Course Syllabus

GRAD-C11/GRAD-E1339: Introduction to Data Science
Concentration: Policy Analysis

1. General information

Course Format	Onsite
Instructor	Prof. Simon Munzert
Instructor's e-mail	munzert@hertie-school.org
Assistant	Alex Karras karras@hertie-school.org
Instructor's Office Hours	Please use the online document to book an appointment

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

For information on **course times, session dates and course locations** please consult the Course List on *MyStudies*.

Instructor Information:

Simon Munzert is Professor of Data Science and Public Policy at the Hertie School and Director 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:

N.N.

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; 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 other technical courses and the master's thesis.

Software:
We will work with RStudio, Quarto/Markdown, 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., Çetinkaya-Rundel M., & Grolemund, G. (2023). R for Data Science, 2nd edition. Available at: https://r4ds.hadley.nz/. This book is also the main textbook for this class, but you might want to dive in already before the course starts.
2. Ismay, Ch., & Kim, A.Y. (2023). Statistical Inference via Data Science. A Modern Dive into R and the Tidyverse. Available at: https://moderndive.com/
3. 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; MPP/MIA 2nd year students who have successfully completed Statistics II.

Teaching style:
The sessions will feature a 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:
For MPP/MIA 2nd year students: 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.)

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 (H)	Deadline: Tuesdays in the week after the lecture at 12 noon	Submit via GitHub	4 x 10%

Assignment 2: Quizzes (Q)	Deadline: Wednesdays after the lecture at 12 noon	Moodle	4 × 5%
Assignment 3: Workshop presentation	Deadline for materials: One week before the workshop (presentation provided by core course students only; elective students just participate in the workshop)	Submit via GitHub/ present live/participate	10%
Assignment 4: Hackathon project	Deadline: 48h to work on project after the on-site hackathon (start: December 19, 10am, end: December 21, 10am)	Submit via GitHub	30%

Assignment Details:

Evaluation is conducted via a combination of a series of homework assignments (counting towards 40% of the final grade), online quizzes (counting towards 20% of the final grade), preparation and execution of a live workshop session (counting towards 10% of the final grade), and one hackathon project (counting towards 30% 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, I 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. The session overview below shows you which sessions are accompanied by a homework (H).

Assignment 2: Quizzes

In 5 online quizzes, you will be tested on the concepts learned in class. The 4 best out of 5 quizzes will contribute to the final grade. The session overview below shows you which sessions are accompanied by a quiz (Q).

Assignment 3: Workshop presentation

Near the end of 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 audience members (also you). Students who take the course as a core course are required to prepare one of the sessions and for them the prepared materials will be graded. Students who take the course as an elective are required to participate in a minimum number of workshop sessions and submit solutions to exercises as proof.

Assignment 4: Hackathon project
In the hackathon project, you get to solve a set of problems related using one given dataset, and also submit a creative project of your choice using the same dataset. After the instructions have been made available, you will have 48 hours to submit solutions as a team of three. The task is similar to the homework assignments but put more emphasis on creative problem-solving using the tools and techniques you have learned in class.

Late submission of assignments: For each day the homework assignments and the hackathon 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:

> Wickham, H., Çetinkaya-Rundel M., & Grolemund, G. (2023). *R for Data Science, 2nd edition*. O'Reilly. Free online version available at: https://r4ds.hadley.nz/ [**R4DS**]

If you already have substantial experience with R, you're recommended to also check out the following textbook, which will be considered optional reading in some sessions:

> Wickham, H., (2023). *Advanced R, 2nd edition*. O'Reilly. Free online version available at: https://adv-r.hadley.nz/ [**AdvR**]

5. Session Overview

Course session times and dates can be found in the [Course Plan](#) on *MyStudies*.

Session	Session Title	Assignment
	Fundamentals	
1	**What is data science?**	H
2	**Version control and project management**	Q
3	**Programming: Functions and debugging**	H
	Collecting and wrangling data	
4	**Relational databases and SQL**	Q
5	**Web data and technologies**	Q
6	**Web scraping and APIs**	H
	Analyzing data	
7	**Model fitting and evaluation**	Q
Mid-term Exam Week: no class		
8	**Visualization**	H
9	**Workshop: Tools for Data Science**	-
	Fine-tuning the workflow	
10	**Automation, scheduling, and packages**	Q
11	**Monitoring and communication**	H
12	**Data science ethics**	-
Final Exam Week: 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. In the accompanying lab, you will be geared up to wrangle data in R the "modern", tidyverse way.

Required Readings	1. **R4DS**. Chapters 1—4; 6; 8.
Optional Readings	2. Kelleher, John and Brendan Tierney. 2018. Data Science. MIT Press. Chapter 1: What Is Data Science?
	3. Salganik, Matt. *Bit by Bit: Social Research in the Digital Age*. Princeton University Press.
	4. **AdvR.** Chapters 1—5.

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. **R4DS**. Chapters 5; 7.
	2. Bryan, Jenny and Jim Hester. 2018. Happy Git and GitHub for the user. https://happygitwithr.com/
Optional Readings	3. Learn Git Branching – the Game. https://learngitbranching.js.org/
	4. Thoen, Edwin. Agile Data Science with R. https://edwinth.github.io/ADSwR/.

Session 3: Programming: Functions and debugging

Learning Objective	After this session, you (a) are able to make your code more efficient using functional programming tools, and (b) have learned the art of debugging, starting with a general strategy, then following up with specific tools.
Required Readings	1. **R4DS**. Chapters 26—28.
	2. **AdvR.** Chapter 22.
Optional Readings	3. Grolemund, G. Hands-On Programming with R. https://rstudio-education.github.io/hopr/.
	4. **AdvR.** Chapters 6-11.

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 21—23.
Optional Readings	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)

	3. Weidmann, Nils. 2023. *Data Management for Social Scientists. From Files to Databases.* Cambridge University Press.
	4. Lord, Robin. Lost at SQL – the learning game. https://lost-at-sql.therobinlord.com/

Session 5: Web data and technologies	
Learning Objective	After this session, you have acquired basic knowledge of web technologies and web data in its various forms.
Required Readings	1. **R4DS.** Chapters 13—16. 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. Chapters 1 (Introduction), 2 (HTML), 3 (XML and JSON), 4 (XPath), 8 (Regular expressions)
Optional Readings	3. Regexcrossword – the regular expressions game. https://regexcrossword.com/ 4. https://cran.r-project.org/web/views/WebTechnologies.html 5. https://github.com/jeroen/jsonlite

Session 6: Web scraping and APIs	
Learning Objective	After this session, you (a) have acquired basic knowledge of classic web scraping and (b) web APIs.
Required Readings	1. **R4DS**. Chapter 25. 2. Bauer, Paul, Lion Behrens, and Camille Landesvatter (eds.). 2022. *APIs for social scientists: A collaborative review.* https://bookdown.org/paul/apis_for_social_scientists/ Chapters 1 (Introduction), 2 (Best Practices), and any of the API-specific chapters you find interesting (at least one to get an impression!)
Optional 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. Chapter 9 (Scraping the web)

Session 7: Model fitting and evaluation	
Learning Objective	After this session you will learned the basics of model fitting and evaluation techniques with R.
Required Reading	1. **R4Ds (1st edition; not available yet in 2nd edition).** Chapters 22—25. https://r4ds.had.co.nz/model-intro.html

	2. Simonsohn, Uri, Joseph P. Simmons, Leif D. Nelson. 2020. Specification curve analysis. *Nature Human Behaviour* 4, 1208—1214.
Optional Readings	3. **AdvR**. Chapters 23—24. 4. Dani Cosme. 2022. Specification Curve Analysis (tutorial) https://dcosme.github.io/specification-curves/SCA_tutorial_2022 5. https://www.tidymodels.org/

Session 8: 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 complementary 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://clauswilke.com/dataviz/. 2. **R4DS**, Chapters 10-12 (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 9: 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 particular data science tool or topic and prepare a workshop session. The session will be run as part of a workshop on tools for data science. Topics have to be selected from a pre-compiled list.
Required Readings	none
Optional Readings	none

Session 10: Automation, scheduling, and packages	
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. 2015. *R Packages, 2nd edition*. O'Reilly. Chapters 1—5. https://r-pkgs.org/
	2. Bryan, Jenny. 2019. All the automation things. https://stat545.com/automation-overview.html
Optional Readings	none

Session 11: Monitoring and communication

Learning Objective	After this session, you will learn how to (a) build interactive web apps and dashboards using Quarty/Markdown and Shiny, and (b) communicate your results and products.
Required Readings	1. **R4DS**. Chapters 29—30.
	2. https://shiny.rstudio.com/
Optional Readings	3. van der Bles AM, van der Linden S, Freeman ALJ, Mitchell J, Galvao AB, Zaval L, Spiegelhalter DJ. 2019 Communicating uncertainty about facts, numbers and science. *R. Soc. open sci.* 6: 181870. http://dx.doi.org/10.1098/rsos.181870
	4. 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.*