Of Mining and Mine Fields: Revolution in Paradigms of Data Analysis and Interpretation[1]

Dipak K. Gupta
Distinguished Professor Emeritus, Department of Political Science
San Diego State University

Brian Spitzberg
Distinguished Professor School of Communication
San Diego State University

Ming-Hsiang Tsou
Professor, Department of Geography
San Diego State University

Mark Gawron
Professor, Department of Linguistics
San Diego State University

An Li
Professor, Department of Geography
San Diego State University

Introduction

There is a revolution taking placing in how social scientists research aggregate and synthesize the data and observations that constitute our understanding of social phenomena (Lazer et al., 2009; Shaw, Yu, & Bombom, 2008). Analyses of social phenomena received an unprecedented boost after the invention of high-speed computers. Prior to the 1960s most significant advances in theory building in social sciences were relegated largely to what we may charitably called, "informed speculation." Karl Marx did not have to bother with the problem of falsifiability of

[1] Research for this article was supported by "Mapping ideas from Cyberspace to Realspace," funded by National Science Foundation, Division of Computer and Network Systems, NSF Program CDI-Type II Award # 1028177. Opinions expressed are those of the authors and not necessarily those of the National Science Foundation.

his proposed theory (Popper, 1980). John Maynard Keynes advanced his general theory without formally statistically testing his hypotheses; he could not empirically demonstrate the link between aggregate demand and economic activities, the central tenet of his argument. Charles Darwin required months at sea, years more of correspondence with fellow scientists and physical examination of specimens, as well as experience with myriad of exemplars to fully formulate the framework of natural selection theory. The social movement theorists of the period had to rely primarily of perceived plausibility of their postulates or attempt to offer proof by conducting small case studies. There were a number of significant problems for empirical verification of posited hypotheses, the hallmark of scientific reasoning (Bostrom, 2003; Chow, 1992; Gilman, 1992; Pavitt, 2004). All of these classic achievements of 20th century social science and theory occurred in the absence of computing technology. It is no accident that they were in their nascent forms rich with ideas and poor of data.

First, without the computer, collection of large-scale aggregate data and their analyses were problematic, extremely time consuming, and frequently impossible. The first installation of commercial computer, UNIVAC I, took place in 1951 in the U.S. Bureau of Census. Soon thereafter, a similar machine, employed by the CBS was used for the first time to predict the 1952 presidential election. With the help of the new technology, 1 percent of the US population was surveyed and the TV news organization was able to correctly predict the outcome of the election: the win of Dwight D. Eisenhower (Gupta, 1994).

As noted by Bell, Hey and Szalay (2009), "Today, some areas of science are facing hundred- to thousand-fold increases in data volumes from satellites, telescopes, high throughput instruments, sensor networks, accelerators, and supercomputers, compared to the volumes generated only a decade ago" (p. 1297; see also Gleick, 2011). One specific attempt to estimate information available to the average U.S. consumer, relative to the amount of information consumed, has grown from 82-to-1 in 1960 to 884-to-1 (Neuman, Park, & Panek, 2012). The problem is no longer obtaining information, but obtaining clarity in the midst of information overload.

A New Era of Computing Capabilities

With such developments in information retrieval, social science has entered a new era of potential empirical verification and prediction. The invention of computers enhanced the capabilities to analyze quantities of numbers beyond any previous human capabilities (Brier & Hopp, 2011; Johnson, Dunlap & Benoit, 2010). Multivariate regressions, which were relegated largely to theoretical development, suddenly became commonplace in estimation. With increased computing capabilities came much more sophisticated statistical techniques. With enhanced ability to crunch numbers, a virtual explosion began to take place in the collection of data on innumerable aspects of life[2]. For instance, today, a nation's development does not need to be defined solely on the based of per capita GDP. It is definable much more broadly with multi-faceted Human Development Index.[3] There are

[2] See, e.g., http://quantifiedself.com/

[33] http://www.nationmaster.com/graph/eco_hum_dev_ind-economy-human-development-index

cross-national indicators of state failure[4], corruption[5], and even gross national happiness[6].

Despite these breathtaking developments in data collection and computing capabilities, traditional social science research methodology suffers from some significant shortcomings and challenges (Ramakrishnan & Grama, 1999; Yang & Wu, 2006). These challenges stem less from the constraints of computational capability or methodological sophistication, but from the process by which social science data are collected.

Traditional Social Science Methodology

Most social science data are collected through one of two ways. They are either collected from observations of direct indicators in the "real world" (e.g., votes counted in an election, instances of violence in the media, etc.) or sample populations are surveyed to learn about their preferences and attitudes (Gupta, 2010). In either case, the process can be extremely expensive, labor-intensive and time-consuming. The decennial census data, the most accurate portrayal of life in the nation takes years to collect, collate, and to publish. Even the quarterly data are lagged by time. Surveys – unless constituted by routine questions such as daily presidential approval ratings – take a long time to design, implement, and to analyze.

---

[4] http://www.foreignpolicy.com/articles/2011/06/17/2011_failed_states_index_interactive_map_and_rankings
[5] http://www.transparency.org/policy_research/surveys_indices/cpi/2010
[6] http://www.nsf.gov/news/newsmedia/pr111725/pr111725.pdf

While data collection traditionally has taken a long time, the world events in the age of super connectivity and spatial mobility are moving at a geometric rate. In the past, it has generally taken years for collective movements to take shape. The Civil Rights movement took at least a couple of decades to reach its height. Al-Qaeda, similarly, had to struggle for years before it was recognized on the global stage. Compared to these, the lightning speed with which the so-called "Arab Spring" spread through the Middle East (Elson et al., 2012; Howard & Hussain, 2001; Murphy, 2006; Spitzberg et al., 2012) and the "Occupy Wall Street" (Ladhani, 2011) spilled over most of the developed world is virtually unprecedented (see, e.g., Hwang et al., 2006; Postmes & Brunsting, 2002; Rohlinger & Brown, 2009).

The rapid diffusion of ideas across various social domains is taking place primarily due to two important factors, both of which relate to the advancement in computer technology and electronic media. The first relates to the physical nature of the innovations whereas the second pertains to the psychology of the users. The ease of communication has indeed made efforts to communicate with a large number of people extremely cost effective and efficient (e.g., Andrés et al., 2010; Huang & Chen, 2010). In the past, people needed to have a "megaphone," offices or institutional pulpits that often came at the heavy expenditure of time and money. As a result, those who were in positions of public prominence had a "bully pulpit" from which to address large crowds or whatever constituted the local or national mass media of the day. Others, attempting a grass-root movement had to slog through years of work. In the contemporary technological context, messages can be spread through the web and myriad social media outlets and can put forward ideas that

have not been previously vetted through years of socio-political group activities (e.g., Flache & Macy, 2011; Postmes & Brunsting, 2002).

Second, as Olson (1968) pointed out, there has traditionally been a psychological impediment in voluntarily organizing large protest movements, known as the "collective action" problem (Hunt & Benford, 2004). Some of the collective action problems resulted from the fact that the costs of exposing contrarian viewpoints would fall primarily on those who would dare to take the first step, whereas the benefits, if the common good is procured, would flow to everyone in the community. As a result, the "collective action problem" would dictate that even when there is a deeply help desire by a large segment of the population for a significant social change, such demand would probably not even be properly articulated (Gupta, 2001). Compared to contemporary capabilities to give voice to risky beliefs or to evidence physical group affiliations, the risks of having to incur stiff costs would weigh heavily in the minds of the early activists (Watts & Dodds, 2007).

In contrast, many activists feel—albeit often erroneously—a sense of anonymity when computers and electronic media are relied upon for the purposes of political mobilization (Lewis et al., 2008; Rohlinger & Brown, 2009; Saldarini & DeRobertis, 2003). Most people tend to have a barrier between their public pronouncements and their privately held beliefs (Kuran, 1988; Petronio, 2002). People possess different levels of their own personal "threshold point" of tolerance for the status quo. When these thresholds are breached, people come out of their cocoons and join a mass movement. The society may appear to be extremely stable

until the day when people would come out and join the revolution. Significant "cascades can only occur when the influence network exhibits a 'critical mass' of early adopters, … who adopt after they are exposed to a single adopting neighbor" or peer (Watts & Dodds, 2007).

Given that neighbors and peers often engage in false expressions of their actual preferences until a sufficient threshold of peers has adopted the actual preferred position (Kuran, 1989, 1995), radical positions are particularly risky in the context of the status quo. That is why Kuran argued, that it is impossible to predict the demise of established political systems, such as the Soviet Union or the Shah's Iran, Mubarak's Egypt or Assad's Syria (see also: Taleb & Blyth, 2011). Since many users, especially the neophytes believe that the computer accords greater anonymity than face-to-face interactions or traditional print and video media, it is possible that in the present condition of technological advancement the threshold level has been reduced considerably (Douglas & McGarty, 2001; Spitzberg et al., 2012). With so many being willing to express their deeply held ambivalence or resistance online, more and more are emboldened to join such voiced causes, thereby forming large political movements almost out of thin air.

The Problem of Traditional Research Methodology

In view of such an altered world, it is clear that the social sciences and policy studies will need, and find, newer and far more sophisticated methods of collecting, extracting, synthesizing, and analyzing data to explain and predict social processes (e.g., Apinar et al., 2006; Bell et al., 2009; Cannataro & Talia, 2003; Lazer et al.,

2009; Shaw et al., 2008; Zimbra et al., 2010). Given the lag in collecting and disseminating traditional data, the traditional research method is akin to astronomers looking at a distant star; the light rays that hit the telescope are light years in the past and say nothing about its current condition. Thus, data mining will be increasingly necessary as an arrow in the arsenals of social science methodology. Human nature may not have accelerated, but the ability of the species to observe it and communicate it has.

Revolution in Social Science Methodology

The word "data mining" was once a dirty word in the social science research, where it was seen as a way of fishing for answers by randomly examining a lot of information without proper theoretical background. However, in the field of computer-assisted research, the term, implying collecting information through monitoring the Internet, is finding new respectability (Leetaru, 2012; Ramakrishnan & Grama, 1999; Yang et al., 2006). As a result, the monitoring of the web sites, the Twitter and the like is open up new methods of understanding social interactions and collective behavior (e.g., Bai, 2011; Crandall et al., 2010; Erickson, 2010; Li & Wu, 2010; Suh, Hong, Pirolli, & Chi, 2010; Takhteyev, et al., 2011; Zimbra et al., 2010).

The monitoring of communications on the Internet creates a deluge of data (Bell et al., 2009). In the past, where researchers used to worry about the sample size, the methods of data mining yield millions of observations. Therefore, the first challenge for the researchers is to classify these in a meaningful way so that we can

make some sense of our collective mood, positions, and opinions (Altaweel, Alessa, & Kliskey, 2010; Brier & Hopp, 2011; Cao, 2010). For instance, by monitoring the Twitter traffic, a group of researchers claimed to have observed the world's mood swings during the week, the year, and in response to momentous world events (Miller, 2011). Other research has demonstrated that Tweets can predict stock market fluctuations (Bollen et al., 2011). Sentiment analysis of online political discussions can accurately predict public opinion shifts in reference to presidential popularity and attitudes toward public policy (González-Bailón, Banchs & Kaltenbrunner, 2012). Such research relies not on the direct observation of moods or peoples' self-reports about how they feel, but on empirical proxies or surrogates that imply what individuals might be feeling. Critics abound; information culled from the cyberspace may indeed be perfunctory, misleading, or worse (Brasch, 2005). This, however, is similar to understanding economic achievement of nations by measuring illumination at night from satellite imagery. Or, for instance, the tonnage of trucking or orders for packaging materials can provide an important clue to the future economic activities of a nation. These may not be perfect indexes, but when used judiciously they can be sufficiently good indicators of what we are attempting to measure.

Web sites, Social Media and the Twitter

Data can now be collected from various new media sources. Among the most informative of these sources for social scientific at this time are (a) the contents of World Wide Web postings, (b) social media, such as open pages of Facebook, and (c) social media such as Twitter, Flickr, Pinterest, etc. There are, however, some

important differences among these sources of information (e.g., Papacharissi, 2009). For example, traditional statistical models are well-suited for analysis of the *point-based* data of the internet, but hyperlink and social media tend to emphasize *link-based* data, which tend to require network-analytic tools (Barabási, 2002; Lewis, 2009).

Are there differences among such media sources in terms of how people reveal their inner thoughts? Some of these differences are more obvious than others (Attrill & Jalil, 2011). For instance, while the Internet allows for expansive explanations of the posters' positions, the Twitter has a strict limitation of 140 characters. While many Internet sites have a "private members only" area or other restrictive privacy settings, the messages sent by Twitter are totally public, although such contents are stored for a relatively short period of time. During this time, the data are available to researchers. Twitter's reliance on hash tags allows search programs to trace tweets and followers' re-tweets, which allows researchers to seek insights into social interaction and networking by the users.

The new technology is finding new uses at an ever-increasing rate. A number of studies, for instance, have found chatter on Twitter to a strong predictor of election outcomes (Huberty, 2012). The current grant from the National Science Foundation, which allowed us to conduct research for this article is an example of concerted effort by various governmental agencies to expand the horizon of social science research based on monitoring of the Internet communications. The Intelligence Advanced Research Project (IARPA), has similarly funded a project to develop early warning systems for predicting protest demonstrations, spread of

epidemics, and sudden economic downturns for Latin American countries, based on open source information on the Internet.[8]

The "Buzz" and the "Narrative"

When eavesdropping on peoples' electronic conversations, interest tends to seek the buzz and the narrative. *Buzz* is defined as a concentration of postings at a certain location. Thus, relatively autonomous search algorithms can identify clusters and trends of postings and their discrepancies form baseline patterns, thereby indicating in almost real time the likelihood of collective activities in real space, such as large-scale events or flash mobs (e.g., Lee, Wakamiya & Sumiya, 2011). Other research on large-scale data sources such as Twitter demonstrate substantial pattern interdependency between social network connection and airflight patterns (Takhteyev, Gruzd, & Wellman, 2011). In such cases the causes of the buzz can be increasingly inferred from statistical methods using a number of demographic and other economic, political, or sociological independent variables (Hand, 2000; Jones et al., 2008).

The problems of geolocation of the Internet data are varied (e.g., González et al., 2008; Jones et al., 2008; Song et al., 2010; Takhteyev et al., 2011). The first problem is to know where the poster of the information is actually located. The task of pinpointing the geolocation depends on platform on which it is being done. For instance, if web sites are searched, each web site contains registration information, which can then be retrieved and by converting these into latitude and

[8] See http://www.economist.com/node/21553006.

longitude data, they be plotted on a map (Tsou et al., 2011). A number of commercial outfits provide such services. Unfortunately, simply because an IP address can be located on the world map, it does not necessarily imply that it reflects the true location of the person posting a message. However, our own research suggests that unless the posters deliberately want to hide their location, locational inferences from the raw data are approximately 90% accurate (Tsou et al., 2011).

Semantic sensitivity of search

A particularly challenging aspect of such megadata-based social scientific methodologies is casting the appropriate level of semantic discrimination in the processes of search, extraction, and interpretation (Alexander, 2009; Arpinar et al., 2006; Corman et al., 2002; Nebot & Berlanga, 2012). A small variation in the spelling or the keyword can provide radically different results. For instance, a search for activism or terrorism predicated on radicalized interpretations of religious texts can use either the keywords of "Koran" or "Quran" or "Qur'an." The former yields websites that are less religious or even anti-Islamic sites. These sites are located mostly in the Western countries. In contrast, a search by "Quran" or "Qur'an" provides more Islamic sites. Similarly, by adding a single term, the charter of the search is changed substantively. Thus, for example, the name al-Awlaki, a firebrand American born Yemeni cleric who was killed by US drone attack can be searched simply by name, or the honorific title can be added to his name. It turns out that"Sheikh al-Awlaki" returns a completely different sets of sites and postings than "al-Awlaki" alone. Searching the internet for "Arab Spring" in English presents a

substantially different geo-spatial density map of web content than searching for the same term in Arabic (Spitzberg et al., 2012). Finally, our research demonstrates that while the term "Global Warming" brings in a lot of postings by the skeptics, we find a different crop of people using the term "Climate Change" (An et al., 2011).

Beyond Sentiment analysis

In this new methodological revolution, sentiment analyses become crucial in understanding the evolving narrative (Fortunati, 2009; Li & Wu, 2010). They can provide invaluable information regarding variety of topics from product placement (e.g., "AT&T network is slower and has more dropped calls") to the outcome of an election, where each contestant wants to pin a narrative around an opponent (e.g., "Romney is an out-of-touch billionaire" or "Obama is a socialist"). Yet, when it comes to collective actions in terms of a political movement, we may dig a bit deeper into human motivation.

The "rational actor" hypothesis, arguably the most widely accepted assumption in the social sciences, tries to explain patterns of human behavior as the natural result of individuals acting in their own individual interest. The idea is an organizing principle in disciplines as diverse as economics, artificial intelligence, psychology, and linguistics. Yet the growing field of social psychology starting, *inter alia*, with the seminal work of Tajfel (1978, 1981; see also: Brown, 2000; Hogg et al., 1995; Korte, 2006; Turner & Reynolds, 2001) is busily accumulating evidence of the importance of groups and group in people's decision-making processes (Watts & Dodds, 2007). Decisions are heavily influenced by the group(s) in which

membership is claimed (McGlone & Giles, 2011; McFarland & Pals, 2005; Mullen et al., 2001; Reid & Giles, 2005). Group and collective identities can supersede individual identity; people often embark upon courses of action detrimental to their personal economic well-being, liberty, and life itself. This possibility is strongest in groups in which the sense of self-identification is strongest. Evidence abounds that "a strong identification with a collectivity makes participation on behalf of that collectivity more likely" (Hunt & Benford, 2004, p. 437). The key implication is that *individual and group identities are manifest in symbolic expression in electronic texts, and such texts are systematically searchable and reliably identifiable en masse* (Gawron et al., 2012).

A reasonable starting assumption is that intense group identification requires a clear articulation not only of who "we" are, but also who "they" are— the outsiders, the other, the "out group", often, the enemies — an articulation that is central to all large-scale collective action from nationalism (Anderson, 2003) to terrorism (Gupta, 2008). Collective identity implies multidimensional features (Hunt & Benford, 2004), including: the identification of boundaries, consciousness (i.e., common interpretive frameworks), and negotiation (i.e., developing and sustaining symbols, artifacts, and interactions that reinforce such identities). In some cases a group is defined by a pre-existing language, but in most cases it is not; whether it is or not, an essential part of the process of dividing us from them is developing a group sublanguage. This may have a complex array of linguistic components, ranging from phonological to syntactic features, but an essential part of it is evaluative language referring to us and to them, as well as language referring

to properties of us and properties of them (Little et al., 2003).

For well-established groups with a longer history the language includes a complex set of references to heroes, leaders, victims, and artists, as well as to subgroups, key events, key dates, and key writings and key works of art, including music and games. Although group formation requires identification of "us" and "them," the mobilization of a large number of people for collective action requires a third factor: a clear articulation of an impending existential threat (Gupta 2008). This is not an unintuitive result. Behavioral research by likes of Kahenman, Slovic and Tversky (1982) demonstrates the dominance of prospective losses over gains in the evaluation of uncertain futures. In brief, the prospect of loss of what is currently possessed is a far more potent motivator than the prospect of gains (Kahneman & Tversky, 1979; Novemsky & Kahneman, 2005). As a result, from political extremism to electoral politics, fear tactics are a winning strategy (de Hoog, Stroebe & de Wit, 2007; Witte & Allen, 2000). In accord with this idea, a well conceived "us-versus-them" analysis will target language articulating threats as well as language referring to the enemy.

For example, the in-group for white militant and hate groups is members of the white race, whereas the out-group or enemies are the non-white population, including Jews and Catholics for some groups, but significantly, also a group of white people who are perceived as traitors to the race. The general existential threat is the degradation and pollution of pure white stock, but there are many more specific instantiations because degradation has many potential dimensions and strategies. A semantic ontology can be developed, relying on domain expertise, incorporating

elements of the militant group argot and linguistic iconography referring to us-versus-them (Chau & Xu, 2007). The ontology is further expanded to include properties and products of us-versus-them and to existential threats to "us." The hypothesis is that the elements of the us-versus-them language are strong markers of group identity. Moreover, the us-versus-them language is largely learned, with more experienced speakers using it more fluently and more frequently. Communicators who control a significant subset of this language are likely to be well-established in the group. Identifying a significant set of such markers in a text provides strong *prima facie* evidence of core group membership, such as high degrees of militancy. Subsequent research that directly inspects the websites extracted from the search processes can help to validate the accuracy of such attributions.

Another example of group identification and its social implications is entailed in the analysis of social movements. From a diffusion of innovations perspective (Compeau et al., 2007; Elkink, 2011; Meade & Islam, 2006; Rogers, 1983; Rogers & Kincaid, 1981; Vishwanath & Chen, 2011; Young, 2009), ideas are a form of diffusion innovation, and as such, can be understood in terms of their adoption curves within populations (Earl, 2010; Marquette, 1981). Social movements represent collective efforts to diffuse a particular vector of ideas, beliefs and values. Given the increasing use and reliance upon various electronic media for the mass diffusion of ideas (e.g., Carty, 2010; Diani, 2000; Earl, 2010; Oliver & Meyers, 2004; Soule, 2004; Stein, 2009; Strodthoff, 1985; Van Laer, 2010), the diffusion of social movements may be potentially mapped in almost real time, if the semantic contents (i.e., ontology) of

such ideas can be discriminated from the background of other ideas. The key elements of this process include ontology development, search processes, and subsequent iterations of interpretation and refinement.

Preliminary investigations of variants of the phase "Arab Spring" in English and Arabic have demonstrated an ability to reveal sensitive changes in the diffusion of democratic concepts throughout various geographic regions in the Middle East (Spitzberg et al., 2012; see also Etling, et al., 2010; Howard & Hussein, 2011). Eventually, correlation of such patterns over time and space, carefully intersected with the features of (Hwang, Schmierbach, Paek, Zuniga, & Shah, 2006), and communication strategies employed by, the protesters (Van Laer & Van Aelst, 2010) and the thresholds of communication activity achieved, may well provide reasonably predictive models that can differentiate stalled from successful revolutions, policy or authority shifts.

Political suppression and Issues of Civil Liberty

As this new methodology in social science research develops, it opens up new possibilities along with possible danger of its misuse (Gandy, 1993). For obvious reasons, law enforcement authorities from all over the world have taken notice of the potential power of data-mining. There are risks that the most powerful processes of data mining may "become the exclusive domain of private companies and government agencies" (Lazer et al., 2009, p. 721). The power with which such media can be increasingly used to geo-locate individual and group activity with spatial referents (Erickson, 2010; Tillema et al., 2010) portend dark possibilities of abuse.

Apart from the Arab Spring, which is changing the political landscape of the entire Arab/Muslim world and the "Occupy Wall Street" movement, changing the course of national discussion about poverty and income inequality, social media and twitter have also helped spawn "flash mobs" with quick and instantly spreading violence among several of the world's capitals (Massaro & Mullaney, 2011). This has sounded alarm bells, particularly in the authoritarian nations. The Egyptian authorities realized the power of the social media a bit too late. By the time they attempted to pull the plug on the Internet traffic, the damage had already been done. Therefore, the surveillance of the Internet traffic and their regulation have become an essential part of the authoritarian nations' law enforcement apparatus, where they not only try to control the flow of the information, but also aim at suppressing legitimate political dissent.

The Infancy of Research and the Need

The world is wired, wireless, and increasingly linked (Barabási, 2002; Dodds et al., 2003; Watts, 2003) and therefore, increasingly interdependent in complex ways (Barabási, 2010). People's behaviors are already revealing far-reaching insights into their everyday patterns of spatial (González et al., 2008; Jones et al., 2008; Song et al., 2010) and communicative (Lewis et al., 2011; Suh et al., 2010; Walther & Bazarova, 2008; Watts et al., 2002) organization. Theorists are increasingly understanding that *cyberspace* is beginning to map into, onto, and through *realspace* (see, e.g., Adams, 2010; Breese, 2011), fulfilling Hägerstrand's (1965, 1967; Gale, 1978) envisioning of an interdisciplinary theory of time-space geography for the cartography of human behavior.

The exponentially increasing computing capabilities along with the use of the Internet and social media by the general public all over the world has opened up a brand new area of inquiry. The very nature of this kind of research effort requires going beyond not only any single discipline in social sciences but also the creation of a true multidisciplinary area that integrates many areas far afield from traditional social sciences. Such developments are likely to revolutionize our analytical understanding of human behavior and blaze paths for new academic disciplines.

Social science research based data mining is still at its infancy. As it develops its possibilities are truly enormous and yet to be realized. On the one hand, the spread of Internet technology is empowering common people from all over the world it is also being viewed with extreme suspicion by others. There is no question about the fact that the new technology is ushering in a new era of social science research (Chandler & Cortada, 2000; Watts, 2003); in order for us to fully harness its power from disaster preparedness and mitigation of threats of pandemics, as social scientists we must be cognizant of its abilities for good and evil.

References

Adams, Paul C. "A taxonomy for communication geography. *Progress in Human Geography, 35*(1) (2010): 37-57.

Alexander, R. J.. *Framing discourse on the environment: A critical discourse approach.* (New York, NY: 2009) Routledge.

Altaweel, M. R., Alessa, L. N., & Kliskey, A. D.. Visualizing situational data: Applying information fusion for detecting social-ecological events. *Social Science Computer Review*, *28*(4), (2010) 497-514. doi:10.1177/0894439309360837

An, L., Tsou, M-T., Wandersee, S., Gupta, D. K., Spitzberg, B. H., & Gawron, M.. *Who Is concerned about climate change? Evidence from space-time analysis.* Paper submitted to the American Association of Geographers Conference, (2012, February) New York, NY.

Andrés, Luis, Cuberes, David, Diouf, Mame, & Serebrisky, Tomás.. The diffusion of the internet: A cross-country analysis. *Telecommunications Policy, 34*, (2010) 323-340.

Arpinar, I. Budak, Sheth, Amit, & Ramakrishnan, Cartic, Usery, Lynn, Azami, Molly, & Kwan, Mei-Po. Geospatial ontology development and semantic analytics. *Transactions in GIS, 10*(4), (2006) 551-575.

Attrill, A., & Jalil, R.. Revealing only the superficial me: Exploring categorical self-disclosure online. *Computers In Human Behavior*, (2011) *27*(5), 1634-1642. doi:10.1016/j.chb.2011.02.001

Bai, X.. Predicting consumer sentiments from online text. *Decision Support Systems, 50*, (2010) 732-742.

Barabási, Albert-László. *Linked: The new science of networks*. (Cambridge, MA: 2002) Perseus.

Barabási, Albert-László. *Bursts: The hidden pattern behind everything we do.* (New York, NY: 2010) Dutton.

Barrett, C., K. Bisset, J. Leidig, A. Marathe, A., & M. Marathe. An Integrated Modeling Environment to Study the Coevolution of Networks, Individual Behavior, and Epidemics. *AI Magazine*, *31*(1), (2010) 75-87.

Bell, Gordon, Tony Hey, and Alex Szalay. Beyond the data deluge. *Science*, *323*, (2009) 1297-1298.

Bollen, Johan, Huina Mao, & Xiaojun Zeng. Twitter mood predicts the stock market. *Journal of Computational Science, 2*, (2011) 1-8.

Bostrom, R. N. Theories, data, and communication research. *Communication Monographs, 70*, (2003) 275-294.

Brasch, W. M. Fool's Gold in the Nation's Data-Mining Programs. *Social Science Computer Review*, *23*(4), (2005) 401-428. doi:10.1177/0894439305278869

Breese, Elizabeth Butler. Mapping the variety of public spheres. *Communication Theory, 21* (2011) 130-149.

Brier, A., & B. Hopp. Computer assisted text analysis in the social sciences. *Quality & Quantity: International Journal Of Methodology*, *45*(1), (2011). 103-128. doi:10.1007/s11135-010-9350-8

Brown, Rupert. Social identity theory: Past achievements, current problems and future challenges. *European Journal of Social Psychology, 30*, (2000) 745-778.

Cannataro, M., & D. Talia. The knowledge grid. *Communications of the ACM, 46*(1), (2003). 89-93.

Cao, L. In-depth behavior understanding and use: The behavior informatics approach. *Information Sciences*, *180*(17), (2010) 3067-3085. doi:10.1016/j.ins.2010.03.025

Carrasco, Juan Antonio, Bernie Hogan, Barry Wellman, & Eric J. Miller. Agency in social actdivity interactions: The role of social networks in time and space. *Tijdschrift voor Economische en Sociale Geografie, 99* (5), (2008). 562-583.

Carty, V. New information communication technologies and grassroots mobilization. *Information, Communication and Society,* 13(2) (2010) 155-173.

Chandler, Alfred D., Jr., & J. W. Cortada, (eds.) *A nation transformed by information: How information has shaped the United States from colonial times to the present*. (New York: 2000) Oxford.

Chau, M., & J. Xu. Mining communities and their relationships in blogs: A study of online hate groups. *International Journal of Human-Computer Studies,* 65(1) (2007) 57-70. doi:10.1016/j.ijhcs.2006.08.009

Chow, S.L. Acceptance of a theory: Justification or rhetoric? *Journal for the Theory of Social Behaviour, 22*, (1992) 447-474.

Copmpeau, Deborah R., Darren B. Meisher, , & Christopher A. Higgins. From prediction to explination: Reconceptualizing and extending the perceived characteristics of innovating. *Journal of the Association for Information Systems, 8*(8) (2007) 409-439.

Corman, Stephen R., T. Kuhn, R. D. McPhee, & K. J. Dooley. Studying complex discursive systems: Centering resonance analysis of communication. *Human Communication Research,* 28(2), (2002) 157-206.

Crandall, David J., Lars Backstrom, Dan Cosley, Siddharth, Suri, Daniel Huttenlocher, & Jon Kleinberg. Inferring social ties from geographic coincidences. *Proceedings of the National Academy of Sciences, 107*(52): (2010) 22436-22441.

de Hoog, N., Stroebe, W., & de Wit, J. B. F. (2007). The impact of vulnerability to and severity of a health risk on processing and acceptance of fear-arousing communications: A meta-analysis. *Review of General Psychology, 11*, 258-285.

Diani, M. Social movement networks: Virtual and real. *Information, Communication and Society,* 3(3) (2000) 386-401.

Dodds, Peter Sheridan, Roby Muhamad, & Duncan J. Watts. An experimental study of search in global social networks. *Science,* (2003) 301(5634), 827-829.

Douglas, K. M., & C. McGarty. Identifiability and self-presentation: Computer-mediated communication and intergroup interaction. *British Journal of Social Psychology, 40*, (2001) 399-416.

Earl, J. The dynamics of protest-related diffusion on the web. *Information, Communication and Society,* 13(2) (2010) 209-225.

Elkink, Johan A. The international diffusion of democracy. *Comparative Political Studies, 44*(12) (2011) 1651-1674.

Elson, Sara Beth, Douglas Yeung, Roshan Parisa, S. R. Bohandy, & Alireza Nader. *Using social media to gauge Iranian public opinion and mood after the 2009 election.* (Santa Monica, CA: Rand/National Security Research Division 2012).

Erickson, Ingrid. Geography and community: New forms of interaction among people and places. *American Behavioral Scientist, 53*(8), (2010) 1194-1207.

Etling, B., J. Kelly, R. Faris, & J. Palfrey. Mapping the Arabic blogosphere: Politics and dissent online. *New Media and Society,* 12(8) (2010) 1225-1243.

Flache, A., & M. W. Macy. Local convergence and global diversity: From interpersonal to social influence. *Journal of Conflict Resolution, 55* (2011) 970-995.

Fortunati, Leopold. Old and new media, old emotion. In L. Fortunati & J. Vincent (Eds.), *Electronic emotion: The mediationof emotion via information and communication technologies* (pp. 35-62). (Bern, Switzerland: 2009) Peter Lang.

Gale, Stephen. Some formal properties of Hägerstrand's model of spatial interactions. *Journal of Regional Science, 12*(2) (1972) 199-217.

Gandy, Oscar H., Jr. *The panoptic sort: A political economy of personal information*. (Boulder, CO: 1993) Westview.

Gawron, J. M., Gupta, D. K., Stephens, K., Tsou, M-H., Spitzberg, B. H., & Li, A.. *Using group membership markers for group identification in web texts*. Paper presented at the Sixth International AAAI Conference on Weblogs and Social Media Conference, Dublin, Ireland (2012, June).

Gilman, D. What's a theory to do...with seeing? or some empirical considerations for observation and theory. *British Journal for the Philosophy of Science, 43* (1992). 287-309.

Gleick, James. *The information*. (New York: 2011). Pantheon.

González, Marta C., Hidalgo, César A., & Barabási, Albert-Lászlo. Understanding individual human mobility patterns. *Nature, 453* (2008) 779-782.

González-Bailón, Sandra, Rafael E. Banchs, & Andreas Kaltenbrunner. Emotions, public opinion, and U.S. presidential approval rates: A5-year analysis of online political discussions. *Human Communication Research, 38* (2012) 121-143.

Gupta, Dipak K. *Decisions by the Numbers: An Introduction to Quantitative Techniques for Public Policy Analysis and Management*. (Englewood Cliffs, N. J.: 1994) Prentice Hall.

Gupta, Dipak K. *Path to Collective Madness: A Study In Social Order and Political Pathology*. (Westport, CT: 2001) Praeger.

Gupta, Dipak K. *Understanding Terrorism and Political Violence: The Life Cycle of Birth, Growth, Transformation, and Demise*. (London: 2008) Routledge.

Gupta, Dipak K. *Analyzing Public Policy: Concepts, Tools, and Techniques*. 2nd ed. (Washington D. C.: 2010) CQ Press.

Hägerstrand, Torsten. Aspects of the spatial structure of social communication and the diffusion of information. *Papers in Regional Science, 16*(1) (1966) 27-42.

Hägerstrand, Torsten. *Innovation diffusion as a spatial process*. (Chicago: 1967). University of Chicago Press.

Hägerstrand, Torsten. What about people in regional science? *Papers in Regional Science, 24*(1) (1970) 7-24.

Hand, D. J. Data Mining: New Challenges for Statisticians. *Social Science Computer Review*, *18*(4) (2000) 442.

Ho, Y., Chung, Y., & K. Lau. Unfolding large-scale marketing data. *International Journal Of Research In Marketing*. (2010) doi:10.1016/j.ijresmar.2009.12.009

Hogg, Michael A., Deborah J. Terry, & Katherine M. White. A tale of two theories: A critical comparison of identity theory with social identity theory. *Social Psychology Quarterly, 58*, (1995) 255-269.

Howard, Philip N., & Muzammil M. Hussain. The upheavals in Egypt and Tunisia: The role of digital media. *Journal of Democracy, 22*(3) (2011) 35-48.

Huang, Chun-Yao, & Hau-Ning Chen. Global digital divide: A dynamic analysis based on the Bass model. *Journal of Public Policy & Marketing, 29*(2) (2010) 248-264.

Huberty, Mark "Voting With Your Tweet: Predictive Modeling of Election Outcomes Using Social Media Data." (2012) International Studies Association Annual Meeting. San Diego.

Hunt, S. A., & R. D. Benford. Collective identity, solidarity, and commitment. In D. A. Snow, S. A. Soule, & H. Driesi (Eds.), *The Blackwell companion to social movements* (pp. 433-457). (Malden, MA: 2004). Blackwell.

Hwang, Hyunseo, M. et al. Media dissociation, Internet use, and antiwar political participation: A case study of political dissent and action against the war in Iraq. *Mass Communication & Society, 9*(4) (2006) 461-483.

Johnson, B. D., E. Dunlap, & E. Benoit. Organizing "mountains of words" for data analysis, both qualitative and quantitative. *Substance Use & Misuse*, *45*(5), (2010) 648-670. doi:10.3109/10826081003594757

Jones, Quenton, et al. Geographic 'place' and 'community information' preferences. *Computer Supported Cooperative Work, 17*(2-3), . (2007) 137-167.

Korte, Russell F. A review of social identity theory with implications for training and development. *Journal of European Industrial Training, 31* (2006) 166-130.

Kumar, R., J. Novak, and A. Tomkins in P.S. Yu, et al. (eds.), *Link Mining: Models, Algorithms, and Applications*, (Springer Science+Business Media, LLC. 2010)

Ladhani, N. Occupy social media. *Social Policy,* 83. (2011, Winter). Available: http://www.socialpolicy.org/index.php/component/content/article/4-latest-issue/503-occupy-social-media

Lazer, David, et al. Computational social science. *Science, 323* (2009) 721-723.

Lee, R., S. Wakamiya, & K. Sumiya. Discovery of unusual regional social activities using geo-tagged microblogs. *World Wide Web, 14*, (2011). 321-349.

Leetaru, Kalev Hannes. *Data mining methods for the content analyst: An introduction to the computational analysis of content*. (New York, NY: 2012) Routledge.

Lewis, Kevin, Marco Gonzalez, & Jason Kaufman. Social selection and peer influence in an online social network. *PNAS Early Edition*. (2011). Available: www.pnas.org/cgi/doi/10.1073/pnas.1109739109.

Lewis, Kevin, Jason Kaufman, & Nicholas Christakis. The taste for privacy: An analysis of college student privacy settings in an online network. *Journal of Computer-Mediated Communication, 14*, (2008). 79-100.

Lewis, Ted. *Network science: Theory and applications*. (New York: 2009). Wiley.

Li, Nan & Desheng Dash Wu. Using text mining and sentiment analysis for online forums hotspot detection and forecast. *Decision Support Systems,* 48(2) (2010) 354-368.

Little, Miles, Jordens, F. C. Christopher, & E. Sayers. Discourse communities and the discourse of experience. *Health: An interdisciplinary Journal for the Social Study of Health, Illness and Medicine, 7*(1) (2003). 73-86.

Massaro, Vanessa A., & Emma Gaalaas Mullaney. The war on teenage terrorists. *City, 15* (5) (2011) 591-604.

Marquette, J.F. A logistic diffusion model of political mobilization. *Political Behavior, 3* (1981) 7-30.

McFarland, Daniel, & Heili Pals. Motives and contexts of identity change: A case for network effects. *Social Psychology Quarterly, 68*, (2005) 289-315.

McKenna, K. Y. A., & J. A. Bargh. Plan 9 from cyberspace: The implications of the internet for personality and social psychology. *Personality and Social Psychology Review, 4*(1) (2000) 47-75.

Meade, N., & T. Islam. Modeling and forecasting the diffusion of innovation – A 25-year review. *International Journal of Forecasting, 22* (2006) 519-525. doi:10.1016/j.ijforecast.2006.01.005

Mullen, B., M. J. Migdal, & M. Hewstone. Crossed categorization versus simple categorization and intergroup evaluations: A meta-analysis. *European Journal of Social Psychology, 31* (2001). 721-736.

Murphy, Emma C. Agency and space: The political impact of information technologies in the Gulf Arab states. *Third World Quarterly, 27* (2006) 1059-1083. DOI: 10.1080/01436590600850376

Nebot, V., & R. Berlanga. Finding association rules in semantic web data. *Knowledge-Based Systems*, *25*(1) (2012) 51-62. doi:10.1016/j.knosys.2011.05.009

Neuman, W. R., Park, Y. J., & E. Panek. Tracking the flow of information into the home: An empirical assessment of the digital revolution in the United States, 1960-2005. *International Journal of Communication, 6* (2012). 1022-1041.

Novemsky, Nathan, & Daniel Kahneman. The boundaries of loss aversion. *Journal of Marketing Research, 42*(2) (2005). 119-128.

Oliver, Pamela E., & Daniel J. Myers. Networks, diffusion, and cycles of collective action. In M. Diani & D. McAdam (Eds.), *Social movements and networks: Relational approaches to collective action* (pp. 173-205). (New York, NY: 2003) Oxford.

Papacharissi, Z. The virtual geographies of social networks: A comparative analysis of Facebook, LinedIn and ASmallWorld. *New Media and Society, 11* (2009) 199-220.

Pavitt, C. Theory-data interaction from the standpoint of scientific realism: A reaction to Bostrom. *Communication Monographs, 71* (2004) 333-342.

Petronio, S. *Boundaries of privacy: Dialectics of disclosure.* (Albany, NY: 2002) State University of New York Press.

Popper, K. Science: Conjectures and refutations. In E. D. Klemke, R. Hollinger, & A. D. Kline (Eds.), *Introductory readings in the philosophy of science* (Buffalo, NY: 1980). Prometheus: 19-34.

Postmes, Tom, & S. Brunsting. Collective action in the age of the internet: Mass communication and online mobilization. *Social Science Computer Review, 20* (3) (2002) 290-301.

Ramakrishnan, N., & A. Y. Grama. Data mining: From serendipity to science. *Computer,* (1999, August) 34-37.

Reid, Scott A., & Howard Giles. Intergroup relations: Its linguistic and communicative parameters. *Group Processes & Intergroup Relations, 8* (3) (2005) 211-214. DOI: 10.1177/1368430205053938

Rogers, Everett. M. *Diffusion of innovations* (5th ed.). (New York: 2003) Free Press.

Rogers, Everett. M., & D. L. Kincaid. *Communication networks: Toward a new paradigm for research*. (New York: (1981) Free Press.

Rohlinger, D. A., & J. Brown. Democracy, action, and the Internet after 9/11. *American Behavioral Scientist, 53(1)* (2009) 133-150.

Saldarini, Robert A., & E. M. DeRobertis. The impact of technology induced anonymity on communications and ethics: New challenges for IT pedagogy. *Journal of Information Technology Impact, 3*(1) (2003) 3-10.

Shaw, Shih-Lung, Hongbo Yu & Leonard S. Bombom. A space-time GIS approach to exploring large individual-based spatiotemporal datasets. *Transactions in GIS, 12*(4) (2008) 425-441.

Song, C, Z. Qu, N. Blumm, N and A. L. Barabási. Limits of predictability in human mobility. *Science, 327*, (2010) 1018-1021. DOI: 10.1126/science.1177170

Soule, Sarah A. Diffusion processes within and across movements. In D. A. Snow, S. A. Soule, & H. Kriesi (Eds.), *The Blackwell companion to social movements* (pp. 294-310). (Malden, MA: 2004). Blackwell.

Strodthoff, G. G., R. P. Hawkins, & A. C. Schoenfeld. Media roles in a social movement: A model of ideology diffusion. *Journal of Communication, 35* (1985) 134-153.

Suh, Bongwon, L. Hong, P. Pirolli, & E. H. Chi. Want to be retweeted? Large scale analytics on factors impacting retweet in Twitter network. Second International IEEE International Conference on Social Computing, (2010).177-184.

Spitzberg, B. H., M-H Tsou, L. An, D. K. Gupta, & J. M. Gawron. *The map is not which territory?: Speculating on the geo-spatial diffusion of ideas in the Arab Spring of 2011*. Paper presented at the International Communication Association Conference. (2012, May*)* Phoenix, AZ.

Stein, L. Social movement web use in theory and practice: A content analysis of US movement websites. *New Media and Society, 11*(5): (2009) 749-771. DOI: 10.1177/1461444809105350

Takhteyev, Yuri, A. Gruzd, & B. Wellman. Geography of Twitter networks. (2011). DOI: 10.1016/j.socnet.2011.05.006

Taleb, N. N., & M. Blyth. The black swan of Cairo: How suppressing volatility makes the world less predictable and more dangerous. *Foreign Affairs, 90*(3) (2011) 33-39.

Tillema, Taede, M. Dijst, & T. Schwanen. Face-to-face and electronic communications in maintaining social networks: The influence of geographical and relational distance and of information content. *New Media & Society, 12*(6), (2010) 965-983.

Tsou, M-H., An, L., S. Wandersee, I-H Kim, B. H. Spitzberg, D. K. Gupta, J. M. Gawron, J. Smith, T-H Lee. Mapping ideas from cyberspace to realspace: Visualizing hidden geospatial fingerprints on web information landscapes. *Annals of the Association of American Geographers* (2011)**.**

Turner, John C., & Katherin J. Reynolds. The social identity perspective in intergroup relations: Theories, themes, and controversies. In R. Brown & S. L. Gaertner (Eds.), *Blackwell handbook of social psychology: Intergroup processes* (pp.133-152). (Oxford, England: 2001). Blackwell.

Van Laer, J. Activists online and offline: The internet as an information channel for protest demonstrations. *Mobilization: An International Journal, 15*, (2010). 347-366.

Van Laer, J., & P. Van Aelst. Internet and social movement action repertoires: Opportunities and limitations. *Information, Communication and Society, 13*, (2010) 1146-1171. DOI: 10.1080/1369118100368307

Vishwanath, Arun, & Hao Chen. Towards a comprehensive understanding of the innovation-decision process. *The diffusion of innovations: A communication science perspective* (pp. 9-32). (New York, NY: 2011). Peter Lang.

Walther, Joseph B. & Natalya N. Bazarova. Validation and application of electronic propinquity theory to computer-mediated communication in groups. *Communication Research, 35*, (2008) 622-645.

Watts, Duncan J. *Six degrees: The science of a connected age.* (New York, NY: 2003) W. W. Norton.

Watts, Duncan J. The "new" science of networks. *Annual Review of Sociology, 30*, 243-270. (2004) DOI: 10.1146/annurev.soc.30.020404.104342

Watts, Duncan J., & Peter S. Dodds. Influentials, networks, and public opinion formation. *Journal of Consumer Research, 34*, (2007) 441-458.

Watts, Duncan J., Peter S.Dodds, & M. E. J. Newman. Identity and search in social networks. *Science, 296* (2002) 1302-1305.

Witte, K., & M. Allen. A meta-analysis of fear appeals: Implications for effective public health campaigns. *Health Education & Behavior, 27*, (2000). 591-615.

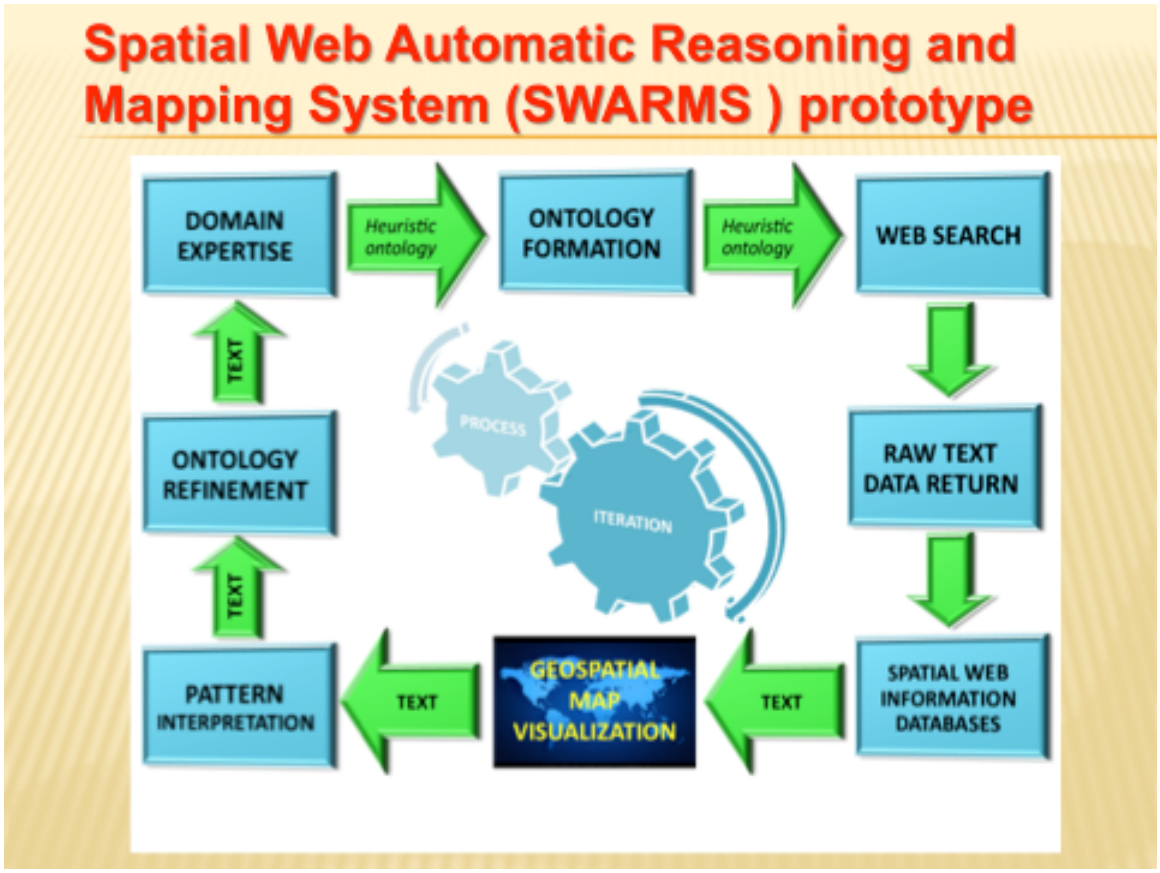
Yang, Qiang, & Xindong Wu. 10 challenging problems in data mining research. *International Journal of Information Technology & Decision Making, 5* (4), (2006) 597–604.

Young, H. Peyton. Innovation diffusion in heterogeneous populations: Contagion, social influence, and social learning. *American Economic Review, 99*(5), (2009). 1899-1924.

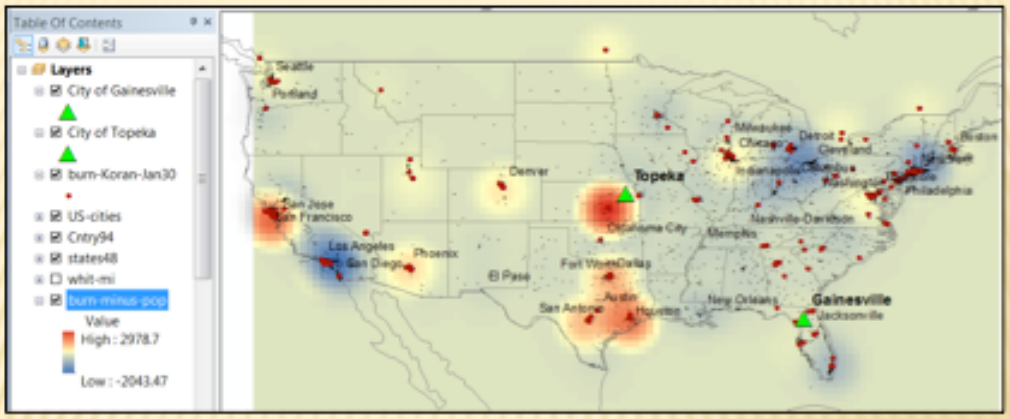
Zimbra, D, A. Abbasi, A., and H. Chen. A cyber-archaeology approach to social movement research: Framework and case study. *Journal of Computer-Mediated Communication, 16*, (2010) 48-70. DOI: 10.1111/j.1083-6101.2010.01531.x

Figure 1.

A basic heuristic flow diagram depicting the use of search and interpretation processes to geo-locate the diffusion of symbolic content on the web



"Burn Koran" – minus - [US. Population Density]



The U.S population density map was used to standardize the popularity density map of "burn Koran". After the standardization, the **red color hot spots** indicate that San Jose, Houston, and the middle of Kansas State are the popular areas of "burn Koran" keywords. **The blue color hot spots** indicate the **negative value (less popular)** of "burn Koran" standardized by city population density.

WHY the hotspot in the middle of Kansas ? Near the City of Topeka? (after the original event happen in the church located in Gainesville, Florida (green symbol), another church in the city of Topeka, Kansas claimed that they will continue the action of "burn Koran".)