Understanding the Origins of Bias in Word Embeddings

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

Facilitators: Elnaz Barshan & Waseem Gharbieh

VECTOR INSTITUTE | INSTITUT VECTEUR
ELEMENT AI
UNIVERSITY OF TORONTO

Introduction

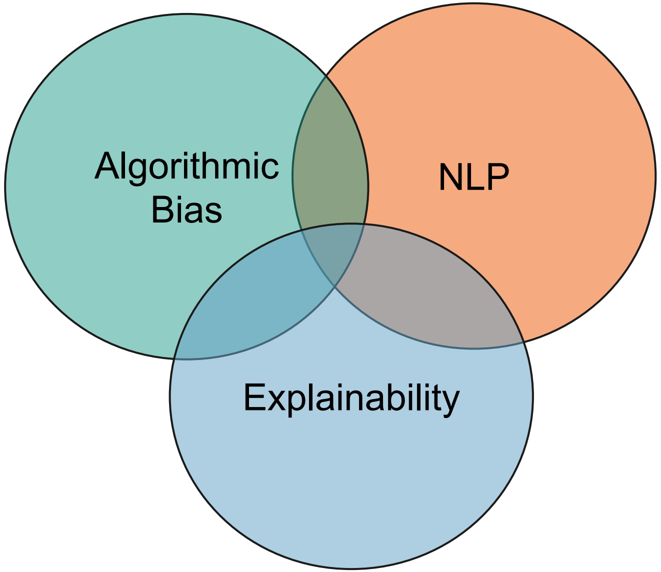
Research Intern at Element AI

Graduate student at Vector Institute (Toronto)

Supervised by Richard Zemel and Ashton Anderson

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

Collaborated with Colleen Alkalay-Houlihan



VECTOR INSTITUTE | INSTITUT VECTEUR

ELEMENT AI

UNIVERSITY OF TORONTO

Presentation Structure

What's on the menu?

1. Motivation
2. Background
3. Overview of Method
4. Technical Details
5. -- Break --
6. Results
7. Discussion

Motivation

A Motivating Example

green
PARTY OF CANADA

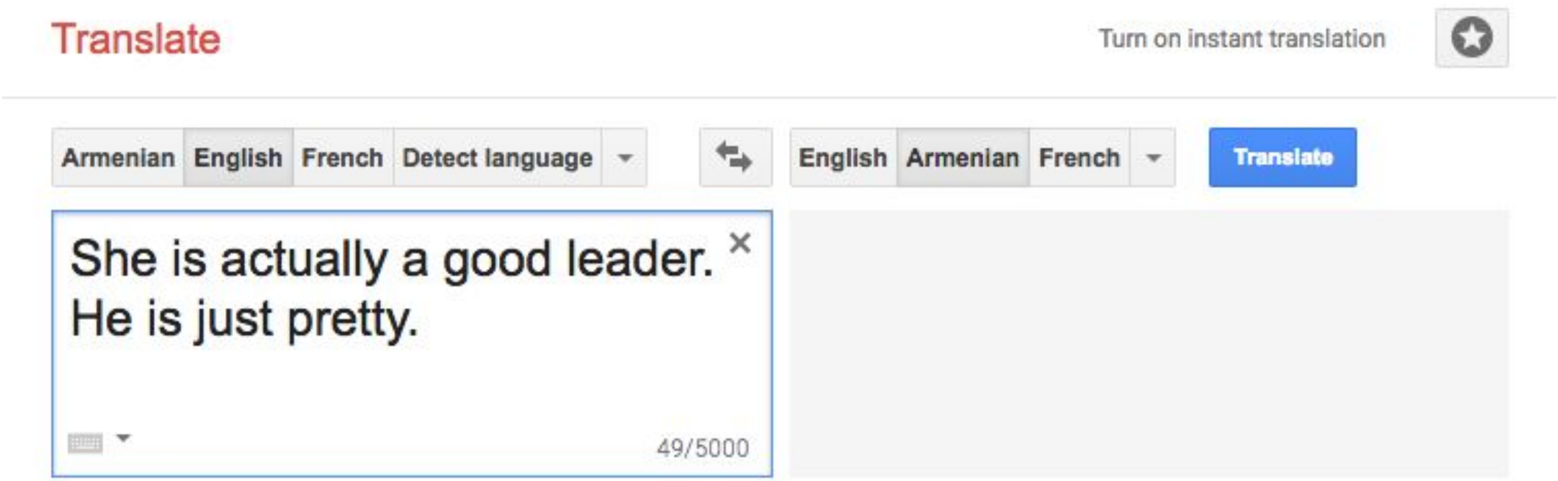
LIBERAL
PARTY OF
CANADA

A Motivating Example

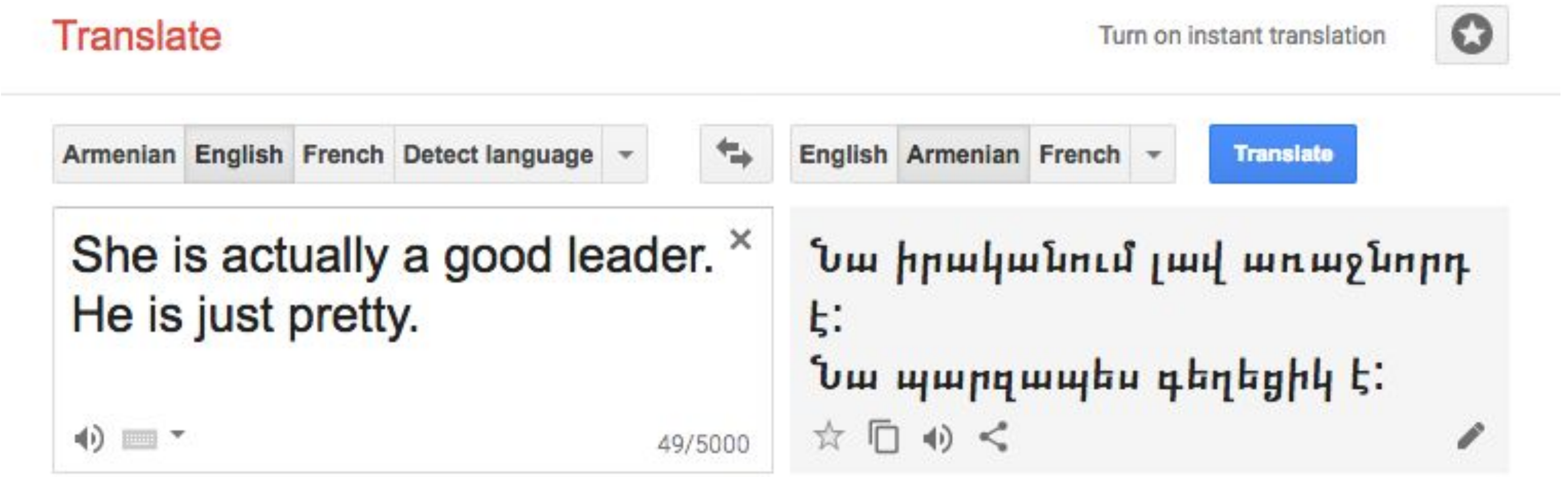
"She is actually a good leader. He is just pretty." #NoPlanetB

LIBERAL PARTY OF CANADA

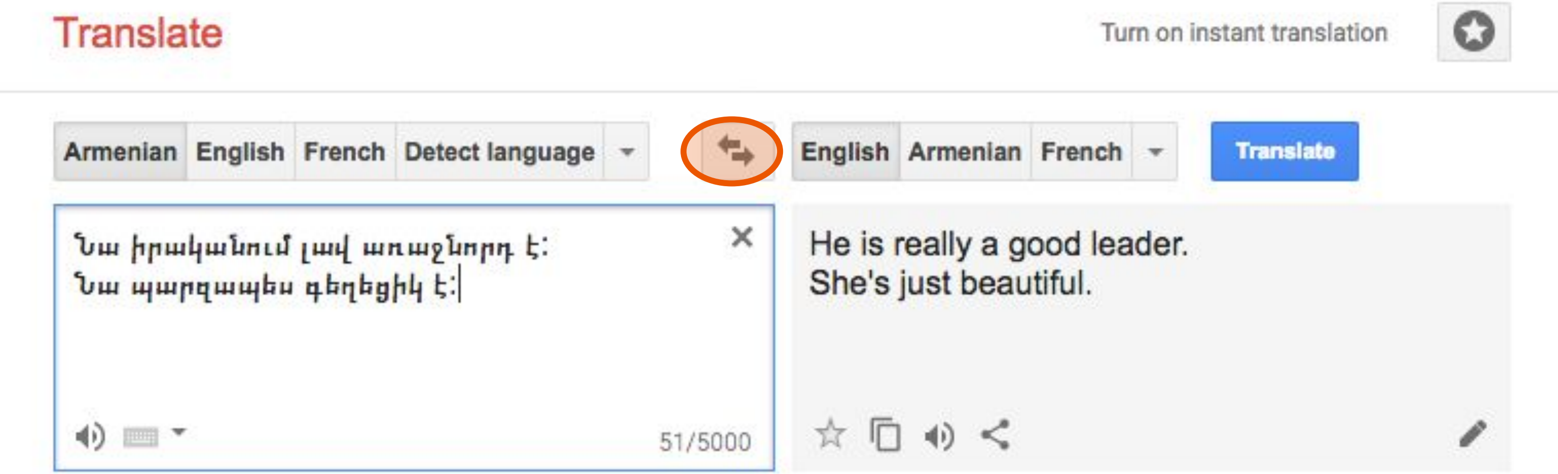
Presumptuous Translation

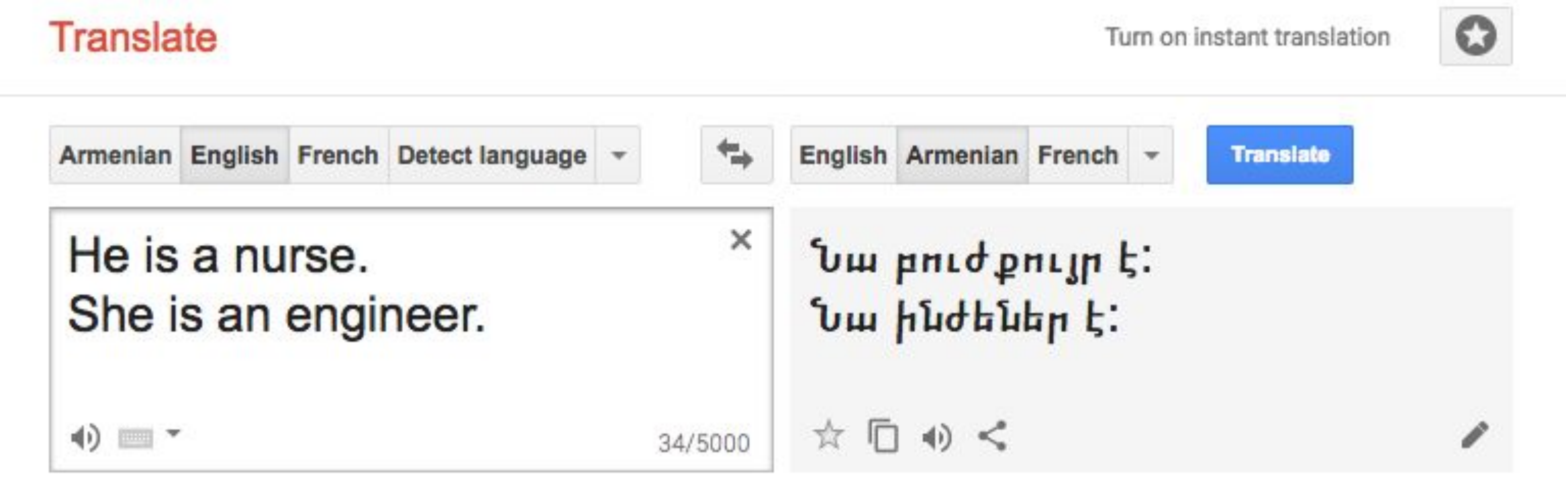


Presumptuous Translation



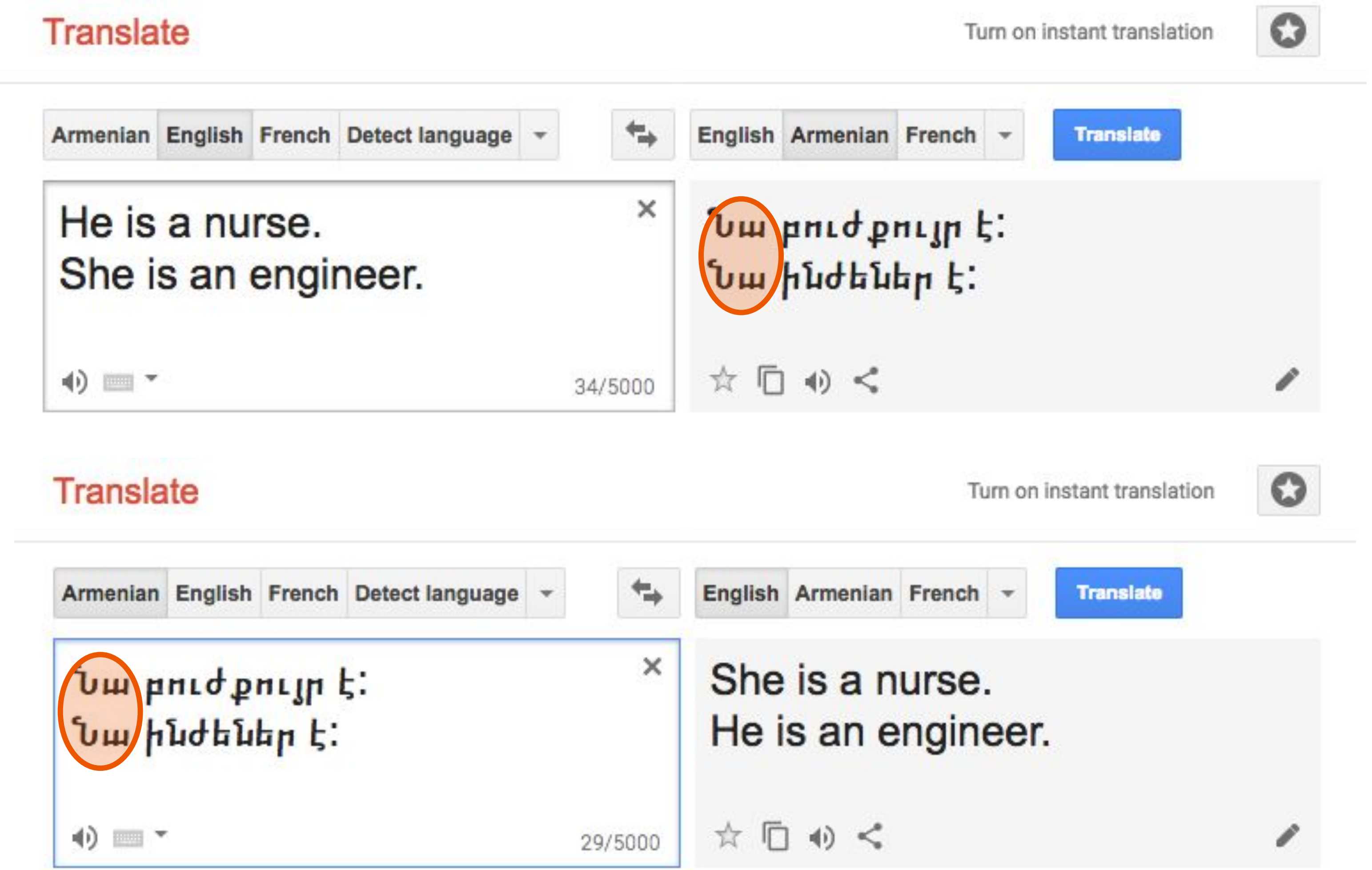
Presumptuous Translation




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)

We want a more detailed understanding.

1) To adjust the models
2) To learn about bias generally

Background:
Word Embeddings & Bias

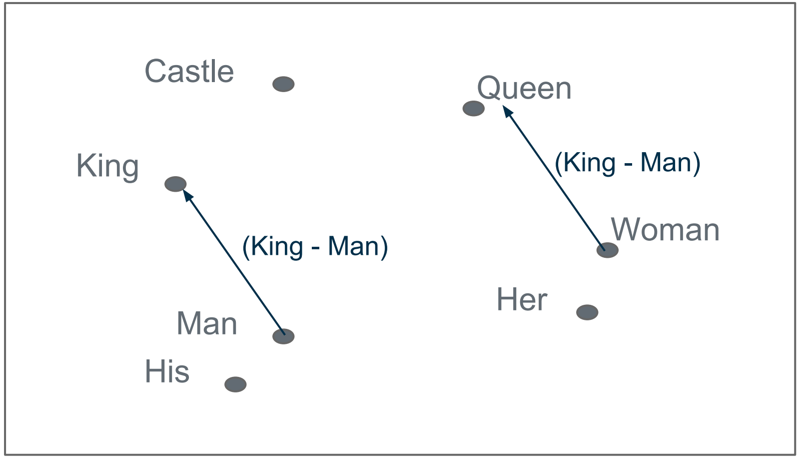
Word Embeddings

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:

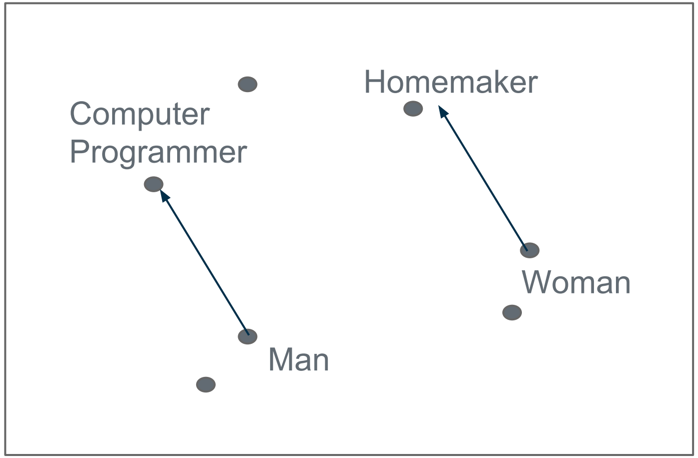
- Sentiment analysis
- Document classification / summarization
- Translation
- Temporal semantic trajectories



“King” - “Man” + “Woman” ≈ “Queen”

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 (NIPS 2016)

Measuring Bias

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

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

		IAT		WEAT	
Target Words	**Attribute Words**	**d**	**P**	**d**	**P**
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)

WEAT

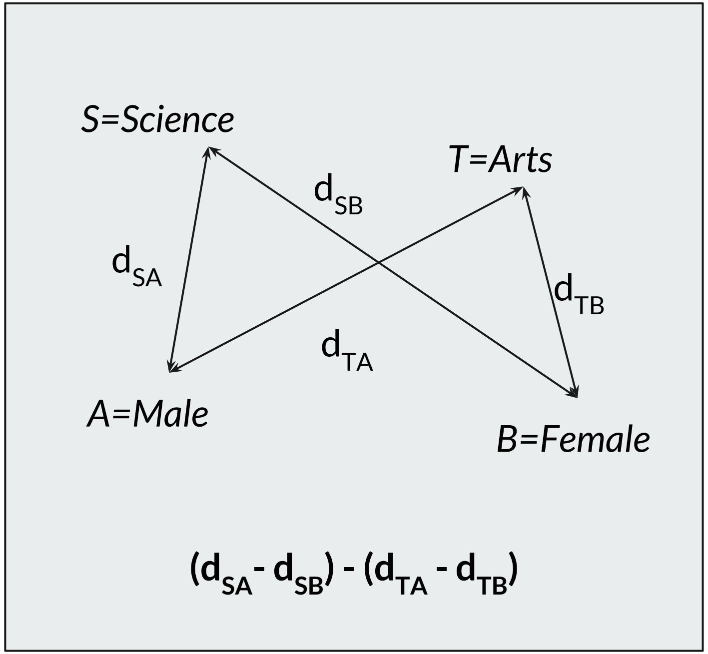
Target Word Sets:
S = {physics, chemistry… } ≈ *Science*
T = {poetry, litterature… } ≈ *Arts*

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

Measures relative association between four concepts

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

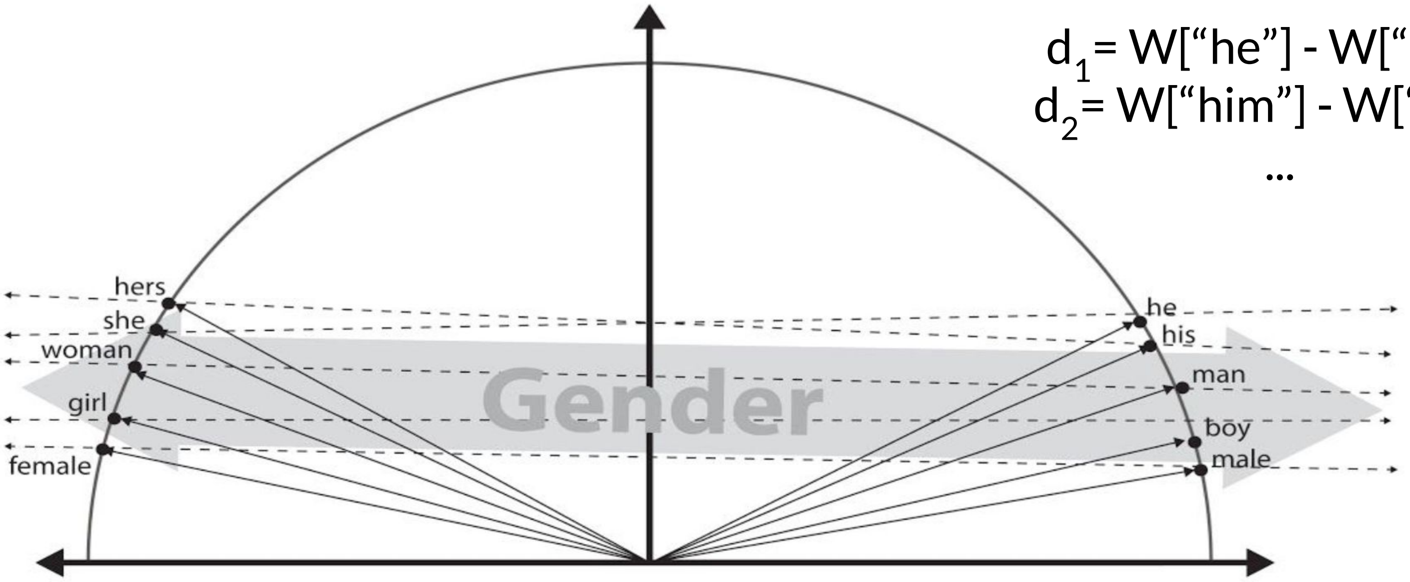
Effect Size = $\dfrac{\underset{s \in S}{\text{mean}}\, f(s, A, B) - \underset{t \in T}{\text{mean}}\, f(t, A, B)}{\underset{w \in S \cup T}{\text{std-dev}}\, f(w, A, B)}$



The Geometry of Bias

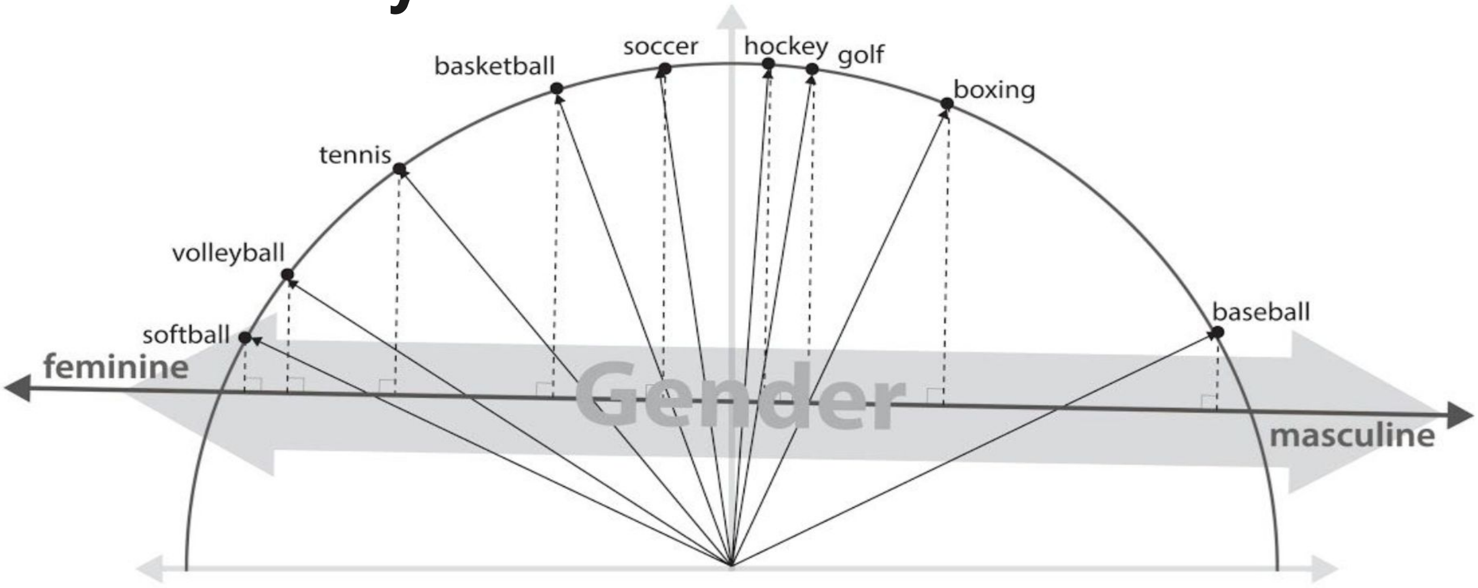
Find axis by running PCA on definitional sets of words:

d_1= W["he"] - W["she"]
d_2= W["him"] - W["her"]
...

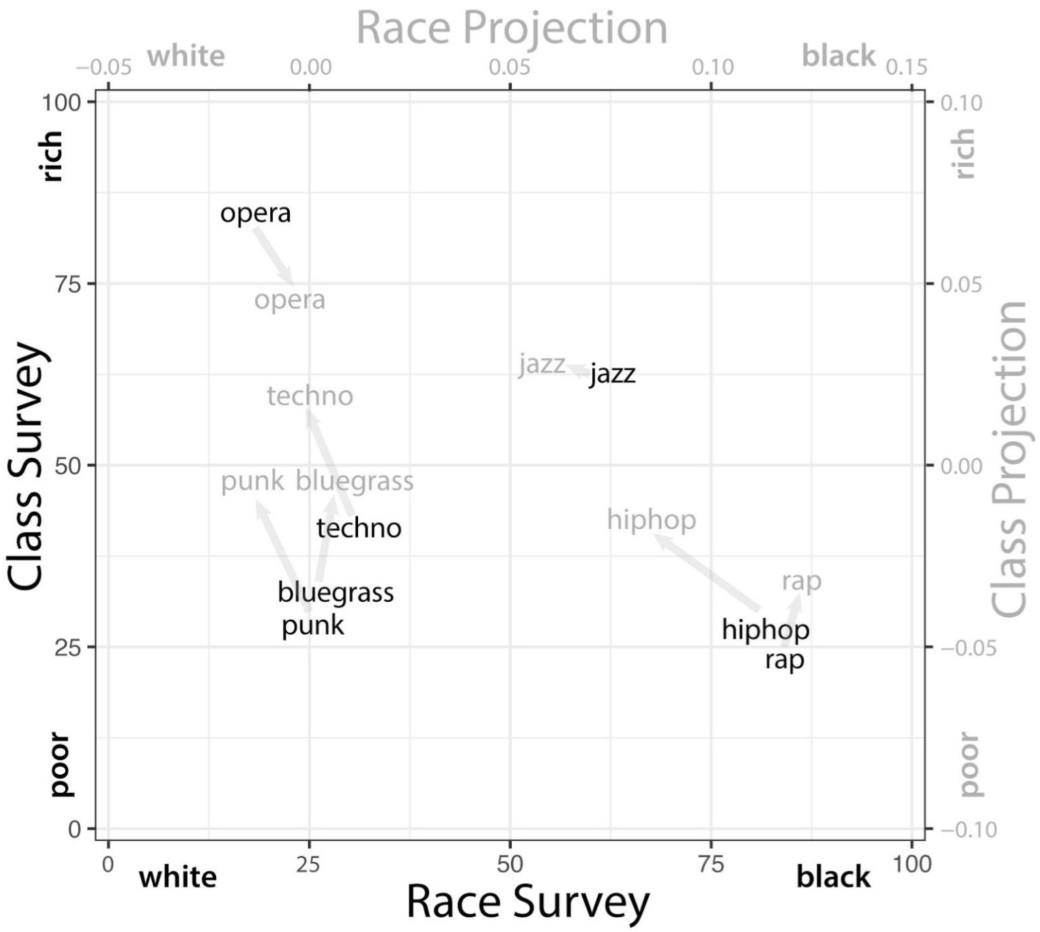


Austin C. Kozlowski, Matt Taddy, James A. Evans (2018)

The Geometry of Bias

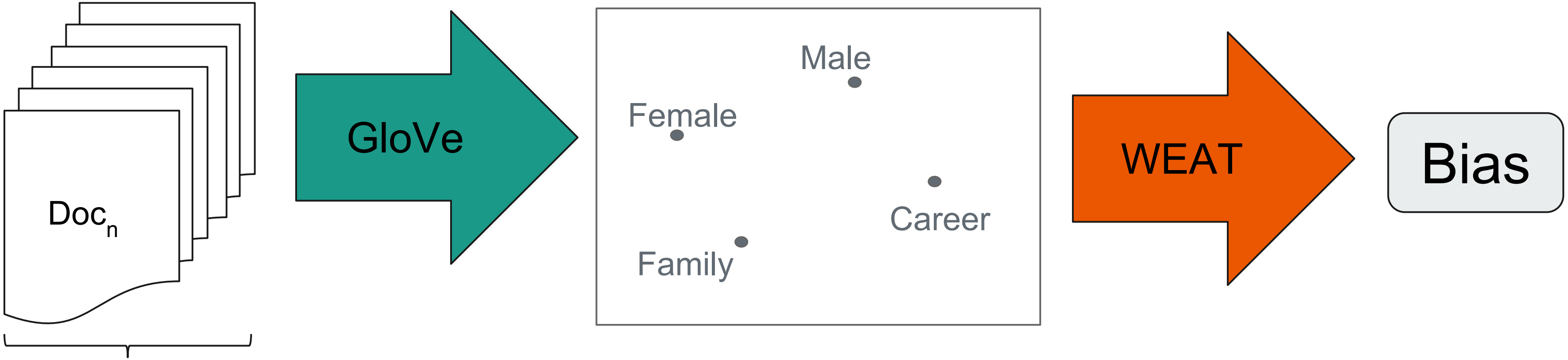


Austin C. Kozlowski, Matt Taddy, James A. Evans (2018)



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

Word2Bias Pipeline



X: Corpus
(e.g. Wikipedia, NYT)

W: Word Embedding

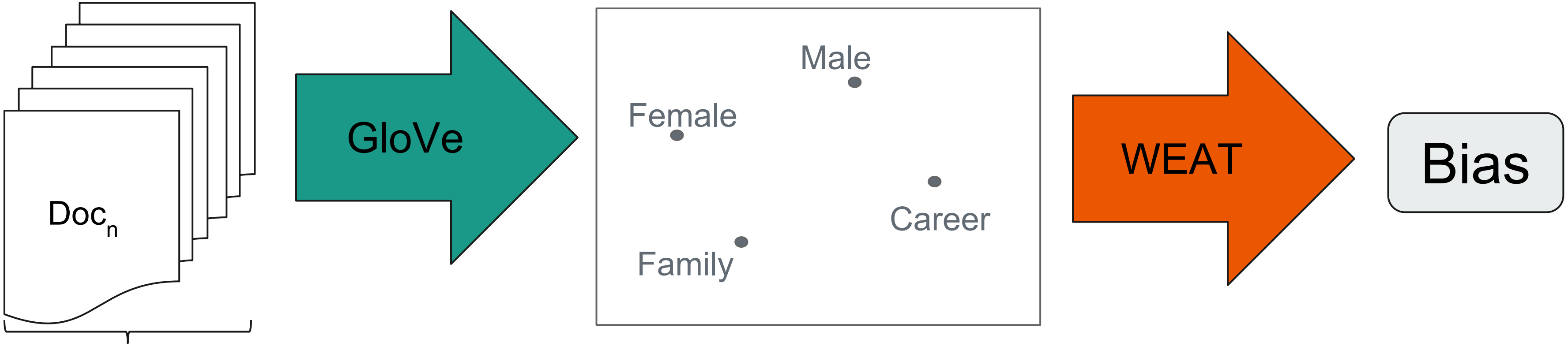
B: Bias

How do individual (sets of) documents within the corpus contribute to this measured bias?

Overview of Methodology

Word2Bias Pipeline

$$\frac{dB}{dX} = \frac{dW}{dX}\frac{dB}{dW}$$



X: Corpus
(e.g. Wikipedia, NYT)

W: Word Embedding

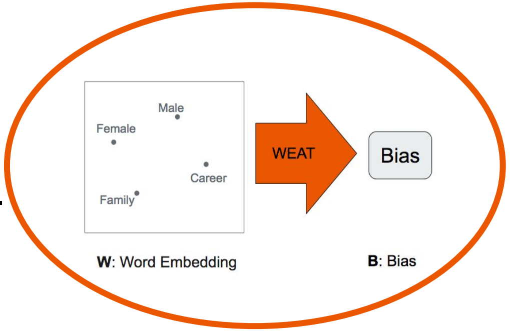
B: Bias

Computing the Components

$$\frac{dB}{dW}$$

Easy. Do the math, or use automatic differentiation.

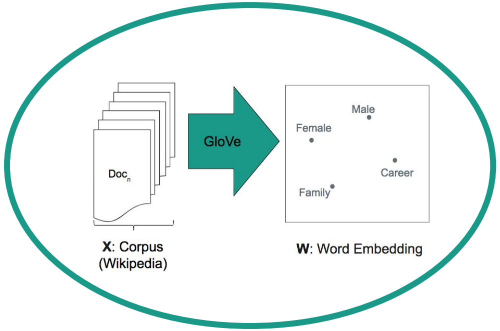
Alternatively consider: $\Delta B = B(\tilde{w}) - B(w)$



$$\frac{dW}{dX}$$

Hard. Differentiate through an entire training procedure… options:

- Leave-one-out retraining (*very slow*)
- Backprop?
- Approximate using **Influence Functions** reintroduced by *Pang Wei Koh & Percy Liang (ICML 2017)*

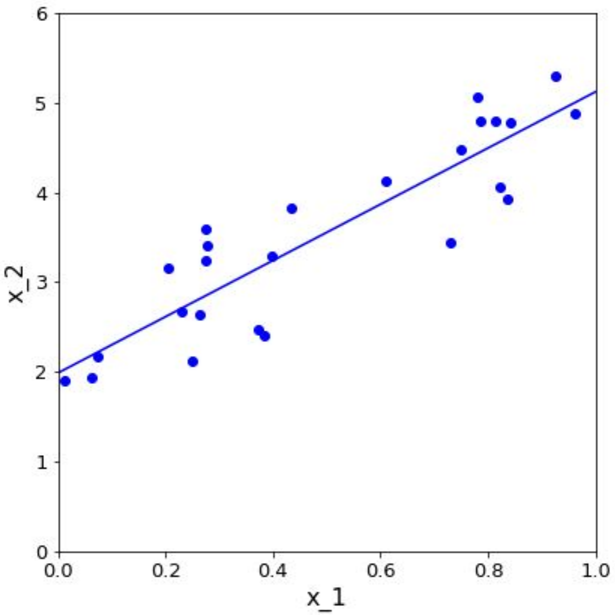


Influence Functions

Optimal model parameters

$$\hat{W} = \underset{W}{\operatorname{argmin}}\ L(W, X)$$

$$y = a\,x + b$$

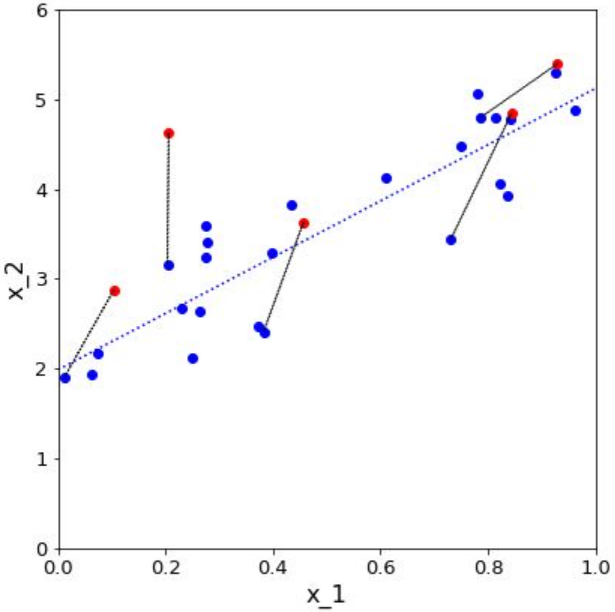


Influence Functions

What happens when you perturb the data?

$$\tilde{W} = \underset{W}{\operatorname{argmin}}\ L(W, \tilde{X})$$

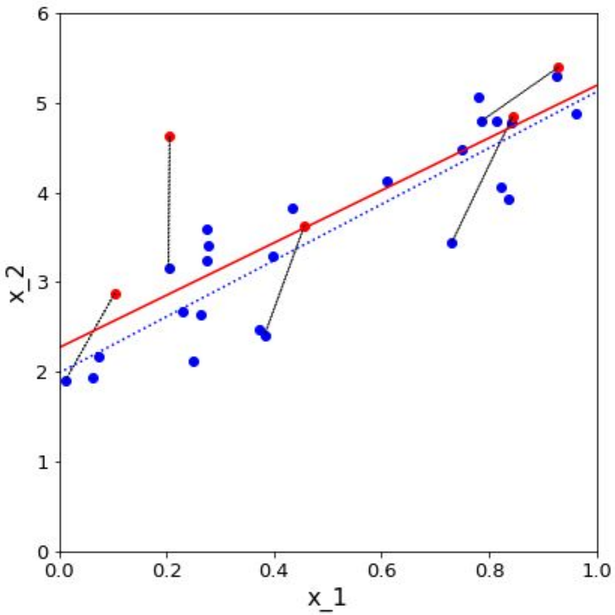
$$y = \tilde{a}\, x + \tilde{b} \quad ?$$



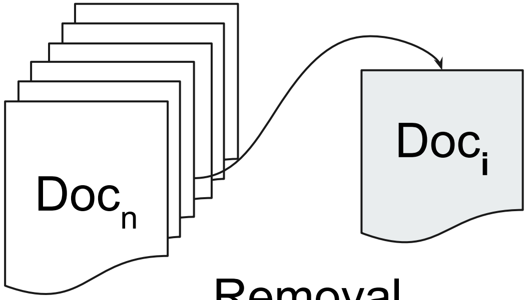
Influence Functions

Gives us a way to approximate the change in model parameters

$$\Delta \hat{W} \approx \frac{d\hat{W}}{dX} \Delta X$$



Applied to Word Embeddings



$$\tilde{X} = \overbrace{X - X_i}^{\text{Removal}}$$

$$\hat{W} = \underset{W}{\operatorname{argmin}}\, L(W, X)$$

$$\Delta\hat{W} \approx [\underbrace{\nabla^2 L(\hat{W}, X)}_{\text{Hessian (very big…)}}]^{-1}\Big(\nabla L(\hat{W}, X) - \nabla L(\hat{W}, \tilde{X})\Big)$$

$$\frac{dB}{dW} \times \Delta \hat{W}$$

$$\Delta B_i$$

Differential Bias
(of document *i*)

Technical Details

Influence Function (IF) Derivation

Generic ML Problem:

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

Optimal params under perturbation

$$\tilde{\theta} = \underset{\theta}{\operatorname{argmin}} \left\{ J(z,\theta) + \underbrace{\varepsilon L(\tilde{z}_k,\theta)}_{\text{perturbed pt.}} - \underbrace{\varepsilon L(z_k,\theta)}_{\text{original pt.}} \right\}$$

$$\text{we seek } \tilde{\theta}|_{\varepsilon=\frac{1}{n}}, \text{ noting that } \tilde{\theta}|_{\varepsilon=0} = \theta^*$$

$$\tilde{\theta} = \underset{\theta}{\operatorname{argmin}} \left\{ J(z, \theta) + \varepsilon L(\tilde{z}_k, \theta) - \varepsilon L(z_k, \theta) \right\}$$

IF Derivation

we seek $\tilde{\theta}|_{\varepsilon = \frac{1}{n}}$, noting that $\tilde{\theta}|_{\varepsilon = 0} = \theta^*$

1st Order Opt.

$$0 = \nabla_\theta J(z, \tilde{\theta}) + \varepsilon \nabla_\theta L(\tilde{z}_k, \tilde{\theta}) - \varepsilon \nabla_\theta L(z_k, \tilde{\theta})$$

Taylor Expand in θ (around θ*)

$$0 \approx \cancel{\nabla_\theta J(z, \theta^*)} + \varepsilon \nabla_\theta L(\tilde{z}_k, \theta^*) - \varepsilon \nabla_\theta L(z_k, \theta^*) + \left[\nabla^2_\theta J(z, \theta^*) + \underbrace{\varepsilon \nabla^2_\theta L(\tilde{z}_k, \theta^*) - \varepsilon \nabla^2_\theta L(z_k, \theta^*)}_{\text{Relatively Small}}\right](\tilde{\theta} - \theta^*)$$

$$\boxed{\tilde{\theta} - \theta^* \approx \left(\frac{-1}{n}\right) H_{\theta^*}^{-1} \left[\nabla_\theta L(\tilde{z}_k, \theta^*) - \nabla_\theta L(z_k, \theta^*)\right]}$$

Hessian of Total Loss:

$$H_{\theta^*} = \left[\nabla^2_\theta J(z, \theta^*)\right]$$

GloVe

GloVe: Global Vectors for Word Representations*

Learns an embedding from a corpus by:

1) Extracting a vocabulary of size V
2) Constructing a co-occurrence matrix **X** (V by V)
3) Learning an embedding $\{\mathbf{w}_i\}$ (V by D)

Constructing X:

w_2 w_{31} w_{42} w_{68} w_{25} w_{18}

$$\underbrace{\textit{The quick brown fox jumped over the fence.}}_{\text{window size: 6}}$$

X[2, 31] += 1	X[31, 2] += 1
X[2, 42] += ½	X[42, 2] += ½
X[2, 68] += ⅓	X[68, 2] += ⅓
X[2, 25] += ¼	X[25, 2] += ¼
…	…

Note we can sum coocs from all docs: $\mathbf{X} = \sum X^{(k)}$

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

Applying IF to GloVe

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

GloVe Loss :

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

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

δ: set of perturbed points

Applying IF to GloVe

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

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

Pointwise Loss :

$$L(X_i, w) = \sum_{j=1}^{V} V f(X_{ij})(w_i^T u_j + b_i + c_j - \log X_{ij})^2$$

Total Loss :

$$J(X, w) = \frac{1}{V} \sum_{i=1}^{V} L(X_i, w)$$

Note: “datapoints” are now the rows of X

Applying IF to GloVe

Pointwise Loss

$$L(X_i, w) = \sum_{j=1}^{V} V f(X_{ij})(w_i^T u_j + b_i + c_j - \log X_{ij})^2$$

Pointwise Gradient

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

$$\nabla_{w_i} L(X_i, w) = \sum_{j=1} 2V f(X_{ij})(w_i^T u_j + b_i + c_j - \log X_{ij})\, u_j$$

$$\nabla^2_{w_i} L(X_i, w) = \sum_{j=1}^{V} 2V f(X_{ij}) u_j u_j^T$$

Total Hessian will be **Block Diagonal!**

Applying IF to GloVe

$$\tilde{w}_i - w_i^* = \underbrace{\left[\nabla^2_{w_i} L(X_i, w^*)\right]}_{\text{Computed once per WEAT word}}^{-1} \Big(\overbrace{\nabla_{w_i} L(\tilde{X}_i, w^*)}^{\text{Computed for every perturbation of interest}} - \underbrace{\nabla_{w_i} L(X_i, w^*)}_{\text{Computed once per WEAT word}}\Big)$$

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

Applied to GloVe

$$\tilde{w_i} - w_i^* = \left[\nabla^2_{w_i} L(X_i, w^*)\right]^{-1} \left(\nabla_{w_i} L(\tilde{X}_i, w^*) - \nabla_{w_i} L(X_i, w^*)\right)$$

For every perturbation (i.e. document or document set removal), compute:

1. $\{\tilde{w}_i\}$ for all i affecting WEAT
2. $\Delta B = B(\{\tilde{w}\}) - B(\{w^*\})$

Main Experimental Method

1. Train baseline embedding (10 different seeds)
2. Calculate differential bias for every document
3. Form **document sets** from most bias influencing documents
4. Predict differential bias of each document set
5. Remove sets and retrain to get ground truth (5 different seeds)
6. Compare with prediction
7. (Make other comparisons)

Break

Results

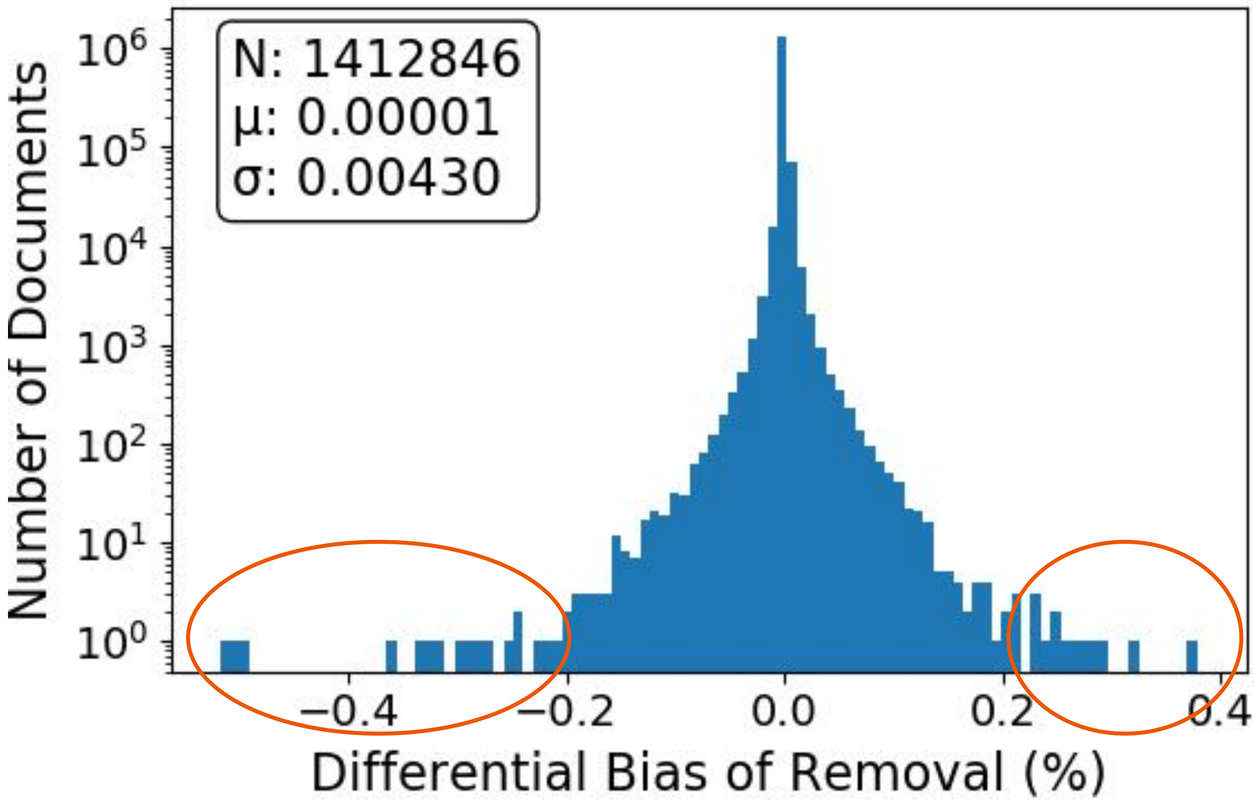
Bias: WEAT 1

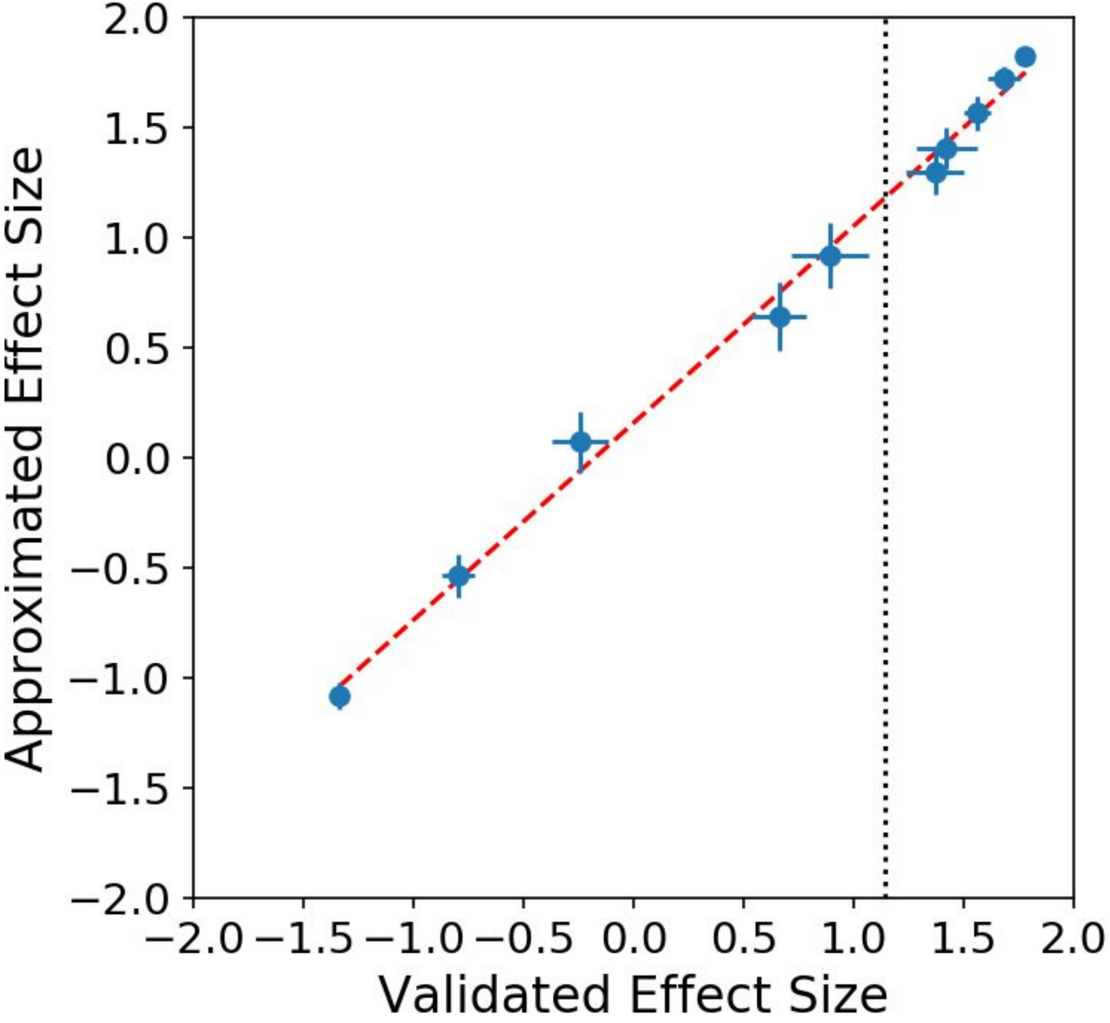
S	science	science, technology, physics, chemistry, einstein, nasa, experiment, astronomy
T	arts	poetry, art, shakespeare, dance, literature, novel, symphony, drama
A	male	male, man, boy, brother, he, him, his, son
B	female	female, woman, girl, sister, she, her, hers, daughter

Corpus:

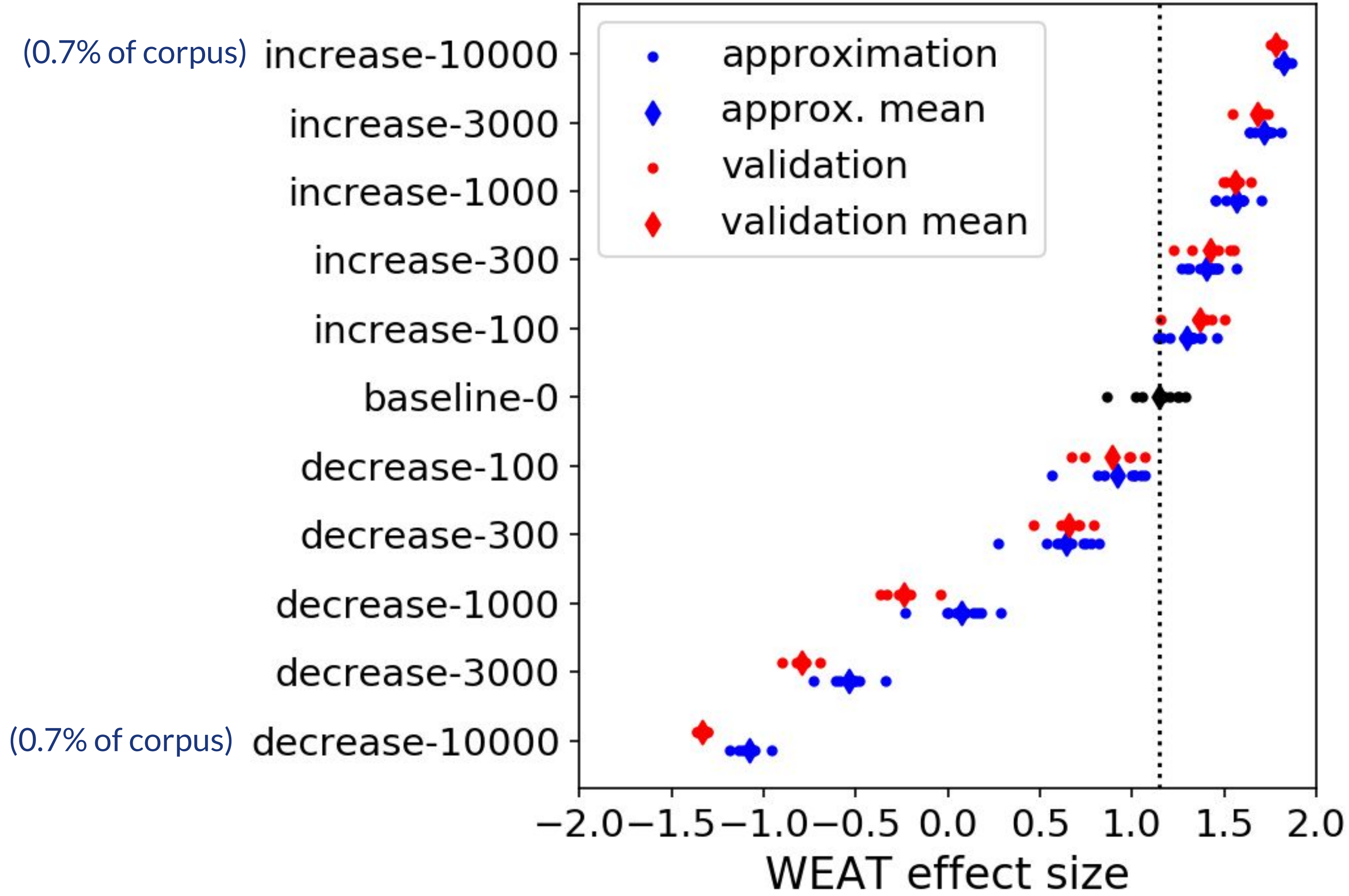
The New York Times

Differential Bias





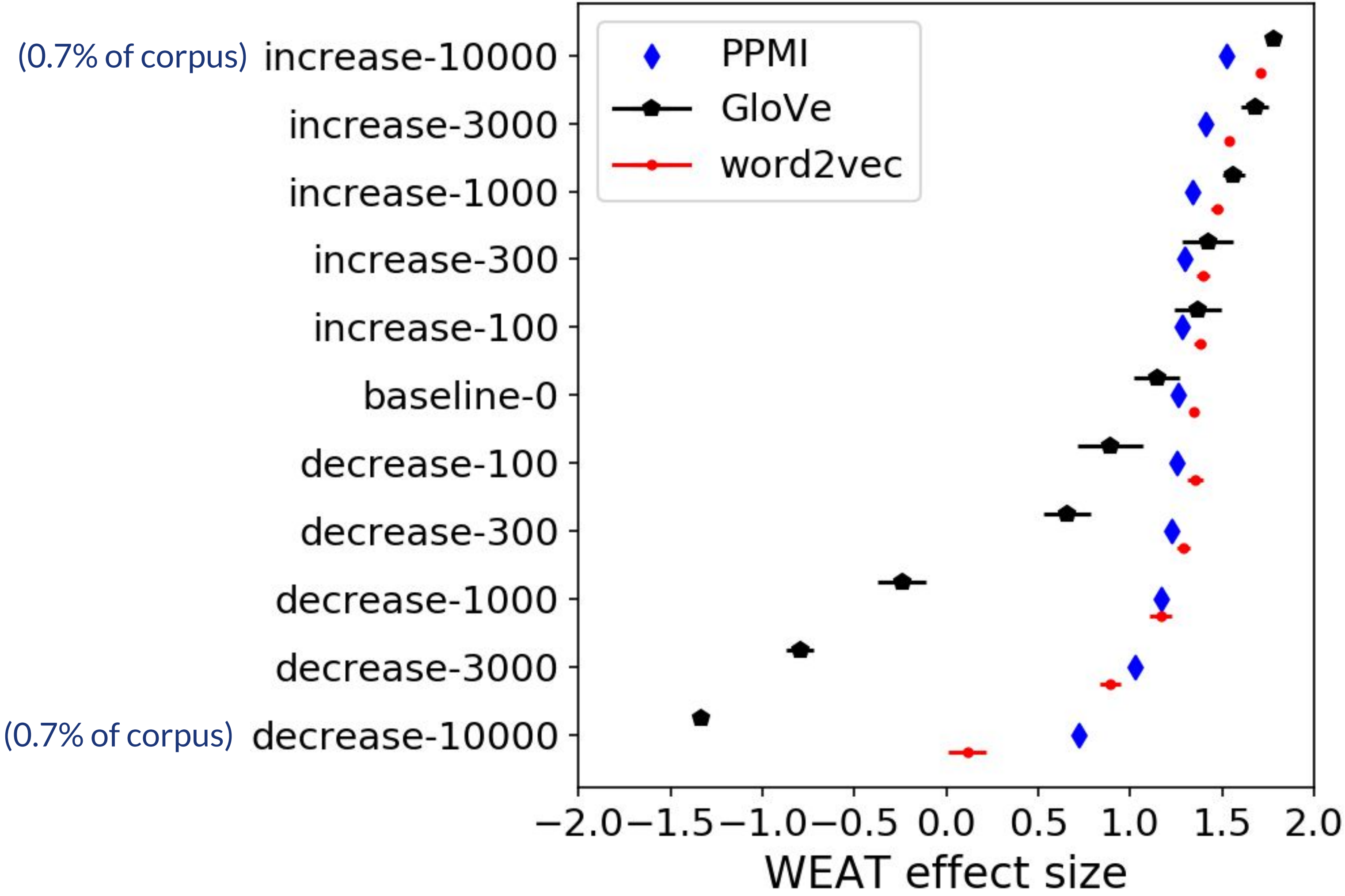
Approximated Effect Size
Validated Effect Size
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
−2.0 −1.5 −1.0 −0.5 0.0 0.5 1.0 1.5 2.0



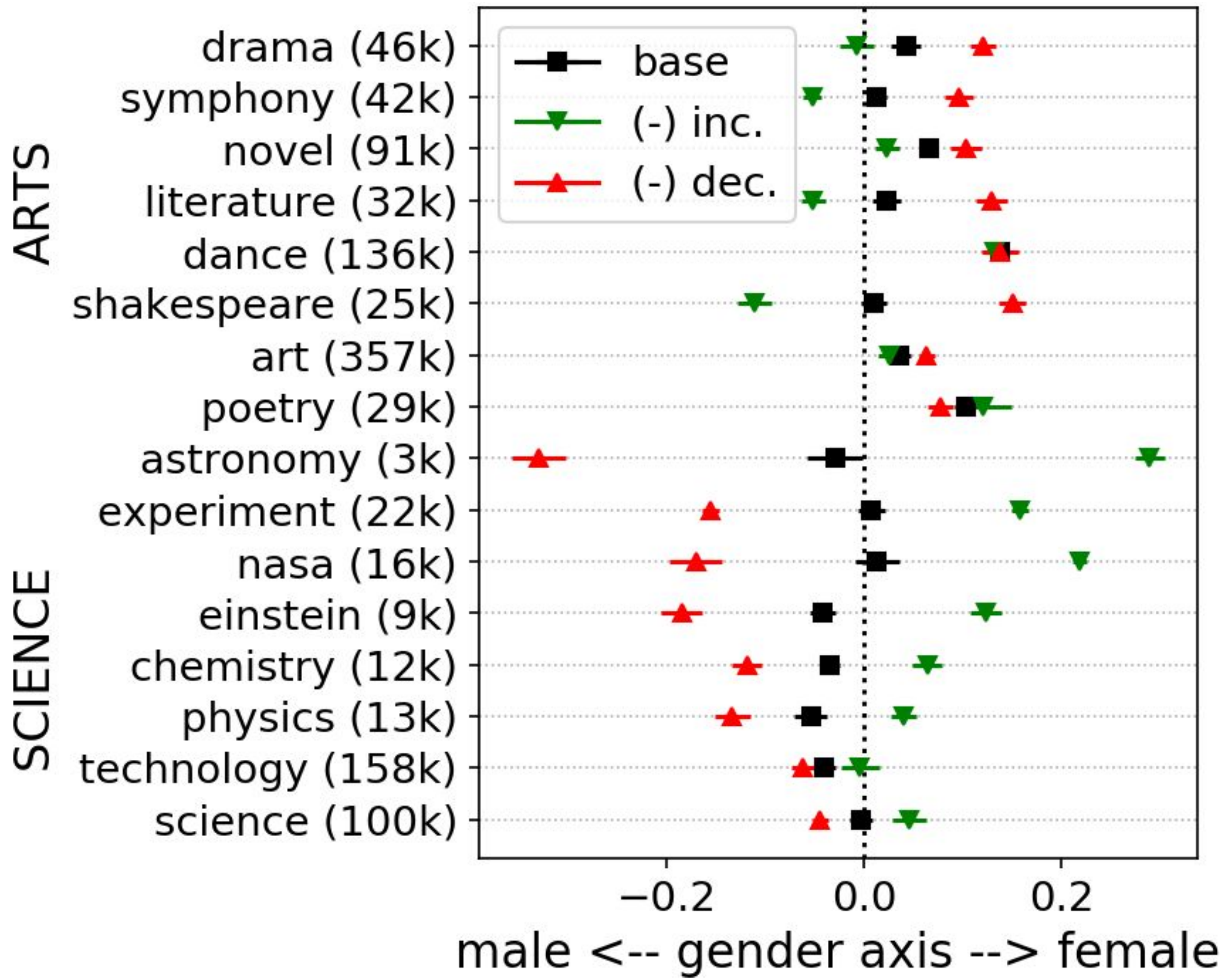
(0.7% of corpus)
increase-10000
increase-3000
increase-1000
increase-300
increase-100
baseline-0
decrease-100
decrease-300
decrease-1000
decrease-3000
(0.7% of corpus)
decrease-10000
approximation
approx. mean
validation
validation mean
−2.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
WEAT effect size

$\Delta_d B$	**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



(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
PPMI
GloVe
word2vec
−2.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
WEAT effect size



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)
base
(-) inc.
(-) dec.
−0.2
0.0
0.2
male <-- gender axis --> female

Recap

- Bias can be quantified in word embeddings, and has been shown to correlate with known human biases.
- We can approximate how corpus perturbations affect these biases using influence functions.
- We can identify the (sets of) documents most responsible for any given bias.
- These documents impact other embedding methods and other bias metrics, they also seem to be qualitatively meaningful.

Discussion Points

- How do we define “bias”? Not all biases are harmful or problematic.
- How should we remove unwanted biases in AI models? e.g.
  - Remove “bias increasing” training samples
  - Remove of “bias increasing” features (protected attributes)
  - Training models with fairness constraints
- How to search for new/other biases?
- Bias is exacerbated by extreme data points (outliers in data and ideologies)
- Bias depends on cultural norms, what is considered problematic today may have not been 100 years ago
- You have to look for bias to see it (i.e. it requires a critical lens)

Other Questions?

Updated paper under review at ICML

mebrunet@cs.toronto.edu

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.
- A. C. Kozlowski, M. Taddy, and J. A. Evans. The Geometry of Culture: Analyzing Meaning through Word Embeddings. arXiv preprint arXiv:1803.09288, 2018.

Waseem

- Bias depends on cultural norms, what is biased today may have not been 100 years ago
- Exposing bias is an iterative process
- You have to look for bias to see it (critical lens)
- Bias is exacerbated by extreme data points (outliers in data and ideologies)

Elnaz

- How do we define “bias”?
  - Not all biases are harmful
- How to detect bias?
  - Predefined bias vs unknown bias
  - Bias measure: A note on WEAT
- How to remove bias in AI models?
  - Removal of “bias increasing” training samples
  - Removal of “bias increasing” features (protected attributes)
  - Boosting the effect of “bias decreasing” training samples/features
  - The real source of bias is us!
- Following the source of bias all the way to the attributes of the training data
  - Which sentences are responsible for “bias-increasing” behaviour of this document