

INTRODUCTION TO BIG DATA FOR SOCIAL SCIENCE
JPSM Short Course 2024
November 13-15

Jun.-Prof. Dr. Christoph Kern
Junior Professor at LMU Munich, Assistant Research Professor at UMD
christoph.kern.kern@stat.uni-muenchen.de

Dr. Anna-Carolina Haensch
Assistant Research Professor at UMD, Post-Doc at LMU Munich
anna-carolina.haensch@stat.uni-muenchen.de

Course Abstract: Social scientists and survey researchers are increasingly faced with integrating data from multiple data sources and are expanding their activities beyond experiments and surveys. This course will discuss these developments and provide practical guidance on combining methods and tools from computer science, statistics, and social science. In particular, techniques such as web scraping and machine learning applications will be discussed, but also methods for working with text data, collecting data with APIs and questions around big data quality. The course presents the key big data tools in a non-intimidating way to social and data scientists while not neglecting research questions and purposes. The course aims to illustrate social science and data science principles through real-world applications, and links computer science concepts to real social science as well as survey research. Data and code to practice will be available on GitHub.

Who should attend: The course is designed for individuals in government, universities, private sector, and nonprofit organizations who are interested in understanding the fundamental concepts applying new analytical techniques to social science issues. This course gives practical guidance to those who have never performed their own data analytics project but who are planning to do so, and those who have performed data analytics projects but would like to learn more the underlying methodology. The instructors will operate under the assumption that students have no prior knowledge of the topic.

The Instructors: *Christoph Kern* is Junior Professor of Social Data Science and Statistical Learning at the Ludwig-Maximilians-University of Munich and Project Director at the Mannheim Centre for European Social Research (MZES). He received his PhD in social science (Dr. rer. pol.) from the University of Duisburg-Essen in 2016. Before joining LMU Munich, he was a Post-Doctoral Researcher at the Professorship for Statistics and Methodology at the University of Mannheim and Research Assistant Professor at the Joint Program in Survey Methodology (JPSM) at the University of Maryland. His work focuses on the reliable use of machine learning methods and new data sources in social science, survey research, and algorithmic fairness.

Anna-Carolina Haensch is an Assistant Research Professor at the University of Maryland, where she teaches statistics and method courses for the Joint Program in Survey Methodology (JPSM). She is also a Researcher ("Akademische Rätin") at Ludwig-Maximilians-Universität München (LMU), where she leads the Introduction to Statistics lecture and

contributes to various projects at the Chair for Statistics and Data Science in Social Sciences and the Humanities. Her research interests span survey methodology, synthetic data, and large language models, and their implications for society. **Textbook:** Foster, I., Ghani, R., Jarmin, R. S., Kreuter, F., and Lane, J. (Eds.). (2020). *Big Data and Social Science: Data Science Methods and Tools for Research and Practice*. 2nd edition. Boca Raton, FL: CRC Press Taylor & Francis Group.
<https://textbook.coleridgeinitiative.org/>

Course Outline:

Date	Content
Nov 13, 8.30-9.30am	Introduction – New Data and Methods for Social Science
Nov 13, 9.30-10.30am	Web Scraping and APIs I
Nov 13, 10.30-10.45am	<i>Break</i>
Nov 13, 10.45-11.30am	Web Scraping and APIs II (practical session)
Nov 13, 11.30am-12.30pm	Big Data Visualization and Interactive Graphs
Nov 14, 8.30-9.30am	Text Mining and Topic Modelling I
Nov 14, 9.30-10.30am	Text Mining and Topic Modelling II (practical session)
Nov 14, 10.30-10.45am	<i>Break</i>
Nov 14, 10.45-11.30am	Foundations of Machine Learning I
Nov 14, 11.30am-12.30pm	Foundations of Machine Learning II (practical session)
Nov 15, 8.30-9.30am	Machine Learning Methods I
Nov 15, 9.30-10.30am	Machine Learning Methods II (practical session)
Nov 15, 10.30-10.45am	<i>Break</i>
Nov 15, 10.45-11.30am	Applications of Machine Learning
Nov 15, 11.30am-12.30pm	Big Data Quality