White Paper Report

Report ID: 105464

Application Number: HJ-50067-12

Project Director: E. Thomas Ewing (etewing@vt.edu)

Institution: Virginia Polytechnic Institute and State University

Reporting Period: 1/1/2012-6/30/2014

Report Due: 9/30/2014

Date Submitted: 9/4/2014

An Epidemiology of Information

Data Mining the 1918 Influenza Pandemic

Project Research Report
March 25, 2014

Research Team:
E. Thomas Ewing
Samah Gad
Bernice L. Hausman
Kathleen Kerr
Bruce Pencek
Naren Ramakrishnan

An Epidemiology of Information
Data Mining the 1918 Influenza Pandemic

TABLE OF CONTENTS

Introduction: Using Newspapers to Study 1918 Influenza

In 1918, pandemic influenza (so-called Spanish flu) moved through the world in two waves, the first mild and the second deadly. Scholars continue to analyze the disease's pathogenesis and the social, historical, and policy-related implications of the pandemic. A 2010 special issue of *Public Health Reports* offers a broad historical overview of the pandemic and its social and cultural impacts [1]. Scholarly books, most recently, Nancy Bristow's *American Pandemic* [2] and Mark Osborne Humphries's *The Last Plague* [3], detail the effects of the pandemic on developing national public health programs in the United States and Canada. Epidemiological research on the disease continues to reveal critical elements of the flu's pathogenesis and abnormal mortality.

This project research report describes the results of four case studies undertaken as part of Virginia Tech's "An Epidemiology of Information: Data Mining the 1918 Flu Pandemic," which was funded through the Digging into Data Challenge of the National Endowment for the Humanities. Although most historical accounts of the Spanish flu make extensive use of newspapers, our project was the first to ask how looking at these texts as a large data source can contribute to historical understanding of the pandemic.

When we wrote the grant proposal for "An Epidemiology of Information" three years ago, we thought that data mining would allow us to analyze all *Chronicling America*'s newspapers for 1918 and 1919. Even given the tempered assessments of the capacity of computational analytics to transform the humanities with quantitative data [4 and 5], we were confident in our capacity to develop methods that utilize digitized newspapers and provide a fuller accounting of the impact of the influenza pandemic than previous smaller-scaled studies.

What we found is that digitized historical sources such as the newspapers in the *Chronicling America* collection present significant limitations to computationally based research and data mining cannot substitute for close or "manual" readings of discrete texts. For example, the greatest value of newspapers for historical analysis is also their greatest challenge to data mining methods. Newspapers illustrate on a daily basis how society comes to grip with an emerging crisis (see **Figure 1**, for example). Newspapers offer a textured and layered body of text that is best understood in context through careful reading across multiple titles over a sustained period of time. By decontextualizing the text, disconnecting the language, and simplifying the categories, data mining methods compromise both the integrity and the complexity of these sources. In addition, the process of digitizing print sources creates problems in terms of the data source itself; poor optical character recognition (OCR) made tone analysis all but impossible unless we manually edited the text, an extremely time-consuming task.

We also found that the rationalist/social constructionist divide plays an important role in how we developed our methods, framed data sets, and visualized the computational outputs. A rationalist approach suggests that there is an achievable ideal: better data, more appropriate criteria, or more accurate representations. The social constructionist approach, however, suggests that these elements are negotiated and emphasizes the naturalization of practices, conventions, and assumptions. The tension between rationalist and social constructionist approaches surfaces as an interdisciplinary conflict. In this project, the rhetorical approach to data mining conventions as

naturalizations conflicted with the rationalist approach of our data mining colleagues, who wanted to adjust the algorithms in order to achieve "better" results.

Nevertheless, analysis of newspapers can benefit a great deal from data mining methods. By looking across large data sets, identifying emerging themes, and identifying broad categories, these methods can guide humanities scholars into new approaches to textual analysis. In this way, data analytics enhances traditional humanities research into primary sources.


NATION'S CAPITAL BATTLES FOR LIFE AGAINST "FLU"

Fig. 1: Photograph from *The Bismarck Tribune*, 10/15/1918, page 2

The greatest challenge for humanistic researchers committed to using new methods such as data mining is striking a balance between the broad overview that is its strength and traditional forms of close reading that produce more complex and nuanced interpretations of source materials. Our findings indicate that topic modeling is most useful for identifying broad patterns while tone classification can suggest insights into how newspapers shaped the meaning of the epidemic. Topic modeling and segmentation together, the "dynamic temporal segmentation" algorithm discussed in the **Project Methods** section, reveal change points in reporting on the disease in specific localities. Our investigation into visualization conventions suggests a rich avenue of collaboration between computer scientists and humanists as both groups seek to find effective ways to integrate research methods and goals. We also note that scale matters in data mining—the outputs for aggregate data differ from the outputs of individual newspapers, suggesting that what rises to significance in the topic modeling of aggregate data is not always representative of reporting in specific communities.

Integrating computational methods with more traditional humanistic approaches to historical analysis offers much in the way of innovation. We find ourselves continually compelled by the possibilities of the next step.

Weekly Newspapers Case Study

I. INTRODUCTION

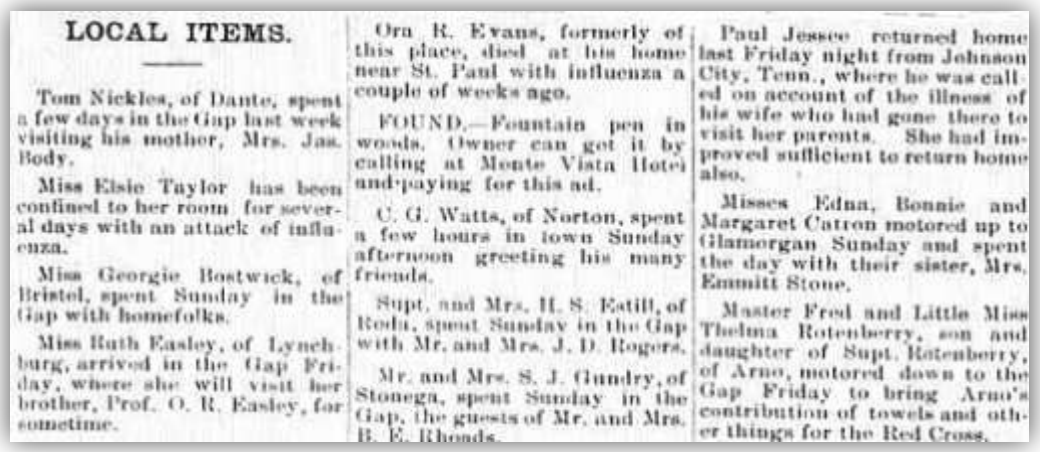
The application of large-scale analytical methods to weekly newspapers provides new insights into the ways that information was communicated across a wide geographical region as well as more intensively within a particular community. This case study uses a sample of 24 weekly newspapers to examine the ways that information about influenza was distributed to and within predominantly rural communities in the United States and Canada.[1] Topic modeling and segmentation was used to analyze all the titles, and tone classification was used to examine a selection of texts from a sample of eight titles.

Table 1. Newspaper List: *Title* (Location); * *Titles* Used for Tone Classification Study

Alliance Herald (Alliance, NE)	*Jasper News* (Jasper, MO)
Ashland Tidings (Ashland, OR)	*Liberal Democrat* (Liberal, KS)
* *Big Stone Gap Post* (Big Stone Gap, VA)	*Lynden Tribune* (Lynden, WA)
Bow Island Review (Bow Island, AB, CA)	*Manning Times* (Manning, SC)
Central Record (Lancaster, KY)	* *Middlebury Register* (Middlebury, VT)
Clinch Valley News (Clinch Valley, VA)	*Mount Vernon Signal* (Mt Vernon, KY)
Colfax Chronicle (Colfax, LA)	* *Perrysburg Journal* (Perrysburg, OH)
* *Colville Examiner* (Colville, WA)	*Princeton Union* (Princeton, MN)
Dakota County Herald (Dakota City, NE)	* *Red Deer News* (Red Deer, AB, CA)
Graham Guardian (Safford, AZ)	*St Joseph Observer* (St. Joseph, MO)
* *Hays Free Press* (Hays, KS)	* *The Era-Leader* (Franklinton, LA)
* *Iron County Record* (Cedar City, UT)	*Warren Sheaf* (Warren, MN)

This case study reveals important insights that can inform both historical and epidemiological understanding of the Spanish flu. It was designed to answer these research questions: 1) When and how did news about the influenza pandemic reach a community? 2) How did the spread of this information anticipate and shape responses to the outbreak of disease within the community? 3) How did the tone of reporting about influenza change before, during, and after the peak of the epidemic?

Fig. 2: *The Big Stone Gap Post*, November 20, 1918, page 2



LOCAL ITEMS.

Tom Nickles, of Dante, spent a few days in the Gap last week visiting his mother, Mrs. Jas. Body.

Miss Elsie Taylor has been confined to her room for several days with an attack of influenza.

Miss Georgie Bostwick, of Bristol, spent Sunday in the Gap with homefolks.

Miss Ruth Easley, of Lynchburg, arrived in the Gap Friday, where she will visit her brother, Prof. O. R. Easley, for sometime.

Ora R. Evans, formerly of this place, died at his home near St. Paul with influenza a couple of weeks ago.

FOUND.—Fountain pen in woods. Owner can get it by calling at Monte Vista Hotel and paying for this ad.

C. G. Watts, of Norton, spent a few hours in town Sunday afternoon greeting his many friends.

Supt. and Mrs. H. S. Estill, of Roda, spent Sunday in the Gap with Mr. and Mrs. J. D. Rogers.

Mr. and Mrs. S. J. Gundry, of Stonega, spent Sunday in the Gap, the guests of Mr. and Mrs. B. E. Rhoads.

Paul Jessee returned home last Friday night from Johnson City, Tenn., where he was called on account of the illness of his wife who had gone there to visit her parents. She had improved sufficient to return home also.

Misses Edna, Bonnie and Margaret Catron motored up to Glamorgan Sunday and spent the day with their sister, Mrs. Emmitt Stone.

Master Fred and Little Miss Thelma Rotenberry, son and daughter of Supt. Rotenberry, of Arno, motored down to the Gap Friday to bring Arno's contribution of towels and other things for the Red Cross.

Analysis of weekly newspapers indicates that news about the influenza pandemic reached a community through a combination of local, even personal, reports of sick family members in other locations (usually army camps) and news reports of disease from cities already afflicted. In most cases, these reports preceded the actual arrival of disease within the community. Once the disease reached a community, newspapers provided extensive coverage of individual victims (see **Figure 2**), endorsements of public health measures, and, less frequently, comprehensive numbers on the death toll. As the disease receded, the level of coverage declined slightly, with a resurgence of coverage in anticipation of the disease's return. Influenza was a significant presence in local news reporting, editorial commentary, statements from public health

authorities, advertisements for local and national products, and reports of individuals who were sick, had died, or were recovering.

II. METHODS

Data Sources

This case study examines 24 weekly newspapers for 1918 from the *Chronicling America* collection and *Peel's Prairie Provinces*. Weekly newspapers were selected because their scale makes it possible to trace both the spread of information and the impact of the disease within a particular community over a four-month period, from the first reports of disease, usually in early October, through the decline and then resurgence—in most cases, in late November, throughout December, and into early January 1919. Weekly newspapers were also selected because a broad range of titles is available through *Chronicling America* for 1918. Twenty two titles came from geographically diverse locations in the United States, with two from Western Canada (see **Table 1**). In addition to geographical variability, this distribution means that newspapers came from regions that were affected by the influenza epidemic at different times during the final months of 1918. Across these 24 titles, a keyword search from the *Chronicling America* collection generates more than 1,000 pages of newspaper copy that include the words *influenza* or *grippe*, which adds up to several million words that can be analyzed with our data mining methods.

III. DISCUSSION

Topic Modeling

For this case study, 25 topic models were generated: one for each of the 24 titles and one with combined data from all the titles. The combined topic model, analyzed in detail below, illustrates three broad trends in how weekly newspapers reported on influenza. First, reports about influenza from outside the community, which appeared early in the epidemic, included both health reports from national authorities (including the surgeon general) and reports on individual soldiers who fell ill and died in the camps. Second, during the middle segment, which is only one week, the reporting shifted to the local level in terms of public health measures, while also referencing national reports. Finally, the presence throughout these tag clouds of familial relations and activities indicates that much of the reporting on the influenza epidemic came in the form of "local news" items, including reports on illness and deaths on an individual basis. In this sense, weekly newspapers, perhaps more than daily newspapers, served as a kind of social media with user-created content offering the bulk of the information about the rise, peak, and decline of the epidemic.

Fig. 3: 8/23/1918-10/18/1918

The tag clouds for August 23 to October 18 for all titles (**Figure 3**) show three distinct reporting patterns during the period prior to the arrival of the epidemic in most rural communities (with the noted exception of Middlebury, VT). The top left cloud (1) indicates the reporting on disease in relation to domestic and familial connections (*wife, son, family, home, visit*, and *miss*). The word *camp* suggests victims in army camps, which is also indicated by the word *son*. The bottom right cloud (5) provides further evidence of reporting on disease in camps with terms such as *army* and *camp*. The word *school* in clouds 1 and 5 suggest public health measures (school closure was a common practice across the country). The middle cloud (3) has more medical terms, including *germ, cough, spread, fever, cold, feel*, and *severe*. These terms probably come from reports issued by public health officials, either locally or nationally.

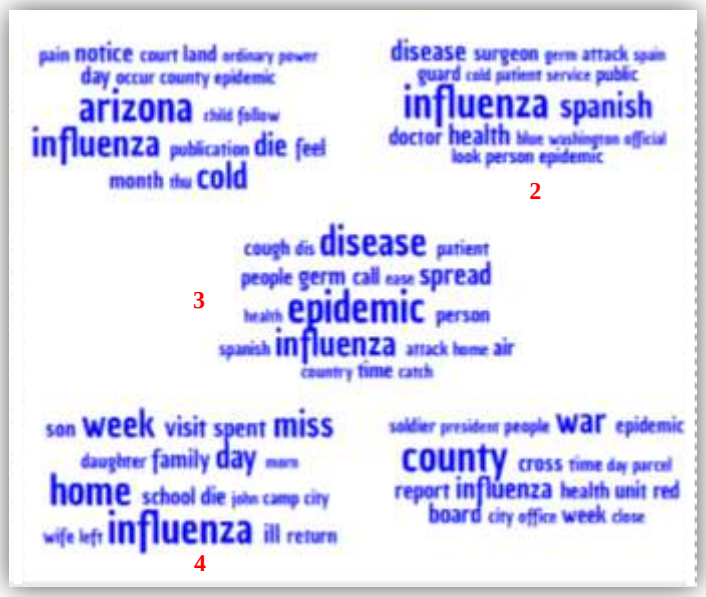


Fig. 4: 10/19/1918-10/26/1918

The tag clouds for October 19-26 (**Figure 4**) suggest a shift in reporting to local conditions, with terms such as *office* and *board* suggesting a county or city health board, while temporal terms such as *day* or *week* may have referred to the duration of public health measures such as closing schools. Familial connections are visible in the lower left cloud (4), including *family, wife, daughter,* and *son* as well as *visit, return*, and *miss*. The middle cloud (3) features medical terms, including *patient, disease, germ, spread, cough,* and *catch*, suggesting warnings about the further spread of the epidemic. Military terms such as *camp* do not appear in this segment, but *washington* appears in the same cloud as *surgeon* and *blue* (suggesting Surgeon General Rupert Blue), *public, spanish,* and *health*, all suggesting the publication of national warnings.

Fig. 5: 10/27/1918-12/22/1918

Finally, the tag clouds for October 27 to December 22 (**Figure 5**) show continued reporting on public health measures with terms such as *county, school, city, public, board*, and *meet* in the bottom right cloud (5). The top right (2) cloud has family terms that indicate the continued toll of the disease. The top left cloud (1) has medical terms, but the appearance of *tablet* suggests that some of these words are in advertisements rather than articles.

The tag clouds for the combined data suggest three broad interpretations. First, weekly newspapers began to report on influenza when the disease reached their communities, even though news about disease in other locations was widely available weeks earlier. The fact that the middle segment (10/19/1918-10/26/1918) makes up only one week indicates how intensively the disease was being reported in the middle of October, which was the point when the influenza was both the most widespread and the most devastating across the country. Second, family relations and activities were an important part of reporting on the influenza, as news about individual victims was widely disseminated through these newspapers. These individual victims can be identified as soldiers in the tag clouds, a representation that is consistent with both patterns of disease transmission and the impact of the draft, as soldiers drawn from all over the country were concentrated in military camps that had the first outbreaks of serious disease. Finally, these tag clouds emphasize the importance of public health reporting in the newspapers, which can be seen both in the peak period, when measures such as closing schools were first implemented, and again in the final cloud, when these measures were discussed either because they were lifted or because they were being reinstated in order to prevent another wave of disease.

Fig. 6: *Middlebury Regi*ster/*Iron County Record* Comparison

The topic models for individual titles allow for more focused analysis. A comparison of the *Middlebury Register*, located in an eastern community that had its first reported cases in the September 27 issue, and the *Iron County Record*, which reported the first cases two weeks later on October 10, offers some points of comparison. In these newspapers, the segment representing the peak of reporting begins with the issue that followed the report of the first cases (September 28 for the *Middlebury Register* and October 18 for the *Iron County Record*). In looking at the peak reporting segments (**Figure 6**), certain patterns are evident: reports on public health actions, indications of family relations and activities, terms associated with advertising, and the names of locations in the vicinity of the publishing locations. In both cases, the peak segment extended well beyond the end of the epidemic, which had ceased by November 1 in Middlebury (with the lifting of all bans on public activities) and by December 20 in Cedar City.

Topic modeling is most effective as a method for identifying broad trends and guiding future research for textual analysis. The insights from the tag clouds and the potential for proximity searching offered by *Chronicling America* allow for more focused exploration of specific titles. In the case of the *Colville Examiner,* a weekly newspaper published in Washington state, a search for *influenza* and *camps* within 50 words of each other generates 16 results, including the first mention of influenza in camps as a factor causing the cancellation of the draft call (September 28); a report from the Washington State Commissioner of Public Health on October 12 urging the population to take measures to prevent the further spread of disease that is spreading "very rapidly" in the Eastern United States and in army camps; two reports on October 19 about men from Colville who are sick at training camps; and, finally, the report of a soldier who died from influenza at Camp Lewis (October 26). This trajectory of reporting, from more distant cases reported earlier but with little direct engagement to more local cases with greater detail and interest, can be predicted by reading across the word clouds generated by the topic modeling and segmentation algorithm.

Table 2. Sample Sentences from Tone Classification: *Big Stone Gap Post*

Alarmist	In many instances whole families—sometimes three generations were ill at the same time in the same house and starvation almost stared them in the face.
	In many an humble [sic] cabin, far up on a mountain in the more remote sections of Virginia, whole families lay ill when the agent found them.
	In many of the counties here is an appalling lack of doctors and nurses and the agents are being called upon to take their places.
	In several neighborhoods the supply of coffins utterly ran out while almost everywhere there was a shortage of doctors and nurses.
Warning	Most things Spanish are not to be sneezed at, but the enfluenza [sic] is an exception.
	While there are several cases here the disease has not become epidemic and will not if our people will avoid congregating together and keep their children at home until the danger is passed.
	Nevertheless the danger continues and all persons are cautioned not to relax from the requirements of the health authorities to which is attributed in large degree are present comparative immunity
	The State Health authorities have come to the aid of the stricken section, having dispatched some six or eight doctors and ten nurses to aid in the field to stop the spread of the epidemic and give relief to those that already have it.
Explanatory	We understand from a reliable source that there are around thirteen hundred cases of influenza in the St. Charles coal fields and surrounding country.
	Quite a number of the victims have died.
	This municipality, which is far more rural than urban, was the centre of a genuine hotbed of Spanish 'flu' and the State Board of Health was quick to dispatch Dr. W. A. Brumfield of the U.S. Public Health Service to the science of the distress.
	While there are a good many cases of influenza in Big Stone Gap the disease does not seem to be spreading.
Reassuring	There is every hope that it has been checked and that it will not become epidemic here, as it has in some towns in Southwest Virginia.
	The men and women agents have been instructed to do this by organizing their people for service—opening soup kitchens in connection with temporary or permanent hospitals, helping to open and maintain these temporary hospitals, and in any way that seems best to each individual agent.
	Richmond well has occasion to be proud of the speed with which she conceived, planned, and opened her emergency hospital in John Marshall High School, but even so, her record is hardly more creditable than that of the little town of Pennington Gap in Lee County.
	Many a serious illness begins with a simple cold that you can guard against by carrying with you a package of Rexall Cold Tables.

Tone Classification

To analyze the tone of reporting in weekly newspapers, we selected texts from eight titles for further analysis. The selected texts came from local reporting on the disease, including news articles about the disease and related public health measures, statements from county and city health officials published in newspapers, editorials and letters offering a more subjective

analysis, and a few advertisements from local companies that referenced influenza. This sample did not include reports on individual victims, such as obituaries or reports of ill individuals, except as these items occurred in the news reporting described above. We analyzed a total of 723 sentences with the following distribution: *Hays Free Press* (66), *Colville Examiner* (169), *Iron County Record* (142), *Perrysburg Journal* (25), *Red Deer News* (70), *Middlebury Register* (94), *Era Leade*r (35), and *Big Stone Gap Pos*t (122). The variation in the number of sentences was related to the amount of local coverage of the influenza in each newspaper. Certain newspapers, such as *The Big Stone Gap Post* and the *Iron County Record*, provided extensive coverage during the peak of the epidemic. Other titles were more moderate in their local coverage, leaving most of the mentions of influenza to local items about individual victims and their families.

All the sentences were classified according to the following four categories: alarmist, warning, neutral, and reassuring. **Table 2** provides samples for each category identified by the classifier. Across all eight titles, fewer than 2% of sentences were classified as alarmist, 16% were classified as warning, 61% were classified as explanatory, and 21% were classified as reassuring (see **Project Methods** section for descriptions of each category). The high proportion of sentences classified as explanatory can be explained by both the nature of newspapers as documentary evidence and the tool used for tone classification. First, most newspaper reporting was intended to be objective, which makes the high rate of explanatory sentences consistent with the historical evidence. Second, the explanatory classification seems to be a default result for sentences that are not easily classified as "negative" (alarmist or warning) or "positive" (reassuring). Finally, the classifier is trained with sample sentences coded by the project team, which identified many more sentences in this category. Thus, the classifier was more likely to identify sentences with explanatory characteristics.

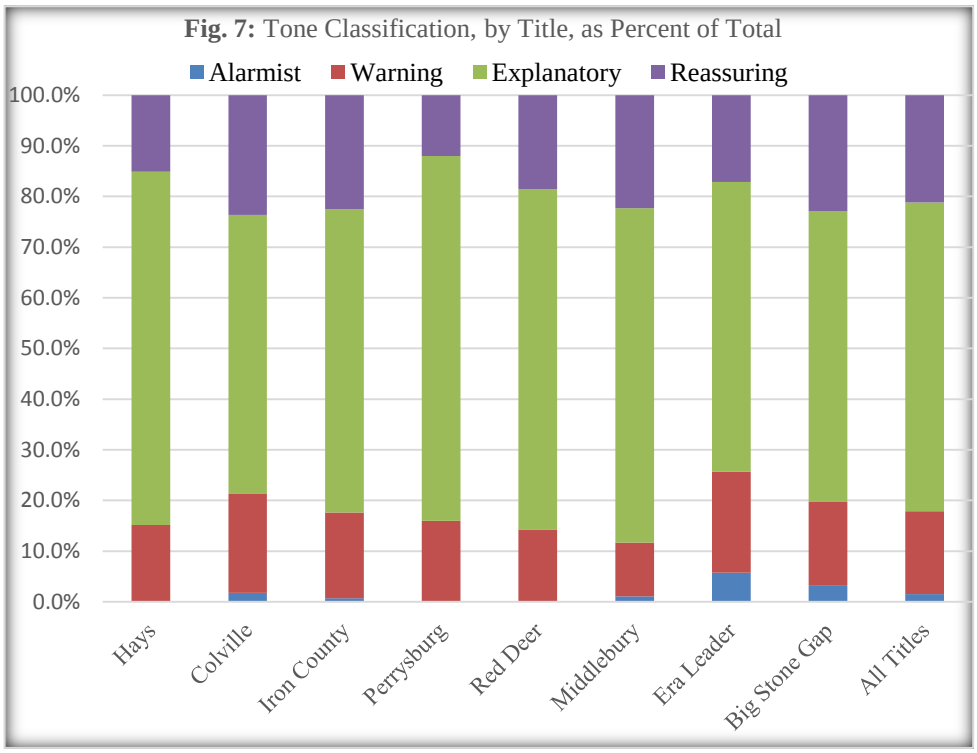


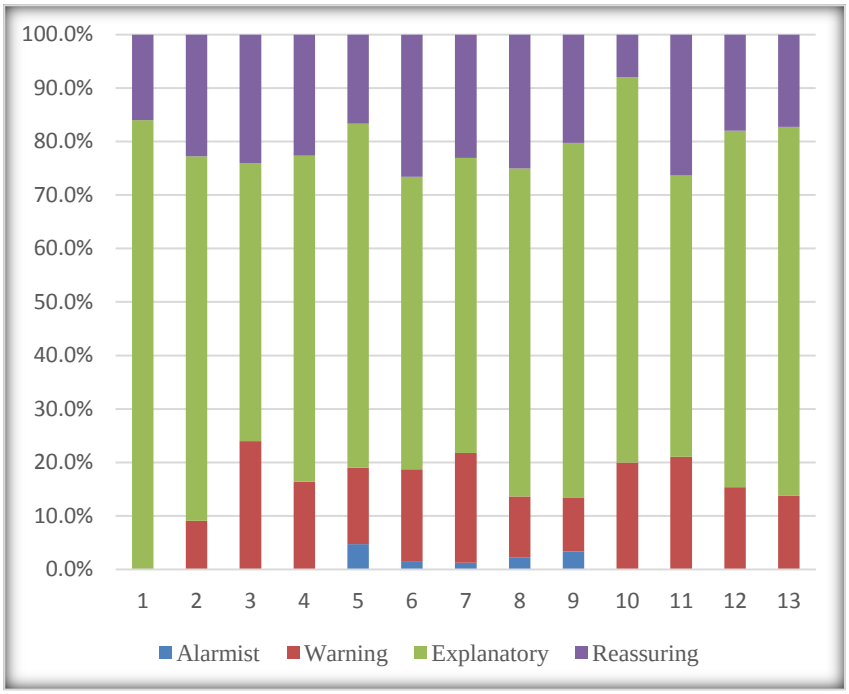
Fig. 7: Tone Classification, by Title, as Percent of Total

Figure 7 shows the distribution of tone categories by title. While explanatory made up the largest share for every title, the proportion varied from a high of 72% to lows of 55-59%. In the cases of three newspapers in the latter category, *The Colville Examiner*, *The Big Stone Gap Post*, and *The Perrysburg Journal*, the selected sentences comprise a large proportion of the data set, and the result suggests that the tone categories even out as the number of sentences classified increases.

A comparison of tone classifications across time (**Figure 8**) shows a similar consistency. Although the proportion of alarmist and warning sentences never exceeded more than one-quarter, the peaks in this form of negative reporting came early in the cycle, October 6-12, during the first two weeks of November, and again in the first week of December. These same weeks also saw the highest proportions of reassuring sentences, suggesting two possible interpretations: either alarmist and warning statements were accompanied by more reassuring statements, or the increased number of sentences being classified produces more even distribution. The former explanation connects to the nature of rhetoric in the historical context; the latter explanation is more related to the analytical tool of the classifier.

Tone classification illustrates the real challenges that the complexity of written language poses for training automatic classifiers, particularly with the type of complicated and multi-layered texts published in newspapers during this era. As an example, consider this sentence from the November 1, 1918, issue of the *Iron County Record*, which reported on the first death from influenza in Cedar City, Utah: "Many of the cases are very mild and require little or no attention, aside from remaining in and avoiding the taking of cold, but one never knows who is going to develop a deadly attack, as so many have done throughout the country." This sentence can be classified differently depending on whether the emphasis is on the first part of the sentence, where a reassuring tone is struck by the reference to the mildness of cases and the limited measures; or a warning tone, as indicated by the urgency of proposed measures to prevent further spread; or the alarmist tone of the latter part of the sentence, with a reference to a deadly attack spreading throughout the country. An argument could be made for classifying this particular sentence as any of these three tones. Given this ambivalence, the classifier chose the neutral tone of explanatory, which was confirmed during the manual verification of tones.

Fig. 8: A Comparison of Tone Categories across Time by Week



IV. CONCLUSION

The application of combined methods to review weekly newspapers illustrates both the richness of information about influenza contained in daily newspapers and the complexity of analyzing textual materials using automated analysis. Weekly newspapers served multiple purposes within communities: first, distributing national news acquired through wire service reports; second, communicating information from local authorities, such as mayors, county and city health officers, and town councils; third, serving a regional audience, usually through news reports from surrounding communities; and finally, providing first hand reports about the immediate community, including reports of individuals who were sick, recovering, or deceased

(**Figure 9**). Particularly in the context of the 1918 influenza pandemic, which involved news reported on both national and local levels, the newspapers exhibited a unilateral transmission pattern (that is, reports transmitted from one location, usually Washington DC, throughout the nation) in a horizontal news network (news generated, communicated, and received across the local or regional community). Our review of weekly newspapers indicates that they served a far more complex function than conveyed by historians Alfred Crosby or John Barry. Our analysis also illustrates the value of newspaper reporting beyond the mortality data extracted by epidemiological studies. Following the historical and anthropological approaches taken by Nancy Bristow and Ann Herring, the Weekly Newspapers Case Study demonstrates the complex information networks that functioned through newspapers. In this sense, the study of weekly newspapers suggests the value of our methods for contemporary contexts, for example, studying the use of social media to track self-reporting on diseases.

Middlebury Register

Save to Buy and Buy to Keep

VOLUME LXXXII — MIDDLEBURY, VERMONT, FRIDAY, OCTOBER 4, 1918 — NUMBER FORTY

MIDD IS MUSTERED INTO U. S. SERVICE

But More Elaborate Ceremony Will Be Held Later

BIGGEST DRAFT LOTTERY DRAWN

Frank James of Vergennes Holds First Number Picked

DEATH TOLL AND SICK LIST IS ONE OF THE HEAVIEST ON RECORD

Practically Every State in the Union is Affected; Innumerable Cases in the County

By government order, all papers not paid in advance must be discontinued. See the label on your paper; if in arrears, pay now. Keep The Register coming

DR. THOMAS IS TO ENTER CAMP TODAY

Left Wednesday Night for Kentucky

GOOD START, BUT LACKING $27,400

Loan Figures up to $72,600 This Noon

Middlebury Needs More Subscribers—Only 280 on Honor Roll

DEATH TOLL AND SICK LIST IS ONE OF THE HEAVIEST ON RECORD

Practically Every State in the Union is Affected; Innumerable Cases in the County

Several Funerals Will be Held Tomorrow—All Public Meetings Have Been Cancelled—College is Under Martial Law

The influenza scourge, which has spread to 43 states, became so threatening in Addison county this week that practically every town put the ban on all public meetings, closing movies, schools and churches until the epidemic shall abate.

In Middlebury the ruling was made by Dr. Eddy last Saturday, and in consequence the stop of the Liberty Loan War Relic train on Monday which many persons had looked forward to was cancelled, the schools were closed all the week, the Boys and Girls county exhibit was postponed indefinitely, and every public gathering called off. At Middlebury college the grounds were put under armed guard, to keep all the boys confined to the campus, and all day and all night the patrols are alert toward enforcing the quarantine. Even the village streets have been patrolled by college men with scouts, making doubly certain that no one escapes the guard on the hill.

While the number of cases at the college have decreased, the epidemic has gained a strong foothold in the village and scores of homes are placarded with the red sign "Influenza." The death toll, made up both of uncertainty cases and deaths from other causes, has been one of the heaviest on record, and up to noon today had included the following deaths for the week:

Mrs. Lewis Hanfield Dead

Mrs. Phillis Hanfield, wife of Lewis J. Hanfield of South street, died Wednesday afternoon at 4:30 o'clock at the Fanny Allen hospital in Winooski, as the result of an illness which became acute last July. Five weeks ago she was taken to the hospital for an operation to relieve a condition of the gall bladder. Her condition, however, did not improve and she gradually failed. Mr. Hanfield went to Burlington Wednesday morning and was with his wife when she died.

Mrs. Hanfield was born in East Middlebury 47 years ago, the daughter of John and Phylinan (Lashua) Boudreau. Beside her husband, she is survived by one son, Paul, aged 12 years, and by two brothers and five sisters. The brothers are Napoleon Boudreau of Middlebury and George Boudreau of Cloverville, N. Y., and the sisters are Mrs. Samuel Lavers of Granville, N. Y., Mrs. John Walsh of Bennington, Mrs. Harry McClellan of St. Albans, Mrs. Daniel McCormack of Arlington, and Mrs. Nevy Dorion of Middlebury.

The funeral will be held at St. Mary's Catholic Church at 9 o'clock tomorrow morning, the Rev. T. J. Leonard officiating. The burial will be in the Catholic cemetery.

Fig. 9: *Middlebury Register*, October 4, 1918, page 1

Daily Newspapers Case Study

I. INTRODUCTION

This case study initially sought to answer two content-related questions using data mining outputs from 15 daily newspapers from the Library of Congress's *Chronicling America* database and one newspaper from the *Peel's Prairie Provinces* database. Our questions related to the outputs for each individual newspaper as well as for all 16 papers in the aggregate, specifically:

- How did newspapers report on influenza nationally, regionally, locally?
- What characterizes reporting on the epidemic during the peak of reporting (identified as the period when most word clusters are about influenza)?

Table 3. List of Daily Newspapers

Newspaper Title	Location
The Evening World *New York Tribune* *The Sun (NY)*	New York, NY
Evening Public Ledger	Philadelphia, PA
The Washington Times *The Washington Herald*	Washington, DC
The Evening Missourian	Columbia, MO
El Paso Herald	El Paso, TX
The Bemidji Daily Pioneer	Bemidji, MN
The Bismarck Tribune	Bismarck, ND
Bisbee Daily Review	Bisbee, AZ
Rogue River Courier	Grants Pass, OR
The Evening Herald	Albuquerque, NM
Tulsa Daily World	Tulsa, OK
The Ogden Standard	Ogden, UT
The Morning Bulletin	Edmonton, AB, CA

Our analysis, however, repeatedly led us to consider questions relating to methods. While some of the discussion provided here includes answers to the two content-related questions above, most of our analysis concerns methods. Specifically, we focus on how reporting at the aggregate level differs from reporting in individual newspapers in terms of what is present in the word clusters. That is, we explore how the scale of the data mining affects the outputs generated by the topic modeling and segmentation algorithm as well as their interpretation.

II. METHODS

After extracting text chunks with the search terms *influenza*, *flu*, *grippe*, and *epidemic*, we ran the topic modeling and segmentation algorithm (see **Project Methods** section in this report for a detailed discussion on data extraction, pre-processing, and processing) on the 16 daily newspapers listed in **Table 3**, both individually and collectively ("all 16 papers"). The data set for this case study was by far the largest in our research project and included almost 21,000 pages from newspapers in *Chronicling America* that contain the search terms identified above. We selected these papers based on several

Fig. 10: Tag Cloud Numbering Conventions

criteria: representation of geographic diversity, daily publication, urban circulation, and availability of digitized versions of the newspapers. The period of analysis was from January 1, 1918, through December 31, 1919. Clusters within a segment are numbered from left to right and top to bottom, as indicated in **Figure 10**.

Fig. 11: Tag Clouds with *iii* Character Set

III. DISCUSSION

Analysis of "All 16 Papers"

There are several interesting phenomena that occur in the "all 16 papers" output. In this discussion, we focus on three. First, a set of characters, *iii*, shows up in many of the segments (**Figure 11**). It initially appeared that this could be attributed to the bad OCR in Albuquerque's *The Evening Herald* (see discussion below); however, *iii* actually shows up in several other newspapers as well. Some version of the *iii* character set appears in 12 of the 23 segments in the "all 16 papers" output. Second, there are many clusters in the segments that contain advertisements. Almost every segment—before, during, and after the epidemic—contains advertisements that include such terms as *calomel*, *quinine*, and *eucalyptus*. Finally, in the 10/9/1918-10/23/1918 segment, the peak of influenza reporting in the "all 16 papers" output, there is a cluster that contains terms suggestive of football (*football*, *college*, *meet*, *play*, *school*, *team*). **Figure 12** is an example of sports reporting from the *Evening Public Ledger* that took place during the peak of the epidemic. As the analysis below will show, however, reporting on football at the height of the influenza was not a national phenomenon—despite its prominence in the "all 16 papers" 10/9/1918-10/23/1918 segment. On the contrary, the outputs for nine of the newspapers do not contain terms related to football—either because the paper did not report on

GRID GAMES ON "INTERAC" LIST WILL BE PLAYED

Delay Caused by Epidemic Will Not Affect Contests in November

Although rumors to the effect that the intercademic football schedule would have to be canceled because of the shattered condition of the teams have been as plentiful as those of the Kaiser's abdication, they appear to be without foundation. Officials of the league contend that the November games undoubtedly will be played although all the October contests may be called off.

Coach Merritt, of Penn Charter, has canceled the game with Swarthmore Prep scheduled for tomorrow and he also has called off practice for this week.

If the health authorities permit the school to be opened on Monday, practice will be re-started immediately in preparation for the game with St. Luke's, dated for next Friday.

Chestnut Hill, Central High, Germantown Academy, Episcopal Academy, Wodbury High and Chester High all have had their squads practicing daily despite the fact that no games have been played.

Chestnut Hill is taking considerable interest in football affairs and should have a good eleven. It was reported that the Hillers were suffering from the epidemic, but such is not the case. The suburbanites are making preparations for their battle with Episcopal Academy.

Fig. 12: *Evening Public Ledger*, October 17, 1918

football in the context of the epidemic or football-related terms in the context of the epidemic were not frequent enough to show up in the topic modeling and segmentation algorithm's outputs.

Analysis of Individual Papers

Table 4: Summary of Characteristics of Topic Modeling and Segmentation Outputs

Paper	Peak of Reporting	"iii" Text	Game	Reporting Focus	Characterization of Public Health Interventions	Ads
All 16 Papers	10/9/1918-10/23/1918	Yes	Yes	Some international prior to epidemic; primarily national (*washington, camp, home*) during and after epidemic; few mentions of location	No names; general references (*public, health, board, government, department, district*).	Many
The Evening World	9/17/1918-11/2/1918	Yes	Yes	Primarily national	general references	Many
The New York Tribune	9/26/1918-11/29/1918	No	Yes	National and international prior to epidemic; localized during epidemic (*york, manhattan, brooklyn*); national afterward	General references; *park* may refer to William Park, chief of NYC's Department of Health laboratory	Some
The Sun	10/3/1918-11/28/1918	Yes	Yes	National (significant) and international reporting; reporting on camps	General terms	Not Evident
Evening Public Ledger	9/18/1918-12/5/1918	Yes	Yes	Reporting on military and camps; shifts to local reporting upon epidemic's approach	General terms; *wilson* likely refers to Woodrow Wilson, although probably in context of war	Not Evident
The Washington Times	9/20/1918-11/2/1918	No	Yes	International and national; local and regional during epidemic	General terms; W.C. Fowler, Chief Health Officer, Washington	Many
The Washington Herald	9/20/1918-11/2/1918	No	No	National; some international	General terms; W.C. Fowler, Chief Health Officer, Washington	Yes
The Evening Missourian	10/23/1918-12/18/1918	No	Yes	International and national reporting; upon epidemic's approach, more local reporting	General terms	Not Evident
El Paso Herald	9/26/1918-12/13/1918	No	Yes	Some international reporting; military reporting as epidemic approaches; local reporting during peak	General terms; Dr. W.L. Brown, local Red Cross Chairman	Yes
The Bemidji Daily Pioneer	9/28/1918-11/23/1918	No	No	Consistently local and regional reporting; some international reporting	General terms	Yes
Bisbee Daily Review	11/8/1918-1/3/1918		No	Some international and national reporting; primarily local during peak	General terms	Yes
The Bismarck Tribune	11/7/1918-11/21/1918	No	No	International reporting prior to epidemic; local reporting during peak	General terms	Yes
Rogue River Courier	9/18/1918-11/13/1918	No	No	Some international and national reporting; primarily local reporting	General terms	Not Evident
The Evening Herald	10/3/1918-11/28/1918	No	No	Mostly illegible until 7/30/1918-9/24/1918 segment. Then primarily short and three-letter words; local and regional reporting	General terms	Possibly
Tulsa Daily World	9/25/1918-10/23/1918	Yes	No	National and international reporting prior to epidemic; local reporting during epidemic	General terms	Yes
The Ogden Standard	9/18/1918-12/19/1918	Yes	No	National and international reporting	General terms	Yes
The Morning Bulletin	11/6/1918-11/20/1918	No	No	National and international reporting prior to and after epidemic; local reporting during epidemic	General terms	Not Evident

Fig. 13: Peak of Reporting, *The Morning Bulletin*

Influenza-related reporting in the selected newspapers had various characteristics. Some papers tended to report locally, regardless of the phase of the epidemic; others shifted to primarily local reporting only during the epidemic. Public health interventions also took different forms. Some newspapers emphasized the role of the authorities and public health measures, while there are few mentions of either public health officials or public health interventions in others. Additionally, some newspapers ran advertisements related to the influenza, while such advertisements did not appear with enough frequency in others for the topic modeling and segmentation algorithm to pick them up. As noted above, a set of characters—*iii*—recurs across segments in some papers, and some reference football while others do not. Finally, as noted below, the peak of reporting on the influenza in the individual newspapers does not always align with the peak identified in the "all 16 papers" output. **Table 4** summarizes key differences relating to influenza reporting for the 16 papers in this case study.

Analysis of the individual newspapers indicates that the topic modeling and segmentation algorithm is able to detect differences and shifts in the reporting of information about the epidemic related to the spread of disease across the country. There are both subtle and obvious differences in the focus of reporting across the newspapers as well as similarities. For example, football discourses, while not evident in *The Morning Bulletin* (Edmonton) or *The Evening Herald* (Albuquerque), appear in the Washington and New York papers as well as in the "all 16 papers" output. Similarly in the "all 16 papers" output, there appears to be a reporting emphasis on the personal impact of the disease (*john, wife, son, daughter*) that is also evident in *The Washington Herald* (*george, wife, home*) but not as apparent in *The Morning Bulletin* (*victim, patient, miss*) and missing in the outputs for the Albuquerque and New York papers.

Additionally, while the individual newspapers' outputs do not indicate the specifics of the public health measures undertaken, they do indicate the general measures (*quarantine, close, open, report, mask*) that were implemented, the types of authorities who were responsible for implementing public health measures (*health board, commissioner, public health officer*), and the time frames within which measures were taken. Furthermore, our analyses show that the topic modeling and segmentation output is sensitive to varying reporting characteristics. For example, **Figure 13** suggests that even during the epidemic's peak, influenza reporting in *The Morning Bulletin* was not as dominant as it was in the U.S. newspapers we examined. Only three of five clusters are clearly about influenza; in one of the three, influenza-related search terms are not even in the top 20 terms—although the cluster relates to public health measures in response to the epidemic (*district, doctor, mask, provincial, board, department*). During the peak of

reporting in each of the individual U.S. papers, however, search terms are frequent enough that they appear, in various combinations, in all five clusters. The algorithm's outputs also suggest that some U.S. newspapers produced more local reporting during the epidemic than others, different newspapers reported on different topics and to varying degrees (e.g., more reporting on football, more or less reporting on the war), and there was great variety in influenza-related advertising across newspapers.

Fig. 14: *The Bismarck Tribune* Ad

Research Findings

This case study highlights how the topic modeling and segmentation algorithm operates at various levels of analysis. For example, as noted in **Table 4** above, the peak of reporting in the "all 16 papers" output is 10/9/1918-10/23/ 1918. However, peak reporting for the individual newspapers occurs between mid-September to mid-December, and only *The Morning Bulletin* and *The Bismarck Tribune* have an approximately two-week peak reporting period. Peak reporting periods for all the other newspapers examined are 30 days or longer. With respect to these two cases, *The Morning Bulletin*'s outputs indicate what appear to be differences in reporting from the U.S. papers but which may, in fact, be attributable to the quality of data in the *Peel's Prairie Provinces* database (see the **Data Quality, Sources, and Management** section for a discussion of the impact of data quality on data mining outputs). In another case, *The Bismarck Tribune* ran a series of subscription-related advertisements that

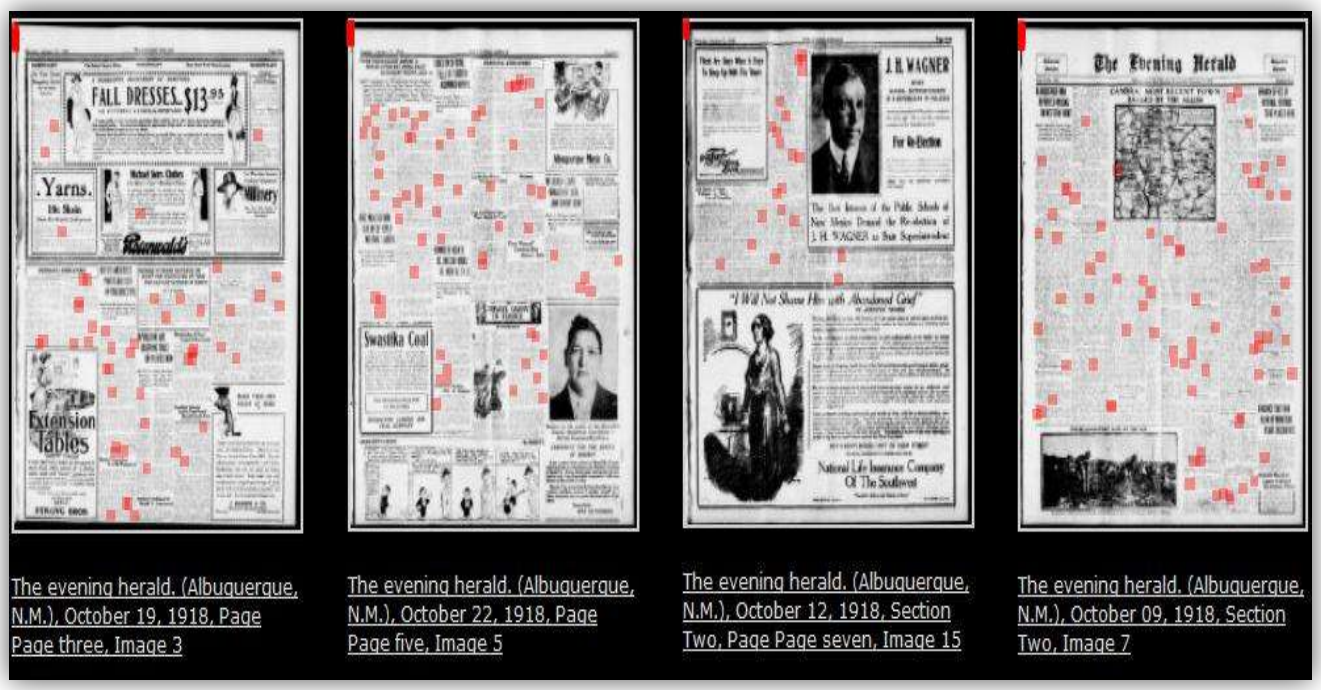


Fig. 15: Partial Results for *iii* Search in *Chronicling America*

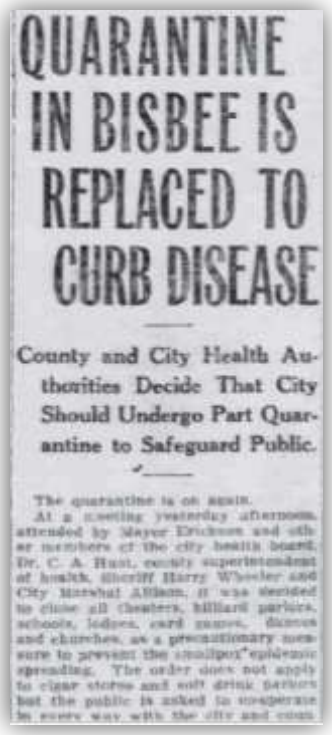
mention the flu (**Figure 14**). These advertisements clearly affect both the duration and placement in the segmentation of what appears to be—but actually is not—the peak reporting for *The Bismarck Tribune*. The topic modeling and segmentation output for this paper indicates a peak of influenza-related reporting from 11/7/1918 to 11/21/1918. A manual search of the *Chronicling America* database, however, indicates that peak influenza reporting for this paper actually occurred in late October.

Additionally, in the "all 16 papers," the *iii* character set appears in 50% of the segments. Many of these occurrences are from the Albuquerque newspaper and result from problems relating to the OCR (see the **Data Quality, Sources, and Management** section). A manual search of the *Chronicling America* database across these 16 newspapers in October 1918 confirms that, indeed, the OCR for the Albuquerque paper is of such poor quality that the *iii* character set frequently appears as a search term—as indicated in **Figure 15**. The red highlighted boxes indicate the occurrence of the *iii* character set on these pages of *The Evening Herald*. This *iii* character set then appears in the extracted text chunks with inordinate frequency, thus impacting the topic modeling and segmentation algorithm's outputs. However, the problem is not limited to *The Evening Herald*; the *iii* character set occurs across the data set when the input is unreadable, for example, when images or parts of images are read as characters (see **Figure 16**).

Although the *iii* character set appears in the top 20 terms of only six of the mined newspapers in the aggregate, it is frequent enough to appear in half of the "all 16 papers" segments, thus displacing other terms that are likely important to the interpretation of the topic modeling and algorithm's outputs.

Fig. 16: *New York Sun*, 10/21/1918

Fig. 17: Bisbee Smallpox Quarantine



QUARANTINE IN BISBEE IS REPLACED TO CURB DISEASE

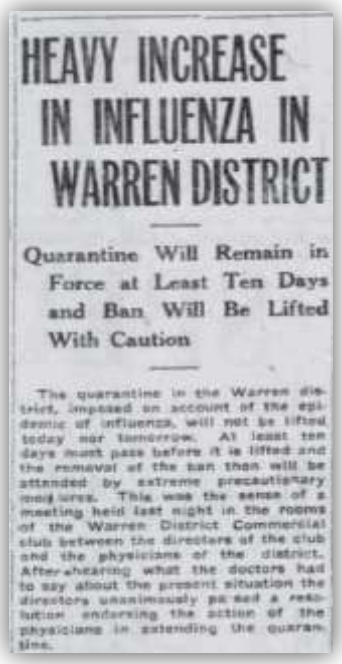
County and City Health Authorities Decide That City Should Undergo Part Quarantine to Safeguard Public.

The quarantine is on again.
At a meeting yesterday afternoon, attended by Mayor Erickson and other members of the city health board, Dr. C. A. Hunt, county superintendent of health, Sheriff Harry Wheeler and City Marshal Aitken, it was decided to close all theaters, billiard parlors, schools, lodges, card rooms, dances and churches, as a precautionary measure to prevent the smallpox epidemic spreading. The order does not apply to cigar stores and soft drink parlors but the public is asked to cooperate in every way with the city and coun-

In this way, the topic modeling and segmentation algorithm's ability to coalesce individual-level reporting into broader themes can pose problems. While eliminating this character set would likely address the problem, a simple exclusion would also eliminate text chunks containing search terms. Thus, dealing with the problem of the *iii* character set will involve either more preprocessing or a finer method of exclusion, both of which would also remove this interesting element from the outputs, an element that allows us to see all the more clearly the relation between reporting in individual papers and reporting across numerous papers.

While the *iii* character set appears to unduly influence the aggregate, there are other discourses that get lost in a larger data set. For example, *quarantine* appears in six of the 16 newspapers, but it does not appear at all in the "all 16 papers" output. These results may suggest that *quarantine* does not rise to any level of prominence in influenza-related discourses, but the outputs for *The Bisbee Daily Review* suggest otherwise. *Quarantine* appears three times in the 1/25/1918-3/22/1918 segment, three times in the 7/2/1918-8/27/1918 segment, and once each in the segments covering the period 9/12/1918-3/1/1919 as well as in the 4/7/1919-6/2/1919 segment. Clearly, this newspaper frequently covered quarantine measures before, during, and after the influenza epidemic, and the frequency with which *quarantine* appears in reporting prior to the epidemic suggests a public already familiar with this measure. Indeed, the February 8, 1918, issue of the paper ran an article emphasizing the need to quarantine *again* because of an epidemic of smallpox (**Figure 17**). Similarly, the paper ran an article on November 16, 1918, telling residents that the quarantine relating to the Spanish influenza would remain in force for at least ten more days (**Figure 18**).

On the other hand, this case study also highlights the topic modeling and segmentation algorithm's ability to identify key discourses in the aggregate that are not necessarily apparent at lower levels of analysis. The variances between the "all 16 papers" output and the outputs for the individual newspapers show that although certain terms or groupings of words may not appear in individual newspapers with sufficient frequency to present as a sustained discourse in a particular community, when examined in a larger context, the same terms and groups of words suggest a broader, recurring theme. Football-related discourses in some newspapers—*The Bemidji Daily Pioneer*, for example—were not generally in the context of the influenza, although such reporting did take place (see **Figure 19**). *The Bemidji Daily Pioneer* ran football-related articles throughout the epidemic, often on its front page; however, in the context of influenza, this reporting was so infrequent that football-related terms are not among the top 20 terms represented in the tag clouds in the topic modeling and segmentation algorithm's output. Nonetheless, as



HEAVY INCREASE IN INFLUENZA IN WARREN DISTRICT

Quarantine Will Remain in Force at Least Ten Days and Ban Will Be Lifted With Caution

The quarantine in the Warren district, imposed on account of the epidemic of influenza, will not be lifted today nor tomorrow. At least ten days must pass before it is lifted and the removal of the ban then will be attended by extreme precautionary measures. This was the sense of a meeting held last night in the rooms of the Warren District Commercial club between the directors of the club and the physicians of the district. After hearing what the doctors had to say about the present situation the directors unanimously passed a resolution endorsing the action of the physicians in extending the quarantine.

Fig. 18: Bisbee Influenza Quarantine

evidenced by its appearance in the "all 16 papers" output, football-related reporting in the context of or in proximity to influenza discourses was a key characteristic of newspaper reporting at the national level during the epidemic. Similarly, influenza-related advertising, while not evident in all the individual newspapers' outputs, was clearly a characteristic of newspaper reporting during the 1918 Spanish influenza epidemic (**Figure 20**).

Fig. 19: *The Bemidji Daily Pioneer*, October 18, 1918

STATE AUTHORITIES CLOSE ILLINOIS TO LAST NOTCH

Chicago, Oct. 18.—All public prohibited by the state authorities gatherings throughout Illinois were today. In explanation of the proclamation. Dr. St. Clair Drake said: "This order means that there will be no football games Saturday, no wrestling bouts and no public entertainment that will draw crowds of people and this includes club meetings, meetings of all kinds, dinners, luncheons, in fact everything that is not absolutely essential to war work."

A keyword search of the 15 daily papers from *Chronicling America* from September through December 1918 comparing *influenza* and *football* within 10 words of each other with *influenza* and *quarantine* yields more pages for *quarantine*. Yet in the text chunks extracted for analysis, there are more mentions of *football* than *quarantine*, which is why it shows up in the "all 16 papers" outputs. Looking at the aggregate data mining outputs, one would think that quarantine was not utilized during the pandemic, but it was, and it was reported on in many newspapers. For whatever reason, *football* was mentioned more frequently, a phenomenon that demands its own cultural analysis.

IV. CONCLUSION

The most important findings of this study, to date, involve the differences that scale makes in analyzing data mining outputs of news reporting. The outputs for the daily papers are processed with far more data than the titles in the Weekly Newspapers Case Study, but for all the increase in the data inputs, the outputs for all the papers together still significantly differ from the outputs for the individual papers. It is clear that some papers influence the output of the "all 16 papers" more than others, either due to poor OCR quality or larger text chunks. Some reporting that is evident (albeit somewhat sporadically) at the local level rises to significance at the national level as well; however, other reporting, does not appear in the "all 16 papers" output, even though the activities reported on were prevalent across the country.

The anomalies catch our interest. The *iii* character set motivates the researcher to discover the origin of this oddity, as does the word cluster on *football* in the outputs for the "all 16 papers." What we expect to see is sometimes absent from the outputs, as is the case with reporting on quarantine measures as compared to the reporting on football. These anomalies (or seeming anomalies) take us back to the newspapers as original sources for an explanation. Thus, analysis of topic modeling outputs for an aggregate of sources must take into account the possibility that reporting at the local level will not necessarily be commensurate with the outputs of these aggregate data. That is, the aggregate outputs will not necessarily be *representative* of local reporting.

INFLUENZA?

Keep your bowels in good condition with Rexall Orderlies and use Rexall Throat Gargle.

DON'T BE CARELESS

BALL'S DRUG STORE

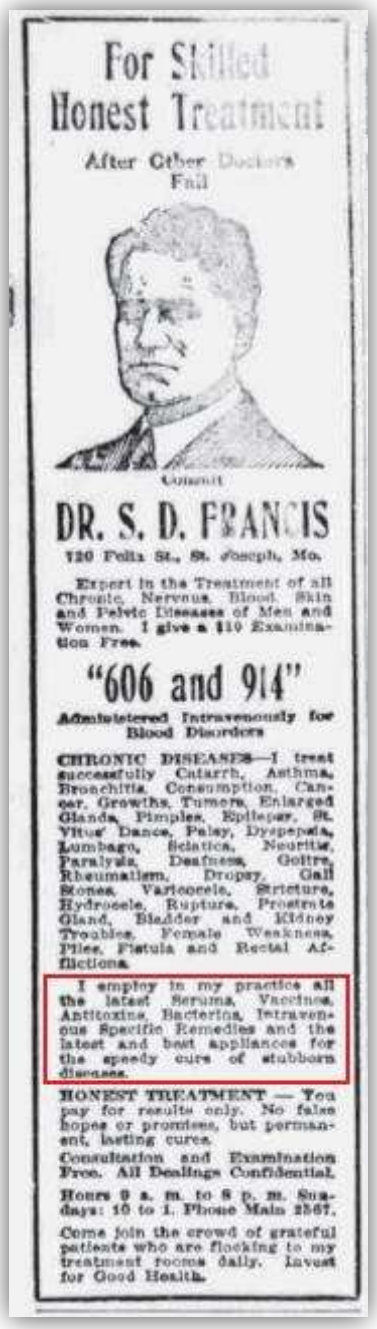
Main St. THE REXALL STORE Phone 307

Fig. 20: *Bisbee Daily Review*, October 10, 1918, page 2

Vaccination-Visualization Case Study

I. INTRODUCTION

To adopt a "big data" approach to studying the 1918 influenza pandemic, 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 [6, 7, and 8].



For Skilled Honest Treatment

After Other Doctors Fail

DR. S. D. FRANCIS

120 Felix St., St. Joseph, Mo.

Expert in the Treatment of all Chronic, Nervous, Blood, Skin and Pelvic Diseases of Men and Women. I give a $10 Examination Free.

"606 and 914"

Administered Intravenously for Blood Disorders

CHRONIC DISEASES—I treat successfully Catarrh, Asthma, Bronchitis, Consumption, Cancer, Growths, Tumors, Enlarged Glands, Pimples, Epilepsy, St. Vitus' Dance, Palsy, Dyspepsia, Lumbago, Sciatica, Neuritis, Paralysis, Deafness, Goitre, Rheumatism, Dropsy, Gall Stones, Varicocele, Stricture, Hydrocele, Rupture, Prostate Gland, Bladder and Kidney Troubles, Female Weakness, Piles, Fistula and Rectal Afflictions.

I employ in my practice all the latest Serums, Vaccines, Antitoxins, Bacterins, Intravenous Specific Remedies and the latest and best appliances for the speedy cure of stubborn diseases.

HONEST TREATMENT — You pay for results only. No false hopes or promises, but permanent, lasting cures.

Consultation and Examination Free. All Dealings Confidential.

Hours 9 a. m. to 6 p. m. Sundays: 10 to 1. Phone Main 2567.

Come join the crowd of grateful patients who are flocking to my treatment rooms daily. Invest for Good Health.

Fig. 21: *St. Joseph Observer,* December 20, 1919, page 3

This vaccination-visualization study involves, first, analysis of the content of vaccination-related newspaper reporting before, during, and after the pandemic.[2] 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. Numerous vaccines were created during the deadly second wave (**Figure 28**), but there are divergent views about their efficacy [9, 10]. Nevertheless, there was significant reporting on vaccines during this period, perhaps because they represent developing confidence "that medical research could provide widespread benefits" [11]. After the war, vaccines became an element of medical practice evident in advertising, in part demonstrating a physician's modern methods of treatment (**Figure 21**).

The second and more significant aspect of our research concerns the conventions that underlie the methods of both data extraction and data visualization practices. We did not set out to ask or answer any questions about visualization in the research project overall. 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 project expanded to include the analysis of visualization conventions as they relate to data mining outputs generally. We seek to better understand 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. The general thrust of this particular study is not to consider how a given visualization can be modified to meet a user's specific requirements, but to move toward an analysis of

the inevitably persuasive effects of any visualization choice. In the discussion presented here, we focus almost exclusively on the methods-related issues of the study.

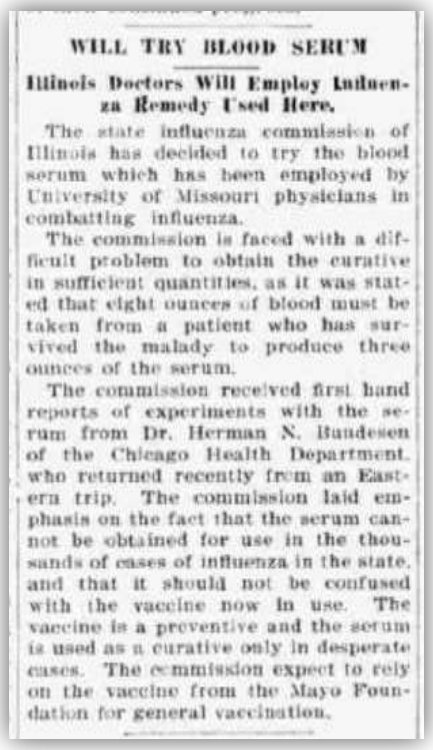
II. METHODS

We applied the integrated topic modeling and segmentation algorithm to 90 titles from January 1, 1918, to December 31, 1919, focusing on vaccination. These were the titles for 1918 and 1919, both daily and weekly, available in the *Chronicling America* database when we began the case study. To begin, we extracted text chunks based on our search terms: the root terms *vaccin* and *inoculat*, as well as *vaccination*, *vaccine*, and *inoculation*. We ran two extractions, the first excluding text chunks containing 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*. We labeled the outputs of the first extraction as "no blackleg/no cholera" and the outputs of the second extraction as "no unwanted terms."

A note about the methodological implications of excluding *serum*: The decision to exclude *serum* was based on an initial observation that it seemed overrepresented in the data output. Technically, a serum is different from a vaccine. A serum is created by separating blood into its solid and liquid components so that antibodies to microbes can be utilized for their immunizing properties. However, *serum* at the time was sometimes used as a synonym for *vaccine*, although not all the time, as we discovered. **Figure 22** is an example of the distinction between *serum* and *vaccine* being upheld in reporting, while **Figures 23** and **27** are examples of the use of the two terms rather indiscriminately.

Our decision to exclude the term *serum* significantly affected the content of the data outputs. In excluding the texts chunks that included *serum*, we lost some reporting on the influenza vaccine, significantly, reporting on the Rosenow vaccine from the Mayo Clinic, which was used in Chicago and widely reported

Fig. 22: *Evening Missourian*, October 22, 1918, page 1



WILL TRY BLOOD SERUM

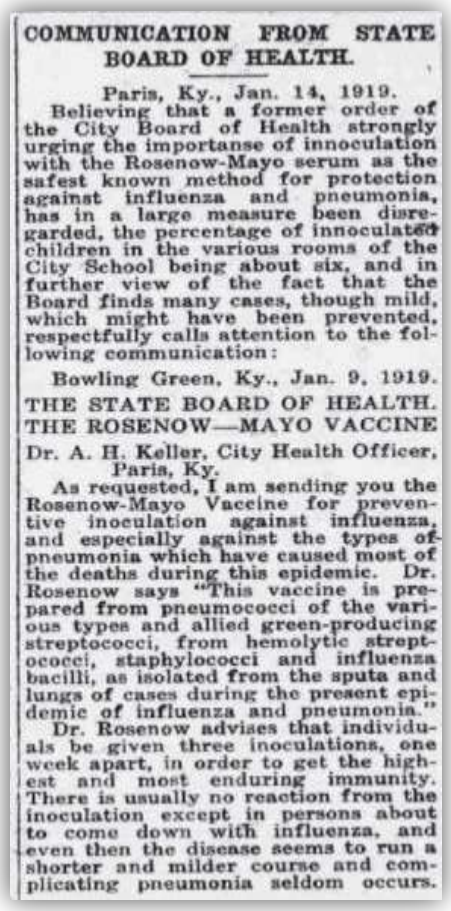
Illinois Doctors Will Employ Influenza Remedy Used Here.

The state influenza commission of Illinois has decided to try the blood serum which has been employed by University of Missouri physicians in combatting influenza.

The commission is faced with a difficult problem to obtain the curative in sufficient quantities, as it was stated that eight ounces of blood must be taken from a patient who has survived the malady to produce three ounces of the serum.

The commission received first hand reports of experiments with the serum from Dr. Herman N. Bundesen of the Chicago Health Department, who returned recently from an Eastern trip. The commission laid emphasis on the fact that the serum cannot be obtained for use in the thousands of cases of influenza in the state, and that it should not be confused with the vaccine now in use. The vaccine is a preventive and the serum is used as a curative only in desperate cases. The commission expect to rely on the vaccine from the Mayo Foundation for general vaccination.

Fig. 23: *Bourbon News*, January 17, 1919, page 8



COMMUNICATION FROM STATE BOARD OF HEALTH.

Paris, Ky., Jan. 14, 1919.

Believing that a former order of the City Board of Health strongly urging the importanse of innoculation with the Rosenow-Mayo serum as the safest known method for protection against influenza and pneumonia, has in a large measure been disregarded, the percentage of innoculated children in the various rooms of the City School being about six, and in further view of the fact that the Board finds many cases, though mild, which might have been prevented, respectfully calls attention to the following communication:

Bowling Green, Ky., Jan. 9, 1919.

THE STATE BOARD OF HEALTH.

THE ROSENOW—MAYO VACCINE

Dr. A. H. Keller, City Health Officer, Paris, Ky.

As requested, I am sending you the Rosenow-Mayo Vaccine for preventive inoculation against influenza, and especially against the types of pneumonia which have caused most of the deaths during this epidemic. Dr. Rosenow says "This vaccine is prepared from pneumococci of the various types and allied green-producing streptococci, from hemolytic streptococci, staphylococci and influenza bacilli, as isolated from the sputa and lungs of cases during the present epidemic of influenza and pneumonia."

Dr. Rosenow advises that individuals be given three inoculations, one week apart, in order to get the highest and most enduring immunity. There is usually no reaction from the inoculation except in persons about to come down with influenza, and even then the disease seems to run a shorter and milder course and complicating pneumonia seldom occurs.

on. **Figure 24** is a representative article on Chicago's use of the Rosenow vaccine, and **Table 5** demonstrates a cluster that includes this kind of reporting in the "no blackleg/no cholera" extraction.

Table 5: Word Frequency List for "No Black Leg/No Cholera" Extraction 9/28/1918-11/23/1918

Topic: 1	Topic: 2	Topic: 3	Topic: 4	Topic: 5
vaccine	vaccine	serum	vaccine	vaccine
cur	influenza	army	influenza	serum
disease	serum	pneumonia	city	county
life	health	vaccine	disease	influenza
administer	dr	war	patient	health
herb	report	camp	inoculation	dr
private	physician	public	treatment	grippe
serum	pneumonia	day	serum	germ
safe	mayo	epidemic	spanish	people
remedy	department	health	health	bureau
ulcer	chicago	medical	germ	call
medicine	death	person	specific	disease
bacteria	receive	officer	dr	cent
charge	supply	influenza	preventive	confidence
prevent	commission	surgeon	effect	successfully
blood	announce	service	lake	furnish
old-fashion	prevent	treatment	free	successful
kidney	city	result	mis	price
low	spanish	follow	injection	time
liver	epidemic	test	measure	contract

The "no unwanted terms" outputs, which are discussed in this report, do not include a word cluster on the Rosenow vaccine or Chicago; however, we see in these outputs more mentions of *smallpox*, which appears to have a negative association with *serum*.

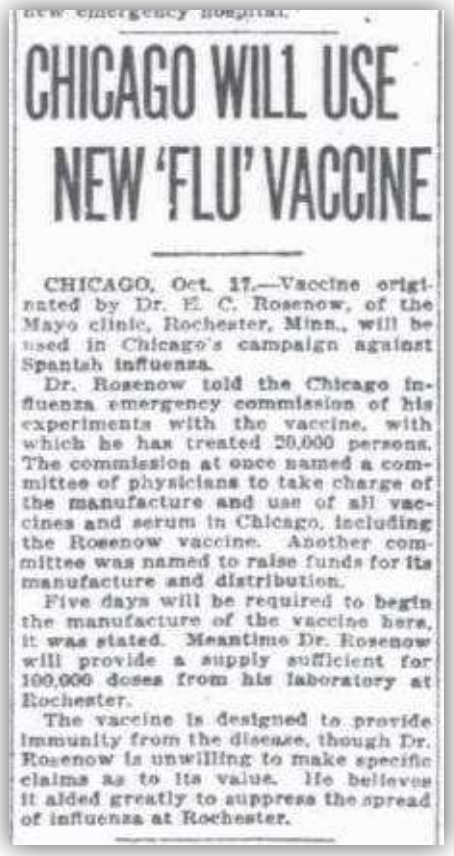
The visualizations of the data that we discuss here all use the same data outputs from the topic modeling and segmentation algorithm, specifically, the "no unwanted terms" outputs. In these outputs, as with all the outputs discussed in this report, there was no norming across or within segments, only within individual word clusters.

III. DISCUSSION

We begin with three methods-related questions:

- How do conventions in reading affect the analysis of visualizations of topic modeling and segmentation outputs?
- What is the rhetorical impact on the interpreter of each specific visualization of topic modeling and segmentation outputs?
- What forms of tacit knowledge are necessary to appropriately read and interpret data mining results using various forms of visualization?

Initially, we conducted a literature review of approaches to visualizations of big data in rhetorical studies [12]. While there is a wealth of scholarship related 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.



CHICAGO WILL USE NEW 'FLU' VACCINE

CHICAGO, Oct. 17.—Vaccine originated by Dr. H. C. Rosenow, of the Mayo clinic, Rochester, Minn., will be used in Chicago's campaign against Spanish influenza.

Dr. Rosenow told the Chicago influenza emergency commission of his experiments with the vaccine, with which he has treated 20,000 persons. The commission at once named a committee of physicians to take charge of the manufacture and use of all vaccines and serum in Chicago, including the Rosenow vaccine. Another committee was named to raise funds for its manufacture and distribution.

Five days will be required to begin the manufacture of the vaccine here, it was stated. Meantime Dr. Rosenow will provide a supply sufficient for 100,000 doses from his laboratory at Rochester.

The vaccine is designed to provide immunity from the disease, though Dr. Rosenow is unwilling to make specific claims as to its value. He believes it aided greatly to suppress the spread of influenza at Rochester.

Fig. 24: *Washington Times*, October 17, 1918, page 3

One important distinction raised by our review of the literature is the rationalist/social constructionist divide. 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, emphasizing design conventions and their naturalization. From the social constructionist perspective, normalizing practices often result in an uncritical and unquestioning acceptance of visualizing conventions without an exploration of their persuasive effects. Rhetoricians will almost always fall into the social constructionist camp, emphasizing the conventions that underlie both the methods of data extraction and data visualization practices. 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 naturalizations conflicted with the rationalist approach of our data mining colleagues, who wanted to adjust the algorithms in order to display the data better.

The three visualizations we used each represent topic modeling and segmentation outputs in useful ways. Because of the nature of the word clusters to index the topics across all visualizations, the visualizations are, indeed, more similar than different and lead to interpretations that are more similar than different. However, subtle differences in interpretation result from the distinct visualization methods, especially with respect to how the researcher approaches the data.

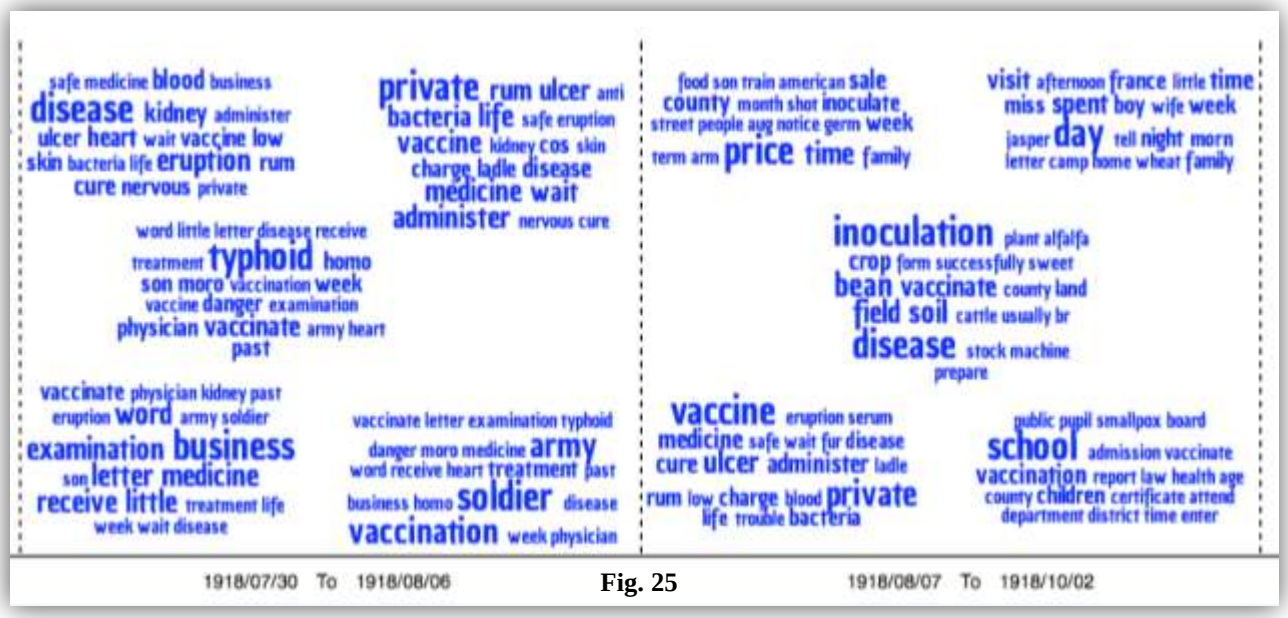


Fig. 25

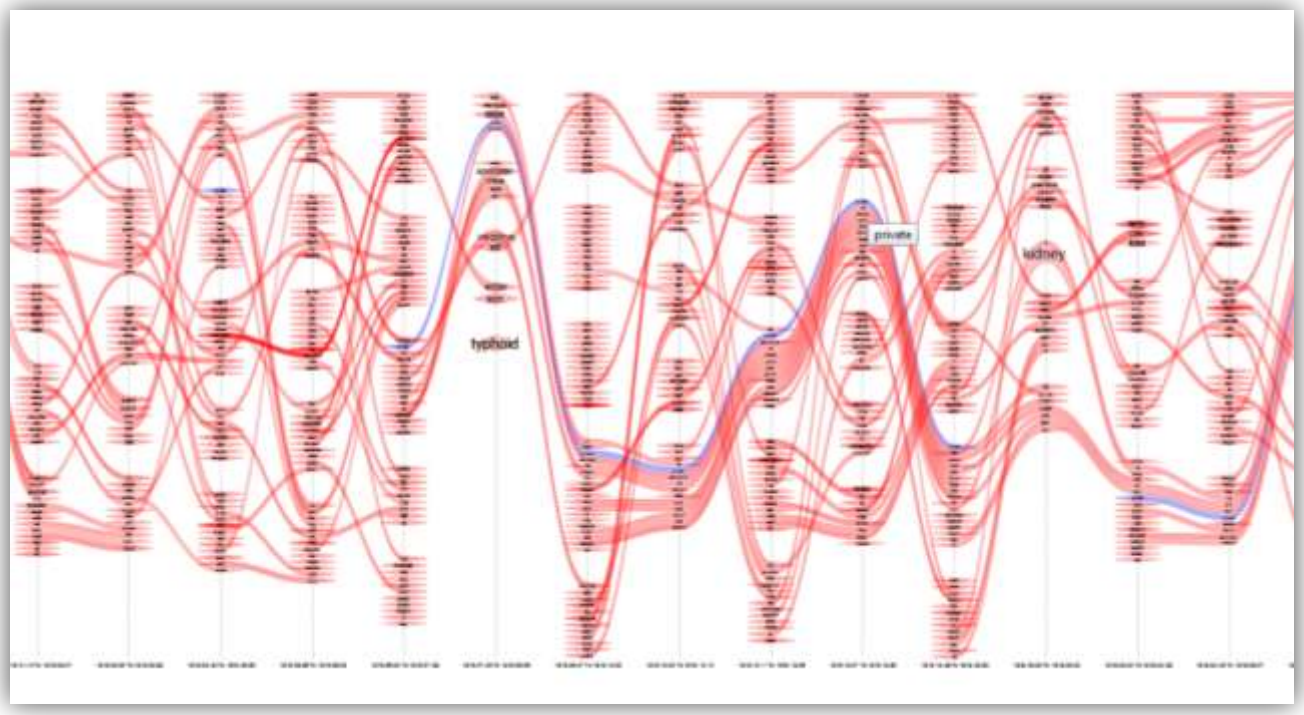
Tag clouds

Our first visualization choice represents 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 and words are arrayed in a non-purposeful manner (**Figure 25**). Tag clouds facilitate an analysis that is bounded and oriented to narrative. Within the segments, each tag cloud seems to tell a particular story. The stories are not linked linearly; rather, they are proximal only insofar as they are located together within a segment. The researcher is motivated to search within each cloud for

thematically related words and to identify the topic that the cluster of words in the cloud represents. That is, the researcher works to narrate a coherent story about the reporting represented in each tag cloud.

Analyzing the tag cloud output on reporting on vaccination and vaccines prior to the flu pandemic suggests that reporting before the pandemic was different from reporting after, at least until the end of 1919. This conclusion is determined from the types of words that are represented together in the clouds as well as the words represented across each segment in the entire time period. Thus, initially in 1918, *vaccine* is linked to *typhoid* and *smallpox* and mentioned in the context of school and the military. In the fall of 1918, *influenza* gains prominence in comparison to *smallpox* and *typhoid*, but it is subsequently supplanted by a return to reporting on *typhoid* and *smallpox*, with seemingly sporadic reporting on *influenza* through 1919.

Fig. 26: ThemeDelta Visualization with Trendline for *Private* Highlighted in Blue



ThemeDelta

ThemeDelta, our second visualization choice, represents topic modeling results over time (**Figure 26**). In this form of visualization, word clusters are arrayed on vertical lines representing the temporal segments. A feature called *trendlines*, variable width lines that branch and merge, follow the term through contiguous segments, demonstrating the persistence or discreteness of particular word clusters over time. The thickness of the line renders the frequency of the term within its cluster.

The ThemeDelta visualization highlights the flow of discrete words across the entire period of analysis. ThemeDelta highlights how a particular word moves through groupings of words across segments. In addition, ThemeDelta draws the eye across the visualization to emphasize the temporal aspect of words. Interpretations of this form of visualization thus trace connections

across time but de-emphasize narrative since the trendlines become the focus, not the clusters of words. The researcher starts with a macro-level look at the trendlines and flows that ThemeDelta emphasizes, then moves into segments, which are vertically arrayed, for more specific analysis of the word clusters.

The ThemeDelta visualization of vaccine-related reporting seems to indicate a more particular and granular form of reporting prior to early August 1918. 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 a one-week early August segment as a large and seemingly discrete topic. There is a very thick band of lines across the segments during the period from early August 1918 through February 1919, after which the patterns resemble those prior to August 1918. The lines noticeably shift again, beginning in the May through July 1919 segment. The most prominent banding in the subsequent sections develops here and dissipates by the mid-September 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.

TABLE 6. 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

A more detailed view of the ThemeDelta visualization shows that vaccine discourses through the first half of 1918 relate variously to smallpox and typhoid in the context of schools and the military, but there are many breaks in the lines. These patterns suggest that reporting on vaccination during and across these segments is sporadic rather than continuous. The visible shift beginning in the June through July 1918 segment initially centers on *typhoid*. However, the patterns further shift during the period the influenza spread across the country and peaked in fall 1918. Reporting beginning in the fall 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.

Word Frequency Lists

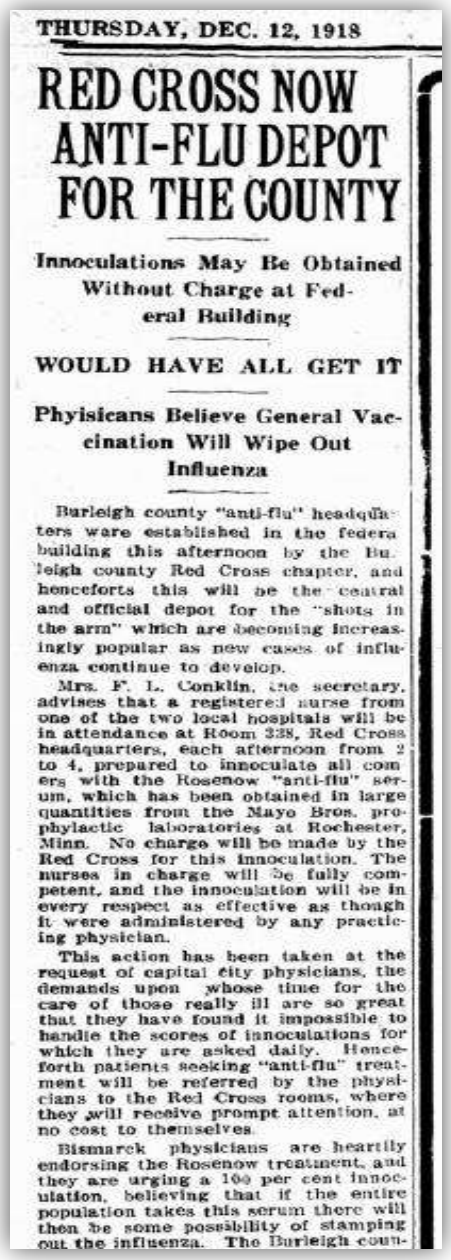
Our third visualization choice is the word frequency list, rendered with standard word processing software. In a word frequency list, words in the cluster are listed in order of frequency and the clusters in a segment are arrayed in columns. Of the three visualizations, word frequency lists are the most indexical (**Table 6**). They encourage interpretation that is based on hierarchical relationships of words. It is very easy to see the most frequent words in the distribution, as they

are at the top of the list. Primary emphasis, then, is on the position of words in the lists, not necessarily on the lists as the context within which words appear. As a result, and especially if the lists have been color-coded by category, the word frequency lists emphasize discrete words in the context of other words of a same or different kind. Many symptom words in conjunction with many treatment words, for example, suggest an advertisement, while many administrative or bureaucratic words with treatment words, in another example, suggest public health measures.

Word frequency lists 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 cluster. Word frequency lists also allow the researcher to manipulate the lists more directly—grouping words, color coding, or otherwise affecting the visualization in the context of interpretation and analysis. Unlike tag clouds, word frequency lists decrease misunderstanding of the relations between words (in tag clouds, word placement is not purposeful but might seem so to the reader). Additionally, word frequency lists facilitate intrasegment analysis; patterns within a segment emerge since the researcher is able to line the topics up and compare them across the segment. Indeed, the researcher is motivated to search the lists in a given segment for particular terms and to identify these terms across clusters in the segment.

On the other hand, if the word frequency lists are rendered in a typical word processing program so that only one or two segments fit on a page, comparison across segments is more difficult. Changes in reporting over time are most easily identified in word frequency lists when they have been color coded by category. At that point, full segments can be analyzed in terms of the distribution of words of different categories and segments can be compared against one another according to differences in distribution.

The different ways these visualizations depict word frequency is another example of how different representations of the same data encourage different interpretations. Tag clouds indicate frequency by the size of the word in a cluster, while ThemeDelta indicates frequency by the thickness of a word's trendline within a cluster, and word frequency lists are hierarchically organized with the most frequent word in the cluster at the top of the list. Without norming across and within segments, relative frequency can only be identified within word clusters. As a result, in our data outputs, none of the representations indicate words' relative frequency between clusters and across segments—that output constraint is a result of decisions made about the algorithm (i.e., an input constraint). Furthermore, in tag clouds and ThemeDelta, it is impossible to tell the relative frequency of words of like size within a cluster. Word frequency lists, on the other hand, clearly indicate relative frequency hierarchically but do not have a mechanism for more finely grained representation unless we attach the actual numerical value of each word, which makes the representation of the lists unwieldy.



THURSDAY, DEC. 12, 1918

RED CROSS NOW ANTI-FLU DEPOT FOR THE COUNTY

Innoculations May Be Obtained Without Charge at Federal Building

WOULD HAVE ALL GET IT

Phyisicans Believe General Vaccination Will Wipe Out Influenza

Burleigh county "anti-flu" headquarters were established in the federa building this afternoon by the Burleigh county Red Cross chapter, and henceforts this will be the central and official depot for the "shots in the arm" which are becoming increasingly popular as new cases of influenza continue to develop.

Mrs. F. L. Conklin, the secretary, advises that a registered nurse from one of the two local hospitals will be in attendance at Room 338, Red Cross headquarters, each afternoon from 2 to 4, prepared to innoculate all comers with the Rosenow "anti-flu" serum, which has been obtained in large quantities from the Mayo Bros. prophylactic laboratories at Rochester, Minn. No charge will be made by the Red Cross for this innoculation. The nurses in charge will be fully competent, and the innoculation will be in every respect as effective as though it were administered by any practicing physician.

This action has been taken at the request of capital city physicians, the demands upon whose time for the care of those really ill are so great that they have found it impossible to handle the scores of innoculations for which they are asked daily. Henceforth patients seeking "anti-flu" treatment will be referred by the physicians to the Red Cross rooms, where they will receive prompt attention, at no cost to themselves.

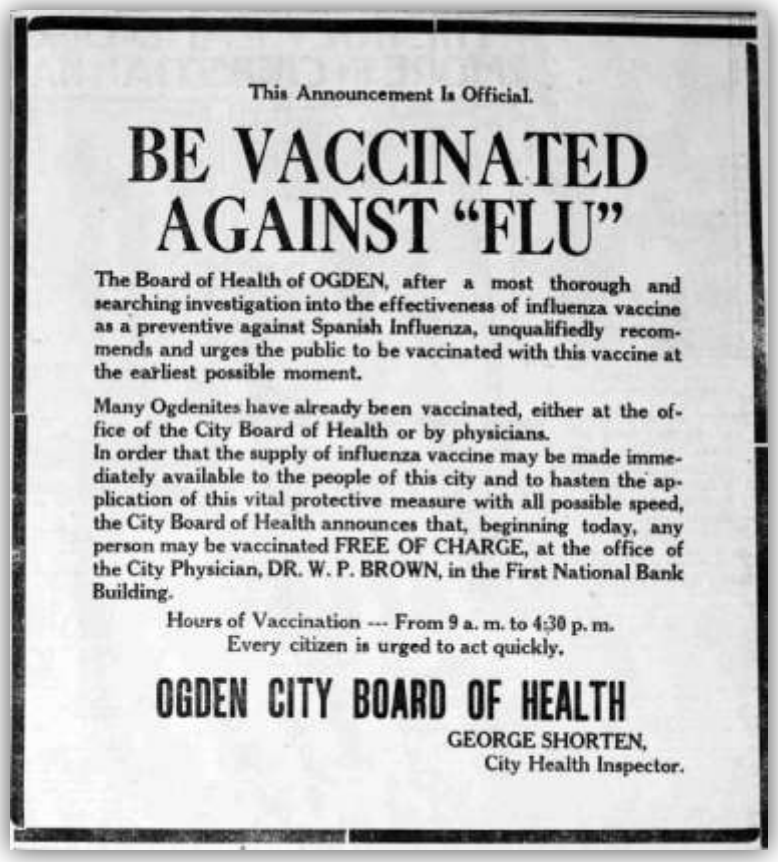
Bismarck physicians are heartily endorsing the Rosenow treatment, and they are urging a 100 per cent innoculation, believing that if the entire population takes this serum there will then be some possibility of stamping out the influenza. The Burleigh coun-

Fig. 27: *Bismarck Daily Tribune*, December 12, 1918, page 5

IV. CONCLUSION

Forms of visualization such as ThemeDelta hold great promise because 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 (by placing the cursor over it) to see it trending across the segments in various word clusters.

The tacit knowledge necessary to interpret ThemeDelta in its current form is crucial to interpretive accuracy, however, as reading conventions generally encourage interpreters to compare trendline thickness across clusters within segments and across segments. In addition, ThemeDelta encourages a focus on the words discretely across segments, rather than the words in clusters. Yet all three of these visualizations are, essentially, words in clusters and words across segments. The analysis of each visualization tends to emphasize the indexical nature of all of them, even as word frequency lists seem more indexical than the others. In the end, all three suggest that reporting on vaccination changed during the pandemic, even as it seems to have shifted back to previous patterns afterward.



This Announcement Is Official.

BE VACCINATED AGAINST "FLU"

The Board of Health of OGDEN, after a most thorough and searching investigation into the effectiveness of influenza vaccine as a preventive against Spanish Influenza, unqualifiedly recommends and urges the public to be vaccinated with this vaccine at the earliest possible moment.

Many Ogdenites have already been vaccinated, either at the office of the City Board of Health or by physicians.
In order that the supply of influenza vaccine may be made immediately available to the people of this city and to hasten the application of this vital protective measure with all possible speed, the City Board of Health announces that, beginning today, any person may be vaccinated FREE OF CHARGE, at the office of the City Physician, DR. W. P. BROWN, in the First National Bank Building.

Hours of Vaccination --- From 9 a. m. to 4:30 p. m.
Every citizen is urged to act quickly.

OGDEN CITY BOARD OF HEALTH

GEORGE SHORTEN,
City Health Inspector.

Fig. 28: *The Ogden Standard*, December 5, 1918, page 9

To confirm and expand on these initial findings, we plan to conduct two future studies. One will more thoroughly document experienced researchers' practices as they formulate interpretations with these visualizations. The second study will collect data across multiple researchers who are relatively inexperienced with the visualizations in order to ascertain what conclusions they come to with the same data visualized differently.

Public Health Officials Case Study: Royal S. Copeland, Commissioner of Health for New York City, and the 1918 Influenza

I. INTRODUCTION

On August 17, 1918, Royal S. Copeland, Health Commissioner for New York City, responded to reports that a ship had arrived from Europe with numerous cases of the "Spanish Flu" with this reassuring statement: "We have not felt and do not feel any anxiety about what people call 'Spanish Influenza,' and we considered it so unimportant that it did not seem necessary to make a public discussion of the situation" [13]. This statement—and others equally reassuring and optimistic in the next several weeks—proved to underestimate the severity of the disease, which, by year's end, resulted in the death of more than 30,000 New Yorkers. Copeland's role in guiding public health responses to the Spanish flu in America's largest city has provoked strikingly opposite evaluations. For his critics, the reassuring statements issued at the start of the epidemic proved false and misleading in ways that potentially undermined more proactive and extensive public health measures [14, 15, 16]. Copeland's defenders, by contrast, offer a more complicated assessment of his efforts in terms of the broader context of the American public health [17, 18, 19, 20, 21, 22]. The divergent opinions about the effect of Copeland's statements notwithstanding, he was a visible figure in how health authorities shaped responses to the 1918 Spanish influenza pandemic because of his role as health commissioner in a large city.

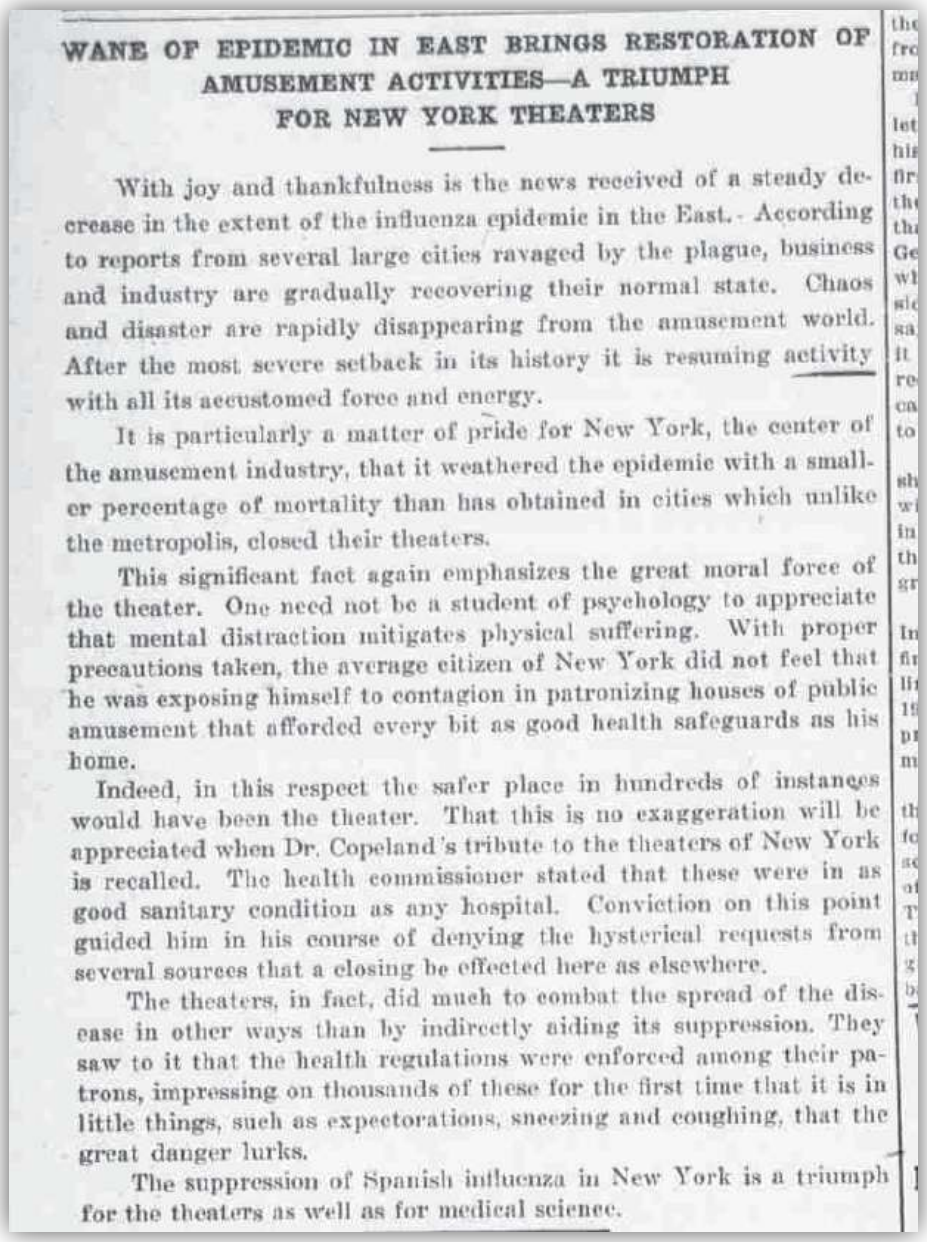
The Public Health Officials Case Study investigates Copeland's role and impact using an integrated method including topic modeling and tone classification. Although existing assessments of Copeland rely heavily on published reports about him and statements he made, mostly from the *New York Times*, this project tests out methods for a more comprehensive and systematic review of his impact locally as well as nationally. A search of *Chronicling America* for the terms *Copeland* and *influenza* produces 305 pages for August to December 1918. Of these results, 208 appear in three New York City newspapers and 97 results appear in national newspapers outside of New York City. Both numbers are revealing.

Fig. 29: *Evening World*, November 21, 1918, page 12

The former indicates the intensity of reporting within New York City on Copeland's efforts to control the influenza. The latter indicates how the example of this major urban center—one of the first cities to experience the influenza epidemic—and of this prominent health official, penetrated into the daily reporting of newspapers across the nation, as shown in a Kentucky newspaper quoting Copeland's assessment of effective public health measures (**Figure 30)**.

II. METHODS

This project uses three methods to analyze newspapers. First, we applied topic modeling and segmentation to the 90 papers initially downloaded from the *Chronicling America* collection to identify text chunks that associate *Copeland* with disease terms (*influenza, grippe, epidemic*). Second, we used tone classification on approximately 150 sentences drawn from published quotations from one newspaper, *The New York Tribune*, in the period from August 17 to December 13. A total of 25 articles were identified that have direct statements from Copeland within this time period. These sentences were classified into four categories: alarmist, warning, explanatory, and reassuring (see the **Project Methods** section in this report for explanations of each tone category). Finally, this case study uses key word searching and manual text analysis to identify ways in which Copeland's impact on public health beyond New York City can be documented.



WANE OF EPIDEMIC IN EAST BRINGS RESTORATION OF AMUSEMENT ACTIVITIES—A TRIUMPH FOR NEW YORK THEATERS

With joy and thankfulness is the news received of a steady decrease in the extent of the influenza epidemic in the East. According to reports from several large cities ravaged by the plague, business and industry are gradually recovering their normal state. Chaos and disaster are rapidly disappearing from the amusement world. After the most severe setback in its history it is resuming activity with all its accustomed force and energy.

It is particularly a matter of pride for New York, the center of the amusement industry, that it weathered the epidemic with a smaller percentage of mortality than has obtained in cities which unlike the metropolis, closed their theaters.

This significant fact again emphasizes the great moral force of the theater. One need not be a student of psychology to appreciate that mental distraction mitigates physical suffering. With proper precautions taken, the average citizen of New York did not feel that he was exposing himself to contagion in patronizing houses of public amusement that afforded every bit as good health safeguards as his home.

Indeed, in this respect the safer place in hundreds of instances would have been the theater. That this is no exaggeration will be appreciated when Dr. Copeland's tribute to the theaters of New York is recalled. The health commissioner stated that these were in as good sanitary condition as any hospital. Conviction on this point guided him in his course of denying the hysterical requests from several sources that a closing be effected here as elsewhere.

The theaters, in fact, did much to combat the spread of the disease in other ways than by indirectly aiding its suppression. They saw to it that the health regulations were enforced among their patrons, impressing on thousands of these for the first time that it is in little things, such as expectorations, sneezing and coughing, that the great danger lurks.

The suppression of Spanish influenza in New York is a triumph for the theaters as well as for medical science.

Fig. 30: *Public Ledger* (Maysville, KY), November 25, 1918

III. DISCUSSION

Topic modeling and segmentation was completed on three sets of results: first, 90 newspapers; second, just the New York City

newspapers; and third, the 90 newspapers from the *Chronicling America* database without the New York City newspapers.

Fig. 31: 9/8/1918-11/3/1918, 90 Newspapers

In the 8/17/1918-9/7/1918 segment for the 90 newspapers output, only two tag clouds appear, one that connects Copeland's name to his title and one to an article in which he recommends preventing the spread of influenza by using a handkerchief and refraining from kissing. As **Figure 31** shows, the 9/8/1918-11/3/1918 segment for the 90 newspapers output includes references to recommended actions, including asking *landlords* to provide *heat* in *apartments* and assigning *physicians* and *nurses* to *hospitals*, refers to warnings to the public (*bacillus, epidemic*, and *spanish*) and suggests public health measures (*close, public, telephone*, and *health department*).

The tag clouds from New York City newspapers demonstrate how closely Copeland's name was connected with reporting on public health measures such as *vaccine*, *quarantine*, *hospital*, and *report*, with city-specific terms such as *brooklyn*, *manhattan*, *bronx*, and *queens*, in addition to processes involved in reporting disease and measuring increases or decreases (**Figure 32**).

The national influence of Copeland's leadership can be illustrated in the tag clouds for the first segment of the national newspapers other than those in New York City. The 9/20/1918-10/4/1918 segment shown in **Figure 33** includes references to medical terms (*vaccine* (three times), *test, serum, preliminary, germ,* and *result*), the appearance of which suggests that reporting on Copeland called attention to the specific health measures being considered by the New York City Health Department and that were also discussed by health officials in the locations of these newspapers.

Fig. 32: 9/8/1918-11/3/1918, NYC Papers

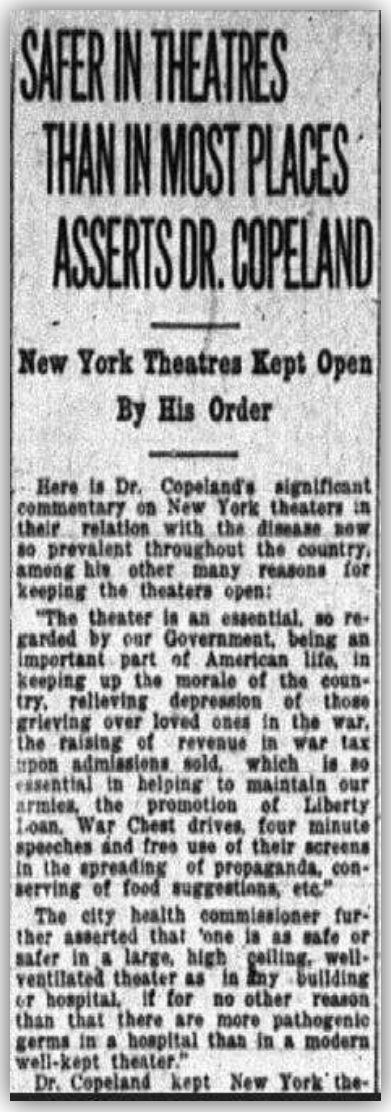
Tone classification for direct statements made by Copeland during the five months of the influenza produced the following results: alarmist, 3.3%; warning, 21.9%; explanatory, 57.6%; and reassuring, 17.2%. This distribution is similar to

Fig. 33: 9/20/1918-10/4/1918, No NYC Papers

epidemic's impact on New York City. This method suggests, at least preliminarily, the need to qualify assertions about Copeland's naïve approach to the Spanish influenza.

Further exploration of Copeland's impact nationally on the spread of information about influenza can be illustrated with key word searching accompanied by manual analysis. As indicated by front page articles in the *Bourbon News* and the *Washington Times*, the public health measures implemented in New York City were cited as examples of aggressive and effective steps to prevent the spread of influenza. Copeland's name was often used in headlines, suggesting that his position and influence were widely acknowledged.

Similar patterns can be identified through a study of influenza advertisements. An advertisement for Lifebuoy Soap, a national brand, included a quoted statement from Copeland on the importance of personal cleanliness as a measure to stop the spread of disease (**Figure 29**). An advertisement for a department store in Bisbee, Arizona, quoted an Associated Press report with a statement from Copeland about the utility of veils for preventing the spread of disease (**Figure 35**). Finally, an advertisement that praised the value of keeping theatres open during this time of crisis, published in *Roanoke Times* without any attribution, quotes Copeland's statement about the importance of the theater in times of crisis and the safety of the overall distribution of statements about influenza found in a sample of newspapers discussed in the **Weekly Papers Case Study**, with slight differences. A lower proportion of Copeland's statements were classified as explanatory and a higher percentage were classified as either warning or alarmist. A breakdown of these classifications by period indicates that all five of the sentences classified as alarmist appeared between October 1 and October 14, the peak of the



SAFER IN THEATRES THAN IN MOST PLACES ASSERTS DR. COPELAND

New York Theatres Kept Open By His Order

Here is Dr. Copeland's significant commentary on New York theaters in their relation with the disease now so prevalent throughout the country, among his other many reasons for keeping the theaters open:

"The theater is an essential, so regarded by our Government, being an important part of American life, in keeping up the morale of the country, relieving depression of those grieving over loved ones in the war, the raising of revenue in war tax upon admissions sold, which is so essential in helping to maintain our armies, the promotion of Liberty Loan, War Chest drives, four minute speeches and free use of their screens in the spreading of propaganda, conserving of food suggestions, etc."

The city health commissioner further asserted that 'one is as safe or safer in a large, high ceiling, well-ventilated theater as in any building or hospital, if for no other reason than that there are more pathogenic germs in a hospital than in a modern well-kept theater."

Dr. Copeland kept New York the-

Fig. 34: *Roanoke Times,* October 30, 1918, page 7

attending a theater that was well ventilated and cleaned (**Figure 34**). Copeland's name in the headline is proof of his national visibility. In each case, a specific public health measure was associated with one individual, New York City Health Commissioner Dr. Royal S. Copeland. Through these articles and advertisements, newspapers served to disseminate Copeland's advice to a national audience, a pattern suggested by topic modeling and confirmed by manual analysis of sample texts.

IV. CONCLUSION

Copeland's prominent role in shaping the public health response to the 1918 influenza epidemic is indicated by the topic modeling methods, which reveal the ways that his name was closely associated with public health measures, both within New York City and nationally. By identifying proximity of medical terms, public health responses, and broad statements of impact, this method quickly identifies word associations that are suggestive of broader interpretations. The tone classification does call into question the claims made by some historians that Copeland's language should be dismissed as too optimistic. The sentences identified by the classifier as alarmist, warning, or explanatory, along with those classified as reassuring, provides a broader mechanism for assessing the tone conveyed by Copeland's published statements. By identifying additional sentences for classification, using a wider range of newspapers, and including reported speech as well as direct quotations, this method could provide further insights into the ways that this one public health expert occupied such a prominent role in newspaper coverage of the national influenza epidemic.

Wear a Chiffon Veil and Stop the "Flu"

—Wearing chiffon veils is advocated by Dr. Royal C. Copeland, health commissioner of New York, as a means of checking New York's continually spreading epidemic of Spanish influenza. The following article by The Associated Press:

"I heartily favor every woman wearing a heavy chiffon veil," Dr. Copeland said. "These veils would serve as an almost absolute preventative. It may become necessary to order every one in New York to adopt this measure."

—Our line of veils and veiling is complete, in chiffon and nets, in dozens of different styles and by the yard.

Fig. 35: *Bisbee Daily Review*, October 27, 1918, page 2

Contemporary Implications of Research Findings

1. Understanding the flow of information is useful in tracking the spread of disease. Just as newspapers reported on the local impact of disease, contemporary forms of media provide first-hand accounts of illness, death, and recovery that are useful for epidemiologists.

2. Disease reporting in real time is valuable. The daily (or weekly) reports on disease found in newspapers demonstrate how people in the midst of a crisis respond to new challenges in ways that are revealing of attitudes, policies, and practices.

3. Understanding diseases means selecting from multiple sources of information. Even though newspapers are one type of media, they contain multiple kinds of information, thus serving as a model for current efforts to sort through numerous types of first-hand accounts, reporting, editorial commentary, and commercial advertising.

4. Social media offer particular insights into disease. Local news columns, especially in rural weekly newspapers, served many of the same functions as contemporary social media, allowing users to create their own content and disseminate it with minimal filtering, thus illustrating both individual and community values in self-reporting.

5. Voices of authority appear in multiple forms. Reporting about influenza allows for the transmission of medical and health authorities, both directly, in quoted statements, and indirectly, through language that is repeated in multiple venues. However, this official discourse is often distorted, qualified, or challenged through these multiple repetitions.

6. Evaluating the tone of reporting is challenging because it is, inherently, an interpretive activity. Identifying the elements of disease reporting that convey certain kinds of tone is challenging because it is interpretive and requires an understanding of context. Nonetheless, tone analysis is essential for tracing the potential impact of reporting.

7. Using media to understand disease involves levels of messiness. Both inaccuracy in reporting and the complexity of texts ensure that newspapers are neither comprehensive nor consistent in reporting on disease, which means that multiple sources of information must be combined with filtering mechanisms to achieve better understanding.

8. Time and space matter in tracking disease. Identifying sources that are specific to geographic location and chronological time provides an essential reminder that diseases also move through time and space, although differently than news and information. Reporting media have their own mechanisms, and knowledge of how a medium operates in time and space is crucial to understanding the relationship between the flow of news and information and the flow of disease.

9. Visualizations are not self-evident representations of data. As ways of rendering data mining outputs, visualizations are rhetorical and require tacit knowledge. The process of research and interpretations of data outputs are affected by various forms of visualization as well as by the data input conventions underlying the algorithms.

10. Scale matters. Topic modeling provides an index of reporting in newspapers, and the output of the aggregate data is different from the outputs for the individual newspapers. Thus, "big data" does not erase issues of scale; rather, it reorients them.

The question of the implications of our research findings provoked a lot of discussion among panelists and attendees at our October 17, 2013, research symposium: "An Epidemiology of Information: New Methods for Interpreting Data and Disease," as well as among project advisors and the grant team. The four case studies described in this report illustrate how the project augments and refines existing historical research on the 1918 pandemic, but they also raise important questions about the relationship between media and disease—in both historical and contemporary contexts.

Historians Nancy Bristow and Mark Humphries, who gave presentations at our research symposium, highlighted the contribution of big data to historical analysis. Bristow, for example, noted that the case studies "raise new questions about the role of the production and consumption of media in the tension between a growing national culture and the persistent schism between urban and rural communities in the early twentieth century." Additionally, Humphries pointed out that the case studies provide new insights about how public health information was spread and acted on during the pandemic, extending understandings of how influenza was reported in newspapers—through sports-related reporting, for example (See **Appendix X** in the *Final Project Report* for Bristow's and Humphries' full comments).

On the other hand, the methods we developed, in addition to their attendant challenges and opportunities, have significant implications in contemporary contexts. In his keynote address at the research symposium, the National Institute of Allergy and Infectious Diseases' David Morens, for example, emphasized that the ability to unearth previously inaccessible information can have a tremendous impact on long-standing problems. Additionally, Humphries pointed out that these methods open up new possibilities for "measuring the significance or pervasiveness of a particular point of view," an opportunity that extends to many disciplines. In short, our work not only provides new insights about the 1918 Spanish influenza pandemic, but the methods we developed are also generalizable and extendable to other fields.

Project Methods

I. INTRODUCTION

Combining algorithmic techniques with interpretive analytics, we make use of more than 90 newspapers for 1918 and 1919 that are available in two repositories: *Chronicling America* at the U.S. Library of Congress, and *Peel's Prairie Provinces* at the University of Alberta. In the preceding case studies, we focus primarily on the former collection, which includes digitized newspapers from across the U.S. and is freely available for public use. The API for this site enables the easy extraction of text captured through OCR technology and is thus amenable to data mining methods.

This section describes two main methods: Dynamic Temporal Segmentation (topic modeling and segmentation) and Tone Analysis. For each method, we discuss data sets, preprocessing steps, the algorithm, and evaluation.

II. DATA MINING METHODS

Dynamic Temporal Segmentation (Topic Modeling and Segmentation)

The analysis of temporal textual data sets is associated with challenges spanning both the need to summarize large textual data sets and the requirement to capture dynamic reorganizations and trends over time. We developed a dynamic temporal segmentation algorithm that wraps around topic modeling algorithms for the purpose of identifying change points where significant shifts in topics occur.

Data Set Details

Our primary data source is a database of historical newspapers called *Chronicling America*, a Library of Congress resource that provides an Internet-based, searchable database of historical U.S. newspapers. The website (http://chroniclingamerica.loc.gov) is maintained by the National Digital Newspaper Program (NDNP), a partnership between the NEH and the Library of Congress. Our secondary data source is the *Peel's Prairie Provinces* database, which is hosted by the University of Alberta, Edmonton, Canada.

In the *Chronicling America* data, there were no metadata associated with the newspaper pages to enable the identification of paragraphs in the OCR text. Hence, we assumed that a paragraph—which is more accurately termed a text chunk in this case—is a sentence that contains a term we are interested in (for example, *influenza, flu, grippe,* and *epidemic*) and three sentences before and after that term. Text chunks for analysis were extracted from the data sets using the search terms indicated in the individual case studies.

Data Preprocessing

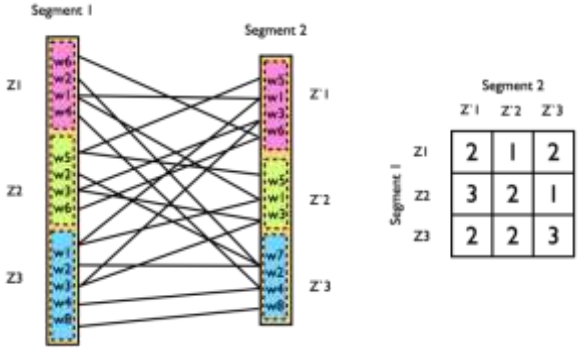
The input data to the segmentation algorithm is intended to be a "bag-of-words" representation

of documents, i.e., an unordered collection of words that does not take into account grammar or syntax. We tokenize (split) the text into individual words, followed by standard processing steps such as lowercase conversion, stemming, and the removal of stop words (commonly used words such as *the*, *and*, etc.) and punctuation.

Algorithm

The main task of the segmentation algorithm is to automatically partition the total time period defined by the documents in the collection such that segment boundaries indicate important periods of temporal evolution and re-organization.

Fig. 36: Contingency Table Used to Evaluate Independence of Topic Distribution for Two Adjacent Windows [23]



The algorithm moves across the data by time and evaluates two adjacent windows, assuming a given segmentation granularity (e.g., discrete days, weeks, or months). This granularity varies from one application to another and is decided by domain experts. We evaluate adjacent windows by comparing their underlying topic distributions and quantifying common terms and their probabilities. We chose to quantify common terms based on the overlap between them. The overlap can be captured using a contingency table. **Figure 36** shows a simplified example of two segments, each comprising three topics, and the corresponding contingency table measuring the overlap between these distributions.

For example, topic 1 (Z_1) in segment 1 and topic 1 (Z'_1) in segment 2 overlap in two words: w_1 and w_6. This results in adding the count 2 in the contingency table that corresponds to the (1, 2) cell. Ideally, the window boundaries are determined such that the topic models of the two adjacent windows are maximally independent, which will happen if the table entries are near uniform.

Formally, given the input data to be indexed over a time series $T = \{t_1, t_2, ..., t_t\}$, the segmentation problem we are trying to tackle is to express T as a sequence of segments or windows: $S_T = (S^{t_a}_{t_1}, S^{t_b}_{t_{a+1}}, ..., S^{t_t}_{t_k})$ where each of the windows $S^{t_e}_{t_s}, t_s \pounds t_e$ denotes a contiguous sequence of time points with t_s as the beginning time point and t_e as the ending time point.

Each window $S^{t_e}_{t_s}$ has a set of topics that is discovered from the set of documents that fall within

this window. The topics are discovered by a standard topic modeling algorithm such as LDA (Latent Dirichlet Allocation) [24]. Applying this algorithm results in two main distributions: document-topic distribution (the distribution of the discovered topics over the documents) and topic-terms distribution (the distribution of the words in each topic).

Topics within each window are represented as $S_{t_s}^{t_e} = \{z_1, z_2, ..., z_n\}$, where n is the number of topics discovered. Each topic is represented by a set of terms, $z_i = \{w_1, w_2, ..., w_m\}$, where m is the number of top terms extracted from the LDA topic-terms distribution. The number of top topics (n) and top terms representing a topic (m) vary from application to another.

We represent two adjacent windows as $S_{t_{s1}}^{t_{e1}}$ and $S_{t_{s2}}^{t_{e2}}$. To evaluate two adjacent windows, we construct the contingency table for two windows. The contingency table is of size r*c, where rows r denote topics in one window and columns c denote topics in the other window. Entry $n_{i\,j}$ in cell (i, j) of the table represents the overlap of terms between topic i of $S_{t_{s1}}^{t_{e1}}$ and topic j of $S_{t_{s2}}^{t_{e2}}$.

To check the uniformity of the table, there are three steps. First, we calculate the following two quantities:

- Column-wise sums $n_{i.} = \sum_j n_{ij}$
- Row-wise sums $n_{.j} = \sum_i n_{ij}$

These two quantities are used to quantify the overlap between the topics discovered from two adjacent windows. Second, we define two probability distributions, one for each row and one for each column.

$$\mathrm{p}(R_i = i) = \frac{n_{ij}}{n_{i.}} \left(1 \le j \le c\right)$$

$$\mathrm{p}(C_j = i) = \frac{n_{ij}}{n_{.j}} \left(1 \le i \le r\right)$$

Third, we calculate the objective function F to capture the deviation of these row-wise and column-wise distributions with regard to the uniform distribution. The objective function is defined as follows:

$$\mathrm{F} = \frac{1}{r} \sum_{i=0}^{r} \mathrm{D}_{KL}(\mathrm{R}_i \| (\frac{1}{c})) + \frac{1}{c} \sum_{j=0}^{c} \mathrm{D}_{KL}(\mathrm{C}_j \| (\frac{1}{c}))$$

where $\mathrm{D}_{KL}(\mathrm{P} \| Q)) = \sum_i \mathrm{P(i)} \log \frac{Q(i)}{P(i)}$

Here, D_{KL} denotes the Kullback-Leibler (KL) divergence that is used to calculate the distance between the row-wise and the uniform distribution. It is also used to calculate the distance between the column-wise distributions and the uniform distribution. The values resulting from using the $\mathrm{D_{KL}}$ are used in calculating the objective function F.

The algorithm repeats the above-mentioned steps for all permutations of the two sliding window

sizes. The goal is to minimize F, in which case the distributions observed in the contingency table are as close to a uniform distribution as possible, which means that the topics are maximally dissimilar. There are two stopping conditions for this algorithm: (1) if convergence of *F* is attained, or (2) the maximum size for both windows is achieved. A detailed description of the algorithm is shown in **Algorithm 1**.

Algorithm 1 Topic Modeling Based Segmentation

$S_T = \{\}$
$T = \{t_1, t_2, t_3, \ldots, t_t\}$
$F =$ Initialize with a big number
$x =$ min. window size.
$y =$ max. window size.
$t_{s_1} = t_{e_1} = t_{s_2} = t_{e_2} = t_1$
while $!(t_{e_1} \leq t_t)$ or $!(t_{e_2} \leq y)$
$t_{e_1} = t_{s_1} + x$
$t_{s_2} = t_{e_1} + 1$
$t_{e_2} = t_{s_2} + x$
  while $!(t_{e_2} \leq t_t)$ or $!(t_{e_2} \leq y)$
    Apply LDA separately on $S_{t_{s_1}}^{t_{e_1}}$ and $S_{t_{s_2}}^{t_{e_2}}$
    Calculate F' for $S_{t_{s_1}}^{t_{e_1}}$ and $S_{t_{e_2}}^{t_{e_2}}$
    if $F \leq F'$
      Add $S_{t_{s_1}}^{t_{e_1}}$ and $S_{t_{e_2}}^{t_{e_2}}$ to S_T
      $t_{s_1} = t_{e_2} + 1$
      Break
    else $F = F'$
    $t_{e_2} = t_{e_2} + x$
return S_T

For each set of topics discovered from each segment, the top 20 terms that represent the topic are tagged with their categories. The categories we identify here are "geo-political," "person," "organization," and "uncategorized." The category identification is accomplished using a named entity recognition algorithm on the documents assigned to each topic. A simple word filtering technique is then applied to the discovered topics within and across segments to eliminate repeated topics and repeated terms.

This algorithm was implemented using Java, and the data preprocessing and post processing was implemented using Python. Algorithm and tone analysis code, as well as instructions, can be found at this link: http://vtechworks.lib.vt.edu/handle/10919/19271.

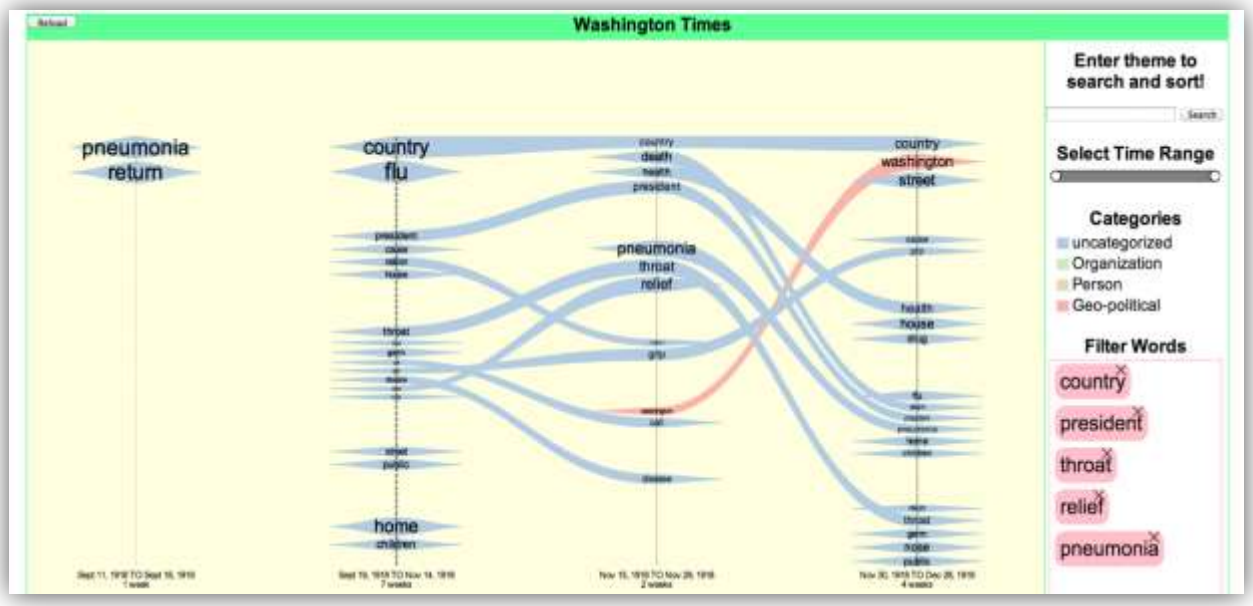
Validation: Expert User Study

To further validate the utility of our algorithm, we developed a visualization called ThemeDelta with colleagues from Purdue University and conducted a qualitative user study involving expert participants. The purpose was to study the suitability of the approach for in-depth expert analysis of dynamic text corpora. For our qualitative evaluation, we extracted data from five U.S. newspapers from New York, Washington, DC, and Philadelphia, using the *Chronicling America* collection. Three graduate students—one from the history department and two from the English

department at Virginia Tech—participated in this study. All were required to have prior knowledge of American history during 1918.

The three participants were given task questions, and model answers were used for evaluation.

Fig. 37: ThemeDelta Representation of the *Washington Times*



Participants correctly reported the sentiment of influenza reporting in the five newspapers. They also successfully described changes in reporting on the spread of influenza. Finally, they all noticed connections between influenza and other categories (e.g., schools, war, and hospitals). **Figure 37** shows a ThemeDelta visualization for the *Washington Times* newspaper in 1918 depicting user interaction wherein the user is searching for specific terms (highlighted in red on the right side in the filter words box).

Tone Analysis

The goal of tone analysis is to identify tone from text. Recognizing and identifying tones across a large data set would mean that computational analysis could actually interpret some forms of textual meaning. Integrating tone classification with other data mining techniques would therefore give humanities and social science researchers the ability to automate the analysis of textual sentiment in large data sets. Understanding shifts of tone in the coverage of influenza across time and space, identifying tones associated with particular aspects of news coverage, and assessing changes in reporting as the disease waxed and waned are elements of interpretation that we would like to see automated through tone classification.

We initially focused on the following tones: alarmist, warning, encouraging, patriotic, explanatory, reassuring, and humorous. We developed these tones through the recursive manual analysis (see discussion in the "**Data Analysis through Close Reading**" sub-section below) that first identified tones in advertisements, revised them, and then applied them to a broader sample of texts.

To detect tones from text automatically, we used a supervised machine learning approach. This is a classic text classification problem and a usual practice in approaching such problems is to first examine text chunks using a Multinomial Naïve Bayes classifier (based on the bag-of-words model). Given that we were using a supervised learning approach to solve this problem, we built a training set for the classifier to learn the differences among tones.

Classifier

We used a Naive Bayes classifier that is based on Bayes's theorem with a feature model that is conditionally independent of the tone. In other words, given a tone, word occurrences are assumed to be independent within that tone. The classifier is first trained using the features extracted from manually tagged text. After training, the classifier predicts tones for newly extracted, previously unseen, text.

Extracted Features

We examined a number of features such as Tfidf (Term frequency–inverse document frequency), a numerical statistic which reflects how important a word is to a document in a collection or corpus. Tf (Term frequency) represents how frequently a term appears in a document. N-grams is a contiguous sequence of n terms from a document. We used 1- to 2-grams, which means we extracted all 1- to 2-grams and got their term frequencies.

We examined these features alone and combined, e.g., Tfidf of 2-grams, Tf of 2-grams, and Tfidf of single grams. We also applied a number of accuracy enhancers such as normalizing frequencies, transforming the term frequencies, and smoothing Tfidf.

Training

In order to train the tone analysis algorithm, we first needed to build a training data set. Initially, paragraphs were extracted from the newspapers for tone tagging, but it was impossible to code these text chunks because of the poor OCR. We then resorted to selecting and transcribing sentences from articles about influenza (see **Figure 38**). Thus, we used manually cleaned up data in order to facilitate the training process. We (1 historian, 1 librarian, and 2 rhetoricians) coded the sentences according the criteria outlined in the "**Data Analysis through Close Reading**" sub-section below, criteria that were modified as we discussed coding discrepancies.

We trained the classifier to detect the four tones (alarmist, explanatory, reassuring, and warning), using approximately 300 cleaned sentences from newspapers and four coders, who attained a moderate level of agreement in their classifications (Kappa=0.47). Kappa is a measure used to assess the degree of agreement among raters. From our seven original categories, we combined reassuring and encouraging into a single category since coders consistently conflated the two, and dropped the patriotic and humorous tone categories because of a lack of samples. The performance of the classifier was estimated using K-Fold Cross Validation, a technique that splits data into K equal-sized subsets so that K-1 subsets are used for training and the remaining portion of the data is used for testing. This is repeated K times and the results averaged to obtain the final performance measure. To date, the tone classifier accuracy is 72%.

Evaluation

We conducted tone analysis on every paragraph within one title to determine how reporting changes in a given geographic location as influenza approaches, peaks, and wanes. Using both cleaned and uncleaned data, we also ran the classifier on several weekly newspapers to determine whether there is a relationship between tone and the status of the epidemic in those regions covered by the newspapers (*Clinch Valley News* and *Big Stone Gap Post*) and, if so, what the outputs could tell us about those relationships. Our findings suggest—as expected—significant differences between the results generated using cleaned data and those generated from uncleaned data.

III. DATA ANALYSIS THROUGH CLOSE READING

The research team utilized traditional scholarly techniques of close reading in three ways: 1) to familiarize ourselves with the newspapers of the period digitized in *Chronicling America*, especially concerning discourses about influenza during the fall of 1918, 2) to corroborate and expand on the analysis developed from data mining outputs, and 3) to code selected sentences for tone in order to train the classifier.

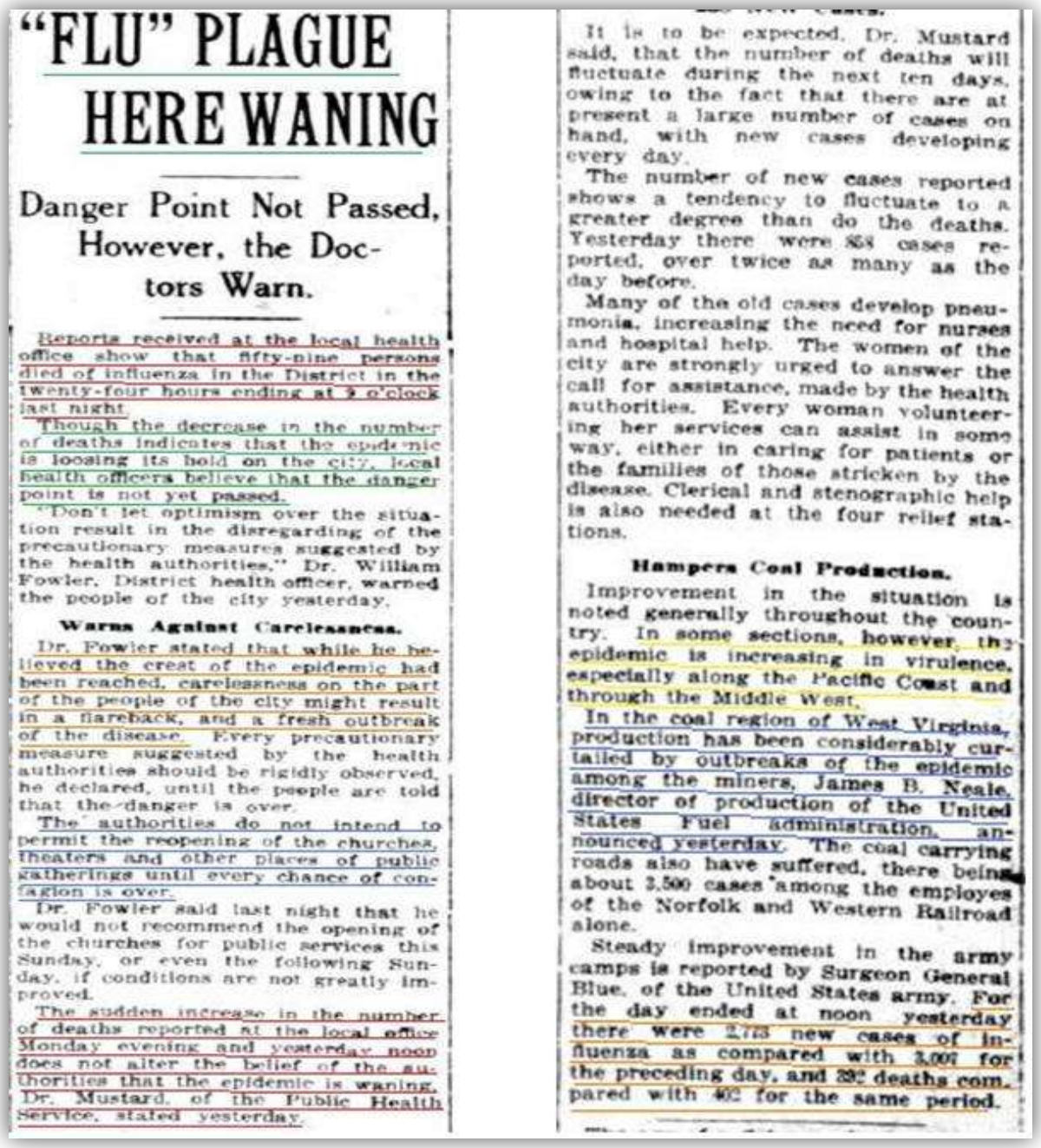
Close (or manual) reading is characterized by an individual researcher or a group of researchers relying on content-area expertise and professional judgment to analyze and evaluate the significance of discrete texts. In this project, researchers began by reading the newspapers during the period of the epidemic in a given location, either by using keyword searching to identify articles of interest or by reading through entire newspaper editions during the period under study. In both instances, the purpose was to develop a feel for the reporting styles of the various newspapers selected, to understand the types of articles (genres) that covered the epidemic, and to identify themes, patterns, and rhetorical features in the reporting on the epidemic. Our analysis was often recursive, beginning with data mining outputs but taking us quickly to the news articles to verify our assumptions and then back to the data mining outputs in order to deepen our understanding of what information they offered. Alternatively, we sometimes began with the newspapers and then read the data mining outputs from our reading of the text itself in context.

The close reader benefits from being able to see the context for the discourse being analyzed—the page in the newspaper, the font, a headline, accompanying articles, images, etc., all affect the meaning of the text as originally produced and read. Close reading involves identification of the genre of text being examined (editorial, obituary, etc.), knowledge of the rhetorical purposes and common rhetorical features of the genre, identification of any seemingly aberrant elements of the text under scrutiny, as well as identification of patterns, themes, and rhetorical features. Content expertise helps the close reader make prudent judgments about which articles to focus on, while an understanding of the genre, in this case news reporting, allows the close reader to make accurate assessments. Close reading allowed the researchers to identify significant differences in the reporting style of the periodicals in the Weekly Newspapers Case Study when compared those in the Daily Newspapers Case Study.

The computational strategies of data mining strip the text of its context; close reading restores the text to that context, answering questions that emerge as a result of confusing data mining

results (e.g., Why do so many results refer to hogs?) and verifying the findings developed from analyzing data mining outputs. Keyword searching available in the digitized *Chronicling America* collection can aid the researcher by allowing her or him to focus on only those news articles including terms like *influenza*, *grippe*, or *epidemic*, but poor OCR transcription may lead to problems as the researcher will miss articles that have misspelled search terms in the OCR.

Preparing text to train the classifier to detect tone necessitated editing of poor OCR transcription and then coding of the resulting sentences according to tone. Four researchers coded cleaned sentences, reading each closely and deciding which tone to assign to it. Researchers either read



"FLU" PLAGUE HERE WANING

Danger Point Not Passed, However, the Doctors Warn.

Reports received at the local health office show that fifty-nine persons died of influenza in the District in the twenty-four hours ending at 9 o'clock last night.

Though the decrease in the number of deaths indicates that the epidemic is loosing its hold on the city, local health officers believe that the danger point is not yet passed.

"Don't let optimism over the situation result in the disregarding of the precautionary measures suggested by the health authorities," Dr. William Fowler, District health officer, warned the people of the city yesterday.

Warns Against Carelessness.

Dr. Fowler stated that while he believed the crest of the epidemic had been reached, carelessness on the part of the people of the city might result in a flareback, and a fresh outbreak of the disease. Every precautionary measure suggested by the health authorities should be rigidly observed, he declared, until the people are told that the danger is over.

The authorities do not intend to permit the reopening of the churches, theaters and other places of public gatherings until every chance of contagion is over.

Dr. Fowler said last night that he would not recommend the opening of the churches for public services this Sunday, or even the following Sunday, if conditions are not greatly improved.

The sudden increase in the number of deaths reported at the local office Monday evening and yesterday noon does not alter the belief of the authorities that the epidemic is waning, Dr. Mustard, of the Public Health Service, stated yesterday.

It is to be expected, Dr. Mustard said, that the number of deaths will fluctuate during the next ten days, owing to the fact that there are at present a large number of cases on hand, with new cases developing every day.

The number of new cases reported shows a tendency to fluctuate to a greater degree than do the deaths. Yesterday there were 858 cases reported, over twice as many as the day before.

Many of the old cases develop pneumonia, increasing the need for nurses and hospital help. The women of the city are strongly urged to answer the call for assistance, made by the health authorities. Every woman volunteering her services can assist in some way, either in caring for patients or the families of those stricken by the disease. Clerical and stenographic help is also needed at the four relief stations.

Hampers Coal Production.

Improvement in the situation is noted generally throughout the country. In some sections, however, the epidemic is increasing in virulence, especially along the Pacific Coast and through the Middle West.

In the coal region of West Virginia, production has been considerably curtailed by outbreaks of the epidemic among the miners, James B. Neale, director of production of the United States Fuel administration, announced yesterday. The coal carrying roads also have suffered, there being about 3,500 cases among the employes of the Norfolk and Western Railroad alone.

Steady improvement in the army camps is reported by Surgeon General Blue, of the United States army. For the day ended at noon yesterday there were 2,713 new cases of influenza as compared with 3,007 for the preceding day, and 282 deaths compared with 402 for the same period.

Fig. 38: *Washington Herald*, October 23, 1918, page 1

the sentences in isolation or linked back to the original publication for more contextual information (**Figure 38**). The coded sentences were then evaluated for consistency across coders (inter-rater reliability, as measured by a Kappa statistic). This form of close reading is more removed from the context of the actual newspapers as described above, but coding in this manner does replicate close reading methods and relies on researcher expertise and training to be successful.

Tone Analysis Categories

Alarmist: uses fear or urgency, often mentioning number (i.e., *numbers* of sick or dead); induces a sense of panic; mentions a number in a comparative context (e.g., 10 *more* deaths today); mentions a seemingly large number for the context (i.e., *hundreds* in a single day).

Warning: text or image refers to the gravity of the situation; serious but not urgent; cautioning; advises the reader what to do; mentions measures being taken but conveys no sense that the threat is diminishing (i.e., if we do these things, things shouldn't get worse, but there is no guarantee).

Reassuring: comforting; implies threat is diminishing; addresses fears with soothing sensibility; typically conveys the idea that if one takes a recommended action, everything will be fine; motivates action with sense of hopefulness, improvement, or possibility of avoidance of disease; involves sense that action will lead to betterment; attempts to downplay or counteract alarmist sensibility or news.

Explanatory: discourse as a source of information; lacks distinctive affect.

Data Quality, Sources, and Management

As originally envisioned, this Digging into Data project was to complement open-access historical newspapers from the U.S. and Canada with historical newspaper digital archives already available on campus under the Virginia Tech Libraries' licenses.

The inclusion of proprietary newspaper archives, of course, would broaden the scope of the coverage of the 1918 Spanish influenza pandemic. The newspapers available under the Virginia Tech Libraries' perpetual-access licenses could introduce into the project such big-city dailies as the *New York Times, Chicago Tribune, Los Angeles Times* (all from ProQuest) and *Philadelphia Inquirer* (part of the Readex *America's Historical Newspapers* series), along with many smaller papers across the country, including the African-American press (ProQuest's *Chicago Defender* and the many newspapers in the Readex *African-American* newspapers collection).

In addition to providing a larger lode of textual evidence to mine, proprietary newspaper databases appeared to offer possible technical insights into indexing and metadata that might overcome shortcomings in the free *Chronicling America* or *Peel's Prairie Provinces* collections: undifferentiated content and poor rendering of printed words into machine-readable text. The *Chronicling America* newspapers' unit of analysis is the full newspaper page; analysts must consult the page images to determine whether flu-related content appears in a news story, an


Vol. 9, No. 187. Albuquerque, N. M., Saturday, Oct

YANKEES MAKE ATTACK

TODAY
IN THE NEWS

Spanish Influenza
Holland Nails a Lie
Back the Bond Sellers

CASES diagnosed by competent physicians as Spanish Influenza have appeared in Albuquerque. This was to have been expected. Albuquerque is a tourist city. The number of transients here each day probably is greater than in any other city of equal population in the west. Many soldiers are passing through. It would have been surprising if this city had escaped a showing of an epidemic that is sweeping the nation.

There is no cause for undue alarm in a meeting of the board of health to consider the situation. There is need of simple precautions.

We have here a climate which is highly efficacious in combating Spanish Influenza or any other kind of influenza. The sunshine of New Mexico is the best protection in the world. If we will take a little extra care of our general physical condition, spend as much time as possible out of doors and observe the simple precautions proposed, to guard against unnecessary exposure to contagion, our

Kicking Off! — By Ripley.

GO DOWN THE FIELD WITH THE BALL !

Albuquerque, N. M.t Saturday, October 5, I9IH.
YANKEE
Price Five Cents
ATTACK
m WIDE FRONT
7 ODA Y
IN THE NEWS
Spanish Influenza
Holland Naili a Lie
Back the Bond Sellers
Kicking Off!
By Ripley.
CASKS cli;iiMHM'll hv cnlllpe- i
tent physicians ns SiiiniHh
Influoii.a have appeared in
Alhu(ueruo. This was to have i
hceii expected. Alliiiiiii'riiii is a !
tourist city. Tin iiuiiiImt f
transients hero each day prnhahly i
is prreator than in miy other i t
of mi I population in thi west. 1
Many soldiers are passion; thrnii.rh.
If would have heen . surprising if
this city hal escaped a showing of
at) epidemic that in sweeping the
lint ion.

Fig. 39: Albuquerque *Evening Herald*, October 5, 1918, page 1
Picture Image and Underlying Machine-Readable Text

advertisement, or a letter. The *Chronicling America* page-level digitization often runs together text from more than one article, conflates words in headlines with words in text, and otherwise makes hash of the elements of design, grammar, and syntax that make the newspaper text meaningful to human readers (see **Figure 39**).

Proprietary newspaper databases index content to the individual article (except for very short articles and classified advertisements) and tag it by genre. The possibility of training the algorithm with the more sophisticated proprietary data and metadata in order to approximate article-level interpretation and comparison in the *Chronicling America* newspapers was appealing to both the humanists and computer scientists.

The research team also hoped that the proprietary newspapers' underlying machine-readable text would be more accurate than the "dirty OCR" of the *Chronicling America* newspapers that was often unintelligible or, worse, ambiguous without extensive manual cleanup and verification (see discussion in the **Daily Newspapers Case Study** and **Project Methods** sections). Cleaner text, we hoped, would facilitate training the algorithm for tone analysis. Finally, acquiring and curating the proprietary newspapers' content for future computational analysis by Virginia Tech researchers would provide an opportunity for developing the Virginia Tech Libraries' new role as the university's hub for research-data management and curation.

Unfortunately, it was not possible to utilize the proprietary newspaper databases for text mining, although they were used in conventional, manual analyses that enriched the grant researchers' knowledge bases. Three related reasons prevented our utilization of the library's historical newspapers: cost, novelty, and contracts. Additional fees for actually acquiring the content for computational analysis were required, even though these were not well spelled out in the library's "perpetual access" licenses. For one vendor, for example, it was not technically possible to extract only the years required for this project from its multiple papers. The additional costs were unanticipated in the initial project budget. In addition, questions about the size of the data transfer and storage concerns complicated acquisition. Researchers, library informaticists, and vendor representatives felt their way across unfamiliar terrain for four months, until there simply was not enough time left to use the proprietary content. Ultimately, it became clear in conversations with the database vendors that text mining of historical periodicals was a use of their content not anticipated in their contracts or their practices.

With regard to data management for the project itself, "An Epidemiology of Information" was awarded without a data management plan (DMP) requirement, but the research team began consultations with the repository library shortly after the grant award was made. These conversations continue and demonstrate that DMP practices developed from laboratory sciences do not necessarily map well to the information sources, conventions, and cultures of scholars in the humanities. Just as the appropriation of proprietary database content proved difficult as a result of contracts that did not anticipate computational analysis, the need for content transfer and storage, and altered licensing agreements, there are challenges to data management when the project is interdisciplinary *and* involves "big data." Avoiding a one-size fits-all model for data management is in the interest of humanities scholars and library informaticists, and including DMP language in proposals for projects like this one is a necessity going forward.

Appendix A: Project Management

The Grant Team and Project Approach

Four principal investigators contributed interdisciplinary perspectives to the project: E. Thomas Ewing (History); Bernice L. Hausman (English); Bruce Pencek (Library); and Naren Ramakrishnan (Computer Science). Three graduate research assistants supported the project: Samah Gad (Computer Science), Michelle Seref (English), and Kathleen Kerr (English). Additionally, Gunther Eysenbach represented our Canadian partners at the University of Toronto's Centre for Global eHealth Innovation, where social scientists, engineers, and health professionals work to develop information metrics for understanding the spread of disease in the contemporary world.

We made use of more than 90 newspapers from 1918 and 1919 that are available in two repositories: *Chronicling America* at the U.S. Library of Congress and *Peel's Prairie Provinces* at the University of Alberta. In the case studies, we focused primarily on the former collection, which includes digitized newspapers from across the U.S. and is freely available for public use. Our research shows that the application of algorithmic techniques enables the domain expert to systematically explore a broad repository of data and identify qualitative features of the pandemic (response, sentiment, and associations) in the small scale as well as the genealogy of information flow in the large scale. By joining traditional close-reading analytics with the techniques of data mining, we demonstrated how humanities and social science disciplines can collaborate with computer science to develop robust interpretive methods for enhanced research.

Milestones and Documenting the Project

We met bi-weekly during the first half of the project, documenting progress, decisions, and challenges in meeting notes. The grant team also collaborated using Virginia Tech's online learning system, *Scholar*, where we stored all our working documents. We established short-term goals for each of the reporting periods, as outlined in the interim reports, developing, testing, and refining our methods. Initially, we focused on three areas: topic modeling and segmentation, tone analysis, and network analysis. However, issues related to data integrity and access (see **Project Methods** section in this report) posed significant challenges, particularly given time and resource constraints. Hence, we scaled back our tone analysis efforts, limiting our application of this computational approach to the Weekly Newspapers Case Study. Similarly, while we made some progress with network analysis, we decided to focus our efforts primarily on dynamic temporal topic modeling and segmentation given our limited time and resources.

During the second half of the grant, we met weekly as we tested our methods in the case studies outlined in this report. We also organized "An Epidemiology of Information: New Methods for Interpreting Data and Disease," which took place at the Virginia Tech Research Center in Arlington, Virginia, on October 17, 2013. The purpose of the symposium was to present our findings and obtain feedback from participants and attendees, who included public health officials, librarians, medical humanities scholars, epidemiologists, computer scientists, and

researchers, among others. The symposium proceedings were also simulcast to the Virginia Bioinformatics Institute at the Virginia Tech campus in Blacksburg, VA.

Feedback on the Project

An interdisciplinary team of advisors helped to guide our research throughout the life of the project:

- Rosalind Eggo (Postdoctoral Fellow, Center for Computational Biology and Bioinformatics, University of Texas at Austin)
- Graeme Hirst (Professor, Department of Computer Science, University of Toronto),
- Jian Pei (Professor of Computing Science, Simon Fraser University),
- Francois Debrix (Professor and Director, Alliance for Social, Political, Ethical, and Cultural Thought, Virginia Tech),
- Ann Herring (Professor, Department of Anthropology, McMaster University)

Our advisors provided regular input on our research that included written responses to all interim reports, meetings via Skype, and comments on the *Symposium Research Report*, which contained our methods and findings. Overall, their comments were both positive and useful, and we made every effort to incorporate their comments into our research. They offered suggestions for publication and conference venues, pointed out the need for additional contextualization, and identified ways we could improve our methods and enhance our analyses. Additionally, symposium commentators responded positively to the research methods and findings we presented, noting specifically the complementary nature of interpretive and computational analytics.

Other Measures of Project Success

Our grant project proposal indicated two primary goals for this research: 1) to augment and refine existing historical research on the 1918 pandemic and 2) to determine how public health information was spread and acted upon during the pandemic. As the case study summaries above illustrate, we accomplished these objectives. Additionally, we measured the success of the project in several other ways: response to the research symposium, publications, and follow-on or spin-off activities. Although symposium attendance was negatively impacted by the government shutdown (many of our registrants are federal employees), we were nonetheless registered at full capacity. The large number of registrants and the diversity of their expertise suggest the positive value of our research to a range of disciplines. Similarly, our many presentations, invited lectures, and publications reinforce the timeliness of and interest in our research. Finally, grant team members are working on a number of projects that extend various aspects of this project:

- A summer seminar for teachers on the Spanish influenza pandemic
- An undergraduate research project on the Russian flu
- A funding proposal for extending the Vaccination-Visualization Case Study
- Vaccination Research Group efforts

Our Evaluation and Lessons Learned

Our two primary goals for this research were to 1) augment and refine existing historical research on the 1918 pandemic and 2) determine how public health information was spread and acted upon during the 1918 Spanish influenza pandemic. As the case study summaries above illustrate, we accomplished these objectives by developing methods that combine computational and interpretive analytics. However, as we note in the **Introduction** section of this report, we learned that digitized data such as the newspapers in the *Chronicling America* collection also present significant limitations to computationally based research and data mining cannot substitute for close or "manual" readings of discrete texts.

We also learned much about this type of research from an operational perspective. Our biggest challenge related to data quality and management, as noted in the **Data Quality, Sources, and Management** section of this report. Additionally, we determined that a project such as "An Epidemiology of Information" needs more staff and funding. It was difficult to estimate time requirements and appropriate funding levels given the labor-intensive nature of, for example, manually correcting the data. Such challenges meant that we had to decide how to allocate resources, often at the expense of other critical tasks. Such adjustments also meant we had to reconsider the scale and scope of our research. Hence, while we hoped to conduct more research, for example, further studying the relationships of time and space to patterns and emphasis of reporting—research our advisors suggested would be very useful—we simply did not have the resources to pursue this line of analysis within the grant period.

Appendix B: List of Figures and Tables

Algorithms

*Figures**

Tables

***Note:** The newspaper articles used in this paper are from the *Chronicling America* collection of historical newspapers.

Appendix C: References

Introduction:

1. *Public Health Reports* 125, supplement 3 (2010).
2. N. Bristow, American Pandemic, New York: Oxford University Press, 2012.
3. M. Humphries, *The Last Plague: Spanish Influenza and the Politics of Public Health* in Canada, Toronto: University of Toronto Press, 2013.
4. K. Crawford, "Big Picture," *Culturomics: Resources,* http://www.culturomics.org/Resources/faq.
5. K. Cukier and V. Mayer-Schoenberger, "Think Again: Big Data," *Foreign Affairs* 92.3 (2013): 27-40, http://www.foreignpolicy.com/articles/2013/05/09/think_again_big_data.

Vaccination-Visualization Case Study

6. B. Barton and M. Barton, "Ideology and the Map," *Central Works in Technical Communication*, New York: Oxford University Press, 2004, pp. 232-252.
7. C. Kostelnick and M. Hassett, *Shaping Information: The Rhetoric of Visual Conventions*, Carbondale: Southern Illinois University Press, 2003.
8. 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.
9. J. M. Eyler,"The State of Science, Microbiology, and Vaccines Circa 1918," *Public Health Reports* 125, supplement 3 (2010): 27-36.
10. Y. Chien, K. Klugman, and D. Morens, "Efficacy of Whole-Cell Killed Bacterial Vaccines in Preventing Pneumonia and Death," *The Journal of Infectious Diseases* 202.11 (2010): 1639-1648.
11. B. Hansen, "New Images of New Medicine: Visual Evidence for the Widespread Popularity of Therapeutic Discoveries in America after 188," *Bulletin of the History of Medicine* 73.4 (1999): 629-678, pg. 631.
12. Additional Literature Review Sources:
    a. R. Arnheim, *Visual Thinking*, Los Angeles and London: University of California Press, 1997.
    b. J. Barry, The Great Influenza: *The Story of the Deadliest Pandemic in History*, New York: Penguin Books, 2005.
    c. N. Bristow, American Pandemic, New York: Oxford University Press, 2012.
    d. 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.
    e. 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.
    f. 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.
    g. S. Hall, Ed., *Cultural Representations and Signifying Practices*, London: Sage Publicatons, 1997.
    h. J. Hullman and N Diakopoulos, "Visualizaton Rhetoric: Framing Effects in Narrative Visualization," *IEEE Transactions on Visualization and Computer Graphics*, vol. 17.2, Dec. 2011, pp 2231-2240.
    i. M. Humphries, *The Last Plague: Spanish Influenza and the Politics of Public Health* in Canada, Toronto: University of Toronto Press, 2013.
    j. E. Morris, *Believing is Seeing*, New York: Penguin Press, 2011.
    k. 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.
    l. C. Spinuzzi, *Tracing Genres Through Organizations*, Cambridge, MA: The MIT Press, 2003.
    m. E. Tufte, *Beautiful Evidence*, Cheshire, CT: Graphics Press, 2006.
    n. E. Tufte, *Visual Explanations*, Cheshire, CT: Graphics Press, 1998.

o. 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.

Public Health Case Study

13. "Fear of Epidemic from Spanish Grip Scouted by Officials," *New York Tribune,* August 17, 1918, p. 12.
14. A. Crosby, *America's Forgotten Pandemic: The Influenza of 1918*, New York: Cambridge University Press, second edition, 2003, pp. 72, 176.
15. J. Barry, *The Great Influenza. The Story of the Deadliest Pandemic in History*, New York: Penguin Books, 2004, pp. 181, 269, 270.
16. N. Bristow, *American Pandemic: The Lost Worlds of the 1918 Influenza Pandemic*, New York: Oxford University Press, 2012, pp. 101-102, 105, 111.
17. For an assessment of Copeland's responses in terms of his advancement of the cause of homeopathic medicine, see Natalie Robins, *Copeland's Cure: Homeopathy and the War Between Conventional and Alternative Medicine*, New York: Alfred A. Knopf, 2005, pp. 151-156.
18. H. Markel, et al., "Nonpharmaceutical Interventions Implemented by US Cities During the 1918-1919 Influenza Pandemic," *Journal of the American Medical Association,* Vol. 298, No. 6 (August 8, 2007), pp. 644-654.
19. N. Tomes, "'Destroyer and Teacher': Managing the Masses During the 1918-1919 Influenza Pandemic," *Public Health Reports,* Vol. 125, Supplement 3 (April 2010), pp. 48-62.
20. A. Minna Stern, et al, "'Better Off in Schools': School Medical Inspection as a Public Health Strategy During the 1918-1919 Influenza Pandemic in the United States," *Public Health Reports,* Vol. 125, Supplement 3 (April 2010), pp. 63-70.
21. F. Aimone, "The 1918 Influenza Epidemic in New York City: A Review of the Public Health Response," *Public Health Reports,* Vol. 125, Supplement 3 (April 2010), pp. 70-79.
22. "New York," *The American Influenza Epidemic of 1918-1919: A Digital Encyclopedia* University of Michigan Center for the History of Medicine and Michigan Publishing, University of Michigan Library (http://www.influenzaarchive.org/index.html)

Project Methods

23. S. Gad, N. Ramakrishnan, K. N. Hampton, and A. Kavanaugh. Bridging the divide in democratic engagement: Studying conversation patterns in advantaged and disadvantaged communities. In *Proceedings of the IEEE Conference on Social Informatics*, 2012.
24. D. M. Blei, A.Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003

---

[1]A version of this case study appears in "Mining Coverage of the Flu: Big Data's Insights into an Epidemic" by E. Thomas Ewing, Samah Gad, Bernice L. Hausman, Kathleen Kerr, Bruce Pencek, and Naren Ramkrishnan (*Perspectives on History* 52.1 (January 2013)).

[2]A version of this case study appears in "Visualization and Rhetoric: Key Concerns for Utilizing Big Data in Humanities Research" by Kathleen Kerr, Bernice L. Hausman, Samah Gad, and Waqas Javed (Proceedings, Big Humanities Workshop, IEEE International Conference on Big Data, October 2013 (doi: 10.1109/BigData.2013.6691666, pp. 25-32).