Understanding the Origins of Bias in Word Embeddings

Marc-Etienne Brunet
Colleen Alkalay-Houlihan
Ashton Anderson
Richard Zemel

UNIVERSITY OF TORONTO
VECTOR INSTITUTE | INSTITUT VECTEUR

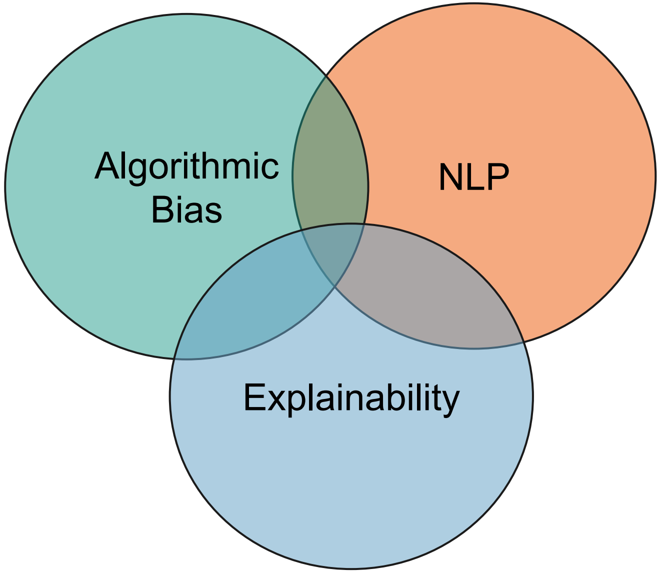
Introduction

Graduate student at U of T (Vector Institute)

Work at the intersection of Bias, Explainability, and Natural Language Processing

Collaborated with Colleen Alkalay-Houlihan

Supervised by Ashton Anderson and Richard Zemel



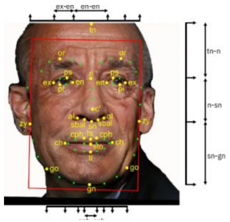
UNIVERSITY OF TORONTO

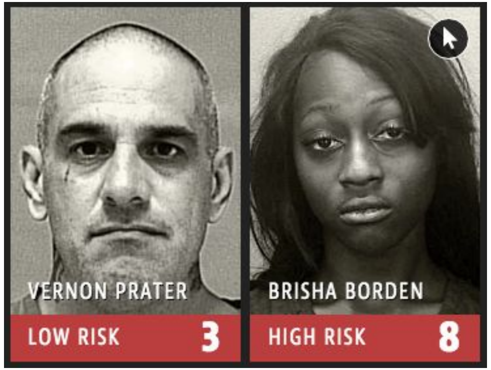
VECTOR INSTITUTE | INSTITUT VECTEUR

Many Forms of Algorithmic Bias

For example:

- Facial Recognition
- Automated Hiring
- Criminal Risk Assessment
- Word Embeddings

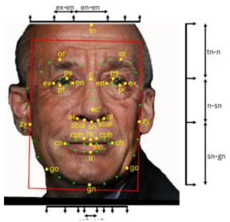


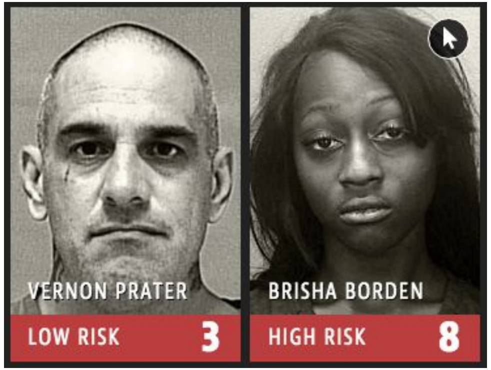


Many Forms of Algorithmic Bias

For example:

- Facial Recognition
- Automated Hiring
- Criminal Risk Assessment
- Word Embeddings





How can we **attribute** the **bias** in word embeddings **to** the individual **documents** in their training corpora?

> **Background**

Method Overview

Critical Details

Experiments

Word Embeddings: Definitions in Vector Space

Definitions encode relationships between words

lead·er
/ˈlēdər/

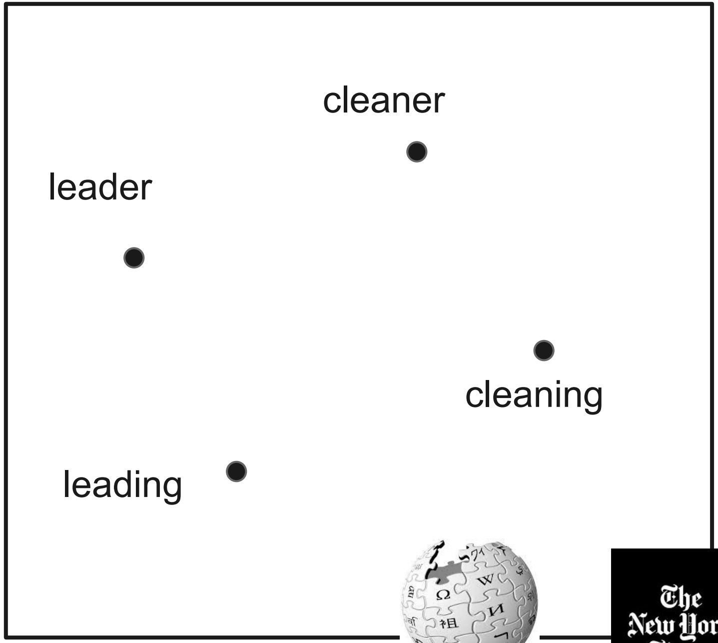
noun

1. the person who leads or commands a group, organization, or country.
"the leader of a protest group"
synonyms: chief, head, principal, boss; More

cleaning
/ˈklēniNG/

noun
noun: **cleaning**

the action of making something clean, especially the inside of a house.
"the housekeeper will help with the cleaning"



WIKIPEDIA
The Free Encyclopedia

The New York Times

Word Embeddings: Definitions in Vector Space

Definitions encode relationships between words

lead·er
/ˈlēdər/

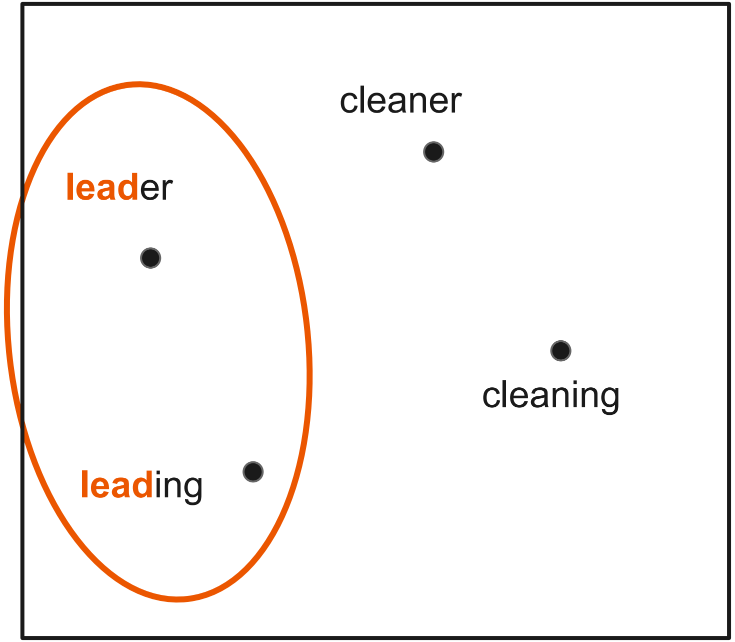
noun

1. the person who leads or commands a group, organization, or country.
   "the leader of a protest group"
   synonyms: chief, head, principal, boss; More

cleaning
/ˈklēniNG/

noun
noun: **cleaning**

the action of making something clean, especially the inside of a house.
"the housekeeper will help with the cleaning"



Word Embeddings: Definitions in Vector Space

lead·er
/ˈlēdər/

noun

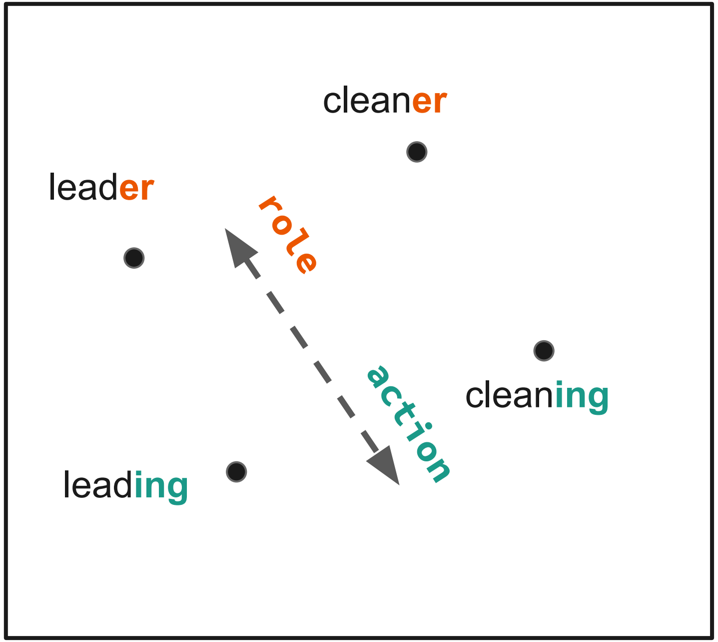
1. the person who leads or commands a group, organization, or country.
   "the leader of a protest group"
   synonyms: chief, head, principal, boss; More

cleaning
/ˈklēniNG/

noun
noun: **cleaning**

the action of making something clean, especially the inside of a house.
"the housekeeper will help with the cleaning"

Definitions encode relationships between words



Problematic Definitions in Vector Space

Definitions encode relationships between words

lead·er
/ˈlēdər/

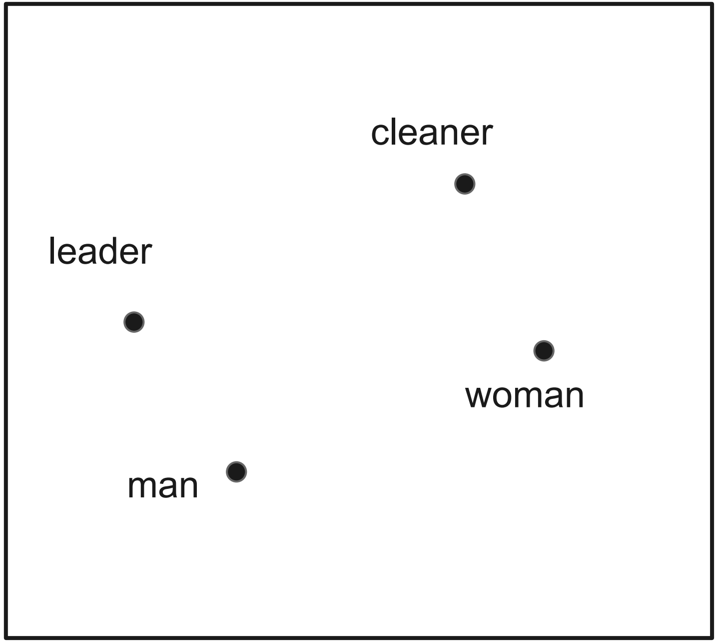
noun

1. the person who leads or commands a group, organization, or country.
   "the leader of a protest group"
   synonyms: chief, head, principal, boss; More

cleaning
/ˈklēniNG/

noun
noun: **cleaning**

the action of making something clean, especially the inside of a house.
"the housekeeper will help with the cleaning"



Problematic Definitions in Vector Space

Definitions encode relationships between words

lead·er
/ˈlēdər/

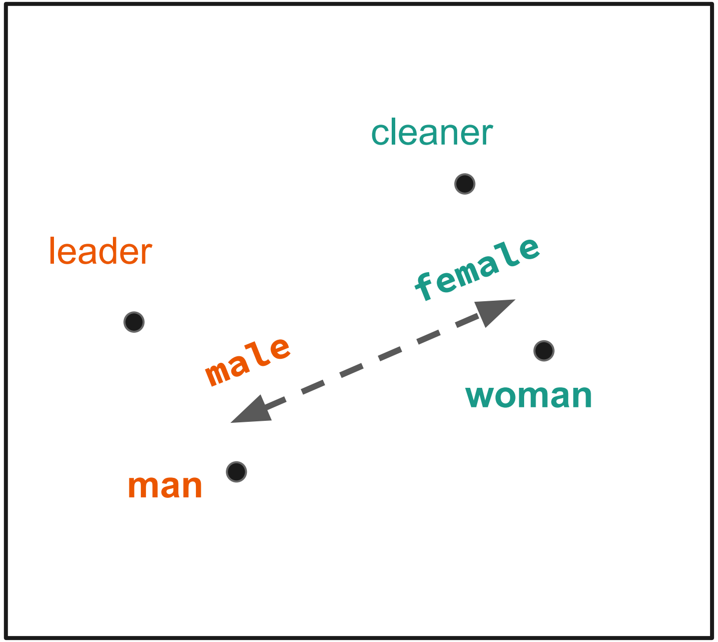
noun

1. the ~~person~~ **man** who leads or commands a group, organization, or country.
   "the leader of a protest group"
   synonyms: chief, head, principal, boss; More

cleaning
/ˈklēniNG/

noun
noun: **cleaning**

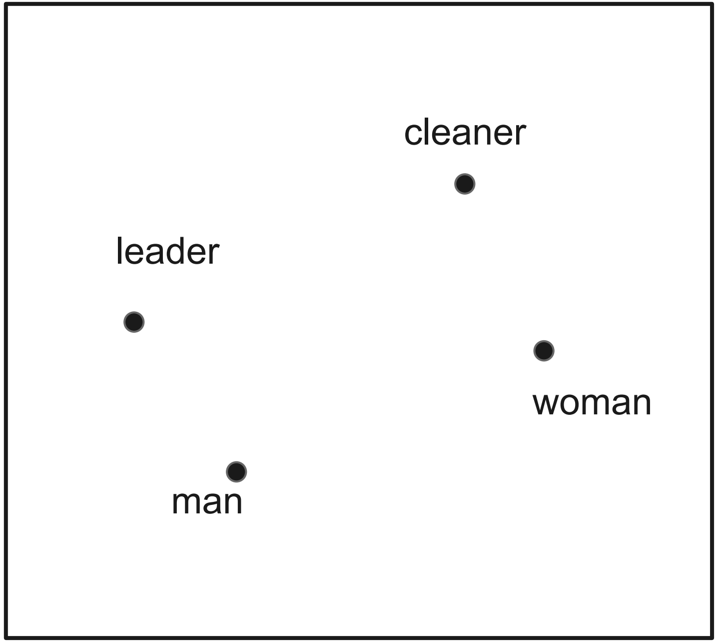
the action of **a woman** making something clean, especially the inside of a house.
"the housekeeper will help with the cleaning"



Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, Adam Kalai (NeurIPS 2016)

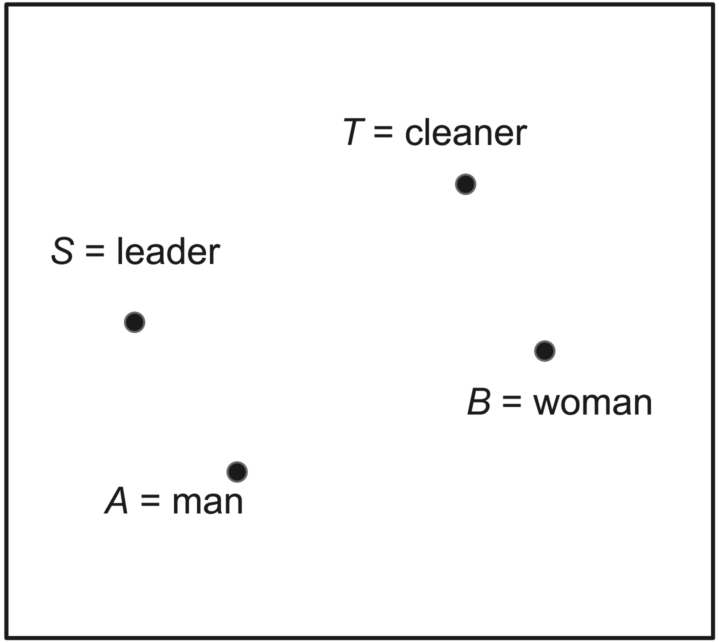
Measuring Bias in Word Embeddings

How can we **measure** bias in word embeddings?



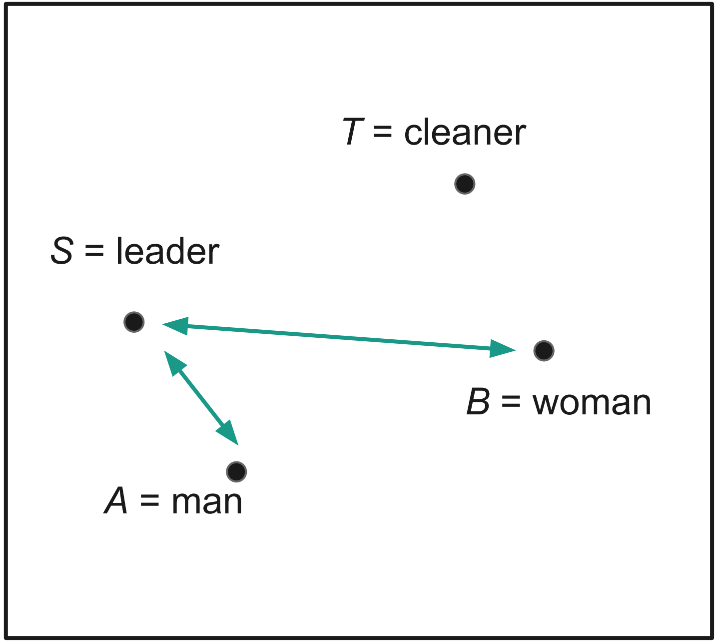
Measuring Bias in Word Embeddings

Implicit **A**ssociation **T**est
(IAT)



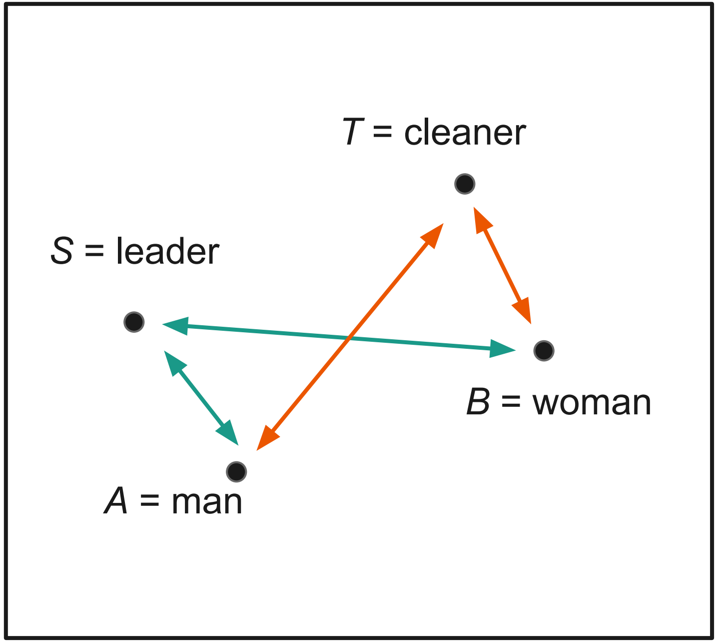
Measuring Bias in Word Embeddings

Implicit **A**ssociation **T**est
(IAT)



Measuring Bias in Word Embeddings

Implicit **A**ssociation **T**est
(IAT)

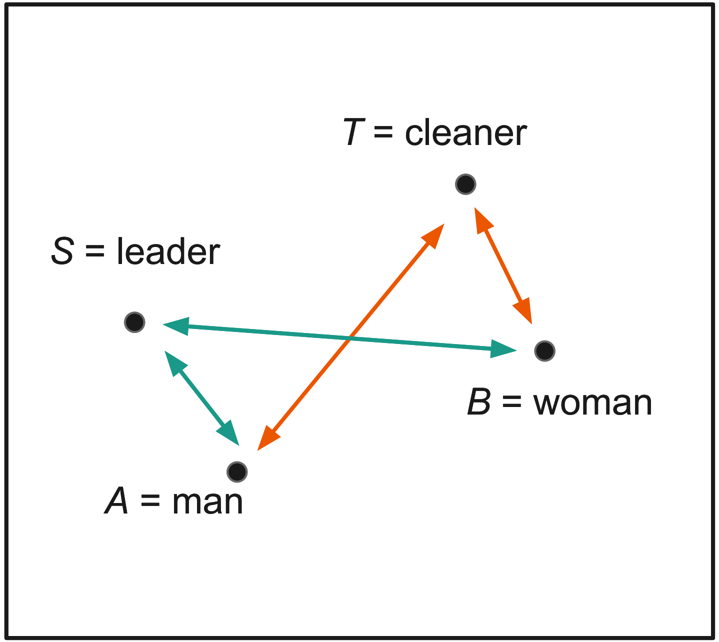


Measuring Bias in Word Embeddings

Implicit **A**ssociation **T**est
(IAT)

Word **E**mbedding **A**ssociation **T**est
(WEAT)

$$\text{Association}_{S,A} \approx \Sigma_{S,A} \cos(s,a)$$



Aylin Caliskan, Joanna J. Bryson, Arvind Narayanan (Science 2017)

Measuring Bias

WEAT on popular corpora matches IAT study results

		IAT		WEAT	
Target Words	**Attribute Words**	**effect size**	**p-val**	**effect size**	**p-val**
Flowers v.s. Insects	Pleasant v.s. Unpleasant	1.35	1.0E-08	1.5	1.0E-07
Math v.s. Arts	Male v.s. Female Terms	0.82	1.0E-02	1.06	1.8E-02
...	...	...		...	

Aylin Caliskan, Joanna J. Bryson, Arvind Narayanan (Science 2017)

Measuring Bias

WEAT on popular corpora matches IAT study results

		IAT		WEAT	
Target Words	**Attribute Words**	**effect size**	**p-val**	**effect size**	**p-val**
Flowers v.s. Insects	Pleasant v.s. Unpleasant	1.35	1.0E-08	1.5	1.0E-07
Math v.s. Arts	Male v.s. Female Terms	0.82	1.0E-02	1.06	1.8E-02
…	…	…		…	

"Semantics derived automatically from language corpora **contain human-like biases**"

Aylin Caliskan, Joanna J. Bryson, Arvind Narayanan (Science 2017)

Background

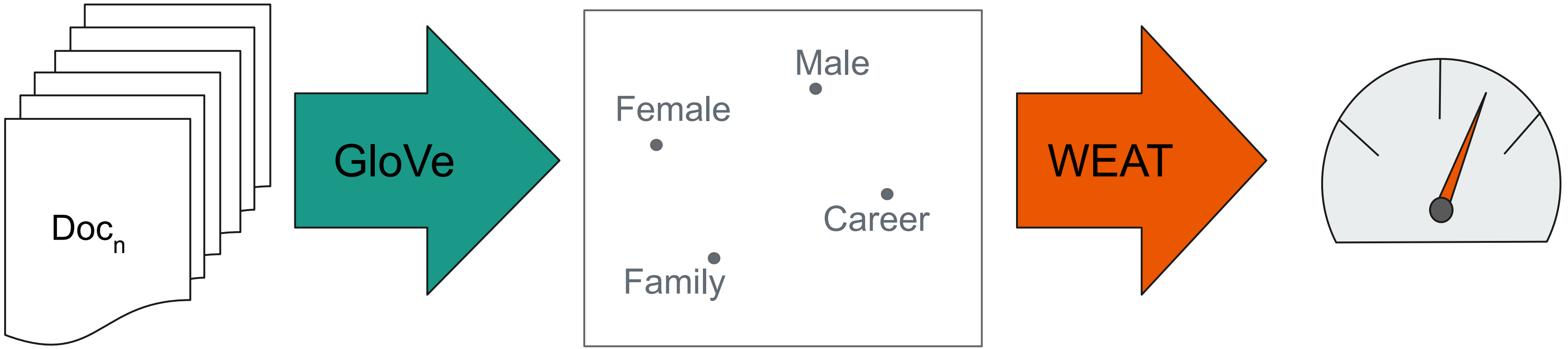
> **Method Overview**

Critical Details

Experiments

How can we **attribute** the **bias** in word embeddings **to** the individual **documents** in their training corpora?

From Word2Bias



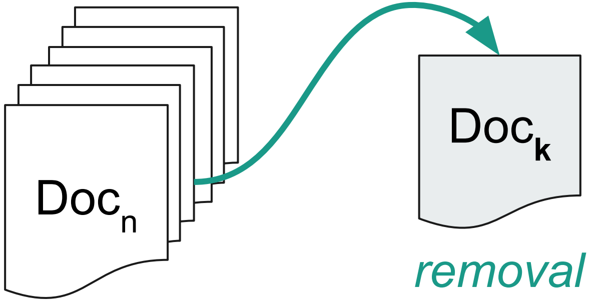
X : Corpus
(e.g. Wikipedia)

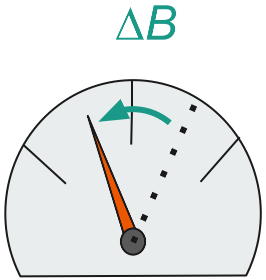
$\{ w_i \} = w(X)$
Word Embedding

$B(w(X))$
Bias Measured

Differential Bias

Idea: Consider the **differential contribution** of each document

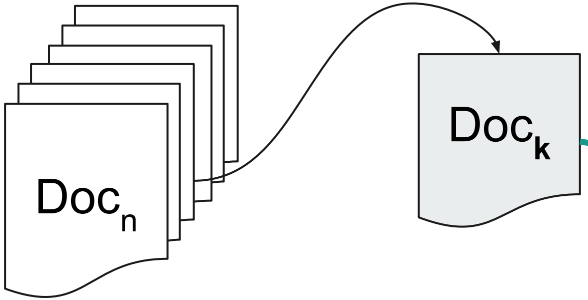




$$X = \sum_{i=1}^{n} X^{(i)} \qquad \tilde{X} = X - X^{(k)}$$

$$\Delta B = B(w(X)) - B(w(\tilde{X}))$$

Differential Bias



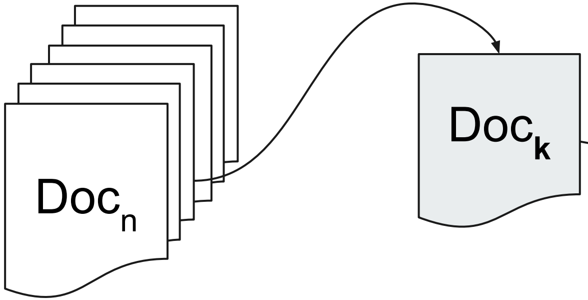
$$X = \sum_{i=1}^{n} X^{(i)} \qquad \tilde{X} = X - X^{(k)}$$

ΔB

Bias Attributed

Document ID	ΔB
1	-0.0014
2	0.0127
...	...
k	**0.0374**
...	...
n	0.0089

Differential Bias



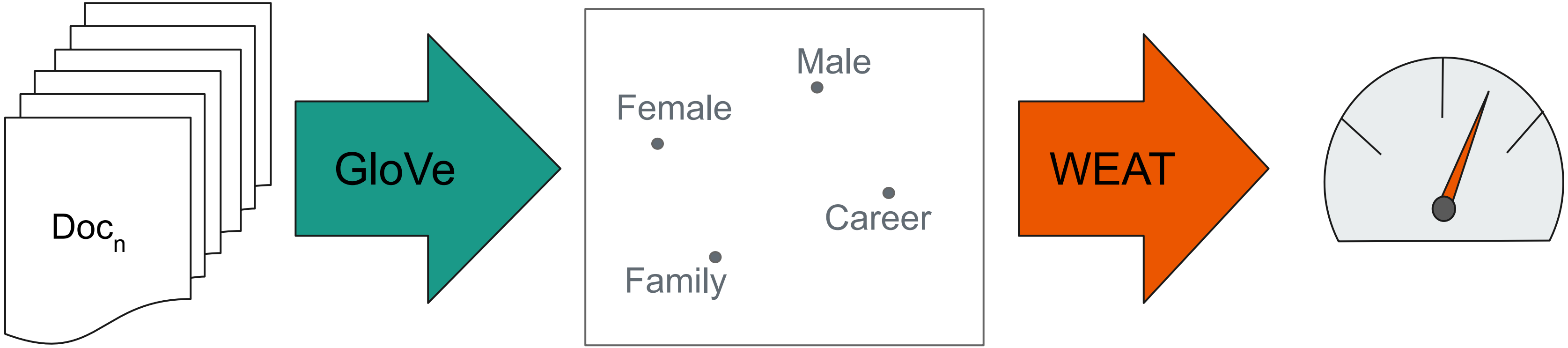
$$X = \sum_{i=1}^{n} X^{(i)} \qquad \tilde{X} = X - X^{(k)}$$

Analyse Metadata?

Document ID	ΔB	Year	Author
1	-0.0014		
2	0.0127		
...	...		
k	**0.0374**	?	?
...	...		
n	0.0089		

Bias Gradient

$$\nabla_X B(w(X)) = \nabla_w B(w) \nabla_X w(X)$$



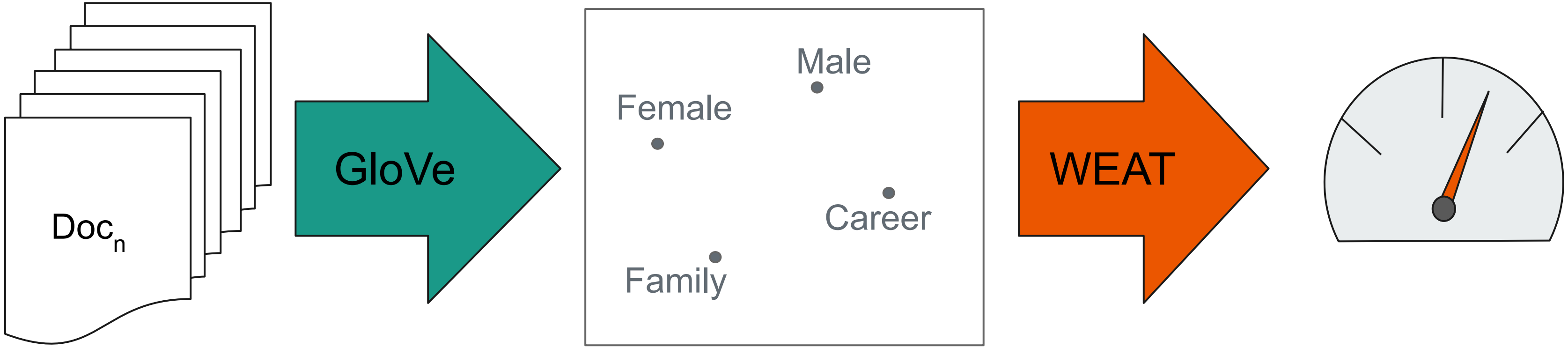
X : Corpus
(e.g. Wikipedia)

$\{ w_i \} = w(X)$
Word Embedding

$B(w(X))$
Bias Measured

Bias Gradient

$$\nabla_X B(w(X)) = \nabla_w B(w) \nabla_X w(X)$$



X : Corpus
(e.g. Wikipedia)

$\{ w_i \} = w(X)$
Word Embedding

$B(w(X))$
Bias Measured

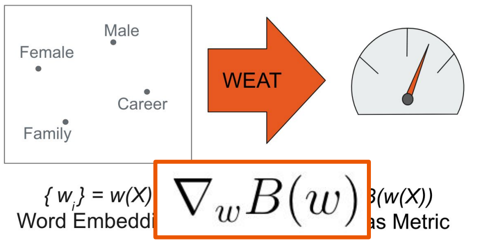
Background

Method Overview

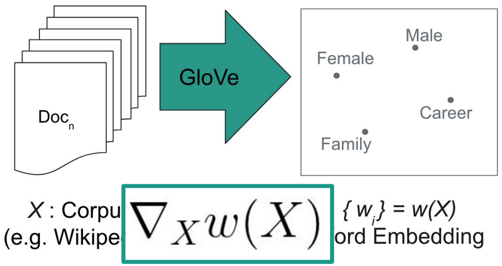
> **Critical Details**

Experiments

Computing the Components



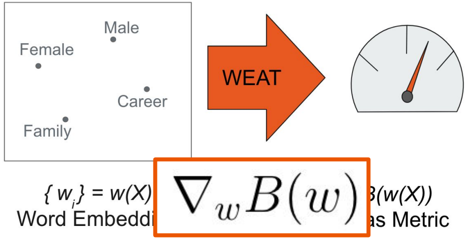
Fast & Easy: Math, Automatic Differentiation, or two evaluations of B(w).



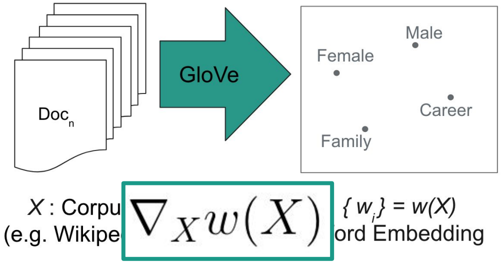
Slow & Hard: Differentiate through an entire training procedure:

- Leave-one-out retraining? (*time-bound*)
- Backprop? (*memory-bound*)
- Approximate using **Influence Functions**
  Koh & Liang (ICML 2017)

Computing the Components



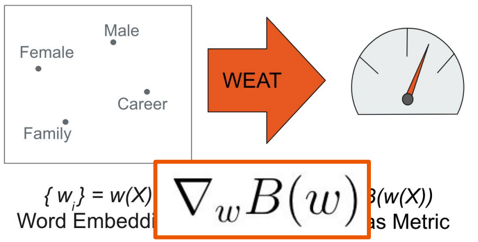
Fast & Easy: Math, Automatic Differentiation, or two evaluations of B(w).



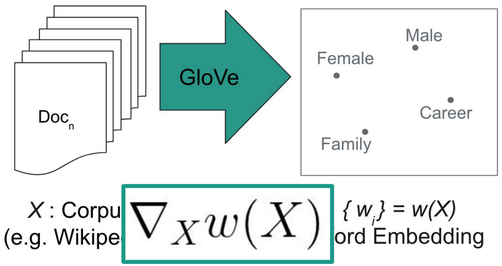
Slow & Hard: Differentiate through an entire training procedure:

- Leave-one-out retraining? (*time-bound*)
- Backprop? (*memory-bound*)
- Approximate using **Influence Functions**
  Koh & Liang (ICML 2017)

Computing the Components



Fast & Easy: Math, Automatic Differentiation, or two evaluations of B(w).



Slow & Hard: Differentiate through an entire training procedure:

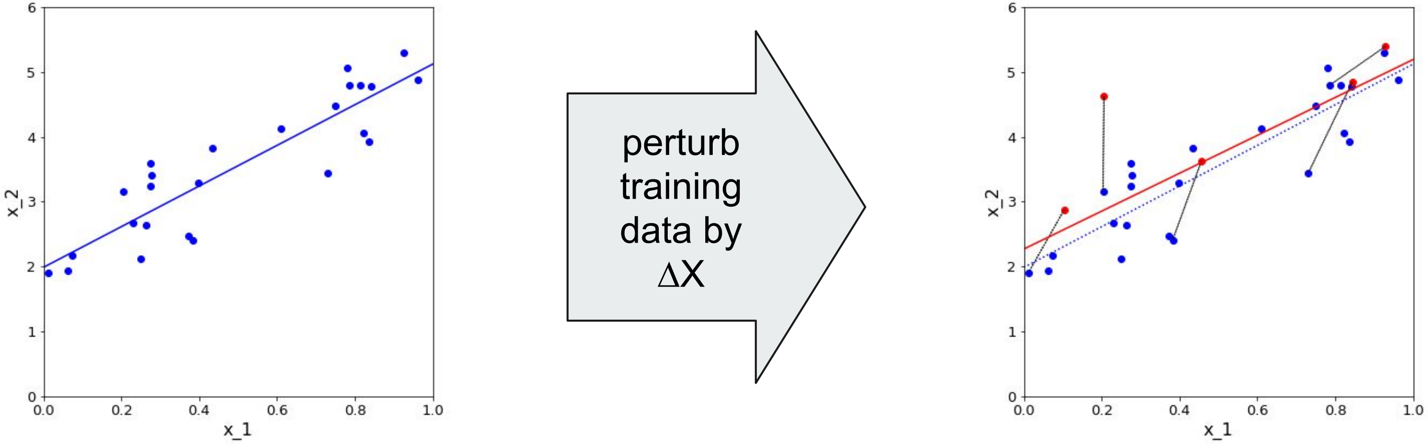
- Leave-one-out retraining? (*time-bound*)
- Backprop? (*memory-bound*)
- Approximate using **Influence Functions**
  Koh & Liang (ICML 2017)

Influence Functions

Give us a way to approximate the change in model parameters

model parameters: θ

new model params: $\tilde{\theta} \approx$ infl_func(θ, ΔX)



Influence Functions

$$\tilde{\theta} \approx \theta^* - \underbrace{\frac{1}{n} H_{\theta^*}^{-1}}_{} \sum_{k \in \delta} \left[\nabla_\theta L(\tilde{z}_k, \theta^*) - \nabla_\theta L(z_k, \theta^*) \right]$$

Inverse Hessian
(GloVe: **2VD** x **2VD** matrix)

$$V = |\text{vocab}| \qquad w_i \in \mathbb{R}^D$$

2VD can easily be > **10^9**

Applying Influence Functions to GloVe

GloVe Loss :

$$J(X, w, \underbrace{u, b, c}_{\text{other params (treat as const)}}) = \sum_{i=1}^{V}\sum_{j=1}^{V} f(X_{ij})(w_i^T u_j + b_i + c_j - \log X_{ij})^2$$

word vectors

Applying Influence Functions to GloVe

Gradient of Pointwise Loss

$$\nabla_w L(X_i, w) = \Big(\underbrace{\overbrace{0, \ldots, 0}^{D(i-1)}, \overbrace{\nabla_{w_i} L(X_i, w)}^{D}, \overbrace{0, \ldots, 0}^{D(V-i)}}_{VD\,\text{dimensions}}\Big)$$

Hessian becomes **block diagonal**!

(**V** Blocks of **D** by **D**)

Allows us to apply influence function approximation to **one word vector at a time!**

Algorithm: Compute Differential Bias

$w^*, u^*, b^*, c^* = \text{GloVe}(X)$ *# Train embedding*

for doc **in** corpus **do**

$\tilde{X} = X - X^{(k)}$ *# Subtract coocs from doc k*

for word i **in** doc $\cap$ WEAT words

 # Only need change in WEAT word vectors

$\tilde{w}_i = w_i^* - \frac{1}{V} H_{w_i}^{-1} \left[\nabla_{w_i} L(\tilde{X}_i, w) - \nabla_{w_i} L(X_i, w) \right]$

end for

$\Delta_{\text{doc}} B \approx B_{\text{weat}}(w^*) - B_{\text{weat}}(\tilde{w})$

end for

Algorithm: Compute Differential Bias

$w^*, u^*, b^*, c^* = \text{GloVe}(X)$ *# Train embedding*

for doc **in** corpus **do**

$\quad \tilde{X} = X - X^{(k)}$ *# Subtract coocs from doc k*

$\quad$ **for** word i **in** doc $\cap$ WEAT words

$\quad\quad$ *# Only need change in WEAT word vectors*

$\quad\quad \tilde{w}_i = w_i^* - \frac{1}{V} H_{w_i}^{-1} \left[\nabla_{w_i} L(\tilde{X}_i, w) - \nabla_{w_i} L(X_i, w) \right]$

$\quad$ **end for**

$\quad \Delta_{\text{doc}} B \approx B_{\text{weat}}(w^*) - B_{\text{weat}}(\tilde{w})$

end for

Algorithm: Compute Differential Bias

w^*, u^*, b^*, c^* = GloVe(X) *# Train embedding*

for doc **in** corpus **do**

$\tilde{X} = X - X^{(k)}$ *# Subtract coocs from doc k*

for word i **in** doc $\cap$ WEAT words

 # Only need change in WEAT word vectors

$$\tilde{w}_i = w_i^* - \frac{1}{V} H_{w_i}^{-1} \left[\nabla_{w_i} L(\tilde{X}_i, w) - \nabla_{w_i} L(X_i, w) \right]$$

end for

$$\Delta_{\text{doc}} B \approx B_{\text{weat}}(w^*) - B_{\text{weat}}(\tilde{w})$$

end for

Algorithm: Compute Differential Bias

$w^*, u^*, b^*, c^* = \text{GloVe}(X)$ *# Train embedding*

for doc **in** corpus **do**

$\tilde{X} = X - X^{(k)}$ *# Subtract coocs from doc k*

for word i **in** doc $\cap$ WEAT words

 # Only need change in WEAT word vectors

$$\tilde{w}_i = w_i^* - \frac{1}{V} H_{w_i}^{-1} \left[\nabla_{w_i} L(\tilde{X}_i, w) - \nabla_{w_i} L(X_i, w) \right]$$

end for

$\Delta_{\text{doc}} B \approx B_{\text{weat}}(w^*) - B_{\text{weat}}(\tilde{w})$

end for

Algorithm: Compute Differential Bias

$w^*, u^*, b^*, c^* = \text{GloVe}(X)$ *# Train embedding*

for doc **in** corpus **do**

$\tilde{X} = X - X^{(k)}$ *# Subtract coocs from doc k*

for word i **in** doc $\cap$ WEAT words

 # Only need change in WEAT word vectors

$\tilde{w}_i = w_i^* - \frac{1}{V} H_{w_i}^{-1} [\nabla_{w_i} L(\tilde{X}_i, w) - \nabla_{w_i} L(X_i, w)]$

end for

$\Delta_{\text{doc}} B \approx B_{\text{weat}}(w^*) - B_{\text{weat}}(\tilde{w})$

end for

Background
Method Overview
Critical Details
> Experiments

Objectives of Experiments

1. Assess the accuracy of our influence function approximation
2. Identify and analyse most bias impacting documents

WEAT

S = Science	T = Arts
A = Male	B = Female

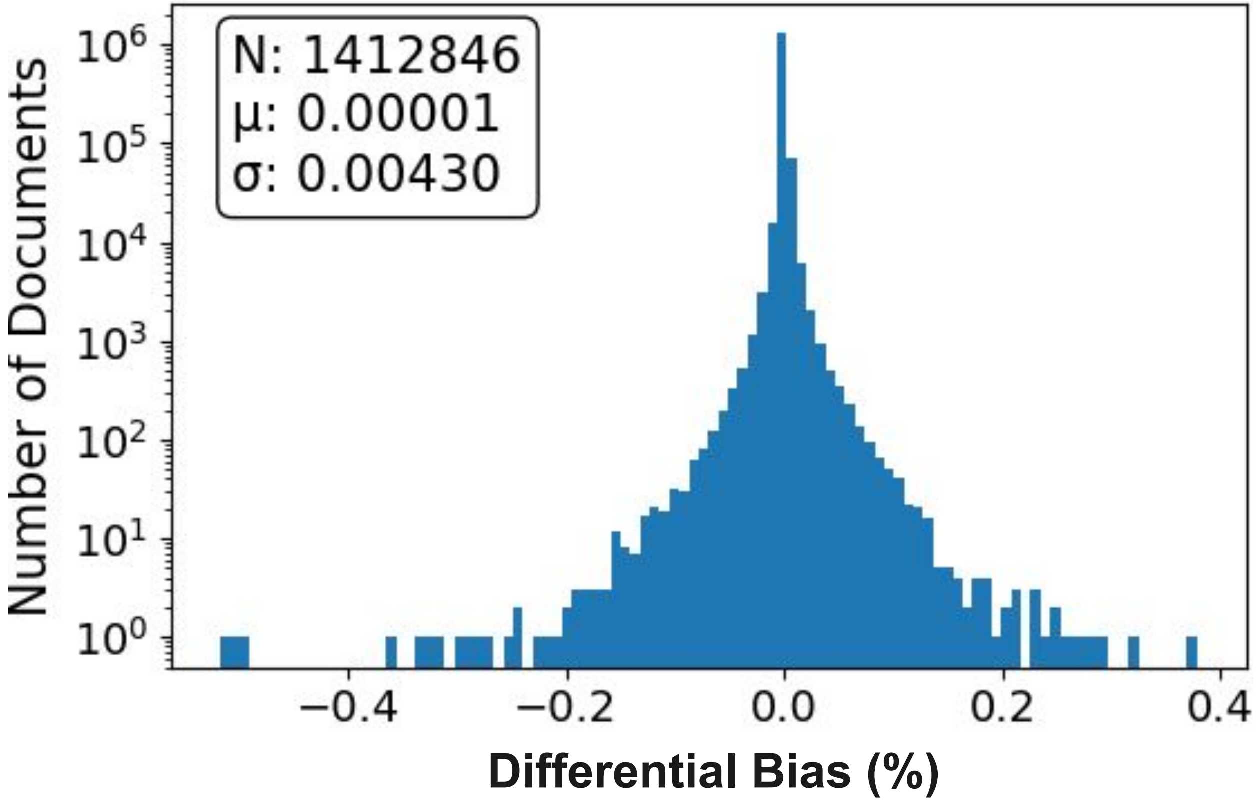
S = Instruments	T = Weapons
A = Pleasant	B = Unpleasant

Corpora

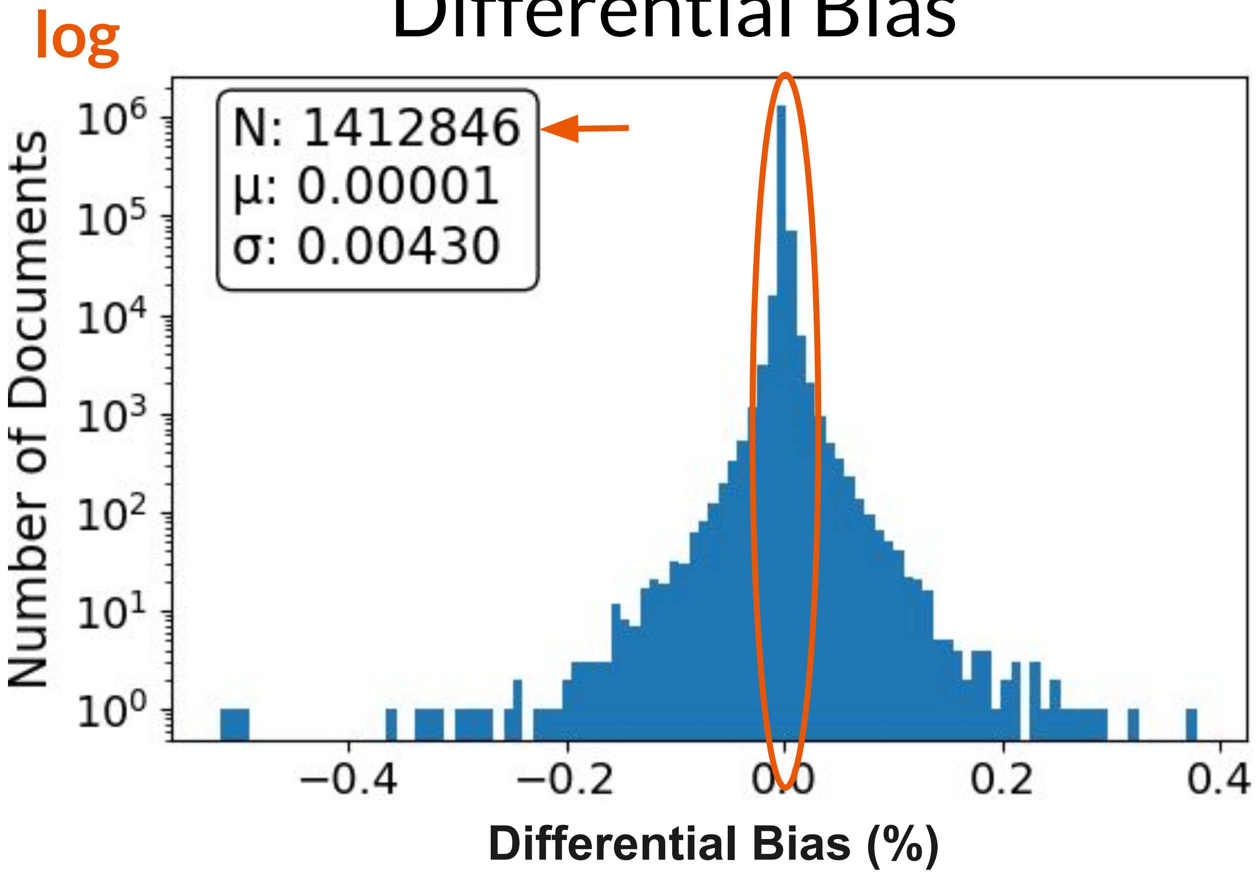
WIKIPEDIA
The Free Encyclopedia

The New York Times

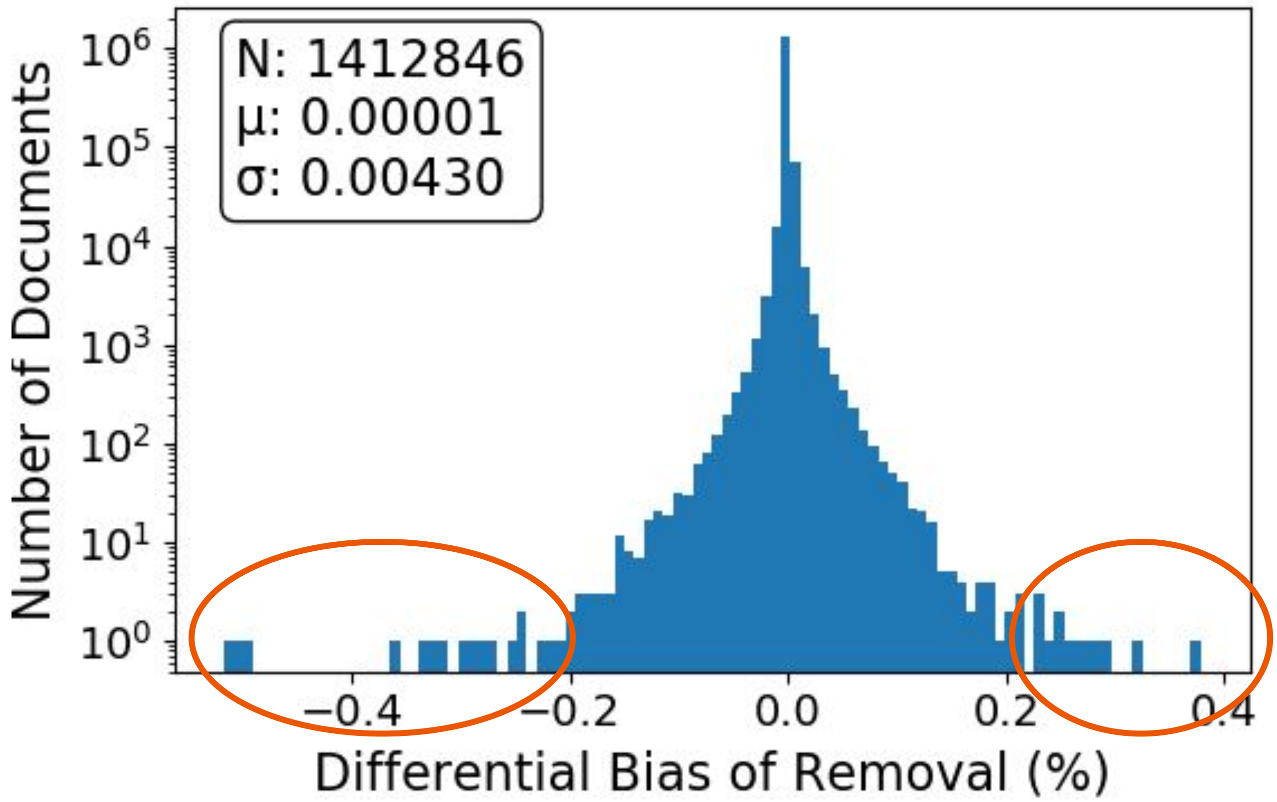
Differential Bias



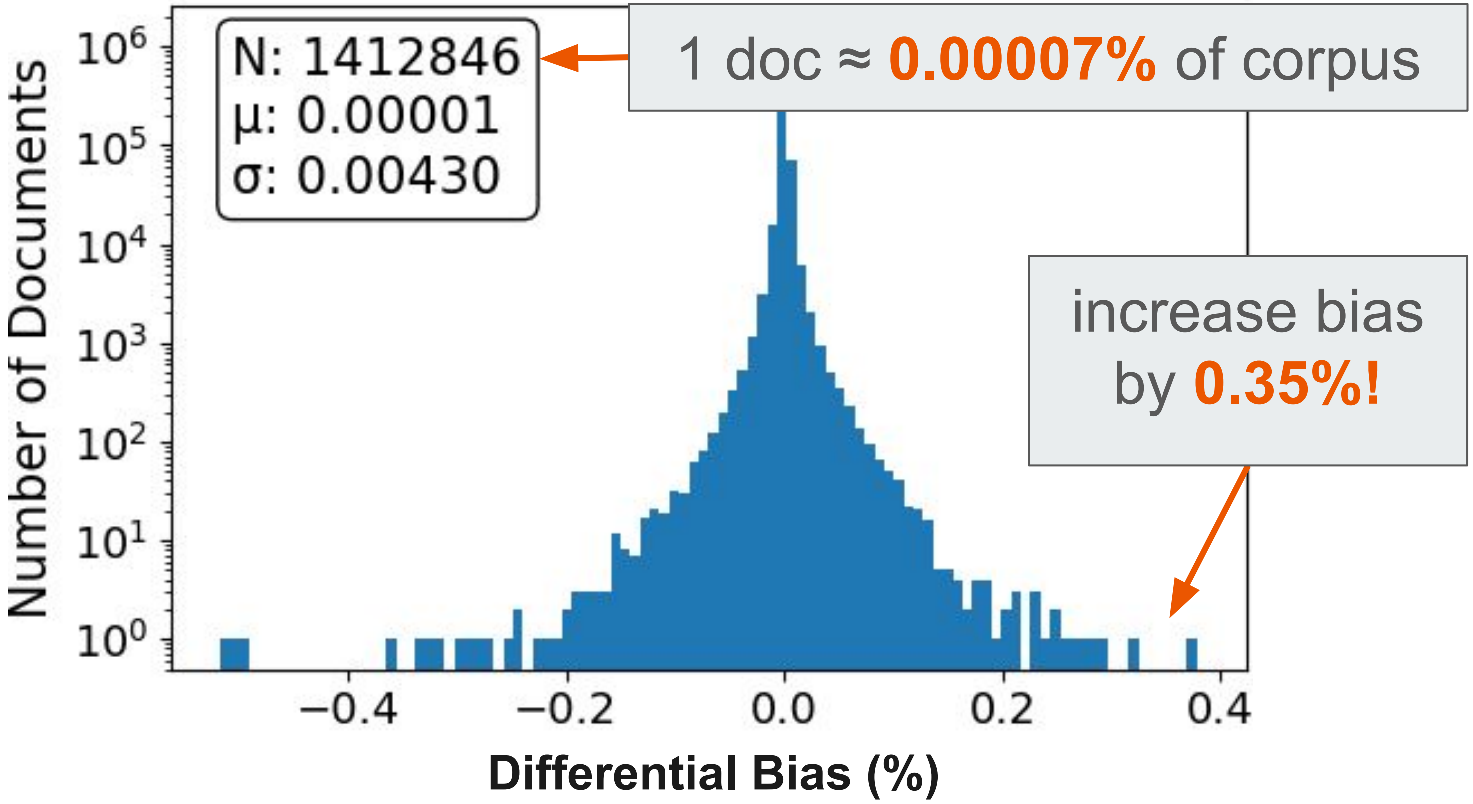
Differential Bias

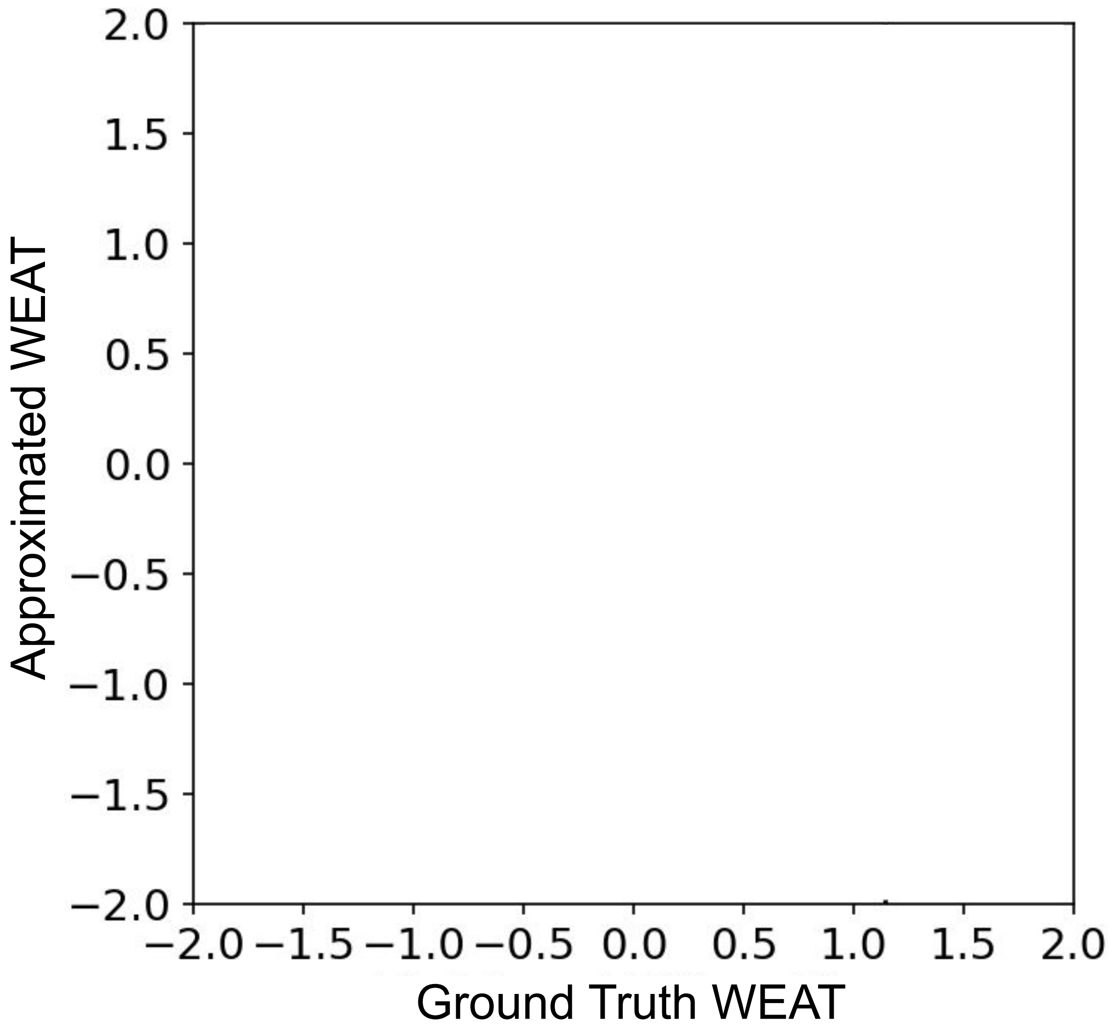


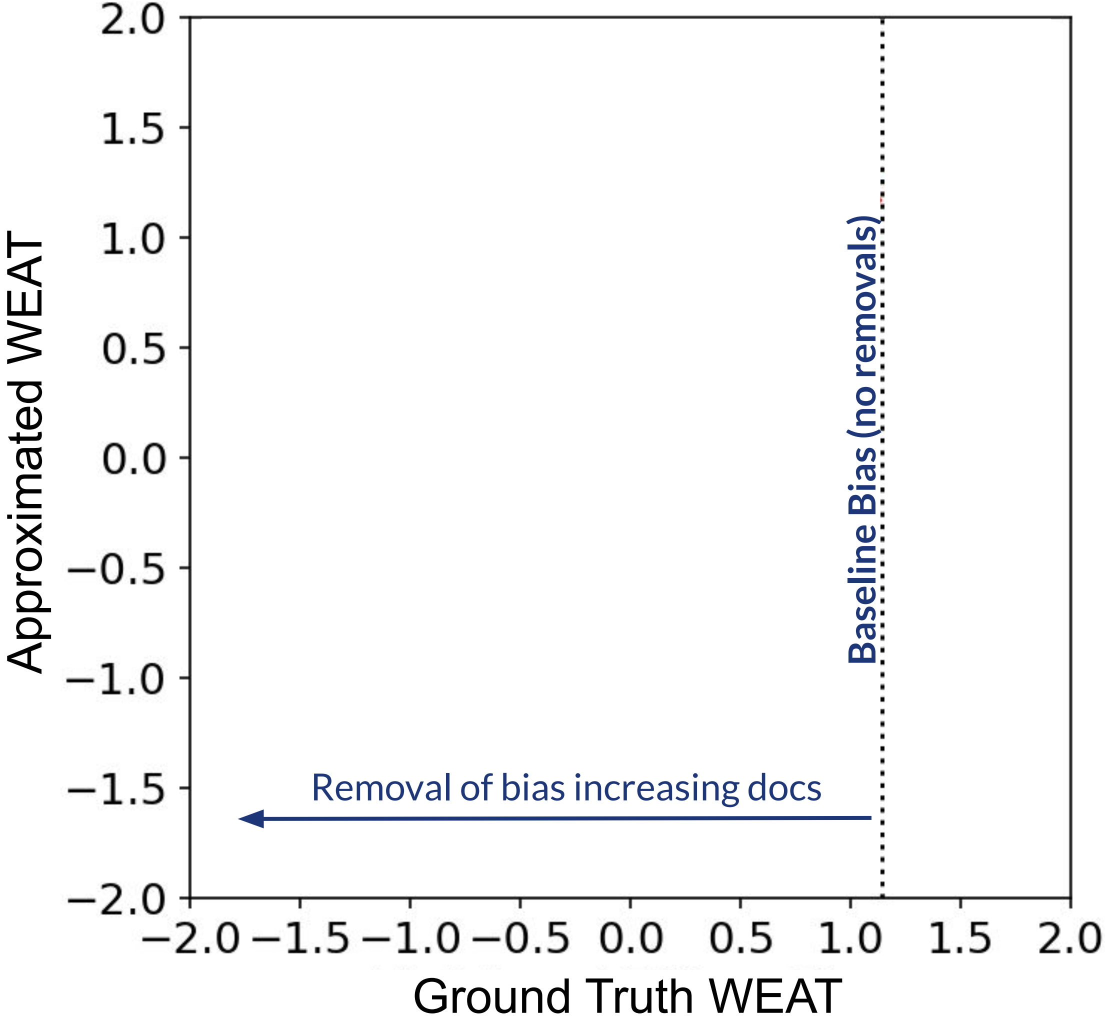
Differential Bias

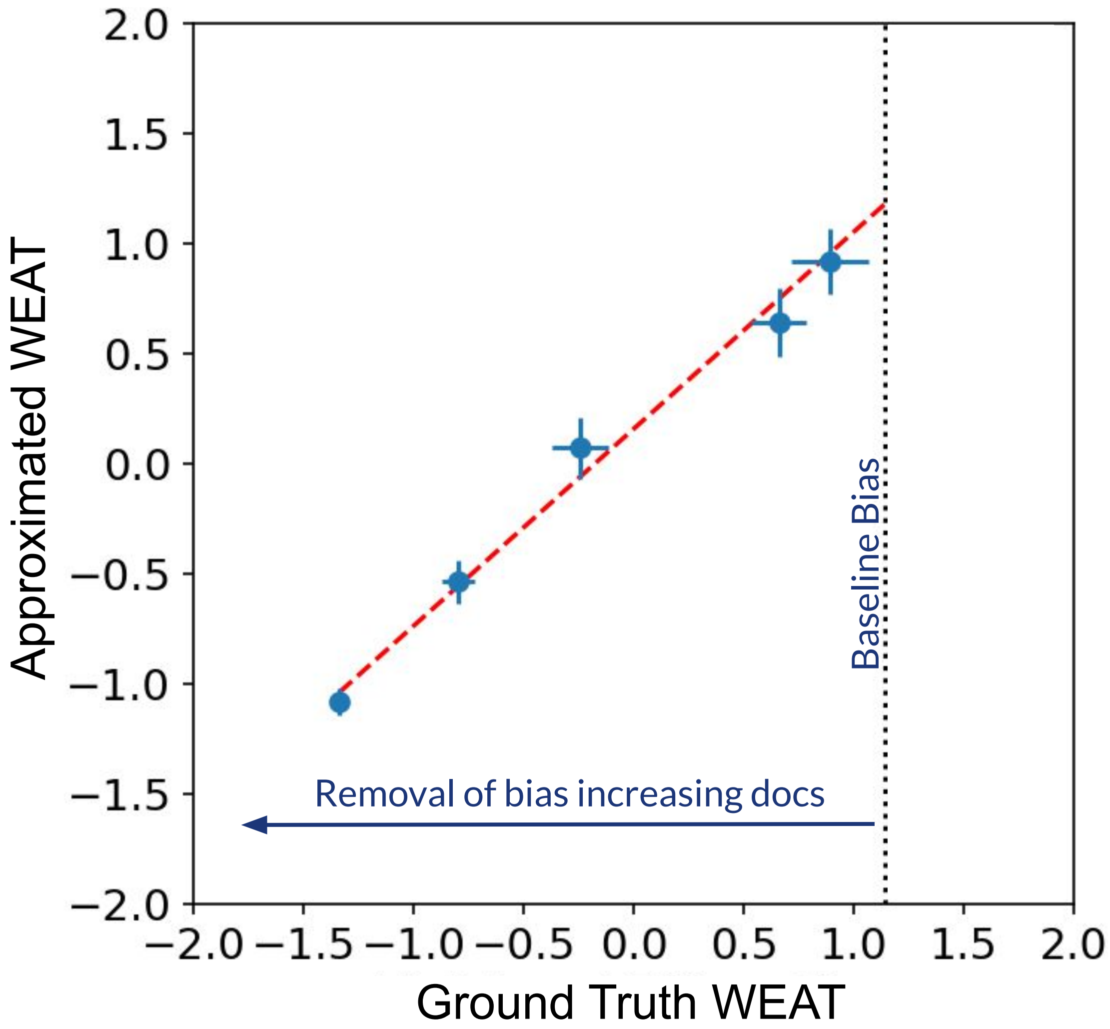


Differential Bias

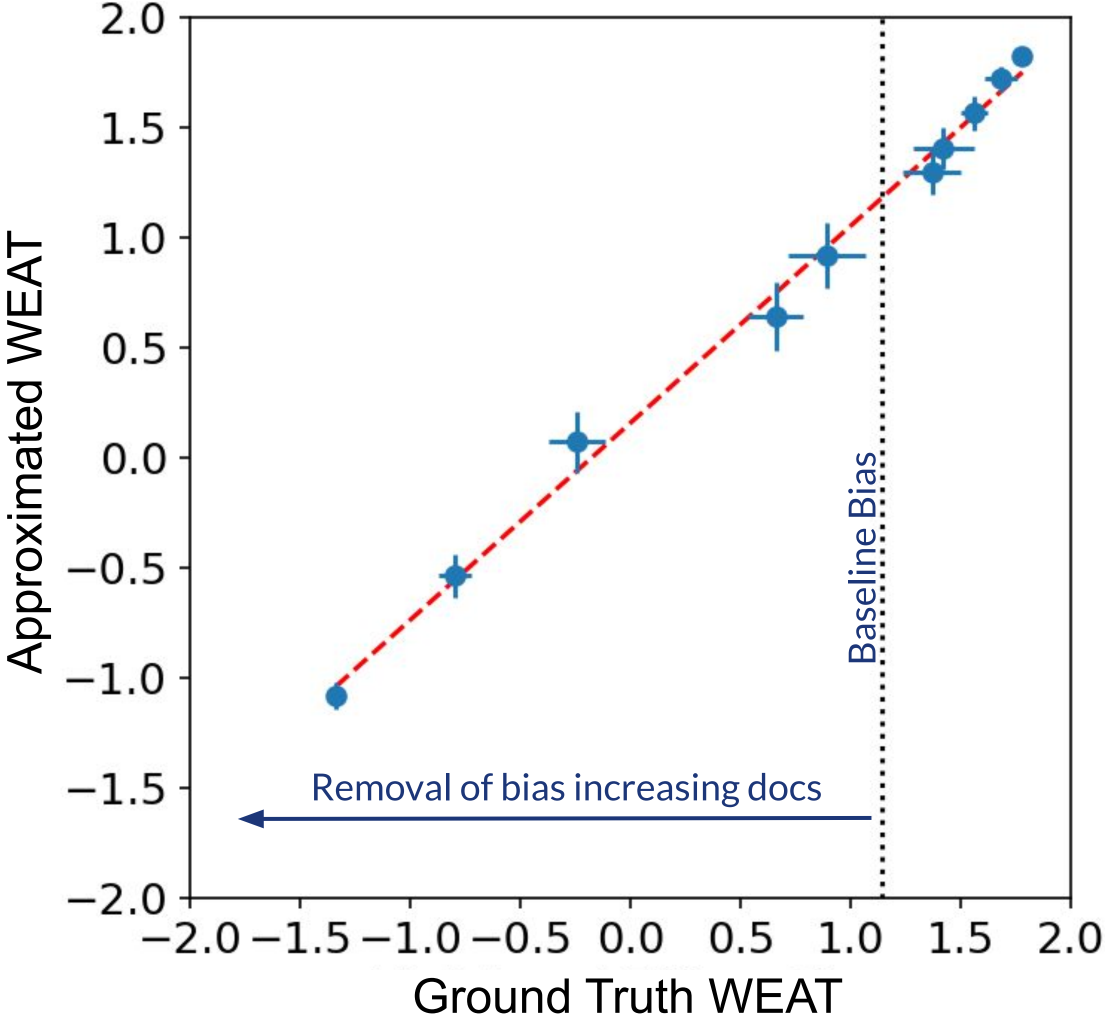


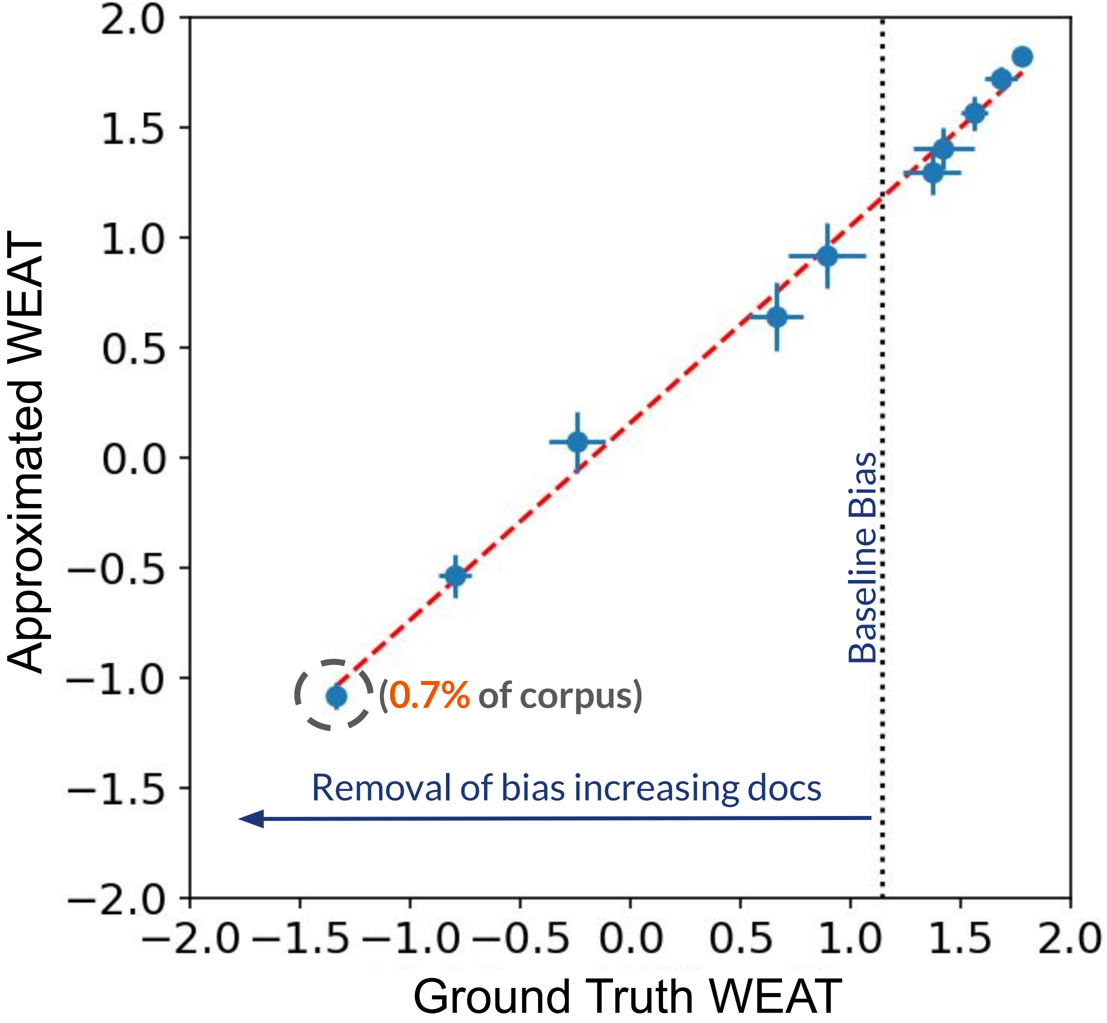

2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
Approximated WEAT
−2.0 −1.5 −1.0 −0.5 0.0 0.5 1.0 1.5 2.0
Ground Truth WEAT


Approximated WEAT
Ground Truth WEAT
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
−2.0 −1.5 −1.0 −0.5 0.0 0.5 1.0 1.5 2.0
Baseline Bias (no removals)
Removal of bias increasing docs



Approximated WEAT
Ground Truth WEAT
Baseline Bias
Removal of bias increasing docs
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
−2.0 −1.5 −1.0 −0.5 0.0 0.5 1.0 1.5 2.0


Approximated WEAT
Ground Truth WEAT
Baseline Bias
Removal of bias increasing docs
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
−2.0 −1.5 −1.0 −0.5 0.0 0.5 1.0 1.5 2.0



2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
Approximated WEAT
Baseline Bias
(0.7% of corpus)
Removal of bias increasing docs
−2.0 −1.5 −1.0 −0.5 0.0 0.5 1.0 1.5 2.0
Ground Truth WEAT

$\Delta_d B$	**Bias Decreasing**
-0.52	Hormone Therapy Study Finds Risk for Some
-0.50	For Women in Astronomy, a Glass Ceiling in the Sky
-0.49	Sorting Through the Confusion Over Estrogen
-0.36	Young Astronomers Scan Night Sky and Help Wanted Ads

$\Delta_d B$	**Bias Increasing**
0.38	Kaj Aage Strand, 93, Astronomer At the U.S. Naval Observatory
0.32	Gunman in Iowa Wrote of Plans In Five Letters
0.29	ENGINEER WARNED ABOUT DIRE IMPACT OF LIFTOFF DAMAGE
0.29	Fred Gillett, 64; Studied Infrared Astronomy
0.27	Robert Harrington, 50, Astronomer in Capital

$\Delta_d B$	**Bias Decreasing**
-0.52	Hormone Therapy Study Finds Risk for Some
-0.50	For Women in Astronomy, a Glass Ceiling in the Sky
-0.49	Sorting Through the Confusion Over Estrogen
-0.36	Young Astronomers Scan Night Sky and Help Wanted Ads

$\Delta_d B$	**Bias Increasing**
0.38	Kaj Aage Strand, 93, Astronomer At the U.S. Naval Observatory
0.32	Gunman in Iowa Wrote of Plans In Five Letters
0.29	ENGINEER WARNED ABOUT DIRE IMPACT OF LIFTOFF DAMAGE
0.29	Fred Gillett, 64; Studied Infrared Astronomy
0.27	Robert Harrington, 50, Astronomer in Capital

$\Delta_d B$	**Bias Decreasing**
-0.52	Hormone Therapy Study Finds Risk for Some
-0.50	For Women in Astronomy, a Glass Ceiling in the Sky
-0.49	Sorting Through the Confusion Over Estrogen
-0.36	Young Astronomers Scan Night Sky and Help Wanted Ads

$\Delta_d B$	**Bias Increasing**
0.38	Kaj Aage Strand, 93, Astronomer At the U.S. Naval Observatory
0.32	Gunman in Iowa Wrote of Plans In Five Letters
0.29	ENGINEER WARNED ABOUT DIRE IMPACT OF LIFTOFF DAMAGE
0.29	Fred Gillett, 64; Studied Infrared Astronomy
0.27	Robert Harrington, 50, Astronomer in Capital

Document Impact Generalizes

WEAT_1 (Science v.s. Arts Gender Bias)

	remove bias **increasing** docs	baseline (no removals)	remove bias **decreasing** docs
GloVe	-1.27	1.14	1.7
word2vec	**0.11**	1.35	**1.6**

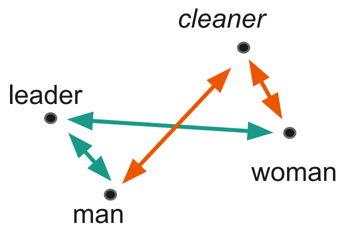
Removal of documents also **affects word2vec**, and other metrics!

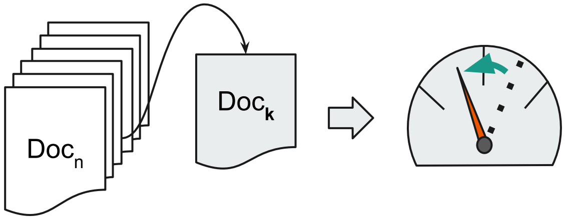
Limitations & Future Work

- Consider **multiple biases** at simultaneously
- Use metrics that depend on **more words**
- Consider bias in **downstream tasks** where embeddings are used
- Does this carry over to **BERT**?

Recap

- Bias can be quantified; correlates with known human biases
- We can identify the documents that most impact bias, and approximate impact
- These documents are qualitatively meaningful, and impact generalizes





$\Delta_d B$	**Bias Decreasing**
-0.52	Hormone Therapy Study Finds Risk for Some
-0.50	For Women in Astronomy, a Glass Ceiling in the Sky
-0.49	Sorting Through the Confusion Over Estrogen
-0.36	Young Astronomers Scan Night Sky and Help Wanted Ads

Thank you!

Poster # 146

mebrunet@cs.toronto.edu

arXiv: 1810.03611

Marc

Colleen

Ashton

Rich

References

- T. Bolukbasi, K.-W. Chang, J. Zou, V. Saligrama, and A. Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In 30th Conference on Neural Information Processing Systems (NIPS), 2016.
- A. Caliskan, J. J. Bryson, and A. Narayanan. Semantics derived automatically from language corpora contain human-like biases. Science, 356(6334):183–186, 2017.
- P. W. Koh and P. Liang. Understanding Black-box Predictions via Influence Functions. In Proceedings of the 34th International Conference on Machine Learning, volume 70 of Proceedings of Machine Learning Research, pages 1885–1894, 2017.

Measuring Bias

"...results raise the possibility that **all** implicit human **biases** are **reflected in** the statistical properties of **language**."

Aylin Caliskan, Joanna J. Bryson, Arvind Narayanan (Science 2017)

Impact on Word2Vec

	Removal of Documents Identified by our Method		
	Decrease (0.7%)	Baseline	Increase (0.7%)
GloVe	-1.27	1.14	1.7
word2vec	**0.11**	1.35	1.6

Word Embeddings

Compact vector representation
(like a dictionary for machines)

Learned from **LARGE** corpora.

Used in many NLP tasks:

- Sentiment Analysis
- Text summarization
- Machine Translation

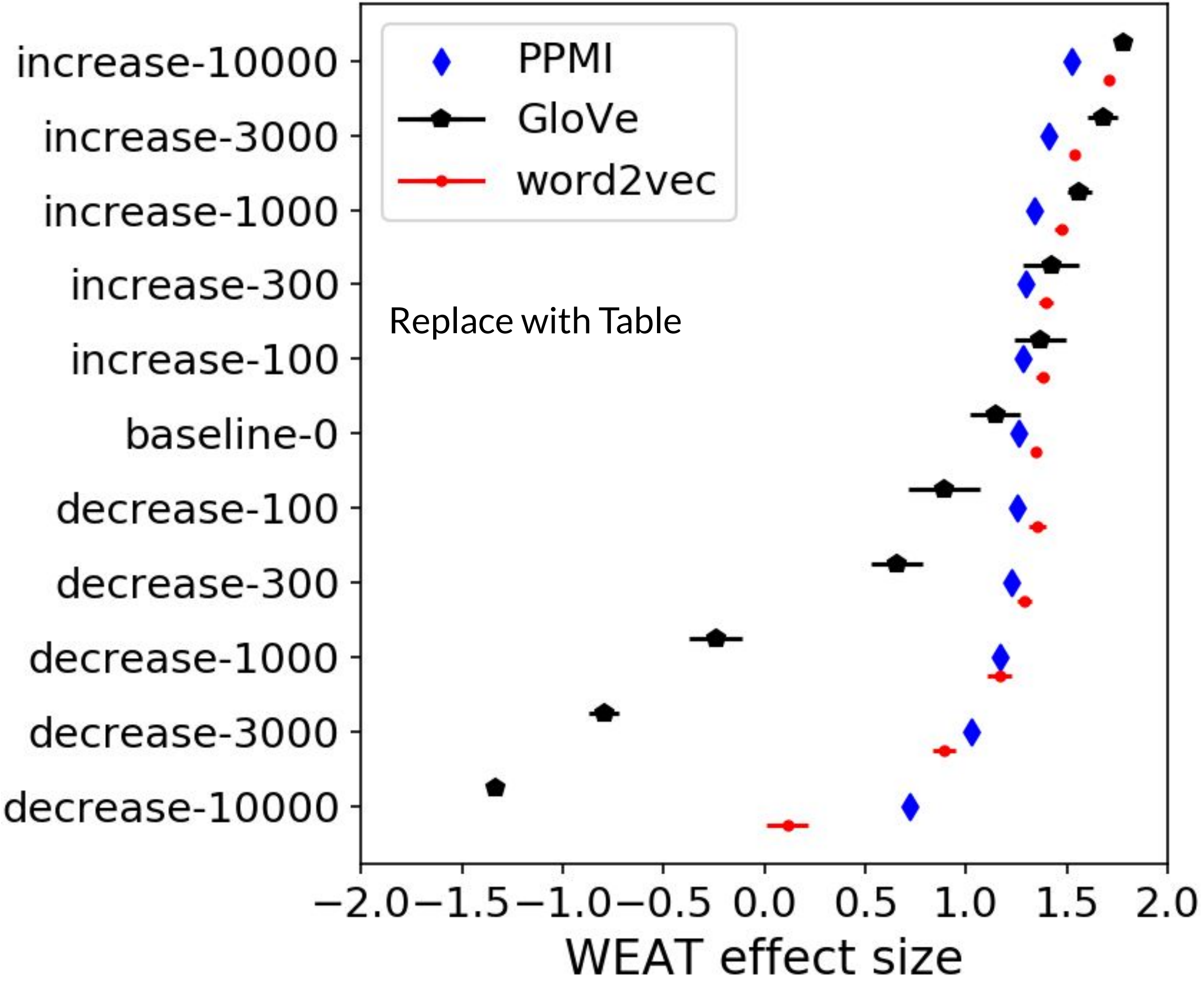
dic·tion·ar·y
/ˈdikSHəˌnerē/

noun

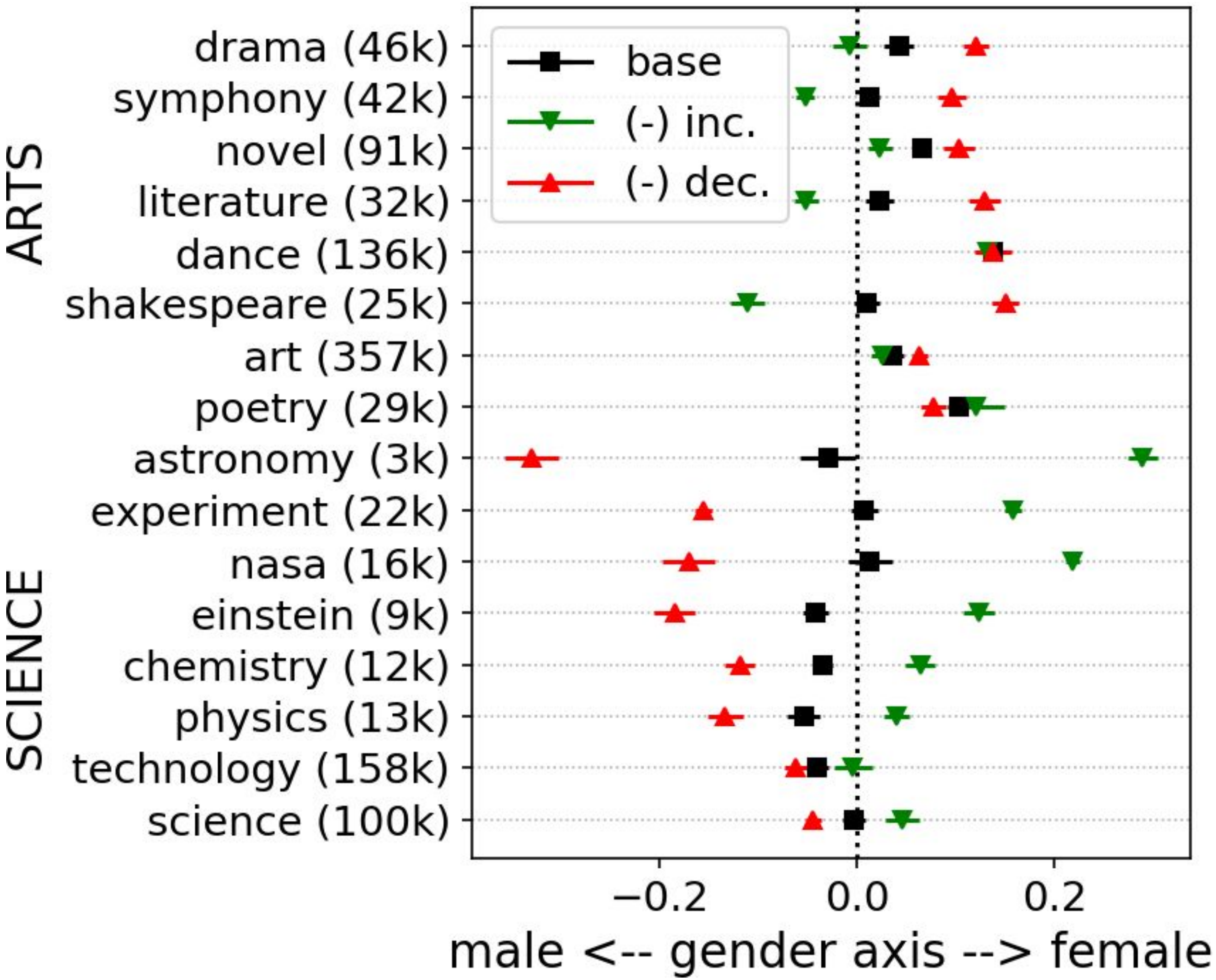
a book or electronic resource that lists the words of a language (typically in alphabetical order) and gives their meaning, or gives the equivalent words in a different language, often also providing information about pronunciation, origin, and usage.
"I'll look up 'love' in the dictionary"
synonyms: lexicon, wordbook, glossary, vocabulary list, vocabulary, word list, wordfinder
"half of the words in his text were not in the dictionary"

```
{
    "dictionally": [1.33,  -0.48,  0.98,  -2.33 ... ],

    "dictionary": [1.23,  -0.52,  1.01,  -2.14 ... ],

    "dictions": [1.04,  -0.63,  0.87,  -2.23 ... ],
    ...
}
```

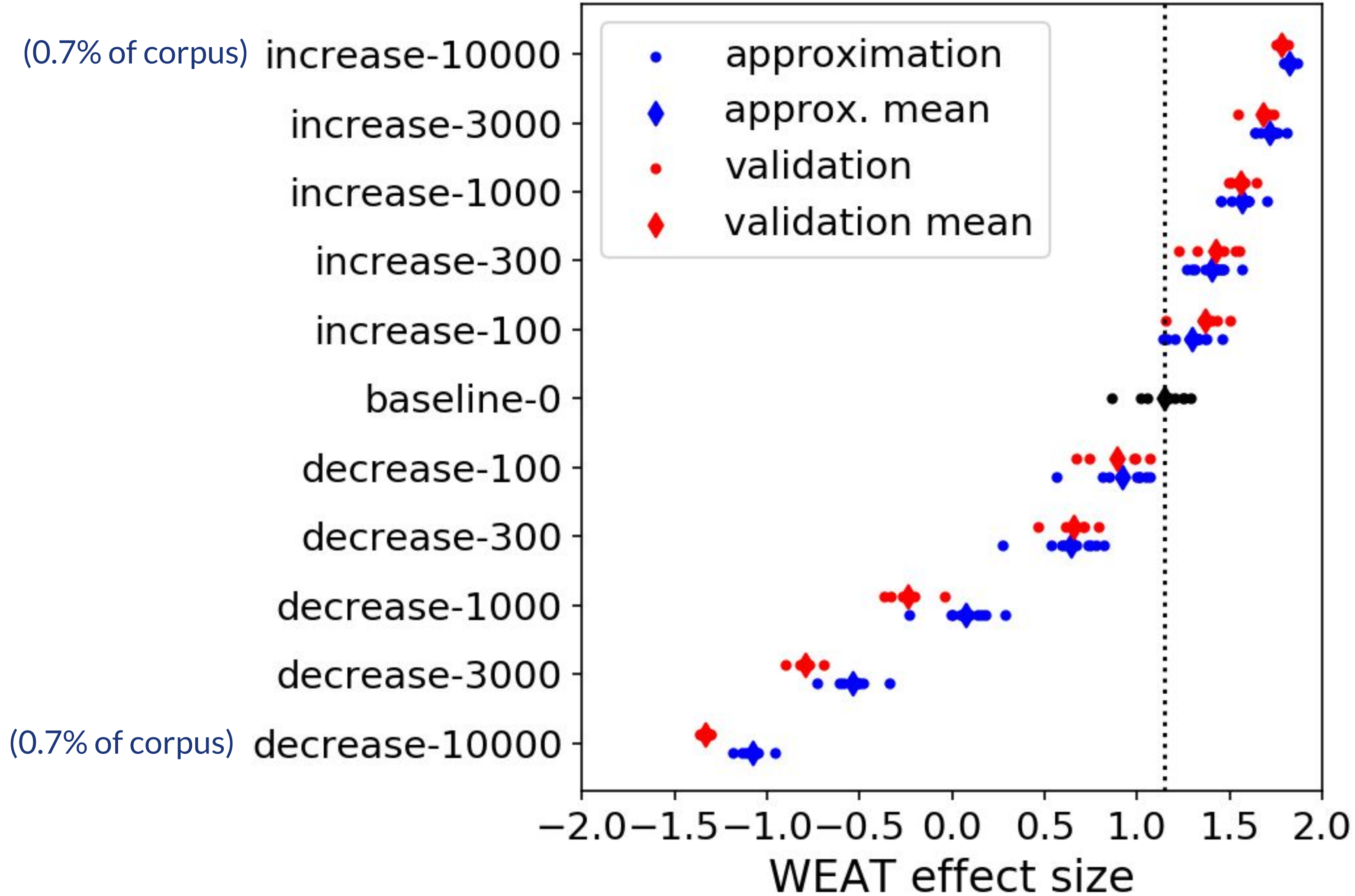
(0.7% of corpus)



(0.7% of corpus)



base
(-) inc.
(-) dec.
ARTS
SCIENCE
drama (46k)
symphony (42k)
novel (91k)
literature (32k)
dance (136k)
shakespeare (25k)
art (357k)
poetry (29k)
astronomy (3k)
experiment (22k)
nasa (16k)
einstein (9k)
chemistry (12k)
physics (13k)
technology (158k)
science (100k)
−0.2
0.0
0.2
male <-- gender axis --> female



(0.7% of corpus)
increase-10000
increase-3000
increase-1000
increase-300
increase-100
baseline-0
decrease-100
decrease-300
decrease-1000
decrease-3000
(0.7% of corpus)
decrease-10000
approximation
approx. mean
validation
validation mean
−2.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
WEAT effect size

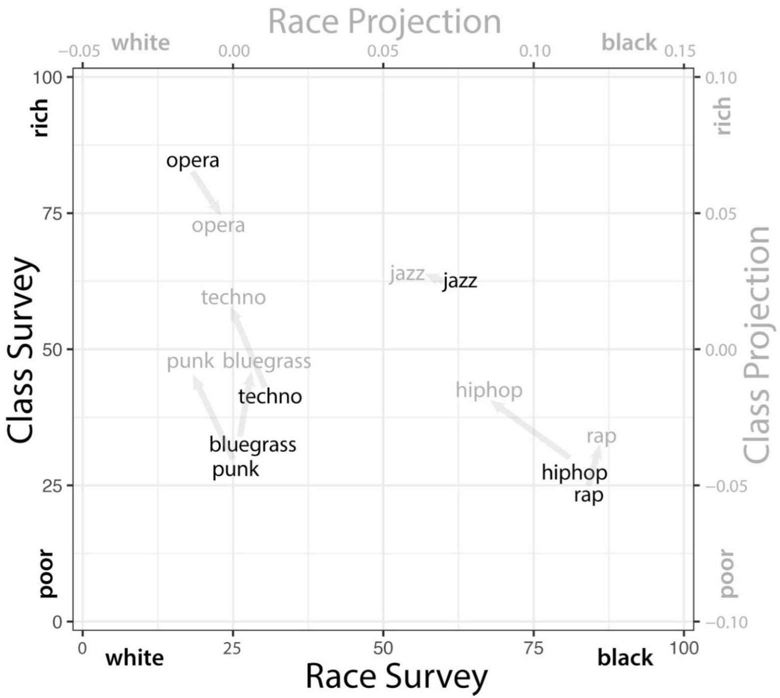
S	science	science, technology, physics, chemistry, einstein, nasa, experiment, astronomy
T	arts	poetry, art, shakespeare, dance, literature, novel, symphony, drama
A	male	male, man, boy, brother, he, him, his, son
B	female	female, woman, girl, sister, she, her, hers, daughter

Psychology, Bias, and Embeddings

One study examined a dozen well- known human biases: all present

Others examined the geometry of

- Class
- Race
- Gender



Austin C. Kozlowski, Matt Taddy, James A. Evans (2018)

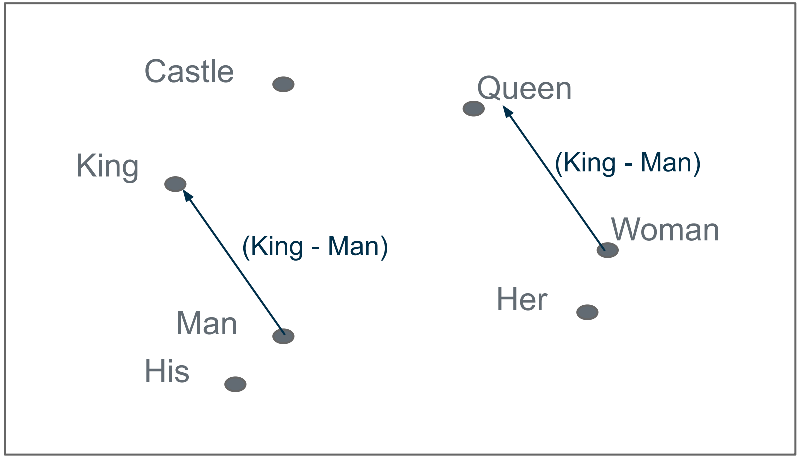
Word Embeddings

What are they?

- A compact vector representation for words
- Learned from a very large corpus of text
- Preserves syntactic and semantic meaning through vector arithmetic (**very useful**)

Applications:

- Sentiment analysis
- Document classification / summarization
- Translation
- Temporal semantic trajectories

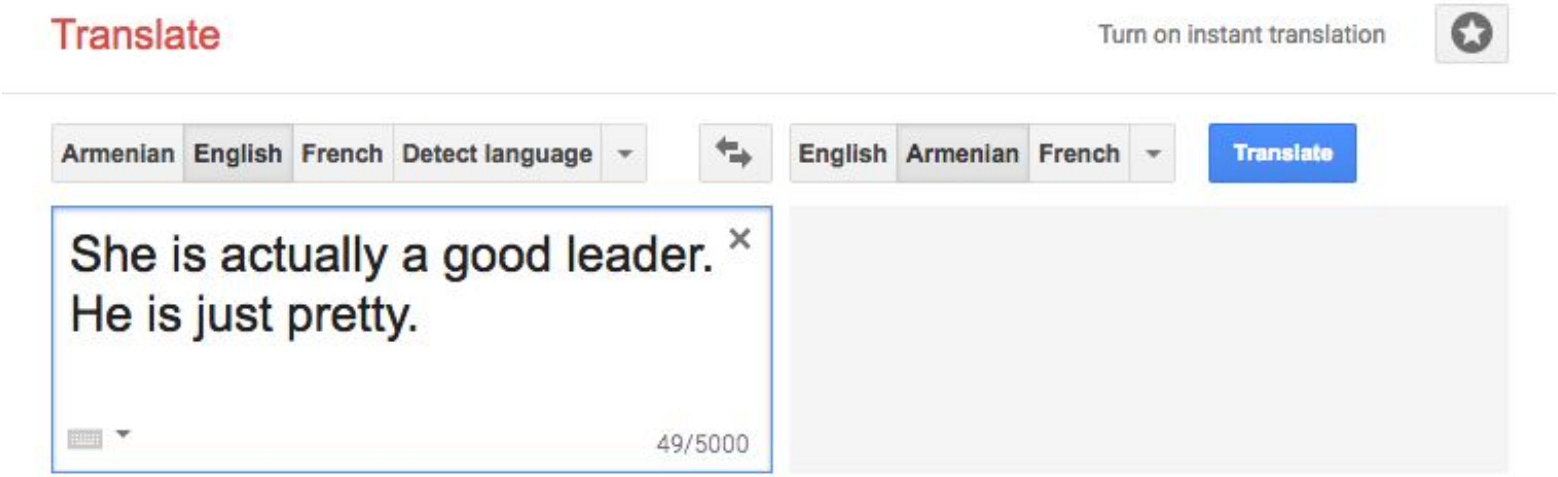


"King" - "Man" + "Woman" ≈ "Queen"

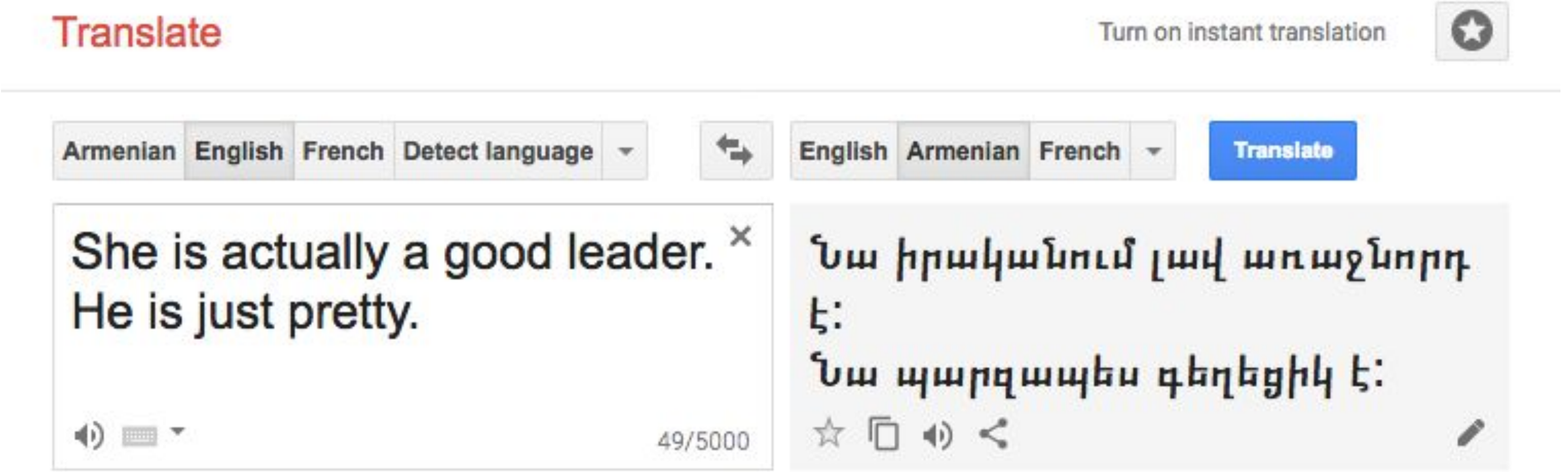
A Motivating Example

"She is actually a good leader. He is just pretty." #NoPlanetB

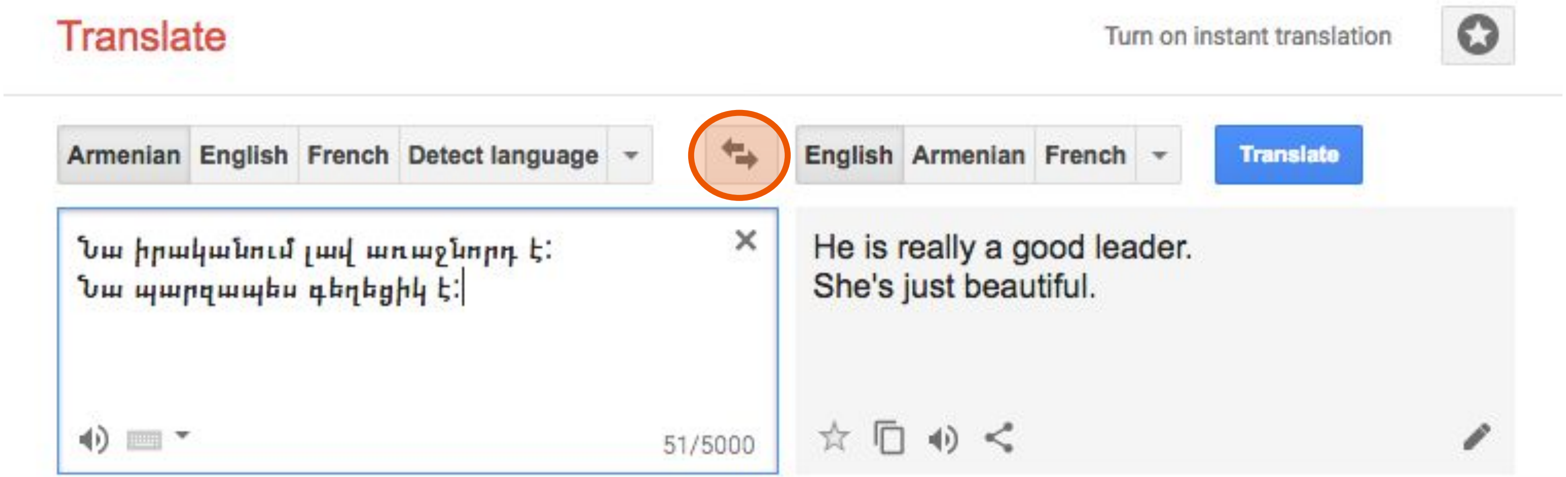
Presumptuous Translation



Presumptuous Translation

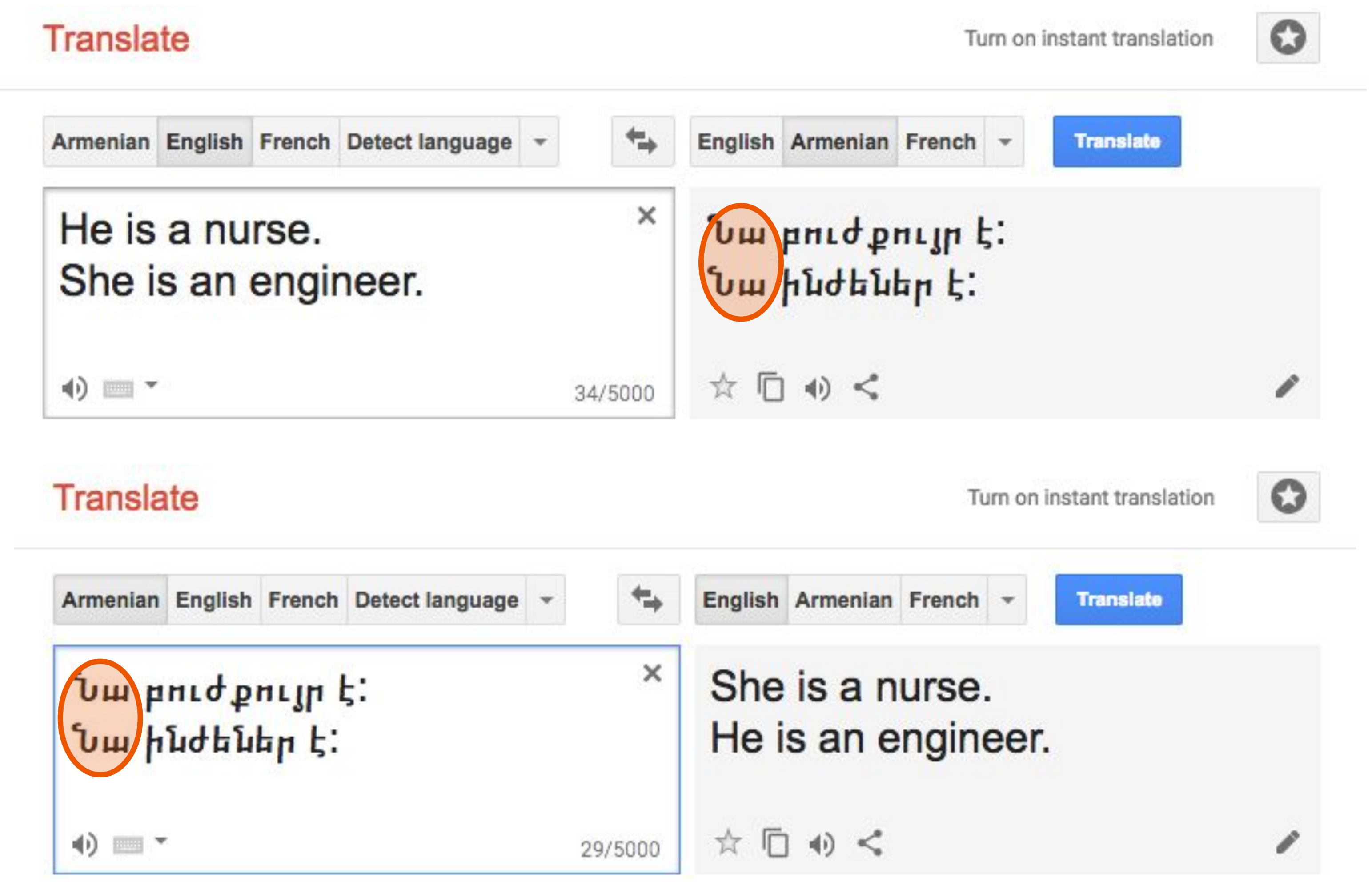


Presumptuous Translation




Translate
Turn on instant translation
Armenian
English
French
Detect language
English
Armenian
French
Translate
He is a nurse.
She is an engineer.
34/5000
Նա բուժքույր է:
Նա ինժեներ է:
Translate
Turn on instant translation
Armenian
English
French
Detect language
English
Armenian
French
Translate
Նա բուժքույր է:
Նա ինժեներ է:
29/5000
She is a nurse.
He is an engineer.

Why does this happen?


Translate
Turn on instant translation
Armenian English French Detect language
English Armenian French
Translate
He is a nurse.
She is an engineer.
34/5000
Նա բուժքույր է:
Նա ինժեներ է:
Translate
Turn on instant translation
Armenian English French Detect language
English Armenian French
Translate
Նա բուժքույր է:
Նա ինժեներ է:
29/5000
She is a nurse.
He is an engineer.

Word Co-Occurrences

	engineer	**nurse**	**leader**	**pretty**	***(all)***
Ratio of **he:she** co-occurrences	6.25	0.550	9.25	3.07	3.53

The New York Times Annotated Corpus (1987-2007, approx. 1B words, context window: 8)

GloVe: Global Vectors for Word Representations

$$J(X, w, u, b, c) = \sum_{i=1}^{V} \sum_{j=1}^{V} f(X_{ij})(w_i^T u_j + b_i + c_j - \log X_{ij})^2$$

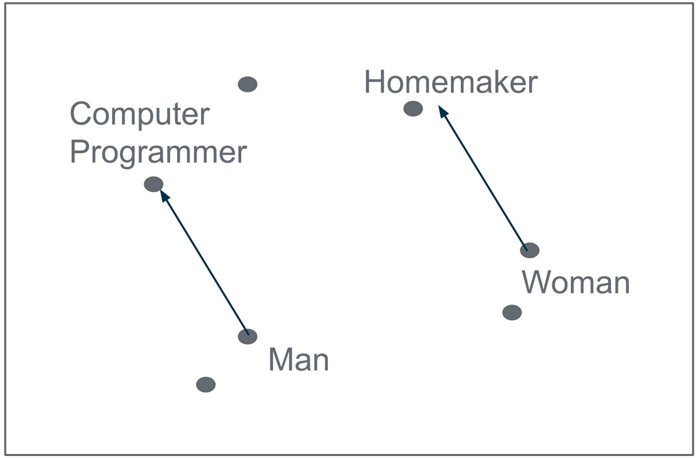
X : co-occurrence Matrix
{ w_i } : set of word vectors
{ u_j }, b, c : other model parameters

$$f(x) = \begin{cases} (x/x_{max})^\alpha \text{ if } x < x_{max} \\ 1 \text{ otherwise} \end{cases}$$

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014.

Bad Analogies

- 🙂 King : Man :: Queen : Woman
- 🙂 Paris : France :: London : England
- 😦 Man : Computer_Programmer :: Woman : Homemaker



Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, Adam Kalai (NeurIPS 2016)

WEAT

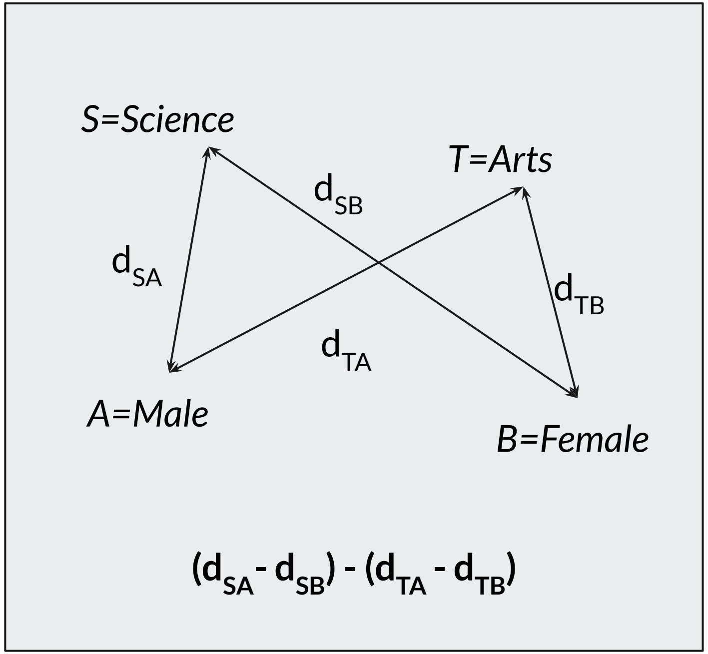
Target Word Sets:
S = {physics, chemistry… } ≈ *Science*
T = {poetry, litterature… } ≈ *Arts*

Attribute Word Sets:
A = {he, him, man… } ≈ *Male*
B = {she, her, woman} ≈ *Female*

Measures relative association between four concepts

$$f(w, A, B) = \underset{a \in A}{\text{mean}} cos(\vec{w}, \vec{a}) - \underset{b \in B}{\text{mean}} cos(\vec{w}, \vec{b})$$

Effect Size = $$\frac{\underset{s \in S}{\text{mean}} f(s, A, B) - \underset{t \in T}{\text{mean}} f(t, A, B)}{\underset{w \in S \cup T}{\text{std-dev}} f(w, A, B)}$$



Applying IF to GloVe

GloVe Loss :

$$J(X, w, u, b, c) = \sum_{i=1}^{V} \sum_{j=1}^{V} f(X_{ij})(w_i^T u_j + b_i + c_j - \log X_{ij})^2$$

Our "datapoints" are NOT documents, but rather the entries of X.
So one document removal: $\tilde{X} = X - X^{(k)}$, perturbs multiple "datapoints".

IF Approx : $$\tilde{\theta} \approx \theta^* - \frac{1}{n} H_{\theta^*}^{-1} \sum_{k \in \delta} [\nabla_\theta L(\tilde{z}_k, \theta^*) - \nabla_\theta L(z_k, \theta^*)]$$

Applying IF to GloVe

$$\tilde{w}_i \approx w_i^* - \frac{1}{V} \underbrace{H_{w_i}^{-1}}_{\text{Computed once per WEAT word}} \Big[\overbrace{\nabla_{w_i} L(\tilde{X}_i, w^*)}^{\text{Computed for every perturbation of interest}} - \underbrace{\nabla_{w_i} L(X_i, w^*)}_{\text{Computed once per WEAT word}} \Big]$$

Notice that for all i where $\tilde{X}_i = X_i$, $\tilde{w}_i = w_i^*$

Influence Functions (IF)

$$R(z,\theta) = \frac{1}{n}\sum_{i=1}^{n} L(z_i,\theta) \qquad \theta^* = \underset{\theta}{\operatorname{argmin}}\, R(z,\theta)$$

$$\tilde{\theta} \approx \theta^* - \underbrace{\frac{1}{n} H_{\theta^*}^{-1}}_{\text{Inverse Hessian}} \sum_{k\in\delta} \underbrace{[\nabla_\theta L(\overbrace{\tilde{z}_k}^{\text{Perturbed}},\theta^*) - \nabla_\theta L(\overbrace{z_k}^{\text{Original}},\theta^*)]}_{\text{Difference of Gradients}}$$

δ: Set of perturbed data points