Discovering the Ebb and Flow of Ideas from Text Corpora

Justin Jee, *New York University*
Lee Case Klippel, *Columbia University*
M. Shahriar Hossain and Naren Ramakrishnan, *Virginia Tech*
Bud Mishra, *New York University*

Changes in word usage patterns in text corpora can yield insight into historic events and discoveries.

The rise and decline in popularity of ideas have a profound effect on human society. Tracing the ebb and flow of ideas has important implications for scientific and historical research because while newer, more accurate, or more useful ideas might be expected to consistently succeed older ones, in reality this isn't always the case.

In the 1840s, for example, when Ignaz Semmelweis discovered that hand sanitation reduced the incidence of childbed (puerperal) fever, his conclusions were disputed because they didn't fit within the context of miasma theory, the prevailing belief system. While following Semmelweis's recommendations would have prevented a widely feared disease from disseminating rapidly, societal forces prevailed against him and many infants died unnecessarily.

Humanist thinkers and observers typically have inferred shifts in ideas' popularity; however, with the advent of resources like Google Books n-grams, it's possible to quantitatively analyze word usage in massive volumes of text. For example, researchers have built models characterizing the spread or decline of scientific concepts (L.M.A. Bettencourt et al., "The Power of a Good Idea: Quantitative Modeling of the Spread of Ideas from Epidemiological Models," *Physica A*, May 2006, pp. 513-536) and tracked the fame of individuals in different categories, such as authors or politicians (J. Michel et al., "Quantitative Analysis of Culture Using Millions of Digitized Books," *Science*, 14 Jan. 2011, pp. 176-182).

Some researchers have dubbed the study of human cultural issues through textual data mining *culturomics*. Culturomics has been characterized as "very high-turbo data analysis" by humanists, who decry the lack of historical expertise and departure from close reading traditions in such analysis (A. Grafton, "Loneliness and Freedom," *Perspectives on History*, Mar. 2011; www.historians.org/Perspectives/issues/2011/1103/1103pre1.cfm). Of particular concern is the existence of artifacts due to selective inclusion in certain databases (T. Schwartz, "Culturomics: Periodicals Gauge Culture's Pulse," *Science*, 1 Apr. 2011, pp. 35-36).

Nevertheless, synergistic collaboration between computational researchers and humanists is possible. Many data mining and machine learning techniques, including *temporal segmentation*, can support the automatic discovery and characterization of features of interest to humanists.

TEMPORAL SEGMENTATION

To gain insight into why new ideas thrive or struggle, we studied the usage patterns of related words, ideas, and people's names in text corpora spanning several decades.

One form of analysis calculates word usage from textual databases to inform analytical models. However, these analyses are vulnerable to lack of coverage specificity, often leading to word frequency variability not connected to popularity shifts.

0018-9162/12/$31.00 © 2012 IEEE Published by the IEEE Computer Society

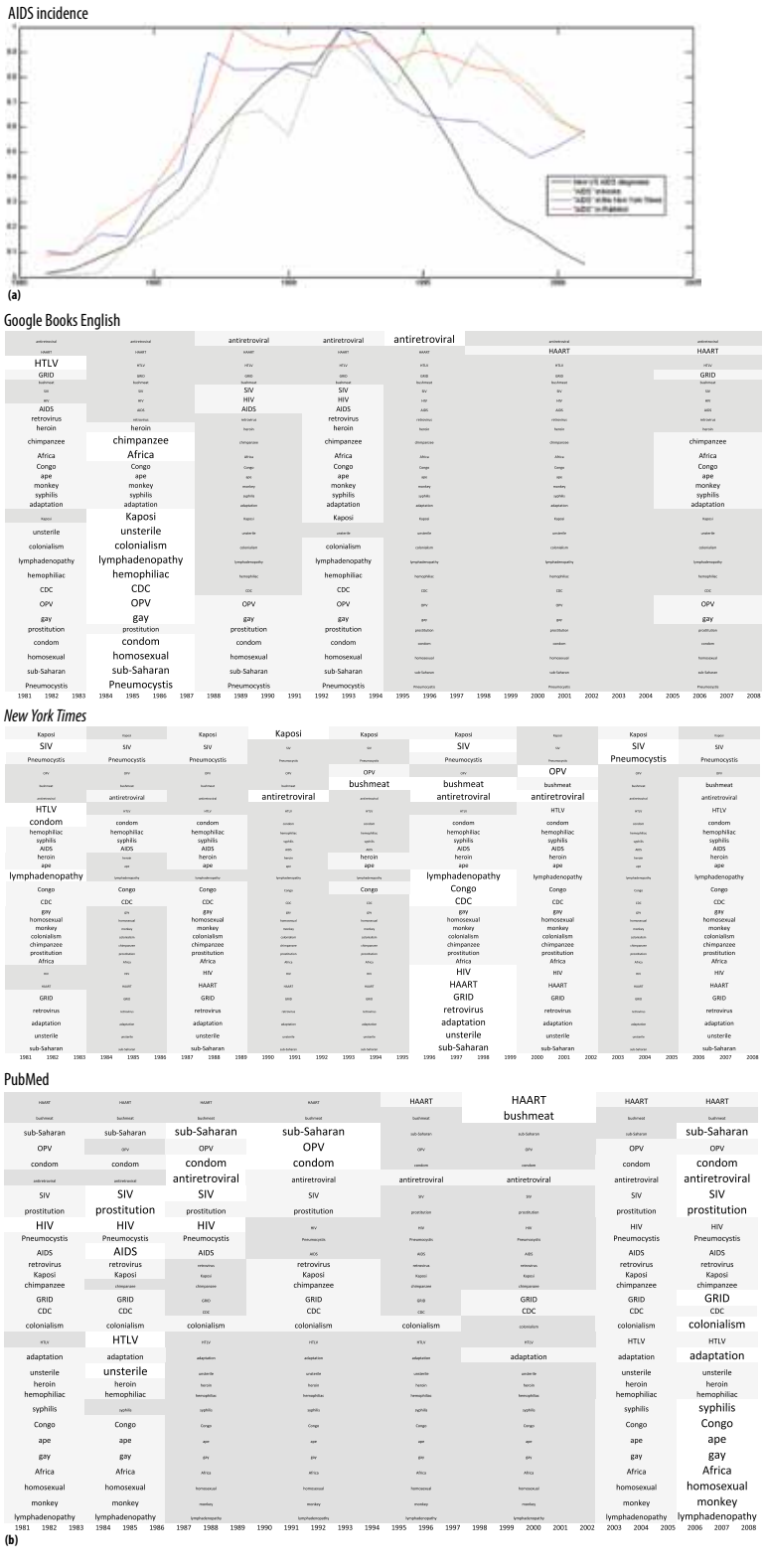

Figure 1. Temporal segmentation of use of the term "AIDS." (a) Plot of new AIDS cases per year alongside frequency of the word in the *New York Times*, Google Books English, and PubMed corpora from 1981 to 2001. (b) Segmentations of the time period 1981-2008 depicting the frequency of 30 AIDS-related words in the three databases.

Instead, we tracked changes in correlation between word-usage frequencies across time using a temporal segmentation framework (N. Ramakrishnan et al., "Reverse Engineering Dynamic Temporal Models of Biological Processes and Their Relationships," *Proc. Nat'l Academy of Sciences*, 13 July 2010, pp. 12511-12516). Temporal segmentation breaks a time course into windows derived from clusters of words rising and falling in popularity together over a given window. Across windows, clusters break up and regroup into different formations, signifying a qualitative behavioral change.

We visualized these segmentations as timelines by displaying the clusters in each time window so that words in the same cluster had the same font size and cell shading. Clusters with the highest rate of increase in a given window had the largest font and brightest cell shade.

We created timelines using frequency values from three corpora—the *New York Times*, Google Books n-grams, and PubMed databases—and concentrated on terms from three historic events: the US AIDS epidemic, Ignaz Semmelweis's discoveries regarding hand sanitation, and the dissemination of Einstein's relativity theory.

Similar to previous analyses of text corpora using Wikipedia as a list-generating source, for each corpus we designed a focused crawler that began with a Wikipedia article ("AIDS," subsection 9: "History and origin"; "Contemporary reaction to Ignaz Semmelweis"; and "History of special relativity," subsection 3.1: "Einstein 1905") and gathered hyperlinked terms for query expansion by following hyperlinks to other related Wikipedia articles (for example, "gay-related immunodeficiency" for the AIDS study, "germ theory of disease" for the Semmelweis study, and "David Hume" for the relativity study). The crawler only considered terms that were represented in more than 1 in

10^8 of the total words printed in the corpus during the years of interest.

For each term, we calculated word usage in the Google Books database as the number of times a word was used in print in a given year divided by the total number of words published in that year. We calculated word usage in the *New York Times* and PubMed databases as the number of times a word was used as a keyword divided by the number of articles included in the database per year. We applied least-squares polynomial smoothing to the raw data.

We segmented the resulting multivariate time-course data. The segmentation algorithm uses dynamic programming to identify segment boundaries such that clusters on either side of the boundary would be maximally disparate (estimated in terms of a mutual information value).

For ease of interpretation, we imposed minimum and maximum length constraints on segments as 3 and 6 years, respectively. We used three clusters in each segment to capture the overlap between clusters in a 3 × 3 contingency table.

Finally, we compared the row-wise and column-wise distributions to the uniform distribution and optimized the cluster groupings using a nonlinear Lagrangian optimization algorithm to make them as close to the uniform distribution as possible. The more uniform the contingency table entries, the more likely the segment boundary reveals a qualitative change of characteristics.

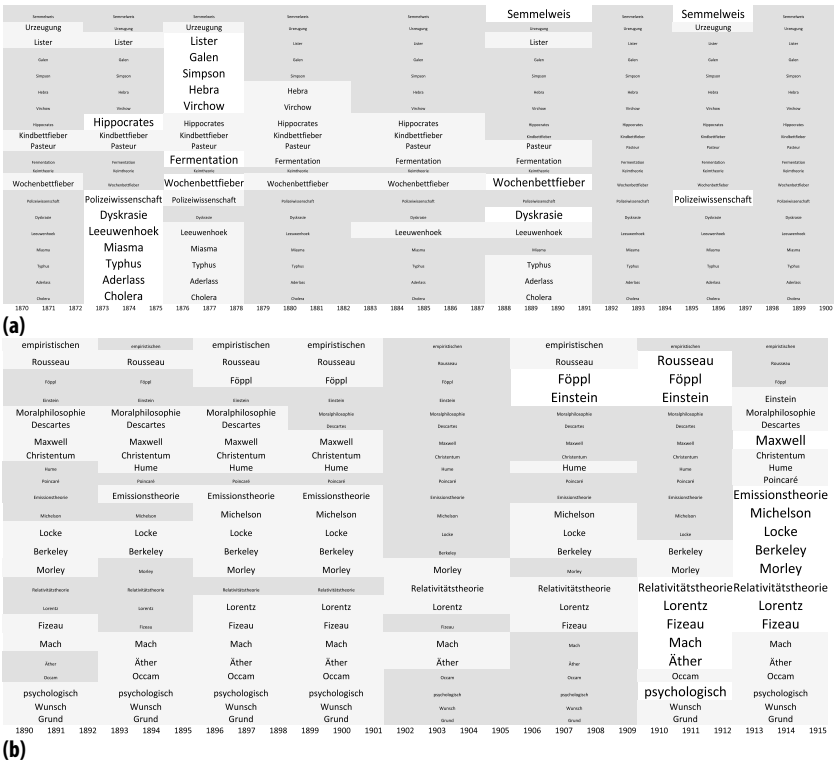

Figure 2. Segmentions of the use of (a) 22 words related to "Semmelweis" from 1870 to 1900 and (b) 22 words related to "relativity" from 1890 to 1915 in Google Books German.

AIDS

As the temporal clusters highlighted in Figure 1 show, in all three corpora, references to AIDS peaked around 1991, when the number of newly diagnosed persons with the disease was highest in the US (www.cdc.gov/mmwr/preview/mmwrhtml/mm6021a2.htm). The pair-wise linear correlation coefficient between incidence of new AIDS cases and use of the term "AIDS" was 0.85 in the *New York Times*, 0.67 in PubMed, and 0.63 in Google Books. Coverage of AIDS in the latter two corpora continued to remain high after 1991.

Google Books references to "AIDS" and "HIV" occur most frequently from 1987 to 1994, when clustered with words such as "gay" and "prostitution." Use of "gay" rose again in 2005, uncoupled from "AIDS." AIDS-related conditions such as "HTLV" (human T-lymphotropic virus) and "Kaposi" (Kaposi's sarcoma) quickly rose and fell in popularity.

In *The New York Times*, contiguous clustering segments were generally shorter than in the other two databases, perhaps indicating that news is a more fickle measure of society's interest in a topic and suggesting the need to aggregate over multiple news sources to reveal more stable trends. In this database, "AIDS" is clustered with words increasing in popularity in the early 1990s, such as "Congo" (one of the first countries to recognize AIDS) and "CDC" (US Centers for Disease Control and Prevention). Interestingly, use of "HIV" doesn't increase until 1997, perhaps because media outlets are slower to adopt new scientific terminology.

In PubMed, "AIDS" is clustered with terms such as "unsterile" as well as "HIV" and "prostitution," but not "gay." Segmentation of this database reveals phases in the understanding of AIDS, from classification in the early 1980s, to the search for antiretrovirals in the late 1980s, to the introduction of highly active antiretroviral therapy ("HAART") in the late 1990s.

SEMMELWEIS DISCOVERIES

Although Semmelweis showed that hand sanitation reduced the incidence of childbed fever as early as 1847, it wasn't until 1888 that the frequency of occurrences of his name in

German books accelerated, as Figure 2a shows.

As might be expected, usage of "miasma" and "polizeiwissenschaft"—a term for public policy that, among other things, embraced community health issues—increased during at least part of that period. However, at the same time, (Antonie van) "Leeuwenhoek," inventor of the microscope, was discussed with increasing frequency even though he died more than a century earlier. Mentions of physicians such as (Robert) "Koch," (Joseph) "Lister," and (Louis) "Pasteur" followed in the successive time window.

When "Semmelweis" did appear frequently in German literature, it coincided with "wochenbettfieber," or puerperal fever. Overall, not until after interest in "miasma" and "Hippocrates" died down did the term "Semmelweis" become truly common.

RELATIVITY THEORY

In the case of relativity theory, empiricist philosophers such as David Hume, George Berkeley, and Ernst Mach experienced levels of fame in Germany similar to or even greater than contemporary physicists prior to Einstein's 1905 paper on special relativity, as Figure 2b shows. Interestingly, increasing use of "relativitätstheorie" corresponded with growing popular interest in physicists such as Hendrik Lorentz, Albert A. Michelson, and Edward Morley.

Einstein himself cited Mach and Hume as inspirations for his work on relativity theory (M. Domsky and M. Dickson, eds., *Discourse on a New Method: Reinvigorating the Marriage of History and Philosophy of Science*, Open Court, 2010). It seems plausible that his theory became popular because of existing societal interest in empiricism.

ANALYTICAL RESULTS

In all three studies, temporal segmentation revealed that the popularity of a scientific event or discovery, as reflected in the Google Books and—in the case of AIDS—*New York Times* corpora, increased when that event or discovery was coupled with societal factors that were similarly famous in either the same or the preceding time window.

The use of terms describing many AIDS-related complications, including "Pneumocystis" and "Kaposi," in books and news reports increased along with references to "AIDS" as the incidence of the disease rose in the US. The frequency of terms such as "condom" and "prostitution," perhaps related to public policy attempts to control the spread of HIV, also increased during this period. It's possible that "gay" was in the same cluster as "AIDS" at this time because the disease became a rallying point for some in the gay rights movement (G. Troy, *Morning in America: How Ronald Reagan Invented the 1980's*, Princeton Univ. Press, 2005).

Semmelweis made his discoveries at a time when childbed fever was common. Why, then, might his findings have been ignored, when those of Leeuwenhoek regarding microbes—also at odds with miasma theory—gained in popularity in the 1870s? One possibility is Pasteur's discovery using a microscope that fermentation was caused by a microbe primed the scientific community for rapid subsequent discoveries tying microbes to diseases. The timing of these later findings coincided with similar discoveries and theories posed by Koch and Lister. In contrast, Semmelweis presented his findings directly to the medical community, and when his claims were disputed, he had little support to draw on (M. Best and D. Neuhauser, "Ignaz Semmelweis and the Birth of Infection Control," *Quality and Safety in Health Care*, June 2004, pp. 233 -234).

Relativity theory became popular, or at least highly debated, soon after its conception, along with "Äther" (aether) theory and the names of physicists such as Lorentz and Einstein. In retrospect, the years leading up to relativity theory created a "perfect storm" for such an idea to flourish. Interest in empiricist philosophers such as Hume and Mach combined with progressive discoveries in the physics community. Although aether theory was still popular during this time, experimentalists such as Michelson and Morley, who had laid the foundation for discrediting the theory, were already well-known. Thus, the way was paved for the previously unknown Albert Einstein to make a discovery that would be instantly popular and widely accepted in the long run.

After the fact, it's often evident when a great idea transforms society. Temporal segmentation of text corpora shows that it's possible to algorithmically infer a timeline of factors correlated with those ideas.

In the case of scientific discoveries, because it often takes years for findings to filter into popular literature, the databases we considered were appropriate for our study. A similar analysis of databases with finer temporal resolution, such as Twitter or Google Trends, might give researchers similar insight into public discussions in more dynamic fields, such as finance or politics. It might also be interesting to consider a set of terms derived entirely from popularity values, without the use of an organizational system such as Wikipedia.

Because our timelines ultimately report qualitative, historical phenomena without truly testing for causality, it's likely that in many cases historians and others with a deep, qualitative understanding will have to interpret the observed results. Thus, our work opens the door for increased cooperation between quantitative analysts and humanists. C

Justin Jee *is a second-year MD/PhD student in the Medical Scientist Training Program at the New York University School of Medicine. Contact him at justin.jee@med.nyu.edu.*

Lee Case Klippel *is a graduate student in the Department of Applied Physics and Applied Mathematics at the Fu Foundation School of Engineering and Applied Science, Columbia University, New York. Contact her at lck2122@columbia.edu.*

M. Shahriar Hossain *is a PhD student in the Department of Computer Science at Virginia Tech. Contact him at msh@cs.vt.edu.*

Naren Ramakrishnan*, Discovery Analytics column editor, is a professor and associate head of graduate studies in the Department of Computer Science at Virginia Tech, where he also directs the Discovery Analytics Center. Contact him at naren@cs.vtu.edu.*

Bud Mishra *is a professor of computer science and mathematics at New York University's Courant Institute of Mathematical Sciences, where he leads the Bioinformatics group, as well as a professor of cell biology at the New York University School of Medicine. Contact him at mishra@nyu.edu.*

This work was supported in part by US NSF grants SES-1111239, DUE-1122609, and CCF- 0937133.

Editor: Naren Ramakrishnan, Dept. of Computer Science, Virginia Tech, Blacksburg, VA; naren@cs.vt.edu