What is **bias**, where does it appear in **text** data, and (how) can we remove it?

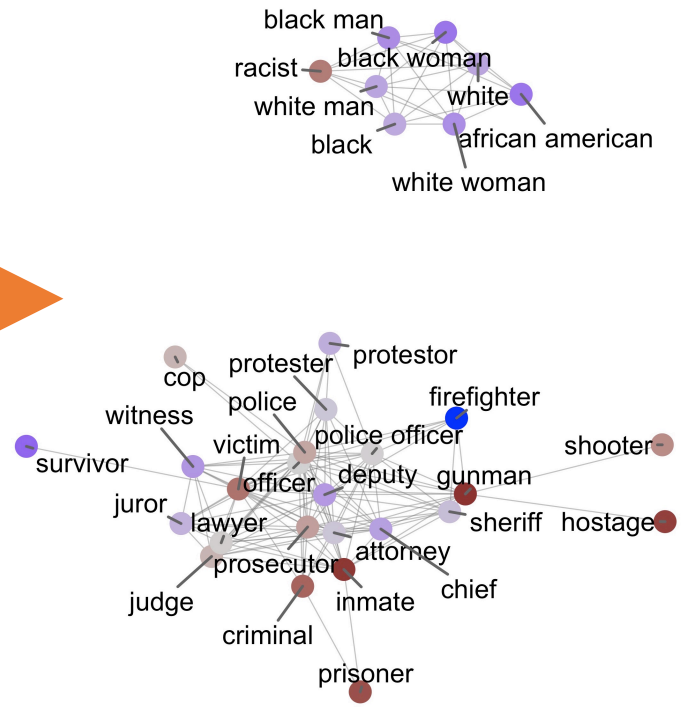
Kenneth (Kenny) Joseph

University at Buffalo
Department of Computer Science and Engineering
School of Engineering and Applied Sciences

5/5/20

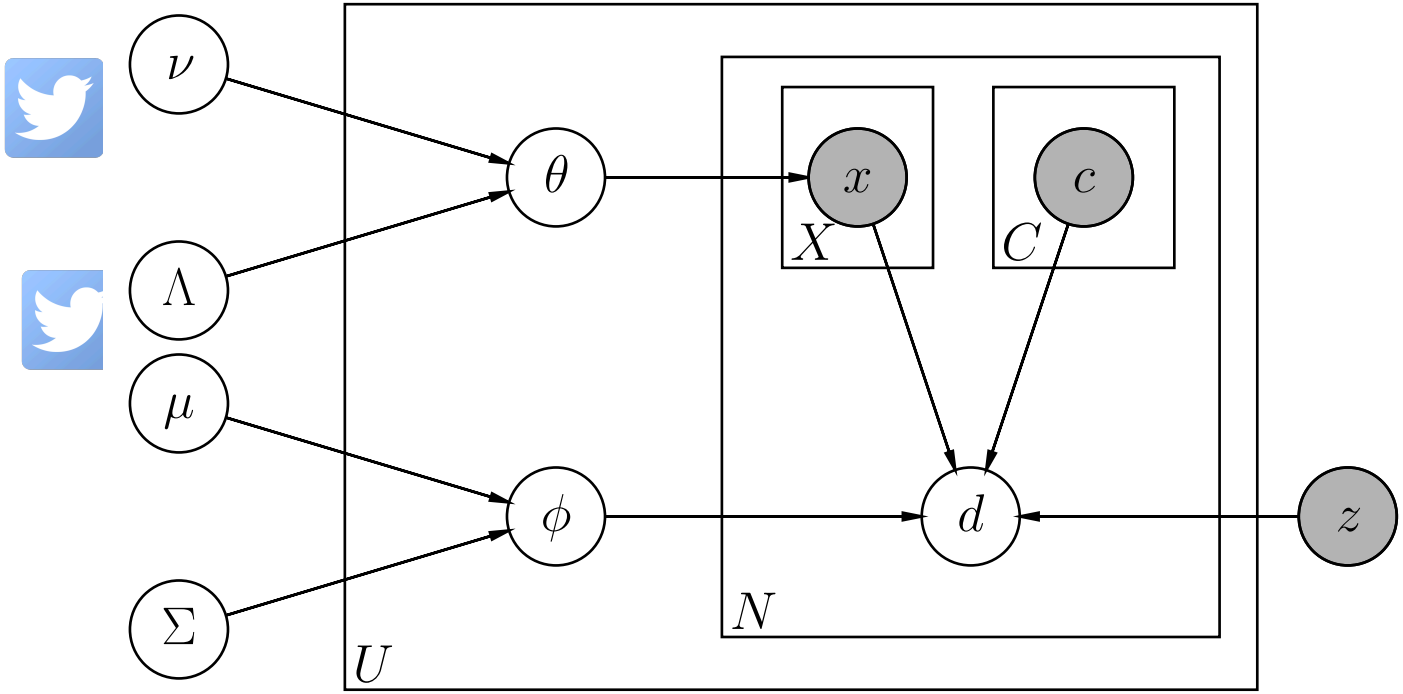
How can we use machine learning and large text corpora to better understand people/culture?

https://www.nbcnews.com/storyline/michael-brown-shooting/michael-brown-death-ferguson-police-chief-asks-calm-n179936

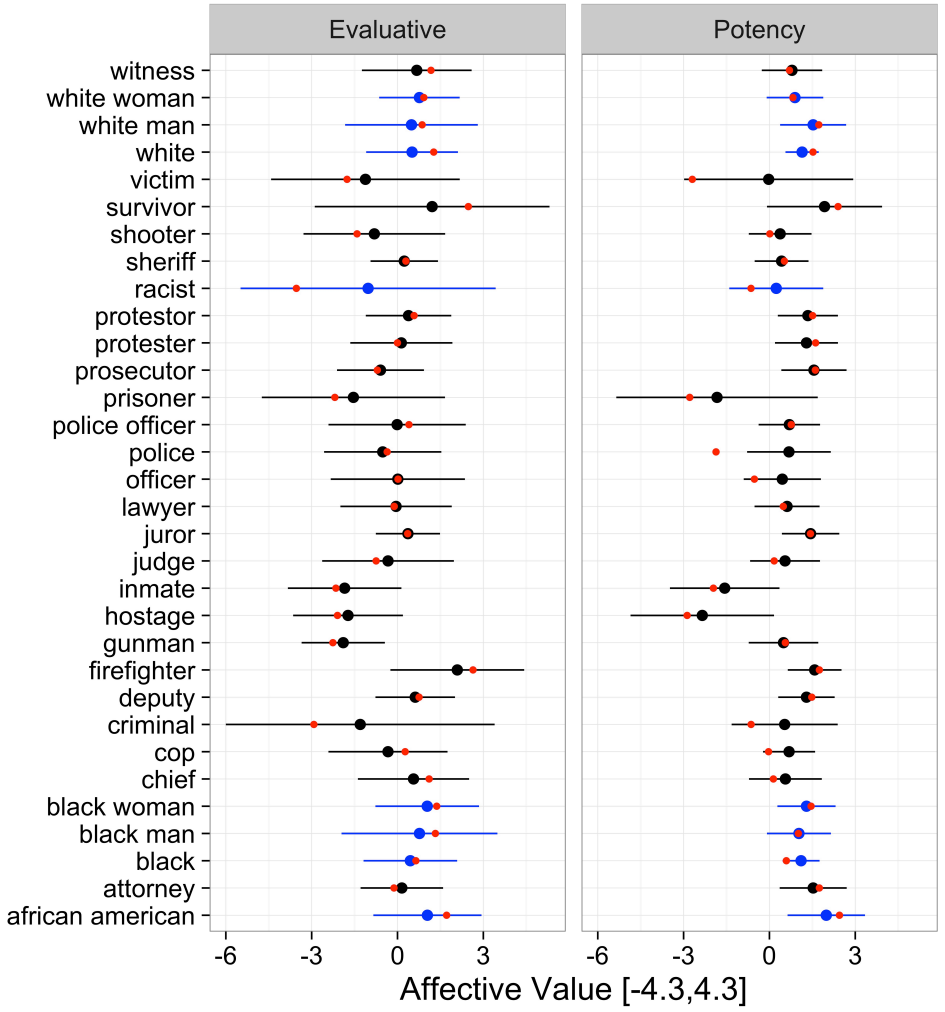


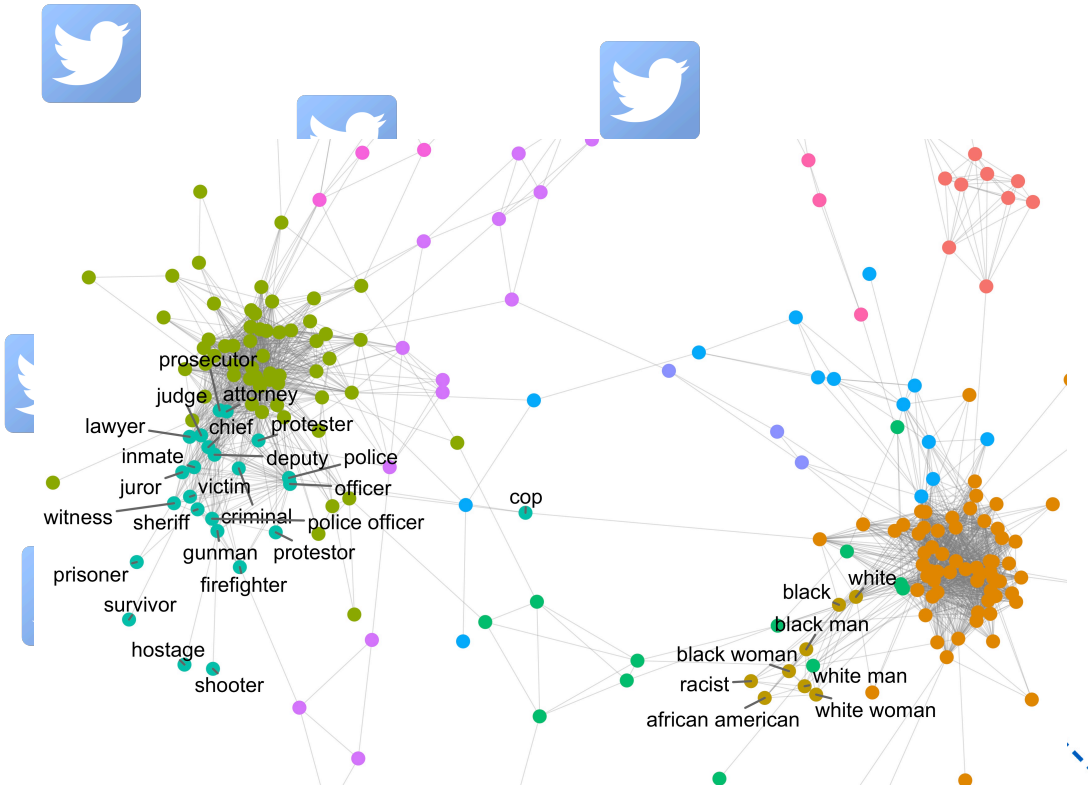
@_kenny_joseph

So terrible that a young man was killed by a police officer.

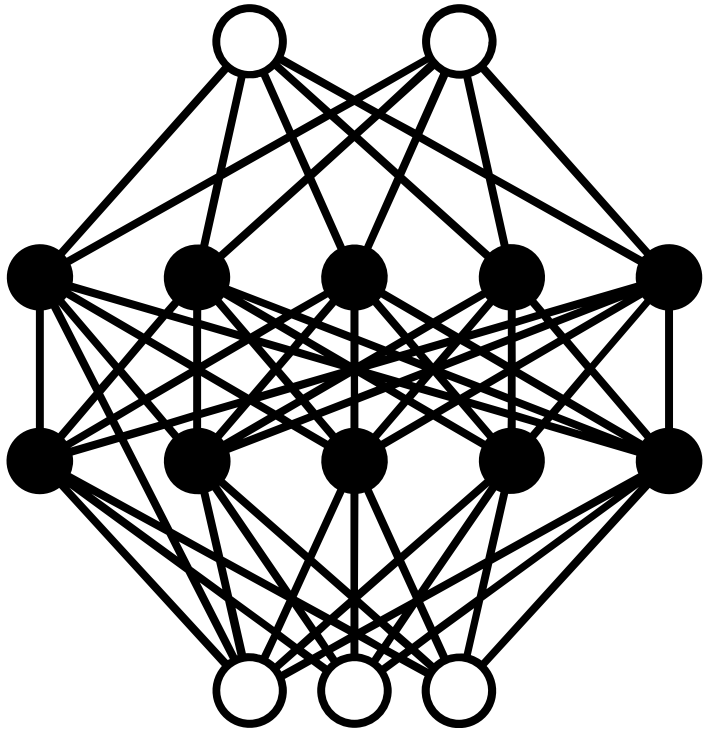


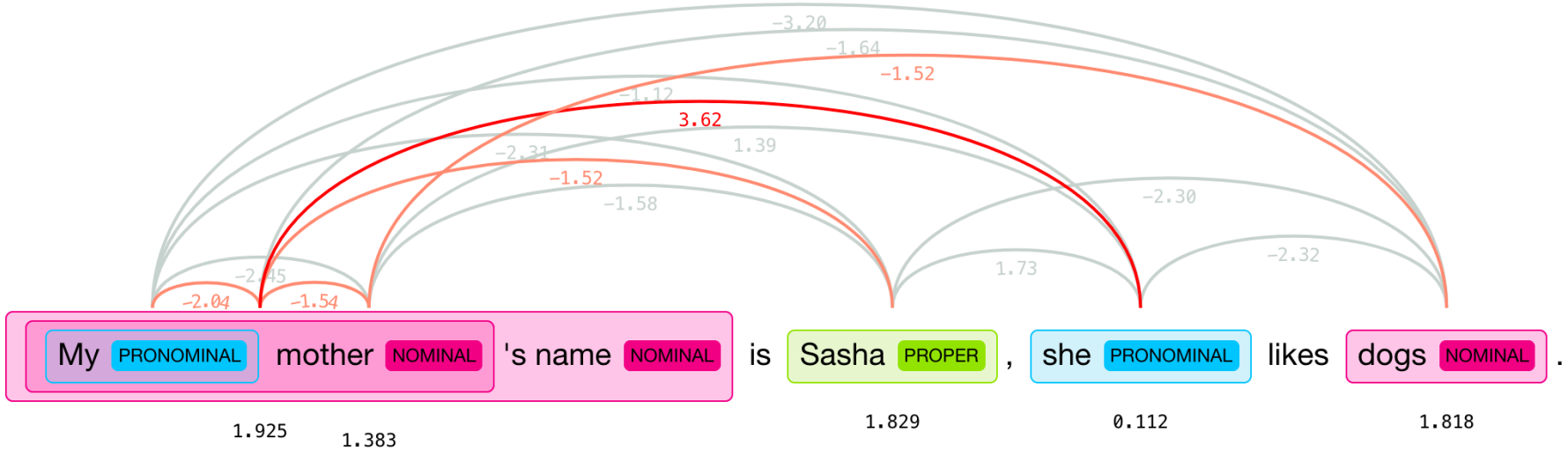
So terrible that a young man was killed by a police officer.





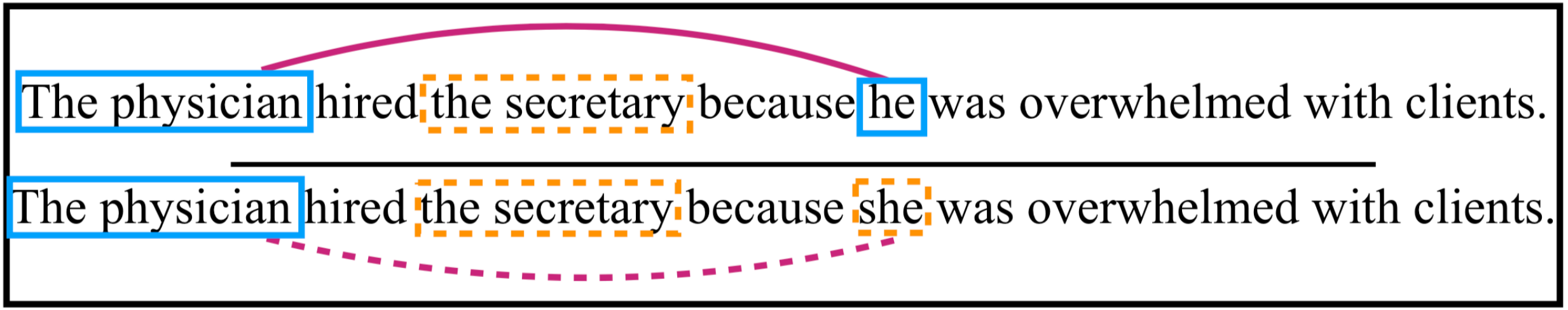
@_kenny_joseph





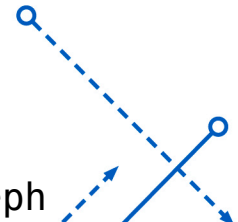
How can we use ML to 1) better understand text to 2) build machines that can do people things?

@_kenny_joseph



Zhao, J., Wang, T., Yatskar, M., Ordonez, V., & Chang, K.-W. (2018). Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods. *ArXiv:1804.06876 [Cs]*.

https://huggingface.co/coref



Kenneth Joseph

Website: kennyjoseph.github.io

Email: josephkena@gmail.com

Github: kennyjoseph

Phone: (716) 983-4115

Address:
Computer Science and Engineering Dept.
University at Buffalo
335 Davis Hall
Buffalo, NY, 14221

Academic Appointments

Asst. Professor	Computer Science	University of Buffalo	2018-
Postdoc	Network Science Institute	Northeastern University	2016-2018
Fellow	Institute for Quantitative Social Science	Harvard University	2016-2018
Fellow	Data Science for Social Good	University of Chicago	2015

Education

Ph.D.	Societal Computing	Carnegie Mellon University	2016
M.S.	Societal Computing	Carnegie Mellon University	2012
B.S.	Computer Science	University of Michigan-Ann Arbor	2010

Thesis: "Latent Cognitive Social Spaces: theory and methods for extracting prejudice from text".
Committee Members: Kathleen Carley (SI, CMU; Chair), Jason Hong (HCII, CMU), Lynn Smith-Lovin (Sociology, Duke), Eric Xing (ML/LTI, CMU)

Publications

Conference

Joseph, K., Swire-Thompson, B., Masuga, H., Baum, M., & Lazer, D. (2019). Polarized, Together: Comparing Partisan Support for Trump's Tweets Using Survey and Platform-based Measures. *ICWSM*.

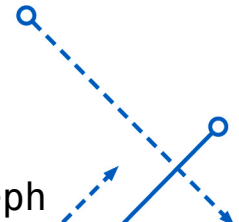
Joseph, K., Wihbey, J. (2019). Breaking News and Younger Twitter Users: Comparing Self-Reported Motivations to Online Behavior. *SMSociety*.

Robertson, R. E., Jiang, S., **Joseph, K.**, Friedland, L., Lazer, D., & Wilson, C. (2018). Auditing Partisan Audience Bias within Google Search. *Proceedings of the ACM on Human-Computer Interaction, 2(CSCW)*, 148. **Best Paper Honorable Mention**

Joseph, K., Friedland, L., Tsur, O., Hobbs, W. & Lazer, D. (2017). Modeling Annotation Context to Improve Stance Classification. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* (pp. 1115-1124).

Hobbs, W., Friedland, L., **Joseph, K.**, Tsur, O., Wojcik, S. & Lazer, D. (2017). "Voters of the Year": 19 Voters Who Were Unintentional Election Poll Sensors on Twitter. *ICWSM*

How can we use ML to 1) better understand text to 2) build machines that can do people things



@_kenny_joseph

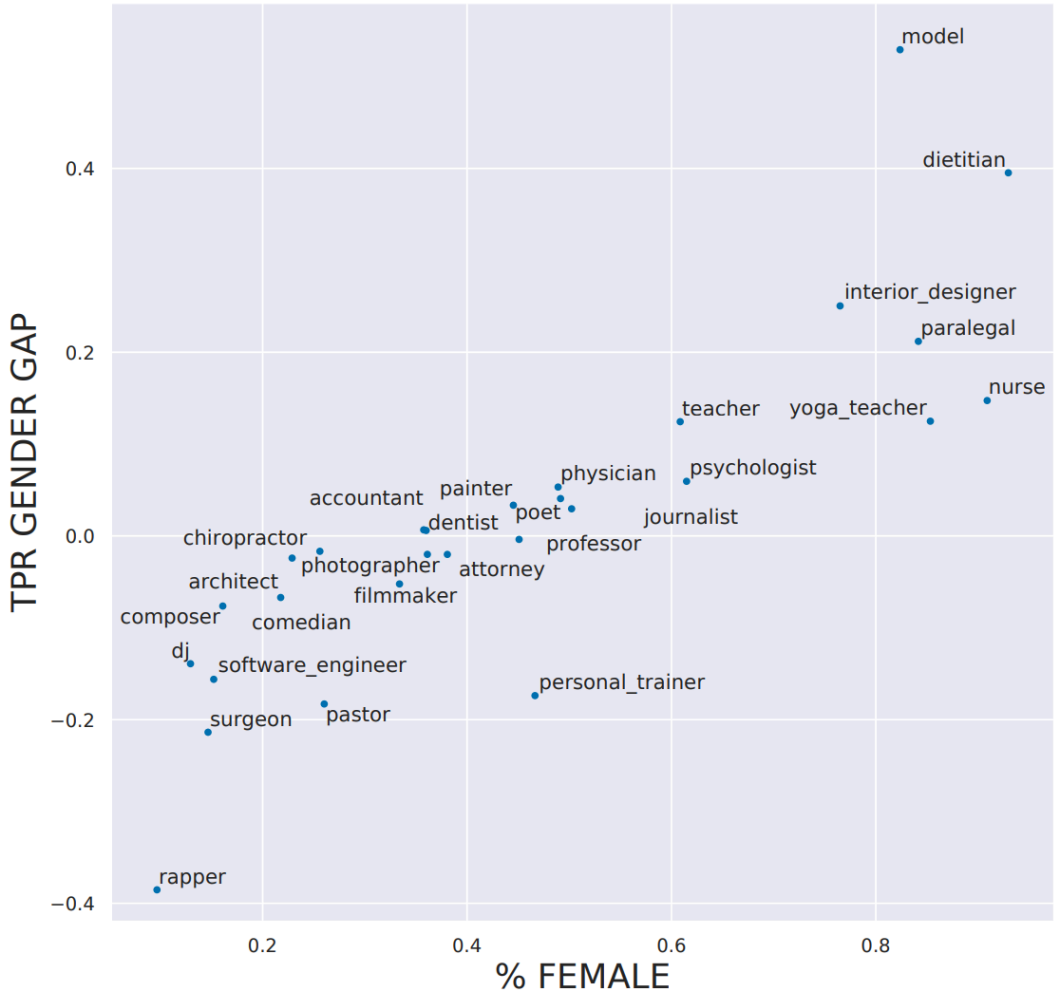
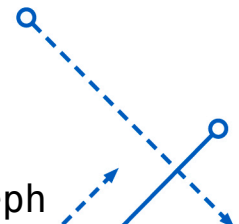
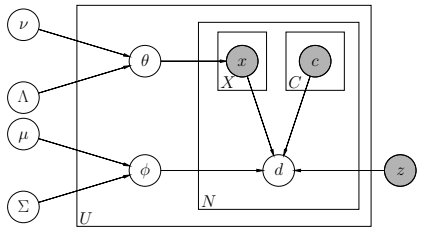


Figure 3: $\text{Gap}_{\text{female},y}$ versus $\pi_{\text{female},y}$ for each occupation y for the BOW representation with explicit gender indicators.

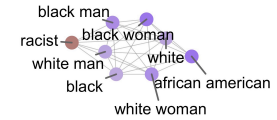
De-Arteaga, M., Romanov, A., Wallach, H., Chayes, J., Borgs, C., Chouldechova, A., ... & Kalai, A. T. (2019, January). Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 120-128). ACM.

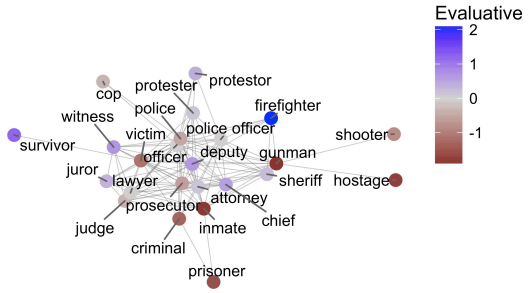


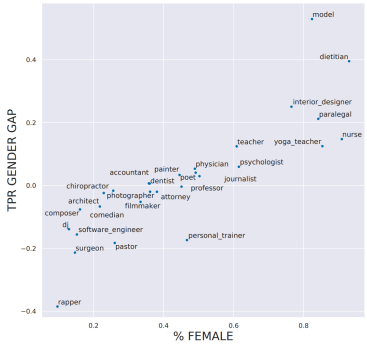
How can we use text to better understand people/ culture?



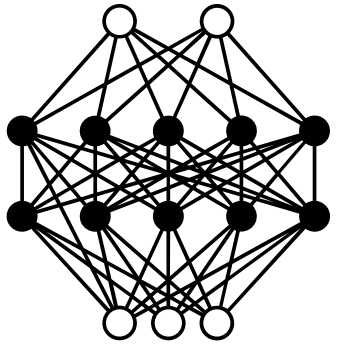
Sweet! Measurement!







How can we use ML to better understand text?



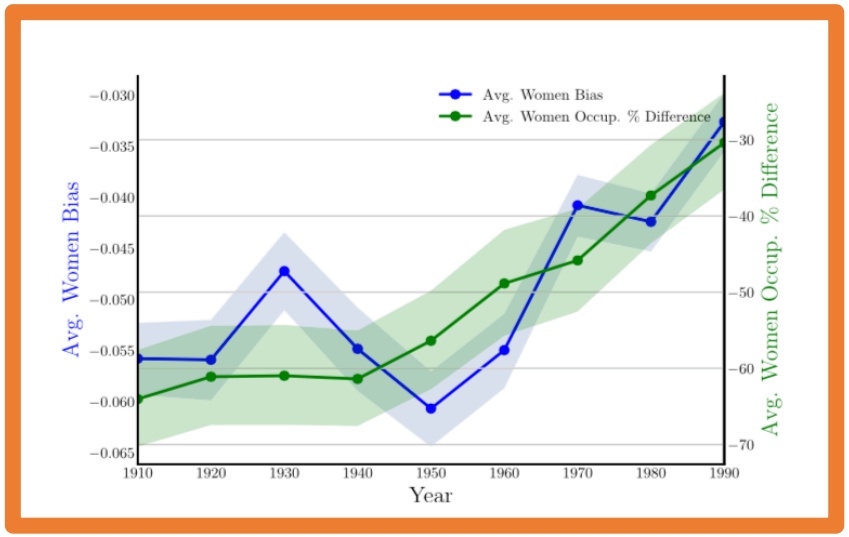
Oh no BIAS oh no oh no oh dear oh no please no

@_kenny_joseph

Today's Talk

Word Embeddings

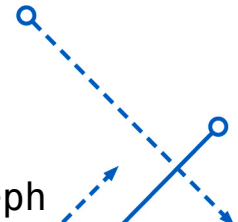
Garg, N. et al. (2018). *PNAS*, *115*(16)



Caliskan et al. (2017). *Science*, *356*(6334).

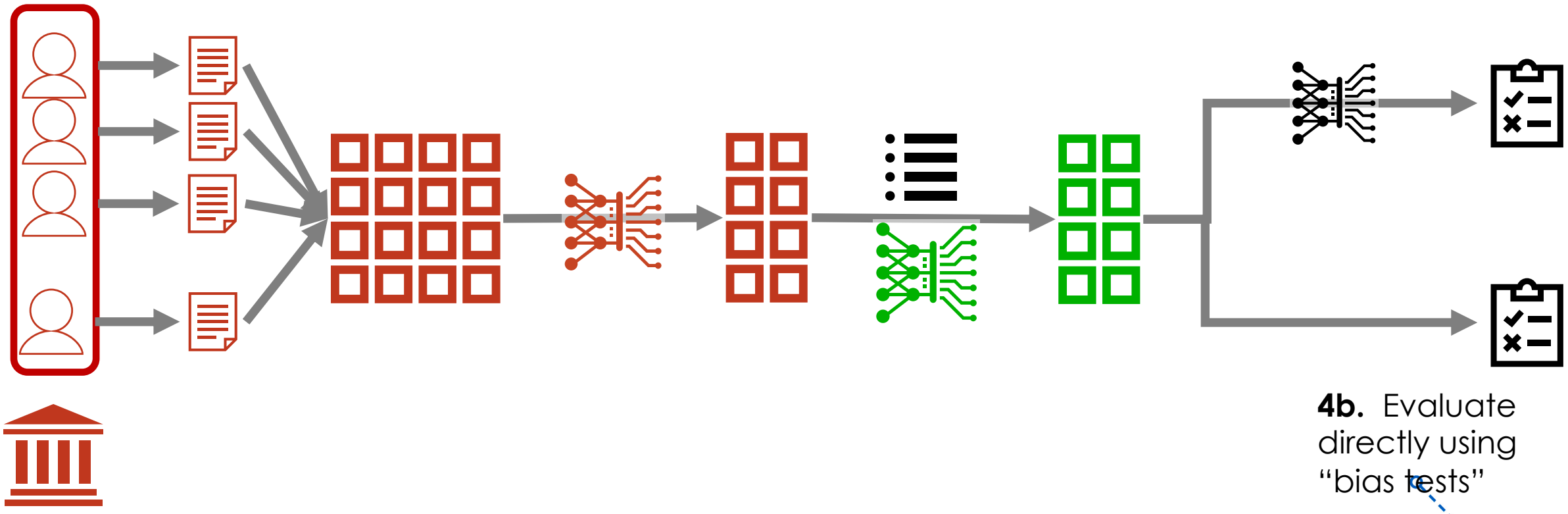
Today's Talk

- What are word embeddings?
- How do we define and measure "bias" (culture) using them?
- How have people tried to *remove* bias from word embeddings?
- Does any of this work or make any sense?
  - Eh. Let's discuss.



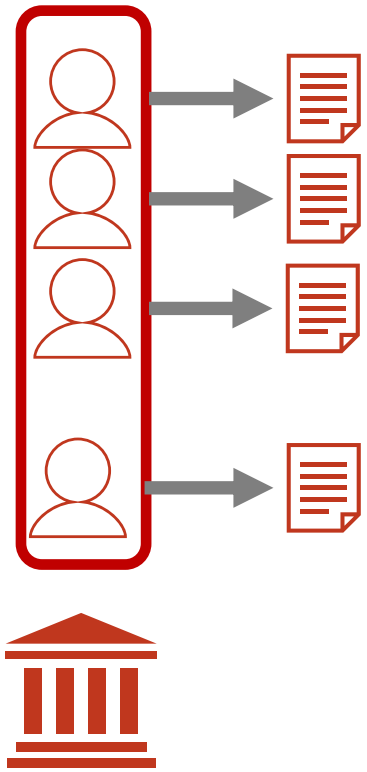
The Debiasing Pipeline

4a. Evaluate on downstream tasks (e.g. coreference resolution)



4b. Evaluate directly using "bias tests"

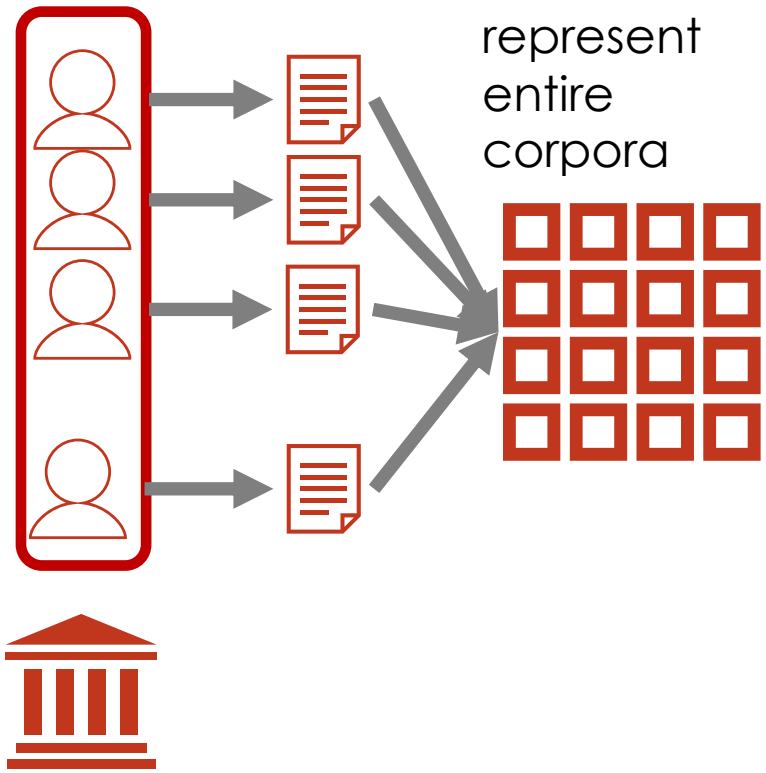
Step 1: Select a corpus

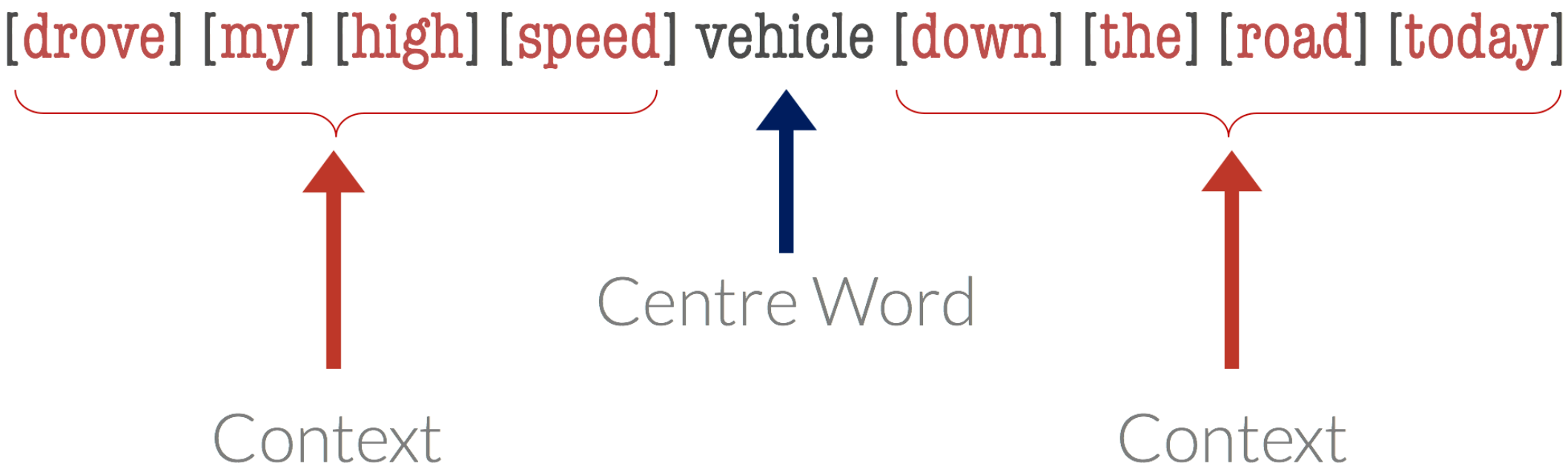


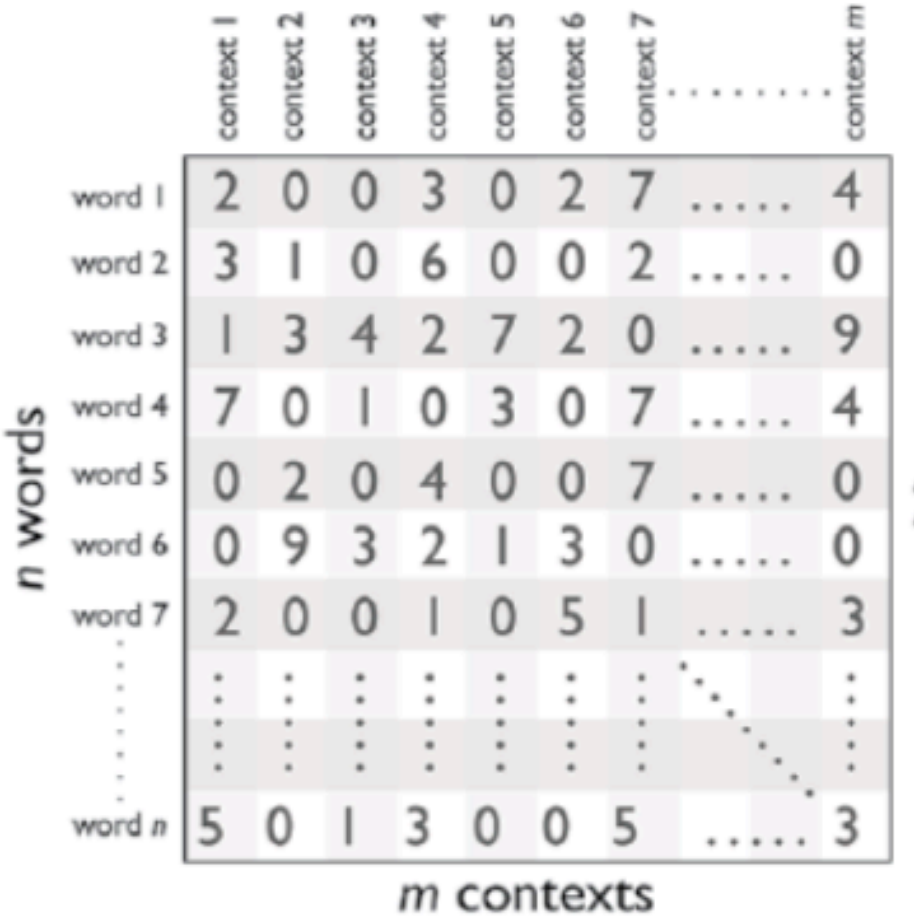
Typical Corpora

- Main requirement: corpus is large.
- All other concerns are secondary
  - Exceptions: **Historical analysis**
    - Garg et al. (PNAS) historical analysis of racial/ethnic stereotypes
    - Kozlowski et al. (ASR) historical analysis of associations w/ affluence
- Some common corpora
  - News articles
  - Common Crawl
  - Twitter
  - Wikipedia

2. Create co-occurence matrix is to represent entire corpora

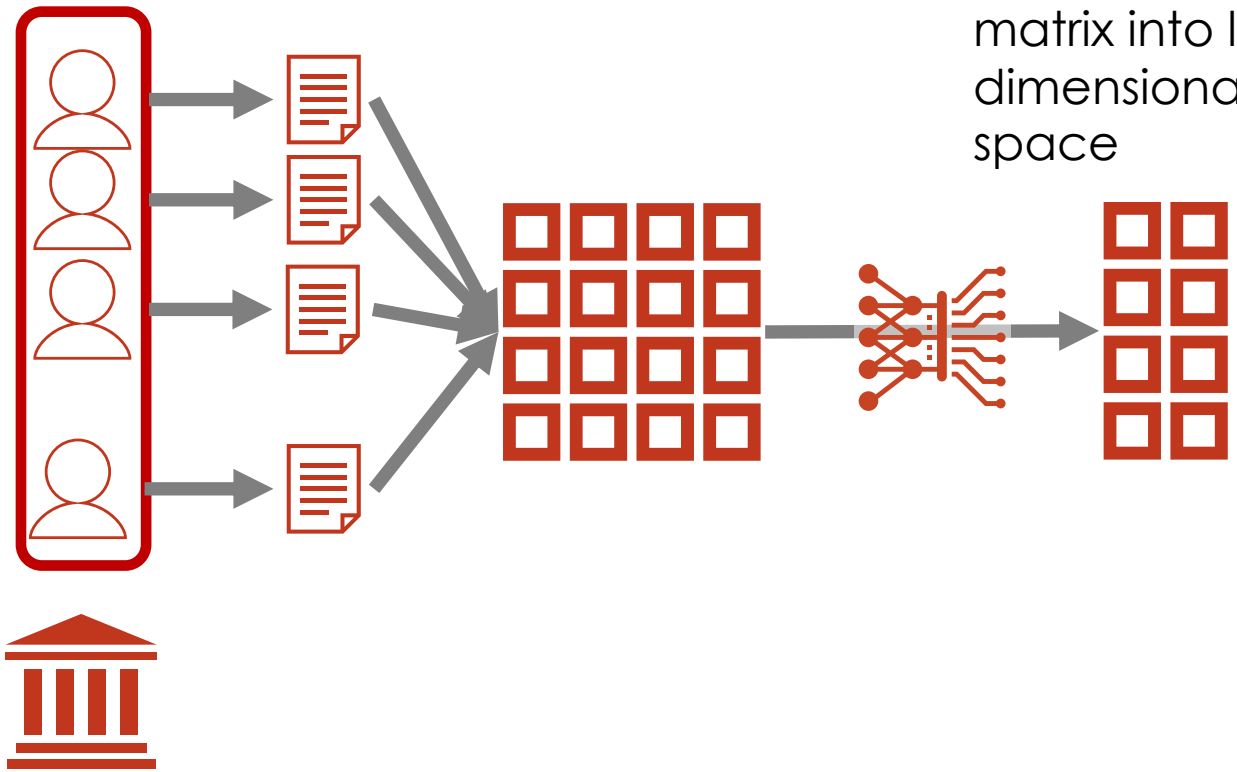


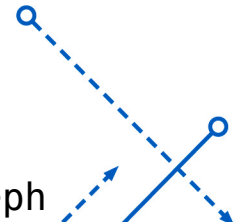

[drove] [my] [high] [speed] vehicle [down] [the] [road] [today]
Centre Word
Context
Context



@_kenny_joseph

3. Run algorithm (e.g. SGNS, GloVe) to embed words in co-occurance matrix into low dimensional space





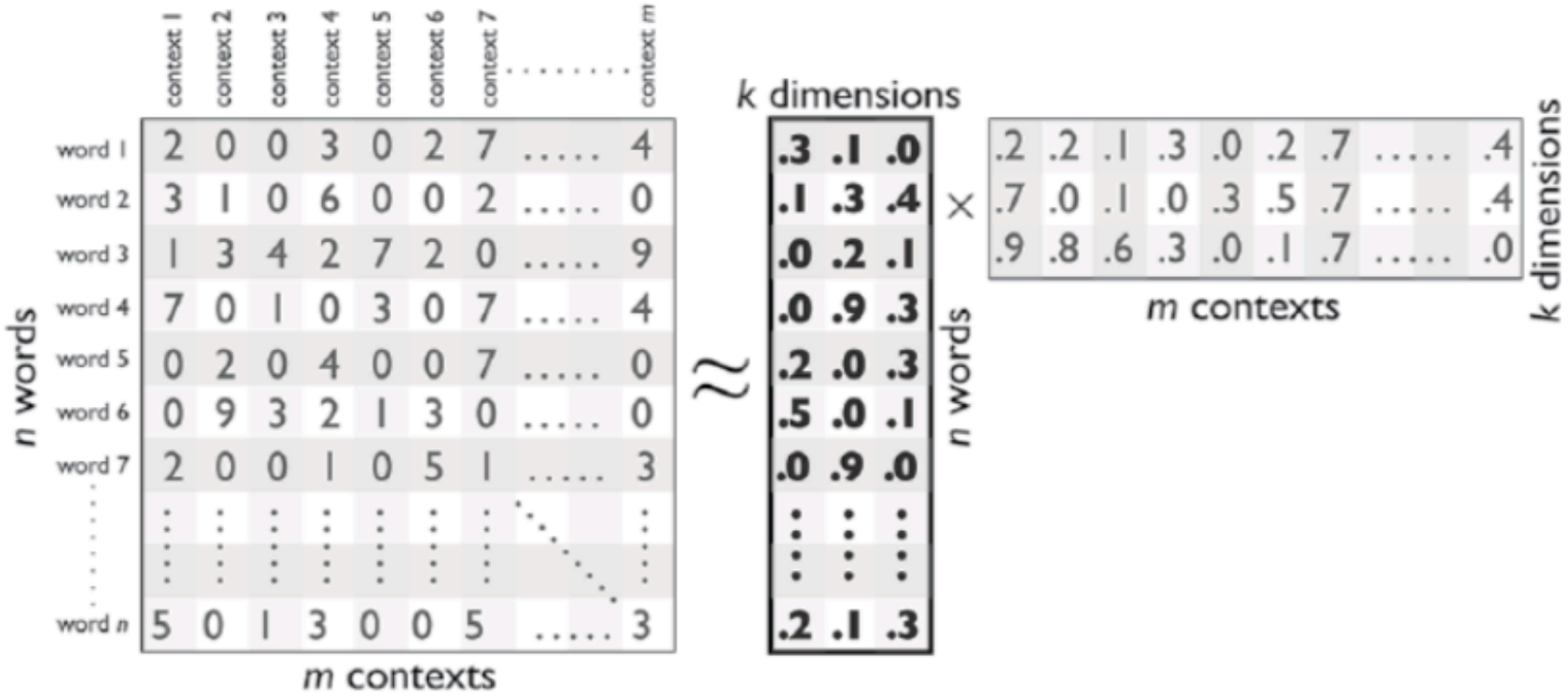
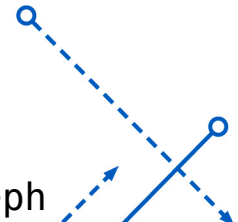
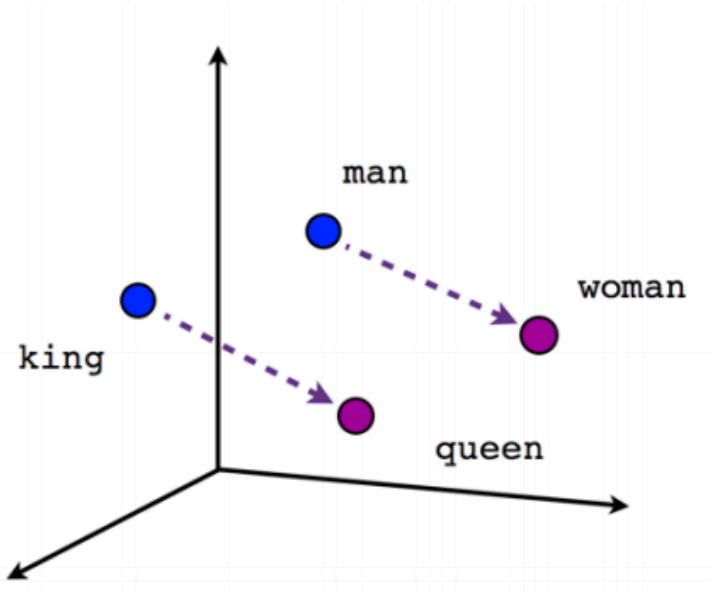


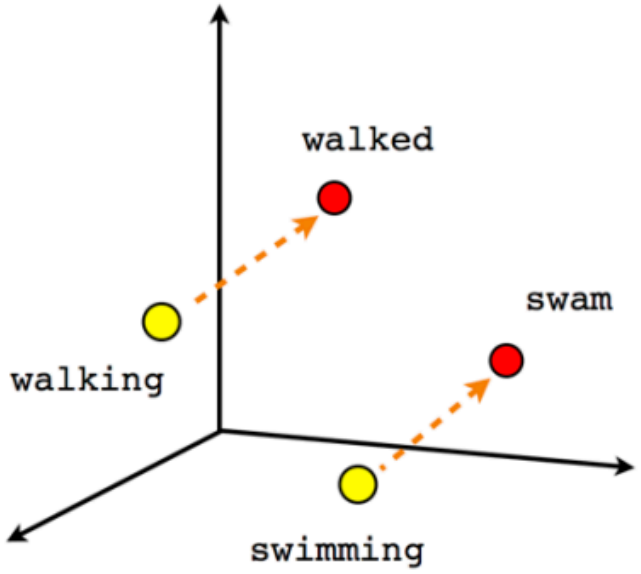
Figure 1. Schematic Illustration of the Descriptive Problem Neural Word Embeddings Solve—How to Represent All Words from a Corpus within a *k*-Dimensional Space That Best Preserves Distances between Words in Their Local Contexts

https://explosion.ai/demos/sense2vec

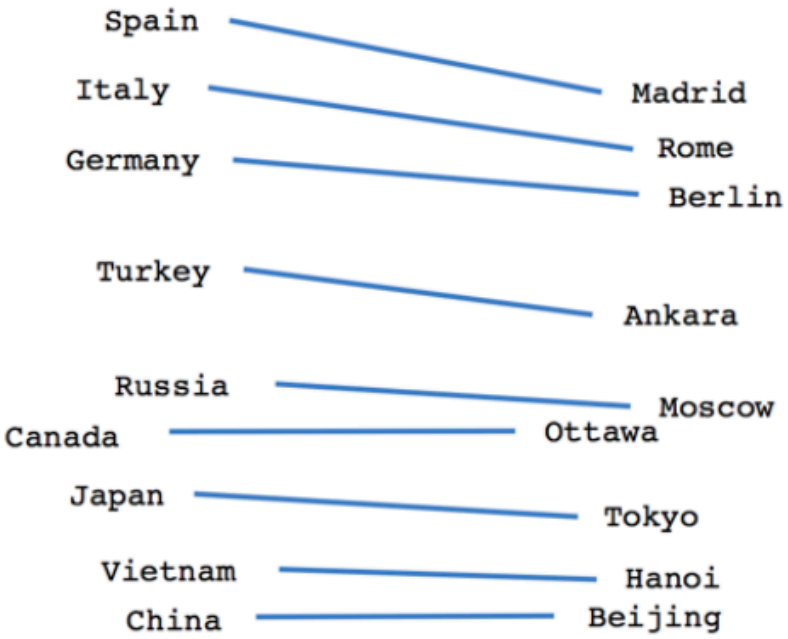




Male-Female

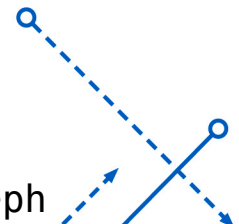


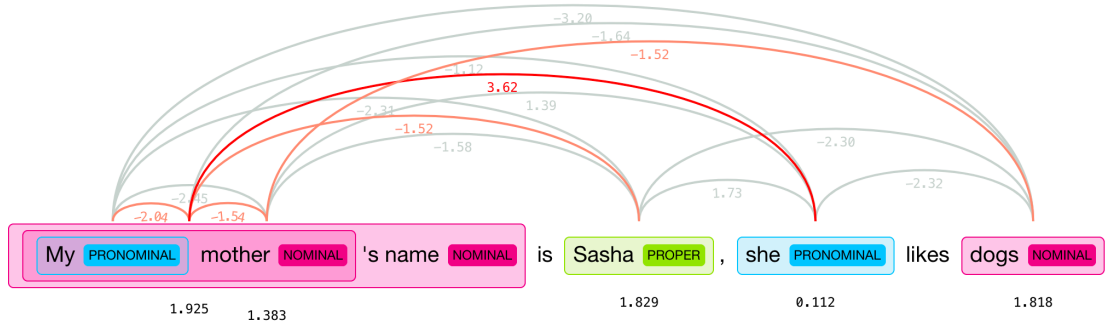
Verb tense

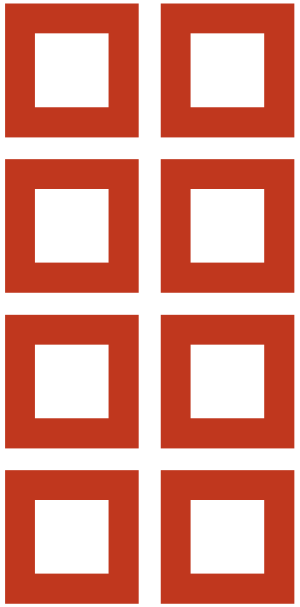


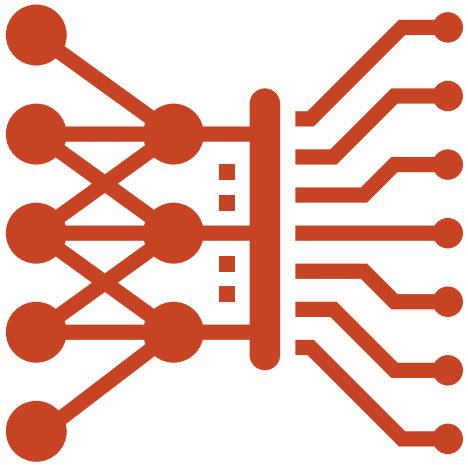
Country-Capital

Mikolov, T., Yih, W., & Zweig, G. (2013). Linguistic Regularities in Continuous Space Word Representations. *HLT-NAACL*, 746–751. Citeseer.









Kenneth Joseph

Website: kennyjoseph.github.io

Email: josephkena@gmail.com

Github: kennyjoseph

Phone: (716) 983-4115

Address:
Computer Science and Engineering Dept.
University at Buffalo
335 Davis Hall
Buffalo, NY, 14221

Academic Appointments

Asst. Professor	Computer Science	University of Buffalo	2018-
Postdoc	Network Science Institute	Northeastern University	2016-2018
Fellow	Institute for Quantitative Social Science	Harvard University	2016-2018
Fellow	Data Science for Social Good	University of Chicago	2015

Education

Ph.D.	Societal Computing	Carnegie Mellon University	2016
M.S.	Societal Computing	Carnegie Mellon University	2012
B.S.	Computer Science	University of Michigan-Ann Arbor	2010

Thesis: "Latent Cognitive Social Spaces: theory and methods for extracting prejudice from text".
Committee Members: Kathleen Carley (SI, CMU; Chair), Jason Hong (HCII, CMU), Lynn Smith-Lovin (Sociology, Duke), Eric Xing (ML/LTI, CMU)

Publications

Conference

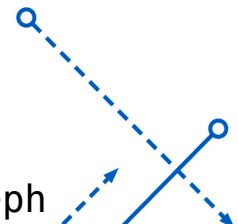
Joseph, K., Swire-Thompson, B., Masuga, H., Baum, M., & Lazer, D. (2019). Polarized, Together: Comparing Partisan Support for Trump's Tweets Using Survey and Platform-based Measures. *ICWSM*.

Joseph, K., Wihbey, J. (2019). Breaking News and Younger Twitter Users: Comparing Self-Reported Motivations to Online Behavior. *SMSociety*.

Robertson, R. E., Jiang, S., **Joseph, K.**, Friedland, L., Lazer, D., & Wilson, C. (2018). Auditing Partisan Audience Bias within Google Search. *Proceedings of the ACM on Human-Computer Interaction, 2(CSCW)*, 148. **Best Paper Honorable Mention**

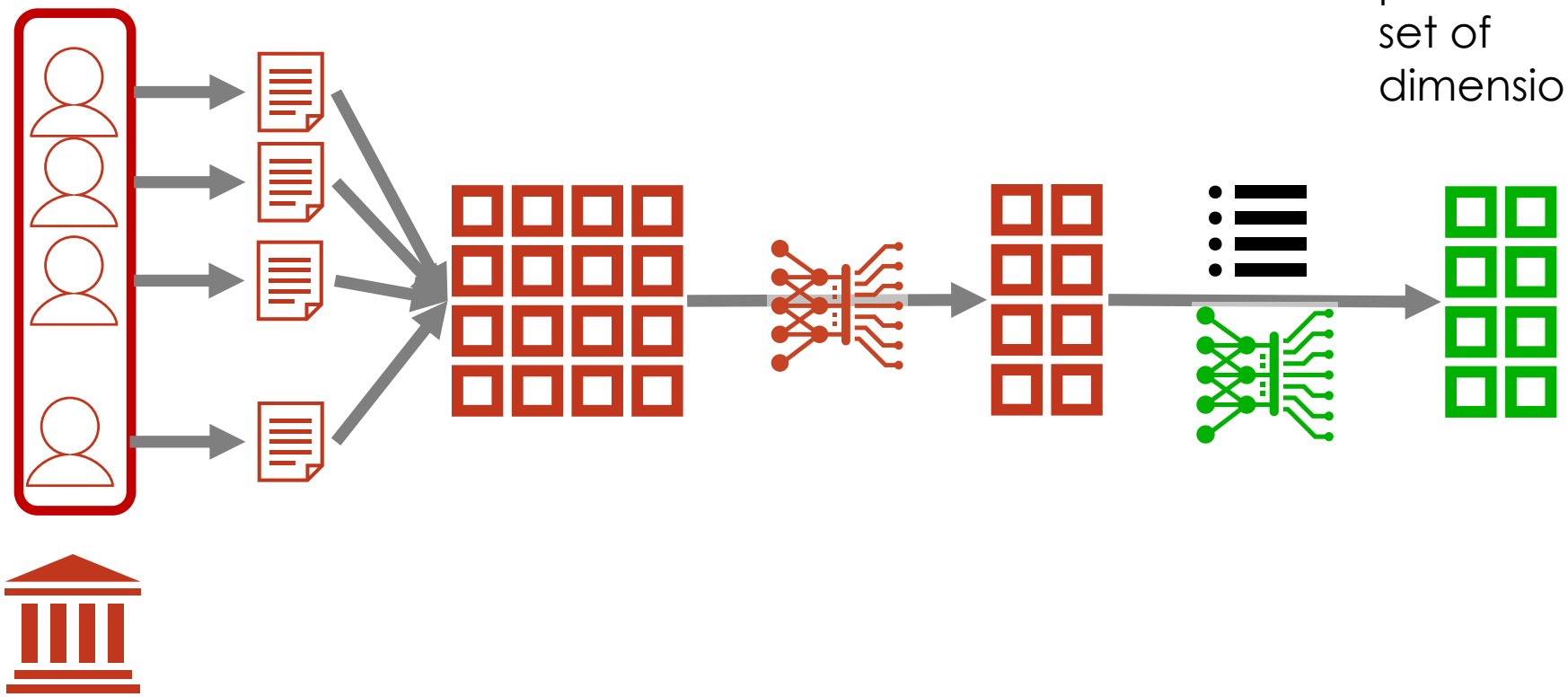
Joseph, K., Friedland, L., Tsur, O., Hobbs, W. & Lazer, D. (2017). Modeling Annotation Context to Improve Stance Classification. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* (pp. 1115-1124).

Hobbs, W., Friedland, L., **Joseph, K.**, Tsur, O., Wojcik, S. & Lazer, D. (2017). "Voters of the Year": 19 Voters Who Were Unintentional Election Poll Sensors on Twitter. *ICWSM*



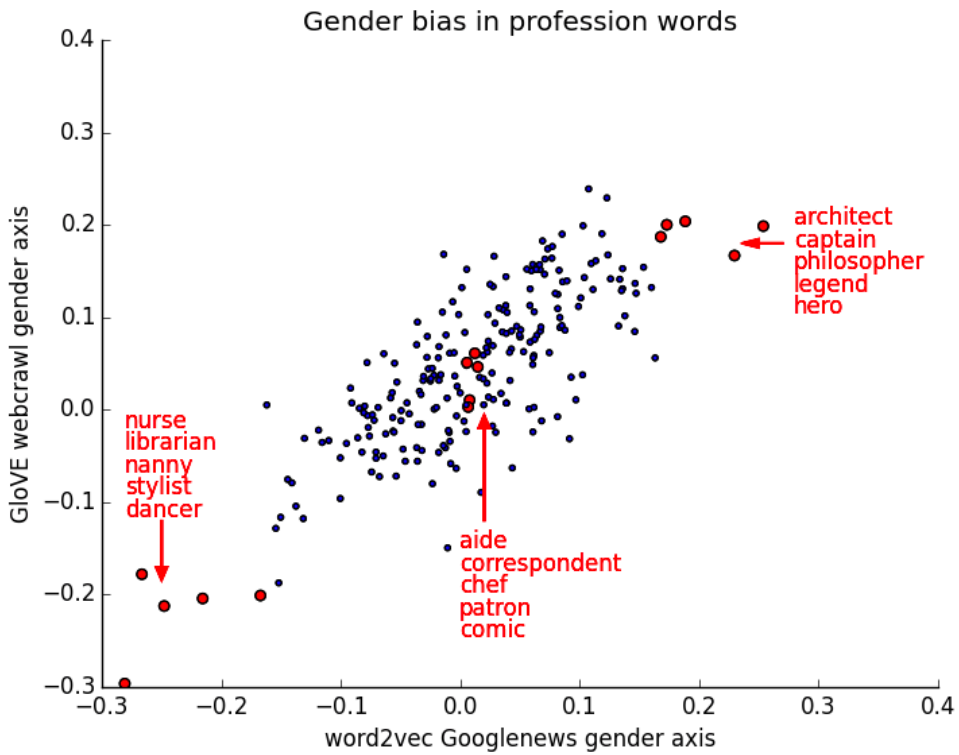
@_kenny_joseph

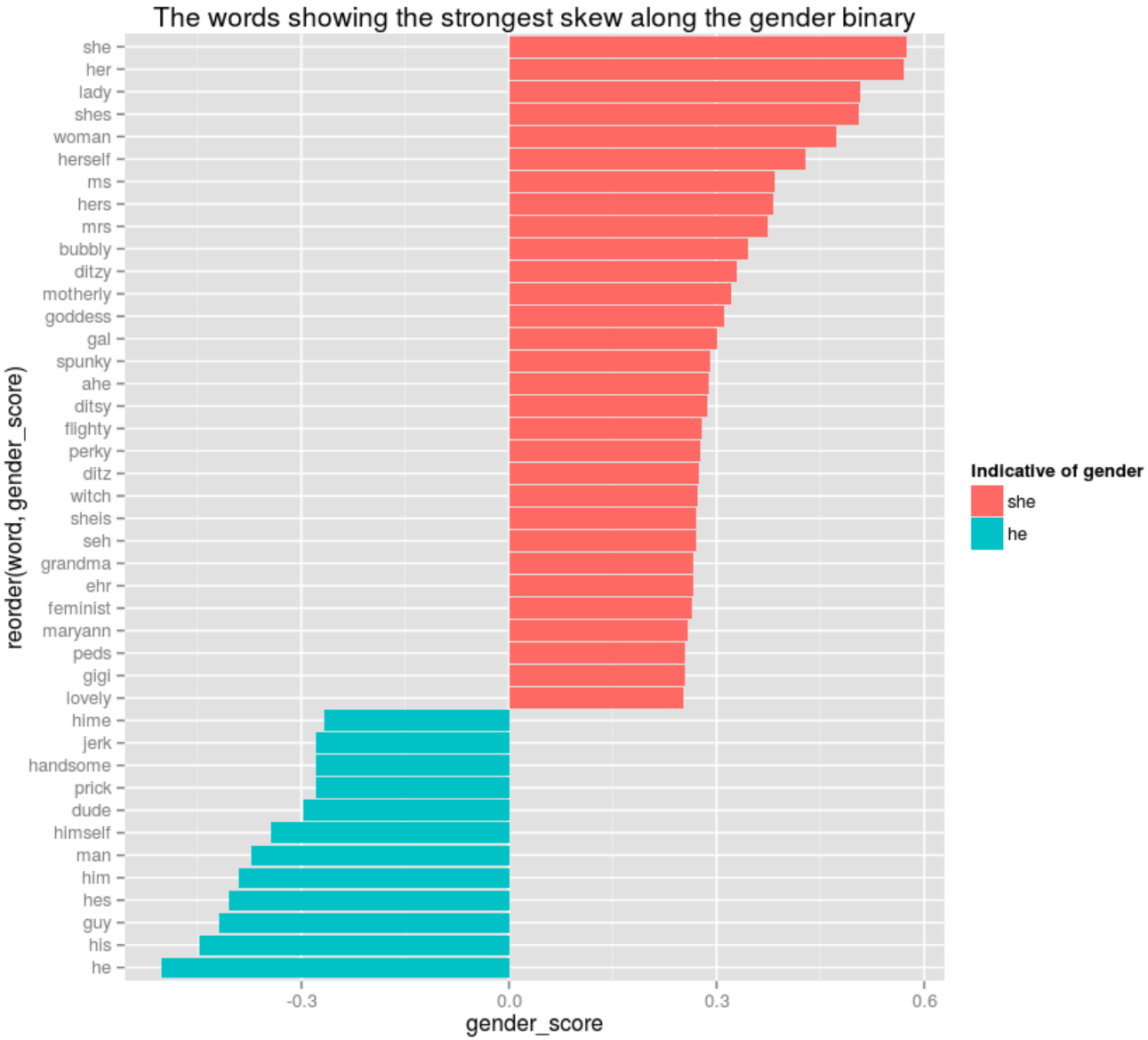
4. Run debiasing algorithm to remove bias along particular set of dimensions



How debiasing (typically) works

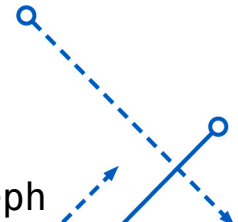
- **Step 1: Pick a dimension of meaning**
  - Usually gender, sometimes race
- **Step 2: Divide words into 3 camps:**
  - *Neutral* – Words that have no relation to gender ("millieu")
  - *Definitional* – Words that are gendered by definition ("sister")
  - *Biased* – Words that are gendered but shouldn't be (secretary)
- **Step 3: Identify "gender direction"**
  - Select words at either ends of a "gender spectrum"
  - Identify direction in the embedding using those words





http://bookworm.benschmidt.org/posts/2015-10-30-rejecting-the-gender-binary.html

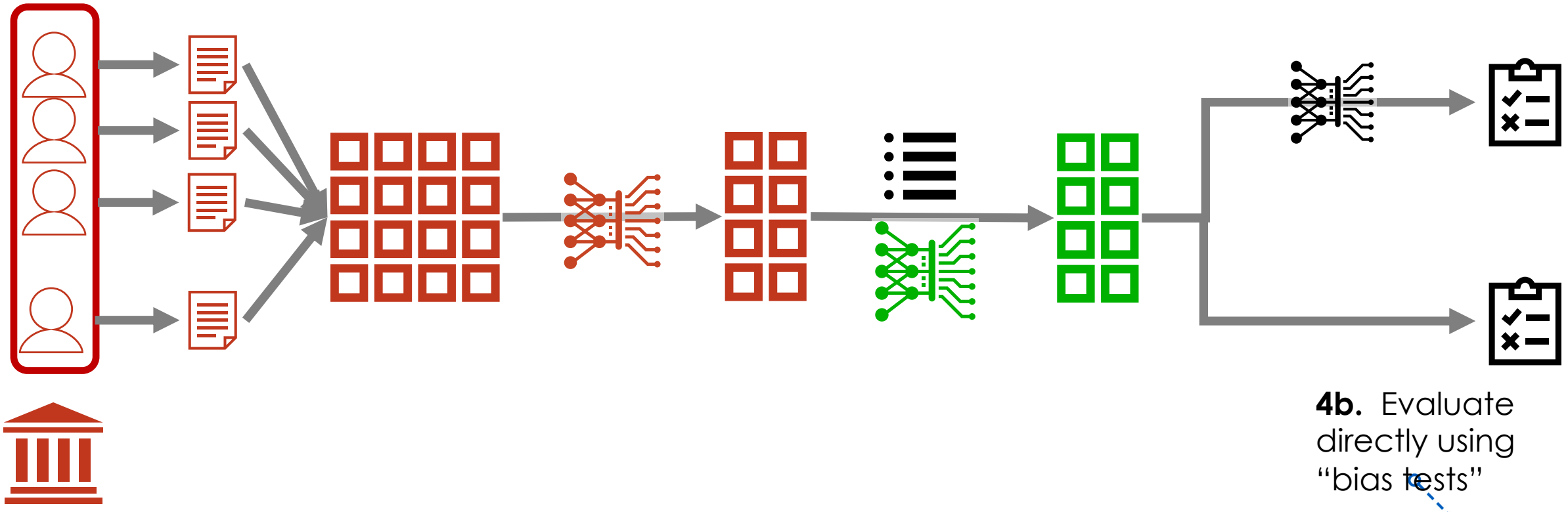
http://wordbias.umiacs.umd.edu/



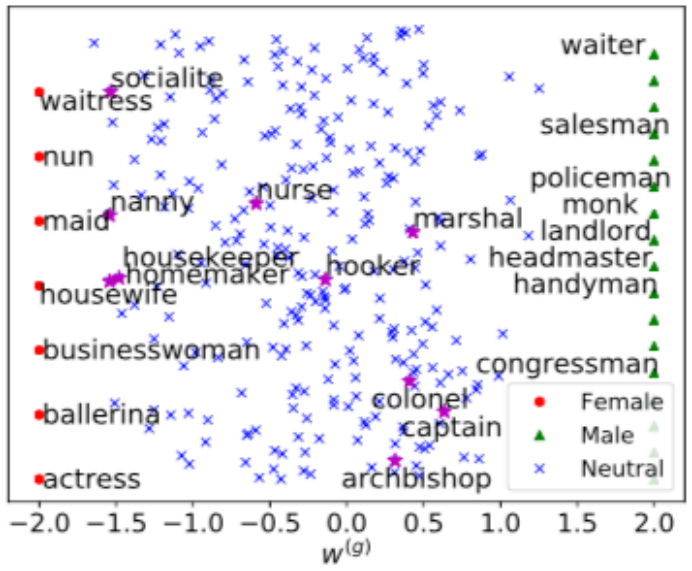
How debiasing (typically) works

- **Step 0:** Usually gender, sometimes race, or good/bad
- **Step 1:** Divide words into 3 camps:
  - *Neutral* – Words that have no relation to gender ("millieu")
  - *Definitional* – Words that are gendered by definition ("sister")
  - *Biased* – Words that are gendered but shouldn't be (secretary)
- **Step 2:** Identify "gender direction"
  - Select words at either ends of a "gender spectrum"
  - Identify direction in the embedding using those words
- **Step 3:** Try to remove gender direction from biased words, keep it for definitional and neutral words

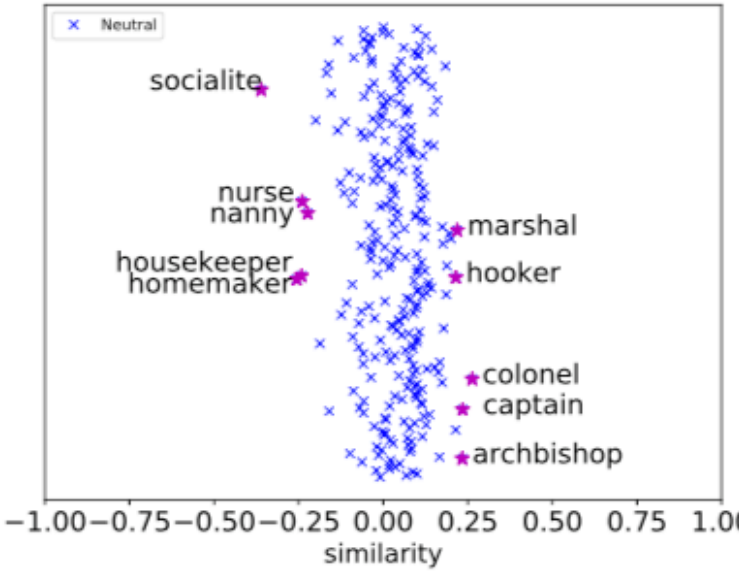
4a. Evaluate on downstream tasks (e.g. coreference resolution)



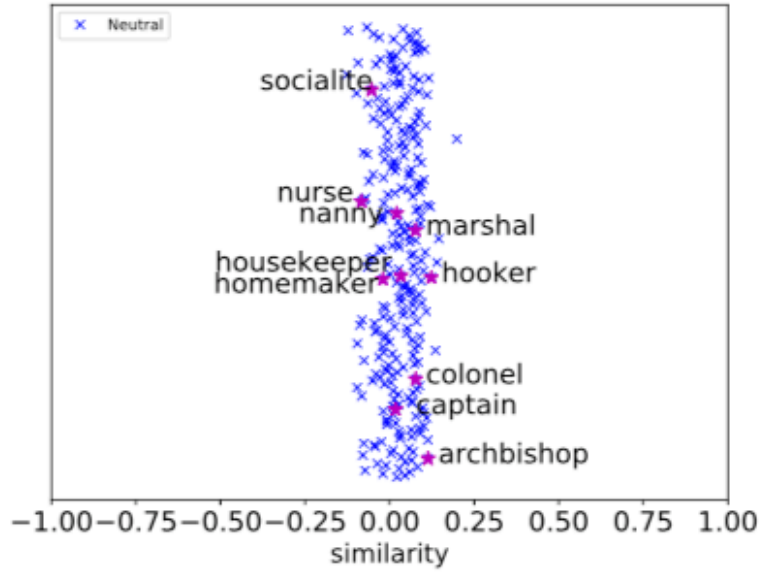
4b. Evaluate directly using "bias tests"



(a) $w^{(g)}$ dimension for all the professions



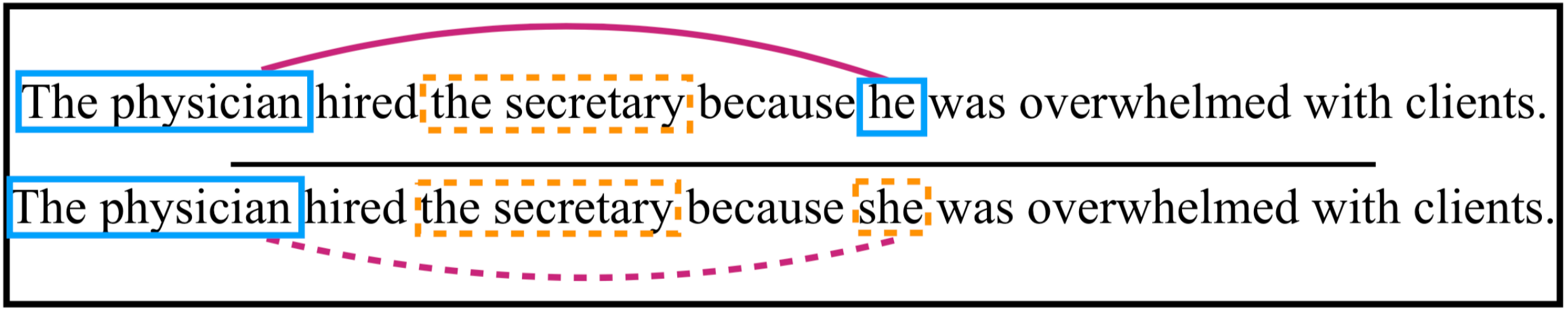
(b) Gender-neutral profession words projected to gender direction in GloVe



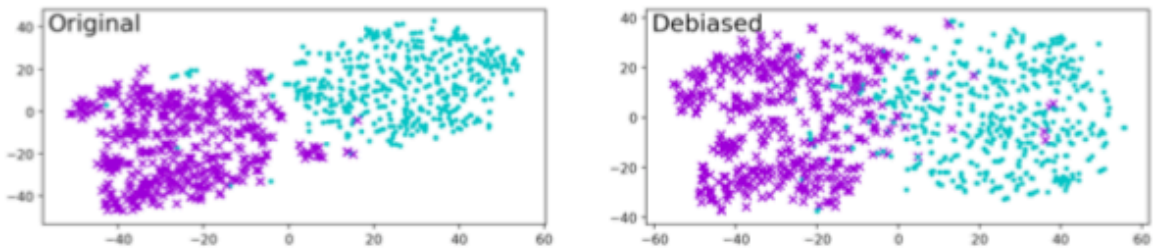
(c) Gender-neutral profession words projected to gender direction in GN-GloVe

Figure 1: Cosine similarity between the gender direction and the embeddings of gender-neutral words. In each figure, negative values represent a bias towards female, otherwise male.

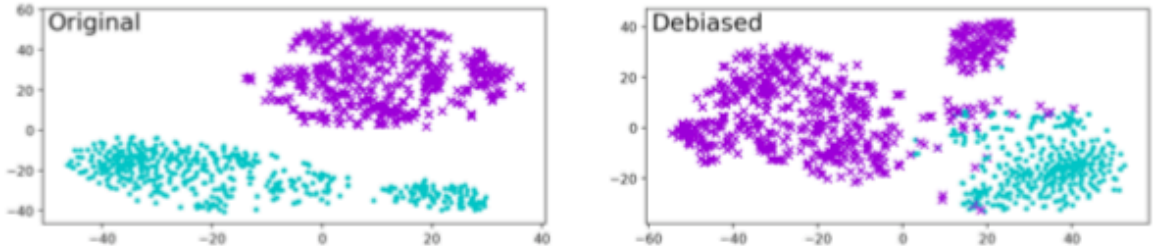
@_kenny_joseph


The physician hired the secretary because he was overwhelmed with clients.
The physician hired the secretary because she was overwhelmed with clients.

Debiasing – does it work?



(a) Clusteing for HARD-DEBIASED embedding, before (left hand-side) and after (right hand-side) debiasing.

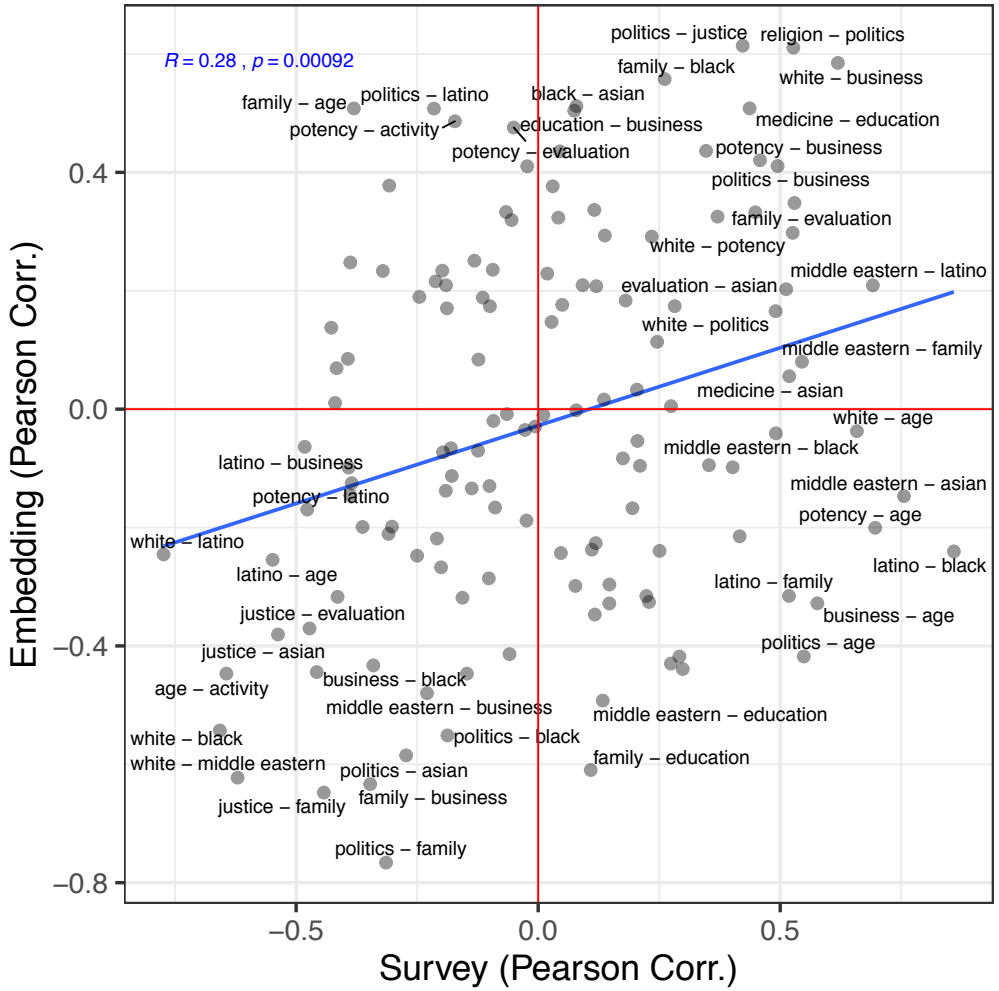


(b) Clusteing for GN-GLOVE embedding, before (left hand-side) and after (right hand-side) debiasing.

- Does it really make sense to treat bias as existing along a "direction"?
  - I think so, actually
  - **But this doesn't mean that it makes sense to *debias* along a direction**

Gonen, H., & Goldberg, Y. (2019). Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them. *ArXiv Preprint ArXiv:1903.03862*.

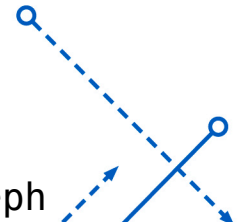
Debiasing



- What dimension are we measuring with that direction?
  - Its not enough to study gender/race/"goodness"
  - E.g. consider secretary
- Indeed, prejudice exists on many other dimensions, some of which are hard to measure
- **And we want to debias along all of them!**

Identifying/evaluating bias and debiasing

- Lots of questions here! Let's discuss!
  1. How do we know to differentiate between "biased" and "definitional"?
  2. What dimension should we be studying?

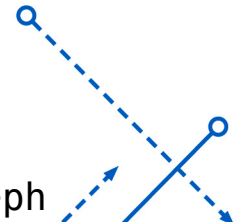


More questions!

- More discussion!
  - How much accuracy/performance should we be willing to lose to debias our embeddings? What does this tradeoff look like in practice?
  - Why is "hiding the bias" bad?
  - What would it mean to remove gender bias from the English language all together? What about other languages?

Recap

1. Does it make sense to treat bias as a "direction"?
   … I think so, but not for "debiasing"
2. What direction should we be studying?
   … Not just gender
3. How do we know to differentiate between "biased" and "definitional"?
   … this is bogus



Summary

- Bias in word embeddings is good or bad, depending on who you ask
- To make progress, we need to synthesize ideas
- Lots of open research questions here!
- Some fun things to think about
  - Is debiasing possible?
  - What does language look like without gender?