Text as Data

Classic Approaches to Quantitative Content Analysis

SICSS

CREST

INSTITUT POLYTECHNIQUE DE PARIS

Introduction

- Reading, and more « distant reading »
- An old endeavor: from the Bible index to content analysis
- An endeavor renewed by the digitization of everyday life.

SICSS CREST INSTITUT POLYTECHNIQUE DE PARIS

Introduction

- ‘Text as data’, ‘Quantitative Content Analysis’, ‘Modeling Text’
- From text, one wants to extract features
  - i) Give it a mathematical representation
  - ii) Apply a statistical method*
  - * from counting words to Transformers
- TaD: The return of a Maverick Method
  - And Old Endeavor
  - Many attempts, no consensus
  - A recent return into favors (AI)

An Overview of Methods

An overview that is necessarily

- Subjective
- Incomplete
- To be continued

Organized by ‘families of methods’

- Lexical statistics
- Dictionary-based methods
- Stylistic Analysis
- Semantic Networks
- Topic Models

An Overview of Methods

Builds on existing reviews:

- Grimmer & Stewart, 2013 [PoliSci]
- Evans & Aceves, 2017 [Soc]
- Gentzkow, Kelly & Taddy, 2017 [Econ]
- Cointet & Parasie, 2018 [Soc] **

Lexical Statistics

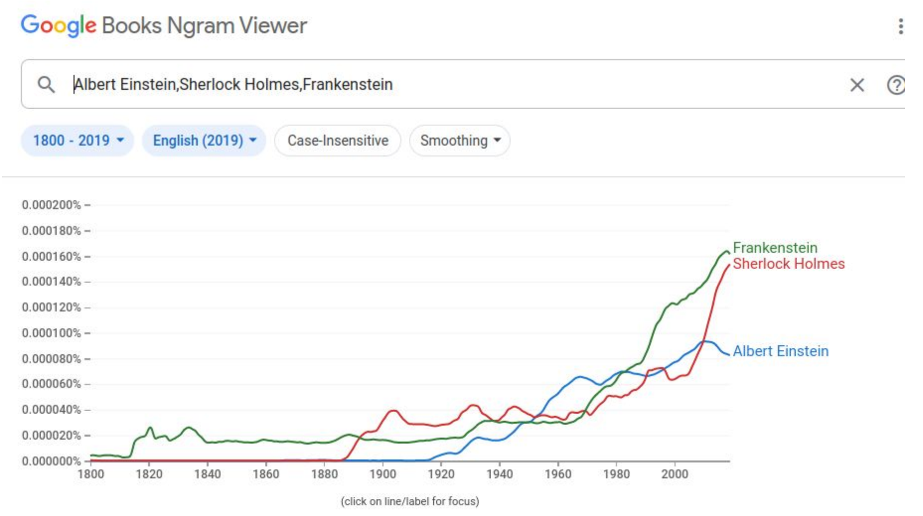
Oldest endeavor, very different options

- **Berelson & Lazarsfeld (1948)**, and before them Weber 1913.
  - A question : role of the media in the shaping of mentalities
  - Method: Counting salient words
  - Intuition: Words capture meaning.
  - Matters because: language contrues our representations (Sapir-Whorf hypothesis)

Lexical Statistics

Oldest endeavor, different options

- [Culturomics](#) (Michel et al., 2011).
  - « Quantitative analysis of culture using millions of digitized books »
  - From Google book archives to the reinvention of the social sciences



Lexical Statistics

Oldest endeavor, very different options

- **BankSpeak** (Pestre & Moretti, 2015)
  - An analysis of the langage in World Bank Reports over 40 years.
  - « Behind this façade of uniformity, a major metamorphosis has taken place »
  - Change in the semantics, from *plainspeak* to *bankspeak*.

Lexical Statistics

Oldest endeavor, very different options

Classic, but criticized

- Purely descriptive
- What about synonyms ?
- Un-natural hypotheses about language
- > Lack of structure, of context (Guerrini 2011)

Dictionary-based Methods

Old idea too (Stone et al., 1966)

Revival in the 2000s. Partly due to commercial interest (Pang *et al.*, 2000 ; Pang & Lee, 2008)

Not a focus on words, but on broader categories the word refers to.

Ex. (global warming, CO^2, greenhouse gas,...) → **climate**

Dictionary-based Methods

Most classic example: **sentiment analysis**

- Determine a sentiment score for a sentence/doc

- Based on certain pre-determined terms denoting positive or negative sentiments
  - > O'Connor *et al.*, 2010: Polls for Obama & Sentiments in Tweets
  - > Tetlock 2007: Sentiment in the *WSJ*

Dictionary-based Methods

Most classic example: **sentiment analysis**

- Flores, Anti-Immigrant Sentiment, *AJS* 2017
  - Does the passing of the law influence public opinion, and if yes, how ?
  - Tweets in Arizona in 2010 after the passing of a restrictive law. Control with Nevada.
- Advanced Sentiment Analysis
  - Scores gradually (from -4 to 4)
  - Distinguishes subject of message
  - Controls for # of active twitter accounts
    > Feeds into regression models

Dictionary-based Methods

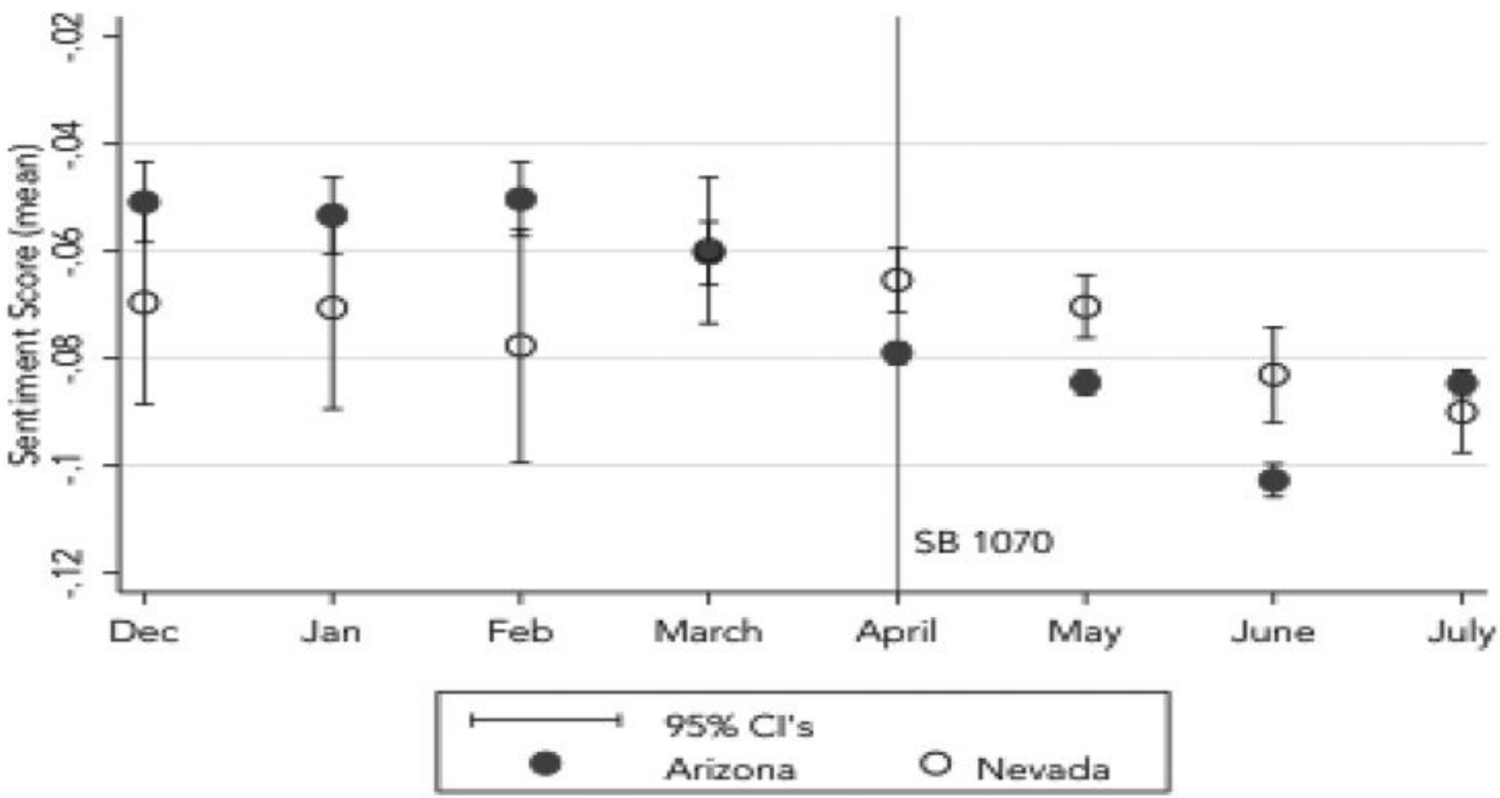


FIG. 3.—Average sentiment score of tweets about immigrants. The vertical lines represent 95% confidence intervals. The vertical line on April 2010 indicates when the Arizona governor approved SB 1070.

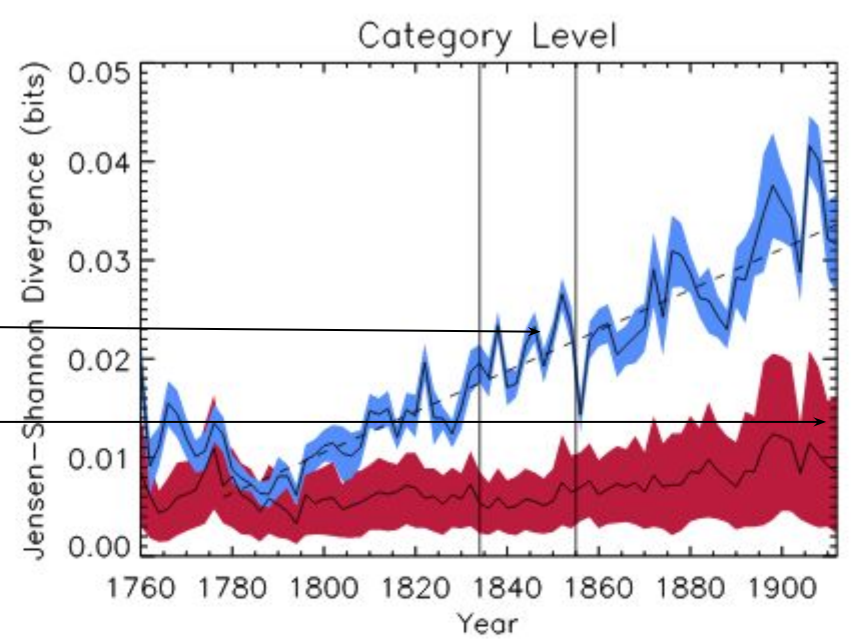
Dictionary-based Methods

But other uses are possible

- **Klingenstein et al.**, *PNAS*, 2014
- How did the judicial vocabulary evolve from 1760 to 1910 ?
  > Invention of the « violent crime » as a judicial category.
- Dictionary-based. Roget Thesa

Jensen-Shannon Divergence for violent vs. non-violent

Null hypothesis: random assignment



Dictionary-based Methods

Known issues

- Better than lexical statistics because more than a word taken into account

- Still no interest in the structure = **bag of words hypothesis**

- Problems of double meaning (‘a <u>formidable</u> regression’), of negation (‘climate change does not exist’)

- Like other methods, does not deal well with irony, second degree, metaphors (Bosco et al., 2013)

Stylistic analysis

Not so frequent but full of potential

- Idea: focus on the « style » (use of langage, deviations from norms) to investigate formality, complexity, politeness, etc.

Stylistic analysis

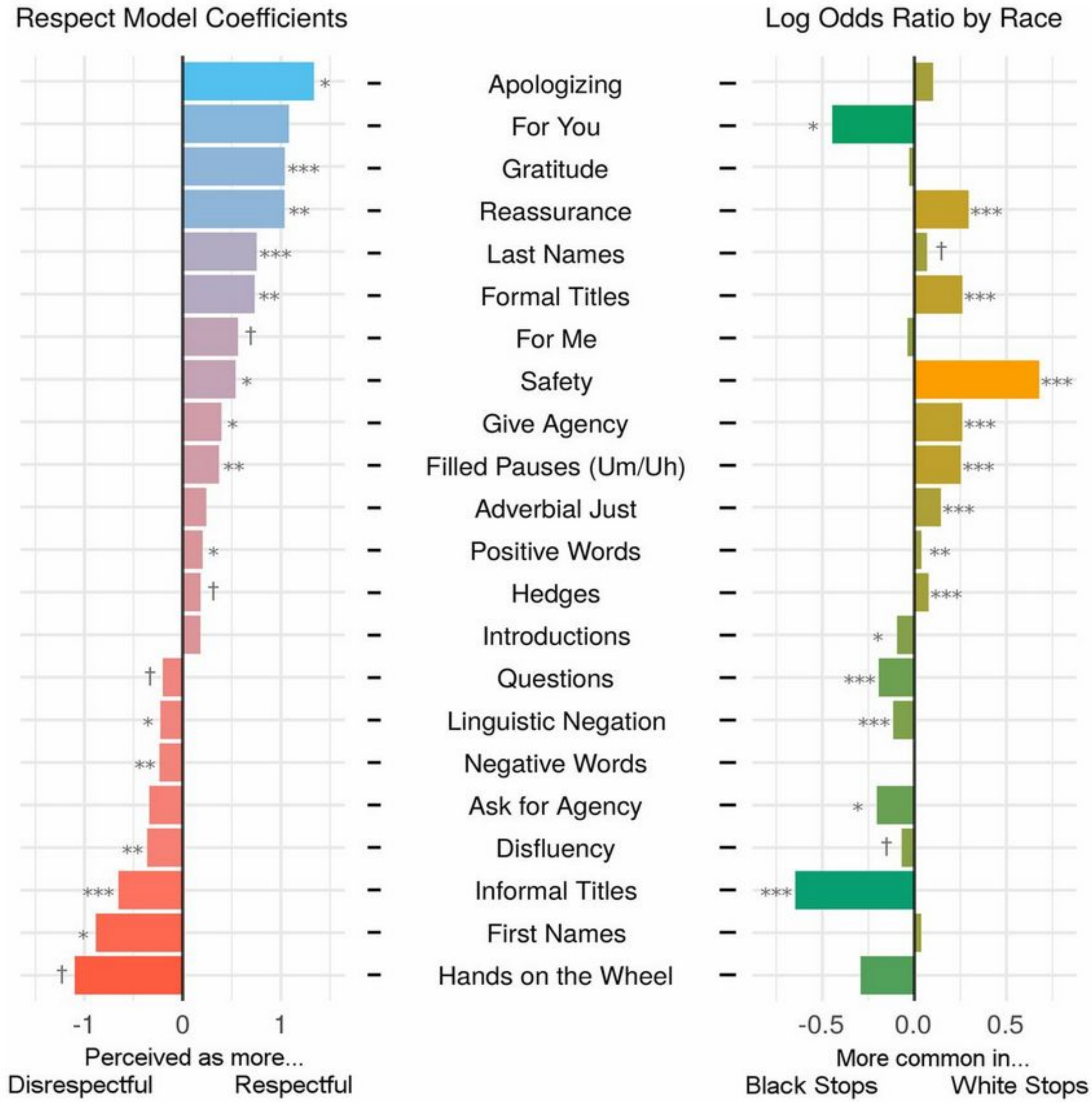
Not so frequent but full of potential

- Voigt et al., *PNAS*, 2017
  - Are police/citizen interactions racialized?
  - Using information from body camera footages.
  - > Analyzing officers' language during vehicle stops of white and black community members.
  - > Controls by place, race of officer, type of suspected infraction, time of the day.

Stylistic analysis

Example	Respect Score
First Name Ask For Agency Questions [name], can I see that driver's license again? It- it's showing suspended. Is that- that's you? Disfluency Negative Word Disfluency	-1.07
Informal Title Ask For Agency Adverbial "Just" All right, my man. Do me a favor. Just keep your hands on the steering wheel real quick. "Hands On The Wheel"	-0.51
Apology Introduction Last Name Sorry to stop you. My name's Officer [name] with the Police Department.	0.84
Formal Title Safety Please There you go, ma'am. Drive safe, please.	1.21
Adverbial "Just" Filled Pause Reassurance It just says that, uh, you've fixed it. No problem. Thank you very much, sir. Gratitude Formal Title	2.07

Stylistic analysis



Stylistic analysis

Full of promise but

- Requires good knowledge in stylistics
- Annotation can be very painful
- (Possibly outsourced to a supervised classifier ?)

Semantic networks

(and other graph-based methods)

- Stems from complex network theory (Barabasi)
- A star method in the 2000s, to circumvent the problem of structure
- Varied uses

Semantic networks

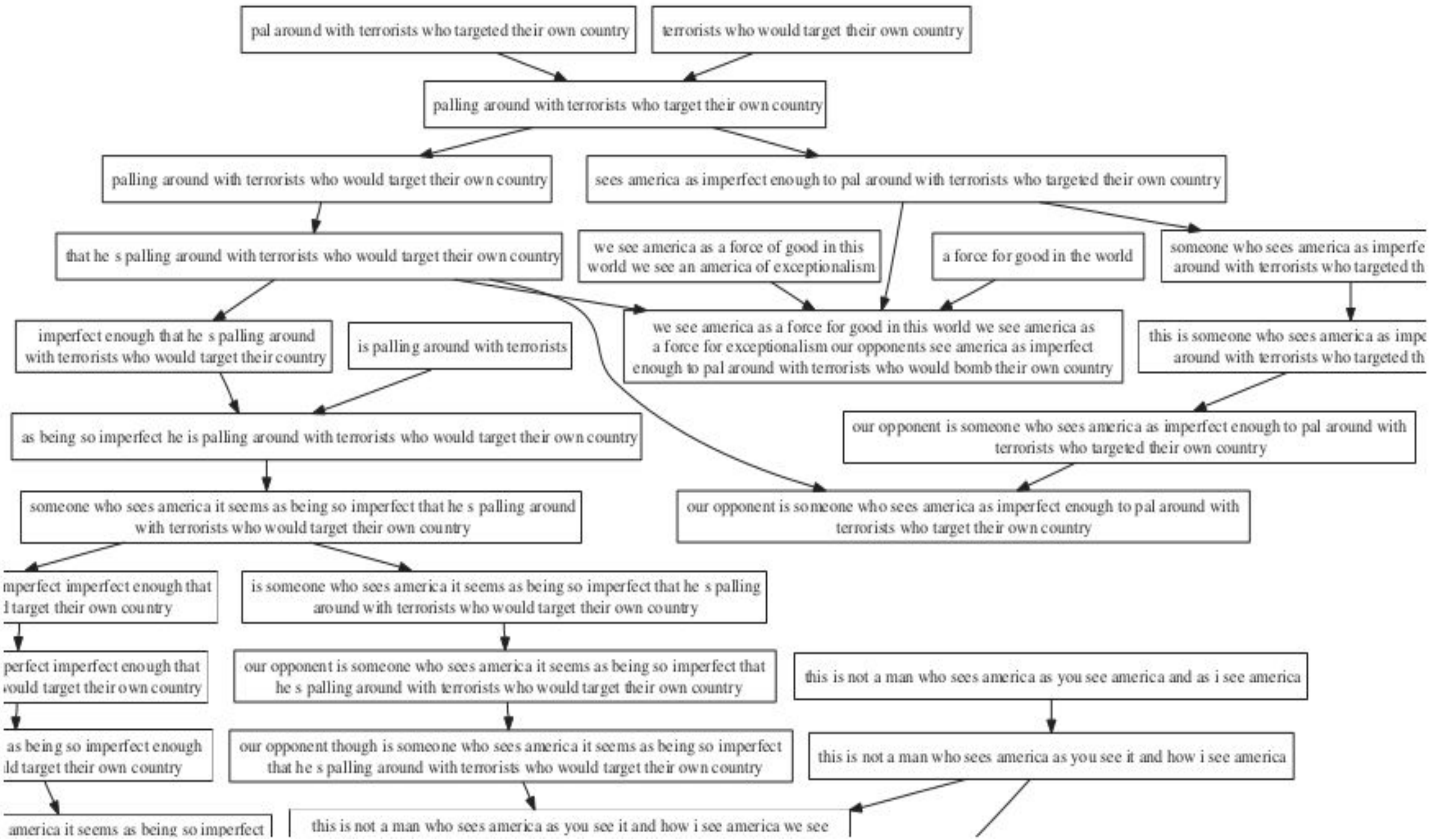
(and other graph-based methods)

- **Leskovec *et al.***, 2009 on *meme dissemination*
  - Which are the most salient quotes in the 2008 campaign ?
  - > A sentence is uttered by a politician
  - > Newsmedia echo it.

- Problem: never the same, and indirect speech.

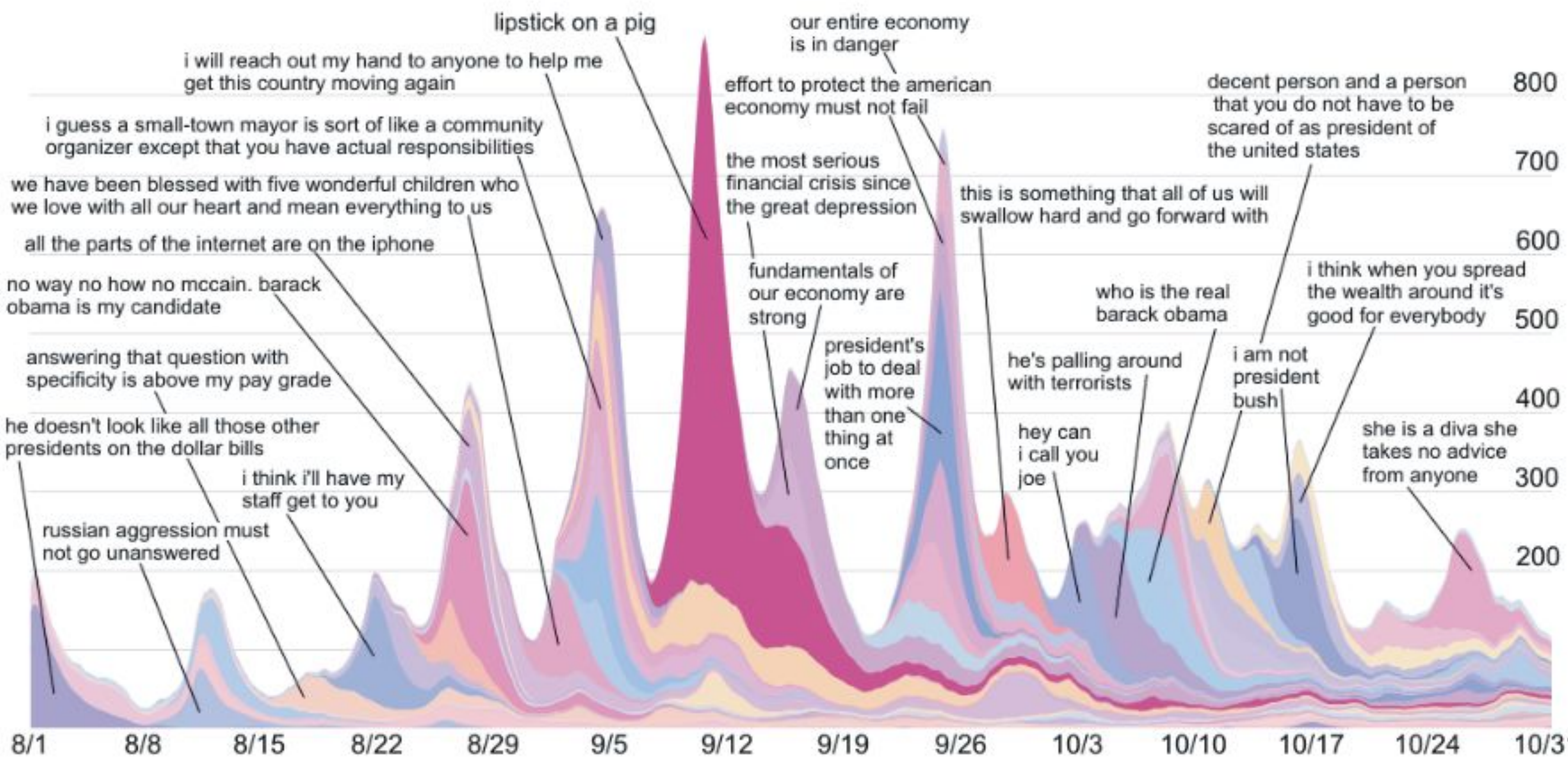
Semantic networks

(and other graph-based methods)



Semantic networks

(and other graph-based methods)



Semantic networks

(and other graph-based methods)

- Takes into account the context, somewhat
- But remain limited to certain words, phrases

Topic models

How to classify themes over a large number of texts ?

- Dictionary-based methods are an option
- Topic models is their **unsupervised counterpart**
  - Unsupervised: opposed to supervised
  - ‘A machine proposes a clustering, which is subsequently interpreted by the scientist’
  - When there is no established coding scheme, nor have we cues to do the classification.

Topic models

How to classify themes over a large number of texts ?

- Topic Models: Blei, circa 2003.
  - Inductively capture clusters of words that co-occur over documents.
  - > « uncover underlying semantic regularities in a set of documents by mapping recurring relationships between words ».
  - Output: a series of « themes » (sets of co-occurring words)

Topic models

In more details

We **assume** there are **K** topics in **n** documents

We want to determine what is the proportion of each topic $K_{1,\ldots,i}$,in each document $n_{1,\ldots,n}$, in a proportion **α** (0<α<1)

Ex. Article 1 is mostly about Economics (k=1, α=.6), a bit about Politics (k=2, α=.2), and not a all about Sport (k=5, α=0).

Topic models

In more details

Most classic method: Latent Dirichlet Allocation (LDA)

See original paper by (Blei, Ng & Jordan, 2003)

i) Assume each word pertains to a topic **k**

ii) For each word **w** in doc n, assume its topic k is wrong but every other word is assigned the correct topic

iii) Assign word **w** to a given topic based on

- what topics are in document **n**
- how many times has **w** been assigned to a particular topic

And repeat

Topic models

How to classify themes over a large number of texts ?

- Many examples in the social sciences
  > Fligstein *et al.*, *ASR*, 2017

-

- Why did the Fed did not foresee the 2008 crisis ?
  ⇒ (macro) frames and confirmation bias

- 72 FOMC meetings, basic LDA on those documents

Topic models

How to classify themes over a large number of texts ?

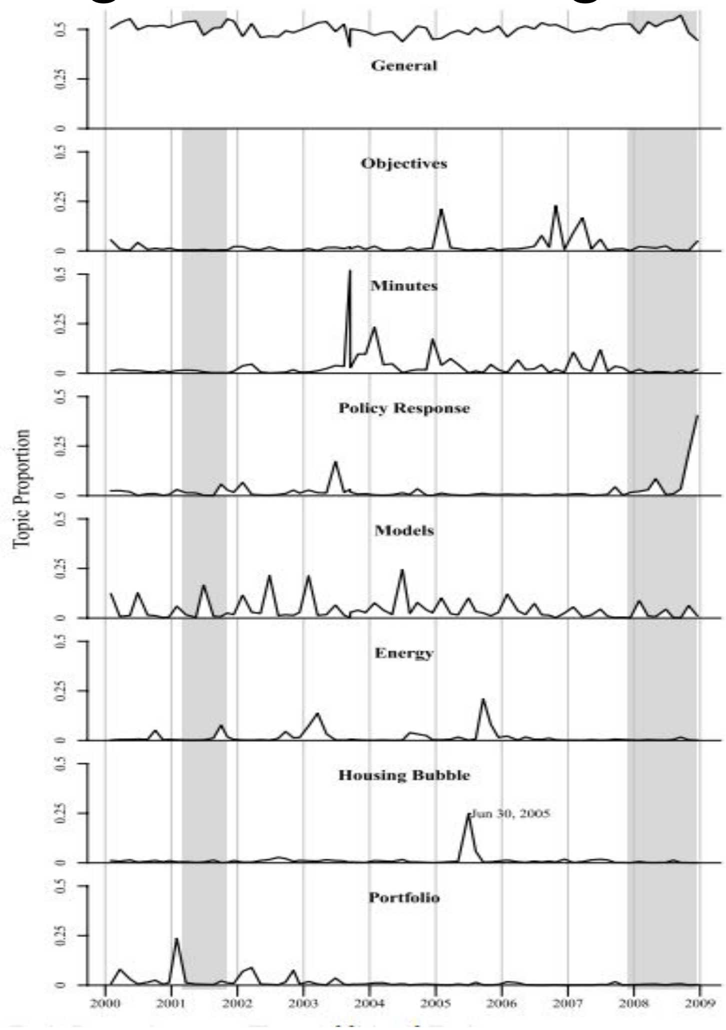
- Fligstein, Brundage, Schultz 2017

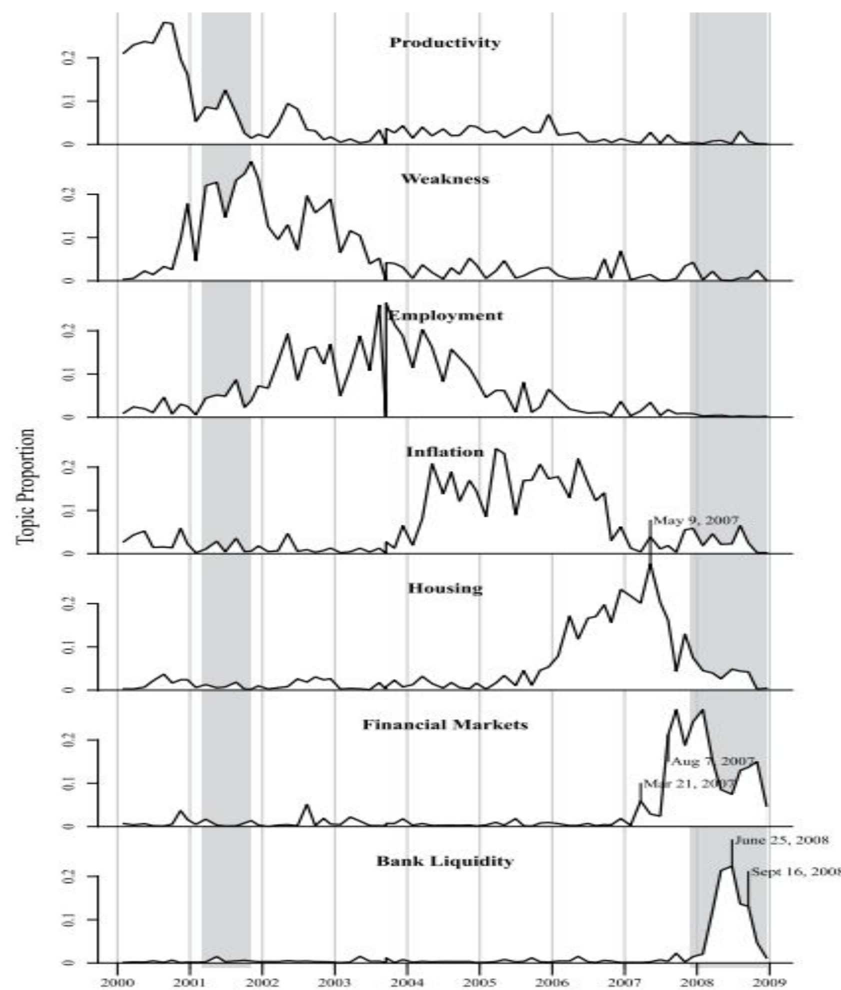
Portfolio	Housing Bubble	Energy	Models	Policy Response	Minutes	Objectives	General
contingency	arms	disruptions	depreciation	treasury	formulaic	Congress	see
Mae	lenders	gas	red	banks	think	horizon	like
window	loans	inertia	present	facility	announcement	adopt	economy
collateral	constant	storm	year	guarantee	meeting	achieve	well
outright	quality	impact	dollar	interest	expediting	anchored	don
sovereign	value	refinery	exports	purchases	vote	public	may
issue	afford	ports	top	ceiling	information	benefits	much
discount	family	barrel	foreign	quantity	process	committee	can
system	ofheo	crude	simulations	effect	decision	run	even
liquidity	percentile	energy	variables	deflation	view	definition	get
Lombard	nonmarket	heating	account	tools	memo	regime	chairman
Freddie	appreciation	effect	bars	excess	oni	prices	say
debt	bond	Venezuela	Taylor	money	give	defined	risk
operations	component	million	structural	policy	issue	specific	because
tally	Francisco	aftermath	rate	fomc	editing	think	look
gnmas	misallocation	stagflation	unemployment	size	use	cpi	come
diversified	shown	damage	productivity	monetary	transparency	consensus	know
disclose	bias	inertial	different	desk	press	diversity	next
Fannie	city	coast	show	alternative	convey	transparency	policy

Topic models

How to classify themes over a large number of texts ?

- Fligstein, Brundage, Schultz, *ASR*, 2017





Topic models

How to classify themes over a large number of texts ?

- Many examples in the social sciences

- Classic criticism: « exploratory analysis » (see Grimmer & Stewart 2013).
  - Problems in long time series with change in meaning of words.
  - Necessary *post hoc* interpretation
  - No good validation criterion.
  - Remain at the level of the word

See: A. Shadrova, ‘Topic models do not model topics’, 2021

Summary

- A wealth of methods
- ...to be used depending on your needs

- Keep in mind that all of these methods rely on very un-natural conceptions of what language is.
  - Almost all are premised on the « bag of word » hypothesis.
  - Arguably, all models are wrong but some are useful.

STILL

Summary

Time flies like an arrow.

But fruit flies like a banana (not an arrow)

⇒ Need to go towards a more realistic description of language

This is what the recent developments in AI promise

References

Berelson, Bernard. 1952: *Content analysis in communication research*. New York: Free Press.
Blei, David, Andrew Ng, and Michael Jordan. 2003. “Latent Dirichlet Allocation.” *Journal of Machine Learning Research* 3:993–1022.
Bosco, Cristina, Viviana Patti, and Andrea Bolioli. 2013. “Developing Corpora for Sentiment Analysis: The Case of Irony and Senti–TUT (Extended Abstract).” *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence (IJCAI 2015)*:4158–62.
Cointet, Jean-Philippe and Sylvain Parasie. 2018. “Ce que le big data fait à l’analyse sociologique des textes: Un panorama critique des recherches contemporaines.” *Revue française de sociologie* 59(3):533.
Evans, James A. and Pedro Aceves. 2016. “Machine Translation: Mining Text for Social Theory.” *Annual Review of Sociology* 42(1):21–50.
Fligstein, Neil, Jonah Stuart Brundage, and Michael Schultz. 2017. “Seeing Like the Fed: Culture, Cognition, and Framing in the Failure to Anticipate the Financial Crisis of 2008.” *American Sociological Review* 82(5): 879–909.
Flores, René. 2017. “Do Anti-Immigrant Laws Shape Public Sentiment? A Study of Arizona’s SB 1070 Using Twitter Data.” *American Journal of Sociology* 123(2):333–84.
Gentzkow, Matthew, Bryan Kelly, and Matt Taddy. 2019. "Text as Data." *Journal of Economic Literature* 57(3): 535-74.
Grimmer, Justin and Brandon M. Stewart. 2013. “Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts.” Political Analysis 21(3):267–97.
Klingenstein, Sara, Tim Hitchcock, and Simon DeDeo. 2014. “The civilizing process in London’s Old Bailey.“ *Proceedings of the National Academy of Sciences* 111(26):9419–24.
Leskovec, Jure, Lars Backstrom, and Jon Kleinberg. 2009. “Meme-tracking and the dynamics of the news cycle.” In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD ‘09)*:497–506.
Michel, Jean-Baptiste, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, The Google Books Team, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak, and Erez Lieberman Aiden. 2011. “Quantitative Analysis of Culture Using Millions of Digitized Books.” Science 331(6014):176–82.
Moretti, Franco, and Dominique Pestre. 2015. “Bankspeak. The Language of World Bank Reports.” *New Left Review* 92: 75–99.
O’Connor, Brendan, Ramnath Balasubramanyan, Bryan R. Routledge, Noah A. Smith. 2010. “From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series.” In *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*:122–9.
Pang, Bo, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? Sentiment Classification using Machine Learning Techniques. In *Proc. 2002 Conf. on Empirical Methods in Natural Language Processing (EMNLP)*:79–86.
Pang, Bo and Lillian Lee. 2008. “Opinion Mining and Sentiment Analysis.” Foundations and Trends in Information Retrieval 2:1-135.
Shadrova, Anna. 2021. “Topic models do not model topics: epistemological remarks and steps towards best practices.” *Journal of Data Mining and Digital Humanities*.
Stone, Philip J., Robert F. Bales, J. Zvi Namenwirth, and Daniel M. Ogilvie. 1962. “The General Inquirer: A Computer System for Content Analysis and Retrieval Based on the Sentence as a Unit of Information.” Behavioral Science 7(4):484–98.
Tetlock, Paul C. 2007. “Giving Content to Investor Sentiment: The Role of Media in the Stock Market.” The Journal of Finance 62(3):1139–68.
Voigt, Rob, et al. 2017. Language from police body camera footage shows racial disparities in officer respect. *Proceedings of the National Academy of Sciences* 114(25):6521–26.