Social Data Science
Rutgers University

Syllabus

Dr. Thomas Davidson

Fall 2021

CONTACT AND LOGISTICS

E-mail: `thomas.davidson@rutgers.edu` or Canvas message.

Website: `https://github.com/t-davidson/social-data-science-fall-2021` and Canvas.

Class meetings: MW 3-4:20 p.m, Loree Classroom Building - Room 020.

Office hours: W 4:30-5:30 p.m, Davison Hall - Room 109 or by appointment.

COURSE DESCRIPTION

This course introduces students to the growing field of computational social science. Students will learn to collect and critically analyze social data using a range of techniques including natural language processing, machine learning, and agent-based modeling. We will discuss how these techniques are used by social scientists and consider the ethical implications of big data and artificial intelligence. Students will complete homework assignments involving coding in the R programming language to analyze several different datasets and will complete a group project to create a web-based application for data analysis and visualization.

PREREQUISITES

Data 101 or equivalent. Enrolled students *must* have experience writing basic programs in a general purpose programming language, e.g. R, Python, Java, C.

We will review the fundamentals for programming and data science in R in weeks 1-3.

ASSESSMENT

- 10% Class participation
- 60% Homework assignments (3 x 20%)
- 30% Group project (R Shiny app and write-up)

READINGS

Most of the readings will consist of chapters from the textbooks listed below. These readings are intended to build familiarity with key concepts and programming skills. Some weeks there will be an additional reading to highlight how data science techniques are used in empirical social scientific research.

Textbooks

* *indicates a required text. All required texts and useful references are available for free online on the listed websites.*

- *Matthew Salganik. 2017. *Bit by Bit*. Princeton University Press. https://www.bitbybitbook.com/en/1st-ed/preface/
- *Wickham, Hadley, and Garrett Grolemund. 2016. *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*. (*R4DS*). O'Reilly Media, Inc. https://r4ds.had.co.nz/
- *Silge, Julia, and David Robinson. 2017. *Text Mining with R: A Tidy Approach.* O'Reilly Media. https://www.tidytextmining.com/dtm.html.
- Healy, Kieran. 2018. *Data Visualization: A Practical Introduction*. Princeton University Press. https://socviz.co/

RESOURCES

The course will be organized using two different tools, Github and Canvas. Canvas will be used for class communication, short quizzes, and for scheduling. Github Classroom will be used for the submission of assignments.

COURSE POLICIES

The Rutgers Sociology Department strives to create an environment that supports and affirms diversity in all manifestations, including race, ethnicity, gender, sexual orientation, religion, age, social class, disability status, region/country of origin, and political orientation. This class will be a space for tolerance, respect, and mutual dialogue. Students must abide by the Code of Student Conduct at all times, including during lectures and in participation online.

All students must abide by the university's Academic Integrity Policy. Violations of academic integrity will result in disciplinary action.

In accordance with University policy, if you have a documented disability and require accommodations to obtain equal access in this course, please contact me during the first week of classes. Students with disabilities must be registered with the Office of Student Disability Services and must provide verification of their eligibility for such accommodations.

I will also be making additional accommodations due to the COVID-19 pandemic. If you or your family are affected in any way that impedes your ability to participate in this course, please contact me as soon as you can so that we can make necessary arrangements.

COURSE OUTLINE

This outline is tentative and subject to change.

Week 1, 9/1 (Wednesday only)

Introduction to social data science

Readings

- *Bit by Bit*, C1
- *R4DS*: C1 & 27 [Note: Chapter numbers correspond to the online book; physical book numbers are different]

Week 2, 9/8 (Wednesday only)

Data structures in R

Readings

- *R4DS*: C2,4, skim 20.

Week 3, 9/13 & 9/15

Programming fundamentals

Readings

- Monday, *R4DS*: C17-19, 21.
- Wednesday, *R4DS*: C5, 9, 10, 13.

Assignment 1 released: Using R for Data Science.

Week 4, 9/20 & 9/22

Data Collection I: Collecting data using Application Programming Interfaces

Readings

- Monday, *Bit by Bit*, C2
- Wednesday, *R4DS*: C3

Week 5, 9/27 & 9/29

Data Collection II: Scraping data from the web

Readings

- Monday, *Bit by Bit*, C6
- Wednesday, *R4DS*: C14, 16

Recommended

- Fiesler, Casey, Nate Beard, and Brian C Keegan. 2020. "No Robots, Spiders, or Scrapers: Legal and Ethical Regulation of Data Collection Methods in Social Media Terms of Service." In *Proceedings of the Fourteenth International AAAI Conference on Web and Social Media*, 187–96.

Week 6, 10/4 & 10/6

Data Collection III: Online experiments and surveys

Assignment 2: Collecting and storing data released.

Readings

- R Shiny tutorial: https://shiny.rstudio.com/tutorial/
- *Bit by Bit*, C3-5

Week 7, 10/11 & 10/12

Natural Language Processing I: The vector-space model

Readings

- *Text Mining with R*, C1 & 3

Recommended

- Evans, James, and Pedro Aceves. 2016. "Machine Translation: Mining Text for Social Theory." *Annual Review of Sociology* 42 (1): 21–50. https://doi.org/10.1146/annurev-soc-081715-074206.

Week 8, 10/18 & 10/20

Natural Language Processing II: Word embeddings

Readings

- *Text Mining with R*: C5.
- Hvitfeldt, Emil and Julia Silge. 2020 *Supervised Machine Learning for Text Analysis in R.* Chapter 5: https://smltar.com/embeddings.html.

Recommended

- Kozlowski, Austin, Matt Taddy, and James Evans. 2019. "The Geometry of Culture: Analyzing the Meanings of Class through Word Embeddings." *American Sociological Review*, September, 000312241987713. https://doi.org/10.1177/0003122419877135.

Week 9, 10/25 & 10/27

Natural Language Processing III: Topic models

Assignment 3: Natural language processing released.

Readings

- *Text Mining with R*: C6.
- Mohr, John, and Petko Bogdanov. 2013. "Introduction—Topic Models: What They Are and Why They Matter." *Poetics* 41 (6): 545–69. https://doi.org/10.1016/j.poetic.2013.10.001.

Recommended

- Roberts, Margaret, Brandon M. Stewart, Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Shana Kushner Gadarian, Bethany Albertson, and David Rand. 2014. "Structural Topic Models for Open-Ended Survey Responses: Structural Topic Models for Survey Responses." *American Journal of Political Science* 58 (4): 1064–82. https://doi.org/10.1111/ajps.12103.

Week 10, 11/1 & 11/3

Machine Learning I: Prediction and explanation

Readings

- Molina, Mario, and Filiz Garip. 2019. "Machine Learning for Sociology." *Annual Review of Sociology* 45: 27–45.

Week 11, 11/8 & 11/10

Machine learning II: Text classification

Readings

- Hanna, Alex. 2013. "Computer-Aided Content Analysis of Digitally Enabled Movements." *Mobilization: An International Quarterly* 18 (4): 367–388.

Recommended

- Barberá, Pablo, Amber E. Boydstun, Suzanna Linn, Ryan McMahon, and Jonathan Nagler. 2020. "Automated Text Classification of News Articles: A Practical Guide." *Political Analysis*, June, 1–24. https://doi.org/10.1017/pan.2020.8.

Week 12, 11/15 & 11/17

Machine learning III: Challenges

Readings

- Salganik, Matthew, Ian Lundberg, Alexander Kindel, et al. 2020. "Measuring the Predictability of Life Outcomes with a Scientific Mass Collaboration." *Proceedings of the National Academy of Sciences*.
- Buolamwini, Joy, and Timnit Gebru. 2018. "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification." In *Proceedings of Machine Learning Research*, 81:1–15.

Week 13, 11/22 (No class Wednesday for Thanksgiving)

Machine learning IV: Image classification

Readings

- Torres, Michelle, and Francisco Cantú. 2021. "Learning to See: Convolutional Neural Networks for the Analysis of Social Science Data." *Political Analysis*, April, 1–19. https://doi.org/10.1017/pan.2021.9.
- Gebru, Timnit, Jonathan Krause, Yilun Wang, Duyun Chen, Jia Deng, Erez Lieberman Aiden, and Li Fei-Fei. 2017. "Using Deep Learning and Google Street View to Estimate the Demographic Makeup of Neighborhoods across the United States." *Proceedings of the National Academy of Sciences* 114 (50): 13108–13. https://doi.org/10.1073/pnas.1700035114.

Week 14, 11/29 & 11/1

Simulation and agent-based models

Readings

- Macy, Michael, and Robert Willer. 2002. "From Factors to Factors: Computational Sociology and Agent-Based Modeling." *Annual Review of Sociology* 28 (1): 143–66. https://doi.org/10.1146/annurev.soc.28.110601.141117.
- https://cran.r-project.org/web/packages/shinySIR/vignettes/Vignette.html

Week 15, 12/6 & 12/7

Presentations

Final projects due 12/16 at 5pm