SURV727, SurvMeth727
Fundamentals of Computing and Data Display
Mo 12:30pm – 3:00pm, 2208 LEF (UMD), 300 Perry (U-M)
Fall 2019

Christoph Kern
c.kern@uni-mannheim.de
Office: 1218 LeFrak Hall (UMD)
Office Hours: By appointment

Course Description: Empirical social scientists are often confronted with a variety of data sources and formats that extend beyond structured and handleable survey data. With the emergence of Big Data, especially data from web sources play an increasingly important role in scientific research. However, the potential of new data sources comes with the need for comprehensive computational skills in order to deal with loads of potentially unstructured information. Against this background, the first part of this course provides an introduction to web scraping and APIs for gathering data from the web and then discusses how to store and manage (big) data from diverse sources efficiently. The second part of the course demonstrates techniques for exploring and finding patterns in (non-standard) data, with a focus on data visualization. Tools for reproducible research will be introduced to facilitate transparent and collaborative programming. The course focuses on R as the primary computing environment, with excursus into SQL and Big Data processing tools.

Prerequisites: Some basic experience with programming in R or Python is helpful, but not strictly necessary. Students without any R knowledge are encouraged to work through one or more R tutorials prior or during the first weeks of the course. Some resources can be found here:

```
https://www.rstudio.com/online-learning/#R
https://rstudio.cloud/learn/primers
http://www.statmethods.net/
https://swirlstats.com/
```

Course Grades: R exercises, web data project (mid-term presentation, final presentation, term paper).

Grade Distribution: Each exercise, presentation and paper will be given a grade between 0 and 100. A missing submission will be scored as zero. A submitted term paper is a precondition for passing the course. Final grade: 35% R exercises (averaged, without lowest score), 10% mid-term presentation, 15% final presentation, 40% term paper.

Letter Grade Distribution: A+ [98–100], A [93–97], A– [90–92], B+ [88–89], B [83–87], B– [80–82], C+ [78–79], C [73–77], C– [70–72], D+ [68–69], D [63–67], D– [60–62], F < 60.

Course Objectives: At the completion of this course, students will have the compu-

tational skills to gather and process data from various (web) sources. Students will also learn how to deal with vast amounts of data and how to extract information from unstructured data through exploratory data analysis and visualizations. The course also includes how to write reproducible code and reports.

Recommended Textbooks:

Baumer, B. S., Kaplan, D. T., and Horton, N. J. (2017). *Modern Data Science with R.* Boca Raton, FL: Chapman & Hall/CRC Press.

Foster, I., Ghani, R., Jarmin, R. S., Kreuter, F., and Lane, J. (Eds.). (2017). *Big Data and Social Science: A Practical Guide to Methods and Tools.* Boca Raton, FL: CRC Press Taylor & Francis Group.

Wickham, H. and Grolemund, G. (2017). *R for Data Science.* O'Reilly.

Wickham, H. (2015). *Advanced R.* Boca Raton, FL: CRC Press Taylor & Francis Group.

Course Policies & Requirements: Students are required to bring a laptop with R (`www.R-project.org`) and RStudio (`www.rstudio.com`) installed for the in-class labs. Students are asked to attend class on time and remain through the entire class. Regular attendance is highly recommended as the graded exercises build heavily on the in-class lab sessions. Students are required to inform the instructor about absences that may conflict with submitting graded course work in a timely manner. More than three absences may cause deduction of points from the course score. Further information on attendance policies can be found at the University of Maryland, Office of Faculty Affairs website. `https://faculty.umd.edu/teach/#attend`. University of Michigan: `http://www.provost.umich.edu/calendar/`.

Academic Honesty Policy: Clear definitions of the forms of academic misconduct, including cheating and plagiarism, as well as information about disciplinary sanctions for academic misconduct, may be found at the University of Maryland, Office of the President's website. `http://www.president.umd.edu/policies/docs/III-100A.pdf`. University of Michigan: `http://www.rackham.umich.edu/policies/academic-policies/section11`.

Disability Accommodation: In order to receive services you must contact the Accessibility and Disability Service (ADS) office to register in person for services. Please call the office to set up an appointment to register with an ADS counselor. Contact the ADS office: `https://www.counseling.umd.edu/ads/`. University of Michigan: `https://ssd.umich.edu/`.

Student Health and Medical Emergency: For mental and physical health resources, visit `http://www.health.umd.edu/`. University of Michigan: `https://www.uhs.umich.edu/`.

Course Outline (Tentative):

Date	Content
09/09/2019	Introduction to git and GitHub
09/16/2019	Web scraping, html, xml, json, APIs, regular expressions • Readings: BD 2 • R exercise 1 assigned (due 11:59 p.m. on 09/29/2019)
09/23/2019	R style guide, data structures in R, functional programming • Readings: AR 2, 10
09/30/2019	Data Wrangling: split-apply-combine, dplyr, tidyr APIs I: Google Trends, censusapi • Readings: RDS 5, 10, 12 • R exercise 2 assigned (due 11:59 p.m. on 10/13/2019)
10/07/2019	Collecting Twitter data APIs II: rtweet • Readings: BD 10
10/14/2019	Databases, SQL, bigrquery • Readings: BD 4, MDS 12 • R exercise 3 assigned (due 11:59 p.m. on 10/27/2019)
10/21/2019	Big data processing, data.table, doParallel • Readings: BD 5, MDS 17
10/28/2019	Mid-term presentations
11/04/2019	Data exploration, Clustering, PCA • Readings: BD 6.5.1 • R exercise 4 assigned (due 11:59 p.m. on 11/10/2019)
11/11/2019	Data display with ggplot2 • Readings: BD 9, GGP 1, 2 • R exercise 5 assigned (due 11:59 p.m. on 11/17/2019)
11/18/2019	Interactive graphs, shiny, plotly, ggvis • Readings: MDS 11 • R exercise 6 assigned (due 11:59 p.m. on 12/01/2019)
11/25/2019	Communicate with RMarkdown • Readings: RDS 27
12/02/2019	Analysis and programming tools, purrr, broom, simulations • Readings: RDS 25
12/09/2019	Final presentations