Political Science 452: Text as Data

Justin Grimmer

Assistant Professor
Department of Political Science
Stanford University

April 6th, 2011

Text and Political Science

A pre-2000's view of text in political science

- Political conflict often occurs in texts

Text and Political Science

A pre-2000's view of text in political science

- Political conflict often occurs in texts
- Political Scientists avoided studying texts/speech

Text and Political Science

A pre-2000's view of text in political science

- Political conflict often occurs in texts
- Political Scientists avoided studying texts/speech
- Why?

Text and Political Science

A pre-2000's view of text in political science

- Political conflict often occurs in texts
- Political Scientists avoided studying texts/speech
- Why?
    - Hard to find

Text and Political Science

A pre-2000's view of text in political science

- Political conflict often occurs in texts
- Political Scientists avoided studying texts/speech
- Why?
  - Hard to find
  - Time Consuming

Text and Political Science

A pre-2000's view of text in political science

- Political conflict often occurs in texts
- Political Scientists avoided studying texts/speech
- Why?
  - Hard to find
  - Time Consuming
  - Not generalizable (each new data set...new coding scheme)

Text and Political Science

A pre-2000's view of text in political science

- Political conflict often occurs in texts
- Political Scientists avoided studying texts/speech
- Why?
  - Hard to find
  - Time Consuming
  - Not generalizable (each new data set...new coding scheme)
  - Difficult to store/search

Text and Political Science

A pre-2000's view of text in political science

- Political conflict often occurs in texts
- Political Scientists avoided studying texts/speech
- Why?
  - Hard to find
  - Time Consuming
  - Not generalizable (each new data set...new coding scheme)
  - Difficult to store/search
  - Idiosyncratic to coders/researcher

A post-2000's view of text in political science:

A post-2000's view of text in political science:

Massive collections of texts are increasingly used as a data source in political science:

A post-2000's view of text in political science:

Massive collections of texts are increasingly used as a data source in political science:

American Politics

- Policy Agendas Project
    - Congressional Bills Project
- LYDIA

Comparative Politics

- Legislative Speech Project
- Comparative Manifesto Project

International Relations

- Penn State Event Data Project

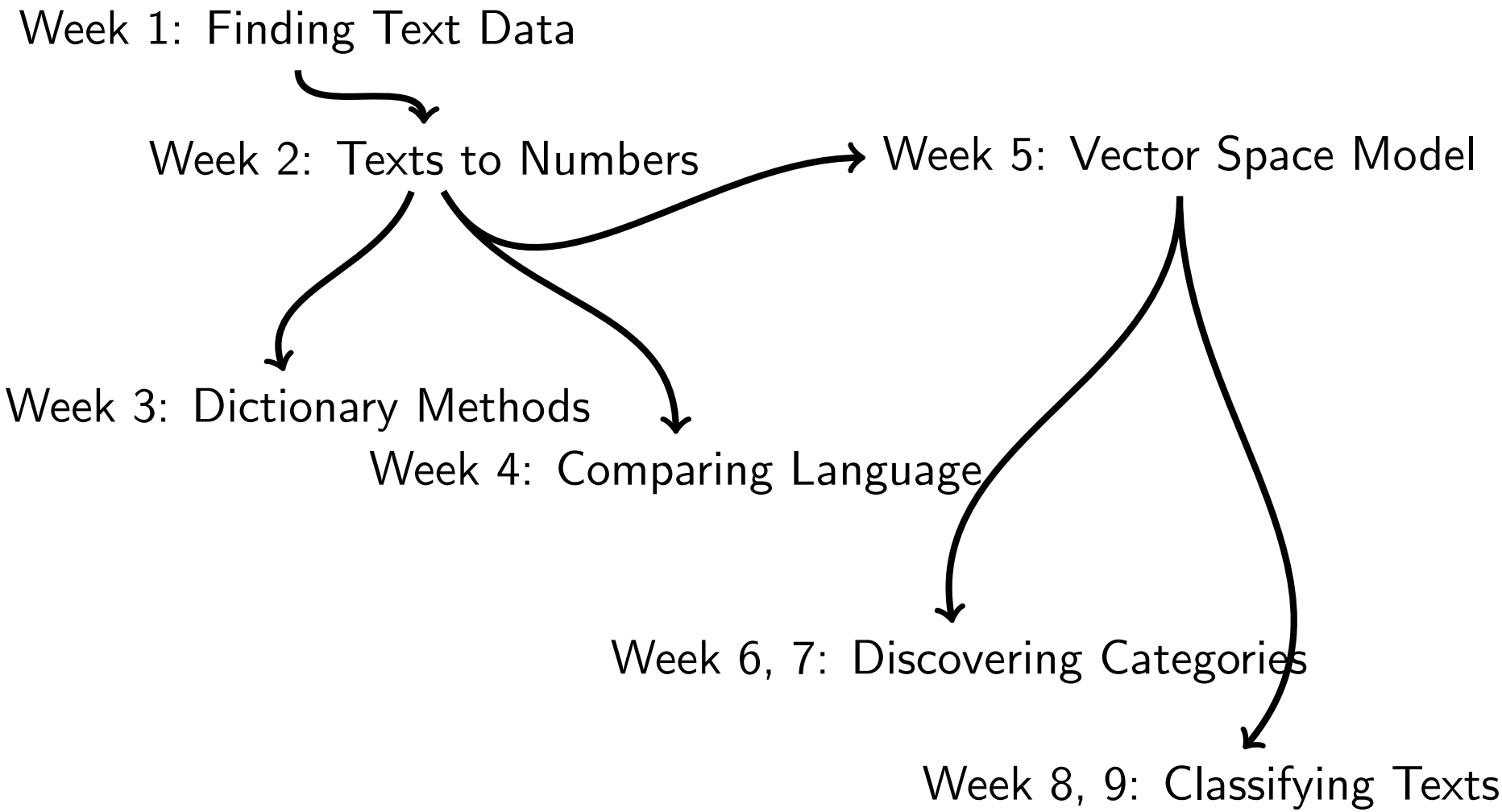
Why?

- Massive increase in availability of unstructured text (10 minutes of worldwide email = 1 LOC)
- Cheap storage: 1956: $10,000 megabyte. 2011: <<< $0.0001 per megabyte (Unless you're sending an SMS)
- Explosion in methods and programs to analyze texts
  - Generalizable: one method can be used across many methods and to unify collections of texts
  - Systematic:parameters/statistics demonstrate how models make coding decisions
  - Cheap: easily applied to many new collections of texts
- Unchanged Demand: Political conflict is expressed (or occurs over) texts
  - Laws
  - Treaties
  - News media
  - Campaigns
  - Political pundits
  - Petitions
  - Press Releases
  - ...

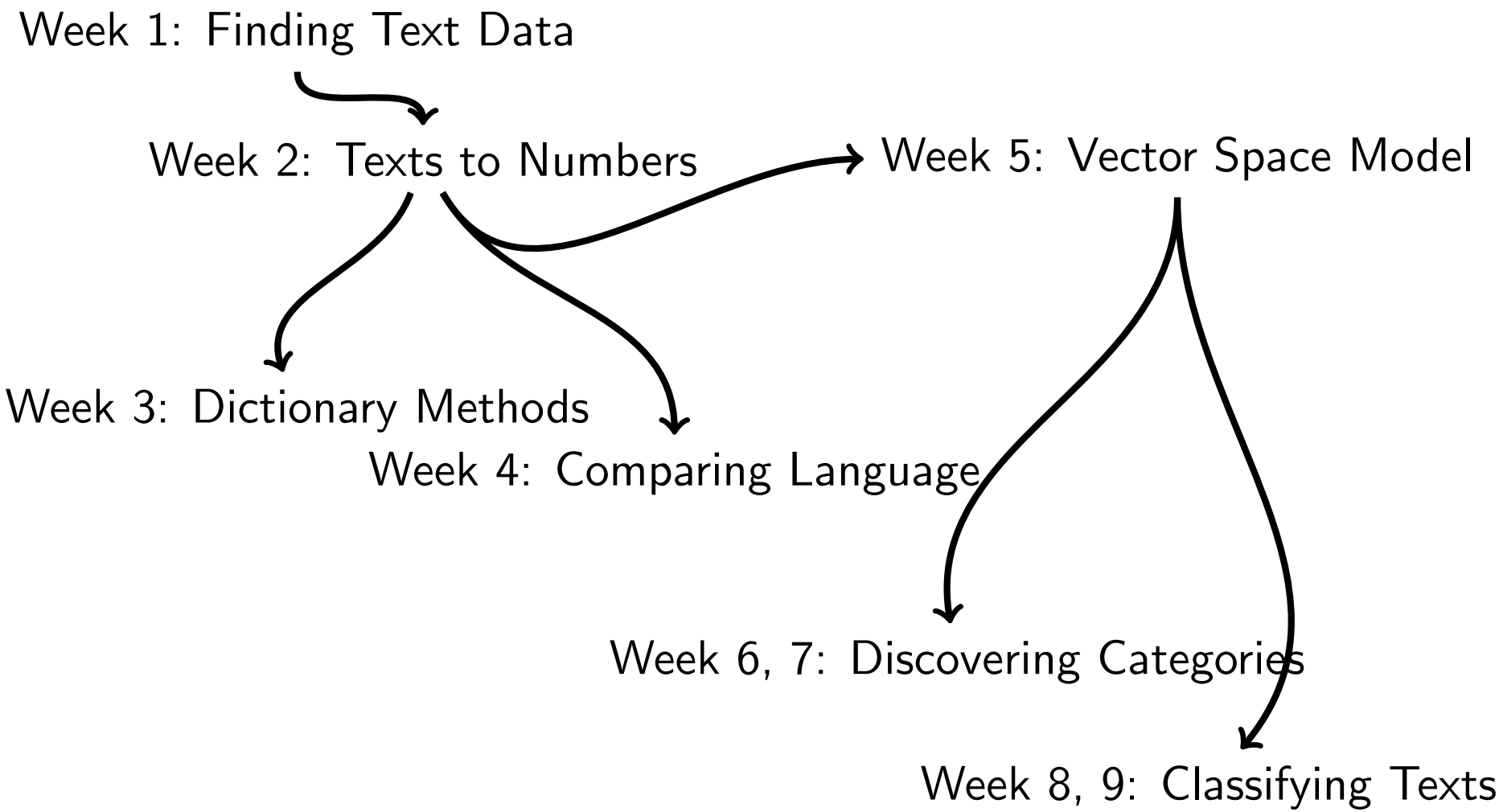
We will study a set of tools for the useful and principled analysis of massive text corpora

1) Methods for inferences about texts using pre-determined words/phrases
2) Methods for comparing language use across groups
3) Methods for discovering new organizations of text
4) Methods for efficiently classifying texts to a predetermined classification scheme

Plan for Course



Plan for Course: Applied Focus



More on What Class is Covering

We're not covering

- Machine Translation
- Word-sense disambiguation
- Collaborative Filtering
- Deep sentence parsing
- The IBM Jeopardy answer machine
- Self-aware machines (SkyNet...)

Examples/Methods draw heavily from what I find useful in my research

Enrolled Students and Motivated Auditors

Two components of evaluation:

1) 50%: Weekly assignments (on your own data)
2) 50%: A final paper analyzing text (broadly defined) using techniques from the class

Analyzing Text Can Be Hard (It Ain't Magic)

Two simple problems: identify words and sentences in the following text

Analyzing Text Can Be Hard (It Ain't Magic)

Two simple problems: identify words and sentences in the following text

```
At least $53 million in federal funds have gone to ACORN
activists since 1994, and the controversial group could get
up to $8.5 billion more tax dollars despite being under
investigation for voter registration fraud in a dozen
states.  The economic stimulus bill enacted in February
contains $3 billion that the non-profit activist group
known more formally as the Association for Community
Organizations for Reform Now could receive, and 2010
federal budget contains another $5.5 billion that could
also find its way into the group's coffers...  A
downloadable spreadsheet of the $53 million is posted on
washingtonexaminer.com.  Scott Levenson, ACORN's national
spokesman, said "we have received no significant federal
funding."  Michelle Bachmann (R-MN)
```

Analyzing Text Can Be Surprisingly Easy (It can seem magical)

(Lamar Alexander (R-TN) Feb 10, 2005)

Word	No. Times Used in Press Release
department	12
grant	9
program	7
firefight	7
secure	5
homeland	4
fund	3
award	2
safety	2
service	2
AFGP	2
support	2
equip	2
applaud	2

What Can Text Methods Do?

Haystack metaphor:

What Can Text Methods Do?

Haystack metaphor:Improve Reading

What Can Text Methods Do?

Haystack metaphor:Improve Reading

- Interpreting the meaning of a sentence or phrase $\rightsquigarrow$ Analyzing a straw of hay

What Can Text Methods Do?

Haystack metaphor:Improve Reading

- Interpreting the meaning of a sentence or phrase $\rightsquigarrow$ Analyzing a straw of hay
  - Humans: amazing (Straussian political theory, analysis of English poetry)
  - Computers: struggle

What Can Text Methods Do?

Haystack metaphor:Improve Reading

- Interpreting the meaning of a sentence or phrase $\rightsquigarrow$ Analyzing a straw of hay
  - Humans: amazing (Straussian political theory, analysis of English poetry)
  - Computers: struggle
- Comparing, Organizing, and Classifying Texts$\rightsquigarrow$ Organizing hay stack

What Can Text Methods Do?

Haystack metaphor:Improve Reading

- Interpreting the meaning of a sentence or phrase $\rightsquigarrow$ Analyzing a straw of hay
  - Humans: amazing (Straussian political theory, analysis of English poetry)
  - Computers: struggle
- Comparing, Organizing, and Classifying Texts$\rightsquigarrow$ Organizing hay stack
  - Humans: terrible. Tiny active memories
  - Computers: amazing$\rightsquigarrow$ the subject of this course

What Can Text Methods Do?

Haystack metaphor:Improve Reading

- Interpreting the meaning of a sentence or phrase $\rightsquigarrow$ Analyzing a straw of hay
  - Humans: amazing (Straussian political theory, analysis of English poetry)
  - Computers: struggle
- Comparing, Organizing, and Classifying Texts$\rightsquigarrow$ Organizing hay stack
  - Humans: terrible. Tiny active memories
  - Computers: amazing$\rightsquigarrow$ the subject of this course

What we won't do:

What Can Text Methods Do?

Haystack metaphor:Improve Reading

- Interpreting the meaning of a sentence or phrase $\rightsquigarrow$ Analyzing a straw of hay
  - Humans: amazing (Straussian political theory, analysis of English poetry)
  - Computers: struggle
- Comparing, Organizing, and Classifying Texts$\rightsquigarrow$ Organizing hay stack
  - Humans: terrible. Tiny active memories
  - Computers: amazing$\rightsquigarrow$ the subject of this course

What we won't do:

- Develop a comprehensive statistical model of language
- Replace the need to read
- Develop a single tool + evaluation for all tasks

Small Problems are harder than you think

Manually develop categorization scheme for partitioning small (100) set of documents

Small Problems are harder than you think

Manually develop categorization scheme for partitioning small (100) set of documents

- Bell(n) = number of ways of partitioning n objects

Small Problems are harder than you think

Manually develop categorization scheme for partitioning small (100) set of documents

- Bell(n) = number of ways of partitioning n objects
- Bell(2) = 2 (AB, A B)

Small Problems are harder than you think

Manually develop categorization scheme for partitioning small (100) set of documents

- Bell(n) = number of ways of partitioning n objects
- Bell(2) = 2 (AB, A B)
- Bell(3) = 5 (ABC, AB C, A BC, AC B, A B C)

Small Problems are harder than you think

Manually develop categorization scheme for partitioning small (100) set of documents

- Bell(n) = number of ways of partitioning n objects
- Bell(2) = 2 (AB, A B)
- Bell(3) = 5 (ABC, AB C, A BC, AC B, A B C)
- Bell(5) = 52

Small Problems are harder than you think

Manually develop categorization scheme for partitioning small (100) set of documents

- Bell(n) = number of ways of partitioning n objects
- Bell(2) = 2 (AB, A B)
- Bell(3) = 5 (ABC, AB C, A BC, AC B, A B C)
- Bell(5) = 52
- Bell(100)

Small Problems are harder than you think

Manually develop categorization scheme for partitioning small (100) set of documents

- Bell(n) = number of ways of partitioning n objects
- Bell(2) = 2 (AB, A B)
- Bell(3) = 5 (ABC, AB C, A BC, AC B, A B C)
- Bell(5) = 52
- Bell(100)$\approx 4.75 \times 10^{115}$ partitions

Small Problems are harder than you think

Manually develop categorization scheme for partitioning small (100) set of documents

- Bell(n) = number of ways of partitioning n objects
- Bell(2) = 2 (AB, A B)
- Bell(3) = 5 (ABC, AB C, A BC, AC B, A B C)
- Bell(5) = 52
- Bell(100)$\approx 4.75 \times 10^{115}$ partitions
- Big Number:

Small Problems are harder than you think

Manually develop categorization scheme for partitioning small (100) set of documents

- Bell(n) = number of ways of partitioning n objects
- Bell(2) = 2 (AB, A B)
- Bell(3) = 5 (ABC, AB C, A BC, AC B, A B C)
- Bell(5) = 52
- Bell(100)$\approx 4.75 \times 10^{115}$ partitions
- Big Number:
  7 Billion RAs

Small Problems are harder than you think

Manually develop categorization scheme for partitioning small (100) set of documents

- Bell(n) = number of ways of partitioning n objects
- Bell(2) = 2 (AB, A B)
- Bell(3) = 5 (ABC, AB C, A BC, AC B, A B C)
- Bell(5) = 52
- Bell(100)$\approx 4.75 \times 10^{115}$ partitions
- Big Number:
  7 Billion RAs
  Impossibly Fast (enumerate one clustering every millisecond)

Small Problems are harder than you think

Manually develop categorization scheme for partitioning small (100) set of documents

- Bell(n) = number of ways of partitioning n objects
- Bell(2) = 2 (AB, A B)
- Bell(3) = 5 (ABC, AB C, A BC, AC B, A B C)
- Bell(5) = 52
- Bell(100)$\approx 4.75 \times 10^{115}$ partitions
- Big Number:
  7 Billion RAs
  Impossibly Fast (enumerate one clustering every millisecond)
  Working around the clock (24/7/365)

Small Problems are harder than you think

Manually develop categorization scheme for partitioning small (100) set of documents

- Bell(n) = number of ways of partitioning n objects
- Bell(2) = 2 (AB, A B)
- Bell(3) = 5 (ABC, AB C, A BC, AC B, A B C)
- Bell(5) = 52
- Bell(100)$\approx 4.75 \times 10^{115}$ partitions
- Big Number:
  7 Billion RAs
  Impossibly Fast (enumerate one clustering every millisecond)
  Working around the clock (24/7/365)
  $\approx 1.54 \times 10^{84} \times$

Small Problems are harder than you think

Manually develop categorization scheme for partitioning small (100) set of documents

- Bell(n) = number of ways of partitioning n objects
- Bell(2) = 2 (AB, A B)
- Bell(3) = 5 (ABC, AB C, A BC, AC B, A B C)
- Bell(5) = 52
- Bell(100)$\approx 4.75 \times 10^{115}$ partitions
- Big Number:
  7 Billion RAs
  Impossibly Fast (enumerate one clustering every millisecond)
  Working around the clock (24/7/365)
  $\approx 1.54 \times 10^{84} \times (14,000,000,000)$

Small Problems are harder than you think

Manually develop categorization scheme for partitioning small (100) set of documents

- Bell(n) = number of ways of partitioning n objects
- Bell(2) = 2 (AB, A B)
- Bell(3) = 5 (ABC, AB C, A BC, AC B, A B C)
- Bell(5) = 52
- Bell(100)$\approx 4.75 \times 10^{115}$ partitions
- Big Number:
  7 Billion RAs
  Impossibly Fast (enumerate one clustering every millisecond)
  Working around the clock (24/7/365)
  $\approx 1.54 \times 10^{84} \times (14,000,000,000)$ years

Small Problems are harder than you think

Manually develop categorization scheme for partitioning small (100) set of documents

- Bell(n) = number of ways of partitioning n objects
- Bell(2) = 2 (AB, A B)
- Bell(3) = 5 (ABC, AB C, A BC, AC B, A B C)
- Bell(5) = 52
- Bell(100)$\approx 4.75 \times 10^{115}$ partitions
- Big Number:
  7 Billion RAs
  Impossibly Fast (enumerate one clustering every millisecond)
  Working around the clock (24/7/365)
  $\approx 1.54 \times 10^{84} \times (14,000,000,000)$ years

Automated methods can help with even small problems

A Brief Digression on Computing Languages

- Texts processing is active across many computer languages
- Goal: useful tools in `R`
- Reality: if you want automated texts to occupy a central part of research, you need to know a little
  - `HTML`
  - `Python` (or `PERL`)
  - If you knew `JAVA` or `C`, you'd be a step ahead
- We'll talk about how to hire programmers to eliminate language gap
- I'll post `python` code on course site [note: Kludgey Python Code]

Where and How to Find Text Data?

Internet and archives have massive stores of text data (and growing!)

- Prepackaged Data
- Computer and Human intensive Web Scraping
- Archive Materials and Optical Character Recognition

Goal: plan text (.txt) file. (UTF-8, ASCII)

An obvious plan for data acquisition

- Check prepackaged resources
  - Lexis-Nexis (Batch Downloads)
  - Proquest
  - Research Librarians
- Move to web based search
  - Before deciding to scrape a data set:
    - Is the HTML standardized? (Our example today: no [Xtreme webscraping])
    - Does the website allow you to scrape? (Not always)
    - Can you do it faster by hand? With Mturk?
- Archival research
  - Invest in a scanner that allows OCR
  - Before making plans to scan, be sure archives allows scanning

Two examples from prepackaged data sources

Automated Literature Reviews

How do we conduct literature reviews?

Automated Literature Reviews

How do we conduct literature reviews?

- Think about literature (ask graduate student working in area for help)

Automated Literature Reviews

How do we conduct literature reviews?

- Think about literature (ask graduate student working in area for help)
- Make an argument about the deficiency/gaps in that literature

Automated Literature Reviews

How do we conduct literature reviews?

- Think about literature (ask graduate student working in area for help)
- Make an argument about the deficiency/gaps in that literature
- Cite the prominent articles, make an argument about "conventional wisdom" (which is always wrong), call it a day

Automated Literature Reviews

How do we conduct literature reviews?

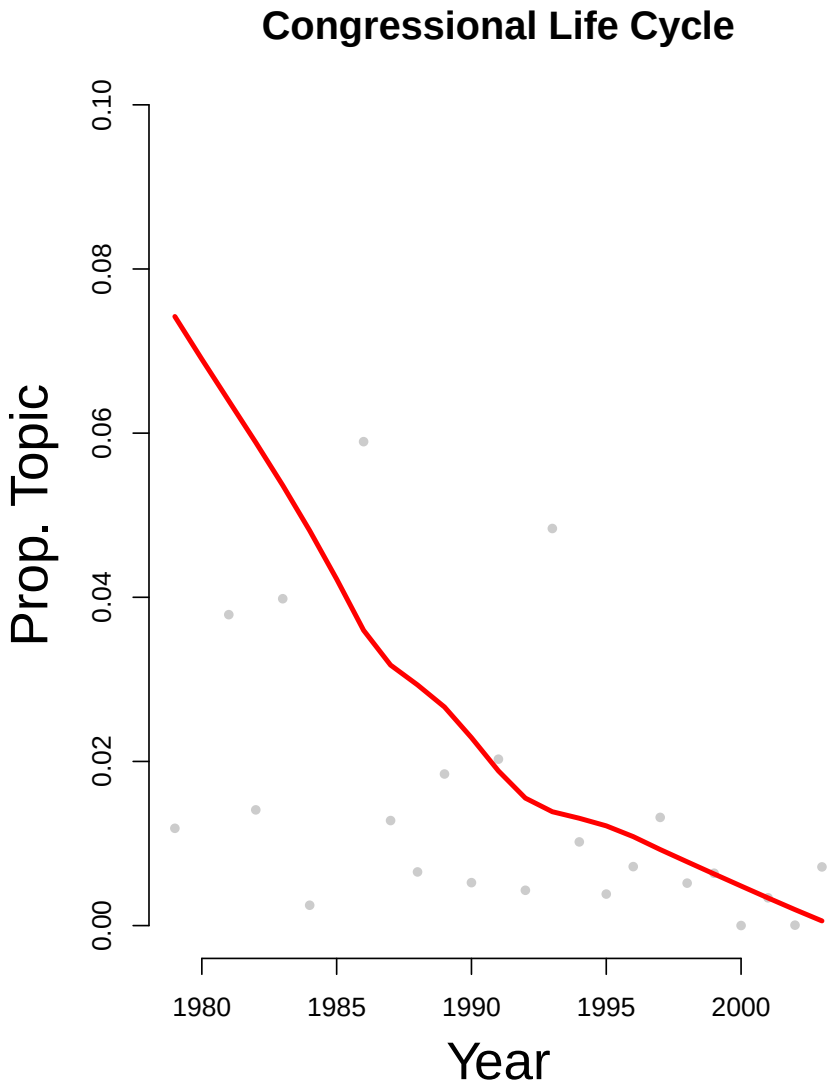
- Think about literature (ask graduate student working in area for help)
- Make an argument about the deficiency/gaps in that literature
- Cite the prominent articles, make an argument about "conventional wisdom" (which is always wrong), call it a day

Literature reviews and analysis of concept development are difficult text problems

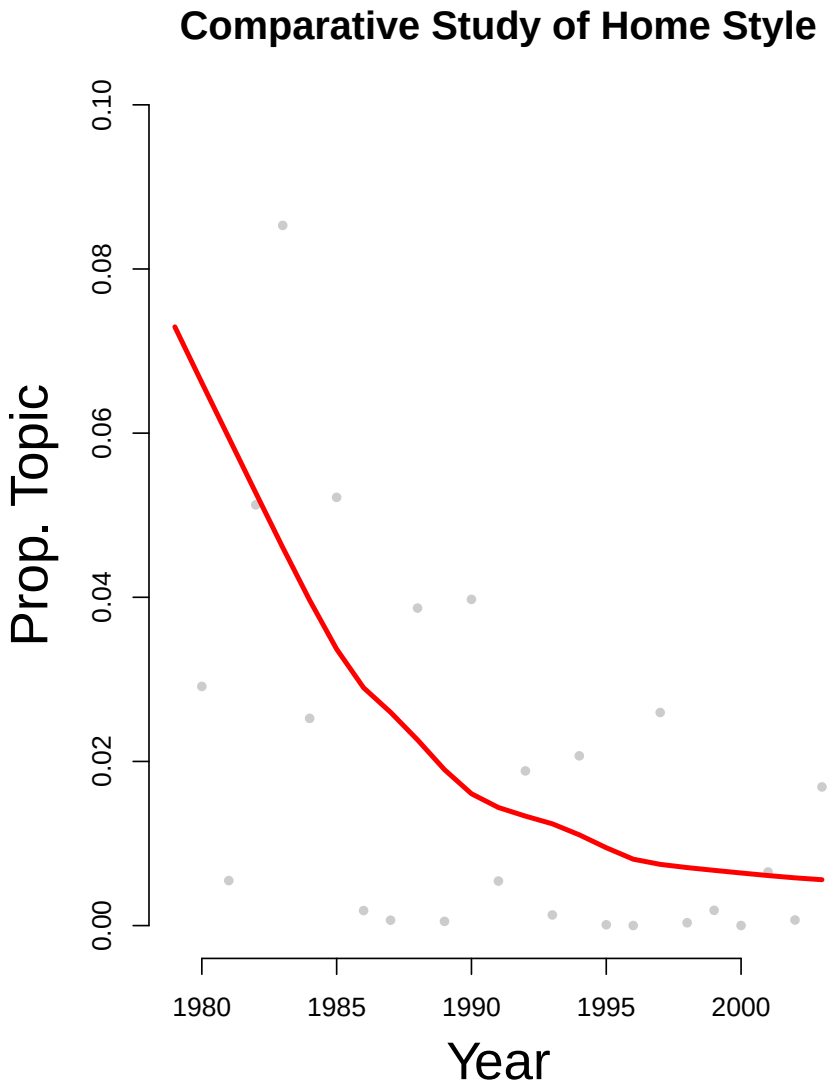
Automated Literature Reviews

JSTOR data, now available for download
`http://dfr.jstor.org`
Live example

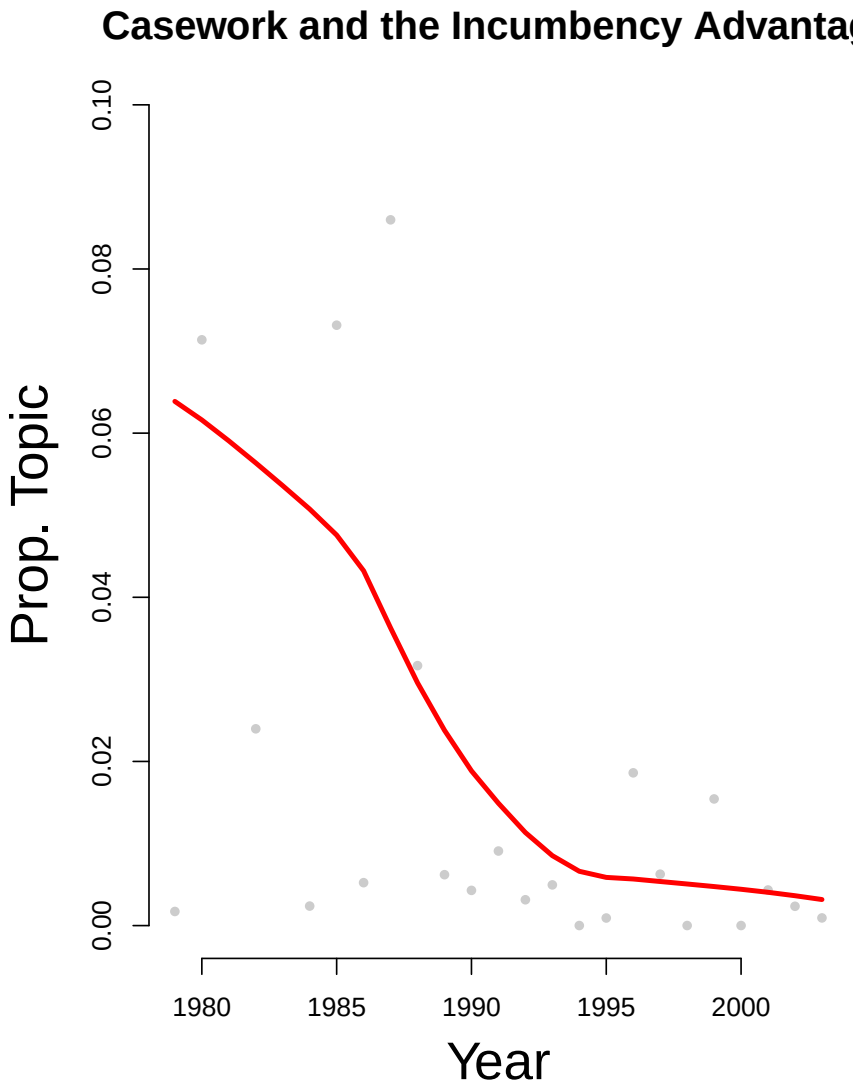
History of Home Style



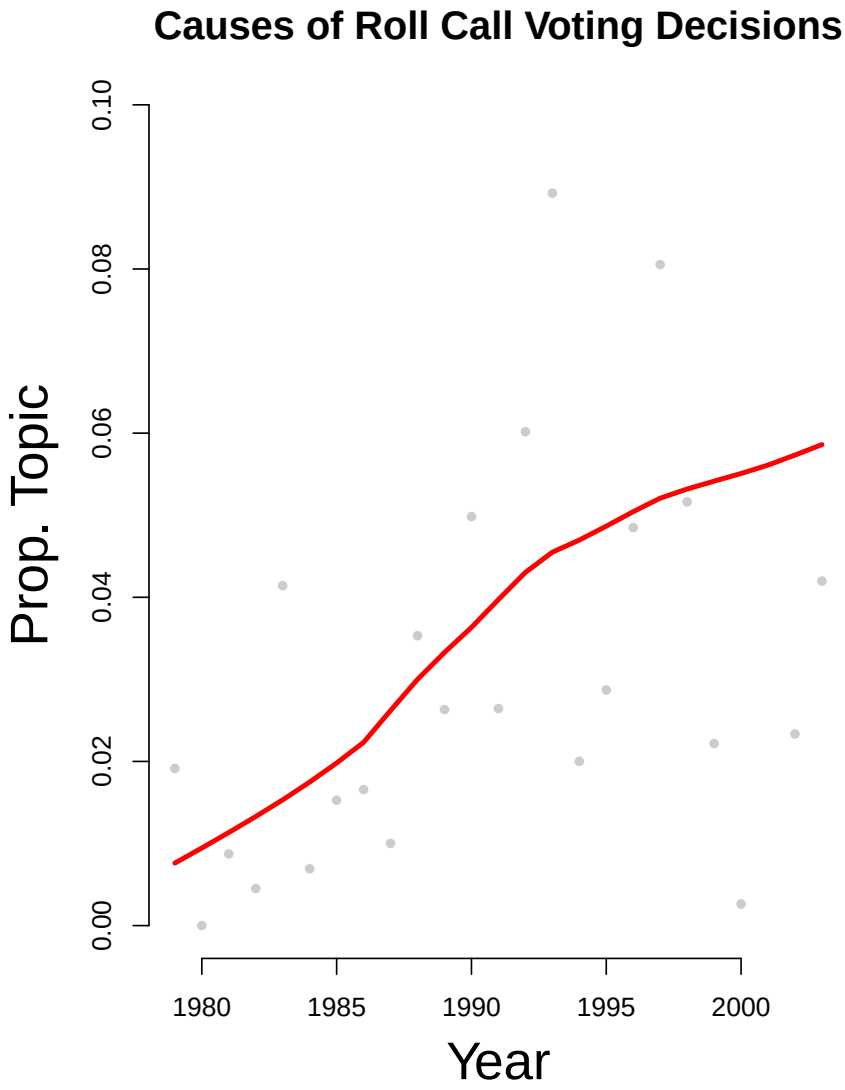
History of Home Style



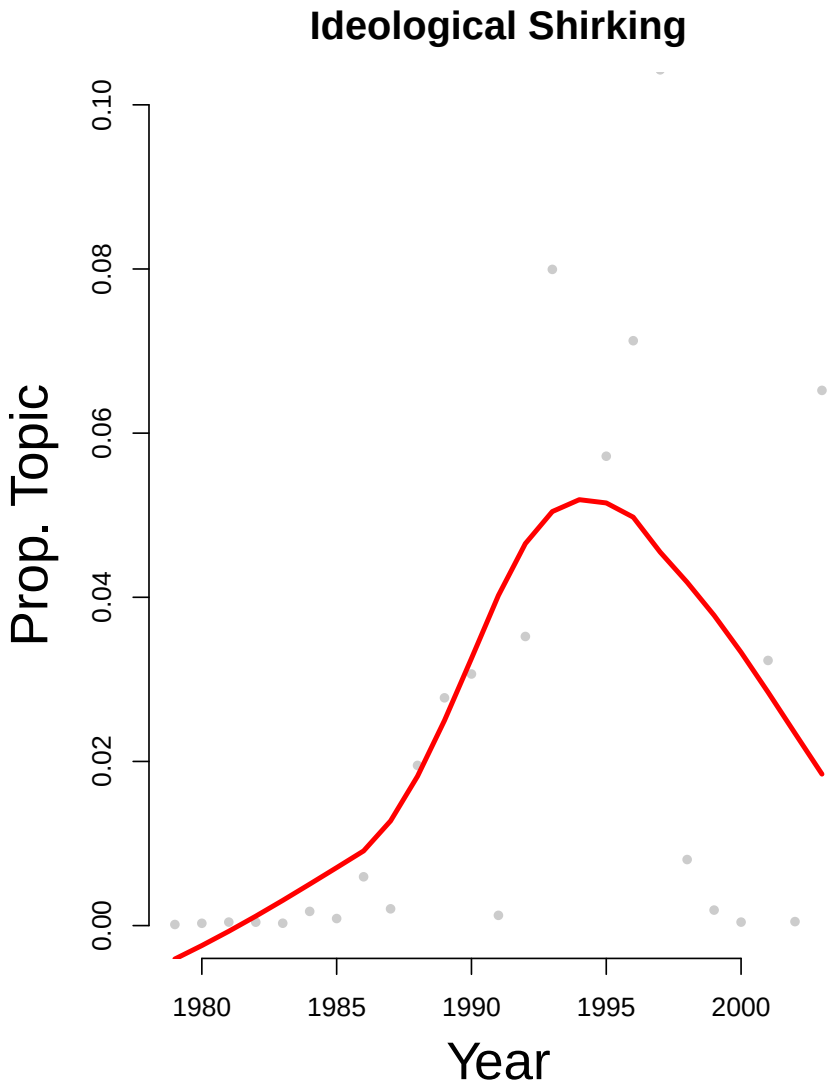
History of Home Style



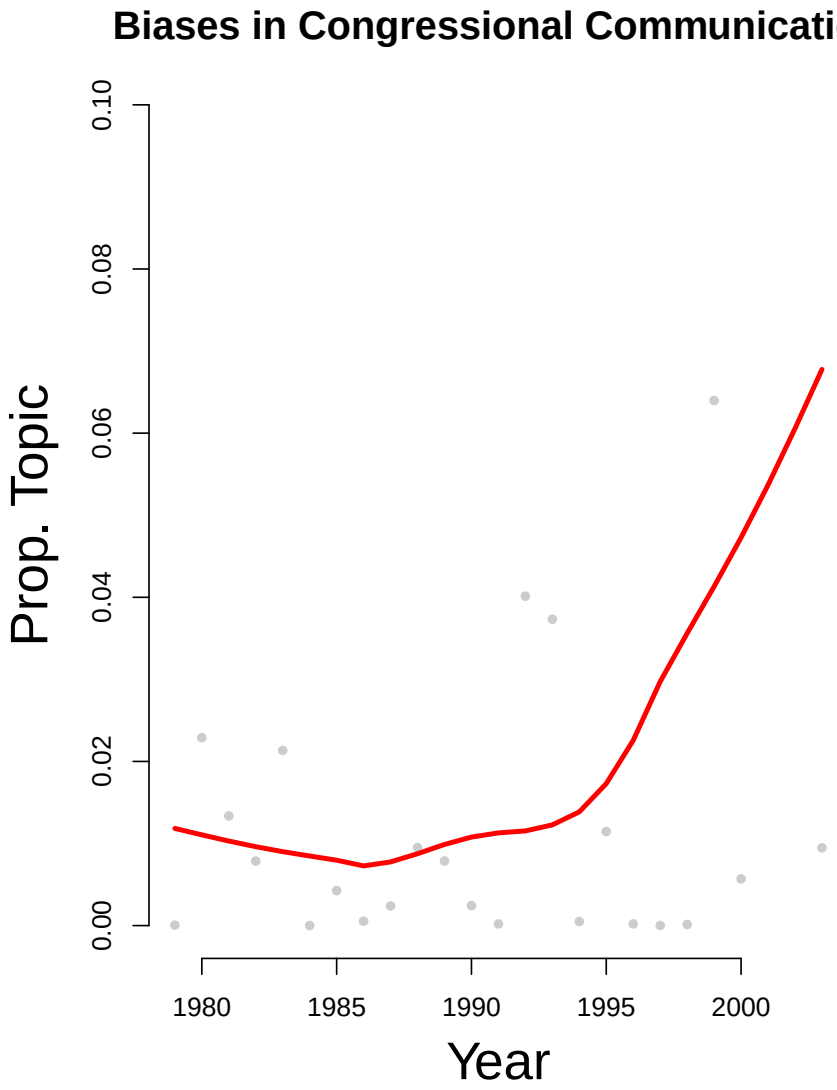
History of Home Style



History of Home Style



History of Home Style



Intellectual History and Google Books

Scholars of political thought:

Live Example, Part 2.

Intellectual History and Google Books

Scholars of political thought: Careful (manual) analysis of texts.

Live Example, Part 2.

Intellectual History and Google Books

Scholars of political thought: Careful (manual) analysis of texts.
Many books now available for download thanks to google

Live Example, Part 2.

Intellectual History and Google Books

Scholars of political thought: Careful (manual) analysis of texts.
Many books now available for download thanks to google

- `books.google.com`
- Advanced Search
- Full view only
- Download ".epub"
- Use a converter

Live Example, Part 2.

Human and Computer Based Web Scraping

Mechanical Turk

Mechanical Turk

- Mechanical Turk is an amazon run marketplace for workers (humans)

Mechanical Turk

- Mechanical Turk is an amazon run marketplace for workers (humans)
- We can replicate this task by asking (bored, poor, bored & poor) workers to do the task

Mechanical Turk

- Mechanical Turk is an amazon run marketplace for workers (humans)
- We can replicate this task by asking (bored, poor, bored & poor) workers to do the task
- Distribute Small Tasks Across Workers

Mechanical Turk

- Mechanical Turk is an amazon run marketplace for workers (humans)
- We can replicate this task by asking (bored, poor, bored & poor) workers to do the task
- Distribute Small Tasks Across Workers

Live Example 3: Paul Tonko (D-NY)

Mechanical Turk

- Mechanical Turk is an amazon run marketplace for workers (humans)
- We can replicate this task by asking (bored, poor, bored & poor) workers to do the task
- Distribute Small Tasks Across Workers

Live Example 3: Paul Tonko (D-NY)

A Brief Introduction to Web Scraping

How do we get other data?

A Brief Introduction to Web Scraping

How do we get other data?

- Web pages are loaded with text data

A Brief Introduction to Web Scraping

How do we get other data?

- Web pages are loaded with text data
- But not necessarily prepared for download

A Brief Introduction to Web Scraping

How do we get other data?

- Web pages are loaded with text data
- But not necessarily prepared for download
- Web scraping:
    - Interact with `html` to extract text from web pages

A Brief Introduction to Web Scraping

How do we get other data?

- Web pages are loaded with text data
- But not necessarily prepared for download
- Web scraping:
  - Interact with `html` to extract text from web pages
- Requires some programming expertise

Xtreme Web Scraping

- Congressional Web Sites and Press Releases
- Web Sites: Mix of professionals, capable amateurs, and horrible html writers
- No coherent structure across websites
- Difficult scraping problem: collecting press releases from a web site

Live Example 4, Paul Tonko (D-NY)

- Identify pages with press releases
- Extract press releases from page

Other Data Sources/Acquisition

Programming/Data Acquisition Help

- ODesk: submit programming tasks to coders (must be very specific)
- Elance: submit many small tasks to dedicated workers (don't mind outsourcing work to India/OK with not paying minimum wage)
- Guru: "World's largest online marketplace"

Conclusion

Today: Introduction and where to get text
Next week: how to represent text quantitatively