Political Science UG4716: Data Science for Political Analytics

Fall 2024

Tuesdays and Thursdays, 2:40–3:55pm

Prof. Gregory Wawro (he/him)
gjw10@columbia.edu

Office Hours: Tues. and Thurs., 1:30–2:30 & by appt.
814 International Affairs Building

Teaching Assistant: TBD

Office Hours: TBD

The digital revolution has created previously unimaginable opportunities to learn about political behavior and institutions. It has also created new challenges for analyzing the massive amounts of data that are now easily accessible. Open source software has reduced barriers and inequities in coding, but it also requires different kinds of effort to employ the latest innovations optimally. Harnessing the power of political data is more critical than ever, given the threats that misinformation and alternative "facts" present to democratic forms of government.

This course will teach students both essential tools and general strategies of data science within the domain of politics. Whether students' goals are to analyze political behavior for academic or professional purposes, successful analysis requires skills for handling a wide array of issues that stand in the way of creating knowledge and insights from data.

This course prioritizes breadth over depth in the sense that we will introduce a broad range of topics relevant for data science to develop basic skills and form a foundation that students can build on. More complete mastery of these skills will require additional engagement beyond this course.

Learning Objectives

Students in this course will learn:

- the fundamental concepts and principles of data science, including data collection and transformation, analysis, and interpretation;
- how to tackle a range of problems that typically arise when working with data of various forms, including both numerical and text data;
- coding skills that are essential for working with political data to provide insights and make predictions;
- how to use visualizations to create narratives for effectively communicating data-driven insights;
- basic machine learning algorithms and techniques;
- best practices for ethical data usage and analysis.

Course organization

Lectures in the course will include live coding demonstrations and problem solving, with students working in teams to create solutions to a variety of problems that are encountered regularly with data collection, wrangling, and analysis. To enhance project-based learning, we will regularly flip

the classroom by asking teams to present their solutions to facilitate conversations about options and improving code. We have produced a suite of videos that will walk students through specific coding challenges. Students will be expected to engage with the videos prior to lectures but also to use them as they work on problem sets and projects. Each video will be accompanied by a short quiz that students can take after viewing with instant feedback to determine whether or not they have absorbed the key pieces of information conveyed in the video. These quizzes will not be part of the final grade for the course.

Course Requirements

The grading for the course is based on five problem sets (40%), in class pop quizzes (20%) and a final project (40%). Students will work in teams for the project, which will involve demonstration of facility with different skill sets in the course (e.g., data wrangling, manipulation, visualization, and reporting results). More details about the projects will follow. While we will not take attendance, students should consider attending class a requirement, given the emphasis on live coding demonstrations.

Late work will not be accepted except for reasons of certified medical necessity or family emergency.

Computing and Software

Students are required to bring a laptop to class. It is essential that students be able to follow along with the live coding demonstrations on their own computers. Student teams will also be called upon to demonstrate coding solutions using their own laptops. Students must refrain from using the laptops in class for activities that are not related to the class.

The primary software for the course will be `R`, but we will discuss other software (in particular `python`) where relevant. We assume no prior knowledge or experience with the software used in this class.

Course Readings

The following texts will be a focus in the class and students may want to purchase them.

Wickham, Hadley, and Garrett Grolemund. *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data* O'Reilly Media, Inc., 2016 (available for free by clicking the links embedded below).

Tufte, Edward R. *Visual Explanations: Images and Quantities, Evidence and Narrative*, Graphics Press, 1997.

Imai, Kosuke and Nora Webb Williams. *Quantitative Social Science: An Introduction in tidyverse.* Princeton University Press, 2022.

Grimmer, Justin, Margaret E. Roberts, and Brandon M. Stewart. *Text As Data: A New Framework for Machine Learning and the Social Sciences.* Princeton University Press, 2022.

Müller, Andreas C., and Sarah Guido. *Introduction to Machine Learning with Python: A Guide for Data Scientists.* O'Reilly Media, Inc., 2016 (available for free through Clio).

Academic integrity

Columbia University expects its students to act with honesty and propriety at all times and to respect the rights of others. It is fundamental University policy that academic dishonesty in any guise is unacceptable. It is essential to the academic integrity and vitality of this community that individuals do their own work and properly acknowledge the circumstances, ideas, sources, and assistance upon which that work is based. Academic honesty in class assignments and exams is expected of all students at all times.

The course is designed to teach students to learn how to learn, as assignments will push them to extend specific skills discussed into new territory. Key to this is learning how to use the wealth of online resources available (such as `stackoverflow` and `ChatGPT`) to figure out how to articulate questions about coding and adapt solutions that others have found for similar problems. The software used in the course, given their open source nature, have vast communities of users who share information about how to solve coding problems. It is important for students to learn how to use these resources while respecting academic integrity and intellectual property rights. Students will be asked to document usage of online resources with specific references in code and write-ups of analysis.

Accommodations

In order to receive disability-related academic accommodations for this course, students must first be registered with the Disability Services (DS) office. Detailed information is available online for the registration processes.

Course Outline

I have not included dates for the topics that we will cover in order to allow for maximum flexibility in the progress of the course. In my experience, classes of this nature require mid-semester corrections as we learn about students abilities and interests. I will regularly update students about what sections we are covering each week and what readings students should focus on. The reading lists includes clickable links. All journal articles and cheat sheets are available on CourseWorks.

I. **Introduction**

- What is data science and why does it matter for politics?

II. **Basic statistics and causal inference**

- Frameworks for causal inference
- Means, medians, variance
- Linear models
- Probability distributions, uncertainty, and hypothesis testing

Readings

- Imai, Chapters 1, 2, 4, 6, and 7

III. **R basics**

- Philosophy, syntax, packages

Readings

- Wickham and Grolemund, Chs. 1, 4, 6

IV. **Data wrangling**

- Importing, tidying, merging, preparing for analysis
- Functions and looping

Readings

- Wickham and Grolemund, Chs. 5, 10, 11, 12, 13, 17
- Data Wrangling Cheat Sheet
- dplyr Cheat Sheet
- forcats Cheat Sheet
- lubridate Cheat Sheet
- stringr Cheat Sheet
- Web scraping 101

V. **Data visualization**

- Theories of visualization
- Distributions, scatterplots, maps
- Interactive plots with Shiny

Readings

- Tufte, *Visual Explanations*, Chs. 1–3, 7
- Wickham and Grolemund, Ch. 3
- Traunmüller, Richard. "Visualizing Data in Political Science." *The SAGE Handbook of Research Methods in Political Science and International Relations* (2020): 436–460.

VI. **Big data**

- Definitions and special concerns
- Voter files; campaign finance data; Census data; social media data

Readings

- Grimmer, Justin. "We are all social scientists now: How big data, machine learning, and causal inference work together." *PS: Political Science & Politics* 48, no. 1 (2015): 80–83.
- Konitzer, Tobias, David Rothschild, Shawndra Hill & Kenneth C. Wilbur,"Using Big Data and Algorithms to Determine the Effect of Geographically Targeted Advertising on Vote Intention: Evidence From the 2012 U.S. Presidential Election." *Political Communication* 36 (2019):1–16.

VII. **Machine learning**

- Supervised and unsupervised learning
- Model evaluation and improvement
- Algorithm chains and pipelines

Readings

- Müller and Guido, Chs. 1–6.

VIII. **Text as data**

- Basics: working with strings and regular expressions
- Natural language processing: tokenization, stemming, word clouds, n-grams, sentiment analysis

Readings

- Grimmer, Roberts, and Stewart, Chs. 1–7, 13
- Ward, Brian, "A Light Introduction to Text Analysis in R"
- Wickham and Grolemund, Ch. 14

IX. **Data ethics**

- Respecting subjects
- Privacy and Confidentiality
- Data Ownership and Control
- Transparency and Accountability

Readings

- Ward, Ken. "Social networks, the 2016 US presidential election, and Kantian ethics: applying the categorical imperative to Cambridge Analytica's behavioral microtargeting." *Journal of Media Ethics* 33, no. 3 (2018): 133–148.
- Persily, Nathaniel. "Can Democracy Survive the Internet?" *Journal of Democracy* 28, no. 2 (2017): 63–76 (CW).
- Rose, Jeremy, and Oskar MacGregor. "The Architecture of Algorithm-Driven Persuasion." *Journal of Information Architecture* 6, no. 1 (2021): 7–40.

Acknowledgments: In addition to the assigned reading materials, this course draws heavily from two outstanding courses: Professor Matthew Blackwell's "Data Science for the Social Sciences" and Professor Brandon Stewart's "Applied Social Statistics."