# POIR 613: Computational Social Science

**Pablo Barberá**

University of Southern California
pablobarbera.com

Course website:
pablobarbera.com/POIR613/

# Quantitative Text Analysis

# Outline

- Motivation
- Workflow
- Overview of QTA methods
- Key concepts
- Selecting documents
- Selecting features
- Where to obtain textual data

# Text as data

# Text as data


Kukuriku
koalicija
Plan
21


the TRUE NORTH
STRONG AND FREE
Stephen Harper's plan for Canadians
Conservative


Et retfærdigt
Danmark


ΕΚΛΟΓΙΚΗ ΔΙΑΚΗΡΥΞΗ
ΣΥΡΙΖΑ
ισχυρός
στη Βουλή
και στους αγώνες
Είναι στο χέρι μας!
ΣΥΡΙΖΑ
ΣΥΝΑΣΠΙΣΜΟΣ ΡΙΖΟΣΠΑΣΤΙΚΗΣ ΑΡΙΣΤΕΡΑΣ


SPÖ
WAHLMANIFEST


GRÜNE NEUE
GESELLSCHAFTS-
VERTRAG
KLIMA • ARBEIT • GERECHTIGKEIT • FREIHEIT


WIR HABEN DIE KRAFT
GEMEINSAM
FÜR UNSER LAND.
REGIERUNGSPROGRAMM 2009 – 2013
CDU CSU

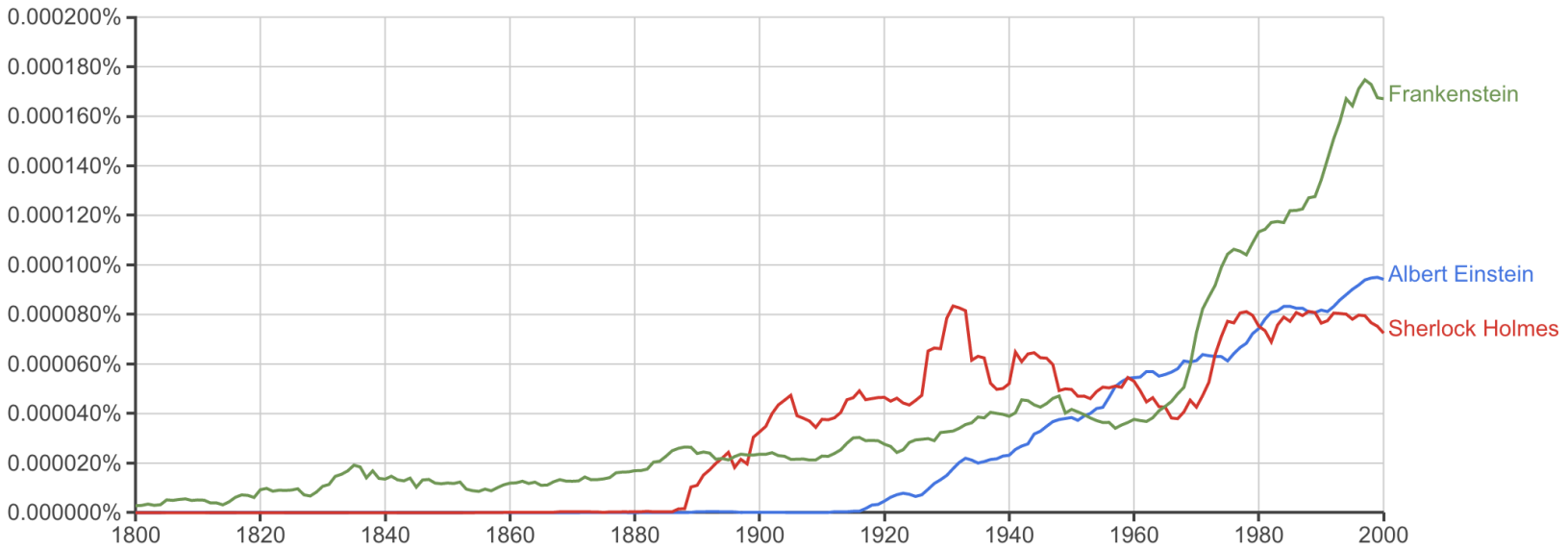

Sozial und
Demokratisch.
Anpacken. Für Deutschland.
Das Regierungsprogramm der SPD
SPD

# Text as data

Google Books Ngram Viewer

Graph these comma-separated phrases: Albert Einstein,Sherlock Holmes,Frankenstein case-insensitive

between 1800 and 2000 from the corpus English with smoothing of 3. Search lots of books



# Text as data


facebook
XING
Pinterest
Instagram
Gefällt mir
twitter
LinkedIn
flickr
skype
YouTube

# Why quantitative text analysis?

Justin Grimmer's haystack metaphor: QTA improves reading

- Analyzing a straw of hay: understanding the meaning of a sentence
  - Humans are great! But computer struggle
- Organizing the haystack: describing, classifying, scaling texts
  - Humans struggle. But computers are great!
  - (What this course is about)

Principles of quantitative text analysis (Grimmer & Stewart, 2013)

1. All quantitative models are wrong – but some are useful
2. Quantitative methods for text amplify resources and augment humans
3. There is no globally best method for automated text analysis
4. Validate, validate, validate

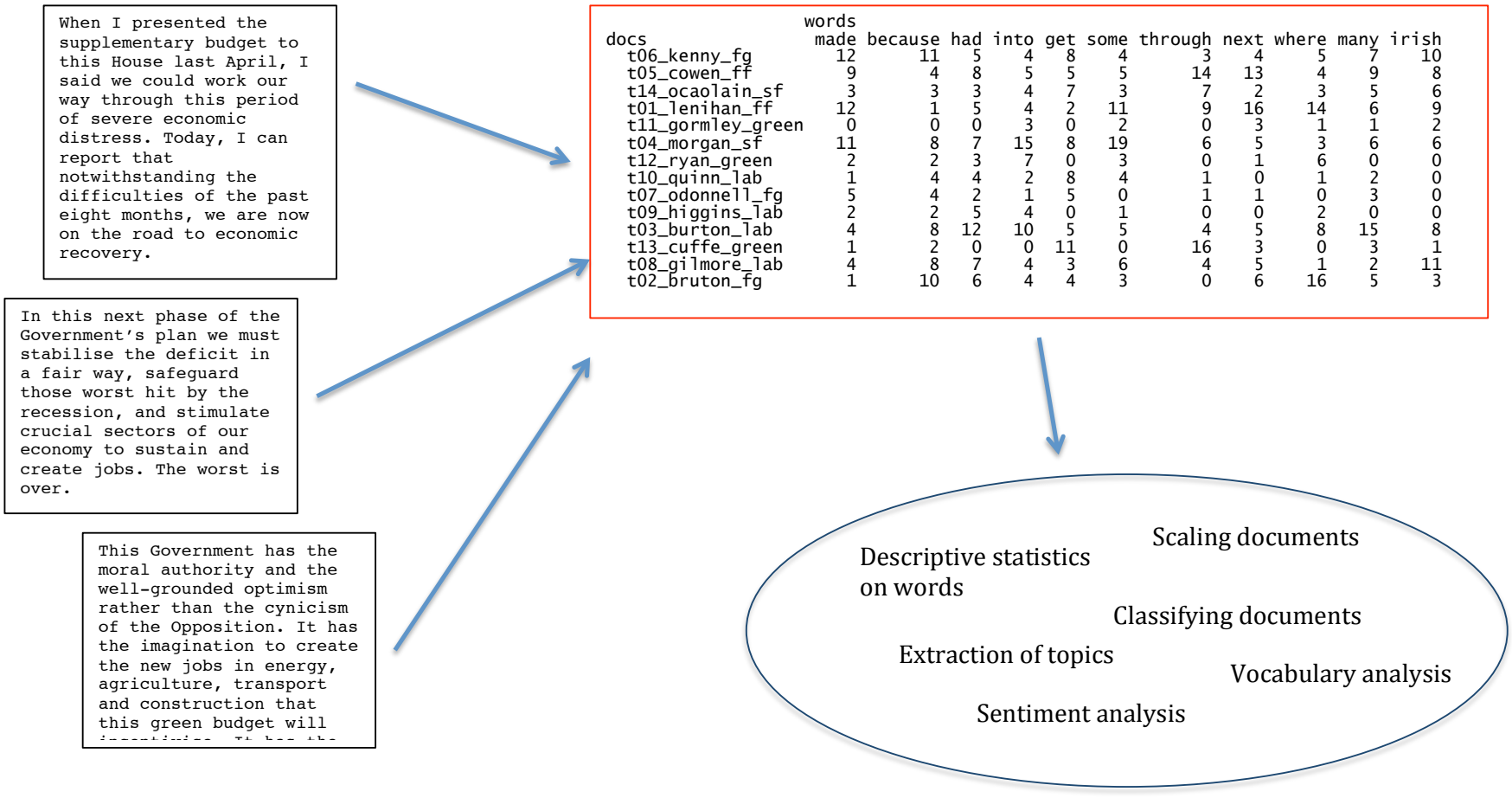
# Quantitative text analysis requires assumptions

1. Texts represent an observable implication of some underlying characteristic of interest
   - An attribute of the author
   - A sentiment or emotion
   - Salience of a political issue
2. Texts can be represented through extracting their *features*
   - most common is the bag of words assumption
   - many other possible definitions of "features" (e.g. word embeddings)
3. A document-feature matrix can be analyzed using quantitative methods to produce meaningful and valid estimates of the underlying characteristic of interest

# Outline

- Motivation
- Workflow
- Overview of QTA methods
- Key concepts
- Selecting documents
- Selecting features
- Where to obtain textual data

# Basic QTA Process: Texts → Feature matrix → Analysis



# Key feature of quantitative text analysis

1. Selecting texts: Defining the *corpus*
2. Conversion of texts into a common electronic format
3. Defining documents: deciding what will be the documentary unit of analysis

# Key feature of quantitative text analysis (cont.)

4. Defining features. These can take a variety of forms, including tokens, equivalence classes of tokens (dictionaries), selected phrases, human-coded segments (of possibly variable length), linguistic features, and more.
5. Conversion of textual features into a quantitative matrix
6. A quantitative or statistical procedure to extract information from the quantitative matrix
7. Summary and interpretation of the quantitative results

# Outline

- Motivation
- Workflow
- Overview of QTA methods
- Key concepts
- Selecting documents
- Selecting features
- Where to obtain textual data

# Overview of QTA (Grimmer and Stewart, 2013)

1. Acquire textual data:
   - Existing corpora; scraped data; digitized text
2. Preprocess the data:
   - Bag-of-words vs word embeddings
3. Apply method appropriate to research goal:
   - Describe and compare documents
     - Readability; similarity; keyness metrics
   - Classify documents into known categories
     - Dictionary methods
     - Supervised machine learning
   - Classify documents into unknown categories
     - Document clustering
     - Topic models
   - Scale documents on latent dimension
     - Known dimension: wordscores
     - Unknown dimensions: wordfish

# Descriptive text analysis

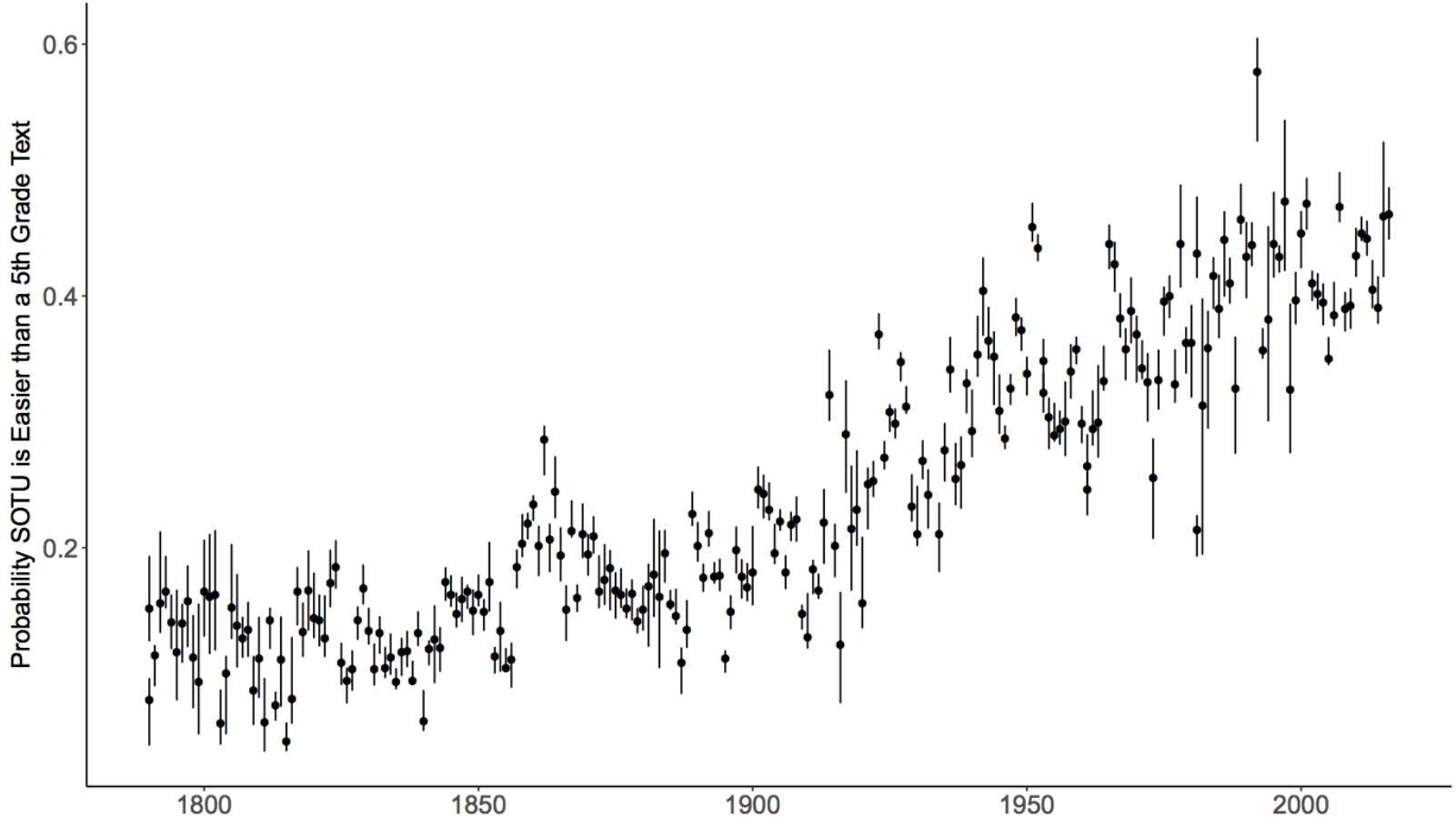
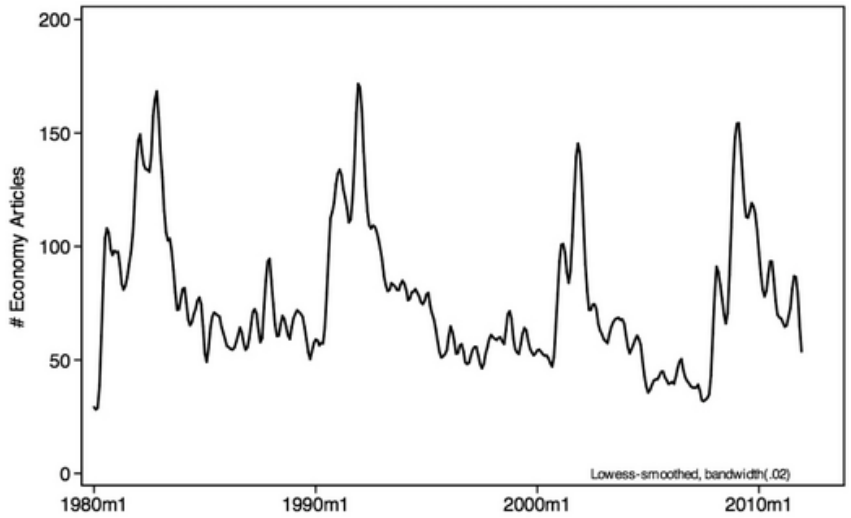
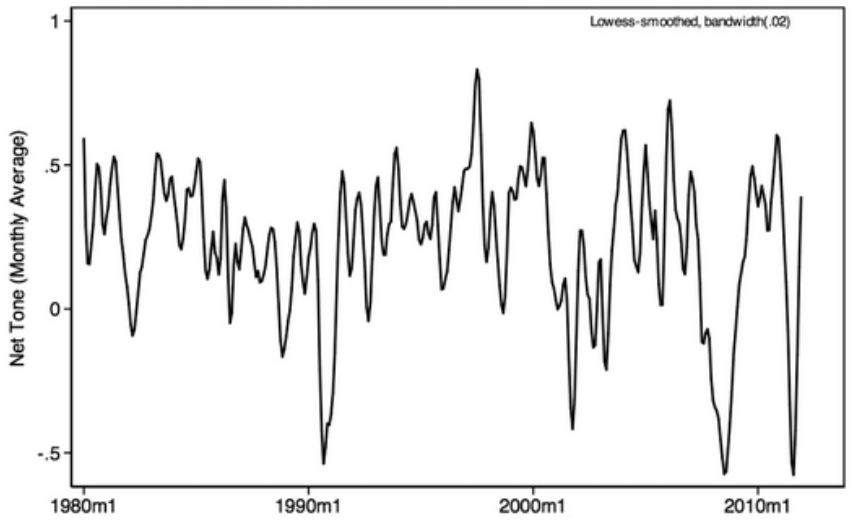


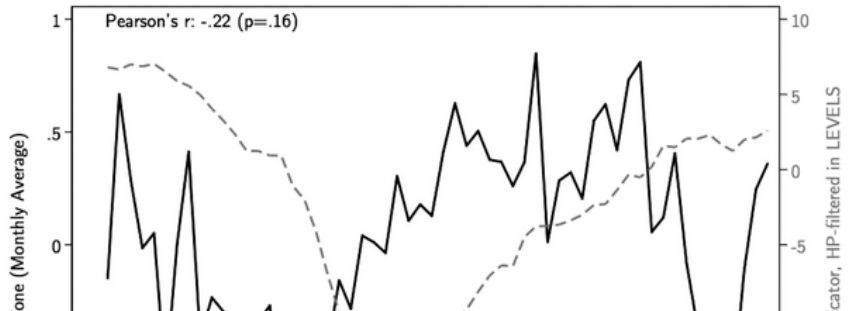

Figure 2: The probability that a State of the Union address is easier to understand than a fifth grade text baseline.

Benoit, Munger & Spirling (2017)

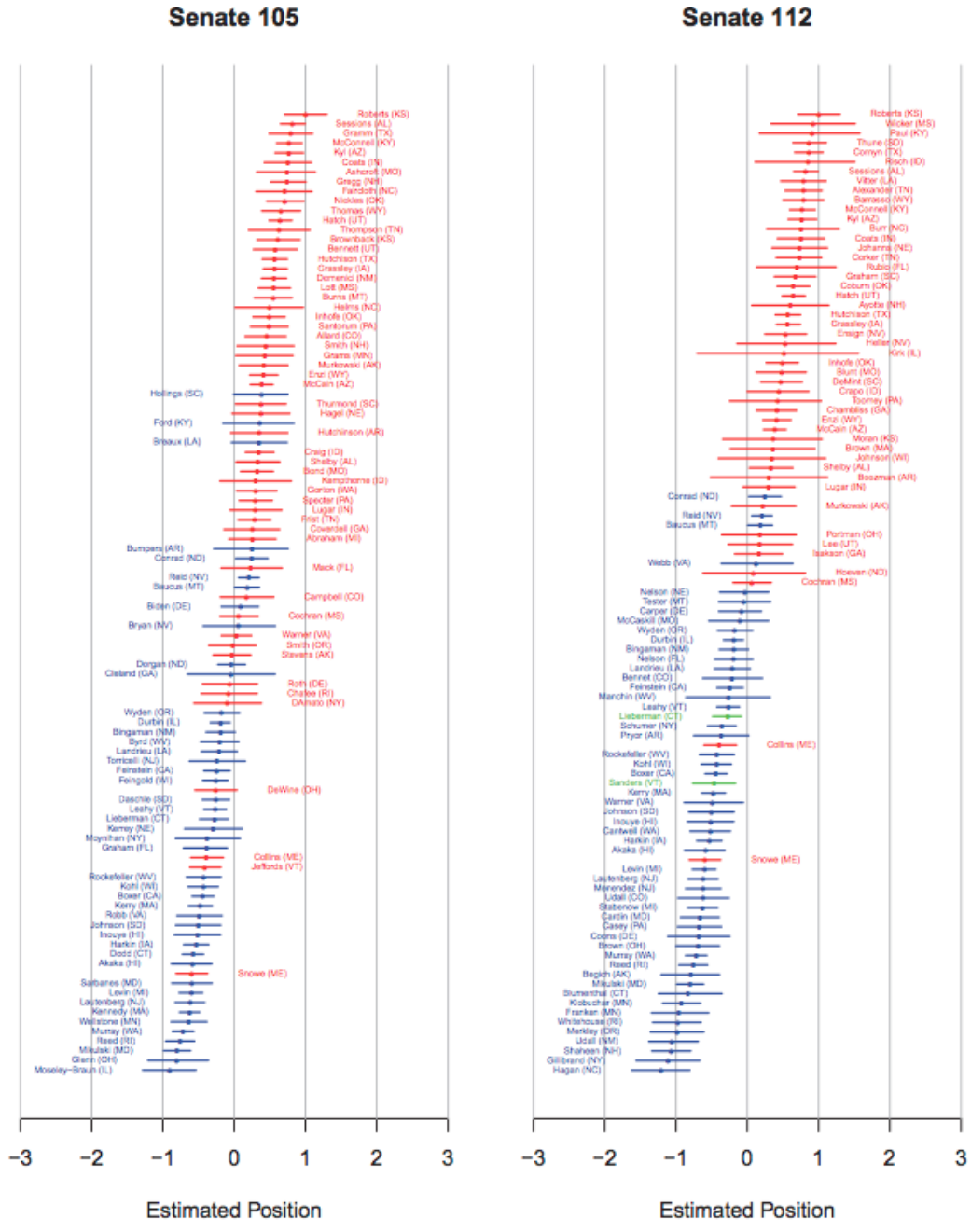
# Document classification into known categories



# Ideological scaling (Lauderdale & Herzog, PA 2016)



# Document classification into unknown categories

Bauer, Barberá *et al*, *Political Behavior*, 2016.

- Data: General Social Survey (2008) in Germany
- Responses to questions: *Would you please tell me what you associate with the term "left"? and would you please tell me what you associate with the term "right"?*
- Open-ended questions minimize priming and potential interviewer effects
- Automated text analysis to discover unknown categories and classify responses

# Document classification into unknown categories

Table 1: Top scoring words associated with each topic, and English translations)

| |
|---|
| Left topic 1: **Parties** (proportion = .26, average lr-scale value = 5.38)<br>linke, spd, partei, linken, pds, politik, kommunisten, parteien, grünen, punks<br>*the left, spd, party, the left, pds, politics, communists, parties, greens, punks* |
| Left topic 2: **Ideologies** (proportion = .26, average lr-scale value = 5.36)<br>kommunismus, links, sozialismus, lafontaine, rechts, aber, gysi, linkspartei, richtung, gleichmacherei<br>*communism, left, socialism, lafontaine, right, but, gysi, left party, direction, levelling* |
| Left topic 3: **Values** (proportion = .24, average lr-scale value = 4.06)<br>soziale, gerechtigkeit, demokratie, soziales, bürger, gleichheit, gleiche, freiheit, rechte, gleichberechtigung<br>*social, justice, democracy, social, citizen, equality, equal, freedom, rights, equal rights* |
| Left topic 4: **Policies** (proportion = .24, average lr-scale value =4.89)<br>sozial, menschen, leute, ddr, verbinde, kleinen, einstellung, umverteilung, sozialen, vertreten<br>*social, humans, people, ddr, associate, the little, attitude, redistribution, social, represent* |
| Right topic 1: **Ideologies** (proportion = .27, average lr-scale value = 5.00)<br>konservativ, nationalsozialismus, rechtsradikal, radikal, ordnung, politik, nazi, recht, menschen, konservative<br>*conservative, national socialism, right-wing radicalism, radical, order, politics, nazi, right, people, conservatives* |
| Right topic 2: **Parties** (proportion = .25, average lr-scale value = 5.26)<br>npd, rechts, cdu, csu, rechten, parteien, leute, aber, verbinde, rechtsradikalen<br>*npd, right, cdu, csu, the right, parties, people, but, associate, right-wing radicalists* |
| Right topic 3: **Xenophobia** (proportion = .25, average lr-scale value = 4.55)<br>ausländerfeindlichkeit, gewalt, ausländer, demokratie, nationalismus, rechtsradikalismus, diktatur, national, intoleranz, faschismus<br>*xenophobia, violence, foreigners, democracy, nationalism, right-wing radicalism, dictatorship, national, intolerance, fascism* |
| Right topic 4: **Right-wing extremists** (proportion = .23, average lr-scale value = 4.90)<br>nazis, neonazis, rechtsradikale, rechte, radikale, radikalismus, partei, ausländerfeindlich, reich, nationale<br>*nazis, neonazis, right-wing radicalists, rightists, radicals, radicalism, party, xenophobia, rich, national* |

**Note:** "proportion" indicates the average estimated probability that any given response is assigned to a topic. "average lr-scale value" is the mean position on the left-right scale (from 0 to 10) of individuals whose highest probability belongs to that particular topic.

Bauer, Barberá *et al*, *Political Behavior*, 2016.

# Document classification into unknown categories

Fig. 6: Left-right scale means for different subsamples of associations with **left** (dashed = sample mean, bars = 95% Cis)

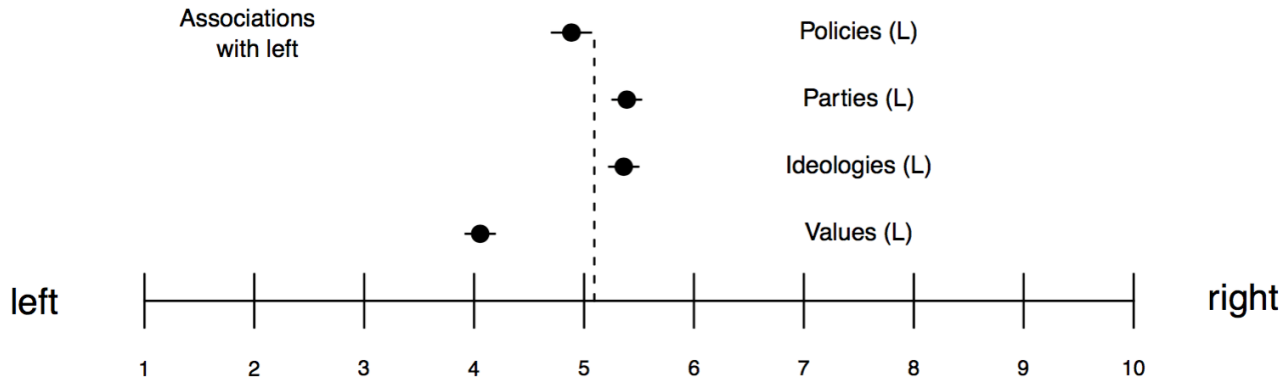
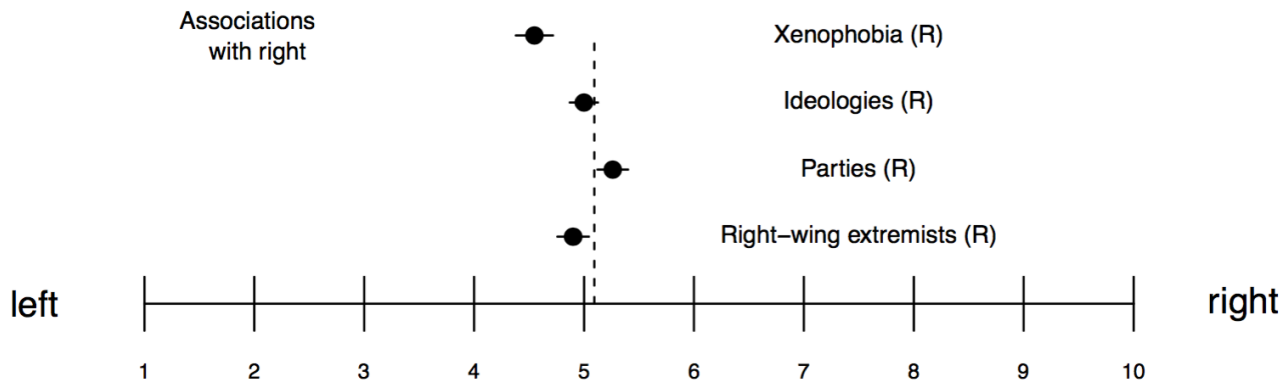


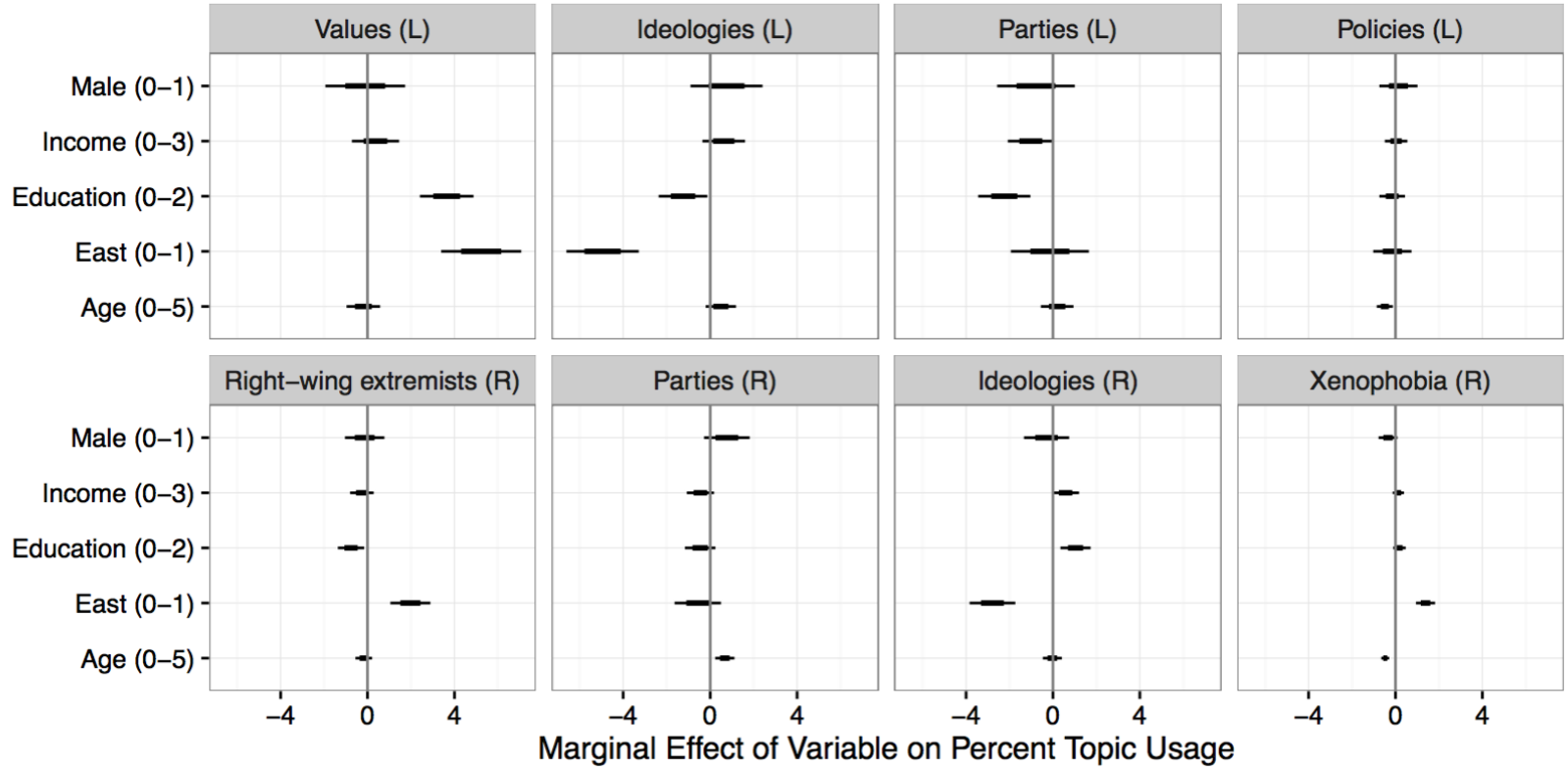

Fig. 7: Left-right scale means for different subsamples of associations with **right** (dashed = sample mean, bars = 95% Cis)




Bauer, Barberá *et al*, *Political Behavior*, 2016.

# Document classification into unknown categories

Fig. 9: Systematic relationship between associations with "left" and "right" and characteristics of respondents




**Note:** Each line indicates a 95% confidence interval (and 66% confidence interval in darker color) for the coefficient of eight different regressions of topic usage (in a scale from 0 to 100) at the respondent level on seven individual-level characteristics. The line on the bottom right corner (second row, second plot), for example, shows that individual a one-category change in age is associated with around one percentage point increase in the probability that the individual associated "right" with political parties.

Bauer, Barberá *et al*, *Political Behavior*, 2016.

# Document classification into unknown categories

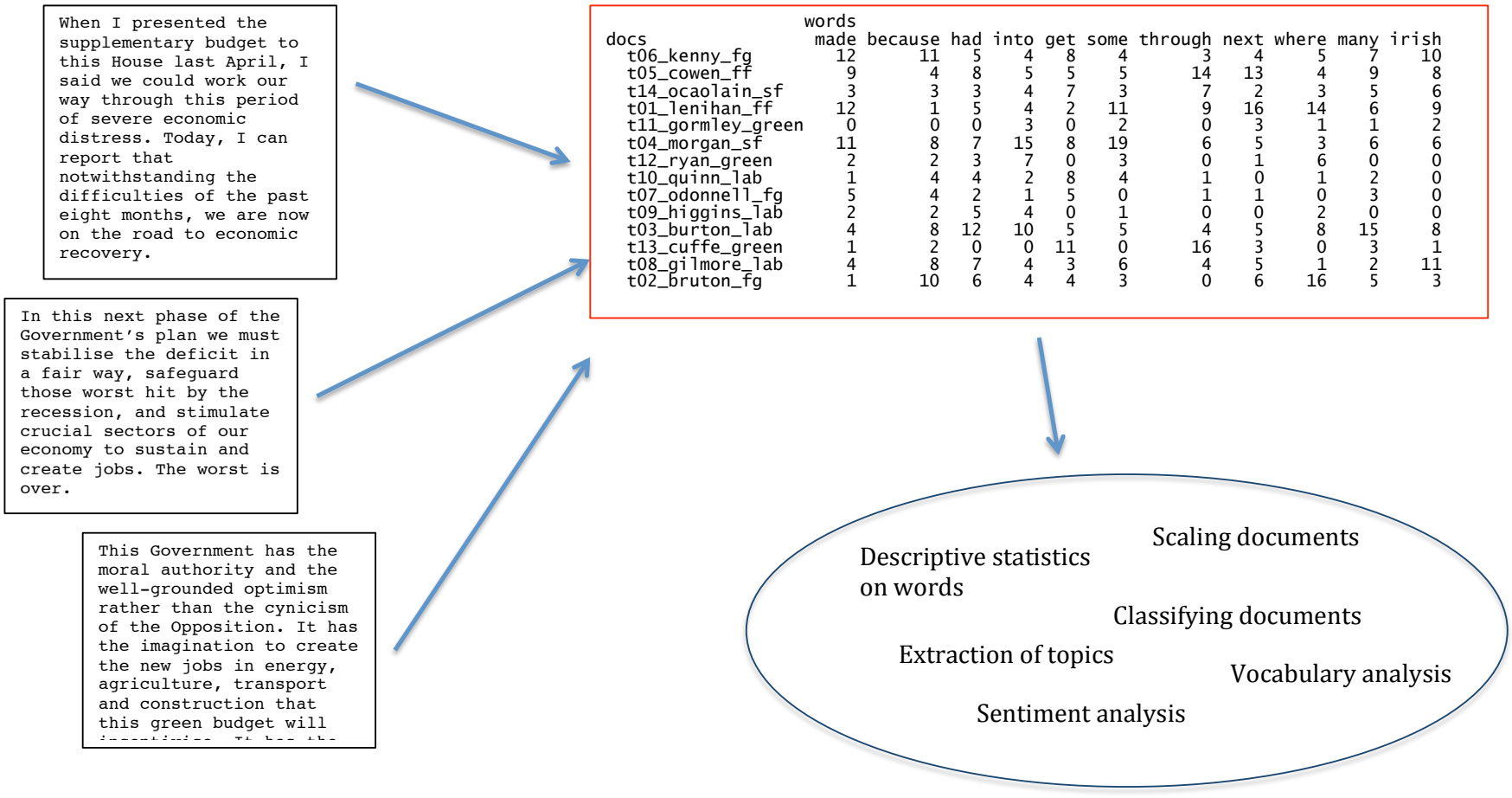
**What political issues do U.S. legislators emphasize on Twitter?** (Barberá et al, APSR 2020)

- Data: 651,116 tweets sent by US legislators from January 2013 to December 2014.
- Unit of analysis: tweets aggregated by day, party, and chamber
- 2,920 documents = 730 days $\times$ 2 chambers $\times$ 2 parties
- Automated text analysis to discover unknown categories and classify responses
- Validation: http://j.mp/lda-congress-demo

# Outline

- Motivation
- Workflow
- Overview of QTA methods
- Key concepts
- Selecting documents
- Selecting features
- Where to obtain textual data

# Basic QTA Process: Texts → Feature matrix → Analysis



# Some key basic concepts

**(text) corpus** a large and structured set of texts for analysis

**document** each of the units of the corpus

**types** for our purposes, a unique word

**tokens** any word – so token count is total words

**e.g.**

```
A corpus is a set of documents.
This is the second document in the corpus.
```

is a corpus with 2 documents, where each document is a sentence. The first document has 6 types and 7 tokens. The second has 7 types and 8 tokens. (We ignore punctuation for now.)

# Some more key basic concepts

**stems** words with suffixes removed (using set of rules)

**lemmas** canonical word form (the base form of a word that has the same meaning even when different suffixes or prefixes are attached)

| **word** | win | winning | wins | won | winner |
|---|---|---|---|---|---|
| **stem** | win | win | win | won | winner |
| **lemma** | win | win | win | win | win |

**keys** such as dictionary entries, where the user defines a set of equivalence classes that group different word types

**"key" words** Words selected because of special attributes, meanings, or rates of occurrence

**stop words** Words that are designated for exclusion from any analysis of a text

# Outline

- Motivation
- Workflow
- Overview of QTA methods
- Key concepts
- Selecting documents
- Selecting features
- Where to obtain textual data

# Strategies for selecting units of textual analysis

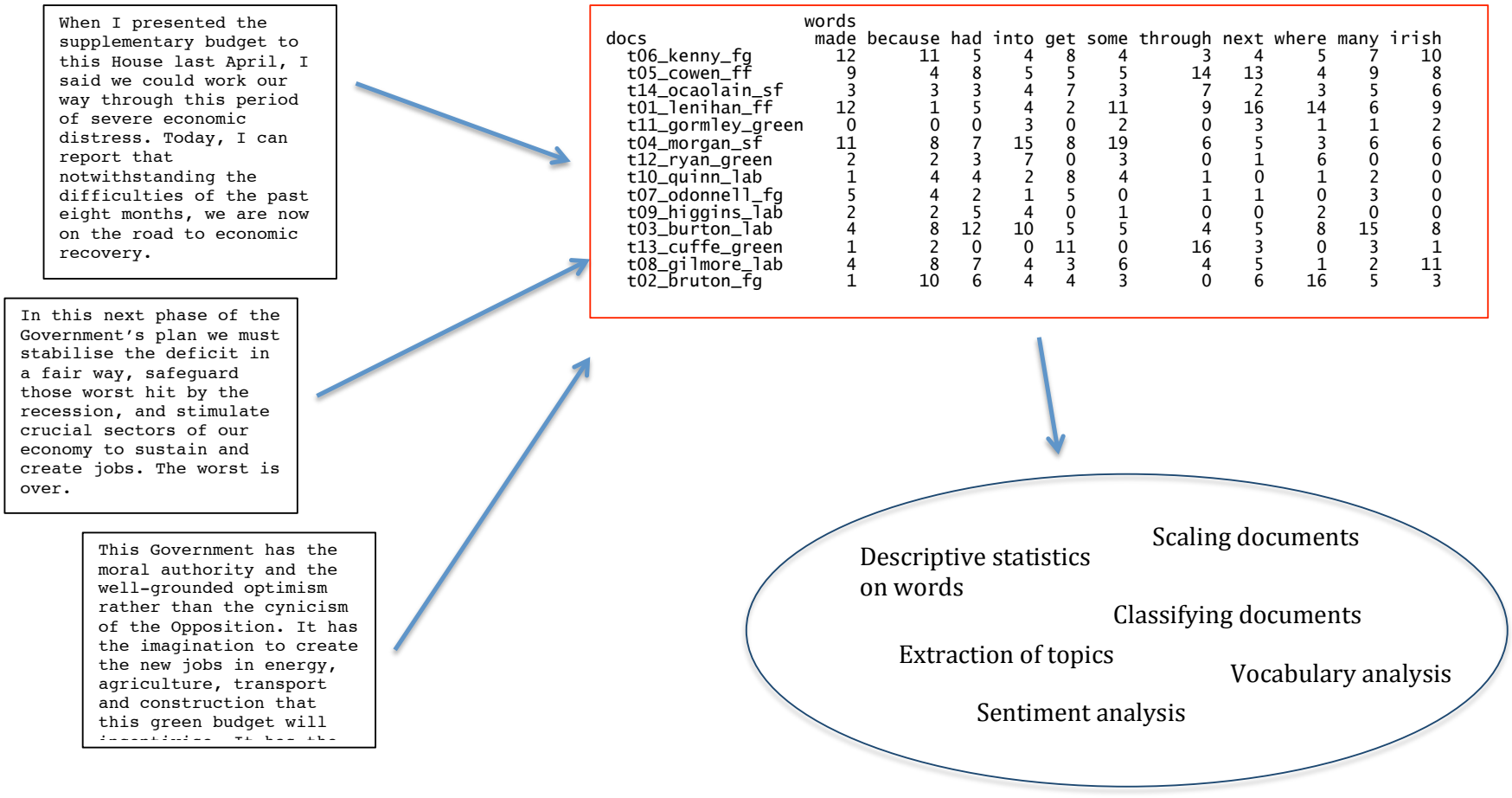
What can the document be?

- Words
- *n*-word sequences
- Sentences
- Pages
- Paragraphs
- Natural units (a speech, a poem, a manifesto)
- Aggregation of units (e.g. all speeches by party and year)
- Key: depends on the research design
- Frequent trade-off between cost and accuracy

# Outline

- Motivation
- Workflow
- Overview of QTA methods
- Key concepts
- Selecting documents
- Selecting features
- Where to obtain textual data

# Basic QTA adopts a bag-of-words approach




| docs | words made | because | had | into | get | some | through | next | where | many | irish |
|---|---|---|---|---|---|---|---|---|---|---|---|
| t06_kenny_fg | 12 | 11 | 5 | 4 | 8 | 4 | 3 | 4 | 5 | 7 | 10 |
| t05_cowen_ff | 9 | 4 | 8 | 5 | 5 | 5 | 14 | 13 | 4 | 9 | 8 |
| t14_ocaolain_sf | 3 | 3 | 3 | 4 | 7 | 3 | 7 | 2 | 3 | 5 | 6 |
| t01_lenihan_ff | 12 | 1 | 5 | 4 | 2 | 11 | 9 | 16 | 14 | 6 | 9 |
| t11_gormley_green | 0 | 0 | 0 | 3 | 0 | 2 | 0 | 3 | 1 | 1 | 2 |
| t04_morgan_sf | 11 | 8 | 7 | 15 | 8 | 19 | 6 | 5 | 3 | 6 | 6 |
| t12_ryan_green | 2 | 2 | 3 | 7 | 0 | 3 | 0 | 1 | 6 | 0 | 0 |
| t10_quinn_lab | 1 | 4 | 4 | 2 | 8 | 4 | 1 | 0 | 1 | 2 | 0 |
| t07_odonnell_fg | 5 | 4 | 2 | 1 | 5 | 0 | 1 | 1 | 0 | 3 | 0 |
| t09_higgins_lab | 2 | 2 | 5 | 4 | 0 | 1 | 0 | 0 | 2 | 0 | 0 |
| t03_burton_lab | 4 | 8 | 12 | 10 | 5 | 5 | 4 | 5 | 8 | 15 | 8 |
| t13_cuffe_green | 1 | 2 | 0 | 0 | 11 | 0 | 16 | 3 | 0 | 3 | 1 |
| t08_gilmore_lab | 4 | 8 | 7 | 4 | 3 | 6 | 4 | 5 | 1 | 2 | 11 |
| t02_bruton_fg | 1 | 10 | 6 | 4 | 4 | 3 | 0 | 6 | 16 | 5 | 3 |

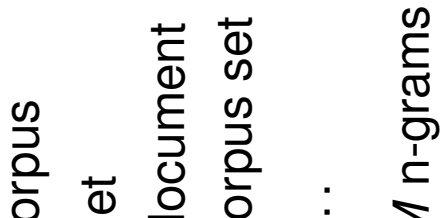
# Bag-of-words approach

From words to numbers:

1. Preprocess text: lowercase, remove stopwords and punctuation, stem, tokenize into unigrams and bigrams (bag-of-words assumption)

   "A corpus is a set of documents."
   "This is the second document in the corpus." "a corpus is a set of documents."
   "this is the second document in the corpus." "~~a~~ corpus ~~is~~ ~~a~~ set ~~of~~ documents~~.~~"
   "~~this~~ ~~is~~ ~~the~~ second document ~~in~~ ~~the~~ corpus~~.~~" "corpus set document~~s~~"
   "second document corpus" [corpus, set, document, corpus set, set document]
   [second, document, corpus, second document, document corpus]

2. Document-feature matrix:
   - **W**: matrix of $N$ documents by $M$ unique n-grams
   - $w_{im}$= number of times $m$-th n-gram appears in $i$-th document.



# Word frequencies and their properties

Bag-of-words approach disregards grammar and word order and uses word frequencies as features. Why?

- *Context is often uninformative*, conditional on presence of words:
  - Individual word usage tends to be associated with a particular degree of affect, position, etc. without regard to context of word usage
- Single words tend to be the most informative, as co-occurrences of multiple words (*n*-grams) are rare
- Some approaches focus on occurrence of a word as a binary variable, irrespective of frequency: a binary outcome
- Other approaches use frequencies: Poisson, multinomial, and related distributions

# Defining Features

- characters
- words
- word stems or lemmas: this is a form of defining *equivalence classes* for word features
- word segments, especially for languages using compound words, such as German, e.g. *Rindfleischetikettierungsüberwachungsaufgabenübertragungsge*
  (the law concerning the delegation of duties for the supervision of cattle marking and the labelling of beef)

# Defining Features (cont.)

- “word” sequences, especially when inter-word delimiters (usually white space) are not commonly used, as in Chinese
- linguistic features, such as parts of speech
- (if qualitative coding is used) coded or annotated text segments
- word embeddings (more on this later in the course)

# Parts of speech

- the Penn "Treebank" is the standard scheme for tagging POS

| Number | Tag | Description |
|---|---|---|
| 1. | CC | Coordinating conjunction |
| 2. | CD | Cardinal number |
| 3. | DT | Determiner |
| 4. | EX | Existential *there* |
| 5. | FW | Foreign word |
| 6. | IN | Preposition or subordinating conjunction |
| 7. | JJ | Adjective |
| 8. | JJR | Adjective, comparative |
| 9. | JJS | Adjective, superlative |
| 10. | LS | List item marker |
| 11. | MD | Modal |
| 12. | NN | Noun, singular or mass |
| 13. | NNS | Noun, plural |
| 14. | NNP | Proper noun, singular |
| 15. | NNPS | Proper noun, plural |
| 16. | PDT | Predeterminer |
| 17. | POS | Possessive ending |
| 18. | PRP | Personal pronoun |
| 19. | PRP$ | Possessive pronoun |
| 20. | RB | Adverb |
| 21. | RBR | Adverb, comparative |
| 22. | RBS | Adverb, superlative |
| 23. | RP | Particle |

# Parts of speech (cont.)

- several open-source projects make it possible to tag POS in text, such as Apache's OpenNLP (and R package `openNLP` wrapper) or TreeTagger

```
> s
Pierre Vinken, 61 years old, will join the board as a nonexecutive dir
Nov. 29. Mr. Vinken is chairman of Elsevier N.V., the Dutch publishing
group.
> sprintf("%s/%s", s[a3w], tags)
 [1] "Pierre/NNP"      "Vinken/NNP"      ",/,"             "61/CD"
 [5] "years/NNS"       "old/JJ"          ",/,"             "will/MD"
 [9] "join/VB"         "the/DT"          "board/NN"        "as/IN"
[13] "a/DT"            "nonexecutive/JJ" "director/NN"     "Nov./NNP"
[17] "29/CD"           "./."             "Mr./NNP"         "Vinken/NNP
[21] "is/VBZ"          "chairman/NN"     "of/IN"           "Elsevier/N
[25] "N.V./NNP"        ",/,"             "the/DT"          "Dutch/JJ"
[29] "publishing/NN"   "group/NN"        "./."
```

# Parts of speech (cont.)

Example: Creating an index of editorialization of journalists' and media outlets' political news coverage.

Proportion of tweets that: (1) mention a major party or candidate, (2) include at least one adjective.

**Table 2.4** Determinants of editorialisation and popularity of news accounts on twitter (OLS regressions)

| | DV = Editorialisation | | DV = Popularity | |
|---|---|---|---|---|
| | Model 1 | Model 2 | Model 3 | Model 4 |
| Type: journalist | 5.10*** | 4.32*** | 2.70*** | 2.49*** |
| | (1.13) | (1.26) | (0.22) | (0.30) |
| Tweets about Europe (%) | –0.03+ | –0.03+ | 0.01*** | 0.01*** |
| | (0.02) | (0.02) | (0.002) | (0.002) |
| Editorialisation Index | | | 0.02*** | 0.02*** |
| | | | (0.004) | (0.004) |
| (Intercept) | 7.58** | 7.94** | –4.03*** | –3.92*** |
| | (2.59) | (2.47) | (0.40) | (0.41) |
| Country fixed effects | YES | YES | YES | YES |
| Outlet fixed effects | YES | YES | YES | YES |
| $R^2$ | 0.12 | 0.12 | 0.71 | 0.71 |
| Adj. $R^2$ | 0.08 | 0.08 | 0.70 | 0.70 |
| Num. obs. | 2662 | 2662 | 2662 | 2662 |
| RMSE | 7.63 | 7.63 | 1.08 | 1.08 |

Barberá, Vaccari, Valeriani (2016) [control variables ommitted]

# Strategies for feature selection

How to choose which features to include?

- All? Computationally inefficient, and rare words are generally uninformative

Potential criteria to select features ("trim" the DFM):

- document frequency: How many documents in which a term appears
- term frequency How many times does the term appear in the corpus
- deliberate disregard Use of "stop words" – words excluded because they represent linguistic connectors of no substantive content
- purposive selection Use of a *dictionary* of words or phrases
- declared equivalency classes Non-exclusive synonyms, also known as *thesaurus* (more on this later)

# Common English stop words

```
a, able, about, across, after, all, almost, also, am, am
an, and, any, are, as, at, be, because, been, but, by, c
cannot, could, dear, did, do, does, either, else, ever,
every, for, from, get, got, had, has, have, he, her, her
him, his, how, however, I, if, in, into, is, it, its, ju
least, let, like, likely, may, me, might, most, must, my
neither, no, nor, not, of, off, often, on, only, or, oth
our, own, rather, said, say, says, she, should, since, s
some, than, that, the, their, them, then, there, these,
they, this, tis, to, too, twas, us, wants, was, we, were
what, when, where, which, while, who, whom, why, will, w
would, yet, you, your
```

- But no list should be considered universal

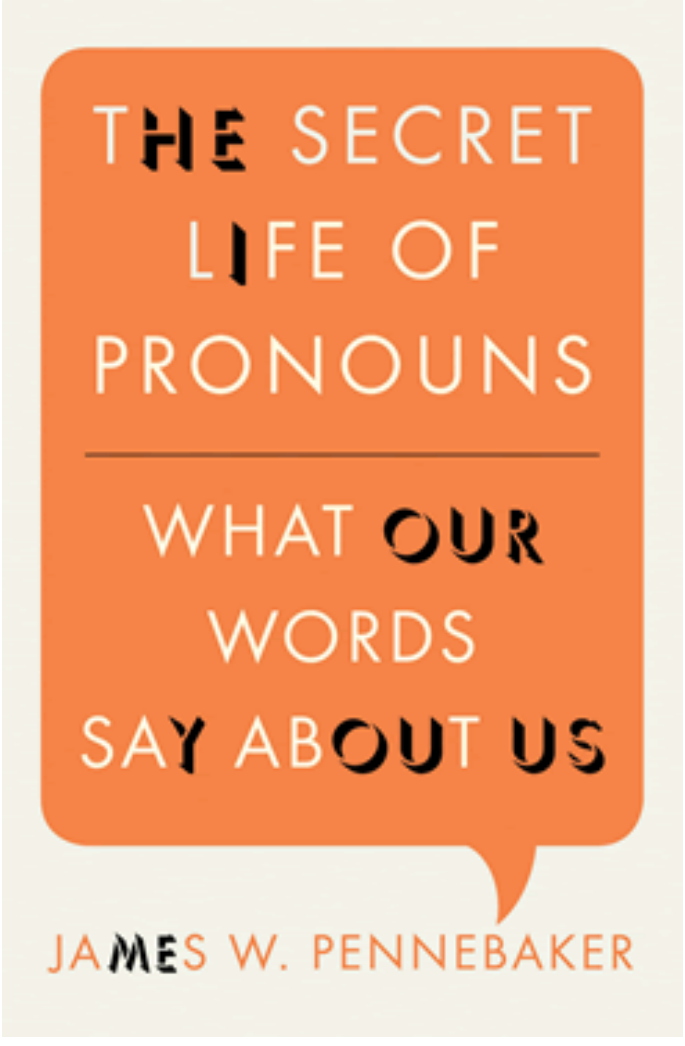
# A more comprehensive list of stop words

as, able, about, above, according, accordingly, across, actually, after, afterwards, again, against, ain't, all, allow, allows, almost, alone, along, already, also, although, always, am, among, amongst, an, and, another, any, anybody, anyhow, anyone, anything, anyway, anyways, anywhere, apart, appear, appreciate, appropriate, are, aren't, around, as, aside, ask, asking, associated, at, available, away, awfully, be, became, because, become, becomes, becoming, been, before, beforehand, behind, being, believe, below, beside, besides, best, better, between, beyond, both, brief, but, by, c'mon, c's, came, can, can't, cannot, cant, cause, causes, certain, certainly, changes, clearly, co, com, come, comes, concerning, consequently, consider, considering, contain, containing, contains, corresponding, could, couldn't, course, currently, definitely, described, despite, did, didn't, different, do, does, doesn't, doing, don't, done, down, downwards, during, each, edu, eg, eight, either, else, elsewhere, enough, entirely, especially, et, etc, even, ever, every, everybody, everyone, everything, everywhere, ex, exactly, example, except, far, few, fifth, first, five, followed, following, follows, for, former, formerly, forth, four, from, further, furthermore, get, gets, getting, given, gives, go, goes, going, gone, got, gotten, greetings, had, hadn't, happens, hardly, has, hasn't, have, haven't, having, he, he's, hello, help, hence, her, here, here's, hereafter, hereby, herein, hereupon, hers, herself, hi, him, himself, his, hither, hopefully, how, howbeit, however, i'd, i'll, i'm, i've, ie, if, ignored, immediate, in, inasmuch, inc, indeed, indicate, indicated, indicates, inner, insofar, instead, into, inward, is, isn't, it, it'd, it'll, it's, its, itself, just, keep, keeps, kept, know, knows, known, last, lately, later, latter, latterly, least, less, lest, let, let's, like, liked, likely, little, look, looking, looks, ltd, mainly, many, may, maybe, me, mean, meanwhile, merely, might, more, moreover, most, mostly, much, must, my, myself, name, namely, nd, near, nearly, necessary, need, needs, neither, never, nevertheless, new, next, nine, no, nobody, non, none, noone, nor, normally, not, nothing, novel, now, nowhere, obviously, of, off, often, oh, ok, okay, old, on, once, one, ones, only, onto, or, other, others, otherwise, ought, our, ours, ourselves, out, outside, over, overall, own, particular, particularly, per, perhaps, placed, please, plus, possible, presumably, probably, provides, que, quite, qv, rather, rd, re, really, reasonably, regarding, regardless, regards, relatively, respectively, right, said, same, saw, say, saying, says, second, secondly, see, seeing, seem, seemed, seeming, seems, seen, self, selves, sensible, sent, serious, seriously, seven, several, shall, she, should, shouldn't, since, six, so, some, somebody, somehow, someone, something, sometime, sometimes, somewhat, somewhere, soon, sorry, specified, specify, specifying, still, sub, such, sup, sure, t's, take, taken, tell, tends, th, than, thank, thanks, thanx, that, that's, thats, the, their, theirs, them, themselves, then, thence, there, there's, thereafter, thereby, therefore, therein, theres, thereupon, these, they, they'd, they'll, they're, they've, think, third, this, thorough, thoroughly, those, though, three, through, throughout, thru, thus, to, together, too, took, toward, towards, tried, tries, truly, try, trying, twice, two, un, under, unfortunately, unless, unlikely, until, unto, up, upon, us, use, used, useful, uses, using, usually, value, various, very, via, viz, vs, want, wants, was, wasn't, way, we, we'd, we'll, we're, we've, welcome, well, went, were, weren't, what, what's, whatever, when, whence, whenever, where, where's, whereafter, whereas, whereby, wherein, whereupon, wherever, whether, which, while, whither, who, who's, whoever, whole, whom, whose, why, will, willing, wish, with, within, without, won't, wonder, would, would, wouldn't, yes, yet, you, you'd, you'll, you're, you've, your, yours, yourself, yourselves, zero

# Stopwords

Are there cases in which we would want to keep stopwords? Or should we always exclude them from our analysis?

# Stopwords sometimes can be informative!




But sometimes we want to add/remove our own new stopwords (e.g. female pronouns, legislative terms, directional terms)

# Stemming words

**Lemmatization** refers to the algorithmic process of converting words to their lemma forms.

**stemming** the process for reducing inflected (or sometimes derived) words to their stem, base or root form. Different from *lemmatization* in that stemmers operate on single words without knowledge of the context.

**both** convert the morphological variants into stem or root terms

**example:** `produc` from `production, producer, produce, produces, produced`

**Why?** Reduce feature space by collapsing different words into a stem (e.g. “happier” and “happily” convey same meaning as “happy”)

# Outline

- Motivation
- Workflow
- Overview of QTA methods
- Key concepts
- Selecting documents
- Selecting features
- Where to obtain textual data

# Where to obtain textual data?

Some tips...

- Existing datasets, e.g.
  - UCD's EuroParl project
  - Hansard Archive of parliamentary debates in UK
  - Media archives (newspaper articles, TV transcripts...) at LexisNexis, ProQuest, Factiva...
  - Academic articles (JSTOR Data for Research)
  - Open-ended responses to survey questions
- Collect your own data:
  - From social media (Twitter, FB) and blogs
  - Scraping other websites
- Digitize your own text data using OCR (optical character recognition) software
  - Options: Tesseract (open-source), Abbyy FineReader

# Problems you are likely to encounter

- Problems with encoding
- File formats that cannot be read as plain text
- Extraneous junk (page footers, numbers, titles, etc)
- Misspelllings
- Different normalizations (e.g. for Japanese)