Automated Text Analysis in Political Science

Lecture 1: Introduction to text as data

May 6, 2019

---

dr. Martijn Schoonvelde

School of Politics and International Relations, UCD

Automated Text Analysis

Today's class

- Why automated text analysis?
- Setting up the course
- Steps in a canonical text analysis project

Who am I?

- Assistant professor in political science at University College Dublin
- Work on automated text analysis of political rhetoric of leaders
  - Also: psychology and text; validation of methods; and media effects
- Before at Vrije Universiteit, University of Exeter, European University Institute, and Stony Brook University

Contact

- Available throughout the course
- Ask questions, come talk to me – happy to help
- mschoonvelde@gmail.com, @hjms
- All materials (updated as we go along):
  `https://github.com/hjmschoonvelde/CEU_ATA_2019`

Who are you?

- Background, interests, experience with R
- What do you expect from this course?

Why Automated Text Analysis for Political Scientists?

- Political actors (politicians, political parties, voters, etc) produce huge amounts of text, much of which is available and stored online
  - Parliamentary speeches, interviews, blog posts, manifestos, tweets, etc.

Why Automated Text Analysis for Political Scientists?

- Political actors (politicians, political parties, voters, etc) produce huge amounts of text, much of which is available and stored online
  - Parliamentary speeches, interviews, blog posts, manifestos, tweets, etc.
- Exciting possibilities to analyze politics beyond elections / surveys
  - Fine-grained analysis of public opinion, political behavior, social networks, etc.

Why Automated Text Analysis for Political Scientists?

- Political actors (politicians, political parties, voters, etc) produce huge amounts of text, much of which is available and stored online
  - Parliamentary speeches, interviews, blog posts, manifestos, tweets, etc.
- Exciting possibilities to analyze politics beyond elections / surveys
  - Fine-grained analysis of public opinion, political behavior, social networks, etc.
- This requires a new set of analytical tools, which automated text methods provide

Why Automated Text Analysis for Political Scientists?

- But: automated text analysis comes with many assumptions (for example, bag of words)
- These can bring us a lot but should not be taken for granted (be mindful of the decisions you make)
- Automated methods should never replace reading, but can help us reading better (and more)
    - "Amplify resources and augment humans" (Grimmer and Stewart, 2013)

Culturomics, Michel et al. (2011)

RESEARCH ARTICLE

Quantitative Analysis of Culture Using Millions of Digitized Books

Jean-Baptiste Michel,[1,2,3,4,5]*† Yuan Kui Shen,[2,6,7] Aviva Presser Aiden,[2,6,8] Adrian Veres,[2,6,9] Matthew K. Gray,[10] The Google Books Team,[10] Joseph P. Pickett,[11] Dale Hoiberg,[12] Dan Clancy,[10] Peter Norvig,[10] Jon Orwant,[10] Steven Pinker,[5] Martin A. Nowak,[1,13,14] Erez Lieberman Aiden[1,2,6,14,15,16,17]*†

We constructed a corpus of digitized texts containing about 4% of all books ever printed. Analysis of this corpus enables us to investigate cultural trends quantitatively. We survey the vast terrain of 'culturomics,' focusing on linguistic and cultural phenomena that were reflected in the English language between 1800 and 2000. We show how this approach can provide insights about fields as diverse as lexicography, the evolution of grammar, collective memory, the adoption of technology, the pursuit of fame, censorship, and historical epidemiology. Culturomics extends the boundaries of rigorous quantitative inquiry to a wide array of new phenomena spanning the social sciences and the humanities.

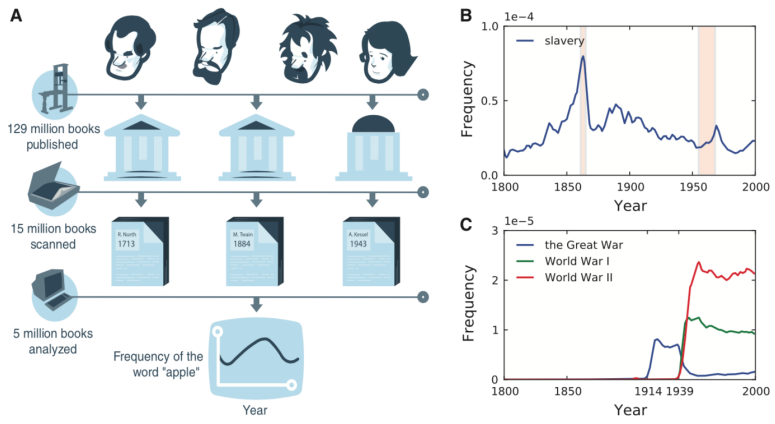


Fig. 1. Culturomic analyses study millions of books at once. (**A**) Top row: Authors have been writing for millennia; ~129 million book editions have been published since the advent of the printing press (upper left). Second row: Libraries and publishing houses provide books to Google for scanning (middle left). Over 15 million books have been digitized. Third row: Each book is associated with metadata. Five million books are chosen for computational analysis (bottom left). Bottom row: A culturomic time line shows the frequency of "apple" in English books over time (1800–2000). (**B**) Usage frequency of "slavery". The Civil War (1861–1865) and the civil rights movement (1955–1968) are highlighted in red. The number in the upper left (1e-4 = 10^{-4}) is the unit of frequency. (**C**) Usage frequency over time for "the Great War" (blue), "World War I" (green), and "World War II" (red).

Culturomics didn't really catch on, became digital humanities instead

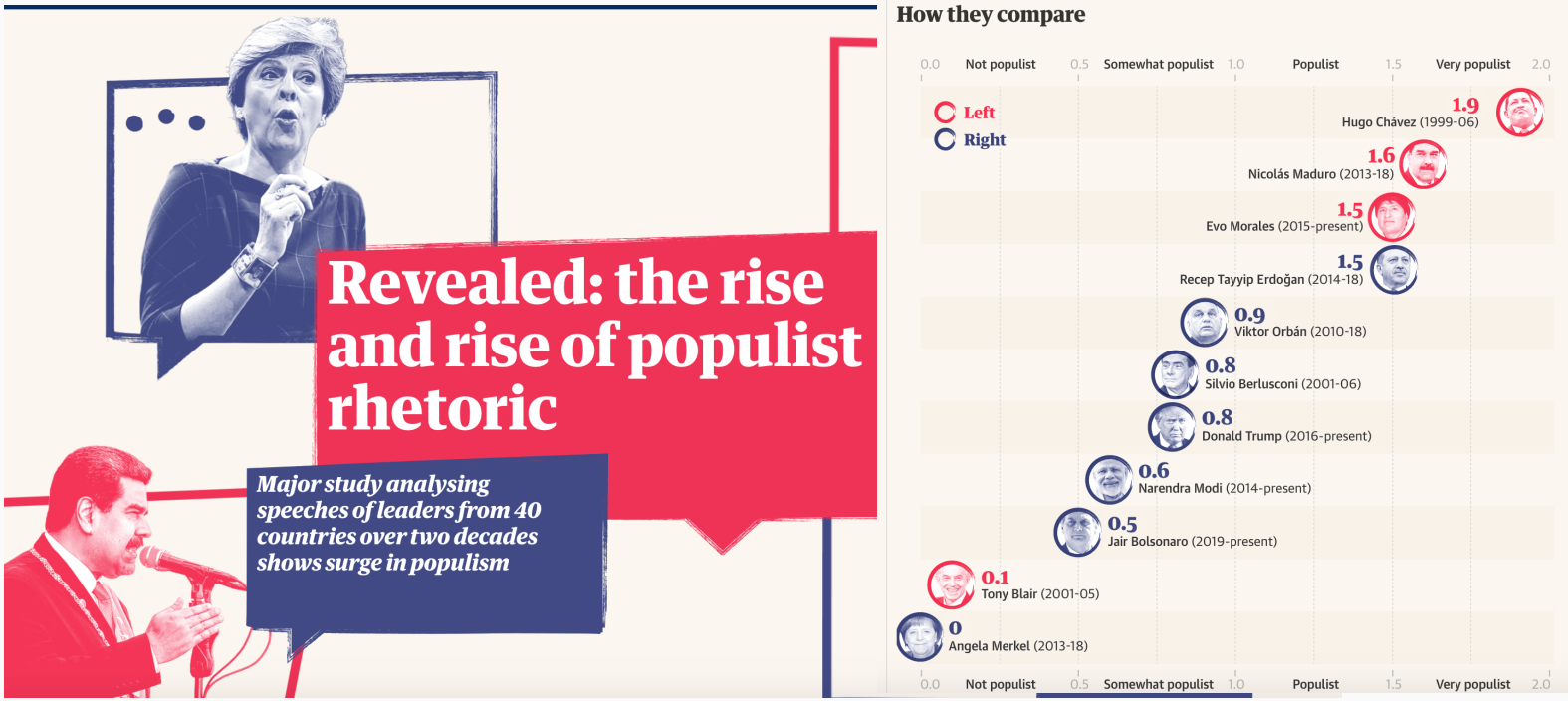
This course

- Introduction of automated text analysis methods in political science using R
- We'll cover the bigger picture of doing research using text
  - However, each step of the research design could fill its own course
  - In reality automated text analysis is not a one-size-fits all type situation, but highly idiosyncratic
- Use this course to figure out what you find interesting and want to pursue further
- Ask questions – and help each other out

This course

- No better time to learn these methods than now
- We'll mostly focus on bag-of-words models but spend some time on word embeddings next week
- Lots of cool developments across disciplines!
  - In computer science and linguistics
  - But also communication science and psychology, economics and history

Text and populism



Collaboration between The Guardian and a large group of academics, including CEU's Leventte Litvay and Erin Jenne

Personality in text

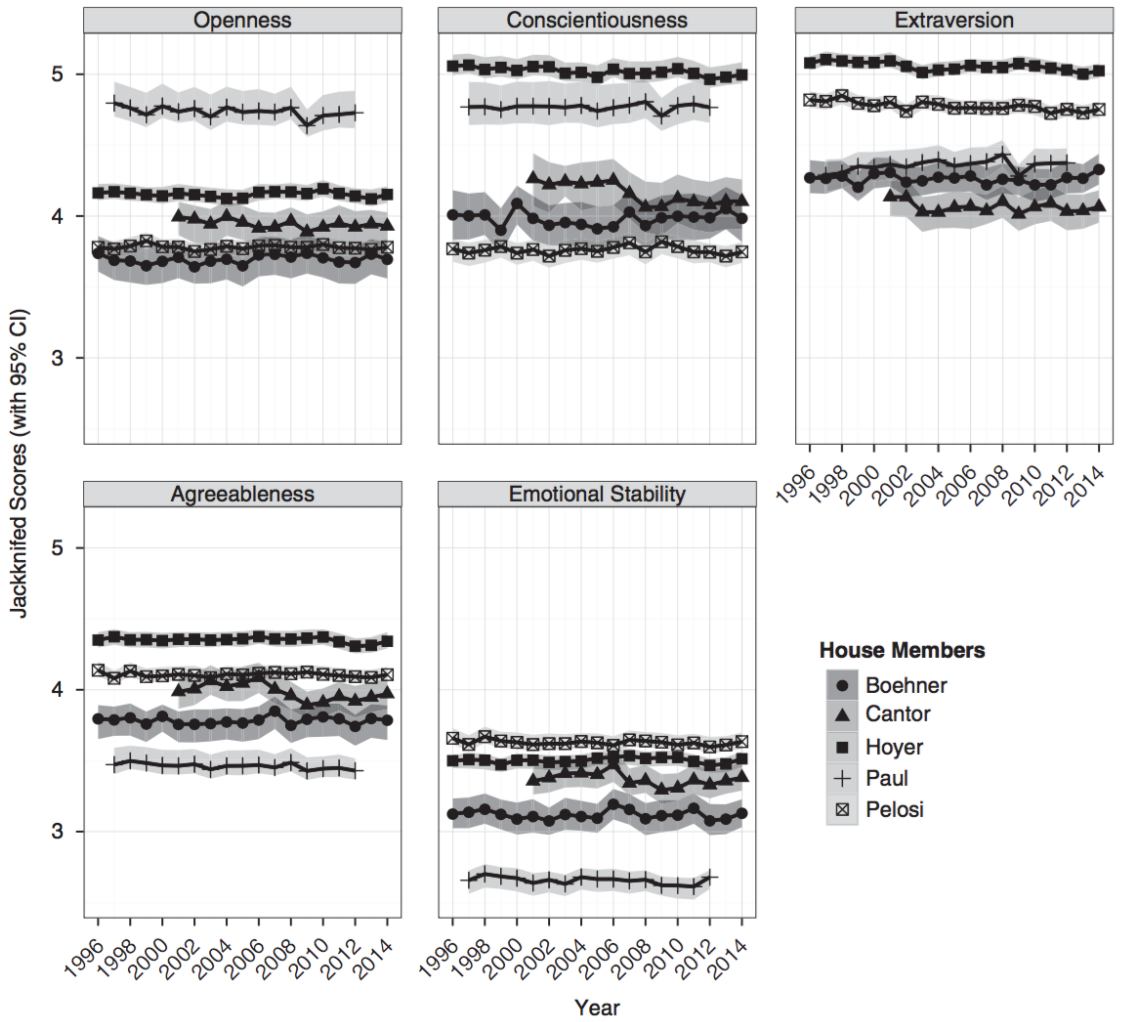
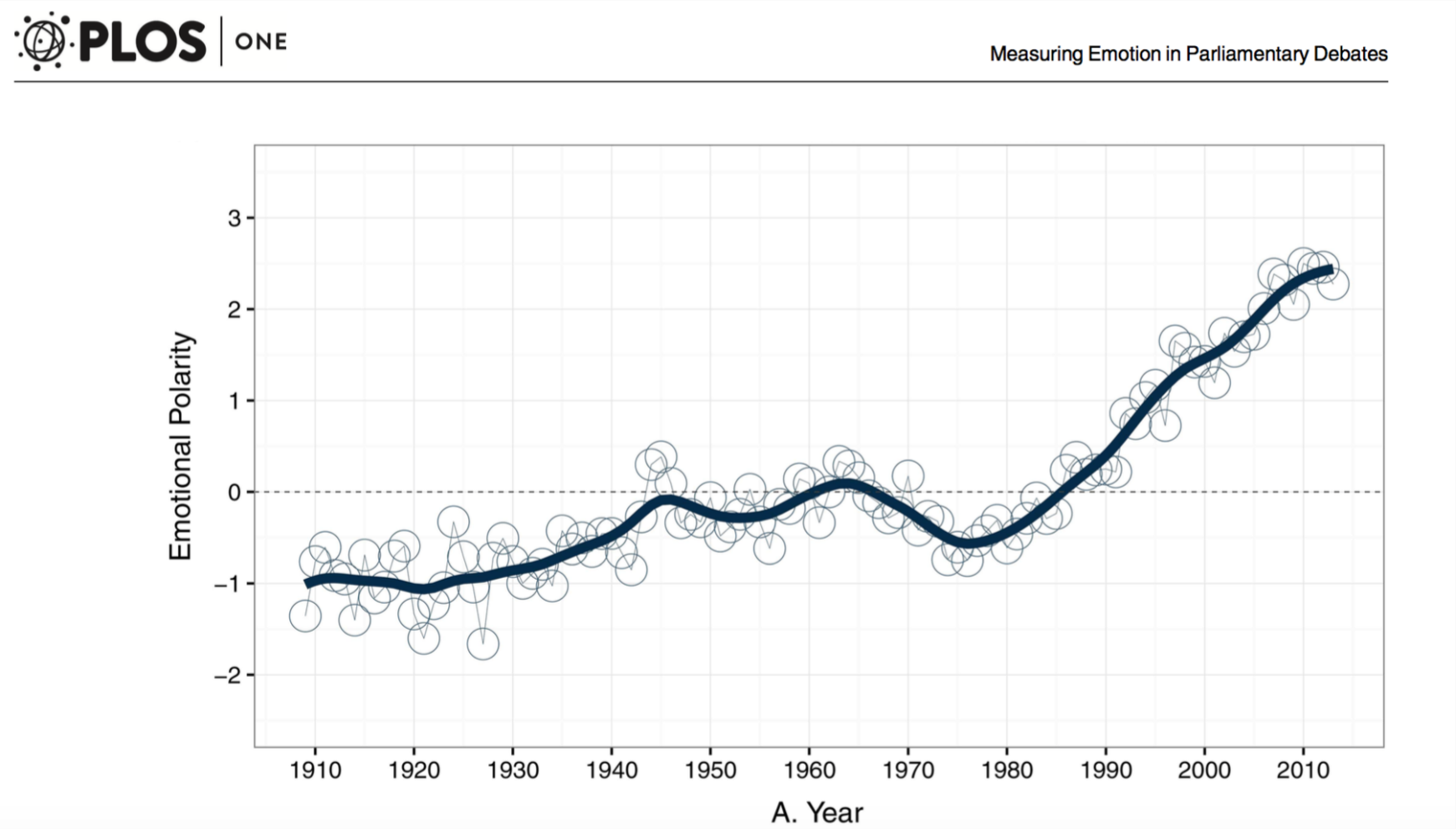


Fig. 2. House scores over time (selected members)
Note: CI = confidence interval.

Sentiment in text



Requirements: grit

Requirements: fun

Course objectives

- Learn about automated text analysis methods in political science
  - **But**: very much an interdisciplinary field
- Practice preprocessing and analyzing text using R
- Think about and set up a research design using text as data
- Critically evaluate existing text as data research

Assessment

- Attendance & participation in class (10%)
- Two coding assignments (30 %)
  - EUSpeech dataset
- Presentation of a research design (15 %)
- Research note (45%)

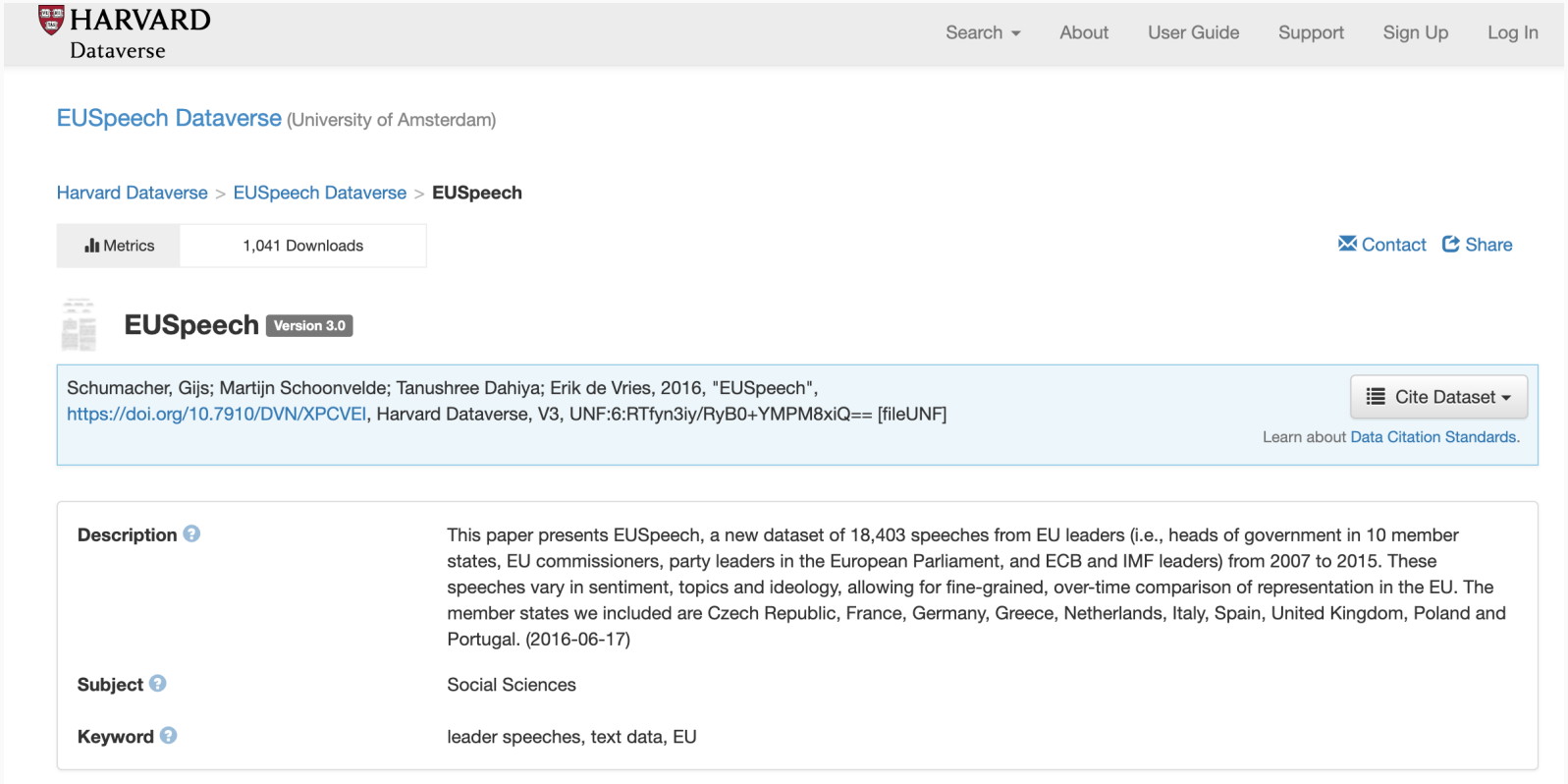
Next Thursday: Flash Talks

- Presentation of an interesting paper / R package / finding that you want to share
  - Why could it be useful for fellow students?
- What is the approach? Results? Is there an R package available?

Why R?

- Encompasses all steps of the research process (from scraping to data viz / analysis)
- Helpful user community
- Lots of development, new packages
- Other languages possible as well of course
    - E.g., Python for some tasks and R for other tasks

EUSpeech dataset



EUSpeech

- 18000+ speeches from leaders in the EU (national and transnational), 2007–2015
- We'll use this dataset for the two coding assignments and as a practice dataset

Steps in a typical text project

1. Sample texts / get data
    - Develop the **corpus** to be analyzed
2. Clean text
    - E.g., remove all bycatch from web scraping
    - Make machine-readable
    - Define **documents**, the unit of analysis
        - Paragraphs, documents, sentences
3. Preprocess text
    - From words to numbers
4. Analyze text
5. Visualize outcomes
6. Write up results

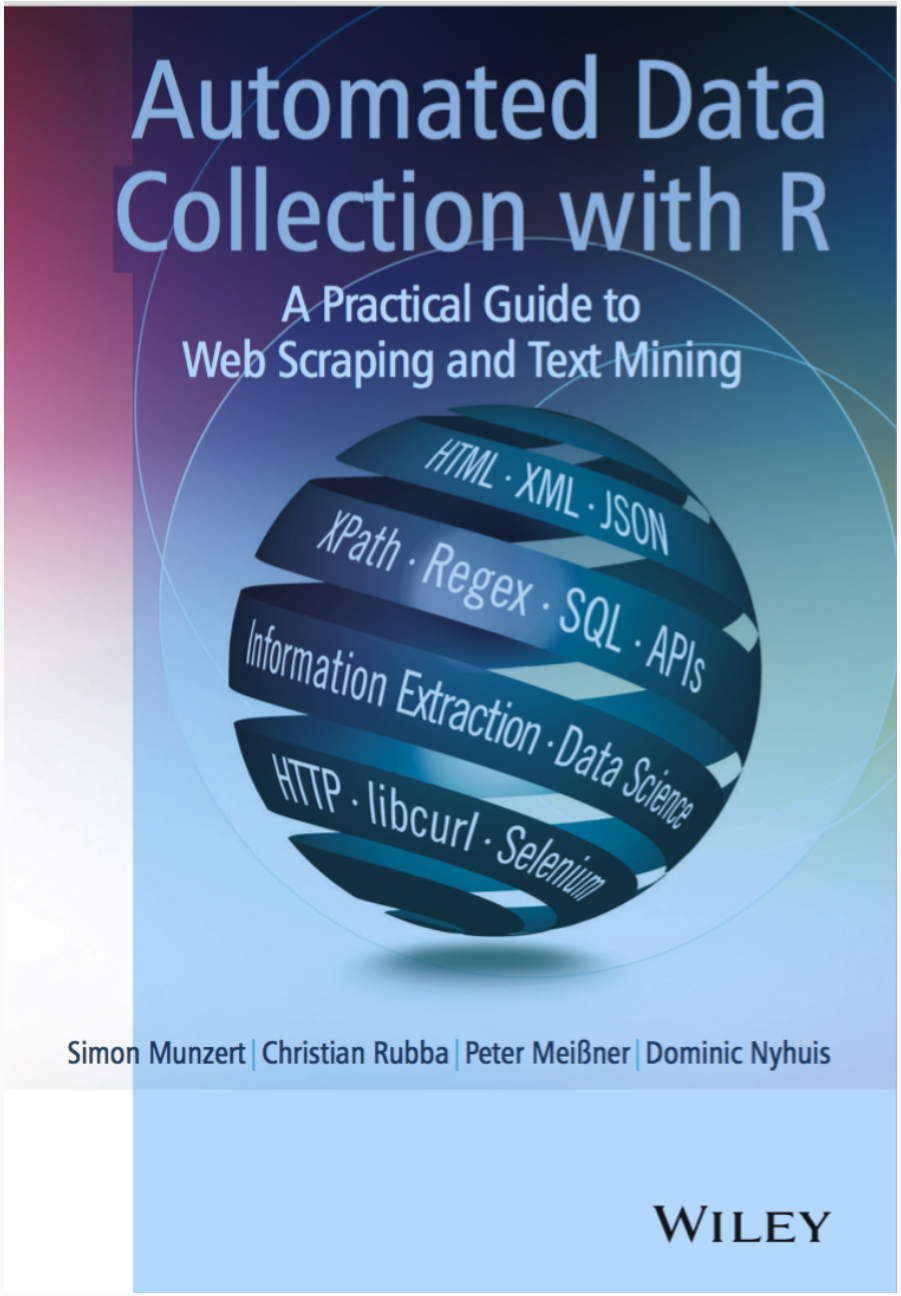
Where to find text data?

- Repositories such as Lexis Nexis (newspaper data)
- Existing text datasets. For example:
  - EUSpeech (Schumacher et al): Harvard Dataverse
  - Parlspeech (Rauh et al): Harvard Dataverse
  - Party manifestos: `https://manifesto-project.wzb.eu/`
- Replication data repositories
- Getting data from the web

Getting data from the web

- Download structured text data directly from the web
  - E.g., if text is stored in csv files or Rdata files in repositories
- API (Application Program Interface)
  - Makes parts of a website available to your computer
  - Google 'CRAN website url' – R packages as API clients
  - `httr`
  - Oftentimes data stored in JSON and XML format
- Web scraping / screen scraping
  - `rvest`
  - Oftentimes data stored in HTML and XML format
- Last resort: copy-paste text
  - Not recommended

Getting data from the web



Cleaning data

- Expect a lot of trial and error
- Important consideration: character encoding
  - Mapping of bits to understandable characters
  - Many coding schemes exists, with different methods of encoding "extended" characters
- Some background: `http://kunststube.net/encoding/`
- From Welbers *et al.* (2017):

tion is simple: in R, ensure that all texts are encoded as UTF-8, either by reading in UTF-8 texts, or converting them from a known encoding upon import. If the encoding is unknown, *readtext's encoding* function can be used to guess the encoding. *readtext* can convert most known encodings (such as ISO-8859-2 for Central and Eastern European languages, or Windows-1250 for Cyrillic—although there are hundreds of others) into the common UTF-8 standard. R also offers additional low-level tools for converting character encodings, such as a bundled version of the GNU *libiconv* library, or conversion though the *stringi* package.

Preprocessing data

- Stemming, lemmatization, number removal, stopword removal, etc.
- We'll discuss these steps (and how to do them in R) in detail tomorrow
- Goal: select most relevant **features**, and allow for comparisons between documents in a corpus

Analyze text

- Three broad types of analysis (Boumans & Trilling 2016), from most deductive to most inductive:
  - counting and dictionary methods
  - supervised machine learning
  - unsupervised machine learning
- We'll encounter many applications in the next two weeks
  - Sentiment analysis using dictionary methods and supervised machine learning methods
  - Scaling methods (Wordscores and Wordfish)
  - Topic models (LDA, structural topic model)

Tools in R

- Quanteda library

  `In R > Install.packages("Quanteda")`

- Tidytext

  `In R > install.packages("Tidytext")`

- These packages are designed to analyze text and build on other packages and functions
- Check them out; see what they can do; and be flexible with using one or the other
- They are installed on the lab computers but you might need to install them on your work computers

Research Note

Tomorrow

- Read the assigned papers
- If you are using a laptop, make sure you have a recent version of R and RStudio installed
- Look at the following snippet of text and list all the ways (you can think of) that it needs to be cleaned:

```
<p>Ladies and gentlemen,</p><p>It is an honour to be here today to introduce the
theme of 'recession and recovery'. If you will permit,  I would like to suggest that
this afternoon we focus more on recovery than on recession. I think we know enough
about the recession side  of the story.</p><p>It started with the fall of Lehman
Brothers on 15 September 2008.. I happened to be here, at the Blouin Creative
Leadership Summit, only ten days later. Everyone was talking about the collapse of
Lehman. They were shocked and alarmed.  But even then we could hardly imagine that
its impact would be so dramatic, so historic.</p><p>As we now know, this event
triggered a global financial and economic crisis. Governments were forced to give
cash injections running into billions to prevent an economic and financial meltdown.
When credit dried up and demand fell, businesses struggled  to keep their heads above
water, and many went under. Ordinary people's jobs, homes and pensions were at
risk.</p><p>
```

- Familiarize yourself with R Markdown in RStudio: Link