Applied Natural Language Processing

Info 256

Lecture 5: Bias in word embeddings (Sept. 11, 2023)

David Bamman, UC Berkeley

Analogical inference

- Mikolov et al. 2013 show that vector representations have some potential for analogical reasoning through vector arithmetic.

$$\text{apple} - \text{apples} \approx \text{car} - \text{cars}$$

$$\text{king} - \text{man} + \text{woman} \approx \text{queen}$$

Mikolov et al., (2013), "Linguistic Regularities in Continuous Space Word Representations" (NAACL)

Science AAAS

REPORT

Semantics derived automatically from language corpora contain human-like biases

Aylin Caliskan[1,*], Joanna J. Bryson[1,2,*], Arvind Narayanan[1,*]

+ See all authors and affiliations

Science 14 Apr 2017:
Vol. 356, Issue 6334, pp. 183-186
DOI: 10.1126/science.aal4230

Bias

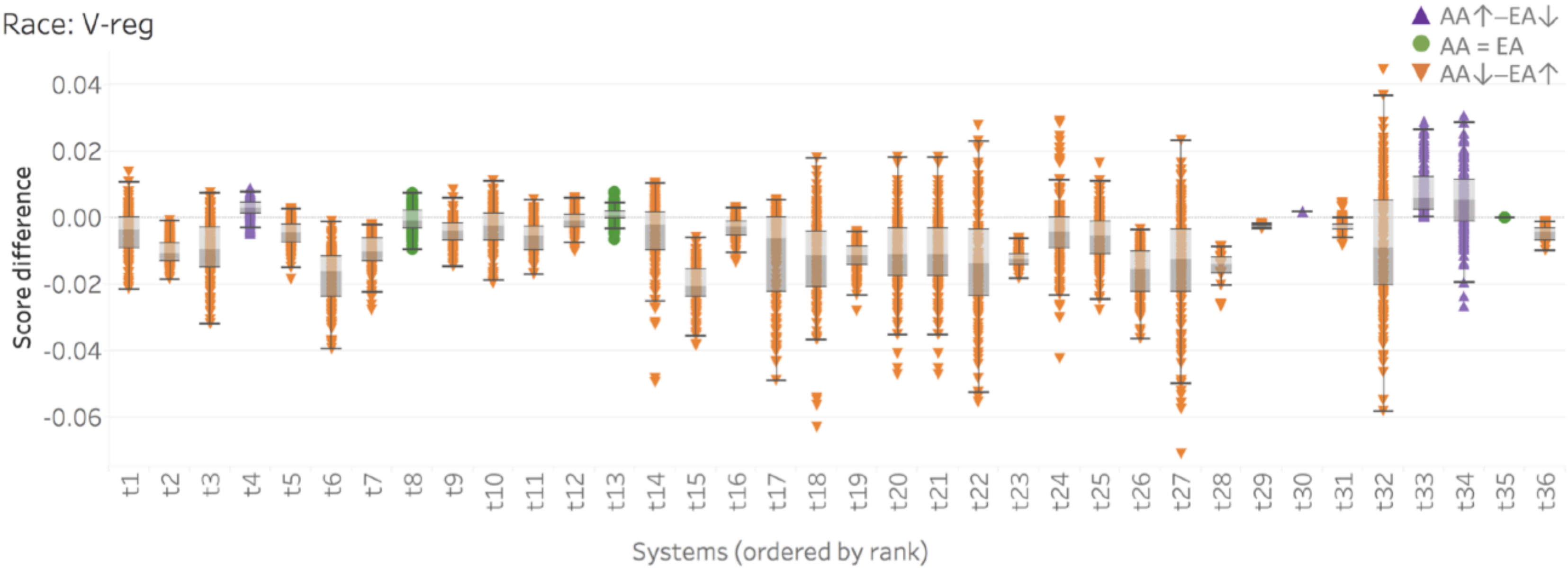
- Allocational harms: automated systems allocate resources unfairly to different groups (access to housing, credit, parole).
- Representational harms: automated systems represent one group less favorably than another (including demeaning them or erasing their existence).

Blodgett et al. (2020), “Language (Technology) is Power: A Critical Survey of “Bias” in NLP”

Representations

- **Pleasant**: caress, freedom, health, love, peace, cheer, friend, heaven, loyal, pleasure, diamond, gentle, honest, lucky, rainbow, diploma, gift, honor, miracle, sunrise, family, happy, laughter, paradise, vacation.
- **Unpleasant**: abuse, crash, filth, murder, sickness, accident, death, grief, poison, stink, assault, disaster, hatred, pollute, tragedy, bomb, divorce, jail, poverty, ugly, cancer, evil, kill, rotten, vomit.

- Embeddings for African-American first names are closer to “unpleasant” words than European-American names (Caliskan et al. 2017)

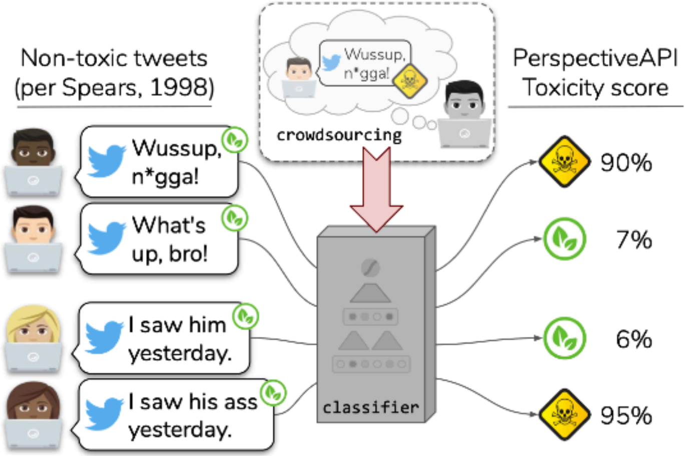


- Sentiment analysis over sentences containing African-American first names are more negative than identical sentences with European-American names.

Kiritchenko and Mohammad (2018), "Examining Gender and Race Bias in Two Hundred Sentiment Analysis Systems"

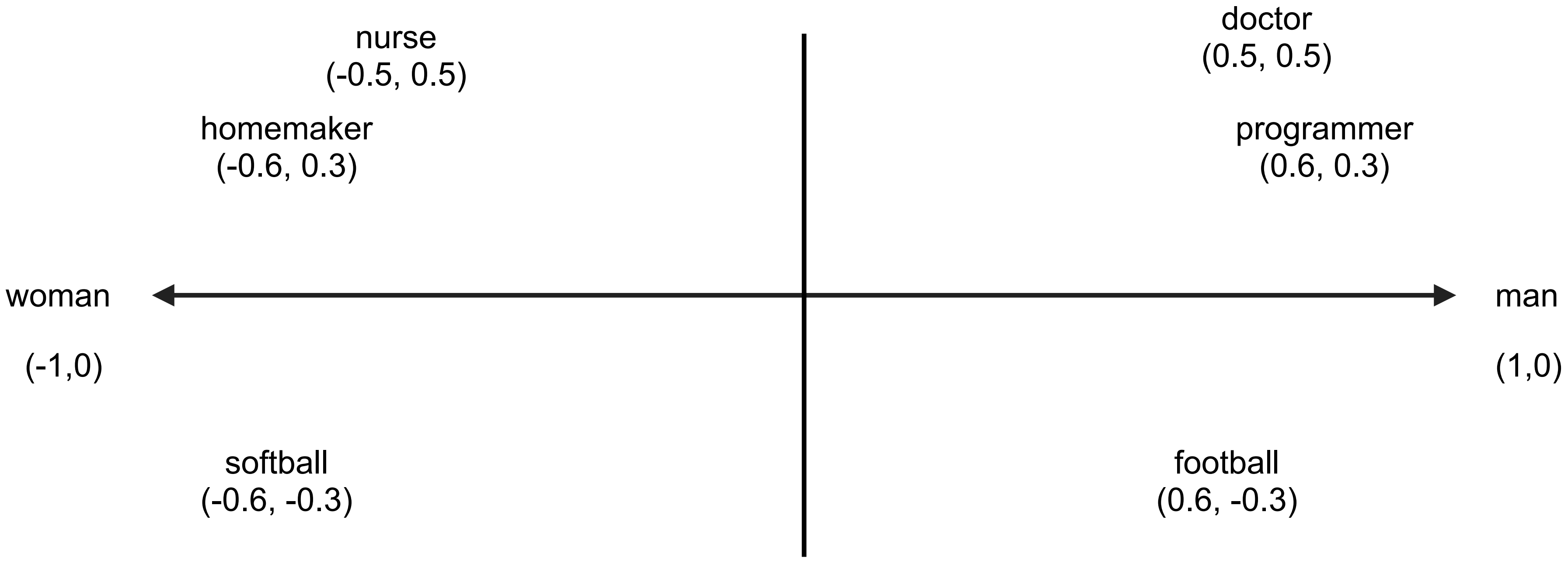
- Toxicity detection systems score text with African-American English as more offensive

- Implicit negative perception of AAE → more AAE tweets are removed → users change language practices



Blodgett et al. (2020); Sap et al. (2019), "The risk of racial bias in hate speech detection"

Bias


nurse
(-0.5, 0.5)
doctor
(0.5, 0.5)
homemaker
(-0.6, 0.3)
programmer
(0.6, 0.3)
woman
(-1,0)
man
(1,0)
softball
(-0.6, -0.3)
football
(0.6, -0.3)

Notation

$$x^\top b$$

$$x \cdot b$$

$$\mathrm{dot}(x, b)$$

$$\sum_i x_i b_i$$

$$x = [3,\ 1,\ 2]$$
$$y = [0,\ 5,\ 2]$$

$$x \cdot y =$$

$$3 \times 0 + 1 \times 5 + 2 \times 2$$

$$= 9$$

Cosine Similarity

$$cos(x, y) = \frac{\sum_{i=1}^{F} x_i y_i}{\sqrt{\sum_{i=1}^{F} x_i^2}\sqrt{\sum_{i=1}^{F} y_i^2}}$$

Cosine similarity

$$\cos(x, y) = \frac{\text{dot}(x, y)}{\sqrt{\text{dot}(x, x)} \times \sqrt{\text{dot}(y, y)}}$$

Cosine similarity

$$\cos(x, y) = \frac{\text{dot}(x, y)}{\boxed{\sqrt{\text{dot}(x, x)} \times \sqrt{\text{dot}(y, y)}}}$$

This part can be done ahead of time by normalizing all vectors:

$$v = \frac{v}{\sqrt{\text{dot}(v, v)}}$$

If all vectors have been normalized in this way, cosine similarity is just the dot product:

$$\cos(x, y) = \text{dot}(x, y)$$

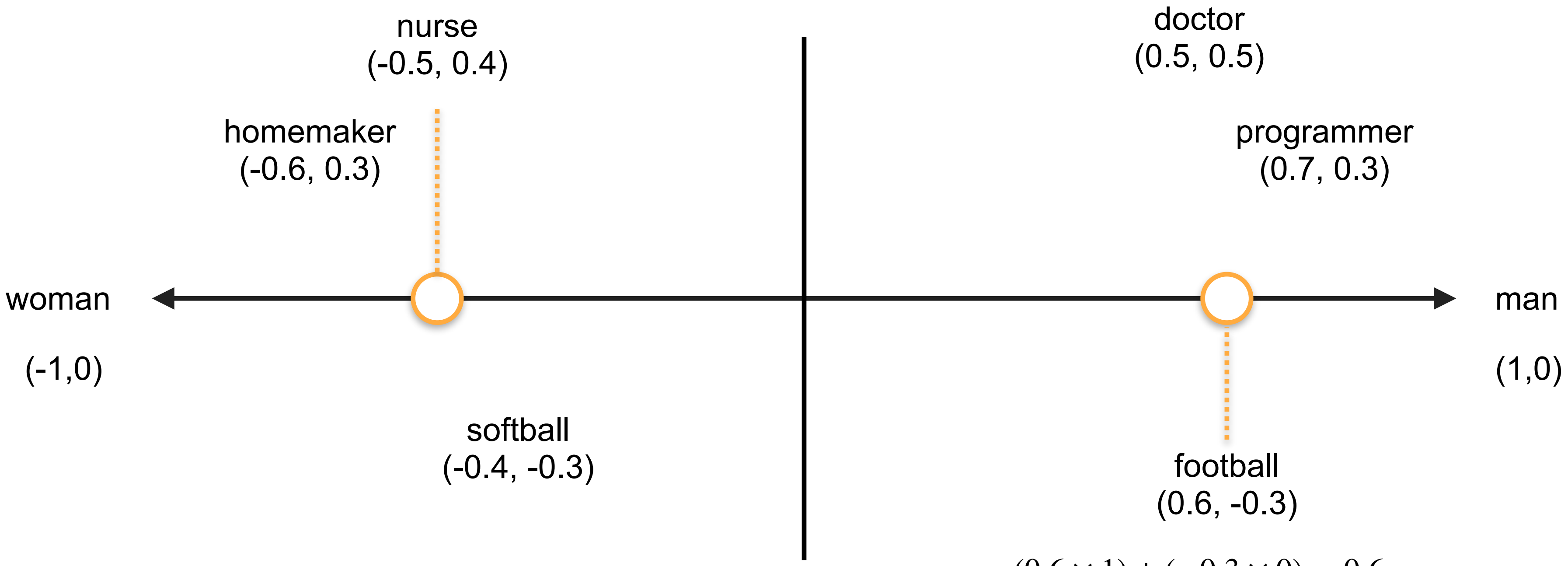
Orthogonal projection

Assume all the vectors have been normalized to unit length

$$v = \frac{v}{\sqrt{\mathrm{dot}(v, v)}}$$

$$x_b = \mathrm{dot}(x, b)\ b$$

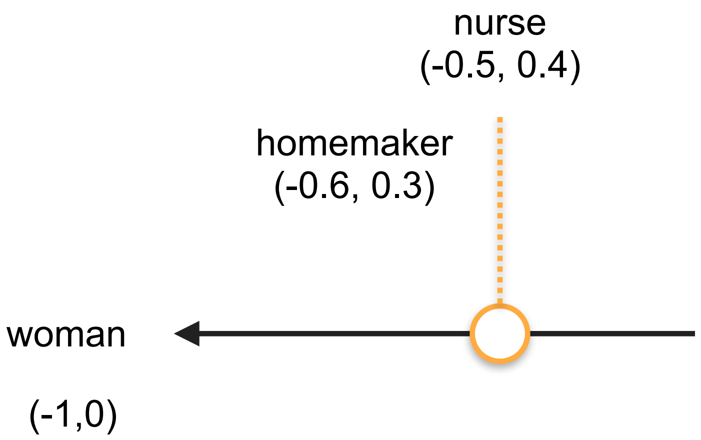
$$x_b = (x^\top b)\, b$$



$$(0.6 \times 1) + (-0.3 \times 0) = 0.6$$

$$0.6 \times [1,0] = [0.6,0]$$

Orthogonal decomposition

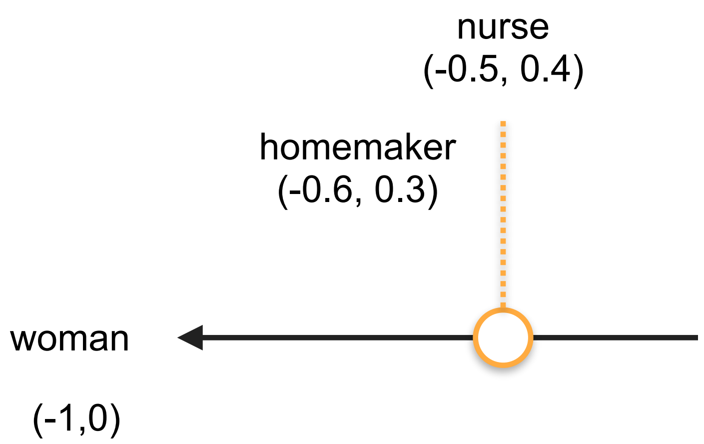


Orthogonal projection w_b of “nurse” onto gender subspace = (-0.5, 0)

$$w = w_b + w_{b^\perp}$$

$$\begin{bmatrix} -0.5 \\ 0.4 \end{bmatrix} = \underbrace{\begin{bmatrix} -0.5 \\ 0 \end{bmatrix}}_{\text{gender part}} + \underbrace{\begin{bmatrix} 0 \\ 0.4 \end{bmatrix}}_{\text{everything else}}$$

Debiasing



$$w_d = w_{b^\perp}$$
$$= w - w_b$$

$$\begin{bmatrix} -0.5 \\ 0.4 \end{bmatrix} = \underbrace{\begin{bmatrix} -0.5 \\ 0 \end{bmatrix}}_{\text{gender part}} + \underbrace{\begin{bmatrix} 0 \\ 0.4 \end{bmatrix}}_{\text{everything else}}$$

Bias

- The last slides illustrate this with a simple 2D subspace (where gender is effectively a 1D line).

- But the same principle (and procedure applies to any dimensionality (e.g., word embeddings of 100 dimensions).

projection onto gender subspace

$$x_b = (x^\top b)\, b$$

debiasing by subtracting gender projection

$$x_d = x - x_b$$

What's the gender subspace?

- Caliskan et al. 2018 construct this by first creating **defining sets** of gendered terms, e.g.
  - D_1 = {man, woman}
  - D_2 = {he, she}
- Performing SVD over a covariance matrix within over all terms in the defining sets (mean-normalized)
- And defining a gender subspace to be the first row of the resulting SVD.

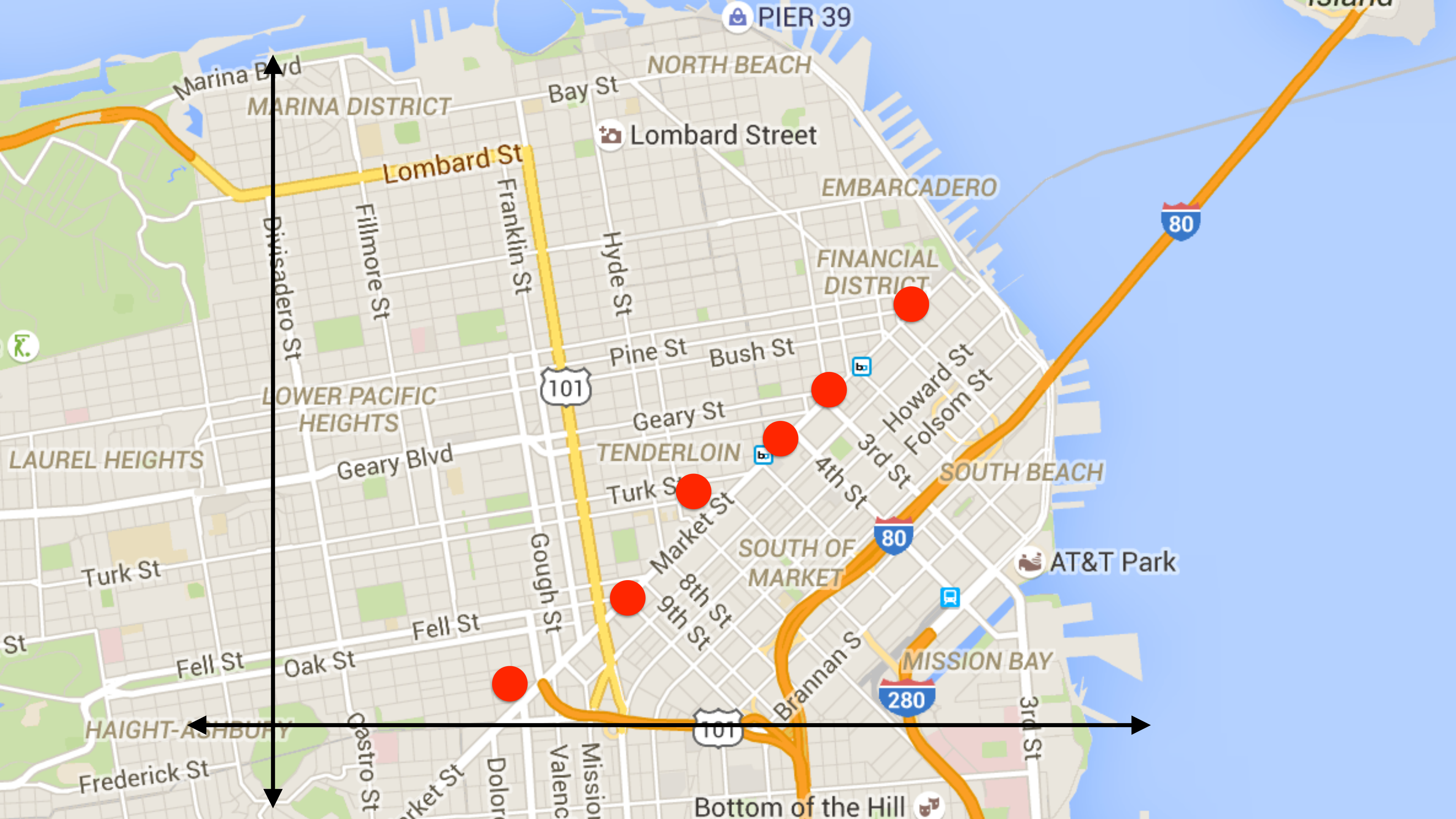
Gender subspace

Vargas and Cotterell (2020) show that this is equivalent to PCA over the following matrix →

If each embedding is 100 dimensions, this matrix is [4 x 100] in size.

The gender subspace is then the first principle component (a 100-dimensional vector in this scenario).

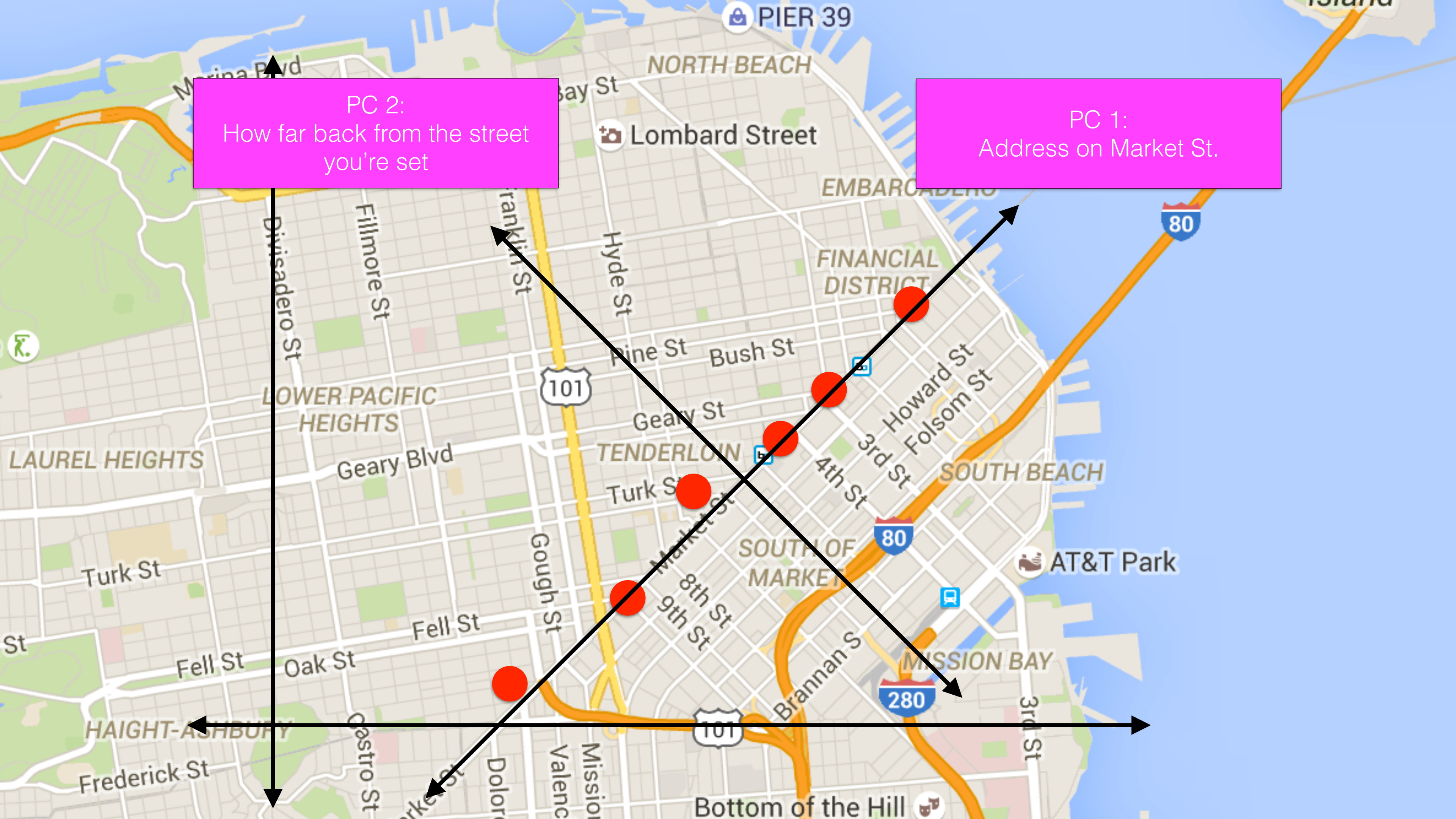
man-mean(man, woman)
woman-mean(man, woman)
he-mean(he, she)
she-mean(he, she)



PIER 39
NORTH BEACH
Marina Blvd
MARINA DISTRICT
Bay St
Lombard Street
Lombard St
EMBARCADERO
Divisadero St
Fillmore St
Franklin St
Hyde St
FINANCIAL DISTRICT
80
Pine St
Bush St
101
Howard St
Folsom St
LOWER PACIFIC HEIGHTS
Geary St
LAUREL HEIGHTS
Geary Blvd
TENDERLOIN
3rd St
4th St
SOUTH BEACH
Turk St
Market St
80
SOUTH OF MARKET
AT&T Park
Turk St
Gough St
8th St
9th St
Fell St
Fell St
Oak St
Brannan S
MISSION BAY
280
HAIGHT-ASHBURY
101
Castro St
3rd St
Frederick St
Bottom of the Hill

Principal Component Analysis

- Method for transforming a set of original (possible correlated) observations into new (uncorrelated) values.


PC 2:
How far back from the street you're set
PC 1:
Address on Market St.
PIER 39
NORTH BEACH
Bay St
Lombard Street
EMBARCADERO
FINANCIAL DISTRICT
Divisadero St
Fillmore St
Franklin St
Hyde St
Pine St
Bush St
Geary St
TENDERLOIN
Turk St
Market St
LOWER PACIFIC HEIGHTS
LAUREL HEIGHTS
Geary Blvd
Howard St
Folsom St
3rd St
4th St
SOUTH BEACH
SOUTH OF MARKET
8th St
9th St
Gough St
Turk St
Fell St
Fell St
Oak St
AT&T Park
Brannan St
MISSION BAY
HAIGHT-ASHBURY
Frederick St
Castro St
Dolores
Valencia
Mission
3rd St
Bottom of the Hill
101
80
280

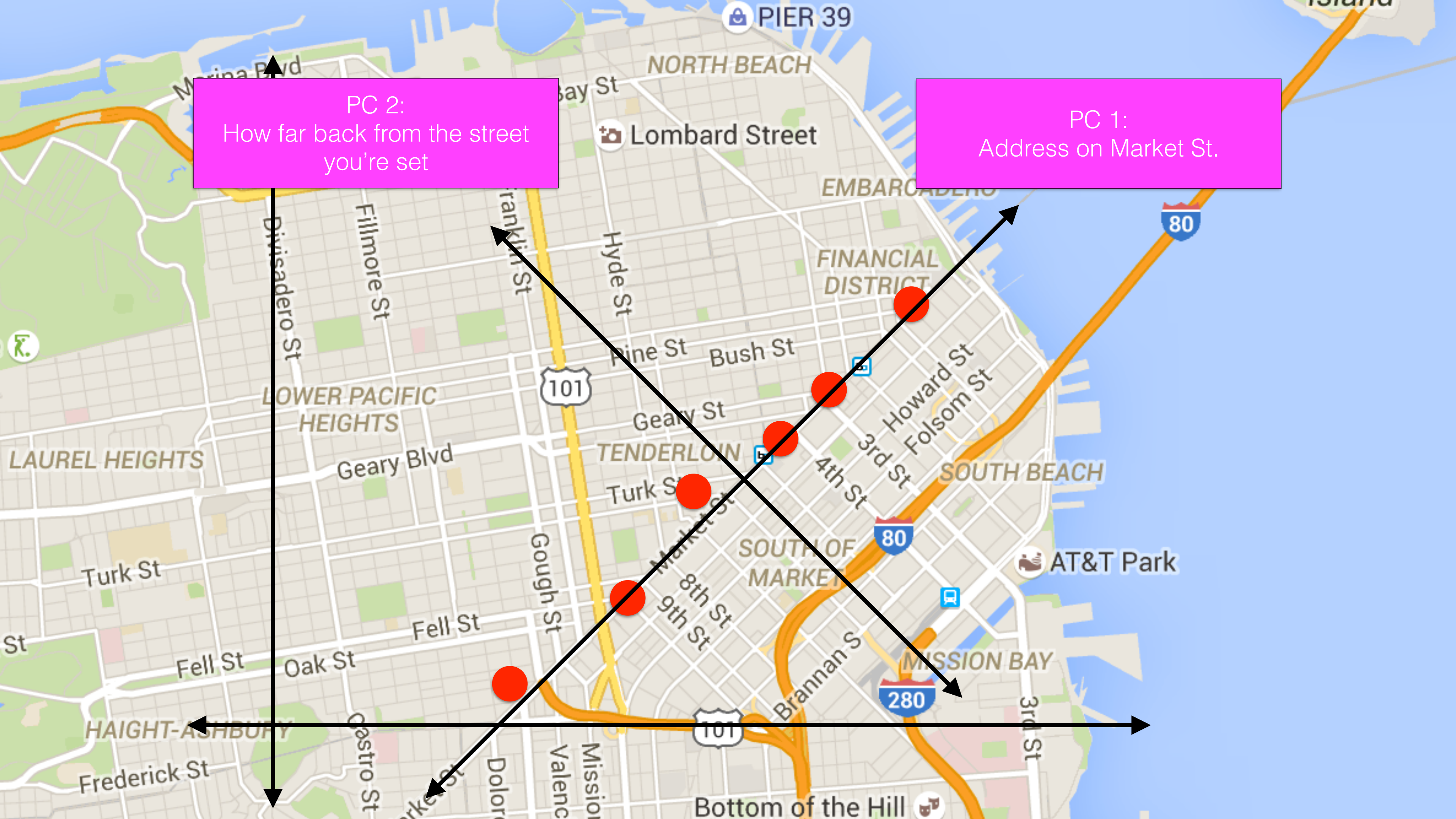
- Original values: latitude and longitude (very strong correlation for these data points)

- Transformed values: street address and distance from street (no correlation)

Main idea

- Each principal component (1 … F) is the axis that exhibits them most variance in the data and is uncorrelated (orthogonal) with earlier PCs

- The first PC explains the most variance; the second PC explains the most remaining variance, etc.


PC 2:
How far back from the street you're set
PC 1:
Address on Market St.
PIER 39
NORTH BEACH
Bay St
Lombard Street
EMBARCADERO
FINANCIAL DISTRICT
Fillmore St
Divisadero St
Hyde St
Pine St
Bush St
LOWER PACIFIC HEIGHTS
Geary St
TENDERLOIN
Geary Blvd
LAUREL HEIGHTS
Turk St
Howard St
Folsom St
3rd St
4th St
SOUTH BEACH
Market St
8th St
9th St
SOUTH OF MARKET
AT&T Park
Gough St
Fell St
Oak St
Brannan St
MISSION BAY
HAIGHT-ASHBURY
Frederick St
Castro St
Bottom of the Hill
101
80
280

Gender subspace

Vargas and Cotterell (2020) show that this is equivalent to PCA over the following matrix →

If each embedding is 100 dimensions, this matrix is [4 x 100] in size.

The gender subspace is then the first principle component (a 100-dimensional vector in this scenario).

man-mean(man, woman)
woman-mean(man, woman)
he-mean(he, she)
she-mean(he, she)

The **mean** of “man” and “woman” captures information that is common to both terms/embeddings (e.g. being people, animate, etc.). The difference is what’s left over to be explained.

- Ryan Heuser (2017), “Word Vectors in the Eighteenth Century” *DH*

SemAxis

- Define a set of terms that comprise the endpoints of an axis of interest and average them up to form axis endpoint vectors.

$$S^- = \{v_1^-, \ldots, v_n^-\}$$

{man, he, mr.}

$$V^- = \frac{1}{n}\sum_1^N v_i^-$$

$$S^+ = \{v_1^+, \ldots, v_m^+\}$$

{woman, she, miss, mrs.}

$$V^+ = \frac{1}{M}\sum_1^M v_i^+$$

An et al. 2018, "SemAxis: A Lightweight Framework to Characterize Domain-Specific Word Semantics Beyond Sentiment"

SemAxis

- The axis vector is then the difference between the two endpoint vectors

{man, he, mr.}

{woman, she, miss, mrs.}

$$V^{-} = \frac{1}{n}\sum_{1}^{N} v_i^{-}$$

$$V^{+} = \frac{1}{M}\sum_{1}^{M} v_i^{+}$$

$$V_{\text{axis}} = V^{+} - V^{-}$$

An et al. 2018, “SemAxis: A Lightweight Framework to Characterize Domain-Specific Word Semantics Beyond Sentiment”

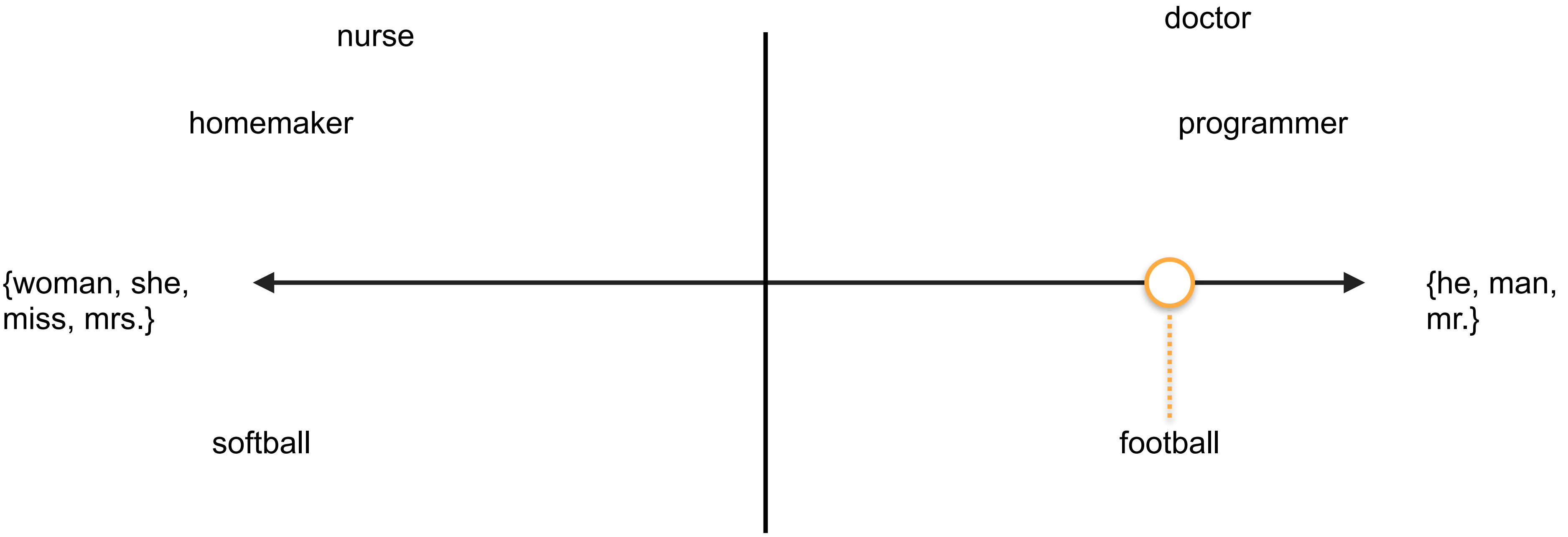
SemAxis

- For any vector, we can find its position along this axis by taking the cosine similarity with it (or dot product if all the vectors are normalized to unit length)

$$\text{Semaxis score} = \cos\left(\text{football}, v_{\text{axis}}\right)$$

An et al. 2018, “SemAxis: A Lightweight Framework to Characterize Domain-Specific Word Semantics Beyond Sentiment”

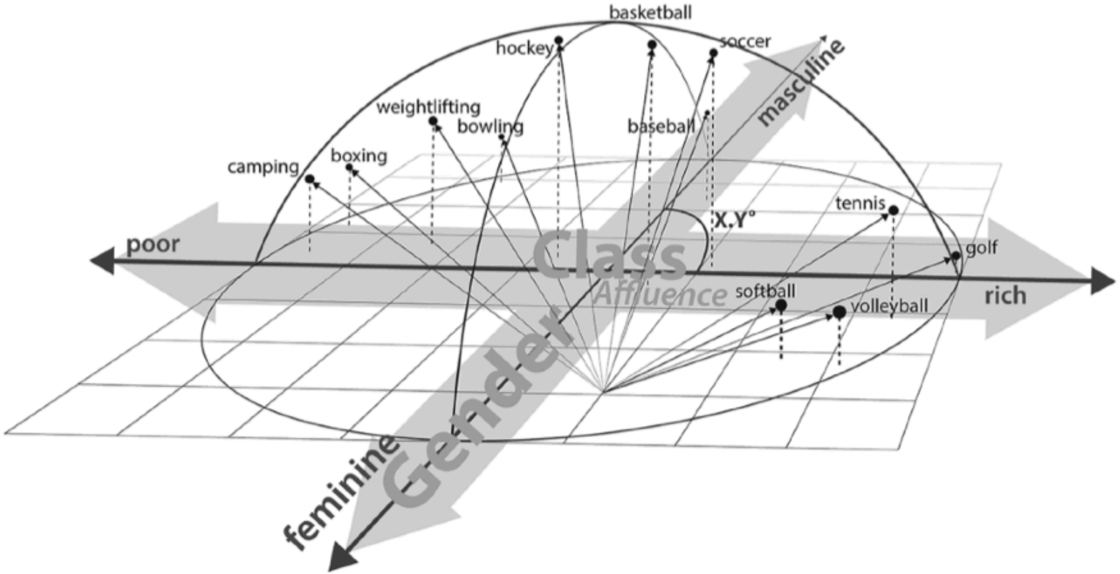
$$x_b = (x^\top b)\, b$$

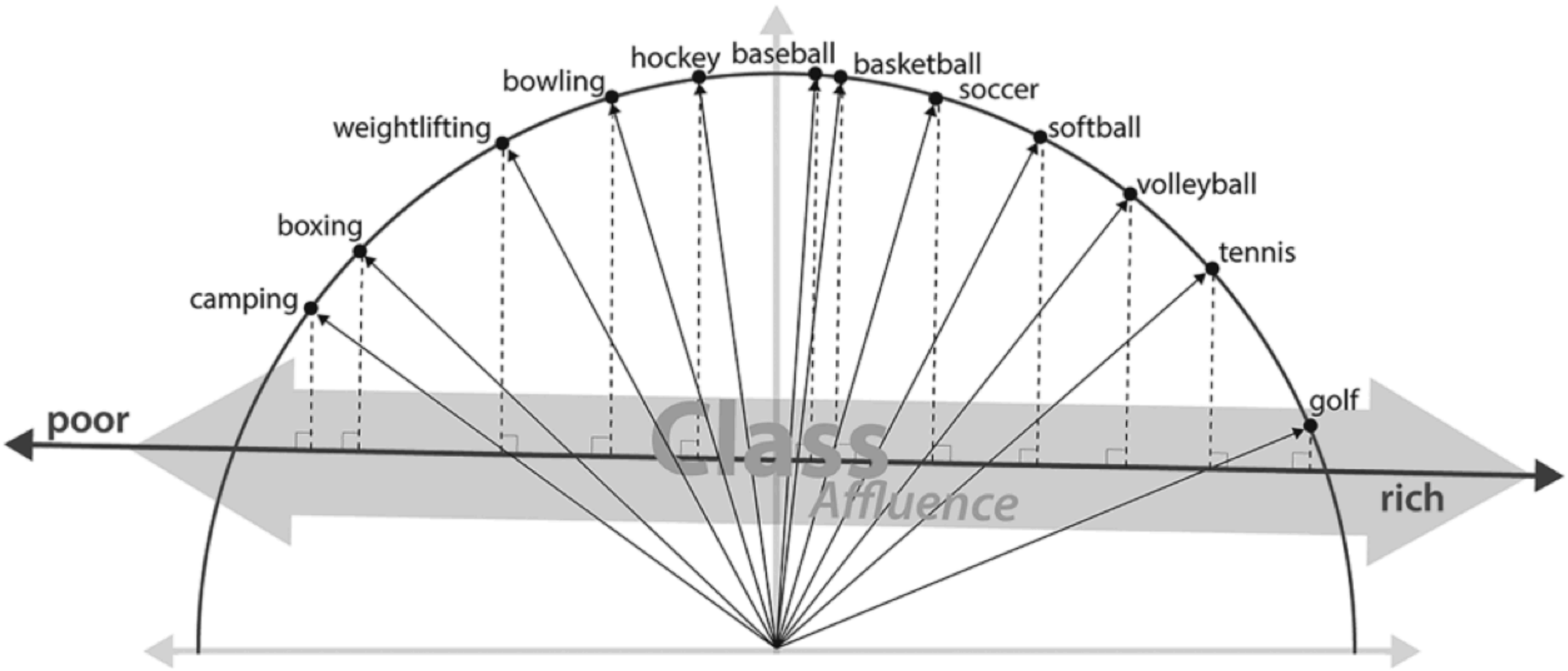

nurse
doctor
homemaker
programmer
{woman, she, miss, mrs.}
{he, man, mr.}
softball
football

Interrogating “bias”

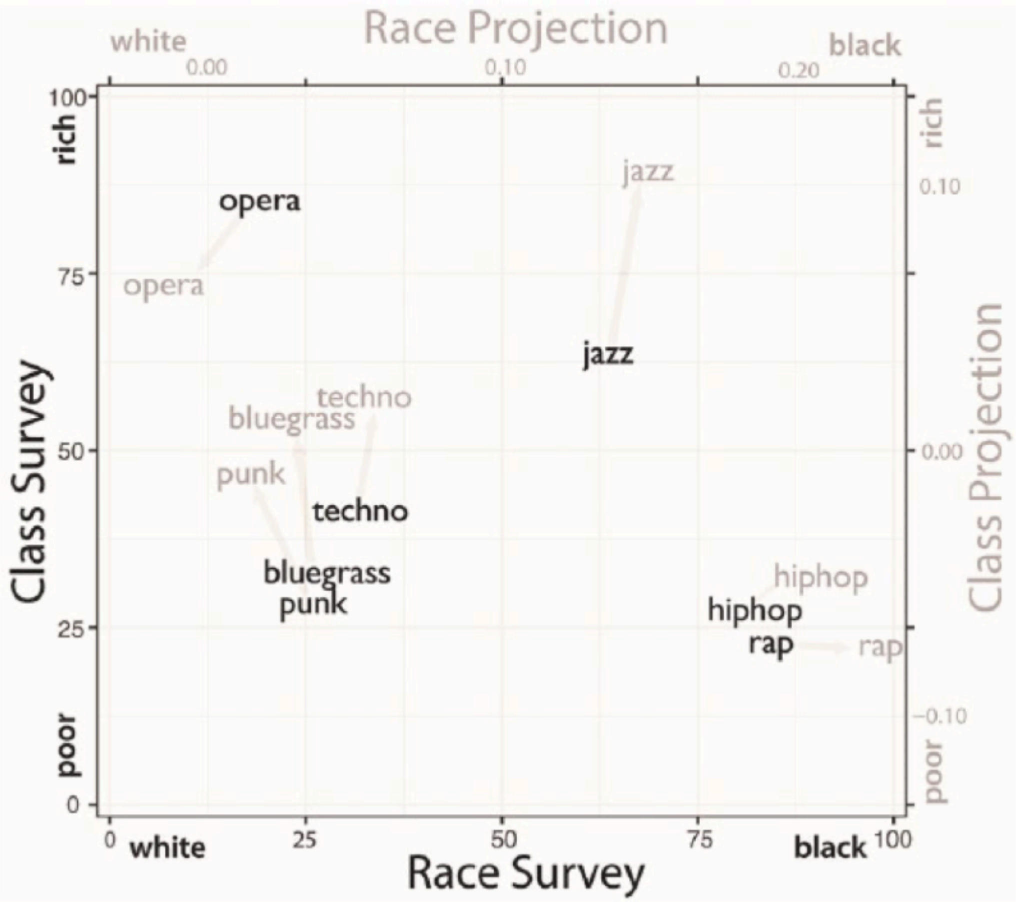
- Kozlowski et al. (2019), “The Geometry of Culture: Analyzing the Meanings of Class through Word Embeddings,” *American Sociological Review*.

- An et al. 2018, “SemAxis: A Lightweight Framework to Characterize Domain-Specific Word Semantics Beyond Sentiment”





Kozlowski et al. (2019); conceptual diagram (not real data)



Kozlowski et al. (2019)

Table D2. Word Pairs Used to Reconstruct 20 Semantic Differential Dimensions from Jenkins and Colleagues (1958) for Historical Survey Validation

soft-hard	**foolish-wise**	**unimportant-important**	**fast-slow**
supple-tough	dumb-smart	inconsequential-consequential	quick-lagging
delicate-dense	irrational-rational	secondary-principal	rapid-unhurried
pliable-rigid	stupid-thoughtful	irrelevant-major	speedy-sluggish
fluffy-firm	unwise-sensible	trivial-crucial	swift-gradual
mushy-solid	silly-reasonable	negligible-critical	quickly-slowly
softer-harder	ridiculous-enlightened	insignificant-significant	swiftly-gradually
softest-hardest	unintelligent-intelligent	unnecessary-essential	faster-slower
		peripheral-central	fastest-slowest
unusual-usual	**excitable-calm**	**strong-weak**	**colorful-colorless**
different-customary	volatile-tranquil	powerful-powerless	brilliant-uncolored
abnormal-normal	nervous-still	muscular-frail	bright-pale
irregular-regular	tempestuous-serene	brawny-feeble	radiant-drab
odd-standard	fiery-peaceful	strapping-puny	vivid-pallid
atypical-typical	emotional-restful	sturdy-fragile	vibrant-lackluster
unexpected-expected	jumpy-sedate	robust-flimsy	colored-bleached
unconventional-conventional	unsettled-settled	vigorous-languid	
rounded-angular	**passive-active**	**true-false**	**ugly-beautiful**
circular-cornered	immobile-mobile	true-untrue	unattractive-attractive
round-pointed	lethargic-energetic	verifiable-erroneous	unsightly-pretty
dull-sharp	frail-vital	veracious-fallacious	hideous-handsome
smooth-jagged	subdued-vigorous	accurate-inaccurate	grotesque-gorgeous
spherical-edged	static-dynamic	faithful-fraudulent	repulsive-cute
	subdued-lively	correct-incorrect	

Activity

- `SemAxis_TODO`: Implement the SemAxis method to define a conceptual axis using word embeddings and situate any word along that axis.
- Brainstorm other axes