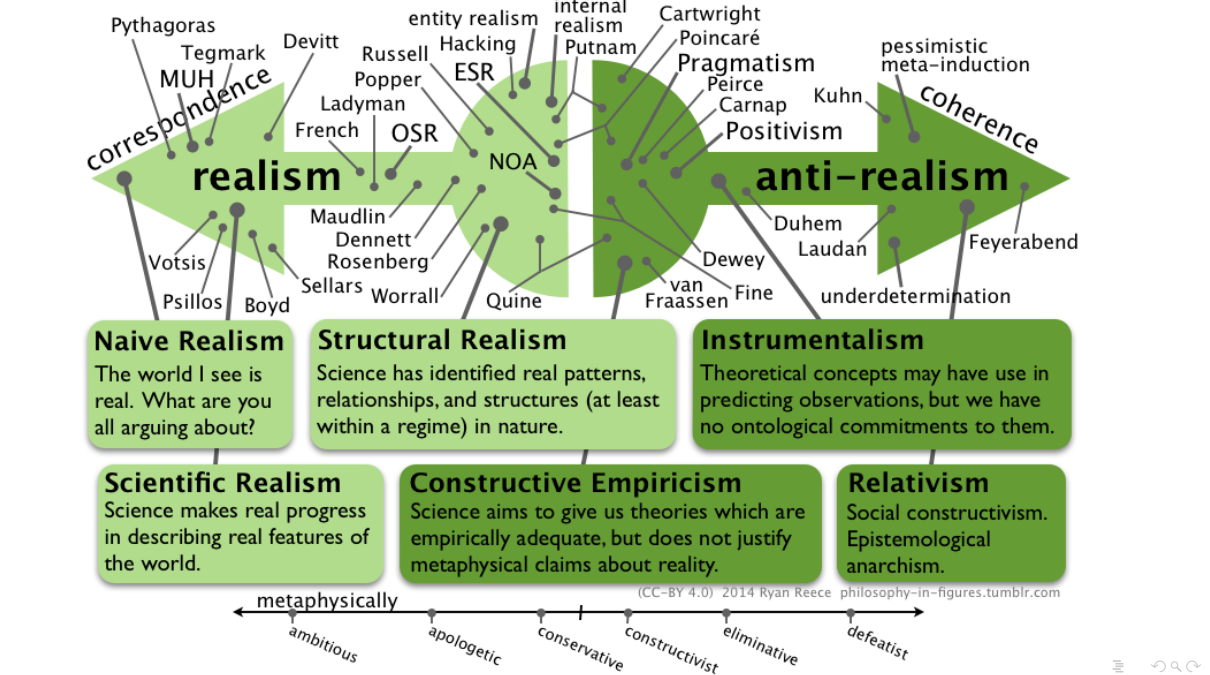
Text as Data

Juraj Medzihorsky

2016-11-28

?

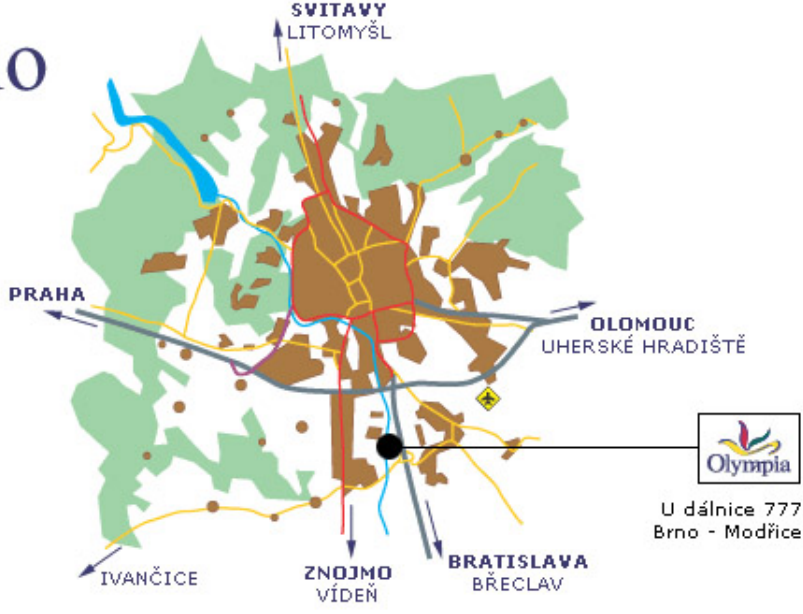

Pythagoras
Tegmark
MUH
correspondence
Devitt
entity realism
Hacking
Russell
ESR
internal realism
Putnam
Cartwright
Poincaré
Pragmatism
Peirce
Carnap
Positivism
pessimistic meta-induction
Kuhn
coherence
Popper
Ladyman
French
OSR
NOA
realism
anti-realism
Maudlin
Dennett
Rosenberg
Worrall
Quine
Votsis
Psillos
Boyd
Sellars
Dewey
van Fraassen
Fine
Duhem
Laudan
Feyerabend
underdetermination
Naive Realism
The world I see is real. What are you all arguing about?
Structural Realism
Science has identified real patterns, relationships, and structures (at least within a regime) in nature.
Instrumentalism
Theoretical concepts may have use in predicting observations, but we have no ontological commitments to them.
Scientific Realism
Science makes real progress in describing real features of the world.
Constructive Empiricism
Science aims to give us theories which are empirically adequate, but does not justify metaphysical claims about reality.
Relativism
Social constructivism. Epistemological anarchism.
(CC-BY 4.0) 2014 Ryan Reece philosophy-in-figures.tumblr.com
metaphysically
ambitious
apologetic
conservative
constructivist
eliminative
defeatist

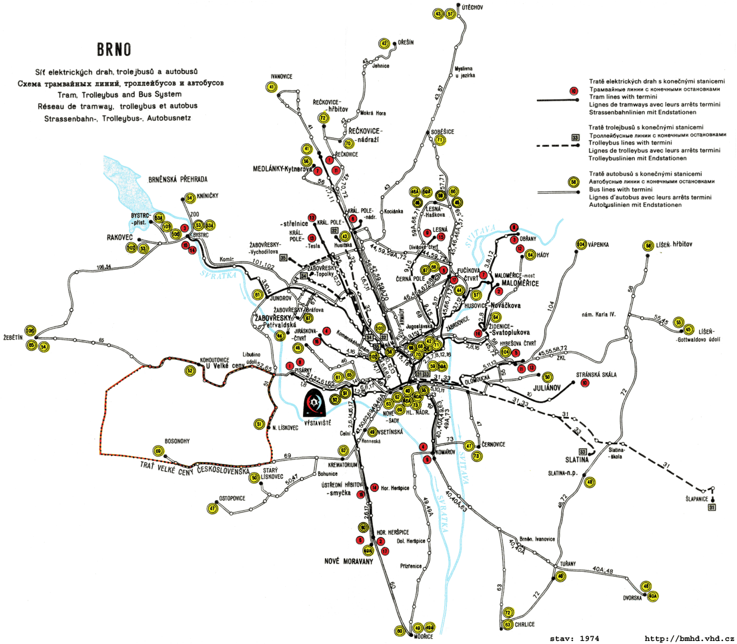
Motivation

?

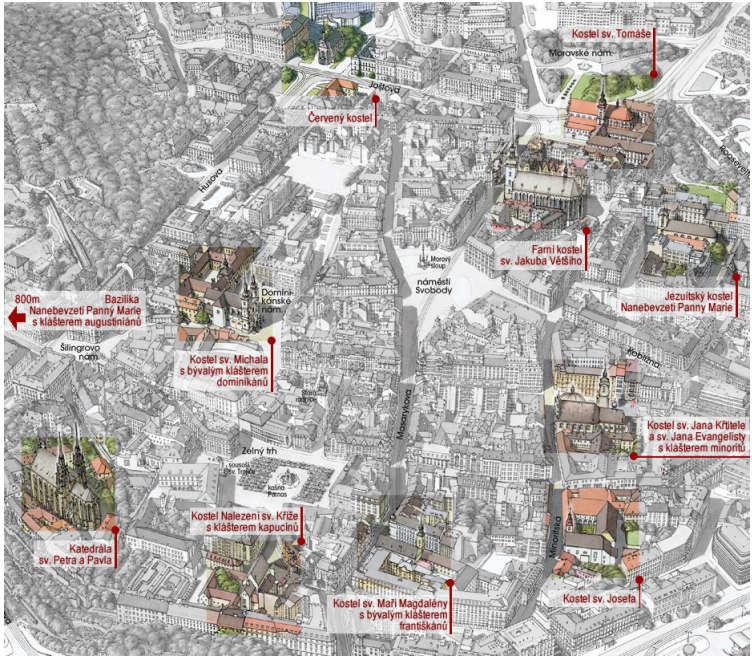
all models are false,
but some are useful
George E. P. Box

Brno

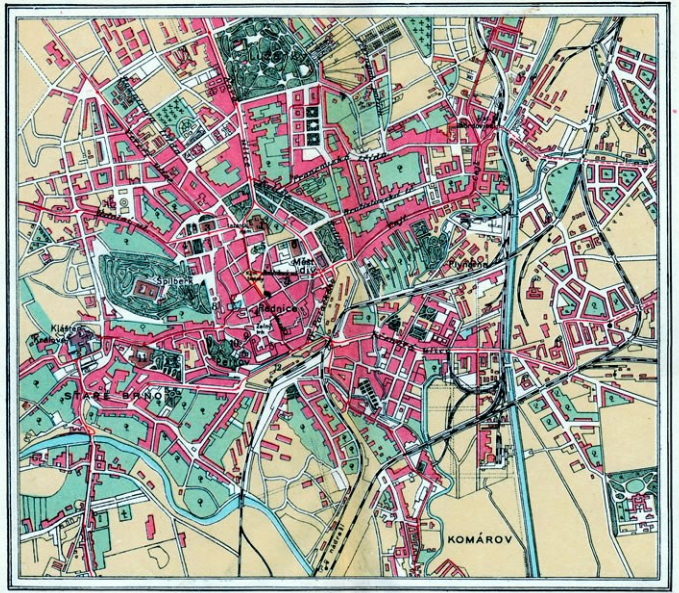

SVITAVY
LITOMYŠL
PRAHA
OLOMOUC
UHERSKÉ HRADIŠTĚ
Olympia
U dálnice 777
Brno - Modřice
IVANČICE
ZNOJMO
VÍDEŇ
BRATISLAVA
BŘECLAV



stav: 1974 http://bmhd.vhd.cz


Kostel sv. Tomáše
Moravské nám.
Jostova
Červený kostel
Husova
Farní kostel
sv. Jakuba Většího
Jezuitský kostel
Nanebevzetí Panny Marie
Morový
sloup
náměstí
Svobody
800m
Bazilika
Nanebevzetí Panny Marie
s klášterem augustiniánů
Dominikánské
nám.
Šilingrovo
nám.
Kostel sv. Michala
s bývalým klášterem
dominikánů
Masarykova
Zelný trh
Kostel sv. Jana Křtitele
a sv. Jana Evangelisty
s klášterem minoritů
Kostel Nalezení sv. Kříže
s klášterem kapucínů
Katedrála
sv. Petra a Pavla
Kostel sv. Maří Magdalény
s bývalým klášterem
františkánů
Kostel sv. Josefa

PLÁN MĚSTA BRNA.



Měřítko 1 : 25.000

Významné budovy:

1 Zem. správa politická.
2 Národní divadlo.
3 Starý zemský dům.
4 Radnice.
5 Chrám sv. Jakuba.
6 Nejvyš. s. dvůr.
7 Ředitelství pošt a telegrafů.
8 Průmyslové museum.
9 Zemské museum.
10 Dóm.
11 Zem. nemocnice.
12 Státní nádraží.

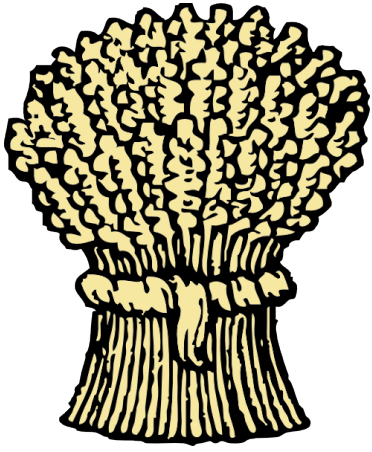
Document Scraping

Document Scraping

- Numbers and text in files
- Local
- Web

Web

- eXtensible Markup Language (XML)
- APIs (e.g. for Twitter)



Text Analysis

눈물 아롱아롱
피리 불고 가신 님의 밟으신 길은
진달래 꽃비 오는 서역 삼만리
흰 옷깃 여며 여며 가옵신 님의
다시 오진 못하는 파촉 삼만 리

ᛁᚴ ᛋᛉᛁᚢᛖᚾ ᚢᚱᚺᚠ ᛁᚱᚾ ᚢᛖᛦ ᚲᚱᛋᚴᚱᛦᚠ

ᛁᛖᛦᚢᚱᛋᚴᛖᛁ ᚳᛖᚱᛉᛖ ᛦᚴᛦ ᚳᚴᛁᚠ ᚢᚳᛋᚴᚢᛖᚢᛖᚾ ᚾᛖᚾ ᛦᚴᛖ ᛗᛖᛋᚠ ᚾᛖᚱ ᛗᚳᛁᚴᚱᚱᛦ ᚢᛖᛁᛉᚾᚢᚳᛦ ᚢᚴᛖ ᚢᚳᚾᚢᚳᛦᚴᚱᛦ ᚢᛖᛁᚱᛗᚢᚳᛦᚠ ᚠᚴᛖ ᛦᚱᛗᚴᚴᛦ ᚳᛖᚱᛉᛖᚱ

ᚱᚱᛗᛗᛖᛁ ᛖᚴᛁᚠ

ᛦᛉᛁᚱᛦ ᚾᚴᛁᛋᛉᛁ ᚠᚴᛖᚱ ᛦᚴᛖ ᛗᛖᛋᚾᚳᛦ ᛗᚳᛁᚱᛖᛗ ᚢᛖᛁᛉᚾᚱᚴᛁ ᚢᚱᛋ ᚢᛖᛁᛖᛗᛋᚴᛦ ᚾᛖᚾᛖᛦᚾᚳᛦ ᛋᛖᚢᚱᛁᛋᚴᛁᛦᚠ ᚱᚱᛋᚢᚴᛁᛋᚴᛁ ᛦᛖᛗ ᚢᛖᚹᛖᚱᛋᚳᛦᛗᛖᛁᚠ ᚱᚱᚴ ᚠᛖᛈ ᛦᚱᛈᛈᚱᛦ ᛦᚱᛖᛁᛋᛁᚴᛦᚳᛦ ᛗᛖᚱ ᛦᚴᛖ ᛦᚳᚾᚢᛖᛗ ᚲᚴᚲᚠ ᛦᛖᚹ ᛗᚳᛁ ᚢᚳᛋᚴᚾᚱᛦ ᛦᚱᛁᛉᚠᚾᚳᛦ ᛗᛖᚱᚠ

ᚱᚱᛗᛗᛖᛁ ᛖᚴᛁᚠ

ᛦᛉᛁᚱᛦ ᚾᚴᛁᛋᛉᛁ ᚠᚴᛖᚱ ᛦᚴᛖ ᛗᛖᛋᚾᚳᛦ ᛗᚳᛁᚱᛖᛗ ᚢᛖᛁᛉᚾᚱᚴᛁ ᚢᚱᛋ ᚢᛖᛁᛖᛗᛋᚴᛦ ᚾᛖᚾᛖᛦᚾᚳᛦ ᛋᛖᚢᚱᛁᛋᚴᛁᛦᚠ ᚱᚱᛋᚢᚴᛁᛋᚴᛁ ᛦᛖᛗ ᚢᛖᚹᛖᚱᛋᚳᛦᛗᛖᛁᚠ ᚱᚱᚴ ᚠᛖᛈ ᛦᚱᛈᛈᚱᛦ ᛦᚱᛖᛁᛋᛁᚴᛦᚳᛦ ᛗᛖᚱᚠ ᚱᚱᛗᛗᛖᛁ ᛖᚴᛁᚠ ᛦᛉᛁᚱᛦ ᚾᚴᛁᛋᛉᛁ ᚠᚴᛖᚱ ᛦᚴᛖ ᛗᛖᛋᚾᚳᛦ ᛗᚳᛁᚱᛖᛗ ᚢᛖᛁᛉᚾᚱᚴᛁ ᚢᚱᛋ ᚢᛖᛁᛖᛗᛋᚴᛦ ᚾᛖᚾᛖᛦᚾᚳᛦ ᛋᛖᚢᚱᛁᛋᚴᛁᛦᚠ ᚱᚱᛋᚢᚴᛁᛋᚴᛁ ᛦᛖᛗ ᚢᛖᚹᛖᚱᛋᚳᛦᛗᛖᛁᚠ ᚱᚱᚴ ᚠᛖᛈ ᛦᚱᛈᛈᚱᛦ ᛦᚱᛖᛁᛋᛁᚴᛦᚳᛦ ᛗᛖᚱᚠ ᚱᚱᛗᛗᛖᛁ ᛖᚴᛁᚠ ᛦᛉᛁᚱᛦ ᚾᚴᛁᛋᛉᛁ ᚠᚴᛖᚱ ᛦᚴᛖ ᛗᛖᛋᚾᚳᛦ ᛗᚳᛁᚱᛖᛗ ᚢᛖᛁᛉᚾᚱᚴᛁ ᚢᚱᛋ ᚢᛖᛁᛖᛗᛋᚴᛦ ᚾᛖᚾᛖᛦᚾᚳᛦ ᛋᛖᚢᚱᛁᛋᚴᛁᛦᚠ ᚱᚱᛋᚢᚴᛁᛋᚴᛁ ᛦᛖᛗ ᚢᛖᚹᛖᚱᛋᚳᛦᛗᛖᛁᚠ ᚱᚱᚴ ᚠᛖᛈ ᛦᚱᛈᛈᚱᛦ ᛦᚱᛖᛁᛋᛁᚴᛦᚳᛦ ᛗᛖᚱᚠ

?

Text Analysis: The Big Picture

Text Analysis

- Discourse
- Content

Content Analysis

- 'Manual' Text Analysis
- Computer-Assisted Text Analysis (CATA)

‘Manual’ Text Analysis

- Humans do most of the work
- Expensive
- Slow
- Reliability issues

Computer-Assisted Text Analysis

- Computers do most of the work
- Boom
  - Huge amount of digitized text available
  - Cheap computing power
  - New methods – CS & PS

Political Text in CATA

Examples:

- Manifestos & platforms
- Press releases
- Social media content
- Floor and debate speeches


ith
orrow

‘Bag of Words’

- Common assumption in CATA
- Order of words (**n-ngrams** of words) does not matter
- Texts as vectors of word counts

Table 2: A word-frequency matrix for two randomly selected sentences from the corpus: a) ‘Let’s enforce the laws already on the book.’ (Cain in 2012); b) ‘Now, the administration has $800 million on hand right now, cash on hand.’ (Hunter in 2008)

	Sentence	
Stem	a	b
administr	0	1
book	1	0
cash	0	1
enforc	1	0
hand	0	2
law	1	0
million	0	1

CATA in Political Science

- Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, mps028.

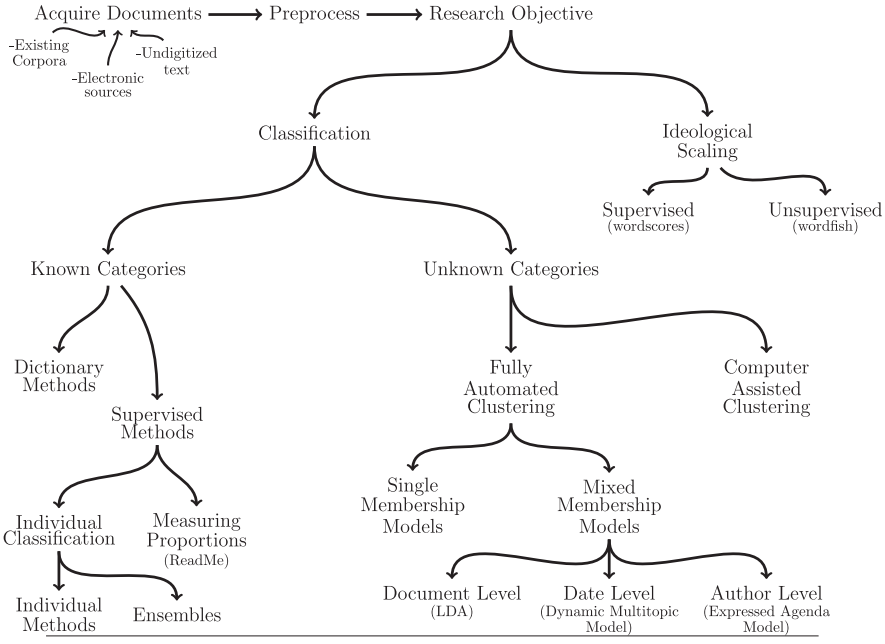


Fig. 1 An overview of text as data methods.

Dictionary-Based Methods

Google N-Grams

Dictionary-Based Methods

- Build a dictionary
- Statistical models are not necessary

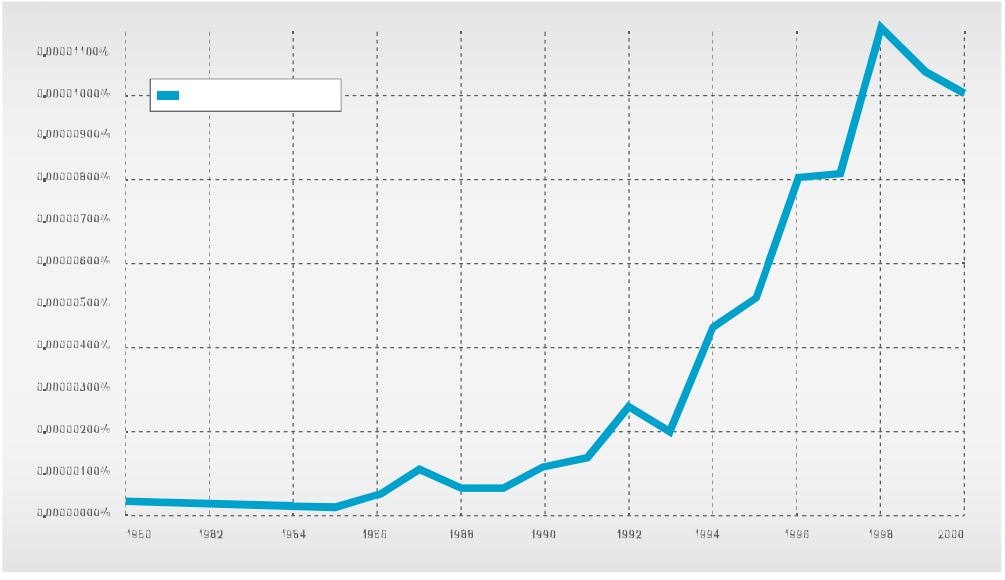
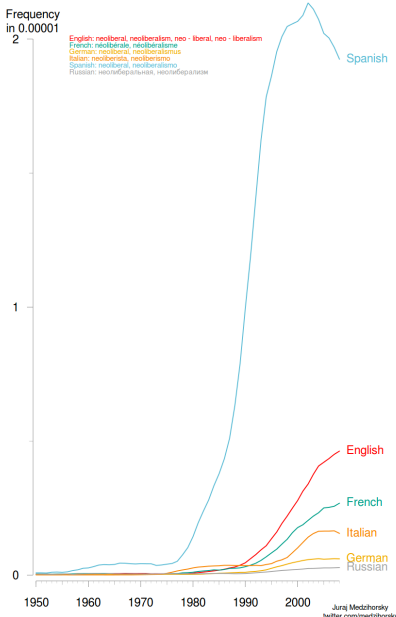
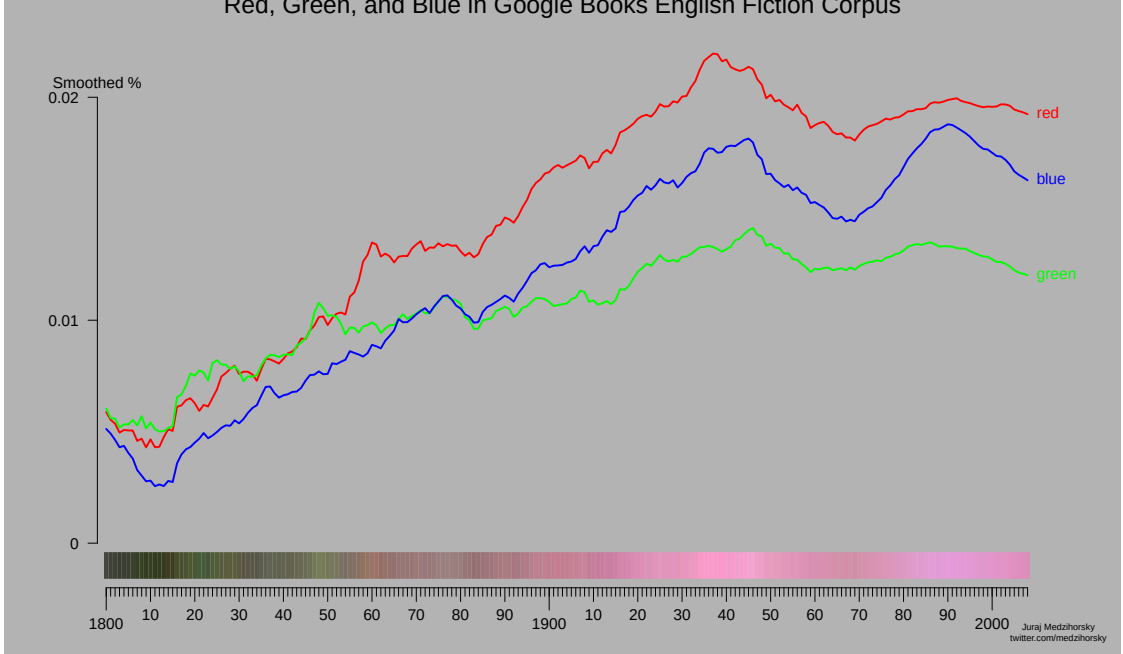


Figure 1: Frequency of appearance of "public intellectual" in Google Books from 1980 to 2000
Source: Graph. *Google Books*. Google, 18 June 2012. Web. 18 June 2012.

Figure 1: Frequency of appearance of “Expertise” and “Expert” in Google Books from 1800 to 2000

'Neoliberalism' in Google Book Corpora




Red, Green, and Blue in Google Books English Fiction Corpus
Smoothed %
0.02
0.01
0
red
blue
green
1800
10
20
30
40
50
60
70
80
90
1900
10
20
30
40
50
60
70
80
90
2000
Juraj Medzihorsky
twitter.com/medzihorsky

Scaling

Scaling Goals

- One or more dimensions
- Place documents (texts, speeches) in a space
- Place words in the same space

Common Scaling Methods

Supervised: Wordscores

Unsupervised: Wordfish

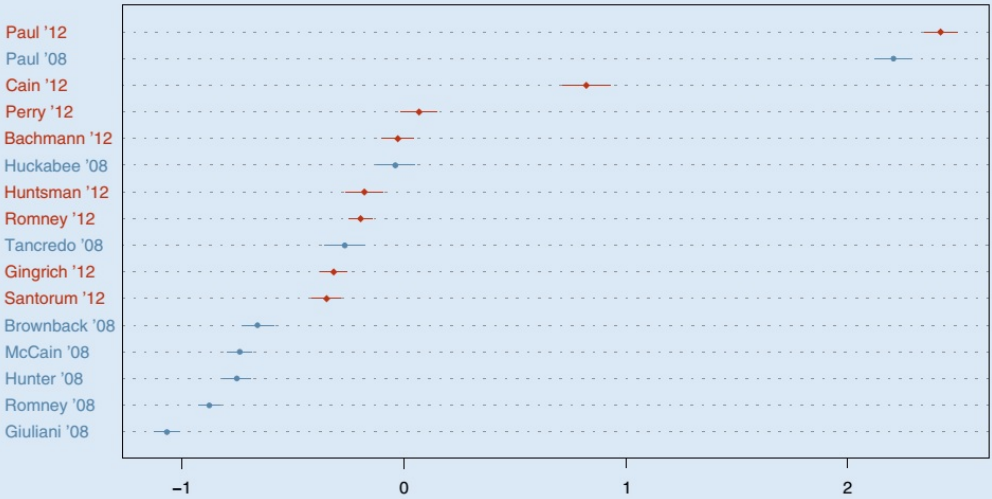
Unsupervised: Correspondence Analysis

A Scaling Example

- 2008 and 2012 Republican presidential primaries
- Debate transcripts from a UCSB website
- Expected move towards Tea Party positions
- Unsupervised scaling: Wordfish

Figure 1

Candidate Positions



Candidate positions extracted from their pre-Iowa debate speeches with bootstrapped 95% confidence intervals (1,000 replications). 2008 candidates denoted by circles (blue) and 2012 candidates by diamonds (red).

Table 4

Selected Five-Sentence Sequences Spoken by a Single Candidate in a Single Debate

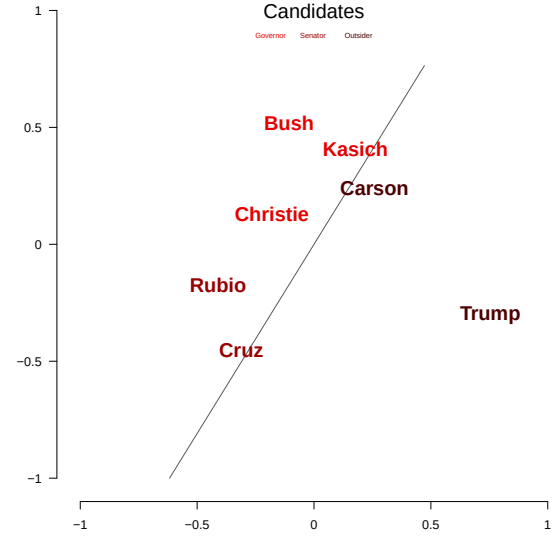
NEGATIVE, −1±0.1	POSITIVE, +1±0.1
"I have joined together across the aisle on a number of pieces of legislation, many of them very important. I'm proud of my legislative record of conserving my ideals and my conservative principles and getting things done in Washington. And I am proud of that, and I will continue to hold to those ideals. But I will reach across the aisle to the Democrats who I have worked with, who know me, and we know we can work together for the good of this country. Let's raise the level of dialogue and discussion and debate in this campaign." (McCain on December 12, 2007; score: −1.1)	"Repeal Dodd-Frank, repeal Obamacare. It really isn't that tough if you try. It is easy to turn around this economy, just have the backbone to do it. Well, as president of the United States, I would not be reappointing Ben Bernanke, but I want to say this. During the bailout, the $700 billion bailout, I worked behind the scenes against the bailout, because one of the things that I saw from the Federal Reserve, the enabling act legislation is written so broadly that, quite literally, Congress has given the Federal Reserve almost unlimited power over the economy." (Bachmann on September 12, 2011; score: 1.9)
"It's the one place I found to agree with President Obama. If every parent in America had a choice of the school their child went to, if that school had to report its scores, if there was a real opportunity, you'd have a dramatic improvement. I visited schools where, three years earlier, there were fights, there were dropouts, there was no hope. They were taken over by a charter school in downtown Philadelphia, and all of a sudden the kids didn't fight anymore, because they were disciplined. They were all asked every day, what college are you going to? Not are you going to go to college, what college are you going." (Gingrich on September 7, 2011; score: −1)	"If we look for it, you'll realize that our national sovereignty is under threat. Yes, and I would like to state that, to the statement earlier made that we all went to Washington to change Washington and Washington changed us, I don't think that applies to me; Washington did not change me. I would like to change Washington, and we could by cutting three programs, such as the Department of Education–Ronald Reagan used to talk about that–Department of Energy, Department of Homeland Security is the biggest bureaucracy we ever had. And besides, what we can do is we can have a stronger national defense by changing our foreign policy. Our foreign policy is costing us a trillion dollars, and we can spend most of that or a lot of that money home if we would bring our troops home." (Paul on November 28, 2007; score: 2)
"I can tell you a good union, the Steel Workers Union. When last year, Chris, we had a strike in a Kansas plant that made the tires for our humvees, I called up the president of the Steelworkers and the president of Goodyear, and within a very short period of time, they were working together, they got that thing done for the good of the country. A union is a receptacle of power, just like management. But those folks love this country, they love their family, and they helped to build a middle class, which has been important for America and for our party. We need to work with unions to win this presidency." (Hunter on October 9, 2007; score: −0.9)	"There's a responsible way for the federal government to do the things that it should do. Running organizations like the TSA, I would agree with Representative Paul, no. Having the federal government responsible for trying to micromanage Medicare, no, trying to micromanage education, no. The federal government is not good at micromanaging anything. This is why I believe in empowering the states to do more and limit what the federal government does with regard to those kinds of program." (Cain on August 11, 2011; score: 2.1)

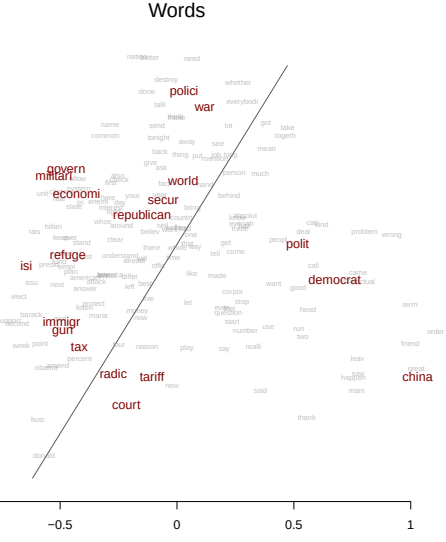
Sequences were selected from all such sequences longer than 500 characters and within ±0.1 to two points on the dimension: −1 and +2. Sequences in the first column characterize the negative end of the extracted dimension and sequences in the second column the positive end.

Another Scaling Example

- Transcripts of US presidential debates
- Unsupervised scaling with correspondence analysis
- Two dimensions

Sixth GOP'16 Presidential Debate





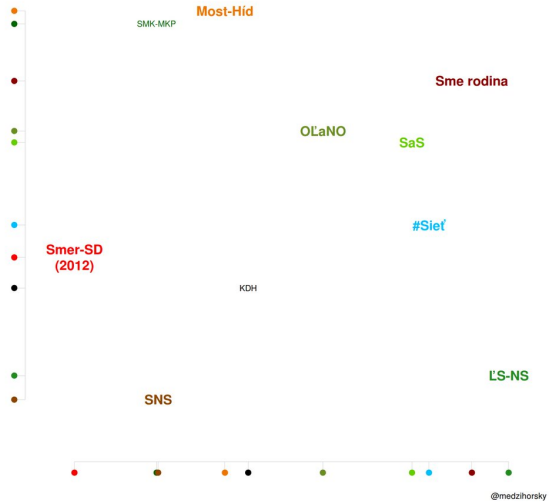
Yet Another Scaling Example

- Slovakian party manifestos: text and CMP codes
- Unsupervised scaling with correspondence analysis
- Two dimensions

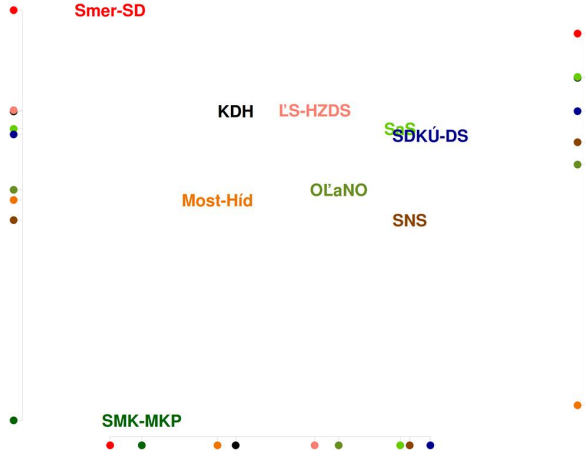
Slovakian Elections 2016: Manifesto Vocabularies

* Smer-SD ran without a new manifesto

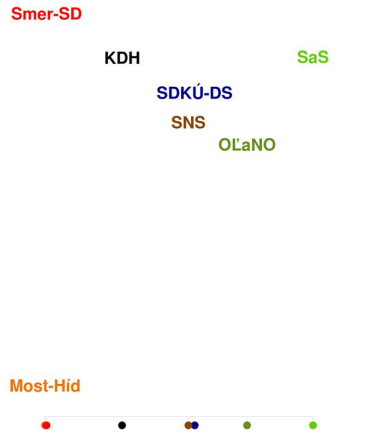
Parties with seats in bold type



Voľby 2012: programy



Voľby 2012: CMP kódy

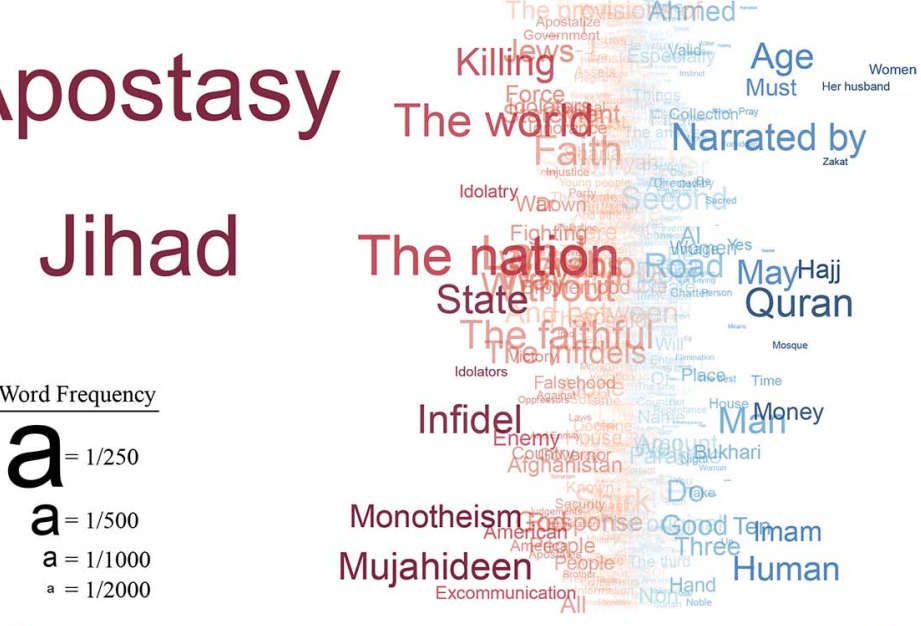


And One More Scaling Example

- Nielsen's (2013) dissertation
- 25,000 + documents by $\sim$ 100 clerics
- *Jihad Score*
- Supervised scaling with a training set

Apostasy

Jihad



If a person arrives while the Imam is preaching at Friday prayers, he should pray two brief prostrations and sit without greeting anyone as greeting people in this circumstance is forbidden because the Prophet, peace be upon him, says, "If your friend speaks to you during the Friday prayers, silence him while the Imam preaches because it is idle talk."

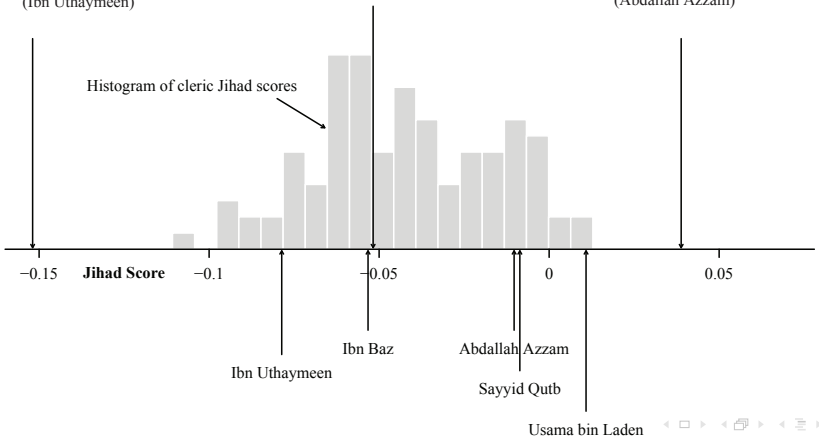
(Ibn Uthaymeen)

There is a fundamental fact about the nature of this religion and the way it works in people's lives. A fundamental, simple fact, but although it is simple, it is often forgotten or not realized at all. Forgetting this fact, or failing to recognize it arises from a serious omission from views of this religion: its truthfulness and historical, present, and future reality.

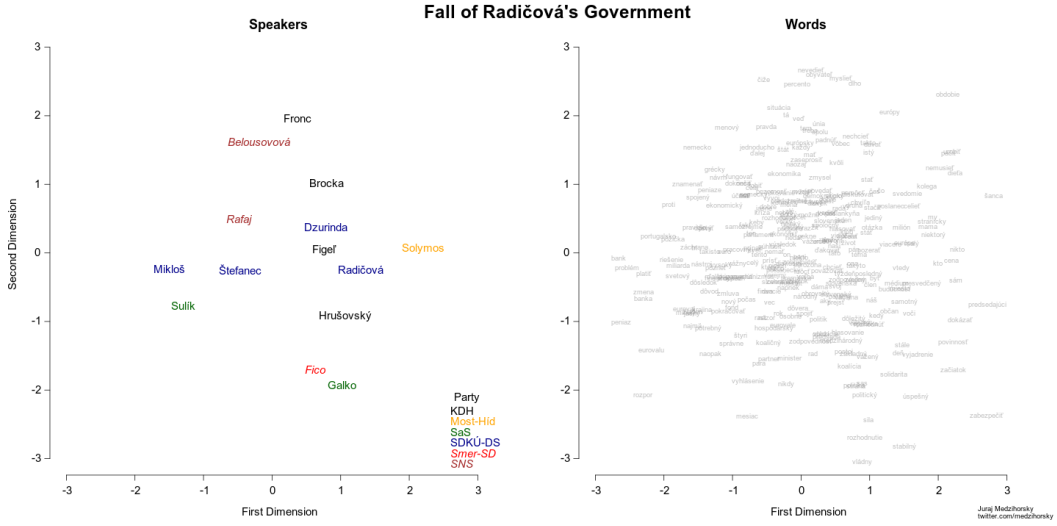
(Sayyid Qutb)

Ruling on Fighting Now in Palestine and Afghanistan. The foregoing has clarified that if an inch of Muslim lands are attacked, then Jihad is obligatory for the people of that area, and those near by. If they do not succeed or are incapable or lazy, the individual obligation widens to those behind them and then gradually the individual obligation expands until it is general for the whole land, from East to West.

(Abdallah Azzam)



Failed Unsupervised Scaling



‘Topic’ Models

Topics

gene	0.04
dna	0.02
genetic	0.01
...	

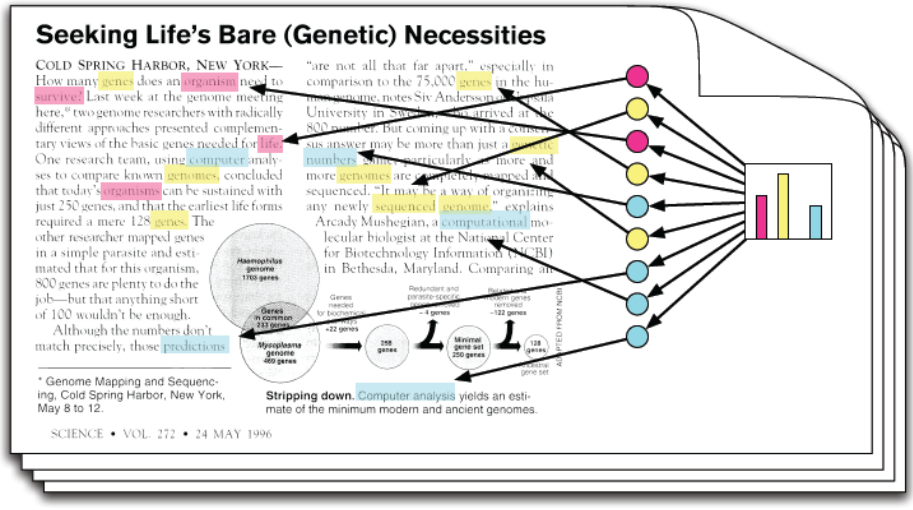
life	0.02
evolve	0.01
organism	0.01
...	

brain	0.04
neuron	0.02
nerve	0.01
...	

data	0.02
number	0.02
computer	0.01
...	

Documents

Topic proportions and assignments



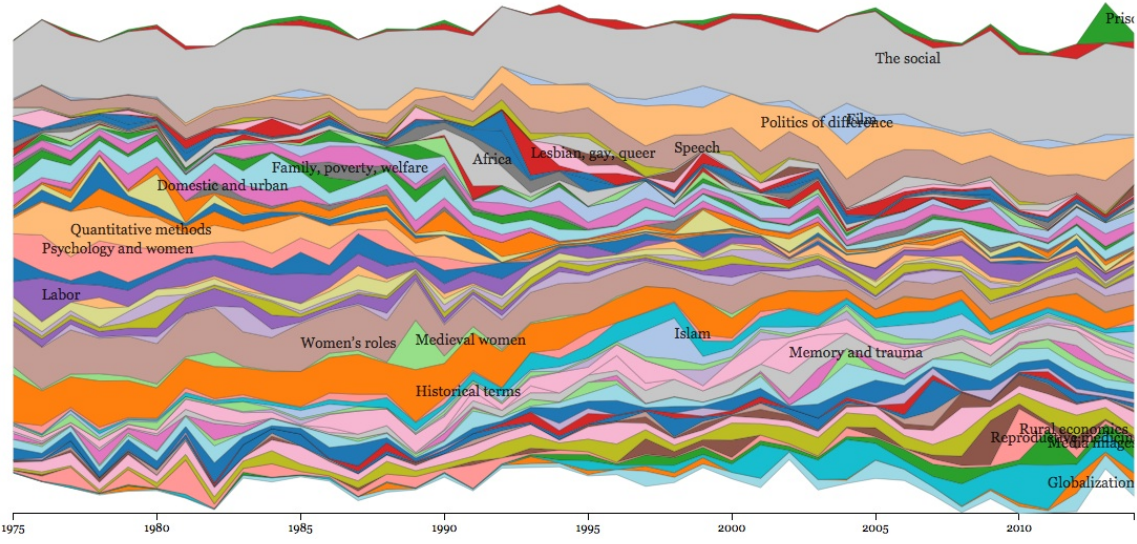
Seeking Life's Bare (Genetic) Necessities

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

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

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

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


Pris
The social
Politics of difference
Film
Africa
Lesbian, gay, queer
Speech
Family, poverty, welfare
Domestic and urban
Quantitative methods
Psychology and women
Labor
Women's roles
Medieval women
Islam
Memory and trauma
Historical terms
Rural economics
Globalization
1975
1980
1985
1990
1995
2000
2005
2010

VisArgue - A Visual Text Analytics Framework for the Study of Deliberative Communication

Mennatallah El-Assady[1], **Valentin Gold**[2], **Annette Hautli-Janisz**[3], **Wolfgang Jentner**[1], **Miriam Butt**[2], **Katharina Holzinger**[2], **Daniel Keim**[3]
[1]Department of Computer and Information Science
[2]Department of Politics and Public Administration
[3]Department of Linguistics
University of Konstanz, Germany
valentin.gold@uni-konstanz.de*

Abstract

For the last two decades, deliberative democracy has been intensively debated within political science and other related fields. Only recently, deliberation research has experienced a computational turn. In this paper, we present a linguistic and visual framework for the study of deliberative communication. The framework

manding and time-consuming resulting in a limited set of debate corpora. Moreover, the coding is often subjective making it subject to critical judgments of other researchers (King, 2009; Black et al., 2010; Dacombe, 2013). As a result, manual coding poses challenges with respect to both validity and reliability.

Only recently, the computational turn in deliberation research allows to analyze large quantities of

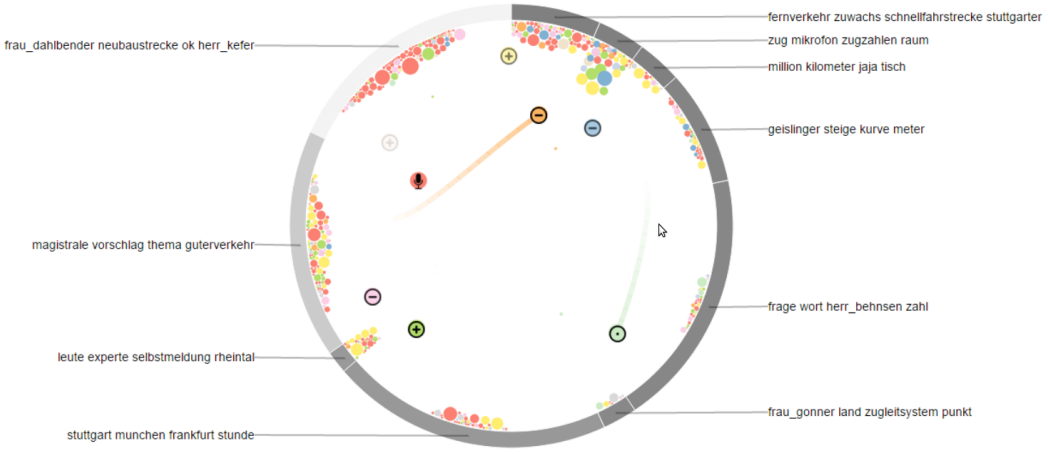


Figure 2: ConToVi Visualization

Back to the Big Picture

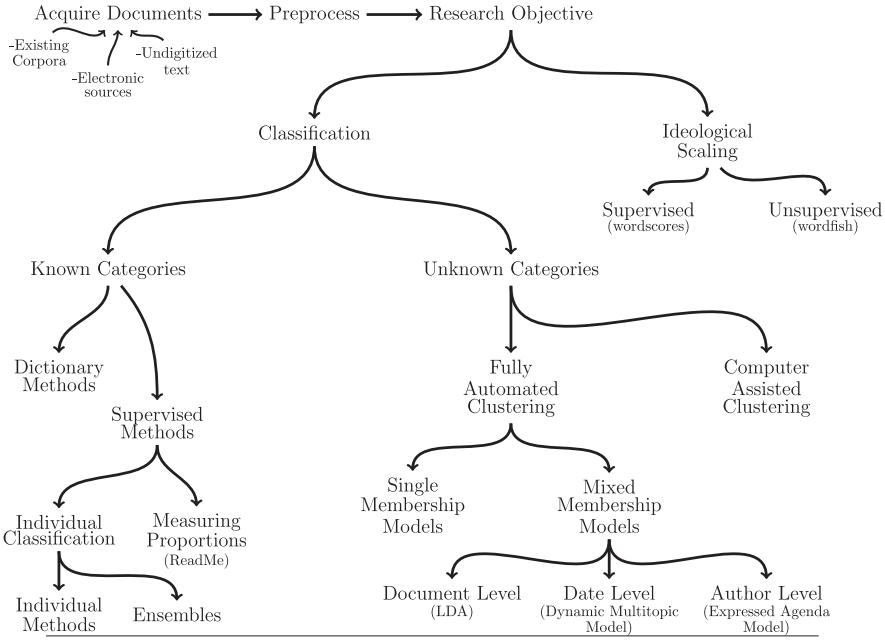


Fig. 1 An overview of text as data methods.

Beyond Money in Politics: Automatic Detection of Legislative Text Re-Use

Eugenia Giraudy
UC Berkeley & YouGov
eugenia.giraudy@gmail.com

Matthew Burgess
University of Michigan
mattburg@umich.edu

Julian Katz Samuels
University of Michigan
jkatzsam@umich.edu

Joe Walsh
University of Chicago
jtwalsh@uchicago.edu

Abstract

State legislatures introduce at least 45,000 bills each year. A large number of these bills are not drafted by legislators but by interest groups. However, existing approaches to detect the source of these bills are slow, biased, and incomplete.

This paper presents the Legislative Influence Detector (LID). LID uses the Smith-

drafted by interest groups. While state legislators normally adapt some of the language to meet their state's legal needs, such as references to existing state code, it is common practice to introduce bills that look the same as the document they borrowed from.

While legislative text re-use is a common phenomenon across states, Americans rely on watchdogs to find legislation of questionable origin, but with at least 45,000 bills introduced in state capi-

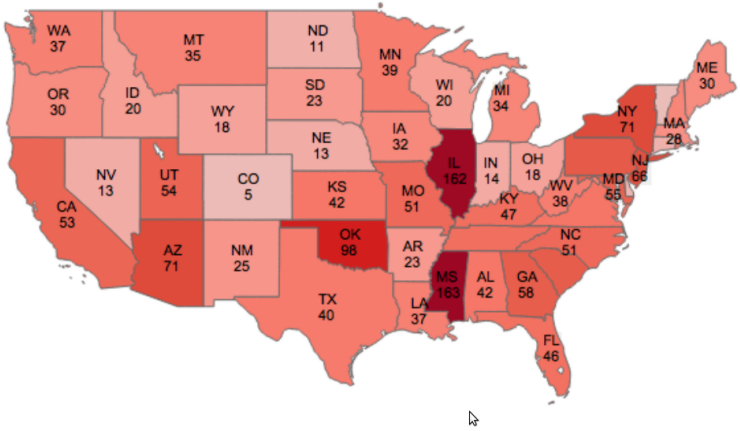


Figure 2: Introduced bills by state from ALEC model legislation

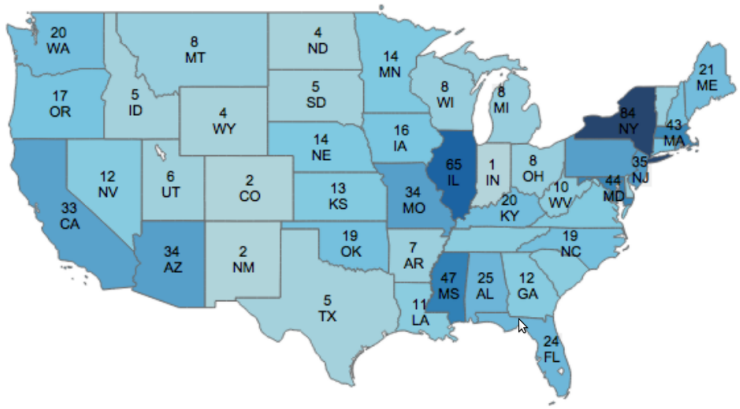


Figure 3: Introduced bills by state from ALICE model legislation