

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

*Concentration: Policy Analysis*

**1. General information**

Instructor	Prof. Simon Munzert
Instructor's e-mail	<a href="mailto:munzert@hertie-school.org">munzert@hertie-school.org</a>
Assistant	Ayamba Kwoyila <a href="mailto:kwoyila@hertie-school.org">kwoyila@hertie-school.org</a>
Teaching assistants	Lisa Oswald, <a href="mailto:l.oswald@phd.hertie-school.org">l.oswald@phd.hertie-school.org</a> Tom Arend, <a href="mailto:t.arend@phd.hertie-school.org">t.arend@phd.hertie-school.org</a>
Instructor's Office Hours	Please email Ayamba Kwoyila in advance to make an appointment and to let me know what you would like to talk about.

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 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 Masters 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; 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, 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. Ismay, Ch., & Kim, A.Y. (2021). 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 1<sup>st</sup> year students; MPP/MIA 2<sup>nd</sup> 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 2<sup>nd</sup> 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
<b>Assignment 1: Homework assignments</b>	Deadline: 11.59pm on the day before class	Submit via GitHub	4 x 10%
<b>Assignment 2: Workshop session</b>	Deadline: November 7 (materials), November 14 (workshop)	Submit via GitHub + present live	30%
<b>Assignment 3: Data science project exam</b>	Deadline: 48h to work on take-home exam (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), preparation and execution of a live workshop session (counting towards 30% of the final grade), and one data science 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, 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.

#### **Assignment 2: Workshop session**

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 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 exam**

In the final data science project exam, you get to solve a set of coding problems related to one particular data setup and associated research questions. After the instructions have been made available, you will have 48 hours to submit your individual solutions. 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 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)	<a href="https://stat545.com/">https://stat545.com/</a>
Intro to data science (Mine Cetinkaya-Rundel)	<a href="http://www2.stat.duke.edu/courses/Spring18/Sta199/">http://www2.stat.duke.edu/courses/Spring18/Sta199/</a>
Data science in a box (RStudio)	<a href="https://datasciencebox.org/">https://datasciencebox.org/</a>
R for Data Science Instructor's Guide (Greg Wilson)	<a href="https://github.com/rstudio-education/r4ds-instructors">https://github.com/rstudio-education/r4ds-instructors</a>

Hands-On Programming with R (Garrett Grolemund)	<a href="https://rstudio-education.github.io/hopr/">https://rstudio-education.github.io/hopr/</a>
R Packages (Hadley Wickham)	<a href="http://r-pkgs.had.co.nz/">http://r-pkgs.had.co.nz/</a>
Agile Data Science with R (Edwin Thoen)	<a href="https://edwinth.github.io/ADSwR/index.html">https://edwinth.github.io/ADSwR/index.html</a>
Statistical Programming (Colin Rundel)	<a href="http://www2.stat.duke.edu/~cr173/Sta523_Fa17/">http://www2.stat.duke.edu/~cr173/Sta523_Fa17/</a>

## 5. Session Overview

Session	Session Title
1	What is data science?
2	Version control and project management
3	R and the Tidyverse
4	Debugging, automation, functions
<u>Collecting and wrangling data</u>	
5	Relational databases and SQL
6	Web data and technologies
7	Web scraping and APIs
<u>Analyzing data</u>	
8	Model fitting and evaluation
9	Visualization
10	Workshop: Tools for Data Science
<u>Fine-tuning the workflow</u>	
11	Monitoring and communication
12	Data science ethics

Midterm Exam Week: 24—28 Oct 2022 – no class

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.
Required Readings	<ol style="list-style-type: none"><li>1. Kelleher, John and Brendan Tierney. 2018. <i>Data Science</i>. MIT Press. Chapter 1: What Is Data Science?</li><li>2. <b>R4DS</b>. Chapters 1—8.</li></ol>
Optional Readings	<ol style="list-style-type: none"><li>3. Salganik, Matt. <i>Bit by Bit: Social Research in the Digital Age</i>. Princeton University Press.</li></ol>

### 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	<ol style="list-style-type: none"><li>1. Bryan, Jenny and Jim Hester. 2018. Happy Git and GitHub for the user. <a href="https://happygitwithr.com/">https://happygitwithr.com/</a></li><li>2. Learn Git Branching – the Game. <a href="https://learnitbranching.js.org/">https://learnitbranching.js.org/</a></li></ol>
Optional Readings	<ol style="list-style-type: none"><li>3. Wickham, H. and Jenny Bryan. 2015. <i>R Packages</i>. O'Reilly Media. Chapter 13 (Git and GitHub). <a href="http://r-pkgs.had.co.nz/git.html">http://r-pkgs.had.co.nz/git.html</a></li></ol>

### 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	<ol style="list-style-type: none"><li>1. <b>R4DS</b>. Chapters 9—11, 17—18.</li></ol>
Optional Readings	<ol style="list-style-type: none"><li>2. Wickham, Hadley. The tidyverse style guide. <a href="https://style.tidyverse.org/">https://style.tidyverse.org/</a></li><li>3. Bryan, Jenny. Stat545, Chapters 17-21 (R as a programming Language). <a href="https://stat545.com/r-objects.html">https://stat545.com/r-objects.html</a></li></ol>

### Session 4: Debugging, automation, functions

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 make your code more efficient using functional programming tools.
--------------------	---

<b>Required Readings</b>	<ol style="list-style-type: none"> <li>1. Wickham, H. 2019. <i>Advanced R</i>. CRC Press. Chapter 22 (Debugging) <a href="https://adv-r.hadley.nz/">https://adv-r.hadley.nz/</a>.</li> <li>2. <b>R4DS</b>. Chapters 19-21.</li> </ol>
<b>Optional Readings</b>	<ol style="list-style-type: none"> <li>3. Bryan, Jenny. 2019. All the automation things. <a href="https://stat545.com/automation-overview.html">https://stat545.com/automation-overview.html</a></li> </ol>

## Session 5: Relational databases and SQL

<b>Learning Objective</b>	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.
<b>Required Readings</b>	<ol style="list-style-type: none"> <li>1. <b>R4DS</b>. Chapters 12—13.</li> </ol>
<b>Optional Readings</b>	<ol style="list-style-type: none"> <li>2. Munzert, Simon, Christian Rubba, Peter Meißner, and Dominic Nyhuis, 2015: <i>Automated Data Collection with R. A Practical Guide to Web Scraping and Text Mining</i>. Chichester: John Wiley &amp; Sons. Chapter 7 (SQL and relational databases)</li> </ol>

## Session 6: Web data and technologies

<b>Learning Objective</b>	After this session, you have acquired basic knowledge of web technologies and web data in its various forms.
<b>Required Readings</b>	<ol style="list-style-type: none"> <li>1. Munzert, Simon, Christian Rubba, Peter Meißner, and Dominic Nyhuis, 2015: <i>Automated Data Collection with R. A Practical Guide to Web Scraping and Text Mining</i>. Chichester: John Wiley &amp; Sons. Chapters 1 (Introduction), 2 (HTML), 3 (XML and JSON), 4 (XPath)</li> </ol>
<b>Optional Readings</b>	<ol style="list-style-type: none"> <li>2. <a href="https://cran.r-project.org/web/views/WebTechnologies.html">https://cran.r-project.org/web/views/WebTechnologies.html</a></li> <li>3. <a href="https://github.com/jeroen/jsonlite">https://github.com/jeroen/jsonlite</a></li> </ol>

## Session 7: Web scraping and APIs

<b>Learning Objective</b>	After this session, you (a) have acquired basic knowledge of classic web scraping and (b) web APIs.
<b>Required Readings</b>	<ol style="list-style-type: none"> <li>1. <b>R4DS</b>. Chapter 14.</li> <li>2. Munzert, Simon, Christian Rubba, Peter Meißner, and Dominic Nyhuis, 2015: <i>Automated Data Collection with R. A Practical Guide to Web Scraping and Text Mining</i>. Chichester: John Wiley &amp; Sons. Chapters 8 (Regular expressions), 9 (Scraping the web)</li> <li>3. Bauer, Paul, Lion Behrens, and Camille Landesvatter (eds.). 2022. APIs for social scientists: A collaborative review. <a href="https://bookdown.org/paul/apis_for_social_scientists/">https://bookdown.org/paul/apis_for_social_scientists/</a> Chapters</li> </ol>

	1 (Introduction), 2 (Best Practices), and any of the API-specific chapters you find interesting (at least one to get an impression!)
Optional Readings	<ol style="list-style-type: none"> <li>1. <a href="https://cran.r-project.org/web/views/WebTechnologies.html">https://cran.r-project.org/web/views/WebTechnologies.html</a><a href="https://cran.r-project.org/web/views/WebTechnologies.html">https://cran.r-project.org/web/views/WebTechnologies.html</a></li> <li>2. <a href="https://github.com/tidyverse/rvest">https://github.com/tidyverse/rvest</a></li> </ol>

## Session 8: Model fitting and evaluation

Learning Objective	After this session you will (a) have refreshed your memories of regression modeling and (b) learned the basics of model fitting and evaluation techniques with R.
Required Reading	1. <b>R4Ds</b> . Chapters 22—25.
Optional Readings	2. <a href="https://www.tidymodels.org/">https://www.tidymodels.org/</a>

## Session 9: 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	<ol style="list-style-type: none"> <li>1. Wilke, Claus. O. 2019. <i>Fundamentals of data visualization: a primer on making informative and compelling figures</i>. O'Reilly Media. <a href="https://serialmentor.com/dataviz/">https://serialmentor.com/dataviz/</a></li> <li>2. <b>R4DS</b>, Chapter 3 (Data visualization).</li> </ol>
Optional Readings	<ol style="list-style-type: none"> <li>3. Healy, K. 2018. <i>Data visualization: a practical introduction</i>. Princeton University Press. <a href="https://socviz.co/">https://socviz.co/</a></li> <li>4. Traunmüller, R. 2020. Visualizing Data in Political Science. In: L. Curini &amp; R. Franzese (eds.) <i>The SAGE Handbook of Research Methods in Political Science and International Relations</i>. Sage. <a href="https://tinyurl.com/visualization-polisci">https://tinyurl.com/visualization-polisci</a></li> </ol>

## Session 10: 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 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	<ol style="list-style-type: none"> <li>1. <b>R4DS</b>. Chapters 26—30.</li> <li>2. <a href="https://shiny.rstudio.com/">https://shiny.rstudio.com/</a></li> </ol>
Optional Readings	<ol style="list-style-type: none"> <li>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. <i>R. Soc. open sci.</i> 6: 181870. <a href="http://dx.doi.org/10.1098/rsos.181870">http://dx.doi.org/10.1098/rsos.181870</a></li> <li>4. Wickham, Hadley. 2020. <i>Mastering Shiny</i>. O'Reilly. <a href="https://mastering-shiny.org/">https://mastering-shiny.org/</a></li> </ol>

### 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	<ol style="list-style-type: none"> <li>1. Kelleher, John and Brendan Tierney. 2018. <i>Data Science</i>. MIT Press. Chapter 6 (Privacy and Ethics)</li> <li>2. Salganik, Matt. <i>Bit by Bit: Social Research in the Digital Age</i>. Princeton University Press. Chapter 6 (Ethics)</li> </ol>
Optional Readings	<ol style="list-style-type: none"> <li>3. Floridi, Luciano and Mariarosaria Taddeo. 2016. What is Data Ethics? <i>Phil. Trans. R. Soc. A.</i> 374:20160360.</li> </ol>

**Final Exam Week: 12 - 16.12.2022 – no class**