

# Taking uncertainty seriously

A Bayesian approach to word embedding bias estimation

Alicja Dobrzeniecka & Rafal Urbaniak

(LoPSE research group, University of Gdansk, Vrij Universiteit Amsterdam)

Boston, April Fools' Day

# Presentation plan

- Bias in word embeddings
- WEAT and MAC methods
- Methodological problems
- Limitations of pre-averaging in bias detection methods
- Accounting for uncertainty with Bayesian approach

# Word2vec

## Question

How to sensibly represent words with numbers?

# Word2vec

## Question

How to sensibly represent words with numbers?

## One-hot encoding

Well, you could use 30k binary vectors with a slot for each lexical unit. . .

# Word2vec

## Question

How to sensibly represent words with numbers?

## One-hot encoding

Well, you could use 30k binary vectors with a slot for each lexical unit. . .  
. . . . . but this would be inefficient and wouldn't capture any relations  
between words.

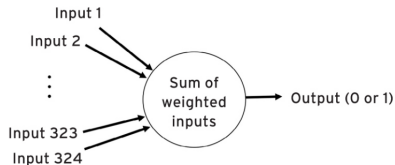
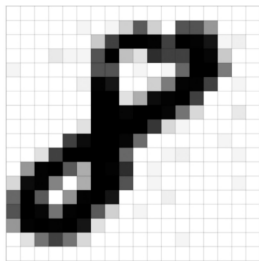


Illustration: M. Mitchell

## Rosenblatt's perceptron

- Inputs (pixel intensities) with weights
- Nodes with activation levels from 0-1
- (Perhaps) 0-1 output based on a threshold

# Word2vec

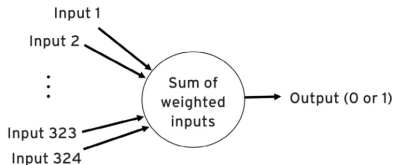
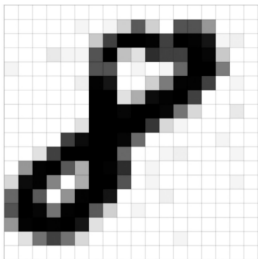


Illustration: M. Mitchell

## Learning

- Start with random weights
- Test on a case:
  - If right, don't change weights.
  - If wrong, change weights a bit, with focus on the ones more responsible for the judgment:

$$w_j \leftarrow w_j = \underbrace{\eta}_{\text{learning rate}} \left( \underbrace{t}_{\text{correct output}} - \underbrace{y}_{\text{actual output}} \right) \underbrace{x_j}_{\text{actual input}}$$

# Word2vec

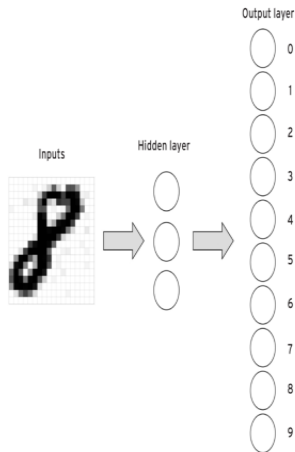


Illustration: M. Mitchell

- Each hidden unit takes a weighted sum of 324 inputs and passes on its activation level as input to outer layer units.
- Activation levels of outer layers are interpreted as network's levels of confidence in a classification problem.
- Learning: back-propagation (gradient descent: approximate the direction of steepest descent in the error surface w.r.t to weights, modify accordingly).



# Word2vec

## Distributional semantics

- "You shall know a word by the company it keeps" (John Firth, 1957)
- "the degree of semantic similarity between two linguistic expressions  $A$  and  $B$  is a function of the similarity of the linguistic contexts in which  $A$  and  $B$  can appear." (A. Lenci, 2008)

# Word2vec

## Distributional semantics

- "You shall know a word by the company it keeps" (John Firth, 1957)
- "the degree of semantic similarity between two linguistic expressions  $A$  and  $B$  is a function of the similarity of the linguistic contexts in which  $A$  and  $B$  can appear." (A. Lenci, 2008)

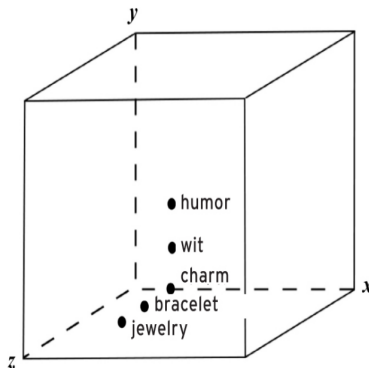


Illustration: M. Mitchell

# Word2vec

## Google and Mikolov

*Efficient Estimation of Word Representation in Vector Space*, 2013

Let's train a neural network and use vectors of weights!

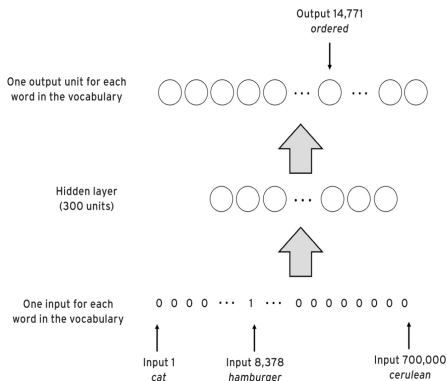


Illustration: M. Mitchell

# Word2vec

## Google and Mikolov

*Efficient Estimation of Word Representation in Vector Space*, 2013

Let's train a neural network and use vectors of weights!

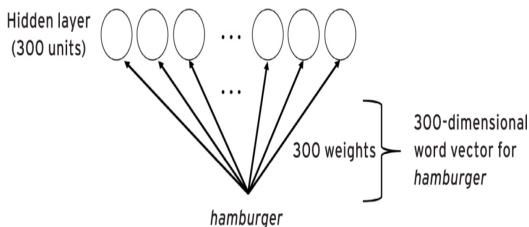


Illustration: M. Mitchell

# Word2vec

| word  | 1     | 2     | 3     | 4     | ... |
|-------|-------|-------|-------|-------|-----|
| woman | 0.456 | 0.267 | 0.675 | 0.131 | ... |
| man   | 0.451 | 0.897 | 0.472 | 0.088 | ... |

## Question

How is this supposed to capture semantic relations?

# Word2vec

| word  | 1     | 2     | 3     | 4     | ... |
|-------|-------|-------|-------|-------|-----|
| woman | 0.456 | 0.267 | 0.675 | 0.131 | ... |
| man   | 0.451 | 0.897 | 0.472 | 0.088 | ... |

## Question

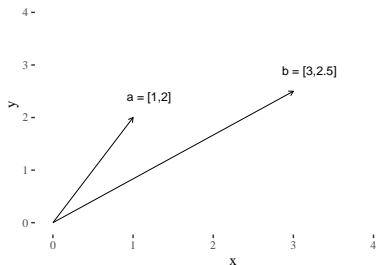
How is this supposed to capture semantic relations?

## General idea

Similarity in vector direction.

# Cosine similarity

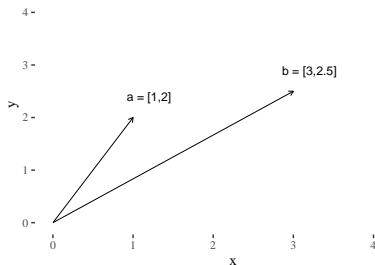
## Vectors



$$a = [1, 2]$$

$$b = [3, 2.5]$$

# Cosine similarity



## Vectors

$$a = [1, 2]$$

$$b = [3, 2.5]$$

## Dot product

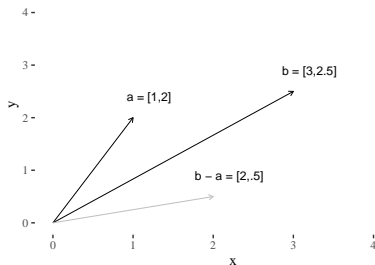
$$a \cdot b = a_1 b_1 + a_2 b_2$$

$$a \cdot a = a_1^2 + a_2^2$$

$$\|a\| = \sqrt{(a \cdot a)}$$



# Cosine similarity



## Vectors

$$a = [1, 2]$$

$$b = [3, 2.5]$$

## Dot product

$$a \cdot b = a_1 b_1 + a_2 b_2$$

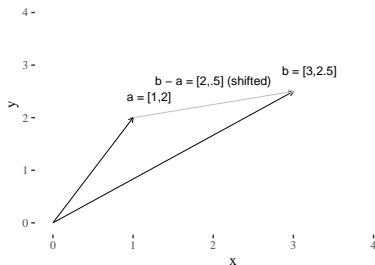
$$a \cdot a = a_1^2 + a_2^2$$

$$\|a\| = \sqrt{(a \cdot a)}$$

## Vector difference

$$b - a = [b_1 - a_1, b_2 - a_2]$$

# Cosine similarity



## Vectors

$$a = [1, 2]$$

$$b = [3, 2.5]$$

## Dot product

$$a \cdot b = a_1 b_1 + a_2 b_2$$

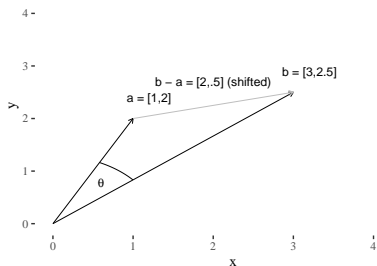
$$a \cdot a = a_1^2 + a_2^2$$

$$\|a\| = \sqrt{(a \cdot a)}$$

## Vector difference

$$b - a = [b_1 - a_1, b_2 - a_2]$$

# Cosine similarity



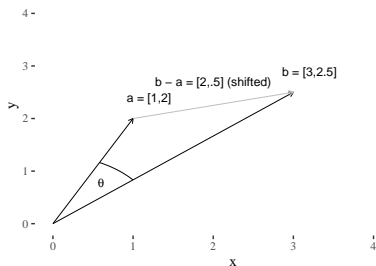
## Angle

$$\|b - a\|^2 = \|b\|^2 + \|a\|^2 - 2\|b\|\|a\| \cos \theta$$

$$b \cdot a = \|b\|\|a\| \cos \theta$$

$$\cos \theta = \frac{b \cdot a}{\|b\|\|a\|}$$

# Cosine similarity



## Angle

$$\|b - a\|^2 = \|b\|^2 + \|a\|^2 - 2\|b\|\|a\| \cos \theta$$

$$b \cdot a = \|b\|\|a\| \cos \theta$$

$$\cos \theta = \frac{b \cdot a}{\|b\|\|a\|}$$

## Orthogonality

$$\cos(90^\circ) = 0$$

$$\frac{b \cdot a}{\|b\|\|a\|} = 0$$

$$b \cdot a = 0$$

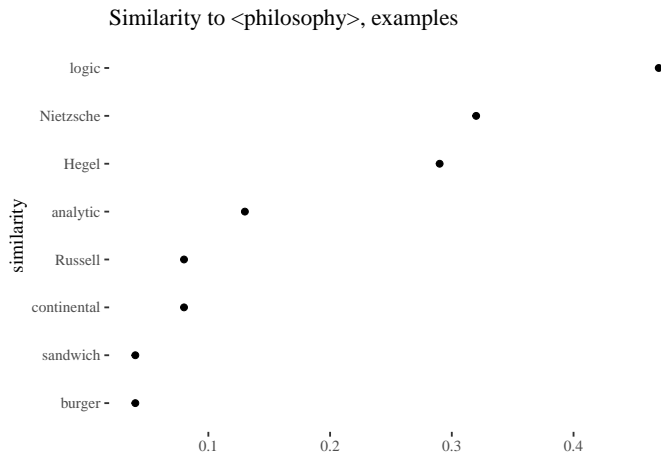
# Cosine similarity & distance

$$\text{cosineSimilarity}(A, B) = \frac{A \cdot B}{\|A\| \|B\|} \quad (\text{Sim})$$

$$\text{cosineDistance}(A, B) = 1 - \text{cosineSimilarity}(A, B) \quad (\text{Distance})$$

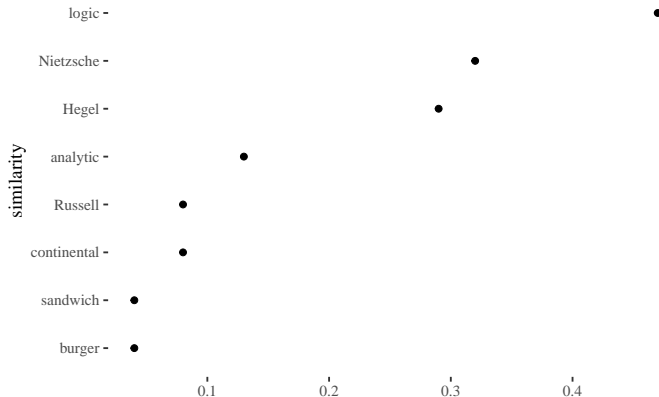
- Naive interpretation: proximity corresponds to semantic similarity
- Geometric interpretation: direction  $\cos \in (-1, 1)$ 
  - 1: maximally similar
  - -1: opposites
  - 0: dissimilar
- $\text{cosineDistance} \in (0, 2)$

# Cosine similarity & distance



# Cosine similarity & distance

Similarity to <philosophy>, examples



The only “jobs” in top-tens

- Man: robber (.55)
- Woman: policewoman (.6)

# Cosine-based measures of bias

## The worry

Word embeddings can learn implicit harmful biases



# Cosine-based measures of bias

## The worry

Word embeddings can learn implicit harmful biases

## The basic intuition

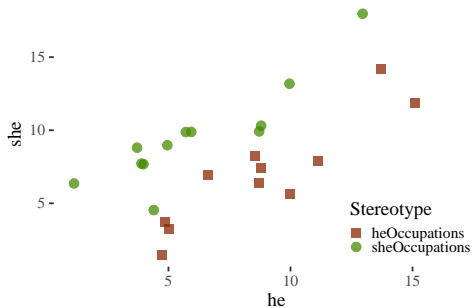
Stereotypically connected words are cosine-close

# Cosine-based measures of bias

## A visual example

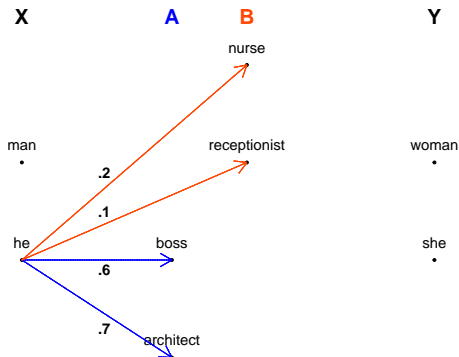
- “feminine” occupations: “homemaker,” “nurse,” “receptionist,” “librarian,” etc.
- “masculine” occupations: “maestro,” “captain,” “architect,” “boss,” etc.

GloVe on Wikipedia 2014 and Gigaword 5th ed.



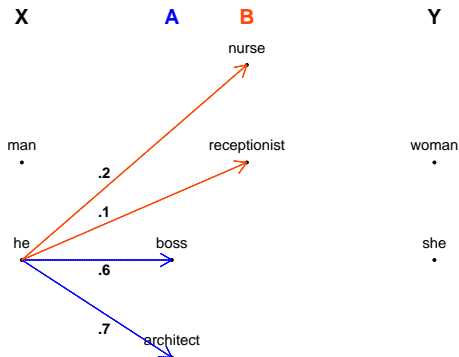
# Cosine-based measures of bias

## Example: Word Embedding Association Test (WEAT)



# Cosine-based measures of bias

## Example: Word Embedding Association Test (WEAT)



- $s_1 = s(\text{he}, A, B) = \frac{.6+.7}{2} - \frac{.2+.1}{2} = .65 - .15 = .5$
- $s_2 = s(\text{man}, A, B) = .3,$   
 $s_3 = s(\text{woman}, A, B) = -.6, s_4 = s(\text{she}, A, B) = -.3$

$$\text{WEAT}(A, B) = \frac{(s_1 + s_2)/2 - (s_3 + s_4)/2}{sd(\{s_1, s_2, s_3, s_4\})} \approx 1.93$$

# Cosine-based measures of bias

## Example: Word Embedding Association Test (WEAT)

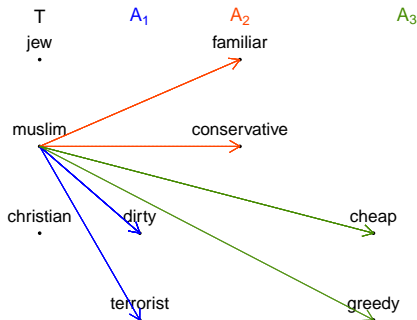
$$s(t, A, B) = \frac{\sum_{a \in A} f(t, a)}{|A|} - \frac{\sum_{b \in B} f(t, b)}{|B|}$$
$$WEAT(A, B) = \frac{\mu(\{s(x, A, B)\}_{x \in X}) - \mu(\{s(y, A, B)\}_{y \in Y})}{\sigma(\{s(w, A, B)\}_{w \in X \cup Y})}$$

- $t$  is a term,  $A, B$  are sets of stereotype attribute words,  $X, Y$  are protected group words
- For instance,  $X$  might be a set of male names,  $Y$  a set of female names,  $A$  might contain stereotypically male-related career words, and  $B$  stereotypically female-related family words
- $s$ -values are used as datapoints in statistical significance tests

(Caliskan, Bryson, & Narayanan, 2017) with extensions in (Lauscher & Glavas, 2019) and applications in (Garg, Schiebinger, Jurafsky, & Zou, 2018)

# Cosine-based measures of bias

Our main target: Mean Average Cosine Similarity (MAC)



$$s_1 = s(\text{muslim}, A_1) = \frac{\cos(\text{muslim}, \text{dirty}) + \cos(\text{muslim}, \text{terrorist})}{2}$$

$$s_2 = s(\text{muslim}, A_2) = \frac{\cos(\text{muslim}, \text{familiar}) + \cos(\text{muslim}, \text{conservative})}{2}$$

$\vdots$

$$\text{MAC}(T, A) = \text{mean}(\{s_i | i \in 1, \dots, k\})$$

# Cosine-based measures of bias

Our main target: Mean Average Cosine Similarity (MAC)

$$S(t_i, A_j) = \frac{1}{|A_j|} \sum_{a \in A_j} \cos(t, a)$$

$$MAC(T, A) = \frac{1}{|T||A|} \sum_{t_i \in T} \sum_{A_j \in A} S(t_i, A_j)$$

- $T = \{t_1, \dots, t_k\}$  is a class of protected words
- each  $A_j \in A$  is a set of attributes stereotypically associated with a protected word
- The t-tests they employ are run on average cosines used to calculate MAC

(Manzini, Lim, Tsvetkov, & Black, 2019)

# Cosine-based measures of bias

Our main target: Mean Average Cosine Similarity (MAC)

Table 2: A few rows from the religion dataset

| protectedWord | wordToCompare | cosineDistance | cosineSimilarity |
|---------------|---------------|----------------|------------------|
| jew           | greedy        | 0.6947042      | 0.3052958        |
| rabbi         | greedy        | 1.0306175      | -0.0306175       |
| rabbi         | conservative  | 0.7175887      | 0.2824113        |
| christian     | uneducated    | 0.5081939      | 0.4918061        |
| christianity  | cheap         | 1.2816164      | -0.2816164       |
| muslim        | terrorist     | 0.2726106      | 0.7273894        |



# Cosine-based measures of bias

## General challenges

- Gender-direction: insufficient indicator of bias (Gonen & Goldberg, 2019)
- Use of analogies: unreliable (Nissim, Noord, & Goot, 2020)
- High sensitivity to irrelevant factors (Zhang, Sneyd, & Stevenson, 2020)

# Some methodological problems

## Word list choice is unprincipled

- We run with it for comparison.

# Some methodological problems

## Word list choice is unprincipled

- We run with it for comparison.

## No design considerations to sample size

- **Ethayarajh (2020)** criticizes WEAT uses Bernstein bounds and argues that we would need a bias specific dataset of size at least 11903 to claim that the system is biased (three times larger than WinoBias).

# Some methodological problems

## Word list choice is unprincipled

- We run with it for comparison.

## No design considerations to sample size

- Ethayarajh (2020) criticizes WEAT uses Bernstein bounds and argues that we would need a bias specific dataset of size at least 11903 to claim that the system is biased (three times larger than WinoBias).
- We show progress can be made with more sensitive Bayesian methods.

## The form of the definition is suspicious

- Ethayarajh, Duvenaud, & Hirst (2019) show that if there are two target words only WEAT is always maximal in one direction.

# Some methodological problems

## Word list choice is unprincipled

- We run with it for comparison.

## No design considerations to sample size

- Ethayarajh (2020) criticizes WEAT uses Bernstein bounds and argues that we would need a bias specific dataset of size at least 11903 to claim that the system is biased (three times larger than WinoBias).
- We show progress can be made with more sensitive Bayesian methods.

## The form of the definition is suspicious

- Ethayarajh, Duvenaud, & Hirst (2019) show that if there are two target words only WEAT is always maximal in one direction.
- We show the problem runs deeper and stems from pre-averaging, and we statistically gauge the uncertainty that arises from raw sample sizes.

# Some methodological problems

## No word class distinction and no control group

We make the subclasses clear, add human neutral predicates and neutral predicates for control. We used L2-Reddit corpus and GoogleNews (we present the results for Reddit for brevity).

Table 3: Rows from extended religion dataset.

| protectedWord | wordToCompare | wordClass | cosineDistance | cosineSimilarity | connection |
|---------------|---------------|-----------|----------------|------------------|------------|
| torah         | hairry        | jewish    | 1.170          | -0.170           | associated |
| christian     | dirty         | muslim    | 0.949          | 0.051            | different  |
| judaism       | cheap         | jewish    | 1.232          | -0.232           | associated |
| christianity  | familial      | christian | 0.645          | 0.355            | associated |
| mosque        | approve       | neutral   | 0.995          | 0.005            | none       |
| imam          | carry         | human     | 0.993          | 0.007            | human      |
| mosque        | merging       | neutral   | 0.868          | 0.132            | none       |
| muslim        | nationalized  | neutral   | 0.870          | 0.130            | none       |

# Some methodological problems

## Our neutral words (examples, full list size = ADD)

liquor, pow, ballpark, glitchy, billy, dallas, rip, called, outlooks, viet, floater, rattlesnake, exports, peruvian, recursion, shortfall, corrected, amicable, solutions, diagnostic, patently, flops, approx, percents, lox, catapults, hamburger, engulfed, households, north, snubbed, playtest

## Our human-related words (examples, full list size = ADD)

switch, studio, stick, soup, sometimes, signal, prior, plant, photo, path, park, near, menu, latter, grass, clock, wear, walk, visitor, toy, tissue, throw, talk, speak, sleep, eye, enjoy, blogger, character, candidate, breakfast, supper, dinner, eat, drink, carry, run, cast, ask, awake, ear, nose, lunch

# Some methodological problems

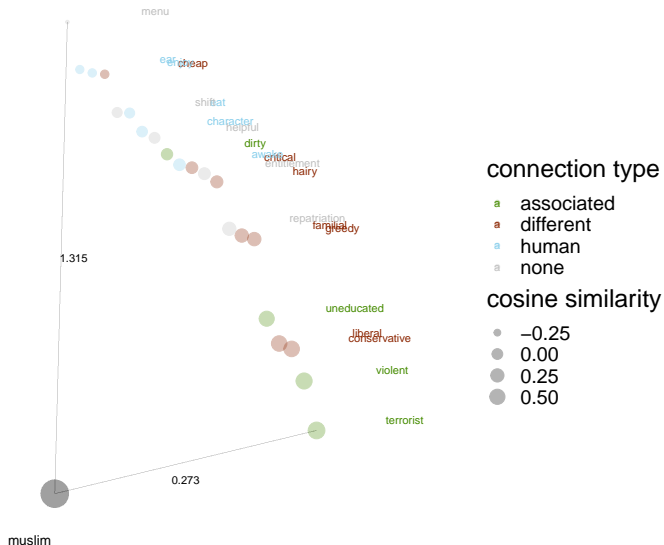
## Outliers and surprisingly dissimilar words

We study those by visualizations and uncertainty estimates.



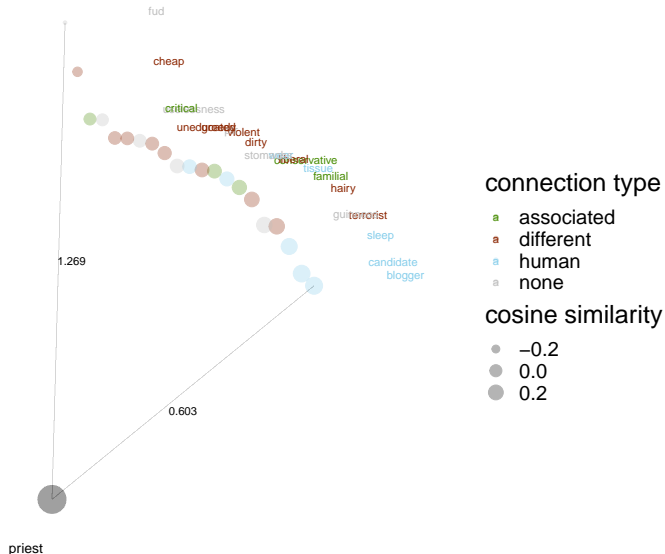
# Some methodological problems

## Distances for “muslim”



# Some methodological problems

## Distances for “priest”



# Some methodological problems

## No principled interpretation

| Category | Biased | Hard Debaised | Soft Debaised | Diff  |
|----------|--------|---------------|---------------|-------|
| Religion | 0.859  | 0.934         | 0.894         | 0.075 |
| Race     | 0.892  | 0.925         | 0.985         | 0.033 |
| Gender   | 0.623  | 0.700         | 0.747         | 0.077 |

- What values are sufficient for the presence of bias and what differences are sign of real improvement?
- Low  $p$ -values are not high effect indicators!
- We compare HPDIs.

# The problem with pre-averaging

## Key conceptual issues

- It throws away information about sample sizes
- It ignores variation in the raw data, which leads to false confidence

# The problem with pre-averaging

## Key conceptual issues

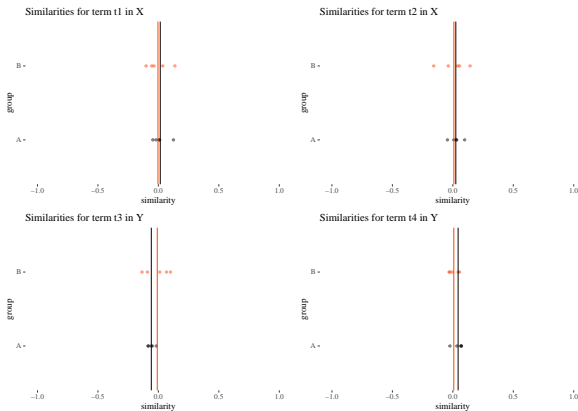
- It throws away information about sample sizes
- It ignores variation in the raw data, which leads to false confidence

## Our simulations

Suppose all similarities for two classes are randomly drawn from the same distribution,  $\text{Normal}(\mu = 0, \sigma = .08)$ , you still can get a really high WEAT!

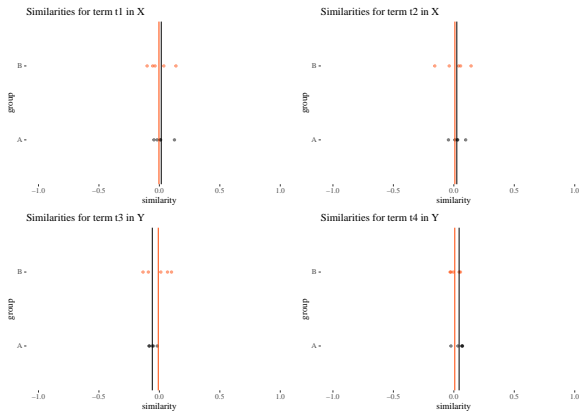
# The problem with pre-averaging

Simple case: two pws, four terms



# The problem with pre-averaging

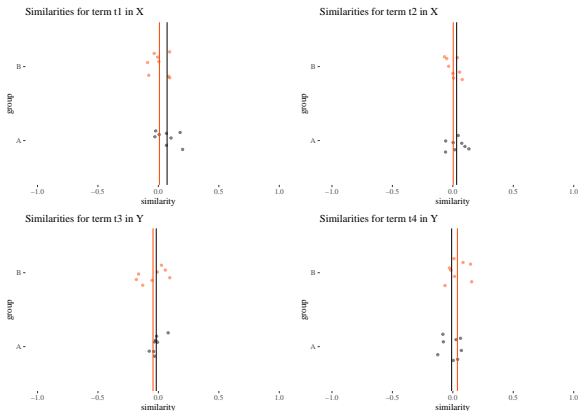
## Simple case: two pws, four terms



- Raw sd in data is 0.072
- The sd of means is 0.037
- The WEAT score is 1.825
- The largest effect size reported by Caliskan, Bryson, & Narayanan (2017) is 1.81!

# The problem with pre-averaging on realistic set-up

## Simulation with realistic set-up (16 predicates)

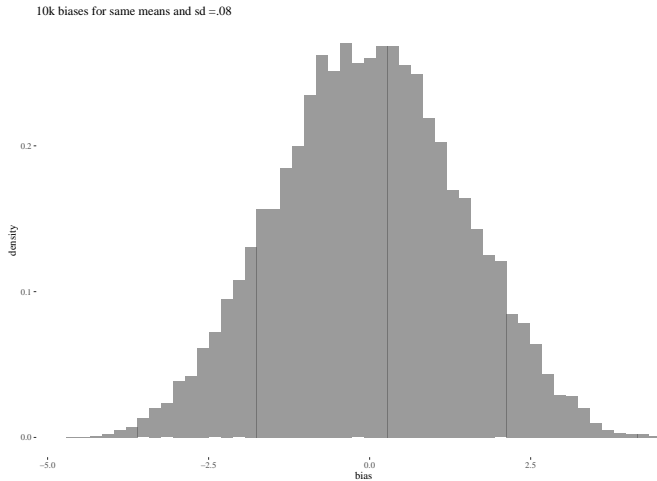


Raw sd: 0.082, sd of means: 0.031, WEAT: 2.337.



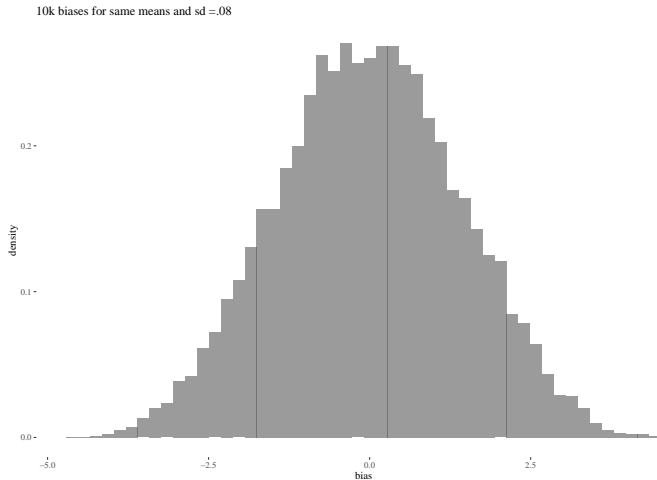
# The problem with pre-averaging on realistic set-up

## 10k simulations (same parameters)



# The problem with pre-averaging on realistic set-up

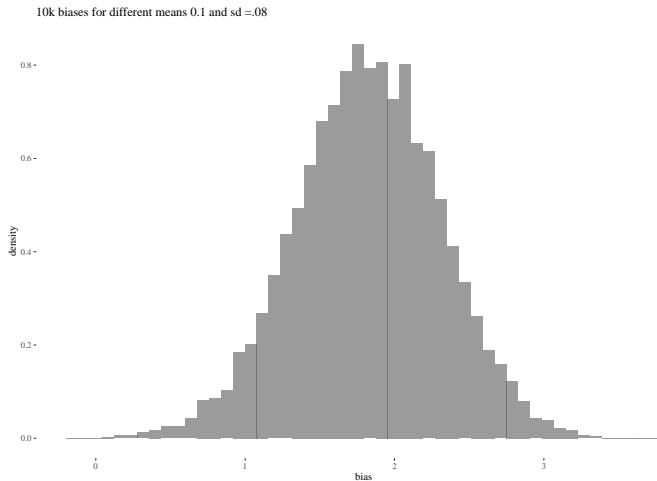
## 10k simulations (same parameters)



- 95% of the scores are in range -2.763, 2.698
- 21.38% of the absolute values are above 1.81

# The problem with pre-averaging on realistic set-up

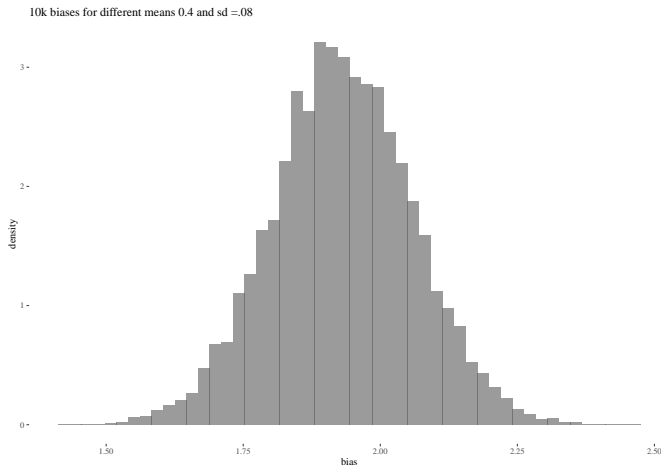
## 10k simulations with mean similarity 0.1



- 95% of the scores are in range 0.851, 2.764
- 51.3% of the absolute values are above 1.81

# The problem with pre-averaging on realistic set-up

## 10k simulations with mean similarity 0.4



- 95% of the scores are in range 1.679, 2.185
- 82.9% of the absolute values are above 1.81

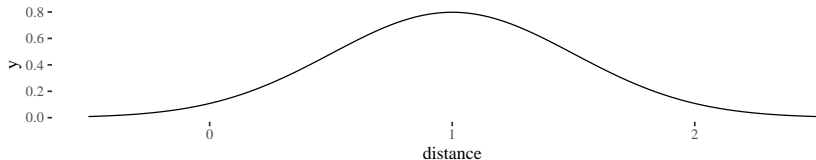
# Advantages of the Bayesian way

- Direct impact of sample sizes
- Straightforward interpretation in terms of posterior probabilities
- Freedom to choose granularity level
- More honest risk assessment and decision making

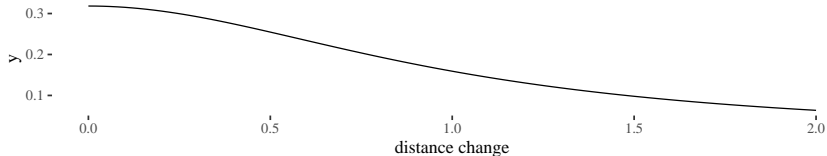
# Bayesian model

## Choosing priors

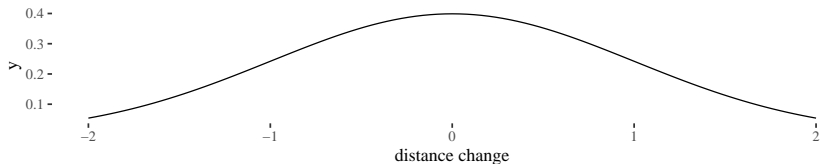
Prior for mean distances



Prior for standard deviation



Prior for coefficients

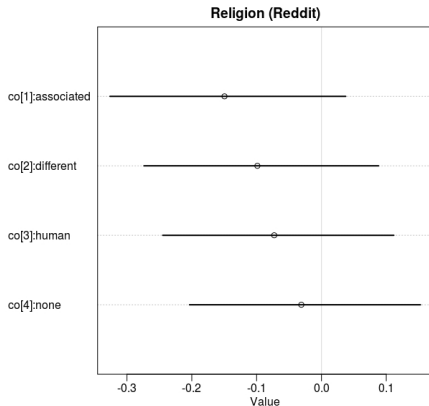


# Bayesian model architecture

```
library(rethinking)
options(buildtools.check = function(action) TRUE )
religionCoefs <- ulam(
  alist(
    cosineDistance ~ dnorm(mu,sigma),
    mu <- m + co[con],
    m ~ dnorm(1,.5),
    co[con] ~ dnorm(0,.5),
    sigma ~ dcauchy(0,1)
  ),
  data = religion,
  chains=2 , iter=8000 , warmup=1000,
  log_lik = TRUE
)
```

# Dataset-level coefficients

## Religion with 89%-compatibility intervals (HPDI)

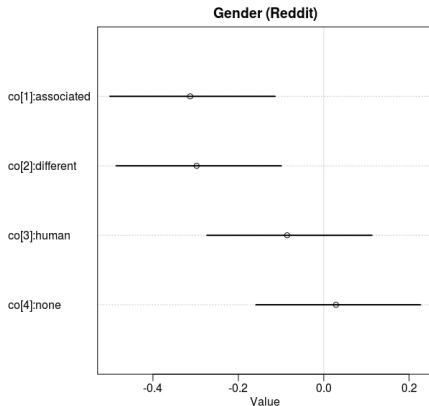


- All HPDIs overlap with 0
- Differences between classes are relatively small
- Coefficients for Race are similar



# Dataset-level coefficients

## Gender with 89%-compatibility intervals (HPDI)

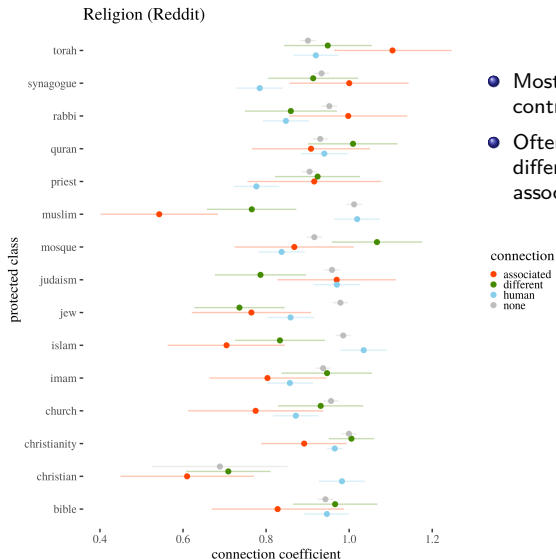


- Associated and different are away from 0
- But they were supposed to be opposites and are very close to each other (co-occurrence?)
- Differences between classes are still relatively small

# Bayesian model architecture

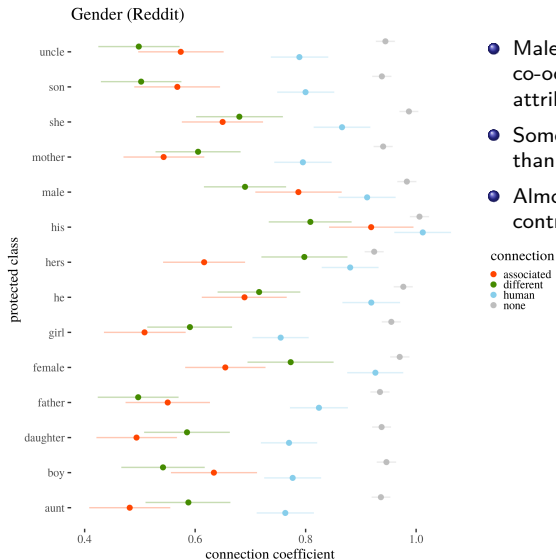
```
library(rethinking)
options(buildtools.check = function(action) TRUE )
religionCoefs <- ulam(
  alist(
    cosineDistance ~ dnorm(mu,sigma),
    mu <- m[pw] + co[con],
    m[pw] ~ dnorm(1,.5),
    co[con] ~ dnorm(0,.5),
    sigma ~ dcauchy(0,1)
  ),
  data = religion,
  chains=2 , iter=8000 , warmup=1000,
  log_lik = TRUE
)
```

# Word-level coefficients



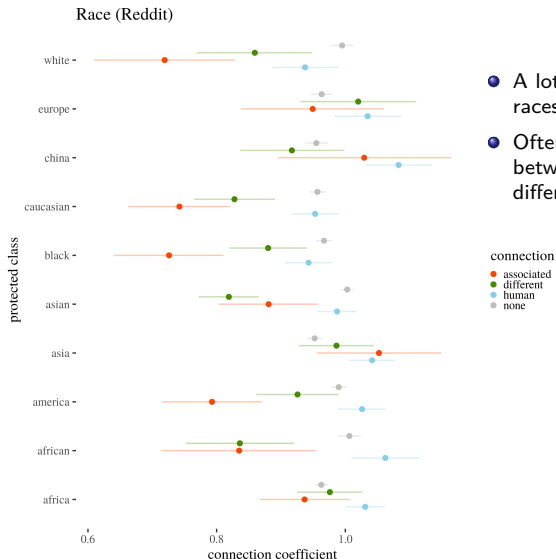
- Most intervals overlap with control groups
- Often not too much difference between associated and different

# Word-level coefficients



- Male attributes: strong co-occurrence with female attributes
- Sometimes different is closer than associated
- Almost no overlap with control groups

# Word-level coefficients



- A lot of variation between races
- Often not much difference between associated and different

# Thank you!

## Summary

- Bias in word embeddings
- WEAT and MAC methods
- Methodological problems
- Limitations of pre-averaging in bias detection methods
- Accounting for uncertainty with Bayesian approach

## Further work

- Including contrasts in Bayesian calculation
- Performance cross-validation in comparison to other methods (regular linear regression, KNN, ...)
- Downstream tasks and connection with intrinsic evaluation
- Testing data from the original Implicit Association Test (IAT)
- Applying uncertainty to WEAT and better word lists
- Looking at other similarity measures

# References

- Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334), 183–186. <https://doi.org/10.1126/science.aal4230>
- Ethayarajh, K. (2020). *Is your classifier actually biased? Measuring fairness under uncertainty with bernstein bounds*.
- Ethayarajh, K., Duvenaud, D., & Hirst, G. (2019). Understanding undesirable word embedding associations. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 1696–1705.
- Garg, N., Schiebinger, L., Jurafsky, D., & Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16), E3635–E3644. <https://doi.org/10.1073/pnas.1720347115>
- Gonen, H., & Goldberg, Y. (2019). Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 609–614. Minneapolis, Minnesota: Association for Computational Linguistics. <https://doi.org/10.18653/v1/N19-1061>
- Lauscher, A., & Glavas, G. (2019). Are we consistently biased? Multidimensional analysis of biases in distributional word vectors. *CoRR*, abs/1904.11783. Retrieved from <http://arxiv.org/abs/1904.11783>
- Manzini, T., Lim, Y. C., Tsvetkov, Y., & Black, A. W. (2019). *Black is to criminal as caucasian is to police: Detecting and removing multiclass bias in word embeddings*. Retrieved from <http://arxiv.org/abs/1904.04047>
- Nissim, M., Noord, R. van, & Goot, R. van der. (2020). Fair is better than sensational: Man is to doctor as woman is to doctor. *Computational Linguistics*, 46(2), 487–497. [https://doi.org/10.1162/coli\\_a\\_00379](https://doi.org/10.1162/coli_a_00379)
- Zhang, H., Sneyd, A., & Stevenson, M. (2020). *Robustness and reliability of gender bias assessment in word embeddings: The role of base pairs*. Retrieved from <http://arxiv.org/abs/2010.02847>