

Content Analysis

Lecture 1: Turning text into data

25 April, 2016

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 content analysis, 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 content analysis, we're (mostly) talking about efforts to solve **measurement problems involving text.**

Otherwise, asking questions about the discourse itself that we are far from being able to answer with automated methods.

Components of measurement

Components of measurement

Conceptualization: precisely characterizing what it is you are trying to measure

e.g. *“democracy is a system of government in which elites compete for power”*



Components of measurement

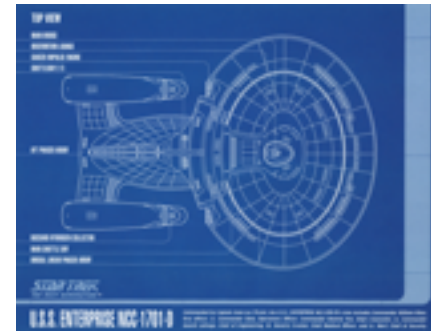
Conceptualization: precisely characterizing what it is you are trying to measure

e.g. “democracy is a system of government in which elites compete for power”



Operationalization: developing specific research procedures that will produce a valid measure of the concept

e.g. “we call it a democracy if chief exec and legislature elected, more than one party, and at least one past episode of power alternation” (paraphrasing Alvarez et al 1996)



Components of measurement

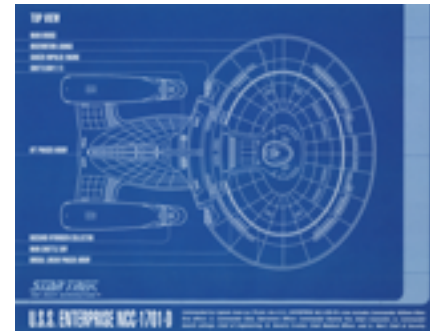
Conceptualization: precisely characterizing what it is you are trying to measure

e.g. “democracy is a system of government in which elites compete for power”



Operationalization: developing specific research procedures that will produce a valid measure of the concept

e.g. “we call it a democracy if chief exec and legislature elected, more than one party, and at least one past episode of power alternation” (paraphrasing Alvarez et al 1996)



Implementation: executing those procedures

e.g. correspondence with country experts, reports from RAs

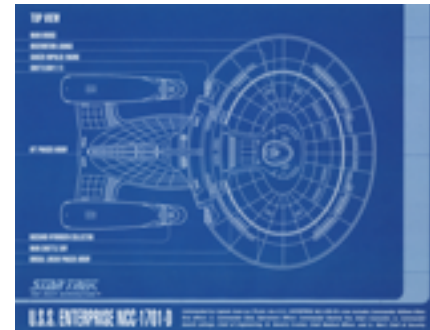


What's new?

Conceptualization



Operationalization



Implementation



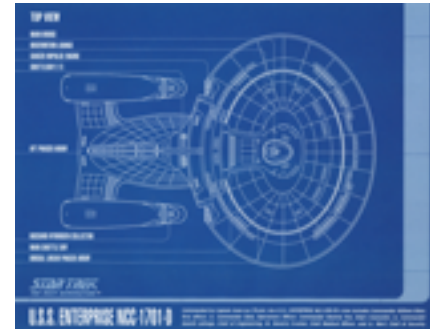
What's new?

Using text for measurement has become cheaper at the implementation stage

Conceptualization



Operationalization



Implementation



What's new?

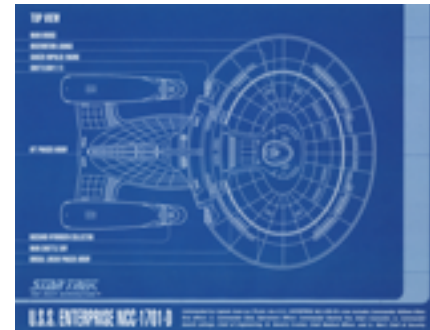
Using text for measurement has become cheaper at the **implementation** stage

- Machine-readable text exponentially proliferating

Conceptualization



Operationalization



Implementation



What's new?

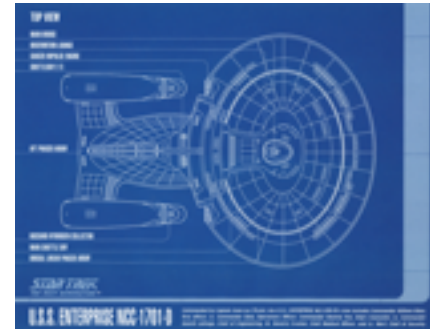
Using text for measurement has become cheaper at the **implementation** stage

- Machine-readable text exponentially proliferating
- Cost of memory and storage exponentially dropping

Conceptualization



Operationalization



Implementation



What's new?

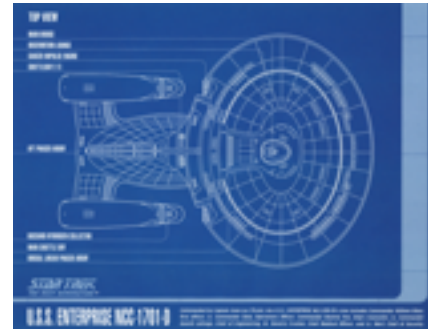
Using text for measurement has become cheaper at the **implementation** stage

- Machine-readable text exponentially proliferating
- Cost of memory and storage exponentially dropping
- Start-up costs for writing code much lower

Conceptualization



Operationalization



Implementation



What's new?

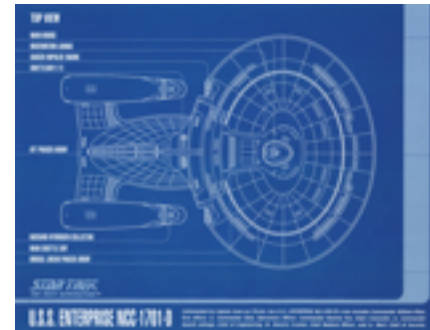
Using text for measurement has become cheaper at the **implementation** stage

- Machine-readable text exponentially proliferating
- Cost of memory and storage exponentially dropping
- Start-up costs for writing code much lower
- Software packages proliferating, helping with collection, manipulation, analysis

Conceptualization



Operationalization



Implementation



What's new?

Using text for measurement has become cheaper at the **implementation** stage

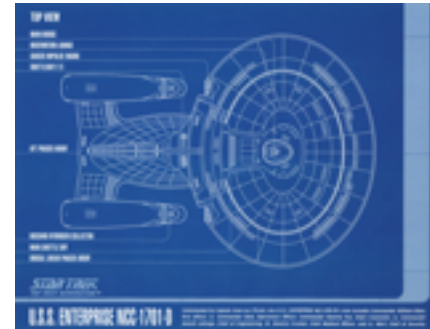
- Machine-readable text exponentially proliferating
- Cost of memory and storage exponentially dropping
- Start-up costs for writing code much lower
- Software packages proliferating, helping with collection, manipulation, analysis

Also, new models/techniques from statistics and machine learning affect the **operationalization** stage

Conceptualization



Operationalization



Implementation



What's new?

Using text for measurement has become cheaper at the **implementation** stage

- Machine-readable text exponentially proliferating
- Cost of memory and storage exponentially dropping
- Start-up costs for writing code much lower
- Software packages proliferating, helping with collection, manipulation, analysis

Also, new models/techniques from statistics and machine learning affect the **operationalization** stage

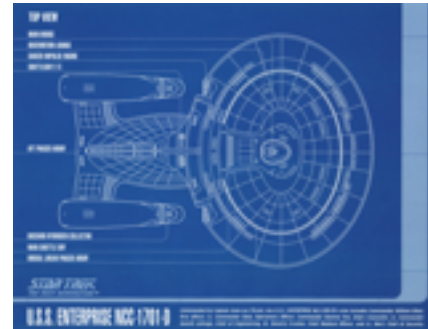


Text being used more widely as source of measures

Conceptualization



Operationalization



Implementation



What's new?

Using text for measurement has become cheaper at the **implementation** stage

- Machine-readable text exponentially proliferating
- Cost of memory and storage exponentially dropping
- Start-up costs for writing code much lower
- Software packages proliferating, helping with collection, manipulation, analysis

Also, new models/techniques from statistics and machine learning affect the **operationalization** stage



Text being used more widely as source of measures

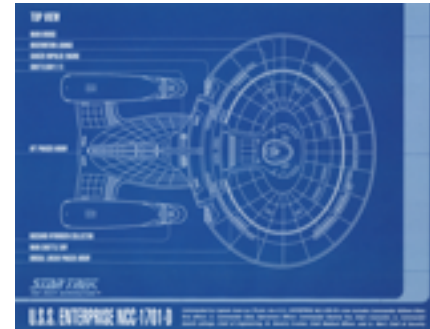


Text-based measures increasingly built on *counting* rather than *reading*

Conceptualization



Operationalization



Implementation



What's new?

Using text for measurement has become cheaper at the **implementation** stage

- Machine-readable text exponentially proliferating
- Cost of memory and storage exponentially dropping
- Start-up costs for writing code much lower
- Software packages proliferating, helping with collection, manipulation, analysis

Also, new models/techniques from statistics and machine learning affect the **operationalization** stage



Text being used more widely as source of measures



Text-based measures increasingly built on *counting* rather than *reading*

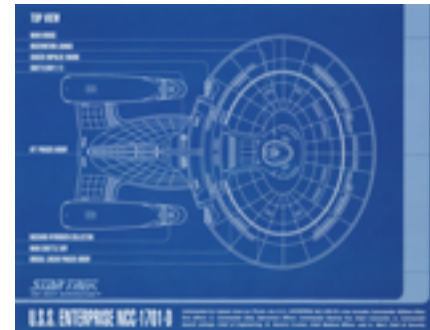


Statistics used in text-based measures has gotten more sophisticated

Conceptualization



Operationalization



Implementation

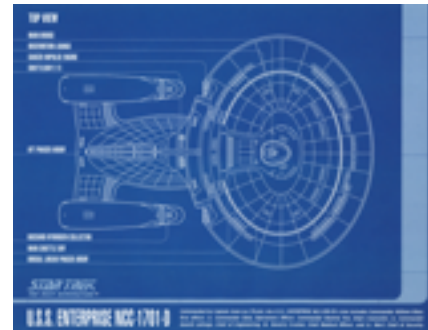


What's not new?

Conceptualization



Operationalization



Implementation



What's not new?

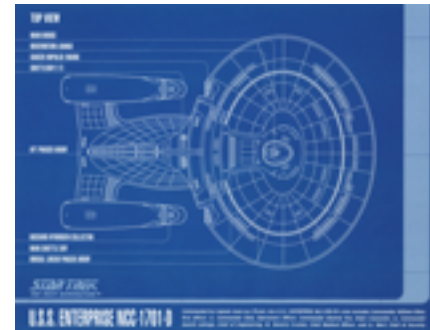
- Uninteresting concepts produce uninteresting measures, regardless of quality of operationalization and implementation

(What is new? More boring papers involving text!)

Conceptualization



Operationalization



Implementation



What's not new?

- Uninteresting concepts produce uninteresting measures, regardless of quality of operationalization and implementation

(What is new? More boring papers involving text!)

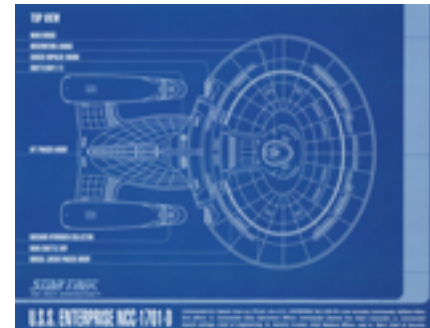
- Bad research designs produce few insights, regardless of quality of question and measurements

(Causal inference has not gotten easier — though maybe some covariates can now be measured?)

Conceptualization



Operationalization



Implementation



What's not new?

- Uninteresting concepts produce uninteresting measures, regardless of quality of operationalization and implementation

(What is new? More boring papers involving text!)

- Bad research designs produce few insights, regardless of quality of question and measurements

(Causal inference has not gotten easier — though maybe some covariates can now be measured?)

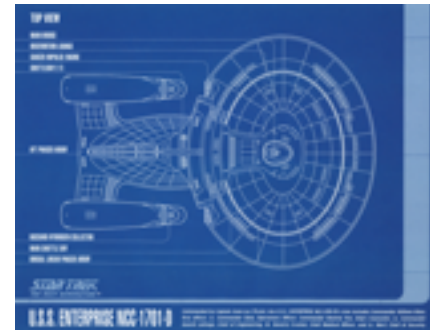
- Description *per se* is often very important!

(Not everyone agrees with that, or realizes they do.)

Conceptualization



Operationalization



Implementation



What's not new?

- Uninteresting concepts produce uninteresting measures, regardless of quality of operationalization and implementation

(What is new? More boring papers involving text!)

- Bad research designs produce few insights, regardless of quality of question and measurements

(Causal inference has not gotten easier — though maybe some covariates can now be measured?)

- Description *per se* is often very important!

(Not everyone agrees with that, or realizes they do.)

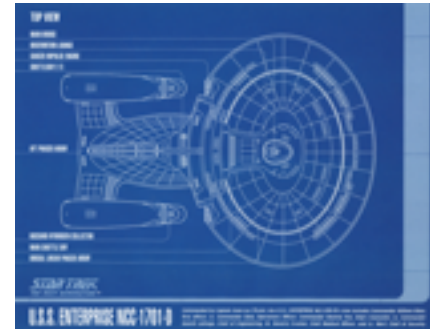
- Slippage at each stage of measurement process is problematic

- Do the different words used by Republicans and Democrats reflect different ideologies?
- What is the text you are analyzing representative of?

Conceptualization



Operationalization



Implementation



We want to help you become...

...more literate & critical as **consumers** of research that uses text.

- How confident should I be in a given text-based measure?
- How should these measures be validated?
- What is topic modeling?
- What is text scaling?

...more effective as **users** of text in your own research.

- Is there a way I could use text for my research problem?
- Is there a cheaper way to do what I'm trying to do with text?
- Are there interesting research questions I haven't considered that involve text?

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 in a given text? (keyness, specificity, weirdness, e.g. “Fightin’ Words”)

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 in a given text? (keyness, specificity, weirdness, e.g. “Fightin’ Words”)
- **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

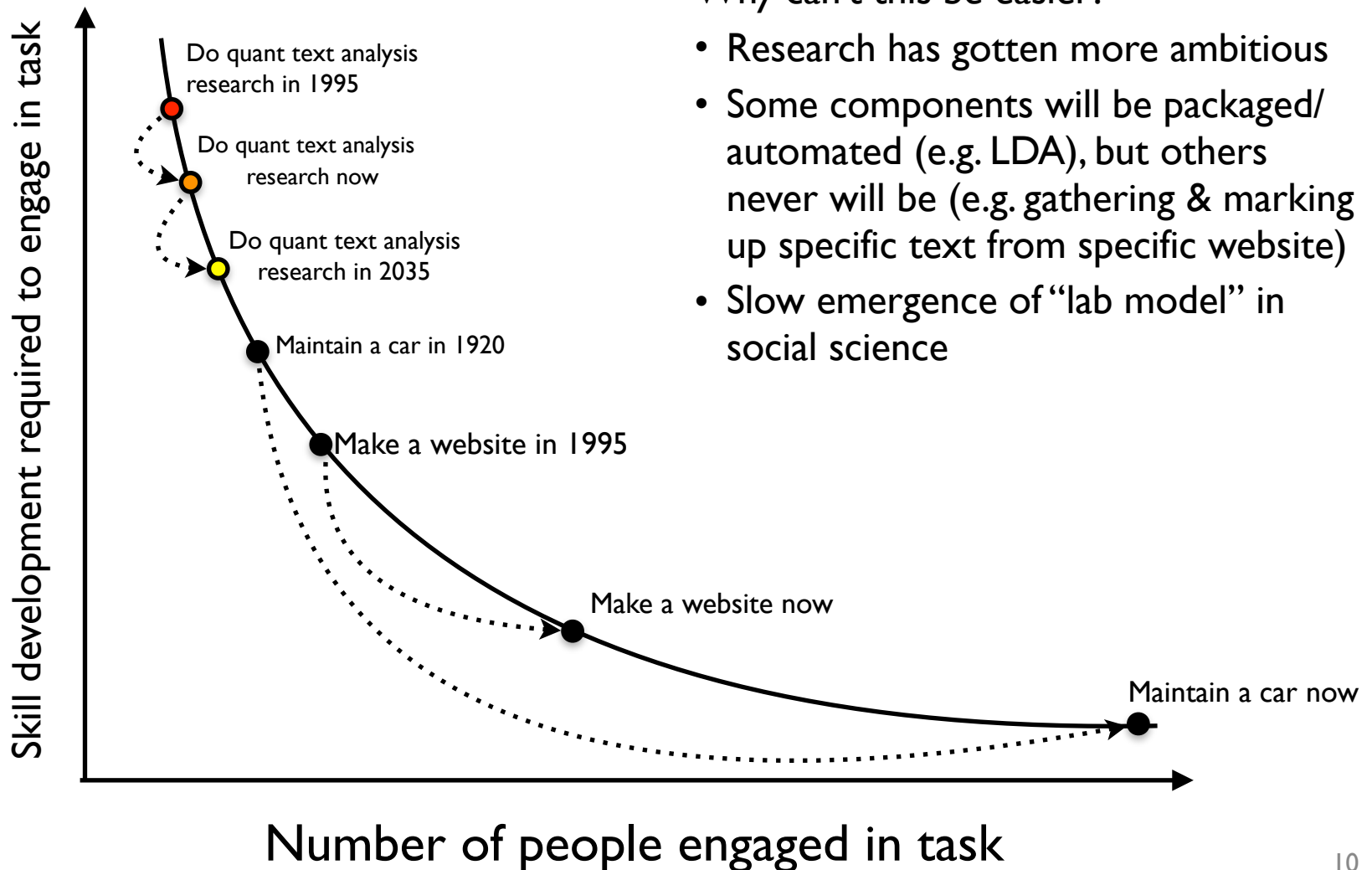
- **Frequencies:** How often does this term/theme, or set of terms/themes, appear in the text?
 - Are the themes identified in the text by readers? → qualitative data analysis/QDA, software like MaxQDA, NVivo, Atlas.ti
 - Are the terms/themes identified in the text via software → dictionary methods, 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 in a given text? (keyness, specificity, weirdness, e.g. “Fightin’ Words”)
- **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)

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
 - 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.

The continued importance of programming skills



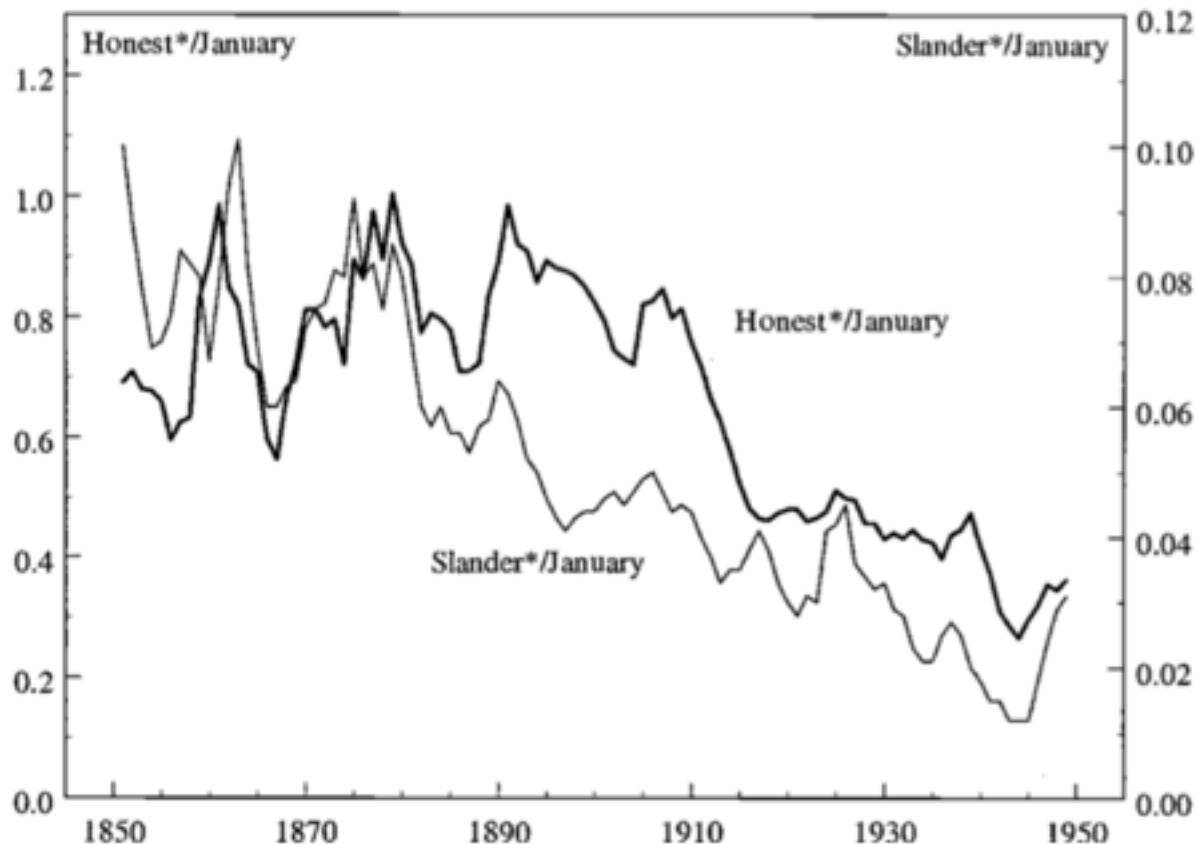
Why can't this be easier?

- Research has gotten more ambitious
- Some components will be packaged/ automated (e.g. LDA), but others never will be (e.g. gathering & marking up specific text from specific website)
- Slow emergence of “lab model” in social science

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, 1995)

Google books Ngram Viewer

(released Dec 2010)

Graph these comma-separated phrases: case-insensitive
between and from the corpus with smoothing of [Search lots of books](#)



(click on line/label for focus, right click to expand/contract wildcards)

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

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

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))

      16      ;      ask      can country      do      for      not
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
```

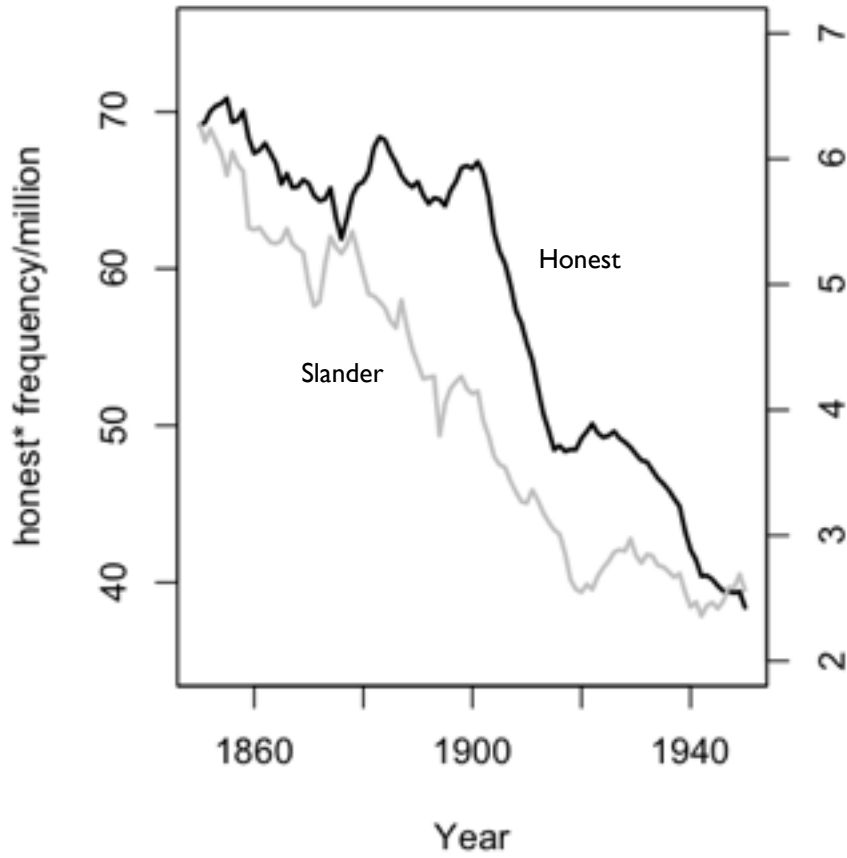
ngramr: an R interface for Google Ngram database

```
> 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-sensitive: TRUE
Corpora: eng_us_2012
Smoothing: 3

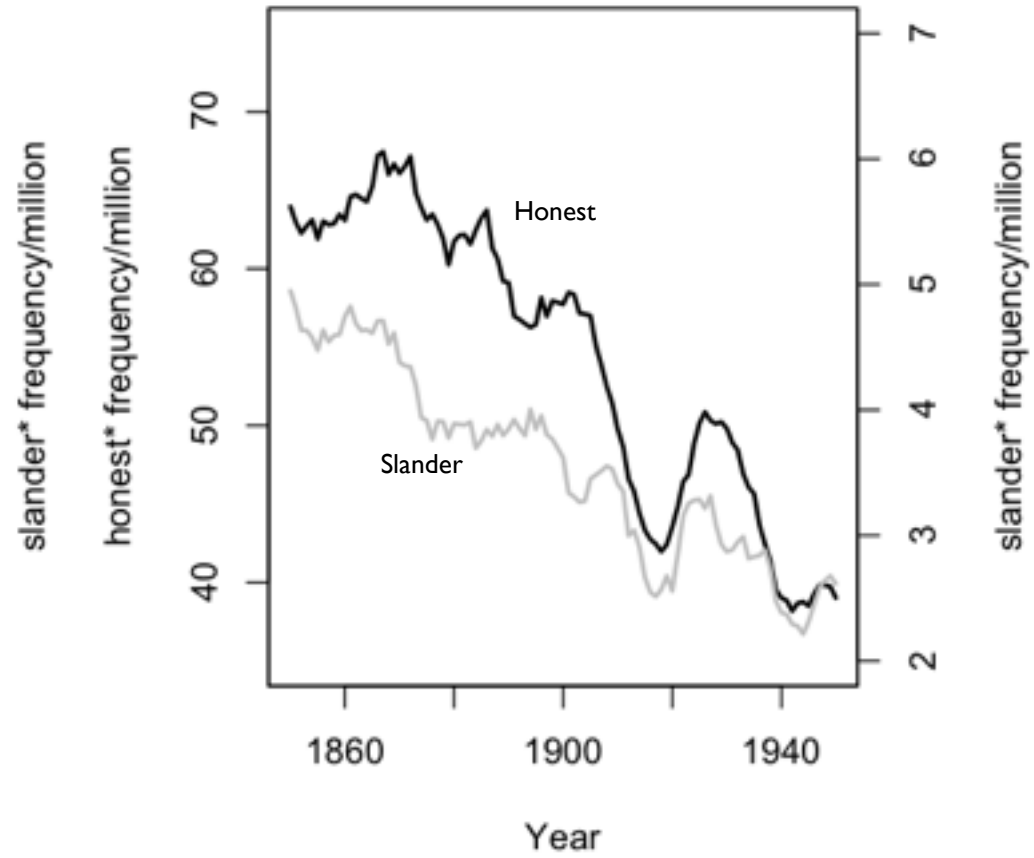
  Year Phrase Frequency      Corpus
1 1850 honest 5.019837e-05 eng_us_2012
2 1851 honest 5.058463e-05 eng_us_2012
3 1852 honest 5.133951e-05 eng_us_2012
4 1853 honest 5.175438e-05 eng_us_2012
5 1854 honest 5.200948e-05 eng_us_2012
6 1855 honest 5.235635e-05 eng_us_2012
> table(us_eng$Phrase)

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

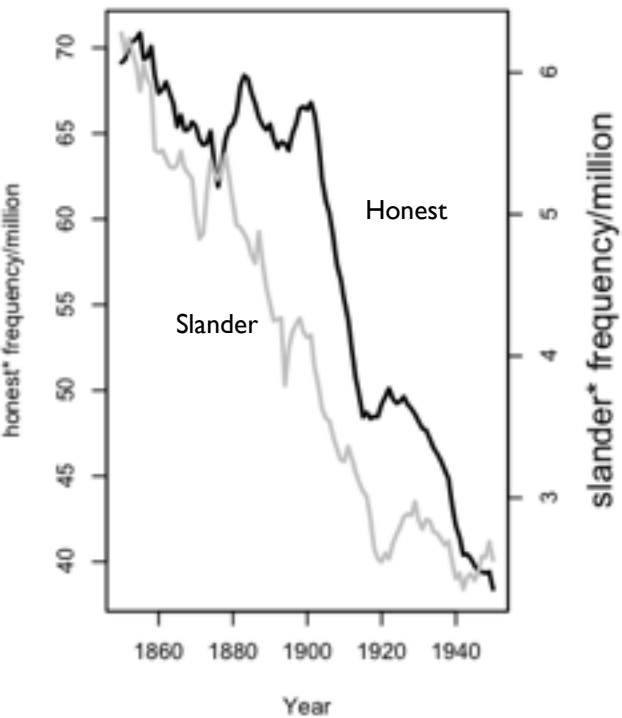

Books (American English)



Books (British English)



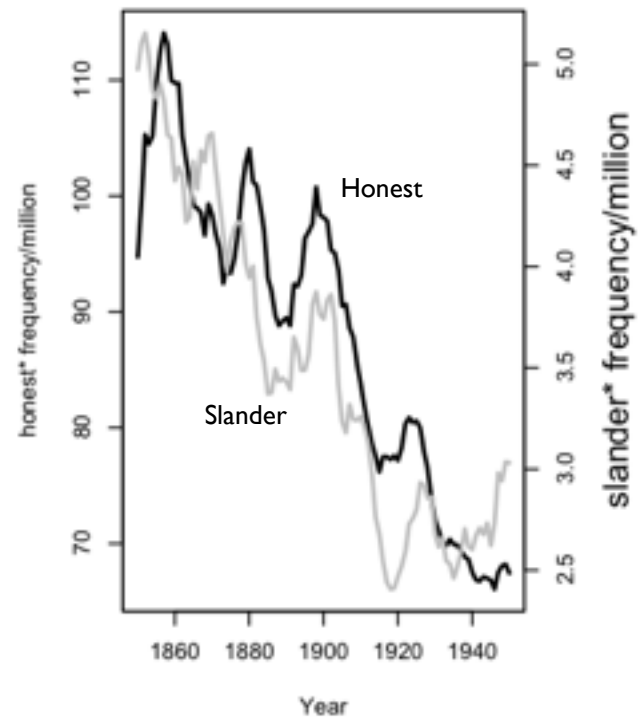
Books (American English)



Books (British English)

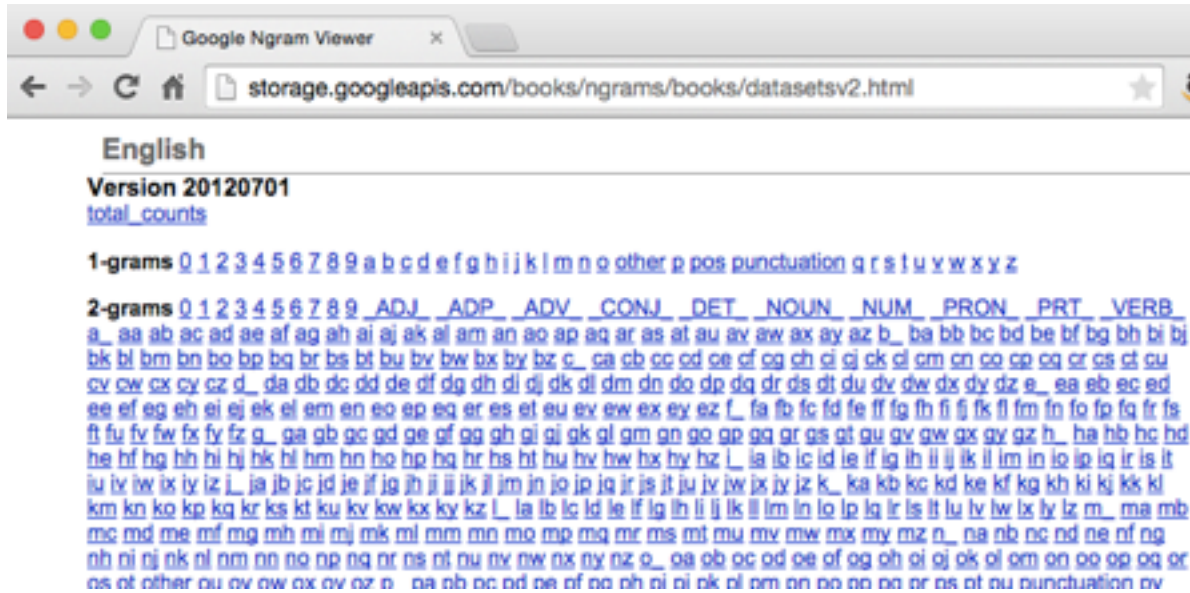


Books (fiction)



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!

Chinese (simplified)

Version 20120701

[total_counts](#)

1-grams 0 1 2 3 4 5 6 7 8 9 a b c d e f g h i j k l m n o other p pos punctuation q r s t u v w x y z

2-grams 0 1 2 3 4 5 6 7 8 9 _ADJ_ _ADP_ _ADV_ _CONJ_ _DET_ _NOUN_ _NUM_ _PRON_ _PRT_ _VERB_ a_ aa ab ac ad ae af ag ah ai aj ak al am an ao ap aq ar as at au av aw ax ay az b_ ba bb bc bd be bf bg bh bi bj bk bl bm bn bo bp bq br bs bt bu bv bw bx by bz c_ ca cb cc cd ce cf cg ch ci cj ck cl cm cn co cp cq cr cs ct cu cv cw cx cy cz d_ da db dc dd de df dg dh di dj dk dl dm dn do dp dr ds dt du dv dw dx dy dz e_ ea eb ec ed ee ef eg eh ei ej ek el em en eo ep eq er es et eu ev ew ex ey ez f_ fa fb fc fd fe ff fg fh fi fj fl fm fn fo fp fr fs ft fu fv fw fx fy g_ ga gb gc gd ge gf gg gh gi gj gl gm gn go gp gr gs gt gu gv gw gx gy h_ ha hb hc hd he hf hg hh hi hj hk hl hm hn ho hp hq hr hs ht hu hv hw hx hy hz l_ la lb lc ld le lf lg lh li lj lk ll lm ln lo lp lr ls lt lu lv lw lx ly lz m_ ma mb mc md me mf mg mh mi mj mk ml mm mn

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!

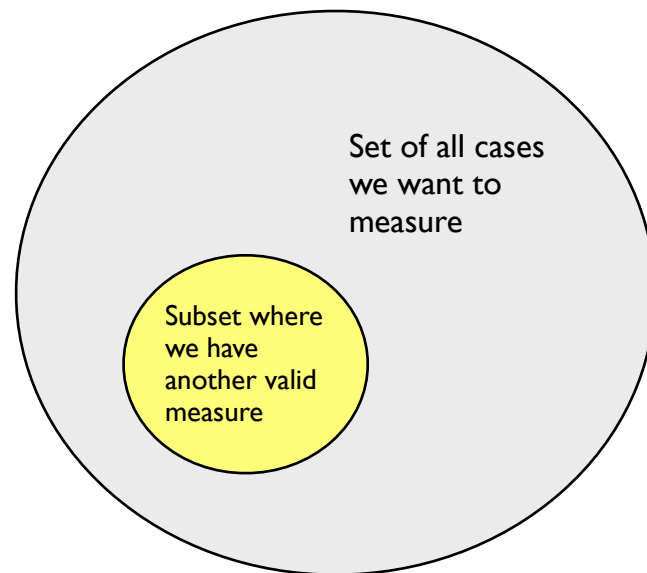
“We have no valid measures of the informativeness of media, so I propose X.”

“Does it work?”

“I don’t know, because we have no valid measures of the informativeness of media.”

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, Fournaies, 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?

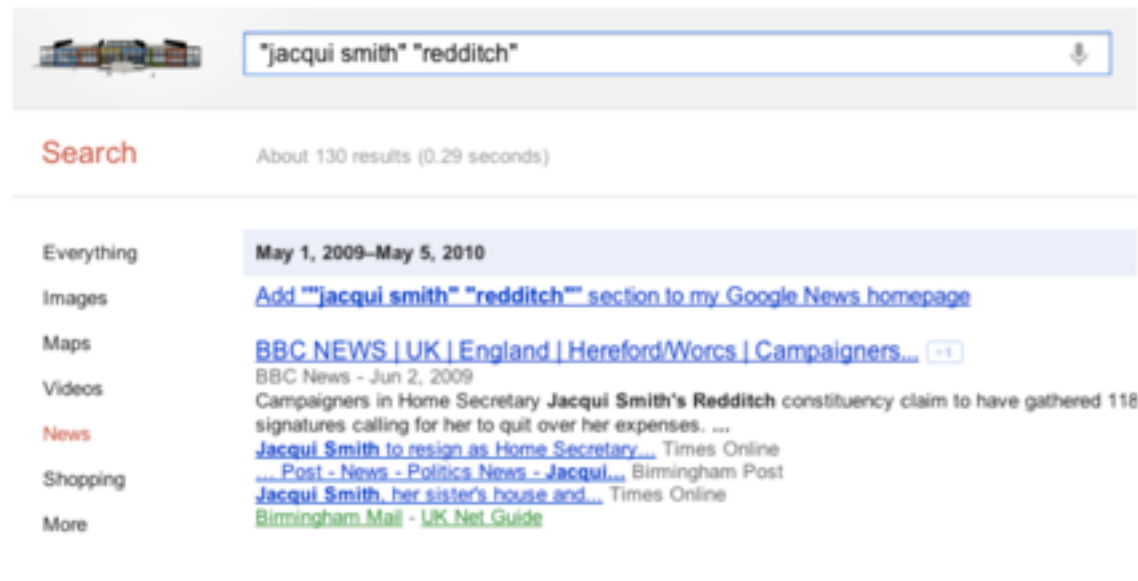
Measurement problem: How much was each MP implicated?

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

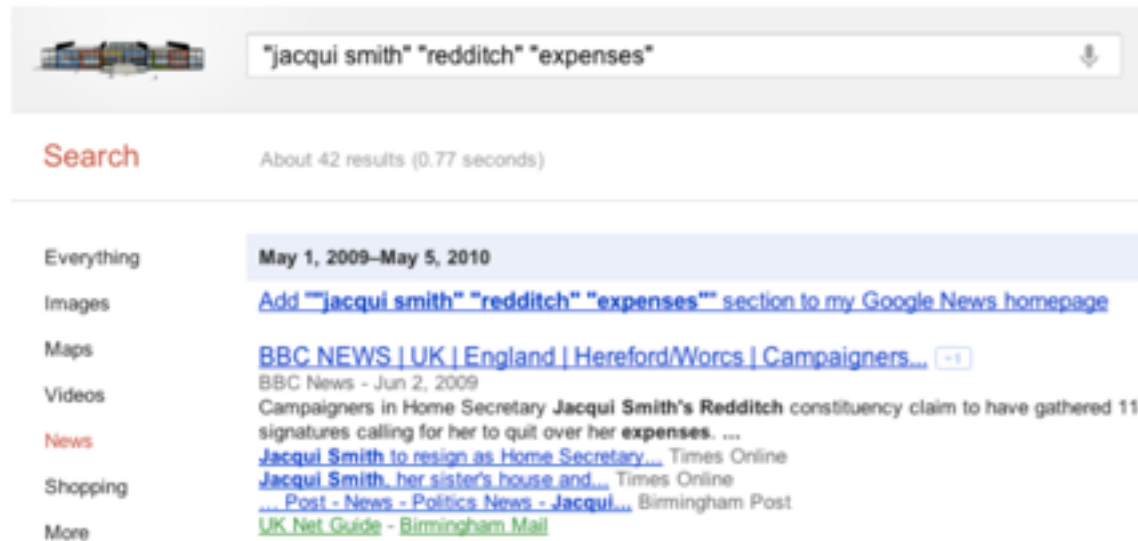


Step 1: count Google News hits for MP's name and constituency between scandal and election



The screenshot shows a Google search interface. The search bar contains the text "jacqui smith redditch". Below the search bar, the word "Search" is displayed in red, followed by the text "About 130 results (0.29 seconds)". On the left side, there is a vertical menu with options: "Everything", "Images", "Maps", "Videos", "News" (highlighted in red), "Shopping", and "More". The main content area shows a date range "May 1, 2009–May 5, 2010" and a link to "Add 'jacqui smith redditch' section to my Google News homepage". Below this, there are several news snippets, including one from BBC News dated June 2, 2009, titled "Campaigners in Home Secretary Jacqui Smith's Redditch constituency claim to have gathered 118 signatures calling for her to quit over her expenses." Other snippets include "Jacqui Smith to resign as Home Secretary..." from Times Online, "Jacqui Smith, her sister's house and..." from Times Online, and "UK Net Guide - Birmingham Mail" from Birmingham Mail.

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



The screenshot shows a Google search interface. The search bar contains the text "jacqui smith redditch expenses". Below the search bar, the word "Search" is displayed in red, followed by the text "About 42 results (0.77 seconds)". On the left side, there is a vertical menu with options: "Everything", "Images", "Maps", "Videos", "News" (highlighted in red), "Shopping", and "More". The main content area shows a date range "May 1, 2009–May 5, 2010" and a link to "Add 'jacqui smith redditch expenses' section to my Google News homepage". Below this, there are several news snippets, including one from BBC News dated June 2, 2009, titled "Campaigners in Home Secretary Jacqui Smith's Redditch constituency claim to have gathered 11 signatures calling for her to quit over her expenses." Other snippets include "Jacqui Smith to resign as Home Secretary..." from Times Online, "Jacqui Smith, her sister's house and..." from Times Online, "UK Net Guide - Birmingham Mail" from Birmingham Mail, and "UK Net Guide - Birmingham Mail" from Birmingham Mail.

Step 3: divide to get implication score

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

How to validate?

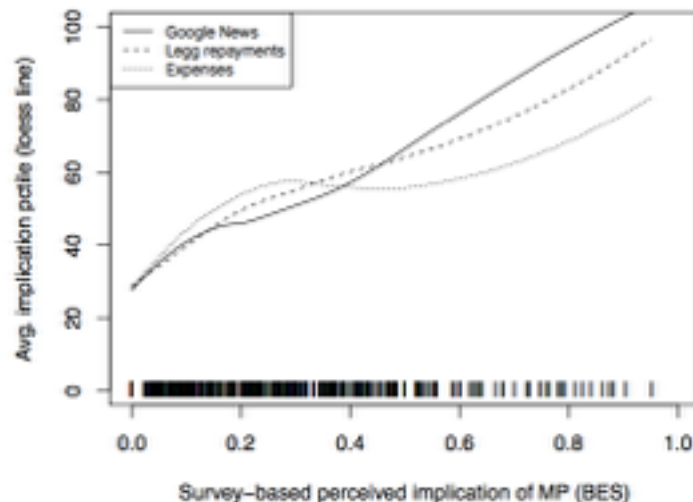
1. Compare with Telegraph's list of "saints" and "sinners"

2. Check list against substantive knowledge

(3. Assess correlation with other possible measures)

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



Example: Ban, Fournaies, 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



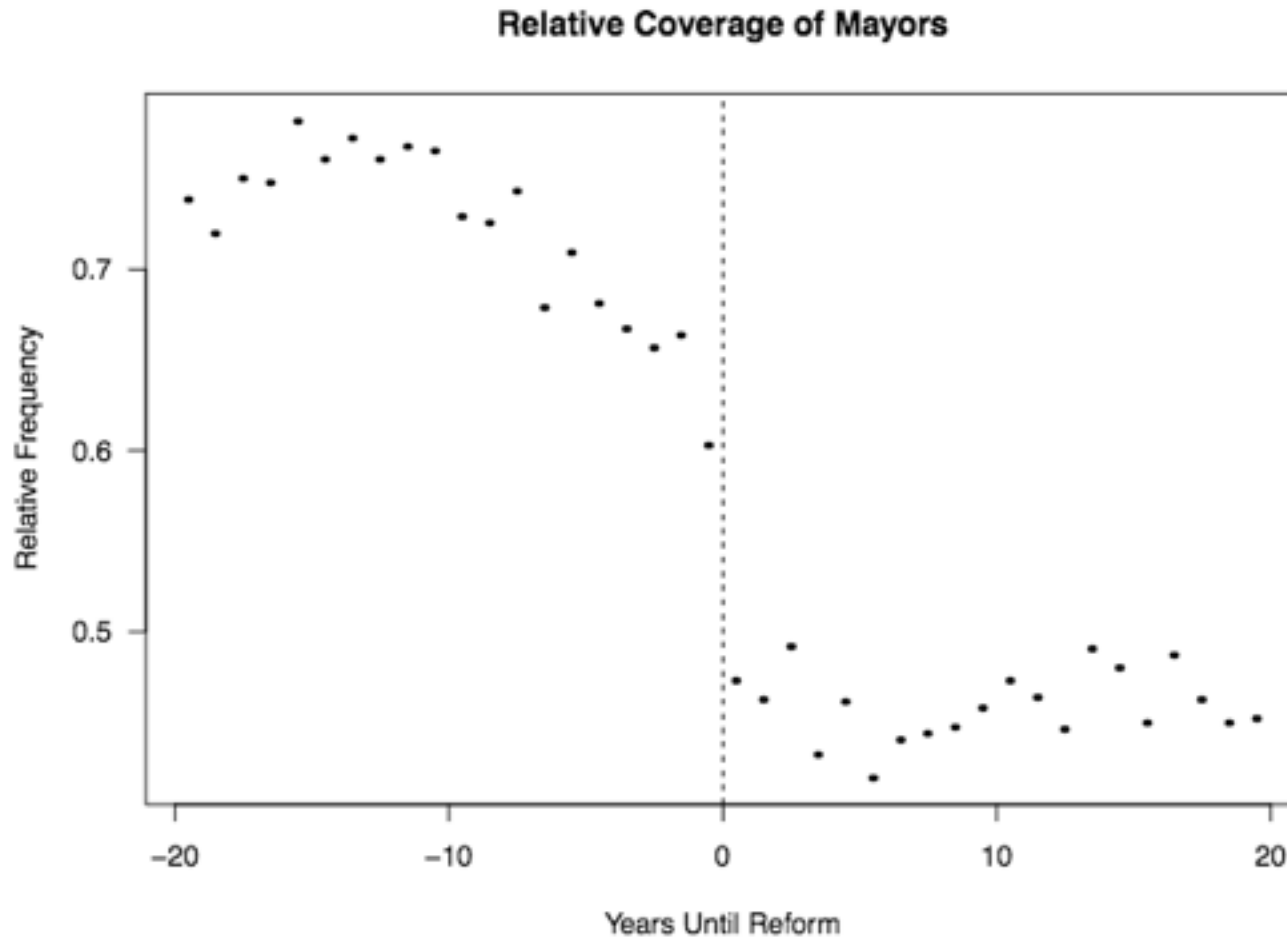
Ban et al (2015): Using newspaper mentions to measure power

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”

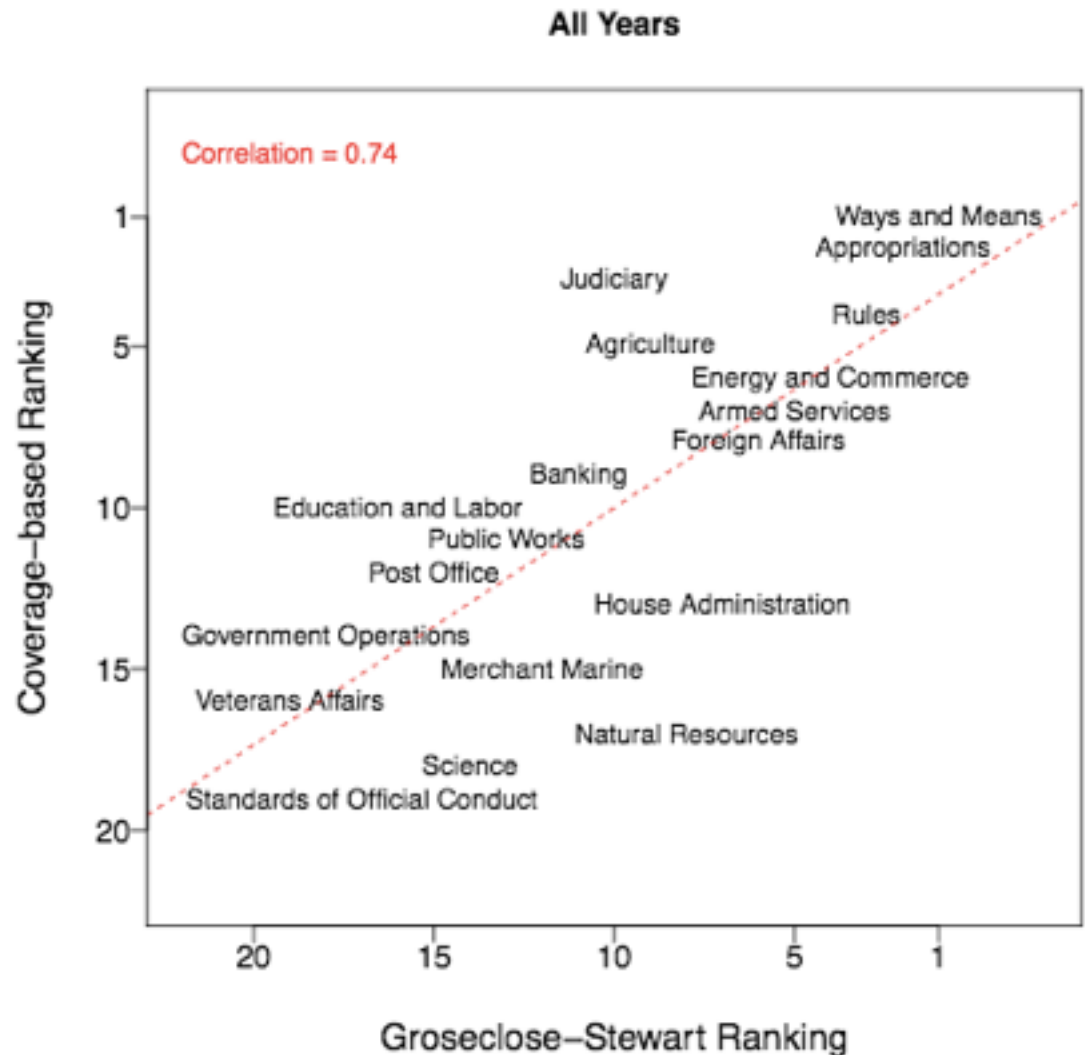
Ban et al (2015): validation: do “mentions” measure power?

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



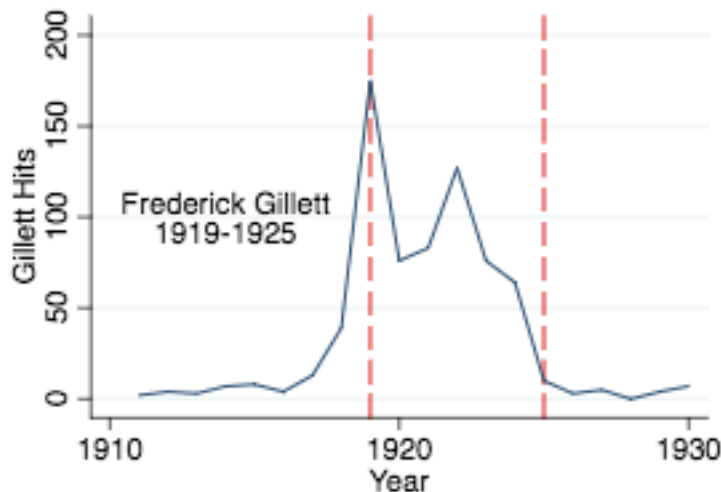
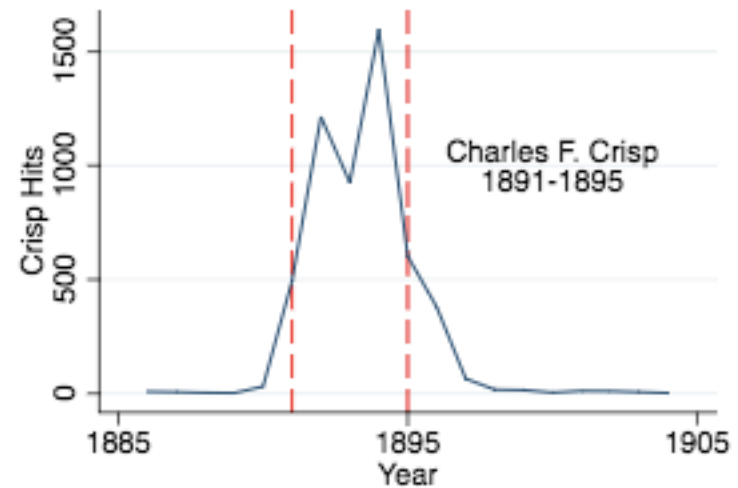
Ban et al (2015): validation: do “mentions” measure power?

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



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

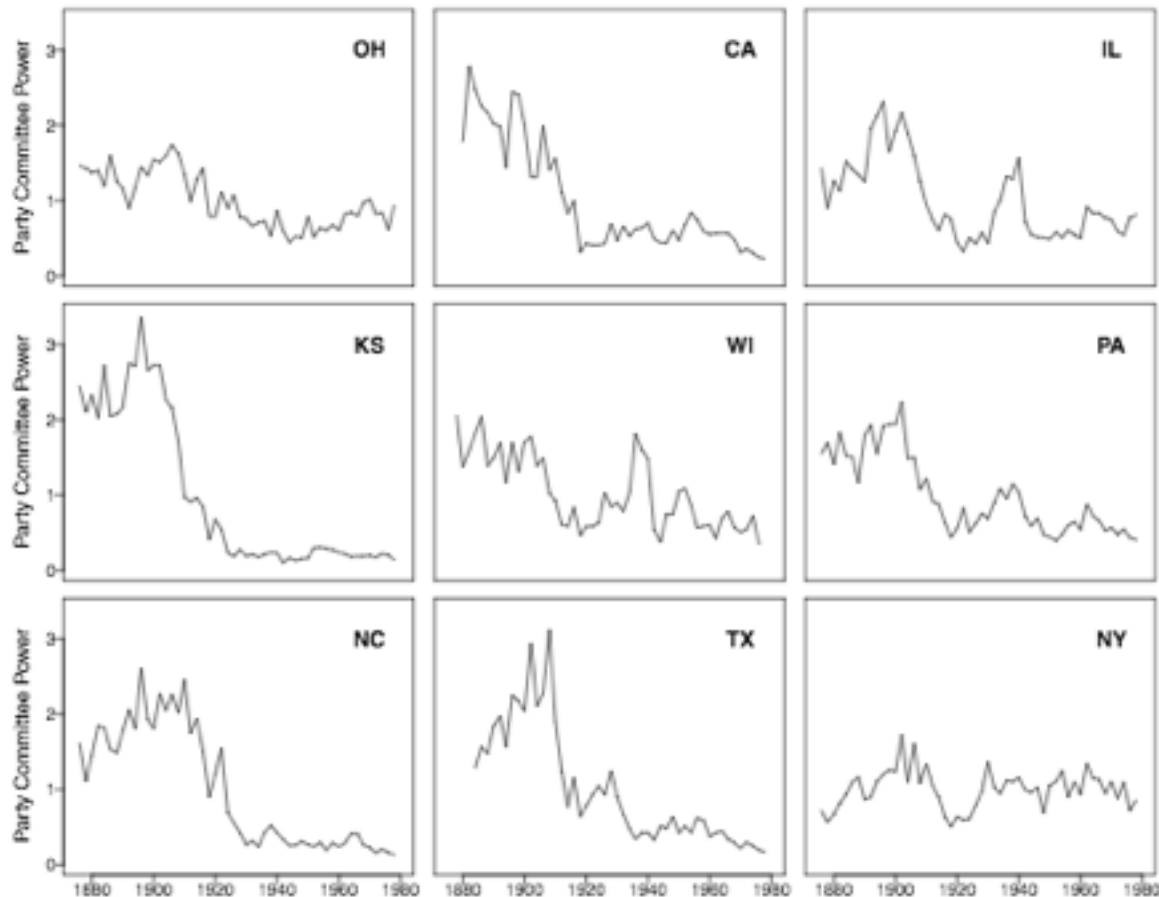
3. Do members of Congress get mentioned more when they occupy leadership positions?



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]

Party Committee Power Over Time in Nine U.S. States



Helpful skills for working with text

Collecting text

- Web scraping
 - Programming ability in something other than Stata: “for” loops, data structures, if-else logic, etc
 - Ability with packages that interact with the web: `RCurl` for R, `selenium` for Python, `mechanize` for Ruby, etc.
- Optical character recognition: getting data from books, files, etc

Working with text

- “Regular expressions”: absolutely indispensable; possible in Stata too! (key commands: `regexm`, `regexs`, `subinstr`)

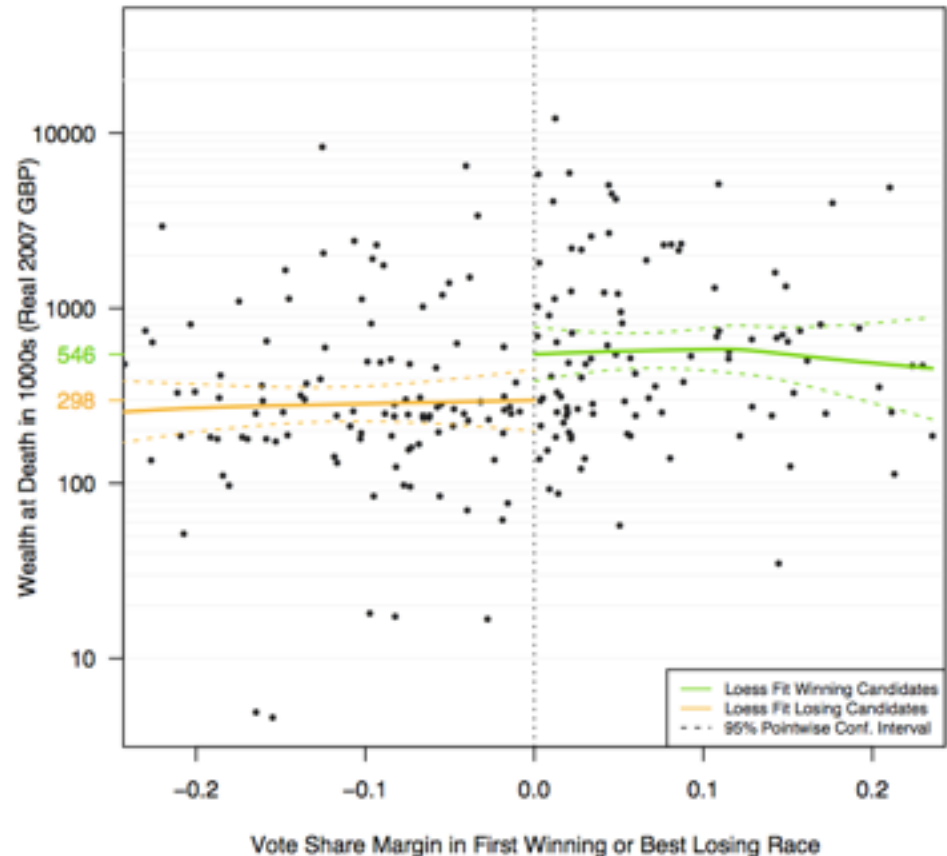
Use case: given a short biography for each observation, create dummy variable that is 1 if “barrister” “solicitor” “lawyer” or “attorney” appears in the text and 0 otherwise

Example: Eggers and Hainmueller (2009) “MPs for Sale?”

Research question: Was serving in U.K. House of Commons financially rewarding?

Research design: Compare wealth at death of narrowly successful and unsuccessful candidates from 1950-1970

Use of text: To make data collection cheaper and more reliable



Example: Eggers and Hainmueller (2009) “MPs for Sale?”

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
<hr/>	
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.	

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.



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.

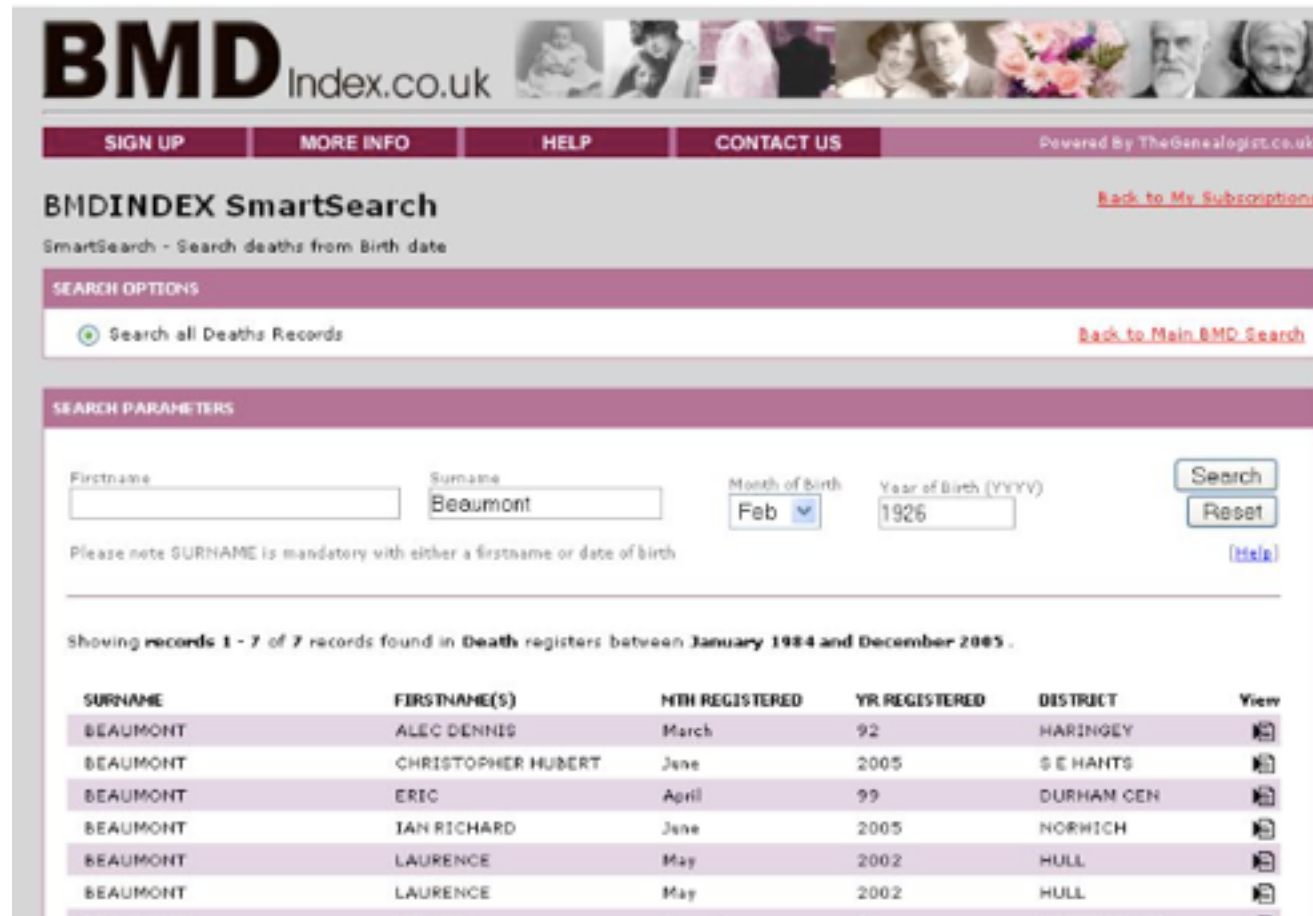
Mr. Dudley Smith, a journalist, is assistant news editor of a national Sunday newspaper. Has been crime reporter, sports writer, and special correspondent. Born 1926; educated at Chichester High School.

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

Web scraping example: Eggers and Hainmueller (2009) “MPs for Sale?”

Pseudocode: For each candidate,

- go to search form
- enter surname and date of birth
- click search button
- collect results
- identify matches based on first name/initials using **regular expressions**



The screenshot shows the BMD Index website interface. At the top, there is a navigation bar with links for SIGN UP, MORE INFO, HELP, and CONTACT US. Below this is the BMDINDEX SmartSearch section, which includes a search form and a results table. The search form has fields for Firstname, Surname (filled with 'Beaumont'), Month of Birth (set to 'Feb'), and Year of Birth (set to '1926'). There are Search and Reset buttons. Below the search form, a message states: 'Please note SURNAME is mandatory with either a first name or date of birth'. The results section shows 7 records found in Death registers between January 1984 and December 2005. The table below lists the search results.

SURNAME	FIRSTNAME(S)	MTH REGISTERED	YR REGISTERED	DISTRICT	View
BEAUMONT	ALEC DENNIS	March	92	HARINGEY	
BEAUMONT	CHRISTOPHER HUBERT	June	2005	S E HANTS	
BEAUMONT	ERIC	April	99	DURHAM CEN	
BEAUMONT	IAN RICHARD	June	2005	NORWICH	
BEAUMONT	LAURENCE	May	2002	HULL	
BEAUMONT	LAURENCE	May	2002	HULL	

Resources for learning these tools

- Google and the internet: endless tutorials, help pages, etc
- Standard texts for getting started in R, Ruby, Python etc
- Simon Jackman (2006), “Data from the web into R” [still good on basic process]
- Ingo Feinerer (2008), “An introduction to text mining in R”
- 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