Visualization for Ethics
Visually Mitigating Biases in Word Embeddings

Bei Wang Phillips

School of Computing
Scientific Computing and Imaging Institute (SCI)
University of Utah
www.sci.utah.edu/~beiwang
beiwang@sci.utah.edu
Joint work with
Archit Rathore, Sunipa Dev (UCLA), Jeff M. Phillips, Vivek Srikumar
Yan Zheng, Chin-Chia Michael Yeh, Junpeng Wang, Wei Zhang (Visa Research)

COMP 5360 / MATH 4100
Lecture 13: Visualization for Ethics
Feb 22, 2022

Related Materials: AAAI, KDD tutorials, NeurIPS Demo

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

http://www.sci.utah.edu/~beiwang/kddbias2021/

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

For this lecture: http://archit.sci.utah.edu:5001/

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

Kate Crawford's NeurIPS 2017 Keynote
The Trouble with Bias

`https://www.youtube.com/watch?v=fMym_BKWQzk`

First 7 minutes: introduction

At 9:50 minutes: what is bias?

At 12:00 minutes: example of bias

Up to 15:00 minutes.

Word Embeddings

What are word representations?

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?

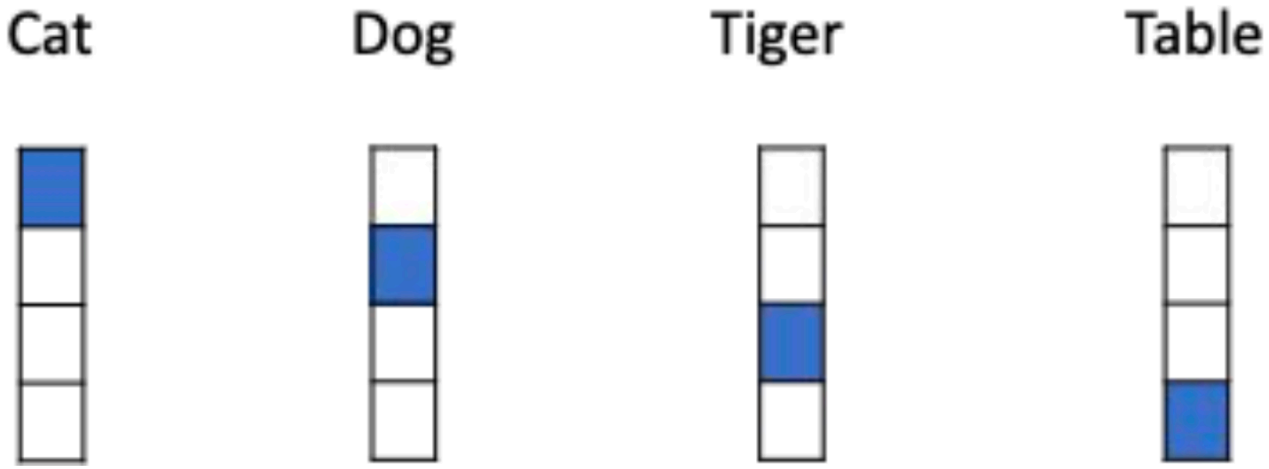
The *Distributional Hypothesis*: words that occur in the same contexts have similar meanings.

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.
  - *Cat* and *tiger* are conceptually closer to each other than to dog or table
  - *Cat*, *tiger* and *dog* are closer to each other than *table*

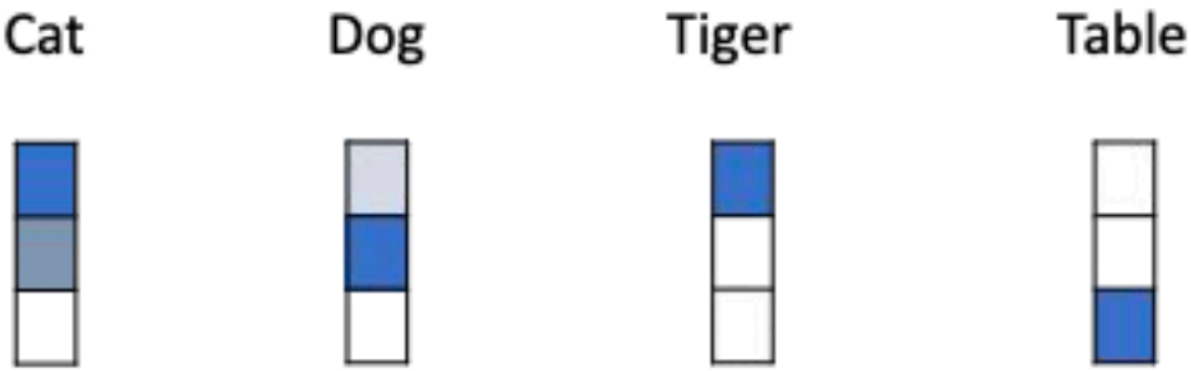
Symbolic vs. Distributed representations

- These *one-hot vectors* do not capture inherent similarities
- Distances or dot products are all equal
- In NLP, a one-hot vector is a $1 \times N$ matrix (vector) used to distinguish each word in a vocabulary from every other word in the vocabulary. The vector consists of 0s in all cells with the exception of a single 1 in a cell used uniquely to identify the word.
- `https://en.wikipedia.org/wiki/One-hot`



Symbolic vs. Distributed representations

- Distributed representations capture concept similarities better
- Vector-valued representations that coalesce superficially distinct concepts



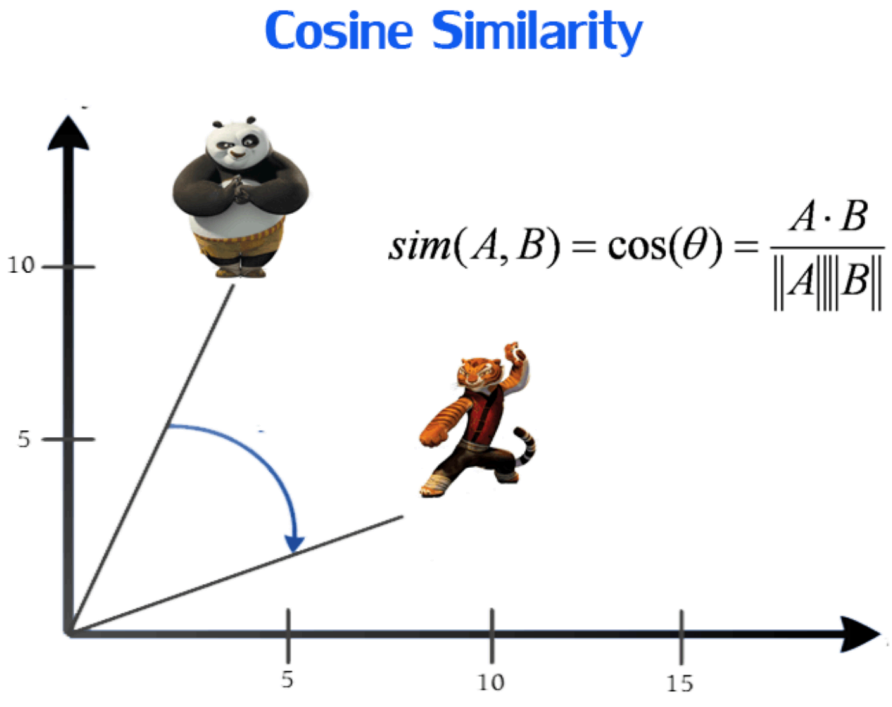
An example of Word2Vect

- *"Have a good day"*
- *"Have a great day"*
- Construct an exhaustive vocabulary $V = \{\text{Have, a, good, great, day}\}$.
- Have = [1,0,0,0,0]; a=[0,1,0,0,0]; good=[0,0,1,0,0]; great=[0,0,0,1,0]; day=[0,0,0,0,1]
- This means 'good' and 'great' are as different as 'day' and 'have', which is not true

https://towardsdatascience.com/introduction-to-word-embedding-and-word2vec-652d0c2060fa

An example of Word2Vect

- Objective: words with similar context occupy close spatial positions.
- Mathematically, the cosine of the angle between such vectors should be close to 1, i.e. angle close to 0.

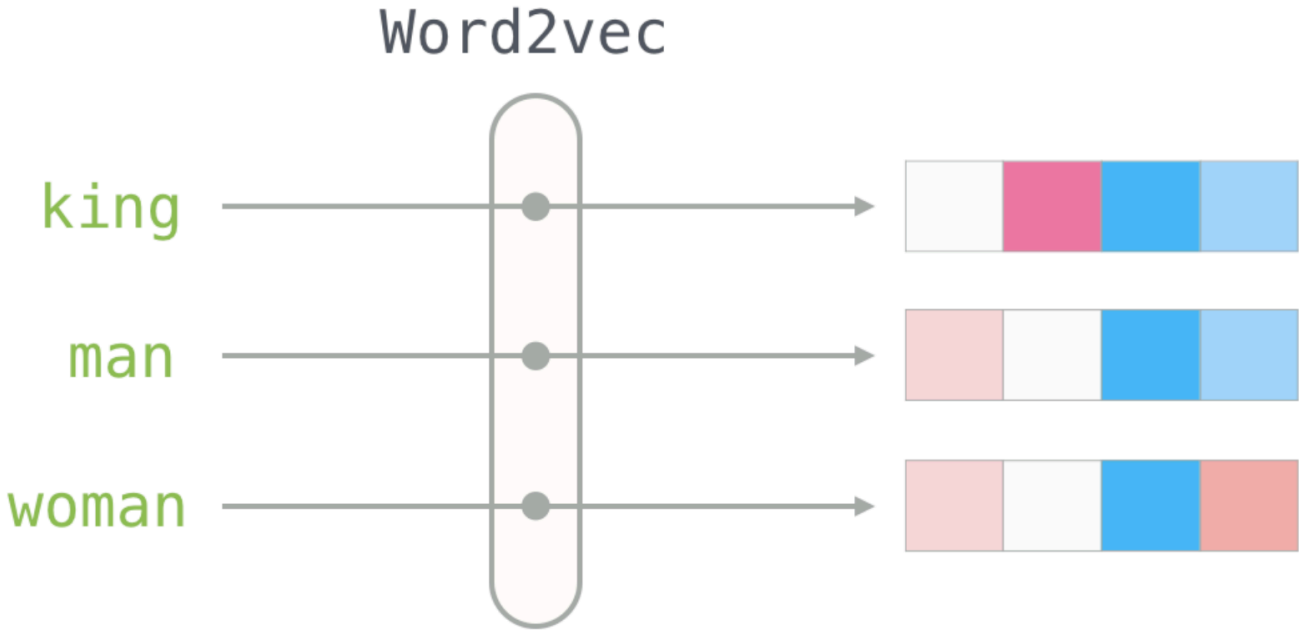


Google Images

https://towardsdatascience.com/introduction-to-word-embedding-and-word2vec-652d0c2060fa

An example of Word2Vect

- Intuitively: introduce some dependence of one word on the other words.
- Word2Vec is a method to construct such an embedding.



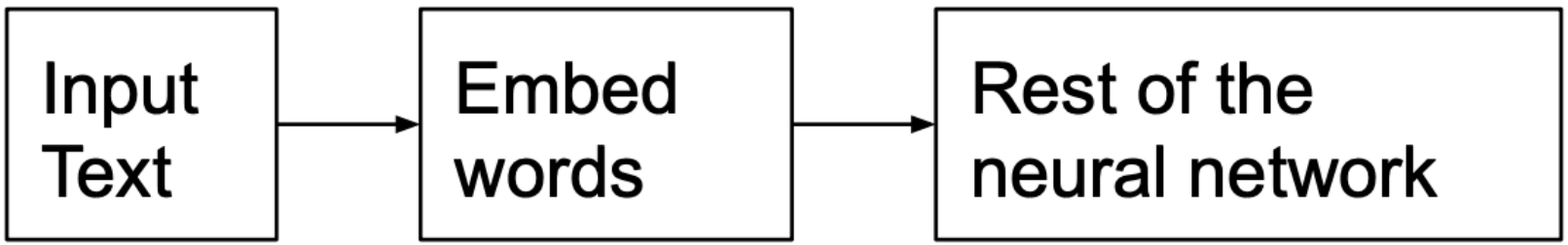
https://jalammar.github.io/illustrated-word2vec/

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.
- A word vector is a row of real valued numbers (as opposed to dummy numbers) where each point captures a dimension of the word's meaning and where semantically similar words have similar vectors[1].

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



[1] https://medium.com/@jayeshbahire/introduction-to-word-vectors-ea1d4e4b84bf

Perspectives on word embeddings

A mapping from words to a vector space could be:

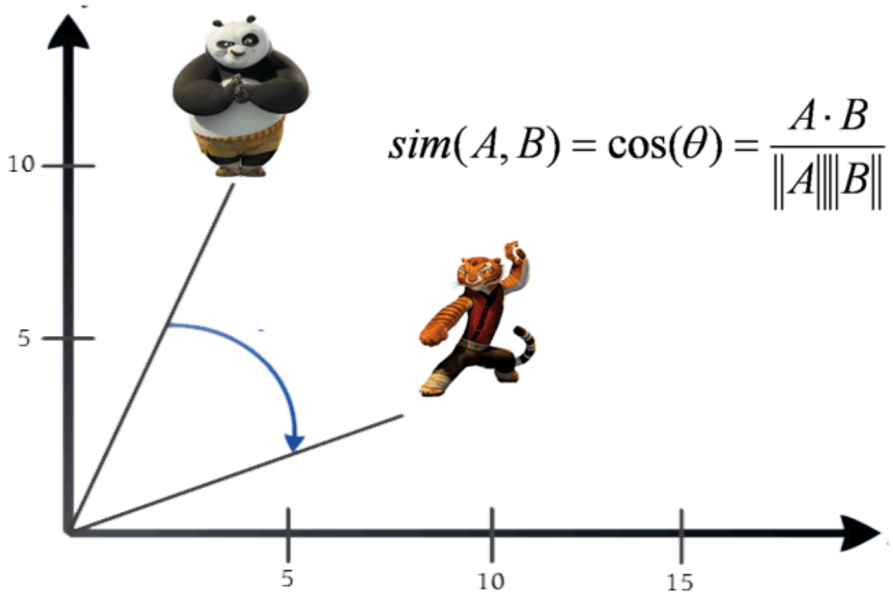
- *They capture distributional semantics*: Embeddings are low dimensional vectors that are constructed by appealing to the distributional hypothesis.
- *They are distributed representations of words*: The embedding dimensions represent underlying aspects of meaning, and words are characterized by membership to these latent dimensions.
- *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:

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

Cosine Similarity



Google Images

How are word embeddings trained?

Example: For the text

"It was a dark and _____ night and …"

1. Define a neural network of the form

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

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

Evaluating word embeddings: Two broad approaches

- *Intrinsic evaluation:* Evaluate the representation directly without training another model
  - Typically simple tasks where success or failure is (almost) entirely a function of the representation
  - Easy to compute, but doesn't say much about the embeddings as features
- *Extrinsic evaluation:* Evaluate the impact of the representation on another task
  - Typically, a neural network
  - Can be more practically useful, but slow and depends on the quality of the model for the task being tested

Word embeddings are great, but...

Word Biases

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

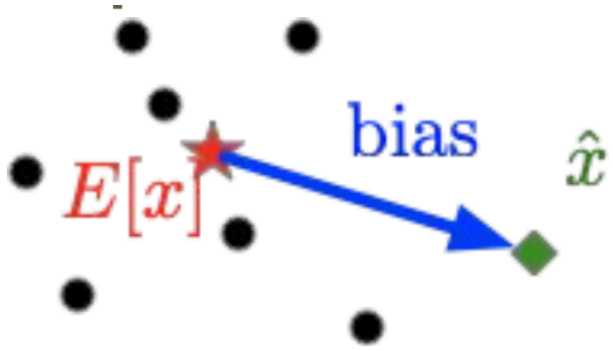
A fast moving field, with new techniques and perspectives being introduced almost every month. Two related lines of work:

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

What is bias?

- Definition: difference between an estimator and its expected value

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



What is bias?

- Definition: an instance of prejudice, especially a personal and sometimes unreasonable outlook. E.g. In machine learning .. a stereotype...

What is bias?

- Definition: 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

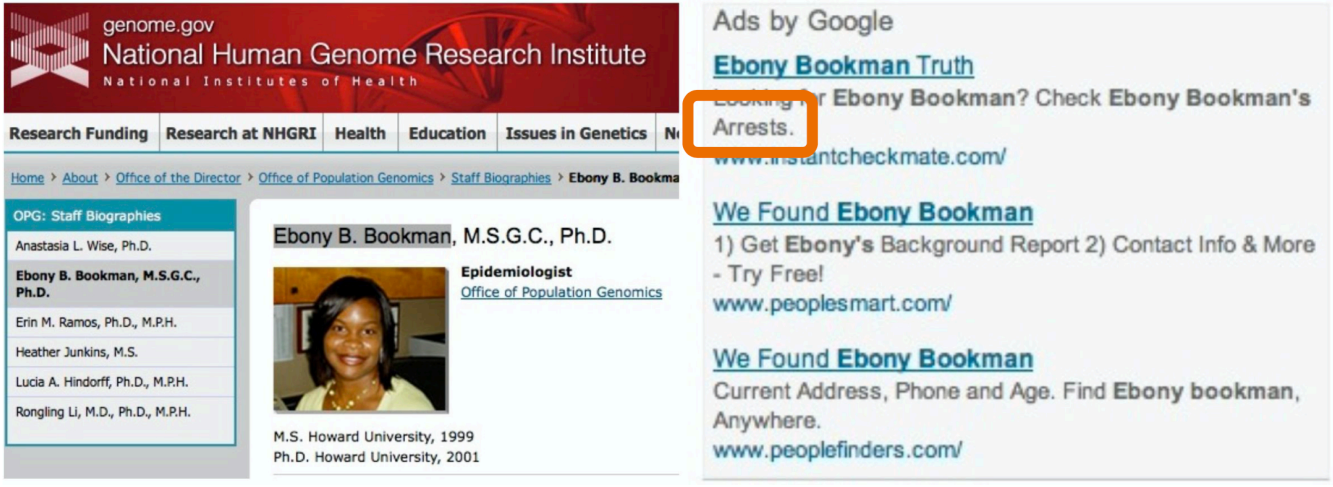
An oversimplification of a concept

- 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`

- Allocational Harms
  - College acceptance
  - Bank loan applications
  - Recidivism prediction and parole
- Representational Harms
  - More subtle. How data is represented which leads to negative stereotypes / bias
  - Sweeney; Discrimination in Online Ad Delivery. CACM 2013



What is bias? In the context of word embeddings...

- We must emphasize that biases are *multifaceted*.
- Biases (in the context of word representations) refer to *stereotypical associations* between words or groups of words that may cause representational harm (i.e. the subordination of a certain social group along the lines of identity).

But knowledge representation is a big part of AI!

Bias + Machine Learning: given bias

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

  . . . can translate into representational or allocational harm

Sources of bias

- Bias in data for training representations: e.g. wikipedia (jobs associated with gender)
- Algorithmic bias: bias retained and amplified by algorithms.
- Bias in data for training specific tasks.

  ... can translate into representational or allocational harm

WEAT Implicit Association Test

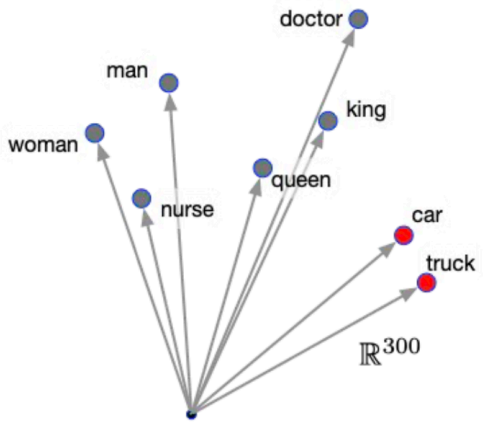
- $X = \{man, male, \dots\}$ (definitionally male words)
- $Y = \{woman, female, \dots\}$ (definitionally female words)
- $A = \{programmer, engineer, scientist, \dots\}$ (stereotypical male professions)
- $B = \{nurse, teacher, librarian, \dots\}$ (stereotypical female professions)

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

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**



Debiasing Methods for Word Embeddings

Four methods

Debiasing word embeddings

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

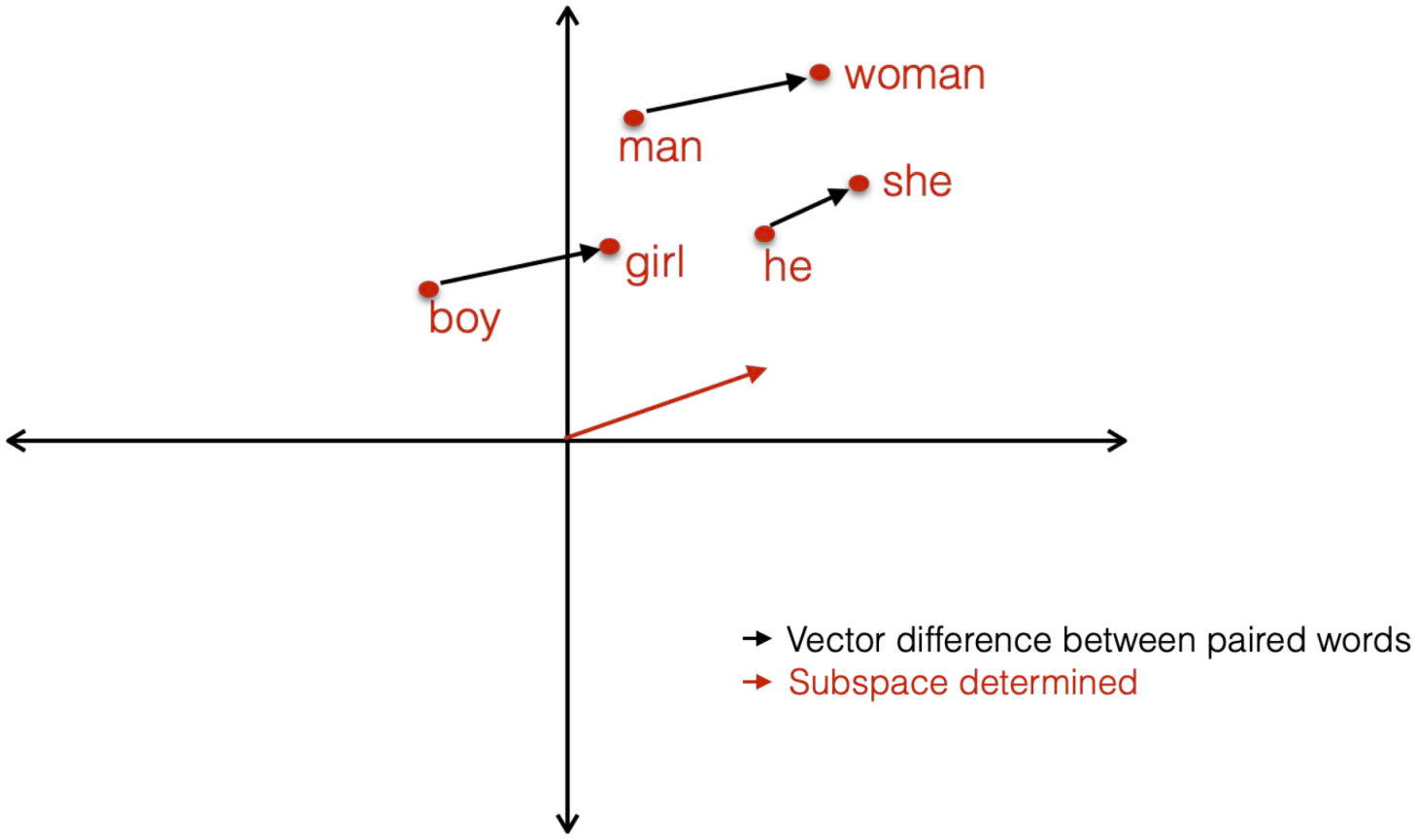
Debiasing by post processing representations

- Modulates representations to mitigate stereotypical associations.
- Easy to extend to different biases.
- Inexpensive!
- Geometric operations computed on the fly...

Feature Subspace Determination

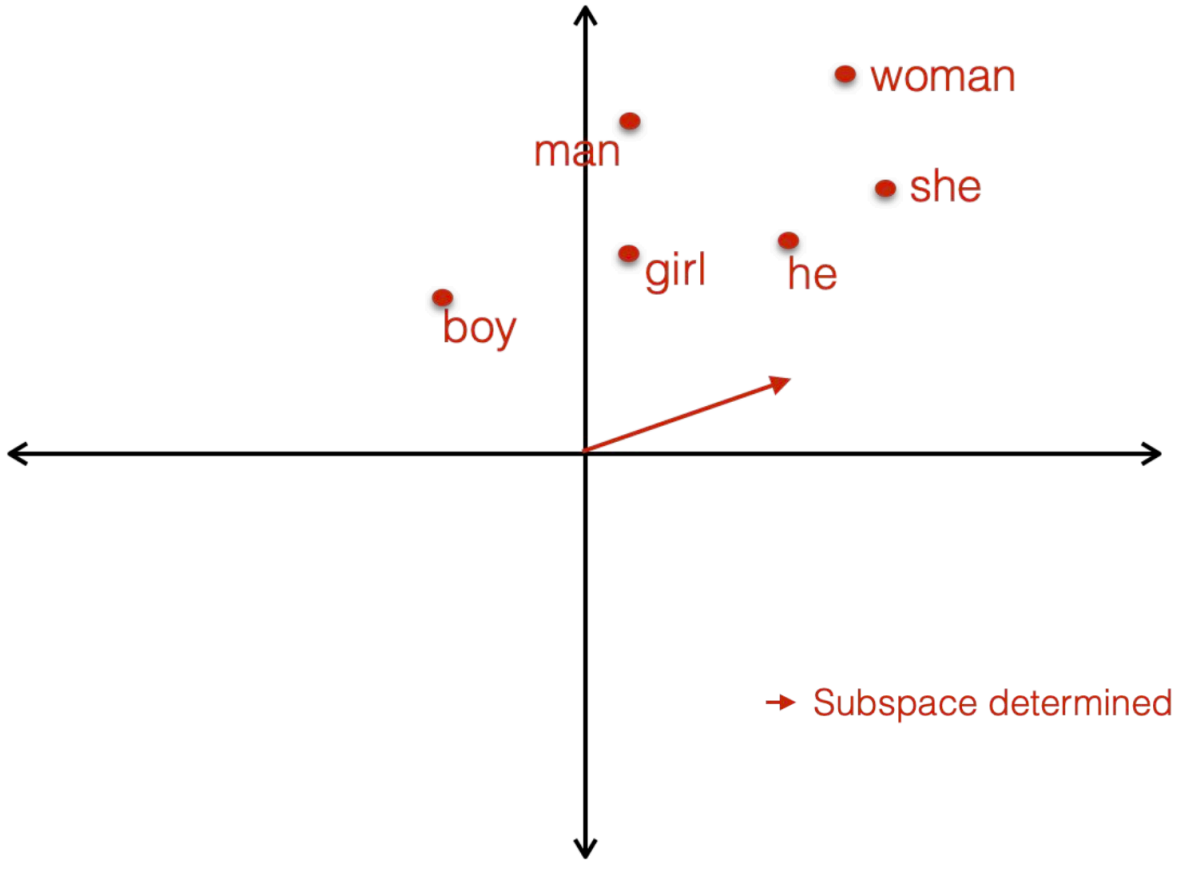
Four methods: finding bias direction/subspace!

1. PCA Paired



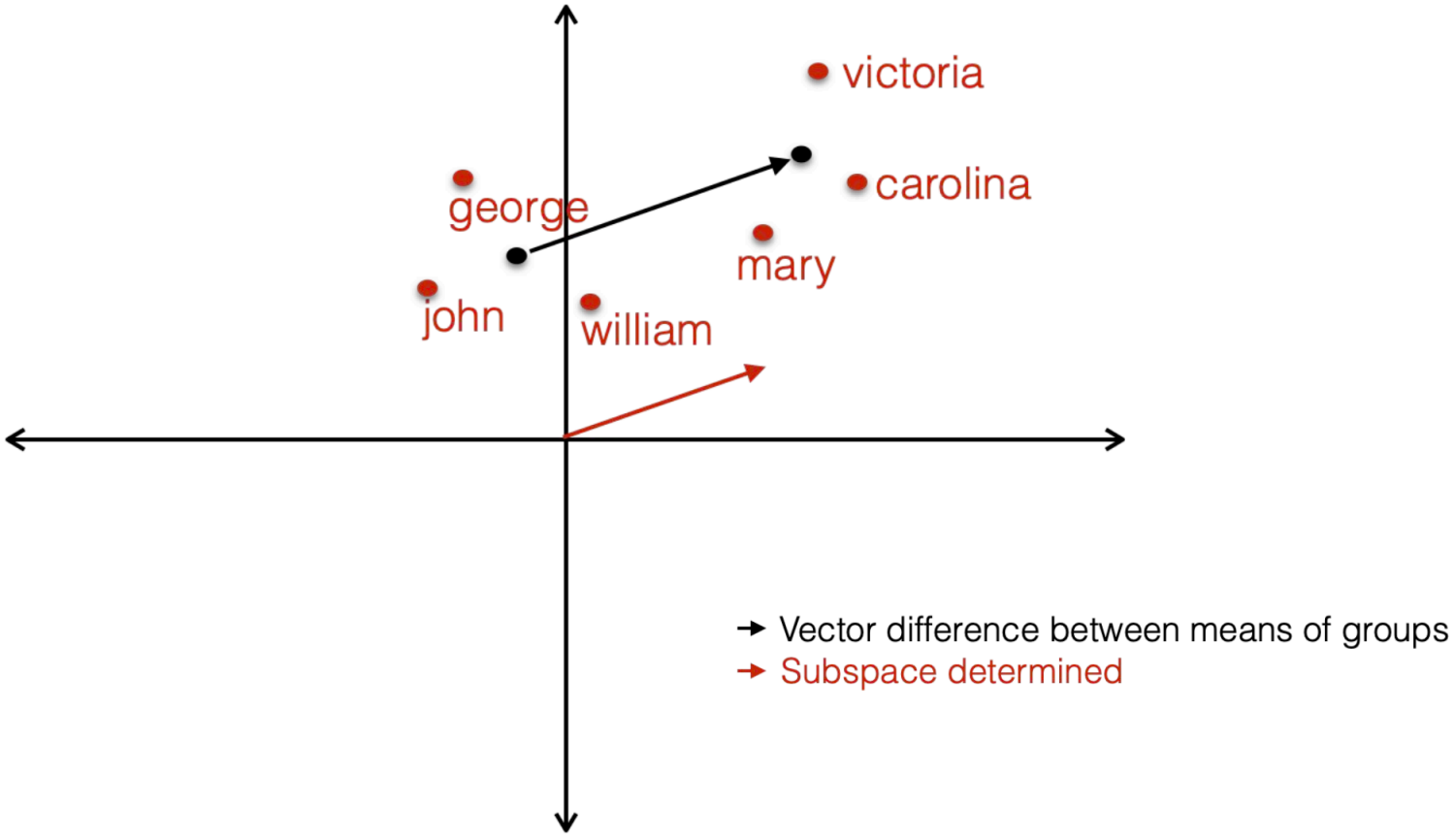
- PCA: principal component analysis
- Take pairs of words: boy-girl, he-she, ...
- Find difference vectors between them
- PC1 is the direction of gender difference (where bias lies)

2. PCA



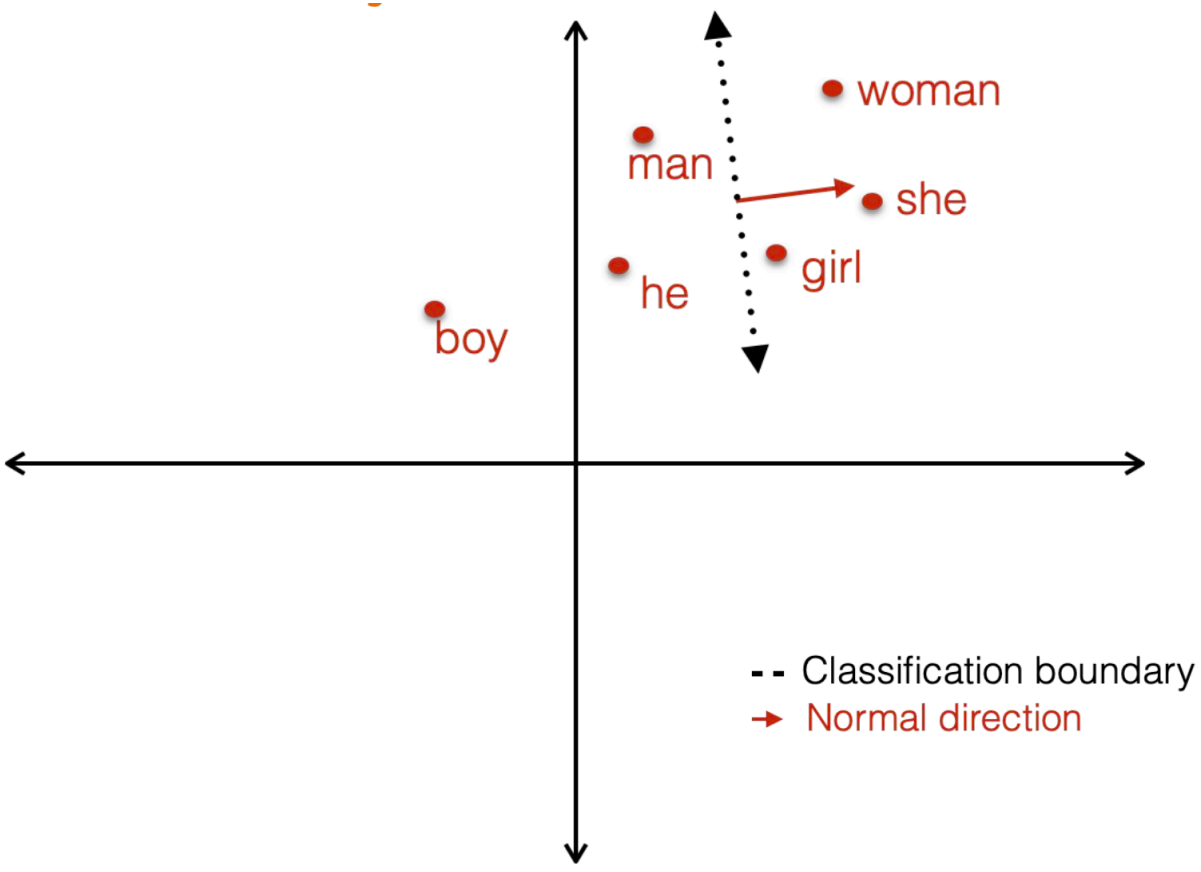
- Take words themselves: boy, girl, he, she, ...
- Find PC1: direction of gender
- More general: captures direction where a lot of bias might be associated with

3. 2-means



- Take 2 groups of words that statistically associated with social groups where we try to determine biases between, e.g. statistical/stereotypical female/male names
- Find means of 2 groups of words, and direction of difference

4. Classification boundary based



- Take 2 groups of words that statistically associated with social groups in question (requires more words)
- Find normal direction to the classification boundary as the direction of bias

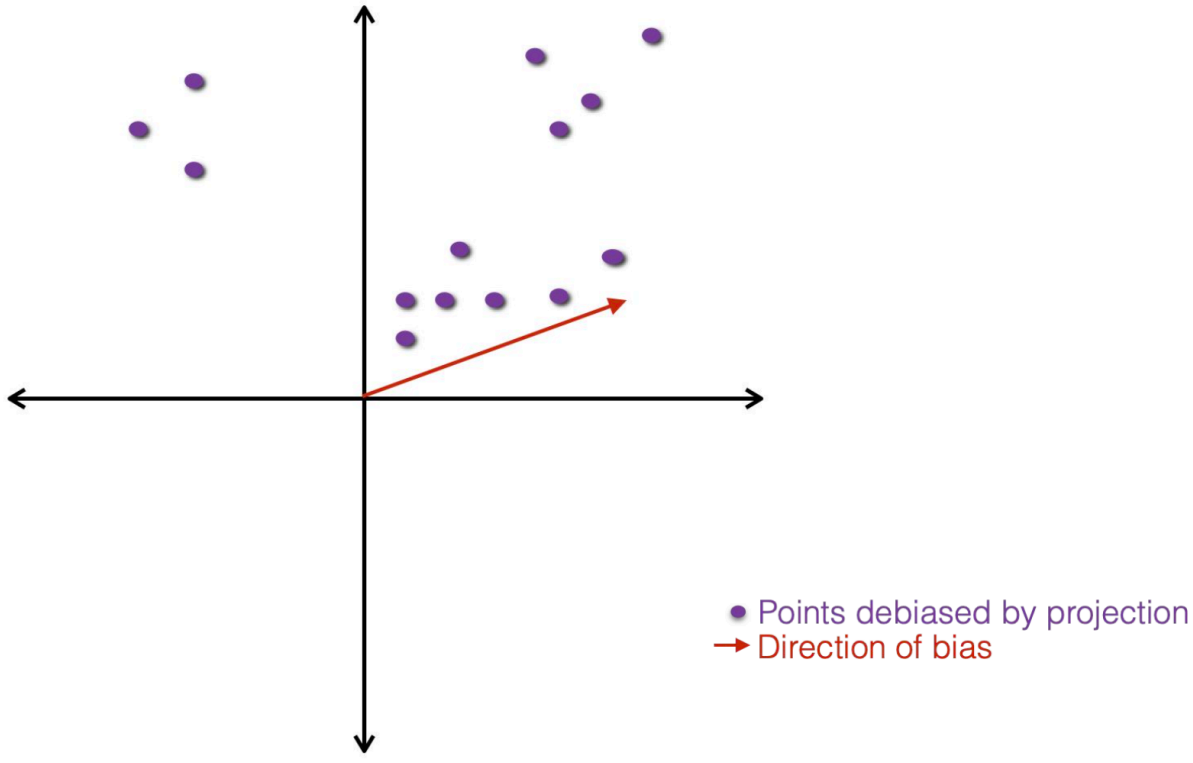
Methods to Debias Embeddings

Modify the embeddings to mitigate bias:

Linear Projection, Hard Debiasing, Iterative Nullspace Projection (INLP),

Orthogonal Subspace Correction and Rectification (OSCaR)

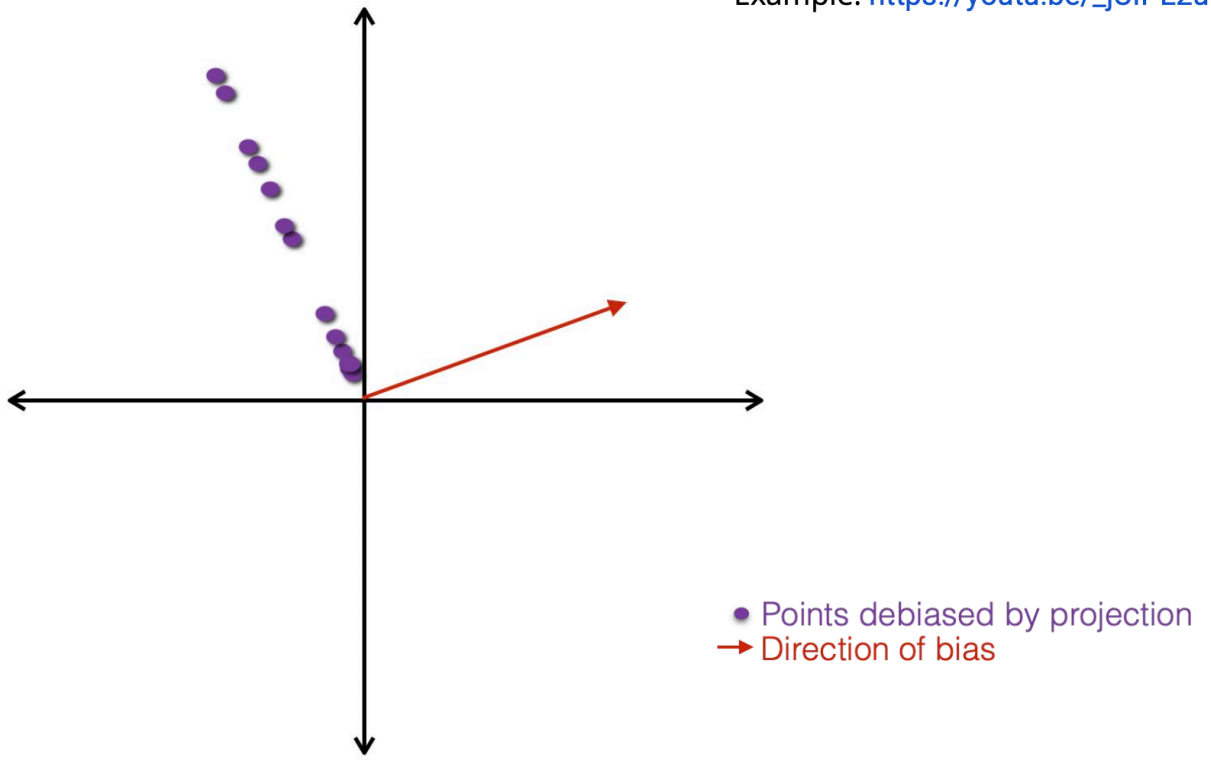
A. Linear projection



- Given a pre-determined bias directions
- Find component of *all words* in the direction of bias and remove it from the word representations. `https://youtu.be/_jUIPL2uM9M`
- Dev and Phillips; Attenuating Bias in Word Vectors. AIStats 2019

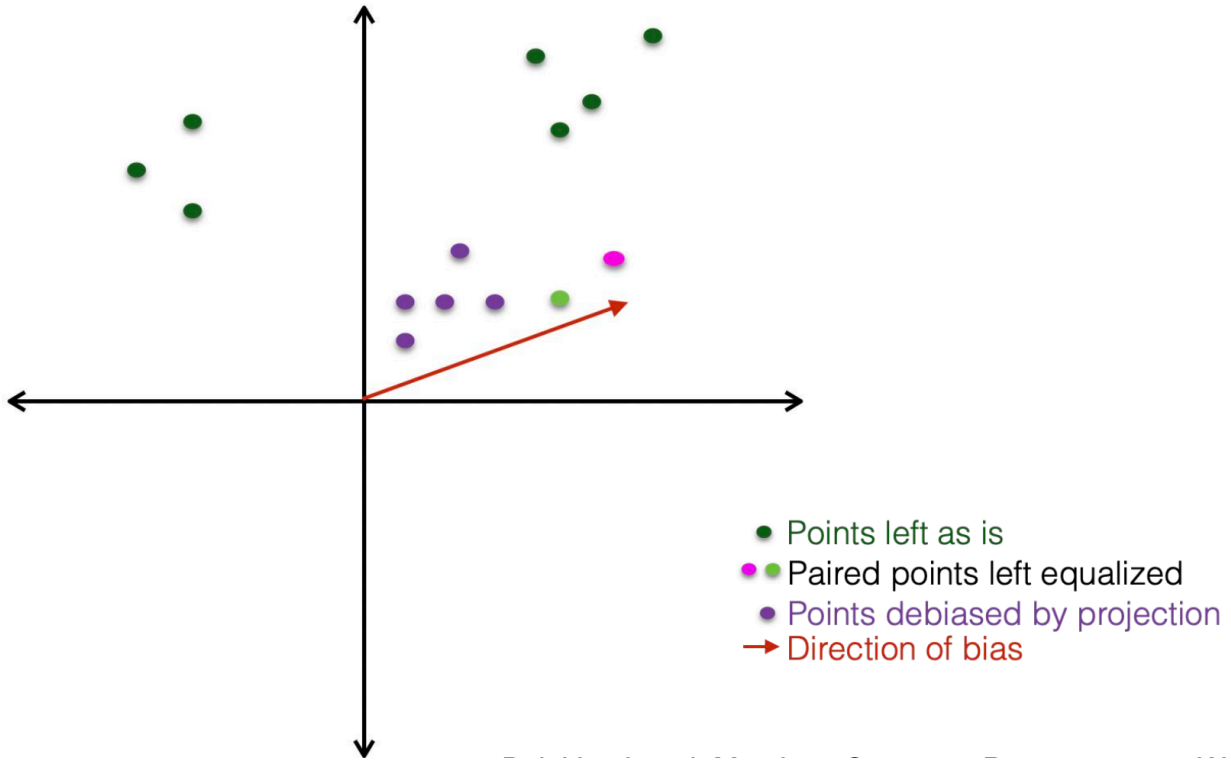
A. Linear projection

Example: https://youtu.be/_jUIPL2uM9M



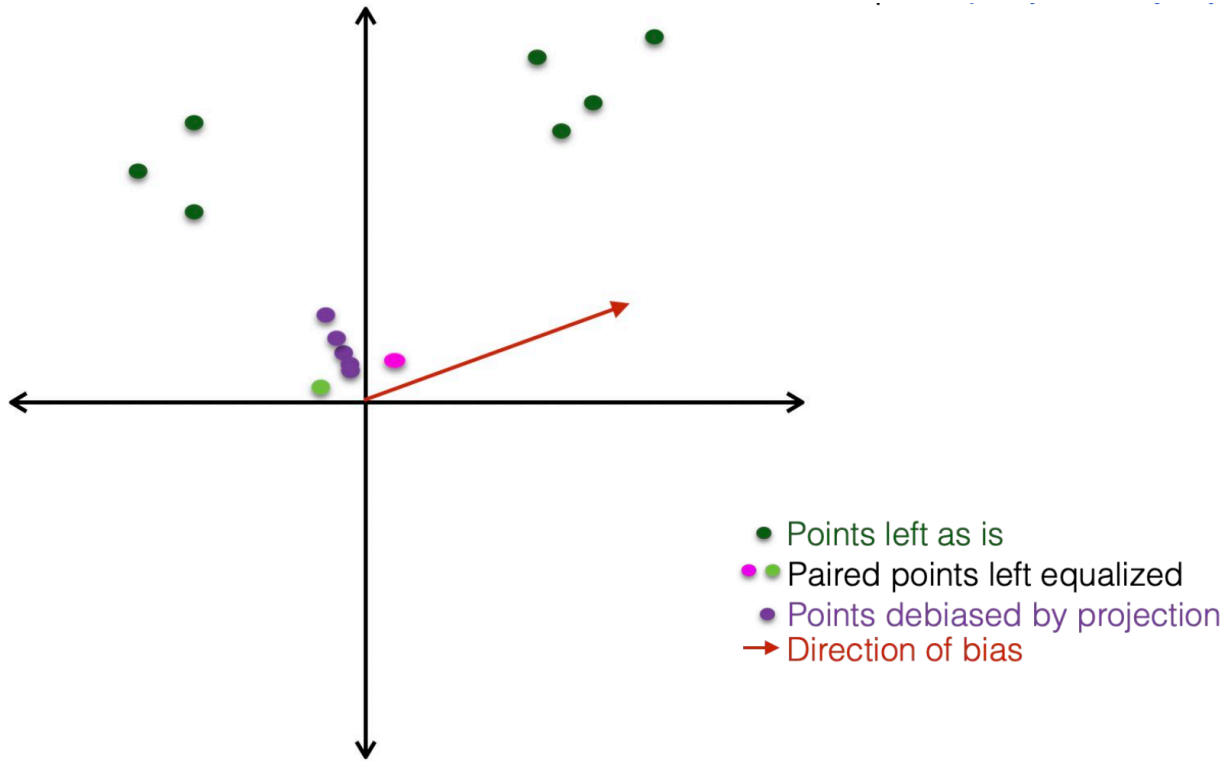
- Given a pre-determined bias directions
- Find component of *all words* in the direction of bias and remove it from the word representations. `https://youtu.be/_jUIPL2uM9M`
- Dev and Phillips; Attenuating Bias in Word Vectors. AIStats 2019

B. Hard debiasing



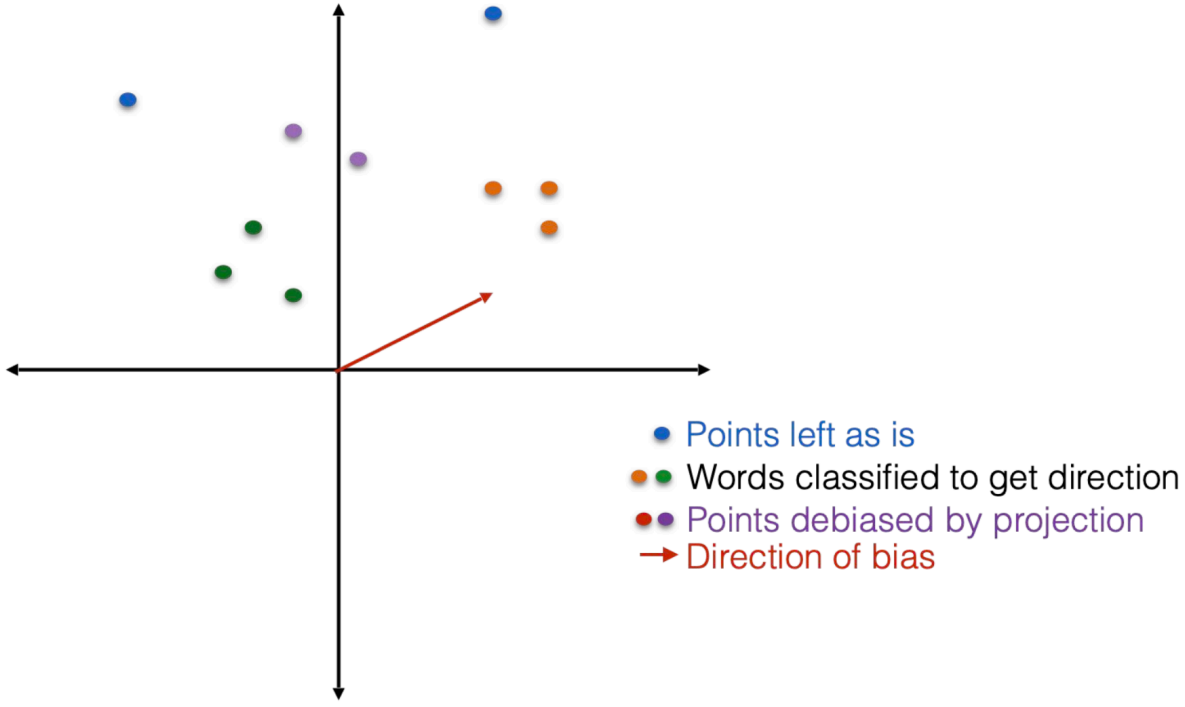
- Dark green words used to determine bias direction are left as is
- Green/pink words are used to retain (reintroduce) meaningful gender information in a controlled way `https://youtu.be/jHlFyqRAsuU`
- Bolukbasi et al; Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. NeurIPS 2016

B. Hard debiasing



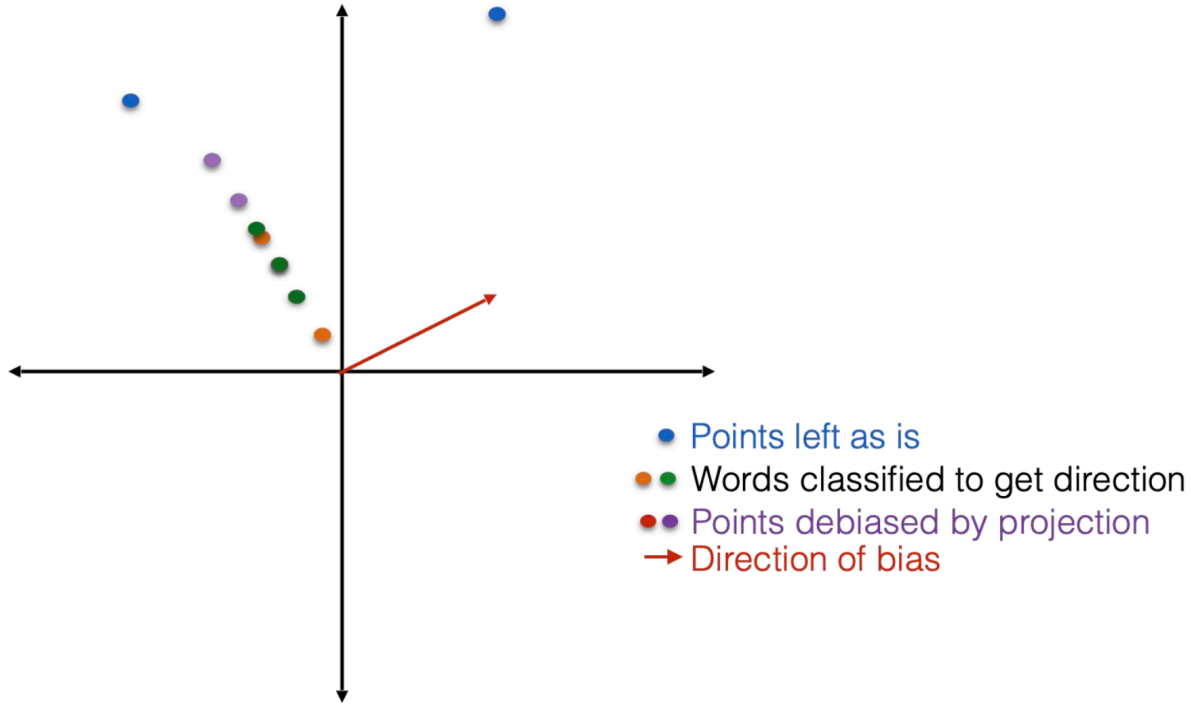
- Dark green words used to determine bias direction are left as is
- Green/pink words are used to retain (reintroduce) meaningful gender information in a controlled way `https://youtu.be/jHlFyqRAsuU`
- Bolukbasi et al; Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. NeurIPS 2016

C. Iterative Nullspace Projection (INLP)



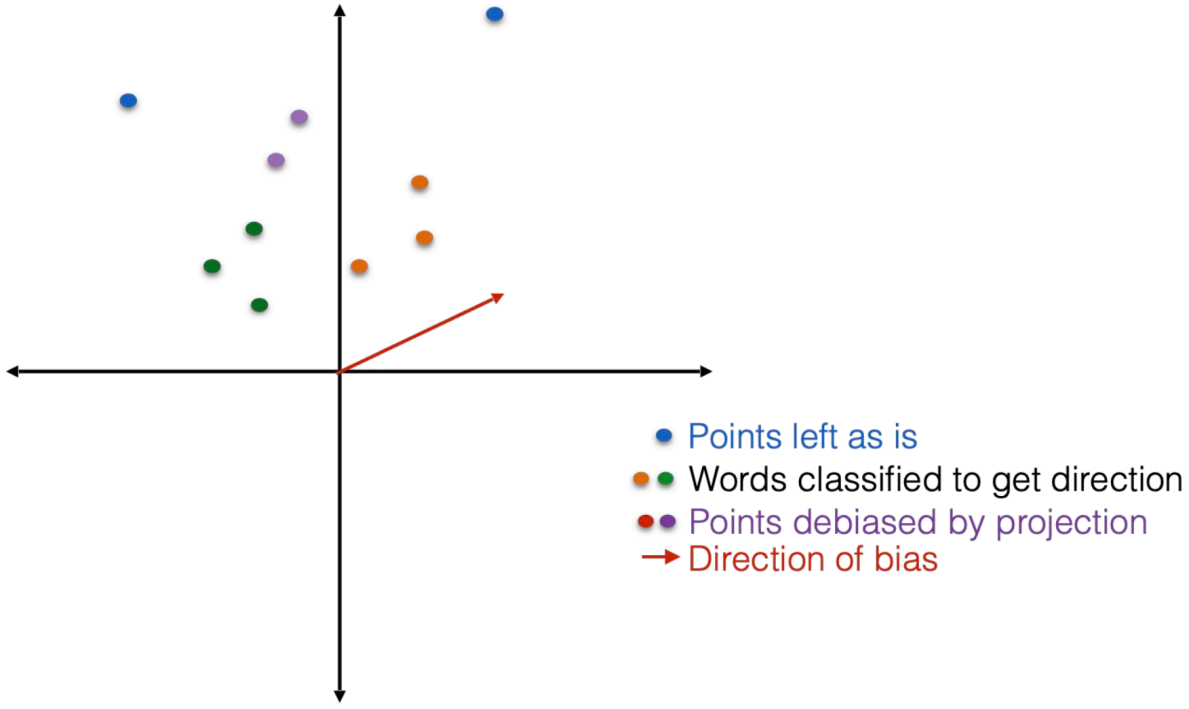
- Blue points are left as is; blue and orange groups are used to decide on a classification boundary thus a bias direction
- All purple, green and orange points are projected to remove components along the bias direction;
  `https://youtu.be/QPnioBIszxE`
- Do this iteratively until sufficient residual biases are removed
- Ravfogel et al; ACL 2020.

C. Iterative Nullspace Projection (INLP)



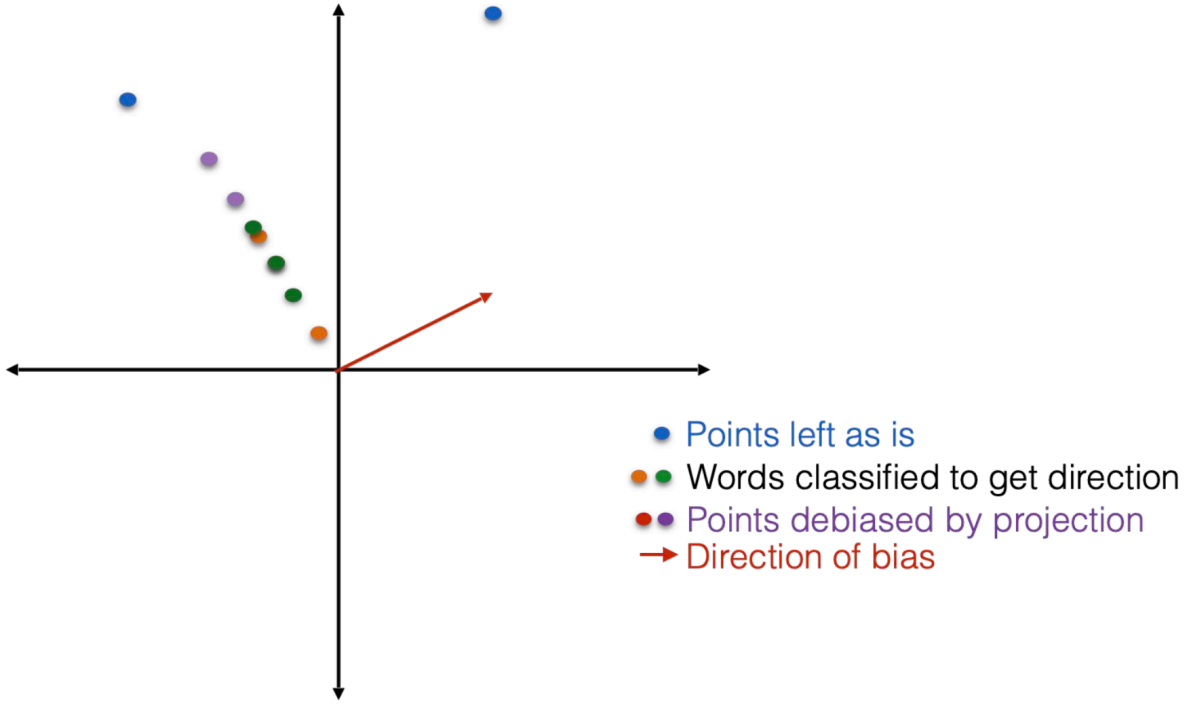
- Blue points are left as is; blue and orange groups are used to decide on a classification boundary thus a bias direction
- All purple, green and orange points are projected to remove components along the bias direction;
  `https://youtu.be/QPnioBIszxE`
- Do this iteratively until sufficient residual biases are removed
- Ravfogel et al; ACL 2020.

C. Iterative Nullspace Projection (INLP)



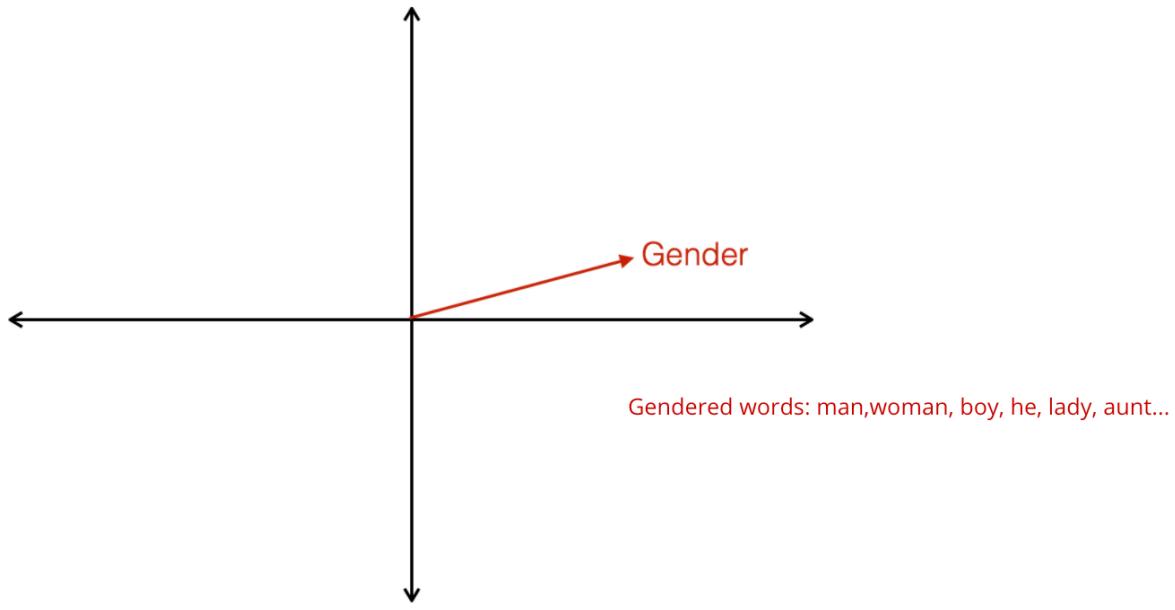
- Blue points are left as is; blue and orange groups are used to decide on a classification boundary thus a bias direction
- All purple, green and orange points are projected to remove components along the bias direction;
  `https://youtu.be/QPnioBIszxE`
- Do this iteratively until sufficient residual biases are removed
- Ravfogel et al; ACL 2020.

C. Iterative Nullspace Projection (INLP)



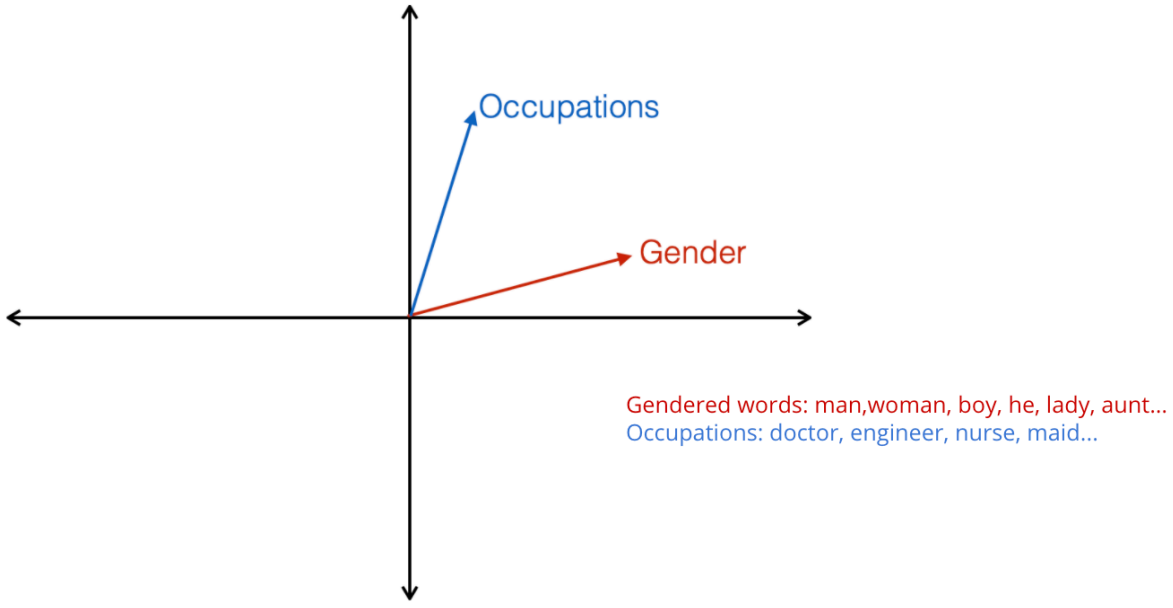
- Blue points are left as is; blue and orange groups are used to decide on a classification boundary thus a bias direction
- All purple, green and orange points are projected to remove components along the bias direction; `https://youtu.be/QPnioBIszxE`
- Do this iteratively until sufficient residual biases are removed
- Ravfogel et al; ACL 2020.

D. Orthogonal Subspace Correction and Rectification



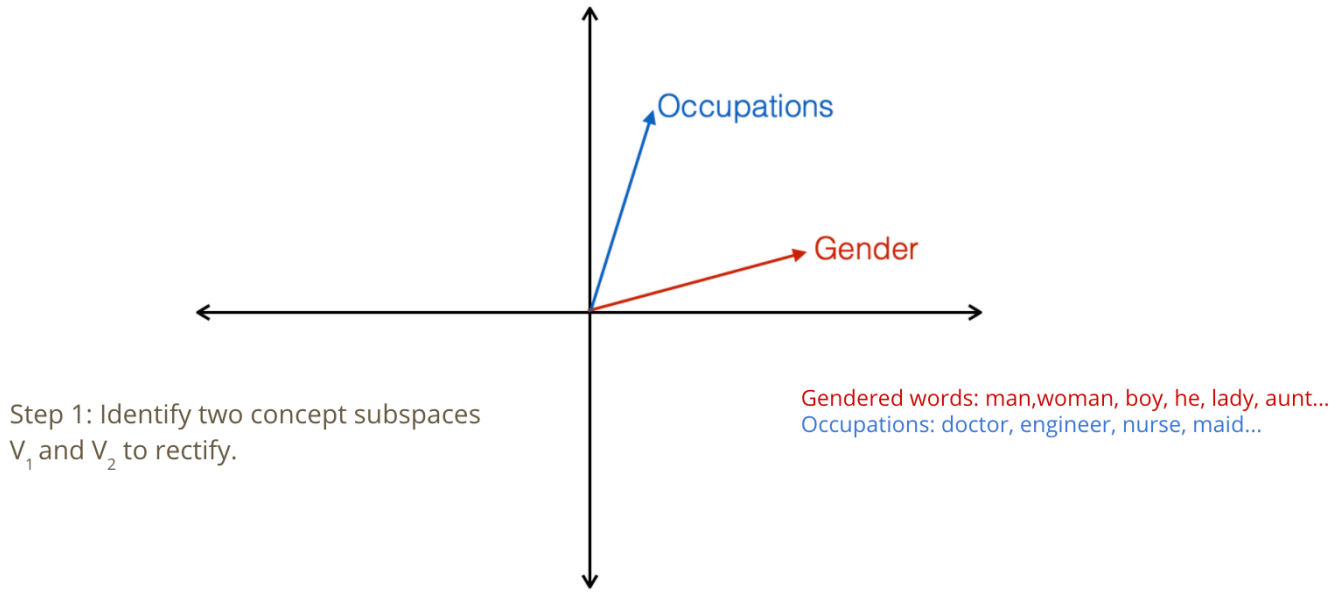
- Dev et al; OSCaR: Orthogonal Subspace Correction and Rectification of Biases in Word Embeddings. arXiv:2007.00049. 2020
- `https://youtu.be/H1sa_GsxQdc`
- Biase are not single-directional; instead of removing bias direction, dis-associated two subspaces (gender vs occupation) by making them orthogonal to teach other.

D. Orthogonal Subspace Correction and Rectification



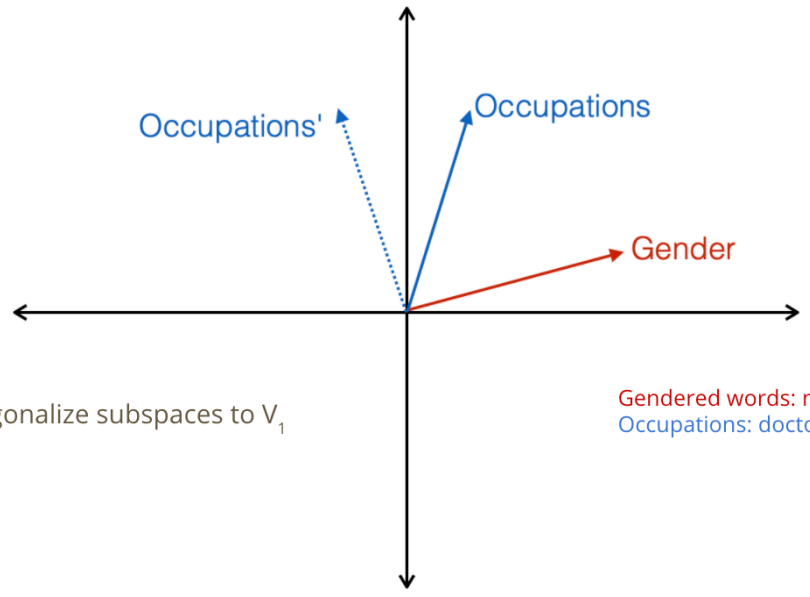
- Dev et al; OSCaR: Orthogonal Subspace Correction and Rectification of Biases in Word Embeddings. arXiv:2007.00049. 2020
- `https://youtu.be/H1sa_GsxQdc`
- Biase are not single-directional; instead of removing bias direction, dis-associated two subspaces (gender vs occupation) by making them orthogonal to teach other.

D. Orthogonal Subspace Correction and Rectification



- Dev et al; OSCaR: Orthogonal Subspace Correction and Rectification of Biases in Word Embeddings. arXiv:2007.00049. 2020
- `https://youtu.be/H1sa_GsxQdc`
- Biase are not single-directional; instead of removing bias direction, dis-associated two subspaces (gender vs occupation) by making them orthogonal to teach other.

D. Orthogonal Subspace Correction and Rectification

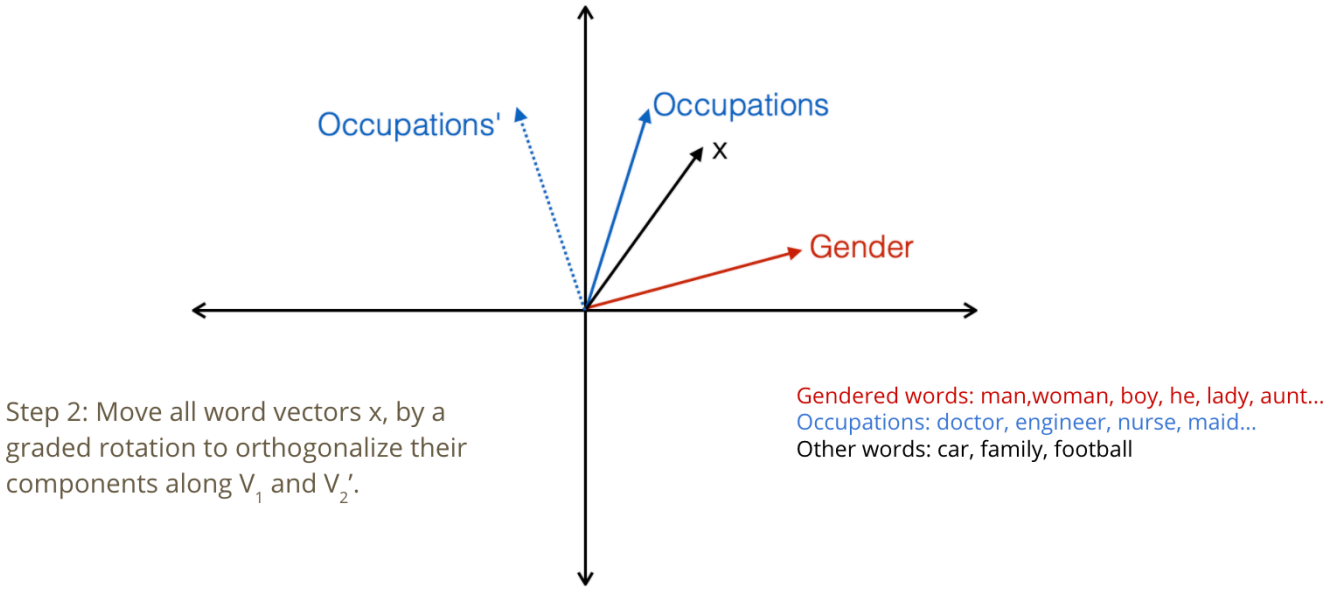


Step 2: Orthogonalize subspaces to V_1 and V_2'.

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

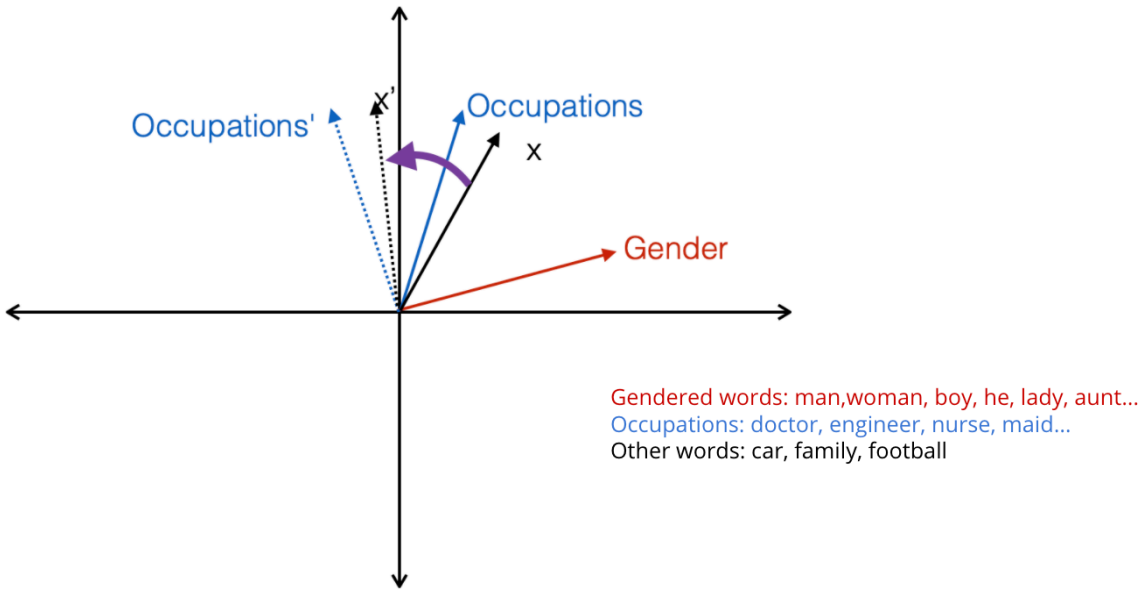
- Dev et al; OSCaR: Orthogonal Subspace Correction and Rectification of Biases in Word Embeddings. arXiv:2007.00049. 2020
- `https://youtu.be/H1sa_GsxQdc`
- Biase are not single-directional; instead of removing bias direction, dis-associated two subspaces (gender vs occupation) by making them orthogonal to teach other.

D. Orthogonal Subspace Correction and Rectification



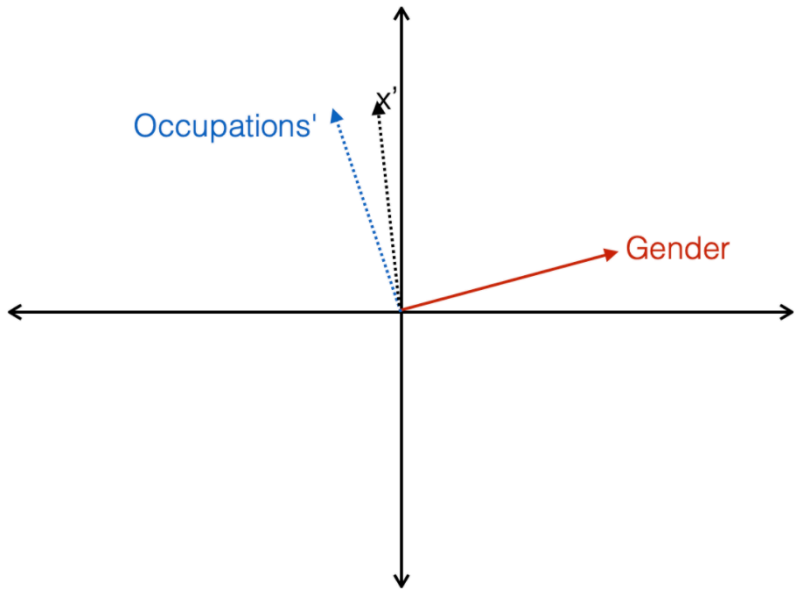
- Dev et al; OSCaR: Orthogonal Subspace Correction and Rectification of Biases in Word Embeddings. arXiv:2007.00049. 2020
- `https://youtu.be/H1sa_GsxQdc`
- Biase are not single-directional; instead of removing bias direction, dis-associated two subspaces (gender vs occupation) by making them orthogonal to teach other.

D. Orthogonal Subspace Correction and Rectification



- Dev et al; OSCaR: Orthogonal Subspace Correction and Rectification of Biases in Word Embeddings. arXiv:2007.00049. 2020
- `https://youtu.be/H1sa_GsxQdc`
- Biase are not single-directional; instead of removing bias direction, dis-associated two subspaces (gender vs occupation) by making them orthogonal to teach other.

D. Orthogonal Subspace Correction and Rectification



- Dev et al; OSCaR: Orthogonal Subspace Correction and Rectification of Biases in Word Embeddings. arXiv:2007.00049. 2020
- `https://youtu.be/H1sa_GsxQdc`
- Biase are not single-directional; instead of removing bias direction, dis-associated two subspaces (gender vs occupation) by making them orthogonal to teach other.

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

Visual Demo

Hands on examples!

`https://github.com/tdavislab/verb`

`http://archit.sci.utah.edu:5001/`

Example: linear projection

- Linear projection: 2-means. Gender. Rename concept 1 to gender.
- Define gender direction using two groups of words. Group 1, *he, him*, Group 2, *she, her*
- How does gender direction defined by these two groups of words visually correspond to occupation?
- Evaluation set: *engineer, banker, nurse, receptionist*
- Or preload Example 1
- Initial embedding with gender direction (determined by 2-means)
- Step by step animation (default: 50D Glove embeddings)
  1. Reorient camera view: see how occupation embeddings encode stereotype on gender
  2. Apply LP to remove this bias (remove gender components from every words in the 50D embedding)
  3. Reorient camera view of 50D debiased embedding (can not separate green from orange points)

Example: hard debiasing

- What to de-bias, not to de-bias, to equalize
- Add *man* to the male word list
- Equalize set: de-bias and reintroduce gender information in a controlled way (equal along the bias direction): *himself-herself, boy-girl*
- Evaluation set: see how de-biasing change its association with the gender direction
- Step by step animation
  1. Reorient camera view: see how occupation embeddings encode stereotype on gender
  2. Apply HD to remove this bias
  3. Reintroduce and retain gender information in a controlled way
  4. Reorient camera view of 50D debiased embedding

Example: OSCaR

- Pre-load example 8 (remove name and gender association; small angle means correlation; rectify correlation)
- Two subspaces: gender and occupation
- Step by step animation
  1. Reorient camera view so gender is along the x-axes; observe the components defined by these two concepts; genders are horizontal; occupation along the other axes.
  2. Apply OSCaR to smooth, graded rotation of the space; gender and occupation axes are orthogonal in the span; the other 48 dimension stays the same; gendered words are still horizontal aligned; occupations are not correlated with gender any longer (vertically aligned); see the word *engineer* moves with the occupation axes; we are making sure components are orthogonal with each other.
  3. Reorient camera view of 50D debiased embedding using PCA; tiger and tigeress gender information is retained. We do not loose gender information completely.

Example: INLP

- Pre-load example, gender (not seeding with 10K words)
- Hard to understand, use larger word set (paper 300D; here 50D); evaluation set: engineer and homemaker
- Find classification boundary based gender direction using two gendered groups
- Step by step animation
  1. Reorient camera view using PCA (PC1 vs PC2); green-orange direction
  2. Apply INLP to find gender classification boundary (we are see the projected view);
  3. Reorient camera view: x-axes is the gender direction, y-axes best PCA, green words are separated from orange sets, jack and Jill are in the mix; and apply linear projection along the gender axes; we now have 49D subspace; reorient back to 49 dimension; we reduce some correlation; still have residual bias, find another correlation. 2nd direction separating Jack and Jill (residual gender information). Paper: 35 iteration for 300D space.
  4. Repeat until convergence. Points have no correlation and spread out as much as possible.

Example: Nationality with LP

- Pre-load example 9 and example 10
- Age, nationality biases, etc.

Example: royalty

- Are loyal people better than common people? Remove that idea from embedding?
- LP with 2-means
- *Royalty*, definitionally *royal*: *king, queen, prince, princess*; definitionally *not royal*: *man, woman, boy, girl*; evaluation set: sentimatal, *great, mediocre*
- observe separation from orange and green
- Reorient: great more on the royalty side; want to remove
- Apply LP: remove correlation between royalty and the aces between great and mediocre
- man $\to$ king: point up; girl $\to$ princess: point down
- Fun toy example: not just gender bias

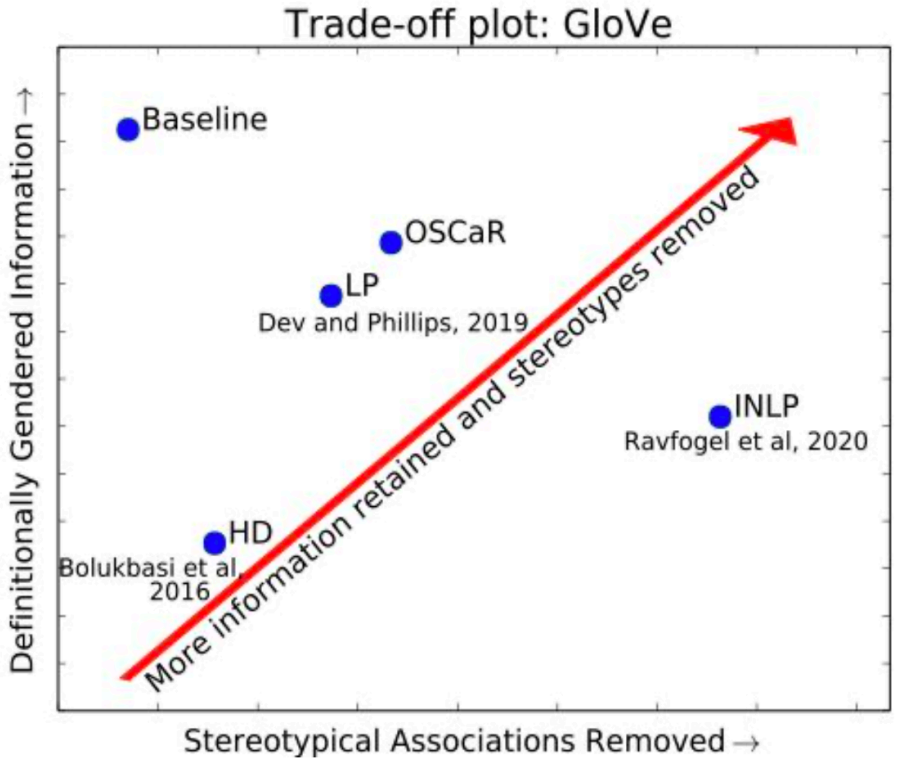
Discussions and Moving Ahead

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.

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.



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

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
- Financial data
- 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?

Want to learn more?

- VERB: Visualizing and Interpreting Bias Mitigation Techniques for Word Representations. Archit Rathore, Sunipa Dev, Jeff M. Phillips, Vivek Srikumar, Yan Zheng, Chin-Chia Michael Yeh, Junpeng Wang, Wei Zhang, Bei Wang. Manuscript, 2021. arXiv:2104.02797.