# 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.

he 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.

Human 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*. While promising, culturomics is viewed skeptically by humanists, who worry about the loss of critical traditions and the serendipitous role of the lone researcher who identifies valuable texts and develops original interpretations (A. Grafton, "Loneliness and Freedom," *Perspectives on History*, Mar. 2011; www.historians.org/Perspectives/ issues/2011/1103/1103pre1.cfm). Particularly concerning 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.

# **DISCOVERY ANALYTICS**



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, "gayrelated 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

108 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, respectively. We used three clusters in each segment to capture the overlap between clusters in a  $3 \times 3$  contingency table.

Finally, we compared the rowwise 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.

#### 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* 

Lemminum.	Seminariantia	Investorio.	Lerveriere.	Immediatio	9	Semmelweis	Secondari	. Semmelwe	eis terreters
Urzeugung	Granging	Urzeugung	innapog	Lhompung		Evenaport	innging	Urzeugung	Uningung
Lister	Lister	Lister	Later	Enter		Lister	Later	later	Later
Galer	lider.	Galen	Eden	Galer		Galen	Eden	Galeri	later.
Simpler	Singson	Simpson	Singson	Simpson		Singular	Simpson	Simpler	Simpson
Hebra	1902	Hebra	Hebra	1400		Robos	MDD	Hebra	1480
Vishes	Tinhow	Virchow	Virchow	Value		Weban	Tribow	Vishan	Vishow
Pippenstas	Hippocrates	Hippocrates	Hippocrates	Hippocrates		Reportates	Sporate	s Hoperados	Hypervies
Kindbettfieber Pasteur	Kindbettfieber Pasteur	Kindbettfieber Pasteur	Kindbettfieber Pasteur	Kindbettfieber Pasteur		Pasteur	Krales Rela	er Kodsettleber	Kecketficker
Pasteur	Pasteur	Fermentation	Fermentation	Fermentation		Fermentation	Padeut	Fatlaw	Padeu
Keimbearie	Reinstheorie	faimtheorie	Keintheole	fernenk		Keintheorie	Pertentado Keintheoria		Fernestzion Reimbeorie
ochenbettfieber	Mischenbertifieber	Wochenbettfieber	Wochenbettfieber	Wochenbettfieber	Wo	chenbettfieber	Washerberth	iber Wochenbetfleber	Notesberzfieber
Noieisenzhit	Polizeiwissenschaft	Polizeiwissenschaft	Polosialastadath	Pokasiwiaana3aR		Poloowiaanschaft	Polasiwiasnac	nn Polizeiwissense	chaft rotanianaturt
Dyelsciedar	Dyskrasie	Dyvlatio	DyAnnie	Dystatuse		Dyskrasie	Dydamie	Dysteriosise	Dyskruse
Leeventeek	Leeuwenhoek	Leeuwenhoek	Lenumenhoek	Leeuwenhoek		Leeuwenhoek	Lenunerbor	ri Leevaenteek	krevaranhork
Macmu	Miasma	Miasma	bizare	Manu		History	Marre	Maona	Mauna
Tahu	Typhus	Typhus	Typhia	Typha		Typhus	Typhus	Salva	Typhes
Admina	Aderlass	Aderlass	Advites	Adedas.		Aderlass	Advites	Adedata	Aderback
Dates 670 1871 1872	Cholera 1873 1874 1875	Cholera 1876 1877 1878	Outres 1879 1880 1881 188	Dates 2 1883 1884 1885 1886 1	887 1888	Cholera	Outru 1892 189	Chatew 3 1894 1895 1896	Doline 1897 1898 1899
empiristischen Rousseau	Rousseau	empirist Rouse				empiristisch Rousseau	en	Rousseau	empirizăzbos Reucosta
									Paucosa
Fópal	Fópel	Föp	ipl Föppl						
				Fogol		Föppl		Föppl	Fóppi
tinden. An en la bila en a bia	lintein A An an Imhillenn	tiez	ein Einzein	fiedes		Einsteir	ı	Einstein	Einstein
none Aoralphilosophie Descartes	Moralphiloso Descartes	phie Moralphil	en tierten Iosophie Menghlong	fiedes			ı		Einstein
Aoralphilosophie Descartes	Moralphiloso Descartes	phie Moralphil s Desca	en sinden losophie Moniphicas artes becartes	Einden Mosiphiosphe Desates		Einstein Merskhilosophie Descartes	1	Einstein Moniphiomphie Descartes	Einstein Moralphilosophie Descartes
Aoralphilosophie	Moralphiloso	phie Moralphil s Desca Maxy	en Sauser losophie Messivilos artes becares well Maxwe	tikase Monajhioaphe Deucane 211 Massel		Einstein	ı	Einstein	Einstein Moralphilosophie
Aoralphilosophie Descartes Maxwell	Moralphiloso Descartes Maxwell	phie Moralphil s Desca Maxy	es Exclesi losophie Moniphiloso artes Descret well Maxwe untum Christent	Excela bis bis bis bis bis bis bis bis bis bis		Einstein Monthilosphie Descartes Massel	ı	Einstein Monghilosophie Descarter Masseel	Einstein Moralphilosophie Descartes Maxwell Christentum Hume
doralphilosophie Descartes Maxwell Christentum	Moralphiloso Descartes Maxwell Christentur Hume Peinait	phie Moralphi s Desca Maxu m Christe Hur	es tautes losophie Monghoug artes Deacter well Maxwe intum Christent me Hume as Natari	bie Bioday Biographicagolia Bill Massel UIII Charae Second Harae Nacari		Einstein Mosthilosphie Desansi Massell Oriterium	1	Einstein Mosyblouphie Becanse Mitswell Onternom Hanne Folicard	Einstein Moralphilosophie Descartes Maxwell Christentum Hume Poincaré
Aoralphilosophie Descartes Maxwell Christentum	Moralphiloso Descartes Maxwell Christentu Hume	phie Moralphi s Desca Maxu m Christe Hur	es tautes losophie Monghoug artes Deacter well Maxwe intum Christent me Hume as Natari	bie Bioday Biographicagolia Bill Massel UIII Charae Second Harae Nacari		Einstein Monthilosphe Becartes Moneri Childentury Hume	1	Einstein Mosyblouphie Becanse Mitswell Onternom Hanne Folicard	Einstein Moralphilosophie Descartes Maxwell Christentum Hume Poincaré Emissionstheo
Moralphilosophie Descartes Maxwell Christentum Farce Farcert	Moralphiloso Descartes Maxwell Christentur Hume Peinait	phie Moralphi s Desca Maxu m Christe Hur	es taces losophie test intes becare well Maxwe intum Christent ne Hume as races stheorie Emissionst	terane des Montplotouples Desarter turm contention contention iteme nicone		Einstein Hosphilosphe Hosphilosphe Hosphilosphe Hosphilosphe Hosphilosphelosph		Einstein Mosyblouphie Becanse Mitswell Onternom Hanne Folicard	Einstein Moralphilosophie Descartes Maxwell Christentum Hume Poincaré Emissionstheo Michelson
Moralphilosophie Descartes Maxwell Christentum Farm Farm Farm	Moralphiloso Descartes Maxwell Christentur Hume Painait Emissionsthe	phie Moralphil s Desca Maxu m Christe Hun wee	es taces losophie temploag prites becare well Maxwe intum Christent me Hume as temploag stheorie Emissionsti elson Michels	te analise analise International analise an Internatise analise anali		Einstein Hersighlesphie Decares Unteretur Hume Naces Einakendherfe		Einstein Mosphiosphie Decare Massel Oriterium Romani Einschert	Einstein Moralphilosophie Descartes Maxwell Christentum Hume Poincaré Emissionstheo
Moralphilosophie Descartes Maxwell Christentum Iarre Parcent Envicentees Mitchen	Moralphiloso Descartes Maxwell Christentu Hume Penant Emissionsthe Webbo	phie Moralphil s Desca Maxu m Christen Hur Necorie Emissions Michee	en boom Josophie Mengehous urtes bewann well Maxwe intum Christent me Hum es Nauer stheorie Emissionsti vison Michels ke Locke	Ar Kenghangan Um Control Um Control Um Control Anac		Einstein Merskihlesphie Decater Missell Oktown Hume Nacow Kasawathewite Michelson		Einstein Monghloogha Decore Monet Monet Monet Monet Monet	Einstein Moralphilosophie Descartes Maxwell Christentum Hume Poincaré Emissionstheo Michelson
Aralphilosophie Descartes Maxwell Christentum Pares Parcea	Moralphiloso Descartes Maxwell Christentu Hume Penade Emissionsthe Webbo	phie Moralphi s Desca m Christe orie Emission: Miche Loc Berke	es Laora Josophie trets Muegewan well Maxwe ntum Christent Hume Hume stheorie Emissionst tskon Michels ke Locke eley Berkele	Are Executes Are Kennebhannete Executes Ell Stouest Comment Frances F		Einstein Merskinkashke Decare Misseel Oktown Hume Nacow Kasawashke ke Michelson Locke		Einstein Norwiphoughs Decore Contention Norwin Norwin Mitchen Berkeley	Einstein Moralphilosophie Descartes Maxwell Christentum Hume Poincaré Emissionstheo Michelson Locke Berkeley
Aoralphilosophie Descartes Maxwell Christentum Incon I	Moralphiloso Descartes Maxwell Christentuu Hume Emissionsthe Locke Berkeley Moteo	m Moralphile s Moralphile m Christe orie Emission Miche Emission Miche Berke	es taous Josophie Usersen well Maxwe ntum Christen as Ausers stheorie Emissionst ke Locke eley Berkele ley Morie	ter bener Hit bener Hit bener ter bener	rie	Einstein Neustrikauska Danars Hume Nauars Macars Michelson Locke Berkeley		Einstein Norderbesehe Dessense Norder Norder Norder Berkeley Morley	Einstein Moralphilosophie Descartes Maxwell Christentum Hume Poincaré Emissionstheo Michelson Locke Berkeley Morley
Acralphilosophie Descartes Maxwell Christentum record record Christentum technology record debten Locke Berkeley Morley	Moralphiloso Descartes Maxwell Christentu Hume Emissionsthe Locke Berkeley Nation	m Moralphile s Moralphile m Christer Hur reso orie Emission: Miche Loc Loc	n Goophie Goophie trets well Maxwe mutur Christent Hume mutur Historie Estissionsti Historie Estissionsti Historie Estissionsti Historie Estissionsti Historie Entissionsti Historie Estissionsti Historie Histori Historie Historie Historie Historie Historie Historie Historie His	na tanan Haran Sana Haran Sa	rie	Einstein Weinstein Deurstein Octoren Assure Assure Michelson Locke Berkeley Relativitätsthe		Einstein Waadhaaan Ordenen Naaroo Nao	Einstein Moralghilosophi Descartes Maxwell Christentum Humm Poincaré Emissionstheo Michelson Locke Berkeley Morley Relativitätstheo
Acralphilosophie Descartes Maxwell Christentum Insee Parced Executed Serkeley Morley Partstätzbeurs Losse	Moralphiloso Descartes Maxwell Christentuu Hume Emissionsthe Locke Berkeley Moteo	m Moralphile Besca m Christe Hur corie Emission: Miche Loce Berk Mor	n team coscyble coscyble coscyble well Maxwe well Maxwe well Maxwe Hurne Hu	an bankar Handright State Handright State Hearding H	rie	Einstein Weinstein Maart Connor Hume Naart Michelson Locke Berkeley May Relativitästhe Lorentz		Einstein Neuerin Stearen Stearen Stearen Stearen Stearen Morley Relativitätstheorie Lorentz	Einstein Moralghillosophi Descartes Maxwell Christentum Poincaré Emissionstheo Michelson Locke Berkeley Morley PRelativitätstheo Lorentz
Acralphilosophie Descartes Maxwell Christentum record record Christentum technology record debten Locke Berkeley Morley	Moralphiloso Descartes Maxwell Christentu Hume Emissionsthe Locke Berkeley Nation	m Moralphile s Moralphile m Christer Hur reso orie Emission: Miche Loc Loc	n team coscyble coscyble coscyble well Maxwe well Maxwe well Maxwe Hurne Hu	an bankar Handright State Handright State Hearding H	rie	Einstein Weinstein Deurstein Octoren Assure Assure Michelson Locke Berkeley Relativitätsthe		Einstein Nordenseke Konore Konore Konore Konore Berkeley Morley Korley Eleativitästheorie Elorentz Fizeau	Einstein Moralghilosophi Descartes Maxwell Christentum Humm Poincaré Emissionstheo Michelson Locke Berkeley Morley Relativitätstheo
Acralphilosophie Descartes Maxwell Christentum Insee Parced Executed Serkeley Morley Partstätzbeurs Losse	Moralphiloso Descartes Maxwell Christentu Hume Emissionsthe Emissionsthe Emissionsthe Emissionsthe Berkeley Notes Berkeley Notes	m Moralphile Besca m Christe Hur corie Emission: Miche Loce Berk Mor	n tour cosphie cosphie weekl Maxwe weekl Maxwe Hum thum	se biogramme se biogramme se biogramme se biogramme se biogramme se bi	rie	Einstein Weinstein Maart Connor Hume Naart Michelson Locke Berkeley May Relativitästhe Lorentz		Einstein Norman Second	Einstein Moralghillosophi Descartes Maxwell Christentum Poincaré Emissionstheo Michelson Locke Berkeley Morley PRelativitätstheo Lorentz
Acralphilosophie Descartes Maxwell Christentum Insee Parced Execute Berkeley Morley Noteve Exerce Fizeau	Moralphiloso Descartes Maxwell Christentu Hume Emissionsthe Locke Berkeley Mathatareas Serkeley Mathatareas	nterna de la companya	n town osophie Useewee well Mawe well Mawe well Mawe Himm e Emissionst ison Michels ke Locke teley Berkelt ke Locke teley Model that that the Make that ison Athen that the Make that the Make the Maket the	eerie beween week week week week week week week	rie	Einstein Weinpleiden Marker Marker Marker Michelson Locke Berkeley Relativitästhe Larentz Fizeau		Einstein Neidensee Second Orderan Neidensee Second Berkeley Berkeley Berkeley Lorentz Lorentz Fizeau Mach Äther	Einstein Moralphilosophilo Desartes Maxwell Christentum Hunaré Emissionstheo Michelson Locke Berkeley Morley Relativitätstheo Lorentz Fizeau Mach Ather
Acralphilosophie Descartes Maxwell Christentum Inves Inves Mohere ILocke Berkeley Morley Investigentes Esteres Interestigentes	Moralphiloso Descartos Maxwell Christentu Emissionsthe Berkeley Berkeley Nature Berkeley Nature Berkeley Nature Na	minimized in the second	n town osophie Useewee weell Mawe weell Mawe weell Mawe Hum e Emissionst isson Michels ke Locke ley Berkelt ke Monenter town t	eerie beween week week week week week week week	rie	Einstein Weinstein Waars Hume Fearentein Michelson Locke Berkeley Many Relativitätsthe Lorentz Fizeau Many		Einstein Norman Second	Einstein Moralaphilosophi Desartes Maxwell Christentum Poincaré Emissionstheo Michelson Locke Berkeley Morley Relativitätstheo Lorentz Fizeau Mach
Acralphilosophie Descartes Maxwell Christentum Energy More Berkeley Morley Morley Eleseu Fizeau Acres	Moralphiloso Descrites Maxwell Christentu Hume Emissionsthe Locke Berkeley Suite Suite Suite Nation	m Moralphi Berker m Christe orte Emission Emission Loc Loc Loc Erize Micht Mor Fize Ma Aa	torophie	se bier bier bier bier bier bier bier bie	rie	Einstein Burgehause Burgehause Burgehause Hurne Burgehause Berkeley Berkeley Berkeley Berkeley Berkeley Berkeley Burgehause Fizeau Burgehause Bargehause B	orie F	Einstein Neidensee Second Orderan Neidensee Second Berkeley Berkeley Berkeley Lorentz Lorentz Fizeau Mach Äther	Einstein Moralphilosophilo Desartes Maxwell Christentum Hunaré Emissionstheo Michelson Locke Berkeley Morley Relativitätstheo Lorentz Fizeau Mach Ather

(b)

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.

*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." AIDSrelated 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 popu-

larity 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.

fter 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. 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

CN Selected CS articles and columns are available for free at http:// ComputingNow.computer.org.