

# Web Scraping mit R

## Eine kleine Einführung

Simon Munzert  
Universität Konstanz

r-datacollection.com  
@RDataCollection  
#rwebscraping

Juni 2015

Vorab: Stellen Sie Fragen! Egal, welche...



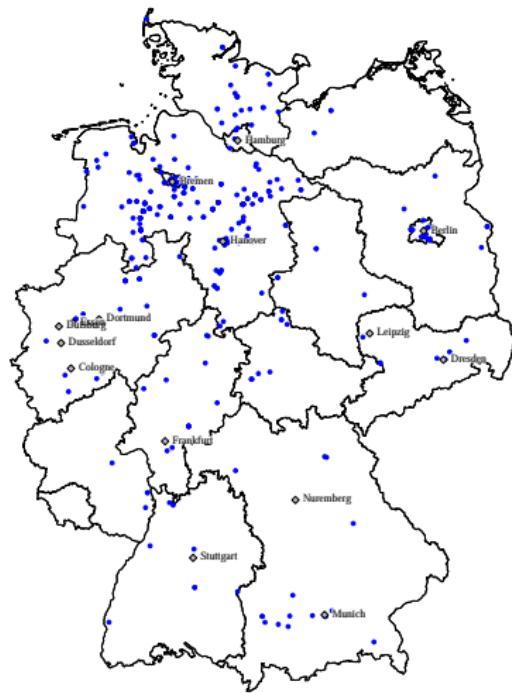
**"Excuse me, but is this The  
Society for Asking Stupid  
Questions?"**

# Ein kleiner Vorgeschmack

People named FLÖSCH in Germany



People named RIPKE in Germany



# Warum Datenerhebung im World Wide Web?

## Web scraping

*A.k.a. screen scraping, crawling, web harvesting;*  
computergestütztes, teilautomatisiertes Sammeln vornehmlich  
unstrukturierter Daten (z.B. aus HTML-Code)

# Warum Datenerhebung im World Wide Web?

## Web scraping

*A.k.a. screen scraping, crawling, web harvesting;*  
computergestütztes, teilautomatisiertes Sammeln vornehmlich  
unstrukturierter Daten (z.B. aus HTML-Code)

Das Web hält eine Fülle relevanter Daten bereit, z.B.

- Verhaltensdaten: Suchmaschinenabfragen, Kaufverhalten, Kommunikation
- Webseiten politischer Akteure
- Social Media (Blogs, Twitter)

# Warum Datenerhebung im World Wide Web?

## Web scraping

*A.k.a. screen scraping, crawling, web harvesting;*  
computergestütztes, teilautomatisiertes Sammeln vornehmlich  
unstrukturierter Daten (z.B. aus HTML-Code)

Das Web hält eine Fülle relevanter Daten bereit, z.B.

- Verhaltensdaten: Suchmaschinenabfragen, Kaufverhalten, Kommunikation
- Webseiten politischer Akteure
- Social Media (Blogs, Twitter)

Warum ist das für uns von besonderem Interesse?

- neue Forschungsfragen!
- spärliche finanzielle und zeitliche Ressourcen
- Reproduzierbarkeit des Forschungsprozesses

# Web Scraping mit R – wie geht das?

- Grundsätzlich: **kein point-and-click-Verfahren**, sondern Automatisierung der Datenerhebung durch eigene kleine Skripte
- Klassisches **Screen-Scraping**: HTML-Seiten als Text, Extraktion von Teilen des Texts mithilfe spezieller Techniken, die sich über viele Seiten automatisieren lassen
- Nutzung von **Web Services/Web APIs**: Software, die Daten bereits aufbereitet zur Verfügung stellt
- Weiterverarbeitung gewonnener Textinformation mit **Textmining-Verfahren**

# Beispiel: ein Wikipedia-basiertes Netzwerk bekannter Psychologen

**Ziel:** Konstruktion eines Netzwerks  
bekannter Psychologen

## Tasks:

- erhebe Liste von Psychologen
- greife auf einzelne Wikipediaeinträge zu
- identifiziere Querverweise
- erstelle die Konnektivitätsmatrix
- visualisiere das Netzwerk



# Was man mit R alles scrapen kann

## Semi-strukturierte Daten

The screenshot shows a web browser window displaying the German Wikipedia page for 'Liste bedeutender Psychologen'. The page lists various psychologists, categorized by letter (A, B, C, D, E, F, G). The browser interface includes a sidebar with navigation links like 'Hauptseite', 'Themenportale', and 'Artikel verfassen', and a bottom bar with search and navigation icons.

**A [Bearbeiten]**  
Alfred Adler – Mary Ainsworth – Gustav Johannes von Allesch – Gordon Allport – Rudolf Arnheim – Wilhelm Karl Arnold – Elliot Aronson – Solomon Asch – John William Atkinson

**B [Bearbeiten]**  
Robert Freed Bales – James Mark Baldwin – Michael Balint – Paul B. Baltes – Albert Bandura – Aaron T. Beck – Dieter Beckmann – Daryl Bem – Hans Bender – Friedrich Eduard Beneke – George Berkeley – Siegfried Bernfeld – Bruno Bettelheim – Alfred Binet – Niels Birbaumer – Ernst E. Boesch – Ernst Bömmann – John Bowby – Elmar Brähler – Jürgen Bredenkamp – Franz Brentano – Uri Bronfenbrenner – Charlotte Bühler – Kari Bühner – Cyril Burt – David Buss

**C [Bearbeiten]**  
Donald T. Campbell – James McKeen Cattell – Raymond Bernard Cattell – Ludwig Ferdinand Claus – Ruth Cohn – Mihaly Csikszentmihalyi

**D [Bearbeiten]**  
Karl Duncker – Dietrich Dörmer – Rudolf Dreikurs – Heinrich Düker –

**E [Bearbeiten]**  
Hermann Ebbinghaus – Danil Borisowitsch Elkonin – Albert Ellis – Erik H. Erikson – Milton H. Erickson – Hans J. Eysenck

**F [Bearbeiten]**  
Jochen Fahrenberg – Gustav Theodor Fechner – Leon Festinger – Peter Fiedler – Kurt W. Fischer – Peter Fonagy – Peter Frensch – Erich Fromm – Erika Fromm – Sigmund Freud

**G [Bearbeiten]**  
Franz Josef Gall – Francis Galton – Howard Gardner – Eugene T. Gendlin – Eleanor J. Gibson – James J. Gibson – Gerd Gigerenzer – Daniel Goleman – Kurt Gottschaldt – Carl Friedrich Graumann – Klaus Graae – Amo Grun – Jürgen Guthke –

# Was man mit R alles scrapen kann

## Suchmaschinendaten

The screenshot shows a Google search results page with the query "web scraping with r". The results are filtered by "Web" and show approximately 4,720,000 results. The first few results are:

- Web Scraping and R - R-bloggers**  
www.r-bloggers.com/search/web/scraping • Diese Seite übersetzen  
12.03.2014 · The latest slides from my web scraping through R! Web scraping for the humanities and social sciences
- Web-Scraping: the Basics | (R news & tutorials) - R-bloggers**  
www.r-bloggers.com/web-scraping-the-basics/ • Diese Seite übersetzen  
19.02.2014 · Slides from the first session of my course about web scraping through R. Web scraping for the humanities and social sciences
- purrr Package 'scraper'**  
cran.r-project.org/web/packages/scraper/scraper.pdf • Diese Seite übersetzen  
03.02.2010 · Description Tools for Scraping Data from Web-Based Documents. License : XML Tools for parsing and generating XML, within R and S-Plus.
- Web-Scraping in R | DiffusePrior**  
diffuseprior.wordpress.com/2012/04/04/web-scraping-in-r/ • Diese Seite übersetzen  
02.04.2012 · Web-scraping, or web-crawling, sounds like a seedy activity worthy of an Interpol investigative department. The reality, however, is far less ...
- Webscraping using readLines and RCurl - ProgrammingR**  
www.programmering.com/webscraping-using-re... • Diese Seite übersetzen  
02.01.2013 · Accessing online data of this sort is sometimes referred to as "webscraping". R facilities, readLines() from the base package and getURL() ...
- WebScraping with R | PearceTrees**  
www.pearcetrees.com/.../DataTools - R-project • Diese Seite übersetzen  
Web Scraper for Google Scholar updated Scraping table from any web page with R or CloudStat. Retrieving RSS Feeds Using Google Reader. Scraping ...
- GivenTheData: R and the web (for beginners), Part II ...**  
giventhedata.blogspot.com/.../r-and-web-for-begin... • Diese Seite übersetzen  
23.08.2012 · In this last post of my little series (see my latest post) on R and the web I explain how to extract data of a website (web scraping) ...
- purrr Web Scraping with R**  
cos.name/web-scraping-with-R-xiaolan.pdf • Diese Seite übersetzen  
Web Scraper with R. Xiao Nan @soa22@fudan 6th China R Beijing. Xiao Nan @ roadstat. Web Scraper with R. 1/43. 1/43 ...
- Scraping data from the Web with R - Revolutions**  
blog.revolutionanalytics.com/scraping-data-from... • Diese Seite übersetzen  
blog.revolutionanalytics.com/2010/07/scraping-data-from-the-web-with-r.html Sometimes the data we need is not packaged up nicely into a simple command-separated file or database.

# Was man mit R alles scrapen kann

## Dynamische Daten

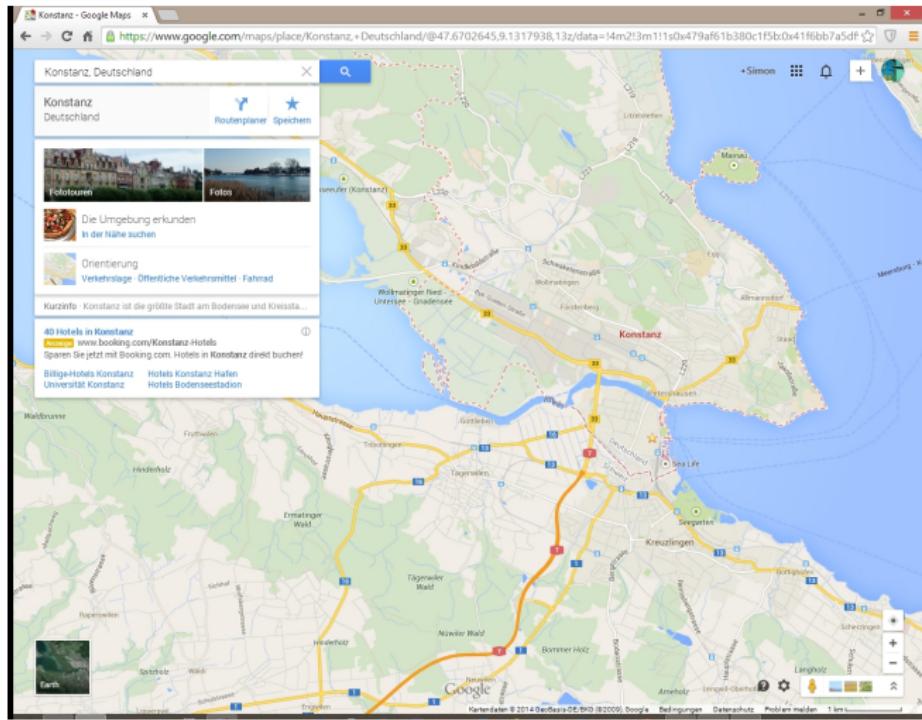
The screenshot shows a web browser window displaying a news article from SPIEGEL ONLINE. The URL in the address bar is [www.spiegel.de/sport/fussball/philipp-lahm-ruecktritt-reaktionen-von-niersbach-sammer-rummenigge-a-981751.html](http://www.spiegel.de/sport/fussball/philipp-lahm-ruecktritt-reaktionen-von-niersbach-sammer-rummenigge-a-981751.html). The page title is "Reaktionen auf Philipp-Lahm-Rücktritt: 'Ich dachte, er wartet noch ein bisschen'".

The main image shows three German national football team players (Lukas Podolski, Philipp Lahm, and Thomas Müller) cheering and shouting. Below the image, a caption reads: "Er war partout nicht zum Bleiben zu bewegen: Der Rücktritt von Philipp Lahm aus der Nationalmannschaft löst unterschiedliche Reaktionen aus. Viel Lob - und Sorge ums DFB-Team."

At the bottom of the page, there is a sidebar with a "THEMA Philipp Lahm" section and a quote from Wolfgang Niersbach (DFB-President): "Philipp hat mich heute Morgen angerufen und persönlich über diesen Schritt informiert. Ich habe in dem Gespräch sehr

# Was man mit R alles scrapen kann

## Dynamische Webseiten



# Was man mit R alles scrapen kann

## Social-Media-Daten

The screenshot shows a Twitter search results page for the query "Automated Data Collection with R". The results include tweets from various accounts:

- ECPR (@ECPR)**: "Who's coming to the #ECPR Summer School? #ssmt14 on.fb.me/1mvpvhg #notting #excited" (17s ago)
- PagerDuty (@pagerduty)**: "Wake up, your server is on fire! Always get your alerts with PagerDuty. Try it for free: pduty.me/r1faicR" (Jun 26)
- Hortonworks (@hortonwo...)**: Followed by Robert Schmidt a. (Follow, Promoted)
- DUKE UniversityPress (@DUK...)**: Followed by Robert Schmidt a. (Follow)
- U of MN Press (@UMinnPress)**: Followed by Robert Schmidt a. (Follow)

The tweet from PagerDuty includes an image of a person wearing a black t-shirt with the text "I never sleep through outages." Below the tweets, there is a "DID YOU KNOW?" box with the following text:

**DID YOU KNOW?**

The idea of the seaside holiday was popularised by Dr. Richard Russell's Brighton clinic and his *Dissertation Concerning the Use of Sea Water in Diseases of the Glands*, printed by the Press in 1752.

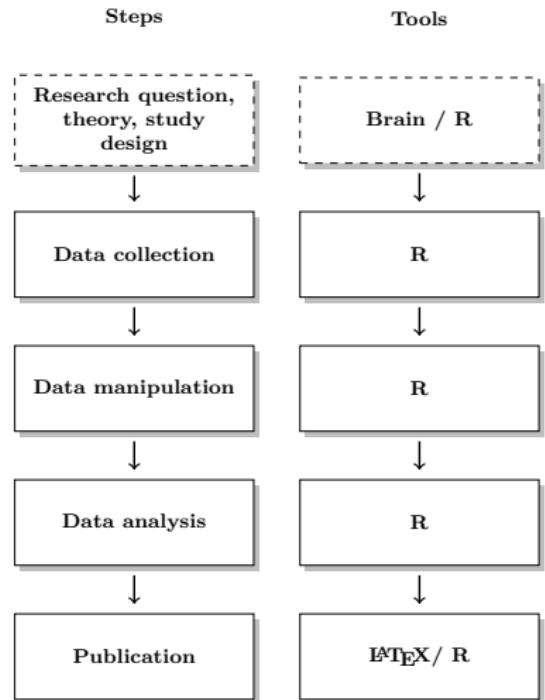
# Was man mit R alles scrapen kann

... und vieles mehr:

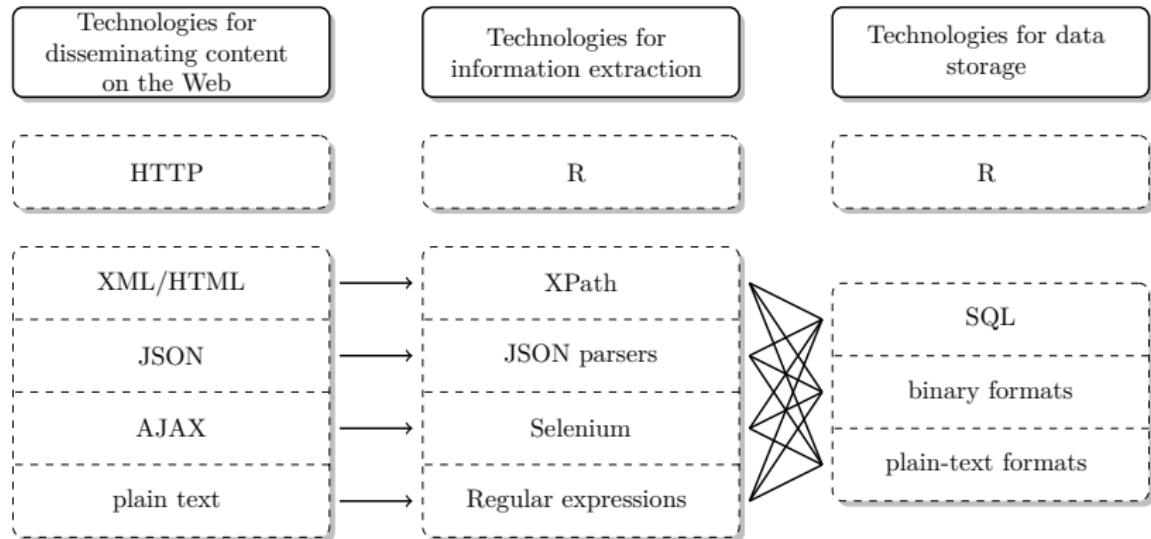
- R als automatisierter Browser
- R als Download-Manager großer Dateienmengen
- Import und Parsing von HTML-, XML- und JSON-Inhalten
- wiederholte Abfrage dynamischer Inhalte
- Anbindung an REST-basierte Web Services problemlos möglich
- Authentifizierung (auch via OAuth), diverse Protokolle (HTTP, HTTPS, FTP, ...)

# Warum R?

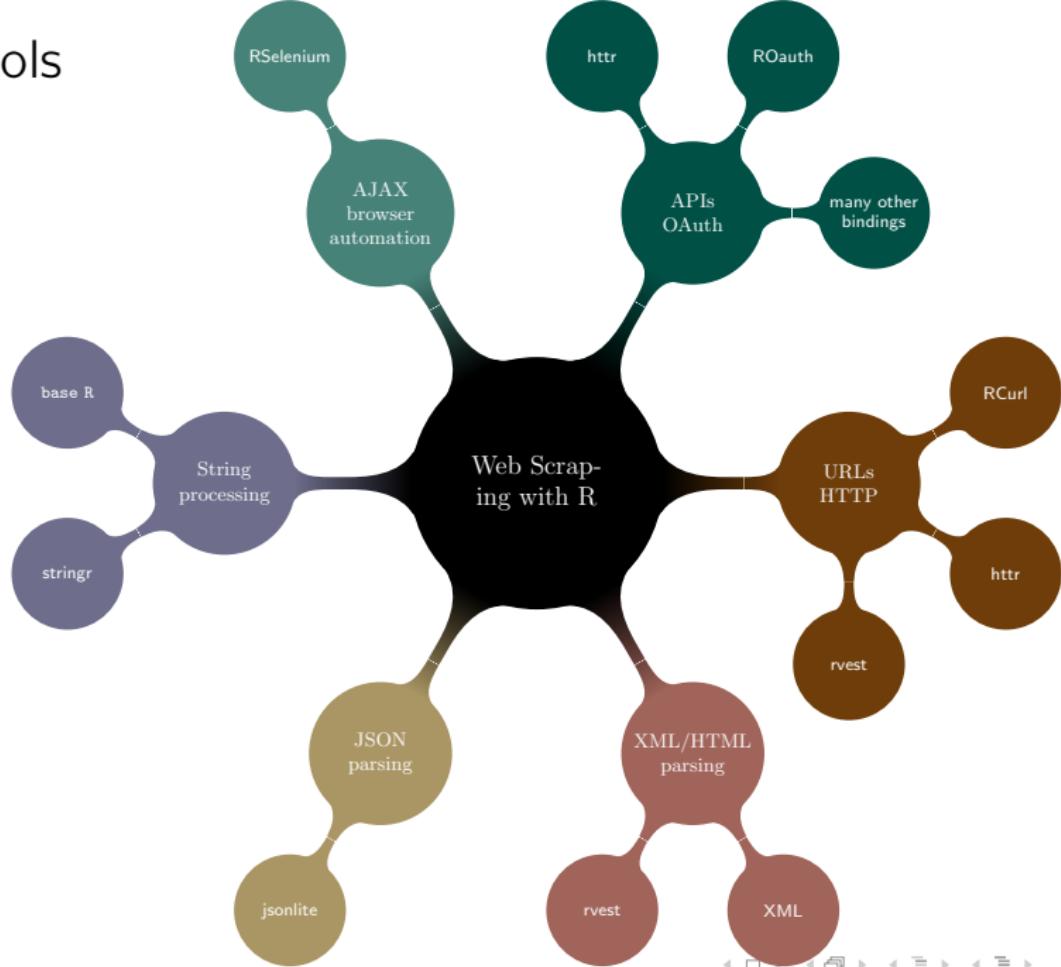
- kostenlos
- Open Source
- große Community
- stark in statistischer Analyse
- stark in Datenvisualisierung
- Verfügbarkeit auf allen Plattformen
- bedient den Workflow vom Anfang bis zum Ende



# Technologien des World Wide Web



# R-Tools



# Und jetzt: Zeit für etwas Praxis mit R!

Sie benötigen:

- R + Editor (RStudio)
- funktionierende Internetverbindung
- Daten und Code, abrufbar unter

[https://github.com/simonmunzert/  
workshopKNPsych2015](https://github.com/simonmunzert/workshopKNPsych2015)

