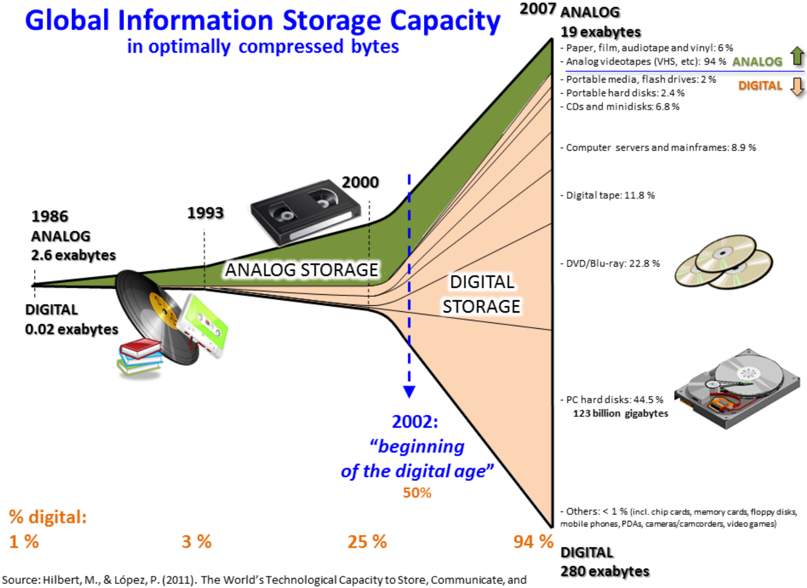
Building a Robot Judge: Data Science for the Law

3. Text Data Essentials

Elliott Ash

Global Information Storage Capacity

in optimally compressed bytes



1986
ANALOG
2.6 exabytes

DIGITAL
0.02 exabytes

1993

2000

ANALOG STORAGE

DIGITAL STORAGE

2002:
"beginning of the digital age"
50%

2007 ANALOG
19 exabytes

- Paper, film, audiotape and vinyl: 6 %
- Analog videotapes (VHS, etc): 94 % ANALOG
- Portable media, flash drives: 2 % DIGITAL
- Portable hard disks: 2.4 %
- CDs and minidisks: 6.8 %
- Computer servers and mainframes: 8.9 %
- Digital tape: 11.8 %
- DVD/Blu-ray: 22.8 %
- PC hard disks: 44.5 %
  123 billion gigabytes
- Others: < 1 % (incl. chip cards, memory cards, floppy disks, mobile phones, PDAs, cameras/camcorders, video games)

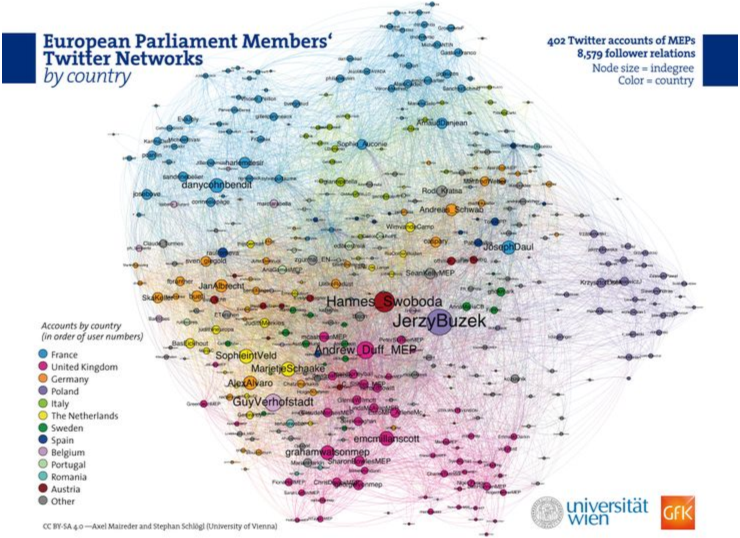
DIGITAL
280 exabytes

% digital:

1 %　　3 %　　25 %　　94 %

Source: Hilbert, M., & López, P. (2011). The World's Technological Capacity to Store, Communicate, and Compute Information. *Science*, 332(6025), 60 –65. http://www.martinhilbert.net/WorldInfoCapacity.html

New Data, New Possibilities



CC BY-SA 4.0 —Axel Maireder and Stephan Schlögl (University of Vienna)

Diversification of Text Data Methods

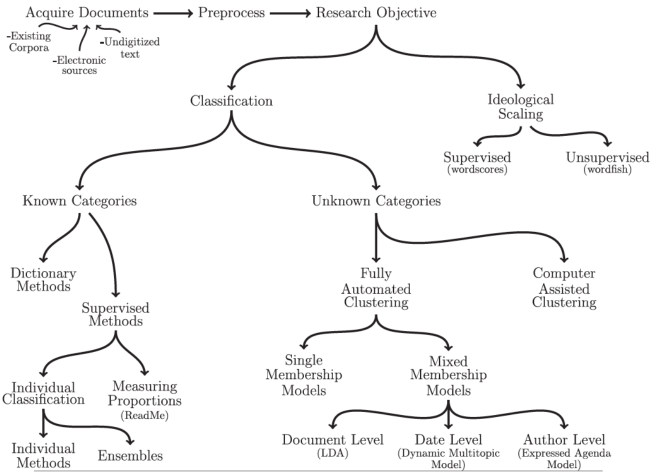


Fig. 1 An overview of text as data methods.

Source: Stewart and Grimmer (2013).

Overview

- Input:
    - A set of documents (e.g. text files), D.
- Output:
    - A matrix, X, containing statistics about phrase frequencies in those documents.

Text as Data

- Text data is a sequence of characters called **documents**.
- The set of documents is the **corpus**.
- Text data is **unstructured**:
  - the information we want is mixed together with (lots of) information we don't.
  - How to separate the two?
- All text data approaches will throw away some information:
  - The trick is figuring out how to retain valuable information.

Documents and metadata

- For small corpora, you might have the text and metadata together in a spreadsheet.
- For larger corpora, you might have:
  - A document is a text file (or an item in a relational database).
  - A corpus is a folder of text files.
  - The filenames for the text files should contain an identifier for linking to metadata.

What counts as a document?

- The unit of document analysis will vary depending on your question.
- If you are looking at how judges decide different types of cases, then a case would be a document.
- If you are looking at how judges differ within a court, then you might aggregate all of a judge's cases as a document.
- If you are looking at the impact of court cases on crime in a year, you might aggregate all the cases in a single year as a single document.
- If you are looking at how different topics are discussed within single cases, then a document might be a section or a paragraph.

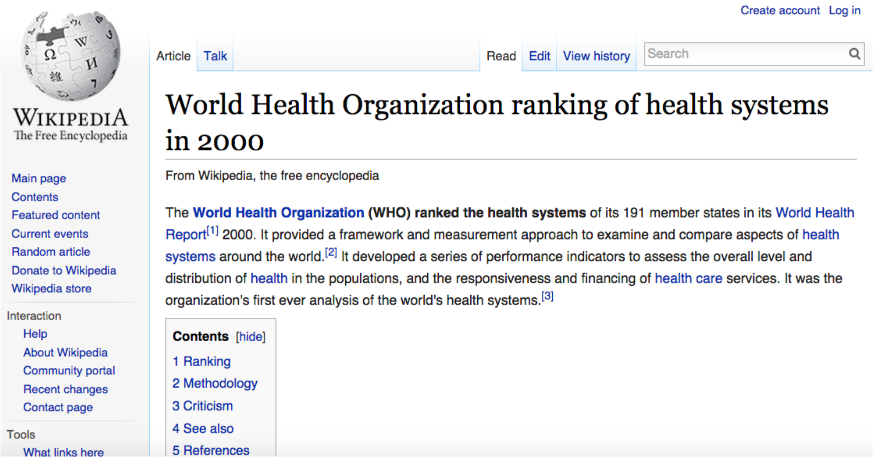
Publicly Available Corpora

- There is already a vast amount of data out there that has already been compiled (e.g. CourtListener, Twitter, New York Times, Reuters, Google, Wikipedia).
- Chris Bail curates a list of these datasets:
  - `https://docs.google.com/spreadsheets/d/1I7cvuCBQxosQK2evTcdL3qtglaEPc0WFEs6rZMx-xiE/edit`
- Some interesting corpora described in NLTK Book Chapter 2.
- Many proprietary corpora are becoming available for research:
  - Lexis
  - Web of Science

Screen Scraping

- A screen scraper is a computer program that:
  - loads/reads in a web page
  - finds some information on it
  - grabs the information
  - stores it in a dataset
- Once upon a time you could collect virtually any piece of information from the internet by screen scraping.
  - But now web sites make it difficult with restrictive terms of use, bot-blockers, javascript, etc.
  - Still, a little creativity goes a long way.

What a web site looks like to us



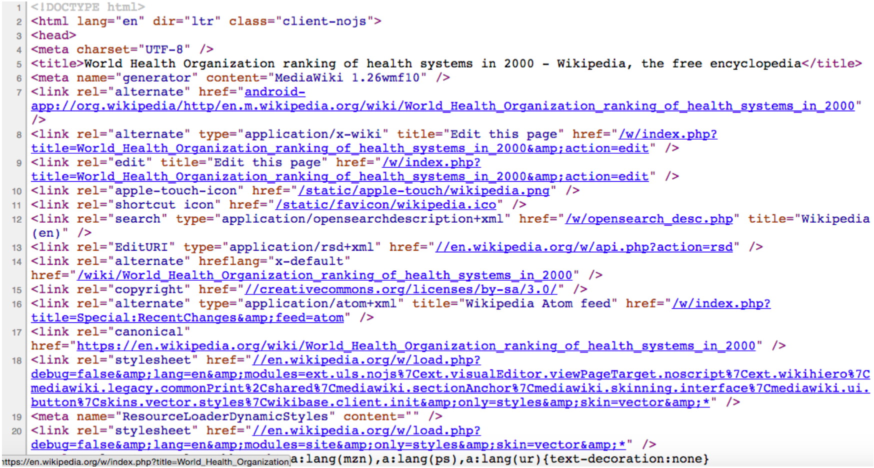
World Health Organization ranking of health systems in 2000

From Wikipedia, the free encyclopedia

The World Health Organization **(WHO) ranked the health systems** of its 191 member states in its World Health Report[1] 2000. It provided a framework and measurement approach to examine and compare aspects of health systems around the world.[2] It developed a series of performance indicators to assess the overall level and distribution of health in the populations, and the responsiveness and financing of health care services. It was the organization's first ever analysis of the world's health systems.[3]

Contents [hide]

What a web site looks like to a computer



```
<!DOCTYPE html>
<html lang="en" dir="ltr" class="client-nojs">
<head>
<meta charset="UTF-8" />
<title>World Health Organization ranking of health systems in 2000 - Wikipedia, the free encyclopedia</title>
<meta name="generator" content="MediaWiki 1.26wmf10" />
<link rel="alternate" href="android-app://org.wikipedia/http/en.m.wikipedia.org/wiki/World_Health_Organization_ranking_of_health_systems_in_2000" />
<link rel="alternate" type="application/x-wiki" title="Edit this page" href="/w/index.php?title=World_Health_Organization_ranking_of_health_systems_in_2000&amp;action=edit" />
<link rel="edit" title="Edit this page" href="/w/index.php?title=World_Health_Organization_ranking_of_health_systems_in_2000&amp;action=edit" />
<link rel="apple-touch-icon" href="/static/apple-touch/wikipedia.png" />
<link rel="shortcut icon" href="/static/favicon/wikipedia.ico" />
<link rel="search" type="application/opensearchdescription+xml" href="/w/opensearch_desc.php" title="Wikipedia (en)" />
<link rel="EditURI" type="application/rsd+xml" href="//en.wikipedia.org/w/api.php?action=rsd" />
<link rel="alternate" hreflang="x-default" href="/wiki/World_Health_Organization_ranking_of_health_systems_in_2000" />
<link rel="copyright" href="//creativecommons.org/licenses/by-sa/3.0/" />
<link rel="alternate" type="application/atom+xml" title="Wikipedia Atom feed" href="/w/index.php?title=Special:RecentChanges&amp;feed=atom" />
<link rel="canonical" href="https://en.wikipedia.org/wiki/World_Health_Organization_ranking_of_health_systems_in_2000" />
<link rel="stylesheet" href="//en.wikipedia.org/w/load.php?debug=false&amp;lang=en&amp;modules=ext.uls.nojs%7Cext.visualEditor.viewPageTarget.noscript%7Cext.wikihiero%7Cmediawiki.legacy.commonPrint%2Cshared%7Cmediawiki.sectionAnchor%7Cmediawiki.skinning.interface%7Cmediawiki.ui.button%7Cskins.vector.styles%7Cwikibase.client.init&amp;only=styles&amp;skin=vector&amp;*" />
<meta name="ResourceLoaderDynamicStyles" content="" />
<link rel="stylesheet" href="//en.wikipedia.org/w/load.php?debug=false&amp;lang=en&amp;modules=site&amp;only=styles&amp;skin=vector&amp;*" />
ttps://en.wikipedia.org/w/index.php?title=World_Health_Organization a:lang(mzn),a:lang(ps),a:lang(ur){text-decoration:none}
```

Browser Automation

- Many web sites are designed to be difficult to scrape.
- Python has solutions for simulating a human browser:
  - selenium (chromedriver, phantomjs)
- Other solutions if all else fails:
  - DownThemAll! plug-in for Firefox
  - Hire mechanical turkers to manually download data.

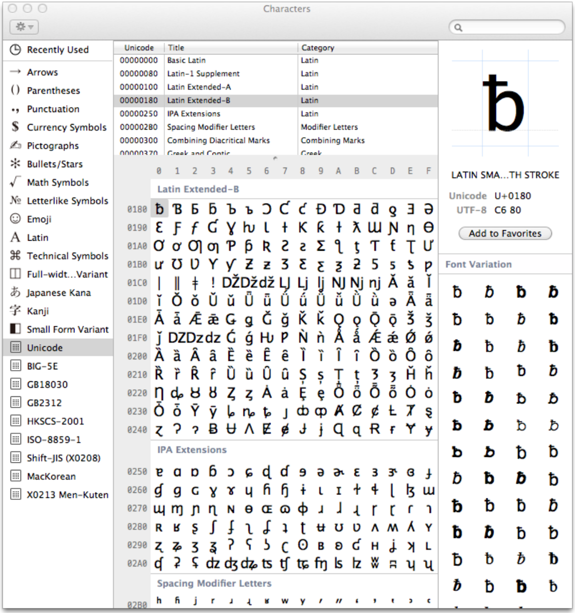
API's

- API = Application Programming Interface
  - These are developer-oriented tools that provide access to cleaner data.
- Chris Bail's list of API's that could be interesting for research:
  - https://docs.google.com/spreadsheets/d/1ZEr3okdlb0zctmXOMZKo-gZKPsq5WGn1nJOxPV7al-Q/edit

Other Languages

- All of the tools that we discuss in this class are available in many languages.
- spaCy has full functionality in English, German, Spanish, Portuguese, French, Italian, and Dutch.
  - beta functionality in dozens of other languages including Chinese and Arabic
  - See `https://spacy.io/usage/models`.
- The machine learning models are language-independent.

Character Encodings



Unicode	Title	Category
00000000	Basic Latin	Latin
00000080	Latin-1 Supplement	Latin
00000100	Latin Extended-A	Latin
00000180	Latin Extended-B	Latin
00000250	IPA Extensions	Latin
000002B0	Spacing Modifier Letters	Modifier Letters
00000300	Combining Diacritical Marks	Combining Marks
00000370	Greek and Coptic	Greek

LATIN SMA…TH STROKE

Unicode U+0180

UTF-8 C6 80

	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
	Latin Extended-B															
0180	ƀ	Ɓ	Ƃ	ƃ	Ƅ	ƅ	Ɔ	Ƈ	ƈ	Ɖ	Ɗ	Ƌ	ƌ	ƍ	Ǝ	Ə
0190	Ɛ	Ƒ	ƒ	Ɠ	Ɣ	ƕ	Ɩ	Ɨ	Ƙ	ƙ	ƚ	ƛ	Ɯ	Ɲ	ƞ	Ɵ
01A0	Ơ	ơ	Ƣ	ƣ	Ƥ	ƥ	Ʀ	Ƨ	ƨ	Ʃ	ƪ	ƫ	Ƭ	ƭ	Ʈ	Ư
01B0	ư	Ʊ	Ʋ	Ƴ	ƴ	Ƶ	ƶ	Ʒ	Ƹ	ƹ	ƺ	ƻ	Ƽ	ƽ	ƾ	ƿ
01C0	ǀ	ǁ	ǂ	ǃ	Ǆ	ǅ	ǆ	Ǉ	ǈ	ǉ	Ǌ	ǋ	ǌ	Ǎ	ǎ	Ǐ
01D0	ǐ	Ǒ	ǒ	Ǔ	ǔ	Ǖ	ǖ	Ǘ	ǘ	Ǚ	ǚ	Ǜ	ǜ	ǝ	Ǟ	ǟ
01E0	Ǡ	ǡ	Ǣ	ǣ	Ǥ	ǥ	Ǧ	ǧ	Ǩ	ǩ	Ǫ	ǫ	Ǭ	ǭ	Ǯ	ǯ
01F0	ǰ	Ǳ	ǲ	ǳ	Ǵ	ǵ	Ƕ	Ƿ	Ǹ	ǹ	Ǻ	ǻ	Ǽ	ǽ	Ǿ	ǿ
0200	Ȁ	ȁ	Ȃ	ȃ	Ȅ	ȅ	Ȇ	ȇ	Ȉ	ȉ	Ȋ	ȋ	Ȍ	ȍ	Ȏ	ȏ
0210	Ȑ	ȑ	Ȓ	ȓ	Ȕ	ȕ	Ȗ	ȗ	Ș	ș	Ț	ț	Ȝ	ȝ	Ȟ	ȟ
0220	Ƞ	ȡ	Ȣ	ȣ	Ȥ	ȥ	Ȧ	ȧ	Ȩ	ȩ	Ȫ	ȫ	Ȭ	ȭ	Ȯ	ȯ
0230	Ȱ	ȱ	Ȳ	ȳ	ȴ	ȵ	ȶ	ȷ	ȸ	ȹ	Ⱥ	Ȼ	ȼ	Ƚ	Ⱦ	ȿ
0240	ɀ	Ɂ	ɂ	Ƀ	Ʉ	Ʌ	Ɇ	ɇ	Ɉ	ɉ	Ɋ	ɋ	Ɍ	ɍ	Ɏ	ɏ
	IPA Extensions															
0250	ɐ	ɑ	ɒ	ɓ	ɔ	ɕ	ɖ	ɗ	ɘ	ə	ɚ	ɛ	ɜ	ɝ	ɞ	ɟ
0260	ɠ	ɡ	ɢ	ɣ	ɤ	ɥ	ɦ	ɧ	ɨ	ɩ	ɪ	ɫ	ɬ	ɭ	ɮ	ɯ
0270	ɰ	ɱ	ɲ	ɳ	ɴ	ɵ	ɶ	ɷ	ɸ	ɹ	ɺ	ɻ	ɼ	ɽ	ɾ	ɿ
0280	ʀ	ʁ	ʂ	ʃ	ʄ	ʅ	ʆ	ʇ	ʈ	ʉ	ʊ	ʋ	ʌ	ʍ	ʎ	ʏ
0290	ʐ	ʑ	ʒ	ʓ	ʔ	ʕ	ʖ	ʗ	ʘ	ʙ	ʚ	ʛ	ʜ	ʝ	ʞ	ʟ
02A0	ʠ	ʡ	ʢ	ʣ	ʤ	ʥ	ʦ	ʧ	ʨ	ʩ	ʪ	ʫ	ʬ	ʭ	ʮ	ʯ
	Spacing Modifier Letters															
02B0	ʰ	ʱ	ʲ	ʳ	ʴ	ʵ	ʶ	ʷ	ʸ	ʹ	ʺ	ʻ	ʼ	ʽ	ʾ	ʿ

Corpus cleaning

- What we've already done:
  - removed HTML markup, extra white space, and unicode
- But HTML markup is often valuable:
  - HTML markup for section header names.
  - Legal database web sites often have HTML tags for citations to other cases.
- Other cleaning steps:
  - page numbers
  - hyphenations at line breaks
  - table of contents, indexes, etc.
- These are all corpus-specific, so inspect ahead of time.

Regular Expressions

- Regular Expressions, implemented in the Python package **re**, provide a powerful string matching tool.
  - A systematic string matching protocal – can match arbitrary string patterns
  - e.g., use utilit* to match utility, utilities, utilitarian, ...
  - Important for identifying speaker names (in political documents) section headers (in statutes), citations (in judicial opinions), etc.
- Also quite tedious, so we will not cover it here.
  - See NLTK book Chapter 3.4-3.5 for an introduction.

OCR (Optical Character Recognition)

- Your data might be in PDF's or images. Needs to be converted to text
- The best solution (that I know of) is ABBYY FineReader, which is expensive but might be available at your university library.
- My colleague Joe Sutherland at Columbia has a nice open-source package for OCR:
    - `https://github.com/jlsutherland/doc2text`

Should you run a spell checker?

- The short answer is no:
  - Most corpora have important specialized vocabulary that would be flagged by standard spell-checkers.
  - They are also very slow to run on large corpora.
  - In most empirical contexts, it's safe to assume that spelling errors (especially OCR errors) are uncorrelated with treatment assignment.
- Better solutions:
  - drop short (one or two letters) and long words (over 12 letters).
  - get doc frequencies for each word and filter out rare words
    - or use word embeddings and trust that misspellings will be nearby the true word.
- But:
  - There are cases where spelling errors could be correlated with treatment (for example, increasing legislator salaries might change both policy priorities and spelling error rates)

Measuring Judicial Output using Decision Texts

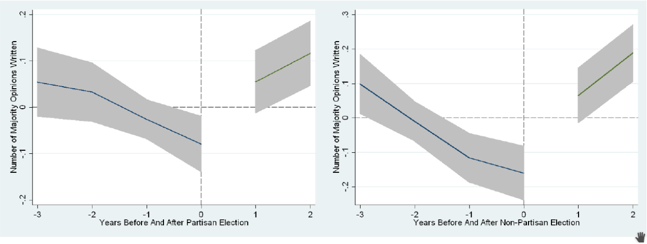
- The number of documents, and the length of those documents, already provide an interesting set of variables for analysis.
- For example:
  - How do electoral incentives affect judge effort?
  - How does the biological aging process affect effort and writing style?

Empirical Setting

- The setting for Ash and MacLeod (2015, 2017, 2018):
  - State supreme courts: the highest appellate court for each of the 50 states in the USA.
  - Data set has 1.1 million judicial opinions for 1947-1994
- States are a nice place to look at natural experiments:
  - Unlike most jurisdictions, state judges are often elected, and the rules for election change over time.

Elections Reduce Number of Opinions Written

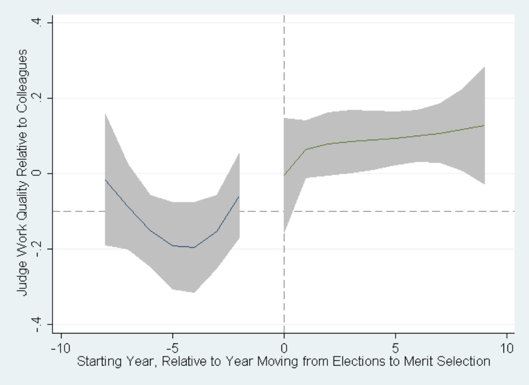
- Left panel: Partisan Elections, Right panel: Non-Partisan Elections



Fractional-polynomial prediction plots with y = outcomes and x = years before and after election year; outcomes residualized on judge and year fixed effects and standardized by judge; gray bars give 95% confidence intervals.

Effect of Merit-Selection Reform on Work Quality

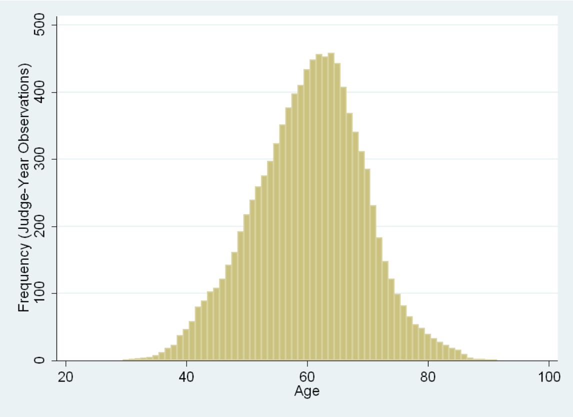
- Quality of judges, residualized on state-year fixed effects, plotted by starting year, relative to merit reform:



Fractional-polynomial prediction plots with y = judge quality and x = judge starting year - reform year; outcomes residualized on state×year fixed effects and standardized by state×year; gray bars give 95% confidence intervals.

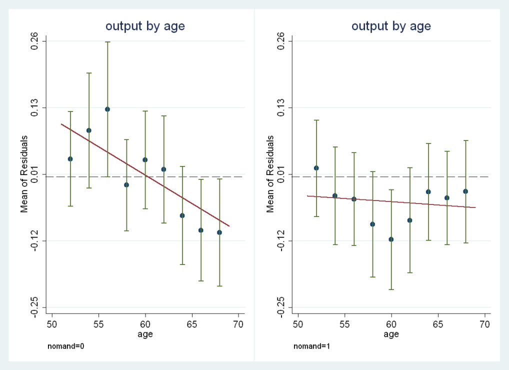
- Judges selected after the reform write higher-quality decisions than judges selected before the reform.

Judge Age Distribution



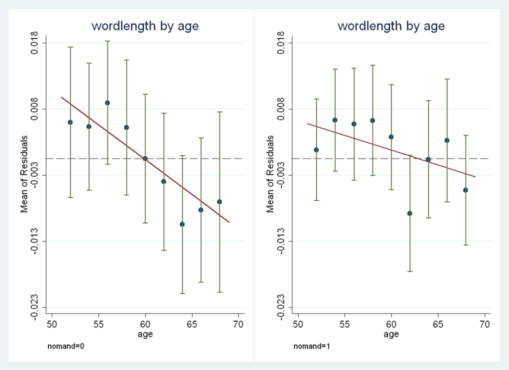
- State supreme court judges have a wide age range but all do the same work task.

Judge Age and Output



- Judge output decreases with age, but only under mandatory retirement (left panel).
    - Consistent with an incentive rather than physiological effect on productivity.

Characters-per-Word and Judge Age



- Older judges use shorter words (fewer characters per word).

Overview of Dictionary-Based Methods

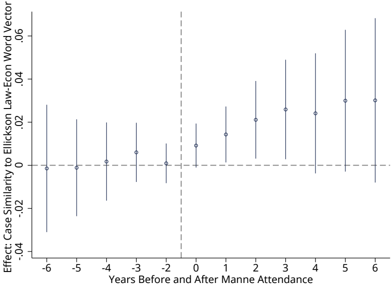
- Dictionary-based text methods use a pre-selected list of words or phrases to analyze a corpus.
- Three major categories:
  - Corpus-specific (e.g., number of times a judge says “justice” vs “efficiency”)
  - General (e.g. LIWC)
  - Sentiment Analysis

Corpus-specific words

- Sometimes counting sets of words or phrases across documents can provide useful evidence.
- Ash, Chen, and Naidu (2017):
  - We analyze the use of economics reasoning in the judiciary.
  - For example, use of the word “efficiency” or “deterrence” after attending a two-week intensive summer course in economics.

Impact of Economics Training on Economics Language

efficiency
practicality
utility
deterrence
optimal
productivity
efficient
economical
economics
betterment
cost
prosperity
predictability
standardization
stability
efficiencies
tranquility
practicability
worthwhile
incremental
affordable
economies
transparency
desirability
infrastructure
openness
interbrand
objective
flexibility
liquidity
marketplace
institutional
desirable
economy
uniformity
profitability
capital
parity
attractiveness
attractive
expediency
fairness
convenience
evenhanded
deterrent
excellence
tourism
industrial
achievement
energy
ideal
competitiveness
diversification
durability
equality
improvement
orderly
innovative
federalism
beneficial
efficacy
simplicity
morale
accessibility
market
Efficiency
practicalities
robust
rehabilitating
overall
welfare
dependable
safety
conveniences
workable
economic
usefulness
rehabilitation
expeditious
necessities
objectivity
capitalist
economically



After attendance, Economics Trained Judges increase use of a selection of terms related to law and economics

Measuring uncertainty in macroeconomy

- Baker, Bloom, and Davis measure economic policy uncertainty using Boolean search of newspaper articles. (See `http://www.policyuncertainty.com/`).
- For each paper on each day since 1985, submit the following query:
  - 1. Article contains “uncertain” OR “uncertainty”, AND
  - 2. Article contains “economic” OR “economy”, AND
  - 3. Article contains “congress” OR “deficit” OR “federal reserve” OR “legislation” OR “regulation” OR “white house”
- Normalize resulting article counts by total newspaper articles that month.

Measuring uncertainty in macroeconomy

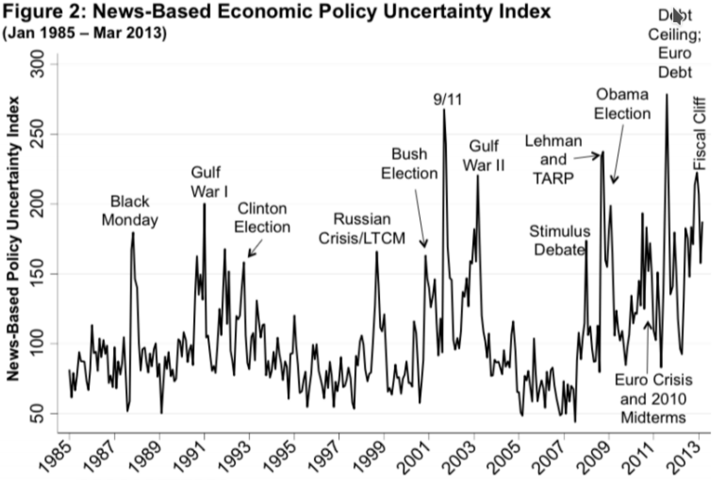
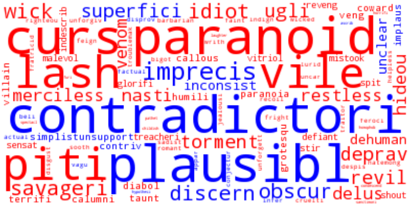


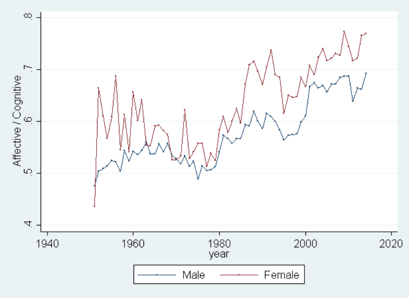
Figure 2: News-Based Economic Policy Uncertainty Index
(Jan 1985 – Mar 2013)

LIWC

- LIWC (pronounced “Luke”) stands for Linguistic Inquiry and Word Counts
    - Info and publications at `liwc.net`
    - Invented in 1980s, now in third version
- Word List Poster: `http://elliottash.com/wp-content/uploads/2017/07/LIWC2015-dictionary-poster.pdf`

Emotive vs. Cognitive Processing in U.S. Congress





Source: Gennaro, Ash, and Loewen (2019)

Sentiment Analysis in Python

- The vader class in nltk provides positive, negative, and neutral scores for a document, and a composite score that combines all three.
  - vader works best on raw text – capitalization and punctuation are used in the calculus.
- Designed for online writing – hard to say how well it works on legal text, for example.
  - Hamilton-Clark-Leskovec-Jurafsky (2016) provide a method for making domain-specific sentiment lexicons using word embeddings (more on this later).

Limitations of sentiment analysis

I’d hate to be the president

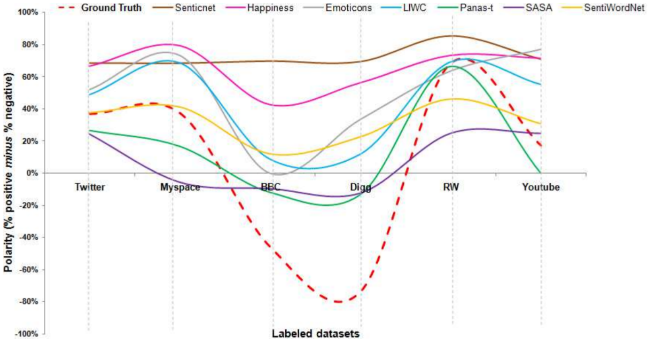


Figure 2: Polarity of the eight sentiment methods across the labeled datasets, indicating that existing methods vary widely in their agreement.