Syllabus for "Big Data Analysis with R" (33335j)

Felix Lennert

Winter term 2020/21

Contact

As the entire course will take place online, we will only be able to communicate through our laptops and phones. This is fairly unfortunate, especially because I cannot simply do some troubleshooting on your computer or anything similar.

- E-mail: felix.lennert@liu.se – please use my LiU mail address for whatever inquiries you have. I have given the Uni Regensburg mail multiple shots, but I simply cannot get it to work with several mail clients on my computer. Hence, to make sure that mails reach me within a reasonable time, use the address stated here.
- Website: Everything will be put on *GRIPS*.
- Office hours: on Fridays, 10–11; will take place online. Please sign up *here* beforehand. If no one's signed up, I will not open the Zoom room. If you cannot make it to one of the office hours, just drop me an e-mail and we set something up. This especially applies to students who live in different time zones. The Zoom room has the meeting ID *911 0279 5820*. The password is "office". Join meeting using this *link*.
- Class hours: Zoom meetings every other Thursday, 16–18, join meeting *here* – meeting ID is *972 9681 0305*, password is "R". The rest of the sessions consists of pre-recorded videos.

Course description

"[. . .] when you think about social research in the digital age, you should not just think *online*, you should think *everywhere.*" (Salganik 2017: 5, emphasis in original)

According to Google Trends, that use "Big Data," the search term "what is big data" was not really in people's scope—right until about 2011/2012. Then public's interest really took off, peaked in spring 2017 and now flattens off. What does that tell us? We do not know. First learning: Big Data itself is not a universal answer if there are no questions.

But since everybody is talking about it, what is Big Data actually? According to the US National Institute of Standards and Technology "Big data and data science are being used as buzzwords and are composites of many concepts." (NIST 2015: 2) The probably most cited concept states that its features are the three Vs: Volume, Velocity, Variety. What do they imply? Volume relates to the sheer size. Usual Laptops are not suitable to analyze those datasets. Sometimes it is stated that data can be considered big as soon as it does require more than one usual computer to store and analyze it. Velocity refers to the rate at which it is produced. These days, everybody leaves digital traces everywhere, or, as Lazer et al. (2009) put it, "we live life in the network" (Lazer at al. 2009: 721) and all these little digital breadcrumbs we scatter while leading our lives can be analyzed. Akin to that is data's variety. Since most of it is "found data" that has not been collected for research purposes, researchers have to make increasing efforts reshaping their material to meet their research requirements. Now you may wonder what the upsides of Big Data are. A more recently introduced new V might give a quick glimpse of that: veracity. Because people are mostly not aware of the extent of the data they produce, it is presumably rather trustworthy. Second learning: you, prospective

student, are always prone to become a guinea pig for hot new social science research.

Data Scientist is, according to the Harvard Business Review, "The Sexiest Job of the 21st Century" (Davenport/Patil 2012). One of their native languages is R. Hence, it follows naturally that we are going to take two directions when talking about Big Data. On the one hand, the course is going to cover data manipulation in R up to an intermediate level. On the other hand, we are going to talk about the implications of Big Data on social sciences in general and Political science in particular. In the end, we may find a synthesis by replicating two papers with the original data material together. That is going to be your third learning: As impressive as the terms AI and Unsupervised Machine Learning may sound – in the end, the computer is nothing more than a bare tool and relies heavily on its user's capabilities

As you may have already noticed, our language of instruction is going to be English. Since I assume that we are all not going to be native speakers, that should not do any harm and nobody has to feel embarrassed if their command is not top-notch. Former experience in R, Python, or even SPSS is not required yet comes in handy.

Course Objectives

The course is mainly designed as a computer course. But working with some data without context might be pretty boring (and where do the data come from? What are their limitations?). Therefore, we will put it into a bigger perspective together, trying to explore what new resources the rise of the world wide web and digital devices has brought to us, social scientists.

R

Introduction to R

This course will provide you with an introduction to the R language. R is a programming environment for statistical analyses and graphics. Bonus: it's free. We will code in R using the IDE RStudio which is free as well. R and RStudio have some neat features, such as RMarkdown[1], RStudio Projects, or its GitHub GUI. They will be introduced as well.

Introduction to data wrangling with the `tidyverse`

The data this course is about is seldom custom-made and clean, most of the time it is dirty, raw, organic data. Hence, before you can draw inferences from it, you need to get it neat and clean. The so-called `tidyverse` is an incredibly versatile collection of R packages to work with data in a *tidy* manner. First, you will get an introduction to the theory behind "tidy data." Afterward, you will "get your hands dirty" and work with some data sets I will provide you with using the `tidyverse` packages.

Visualization using `ggplot2`

Quantitative information is usually reported using graphs. You will learn how to accomplish such tasks using `ggplot2`.

Functional programming in R

Copy-Pasting code is incredibly tedious and will, almost inevitably, lead to errors. You will learn techniques to automate your processes by using functions, loops, and flow-control.

[1] which was used for writing this syllabus BTW.

The new forms of data

Implications of big data

What kinds of data do we have at our hands? What is there for us? What are pitfalls and how can we avoid them?

How can we draw inferences?

Two analysis techniques suitable for the analysis of these new types of data are introduced: text mining and social network analysis. You will get brief introductions and demonstrations in R.

Commented literature list

The course's foundation is two books: Hadley Wickham's and Garrett Grolemund's "R for Data Science" and Matthew Salganik's "Bit By Bit." Both can be obtained online.

- Salganik, Matthew 2017: *Bit By Bit. Social Research in the Digital Age*. Princeton/Oxford: Princeton University Press.

Matthew Salganik is a professor at Princeton and one of the founders of the Summer Institute in Computational Social Science. His book provides you with an overview of the new possibilities the "Digital Age" has in stock for Social Scientists. It can be read online, so you do not have to obtain a copy yourself.

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

Hadley Wickham is the founder of the `tidyverse` and the chief scientist of RStudio. He "is a legend in the data science field for having invented a completely new way of doing data analysis that no one had thought of before." (Roger Peng) You will become familiar with his approach. The book can be read online.

Course Policies

Below you can find some basic rules of behavior for the course and what you will have to do to pass it.

Basic rules of behavior

- Mute your microphone during Zoom lectures at all times – except for when you want to say something.
- Do not hesitate to interrupt me whenever.
- Never discriminate any of your classmates due to whatever.
- If anything course-related bothers you, send me an e-mail.
- Feel free to copy code from Stackoverflow or whatever resources pop up in your google searches. I do not consider this cheating. If anybody considers this cheating, they have probably never coded themselves. You are, however, responsible for the outcome, so make sure the code does what you want to accomplish.

Attendance policies

This semester attending class is probably easier than ever – in theory, you can just stay in bed. However, if you are not able to make it to class due to technical problems, send me an e-mail. I can provide you with information on how to dial in via your cell phone. However, you will only have to participate every other week, since Zoom sessions with theoretical content and pre-recorded R sessions will alternate.

For the (more theory-heavy) Zoom sessions, I expect you to have at least a rough idea of what we are going to talk about. I do not expect you to have understood the literature in depth. I rather aim to spark you interest for how research in the 21st century has changed – due to the data newly available.

In weeks with no theoretical sessions, you are supposed to work through R-related videos and accompanying scripts on your own. At the end of the scripts, you will find accompanying exercises. Please hand them in before the subsequent theoretical session. After the deadline has passed, I will release a script with a suggested solution and a video going through it step-by-step.

Due to the fact that I am no full-time teacher, I cannot and will not provide extensive and individual feedback on your solutions. If questions remain, we can talk about them during office hours (find instructions regarding office hours in the first section). If you cannot make it to office hours, I can make time for your inquiries sometime else – just shoot me an e-mail and we can schedule a meeting. Please note that there will be no in-person meetings due to the current pandemic situation.

These assignments with mandatory deadlines are not supposed to torture you. When learning how to code, practice is key – similar to learning a new language. The reason for my decisions to make practicing mandatory comes from my experience teaching the course during the former summer term: people were dropping out one after the other because they had lost track of the R content. Applying external pressure might prevent you from that. Furthermore, without proper feedback it was incredibly hard for me to make the adaptations necessary to enhance their learning experience. Hence, I am also going to ask you to fill out a brief (anonymous) survey after every R session. Please do always contribute there, not only if you disliked the session. The quality of the handed-in exercises then enables me to make sure that each of you has actually understood the contents – and not only claims it. If the videos and sessions do not work for a significant number of you, we can discuss and decide upon different solutions.

E-mail Policy

E-mails will probably be our main means of communication. As my time is limited, please stick to the following rules. Before sending an e-mail, consider the following problem-solving strategies:

- Give this syllabus a second look.
- Google your problem. Try at least three solutions from Stackoverflow.
- Give the script I provided you with (and its references) a second look. I take most of the inspiration for the exercises from the resources I use for writing the script. Hence, the solution should be in there.

If you were not able to solve the problem by following the aforementioned strategies, please stick to the following rules:

- Start the mail's subject with *bdawr* – that's how I know which e-mails need to be answered first ;-).
- If there is an R-related problem, try to provide me with a *reprex*.
- Please try to be concise about your problem.

Examination

I will not distinguish between what program you are in or what module you chose the course for. There are a couple of things you will have to hand in:

- The biweekly R assignments. Handed in as .R or .RMD files – more on that in the first R session. Please do them in pairs, it will make the working on them more fun and you connected to others throughout this lonely winter.
- A final data challenge. It will be circulated after the end of the regular sessions. You can team up with one of your peers. The deadline for the will be the end of the semester (i.e., 2021-03-31). While the biweekly R assignments can be solved by simply coding, passing the final challenge will require some sort of story-telling. Analyzing data means formulating a question, digging into datasets to find traces for answering it, and then drawing decent pictures (graphs) and/or tables to bring the story to the reader. Both data challenges need to be handed in as an RMD and a PDF file. Everything I will need to knit it must be included. This includes, for instance, .bib files, data sets – in .csv format, and .csl files.
- An article review. As the course also has a theoretical component, you will be randomly assigned an article that you will then have to review following Maurice Zeitlin's *The Four Questions*. The deadline,

again, will be the end of the semester (i.e., 2021-03-31). The assignment will take place in the final session.

For the assignments that are due on 2021-03-31, you will get a one-week extension pretty easily. Just drop me an e-mail, saying "I won't make it," and provide me with whatever reason – "...due to my poor time-management" suffices. If you need more time beyond this one week, you will have to provide me a valid reason.

Class Schedule

As mentioned earlier, the course will follow a biweekly structure – theory, R, theory, R, etc....

2020-11-05: Kick off!

- The "data deluge"
- Why coding?
- Chores (FlexNow, etc.)
- Getting to know each other
- Course objectives

2020-11-12: Let's get R started!

- Introduction to R, RStudio, RStudio Projects, RMarkdown, GitHub
- R as a scientific calculator
- All about vectors
- Introducing the *Tibble*

Accompanying readings:

- Cotton, Richard. 2013. Learning R. First Edition. Beijing ; Sebastopol, CA: O'Reilly.
- Wickham, Hadley and Garrett Grolemund. 2016. R for Data Science: Import, Tidy, Transform, Visualize, and Model Data. Sebastopol et al.: O'Reilly. Chapters 2, 4, 6, and 16. (–> available *online*)
- Xie, Yihui, J. J. Allaire, and Garrett Grolemund. 2018. R Markdown: The Definitive Guide. Boca Raton: Taylor & Francis, CRC Press. (–> available *online*)

2020-11-19: The new data

Before wrangling data, let's talk about them!

- What is Big Data?
- What is new now?
- Readymade – Custommade/Organic – Designed
- Observing behavior and digital trace data

Readings:

- Salganik, Matthew 2017: *Bit By Bit. Social Research in the Digital Age*. Princeton/Oxford: Princeton University Press, pp. 13–84. (–> available *online*)
- Brady, Henry 2019: The Challenge of Big Data and Data Science. In *Annual Review of Political Science 22(1)*, pp. 297–323.
- Lazer, David, et al. 2009: Life in the network: the coming age of computational social science. In *Science 323(5915)*, pp. 721–723.
- Lazer, David/Radford, Jason 2017: Data ex Machina: Introduction to Big Data. In *Annual Review of Sociology 43(1)*, pp. 19–39.

Try to obtain the readings yourself first. If you fail, drop me a mail.

2020-11-26: Tidy data!

- The concept (read: Wickham, Hadley 2014: Tidy Data. *Journal of Statistical Software 59(10)*. Obtain it *here*)

But before we can make a data set tidy...

- Getting data into R: `readr`, `haven`, `readxl`
- How to store it: let's revisit `tibble`
- A data science conjunction: `magrittr` and the pipe
- Making data tidy: `tidyr`
- A general look at the different `tidyverse` packages

Accompanying readings:

- Wickham, Hadley and Garrett Grolemund. 2016. R for Data Science: Import, Tidy, Transform, Visualize, and Model Data. Sebastopol et al.: O'Reilly. Chapters 7, 8, 9, and 14. (–> available *online*)

2020-12-03: Potential data flaws

- Data biases
- How can we address them?

Readings:

- Brady, Henry 2019: The Challenge of Big Data and Data Science. In *Annual Review of Political Science 22(1)*, pp. 297–323.
- Ruths, Derek/Pfeffer, Jürgen 2014: Social Media for Large Studies of Behavior. In *Science 346(6213)*, pp. 1063–1064.
- Stevens-Dawidowitz, Seth 2014: The cost of racial animus on a black candidate: Evidence using Google search data. In *Journal of Public Economics 118*, pp. 26–40.

2020-12-10: A dive into the data!

- `dplyr`
- `lubridate`
- `forcats`

This session will be intense and many things will be overwhelming at first. However, bear with me, do the exercises I provide you, and feel more than free to reach out for help! Deadline for the assignment is the first session *after* christmas (i.e., 2021-01-07).

Accompanying readings:

- Wickham, Hadley and Garrett Grolemund. 2016. R for Data Science: Import, Tidy, Transform, Visualize, and Model Data. Sebastopol et al.: O'Reilly. Chapters 3, 12, and 13. (–> available *online*)

2020-12-17: Easier before Christmas

Let's discuss, present, and discuss a couple of papers in breakout rooms. You will be assigned to groups at the end of the meeting on 2020-12-03.

Papers for the discussion:

- Garcia, David/Rimé, Bernard 2019: Collective Emotions and Social Resilience in the Digital Traces After a Terrorist Attack. *Psychological Science 30(4)*, pp. 617–628. http://dx.doi.org/10.1177/0956797619831964.
- González-Bailón, Sandra/Banchs, Rafael/Kaltenbrunner, Andreas 2012: Emotions, Public Opinion, and U.S. Presidential Approval Rates. A 5-Year Analysis of Online Political Discussions. *Human Communication Research 38(2)*, pp. 121–43. http://dx.doi.org/10.1111/j.1468-2958.2011.01423.x.

- Halpern, Daniel/Gibbs, Jennifer 2012: Social media as a catalyst for online deliberation? Exploring the affordances of Facebook and YouTube for political expression. *Computers in Human Behavior (2012)*, pp. 1–10. http://dx.doi.org/10.1016/j.chb.2012.10.008.

2021-01-07: Visualize your data!

- `ggplot2`

Note: For this session, it is crucial to have thoroughly understood the concepts introduced in the sessions before. In data analysis, visualization comes in the vast majority of cases after the wrangling, meaning that the data set is tailored towards the visualization function.

Accompanying readings:

- Wickham, Hadley and Garrett Grolemund. 2016. R for Data Science: Import, Tidy, Transform, Visualize, and Model Data. Sebastopol et al.: O'Reilly. Chapter 1. (–> available *online*)
- Chang, Winston 2013: R Graphics Cookbook. Sebastopol et al.: O'Reilly.

2021-01-14: Introduction to Social Network Analysis

First half: a quick theoretical introduction to SNA

- Readings:
  - Borgatti, Stephen/Mehra, Ajay/Brass, Daniel/Labianca, Giuseppe, 2009: Network Analysis in the Social Sciences. In *Science (323)*, pp. 892–895. http://dx.doi.org/10.1126/science.1165821.
  - Marin, Alexandra/Wellman, Barry 2011: Social Network Analysis: An Introduction. In: Scott, John/Carrington, Peter (eds.): *The SAGE Handbook of Social Network Analysis.* London et al.: SAGE Publications Ltd, pp. 12–25. (–> will be uploaded to GRIPS)
  - Watts, Duncan 2004: The "New" Science of Networks. In *Annual Review of Sociology 30(1)*, pp. 243–270. https://doi.org/10.1146/annurev.soc.30.020404.104342.

Second half: doing it in R

- Analyzing networks the tidy way with tidygraph
- Visualization with ggraph

2021-01-21: Introduction to Text Mining

First half: a quick theoretical introduction to TM

- Readings:
  - Grimmer, Justin/Stewart, Brandon 2013: Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. In *Political Analysis (2013)*, pp. 1-–31. https://doi.org/10.1093/pan/mps028.
  - Silge, Julia/Robinson, David 2019: *Text Mining with R. A Tidy Approach.* Sebastopol et al.: O'Reilly Media. (–> can be read online)

Second half: doing it in R

- wrangling text: how "tidy" principles facilitate text mining
- sentiment analysis using AFINN
- basic topic modeling

2021-01-28: How you could go beyond

Let's get into functional programming and some a bit more advanced R things!

- What is functional programming?

- How and when to write a `function`?
- `for`- and `while`-loops – apply functions iteratively
- the `purrr` package

Readings:

- Wickham, Hadley 2019: Advanced R. Boca Raton: CRC Press, pp. 205–207. (–> can be read online)
- Wickham, Hadley and Garrett Grolemund. 2016. R for Data Science: Import, Tidy, Transform, Visualize, and Model Data. Sebastopol et al.: O'Reilly. Chapters 15 and 17. (–> available *online*)

2021-02-04: Wrap-up time!

- Feedback
- What should you do next?