The Bionic Social Scientist:
Human Sciences and Emerging Ways of Knowing

Jonah Bossewitch
POLS G8247: Mass Mediated American/Global Politics
Prof. Brigitte Nacos
April 30, 2008

We have found a strange footprint on the shores of the unknown. We have devised profound theories, one after another, to account for its origin. At last we have succeeded in reconstructing the creature that made the footprint. And lo! It is our own.

Sir Arthur S. Eddington[1]

A few years ago I attended a presentation on the preliminary findings of an ethnographic study examining the impact of surveillance in high schools. The researcher highlighted numerous student interviews to convincingly argue that these surveillance measures were actually fomenting stress, aggression, and violence, contrary to their purported purpose. They based their claims on field notes gathered over the past few years, but had not yet published their research since they had not yet crunched their data into a numerical form demanded by most journals. I understood that an anecdote is not evidence, but I thought it was unfortunate that these findings were not more widely available. I began to wonder how some of the transformations in the analysis and explanation of complex qualitative data occurring in various sectors of the economy might apply to social science research.

Reformulating this question in more abstract terms I started thinking about the effects that technology is having upon epistemology, specifically the ways it is changing what is knowable and our ways of knowing. Some of these changes are more obvious than others, as researchers across disciplines are experiencing disruptive shifts as digital technologies accelerate multimedia production and communicative potentials. This essay does not focus on the epistemic shifts fostered by increasing in collaboration and peer-production. Instead, I concentrate on some of the ways that computational progress is assisting us in understanding complex phenomena. Engineering, mathematics, and the natural sciences have already witnessed significant methodological transformations with their increased reliance on computational simulation, data visualization, and exhaustive bruteforce explorations of problem spaces[2]. However, the social sciences have been slow to recognize the tectonic shift that could greatly transform their research designs and investigations.

A central activity of social scientists – encompassing disciplines including sociology, anthropology, cultural studies, educational research, and political science – is the transformation of qualitative data into evidence which supports a hypothesis. There are different ways of establishing social scientific confirmation, but at its core, a persuasive finding must provide a meaningful narrative which explains its relations. For over a century social scientific methods have relied heavily upon statistics as a tool for describing and establishing causal relationships in the study of human behavior[3].

In order to employ statistical implements, qualitative data must be exposed as discrete, measurable, elements which are capable of being manipulated by these tools. There is a certain degree of signal loss when complex data is reduced to the kinds of buckets that current statistical instruments are capable of manipulating. There has been a great deal of advancement in the sophistication of the instruments used to establish and study these the relationships between these distilled elements. There have also been many improvements in the tools used to collect and categorize the raw data. But, these improvements have been largely incremental, and there is an increasing need for social scientists to reexamine the process of evidence confirmation in the computationally rich context that we now inhabit

In this essay I consider three types of investigative practices common to the social sciences and contrast them them with similar types of analysis being conducted outside of the academy. I

consider the analysis of qualitative data in the form of interviews and ethnographic research, the structurally similar task of content analysis of the media, as well as the use of surveys and polling instruments that measure public beliefs and opinions . I also discuss some emerging forms of inquiry, such as multi-dimensional visualizations, network analysis, and mapping, that vast increases in storage and computational power have begun to make possible. I argue that asymmetrical efficiency is beginning to disempower social scientists, as they are increasingly falling behind advertisers, marketing professionals, and campaign strategists in their understanding of human motivation and behavior. Social scientists should actively pursue the adaption and adoption of the techniques being developed for the natural sciences and for the corporate and government sectors.

Qualitative Clarifications

It is important to qualify a few points before we embark on this exploration. First, the kinds of research instruments imagined in this essay do not rely on sentient artificial intelligence replacing the nuanced judgment of a human researcher. Instead, these tools should be conceived as cognitive prostheses – extending human capacities in ways that allow for a redistribution of attention and labor. It is likely that researchers using these tools will spend as much effort, or more, confirming their hypotheses based on their data, but they will be asking different kinds of questions, and presenting different kinds of evidence in support of their hypotheses. The technologies discussed are not science fiction, but are being used successfully in many different settings.

Second, the imagined methodological shifts should not be reduced to the perennial qualitative versus quantitative debate. Instead, these new styles of research may represent a middle path, where the quantitative calculations are rendered largely transparent, incorporated into the computational tools used to perform the research. We have the opportunity to introduce new ways of learning about human behavior that can do for the social sciences what the automatic gene sequencer has done for molecular biology. The automatic gene sequencer doesn't provide answers to questions about the function of a particular gene is, but it does free the researchers to devote their resources to *how* and *why*, instead of spending as much effort on the *what.* This direction does not dismiss the importance of empirical observation and quantitative representations - rather we can reimagine the tools that are used to mediate the raw data and inform our interpretations. In some cases, these new techniques may be used to help us target the questions and domains where traditional methods of analysis can be subsequently marshaled. Hopefully, as with the microscope and telescope, these new instruments will help us see deeper and farther than ever before.

The shifts imagined in this essay will involve a reorientation in the standards of evidence. Modern data visualization can be characterized as a form of media, capable of expressing a full range of subjective emotions, from joy to fear, excitement to dread. As Edward Tufte puts it, this story is "about how seeing is turning into showing, [and] how empirical observations turn into explanations and evidence"[4]. The nature of quantitative analysis will change as researchers begin to interact directly with visualizations of their data. The power of the spectacle to influence and persuade raises complex philosophical questions about the kinds of objective standards the social sciences aspire towards. Arguably, the current range of statistical implements are subject to many of the same critiques – "torture numbers and they will tell you anything," as the saying goes. But we are entering an era where publications can include the raw data, as well as the tools of analysis, allowing readers to examine the data and verify the conclusions for themselves. With abundant storage and minimal distribution costs, it is becoming feasible for researchers to show

their work in more detail, as opposed to merely describing their procedures and summarizing their results. Some of these methods may allow us to preserve more of the richness and complexity of the raw data, while still being able to observe comprehensible patterns in its assembly.

The Poverty of Rich Content

A great deal of valuable social research revolves around the study of people's expressions conveyed using natural language. Subjects are interviewed, student assignments are assessed, and transcripts of conversations are analyzed. Similarly, researchers studying phenomena in media and politics confront corresponding requirements, as news stories, public speeches, and historical documents are objects of study whose original form is natural language. The challenges involved in studying these materials are only compounded when the objects of study are multimedia assets, composed of audio, images, and/or video. Depending upon the nature of the research, it may be sufficient to study a text-based transcript, but often the subject's actions and expressions are relevant for interpretation.

These objects of study are often loosely structured and non-uniform, and are not been directly amenable to statistical analysis in their raw form. Traditionally, this kind of data is transformed into a form that can be subjected to quantitative treatment by systematic *coding*. Free form expressions are organized and categorized according to a schema determined by the researcher. The process for developing this schema may be iterative, as continuous and reflective analysis of the data may require an adjustment to the schema used to describe it.

The effort involved in coding a large corpus of content is very resource intensive, and multiple researchers often participate in these efforts. Portions of the corpus are often recoded by different researchers to verify uniformity and consistency in coding judgments. Software applications such as NVivo[5] have been developed to assist with the organization and tracking of these manual coding efforts, but the traditional coding workflow, requiring researchers to read, judge, and mark the entire corpus, remains essentially unchanged. Academic researchers have turned to software to streamline the bureaucratic management and logistical demands of working with large, complex, data sets, but have been slow to explore methodological shifts these technologies suggest.

Content Fantasies

To set the stage for this discussion it is worthwhile to consider a few non-traditional approaches to extracting meaning from large volumes of data, and then consider how technology may facilitate and accelerate these kinds of investigations. A point of departure for exploring the kinds of alternative methodologies that computational advancements may enable on larger scale is prefigured in a technique of content analysis practiced by the fringe discipline of psychohistory. Psychohistorians take a psychoanalytic approach to understanding the behavior and motivations of large groups. Their theories rely extensively on the influence of child rearing on the subsequent attitudes of adults, and they are concerned with questions of power and violence that dominate society. An important psychohistorical method is known as *fantasy analysis* where only the emotionally vivid metaphors are extracted from public documents and speeches, typically reducing the primary document to one percent of its original length[6]. The analysis of this verbiage, which is imbued with feeling and action, captures a sense of meaning in these primary sources that is sometimes lost in secondary analysis or spin. The technique differers from traditional coding since the representation of the primary object of analysis comes

directly from the original text, and is not constrained to a fixed vocabulary.

In recent years, a similar method has gained more visibility. Newspapers such as the New York Times have begun to regularly publish word frequency analysis of primary texts such as the President's State of the Union address, and have even released interactive tools that allow readers to explore these texts according to particular words or concepts[7]. Certain patterns of political rhetoric are easier to discern using this type of word frequency analysis, as themes such as peace and war, freedom and justice, and fear and love, emerge from the clutter as the President's message is carefully reinforced.

These representations visually rhyme with the culturally popular *tag cloud*, which has become a popular means of organizing and identifying information across the internet[8]. A tag cloud visually represents words with different weights by varying the font size, color, and placement of the word. Tag clouds are capable of comfortably representing dozens of weighted concepts, far more than a typical pie chart or histogram. As this form of representation gains recognition, it has introduced a new kind of shorthand for summarizing and capturing the important concepts in a text. The presentation of these findings do not take the form of tables of percentages or simple charts. Instead, recognizable patterns emerge without applying standard statistical techniques. As we begin to adjust to these new forms of presentation they allow us to ask different kinds of questions, and compare texts to each other in new ways.

Word frequencies, character, line, and paragraph counts, and new ways of displaying metadata represent different kinds of syntactic analyses, but do they lead to meaningful conclusions? Do these techniques merely present curiosities, or do they help us make convincing arguments that support substantive findings?

Perpetual Negotiation Machines

With its explosive growth and surprising successes, the Wikipedia project has gained a great deal of attention from researchers who have flocked to the project to study issues such as motivation, governance, collaboration, negotiation, consensus building, and conflict resolution. One of the attractions for studying the project is the fact that the entire history of editing activity has been saved and is readily available. This attraction simultaneously presents formidable challenges, as making sense of very large volumes of data can be extremely difficult.

IBM's Collaborative User Experience Research Group has developed a series of tools that offer new approaches to these challenges. The *History Flow* tool was designed to help understand how the Wikipedia community achieved its success through a in depth study of the evolution of its articles over time. The researchers were interested in developing a better understanding of cooperation and conflict, authorship and vandalism, and the dynamic interactions within the community.

> "Without the aid of history flow, it would have been a daunting task to piece together the collaboration patterns described here. The efficacy of history flow in highlighting patterns of behavior suggests that visualization is a technique well-suited to records of social behavior. One speculation is that social interaction is often characterized by mostly normal behavior punctuated by outlying abnormal episodes, and information visualization can be an excellent way to simultaneously show broad trends and outlying data points."[9]

The specific findings of this research are outside the scope of this essay, but it is significant to note that their methods did not involve traditional coding approaches. They created tools which allowed the researchers to visually examine and interact with various aspects of the raw data, and drew their conclusions based on the patterns that were apparent in these visualizations. It is hard to imagine a series of percentages or charts that would have readily exposed and supported these conclusions. It is also apparent that the dynamic relationship between the researchers and the raw data suggested multiple vectors of inquiry that would have been easy to miss in just a single pass over the data.

In theory, these findings could be corroborated by statistical techniques, but that step may not even be necessary if their conclusions are reproducible, consistent, and convincing without that style of analysis. Social research is often interested in trends and tendencies, and many statistical techniques offer a false degree of precision that is largely irrelevant for establishing these claims. Representing the relative stability or instability of an article in Wikipedia as a floating point number may be necessary as an intermediate computational step, but does not contribute much significance in the final presentation of the analysis.

Over the Horizon

The problem of making sense of very large volumes of qualitative data is not confined to the social sciences. The digital age has ushered in an explosion in record keeping, since 'remembering' is something that software is very good at. There is more data available for analysis than ever before in human history, as we are producing, capturing, and storing information at unprecedented rates. Lawyers, advertisers, intelligence analysts, and individuals are confronted with this challenge on a daily basis. It is not uncommon for litigants to introduce millions of documents into discovery, for marketing firms to analyze petabytes of digital footprints left behind by consumer's behavior, for intelligence operatives to wrestle with the deluge of communications, and for our own email inboxes to overwhelm us with torrents of correspondence. These sensemaking tasks share a great deal of structural similarity with typical social science and it is worthwhile to look at some of the tools being used to contend with this onslaught. Analogies across domains often inspire unexpected applications.

The promise of natural language processing has been an elusive dream of artificial intelligence researchers for decades. In the 1960's computer scientists believed that strong artificial intelligence was just around the corner. Although these grandiose visions have not materialized, the power of search, recommendation engines, and machine learning are beginning to fulfill, at least to a first approximation, some of these original fantasies.

To offer some perspective on the commonplace importance of these approaches, consider that Google's co-founder Larry Page considers artificial intelligence to be their core business[10]. Their stated mission, “to organize the world's information and make it universally accessible and useful,”[11] implies being able to automatically retrieve the answers to questions formulated using natural language. Their AdSense program[12], which programmatically analyzes content and attempts to place relevant advertisements alongside it, is a precursor to the prevalence of these techniques. While many examples of absurd placements have been reported, the system continues to improve and demonstrates the power and ubiquity of automated content analysis. Amazon's book recommendations and Netflix's movie recommendations are kindred spirits – personal recommendations that are based upon automatic classification, categorization, and conceptual proximity (of both users and content).

Large scale news organizations have also begun to rely on automated classification systems to relate similar stories to each other. A standard feature on the websites of major news outlets are the links to other related stories. These connections are increasingly being created automatically, since no single editor has a sufficient mastery of the entire archive to establish these connections. Some of these connections are established based on common human-generated metadata, but automated semantic analysis of news is improving quickly. In January 2008 Reuters opened up free access to its Calais system[13], which automatically extracts people, places, and businesses from news stories and annotates this information as machine readable metadata. This metadata is currently used to construct connections between related stories, and provide visual concept maps of the networks of ideas connected to the stories. The Open Text Summarizer (OTS)[14] is an open source text summarizer and auto-classifier that has received praise in several reputable computer science journals. It has been deployed in combination with the WordNet[15] semantic lexicon to create free software equivalents of the Calais system[16].

In the legal sector, high-end corporate law firms have embraced eDiscovery tools to help them grapple with the massive amounts of information that have become ordinary in litigation. These tools help manage the logistical challenges surrounding the management of complex workflows involving many collaborators, but they also help lawyers organize and analyze these primary sources. They are more than just glorified search engines, as they have begun to classify and group documents according to criteria specified by their users. They have been used to help pinpoint elusive needles in voluminous haystacks, and have also surprised researchers with unexpected connections.

In the marketing sector, massive data warehouses of consumer behavior are being assembled by companies such as ChoicePoint[17], Lexis-Nexis[18]. We will return to the potential research opportunities of these warehouses shortly, but it is useful to become familiar with the analysis tools that these companies promote to develop a better feel for how sectors such as law enforcement, campaign strategists, and commercial marketing firms are interacting with this kind of data. ChoicePoint's marketing materials claim that their tools set the standards for investigative analysis “from combating terrorism or fraud, to underpinning the development of sound business intelligence, our products are proven in law enforcement and the commercial sector worldwide. We offer solutions that give you the ability to meaningfully visualize, query and analyze even the most complex data. Our products facilitate a range of analytical techniques and supporting activities.”[19]

The security sector invests heavily in automated multimedia analysis, with applications in the analysis of foreign media broadcasts, automated real-time surveillance, and computer vision. Automated image classification, facial recognition and identity resolution has already gone public to consumers[20], and machine-based behavioral classification is following closely behind. Research groups such Carnegie Mellon's Informedia project[21] are working to automatically classify human behavior based exclusively on the optical information present in video streams (i.e. without relying on sensors or ubiquitous computing). They have deployed their system in nursing homes, and have been able to train it to automatically recognize behaviors such as eating, sleeping, and grooming, with more abstract behavioral patterns, like acting aggressively or suspiciously, in the works. This same group has also created a system for intelligence organizations to automatically index video news footage. Purportedly, casinos have deployed the most advanced automated behavioral analysis systems, but major cities such as Chicago and Beijing are rapidly following suit. They are currently being outfitted with automated networked surveillance systems, and though the behavioral classification components are in their infancy, the foundations of this infrastructure are in place.

The proliferation of these techniques is unsurprising given the enormous growth in the amount of records that are generated and recorded, and the interest and value in understanding the underlying behaviors. Social scientists need to be paying close attention to these developments for a number of reasons. First, it seems that many of these techniques can be adopted by researchers to assist in their own research analysis. Second, the politics of algorithms is subtle and insidious, but demands closer scrutiny. The prioritizing, omitting, and favoring of certain kinds of data over others is something that needs to be closely monitored. These automated systems are also effectively framing large sets of data – the authority to control the vocabulary used to classify human activities should not go unchallenged. Media scholars have long recognized the power wielded by the press in framing stories and issues – a parallel effect is being perpetrated by automated tools that are being used to describe behavior and media. Humanistic scholars must keep critical and vigilant watch over these developments if they want to shape alternative trajectories.

Mining for Precious Resources

In parallel to the accelerating advancements in computational analysis, the warehousing of data is also shifting the nature of investigation in research and industry. Large-scale scientific experiments such as supercolliders, space-based telescopes, and genome projects are generating massive data sets that are of significant interest to a wide range of scholars. The phrase *triple-bind experiments* is sometimes used to describe this emerging form of data collection"

> "In a double blind experiment neither researcher nor subject are aware of the controls, but both are aware of the experiment. In a triple blind experiment all participants are blind to the controls and to the very fact of the experiment itself. The way of science depends on cheap non-invasive sensor running continuously for years generating immense streams of data. While ordinary life continues for the subjects, massive amounts of constant data about their lifestyles are drawn and archived. Out of this huge database, specific controls, measurements and variables can be 'isolated' afterwards... This post-hoc analysis depends on pattern recognition abilities of supercomputers"[22]

Professor Deb Roy at the MIT Media Lab is currently capturing the first three years of her childs's life in an attempt to crack the mystery of language acquisition. The Human Speechome Project[23] is one of the first efforts of this sort to cross over to the psychological and social sciences, but it surely won't be the last. The raw data in this research will not only be of interest to linguists, but to child psychologists, educators, anthropologists, sociologists, etc. In many respects, this is an entirely new form of data collection, as in the past it was extremely rare for a particular collection protocol to provide useful material to other research agendas. But, as the scale of data collection increases, and the tools for mining ,extraction, and analysis improve, it is likely that these kinds of experiments will provide the primordial ingredients for a wide range of future, unanticipated research.

In a similar vein, Matt Pinsky the CEO of Blackboard, a popular Course Management System, has announced plans to develop a Networked Education Database (NED)[24]. This database will allow researchers to easily distribute and capture research surveys, where they will be shared and pooled for other researchers to analyze. While NED still relies on self-reported data, and doesn't intend to capture implicit preferences revealed by a student's activities within the system, it still represents a new way of thinking about data collection and analysis.

Consumer data is also being collected and warehoused at a scale that would shock most people who have not been paying attention to the profusion of privacy concerns that surface regularly in the news[25]. Advertisers, political strategists, and law enforcement officials are all interested in the identification and prediction of behavioral patterns based on the trail of digital breadcrumbs we leave behind whenever we interact with networked computers. The stakes in this game are high enough that market forces will continue to hone these techniques. Predictions need to be 100% accurate to stop terrorism, but only need to increase sales or voter turnout by a few percentage points to justify the expenditures necessary to underwrite these systems.

People's beliefs and opinions are explicitly and implicitly expressed across the mediascape, and advertisers have already fixated on these new targeted approaches to direct consumer marketing. The market value of internet companies like MySpace and Facebook are largely based on the perceived value of this consumer data. It is likely that Google has a better idea of where I will surf tomorrow than I do, and that T-Mobile has a better idea of where I will walk. And by now, Facebook may have discovered strong correlations between preferences for vampires or pirates and political affiliations. In the hands of advertisers, moments of desire and temptation are being surgically targeted with increasing precision. But, to effectively capitalize these opportunities, human motivations and behavior needs to be closely studied and explained.

Michael Turk, a GOP campaign strategist recently claimed that the Republican National Committee can pinpoint their voters based on the brand of whiskey they consume, the make of car they drive, and the kind of music they listen to[26]. Correlations between these variables may be established initially with self-reported survey data, but they begin to suggest ways in which surveys and polling instruments will soon be displaced by large scale data mining efforts.

The methods of social research pioneered by Paul Lazersfeld in the 1920s at Columbia University's Bureau for Applied Social Research is being quietly overturned, at least within industrial and government research. Empirical social researchers must pay close attention to these advancements if they want to remain accurate and relevant.

Critical Automatons

By interpolating this constellation of examples, we are in a better position to envision the kinds of instruments that social scientists should leverage in the twentyfirst century. Taking the natural sciences as an example, subdisciplines such as computational astrophysics and biology have sprung up to deal specifically with the increasing demands on researchers to customize the implements used to conduct research. We are beginning to see hints of these practices emerging in the social sciences, but mostly on an ad-hoc basis, not yet systematically. Many social scientists have not seriously considered the alternative research methodologies these technologies afford.

A researcher's ability to dynamically interact with their data with minimal effort will allow them to explore hypotheses within their problem space they may have otherwise overlooked. Playful immersion in data enables greater theoretical agility, creates tighter feedback loops between theory and observation, and helps cultivate stronger intuitions about the contours and dimensions of the underlying models. Processes which encourage researchers to iteratively and continuously refine the clusters of words they wish to code will catalyze increased flexibility and exploration.

The shift is analogous to the difference between completing your taxes using paper and pencil versus working on them using an application like TurboTax. A wider space of possibilities may

be explored, and new avenues of research may suggest themselves during these interactions. Ultimately, the researcher must still deliver a meaningful narrative from a point of view that is based on theoretical assumptions. But the as the process for generating and evaluating these narratives changes, we can expect that the kinds of questions and answers we are interested in will change as well.

Similarly, as social and communicative relations continue to change around us, research agendas are changing along with them. The analysis of networks is a recent example of shifting methods which relies heavily on the relatively new mathematics of networks.[27] A recent study of the Persian blogosphere confirms the emergence of some of the hybrid methodologies outlined in this essay. The study utilized network visualizations to target and guide the content analysis, which in turn, generated feedback used to refine and reconceptualize the visualizations themselves. Some of the studies' conclusions are supported by simply inspecting the patterns of clustering apparent in the visualization. The content analysis included a combination of traditional and computational methods:

> "Several types of content analysis are used to help interpret the cultural and political meaning of the Iranian blogosphere's structure. We worked with a team of Persian speakers to read and code hundreds of blogs using two questionnaires. We analyzed the frequencies of words and phrases in the posts of Iranian bloggers. And we spent hours sitting with culturally knowledgeable Iranians, looking at dozens of blogs in key positions on the map, as well as dozens of the news sites, organizations, and other online resources these bloggers link to. The results, quantitative and qualitative, portray a diverse network of online discourse, in which one can see the richness of Iranian culture and the clear footprint of political contention. The society's broad ideological divide is visible, as well as the more focused role of practical politics."[28]

Mapping the Mines

Alongside social network visualizations, another emerging form of analysis relies on geographical mapping to help discern meaningful patterns in environmental, spatial and temporal relations. Amy Hillier is a researcher on city planning, urban studies, and social work whose recent work has used maps to demonstrate the relationship between the availability of healthy foods and the incidence of obesity in poor neighborhoods. When this data was presented to local leaders, they claimed "I already knew that there was a problem, the map just made it real. It put a face on it. It was like an exhibit in a courtroom."[29] As Hillier argues

> "Mapping is critical to understanding how and why the environment impacts individuals, but at an even more basic level, maps can provide powerful evidence of disparity. Patterns that may not emerge in tables or that make much less of an impression when represented as summary statistics may be compelling to a wide audience when mapped."[30]

.The technological support for data mapping has improved tremendously in recent years, and it is likely that this form of expression will play an important role in constructing convincing arguments in future discourse. Maps have a long history of use in the social and political sciences, but their ubiquity and the increasing ability to overlay very rich and detailed information is revealing new insights in business and politics[31].

Changing the Guard

The changes anticipated in this essay are not the first time that technology has influenced the methods of social science. The widespread adoption of survey instruments, as well as the calculating mainframes which tallied up these results are an example of a similar process in the twentieth century. A more recent transformation is the reliance on programmatic 'search' in the study of historical archives. Many quantitative research papers in political communication studies begin with a description of the search terms used to locate their data set within Lexis-Nexis. An interesting followup study to this essay could examine the history of this adoption. What forces and possible controversies led to the widespread acceptance of 'search' as a legitimate method of discovery in scholarship? Perhaps an account of the adoption of search could provide insight and direction for the smoother adoption of newer methods.

Well established scientific standards insist that research be reproducible, raising the question of transparency and reproducibility when it comes to proprietary search engines. Without access to the underlying source code, it is impossible to verify that these engines are doing what they claim, reliably and consistently over time. Additionally, the matching rules are often opaque, fostering a lingering ambiguity over stemming, alternate spellings, and near misses.

Given these concerns it is imperative that social scientists embracing these methods demand the publication of raw data and the source code corresponding to the tools of analysis alongside the results. An explicit representation of the analytic rules will also radically change research opportunities. It is imaginable that identical methods could be applied to much larger data sets, as well as to future data sets that have not yet been gathered.

Conclusions

The transitions described in this essay represent a methodological paradigm shift, which will likely be subject to the contentious forces that characterize epistemic upheavals. Standards of confirmation have changed throughout the history of science, as the acceptance of new theories and instruments have disrupted the status quo. These periods of transition have often been difficult, as the conservative incumbents continue to cling to their traditional methods of thinking. They are reluctant to relinquish their authority, and hesitant to learn new ways of knowing.

As with other historical revolutions in epistemology, generational turnover will eventually lead to adoption. Until that point, innovative researchers will be under pressure to conform to older styles of analysis. Young scholars hoping to publish in respectable journals will be discouraged from taking risks on unproven methods. It is the tenured faculty who must lead the charge, using their authority and influence to help clear the path for innovation.

In the great debates between qualitative versus quantitative researchers, qualitative researchers have accused the bean counters of colonizing the human sciences with the tyranny of statistics. In turn, quantitative researchers have accused the armchair theorists of ephemeral speculation. These polar viewpoints are caricatures of a more sophisticated exchange, and most would concur that quantitative and qualitative studies inform one another in essential ways. However, we can imagine a greater reconciliation of these approaches – perhaps this conceptual revolution does not need to be bloody, but can help lead to greater peace and understanding.

[1]Eddington, Arthur. (1920). *Space, Time, and Gravitation: An Outline of the General Relativity Theory.* Cambridge, UK: Cambridge University Press.
[2]Ben Schneiderman (<http://www.cs.umd.edu/~ben/>) is a leading researcher in the area of information visualization, and companies like EnThought, Inc (<http://www.enthought.com/>) offer services focused on Scientific Computing.
[3] Durkeim, Emile.(1903) 1951. *The Rules of Sociological Method,* edited and introduced by Steven Lukes. Translated by W. D. Halls. London: Macmillan. Also, for a more detailed history, see Coven , Victoria (200). A History of Statistics in the Social Sciences. <http://grad.usask.ca/gateway/art_Coven_spr_03.pdf>.
[4]Tufte, Edward. (2006). *Beautiful Evidence.* Cheshire, CT: Graphics Press LLC. p. 9.
[5]<http://www.qsrinternational.com/>, formerly known as NUD*IST.
[6]deMause, Lloyd (1982). Foundations of Psychohistory. New York: Creative Roots.
[7]<http://www.nytimes.com/ref/washington/20070123_STATEOFUNION.html>. Also Brad Borevitz has released an innovative web project <http://stateoftheunion.onetwothree.net> which provides tools for visualizing the entire history of State of the Union addresses.
[8]The internet photo sharing site, Flickr <http://flickr.com> has been credited with the popularization of the tag cloud as a form of organizing large, diverse sets of photographs.
[9]Viégas, F., Wattenberg, M. & Kushal, D. (2004). Studying Cooperation and Conflict between Authors with History Flow Visualization. In *Proceedings of the 2004 conference on Human factors in computing systems.* New York: ACM. <http://alumni.media.mit.edu/~fviegas/papers/history_flow.pdf>
[10]For example, Stefanie Olsen, CNET News.com, (February 17, 2007), "Google's Page urges scientists to market themselves <http://www.news.com/2100-11395_3-6160372.html> (May 1, 2008).
[11]<http://www.google.com/corporate/> (May 1, 2008).
[12]<https://www.google.com/adsense/> (May 1, 2008).
[13]<http://www.opencalais.com/> (May 1, 2008).
[14]OTS can be downloaded at <http://libots.sourceforge.net/> (May 1, 2008).
[15]<http://wordnet.princeton.edu/> (May 1, 2008).
[16]For example, Mexico's leading daily newspaper, *La Jordana*, has implemented such a system using the Plone Content Mangement System <http://plone.org> and Object Realms' free Haystack product <http://svn.objectrealms.net/view/public/browser/ore.haystack/trunk>.(May 1, 2008).
[17]<http://www.choicepoint.com> (May 1, 2008).
[18]<http://www.lexisnexis.com> (May 1, 2008).
[19]<http://www.i2.co.uk/products/> (May 1, 2008).
[20]Greene, Kate. May 17, 2006, "Face Recognition Software Goes Public" <https://www.technologyreview.com/read_article.aspx?ch=specialsections&sc=security&id=16882> (May 1, 2008). At the time of writing this software was confined to the walled gardens of a few specific web sites, but it is bound to soon be unleashed across the wider internet.
[21]<http://www.informedia.cs.cmu.edu/> (May 1, 2008).
[22]Kelly, Kevin. (March 10, 2006). "Speculations on the Future of Science"<http://www.edge.org/3rd_culture/kelly06/kelly06_index.html> (May 1, 2008).
[23]<http://www.media.mit.edu/cogmac/projects/hsp.html> (May 1, 2008).
[24]Networked Education Database <http://edlab.tc.columbia.edu/index.php?q=node/904> (May 1, 2008).
[25] Stanley, Jay and Steinhardt, Barry. (2003) Bigger Monster, Weaker Chains, The Growth of an American Surveillance Society, ACLU Technology and Liberty Program. <http://www.aclu.org/FilesPDFs/aclu_report_bigger_monster_weaker_chains.pdf> (May 1, 2008).
[26]Michael Turk, (Keynote address at *Politics: Web 2.0*, London April 19, 2008) http://newpolcom.rhul.ac.uk/politics-web-2-0-conference/ (May 1, 2008).
[27]Barabasi, A. L. (2002), *Linked: How Everything is Connected to Everything Else and What It Means for Business, Science and Everyday Life*, New York: Plume.
[28]Kelly, John Etling, Brucee. Mapping Iran's Online Public: Politics and Culture in the Persian Blogoshere. 2008. The Berkman Center for Internet and Society Publication Series. Research Publication No. 2008-01 <http://cyber.law.harvard.edu/sites/cyber.law.harvard.edu/files/Kelly&Etling_Mapping_Irans_Online_Public_2008.pdf> p. 9.
[29] Charlse, Dan (January 31, 2007). "Group Maps City Access to Healty Foods" NPR. <http://www.npr.org/templates/story/story.php?storyId=7097476> (May 1, 2008).
[30]Hillier, Amy (2007) Why Social Work Needs Mapping. *Journal of Social Work Education Volume 43, Issue 2, July 2007, pages 205-221.* p. 210. <http://works.bepress.com/amy_hillier/9> (May 1, 2008).
[31]Stamen Design's projects are good examples of work along these lines <http://stamen.com> (May 1, 2008). They a boutique design agency that specializes in creating live views of complex data sets. Their clients include Fortune 500 companies, government agencies, and museums.