Quantitative Text Analysis

Bruno Castanho Silva

Day 1 - Introductions

Cologne Center for
Comparative Politics

Who am I?

- Post-doctoral researcher at the Chair of European Politics at the Cologne Center for Comparative Politics (CCCP)
- PhD in Political Science (Comparative Politics, 2017) at the Central European University (Budapest)
- Research on: populism, social media and politics; legislative politics
- Have been teaching various quantitative methods for a few years, including SEM, machine learning, causal inference

Who are you?

- What program are you in?
- Do you have a specific data/application of text analysis in mind?
- What's your experience with quantitative text analysis?
- What's your experience with R?

Organization

- Ask questions as they come up;
- Write me an email if you'd like a meeting to discuss your project;
- bcsilva@wiso.uni-koeln.de

Structure

- Lecture + lab
- For lab exercises, you'll be sent to breakout rooms;
- One or two breaks each session, depending on how it's going;

What this course is about

- Learn various state-of-the-art text-as-data approaches
- Learn how to analyze textual data and present results
- Discuss the promises, pitfalls, and limitations of quantitative text analysis in social sciences
- Help to identify the most appropriate methods/techniques for your data and question;
- Getting very annoyed at quanteda every once in a while

What this course is **not** about

- Programming. We're not software engineers developing new methods
- A replacement for qualitative and interpretivist approaches
- An exhaustive coverage of text-as-data approaches. There's much more out there, every day

Getting started

People and politicians talk a lot

We want to make sense of it

- There is a **message** or **content** that cannot be directly observed
    - Position on issues, topics discussed, tone, etc
- and **behaviour**, including **linguistic behaviour**, which can be directly observed
    - Expressed **words** and **sentences**
- But needs to be **interpreted**

We spend a lot of time learning it

And get pretty good…

- Takes us little effort to:
    - Spot the topic of a text
    - Get its content
    - Understand the meaning
- But in science we want something systematic, that can be replicated by others

Classical content analysis

- A research technique for making **replicable** and **valid** inferences from texts
- Apply explicit **coding rules** to **classify content** and **summarize the results** numerically.
- Examples:
  - Frequency analysis of topics in a newspaper
  - Determine the tone or position in speeches

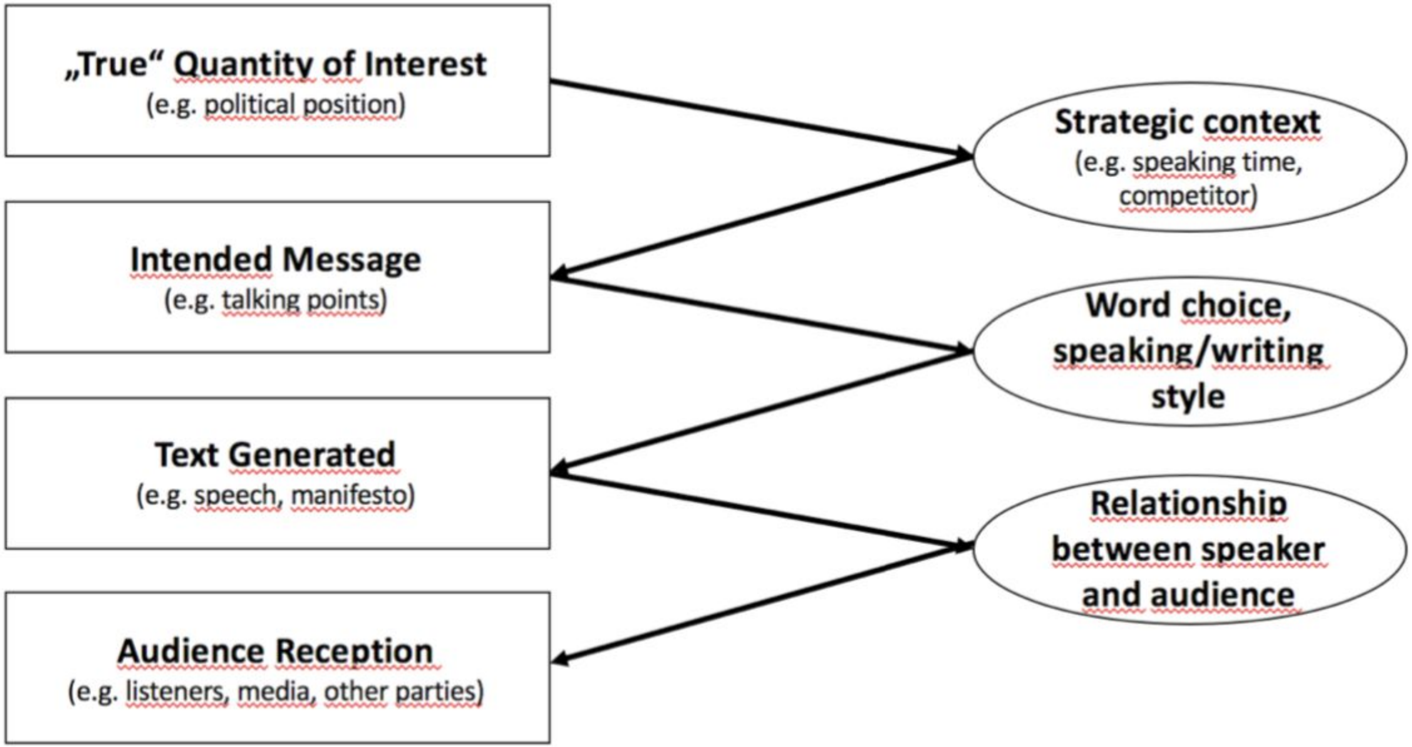
We eventually reach a limit

- There’s just so much text a human can read
- With the increased availability of text data, we need to scale up
- Enters **quantitative text analysis**

Qual v. Quant text analysis

- All reading of texts is **qualitative**
- Quantitative: using computers to assist us in making sense from a large body of text;
- Qualitative: close reading of a few texts by human coders;
- **Not mutually exclusive**: we often combine both approaches

What’s in a speech?



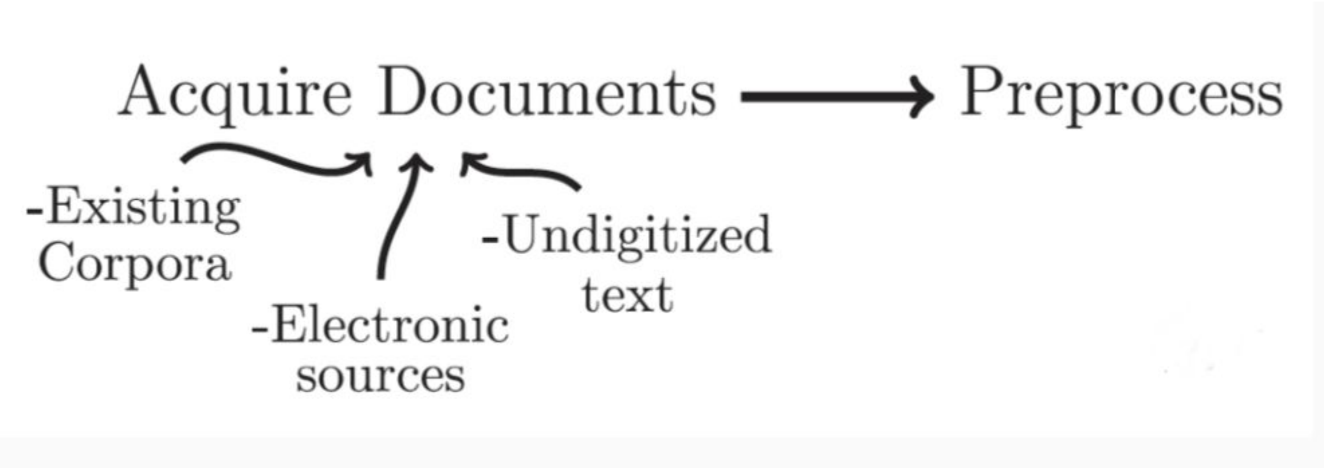
QTA is about **inference**

- Based on what is **expressed**, we want to **infer** a **true quantity**
- As with all inference, there's always **measurement error**
- Some concepts are easier to capture
    - e.g. topic; positive/negative tone
- Some are more difficult, even for humans
    - e.g. populism

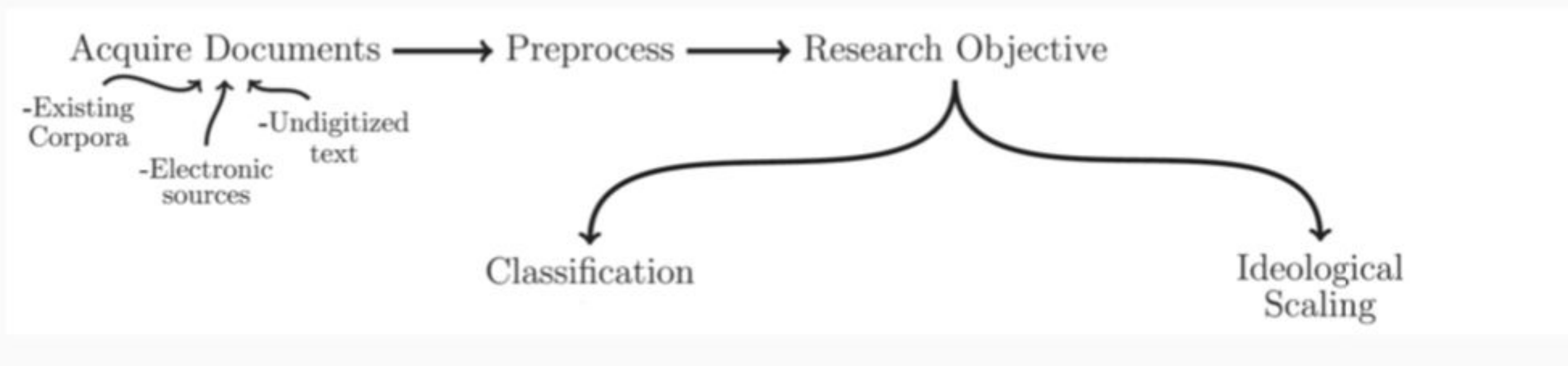
Limitations of text analysis

- We need textual record (potential **selection bias**)
  - e.g. Twitter is overrepresented in studies of social media and politics because it's the easiest to get data from
- Easier to establish **reliability** than **validity**
- Language is never 100% precise
  - No matter what Germans tell you about theirs…
- Recommendation: **validate, validate, validate**

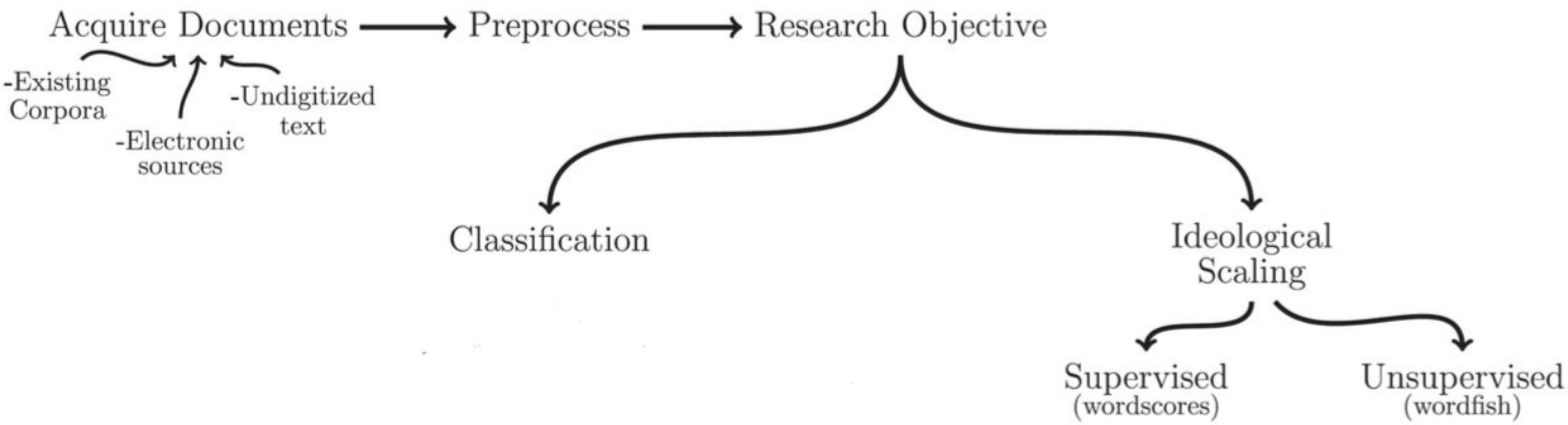
Text as data, an overview (Grimmer and Stewart, 2013)



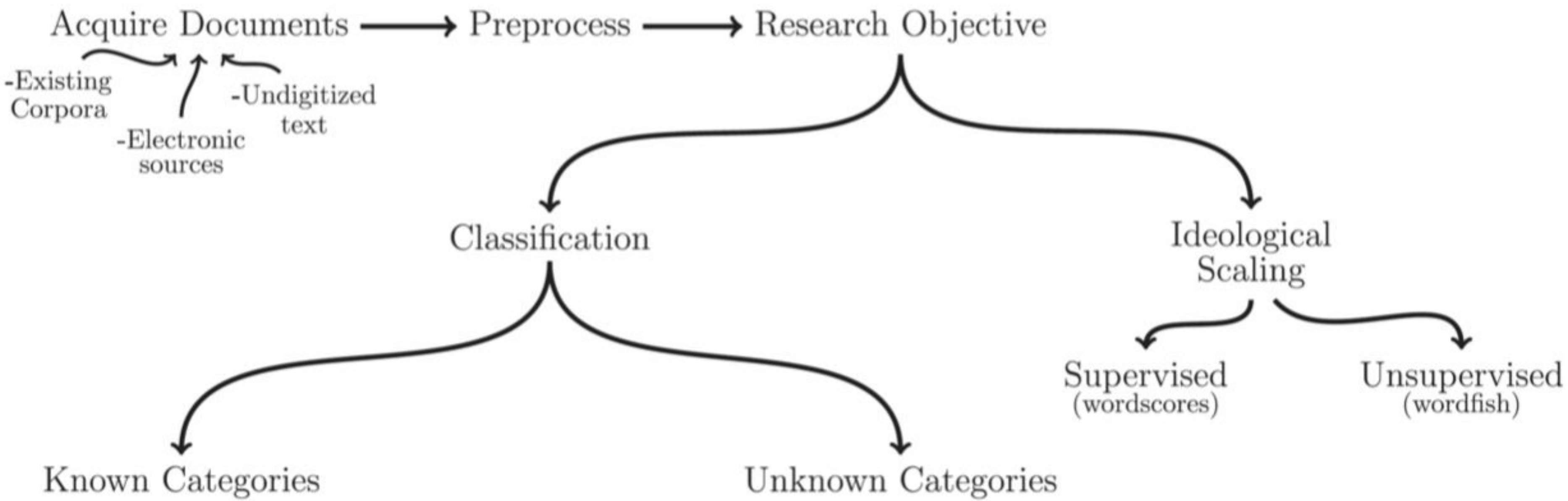
Text as data, an overview (Grimmer and Stewart, 2013)

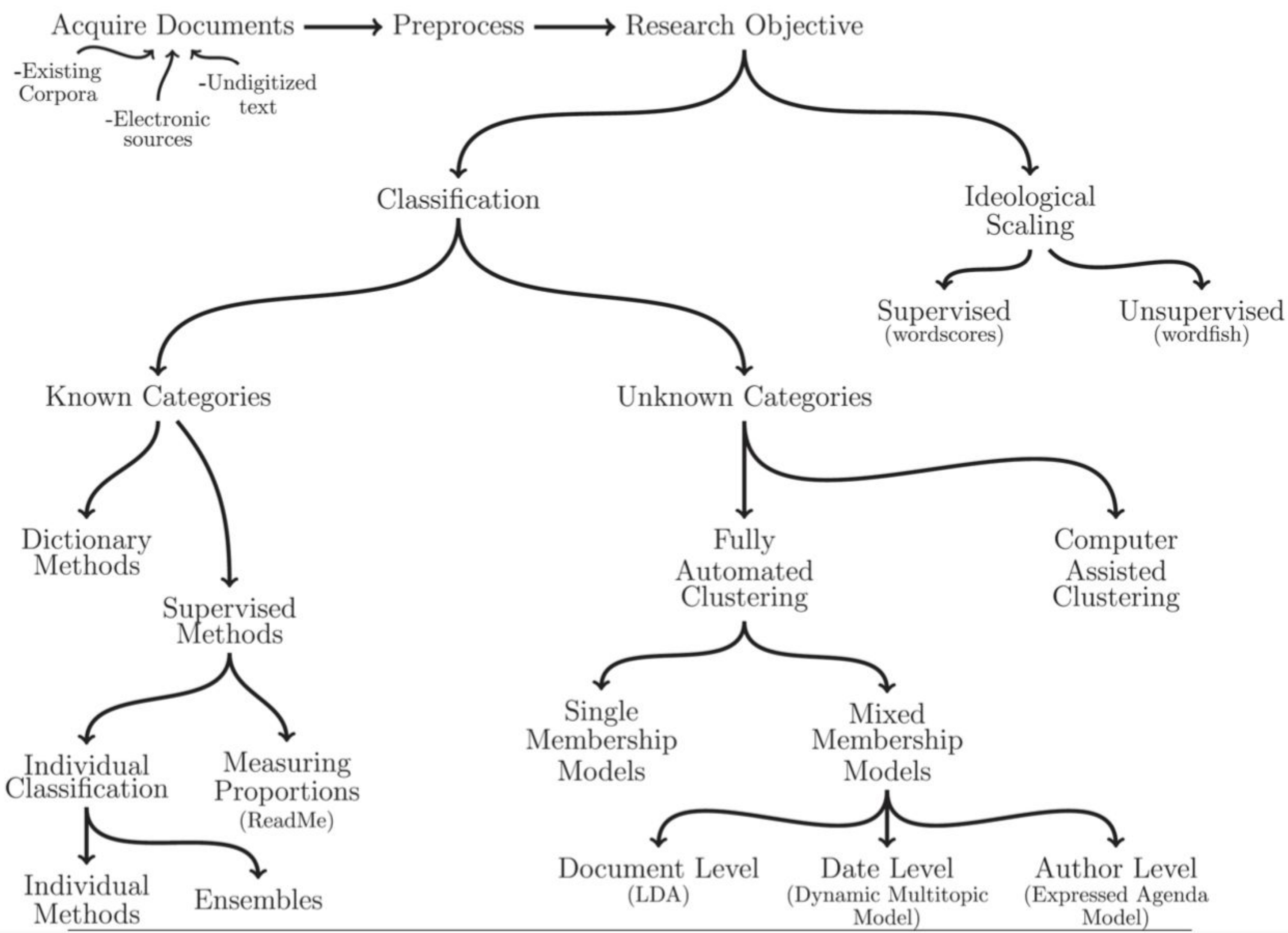


Text as data, an overview (Grimmer and Stewart, 2013)



Text as data, an overview (Grimmer and Stewart, 2013)





Acquire Documents
Preprocess
Research Objective
-Existing Corpora
-Electronic sources
-Undigitized text
Classification
Ideological Scaling
Supervised (wordscores)
Unsupervised (wordfish)
Known Categories
Unknown Categories
Dictionary Methods
Supervised Methods
Fully Automated Clustering
Computer Assisted Clustering
Individual Classification
Measuring Proportions (ReadMe)
Single Membership Models
Mixed Membership Models
Individual Methods
Ensembles
Document Level (LDA)
Date Level (Dynamic Multitopic Model)
Author Level (Expressed Agenda Model)

Text-as-data, important considerations

- Consider the **strategic data-generating process**
  - Actors, medium, target-audience, constraints, ...
- Find the relevant **unit of analysis** and fit the measurement accordingly
  - Individual texts/speeches, politicians, parties, etc
- Concepts are **latent;** the words are a visible manifestation


SO ARE THESE "CONCEPTS"
IN THE ROOM WITH US RIGHT NOW?

Assumptions

- Texts represent an **observable implication** of a characteristic of interest (e.g. a political position)
- Political speech is meaningful and **not just cheap talk**
- Texts can be represented by extracting **features** (most common is bag of words)
- The exact sentences said are (usually) not that interesting - what counts is the **latent dimension** we are interested in

QTA, step-by-step

1. **Define a research question**
2. Selecting texts and defining the corpus making sure there is no sampling bias
3. Conversion into electronic format (if needed)
4. Define documents
5. Define features
6. Convert features into matrices that can be quantitatively processed
7. Analysis and summary of results

Questions?

Getting started - Basic concepts

Important concepts

- **Corpus**: large and structured set of texts
- **tokens:** any word
- Preprocessing:
    - **stems:** words with suffixes removed (e.g. “walking” reduced to “walk”)
    - **lemmas:** canonical word form (e.g. “better” has “good” as its lemma)
    - **stop-words:** words that don’t convey meaning and are often removed from analysis (prepositions, articles,...)
- **bag-of-words**: assumption that word order doesn’t matter. Present in several QTA applications

Corpora

- A large and structured set of texts
- Associates the full text to metadata for each text
  - E.g. author, date, etc
- Can be created from a set of text files, from splitting a long text into segments, or any other way
- Basic data structure for QTA

Goes from here…

Sample Texts

```
[1] "When I presented the supplementary budget to this House l
ast April, I said we could work our way through this period of
severe economic distress. Today, I can report that notwithstan
ding the difficulties of the past eight months, we are now on
the road to economic recovery.\nIt is of enormous benefit that
the main political parties in this House share a common unders
tanding of the extent of our difficulties and even if we disag
ree on how to solve our problems, our agreement on the amount
of savings required sends a powerful signal to the rest of the
world that we are able and willing to put our own house in ord
er.\nToday, I want to tell the Irish people that even though o
ur economy is still in a weakened condition, and our self-conf
idence as a nation has been shaken, the Government's strategy
over the past 18 months is working and we can now see the firs
t signs of a recovery here at home and in our main internation
al markets.\nWe have taken bold, decisive and innovative steps
```

```
[1] "This draconian budget should not be happening today. It i
s happening, however, because Fianna Fáil failed to heed the w
arnings and drove this economy on to the rocks. Even now, the
thinking behind this budget is short-sighted. It is sucking us
into a cycle of more job losses and higher debt. People will h
urt badly after this budget, people who had no hand, act or pa
rt in creating the problem that we now face.\nThis is a jobles
s and a joyless budget. It offers no vision that would rebuild
confidence, it serves only to get the Taoiseach and his Minist
ers to the end of this week.\nThe only way to break out of the
cycle that has been created is with a convincing jobs strategy
and that strategy is simply missing. This requires real leader
ship from Government. We saw that sort of real leadership 50 y
```

… to here

Corpus consisting of 14 documents:

Text	Types	Tokens	Sentences	year	debate	number	foren	name	party
2010_BUDGET_01_Brian_Lenihan_FF	1953	8641	374	2010	BUDGET	01	Brian	Lenihan	FF
2010_BUDGET_02_Richard_Bruton_FG	1040	4446	217	2010	BUDGET	02	Richard	Bruton	FG
2010_BUDGET_03_Joan_Burton_LAB	1624	6393	307	2010	BUDGET	03	Joan	Burton	LAB
2010_BUDGET_04_Arthur_Morgan_SF	1595	7107	343	2010	BUDGET	04	Arthur	Morgan	SF
2010_BUDGET_05_Brian_Cowen_FF	1629	6599	250	2010	BUDGET	05	Brian	Cowen	FF
2010_BUDGET_06_Enda_Kenny_FG	1148	4232	153	2010	BUDGET	06	Enda	Kenny	FG
2010_BUDGET_07_Kieran_ODonnell_FG	678	2297	133	2010	BUDGET	07	Kieran	ODonnell	FG
2010_BUDGET_08_Eamon_Gilmore_LAB	1181	4177	201	2010	BUDGET	08	Eamon	Gilmore	LAB
2010_BUDGET_09_Michael_Higgins_LAB	488	1286	44	2010	BUDGET	09	Michael	Higgins	LAB
2010_BUDGET_10_Ruairi_Quinn_LAB	439	1284	59	2010	BUDGET	10	Ruairi	Quinn	LAB
2010_BUDGET_11_John_Gormley_Green	401	1030	49	2010	BUDGET	11	John	Gormley	Green
2010_BUDGET_12_Eamon_Ryan_Green	510	1643	90	2010	BUDGET	12	Eamon	Ryan	Green
2010_BUDGET_13_Ciaran_Cuffe_Green	442	1240	45	2010	BUDGET	13	Ciaran	Cuffe	Green
2010_BUDGET_14_Caoimhghin_OCaolain_SF	1188	4044	176	2010	BUDGET	14	Caoimhghin	OCaolain	SF

Tokens

- Every word that appears in the corpus
- Common to exclude infrequent words or words that appear in very few documents;
- Sometimes we use **bigrams** or **n-grams** (combination of two or more words)
- Tokenization is very tricky when working with East Asian languages
  - Tokenizing Japanese or Mandarin texts requires pre-trained deep learning models

Document-feature-matrix (dfm)

- aka dtm: document-term-matrix
- A matrix where:
    - each row is a document *i*
    - each column one unique token *k* from the entire corpus
    - each cell is filled by the nr of times the token *k* occurred in document *i*
- docvars: the other respective variables that describe each document (author, length, date, etc)

Stemming

- Process for reducing inflected (or sometimes derived) words to their stem, base or root form.
- Example: **comput**, from: *computer, compute, computation*
- Assumption: The concept is in the stemmed word and not in its specific case or usage
- There are different stemming algorithms and they might work better for some languages than others
- Stemming might sometimes create problems (e.g. when police and policy are reduced to the same word)

Stopwords

- Words that are excluded from the analysis, because they do not have substantial meaning and would not help us decipher the meaning of a text.
- Pre-installed in R for many languages, but with varying quality and depth
- Think before excluding reflexively! They may carry important information for your research question!

Example from quanteda

```
  [1] "i"          "me"         "my"         "myself"     "we"         "our"        "ours"       "ourselves"  "you"
 [10] "your"       "yours"      "yourself"   "yourselves" "he"         "him"        "his"        "himself"    "she"
 [19] "her"        "hers"       "herself"    "it"         "its"        "itself"     "they"       "them"       "their"
 [28] "theirs"     "themselves" "what"       "which"      "who"        "whom"       "this"       "that"       "these"
 [37] "those"      "am"         "is"         "are"        "was"        "were"       "be"         "been"       "being"
 [46] "have"       "has"        "had"        "having"     "do"         "does"       "did"        "doing"      "would"
 [55] "should"     "could"      "ought"      "i'm"        "you're"     "he's"       "she's"      "it's"       "we're"
 [64] "they're"    "i've"       "you've"     "we've"      "they've"    "i'd"        "you'd"      "he'd"       "she'd"
 [73] "we'd"       "they'd"     "i'll"       "you'll"     "he'll"      "she'll"     "we'll"      "they'll"    "isn't"
 [82] "aren't"     "wasn't"     "weren't"    "hasn't"     "haven't"    "hadn't"     "doesn't"    "don't"      "didn't"
 [91] "won't"      "wouldn't"   "shan't"     "shouldn't"  "can't"      "cannot"     "couldn't"   "mustn't"    "let's"
[100] "that's"     "who's"      "what's"     "here's"     "there's"    "when's"     "where's"    "why's"      "how's"
[109] "a"          "an"         "the"        "and"        "but"        "if"         "or"         "because"    "as"
[118] "until"      "while"      "of"         "at"         "by"         "for"        "with"       "about"      "against"
[127] "between"    "into"       "through"    "during"     "before"     "after"      "above"      "below"      "to"
[136] "from"       "up"         "down"       "in"         "out"        "on"         "off"        "over"       "under"
[145] "again"      "further"    "then"       "once"       "here"       "there"      "when"       "where"      "why"
[154] "how"        "all"        "any"        "both"       "each"       "few"        "more"       "most"       "other"
[163] "some"       "such"       "no"         "nor"        "not"        "only"       "own"        "same"       "so"
[172] "than"       "too"        "very"       "will"
```

Bag-of-words

- Approach that disregards grammar and word order, but keeps number of times a word is present.
- Advantage: Computationally easy and straightforward to understand and interpret
- Problem: We lose information and need to assume that ignoring word order does not bias results
- “War is good, peace is bad” and “Peace is good, war is bad” are exactly the same text for most text analysis approaches.

Usual pre-processing

- Commonly, researchers exclude numbers and punctuation and convert everything to lower case.
- It is common to exclude very rare words and words that only occur in some of the texts. Also, stemming makes sense and stopwords are mostly removed.
- However, we should always test whether the substantial results hold if we change these preprocessing steps!

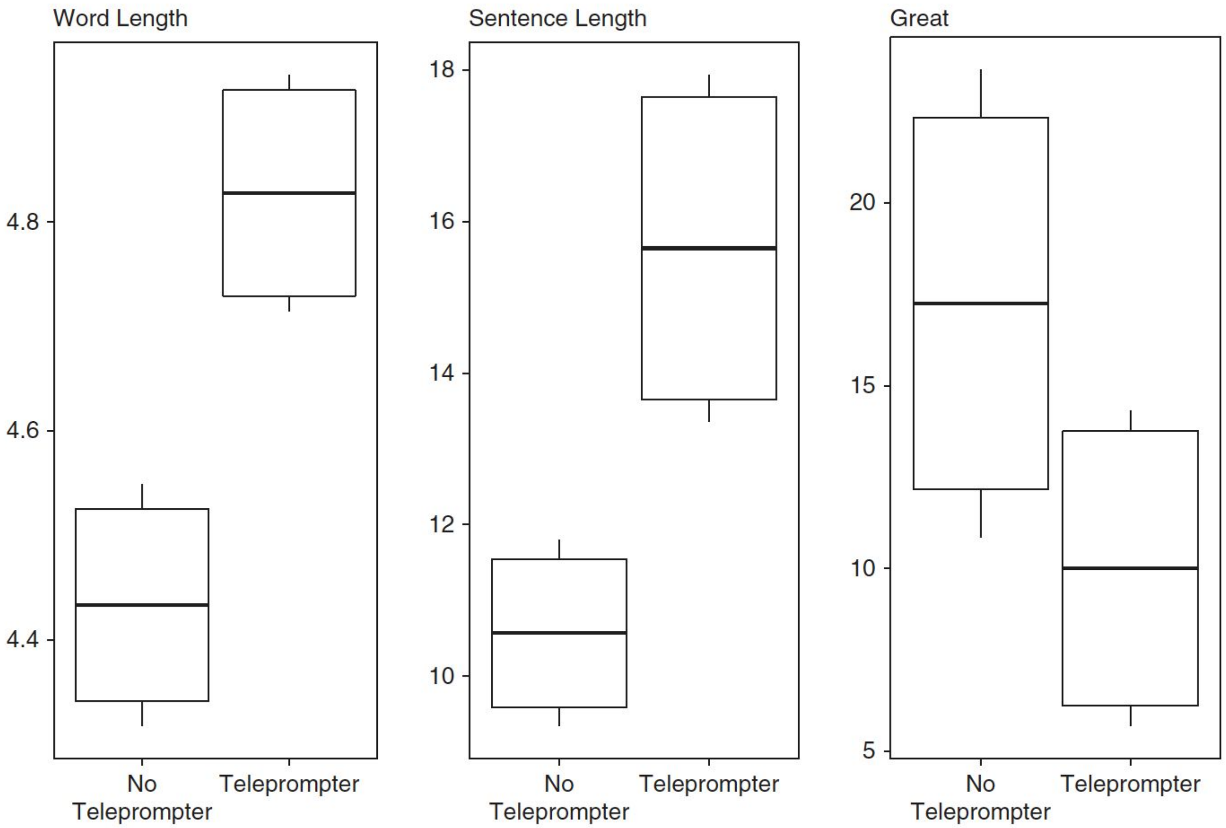
Questions?

Description

Describing the corpus can be substantively interesting

- Do left or right (or populist/non-populist) politicians have simpler speech?
- Do male MPs speak more and longer than female MPs?
- How similar are two documents?
    - E.g. two opinions by Supreme Court justices?

Trump speeches (Hawkins/Littvay 2019)



Descriptive methods

- **Readability:** use a combination of syllables and sentence length to indicate complexity of speech
- **Lexical diversity**: examples are type-to-token ratio (unique words as types, total words as tokens)
- **Length**: characters, words, sentences, paragraphs, etc.

Example - Flesch Reading Ease Index

- Optimized for English
- The higher FRE, the easier the text;

Formula:

FRE = 206.835 - 1.015(# words/#sentences) - 84.6(# syllables/# words)

- Most values between 0 and 100
- Default in quanteda, but there’s many more formulas

Comparing texts

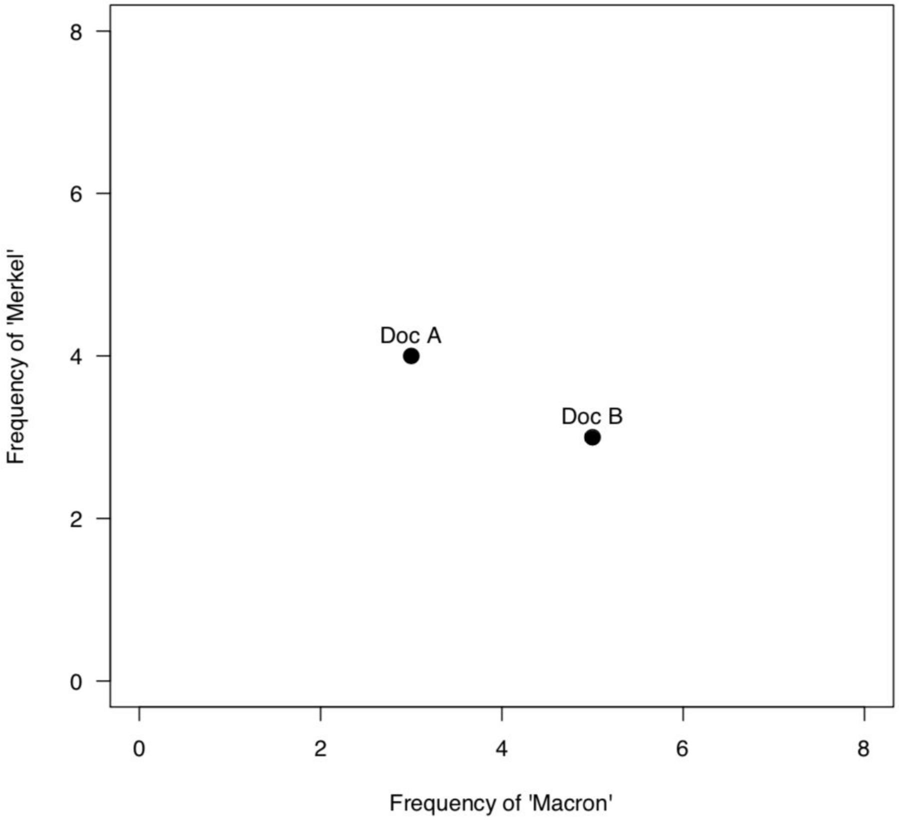
- Basic approach: features form a vector for each document
- In most applications: the row of a document-feature matrix (e.g. document-term matrix)
- We may be interested in evaluating the pairwise similarity between documents

Illustration

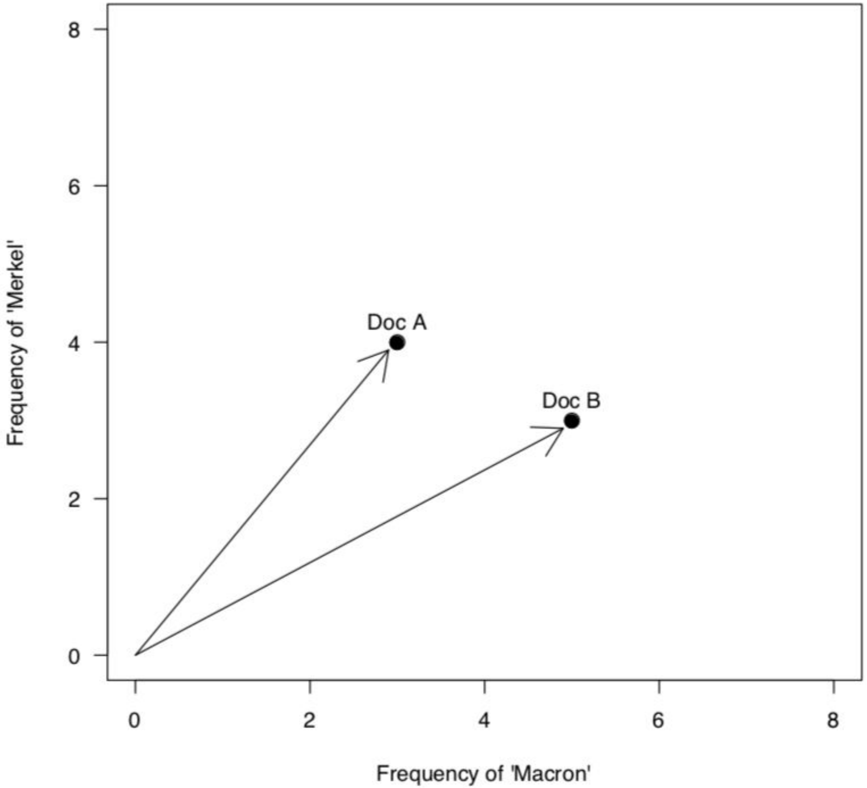
- Suppose we have two documents and they only mention two words: “Macron” and “Merkel”. Our DTM look as follows:

	Macron	Merkel
Doc A	3	4
Doc B	5	3

We can represent these documents as vectors



Drawing arrows

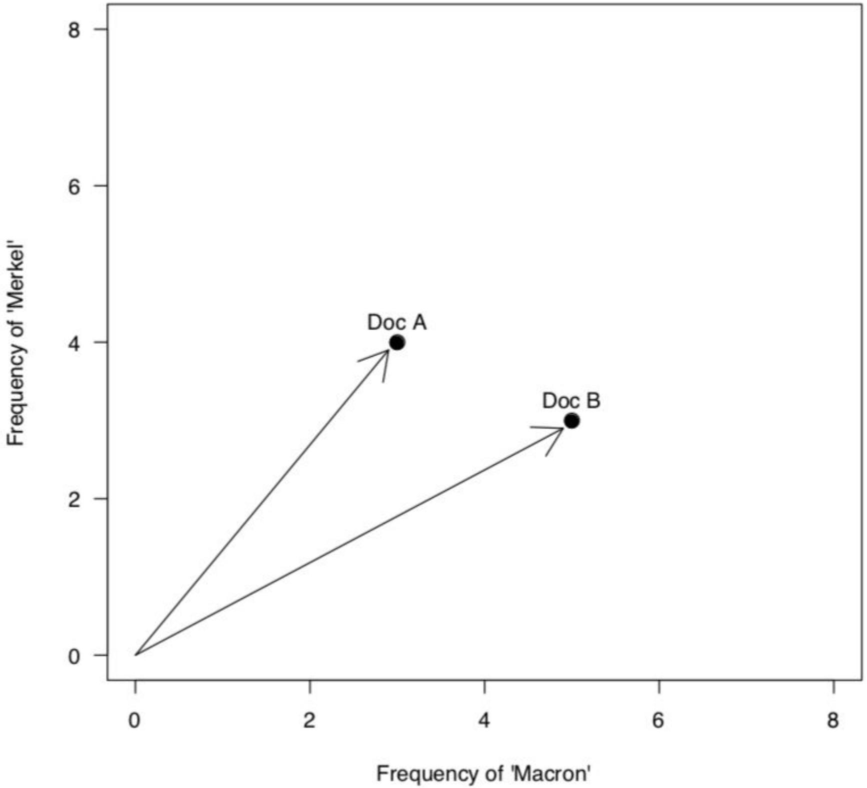


Cosine similarity

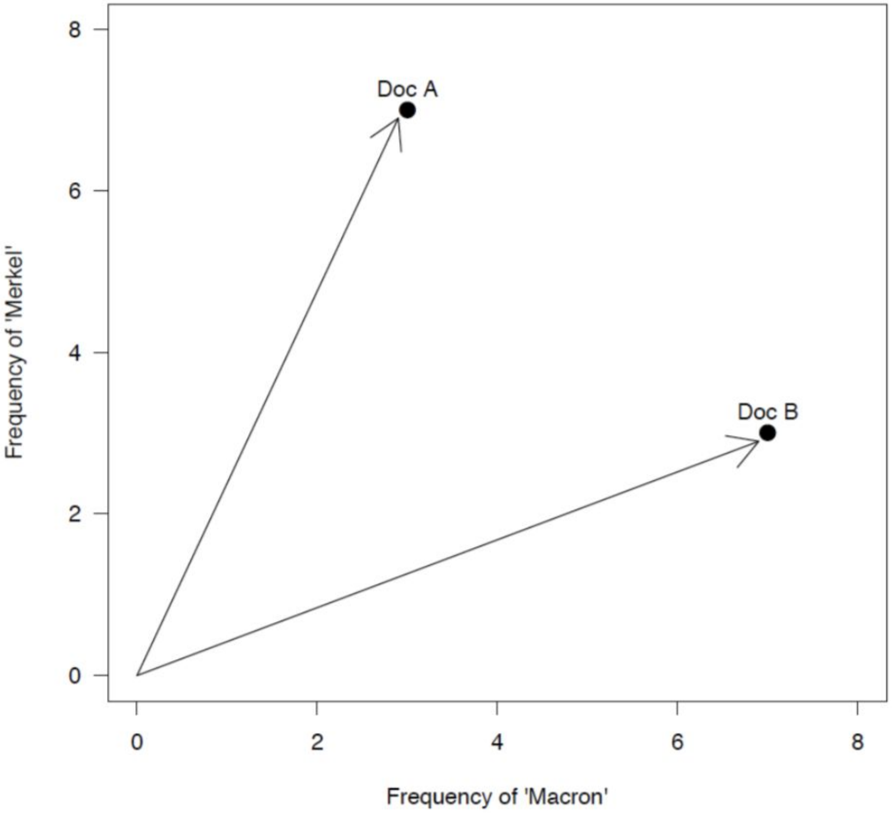
- The more similar the frequencies, the smaller the angle between the vectors
- Based on the size of the angle between vectors
- Calculated by the dot product of two vectors, normalized by the product of the vector lengths
- Output values close to 1 indicate high similarity.

$$cos(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x} \cdot \mathbf{y}}{||\mathbf{x}|| \cdot ||\mathbf{y}||}$$

Illustration



Illustration



Issues with cosine similarity

- In most applications we are not interested in pairwise comparisons;
- Would rather put a large number of documents into some kind of scale;
- Still reliant on words being exactly the same. The following sentences would be considered entirely different:
  - Obama spoke to followers in his hometown
  - The former president addressed supporters in Chicago

Enough talking. R