LITERARY DATA: SOME APPROACHES

http://www.rci.rutgers.edu/~ag978/litdata

Thursdays, 1:10–4:10 p.m. in Murray 305 (Plangere Seminar Room) Professor Andrew Goldstone (andrew.goldstone@rutgers.edu)

Office hours: Mondays, 2:00–4:00 p.m. in Murray 019

COURSE DESCRIPTION

In the last ten years, the strange quasi-disciplinary formation known as DH or $Digital\ Human-ities$ has renewed the struggle over methods in literary studies. Analyses of digitized texts using computer-assisted techniques promise to transform the kinds of evidence, the methods of interpretation, and the modes of argument which matter to literary scholarship. Data is now a subject of energetic debate in literary studies: what constitutes literary data, and how should it be analyzed and interpreted? How might aggregation and quantification produce new knowledge in literary scholarship? What methods are most appropriate for grappling with the enormous, and enormously messy, world of digitized literary texts and data about literature?

This course pursues two aims in parallel: to engage with the history and current practice of literary data analysis, and to introduce the foundational skills of literary data analysis in the R programming language. Class time will be divided between seminar and practical instruction. The seminar discussions trace theoretical debates about literary data from structuralism and scientific bibliography, to experiments in computational stylistics, to contemporary scholarly controversies in and around DH. The practicum surveys the fundamentals of programming and data manipulation, with an introduction to selected numerical techniques and data visualizations. Short homework exercises supplement the in-class instruction, with an emphasis on handling actual literary data of various kinds.

There are two major assignments. A short position paper on a theoretical question about literary data and DH is due at midterm. The final assignment is to plan, carry out, and report on a small-scale project in literary data analysis. This project is to be undertaken in small groups; the report will detail methods and interpretations together with code and data.

No special technical expertise of any kind is expected; instruction begins from first principles. However, the work of programming does require willingness to experiment, patience in the face of frustration, and the nerve to ask for help as often as needed.

Bring your own laptop to class, if you have one; loaner laptops will also be available for in-class workshops. MacOS X and Linux are the preferred operating systems for work in the course, but Windows will be accommodated as well.

LEARNING GOALS

- 1. Engage critically, through discussion and formal writing, with contemporary debates about computational methods in literary study.
- 2. Understand contemporary discussions of the digital humanities in the context of arguments about data in the human sciences in the last fifty years.
- 3. Develop basic competence in analyzing data using the R environment, including obtaining, reformatting, tabulating, and visualizing multiple forms of data.
- 4. Understand the fundamentals of computation by mastering the most important constructs in the R programming language.
- 5. Gain practical experience in analyzing literary data for the purpose of answering research questions in literary scholarship through a collaborative data-analysis research project.

This course fulfills the department's B distribution requirement.

REQUIREMENTS AND GRADING

10% SEMINAR PARTICIPATION IN CLASS

Seminar discussion: your thoughtful participation is required. I do not expect everyone's manner of thoughtful participation to be identical.

Practicum: I expect your best effort, not instant mastery. I also expect that you will seek my help or the help of classmates when you are confused or stuck.

10% PRACTICUM HOMEWORK

Programming homework sets are due each week at 6 a.m. on Thursday.

30% SHORT PAPER

A conference-length paper (7-8 pp.) on some of the theoretical questions raised in the first half of the course, engaging with several scholarly essays (at least one of which must be from the syllabus).

50% FINAL PROJECT

A collaborative project making use of the techniques discussed in the course. The ideal project identifies a research question, obtains a relevant dataset, analyzes the data, and develops a written argument on the basis of the analysis. Given the new skills the course teaches and the difficulties of data-gathering, however, the project may also consist of a proposal for such a research project

together with a pilot study representing the progress the group was able to make. Working computer code, data, and a report are submitted together. I will work closely with each group.

STUDENTS WITH DISABILITIES

All reasonable accommodation will be given to students with disabilities. Students who may require accommodation should speak with me at the start of the semester. You may also contact the Office of Disability Services (disability services.rutgers.edu; 848-445-6800).

SCHEDULE

In addition to the assigned readings from Jockers, *Text Analysis*, this schedule also notes suggested sections from Teetor, *R Cookbook*, which is useful for review. Two additional, optional texts with relevant reference material on the R language are Spector, *Data Manipulation with R* and Navarro, *Learning Statistics with R*. Finally, two more advanced texts on R available online are Wickham, *Advanced R* and Chambers, *Software for Data Analysis*; by the end of the course, the former will be approachable (the latter is exhaustive).

Many readings are available online; in digital versions of this syllabus, click a title to go to its bibliography entry, where, in most cases, you'll find URLs.

Homework is listed under the date it is due.

JANUARY 22. INTRODUCTION. WHAT IS LITERARY DATA?

Rosenberg, "Data before the Fact."

Petzold, *Code*, chap. 20.

Optional: "What Is Data in Literary Studies?"

Practicum. Introduction to programming: transforming inputs to outputs; data and data types; expressions, statements, assignment. Documenting code and literate programming. Some handson time for R and RStudio setup.

JANUARY 29. THE DATA OF THE HUMAN SCIENCES.

Lévi-Strauss, "The Structural Study of Myth."

McKenzie, "Printers of the Mind," 1−13.

Darnton, "Reading, Writing, and Publishing in Eighteenth-Century France," 214-28 and nn.

Homework. Jockers, *Text Analysis*, chap. 1. Additionally: literate-programming practice with expressions and assignment. (Teetor, *R Cookbook*, 2.1–2, 2.5–10, 2.13.)

Practicum. Boolean logic and logical indexing. String data and first text functions.

FEBRUARY 5. STYLE AND STYLISTICS.

Burrows, *Computation into Criticism*, 1–10, 34–44, 98–106.

Moretti, "Style, Inc."

Trumpener, "Paratext and Genre System."

Homework. Initial formation of research groups for final project. Jockers, *Text Analysis*, chap. 2 (text data, vectors, tabulating). Additionally: more practice with vector wrangling. (Teetor, *R Cookbook*, 5.1–3, 7.1–4, 9.2–3.)

Practicum. Conditionals and loops.

FEBRUARY 12. MORE COUNTING. NO SEMINAR MEETING.

Hourlong group tutorial meetings are on Monday the 9th at noon, Monday the 16th, at 1 p.m., and Wednesday the 18th at 10 a.m. in Murray 302.

Homework. Jockers, Text Analysis, chap. 3 (word frequency data). Additionally: control flow practice.

FEBRUARY 19. SOCIOLOGY OF FORMS.

Moretti, Graphs, Maps, Trees.

Jockers, "Metadata."

Hoover, "Quantitative Analysis and Literary Studies."

Griswold, "Number Magic in Nigeria."

Michel et al., "Quantitative Analysis of Culture Using Millions of Digitized Books."

Homework. Lists, data frames, factors. (Teetor, *R Cookbook*, introduction to chap. 5, 5.4, 5.22–25. Jockers, *Text Analysis*, 5§3.)

Practicum. Data structures wrap-up. Introduction to regular expressions.

FEBRUARY 26. THE FORMATION OF DH.

Hockey, "The History of Humanities Computing."

McCarty, "Knowing...: Modeling in Literary Studies."

McPherson, "Why Are the Digital Humanities So White?"

Kirschenbaum, "What Is 'Digital Humanities,' and Why Are They Saying Such Terrible Things about It?"

Homework. Gries, *Quantitative Corpus Linguistics with R*, §3.7. (Read and think in the "think breaks.") Additionally: regular expression practice.

Practicum. Functions. Abstraction principles: encapsulation and modularity.

MARCH 5. CLASS CANCELED.

Missed seminar guest: Natalia Cecire (University of Sussex).

Homework. Jockers, Text Analysis, chap 8 (KWIC, function definition). Function practice: abstracting what we've done so far. (Teetor, *R Cookbook*, 2.12.)

MARCH 12. READING LITERARY DATA.

Ramsay, *Reading Machines*, chap. 1 (and optionally chap. 5).

Underwood, "Theorizing Research Practices We Forgot to Theorize Twenty Years Ago."

Cecire, "Ways of Not Reading Gertrude Stein."

(Optionally: Clement, "'A thing not beginning and not ending.")

Homework. Work through vignette ("introduction", "dplyr"). (Related, but not about dplyr: Teetor, R Cookbook, 5.27–33 and all of chap. 6. Expect confusion the first time.)

Practicum. Higher-order functions. Data reshaping: tidy and untidy data; dplyr basics.

(MARCH 19. SPRING RECESS.)

(MARCH 25.) SHORT PAPER DUE.

MARCH 26. DESIGN, VISUAL METHOD, GRAMMATICAL VISUALIZATION.

Seminar guest: Francesca Giannetti, Digital Humanities Librarian.

Klein, "The Image of Absence."

Lunenfeld et al., *Digital_Humanities*, 40–45.

Manovich, "What Is Visualisation?"

Wilkinson, *The Grammar of Graphics*, chap. 2, "How to Make a Pie" (optional).

Healy and Moody, "Data Visualization in Sociology" (optional).

Homework (due March 30). Reshaping and aggregating practice; visualization introduced.. Optionally: read ahead in the sections of Wickham, ggplot2, assigned for next week.

Practicum. Visualization; descriptive statistics; their relationship.

APRIL 2. MARKUP AND HYPERTEXT.

McGann, *Radiant Textuality*, 88–97, 106–35 (optional), 137–53.

Folsom, "Database as Genre."

Download the sample TEI files and follow the browsing guide.

Homework (due April 6). Plots, rebooted. Wickham, *ggplot2*, chaps. 2–3 (qplot and the ggplot grammar). Optional: Jockers, *Text Analysis*, chaps. 6–7 (lexical variety; *hapax* richness).

Practicum. HTML and XML. XML in R.

APRIL 9. CONTENT ANALYSIS AND TOPIC MODELING.

Blei, "Probabilistic Topic Models."

Grimmer and Stewart, "Text as Data."

Goldstone and Underwood, "The Quiet Transformations of Literary Studies."

Goldstone and Underwood, "Quiet Transformations."

Homework (due April 11). Jockers, Text Analysis, chap. 10 (XML). Additionally: more XML practice; HTML and web scraping.

Practicum. What in the name of Michel Foucault is topic modeling?

APRIL 16. CONTENT ANALYSIS AND TOPIC MODELING, CONT.

Schmidt, "Words Alone."

DiMaggio, Nag, and Blei, "Exploiting Affinities."

Krippendorff, Content Analysis, 24-35.

Liu, "The Meaning of the Digital Humanities" (optional but highly recommended).

Homework (due April 16). Exploring a topic model. Additionally, read Jockers, *Text Analysis*, 13.1–8 (generating a model), but skip the exercises.

Practicum. Even more modeling possibilities (LSA to LDA and what lies beyond).

APRIL 23. NETWORK ANALYSIS AND TIE MEANING.

So and Long, "Network Analysis and the Sociology of Modernism."

DeWitt, "Advances in the Visualization of Data."

Finn, Becoming Yourself.

Optional: Houston, "Toward a Computational Analysis of Victorian Poetics."

Optional: Elson, Dames, and McKeown, "Extracting Social Networks from Literary Fiction."

Practicum. Basic network visualization in R.

APRIL 30. LITERARY GEOGRAPHY.

Wilkens, "The Geographic Imagination of Civil War-Era American Fiction."

Bourdieu, *Rules of Art* (excerpt).

Homework. Long project abstract due.

Practicum. Informal project presentations and workshop.

MAY 25. FINAL PROJECTS DUE.

READINGS

- Blei, David M. "Probabilistic Topic Models." *Communications of the ACM* 55, no. 4 (April 2012): 77–84. http://www.cs.princeton.edu/~blei/papers/Blei2012.pdf.
- Bourdieu, Pierre. *The Rules of Art: Genesis and Structure of the Literary Field.* Translated by Susan Emanuel. Cambridge: Polity Press, 1996. Excerpt on Sakai.
- Burrows, J. F. Computation into Criticism: A Study of Jane Austen's Novels and an Experiment in Method. Oxford: Clarendon, 1987. Excerpt on Sakai.
- Cecire, Natalia. "Ways of Not Reading Gertrude Stein." ELH (2014).
- Chambers, John M. *Software for Data Analysis: Programming with R.* New York: Springer, 2008. http://dx.doi.org.proxy.libraries.rutgers.edu/10.1007/978-0-387-75936-4. Not required. An advanced reference on the programming language.
- Clement, Tanya E. "'A thing not beginning and not ending': Using Digital Tools to Distant-Read Gertrude Stein's *The Making of Americans*." *Literary and Linguistic Computing* 23, no. 3 (September 2008): 361-81.
- Darnton, Robert. "Reading, Writing, and Publishing in Eighteenth-Century France: A Case Study in the Sociology of Literature." *Daedalus* 100, no. 1 (1971): 214–56. http://www.jstor.org/stable/20023999.
- DeWitt, Anne. "Advances in the Visualization of Data: The Network of Genre in the Victorian Periodical Press." *Victorian Periodicals Review* 48, no. 2 (Summer 2015): forthcoming. Available on Sakai.
- DiMaggio, Paul, Manish Nag, and David Blei. "Exploiting Affinities between Topic Modeling and the Sociological Perspective on Culture: Application to Newspaper Coverage of U.S. Government Arts Funding." *Poetics* 41, no. 6 (December 2013). http://dx.doi.org.proxy.libraries.rutgers.edu/10.1016/j.poetic.2013.08.004.
- Elson, David, Nicholas Dames, and Kathleen McKeown. "Extracting Social Networks from Literary Fiction." In *ACL 2010*, 138–47. 2010. http://www.aclweb.org/anthology/P10-1015.
- Finn, Ed. *Becoming Yourself: The Afterlife of Reception*. Pamphlets of the Stanford Literary Lab 3. September 15, 2011. http://litlab.stanford.edu/LiteraryLabPamphlet3.pdf.
- Folsom, Ed. "Database as Genre: The Epic Transformation of Archives." *PMLA* 122, no. 5 (October 2007): 1571–1579. http://www.mlajournals.org.proxy.libraries.rutgers.edu/doi/abs/10.1632/pmla.2007. 122.5.1571.
- Goldstone, Andrew, and Ted Underwood. "Quiet Transformations: A Topic Model of Literary Studies Journals." 2014. http://www.rci.rutgers.edu/~ag978/quiet. This site accompanies the authors' essay.

- Goldstone, Andrew, and Ted Underwood. "The Quiet Transformations of Literary Studies: What Thirteen Thousand Scholars Could Tell Us." *NLH* 45, no. 3 (Summer 2014): 359–84. http://muse.jhu.edu.proxy.libraries.rutgers.edu/journals/new_literary_history/vo45/45.3.goldstone.pdf.
- Gries, Stefan Thomas. *Quantitative Corpus Linguistics with R: A Practical Introduction*. New York: Routledge, 2009. Excerpt on Sakai.
- Grimmer, Justin, and Brandon M. Stewart. "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts." *Political Analysis* 21, no. 3 (Summer 2013): 267–97. http://dx.doi.org.proxy.libraries.rutgers.edu/10.1093/pan/mps028.
- Griswold, Wendy. "Number Magic in Nigeria." *Book History* 5, no. 1 (2002): 275–82. http://dx.doi.org. proxy.libraries.rutgers.edu/10.1353/bh.2002.0008.
- Healy, Kieran, and James Moody. "Data Visualization in Sociology." *Annual Review of Sociology* 40, no. 1 (2014). http://www.annualreviews.org.proxy.libraries.rutgers.edu/doi/abs/10.1146/annurev-soc-071312-145551.
- Hockey, Susan. "The History of Humanities Computing." Chap. 1 in *A Companion to Digital Humanities*, edited by Susan Schreibman, Ray Siemens, and John Unsworth. Oxford: Blackwell, 2004. http://www.digitalhumanities.org/companion/. Available online.
- Hoover, David. "Quantitative Analysis and Literary Studies." Chap. 28 in Schreibman, Siemens, and Unsworth, *A Companion to Digital Literary Studies*. Available online.
- Houston, Natalie M. "Toward a Computational Analysis of Victorian Poetics." *Victorian Studies* 56, no. 3 (Spring 2014): 498–510. http://muse.jhu.edu.proxy.libraries.rutgers.edu/journals/victorian_studies/vo56/56.3.houston.pdf.
- Jockers, Matthew L. "Metadata." Chap. 5 in *Macroanalysis: Digital Methods and Literary History*. Urbana: University of Illinois Press, 2013. Excerpt on Sakai. The rest of the book, though not required, gives a striking range of different kinds of quantitative text analyses.
- ———. Text Analysis with R for Students of Literature. New York: Springer, 2014. ISBN: 9783319031637. http://dx.doi.org.proxy.libraries.rutgers.edu/10.1007/978-3-319-03164-4. This textbook is available to you in PDF through the library. You are likely to find it easier to work from in paper. Springer sells both an ordinary hardcover and a print-on-demand softcover.
- Kirschenbaum, Matthew. "What Is 'Digital Humanities,' and Why Are They Saying Such Terrible Things about It?" *differences* 25, no. 1 (2014): 46–63. http://differences.dukejournals.org.proxy.libraries.rutgers.edu/content/25/1/46.
- Klein, Lauren F. "The Image of Absence: Archival Silence, Data Visualization, and James Hemings." *American Literature* 85, no. 4 (December 2013): 661–88. http://americanliterature.dukejournals.org.proxy.libraries.rutgers.edu/content/85/4/661.
- Krippendorff, Klaus. *Content Analysis: An Introduction to Its Methodology*. Thousand Oaks, CA: SAGE, 2013. Excerpt on Sakai.

- Lévi-Strauss, Claude. "The Structural Study of Myth." *The Journal of American Folklore* 68, no. 270 (October 1955): 428–444. http://www.jstor.org.proxy.libraries.rutgers.edu/stable/536768.
- Liu, Alan. "The Meaning of the Digital Humanities." *PMLA* 128, no. 2 (2013): 409–23. http://www.mlajournals.org/doi/abs/10.1632/pmla.2013.128.2.409.
- Lunenfeld, Peter, Anne Burdick, Johanna Drucker, Todd Presner, and Jeffrey Schnapp, eds. *Digital_Humanities*. Open access ed. Cambridge: MIT Press, 2012. http://mitpress.mit.edu/sites/default/files/titles/content/9780262018470_Open_Access_Edition.pdf.
- Manovich, Lev. "What Is Visualisation?" *Visual Studies* 26, no. 1 (2011): 36–49. http://dx.doi.org.proxy. libraries.rutgers.edu/10.1080/1472586X.2011.548488.
- McCarty, Willard. "Knowing...: Modeling in Literary Studies." Chap. 21 in Schreibman, Siemens, and Unsworth, *A Companion to Digital Literary Studies*. Available online.
- McGann, Jerome J. Radiant Textuality: Literature after the World Wide Web. New York: Palgrave, 2001. Excerpt on Sakai.
- McKenzie, D. F. "Printers of the Mind: Some Notes on Bibliographical Theories and Printing-House Practices." *Studies in Bibliography* 22 (1969): 1–75. http://www.jstor.org.proxy.libraries.rutgers.edu/stable/40371475. Only an excerpt of this long essay is assigned.
- McPherson, Tara. "Why Are the Digital Humanities So White? or Thinking the Histories of Race and Computation." In *Debates in the Digital Humanities*, open access ed., edited by Matthew K. Gold. Minneapolis: University of Minnesota Press, 2013. http://dhdebates.gc.cuny.edu/debates/text/29.
- Michel, Jean-Baptiste, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, Joseph P. Pickett, Dale Hoiberg, et al. "Quantitative Analysis of Culture Using Millions of Digitized Books." *Science* 331, no. 6014 (2011): 176–182. http://www.sciencemag.org.proxy.libraries.rutgers.edu/content/331/6014/176.
- Moretti, Franco. *Graphs, Maps, Trees: Abstract Models for Literary History*. London: Verso, 2005. ISBN: 9781844671854. Available at the bookstore.
- Navarro, Dan. *Learning Statistics with R*. http://health.adelaide.edu.au/psychology/ccs/teaching/lsr/. Not required. This statistics textbook, free online, includes a good introductory discussion of R.
- Petzold, Charles. *Code: The Hidden Language of Computer Hardware and Software*. Redmond, WA: Microsoft Press, 2000. Excerpt on Sakai.
- Ramsay, Stephen. *Reading Machines: Toward an Algorithmic Criticism*. Urbana: University of Illinois Press, 2011. Excerpt on Sakai.
- Rosenberg, Daniel. "Data before the Fact." Chap. 1 in "Raw Data" Is an Oxymoron, edited by Lisa Gitelman. Cambridge: MIT Press, 2013. Available on Sakai.

- Schmidt, Benjamin M. "Words Alone: Dismantling Topic Models in the Humanities." Journal of Digital Humanities. 2, no. 1 (Winter 2012). http://journalofdigitalhumanities.org/2-1/words-alone-by-benjamin-m-schmidt/.
- Schreibman, Susan, Ray Siemens, and John Unsworth, eds. *A Companion to Digital Literary Studies*. Hardcover. Oxford: Blackwell, 2008. http://www.digitalhumanities.org/companionDLS/. Not required, except for the chapters by Hoover and McCarty. This anthology, freely available online, gives a somewhat dated overview of literary humanities computing.
- So, Richard Jean, and Hoyt Long. "Network Analysis and the Sociology of Modernism." *boundary 2* 40, no. 2 (June 2013): 147–182. http://boundary2.dukejournals.org.proxy.libraries.rutgers.edu/content/40/2/147.
- Spector, Phil. *Data Manipulation with R*. New York: Springer, 2008. http://dx.doi.org.proxy.libraries.rutgers.edu/10.1007/978-0-387-74731-6. Not required. A further reference on R.
- Teetor, Paul. *R Cookbook*. Sebastopol, CA: O'Reilly, 2011. ISBN: 9780596809157. Strongly recommended: a clear reference with lots of examples.
- Trumpener, Katie. "Paratext and Genre System: A Response to Franco Moretti." *Critical Inquiry* 36, no. 1 (2009): 159–71. http://www.jstor.org.proxy.libraries.rutgers.edu/stable/10.1086/606126.
- Underwood, Ted. "Theorizing Research Practices We Forgot to Theorize Twenty Years Ago." *Representations* 127, no. 1 (Summer 2014): 64–72. http://www.jstor.org.proxy.libraries.rutgers.edu/stable/10. 1525/rep.2014.127.1.64.
- "What Is Data in Literary Studies?" *Arcade* (January 16–20, 2014). http://arcade.stanford.edu/content/what-data-literary-studies-1. Not required. A collection of statements from an MLA 2014 Roundtable.
- Wickham, Hadley. Advanced R. Chapman & Hall, 2014. http://adv-r.had.co.nz/. Not required.
- ——. ggplot2: Elegant Graphics for Data Analysis. New York: Springer, 2009. http://dx.doi.org.proxy. libraries.rutgers.edu/10.1007/978-0-387-98141-3.
- Wilkens, Matthew. "The Geographic Imagination of Civil War-Era American Fiction." *American Literary History* 25, no. 4 (2013): 803–840. http://muse.jhu.edu.proxy.libraries.rutgers.edu/journals/american_literary_history/vo25/25.4.wilkens.html.
- Wilkinson, Leland. *The Grammar of Graphics*. 2nd ed. New York: Springer, 2005. http://dx.doi.org.proxy.libraries.rutgers.edu/10.1007/0-387-28695-0.

ACKNOWLEDGMENTS

I have drawn on contributions by the following people, either in discussion with me or through publicly-available course syllabuses or both: David Bamman, Natalia Cecire, Anne DeWitt, Lauren Klein, John Kucich, Meredith McGill, Benjamin Schmidt, Scott Selisker, Cosma Shalizi, Ted Underwood, and Matthew Wilkens.

This syllabus is available for duplication or modification for other courses and non-commercial uses under a Creative Commons Attribution-NonCommercial 4.0 International License. Acknowledgment with attribution is requested.