Composition Forum 37, Fall 2017

Writing through Big Data: New Challenges and Possibilities for Data-Driven Arguments

Aaron Beveridge

Abstract: As multimodal writing continues to shift and expand in the era of Big Data, writing studies must confront the new challenges and possibilities emerging from data mining, data visualization, and data-driven arguments. Often collected under the broad banner of *data literacy*, students' experiences of data visualization and data-driven arguments are far more diverse than the phrase *data literacy* suggests. Whether it is the quantitative rhetoric of "likes" in entertainment media, the mapping of social sentiment on cable news, the use of statistical predictions in political elections, or the pervasiveness of the algorithmic phrase "this is trending," data-driven arguments and their accompanying visualizations are now a prevalent form of multimodal writing. Students need to understand how to read data-driven arguments, and, of equal importance, produce such arguments themselves. In Writing through Big Data, a newly developed writing course, students confront Big Data's political and ethical concerns head-on (surveillance, privacy, and algorithmic filtering) by collecting social network data and producing their own data-driven arguments.

As multimodal writing continues to shift and expand in the era of Big Data, writing studies must confront the new challenges and possibilities emerging from data mining, data visualization, and data-driven arguments. Often these challenges and possibilities are collected under the broader banner of *data literacy*, but for students, their everyday experience of data visualization and data-driven arguments is far more diverse than the phrase *data literacy* suggests. Whether it is the quantitative rhetoric of "likes" in entertainment media, the mapping of social sentiment on cable news, the use of statistical predictions in political elections, or the pervasiveness of the algorithmic phrase "this is trending," data-driven arguments and their accompanying visualizations are now a prevalent form of multimodal writing. As the era of Big Data continues to reshape the invention, delivery, and discovery of digital content, which now includes the Internet of Things and the vast expansion of sensor-driven technologies, visualized streams of data will continue to increase the amount of automated decision making in everything from business management to refrigerators that sense when to order another carton of milk. The collecting, processing, and visualizing of data—the creation of *data-driven arguments*—should be understood as *both* statistical *and* rhetorical in nature. For classrooms centered on multimodal forms of writing, this means that students need to learn how to read data-driven arguments, and, of equal importance, how to produce such arguments themselves. As students read and produce data-driven arguments, Big Data's larger political and ethical concerns come into frame. A recent collection of articles for *Kairos*, titled "Writing in an Age of Surveillance, Privacy, & Net Neutrality," calls for interventions that address the "natural language systems that code and collect billions of posts, and tracking systems that follow our every click." As the authors explain, these writing systems "have fundamentally changed the spaces and places in which we compose, create, interact, research, and teach." As a response to such concerns, this article describes a newly developed writing course called "Writing through Big Data"{1} (WtBD). In WtBD students confront the political and ethical concerns of Big Data head-on by collecting writing data themselves and producing their own data-driven arguments.

In Writing through Big Data social network trends are the primary object of study. In the context of this course, writing is defined under its broader conception that includes text, image, audio, video and any other digital artifact used to make or remix meaning. Social network trends are popular and sometimes viral topics that circulate among social networks such as Facebook, Twitter, YouTube, Reddit, Tumblr, Instagram, Snapchat, and many others. Specifically, for this course, students investigate trends occurring within Twitter's network and visualize trend data to study the way writing moves and changes over time. Weekly workshops teach students how to collect and visualize writing data to produce multimodal, data-driven arguments about trends, and all of the work completed by students during

the semester builds toward the final project in which students deliver their semester-long trend research in an oral slideware presentation. Through the use of online discussion forums, the trending topics and digital artifacts chosen by students provide thematic content for the course, allowing collaborative invention to emerge from a combination of technical practice, visual design, and engaged discussion.

Two key reciprocal concepts undergird the rationale for Writing through Big Data. First and foremost, WtBD is a rhetoric and writing studies course where students research and write about *networked writing*. As Doug Eyman defines it in *Digital Rhetoric: Theory, Method, Practice*, networked writing functions as an "an ecology of circulation," where digital networks provide a "framework that situates digital circulation within specific ecologies and economies of production: while circulation ecologies represent the places, spaces, movements, and complex interactions of digital texts as they are produced, reproduced, exchanged, or used" (84). Networked writing has taken many various forms since the invention of the Internet, and many new forms will no doubt continue to emerge. Due to the immense growth in worldwide users, and the influence that social networks have on culture, politics, economics, journalism, and entertainment, social networks provide a valuable resource for studying networked writing. By researching and writing about trending topics and digital artifacts they find meaningful, students develop a better understanding of their own networked writing practices.

While social network trends provide a macroscopic view of large-scale patterns in networked writing, they also provide a way of understanding the networks themselves. As Sidney Dobrin explains regarding networked writing in *Postcomposition*, "writing-as-system interrelates with networks rather than suggest writing is itself an identifiable network" (181). According to Dobrin's understanding of "writing-as-system," there is an indistinguishable reciprocity between writing and system—writing is structured by the technologies, systems, and networks that enable its production and circulation, but writing also restructures those same systems and networks as it saturates and permeates them: "Network emerges from writing and depends on its saturation, its fluctuation, and its mass. Without being saturated by and within writing, the network would neither emerge nor evolve, nor would the connections between the nodes and knots serve the network to any degree of circulation" (184). With #hashtags providing an organic reference tag functionality, trends often arise from the connections, topics, and digital artifacts already available within a network, but writing also moves and changes those systems by circulating new ideas and texts and forging new connections. While complex descriptions of networked writing may seem theoretically difficult for undergraduate students to grasp, trending topics and highly-circulated digital artifacts provide a non-threatening and familiar starting point for introducing more complex descriptions of networked writing.

In addition to networked writing, the second key concept already mentioned is data literacy. Data literacy, as an offshoot of digital literacy, focuses specifically on the tools, techniques, and rhetorical practices of data-driven arguments. WtBD fosters data literacy by tuning students into the pervasive and ubiquitous uses of data visualization in academia, politics, journalism, and entertainment media as well as its rhetorical dimensions. In "Rhetorical Numbers: A Case for Quantitative Writing in the Composition Classroom," Joanna Wolfe explains that "there is a paradox in that on one hand our culture tends to represent statistical evidence as a type of 'fact' and therefore immune to the arts of rhetoric, but on the other hand we are deeply aware and suspicious of the ability of statistics to be 'cooked,' 'massaged,' 'spun,' or otherwise manipulated." In other words, Wolfe argues, "Treating numbers as inherently truthful or inherently deceptive is equally naive" (453). In WtBD, students arrive at a rhetorical awareness of data visualization by collecting data on networked writing (trends) and by forming arguments about that data through infographics and visualizations they create themselves.

For Writing through Big Data, data literacy is as much about finding productive intersections of rhetoric and statistical reasoning as it is about bridging the digital divide, increasing the availability and accessibility of data-intensive tools and methods, and putting all of this to work within topics and arguments that students find meaningful. Students spend roughly the first third of the course determining which trend they want to research throughout the rest of the semester. While trends are complex in how they emerge and how they may be observed within various networks, trends are often understood simply as popular topics or digital artifacts (such as memes or viral videos) that are written about or shared by a large number of users within a network. Trends may also arise from particular social or political events and then transform into broader issues or concepts that users continue to discuss. Often, these broader concepts become categorical tags that are used to create connections among other newer topics or events (associative connections made through the use of #hashtags or common words/phrases). For example, in the case of #blacklivesmatter, the protests in the town of Ferguson (also leading to the #ferguson tag) transformed into a much larger trend focused on police violence and institutional racism. The tag #blacklivesmatter was then later used to identify other similar acts of police violence, to motivate protests against institutional racism, and to expand a broader conversation about racism in contemporary society. Certainly, for trends like #blacklivesmatter, the amount of quantitative attention it receives within a single network like Twitter (or the extent to which it appears to be "trending") should not reductively determine its broader social impact and relevance. In WtBD, students attend to both concerns.

Because individual user feeds and individual “friend” or "follower” timelines are incapable of capturing the vast amount of trends occurring across entire social networks, the methods provided by data mining allow for broad macroscopic “readings" of trend circulation. Such an approach to writing studies combines Laurie E. Gries’ understanding of circulation studies, what she often refers to as “rhetoric in motion,” with digital humanities methods of macroanalysis, where the interaction of data-intensive methods and rhetorical theory motivate student research and production without determining its end. Indeed, because of the algorithmic filtering of trends and the way many topics may be suppressed as a result (more on this later), students are encouraged to choose topics that may *not* be trending at a high level and therefore to transform their research questions into explicitly persuasive projects: why *should* people pay attention to (or participate in) a particular trend? While a course of 15 to 30 students may have little effect on what trends within a network like Twitter, students in Writing through Big Data are instructed to not approach their projects as mere “observational” research. Rather, as Gries helps us understand, the very act of collecting data, as a way of interacting with a trending topic or digital artifact, may help to “accelerate” its circulation and “intensi[fy] its consequentiality” (345). In this sense, students learn that their research becomes part of the very networks they are studying and, even if not directly intended, contributes to changes in a network’s configuration.

While such macroscopic methods may be associated with the digital humanities and/or data science,{2} recent work in writing studies has begun to apply data mining methods to digital rhetoric and writing studies research. These methods are exemplified in Eric Detweiler’s text mining analysis of rhetoric and composition journals, in Derek N. Mueller’s word cloud visualization of CCCC chair’s addresses, in Ryan Ormizo's and Bill Hart-Davidson's use of computational rhetorical analysis to detect citation patterns in academic writing, and in my own work with the mining and visualization of text data from Wikipedia articles. In a recent edited collection titled *Rhetoric and the Digital Humanities*, Jim Ridolfo and Bill Hart-Davidson argue that many similar theoretical and methodological concerns are held by these two fields and that rhetoric and the digital humanities have been moving forward in parallel and complimentary trajectories. Writing through Big Data provides an example for how rhetorical methods and DH research practices may be combined to produce data-driven arguments about networked writing, and, in so doing, attends to the broader ethical and political concerns of data literacy.

Tools and Methods

Writing through Big Data was first developed in 2014 after the University of Florida English department, in partnership with UF's Research Computing Department and the UF Libraries, applied for (and eventually was awarded) a Level II Startup Grant from the National Endowment of the Humanities.{3} The Startup Grant was pursued in order to fund the continued development of an open source research software called MassMine—a social media data mining and archiving application that simplifies the process of collecting and analyzing social network data. As the co-creator of MassMine, my responsibilities include testing MassMine's core functionality, the development of additional data processing and analysis applications, and the creation of training materials and tutorials for interdisciplinary research.

MassMine was developed to address the lack of accessibility for social network data. Social network data is often difficult to access without paying high fees to data-resellers and cloud analytics companies. Because of its use for marketing, brand management, product development, and customer service applications, social network data is often too expensive for humanities scholars to afford. In response to this problem MassMine accesses and collects the no-cost API (application programming interface) data provided freely by social networks for software developers. However, accessing API data often requires a high level of technical knowledge and advanced programming skills. The MassMine project works to greatly reduce the technical knowledge required to collect data from APIs, enabling users to mine data from social networks without any programming knowledge. The current data sources for MassMine include Twitter, Tumblr, Wikipedia, Google Trends, and any general website (URL web scraping). All of MassMine’s code remains open and freely available on GitHub, and this allows anyone to “fork” the project and develop new features or uses. Following the open-source release of version 1.0.0 in July of 2016, in fulfillment NEH's funding and support, full documentation and tutorial videos are available at www.massmine.org.

MassMine’s core functionality archives and curates data by collecting raw JSON (JavaScript Object Notation) data from social network APIs, and then data are transformed into a structured spreadsheet (CSV file) that researchers can easily access for their individual projects. The analyses and data visualizations discussed throughout the rest of this section were developed to fit the particular type of trend research students conduct during WtBD, but many other data types and research projects are possible using data collected by MassMine. All of the analyses and data visualizations provided below are based on code snippets and tutorials available here on the MassMine website. However, because MassMine exports data in universally recognized formats (JSON, CSV), archived data will work with any analysis suite or programming language (Excel, SPSS, SAS, Matlab, Python). As more teachers/researchers use MassMine, develop and share new analysis scripts, or provide tutorials and use cases for

other data collection tools and analytics software, the research possibilities for the computational study of networked writing will likewise expand.

Social network data collection can result in large datasets, and therefore human reading must be augmented with machine reading (data mining) tools. For example, it would be possible for a student to collect, read, and summarize up to a 1000 tweets during a semester (and maybe more depending on the manual process employed), but once datasets grow to 50,000 or even 100,000 tweets in size, analyzing a corpora of tweets that large requires other methods. Personally, I prefer R for data mining projects, but there are many other tools{4} available for data mining tweet corpora. R{5} is an open source programming language designed for statistical analysis and data visualization. As of 2015, R was second only to SPSS as the top research software for data analysis, and it recently surpassed SAS to become the most-used open source statistics tool among academics.{6} R's popularity stems from its package system (the CRAN repository) that includes 8,745 different tools, libraries, and add-ons to extend R's core functionality.{7} Many of the top R packages have come from academic publications and are maintained and updated through ongoing interdisciplinary research and collaboration. For example, the TM package, which stands for "Text Mining," was originally published in the *Journal of Statistical Software* (Feinerer 2008), and in the eight years since its development TM has integrated many new functions and capabilities. In WtBD, many of the scripts and code snippets used for student analyses and data visualizations are built on top of R's TM package. Packages like TM go a long way toward improving the accessibility of programming languages for scholars and teachers who are not formally trained in programming or computational methods. Like a scientific calculator, where there is a button specifically designated for calculating the square root of a number, programming packages provide built-in functionality that reduces (and in some cases removes altogether) a significant amount of the statistical and procedural knowledge required of novice programmers.

To be clear, students do not learn R or statistical programming during WtBD—students merely download scripts and run them. With well-commented code, R's functional design simplifies the explanation of TM's basic text mining features. As Figure 1 below shows, in order to define each step of the "data scrubbing" process for students, I combine a short description with a programming function named according to its task.{8} Text "scrubbing" or "cleaning" removes those items from a dataset that are not essential to the analysis being conducted. For example, when counting the most frequent words in a corpus of tweets, words like "the" or "and" often show up as the most frequent words. Therefore, by removing or "scrubbing" these words (and other similar "stopwords") from text data, the analyses conducted by students will focus on those terms that are most relevant to their research questions. Figure 1 displays some of the basic text-scrubbing processes used with Twitter data in WtBD:

```
16
17 ## Changes all of the characters in the corpus
18 ## of Twitter texts to lower case. Allowing, for
19 ## example, the two separate character strings of
20 ## "Writer" and "writer" to be coutned as the
21 ## same word.
22 d <- tolower(d)
23
24 ## Removes stopwords from the corpus of Twitter
25 ## texts. For example, words such as "at",
26 ## "it", "the", "is", "are", etc. are removed.
27 d <- removeWords(d, stopwords(kind="en"))
28
29 ## Removes URLs from the corpus of Twitter texts.
30 d <- gsub("(http|https)([^/]+).*", " ", d)
31
```

Figure 1. Code Example for Text Mining in R

For the functions `tolower()` and `removeWords()` in Figure 1, the variable `d` (Twitter text data) is modified as described in the comments above each line of code. While the function `gsub()` isn't quite as self-explanatory as the other two functions, the comment above this line of code explains its purpose for removing URLs. In addition to the data-janitorial steps displayed in Figure 1, punctuation is removed (except for @ and #), numbers are removed, and all other non-semantic objects are removed. This reduces Twitter texts to their most basic features: semantic words, #hashtags, and @usernames.

It takes no knowledge of programming to conceptually understand what is "cleaned" or "scrubbed" from the text data, but too often the *decisions* involved in data janitorial work, and how such choices effect visualizations produced from "scrubbed" data, are overlooked or not disclosed. This needlessly reduces the *rhetorical* dimensions of data-driven arguments to visual color choices for graphs and plots, proportional differences among charted objects, and the designs of surface features for data visualizations. Having students read through the script comments provides not only the opportunity to discuss the application of text mining methods for their own research projects, but it also

enables broader discussions regarding cloud data visualization tools and dynamic online graphs/plots. Rarely do such tools disclose the methods involved in cleaning and processing data. Using a tool like R demystifies this aspect of data-driven arguments.{9}

While substantial attention is paid to the texts of the Tweets themselves, through basic quantitative analyses such as word frequency counts (determining which words, #hashtags, and @usernames appeared the most in a corpus of Twitter texts), non-text data is also useful for trend research. A time series analysis can investigate *when* a trend was *more or less active*—to identify specific periods of time for more detailed analysis.{10} While a single list of the most frequent words for an entire trend dataset may be useful for general descriptions of a trend, this list on its own cannot describe the changes that occurred in the trend throughout the data collection. Further analyses are needed to describe the other various words, #hashtags, @usernames, or associated trends that participate in a trend's overall trajectory. Students compare among relative word frequencies for individual days in their dataset (or hours/weeks, depending on their research project), comparing word frequencies from individual days to the overall most frequent words. If a time series analysis reveals certain days with much higher activity, the most frequent words from these days may be considered as well. Or an individual day may be selected for comparison because of the occurrence of a particular event that is associated with a trend. Regardless of how students decide to compare data, they are encouraged to avoid cause/effect claims for their analyses, and this is true for all data analysis conducted in WtBD. While the data provided from Twitter's API by MassMine is provided as random sample, it cannot be said to be "representative" in any way that would allow for strong claims of causality between @usernames and the tweets, trends, and topics that might be associated with them.

Students may also analyze correlations among the words, #hashtags, and @usernames within their datasets. As with frequency analysis, the correlation analysis returns a simple list of words and numbers. The words returned correlate the highest across all of the Tweets and the numbers show the strength of the correlation.

Table 1. Top Word Frequencies and Correlations with "Police"

	Word	**Freq**		**Police**	**CC**
1	#blacklivesmatter	8736	1	officer	0.34
2	black	1823	2	stachowiak	0.30
3	man	479	3	threatens	0.30
4	matter	450	4	brutality	0.29
5	police	446	5	@delotaylor	0.29
6	white	442	6	jim	0.29
7	@gloedup	415	7	former	0.28
8	just	411	8	supporter	0.27
9	people	400	9	muslims	0.23
10	flood	394	10	officers	0.23

Table 1 displays the top word frequencies for one day of Tweets collected from Twitter's Rest API.{11} Using the query "blacklivesmatter," the columns on the left side of the figure show the most frequent words, #hashtags, and @usernames for the corpus of tweets, and the columns on the right display the top 10 words correlating the highest with "police" in the dataset. The column labeled "CC" displays the correlation coefficient for each word. The TM package for R allows users to designate a correlation threshold, increasing the likelihood that correlations are not a coincidental co-occurrence. During initial exploratory analyses, a correlation coefficient of .2 is used, but if particular associations among specific terms become a focus of a student's project then a minimum of .5 is recommended.{12} Whereas relative frequency totals allow students to consider how a trend changes over time, the correlations describe the associative relationships among various words, #hashtags, and @usernames within a corpus of tweets.

Certainly, other more advanced analyses are possible (and preferable depending on the research question) such as topic modeling{13} (finding the latent topics/ideas/concepts within a corpus) and semantic analysis{14} (determining whether Tweets are positive or negative). For a graduate course exploring similar research questions to WtBD, these more advanced analyses should be considered, but for an undergraduate course, frequency and correlation provide sufficient content analysis for students' trend research.

As for the types of data visualizations produced in Writing through Big Data, ordered lists and tables are the most common, and where appropriate, simple barcharts and linecharts are also effective. Some students produce wordclouds{15} as visual aids to go alongside word frequency and correlation tables. And for more complex data visualizations, other students may attempt to produce a network graph of the most highly correlated terms in their corpus, as illustrated in Figure 2.

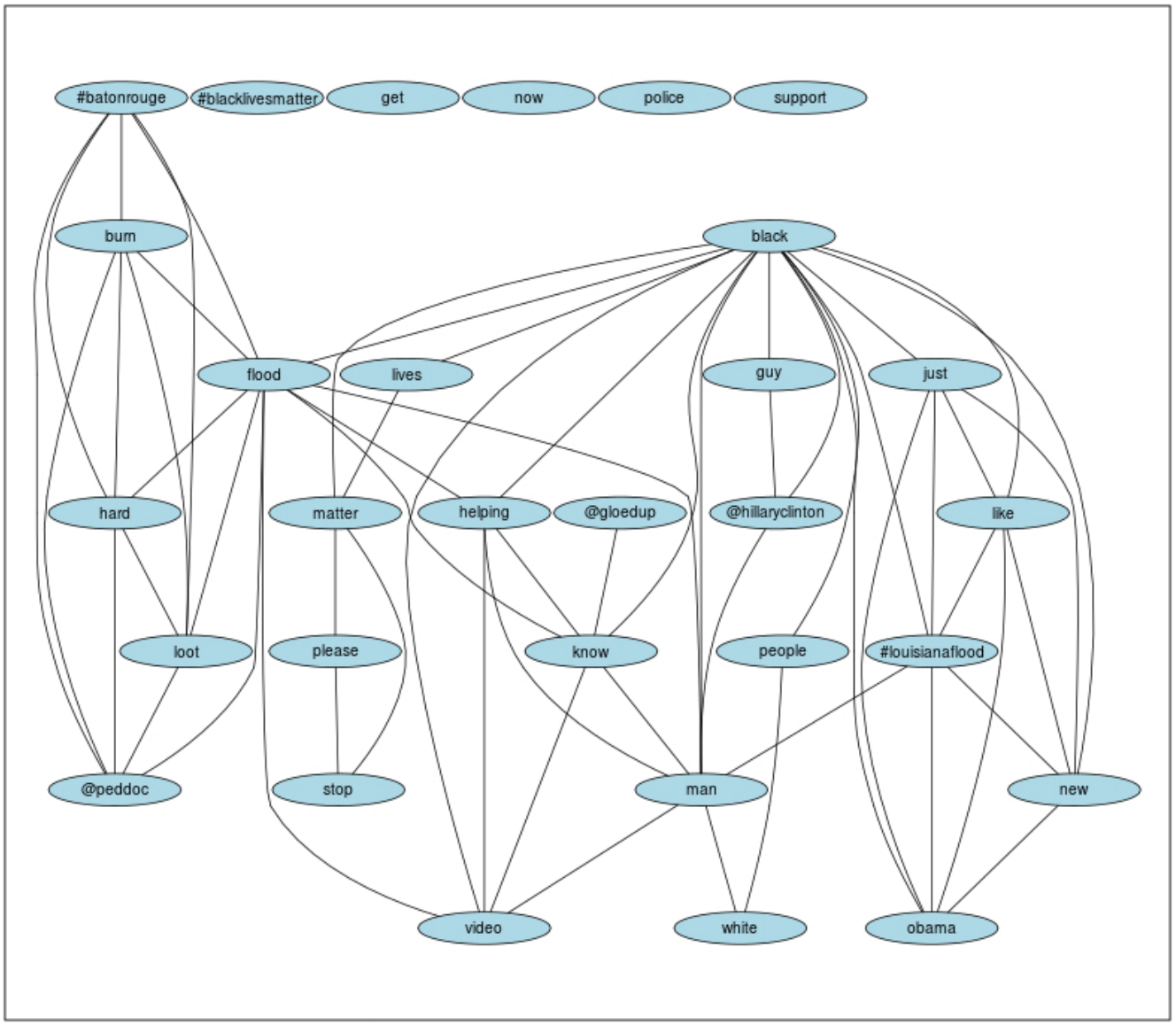

Figure 2. Network Graph of Top Correlations in #blacklivesmatter Corpus

As Figure 2 shows, using the same dataset as Table 1, students can visualize the associations among the top words, #hashtags, and @usernames in their dataset. Depending on the direction of the project, students may isolate just words and #hashtags, or they may focus on which public @usernames are most associated among the mentions{16} in their dataset. For WtBD, trend research is descriptive and exploratory; no strict hypotheses are made/tested regarding any of the trends researched. Yet student analyses of social network data are neither contrived to fit their preliminary expectations nor ignored if they disrupt students' desired claims about their trends. Like close-readings of text, rhetorical analyses of discourses and genres, materialist readings of institutions and economies, ethnographic reconstructions of communities and activities, ecological analyses of environments and networks, or posthuman descriptions of objects and technologies, descriptive and exploratory data analyses should complement writing studies research without essentializing it. To borrow a phrase from Susan Miller, this approach to data analysis

opens additional avenues for “empirical but not positivist” research of networked writing (500).

Regardless of how the tools and methods may be framed, calling the datasets collected by students “Big Data” would be an exaggeration. Most datasets were around 100k tweets,{17} and the largest in the course was 1.4 million tweets. When using R, the datasets in that range may be processed and analyzed on a standard quad-core desktop/laptop in a reasonable amount of time (depending on the various analyses conducted), and certainly do not qualify as “Big Data.” While there is no fixed size-threshold determining if a dataset counts as “Big Data,” this phrase generally refers to the tools and methods that are used when a dataset is too big for conventional methods of data processing and analysis—a relative determination. For example, if a dataset is too large to process with an individual computer, a supercomputer can be used to process the data. Because of the *potential* for large Twitter datasets collected by MassMine, students in Writing through Big Data were given free remote access to HiperGator (UF’s supercomputer) for the entire semester. MassMine was pre-installed on HiperGator, and students regularly logged into the system to collect and process their data.{18} Beyond hardware and processing considerations, the phrase “Big Data” is also relative to an individual’s access to software. For instance, 1.4 million Tweets may not be considered “Big Data” in any general comparison, but for researchers who only have access to Excel for data analysis, 1.4 million Tweets certainly counts as “Big” (Excel’s current limit is 1.04 million rows of data). Of course, the phrase “Big Data” has taken on meanings far broader than mere comparisons of datasets, hardware, and software/methods. The phrase “Big Data” is often used to refer to our current technological era, where data literacy and the ethics of surveillance and privacy require increased attention and criticism. In Writing through Big Data, the tools and methods deployed provide a pragmatic starting point for engaging these broader concerns.

Student Research and Data-Driven Arguments

One of the first activities students complete in WtBD is collecting an arbitrary dataset from Twitter with MassMine. Using a generic search query like “love,” students collect 100 tweets from Twitter's Rest API. Beyond the text data from the 100 tweets, students are usually shocked to see the additional data available on each of the public Twitter users who wrote the tweets: location (city, state), personal descriptions (sometimes including demographic data such as age, gender, race), profile pictures, the total number of “favorites” and “retweets” user accounts have received, the total number of posts by each user, and many other types of data associated with either the tweet or the user who wrote the tweet. While geolocation data is usually sparse, for a few of the tweets students can actually look up the exact location where the tweet was posted. Beyond introducing students to the data they collect throughout the semester, this first activity facilitates a broader discussion of surveillance and privacy rights. I ask the students: *Would you feel differently about your own social network postings if you knew they were analyzed outside of your own peer/friend networks? Do you have an expectation of privacy from outside analysis? Have you ever read the terms of service from your preferred social network?* This introductory activity exemplifies how the pragmatics of networked writing research opens the classroom to broader ethical discussions.

Once students have a basic understanding of the data, they decide which Twitter trend to research throughout the semester. Students are provided an introductory set of research questions to motivate their work for the course: *What is the exigence of your trend? Why are people paying attention to the trend (or why should they pay attention)? Why does the trend matter (or why should it matter)? What may have caused the trend to gain such a following? Why does it appear to have such momentum? Will this momentum last?* When students are in the trend discovery phase of the course, they are encouraged to look outside of Twitter. Facebook's trend lists may be useful. Other social networks such as Reddit, Instagram, Tumblr, or Snapchat may help students discover a trend they are interested in researching. News aggregation sites may also provide topics or events of interest.

Once a trend is identified, students go to Twitter's search page to determine which combination of terms or which #hashtag returns the most data for their trend. Because Twitter now makes all of their data available through https://twitter.com/search-home, scrolling down the page of returned search results quickly reveals how active a trend has been. If it takes a while to scroll to the previous day or week of tweets (identified with the time stamps on tweets), then a trend remains active. However, if scrolling down through the search results quickly moves to tweets from prior months and years, then a trend is less active. Students may search for multiple variations of a trend before finding the most effective query for starting their semester-long research. Even though students begin collecting data with MassMine as soon as they identify a trend of interest, they continue to conduct manual searches through Twitter's search page as well, looking for example Tweets of interest, news articles about their trend circulating within Twitter, or other digital artifacts such as images or videos that may help to describe their trend for their final presentation. As they begin to piece together the “story” of their trend, they design a single page infographic to visually display their trend's exigence. Using one of the many available cloud tools,{19} students are encouraged to find a template or design they like and then modify it for the information they want to convey.

Students' weekly discussion postings on the course's online forum act as a collaborative archive for various articles and artifacts related to their individual research projects. In addition to posting about some additional aspect or change in their own trend, students produce substantive responses to two of their classmates' posts every week. A substantive response may suggest other associated trends or topics of interest related to their classmates' research, they may share additional articles or artifacts, or they may provide helpful feedback. The goal for the weekly discussions is to collect regular non-MassMine observations and work toward possible responses to the question that frames the first half of the semester: *What is the exigence of your trend?* Certainly, during the first couple of weeks of collaborative archiving, students may change which trend they are researching for the course, but the goal of the research paper due mid-way through the semester is to push students toward a "final" decision. If the weekly discussion postings provide an informal way to think about their trend research collaboratively, then the mid-semester research paper and infographic design project provide the opportunity to carefully assemble the exigence for their chosen trend, including possible sub-trends and counter trends that may be circulating along with their main trend.{20}

For the final project, students produce a slideware presentation similar to an academic conference presentation. Their presentations included example Tweets and common images or videos from their trend to "tell the story" or help "introduce" their research project. Following a brief introduction, students explain what they initially described as the exigence for their trend, using the infographic they created earlier in the semester. While their initial expectations or ideas about trend exigence may be similar in some ways to a hypothesis—providing something to think about, "test," or question when analyzing their data—these initial expectations/ideas are too broad to be tested or falsified through traditional statistical methods. Trend exigence is useful for explaining students' interests in the trends they choose to research—for explaining *why* they researched a particular trend—and for connecting their trend to broader issues, events, and contexts. The benefit of exploratory analysis is that it allows a broad view of associations and connections among descriptions of data, but the limitation of this approach is that nothing is "proved" or "validated" through such descriptions. After explaining trend exigence, students' presentations display the data visualizations they produce from their data collections. If a student's trend dataset ended up sparse or relatively uninteresting, then the visuals produced for their final project provide additional descriptions of the overall trend. For students with consistent/large datasets, the more advanced analyses and visualizations are added to their presentations, allowing them to identify key days/times in their trend dataset, and correlations between @username mentions, #hashtags, and words in their trend. After describing their data, students conclude by reflecting on the differences between their initial expectations/ideas about their trend and what they may have learned about their trend from the data analysis.

While most students did well on the singular project-focused approach to WtBD, a limitation to such a course design was that some students chose trends with less activity (trends that were receiving less attention on Twitter), and other students chose trends that eventually became (or were revealed as) less interesting to them as the semester worked toward the end. For example, one of the research projects focused on the Ebola virus, and the student was initially interested in the way the transmission and viral spread of Ebola were discussed on Twitter. However, toward the middle of the semester most of the conversation on Twitter shifted to the political and financial issues regarding vaccine deployment in Africa. The distinct possibility of trends changing in unexpected ways provides an excellent context for discussing positive conceptions of "failure" for academic research. Rather than working to "confirm" their ideas about the trends they research, the data students collect about trends force many students to confront their projects in ways they had not originally intended. As a result, some of the best presentations at the end of the semester turn out to be "failure" projects. For other presentations, students simply conclude that nothing interesting or unexpected changed with their trend, and those projects become informative presentations where the data provides additional descriptions about the trend as a whole.

One of the more effective "failure" presentations researched the "#booty" trend on Twitter. The student's initial description of the exigence for the "#booty" trend was that it represented a move toward "healthy" or "more realistic" body representations for women in popular culture. While the student was able to find plenty of individual tweets that supported her initial assumption, and while there were recent articles from news outlets and entertainment media sources motivating her claim, the ten week data collection showed word frequency totals and word correlations that portrayed a largely misogynistic and sexually objectified approach to female bodies in her dataset. The difference between the student's initial expectations and the findings from the data collection in this project provided the basis for an effective presentation about feminism and body image. Prior to presenting the table of top word frequencies on the screen for her final presentation, this student began by warning her fellow classmates that the most frequent words associated with her trend were offensive. However, during the presentation she argued that the appearance of the words should not be offensive in and of themselves, but rather, she insisted, the words should shock her audience into realizing how much work remains to change this trend. Like this project, many of the presentations "failed" to align with students' initial expectations or assumptions. By having the data play a critical role in affecting student assumptions and mediating their "reading" of a trend's exigence, it allowed me to act as a facilitator and collaborator as students created their archives, designed their infographics, collected and analyzed data, and

produced data visualizations for their final presentations.

Critical Reflection

Because a significant amount of class time and focus was spent learning MassMine, researching and writing about students' individual trends, and producing their final presentations, the issues surrounding the algorithmic filtering of trends, privacy rights, and the surveillance economy were not a significant aspect of the course's first iteration. Initially, I approached these issues with a lecture and classroom discussion to engage with students on the issues of algorithmic filtering and privacy rights, but in the future I will substantially increase the time and attention given to these issues. Any course that intersects data literacy with social network data must question the quantitative assumptions underlying the phrases "this is trending" and "this is viral." In an article that was published after the Writing through Big Data course was first taught, Eunsong Kim's "The Politics of Trending" addresses the issue of Twitter's algorithmic blackbox: "We don't know why something trends. The algorithm is a locked secret, a 'black box' [...] Trending visibility is granted by a closed, private corporation and their proprietary algorithms." Kim's article shows how Twitter's trend algorithm ignored important and highly tweeted social movement topics, and instead, Twitter's algorithm identified as "trending" less controversial topics that were producing lower numbers of tweets. For example, Kim shows how the trend #ExplainAMovieByItsTitle was trending according to Twitter, but the #Ferguson tag (associated with #blacklivesmatter), according to Twitter at the time, was not considered a trend. Yet as Kim's analysis shows, #Ferguson was producing substantially more tweets than #ExplainAMovieByItsTitle. In future semesters, I plan to have students read Kim's article and discuss the problems raised by her quantitative comparisons of trend activity. Asking students: *If trends are not simply the objects receiving the "most" attention or the "highest" engagement, then what are they?*

More recently, Twitter has increased the amount of algorithmic filtering they do within their network—not only do they filter trends, but, following Facebook's lead, they are now filtering users' newsfeeds as well. Once networks begin filtering both feeds and trends, then they are no longer strictly *social* networks—they are reduced to *user-generated broadcast networks*. In other words, the newsfeeds and timelines for such networks do not merely show a user their friends'/peers' posts in reverse chronological order, but instead newsfeeds and timelines show users the content from their friends/peers that an algorithm determines as "relevant." Students must understand the difference between a social network and a user-generated broadcast network and come to a critical understanding of how trend feedback/manipulation occurs within a network-controlled broadcast environment. Like television broadcasts, where user attention is controlled by the channel they are tuned into, social networks have taken similar control of user newsfeeds and timelines. The content viewed within filtered networks is directed toward one primary purpose: to sell advertising. When relevance algorithms place advertising within users' newsfeeds and timelines, according to companies/products that are closely associated with the topics that users frequently "like" or engage, it appears more organic—like just another post from a friend. This is not all that different from toy companies placing television commercials within children's cartoons or a beer companies purchasing advertising during football broadcasts, but it is more refined because rather than targeting broad audiences based on programming content, social network users are targeted on an individual basis.

In future semesters, I will likely let students continue to research Twitter, if they so decide, but I may also encourage students to consider Tumblr or Reddit as alternatives and explain the differences in how writing circulates among these various networks. Not only do students need to understand the potential ways in which their networked writing may be shared, remixed, or redelivered—what Jim Ridolfo and Danielle Nicole DeVoss call *rhetorical velocity*—but students also need to know that not all networks share, remix, and redeliver writing equally. Networks are always changing based on both the users that write within those networks and based on the programs and algorithms that filter the content within those digital spaces. While the technical instruction and the data-driven production experiences from WtBD are no doubt important, I hope students finish the course with an understanding that trends do not simply emerge from networks, nor do they arise from the basic peer-to-peer sharing (or through "likes" or "favorites") of digital artifacts. Certainly, user-generated content within digital networks influences and adds to the momentum of trends within those networks, but those networks also filter and control the writing delivered within them in order to privilege sponsored content. Writing through Big Data works to provide students with both perspectives—with an awareness that user-generated content may help to create momentum for trends, but that such content may also be appropriated, filtered, or systematically ignored.

That said, it is important to note that *algorithms* should not be reductively associated with notions of "control." In general, algorithms are programs that systematically complete a task. Too often, this term carries the ethos of advanced mathematics or true Artificial Intelligence, and this reinforces the problematic black boxing of social network algorithms. Algorithms exist for a wide range of reasons, many of which are productive and useful. For example, rather than endlessly scrolling through Netflix movies chronologically or alphabetically, Netflix's algorithm

helps determine which new movies may be preferable to a user based on previous viewing choices or ratings input. Pandora's radio application relies on an algorithm to combine song metadata with sonic analyses to create custom radio stations for listeners. And even Facebook's newsfeed algorithm was initially implemented to provide users a service they needed. In 2009 when they started to filter newsfeeds based on the popularity of posts and the amount of user engagement, the goal was to help users quickly identify the most "important" posts in their newsfeed. As social network users acquire large amounts of friends/peers within a network, it becomes difficult to keep up with all of the various postings and digital artifacts shared/liked within a network. Feed-filtering algorithms serve a valuable purpose by automatically filtering the continuously-flowing feed of posts so that users can more easily find recent posts of interest.

For social networks, filtering algorithms become a problem when (1) they stop remaining optional; (2) when they are not open and accessible; (3) when the same network also filters its "trending" lists with algorithms that are likewise not open and accessible. If a social network filters its newsfeeds, then its trending lists should simply display the posts/topics/artifacts receiving the most attention and highest engagement. Or, just the opposite, if a social network filters its trends, then its newsfeed should work in a reverse chronological fashion—or, at least retain the *option* to work in this way. When both trending lists and newsfeeds are algorithmically filtered in ways that users can neither question nor understand, then there is no way of knowing how network interactions are controlled, manufactured, and manipulated.

Certainly, the many new challenges and possibilities emerging from data-mining, data visualization, and data-driven arguments cannot be confronted or explored entirely in any one course. Writing through Big Data, in its first iteration, was an attempt to balance many of the issues raised in this article by focusing students on a particular research project: producing data-driven arguments about networked writing. As I prepare to teach a new variation of the course, I am forced to reconsider the time-economy of the fifteen-week semester and decide which aspects of the course to reduce and which aspects to increase or revise. I plan to reduce the amount of technical instruction related to data collection and analysis, as students easily exceeded my expectations in this area. The first half of the course will continue to focus on questions of exigence as students create a collaborative archive, but as students transition to data collection in the second half of the course their weekly postings in the digital forum will focus on the issues surrounding algorithmic filtering, privacy rights, and the surveillance economy. In addition to reading Eunsong Kim's "The Politics of Trending," students will watch the PBS documentary called "Generation Like" and Aral Balkan's "Beyond the Camera Panopticon"—an anti-Ted Talk that explains the business model of the surveillance economy for companies like Facebook and Google. Following the viewing of these videos, I will have the students read Clay Shirky's 2009 book, *Here Comes Everyone: The Power of Organizing without Organizations*. In the seven years since it was published, Shirky's optimistic claims about social networks provide an effective contrast for students to consider. As social networks are becoming user-generated broadcast networks, the questions are as follows: *Who (or what) controls the broadcast? Who controls what users see within a social network?* As advertising and shareholder value take precedence over peer-to-peer sharing, the "Power to Organize" in many networks has been greatly reduced.

Of course, students are not powerless; new networks for sharing content and interacting digitally continue to emerge. Like changing the channel on the television to avoid the commercials (or, better yet, using a DVR to remove commercials altogether) students are free to move their writing to other systems and networks, and they need to understand this possibility as an important choice worth considering. In addition to responding to the above readings during the second half of the semester, students will attempt to participate in new networks and discuss their experiences of sharing content and interacting with peers. One example of a new social network with promising technology that ensures users privacy is the Akasha{21} social network. Based on cryptographic-network technology (block-chain based networks where data is distributed rather than located on central servers{22}), users control all of their data and can permanently delete posts without worrying that the network keeps permanent copies on central servers. There are countless new networks and technologies for content distribution emerging every day. Students will take the above readings into account as they test new networks and share their experiences with their classmates in their weekly digital posts for the second half of the semester.

Whatever *data literacy* may mean in various contexts, for WtBD the meaning of the phrase evolves throughout the semester as students complete the productive activities of the course. From collecting and analyzing social network data to designing multimodal presentations with data visualizations and infographics, data literacy is taught through data-driven research about networked writing. For the next iteration of Writing through Big Data, students will complete their research while giving contextualized attention to the broader ethical and political concerns of the Big Data era. While having students collect data and produce their own analyses extends the data literacy objectives for the course, many aspects of this course (the trend-exigence research, the infographic design project, and the ethics of algorithmic filtering and the surveillance economy) could be replicated in other courses without collecting and analyzing data. Regardless of what readers may find useful in this article, issues surrounding data literacy, as they

relate to ever-expanding notions of multimodal writing, will continue to produce new challenges and possibilities for digital rhetoric and writing studies. In Writing through Big Data, students' data-driven arguments disrupt the self-evident descriptions that current networks produce about students' data and content, and in future iterations of the course, students will work to discover new systems and networks that facilitate social interaction rather than manipulating it.

Acknowledgments: A special thanks to Sidney Dobrin, Nicholas Van Horn, Matthew Gitzendanner, and Laurie Gries for their valuable guidance and feedback in developing this course and revising this article.

Notes

1. Writing through Big Data is based on a course series developed for University of Florida by Gregory Ulmer called Writing through Media. Here's a link to the page for Ulmer's description of the series: http://www.english.ufl.edu/resources/grad/teaching/ENG_1131.html. (Return to text.)

2. For such DH approaches to literary analysis see Matthew Jockers' *Text Analysis with R for Students of Literature.* For more traditional data science approaches to text mining see "TM: Text Mining Package" for R. In WtBD, R's TM package was used to help students data mine text data collected from Twitter. (Return to text.)

3. MassMine NEH application: http://ufdc.ufl.edu/AA00025642/00001. (Return to text.)

4. See the DiRT Directory for a comprehensive index of data mining and digital research tools: https://dirtdirectory.org/. (Return to text.)

5. https://cran.r-project.org/ (Return to text.)

6. Figure 2a in the following article displays the research software used in publications from 2015: http://r4stats.com/articles/popularity/ (Return to text.)

7. https://cran.r-project.org/web/packages/ (Return to text.)

8. While the concept of "semantic" coding or software design can take on many different meanings in various programming languages and development environments, the "semantic" naming of functions and variables means using names that describe what they *do* (functions) or what they *are* (variables, data types, objects, etc.). (Return to text.)

9. For an alternative to R for text mining, see: http://voyant-tools.org/. Voyant is an open-source cloud tool developed and maintained by digital humanities scholars for text analysis. Keep in mind that the cloud tool will only be useful for exploratory datasets of less than 100K Tweets or so (based on my own personal experience with the web-app). For larger datasets, it is possible to download Voyant and use it on your local machine: https://github.com/sgsinclair/VoyantServer (Return to text.)

10. This is only possible for Twitter data collected from the Streaming API—where Twitter provides a consistent and randomized 1% of all Tweets for a particular query. The Rest API may be useful for "constructed week" *content* analyses, but not for analyses of activity that consider *how much* a topic/trend was tweeted about within Twitter's network. (Return to text.)

11. The dataset is 9,948 Tweets from 8/18/2016—8/19/2016. Using MassMine, the dataset was collected with the **twitter-search** task. Full instructions on collecting data with this task are available here: http://www.massmine.org/docs/twitter.html (Return to text.)

12. The recommended threshold of .5 comes from the examples provided in Ingo Feinerer's initial publication about the TM package: "Text Mining Infrastructure in R." (Return to text.)

13. For an overview of topic modeling, see the "Topic Modeling and Digital Humanities" special issue for *Journal of Digital Humanities*: http://journalofdigitalhumanities.org/2-1/topic-modeling-and-digital-humanities-by-david-m-blei/ (Return to text.)

14. A recent article uses R with Twitter data to analyze Donald Trump's tweets. It's called "Text analysis of Trump's tweets confirms he writes only the (angrier) half." The article provides code snippets and examples and functions both as effective data journalism and as a working tutorial for sentiment analysis. Article is

available here: http://varianceexplained.org/r/trump-tweets/ (Return to text.)

15. My article, "Looking in the Dustbin: Data Janitorial Work, Statistical Reasoning, and Information Rhetorics," provides examples and R scripts for how to produce word frequency bar charts and wordclouds with R: https://casit.bgsu.edu/cconline/fall15/beveridge/index.html. There are also many good cloud tools available for producing wordclouds—be sure to use one similar to WordItOut, where students are provided the ability to adjust *how* the text is filtered and displayed: http://worditout.com/word-cloud/make-a-new-one (Return to text.)

16. The @usernames appearing in students' Twitter text datasets are public @usernames "mentioned" or included in the text of a tweet by another user. This is one of the main reasons that cause/effect conclusions should not be drawn regarding the association of @usernames with particular terms and #hashtags. And while Twitter's API data is public/historical data and does not require IRB approval, @usernames are considered textual components for the sake of WtBD. If researching individual user accounts or groups of Twitter user accounts makes sense for more advanced writing studies courses where more formal research is conducted, privacy ethics and potential IRB approval should be considered. (Return to text.)

17. There is no official documentation on Twitter's developer page to confirm this, but it appears as if they have increased the amount of data coming from their Rest API since WtBD was first taught, which should significantly increase the student average datasets from 100k tweets to around 500k to 1 million tweets. (Return to text.)

18. A supercomputer is not necessary for using MassMine or working with Twitter data. The class as currently conceived would work just the same with students using "lab" computers or their own personal computers for data collection and analysis. (Return to text.)

19. I usually use http://www.easel.ly/ for creating infographics with students. It is free and has lots of various embedding, sharing, and exporting options. (Return to text.)

20. For example, in a trend like #blacklivesmatter, subtrends like #icantbreathe and counter trends like #bluelivesmatter may be considered when assembling a trend's broader exigence. (Return to text.)

21. http://akasha.world/ (Return to text.)

22. As the following video explains, new cryptographic network technologies may possibly change the web as we know it and enable distributed networks rather than those housed within server farms and controlled by massive technology companies. https://youtu.be/HUVmypx9HGI (Return to text.)

Works Cited

Balkan, Aral. "Beyond The Camera Panopticon" *Re:publica*, 5 May, 2015, https://www.youtube.com/watch?v=jh8supIUj6c. Accessed 15 March, 2017.

Beck, Estee N., Angela Crow, Heidi A. McKee, Colleen A. Reilly, Jennifer deWinter, Stephanie Vie, Laura Gonzales, and Danielle Nicole DeVoss. "Writing in an Age of Surveillance, Privacy, and Net Neutrality." *Kairos: A Journal of Rhetoric, Technology, and Pedagogy*, vol. 20, no. 2, 2016, http://kairos.technorhetoric.net/20.2/topoi/beck-et-al/index.html. Accessed 15 March, 2017.

Beveridge, Aaron. "Looking in the Dustbin: Data Janitorial Work, Statistical Reasoning, and Information Rhetorics." *Computers and Composition Online*, 2015, http://casit.bgsu.edu/cconline/fall15/beveridge/. Accessed 15 March, 2017.

Detweiler, Eric. "'/' 'And' '-'?: An Empirical Consideration of the Relationship Between 'Rhetoric' and 'Composition.'" *Enculturation*, vol. 20, 2015, http://enculturation.net/an-empirical-consideration. Accessed 15 March, 2017.

DeVoss, Jim Ridolfo, and Dànielle Nicole. "Composing for Recomposition: Rhetorical Velocity and Delivery." *Kairo: A Journal of Rhetoric, Technology, and Pedagogy*, vol. 13, no. 2, 2009, http://kairos.technorhetoric.net/13.2/topoi/ridolfo_devoss/. Accessed 15 March, 2017.

Dobrin, Sidney I. *Postcomposition*. SIU Press, 2011.

Eunsong, Kim. "The Politics of Trending." *Model View Culture*, vol. 18, 2015,

https://modelviewculture.com/pieces/the-politics-of-trending. Accessed 15 March, 2017.

Eyman, Douglas. *Digital Rhetoric: Theory, Method, Practice*. University of Michigan Press, 2015.

Feinerer, Ingo. "Introduction to the TM Package Text Mining in R." 2015, https://cran.r-project.org/web/packages/tm/vignettes/tm.pdf. Accessed 15 March, 2017.

Feinerer, Ingo, Kurt Hornik, and David Meyer. "Text Mining Infrastructure in R." *Journal of Statistical Software*, vol. 25, no. 5, 2008, pp. 1-54.

Gries, Laurie E. "Iconographic Tracking: A Digital Research Method for Visual Rhetoric and Circulation Studies." *Computers and Composition* vol. 30, no. 4, 2013, pp. 332-348.

Jockers, Matthew. *Text Analysis with R for Students of Literature*. Springer, 2014.

Miller, Susan. "Technologies of Self?-Formation." *JAC*, vol. 17, no. 3, 1997, pp. 497-500.

Mueller, Derek N. "Views from a Distance: A Nephological Model of the CCCC Chairs' Addresses, 1977-2011." *Kairos*, vol. 16, no. 2, 2012, http://kairos.technorhetoric.net/16.2/topoi/mueller/. Accessed 23 Oct. 2015.

Omizo, Ryan, and William Hart-Davidson. "Finding Genre Signals in Academic Writing." *Journal of Writing Research*, vol. 7, no. 3, 2016, pp. 453-483.

Ridolfo, Jim, and William Hart-Davidson, eds. *Rhetoric and the Digital Humanities*. University of Chicago Press, 2015.

Rushkoff, Douglas. "Generation Like." *Frontline/PBS*. 18 Feb. 2014.

Shirky, Clay. *Here Comes Everybody: The Power of Organizing Without Organizations*. Penguin, 2008.

Ulmer, Gregory L. "Syllabus Guidelines: ENG 1131." 29 Mar. 2010, http://www.english.ufl.edu/resources/grad/teaching/ENG_1131.html. Accessed 4 Mar. 2016.

Van Horn, Nicholas M., and Aaron Beveridge. "MassMine: Your Access to Data." *The Journal of Open Source Software*, vol. 1, no. 8, 2016. http://dx.doi.org/10.21105/joss.00050.

Wolfe, Joanna. "Rhetorical Numbers: A Case for Quantitative Writing in the Composition Classroom." *College Composition and Communication*, vol. 61, no .3, 2010, pp. 452-475.

"Writing through Big Data" from *Composition Forum* 37 (Fall 2017)

Return to *Composition Forum* 37 table of contents.

Composition Forum is published by the Association of Teachers of Advanced Composition with the support and generous financial assistance of Penn State University. *Composition Forum* ISSN: 1522-7502.