A Primer to Web Scraping with R

Simon Munzert

Department of Social Sciences

Humboldt University of Berlin

simonmunzert.github.io

github.com/simonmunzert

@simonsaysnothin

Workshop outline

	Time	Topic
Slot 1	Thu, 10.00 - 11.30	Introduction and setup
Slot 2	Thu, 11.45 - 13.00	Regular expressions
Slot 3	Thu, 14.00 - 15.30	Scraping static webpages
Slot 4	Thu, 15.45 - 17.00	Advanced scraping of static webpages
Slot 5	Fri, 10.00 - 11.30	Scraping dynamic webpages
Slot 6	Fri, 11.45 - 13.00	Tapping APIs
Slot 7	Fri, 14.00 - 15.30	Gathering social media data
Slot 8	Fri, 15.45 - 17.00	Workflow and scraping etiquette

Who are you?

- ? name
- ? areas of interest
- ? experience in web scraping / ideas for applications
- ? knowledge on web technologies (HTML, XML, JSON, HTTP, ...)
- ? operating system (Windows, Mac OS, Linux)

Why web scraping?

Web 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

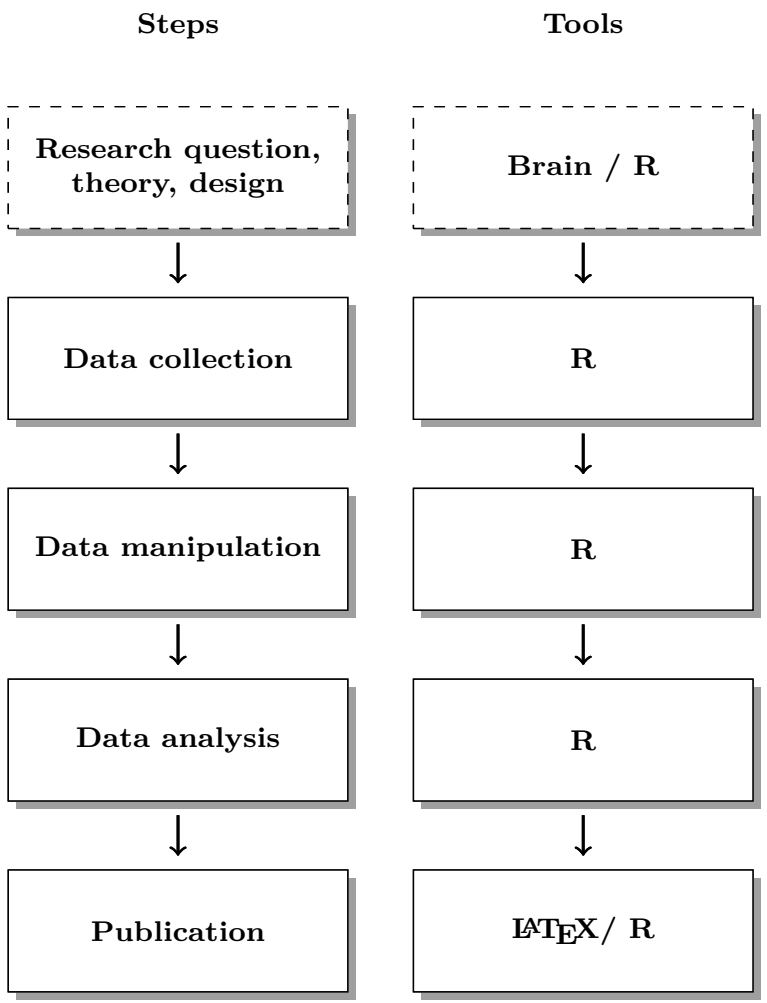
- data abundance online
- social interaction online
- services track social behavior
- online data meant for display, not download

A pragmatists‘s view

- financial resources
- time resources
- reproducibility
- updateability

Why R?

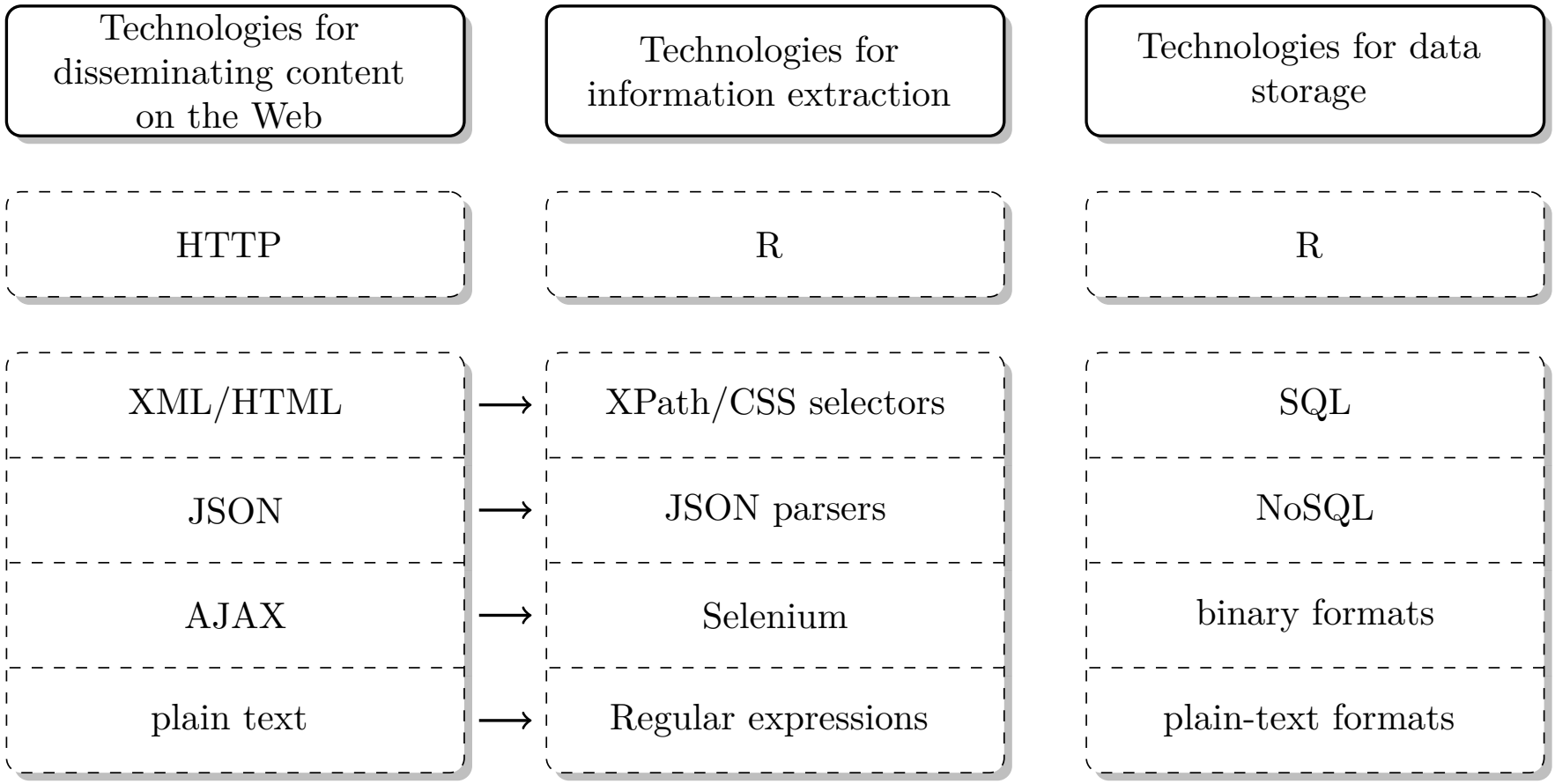
- free
- open source
- large community
- powerful for statistical analysis
- powerful for visualization
- flexible in processing all kinds of data types
- useful in every step of the workflow



The philosophy behind web data collection with R

- no point-and-click procedure
- automation of download, parsing, and data extraction
- classical screen scraping
- tapping of web services and APIs
- post-processing of text data
- reproducibility

Web technologies



Midterm goals to become a seasoned web scraper (takes 3-5 projects)

1. Learn to understand and construct regular expressions

   http://stat545.com/block022_regular-expression.html

2. Learn to build CSS selectors (or XPath expressions) by hand

   http://flukeout.github.io/

3. Integrate more packages in your scraping toolbox

   https://cran.r-project.org/ web/views/WebTechnologies.html

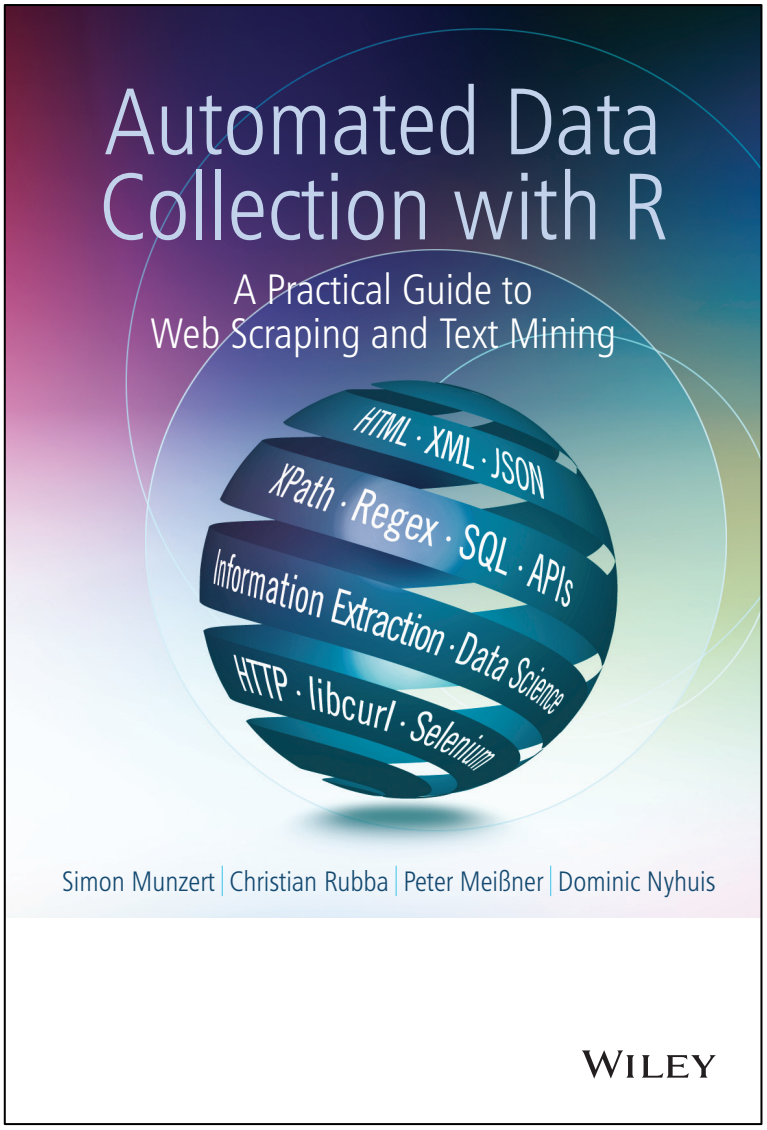
An accompanying book

- covers the entire scraping workflow in R and:
  - fundamentals of web technologies (HTTP, HTML, XML, JSON, XPath)
  - regular expressions
  - text mining
  - much more

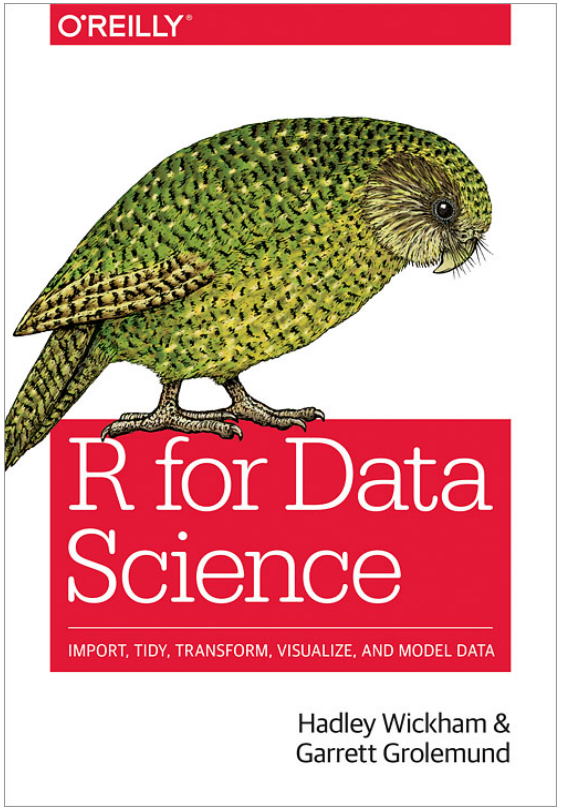
- published in late 2014 → not entirely up-to-date anymore (we‘re working on 2nd ed.)

- homepage with materials:

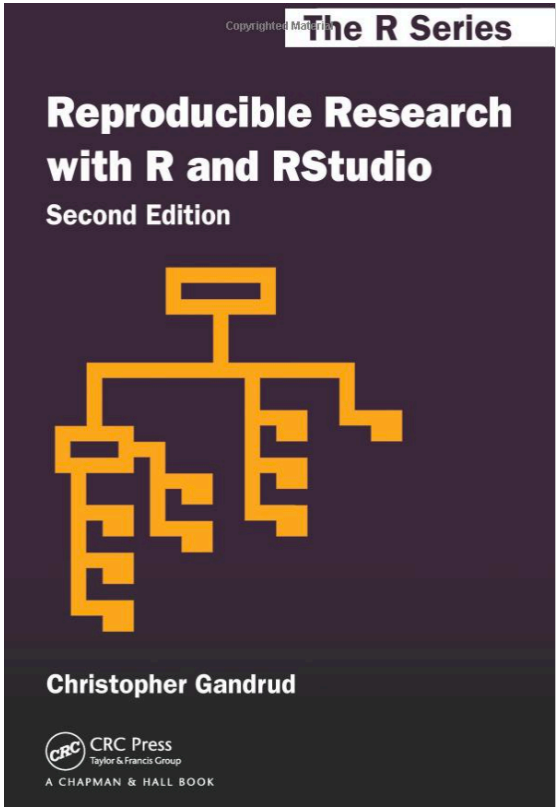
  r-datacollection.com



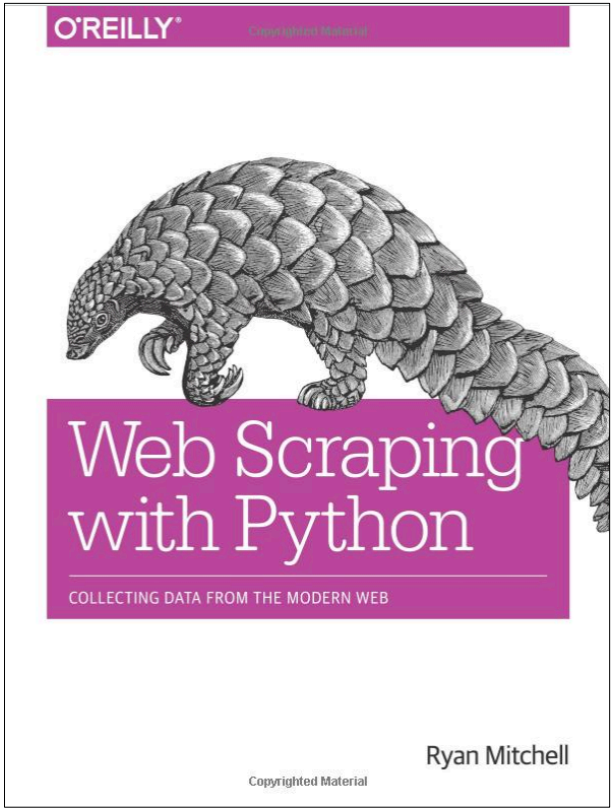
Other useful books on the market



- modern intro to R
- available for free at: http://r4ds.had.co.nz/



- helpful at establishing robust (scraping) workflow



- web scraping works with other programming languages, too!

Materials:

https://github.com/simonmunzert/rscraping-eui-2017

Book:

r-datacollection.com
@RDataCollection

Me:

simonmunzert.github.io
github.com/simonmunzert
@simonsaysnothin