Working with Data
Professor Benjamin Schmidt
benschmidt.org/WWD20
Email: bs145@nyu.edu
Office Hours: Tuesday, 10-12, 20 Cooper Sq Room 520

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* in the context of other methods. 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 of data. 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.

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?

First off: you don't *have* to do R. If you already know python and want to build on that, it is possible to do almost everything in this class using pandas for dataframe analysis and altair for visualization. At some point, I may try to rewrite the whole class to support *either*.

In this class, you will have to do some coding as wellg about *data*. Exploratory data analysis which operates on 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 languauge, 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 with pandas you'll be confronted with a variety of problems;[1] R for data science has been a little less confusing.

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.

Textbook

This course works alongside of an online textbook that I will try to keep up to date with what we're working on.

The exercises exist as part of that textbook.

In general, I tell you to read a chapter or two each week. These are going to messy, and there are some weeks they might be almost missing. It's somewhere between an outline, a textbook, and a reference. But it keeps in one place both the topics you need to cover and the worksets we'll be using, alongside code that you can use, copy, and paste.

[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

Do not expect the sections to be online before Tuesday around 5:00pm.

R Package

Since we're using R for this course, I'm putting the materials for it online as an R "package." That means a bundle of code and data that you can install to your local computer and use.

To use this, you'll need to install R and the RStudio environment that it uses. Some instructions I think should work for this are here

That package is online at https://github.com/HumanitiesDataAnalysis/HumanitiesDataAnalysis. But in general, there should be no reason to access it there; instead, you will install it within Rstudio on your machine or on any machine you work on. (The computers in Data Services at Bobst, for example, should have RStudio on them; you can just pull up this website, run the code below, and have the basic things you need to work with in this class.

The purpose of this is twofold.

1. We're going to learn how to do some basic things that it will be simpler to just have code that does again and again.
2. I'm going to distribute a lot of *data* of various sorts, and this provides a simple way for getting it on to your computer.

```
if(!require(remotes)) {install.packages("remotes")}
remotes::install_github("HumanitiesDataAnalysis/HumanitiesDataAnalysis", update = FALSE)
```

There are a few other R packages associated with this course that we'll use as we go along, by me and others. They will be installed in similar ways.

In general, you'll then be able to start working by adding the following text to the beginning of your code (make sure it's inside an R Markdown block.)

```
library(HumanitiesDataAnalysis)
library(tidyverse)
```

Remember to keep this on hand; you will probably need to rerun some of these command quite often.

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. This may be a tiny class; let me know in advance if you are going to be missing.

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. The purpose is to try.

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

Later 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 *Digital Scholarship in the Humanities* or *Cultural Analytics*) 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 environment **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 a poster proposal.

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.

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.

Grading is based on how fully you do the following:

1. Submit assignments weekly, both text exercises and more freeform ones. (50%) You do not need to complete each individual problem each week–that's going to be easier for some people than others, but you must show evidence of having attempted them.
2. Participate actively and supportively in classroom discussions. (20%)
3. Hand in a final project that builds on the skills in this class. (30%)

Schedule

Notes: mostly we'll be reading articles in this course available online. One 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 in a CS paper. 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.

Defining and transforming

Mon, Jan 24 Introductions

Due Mon, Jan 24: - "Install R on your computer." - "The programs R and Rstudio Rstudio is a wrapper program around the R language that we'll be using for almost every assignment."

Mon, Jan 31 What is (could be) Humanities Data Analysis?

Readings

- Jean-Baptiste Michel et al. "Quantitative Analysis of Culture Using Millions of Digitized Books," *Science (New York, N.Y.)* 331, no. 6014 (January 14, 2011): 176–82, https://doi.org/10.1126/science.1199644.
- Johanna Drucker "Humanities Approaches to Graphical Display" 5, no. 1 (2011), http://www.digitalhumanities.org/dhq/vol/5/1/000091/000091.html.

Online text

- Introduction
- Working in a Programming Language
- "Introduction to Data"

Due Mon, Jan 31: Choose two datasets to discuss in class that are relevant to your research interests, so far as you're able to find them.

One should be something that you can actually download, almost certainly in the form of a CSV or Excel File.

The other should be something that you know exists, but that you might not be fully able to work with yet.

For both of them, fill out the online spreadsheet. The goal here is to reduce this to a *tabular* dataset. Describe what each of the columns in this dataset would be.

Do not describe the dataset as a whole aside from the columns–see if you can capture it in the individual elements.

Due Mon, Jan 31: Try to finish the exercises for "Working in a Programming Language," installing R and Rstudio.

Mon, Feb 07 Information %>% Data

Readings

- John Unsworth "Knowledge Representation in Humanities Computing," 2001, http://www.people.virginia.edu/~jmu2m/KR/KRinHC.html.
- Daniel Rosenberg "Data Before the Fact," in *Raw Data Is an Oxymoron*, ed. Lisa Gitelman (Cambridge: MIT Press, 2013).
- (Much more practical) Hadley Wickham, "Tidy Data," Journal of Statistical Software, 2015

Online text

- Counting things
- Please report issues, and use the "Troubleshooting Guide

agenda: Class agenda

Mon, Feb 14 Data Visualization

Readings

- Jacques Bertin, Semiology of Graphics: Diagrams, Networks, Maps, selections.
- Scan http://docs.ggplot2.org/current/, paying particular attention to the first section; the different "geoms.".
- "Streetscapes: Mozart, Marx, and a Dictator." Die Zeit online, 2018-02-13. link

Online text

- Visualizing Data

Practicum for next class

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

Related texts not to read

- Lorraine Daston and Peter Galison *Objectivity* (New York; Cambridge, Mass.: Zone Books ; Distributed by the MIT Press, 2007)., Chapter 7 (On the Sciences, today.)

Due Wed, Feb 16: Counting things

Mon, Feb 21 No class: President's Day

Mon, Feb 28 Counting, grouping, and accounting for how only things that get counted count.

description: A huge amount of work is just about 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.

Readings

- Trevon D. Logan and John M. Parman "The National Rise in Residential Segregation," *The Journal of Economic History* 77, no. 1 (March 2017): 127–70, https://doi.org/10.1017/S0022050717000079. As with all econ in this class, read for the general findings, and data, *not* the methodology.
- Ted Underwood, David Bamman, and Sabrina Lee "The Transformation of Gender in English-Language Fiction," *Journal of Cultural Analytics*, 2018, https://doi.org/10.22148/16.019. (This is basically chapter 4 of the other Underwood, but I like the order in the article version better for this week.)
- *What Gets Counted Counts* From D'Ignazio and Klein, *Data Feminism*, MIT press 2020.

Online text

- Cleaning Data

Related texts not to read

- Too long, too old, and too out of print to assign: but be aware of the granddaddy of them all: John W. Tukey *Exploratory Data Analysis*, Addison-Wesley Series in Behavioral Science (Reading, Mass: Addison-Wesley Pub. Co, 1977).

Practicum for next class

- Circle back to the analysis set. Do something more with the collection of book titles.
- If you successful finished much of the last problem set:
- Take a stab at the problems for Cleaning Data and tidying data

Mon, Mar 07 Making data work together

Readings

- TBD. Topics including the Textual Encoding Initiative, Codd normal forms, and maybe Underwood.

Online text

- Combining datasets: Merges, joins, and standards.

Practicum for next class

- Combining datasets: Merges, joins, and standards.

Mon, Mar 14 No class: Spring Break

Texts, maps, and data

Mon, Mar 21 Text as Data, 1

Readings

- Gentzkow et al, Journal of Economic Literature, https://doi.org/10.1257/jel.20181020
- 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*.

practicum for next class: -"Texts as Data, exercises."

Due Fri, Mar 25: Place on the course Slack two ggplot visualizations results from a **join** between two different datasets. Try to be goofy on one and serious with the others. You may use text fields if you want.

Mon, Mar 28 Text as Data, 2

Readings

- Stephen Ramsay, "The Hermeneutics of Screwing Around," 2010. https://libraries.uh.edu/wp-content/uploads/Ramsay-The-Hermeneutics-of-Screwing-Around.pdf.
- Michael Witmore "Text: A Massively Addressable Object," December 31, 2010, `http://winedarksea.org/?p=926`.

Online text for this class session

- Chapter 9.1, The Variable Document Model
- Chapter 10.1, Three Metrics
- Chapter 9.2, Getting Data.

agenda: Class agenda

Due Fri, Apr 01: Place on the course Slack two ggplot visualizations results from a **join** between two different datasets. Try to be goofy on one and serious with the others. You may use text fields if you want.

Mon, Apr 04 Space as Data

Readings

- C. 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.
- Anbinder et al, Networks and Opportunities: A Digital History of Ireland's Great Famine Refugees in New York. American Historical Review, 2019. Be sure to spend a good amount of time in the online map as well as the printed article.
- Pleiades project site. Browse in general: and read https://pleiades.stoa.org/help/data-structure, and think about types of uncertainty/ambiguity in geovisualization.

Online text for this class

- Space as Data (complete).
- For a full reference information, see Lovelace et al, Geocomputation in R. Note that Lovelace uses the tmap package for mapping while we stick to ggplot2 with the spatial geometries function geom_sf. If you really want to make–say–a zoomable map, you may want to explore tmap on your own.

Due Wed, Apr 06: Free exercise: use some bag of words on the texts of your own choosing and explore comparisons between subsets using PMI or Dunning. These can be full-text, XML, or–if–you prefer–wordcounts for books from the HathiTrust as described in the online text. Post as images or tables to the slack channel #getting-text-files."

Mon, Apr 11 Dogs as Data

description: I think we need a little reboot, so we'll focus on dogs for a little bit. Claim a possible question in the slack as described there. It's OK if you can't fully realize what you want to do, but you *must* try something, post your questions, your broken code.

Readings

- Read the Space as Data chapter in the online text.

Due Mon, Apr 11: Download a shapefile or geojson from the Internet, read it into R, and make a map that you are confident no one has made before. Post in Slack.

Due Mon, Apr 11: Identify data/datasets you'll be working with for the rest of the class

The algorithmic toolkit for exploring humanities datasets.

Mon, Apr 18 Supervised Learning and Predictive Models

note: From this point on, the weekly readings and topics are about specific applications of algorithms to different types of problems. To this point, everything we've done has been foundational–from here on out, it's more about specific applications that you *can* do if you want, but don't necessarily need to.

Class agenda

- For the four people who didn't bring a description in class, please talk through what you're thinking about investigating.
- Discussion of Noble. * Black Feminist Technology Studies

Readings

- Safiya Noble, *Algorithms of Oppression*. This is long, and we won't be able to give the time it deserves, so try to give it just a bit..

online text: Classification.

Mon, Apr 25 Clustering, topic modeling, and unsupervised approaches

Readings

- Sarah Allison et al. "Quantitative Formalism: An Experiment (Stanford Literary Lab, Pamphlet 1)" (Stanford: Standford Literary Lab, January 15, 2011). Link

In class agenda

- Talk about Allison and about Underwood. Distinguishing clustering and classification.
- Walk through of the vector space model concept in R.
- Pointing towards how to do classification in R.
- Walk through of basic clustering strategies.

Due Mon, Apr 25: due

Due Mon, Apr 25: text

Mon, May 02 The Embedding Strategy and representation learning.

description: 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.

Readings

- Tukey, Friedman, and Fisherkeller: Introduction to Prim-9. This is on YouTube, or I have a local copy.
- Ryan Heuser, Abstraction: A Literary History
- If you want to exist in this modern world, you should know something about deep learning. Yann LeCun, Yoshua Bengio, and Geoffrey Hinton "Deep Learning," *Nature* 521, no. 7553 (May 2015): 436–44, `https://doi.org/10.1038/nature14539`. is the single central text. It's hard to understand, but we'll take a stab. Come with questions, not understanding?
- Either Michael Gavin, "Word Space," *Critical Inquiry*, 2018. Very English about words; or Tilton and Arnold, Distant Viewing: Analyzing Large Visual Corpora.
- PixPlot, Yale DH Lab: http://dhlab.yale.edu/projects/pixplot/

- Look at http://benschmidt.org/profCloud/ to see if you want to try and build one of these.

Assignment for this class

- Submit a draft of your dog article

Online text

- Let's try this again: online text on clustering.
- Vector Space Models, Principal Components Analysis, and similarity.

Mon, May 09 Going deep

Readings

- Lauren Tilton and Taylor Arnold, "Distant Viewing," Digital Scholarship in the Humanities, 2019. https://www.distantviewing.org/pdf/distant-viewing.pdf
- 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
- Bengio, Lecunn, and Hinton, "Deep Learning"

Class agenda

- General check-in
- What can Deep Learning do?
- What can Deep Learning do for you?
- What about Word Embeddings?
- What about different forms of storytelling?

Agenda Notes

Notes for Mon, Feb 07

1. Rstudio installation and debug issues. What are packages, etc.
2. Any python holdouts?
3. It's to use Jupyter instead of RStudio if you prefer; but you will to install locally, because there are too many dependencies to re- download to Google Colab each time.
4. Drucker and Michel.
5. Polar opposites, so I find it helpful to find out which one you all find more amenable.
6. The question of where data comes from. Google Ngrams.

    - 02138
    - If
    - Four-letter words

7. Issues of representation and the gift of data.
    - Wild ways of thinking about datasets.
8. Your datasets
9. A new section: Ontologies are formal languages for particular domains.
10. Categorical fields.
11. Introduction to Counting.

Notes for Mon, Mar 28

- "Collaboration?"
- Finding Texts–pushing mostly to Wednesday
    - Tokenization alternatives.
- "Discuss Shore and Ramsay–can we have fun?"
- "Discuss Witmore and free discussion of problems that can be approached as different sets of documents."

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.

I've leaned especially heavily on Ryan Cordell's 2017 offering of a version of this course at Northeastern University. Thanks also to the graduate students who took it in 2015 and 2019 at that institution.

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

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, January 15, 2011.

Blevins, C. "Space, Nation, and the Triumph of Region: A View of the World from Houston." *Journal of American History* 101, no. 1 (2014): 122–47. https://doi.org/10.1093/jahist/jau184.

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

Drucker, Johanna. "Humanities Approaches to Graphical Display" 5, no. 1 (2011). http://www.digitalhumanities.org/dhq/vol/5/1/000091/000091.html.

LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. "Deep Learning." *Nature* 521, no. 7553 (May 2015): 436–44. https://doi.org/10.1038/nature14539.

Logan, Trevon D., and John M. Parman. "The National Rise in Residential Segregation." *The Journal of Economic History* 77, no. 1 (March 2017): 127–70. https://doi.org/10.1017/S0022050717000079.

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 (New York, N.Y.)* 331, no. 6014 (January 14, 2011): 176–82. https://doi.org/10.1126/science.1199644.

Rosenberg, Daniel. "Data Before the Fact." In *Raw Data Is an Oxymoron*, edited by Lisa Gitelman. Cambridge: MIT Press, 2013.

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. https://doi.org/10.22148/16.019.

Unsworth, John. "Knowledge Representation in Humanities Computing," 2001. http://www.people.virginia.edu/~jmu2m/KR/KRinHC.html.

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