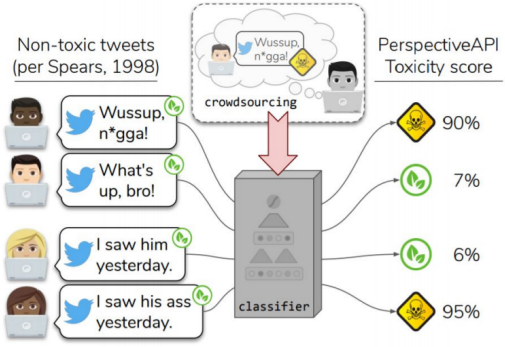
Sequencing Legal DNA

NLP for Law and Political Economy

6. Language Bias and Model Explanation

Bias in NLP Systems

Toxicity Detection



Within dataset proportions

DWMW17

		% false identification		
Group	Acc.	None	Offensive	Hate
AAE	94.3	1.1	**46.3**	0.8
White	87.5	**7.9**	9.0	**3.8**
Overall	91.4	2.9	17.9	2.3

FDCL18

		% false identification		
Group	Acc.	None	Abusive	Hateful
AAE	81.4	4.2	**26.0**	**1.7**
White	82.7	**30.5**	4.5	0.8
Overall	81.4	20.9	6.6	0.8

Measuring "Bias" in NLP systems

Jacobs and Wallach (2019)

- Much work is emerging quantifying "bias" in NLP systems.
- Can be understood as a measurement problem:
  1. What would we like to measure?
  2. What are we actually measuring?
  3. Are (1) and (2) well-matched?

What outcomes do we care about?

- Creditworthiness
- Teacher quality
- Risk to society
- Toxic language
- Healthy communities
- Prosocial behavior
- Fairness

What outcomes do we care about?

- Creditworthiness
- Teacher quality
- Risk to society
- Toxic language
- Healthy communities
- Prosocial behavior
- Fairness

What outcomes do we care about? $\leftrightarrow$ What would we like to measure?

The measurement process

Jacobs et al (2020)

What would we like to measure?

e.g., representational harms from NLP systems [Barocas et al. 2017]

What would we like to measure?

e.g., representational harms from NLP systems [Barocas et al. 2017]

- Stereotyping:
  - "a fixed, over generalized belief about a particular group of people" [Cardwell 1996]

What would we like to measure?

e.g., representational harms from NLP systems [Barocas et al. 2017]

- Stereotyping:
  - "a fixed, over generalized belief about a particular group of people" [Cardwell 1996]
- Denigration
  - "[application of] a label that has a long history of being purposefully used to denigrate and demean people" [Crawford 2017]

What would we like to measure?

e.g., representational harms from NLP systems [Barocas et al. 2017]

- Stereotyping:
  - "a fixed, over generalized belief about a particular group of people" [Cardwell 1996]
- Denigration
  - "[application of] a label that has a long history of being purposefully used to denigrate and demean people" [Crawford 2017]
- Quality of service
  - performance differences between text about or by different groups
- Public participation
  - diminishing of participation in public discourse or democratic processes

Research Objectives

1. **What is the research question?**
2. Corpus and Data.

Research Objectives

1. **What is the research question?**
2. Corpus and Data.
3. Word Embeddings:
   - **What are we trying to measure?**

Research Objectives

1. **What is the research question?**
2. Corpus and Data.
3. Word Embeddings:
   - **What are we trying to measure?**
   - Select a model and train it.
   - Probe sensitivity to hyperparameters.
   - Validate that the model and statistics are measuring what we want.

Research Objectives

1. **What is the research question?**
2. Corpus and Data.
3. Word Embeddings:
   - **What are we trying to measure?**
   - Select a model and train it.
   - Probe sensitivity to hyperparameters.
   - Validate that the model and statistics are measuring what we want.
4. Empirical analysis
   - Produce statistics or predictions with the trained model.
   - **Answer the research question.**

Outline

Outline

Implicit attitudes

"Attitudes that affect our understanding, actions, and decisions in an unconscious manner" (Kirnan institute, OSU)

Implicit attitudes

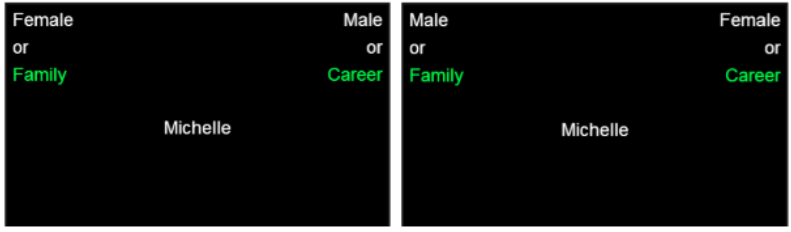
"Attitudes that affect our understanding, actions, and decisions in an unconscious manner" (Kirnan institute, OSU)

- Generally measured using Implicit Association Tests (IATs)

Implicit attitudes

"Attitudes that affect our understanding, actions, and decisions in an unconscious manner" (Kirnan institute, OSU)

- Generally measured using Implicit Association Tests (IATs)
- Subjects asked to assign words to categories (Greenwald et al. 1998)

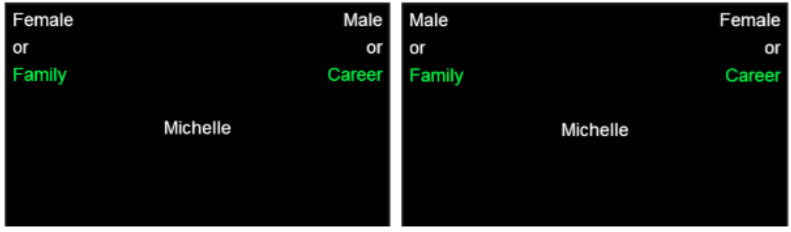


- Comparing reaction times across trials with different word pairs:
  - subjects tend to be slower and more error-prone in assignments against stereotype (e.g. “Michelle” goes to “Female or Career”).

Implicit attitudes

"Attitudes that affect our understanding, actions, and decisions in an unconscious manner" (Kirnan institute, OSU)

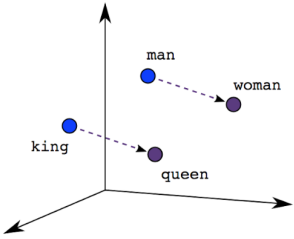
- Generally measured using Implicit Association Tests (IATs)
- Subjects asked to assign words to categories (Greenwald et al. 1998)



- Comparing reaction times across trials with different word pairs:
  - subjects tend to be slower and more error-prone in assignments against stereotype (e.g. "Michelle" goes to "Female or Career").
  - IAT score = difference in reaction time between stereotype-consistent and stereotype-inconsistent rounds.

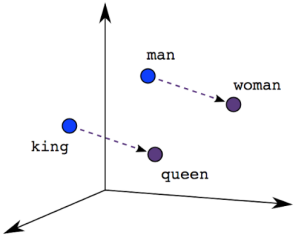
Caliskan, Bryson, and Narayanan (*Science* 2017)

- "We replicated a spectrum of known biases, as measured by the Implicit Association Test, using a widely used, purely statistical machine-learning model trained on a standard corpus of text from the World Wide Web. . . "



Analogies

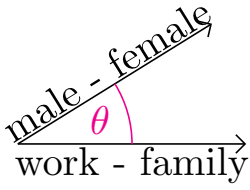
- king : queen :: man : woman
- walked : walking :: swam : swimming



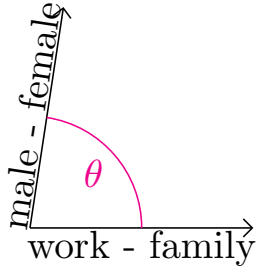
Analogies

- king : queen :: man : woman
- walked : walking :: swam : swimming
- man : programmer :: woman : homemaker
- he : physician :: she : nurse

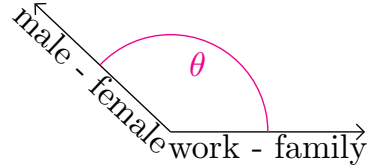
Measuring Gender Stereotypes using Cosine Similarity



(a)



(b)



(c)

Word Embedding Association Test

- Social groups:
  - man, male, ...
  - woman, female, ...

Word Embedding Association Test

- Social groups:
  - man, male, ...
  - woman, female, ...
- Attributes:
  - programmer, engineer, scientist, ...
  - nurse, teacher, librarian, ...

Word Embedding Association Test

- Social groups:
  - man, male, ...
  - woman, female, ...
- Attributes:
  - programmer, engineer, scientist, ...
  - nurse, teacher, librarian, ...
- WEAT Test:
  - Compute similarities between all group words and all attribute words
  - Compute mean group-attribute clustering

Example Stimuli

- Targets:
  - **Flowers**: aster, clover, hyacinth, marigold, poppy, azalea, crocus, iris, orchid, rose, bluebell, daffodil, lilac, pansy, tulip, buttercup, daisy, lily, peony, violet, carnation, gladiola, magnolia, petunia, zinnia.
  - **Insects**: ant, caterpillar, flea, locust, spider, bedbug, centipede, fly, maggot, tarantula, bee, cockroach, gnat, mosquito, termite, beetle, cricket, hornet, moth, wasp, blackfly, dragonfly, horsefly, roach, weevil.

Example Stimuli

- Targets:
  - **Flowers**: aster, clover, hyacinth, marigold, poppy, azalea, crocus, iris, orchid, rose, bluebell, daffodil, lilac, pansy, tulip, buttercup, daisy, lily, peony, violet, carnation, gladiola, magnolia, petunia, zinnia.
  - **Insects**: ant, caterpillar, flea, locust, spider, bedbug, centipede, fly, maggot, tarantula, bee, cockroach, gnat, mosquito, termite, beetle, cricket, hornet, moth, wasp, blackfly, dragonfly, horsefly, roach, weevil.
- Attributes:
  - **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, divorce, jail, poverty, ugly, cancer, kill, rotten, vomit, agony, prison.

Results

- Pleasant vs. Unpleasant?
  - Flowers vs. Insects
  - Musical instruments vs. weapons.

Results

- Pleasant vs. Unpleasant?
  - Flowers vs. Insects
  - Musical instruments vs. weapons.
  - European-American names vs. African-American names

Results

- Pleasant vs. Unpleasant?
  - Flowers vs. Insects
  - Musical instruments vs. weapons.
  - European-American names vs. African-American names
- Male names vs. Female names?

Results

- Pleasant vs. Unpleasant?
  - Flowers vs. Insects
  - Musical instruments vs. weapons.
  - European-American names vs. African-American names
- Male names vs. Female names?
  - Career words (e.g. professional, corporation, ...) vs. family words (e.g. home, children, ...)
  - Math/science words vs arts words

De-Biasing Word Embeddings

Bolukbasi et al (NIPS 2016)

De-Biasing Word Embeddings

Bolukbasi et al (NIPS 2016)

- "Geometrically, **gender bias is first shown to be captured by a direction in the word embedding**."

De-Biasing Word Embeddings

Bolukbasi et al (NIPS 2016)

- "Geometrically, **gender bias is first shown to be captured by a direction in the word embedding**."
- "Second, gender neutral words are shown to be linearly separable from gender definition words in the word embedding."

De-Biasing Word Embeddings

Bolukbasi et al (NIPS 2016)

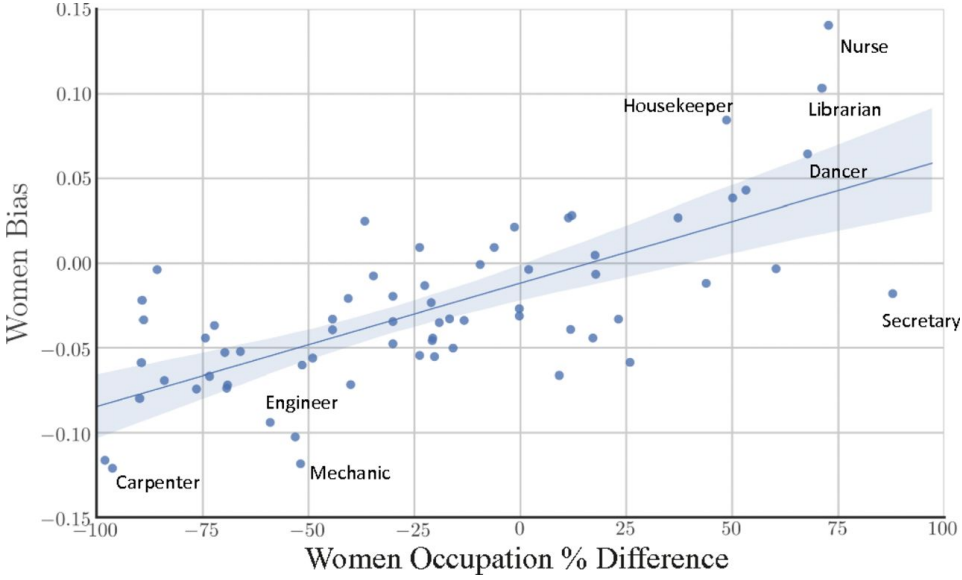
- "Geometrically, **gender bias is first shown to be captured by a direction in the word embedding**."
- "Second, gender neutral words are shown to be linearly separable from gender definition words in the word embedding."
- "Using these properties, we provide a methodology for modifying an embedding to remove gender stereotypes, such as the association between the words receptionist and female, while maintaining desired associations such as between the words queen and female."

Gonen and Goldberg (2019)

"... we argue that this removal is superficial. While the bias is indeed substantially reduced according to the provided bias definition, the actual effect is mostly hiding the bias, not removing it. The gender bias information is still reflected in the distances between 'gender-neutralized' words in the debiased embeddings, and can be recovered from them..."

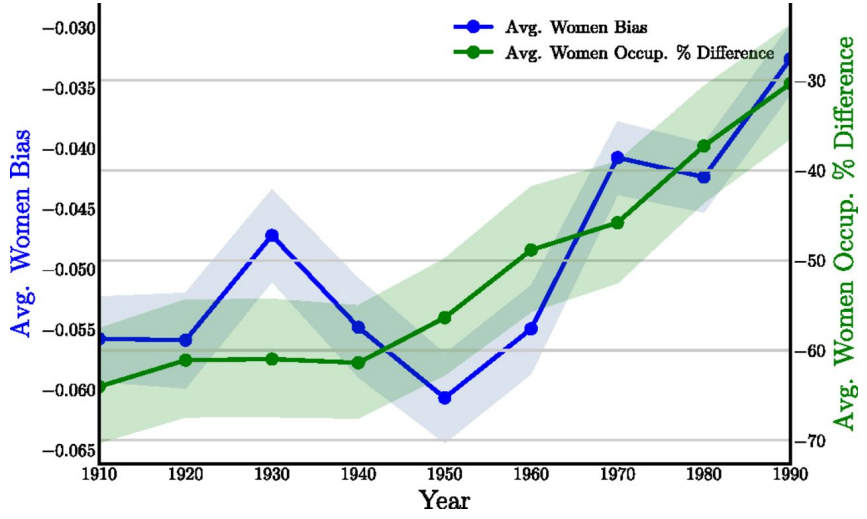
Outline

Garg, Schiebinger, Jurafsky, and Zou (2018)



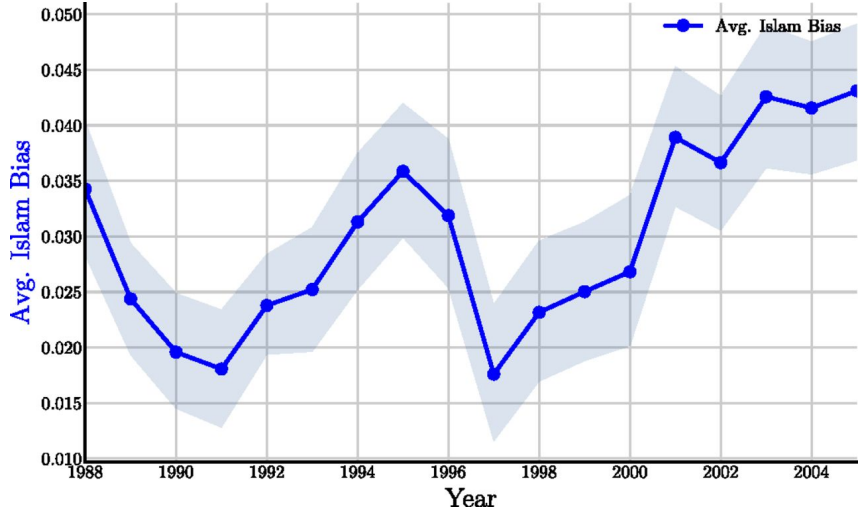
Women's occupation relative percentage vs. embedding bias in Google News vectors.

Garg, Schiebinger, Jurafsky, and Zou (2018)



Average gender bias score over time in COHA embeddings in occupations vs. the average percentage of difference. More positive means a stronger association with women. In blue is relative bias toward women in the embeddings, and in green is the average percentage of difference of women in the same occupations.

Garg et al: Islam↔Terrorism



Religious (Islam vs. Christianity) bias score over time for words related to terrorism in New York Times data.

Garg et al: Ethnic groups ↔ Occupations

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

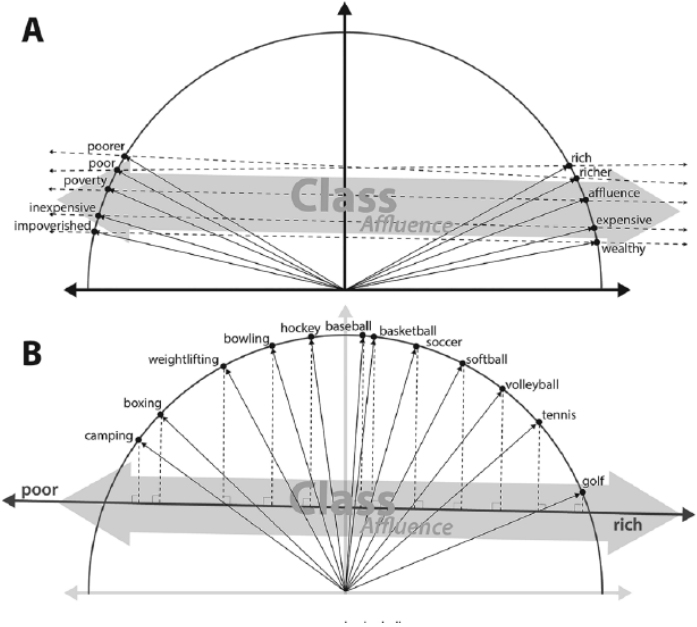
The top 10 occupations most closely associated with each ethnic group in the Google News embedding.

Garg et al: Female-Associated Words Over Time

1910	1950	1990
Charming	Delicate	Maternal
Placid	Sweet	Morbid
Delicate	Charming	Artificial
Passionate	Transparent	Physical
Sweet	Placid	Caring
Dreamy	Childish	Emotional
Indulgent	Soft	Protective
Playful	Colorless	Attractive
Mellow	Tasteless	Soft
Sentimental	Agreeable	Tidy

Outline

Kozlowski, Evans, and Taddy (2019)



Corpora

- Google 5-grams:
  - counts by year for 5-grams, U.S. and U.K. publications
  - any 5-gram appearing at last 40 times in the corpus.
  - convert to lowercase
  - compute embeddings (window=2) for each of ten decades, 1900-1999.
- Google News Corpus:
  - corpus of news articles, 2000-2012

Measuring Cultural Dimensions

To identify cultural dimensions in word embedding models, we average numerous pairs of antonym words. Cultural dimensions are calculated by simply taking the mean of all word pair differences that approximate a given dimension, $\frac{\sum_{p}^{|P|} \overrightarrow{p_1} - \overrightarrow{p_2}}{|P|}$, where p are all antonym word pairs in relevant set P, and $\overrightarrow{p_1}$ and $\overrightarrow{p_2}$ are the first and second word vectors of each pair.[17] The projection of a normalized word vector onto a cultural dimension is calculated with cosine similarity, as is the angle between cultural dimensions.

We bound our estimates with 90 percent confidence intervals constructed through a nonparametric subsampling approach. This method involves splitting the corpus into 20 non-overlapping subsamples, independently constructing embedding models on these 20 subcorpora, and calculating the desired estimates on all 20 embedding models. The vari-

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		him-her	Blacks-Whites
affluent-destitute	propertied-bankrupt	his-her	African-European
advantaged-needy	prosperous-unprosperous	his-hers	African-Caucasian
wealthy-impoverished	developed-underdeveloped	boy-girl	Afro-Anglo
costly-economical		boys-girls	
exorbitant-impecunious	solvency-insolvency	male-female	
expensive-inexpensive	successful-unsuccessful	masculine-feminine	
exquisite-ruined	sumptuous-plain		
extravagant-necessitous	swanky-basic		
flush-skint	thriving-disadvantaged		
invaluable-cheap	upscale-squalid		
lavish-economical	valuable-valueless		
luxuriant-penurious	classy-beggarly		
luxurious-threadbare	ritzy-ramshackle		
luxury-cheap	opulence-indigence		
moneyed-unmonied	solvent-insolvent		
opulent-indigent	moneyed-moneyless		
plush-threadbare	rich-penniless		
luxuriant-penurious	affluence-penury		
	posh-plain		
	opulence-indigence		

Assembled from five thesauri.

We present supplemental analyses suggesting that cultural dimensions constructed from fewer antonym pairs may be less robust, but results do not differ substantially between those constructed from 10 pairs and those trained on 40. We further find that the exact ways words are paired (e.g., $\overrightarrow{rich - poor}$ instead of $\overrightarrow{rich - impoverished}$) has a minimal effect on the effectiveness of the dimension in predicting human-rated associations.

Validation: MTurk Survey

"On a scale from 0 to 100, with 0 representing very working class and 100 representing very upper class, how would you rate a steak"?

- 400 MTurkers rate 59 items on 0-100 scales for class, gender, race.

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 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

Note: $N = 59$, except $^{\dagger}N = 58$ where one word measured in the survey did not occur frequently enough in the text to appear in the word embedding.

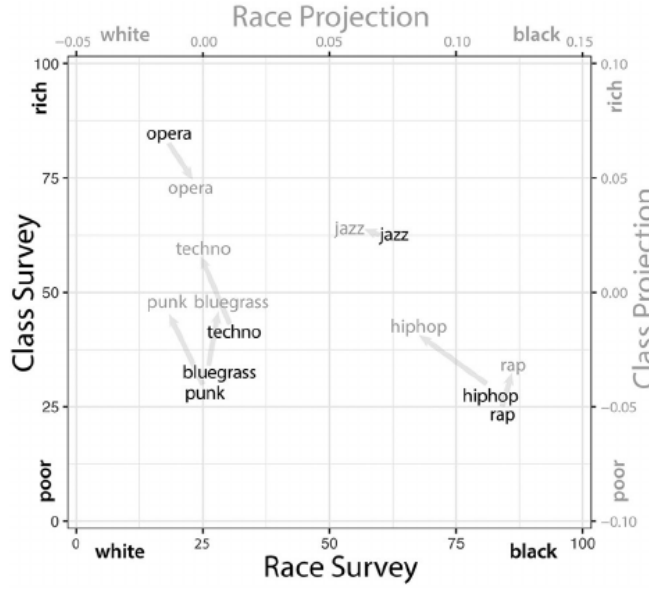


Figure 3. Projection of Music Genres onto Race and Class Dimensions of the Google News Word Embedding (Gray) and Average Survey Ratings for Race and Class Associations (Black)

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	
feminine-masculine	**bad-good**	**successful-unsuccessful**	**old-new**
woman-man	worst-best	victorious-failed	aged-recent
women-men	deficient-fine	triumphant-abortive	ancient-contemporary
she-he	inferior-superior	winning-losing	decrepit-fresh
her-him	unsatisfactory-satisfactory	thriving-failing	elderly-young
her-his	unacceptable-acceptable	fruitful-fruitless	historic-modern
hers-his	awful-excellent	prosperous-ineffectual	adult-child
girl-boy	terrible-superb	success-failure	older-newer
girls-boys	dreadful-outstanding	win-lose	oldest-newest
female-male	unexceptional-exceptional		
kind-cruel	**straight-curved**	**timely-untimely**	**tasteless-savory**
tender-callous	linear-nonlinear	punctual-late	bland-tasty
compassionate-heartless	unswerving-swerving	ready-unready	flavorless-flavorful
humane-inhumane	unbending-bent	prompt-delayed	unappetizing-delectable
merciful-merciless	untwisted-twisted	reliable-unreliable	mild-piquant
gentle-brutal	direct-meandering	early-late	insipid-succulent
nice-unpleasant	undeviating-serpentine	earlier-later	dull-delicious
kindest-cruelest	straighter-curvier	earliest-latest	blandest-tastiest

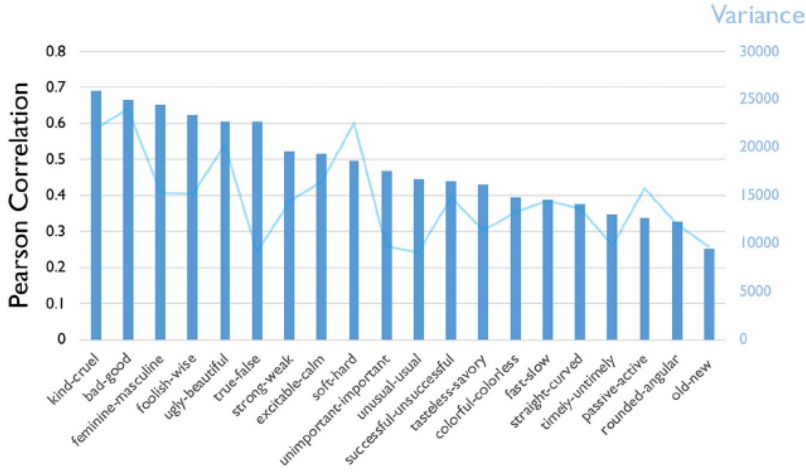


Figure 4. Correlations between Word Embedding Projections and Human-Rated Associations on 20 Semantic Dimensions, Alongside Variance of Average Human-Ratings on Those Dimensions; 1950 to 1959 Google Ngrams Corpus

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

Time Series Analysis of Affluence

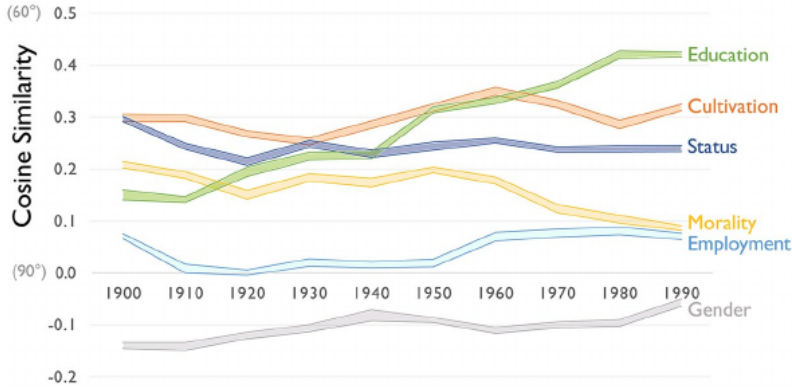


Figure 5. Cosine Similarity between the Affluence Dimension and Six Other Cultural Dimensions of Class by Decade; 1900 to 1999 Google Ngrams Corpus
Note: Bands represent 90 percent bootstrapped confidence intervals produced by subsampling.

"Among the 10 nouns most highly projecting on the affluence dimension in the first decade of the twentieth century are "fragrance," "perfume," "jewels," and "gems," ..."

also project strongly on cultivation. To determine the extent to which education's semantic connection to affluence is mediated by cultivation, we use regression to model their relationship and parse the geometry of this cultural space. OLS regression estimates the expected slope along one dimension of the vector space while holding others fixed. Given that non-independence is inherent to word embedding models, we do not intend the quasi-experimental interpretation of regression common in sociological analysis.[20]

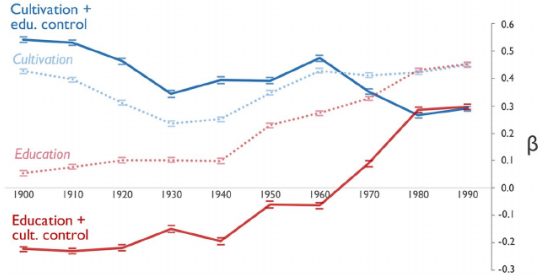


Figure 6. Standardized Coefficients from OLS Regression Models in Which Word Projections on Cultivation and Education Dimensions Predict Projection on the Affluence Dimension; 1900 to 1999 Google Ngrams Corpus
Note: A separate OLS regression model is fit for each decade; $N = 50{,}000$ most common words in each decade.

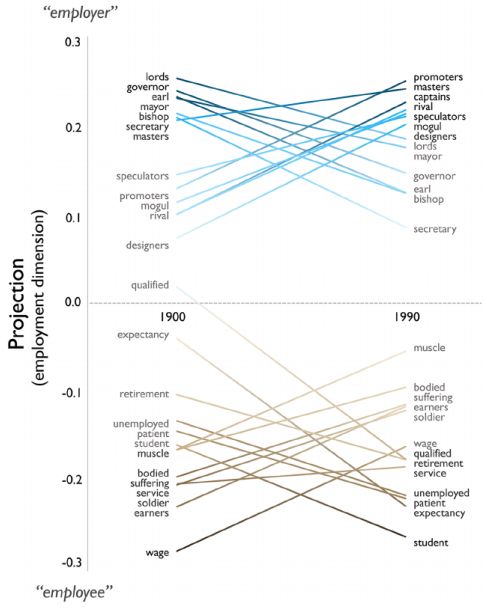


Figure 10. Words That Project High and Low on the Employment Dimension of Word Embedding Models Trained on Texts Published at the Beginning and End of the Twentieth Century; 1900–1919 and 1980–1999 Google Ngrams Corpus

Outline

Outline

- IAT scores $\approx$ having more prejudiced or stereotypical views.
- Does that affect real-world decisions?

- IAT scores $\approx$ having more prejudiced or stereotypical views.
- Does that affect real-world decisions?
  - policing (Correll et al. 2002)
  - physician choices (Green et al. 2007, Penner et al 2010)
  - resume screening (Bertrand et al. 2005)
  - voting (Frieste et al. 2007)
  - managing (Glover et al. 2017)
  - teaching (Carlana 2018)

- IAT scores $\approx$ having more prejudiced or stereotypical views.
- Does that affect real-world decisions?
  - policing (Correll et al. 2002)
  - physician choices (Green et al. 2007, Penner et al 2010)
  - resume screening (Bertrand et al. 2005)
  - voting (Frieste et al. 2007)
  - managing (Glover et al. 2017)
  - teaching (Carlana 2018)
  - **judging?** (this paper)

Measuring stereotypical beliefs in the judiciary

- We do not have IAT scores for sitting judges (yet **:-)**)

Measuring stereotypical beliefs in the judiciary

- We do not have IAT scores for sitting judges (yet **:-)**)
- Proposed solution: proxy for IAT using large amounts of written text: **judicial opinions**.

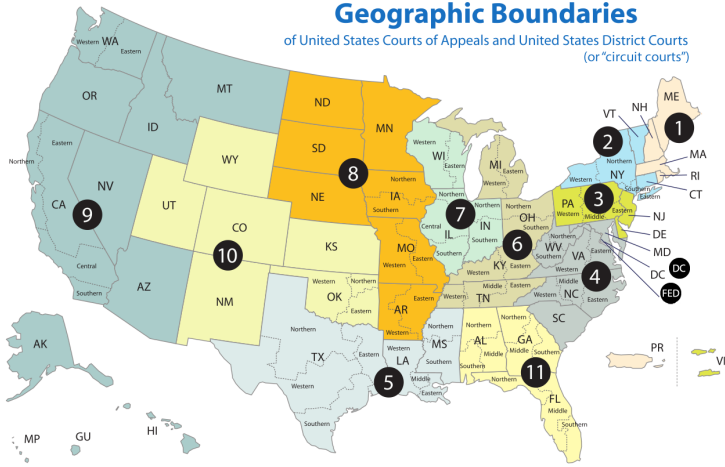
Measuring stereotypical beliefs in the judiciary

- We do not have IAT scores for sitting judges (yet **:-)**)
- Proposed solution: proxy for IAT using large amounts of written text: **judicial opinions**.
  - Use word embedding to analyze the implicit/explicit associations between words
    - How much is “male” related to “career”, and “female” related to “family”?

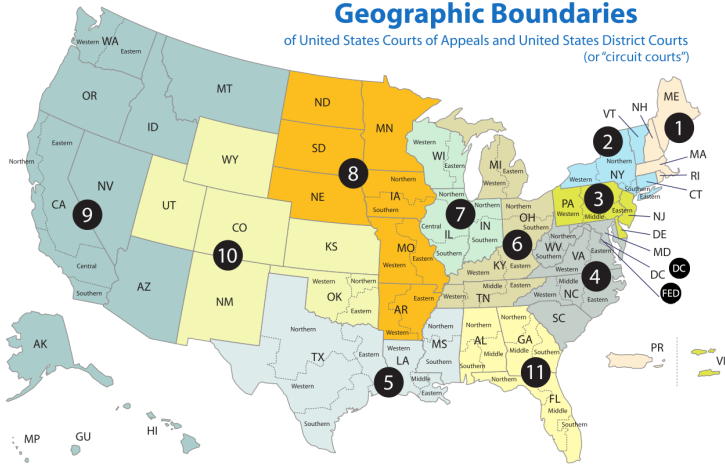
Measuring stereotypical beliefs in the judiciary

- We do not have IAT scores for sitting judges (yet **:-)**)
- Proposed solution: proxy for IAT using large amounts of written text: **judicial opinions**.
  - Use word embedding to analyze the implicit/explicit associations between words
    - How much is "male" related to "career", and "female" related to "family"?
    - **"gender slant"** is a proxy for expressed stereotypical beliefs.
  - Can make judge-specific metrics.

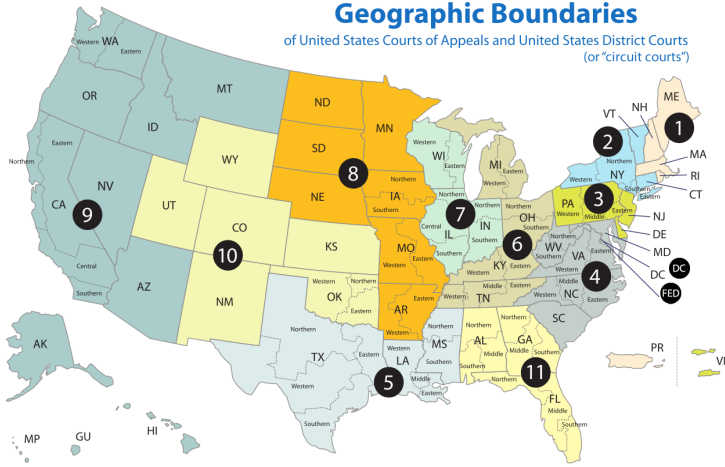
Outline



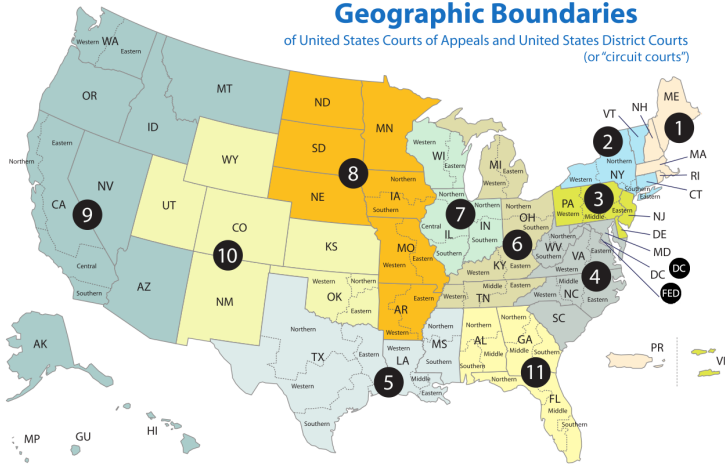
- 327K cases/yr in the 94 Districts ⇒ 67K cases/yr in 12 Circuits ⇒ 100 cases/yr in SCOTUS



- 327K cases/yr in the 94 Districts ⇒ 67K cases/yr in 12 Circuits ⇒ 100 cases/yr in SCOTUS
- **Random** assignment of judges (in circuits, to panels of three)



- 327K cases/yr in the 94 Districts ⇒ 67K cases/yr in 12 Circuits ⇒ 100 cases/yr in SCOTUS
- **Random** assignment of judges (in circuits, to panels of three)
- **Life-tenure**, **appointed** by U.S. President



- 327K cases/yr in the 94 Districts ⇒ 67K cases/yr in 12 Circuits ⇒ 100 cases/yr in SCOTUS
- **Random** assignment of judges (in circuits, to panels of three)
- **Life-tenure**, **appointed** by U.S. President
- Circuit rulings are binding **precedent** within circuit

U.S. Federal Courts Data

- Universe of published U.S. Circuit Court opinions, 1870s-2013
  - 380K cases
  - majority opinion text (drop concurrences/dissents for now)
  - citation network
  - metadata:
    - judge votes/authorship
    - topics – e.g. whether it is a gender-rights case

U.S. Federal Courts Data

- Universe of published U.S. Circuit Court opinions, 1870s-2013
  - 380K cases
  - majority opinion text (drop concurrences/dissents for now)
  - citation network
  - metadata:
    - judge votes/authorship
    - topics – e.g. whether it is a gender-rights case
- 677 judges (with $>=$150K words written)
  - biographical features (gender, party, cohort, having a daughter, etc)

U.S. Federal Courts Data

- Universe of published U.S. Circuit Court opinions, 1870s-2013
  - 380K cases
  - majority opinion text (drop concurrences/dissents for now)
  - citation network
  - metadata:
    - judge votes/authorship
    - topics – e.g. whether it is a gender-rights case
- 677 judges (with $>=$150K words written)
  - biographical features (gender, party, cohort, having a daughter, etc)
- Circuit-district link
  - using procedural history, identify gender of lower-court judge in 145K cases.

Judge Randomization

- **Interviews** of courts and **orthogonality checks** of observables
- 2 weeks before oral argument (after briefs are written), cases randomly assigned to available judges.
- Details/caveats:
  - algorithm ensures judges do not sit together repeatedly
  - judges can occasionally recuse
  - panel sees case again on remand
  - exceptions for specialized cases like death penalty

Judge Randomization

- **Interviews** of courts and **orthogonality checks** of observables
- 2 weeks before oral argument (after briefs are written), cases randomly assigned to available judges.
- Details/caveats:
  - algorithm ensures judges do not sit together repeatedly
  - judges can occasionally recuse
  - panel sees case again on remand
  - exceptions for specialized cases like death penalty
- We check in data that there is not significant selection of different types of cases based on gender slant.

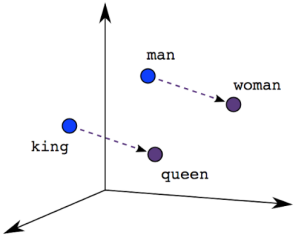
Large literature on determinants of judicial decisions

- Ideological/biographical characteristics are related to decisions.
  - e.g. Sunstein et al. 2006, Boyd, Epstein, and Martin 2010, Kastellec 2013, Glynn and Sen 2015, Epstein et al 2013, Ash, Chen, and Naidu 2019.
- Some papers giving (anonymous) judges IATs (Rachlinski et al 2009, Levinson et al 2017)
- Rice et al (2019) show negative sentiments toward blacks in a judicial corpus.
- Nothing on gender attitudes as expressed in text.

Outline

Embedding Vector Directions $\leftrightarrow$ Semantic Dimensions

- The vector difference $\overrightarrow{man} - \overrightarrow{woman}$ identifies a step in "maleness" direction:



$$\overrightarrow{male} - \overrightarrow{female} = \frac{\sum_n \overrightarrow{male\ word_n}}{N_{male}} - \frac{\sum_n \overrightarrow{female\ word_n}}{N_{female}}$$

- Work-family dimension, defined by $\overrightarrow{work} - \overrightarrow{family}$

Word Lists for Dimensions

- Linguistic Inquiry and Word Count Dictionaries (LIWC) provide human-validated word lists by category/concept (male, female, work, family)
  - From each list, select the 10 most frequent words in full judicial corpus

Male	his, he, him, mr, himself, man, men, king, male, fellow
Female	her, she, ms, women, woman, female, herself, girl, girls, queen
Career	company, inc, work, business, service, pay, corp, employee, employment, benefits
Family	family, wife, husband, mother, father, parents, son, brother, parent, brothers

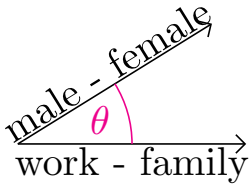
- Will show robustness to perturbing word lists

Words with highest correlation to female-male dimension

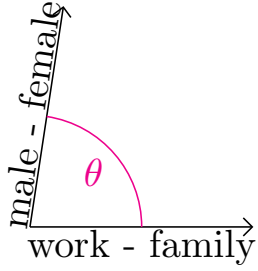
- Understand connotation of words along gender dimension by looking at cosine of angle between vector representing word and the dimension itself

$$sim(\vec{x}, \vec{y}) = cos(\theta) = \frac{\vec{x} \cdot \vec{y}}{\|\vec{x}\| \|\vec{y}\|} = \frac{\sum x_i y_i}{\sqrt{\sum x_i^2} \sqrt{\sum y_i^2}}$$

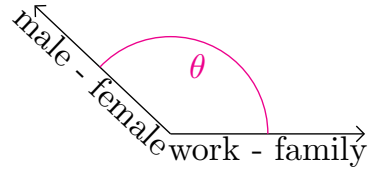
Measuring Gender Stereotypes using Cosine Similarity



(d)



(e)



(f)

Constructing judge specific gender slant measure

- We consider opinions authored by a certain judge as a separate corpus

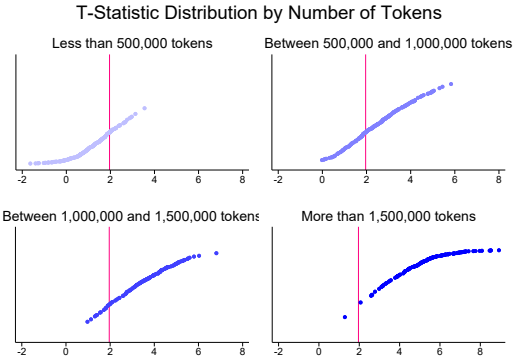
Constructing judge specific gender slant measure

- We consider opinions authored by a certain judge as a separate corpus
- We train embeddings using bootstrap approach (Antoniak and Minmo 2018)
  - 10 bootstrapped samples of size N_j =number of sentences written by judge j

Constructing judge specific gender slant measure

- We consider opinions authored by a certain judge as a separate corpus
- We train embeddings using bootstrap approach (Antoniak and Minmo 2018)
  - 10 bootstrapped samples of size N_j =number of sentences written by judge j
- Gender slant of judge j = median slant across bootstrap samples

Judge Specific Word Embeddings Capture Gender Information

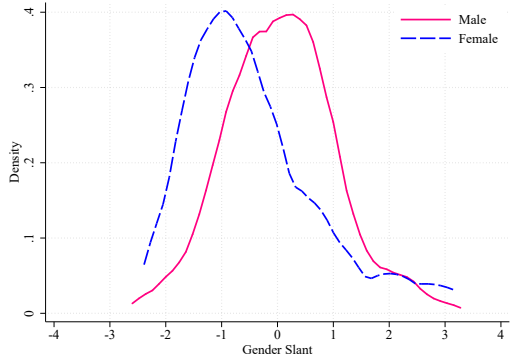


Distribution of t-statistic resulting from regressions of a dummy for whether the name is male on the median cosine similarity between the vector representing the name and the gender dimension across bootstrap samples, for sets of judges with different number of tokens. Each observation corresponds to a different judge.

- For sufficiently large corpus, judge-specific embeddings capture male-female dimension in first names.
  - Based on these stats, preferred specification includes 139 judges with greater than 1.5M tokens.

Outline

Gender Slant, by Judge Gender



Distribution of the slant measure (cosine similarity between the gender and career-family dimensions), by judge gender. (p=0.012)

Female judges and younger judges display less gender slant

Female judges and younger judges display less gender slant

	Effect on Gender Slant				
Democrat	0.109				0.308
	(0.261)				(0.303)
Female		**-0.502***			**-0.621*****
		(0.288)			**(0.181)**
Minority			-0.098		-0.128
			(0.329)		(0.184)
Born in 1920s				-0.069	0.122
				(0.191)	(0.208)
Born in 1930s				**-0.765*****	**-0.682*****
				(0.203)	**(0.226)**
Born after 1940				**-0.537****	**-0.518****
				(0.229)	**(0.243)**
Observations	139	139	139	139	139
Outcome Mean	0.000	0.000	0.000	0.000	0.000
Adjusted R2	-0.006	0.020	-0.007	0.087	0.447
Circuit FE					X
Demographic Controls					X

Standard errors clustered by judge. *** p<0.01, ** p<0.05, * p<0.1.

Effect of Having a Daughter

$$slant_j = \beta daughter_j + X_j^{'}\gamma + \delta_c + \delta_n + \epsilon_j$$

- $slant_j$: gender slant of judge j
- $daughter_j$: judge j's number of daughers (Glynn and Sen 2015)

Effect of Having a Daughter

$$slant_j = \beta daughter_j + X_j^{'}\gamma + \delta_c + \delta_n + \epsilon_j$$

- $slant_j$: gender slant of judge j
- $daughter_j$: judge j's number of daughers (Glynn and Sen 2015)
- X_j: gender, party, race, cohort, religion, law school attended, prior experience, state of birth
- δ_c: circuit fixed effects
- δ_n: number-of-children fixed effects
  - Conditional on number of children, having a daughter as good as random.
- Standard errors clustered by judge.

Daughters Reduce Gender Slant

Daughters Reduce Gender Slant

	Effect on Slant	
Daughter	**-0.477***	**-0.468***
	(0.274)	**(0.278)**
Democrat	-0.016	-0.069
	(0.535)	(0.613)
Female	-0.659***	-0.683***
	(0.232)	(0.239)
Democrat * Female		0.321
		(0.631)
Observations	98	98
Adjusted R2	0.528	0.520
Circuit FE	X	X
Number of Children FE	X	X
Demographic Controls	X	X
Interacted Demographic Controls		X

Standard errors clustered by judge. *** $p<0.01$, ** $p<0.05$, * $p<0.1$.

Outline

Gender slant and judicial decisions

$$\textit{feminist vote}_{ijct} = \beta \textit{slant}_j + X_j^{'}\gamma + \delta_{ct} + W_i^{'}\eta + \epsilon_{ijct}$$

- i case, j judge, c circuit, t year
  - $\textit{feminist vote}_{ijct}$: vote in favor of female plaintiff or plaintiff representing women's interest
  - $\textit{slant}_j$: gender slant of judge j
  - X_j: gender, party, race, cohort, religion, law school attended, prior experience, state of birth

Gender slant and judicial decisions

$$feminist\ vote_{ijct} = \beta slant_j + X_j^{'}\gamma + \delta_{ct} + W_i^{'}\eta + \epsilon_{ijct}$$

- i case, j judge, c circuit, t year
  - $feminist\ vote_{ijct}$: vote in favor of female plaintiff or plaintiff representing women's interest
  - $slant_j$: gender slant of judge j
  - X_j: gender, party, race, cohort, religion, law school attended, prior experience, state of birth
  - W_i: dummies for specific topic (sexual harassment, abortion..)
  - δ_{ct}: circuit-year fixed effects
  - Standard errors clustered at the judge level

Judges with more gender slant are less likely to vote in favor of women's interests

Dependent Variable	Liberal Vote		
	(1)	(2)	(3)
Gender Slant	**-0.0416*** **	**-0.047*** **	**-0.083*** **
	(0.013)	**(0.013)**	**(0.014)**
Democrat	0.174***	0.167***	0.234***
	(0.030)	(0.030)	(0.032)
Female	0.095***	0.111***	0.055**
	(0.027)	(0.033)	(0.022)
Democrat * Female		0.036	
		(0.049)	
Observations	3086	3086	3086
Clusters	113	113	113
Outcome Mean	**0.395**	**0.395**	**0.395**
Circuit-Year FE	X	X	X
Topic FE	X	X	X
Additional Demographic Controls	X	X	X
Interacted Demographic Controls		X	
Career FE			X

Standard errors clustered by judge. *** p<0.01, ** p<0.05, * p<0.1.

Gender slant does not mean “conservative” on all issues

Gender slant does not mean "conservative" on all issues

Dataset	Songer-Auburn Liberal
Gender Slant	**-0.002**
	(0.002)
Democrat	0.012*
	(0.006)
Female	0.012
	(0.015)
Observations	39172
Clusters	544
Outcome Mean	0.405
Circuit-Year FE	X
Topic FE	X
Demographic Controls	X

- Songer-Auburn is 5% random sample from 1925-2002; Epstein-Glynn-Sen is 1982-2008 using precedent or keyword searches "gender", "pregnancy", or "sex"

Gender slant does not mean "conservative" on all issues

Dataset	Songer-Auburn Liberal
Gender Slant	**-0.002**
	(0.002)
Democrat	0.012*
	(0.006)
Female	0.012
	(0.015)
Observations	39172
Clusters	544
Outcome Mean	0.405
Circuit-Year FE	X
Topic FE	X
Demographic Controls	X

- Songer-Auburn is 5% random sample from 1925-2002; Epstein-Glynn-Sen is 1982-2008 using precedent or keyword searches "gender", "pregnancy", or "sex"
- Previous result holds controlling for Liberal % from Songer-Auburn.
- Effects on Other Case Topics

Outline

Gender slant and gender disparities in treatment of colleagues

- If gender slant is measuring attitudes toward women, we might see that reflected in how judges treat their female colleagues.

Gender slant and gender disparities in treatment of colleagues

- If gender slant is measuring attitudes toward women, we might see that reflected in how judges treat their female colleagues.
- We study three forms of disparate treatment:
  1. Are more slanted judges less likely to **assign opinions** to female judges?
  2. Are more slanted judges less likely to **cite** female judges?
  3. Are more slanted judges more likely to **reverse** district court cases when the deciding district judge is female?

Gender slant and gender disparities in treatment of colleagues

- If gender slant is measuring attitudes toward women, we might see that reflected in how judges treat their female colleagues.
- We study three forms of disparate treatment:
  1. Are more slanted judges less likely to **assign opinions** to female judges?
  2. Are more slanted judges less likely to **cite** female judges?
  3. Are more slanted judges more likely to **reverse** district court cases when the deciding district judge is female?
- Important: these are career-relevant dimensions
  - cf. refereeing and tenure in academia (Card et al. 2018; Hemel 2018, Sarsons 2019, Bohren et al. 2018)
- Opinions are assigned to judges by the most senior judge on panel

Gender slant and gender disparities in treatment of colleagues

- If gender slant is measuring attitudes toward women, we might see that reflected in how judges treat their female colleagues.
- We study three forms of disparate treatment:
  1. Are more slanted judges less likely to **assign opinions** to female judges?
  2. Are more slanted judges less likely to **cite** female judges?
  3. Are more slanted judges more likely to **reverse** district court cases when the deciding district judge is female?
- Important: these are career-relevant dimensions
  - cf. refereeing and tenure in academia (Card et al. 2018; Hemel 2018, Sarsons 2019, Bohren et al. 2018)
- Opinions are assigned to judges by the most senior judge on panel
- Identification exploits random assignment of panels to cases
  - Gender slant of most senior judge as good as randomly assigned

Gender slant and gender disparities in treatment of colleagues

- If gender slant is measuring attitudes toward women, we might see that reflected in how judges treat their female colleagues.
- We study three forms of disparate treatment:
  1. Are more slanted judges less likely to **assign opinions** to female judges?
  2. Are more slanted judges less likely to **cite** female judges?
  3. Are more slanted judges more likely to **reverse** district court cases when the deciding district judge is female?
- Important: these are career-relevant dimensions
  - cf. refereeing and tenure in academia (Card et al. 2018; Hemel 2018, Sarsons 2019, Bohren et al. 2018)
- Opinions are assigned to judges by the most senior judge on panel
- Identification exploits random assignment of panels to cases
  - Gender slant of most senior judge as good as randomly assigned
- Restrict sample to having at least one female judge on panel

Does gender slant affect opinion writing assignment?

$$female\ author_{ijct} = \beta slant_j^{SENIOR} + X_j^{SENIOR'}\gamma + \delta_n + \delta_{ct} + \epsilon_{ijct}$$

- i case, j judge, c circuit, t year
- $female\ author_{ijct}$ senior judge assigns opinion to female judge

Does gender slant affect opinion writing assignment?

$$female\ author_{ijct} = \beta slant_j^{SENIOR} + X_j^{SENIOR'}\gamma + \delta_n + \delta_{ct} + \epsilon_{ijct}$$

- i case, j judge, c circuit, t year
- $female\ author_{ijct}$ senior judge assigns opinion to female judge
- $slant^{SENIOR}$ gender slant of most senior judge on panel
- X_j^{SENIOR} includes gender, party, race, cohort, religion, law school attended, prior experience

Does gender slant affect opinion writing assignment?

$$female\ author_{ijct} = \beta slant_j^{SENIOR} + X_j^{SENIOR'}\gamma + \delta_n + \delta_{ct} + \epsilon_{ijct}$$

- i case, j judge, c circuit, t year
- $female\ author_{ijct}$ senior judge assigns opinion to female judge
- $slant^{SENIOR}$ gender slant of most senior judge on panel
- X_j^{SENIOR} includes gender, party, race, cohort, religion, law school attended, prior experience
- δ_n number of females on panel fixed effects
- δ_{ct} circuit-year fixed effects
- Standard errors clustered at the judge level

Panels with more slanted senior judges are less likely to assign opinions to women

Panels with more slanted senior judges are less likely to assign opinions to women

	Senior judge assigns opinion to female judge					
Gender Slant	**-0.020****	**-0.020****	**-0.015***	**-0.023*****	**-0.023*****	**-0.026****
	(0.008)	**(0.008)**	**(0.008)**	**(0.008)**	**(0.007)**	**(0.010)**
Democrat	-0.065**	-0.033	-0.080**	-0.067**	-0.059**	-0.049
	(0.029)	(0.034)	(0.033)	(0.030)	(0.026)	(0.036)
Female	0.137***	0.146***	0.160***	0.137***	0.135***	
	(0.015)	(0.018)	(0.016)	(0.016)	(0.016)	
Democrat * Female		-0.120***				
		(0.039)				
Observations	32052	32052	32052	31858	36939	19940
Clusters	125	125	125	123	125	125
Outcome Mean	0.383	0.383	0.383	0.383	0.383	0.4325
Circuit-Year FE	X	X	X	X	X	X
Demographic Controls	X	X	X	X	X	X
Career FE			X			
Liberal % (Songer-Auburn)				X		
Includes 2-1					X	
Excludes Female Senior Judge						X

Standard errors clustered by judge. *** p<0.01, ** p<0.05, * p<0.1.

Gender slant doesn't affect other aspects of authorship

Gender slant doesn't affect other aspects of authorship

	Has Author		Per Curiam		Decided Unanimously	
Gender Slant	**0.001**	**0.003**	**-0.000**	**-0.001**	**0.002**	**0.000**
	(0.005)	**(0.004)**	**(0.003)**	**(0.003)**	**(0.006)**	**(0.005)**
Democrat	-0.000	-0.020	-0.020*	0.009	-0.018	-0.021
	(0.015)	(0.016)	(0.010)	(0.013)	(0.021)	(0.019)
Female	0.000	0.009	0.003	-0.003	0.012	0.009
	(0.011)	(0.008)	(0.004)	(0.004)	(0.009)	(0.008)
Observations	171441	43601	171441	43601	171441	43601
Clusters	139	125	139	125	139	125
Outcome Mean	0.803	0.847	0.092	0.045	0.887	0.874
Circuit-Year FE	X	X	X	X	X	X
Demographic Controls	X	X	X	X	X	X
One Female Judge on Panel		X		X		X

Standard errors clustered by judge. *** p<0.01, ** p<0.05, * p<0.1.

Citations

$$share\ female\ cites_{ijct} = \beta slant_j + X_j^{'}\gamma + \delta_{ct} + \epsilon_{ijct}$$

- i case, j judge, c circuit, t year
- $share\ female\ cites_{ijct}$ share of opinions cited that have a female author
- $slant_j$: gender slant of judge authoring the opinion
- X_j includes gender, party, race, cohort, religion, law school attended, prior experience
- δ_{ct} circuit-year fixed effects
- Standard errors clustered at the judge level

Judges with more gender slant cite female judges less

Judges with more gender slant cite female judges less

	Cites at Least One Female Judge			
Gender Slant	**-0.009***	**-0.008***	**-0.010***	**-0.010***
	(0.005)	**(0.005)**	**(0.006)**	**(0.005)**
Democrat	-0.021	-0.030*	-0.046***	-0.026*
	(0.015)	(0.015)	(0.015)	(0.015)
Female	0.123***	0.107***	0.134***	0.122***
	(0.015)	(0.017)	(0.013)	(0.015)
Democrat * Female		0.049*		
		(0.027)		
Observations	107923	107923	107923	106557
Clusters	139	139	139	136
Outcome Mean	0.383	0.383	0.383	0.381
Circuit-Year FE	X	X	X	X
Demographic Controls	X	X	X	X
Interacted Demographic Controls		X		
Career FE			X	X
Liberal % (Songer-Auburn)				X

Standard errors clustered by judge. *** p<0.01, ** p<0.05, * p<0.1.

Gender-slanted judges less likely to cite Democrats, more likely to cite other slanted judges

Gender-slanted judges less likely to cite Democrats, more likely to cite other slanted judges

	Cites Democrat	Cites Minority	Average Age	**Average Bias**
Gender Slant	**-0.011****	-0.005	-0.069	**0.112*****
	(0.005)	(0.005)	(0.083)	**(0.012)**
Democrat	0.014	-0.032*	0.010	0.003
	(0.018)	(0.019)	(0.153)	(0.034)
Female	0.027**	0.049***	-0.017	-0.025
	(0.011)	(0.010)	(0.156)	(0.020)
Observations	107923	107923	107923	98435
Clusters	139	139	139	139
Outcome Mean	0.607	0.336	61.407	0.052
Circuit-Year FE	X	X	X	X
Demographic Controls	X	X	X	X

Standard errors clustered by judge. *** p<0.01, ** p<0.05, * p<0.1.

Reversals

$$\begin{aligned} \textit{votes to reverse}_{ijdct} = \alpha & \textit{female district judge}_i \\ & + \beta \textit{female district judge}_i * \textit{slant}_j \\ & + \textit{female district judge}_i * X_j^{'} \gamma \\ & + \delta_j + \delta_{dt} + \epsilon_{ijdct} \end{aligned}$$

- $\textit{votes to reverse}_{ijdct}$, circuit judge j reviewing district court case i votes to reverse.
- $\textit{female district judge}_i$, lower-court judge is female
- $\textit{slant}_j$, judge j's slant measure.
- X_j, judge j's characteristics.
- δ_j, circuit judge fixed effects
- δ_{dt}, District-year fixed effects

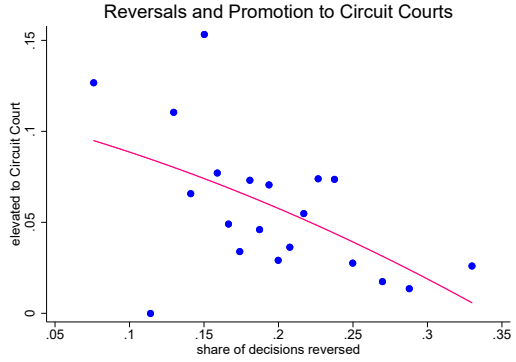
Judges with more gender slant reverse female district judges more

Judges with more gender slant reverse female district judges more

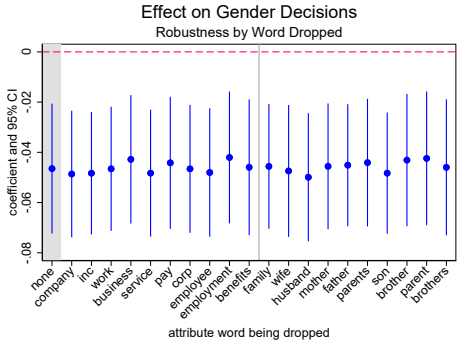
Gender Slant * Female District Judge	**0.010*** **	**0.010*** **	**0.012*** **	**0.012*** **
	(0.004)	**(0.004)**	**(0.004)**	**(0.004)**
Democrat * Female District Judge	-0.009	-0.024**	-0.006	-0.007
	(0.014)	(0.009)	(0.014)	(0.013)
Female * Female District Judge	-0.009	-0.022***	-0.007	-0.011
	(0.009)	(0.008)	(0.009)	(0.010)
Democrat * Female * Female District Judge		0.152***		
		(0.015)		
Liberal Score * Female District Judge			-0.051	
			(0.036)	
Observations	145862	145862	144965	145563
Clusters	133	133	130	133
Outcome Mean	0.177	0.177	0.177	0.177
Circuit-Year FE	X	X	X	X
Judge FE	X	X	X	X
District Judge FE	X	X	X	X
Demographic Controls	X	X	X	X
+ Interactions		X		
District-Year FE				X

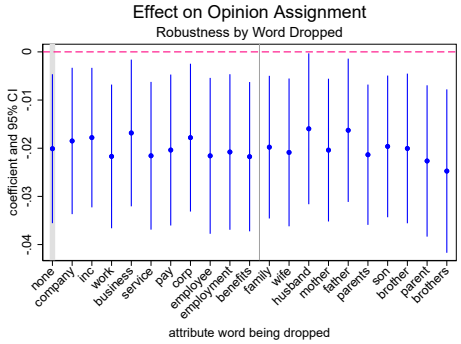
Standard errors clustered by circuit judge. *** p<0.01, ** p<0.05, * p<0.1.

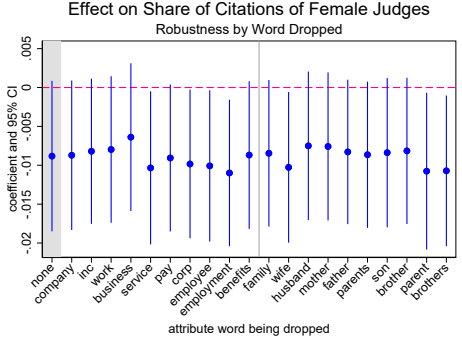
Figure: Reversals and Promotions from District to Circuit Courts

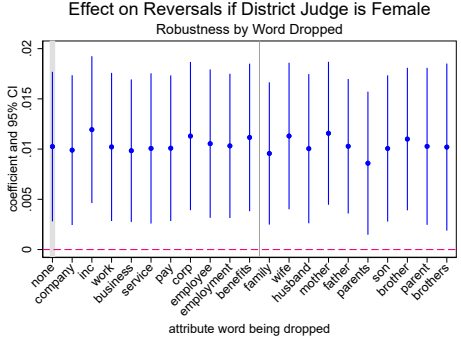


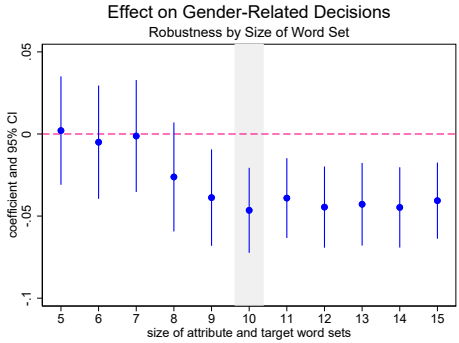
Notes: The graph shows the relationship between the probability of being elevated from a District to a Circuit Court and the share of decisions that were reversed on appeal, conditional on demographic controls and circuit fixed effects. The sample is restricted to district judges for which we observe at least 50 cases.

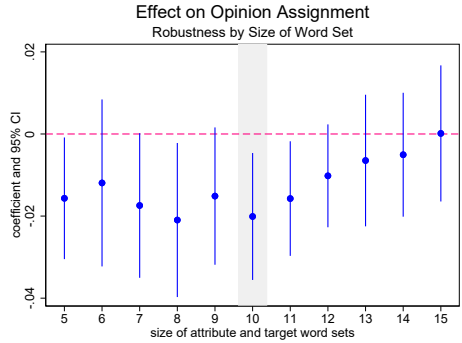

Effect on Gender Decisions
Robustness by Word Dropped
coefficient and 95% CI
0
-.02
-.04
-.06
-.08
none
company
inc
work
business
service
pay
corp
employee
employment
benefits
family
wife
husband
mother
father
parents
son
brother
parent
brothers
attribute word being dropped

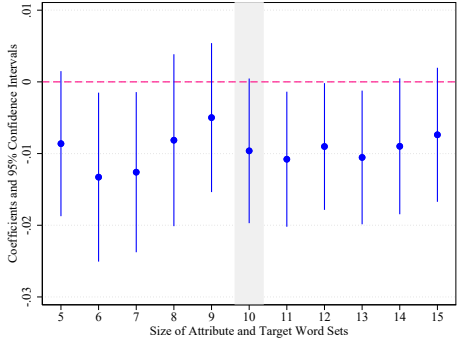

Effect on Opinion Assignment
Robustness by Word Dropped
coefficient and 95% CI
0
-.01
-.02
-.03
-.04
none
company
inc
work
business
service
pay
corp
employee
employment
benefits
family
wife
husband
mother
father
parents
son
brother
parent
brothers
attribute word being dropped

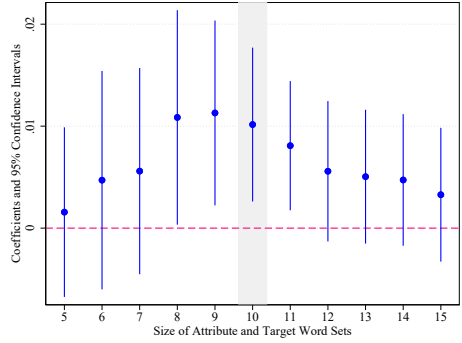

Effect on Share of Citations of Female Judges
Robustness by Word Dropped
coefficient and 95% CI
.005
0
-.005
-.01
-.015
-.02
none
company
inc
work
business
service
pay
corp
employee
employment
benefits
family
wife
husband
mother
father
parents
son
brother
parent
brothers
attribute word being dropped


Effect on Reversals if District Judge is Female
Robustness by Word Dropped
coefficient and 95% CI
.02
.015
.01
.005
0
none
company
inc
work
business
service
pay
corp
employee
employment
benefits
family
wife
husband
mother
father
parents
son
brother
parent
brothers
attribute word being dropped


Effect on Gender-Related Decisions
Robustness by Size of Word Set
coefficient and 95% CI
.05
0
-.05
-.1
5 6 7 8 9 10 11 12 13 14 15
size of attribute and target word sets


Effect on Opinion Assignment
Robustness by Size of Word Set
coefficient and 95% CI
.02
0
-.02
-.04
5 6 7 8 9 10 11 12 13 14 15
size of attribute and target word sets


Coefficients and 95% Confidence Intervals
.01
0
-.01
-.02
-.03
5 6 7 8 9 10 11 12 13 14 15
Size of Attribute and Target Word Sets


Coefficients and 95% Confidence Intervals
.02
.01
0
5 6 7 8 9 10 11 12 13 14 15
Size of Attribute and Target Word Sets

Magnitudes

- Having a daughter $\rightarrow$

Magnitudes

- Having a daughter $\rightarrow$
  - 0.5 standard deviation lower gender slant
  - $>$ Democrat effect; $\sim$ female effect

Magnitudes

- Having a daughter $\rightarrow$
  - 0.5 standard deviation lower gender slant
  - $>$ Democrat effect; $\sim$ female effect
- Two standard deviations increase in gender slant $\rightarrow$

Magnitudes

- Having a daughter $\rightarrow$
  - 0.5 standard deviation lower gender slant
  - $>$ Democrat effect; $\sim$ female effect
- Two standard deviations increase in gender slant $\rightarrow$
  1. 20% lower likelihood of pro-women's rights vote
  - $\sim\frac{2}{3}$ of Democrat effect; $\sim$ female effect

Magnitudes

- Having a daughter $\rightarrow$
  - 0.5 standard deviation lower gender slant
  - $>$ Democrat effect; $\sim$ female effect
- Two standard deviations increase in gender slant $\rightarrow$
  1. 20% lower likelihood of pro-women's rights vote
     - $\sim\frac{2}{3}$ of Democrat effect; $\sim$ female effect
  2. 10% lower likelihood of female assigned authorship
     - $\sim$ Democrat effect; $\sim\frac{1}{3}$ of female effect

Magnitudes

- Having a daughter $\rightarrow$
  - 0.5 standard deviation lower gender slant
  - $>$ Democrat effect; $\sim$ female effect
- Two standard deviations increase in gender slant $\rightarrow$
  1. 20% lower likelihood of pro-women's rights vote
     - $\sim\frac{2}{3}$ of Democrat effect; $\sim$ female effect
  2. 10% lower likelihood of female assigned authorship
     - $\sim$ Democrat effect; $\sim\frac{1}{3}$ of female effect
  3. 6% lower likelihood of citing a female
     - $\sim$ Democrat effect; $\sim\frac{1}{6}$ of female effect

Magnitudes

- Having a daughter $\rightarrow$
  - 0.5 standard deviation lower gender slant
  - $>$ Democrat effect; $\sim$ female effect
- Two standard deviations increase in gender slant $\rightarrow$

1. 20% lower likelihood of pro-women's rights vote
   - $\sim\frac{2}{3}$ of Democrat effect; $\sim$ female effect
2. 10% lower likelihood of female assigned authorship
   - $\sim$ Democrat effect; $\sim\frac{1}{3}$ of female effect
3. 6% lower likelihood of citing a female
   - $\sim$ Democrat effect; $\sim\frac{1}{6}$ of female effect
4. 10% more likely to reverse a female
   - $>$ Democrat effect, $>$ female effect $\rightarrow$ could be important for career prospects.

Discussion

Discussion

- What are we measuring?
  - Are these implicit attitudes?
  - How does our measure correlate with actual IAT scores?

Discussion

- What are we measuring?
  - Are these implicit attitudes?
  - How does our measure correlate with actual IAT scores?
- Other forms of slant?
  - e.g. racial sentiment

Discussion

- What are we measuring?
  - Are these implicit attitudes?
  - How does our measure correlate with actual IAT scores?
- Other forms of slant?
  - e.g. racial sentiment
- Relevant in other domains?
  - Preliminary analysis on congressional speech shows similar results
  - Also looking at district judges to relate to sentencing

Discussion

- What are we measuring?
  - Are these implicit attitudes?
  - How does our measure correlate with actual IAT scores?
- Other forms of slant?
  - e.g. racial sentiment
- Relevant in other domains?
  - Preliminary analysis on congressional speech shows similar results
  - Also looking at district judges to relate to sentencing
- Does language matter?
  - Djourelova (2019): style change from "illegal" to "undocumented" immigrant softened attitudes toward immigration.
  - what about slanted law? on judges, lawyers, law students, etc.

Outline

https://xkcd.com/

Why interpretability?

- In machine learning, helps with debugging.

Why interpretability?

- In machine learning, helps with debugging.
- In research, helps with measurement validity.

Why interpretability?

- In machine learning, helps with debugging.
- In research, helps with measurement validity.
- In applications, can be helpful to users.

Why interpretability?

- In machine learning, helps with debugging.
- In research, helps with measurement validity.
- In applications, can be helpful to users.
- In decision systems, can help subjects feel fairly treated.

Some features of “good” explanations

- A good explanation answers a “why” question.

Some features of “good” explanations

- A good explanation answers a “why” question.
- **Contrastive**: they explain not just why a certain prediction was made, but why it was made instead of other predictions.

Some features of “good” explanations

- A good explanation answers a “why” question.
- **Contrastive**: they explain not just why a certain prediction was made, but why it was made instead of other predictions.
- **Selective**: explanations should be short.

Some features of "good" explanations

- A good explanation answers a "why" question.
- **Contrastive**: they explain not just why a certain prediction was made, but why it was made instead of other predictions.
- **Selective**: explanations should be short.
- **Social**: explanations should be targeted to the relevant audience.

Some features of "good" explanations

- A good explanation answers a "why" question.
- **Contrastive**: they explain not just why a certain prediction was made, but why it was made instead of other predictions.
- **Selective**: explanations should be short.
- **Social**: explanations should be targeted to the relevant audience.
- **Outlier-focused**: if one of the input features is abnormal, that should be the focus of the explanation.

Interpretable Models

Algorithm	Linear	Monotone	Interaction
Linear regression	X	X	
Logistic regression		X	
Decision trees		~	X
k-nearest neighbors			

- **Linearity**: association between features and target is modelled linearly.
  - in addition, L1 penalty can enforce sparsity.

Interpretable Models

Algorithm	Linear	Monotone	Interaction
Linear regression	X	X	
Logistic regression		X	
Decision trees		~	X
k-nearest neighbors			

- **Linearity**: association between features and target is modelled linearly.
  - in addition, L1 penalty can enforce sparsity.
- **Monotonicity**: the relationship between a feature and the target outcome always goes in the same direction over the entire range of the feature.

Interpretable Models

Algorithm	Linear	Monotone	Interaction
Linear regression	X	X	
Logistic regression		X	
Decision trees		~	X
k-nearest neighbors			

- **Linearity**: association between features and target is modelled linearly.
  - in addition, L1 penalty can enforce sparsity.
- **Monotonicity**: the relationship between a feature and the target outcome always goes in the same direction over the entire range of the feature.
- **No interactions:** allowing interactions between features improves predictive performance but hurts interpretability.

Global Surrogate Model

1. Get predictions $\hat{y}$ of the black box model from the data X.

Global Surrogate Model

1. Get predictions $\hat{y}$ of the black box model from the data X.
2. Train an interpretable model (lasso, decision tree, etc) on X with $\hat{y}$ as the label.
   - This is the surrogate model!

Global Surrogate Model

1. Get predictions $\hat{y}$ of the black box model from the data X.
2. Train an interpretable model (lasso, decision tree, etc) on X with $\hat{y}$ as the label.
   - This is the surrogate model!
3. Validate that the surrogate model replicates the predictions of the black box model
   - e.g., compute R^2 or $F1$ between black box $\hat{y}$ and surrogate $\hat{\hat{y}}$.
   - doesn't need to be in held out test set.

Local Surrogate (LIME)

- LIME = local interpretable model-agnostic explanations.
- Isolates the features which are most important at a particular data point.

Local Surrogate (LIME)

- LIME = local interpretable model-agnostic explanations.
- Isolates the features which are most important at a particular data point.

1. Select data point to explain

Local Surrogate (LIME)

- LIME = local interpretable model-agnostic explanations.
- Isolates the features which are most important at a particular data point.

1. Select data point to explain
2. Perturb dataset (locally) and get black box predictions for the new points.

Local Surrogate (LIME)

- LIME = local interpretable model-agnostic explanations.
- Isolates the features which are most important at a particular data point.

1. Select data point to explain
2. Perturb dataset (locally) and get black box predictions for the new points.
3. Train an interpretable surrogate model on the perturbed dataset (weighted by proximity to initial data point).
   - This is the "local" surrogate model.
   - use lasso with high L1 penalty to get a sparse explanation.

LIME for Text

1. Generate new texts by randomly *removing* words from the original document.

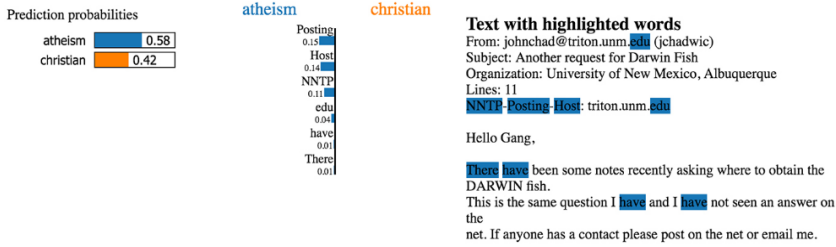
LIME for Text

1. Generate new texts by randomly *removing* words from the original document.
2. Form predictions $\hat{y}$ from black box model for these perturbed documents.

LIME for Text

1. Generate new texts by randomly *removing* words from the original document.
2. Form predictions $\hat{y}$ from black box model for these perturbed documents.
3. Train lasso on dataset of binary features for each word, equaling one if word appears, to predict $\hat{y}$.
   - weight by proximity to initial data point (one minus the proportion of words dropped)

```
exp = explainer.explain_instance(test_example,
classifier.predict_proba, num_features=6)
```



Practical Advice for Research and Applications

1. for gradient boosting, use the contained feature importance.
2. for regression, examine coefficients
3. look at highest and lowest ranked documents for $\hat{y}$
4. report a few example documents with LIME highlighting