

A Primer to Web Scraping with R

Simon Munzert

Department of Social Sciences
Humboldt University of Berlin

simonmunzert.github.io
github.com/simonmunzert
[@simonsaysnothin](https://twitter.com/simonsaysnothin)

Slides and Materials: <https://git.io/vyAde>

Overview

1. Introductory example
2. Why web scraping?
3. Web scraping with R
4. Tapping APIs with R
5. Outlook and helpful resources

Introductory example

A word cloud centered around the word "data". The word "data" is the largest and most prominent word, colored red. Other words are arranged around it, some overlapping, in various colors including green, blue, and orange. The words represent concepts related to data, such as technologies, sources, quality, index, margin, note, text, ref, web, ajax, might, scraping, provide, part, collection, one, xml, chapter, html, use, example, information, json, techniques, documents, can, book, and general.

data technologies
source sources
quality index margin note text
ref web ajax might scraping
provide part collection one
xml chapter html use example
information json techniques
documents can book general

Data on the web

W Berlin - Wikipedia Simon

Sicher <https://en.wikipedia.org/wiki/Berlin>

Not logged in Talk Contributions Create account Log in

Article Talk Read Edit View history Search Wikipedia

Berlin

From Wikipedia, the free encyclopedia

This article is about the capital of Germany. For other uses, see Berlin (disambiguation).

Berlin (/ber'lin/, German: [be̥rl̩ɪn] (listen)) is the capital and the largest city of Germany as well as one of its constituent 16 states. With a population of approximately 3.5 million people,^[4] Berlin is the second most populous city proper and the seventh most populous urban area in the European Union.^[5] Located in northeastern Germany on the banks of rivers Spree and Havel, it is the centre of the Berlin-Brandenburg Metropolitan Region, which has about 6 million residents from more than 180 nations.^{[6][7][8][9]} Due to its location in the European Plain, Berlin is influenced by a temperate seasonal climate. Around one-third of the city's area is composed of forests, parks, gardens, rivers and lakes.^[10]

First documented in the 13th century and situated at the crossing of two important historic trade routes,^[11] Berlin became the capital of the Margraviate of Brandenburg (1417–1701), the Kingdom of Prussia (1701–1918), the German Empire (1871–1918), the Weimar Republic (1919–1933) and the Third Reich (1933–1945).^[12] Berlin in the 1920s was the third largest municipality in the world.^[13] After World War II and its consequent occupation by the victorious countries, the city was divided; East Berlin became the capital of East Germany while West Berlin became a de facto West German enclave, surrounded by the Berlin Wall (1961–1989). Following German reunification in 1990, Berlin once again became the capital of统一的Germany.

Berlin
State of Germany



https://en.wikipedia.org/wiki/File:Alte_Nationalgalerie_Berlin,_2011.jpg

Data on the web

W Berlin - Wikipedia Simon

← → ⌛ ⌂ ⌃ Sicher https://en.wikipedia.org/wiki/Berlin

Borough mayors make up the council of mayors (*rat der Bürgermeister*), which is led by the city's Governing Mayor and advises the Senate. The neighborhoods have no local government bodies.

Twin towns – sister cities [edit]

See also: *List of twin towns and sister cities in Germany*

Berlin maintains official partnerships with 17 cities.^[100] Town twinning between Berlin and other cities began with its sister city Los Angeles in 1967. East Berlin's partnerships were canceled at the time of German reunification but later partially reestablished. West Berlin's partnerships had previously been restricted to the borough level. During the Cold War era, the partnerships had reflected the different power blocs, with West Berlin partnering with capitals in the Western World, and East Berlin mostly partnering with cities from the Warsaw Pact and its allies.

There are several joint projects with many other cities, such as Beirut, Belgrade, São Paulo, Copenhagen, Helsinki, Johannesburg, Mumbai, Oslo, Shanghai, Seoul, Sofia, Sydney, New York City and Vienna. Berlin participates in international city associations such as the Union of the Capitals of the European Union, Eurocities, Network of European Cities of Culture, Metropolis, Summit Conference of the World's Major Cities, and Conference of the World's Capital Cities. Berlin's official sister cities are:^[100]

- 1967  Los Angeles, United States
- 1987  Paris, France
- 1988  Madrid, Spain
- 1989  Istanbul, Turkey
- 1991  Warsaw, Poland^[101]
- 1991  Moscow, Russia
- 1992  Brussels, Belgium
- 1992  Budapest, Hungary^[102]
- 1993  Tashkent, Uzbekistan
- 1993  Mexico City, Mexico
- 1993  Jakarta, Indonesia
- 1994  Beijing, China
- 1994  Tokyo, Japan
- 1994  Buenos Aires, Argentina
- 1995  Prague, Czech Republic^[103]
- 2000  Windhoek, Namibia
- 2000  London, United Kingdom

Capital city [edit]

Berlin is the capital of the Federal Republic of Germany. The [President of Germany](#), whose functions are mainly ceremonial under the [German constitution](#), has his official residence in [Schloss Bellevue](#).^[104] Berlin is the seat of the [German executive](#), housed in the [Chancellery](#), the [Bundeskanzleramt](#). Facing the Chancellery is the [Bundestag](#), the German Parliament, housed in the renovated [Reichstag building](#) since the government relocated to Berlin in 1998. The [Bundesrat](#) ("federal council", performing the function of an upper house) is the representation of the Federal States (*Bundesländer*) of Germany and has its

Let's grab these data!

```
# load packages
library(rvest)
library(stringr)

# scrape page source
cities_string <- read_html("https://en.wikipedia.org/wiki/Berlin") %>%
# extract list items
  html_nodes(css = ".column-count-3 li") %>%
  html_text()
cities_string
```

original URL

CSS selector

```
[1] "1967 Los Angeles, United States"  "1987 Paris, France"
[3] "1988 Madrid, Spain"                  "1989 Istanbul, Turkey"
[5] "1991 Warsaw, Poland[101]"           "1991 Moscow, Russia"
[7] "1992 Brussels, Belgium"              "1992 Budapest, Hungary[102]"
[9] "1993 Tashkent, Uzbekistan"          "1993 Mexico City, Mexico"
[11] "1993 Jakarta, Indonesia"            "1994 Beijing, China"
[13] "1994 Tokyo, Japan"                  "1994 Buenos Aires, Argentina"
[15] "1995 Prague, Czech Republic[103]"   "2000 Windhoek, Namibia"
[17] "2000 London, United Kingdom"
```

Why was this so easy?

view HTML source code

right-click,
„inspect element...“

view HTML source code

W Berlin - Wikipedia view-source:https://en.wikipedia.org/w/index.php?title=Berlin&oldid=92295230052

Sicher https://en.wikipedia.org/wiki/Berlin

the Union of the Capitals of the European Union, Eurocities, Network of European Cities of Culture, Metropolis, Summit Conference of the World's Major Cities, and Conference of the World's Capital Cities.

Berlin | 82.94 × 88 other cities are:^[100]

- 1967 Los Angeles, United States
- 1992 Brussels, Belgium
- 1994 Tokyo, Japan
- 1992 Budapest, Hungary^[102]
- 1993 Buenos Aires, Argentina
- 1995 Prague, Czech Republic^[103]
- 1993 Mexico City, Mexico
- 2000 Windhoek, Namibia
- 1993 Jakarta, Indonesia
- 2000 London, United Kingdom
- 1994 Beijing, China
- 1987 Paris, France
- 1988 Madrid, Spain
- 1989 Istanbul, Turkey
- 1991 Warsaw, Poland^[101]
- 1991 Moscow, Russia

Capital city [edit]

Berlin is the capital of the Federal Republic of Germany. The President of Germany, whose functions are mainly ceremonial under the German constitution,

Elements Console Sources Network Timeline Profiler

<div class="div-col columns column-count column-count-3" style="--moz-column-count: 3; --webkit-column-count: 3; column-count: 3;>

 1967 ... Los Angeles, United States

 1994 ... Paris, France

 1992 ... Budapest, Hungary

 1993 ... Tashkent, Uzbekistan

 1995 ... Prague, Czech Republic

 1993 ... Mexico City, Mexico

 2000 ... Windhoek, Namibia

 1991 ... Jakarta, Indonesia

 2000 ... London, United Kingdom

 1994 ... Beijing, China

html body #content #mw-content-text ul

Styles Event Listeners DOM Breakpoints Properties

Filter :hover .cls +

element.style { }

div.columns dl, _load.php?debug=&skin=vector:1 div.columns ol, div.columns ul { margin-top: 0; }

.mw-content-ltr ul, _load.php?debug=_terlanguage:_1 .mw-content-rtl .mw-content-ltr ul { margin: 0.3em 0 1.6em; padding: 0; }

ul { _load.php?debug=_terlanguage:_1 list-style-type: disc; list-style-image: url(data:image/svg+xml,%3C%3Fxml%20version%3D8%22%20%3D%22%20r%3D222.5%22%20fill%3D%22%230052 list-style-image: }

margin border padding 22.400 82.938 × 1229.470

Filter Show all

color ► **rgb(3...**
direction ► ltr
display ► block
font-family ► sans-se.
font-size ► 14px

Let's clean up these data!

```
# remove footnotes ([101], [102], ...)
cities_string <- str_replace(cities_string, "\\[\\d+\\]", "")

# extract data
year <- str_extract(cities_string, "\\d{4}")
city <- str_extract(cities_string, "[[:alpha:] ]+") %>% str_trim
country <- str_extract(cities_string, "[[:alpha:] ]+\$") %>% str_trim
```

regular expression

```
[1] "1967 Los Angeles, United States"    "1987 Paris, France"
[3] "1988 Madrid, Spain"                  "1989 Istanbul, Turkey"
[5] "1991 Warsaw, Poland"                 "1991 Moscow, Russia"
[7] "1992 Brussels, Belgium"               "1992 Budapest, Hungary"
[9] "1993 Tashkent, Uzbekistan"            "1993 Mexico City, Mexico"
[11] "1993 Jakarta, Indonesia"              "1994 Beijing, China"
[13] "1994 Tokyo, Japan"                   "1994 Buenos Aires, Argentina"
[15] "1995 Prague, Czech Republic"         "2000 Windhoek, Namibia"
[17] "2000 London, United Kingdom"
```

Let's clean up these data!

```
# remove footnotes ([101], [102], ...)
cities_string <- str_replace(cities_string, "\\[\\d+\\]", "")

# extract data
year <- str_extract(cities_string, "\\d{4}")
city <- str_extract(cities_string, "[[:alpha:] ]+") %>% str_trim
country <- str_extract(cities_string, "[[:alpha:] ]+\$") %>% str_trim
```

regular expression

```
year[1:5]
[1] "1967" "1987" "1988" "1989" "1991"
```

```
city[1:5]
[1] "Los Angeles"    "Paris"          "Madrid"        "Istanbul"      "Warsaw"
```

```
country[1:5]
[1] "United States"  "France"        "Spain"         "Turkey"        "Poland"
```

Let's clean up these data!

```
# remove footnotes ([101], [102], ...)
cities_string <- str_replace(cities_string, "\\[\\d+\\]", "")

# extract data
year <- str_extract(cities_string, "\\d{4}")
city <- str_extract(cities_string, "[[:alpha:]]+") %>% str_trim
country <- str_extract(cities_string, "[[:alpha:]]+\$") %>% str_trim

# put everything into data.frame
cities_df <- data.frame(year, city, country)
head(cities_df)
```

regular
expression

	year	city	country
1	1967	Los Angeles	United States
2	1987	Paris	France
3	1988	Madrid	Spain
4	1989	Istanbul	Turkey
5	1991	Warsaw	Poland
6	1991	Moscow	Russia

Let's map these data!

```
# geocode cities
library(ggmap)
cities_coords <- geocode(paste0(cities_df$city, ", ", cities_df$country))
cities_df$lon <- cities_coords$lon
cities_df$lat <- cities_coords$lat
```

query Google Maps API


```
# plot world map, add coordinates
map_world <- borders("world", colour = "gray50", fill = "white")
ggplot() + map_world + geom_point(aes(x = cities_df$lon, y = cities_df$lat),
color = "red", size = 1) + theme_void()
```



Why web scraping with R?

A word cloud centered around the words "scraping" and "web". The words are in various sizes and colors (black, blue, green, red). Other visible words include "information", "site", "also", "common", "law", "browser", "pages", "states", "the", "use", "sites", "using", "access", "related", "websites", "this", "court", "bots", "case", "may", "terms", "software", "data", "can", "human", "content", "farechase", and "computer".

Why web scraping?

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

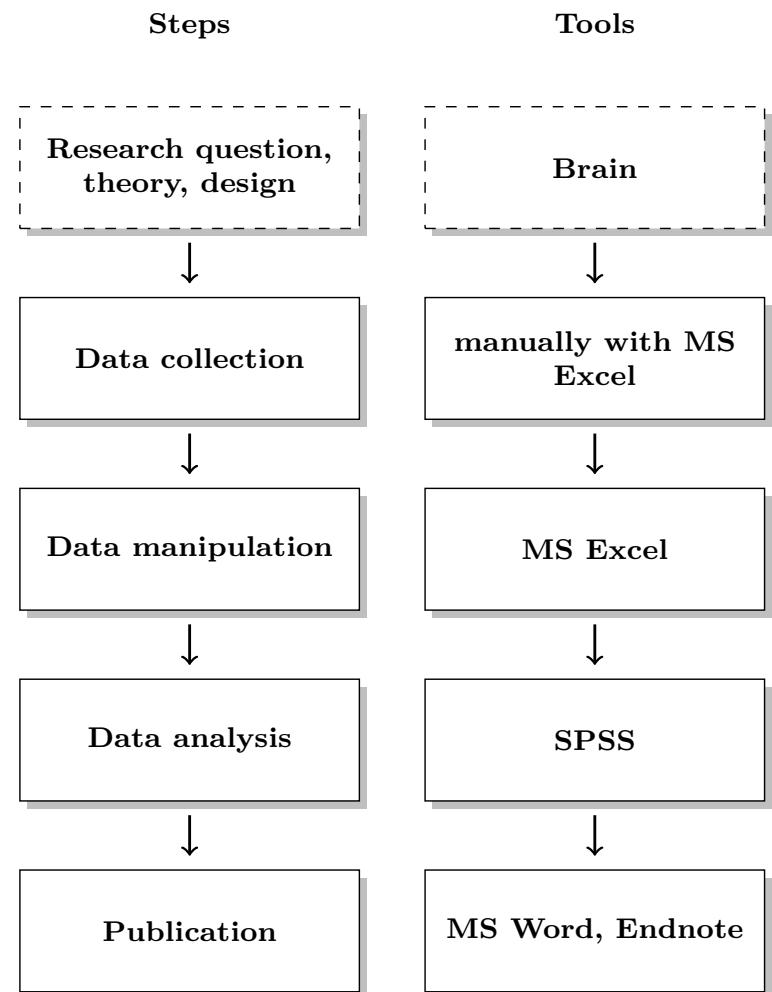
- financial resources
- time resources
- reproducibility
- updateability

Web scraping is the business of

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

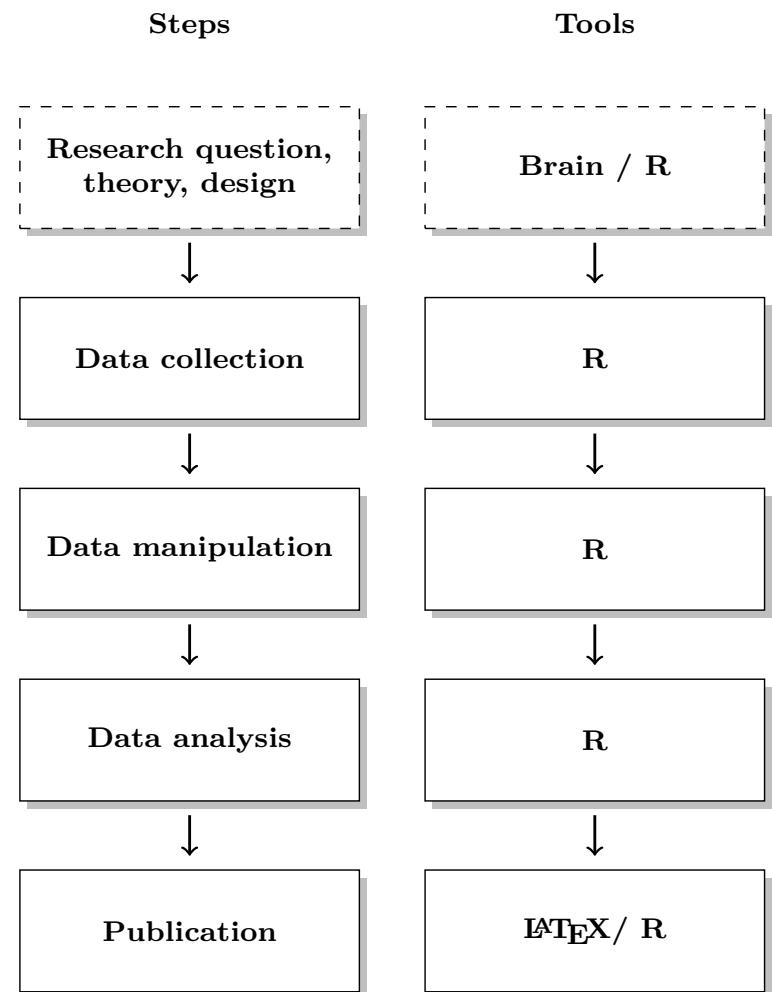
Why R?

- 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



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



Web scraping with R at a glance

A word cloud centered around the word "request". Other prominent words include "response", "resource", "web", "client message", "server", and "methods". Smaller words surrounding the center include "data", "tcp", "html", "example", "connection", "status", "header", "user information", "information", "line", "content", "also", "method", "protocol", and "requests".

response resource
data web request
tcp client message rfc
the can get html example connection
servers status header user information
may line content also method
methods server
protocol requests

A beginner’s toolbox (also helpful to experts)

- **rvest**: easy web scraping (“harvesting”) suite
(by Hadley Wickham)
- **xml2**: engine to parse XML-style files
(by Hadley Wickham and James Hester)
- **stringr**: tools for string processing
(by Hadley Wickham)

Scraping HTML tables

Wikipedia The Free Encyclopedia

Main page Contents Featured content Current events Random article Donate to Wikipedia Wikipedia store

Interaction Help About Wikipedia Community portal Recent changes Contact page

Tools What links here Related changes Upload file Special pages Permanent link Page information Wikidata item Cite this page

Print/export Create a book Download as PDF Printable version

Languages Català Čeština Dansk ★ Deutsch Español Italiano Lëtzebuergesch Lietuvių Nederlands 日本語 Norsk bokmål

W List of human spaceflights - W Sicher https://en.wikipedia.org/wiki/List_of_human_spaceflights Simon

List of human spaceflights

From Wikipedia, the free encyclopedia

For a list of spaceflights with human crews organised by program, see [List of human spaceflights by program](#). For a list of spaceports with achieved launches of humans to space, see [Spaceport](#).

These chronological lists include all crewed spaceflights that reached an altitude of at least 100 km (the FAI definition of spaceflight, see [Kármán line](#)), or were launched with that intention but failed. The USA has adopted a slightly different definition of spaceflight, requiring an altitude of only 50 miles (80 km). During the 1960s, 13 flights of the US [X-15 rocket plane](#) met the US criteria, but only two met the FAI's. These lists include only the latter two flights; see the [list of highest X-15 flights](#) for all 13. As of 29 January 2017, there have been 314 manned spaceflights that reached 100 km or more in altitude (316 including two failed attempts), 8 of which were [sub-orbital spaceflights](#).

To date, there have been four [fatal missions](#) in which 18(19 if you count the US Air Force Definition) astronauts died.

Contents [hide]

- 1 Summary
- 2 Detailed lists
- 3 Timeline
- 4 See also

Summary [edit]

	Russia USSR	United States	China	Total
1961–1970	16	25		41
1971–1980	30	8		38
1981–1990	*25	*38		*63
1991–2000	20	63		83
2001–2010	24	34	3	61
2011–2020	24	3	3	30
Total	*139	*171	6	*316

*Includes the two failed launches of [STS-51-L](#) and [Soyuz T-10-1](#).



Apollo 7 heads into orbit with its crew of three, 1968

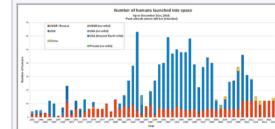


Chart of all humans launched into space as of December the 31st, 2016, Including unsuccessful launches (STS-51-L and Soyuz T-10-1).

Scraping HTML tables

- the [web developer tools](#) and [source code inspector](#) are your friends (remember: right-click, “inspect element...”)

The screenshot shows a web browser window displaying a table from Wikipedia about human spaceflights. The table includes columns for Country/USSR, Total number of flights, and the total count. The developer tools panel is open, highlighting the table's source code with a red box. The code uses the class "wikitable" and a style attribute with "text-align:right;". Below the table, a note states: "Includes the two failed launches of STS-51-L and Soyuz T-10-1." The developer tools also show the element's bounding box dimensions: width 412px, height 238px, and a margin of 14px.

	Russia USSR	United States	China	Total
1961–1970	16	25		41
1971–1980	30	8		38
1981–1990	*25	*38		*63
1991–2000	20	63		83
2001–2010	24	34	3	61
2011–2020	24	3	3	30
Total	*139	*171	6	*316

*Includes the two failed launches of STS-51-L and Soyuz T-10-1.

Detailed lists [edit]

The [Salyut](#) series, [Skylab](#), [Mir](#), [ISS](#), and [Tiangong](#) series space stations, with which various of these flights docked in orbit, are not listed separately here. See the detailed lists (links above) for information.

Missions which were intended to reach space but which failed to do are listed in *italics*, and fatal missions are marked with asterisk.

1961	Vostok 1 — Mercury-Redstone 3 — Mercury-Redstone 4 — Vostok 2
1962	Mercury-Atlas 6 — Mercury-Atlas 7 — Vostok 3 — Vostok 4 — Mercury-Atlas 8
1963	Mercury-Atlas 9 — Vostok 5 — Vostok 6 — X-15 Flight 90 — X-15 Flight 91

Scraping HTML tables

- the [web developer tools](#) and [source code inspector](#) are your friends (remember: right-click, “inspect element...”)
- HTML tables use to share a common format

```
<table class="wikitable">
  <tr>
    <td></td>
    <td>Russia USSR</td>
    <td>United States</td>
    <td>China</td>
    <td>Total</td>
  </tr>
  <tr>
    <td>1961–1970</td>
    <td>16</td>
    <td>25</td>
    <td></td>
    <td>41</td>
  </tr>
  ...
</table>
```

Let's grab these data!

```
# load package
library(rvest)

# import webpage
url_p <- read_html("https://en.wikipedia.org/wiki/List_of_human_spaceflights")

# extract HTML tables
tables <- html_table(url_p, header = TRUE, fill = TRUE)
spaceflights <- tables[[1]]
spaceflights
```

	Russia	USSR	United States	China	Total	
1	1961-1970		16	25	NA	41
2	1971-1980		30	8	NA	38
3	1981-1990		*25	*38	NA	*63
4	1991-2000		20	63	NA	83
5	2001-2010		24	34	3	61
6	2011-2020		24	3	3	30
7	Total		*139	*171	6	*316

What if data do not come in nice HTML tables?

The New York Times - Breaking x Simon

← → ⌂ ⌂ Sicher https://www.nytimes.com

SUBSCRIBE NOW LOG IN ⚙

SECTIONS SEARCH ENGLISH 中文 (CHINESE) ESPAÑOL

The New York Times

Tuesday, March 21, 2017 | Today's Paper | Video | 54°F | DAX +0.02% ↑

World U.S. Politics N.Y. Business Opinion Tech Science Health Sports Arts Style Food Travel Magazine T Magazine Real Estate ALL

THE 45TH PRESIDENT

Fresh Worries on Russia for Trump's Weary Defenders

By GLENN THRUSH and MAGGIE HABERMAN 5:24 AM ET

• The obsessiveness and ferocity of President Trump's pushback against the Russia allegations, often untethered from fact, are making an uncertain situation worse.

• "The tweets make it much more difficult for us as we try to build a case against these leakers," said Representative Peter T. King, a New York Republican who sits on the

Support CO₂ limits on power plants

LESS → MORE

How Americans Think About Climate Change

Americans overwhelmingly believe that global warming is happening, and that carbon emissions should be scaled back. But fewer are sure that the changes will harm them.

The Opinion Pages

EDITORIALS

Comey's Haunting News on Trump and Russia

Possible election collusion makes a special prosecutor and a select congressional committee crucial.

Gorsuch Faces the Senate

Here's a good question for Judge Gorsuch: Why are you here?

OP-ED CONTRIBUTOR

How to Fix Health Care

By BENJAMIN DOMENECH

The plan to replace Obamacare has won few fans on the right or left. Here's a proposal to unite them: catastrophic coverage for all.

Brooks: The Unifying American Story

Leonhardt: All the President's Lies

'W.' and the Art of Redemption

Follow us on Twitter »

TIMES INSIDER »

Why Talk to a Reporter? Especially about Something Difficult and Personal?

THE CROSSWORD »

Play Today's Puzzle

Easy solution: SelectorGadget

- available at <http://selectorgadget.com/>, works as browser plugin
- helps you construct CSS selectors (or XPath expressions)
- with SelectorGadget, you can
 - click on elements you want to scrape
 - un-click elements you do not want to scrape
 - “point-and-click” the information you are interested in
- caveat: solves 80% of your standard scraping tasks – for the rest, learn how to build CSS selectors / XPath expressions

SelectorGadget

The New York Times - Breaking x Simon

Sicher https://www.nytimes.com

SECTIONS SEARCH ENGLISH 中文 (CHINESE) ESPAÑOL SUBSCRIBE NOW LOG IN :

select elements via point-and-click

THE 45TH PRESIDENT

FRESH Worries on Russia for Trump's Weary Defenders

LESS → MORE

Discuss global warming occasionally

How Americans Think About Climate Change

No valid path found.

The Opinion Pages

EDITORIALS

Comey's Haunting News on Trump and Russia

Possible election collusion makes a special prosecutor and a select congressional committee crucial.

GORSUCH Faces the Senate

Here's a good question for Judge Gorsuch: Why are you here?

U.N. Accepts Blame but Dodges the Bill in Haiti

OP-ED CONTRIBUTOR

How to Fix Health Care

By BENJAMIN DOMENECH

The plan to replace Obamacare has won few fans on the right or left. Here's a proposal to unite them: catastrophic coverage for all.

Brooks: The Unifying American Story

Leonhardt: All the President's Lies

'W.' and the Art of Redemption

Follow us on Twitter »

TIMES INSIDER »

Why Talk to a Reporter? Especially about Something

THE CROSSWORD »

Play Today's Puzzle

Clear Toggle Position XPath Help X

The screenshot shows a web browser window displaying the New York Times homepage. The SelectorGadget extension is active, as indicated by the red annotation. The main headline 'Fresh Worries on Russia for Trump's Weary Defenders' is highlighted with a yellow box and a red arrow points to it from the text 'select elements via point-and-click'. The page features a map of the United States where each county is colored according to its attitude towards discussing global warming. Below the map, the subtitle 'How Americans Think About Climate Change' is visible. To the right, there are columns for 'The Opinion Pages', 'Editorials', and 'Op-Ed Contributor'. At the bottom, there are links for 'Times Insider' and 'The Crossword'. A status bar at the bottom of the extension window says 'No valid path found.' and includes buttons for 'Clear', 'Toggle Position', 'XPath', 'Help', and 'X'.

SelectorGadget

The New York Times - Breaking x Simon

Sicher https://www.nytimes.com

SECTIONS SEARCH ENGLISH 中文 (CHINESE) ESPAÑOL SUBSCRIBE NOW LOG IN

The New York Times

Tuesday, March 21, 2017 | Today's Paper | Video | 54°F | FTSE 100 -0.16% ↓

World U.S. Politics N.Y. Business Opinion Tech Science Health Sports Arts Style Food Travel Magazine T Magazine Real Estate ALL

THE 45TH PRESIDENT

NEWS ANALYSIS **Fresh Worries on Russia for Trump's Weary Defenders**

By GLENN THRUSH and MAGGIE HABERMAN 5:24 AM ET

The obsessiveness and ferocity of President Trump's pushback against the Russia allegations, often untethered from fact, are making an uncertain situation worse.

"The tweets make it much more difficult for us as we try to build a case against these leakers," said Representative Peter T. King, a New York Republican who sits on the Intelligence Committee.

Support CO₂ limits on power plants

LESS → MORE

How Americans Think About Climate Change

Americans overwhelmingly believe that global warming is

The Opinion Pages

EDITORIALS

Comey's Haunting News on Trump and Russia

Possible election collusion makes a special prosecutor and a select congressional committee crucial.

Gorsuch Faces the Senate

Here's a good question for Judge Gorsuch: Why are you here?

OP-ED CONTRIBUTOR

How to Fix Health Care

By BENJAMIN DOMENECH

The plan to replace Obamacare has won few fans on the right or left. Here's a proposal to unite them: catastrophic coverage for all.

Brooks: The Unifying American Story

Leonhardt: All the President's Lies

'W.' and the Art of Redemption

Follow us on Twitter »

TIKES INSIDER »

Why Talk to a Reporter? Especially about Something

THE CROSSWORD »

Play Today's Puzzle

story-heading

Clear (145)

Toggle Position

XPath

Help

X

magic!

CSS selector # elements detected

Let's grab these data!

```
# load package
library(rvest)

# import webpage
url_p <- read_html("https://www.nytimes.com")

# look for nodes (parts of the HTML document that we want to extract)
headlines <- html_nodes(url_p, css = ".story-heading")

# extract the headline text
headlines_raw <- html_text(headlines)
head(headlines_raw)

[1] "Fresh Worries on Russia for Trump's Weary Defenders"
[2] "F.B.I. Confirms Inquiry on Trump Team's Russia Ties"
[3] "Takeaways From the Hearing "
[4] "\n                                Trump and the Russians: Links? No Links?"
[5] "G.O.P. Responds by Changing Subject"
[6] "U.S. Limits Devices on Foreign Airlines From 8 Countries"
```

Let's clean up these data!

```
# remove line breaks and empty spaces  
headlines_clean <- headlines_raw %>% str_replace_all("\\n", "") %>% str_trim()  
  
# how many headlines contain the word "Trump"?  
str_detect(headlines_clean, "Trump") %>% table()
```

FALSE	TRUE
133	12

Other scraping scenarios

- pull information out of lists
- geocode cities

Screenshot of a web browser showing the "Brauerei Liste Deutschlands Nr. 1" page from www.biermap24.de/brauereiliste.php. The page displays a list of breweries in Germany, categorized by state. A red arrow points from the list of cities on the right side of the slide towards this screenshot.

Brauerei Liste Deutschlands Nr. 1

Impressum Kontakt Home Wir über uns AddThis

Mein biermap24.de

Login Neu hier, gleich anmelden...

Biermenü

Startseite

Biergeschichtliches

Bierkarte

Brauereiliste (1025)

Biersorten Liste (7771)

alkoholfreies Bier (266)

Top 10 Bierliste

5,- EUR Biergutschein

Bier verdienen

Bier online bestellen

Biergeschenke

Gästebuch

unsere News

eine kleine Spende an uns - DANKE und PROST

Spenden

Brauereien in Deutschland

Baden-Württemberg (138) Bayern (462) Berlin (19) Brandenburg (17)
Bremen (4) Hamburg (4) Hessen (53) Mecklenburg-Vorpommern (13)
Niedersachsen (39) Nordrhein-Westfalen (109) Rheinland-Pfalz (36) Saarland (11)
Sachsen (48) Sachsen-Anhalt (17) Schleswig-Holstein (13) Thüringen (35)
alle Brauereien (1025)

1. Brauereigasthof Stadter Sachsendorf

2. Dampfbierbrauerei Zwiesel GmbH & Co.KG Zwiesel

3. Flensburger Gasthausbraueriebetriebs GmbH Flensburg

4. Mainzer Gasthausbrauerei GmbH Mainz

5. Stuttgarter Lokalbrauerei Calwer-Eck-Bräu Familie Breitmayer Stuttgart

A

6. A+V Brauhaus GmbH Bad Wildungen

7. Aachener Spezial-Biere Baesweiler

8. Aalener Löwenbräu Aalen

9. Accente Bremen GmbH Bremen

10. Adler Bräu GbR Inhaber Ramona Jentzsch-Volk & Robert Volk Wiernsheim

11. Adler-Bräu Stettfeld

12. Adlerbrauerei Göppingen Vertriebs GmbH Göppingen

13. Adlerbrauerei Götz Geislingen Steige

14. Adlerbrauerei Heribert Werner GmbH & Co KG Zuzenhausen

15. Ahornberger Landbrauerei Strössner-Bräu KG Konradsreuth

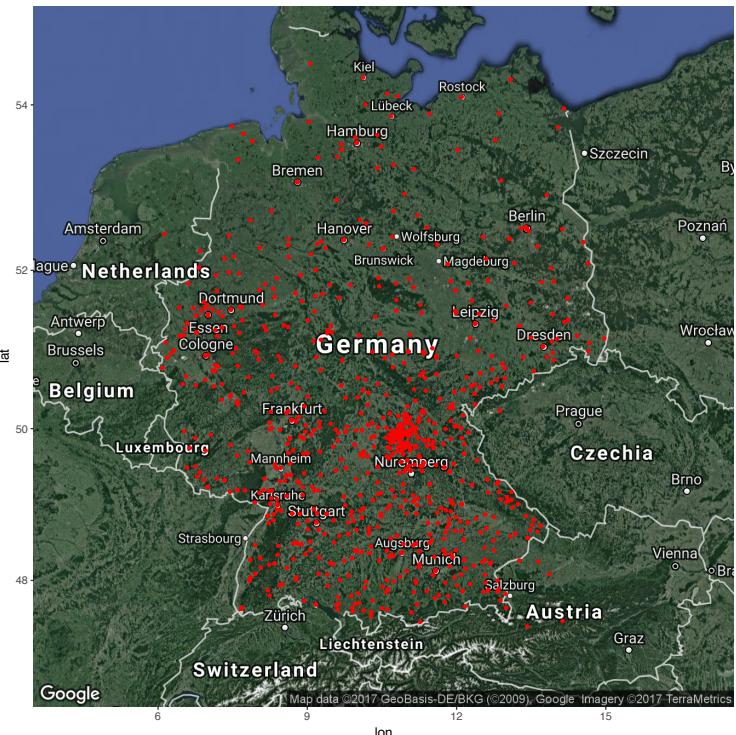
16. Aktienbrauerei Kaufbeuren AG Kaufbeuren

17. Aktiv- und Wanderhotel Adam-Bräu Bodenmais

`cities[1:15]`

```
[1] "Flensburg"
[3] "Stuttgart"
[5] "Baesweiler"
[7] "Bremen"
[9] "Stettfeld"
[11] "Geislingen Steige"
[13] "Konradsreuth"
[15] "Bodenmais"
```

"Mainz"
"Bad Wildungen"
"Aalen"
"Wiernsheim"
"Göppingen"
"Zuzenhausen"
"Kaufbeuren"



Other scraping scenarios

- download hundreds of (thousands of) HTML pages
- pull information from multiple websites
- identify links between HTML documents

The screenshot shows two browser tabs side-by-side. The left tab is titled "List of statisticians - Wikipedia" and displays the main content page for statisticians. The right tab is titled "Odd Aalen - Wikipedia" and displays the biography of Odd Aalen. A large red arrow points from the URL bar of the right tab to the URL bar of the left tab, illustrating how a web scraper might follow links between different pages on a website.

List of statisticians
From Wikipedia, the free encyclopedia
This list of **statisticians** lists people who have made notable contributions to the theories or application of **statistics**, or to the related fields of **probability** or **machine learning**. Also included are **actuaries** and **demographers**.

Contents :
A · B · C · D · E · F · G · H · I · J · K · L · M · N · O · P · Q · R · S · T · U · V

A [edit]
• Aalen, Odd Olai (1947–1987)
• Abbott, Edith (1876–1957)
• Abelson, Robert P. (1928–2005)
• Abramovitz, Moses (1912–2000)
• Achenwall, Gottfried (1719–1772)
• Adelstein, Abraham Manie (1916–1992)
• Ahsan, Riaz (1951–2008)

Odd Aalen
From Wikipedia, the free encyclopedia
Odd Olai Aalen (born 6 May 1947, in Oslo) is a Norwegian statistician and a professor at the Department of Biostatistics at the Institute of Basic Medical Sciences at the University of Oslo.^[1]

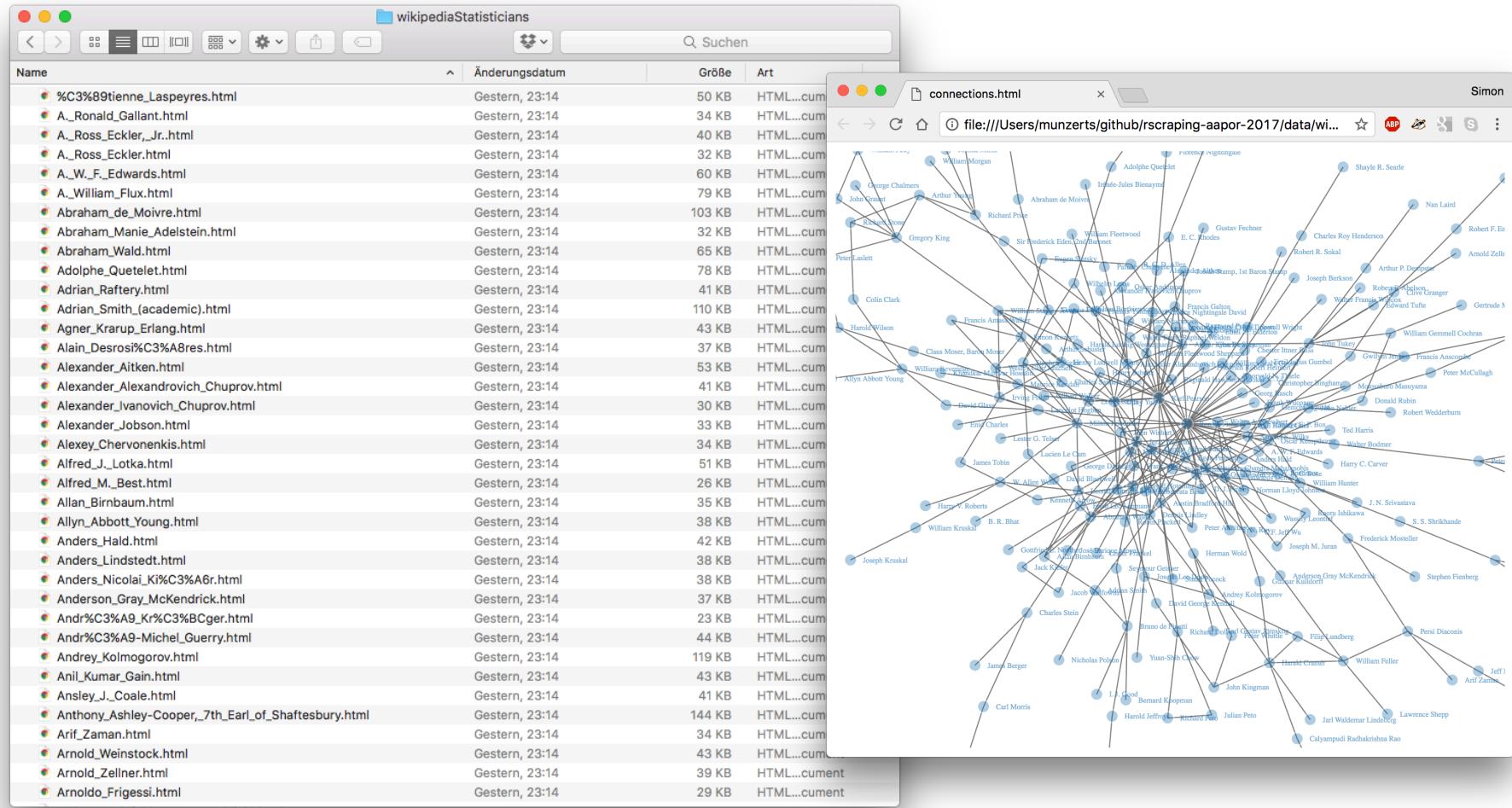
Contents [hide]
1 Life
2 Work
3 Honors and awards
4 References
5 External links

Life [edit]
Aalen completed his **examen artium** in 1966 at **Oslo Cathedral School** before studying first mathematics and physics and then statistics in which he graduated at the **University of Oslo** in 1972.^[2]

Work [edit]
His research work is geared towards applications in **biosciences**. Aalen's early work on **counting processes** and **martingales**, starting with his 1976 Ph.D. thesis at the **University of California, Berkeley**, has had profound influence in **biostatistics**. Inferences for fundamental quantities associated with cumulative hazard rates, in **survival analysis** and models for analysis of event histories, are typically based on the **Nelson–Aalen estimator** or appropriate related statistics. The **Nelson–Aalen estimator** is related to the **Kaplan–Meier estimator** and

Other scraping scenarios

- download hundreds of (thousands of) HTML pages
- pull information from multiple websites
- identify links between HTML documents



Other scraping scenarios

- gather data from dynamically rendered websites
- ...one of the more complex scraping scenarios

A screenshot of a web browser showing the 'IEA - Renewable Energy' advanced search page at www.iea.org/policiesandmeasures/renewableenergy/. The page has a sidebar on the left with sections for 'Countries', 'Policy Type', 'Renewable Energy Policy Target', 'Effective between', 'Size of Plant', and search options. A red arrow points to the 'Countries' section, which includes '+ Regions' and '+ Countries' buttons. The main content area shows 'Policy Type' filters like Economic Instruments, Information and Education, etc., and 'Renewable Energy Policy Target' filters for Bioenergy, Geothermal, Hydropower, etc. At the bottom are 'RESET' and 'SEARCH' buttons.

IEA - Renewable Energy

Simon

www.iea.org/policiesandmeasures/renewableenergy/

iea International Energy Agency

Advanced search click here

Countries

+ Regions
+ Countries

Policy Type

- + Economic Instruments
- + Information and Education
- + Policy Support
- + Regulatory Instruments
- + Research, Development and Deployment (R&D)
- + Voluntary Approaches

Renewable Energy Policy Target

- + Bioenergy
- + Geothermal
- Hydropower
- + Multiple Renewable Energy Sources
- + Ocean
- + Solar
- + Solar Thermal
- + Wind

Sector

- Electricity
- Framework Policy
- Heating and Cooling
- Multi-sectoral Policy
- Transport

Effective between

Select and Select

Jurisdiction

- International
- National
- State/Regional
- Municipal

Policy Status

- Ended
- In Force
- Planned
- Superseded
- Under Review

Size of Plant

Large
 Small

Search by keyword(s)

Search only recently updated policies

RESET **SEARCH**

Other scraping scenarios

- gather data from dynamically rendered websites
- ...one of the more complex scraping scenarios

The screenshot shows a web browser window for the International Energy Agency's (IEA) Renewable Energy policies and measures database. The URL in the address bar is www.iea.org/policiesandmeasures/renewableenergy/. A red arrow points from the address bar to the status bar at the bottom of the browser, which displays "URL does not change".

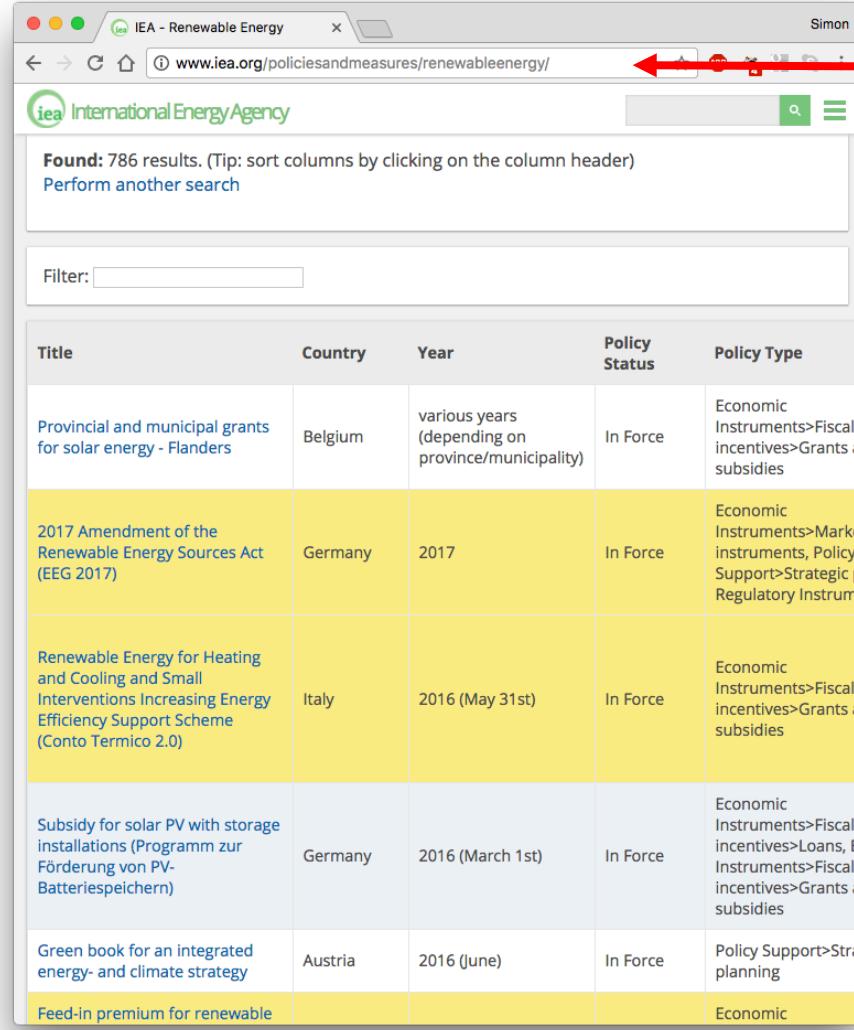
The page title is "Advanced search". The search interface includes several filter sections:

- Countries**: Includes checkboxes for Regions, OECD, Non-OECD, IEA, IRENA, EU (which is checked), Caribbean, Island countries, and a "+ Countries" link.
- Policy Type**: Includes checkboxes for Economic Instruments, Information and Education, Policy Support, Regulatory Instruments, Research, Development and Deployment (RD&D), and Voluntary Approaches.
- Renewable Energy Policy Target**: Includes checkboxes for Bioenergy, Geothermal, Hydropower, Multiple Renewable Energy Sources, Ocean, Solar, Solar Thermal, and Wind.
- Sector**: Includes checkboxes for Electricity, Framework Policy, Heating and Cooling, Multi-sectoral Policy, and Transport.
- Effective between**: Contains two dropdown menus labeled "Select" and "Select".
- Jurisdiction**: Includes checkboxes for International, National, State/Regional, and Municipal.
- Policy Status**: Includes checkboxes for Ended, In Force, Planned, Superseded, and Under Review.
- Size of Plant**: Includes checkboxes for Large and Small.
- Search by keyword(s)**: A text input field.
- Checkboxes at the bottom**: "Search only recently updated policies", "RESET" button, and a large blue "SEARCH" button.

A red arrow labeled "click here" points to the "EU" checkbox under the Countries section. Another red arrow points from the "EU" checkbox towards the "SEARCH" button at the bottom right.

Other scraping scenarios

- gather data from dynamically rendered websites
- ...one of the more complex scraping scenarios



Title	Country	Year	Policy Status	Policy Type
Provincial and municipal grants for solar energy - Flanders	Belgium	various years (depending on province/municipality)	In Force	Economic Instruments>Fiscal/incentives>Grants and subsidies
2017 Amendment of the Renewable Energy Sources Act (EEG 2017)	Germany	2017	In Force	Economic Instruments>Market instruments, Policy Support>Strategic policy instruments, Regulatory Instruments
Renewable Energy for Heating and Cooling and Small Interventions Increasing Energy Efficiency Support Scheme (Conto Termico 2.0)	Italy	2016 (May 31st)	In Force	Economic Instruments>Fiscal/incentives>Grants and subsidies
Subsidy for solar PV with storage installations (Programm zur Förderung von PV-Batteriespeichern)	Germany	2016 (March 1st)	In Force	Economic Instruments>Fiscal/incentives>Loans, Economic Instruments>Fiscal/incentives>Grants and subsidies
Green book for an integrated energy- and climate strategy	Austria	2016 (June)	In Force	Policy Support>Strategic planning
Feed-in premium for renewable				Economic

URL does not change

- make use of **RSelenium** together with Selenium Webdriver
- the basic idea: simulate activities in the browser with R, e.g. “click here”, “type something there”, and log the results using external software

Wrap-up

- the `rvest` package is your friend (plus `SelectorGadget` plus `regular expressions`)
- you don't have to become an HTML expert, but looking into a website's source code helps
- scraping data from static HTML tables can be as easy as this:

```
# load package
library(rvest)

# parse html
url_parsed <- read_html("https://www.example.com")

# extract table(s)
tables <- html_table(url_parsed, fill = TRUE)
your_table <- tables[[1]]

# tidy data
your_clean_table <- magic_data_tidier(your_table)
```

Wrap-up

- the `rvest` package is your friend (plus `SelectorGadget` plus `regular expressions`)
- you don't have to become an HTML expert, but looking into a website's source code helps
- scraping data from static HTML pages can be as easy as this:

```
# load package
library(rvest)

# parse html
url_parsed <- read_html("https://www.example.com")

# extract content you like with CSS selector
nodes <- html_nodes(url_parsed, css = "CSS selector")
your_data <- html_text(nodes)

# tidy data
your_clean_data <- magic_data_tidier(your_data)
```

Tapping APIs with R at a glance

javascript
format example xml the number web browsers also yaml code type use can schema this used eval types object request native language data standard using json

What are APIs?

- Application Programming Interface
- many web services provide APIs to access their data and services (Google, Twitter, Facebook, Wikipedia, ...)
- often works like a “data search engine”: you pose a request, the API answers with a bulk of data
- common data return formats: JSON, XML

APIs in the context of web scraping

- instant access to clean data
 - frees you from building manual scraping programs
- collecting data from the web using APIs provided by the data owner represents the **gold standard of web-based data retrieval**

Example: Google Maps API

- Google provides access to powerful location services
- free service (at least when used modestly)
- input/output: places, names, coordinates, meta information, maps, ...
- see also: <https://developers.google.com/maps/documentation/>

Potential use cases

- geocode observations based on address, city, post code, ...
- calculate distances between observations
- map observations

Example: Google Maps API

```
library(ggmap)  
geocode("Humboldt University, Berlin")
```

Source : <https://maps.googleapis.com/maps/api/geocode/json?address=Humboldt%20University%2C%20Berlin>

	lon	lat
1	13.39366	52.51788



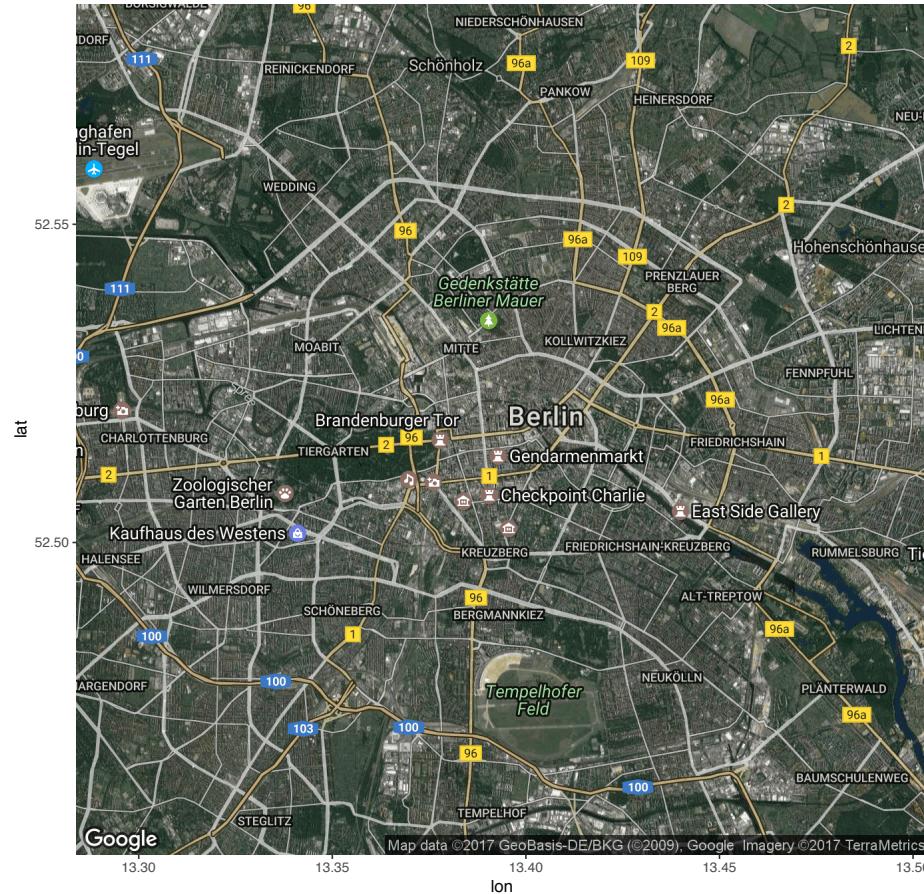
Excerpt from raw JSON

```
{  
  "address_components" : [  
    { "long_name" : "Germany",  
      "short_name" : "DE",  
      "types" : [ "country", "political" ]},  
    { "long_name" : "10099",  
      "short_name" : "10099",  
      "types" : [ "postal_code" ]}  
,  
  "formatted_address" : "Unter den Linden 6, 10099 Berlin, Germany",  
  "location" : { "lat" : 52.517883, "lng" : 13.3936551 },  
  "place_id" : "ChIJI0K8KtZNqEcRxSbyV3Fx5Kk",  
  "types" : [ "establishment", "point_of_interest", "university" ]  

```

Example: Google Maps API

```
library(ggmap)
geocode("Humboldt University, Berlin")
get_googlemap("Humboldt University, Berlin", zoom = 12, maptype = "hybrid") %>%
ggmap()
```



Example: Google Maps API

```
library(ggmap)
geocode("Humboldt University, Berlin")
get_googlemap("Humboldt University, Berlin", zoom = 12, maptype = "hybrid") %>%
ggmap()
ggmap_credentials()
```

```
Google -
key :
account_type : standard
day_limit : 2500
second_limit : 50
client :
signature :
```

Example: Twitter APIs

- Twitter provides access to much of their data for free
- output: user profile, follower data, tweet collections (keyword-based), ...
- registration mandatory (see, e.g., <http://bit.ly/2mM4WwV>)
- see also: <https://dev.twitter.com/rest/public>

Potential use cases

- analyze users' output (tweets, retweets)
- analyze follower networks
- real-time tracking of activity on the platform

Example: Twitter APIs

```
library(rtweet)

## search for tweets
tweets_aapor <- search_tweets("#AAPOR", n = 100)
names(tweets_aapor)
tweets_aapor$text[1:3]
```

Searching for tweets...
Finished collecting tweets!

```
[1] "screen_name"
[4] "status_id"
[7] "favorite_count"
[10] "is_retweet"
[13] "in_reply_to_status_user_id"
[16] "source"
[19] "media_url_expanded"
[22] "urls_expanded"
[25] "symbols"
[28] "place_id"
[31] "place_full_name"
[34] "bounding_box_coordinates"
[       ] "user_id"
[       ] "text"
[       ] "is_quote_status"
[       ] "retweet_status_id"
[       ] "in_reply_to_status_screen_name"
[       ] "media_id"
[       ] "urls"
[       ] "mentions_screen_name"
[       ] "hashtags"
[       ] "place_type"
[       ] "country_code"
[       ] "bounding_box_type"
[       ] "created_at"
[       ] "retweet_count"
[       ] "quote_status_id"
[       ] "in_reply_to_status_status_id"
[       ] "lang"
[       ] "media_url"
[       ] "urls_display"
[       ] "mentions_user_id"
[       ] "coordinates"
[       ] "place_name"
[       ] "country"

[1] "Come on, #AAPOR research nerds. I know you want to register for this. https://t.co/GpS7vGX6pb"
[2] "@kyleymcg: Have you registered for the next #AAPOR Webinar on 4/12? A Primer to Web Scraping w/ R
by @simonsaysnothin https://t.co/exFDD... "
[3] "@AAPOR: 2017 #AAPOR Award Winners Announced and will be honored during the Annual Conference in New
Orleans. https://t.co/eLfP0F5iNR "
```

Wrap-up

Advantages of data gathering with APIs

- pure data collection without “layout waste”
- standardized data access, robustness of calls
- de facto automatic agreement of data owner

Potential pitfalls

- requires you to understand the basics of the API architecture (or to the piece of software that links to it, e.g., an R package)
- dependent upon API suppliers; sometimes requires registration
- not always free

Wrap-up

Resources

- Collection of APIs on the web:
<http://www.programmableweb.com/apis>
- rOpenSci: Collection of R-API interfaces:
<http://ropensci.org/packages/>
- Twitter mining with R:
<https://mkearney.github.io/rtweet/index.html>

Outlook and helpful resources



A word cloud centered around the word "Workflow". The word "Workflow" is the largest and most prominent word, colored red. Other significant words include "process" (blue), "management" (dark blue), and "systems" (teal). Smaller words surrounding the center include "workflows" (dark blue), "output" (light blue), "flow" (light blue), "work" (light blue), "information" (green), "processing" (green), "business" (green), "development" (green), "concept" (green), "components" (green), "processes" (green), "description" (green), "may" (green), "analysis" (green), "data" (green), "the" (light blue), "control" (light blue), "related" (light blue), "one" (light blue), "systems" (teal), "scientific" (teal), "concepts" (green), "can" (teal), and "process" (blue).

Web technologies

Technologies for disseminating content on the Web

Technologies for information extraction

Technologies for data storage

HTTP

R

R

XML/HTML

XPath/CSS selectors

JSON

JSON parsers

AJAX

Selenium

plain text

Regular expressions

SQL

NoSQL

binary formats

plain-text formats

Three next steps for you (takes less than 3 hours)

1. Go to the following website and  the repository to your desktop:
<https://git.io/vyAde>
2. Start with opening `01-intro.r` and work through the commented R code. Then do the same with `02-scraping-with-rvest.r`
3. By modifying existing examples in the code, get better at understanding the individual steps in the process and ultimately start your own scraping projects

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/>

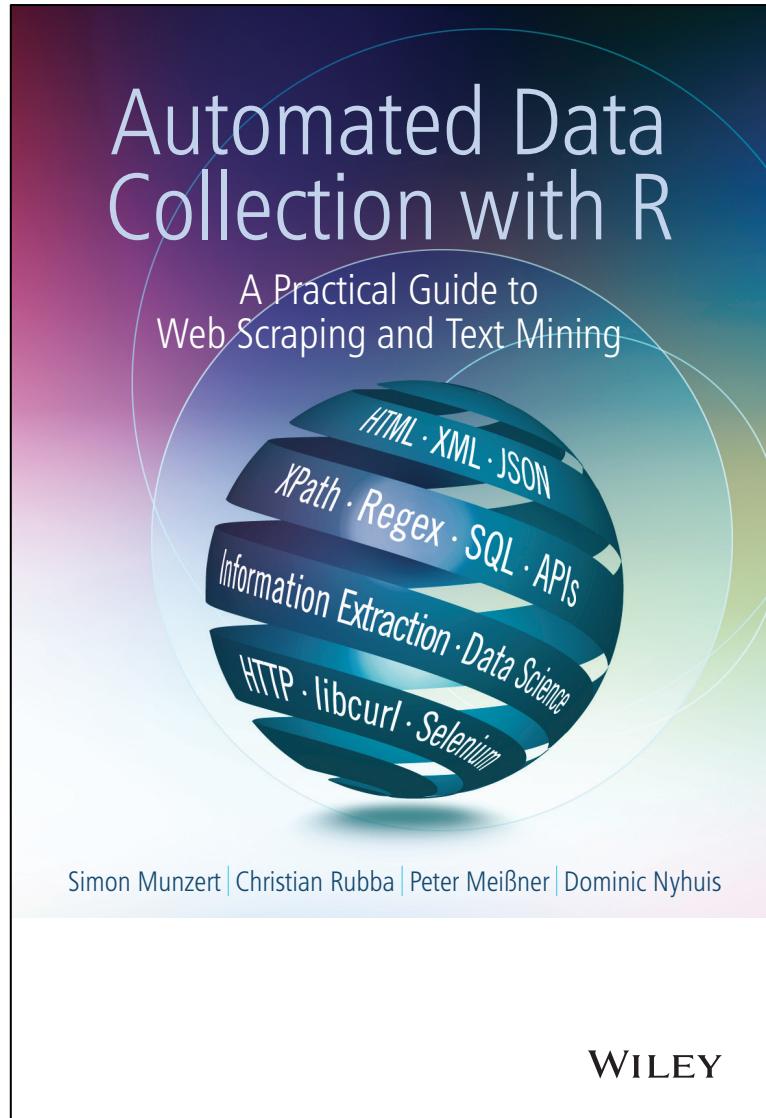
3. Integrate more packages in your scraping toolbox

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

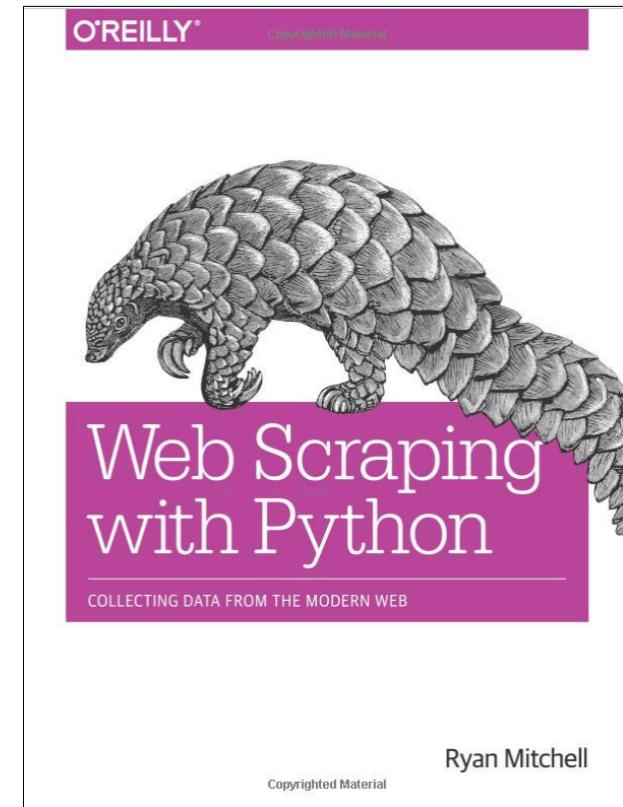
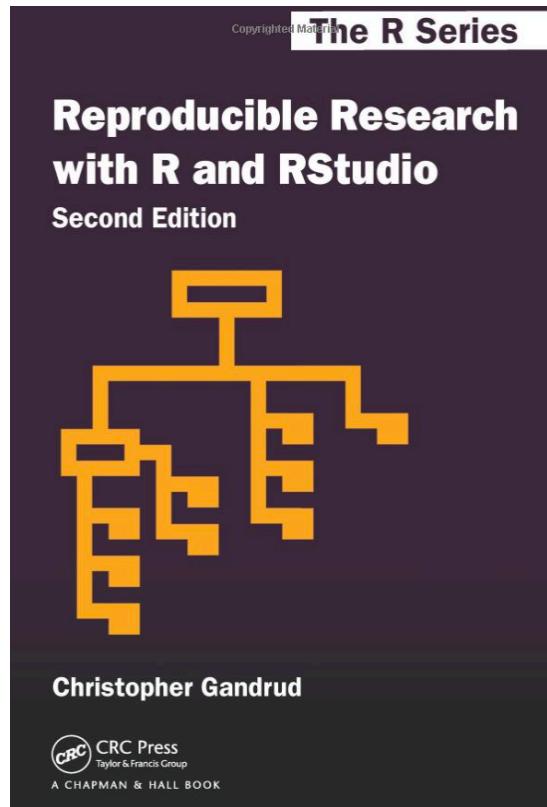
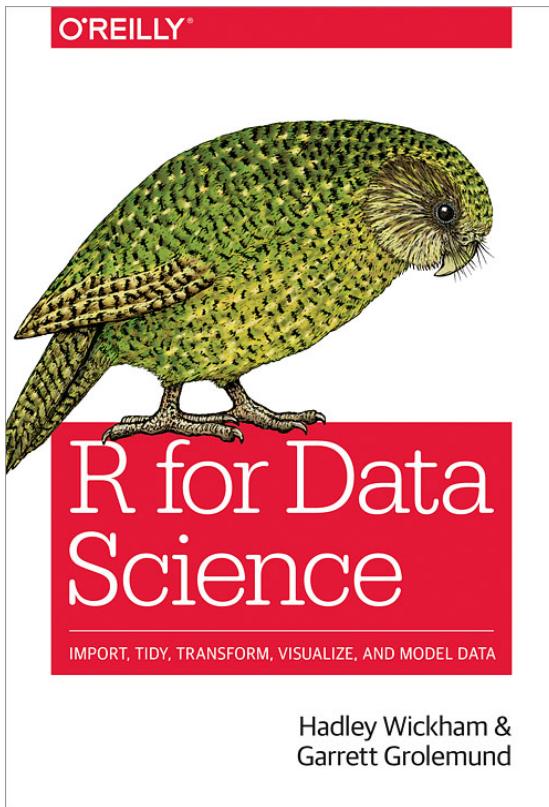
An accompanying book

- 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!

Slides and Materials:

<https://git.io/vyAde>

Book:

r-datacollection.com
@RDataCollection

Me:

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