Humanities Data Analysis
Professor Benjamin Schmidt
benschmidt.org/HDA19
Email: b.schmidt@northeastern.edu
Office Hours: TBA

Overview

Data analysis in the humanities presents challenges of scale, interpretation, and communication distinct from the social sciences or sciences.

This seminar will explore the emerging practices of data analysis in the digital humanities from two sides: a critical perspective aiming to be more responsible readers of cultural analytics, and a creative perspective to equip you to perform new forms of data analysis yourself.

Our goal is to make it possible to merging forms of data analysis taking place in humanities scholarship, both in terms of applying algorithms and in terms of better investigating the presuppositions and biases of the digital object. We'll aim to come out much more sophisticated in the use of computational techniques and much more informed about how others might use them.

Some of the key questions we'll aim to answer are:

1. What light can algorithmic approaches shed on live questions in humanistic scholarship?
2. How can you come to understand a new algorithm?
3. What new forms of research are enabled by the use of data?
4. What sort of data do practicing humanists want museums and libraries to make available?

A wide variety of types of data will be used but we will focus particularly on methods for analyzing *texts*. If your interests lie elsewhere, don't worry too much–as you'll learn, most of the textual approaches we'll consider are easily adaptable for (and in many cases, orginally developed for) other sources of data.

Over the course of the semester, you should work to develop your own collection; ideally it will be of texts. Working with these texts will allow us to ask more sophisticated questions on large documents of scholarly importance.

Course Goals

1. Be able to contribute to debates about the place of data analysis in the humanities from both a technical and theoretical perspective, in a way that lets you responsibly elicit "data" (or as Johanna Drucker would have it, "capta") out of more humanistic stores of knowledge.

2. Acquire proficiency in the manipulation, transformation, and graphic presentation of data in the R programming language for use in the context of exploratory data analysis and presentation.

3. Know the appropriate conditions for using, and be able to use, some of the major machine learning algorithm for data classification, clustering, and dimensionality reduction.

4. Execute projects creatively deploying and combining these methods in ways that contribute to humanistic understanding.

Requirements

This is an unconventional methods course, because we'll be looking at two very different kinds of methods; literature and code

Readings and Attendance

You must attend class each week having completed the assigned readings and ready to discuss them. Note that the schedule also includes a number of *optional* explorations of the algorithms; these are for your edification only.

Problem Sets

To help consolidate your programming abilities, there will be weekly problem sets. The first can be completed in any form you like; later ones should be mailed to me as R Markdown Documents. (This is the format I'll be sending them to you as.)

They should be completed weekly, and e-mailed to me before the start of class. Problem sets are required but ungraded–I understand that you may not be able to complete them every week.

They also

Data Exploration

Once we get our feet wet, I'll ask you to post results of data explorations online. If I were you, I'd do this on a blog or straight to social media. Also keep them coming to class. This can be on one of the large unstructured sets I provide, or another data source you work out with me in class.

Algorithmic application

After break, we'll be exploring a series of specific algorithms. Some we'll go over in depth in class; others we'll only touch on obliquely. Based on the data from class you find most interesting (or data of your own, we'll work to determine which of those algorithms may make sense as a transformation. You'll then write up a short version of the exploration for the course blog, present it briefly in class, and then revise your post in response to comments.

Final Projects

If you are taking this under the guise of a research seminar with your own materials, you should produce a multifaceted analysis with a reflective, methodological take on the data you bring to the class. This could either take the form of an explicit journal article for a digital humanities audience (I would mirror those in what used to be called *Literary and Linguistic Computing*) or a 10-20 methodological appendix to a larger work (such as a dissertation) giving the details of an analysis that may take only a few pages in a more traditional work. You'll consult with me over the semester about how best to integrate your sources with materials from class

If you are taking it as a readings course, you still should create something. As we move into the later weeks of the semester, you should figure out which of the the various data sets we've used may be particularly interesting and find a way to build out on the techniques and strategies to create something novel. Most likely, this will be an *experiment* along the lines of the "Quantitative Formalism" pamphlet we read later. In it, you will take the

fundamental advantages of a programming environemnt **combine** some of the various methods and strategies we've learned in a programming environment, or build out some new ones. Appropriate products might include a large-format print map, a set of blog posts exploring generic distances, or

Keeping up to date

This is an advanced graduate seminar; I hope that the syllabus will change in response to your own interests and readings. The last week, in particular, will be decided by vote.

This flexibility may cause problems of "versioning": what version of the syllabus should you believe? So for the record, the priority for what to do consists of:

1. Any e-mails or announcements in class.
2. The current version of the syllabus on the course web site.
3. The most recent paper copy of the syllabus handed out.

Grading

Grading in this course is thorny. As graduate students, you should be starting to get a sense of what is important to *you*; I'm not going to quiz you on the parts of books that you find interesting.

At the same time, a small seminar requires focused engagement from all students (and from the faculty member, it should go without saying). So we can't just go all loosy-goosey here.

Instead, we'll use a form of "contract grading." That means: you tell me affirmatively what you want to accomplish in this class and for what grade, and I'll contract with you to allow that.

Example Template

> I will:
>
> 1. Attend every class
> 2. Read all of the assigned texts to the point and come with established opinions about them.
> 3. Work to complete all of each distributed workset before class.
> 4. Support my peers in the class in the way that best supports their learning (which sometimes entails giving "the answer," and sometimes does not, but always entails treating them respectfully).
> 5. Work to develop my own data source, and find 4 jumping off points where I apply methods from class to it.
> 6. Read and review one book or article *not* on the syllabus and present in one week.
> 7. Produce a web-based final project looking at data intended for my dissertation, improving the jumping-off points.

You shouldn't use precisely this template! It might be too much for you. Or you might want to move in slightly a different direction. You might not have a dissertation data, say. But conversely, you can't remove the bit about treating your peers respectfully.

You can also demand additional things out of me. Generally I'll give a once-over to the problem sets–but let me know if you want detailed comments on your code.

E-mail a proposed contract to me by Wednesday in the second week of classes.

Coding and Scripting in R

This course will have you writing some code in the R language. There is an extensive debate about whether digital humanists need to learn to code which we're not going to engage in; the fact of the matter is simply that if you want to either *do* data analysis in the humanities, coding will often be the only way to realize your personal vision; and if you want to *build resources in the humanities* that *others* might want analyze, you'll need to know what sophisticated users want to do with your tools to make them work for them.

I have no expectation that anyone will come out of this a full-fledged developer. In fact, I hope that by doing some actual scripting, you'll come to see that these debates over learning to code brush over a lot of intermediate stages. We'll be focusing in particular in developing skills less in full-fledged "programming," but in "scripting." That means instructing a computer in every stage of your work flow; using a language rather than a Graphical User Interface (GUI), which may be almost all the program you've used before. This takes more time at first, but has some extraordinary advantages over working in a GUI:

1. Your work is saved and open for inspection.
2. If you want to discover an error, you can correct it without losing the work done after.
3. If you want to amend your process (analyze a hundred books instead of ten, for instance) but do the same analysis, you can alter the code only slightly.

Why R?

In this class, you will have to do some coding as well as just thinking about data analysis in the humanities. If you've never coded before, this will be frustrating from time to time. (In fact, if you've done a lot of coding before, it will *still* be frustrating!)

We'll be working entirely in the "R" language, developed specifically for statistical computing. This has three main advantages for the sort of work that historians do:

1. It is easy to download and install, though the program `RStudio`. This makes it easy to do "scripting," rather than true programming, where you can test your results step by step. It also means that R takes the least time to get from raw data to pretty plots of anything this side of Excel. RStudio also offers a number of features that make it easier to explore data interactively.

2. It has a set of packages we'll be using for data analysis. These packages, whose names you will scattered through this text, are `ggplot2`, `tidyr`, `dplyr`, and the like. These are not core R libraries, but they are widely used and offer the most intellectually coherent approach to data analysis and presentation of any computing framework in existence. That means that even if you don't use these particular tools in the future, working with them should help you develop a more coherent way of thinking about what data is from the computational side, and what you as a humanist might be able to do with it. These tools are rooted in a long line of software based on making it easy for individuals to manipulate data: read the optional source on the history of database populism to see more. The ways of thinking you get from this will serve you will in thinking about relational databases, structured data for archives, and a welter of other sources.

3. It is free: both "free as in beer," and "free as in speech," in the mantra of the Free Software Foundation. That means that it–like the rest of the peripheral tools we'll talk about–won't suddenly become inaccessible if you lose a university affiliation.

R vs. Python vs. Javascript: which is the best language for humanities computing? Different computer languages serve different purposes. If you have taken ever taken an introductory computer science course, you might have learned a different language, like python, Java, C, or Lisp.

Although computing languages are equivalent in a certain, abstract sense, they each channel you towards thinking in particular ways. So when we

Which of these languages is best? It depends, obviously, on what you want to do. If you only learn a single language, there's a strong argument that it should Python, which is a widespread, swiss-army-knife type language that can frequently run quite quickly. But python generally promotes a kind of thinking about how you can get a problem done.

What R–especially tidyverse R–does best is let you abstract back from thinking about *programming* to thinking about *data*. Exploratory data analysis which operates on a particular base class, the "data frame." We'll talk about this more in Chapter 3; but the point is that it provides a coherent, basic language for describing any data set in terms of groupings, summary statistics, and visualization.

The closest analogues to these in other languages are less elegant and less well thought out. Python has widely used tool called `pandas` for analyzing data that is fast, powerful, and effective. But it is also more challenging for beginners than it need be. If you Google problems you'll be confronted with a variety of problems.[1]

The place of pre-packaged software.

One thing you can't do in this course, though, is rely on the out-the-box approaches prevalent in many DH programs. ArcGIS or QGIS may be the best way to make maps, and Gephi the best way to do network analysis. But as this is a course in *data analysis,* I want you to think about the fundamental operations of cartography and network analysis as simply subsets of a broader field, which is hard to see from the confines. All of these things are possible in R. And unlike graphical tools, working in a language *saves your workflow*. If you make a map with laboriously poisitioned points in ArcGIS, your operations aren't open for inspection. In R, though, every step you take and every move you make can be preserved. This is called *reproducible research*, and it is among the most important contributions you can make when working collaboratively.

Schedule

I'm updating this from the 2015 syllabus–but things have changed in this field, a lot, and so will the syllabus before we start. I've also stolen a lot from Ryan Cordell's 2017 offering of this course

Notes: mostly we'll be reading articles in this course available online. A few books are required for purchase. If you have difficulty obtaining any texts, please let me know as soon as possible.

In week 1, you'll read my spiel about what humanists need to understand when they read CS. My answer is, in general–you need to know *what they did*, but not *how they did it*. I've put some CS papers in this syllabus to expand your thinking about what's possible. You should absolutely, positively, not aim to understand the process. As a rule, if you see a fancy equation in an article *not* written by a humanist, you can probably skip the whole section for the time being.

---

[1] Ten years ago, experts tended to pontificate that python was better than R because it had a small standard library, cleaner syntax, and promoted a single way to do things effectively. One of the great ironies of modern data science is that, for programming with data, the situation has almost completely reversed; the pandas library presents a

Required books

- Ramsay *Reading Machines*.
- TBD. One of Andrew Piper or Ted Underwood's new books in literary history, probably.

Recommended supplementary materials (free online)

- James *An Introduction to Statistical Learning with Applications in R.*[2]

- Jockers *Text Analysis with R for Students of Literature.*[3]

Unit 1: Defining and transforming data

Thursday, January 10

Week 1: Introduction. What is (could be) Humanities Data Analysis?

Reading

- (optionally) Debates in Digital Humanities 2016, forum on text analysis.

Practicum: regular expressions

- Software: Please come to class having installed:

    2. The programs R and Rstudio

    – Rstudio is wrapper program around the R language that we'll be using for almost every assignment – save this first week – in this class.

Problem set: Regex practice. *Note*: Regular expressions embody pretty much everything that is miserable, ugly, and inelegant about computer programming. But they're basically indispensable for actually manipulating data in the real world. So we baptize by fire!

Week 2: turning Information %>% Data

Thursday, January 17

---

[2]The algorithms we will discuss in the second half of the semester are discussed in greater length in this text. If you wish to come to more mathematical understanding, this provides a relatively gentle introduction in machine-learning terms, but with some levels of math we'll gloss over in this class, also based in the R language. All chapters are available for download, for free, from the Northeastern library; download any now that you find helpful.

[3]For those interested solely in text analysis and not census, bibliographic, or other forms of "humanities data," this may be valuable. But be aware it uses a different set of libraries and data models for visualization and analysis than the ones we are using in this class, so the code is unlikely to work immediately

Reading

- Unsworth: Knowledge Representation in Humanities Computing. http://people.brandeis.edu/~unsworth/KR/KRinHC.html
- Rosenberg, "Data before the Fact", in Gitelman *"Raw Data" Is an Oxymoron*.
- "Becoming Digital," from Daniel J. Cohen and Roy Rosenzweig, Digital History: A Guide to Gathering, Preserving, and Presenting the Past on the Web.
- Katie Rawson and Trevor Muñoz, "Against Cleaning", Curating Menus, 7 July, 2016.
- (Much more practical) Hadley Wickham, "Tidy Data," Journal of Statistical Software, 2015

Practicum

- In class: introduction to R: data types, data.frames, and manipulation; using functions. The pipe.

Problem set:

- Census data manipulation with `read_table` and `tidyr`
- Optional reach: use `tidyr` to manipulate a data set into a form for network representations.

Week 3: Exploratory Data Analysis = split %>% apply %>% combine

Thursday, January 24

A huge amount of work is just: finding interesting things to count. Often, sophisticated work can just be figuring how to count something new. Here we look a little bit at how you can, simply count something.

Reading

- Behrens "Principles and Procedures of Exploratory Data Analysis."
- Too long, too old, and too out of print to assign: but be aware of the granddaddy of them all: Tukey *Exploratory Data Analysis*.
- Logan, T., & Parman, J. (2017). The National Rise in Residential Segregation. The Journal of Economic History, 77(1), 127-170. doi:10.1017/S0022050717000079. (We're reading this for the general findings, *not* the methodology.
- Underwood, Bamman, and Lee "The Transformation of Gender in English-Language Fiction."

Problem Set: Split/Apply/Combine

- Skill: the split/apply/combine strategy.
- Summary statistics and exploratory data analysis on the census files.

Doing   Do a bit of a research to try to find some tabular data that you can bring to class about something you're interested in.

Good data for these purposes.

1. Be multivariate or have some sensible way to *make* it have multiple columns.
2. Be fairly big. (More than 100 things, for sure. More than 1,000 would be good).

Exception/special case: is there textual data you can work with?

Week 4: data %>% visualization

Thursday, January 31

- Scan http://docs.ggplot2.org/current/, paying particular attention to the first section; the different "geoms".
- Find Microsoft Excel: compare this to their different "chart types."
- selfiecity.net
- Klein "The Image of Absence."
- Drucker, "Humanities Approaches to Graphical Display" (Digital Humanities Quarterly, 2011)
- Klein and D'Ignacio, Feminist Data Visualization. (Read the short paper–browse the reviewable book at MIT press if it's still online).

Problem set: Visualization

- R package: `ggplot2`
- Multiple geoms in ggplot.

Some documentation is available at Wickham "Ggplot2."

Penumbral:

- Daston and Galison *Objectivity*., Chapter 7 (On the Sciences, today.)

Week 5: History Dept/NuLab event with "Uncivil" Hosts.

Reading postponed–keep working on the visualization task with the HathiTrust books.

Week 6: text %>% data

Thursday, February 14.

- Daniel Shore, Shakespeare's Constructicon." *Shakespeare Quarterly,* 2015.
- Andrew Piper, *Enumerations*, 2018. Introduction and Chapter 3.
- Julia Silge and David Robinson, *Text Mining with R: A Tidy Approach*. Read online at https://www.tidytextmining.com. Preface and Chapters 1-4. This is technical–think about using it with some corpus of texts of interest to *you*.
- Stephen Ramsay, "The Hermeneutics of Screwing Around," 2010.   https://libraries.uh.edu/wp-content/uploads/Ramsay-The-Hermeneutics-of-Screwing-Around.pdf.
- Reread because we didn't talk about it enough: Witmore "Text."

Week 7: data %>% the Embedding Strategy.

Thursday, February 21

Modern machine learning requires data, but it doesn't just look like an XML or TEI representation. Instead, a particular trick for turning items into strings of numbers–the **embedding strategy**–has emerged as the dominant ways for computers to represent information to themselves. So we'll talk about that strategy, and how to get things in and out of it.

- Michael Gavin, Word Space article.
- Schmidt "Stable Random Projection." (Skip the first section about the algorithm, read starting with the overview visualization of Hathi Trust.)
- Tukey, Friedman, and Fisherkeller: Introduction to Prim-9. This is on YouTube, or I have a local copy.
- Allison et al. *Quantitative Formalism*. Link
- Ryan Heuser, something, maybe just visualizations.

Problem Set.

- Vector Space Models
- Principal Components Analysis
- Cosine similarity

Supplemental readings (technical explanations):

- James *An Introduction to Statistical Learning with Applications in R.*, chapters 10-10.2.

Week 8: (image, text, data) %>% embeddings

Thursday, February 28

- Tanya Clement and Stephen McLaughlin, "Measured Applause: Toward a Cultural Analysis of Audio Collections," CA: Journal of Cultural Analytics (23 May 2016)
- Melvin Wevers, Tomas Smits. "The visual digital turn: Using neural networks to study historical images". https://doi.org/10.1093/llc/fqy085. 18 January 2019
- PixPlot, Yale DH Lab: http://dhlab.yale.edu/projects/pixplot/
- CS paper: read for a sense of possibilities, not method. The Shape of Art History in the Eyes of the Machine: Ahmed Elgammal, Marian Mazzone, Bingchen Liu, Diana Kim, Mohamed Elhoseiny. https://arxiv.org/abs/1801.07729

(For background, you could read this generally useful, slightly hyperventilating introduction to neural networks from the NY Times)

Week 8: Supervised Learning and predictive models

Thursday, March 14

- Mosteller and Wallace "Inference in an Authorship Problem."
- Lena Hettinger, Martin Becker, Isabella Reger, Fotis Jannidis, and Andreas Hotho "Genre Classification on German Novels" (2015)
- Ted Underwood, Distant reading and the blurry edges of genre"
- Ted Underwood, Understanding Genre in a Collection of a Million Volumes, Interim report
- Maybe instead Ted Underwood, *Distant Horizons*; not sure how the publication date will align here.

Problem sets:

- Naive Bayes
- Logit classifiers
- K-nearest-neighbor classification.

Methods:

- James *An Introduction to Statistical Learning with Applications in R.*, Chapter 4.
- Chapter 8, Decision trees, particularly 8.1: optional.
- Reach–Support Vector Machines, James Chapter 9. Only for the foolhardy.

Week 9: Unsupervised clustering

Thursday, March 21

- Blei, "Probabilistic Topic Models."
- Goldstone and Underwood "The Quiet Transformations of Literary Studies." and online supplement
- Ted Underwood, "Topic Modeling Made Just Simple Enough," April 7, 2012
- Rhody "Topic Modeling and Figurative Language."

Problem sets and methods:

- K-means Clustering.
- Hierarchical Clustering.
- Topic Modeling using the R-Mallet package.

Week 10: space %>% data

Thursday, March 28

- Richard White "What is Spatial History?"   http://web.stanford.edu/group/spatialhistory/cgi-bin/site/pub.php?id=29
- Wilkens "The Geographic Imagination of Civil War-Era American Fiction."
- Blevins "Space, Nation, and the Triumph of Region."
- Present: Colleen on Philip Ethington's "Placing the past: 'Groundwork' for a spatial theory of history."

Problem Set: Geographic binning and visualization.

Week 11:

Thursday, April 4

- Elliott Ash, Daniel L. Chen, Suresh Naidu. "Ideas Have Consequences: The Impact of Law and Economics on American Justice." Working paper: http://elliottash.com/wp-content/uploads/2018/08/ash-chen-naidu-2018-07-15.pdf
- Choose 4 articles from the Current Research in Digital History forum/journal. Read them, and come prepared to talk about 2 in particular. How do they work? What do they succeed at?

Week 12:

Wednesday, April 11

- Roopika Risam, *New Digital Worlds: Postcolonial Digital Humanities in Theory, Praxis, and Pedagogy* (Northwestern University Press, 2019). Chapter 5. Available online: https://www-jstor-org.ezproxy.neu.edu/stable/j.ctv7tq4hg.9
- Lauren Tilton and Taylor Arnold, "Distant Viewing," Digital Scholarship in the Humanities, 2019. https://www.distantviewing.org/pdf/distant-viewing.pdf
- Daniel Rodgers, *Age of Fracture*, Chapter 1: Losing the words of the Cold War. Available online: https://ebookcentral.proquest.com/lib/northeastern-ebooks/detail.action?docID=3300914 With this, think about all the ways we've looked at the language in State of the Union addresses, and think–as specifically as possible–about what digital-textual methods could bring to this problem. I am indebted to a variety of people for contributions to this class. Those whose syllabi I have taken readings, ideas, and (in one case) a unit title from include Andrew Goldstone, Johanna Drucker, Lev Manovich, Jason Heppler, and Ted Underwood.

In revising it, I've leaned on Ryan Cordell's 2017 offering of the course.

I also gratefully acknowledge Andrew Goldstone's contribution to the syllabus template.

#Full Citations

Allison, Sarah, Ryan Heuser, Matthew L. Jockers, Franco Moretti, and Michael Witmore. *Quantitative Formalism: An Experiment (Stanford Literary Lab, Pamphlet 1)*. Stanford: Standford Literary Lab, n.d.

Behrens, John T. "Principles and Procedures of Exploratory Data Analysis." *Psychological Methods* 2, no. 2 (1997): 131. http://psycnet.apa.org/journals/met/2/2/131/.

Blevins, C. "Space, Nation, and the Triumph of Region: A View of the World from Houston." *Journal of American History* 101, no. 1 (2014): 122–147. doi:10.1093/jahist/jau184.

Daston, Lorraine, and Peter Galison. *Objectivity*. New York: Zone Books ; Distributed by the MIT Press, 2007.

Gitelman, Lisa. *"Raw Data" Is an Oxymoron*. Infrastructures Series. Cambridge, Massachusetts: The MIT Press, 2013.

Goldstone, Andrew, and Ted Underwood. "The Quiet Transformations of Literary Studies: What Thirteen Thousand Scholars Could Tell Us." *New Literary History* 45, no. 3 (2014): 359–384. doi:10.1353/nlh.2014.0025.

James, Gareth. *An Introduction to Statistical Learning with Applications in R*, 2013. http://dx.doi.org/10.1007/978-1-4614-7138-7.

Jockers, Matt. *Text Analysis with R for Students of Literature*. Springer, 2014. http://www.springer.com/statistics/computational+statistics/book/978-3-319-03163-7.

Klein, Lauren F. "The Image of Absence: Archival Silence, Data Visualization, and James Hemings." *American Literature* 85, no. 4: 661–688. Accessed January 14, 2015. doi:10.1215/00029831-2367310.

Mosteller, Frederick, and David L. Wallace. "Inference in an Authorship Problem: A Comparative Study of Discrimination Methods Applied to the Authorship of the Disputed Federalist Papers." *Journal of the American Statistical Association* 58, no. 302 (1963): 275–309. http://www.tandfonline.com/doi/abs/10.1080/01621459.1963.10500849.

Ramsay, Stephen. *Reading Machines: Toward an Algorithmic Criticism*. Topics in the Digital Humanities. Urbana: University of Illinois Press, 2011.

Rhody, Lisa M. "Topic Modeling and Figurative Language." *Journal of Digital Humanities* 2, no. 1 (April 7, 2013). http://journalofdigitalhumanities.org/2-1/topic-modeling-and-figurative-language-by-lisa-m-rhody/.

Schmidt, Benjamin. "Stable Random Projection: Lightweight, General-Purpose Dimensionality Reduction for Digitized Libraries." *Journal of Cultural Analytics* (2018). doi:10.22148/16.025.

Tukey, John W. *Exploratory Data Analysis*. Addison-Wesley Series in Behavioral Science. Reading, Mass: Addison-Wesley Pub. Co, 1977.

Underwood, Ted, David Bamman, and Sabrina Lee. "The Transformation of Gender in English-Language Fiction." *Journal of Cultural Analytics* (2018). doi:10.22148/16.019.

Wickham, Hadley. "Ggplot2." *Wiley Interdisciplinary Reviews: Computational Statistics* 3, no. 2 (2011): 180–185. doi:10.1002/wics.147.

Wilkens, Matthew. "The Geographic Imagination of Civil War-Era American Fiction." *American Literary History* 25, no. 4: 803–840. Accessed January 15, 2015. doi:10.1093/alh/ajt045.

Witmore, Michael. "Text: A Massively Addressable Object," December 31, 2010. http://winedarksea.org/?p=926.