Text as Data

Brandon M. Stewart[1]

Department of Government, Harvard University

Camp Resources XX, August 5, 2013

[1] Thanks to Gary King, Justin Grimmer, Rich Nielsen and Molly Roberts for permission to include figures here.

Overview

Two points and an application.

Overview

Two points and an application.

1. Large bodies of text can provide a new source of *data* for social science research.

Overview

Two points and an application.

1. Large bodies of text can provide a new source of *data* for social science research.
2. Big data isn't about the data (it's about the methods).

Overview

Two points and an application.

1. Large bodies of text can provide a new source of *data* for social science research.
2. Big data isn't about the data (it's about the methods).
3. Application to censorship and media control in China.

Overview

Two points and an application.

1. Large bodies of text can provide a new source of *data* for social science research.
2. Big data isn't about the data (it's about the methods).
3. Application to censorship and media control in China.

References

Grimmer Justin and Brandon M. Stewart. (2013) "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts." *Political Analysis*.

Overview

Text as Data

Massive quantities of unstructured text are increasingly available and open up new possibilities.

Text as Data

Massive quantities of unstructured text are increasingly available and open up new possibilities.

1. More Systematic and Replicable Data collection

Text as Data

Massive quantities of unstructured text are increasingly available and open up new possibilities.

1. More Systematic and Replicable Data collection
2. Cheaper to collect!

Text as Data

Massive quantities of unstructured text are increasingly available and open up new possibilities.

1. More Systematic and Replicable Data collection
2. Cheaper to collect!
3. New Quantities of Interest

What Changed?

- Massive increase in availability of unstructured text (10 minutes of worldwide email = 1 LOC)

What Changed?

- 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)

What Changed?

- 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

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$\rightsquigarrow$ and getting better all the time!

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$\rightsquigarrow$ and getting better all the time!

What They Don'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$\rightsquigarrow$ and getting better all the time!

What They Don't Do:

- Develop a comprehensive statistical model of language

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$\rightsquigarrow$ and getting better all the time!

What They Don't Do:

- Develop a comprehensive statistical model of language
- Replace the need to read

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$\rightsquigarrow$ and getting better all the time!

What They Don't Do:

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

How to Use Text as Data

A simple recipe for automated content analysis:

How to Use Text as Data

A simple recipe for automated content analysis:

1. Define the Problem

How to Use Text as Data

A simple recipe for automated content analysis:

1. Define the Problem
2. Find a Text Corpus

How to Use Text as Data

A simple recipe for automated content analysis:

1. Define the Problem
2. Find a Text Corpus
3. Analyze with an Appropriate Model

How to Use Text as Data

A simple recipe for automated content analysis:

1. Define the Problem
2. Find a Text Corpus
3. Analyze with an Appropriate Model
4. Validate and Visualize

How to Use Text as Data

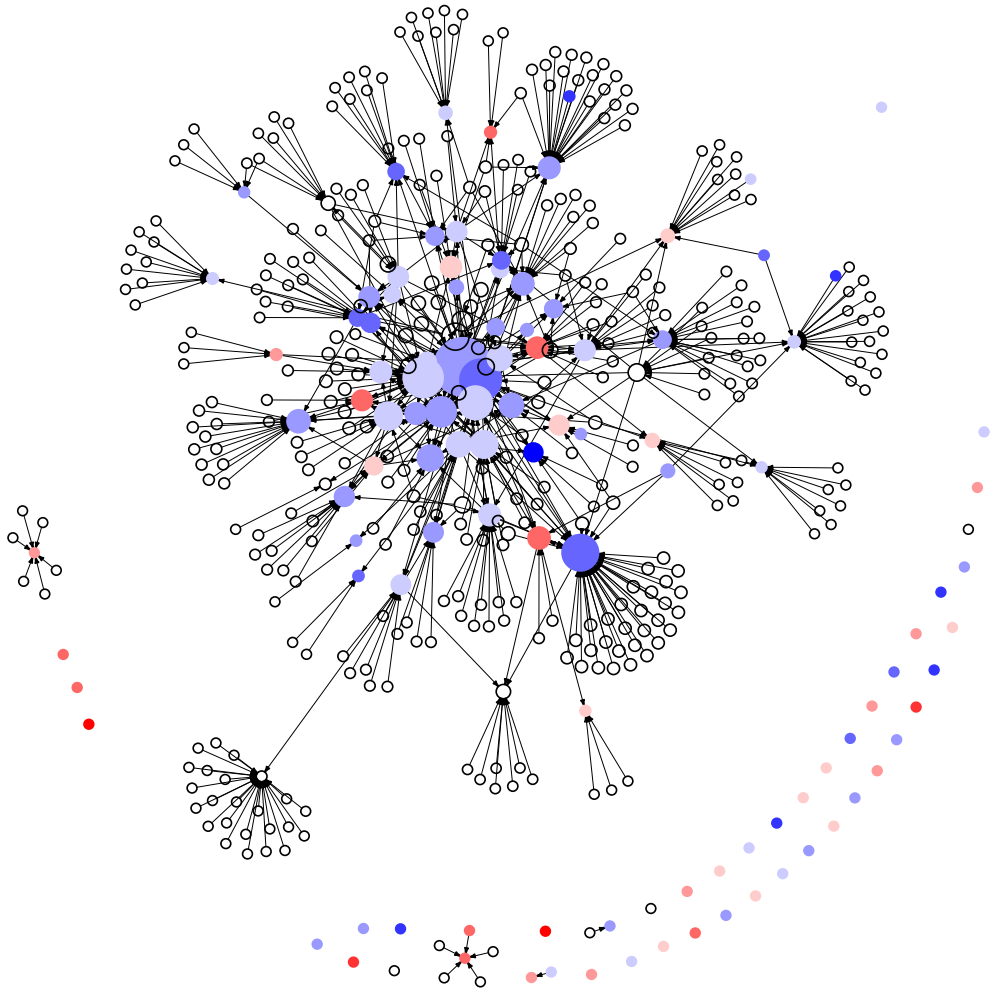
A simple recipe for automated content analysis:

1. Define the Problem
2. Find a Text Corpus
3. Analyze with an Appropriate Model
4. Validate and Visualize

But seeing is believing...

Islamic Clerics and Jihad (Nielsen)

Why do some Islamic Clerics support militant Jihad?

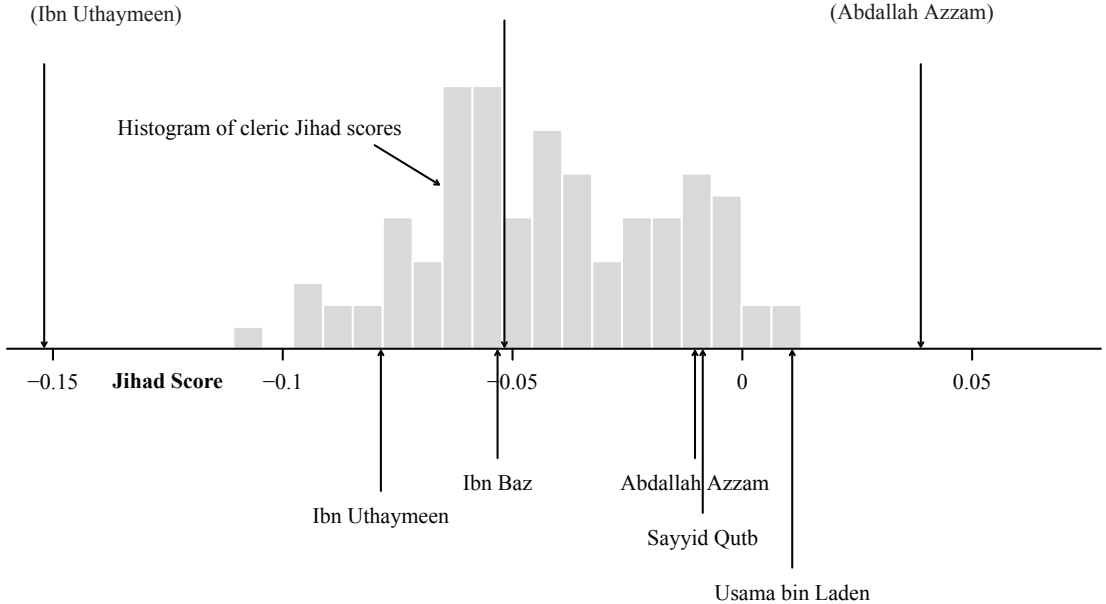


Islamic Clerics and Jihad (Nielsen)

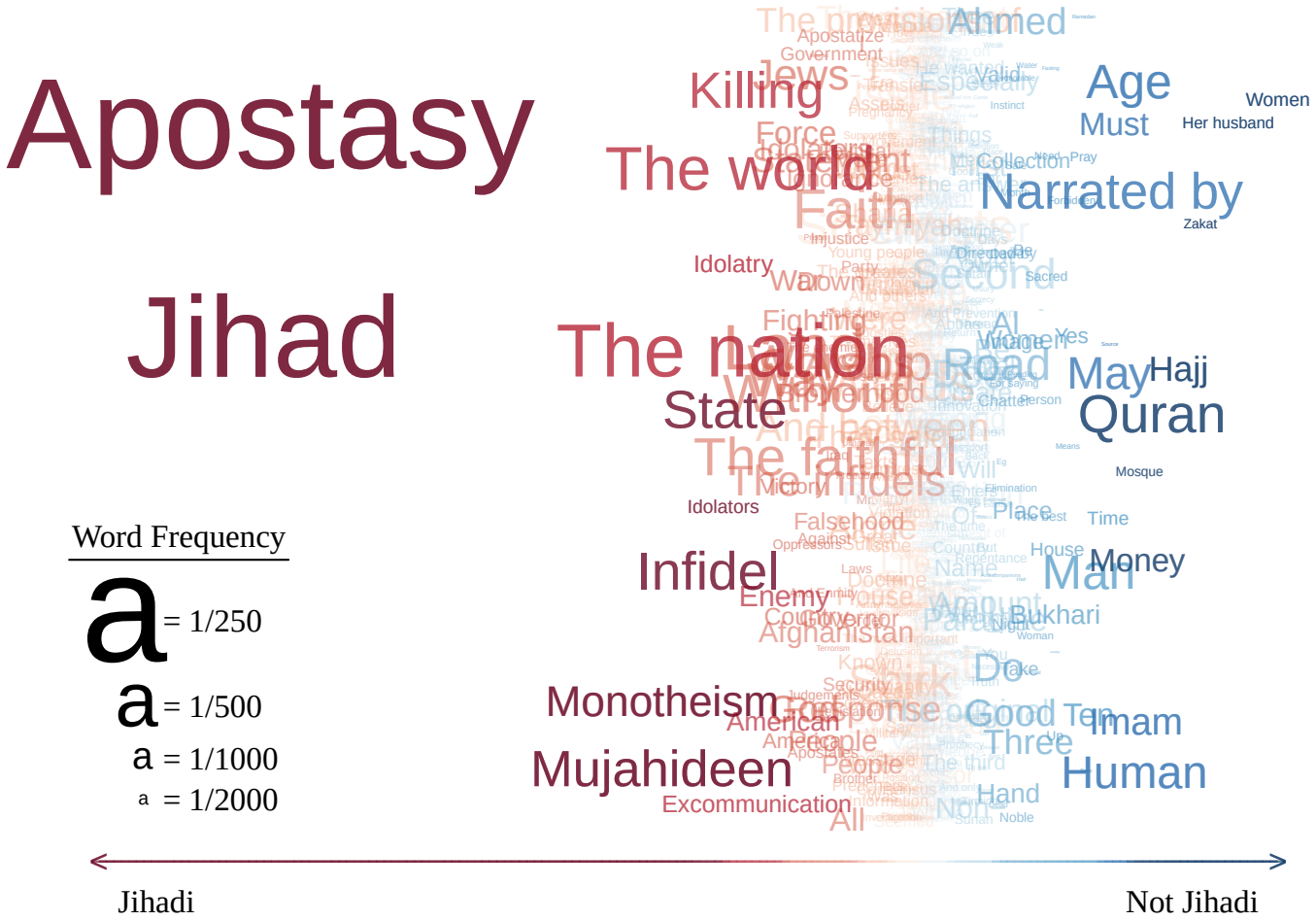
If a person arrives while the Imam is preaching at Friday prayers, he should pray two brief prostrations and sit without greeting anyone as greeting people in this circumstance is forbidden because the Prophet, peace be upon him, says, "If your friend speaks to you during the Friday prayers, silence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

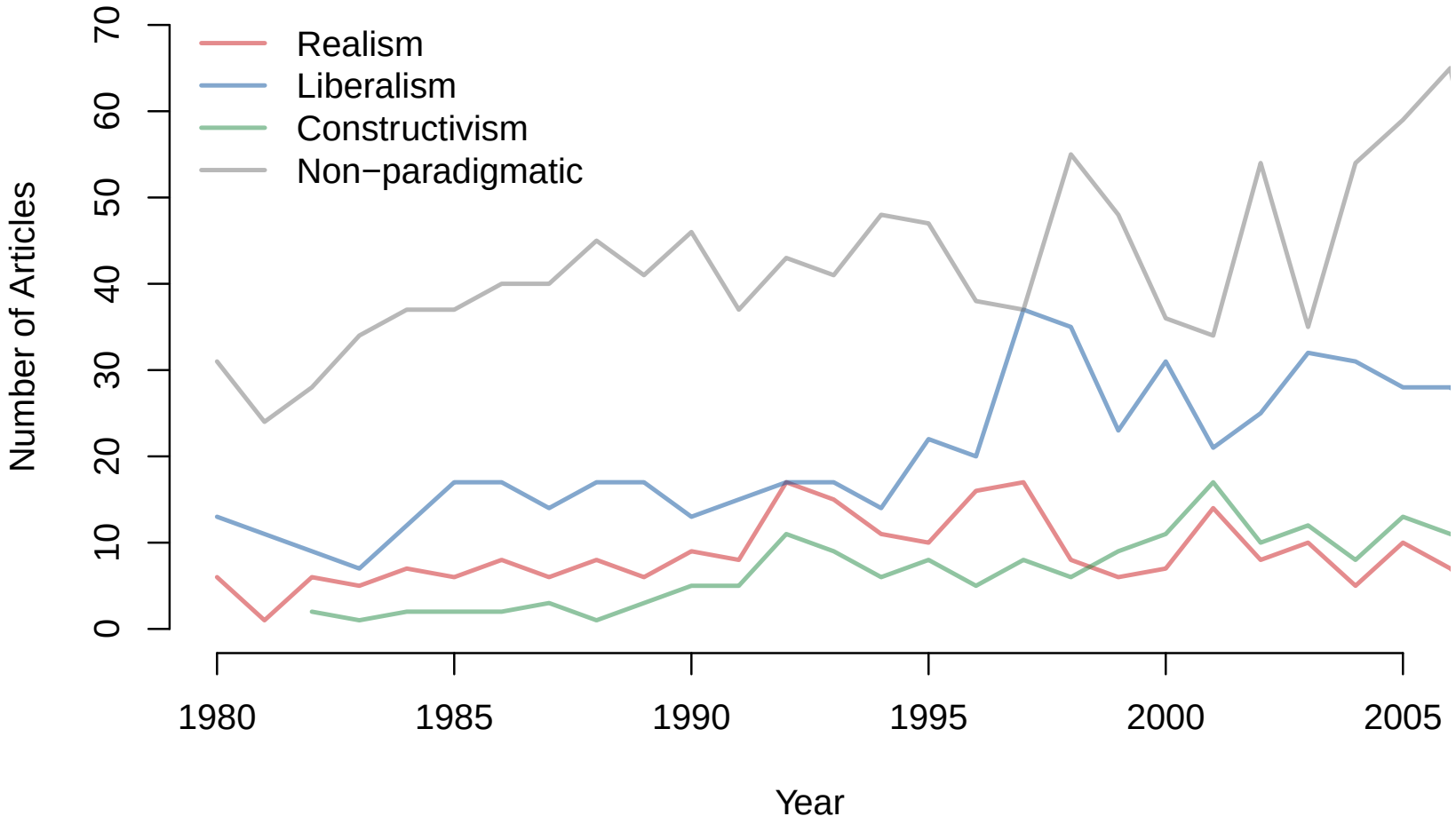


Islamic Clerics and Jihad (Nielsen)



Digital Literature Reviews (Nielsen & Stewart)

The International Relations Literature



Digital Literature Reviews (Nielsen & Stewart)

Realism

Digital Literature Reviews (Nielsen & Stewart)

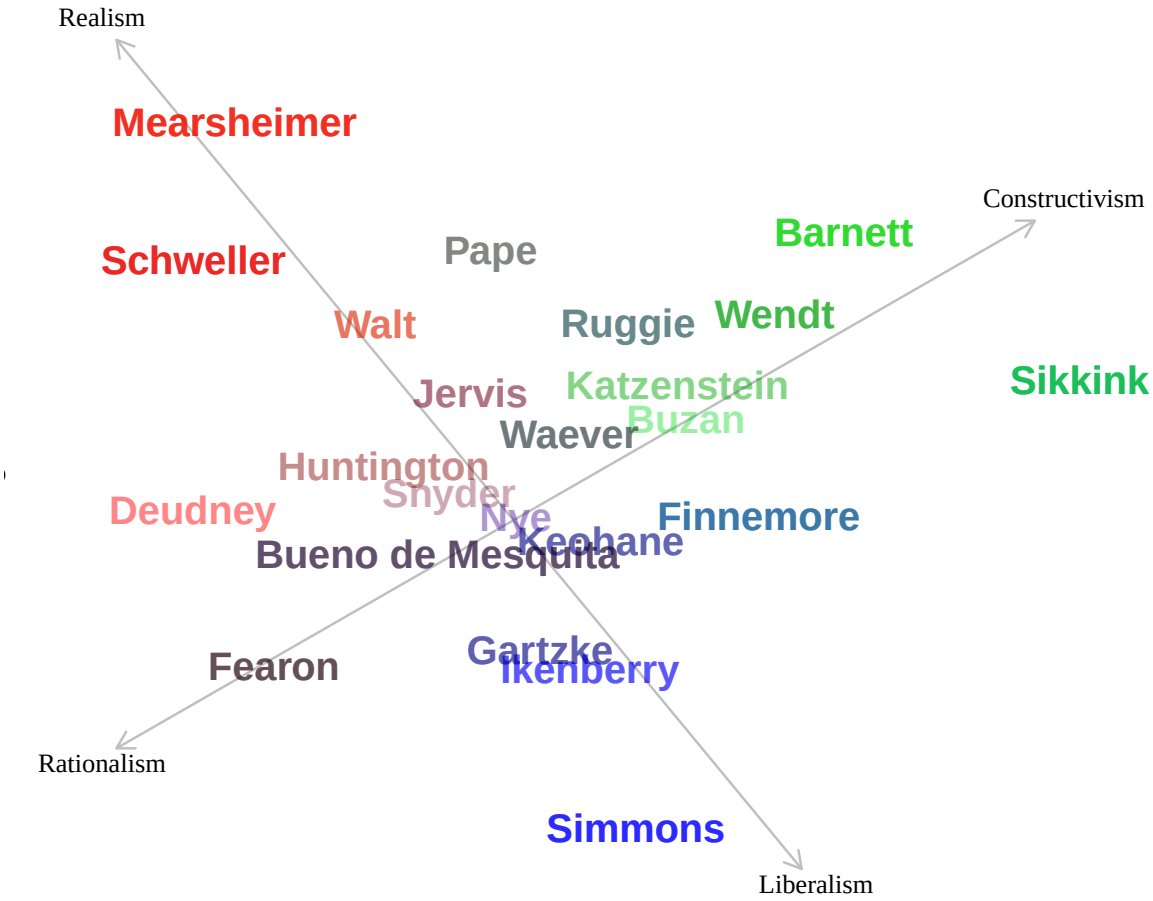
Liberalism

Digital Literature Reviews (Nielsen & Stewart)

Constructivism

Digital Literature Reviews (Nielsen & Stewart)

Who are the big names?



International Events (O'Connor, Stewart & Smith)

Want to code international events of form "someone did something to someone else"

International Events (O'Connor, Stewart & Smith)

Want to code international events of form "someone did something to someone else"

- Original Approach: Manual Coding

International Events (O'Connor, Stewart & Smith)

Want to code international events of form "someone did something to someone else"

- Original Approach: Manual Coding
- First Automated Approach: 15,000 patterns, 200 event classes

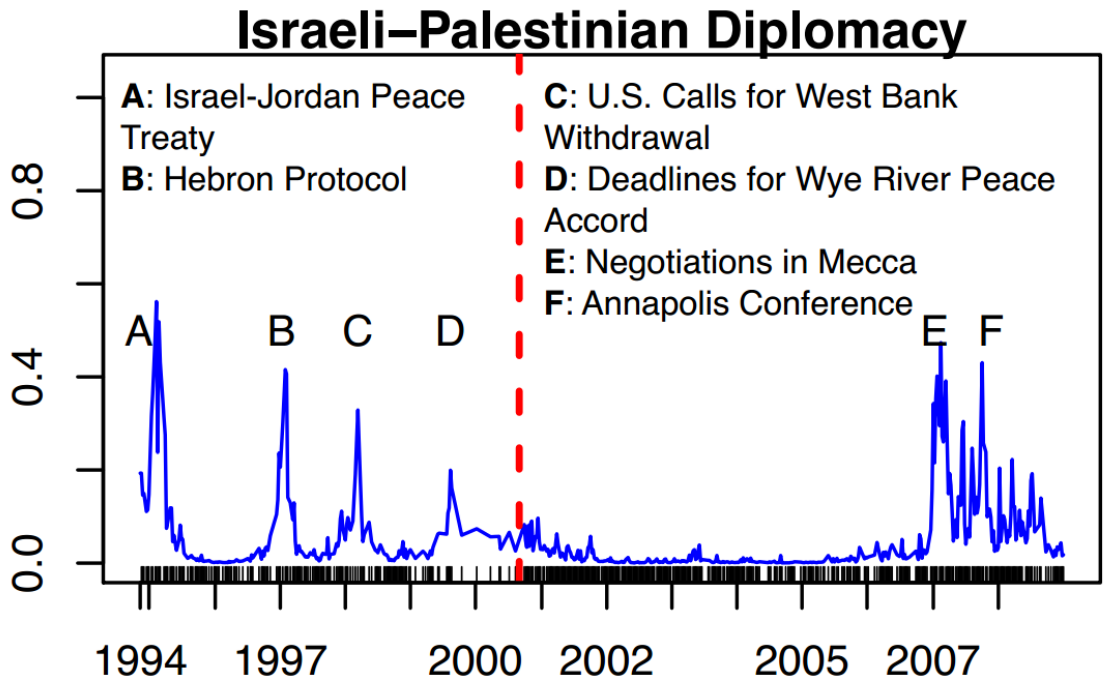
International Events (O'Connor, Stewart & Smith)

Want to code international events of form "someone did something to someone else"

- Original Approach: Manual Coding
- First Automated Approach: 15,000 patterns, 200 event classes
- Our Approach: Learn Event Types from the Text with a Probabilistic Model!

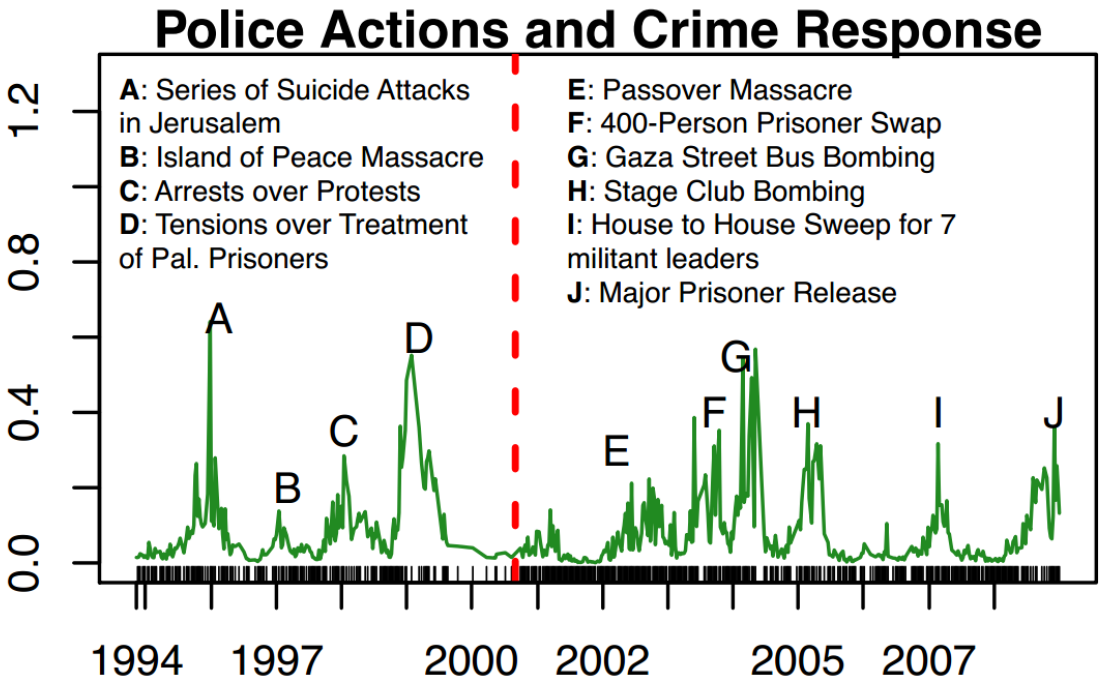
International Events (O'Connor, Stewart & Smith)

meet with, sign with, praise, say with, arrive in, host, tell, welcome, join, thank, meet, travel to, criticize, leave, take to, begin to, begin with, summon, reach with, hold with



International Events (O'Connor, Stewart & Smith)

accuse, criticize, reject, tell, hand to, warn, ask, detain, release, order, deny, arrest, expel, convict, free, extradite to, allow, sign with, charge, urge



Many More Possibilities!

Some other interesting questions addressed with Text Data

Many More Possibilities!

Some other interesting questions addressed with Text Data

- How do civilians differ from military leaders on foreign policy? (Stewart and Zhukov 2009)

Many More Possibilities!

Some other interesting questions addressed with Text Data

- How do civilians differ from military leaders on foreign policy? (Stewart and Zhukov 2009)
- How do legislators relate to their constituents? (Grimmer 2010,2013)

Many More Possibilities!

Some other interesting questions addressed with Text Data

- How do civilians differ from military leaders on foreign policy? (Stewart and Zhukov 2009)
- How do legislators relate to their constituents? (Grimmer 2010,2013)
- What drives media slant? (Gentkow and Shapiro 2010)

Many More Possibilities!

Some other interesting questions addressed with Text Data

- How do civilians differ from military leaders on foreign policy? (Stewart and Zhukov 2009)
- How do legislators relate to their constituents? (Grimmer 2010,2013)
- What drives media slant? (Gentkow and Shapiro 2010)
- Was there a constitutional moment in the 1860s? (Stewart and Young 2013)

Many More Possibilities!

Some other interesting questions addressed with Text Data

- How do civilians differ from military leaders on foreign policy? (Stewart and Zhukov 2009)
- How do legislators relate to their constituents? (Grimmer 2010,2013)
- What drives media slant? (Gentkow and Shapiro 2010)
- Was there a constitutional moment in the 1860s? (Stewart and Young 2013)
- Where does [insert party/legislator] fall on the ideological spectrum?

Many More Possibilities!

Some other interesting questions addressed with Text Data

- How do civilians differ from military leaders on foreign policy? (Stewart and Zhukov 2009)
- How do legislators relate to their constituents? (Grimmer 2010,2013)
- What drives media slant? (Gentkow and Shapiro 2010)
- Was there a constitutional moment in the 1860s? (Stewart and Young 2013)
- Where does [insert party/legislator] fall on the ideological spectrum?
- Plus: Open-ended survey analysis (Roberts et. al 2013), Digital Humanities (Jockers 2012), Digital Historiography (Mimno 2012), Public Opinion (Hopkins and King 2010), Congressional Discourse (Quinn et al 2010), Legal Analysis (Gill and Hall 2013) etc.

Overview

Big Data Isn't About the Data

Data only gets you so far.

Big Data Isn't About the Data

Data only gets you so far.
Four principles for automated text analysis:

Big Data Isn't About the Data

Data only gets you so far.
Four principles for automated text analysis:

1. Statistical models of language are wrong but useful.

Big Data Isn't About the Data

Data only gets you so far.
Four principles for automated text analysis:

1. Statistical models of language are wrong but useful.
2. Best methods *amplify* resources and augment humans.

Big Data Isn't About the Data

Data only gets you so far.
Four principles for automated text analysis:

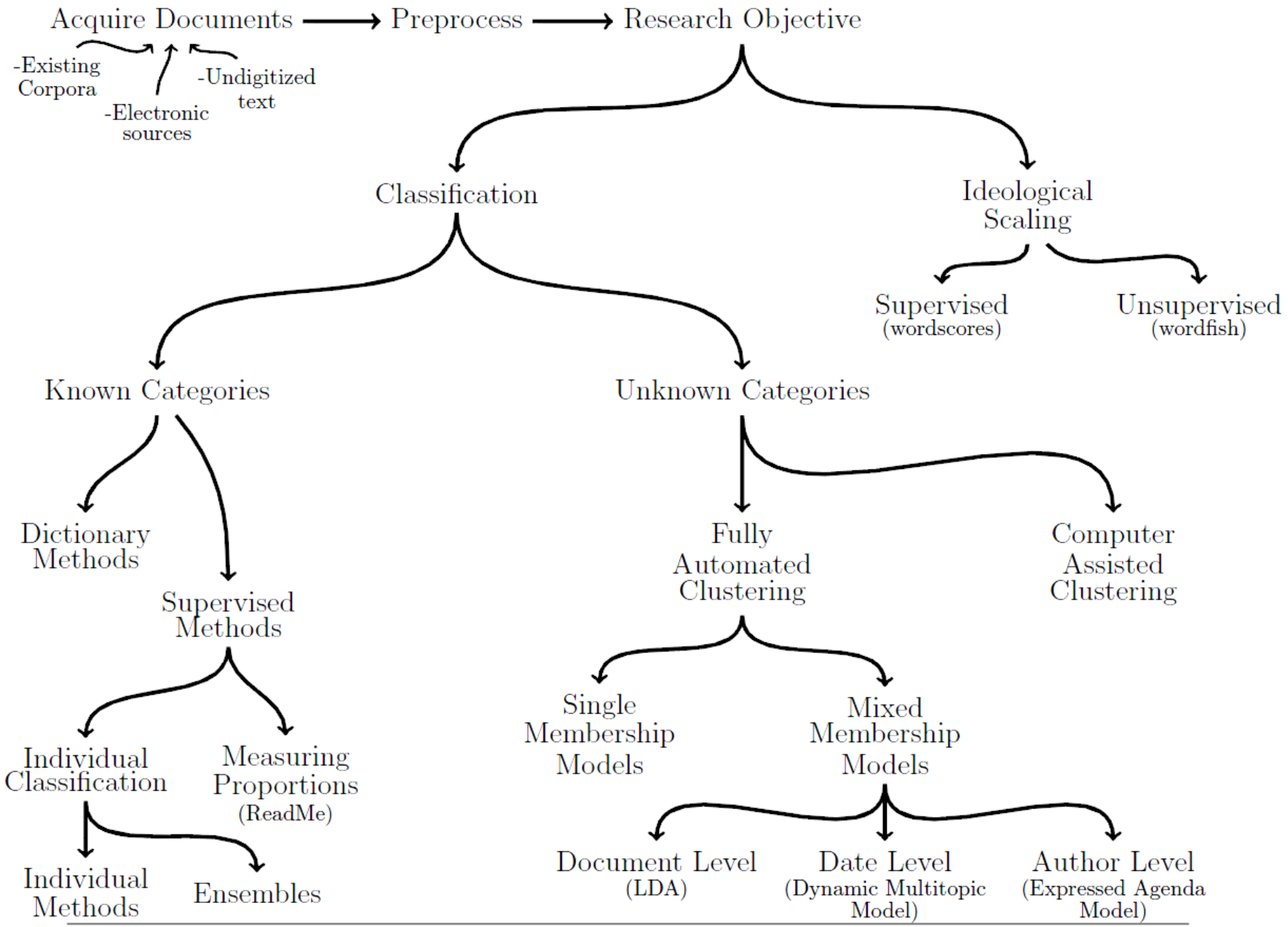
1. Statistical models of language are wrong but useful.
2. Best methods *amplify* resources and augment humans.
3. There is no globally best method.

Big Data Isn't About the Data

Data only gets you so far.
Four principles for automated text analysis:

1. Statistical models of language are wrong but useful.
2. Best methods *amplify* resources and augment humans.
3. There is no globally best method.
4. Validate, Validate, Validate.

A unifying view of text methods



The Value of Better Methods

- Moores Law (doubling speed/power every 18 months) vs. Better Algorithm (1000x speed increase in 1 day)

The Value of Better Methods

- Moores Law (doubling speed/power every 18 months) vs. Better Algorithm (1000x speed increase in 1 day)
- Statistics can see the needle in the haystack

The Value of Better Methods

- Moores Law (doubling speed/power every 18 months) vs. Better Algorithm (1000x speed increase in 1 day)
- Statistics can see the needle in the haystack
- Simple Methods can lead you astray.

The Value of Better Methods

- Moores Law (doubling speed/power every 18 months) vs. Better Algorithm (1000x speed increase in 1 day)
- Statistics can see the needle in the haystack
- Simple Methods can lead you astray.

Examples of bad analytics (Jobs and Jordan).

Example: International Events

200 Million International Events

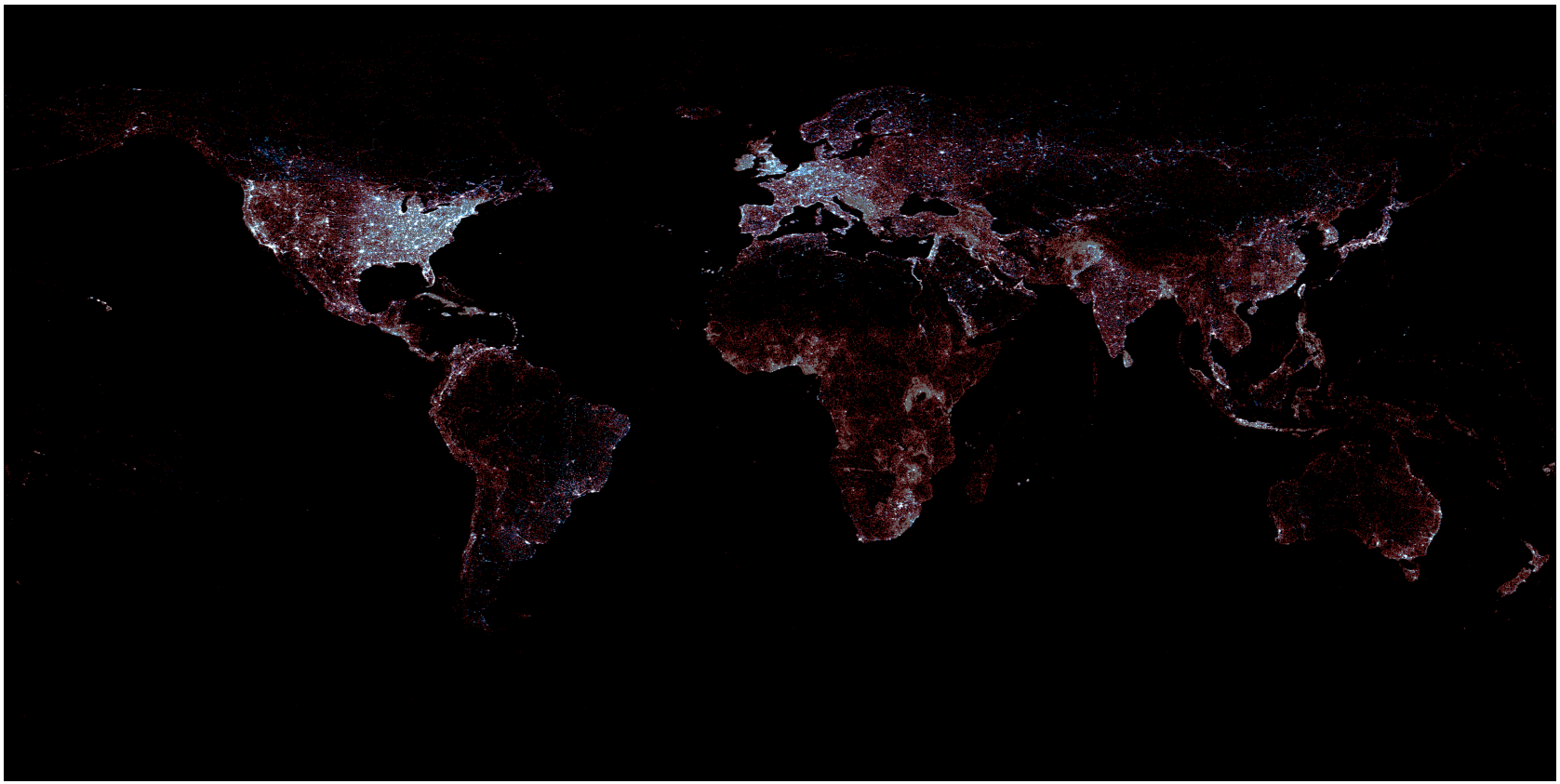
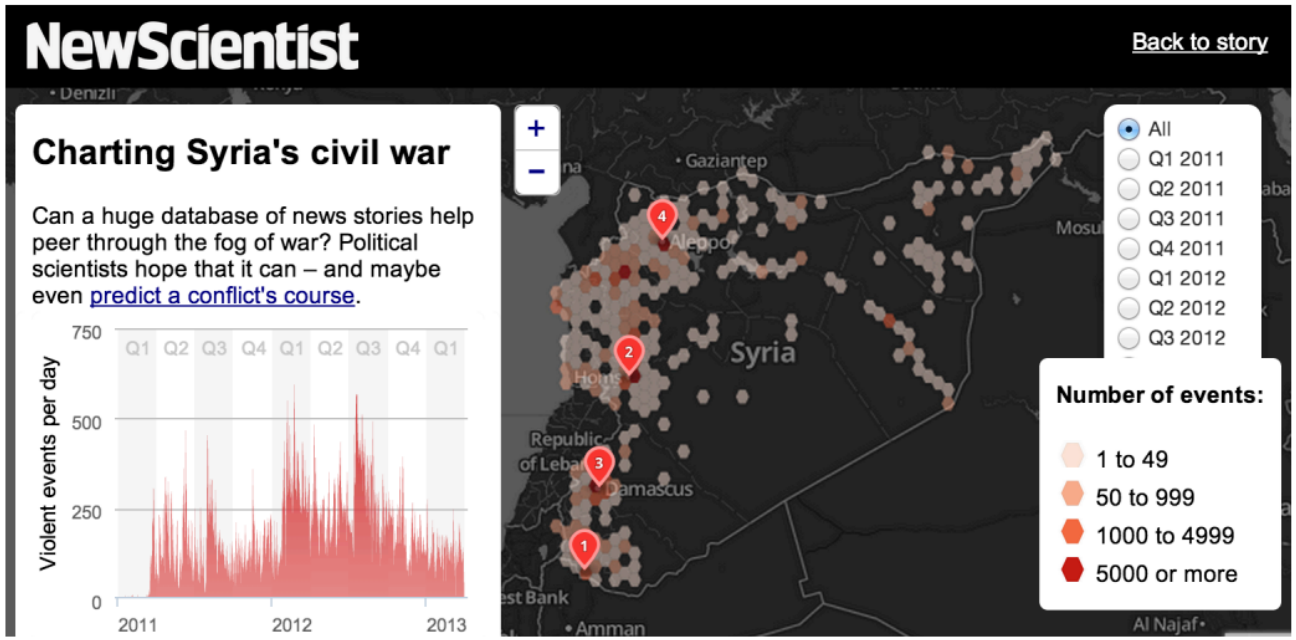


Image from GDELT data by Kalev Leetaru

Example: International Events

Able to track violence over time.



Example: International Events

Suddenly in the 1990's Jordan began attacking everyone.

Example: International Events

Because Michael Jordan was attacking the net.

Overview

Chinese Censorship (King, Roberts and Pan 2013)

- Largest selective suppression of human expression in history

Chinese Censorship (King, Roberts and Pan 2013)

- Largest selective suppression of human expression in history
- Press freedom ranking: 187th out of 197

Chinese Censorship (King, Roberts and Pan 2013)

- Largest selective suppression of human expression in history
- Press freedom ranking: 187th out of 197
- Scale of Effort:

Chinese Censorship (King, Roberts and Pan 2013)

- Largest selective suppression of human expression in history
- Press freedom ranking: 187th out of 197
- Scale of Effort:
  - Total Government Censors: $\approx$ 200,000
  - Internet Police: $20,000 - 50,000$

Chinese Censorship (King, Roberts and Pan 2013)

- Largest selective suppression of human expression in history
- Press freedom ranking: 187th out of 197
- Scale of Effort:
  - Total Government Censors: $\approx$ 200,000
  - Internet Police: $20,000 - 50,000$
- 11 Million Posts analyzed, about 13% censored

What Do They Censor?

- Previous understanding: they censor criticisms of the government

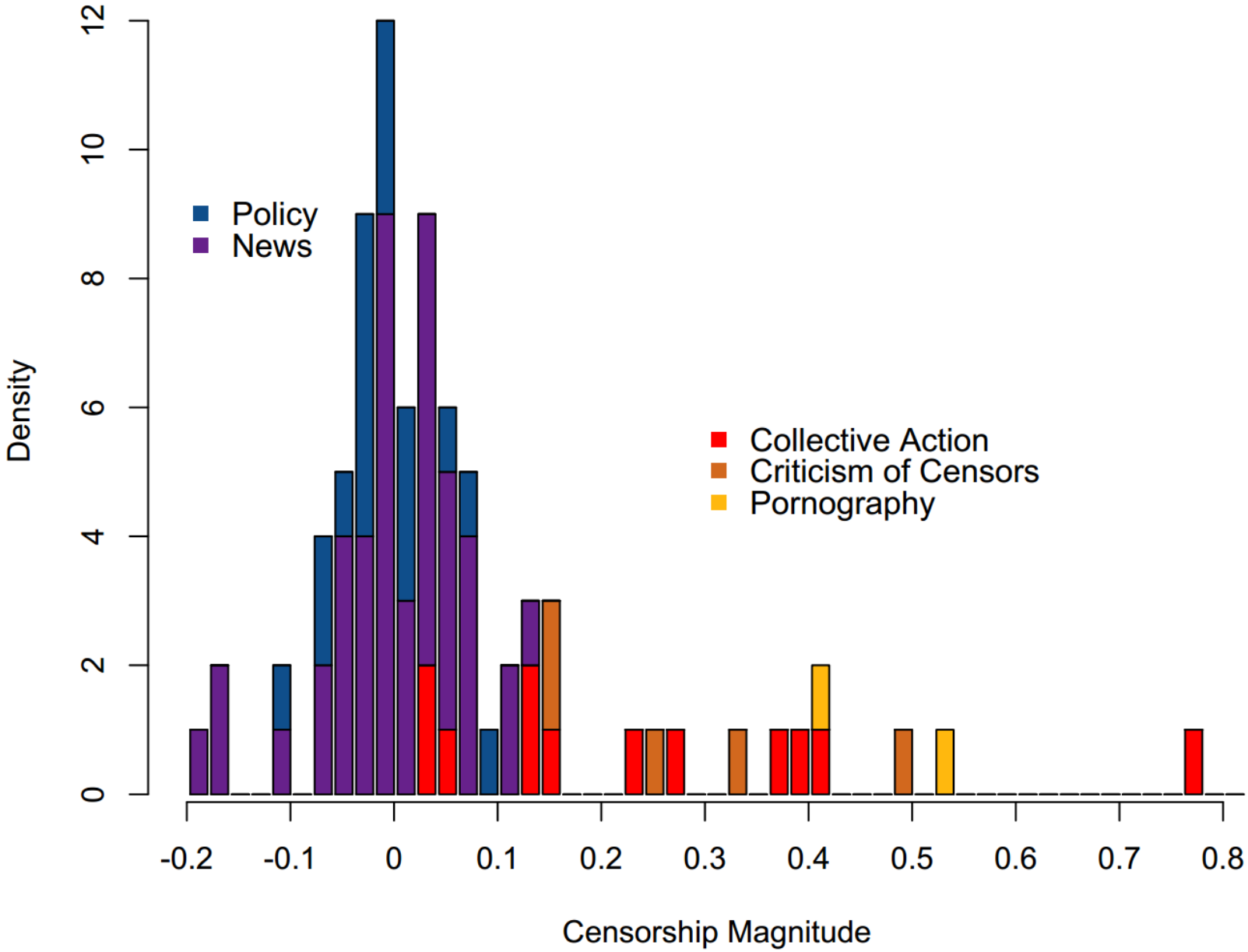
What Do They Censor?

- Previous understanding: they censor criticisms of the government
- New Results: they silence *collective action*

What Do They Censor?

- Previous understanding: they censor criticisms of the government
- New Results: they silence *collective action*
- Criticism relatively uncensored.

What Do They Censor?



China and the Media (Roberts, Stewart and Airoldi 2013)

- Want to understand how the traditional media covers China's rise.

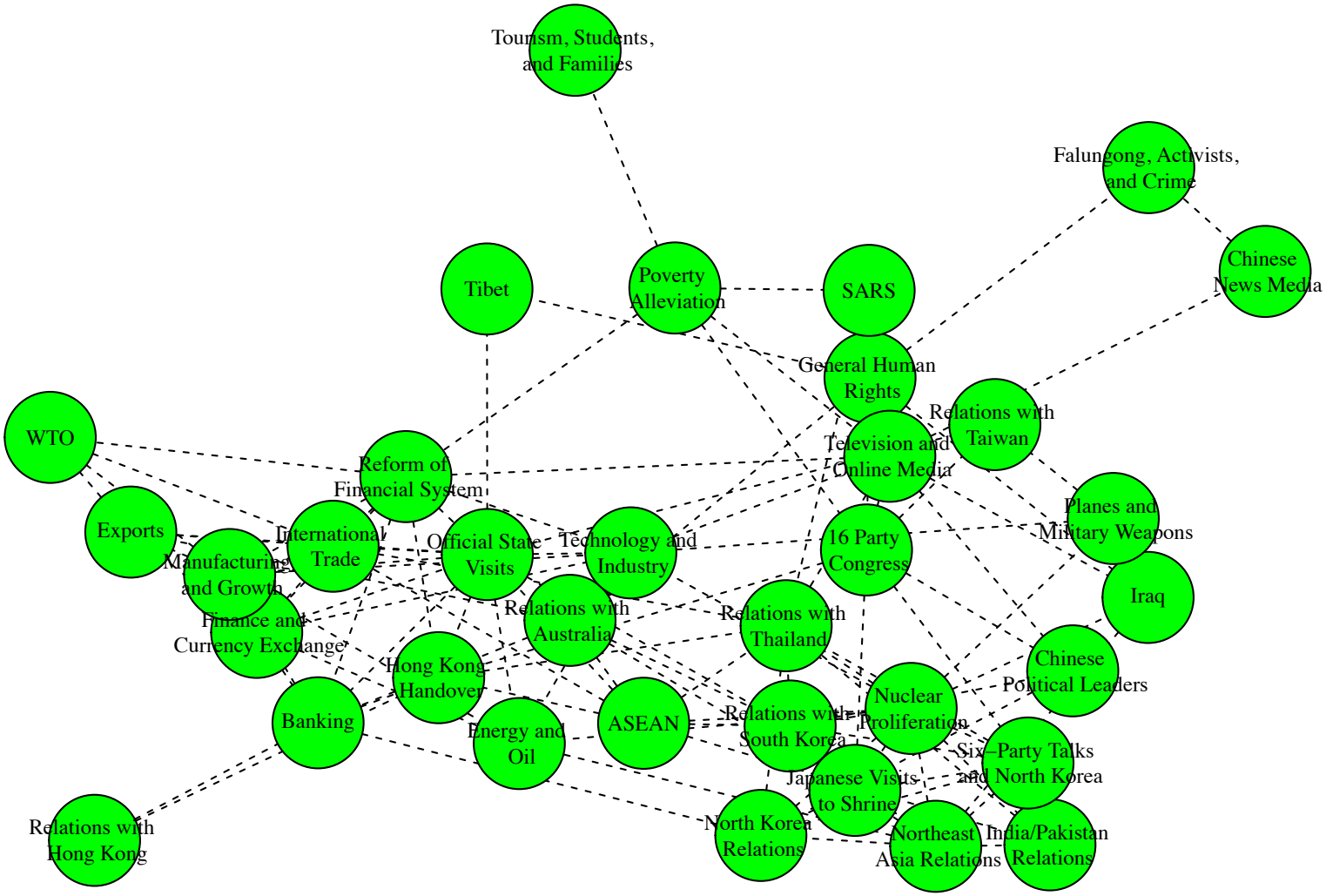
China and the Media (Roberts, Stewart and Airoldi 2013)

- Want to understand how the traditional media covers China's rise.
- Topic Models to estimate and track topics in millions of news reports.

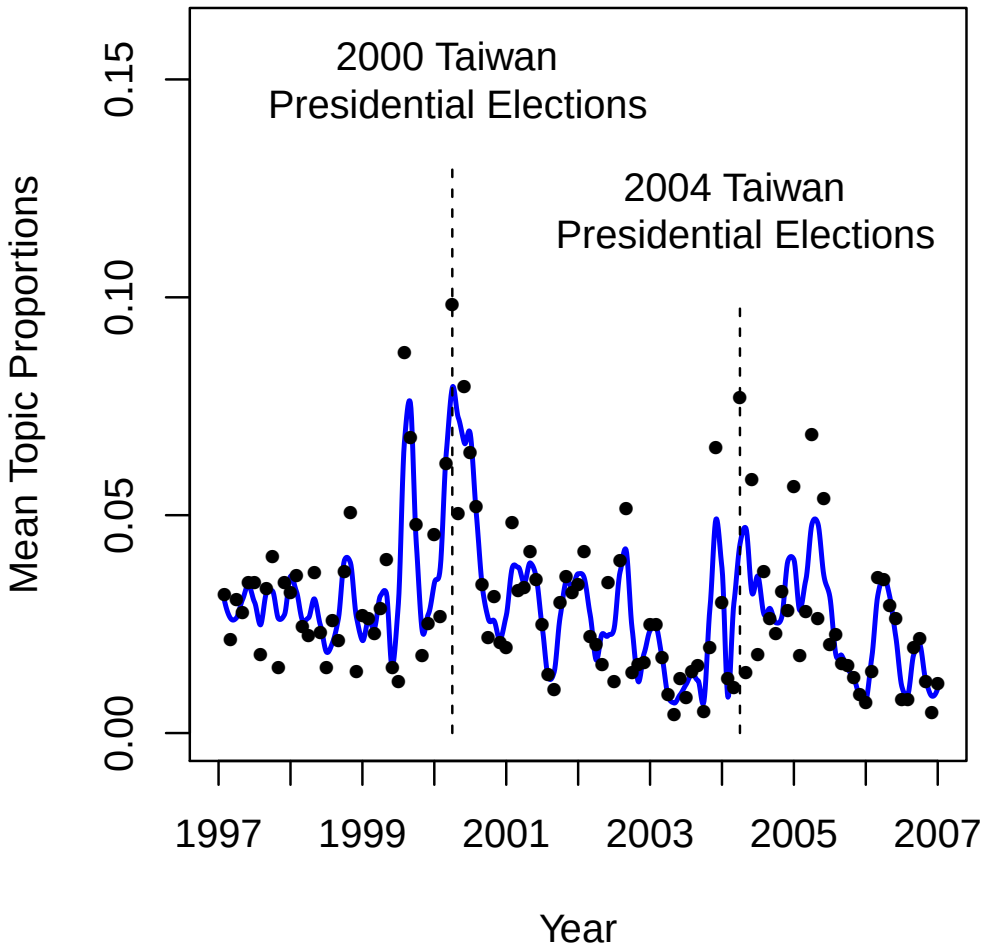
China and the Media (Roberts, Stewart and Airoldi 2013)

- Want to understand how the traditional media covers China's rise.
- Topic Models to estimate and track topics in millions of news reports.
- Uses the Structural Topic Model (an extension of Latent Dirichlet Allocation)

China and the Media (Roberts, Stewart and Airoldi 2013)



Topics by Time



Comparing Coverage by Source

- Compare Xinhua to Western News Wires

Comparing Coverage by Source

- Compare Xinhua to Western News Wires
- Same topics covered but in different ways.

Comparing Coverage by Source

- Compare Xinhua to Western News Wires
- Same topics covered but in different ways.
  - Taiwan: 'one-china', 'reunification' vs. 'elections', 'democratic'

Comparing Coverage by Source

- Compare Xinhua to Western News Wires
- Same topics covered but in different ways.
  - Taiwan: 'one-china', 'reunification' vs. 'elections', 'democratic'
  - Falungong: 'crime', 'illegal' vs. 'protest', 'crackdown'

Comparing Coverage by Source

- Compare Xinhua to Western News Wires
- Same topics covered but in different ways.
  - Taiwan: 'one-china', 'reunification' vs. 'elections', 'democratic'
  - Falungong: 'crime', 'illegal' vs. 'protest', 'crackdown'
- Delayed coverage: SARS

Conclusion

Two Points:

Conclusion

Two Points:

- Text as Data

Conclusion

Two Points:

- Text as Data
- Data + Analysis $>$ Data

Conclusion

Two Points:

- Text as Data
- Data + Analysis $>$ Data

More Info: `scholar.harvard.edu/bstewart`

Conclusion

Thank You!