Fall 2023
SURV727, SurvMeth727
Fundamentals of Computing and Data Display
Wed 1:30pm – 4:00pm
LeFrak 1208
Zoom Link: `https://umich.zoom.us/j/99175049270`

Brian Kim
kimbrian@umd.edu
Office: LeFrak 1218S
Office Hours: By appointment only. Tuesday 2-3pm, Wednesdays 12-1pm, or when available

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.

Learning Outcomes: After taking this course, student will be able to:

- Use Git and GitHub to collaborate on projects and disseminate reproducible research.
- Implement best practices in R programming and version control.
- Obtain web data from a variety of sources using R.
- Produce effective, presentation-quality visualizations

Class Meetings: This course uses a flipped classroom design. In this course, you are responsible for watching video recorded lectures and going through the readings, and then attending class meetings where students have the chance to discuss the materials from a unit with the instructor. In general, the class time will be used for in-class activities, answering questions, and walking through examples. In preparation for class meetings, students are expected to watch the lecture videos and go over the readings before the start of the meeting. In addition, **students must answer any prompts and post questions or comments about the materials covered in the videos and readings of the week to the instructor in the forum before the meetings**. If you do not have questions, you may respond to another student, or comment about possible topics of interest to you, such as different applications of the tools learned in class. The deadline for posting is three hours before the online meeting. **Posting weekly will account for 10% of your final grade.**

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://rstudio.cloud/learn/primers`
`http://www.statmethods.net/`
`https://swirlstats.com/`

Course Grades: Forum Posts, R exercises, Midterm Exam, Web Data Project (project proposal, 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 final paper is a precondition for passing the course.
The distribution for the final grade will be:

- 10% Forum posts
- 30% Assignments
- 20% Online Midterm Exam
- 40% Web Data Project (5% final paper proposal, 10% final presentation, 25% 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 computational 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.

Course Textbooks & Literature:

Amaya, A., Bach, R., Keusch, F., and Kreuter, F. (2019). New Data Sources in Social Science Research: Things to Know Before Working With Reddit Data. *Social Science Computer Review*, online first.
`https://doi.org/10.1177/0894439319893305`

Baumer, B. S., Kaplan, D. T., and Horton, N. J. (2021). *Modern Data Science with R, Second Edition.* Boca Raton, FL: Chapman & Hall/CRC Press. [**MDS**]
`https://mdsr-book.github.io/mdsr2e/`

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. [**BD**]
https://textbook.coleridgeinitiative.org/

Silge, J., and Robinson, D. (2017). *Text Mining with R: A Tidy Approach.* O'Reilly. [**TMR**]
https://www.tidytextmining.com/

Stephens-Davidowitz, S., and Varian, H. (2015). *A Hands-on Guide to Google Data.*
http://people.ischool.berkeley.edu/~hal/Papers/2015/primer.pdf

Wickham, H. and Grolemund, G. (2017). *R for Data Science.* O'Reilly. [**RDS**]
https://r4ds.had.co.nz/

Wickham, H. (2015). *Advanced R.* 1st edition. Boca Raton, FL: CRC Press Taylor & Francis Group. [**AR**]
http://adv-r.had.co.nz/

Additional Literature:
These books cover some of the topics of this course in much more detail, but are highly recommended for those who want to learn more.

Luraschi, J, Kuo, K., and Ruiz, E. (2019). *Mastering Spark with R. The Complete Guide to Large-Scale Analysis and Modeling.* O'Reilly.
https://therinspark.com/

Salganik, M. J. (2017). *Bit By Bit: Social Research in the Digital Age.* Princeton University Press.
https://www.bitbybitbook.com/en/1st-ed/preface/

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 crucial for success in the course because this is a flipped classroom format and much of the in-class time will involve working on assignments. Students are required to inform the instructor about absences that may conflict with submitting graded course work in a timely manner. 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):

Week	Date	Content
Week 1	08/30	Introduction and presentation of the course, Git and GitHub
Week 2	09/06	R basics, functional programming, R Markdown - Readings: Readings: AR 2 & 10 - Assignment 1 (due 11:59 p.m. on 09/19)
Week 3	09/13	Communicate with R Markdown and Quarto - Readings: RDS 27
Week 4	09/20	Data wrangling: split-apply-combine, dplyr, tidyr - Readings: RDS 5, 10, 12; Stephens-Davidowitz & Varian (2015) - Assignment 2 (due 11:59 p.m. on 10/03)
Week 5	9/27	APIs: gtrendsR, censusapi - Readings: BD 10, Amaya et al. (2020)
Week 6	10/04	Web scraping, html, xml, json, more APIs, regular expressions - Readings: BD 2 - Assignment 3 (due 11:59 p.m. on 10/17)
Week 7	10/11	Text mining - Readings: TMR 6, BD 8, MDS 19
Week 8	10/18	Online Midterm Exam (No Class)
Week 9	10/25	Databases, SQL, bigrquery - Readings: BD 4, MDS 15 - Assignment 4 (due 11:59 p.m. on 11/7)
Week 10	11/1	Data display with ggplot2 - Readings: RDS 3, 28; MDS 14
Week 11	11/8	Interactive graphs, shiny, plotly, ggvis - Readings: RDS 3, 28; MDS 14 - R exercise 5 (due 11:59 p.m. on 11/28)
Week 12	11/15	Data exploration, Clustering, PCA - Readings: BD 7.6.1
	11/22	Thanksgiving
Week 13	11/29	Big data processing, data.table, doParallel, Virtual machines - Readings: BD 5, MDS 21
Week 14	12/06	Final presentations
	12/13	Final project due at 11:59 p.m.

- AR 2 "Data Structures"
- AR 10 "Functional Programming"
- RDS 27 "R Markdown"
- RDS 5 "Data transformation"
- RDS 10 "Tibbles"
- RDS 12 "Tidy Data"
- BD 10 "Data Quality and Inference Errors"
- BD 2 "Working with Web Data and APIs"
- TMR 6 "Topic Modeling"
- BD 8 "Text Analysis"
- MDS 19 "Text as Data"
- BD 4 "Databases"
- MDS 15 "Database Querying Using SQL"
- BD 5 "Scaling up through Parallel and Distributed Computing"
- MDS 21 "Epilogue: Towards Big Data"
- BD 7.6.1 "Machine Learning — Unsupervised learning methods"
- RDS 3 "Data Visualisation"
- RDS 28 "Graphics for Communication"
- MDS 14 "Interactive Data Graphics"