06-002-107-3: Forschungsseminar – Experimentelle Soziologie und Computational Social Science

Felix Lennert

Winter 2024/25

- E-mail: felix.lennert@uni-leipzig.de
- Course hours: Tuesdays, 15:15 – 16:45, NSG, SR 423, and Thursdays 11:15 – 12:45, NSG, SR 329; for exact dates, see schedule
- Readings: see schedule, available through Moodle
- Course materials: see website
- Student hours: are to be set up individually via email; there is a multitude of valid reasons why you should come over – some are listed here:
  - things are unclear and you need help with the material
  - you want to discuss a research idea
  - you have come across a cool new paper that I might deem interesting
  - you have recommendations for me in terms of general course resources/references
  - you want some career advice from someone roughly your age
  - you forgot your mensa card at home and want to steal some coffee/tea
  - in case you need some free period supplies, no email is required, you can just get them at Leonie Steinbrinker's office (I also have a key if she's not there), H3 1.06

Course description

The Forschungsseminar in Computational Social Science (CSS) equips you with the tools to analyze human behavior, predict social trends, and tackle complex societal issues using state-of-the-art data science techniques. From web scraping to AI-powered text analysis, you'll learn to harness the power of computation to gain new insights into social phenomena. The curriculum covers a range of topics including data management, web scraping, text analysis, analyzing spatial data, and agent-based simulation. Students will hone their R and develop skills in Python, applying these languages to real-world social science problems. The course progresses from fundamental concepts to advanced techniques, including the use of state-of-the-art AI models for text analysis. The course structure consists of one lecture and one workshop per week, providing a balance of theoretical knowledge and practical application. Throughout the semester, students will benefit from hands-on coding exercises, one-on-one mentoring, and collaborative projects. The course culminates in a group research paper, allowing students to apply their new skills to a topic of their choice. This course is ideal for social scientists looking to enhance their computational skills. It is geared towards 2nd year Master's students who are enrolled in the reformed Sociology Master's program. Interested Bachelor's students (semester 5 or higher) are also very welcome to attend, however, they will not be able to earn credits with their attendance.

What to expect

This course is structured so that it provides you – the student – with theory-heavy lectures (on Tuesdays) and hands-on R sessions (Thursdays). The lectures aim to introduce you to how you can use techniques to conduct empirical research and will mostly consist of the presentation of innovative and/or cutting-edge research that has harnessed digital data to solve exciting research puzzles. The practical sessions will be

delivered as videos. You are expected to watch these and work on exercises during the session. This ensures that students can work at their own pace and rewatch the content that might be unclear if needed.

The main objective of this course is that students **perform their own research as the course unfolds**. To this end, we will talk about your interests and then assign you to groups of matching research interests. Students are then expected to continue working on their projects as the course unfolds. Students can of course also work individually.

To ensure that nobody is stuck with their research, I require one-on-one meetings with each group (including groups of 1) during each week. These meetings are very casual, not graded, and can be thought of as a mere check-in. They also do not replace but rather complement office hours.

At the very end of the course, it is time for "peer reviewed presentations" of your projects. One of your peers' groups will be assigned your project and provide comments on it. This peer review ensures that you are right on track and that everyone has accomplished the course's learning goals. After that, you are all good to go and can check your analyses and write up your resulting papers so that they resemble a proper empirical research paper (see more in "Expectations" below).

The actual "peer-review" will work as follows: first, you present your work briefly (10 minutes). Then, the assigned reviewer group will provide their comments (5 minutes). They are supposed to give feedback on the question/motivation, the chosen theoretical angle, the data, and the method. The reviewers shall summarize each section briefly and point out what they perceive as strengths and weaknesses. Moreover, they can suggest how to frame the questions, provide further literature, or point out alternate methodological approaches that may be better suited to test the hypotheses.

To facilitate this, you must send presentations presenting the main idea, some theory (including hypotheses), a preliminary testing strategy, and the first results to me by **January 29, 2025**, so I can distribute them to your peers.

The deadline for the paper ("Forschungsbericht") is **March 21, 2025**. Please send it to Simone Müller (muellers@uni-leipzig.de) – and feel also free to CC me. Code and data have to be sent to me via email (depending on data size, data can also be shared via Google Drive/Dropbox/Uni Leipzig Wolke). The code should run "out of the box" and contain everything necessary to replicate the results and graphs in the paper. Preferably, this is in chronological order and structured into sections with descriptive titles.

Extensions

Extensions can be granted for particular reasons. These involve, among others, internships and sickness. In the case of the former, please give me a quick heads-up so that I can arrange it (preferably with some sort of proof). If you need an extension for a different reason than the ones mentioned above, feel free to reach out anytime, and I will do my best to accommodate your needs.

Expectations

- the basics
  - font size 12 pt, 1.5 line spacing
  - no typos, grammatical flaws, etc. (you are living in the age of helpers such as Grammarly, there are no more excuses)
  - length: between 4,000 and 8,000 words
  - cite correctly and in a uniform manner; my preferred citation style: ASA (it's strongly advised to use Zotero and Quarto/Overleaf; resources can be provided upon request)
- structured as an empirical research paper:
  - the *introduction* contains an empirical social scientific research question that is theoretically and practically motivated (i.e., showing its scientific and real-world relevance)
  - the *theory section* provides a **brief** overview of relevant prior research; clearly testable hypotheses are derived from the literature/goals for exploratory analyses are formulated

  - in *data and methods*, the data (including acquisition strategy), as well as the analysis strategy, are described; in our case, the data consist of text, the analyses are related to the course content; data and methods need to enable valid results
  - *results* need to be visualized through tables and/or (gg)plots and described in the text; tables and visualizations need to be properly labeled so that they can "stand on their own"
  - *discussion* of the results is performed in lieu of the theoretical foundations; potential shortcomings and reach of the paper are outlined
  - the *conclusion* circles back to the introduction and connects it to the results; it needs to clearly answer the research question

Basic rules of behavior

- If anything is unclear, ask me. This probably means that I have failed my job, and your question offers me a second chance to fix this.
- No discrimination. Never. If you witness any, tell me. I will find a way to deal with it.
- THIS IS IMPORTANT: If there are problems, reach out whenever. Do not let them become too big.
- Copy code from the internet – but you are responsible for the solution, so please make sure it works and solves your problem.
- Generative AI (i.e., ChatGPT et al.) is explicitly allowed. In my opinion, it is a tool that is here to stay, and you should use whatever resource you have to get the job(s) done. Plus, writing the right prompt is a skill in itself that you should definitely hone. If you use it for your writing, please make sure to proof-read everything properly, since you will be held accountable for both content and style.
- Form groups with your peers for working on the material. Everything will be easier and more fun. Except for when you have free riders. Kick them out of your group.
- AGAIN: ask questions if needed. Anytime.

Schedule

As stated above, Tuesdays are lectures and Thursdays are lab sessions. Please bring a laptop to all of our meetings (if you don't have one, feel free to reach out and we will try our best to lend you one).

Literature-wise, we will use a mix of textbooks and review papers to introduce theoretical concepts and related studies to illustrate. In terms of programming, we will mostly rely on online resources. However, everything that is relevant in terms of R content (and more!) can be found in the R script.

Every reading will be either provided online or linked to in this syllabus (just click on *"online"* – the link is hidden behind it.

I do not expect you to read the literature and will do the theoretical sessions in a "lecture" style. This is because this is an applied course and not every piece of content has the same relevance for everyone. Having been a student myself, I think that students should not be overwhelmed by having to read everything while working on their projects. I will upload additional readings on top of the ones this syllabus mentions in case you want to read more and need some inspo.

Week 1: Kick Off

Welcome & Housekeeping (Tue, 15 October 2024; 15:15 – 16:45; NSG, SR 423)

No readings.

Setting up your workstation (Thu, 17 October 2024; 11:15 – 12:45; NSG, SR 329)

- Acquire access to *sc.uni-leipzig.de*
- R recap (the corresponding chapters can be found in the R4DS book – *online*
  - `RMarkdown/Quarto` – chapters 28 & 29

– `dplyr` – chapter 4
– `tidyr` – chapter 6
– `ggplot2` – chapters 2 & 10 & 11 & 12
– `purrr` & loops in different flavors – chapter 27
– functional programming – chapter 26

Week 2: New Possibilities of CSS

Why CSS? (Tue, 22 October 2024; 15:15 – 16:45; NSG, SR 423)

- Jarvis, Keuschnigg, and Hedström (2021)
- Salganik (2018) – *online*, chapter 2

Regular Expressions (Thu, 24 October 2024; 11:15 – 12:45; NSG, SR 329)

- `stringr` & Regular Expressions – R4DS book, *online*, chapters 15 & 16

Week 3: Data Acquisition I

How the Web is Written (Tue, 29 October 2024; 15:15 – 16:45; NSG, SR 423)

- Stoltz and Taylor (2024) – chapter 5
- Blog post on CSS selectors – *online*
- Blog posts on API calls – *online*

`rvest/selenium` (Thu, 31 October 2024; 11:15 – 12:45; NSG, SR 329)

– This seems to be a public holiday, so work on the material if you find the time; if not, there should be enough time in the remaining weeks –

- set up `reticulate` in RStudio – *online*
- `httr2` documentation – *online*
- `rvest` Web scraping 101 – *online*
- `selenium` documentation – *online*

Week 4: Data Acquisition II

Optical Character Recognition and Transcription (Tue, 05 November 2024; 15:15 – 16:45; NSG, SR 423)

- Stoltz and Taylor (2024) – chapter 5

OpenAI Whisper/OCR (Thu, 07 November 2024; 11:15 – 12:45; NSG, SR 329)

- `Tesseract` documentation – *online*
- OpenAI Whisper Python package documentation – *online*

Week 5: Text as Data I

Bag of Words(Tue, 12 November 2024; 15:15 – 16:45; NSG, SR 423)

- Evans and Aceves (2016)
- Grimmer, Roberts, and Stewart (2022), chapters 3–5, 11, & 15
- Stoltz and Taylor (2024), chapters 4–9

Sentiment Analysis, TF-IDF, and NER/POS (Thu, 14 November 2024; 11:15 – 12:45; NSG, SR 329)

- Grimmer et al. (2022), chapter 11
- Jurafsky and Martin (n.d.), chapter 21 – *online*
- Silge and Robinson (2017) – *online*, chapters 2 & 3

Week 6: Text as Data II

Machine Learning (Tue, 19 November 2024; 15:15 – 16:45; NSG, SR 423)

Supervised ML

- Barberá et al. (2021)
- Grimmer et al. (2022), chapters 17–20
- Stoltz and Taylor (2024), chapters 9 & 12

Unsupervised ML

- Blei (2012)
- DiMaggio, Nag, and Blei (2013)
- Grimmer et al. (2022), chapters 10, 12–3
- Stoltz and Taylor (2024), chapters 10 & 11

Classification and Topic Modeling (Thu, 21 November 2024; 11:15 – 12:45; NSG, SR 329)

- Hvitfeldt and Silge (2022) – *online*, chapters 6 & 7
- Silge and Robinson (2017) – *online*, chapter 6
- Silge and Hvitfeldt (2019) – *online*

Week 7

No classes.

Week 8: Text as Data III

Distributional Hypothesis (Tue, 03 December 2024; 15:15 – 16:45; NSG, SR 423)

- Jurafsky and Martin (n.d.), chapter 6 – *online*
- Stoltz and Taylor (2021)

Word Embeddings (Thu, 05 December 2024; 11:15 – 12:45; NSG, SR 329)

- Hvitfeldt and Silge (2022) – *online*, chapter 5
- Stoltz and Taylor (2024), chapter 11
- `text2map`: R Tools for Text Matrices – *online*

Week 9: Text as Data IV

New Developments in NLP (Tue, 10 December 2024; 15:15 – 16:45; NSG, SR 423)

- Do, Ollion, and Shen (2022)
- Laurer et al. (2024)
- Törnberg (2023)
- Wankmüller (2022)

BERT/GPT/NLI (Thu, 12 December 2024; 11:15 – 12:45; NSG, SR 329)

- set up environments
- Augmented Social Scientist tutorial – *online*
- BERTopic – *online*
- `gptstudio` plugin for RStudio – *online* and `gpttools` plugin for RStudio – *online*
- Repository for Laurer et al. 2024 – *online*

Week 10: Spatial Data I

Basics in Spatial Data Analysis (Tue, 17 December 2024; 15:15 – 16:45; NSG, SR 423)

- Logan (2012)

Working with Geo Data (Thu, 19 December 2024; 11:15 – 12:45; NSG, SR 329)

- Lovelace, Nowosad, and Muenchow (2025) – *online*

Weeks 11 & 12

Christmas break.

Week 13: Spatial Data II

Modeling Spatial Data (Tue, 07 January 2025; 15:15 – 16:45; NSG, SR 423)

- LeSage (2014)

Weighting/Autocorrelation/Regression (Thu, 09 January 2025; 11:15 – 12:45; NSG, SR 329)

- Lovelace et al. (2025) – *online*
- Tutorial on Spatial Regression Analysis – *online*

Week 14: Simulation I

Agent-based Modeling (Tue, 14 January 2025; 15:15 – 16:45; NSG, SR 423)

- Arvidsson et al. (2024)
- Flache and Macy (2011)

ABMs in R (Thu, 16 January 2025; 11:15 – 12:45; NSG, SR 329)

- Acerbi, Mesoudi, and Smolla (2020) – *online*

Week 15: Simulation II

Empirical Calibration (Tue, 21 January 2025; 15:15 – 16:45; NSG, SR 423)

- Bruch and Atwell (2015)

Empirically Calibrated ABMs in R (Thu, 23 January 2025; 11:15 – 12:45; NSG, SR 329)

- Axelrod (1997)
- Acerbi et al. (2020) – *online*

Week 16: Presentation Preparation Week

No classes. Deadline for sending presentations: January 29, 6PM.

Week 17: Presentation & Wrap Up Week

Presentations (Tue, 04 February 2025; 15:15 – 16:45; NSG, SR 423)

No readings.

Presentations & Wrap-up (Thu, 06 February 2025; 11:15 – 12:45; NSG, SR 329)

No readings.

Deadline Forschungsbericht

March 21, 2025.

References

Acerbi, Alberto, Alex Mesoudi, and Marco Smolla. 2020. "Individual-Based Models of Cultural Evolution. A Step-by-Step Guide Using R."

Arvidsson, Martin, Peter Hedström, Benjamin Jarvis, and Marc Keuschnigg. 2024. "On the Intersection of Analytical Sociology and Computational Social Science."

Axelrod, Robert. 1997. "The Dissemination of Culture: A Model with Local Convergence and Global Polarization." *The Journal of Conflict Resolution* 41(2):203–26.

Barberá, Pablo, Amber E. Boydstun, Suzanna Linn, Ryan McMahon, and Jonathan Nagler. 2021. "Automated Text Classification of News Articles: A Practical Guide." *Political Analysis* 29(1):19–42.

Blei, David. 2012. "Probabilistic Topic Models." *Communications of the ACM* 55(4):77–84.

Bruch, Elizabeth and Jon Atwell. 2015. "Agent-Based Models in Empirical Social Research." *Sociological Methods & Research* 44(2):186–221.

DiMaggio, Paul, Manish Nag, and David Blei. 2013. "Exploiting Affinities Between Topic Modeling and the Sociological Perspective on Culture: Application to Newspaper Coverage of U.S. Government Arts Funding." *Poetics* 41(6):570–606.

Do, Salomé, Étienne Ollion, and Rubing Shen. 2022. "The Augmented Social Scientist: Using Sequential Transfer Learning to Annotate Millions of Texts with Human-Level Accuracy." *Sociological Methods & Research* 004912412211345.

Evans, James A. and Pedro Aceves. 2016. "Machine Translation: Mining Text for Social Theory." *Annual Review of Sociology* 42(1):21–50.

Flache, Andreas and Michael Macy. 2011. "Social Dynamics from the Bottom Up: Agent-Based Models of Social Interaction." in *The Oxford Handbook of Analytical Sociology*. Oxford University Press.

Grimmer, Justin, Margaret Roberts, and Brandon Stewart. 2022. *Text as Data: A New Framework for Machine Learning and the Social Sciences*. Princeton: Princeton University Press.

Hvitfeldt, Emil and Julia Silge. 2022. *Supervised Machine Learning for Text Analysis in R*. First edition. Boca Raton London New York: CRC Press, Taylor & Francis Group.

Jarvis, Benjamin F., Marc Keuschnigg, and Peter Hedström. 2021. "Analytical Sociology Amidst a Computational Social Science Revolution." Pp. 33–52 in *Handbook of Computational Social Science, Volume 1*. London: Routledge.

Jurafsky, Dan and James Martin. n.d. "Speech and Language Processing."

Laurer, Moritz, Wouter Van Atteveldt, Andreu Casas, and Kasper Welbers. 2024. "Less Annotating, More Classifying: Addressing the Data Scarcity Issue of Supervised Machine Learning with Deep Transfer Learning and BERT-NLI." *Political Analysis* 32(1):84–100.

LeSage, James P. 2014. "What Regional Scientists Need to Know About Spatial Econometrics." *SSRN Electronic Journal*.

Logan, John R. 2012. "Making a Place for Space: Spatial Thinking in Social Science." *Annual Review of Sociology* 38(1):507–24.

Lovelace, Robin, Jakub Nowosad, and Jannes Muenchow. 2025. *Geocomputation with R*. 2nd ed. CRC Press.

Salganik, Matthew J. 2018. *Bit By Bit: Social Research in the Digital Age*. Princeton: Princeton University Press.

Silge, Julia and Emil Hvitfeldt. 2019. "Predictive Modeling with Text Using Tidy Data Principles." in *useR2020*.

Silge, Julia and David Robinson. 2017. *Text Mining with R: A Tidy Approach*. First edition. Beijing ; Boston: O'Reilly.

Stoltz, Dustin S. and Marshall A. Taylor. 2021. "Cultural Cartography with Word Embeddings." *Poetics* 88:101567.

Stoltz, Dustin S. and Marshall A. Taylor. 2024. *Mapping Texts: Computational Text Analysis for the Social Sciences*. New York, NY: Oxford University Press.

Törnberg, Petter. 2023. "How to Use LLMs for Text Analysis."

Wankmüller, Sandra. 2022. "Introduction to Neural Transfer Learning With Transformers for Social Science Text Analysis." *Sociological Methods & Research* 004912412211345.