Introduction to Computational Social Science methods with Python

Haiko Lietz
(Editor and author)
GESIS, Cologne, Germany
haiko.lietz@gesis.org

N. Gizem Bacaksizlar Turbic
(Editor and author)
GESIS, Cologne, Germany
ngizembacaksizlar@gmail.com

Pouria Mirelmi
(Author)
GESIS, Cologne, Germany
pouria.mirelmi@gesis.org

Olga Zagovora
(Author)
GESIS, Cologne, Germany
olgazgovora@gmail.com

Nicolò Gozzi
(Author)
ISI Foundation, Torino, Italy
nic.gozzi@gmail.com

August 18, 2023

Syllabus

The digitization of large domains of society is also transforming the social sciences. Computational Social Science (CSS) is the field concerned with computational approaches to social science problems. In its current form, it combines data, computation, and theory. As such, it is a trading ground for the fields of Quantitative Social Science (QSS), Data Science (DS), and Social Simulation. This paradigm shift requires researchers to rethink fundamental assumptions and to have skills classically not taught in social science curricula: the individual is not the fundamental unit of observation anymore; the repertoire of statistical and analytical modeling is extended by methods from the areas of Machine Learning (ML), Natural Language Processing (NLP), network analysis, and numerical modeling; and theories must be adapted to a changing world.

This syllabus contains a full introductory course to CSS methods with Python. Teaching materials meet the criteria of a gradable university course, are fully online, self-explanatory, and freely available: the materials combine coding tutorials with recom-

mended readings, specific teaching lessons, and experience-based guidelines; they are housed in a public GitHub repository at `https://github.com/gesiscss/css_methods_python`, which means, everybody can study them; they have the form of Jupyter Notebooks, which means, they have the look and feel of a manuscript, yet, they contain Python code that is fully executable in a browser window, potentially without the need to locally install Python; and they are available under a Creative Commons license which allows you to freely share and adapt them. The course consists of sessions that gradually lead participants to acquire more skills in Python.

The course consists of four sections. The first section teaches how to set up a computing infrastructure and conveys basic data management and scientific computing skills. The second section teaches students how to collect data using dedicated Python packages for using Application Programming Interfaces (APIs) and web scraping. The third section focusses on data preprocessing methods from network analysis and NLP and includes applications of Large Language Models (LLMs). The fourth section is about data analysis methods and goes into depth with network analysis and modeling, unsupervised and supervised ML, as well as topic modeling. Some datasets are repeatedly used throughout the course, among them a corpus of tweets on the tropic of COVID (TweetsCOV19) from May 2020, social networks from the Copenhagen Networks Study (CNS), and the Varieties of Democracy (V-Dem) dataset on countries and principles of democracy. Whenever possible, sessions are interlinked and build on top of each other.

The main textbook used in the course is McLevey's *Doing Computational Social Science: A Practical Introduction* (2022), a rather complete introduction to methods in the field with well-structured and insightful chapters. The network preprocessing and analysis sessions benefit from Platt's *Network Science with Python and NetworkX Quick Start Guide* (2019) and Menczer et al.'s *A First Course in Network Science* (2020). All three references come with publicly available Jupyter Notebooks that can serve as additional teaching resources. In sessions on DS, statistical modeling, and ML methods, chapters from the *Handbook of Computational Social Science*, volume 2, by (Engel et al., 2022) give depth. Python is chosen as the programing language because the course has a much stronger focus on ML than on statistical modeling.

These resources have been developed as part of the Social ComQuant project (`https://socialcomquant.ku.edu.tr/`) which had been funded as a twinning project among Koç University (Istanbul, Turkey), GESIS – Leibniz Institute for the Social Sciences (Cologne, Germany), and the ISI Foundation (Torino, Italy) under the European Commission's Horizon 2020 funding line.

Section A: Introduction

Session A1: Computing infrastructure

In this session, you will learn how to set up your computing infrastructure. All tools you will use are free of charge. In subsession **A1.1**, we will guide you through the installation of the Anaconda distribution. We specify a particular release of the distribution so all users of the course materials will work with the same set of package versions and not

experience any problems. You will learn how to create a custom environment that has all packages installed in the required versions and to open Jupyter Notebooks. In subsession **A1.2**, we will teach the neccessary steps to use Git and download the teaching materials from GitHub to your local computer. If you want to work in the cloud – potentially skipping the installation of Anaconda and Git –, either launch Binder or go to subsession **A1.3** to see how you can set yourself up on Google Colab.

Readings

- Heiberger and Riebling (2016)
- McLevey (2022, ch. 1)

Session A2: Data management and relational databases

In this session, you will learn how to manage your data and keep it tidy while keeping a focus on your research questions. We will have a deep look at the 2-dimensional table as the fundamental data structure we will work with throughout all sessions. In subsession **A2.1**, we start with some illustrative toy examples about important dataframe properties. In subsession **A2.2**, we enter the almost-big-data world using the TweetsCOV19 dataset. You will experience how you can use Pandas to handle tables and mimic a relational database in such a way that your data gets ready for analysis. You will see what it means that relational databases eliminate redundancy and ensure consistency. The TweetsCOV19 dataset will function as an example that will shine up repeatedly in this and subsequent sessions. In subsession **A2.3**, you will see how you can save processed data to multiple files, an SQL database, or Excel. Subsession **A2.4** is dedicated to how data can be retrieved from the relational data structure we have created.

Readings

- McLevey (2022, ch. 6)
- Weidmann (2023, chs. 1, 3, and 8)

Session A3: Scientific computing and data visualization

In this session, you will learn basic steps of scientific computing and data visualization. The TweetsCOV19 dataset will continue to function as the example. In subsession **A3.1**, we will introduce the NumPy package to work with n-dimensional arrays which are typically needed in the data processing. In subsession **A3.2**, we will get to know SciPy and how to handle sparse data with it. Finally, in subsession **A3.3**, you will learn how to use the Matplotlib and Seaborn packages to explore data visually.

Readings

- McLevey (2022, ch. 7)

- Sundnes (2020, ch. 6)

Section B: Data collection methods

Session B1: API harvesting

In this session, you will learn how to collect Digital Behavioral Data via API harvesting. In subsession **B1.1**, we will list resources for social media APIs and for datasets that have already been collected. In subsession **B1.2**, we will dive into harvesting Wikipedia, introducing a few APIs that help with collecting various parts of Wikipedia pages. Finally, in subsession **B1.3**, we will discuss the Total Error Sheets for Datasets (TES-D) framework to document a Twitter dataset.

Readings

- Jünger (2022)
- McLevey (2022, ch. 4)

Session B2: Data parsing and static web scraping

In this session, you will learn how to do web scraping with Python from scratch. In subsession **B2.1**, we will convey the basic knowledge you need for building a web scraper by taking a deep look at how HTML is structured. In subsession **B2.2**, we will introduce the Beautiful Soup package. You will learn its basic functions and how you can scrape a news site, taking Aljazeera as an example. We close, in subsession **B2.3**, with a demonstration how you can parse an RSS feed.

Readings

- Bosse et al. (2022)
- McLevey (2022, ch. 5)

Session B3: Dynamic web scraping

In this session, you will learn how to collect webpage content that is dynamically generated. In subsession **B3.1**, we will get to know the Selenium package and learn how to automate web browsing with it. In subsession **B3.2**, we will work on practical examples: scraping questions and answers from the Quora platform.

Readings

- Bosse et al. (2022)
- McLevey (2022, ch. 5)

Section C: Data preprocessing methods

Session C1: Network construction and visualization

In this session, you will learn how you can construct undirected and directed networks, enrich them with attributes, and draw them such that they are ready for publication, all using the NetworkX package. In subsession **C1.1**, we will get to know graphs with undirected or directed and unweighted or weighted edges. We will use Pandas package to organize network data as nodelists and edgelists and see how attributes stored in those tables can be visualized and internalized in graph objects. Almost all throughout the whole session, we will be working with the CNS interaction data. In subsession **C1.2**, you will learn the essentials of network drawing and relational information visualization. Finally, in subsession **C1.3**, you will learn how to make networks simpler by removing certain pieces of information and how you can save graph objects to use them in another notebook or software.

Readings

- Krempel (2011)
- Platt (2019, chs. 2 and 3)

Session C2: Multilayer and multimodal network construction

In this session, you will learn to work with multilayer and multimodal networks and how those enable you to think relationally. We will continue using the NetworkX package. Subsession **C2.1** is dedicated to multilayer networks, creating a communication network with text message and phone call layers. In subsession **C2.2**, we approach the dynamic nature of networks and see how different kinds of snapshots can be constructed from temporal aggregation and combined in a multilayer graph. Subsession **C2.3** is about bipartite networks and about how nodes of one mode can be projected into edges among nodes of the other mode. Finally, in subsession **C2.4**, we will see how we remove the layer information from graphs.

Readings

- Platt (2019, chs. 4 and 10)

Session C3: Natural Language Processing

...

Readings

- ...

Section D: Data analysis methods

Session D1: Micro-level network analysis and community detection

In this session, you will learn how to perform micro- and meso-level network analysis. We will continue to use the NetworkX package and the networks from the Copenhagen Networks Study constructed in section C. In subsession **D1.1**, we will demonstrate the "big 3" algorithms of centrality analysis. We will get to know basic graph-theoretical concepts to define them, and we will use variants of a social communication network to see how that effects centrality scores. Subsession **D1.2** is dedicated to closure and brokerage and how two scores from that area relate to centrality indices. Finally, in subsession **D1.3**, you will learn to handle two main community detection algorithms and the pitfalls and options that come with their use. Throughout the session, we will encounter multilayer networks both as a way to store data and as a representation of the social world.

Readings

- Fuhse (2021, ch. 2)
- McLevey (2022, ch. 15)
- Menczer et al. (2020, ch. 6)

Session D2: Macro-level network analysis and network modeling

In this session, you will learn how to analyze the macro level of networks and model them. We will continue to use the NetworkX package and the networks from the Copenhagen Networks Study constructed in section C. The first three subsessions are dedicated to notions of network connectivity. Subsession **D2.1** revolves around the structural study of social cohesion. You will learn to detect core/periphery structure and quantify the extent of segregation, both numerically and categorically. In subsession **D2.2**, we take a look at the Small World and how to measure the extent of separation in a network. Subsession **D2.3** deals with inequality. We will see how unevenly degrees are distributed in a network and how such distributions can be characterized. Subsession **D2.4** is dedicated to network modeling. It builds on the previous subsessions and introduces three very different network models that exhibit self-organization or emergence. Finally, in subsession **D2.5**, we extent pattern analysis to the cultural case of a large hashtag co-occurrence network.

Readings

- Fuhse (2021, ch. 7)
- Menczer et al. (2020, chs. 2 and 5)
- Page (2015)

Session D3: Unsupervised machine learning

In this session, you will learn how to perform unsupervised learning. In subsession **D3.1**, we will introduce the V-Dem dataset, its hierarchy of variables, and how they will be used for ML. This dataset will serve to demonstrate all algorithms in this session. In subsession **D3.2** on dimensionality reduction, you will learn how to do infer latent variables using Factor Analysis and Principal Component Analysis, how these methods differ in the way theory is used, and how we can think of them as abductive inference tools. Subsession **D3.3** is dedicated to clustering, and you will see how different algorithms, namely K-Means and Agglomerative Clustering, lead to different solutions and interpreations.

Readings

- Bacher et al. (2022)
- McLevey (2022, chs. 8 and 19)

Session D4: Topic modeling

...

Readings

- ...

Session D5: Supervised machine learning

In this session, you will learn how to perform supervised learning and how it differs from the way algorithms are used in SSH. The V-Dem dataset will serve to demonstrate all algorithms in this session. Subsession **D5.1** focuses on explanation. We will demonstrate how the traditional way of statistical modeling can be done in statsmodels. We will guide you through Linear and Logistic Regression and how coefficients and significance scores can be obtained. The major part of the session, subsession **D5.2**, focuses on prediction and the scikit-learn package. You will learn about the order of operations (training, cross-validation, and testing), evaluation metrics, how to identify overfitting, and how to avoid it using model tuning techniques. In terms of algorithms, we will cover Linear Regression, Deep Learning, Decision Trees, and Gradient Boosting, first for regression and then for classification.

Readings

- De Veaux and Eck (2022)
- McLevey (2022, chs. 20, 21, and 22)

References

J. Bacher, A. Pöge, and K. Wenzig. Unsupervised methods: Clustering methods. In U. Engel, A. Quan-Haase, S. Liu, and L. Lyberg, editors, *Handbook of Computational Social Science*, volume 2, pages 334–351. Routledge, 2022.

S. Bosse, L. Dahlhaus, and U. Engel. Web data mining: Collecting textual data from web pages using R. In U. Engel, A. Quan-Haase, S. Liu, and L. Lyberg, editors, *Handbook of Computational Social Science*, volume 2, pages 46–70. Routledge, 2022.

R. D. De Veaux and A. Eck. Machine Learning methods for Computational Social Science. In U. Engel, A. Quan-Haase, S. Liu, and L. Lyberg, editors, *Handbook of Computational Social Science*, volume 2, pages 291–321. Routledge, 2022.

U. Engel, A. Quan-Haase, S. X. Liu, and L. Lyberg. *Handbook of Computational Social Science, Volume 2: Data Science, Statistical Modelling, and Machine Learning Methods*. Routledge, London, 2022. doi: 10.4324/9781003025245.

J. A. Fuhse. *Social Networks of Meaning and Communication*. Oxford University Press, New York, NY, 2021. **Accessible account of sociological network theory that delves deep into the mutual constitution of social networks and communication networks.**

R. H. Heiberger and J. R. Riebling. Installing computational social science: Facing the challenges of new information and communication technologies in social science. *Methodological Innovations*, 9(0), 2016. doi: 10.1177/2059799115622763. **A comparison of network analysis libraries in Python and R in terms of speed and a general discussion of computation in social science.**

J. Jünger. A Brief History of APIs: Limitations and opportunities for online research. In U. Engel, A. Quan-Haase, S. Liu, and L. Lyberg, editors, *Handbook of Computational Social Science*, volume 2, pages 17–32. Routledge, 2022.

L. Krempel. Network visualization. In J. Scott and P. J. Carrington, editors, *The SAGE Handbook of Social Network Analysis*, pages 558–577. Sage, 2011. **A systematic overview how to visualize rich network information.**

J. McLevey. *Doing Computational Social Science: A Practical Introduction*. SAGE, Los Angeles, CA, 2022. **A rather complete introduction to the field with well-structured and insightful chapters. The website (`https://github.com/UWNETLAB/dcss_supplementary`) offers the code used in the book.**

F. Menczer, S. Fortunato, and C. A. Davis. *A First Course in Network Science*. Cambridge University Press, 2020. **An introductory course with exercises that emerged from years of teaching network analysis using NetworkX. Written from the physicist's perspective, this book is focused on the network**

science paradigms of small-world networks, scale-free networks, community detection, and complex systems modeling. The website (`https://cambridgeuniversitypress.github.io/FirstCourseNetworkScience/`) provides all code in the form of Jupyter Notebooks, data, and solutions to the exercises.

S. E. Page. What sociologists should know about complexity. *Annual Review of Sociology*, 41(1):21–41, 2015. doi: 10.1146/annurev-soc-073014-112230. **The title is what you get. The author reviews concepts from complexity science that create new perspectives on the structure and dynamics of (not just) networks.**

E. L. Platt. *Network Science With Python and NetworkX Quick Start Guide: Explore and Visualize Network Data Effectively.* Packt Publishing, Birmingham, UK, 2019. **Systematic introduction to the practice of network preprocessing and analysis. All Jupyter Notebooks are publically available on the website (`https://github.com/PacktPublishing/Network-Science-with-Python-and-NetworkX-Quick-Start-Guide`).**

J. Sundnes. *Introduction to Scientific Programming with Python.* Springer International Publishing, Cham, 2020. **An openly accessible introduction covering the basic functionalities of Python. The website (`https://sundnes.github.io/python_intro/`) offers the code used in the book.**

N. B. Weidmann. *Data Management for Social Scientists: From Files to Databases.* Cambridge University Press, 2023. **A fresh account of data management and processing. The book uses R, but general insights also apply to Python.**