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

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.