences of Word Embeddings. AAAI 2020

NLI Templates

Premise : a **doctor** **bought** a **bagel**
Hypothesis 1: a **woman** **bought** a **bagel**
Hypothesis 2: a **man** **bought** a **bagel**

164 Occupations (e.g. **doctor**)
27 Verbs (e.g., **bought**)
184 Objects (e.g., **bagel**)
3 gendered word pairs (e.g., **man-woman**)

Entailment	Neutral	Contradiction
0.87	0.11	0.02

Statistics on results

net neutral = **average** **neutral** value on all 1.9M templates

frac neutral = fraction of 1.9M templates with **neutral** **>** **entail**, **contradict**

Dev *et.al.*; On Measuring and Mitigating Biased Inferences of Word Embeddings. AAAI 2020

Debiasing Methods for Word Embeddings

Sources of Bias

- Bias in data for training representations.
- Algorithmic bias.
- Bias in data for training specific tasks.

Debiasing word embeddings

- Data augmentation/balancing.
- Modifying embedding generating algorithm.
- Post-processing of embeddings.
- Additionally: debias/balance task specific data.

Data Balancing

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.

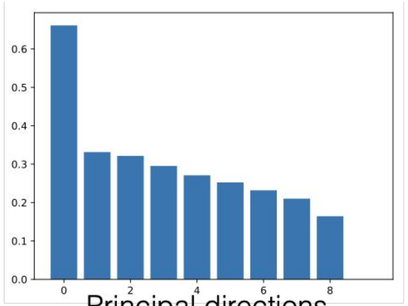
She was talking to the boy.

He was talking to the girl.

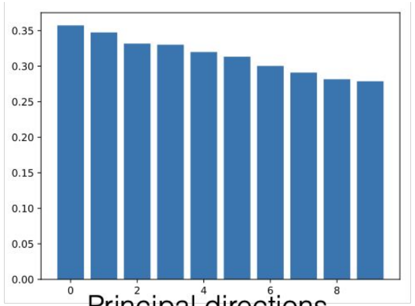
Dev and Phillips; Attenuating Bias in Word Vectors. AIStats 2019

Data Balancing

Principal Values

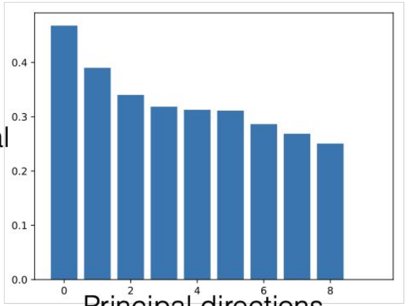


Principal directions
p = 0.0

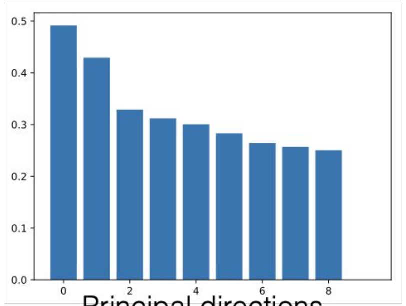


Principal directions
Random gaussian vectors

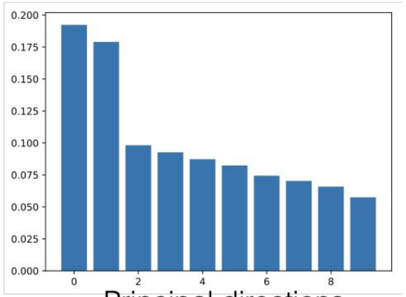
Principal Values



Principal directions
p = 0.50



Principal directions
p = 0.75



Principal directions
p = 1.0

Data Balancing

- Implicit residual bias still large - some cases worse
- Not easy to generalize
- Requires retraining - expensive!

Dev and Phillips; Attenuating Bias in Word Vectors. AIStats 2019

Gender Neutral GloVe

- Learns a protected attribute - gender - in specific dimensions and neutralizes everywhere else
- Not easy to generalize
- Requires retraining of whole embedding - expensive!

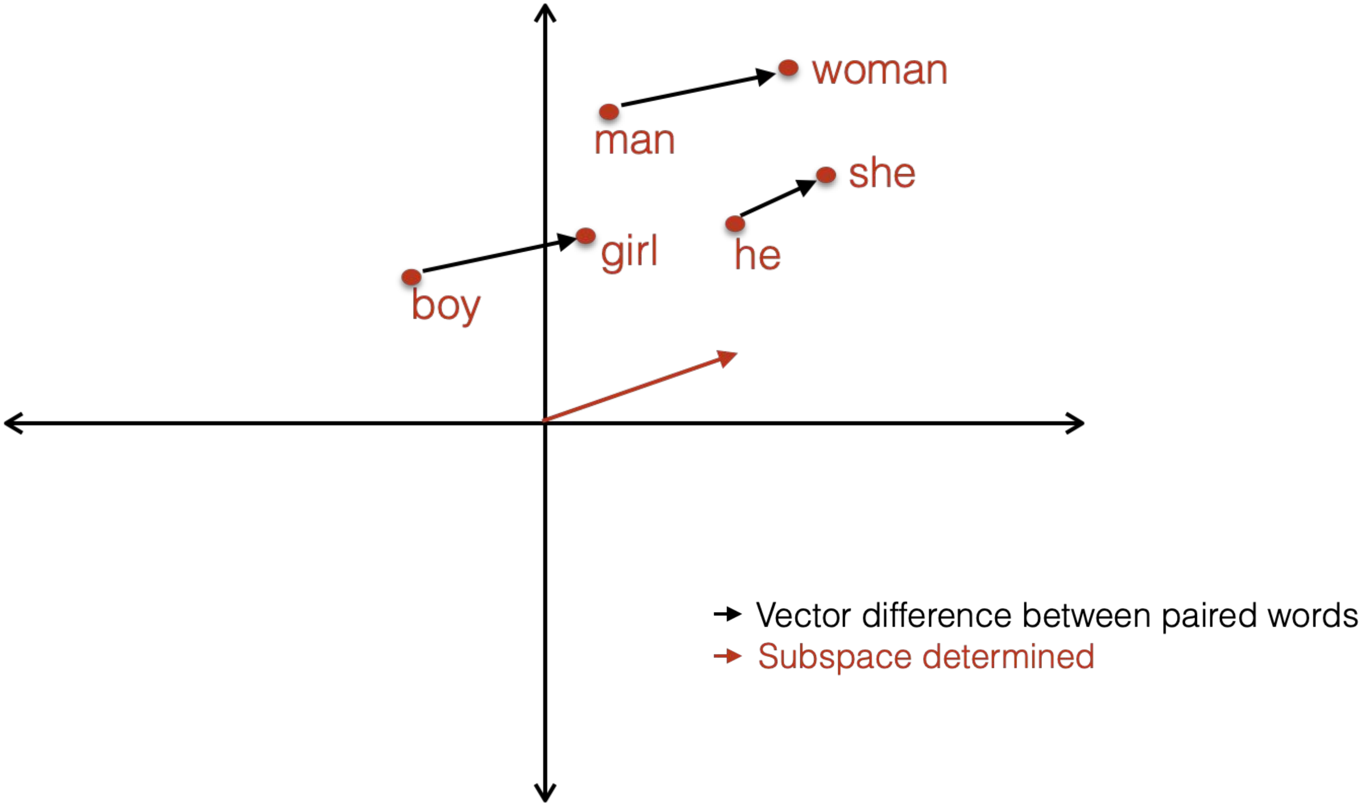
Zhao et al; Learning gender neutral word embeddings. EMNLP 2018

Debiasing by Post Processing Representations

- Modulates representations to mitigate stereotypical associations.
- Easy to extend to different biases.
- Inexpensive!

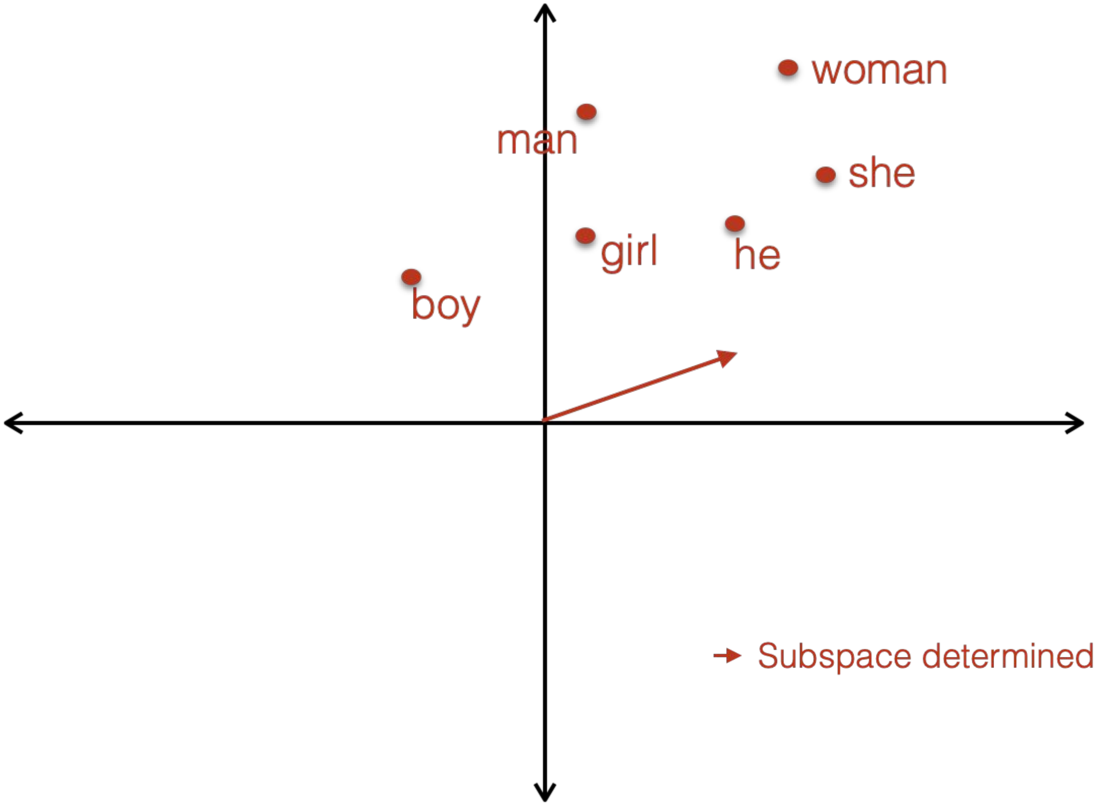
Feature Subspace Determination

PCA Paired

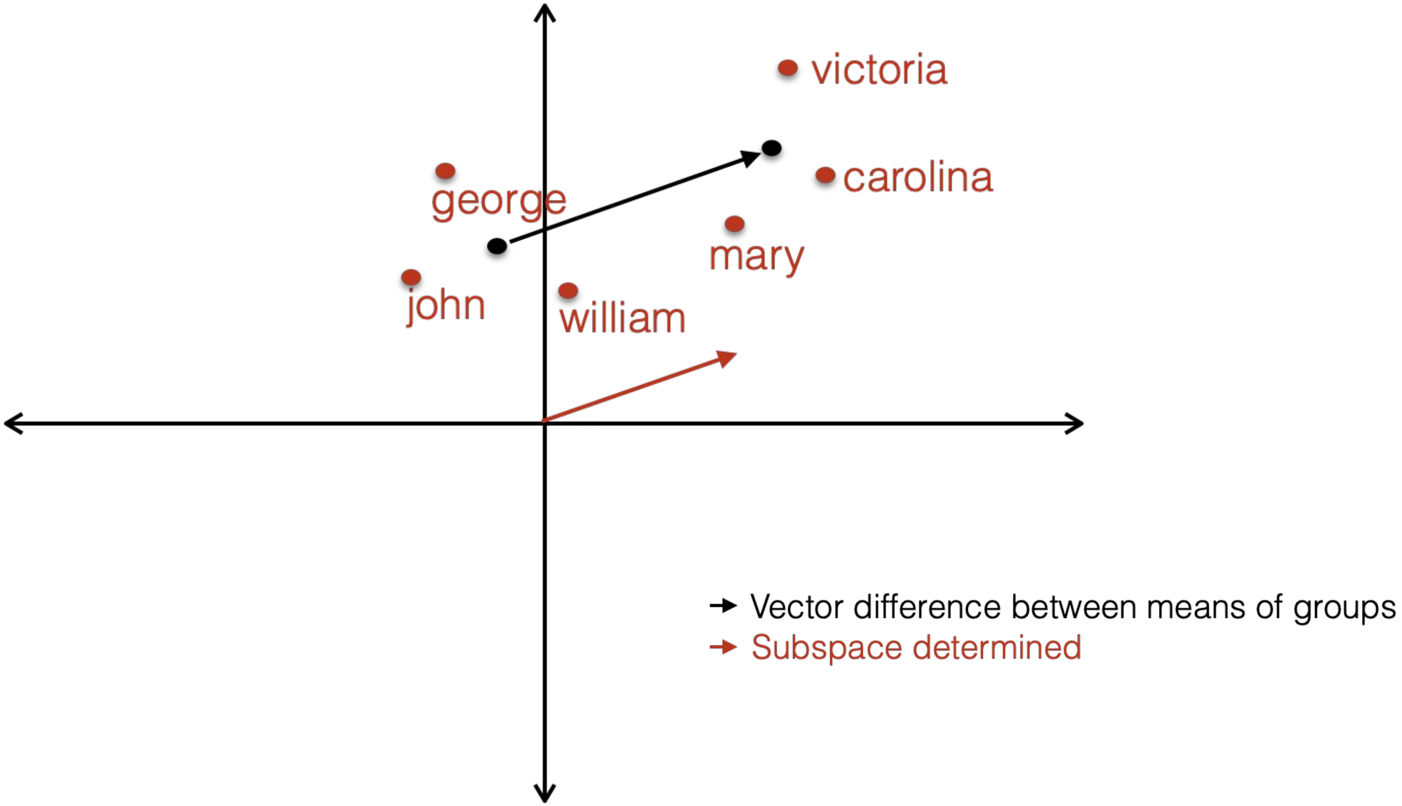


woman
man
she
girl
he
boy
Vector difference between paired words
Subspace determined

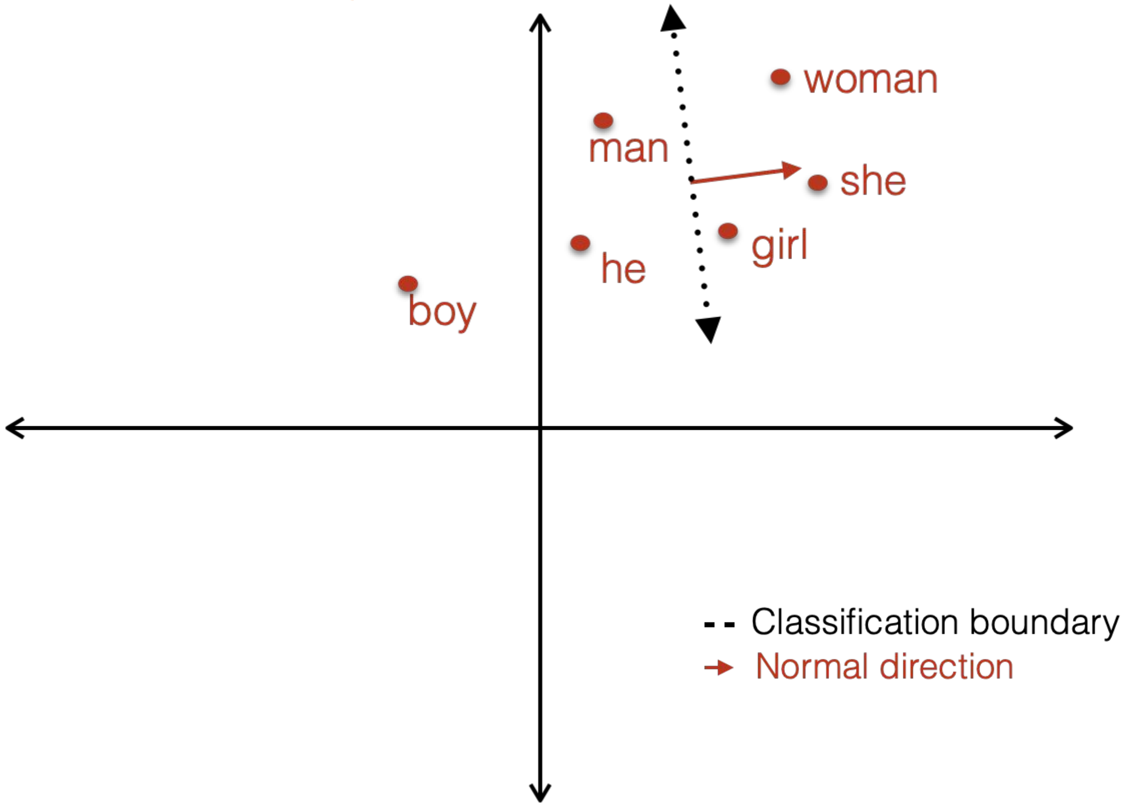
PCA



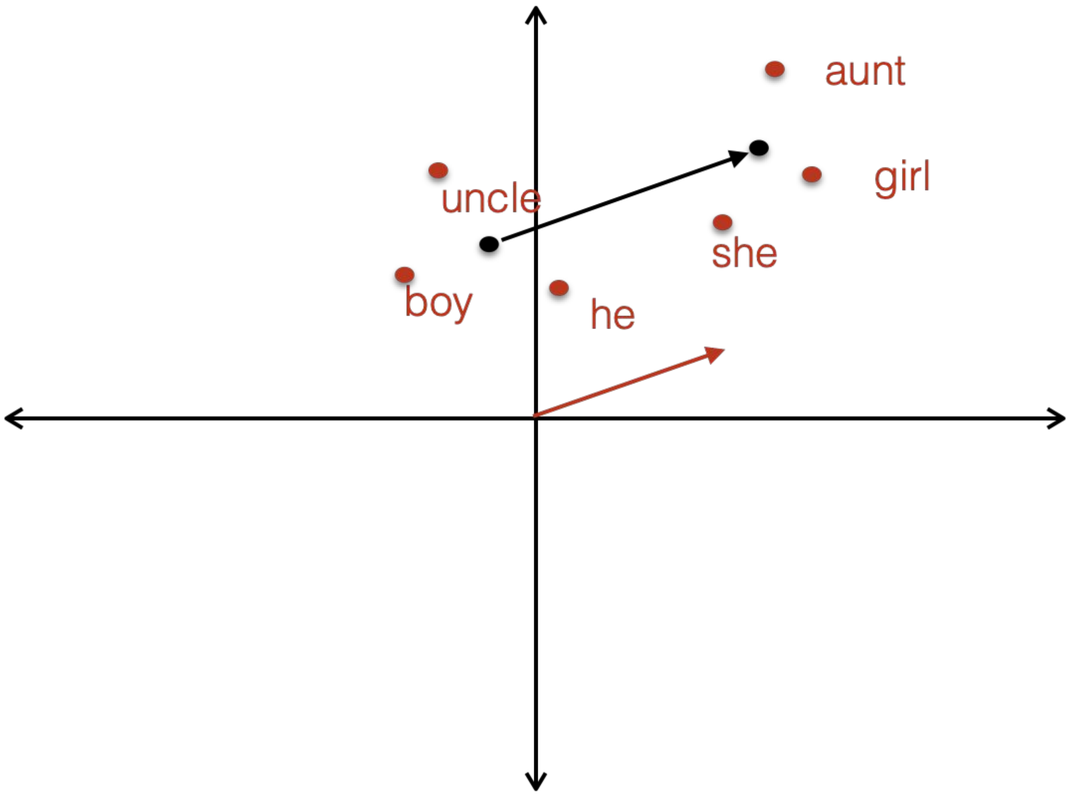
2 - Means Method



Classification Boundary Based

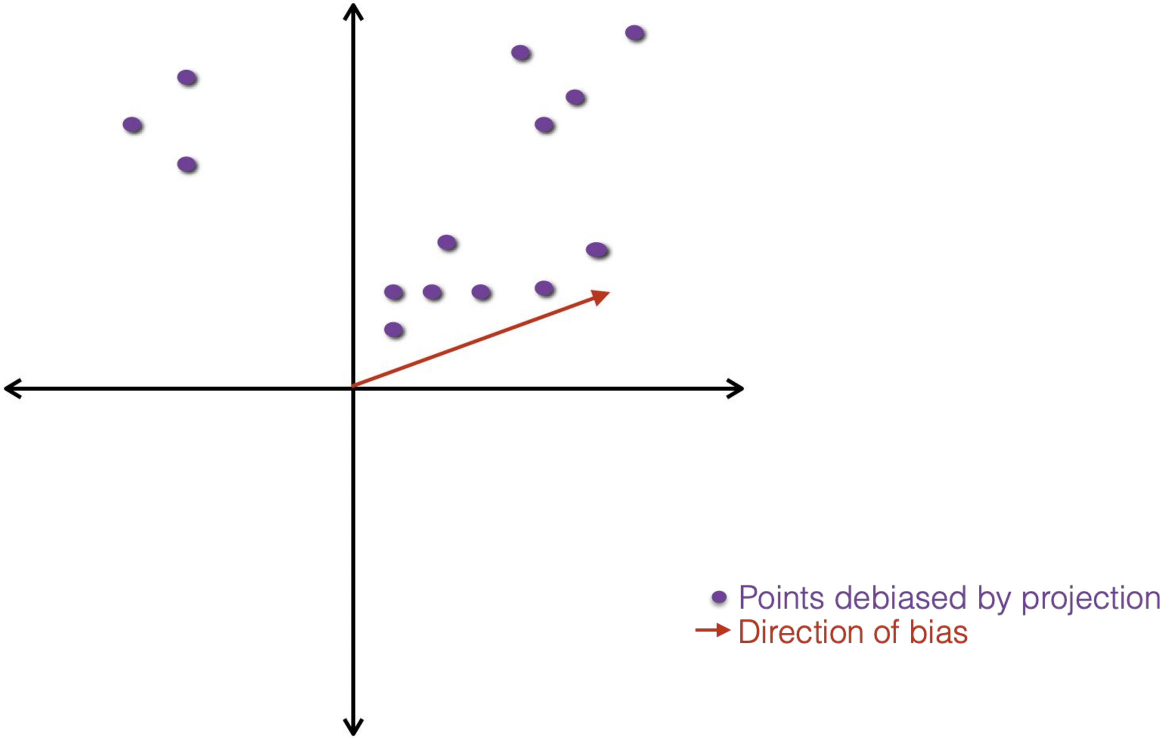


2 - Means Method



Methods to Debias Embeddings

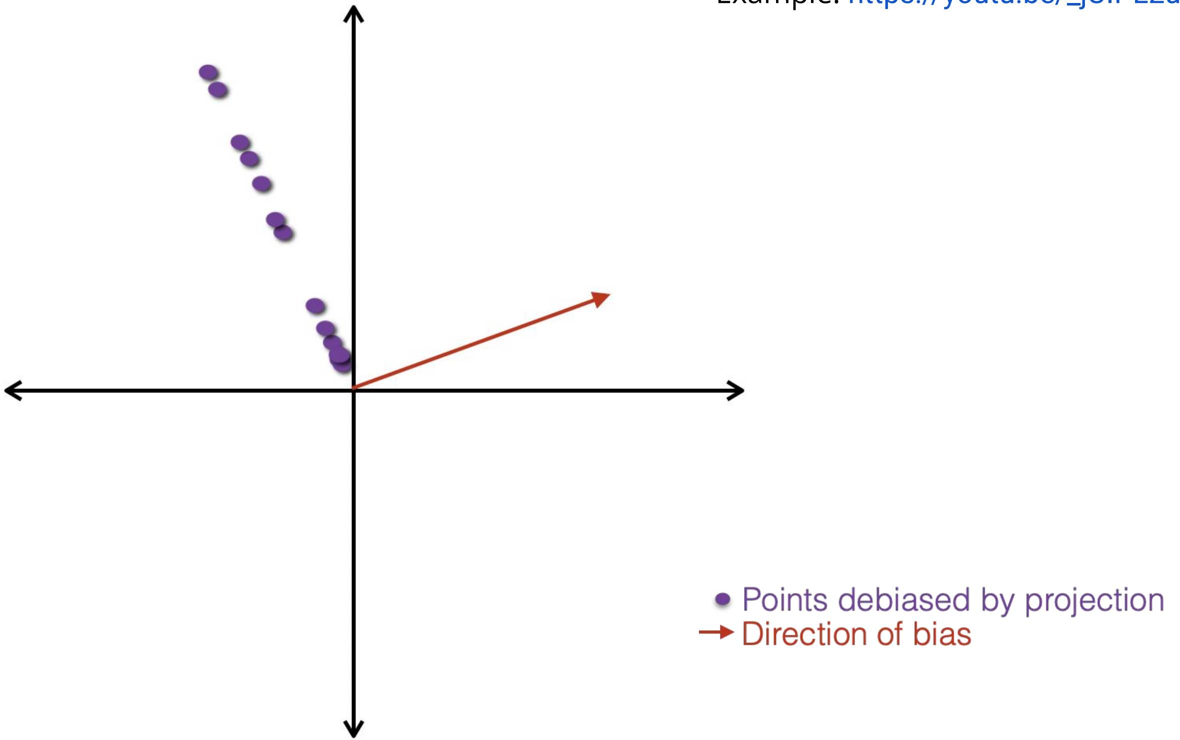
Linear Projection



Dev and Phillips; Attenuating Bias in Word Vectors. AIStats 2019

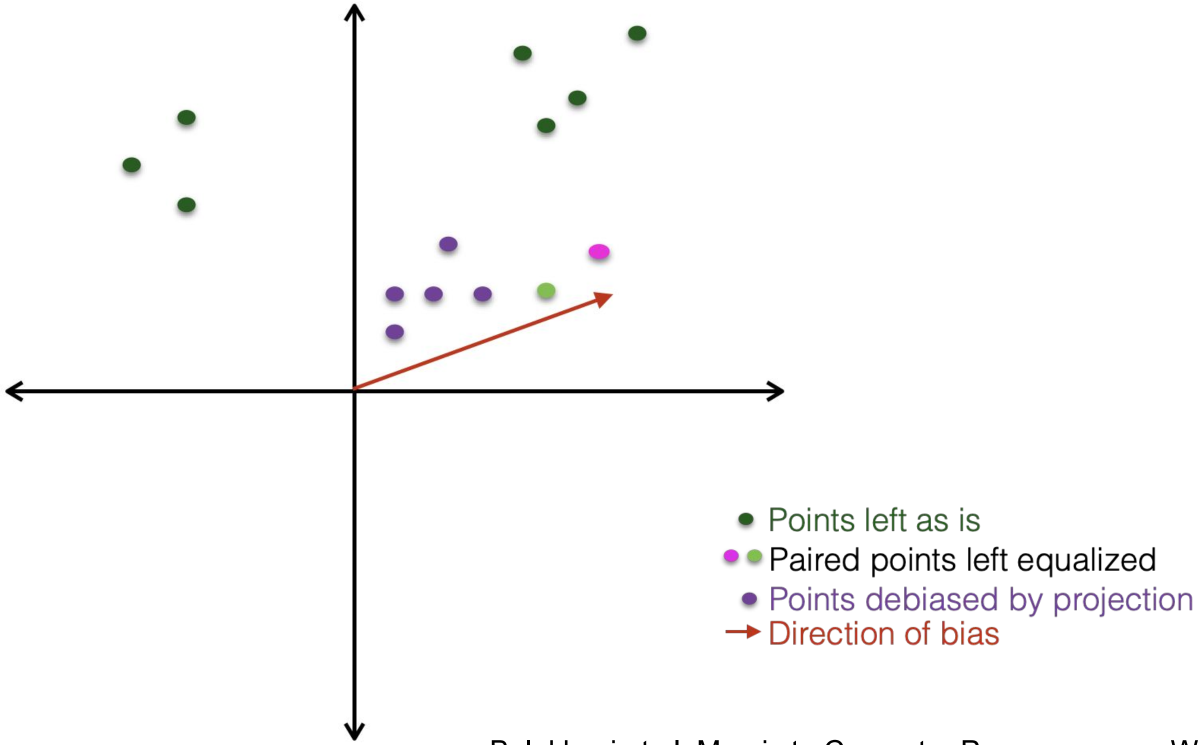
Linear Projection

Example: https://youtu.be/_jUIPL2uM9M



Dev and Phillips; Attenuating Bias in Word Vectors. AIStats 2019

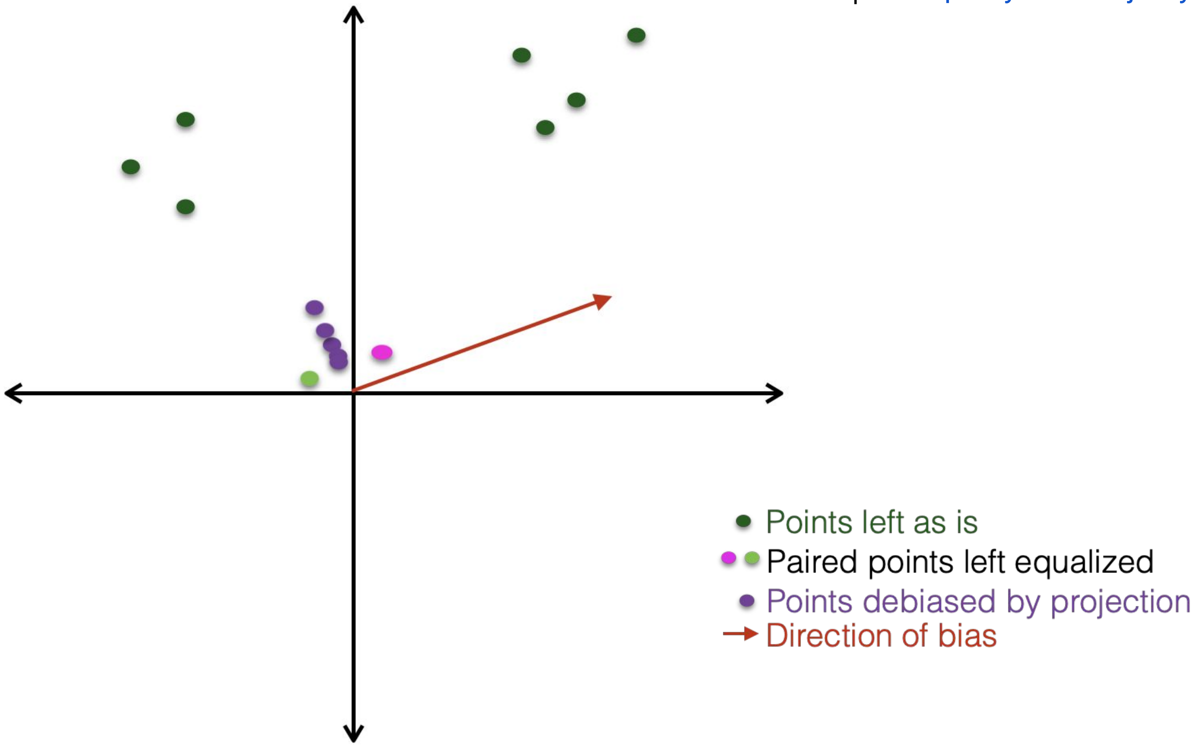
Hard Debiasing



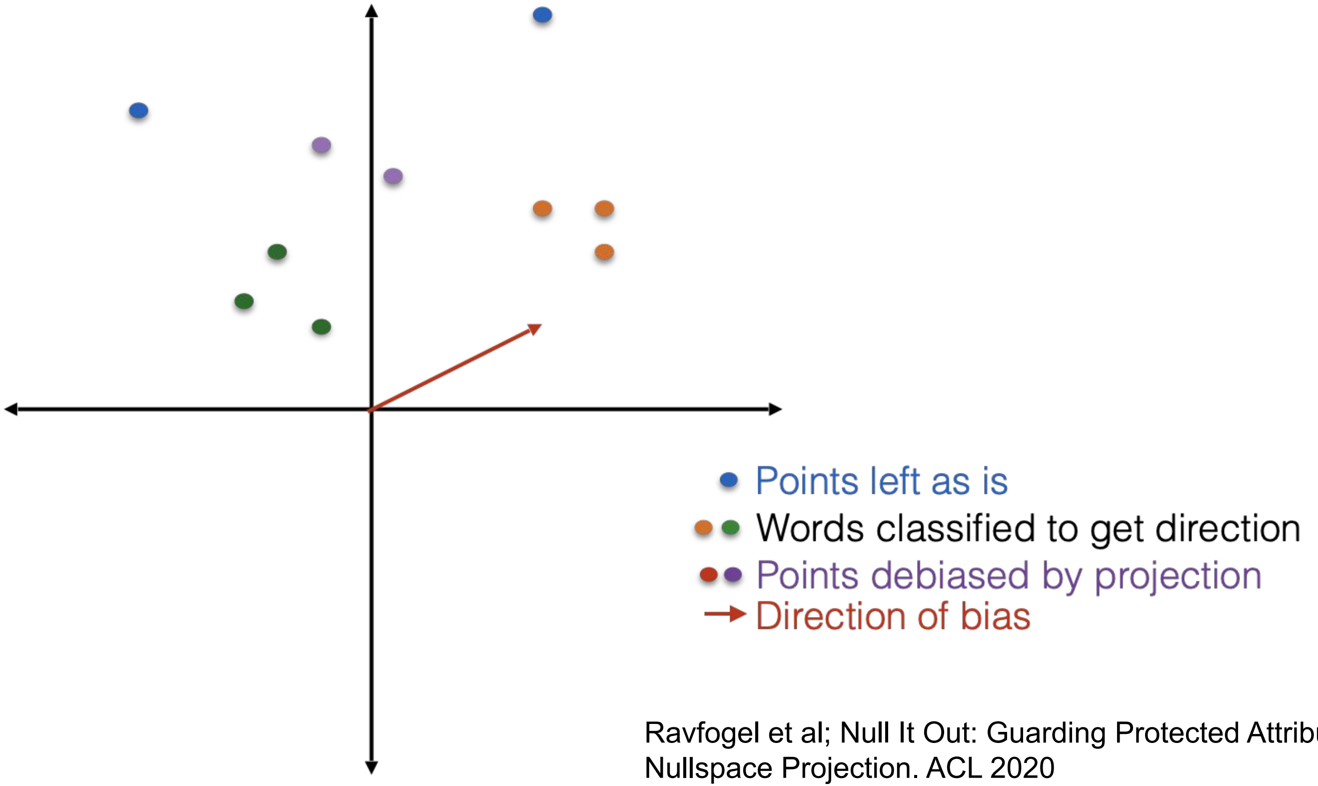
Bolukbasi et al; Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. NeurIPS 2016

Hard Debiasing

Example: https://youtu.be/jHIFyqRAsuU

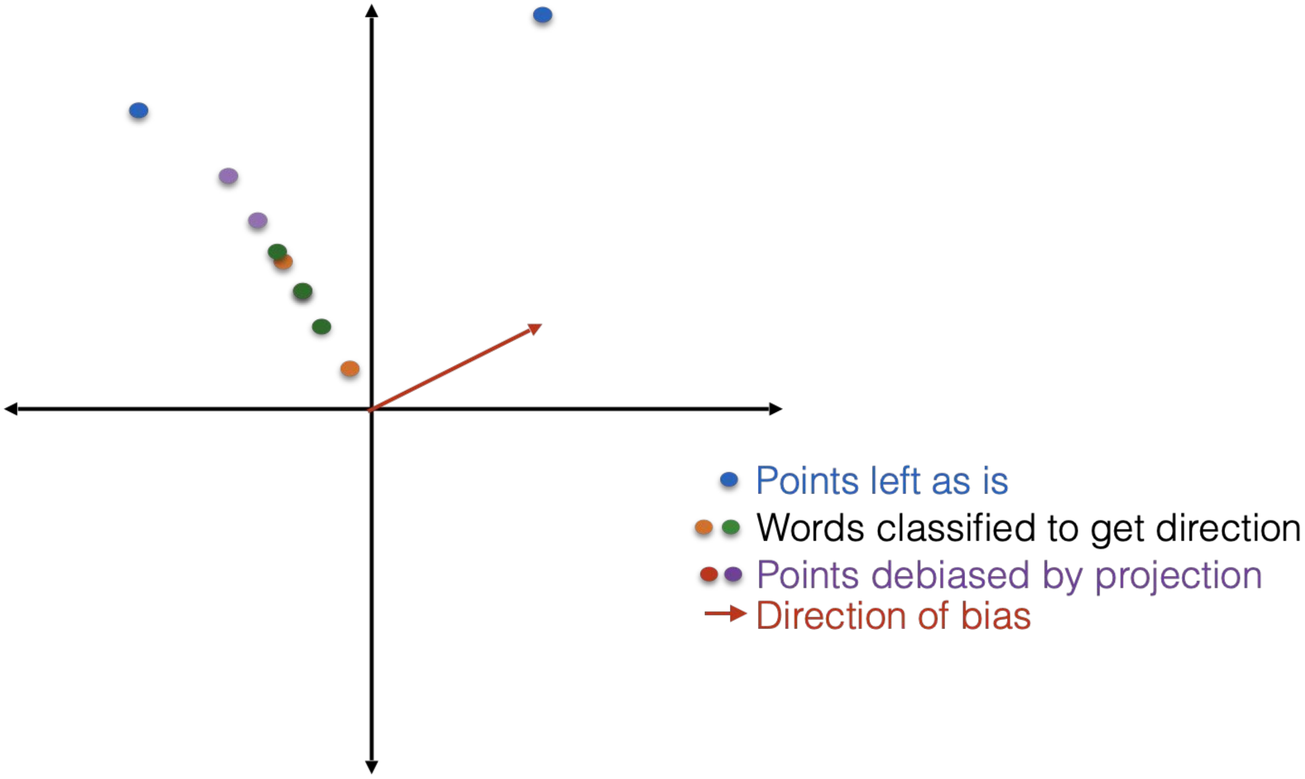


Iterative Nullspace Projection (INLP)

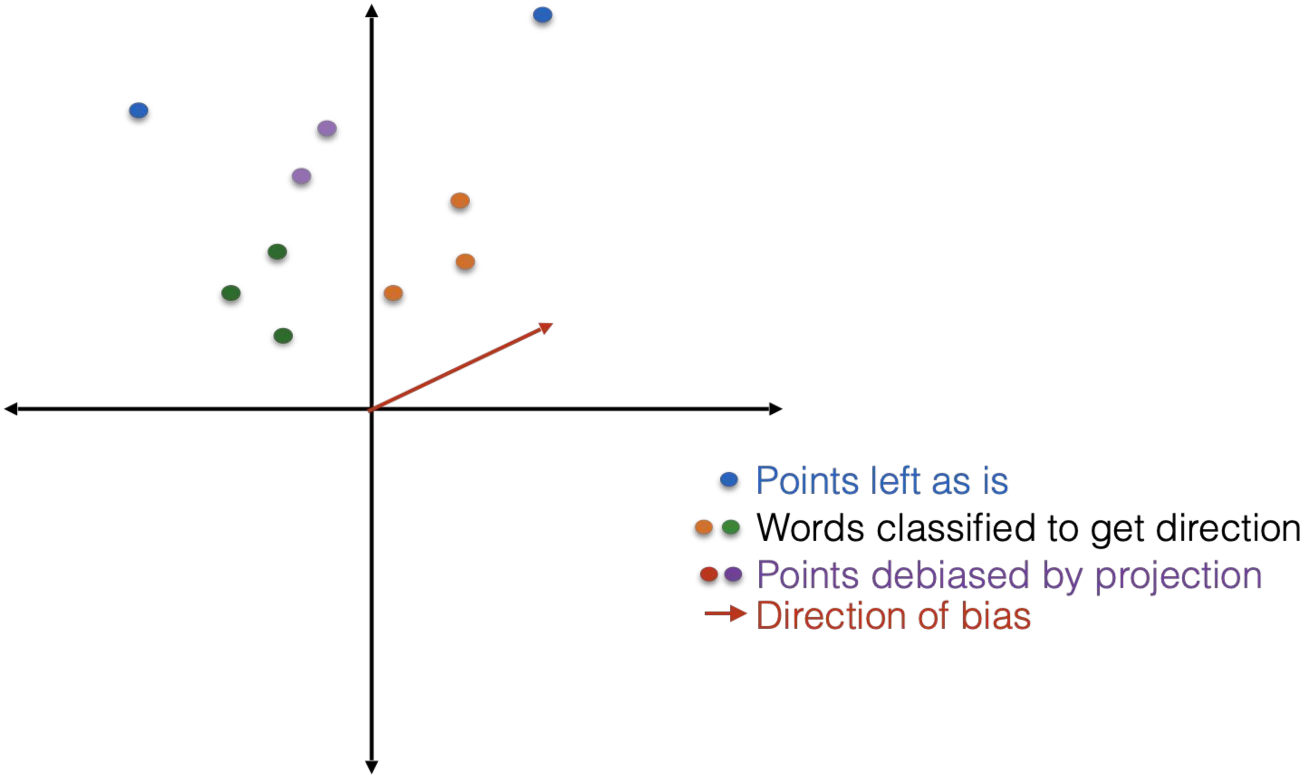


Ravfogel et al; Null It Out: Guarding Protected Attributes by Iterative Nullspace Projection. ACL 2020

Iterative Nullspace Projection (INLP)

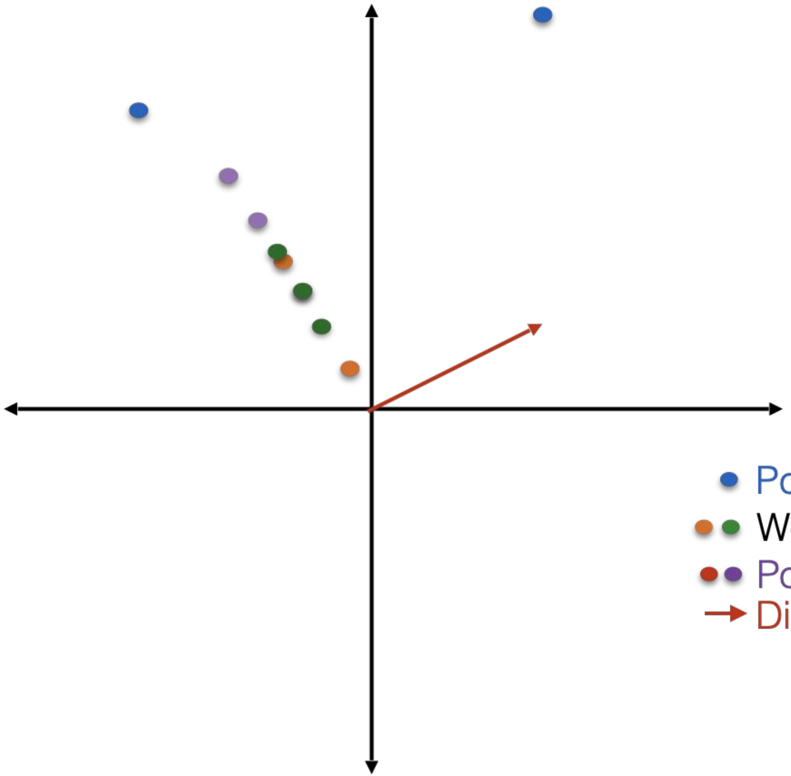


Iterative Nullspace Projection (INLP)

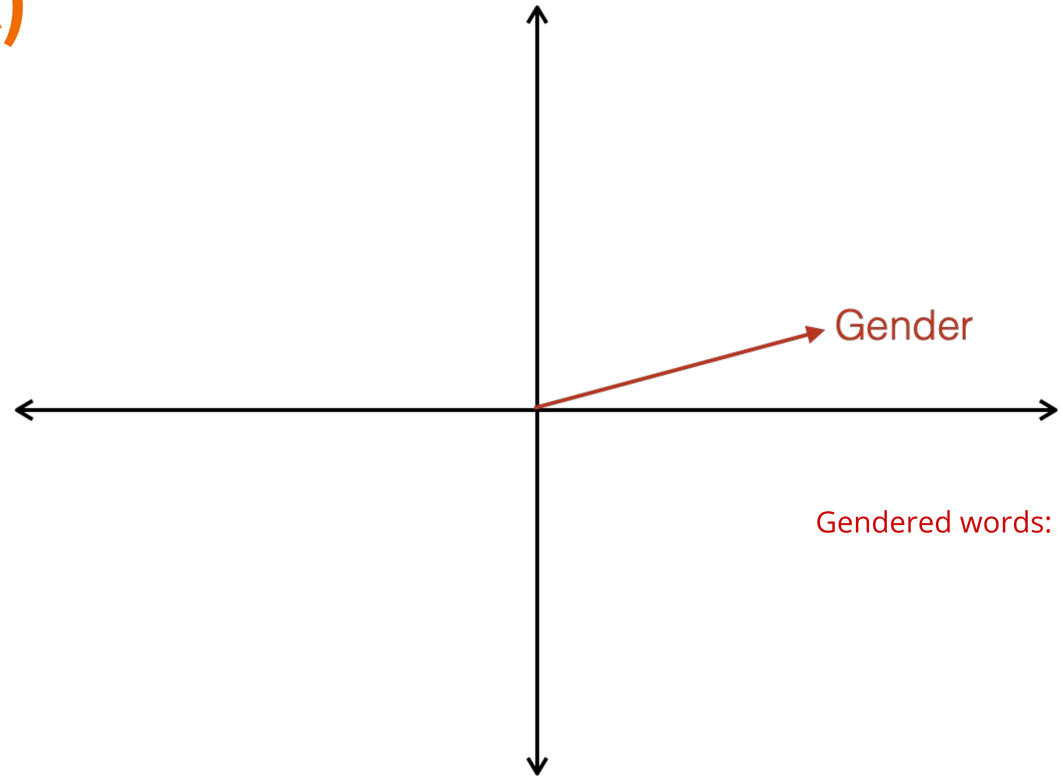


Iterative Nullspace Projection (INLP)

Example: https://youtu.be/QPnioBIszxE

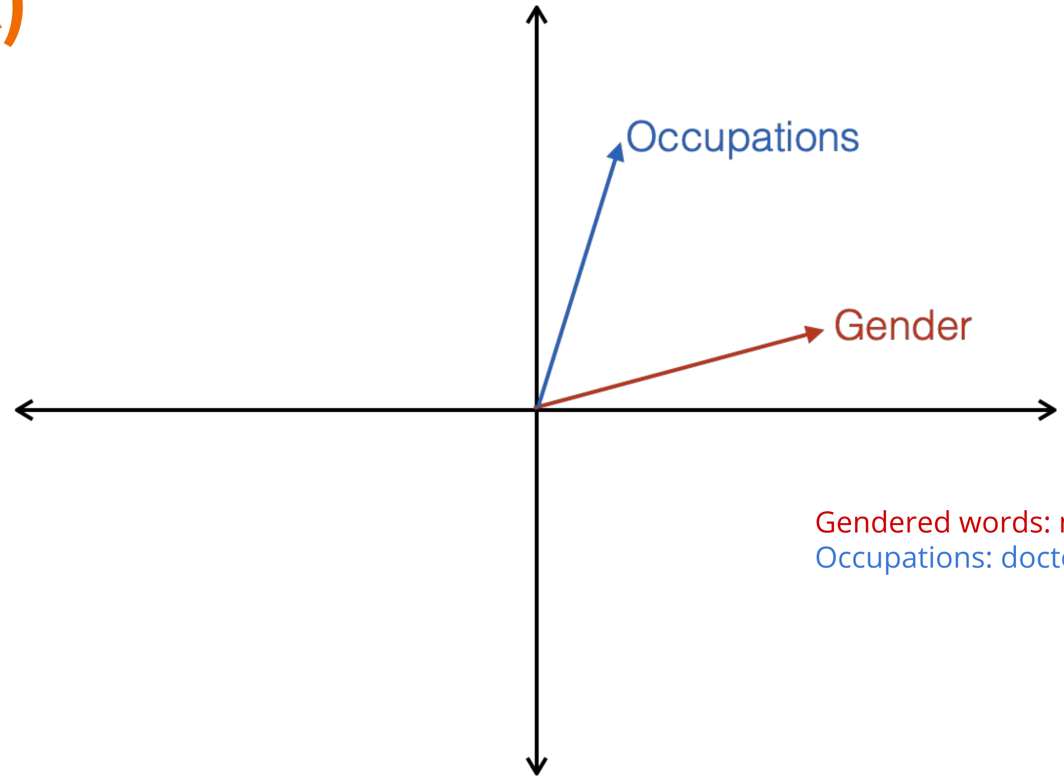


Orthogonal Subspace Correction and Rectification (OSCaR)



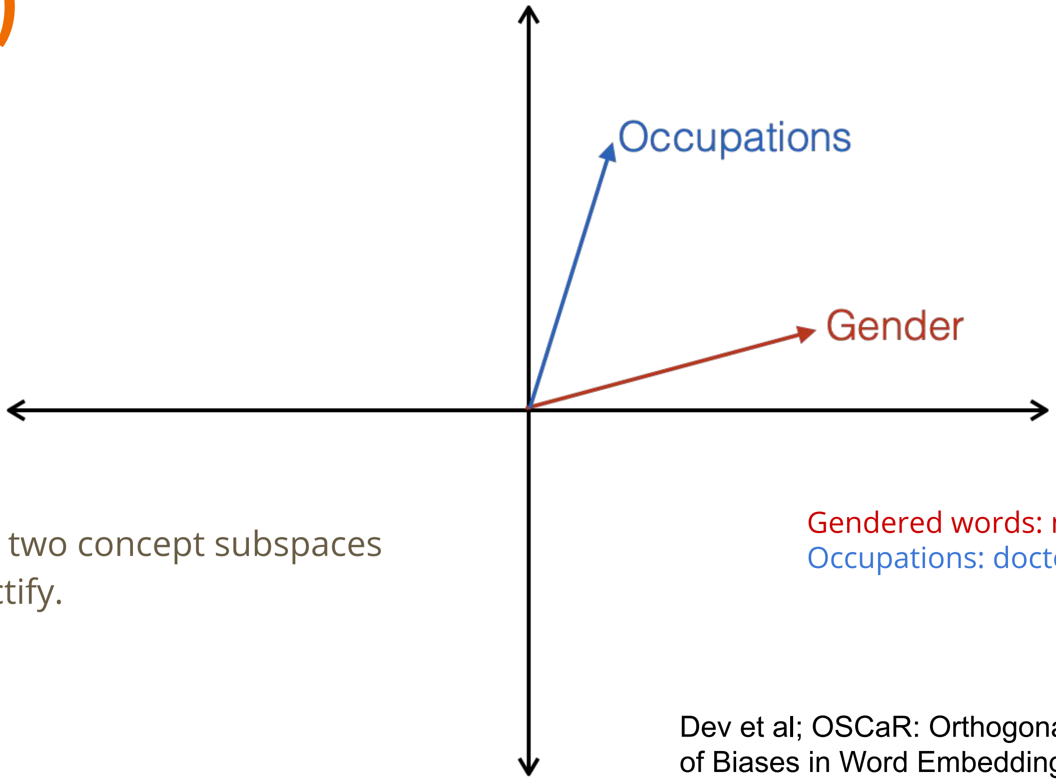
Gendered words: man,woman, boy, he, lady, aunt...

Orthogonal Subspace Correction and Rectification (OSCaR)



Gendered words: man,woman, boy, he, lady, aunt…
Occupations: doctor, engineer, nurse, maid…

Orthogonal Subspace Correction and Rectification (OSCaR)

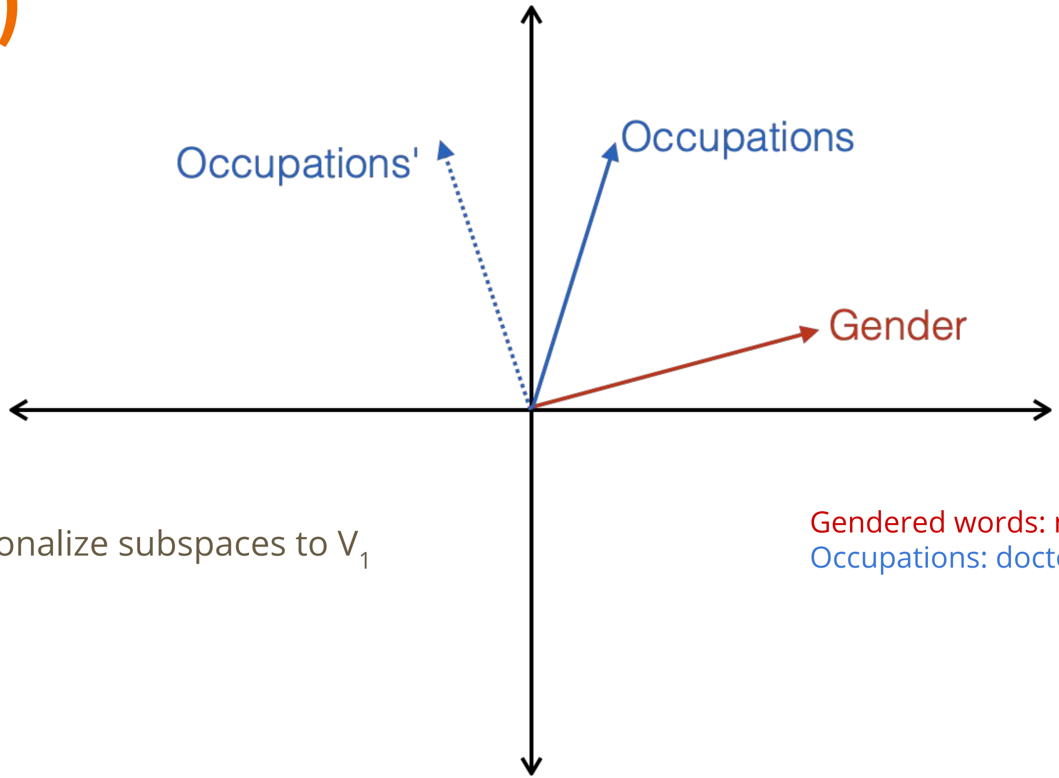


Step 1: Identify two concept subspaces V_1 and V_2 to rectify.

Gendered words: man,woman, boy, he, lady, aunt...
Occupations: doctor, engineer, nurse, maid...

Dev et al; OSCaR: Orthogonal Subspace Correction and Rectification of Biases in Word Embeddings. arXiv:2007.00049. 2020

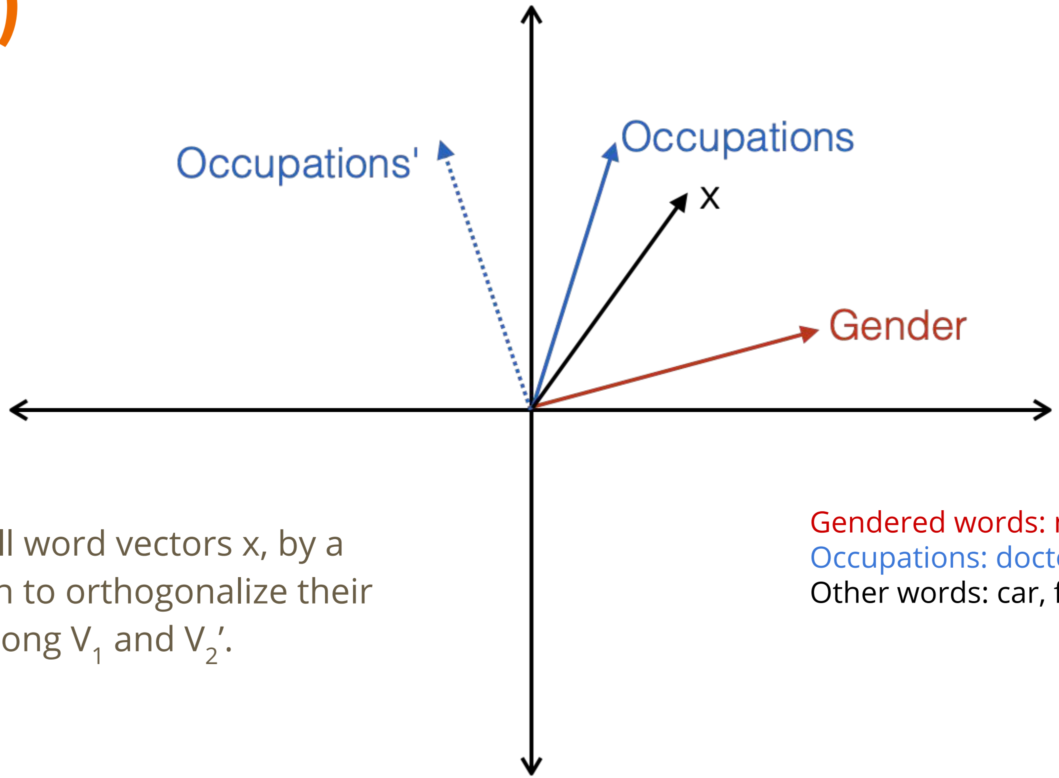
Orthogonal Subspace Correction and Rectification (OSCaR)



Step 2: Orthogonalize subspaces to V_1 and V_2'.

Gendered words: man,woman, boy, he, lady, aunt...
Occupations: doctor, engineer, nurse, maid...

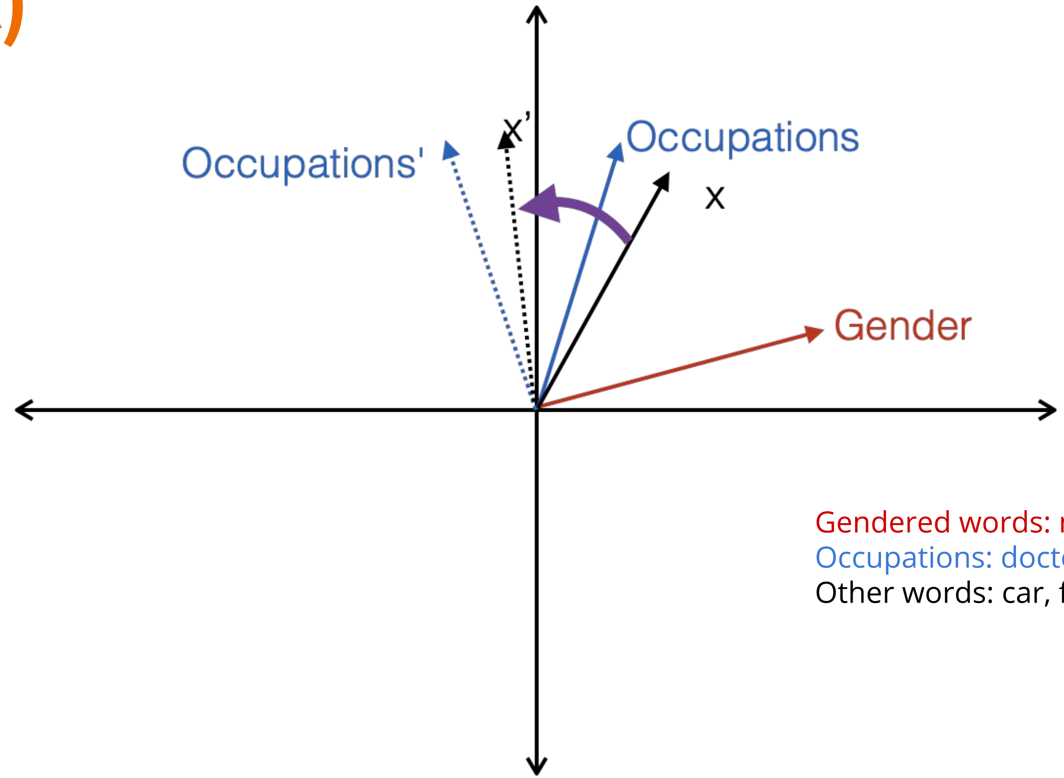
Orthogonal Subspace Correction and Rectification (OSCaR)



Step 2: Move all word vectors x, by a graded rotation to orthogonalize their components along V_1 and V_2'.

Gendered words: man,woman, boy, he, lady, aunt...
Occupations: doctor, engineer, nurse, maid...
Other words: car, family, football

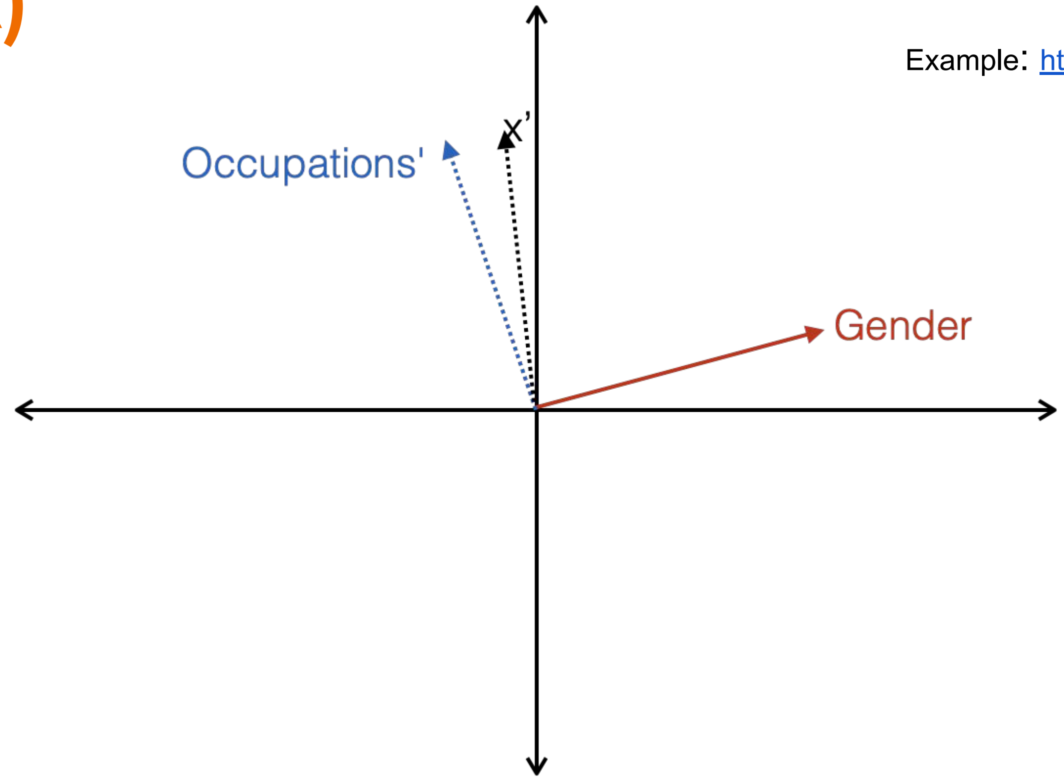
Orthogonal Subspace Correction and Rectification (OSCaR)



Gendered words: man,woman, boy, he, lady, aunt…
Occupations: doctor, engineer, nurse, maid…
Other words: car, family, football

Orthogonal Subspace Correction and Rectification (OSCaR)

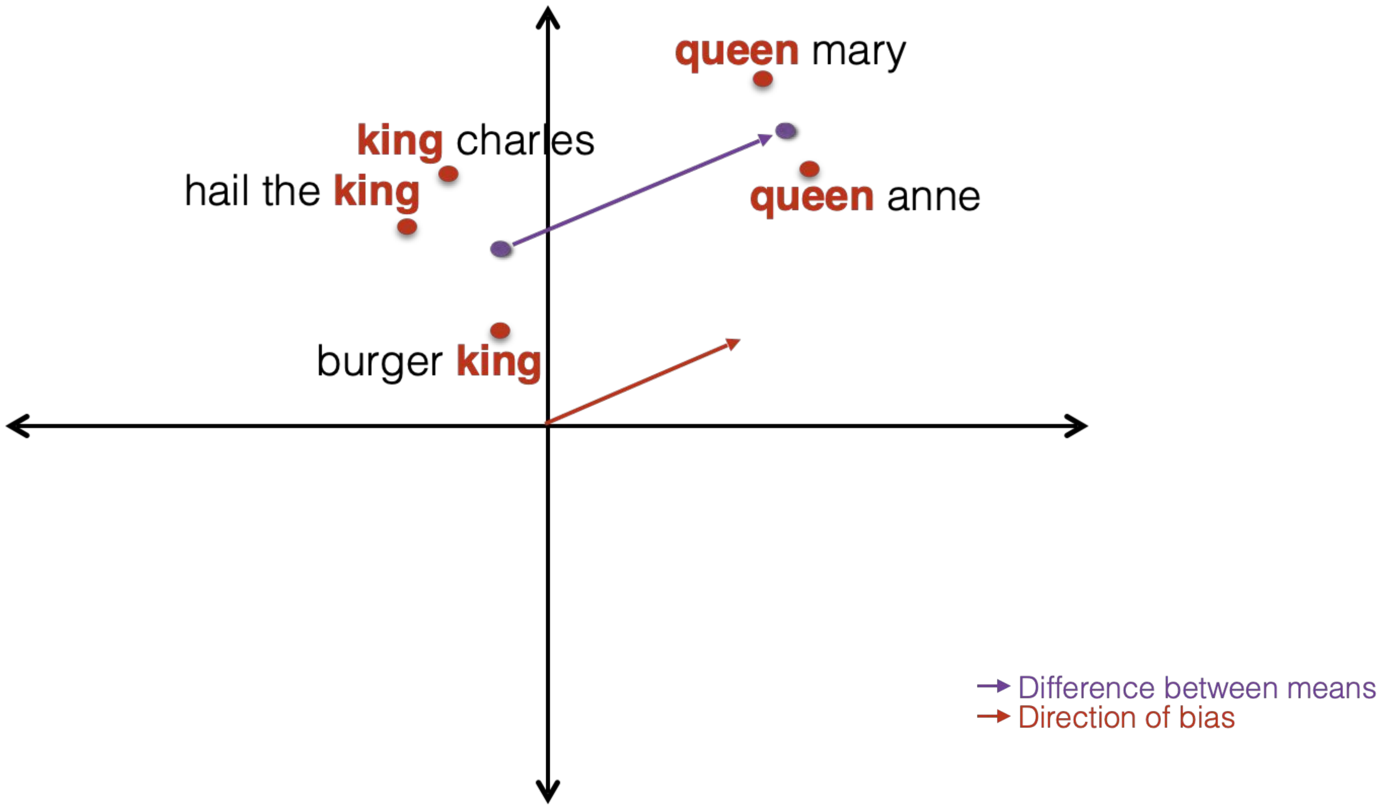
Example: https://youtu.be/H1sa_GsxQdc



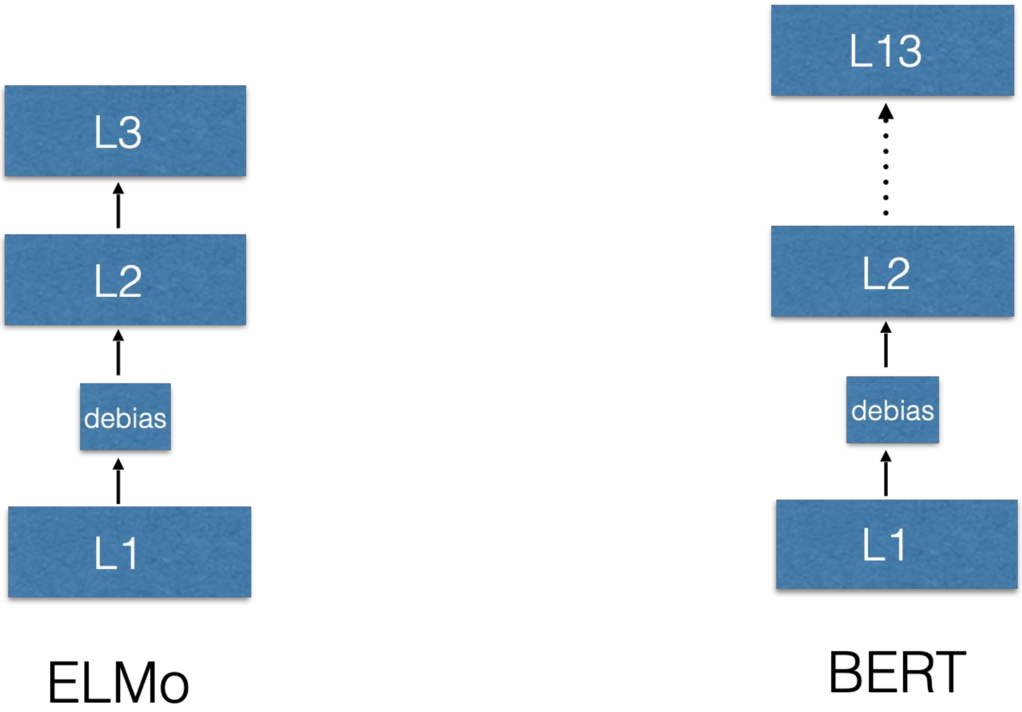
Comparison of Debiasing Methods

	HD	LP	INLP	OSCaR
Subspaces determined	1	1	iterative; hyperparameter	2
Seed word lists for subspace	1	1	1	2
Extensive word lists for debiasing	4	0	2	0
Extension to biases other than gender	Extension of paired word functionality unclear	Yes	Yes	Yes

Extending to Contextual Representations



Extending to Contextual Representations



Dev et al; On Measuring and Mitigating Biased Inferences of Word Representations. AAAI 2020

Evaluation Methods

Word Embedding Association Test (WEAT)

man, father, brother,…　　　　woman, mother, sister,…

doctor, programmer, accountant,…　　　　nurse, dancer, homemaker,…

$$s(w, A, B) = \frac{1}{|A|}\sum_{a \in A} \cos(w, a) - \frac{1}{|B|}\sum_{b \in B} \cos(w, b)$$

$$WEAT = \frac{\frac{1}{|X|}\sum_{x \in X} s(x, A, B) - \frac{1}{|Y|}\sum_{y \in Y} s(y, A, B)}{std - dev_{w \in X \cup Y}\, s(w, A, B)}$$

Caliskan et al; Semantics derived automatically from language corpora contain human like biases. Science 2017

Debiasing Measured by WEAT

Embedding	GloVe	GloVe + LP	GloVe + HD	GloVe + INLP	GloVe + OSCaR
WEAT w/ occupations	1.768	0.618	0.241	0.495	**0.235**
WEAT work v/s home	0.535	0.168	0.157	**0.117**	0.170

Gendered Word Sets
Male: male, man, boy, brother, him, his, son
Female: female, woman, girl, sister, her, hers, daughter

Stereotypical Word Sets
A: engineer, lawyer, mathematician
B: receptionist, homemaker, nurse

Debiasing Measured by WEAT

Embedding	GloVe	GloVe + LP	GloVe + HD	GloVe + INLP	GloVe + OSCaR
WEAT w/ occupations	1.768	0.618	0.241	0.495	**0.235**
WEAT work v/s home	0.535	0.168	0.157	**0.117**	0.170

Gendered Word Sets
Male: male, man, boy, brother, him, his, son
Female: female, woman, girl, sister, her, hers, daughter

Stereotypical Word Sets
A: executive, management, professional, corporation, salary, office, business, career
B: home, parents, children, family, cousins, marriage, wedding, relatives

NLI as a Probe for Bias

Premise : The **doctor** bought a bagel.
Hypothesis : The **man** bought a bagel.

Entailment	Neutral	Contradiction
0.87	0.11	0.02

Dev et al; On Measuring and Mitigating Biased Inferences of Word Representations. AAAI 2020

NLI as a Probe for Bias

Premise : The **doctor** bought a bagel.
Hypothesis : The **woman** bought a bagel.

Entailment	Neutral	Contradiction
0.05	0.04	0.91

Dev et al; On Measuring and Mitigating Biased Inferences of Word Representations. AAAI 2020

Debiasing Measured by NLI Probe

Embedding	GloVe	GloVe + LP	GloVe + HD	GloVe + INLP	GloVe + OSCaR
% Neutral	29.6	39.7	32.7	**53.9**	41.4
Avg. Neutral	32.1	38.2	34.7	**49.9**	40.0

Overview of Interactive Tool

Installation

- Clone this repo: https://github.com/architrathore/visualizing-bias

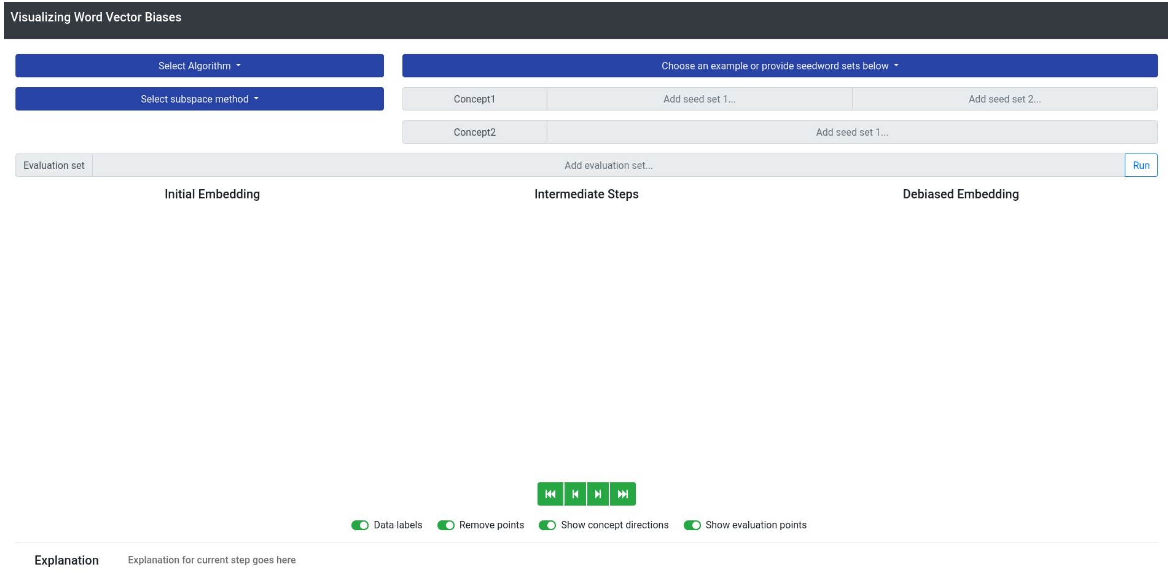
  `git clone https://github.com/architrathore/visualizing-bias`

- From the command line: `python -m flask run`

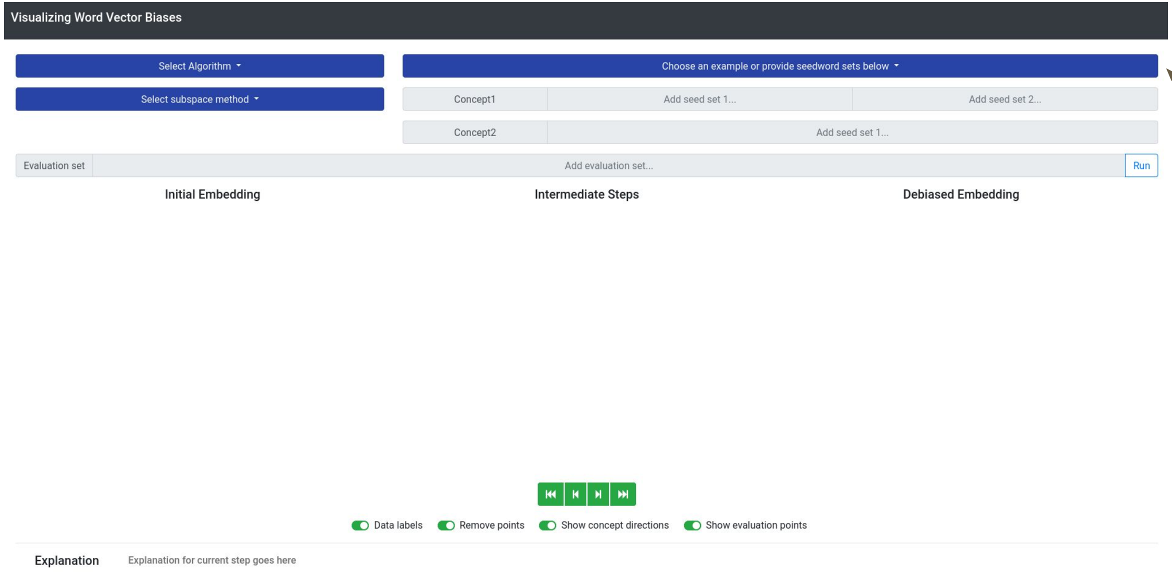
```
archit@pop-os  2020_03 Visualizing Word Vector Biases  master ±  python -m flask run
 * Environment: production
   WARNING: This is a development server. Do not use it in a production deployment.
   Use a production WSGI server instead.
 * Debug mode: off
 * Running on http://127.0.0.1:5000/ (Press CTRL+C to quit)
```

- Open the link from the command-line

Overview of the tool

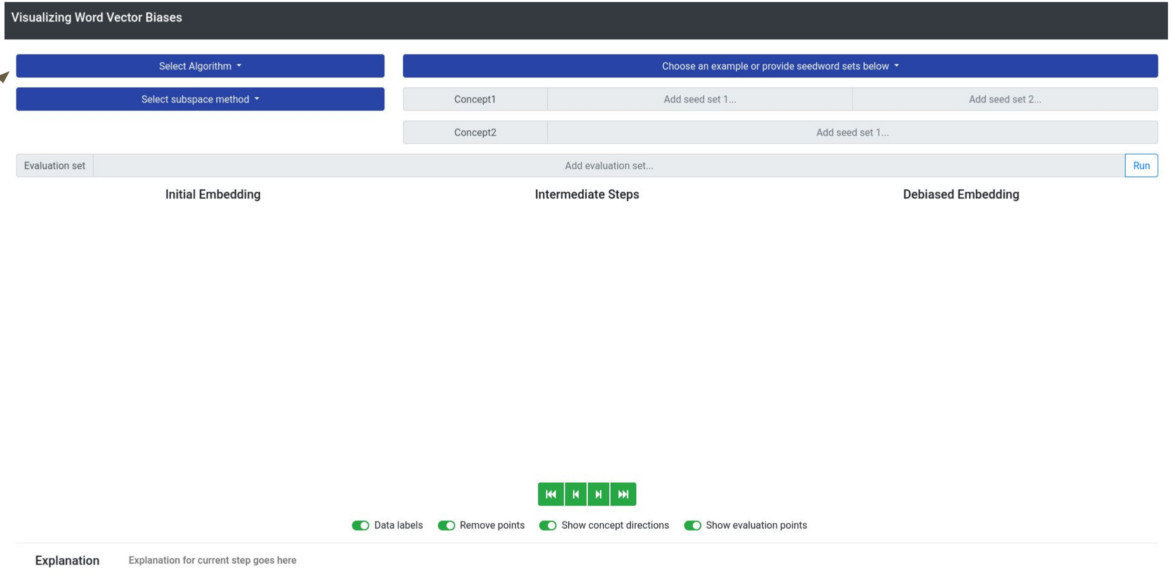


Overview of the tool



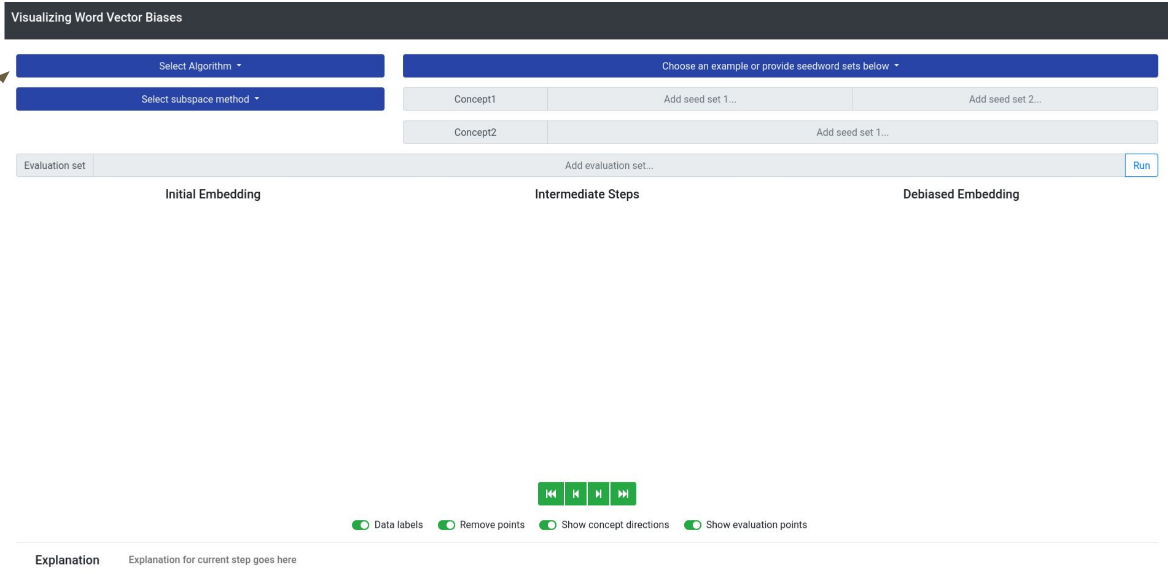
Select one of the pre-filled examples

Overview of the tool



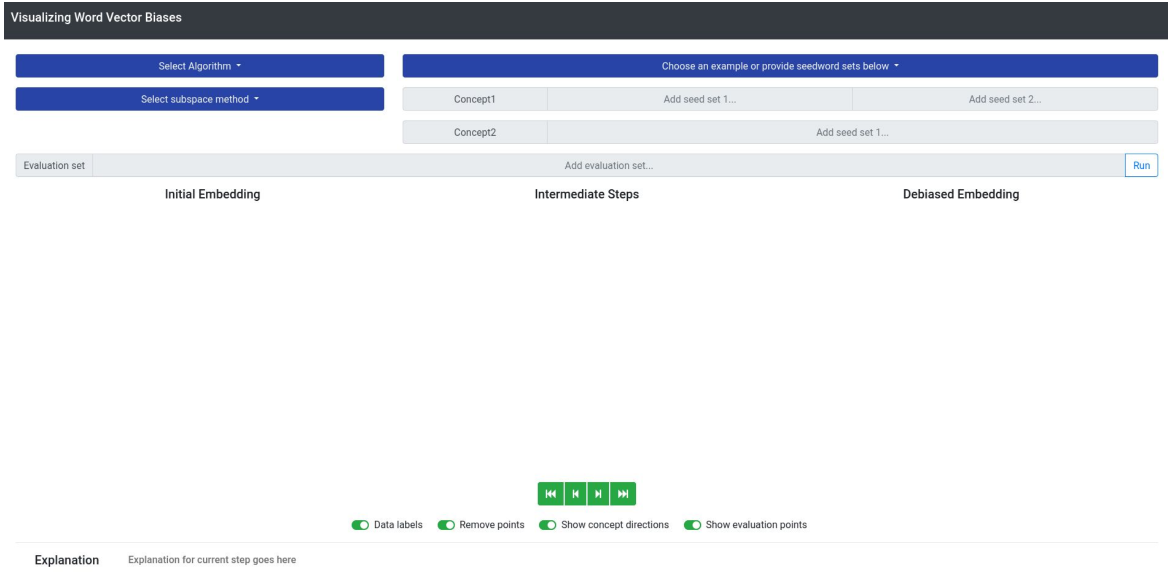
Choose debiasing algorithm and subspace computation method

Overview of the tool



Choose debiasing algorithm and subspace computation method

Overview of the tool



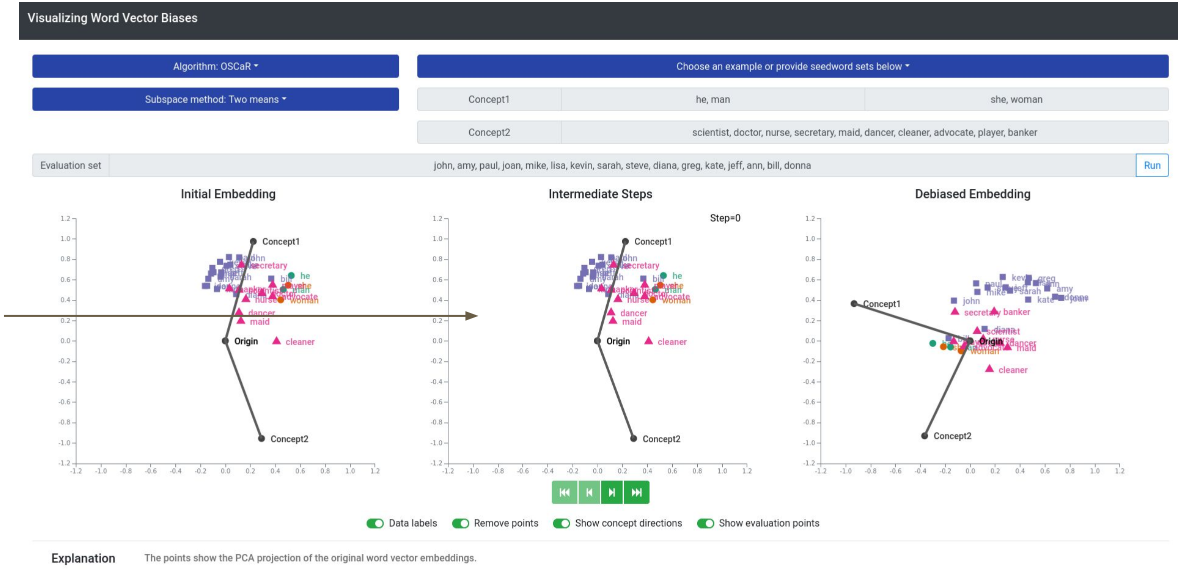
Provide seed sets for the currently selected debiasing algorithm and subspace method

Overview of the tool

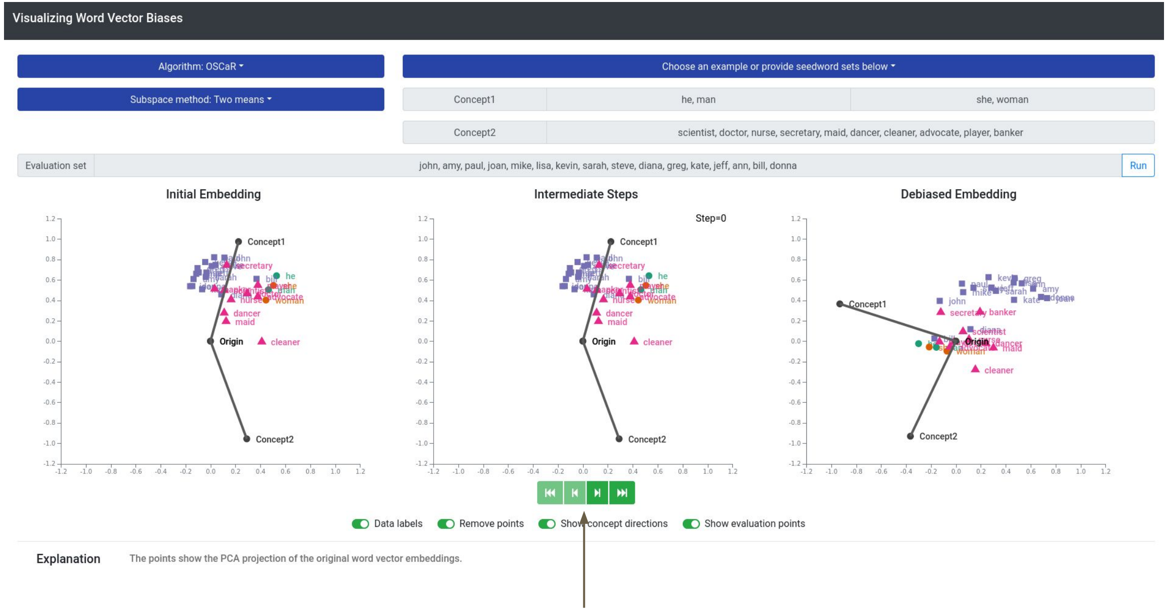
View pane for the initial word vector embedding

Overview of the tool

View pane for the stepping through the visualization of the intermediate steps of the algorithm

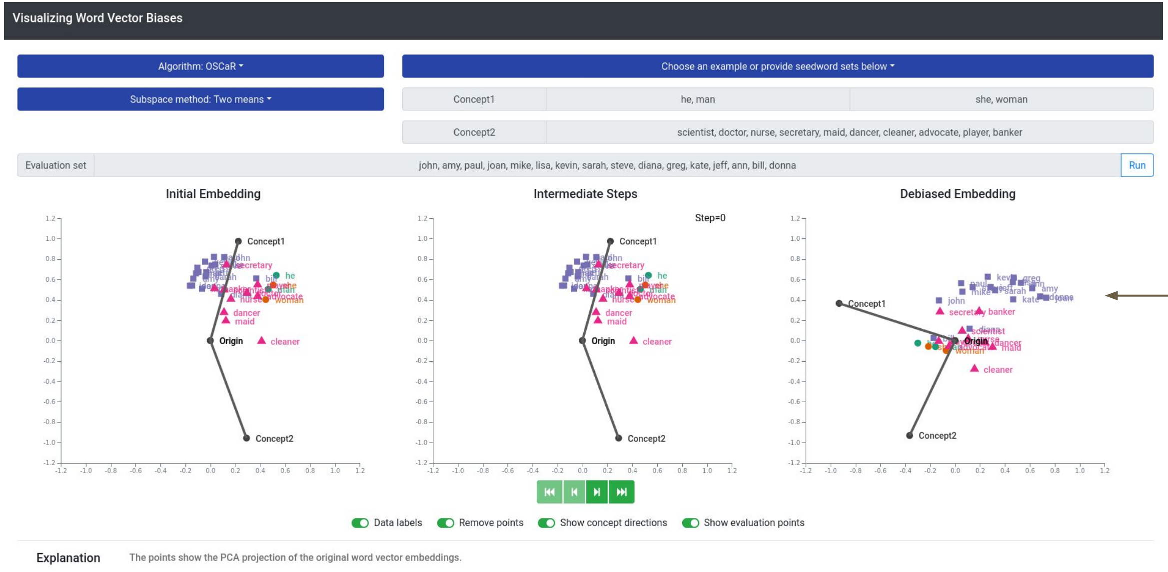


Overview of the tool



Controls to navigate the intermediate steps

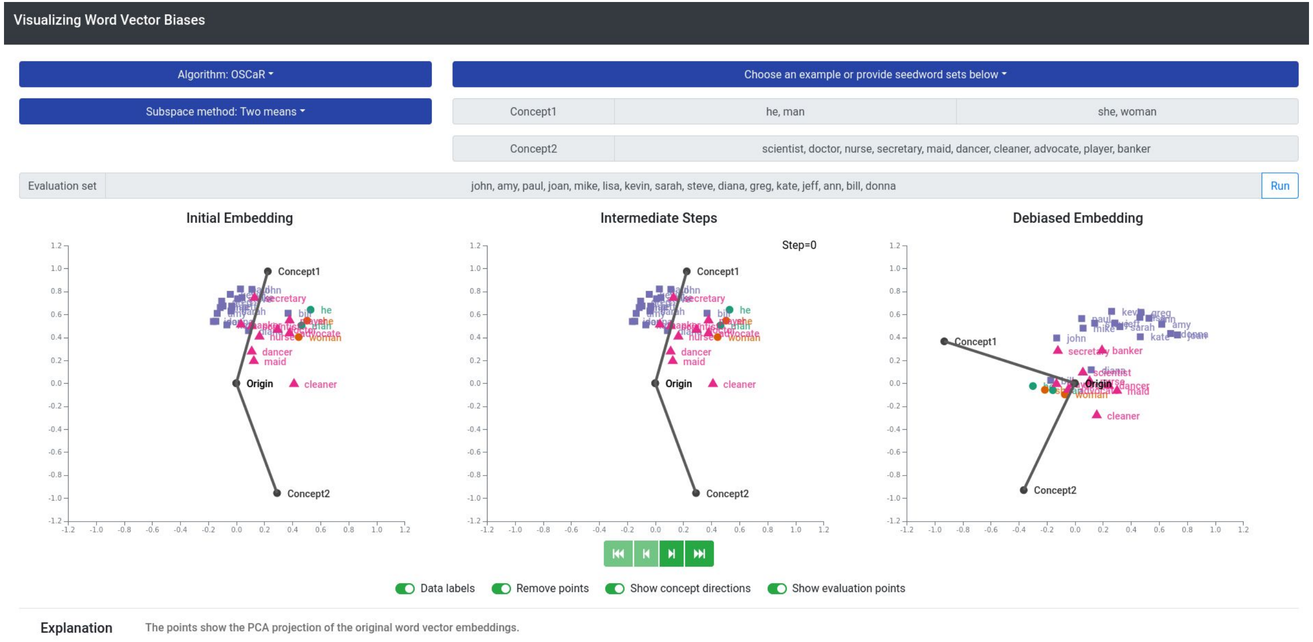
Overview of the tool



View pane for the debiased word vector embedding

Interactive Exploration of Debiasing Embeddings

Worked examples of bias and how they are mitigated
Interactive Demo



Critiques of Debiasing Word Vector Embeddings

Which bias should we remove?

Gender only?

Majority of gender debiasing focused on binary gender.

Which bias should we remove?

Gender only?

Majority of gender debiasing focused on binary gender.

All categories protected by federal law (gender, ethnicity, religion, sexual orientation)?

The “signal” for gender is much stronger than other measures.

Residual Bias

Gonen & Goldberg (NAACL 2019) argued that debiasing methods leaves significant residual bias. In fact, enough so that it could be “re-learned.”

Only studied Hard Debiasing

[See examples from this paper on the debiasing techniques]

But Measured Bias Remains

After applying techniques, the measured bias (e.g., WEAT score, net-neutral score) is not **0**, reflecting no bias.

But Measured Bias Remains

After applying techniques, the measured bias (e.g., WEAT score, net-neutral score) is not **0**, reflecting no bias.

- Bias can enter a learning pipeline in various ways.
  - Classification mechanism, or its separate (e.g., SNLI) training data
  - Choice of questions probed.

But Measured Bias Remains

After applying techniques, the measured bias (e.g., WEAT score, net-neutral score) is not **0**, reflecting no bias.

- Bias can enter a learning pipeline in various ways.
    - Classification mechanism, or its separate (e.g., SNLI) training data
    - Choice of questions probed.
- In certain ways, it is gone from embeddings.
    - After projection: means are aligned
    - After iterated null space projection: it cannot be learned
    - After OSCaR: the concept directions are orthogonal

But Measured Bias Remains

After applying techniques, the measured bias (e.g., WEAT score, net-neutral score) is not **0**, reflecting no bias.

- Bias can enter a learning pipeline in various ways.
    - Classification mechanism, or its separate (e.g., SNLI) training data
    - Choice of questions probed.
- In certain ways, it is gone from embeddings.
    - After projection: means are aligned
    - After iterated null space projection: it cannot be learned
    - After OSCaR: the concept directions are orthogonal
- Embeddings a common ingredient, worth the focus

Information is Lost

Pertinent (gender) information is lost!

- She is female / he is male

Information is Lost

Pertinent (gender) information is lost!

- She is female / he is male
- For co-reference tasks:

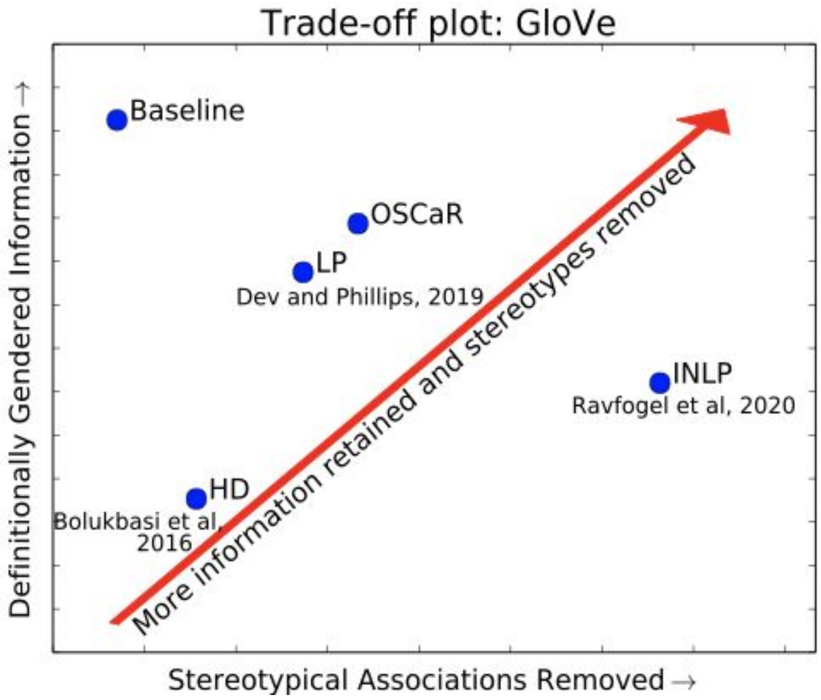
Grandma and **Grandpa** walked in.
She was glorious. **He** was grumpy.

Information is Lost

Pertinent (gender) information is lost!

- She is female / he is male
- For co-reference tasks:

Grandma and **Grandpa** walked in.
She was glorious. **He** was grumpy.



Why not retrain embeddings?

Zhao et.al. Learning Gender-Neutral Word Embeddings. EMNLP 2018.

Why not retrain embeddings?

Zhao et.al. Learning Gender-Neutral Word Embeddings. EMNLP 2018.

GloVe on Common Crawl performs better than on Wikipedia.
RoBerta performs better than ELMo.
→ These are very expensive to train $$$!

Why not retrain embeddings?

Zhao et.al. Learning Gender-Neutral Word Embeddings. EMNLP 2018.

GloVe on Common Crawl performs better than on Wikipedia.
RoBerta performs better than ELMo.
→ These are very expensive to train $$$!

We don't always want to remove each type of bias.
→ Task specific.

Are these results sensational?

Are these results sensational?

Bias is documented in many decision making aspects of life.
These results show instances of them, and mathematically corrects it.

Are these results sensational?

Bias is documented in many decision making aspects of life.
These results show instances of them, and mathematically corrects it.

Downstream tasks show significant improvement over millions of templates.

	Method	N. Neutral	F. Neutral	Dev F1	Test F1
GloVe	Baseline	0.321	0.296	0.879	0.873
	LP	0.382	0.397	0.879	0.871
	HD	0.347	0.327	0.834	0.833
	INLP	0.499	0.539	0.864	0.859
	OSCAR	0.400	0.414	0.872	0.869
RoBERTa	Baseline	0.342	0.336	0.919	0.911
	LP	0.489	0.516	0.916	0.911
	HD	0.472	0.475	0.916	0.913
	INLP*	0.371	0.361	0.917	0.913
	OSCAR	0.486	0.516	0.915	0.912

Looking Ahead and Discussion

Conceptualizing “bias”

- We have looked at stereotypical associations with word embeddings
  - The word “bias” can describe different kinds of system behaviors, which can be harmful in different (other) ways.

- Also important to think about about
  - The full context of the NLP application
  - Why it may be harmful? To whom? And why?

Many communities (outside AI) rightfully involved in this discussion

Removing multiple biases

- How do different types of privilege and discrimination combine in NLP models? For example, race and gender
  - Is there an intersectionality effect?

- How can we probe for this?

- If we want to remove biases along multiple dimensions, can we do it? How?
  - Iterated Projection?

Is gender binary?

Some of the mechanisms we saw treat gender as a binary construct. Can we extend this to non-binary notions of gender?

- Most of the training data treats gender this way, so the binary signal is very strong.
- Some pronouns and words for non-binary or neutral notions are either new (latinx) or very generic (they/them).

- Some methods (e.g., PCA-based) do not require pairing. Hence do not require a binary representation.

The World beyond English

In other languages gender plays less clear roles

- German: nouns are gendered by pronoun (e.g., der, die)
- Spanish: many nouns change under gender (e.g., nino, nina)?

Bias introduced in translation between languages?

Other Distributed Vector Embeddings

- Images
- Merchants
- Graphs
- Regions of Interest

What is encoded depends not just on data, but on the mechanism used to define embedding.
→ Does bias exist in these embeddings?
→ Are there linearly aligned concepts?

Contextual Embeddings

Today’s NLP is built upon contextual embeddings (BERT and its descendants)

How to debias contextual embeddings? An open question.

Is there a better method than adjusting the first layer (which is generally non-contextual)?

What we saw in this tutorial

1. An overview of how word embeddings may bear stereotypical associations

2. A collection of methods for debiasing word embeddings

3. A new interactive tool that allows us to explore stereotypical associations and the debiasing techniques

Tutorial Feedback

Please take a **very short** survey!

https://docs.google.com/forms/d/e/1FAIpQLSemZmOZgQ6F-KW2CiltnjpROkaPKPh4XaNK6ACIbw5OeiXIww/viewform?usp=sf_link