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

Felix Lennert

Summer term 2020

Contact

As the entire course will take place online, we can only 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: **TBA** – my website is currently work in progress. As soon as it is ready, I will make an announcement. You will then find the course materials there. Besides, everything will be put on GRIPS as well.
- 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. The Zoom Room has the meeting ID: 911 0279 5820, password: office. Join meeting here

- Class hours: Mondays and Thursdays, 14-16; Join meeting here – meeting ID: 972 9681 0305, password: bigdata

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. However, 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 adventure 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 this completely new 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.

Attendance policies

This summer attending class is probably easier than ever – 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.
Mondays are for R coding. I will show you some stuff on my laptop and am very happy to answer your questions. I will not give you time to try stuff yourself, please do that afterward. In exchange, some sessions

might not last the entire 1.5 hours. At the end of the scripts, you will find exercises that are related to the things you will have heard in class. I am more than happy to have a look at your solutions and talk about them during office hours (which can, of course, be extended as well). Please contact me beforehand and include your code.

If you, for whatever reason, miss Monday's session, you are obliged to hand in the exercises until the following R session.

I do not care if you do not hand in anything every week. However, you will be obliged to hand in two take-home assignments.

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 three things you will have to hand in:

- Two data challenges. The first one will be circulated after four R sessions (i.e., on 2020-06-29). From then on, you have one week to finish it. You can team up with one of your peers. The second one will be given to you at the end of our last R session (i.e., 2020-07-13). The deadline for the latter is the end of the semester (i.e., 2020-09-30). While the first one can be solved by simply coding, for the latter one some sort of story-telling is necessary. 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 tell 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 obliged to choose an article and review it following Maurice Zeitlin's *The Four Questions*. The deadline is the end of the semester (i.e., 2020-09-30). The list of articles is yet to be announced.

For the assignments that are due on 2020-09-30, 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, R stuff generally takes place on Mondays, more substantial stuff on Thursdays. I will link the R resources in the R script (and later add links to the scripts as soon as my website is set up properly). The resources for the literature-heavy Thursday sessions can be found below.

2020-06-08

Let's get it started!

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

2020-06-11

Fronleichnam. Enjoy your day.

2020-06-15

Let's get into (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 Tibbles
- A data science conjunction: the pipe
- Making data tidy: tidyr
- A general look at the different tidyverse packages

This session might be shorter! If you have any questions related to R, pose them after the session! I want to make sure that we are all on the same page before we dive into the actual data wrangling.

2020-06-18

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.

- 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.

2020-06-22

Now we can 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!

2020-06-25

How can our data be flawed?

- 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.

2020-06-29

Visualize your data!

- ggplot2

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

Furthermore: Data challenge #1 is handed out after the session. Due on 2020-07-06.

2020-07-02

Social Network Analysis and Text Mining – brace yourself for two quick 45 minute lectures

Readings:

For Text Mining: * 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.
* Silge, Julia/Robinson, David 2019: *Text Mining with R. A Tidy Approach.* Sebastopol et al.: O'Reilly Media. (–> can be read online)

Social Network Analysis: * Borgatti, Stephen/Mehra, Ajay/Brass, Daniel/Labianca, Giuseppe, 2009: Network Analysis in the Social Sciences. In *Science (323)*, pp. 892–895.
* 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.

2020-07-06

TAKE-HOME ASSIGNMENT #1 DUE!

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

- What is functional programming?
- How 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)

2020-07-09

Even more R: looking into Text Mining with R

- wrangling text: how "tidy" principles facilitate text mining
- sentiment analysis using AFINN – an example from Sweden
- topic modeling

2020-07-13

Social Network Analysis with R

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

2020-07-16

We're done, wrap-up time!

- Feedback
- What should you do next?