Visualization and Rhetoric: Key Concerns for Utilizing Big Data in Humanities Research

A Case Study of Vaccination Discourses: 1918-1919

Kathleen Kerr, *Graduate Research Assistant, English,* Bernice L. Hausman, *Professor, English,* Samah Gad, *Graduate Research Assistant, Computer Science, Virginia Tech*

Waqas Javen, *Lead HCI Researcher, General Electric and Global Research*

***Abstract*— Visualization of data mining results is the linchpin of successful research in the humanities that uses computational techniques. This paper describes efforts to utilize "big data" in a case study of news reporting on vaccination before, during, and after the 1918 influenza pandemic, focusing primarily on the conventions underlying methods of data extraction, data visualization practices, and the rhetorical impact of visualization design choices on researchers' observations and interpretive decisions. Purposeful attention to visualization and the methodological conventions that are embedded in particular visualization practices will allow humanists to have more confidence in their interpretations of big data, a key element in the acceptance of data mining as a valuable method for humanities research.**

Keywords-big data; data design; data mining; interpretation; rhetoric; visualization

I. INTRODUCTION

In 1918, pandemic influenza (so-called Spanish flu) took countless lives across the globe. Scholars continue to analyze the pandemic, the disease's pathogenesis, and the social, historical, and policy-related implications of the pandemic, relying largely on public health reports generated during and subsequent to the epidemic, archives of the era's newspapers, and other historical artifacts. This scholarship examines how public authorities responded to the epidemic [2, 13], changes in public health policy as a result of the disease [4, 10], the spatial dynamics of the epidemic [5], and bioethics-related issues [7]. However, as Mark Osborne Humphries points out in *The Last Plague: Spanish Influenza and the Politics of Public Health in Canada*, "most historians have taken a community case study approach, which localizes the flu's impact" [10]. In other words, there are limitations to research that relies on traditional interpretive analytics—close readings of discrete texts.

We can draw inferences about how the Spanish influenza behaved, its effects, and the efficacy of public health interventions based on anecdotal evidence from textual artifacts and case studies, but we cannot systematically explore either the qualitative features of the pandemic or the reticulate nature of information flow on a large scale. With the increasing digitization of archival texts, however, computational analytics provide new opportunities to answer lingering questions about the pandemic that close textual analysis and localized case studies do not. To adopt such a "big data" approach, we need a good methodological understanding of data mining algorithms, i.e., their modeling assumptions, as well as visualizations of the data mining results that are legible to and utilizable by the humanists trying to interpret them. Without thoughtful attention to the rhetorical impacts of various forms of visualization of the same data, the research results will continue to obscure assumptions and biases inherent in the simplifications that such methods involve [3, 11, 14].

This paper describes efforts to utilize "big data" in a case study of news reporting on vaccination before, during, and after the 1918 influenza pandemic. One aspect of our research addresses the content of vaccination-related newspaper reporting and whether and how it changed during and directly after the pandemic. The 1918 influenza pandemic occurred at an important juncture in the history of vaccine development—before it was possible to create vaccines for influenza viruses, but after some vaccinations had been developed for other diseases. As a result, vaccines were developed during the pandemic's deadly second wave, although none proved, in retrospect, to be effective. Nevertheless, there was significant reporting on vaccines during this period.

The second and more significant aspect of our research concerns the conventions that underlie both the methods of data extraction and data visualization practices. We did not set out to ask or answer any questions about visualization when we undertook this case study. Rather, these questions arose during the analysis of data mining outputs, by which point decisions relating to data mining algorithms had already been made. As a result, our aim for the study expanded to include the analysis of visualization *conventions* as they relate to data mining outputs generally—not to evaluate the effectiveness of a specific visualization of data mining outputs compared to another. Indeed, in this rhetorical analysis, we seek to better understand what visualizations do, the persuasive effects of visualization conventions, the underlying assumptions that influence or interfere with researchers' interpretations of

visualizations, and how different design choices suggest different interpretive possibilities for humanists.

We recognize that many, if not most, design elements in a visualization can be modified to meet a user's specific requirements, and this rhetorical analysis aims to uncover how these design elements operate within the overall visualization and to what ends. Some of the discussion will undoubtedly lead to improved visualizations that meet the needs of humanists more precisely, but the general thrust of the paper is to demonstrate that there are inevitable persuasive effects of any visualization choice; thus, there is a need to consider the broader rhetorical impacts of visualization itself on data mining collaborations with humanists.

The case study is part of Virginia Tech's "An Epidemiology of Information: Data Mining the 1918 Flu Pandemic," which is funded through the Digging Into Data Challenge of the National Endowment for the Humanities. "An Epidemiology of Information" applies data mining tools to digitized historical newspapers in the Library of Congress's *Chronicling America* database, addressing news reporting on the 1918 pandemic as big data. The research team includes humanists and computer scientists, who collaborated to explore the implications of various visualization conventions and design choices in the representation of data mining outputs. Recognizing that our understandings of certain terms and concepts occasionally do not translate across disciplines, we define here any terms with discipline-specific meanings that may cause confusion for readers.

II. METHODS

We have applied an integrated topic modeling and segmentation algorithm to 90 titles from January 1, 1918, to December 31, 1919, to investigate whether and to what extent computational analysis supports or challenges the findings of traditional interpretive analytics of newspaper reporting on vaccination. Topic modeling infers key distributions of words (the "topics") underlying a given corpus (e.g., 90 papers in the *Chronicling America* database). Topic modeling and segmentation works with a time-varying corpus (e.g., the 90 newspapers over a two-year period) and identifies time segments such that topics are stable within a segment but vary significantly across neighboring segments.

To define our corpus, we extracted "text chunks" from 90 papers in the *Chronicling America* database. We define a text chunk as three sentences before and after a desired search term. The search terms in this study included the root terms *vaccin* and *inoculat*, as well as *vaccination*, *vaccine*, and *inoculation*. We ran two extractions, the first blocking the terms *blackleg* (or *black leg*, both of which refer to a disease in cattle for which a vaccine had been developed) and *cholera* (which at the time was a reference to hog cholera). In the second extraction, we blocked those terms as well as *serum*. In excluding terms, we discarded any text chunks that included those terms as we reasoned they included reporting that was not relevant to our study. We labeled the outputs of the first extraction as "no blackleg/no cholera" and the outputs of the second extraction as "no unwanted terms."

The decision to exclude *serum* was based on an initial observation that it seemed overrepresented in the data output. Excluding *serum* significantly affected the content of the data outputs, however. *Serum* at the time was sometimes used as a synonym for *vaccine*, even though it refers to passive, rather than active, immunization. Inadvertently, then, a data input decision negatively affected the outputs, because in excluding the texts chunks that included *serum*, we lost some reporting on the influenza vaccine. In the "no unwanted terms" outputs, however, we see more mentions of *smallpox*, which appears to have a negative association with *serum*.

The extracted text chunks were then used as input for our integrated topic modeling and segmentation algorithm. The algorithm designates a specific number of topics identified in each time segment; we chose five in this study. Word clusters represent the identified topics and are comprised of the 20 terms most likely to be found in that cluster by frequency. The minimum window size for each segment was one week. It is important to note that this does not mean that each topic contains only 20 terms; rather, we are choosing to represent each topic with a cluster of 20 words most likely to be found in it. The algorithm's ability to dynamically identify boundaries allows us to see how the reporting on vaccination in the 90 newspapers changed over time as the clusters shift and new arrangements of words emerge.

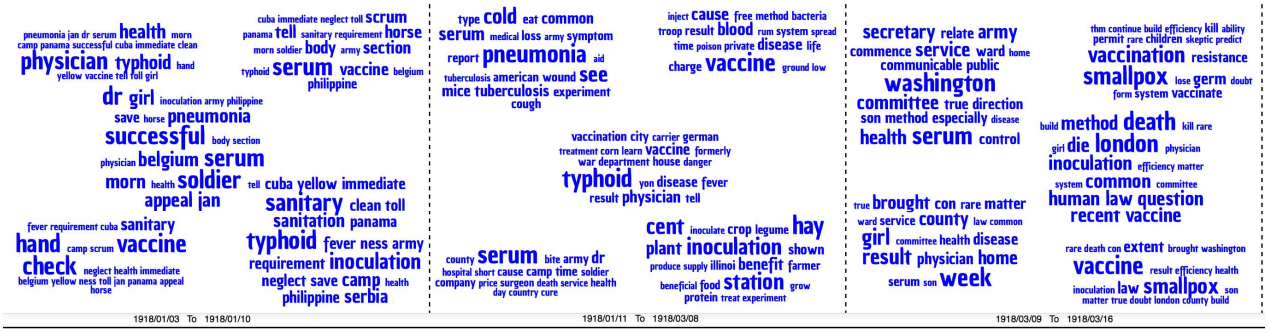

Figure1: Tag Clouds Visualization

Each cluster of words in the algorithm's output represents proximal relationships of words within the reporting and, therefore, suggests potential thematic relationships; the word clusters thus function as an index to news reporting on vaccination. Our first visualization choice is web-based and was built to represent the word clusters as tag clouds in the typical manner: the size of the words in each cloud represents their relative frequency within that cluster (see Fig. 1).

Our ability to manipulate the visualizations of the topic modeling and segmentation algorithm's output impacts how and the degree to which we are able to interpret the data. Tag clouds do not provide the user with an opportunity to interact with the visualization and modify how the data are represented. Hence, analysis is limited by design choices as well as to what the conventions of this type of data visualization can support in a fixed form.

ThemeDelta, our second visualization choice (Fig. 2), is a novel web-based visualization built specifically to represent topic modeling results over time. It uses a representation called *trendlines*, variable width lines that branch and merge. The width of each trendline renders the frequency of that term in the cluster at that particular time. The visualization is also segmented such that each vertical line represents a discovered segment and clusters are arrayed vertically in groups. Unlike tag clouds, ThemeDelta is interactive, allowing searching and the rearrangement of the data (to a certain extent).

One of the benefits of the ThemeDelta visualization is that the trendlines connecting terms from one segment to the next can demonstrate a cluster of words that stays together across segments, thereby revealing certain consistencies in the reporting across time. The researcher has to pay close attention to how the trendlines coalesce, however, since a cluster in one segment can seem to be consistent or the same as a cluster in the previous segment when, in actuality, the lines are coming together from various segments. Once this issue is attended to, it is clear that ThemeDelta may make it easier to identify recurrent clusters across segments than tag clouds. Many of the recurrent clusters are actually advertisements, which may run for several weeks in one newspaper with the same text, thereby dominating the reporting in particular segments and appearing repeatedly as the same word cluster.

Our third visualization choice is the word frequency list, a simple non-interactive visualization developed with standard word processing software (see Table 1). In a word frequency list, words in the cluster are listed in order of frequency, and the clusters in a segment are arrayed in columns. Analysts can easily customize the presentation of the data, for example by manually coding key terms by color, which can facilitate the interpretive process.

TABLE I. Word Frequency Lists: 1/10/1918 – 3/7/1918 Segment

Topic: 1	Topic: 2	Topic: 3	Topic: 4	Topic: 5
call	school	spent	street	typhoid
german	county	day	camp	vaccination
club	vaccinate	week	war	smallpox
inoculate	board	home	time	physician
cent	farm	miss	day	disease
hay	day	youngstown	committee	vaccine
jesu	red	visit	arm	vaccinate
free	health	john	ohio	health
people	color	night	special	fever
inoculation	cent	daughter	week	house
government	week	family	lie	result
kidney	smallpox	son	son	danger
propaganda	price	feb	school	city
poison	city	north	doctor	ease
life	children	guest	little	army
country	bank	church	vaccinate	american
world	jasper	call	company	dis
john	lost	parent	town	medical
house	public	school	home	tell
record	ship	entertain	receive	bad

III. Discussion

Our analyses of these visualizations of the algorithm's outputs raise important questions about how design conventions and visualization practices affect the interpretation of big data:

- How does the method of visualizing the results of topic modeling with segmentation affect interpretation of those results?
- How do forms of visualizing segmented topic models affect conventions in reading and interpretation that may positively or negatively

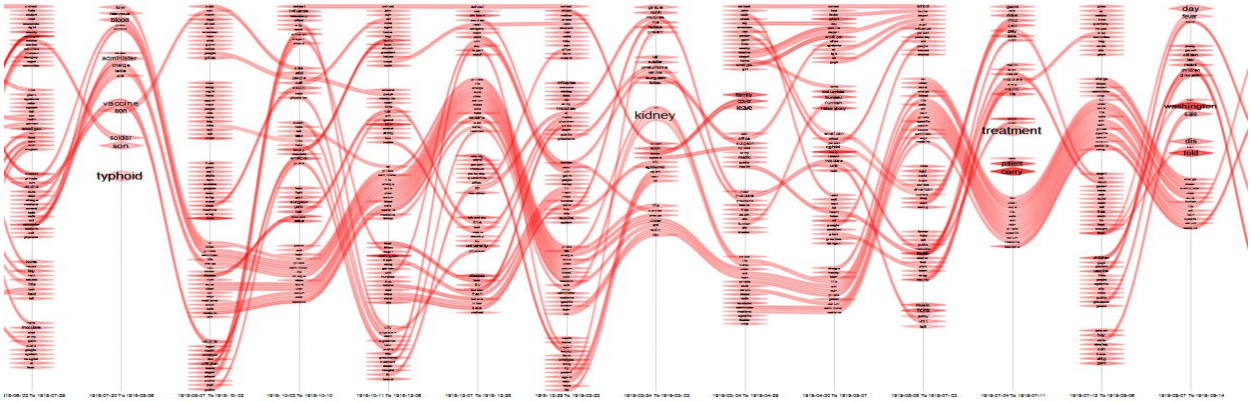

Figure2: ThemeDelta Visualization

influence research findings?

- What forms of tacit knowledge are necessary to appropriately read and interpret data mining results using various forms of visualization?

A. *Literature Review*

To help us answer these questions, we conducted a literature review of approaches to visualizations of big data in rhetorical studies. While there is a wealth of scholarship relating to data visualization, visual and digital rhetoric, and big data, there is little research that brings these disciplines together or examines the implications of big data visualization. What exists reminds us to pay attention to the conventions that underlie both the methods of data extraction and the forms of visualization. These conventions are elements of the tacit knowledge that lead to the naturalization of certain visualizing practices based on unquestioned assumptions and biases, whether or not such conventions are helpful in making sense of the data. Indeed, some design conventions may lead to significant misunderstanding of the data if the conventions organizing the visualization are not known to humanities researchers or are not readily evident to the user. All in all, the literature review, discussed in detail below, suggests that exploring the significance of visualization of data mining results is necessary to understanding how to utilize "big data" for humanities research.

Researchers such as Clark Freifeld et al., acknowledge that visualization tools—HealthMap, for example, which is an online health data information tool—must make certain assumptions in order to find the balance between flexibility and simplicity [6]. How, then, do assumptions regarding information visualization affect the design and interpretation of an artifact? Jessica Hullman and Nicholas Diakopoulos attempt to answer this question by examining the rhetorical effects of information design decisions relating to visualizations, pointing out "under-acknowledged facets of design and interpretation," including the fact that there is always, inherently, an element of simplification in visualizations [9]. Thus, a visualization can be said to operate metonymically—that is, to be understood *as* the data it represents. According to Madeleine Sorapure, another aspect of information visualization that is not always apparent is that "arguments and ideologies are embedded in particular design choices" [14], a point that Ben Barton and Marthalee Barton also emphasize when they argue that what is included—*as well as* what is excluded—in a visualization (maps, in their analysis) is a function of ideology [3].

Although Hullman and Diakopoulos's visualization rhetoric lacks a coherent theoretical underpinning, they usefully identify four editorial layers that impact meaning in information design: data, visual representation, textual annotations, and interactivity [9]. Many other scholars have articulated the complexity of meaning in the design and interpretation of visual representations. Stuart Hall, for example emphasizes the role of "conventional conceptual classifications" in the connotative process [8], while Rudolf Arnheim points out the pervasiveness of the notion that reason makes clear what is difficult to understand [1]. Arnheim's concern is echoed in Clay Spinuzzi's discussion of the tendency to place user-centered information design in opposition to system-centered design of artifacts [15]. According to Spinuzzi, system-centered design is presented as formalist and rational, while user-centered design is understood as social constructionist and post rationalist [15]. Much of the current literature on visualizations falls into one or another of these categories.

This distinction is important because the rationalist/social constructionist divide suggests divergent positions with regard to the purpose of producing visualizations. A rationalist approach suggests that the goal is to represent data credibly and accurately—the ideal is achievable with the right format. The social constructionist approach, on the other hand, suggests that the right format is negotiated in relation to the needs of both producer and user—norms are created in the context of producing the visualization, and what is needed is a user who knows those norms or can be educated in them.

Rhetoricians Charles Kostelnick and Michael Hassett address information visualization from a social constructionist perspective, cautioning that generally accepted conventions are not universals but, rather, social constructs that are situated in given social, historical, and rhetorical contexts to meet the needs both of designers and users. Their rhetorical approach understands visualization as based on a set of assumptions—whether cultural, organizational, disciplinary, or technical, etc.—that can (and do) have both positive and negative implications in terms of design and interpretation [11]. Similarly, Errol Morris states that what we see is often determined by what we believe (or think) [12].

On the other hand, Edward Tufte forwards a rationalist approach, theorizing visualizations as evidence [18] and notes in *Visual Explanations*: "When we reason about quantitative evidence, certain methods for displaying and analyzing data are better than others" [17]. From Tufte's perspective, the best methods for visualizing data produce "truthful, credible, and precise findings" [17], although like Kostelnick and Hasset [11], he notes that credibility also depends on the author and source—their quality and integrity [18]. Sorapure, however, argues that "[w]e are less likely to question the authority of data or to see the potential biases in how it was gathered, organized, and visualized" [14]. This tendency does not necessarily correlate to truthfulness, credibility, or precise findings; rather, it could be a result of the "naturalization" of certain design conventions, greater confidence in computational approaches, or the conflation of large-scale data and data integrity. That is, as a result of normalizing practices, there is often an uncritical and unquestioning acceptance of conventions such as tag clouds, word trees, certain text

forms and placements, and the size and shape of data displays.

What we learn from the literature review is that we must be attentive to the conventions that underlie both the methods of data extraction and data visualization practices. These conventions reflect commonly held assumptions, biases, and tacit knowledge about visualization practices, resulting in the tendency to naturalize certain design conventions. Because not all users have access to this tacit knowledge, such conventions do not always facilitate data interpretation. Indeed, some design conventions may even lead to significant misunderstanding of the data.

In addition, the tension between rationalist and social constructionist approaches to data display surfaces as an interdisciplinary conflict. For example, in our research group, the rhetorical approach to data mining *conventions* as constraints repeatedly elicits a "let's fix it" response from the data miners. From the rhetorical perspective, some data visualizations are better than others, but evaluation is always based on the constraints that make the information useful. There is no absolute better or worse display; what is interesting is what can be said in a particular context. In that sense, rhetorical approaches are almost always social constructionist, in tension with the objectivism of the rationalist approach.

What follows are analyses of the visualizations outlined in the methods section above. Following the rhetorical tradition, we are interested in the analytic observations that differing formats impel us toward and the interpretations that result. Our main research question with respect to visualization as a methodology concerns the persuasive elements of various forms of visualization of the same data. In addition, our focus on visualization has led us to more closely scrutinize the methods for extracting data and organizing it in particular ways. As a result, we are paying close attention to the ways in which decisions made at various points in the data mining process affect the outputs that we intend to analyze. For these analyses, we used the outputs from the "no unwanted terms" extraction.

B. Analysis of Tag Clouds

An analysis of the tag cloud outputs of vaccine-related terms across 90 papers in 1918-1919 indicates that reporting on vaccines and vaccination prior to the flu pandemic was different from reporting after the pandemic—at least until the end of 1919. This form of visualization encourages a narrative interpretation of the topic modeling and segmentation algorithm's outputs, showing, for example, how authoritative voices were represented in vaccine-related reporting following the 1918 flu pandemic as well as the context of reporting.

Initially in 1918, January through May, *vaccine* is linked to *typhoid* and *smallpox* and in the context of school and the military. Several tag clouds in several segments—although not in all—also refer to vaccination in the context of public health. The first mention of the term *epidemic* is in the 6/3/1918-7/29/1918 segment, although it is unclear if the term refers to smallpox or some other plague (*plant* is in the same topic cloud, for example, as is *water*). *School* shows up less often in the tag clouds, but military-related terms (*war, camp, army*, etc.) appear in more of the clouds than earlier in the year. The terms *disease* and *typhoid* appear throughout segments in May through October 1918, as do *germ* and *bacteria*.

The term *laboratory* also shows up across the segments from October 1918 through February 1919, as do terms such as *preventive, effective, cure, anti-toxin, save*, and *remedy*, all of which suggest vaccination is represented in news reporting as an effective way to prevent or cure influenza—a representation whose ethos is bolstered by the actions and statements of the authorities.

The tag clouds in this period frequently include such terms as *inoculate* and *administer*—more so than previously—suggesting the use of more forceful language ("You must get vaccinated to be protected against disease."), more discussion and debate about vaccines and vaccination, greater availability of vaccines, more people getting vaccinated, or some combination thereof. *Smallpox* shows up once in the segment from late December 1918 to February 1919, then reappears, with *typhoid*, in March 1919. On the other hand, *influenza* appears in two topic clouds in segment 3/4/1919-4/29/1919 but not again until 5/8/1919-7/3/1919. This pattern, of course, suggests that *influenza* became less important in vaccine and vaccination discourses and *smallpox* and *typhoid* regained their former prominence after the end of the pandemic's second wave in fall 1918.

The tag clouds July through November 1919 reflect, for the most part, those from February to July—until the segment 9/15/1919-11/10-1919, when *influenza* reappears in a single cloud. The other segments link *vaccine* to *smallpox* and *typhoid*, continue to contain references to authorities (*office, officer, board, council, prohibit, compulsory, mayor, government*), and emphasize the positive aspects of vaccination (*save, treatment, free* (possibly referring to disease-free or no cost), *cure, safe*). At the same time, there are tag clouds containing terms that might suggest anti-vaccination discourses: *failure, opinion, susceptible, artificial, prove, low, toxin, value, human, doubtful, risk*.

Overall, tag clouds encourage a bounded, narrative interpretation of the data. Tag clouds convey discrete stories; however, the researcher may be able to link the various stories to form a type of "meta-narrative" relating to specific aspects of the reporting. For example, the tone of reporting on vaccination appears to be more positive than negative, although the tag clouds suggest that the positive reporting (*cure, antidote, prevent*, etc.) changed to what appears to be an imperative tone (i.e., action verbs: *inoculate, vaccinate*) in early 1919. Additionally, the terms in some tag clouds suggest that not all reporting was positive—that there may have been some concern about the safety and effectiveness of vaccines and even resistance to what appears as compulsory vaccination (9/15/1919-

11/10/1919). The data mining outputs rendered in tag clouds alone, however, cannot validate this interpretation because of the absence of context, so a close reading of a sampling of newspapers is necessary to determine the validity of the analysis.

C. *Analysis of ThemeDelta*

Unlike the tag clouds, ThemeDelta encourages an analysis of terms and clusters from a diachronic perspective. Hence, interpretation is based not on the terms themselves but, instead, on the patterns of the trendlines representing them, where they start and stop, etc. ThemeDelta seems to indicate a more particular and granular form of reporting prior to 8/7/1918 (see Fig.2). There are many and complex linear relationships such that there does not appear to be one or even several dominant discourses. Rather, there appear to be multiple, equally "frequent" discourses—with one exception. *Typhoid* appears in the 7/30/1918-8/6/1918 segment as a large and seemingly discrete topic. There is a very thick band of lines across the segments during the period from 8/7/1918 through 2/24/1919, after which the patterns resemble those in the period before 7/30/918. The lines noticeably shift again, beginning 5/8/1919-7/3/1919. The most prominent banding in the subsequent sections develops here—although some terms carry over from earlier segments—and dissipates by the 9/7/1919-9/14/1919 segment, at which point, the topics seem, once again, largely discrete. This high-level analysis suggests that vaccination discourses did shift with the onset, peak, and dissipation of the epidemic, and that they shifted again in early September 1919 before returning to pre-epidemic patterns.

A more detailed view of this visualization shows that early vaccine discourses (up to the 06/03/1918-07/29/1918 segment) relate variously to smallpox and typhoid in the context of schools and the military, but there are many breaks in the lines. That is to say, *smallpox* and *vaccine*, for example, do not seem to cross segments with any consistency. These patterns suggest that reporting on vaccination during and across these segments is sporadic rather than continuous and, again, relates primarily to smallpox and typhoid. ThemeDelta also shows clearly how the discourses were constructed around certain terms. For example, words such as *safe, free, son, price,* and *private* appear across this part of the visualization and suggest efforts to persuade parents to protect their children's health through free and safe vaccination (*private* here probably refers to a private in the army or private schools, a usage that might also affect the way we interpret other words).

The visible shift beginning in the 6/3/1918-7/29/1918 segment initially centers on *typhoid*. However, although the patterns further shift—and noticeably—during the period the influenza spread across the country and peaked (8/1918-12/1918), *influenza* appears quite infrequently in the visualization overall, and relatively little during this particular period. Instead of flu-related terms, such words as *private, cure, administer, live, charge, ulcer, safe, kidney, life*, and *bacteria* appear frequently. Reporting beginning in the 9/15/1919-11/10/1919 segment appears to repeat the patterns of reporting in early 1918 that are characterized by discrete topics with little overlap of terms across segments. This reporting appears to include discussions of vaccine-related experiments (*human, rat, result, success*) and smallpox.

D. *Analysis of Word Frequency Lists*

Of the three visualizations, word frequency lists are the most indexical. They encourage interpretation that is based on hierarchical relationships of words, an approach that sometimes obscures the importance of terms that are lower in a topic's hierarchy. Primary emphasis, then, is on the position of words, not necessarily the context within which they occur. For example, up to the 5/26/1918-6/2/1918 segment, reporting specifically on vaccine and vaccination appears to be, if not urgent, at least prominent given the term's (or related terms') appearance in 12 of 16 total topics. That is, prior to 5/26/1918, *vaccine, vaccinate, vaccination, inoculate*, or *inoculation* occur at frequencies that are high enough to place them among the top 20 terms in three or four of five topics in each segment as well as in the single topic in the 1/2/1918-1/9/1918 segment.

The frequency of *vaccine* and its related terms in reporting increases again in the 6/3/1918-7/29/1918 segment, this time in conjunction with numerous disease terms—both general and specific: *smallpox, epidemic, spread* (Topic 2); *typhoid, disease, bacteria* (Topic 3); and *germ* (Topic 5). As the influenza epidemic spreads, there seems to be a shift in the types of terms that appear in the word clusters. In the 8/7/1918-10/2/1918 segment, for example, there is no indication that any of the vaccine discourses relate specifically to the Spanish flu. Rather, the most cohesive group of words seems to be in Topic 5, in which the disease term *smallpox* appears along with *school, children, public, pupil, board*, and *admission*. These are all terms that suggest a conversation about vaccines for school-aged children as a necessity for admission to school. In the 10/3/1918-10/10/1918 segment, however, the topics contain terms such as *total* and *effective* (Topic 1), *quality* (Topic 2), *death* (Topic 3), *ease* (Topic 4), and *cure, low, trust,* and *poison* (Topic 5), all of which suggest discourses relating to the safety and efficacy of vaccination. *Influenza* first appears in the 10/11/1918-12/6/1918 segment.

By segment 12/29/1918-2/23/1919, *vaccine, vaccinate, vaccination, inoculate*, and *inoculation* have become the most frequent terms in all five topics, appearing several times in four of the five topics in that segment. Vaccine reporting in the 3/4/1919-4/29/1919 segment seems to take place in the context of disease prevention (*avoid, cover, crowd, spread*), and in the 4/30/1919-5/7/1919 segment, there seems to be an effort to convince the public to get vaccinated (*administer, cure, safe, effective, pro, urge*). The word *danger* appears in Topic 4 and could refer either to the danger of disease or the danger of vaccination. *Fowler*

shows up in Topic 5, the first time the name of a public health official (W.C. Fowler was the district health officer in Washington, D.C. at the time) appears in these word frequency lists—although *surgeon*, which likely refers to Surgeon General Rupert Blue, appears in Topic 4 in segment 10/3/1918-10/10/1918 and in several other subsequent segments.

By segment 5/8/1919-7/3/1919, the frequency with which *vaccine* and its related terms appear in reporting has diminished. However, in the 9/15/1919-11/10/1919 segment, there is a spike in the occurrence of *vaccine, vaccinate, vaccination, inoculate*, and *inoculation*, which appear in every topic, sometimes in combination, and as the most or second-most frequent term in three of five topics. It should also be noted the Topic 5 seems to reflect some anti-vaccine sentiment (*doubtful, risk*). By the end of 1919 (segment 11/19/1919-12/24/1919), vaccine-related reporting seems to have returned to its pre-epidemic characteristics, that is, reflecting the issue of vaccinating school-aged children (Topic 1) as well as inoculation as it relates to animals (Topic 2) and plants (Topic 3), although there was also reporting on vaccine development (Topic 4).

E. Rhetorical Effects of Visualization Choices

The three visualizations each represent topic modeling and segmentation outputs in useful ways. Because of the nature of the word clusters to index the topics, these representations are indeed more similar than different. However, they each encourage different types of analysis. As a result, interpretations of the visualizations provide unique insights, but with significant overlap. The ThemeDelta visualization, for example, highlights the flow of discrete words across the entire period of analysis. Consider the word *private*, which might suggest vaccination discourses that emphasize personal choice, refer to a private in the military or a private school, or advertise a private medical practice. When the cursor is placed over the term—regardless of where it is located in the visualization, the line representing it turns blue, and it becomes immediately apparent that vaccination-related reporting consistently included this term from the 6/3/1918-7/29/1918 segment through the 12/29/1918-2/23/1919 segment and again beginning in the 3/4/1919-4/29/1919 segment. In short, the ThemeDelta visualization represents terms and themes across the horizon of analysis and makes more visible how a particular word moves through groupings of words.

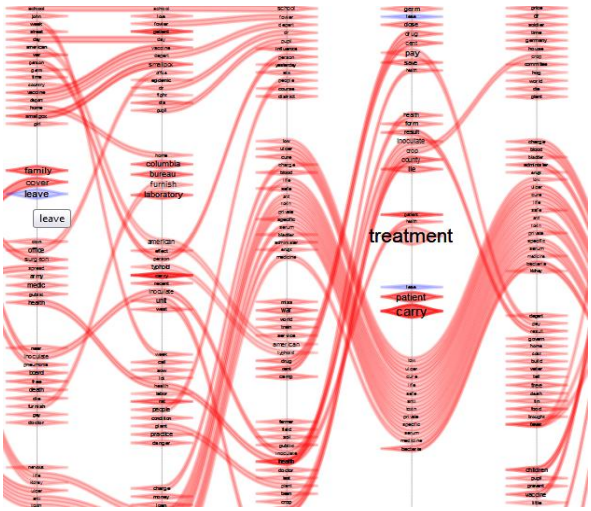

Figure 3: ThemeDelta Frequency Representation

Tag clouds facilitate an analysis that is bounded and more oriented to narrative. Within the segments, each tag cloud seems to tell a particular story; however, those stories are not linked linearly; rather, they are proximal only insofar as they are co-located within a particular segment. That is to say, tag clouds emphasize the *spatial* aspect of words *and groups of words*. As a result, interpretations of the topic modeling and segmentation outputs, when visualized as tag clouds, suggest stories. While the stories can be linked to form a meta-narrative about a particular theme or series of events, they are not necessarily related, either across time or within a segment. On the other hand, ThemeDelta draws the eye across the visualization to emphasize the *temporal* aspect of words—although not groups of words since the groupings continually shift (except for advertisements). Hence, interpretations of this form of visualization trace connections across time, but it de-emphasizes the narrative since the lines become the focus, not the clusters.

The different ways these two visualizations depict word frequency is another example of how different representations of the same data encourage different interpretations, that is, persuade differently. Tag clouds indicate frequency by the size of the word in a cluster, while ThemeDelta indicates frequency by the thickness of a word's lines within a cluster. For example, in the 7/4/1919-7/11/1919 segment, the line representing *leave* (highlighted in blue in Fig. 3) is significantly smaller than those representing *patient* and *carry*, indicating that *leave* occurs less frequently in this cluster than does *carry*.

In the tag clouds, *vaccine* is often the largest word in clusters where it occurs (See Fig. 4). It is clear that *vaccine* is relatively more frequent than some of the other words in its tag cloud, but because in this output (as with all the outputs we are analyzing) the data is not normed within or across segments, it is impossible to gauge the relative frequency of *vaccine* in one segment versus in another. Similarly, the size of *serum* in the segment on the left in Fig.4 is approximately the same as the size of *serum* in the segment on the right, yet their relative frequency cannot be

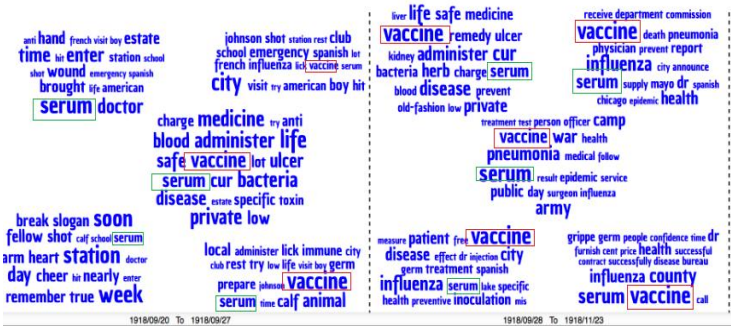

Figure 4: Tag Clouds Frequency Representation

inferred from this size similarity—despite the researcher's tendency to do so.

ThemeDelta and tag clouds tend to obscure words' relative frequency outside, within, and across segments. Furthermore, unless the word frequencies differ enough within a cluster to change word size or trendline thickness, it is impossible to tell the relative importance of words of like size within a cluster. Word frequency lists, however, allow the researcher to easily identify the key word(s) in each cluster as well as the relative importance of every other word within the topic. Additionally, word frequency lists facilitate inter-segment analysis; patterns within a segment emerge since the researcher is able to line the topics up and compare them across the segment. On the other hand, the word frequency lists appear more indexical than either ThemeDelta or tag clouds, and their linear presentation makes intra-segment analysis difficult.

IV. Conclusion

Forms of visualization like ThemeDelta hold great promise, as they represent through the trendlines the recurrence of words and word clusters from one time segment to the next. The trendlines allow the user to see clusters that remain relatively similar across time, indicating that reporting on a particular topic is consistent or that an advertisement is repeatedly published across several weeks or months. As an index, then, ThemeDelta offers more information more directly to the reader, who only needs to highlight a particular word to see it trending across the segments in various word clusters. Tag clouds, on the other hand, facilitate a narrative analysis of topics within a segment. They encourage the researcher to find the story in the data mining output at the same time they present that story as a "bounded" narrative. Word frequency lists help the researcher to identify the relative importance of terms within a topic as well as to better develop themes across a specific segment. The word frequency lists also appear to facilitate the identification of recurrent clusters across segments. However, they only indicate relative frequency hierarchically, with the most frequent term at the top of each list and terms of lesser frequency lower on the list; they do not have a mechanism for more finely grained representation of relative frequency unless we attach the actual numerical value of each word, which makes the representation of the lists unwieldy.

This analysis shows that different visualizations help to persuade the researcher toward different ends. In each analysis, the researcher understood and presented findings with different emphases—as trends (ThemeDelta), as narratives (tag clouds), and as indices (word frequency lists). Whether implicit or explicit, the context in which design conventions are derived and become unquestioningly accepted—naturalized—impacts how visualizations operate rhetorically toward certain ends. Purposeful attention to visualization and the methodological conventions that are embedded in particular visualization practices will allow humanists to have more confidence in their interpretations of big data, a key element in the acceptance of data mining as a valuable method for humanities research.

References

[1] R. Arnheim, *Visual Thinking*, Los Angeles and London: University of California Press, 1997.

[2] J. Barry, The Great Influenza: *The Story of the Deadliest Pandemic in History*, New York: Penguin Books, 2005.

[3] B. Barton and M. Barton, "Ideology and the Map," *Central Works in Technical Communication*, New York: Oxford University Press, 2004, pp. 232-252.

[4] N. Bristow, American Pandemic, New York: Oxford University Press, 2012.

[5] R. Eggo, S. Cauchemez, and N Ferguson, "Spatial dynamics of the 1918 influenza pandemic in England, Wales and the United States," *Interface*, vol. 8, 23 Jun. 2010, pp. 233-243, doi: 10.1098/rsif.2010.0216.

[6] C. Freifeld, K. Mandi, B. Reis, and J. Brownstein, "HealthMap: Global Infectious Disease Monitoring through Automated Classification and Visualizatin of Internet Media Reports," *Journal of the American Medical Informatics Association*, vol. 15.2, Mar. 2008, pp. 150-157, doi: 10.1197/jamia.M2544.

[7] R. Godderis and K. Rossiter, "'If you have a soul, you will volunteer at once': Gendered expectations of duty to care during pandemics," *Sociology of Health and Illness*, vol. 35, Feb. 2013, pp. 304-308, doi:10.1111/j.1467-9566.2012.01495.x.

[8] S. Hall, Ed., *Cultural Representations and Signifying Practices*, London: Sage Publicatons, 1997.

[9] J. Hullman and N Diakopoulos, "Visualizaton Rhetoric: Framing Effects in Narratie Visualization," *IEEE Transactions on Visualization and Computer Graphics*, vol. 17.2, Dec. 2011, pp 2231-2240.

[10] M. Humphries, *The Last Plague: Spanish Influenza and the Politics of Public Health* in Canada, Toronto: University of Toronto Press, 2013.

[11] C. Kostelnick and M. Hassett, *Shaping Information: The Rhetoric of Visual Conventions*, Carbondale: Southern Illinois University Press, 2003.

[12] E. Morris, *Believing is Seeing*, New York: Penguin Press, 2011.

[13] V. Northington Gamble, "'There wasn't a lot of comforts in those days': African Americans, public health, and the 1918 influenza pandemic," *Public Health Reports*, vol. 125, Mar. 2010, pp. 114-122.

[14] M. Sorapure, "Information Visualization, Web 2.0, and the Teaching of Writing," *Computers and Composition*, vol. 27, 2010, doi:10.1016/j.compcom.2009.12.003.

[15] C. Spinuzzi, *Tracing Genres Through Organizations*, Cambridge, MA: The MIT Press, 2003.

[16] E. Tufte, *Visual Explanations*, Cheshire, CT: Graphics Press, 1998.

[17] E. Tufte, *Beautiful Evidence*, Cheshire, CT: Graphics Press, 2006.

[18] M. Zachry and C. Thralls, "Cross-Disciplinary Exchanges: An Interview with Edward R. Tufte," *Technical Communication Quarterly*, vol. 13.4, Fall 2004, pp 447-462.