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

Will Lowe, Simon Munzert

1. General information

<table>
<thead>
<tr>
<th>Class time</th>
<th>Thursday, 14-16h</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class Format</td>
<td>This course uses a “flipped classroom” format and combines 50 minutes of pre-recorded material (audio or video) with a 50-minute interactive seminar. Students will use the pre-recorded material to prepare for the seminar. The seminar is taught onsite at the Hertie School, or online via the platform Clickmeeting, depending upon your location. For those attending the online seminar, Clickmeeting allows for interactive, participatory seminar style teaching.</td>
</tr>
</tbody>
</table>
| Instructor               | Will Lowe, PhD (WL)  
|                          | Prof. Simon Munzert (SM) |
| Instructor’s office      | 3.14 (WL)  
|                          | 3.13.1 (SM) |
| Instructor’s e-mail      | lowe@hertie-school.org  
|                          | munzert@hertie-school.org |
| Instructor’s phone number| +49 (0)30 259 219 XXX (WL)  
|                          | +49 (0)30 259 219 450 (SM) |
| Assistant                | Ayamba Kwoyila  
|                          | kwoyila@hertie-school.org  
|                          | +49 (0)30 259 219 121  
|                          | 3:59 |
| Instructor’s Office Hours| XXXXdays, X-X  
|                          | Tuesdays, 2-3pm (SM) |

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

Instructor Information:  
Dr. William Lowe is a political methodologist with interests in text analysis, causal inference, and machine learning and is Senior Research Scientist at the Hertie School Data Science Lab. He received his Ph.D. in Cognitive Science and Natural Language Processing but has, save for a short period in the technology industry, worked in Political Science ever since.

Simon Munzert is Assistant Professor of Data Science and Public Policy at the Hertie School and part of the Hertie School Data Science Lab. His research interests include opinion
formation in the digital age, public opinion, and the use of online data in social research. He received his Doctoral Degree in Political Science from the University of Konstanz.

2. Course Contents and Learning Objectives

Course contents:
This course will teach you how to do data science 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, medicine, and social science. This course provides students with advanced data science skills using the powerful R programming language.

The course is organized in five parts. The first part covers basic workflow with R and RStudio, version control with Git/GitHub and basic rules of efficient coding. The second part focuses on data wrangling of relationally structured data as well as non-relationally structured data, such as spatial or text data. In the third part, students learn how to collect data from the web using scraping technology and web APIs as well as online experiments. The fourth part turns to the big picture of data analysis, covering model fitting techniques and data visualization. The last part addresses advanced workflow issues, including solutions to big data, how to speed things up, debugging and automation, and data communication and tool-building.

The course is intended for students with some experience in working with R. If you have had little to no exposure to R before, but nevertheless want to take this course, then you are strongly recommended to complete the Swirl course “R programming” (see https://swirlstats.com/students.html and https://github.com/swirldev/R_Programming_E) before the course starts.

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

Target group:
MPP and MIA 2nd year students

Teaching style:
The sessions will mainly feature an interactive lecture on the session’s topic led by the instructor. From time to time, you will work on conceptual and coding problems in small groups.

Prerequisites:
Statistics I, 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.
2. Grading and Assignments

Composition of Final Grade:

| Assignment 1: Series of weekly assignments | Deadline: 11.59pm on the day before class | Submit via GitHub | 6 x 10% |
| Assignment 2: Final data analysis project | Deadline: 21.12.2020, 11.59pm | Submit via GitHub | 40% |

Assignment Details

Evaluation is conducted via a combination of a series of weekly assignments (counting towards 60% of the final grade) and one data analysis project (counting towards 40% of the final grade). While you should submit your own, individual solutions to both the weekly assignments and the final data analysis project, we generally encourage you to study and learn to use the software together.

Assignment 1: Weekly assignments
In the weekly 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. Not all sessions will be accompanied by an assignment. The first week’s assignment will serve as a non-graded test run. The 6 best out of the remaining 7 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: Data analysis project
In the final data analysis project, to be submitted a couple of weeks after classes have finished, you will design and implement your own data analysis project. You are supposed to collaborate in groups of two or three students. Student groups choose their topic subject to approval by the instructors. Grades will be based on a group presentation and report, weighted equally.

Late submission of assignments: For each day the assignment 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.

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.

3. General Readings

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


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:

<table>
<thead>
<tr>
<th>STAT 545: Data wrangling, exploration, and analysis with R (Jenny Bryan)</th>
<th><a href="https://stat545.com/">https://stat545.com/</a></th>
</tr>
</thead>
<tbody>
<tr>
<td>STA 199: Intro to data science (Mine Cetinkaya-Rundel)</td>
<td><a href="http://www2.stat.duke.edu/courses/Spring18/Sta199/">http://www2.stat.duke.edu/courses/Spring18/Sta199/</a></td>
</tr>
<tr>
<td>Data science in a box (RStudio)</td>
<td><a href="https://datasciencebox.org/">https://datasciencebox.org/</a></td>
</tr>
<tr>
<td>R Packages (Hadley Wickham)</td>
<td><a href="http://r-pkgs.had.co.nz/">http://r-pkgs.had.co.nz/</a></td>
</tr>
<tr>
<td>Statistical Programming (Colin Rundel)</td>
<td><a href="http://www2.stat.duke.edu/~cr173/Sta523_Fa17/">http://www2.stat.duke.edu/~cr173/Sta523_Fa17/</a></td>
</tr>
</tbody>
</table>

4. Session Overview

<table>
<thead>
<tr>
<th>Session</th>
<th>Session Date</th>
<th>Session Title</th>
<th>Instructor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Setting up the workflow</td>
<td>10.09.2020</td>
<td>Overview and basic workflow</td>
<td>Will</td>
</tr>
</tbody>
</table>
5. **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: Overview and basic workflow (Will)

<table>
<thead>
<tr>
<th>Learning Objective</th>
<th>Required Readings</th>
</tr>
</thead>
<tbody>
<tr>
<td>After this session you have learned the principles of a replicable data science workflow.</td>
<td>1. AdvR. Chapters 4 &amp; 8</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Optional Readings</th>
</tr>
</thead>
</table>

---

### Required Readings

1. AdvR. Chapters 4 & 8

---

### Optional Readings

---
### Session 2: Version control and coding style (Simon)

<table>
<thead>
<tr>
<th>Learning Objective</th>
<th>After this session, you (a) have learned about the virtues of a robust version control workflow, (b) have learned the basics of functional programming to support an efficient coding style, and (c) are able to implement that workflow with Git and GitHub.</th>
</tr>
</thead>
</table>
2. Wickham, Hadley. The tidyverse style guide. [https://style.tidyverse.org/](https://style.tidyverse.org/)  

### Session 3: Relationally structured data (Will)

<table>
<thead>
<tr>
<th>Learning Objective</th>
<th>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.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Optional Readings</td>
<td></td>
</tr>
</tbody>
</table>

### Session 4: Spatial data (Simon)

<table>
<thead>
<tr>
<th>Learning Objective</th>
<th>After this session, you (a) have learned how spatial information can be encoded in spatial features (points, lines, polygons), (b) are able to set up and manage spatial datasets, and (c) can visualize spatial data with R.</th>
</tr>
</thead>
</table>
| Optional Readings  | 3. CRAN Task View: Analysis of Spatial Data. [https://cran.r-project.org/web/views/Spatial.html](https://cran.r-project.org/web/views/Spatial.html)  
### Session 5: Text data (Will)

**Learning Objective**
After this session you will (a) understand the essential problems of working with text as data and their solutions and (b) have a basic understanding of and ability to apply topic and scaling models, and their manual counterparts.

**Required Readings**

**Optional Readings**

### Session 6: Web data (Simon)

**Learning Objective**
After this session, you (a) have acquired basic knowledge of web technologies, (b) are able to scrape information from static and dynamic websites using R, and (c) are able to access web services (APIs) with R.

**Required Readings**

**Optional Readings**
2. [https://cran.r-project.org/web/views/WebTechnologies.html](https://cran.r-project.org/web/views/WebTechnologies.html)
3. [https://github.com/tidyverse/rvest](https://github.com/tidyverse/rvest)
4. [https://github.com/jeroen/jsonlite](https://github.com/jeroen/jsonlite)

**Mid-term Exam Week: 19 – 23.10.2020 – no class**
### Session 7: Experimental and crowdsourced data (Simon)

**Learning Objective**

After this session, you (a) have learned the basics about setting up and monitoring experimental analyses using R and (b) are able to design your own experiments and crowdsourced tasks online.

**Required Readings**


**Optional Readings**

3. CRAN Task View: Design of Experiments (DoE) & Analysis of Experimental Data. https://cran.r-project.org/web/views/ExperimentalDesign.html

### Session 8: Fitting models (Will)

**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**


**Optional Readings**

2. CRAN Task View: Reproducible Research. https://cran.r-project.org/web/views/MachineLearning.html

### Session 9: Visualization (Simon)

**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**

2. R4DS, Chapter 3 (Data visualisation).

**Optional Readings**

6. htmlwidgets for R. [https://www.htmlwidgets.org/](https://www.htmlwidgets.org/)

<table>
<thead>
<tr>
<th>Session 10: Trouble with big data (Will)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Learning Objective</strong></td>
</tr>
<tr>
<td><strong>Optional Readings</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Session 11: Debugging, automation, and packaging (Simon)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Learning Objective</strong></td>
</tr>
</tbody>
</table>
| **Required Readings** | 1. AdvR. Chapter 22 (Debugging).  

<table>
<thead>
<tr>
<th>Session 12: Special Topics</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Learning Objective</strong></td>
</tr>
<tr>
<td>Required Readings</td>
</tr>
<tr>
<td>-------------------</td>
</tr>
<tr>
<td>Optional Readings</td>
</tr>
</tbody>
</table>

Final Exam Week: 14 - 18.12.2020 – no class