



# **Week 4**

MORPH Algorithmic Fairness

# **Agenda:**

## **Applying Fairness to Algorithms I**

- Feedback
- Fairness in NLP context
- Week 5 preview
- Group presentation

The background features a large white circle centered on a black field. To the left of the white circle, there is a series of overlapping circles in shades of gray, creating a tunnel-like effect. To the right, there are several concentric white circles of varying diameters, also creating a tunnel-like effect.

# **Intro to Fairness in NLP**

**What do you see?**



Exercise inspired by V. Prabhakaran



# What do you see?

- Bananas
- Stickers
- Dole bananas
- Bananas at a store
- Bunches of bananas

We don't tend to say  
**yellow bananas**



**What do you see?**



Exercise inspired by V. Prabhakaran



# What do you see?

- **Green** bananas
- **Raw/unripe** bananas



# What do you see?

**Yellow** bananas

**Yellow** is “the norm” for bananas





A series of four concentric circles in a light gray color, centered on the left side of the slide.

# Storytime

A man and his son are in a terrible accident and are rushed to the hospital in critical care.

The doctor looks at the boy and exclaims "I can't operate on this boy, he's my son!"

How could this be?




A series of four concentric circles in a light gray color, centered on the left side of the slide.

# Storytime

A man and his son are in a terrible accident and are rushed to the hospital in critical care.

The doctor looks at the boy and exclaims "I can't operate on this boy, he's my son!"

How could this be?

A large, light gray circle with a thin white border, positioned on the right side of the slide.

Distinction  
between  
**doctor** and  
**female doctor**



"male doctor"



Images

Videos

News

Shopping

About 7,260,000 results (0.41 seconds)

"female doctor"



Images

Videos

News

Maps



About 13,600,000 results (0.45 seconds)



**The majority of test subjects overlooked the possibility that the doctor is a she – including men, women, and self-described feminists.**

Wapman and Belle (2014)



# Word Embeddings

- A method of representing words, useful in deep learning (which is unsupervised)
  - Unable to process strings or raw data
- Represent words with a **vector of numbers**
- Learn these representations from a large data corpus — word2vec (Google), GloVE (Stanford)

# Debiasing Word Embeddings: Part 1

Bolukbaski et al. "Man is to Computer Programmer as Woman is to Homemaker." (2016)

Main idea: Word embeddings embed sexism. In fact, we can identify the gender subspace  $g$ .

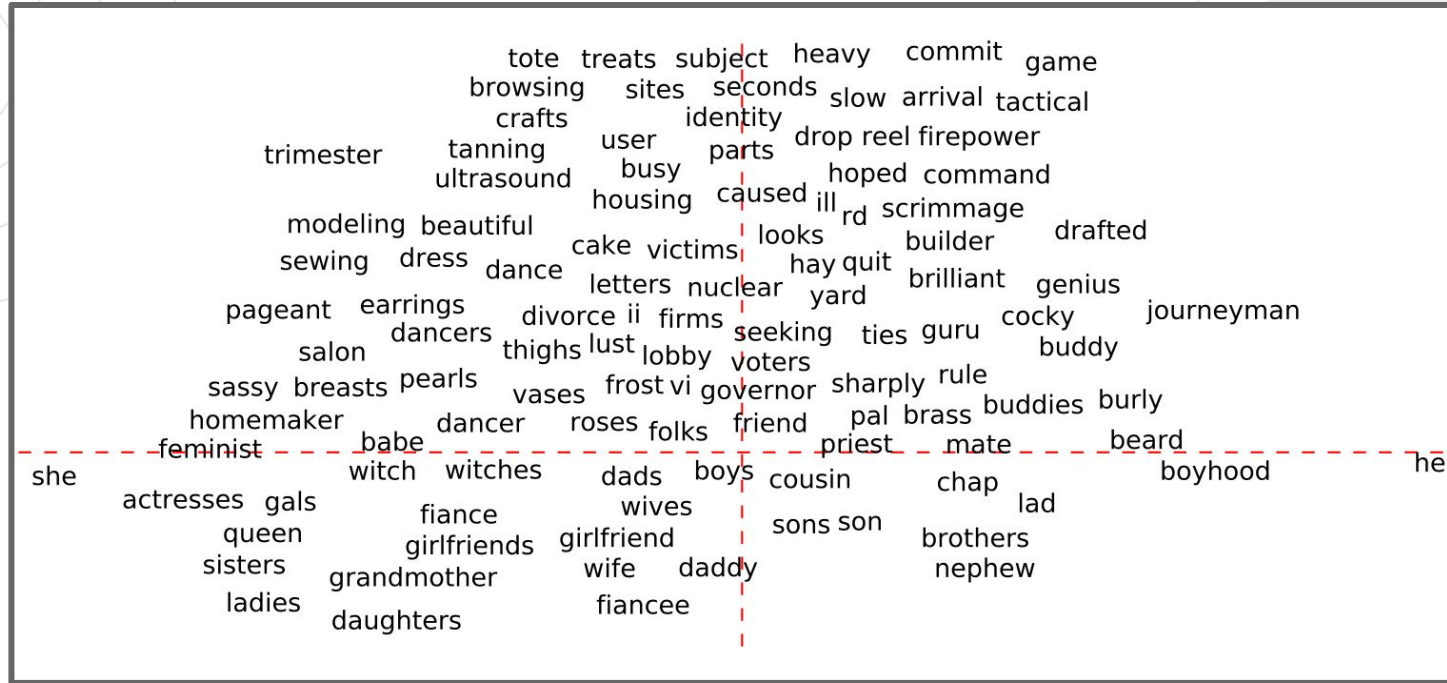
# Debiasing Word Embeddings: Part 1

Finding  $g$ , the gender subspace:

$$\overrightarrow{\text{grandmother}} - \overrightarrow{\text{grandfather}} = \overrightarrow{\text{gal}} - \overrightarrow{\text{guý}} = g$$

Use  $g$  to identify bias of embeddings:  $\cos(v, g)$  (or, equivalently, the dot product)

- Project word vectors onto gender dimension to get a quantitative bias score





# Debiasing Word Embeddings: Part 1

- Proposed debiasing methods (hard and soft) essentially subtracts gender direction from gender-neutral words to remove bias
- Happens in the postprocessing step
- After debiasing, these analogies should have a lower bias score



# Week 5 Preview

## ▼ Step 4: Analyzing gender bias in word vectors associated with professions

Next, we show that many occupations are unintendedly associated with either male or female by projecting their word vectors onto the gender dimension.

The script will output the profession words sorted with respect to the projection score in the direction of gender.

```
[ ] # profession analysis gender
    sp = sorted([(E.v(w).dot(v_gender), w) for w in profession_words])

    sp[0:20], sp[-20:]
```

```
↳ ((-0.23798442, 'maestro'),
    (-0.21665451, 'statesman'),
    (-0.20758669, 'skipper'),
    (-0.20267202, 'protege'),
    (-0.2020676, 'businessman'),
    (-0.19492392, 'sportsman'),
    (-0.18836352, 'philosopher'),
    (-0.1807366, 'marksman'),
    (-0.1728986, 'captain'),
    (-0.16785555, 'architect'),
```

# Debiasing Word Embeddings: Part 2

- Rather than postprocess, trains debiased embeddings from scratch (learns gender neutral embeddings)
- Add new dimensions to capture
  - 1) gender information (definition and stereotype),
  - 2) gender neutral information



The background is black. A large white circle is centered on the slide. To the left of the white circle, there are two overlapping circles: a dark gray one in front of a lighter gray one. To the right of the white circle, there are several concentric white circles of varying sizes.

# **Group Presentation**

Lipstick on a Pig