HTW Berlin – GOR 2020

A Primer to Web Scraping with R

Introduction

Simon Munzert
Hertie School, Berlin

March 11, 2020

Welcome

Welcome!

Thank you for your interest in the course!

I'm glad that I will have the opportunity to give you a gentle introduction to web scraping with R.

Just a few words on my personal background:

- Assistant Professor in Data Science and Public Policy at the Hertie School
- Political scientist by training
- Working with web-based data since about 2010
- What I do with web data: measure public awareness, news consumption, political behavior

Goals and outline

Goals

After attending this course, ...

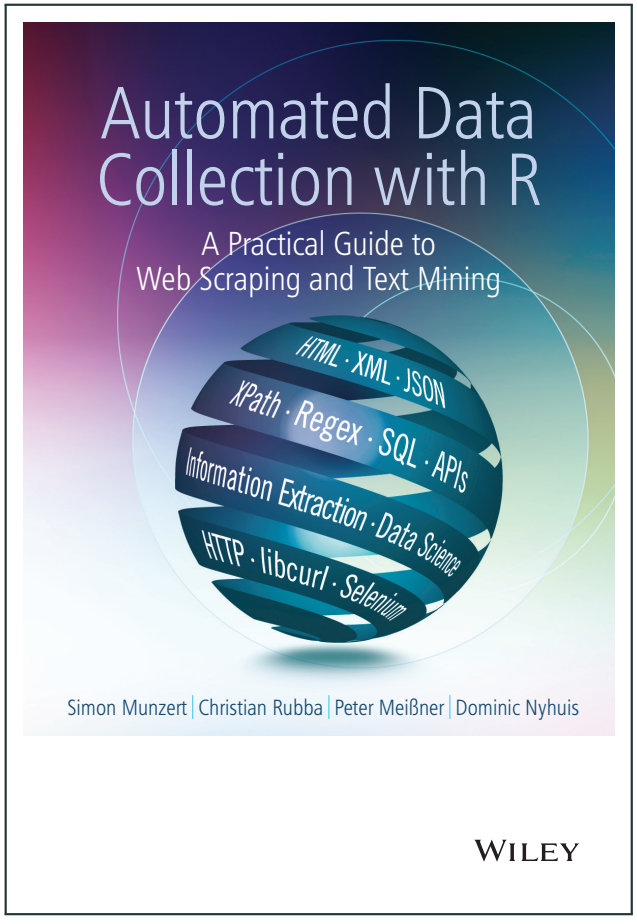
- you have acquired basic knowledge of web technologies
- you are able to scrape information from static and dynamic websites using R
- you are able to access web services (APIs) with R
- you can build up and maintain your own original sets of web-based data

Course outline

Unit	Topic
1	Introduction and setup
2	Regular expressions
3	Scraping static webpages using XPath
4	Tapping APIs

The accompanying book

- contains most of which I tell you during the course (but much more, and at times more accurate)
- written between 2012 and 2014 → not entirely up-to-date anymore—but the course material reflects the state of the art
- homepage with materials: www.r-datacollection.com
- if you find any errors in the book, please let me know!



Web scraping with R

Web scraping. What? Why?

Web scraping

A.k.a. screen scraping, is the business of

- getting (unstructured) data from the web and
- bringing it into shape (e.g., clean, make tabular format)

A data analyst's view

- data abundance online
- social interaction online
- services track social behavior

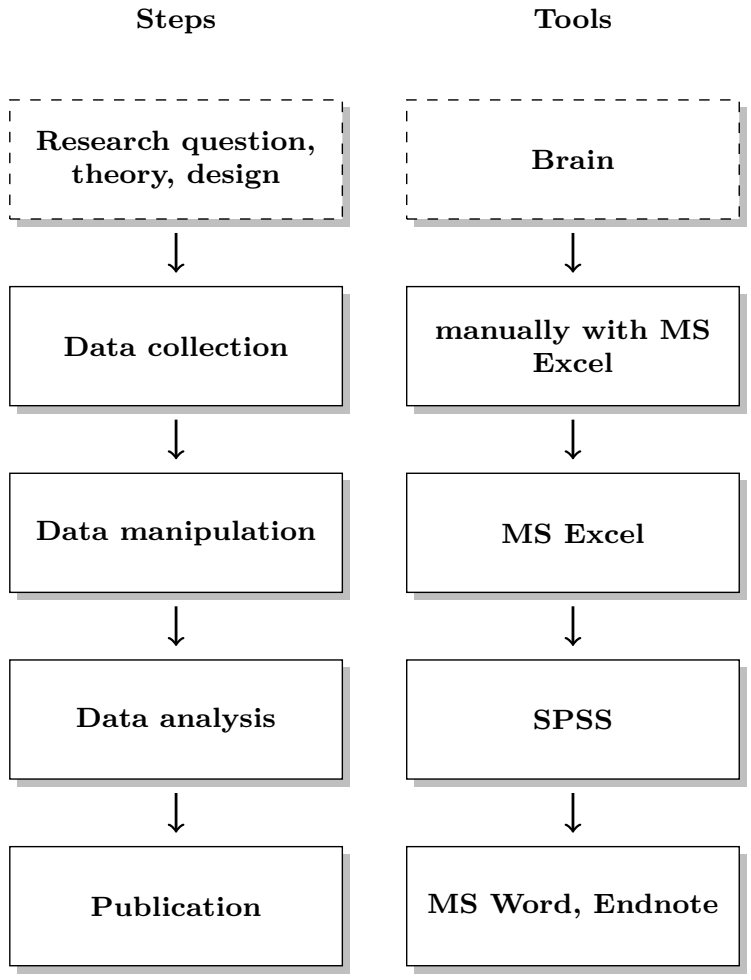
A pragmatist's view

- financial resources
- time resources
- reproducibility
- updateability

Why R?

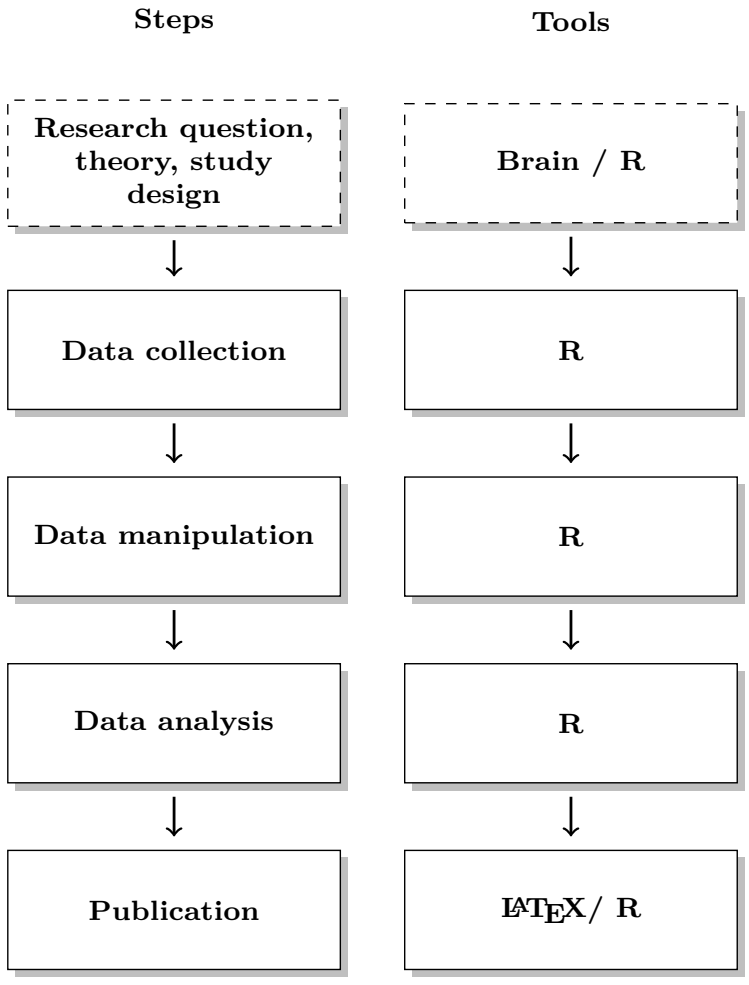
Why R?

- free
- open source
- large community
- powerful tools for statistical analysis
- powerful tools for visualization
- flexible in processing all kinds of data/languages
- useful in every step of the workflow



Why R?

- free
- open source
- large community
- powerful tools for statistical analysis
- powerful tools for visualization
- flexible in processing all kinds of data/languages
- useful in every step of the workflow

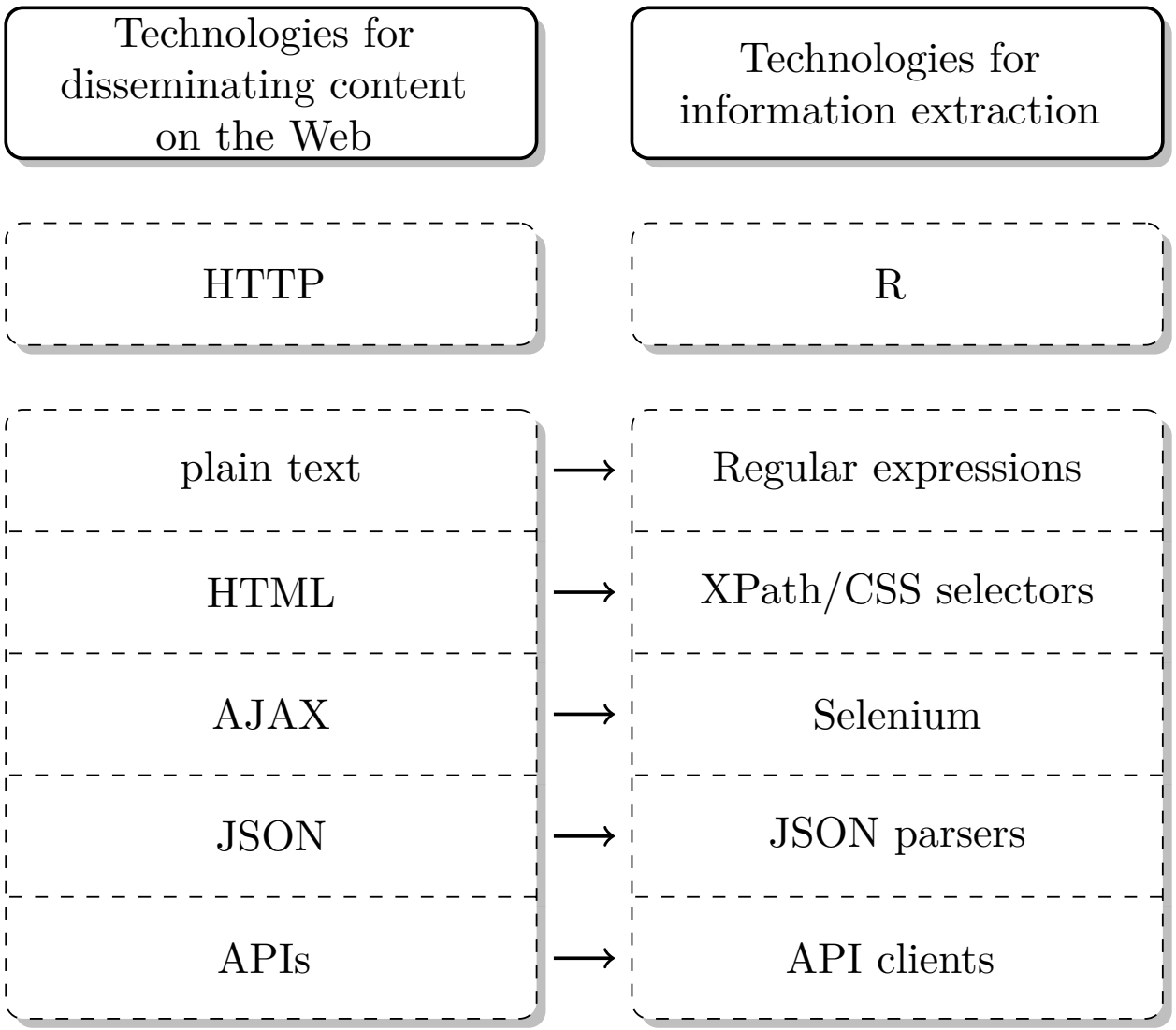


The philosophy behind web data collection with R

- no point-and-click procedure
- script the entire process from start to finish
- automation of
  - downloading
  - classical screen scraping
  - tapping APIs
  - parsing
  - data tidying, text data processing
- scaling up scraping procedures
- scheduling of scraping tasks

Technologies of the World Wide Web

Technologies of the World Wide Web



Technical setup

1. make sure that the newest version of R is installed on your computer (available here: https://cran.r-project.org/)
2. install the newest stable version of *RStudio Desktop* (available here: https://www.rstudio.com/products/rstudio/download)