TEXT AS DATA: WEEK 1

MATTHIAS HABER

8 SEPTEMBER 2021

GOALS FOR TODAY

GOALS

- General introduction
- Text as data in the social sciences
- Getting you setup

INTRODUCTION

ABOUT MYSELF

- Partner & Senior Director for Data & Analytics at Looping Group since 2018
- Postdoc at Hertie 2015-2017 (Governance Report)
- PhD in PolSci (University of Mannheim)
- Research on parties, legislative politics, electoral behavior
- First started programming in 2011

Contact:

haber.matthias@gmail.com

ABOUT YOURSELF

- Who are you?
- Why did you take this class?
- What kind of textual data are you interested in and which methods/techniques would make your life easier?

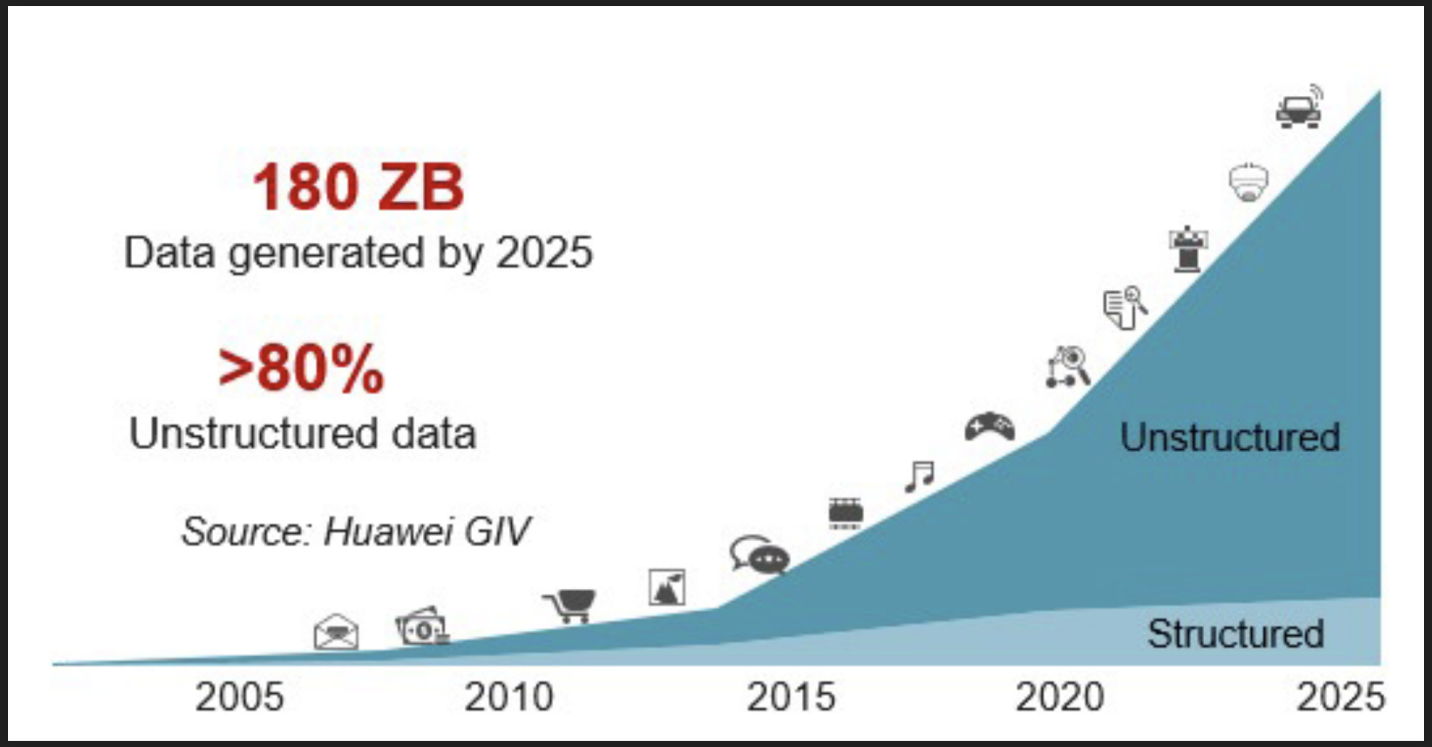
TEXT AS DATA

TEXT IN THE SOCIAL SCIENCES

> *Text is the most pervasive — and certainly the most persistent — artifact of political behavior (Monroe and Schrodt 2009)*

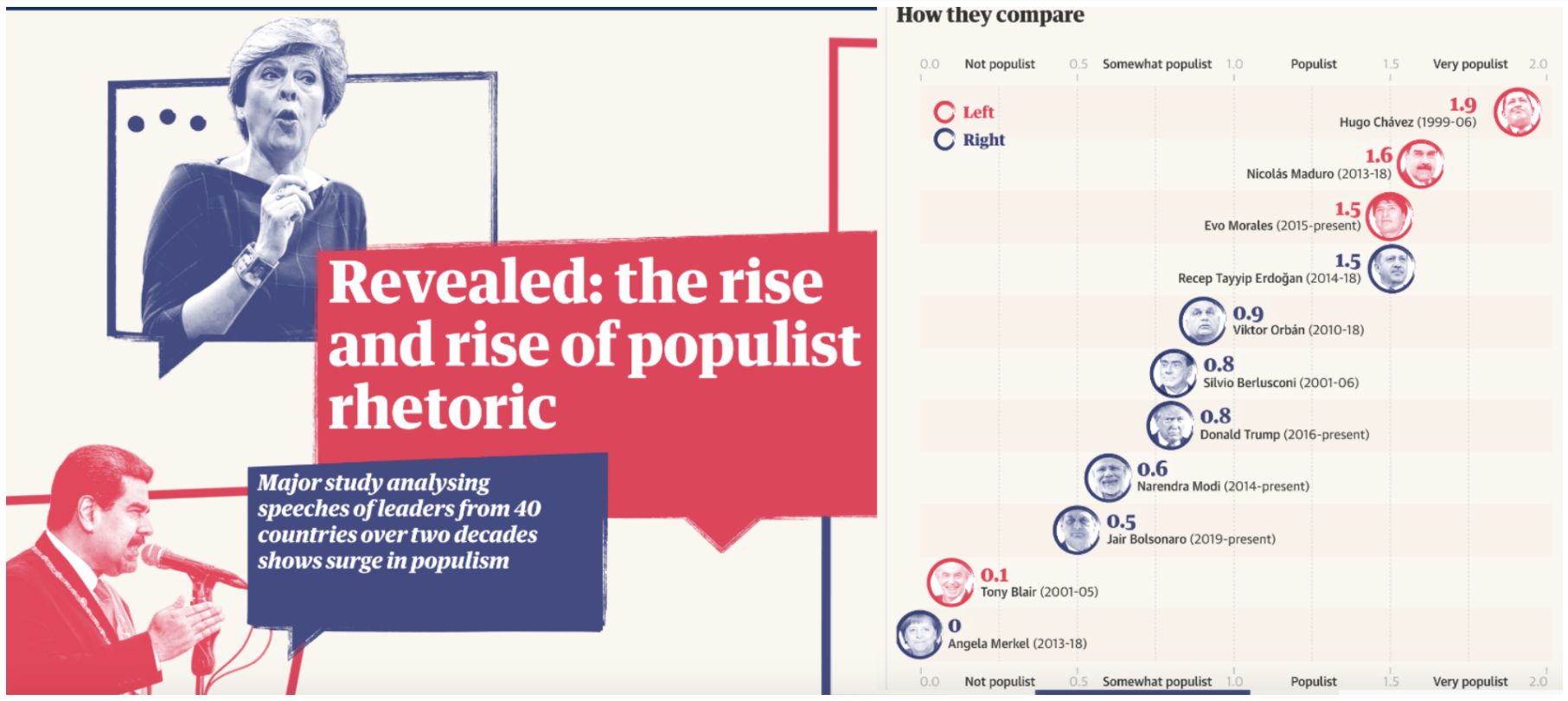
Never before has so much text been so readily available in the social sciences: Legislative debates, party manifestos, political speeches, committee transcripts, lobbying documents, court opinions, laws, …

UNSTRUCTURED DATA



There is no better time to start learning how to turn text into data than now!

TEXT AS DATA EXAMPLE: SCALING



TEXT AS DATA EXAMPLE: PERSONALITY

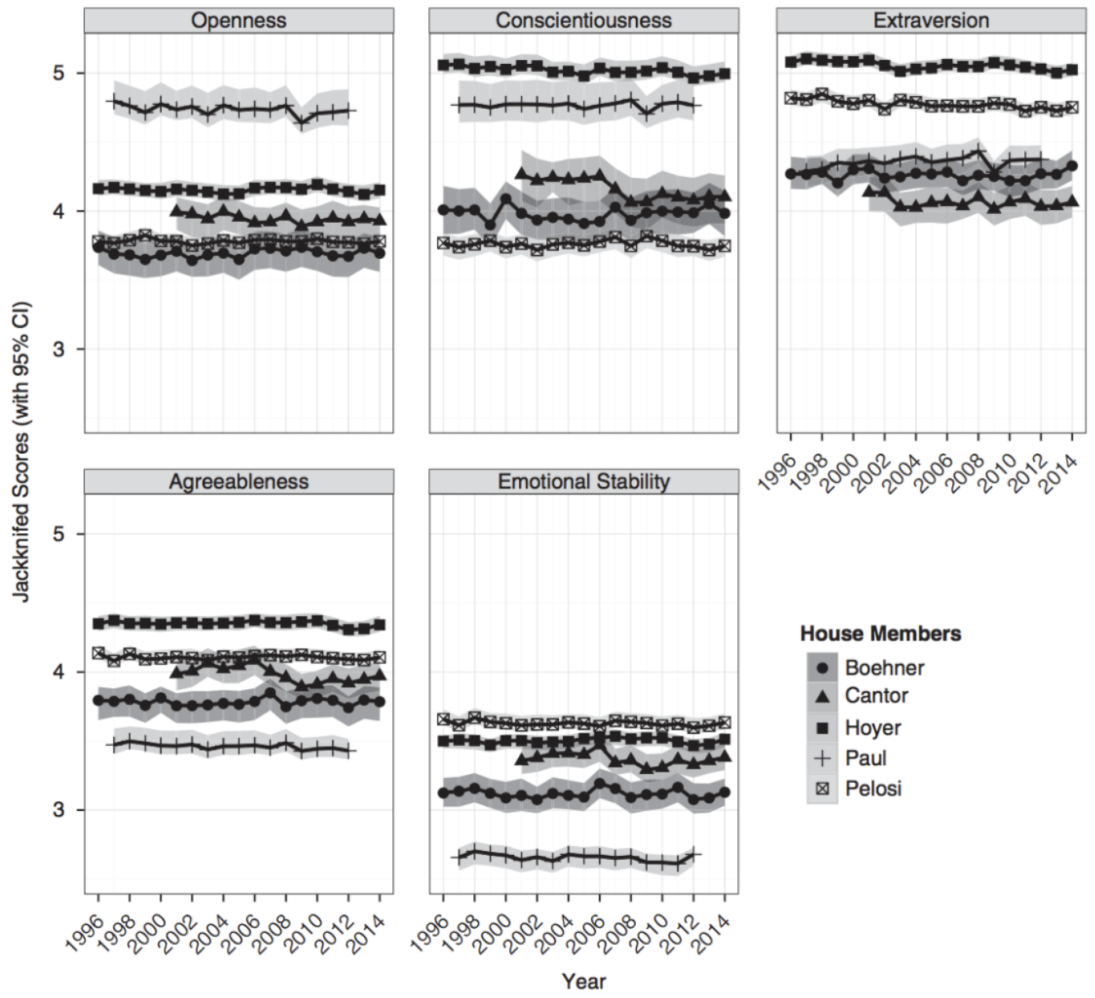
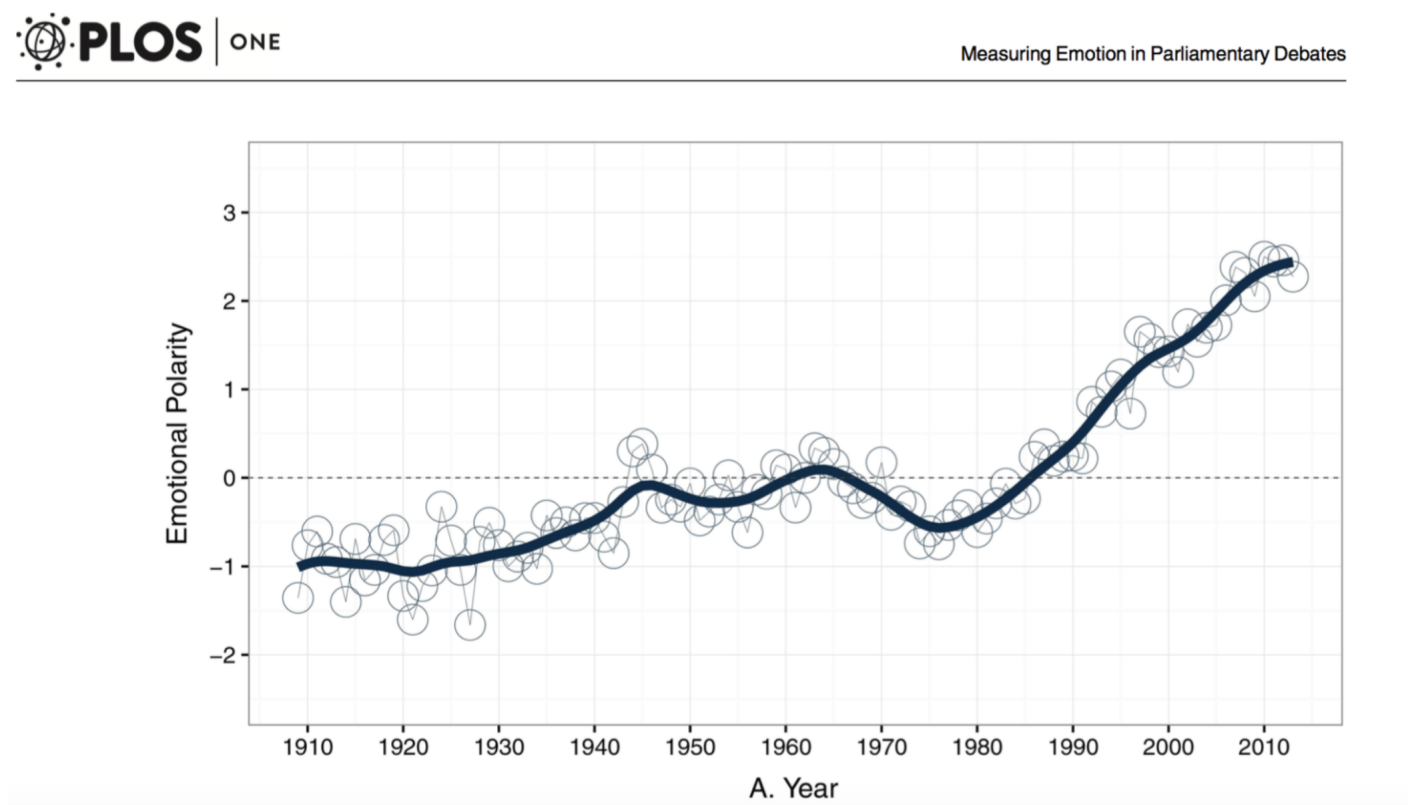
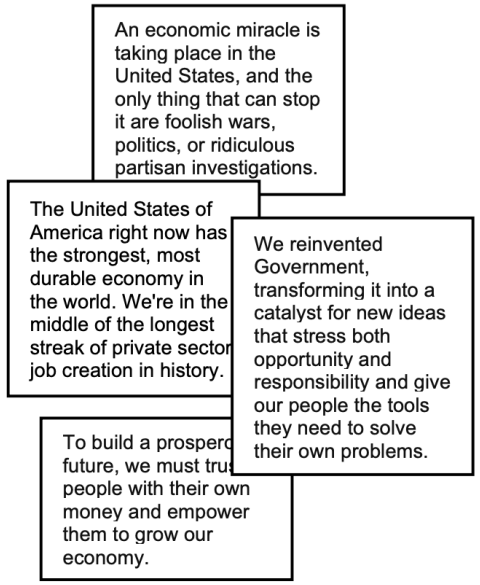


Fig. 2. House scores over time (selected members)
Note: CI = confidence interval.

TEXT AS DATA EXAMPLE: SENTIMENT



FROM TEXT TO DATA



Source texts

Processed text as a document-feature matrix

documents	features				
	economy	united	wall	crime	climate
Clinton–2000	10	4	1	5	1
Bush–2008	6	4	0	0	1
Obama–2016	16	4	1	0	4
Trump–2019	5	19	6	2	0

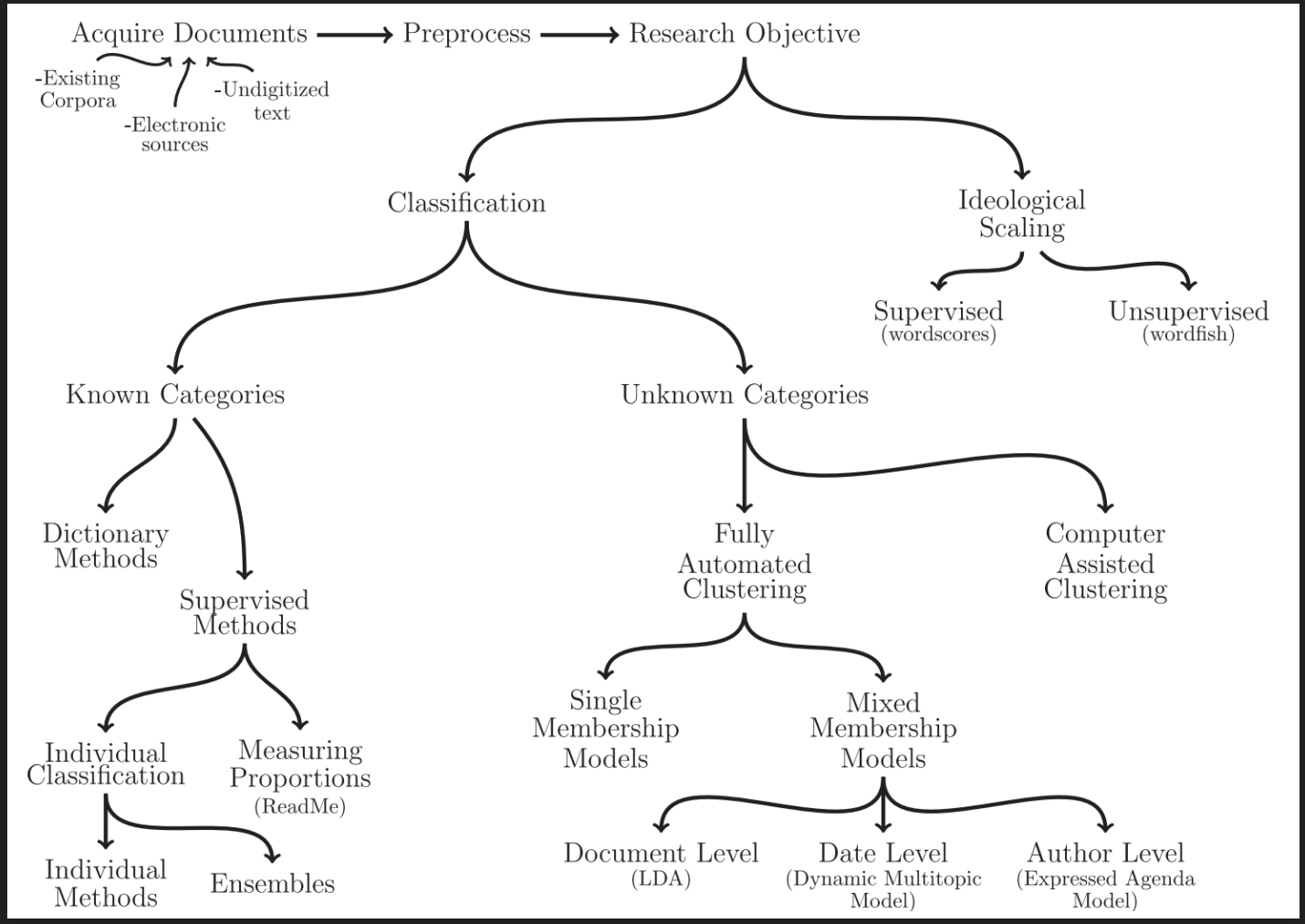
Quantitative analysis and inference

Describing texts quantitatively or stylistically
Identifying keywords
Measuring ideology or sentiment in documents
Mapping semantic networks
Identifying topics and estimating their prevalence
Measuring document or term similarities
Classifying documents

THERE ARE BASICALLY THREE VARIETIES OF TEXT ANALYSIS

1. Literary (Discourse) analysis -> text as text
2. Qualitative text analysis -> human generated textual data
3. Quantitative text analysis -> machine generated textual data

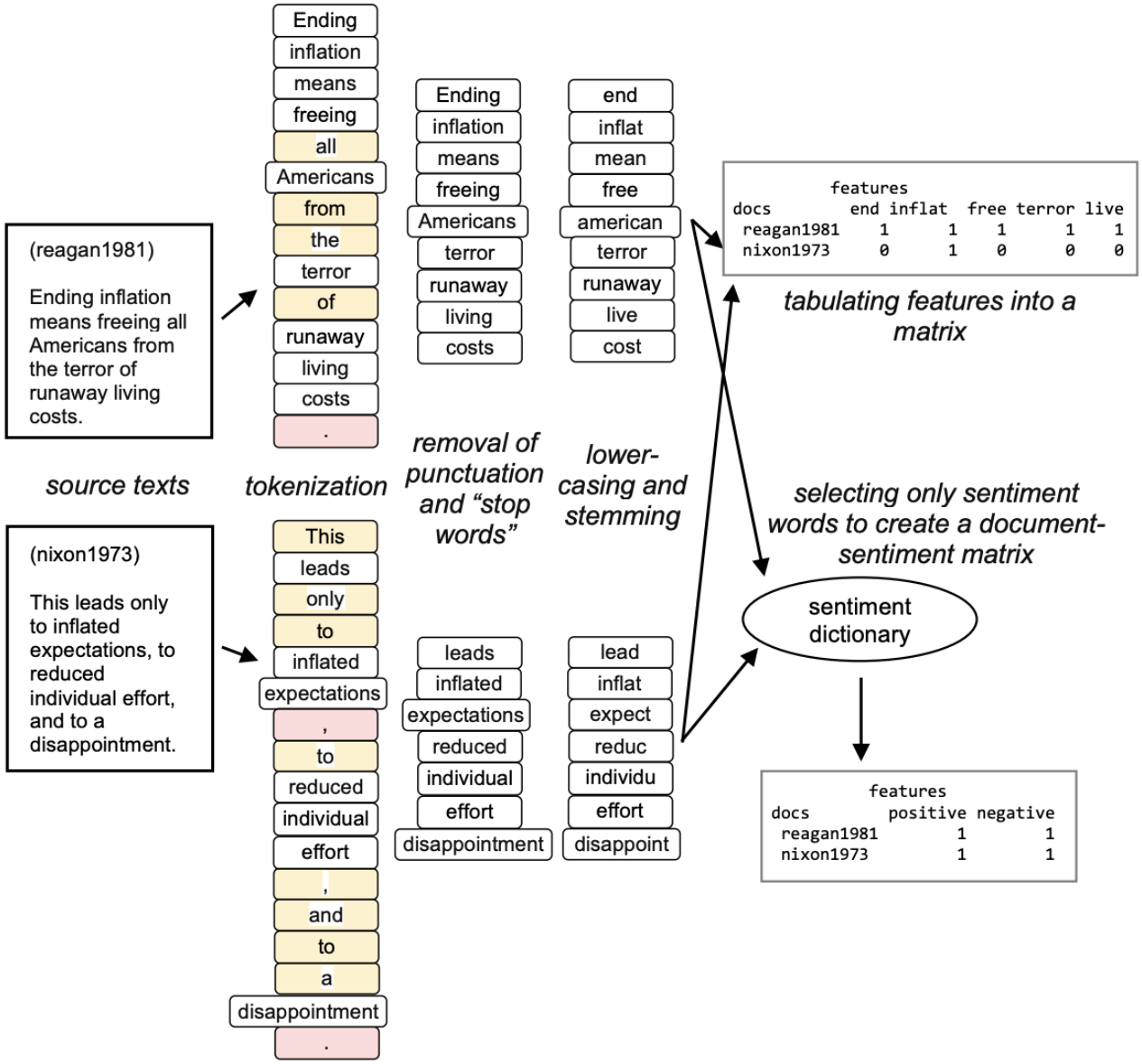
TEXT AS DATA METHODS



STAGES IN ANALYZING TEXT AS DATA (BENOIT 2019)

1. Selecting texts and defining the corpus.
2. Converting of texts into a common electronic format.
3. Defining documents and choosing the unit of analysis.
4. Defining and refining features.
5. Converting of textual features into a quantitative matrix.
6. Analyzing the (matrix) data using an appropriate statistical procedure.
7. Interpreting and reporting the results

FROM TEXT TO MATRIX



4 PRINCIPLES OF AUTOMATED TEXT ANALYSIS

1. All quantitative models of language 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

TEXT AS DATA COURSE

THIS COURSE

- Introduction to automated text analysis methods using R
- We focus on application, not on theory
- This course should give you a skill set that you can use in your own projects and research
- Ask questions - and help each other out!

FIRST HALF - DATA WRANGLING FUNDAMENTALS

Session	Session Date	Session Title
1	08.09.2021	Text as data (Room 2.32)
2	15.09.2021	Importing Textual Data (Room 2.32)
3	22.09.2021	Cleaning and Transforming Text
4	29.09.2021	String Manipulation
5	06.10.2021	Preparing text for analysis
6	13.10.2021	Descriptive analysis and visualization of text

Mid-term Exam Week

SECOND HALF - SUPERVISED AND UNSUPERVISED LEARNING

Session	Session Date	Session Title
7	27.10.2021	Dictionary Approaches
8	03.11.2021	Sentiment Analysis
9	10.11.2021	Topic Models
10	17.11.2021	Scaling Models
11	24.11.2021	spaCy - An Introduction to Industry NLP (Room 2.32)
12	01.12.2021	Transformers and the State of the Art of NLP (Room 2.32)

ASSIGNMENTS

Over the course of the semester, you will have to complete three assignments.

ASSIGNMENT 1

The first assignment will be a code completion exercise in which you'll be given a partially complete code. Full points will be awarded if you get the whole script to work.

ASSIGNMENT 2

The second assignment will involve the construction of a topic model, including a visualization and brief discussion of the results. Your grade for this assignment will be based on how you approach this task and if you follow the steps discussed in class.

ASSIGNMENT 3

Your final assignment is a presentation of a research project, or part thereof, of your own choosing. The only requirement is that the project involves the analysis of text and uses the methods that we cover in the course. Please record your presentation and submit it via Moodle. Grading will be determined by the quality of the presentation and the degree to which you manage to apply the skills what you have learned during the course.

GRADING

Composition of the Final Grade

Name	Percent of Final Mark	Due
Assignment 1: code completion exercise	30%	October 22, 8 am
Assignment 2: data analysis exercise	30%	November 20, 8 am
Assignment 3: Oral presentation of your own research project	40%	December 17, 8am

GENERAL READINGS

- Wickham, H. and G. Grolemund. 2017. R for Data Science: Import, Tidy, Transform, Visualize, and Model Data. O'Reilly.
- Silge, Julia, and David Robinson. 2017. Text mining with R: A tidy approach. O'Reilly.
- quanteda R package

RECAP R & RSTUDIO

R

- Based on the statistical programming language S (1976)
- R was developed by Ross Ihaka and Robert Gentleman (1995)
- R was intentionally developed to be a data analysis language

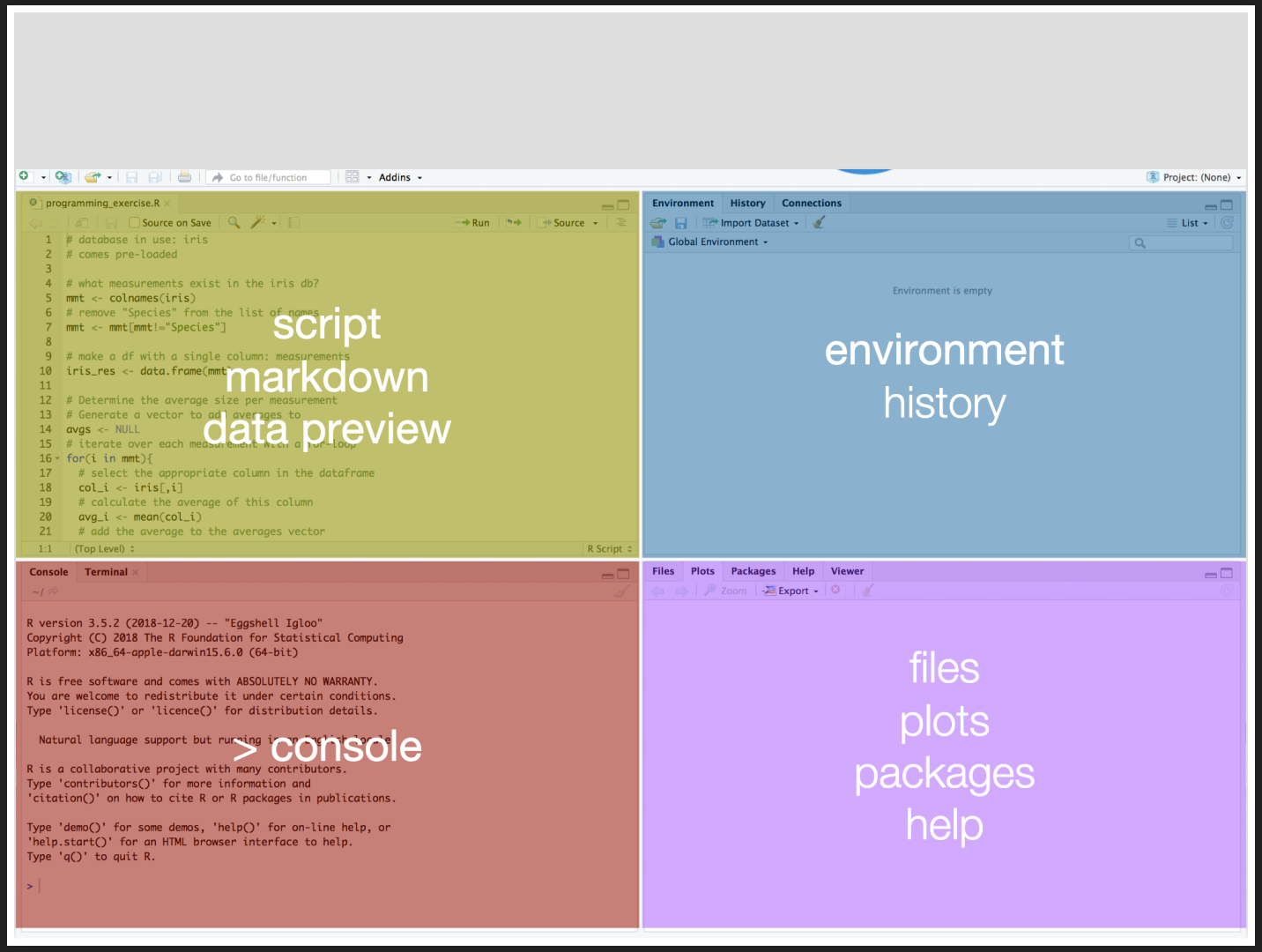
WHY R

- Open source: makes it highly customizable and easily extensible
- Over 18,000 packages and counting
- Used by many social scientists interested in data analysis
- Powerful tool to generate elegant and effective plots
- Command-line interface and scripts favors reproducibility
- Excellent documentation and online help resources

WE WILL WORK IN RSTUDIO

- RStudio is an Integrated Developer Environment (IDE) and serves as:
  - Code editor
  - Project manager
  - Workspace viewer
  - Data browser
  - Enhanced output viewer
  - Help browser

THE RSTUDIO INTERFACE



BASIC WORKFLOW

- Edit in code editor (.r-file)
- Paste to console
- Save Workspace/Datasets (.Rdata-file)
- Save code routinely (no auto-save!)
- Press TAB to use RStudio’s autocompletion feature

SHORTCUTS

- Run code from editor: Select line and ctrl+Enter
- Switch between source and console: ctrl+1, ctrl+2
- Clear console: ctrl+L
- ‘Arrow up’ gives you the last line of code in the console
- Press Alt+Shift+K to see all keyboard shortcuts

SETUP

SOFTWARE

- Download and install latest versions of R (4.1.1) and RStudio (1.4.1717)
- Download and install Git
  - for Mac OS: `brew install git`

PACKAGES

- Install the following packages:
  - *tidyverse*
  - *quanteda*

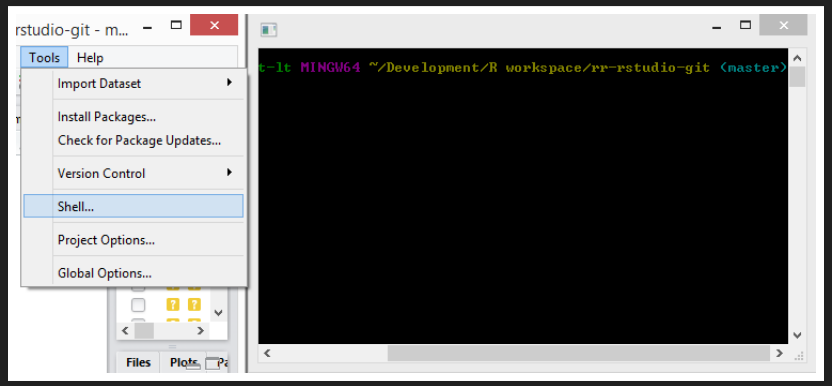
SETUP GITHUB

1. Create an account on GitHub
2. Open RStudio and go to *Tools* > *Global Options…* click on *Git/SVN*
3. Check *Enable version control interface for RStudio projects*
4. Set the path to the Git executable that you just installed. Open a shell, if you don't know where Git is installed. Windows: type `where git` and hit enter. The path should be something like:
   `C:/Program Files (x86)/Git/bin/git.exe`
   Linux/OS X: type `which git` and hit enter. The path should be something like: `/usr/bin/git`
5. Restart RStudio, if it worked out you will find the Git icon on the top toolbar.

SETUP GITHUB CONT.

6. Configure Git and set your *user name* and *email* (the email address you used to register on GitHub). You can directly open the Git prompt from within RStudio. User name and email needs to be set only once. Go to *Tools* > *Shell* to open the Git Shell to tell Git your username and GitHub email.

```
git config --global user.name 'yourGitHubUsername'
git config --global user.email 'name@provider.com'
```



CLONE A GITHUB REPO

7. Copy the repository HTTPS urlfrom the text as data repository located on https://github.com/mhaber/Text_as_Data_21. Click on *Code* and copy the HTTPS link of the project by clicking the little icon to the right of the URL.
8. In RStudio Select *File* > *New Project..*, select *Version Control*, Choose *Git*, then provide the repository HTTPS link, select the R workspace folder and create the project. RStudio now clones the content of the repository to your project folder. The content of the GitHub repository should now appear in the Files pane of RStudio.

CREATE YOUR OWN GITHUB REPO

Task: Create your own GitHub repository for this course with an appropriate *README.md* file and clone the repository in RStudio. Then create an *.R* script file with a function to print *Hello World!* and commit and push the new file to your GitHub repository.

WRAPPING UP

QUESTIONS?

OUTLOOK FOR OUR NEXT SESSION

- Next week we will learn how to import textual data into R
- We will again meet in room 2.32

THAT’S IT FOR TODAY

Thanks for your attention!