1. Introduction

23.04.2020

Übung: “Big Data Analysis” Using R
Felix Lennert, B.A.
felix.lennert@politik.uni-regensburg.de

Universität Regensburg

Some rules for the meeting

- Keep your microphone muted at all times—unmute it only if you want to say something
- Try to speak in a slow, loud, and clear manner
- Please indicate that you can hear and see me properly by sending a “thumbs up” when I ask you to do so
- Interrupt me if anything is unclear
- Interrupt me if I am too fast
- Interrupt me if there are technical problems—we will not leave anyone behind…
- If you have any requests or whatever: feel free to simply use the chat

Outline

- The “Data Deluge”
- Coding as an existential skill of a 21st century researcher
- Chores
- Who are you?
- What are you going to learn?
- How are we going to go about that?

The “Data Deluge”

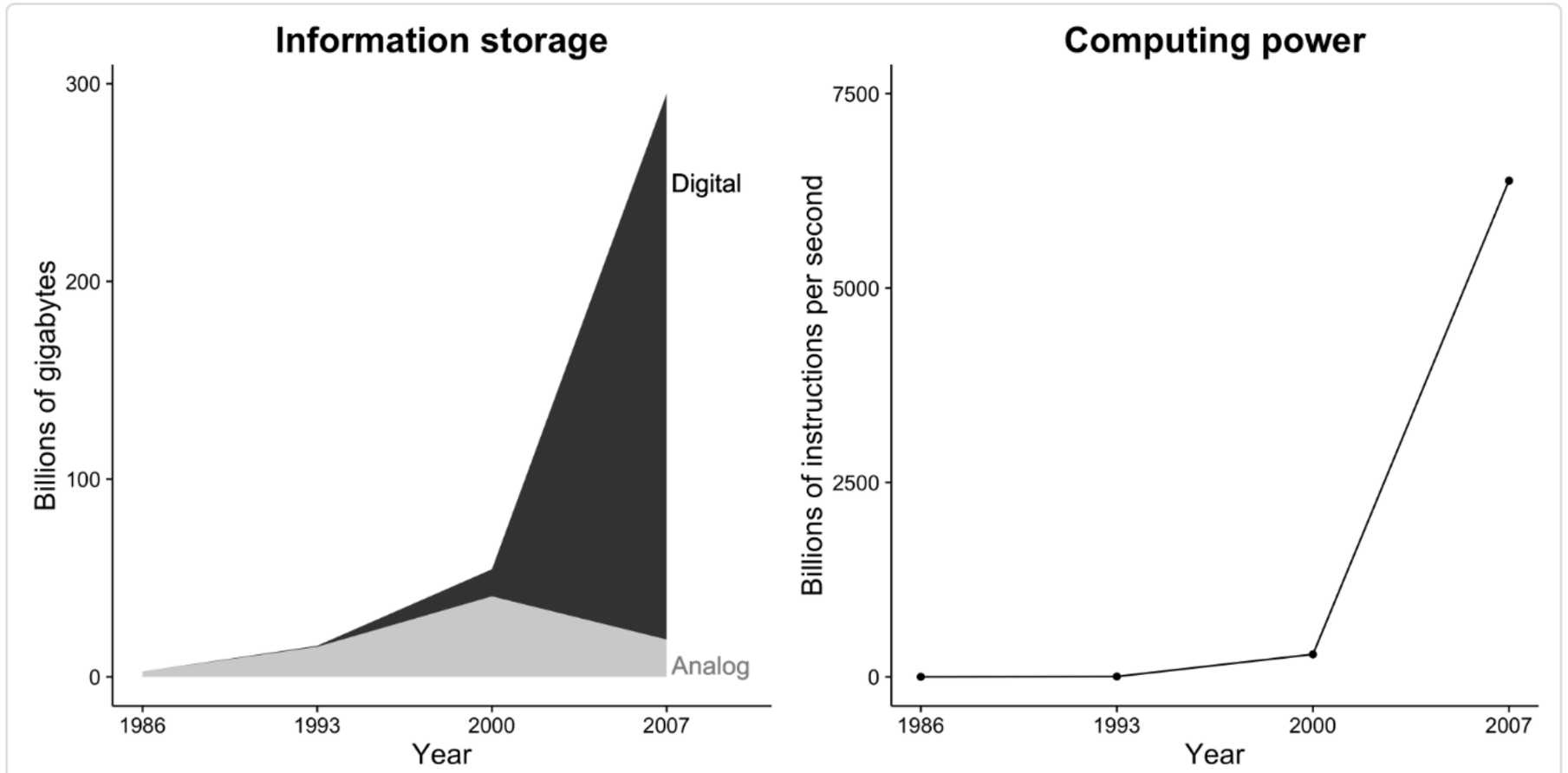


Fig. 1: Development of information storage and computing power

The “Data Deluge”

Fig. 2: Economist 02/27/2010

Fig. 3: Economist 05/06/2017

The “Data Deluge”

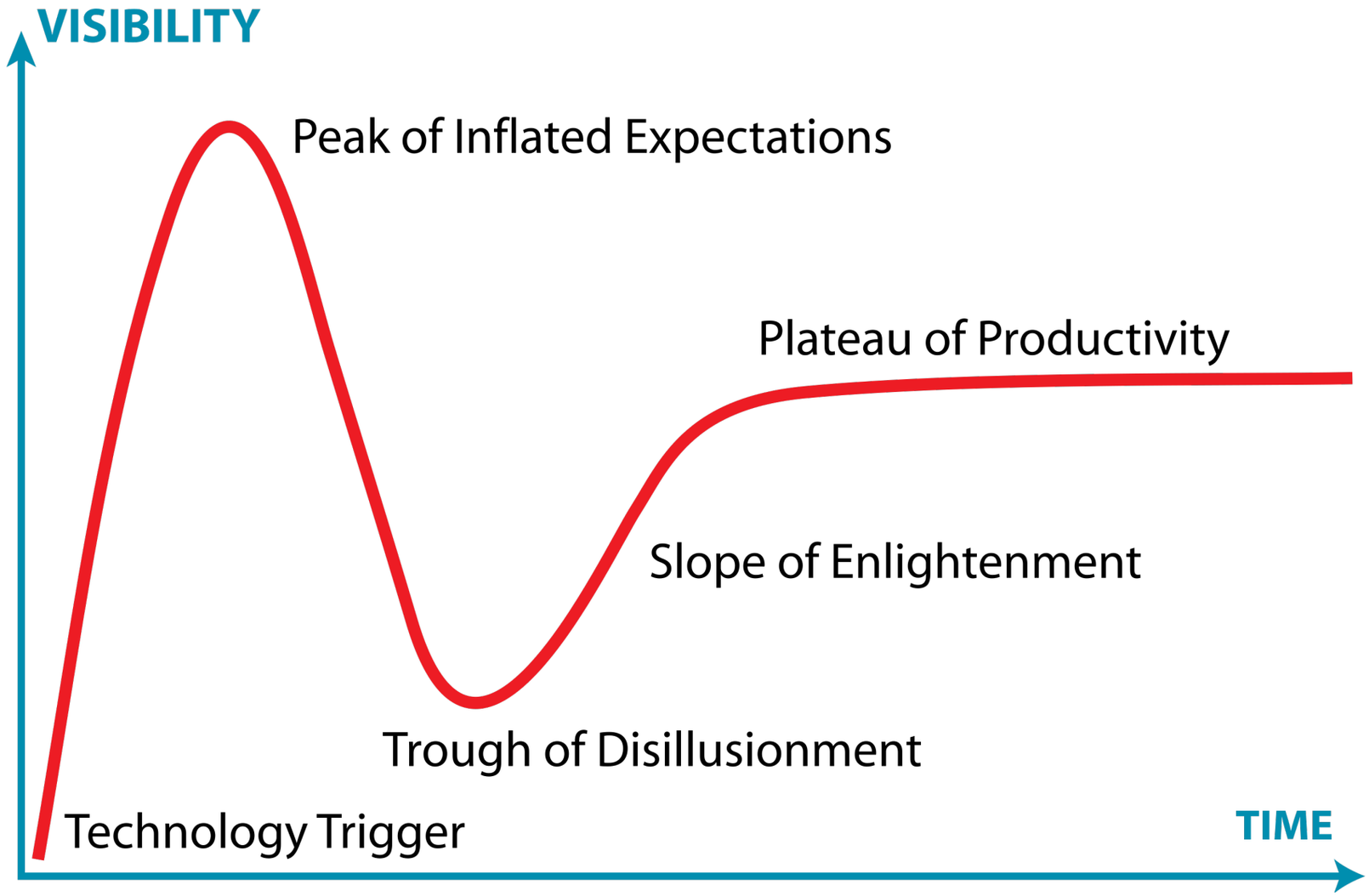


Fig. 4. Gartner Hype Cycle

The “new oil”?

“Data is the new oil. It’s valuable, but if unrefined it cannot really be used. It has to be changed into gas, plastic, chemicals, etc to create a valuable entity that drives profitable activity; so must data be broken down, analyzed for it to have value.”

— Clive Humby

Why coding skills matter

Fig. 5: Tweet

What R coding facilitates

- **Reproduction of existing research**
- **Dynamic development of state-of-the-art techniques**
- **Being part of a bigger community**
- **Getting a job**
- **The entire process of doing research**

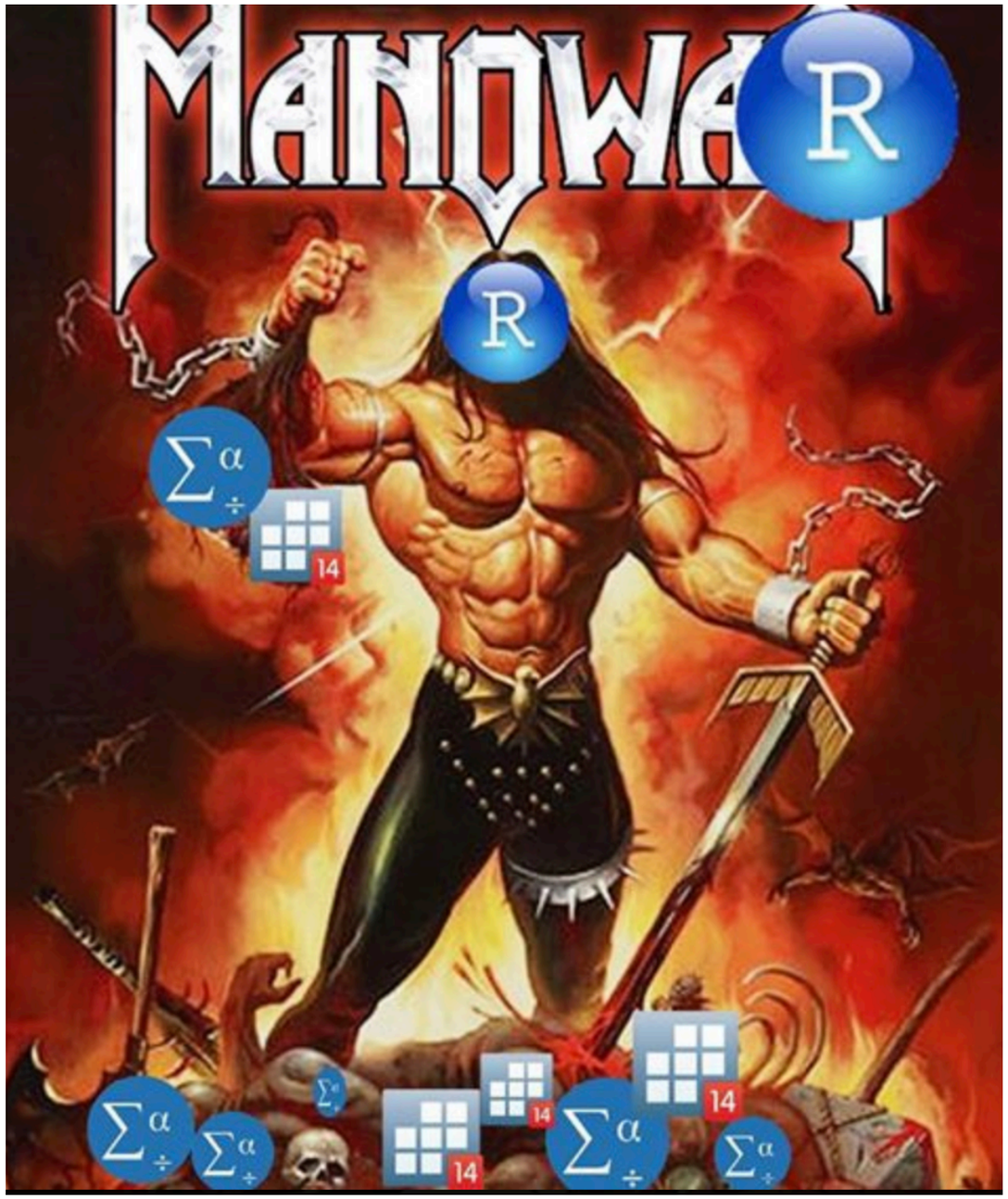


Fig. 6: Random R Meme

Reproducibility

Example:

Voices from the far right: a text analysis of Swedish parliamentary debates

Måns Magnusson
Linköping University
mans.magnusson@liu.se

Katarina Barrling
Uppsala University

Richard Öhrvall
Linköping University and
Research Institute of Industrial Economics

David Mimno
Cornell University

Abstract

In this paper we study the effects of a radical right party entering a national parliament, on the parliament discourse. We follow the classification developed by Meguid (2008) and use a probabilistic topic model approach to analyze the 300,000 speeches delivered in the Swedish parliament between 1994 and 2017. Our results indicate that immigration became a more prevalent topic in party leader debates when the Sweden Democrats entered the parliament in 2010. The other parties started to address immigration more, but still not to the extent that the Sweden Democrats did. In 2015, as Sweden faced a migration crisis, immigration became a more salient issue in the parliamentary debates. This could be seen as an external shock that forced the mainstream parties to put more emphasis on the topic of immigration. We conclude that the mainstream parties used a partly dismissive, partly adversarial strategy in their speeches when the SD entered the parliament. The migration crises in 2015 made them focus more on immigration and they thereby adopted a more adversarial strategy.

[1]The corpus and all details on how the data has been curated can be found at https://github.com/MansMeg/rcrpsriksdag. The open data of the Swedish parliament can be found at https://data.riksdagen.se/.

[2]All details (list of collocation words, stop words etc.) and the final corpus can be found at https://github.com/MansMeg/rcrpsriksdag.

Magnusson et al. 2018

Dynamic development

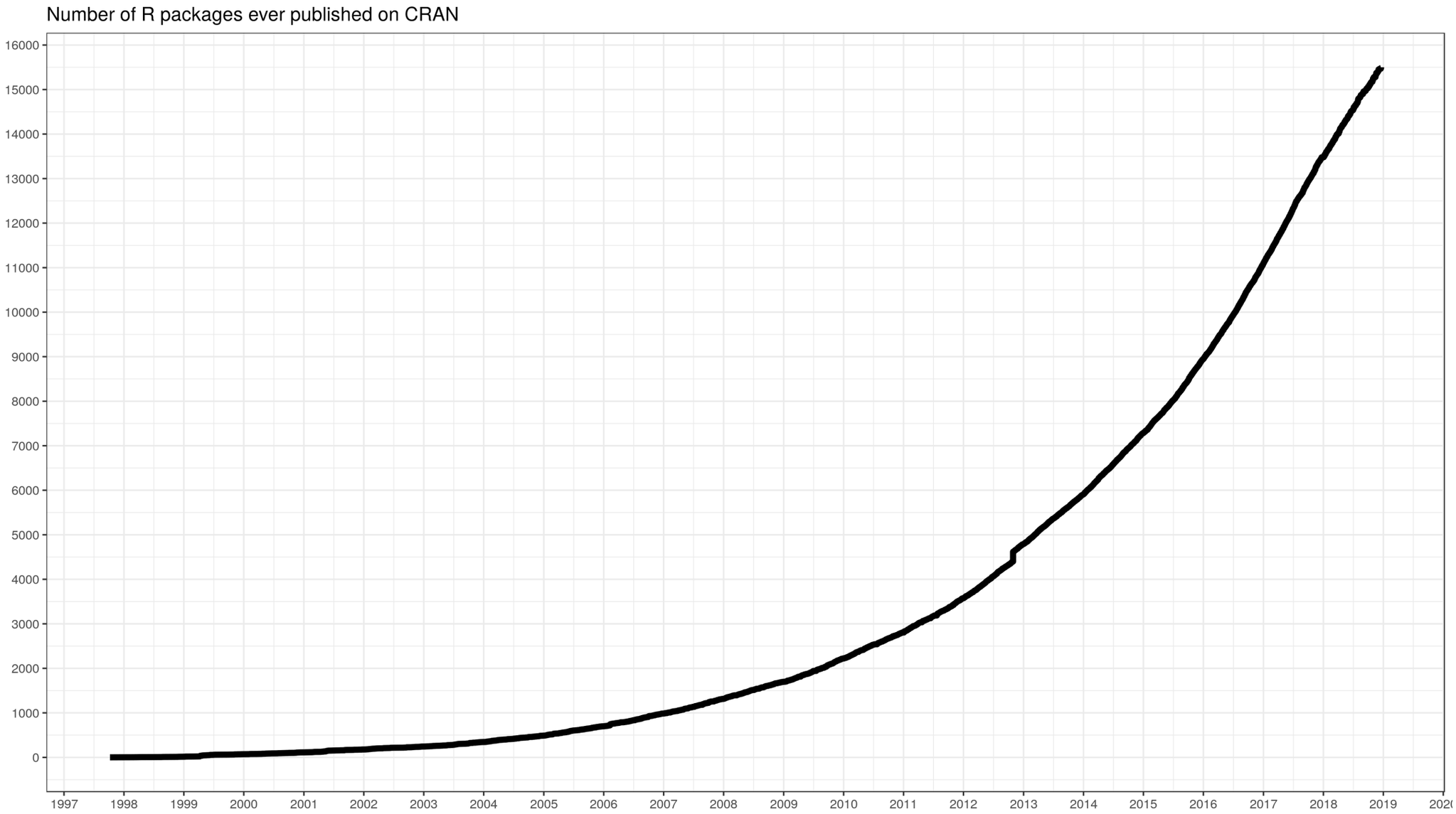


Fig. 7: Number of R packages on CRAN—development

The R community

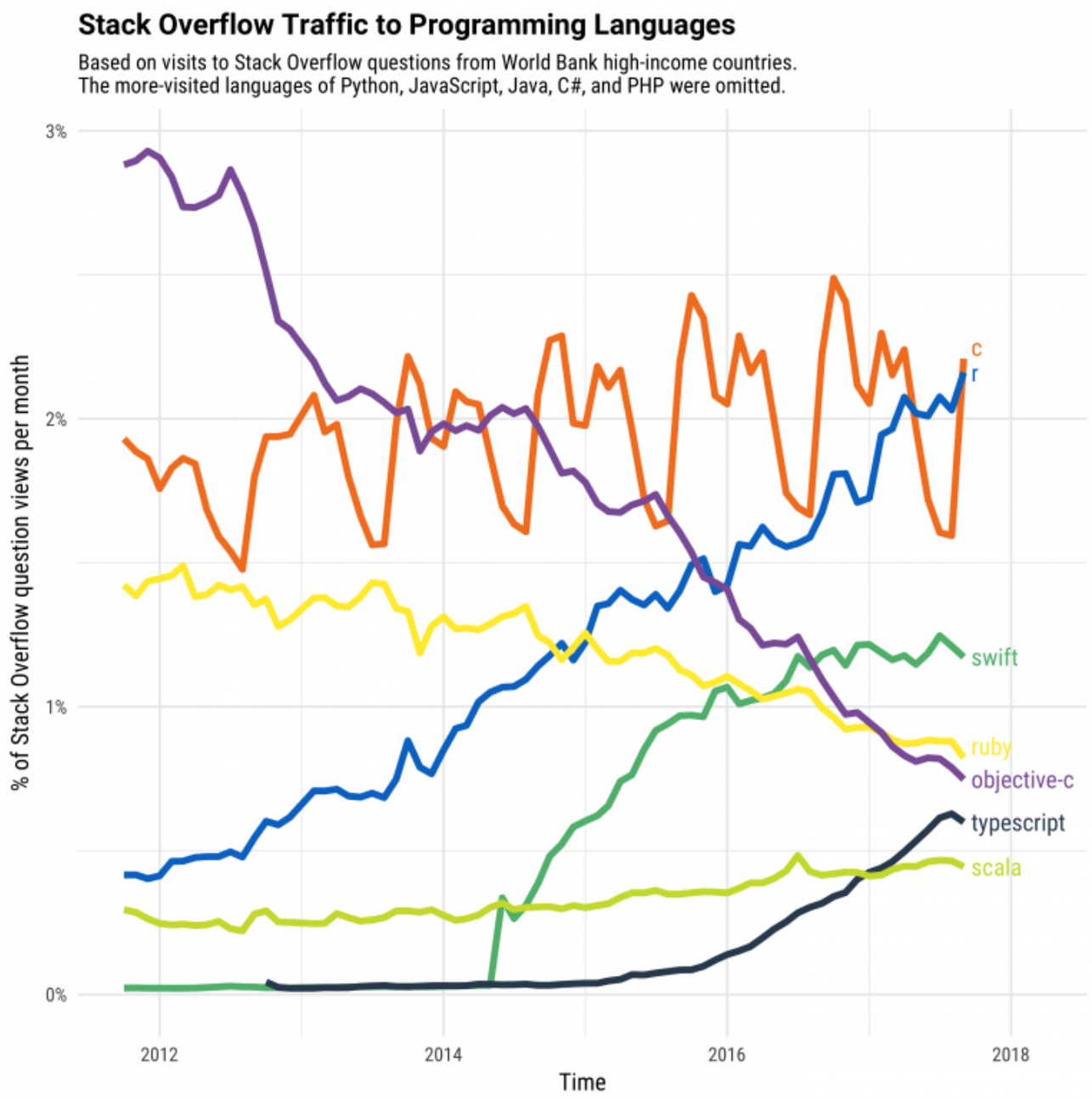


Fig. 8: StackOverflow traffic

Übung: “Big Data Analysis” Using R
Felix Lennert, B.A.
felix.lennert@politik.uni-regensburg.de

The R community

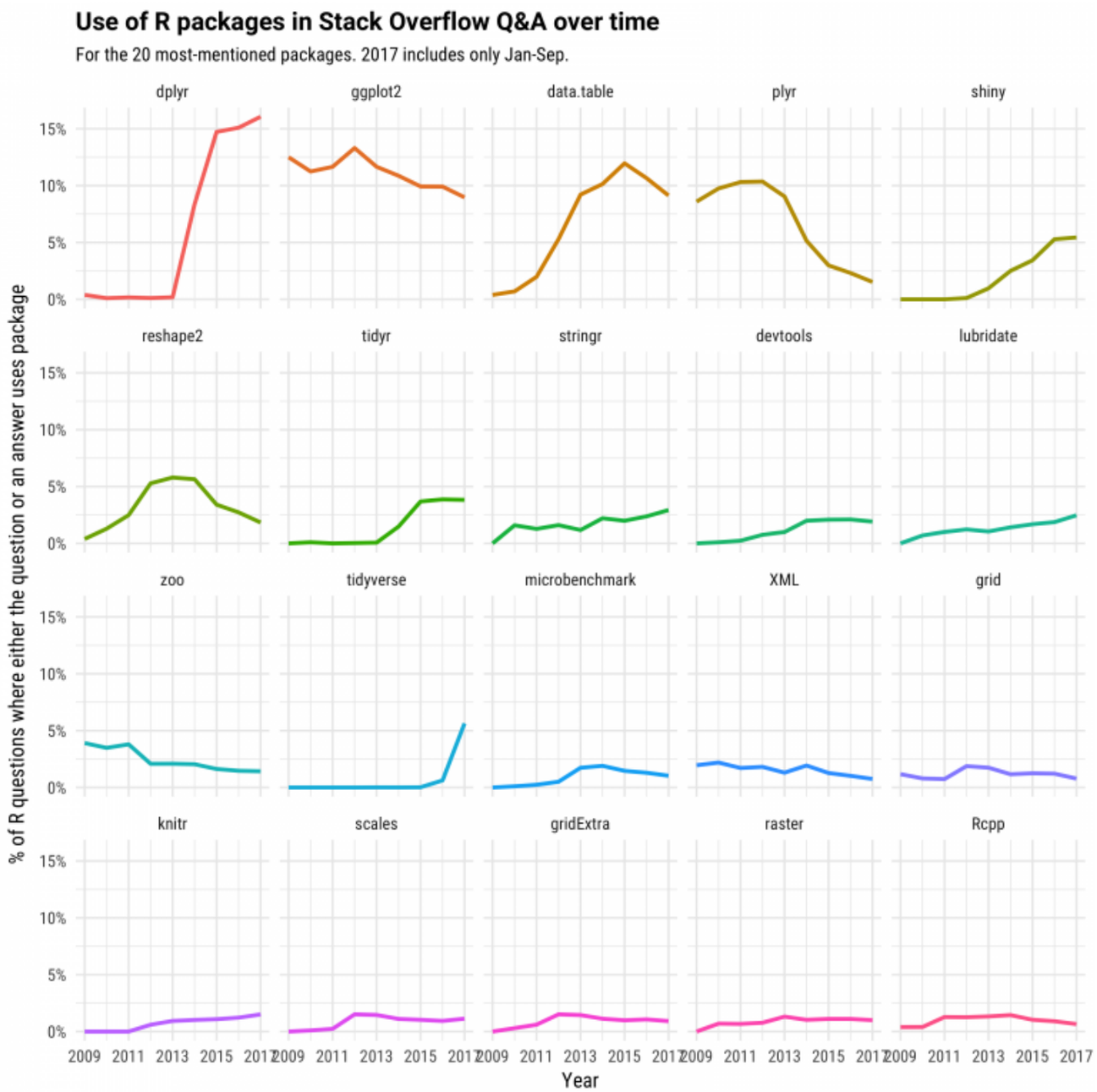


Fig. 9: StackOverflow traffic

The R community

Software	Number of Blogs	Source
R	550	R-Bloggers.com
Python	60	SciPy.org
SAS	40	PROC-X.com, sasCommunity.org Planet
Stata	11	Stata-Bloggers.com

Fig. 10: Number of blogs

The R community

#rstats

Fig. 11: RStats on Twitter

Ladies

Fig. 12: RLadies

Fig. 13: TidyTuesday

Job offers that require proficiency in R

Harvard Business Review

Data | **Data Scientist: The Sexiest Job of the 21st Century**

Fig. 14: Data Scientist…

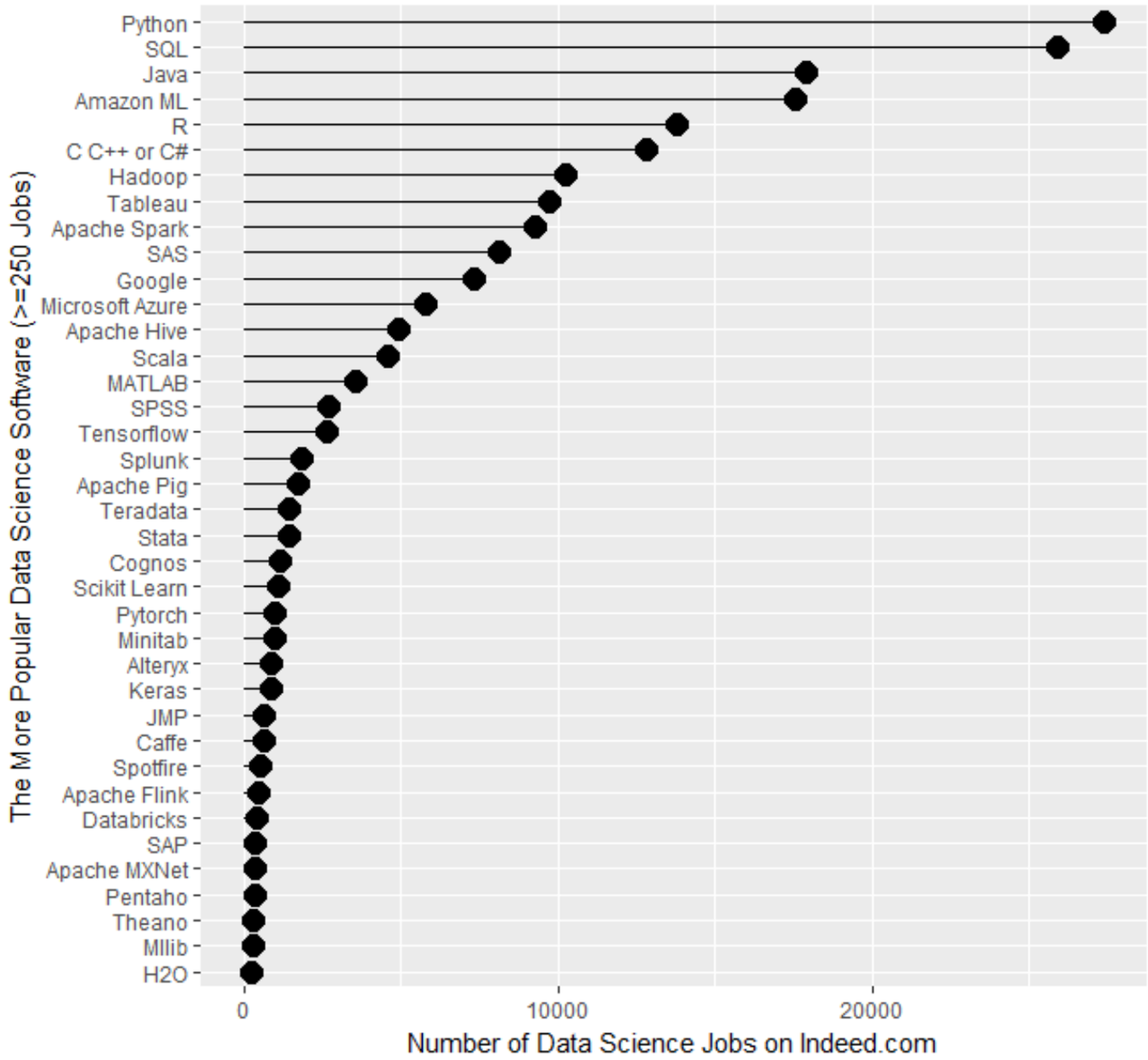


Fig. 15: Job offers

Übung: "Big Data Analysis" Using R

Felix Lennert, B.A.
felix.lennert@politik.uni-regensburg.de

How does R change the process of doing research?

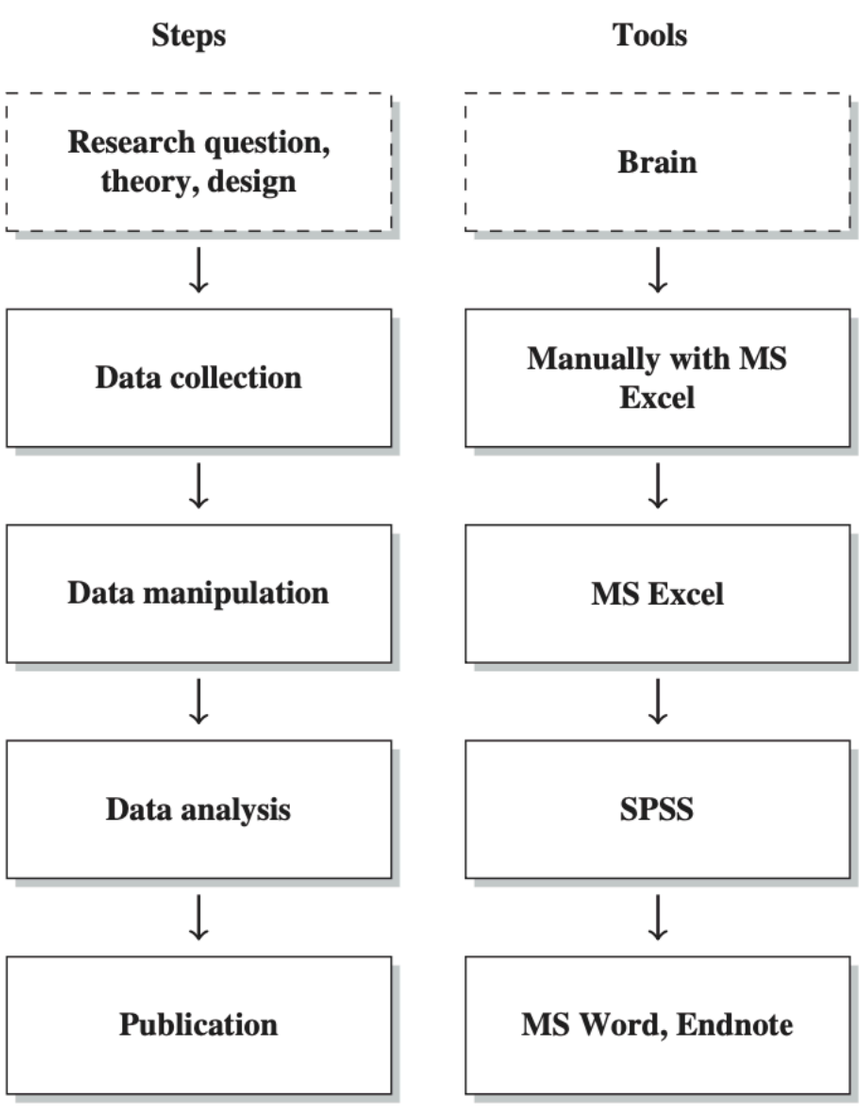


Figure 1 The research process **not** using R—stylized example

Fig. 16: Work flow without R…

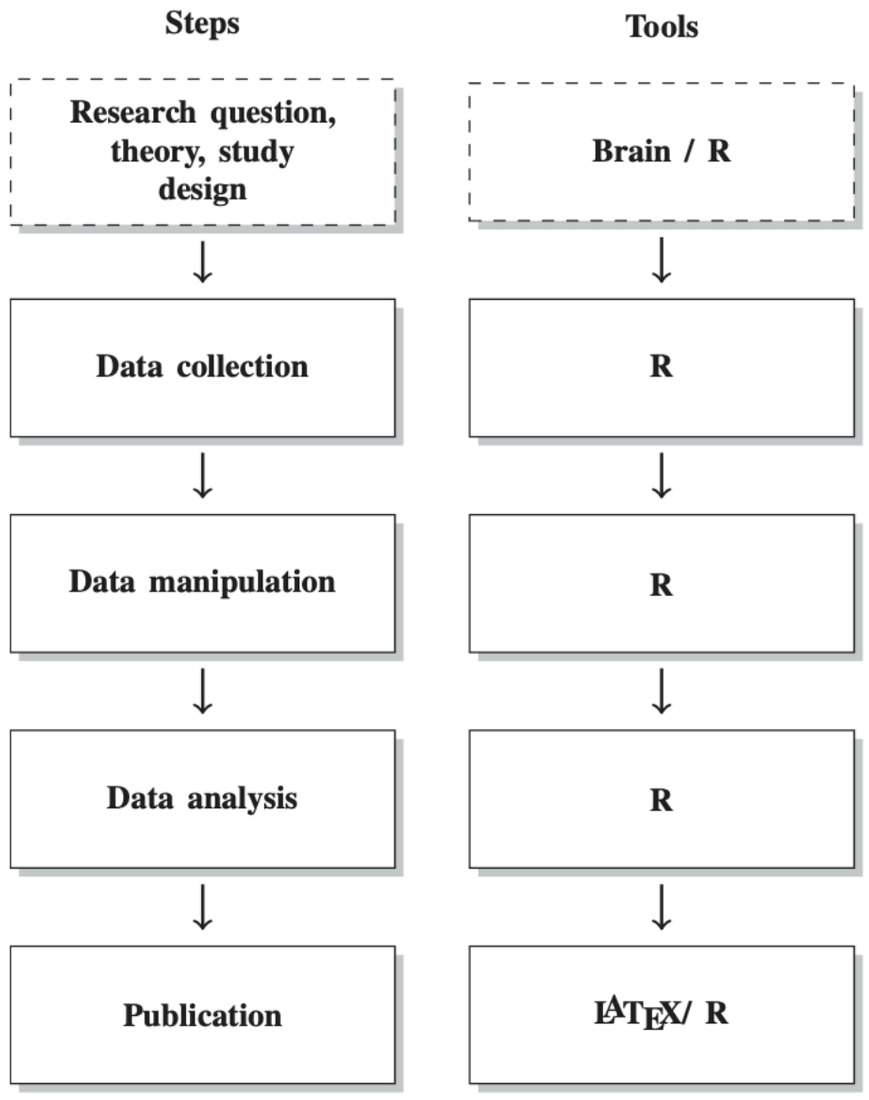


Figure 2 The research process using R—stylized example

… and with R.

The chores—FlexNow

Modulkürzel	Modulposition	Leistungspunkte	ab Semester
EUST-M04.1	Themen aus der Geschichts- oder Politikwissenschaft mit europäischem Bezug	7	WS 2019/20
POL-BA-27.2	Seminar	5	WS 2012/13
POL-BA-WB-fachintern	Politikwissenschaft Bachelor of Arts Wahlbereich fachintern	5	WS 2008/09
POL-MA-28.1	Praxisseminar	5	WS 2012/13
POL-MA-29.2	Praxisseminar	5	WS 2012/13
POL-MA-30.1	Nach Wahl		WS 2012/13
POL-MA-30.2	Nach Wahl		WS 2012/13
POL-MA-ZfL	zusätzliche freiwillige Leistungen Master Demokratiewissenschaft	5	WS 2012/13

Fig. 17: Check out LSF

DEADLINE: 26.04.2020

If there are any problems: send an email to michaela.schmid@politik.uni-regensburg.de

The chores—how to pass the course?

- Two take-home assignments
- One article review (3–5 pages)

Fig. 18: Coffee GIF

BREAK ANYONE?

Fig. 19: More Coffee GIF

Who I am

- Studied Political Science at University of Regensburg
- Worked as Research and Teaching Assistant for Prof. Dr. Walter-Rogg (responsibility for one course about SNA and an introduction to Computational Social Science)

Who I am

- Currently: Graduate Student of Computational Social Science at Linköping University

- Beyond that: Graduate Research Assistant at Institute for Analytical Sociology; part of the “Mining for Meaning: The Dynamics of Public Discourse on Migration” project

Fig. 20: Campus Norrköping

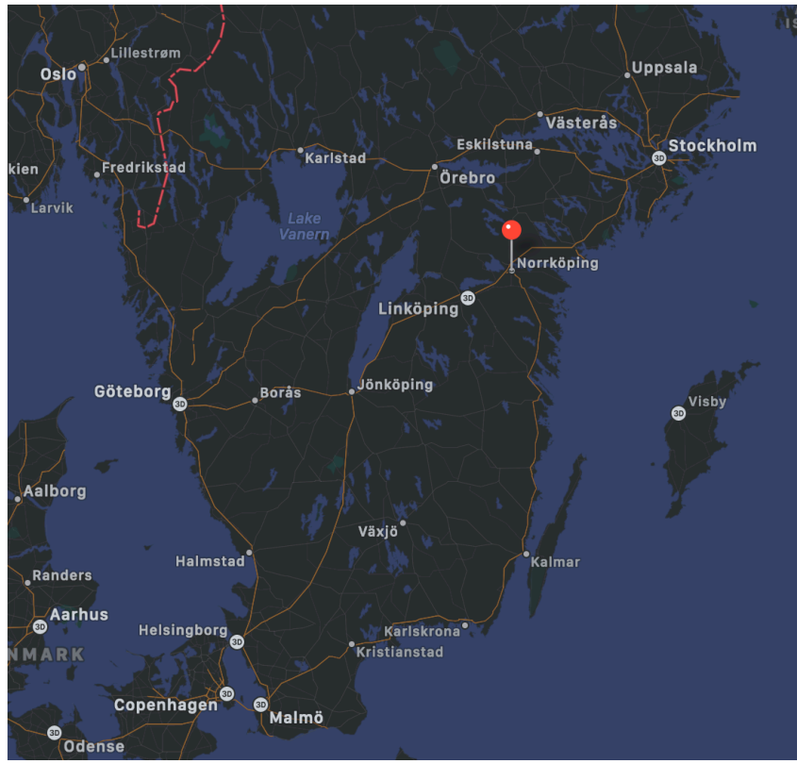


Fig. 21: On a map

Who I am

Computational Text Analysis

Computational analysis offers new ways to derive meaning from text. Our project "Mining for Meaning: The Dynamics of Public Discourse on Migration" uses large corpora of text as social sensors to measure what people feel, think, and talk about, which allows us to track the emergence of shared social understandings.

Fig. 22: What I am currently on…

Who I am

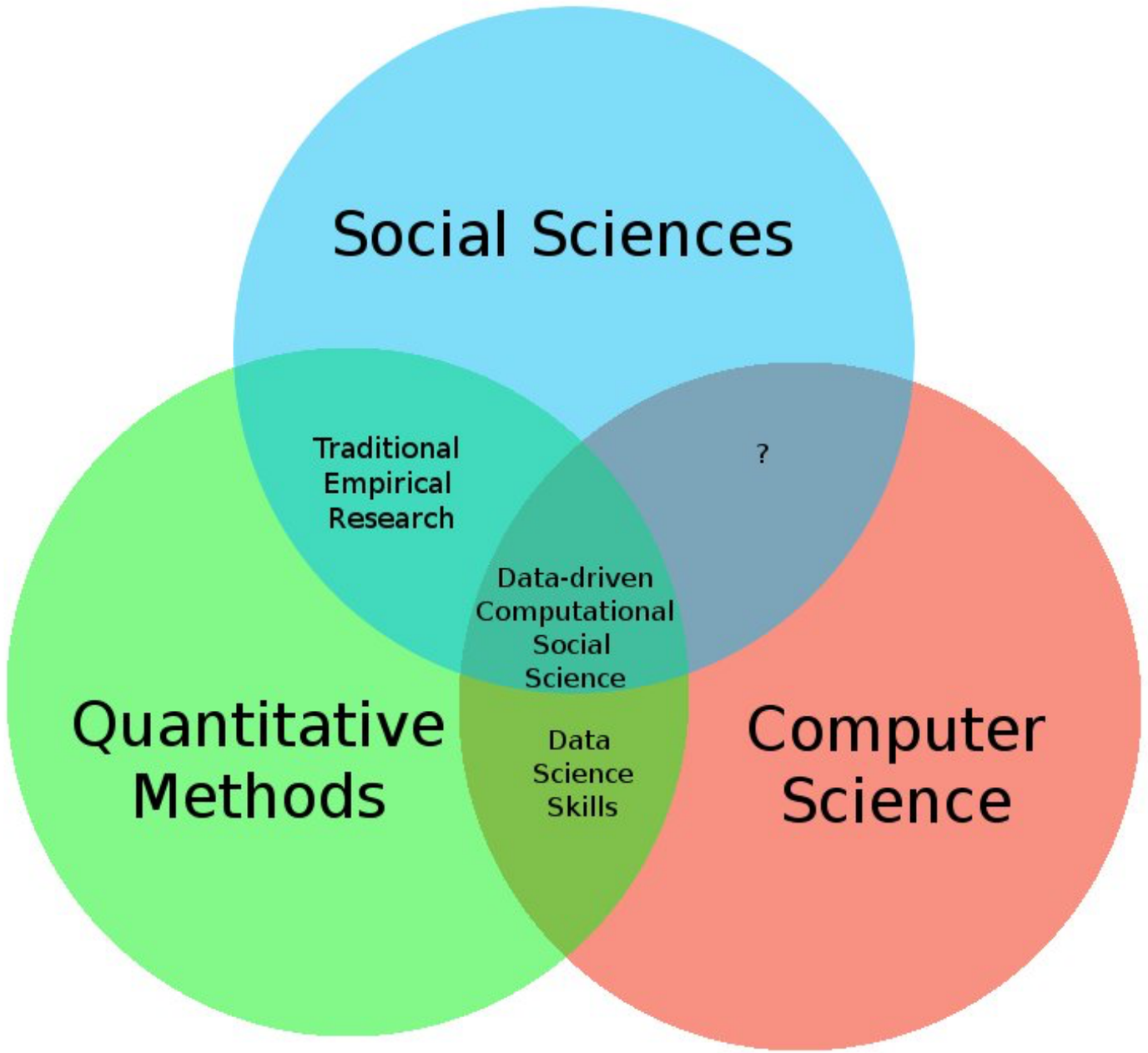


Fig. 23: CSS in a Venn diagram

How you can get in touch

felix.lennert@politik.uni-regensburg.de

Just send me an email so we can arrange a zoom meeting or whatever; I basically work every day, so feel free to reach out whenever

For the introductory R stuff you can do beforehand, I will arrange weekly “office hours”; I suggest Thursdays. However, an “official” announcement will follow which encompasses a Google Doc where you can sign up for them

Who you are—in (quantitative) terms of your R capabilities

let's go to R and check out how we would do it there…

Who are you?

- Activate your Webcam—so we can see you
- What are you studying?
- Where do your interests lie?
- How about your experience in terms of programming so far?
- Your expectations for the course?
- And, given the times: share your best hacks for coping with the lockdown.

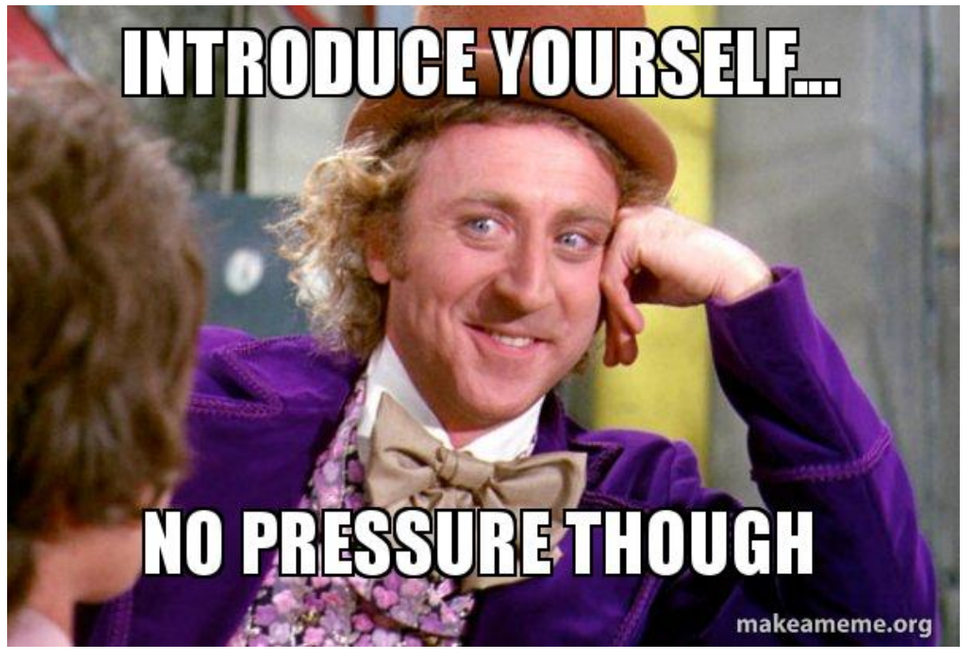


Fig. 24: Your turn

What are you going to learn?

80% of data analysis is *data wrangling*

From Wikipedia:
“**Data wrangling**, sometimes referred to as **data munging**, is the process of transforming and mapping data from one "raw" data form into another format with the intent of making it more appropriate and valuable for a variety of downstream purposes such as analytics. A **data wrangler** is a person who performs these transformation operations.” (Wikipedia 2020)

The good thing about data wrangling in R: it’s scalable. Hence, if your code works for small data, it will also work for “Big Data”

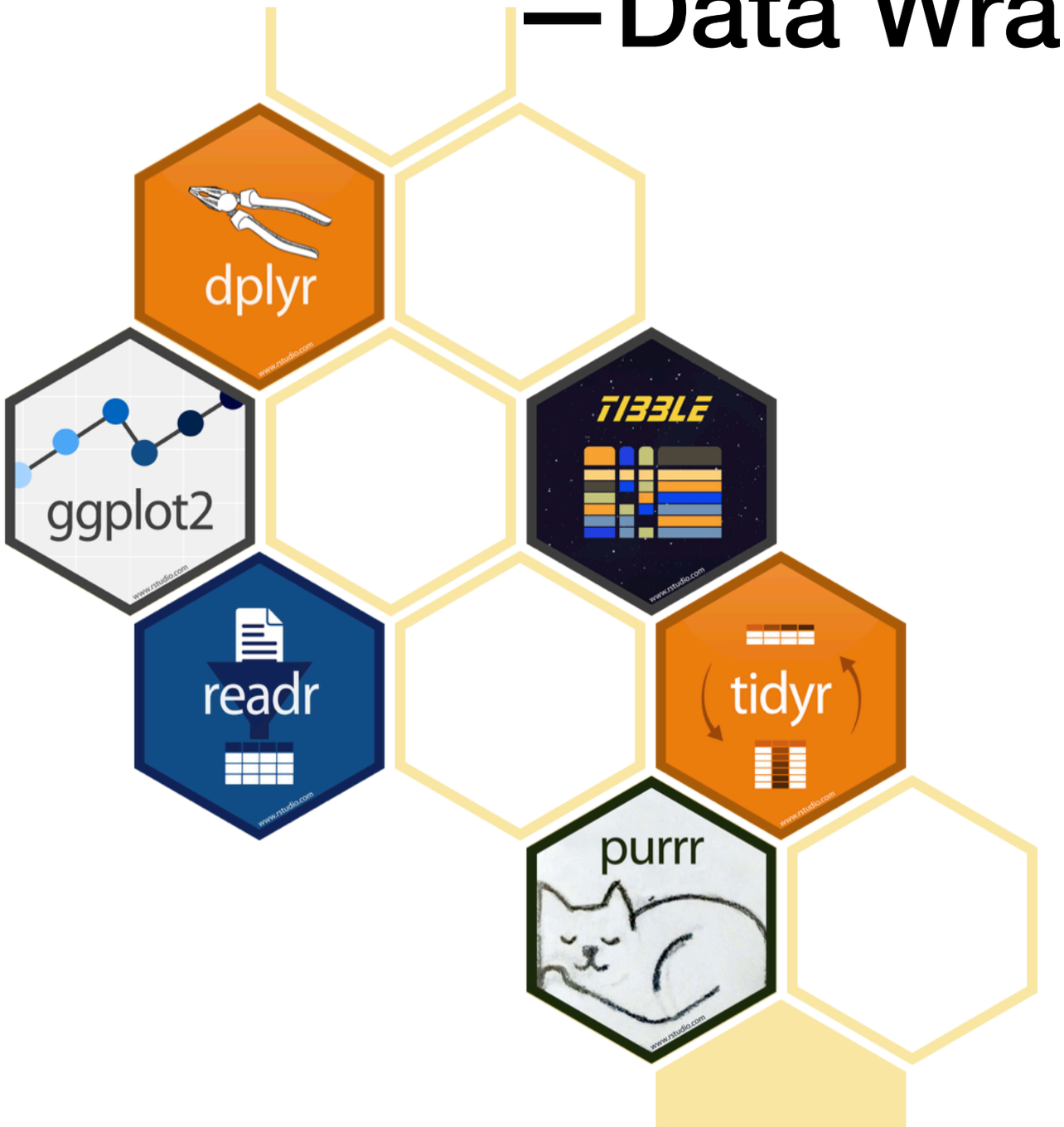
Felix Lennert, B.A.
felix.lennert@politik.uni-regensburg.de

What are you going to learn?
—in terms of R—

80% of data analysis is *data wrangling*

—> Approximately 80% of the R Stuff we are going to learn helps you with data wrangling

What are you going to learn?
—Data Wrangling—



- The probably most popular collection for packages for data wrangling in R

- “All packages share an underlying design philosophy, grammar, and data structures”

Fig. 25: The tidyverse

What are you going to learn?
—Data Visualization—

Fig. 26: ggplot2

- An enormously powerful tool for visualizing basically everything—as long as it is in tidy format

- Based on “The Grammar of Graphics” (Wilkinson et al. 2000)

What are you going to learn?
—Data Analysis—

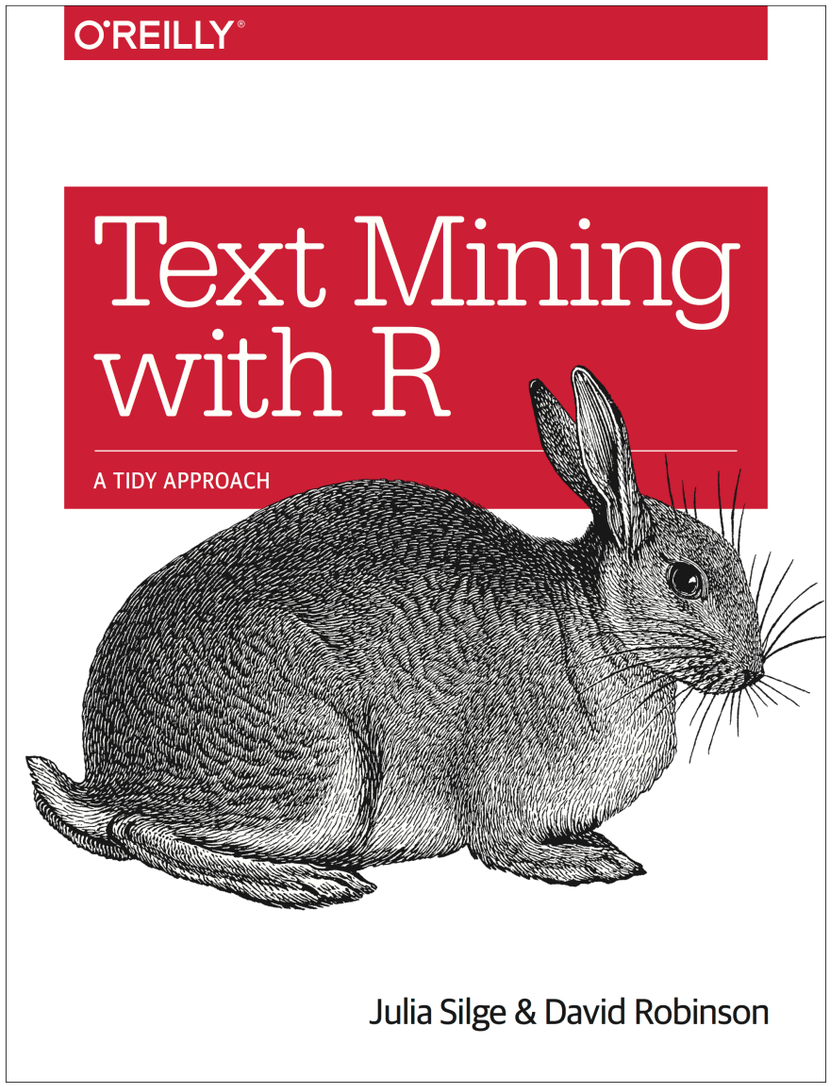


Fig. 27: Tidy text analysis

- You can extend the underlying philosophy of “tidy data” to, for instance, text
- This course will cover the basics of “tidy text mining”
  - bringing text into the right shape
  - sentiment analysis
  - topic modeling

What are you going to learn?
—Data Analysis—

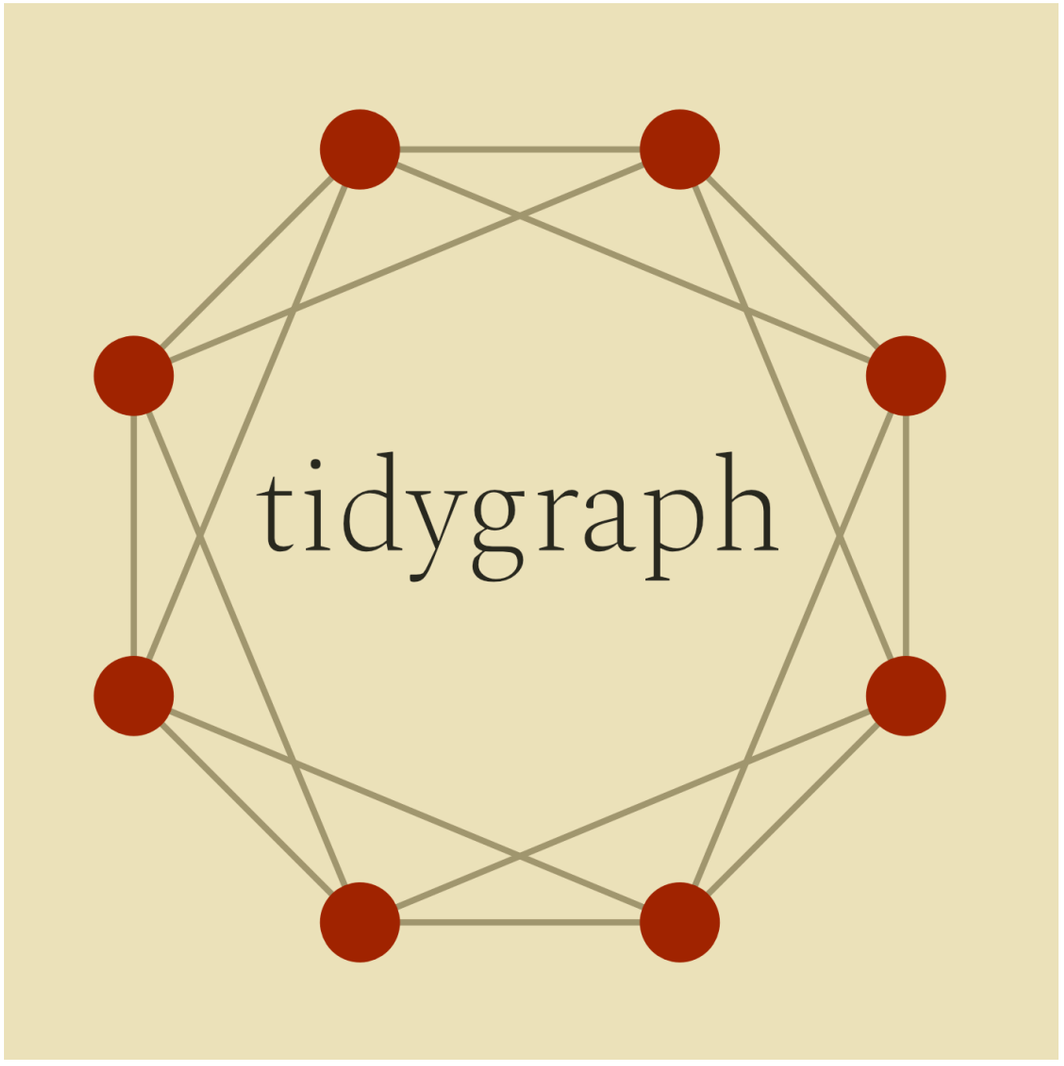


Fig. 28: Tidy SNA

- Social Network Analysis can also be performed using tidy data
- One lesson will be about doing exactly this
  - bringing data into the right shape (node- and edgelists)
  - basic visualization (with ggraph)
  - basic measures

What are you going to learn?
—an overview—

In general:

- **Mondays are for R coding**
- **Thursdays are for (hopefully fruitful) discussions**
- **(and every day of the week is for R practicing)**

What are you going to learn?
—in terms of R: an overview—

08.06.2020	Introduction to R: RStudio (projects, GitHub, RMarkdown); formats and classes; packages; basic operations (“R as a calculator”)
15.06.2020	Introduction to tidy data: the concept itself; import of data in diverse formats; the pipe
22.06.2020	Introduction to data wrangling: tibble, dplyr, and tidyr
29.06.2020	Introduction to data visualization: ggplot2—first take-home exam (deadline yet to be announced)
02.07.2020	More advanced programming with R: iteration; functional programming; conditions; purrr
09.07.2020	Social Network Analysis: tidygraph and ggraph
13.07.2020	Text Mining: tidytext—second take-home exam (deadline: end of the semester—i.e., 30.09.2020)

What are you going to learn?
—what are we going to talk about?—

11.06.2020	Fronleichnam: no meeting
18.06.2020	Up- and downsides of Big Data
25.06.2020	Data biases and how to address them
09.07.2020	Social Network Analysis and Text Mining: brief theoretical introductions
16.07.2020	Wrap up: what can you do now; what should you do next; feedback

What are you going to learn?
—our main resources—

Remember: one of the best things about R is its community
—> this, for instance, indicates that many resources are free

Fig. 29: The intro

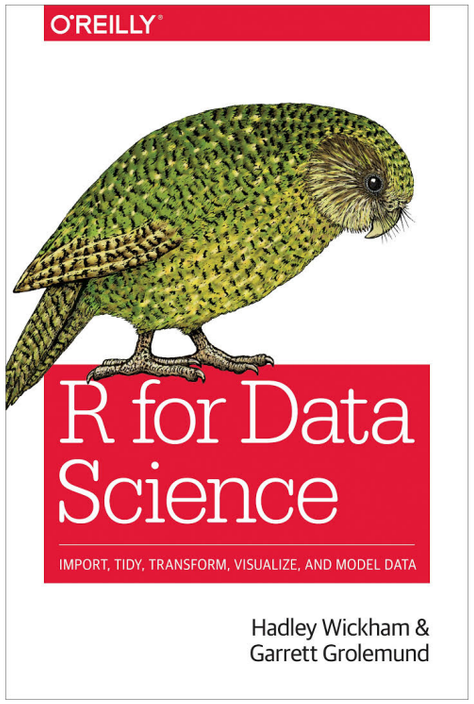


Fig. 30: The bible

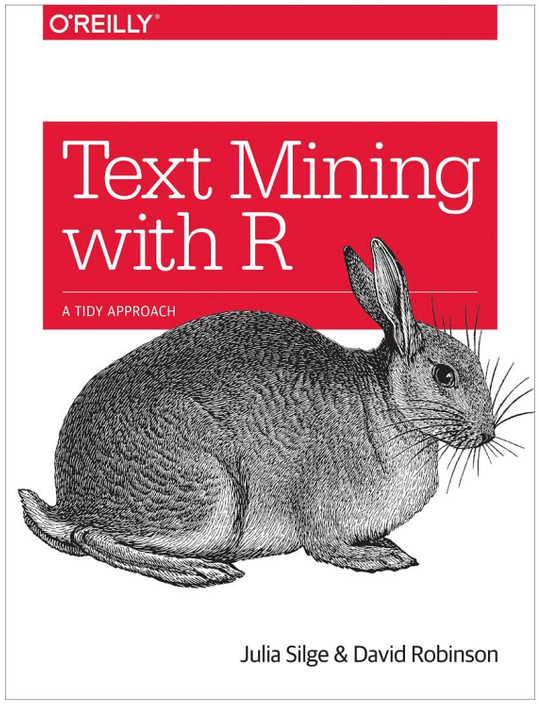


Fig. 31: Text analysis

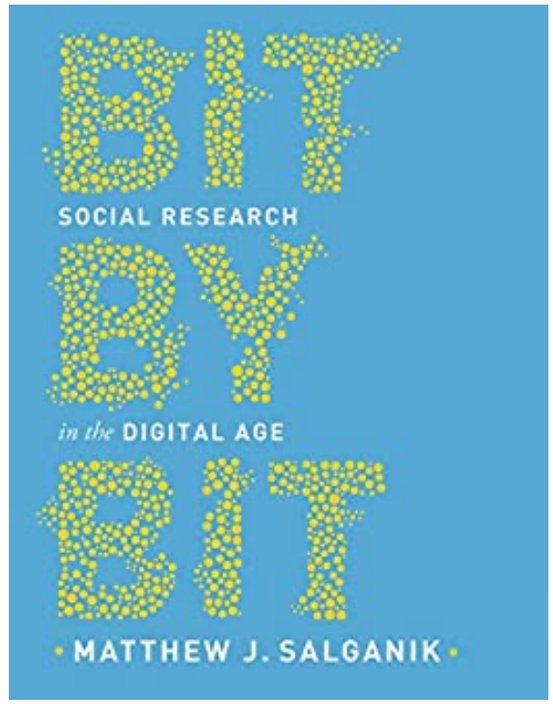


Fig. 32: Theoretical foundations

How are we going to go about it?
—let's make a deal—

- The hardest thing for me about the lockdown is the fact that I really struggle to find motivation (even though I get paid money for doing those things)

- Our class is fairly diverse in terms of proficiency in R

- The hardest thing for some of you might be getting their head around the "basics" of R; the hardest thing for some of you might be not being completely bored out in the beginning

How are we going to go about it?
—let's make a deal—

So here's the deal:

- I upload the scripts with the things we are going to learn every week—with exercises
- Nobody is forced to work on them immediately, but s/he will have to do so at latest in the week after we have covered it in the seminar
- I will provide office hours every week (exact dates are yet to be announced, also depends on the other course I offer) for the ones who have worked on the material

This is good because…

…it forces me to do something on a weekly basis (if not, I might end up preparing your stuff the night before the respective class)

…it enables you to study on a more decent/your self-chosen pace

…we can use our precious sessions for questions/clarifications etc. and may be able to spend less time on explanations

How are we going to go about it?
—next steps (mostly notes to myself…)—

I will…

- Upload a more detailed syllabus (including literature for every session, deadlines, etc.)
- Create a Google Doc (basically a spreadsheet) where you can sign up for the weekly “office hours” on zoom
- Create R Scripts and upload them on a weekly basis (the order will be akin to the one in the syllabus)

I ask you to…

- Sign up for the course on FlexNow (deadline: 26.04.2020)
- Check out the scripts
- Provide feedback on them (if possible)
- Check in with me if there are any concerns/questions/etc.
- Take the utmost care of you and others

Questions?

Übung: “Big Data Analysis” Using R
Felix Lennert, B.A.
felix.lennert@politik.uni-regensburg.de

Figures 1/2

Fig. 1: Bail, Christopher n. d.: An Introduction to Text As Data. URL: https://cbail.github.io/textasdata/strengths-weaknesses/rmarkdown/Strengths_and_Weaknesses.html, last access: 04/22/2020.

Fig. 2: The Economist 2010: Cover. URL: https://www.economist.com/node/21521550, last access: 04/22/2020.

Fig. 3: The Economist 2017: Cover. URL: https://www.economist.com/printedition/covers/2017-05-04/ap-e-eu-la-me-na-uk-0, last access: 04/22/2020.

Fig. 4: Kemp, Jeremy 2007: Gartner Research's Hype Cycle diagram. URL: https://commons.wikimedia.org/wiki/File:Gartner_Hype_Cycle.svg?uselang=de, last access: 04/22/2020.

Fig. 5: Van Smeden, Marten 2018: „In the last couple of years I've met many PhD student and early career researchers in doubt or discouraged to learn a statistical programming language (R, Python, Julia, …). I've made a list with 10 reasons to start programming [THREAD].“, Tweet, URL: https://twitter.com/MaartenvSmeden/status/995791001825431552, last access: 25.01.2019.

Fig. 6: R Memes for Statistical Fiends 2018: manowaR! Meme sent in by @Ramush Çeku ~Rashnutin. URL: https://www.facebook.com/Rmemes0/photos/a.1230204967031792/2158411574211122/?type=3&theater, last access: 01/28/2019.

Fig. 7: Smith, David 2017: CRAN now has 10,000 R packages. Here's how to find the ones you need. URL: https://blog.revolutionanalytics.com/2017/01/cran-10000.html, last access: 04/22/2020.

Fig. 8: Smith, David 2017: The Impressive Growth of R. URL: https://stackoverflow.blog/2017/10/10/impressive-growth-r/, last access: 04/22/2020.

Fig. 9: Smith, David 2017: The Impressive Growth of R. URL: https://stackoverflow.blog/2017/10/10/impressive-growth-r/, last access: 04/22/2020.

Fig. 10: Muenchen, Robert 2019: The Popularity of Data Science Software. URL: http://r4stats.com/articles/popularity/, last access: 04/22/2020.

Fig. 11: Find them here: https://twitter.com/search?q=%23rstats&src=typed_query.

Fig. 12: R-LADIES GLOBAL 2020: No title. URL: https://rladies.org/, last access: 04/22/2020.

Fig. 13: Tidy Tuesday 2020: No title. URL: https://www.tidytuesday.com, last access: 04/22/2020.

Fig. 14: Davenport, Thomas/Patil, D.J. 2012: Data Scientist: The Sexiest Job of the 21st Century. URL: https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century, last access: 04/22/2020.

Fig. 15: Muenchen, Robert 2019: The Popularity of Data Science Software. URL: http://r4stats.com/articles/popularity/, last access: 04/22/2020.

Fig. 16: Munzert, Simon/Rubba, Christian/Meißner, Peter/Nyhuis, Dominic 2015: Automated Data Collection with R, Chichester: John Wiley & Sons. pp. xviii–xix

Figures 2/2

Fig. 17: LSF 2020: Big Data Analysis with R - Einzelansicht. URL: https://lsf.uni-regensburg.de/qisserver/rds?state=verpublish&status=init&vmfile=no&publishid=147856&moduleCall=webInfo&publishConfFile=webInfo&publishSubDir=veranstaltung&expand=0#auswahlBaum%7Cabschluss%3Aabschl%3D82%7Cstudiengang%3Astg%3D129%7CstgSpecials%3Apordnr%3D9383%2Cvert%3D999%2Cschwp%3D%C2%A0%2Ckzfa%3DA%7CkontoOnTop%3Apordnr%3D9385%2Cpversion%3D20122%7Ckonto%3Apordnr%3D9389%7Cveranst%3Averanstid%3D147856, last access: 04/22/2020.

Fig. 18: @theoriginaldonutshopcoffee 2018: Coffee Break Yes. URL: https://giphy.com/gifs/theoriginaldonutshopcoffee-fun-yes-vgzA1vS0cP7rW4oFtJ, last access: 04/22/2020.

Fig. 19: @stickergiant 2018: Coffee Time. URL: https://giphy.com/gifs/9rnJy5BnaN5wER7NRj, last access: 04/22/2020.

Fig. 20: Akademiska Hus 2017: Campus Norrköping fyller 20 år. URL: https://www.akademiskahus.se/aktuellt/nyheter/2017/03/campus-norrkoping-fyller-20-ar/, last access: 04/22/2020.

Fig. 21: Screenshot from Apple Maps.

Fig. 22: Linköping University 2020: Computational Text Analysis. URL: https://liu.se/en/research/computational-text-analysis, last access: 04/22/2020.

Fig. 23: Matter, Ulrich 2013: Data Science in Business/Computational Social Science in Academia?. URL: http://giventhedata.blogspot.com/2013/03/data-science-in-businesscomputational.html, last access: 04/22/2020.

Fig. 24: Makeameme o. J.: INTRODUCE YOURSELF…- NO PRESSURE THOUGH. URL: https://makeameme.org/meme/introduce-yourself-no, last access: 04/04/2019.

Fig. 25: Wickham, Hadley et al. 2019: Welcome to the tidyverse. Journal of Open Source Software, 4(43): 1686.

Fig. 26: Wickham, Hadley 2016: ggplot2: Elegant Graphics for Data Analysis. New York: Springer-Verlag.

Fig. 27: Silge, Julia/Robinson, David 2017: Text Mining with R. Sebastopol et al.: O'Reilly Media.

Fig. 28: Pedersen, Thomas 2018: tidygraph 1.1 – A tidy hope. URL: https://www.data-imaginist.com/2018/tidygraph-1-1-a-tidy-hope/, last access: 04/22/2020.

Fig. 29: Cotton, Richard 2013: Learning R. Sebastopol et al.: O'Reilly Media.

Fig. 30: Wickham, Hadley/Grolemund, Garrett 2017: R for Data Science. Sebastopol et al.: O'Reilly Media.

Fig. 31: Silge, Julia/Robinson, David 2017: Text Mining with R. Sebastopol et al.: O'Reilly Media.

Fig. 32: Salganik, Matthew 2018: Bit By Bit. Princeton: Princeton University Press.

Literature

- Cotton, Richard 2013: Learning R. Sebastopol et al.: O'Reilly Media.
- Magnusson, Måns/Öhrvall, Richard/Barrling, Katarina/Mimno, David 2018: Voices from the Far Right: A Text Analysis of Swedish Parliamentary Debates. Preprint. SocArXiv.
- Salganik, Matthew 2018: Bit By Bit. Princeton: Princeton University Press.
- Silge, Julia/Robinson, David 2017: Text Mining with R. Sebastopol et al.: O'Reilly Media.
- Wickham, Hadley/Grolemund, Garrett 2017: R for Data Science. Sebastopol et al.: O'Reilly Media.

Felix Lennert, B.A.
felix.lennert@politik.uni-regensburg.de