Instructor: Matthew Simonson
Northeastern University, Spring 2020
simonson.m@northeastern.edu

DS 2001: Data Science Programming Practicum

Data Science for Social Scientists

Course Description

The goal of this course is to teach you how to use data science to conduct research in the social sciences. Throughout the course, you will be conducting an original research project to get a taste of what it is like to do data-driven social science research. The course will teach you how to read scientific articles efficiently and to evaluate their use of data and statistics. We will practice our Python programming skills on data sets similar to those in the articles, reinforcing coding and analysis skills that can be used in future courses, internships, and jobs. Finally, for every topic, we will examine the ethical dimensions of how social scientists obtain and use data and the positive and negative implications of their discoveries.

Structure:

This two-credit class (half a standard class) meets weekly for two hours. Most weeks students read 1–3 academic articles before class and submit a 400–600-word article analysis addressing the thesis, evidence, internal validity, external validity, and ethics for one of the studies. Students begin class by discussing the readings in groups, then as a class, followed by a lesson covering any new statistics or research design concepts that arise in the readings. Some weeks there will be a quiz on these concepts; other weeks practice exercises are provided. The second half of class is spent practicing coding skills covered earlier that week in this class's two-credit companion course, "DS 2000: Programming for Data Science" which students take concurrently. In addition to reinforcing Python programming skills, the exercises are meant to demonstrate how these skills can be used in social science and are usually inspired on the topics and techniques covered in readings. As the course progresses, emphasis shifts toward the students' original group research projects.

Week 1: What is Data Science? Why Learn It?

In our first meeting, we dive into data straightaway, conducting a class-wide survey experiment to measure how race and gender affect our reactions to #MeToo allegations. We analyze our responses in a spreadsheet, using Fisher's randomization inference to determine the statistical significance of our results. Along the way, we learn about p-values, vignette experiments, and our own implicit biases, while discovering why writing code can save a lot of time compared to doing statistics in Google Sheets or Excel. In the second half of class, we go over course logistics, install Python 3 and Atom code editor on our laptops, and write our first script in Python.

Week 2: The Promise and Peril of Cell Phone Data

Readings

Berger, Daniel, Shankar Kalyanaraman, and Sera Linardi. 2014. "Violence and Cell Phone Communication: Behavior and Prediction in Côte d'Ivoire." *Available at SSRN 2526336*. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2526336.

In our second meeting, we discuss how cell phone metadata can be used to predict outbreaks of violence in the aftermath of civil war, but why that's ethically a double-edged sword. We take a deep dive into vignette experiments, examining ongoing research into ethnic/religious bias and looking for clever ways to minimize experimenter demand effects. We also practice using **user input, print statements, variables, formatted strings, lists, conditional evaluation, and for loops** to create a basic caller ID program. Finally, we review basic statistics terminology and discuss the challenges of designing survey questionnaires and survey experiments.

Week 3: Matching Up Big Data with Small Data

Readings

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.

Eagle, Nathan, Alex (Sandy) Pentland, and David Lazer. 2009. "Inferring Friendship Network Structure by Using Mobile Phone Data." Proceedings of the National Academy of Sciences 106 (36): 15274–78. https://doi.org/10.1073/pnas.0900282106.

Lazer, David, Ryan Kennedy, Gary King, and Alessandro Vespignani. 2014. "The Parable of Google Flu: Traps in Big Data Analysis." Science 343 (6176): 1203–1205. https://doi.org/10.1126/science.1248506.

With all the hype about big data in the past five years, have surveys and other traditional forms of "small data" become obsolete? We examine three papers that demonstrate that when we match up these two types of data sources, we can learn new things that neither source could teach us on its own. Applications include estimating regional patterns of wealth in developing countries, identifying friendships among cell phone users, and forecasting the spread of the flu (or so we thought…). We also learn about the pitfalls of "big data hubris" and the difficulties in basing scientific discoveries on search engines and apps whose dynamic algorithms are constantly changing. All of these studies raise questions about privacy, but they also open the door to new opportunities to help people live happier, healthier lives. How do we get the balance right?

We get a taste of how this research is performed by searching for outliers in our own cell phone data, inferring a friendship network from call records, and identifying search terms that best predict pregnancy. To do so, we practice using **for/while loops, range, if/elif, and iterating through zipped and nested lists**. New statistics concepts covered include outliers, percentiles, regression, variable selection, confidence intervals, false positives, ROC curves, and out-of-sample prediction.

Week 4: Censuses: The Original Big Data

Readings

Fryer, Roland. 2007. "Guess Who's Been Coming to Dinner? Trends in Interracial Marriage over the 20th Century." Journal of Economic Perspectives 21 (2): 71–90. https://doi.org/10.1257/jep.21.2.71.

In-Class Quiz

True & False positives and ROC curves; p-values, confidence intervals and significance testing; internal/external validity, types of inference.

Which racial groups are most likely to intermarry? Simply counting the number of intermarriages may not give us the best answer, even if that count comes from as reliable a source as the U.S. Census. Is intermarriage between African Americans and Asian-American so rare because of racial preference/prejudice, or because they tend to live in different communities? To answer these questions, we learn about re-weighting groups by population. We also discuss the historical misuses of census data to prosecute draft-dodgers in WWI and intern Japanese citizens in WWII, as well as the risk of misuse today against immigrants and suspected terrorists. We consider who is most likely to be undercounted in a census and the political implications of including a citizenship question. Finally, we look at censuses in other countries and how breakdowns in data collection can themselves be a useful form of data for measuring state capacity. To practice **function syntax, variable scope, and indexing and slicing in lists,** we write a script that adjusts intermarriage rates to reflect each group's share of the population.

Week 5: Applying Data Science to Historical and Archival Data

Readings

Braun, Robert. 2016. "Religious Minorities and Resistance to Genocide: The Collective Rescue of Jews in the Netherlands during the Holocaust." American Political Science Review 110 (01): 127–47. https://doi.org/10.1017/S0003055415000544.

Pietryka, Matthew T., and Donald A. Debats. 2017. "It's Not Just What You Have, but Who You Know: Networks, Social Proximity to Elites, and Voting in State and Local Elections." American Political Science Review 111 (02): 360–78. https://doi.org/10.1017/S000305541600071X.

Can data science help us understand history? Do archives and historical records contain data we might be able to use Python programming to analyze? In our first article, we consider how using geolocated data from Nazi records of Jewish residences in the occupied Netherlands can lend us insight into who is likely to protect potential victims during persecution and genocide. In our second article, we take a deep dive into the social lives of voters in two 19th century American towns to see how social networks influence voting.

In our discussion, we consider the ethical tradeoffs of studying the dead without their consent. The Nazis collected this data in order to round up Jews for execution. Does that make it problematic for researches to use it? Is there a difference between using data about where their victims lived and using data collected from sadistic Nazi medical experiments? Meanwhile, the authors of the voting article used public records and archival data to find out not merely where voters lived but who they socialized with. Is it ethical to try to probe the personal lives of the

deceased for research purposes? Does it matter if these are ordinary citizens and not public figures? Does it matter how much time has passed?

Instead of a new in-class coding exercises, we practice **commenting, unit-testing, and debugging** three previous exercises as described below.

Due 24 hours after class: Programming Portfolio I. Select three of the coding exercises we have done in-class this semester to tidy up, add extra bells and whistles to, and turn them in (in-class exercises are not otherwise graded).

Week 6: Statistics, Networks, and Asking Good Research Questions

Readings: None

Building on last week's readings, we take a foray into the fast-growing field of network analysis and try our hand at some network exercises. We also discuss regression coefficients, standard errors, and the differences between description, prediction, and causal inference. In preparation for our semester-long project, we examine the differences between a topic, research question, and hypothesis and how to proceed from one to the next. Groups will have time to meet and refine their research questions with instructor feedback. Lastly, to reinforce our Python skills and better understand geospatial data analysis, we replicate Robert Braun's approach to analyzing the role of minority churches in saving Jews during the Holocaust using a simplified, simulated data set. In doing so, we practice writing **list comprehensions, indexing and slicing, and importing Python libraries.**

Due 24 hours after class: Each group must submit a 300-word abstract laying out their research question, accompanied by a bibliography.

Week 7: Twitter: Too Swift a Moving Target?

Readings

Munger, Kevin. 2017. "Tweetment Effects on the Tweeted: Experimentally Reducing Racist Harassment." Political Behavior 39 (3): 629–49. https://doi.org/10.1007/s11109-016-9373-5.

Munger, Kevin. 2019. "The Limited Value of Non-Replicable Field Experiments in Contexts With Low Temporal Validity." Social Media + Society 5 (3): 2056305119859294. https://doi.org/10.1177/2056305119859294.

Larson, Jennifer M., Jonathan Nagler, Jonathan Ronen, and Joshua A. Tucker. 2019. "Social Networks and Protest Participation: Evidence from 130 Million Twitter Users." American Journal of Political Science 63 (3): 690–705. https://doi.org/10.1111/ajps.12436.

In the past decade, Twitter has revolutionized the public sphere from political mass mobilization to middle school bullying to corporate marketing strategies. Yet the pace at which behavior on Twitter is changing and the evolution of the platform itself present challenges to social scientists.

Will a dataset collected today still be relevant by the time its results are published? Can those findings be verified or will changes to the underlying software render replication impossible? Are we learning anything useful about humanity if the phenomena we study are barely recognizable five years hence?

In addition to the issues of temporal validity described above, we also discuss the difference between experiments and large-scale observational studies, as well as what can be learned in both cases from a non-representative sample of the population. Are Twitter users a population worth studying in their own right? Do they have something to teach us about people in general? To what extent do behaviors on Twitter reflect human behavior offline? For our in-class Python exercise, we analyze a simulated Twitter experiment aimed at reducing middle school bullying to practice **working with text as data, retrieving data stored in dictionaries, and reading data from files.**

Due 24 hours after class: Each group must submit a 300-word pre-analysis plan detailing the data sources and methods they intend to use in their project.

Week 8: Proposal Presentations

Readings: None

We begin by discussing the oft-overlooked skill of how to deal with feedback when presenting—namely, how to respond to critiques while simultaneously taking them to heart. Groups then present their project proposals, including research question, hypotheses, data sources, and analysis plan. Providing useful feedback to other groups is included in the grade. Time permitting, we do an **object-oriented programming exercise** related to last-week's Twitter studies.

Week 9: Harnessing Data from Scientific Agencies and Natural Experiments

Readings

Gomez, Brad T., Thomas G. Hansford, and George A. Krause. 2007. "The Republicans Should Pray for Rain: Weather, Turnout, and Voting in U.S. Presidential Elections." *The Journal of Politics* 69 (3): 649–63. https://doi.org/10.1111/j.1468-2508.2007.00565.x.

Corbane, Christina, Thomas Kemper, Sergio Freire, Christophe Louvrier, Martino Pesaresi, European Commission, and Joint Research Centre. 2016. "Monitoring the Syrian Humanitarian Crisis with the JRC's Global Human Settlement Layer and Night-Time Satellite." Luxembourg: Publications Office. http://bookshop.europa.eu/uri?target=EUB:NOTICE:LBNA27933:EN:HTML.

Fowler, Anthony, and Andrew B. Hall. 2018. "Do Shark Attacks Influence Presidential Elections? Reassessing a Prominent Finding on Voter Competence." *The Journal of Politics* 80 (4): 1423–37. https://doi.org/10.1086/699244.

Just as we've seen historical and archival data repurposed, so too can we make use of data collected by governments and physical/natural scientists to answer social science questions. We first look the impact of rainfall on American elections and learn about natural and quasi-experiments. We contrast this framework with an observational study that uses images of the earth at night to monitor the Syrian refugee crisis. Returning to our theme of prediction versus causal inference, we discuss how each type of research design is best suited to different questions and where their weaknesses lie. We also discuss the problems with forecasting the effects of climate change on social problems. As we saw earlier in the semester when we looked at Twitter, it is difficult to use past or even present-day data when the phenomena we are studying are rapidly changing in ways we have never experienced. And yet to not study these pressing issues is unconscionable. How should social scientists go about studying an ever-changing world? Are there fundamental laws to be derived or only trends to observe and extrapolate?

In preparation for our research projects, we briefly examine how a scholarly report aimed at practitioners (second reading) differs stylistically from articles aimed at other academics. Lastly, we see why hungry sharks did *not*, in all likelihood, get Woodrow Wilson elected, and why the world would be hopelessly unpredictable place if this finding, and a hundred others like it, were all true simultaneously. We **visualize** this finding **using Mathplotlib in a Jupyter Notebook.**

Due this week or next: All groups must meet with the instructor during office hours to discuss the first draft of their research project (preferably in the form of a Jupyter Notebook).

Week 10: Surveys, Multiple Regression, and Bias

Readings

Mitchell, Kirstin R., Catherine H. Mercer, Philip Prah, Soazig Clifton, Clare Tanton, Kaye Wellings, and Andrew Copas. 2019. "Why Do Men Report More Opposite-Sex Sexual Partners Than Women? Analysis of the Gender Discrepancy in a British National Probability Survey." *The Journal of Sex Research* 56 (1): 1–8. https://doi.org/10.1080/00224499.2018.1481193.

Bakija, Jon. 2013. "A Non-Technical Introduction to Regression." https://web.williams.edu/Economics/wp/Bakija-Non-Technical-Introduction-to-Regression.pdf.

Surveys are ubiquitous in our society: university offices, stores, and apps are constantly asking us to evaluate them, while the media frequently report on public opinion and election polls. When it comes to interpreting the answers, we're often so focused on what's that we often neglect three equally fundamental questions: Who's being asked? How are they being asked? Who is doing the asking? After enumerating the many biases that can arise in the survey process (selection bias, social desirability bias, recall & salience bias, etc.) we learn about at how those biases can best be averted, minimized, or corrected. Multiple regression, though far from a panacea, allows us to control for observed confounders, which is an important step toward improving our estimates. **We practice working with survey data in Pandas data frames (imported with the csv module), examine how biases affect inference, and try out the multiple regression tools available in the scikit-learn and NumPy libraries.**

Week 11: Facebook and Proprietary Data Partnerships

Readings

Kramer, Adam D. I., Jamie E. Guillory, and Jeffrey T. Hancock. 2014. "Experimental Evidence of Massive-Scale Emotional Contagion through Social Networks." *Proceedings of the National Academy of Sciences* 111 (24): 8788–90. https://doi.org/10.1073/pnas.1320040111.

Bond, Robert M., Christopher J. Fariss, Jason J. Jones, Adam D. I. Kramer, Cameron Marlow, Jaime E. Settle, and James H. Fowler. 2012. "A 61-Million-Person Experiment in Social Influence and Political Mobilization." *Nature* 489 (7415): 295–98. https://doi.org/10.1038/nature11421.

Bakshy, Eytan, Solomon Messing, and Lada A. Adamic. 2015. "Exposure to Ideologically Diverse News and Opinion on Facebook." *Science* 348 (6239): 1130–1132. https://doi.org/10.1126/science.aaa1160.

Companies like Facebook have an unprecedented capacity to conduct social science research. For one thing, they are conducting *market* research on us every time we use their services. Is it ethical for them to also manipulate their users to address questions of scientific interest, given that they're already doing so to maximize profits? Facebook faced a massive PR debacle when it reportedly made millions of its users ever-so-slightly less happy to answer questions about how emotions are contagious (first article). Yet the media hardly commented when, in the second article, it ever-so-slightly nudged millions of its users to go vote. Is there a difference? What should our ethical guidelines be here?

There's a deeper scientific problem when it comes to working with Facebook data, however: you have to work with Facebook to get it. Our third article, published a full year before the 2016 election, treats Facebook merely as a microscope through which we watch polarization and media bubbles. Yet if the authors had found that Facebook *caused* these negative trends, would the company have allowed them to publish their findings? How can other researchers hope to replicate these findings if the data and code are not publicly available? Ultimately, Facebook and other proprietary platforms are too important to ignore, both as tools and objects of study. We discuss new partnership models, such as the Social Science One partnership, that seek to create greater transparency and replicability, while still protecting the anonymity of users (and, presumably, Facebook's trade secrets).

No new coding exercises today; students prepare previous exercises to be graded.

Due 24 hours after class: Programming Portfolio II (polish and enhance 2 in-class exercises from the past 5 weeks.

Due this week or next: All groups must meet with instructor during office hours to discuss the second draft of their research project Jupyter Notebook.

Week 12: Crowdsourcing, Event Data, and Deep Learning

Readings

Mellers, Barbara, Eric Stone, Terry Murray, Angela Minster, Nick Rohrbaugh, Michael Bishop, Eva Chen, et al. 2015. "Identifying and Cultivating Superforecasters as a Method of Improving Probabilistic

Predictions." *Perspectives on Psychological Science* 10 (3): 267–81. https://doi.org/10.1177/1745691615577794.
Jäger, Kai. 2016. "Not a New Gold Standard: Even Big Data Cannot Predict the Future." *Critical Review* 28 (3–4): 335–55. https://doi.org/10.1080/08913811.2016.1237704.

In-Class Quiz

All research design and stats concepts covered to date, with emphasis placed on regression (control variables, coefficients, confounders, etc.), (quasi)-experiments (control v. treatment, natural experiment v. RCT, randomization, stratification, ATE), and types of bias.

With enough sources of data, can we finally succeed in reliably predicting rare and dramatic events? Psychologist Philipp Tetlock and his collaborators believe the answer is yes, but by using people rather than machines. In our final regular class meeting, we discuss the successes of Tetlock's so-called superforecasters, the reliability of betting markets, and the ethics of Amazon Mechanical Turk. In doing so, we contrast the wisdom of crowds experts, the wisdom of experts, and the wisdom of "crowds of experts." We then turn to critiques by Kai Jäger (and many other scholars) of why fancy computer algorithms, even with human supervision, face fundamental limitations when it comes to predicting human activity. We go over various categories of machine learning—supervised, unsupervised, deep neural networks—and try to get a sense of why they may be highly reliable for some tasks but very bad at others. **As we implement our own machine learning algorithm in Python**, we consider the common critique that such algorithms present a "blackbox" which fails to shed light on causal mechanisms. In the end, the machines have learned a lot about us, but what have we managed to learn from them? We conclude by reviewing the differences between description, prediction, and causal inference.

Week 13: Final Draft Presentations of Research Projects

Readings: None

Groups will present their Jupyter notebooks to the class as if soliciting feedback at a conference prior to submitting a journal article for publication. Students are expected to give and incorporate feedback. Final version of Jupyter notebook due 5 days after final class.