Sequencing Legal DNA

NLP for Law and Political Economy

3. Dimensionality and Distance

Weekly Q&A Page

bit.ly/NLP-QA03

Different Goals, Different Methods

- Supervised Learning (next week)
  - pursuing a known goal, e.g., predicting whether a political speech is from a Democrat or a Republican.
  - machine learns to replicate labels for new data points

Different Goals, Different Methods

- Supervised Learning (next week)
  - pursuing a known goal, e.g., predicting whether a political speech is from a Democrat or a Republican.
  - machine learns to replicate labels for new data points
- Unsupervised Learning (today)
  - algorithm discovers themes/patterns in text (or other high-dimensional data)
  - human interprets the results (e.g. inspect content of topics or clsuters)

Different Goals, Different Methods

- Supervised Learning (next week)
  - pursuing a known goal, e.g., predicting whether a political speech is from a Democrat or a Republican.
  - machine learns to replicate labels for new data points
- Unsupervised Learning (today)
  - algorithm discovers themes/patterns in text (or other high-dimensional data)
  - human interprets the results (e.g. inspect content of topics or clsuters)
- Both strategies amplify human effort, each in different ways.

Different Goals, Different Methods

- Supervised Learning (next week)
  - pursuing a known goal, e.g., predicting whether a political speech is from a Democrat or a Republican.
  - machine learns to replicate labels for new data points
- Unsupervised Learning (today)
  - algorithm discovers themes/patterns in text (or other high-dimensional data)
  - human interprets the results (e.g. inspect content of topics or clsuters)
- Both strategies amplify human effort, each in different ways.
- Distinctions are not clear-cut:
  - supervised learning models can be used to discover themes/patterns
  - unsupervised learning models can be used in service of prediction or known goals.

Objectives: Social-Science Research using Unsupervised Learning

1. **What is the research question?**

Objectives: Social-Science Research using Unsupervised Learning

1. **What is the research question?**
2. Corpus and Data:
   - obtain, clean, preprocess, and link.
   - Produce descriptive visuals and statistics on the text and metadata

Objectives: Social-Science Research using Unsupervised Learning

1. **What is the research question?**
2. Corpus and Data:
   - obtain, clean, preprocess, and link.
   - Produce descriptive visuals and statistics on the text and metadata
3. Unsupervised learning:
   - **What are we trying to measure?**

Objectives: Social-Science Research using Unsupervised Learning

1. **What is the research question?**
2. Corpus and Data:
   - obtain, clean, preprocess, and link.
   - Produce descriptive visuals and statistics on the text and metadata
3. Unsupervised learning:
   - **What are we trying to measure?**
   - Select a model and train it.
   - Probe sensitivity to hyperparameters.
   - Validate that the model is measuring what we want.

Objectives: Social-Science Research using Unsupervised Learning

1. **What is the research question?**
2. Corpus and Data:
   - obtain, clean, preprocess, and link.
   - Produce descriptive visuals and statistics on the text and metadata
3. Unsupervised learning:
   - **What are we trying to measure?**
   - Select a model and train it.
   - Probe sensitivity to hyperparameters.
   - Validate that the model is measuring what we want.
4. Empirical analysis
   - Produce statistics or predictions with the trained model.
   - **Answer the research question.**

Outline

Text Re-Use

- Text Re-Use algorithms (like “Smith-Waterman”) measure similarity by finding and counting shared sequences in two texts above some minimum length, e.g. 10 words.
  - useful for plagiarism detection, for example.
- precise but slow
  - shortcut: look at proportion of shared (hashed) 5-grams across texts

Cosine Similarity

- Recall in the previous lecture we represented each document i as a vector x_i,
  - for example $x_i =$ term counts or $x_i =$ IDF-weighted term frequencies.

Cosine Similarity

- Recall in the previous lecture we represented each document i as a vector x_i,
  - for example $x_i =$ term counts or $x_i =$ IDF-weighted term frequencies.
- Each document is a non-negative vector in an n_x-space, where $n_x =$ vocabulary size.
  - that is, documents are rays, and similar documents have similar vectors.

Cosine Similarity

- Recall in the previous lecture we represented each document i as a vector x_i,
  - for example x_i = term counts or x_i = IDF-weighted term frequencies.
- Each document is a non-negative vector in an n_x-space, where n_x = vocabulary size.
  - that is, documents are rays, and similar documents have similar vectors.
- Can measure similarity between documents i and j by the cosine of the angle between x_i and x_j :
  - With perfectly collinear documents (that is, $x_i = \alpha x_j$, $\alpha > 0$), $\cos(0) = 1$
  - For orthogonal documents (no words in common), $\cos(\pi/2)=0$

Cosine Similarity

- Recall in the previous lecture we represented each document i as a vector x_i,
  - for example $x_i =$ term counts or $x_i =$ IDF-weighted term frequencies.
- Each document is a non-negative vector in an n_x-space, where $n_x =$ vocabulary size.
  - that is, documents are rays, and similar documents have similar vectors.
- Can measure similarity between documents i and j by the cosine of the angle between x_i and x_j :
  - With perfectly collinear documents (that is, $x_i = \alpha x_j$, $\alpha > 0$), $\cos(0) = 1$
  - For orthogonal documents (no words in common), $\cos(\pi/2)=0$

Cosine similarity is computable as the normalized dot product between the vectors:

$$\text{cos\_sim}(x_1, x_2) = \frac{x_1 \cdot x_2}{||x_1|| ||x_2||}$$

```
from sklearn.metrics.pairwise import
cosine_similarity
# between two vectors:
sim = cosine_similarity(x, y)[0,0]
# between all rows of a matrix:
sims = cosine_similarity(X)
```

Notes on Cosine Similarity

- For a corpus with n rows, the pairwise similarities give $n \times (n-1)$ similarity scores.

Notes on Cosine Similarity

- For a corpus with n rows, the pairwise similarities give $n \times (n-1)$ similarity scores.
- tf-idf down-weights terms that appear in many documents, usually gives better results.
  - tf-idf similarity is the workhorse, used for example in elasticsearch.

Notes on Cosine Similarity

- For a corpus with n rows, the pairwise similarities give $n \times (n-1)$ similarity scores.
- tf-idf down-weights terms that appear in many documents, usually gives better results.
  - tf-idf similarity is the workhorse, used for example in elasticsearch.

Alternative distance metrics:

- dot product (sensitive to document length)
- Euclidean distance, $||v_1 - v_2||$
- Jensen-Shannon Divergence
- etc.

hopefully empirical results are not sensitive to choice of distance metric.

Burgess et al, "Legislative Influence Detectors"

- Compare bill texts across states in two-step process: (1) find candidates using elasticsearch (tf-idf similarlity); (2) compare candidates using text reuse score.

Burgess et al, “Legislative Influence Detectors”

- Compare bill texts across states in two-step process:
  (1) find candidates using elasticsearch (tf-idf similarlity);
  (2) compare candidates using text reuse score.

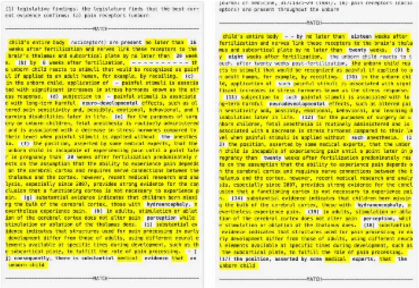


Figure 10: Match between Scott Walker’s bill and a highly similar bill from Louisiana. For a detailed view, please visit http://dssg.uchicago.edu/lid/.

Burgess et al, “Legislative Influence Detectors”

- Compare bill texts across states in two-step process:
  (1) find candidates using elasticsearch (tf-idf similarlity);
  (2) compare candidates using text reuse score.

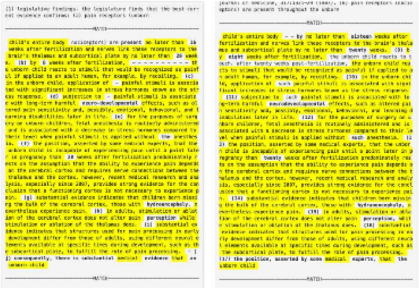


Figure 10: Match between Scott Walker's bill and a highly similar bill from Louisiana. For a detailed view, please visit http://dssg.uchicago.edu/lid/.

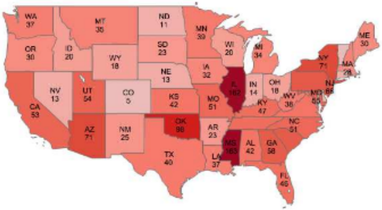


Figure 7: Introduced bills by state from ALEC model legislation

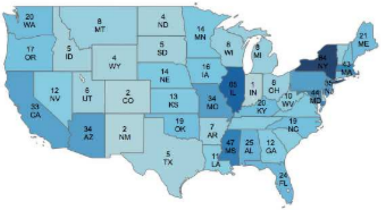


Figure 8: Introduced bills by state from ALICE model legislation

ABSTRACT

State legislatures introduce at least 45,000 bills each year. However, we lack a clear understanding of who is actually writing those bills. As legislators often lack the time and staff to draft each bill, they frequently copy text written by other states or interest groups.

However, existing approaches to detect text reuse are slow, biased, and incomplete. Journalists or researchers who want to know where a particular bill originated must perform a largely manual search. Watchdog organizations even hire armies of volunteers to monitor legislation for matches. Given the time-consuming nature of the analysis, journalists and researchers tend to limit their analysis to a subset of topics (e.g. abortion or gun control) or a few interest groups.

This paper presents the Legislative Influence Detector (LID). LID uses the Smith-Waterman local alignment algorithm to detect sequences of text that occur in model legislation and state bills. As it is computationally too expensive to run this algorithm on a large corpus of data, we use a search engine built using Elasticsearch to limit the number of comparisons. We show how LID has found 45,405 instances of bill-to-bill text reuse and 14,137 instances of model-legislation-to-bill text reuse. LID reduces the time it takes to manually find text reuse from days to seconds.

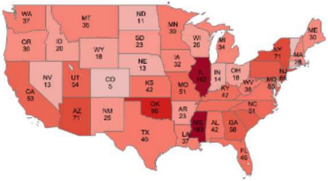


Figure 7: Introduced bills by state from ALEC model legislation

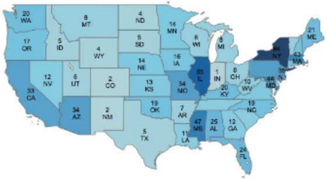


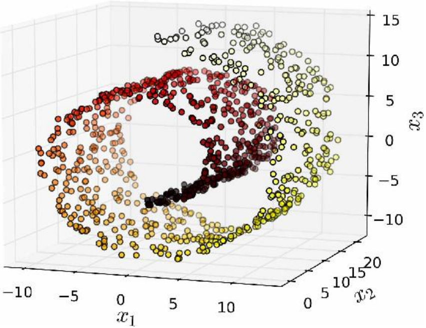
Figure 8: Introduced bills by state from ALICE model legislation

1. What is the research question?
2. Why is it important?
3. What is the problem solved?
4. What is being measured?
5. How does the measurement help answer the research question?

Outline

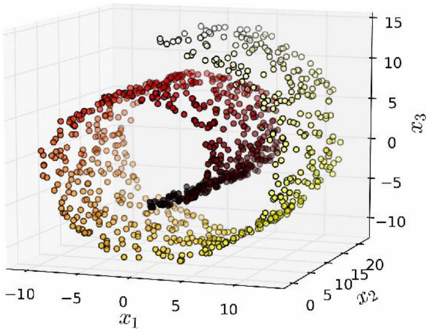
- Datasets are not distributed uniformly across the feature space.
- They have a lower-dimensional latent structure – a **manifold** – that can be learned.

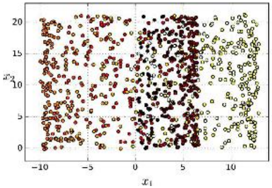
"The Swiss Roll"

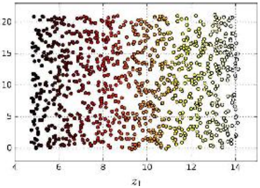


- Datasets are not distributed uniformly across the feature space.
- They have a lower-dimensional latent structure – a **manifold** – that can be learned.

“The Swiss Roll”





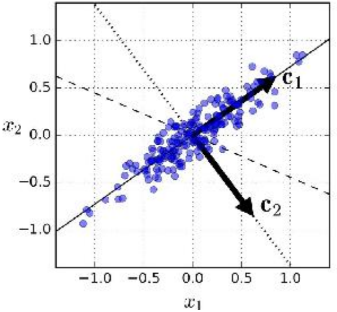


- **Dimensonality reduction** makes data more interpretable – for example by projecting down to two dimensions for visualization.
- improves computational tractability.
- can improve model performance.

What dimension reductions have we already tried?

PCA (principal component analysis) / SVD (singular value decomposition)

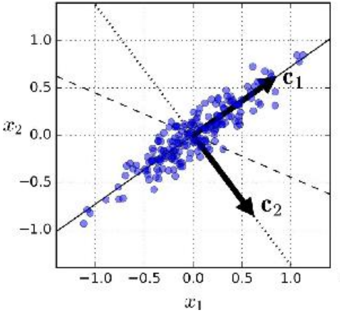
PCA (principal component analysis) / SVD (singular value decomposition)



- PCA computes the dimension in data explaining most variance.

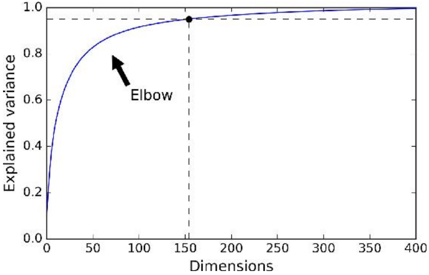
```
from sklearn.decomposition import PCA
pca = PCA(n_components=10)
X_train_pca = pca.fit_transform(X_train)
```

PCA (principal component analysis) / SVD (singular value decomposition)



- PCA computes the dimension in data explaining most variance.

```
from sklearn.decomposition import PCA
pca = PCA(n_components=10)
X_train_pca = pca.fit_transform(X_train)
```



- after the first component, subsequent components learn the (orthogonal) dimensions explaining most variance in dataset after projecting out first component.

PCA/NMF for Dimension Reduction

Data can be reduced by projecting down to first principal component dimensions.

- Distance metrics between observations (e.g. cosine similarity) are approximately preserved.

PCA/NMF for Dimension Reduction

Data can be reduced by projecting down to first principal component dimensions.

- Distance metrics between observations (e.g. cosine similarity) are approximately preserved.
- For supervised learning, reduced matrix be used as predictors instead of the original matrix.
  - but might destroy (a lot of) predictive information in your dataset.
  - compromise: use feature selection to keep strong predictors, and take principal components of weak predictors.

PCA/NMF for Dimension Reduction

Data can be reduced by projecting down to first principal component dimensions.

- Distance metrics between observations (e.g. cosine similarity) are approximately preserved.
- For supervised learning, reduced matrix be used as predictors instead of the original matrix.
  - but might destroy (a lot of) predictive information in your dataset.
  - compromise: use feature selection to keep strong predictors, and take principal components of weak predictors.
- PCA dimensions are not interpretable.
  - For non-negative data (e.g. counts or frequencies), **Non-negative Matrix Factorization (NMF)** provides more interpretable factors than PCA.

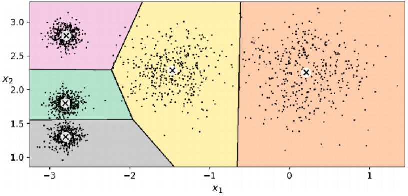
k-means clustering separates observations into k groups

k-means clustering separates observations into k groups

- Matrix of predictors treated as a Euclidean space (should standardize all columns)
- algorithm: initialize cluster centroids randomly, then shift around to minimize sum of within-cluster squared distance

k-means clustering separates observations into k groups

- Matrix of predictors treated as a Euclidean space (should standardize all columns)
- algorithm: initialize cluster centroids randomly, then shift around to minimize sum of within-cluster squared distance

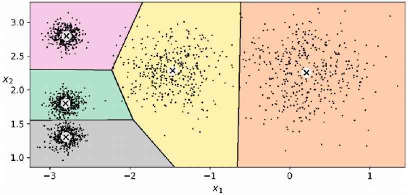


K-Means decision boundaries (Voronoi tessellation)

```
from sklearn.cluster import Kmeans
kmeans = KMeans(n_clusters=10)
kmeans.fit(X)
assigned_cluster = kmeans.labels_
```

k-means clustering separates observations into k groups

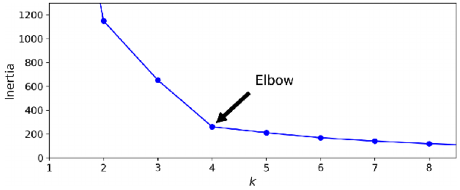
- Matrix of predictors treated as a Euclidean space (should standardize all columns)
- algorithm: initialize cluster centroids randomly, then shift around to minimize sum of within-cluster squared distance



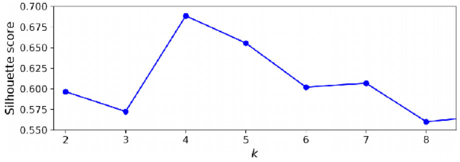
K-Means decision boundaries (Voronoi tessellation)

```
from sklearn.cluster import Kmeans
kmeans = KMeans(n_clusters=10)
kmeans.fit(X)
assigned_cluster = kmeans.labels_
```

k (number of clusters) is the only hyperparameter, can select using:



Selecting the number of clusters k using the "elbow rule"



Selecting the number of clusters k using the silhouette score

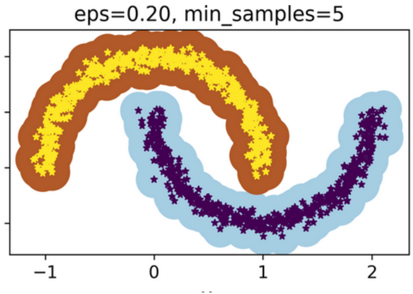
Other clustering algorithms

- “k-medoid” clustering use L1 distance rather than Euclidean distance; produces the “medoid” (median vector) for each cluster rather than “centroid” (mean vector).
  - less sensitive to outliers, and medoid can be used as representative data point.

Other clustering algorithms

- "k-medoid" clustering use L1 distance rather than Euclidean distance; produces the "medoid" (median vector) for each cluster rather than "centroid" (mean vector).
    - less sensitive to outliers, and medoid can be used as representative data point.

- DBSCAN defines clusters as continous regions of high density.
    - detects and excludes outliers automatically



- Agglomerative (hierarchical) clustering makes nested clusters.

Applications

Ganglmair and Wardlaw, "Complexity, Standardization, and the Design of Loan Agreements"

- use k-medoid clustering to identify different types of debt contracts, and analyze customization.
- used for descriptive analysis $\rightarrow$ e.g., that larger deals have more customization.

Applications

Ganglmair and Wardlaw, "Complexity, Standardization, and the Design of Loan Agreements"

- use k-medoid clustering to identify different types of debt contracts, and analyze customization.
- used for descriptive analysis $\rightarrow$ e.g., that larger deals have more customization.

Hoberg and Phillips, "Text-Based Network Industries and Endogenous Product Differentiation"

- "business description" section from annual regulatory filings, preprocessed by extracting nouns, drop words appearing in more than 25% of documents.
- vector represetnation: binary for whether word appears (rather than counts)
- clusters of these vectors are "industries" – sets of firms with similar lists of nouns in their business descriptions.

Outline

Topic Models in Social Science

- Core methods for topic models were developed in computer science and statistics
  - summarize unstructured text
  - use words within document to infer subject
  - useful for dimension reduction

Topic Models in Social Science

- Core methods for topic models were developed in computer science and statistics
  - summarize unstructured text
  - use words within document to infer subject
  - useful for dimension reduction
- Social scientists use topics as a form of measurement
  - how observed covariates drive trends in language
  - tell a story not just about what, but how and why

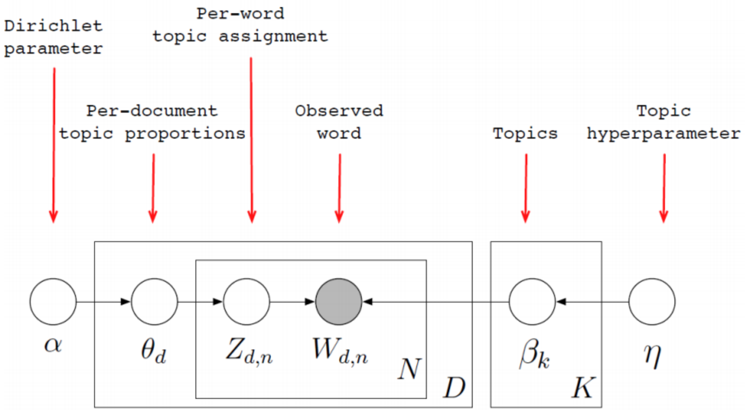
Topic Models in Social Science

- Core methods for topic models were developed in computer science and statistics
  - summarize unstructured text
  - use words within document to infer subject
  - useful for dimension reduction
- Social scientists use topics as a form of measurement
  - how observed covariates drive trends in language
  - tell a story not just about what, but how and why
  - **topic models are more interpretable** than other dimension reduction methods, such as PCA.

- Latent Dirichlet Allocation (LDA):
  - Each topic is a distribution over words.
  - Each document is a distribution over topics.

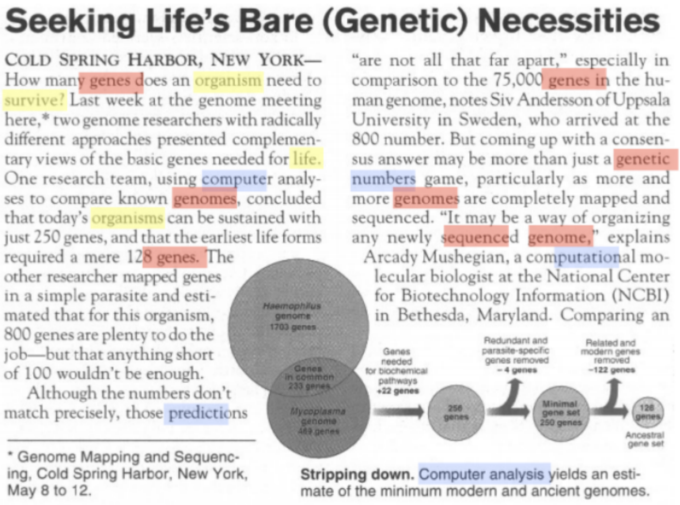
- Latent Dirichlet Allocation (LDA):
  - Each topic is a distribution over words.
  - Each document is a distribution over topics.
- Input: $N \times M$ document-term count matrix X
- Assume: there are K topics (tunable hyperparameter, use coherence).
- Like PCA or NMF, LDA works by factorizing X into:
  - an $N \times K$ document-topic matrix
  - an $K \times M$ topic-term matrix.

- Latent Dirichlet Allocation (LDA):
  - Each topic is a distribution over words.
  - Each document is a distribution over topics.
- Input: $N \times M$ document-term count matrix X
- Assume: there are K topics (tunable hyperparameter, use coherence).
- Like PCA or NMF, LDA works by factorizing X into:
  - an $N \times K$ document-topic matrix
  - an $K \times M$ topic-term matrix.



Variational inference setup (Brandon Stewart slides).

A statistical highlighter



Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

Image from Hanna Wallach

Using an LDA Model

Once trained, can easily get topic proportions for a corpus.

- for any document – doesn't have to be in training corpus.
- main topic is the highest-probability topic

Using an LDA Model

Once trained, can easily get topic proportions for a corpus.

- for any document – doesn't have to be in training corpus.
- main topic is the highest-probability topic
- documents with highest share in a topic can work as representative documents for the topic.

Using an LDA Model

Once trained, can easily get topic proportions for a corpus.

- for any document – doesn't have to be in training corpus.
- main topic is the highest-probability topic
- documents with highest share in a topic can work as representative documents for the topic.

Can then use the topic proportions as variables in a social science analysis.

- e.g., Catalinac (2016) shows that after a Japanese political reform that reduced intraparty competition, candidate platforms reduced local pork and increased national policy.

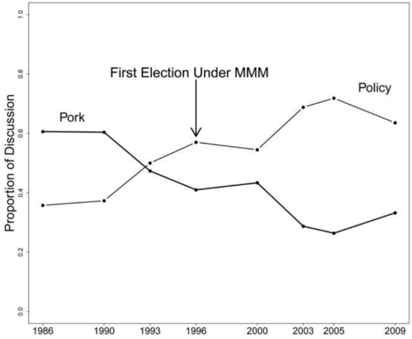


Table 1 A Summary of Common Assumptions and Relative Costs Across Different Methods of Discrete Text Categorization

	Method				
A. Assumptions	*Reading*	*Human Coding*	*Dictionaries*	*Supervised Learning*	*Topic Model*
Categories are known	No	Yes	Yes	Yes	No
Category nesting, if any, is known	No	Yes	Yes	Yes	No
Relevant text features are known	No	No	Yes	Yes	Yes
Mapping is known	No	No	Yes	No	No
Coding can be automated	No	No	Yes	Yes	Yes
B. Costs					
Preanalysis Costs					
Person-hours spent conceptualizing	Low	High	High	High	Low
Level of substantive knowledge	Moderate/High	High	High	High	Low
Analysis Costs					
Person hours spent per text	High	High	Low	Low	Low
Level of substantive knowledge	Moderate/High	Moderate	Low	Low	Low
Postanalysis Costs					
Person-hours spent interpreting	High	Low	Low	Low	Moderate
Level of substantive knowledge	High	High	High	High	High

Recommended: read this part of Quinn, Monroe, Colaresi, Crespin, and Radev (2010).

Topic modeling Federal Reserve Bank transcripts

Hansen, McMahon, and Prat (QJE 2017)

- Analyze speech transcripts from FOMC (Federal Open Market Committee).
  - private discussions among committee members at Federal Reserve (U.S. Central Bank)
  - 150 meetings, 20 years, 26,000 speeches, 24,000 unique words.

Topic modeling Federal Reserve Bank transcripts

Hansen, McMahon, and Prat (QJE 2017)

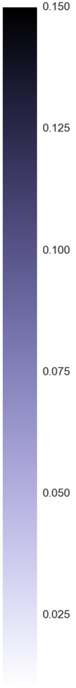
- Analyze speech transcripts from FOMC (Federal Open Market Committee).
  - private discussions among committee members at Federal Reserve (U.S. Central Bank)
  - 150 meetings, 20 years, 26,000 speeches, 24,000 unique words.
- Pre-processing:
  - drop stopwords, stems; vocab = 10,000 words

Topic modeling Federal Reserve Bank transcripts

Hansen, McMahon, and Prat (QJE 2017)

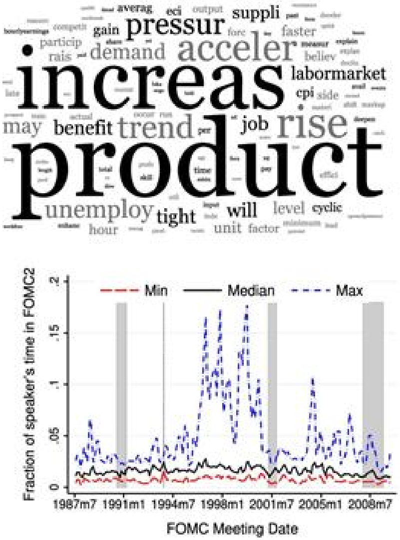
- Analyze speech transcripts from FOMC (Federal Open Market Committee).
  - private discussions among committee members at Federal Reserve (U.S. Central Bank)
  - 150 meetings, 20 years, 26,000 speeches, 24,000 unique words.
- Pre-processing:
  - drop stopwords, stems; vocab $=$ 10,000 words
- LDA:
  - $K = 40$ topics selected for interpretability / topic coherence.

													Pro-cyclicality
Topic0[1]	product	increas	wage	price	cost	labor	rise	acceler	inflat	pressur	trend	compens	0.024
Topic1[1,2]	growth	slow	economi	continu	expans	strong	trend	inflat	will	recent	slowdown	moder	0.023
Topic2[2]	inflat	expect	core	measur	higher	path	slack	gradual	continu	remain	view	suggest	0.017
Topic3[1]	percent	year	quarter	growth	month	rate	last	next	state	averag	california	employ	0.007
Topic4	number	data	look	chang	measur	use	point	show	revis	estim	gdp	actual	0.007
Topic5[1,2]	polici	inflat	monetarpol	need	time	can	monetari	move	tighten	view	action	believ	0.005
Topic6[2]	rate	term	expect	real	lower	increas	rise	level	declin	short	nomin	year	0.005
Topic7	statement	word	chang	meet	languag	discuss	issu	want	read	sentenc	view	use	0.005
Topic8[2]	chairman	support	mr	direct	recommend	agre	asymmetr	prefer	symmetr	move	toward	favor	0.004
Topic9[1]	employ	continu	growth	job	nation	region	seem	state	manufactur	greenbook	busi	bit	0.004
Topic10	dollar	unitedstates	export	countri	import	foreign	japan	growth	abroad	trade	develop	currenc	0.003
Topic11	model	use	simul	shock	effect	scenario	nairu	differ	rule	chang	baselin	altern	0.003
Topic12[2]	risk	may	balanc	seem	side	uncertainti	possibl	economi	probabl	reason	upsid	much	0.003
Topic13	forecast	greenbook	staff	project	differ	assumpt	littl	assum	somewhat	lower	end	period	0.002
Topic14	period	committe	consist	econom	run	maintain	futur	read	slightli	stabil	expect	develop	0.002
Topic15	invest	incom	spend	capit	household	consum	busi	hous	consumpt	sector	stock	stockmarket	0.002
Topic16[1]	month	report	increas	survey	expect	indic	remain	continu	last	recent	data	activ	0.002
Topic17[1]	project	forecast	year	quarter	expect	will	percent	revis	anticip	growth	next	recent	0.002
Topic18	question	ask	issu	let	want	answer	rais	discuss	don	start	without	okay	0.001
Topic19	peopl	talk	lot	much	comment	around	differ	number	realli	look	thing	hear	0.001
Topic20	presid	ye	governor	parri	stern	vice	hoenig	minehan	kelley	jordan	moskow	mcteer	0.001
Topic21	move	can	evid	signific	stage	inde	will	issu	economi	may	quit	clearli	0.001
Topic22[2]	chairman	thank	mr	time	meet	laughter	comment	let	will	point	call	may	0.0
Topic23[1]	year	panel	line	shown	right	chart	expect	project	percent	middl	left	next	0.0
Topic24	district	nation	area	continu	sector	construct	manufactur	report	activ	region	economi	remain	0.0
Topic25	know	someth	happen	right	thing	want	look	sure	can	realli	anyth	els	0.0
Topic26[1,2]	polici	might	committe	market	may	tighten	eas	risk	action	staff	possibl	potenti	-0.001
Topic27	year	continu	product	price	level	industri	will	sale	increas	auto	last	district	-0.001
Topic28[1]	inventori	product	sale	level	order	will	sector	come	good	quarter	much	adjust	-0.001
Topic29	price	oil	increas	energi	effect	import	suppli	product	demand	will	market	oilprices	-0.002
Topic30	term	might	point	can	sens	run	short	probabl	time	longer	tri	someth	-0.002
Topic31	seem	may	time	certainli	bit	littl	quit	much	far	perhap	better	might	-0.003
Topic32	money	aggreg	borrow	seem	rang	reserv	rate	target	time	altern	suggest	million	-0.003
Topic33[2]	move	market	point	will	fundsrate	rate	basispoints	need	fed	today	basi	time	-0.004
Topic34[1]	report	busi	compani	year	contact	firm	sale	worker	expect	plan	director	industri	-0.004
Topic35	will	fiscal	ta	budget	cut	govern	effect	billion	state	spend	deficit	year	-0.005
Topic36	will	economi	world	rather	problem	believ	can	situat	much	seem	view	good	-0.008
Topic37	realli	look	side	thing	lot	problem	concern	littl	pretti	situat	kind	much	-0.012
Topic38	bank	credit	market	loan	financi	debt	lend	fund	concern	financ	problem	spread	-0.018
Topic39[1,2]	economi	weak	recoveri	recess	confid	eas	neg	econom	will	turn	declin	period	-0.059

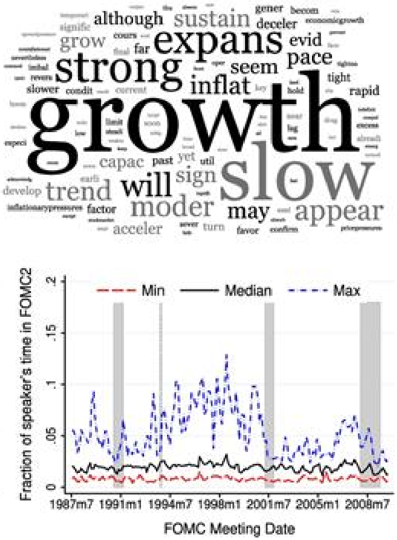


Pro-Cyclical Topics

Hansen, McMahon, and Prat (QJE 2017)



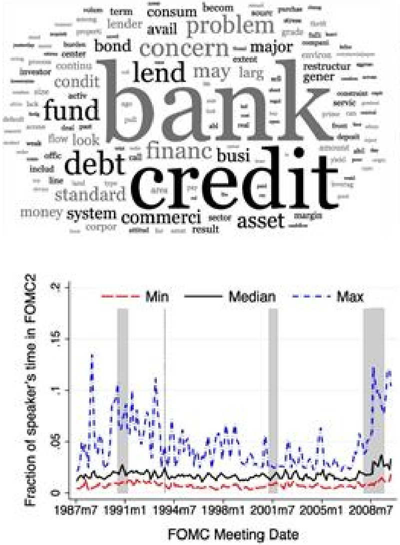
(A) Topic 0 'Productivity'



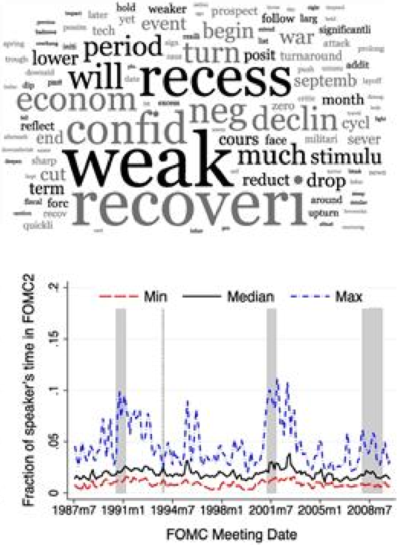
(B) Topic 1 'Growth'

Counter-Cyclical Topics

Hansen, McMahon, and Prat (QJE 2017)



(A) Topic 38 'Financial sector'



(B) Topic 39 'Economic weakness'

Effect of Transparency

Hansen, McMahon, and Prat (QJE 2017)

- In 1993, there was an unexpected transparency shock where transcripts became public.

Effect of Transparency

Hansen, McMahon, and Prat (QJE 2017)

- In 1993, there was an unexpected transparency shock where transcripts became public.
- Increasing transparency results in:
  - higher discipline / technocratic language (probably beneficial)
  - higher conformity (probably costly)
- Highlights tradeoffs from transparency in bureaucratic organizations.

Structural Topic Model = LDA + Metadata

Roberts, Stewart, and Tingley

STM provides two ways to include contextual information:

- Topic prevalence can vary by metadata
  - e.g. Republicans talk about military issues more then Democrats

Structural Topic Model = LDA + Metadata

Roberts, Stewart, and Tingley

STM provides two ways to include contextual information:

- Topic prevalence can vary by metadata
  - e.g. Republicans talk about military issues more then Democrats
- Topic content can vary by metadata
  - e.g. Republicans talk about military issues more patriotically than Democrats.

Structural Topic Model = LDA + Metadata

Roberts, Stewart, and Tingley

STM provides two ways to include contextual information:

- Topic prevalence can vary by metadata
  - e.g. Republicans talk about military issues more then Democrats
- Topic content can vary by metadata
  - e.g. Republicans talk about military issues more patriotically than Democrats.

- Structural topic model is not a prediction model:
  - it will tell you which topics or features correlate with an outcome, but it will not provide an in-sample or out-of-sample prediction for an outcome

Structural Topic Model = LDA + Metadata

Roberts, Stewart, and Tingley

STM provides two ways to include contextual information:

- Topic prevalence can vary by metadata
  - e.g. Republicans talk about military issues more then Democrats
- Topic content can vary by metadata
  - e.g. Republicans talk about military issues more patriotically than Democrats.

- Structural topic model is not a prediction model:
  - it will tell you which topics or features correlate with an outcome, but it will not provide an in-sample or out-of-sample prediction for an outcome
- The main implementation is in R. gensim has a light-weight version called "author topic model" (see this week's notebook).

Text-Based Ideal Points

Vafa, Naidu, and Blei

Vote-based ideal points (from political science):

- infer ideology dimension of politican i based on **vote** differences across **bills** j.

Text-Based Ideal Points

Vafa, Naidu, and Blei

Vote-based ideal points (from political science):

- infer ideology dimension of politican i based on **vote** differences across **bills** j.
- vote $v_{ij} \in \{0, 1\}$ modeled as

$$v_{ij} = \text{Bernoulli}(\text{sigmoid}(\beta_j + x_i \eta_j))$$

  - β_j is bill effect
  - x_i **= ideal point**
  - η_j is the bill's party polarity

Text-Based Ideal Points

Vafa, Naidu, and Blei

Vote-based ideal points (from political science):

- infer ideology dimension of politican i based on **vote** differences across **bills** j.
- vote $v_{ij} \in \{0, 1\}$ modeled as

$$v_{ij} = \text{Bernoulli}(\text{sigmoid}(\beta_j + x_i \eta_j))$$

  - β_j is bill effect
  - x_i **= ideal point**
  - η_j is the bill's party polarity

Text-based ideal points:

- infer ideology dimension of politican i based on **word** differences across **topics** j.

Text-Based Ideal Points

Vafa, Naidu, and Blei

Vote-based ideal points (from political science):

- infer ideology dimension of politican i based on **vote** differences across **bills** j.
- vote $v_{ij} \in \{0,1\}$ modeled as

$$v_{ij} = \text{Bernoulli}(\text{sigmoid}(\beta_j + x_i \eta_j))$$

  - β_j is bill effect
  - x_i **= ideal point**
  - η_j is the bill's party polarity

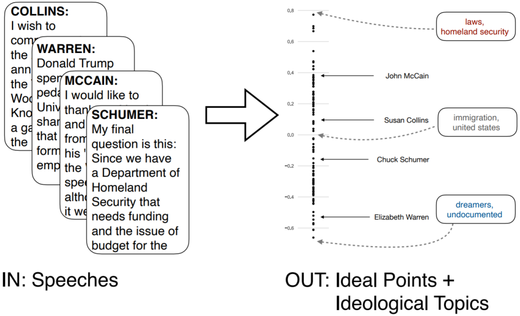
Text-based ideal points:

- infer ideology dimension of politican i based on **word** differences across **topics** j.
- word count $v_{iw} \in \{0,1,...\}$ for word w

$$v_{iw} = \text{Poisson}(\sum_k \theta_{ik}\beta_{kw}\exp(x_i\eta_{kw}))$$

  - k indexes topics
  - θ_{ik}, politician i's topic share.
  - β_{kw}, word w's imporance to topic
  - x_i **= ideal point**
  - η_{kw} word w's topic-specific polarity

Text-Based Ideal Points



Ideology	Top Words
Progressive	class, billionaire, billionaires, walmart, wall street, corporate, executives, government
Neutral	economy, pay, trump, business, tax, corporations, americans, billion
Moderate	trade war, trump, jobs, farmers, economy, economic, tariffs, businesses, promises, job
Progressive	#medicareforall, insurance companies, profit, health care, earth, medical debt, health care system, profits
Neutral	health care, plan, medicare, americans, care, access, housing, millions
Moderate	healthcare, universal healthcare, public option, plan, universal coverage, universal health care, away, choice
Progressive	green new deal, fossil fuel industry, fossil fuel, planet, pass, #greennewdeal, climate crisis, middle ground
Neutral	climate change, climate, climate crisis, plan, planet, crisis, challenges, world
Moderate	solutions, technology, carbon tax, climate change, challenges, climate, negative, durable

Table 3. The TBIP learns topics from 2020 Democratic presidential candidate tweets that vary as a function of the candidate's political positions. The neutral topics are for an ideal point of 0; the ideological topics fix ideal points at -1 and $+1$. We interpret one extreme as progressive and the other as moderate.

- Vafa et al show that text-based ideal points are correlated with vote-based ideal points.

Outline

Zoom Poll: Correlation $\neq$ Causation

Which of these research designs is not like the others?

1. Ganglmair-Wardlaw:
   - k-medoids on debt contracts
   - larger deals increase contract customization.
2. Catalinac:
   - LDA on Japanese party platforms
   - after reform, local pork decreases and national policy increases.
3. Hansen-McMahon-Prat:
   - LDA on Central Bank transcripts
   - productivity/growth topics increase economic growth.

Causal inference is needed to improve the world

Consider important policy questions like:

- In light of coronavirus, should schools reopen or not for in-person teaching?

Causal inference is needed to improve the world

Consider important policy questions like:

- In light of coronavirus, should schools reopen or not for in-person teaching?
  - No matter how much we know from lab experiments about the biology/epidemiology of the virus, there will be too much uncertainty about costs/benefits to answer this.
  - We need real-world evidence, but we can't experimentally force schools to reopen or not.

Causal inference is needed to improve the world

Consider important policy questions like:

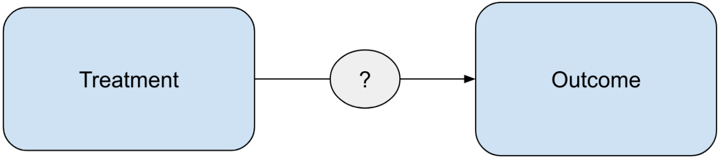
- In light of coronavirus, should schools reopen or not for in-person teaching?
  - No matter how much we know from lab experiments about the biology/epidemiology of the virus, there will be too much uncertainty about costs/benefits to answer this.
  - We need real-world evidence, but we can't experimentally force schools to reopen or not.
- Can use a **natural experiment** to produce causal estimates:
  - e.g., variation in number of coronavirus cases before/after openings, using differences in the timing of openings (differences-in-differences).

Causal inference is needed to improve the world

Consider important policy questions like:

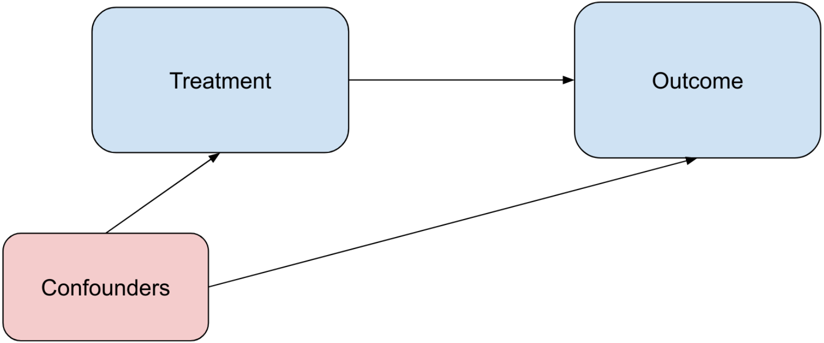
- In light of coronavirus, should schools reopen or not for in-person teaching?
  - No matter how much we know from lab experiments about the biology/epidemiology of the virus, there will be too much uncertainty about costs/benefits to answer this.
  - We need real-world evidence, but we can't experimentally force schools to reopen or not.
- Can use a **natural experiment** to produce causal estimates:
  - e.g., variation in number of coronavirus cases before/after openings, using differences in the timing of openings (differences-in-differences).
- Google/Facebook understand the importance of causal inference with A/B testing; social scientists want to use it to assist public policy.

Causal Graphs



- We are interested in estimating a causal effect (if any) of a "treatment" on an "outcome".

- ***Unobserved Confounders*** are variables that affect both the treatment and the outcome, which we don't have in our dataset:



- **Observed confounders** are not a problem, because we can adjust (control) for them in causal inference analysis (that is, including them in a regression).

- **Reverse causation**: "the outcome" affects "the "treatment".
  Joint causation: there is bidirectional causation.

- e.g., effect of tax collections on economic growth.

- **Reverse causation**: “the outcome” affects “the “treatment”.
  Joint causation: there is bidirectional causation.

- e.g., effect of tax collections on economic growth.
- Resulting estimates are biased (not causal), and cannot be fixed by adjusting for observed confounders.

With joint causality, or with unobserved confounders, it is often impossible to produce statistical estimates with a causal interpretation.

With joint causality, or with unobserved confounders, it is often impossible to produce statistical estimates with a causal interpretation.

- The gold standard: randomized control trials.
  - often not available, e.g. with opening/closing schools under covid-19.

With joint causality, or with unobserved confounders, it is often impossible to produce statistical estimates with a causal interpretation.

- The gold standard: randomized control trials.
  - often not available, e.g. with opening/closing schools under covid-19.
- Second best: natural experiments.
  - differences-in-differences: use longitudinal data and look at groups or places that adopted treatment at different times.

With joint causality, or with unobserved confounders, it is often impossible to produce statistical estimates with a causal interpretation.

- The gold standard: randomized control trials.
  - often not available, e.g. with opening/closing schools under covid-19.
- Second best: natural experiments.
  - differences-in-differences: use longitudinal data and look at groups or places that adopted treatment at different times.
  - regression discontinuity: compare individuals just above or just below some discrete scoring threshold.

With joint causality, or with unobserved confounders, it is often impossible to produce statistical estimates with a causal interpretation.

- The gold standard: randomized control trials.
  - often not available, e.g. with opening/closing schools under covid-19.
- Second best: natural experiments.
  - differences-in-differences: use longitudinal data and look at groups or places that adopted treatment at different times.
  - regression discontinuity: compare individuals just above or just below some discrete scoring threshold.
  - instrumental variables: use a third variable ("instrument") that randomly shifts the probability of treatment.

Fong and Grimmer (2016): Causal effect of political messaging

- What biographical characteristics of politicians influence voter evaluations?

Fong and Grimmer (2016): Causal effect of political messaging

- What biographical characteristics of politicians influence voter evaluations?
- Could run a survey experiment:
  - `Document 1: He earned his Juris Doctor in 1997 from Yale Law School, where he operated free legal clinics for low-income residents of New Haven, Connecticut...`
  - `Document 2: He served in South Vietnam from 1970 to 1971 during the Vietnam War in the Army Rangers' 75th Ranger Regiment, attached to the 173rd Airborne Brigade. He participated in 24 helicopter assaults...`

Fong and Grimmer (2016): Causal effect of political messaging

- What biographical characteristics of politicians influence voter evaluations?
- Could run a survey experiment:
  - `Document 1: He earned his Juris Doctor in 1997 from Yale Law School, where he operated free legal clinics for low-income residents of New Haven, Connecticut...`
  - `Document 2: He served in South Vietnam from 1970 to 1971 during the Vietnam War in the Army Rangers' 75th Ranger Regiment, attached to the 173rd Airborne Brigade. He participated in 24 helicopter assaults...`
- But hard to generalize what features drive differences.

Fong and Grimmer (2016): Approach

- Lab experiment: 1,886 participants, 5,303 responses

1. Randomly assign texts, X_i , to respondents i
   - Sees up to 3 texts from the corpus of $>$ 2200 Wikipedia biographies
2. Obtain responses Y_i for each respondent
   - Feeling thermometer rating: 0-100

Fong and Grimmer (2016): Approach

- Lab experiment: 1,886 participants, 5,303 responses

1. Randomly assign texts, X_i , to respondents i
   - Sees up to 3 texts from the corpus of $>$ 2200 Wikipedia biographies
2. Obtain responses Y_i for each respondent
   - Feeling thermometer rating: 0-100
3. Structural topic model variant ("supervised indian buffet process"):
   - Discover mapping from texts X to latent topic treatments $\vec{D}$ based on their effect on Y.

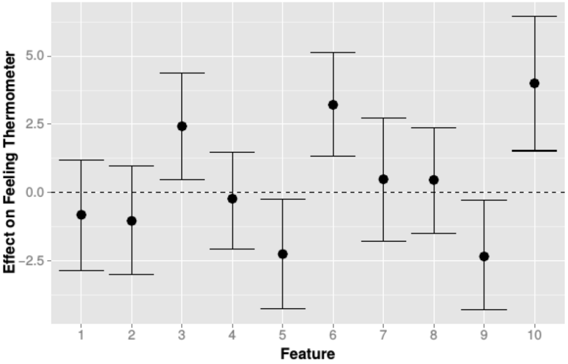
Fong and Grimmer (2016): Approach

- Lab experiment: 1,886 participants, 5,303 responses

1. Randomly assign texts, X_i , to respondents i
   - Sees up to 3 texts from the corpus of $>$ 2200 Wikipedia biographies
2. Obtain responses Y_i for each respondent
   - Feeling thermometer rating: 0-100
3. Structural topic model variant ("supervised indian buffet process"):
   - Discover mapping from texts X to latent topic treatments $\vec{D}$ based on their effect on Y.
4. Measure causal effects of these treatments on Y_i

Fong and Grimmer (2016): Results

Treatment	Keywords
3	director, university, received, president, phd, policy
5	elected, house, democratic, seat
6	united_states, military, combat, rank
9	law, school_law, law_school, juris_doctor, student
10	war, enlisted, united_states, assigned, army



Outline

Document Distance

Dimensionality Reduction

Topic Models

Social Science Research with Text

Wrapping Up

First Response Essay due in two weeks

- Critically read and review an application paper.
- 300 words is the minimum for a passing grade (3+) but 500+ words would be expected for high grade (7+).
- Anonymize your submission – do not include your name anywhere in the document. Submit as TXT or PDF to EduFlow.

What can I write about?

- Any “Applications” reading from weeks 1 through 4 (see reading list).

What can I write about?

- Any "Applications" reading from weeks 1 through 4 (see reading list).
- Can read/respond to an off-syllabus paper if it applies tools from one of the first four lectures.
  - dictionary methods, text complexity, n-grams, document similarity, topic models, machine learning.
  - Please confirm off-syllabus readings with me by email.

What can I write about?

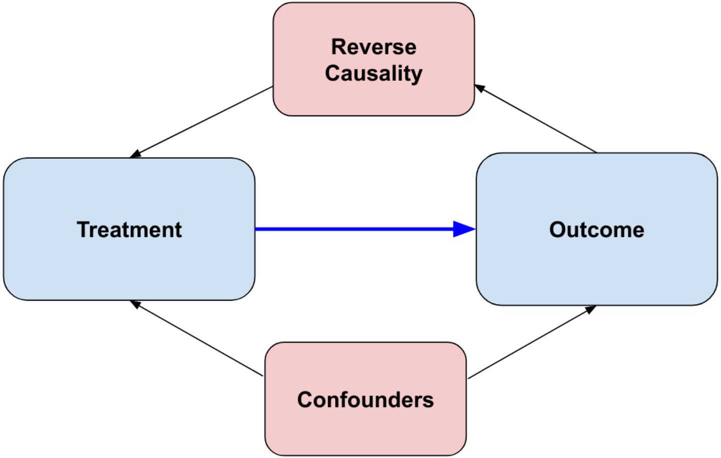
- Any "Applications" reading from weeks 1 through 4 (see reading list).
- Can read/respond to an off-syllabus paper if it applies tools from one of the first four lectures.
  - dictionary methods, text complexity, n-grams, document similarity, topic models, machine learning.
  - Please confirm off-syllabus readings with me by email.
- Note:
  - for conference-style articles 14 pages or shorter, you have to do two (shorter) response essays.
  - if a paper has an appendix, you are responsible for reading it!
  - can also compare/contrast two readings.

What to think about

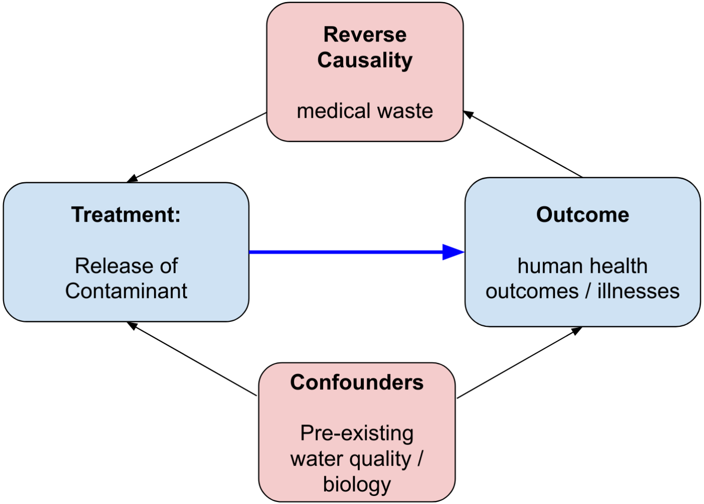
See last page of homework assignments document.

- Example essays from previous years are available from homework page.
- We will practice criticizing the required reading next week.

Causal Graphs



Causal Graph Example: Pollution of a River



Activity: Practice with Causal Graphs

- Think of two example causal inference questions:
  1. where you have **language as an outcome**
  2. where you have **language as a treatment**
- Try to personalize it:
  - a research question from your field
  - a policy you are interested in
  - a mystery you are fascinated by

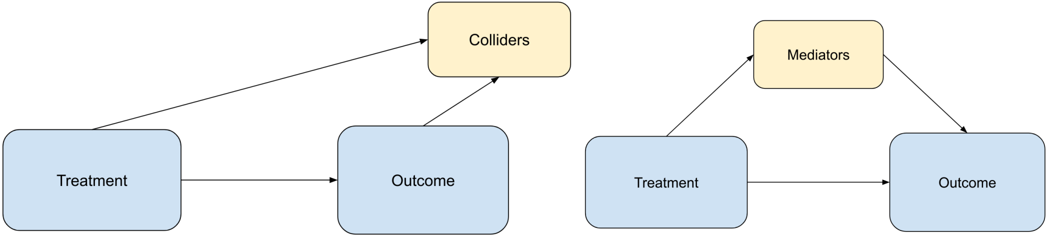
Activity: Practice with Causal Graphs

- Think of two example causal inference questions:
  1. where you have **language as an outcome**
  2. where you have **language as a treatment**
- Try to personalize it:
  - a research question from your field
  - a policy you are interested in
  - a mystery you are fascinated by
- Link to causal graph template posted in zoom chat:
  - make a copy, fill it in
  - make your doc viewable and paste link into padlet (also in zoom chat).
  - will review these at beginning of next lecture.

Extra Slides

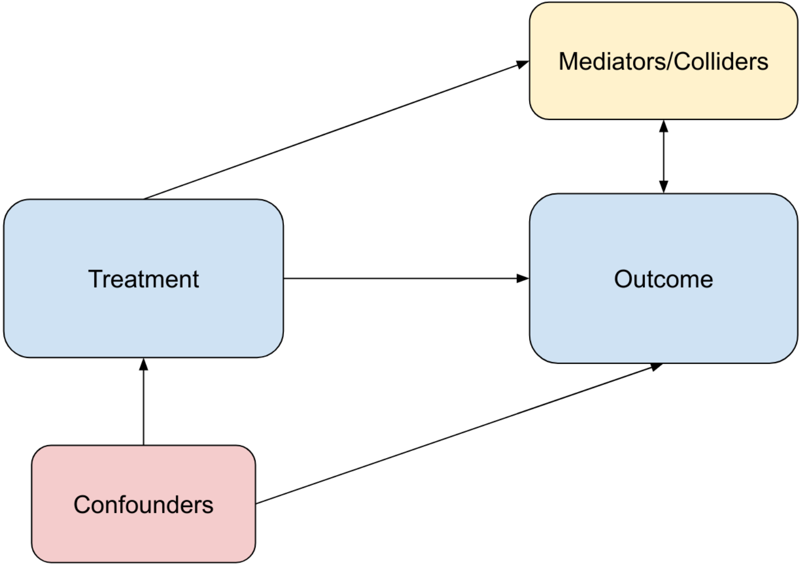
Colliders and Mediators

- **Colliders** are affected by both the treatment and the outcome.
  Mediators are intermediate outcomes / mechanisms.



- Adjusting for colliders or mediators will add bias.

Causal Graphs: Overview



Causal Graph Example: Pollution of a River