1878

Western

UNIVERSITY · CANADA

Using Decision-Directed Data Decomposition to Modify Neural Representations

https://arxiv.org/abs/1909.08159

Western Science

By Brent Davis, Ethan Jackson & Dan Lizotte

Table of Contents

- Background
- D^4 Formulation and Algorithm
- Applications:
  - Targeted Concept Removal in Neural Image Space
  - Improving Generalization
  - De-biasing Vector Representations of Words
- Future Work:
  - Theory: Kernel-based D^4
  - Applied: Modifying Natural Language Generation

Background

- Unnecessary or noisy data has long been a problem; it can typically be removed to some extent by using dimensionality reduction and discarding dimensions with lower variance.
- Another approach to information being entangled in a neural space is to disentangle only the relevant information.
- Orthogonal projections have been explored to debias neural representations by discarding information tied to bias.
- Bias has been known to be deeply entrenched and resistant to attempts to remove it; more thorough techniques are required.

High-Level D^4

- D^4 is an algorithm that performs repeated orthogonal projections until there is no discriminability left between the two classes.

- This is done by repeatedly disentangling a component from the full neural information space, resulting in many disentangled components that are undesirable and one modified information space.

- While initial experiments show promise in preventing recoverability, there is no guarantee that this will hold for any given case D^4 is applied to.

D^4 Basic Operations

Generalized linear decision function

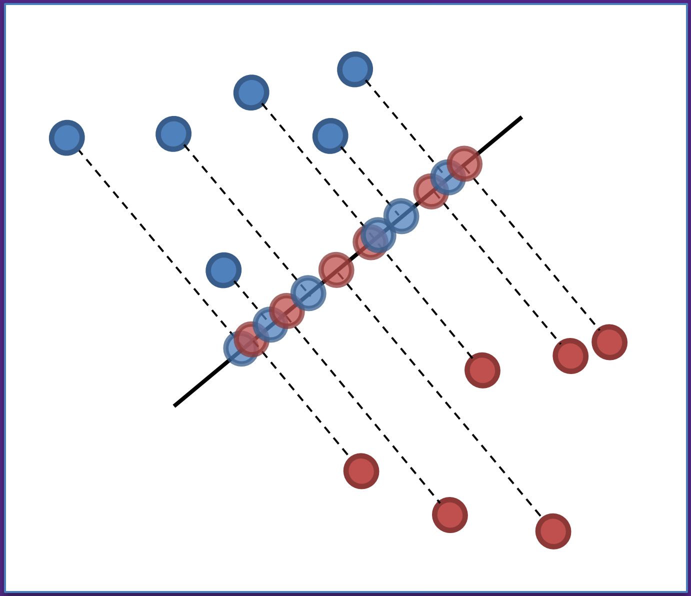
$$h(\boldsymbol{x}) = g\left(\boldsymbol{x}^{\mathsf{T}}\boldsymbol{w}\right)$$

Projection of $\boldsymbol{x}$ onto unit vector $\boldsymbol{\omega}$

$$\boldsymbol{X}_{||} = \boldsymbol{X}\boldsymbol{\omega}\boldsymbol{\omega}^{\mathsf{T}}$$

Projection of $\boldsymbol{x}$ onto orthogonal complement of $\boldsymbol{\omega}$

$$\boldsymbol{X}_{\perp} = \boldsymbol{X}(\boldsymbol{I} - \boldsymbol{\omega}\boldsymbol{\omega}^{\mathsf{T}})$$



Western Science

D[4] Algorithm

Data: Full-rank feature matrix $\boldsymbol{X}$ $(n \times p)$ of training points, targets $\boldsymbol{y}$ $(n \times 1)$
Result: Orthogonal basis vectors $\boldsymbol{\omega}^{(\mathbf{1})}$, $\boldsymbol{\omega}^{(\mathbf{2})}$, …, $\boldsymbol{\omega}^{(\boldsymbol{p})}$

for $\boldsymbol{i}$ *from* **1** *to* $\boldsymbol{p}$ **do**

$\boldsymbol{\Omega} \leftarrow \boldsymbol{I} - \sum_{\boldsymbol{j=0}}^{\boldsymbol{i-1}} \boldsymbol{\omega}^{(j)} \boldsymbol{\omega}^{(j)\mathsf{T}}$ (Sum Projections)

$\boldsymbol{w} \leftarrow \text{learn}(\boldsymbol{X\Omega}, \boldsymbol{y})$ (Project and Fit)

$\boldsymbol{\omega}^{\boldsymbol{i}} \leftarrow w / ||w||$ (Normalize)

D^4 vs PCA

- PCA is an **unsupervised** decomposition method that uses similar operations (projections)
- D^4 is a **supervised** decomposition method that uses a different strategy for identifying components (learned decision boundaries vs. variance maximization)
- PCA can not target specific components for removal, but it can be effective for removing arbitrary non task-oriented information

Considerations and Limitations of D^4

- D^4 is **Supervised** Learning: Limited by the Labels and available data to inform labels
    - *Labels are subject to bias.*
- Here, we perform binary classification:
    - Gender doesn’t exist as a binary split.
    - Nor does gender exist as a non-changing point (for some individuals)
    - We revisit options for multi-class modifications in Future Work

Considerations and Limitations of D^4

Images & D^4: Concepts in Image Space

Deep neural networks learn rich representations of data that may capture non task-oriented concepts.

(a) Husky classified as wolf (b) Explanation

Figure 11: Raw data and explanation of a bad model's prediction in the "Husky vs Wolf" task.

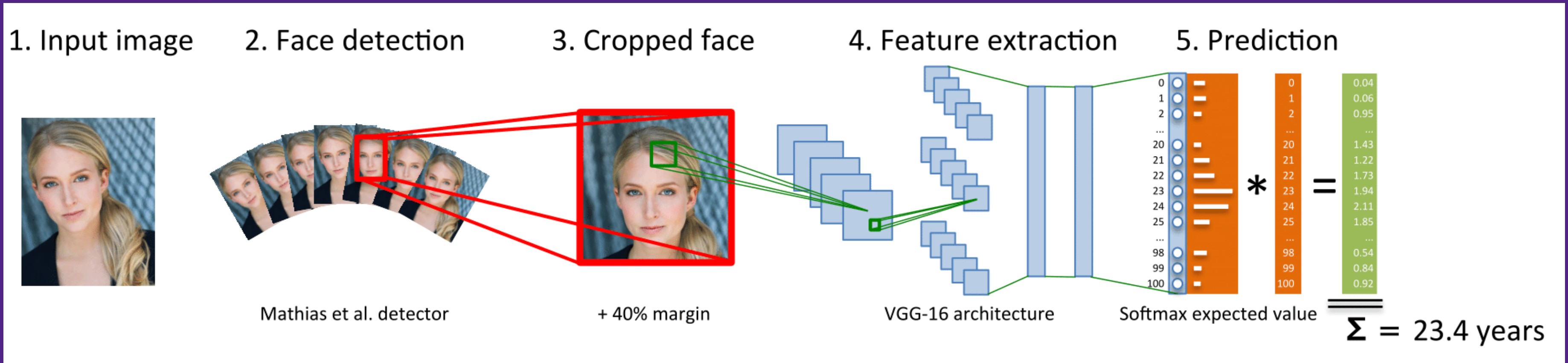
Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). *"Why Should I Trust You?": Explaining the Predictions of Any Classifier*.

Images & D^4: Concepts in Image Space

IMDB-WIKI Dataset

- Images of human faces with age and gender* labels

Deep Expectation of Apparent Age Method



Rothe, R., Timofte, R., & Van Gool, L. (2015). *DEX: Deep EXpectation of Apparent Age from a Single Image*.

Images & D^4: Target Concept Removal

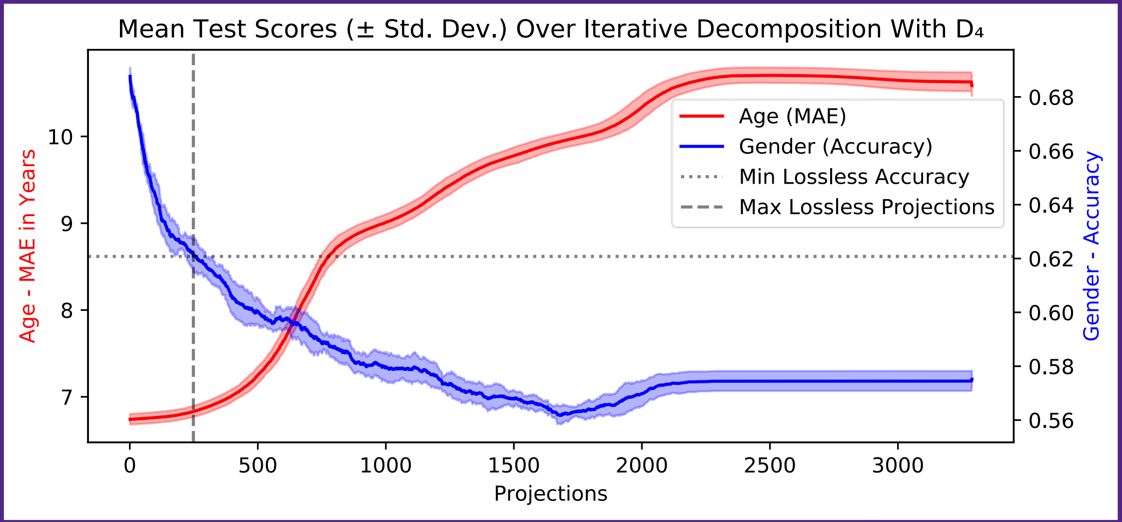
- How much information does DEX capture about gender when it is trained solely on age?
- Can (linear) discriminability on age and gender be disentangled?
- How much information about gender does DEX rely on to achieve target levels of age prediction error?

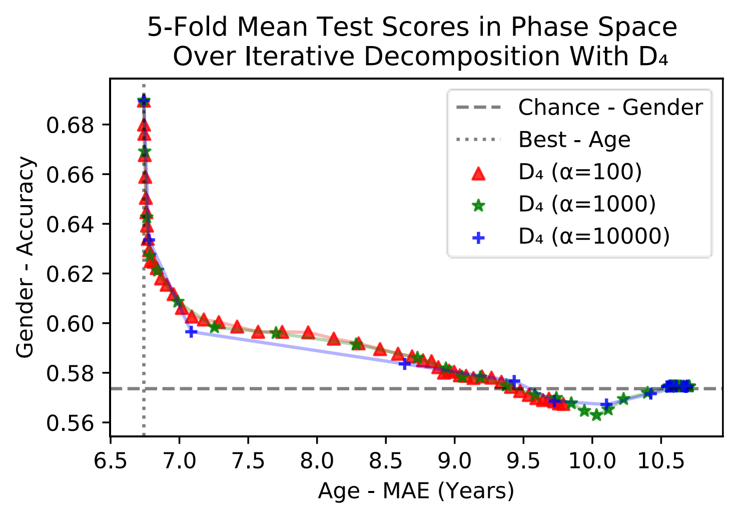
Images & D^4

6.7% reduction in linear gender discriminability with near-zero impact on age prediction

Further information on gender decision directions can be iteratively removed

L2-regularization can significantly reduce the number of D^4 iterations needed to discard information





Western Science

Generalization, Obscuring Information & D^4

- If the presence of snow is more reliable than any extracted image features, why would a classifier not continue to use it?

- The utility of D^4 here comes from being able to remove features like this, forcing the classifier to work with features that we know are more generic.

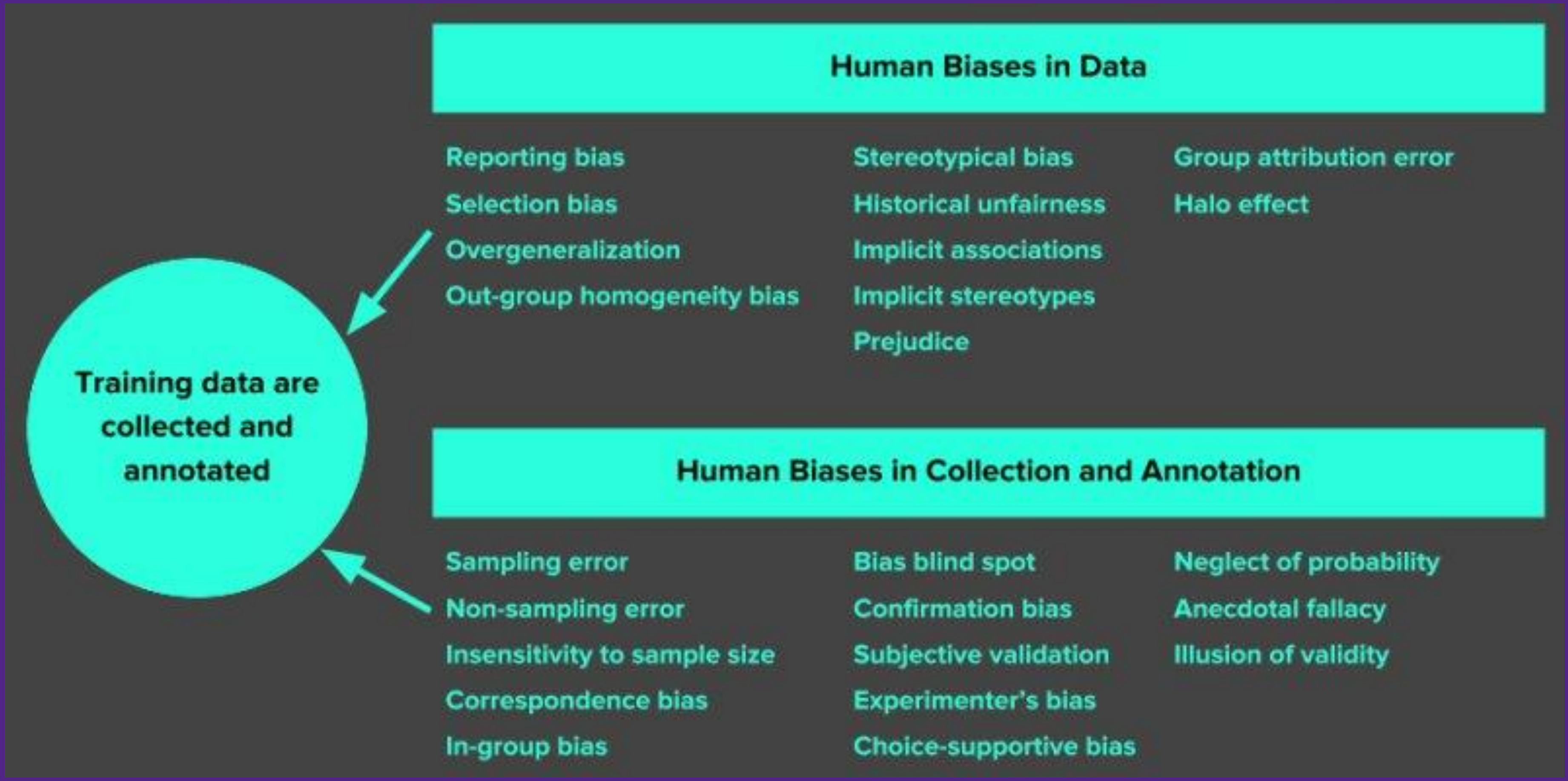
(a) Husky classified as wolf (b) Explanation

Figure 11: Raw data and explanation of a bad model's prediction in the "Husky vs Wolf" task.

Bias

(from: https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1194/slides/cs224n-2019-lecture19-bias.pdf

by Dr. Margaret Mitchell)



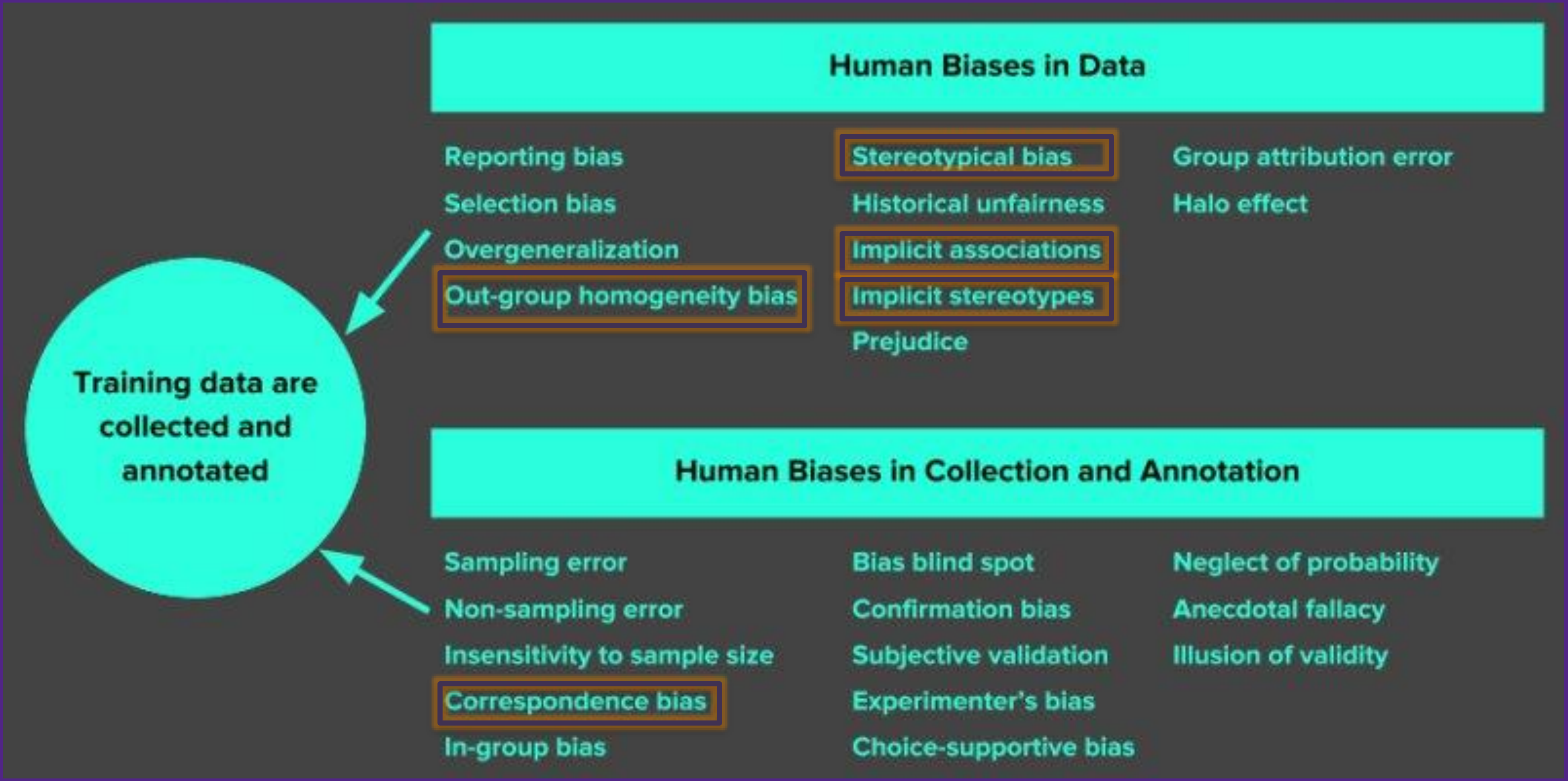
Bias in Word Vectors

- Our goal here is to remove as much bias as we can.
- D^4 (and any other known technique) are not silver bullets for this; there are kinds of bias that it will not be able to mitigate.
- What kinds are we attempting to target then?

Bias in Word Vectors

(from: https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1194/slides/cs224n-2019-lecture19-bias.pdf

by Dr. Margaret Mitchell)

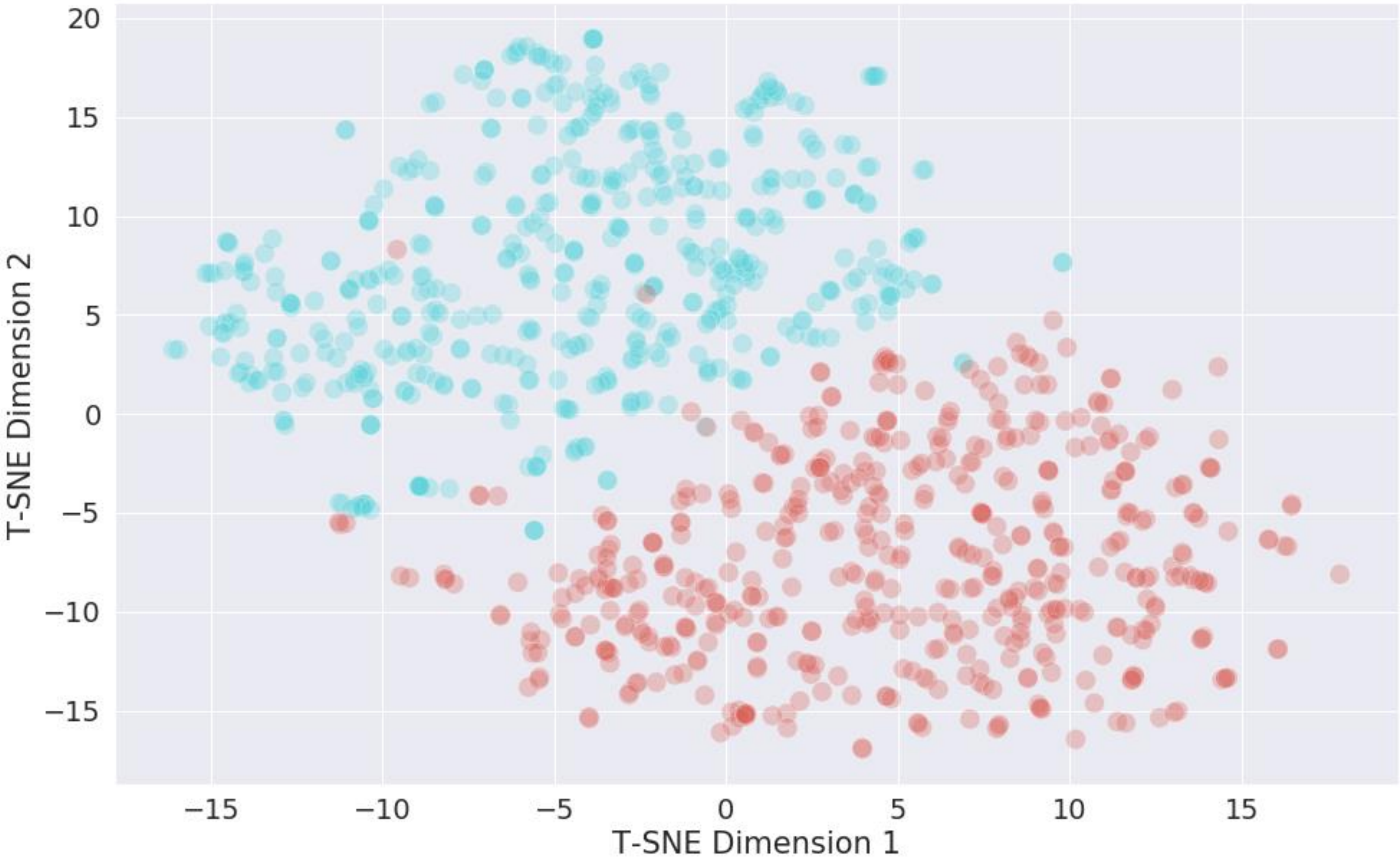


Western Science

Bias in Word Vectors

- Optimistically, we’re targeting 5/24 kinds of bias listed. (*Important: Targeting doesn’t guarantee success*)
- Some kinds of bias have been demonstrated by previous work, such as ‘Man is to Computer Programmer as Woman is to Homemaker?’ (Bolukbasi et al., 2016)
- This is done by taking the vector path from ‘Man’ to ‘Computer Programmer’ and then seeing which word is closest after taking the same path from ‘Woman’.
- We can see evidence of bias by looking at gendered professions, too –

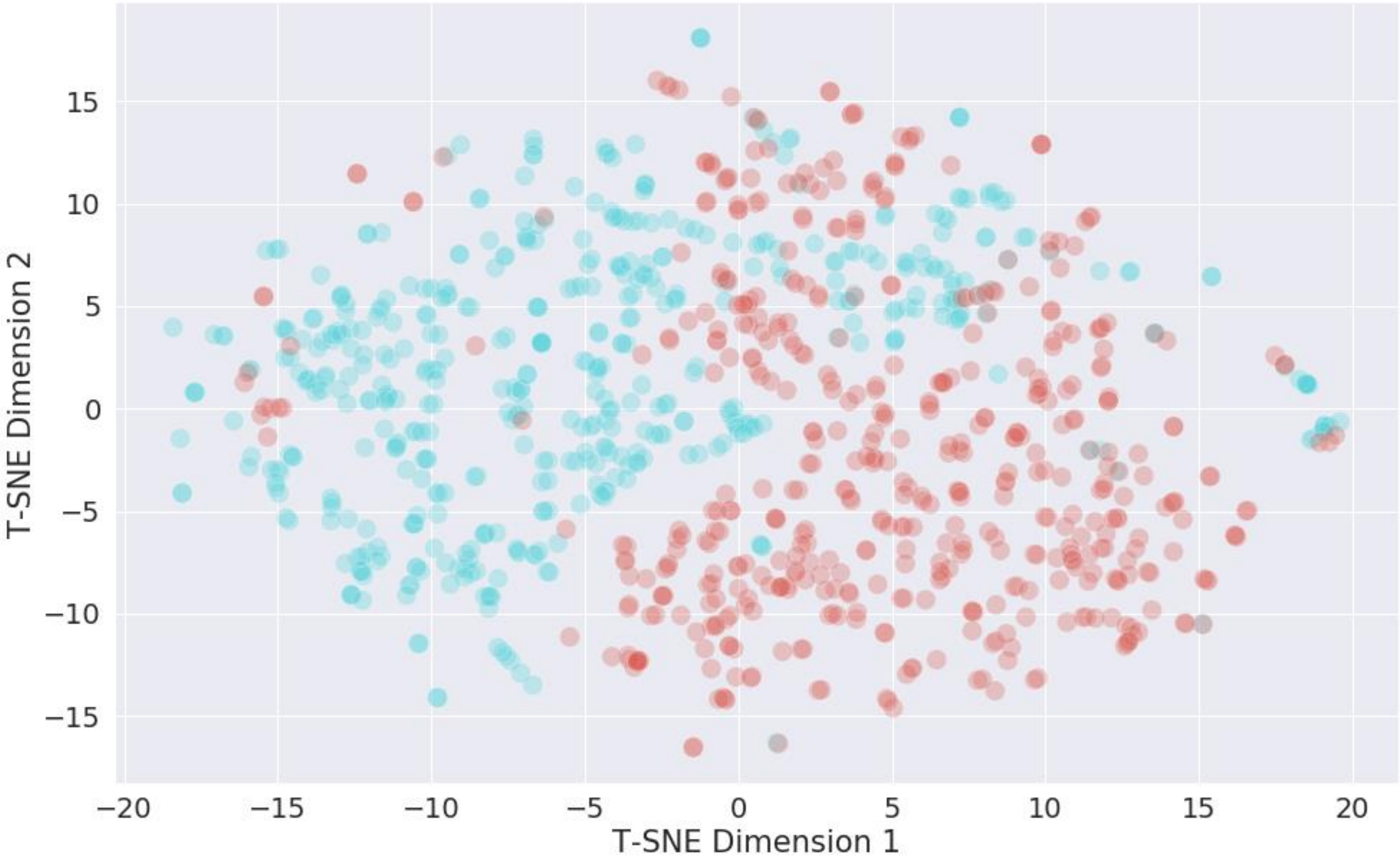
Bias in Word Vectors



Debiasing Word Vectors

- Goals:
  - Remove this kind of division between professions
  - Remove associations learned from this ordering / placement of professions in other, ungendered words
- By removing all associations / discriminability based on the difference between gendered words, we are attempting to enforce a kind of statistical parity (Man -> Programmer ~ Woman -> Programmer)

Debiasing Word Vectors



Debiasing Word Vectors

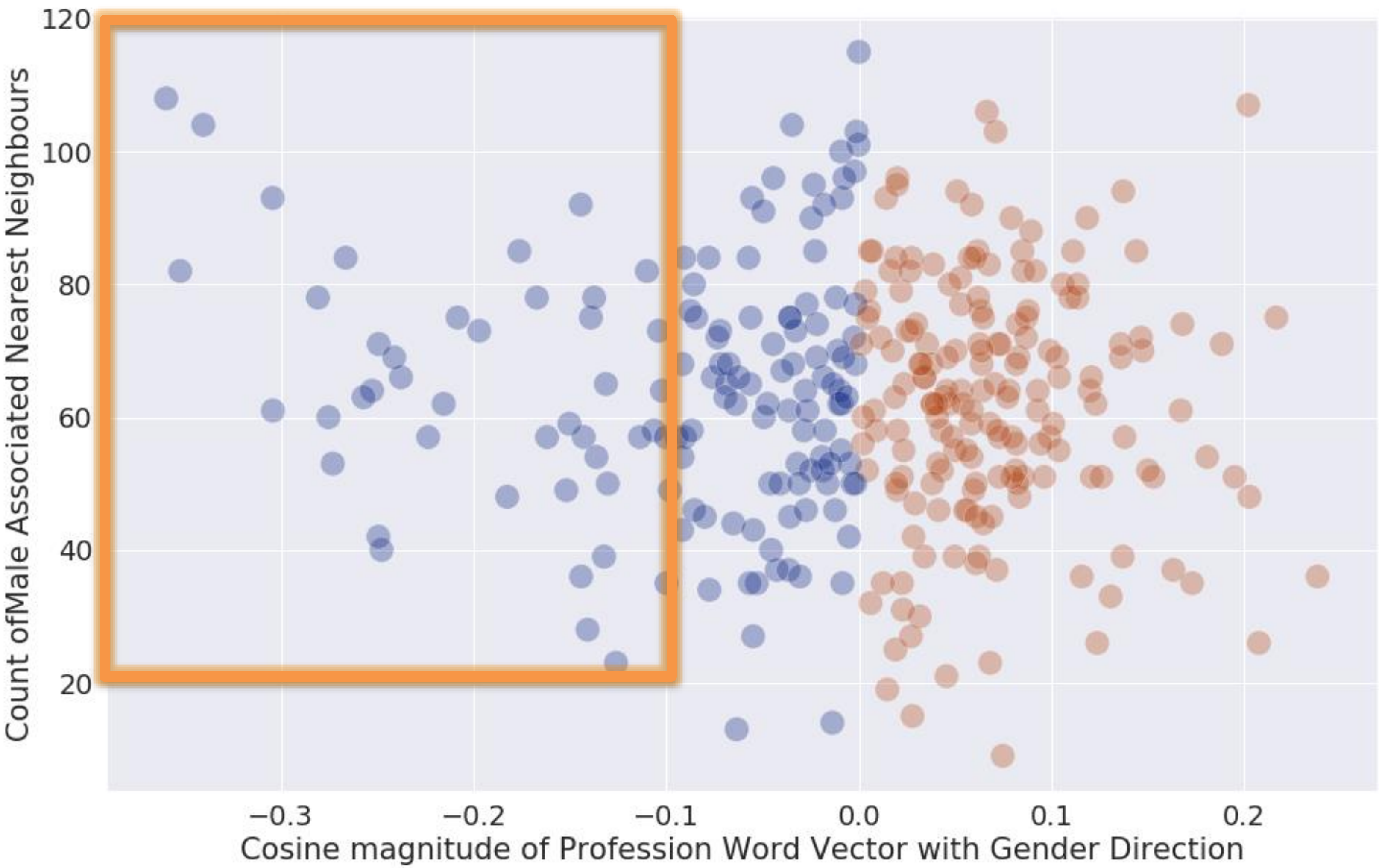
- Methodology:
  - Generate a list of words which have a gendered association (ex., businesswoman, salesman)
  - Train a classifier to maximally separate the two classes.
  - Project **all** word vectors orthogonal to the learned decision boundary
  - Repeat on debiased representations until classifier accuracy converges (typically to an accuracy of labeling all classes as being the same class)

Evaluating Debiasing Word Vectors: Nearest Neighbour Profession Count

Western Science

Evaluating Debiasing Word Vectors:

Debiased Nearest Neighbour Profession Count

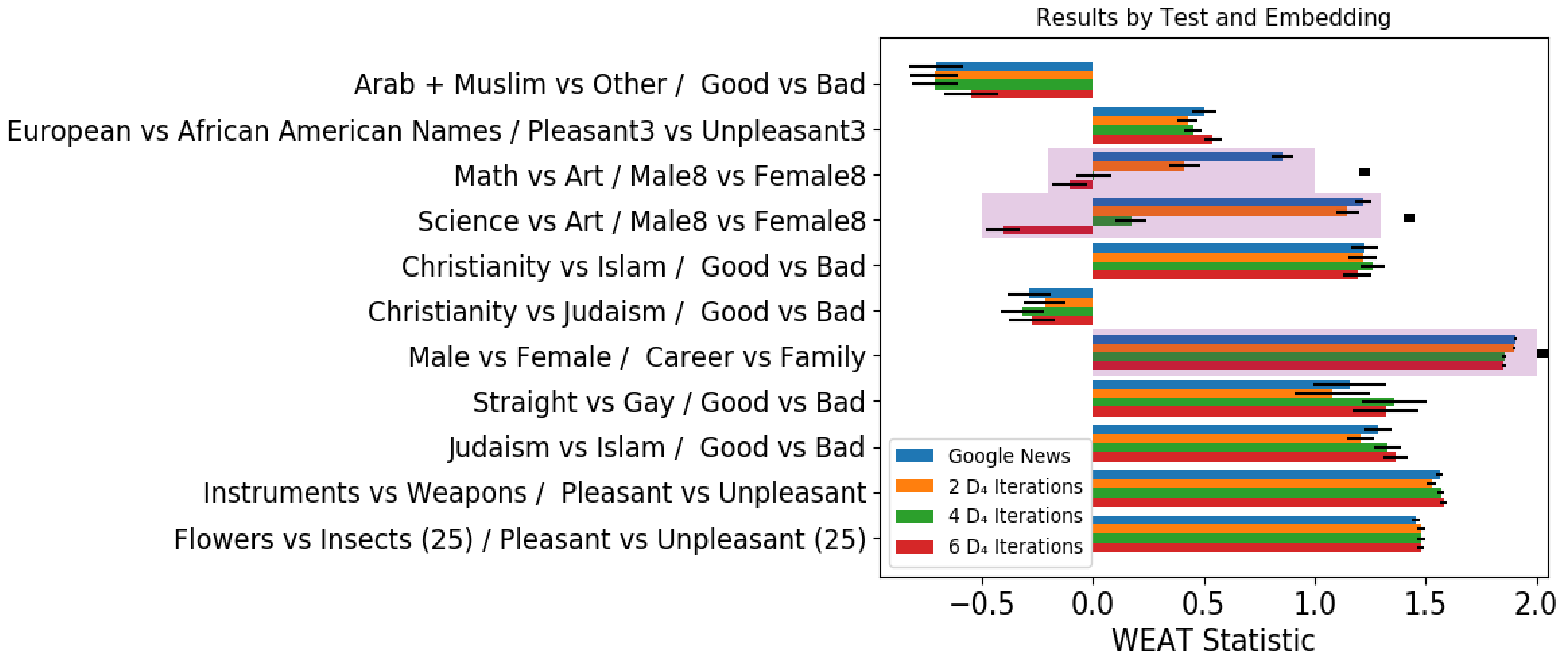


Western Science

Evaluating Debiasing Word Vectors: Word Embedding Association Test (WEAT)

- Demonstrating that bias has been removed is difficult.
  - Showing that bias is not recoverable is similarly difficult.
- We use Caliskan’s (2017) WEAT test to help quantify the effect we’re having.
- This test measures the association of words in each category with various measures (ex., Good & Evil) to test for bias.

Debiasing Word Vectors: Word Embedding Association Test (WEAT)



Western Science

Future Work

Kernel D^4

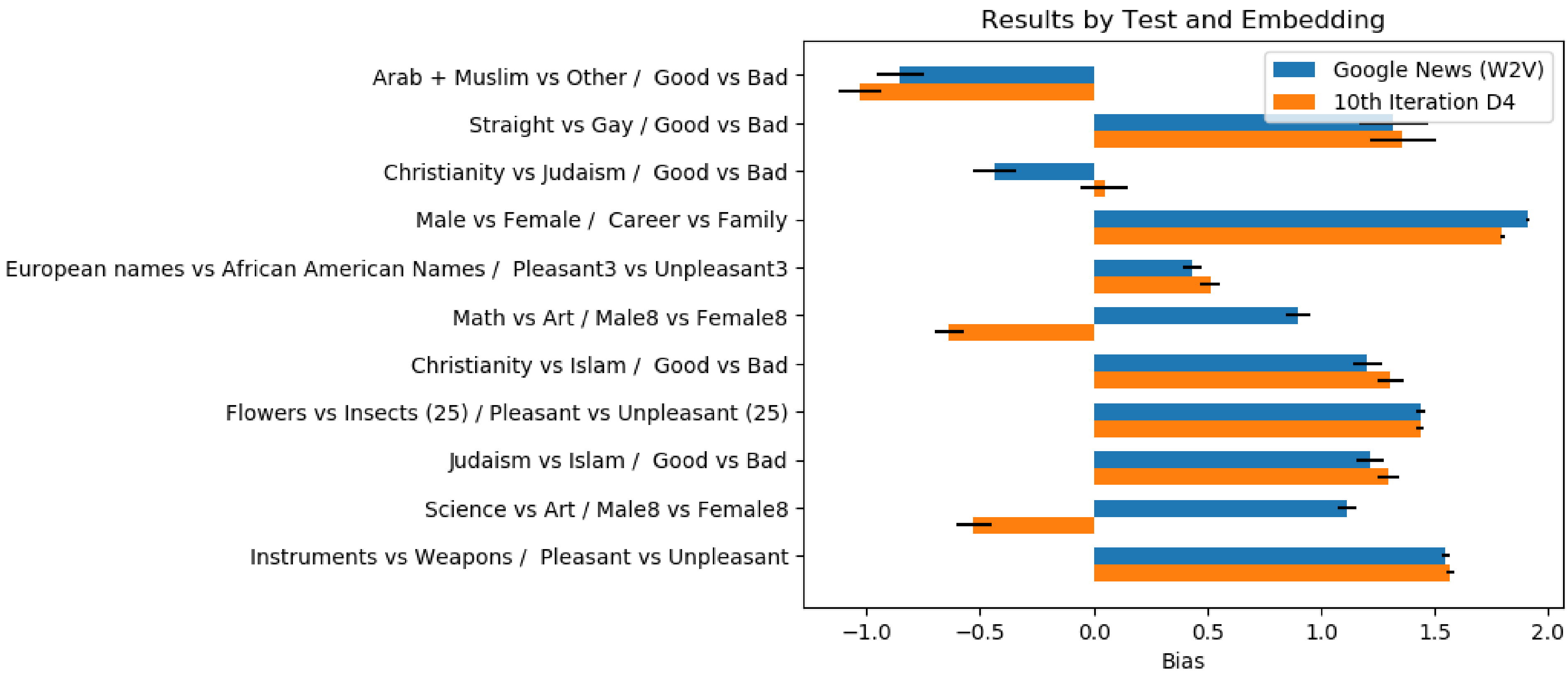
- General formulation with linear operators extensible to kernel spaces.
- Enables projections in non-linear feature spaces.

D^4: The Odd – Recovering Word Frequency? Projecting 300x in a 300 dimensional space.

- Seen here: Words at one extreme of the ‘gender direction’ (defined as the vector between ‘he’ and ‘she’) after applying D^4 300x to a 300 dimensional word embedding.
- We can’t get the original word frequencies, but we suggest that projecting this many times removes almost all information besides magnitude (word frequency).

D4: The Odd

Projecting past initial convergence



Western Science

Natural Language Generation & D^4: The Odd

- If we apply D^4 10x, we see some of the continued trend from the 2, 4 and 6 iterations.
- Intriguingly, the WEAT suggests we can reverse the direction of the bias (although we don't reach the same magnitude)
- This effect disappears when regularization is applied.
- Given D^4 can be applied to arbitrary labels, this could have applications in customizing generative text.

Natural Language Generation & D^4

- An application we are excited in developing for D^4 is in modifying the search space for generative text algorithms
- These methods use various heuristics and techniques involving the neural representation of words to ‘decode’ a choice of words when generating a sentence.
- Undesirable results can come from the associations embedded in pre-trained (or trained during) models.
- How can we modify the search space? (Hopefully D^4)

Natural Language Generation & D^4: Modifying Discovery by Decoders

- Common algorithms (greedy decoder, beam decoder) use a probability or criteria to search for words that are the most likely to appear in a sequence.

- We are proposing to essentially perform an adversarial attack on the generator that targets undesirable associations with D^4.

Multiclass D^4: More Gender Inclusive Debiasing

- We propose the use of multi-class support vector machines (or other multi-class classifiers producing decision boundaries) will have applications in more complicated debiasing situations.
- Unfortunately, we do not feel qualified to make, and have not been able to find, any works which provide multi-class labeled instances of words to test on.
  - Class imbalance, similarly, is likely to be an issue.

Societal Implications & Usage

Societal Implications: NLP

- Integration of any new techniques into real world practice is a complex, dynamic process to figure out what really works.
- One risk of any supervised debiasing comes from information, context and words that are not included.
- We try to mitigate (as much as we can) this by applying our projections to every word in the set, but this could exacerbate a bias blind-spot.
- Gendered slang, particularly new slang, is an example of a blind spot in most NLP work – the kinds of associations there could be unique and thus be missed by this kind of debiasing.
- Highlights debiasing as inherently interdisciplinary activity. We need the linguists and fairness expertise.

Questions?

A Last Question For The Audience:

How would you know if this technique was applied to a dataset by a bad actor without your knowing?

Western

UNIVERSITY · CANADA