Data and Visualization in Digital History

HIST697-001: Creating History in New Media. Spring 2017. Department of History and Art History, George Mason University. 3 credits. Meets Mondays, 7:20 p.m. to 10:00 p.m., at RRCHNM, Research Hall fourth floor.

Instructor: Lincoln Mullen <lmullen@gmu.edu>. Office: Research Hall 457. Office hours on Tuesdays, 1:30 p.m to 2:30 p.m.; Thursdays, 2:00 p.m. to 3:00 p.m.; and by appointment.

Course description

In this methods course you will be introduced to data analysis and visualization for historians. You will learn to work with historical data, including finding, gathering, manipulating, analyzing, visualizing, and arguing from data, with special attention to geospatial, textual, and network data. These methods will be taught primarily through scripting in the R programming language, using other command line tools as appropriate. You will also learn how to present history on the web with HTML and CSS. While historical 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. Therefore, in this course we will examine the historiographical tradition to see how historians have used data and visualization to understand the past. And we will work together to apply these methods to a series of datasets in the history of the nineteenth-century United States, with a focus on religion.

After taking this course, you will be able to

- perform exploratory data analysis; clean, tidy, and manipulate data; gather historical data from print and manuscript sources; use existing historical data sets provided by government or other research groups; 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 use command line programs.
- understand the place of data analysis and visualization within humanities computing, digital history, and the discipline of history.
- conceive of and execute a research project in computational history suitable for treatment in a dissertation chapter or journal article.
- take the course "Programming in History/New Media," a.k.a. Clio 3, should you choose.

Essential information

You are always welcome to come to my office hours or to arrange an appointment. All other communication for this course will happen in our Slack group. Read this getting started guide if you need help.

Please obtain copies of the following books. All other readings are available online or through the GMU libraries.

- Isabel Meirelles, *Design for Information: An Introduction to the Histories, Theories, and Best Practices Behind Effective Information Visualizations* (Rockport Publishers, 2013).
- Susan Schulten, *Mapping the Nation: History and Cartography in Nineteenth-Century America* (University of Chicago Press, 2012).

Bring a computer to each class meeting. You will need the following software.

- We will use R and RStudio. Install them on your own computer. You will also have access to an RStudio Server instance which will let you use R in your browser.
- Get a real text editor and CSV reader. Sublime Text and Atom seem to be popular choices. The one true text editor is Vim, but it has a steep learning curve. Download LibreOffice or use Google Sheets because both do better than Excel at not messing with CSV files.
- Much of your work for the course will go on GitHub. Sign up for an account.
- If you do not already have a web host, I recommend Reclaim Hosting.

Assignments

For any assignment, **you are more than welcome to work on your own research topic,** provided that you check with me first. Ideally your work in this course will advance your dissertation or other research project. Talk to me early in the course and we will figure out together what would work best for you.

Be prepared. Preparation and participation are expected as a matter of course in a graduate class. Complete all readings and assignments before class. If the readings include sample code or questions at the end, work through them as part of doing the readings. By noon on the day of class, please post a brief message to the course Slack group about one thing you learned and one thing you want to see better explained. I will use your messages to tailor my explanations in class.

Worksheets and weekly assignments (45%). Most classes will have an assignment due before class begins. Some will require you to do library research; others will be data analysis worksheets. The worksheets are for you to practice what we have gone over in class. They will also serve as a helpful reference sheet when you need to remember how to perform some kind of analysis. Some of the questions on the worksheet 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. Unless otherwise specified, ask for help in the Slack channel but try not to just give the answers away. Each student must turn in his or her own worksheet. If you attempt a problem and can't solve it, you should still turn in whatever work you did on it. Here is how I will grade these worksheets. Each question will receive a 2 for a correct answer correctly solved; a 1 for an incorrect answer with a good attempt at solving the problem; and otherwise a 0. I will multiply the score for the question by a number for the difficulty of the question, so that more difficult questions are worth more. I will total up your scores as we go, and grade them on a curve at the end of the semester. Students who complete all the easy and moderate difficulty questions, attempt the very difficult questions, and ask me for help as needed will do just fine.

Visualization essays (3 essays @ 15% each). You will write three brief visualization essays, each demonstrating a specific skill in data analysis. At a minimum these essays should include the following: (1) explanatory text written for a scholarly but non-technical audience that frames the visualizations in both historical and historiographical terms and argues some worthwhile point; (2) as many well-crafted visualizations or data tables as are necessary to argue that point; (3) access to the code through GitHub (though the code should not be visible in the essay itself) which can be successfully re-run by me; and (4) citations. These essays should be submitted as R Markdown documents via GitHub, each as its own repository. Ideally you will write these essays using some historical data of interest to you, but I will provide data if you don't have any of your own. Students who wish to attempt some more ambitious project (which must still meet those requirements) are welcome to talk with me about combining two or three of the visualization essays into a single format.

Tutorial (10%). At our second meeting, you will pick from a list of analytical techniques, R packages, and other software not covered in this class. Beginning at week 6, you will be assigned a week to teach the class for 25 minutes about the topic you selected. As part of that teaching, you will prepare a handout with an

explanation of why the package or method is useful, some suggested applications to history, and sample code and results. The Programming Historian tutorials are a helpful model. Submit the human-readable form of the tutorial to the course Slack the day before class, so everyone can read it before class. Submit the GitHub repository with your handout to me.

Schedule

Week 1 (Jan. 23): Introduction

Assignment: Join the Slack group. Install R and other necessary software.

Read:

- Fletcher W. Hewes and Henry Garnett, *Scribner's Statistical Atlas of the United States Showing by Graphic Methods Their Present Condition and Their Political, Social and Industrial Development* (New York: Charles Scribner's Sons, 1883), plates 58 to 61.
- Herman Carl Weber, *Presbyterian Statistics through One Hundred Years, 1826-1926* (Philadelphia: Presbyterian Church in the U.S.A., 1927), part II.
- Robert Baird, *Religion in America; Or, An Account of the Origin, Relation to the State, and Present Condition of the Evangelical Churches in the United States. With Notices of the Unevangelical Denominations* (New York: Harper & Brothers, 1856). Read one of the brief chapters on an evangelical denomination (book 6, starting at page 438) and one of the chapters on the "unevangelical" denominations (book 7, starting at page 540).

In class: HTML, CSS, HTML 5 Boilerplate.

Week 2 (Jan. 30): Data used by historical actors

Assignment: Find visualizations, data tables, datasets, or corpora from the nineteenth-century United States. 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.

Read:

- Schulten, *Mapping the Nation*. Also examine the companion website.
- Shari Rabin, " 'Let Us Endeavor to Count Them Up': Statistics and American Judaism in the Nineteenth Century" (presented at the Association for Jewish Studies, Boston, MA, 2013).
- Rodney Stark, "The Reliability of Historical United States Census Data on Religion," *Sociology of Religion* 53, no. 1 (1992): 91–95, doi:10.2307/3711631.
- "Colleges, Universities and Churches" in Robert K. Nelson, Scott Nesbit, et al., *Atlas of the Historical Geography of the United States* (Digital Scholarship Lab, University of Richmond, 2016).

In class: More HTML and CSS; Bootstrap; web servers.

Week 3 (Feb. 6): Data used by historians

Assignment: Find at least one instance of how data analysis has been used by historians, either from your own area of specialization or from nineteenth-century U.S. religion. Post a scan or photo of at least one visualization or data table and write one paragraph explaining how the historian made use of data.

Read:

- Roger Finke and Rodney Stark, *The Churching of America, 1776-2005: Winners and Losers in Our Religious Economy* (Rutgers University Press, 2005), ch. 1, 3, 5.

- 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.
- Laurie F. Maffly-Kipp, "Putting Religion on the Map." *The Journal of American History* 94, no. 2 (2007): 522–29. doi:10.2307/25094964.
- Frederick W. Gibbs, "New Forms of History: Critiquing Data and Its Representations," *The American Historian* (February 2016).
- William G. Thomas II, "Computing and the Historical Imagination," in *A Companion to Digital Humanities*, ed. Susan Schreibman, Ray Siemens, John Unsworth (Blackwell, 2004).

In class: Basics of R and R Markdown; basics of Git and GitHub.

Week 4 (Feb. 13): Basics of R

Assignment: Getting familiar with R worksheet.

Read:

- Wickham and Grolemund, *R for Data Science*, ch. 1, 4, 6.
- Meirelles, *Design for Information*, ch. 1.
- Shawn Graham, Ian Milligan, and Scott Weingart, *Exploring Big Historical Data: The Historian's Macroscope* (Imperial College Press, 2015), ch. 1–2.

Week 5 (Feb. 20): Visualization with the grammar of graphics

Assignment: Data structures worksheet and functions worksheet.

Read:

- Wickham and Grolemund, *R for Data Science*, ch. 3, 8.
- Meirelles, *Design for Information*, ch. 3.
- Graham, Milligan, Weingart, *Macroscope*, ch 5.
- For reference: Joanna Zhao and Jennifer Bryan, "R Graph Catalog"; Selva Prabhakaran, "Top 50 ggplot2 Visualizations".

Week 6 (Feb. 27): Data manipulation

Assignment: Basics of ggplot2 worksheet.

Read:

- Wickham and Grolemund, *R for Data Science*, ch. 5, 12, 13.
- For reference: The Quartz guide to bad data.

In class: tutorials TBD.

Week 7 (Mar. 6): Exploratory data analysis

Assignment: Basic data manipulation worksheet; more data manipulation worksheet.

Read:

- Wickham and Grolemund, *R for Data Science*, ch. 7.
- Kieran Healy and James Moody, "Data Visualization in Sociology" *Annual Review of Sociology*, 40:105–128.
- Taylor Arnold and Lauren Tilton, *Humanities Data in R* (Springer, 2015), ch. 3–4.

In class: tutorials TBD.

Week 8 (Mar. 20): Craftsmanship

Assignment:

- Exploratory data analysis worksheet.
- Transcribe historical data, chosen in consultation with me. This data should be the basis of at least one of your visualization essays in the class. The data that you transcribe will be used for the week on exploratory data analysis, and it must be amenable to manipulation with dplyr/tidyr and visualization with ggplot2.

Read:

- David Mimno, "Data Carpentry," August 19, 2014.
- Wickham and Grolemund, *R for Data Science*, ch. 27, 29.4, 30,
- Skim: Wickham and Grolemund, *R for Data Science*, ch. 14, 19.
- Jenny Bryan et al, *Happy Git and GitHub for the useR*.
- Greg Wilson et al., "Good Enough Practices in Scientific Computing," *arXiv:1609.00037*, August 31, 2016, http://arxiv.org/abs/1609.00037.
- Ben Marwick, Carl Boettiger, and Lincoln Mullen, "Packaging Data Analytical Work Reproducibly using R (and Friends)."

In class: tutorials TBD.

Week 9 (Mar. 27): Mapping

Assignment: Exploratory data analysis visualization essay (New Nation Votes dataset).

Read:

- Meirelles, *Design for Information*, ch. 4.
- Arnold and Tilton, *Humanities Data in R*, ch. 7.
- Stephen Robertson, "Putting Harlem on the Map," in *Writing History in the Digital Age*, edited by Jack Dougherty and Kristen Nawrotzki (University of Michigan Press, 2012).
- Richard White "What is Spatial History?"
- Todd Samuel Presner, David Shepard, and Yoh Kawano, *HyperCities: Thick Mapping in the Digital Humanities*, MetaLABprojects (Harvard University Press, 2014), 110–27.
- For reference: Spatial Humanities Workshop.
- leaflet package documentation.
- sf package documentation.

In class: tutorials TBD.

Week 10 (Apr. 3): Mapping

Assignment: Basic mapping worksheet.

Read:

- Meirelles, *Design for Information*, ch. 5.
- Cameron Blevins, "Space, Nation, and the Triumph of Region: A View of the World from Houston," *Journal of American History* 101, no. 1 (June 1, 2014): 122–47, doi:10.1093/jahist/jau184.
- Geocoding functions in ggmap package documenation.
- cleanNLP package documentation.

In class: tutorials TBD.

Week 11 (Apr. 10): Text analysis

Assignment: Mapping visualization essay (WPA former slave narratives dataset).

Read:

- Matthew K. Gold et al., "Forum: Text Analysis at Scale," in *Debates in the Digital Humanities 2016* (University of Minnesota Press, 2016), 525–568.
- Silge and Robinson, *Tidy Text Mining with R*, ch. 1–2, 4–7.
- Graham, Milligan, Weingart, *Macroscope*, chs. 3–4.
- Meirelles, *Design for Information*, ch. 6.
- Arnold and Tilton, *Humanities Data in R*, ch. 9–10.

In class: tutorials TBD.

Week 12 (Apr. 17): Text analysis

Assignment: Text analysis worksheet.

Read:

- Ryan Cordell, "Reprinting, Circulation, and the Network Author in Antebellum Newspapers," *American Literary History* 27, no. 3 (2015): 417–445, doi:10.1093/alh/ajv028.
- David A. Smith, Ryan Cordell, and Abby Mullen, "Computational Methods for Uncovering Reprinted Texts in Antebellum Newspapers," *American Literary History* 27, no. 3 (2015): E1–E15, doi:10.1093/alh/ajv029.
- textreuse package documentation.

In class: tutorials TBD.

Week 13 (Apr. 24): Network analysis

Assignment: Text analysis visualization essay (American Tract Society dataset; WPA former slave narratives dataset).

Read:

- Graham, Milligan, Weingart, *Macroscope*, ch. 6–7.
- Meirelles, *Design for Information*, ch. 2.
- Arnold and Tilton, *Humanities Data in R*, ch. 6.
- igraph package documentation.
- ggraph package documentation.

In class: tutorials TBD.

Week 14 (May 1): Conclusion

Assignment: Create a hand-coded portfolio of your work for this class and publish it to the web.

Readings: TBD.

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 (including participation assignments) in order to pass this course. Your attendance is expected at every meeting. If you must be absent, I request that you notify me in advance of the class meeting. I am sometimes willing to grant extensions for cause, but you must request an extension before the assignment's due date. For every day or part of a day that an assignment is late without an extension, I may reduce your grade. No work (other than final exams and final projects) will be accepted later than the last day that the class meets unless I have agreed otherwise. I will discuss grades only in person during office hours.

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 at 703-993-2474 or through their website. You are responsible for verifying your enrollment status. All academic accommodations must be arranged through that office. Please note these dates from the academic calendar.

- Last day to add a class: January 30, 2017.
- Last day to drop a class without penalty: January 30, 2017.
- Last day to drop a class without special permission: February 24, 2017.