Week 6: Bias in Language

Warm-up: Riddle

A young boy and his father are on their way home from soccer practice when a distracted driver crosses the center line and hits them head-on. The father dies at the scene of this horrible car accident, but the boy is still alive when the emergency medical technicians arrive. The injured boy is transported in an ambulance to the hospital, where's he taken immediately into surgery.

However, the awaiting surgeon steps out of the operating room and says, "Call Dr. Baker stat to the operating room. I can't operate on this boy. He's my son!"

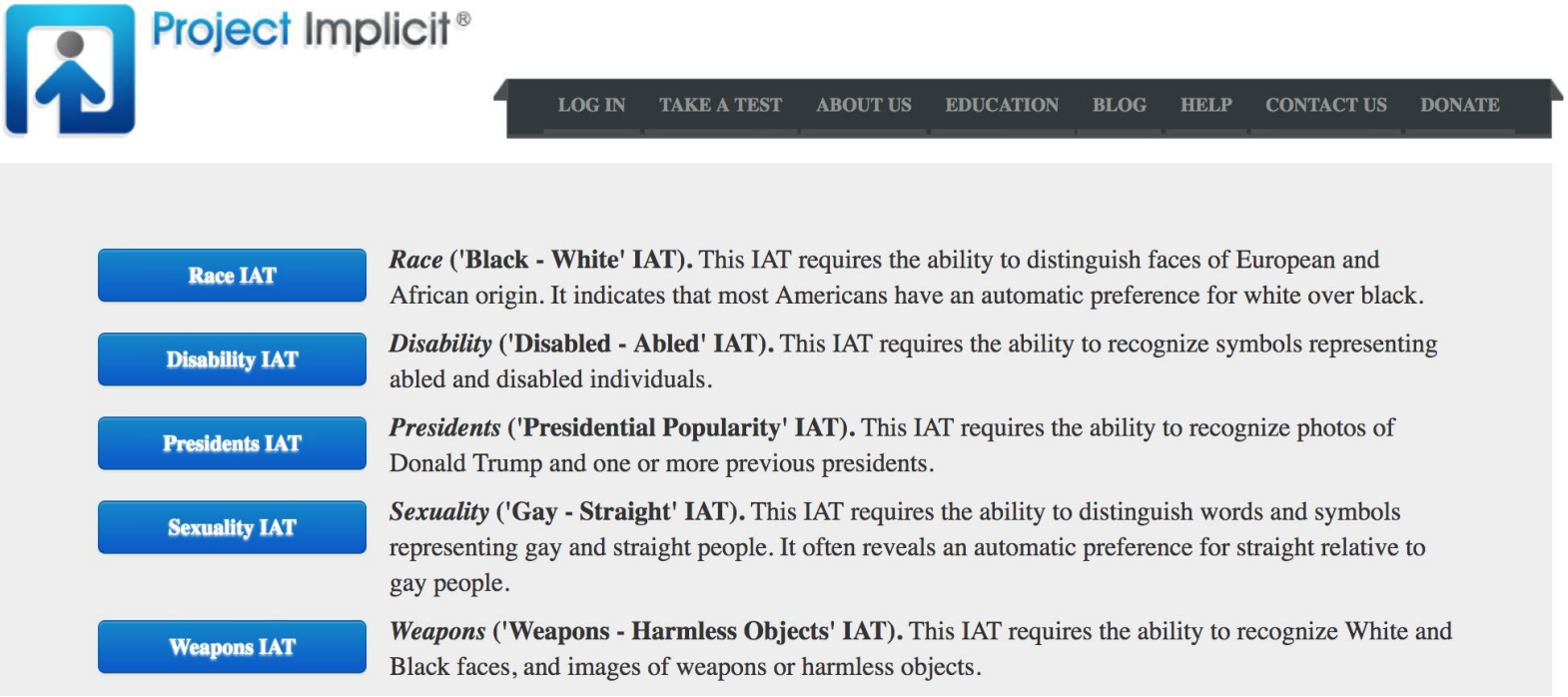
The question: Who is the surgeon?

Bias discussion questions (in pairs)

- What is *bias?*
- Do you think you have *bias*? If so, where?
- Do you think *bias* exists in the world/in data? Where and why?
- Is *bias* always bad?

Implicit association test (in pairs)

https://implicit.harvard.edu/implicit/selectatest.html



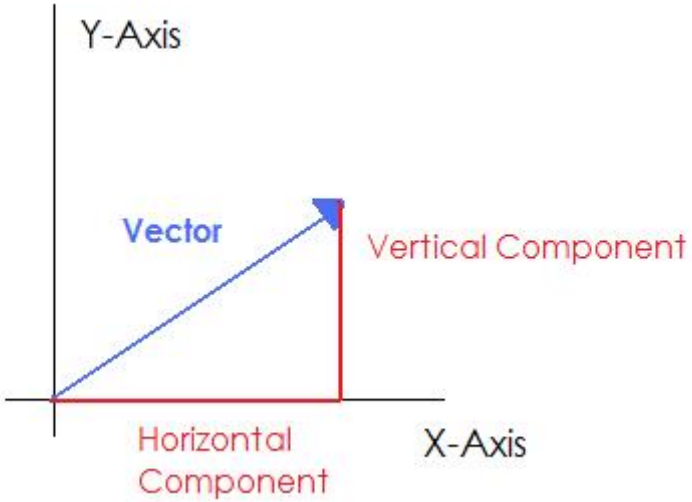
https://implicit.harvard.edu/implicit/takeatest.html

Reading summary

Vectors

Two dimensions

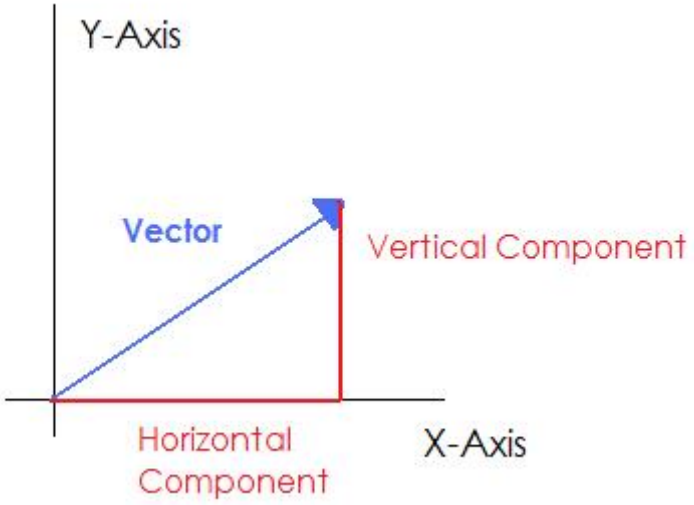
P = (x, y)



Vectors

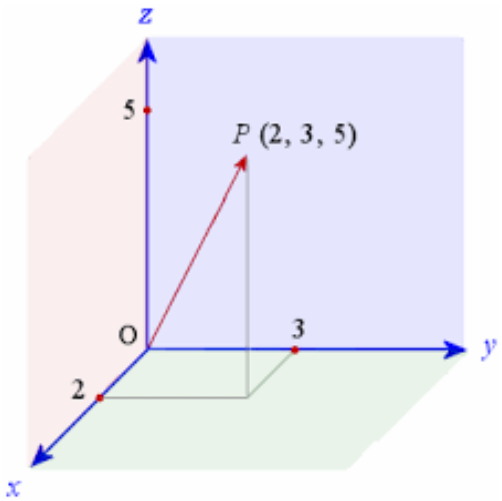
Two dimensions

P = (x, y)



Three dimensions

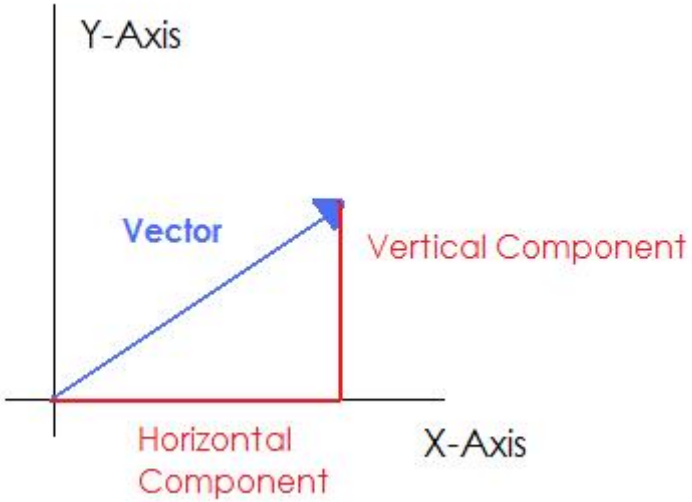
P = (x, y, z)



Vectors

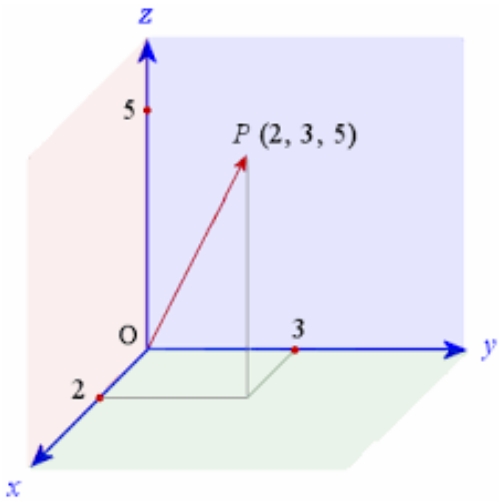
Two dimensions

$P = (x, y)$



Three dimensions

$P = (x, y, z)$



N dimensions

$P = (x_1, x_2, x_3, x_4, x_5, \ldots, x_n)$

Distributional hypothesis (linguistics)

- "a word is characterized by the company it keeps" (Firth, 1957)
- linguistic items with similar distributions have similar meanings

Distributional hypothesis (linguistics)

- "a word is characterized by the company it keeps" (Firth, 1957)
- linguistic items with similar distributions have similar meanings

Example: What is this word?

1. A bottle of ____ is on the table.
2. Everybody likes _____.
3. Don't have _____ before you drive.
4. We make ____ out of corn.

Example credit: Jacob Eisenstein

Distributional hypothesis (linguistics)

- "a word is characterized by the company it keeps" (Firth, 1935)
- linguistic items with similar distributions have similar meanings

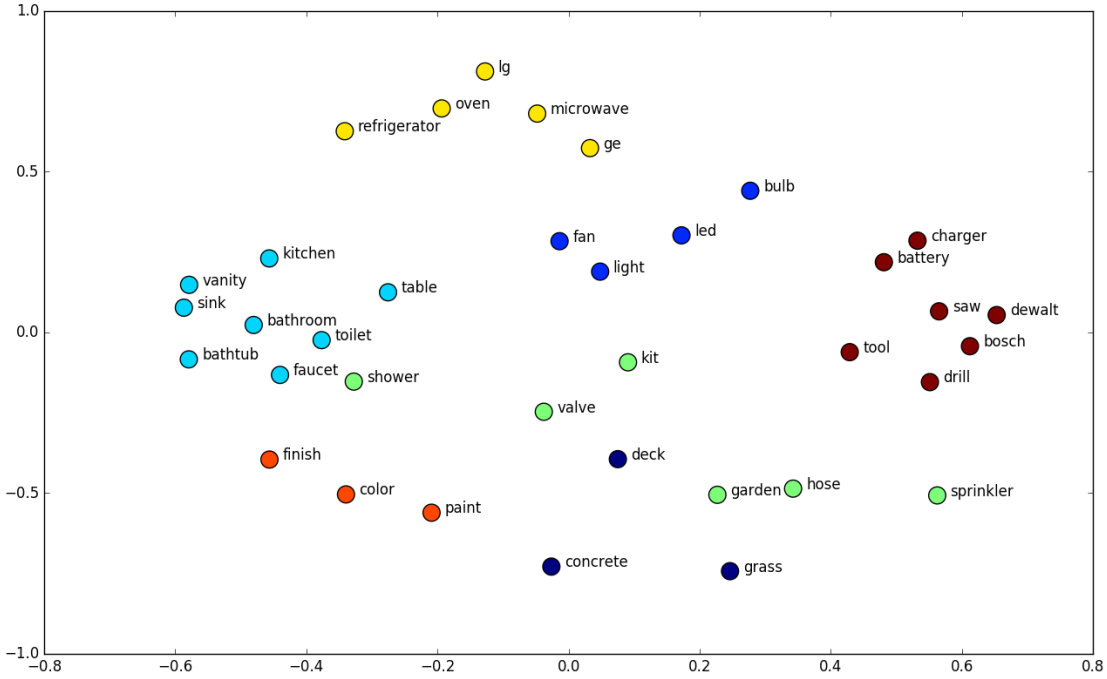
Example: What is this word?

1. A bottle of ____ is on the table.
2. Everybody likes _____.
3. Don't have _____ before you drive.
4. We make ____ out of corn.

Answer: ***Tesgüino*** (corn beer made by the Tarahumara Indians in Mexico)

Example credit: Jacob Eisenstein

Word Embeddings = represent each word as an n-dimensional vector



Sapir-Whorf hypothesis

Strong: language *determines* thought and decisions

Weak: language only *influences* thought and decisions

Discussion questions

- What evidence is there that word embeddings are sexist?
  - What caused this?
- What is the Word Embedding Association Test (WEAT)?
  - How does this extend the IAT?
  - What is the analogy to reaction time in IAT for WEAT?
- Once we've identified bias how do we address it?
  - What should we do about the "imprints of our historic biases"?
- Does the language we use matter?

Word embeddings demos

http://bionlp-www.utu.fi/wv_demo/

https://www.propublica.org/article/breaking-the-black-box-how-machines-learn-to-be-racist

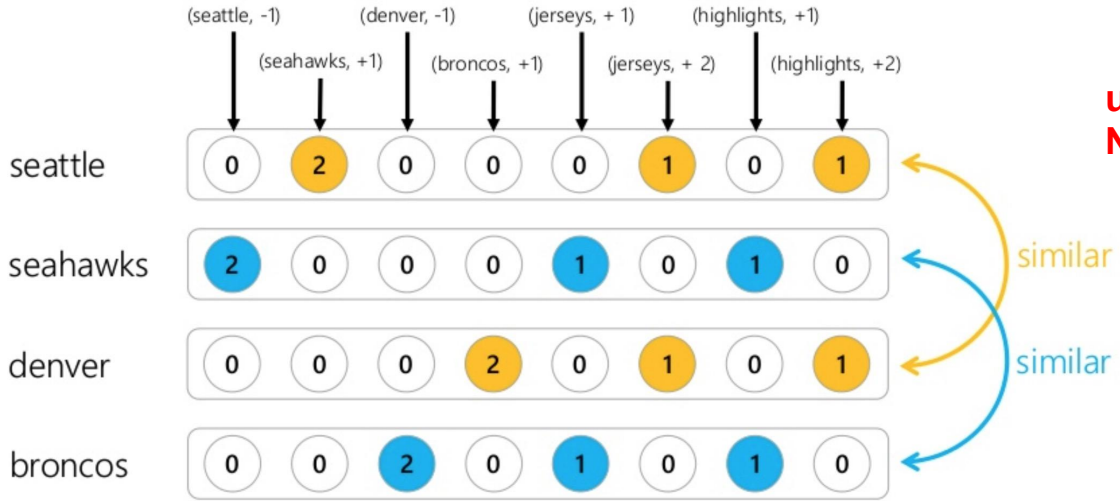
Takeaways

- What are word embeddings?
- People and society are biased (IAT), so language written by people is biased, so word embeddings are biased, so AI will be biased

Survey

Word Embeddings = represent each word as an n-dimensional vector

1. Find co-occurrence of each word with every other word in the vocabulary within a window (e.g. 5 words to the left and right)
2. Get a low-dimensional representation (matrix factorization, neural networks)



used as input into NLP systems!

Figure credit: Bhaskar Mitra