Gov 94JC: Making Big Data

Spring 2024 ∗ Mondays, 12:45-2:45 pm ∗ Location K450

Connor T. Jerzak
Visiting Assistant Professor of Government ∗ Harvard University
Assistant Professor of Government ∗ University of Texas at Austin

In the social sciences, quantitative analyses depend critically on the availability of high-quality data, but these data are rarely readily available at the click of a mouse. Instead, they are often messy, in difficult formats, or in forms not immediately useable in statistical software programs. In this class, we will learn how to find, curate, and create high-quality social science data. We will balance the teaching of practical tools—such as web scraping, crowd-sourcing techniques, online surveys, massive database management, and dataset merging—with the examination of theoretical perspectives on the role of measurement error in the analysis of causality. Code will also be treated as a form of data, and some teaching emphasis placed on how to make software packages that can improve reproducibility and enhance research impact. Projects completed during the course will give students an opportunity to create new, high-quality data contributing not only to their research projects but also to the broader cumulation of knowledge in the social sciences. Geared towards undergraduates with some coding background in R. □

British museum object #92687 showing a 6th century BCE Bablyonian map of the world. With better data gained via exploration, understanding of the earth would have been improved.
URL: `britishmuseum.org/collection/object/W_1882-0714-509`

1 Contact Information

Instructor
NAME: Connor T. Jerzak
EMAIL: connor.jerzak@austin.utexas.edu
URL: ConnorJerzak.com
OFFICE: K223
OFFICE HOURS: Wednesdays, 12:05-1:35 pm (no appointment needed).
Feel free to drop by my office anytime as well:
As long as I'm there and the door is open, I'm happy to talk anytime.
Location will be K223 unless the weather is good,
in which case we will hold office hours in the IQSS courtyard.

2 Prerequisites

By far the most important prerequisite for the course is a desire to learn the material and willingness to embrace the feeling of discomfort that can arise when learning new things.

This course also assumes some knowledge of data analysis using R. If you are unsure of whether your background will be sufficient for the course, reach out to the instructor.

3 Instructional Tools

We will use two tools to run this course.

- **Piazza**: Our course website will be housed on Piazza. On this Piazza, you will be able to find all materials related to the course, as well as to engage in discussions both with other students and instructors about the class content.
  - Piazza URL: https://piazza.com/utexas/spring2024/gov94jc/home
  - Signup: https://piazza.com/utexas/spring2024/gov94jc
    - Access code: xfj9kil6ayl
  - Gradescope URL: https://www.gradescope.com/courses/718362
    - Access code: EJNGZW
- **Gradescope**: Gradescope is the online platform you will use to submit and receive feedback on your assignments. You can access our Gradescope page via Piazza.

4 Course Framework

The course has several main components: in-class discussions, exercises done in pairs, and a collaborative end-of-semester project. We believe it is important to have multiple ways in which students can be evaluated since some students may learn best with each modality (e.g., in group work, in solo work, in project-based work).

Some guidelines.

- **Class Meetings**: In class meetings, key material will be introduced. We will typically begin class with a discussion of the day's content. A break will occur at the class half-way point. After the break, we will usually have a data and coding tutorial. We ask that students observe the following guidelines during class:

    1. *Be on time.* Late entry of students can be disruptive to the focus of students who arrived on time. There may be cases where, due to external circumstances, students may need to regularly arrive late; in that case, let the instructors know and we will think of a plan to minimize the degree of class disruption.
    2. *Keep computer use to essentials only.* Computers can be a useful aide to learning in some situations, but evidence has accumulated that they can be a source of distraction for those around you in others. Therefore, use of laptops and cell phone during the discussion component of the course should be limited.

- **Exercises**: Exercises will be done in partnership with a randomly selected peer pair. These pairs will be rotated every exercise. You and your partner will work together to answer the questions. Submissions are to be made to Gradescope by 5 pm Sunday. See next section for additional details.

5 Course Requirements

The course grade is based on the following components:

- **Class participation** (10% of the course grade): The class participation portion of one's grade is important but can sometimes feel ill-defined. By "participation," we mean engagement broadly defined: Does a student attend class? Is the student engaged in class or section? Is the student helpful to others on the Canvas discussion page? Does the student treat other students with courtesy and respect in collaborations or discussions? The later aspect of the grade is not meant to be punitive; in your career, your success will be affected by your ability to work well with others, so this aspect is also a part of the course evaluation.

- **Class presentations** (10% of the course grade): During the semester, we will close out each class with a discussion led by a student. In particular, each student will be tasked with the following:

    – Select a social science publication or working paper whose main contribution is the production of a new data source. Ideally, the data source will be relevant to the topic being discussed in a given week, although students may use their judgement in this regard.

    – Present the paper during class (15 min. presentation, 10 min. discussion). The presentation should emphasize the main contribution of the data source, describe the mode of data creation, describe the way of data dissemination, and discuss relevant strengths/weaknesses as well as what we can learn from this prior data creation effort.

- **Exercises** (40% of the course grade): Students will be required to complete exercises on a bi-weekly basis. You will be randomly assigned pairs. In these pairs, you will expected to support each other. The motivation for this pairing is to encourage students to (a) get to know a broad set of other students while (b) helping each other in your understanding

of the material. The use of bi-weekly exercises also encourages regular engagement with the material, something critical to learning. The following guidelines will apply to these exercises:

  - *Collaboration policy.* Students are allowed to discuss the exercises with the instructional staff and any other students in class in addition to your assigned pair. After all, teaching one's peers can be helpful in learning the material better oneself. However, you and your assigned partner are required to write up your own answers. You are not allowed to copy the work of other teams. Copying others' work deprives you of the opportunity to learn the material; if you feel like you are struggling in the course, this information is very important for the instructional staff to know so we can support you. If you feel like you are struggling in the course, it is better to let us know sooner rather than later so we can identify impactful solutions.
  - *Online help and office hours.* You are strongly encouraged to reach out to the instructional staff through Canvas and office hours about any questions you might have about the course materials. Students should also feel free to ask questions and answer the questions posed by others at Canvas, which will be considered in the overall evaluation of class participation.
  - *Submission guidelines.* The instructional staff will provide a submission template. For the exercises, each student should submit the PDF file electronically to Gradescope. Once you upload the PDF file, you will see a list of the questions in the assignment and thumbnails of your file. For each assigned question, click the PDF page(s) that contains your answer. No late submission will be accepted unless you obtain a prior approval from the instructor.

- **Final project** (40% of the course grade): The final project can be a collaborative project with another student in this class or can be done on an individual basis. There are pros and cons to either choice. In collaborative projects, you can often implement more ambitious analyses, as you will have the support of a peer. In an individual project, you can tailor it specifically to a part of your dissertation.

  The final project will involve the creation of a new data source, broadly defined. For more methodologically oriented students, the project can be the creation of a software package. For more substantively oriented students, the project can be the production of a new data source for applied research.

  The quality of project will be judged based on the originality of data creation effort and the clarity with which the ideas are presented.

  Each problem set will have one question related to your project progress!

  - Tentative project deadline: May 1. Requirements:
    * Report no more than 15 single-space page including the title, tables, and figures
    * *(Optional but encouraged)* Upload original data to Harvard Dataverse or HuggingFace
      · For instructions on how to upload your data, see support.dataverse.harvard.edu/getting-started

6 Course Materials

There is no single textbook for this course. Relevant readings are listed below and will be updated as needed. For those seeking additional information on the sociology of social science data, the

following is an excellent reference:

- Borgman, C.L., 2017. *Big data, little data, no data: Scholarship in the networked world.* MIT Press.

7 Course Plan

This course intends to provide an introduction to the following topics related to the creation of new data streams.

Module 1. Introduction

Cls. 1 (Mon. Jan. 22) Overview of the course

Reading
- Borgman, C.L., 2017. *Big data, little data, no data: Scholarship in the networked world.* MIT Press. Chapters 2-4.

Module 2. The Web as Data

Cls. 2 (Mon. Jan. 29) Introduction to web scraping + primary web languages: HTML, CSS, JavaScript, etc.

Reading
- Mughal, M.J.H., 2018. Data mining: Web data mining techniques, tools and algorithms: An overview. *International Journal of Advanced Computer Science and Applications*, 9(6).
- Code Academy, HTML Module 1– Elements and Structure. codecademy.com/learn/learn-html

Cls. 3 (Mon. Feb. 5) More on working with web languages; crawling; accessing historical web data

Reading
- Suchomel, V. and Pomikálek, J., 2012, April. Efficient web crawling for large text corpora. In *Proceedings of the Seventh Web as Corpus Workshop (WAC7)* (pp. 39-43).
- GNU Wget 1.21.1-dirty Manual. Sections 1, 3-4, 7. gnu.org/software/wget/manual/wget.html#Recursive-Download

Cls. 4 (Mon. Feb. 12) Data APIs—Twitter, Google Maps, OpenStreetMap, Zillow, etc.

Reading
- Fergusson, A. and Wild, C.J., 2021. On traversing the data landscape: Introducing APIs to data-science students. *Teaching Statistics*, 43, pp.S71-S83.
- Cooksey, B., 2014. An introduction to APIs. doi.org/10.1553/giscience2016_01_s191

Cls. 5 (Mon. Feb. 19) Working with spatial data; accessing remote sensing data; QGIS, rasters

Reading
- Introduction to Raster Data. datacarpentry.org/organization-geospatial/01-intro-raster-data/
- Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D. and Moore, R., 2017. Google Earth Engine: Planetary-scale geospatial analysis for everyone. *Remote sensing of Environment*, 202, pp.18-27.

Module 3. Collective Data Creation

Cls. 6 (Mon. Feb. 26) Online experiments; external collaborations

READING
- Grootswagers, T., 2020. A primer on running human behavioural experiments online. *Behavior Research Methods*, 52(6), pp.2283-2286.
- Chandler, J., Rosenzweig, C., Moss, A.J., Robinson, J. and Litman, L., 2019. Online panels in social science research: Expanding sampling methods beyond Mechanical Turk. *Behavior research methods*, 51(5), pp.2022-2038.

CLS. 7 (MON. MAR. 4) Crowd sourcing, expert coding

READING
- Sumner, J.L., Farris, E.M. and Holman, M.R., 2020. Crowdsourcing reliable local data. *Political Analysis*, 28(2), pp.244-262.
- Marquardt, K.L. and Pemstein, D., 2018. IRT models for expert-coded panel data. *Political Analysis*, 26(4), pp.431-456.
- Pemstein, D., Marquardt, K.L., Tzelgov, E., Wang, Y.T., Krusell, J. and Miri, F., 2018. The V-Dem measurement model: latent variable analysis for cross-national and cross-temporal expert-coded data. V-Dem Working Paper, 21.

MODULE 4. MANAGING BIG DATA

CLS. 8 (MON. MAR. 11) Working with massive data; SQL in R

READING
- Roger D. Peng, Sean Kross, and Brooke Anderson. *Mastering Software Development in R*. Sections 1.8-1.10. bookdown.org/rdpeng/RProgDA/building-r-packages.html
- Code Academy, SQL Module 1 – Manipulation. Try out additional modules if relevant to future research. codecademy.com/learn/learn-sql

CLS. 9 (MON. MAR. 18) Merging data; probabilistic matching; string matching; areal interpolation

READING
- Enamorado, T., Fifield, B. and Imai, K., 2019. Using a probabilistic model to assist merging of large-scale administrative records. *American Political Science Review*, 113(2), pp.353-371.
- Comber, A. and Zeng, W., 2019. Spatial interpolation using areal features: A review of methods and opportunities using new forms of data with coded illustrations. *Geography Compass*, 13(10), p.e12465.

MODULE 5. CODE AS DATA

CLS. 10 (MON. MAR. 25) Making R packages; git

READING
- Roger D. Peng, Sean Kross, and Brooke Anderson. *Mastering Software Development in R*. Chapter 3. bookdown.org/rdpeng/RProgDA/building-r-packages.html
- Blischak, J.D., Davenport, E.R. and Wilson, G., 2016. A quick introduction to version control with Git and GitHub. *PLoS computational biology*, 12(1), p.e1004668.

CLS. 11 (MON. APR. 1) Reproducibility; dockers; Code Ocean; Harvard Dataverse

READING
- "Reproducible Data Science with Docker Containers - Ben Hamner", youtube.com/watch?v=cYavmPa3aFQ
- Borgman, C.L., 2017. *Big data, little data, no data: Scholarship in the networked world.* MIT Press. Chapters 8-10.

MODULE 6. THEORY OF DATA

Cls. 12 (Mon. Apr. 8) Measurement error; uncertainty estimation; proxy variables; measurement error & causality

Reading
- le Cessie, S., Debeij, J., Rosendaal, F.R., Cannegieter, S.C. and Vandenbrouckea, J.P., 2012. Quantification of bias in direct effects estimates due to different types of measurement error in the mediator. *Epidemiology*, pp.551-560.
- Miao, W., Geng, Z. and Tchetgen Tchetgen, E.J., 2018. Identifying causal effects with proxy variables of an unmeasured confounder. *Biometrika*, 105(4), pp. 987-993.
- Cvetojevic, S., Juhasz, L. and Hochmair, H., 2016. Positional accuracy of twitter and instagram images in urban environments. *GI Forum*, 1, pp.191-203.

Module 7. Students Vote

Cls. 13 (Mon. Apr. 15) Students vote on topics to be covered based on student proposals

Reading – TBD

Final Project **Due by 11:59 pm Cambridge time, May 1 (tentative)**

8 Q&A

- *Q: Why is the theory section placed towards the end of the class?*

  A: Student projects will require the use of the computation skills taught in the class. The sooner students learn these skills, the sooner they will be able to apply them in building up their project dataset. Therefore, we have placed the teaching of these skills early on in the class.

- *Q: Will there be guest speakers?*

  A: There may be guest speakers are various points in the course depending on speaker availability.

- *Q: Why might I not want to make my project data publicly available yet?*

  A: Making your data publicly available will help other researchers build on what you've done. You may opt not to make your data publicly available until after your dissertation is released or relevant papers have made it through the peer-review pipeline.

9 A Note on the Readings

When doing the readings, there are several strategies that you might pursue.

1. First, you could start by skimming over the article, looking at the equations (or figures) and trying to understand what they're telling you. After thinking about that, then you may have a better framework for where you should focus your attention in the article, and what material is relatively less important to linger on.

2. A second approach is to read the introduction and conclusion, getting a sense for the big picture, and then diving into the details.

Don't ever feel discouraged if you feel like you aren't understanding parts of the readings!

Get what you can out of it in your first read-through—then talk to friends, consult YouTube videos, or come to Office Hours to work through your understanding. A great way to understand the intuition for something better is to implement it yourself (as Richard Feynman famously said, "What I cannot create, I do not understand.").

No one's understanding of a topic is total, we are all in a learning journey that sometimes involves valleys but sometimes also involves peaks!

10 Other Resources

There are other resources that could be useful to students in different aspects of the course, such as when giving presentations or when undertaking the creative process of formulating a research project.

On giving presentations:

- Winston, Patrick."How to Speak." youtube.com/watch?v=Unzc731iCUY

On writing:

- Pinker, Stephen. *The Sense of Style: The Thinking Person's Guide to Writing in the 21st Century.* Penguin Books.
- Williams, Joseph and Joseph Bizup. *Style: Lessons in Clarity and Grace.* Pearson.

11 Arrangements

Students with arrangements through the Harvard Disability Access Office should ensure that those arrangements are communicated to the instructor so we can set up a constructive environment for you.

12 Integrity

"Members of the Harvard College community commit themselves to producing academic work of integrity—that is, work that adheres to the scholarly and intellectual standards of accurate attribution of sources, appropriate collection and use of data, and transparent acknowledgement of the contribution of others to their ideas, discoveries, interpretations, and conclusions. Cheating on exams or problem sets, plagiarizing or misrepresenting the ideas or language of someone else as one's own, falsifying data, or any other instance of academic dishonesty violates the standards of our community, as well as the standards of the wider world of learning and affairs."

—Harvard College Honor Code

13 Acknowledgements

I am grateful to Matthew Blackwell, Andrés Cruz, Adel Daoud, Frances Hagopian, Kosuke Imai, Stephen Jessee, Siyun Jiang, Gary King, Matthew Martin, Michael Peress, Xiaolong Yang, and Xiang Zhou; the course builds on conversations or prior teaching experience with them. Thanks also go to previous students I have had the opportunity to work with, whose feedback has been very valuable.

14 Final Reflections

The instructors reserve the right to modify aspects of the course as it proceeds, both in response to student feedback and our own assessment of how to improve the course.