Sequencing Legal DNA

NLP for Law and Political Economy

3. Document Distance and Topic Models

Different Goals, Different Methods

- Supervised: Pursuing a known goal.
  - e.g., predicting whether a defendant will win his case.
  - machine learns to replicate human decision process.

Different Goals, Different Methods

- Supervised: Pursuing a known goal.
  - e.g., predicting whether a defendant will win his case.
  - machine learns to replicate human decision process.
- Unsupervised:
  - algorithm discovers themes/patterns in data (e.g. text)
    - e.g., k-means clustering of similar documents.
  - human interprets the results (e.g. clusters)

Different Goals, Different Methods

- Supervised: Pursuing a known goal.
  - e.g., predicting whether a defendant will win his case.
  - machine learns to replicate human decision process.
- Unsupervised:
  - algorithm discovers themes/patterns in data (e.g. text)
    - e.g., k-means clustering of similar documents.
  - human interprets the results (e.g. clusters)
- Both strategies amplify human effort, each in different ways.
- Also: supervised learning models can be used to discover themes/patterns, and unsupervised learning models can be used in service of prediction or known goals.

Different Goals, Different Methods

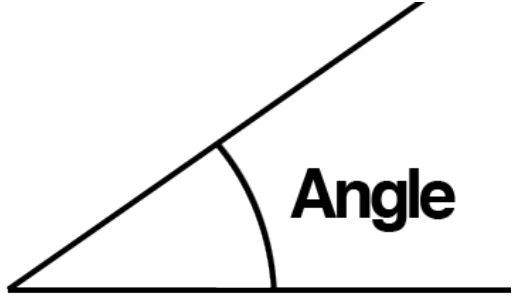
- Supervised: Pursuing a known goal.
  - e.g., predicting whether a defendant will win his case.
  - machine learns to replicate human decision process.
- Unsupervised:
  - algorithm discovers themes/patterns in data (e.g. text)
    - e.g., k-means clustering of similar documents.
  - human interprets the results (e.g. clusters)
- Both strategies amplify human effort, each in different ways.
- Also: supervised learning models can be used to discover themes/patterns, and unsupervised learning models can be used in service of prediction or known goals.

- Today, focus on **unsupervised learning** using **distance metrics** and **topic models**.

Outline

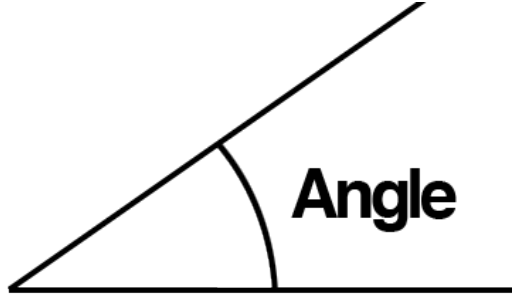
Outline

Cosine Similarity: Idea



- each document is a non-negative vector in an n-space (size of the common dictionary) and it defines a *ray*
  - closer rays form smaller angles
  - the furthest rays are orthogonal

Cosine Similarity: Idea



- each document is a non-negative vector in an *n*-space (size of the common dictionary) and it defines a *ray*
  - closer rays form smaller angles
  - the furthest rays are orthogonal
- $\cos(0) = 1$ and $\cos(\pi/2)=0$
- distance monotonically increases on $\{0,\pi/2\}$ -> cosine or similarity monotonically decreases on $\{1,0\}$

Cosine similarity: Formula

$$\text{cos\_sim}(v_1, v_2) = \frac{v_1 \cdot v_2}{||v_1||||v_2||}$$

where v_1 and v_2 are vectors, representing documents (e.g., tf-idf weighted word counts).

- $+1$ means identical documents; 0 means no words in common.
- Note that for n rows, this gives you $n \times (n-1)$ similarity scores.

Cosine similarity: Formula

$$\text{cos\_sim}(v_1, v_2) = \frac{v_1 \cdot v_2}{||v_1||||v_2||}$$

where v_1 and v_2 are vectors, representing documents (e.g., tf-idf weighted word counts).

- $+1$ means identical documents; 0 means no words in common.
- Note that for n rows, this gives you $n \times (n-1)$ similarity scores.
- tf-idf similarities will down-weight terms that appear in many documents and could give better results.
- dimension-reduction with PCA can also help before computing similarities.

Other distance metrics

- Euclidean distance, $||v_1 - v_2||$
- Jensen-Shannon Divergence
- etc.
- hopefully empirical results are not sensitive to choice of metric.

Outline

Text analysis of patent innovation

"Measuring technological innovation over the very long run," Kelly, Papanikolau, Seru, and Taddy (2018)

- Data:
  - 9 million patents since 1840, from U.S. Patent Office and Google Scholar Patents.
  - date, inventor, backward citations
  - text (abstract, claims, and description)

Text analysis of patent innovation

"Measuring technological innovation over the very long run," Kelly, Papanikolau, Seru, and Taddy (2018)

- Data:
  - 9 million patents since 1840, from U.S. Patent Office and Google Scholar Patents.
  - date, inventor, backward citations
  - text (abstract, claims, and description)
- Text pre-processing:
  - drop HTML markup, punctuation, numbers, capitalization, and stopwords.
  - remove terms that appear in less than 20 patents.
  - 1.6 million words in vocabulary.

Measuring Innovation

- Backward IDF weighting of word w in patent p:

$$\text{BIDF}(w,p) = \frac{\#\text{ of patents prior to } p}{\log\left(1 + \#\text{ documents prior to } p \text{ that include } w\right)}$$

  - down-weights words that appeared frequently before a patent, but up-weights new words.

Measuring Innovation

- Backward IDF weighting of word w in patent p:

$$\text{BIDF}(w, p) = \frac{\#\text{ of patents prior to } p}{\log\left(1 + \#\text{ documents prior to } p \text{ that include } w\right)}$$

  - down-weights words that appeared frequently before a patent, but up-weights new words.
- For each patent:
  - compute cosine similarity to all future patents, using BIDF of earlier patent.

Measuring Innovation

- Backward IDF weighting of word w in patent p:

$$\text{BIDF}(w,p) = \frac{\#\text{ of patents prior to } p}{\log\left(1 + \#\text{ documents prior to } p \text{ that include } w\right)}$$

  - down-weights words that appeared frequently before a patent, but up-weights new words.
- For each patent:
  - compute cosine similarity to all future patents, using BIDF of earlier patent.
- 9m$\times$9m similarity matrix = 30TB of data.
  - enforce sparsity by setting similarity $< .05$ to zero (93.4% of pairs).

Novelty, Impact, and Quality

- "Novelty" is defined by (negative) similarity to previous patents:

$$\text{Novelty}_j = -\sum_{i \in B(j)} \rho_{ij}$$

where $B(j)$ is the set of previous patents (in, e.g., last 20 years).

Novelty, Impact, and Quality

- "Novelty" is defined by (negative) similarity to previous patents:

$$\text{Novelty}_j = -\sum_{i \in B(j)} \rho_{ij}$$

  where $B(j)$ is the set of previous patents (in, e.g., last 20 years).

- "Impact" is defined as similarity to subsequent patents:

$$\text{Impact}_i = \sum_{i \in F(j)} \rho_{ij}$$

  where $F(j)$ is the set of future patents (in, e.g., next 100 years).

Novelty, Impact, and Quality

- "Novelty" is defined by (negative) similarity to previous patents:

$$\text{Novelty}_j = -\sum_{i \in B(j)} \rho_{ij}$$

  where $B(j)$ is the set of previous patents (in, e.g., last 20 years).

- "Impact" is defined as similarity to subsequent patents:

$$\text{Impact}_i = \sum_{i \in F(j)} \rho_{ij}$$

  where $F(j)$ is the set of future patents (in, e.g., next 100 years).

- A patent has high quality if it is novel and impactful:

$$\text{Quality}_i = \frac{\text{Impact}_i}{-\text{Novelty}_i}$$

Validation

- For pairs with higher $\rho_{i,j}$, patent j is more likely to cite patent i.

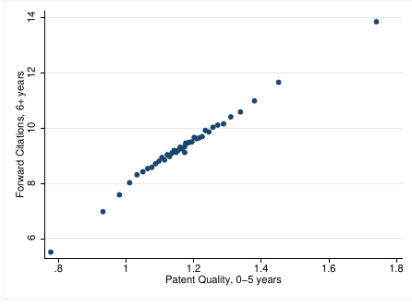
Validation

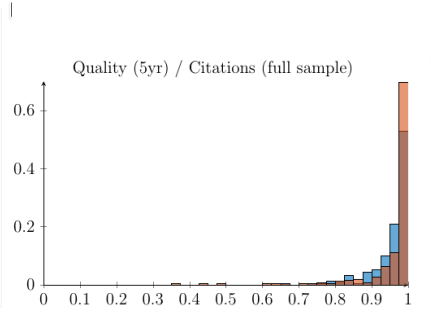
- For pairs with higher $\rho_{i,j}$, patent j is more likely to cite patent i.
- Patent office assigns 3-digit technology class code; similarity is significantly higher within class compared to across class.

Validation

- For pairs with higher $\rho_{i,j}$, patent j is more likely to cite patent i.
- Patent office assigns 3-digit technology class code; similarity is significantly higher within class compared to across class.

- Higher quality patents get more cites:

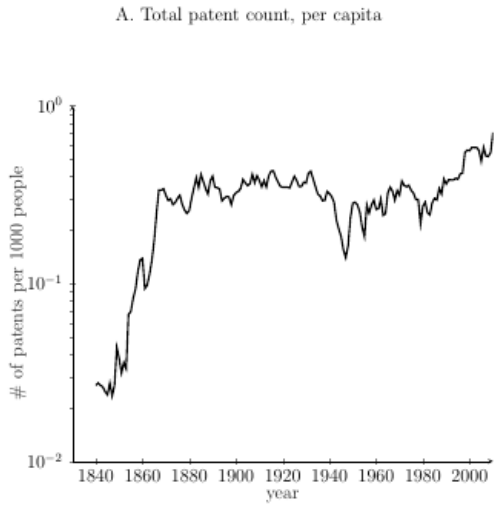


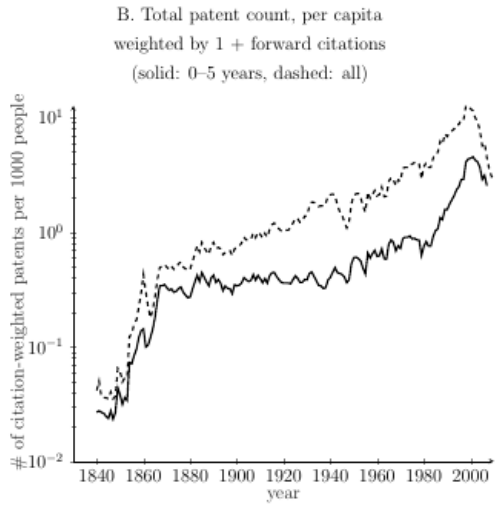


Most Innovative Firms

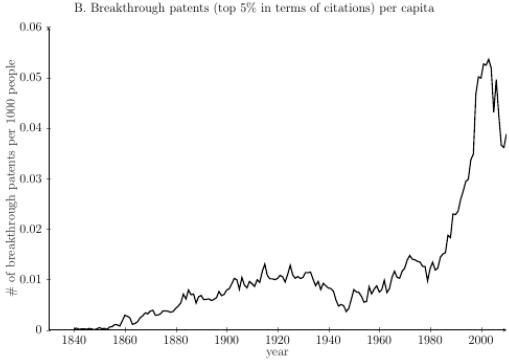
Assignee	First Year	# Breakthroughs
General Electric	1872	3,457
Westinghouse Electric Co.	1889	1,762
Eastman Kodak Co.	1890	2,244
Western Electric Co.	1899	1,222
AT&T (includes Bell Labs)	1899	5,645
Standard Oil Co.	1900	1,212
Dow Chemical Co.	1902	1,235
Du Pont	1905	3,353
International Business Machines	1908	14,913
American Cyanamid Co.	1909	690
Universal Oil Products Co.	1919	590
RCA	1920	3,222
Monsanto Company (inc. Monsanto Chemicals)	1921	902
Honeywell International, inc.	1928	872
General Aniline & Film Corp.	1929	1,181
Massachusetts Institute of Technology	1935	504
Philips	1939	1145
Texas Instruments	1960	2,088
Xerox	1961	2,198
Applied Materials	1971	510
Digital Equipment	1971	1,101
Hewlett-Packard Co.	1971	2,661
Intel	1971	2,629
Motorola, inc.	1971	4,129
Regents of the University of California	1971	823
United States Navy	1945	791
NCR	1973	737
Advanced Micro Devices	1974	1,195
Apple Computer	1978	864

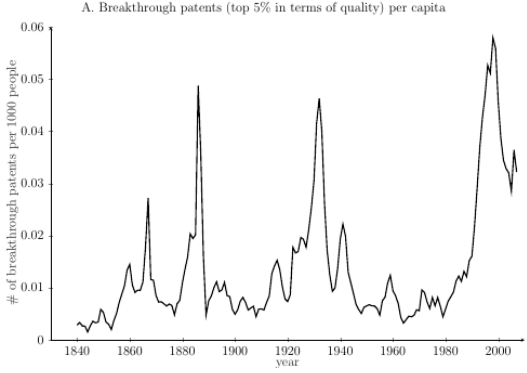
Patents per capita



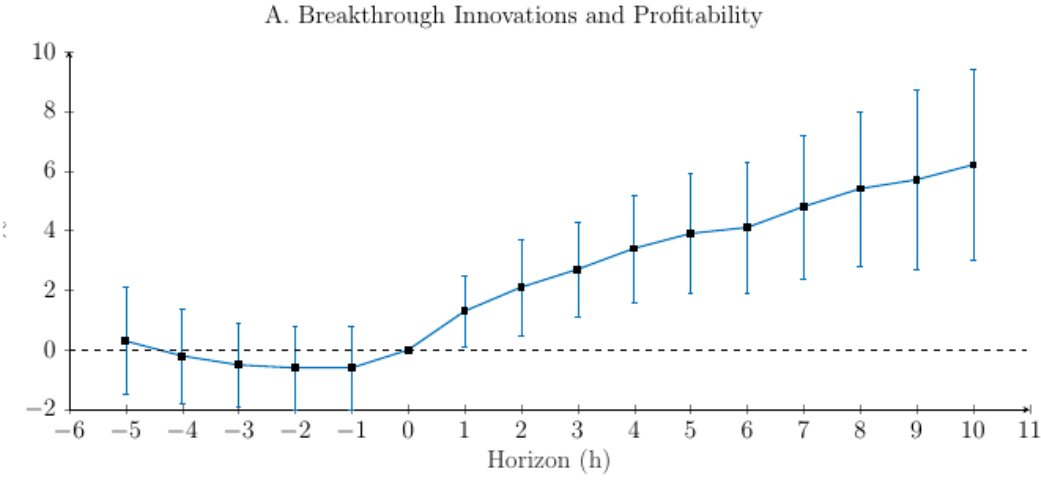


Breakthrough patents per capita





Breakthrough patents and firm profits

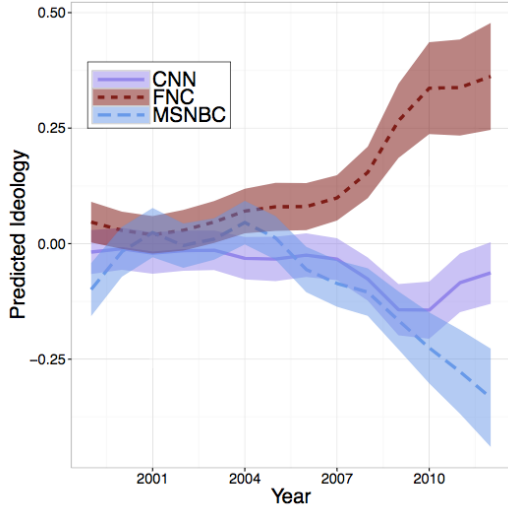


Outline

Cable News and Political Discourse

- Context:
  - U.S. congressional districts ($N = 435$), years 2005-2008
- Data:
  - transcripts for prime time shows in major cable channels
  - transcripts for congressional speeches in U.S House
  - geographical data on U.S House representatives
  - channel position and viewership for cable news channels

Fox News Channel is Politically Conservative



Martin and Yurukoglu (2017): Estimated ideology based on phrase usage for CNN, Fox News Channel (FNC), and MSNBC. Higher is more conservative.

Text Features

- Featurization:
  - Convert to lower case, remove punctuation, stopwords, numbers
  - Do stemming
  - Construct 3-grams for the observations
  - Remove rare 3-grams
- Create frequency matrices for congressional speakers by year, and for cable news transcripts for each channel by year.

Compute similarity of each speech to cable channels

What we have	What we want
frequency matrices	similarity columns
$M_{congress}$, M_{Fox}	$S_{congressFox}$
$M_{congress}$, M_{CNN}	$S_{congressCNN}$
$M_{congress}$, M_{MSNBC}	$S_{congressMSNBC}$

- cosine similarity captures linguistic similarity between TV shows and congress speeches
- need to normalize to get the similarity specific for Fox News Channel:

$$foxsim = \frac{2similarity(fox, congress)}{similarity(cnn, congress) + similarity(msnbc, congress)}$$

Regression Model

$$Y_i = \alpha + \rho V_i + X_i\beta + \epsilon_i$$

- congressional district i
- Y_i, a speeches' similarity to Fox News variable
- V_i, measure of Fox News viewership
- X_i, covariates
  - state-time fixed effects
  - demographic covariates
- ε_i, unobservable factors and randomness
- ρ, effect of Fox News on House speeches similarity to Fox

Cable television channel positions

- In 2000s, majority of American households had paid cable television.
- lineup of channels varies across local cable systems.
- channel positions set in mid to late 1990s, haphazardly, based on order of joining systems, and what channels were being replaced.
  - once channels are set, providers rarely change them.

Cable television channel positions

- In 2000s, majority of American households had paid cable television.
- lineup of channels varies across local cable systems.
- channel positions set in mid to late 1990s, haphazardly, based on order of joining systems, and what channels were being replaced.
  - once channels are set, providers rarely change them.
- Martin and Yurukoglu (2017) show that when Fox News has a lower channel number, that increases viewership.
  - **use channel position as instrumental variable**.

What is a valid instrumental variable?

Instrumental variable (IV) is a variable that:

1. Is correlated with causal variable of interest, V_i:

$$\text{Cov}[Z_i, V_i] \neq 0$$

2. Is uncorrelated with any other determinants of Y_i:

$$\text{Cov}[Z_i, \epsilon_i] = 0$$

What is a valid instrumental variable?

Instrumental variable (IV) is a variable that:

1. Is correlated with causal variable of interest, V_i:

$$\text{Cov}[Z_i, V_i] \neq 0$$

2. Is uncorrelated with any other determinants of Y_i:

$$\text{Cov}[Z_i, \epsilon_i] = 0$$

- The second requirement can be decomposed in two:
  - 2.1: Exogeneity: None of the unobserved factors affects the instrument:

$$\epsilon_i \nrightarrow Z_i$$

  - 2.2 Exclusion restriction: Instrument only affects outcome through treatment variable:

$$Z_i \nrightarrow \epsilon_i$$

- With a valid instrumental variable we can consistently estimate ρ in

$$Y_i = \alpha + \rho V_i + \epsilon_i$$

- With a valid instrumental variable we can consistently estimate ρ in

$$Y_i = \alpha + \rho V_i + \epsilon_i$$

- Write the covariance of Z_i and Y_i as:

$$\text{Cov}[Z_i, Y_i] = \rho \text{Cov}[Z_i, V_i] + \text{Cov}[Z_i, \epsilon_i]$$

- With a valid instrumental variable we can consistently estimate ρ in

$$Y_i = \alpha + \rho V_i + \epsilon_i$$

- Write the covariance of Z_i and Y_i as:

$$\text{Cov}[Z_i, Y_i] = \rho \text{Cov}[Z_i, V_i] + \text{Cov}[Z_i, \epsilon_i]$$

- The **exogeneity/exclusion assumption** is $\text{Cov}[Z_i, \epsilon_i] = 0$.

- With a valid instrumental variable we can consistently estimate ρ in

$$Y_i = \alpha + \rho V_i + \epsilon_i$$

- Write the covariance of Z_i and Y_i as:

$$\text{Cov}[Z_i, Y_i] = \rho \text{Cov}[Z_i, V_i] + \text{Cov}[Z_i, \epsilon_i]$$

- The **exogeneity/exclusion assumption** is $\text{Cov}[Z_i, \epsilon_i] = 0$.
- Thus:

$$\rho = \frac{\text{Cov}[Z_i, Y_i]}{\text{Cov}[Z_i, V_i]}$$

is a consistent population estimate.

Weak Instruments

- The bias of 2SLS can be written as:

$$\text{plim}\hat{\rho} = \rho + \frac{\text{Corr}[Z, \epsilon]}{\text{Cov}[V, Z]} \cdot \frac{\sigma_\epsilon}{\sigma_V}$$

  - When the instrument is weakly correlated with the endogenous regressor, the bias increases.

Weak Instruments

- The bias of 2SLS can be written as:

$$\text{plim}\hat{\rho} = \rho + \frac{\text{Corr}[Z, \epsilon]}{\text{Cov}[V, Z]} \cdot \frac{\sigma_\epsilon}{\sigma_V}$$

  - When the instrument is weakly correlated with the endogenous regressor, the bias increases.
- Can check for a weak instrument with first-stage F-statistic: it should be higher than 10.

Intuition for Instrumental Variables

- OLS Regression based on observables:
  - The consistency of the estimate relies on the "hope" that any unobserved factor that might affect the outcome variable is balanced across the treatment and the control group.
  - Therefore, any difference in outcomes between the control and the treatment group can be attributed to the treatment.

Intuition for Instrumental Variables

- OLS Regression based on observables:
  - The consistency of the estimate relies on the "hope" that any unobserved factor that might affect the outcome variable is balanced across the treatment and the control group.
  - Therefore, any difference in outcomes between the control and the treatment group can be attributed to the treatment.
- Instrumental variables:
  - We identify some source of variation in the assignment to the treatment which, for some reason, we know that it is orthogonal to any relevant unobserved variable which might be affecting the outcome variables.
  - We compare group of individuals that, due to the instrument, are assigned to the control and the treatment group. Any difference in outcomes between these two groups is attributed to the treatment.

Matrix Notation, and Comparison to OLS

With model $Y = X'\beta + U$ and instrument Z, we have

$$\beta_{\text{OLS}} = (X'X)^{-1}(X'Y)$$
$$\beta_{\text{IV}} = (Z'X)^{-1}(Z'Y)$$

Matrix Notation, and Comparison to OLS

With model $Y = X'\beta + U$ and instrument Z, we have

$$\beta_{\text{OLS}} = (X'X)^{-1}(X'Y)$$
$$\beta_{\text{IV}} = (Z'X)^{-1}(Z'Y)$$

$$\begin{aligned}\mathbb{E}[\beta_{\text{OLS}}] &= \mathbb{E}[(X'X)^{-1}(X'Y)] = \mathbb{E}[(X'X)^{-1}(X'(X'\beta + U))] \\ &= \beta + \underbrace{\mathbb{E}[(X'X)^{-1}(X'U)]}_{\text{OLS Bias}}\end{aligned}$$

$$\begin{aligned}\mathbb{E}[\beta_{\text{IV}}] &= \mathbb{E}[(Z'X)^{-1}(Z'Y)] = \mathbb{E}[(Z'X)^{-1}(Z'(X'\beta + U))] \\ &= \beta + \underbrace{\mathbb{E}[(Z'X)^{-1}(Z'U)]}_{\text{IV Bias}}\end{aligned}$$

$$\mathbb{E}[(X'X)^{-1}(X'U)] \gtrless \mathbb{E}[(Z'X)^{-1}(Z'U)]?$$

IV Approach for Cable News Viewership

- The first stage from Martin and Yurukoglu (AER 2017) is

$$V_i = \alpha + \gamma Z_i + \eta_i$$

- Z_i, Fox News channel number in district i

IV Approach for Cable News Viewership

- The first stage from Martin and Yurukoglu (AER 2017) is

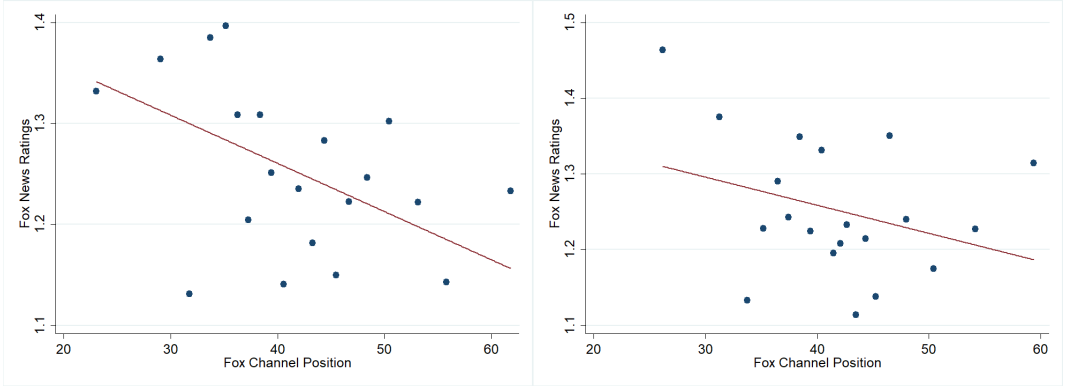
$$V_i = \alpha + \gamma Z_i + \eta_i$$

- Z_i, Fox News channel number in district i
- Second stage is

$$Y_i = \alpha + \rho \hat{V}_i + X_i \beta + \epsilon_i$$

estimated with two-stage least squares (2SLS).

Low Fox Channel Number → High Fox Viewership



Average Fox News viewership share plotted against Fox News channel position (left panel, without state-year controls; right panel, with controls).

Exogeneity Check: Channels Unrelated to Past Republican Vote Shares

Exogeneity Check: Channels Unrelated to Past Republican Vote Shares

	(1)	(2)	(3)	(4)	(5)	(6)
	Effect on 1996 Republican Presidential Vote Share					
FNC Channel	0.000375					
	(0.000558)					
Z_fnc		-0.00321				
		(0.00416)				
CNN Channel			8.78e-05			
			(0.000668)			
Z_cnn				-0.000871		
				(0.00460)		
MSNBC Channel					0.000468	
					(0.000345)	
Z_msnbc						-0.00392
						(0.00362)
Observations	1,398	1,398	1,398	1,398	1,398	1,398
R-squared	0.531	0.532	0.530	0.530	0.533	0.532
F-test	0.451	0.597	0.0173	0.0358	1.840	1.175

Regressions include state-year FEs. SEs in parentheses clustered by district.
*** p<0.01, ** p<0.05, * p<0.1

Reduced Form: Channel Position and Speech Similarity

	(1)	(2)	(3)	(4)	(5)	(6)
VARIABLES	Sim to Fox		Sim to CNN		Sim to MSNBC	
Z_fnc	0.0532**	0.0578**				
	(0.0243)	(0.0256)				
Z_cnn			0.0130	0.0180		
			(0.0207)	(0.0214)		
Z_msnbc					0.0149	0.0127
					(0.0281)	(0.0273)
Demo Controls		X		X		X
Observations	1,398	1,398	1,398	1,398	1,398	1,398
Adj. R^2	0.322	0.336	0.300	0.302	0.266	0.276

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

2SLS Effect of Fox Exposure on similarity to Fox

Table 8: 2SLS Estimates: Fox

	(1)	(2)	(3)	(4)
	2SLS Effect on Sim to Fox			
Fox Ratings	0.537**	0.658**	0.611**	0.703*
	(0.233)	(0.308)	(0.305)	(0.390)
Republican			0.202***	
			(0.0734)	
Prob(Repub|Text)				0.269
				(0.181)
Demo Controls		X		
Observations	1,398	1,398	1,398	1,326
Adj. R^2	-0.309	-0.403		
Kleibergen-Paap F	11.75	8.827	11.28	8.342

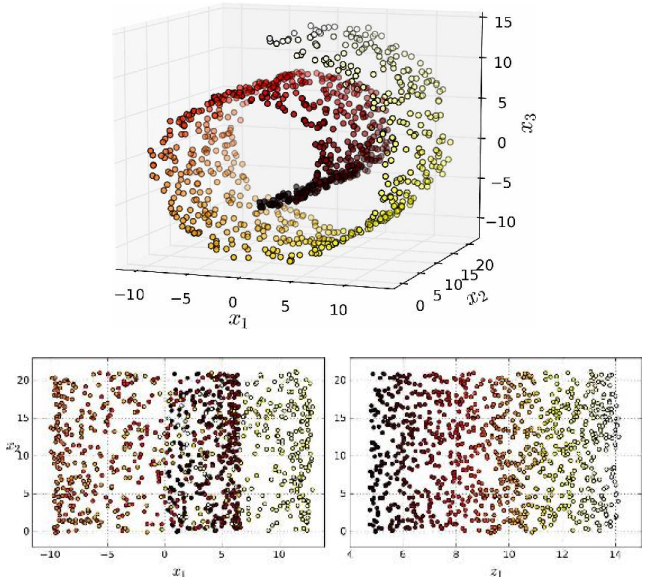
Robust standard errors in parentheses, clustered by district.
*** p<0.01, ** p<0.05, * p<0.1

2SLS estimates of effect of FNC ratings on congress speech similarity to FNC; standard errors in parenthesis clustered by district; * p<.1.

Outline

Dimensionality Reduction

- Especially in the case of text data, machine learning problems often involve thousands of features.
- Dimension reduction methods are needed:
  - not just for computational tractability, but also to help find a good solution.
  - also for data visualization – for example, to plot data in two dimensions.



The dimension reduction process matters: projecting down to two dimensions directly (left panel) might not isolate the variation we are interested in (as done in the right panel, which unrolls the Swiss Roll)

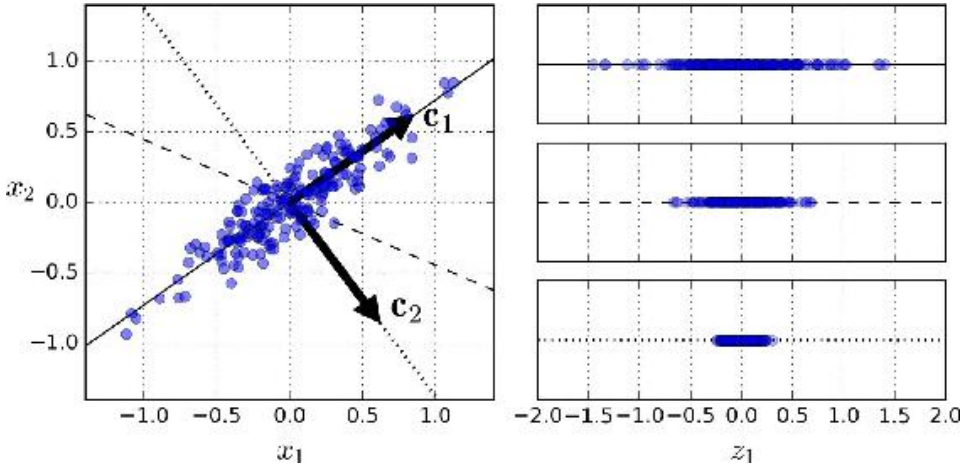
Manifold Learning

- The swiss roll is an example of a 2D manifold:
  - like many real-world data sets, the data are not uniformly distributed across the space.
  - can be modeled in a lower-dimensional subspace while retaining most of the information

Manifold Learning

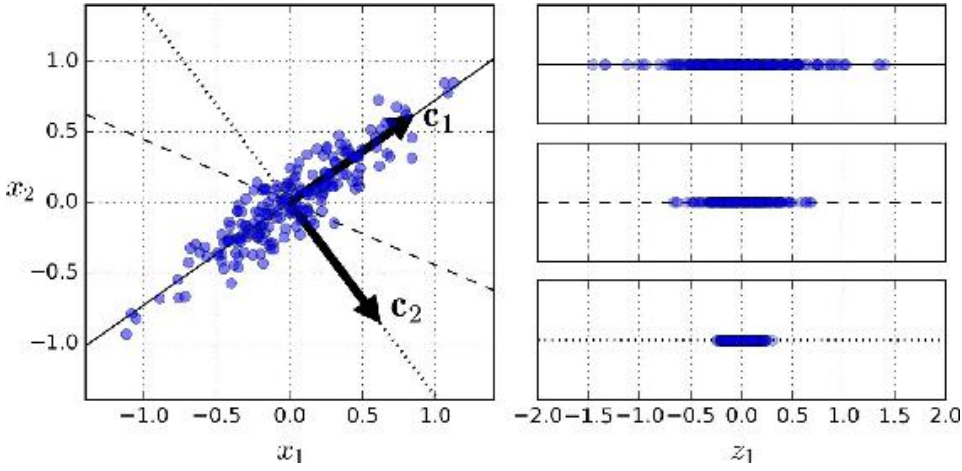
- The swiss roll is an example of a 2D manifold:
  - like many real-world data sets, the data are not uniformly distributed across the space.
  - can be modeled in a lower-dimensional subspace while retaining most of the information
- Dimension reduction methods in machine learning are motivated by this "manifold hypothesis."

PCA



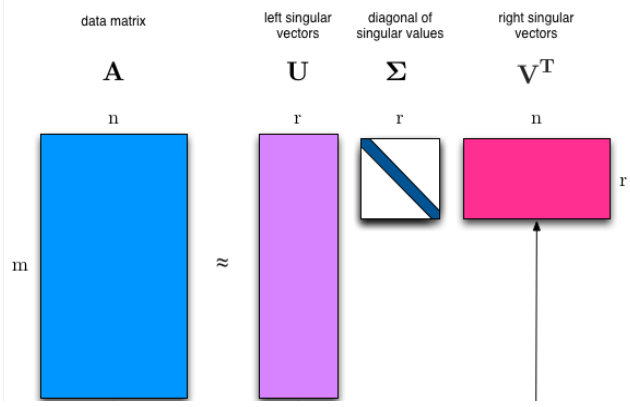
- PCA (principal components analysis), a popular dimension reduction technique.
  - identifies the axis that accounts for the largest amount of variance in the training set.
  - finds a second axis, orthogonal to the first, that accounts for the largest amount of the remaining variance, and so on

PCA



- PCA (principal components analysis), a popular dimension reduction technique.
  - identifies the axis that accounts for the largest amount of variance in the training set.
  - finds a second axis, orthogonal to the first, that accounts for the largest amount of the remaining variance, and so on
- The unit vector defining the ith axis is called the ith principal component.

PCA (SVD)



- SVD factorizes an $m \times n$ matrix A into an $m \times r$ orthogonal matrix U, representing documents, an $n \times r$ orthogonal matrix V, representing features, and the singular values Σ indicating the relative importance of factors.
  - The U matrix is the principal components matrix.

PCA Projection

- For principal components $\boldsymbol{c}_1, \boldsymbol{c}_2, ..., \boldsymbol{c}_n$, define

$$\boldsymbol{V}_n = \begin{bmatrix} \vdots & \vdots & & \vdots \\ \boldsymbol{c}_1 & \boldsymbol{c}_2 & \dots & \boldsymbol{c}_n \\ \vdots & \vdots & & \vdots \end{bmatrix}$$

PCA Projection

- For principal components $\boldsymbol{c}_1, \boldsymbol{c}_2, ..., \boldsymbol{c}_n$, define

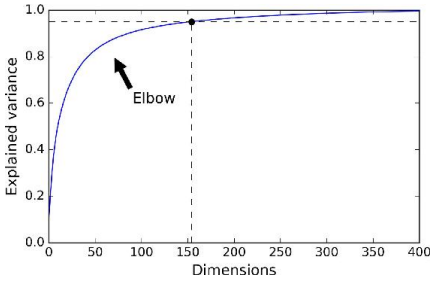
$$\boldsymbol{V}_n = \begin{bmatrix} \vdots & \vdots & & \vdots \\ \boldsymbol{c}_1 & \boldsymbol{c}_2 & \dots & \boldsymbol{c}_n \\ \vdots & \vdots & & \vdots \end{bmatrix}$$

- Dimension reduction works by projecting the data set down to the hyperplane defined by the first d principal components.

$$\boldsymbol{X}_{\text{PCA}} = \boldsymbol{X} \cdot \boldsymbol{V}_d$$

where the d subscript specifies that we have used the first d columns of $\boldsymbol{V}_n$

PCA for dimension reduction



- the principal components can be taken as a compressed representation of the original matrix.
- can be used as predictors instead of the original matrix.
- cosine similarities between PCA-reduced documents often behave better than that between original vectors.

Incremental PCA and numpy `memmap`

- Standard PCA requires you to load the whole data set into memory.
- numpy memmap loads big arrays from disk as needed
- sklearn.IncrementalPCA splits the data into mini-batches and trains gradually.

Pros and Cons of PCA

- Advantages:
  - fast to compute
  - good performance on many tasks in practice
  - components are orthogonal by construction

Pros and Cons of PCA

- Advantages:
  - fast to compute
  - good performance on many tasks in practice
  - components are orthogonal by construction
- Disadvantages:
  - lose (potentially a lot of) predictive information from X
  - Coefficients are not easily interpretable.

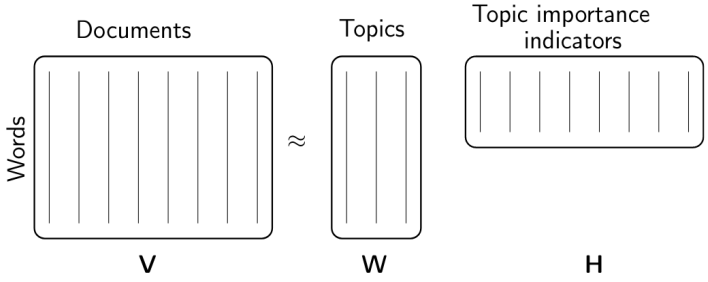
Pros and Cons of PCA

- Advantages:
  - fast to compute
  - good performance on many tasks in practice
  - components are orthogonal by construction
- Disadvantages:
  - lose (potentially a lot of) predictive information from X
  - Coefficients are not easily interpretable.
- Compromise:
  - keep strong predictors in feature set; take principal components of weak predictors and keep those

- For non-negative data (e.g. n-gram counts or frequencies), **Non-negative Matrix Factorization (NMF)** provides more interpretable factors than PCA.

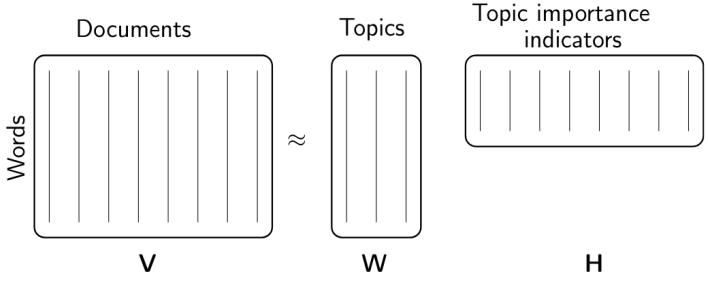
- For non-negative data (e.g. n-gram counts or frequencies), **Non-negative Matrix Factorization (NMF)** provides more interpretable factors than PCA.

assume $\mathbf{V} = [v_{fn}]$ is a (scaled) **term-document** co-occurrence matrix:
v_{fn} is the frequency of occurrences of word m_f in document d_n;



- For non-negative data (e.g. n-gram counts or frequencies), **Non-negative Matrix Factorization (NMF)** provides more interpretable factors than PCA.

assume $\mathbf{V} = [v_{fn}]$ is a (scaled) **term-document** co-occurrence matrix:
v_{fn} is the frequency of occurrences of word m_f in document d_n;

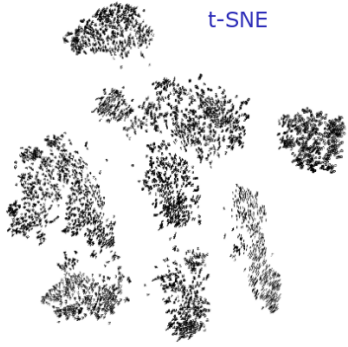


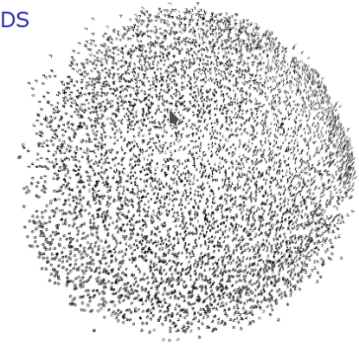
- Given a matrix of features X and a factorization rank r, NMF generates

$$\underbrace{V(:,j)}_{j\text{th document}} \approx \sum_{k=1}^{r} \underbrace{W(:,k)}_{k\text{th topic}} \underbrace{H(k,j)}_{\text{share } k \text{ in } j}$$

  - $W \geq 0$, the (interpretable) topics learned from the corpus
  - $H \geq 0$, the topic shares in each document

t-SNE and MDS





From: L. Van der Maaten & G. Hinton, Visualizing Data using t-SNE, Journal of Machine Learning Research 9 (2008) 2579-2605

- **t-Distributed Stochastic Neighbor Embedding (t-SNE)** reduces dimensionality while trying to keep similar instances close and dissimilar instances apart.
  - Useful for visualizing clusters of instances in high-dimensional space
- **Multidimensional Scaling (MDS)** reduces dimensionality while trying to preserve the distances between the instances.

Outline

Outline

k-means clustering

- k-means clustering separates documents into k groups:
  - Given document vectors $\{\vec{q}_1, \vec{q}_2, ..., \vec{q}_P\}$, the algorithm chooses clusters $Q = \{Q_1, Q_2, ...Q_k\}$, $k > 1$, to minimize the within-cluster sum of squares:

$$\arg\min_{Q} \sum_{i=1}^{k} \sum_{\vec{q} \in Q_i} ||\vec{q} - \mu_i||^2$$

where μ_i is centroid (mean vector) for cluster Q_i.

k-means clustering

- k-means clustering separates documents into k groups:
  - Given document vectors $\{\vec{q}_1, \vec{q}_2, ..., \vec{q}_P\}$, the algorithm chooses clusters $Q = \{Q_1, Q_2, ...Q_k\}$, $k > 1$, to minimize the within-cluster sum of squares:

$$\arg\min_{Q} \sum_{i=1}^{k} \sum_{\vec{q} \in Q_i} ||\vec{q} - \mu_i||^2$$

    where μ_i is centroid (mean vector) for cluster Q_i.
- It is important to scale features before doing k-means clustering.

k-means clustering

- k-means clustering separates documents into k groups:
  - Given document vectors $\{\vec{q}_1, \vec{q}_2, ..., \vec{q}_P\}$, the algorithm chooses clusters $Q = \{Q_1, Q_2, ...Q_k\}$, $k > 1$, to minimize the within-cluster sum of squares:

$$\arg\min_{Q} \sum_{i=1}^{k} \sum_{\vec{q} \in Q_i} ||\vec{q} - \mu_i||^2$$

    where μ_i is centroid (mean vector) for cluster Q_i.
- It is important to scale features before doing k-means clustering.
- Can pick optimal number of clusters with the **silhouette score**:

$$\frac{b_i - a_i}{\max(a_i, b_i)}$$

  - a_i = mean distance to members of i's cluster
  - b_i = mean distance to members of i's second-closest cluster.

How to use clusters

- Clusters of documents are (normally) on related topics.
  - topics are discrete mutually exclusive categories, which could be better than a topic model depending on your research task.

How to use clusters

- Clusters of documents are (normally) on related topics.
  - topics are discrete mutually exclusive categories, which could be better than a topic model depending on your research task.
- Can use a document's distance to all clusters as a feature set.

How to use clusters

- Clusters of documents are (normally) on related topics.
  - topics are discrete mutually exclusive categories, which could be better than a topic model depending on your research task.
- Can use a document's distance to all clusters as a feature set.
- Descriptive statistics: Can print documents closest to cluster centroids as a set of "representative documents".

Other clustering algorithms

- "k-medoid" clusters use the 1-norm:

$$\arg\min_{Q} \sum_{i=1}^{k} \sum_{\vec{q} \in Q_i} ||\vec{q} - \mu_i||$$

and μ_i would give the medoid (the median vector) for the cluster.

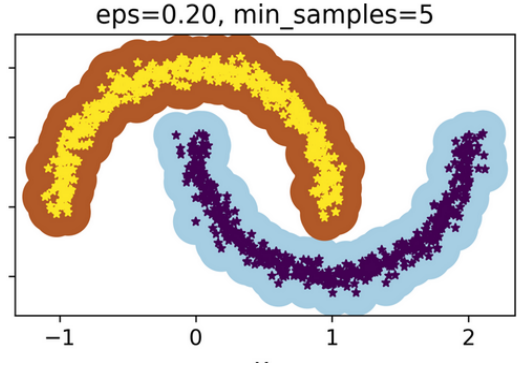
Other clustering algorithms

- "k-medoid" clusters use the 1-norm:

$$\arg\min_{Q} \sum_{i=1}^{k} \sum_{\vec{q} \in Q_i} ||\vec{q} - \mu_i||$$

and μ_i would give the medoid (the median vector) for the cluster.

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



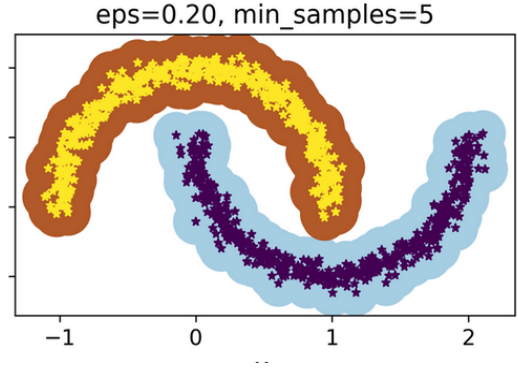
Other clustering algorithms

- "k-medoid" clusters use the 1-norm:

$$\arg\min_{Q} \sum_{i=1}^{k} \sum_{\vec{q} \in Q_i} ||\vec{q} - \mu_i||$$

and μ_i would give the medoid (the median vector) for the cluster.

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



- Agglomerative (hierarchical) clustering makes nested clusters.

Outline

Customization of Debt Contracts

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

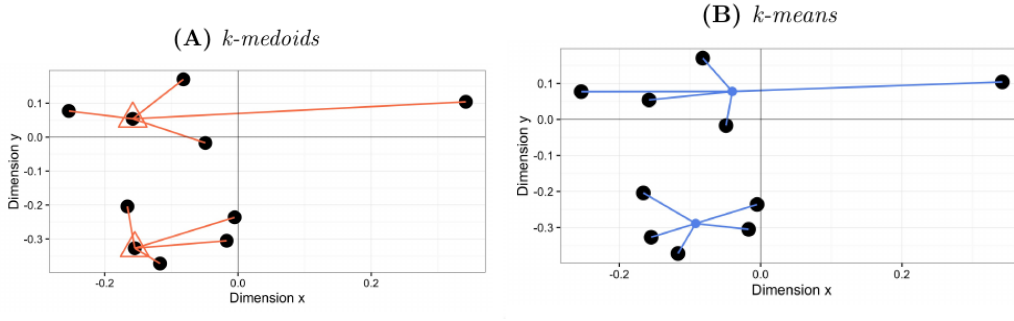
- Substantive question:
  - what explains customization and complexity in debt contracts?

Customization of Debt Contracts

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

- Substantive question:
  - what explains customization and complexity in debt contracts?
- Methodological question:
  - Can we use contract text to analyze customization and complexity?
    - previous work relies on expensive hand-coding

Measuring customization



- Medoides are nice because there is a singular document representing a cluster.
- Measuring **customization** of contracts:
  - distance to the k-medoid for all debt contracts drafted within a two-year window.

Descriptive findings

- Contracts are not boilerplate – there are important differences between contracts.
  - Text differences are driven by borrowers, rather than lenders

Descriptive findings

- Contracts are not boilerplate – there are important differences between contracts.
  - Text differences are driven by borrowers, rather than lenders
- More standardization is associated with larger deals.

Outline

Text analysis of corporate filings

"Text-Based Network Industries and Endogenous Product Differentiation" (2016)

- Data
  - 10-K annual filings from EDGAR, 1996-2008
  - Extract **"business description"** section, where firms are **legally required** to "describe the significant products they offer to the market" for the current fiscal year.

Text analysis of corporate filings

"Text-Based Network Industries and Endogenous Product Differentiation" (2016)

- Data
  - 10-K annual filings from EDGAR, 1996-2008
  - Extract **"business description"** section, where firms are **legally required** to "describe the significant products they offer to the market" for the current fiscal year.
- Text features:
  - nouns (including proper nouns), except location names (state, county, city)
  - drop words appearing in more than 25% of documents.
  - binary for whether word appears (rather than counts)

Text analysis of corporate filings

"Text-Based Network Industries and Endogenous Product Differentiation" (2016)

- Data
  - 10-K annual filings from EDGAR, 1996-2008
  - Extract **"business description"** section, where firms are **legally required** to "describe the significant products they offer to the market" for the current fiscal year.
- Text features:
  - nouns (including proper nouns), except location names (state, county, city)
  - drop words appearing in more than 25% of documents.
  - binary for whether word appears (rather than counts)
- Similarity:
  - cosine similarity between these vectors of "word-appears" indicators

Text-Based Industries

- The paper constructs "industries" as sets of firms with similar lists of nouns in their business descriptions.
  - they use an unusual clustering algorithm that probably ends up being close to k-means.

Text-Based Industries

- The paper constructs "industries" as sets of firms with similar lists of nouns in their business descriptions.
  - they use an unusual clustering algorithm that probably ends up being close to k-means.
- Qualitative validation: Example clusters for "Business Services" (SIC code 737):
  1. entertainment, video, television, royalties, internet, content, creative, promotional, copyright, game, sound, publishing
  2. client, database, solution, patient, copyright, secret, physician, hospital, health care, server, resource, functionality, billing
  3. internet, telecommunications, interface, communication, solution, platform, architecture, call, infrastructure), voice, functionality, server

Text industries predict shared outcomes better than standard groupings

TABLE 3
FIRM CHARACTERISTICS AND INDUSTRY CLASSIFICATIONS

Industry Controls	OI/Sales	OI/Assets	Sales Growth	Market Beta	Asset Beta
	A. Across-Industry Standard Deviations: Firm-Weighted Results; All Industry Classifications				
1. SIC-3 fixed effects	.204	.111	.126	.283	.271
2. NAICS-4 fixed effects	.205	.112	.136	.289	.276
3. 10-K-based 300 fixed effects	.231	.128	.157	.298	.285
4. TNIC equal-weighted average	.248	.142	.163	.332	.324
5. TNIC similarity-weighted average (excluding the focal firm)	.267	.153	.199	.384	.369
	B. Across-Industry Standard Deviations: Industry-Weighted Results; Transitive Industry Classifications Only				
1. SIC-3 fixed effects	.156	.111	.179	.347	.308
2. NAICS-4 fixed effects	.169	.126	.210	.414	.362
3. 10-K-based 300 fixed effects	.202	.139	.224	.469	.432

NOTE.—For a given variable indicated in the left-hand column, across-industry standard deviations are computed as the standard deviation of the industry average of the given variable across all firms in our sample (panel A) and across all industries (panel B). TNIC refers to text-based network industries.

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 wanted to 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 wanted to 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 methods, e.g. principal components analysis.

Some example questions

- How do U.S. politicians present their work to the public? What explains variation in representational style? (Grimmer 2013)

Some example questions

- How do U.S. politicians present their work to the public? What explains variation in representational style? (Grimmer 2013)
- Did electoral reform change the portfolio of issues addressed by politicians in Japan? (Catalinac 2016)

Some example questions

- How do U.S. politicians present their work to the public? What explains variation in representational style? (Grimmer 2013)
- Did electoral reform change the portfolio of issues addressed by politicians in Japan? (Catalinac 2016)
- What are the propaganda strategies of the Chinese government? (Roberts and Stewart 2016)

Some example questions

- How do U.S. politicians present their work to the public? What explains variation in representational style? (Grimmer 2013)
- Did electoral reform change the portfolio of issues addressed by politicians in Japan? (Catalinac 2016)
- What are the propaganda strategies of the Chinese government? (Roberts and Stewart 2016)
- How do central bankers respond to an increase in transparency over their discussions? (Hansen, McMahon, and Pray 2015)

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

Quinn, Monroe, Colaresi, Crespin, and Radev (2010).

Outline

Latent Dirichlet Allocation (LDA)

- Idea: documents exhibit each topic in some proportion.
  - Each document is a distribution over topics.
  - Each topic is a distribution over words.

Latent Dirichlet Allocation (LDA)

- Idea: documents exhibit each topic in some proportion.
  - Each document is a distribution over topics.
  - Each topic is a distribution over words.
- Latent Dirichlet Allocation (e.g. Blei 2012) estimates:
  - The distribution over words for each topic.
  - The proportion of a document in each topic, for each document.

Latent Dirichlet Allocation (LDA)

- Idea: documents exhibit each topic in some proportion.
  - Each document is a distribution over topics.
  - Each topic is a distribution over words.
- Latent Dirichlet Allocation (e.g. Blei 2012) estimates:
  - The distribution over words for each topic.
  - The proportion of a document in each topic, for each document.
- Maintained assumptions: Bag of words/phrases, and fix number of topics ex ante.

Document-term Matrix X

	W1	W2	W3	Wn
D1	0	2	1	3
D2	1	4	0	0
D3	0	2	3	1
Dn	1	1	3	0

- A corpus of N documents $D_1, D_2, D_3 \ldots D_n$

Document-term Matrix X

	W1	W2	W3	Wn
D1	0	2	1	3
D2	1	4	0	0
D3	0	2	3	1
Dn	1	1	3	0

- A corpus of N documents $D_1, D_2, D_3 \ldots D_n$
- Vocabulary of M words $W_1, W_2 \ldots W_m$.

Document-term Matrix X

	W1	W2	W3	Wn
D1	0	2	1	3
D2	1	4	0	0
D3	0	2	3	1
Dn	1	1	3	0

- A corpus of N documents $D_1, D_2, D_3 \ldots D_n$
- Vocabulary of M words $W_1, W_2 \ldots W_m$.
- The value of i, j cell gives the frequency count of word W_j in Document D_i.

Matrix factoring

- LDA factors the document-term matrix into two lower-dimensional matrices, M_1 and M_2:

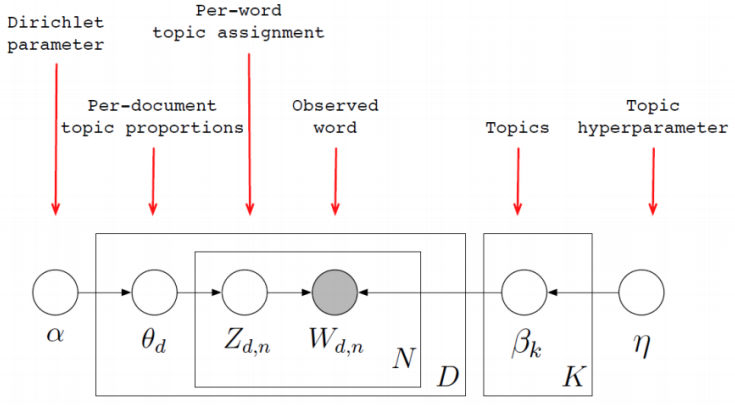
	K1	K2	K3	K
D1	1	0	0	1
D2	1	1	0	0
D3	1	0	0	1
Dn	1	0	1	0

	W1	W2	W3	Wm
K1	0	1	1	1
K2	1	1	1	0
K3	1	0	0	1
K	1	1	0	0

- M_1 is a $N \times K$ document-topic matrix
- M_2 is a $K \times M$ topic-term matrix.

A Bayesian Model

Figure: Plate Notation of Latent Dirichlet Allocation



Source: Brandon Stewart Topic Models Slides.

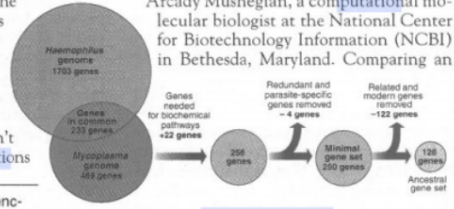
- NMF is equivalent to LDA with a uniform (rather than dirichlet) prior on topics (Faleiros and Lopes 2016).

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



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

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

Image from Hanna Wallach

Why does this work? Co-occurrence

- Where is the information for each word's topic?
  - We are learning the pattern of what words occur together.

Why does this work? Co-occurrence

- Where is the information for each word's topic?
  - We are learning the pattern of what words occur together.
- The model wants a topic to contain as few words as possible, but a document to contain as few topics as possible.
  - This tension is what makes the model work.

Setting the number of topics

- the "statistically optimal" topic count is usually too high for the topics to be interpretable/useful.
- Implementations like Mallet provide coherence scores which work really well.
- In general:
  - if there are duplicate topics, reduce the number
  - if topics aren't specific enough, increase the number

Using an LDA Model

- Once trained, can easily get topic proportions for a document.
  - 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 document.
  - for any document – doesn't have to be in training corpus.
  - main topic is the highest-probability topic
- Can also get representative documents for each topic.
  - documents with highest share in that topic.

Outline

Topic modeling Federal Reserve Bank transcripts

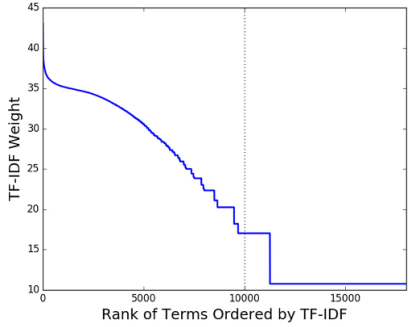
Hansen, McMahon, and Prat (QJE 2017)

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

Topic modeling Federal Reserve Bank transcripts

Hansen, McMahon, and Prat (QJE 2017)

- Use LDA to analyze speech at the FOMC (Federal Open Market Committee).
  - private discussions among committee members at Federal Reserve (U.S. Central Bank)
  - transcripts: 150 meetings, 20 years, 26,000 speeches, 24,000 unique words.
- Pre-processing:
  - drop stopwords, stems, etc.
  - Drop words with low TF-IDF weight



LDA Training

Hansen, McMahon, and Prat (QJE 2017)

- $K = 40$ topics selected for interpretability / topic coherence.
  - the "statistically optimal" $K = 70$, but these were less interpretable.

LDA Training

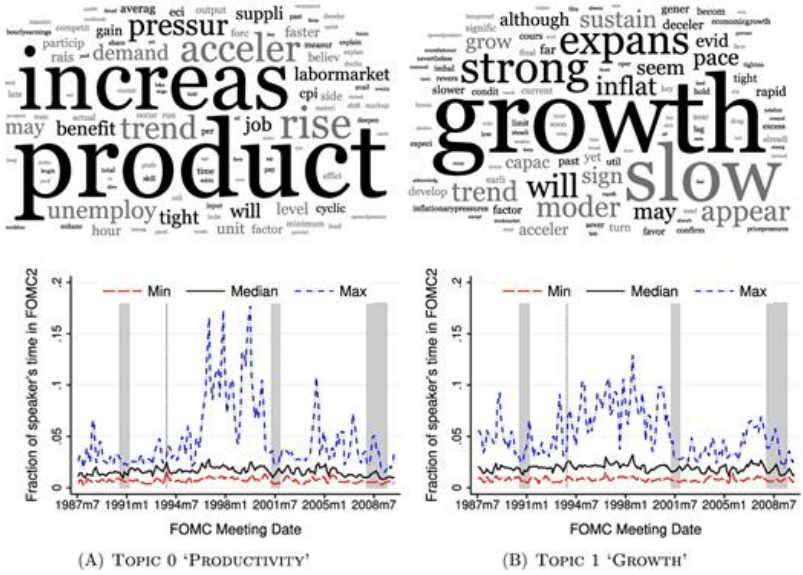
Hansen, McMahon, and Prat (QJE 2017)

- $K = 40$ topics selected for interpretability / topic coherence.
  - the "statistically optimal" $K = 70$, but these were less interpretable.
- hyperparameters $\alpha = 50/K$ and $\eta = .025$ to promote sparse word distributions (and more interpretable topics).

													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

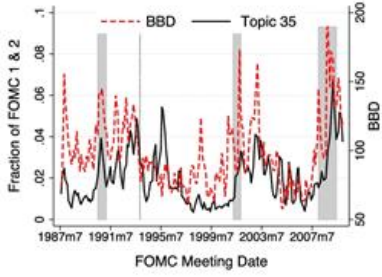
Hansen, McMahon, and Prat (QJE 2017)

(A) Topic 38 'Financial sector'

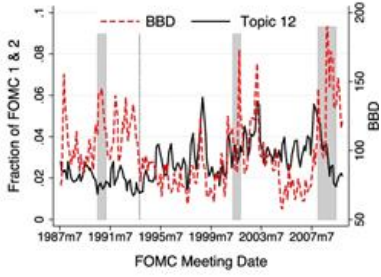
(B) Topic 39 'Economic weakness'

FOMC Topics and Policy Uncertainty

Hansen, McMahon, and Prat (QJE 2017)



(A) Topic 35 'Fiscal issues'



(B) Topic 12 'Risk'

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.

Outline

Extensions

- There are many extensions/variants of LDA.
  - But almost all of them are context-specific.

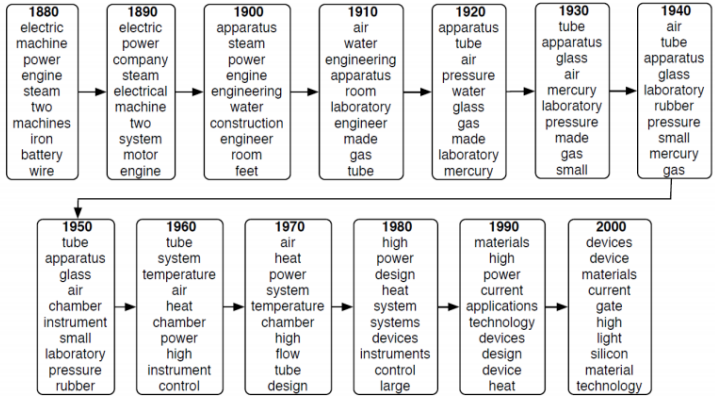
Extensions

- There are many extensions/variants of LDA.
  - But almost all of them are context-specific.
  - LDA is great because it works so well across different domains.

Dynamic Topic Model

Blei and Lafferty 2006.

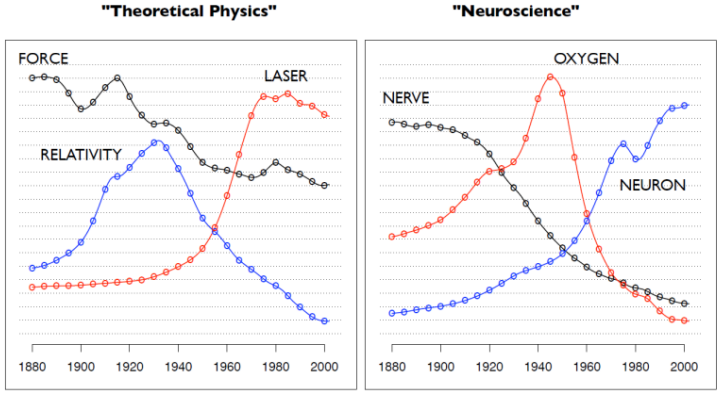
Figure: Topic Evolution over Time



Dynamic Topic Model

Blei and Lafferty 2006.

Figure: Word Use in Topics Over Time



Structural Topic Model = LDA + Metadata

Roberts, Stewart, and Tingley

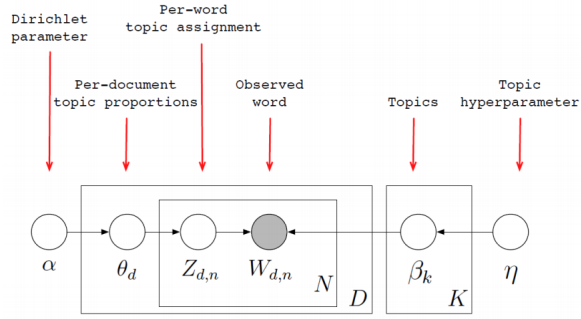
- STM (`structuraltopicmodel.com`) provides two ways to include contextual information:
  - Topic prevalence can vary by metadata
    - e.g. Republicans talk about military issues more then Democrats

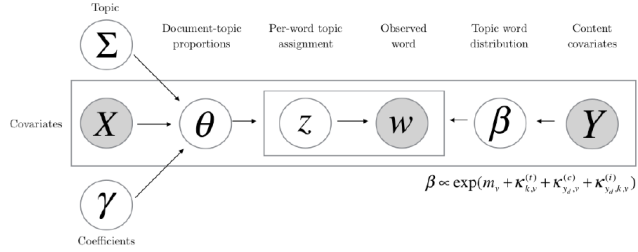
Structural Topic Model = LDA + Metadata

Roberts, Stewart, and Tingley

- STM (`structuraltopicmodel.com`) 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 differently from Democrats.

LDA vs. STM – Illustration





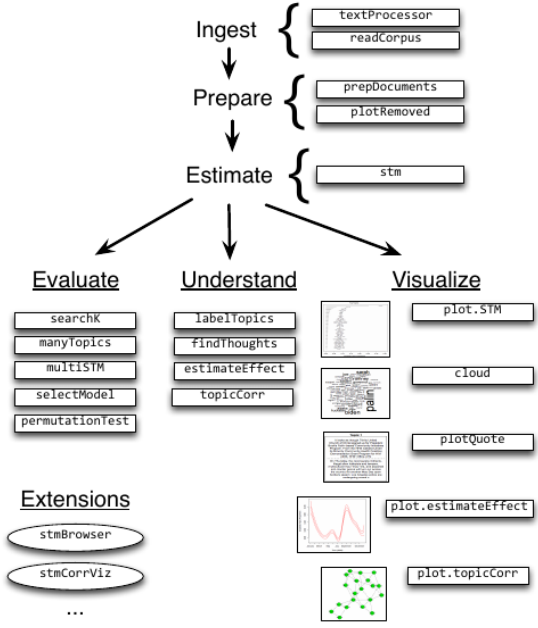
stm Package in R

- Complete workflow: raw texts $\rightarrow$ figures
- Simple regression style syntax using formulas

```
mod.out <- stm(documents,vocab, K=10, prevalence= ~ paper + s(time),
               data=metadata, init.type="Spectral")
```

- many functions for summarization, visualization and checking
- Complete vignette online with examples

stm has great functions/features



Caveats

- 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

Caveats

- 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
- STM does not work with streaming data (yet)
  - have to load the whole corpus into memory