Data Literacy and Data Visualization

Political Science 3780

Syllabus

Instructor Information
Bear F. Braumoeller
The Ohio State University
Department of Political Science
Office: Derby 2168
Office hours: Wednesday, 2:00–4:00 p.m., OBA
e-mail: braumoeller.1@osu.edu

January 21, 2014

Course description

Many, if not most, of the major debates in modern political science revolve around questions that can be addressed with data. The sources of voting behavior, the correlates of war, the determinants of development, political economy, psychology, institutions, and conflict—all are issues that are amenable to data-based analysis.

At the same time, the amount of available data and the number of publicly-available open-source tools for cleaning, transforming, analyzing and visualizing it have increased exponentially since the turn of the millennium. With a few clicks students can compare word frequencies in books over time or construct elaborate size-weighted wordclouds—tasks that would have taken scholars weeks if not months of effort in the past.

This course introduces students to those tools and the principles behind their use in the context of applications in political science. It marries the substance of political theory to the methodologies of data visualization and exploratory data analysis to help students better understand the social world. Moreover, the course introduces students to basic concepts in statistics and probability so that they can understand, not just the most plausible answer to their questions, but the degree of uncertainty surrounding that answer. It is designed to serve either as a standalone course or as a gateway to a more advanced data-analytics class.

Requirements

The format of the course is unusual: the lectures are online, and the Professor and the TA will meet weekly with the students in a small computer lab to work one-on-one with you through homework and exercises. Completing the exercises will require having listened to and viewed the lectures prior to class, but students may do so at whatever time is convenient. Questions about lecture material should be raised in the online Carmen forums, where they can be answered by the Professor and the TA.

The lectures are recorded in Quicktime format, which should be viewable on nearly any computer or mobile device. One advantage to the format is that, on most devices, lectures can be sped up to 1.4x to 1.5x; beyond about 2x, most of them become unintelligible. Students are advised to take advantage of this feature, as humans can generally understand speech at a higher rate than they can produce it. Be advised, though, that higher speeds generally require more focused attention, as important details are easier to miss. It's also possible to rewind and slow the recording back down if a particular section moves quickly or is difficult to understand.

Students will attend all recitation sections. There will not be a traditional midterm or final exam. Rather, weekly short assignments will make up 60% of the grade, and the remaining 40% will come from a final project in which the student finds a

dataset in his or her area of interest that is not already used in the course, analyzes it to assess the structure of the data, and works through the most appropriate, succinct, and informative summaries and visualizations. Students will be given the last 2-3 weeks of recitation sections to work in-class on final projects, with the Professor and the TA present to assist.

The final project is designed to be the foundation of an independent research project that will fulfill the requirements for a senior thesis. Students who are eligible to graduate with research distinction who are interested in writing a senior thesis should discuss doing so with their advisor very soon, as applications are due early in the fall semester. Funding for thesis research is available from the University. Applications for the Undergraduate Research Scholarship (URS) and the International Research Grant for undergraduates in the Arts and Sciences are available at `http://aschonors.osu.edu/opportunities/scholarships/undergrad`.

Again, students who wish to compete for funding for a senior project should speak to their advisors soon. The deadline for applications and project advisor recommendations is September 16, and selections will be made by mid-November. Applicants will compete for approximately 50 scholarships awarded in amounts ranging from $500 to $6,000. University regulations require that the funds be used to meet tuition and fee expenses, with any remaining amount available directly to recipients (provided they have not already exceeded the maximum allowable amount for financial aid).

Books

Three books are required for the class and one is recommended. There is no coursepack. The books are:

> Teetor, Paul. *R Cookbook.* (O'Reilly Cookbooks)
>
> Chang, Winston. *R Graphics Cookbook.* (O'Reilly Cookbooks)
>
> Yau, Nathan. *Visualize This: The FlowingData Guide to Design, Visualization, and Statistics.* (Wiley)
>
> Tufte, Edward. *The Visual Display of Quantitative Information.* (Graphics Pr) (Recommended)

GE Goals and Expected Learning Outcomes

Goals

Students develop skills in drawing conclusions and critically evaluating results based on data.

Achieving Expected Learning Outcomes

Students understand basic concepts of statistics and probability, comprehend methods needed to analyze and critically evaluate statistical arguments, and recognize the importance of statistical ideas.

The course will fulfill the General Education (GE) requirement in Data Analysis by helping students develop skills in drawing conclusions and critically evaluating arguments based on data. It will introduce students to basic concepts in statistics and probability, including sampling, data distributions, and the Central Limit Theorem, and it will teach students how to use iterated simulation and resampling (i.e., Monte Carlo simulation and bootstrapping) to obtain estimates of unknown probabilistic outcomes and to assign measures of accuracy to sample estimates. It will tie these elements together with the logic of research design in order to give students the ability to evaluate statistical arguments, and it will show them examples of how to do so using data on such topics as American partisanship and elections, drone strikes, and international conflict.

Assessment of Expected Learning Outcomes

The effectiveness of this course in achieving the expected learning outcomes outlined above will be determined in three ways.

1. A problem set will be assigned in class every week in weeks 2–13 to help students internalize the material delivered in the online lectures. The problem sets will be graded, students' mastery of the relevant skills will be assessed, and they will be informed of problem areas, if any.

2. During the last two weeks of class, students will work on an independent project of their choosing, one that involves answering questions using the skills they have acquired. This final assignment will count for 40% of the course grade.

3. Students will be urged to fill in the online Student Evaluation of Instruction (SEI) reports to assess the amount that they have learned in this course relative to others.

Grading

Appropriately for a course of this nature, grading takes place via visualization. Course grades are assigned on a curve, with the exact cutoffs depending on the distribution of student performance. Typically, grades in the top quintile receive As, those in the second quintile receive low As or high Bs, those in the third quintile receive Bs and low Bs, and those in the fourth and fifth quintiles receive Cs, Ds, or Es.

Academic Misconduct

It is the responsibility of the Committee on Academic Misconduct to investigate or establish procedures for the investigation of all reported cases of student academic misconduct. The term "academic misconduct" includes all forms of student academic misconduct wherever committed; illustrated by, but not limited to, cases of plagiarism and dishonest practices in connection with examinations. Instructors shall report all instances of alleged academic misconduct to the committee (Faculty Rule 3335-5-487). For additional information, see the Code of Student Conduct (http://studentaffairs.osu.edu/resource_csc.asp).

All students believe that they know how not to plagiarize. Many of them are wrong. Every year, many of them find that out the hard way. Don't be one of them.

The short version is that passing off another person's work or ideas as your own is plagiarism. That includes the unacknowledged word-for-word use or paraphrasing of another person's work or ideas. It is not enough, for example, simply to copy and paste a passage and then cite the source at the end. If the passage is taken word-for-word, it must be in quotes as well to indicate that fact.

There is an excellent video at http://hdl.handle.net/1811/46848, if you have any doubts. You should be crystal clear, as the University's policies exist to ensure fairness, and violators of University regulations on academic integrity will be dealt with severely.

Disability Services

Students with disabilities that have been certified by the Office for Disability Services will be appropriately accommodated, and should inform the instructor as soon as possible of their needs. The Office for Disability Services is located in 150 Pomerene Hall, 1760 Neil Avenue; telephone 292-3307, TDD 292-0901; http://www.ods.ohio-state.edu/.

Lecture Schedule

Date	To view before class
Week 1	Read syllabus
Week 2	Lectures 1 and 2
Week 3	Lectures 3 and 4
Week 4	Lectures 5 and 6
Week 5	Lectures 7 and 8
Week 6	Lectures 9 and 10
Week 7	Lectures 11 and 12
Week 8	Lectures 13 and 14
Week 9	Lectures 15 and 16
Week 10	Lectures 17 and 18
Week 11	Lectures 19 and 20
Week 12	Lectures 21 and 22
Week 13	Lectures 23 and 24
Week 14	Lectures 25 and 26
Week 15	Lectures 27 and 28

Lectures

Lecture 1: Introduction. Nature of the course, its relationship to other courses and students' plans of study; how and why data visualization can be useful.

Lecture 2: Data are All Around Us! Terminology, sources, and advice on what to do if you can't find a preexisting dataset. (Yau, pp. 21–43)

Lecture 3: Data Tools, part I. Open-source solutions that require little or no additional information or effort to produce compelling results.

Lecture 4: The Good. Exemplary representations of data from our field and others, with a discussion of what makes them so compelling. (Tufte, ch. 1; Yau, pp. 2–12)

Lecture 5: The Bad and The Ugly. Ideas that seemed good at the time, why they really aren't, how we can fix them, and what general principles they imply. (Tufte, ch. 2; Yau, 13–20)

Lecture 6: Data Tools, part II. These versatile tools allow you to upload, explore, and visualize your own data. (Yau, pp. 54–62)

Lecture 7: Truth. Data visualization is inherently a reduction of information. The main imperative, in reducing information, is integrity: representing the data as honestly as possible. We look at some good cases and some not-so-good cases. (Tufte, ch. 3)

Lecture 8: Introduction to R. A powerful (and free) general statistical package, `R` is capable of producing a wide range of graphs. In these lectures we demonstrate how to download and install the package, add libraries, and import data. (Yau, pp. 71–75; Teetor, chs. 1–4.)

Lecture 9: Beauty. The difference between a fairly good data visualization and an amazing one often lies in the application of a few straightforward graphical principles. (Tufte, chs. 4–6)

Lecture 10: R, part II. Data structures and data transformations. (Teetor, chs. 5–7.)

Lecture 11: APIs. What APIs are and why you should care.

Lecture 12: Plots in R. Scatterplots, line graphs, boxplots, dot charts, and more. (Teetor, chs. 8–10.)

Lecture 13: Time. Different ways of examining variables over time. (Yau, ch. 4)

Lecture 14: Space. Less common, but more eye-catching, are ways of looking for patterns across space, especially in maps. (Yau, pp. 80–89)

Lecture 15: Space and Time. Representing relationships and trends over both space and time. (Yau, ch. 5)

Lecture 16: Basic Probability. Combinations, permutations, distributions, and densities; generating random numbers. (Teetor, ch. 8.)

Lecture 17: Simulated Worlds. Visualizing probability and probabilistic outcomes via simulation.

Lecture 18: Samples and Populations. How resampling allows you to quantify the certainty of your generalizations.

Lecture 19: Research Design. Theories, hypotheses, research designs, tests, and conclusions.

Lecture 20: Workshop—Election 2012. How do political scientists predict election outcomes with such shocking accuracy?

Lecture 21: Workshop—Partisanship. Is American politics more partisan than it was 50 years ago? If so, why?

Lecture 22: Workshop—Politics and Plankton. A lesson in exploratory data analysis.

Lecture 23: Workshop—Drone Strikes. The pros and cons of warfare via unmanned aerial vehicle.

Lecture 24: Workshop—The End of War? Is war becoming more common, or is it actually on the decline?

Lecture 25: Workshop—The Austerity Debate. Is it better to spend one's way out of recession or cut back further?

Lecture 26: Workshop—Afghanistan. Mapping the WikiLeaks data on improvised explosive devices (IEDs).

Lecture 27: Workshop—Gaza. Exploring patterns of attacks and reciprocity.

Lecture 28: Conclusion: How to Lie Without Statistics. What are the limits of inference via data visualization? What more can data tell us that we can't access visually or intuitively? Worse, how can apparently obvious visual patterns mislead us? In short, why do people ever take statistics courses?