gesis Leibniz Institute for the Social Sciences

GESIS Summer School in Survey Methodology 2022

Syllabus for course:
"Data Science Techniques for Survey Researchers"

Lecturers:	Dr. Christoph Kern	Dr. Malte Schierholz
E-mail:	c.kern@uni-mannheim.de	Malte.schierholz@stat.uni-muenchen.de

Date: 08-12 August 2022
Time: 10:00-13:00 | 14:30-17:30
Venue: Online via Zoom

About the Lecturers:

Christoph Kern is 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. He also is a Project Director at the Mannheim Centre for European Social Research (MZES) and member of the Mannheim Center for Data Science (MCDS). He received his PhD (Dr. rer. pol.) in social science from the University of Duisburg-Essen (UDE) in 2016. Before joining the University of Mannheim, he was a research associate at the Professorship for Empirical Social Science Research and Statistics at UDE.

Malte Schierholz is a Post-Doctoral Researcher at the Chair for Statistics and Data Science in Social Sciences and the Humanities at Ludwigs-Maximilians-Universität München. Prior to that, he worked at the Institute for Employment Research in Nuremberg and was a visiting scholar at the Machine Learning Department at Carnegie Mellon University. With a Master in Statistics and a PhD (Dr rer. Soc.) from the University of Mannheim in sociology, he enjoys thinking about how social scientists can best leverage the promises and benefits of machine learning.

Selected Publications:

- Kim, M. P., Kern, C., Goldwasser, S., Kreuter, F. and Reingold, O. (2022). Universal Adaptability: Target-Independent Inference that Competes with Propensity Scoring. *Proceedings of the National Academy of Sciences of the United States of America (PNAS)* 119(4). https://doi.org/10.1073/pnas.2108097119
- Kern, C., Weiss, B., and Kolb, J.-P. (2021). Predicting Nonresponse in Future Waves of a Probability-based Mixed-mode Panel with Machine Learning. *Journal of Survey Statistics and Methodology*. https://doi.org/10.1093/jssam/smab009
- Kern, C., Klausch, T., and Kreuter, F. (2019). Tree-based Machine Learning Methods for Survey Research. *Survey Research Methods* 13(1), 73-93. https://doi.org/10.18148/srm/2019.v1i1.7395
- Schierholz, M., Gensicke, M., Tschersich, N., and Kreuter, F. (2018). Occupation Coding During the Interview. *Journal of the Royal Statistical Society: Series A*, 181(2), 379–407. https://doi.org/10.1111/rssa.12297
- Schierholz, M. and Schonlau, M. (2021). Machine Learning for Occupation Coding—A Comparison Study, *Journal of Survey Statistics and Methodology* 9(5), 1013–1034, https://doi.org/10.1093/jssam/smaa023
- Fitzenberger, B., Kagerl, C., Schierholz, M., Stegmaier, J. (2021): Zeitnahe Daten in der Corona-Krise: Von der schwierigen Vermessung der Kurzarbeit. (IAB-Kurzbericht, 24/2021), Nürnberg, 12 S.

Course Description:

A variety of digital data sources are providing new avenues for empirical social science research. In order to effectively utilize these data for answering substantive research questions, a modern methodological toolkit

paired with a critical perspective on data quality is needed. Organized and offered in collaboration with BERD@NFDI, this course introduces state-of-the-art data science techniques that are suited for collecting and analyzing digital behavioral data, so-called "big data", and traditional survey data. In addition, aspects of data quality and error frameworks for digital (big) data sources are discussed.

The course will cover the following topics and techniques:

- Overview of Big Data: What is it and why does it matter?
- Total Survey Error for Big Data
- Git and GitHub
- Web Scraping
- Data bases and SQL
- Data quality for gathered data types
- Sampling from online material (e.g., Twitter)
- (Supervised) Machine Learning for Social Scientists, including:
    - Regularized Regression
    - Decision Trees and Random Forest
    - Boosting
    - Applications
- Working with textual data: Text Mining and Topic Models

After the course, participants will have a profound understanding of important methods from the data science toolkit for collecting and analyzing the data types mentioned. They will be able to apply these methods and techniques in their research using statistical software.

Keywords:

Data Science, Big Data, Machine Learning, Text Mining, Total Survey Error, online

Course Prerequisites:

- general knowledge of statistics and statistical modelling (i.e., regression)
- prior experiences with syntax-based software (like R, Stata, or Python)

Some basic experience with programming in R is helpful, but not strictly necessary. For those without prior exposure to R, we will ensure everyone is able to execute R markdown files. Students without any R knowledge are encouraged to work through one or more R tutorials prior of the course. Some resources can be found here:
https://rstudio.cloud/learn/primers
http://www.statmethods.net/
https://swirlstats.com/
https://rmarkdown.rstudio.com/lesson-1.html (for R Markdown)

Target Group:

Participants will find the course useful if:

- they are interested in learning some fundamental techniques in data science
- they want to collect and work with digital behavioural data, be it administrative data or data found online
- they want to understand what (supervised) machine learning is

Course and Learning Objectives:

By the end of the course participants will:

- Understand the challenges when analyzing digital behavioural data
- Be familiar with some of the software tools commonly used to analyze such data
- Know additional data science tools and techniques
- Know the promises and benefits of (supervised) machine learning
- Be able to use (supervised) machine learning for data analysis

- Be able to use common routines for analyzing textual data
- Learn some of the metrics used to assess data quality for gathered data types
- Learn about two different approaches to sampling Twitter.

Organizational Structure of the Course:

The course is partly theoretical, partly practical. Each topic will be introduced in a lecture. The best way to deepen one´s understanding is with practical hands-on exercises. Files written in R Markdown will be provided to help participants execute the prepared scripts on their own computer and complete the assignments. The instructors are available to assist and answer questions during the practical sessions.

Software and Hardware Requirements:

We will use R for the practical sessions. Please have R and RStudio installed on your computer. Both programs are free and open source. We will inform you a few days before the course starts about recommended steps to setup your system. You should be able to access the internet and install additional packages during the course.

Day-to-day Schedule and Literature:

Day	Topic(s)
1 (morning)	**Course introduction** **New data sources in the digital age (Guest speaker: Prof. Dr. Frauke Kreuter)** **New techniques to analyse such data**
	Suggested reading (suggested, yet do not have to be read before class): ▪ Chapter 1 and 10 in Foster, I., Ghani, R., Jarmin, R.S., Kreuter, F., & Lane, J. (2020) ▪ Salganik, Matthew J. 2018. Bit by Bit: Social Research in the Digital Age. Princeton: Princeton University Press. https://www.bitbybitbook.com/en/1st-ed/preface/.
1 (afternoon)	**git and GitHub** **Web scraping, html, xml, json, APIs, regular expressions**
	Suggested reading: ▪ Chapter 2 in Foster, I., Ghani, R., Jarmin, R.S., Kreuter, F., & Lane, J. (2020) ▪ Part One in Munzert, Simon, Christian Rubba, Peter Meißner, and Dominic Nyhuis. 2014. Automated Data Collection with R: A Practical Guide to Web Scraping and Text Mining. Chichester, UK: John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118834732.
2 (morning)	**Data bases and SQL**
	Suggested reading: ▪ Chapter 4 in Foster, I., Ghani, R., Jarmin, R.S., Kreuter, F., & Lane, J. (2020) ▪ Chapter 15 in Baumer, B.S., Kaplan, D.T., & Horton, N.J. (2021)
2 (afternoon)	**Machine Learning: Regression with Lasso penalization, Variable selection, Recursive partitioning and decision trees**
	Suggested reading: ▪ Chapter 6 and 8 in James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). An Introduction to Statistical Learning. New York, NY: Springer. ▪ Hastie, T., Tibshirani, R. and Wainwright, M. (2015). Statistical Learning with Sparsity. Boca Raton: Chapman and Hall/CRC. ▪ Loh, W.-Y. (2014). Fifty Years of Classification and Regression Trees. International Statistical Review 82(3), 329-348.

3 (morning)	**Machine Learning: Supervised learning methodology**
	Suggested reading: ▪ Chapter 2 and 5 in James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). An Introduction to Statistical Learning. New York, NY: Springer. ▪ Chapter 11 in Kuhn, M. and Johnson, K. (2013). Applied Predictive Modeling. New York, NY: Springer.
3 (afternoon)	**Data quality for gathered data types (like Twitter, Sensor etc.) and an Overview of Sampling from Twitter (Guest speaker: Dr. Trent D. Buskirk)**
	Suggested reading: ▪ Xu, H., Zhang, N., & Zhou, L. (2020). Validity concerns in research using organic data. Journal of Management, 46(7), 1257-1274. https://journals.sagepub.com/doi/pdf/10.1177/0149206319862027 ▪ Buskirk, T.D., Blakely, B.P., Eck, A. et al. Sweet tweets! Evaluating a new approach for probability-based sampling of Twitter. EPJ Data Sci. 11, 9 (2022). https://doi.org/10.1140/epjds/s13688-022-00321-1 ▪ Hino, A., & Fahey, R. A. (2019). Representing the Twittersphere: Archiving a representative sample of Twitter data under resource constraints. International journal of information management, 48, 175-184. https://www.sciencedirect.com/science/article/pii/S0268401218306005 ▪ Sen, Indira, Fabian Floeck, Katrin Weller, Bernd Weiss, and Claudia Wagner. "TED-On: A Total Error Framework for Digital Traces of Human Behavior on Online Platforms." arXiv preprint arXiv:1907.08228 (2019). https://arxiv.org/pdf/1907.08228.pdf
4	**Machine Learning: Boosting and Random Forests as general-purpose prediction techniques, Outlook on further methods, Application of supervised learning in the social sciences**
	Suggested reading: ▪ Chapter 10 and 15 in Hastie, T., Tibshirani, R., and Friedman, J. (2009) ▪ Chapter 7 and 11 in Foster, I., Ghani, R., Jarmin, R.S., Kreuter, F., & Lane, J. (2020)
5	**Text mining and Topic modelling, Clustering, Interpretable ML**
	Suggested reading: ▪ Chapter 7 and 8 in Foster, I., Ghani, R., Jarmin, R.S., Kreuter, F., & Lane, J. (2020) ▪ Grimmer, Justin, Margaret E. Roberts, and Brandon M. Stewart. (2021). Text as Data: A New Framework for Machine Learning and the Social Sciences. Princeton University Press. ▪ Silge, Julia and David Robinson (2017). Text mining with R: A tidy approach. O'Reilly ▪ Molnar, C. (2022). Interpretable Machine Learning: A guide for Making Black Box Models Explainable (2nd ed.). Independently published. https://leanpub.com/interpretable-machine-learning

Additional Recommended Literature:

- Hastie, T., Tibshirani, R., and Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. New York, NY: Springer.
- Foster, I., Ghani, R., Jarmin, R.S., Kreuter, F., & Lane, J. (2020). Big Data and Social Science: Data Science Methods and Tools for Research and Practice (2nd ed.). Chapman and Hall/CRC. https://doi.org/10.1201/9780429324383
- Baumer, B.S., Kaplan, D.T., & Horton, N.J. (2021). Modern Data Science with R (2nd ed.). Chapman and Hall/CRC. https://doi.org/10.1201/9780429200717

- Mehrabi, N., Morstatter, F., Saxena, N. Lerman, K., and Galsytan, A. (2021). A Survey on Bias and Fairness in Machine Learning. ACM Comput. Surv. 54, 6. https://doi.org/10.1145/3457607