Hertie School

Master of Data Science for Public Policy/Master of International Affairs/Master of Public Policy
Master of Public Policy
Fall Semester 2021
Course Syllabus, Version 22.09.2021

GRAD-C11/GRAD-E1339: Introduction to Data Science

(Elective *Concentration: Policy Analysis)*

Simon Munzert

1. General information

Class time	Thurs, 10-12h (MPP/MIA 2nd year elective group/PhD) Thurs, 14-16h (MDS core course)
Class room	MPP elective group: 2.61 MDS core course group: Forum
Class Format	Both groups will be taught in the hybrid format. That is, clases will be held onsite (pandemic conditions permitting) but also live-streamed via Zoom. Onsite attendance is regulated via a rotation scheme (MPP group). The event will be recorded published on Moodle (consent of all participants permitting).
Instructor	Prof. Simon Munzert
Instructor's office	3.13.1
Instructor's e-mail	munzert@hertie-school.org
Instructor's phone number	+49 (0)30 259 219 450
Assistant	Ayamba Kwoyila kwoyila@hertie-school.org +49 (0)30 259 219 121 3.10
Teaching assistants	Lisa Oswald, l.oswald@phd.hertie-school.org Tom Arend, t.arend@phd.hertie-school.org
Lab times	Group 1: Wed 8-9:30h, 2.61 (Lisa Oswald) Group 2: Wed 12-13:30h, 3.30 (Lisa Oswald) Group 3: Wed 18-19:30h, online (Tom Arend) Group 4: Wed 8-9:30h, 3.30 (Tom Arend) Group 5: Wed 12-13:30h, 2.34 (Tom Arend)
Instructor's Office Hours	Tuesdays, 15-16h Please email Ayamba Kwoyila in advance to make an appointment and to let me know what you would like to talk about.

Link to Module Handbook MDS, MIA and MPP
Link to Study, Examination and Admission Rules

Instructor Information:

Simon Munzert is Assistant Professor of Data Science and Public Policy at the Hertie School and member of the Hertie School Data Science Lab. His research interests include attitude formation in the digital age, public opinion, and the use of online data in social research. He is principal investigator in third party-funded projects on media exposure and opinion formation as well as on global public opinion on hate speech regulation. He received his Doctoral Degree in Political Science from the University of Konstanz.

TA information:

Lisa Oswald is pursuing a PhD in Governance at the Hertie School in Berlin. She graduated from the University of Oxford with a MSc degree in Social Data Science, and from the University of Kassel with a BSc and MSc degree in Psychology. She is interested in online communication and deliberation, the public perception of climate change, political opinion formation and the emergence of collective behavior.

Tom Arend is a second year PhD candidate at the Dynamics Research Training Group in Berlin (jointly organized by the Hertie School and HU). He obtained his Master in Politics and Public Policy from Sciences Po Paris, and his BA from Lancaster University. His research interests lie in comparative politics, where he focuses particularly on questions of accountability and representation in Western democracies.

2. Course Contents and Learning Objectives

Course contents:
This course will introduce you to the modern data science workflow with R. In recent years, data analysis skills have become essential for those pursuing careers in policy advocacy and evaluation, business consulting and management, or academic research in the fields of education, health, and social science. We will cover topics like version control (Git) and project management; data collection, wrangling, storage, and visualization; model fitting and simulation; advanced workflow issues, debugging, automation; and data science ethics. The course is intended for students with some experience in working with R.

Main learning objectives:
The goals are to (1) equip you with conceptual knowledge about the data science pipeline and coding workflow, data structures, and data wrangling, (2) enable you to apply this knowledge with statistical software, and (3) prepare you for our other methods electives and the master's thesis.

Software:
We will work with RStudio, RMarkdown, and Git/GitHub to implement and practice the learned techniques. It is assumed that you have some basic knowledge in using R. If not, we strongly encourage you to familiarize yourself with R prior to the course so to be able to focus on the substance of the course. Good introductions (with different types of focus) are:

1. Wickham, H., & Grolemund, G. (2017). R for Data Science. Available at: https://r4ds.had.co.nz/.
2. Larsen, E. G., & Fazekas, Z. (2019). Quantitative Politics with R. Available at: http://www.qpolr.com/.
3. Ismay, Ch., & Kim, A.Y. (2021). Statistical Inference via Data Science. A Modern Dive into R and the Tidyverse. Available at: https://moderndive.com/
4. DataQuest (https://www.dataquest.io/), in particular the following courses:
    a. Introduction to Data Analysis in R
    b. Data Structures in R
    c. Control Flow, Iteration, and Functions in R
    d. Data Cleaning in R

Target group:
MDS 1st year students and MPP/MIA 2nd year students, who have successfully completed Statistics II.

Labs:
Curricular Affairs will initially allocate you to labs. If you would like to switch labs, please use the switching partner forum on Moodle to connect with your fellow students. Once you find a switching partner, please fill in the form "switching course request" on MyStudies. It is your responsibility to make sure that by switching courses you do not create time clashes with your other courses.

Teaching style:
The sessions will mainly feature an interactive lecture on the session's topic led by the instructor. For the discussion of exercises and their implementation in R, students will be enrolled in a small group support lab session taught weekly by a teaching assistant.

Prerequisites:
In order to take this course as an elective, it is required that you have successfully completed Statistics II. (Taking Statistics II in parallel is not sufficient.) Basic command of R.

Diversity Statement:
We are passionate about creating an inclusive classroom atmosphere that values diversity. The R community lives these values, and we want you to become part of it. If you have any suggestions that contribute to this goal, we are always grateful for feedback.

3. Grading and Assignments

Composition of Final Grade:

Assignment	Due	Submission	Grade
Assignment 1: Homework assignments	Deadline: 11.59pm on the day before class	Submit via GitHub	4 x 10%
Assignment 2: Workshop session	Deadline: October 29 (materials), November 2-6 (workshop; exact date TBD)	Submit via GitHub + present live	25%
Assignment 3: Final data science project	Deadline: TBD	Submit via GitHub	35%

Assignment Details:
Evaluation is conducted via a combination of a series of homework assignments (counting towards 40% of the final grade), preparation and execution of a live workshop session (counting towards 25% of the final grade), and one data science project (counting towards 35% of the final grade). There will be no midterm and no final exam. While you should submit your own, individual solutions to the homework assignments, we generally encourage you to study and learn to use the software together.

Assignment 1: Homework assignments
In 5 homework assignments, you will apply the concepts learned in class to solve data analytic problem sets using R. While you are encouraged to collaborate, everyone will hand in a separate solution. The 4 best out of 5 assignments will contribute to the final grade. Grades will be based on (1) the accuracy of your solutions and (2) the adherence of a clean and efficient coding style that you will learn in the first sessions.

Assignment 2: Workshop session
About halfway through the semester, we will flip the roles: Students will become instructors. In the workshop "Tools for Data Science" you, in groups of two students, will present a pre-existing tool or package that is useful for the data science workflow. This will include a lightning talk where you briefly introduce and motivate the tool, and a hands-on practice session where you showcase the tool and offer exercises to try it out. The topics for the sessions will be announced a few weeks before the workshop. The workshop itself will then function like a conference with several parallel panels and many audience members. Both the talk and the prepared exercise materials, which must also be hosted openly on GitHub, will be graded.

Assignment 3: Data science project
In the final data science project, to be submitted a couple of weeks after classes have finished, you will design and implement your own data science project. You are supposed to collaborate in groups of two students. Student groups choose their topic subject to approval by the instructor.

Late submission of assignments: For each day the homework assignments and the final data science project is turned in late, the grade will be reduced by 10% (e.g. submission two days after the deadline would result in 20% grade deduction).

Attendance: Students are expected to be present and prepared for every class session. Active participation during lectures and seminar discussions is essential. If unavoidable circumstances arise which prevent attendance or preparation, the instructor should be advised by email with as much advance notice as possible. Please note that students cannot miss more than two out of 12 course sessions. For further information please consult the Examination Rules §10.

Academic Integrity: The Hertie School is committed to the standards of good academic and ethical conduct. Any violation of these standards shall be subject to disciplinary action. Plagiarism, deceitful actions as well as free-riding in group work are not tolerated. See Examination Rules §16 and the Hertie Plagiarism Policy.

Compensation for Disadvantages: If a student furnishes evidence that he or she is not able to take an examination as required in whole or in part due to disability or permanent illness, the Examination Committee may upon written request approve learning accommodation(s). In this respect, the submission of adequate certificates may be required. See Examination Rules §14.

Extenuating circumstances: An extension can be granted due to extenuating circumstances (i.e., for reasons like illness, personal loss or hardship, or caring duties). In such cases, please contact the course instructors and the Examination Office *in advance* of the deadline.

4. General Readings

During this course, we will frequently rely on the following textbook:

> Grolemund, G., & Wickham, H. (2018). *R for Data Science.* O'Reilly. Free online version available at https://r4ds.had.co.nz/. [**R4DS**]

Furthermore, there is an ocean of resources online. We have selected some resources as required reading and optional reading that we find particularly helpful. In addition, there are some resources that you might find generally useful:

Data wrangling, exploration, and analysis with R (Jenny Bryan)	https://stat545.com/
Intro to data science (Mine Cetinkaya-Rundel)	http://www2.stat.duke.edu/courses/Spring18/Sta199/
Data science in a box (RStudio)	https://datasciencebox.org/
R for Data Science Instructor's Guide (Greg Wilson)	https://github.com/rstudio-education/r4ds-instructors
Hands-On Programming with R (Garrett Grolemund)	https://rstudio-education.github.io/hopr/
R Packages (Hadley Wickham)	http://r-pkgs.had.co.nz/
Agile Data Science with R (Edwin Thoen)	https://edwinth.github.io/ADSwR/index.html
Statistical Programming (Colin Rundel)	http://www2.stat.duke.edu/~cr173/Sta523_Fa17/

5. Session Overview

Session	Session Date	Session Title
Setting things up		
1	09.09.2021	What is data science?
2	16.09.2021	Version control and project management
Collecting and wrangling data		
3	23.09.2021	R and the tidyverse
4	30.09.2021	Relational databases and SQL
5	07.10.2021	Web data and technologies
Analyzing data		
6	14.10.2021	Model fitting and simulation
Mid-term Exam Week: 18.10 - 22.10.2021 – no class		
7	28.10.2021	Visualization
8	04.11.2021	Workshop: Tools for Data Science
Fine-tuning the workflow		
9	11.11.2021	Working at the command line
10	18.11.2021	Debugging, automation, packaging
11	25.11.2021	Monitoring and communication
12	02.12.2021	Data science ethics
Final Exam Week: 14.12 - 18.12.2021 – no class		

6. Course Sessions and Readings

If not freely available online (see URLs), all readings will be accessible on the Moodle course site before semester start. In the case that there is a change in readings, students will be notified by email.

Required readings are to be read and analyzed thoroughly. Optional readings are intended to broaden your knowledge in the respective area, and it is highly recommended to at least skim them.

Session 1: What is data science?	
Learning Objective	After this session you have an overview of data science as an interdisciplinary field. Also, you'll have a roadmap of the course to navigate the rest of the course.

Required Readings	1. Kelleher, John and Brendan Tierney. 2018. *Data Science*. MIT Press. Chapter 1: What Is Data Science?
Optional Readings	2. Salganik, Matt. *Bit by Bit: Social Research in the Digital Age.* Princeton University Press.

Session 2: Version control and project management	
Learning Objective	After this session, you (a) have learned about the virtues of a robust version control and project setup workflow, and (b) are able to implement that workflow with Git and GitHub.
Required Readings	1. Bryan, Jenny and Jim Hester. 2018. Happy Git and GitHub for the user. https://happygitwithr.com/ 2. **R4DS**. Chapters 1—2, 4—8.
Optional Readings	3. Wickham, H. and Jenny Bryan. 2015. *R Packages*. O'Reilly Media. Chapter 13 (Git and GitHub). http://r-pkgs.had.co.nz/git.html

Session 3: R and the tidyverse	
Learning Objective	After this session, you will be ready to wrangle data in R the "modern", tidyverse way.
Required Readings	1. **R4DS**. Chapters 9—12, 17—21.
Optional Readings	2. Weidmann, Nils. 2021. *Data Management for Social Scientists: From Files to Databases.* Part One (Introduction) + Part Two (Data in Files) 3. Wickham, Hadley. The tidyverse style guide. https://style.tidyverse.org/ 4. Bryan, Jenny. Stat545, Chapters 17-21 (R as a programming Language). https://stat545.com/r-objects.html

Session 4: Relational databases and SQL	
Learning Objective	After this session you will (a) understand the principles of relational data structures and basic normalization, (b) be able to manipulate and join tables of data, (c) be able to interact with remote relational databases as if local.
Required Readings	1. **R4DS**. Chapters 12—13. 2. Munzert, Simon, Christian Rubba, Peter Meißner, and Dominic Nyhuis, 2015: *Automated Data Collection with R. A Practical Guide to Web Scraping and Text Mining*. Chichester: John Wiley & Sons. Chapter 7 (SQL and relational databases)

Optional Readings	3. Weidmann, Nils. 2021. *Data Management for Social Scientists: From Files to Databases.* Part Three (Data in Databases)

Session 5: Web data and technologies	
Learning Objective	After this session, you (a) have acquired basic knowledge of web technologies and (b) are able to scrape data from static webpages with R.
Required Readings	1. Munzert, Simon, Christian Rubba, Peter Meißner, and Dominic Nyhuis, 2015: *Automated Data Collection with R. A Practical Guide to Web Scraping and Text Mining*. Chichester: John Wiley & Sons. Chapters 1 (Introduction), 2 (HTML), 3 (XML and JSON), 4 (XPath) 2. https://rvest.tidyverse.org/articles/rvest.html
Optional Readings	3. https://cran.r-project.org/web/views/WebTechnologies.html 4. https://github.com/tidyverse/rvest 5. https://github.com/jeroen/jsonlite

Session 6: Model fitting and simulation	
Learning Objective	After this session you will (a) understand the bias/variance tradeoff in model fitting, (b) appreciate the implications of regularization strategies and the role of hyperparameter tuning, and (c) have a range of evaluation measures and strategies.
Required Reading	1. **R4Ds.** Chapters 22—25. 2. Muldoon, Ariel. 2018. Getting started with simulating data in R: some helpful functions and how to use them. https://aosmith.shinyapps.io/tutorial_simulation_helper_functions/
Optional Readings	3. https://www.tidymodels.org/

Mid-term Exam Week: 18 – 22.10.2021 – no class

Session 7: Visualization	
Learning Objective	After this session, you (a) have learned about basic rules to making visualizations that accurately reflect the data, tell a story, and look professional, (b) have learned about popular mistakes in visualization and how to avoid them, and (c) are able to integrate visualization as an alternative means to analyze data into your workflow.

Required Readings	1. Wilke, Claus. O. 2019. *Fundamentals of data visualization: a primer on making informative and compelling figures*. O'Reilly Media. https://serialmentor.com/dataviz/ 2. **R4DS**, Chapter 3 (Data visualization).
Optional Readings	3. Healy, K. 2018. *Data visualization: a practical introduction*. Princeton University Press. https://socviz.co/ 4. Traunmüller, R. 2020. Visualizing Data in Political Science. In: L. Curini & R. Franzese (eds.) *The SAGE Handbook of Research Methods in Political Science and International Relations*. Sage. https://tinyurl.com/visualization-polisci

Session 8: Workshop: Tools for Data Science	
Learning Objective	In this workshop, you will take both an active and a consumer role. You will make yourself familiar with one topic or package and prepare a workshop session. The session will be run as part of a workshop on tools for data science. Topics can be selected from a pre-arranged list, and include working with spatial data, network data, text data, regular expressions, string manipulation, interactive web-based visualization, and more.
Required Readings	none
Optional Readings	none

Session 9: Working at the command line	
Learning Objective	After this session, you will have made yourself familiar with the command line and understand how to use it to interact with your system and integrate it into your data science workflow.
Required Readings	1. Janssens, Jeroen. *Data Science at the Command Line*. 2nd edition. O'Reilly. Chapters 1, 2, 4, 5, 10.1, 10.4, 10.5.
Optional Readings	2. Ritchie, Gary. Using the RStudio Terminal in the RStudio IDE. https://support.rstudio.com/hc/en-us/articles/115010737148-Using-the-RStudio-Terminal

Session 10: Debugging, automation, and packaging	
Learning Objective	After this session, you (a) have learned the art of debugging, starting with a general strategy, then following up with specific tools, and (b) are able to turn your code into packages that others can easily download and use.
Required Readings	1. Wickham, H. 2019. *Advanced R.* CRC Press. Chapter 22 (Debugging) https://adv-r.hadley.nz/.

	2. Wickham, H. 2015. *R Packages*. O'Reilly Media. Chapters 1, 2, 4—9, 14. http://r-pkgs.had.co.nz/
Optional Readings	3. Bryan, Jenny. 2019. All the automation things. https://stat545.com/automation-overview.html

Session 11: Monitoring and communication	
Learning Objective	After this session, you will learn how to (a) build interactive web apps and dashboards using RMarkdown, and Shiny, and (b) communicate your results and products.
Required Readings	1. **R4DS**. Chapters 26—30. 2. https://shiny.rstudio.com/
Optional Readings	Wickham, Hadley. 2020. *Mastering Shiny*. O'Reilly. https://mastering-shiny.org/

Session 12: Data science ethics	
Learning Objective	After this session, you will have a deeper understanding of the ethical issues associated with working with sensitive data or with the construction of tools that - if applied improperly - can do harm to research subjects or society as a whole.
Required Readings	1. Kelleher, John and Brendan Tierney. 2018. *Data Science*. MIT Press. Chapter 6 (Privacy and Ethics) 2. Salganik, Matt. *Bit by Bit: Social Research in the Digital Age.* Princeton University Press. Chapter 6 (Ethics)
Optional Readings	3. Floridi, Luciano and Mariarosaria Taddeo. 2016. What is Data Ethics? *Phil. Trans. R. Soc. A. 37420160360.*

Final Exam Week: 13 - 17.12.2021 – no class