Representing words as discrete symbols

- In the methods studied so far we have regarded words as discrete symbols: hotel vs. motel

Means one 1, the rest 0s

- Words can be represented by one-hot vectors

$$\begin{aligned} \text{motel} &= [0\,0\,0\,0\,0\,0\,0\,0\,1\,0\,0\,0\,0\,0\,0\,0] \\ \text{hotel} &= [0\,0\,0\,0\,1\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0] \end{aligned}$$

- Vector dimension = number of words in vocabulary (e.g. 500,000)

Problem with words as discrete symbols

- Example: in web search, if user searches for "Florence motel", we would like to match documents containing "Florence hotel"
- But:

$$
\begin{aligned}
\text{motel} &= [0\,0\,0\,0\,0\,0\,0\,0\,1\,0\,0\,0\,0\,0\,0] \\
\text{hotel} &= [0\,0\,0\,0\,1\,0\,0\,0\,0\,0\,0\,0\,0\,0\,0]
\end{aligned}
$$

- These two vectors are orthogonal
- There is no natural notion of similarity for one-hot vectors
- Solution:
  - Encode similarity in the vectors themselves

Representing words by their context

- Core idea: A word's meaning is given by the words that frequently appear close-by
  - *"You shall know a word by the company it keeps"* (J.R. Firth 1957)
  - One of the most successful ideas of modern statistical NLP
- When a word w appears in a text, its context is the set of words that appear nearby (within a fixed-size window)
- Use the many contexts of w to build up a representation of w

...government debt problems turning into	*banking*	*crises as happened in 2009 ...*
...saying that Europe needs unified	*banking*	*regulation to replace the hodgepodge ...*
...India has just given its	*banking*	*system a shot in the arm ...*

These context words will represent **banking**

Word vectors

- We will build a dense vector for each word, chosen so that it is similar to vectors of words that appear in similar contexts

$$\text{linguistics} = \begin{pmatrix} 0.286 \\ 0.792 \\ -0.177 \\ -0.107 \\ 0.109 \\ -0.542 \\ 0.349 \\ 0.271 \end{pmatrix}$$

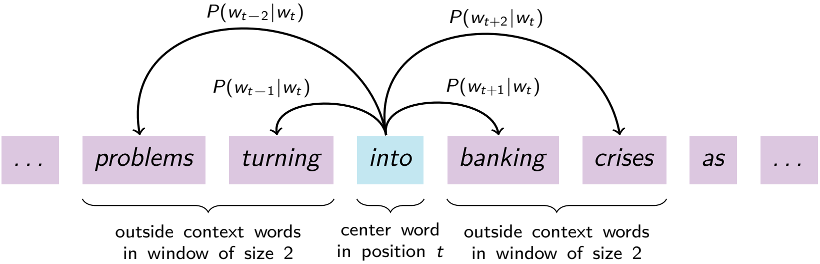
- Note: word vectors are sometimes called word embeddings or word representations

word2vec: overview

- word2vec (Mikolov et al. 2013) is a framework for learning word vectors
- Idea:
  - We have a large corpus of text
  - Every word in a fixed vocabulary is represented by a vector
  - Go through each position t in the text, which has a center word c and context ('outside') words o
  - Use the similarity of the word vectors for c and o to calculate the probability of o given c (or vice versa)
  - Keep adjusting the word vectors to maximize this probability

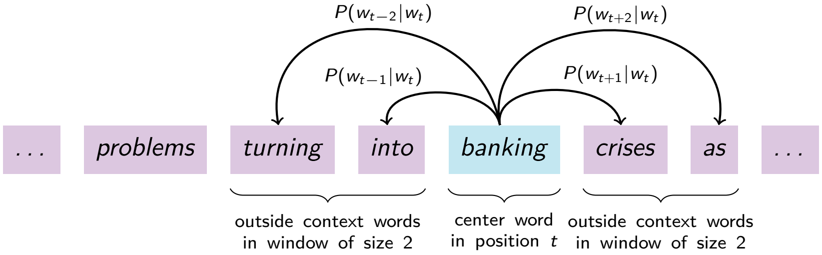
word2vec: overview

- Example windows and process for computing $P(w_{t+j}|w_t)$



word2vec: overview

- Example windows and process for computing $P(w_{t+j}|w_t)$



word2vec: objective function

- For each position $t = 1, \ldots, T$, predict context words within a window of fixed size m, given center word w_j

$$L(\theta) = \prod_{t=1}^{T} \prod_{-m \le j \le m} P\left(w_{t+j} \mid w_t; \theta\right)$$

  - θ is all the variables to be optimized

- The objective function $J(\theta)$ is the average negative log likelihood:

$$J(\theta) = -\frac{1}{T} \log L(\theta) = -\frac{1}{T} \sum_{t=1}^{T} \sum_{m \le j \le m} \log P\left(w_{t+j} \mid w_t; \theta\right)$$

  - Minimizing objective function $\Leftrightarrow$ Maximizing predictive accuracy

word2vec: objective function

- We want to minimize the objective function

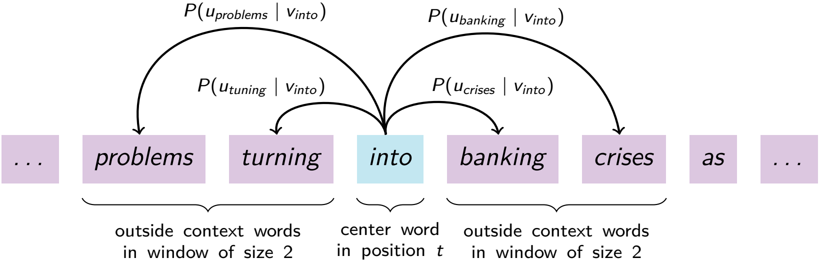
$$J(\theta) = -\frac{1}{T}\sum_{t=1}^{T}\sum_{m \leq j \leq m} \log P\left(w_{t+j} \mid w_t; \theta\right)$$

- Question: How to calculate $P\left(w_{t+j} \mid w_t; \theta\right)$
- Answer: We will use two vectors per word w:
  - v_w when w is a center word
  - u_w when w is a context word
- Then for a center word c and a context word o:

$$P(o \mid c) = \frac{\exp\left(u_o^T v_c\right)}{\sum_{w \in V} \exp\left(u_w^T v_c\right)}$$

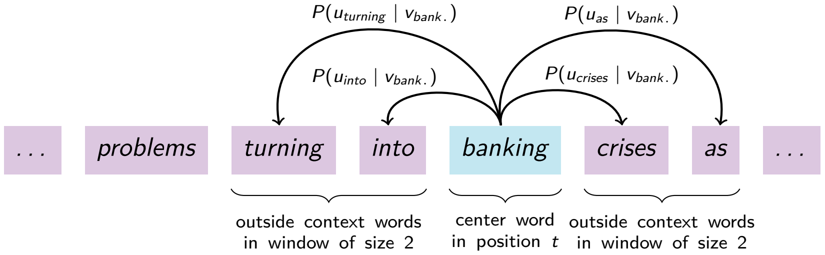
word2vec: overview with vectors

- Example windows and process for computing $P(w_{t+j} \mid w_t)$
- $P(u_{problems} \mid v_{into})$ is short for $P(problems|into, u_{problems}, v_{into}, \theta)$



word2vec: overview with vectors

- Example windows and process for computing $P(w_{t+j} \mid w_t)$



word2vec: prediction function

$$P(o \mid c) = \frac{\exp\left(u_o^T v_c\right)}{\sum_{w \in V} \exp\left(u_w^T v_c\right)}$$

- $u_o^T v_c$: dot product compares similarity of o and c. Larger dot product implies a larger probability
- $\sum_{w \in V} \exp\left(u_w^T v_c\right)$: after taking exponent, normalize over entire vocabulary
- This is an example of the softmax function $\mathbb{R}^n \to \mathbb{R}^n$
- The softmax function maps arbitrary values x_i to a probability distribution p_i
  - "max" because amplifies probability of largest x_i
  - "soft" because still assigns some probability to smaller x_i

To train the model: compute all vector gradients

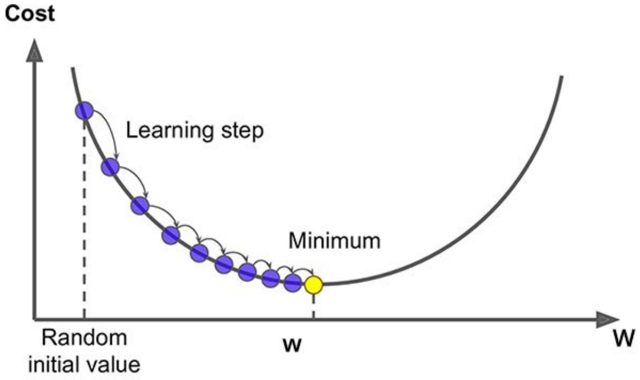
- Recall: θ represents all model parameters, in one long vector
- In our case with d-dimensional vectors and V-many words:

$$\theta = \begin{bmatrix} v_{aardvark} \\ v_a \\ \vdots \\ v_{zebra} \\ u_{aardvark} \\ u_a \\ \vdots \\ u_{zebra} \end{bmatrix} \in \mathbb{R}^{2dV}$$

- Recall: every word has two vectors
- We then optimize these parameters using stochastic gradient descent

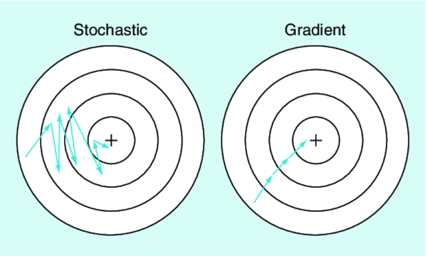
word2vec: gradient descent

- We have a cost function $J(\theta)$ we want to minimize
- Idea: for the current value of θ, calculate the gradient of $J(\theta)$, then take a small step in the direction of the negative gradient. Repeat.



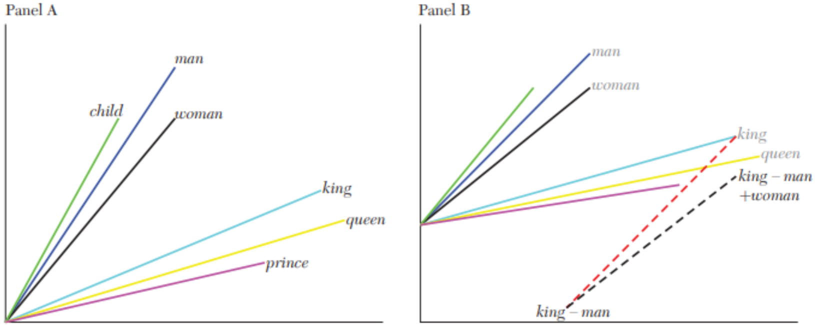
word2vec: stochastic gradient descent

- Problem: $J(\theta)$ is a function of all windows in the corpus
  - $\nabla_\theta J(\theta)$ may be very expensive to compute
- You would wait a very long time before making a single update. A very bad idea for pretty much all neural nets
- Solution: stochastic gradient descent (SGD)
  - Repeatedly sample windows, and update after each one



word2vec: applications

- Once the vectors are constructed, they can be used to represent relations between words



Garg, Schiebinger, Jurafsky, and Zou (2018)

- Goal: develop a systematic framework to analyze word embeddings trained over a century of text data to identify historical patterns of bias and stereotype changes in the US
- Motivation: in word-embedding models, words are assigned to a high-dimensional vector in a way that they capture relationships not found through simple co-occurrence analysis
- Idea: exploit differences in Euclidean distance between ethnic-gender terms and professions-stereotypes words to quantify historical trends
- Main findings: the embedding captures societal shifts and sheds light on how specific adjectives and occupations became more closely associated with certain populations over time

GSJZ (2018): data

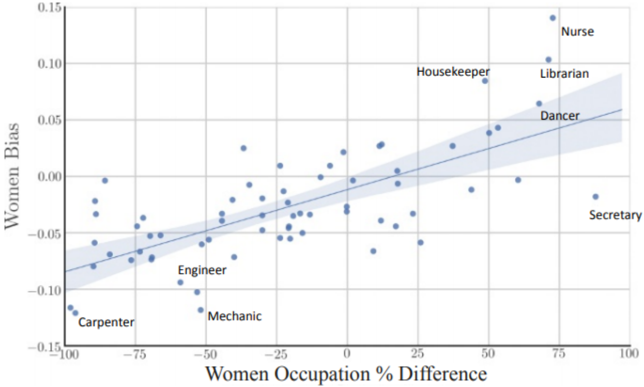
- Word embeddings:
  - word2vec embeddings trained on the Google News dataset
  - Nine decade-specific embeddings trained on text from the Corpus of Historical American English
- Word lists:
  - Gender: *he, she, son, daughter, male, female, boy, girl, etc.*
  - Ethniticty: *harris, ruiz, cho, thompson, gomez, lin, etc.*
  - Occupations: *janitor, teacher, shoemaker, scientist, carpenter, etc.*
  - Adjectives: *headstrong, inventive, enterprising, poised, moody, etc.*

GSJZ (2018): methodology

- Measure the strength of association between occupations or adjectives (neutral words) and a gender or ethnicity
  1. Compute the average vector representation of a gender or ethnic group
  2. Calculate the average Euclidean distance between the representative vector and each vector in a list of neutral words
  3. Use the difference of the average distance between gender or ethnicity pairs as a measure of embedding bias
- e.g. the occupational embedding bias for women
  1. Compute average embedding distance between words *she, female* and occupational words *teacher, lawyer*. Repeat for words he, male
  2. Compute the difference in average distances between group pair

$$\text{relative norm distance } = \sum_{v_m \in M} \|v_m - v_1\|_2 - \|v_m - v_2\|_2$$

GSJZ (2018): gender bias



- Occupation difference as the relative percentage of women in each occupation using data from the Integrated Public Use Microdata Series

GSJZ (2018): gender bias

Occupations		**Adjectives**	
Man	Woman	Man	Woman
carpenter	nurse	honorable	maternal
mechanic	midwife	ascetic	romantic
mason	librarian	amiable	submissive
blacksmith	housekeeper	dissolute	hysterical
retired	dancer	arrogant	elegant
architect	teacher	erratic	caring
engineer	cashier	heroic	delicate
mathematician	student	boyish	superficial
shoemaker	designer	fanatical	neurotic
physicist	weaver	aimless	attractive

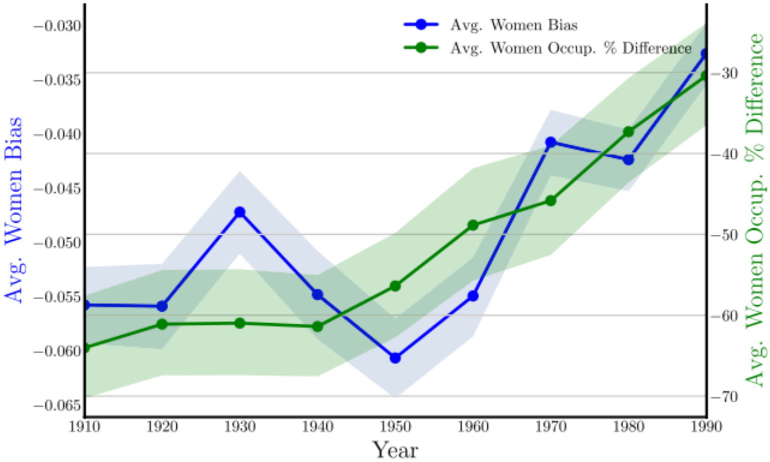
Table B.1: Top occupations and adjectives by gender in the Google News embedding.

GSJZ (2018): gender bias

Table B.6: Most Man and Woman *residuals* when regressing embedding bias vs occupation gender proportion. The more Woman (Man) a residual the more biased the embedding is toward Women (Men) above what occupation proportions would suggest.

Man	Woman
secretary	nurse
mechanic	housekeeper
musician	gardener
architect	clerk
janitor	librarian
carpenter	sailor
broker	judge
geologist	artist
accountant	dancer
economist	painter

GSJZ (2018): gender bias over time



GSJZ (2018): gender bias over time

1910	1920	1930	1940	1950	1960	1970	1980	1990
mathematician	accountant	engineer	surveyor	architect	lawyer	architect	architect	architect
soldier	surveyor	architect	architect	engineer	architect	engineer	auctioneer	mathematician
architect	architect	lawyer	engineer	mathematician	surveyor	judge	judge	surveyor
surveyor	lawyer	surveyor	smith	lawyer	soldier	economist	surveyor	engineer
administrator	mathematician	manager	sheriff	sheriff	engineer	soldier	sheriff	pilot
lawyer	sheriff	pilot	lawyer	postmaster	pilot	author	author	lawyer
judge	engineer	author	scientist	surveyor	scientist	surveyor	engineer	author
scientist	statistician	scientist	author	scientist	economist	administrator	broker	judge
author	mason	mathematician	economist	author	author	mason	inspector	soldier
economist	scientist	accountant	mason	soldier	mason	mathematician	police	blacksmith

Table B.27: Most Man occupations in each decade in the SGNS embeddings.

GSJZ (2018): gender bias over time

1910	1920	1930	1940	1950	1960	1970	1980	1990
nurse	nurse	nurse	nurse	nurse	nurse	nurse	nurse	nurse
attendant	housekeeper	housekeeper	attendant	housekeeper	attendant	dancer	dancer	housekeeper
housekeeper	attendant	attendant	janitor	attendant	dancer	housekeeper	attendant	midwife
cashier	dancer	dancer	housekeeper	dancer	housekeeper	attendant	housekeeper	dentist
cook	teacher	janitor	midwife	cook	photographer	conductor	midwife	student
bailiff	supervisor	midwife	dentist	gardener	midwife	dentist	statistician	dancer
porter	cook	clerical	cook	cashier	dentist	statistician	student	supervisor
operator	doctor	dentist	clerical	midwife	janitor	baker	conductor	bailiff
supervisor	dentist	cook	clergy	musician	cook	clerical	dentist	physician
clergy	mechanic	teacher	sailor	sailor	porter	sailor	supervisor	doctor

Table B.28: Most Woman occupations in each decade in the SGNS embeddings.

GSJZ (2018): gender bias over time

1910	1920	1930	1940	1950	1960	1970	1980	1990
honorable	regimental	honorable	honorable	knowledge	gallant	honorable	honorable	honorable
gallant	honorable	trusting	conservative	gallant	honorable	wise	loyal	regimental
regimental	stoic	courageous	ambitious	honorable	sage	knowledge	petty	unreliable
skillful	political	gallant	shrewd	directed	regimental	gallant	gallant	skillful
disobedient	sage	confident	regimental	regimental	knowledge	insulting	lyrical	gallant
faithful	ambitious	adventurous	knowledge	efficient	wise	trusting	honest	honest
wise	reserved	experimental	destructive	sage	conservative	honest	faithful	loyal
obedient	progressive	efficient	misguided	wise	honest	providential	obedient	wise
obnoxious	unprincipled	predatory	gallant	faithful	adventurous	modern	wise	directed
steadfast	shrewd	modern	petty	creative	efficient	regimental	hostile	courageous

Table B.29: Most Man adjectives in each decade in the SGNS embeddings.

GSJZ (2018): gender bias over time

1910	1920	1930	1940	1950	1960	1970	1980	1990
charming	charming	charming	delicate	delicate	sweet	attractive	maternal	maternal
placid	relaxed	delicate	placid	sweet	charming	maternal	attractive	morbid
delicate	delicate	soft	sweet	charming	soft	charming	masculine	artificial
passionate	amiable	hysterical	gentle	transparent	relaxed	sweet	impassive	physical
sweet	hysterical	transparent	soft	placid	attractive	caring	emotional	caring
dreamy	placid	sweet	warm	childish	placid	venomous	protective	emotional
indulgent	soft	relaxed	charming	soft	delicate	silly	relaxed	protective
playful	gentle	shy	childish	colorless	maternal	neat	charming	attractive
mellow	attractive	maternal	irritable	tasteless	indulgent	delicate	naive	soft
sentimental	sweet	smooth	maternal	agreeable	gentle	sensitive	responsive	tidy

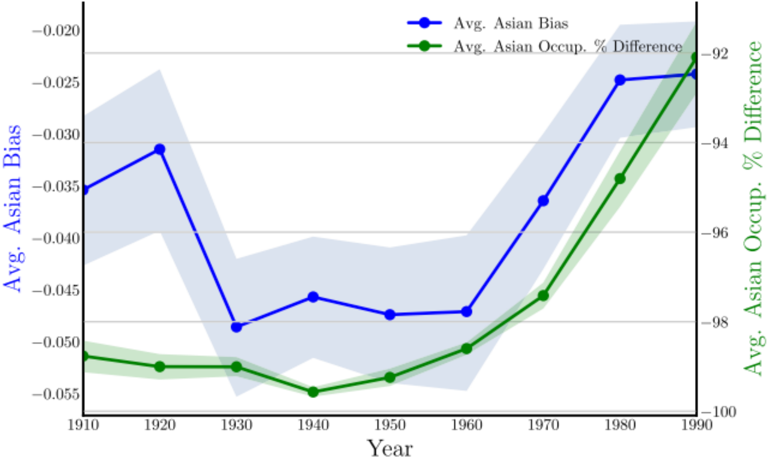
Table B.30: Most Woman adjectives in each decade in the SGNS embeddings.

GSJZ (2018): ethnic bias

Table 1. The top 10 occupations most closely associated with each ethnic group in the Google News embedding

Hispanic	Asian	White
Housekeeper	Professor	Smith
Mason	Official	Blacksmith
Artist	Secretary	Surveyor
Janitor	Conductor	Sheriff
Dancer	Physicist	Weaver
Mechanic	Scientist	Administrator
Photographer	Chemist	Mason
Baker	Tailor	Statistician
Cashier	Accountant	Clergy
Driver	Engineer	Photographer

GSJZ (2018): ethnic bias over time



GSJZ (2018): ethnic bias over time

1910	1920	1930	1940	1950	1960	1970	1980	1990
irresponsible	mellow	hateful	solemn	disorganized	imprudent	cynical	superstitious	inhibited
envious	relaxed	unchanging	reactive	outrageous	pedantic	solemn	upright	passive
barbaric	haughty	oppressed	outrageous	pompous	irresponsible	mellow	providential	dissolute
aggressive	tense	contemptible	bizarre	unstable	inoffensive	discontented	unstable	haughty
transparent	hateful	steadfast	fanatical	effeminate	sensual	dogmatic	forceful	complacent
monstrous	venomous	relaxed	assertive	unprincipled	venomous	aloof	appreciative	forceful
hateful	stubborn	cruel	unprincipled	venomous	active	forgetful	dry	fixed
cruel	pedantic	disorganized	barbaric	disobedient	inert	dominating	reactive	active
greedy	transparent	brutal	haughty	predatory	callous	disconcerting	fixed	sensitive
bizarre	compassionate	intolerant	disconcerting	boisterous	inhibited	inhibited	sensitive	hearty

Table C.8: Top Asian (vs White) Adjectives over time by relative norm difference.

Kozlowski, Taddy and Evans (2019): geometry of culture

- Motivation
  - If text represents culture, can we construct cultural dimensions of class from the dimension of word embedding vectors
  - Has the meaning of these dimensions changed over the XXth century?

Kozlowski, Taddy and Evans (2019): geometry of culture

- Motivation
  - If text represents culture, can we construct cultural dimensions of class from the dimension of word embedding vectors
  - Has the meaning of these dimensions changed over the XXth century?

- Class as the systematic and hierarchical distinction of people and groups in social standing. Dimension-specific nuances:
  - **Money:** easy to convert into various forms of power → affluence
  - **Education:** determines the labor market position → education
  - **Status:** based on authority and social position → status
  - **Cultivated taste:** based on the culture consumed → cultivation
  - **Gender:** misogynistic or patriarchal hierarchies → gender
  - **Race:** reflected in post-colonial, structural racism → race

KTE (2019): the cultural dimensions of class

- These dimensions can be represented through semantic contrasts
  - **Affluence:** rich vs poor, wealthy vs impoverished, luxury vs cheap
  - **Education:** educated vs uneducated, knowledgeable vs ignorant
  - **Status:** acclaimed vs modest, eminent vs mundane
  - **Cultivation:** civil vs uncivil, cultured vs uncultured
  - **Gender:** masculine vs feminine, he vs she, male vs female
  - **Race:** black vs white, African vs European
- Main idea: words that are opposites semantically will display systematic differences in their vector representation

KTE (2019): the cultural dimensions of class

- Intuition: solving an analogy is equivalent to projecting a word vector onto a specific dimension

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

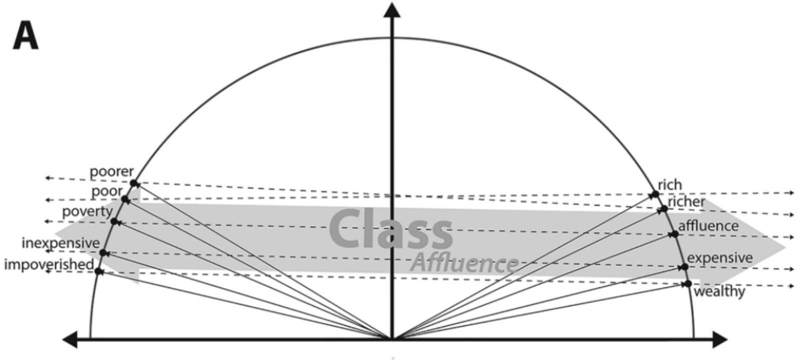
- The projection of the word vector for king onto a gender dimension captured by $\overrightarrow{\text{woman}} - \overrightarrow{\text{man}}$ yields the word vector for queen

- Collate lists of antonyms similar to $\overrightarrow{\text{woman}} - \overrightarrow{\text{man}}$ for the different dimensions of class, i.e. $\overrightarrow{\text{rich}} - \overrightarrow{\text{poor}}$

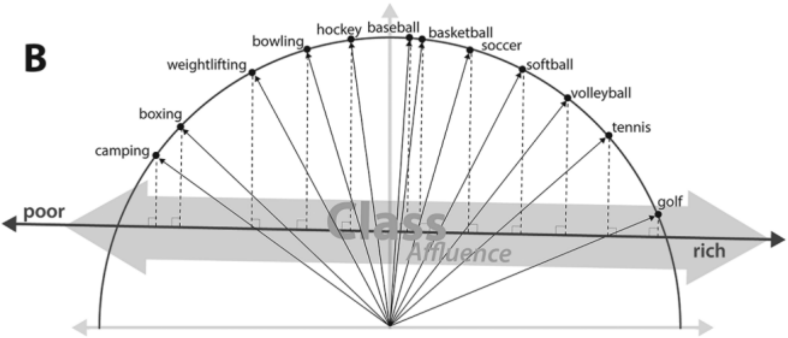
- Project words onto dimension-specific antonym lists to identify the cultural associations embedded in the word

$$\overrightarrow{\text{hockey}} + \overrightarrow{\text{rich}} - \overrightarrow{\text{poor}} \approx \overrightarrow{\text{lacrosse}}$$

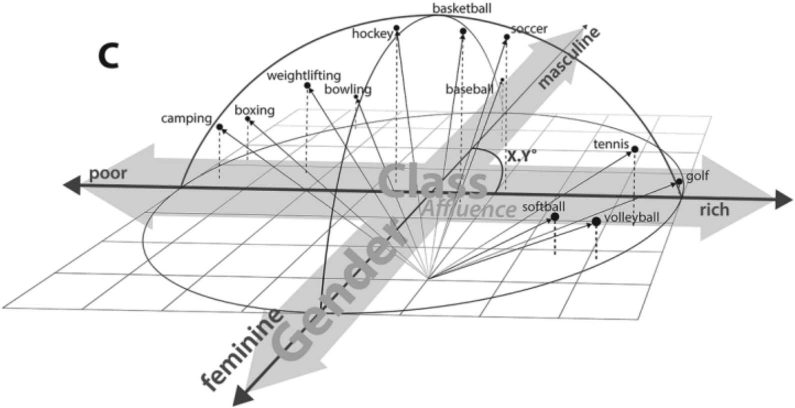
KTE (2019): the cultural dimensions of class



KTE (2019): the cultural dimensions of class



KTE (2019): the cultural dimensions of class



KTE (2019): data and methods

- Use three pre-trained word embedding models:
  - Google Ngrams US
  - Google News embeddings
  - GloVe embeddings
- Trained via Google Ngram corpus
  - 6% of all books ever published
  - Only look at 5-grams
  - Divide corpus by decades
  - Keep only words that appear $>$ 25 times
- Antonym lists compiled from five thesauri

KTE (2019): data

Table D1. Word Pairs Used to Construct Affluence, Gender, and Race Dimensions for Amazon Mechanical Turk Survey Validation

Affluence		Gender	Race
rich-poor	precious-cheap	man-woman	black-white
richer-poorer	priceless-worthless	men-women	blacks-whites
richest-poorest	privileged-underprivileged	he-she	Black-White
affluence-poverty	propertied-bankrupt	him-her	Blacks-Whites
affluent-destitute	prosperous-unprosperous	his-her	African-European
advantaged-needy	developed-underdeveloped	his-hers	African-Caucasian
wealthy-impoverished	solvency-insolvency	boy-girl	Afro-Anglo
costly-economical	successful-unsuccessful	boys-girls	
exorbitant-impecunious	sumptuous-plain	male-female	
expensive-inexpensive	swanky-basic	masculine-feminine	
exquisite-ruined	thriving-disadvantaged		
extravagant-necessitous	upscale-squalid		
flush-skint	valuable-valueless		
invaluable-cheap	classy-beggarly		
lavish-economical	ritzy-ramshackle		
luxuriant-penurious	opulence-indigence		
luxurious-threadbare	solvent-insolvent		
luxury-cheap	moneyed-moneyless		
moneyed-unmonied	rich-penniless		
opulent-indigent	affluence-penury		
plush-threadbare	posh-plain		
luxuriant-penurious	opulence-indigence		

KTE (2019): data

Table D3. Word Pairs Used to Construct Class Dimensions (Along with Affluence and Gender in Table D1)

Cultivation	Employment	Education	Status	Morality
cultivated-uncultivated	employer-employee	educated-uneducated	prestigious-unprestigious	good-evil
cultured-uncultured	employers-employees	learned-unlearned	honorable-dishonorable	moral-immoral
civilized-uncivilized	owner-worker	knowledgeable-ignorant	esteemed-lowly	good-bad
courteous-discourteous	owners-worker	trained-untrained	influential-uninfluential	honest-dishonest
proper-improper	industrialist-laborer	taught-untaught	reputable-disreputable	virtuous-sinful
polite-rude	industrialists-laborers	literate-illiterate	distinguished-commonplace	virtue-vice
cordial-uncordial	proprietor-employee	schooled-unschooled	eminent-mundane	righteous-wicked
formal-informal	proprietors-employees	tutored-untutored	illustrious-humble	chaste-transgressive
courtly-uncourtly	capitalist-proletarian	lettered-unlettered	renowned-prosaic	principled-unprincipled
urbane-boorish	capitalists-proletariat		acclaimed-modest	unquestionable-questionable
polished-unpolished	manager-staff		dignitary-commoner	noble-nefarious
refined-unrefined	managers-staff		venerable-unpretentious	uncorrupt-corrupt
civility-incivility	director-employee		exalted-ordinary	scrupulous-unscrupulous
civil-uncivil	directors-employees		estimable-lowly	altruistic-selfish
urbanity-boorishness	boss-worker		prominent-common	chivalrous-knavish
politesse-rudeness	bosses-workers			honest-crooked
edified-loutish	foreman-laborer			commendable-reprehensible
mannerly-unmannerly	foremen-laborers			pure-impure
polished-gruff	supervisor-staff			dignified-undignified
gracious-ungracious	superintendent-staff			holy-unholy
obliging-unobliging				valiant-fiendish
cultured-uncultured				upstanding-villainous
genteel-ungenteel				guiltless-guilty
mannered-unmannered				decent-indecent
polite-blunt				chaste-unsavory
				righteous-odious
				ethical-unethical

KTE (2019): methods

- For each class dimension, calculate the following:

$$\frac{\sum_{p}^{|P|} \overrightarrow{p_1} - \overrightarrow{p_2}}{|P|}$$

- p are all antonym couples in set P of relevant words by context

- The projection of a word vector onto a dimension is computed using cosine similarity

KTE (2019): methods

Validation

- Use two surveys
  - **Modern Survey:** rate 59 words on different dimensions (class, race, gender)
  - ***Historical* Survey:** rate 360 words on 20 semantic dimensions (good/bad, soft/hard, . . .)
- Example:
  - On a scale from 0 (very working class) to 100 (very upper class), how would you rate a steak?

KTE (2019): methods

Validation

Table B1. List of Words Rated in Cultural Associations Survey

Occupations	Clothing	Sports	Music Genres	Vehicles	Food	First Names
Banker	Blouse	Baseball	Bluegrass	Bicycle	Beer	Aaliyah
Carpenter	Briefcase	Basketball	Hip hop	Limousine	Cheesecake	Amy
Doctor	Dress	Boxing	Jazz	Minivan	Hamburger	Connor
Engineer	Necklace	Golf	Opera	Motorcycle	Pastry	Jake
Hairdresser	Pants	Hockey	Punk	Skateboard	Salad	Jamal
Journalist	Shirt	Soccer	Rap	SUV	Steak	Molly
Lawyer	Shorts	Softball	Techno	Truck	Wine	Shanice[a]
Nanny	Socks	Tennis				Tyrone
Nurse	Suit	Volleyball				
Plumber	Tuxedo					
Scientist						

Table B3. Percentage of Statistically Significant ($p < .01$) Survey Differences Correctly Classified in Google News Word Embedding Model

	Sports	Food	Music	Occupations	Vehicles	Clothes	Names	All Domains
Gender	87.9%	88.2%	72.2%	93.6%	82.4%	74.4%	95.2%	84.8%
Class	96.3%	93.8%	88.9%	60.9%	94.1%	90.0%	77.3%	75.3%
Race	90.0%	68.8%	100%	51.5%	87.5%	55.0%	94.7%	69.1%

KTE (2019): methods

Validation

Table 1. Pearson Correlations between Survey Estimates and Word Embedding Estimates for Gender, Class, and Race Associations

	Class (Affluence)	Gender	Race
Google Ngrams *word2vec* Embedding†	.53	.76	.27
Google News *word2vec* Embedding	.58	.88	.75
Common Crawl *GloVe* Embedding	.57	.90	.44

KTE (2019): methods

Validation

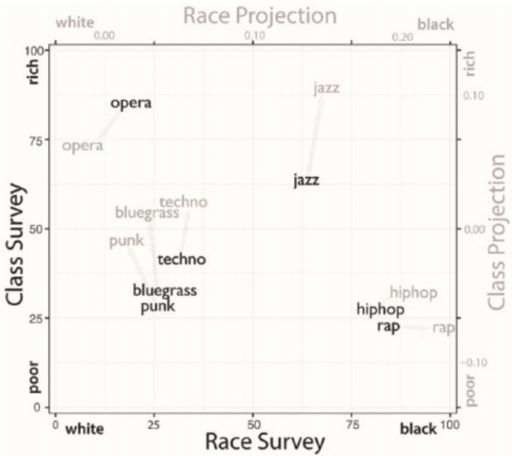


Figure 3. Projection of Music Genres onto Race and Class Dimensions of the Google News Word Embedding (Gray) and Average Survey Ratings for Race and Class Associations (Black)

KTE (2019): results

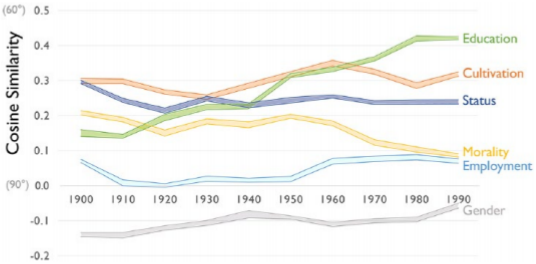


Figure 5. Cosine Similarity between the Affluence Dimension and Six Other Cultural Dimensions of Class by Decade; 1900 to 1999 Google Ngrams Corpus
Note: Bands represent 90 percent bootstrapped confidence intervals produced by subsampling.

- Education has become more synonymous with **affluence**
  - Crucial for a competitive labor market → **signaling**
  - Mediated by cultivation: when controlled, negligible correlation