This syllabus comes from https://lincolnmullen.com/courses/data.2023/.

Only the online version of this syllabus is authoritative, and it may be updated as necessary.

Data Analysis for History (Spring 2023)

Course: HIST 697-001. Spring 2023. Department of History and Art History, George Mason University. 3 credits. Meets Mondays, 7:20–10:00pm in Peterson Hall 2408.

Instructor: Lincoln Mullen <lmullen@gmu.edu>. Office: Research Hall 484. Office hours: By appointment. Book an appointment.

Course description

In this course you will learn to use computational methods to create historical interpretations. You will work with historical data, which includes finding, gathering, manipulating, analyzing, visualizing, and arguing from datasets, with special attention to geospatial, textual, and network data. These methods will be taught primarily using programming languages for data analysis. While data analysis methods can be applied to many topics and time periods, they cannot be understood separate from how the discipline forms meaningful questions and interpretations, nor divorced from the particularities of the sources and histories of some specific topic. You will therefore work through a series of example problems using datasets from the history of the nineteenth-century United States, and then apply the methods to write a research paper using a dataset from your own historical field.

Learning goals

After taking this course, you will be able to

- gather historical data from print and manuscript sources; use existing historical data sets; clean, tidy, and manipulate data; perform exploratory data analysis; create common visualizations; work with geospatial, textual, and network data.
- write scripts using the R programming language and its extensive set of packages, as well as gain a basic understanding of data visualization for the web.
- understand the place of data analysis and visualization within the field of digital history and the discipline of history.
- conceive of and execute a short research project in computational history.

Essential information

Most required readings are available online or through the GMU libraries. These are the main books that we will be using.

- Hadley Wickham and Garrett Grolemund, *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data* (O'Reilly, 2017). ISBN: 978-1491910399.
- Kieran Healy, *Data Visualization: A Practical Introduction* (Princeton University Press, 2018). ISBN: 978-0691181622.

This is a graduate methods course in a field that moves reasonably quickly. The syllabus is likely to change over the course of the semester. In particular, I am likely to send you additional projects or visualizations to look at before class, which should be treated the same as other assigned readings.

All communication for this course will happen in our Slack group. Please signup and join the `#data-2023` channel. Read this getting started guide if you need help. The Slack group is your primary place to ask for help. Please ask for help in the public channels rather than private messages. You are almost certainly not the only person to have your question, and asking and answering questions publicly benefits everyone. When you ask a question, help me help you by

including the code that you are asking about and any error messages that are relevant.

You are always welcome to talk with me during office hours. My office hours page has instructions on how to book an appointment. If the scheduled times don't work for you, please contact me and suggest a few other times that would work for you.

Bring a computer to each class meeting. We will be using both the R programming langauge and JavaScript via Observable notebooks. You can use R on your computer or via GMU's OpenOnDemand instance, which you can use via a web browser. I also recommend that you install some key software on your computer. See the list under the heading for the first week. (I will assume that you have a computer with some kind of Unix-like operating system available. The easiest will be macOS or a Linux distribution. But if you use Windows, good news: R has very good support for Windows.)

In general I have provided datasets and questions for you to work on for all the assignments except the final paper. But for any assignment, you may substitute a dataset from your own historical field after checking with me. The forward-thinking graduate student will try to find such datasets early on in the semester so that you can use the intermediate assignments as preparation for your final assignment. If you can peer even farther into the future, you could try to use the final assignment as a test run for work you might want to do in one of your own research projects, such as a conference presentation, article, or dissertation.

Assignments

For each assignment, you should send me the completed file (not the source code) knit from your RMarkdown document. Please submit the assignments via the Blackboard page for this class. Send the assignments before the start of class on the day on which they are due.

Preparation and participation are expected as a matter of course in a graduate class. Complete all readings and submit all assignments

before class. If the readings include sample code or questions at the end, work through them as part of doing the readings, though you do not need to to submit them and I will not check them. Final grades will be calculated using the typical percentage-based grading scale (A = 93–100, A- = 90–92, B+ = 88–89, B = 83–87, B- = 80–82, … F = 0–59).

Worksheets and weekly assignments (25%). Many classes will have an assignment due before class begins. Some will require you to do library research; others will be practice data analysis worksheets. Some of the questions on the worksheets will be easy; most will be difficult; some you may find nearly impossible. The aim is to *practice*. We will go over the worksheets in class each week. If you attempt a problem and can't solve it, you should still turn in whatever work you did on it. Students who complete all the easy and moderately difficult questions, attempt the very difficult questions, and ask for help as needed will do just fine. These assignments will graded by completion.

Analysis assignments (4 × 10% = 40%). You will do four analysis assignments, each demonstrating a specific skill in data analysis. For these assignments you will be use a historical dataset and asked some interpretative questions. You will prepare a Quarto document containing prose, code, and tables or visualizations to answer the historical questions and, as necessary, explain your methods. For these assignment I will provide a dataset that you can work with (but see below).

Final project (35%). You will designate one of the analysis assignments as a stepping stone to your final project. For that analysis assignment, you will use the same dataset that you will use for the final project. You will try out one of the methods we are learning on that dataset. In addition to the normal feedback that I will provide on an assignment, I will also give you guidance about how to refine and expand your analysis, visualizations, and interpretations. Then, you will expand and revise the work you did in the analysis assignment for the final project. This expanded version should include more prose and citations, not to exceed 1,500 words. You may do this assignment either in a Quarto notebook or in an

Observable notebook. The visualizations and data analysis should be expanded if necessary and refined in each case to the level of quality that would be expected in a published article. Each table and figure must have a caption written in complete sentences. Explain your methods as needed, but write in a way which would be understandable and compelling to any historian working in your field. The final assignment will be evaluated according to two primary criteria: (1) Did the visualizations significantly improve in refinement and quality? (2) Does the combination of prose and visualizations convey a meaningful historical argument? **Due by 12pm on Monday, May 15.**

Schedule

Week 1 (January 23): Introduction to computational history

Assignment:

- Find one example of a digital history project that uses visualization or data analysis. Be prepared with the URL and a three-minute answer to these questions: What is interesting or insightful about this project? What did this project do that you would like to learn how to do for your own research?

Readings:

- Frederick W. Gibbs, "New Forms of History: Critiquing Data and Its Representations," *The American Historian* (February 2016).
- Taylor Arnold and Lauren Tilton, "New Data: The Role of Statistics in DH," in *Debates in DH 2019*, ed. Matthew K. Gold and Lauren F. Klein (University of Minnesota Press, 2019).
- "Introduction," in *Computational Historical Thinking*.

We will set these up the first day of class:

- Join the class Slack group.
- Get a GitHub account and post it to the Slack group (e.g., I am `lmullen` and this is my GitHub profile).
- Install R, a programming language for data analysis.

- Install RStudio, an environment for using R.
- Install Homebrew (only if you use macOS).
- Install Visual Studio Code, a general-purpose text editor for developers.
- Get an account at Observable.

Week 2 (January 30): Data from history and historians

Assignment:

- Find at least three primary source data tables, datasets, or corpora from your field of historical research. These could include sources that are in print or manuscript, as well as datasets that have already been created. Post full citations and URLs in the Slack group, along with a sentence or two explaining what you've found. Examine the links that other people post before class.

Readings:

- Roger Finke and Rodney Stark, *The Churching of America, 1776-2005: Winners and Losers in Our Religious Economy* (Rutgers University Press, 2005), ch. 1.
- Chad Gaffield, "Words, Words, Words: How the Digital Humanities Are Integrating Diverse Research Fields to Study People," *Annual Review of Statistics and Its Application* 5, no. 1 (2018): 119–39, https://doi.org/10.1146/annurev-statistics-031017-100547.
- Abraham Gibson and Cindy Ermus, "The History of Science and the Science of History: Computational Methods, Algorithms, and the Future of the Field," *Isis* 110, no. 3 (2019): 555–66, https://doi.org/10.1086/705543.
- Jessica Marie Johnson, "Markup Bodies: Black [Life] Studies and Slavery [Death] Studies at the Digital Crossroads," *Social Text* 36, no. 4 (2018): 57–79, https://doi.org/10.1215/01642472-7145658.
- Laurie F. Maffly-Kipp, "If It's South Dakota You Must Be Episcopalian: Lies, Truth-Telling, and the Mapping of U.S. Religion" *Church History* 71, no. 1 (2002): 132–42.

- Shari Rabin, "'Let us Endeavor to Count Them Up': The Nineteenth-Century Origins of American Jewish Demography," *American Jewish History* 101, no 4 (2017): 419–440, https://doi.org/10.1353/ajh.2017.0060.

Browse:

- Robert K. Nelson et al., *Atlas of the Historical Geography of the United States* (Digital Scholarship Lab, University of Richmond).
- Herman Carl Weber, *Presbyterian Statistics through One Hundred Years, 1826-1926* (Philadelphia: Presbyterian Church in the U.S.A., 1927), part II.
- Jasmine Weber, "How W.E.B. Du Bois Meticulously Visualized 20th-Century Black America," 5 February 2019.

Week 3 (February 6): Basics of R

Assignment:

- Getting familiar with R worksheet.
- Either use a primary source dataset that you found last week or, as a backup, the *Minutes* of the Methodist Episcopal Church from after 1851. Create a well-structured spreadsheet and transcribe at least 25 rows of the data. Upload a CSV file to Slack before class. Be prepared to describe in class how you decided on the structure of your data, and how you identified what the variables were. Use the Broman and Woo article as a guide.

Readings:

- Wickham and Grolemund, *R for Data Science*, ch. 1, 4, 6, 8, 27.
- Karl W. Broman and Kara H. Woo, "Data Organization in Spreadsheets," *American Statistician* 72, no. 1 (2018): 2–10, https://doi.org/10.1080/00031305.2017.1375989.
- "Getting Started" and "An R Primer" in *Computational Historical Thinking*.
- Quarto documentation.

Week 4 (February 13): Data manipulation

Assignment:

- Data structures worksheet.
- Functions worksheet.

Readings:

- Wickham and Grolemund, *R for Data Science*, ch. 5, 12, 13.
- Documentation for the tidyverse.
- Documentation for databases in R.

Week 5 (February 20): Data visualization

Assignment:

- Data manipulation worksheet.

Readings:

- Healy, *Data Visualization*, ch. 1, 3, 4.
- Wickham and Grolemund, *R for Data Science*, ch. 3, 28.
- Kieran Healy and James Moody, "Data Visualization in Sociology" *Annual Review of Sociology*, 40:105–128.
- Lauren F. Klein, "The Image of Absence: Archival Silence, Data Visualization, and James Hemings," American Literature 85, no. 4 (December 1, 2013): 661–88, https://doi.org/10.1215/00029831-2367310.
- John Theibault, "Visualizations and Historical Arguments," in *Writing History in the Digital Age*, ed. Kristen Nawrotzki and Jack Dougherty (University of Michigan Press, 2013), https://doi.org/10.3998/dh.12230987.0001.001.

Week 6 (February 27): Exploratory data analysis

Assignment:

- Data visualization worksheet.

Readings:

- Wickham and Grolemund, *R for Data Science*, ch. 7, 17–21, 30.

- Healy, *Data Visualization*, ch. 5, 8.
- Roger Peng, *Exploratory Data Analysis with R* (Leanpub, 2016), ch. 1, 4–6.
- Jordan F. Bratt, "Congressional Incumbency in the Early Republic," *Mapping Early American Elections* (RRCHNM, 2019).

Week 7 (March 6): Maps

Assignment:

- Exploratory data analysis assignment.

Readings:

- Healy, *Data Visualization*, ch. 7.
- Richard White, "What is Spatial History?," *Spatial History Project* (Stanford University, 2010).
- Cameron Blevins, "Space, Nation, and the Triumph of Region: A View of the World from Houston," *Journal of American History* 101, no. 1 (2014): 122–47, https://doi.org/10.1093/jahist/jau184.
- Browse: Robert K. Nelson and Edward L. Ayers, eds., *American Panorama: An Atlas of United States History* (Digital Scholarship Lab, University of Richmond).

For reference:

- Documentation for the sf package.
- Documentation for the leaflet package.
- Greta Swain, "Maryland's Political Geography in the Early Republic," *Mapping Early American Elections* (RRCHNM, 2019).

Spring break (March 13)

Week 8 (March 20): Networks

Assignment:

- Mapping assignment.

Readings:

- Mark E. J. Newman, *Networks: An Introduction* (Oxford University Press, 2010), ch. 1, 3, 4. Skim chs. 6, 7.
- Matthew Lincoln, "Social Network Centralization Dynamics in Print Production in the Low Countries, 1550–1750," *International Journal for Digital Art History* 2 (2016): 134–157, https://doi.org/10.11588/dah.2016.2.25337.

Browse:

- Analysis repository for civil procedure codes.

For reference:

- Documentation for the ggraph package.

Week 9 (March 27): Texts

Assignment:

- Network assignment.

Readings:

- Kasper, Welbers, Wouter van Atteveldt, and Kenneth Benoit, "Text analysis in R," *Communications Methods and Measures* 11, no. 4: 245–265, https://doi.org/10.1080/19312458.2017.1387238.
- Taylor Arnold, Nicolas Ballier, Paula Lissón, and Lauren Tilton, "Beyond Lexical Frequencies: Using R for Text Analysis in the Digital Humanities," *Language Resources and Evaluation* 53, no. 4 (2019): 707–733, https://doi.org/10.1007/s10579-019-09456-6.
- Tim Hitchcock and William J. Turkel, "The *Old Bailey Proceedings, 1674–1913*: Text Mining for Evidence of Court Behavior," *Law and History Review* 34, no. 4 (2016): 929–955, https://doi.org/10.1017/S0738248016000304.
- Joshua Catalano, "Digitally Analyzing the Uneven Ground: Language Borrowing Among Indian Treaties," *Current Research in Digital History* 1 (2018): https://doi.org/10.31835/crdh.2018.02.

- Ryan Cordell, “Reprinting, Circulation, and the Network Author in Antebellum Newspapers,” *American Literary History* 27, no. 3 (2015): 417–445, https://doi.org/10.1093/alh/ajv028.
- Browse: Taylor Arnold, Courtney Rivard, Lauren Tilton, *Layered Lives: Rhetoric and Representation in the Southern Life History Project* (Stanford University Press, 2022): https://doi.org/10.21627/2022ll.

For reference:

- Documentation for quanteda package.
- Documentation for cleanNLP documentation.

Week 10 (April 3): Word embeddings

Readings:

- Ben Schmidt, “Vector Space Models for the Digital Humanities” (October 25, 2015).
- Ben Schmidt, “Rejecting the Gender Binary: A Vector-Space Operation” (October 30, 2015).
- Ryan Heuser, “Word Vectors in the Eighteenth Century.”
- Matthew K. Gold and Lauren F. Klein et al., “Forum: Text Analysis at Scale,” in *Debates in the Digital Humanities 2016* (University of Minnesota Press, 2016), 525–568.
- Jo Guldi, “Critical Search: A Procedure for Guided Reading in Large-Scale Textual Corpora,” *Journal of Cultural Analytics* (2018): https://doi.org/10.22148/16.030.

Week 11 (April 10): Supervised and unsupervised classification

Assignment:

- Text analysis assignment.

Readings:

- Roger Peng, *Exploratory Data Analysis with R* (Leanpub, 2016), ch. 12.

- Robert K. Nelson, *Mining the Dispatch* (Digital Scholarship Lab, University of Richmond).
- Benjamin Schmidt, "Stable Random Projection: Lightweight, General-Purpose Dimensionality Reduction for Digitized Libraries," *Journal of Cultural Analytics* (2018): https://doi.org/10.22148/16.025.
- Skim Gareth James, et al., *An Introduction to Statistical Learning: With Applications in R* (Springer, 2013), ch. 10. GMU library
- Wickham and Grolemund, *R for Data Science*, ch. 23–24.
- Matthew L. Jockers and Ted Underwood, "Text-Mining the Humanities" in *A New Companion to Digital Humanities*, ed. Susan Schreibman, Ray Siemens, and John Unsworth (Wiley, 2016), 291–306. GMU library

Week 12 (April 17): Introduction to JavaScript and Observable notebooks

For this week, we will do a deep dive into Observable notebooks, which is a way of doing similar work with JavaScript rather than R or Python. Read as much of the Observable documentation as you can, focusing on the hands-on tutorials.

Week 13 (April 24): Introduction to Observable plot

Read the Observable documentation about their libraries and working with data.

Readings:

- Rebecca Sutton Koeser, "Trusting Others to 'Do the Math'" *Interdisciplinary Science Reviews* 40, no. 4 (2015): 376–392, https://doi.org/10.1080/03080188.2016.1165454.
- Benjamin Schmidt, "Do Digital Humanists Need to Understand Algorithms?" in *Debates in the Digital Humanities 2016*, ed. Matthew K. Gold and Lauren F. Klein (University of Minnesota Press, 2016).

Week 14 (May 1): Final project workshop

Readings:

- Read as much as you can: Jeri Wieringa, "A Gospel of Health and Salvation" (PhD dissertation, George Mason University, 2019). Assignment:

- Circulate a draft of your final project in Slack by Friday, April 28. Be prepared to present your work in class for approximately ten minutes. Read each person's draft and come prepared to offer helpful comments on their work.

Fine print

This syllabus may be updated online as necessary. The online version of this syllabus is the only authoritative version.

Students must satisfactorily complete all assignments in order to pass this course. I am sometimes willing to grant extensions on assignments for cause, but you must request an extension before the assignment's due date. For graduate students, I never penalize late work, but falling behind is not a good idea. No work (other than final projects) will be accepted after the last day that the class meets. I will discuss grades only in person during office hours.

Please submit all assignments in Blackboard.

You are expected to attend each class and to participate actively (exceptions made only for health reasons, religious holidays, and other university-approved excuses). Whether or not students attend class consistently is the best indicator of how well they will do in the class. If you wish to be excused for an absence, please email me before the absence if possible, or as soon as possible after the absence. I understand that life happens, and I will do my best to work with you.

See the George Mason University catalog for general policies, as well as the university statement on diversity. You are expected to know and follow George Mason's policies on academic integrity and the honor code. If you are a student with a disability and you need academic accommodations, please see me and contact the Office of

Disability Services through their website. You are responsible for verifying your enrollment status. All academic accommodations must be arranged through that office. Please note the dates for dropping and adding courses from the GMU academic calendar.

This syllabus draws ideas and assignments from many people and syllabi, including Taylor Arnold, Andrew Goldstone, Jason Heppler, Ben Schmidt, and Lauren Tilton.
