DS 2001: Social Science Practicum

Stefan McCabe

Section 01

Network Science Institute

mccabe.s@northeastern.edu

Office Hours: Fridays 3:00–5:00, 1027I 177 Huntington

Carolina Mattsson

Section 02

Network Science Institute

mattsson.c@northeastern.edu

Office Hours: Thursdays 9:30-11:30am, or by appointment

Course Description

Meetings

Section 01

Wednesdays, 2:50–4:30pm, Hastings Suite 206

Section 02

Wednesdays, 9:50–11:30am, West Village H 212

Websites

Piazza: https://piazza.com/northeastern/fall2019/ds2000
Course website: https://course.ccs.neu.edu/ds2000/index.html
Practicum website: https://sdmccabe.github.io/ds2001/

Teaching Assistants

Section 01

Romil Rathi (rathi.r@husky.neu.edu)
Mondays 3:00–5:00 (WVH 3rd floor), Thursdays 5:00-7:00 (KA 308)

Section 02

Kevin Liang (liang.ke@husky.neu.edu)
Mondays 6:00–9:00 (Ryder 202)

Overview

Large-scale data and computationally complex methods for understanding human behavior are accessible like never before with the emergence of vast archives of passive data collection, online experimentation, and innovative uses of simulation. Digital traces of our daily lives our increasingly recorded, aggregated, analyzed, and used to shape our future experience. These data and methods offer the potential for rich insights into society, while simultaneously introducing new ethical and infrastructural challenges. In this practicum we will (1) practice the skills you learn in the DS2000 lecture using applied examples drawn from the social sciences, (2) read about how data science is impacting society, and (3) develop an intuition for computational social science. The practicum will meet once a week for 1 hour 40 minutes. Class time will be a combination of discussion and reactions to short readings, and hands-on tutorials that practice the skills you learned in lecture. Your grade will be based on your reading reflections, programming exercises, a project proposal, and a final project and presentation.

Programming

Students who have access to a laptop should bring it to class every day. This workshop will be taught in the open source programming language Python 3. The room scheduled for practicum has desktops with Python 3 installed on them for those who do not have access to a laptop. We will begin by writing Python code in the open source text editor Atom. This course will use the Anaconda distribution of Python, which is available on Windows, macOS, and Linux, and includes many of the libraries we will be using in the course. Download the Python 3.7 Anaconda distribution from https://www.anaconda.com/distribution/. Part way through the course we will introduce the programming environment Jupyter, which is included in the Anaconda distribution.

Readings

All of the readings are available online and the links are provided in the syllabus. If you have trouble accessing any of the readings, please let us know as soon as possible.

Course Schedule

September 4: Hello, world!

September 11: Computing

- Kleiman, Kathy. 2016. *Great Unsung Women of Computing, Part 1: The Computers*. San Francisco: Women Make Movies. https://link.ezproxy.neu.edu/login?url=https://northeastern.kanopy.com/node/2288448.

September 18: Big Data

- Quince, Anabelle. 2016. *The Dark Side of Census Collections*. Podcast. Rear Vision. Australian Broadcasting Corporation. https://www.abc.net.au/radionational/programs/rearvision/the-dark-side-of-census-collections/7860908.
- Foucault Welles, Brooke. 2014. "On Minorities and Outliers: The Case for Making Big Data Small." *Big Data & Society* 1 (1): 1–2. https://doi.org/10.1177/2053951714540613.

September 25: Big Digital Data

- Blumenstock, Joshua, Gabriel Cadamuro, and Robert On. 2015. "Predicting Poverty and Wealth from Mobile Phone Metadata." *Science* 350 (6264): 1073–76. https://doi.org/10.1126/science.aac4420.
- Lazer, David, Ryan Kennedy, Gary King, and Alessandro Vespignani. 2014. "The Parable of Google Flu: Traps in Big Data Analysis." *Science* 343(6176): 1203–5. https://science.sciencemag.org/content/343/6176/1203

October 2: Ranking & Hiring

- Hakes, Jahn K., and Raymond D. Sauer. 2006. "An Economic Evaluation of the Moneyball Hypothesis." *Journal of Economic Perspectives* 20(3): 173–86. https://www.aeaweb.org/articles?id=10.1257/jep.20.3.173
- Jeffrey Dastin. 2018. "Amazon Scraps Secret AI Recruiting Tool That Showed Bias against Women." *Reuters*. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G

October 9: Genetic Databases

- Gregory Rodriguez. 2014. "How Genealogy Became Almost as Popular as Porn." *Time*. https://time.com/133811/how-genealogy-became-almost-as-popular-as-porn/
- "DNA Testing Identifies Actual Perpetrator in 1996 Idaho Falls Rape and Murder, Confirming Christopher Tapp's Innocence." 2019. *Innocence Project*. https://www.innocenceproject.org/christopher-tapp-exoneration/
- Human Rights Watch. 2017. "China: Minority Region Collects DNA from Millions." New York: Human Rights Watch. https://www.hrw.org/news/2017/12/13/china-minority-region-collects-dna-millions

Optional:

- Erlich, Yaniv, Tal Shor, Itsik Pe'er, and Shai Carmi. 2018. "Identity Inference of Genomic Data Using Long-Range Familial Searches." *Science* 362(6415): 690–94. https://science.sciencemag.org/content/362/6415/690

October 16: Predictive Policing

- Wolpert, Stuart. 2015. "Predictive Policing Substantially Reduces Crime in Los Angeles during Months-Long Test." *UCLA Newsroom*. http://newsroom.ucla.edu/releases/predictive-policing-substantially-reduces-crime-in-los-angeles-during-months-long-test
- Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. "Machine Bias." *ProPublica*. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Optional:

- Mohler, G. O. et al. 2015. "Randomized Controlled Field Trials of Predictive Policing." *Journal of the American Statistical Association* 110(512): 1399–1411. https://www.tandfonline.com/doi/full/10.1080/01621459.2015.1077710

October 23: Asking Questions

Project teams due

- King, Gary, Robert O. Keohane, and Sidney Verba. 1994. "The *Science* in Social Science" In *Designing Social Inquiry: Scientific Inference in Qualitative Research*, 3–33. Princeton, NJ: Princeton University Press. https://www.jstor.org/stable/j.ctt7sfxj.4

October 30: Search

Project proposals due

- Jensen, Robert. 2007. "The Digital Provide: Information (Technology), Market Performance, and Welfare in the South Indian Fisheries Sector." *The Quarterly Journal of Economics* 122 (3): 879–924. https://doi.org/10.1162/qjec.122.3.879. (**read Introduction and Figs 2–4**)
- Noble, Safiya Umoja. 2018. *Algorithms of Oppression: How Search Engines Reinforce Racism*. NYU Press. **15-29**. https://www-jstor-org.ezproxy.neu.edu/stable/j.ctt1pwt9w5.5

November 6: Replication

- Hall, Andrew B., Connor Huff, and Shiro Kuriwaki. 2019. "Wealth, Slaveownership, and Fighting for the Confederacy: An Empirical Study of the American Civil War." *American Political Science Review* 113 (3): 658–73. https://doi.org/10.1017/S0003055419000170.

November 13: Segregation

- Schelling, Thomas C. 2006. "Sorting and Mixing" In *Micromotives and Macrobehavior*, 137-166. New ed. New York: Norton. https://snap.stanford.edu/class/cs224w-readings/schelling78segregation.pdf. (read up to 155, skim the rest)
- Tripodi, Francesca. 2018. "Searching for Alternative Facts: Analyzing Scriptural Inference in Conservative News Practices." Data & Society Research Institute. https://datasociety.net/output/searching-for-alternative-facts/. (read 27-34 only)

Optional:

- Robertson, Ronald E., Shan Jiang, Kenneth Joseph, Lisa Friedland, David Lazer, and Christo Wilson. 2018. "Auditing Partisan Audience Bias within Google Search." *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW: Article 148. https://doi.org/10.1145/3274417.

November 20: Recommender systems

- Tufekci, Zeynep. 2018. "YouTube, the Great Radicalizer." *The New York Times*, 2018. https://www.nytimes.com/2018/03/10/opinion/sunday/youtube-politics-radical.html.
- Ribeiro, Manoel Horta, Raphael Ottoni, Robert West, Virgílio A. F. Almeida, and Wagner Meira. 2019. "Auditing Radicalization Pathways on YouTube." arXiv:1908.08313 [cs.CY]. http://arxiv.org/abs/1908.08313.

Optional:

- Fisher, Max, and Amanda Taub. 2019. "How YouTube Radicalized Brazil." *The New York Times*, 2019. https://www.nytimes.com/2019/08/11/world/americas/youtube-brazil.html.

November 27: Thanksgiving (no class)

December 4: Presentations

Course Requirements

Structure

This is a hands-on course that will introduce students to the practical application of programming to questions in social science. It will consist of in-class programming exercises, assigned readings, and class discussions. Taught in the programming language Python and the development environment Jupyter, in-class exercises are aimed at getting students processing and analyzing data. The assigned readings explore real-world applications of data science with a particular relevance to social

science, both in their impact on society and on how we study society. It is important that you complete the readings before each class and come prepared to discuss the material. Discussions—both on Piazza and in-class—will be a space to synthesize and critique both materials and methods. It is important that we respect one another's thoughts, give everybody the space to talk, and address our comments at the ideas and not the person.

Grading

The practicum grade will be based on:

- Weekly coding exercises (30%)
- Six reading reflections (15%)
- Class + Piazza discussion (15%)
- Final Project + Presentation (40%)

Attendance and Participation

In each week we will learn skills and develop knowledge that build on what was taught in prior weeks, so it is important to attend every class. While we will not keep attendance, each class period will include in-class coding exercises and discussion of the week's readings. If you miss a class you will lose the opportunity to get those points. If you know you are going to miss a class you should notify us at least four days in advance so we can arrange a way to make up the material. If given less notice, there may not be a way to earn those points.

Programming exercises

Learning Python is like learning a foreign language—the best way to learn it is to use it all the time. Weekly coding exercises are in-class activities that should be submitted at the end of the practicum. See the first Practicum Handout for details on how to submit these practicum assignments.

Reading reflections

Learning how to use data science to answer important questions about society is less straightforward. We will assign short readings each week to get you familiar with real-world applications of data science that have impacted society, impacted how we study society, or both. In class we will have you answer brief questions about the readings. To facilitate discussion, you must post at least 6 reading reflections to Piazza over the course of the semester. These reflections should be a 250-500 words and should be posted by midnight the day before class (11:59pm on Tuesday) in the week that reading is assigned. You are welcome to post your reflections as a new thread or as respectful commentary on someone else's reflection. Note that readings are assigned for 8 weeks of the class, so choose your 6 weeks of reading reflection wisely!

Project

The final project will allow you to creatively combine the techniques you learned in the course to explore a question related to the humanities or social sciences. Through this project you should show that you understand (a) what types of questions are interesting or important to social scientists, (b) what types of questions can be best answered using computational or digital techniques, (c) what types of techniques and evidence are appropriate to best answer your question, and (d) that you can think about how to present your findings and analysis in a reproducible way and in a way that supports, and persuades others of, your (preliminary) conclusion. The final project consists of a project proposal, Jupyter notebook, and final presentation.

Project teams – due Wednesday, October 23rd

A short document listing the members of the team (two or three people) and the planned division of labor among you. The best teams will include people from different disciplinary backgrounds so you can leverage each other's specialized knowledge.

Project proposal – due Wednesday, October 30th

A 1-page document detailing a preliminary plan for your final project. You should include the following in your project proposal:

1. Identify a general question related to the social sciences that you plan to address in your final project. You should outline why this is an interesting or important question and describe why computational methods are necessary and/or helpful in exploring this question. If possible, explain how others have answered/attempted to answer this question using different methods.

2. Identify the data or collection of material you will use to explore this question, and briefly describe why the data/material is appropriate. Additionally describe how you will collect the data/material and whether or not it will need to be cleaned prior to analysis. If possible, import your data and provide a glimpse of its format.

3. Describe the techniques you expect, or would like, to use to analyze the data/material and explore your question. Why these techniques and not others? What kind of evidence will these techniques produce, and how will this help you answer the question and persuade others of your answer? If you already have some preliminary analyses, include these as well.

4. Briefly discuss any data visualization or interpretive techniques you will, or would like to, use to present your findings and convince others of your interpretation.

Project notebook – due Wednesday, November 27th

A Jupyter notebook detailing your final project. Project notebooks combine written analysis with data analysis and visualization in Python. You should expound on the points in your project proposal, present your results, and synthesize your findings. Successful projects will likely include the following:

1. A general question related to the social sciences that you are addressing in your final project. You should motivate your project by describing why this is an interesting or important question and why computational methods are useful in exploring this question. If appropriate for your project, you could mention how others have approached this question using different methods.
2. A description of the data you use to explore your question and the reasons this data is useful for doing so. You should explain how you collected the data and any important choices you made in the process of cleaning and preparing the data. Please include an overview of the dataset itself including such things as what constitutes a record, the number of records, and a discussion of the relevant variables. If appropriate for your project, you could note any data or measurement issues to keep in mind.
3. Analysis of the data using computational techniques in Python and the results of this analysis. You should describe the techniques you are using and the results these techniques are producing in the context of the question you are exploring. Please include descriptive (summary) statistics for key variables. You will likely want to analyze relevant relationships between variables or differences across relevant subgroups. If appropriate for your project, you could identify outliers and discuss their significance or employ statistical or machine learning models (e.g., linear regression, random forests) and explain their outputs.
4. Interpretation that relates specific results from your analysis to the broader question you are exploring. You should discuss the takeaways from your analysis and how your findings contribute towards an understanding of your topic. Where visualizations aid in interpretation, please use effective annotations - axis labels, titles etc. If appropriate for your project, you could discuss the limitations of your work and what might be done in the future.

Project Presentation – due Wednesday, December 4th

An in-class presentation of your final project. Project presentations will be 10 minutes long. You should motivate your project, describe your analysis, present your results, and synthesize your findings. Please do not include code.

Please upload the slides accompanying your presentation before class so that I can have everyone's presentation already loaded onto the classroom computer.

Course Policies

Questions? Discussion Board, Office Hours, and Email

If you come across errors as you run code post them to the discussion board on Piazza (start a new thread for new errors). You may also post questions or comments about the readings or about your final project. We encourage everyone to answer each other's questions, as this is the best way to learn complicated material. Often many people will get the same error or will have similar questions, so check the discussion board for answers before posting your error or question. This is not the comments section on YouTube, so keep your comments respectful. Disrespect will absolutely not be tolerated. You are also encouraged to come to our office hours. Please feel free to email us at mccabe.s@northeastern.edu or mattsson.c@northeastern.edu if you would like to meet outside of scheduled office hours or need to inform us of a planned absence. Please include "(DS 2001)" at the beginning of your email's subject heading so that we know it's about class, or else we may not notice it.

We will respond to emails and Piazza posts within one business day; e.g., between the hours of 9am–5pm, Monday through Friday. We will typically respond to messages the same day, but if you send a message near the end of the day I will most likely respond the next morning, and if you email us or post questions on a Friday afternoon or a weekend, we may not respond until the following Monday.

Statement of Non-Discrimination

As the instructors of the course, we are committed to maintaining a positive learning environment based upon communication, mutual learning, and respect. Any suggestions as to how to further such a positive and open environment in this class will be appreciated and given serious consideration. The university does not discriminate on the basis of race, sex, age, disability, religion, sexual orientation, color, or national origin. If you are a person with a disability and anticipate needing any type of accommodation in order to participate in this class, please advise us and make appropriate arrangement with Disability Resource Center (617) 373-4428. If you need accommodation for any topic not listed here, please let us know.

Consulting Resources

We encourage you to take advantage of the Digital Scholarship Group at Northeastern. They offer a wealth of services—including digital data collections—and can offer advice on collecting and structuring digital data. They also offer a quiet space to work.

Note on Plagiarism

We encourage you to work together to help each other review the readings and to learn the coding. However, all written and coding work must be your own. We take academic honesty seriously, and you should too. The Northeastern University Policy on Academic Integrity can be found at: http://www.northeastern.edu/osccr/

academichonesty.html Since students in this course are often encouraged to work in teams, some specific remarks are in order:

It is not considered cheating if you:

- Work together on homework assignments, as long as you each work out and submit your own final answers
- Get help from professors, tutors, etc. on the homework assignments
- Work together on preparing for quizzes and exams

It is considered cheating if you:

- Submit work done by others (without your participation) as your own
- Copy work on quizzes and exams

Final thoughts

If you are unsure about anything related to the rules guiding this course, please ask!