GESIS Leibniz Institute for the Social Sciences

gesis training

GESIS Fall Seminar in Computational Social Science
"Introduction to Computational Social Science with R"

Lecturer:	Johannes B. Gruber
Affiliation:	GESIS
Email:	johannes.gruber@gesis.org

Course Dates and Workload

Date: September 1-5, 2025
Total estimated workload: 30 hours

About the Lecturer

Johannes B. Gruber is a team lead in the department Data Services for the Social Sciences at GESIS. Previously he worked as a Postdoctoral Researcher at University of Amsterdam, Vrije Universiteit Amsterdam and the European New School of Digital Studies. His expertise is on the collection and analysis of social and media data, with a focus on the information flows in our media system. Besides his research, he has developed or worked on software packages for web-scraping (traktok, paperboy, cookiemonster), text analysis (rollama, spacyr, quanteda.textmodels, stringdist, rwhatsapp, LexisNexisTools), and data storage and sharing (amcat4r).

Course Description

The digital revolution has produced unprecedented amounts of data that are relevant for researchers in the social sciences, from online surveys to social media user data, travel and access data, and digital or digitized text data. How can these masses of raw data be turned into understanding, insight, and knowledge? The goal of this course is to introduce you to topics, methods, and workflows in Computational Social Science (CSS) with R, a powerful programming language that offers a wide variety of tools, used by journalists, data scientists, and researchers alike. Unlike many introductions to programming, e.g., in computer science, the focus of this course is on how to explore, obtain, wrangle, visualize, model, and communicate data to address challenges in the social sciences. The course emphasizes the theoretical and ethical aspects of CSS while covering topics such as web scraping (i.e. obtaining data from the internet), data cleaning (i.e. getting raw data into a rectangular or otherwise easy-to-analyze format), and visualization (i.e. drawing bar, line, scatter plots and more from data), automated/computational text analysis (i.e. using the computer to find patterns in text or sort documents into categories), machine learning (i.e. training algorithms on annotated data and generalizing patterns to unseen data), network analysis (i.e. examining relationships among entities, such as persons, organizations, or documents) and the basics of probabilistic modeling. The course will be held as an **online blended learning** format with video lectures focused on theoretical background and demonstrations accompanied by live online sessions where participants can ask questions and work through projects together.

Organizational Structure of the Course

The course will take place in a blended learning format. That means that you will need to (1.) read the literature listed under each session (if any); (2.) watch the video lecture; (3.) finish the exercises before each live group session. This means that participants will be on roughly the same level of knowledge during the live sessions, and we will be able to focus on open discussion, the answering of questions, and small group exercises.

Keywords

text analysis; network analysis; web scraping; visualization

Target Group

You will find the course useful if:

- you have taken, e.g., a statistics course, know a little bit of R, and now want to explore and get an overview of computational methods, data science, or one of the approaches listed above.

Course and Learning Objectives

By the end of the course you will:

- be able to define what constitutes the field of computational social science
- have a high-level overview of the approaches utilized in computational social science, including advantages and shortcomings
- have basic knowledge and hands-on experience of how to apply the approaches and what tools are considered state-of-the-art
- be equipped to deepen your knowledge on the theory and practice of computational social science.

Course Prerequisites

Working knowledge of R is required. You should take the self-assessment test I prepared here: jbgruber.shinyapps.io/self-assessment. If you do not know enough R to comfortably solve all or at least most of the exercises, there is a list of recommended crash courses at the end of the self-assessment site. If you need more help, you should take the "Introduction to R" course that takes place online from 25-27 August.

Software Requirements

We will work with R and RStudio in the course. But you can of course use an IDE different from RStudio if you prefer. I will provide a script that installs and/or updates all R packages we need during the course. You will need a version of R newer than 4.1.0. I also recommend you use an updated version of the Zoom client for the live sessions (a good camera and mic would also be preferable).

Schedule

1. Introduction to Computational Social Science

Self-learning Session 1: to be completed before Live Session 1 (see below), estimated workload: 3h

This session will give a broad overview of the field of computational social science, related fields, and its development. The material will also provide background on the ethical conduct of CSS research and give participants practical guidelines on how to make sure their research adheres to ethical (and legal) standards. Additionally, it will provide basic skills in Exploratory Data Analysis (EDA) and visualization, and go over how to set up things properly for the course.

Literature (preliminary; details on what is mandatory and what is recommended will follow):

- Salganik, M. J. (2019). *Bit by bit: Social research in the digital age*. https://www.bitbybitbook.com/en/1st-ed/introduction/
- Lazer, D. M., Pentland, A., Watts, D. J., Aral, S., Athey, S., Contractor, N., ... & Wagner, C. (2020). Computational social science: Obstacles and opportunities. *Science*, 369(6507), 1060-1062. https://doi.org/10.1126/science.aaz8170
- Theocharis, Y. & Jungherr, A. (2021). Computational Social Science and the Study of Political Communication. *Political Communication*, 38(1-2), 1-22. https://doi.org/10.1080/10584609.2020.1833121
- van Atteveldt, W. & Peng, T.-Q. (2018). When Communication Meets Computation: Opportunities, Challenges, and Pitfalls in Computational Communication Science. *Communication Methods and Measures*, 12(2-3), 81-92. https://doi.org/10.1080/19312458.2018.1458084

Live Session 1: Monday, 1 September, 2:00-5:30 pm CEST, {Zoom Link}

In this session we will get to know each other. There will also be room for questions and we will make sure that everyone has the software up and running. We will then break into groups to work on a small task.

2. Obtaining Data

Self-learning Session 2: to be completed before Live Session 2 (see below), estimated workload: 3h
A core idea of CSS is that you work with found, rather than with designed data. Designed data would be data that was specifically collected through a survey or experiment, where the researcher controls what questions to ask and in what format the data is returned. Found data, on the other hand, are often traces left behind by people who were doing something that had nothing to do with research. Like writing, liking, sharing or deleting a post on social media, using a website, an app or a service, or doing their jobs, for example, as politicians, journalists or book authors. These data can often tell us more about the actual behavior of people and institutions than what they would share in a survey or experiment. They are also often cheaply available en masse. But to download, wrangle, and clean them can be difficult. This session gives an overview of web scraping, which is the process of downloading, wrangling ,and cleaning found data to make it possible to analyze it.

Literature (preliminary):

- Freelon, D. (2018). Computational Research in the Post-API Age. *Political Communication*, 35(4), 665-668. https://doi.org/10.1080/10584609.2018.1477506
- Hennesy, C., & Samberg, R. (2019). "Law and Literacy in Non-Consumptive Text Mining: Guiding Researchers Through the Landscape of Computational Text Analysis." *Copyright Conversations: Rights Literacy in a Digital World*. https://escholarship.org/uc/item/55j0h74g
- Luscombe, A., Dick, K. & Walby, K. (2022). Algorithmic thinking in the public interest: navigating technical, legal, and ethical hurdles to web scraping in the social sciences. *Qual Quant*, 56, 1023-1044. https://doi.org/10.1007/s11135-021-01164-0
- Tromble, R. (2021). Where have all the data gone? A critical reflection on academic digital research in the post-API age. *Social Media+ Society*, 7(1), 2056305121988929. https://doi.org/10.1177/2056305121988929

Live Session 2: Tuesday, 2 September, 2:00-5:30 pm CEST, {Zoom Link}
After answering some questions, you will split up into breakout groups to work on a small web scraping project.

3. Computational Text Analysis I

Self-learning Session 3: to be completed before Live Session 3 (see below), estimated workload: 3h
The advent of computational text analysis methods has reinvented the field of CSS. A lot of human interaction online is happening through textual data – at a scale that makes efforts to manually analyze it for answering theoretical questions essentially impossible. Computational text analysis thus takes up a prominent role in CSS. In this session, you will get an overview of the most important approaches, namely dictionary methods, and supervised and unsupervised machine learning. We will also introduce deep and transfer learning with a focus on using generative Large Language Models (LLMs) for topic modelling and sentiment analysis.

Literature (preliminary):

- Grimmer, J., & Stewart, B. M. (2013). Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis*, 21(3), 267-297. https://doi.org/10.1093/pan/mps028
- Boumans, J. W., & Trilling, D. (2016). Taking stock of the toolkit: an overview of relevant automated content analysis approaches and techniques for digital journalism scholars. *Digital Journalism*, 4(1), 8-23. https://doi.org/10.1080/21670811.2015.109659
- Atteveldt, W. van, Velden, M. A. C. G. van der, & Boukes, M. (2021). The Validity of Sentiment Analysis: Comparing Manual Annotation, Crowd-Coding, Dictionary Approaches, and Machine Learning Algorithms. *Communication Methods and Measures*, 15(2), 121-140. https://doi.org/10.1080/19312458.2020.1869198
- Welbers, K., Atteveldt, W. V., & Benoit, K. (2017). Text Analysis in R. *Communication Methods and Measures*, 11(4), 245-265. https://doi.org/10.1080/19312458.2017.1387238
- Spirling, A. (2023). Why open-source generative AI models are an ethical way forward for science. *Nature*. https://doi.org/10.1038/d41586-023-01295-4

- Weber, M., Reichardt, M. (2024). *Evaluation is all you need. Prompting Generative Large Language Models for Annotation Tasks in the Social Sciences. A Primer using Open Models*. https://doi.org/10.48550/arXiv.2401.00284

Live Session 3: Wednesday, 3 September, 2:00-5:30 pm CEST, {Zoom Link}
After answering some questions, you will split up into breakout groups to work on a small text analysis project.

4. Computational Text Analysis II

Self-learning Session 4: to be completed before Live Session 4 (see below), estimated workload: 3h
This second module on computational text analysis focuses on deep learning approaches, specifically the use of generative Large Language Models (LLMs) for topic modelling and sentiment analysis, and zero-shot learning for classification tasks.

Literature (preliminary):

- McLevey, J., & Crick, T. (2022). "Machine Learning and Neural Network Language Modelling for Sentiment Analysis" in Quan-Haase, A., & Sloan, L. (eds). *The Sage Handbook of Social Media Research Methods*, pp. 294-306. Routledge. https://doi.org/10.4135/9781529782943
- Spirling, A. (2023). Why open-source generative AI models are an ethical way forward for science. *Nature*, 616, 413. https://doi.org/10.1038/d41586-023-01295-4
- Weber, M., Reichardt, M. (2024). *Evaluation is all you need. Prompting Generative Large Language Models for Annotation Tasks in the Social Sciences. A Primer using Open Models*. https://doi.org/10.48550/arXiv.2401.00284

Live Session 4: Thursday, 4 September, 2:00-5:30 pm CEST, {Zoom Link}
After answering some questions, you will split up into breakout groups to work on a small text analysis project.

5. Computational Network Analysis

Self-learning Session 5: to be completed before Live Session 4 (see below), estimated workload: 3h
Social and political network analysis has been heavily influenced by developments in computational social science, such as the availability of massive timestamped related data (e.g., from social media) and the development of new computationally intensive modelling frameworks. This session will introduce you to the basic language, data, methods, and models used in network analysis. We will emphasize working with social media data and generative approaches to inferential network analysis.

Literature (preliminary):

- Atteveldt, W.v., Trilling D., & Arcila, C (2022). *Computational Analysis of Communication*. Chapter 13. https://v2.cssbook.net/content/chapter13
- Leifeld, P. (2017). "Discourse Network Analysis: Policy Debates as Dynamic Networks" in Victor, J.N., Montgomery, A. H., & Lubell, M. N. (eds). *The Oxford Handbook of Political Networks*, pp. 301-325. Oxford University Press. http://dx.doi.org/10.1093/oxfordhb/9780190228217.013.25
- Kitts, J., Grogan, H., & Lewis, K. (2023). "Social Networks and Computational Social Science" in McLevey, J., Scott, J., & Carrington, P. (eds). *The Sage Handbook of Social Network Analysis*, pp. 44-54. Sage. https://doi.org/10.4135/9781529614695

Live Session 5: Thursday, 4 September, 2:00-5:30 pm CEST, {Zoom Link}
After answering some questions, you will split up into breakout groups to work on a small network analysis project.