Content Analysis

Lecture 1: Turning text into data

24 April, 2017

Prof. Andrew Eggers

An exciting moment

Content analysis is a broad field. **Our focus:** use of text as data in quantitative social science.

Sense of huge potential right now:

Much of social life occurs in texts (speeches, press releases, fatwas, laws, letters, books; emails, tweets)

Already huge content, but growing faster than anyone can read:

- 10 mins of worldwide email = 1 Library of Congress
- 1 min of YouTube uploads = 300 hours of video

Text generically hard to interpret, but we're making progress (?)

But let's not get carried away

Social science is full of measurement problems:

- How can I measure the economic output of a country?
- How can I measure how democratic a country is?
- How can I measure someone's partisanship?

But let's not get carried away

Social science is full of measurement problems:

- How can I measure the economic output of a country?
- How can I measure how democratic a country is?
- How can I measure someone's partisanship?

In this course, we're (mostly) talking about efforts to solve **measurement problems involving text:**

But let's not get carried away

Social science is full of measurement problems:

- How can I measure the economic output of a country?
- How can I measure how democratic a country is?
- How can I measure someone's partisanship?

In this course, we're (mostly) talking about efforts to solve **measurement problems involving text:**

- generating dependent and/or independent variable(s) of a regression from textual sources, or

But let's not get carried away

Social science is full of measurement problems:

- How can I measure the economic output of a country?
- How can I measure how democratic a country is?
- How can I measure someone's partisanship?

In this course, we're (mostly) talking about efforts to solve **measurement problems involving text:**

- generating dependent and/or independent variable(s) of a regression from textual sources, or
- characterizing trends in some other concept (e.g. partisanship, sentiment) using language as the raw data

But let’s not get carried away

Social science is full of measurement problems:

- How can I measure the economic output of a country?
- How can I measure how democratic a country is?
- How can I measure someone’s partisanship?

In this course, we’re (mostly) talking about efforts to solve **measurement problems involving text:**

- generating dependent and/or independent variable(s) of a regression from textual sources, or
- characterizing trends in some other concept (e.g. partisanship, sentiment) using language as the raw data

Broader field of content analysis includes studies of discourse for its own sake. (See chap. 3 of Krippendorff.)

Example 1: Fouirnaies and Hall on regulatory risk

Research question: In the US, do firms contribute money to incumbent politicians in order to obtain preferential treatment?

Research design: If so, we would expect responsiveness of contributions to election outcomes to be higher for firms whose business depends more on government regulation.

Measurement problems:

- Which firms' contributions are more responsive to election outcomes?
- Which firms are more exposed to government regulation?

Example 1: Fouirnaies and Hall (cont'd)

Measurement strategy: Generate government exposure index from keyword counts in companies' official annual reports (10-K), where they describe risks they face.

Example 1: Fouirnaies and Hall (cont'd)

Measurement strategy: Generate government exposure index from keyword counts in companies' official annual reports (10-K), where they describe risks they face.

Excerpts from discussion of risks in a sample 10-K:

Example 1: Fouirnaies and Hall (cont'd)

Measurement strategy: Generate government exposure index from keyword counts in companies' official annual reports (10-K), where they describe risks they face.

Excerpts from discussion of risks in a sample 10-K:

More people are using devices other than desktop computers to access the Internet and accessing new devices to make search queries. If manufacturers and users do not widely adopt versions of our search technology, products, or operating systems developed for these devices, our business could be adversely affected.

Example 1: Fouirnaies and Hall (cont'd)

Measurement strategy: Generate government exposure index from keyword counts in companies' official annual reports (10-K), where they describe risks they face.

Excerpts from discussion of risks in a sample 10-K:

More people are using devices other than desktop computers to access the Internet and accessing new devices to make search queries. If manufacturers and users do not widely adopt versions of our search technology, products, or operating systems developed for these devices, our business could be adversely affected.

We are subject to increasing regulatory scrutiny that may negatively impact our business. Additionally, changes in policies governing a wide range of topics may adversely affect our business.

The growth of our company and our expansion into a variety of new fields involves a variety of new regulatory issues, and we have experienced increased regulatory scrutiny as we have grown. For instance, various regulatory agencies are reviewing aspects of our search and other businesses. We continue to cooperate with the European Commission and other regulatory authorities around the world in investigations they are conducting with respect to our business.

Example 1: Fouirnaies and Hall (cont'd)

Measurement strategy: Generate government exposure index from keyword counts in companies' official annual reports (10-K), where they describe risks they face.

Excerpts from discussion of risks in a sample 10-K:

More people are using devices other than desktop computers to access the Internet and accessing new devices to make search queries. If manufacturers and users do not widely adopt versions of our search technology, products, or operating systems developed for these devices, our business could be adversely affected.

We are subject to increasing regulatory scrutiny that may negatively impact our business. Additionally, changes in policies governing a wide range of topics may adversely affect our business.

The growth of our company and our expansion into a variety of new fields involves a variety of new regulatory issues, and we have experienced increased regulatory scrutiny as we have grown. For instance, various regulatory agencies are reviewing aspects of our search and other businesses. We continue to cooperate with the European Commission and other regulatory authorities around the world in investigations they are conducting with respect to our business.

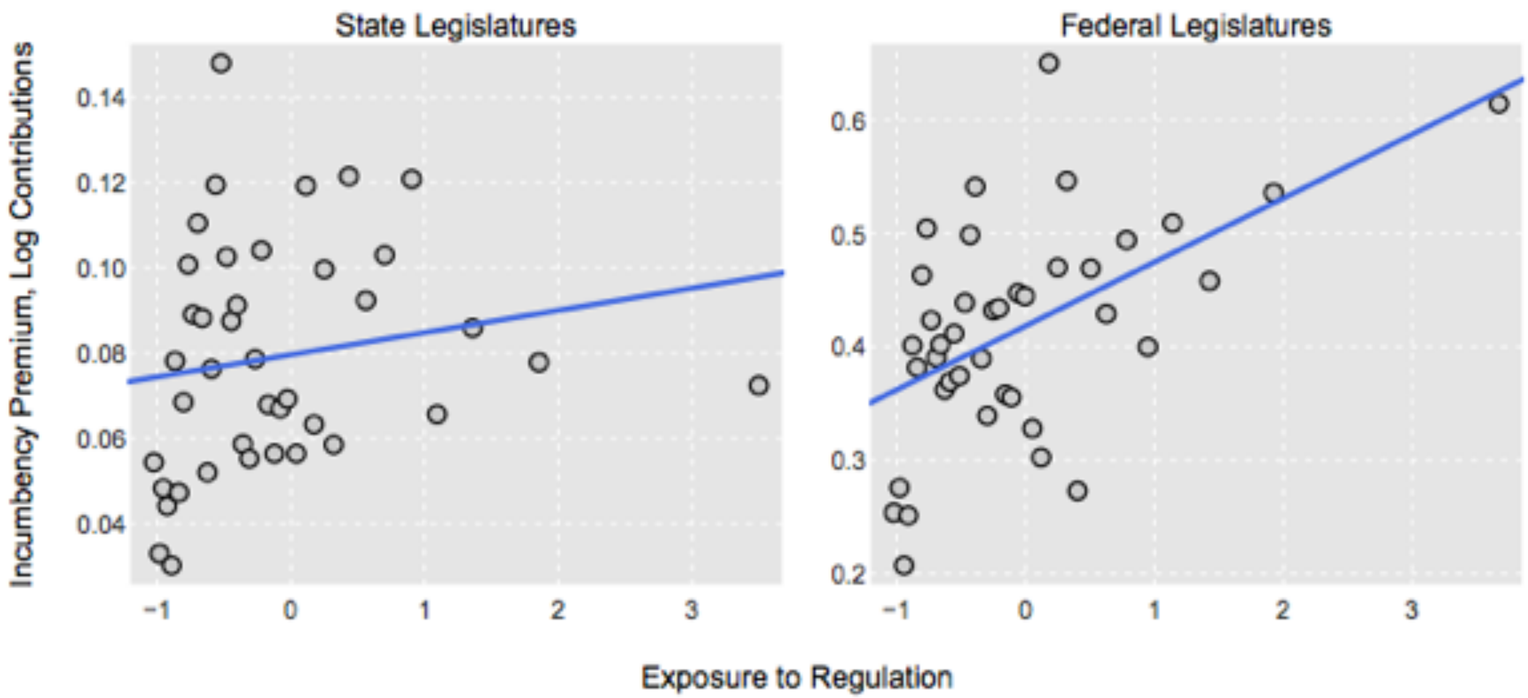
Keywords: *require, regulat, law, polic, federal, …*

Principal components analysis (PCA) on counts => single index.

Example 1: Fouirnaies and Hall (cont'd)

Analysis: Shows that contributions from firms with higher exposure indices (calculated from counts of keywords) respond more to election results.

Figure 8 – Exposure to Regulation and Firm Contributions to Incumbents and Non-Incumbents. Regulated firms are more sensitive to incumbency.



Note: Points represent averages in equal-sample-sized bins of the exposure to regulation variable. Lines are simple OLS predictions from a regression fitted to the binned points.

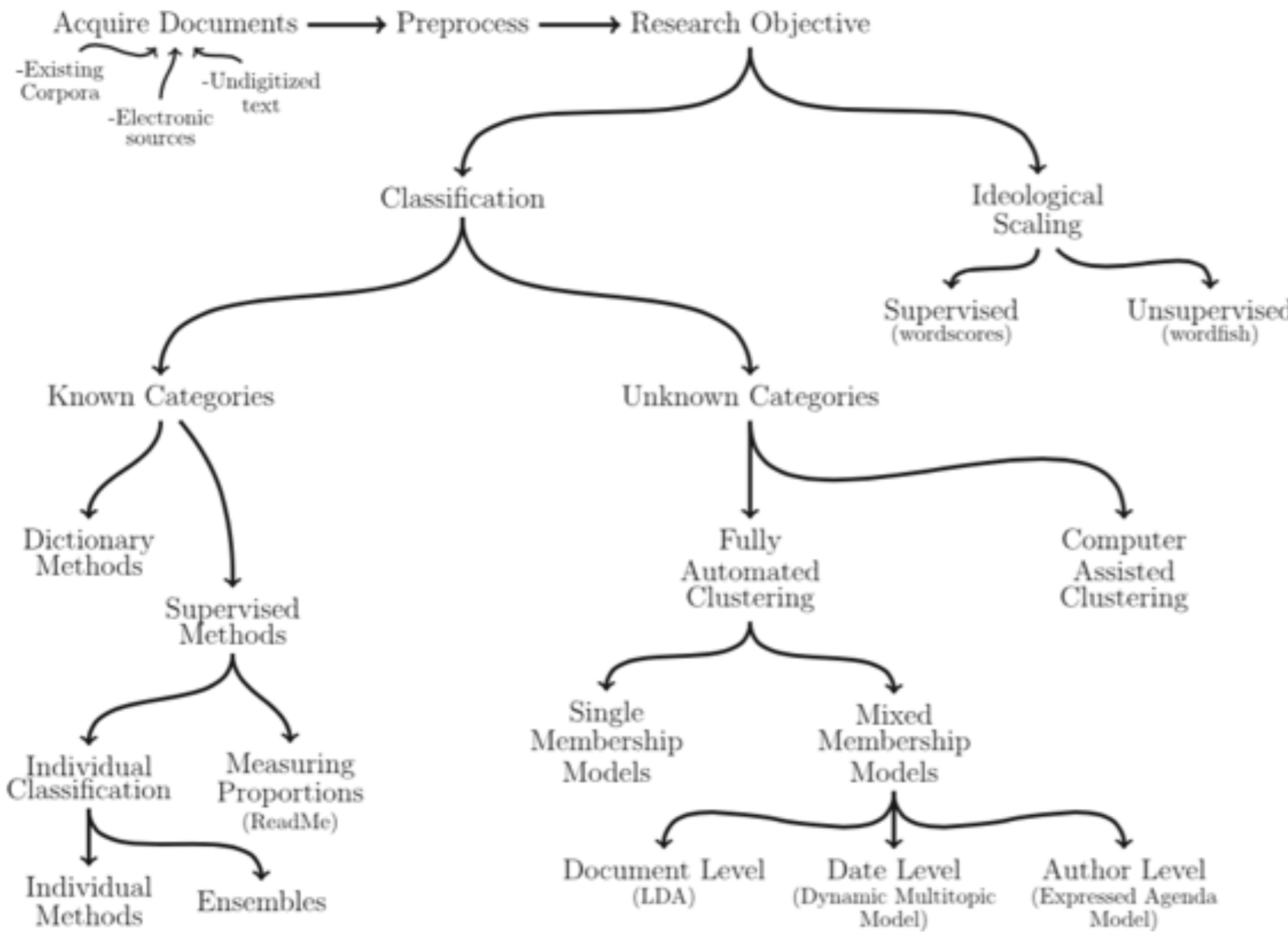
Fouirnaies and Hall in context

This is an example of a **dictionary method:** researcher decides on keywords (perhaps through reading, trial and error, reliance on previous literature) and counts occurrences.

Other examples in this week's reading list:

- Gentzkow et al: counting occurrences of **emotionally charged words** in newspapers as measure of slanted journalism
- Baker et al: counting articles mentioning **keywords relating to economy, policy, and uncertainty** as measure of economic policy uncertainty
- Ban et al: counting **how many times an entity is mentioned** as measure of entity's power

A classification of "text-as-data" methods



Grimmer and Stewart (2013), "Text as Data"

Example 2: Larcker & Zakolyukina on deceptive CEOs

Research question: Can we predict which companies are likely to have financial restatements based on what CEOs/CFOs say in conference calls with investors?

(or)

Is deceptive speech different from truthful speech?

Example 2: Larcker & Zakolyukina, cont'd

Strategy: Using lots of previous research, identify groups of keywords characteristic of deception

TABLE 1—*Continued*

Panel A: Variables, Computation, and Predicted Signs

Category	Abbreviation	Sign	Calculation
Anxiety	anx	+	LIWC category "anx": worried, fearful, nervous, etc. Simple count divided by the number of words ignoring articles (wc) and multiplied by the median wc in the sample. Prior research: Bachenko, Fitzpatrick, and Schonwetter [2008], Bond and Lee [2005], Knapp, Hart, and Dennis [1974], Newman et al. [2003], Vrij [2008].
Anger	anger	+	LIWC category "anger": hate, kill, annoyed, etc. Simple count divided by the number of words ignoring articles (wc) and multiplied by the median wc in the sample. Prior research: Bachenko, Fitzpatrick, and Schonwetter [2008], Bond and Lee [2005], Newman et al. [2003], Vrij [2008].
Swear words	swear	+	LIWC category "swear": screw*, hell, etc. Simple count divided by the number of words ignoring articles (wc) and multiplied by the median wc in the sample. Prior research: Bachenko, Fitzpatrick, and Schonwetter [2008], DePaulo et al. [2003], Vrij [2008].
Extreme negative emotions	negemoextr	+	Self-constructed category: absurd, adverse, awful, etc. For the complete list see panel B. Simple count divided by the number of words ignoring articles (wc) and multiplied by the median wc in the sample. Prior research: Newman et al. [2003], Vrij [2008].
			Cognitive Process
Certainty	certain	−	LIWC category "certain": always, never, etc. Simple count divided by the number of words ignoring articles (wc) and multiplied by the median wc in the sample. Prior research: Bond and Lee [2005], Knapp, Hart, and Dennis [1974], Newman et al. [2003], Vrij [2008].
Tentative	tentat	+	LIWC category "tentat": maybe, perhaps, guess, etc. Simple count divided by the number of words ignoring articles (wc) and multiplied by the median wc in the sample. Prior research: Adams and Jarvis [2006], Bond and Lee [2005], DePaulo et al. [2003], Knapp, Hart, and Dennis [1974], Newman et al. [2003], Vrij [2008].

Then fit predictive logit model, with restatement indicator as DV, 19 linguistic measures as independent variables.

Example 2: Larcker & Zakolyukina, cont'd

TABLE 6

Logit Linguistic-Based Prediction Models for CEO and CFO Narratives During Conference Calls

Panel A: CEO Sample

		NT	IRAI	IR	AAER
		Word Count			
wc‡	±	1.01	1.04	1.16	1.04
		(0.10)	(0.13)	(0.15)	(0.24)
		References			
I	−	0.95	0.89	0.87	0.87
		(0.07)	(0.09)	(0.10)	(0.18)
we	+	0.99	0.92	0.95	1.07
		(0.04)	(0.05)	(0.05)	(0.12)
they	±	1.06	1.10	0.98	0.50**
		(0.12)	(0.16)	(0.16)	(0.15)
ipron	±	0.94	0.97	0.96	1.14
		(0.04)	(0.05)	(0.06)	(0.12)
genknlref	±	1.91***	1.96***	1.99***	1.98**
		(0.33)	(0.33)	(0.36)	(0.64)
		Positives/Negatives			
assent	−	1.10	1.16	1.20	0.36
		(0.28)	(0.36)	(0.43)	(0.28)
posemone	−	0.88**	0.94	0.93	0.97
		(0.05)	(0.07)	(0.08)	(0.16)
posemoextr	±	1.20	1.62***	1.99***	3.51***
		(0.16)	(0.25)	(0.33)	(1.26)
negate	+	0.92	0.86	0.87	1.24
		(0.11)	(0.13)	(0.15)	(0.43)
anx	+	0.38**	0.34**	0.25***	0.08**
		(0.16)	(0.14)	(0.11)	(0.08)
anger	+	0.97	1.16	1.32	0.57
		(0.35)	(0.55)	(0.70)	(0.66)
swear†	+	0.97	0.95	0.94	1.03
		(0.07)	(0.06)	(0.07)	(0.15)
negemoextr	+	0.99	0.84	0.88	0.83
		(0.26)	(0.31)	(0.33)	(0.66)
		Cognitive Mechanism			
certain	−	1.16	0.90	0.88	0.75
		(0.13)	(0.13)	(0.14)	(0.18)
tentat	+	0.96	0.96	1.00	0.99
		(0.07)	(0.08)	(0.10)	(0.19)
		Other Cues			
hesit†	±	1.05	1.04	1.11*	0.99
		(0.05)	(0.05)	(0.06)	(0.16)
shvalue†	±	0.91**	0.90*	0.88**	0.95
		(0.04)	(0.05)	(0.06)	(0.12)
value†	±	0.90	0.87	0.83*	1.11
		(0.07)	(0.08)	(0.09)	(0.17)
Total firm-quarters		17,150	17,150	17,150	17,150
Deceptive firm-quarters		2,325	1,627	1,355	274
Area under the ROC curve		0.58	0.59	0.61	0.66
Log-likelihood value		−6,732.51	−5,294.87	−4,638.95	−1,353.13
Pseudo *R*-squared		0.011	0.016	0.021	0.037

Larcker & Zakolyukina in context

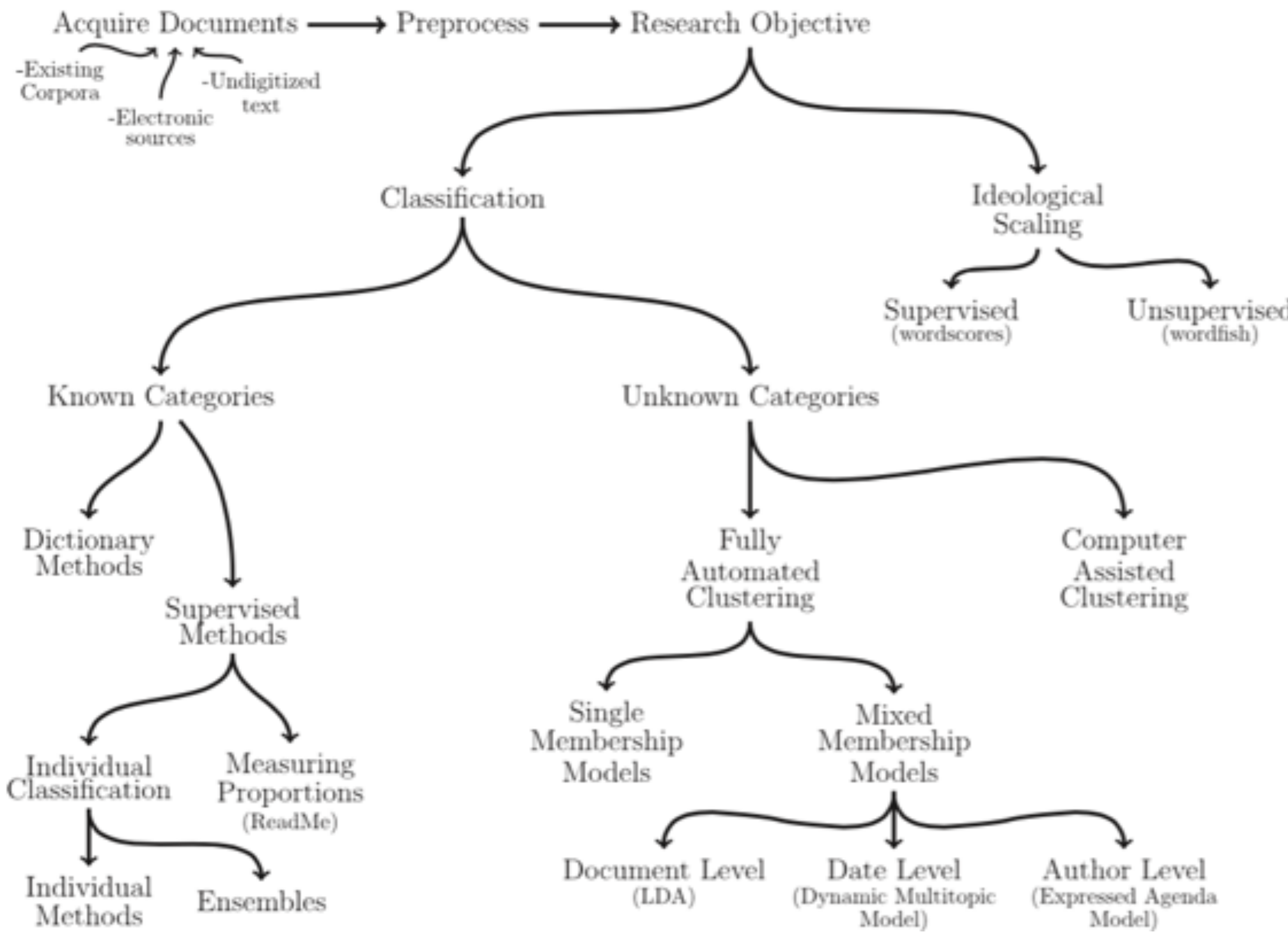
This is an example of **model-based classification**, or **classification via supervised learning**.

This is just like many predictive/explanatory models you have run, except the covariates come from text.

When would this be useful for research?

- When you have a fundamentally predictive problem
  - Future predictions useful
  - There is scholarly interest in showing a connection between linguistic features and some outcome
- When you want to label an enormous amount of data based on a smaller labeled set (e.g. to generate an outcome, or a covariate)

A classification of "text-as-data" methods



Grimmer and Stewart (2013), "Text as Data"

A brief overview of some things to do with text

A brief overview of some things to do with text

- **Frequencies:** How often does this term/theme, or set of terms/themes, appear in the text?
  - Are the themes identified in the text by readers? → qualitative data analysis/QDA, software like MaxQDA, NVivo, Atlas.ti
  - Are the terms/themes identified in the text via software → dictionary methods, sentiment analysis

A brief overview of some things to do with text

- **Frequencies:** How often does this term/theme, or set of terms/themes, appear in the text?
  - Are the themes identified in the text by readers? → qualitative data analysis/QDA, software like MaxQDA, NVivo, Atlas.ti
  - Are the terms/themes identified in the text via software → dictionary methods, sentiment analysis
- **Frequencies of co-occurrence:** What words tend to appear with a given word/phrase? (collocation, co-occurrence, e.g. the work of Paul Baker)

A brief overview of some things to do with text

- **Frequencies:** How often does this term/theme, or set of terms/themes, appear in the text?
  - Are the themes identified in the text by readers? → qualitative data analysis/QDA, software like MaxQDA, NVivo, Atlas.ti
  - Are the terms/themes identified in the text via software → dictionary methods, sentiment analysis
- **Frequencies of co-occurrence:** What words tend to appear with a given word/phrase? (collocation, co-occurrence, e.g. the work of Paul Baker)
- **Distinctive words/phrases:** What words are especially common to a given text/speaker? (keyness, specificity, weirdness, e.g. "Fightin' Words")

A brief overview of some things to do with text (2)

A brief overview of some things to do with text (2)

- **Grouping**: What texts or speakers are similar to each other? (clustering, topic modeling e.g. LDA, scaling e.g. Wordfish)

A brief overview of some things to do with text (2)

- **Grouping**: What texts or speakers are similar to each other? (clustering, topic modeling e.g. LDA, scaling e.g. Wordfish)
- **Classification:** I have labeled some of my texts; tell me what the labels should be on the rest of the texts! (e.g. Naive Bayes, random forests)

A brief overview of some things to do with text (2)

- **Grouping**: What texts or speakers are similar to each other? (clustering, topic modeling e.g. LDA, scaling e.g. Wordfish)
- **Classification:** I have labeled some of my texts; tell me what the labels should be on the rest of the texts! (e.g. Naive Bayes, random forests)
- **Scaling:** Put these texts in some space based on underlying similarities

A brief overview of some things to do with text (3)

Suppose you are doing all of your analysis manually.

- If you're following a **simple rule** to record textual features, the computer can do it better.
- If it is difficult to turn your rule into an algorithm, the computer might be able to help:
  - with data entry/collection (web scraping, keyword counting)
  - visualizing/analyzing the resulting data
  - uncovering the rule that you are actually applying (machine learning)
- Your software may be able to show that something you learn for a subset of texts is probably more generally true.

What do you need in order to do things like this? (1)

- For **collecting text** and **counting features**, you probably need some programming skills. (These problems are too niche for there to be "off-the-shelf" solutions.)
  - **Web scraping** can be very useful for
    - getting the text
    - getting search counts e.g. in Bloom et al, my paper on expenses scandal
  - Given a chunk of text, you need a way to count occurrences (e.g. **regular expressions**)
  - Given many pieces of text, you need to be able to **loop** through them in code and produce output

What do you need in order to do things like this? (2)

- Optical character recognition (OCR) is also useful given printed (e.g. archival) sources
  - e.g. in Eggers & Hainmueller (2009) "MPs for Sale"
  - built into many PDFs; see Text Fairy for phones

7 volumes of *Times Guide to the House of Commons* Converted to text by Widener Library digital services

Peckham
Electorate : 61,050
*Corbet, Mrs. F. K. (Lab.) 26,315
Smith, D. G. (C.) 12,547

Lab. majority 13,768
NO CHANGE
TOTAL VOTE, 38,862.—Lab., 67·7%; C., 32·3%.—Maj., 35·4%.
1951 :—Lab., 33,703 ; C., 14,557.—Lab. maj., 19,146.

MRS. FREDA CORBET represented North-West Camberwell in 1945 and was returned for Peckham in 1950. She contested East Lewisham in 1935. Born 1900 ; educated at Wimbledon County School and University College, London ; became a teacher, lecturer, and barrister. A member of London County Council since 1934 and chief whip of the Labour group. She is interested in education and penal reform.
MR. DUDLEY SMITH, a journalist, is assistant news editor of a national Sunday newspaper. Has been crime reporter, sports writer, and special correspondent. Born 1926 ; educated at Chichester High School.

Peckham
Electorate : 61,050
*Corbet, Mrs. F. K. (Lab.) .. 28,315
Smith, D. G. (C.) 12,547
Lab. majority .. 13,768
NO CHANGE
Total Vote, 38,862 -- Lab., 67-7%; C, 32-3% -- Maj., 35-4%.
1951 :-- Lab., 33,703 ; C, 14,557. -- Lab. maj., 19,146.

Mrs. Freda Corbet
represented North-
West Camberwell in
1945 and was returned
for Peckham in 1950.
She contested East
Lewisham in 1935.
Born 1900 ; educated at
Wimbledon County
School and University
College, London ; be-
came a teacher, lecturer,
and barrister. A mem-
ber of London County
Council since 1934 and chief whip of the
Labour group. She is interested in education
and penal reform.

Mr. Dudley Smith, a journalist, is assis-
tant news editor of a national Sunday news-
paper. Has been crime reporter, sports
writer, and special correspondent. Born
1926 ; educated at Chichester High School.

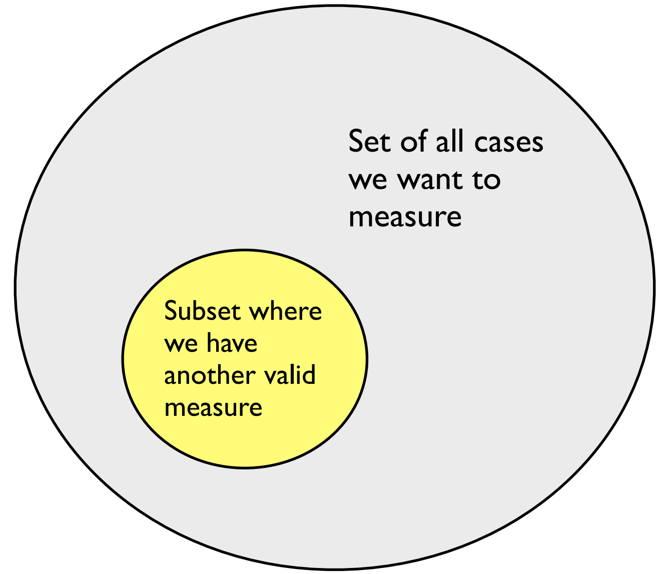
Converted to database using **regular expressions** to identify party, vote count, profession, school, date of birth for each candidate

What do you need in order to do things like this? (3)

- For a new measure (based on dictionary methods or otherwise) you'll need to do **validation**
- For **classification model**,
  - Your **intro stats** skills will be useful!
  - But since we don't focus on prediction/classification, look at *Introduction to Statistical Learning* or elsewhere for discussions of
    - overfitting, test/training sets, cross-validation
    - model selection & what to do when you have too many predictors: regularization, shrinkage, LASSO, support vector machines (SVM), ridge regression, naive Bayes

How we validate, with two examples

Basically, we assess whether a measure works for the subset of cases where we know what it should produce, i.e. where we have another valid measure.



Two examples:

- Measuring implication in 2009 parliamentary expenses scandal with counts of Google News articles (Eggers 2014)
- Measuring political power with mentions in U.S. newspapers (Ban, Fouirnaies, Hall, Snyder 2015)

Example: Eggers (2014) on expenses scandal

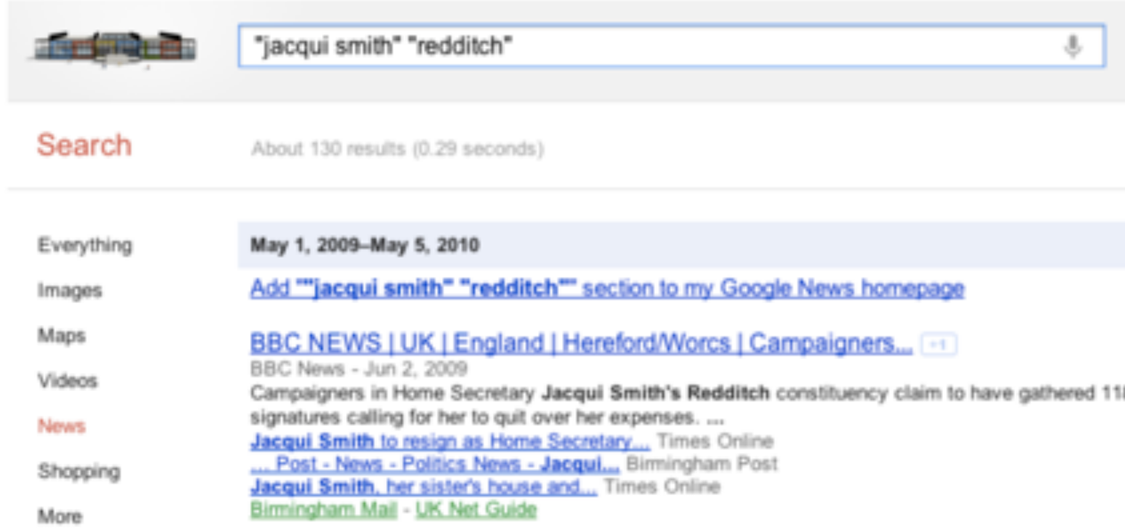
Research question: How did local strength of party preference affect degree to which MPs were punished in expenses scandal?

Measurement problem: How much was each MP implicated?

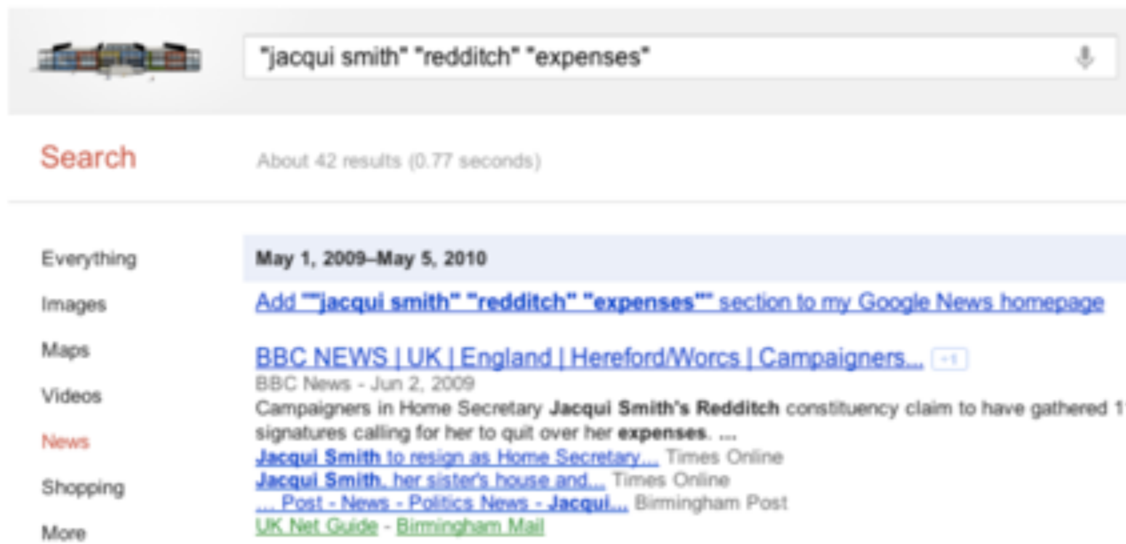
Possible measures:

- Amount of money MP spent
- Amount of money MP was asked to return
- BES survey of voters: "did your MP spend money improperly?"
- Appearance on a list of worst offenders e.g. in the *Telegraph* in May 2009

Step 1: count Google News hits for MP's name and constituency between scandal and election



Step 2: count hits for for MP's name and constituency **and the word "expenses"**



Step 3: divide to get implication score

$$\text{Implication}_i = \frac{\#\text{expenses stories}_i}{\#\text{stories}_i + n_0}$$

How to validate?

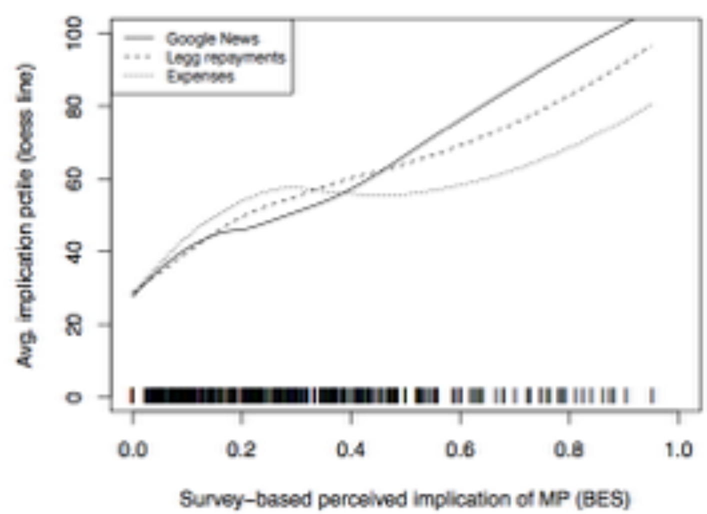
1. Compare with Telegraph's list of "saints" and "sinners"

2. Check list against substantive knowledge

Top 5

MP	Total stories	Expenses stories	Index
Margaret Moran	158	140	0.83
David Chaytor	109	93	0.78
Andrew MacKay	111	89	0.74
Julie Kirkbride	198	147	0.71
Peter Viggers	92	72	0.71

(3. Assess correlation with other possible measures)



Example: Ban, Fouirnaies, Hall, and Snyder (2015) on political power

Research question: Did U.S. Progressive-era reforms weaken state party machines?

Measurement problem: How powerful is the state party machine?

Possible measures:

- Historians' accounts
- Mayhew's measures, which only apply to 1966-1970

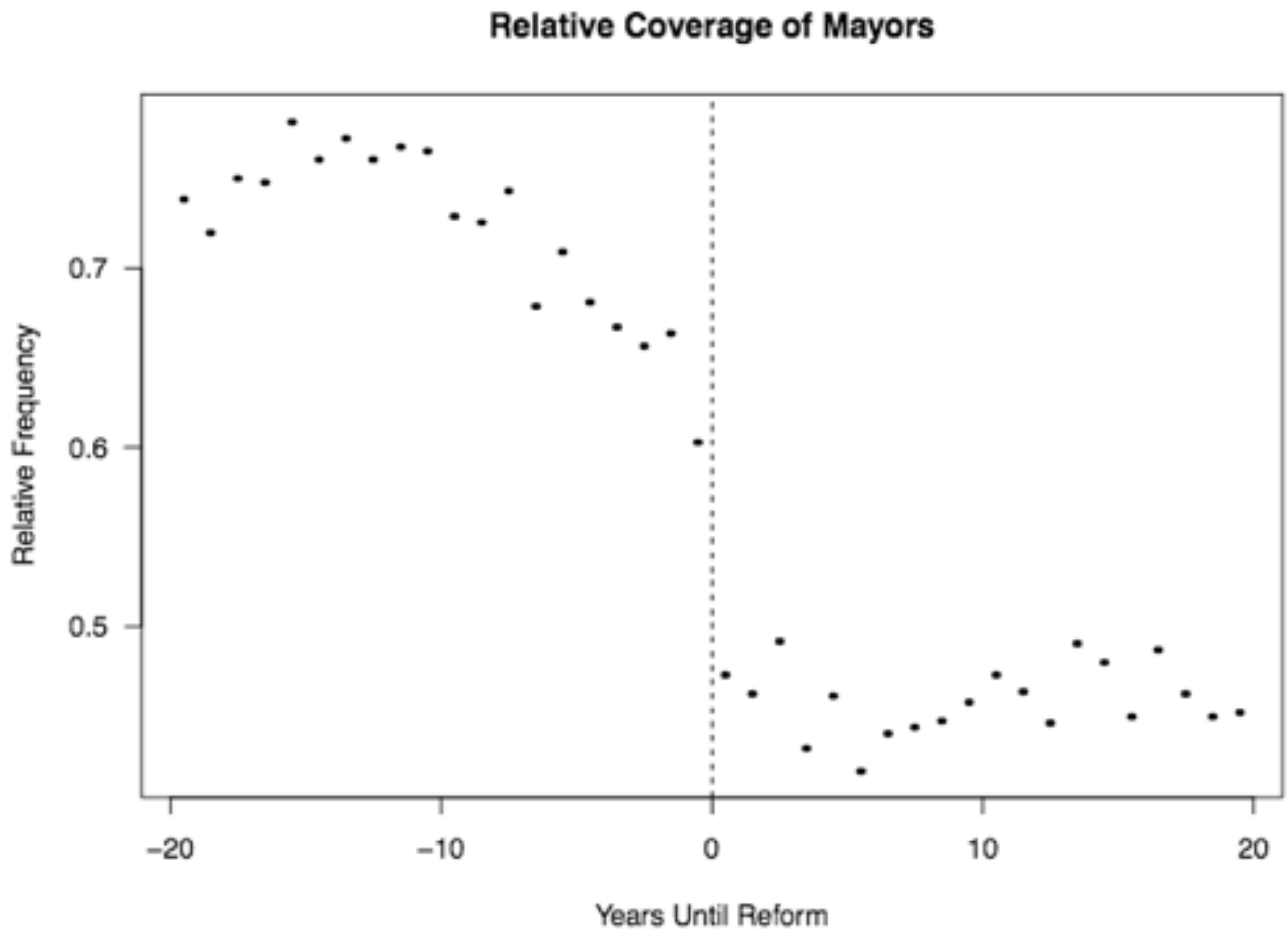
Ban et al (2015): Using newspaper mentions to measure power

Procedure:

- Gather huge newspaper database from online sources
  - 3,000+ newspapers
  - 1877-1977
  - 60+ million pages of text
- Count instances (by state and year) when the word "committee" follows within 5 words of "state", "county", "district", "local" etc **and** "Democratic", "Republican", or "GOP"

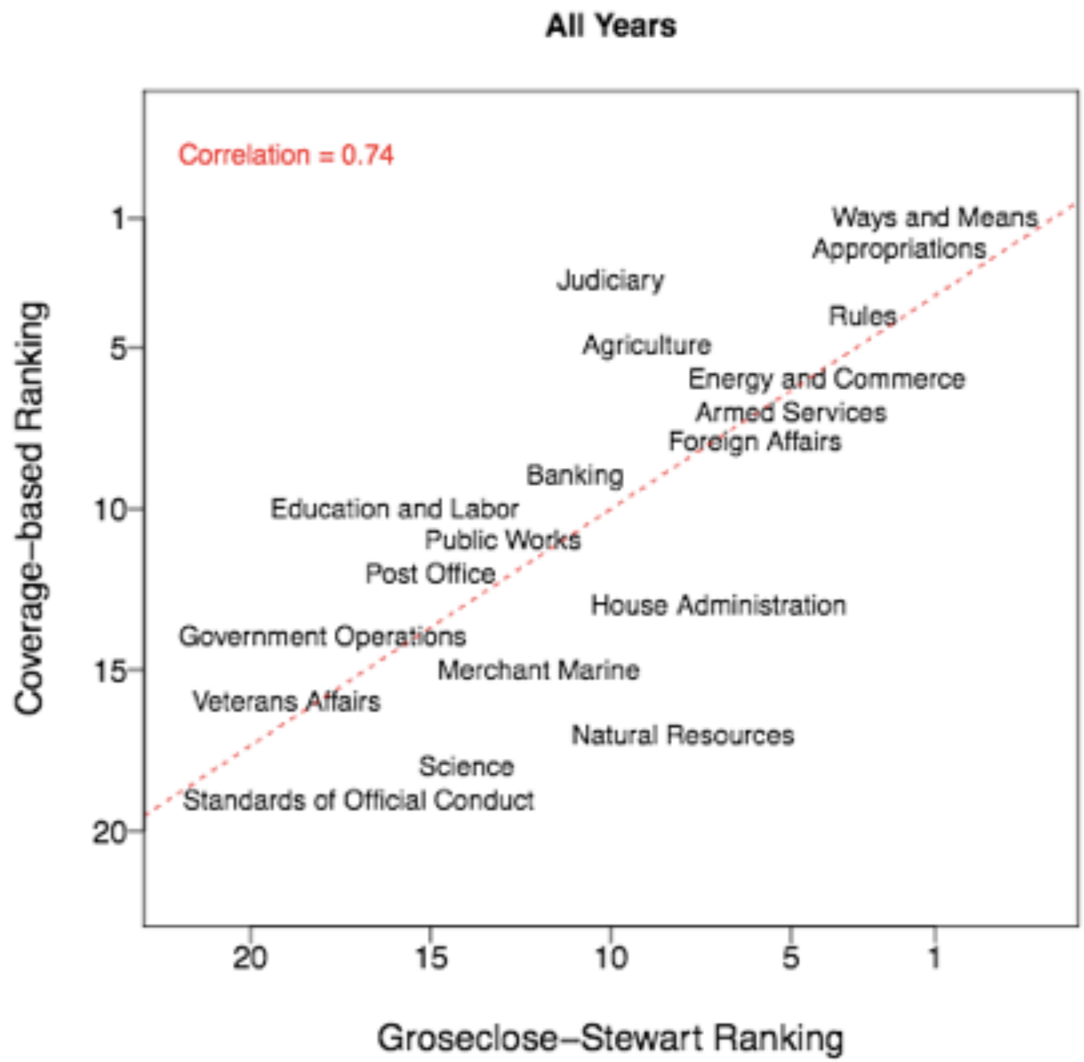
Ban et al (2015): validation: do “mentions” measure power?

1. Do mayor’s mentions go down when city shifts power to a city manager?



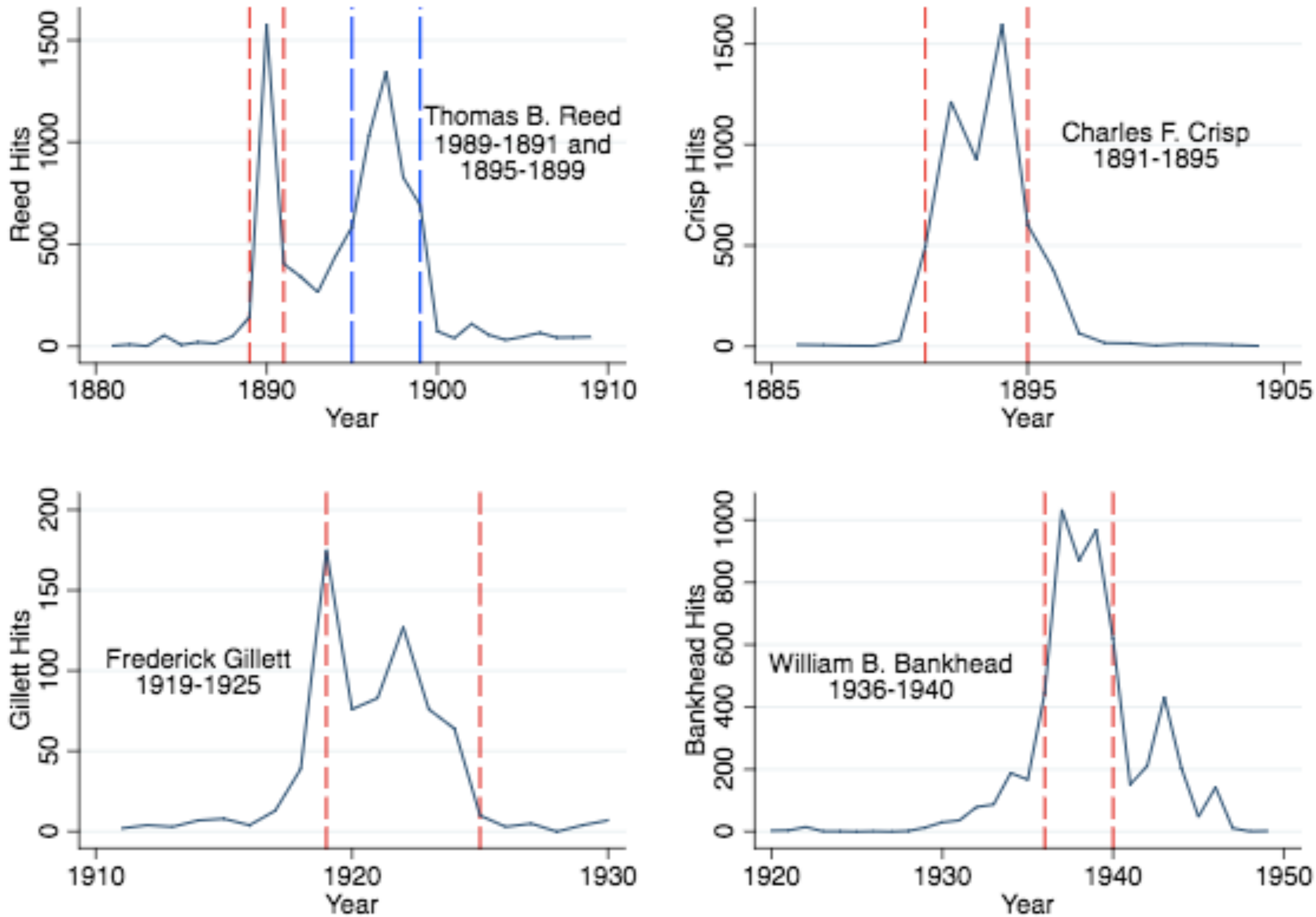
Ban et al (2015): validation: do “mentions” measure power?

2. Do congressional committees recognized as powerful get mentioned more?



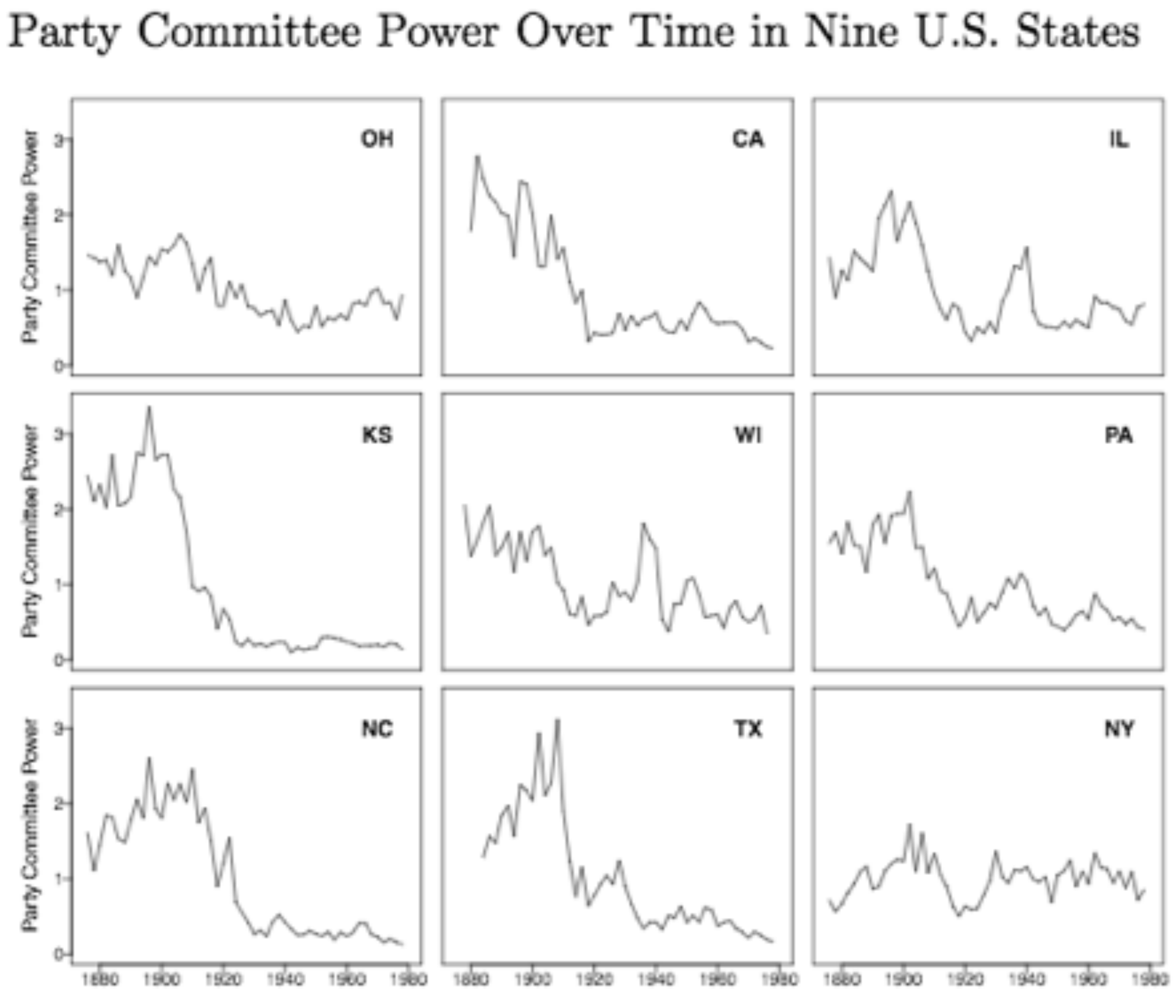
Ban et al (2015): validation: do “mentions” measure power?

3. Do members of Congress get mentioned more when they occupy leadership positions?



Ban et al (2015): validation: do "mentions" measure power?

4. How well does measure of party committee power correlate with Mayhew's TPO scores for 1966-1970? [corr > .5]



Resources for learning these tools

- Google and the internet: endless tutorials, help pages, etc
- Standard texts for getting started in R, Ruby, Python etc
- in R
  - stringr (for basic text stuff, regular expressions)
  - rvest (for web scraping)
  - Simon Jackman (2006), "Data from the web into R" [old school, but still good on basic process]
  - Gaston Sanchez (2013), "Handling and processing strings in R"
  - Pablo Barberá (2013), "Scraping twitter and web data using R"
- Chris Hanretty (2013), "Scraping the web for arts and humanities" [Python]

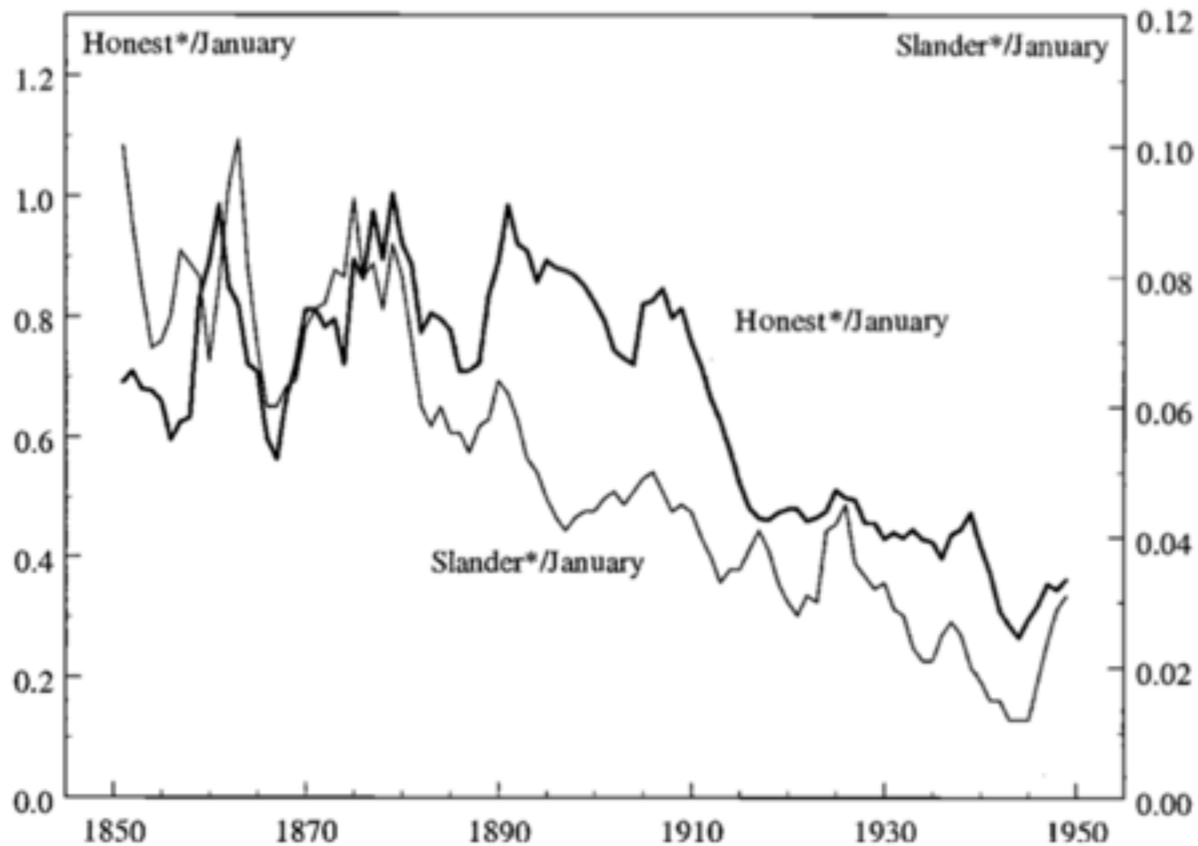
Take-aways for today

- content analysis is exciting and promising
- research is research:
  - big data + amazing stats + boring question = boring
  - big data + amazing stats + bad research design = bad
- there are many fancy things to do (we'll talk about them)
- before doing those things, you often have to un-fancy things: collecting data, counting things
- some of the best research involving text does **nothing** fancy

Simple example of dictionary methods: Gentzkow et al ("How newspapers became informative and why it mattered", 2006)

Evidence for a rise in unbiased/informative reporting in U.S. media 1850-1950:

- more papers without explicit political affiliations
- in ancestry.com's database of scanned newspaper articles, less use of "honest" & "slander" relative to "January":



Gentzkow et al continued

Alternative explanation: general change in use of these words.

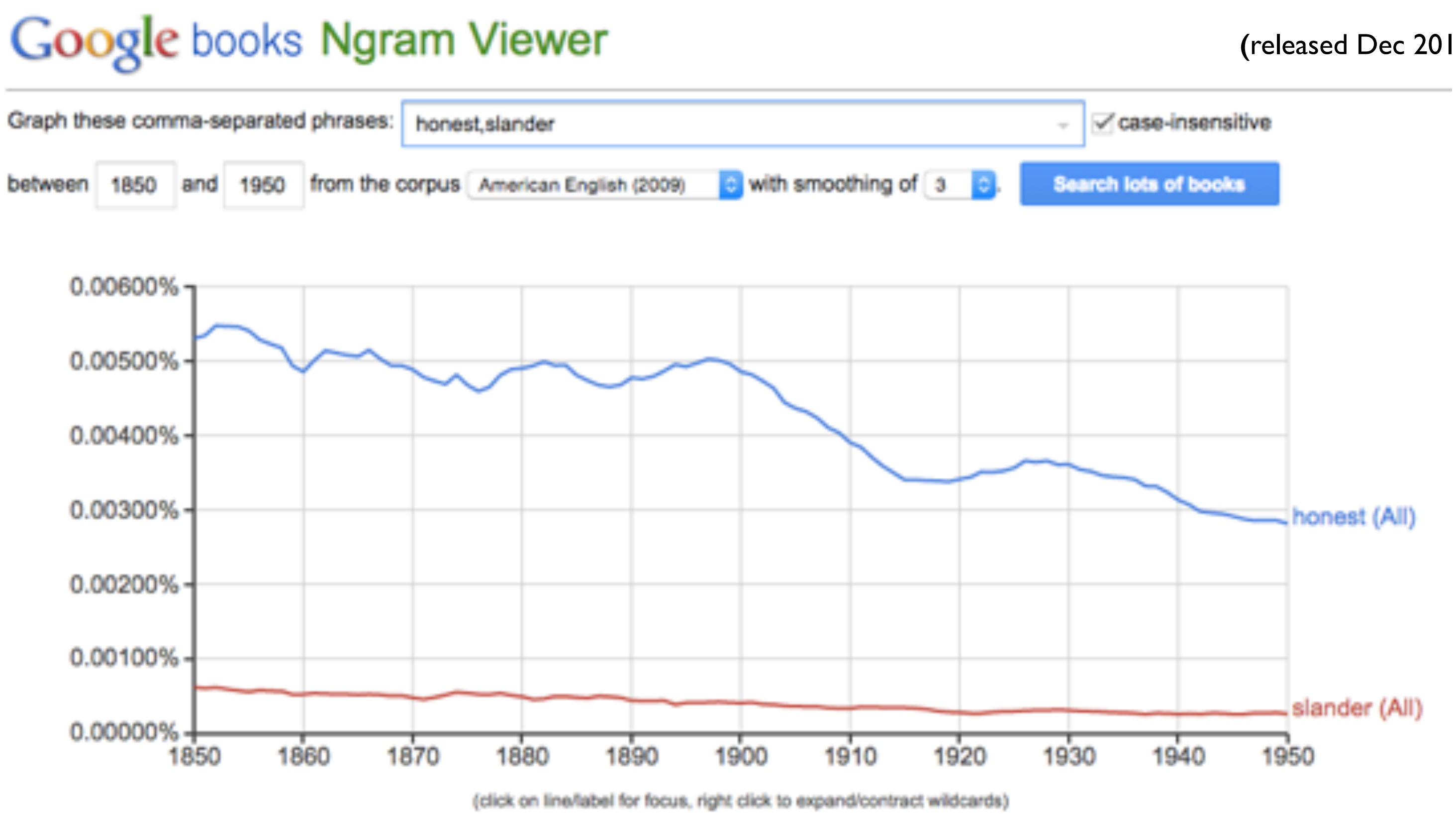
The general usage of charged and emotional words did change in the nineteenth century, but the change preceded that in the political press by about a half century. (Gentzkow et al, 195)

Gentzkow et al continued

Alternative explanation: general change in use of these words.

The general usage of charged and emotional words did change in the nineteenth century, but the change preceded that in the political press by about a half century. (Gentzkow et al, 195)

(released Dec 2010)



A word about n-grams and the “bag of words”

n-gram: continuous sequence of n words

The phrase “continuous sequence of n words” contains the following n-grams:

- **unigrams**: continuous, sequence, of, n, words
- **bigrams**: continuous sequence, sequence of, of n, n words
- **trigrams**: continuous sequence of, sequence of n, of n words

The **bag of words** maintains word order only within n-grams.

A word about n-grams and the "bag of words"

n-gram: continuous sequence of n words

The phrase "continuous sequence of n words" contains the following n-grams:

- **unigrams**: continuous, sequence, of, n, words
- **bigrams**: continuous sequence, sequence of, of n, n words
- **trigrams**: continuous sequence of, sequence of n, of n words

The **bag of words** maintains word order only within n-grams.

```
> t = "ask not what your country can do for you; ask what you can do for
your country"
>
> table(strsplit(t, "\\s+")[[1]])

     ask      can  country       do      for      not     what      you
       2        2        2        2        2        1        2        1
    you;     your
       1        2
>
> require(tau)
> table(tokenize(t))

                ;      ask      can  country       do      for      not
      16        1        2        2        2        2        2        1
    what      you     your
       2        2        2
```

A word about n-grams and the "bag of words"

n-gram: continuous sequence of n words

The phrase "continuous sequence of n words" contains the following n-grams:

- **unigrams**: continuous, sequence, of, n, words
- **bigrams**: continuous sequence, sequence of, of n, n words
- **trigrams**: continuous sequence of, sequence of n, of n words

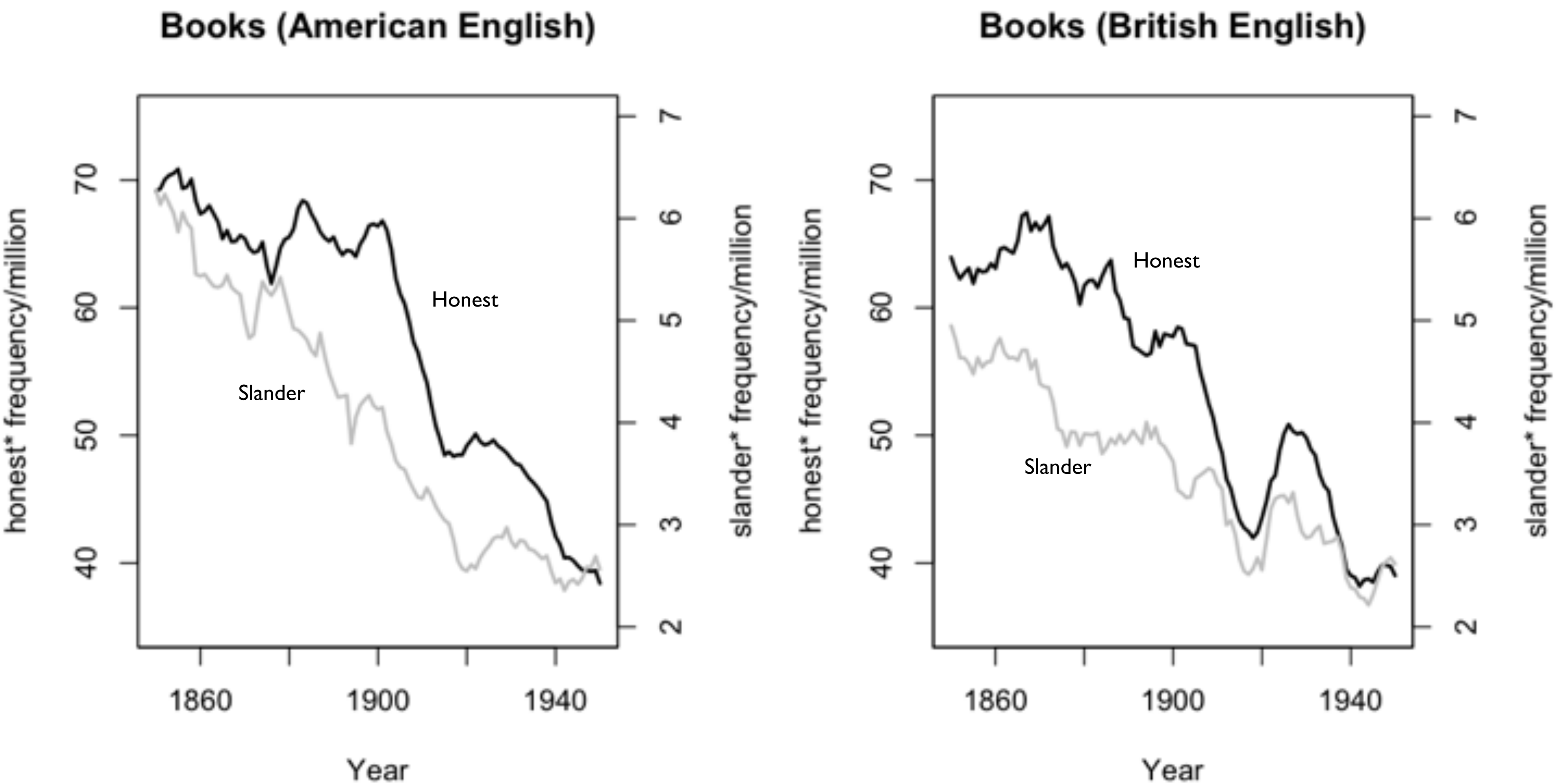
The **bag of words** maintains word order only within n-grams.

```
> t = "ask not what your country can do for you; ask what you can do for your country"
>
> table(strsplit(t, "\\s+")[[1]])

    ask     can country      do     for     not    what     you
      2       2       2       2       2       1       2       1
   you;    your
      1       2
>
> require(tau)
> table(tokenize(t))

              ;     ask     can country      do     for     not
     16       1       2       2       2       2       2       1
   what     you    your
      2       2       2
```

```
> bigrams = function(text){
+   word.vec = strsplit(text, "\\s+")[[1]]
+   out = c()
+   for(i in 1:(length(word.vec) - 1)){
+       out = c(out, paste(word.vec[i], word.vec[i+1]))
+   }
+   out
+ }
>
> table(bigrams(t))

     ask not      ask what        can do   country can        do for
           1             1             2             1             2
    for you;      for your      not what      what you     what your
           1             1             1             1             1
     you can  you; ask your country
           1             1             2
```

ngramr: an R interface for Google Ngram database

```
> require(ngramr)
> phrases = c("honest", "honesty", "honestly", "slander", "slanderous", "january") # the words we want
to look up
> # download these counts from Google for US and GB corpora -- takes a little while
> us_eng = ngram(phrases, corpus = "eng_us_2012", year_start = 1850, year_end = 1950, smoothing = 3,
case_ins = T)
> gb_eng = ngram(phrases, corpus = "eng_gb_2012", year_start = 1850, year_end = 1950, smoothing = 3,
case_ins = T)
>
> head(us_eng)
Phrases: honest, honesty, honestly, slander, slanderous, january
Case-sentitive: TRUE
Corpuses: eng_us_2012
Smoothing: 3

  Year Phrase Frequency    Corpus
1 1850 honest 5.019837e-05 eng_us_2012
2 1851 honest 5.058463e-05 eng_us_2012
3 1852 honest 5.133951e-05 eng_us_2012
4 1853 honest 5.175438e-05 eng_us_2012
5 1854 honest 5.200948e-05 eng_us_2012
6 1855 honest 5.235635e-05 eng_us_2012
> table(us_eng$Phrase)

    honest     honesty    honestly     slander slanderous     january
       101         101         101         101         101         101
```

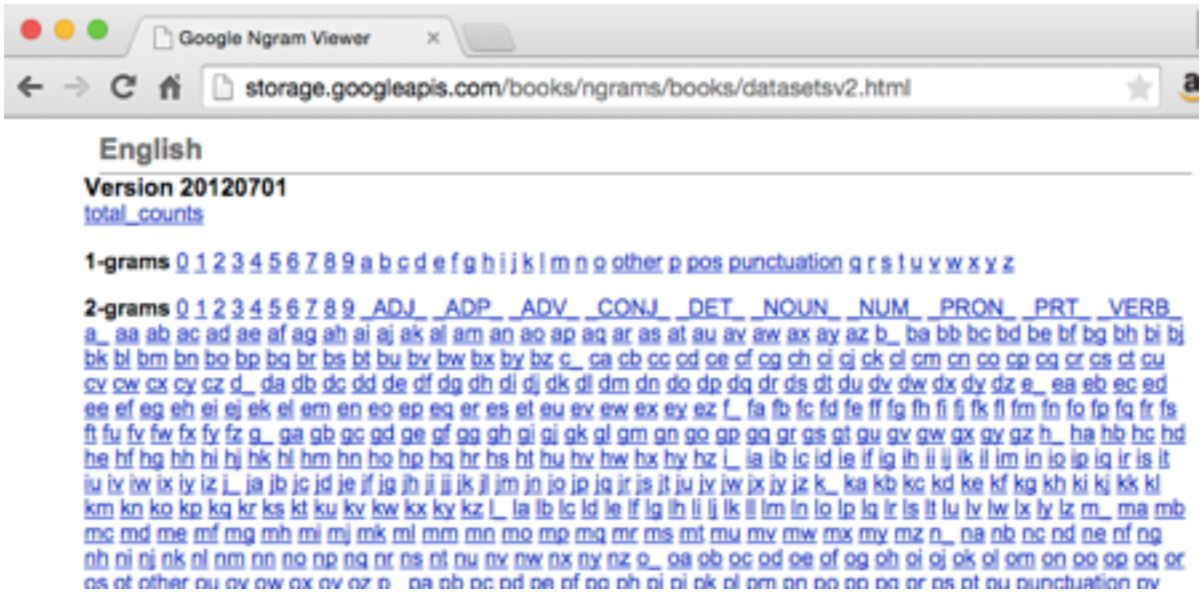


Books (American English)
Honest
Slander
honest* frequency/million
slander* frequency/million
Year
1860
1900
1940
Books (British English)
Honest
Slander
honest* frequency/million
slander* frequency/million
Year
1860
1900
1940

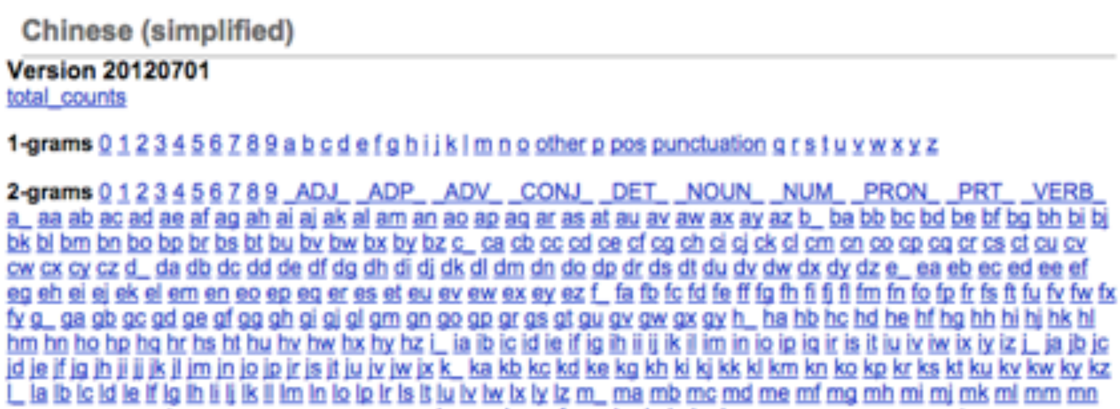

Books (American English)
Books (British English)
Books (fiction)
honest* frequency/million
slander* frequency/million
Honest
Slander
Year
1860
1880
1900
1920
1940

What can Google Ngrams do for you?

A panel dataset: word frequency in the Google Books corpus by ngram-year…



…and not just in English!



But. . .

- Doesn't give you a good research question
- Defines the corpus for you; you may want something narrower
- Handles a huge data processing challenge (scanning, counting) but leaves you with another: "s" unigrams file in English is 2.3G

What lessons to draw from Gentzkow et al?

- **To admire**: creative and sensible-seeming measure, linked to interesting research question
- **To criticize**: *validity* (and *validation*) of the measure

How do we assess the validity of a new measure?

Tricky problem!

"We have no valid measures of the informativeness of media, so I propose X."

"Does it work?"

"I don't know, because we have no valid measures of the informativeness of media."