Text as Data for the Social Sciences (I)

Germain Gauthier, Felix Lennert, Etienne Ollion

germain.gauthier@polytechnique.edu

SICSS Paris

June 2022

The Rise of Text Data

- The digital era generates considerable amounts of text.
  - Social media and internet queries
  - Wikipedia, online newspapers, TV transcripts
  - Digitized books, speeches, laws
- It is matched with a similar increase in computational resources.
  - Moore's law = processing power of computers doubles every two years (since the 70s!)
- Natural language processing is a **data-driven approach to the analysis of text documents**.

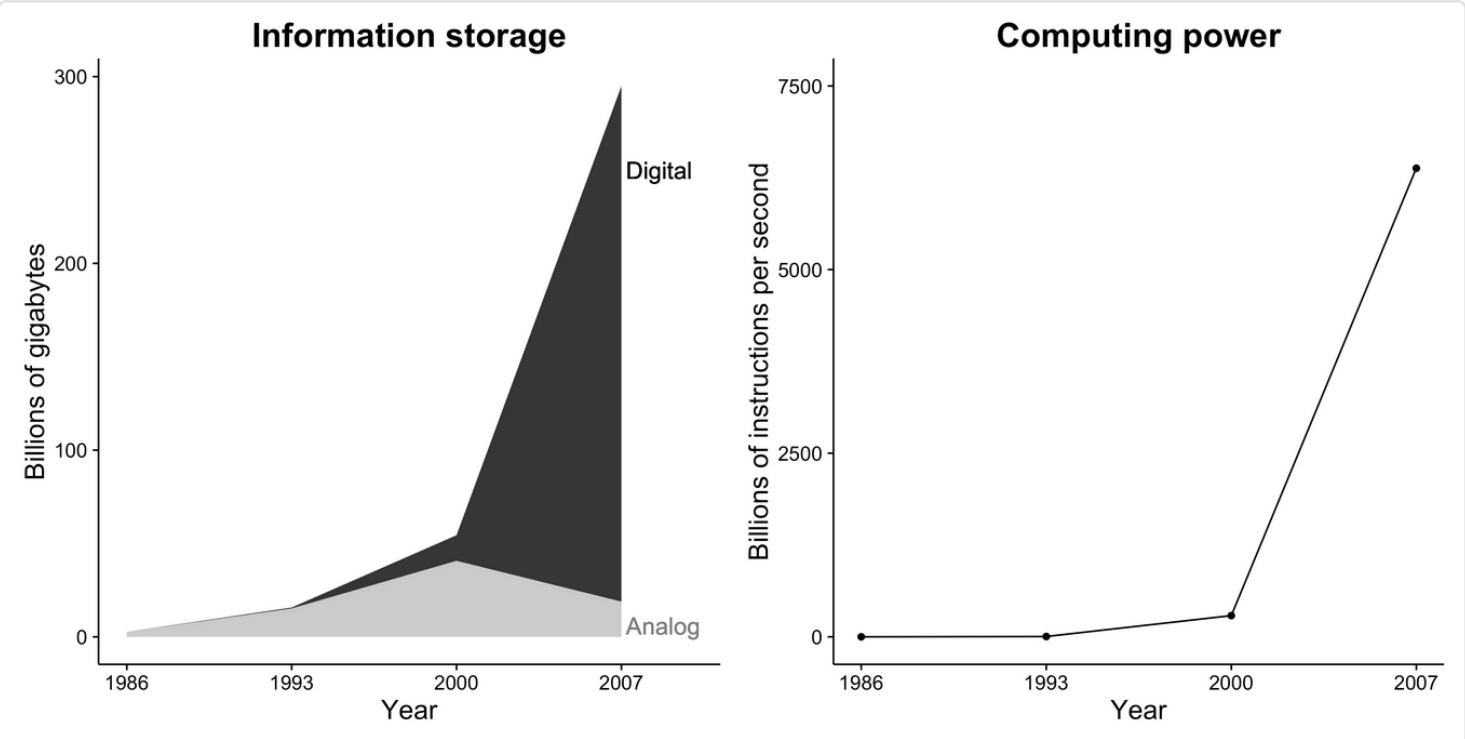


Figure 1.1: Information storage capacity and computing power are increasing dramatically. Further, information storage is now almost exclusively digital. These changes create incredible opportunities for social researchers. Adapted from Hilbert and López (2011), figures 2 and 5.

Source: Bit by Bit: Social Science Research in the Digital Age, Matthew J. Salganik

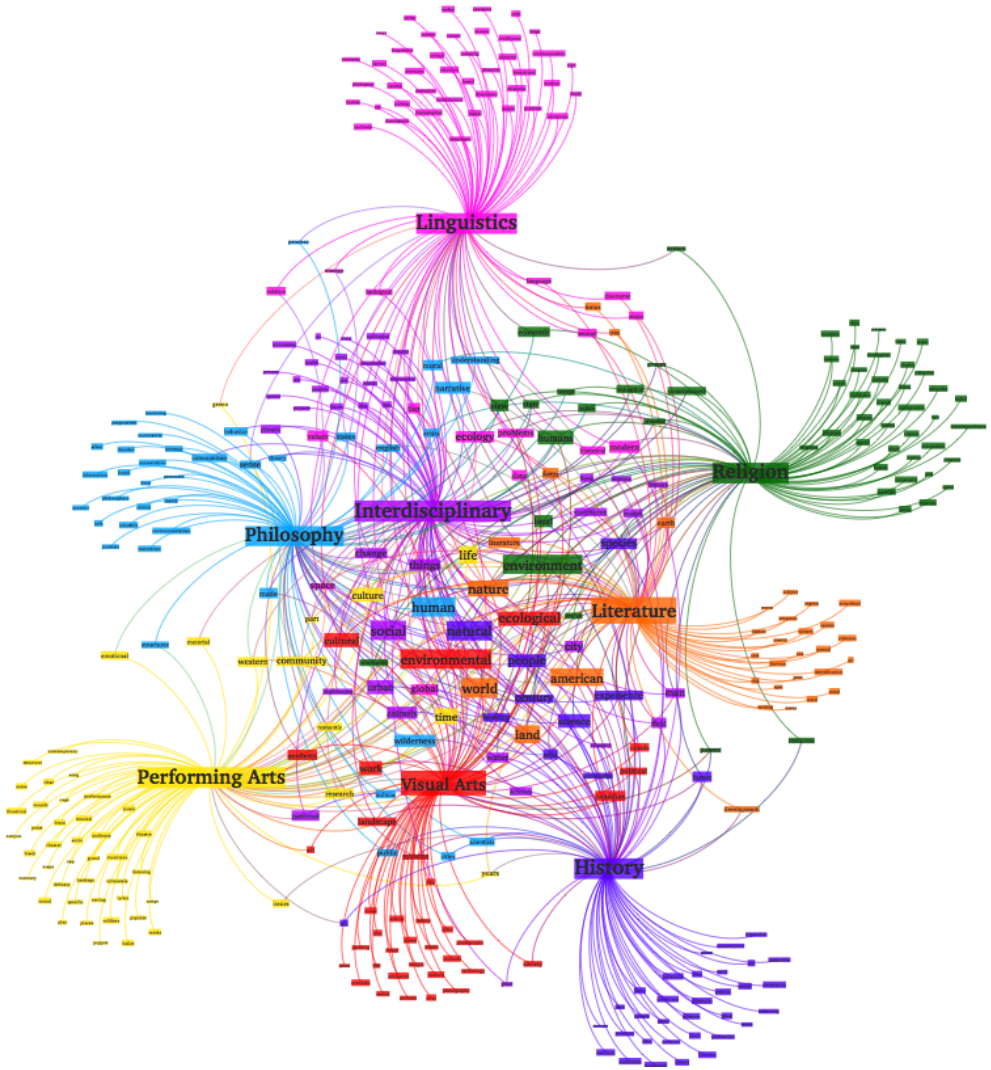
Typical Applications

- In everyday life:
  - Search engines
  - Translation services
  - Spam detection
  - But also bots when you try to terminate a contract...
- In the social sciences:
  - Has political polarization increased?
  - Can we detect Russian trolls or terrorists?
  - Can we predict the occurrence of protests?
  - Or perhaps GDP growth and financial markets?

This Course

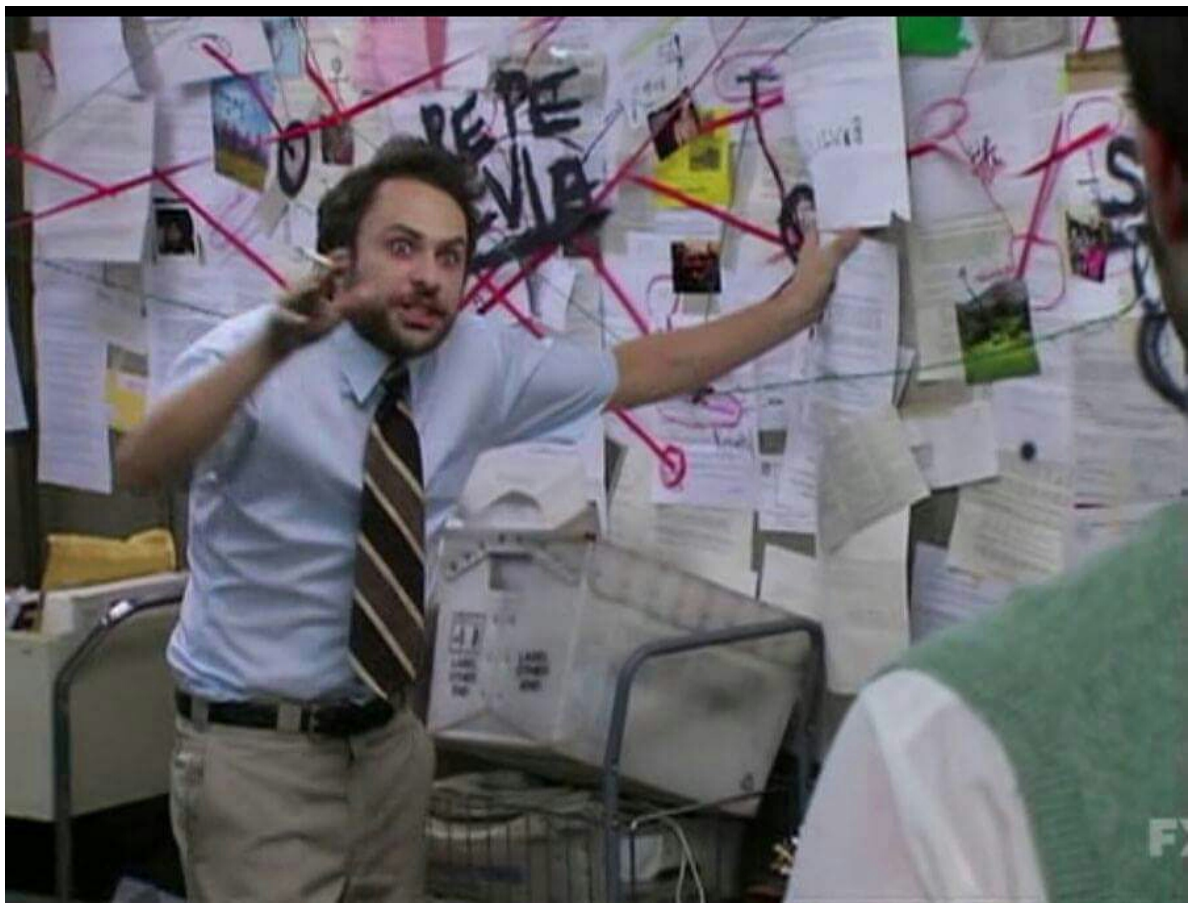
- Overview of text as data **in theory** and **in practice**.
- Rigorous introduction to the **general theoretical framework**.
- Focus on **applications for social scientists**.

Figure: Cool Topic Model Visualization



Source: Joyce Xu on Medium

Figure: Picture of me explaining topic models.



Plan

Modeling Text

- Raw text is typically **unstructured**.
- The information we are after is mixed in with irrelevant information.
- To put some structure on the text, we need a **a statistical model**.
- All models simplify and throw some information away (i.e., make assumptions).
- But good models retain essential information.

The Curse of High-dimensionality

- Raw text is typically **high-dimensional**.
- Suppose each document is composed of w words drawn from a vocabulary of p possible words. Then the unique representation of these documents has dimension p^w.
- This quickly leads to absurdly large numbers.
- We will run into **high-dimensional statistics**.

  e.g., dimension reduction, feature selection, etc.

The Typical Research Pipeline

- Let a corpus $\mathcal{C}$ be a collection of text documents.
- A typical research pipeline involves:
  1. **A Featurization Approach**
     We wish to represent $\mathcal{C}$ as numerical array W.
  2. **A Measurement Approach**
     A mapping f from features W to outcomes of interest Y.
  3. **Some Insights (hopefully!)**
     A causal or descriptive analysis of $\hat{Y}$.

1. Featurization

- First, we need a mathematical representation of the text.
- We wish to represent the corpus $\mathcal{C}$ as a numerical array W.
- This involves **cleaning the text** and removing things we think are uninformative *ex ante*.
- The most common (and simple) representation of clean text is **word frequencies**.
- But we will also see richer (and more demanding) representations in the second part of the course.

2. Measurement

- We have a numerical representation W of our corpus $\mathcal{C}$.
- Next, we need a mapping f from features W to outcomes Y.
- Two common families in machine learning.

1. **Supervised learning**
   We learn f based on *labeled* data.
   e.g., Predict the GDP from newspapers text.
2. **Unsupervised learning**
   We learn f based on *unlabeled* data.
   e.g., Uncovering latent trending topics in news.

3. Insights

- We have a prediction of the outcome: $\hat{Y} = f(W)$.
- Next, we can use $\hat{Y}$ for standard statistical analysis.
- Two common approaches in the social sciences.

1. **Descriptive**

   We study $\hat{Y}$ with no attempt to make causal claims.

   e.g., What narratives do Democrats emphasize in the U.S. Congress?

2. **Causal**

   For a set of observed variables X, we attempt to understand whether $\hat{Y}$ *causes* or *is the result of* X.

   e.g., After attending an economics training program, are judges more likely to use economic jargon in their rulings?

Application: Is media slant contagious?

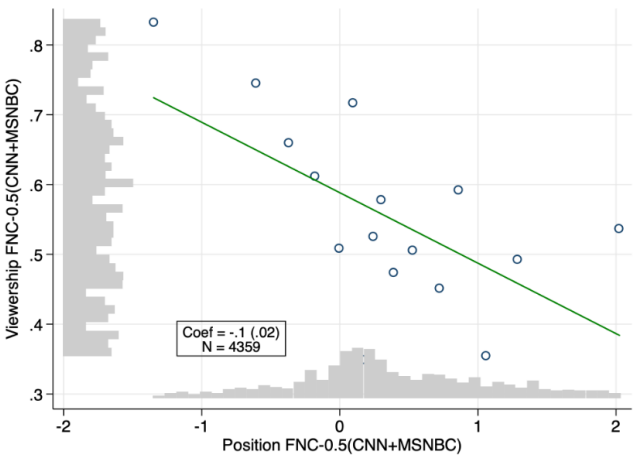
- **Paper:** Widmer, Galletta & Ash (2022) study how Fox News affects local reporting in U.S. newspapers.
- **Data:** News show transcripts from FNC, CNN, and MSNBC + local newspaper articles snippets (2005-2008)

1. **Featurization:** Lowercase, drop common words and non-letter characters, stem, and focus on bigram frequencies.
2. **Measurement:** Using the news show transcripts, they train a penalized regression model to identify Fox News content relative to CNN/MSNBC. They then apply the model to local newspaper snippets.
3. **Main insight:** Instrumental variables strategy for causal inference. More Fox News $\rightarrow$ more Fox News-like reporting in the local newspapers.

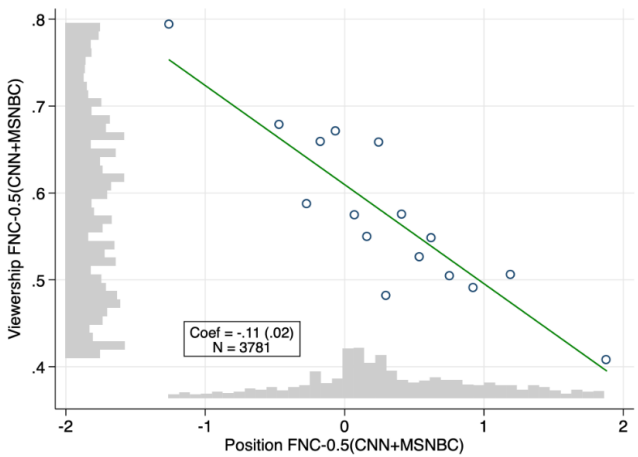
Empirical Results

Figure 3: First Stage: Cable Channel Position and Cable News Viewership

(a) No controls



(b) With controls



Notes: Binned scatterplots (16 bins) of standardized viewership of FNC-0.5(CNN+MSNBC) against standardized position of FNC-0.5(CNN+MSNBC). Cross-section with newspaper-county-level observations weighted by newspaper circulation in each county. On the left, state fixed effects are included. On the right, state fixed effects, as well as demographic controls (see Appendix Table A.2), channel controls (population share with access to each of the three TV channels), and generic newspaper language controls (vocabulary size, avg. word length, avg. sentence length, avg. article length) are included. In grey (next to the axes), we show the distributions of the underlying variables.

Source: Media Slant is Contagious, Widmer et al., Working Paper, (2022)

Empirical Results (Continued)

Table 4: Cable News Effects on Newspaper Content (2SLS)

Dep. variable: Slant_{ijs}=$\Pr(FNC\|\text{Text}_{ijs})$	(1)	(2)	(3)
FNC Viewership (rel. to CNN/MSNBC)	0.314***	0.311***	0.318**
	(0.114)	(0.113)	(0.126)
K-P First-Stage F-stat	36.553	36.298	34.147
N observations	3781	3781	3781
State FE	X	X	X
Demographic controls	X	X	X
Channel controls		X	X
Newspaper language controls			X

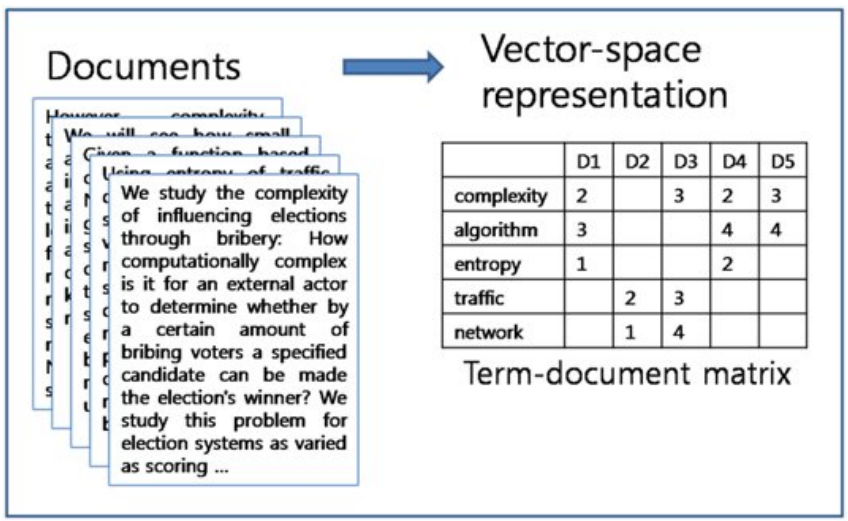
Notes: 2SLS estimates. Cross-section with newspaper-county-level observations weighted by newspaper circulation in each county. The dependent variable is newspaper language similarity with FNC (the average probability that a snippet from a newspaper is predicted to be from FNC), Slant_{ijs}. The right-hand side variable of interest is instrumented FNC viewership relative to averaged CNN and MSNBC viewership: *Viewership (FNC - 0.5(MSNBC - CNN))*. All columns include state fixed effects and demographic controls as listed in Appendix Table A.2. Column 2 also includes channel controls (population shares with access to each of the three TV channels). Column 3 controls for generic newspaper language features (vocabulary size, avg. word length, avg. sentence length, avg. article length). Standard errors are multiway-clustered at the county and at the newspaper level (in parenthesis): * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Source: Media Slant is Contagious, Ash et al., Working Paper, (2022)

Plan

Counting Words

- The simplest way to represent text documents is word frequencies.
- This is referred to as the **bag-of-words** approach.
- Word frequencies are summarized into a **term-document matrix**.
- Documents are columns and frequency counts are rows (or vice versa).



	D1	D2	D3	D4	D5
complexity	2		3	2	3
algorithm	3			4	4
entropy	1			2	
traffic		2	3		
network		1	4		

Some Refinements

- The unit for counts is called a **token**.
- Tokens can also include symbols and digits.

  e.g., #, !, ?, haha, 2008, etc.

- Depending on the application, some tokens can be uninformative and are removed (i.e., stopwords).

  e.g., she, he, the, a, etc.

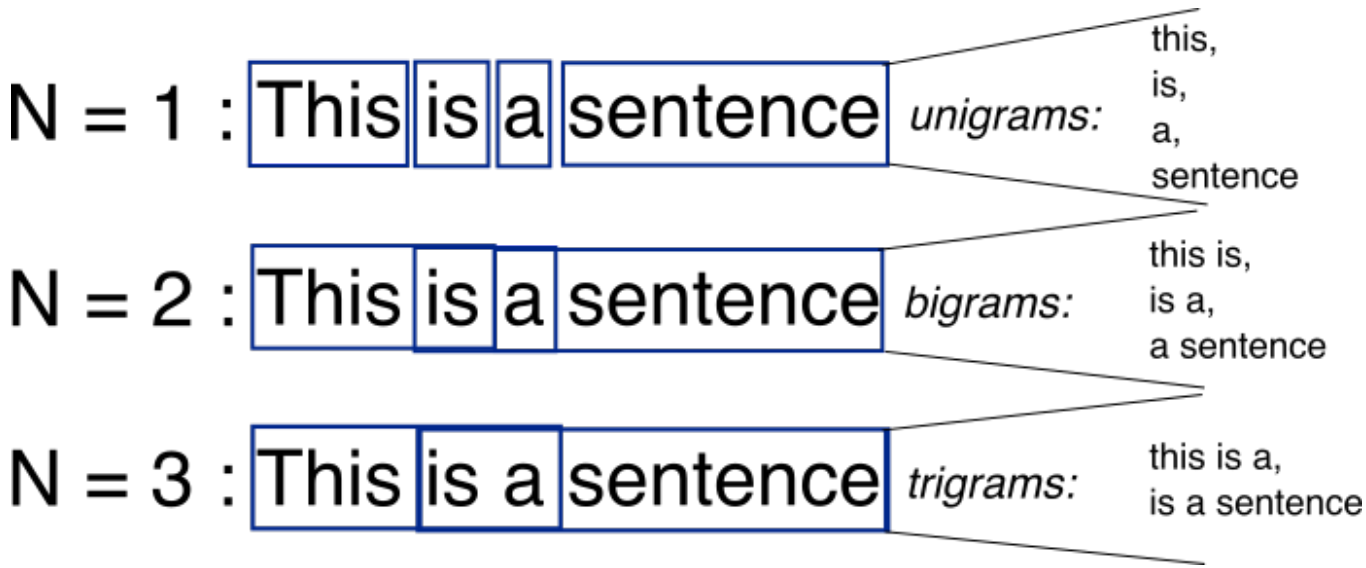
- Some tokens mean the same thing and are grouped together (via stemming or lemmatization).

  e.g., animal and animals, eating and eat, etc.

- The final set of tokens considered is the **vocabulary**.

Some Refinements

- The bag-of-words approach can be generalized to arbitrarily large token sequences called **n-grams**.



What should we do with W?

Plan

Dictionaries

- The **dictionary approach** consists of:
  - A narrow vocabulary based on a set of pre-defined tokens.
  - A deterministic mapping f from the features W to the outcomes Y (i.e., no machine learning involved).
- Extensively used for **sentiment analysis**:
  - Let (w_i, s_i) be pairs of words w_i and their associated sentiment score $s_i \in [-1, 1]$.

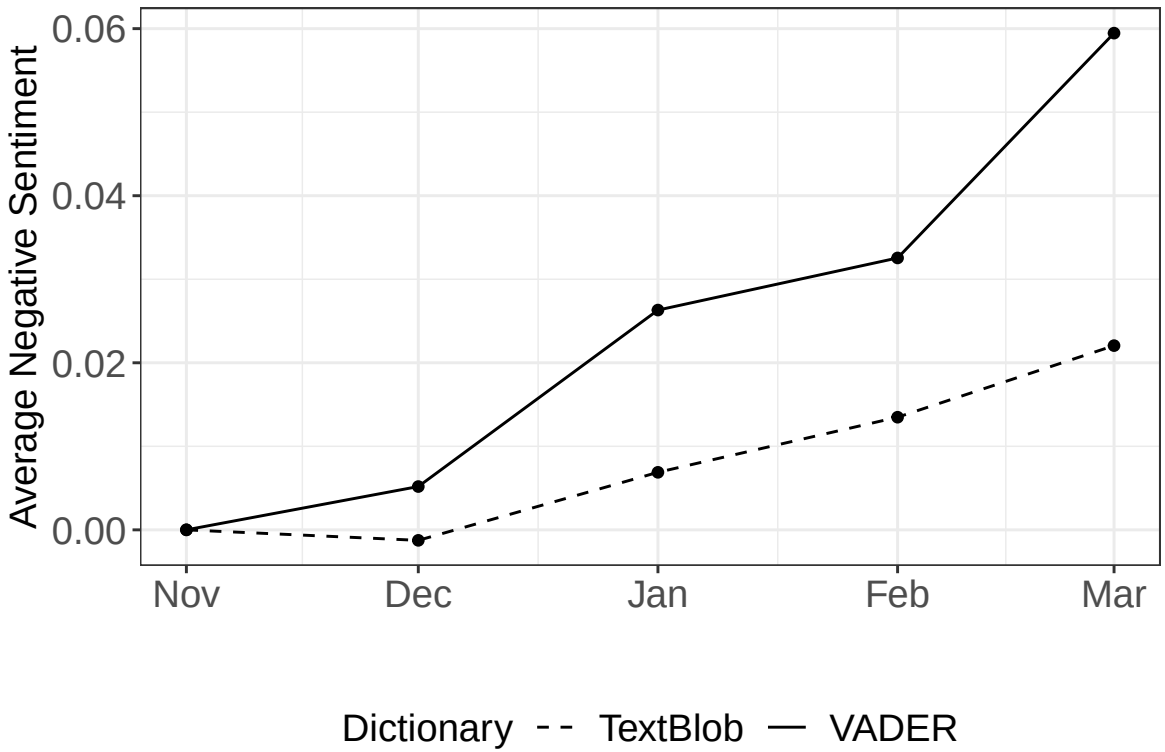
    e.g., ("perfect", 0.8), ("awful", -0.9)
  - The sentiment score for any phrase j of k tokens is a weighted average

$$s_j = \frac{1}{K} \sum_{i=1}^{k} s_i$$

Application: Yellow Vests Facebook Pages

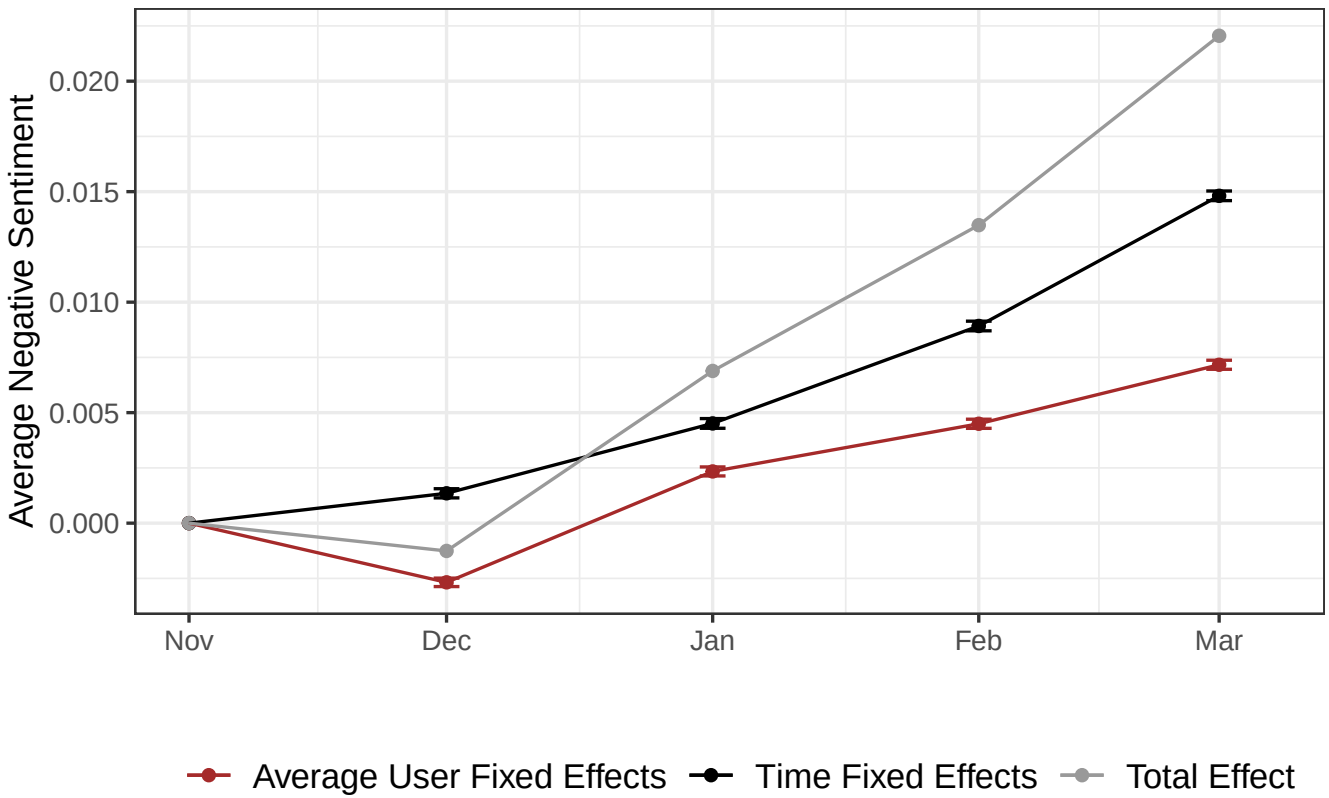
- **Paper:** Boyer et al. (2022) study the role of social media in the Yellow Vests protests that took place in 2018 in France.
- **Data:** 600+ active Yellow Vests Facebook pages between November 2018 and April 2019
- **Method:** Dictionary-based approach to measure sentiment in the posts and comments.
- **Main insights:** Average negative sentiment increased over the period. Moderate users fell into inactivity, but those who stayed radicalized.

Figure: Sentiment on Yellow Vests Facebook Pages (2018-2019)



Source: Mobilization without Consolidation: Social Media and the Yellow Vests Protests, Boyer et al. (2022)

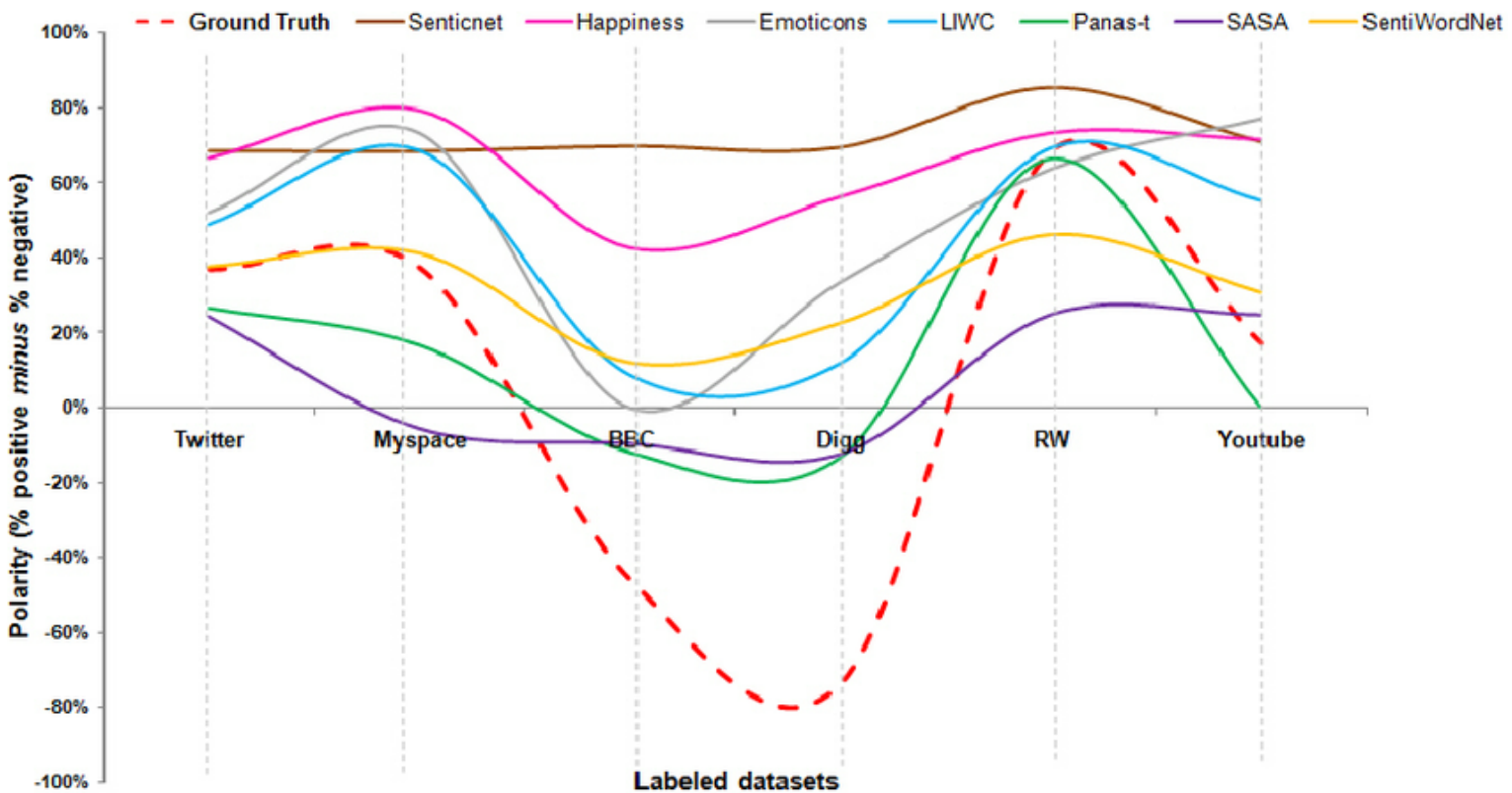
Figure: Intensive and Extensive Margins of Radicalization



Source/Notes: Mobilization without Consolidation: Social Media and the Yellow Vests Protests, Boyer et al. (2022). A simple linear decomposition suggests moderate users left the movement, and those who stayed radicalized.

Pros and Cons

- **Pros**
  - Straightforward and transparent
  - A lot of researcher control over the dictionary
- **Cons**
  - Requires domain-specific knowledge
  - Dictionaries cannot be exported easily to different contexts.
  - Predicted sentiment is sensitive to the choice of the dictionary.
  - Fails to identify irony.
  - No machine learning involved, so the model has no opportunity to discover patterns on its own.



Source/Notes: Comparing and combining sentiment analysis methods, Gonçalves et al. (2013). Polarity of the eight sentiment methods across the labeled datasets, indicating that existing methods vary widely in their agreement.

Plan

Text Regressions

- An alternative intuitive approach is a text regression:

$$Y_i = W_i' \cdot \beta + \varepsilon_i$$

- But W can be very large as the vocabulary size grows.
- A penalized regression is often required to estimate β (e.g., LASSO):

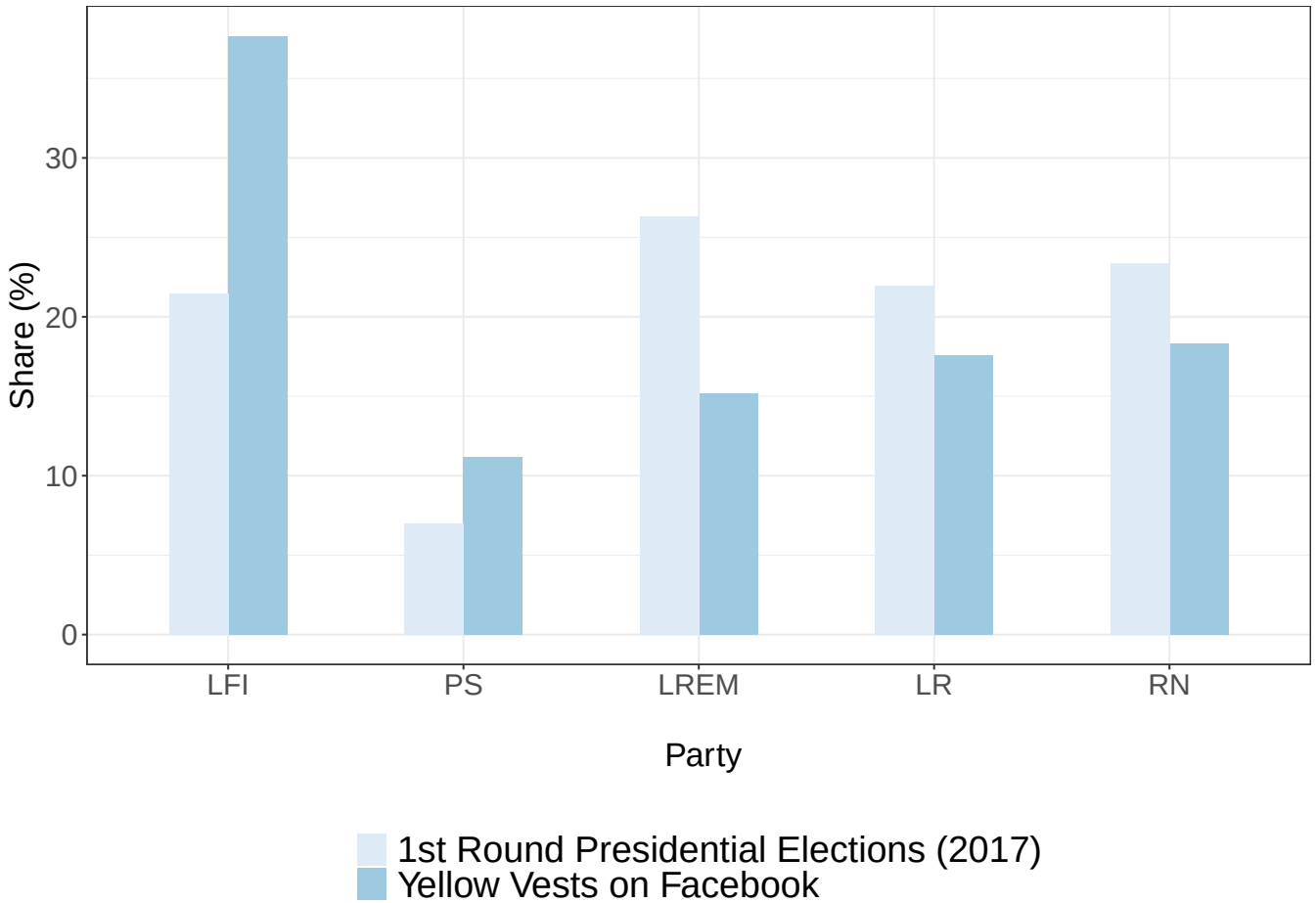
$$min_\beta \left(\sum_{i=1}^{n} Y_i - W_i' \cdot \beta \right) \text{ such that } \sum_{j=1}^{p} |\beta_j| \leq t$$

- Other common approches: random forests and support vector machines (SVMs).
- Other common transforms: logistic and multinomial logistic.

Application: What is the political color of the Yellow Vests?

- **Data:** Over 600 active Yellow Vests Facebook pages between November 2018 and April 2019 + tweets of politicians at the French parliament.
- **Method:** Train a penalized multinomial logisitic regression (LASSO) on tweets ($\approx$ 55% accuracy). Then predict the partisanship of users on the Yellows Vests Facebook pages.
- **Main insights:** Facebook users involved in the movement mainly used left-wing slanted language.

Figure: Political Partisanship on the Yellow Vests Facebook Pages



Source: Mobilization without Consolidation: Social Media and the Yellow Vests Protests, Boyer et al. (2022)

Application: Measuring Speech Polarization

- **Paper:** Peterson and Spirling (2018) measure speech polarization in the U.K. Parliament.
- **Data:** Entire Hansard record of British parliamentary debates (1935-2013)
- **Method:** Penalized Logistic regression (Ridge)
- **Main insight:** Speech polarization remains stable in the post-war period, but sharply increases with Thatcher, before decreasing in the early 21st century.

Application: Predicting Party From Speeches

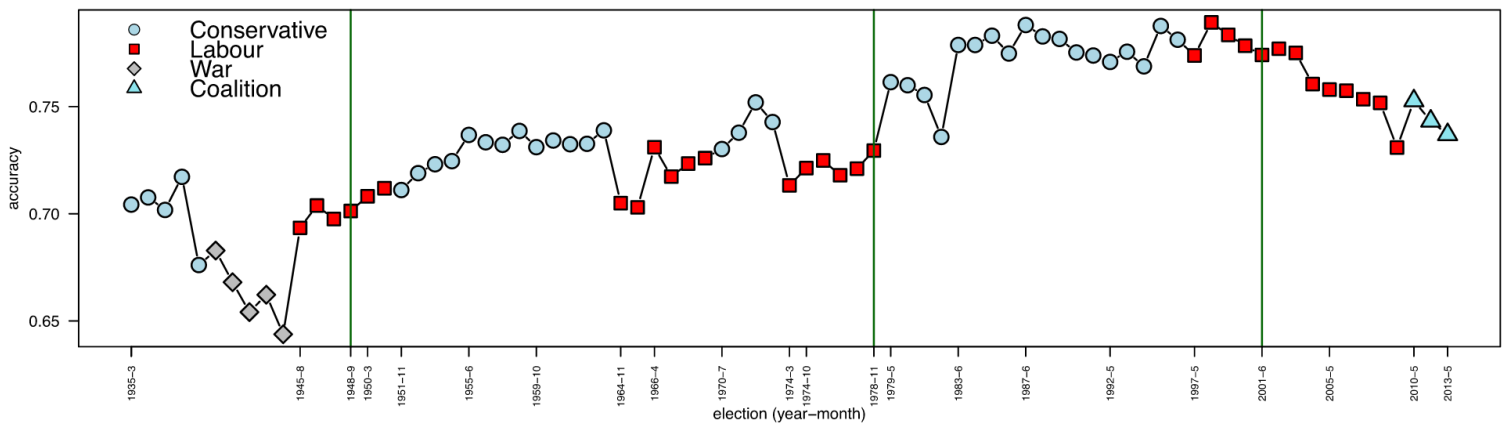


Figure 3. Estimates of parliamentary polarization, by session. Election dates mark x-axis. Estimated change points are [green] vertical lines.

Source: Classification Accuracy as a Substantive Quantity of Interest: Measuring Polarization in Westminster Systems, Peterson and Spirling, Political Analysis (2018)

Pros and Cons

- **Pros**
  - Involves machine learning, so we can learn from the data.
  - No *a priori* knowledge required.
  - Relatively robust to text cleaning and featurization (Denny and Spirling, 2018).
- **Cons**
  - Requires labeled data to train the model.

Plan

- In some cases:
  - We do not have access to labeled data.

    e.g., New unanticipated topics appear all the time in the news.
  - We want to minimize researcher priors.

    e.g., Do we really know all the topics discussed in the U.S. Congress?

Can we let the data "speak for itself"?

Latent Variable Models

- Yes, but... we need **a data-generating process**.
- In other words, we need a story of how the documents are generated.
- Latent variable models are popular for unsupervised learning tasks.
- A famous model is the LDA model for topic mining.
- Before jumping into LDA, let's cover the basics and build some intuition.

A Stylized Example

Doc 1: guns zombies biohazard win lose...
Doc 2: player lose score survival...
Doc 3: zombies survival congress guns...
Doc 4: ...
Doc 5: ...
Doc 100000: congress welfare constitution guns...

What are the topics in these documents?

A Stylized Example

Doc 1: guns zombies biohazard win lose...
Doc 2: player lose score survival...
Doc 3: zombies survival congress guns...
Doc 4: ...
Doc 5: ...
Doc 100000: congress welfare constitution guns...

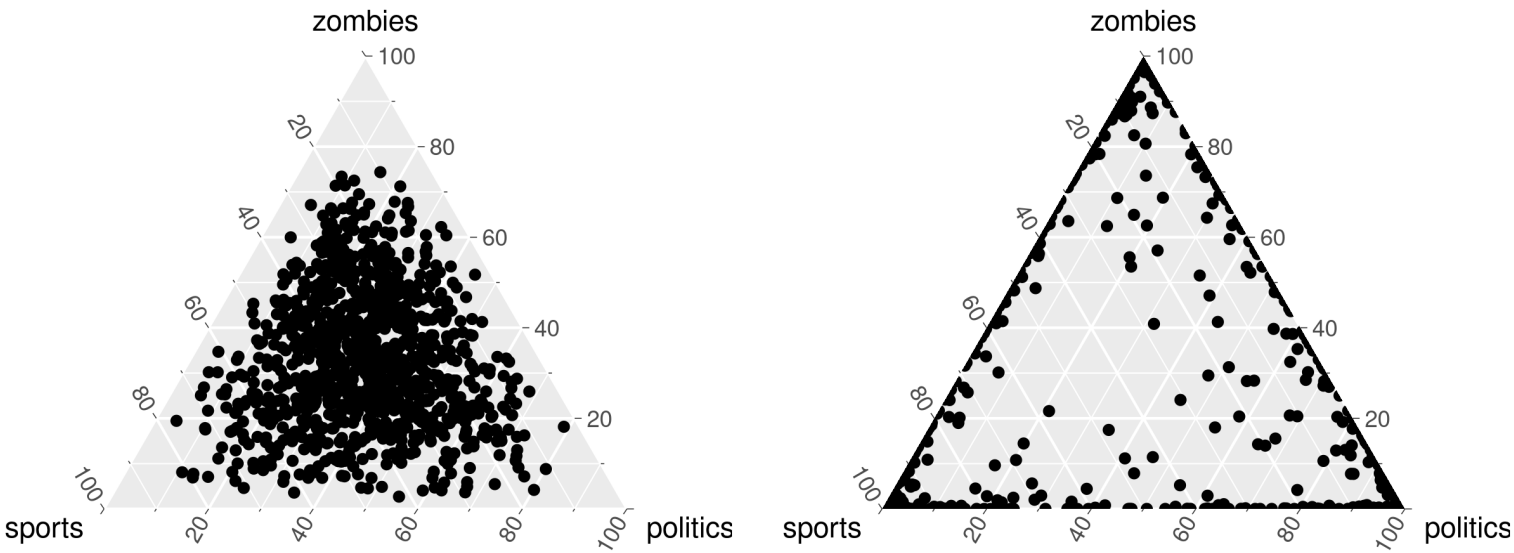
What are the topics in these documents?

Zombies: guns, zombies, biohazard, survival
Sports: player, win , score, lose
Politics: welfare, congress, constitution, guns

Can we infer these topics without specifying them a priori?

Prerequisite: Useful Distribution (I)

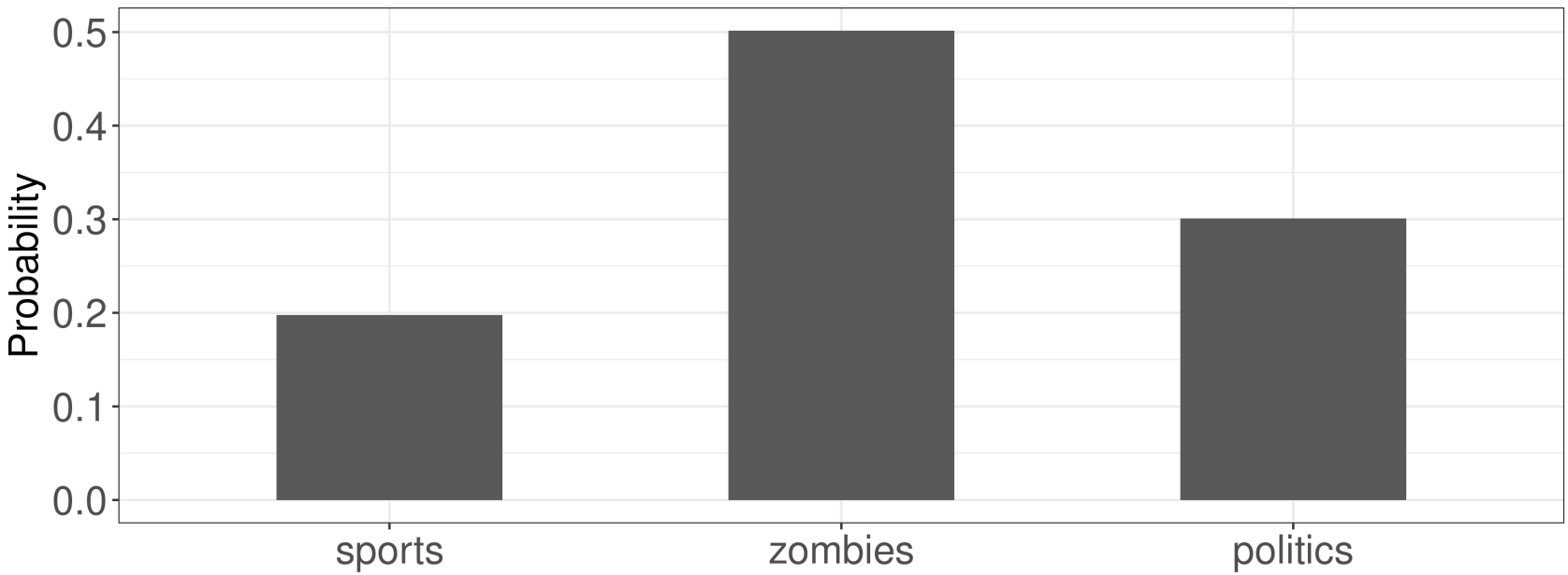
Figure: Dirichlet Distribution



Notes: $\vec{\theta} \sim \text{Dir}(\vec{\alpha})$. $\alpha \geq 1$ leads to many points in the center. $\alpha \leq 1$ leads to many points in the corners.

Prerequisite: Useful Distributions (II)

Figure: Categorical Distribution



Notes: $Z \sim \text{Cat}(0.2, 0.5, 0.3)$. Typical illustrative example is drawing colored balls out of a box. By the way, Dirichlet distributions give us the parameters of Categorical distributions!

Unigram Model

- Words from every document are drawn from a single categorical distribution.
- All words are pronounced independently from one another.
- Data-generating process
  - Distribution of words in a document: $\vec{\phi} \sim \text{Dir}(\vec{\beta})$
  - Distribution of tokens: $W \sim \text{Cat}(\vec{\phi})$

Generating Some Documents

Figure: Simulated Documents from the Unigram Model

```
Doc: constitution win zombies win guns biohazard guns survival biohazard constitution
Doc: zombies win guns zombies win win lose player win win
Doc: guns win welfare zombies guns constitution guns constitution guns welfare
Doc: win win zombies congress zombies survival zombies zombies win win
Doc: biohazard player congress win biohazard constitution congress congress score score
Doc: constitution constitution guns guns biohazard player biohazard welfare welfare constitution
Doc: score lose player score win lose survival score lose lose
Doc: win congress congress win congress congress congress win congress win
Doc: lose win constitution biohazard lose win survival player biohazard win
Doc: welfare survival zombies zombies survival win win survival survival survival
```

Notes: The model is rather uninformative. Tokens are simply more likely to be drawn if they are frequent in a given corpus. By construction, there are no topics.

Mixture of Unigrams Model

- Each document is assigned **one topic** and each topic has its own distribution over words.
- All words are pronounced independently from one another **conditional on topic assignment**.
- Data-generating process
  - Distribution of topics to documents: $\vec{\theta} \sim \text{Dir}(\vec{\alpha})$
  - Distribution of words to topics: $\vec{\phi} \sim \text{Dir}(\vec{\beta})$
  - The topic: $T \sim \text{Cat}(\vec{\theta})$
  - The tokens: $W \sim \text{Cat}(\vec{\phi})$

Generating Some Documents

Figure: Simulated Documents from the Mixture of Unigrams Model

```
Doc: guns welfare welfare welfare welfare welfare welfare welfare welfare constitution
Doc: player score score player score score score score score score
Doc: welfare welfare welfare welfare guns welfare welfare welfare welfare welfare
Doc: survival survival survival survival biohazard biohazard survival survival guns survival
Doc: welfare welfare guns congress welfare welfare welfare welfare welfare welfare
Doc: survival survival survival survival survival biohazard survival biohazard survival guns
Doc: welfare congress welfare welfare welfare guns welfare welfare guns welfare
Doc: welfare welfare welfare welfare welfare welfare welfare welfare welfare guns
Doc: biohazard biohazard biohazard survival survival biohazard biohazard survival survival biohazard
Doc: welfare welfare welfare welfare welfare welfare guns welfare welfare welfare
```

Notes: We now have topics to infer. However, contrary to our original corpus, there is only one topic per document.

Latent Dirichlet Allocation (LDA)

- More often than not documents are composed of **multiple topics**.
- LDA assumes that:
  - Each document is a **distribution over topics**.
  - Each topic is a **distribution over words**.
  - Documents refer to only a relatively **small number of topics**.
  - Topics use a relatively **small number of words**.

Latent Dirichlet Allocation (LDA)

- Data-generating process
  - Distribution of topics for each document: $\vec{\phi} \sim \text{Dir}(\vec{\alpha})$
  - Distribution of words for each topic: $\vec{\phi} \sim \text{Dir}(\vec{\beta})$
  - The topic for each token: $T \sim \text{Cat}(\vec{\theta})$
  - The token: $W \sim \text{Cat}(\vec{\phi})$

Generating Some Documents

Figure: Simulated Documents from the LDA Model

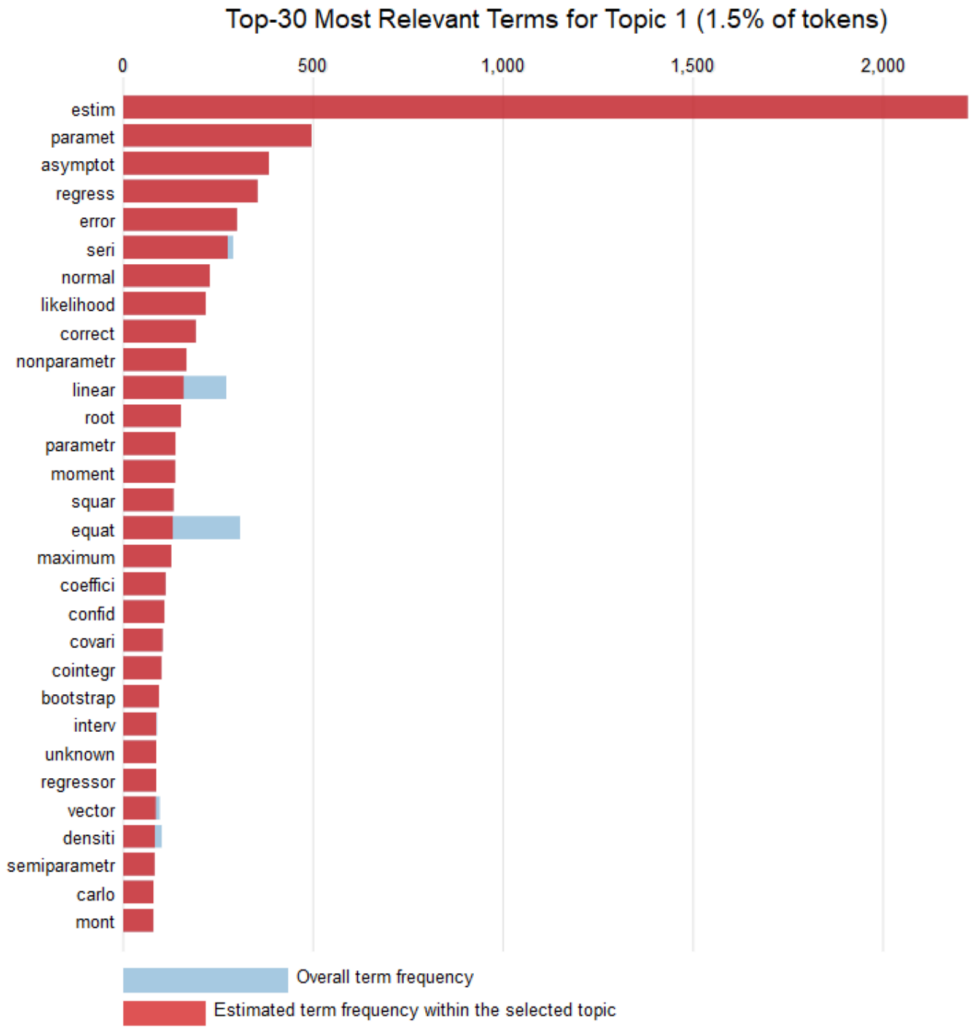
```
Doc: welfare guns constitution welfare congress welfare guns win welfare welfare
Doc: score score win guns score welfare win congress win lose
Doc: score win win score score win score score score score
Doc: win win welfare score score score score win score score
Doc: survival guns survival welfare biohazard survival survival congress zombies welfare
Doc: survival guns biohazard biohazard guns biohazard guns survival biohazard biohazard
Doc: player win score score constitution score welfare welfare score welfare
Doc: guns guns biohazard guns survival guns guns guns biohazard biohazard
Doc: zombies guns congress biohazard guns guns biohazard survival biohazard guns
Doc: guns guns survival biohazard score welfare guns score score congress
```

Notes: The simulated documents are a mixture of topics. As a result, they are closer to our original corpus.

Estimation / Statistical Inference

- In practice, Dirichlet distributions are **priors**.
  - Priors capture our beliefs/knowledge and are hyperparameters (i.e., they are not estimated).
  - Priors are useful for text in so far as we know quite some things about language (*e.g., topic distributions over words are likely skewed*).
  - As the sample size increases, priors become less relevant, but they can largely improve results for small sample sizes (if specified correctly).
- We estimate the Categorical distributions in these models.
- Many methods exist to estimate Bayesian models: MCMC, Variational Bayes, Gibbs sampling, EM algorithm, etc.

Visualizing Results of a Topic Model



Source/Notes: My own dusty master's thesis. Similar graphs can be obtained for the topic-shares within a document.

Pros and Cons

- **Pros**
  - In practice, LDA is very easy to implement.
  - Many variants of LDA that relax some assumptions.

    e.g., Correlated topic model, Dynamic topic model, Bi-term topic model, Structural topic model, etc.
- **Cons**
  - LDA does not really model "topics" but distributions. It is simply a dimension reduction task. Human labeling of resulting distributions can be tricky.
  - The number of topics is the main hyperparameter and may influence results downstream.
  - Some data-driven approaches exist to choose the "optimal" number of clusters but careful human inspection remains the safest bet.

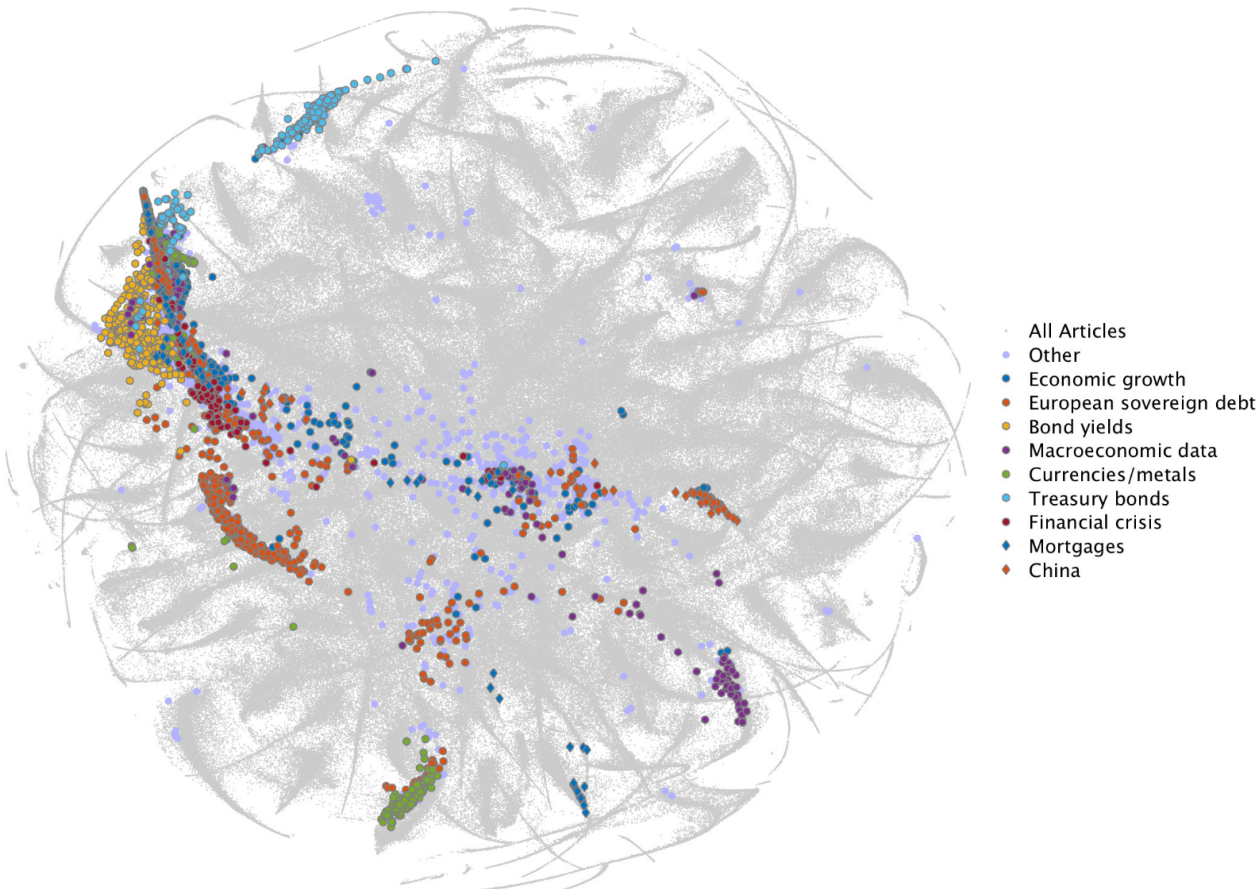
"We refer to the latent multinomial variables in the LDA model as topics, so as to exploit text-oriented intuitions, but we make no epistemological claims regarding these latent variables beyond their utility in representing probability distributions on sets of words."

(Blei et al., 2003)

Application: Predicting Macroeconomic Variables

- **Paper:** Business News and Business Cycles, Bybee et al. (2022)
- **Data:** Wall Street Journal text articles
- **Method:** Latent Dirichlet Allocation combined with a penalized regression
- **Main insight:** A topic model trained on business news closely tracks a broad range of macroeconomic variables.

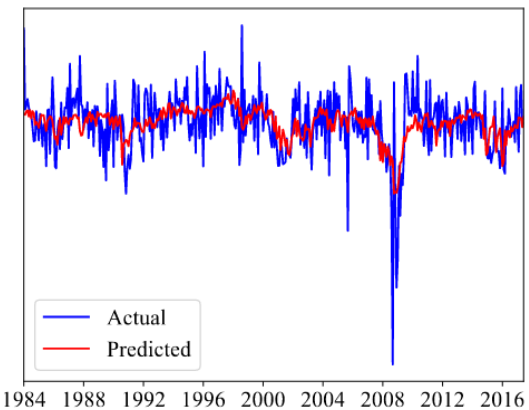
Figure 14: Articles Featuring the Federal Reserve



Source: Business News and Business Cycles, Bybee et al. (2022)

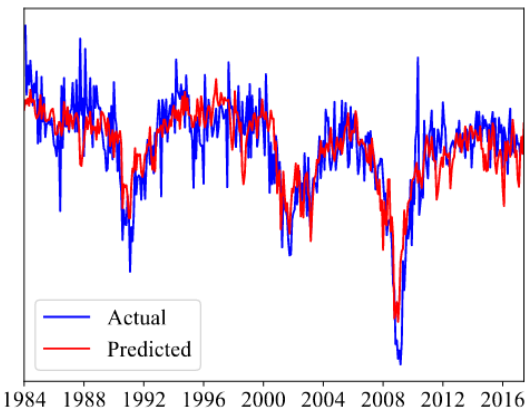
Industrial Production Growth

Topic	Coeff.	p-val.
Recession	-0.38	0.00
Oil market	-0.17	0.00
Southeast Asia	0.11	0.10
Health insurance	0.06	0.93
Clintons	0.03	0.40
In-Sample R^2	0.21	
Out-of-Sample R^2	0.06	



Employment Growth

Topic	Coeff.	p-val.
Recession	-0.61	0.00
Rail/trucking/shipping	0.22	0.01
Bush/Obama/Trump	-0.15	0.09
Iraq	-0.14	0.01
Clintons	0.12	0.01
In-Sample R^2	0.59	
Out-of-Sample R^2	0.48	

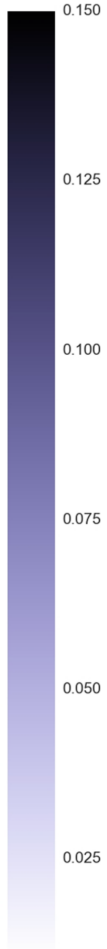


Source: Business News and Business Cycles, Bybee et al. (2022)

Application: Federal Reserve Bank Transcripts

- **Paper:** Hansen et al. (2017) study how transparency affect monetary policy-makers' decisions.
- **Data:** Federal Open Market Committee (FOMC) transcripts
- **Method:** Latent Dirichlet Allocation with 50 topics
- **Main insight:** Using a difference-in-differences design, they empirically validate that transparency has a positive disciplining effect and a negative conformity effect. The positive effect dominates.

													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



Source: Transparency and Deliberation Within the FOMC: A Computational Linguistics Approach, Hansen et al. (2017).

- Many more latent variable models exist.
- For example, **latent ideology models** project word usage differences on K dimensions (interpreted as dimensions of ideology).
- With some practice you can write your own!

Figure: Text-Based Ideal Points for U.S. Congress Members

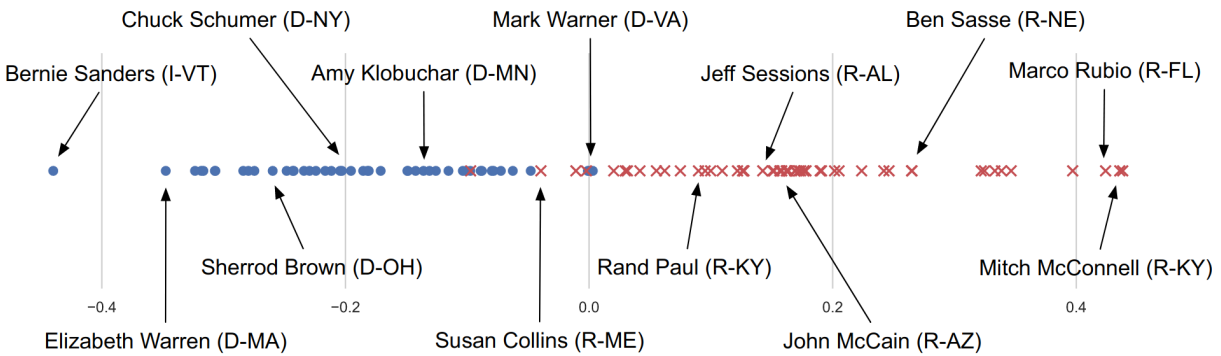


Figure 1. The text-based ideal point model (TBIP) separates senators by political party using only speeches. The algorithm does not have access to party information, but senators are coded by their political party for clarity (Democrats in blue circles, Republicans in red x's). The speeches are from the 114th U.S. Senate.

Source: Text-Based Ideal Points, Vafa et al. (2020)

Plan

- Easy, transparent, the bag-of-words approach has led to tens of thousands of publications.
- It is still widely used in the social sciences.
- Nonetheless, it also comes with clear limitations.

What does the bag-of-words approach discard?

- Basically disregards **word ordering, grammar, and syntax**.

  Example 1

  "The terrorists killed American soldiers."

  "American soldiers killed the terrorists."

  $\rightarrow$ These two sentences have the same representation!

  Example 2

  "This is a sentient being."

  "I love being the way I am."

  $\rightarrow$ "being" has different meanings and grammatical functions, but the same representation.

- No notion of **distance** between tokens.

  For example, *"hi"* and *"hello"* are not considered more similar to one another than to *"sociology"*.

  In fact, tokens are all at the same distance of one another.

- In other words, tokens are **discrete** features.

  Once again, *"hi"* and *"hello"* are completely distinct features for predicting whether a message is greeting somebody.

  "sociology" and *"economics"* are completely distinct features for predicting whether a message is about the social sciences.

What's next?

- In the next course, we will move beyond the bag-of-words approach.
- We will discuss text embeddings that capture:
  - Contextual information
  - Distance between tokens

Thanks for listening!

Text as Data for the Social Sciences (I)

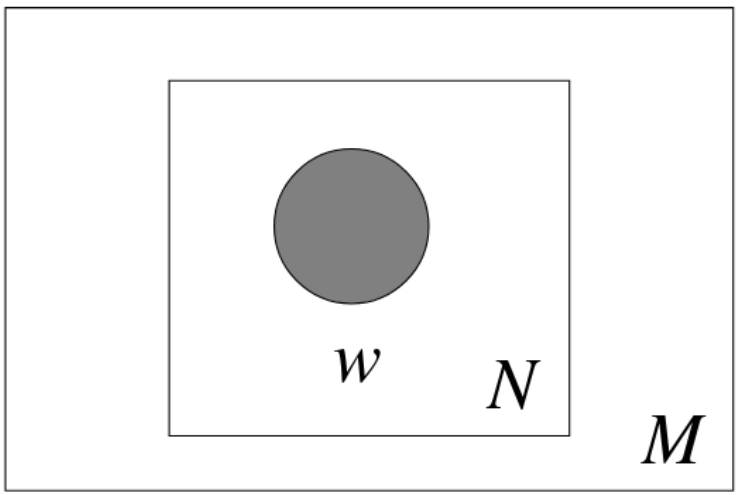
Germain Gauthier, Felix Lennert, Etienne Ollion

germain.gauthier@polytechnique.edu

SICSS Paris

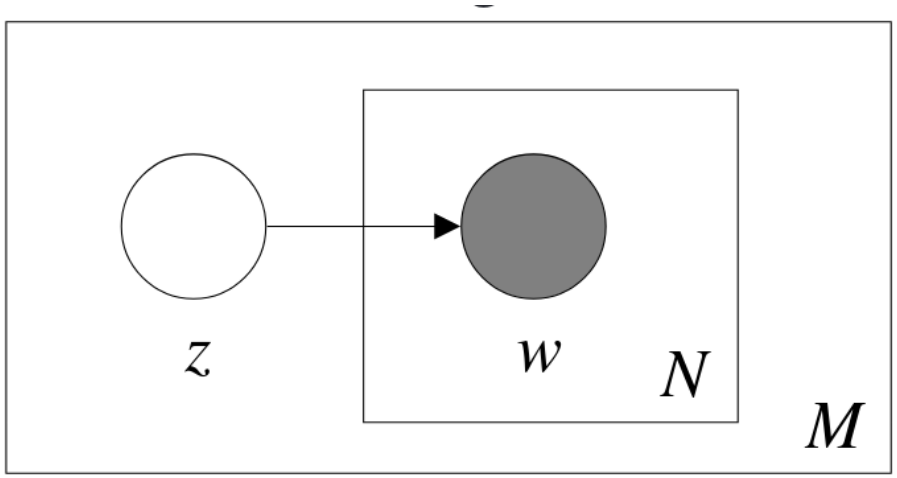
June 2022

Graphical Model – Unigram Model



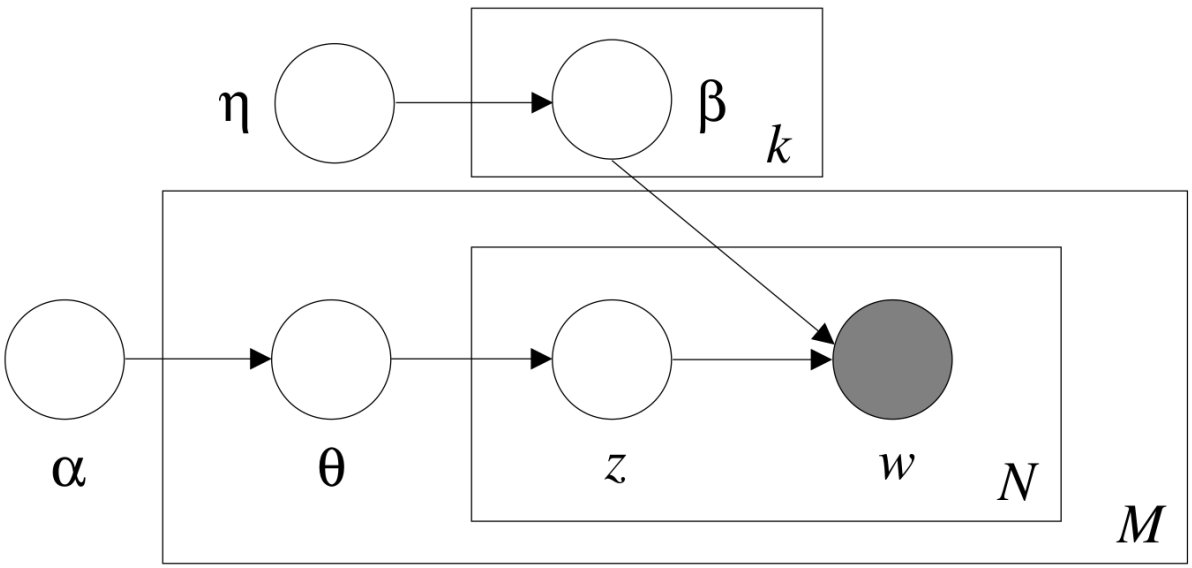
(a) unigram

Graphical Model – Mixture of Unigrams Model



(b) mixture of unigrams

Graphical Model – LDA



Bayesian Statistics

- Bayesian statistics are a set of useful methods to incorporate priors (e.g., beliefs or expert knowledge) into inference problems.
- This is useful for text in so far as we know quite some things about language.
- As the sample size increases, the prior becomes irrelevant, but it can largely improve results for small sample sizes (if specified correctly).
- Given a prior belief that a probability distribution function is $p(\theta)$ and that the observations x have a likelihood $p(x|\theta)$, then the posterior probability is defined as

$$\underbrace{p(\theta|x)}_{\text{posterior probability}} \propto \underbrace{p(x|\theta)}_{\text{likelihood}} \underbrace{p(\theta)}_{\text{prior}}$$