Building a Robot Judge: Data Science for the Law

10. Document Embeddings

Elliott Ash

Vectorizing Documents

- Quantitative analysis of language requires that documents be transformed to numbers – that is, vectors.
- We started with the baseline approach: documents become sparse vectors of token counts/frequencies.
  - high-dimensionality can cause issues, but sparsity mitigates.
  - can use documents of arbitrary length
  - can capture local word order with n-grams, but long-run word order is lost.

Embedding layers

- Last class, we introduced embedding layers:
  - take the whole document as input, pad documents to the same length, and represent the document as a flattened series of embedding vectors.
  - potentially captures information on long-range ordering of features in documents
  - DNNs work better with dense vectors
  - computationally demanding
  - only works with short documents

From Word Vectors to Document Vectors

$$\vec{D} = \sum_{w \in D} a_w \vec{w}$$

- The "continuous bag of words" representation for document D is the sum, or the average (potentially weighted by a_w), of the vectors $\vec{w}$ for each word w in ahe document.
  - word vectors $\vec{w}$ constructed using Word2Vec or GloVe (pre-trained or trained on the corpus).
  - "Document" could be sentence, paragraph, section, etc.

Document Vectors

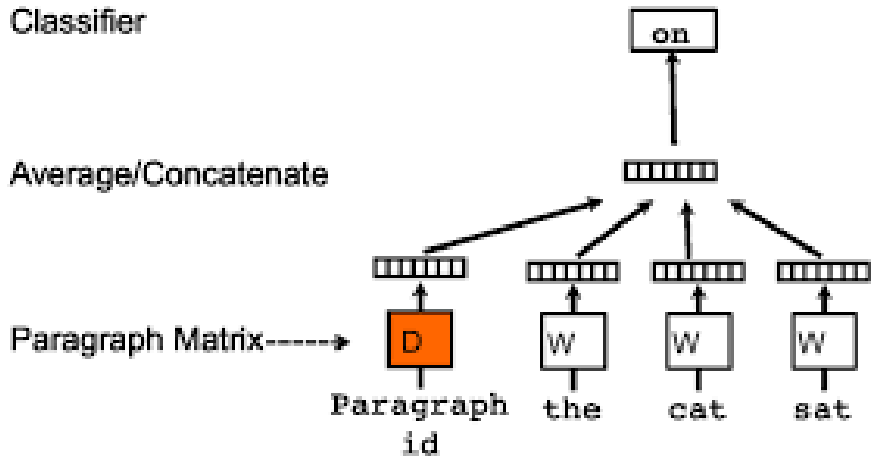
$$\vec{D} = \sum_{w \in D} a_w \vec{w}$$

- Can filter tokens:
  - drop stopwords
  - filter on parts of speech (e.g., keep only nouns, adjectives, and verbs)
- Token weighting:
  - set a_w to weight words by inverse term frequency or inverse document frequency (that is, up-weight rare/informative words)
  - Arora, Liang, and Ma (2016) provide a "tough to beat baseline", the SIF-weighted ("smoothed inverse frequency") average of the vectors:

$$a_w = \frac{\alpha}{\alpha + p_w}$$

where p_w is the probability (frequency) of the word and $\alpha = .001$ is a smoothing parameter.

Doc2Vec (Le and Mikolov)



- Doc2Vec generalizes Word2Vec to documents:
  - predict a word using both the immediate neighbors, as well as **a bag-of-words representation of the whole document**.
- In Doc2Vec, both words **and documents** are assigned a learned vector representation through an embedding layer.

Document Embeddings Geometry

- Just as directions in word space encode semantic information about the words, directions in document space encode topical information about the documents.
- In topic models, each dimension has a topical interpretation; in document embeddings, a direction (might) have a topical interpretation.

Doc2Vec in gensim

- syntax is the same as Word2Vec.
- can train both document vectors and word vectors.
- can get similarity between documents, and use clustering to get groups of related documents.

Doc2Vec on Wikipedia

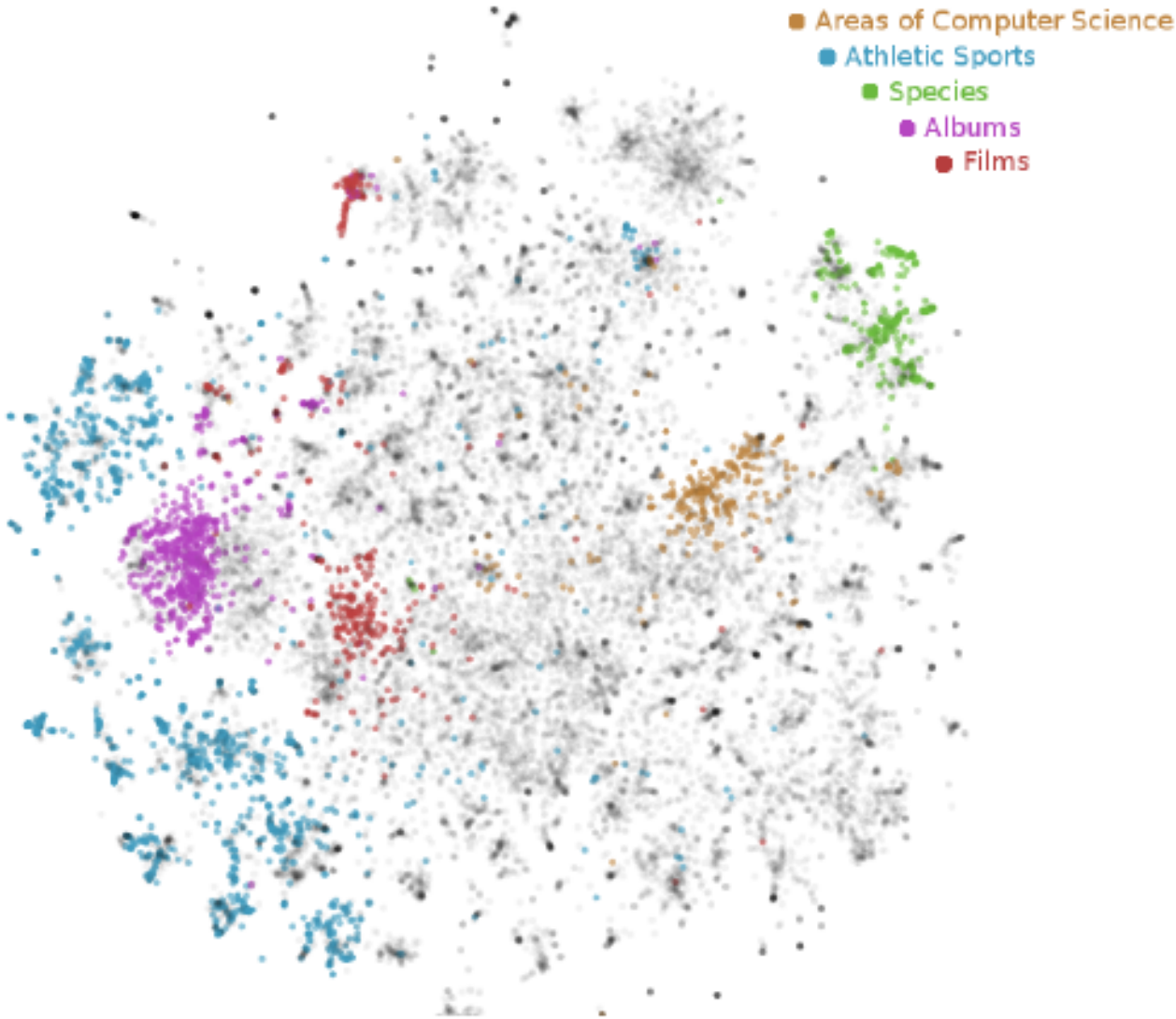


Figure 3: Visualization of Wikipedia paragraph vectors using t-SNE.

Table 1: Nearest neighbours to "Machine learning." Bold face texts are articles we found unrelated to "Machine learning." We use Hellinger distance for LDA and cosine distance for Paragraph Vectors as they work the best for each model.

LDA	**Paragraph Vectors**
Artificial neural network	Artificial neural network
Predictive analytics	Types of artificial neural networks
Structured prediction	Unsupervised learning
Mathematical geophysics	Feature learning
Supervised learning	Predictive analytics
Constrained conditional model	Pattern recognition
Sensitivity analysis	Statistical classification
SXML	Structured prediction
Feature scaling	Training set
Boosting (machine learning)	Meta learning (computer science)
Prior probability	Kernel method
Curse of dimensionality	Supervised learning
Scientific evidence	Generalization error
Online machine learning	Overfitting
N-gram	Multi-task learning
Cluster analysis	Generative model
Dimensionality reduction	Computational learning theory
Functional decomposition	Inductive bias
Bayesian network	Semi-supervised learning

Table 5: arXiv nearest neighbours to "Distributed Representations of Sentences and Documents" using Paragraph Vectors.

Title	Cosine Similarity
Evaluating Neural Word Representations in Tensor-Based Compositional Settings	0.771
Polyglot: Distributed Word Representations for Multilingual NLP	0.764
Lexicon Infused Phrase Embeddings for Named Entity Resolution	0.757
A Convolutional Neural Network for Modelling Sentences	0.747
Distributed Representations of Words and Phrases and their Compositionality	0.740
Convolutional Neural Networks for Sentence Classification	0.735
SimLex-999: Evaluating Semantic Models With (Genuine) Similarity Estimation	0.735
Exploiting Similarities among Languages for Machine Translation	0.731
Efficient Estimation of Word Representations in Vector Space	0.727
Multilingual Distributed Representations without Word Alignment	0.721

Table 2: Wikipedia nearest neighbours

(a) Wikipedia nearest neighbours to "Lady Gaga" using Paragraph Vectors. All articles are relevant.

Article	Cosine Similarity
Christina Aguilera	0.674
Beyonce	0.645
Madonna (entertainer)	0.643
Artpop	0.640
Britney Spears	0.640
Cyndi Lauper	0.632
Rihanna	0.631
Pink (singer)	0.628
Born This Way	0.627
The Monster Ball Tour	0.620

(b) Wikipedia nearest neighbours to "Lady Gaga" - "American" + "Japanese" using Paragraph Vectors. Note that Ayumi Hamasaki is one of the most famous singers, and one of the best selling artists in Japan. She also has an album called "Poker Face" in 1998.

Article	Cosine Similarity
Ayumi Hamasaki	0.539
Shoko Nakagawa	0.531
Izumi Sakai	0.512
Urbangarde	0.505
Ringo Sheena	0.503
Toshiaki Kasuga	0.492
Chihiro Onitsuka	0.487
Namie Amuro	0.485
Yakuza (video game)	0.485
Nozomi Sasaki (model)	0.485

Table 7: arXiv nearest neighbours to "Distributed Representations of Sentences and Documents" - "neural" + "Bayesian". I.e., the Bayesian equivalence of the Paragraph Vector paper.

Title	Cosine Similarity
Content Modeling Using Latent Permutations	0.629
SimLex-999: Evaluating Semantic Models With (Genuine) Similarity Estimation	0.611
Probabilistic Topic and Syntax Modeling with Part-of-Speech LDA	0.579
Evaluating Neural Word Representations in Tensor-Based Compositional Settings	0.572
Syntactic Topic Models	0.548
Training Restricted Boltzmann Machines on Word Observations	0.548
Discrete Component Analysis	0.547
Resolving Lexical Ambiguity in Tensor Regression Models of Meaning	0.546
Measuring political sentiment on Twitter: factor-optimal design for multinomial inverse regression	0.544
Scalable Probabilistic Entity-Topic Modeling	0.541

- An interesting factor in political psychology is the role of **cogition and emotion** in political messaging.
  - What works better: a **logical argument** or an **emotional appeal**?
- In this paper, we build a new measure of cognitive/emotive valence in language and apply it to speeches of U.S. Congress members.

Cognitive/Affective Dictionary

- Dictionary: a new domain-appropriate list of words for:
  - **Cognitive Processing ("thinking")**: insight, causation, discrepancy, tentativeness, certainty, inhibition, inclusion, and exclusion
  - **Affective Processing ("feeling"):** positive and negative emotions, pleasure, pain, happiness, anxiety, anger, and sadness.

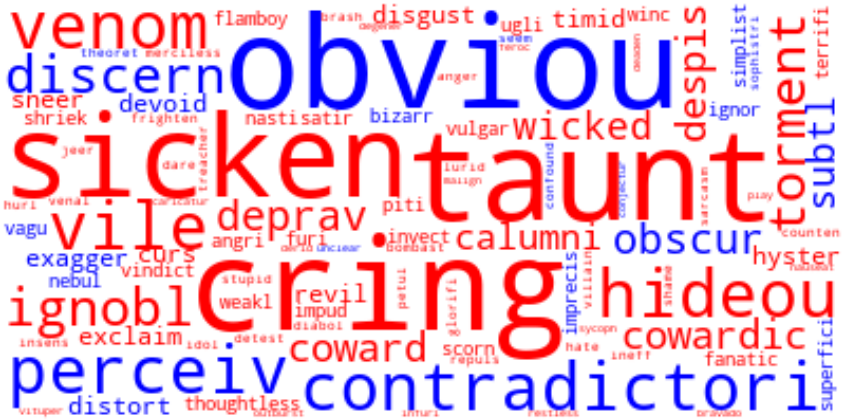
- Drawn from LIWC, but many false positives removed (e.g., "admir*" matches to "admiral", so that's dropped).

Word Embeddings and Metric

- We train word2vec on all speeches in U.S. House of Representatives and Senate for the years 1994-2014.
- We get the embeddings for each word in our dictionaries, and take the centroid (average) for affective $\vec{A}$ and cognitive $\vec{C}$.
- Next: let $\vec{d}_i$ be the SIF-weighted average of the embeddings for the words in document i
- Then: the measure of relative emotionality in i is

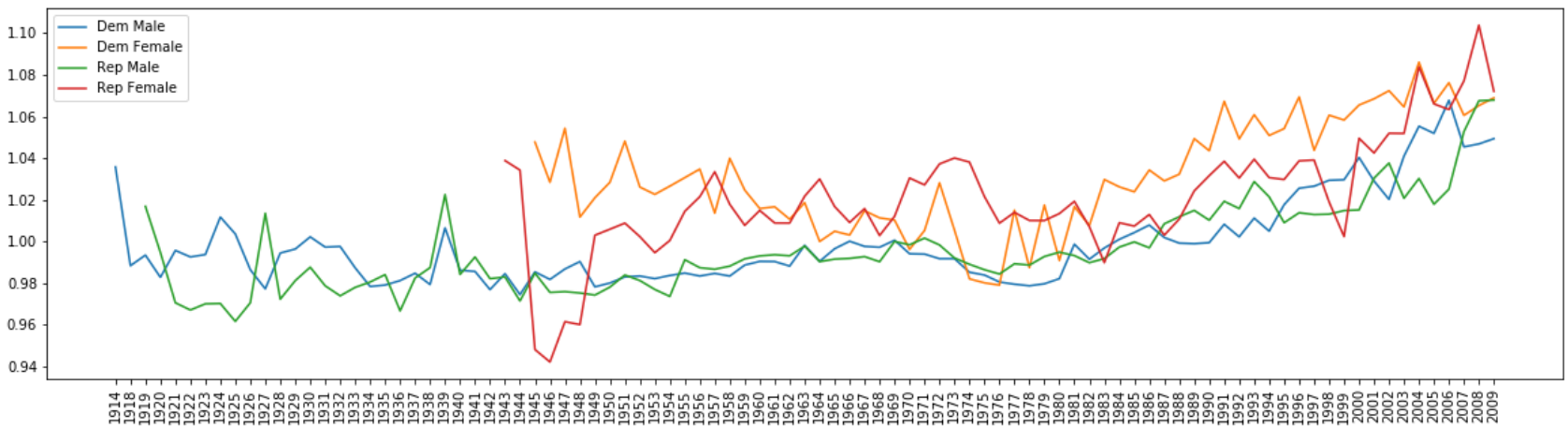
$$Y_i = \frac{\text{sim}(\vec{d}_i, \vec{A}) + 1}{\text{sim}(\vec{d}_i, \vec{C}) + 1}$$

where sim() is cosine similarity.



- Top cognitive sentences:
  - "In my judgment, neither is true in the case of this amendment."
  - "Is that correct?"
  - "R. 15 contains a provision that is similar but, in fact, broader in scope."
- Top emotional sentences:
  - "There is nothing to trouble any heart, nothing to hurt at all; death is only a quiet door, in an old garden wall."
  - "With joy in his heart and a smile on his face he graced practically every social occasion with a song."
  - "We Democrats may disagree, but we love our fellow men and we never hate them."

Emotional language in the very long run



Document Vectors for Judicial Opinions

- Ash and Chen (2018) produce document vectors for each case to understand differences between judges and courts.
  - Corpus: 300,000 cases from U.S. Circuit Courts, 1870-2010.
- We de-mean vectors by group (court, topic, or year) to extract relevant information:
  - de-mean by topic-year to distinguish courts.
  - de-mean by court-topic to distingush years.
  - de-mean by court-year to distinguish topics.

Figure 1: Centered by Topic-Year, Averaged by Judge, Labeled by Court

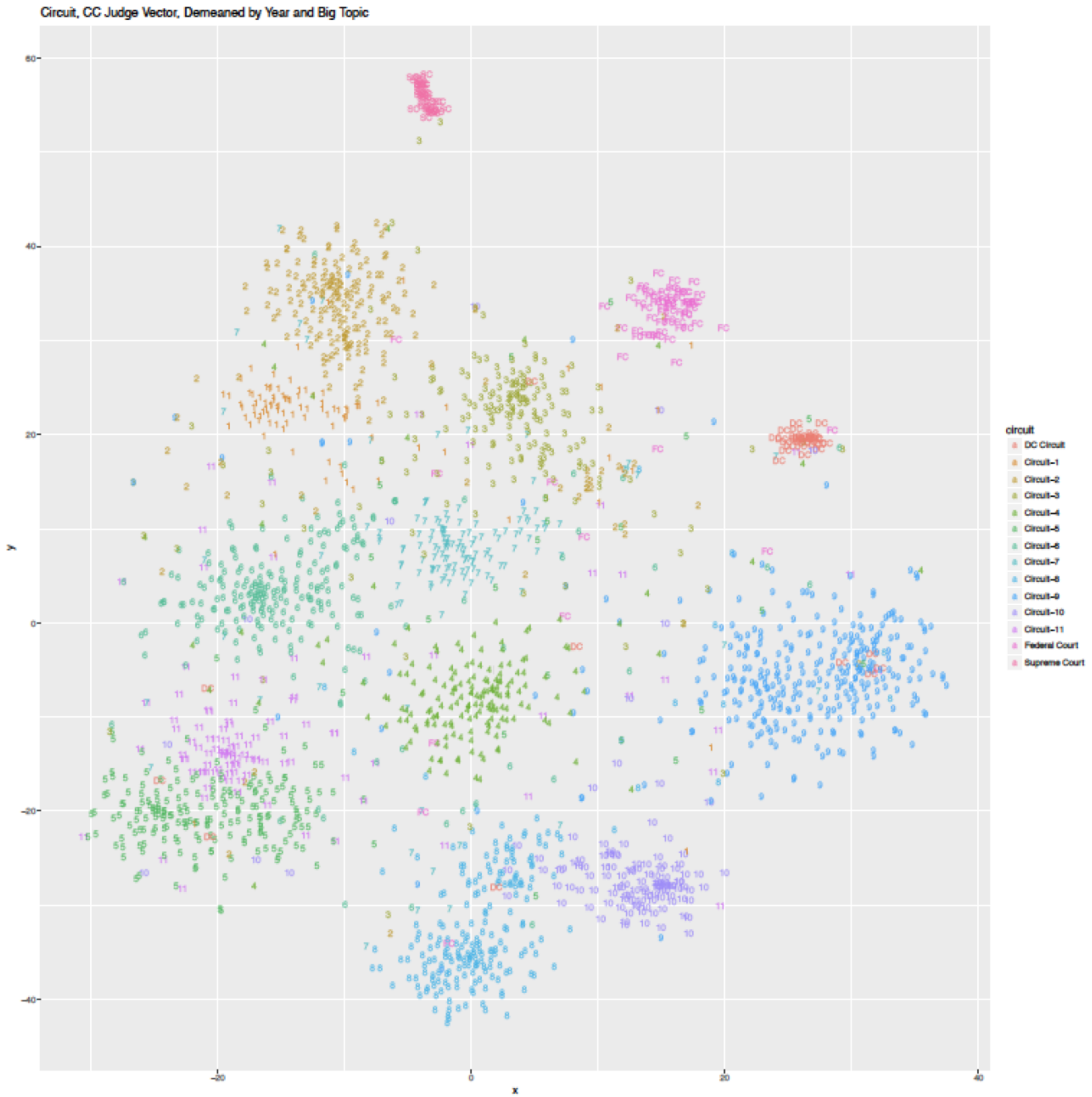


Figure 2: Centered by Court-Topic, Averaged by Court-Year, Labeled by Decade

tsne plot

Figure 3: Centered by Judge-Year, Averaged by Topic-Year, Labeled by Topic

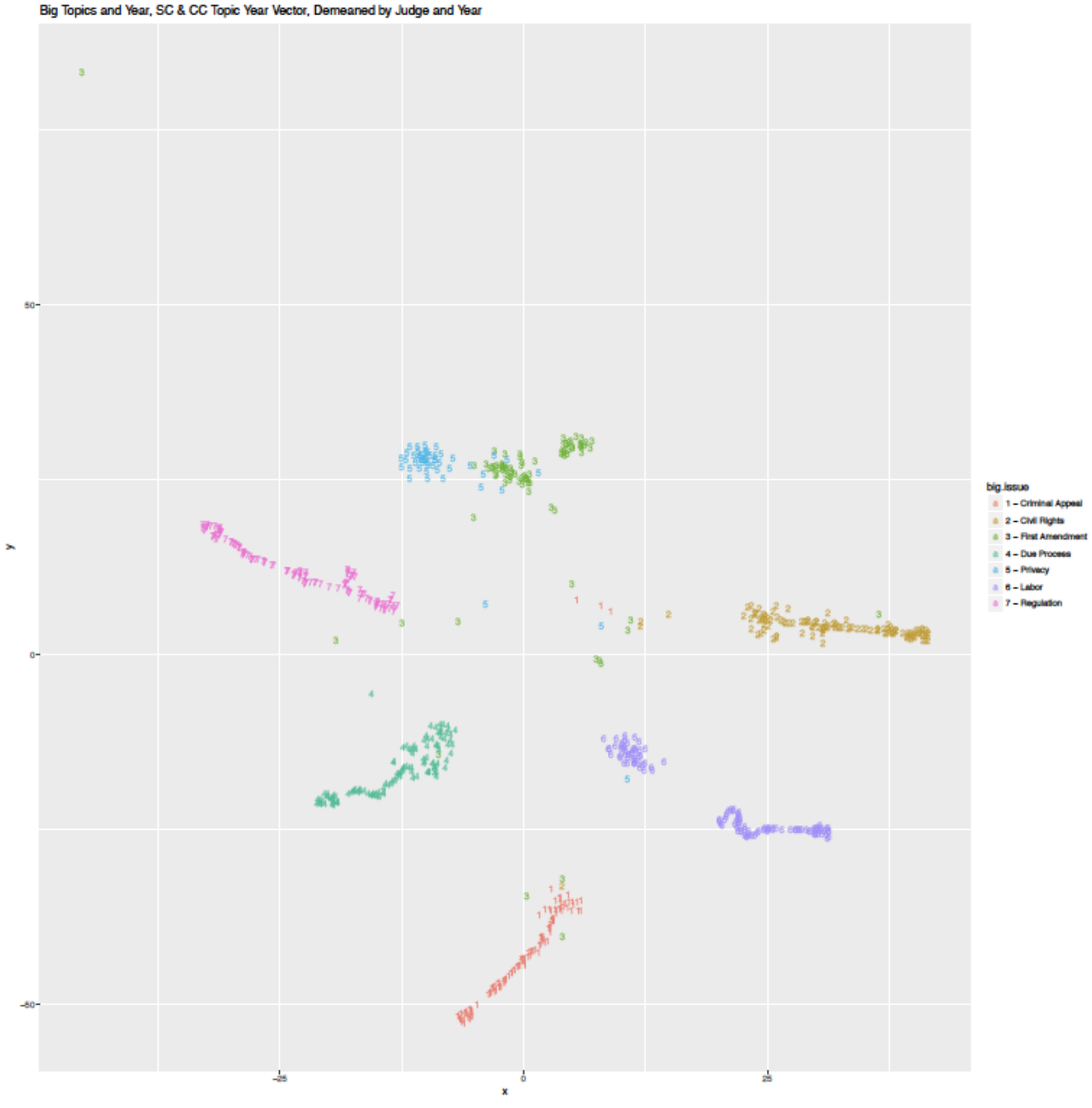
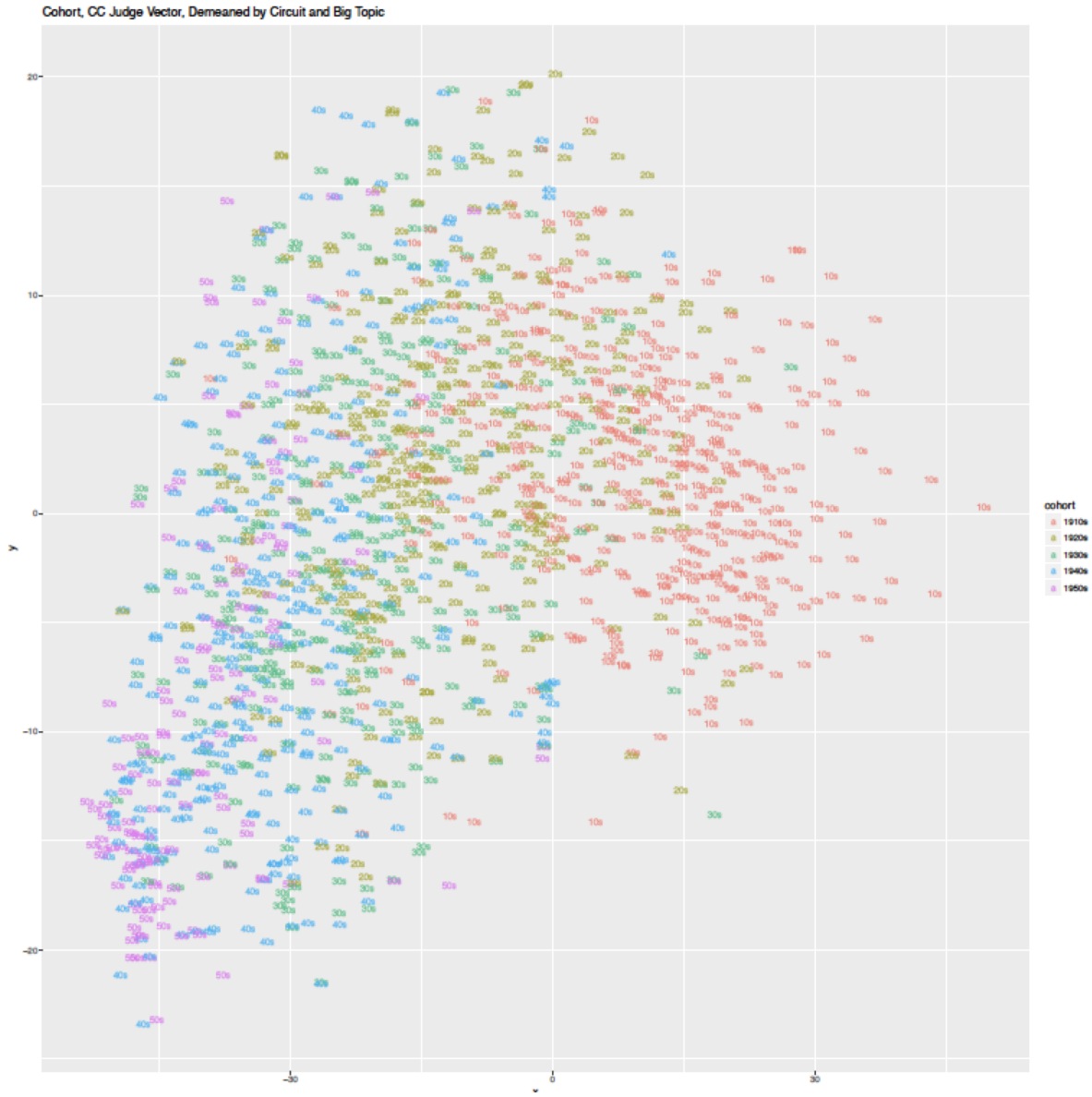


Figure 5: Centered by Court-Topic-Year, Averaged by Judge, Labeled by Judge Birth Cohort



tsne plot

Figure 4: Centered by Court-Topic-Year, Averaged by Judge, Labeled by Political Party

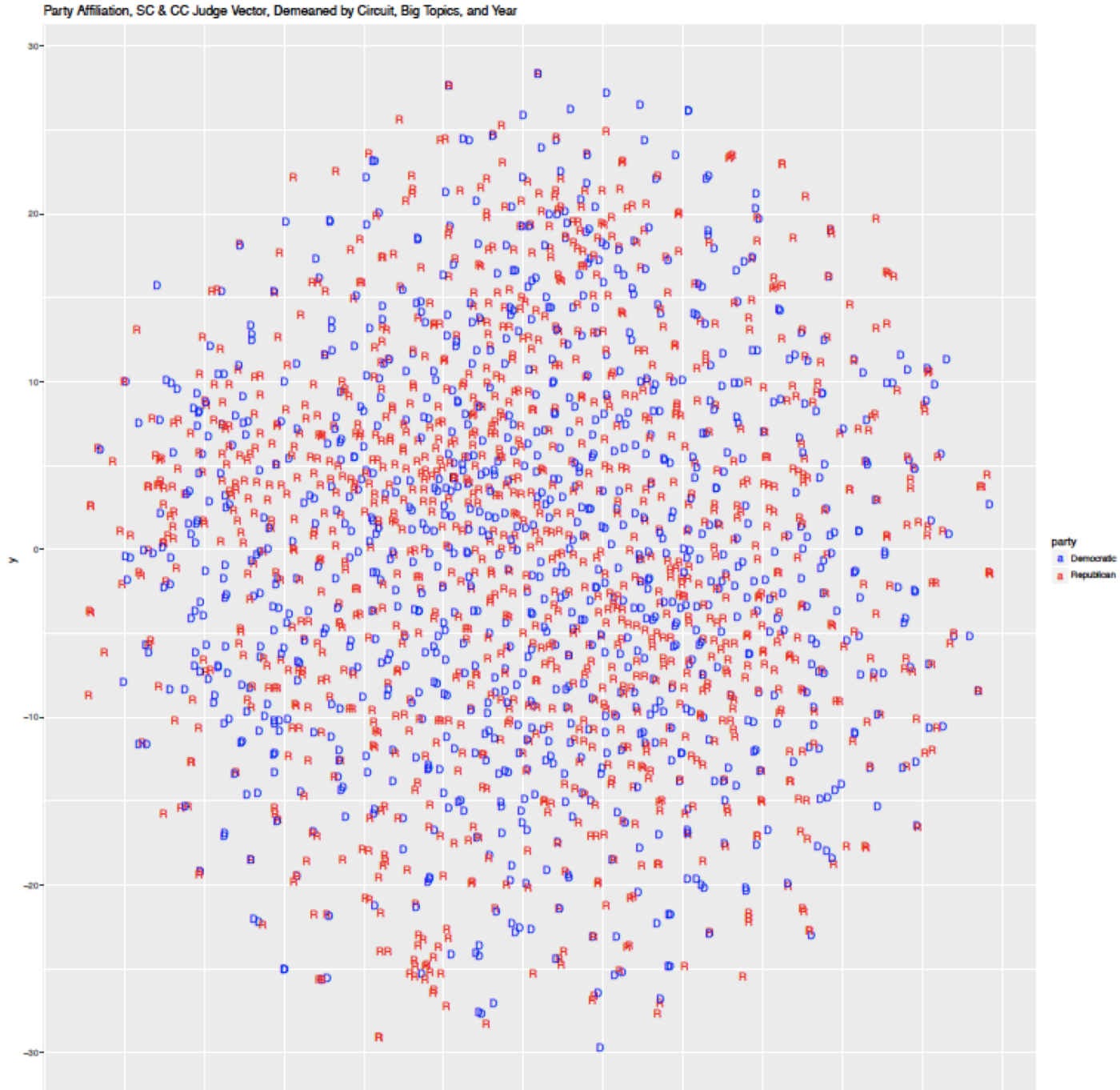
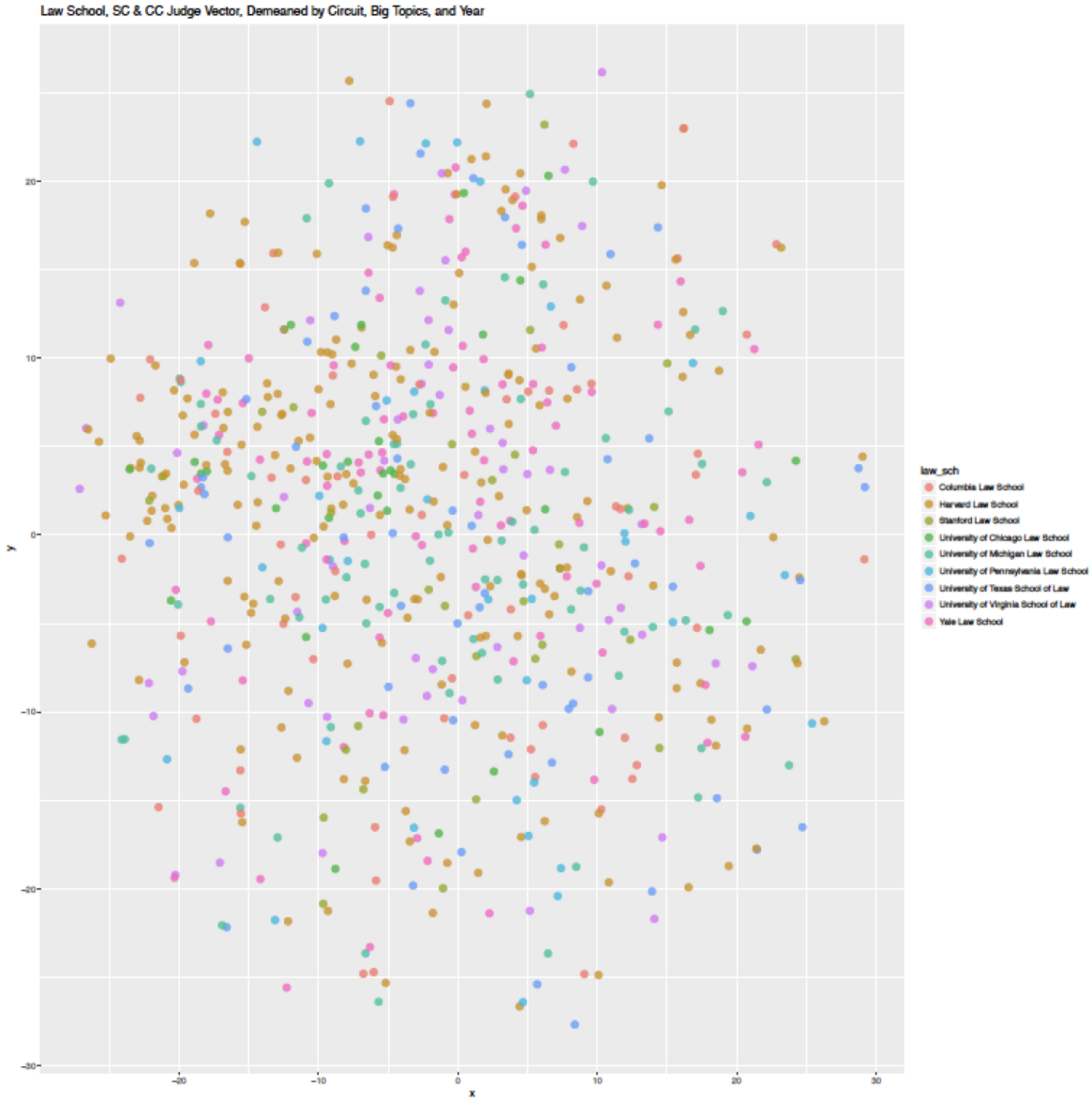


Figure 6: Centered by Court-Topic-Year, Averaged by Judge, Labeled by Law School Attended



Relatedness between judges (e.g. Richard Posner)

Circuit Judge Name	Similarity	Rank	Circuit Judge Name	Similarity	Rank
POSNER, RICHARD A.	1.000	1	TONE, PHILIP W.	0.459	16
EASTERBROOK, FRANK H.	0.663	2	SIBLEY, SAMUEL	0.459	17
SUTTON, JEFFREY S.	0.620	3	SCALIA, ANTONIN	0.456	18
NOONAN, JOHN T.	0.596	4	COLLOTON, STEVEN M.	0.445	19
NELSON, DAVID A.	0.592	5	DUNIWAY, BENJAMIN	0.438	20
CARNES, EDWARD E.	0.567	6	GIBBONS, JOHN J.	0.422	21
FRIENDLY, HENRY	0.566	7	BOGGS, DANNY J.	0.420	22
KOZINSKI, ALEX	0.563	8	BREYER, STEPHEN G.	0.414	23
GORSUCH, NEIL M.	0.559	9	GOODRICH, HERBERT	0.412	24
CHAMBERS, RICHARD H.	0.546	10	LOKEN, JAMES B.	0.410	25
FERNANDEZ, FERDINAND F.	0.503	11	WEIS, JOSEPH F.	0.408	26
EDMONDSON, JAMES L.	0.501	12	SCALIA, ANTONIN (SCOTUS)	0.406	27
KLEINFELD, ANDREW J.	0.491	13	BOUDIN, MICHAEL	0.403	28
WILLIAMS, STEPHEN F.	0.481	14	RANDOLPH, A. RAYMOND	0.397	29
KETHLEDGE, RAYMOND M.	0.459	15	MCCONNELL, MICHAEL W.	0.390	30

Document vectors demeaned by court, year, and topic, then aggregated by judge.

Implicit Association Test (IAT)

Example of a IAT categorization task.
If salary belongs to the right, press "i".
If salary belongs to the left, press "e".

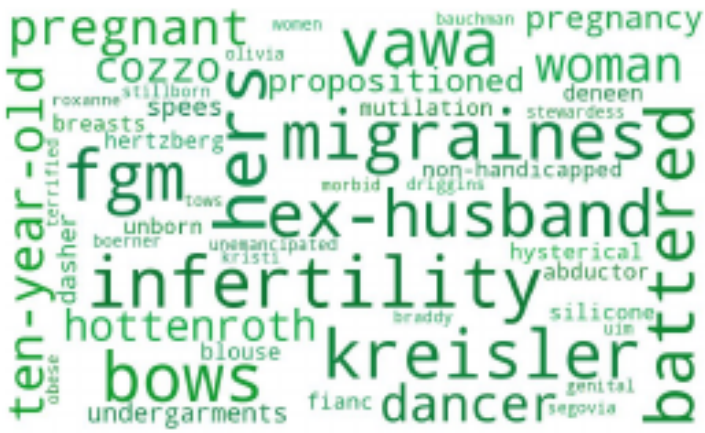
- "Implicit bias": associating attributes with a group more strongly than a comparison group (Greenwald et al, 1998).
- A growing empirical literature in economics has shown that IAT scores are correlated with professionals' behavior:
  - how physicians' make clinical decisions (Green et al, 2007)
  - which candidates receive call-back interviews (Rooth, 2010)
  - manager interactions with staff (Glover et al 2017)
- . . . Could it matter for judges as well?

GloVe Implementation

- Preprocessing:
  - Represent corpus as shuffled list of sentences.
- GloVe embedding hyperparameters:
  - 20 epochs, 300 dimensional vectors, 0.05 learning rate, window of 10 words.
- Steps:
  1. iterate through whole corpus (300K opinions) to get full vocabulary
  2. train judge-specific embeddings
- Bootstrapping approach (Antoniak and Mimno 2018):
  - Sample N_j sentences with replacement, where N_j is the number of sentences written by judge j.
  - 15 bootstrapped samples.

Figure 1: Words with Strongest Male/Female Associations in Judicial Corpus

(a) Male



(b) Female

Figure 2: Words with Strongest Positive/Negative Associations in Judicial Corpus

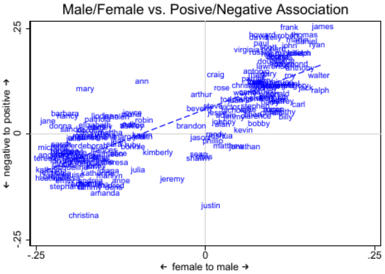
(a) Positive

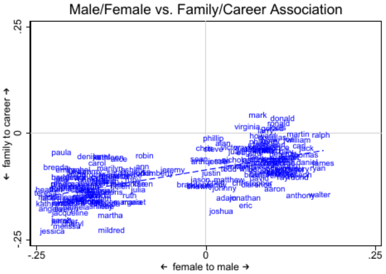
(b) Negative

Figure 3: Words with Strongest Career/Family Associations in Judicial Corpus

(a) Career

(b) Family


Male/Female vs. Posive/Negative Association
← negative to positive →
← female to male →
.25
0
-.25
-.25
0
.25


Male/Female vs. Family/Career Association
← family to career →
← female to male →
.25
0
-.25
-.25
0
.25

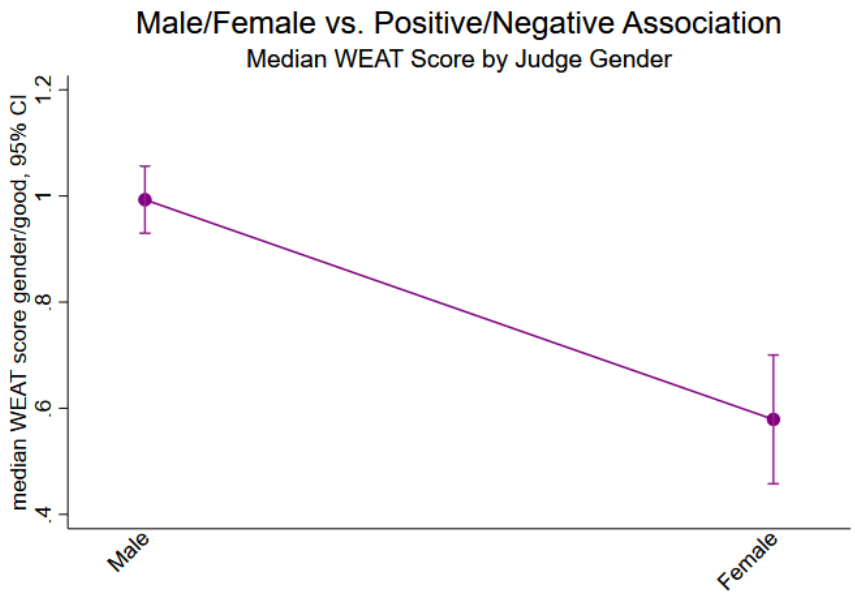
Word Embedding Association Test

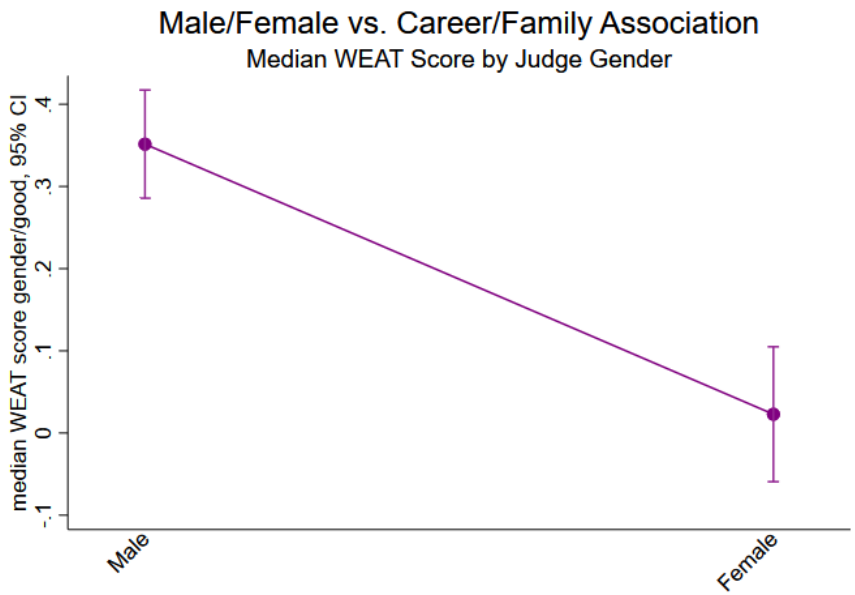
- Schematically, to compare relative association of "woman" to "family" and "man" to "career":

$$\frac{\text{sim(woman,family)}}{\text{sim(woman,career)}} \gtrless \frac{\text{sim(man,family)}}{\text{sim(man,career)}}$$

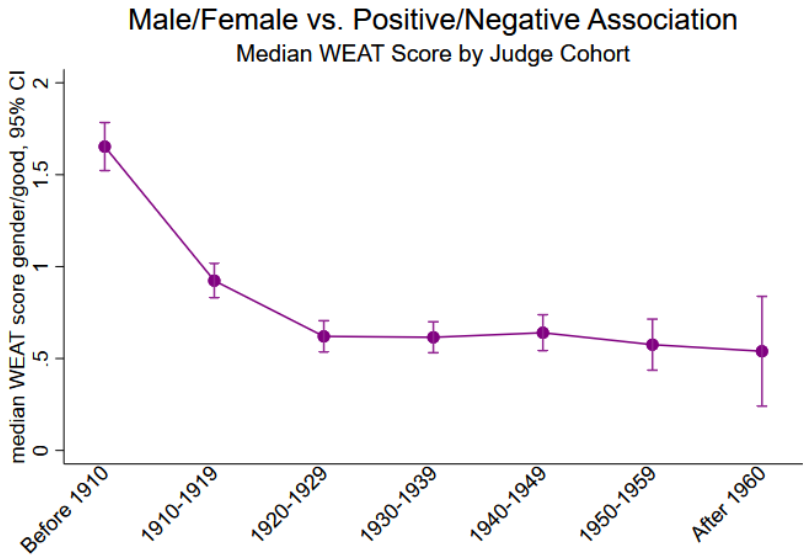
  where sim$(\cdot)$ gives the cosine similarity between the vectors.
- Word Embedding Association Test (WEAT) is a permutation test comparing sets of words in these categories:
  - "woman": woman, female, she, her, ...
  - "man": man, male, he, her, ...
  - "family": family, children, baby, ...
  - "career": career, job, work, ...

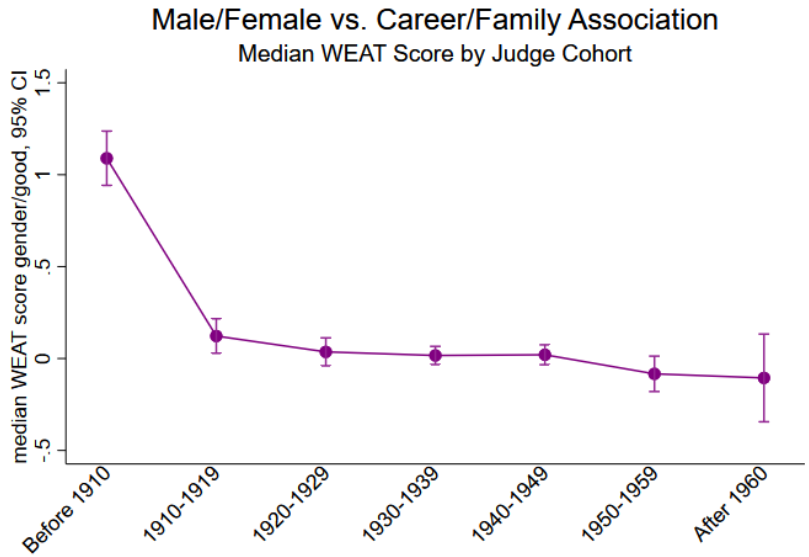
Gender WEAT: By Judge Gender

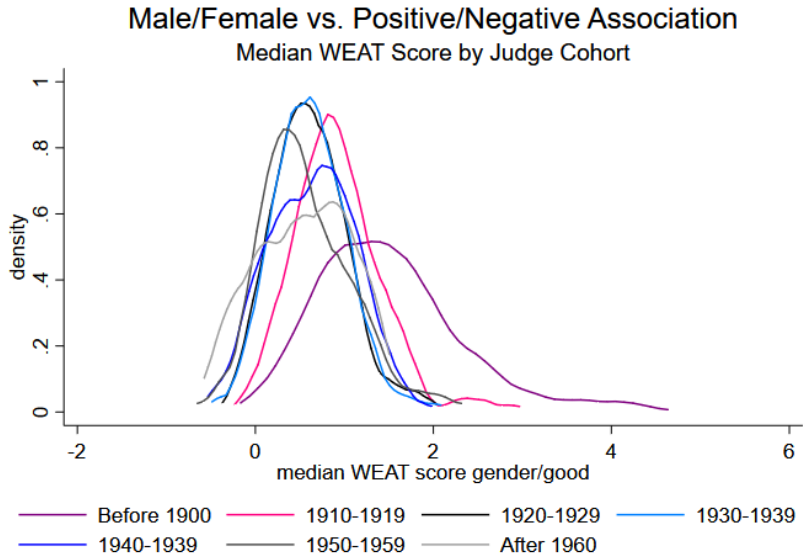


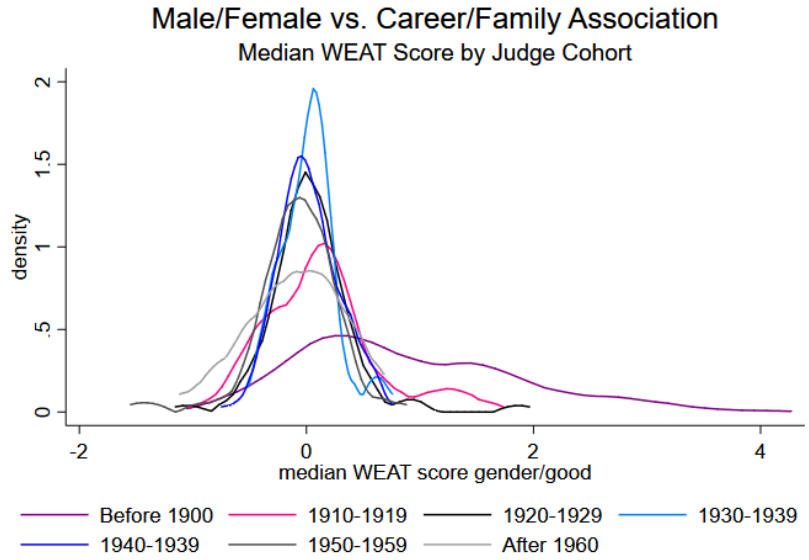


Gender WEAT: By Cohort

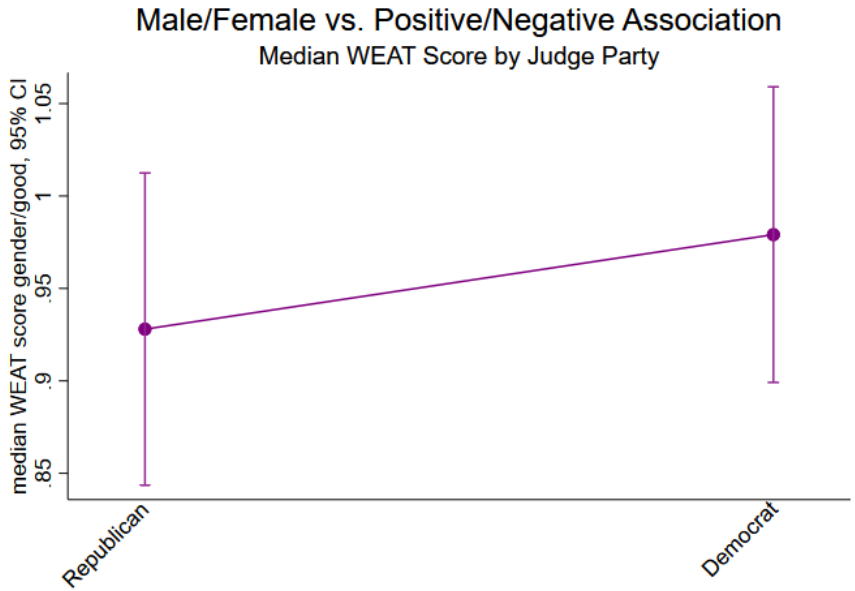


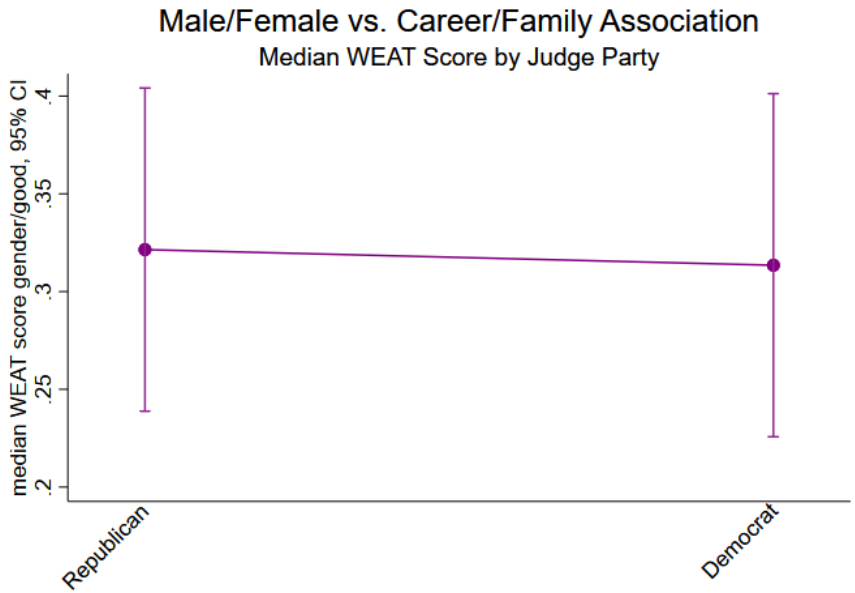




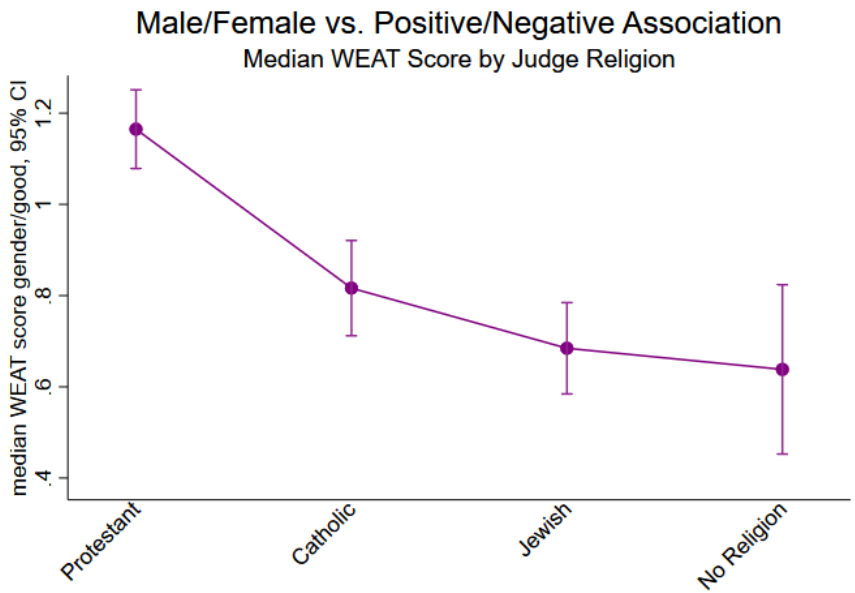


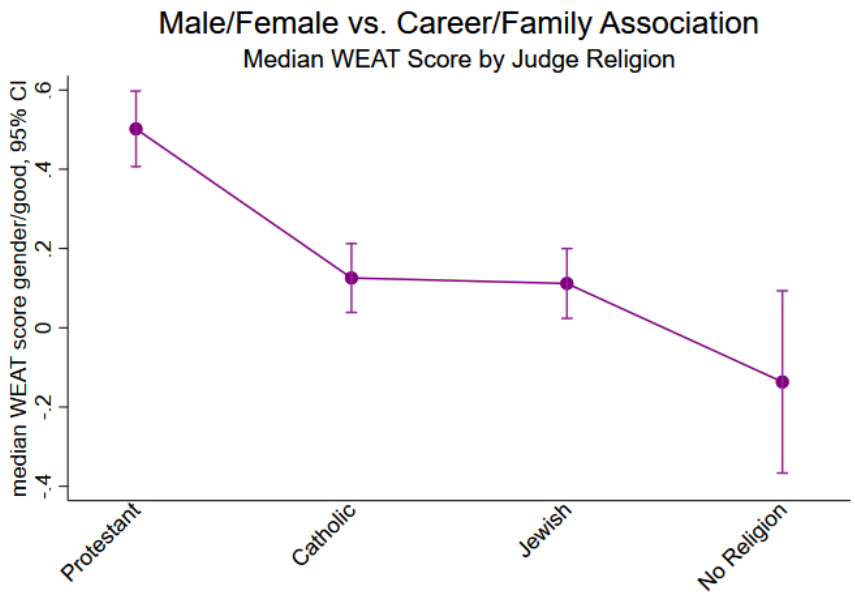
Gender WEAT: By Judge Party



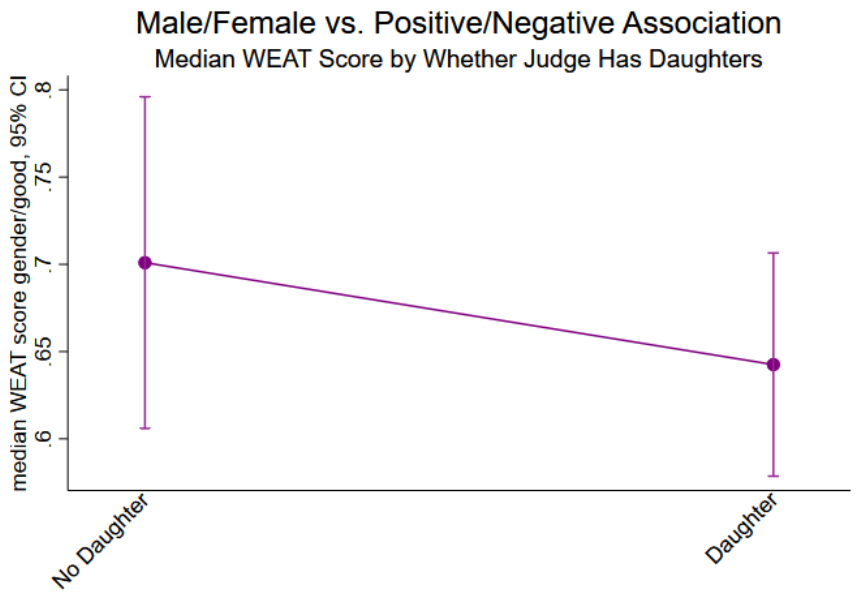


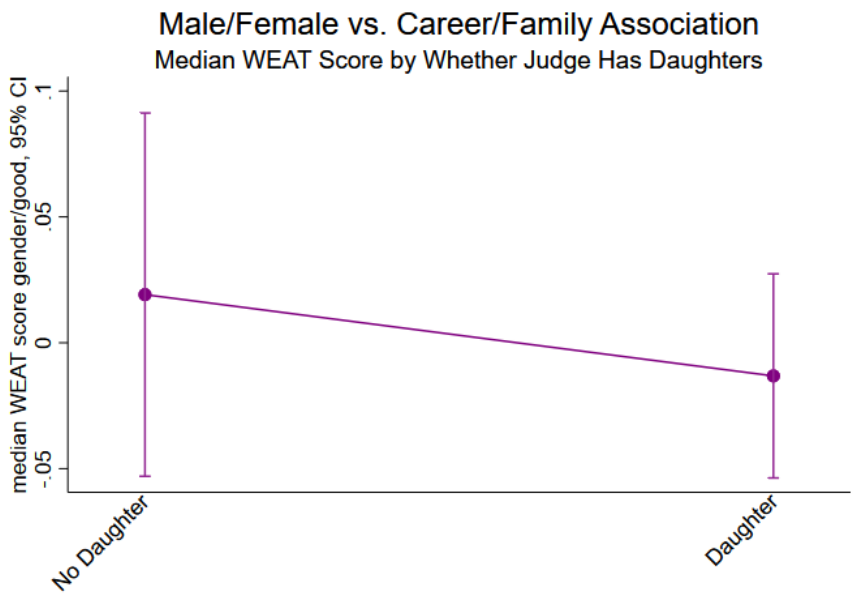
Gender WEAT: By Judge Religion





Gender WEAT: By Judge Having a Daughter





Exposure to female judges reduces language bias

	Male/Female vs. Positive/Negative Association			
	(1)	(2)	(3)	(4)
Share Judge Female	-0.116	-0.361	-0.439*	-0.449*
	(0.129)	(0.200)	(0.167)	(0.150)
Observations	11565	11565	11565	11293
Clusters	10	10	10	10
Year FEs	yes	yes	yes	yes
Circuit FEs		yes	yes	yes
Circuit Trends			yes	yes
Lagged DV				yes

Gender WEAT: Relation to Abortion Decisions

	Abortion		
	(1)	(2)	(3)
Male/Female vs. Positive/Negative Association	-0.098	-0.134*	-0.144*
	(0.072)	(0.077)	(0.084)
Male/Female vs. Career/Family Association	-0.006	0.009	0.062
	(0.111)	(0.111)	(0.124)
Democrat		0.146**	0.185**
		(0.061)	(0.072)
Female		0.135**	0.191**
		(0.064)	(0.078)
Has Daughters			0.078
			(0.076)
Observations	332	330	275
Clusters	176	174	144
Circuit-Year FEs	yes	yes	yes
Issue FEs			
# of Children FEs			yes

Outcome is pro-claimant abortion decision (decision in favor of abortion rights). Standard errors clustered by circuit-year.

Gender WEAT: Relation to Sex Discrimination Decisions

	Sex Discrimination		
	(4)	(5)	(6)
Male/Female vs. Positive/Negative Association	-0.007	-0.020	-0.023
	(0.027)	(0.029)	(0.032)
Male/Female vs. Career/Family Association	-0.099**	-0.102**	-0.105**
	(0.049)	(0.046)	(0.052)
Democrat		0.088***	0.077***
		(0.026)	(0.029)
Female		0.066*	0.074**
		(0.034)	(0.034)
Has Daughters			-0.002
			(0.027)
Observations	1716	1708	1502
Clusters	266	261	210
Circuit-Year FEs	yes	yes	yes
Issue FEs			
# of Children FEs			yes

Outcome is pro-claimant sex-discrimination decision (decision in favor of gender equality). Standard errors clustered by circuit-year.

Reversal of Female Judges

Table 9: Circuit Judge WEAT Scores and Reversal Rates for Female District Judges

	Outcome: Lower Court Reversed				
	(1)	(2)	(3)	(4)	(5)
Female District Judge	-0.0126**	-0.0267**	-0.0255**	-0.0177	-0.0179
	(0.00361)	(0.00568)	(0.00577)	(0.0167)	(0.0168)
WEAT (+/-) Effect Size		-0.000926			
		(0.00976)			
WEAT (+/-) ×		0.0440**	0.0407**	0.0379*	0.0380*
Fem. D. Judge		(0.0145)	(0.0147)	(0.0153)	(0.0154)
WEAT (fam/job) ×					-0.00129
Fem. D. Judge					(0.0101)
N	61014	61014	60956	60956	60956
adj. R-sq	0.029	0.029	0.030	0.030	0.030
Circuit x Year FE	X	X	X	X	X
Judge FE			X	X	X
Bio×Fem. D. Judge				X	X