BEA
Educating for Tomorrow's Media

Routledge
Taylor & Francis Group

The Value of Big Data in Digital Media Research

Merja Mahrt and Michael Scharkow

This article discusses methodological aspects of Big Data analyses with regard to their applicability and usefulness in digital media research. Based on a review of a diverse selection of literature on online methodology, consequences of using Big Data at different stages of the research process are examined. We argue that researchers need to consider whether the analysis of huge quantities of data is theoretically justified, given that it may be limited in validity and scope, and that small-scale analyses of communication content or user behavior can provide equally meaningful inferences when using proper sampling, measurement, and analytical procedures.

Communication research is interdisciplinary and digital media research may be even more so. Every discipline brings its own theories and methods to phenomena related to recently developed forms of computer-mediated communication. This has, for instance, led researchers in the humanities to discover quantitative methods for large data sets (Lazer et al., 2009; Manovich, 2012), while information scientists are exploring the merit of qualitative analyses (Parker, Saundage, & Lee, 2011).

Some of these approaches were developed for the analysis of data that are created or becomes available through people's use of digital media. The resulting data structures are often different from those typically studied in a given field, thus fueling methodological innovation. Among practitioners and applied researchers, the reaction to data available through blogs, Twitter, Facebook, or other social media can be described as a ''data rush,'' promising new insights about consumers' choices and behavior and many other issues (e.g., Kearon & Harrison, 2011; Russom, 2011). For instance, some proponents of data mining see it as a way to study many different phenomena, without requiring ''significant background knowledge of data analysis'' (Russell, 2011, p. xvi). In contrast to such enthusiasts, others express less excitement, asking whether large-scale data analyses actually allow practitioners or researchers to really understand what the findings *mean* and suggesting an approach that combines data-driven and more ethnographic methods (Hooper, 2011).

Merja Mahrt *(Ph.D., University of Amsterdam) is a research associate at Heinrich Heine University, Germany. Her research interests include social functions and effects of mass media, especially with regard to differences between online and offline media.*

Michael Scharkow *(Ph.D., University of the Arts Berlin) is a research associate at the University of Hohenheim, Germany. His research interests include empirical research methods, online communication, and media use.*

DOI: 10.1080/08838151.2012.761700
Journal of Broadcasting & Electronic Media 57(1), 2013, pp. 20–33
ISSN: 0883-8151 print/1550-6878 online

In addition to methodological questions about how to approach new data and new phenomena of digital media, the role of theory for Internet-related research has been called into question (Anderson, 2008; Bailenson, 2012; Bollier, 2010). Some argue that researchers should let the data speak for itself, that ''the data *is* the question!'' (IBM engineer Jeff Jonas, cited by Bollier, 2010, p. 9). Given these different disciplinary and methodological approaches, this paper gives an overview of ongoing debates about digital media research and argues for the lasting merit of established principles of empirical research, regardless of how novel or up-and-coming a medium or research question may seem.

The ''Data Rush''

Since the 1990s and early 2000s, social scientists from various fields have relied more and more on digitized methods in empirical research. Surveys can be administered via Web sites instead of paper or via telephone; digital recordings of interviews or experimental settings make the analysis of content or observed behavior more convenient; and coding of material is supported by more and more sophisticated software. But in addition to using research tools that gather or handle data in digital form, scholars have also started using digital material that was not specifically created for research purposes. The introduction of digital technology in, for example, telephone systems and cash registers, as well as the diffusion of the Internet to large parts of a given population, have created huge quantities of digital data of unknown size and structure. While phone and retail companies usually do not share their clients' data with the academic community (Savage & Burrows, 2007), scholars have concentrated on the massive amounts of publicly available data about Internet users, often giving insight into previously inaccessible subject matters. Subsequently, methodological literature began discussing research practices, opportunities, and drawbacks of online research (Batinic, Reips, & Bosnjak, 2002; Johns, Chen, & Hall, 2004; Jones, 1999a). This chapter gives a brief overview of methodological issues related to Internet research across different disciplines and communities in the social sciences. Two aspects deserve a deeper discussion: the current debate around the concept of ''Big Data'' and the question of finding ''meaning'' in digital media data.

Methodological Issues in the Study of Digital Media Data

Christians and Chen (2004) discuss technological advantages of Internet research, but urge their readers to also consider its inherent disadvantages. Having huge amounts of data available that is ''naturally'' created by Internet users also has a significant limitation: The material is not indexed in any meaningful way, so no comprehensive overview is possible. Thus, there may be great material for many different research interests, but the question of how to access and select it cannot be easily answered. Sampling is therefore probably the issue most often and consistently

raised in the literature on Internet methodology (Erlhofer, 2010; Mitra & Cohen, 1999; Vogt, Gardner, & Haeffele, 2012; Welker et al., 2010).

On the one hand, sampling online content poses technical or practical challenges. Data may be created through digital media use, but it is currently impossible to collect a sample in a way that adheres to the conventions of sample quality established in the social sciences. This is partly due to the vastness of the Internet, but the issue is further complicated by the fact that online content often changes over time (Jones, 1999b; Mitra & Cohen, 1999). On Web sites and, to an even lesser degree, social media sites, content is not as stable or clearly delineated as in most traditional media, which can make sampling and defining units of analysis challenging (Herring, 2010). It seems most common to combine purposive and random sampling techniques, which is what Mazur (2010) recommends.

The problems related to sampling illustrate that tried and tested methods and standards of social science research will not always be applicable to digital media research. But scholars take opposing sides on whether to stick with traditional methods or adopt new ones: Some suggest applying well-established methods to ensure the quality of Internet research (Jankowski & van Selm, 2005; Lally, 2009; McMillan, 2000). On the other hand, Jones (1999b) questions whether conventional methods would be applicable to large data sets of digital media. In Herring's (2010) view, communication scholars trained in conventional content analysis will find they need to adapt their methodological toolbox to digital media, at least to some degree. She pleads for the incorporation of methods from other disciplines to be able to adequately study the structure of Web sites, blogs, or social network sites (see also Christians & Chen, 2004). Scholars may find more appropriate or complementary methods in, for instance, linguistics or discourse analysis.

But methodological adaptation and innovation have their drawbacks, and scholars of new phenomena or data structures find themselves in an area of conflict between old and new methods and issues. While scholars like Herring (2010) or Jones (1999b) argue for a certain level of restraint toward experimenting with new methods and tools, researchers caught in the ''data rush'' seem to have thrown caution to the wind, allowing themselves to be seduced by the appeal of Big Data.

Big Data

The term *Big Data* has a relative meaning and tends to denote bigger and bigger data sets over time. In computer science, it refers to data sets that are too big to be handled by regular storage and processing infrastructures. It is evident that large data sets have to be handled differently than small ones; they require different means of discovering patterns—or sometimes allow analyses that would be impossible on a small scale (Bollier, 2010; Manovich, 2012; Russom, 2011; Savage & Burrows, 2007). In the social sciences and humanities as well as applied fields in business, the size of data sets thus tends to challenge researchers as well as software or hardware. This may be especially an issue for disciplines or applied fields that are more or less unfamiliar with quantitative analysis. Manovich (2012) sees knowledge

of computer science and quantitative data analysis as a determinant for what a group of researchers will be able to study. He fears a ''data analysis divide'' (p. 461) between those equipped with the necessary analytical training and tools to actually make use of the new data sets and those who will inevitably only be able to scratch the surface of this data.

New analytical tools tend to shape and direct scholars' ways of thinking and approaching their data. The focus on data analysis in the study of Big Data has even led some to the assumption that advanced analytical techniques make theories obsolete in the research process (Anderson, 2008; Bailenson, 2012). Research interest could thus be almost completely steered by the data itself.

But being driven by what is possible with the data may cause a researcher to disregard vital aspects of a given research object. boyd and Crawford (2012) underline the importance of the (social) context of data that has to be taken into account in its analysis. The scholars illustrate how analyses of large numbers of messages (''tweets'') from the microblogging service Twitter are currently used to describe aggregated moods or trending topics—without researchers really discussing what and particularly who these tweets represent: Only parts of a given population are even using Twitter, often in very different ways. As boyd and Crawford point out, these contexts are typically unknown to researchers who work with samples of messages captured through Twitter.

In addition, Big Data analyses tend only to show *what* users do, but not *why* they do it. In his discussion of tools for Big Data analysis, Manovich (2012) questions the significance of the subsequent results in terms of their relevance for the individual or society. The issue of *meaning* of the observed data and/or analyses is thus of vital importance to the debate around Big Data.

Meaning

Jones (1999b) already wrote about the belief or hope of scholars involved in Internet research that data collected through Web sites, online games, e-mail, chat, or other modes of Internet usage ''represent ... well, *something*, some semblance of reality, perhaps, or some 'slice of life' on-line'' (p. 12). Yet, Park and Thelwall (2005) illustrate that even the comparatively simple phenomenon of a hyperlink between two Web sites is not easily interpreted. What does the existence of just the connection, as such, between two sites tell a researcher about why it was implemented and what it means for the creators of the two sites or their users?

Studying online behavior through large data sets strongly emphasizes the technological aspect of the behavior (Christians & Chen, 2004) and relies on categories or features of the platforms that generated the data. Yet, the behaviors or relationships thus expressed online may only *seem* similar to their offline counterparts. boyd and Crawford (2012) illustrate that, for instance, network relationships between cell phone users may give an account of who calls whom how often and for how long. Yet, whether frequent conversation partners also find their relationship important for them personally or what relevance they attribute to it cannot be derived from

their connection data without further context. In addition, Mazur (2010) advises researchers to be wary of data collected from social media to a certain degree, because they may be the result of active design efforts by users who purposefully shape their online identities.

It is unclear how close to or removed from their online personas users actually are (Utz, 2010). This means that scholars should be careful to take data for what it truly represents (e.g., traces of behavior), but not infer too much about possible attitudes, emotions, or motivations of those whose behavior created the data—although some seem happy to make such inferences (Kearon & Harrison, 2011). Orgad (2009) and Murthy (2008) urge scholars to scrutinize whether a study can rely on data collected online alone or whether it should be complemented by offline contextual data.

Opportunities of Big Data Research

As has been pointed out above, data collected through use of online media is obviously attractive to many different research branches, both academic and commercial. We will briefly summarize key advantages in this section and subsequently discuss critical aspects of Big Data in more depth.

Focusing on the social sciences, advantages and opportunities include the fact that digital media data are often a by-product of the everyday behavior of users, ensuring a certain degree of ecological validity (Mehl & Gill, 2010). Such behavior can be studied through the traces it automatically left, providing a means to study human behavior without having to observe or record human subjects first. This can also allow examination of aspects of human interaction that could be distorted by more obtrusive methods or more artificial settings, due to observer effects or the subjects' awareness of participating in a study, for instance (Jankowski & van Selm, 2005; Vogt et al., 2012).

Such observational data shares similarities with material used in content analysis since it can be stored or already exists in document form. Thus, content analysis methodology well-established in communication or other research fields can be applied to new research questions (Herring, 2010; McMillan, 2000). When content posted on a platform is analyzed in combination with contextual data, such as time of a series of postings, geographic origin of posters, or relationships between different users of the same platform or profile, digital media data can be used to explore and discover patterns in human behavior, e.g., through visualization (Dodge, 2005). For some equally explorative research questions, the sheer amount of information accessible online seems to fascinate researchers because it provides (or at least seems to provide) ample opportunities for new research questions (Vogt et al., 2012; Welker et al., 2010).

Lastly, the collection of Big Data can also serve as a first step in a study, which can be followed by analyses of sub-samples on a much smaller scale. Groups hard to reach in the real world (Christians & Chen, 2004) or rare and scattered phenomena can be filtered out of huge data sets, thus providing access to the

proverbial needle in the digital haystack. This can be much more efficient than drawing, for instance, a huge sample of people via a traditional method, such as random dialing or random walking, when attempting to identify those who engage in comparatively rare activities.

Challenges of Big Data Research

Although Big Data seems to be promising a golden future, especially to commercial researchers (Kearon & Harrison, 2011; Russom, 2011), the term is viewed much more critically in the academic literature. boyd and Crawford (2012) as well as Manovich (2012) discuss issues related to the use of Big Data in digital media research, some of which have been summarized above. In addition to more general political aspects of ownership of platforms and ''new digital divides'' in terms of data access or questions about the meaning of Big Data, its analysis also poses concrete challenges for researchers in the social sciences. This section discusses aspects of Big Data research that scholars need to address at different stages of the research process. One recurring theme in many studies that make use of Big Data is what we call its availability bias: Rather than theoretically defining units of analysis and measurement strategies, researchers tend to use whatever data is available and then try to provide an ex-post justification or even theorization for its use. This research strategy is in stark contrast to traditional theory-driven research and raises concerns about the validity and generalizability of the results.

Sampling and Data Collection

The problem of sampling in Internet research has already been addressed above and is mentioned in almost every publication on online research (Batinic et al., 2002; Herring, 2010; McMillan, 2000). While there are some promising approaches for applying techniques such as capture-recapture (Engesser & Krämer, 2011) or adaptive cluster sampling to online research, the problem of proper random sampling, on which all statistical inference is based, remains largely unsolved. Most Big Data research is based on nonrandom sampling, such as using snowball techniques or simply by using any data that is technically and legally accessible.

Another problem with many Big Data projects is that even with a large sample or complete data from a specific site, there is often little or no variance in the level of platforms or sites. If researchers are interested in social network sites, multiplayer games, or online news in general, it is problematic to include only data from Facebook and Twitter, World of Warcraft and Everquest II, or a handful of newspaper and broadcast news sites. From a platform perspective, the sample size of these studies is tiny, even with millions of observations per site. This has consequences not only for the inferences that can be drawn from analyses, but also from a validity perspective: Expanding and testing the generalizability of the results

would not require more data from the same source, but information from many different sources. In this respect, the hardest challenge of digital media research might not be to obtain Big Data from a few, although certainly important, Web sites or user groups, but from many different platforms and persons. Given the effort required to sample, collect, and analyze data from even a single source, and the fact that this can rarely be automated or outsourced, this ''horizontal'' expansion of online research remains a difficult task.

A third important aspect of Big Data collection is the development of ethical standards and procedures for using public or semi-public data. Zimmer (2010) provides an excellent account of the problems researchers face when making seemingly public data available to the research community. The possibility of effective de-anonymization of large data sets (Narayanan & Shmatikov, 2008) has made it difficult for researchers to obtain and subsequently publish data from social networks such as YouTube, Facebook, or Twitter. Moreover, the risk of inadvertently revealing sensitive user information has also decreased the willingness of companies to provide third parties with anonymized data sets, even if these companies are generally interested in cooperation with the research community. Researchers who collect their data from publicly available sources are at risk as well because the content providers or individual users may object to the publication of this data for further research, especially after the data has successfully been de-anonymized. The post-hoc withdrawal of research data, in turn, makes replications of the findings impossible and therefore violates a core principle of empirical research.

Finally, basically all Big Data research is based on the assumption that users implicitly consent to the collection and analysis of their data by posting them online. In light of current research on privacy in online communication, it is questionable whether users can effectively distinguish private from public messages and behavior (Barnes, 2006). But even if they can, since it is technically possible to recover private information even from limited public profiles (Zheleva & Getoor, 2009), Big Data research has to solve the problem of guaranteeing privacy and ethical standards while also being replicable and open to scholarly debate (see also Markham & Buchanan, 2012).

Measurement

Concerns about the reliability and validity of measurement have been raised in various critical papers on Big Data research, most recently by boyd and Crawford (2012). Among the most frequently discussed issues are (1) comparatively shallow measures, (2) lack of context awareness, and (3) a dominance of automated methods of analysis. Clearly, these concerns and their causes are related to an implicit or explicit tendency toward data-driven rather than theory-driven operationalization strategies. In addition to the possible ''availability bias'' mentioned above, many prominent Big Data studies seem to either accept the information accessible via digital media as face-valid, e.g., by treating Facebook friendship relations as similar

to actual friendships, or reduce established concepts in communication such as *topic* or *discourse* to simple counts of hashtags or retweets (Romero, Meeder, & Kleinberg, 2011; Xifra & Grau, 2010). While we do not argue that deriving measurement concepts from data rather than theory is problematic, per se, researchers should be aware that the most easily available measure may not be the most valid one, and they should discuss to what degree its validity converges with that of established instruments. For example, both communication research and linguistics have a long tradition of content-analytic techniques that are, at least in principle, easily applicable to digital media content. Of course, it is not possible to manually annotate millions of comments, tweets, or blog posts. However, any scholar who analyzes digital media can and should provide evidence for the validity of measures used, especially if they rely on previously unavailable or untested methods.

The use of shallow, ''available'' measures often coincides with an implicit preference for automatic coding instruments over human judgment. There are several explanations for this phenomenon: First, many Big Data analyses are conducted by scholars who have a computer science or engineering background and may simply be unfamiliar with standard social science methods such as content analysis (but some are discussing the benefits of more qualitative manual analyses; Parker et al., 2011). Moreover, these researchers often have easier access to advanced computing machinery than trained research assistants who are traditionally employed as coders or raters. Second, Big Data proponents often point out that automatic approaches are highly reliable, at least in the technical sense of not making random mistakes, and better suited for larger sample sizes (King & Lowe, 2003; Schrodt, 2010). However, this argument is valid only if there is an inherent advantage to coding thousands of messages rather than a smaller sample, and if this advantage outweighs the decrease of validity in automatic coding that has been established in many domains of content analysis research (Krippendorff, 2004). For example, Thelwall, Buckley, Paltoglou, Cai, and Kappas (2010) report an average correlation of about $r = .5$ between automatic sentiment analysis and human raters, and Scharkow (2013) finds that supervised text classification is on average 20 percent less reliable than manual topic coding. Despite the vast amount of scholarship on these methods, the actual tradeoff between measurement quality and sample quantity is hardly ever discussed in the literature, although it is central to the question of whether and when, for example, we accept shallow lexical measures that are easy to implement and technically reliable as substitutes for established content-analytic categories and human coding.

Data Analysis and Inferences

In addition to sampling, data collection, and measurement, the analysis of large data sets is one of the central issues around the Big Data phenomenon. If a researcher deals with Big Data in the original technical sense, meaning that data sets cannot be analyzed on a desktop computer using conventional tools such as SPSS or SAS, he or she can investigate the possibilities of distributed algorithms and software

that can run analyses on multiple processors or computing nodes. An alternative approach would be to take a step back and ask whether an analysis of a subset of the data could provide enough information to test a hypothesis or make a prediction. Although in general, a larger sample size means more precise estimates and a larger number of indicators or repeated observations leads to less measurement error (Nunnally, 1978), most social science theories do not require that much precision (Gerring, 2001). If the sampling procedure is valid, the laws of probability and the central limit theorem also apply to online research, and even analyses that require much statistical power can still be run on a single machine. In this way, Big Data can safely be reduced to medium-size data and still yield valid and reliable results.

The requirement of larger or smaller data sets is also linked to the question of what inferences one might like to draw from the analysis: Are we interested in aggregate or individual effects, causal explanation or prediction? Predicting individual user behavior, for example on a Web site, requires both reliable and valid measurement of past behavior as well as many observations. Longitudinal analyses of aggregate data, e.g., using search queries (Scharkow & Vogelgesang, 2011) or large collections of tweets (Chew & Eysenbach, 2010), do not necessarily require perfectly reliable coding or large sample sizes: If a blunt coding scheme based on a simple word list has only 50 percent accuracy, it is still possible to analyze correlations between time series of media content and user behavior—as long as the amount of measurement error is the same over time (see Granger, 1986). Moreover, whether a time series is based on hundreds or thousands of observations rarely affects the inferences that can be drawn on the aggregate level, at least if the observations are representative of the same population. If, on the other hand, a researcher is interested in analyzing the specific content of a set of messages or the behavior of a pre-defined group of online users, an instrument that has a reliability of .5 might not be enough. As in other disciplines such as psychology, education, or medicine, individual diagnostics and inferences require far more precision than the detection of aggregate trends.

Finally, one should ask how generalizable the findings of a study can or should be: In-depth analysis, both qualitative and quantitative, might allow for accurate predictions and understanding of a single individual, but it often cannot be generalized for larger samples or the general population. Observing a handful of Internet users in a computer lab can rarely lead to valid inferences about Internet users in general, simply because there is often too little information about individual differences or, more technically, between-person variance. Correlations on the aggregate level, on the other hand, cannot simply be applied to the individual level without the risk of an ecological fallacy, i.e., observing something in aggregate data that never actually occurs on the individual level (Yanovitzky & Greene, 2009).

Interpretation and Theoretical Implications

If researchers have undertaken analyses of Big Data, they need, of course, to interpret their results in light of the decisions they have made along the research

process and the consequences of each of these decisions. The core question should be: What is the theoretical validity and significance of the data? Large samples of digital media are limited in some respects, so scholars have to be careful about what inferences are drawn from them. The problem of determining the meaning of some types of digital media data has already been alluded to above. The number of times a message gets forwarded (''retweeted'') on Twitter, for instance, may show a certain degree of interest by users, but without looking at the content and/or style of a tweet, ''interest'' could stand for popularity and support, revulsion and outrage, or simply the thoughtless routines of Twitter usage behavior. And as boyd and Crawford (2012) point out: No matter how easily available Facebook, YouTube, or Twitter data is, it is based on a small and certainly nonrandom subset of Internet users, and this is even more true when investigating specific Web sites, discussion boards, online games, or devices. If less than 5 percent of Internet users in a given country are active on Twitter, as in Germany, for instance (Busemann & Gscheidle, 2012), an analysis of trending topics on the microblogging service can hardly represent the general population's current concerns.

In addition, a platform's interfaces (or ethical constraints) may not allow researchers to access information that would be most interesting to them, confining them to descriptive exploration of artificial categories. Visualizations based on such categories, for example connections between social media users, may allow the discovery of patterns (Dodge, 2005), but without cases to compare them to, these patterns may not lead to insight. Likewise, we have already underlined that the mere occurrence of certain keywords in a set of social media messages does not constitute ''discourse,'' per se. Such theoretical constructs should not be tweaked beyond recognition to fit the data structure of a given platform.

In sum, researchers should not compromise their original research interests simply because they cannot be as easily approached as others. If after careful scrutiny of the possibilities a certain platform or type of analysis really offers, the scholar decides that a Big Data approach is not advisable, a thorough analysis of smaller data sets may well produce more meaningful results. While similar problems exist in all empirical studies, such issues seem especially pressing in Big Data research.

Conclusion

The opportunities for large-scale digital media research are obvious—as are its pitfalls and downsides. Thus, researchers should differentiate between alternative research approaches carefully and be cautious about the application of unfamiliar tools, analytical techniques, or methodological innovation. With no or few references to compare one's results to, findings will be difficult to interpret and online researchers should ''hold *themselves* to high standards of conceptual clarity, systematicity of sampling and data analysis, and awareness of limitations in interpreting their results'' (Herring, 2010, p. 246).

Both boyd and Crawford (2012) as well as Manovich (2012) assert that methodological training should be part of the answer to the challenges of digital media research. Manovich argues for advanced statistics and computer science methods, which could likely help in furthering an understanding of the underlying algorithms of online platforms as well as analytical tools. Yet, a reflection on and an understanding of what comes before the first data is collected or analyzed is equally or possibly even more important.

Methodological training not only teaches how to handle data, but also allows students and researchers to learn to ask meaningful questions and be aware of how their choices at any given point in the research process will affect all subsequent phases. Theoretical considerations should be narrowly tied to concrete hypotheses or research questions which, in turn, determine operationalizational decisions as well as, to a considerable degree, the types of analyses that are possible and reasonable. Such an understanding of the interconnectedness of a researcher's decision is not easily acquired, but of vital importance. Mehl and Gill (2010), for instance, emphasize that scholars should use software that fits their research question and that the resulting data should be interpreted sensibly for what it is. At every stage of the research process, the value of using big- or small-scale data should be assessed.

In general, we should resist the temptation to let the opportunities and constraints of an application or platform determine the research question; the latter should be based on relevant and interesting issues—regardless of whether something is available through an API of a platform or seems easily manageable with a given analytical tool. Methodological training for upcoming generations of communication researchers should not only focus on computational issues and data management, but also continue to stress the importance of methodological rigor and careful research design. This includes a strong need for theoretical reflection, in clear contrast to the alleged ''end of theory'' (Anderson, 2008; Bailenson, 2012).

New data structures and research opportunities should, of course, not be ignored by media and communication scholars, and there are many relevant and interesting research questions that are well suited to Big Data analysis. On the other hand, established practices of empirical research should not be discarded as they ensure the coherence and quality of a study. After all, this is one of the key contributions that social scientists can bring to the table in interdisciplinary research. It should go without saying that a strong focus on theoretically relevant questions always increases the scientific significance of the research and its results. Yet, some developments in digital media research, particularly those related to Big Data, seem to warrant affirmation of this fundamental principle.

References

Anderson, C. (2008, June 23). The end of theory: The data deluge makes the scientific method obsolete. *Wired*. Retrieved from http://www.wired.com/science/discoveries/magazine/16-07/pb_theory/

Bailenson, J. N. (2012, May). Contribution to the ICA Phoenix closing plenary: ''The Internet is the end of communication theory as we know it.'' 62nd annual convention of the International Communication Association, Phoenix, AZ. Retrieved from http://www.icahdq.org/conf/2012/closing.asp

Barnes, S. (2006). A privacy paradox: Social networking in the United States. *First Monday*, *11*(9). Retrieved from http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/1394/1312

Batinic, B., Reips, U.-D., & Bosnjak, M. (Eds.). (2002). *Online social sciences*. Seattle, WA: Hogrefe & Huber.

Bollier, D. (2010). *The promise and peril of Big Data*. Washington, DC: Aspen Institue. Retrieved November 30, 2012, from http://www.aspeninstitute.org/sites/default/files/content/docs/pubs/The_Promise_and_Peril_of_Big_Data.pdf

boyd, d., & Crawford, K. (2012). Critical questions for Big Data. Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society, 15*(5), 662–679. doi: 10.1080/1369118x.2012.678878

Busemann, K., & Gscheidle, C. (2012). Web 2.0: Habitualisierung der Social Communitys [Web 2.0: Habitualization of social community use]. *Media Perspektiven*, (7–8), 380–390.

Chew, C., & Eysenbach, G. (2010). Pandemics in the age of Twitter: Content analysis of tweets during the 2009 H1N1 outbreak. *PLoS ONE, 5*(11), e14118. doi: 10.1371/journal.pone.0014118

Christians, C. G., & Chen, S.-L. S. (2004). Introduction: Technological environments and the evolution of social research methods. In M. D. Johns, S.-L. S. Chen & G. J. Hall (Eds.), *Online social research. Methods, issues, & ethics* (pp. 15–23). New York, NY: Peter Lang.

Dodge, M. (2005). The role of maps in virtual research methods. In C. Hine (Ed.), *Virtual methods. Issues in social research on the Internet* (pp. 113–127). Oxford, UK: Berg.

Engesser, S., & Krämer, B. (2011). Die Rückfangmethode. Ein Verfahren zur Ermittlung unzugänglicher Grundgesamtheiten in der Journalismusforschung [Capture-recapture. A method to determine hard-to-reach populations in journalism research]. In O. Jandura, T. Quandt & J. Vogelgesang (Eds.), *Methoden der Journalismusforschung* (pp. 171–187). Wiesbaden, Germany: VS Verlag.

Erlhofer, S. (2010). Datenerhebung in der Blogosphäre: Herausforderungen und Lösungswege [Data gathering in the blogosphere: Challenges and solutions]. In M. Welker & C. Wünsch (Eds.), *Die Online-Inhaltsanalyse* (pp. 144–166). Cologne, Germany: Halem.

Gerring, J. (2001). *Social science methodology: A criterial framework*. Cambridge, UK: Cambridge University Press.

Granger, C. W. J. (1986). Developments in the study of cointegrated economic variables. *Oxford Bulletin of Economics and Statistics*, *48*(3), 213–228. doi: 10.1111/j.1468-0084.1986.mp48003002.x

Herring, S. C. (2010). Web content analysis: Expanding the paradigm. In J. Hunsinger, L. Klastrup & M. Allen (Eds.), *International handbook of Internet research* (pp. 233–249). Dordrecht, The Netherlands: Springer.

Hooper, C. S. (2011). Qualitative in context. *Journal of Advertising Research, 51*, 163–166.

Jankowski, N. W., & van Selm, M. (2005). Epilogue: Methodological concerns and innovations in Internet research. In C. Hine (Ed.), *Virtual methods. Issues in social research on the Internet* (pp. 199–207). Oxford, UK: Berg.

Johns, M. D., Chen, S.-L. S., & Hall, G. J. (2004). *Online social research: Methods, issues, & ethics*. New York, NY: Peter Lang.

Jones, S. (1999a). *Doing Internet research. Critical issues and methods for examining the Net*. Thousand Oaks, CA: Sage.

Jones, S. (1999b). Studying the Net. Intricacies and issues. In S. Jones (Ed.), *Doing Internet research. Critical issues and methods for examining the Net* (pp. 1–27). Thousand Oaks, CA: Sage.

Kearon, J., & Harrison, P. (2011). Research robots. A dramatic new way to conduct research & generate insights. Retrieved November 30, 2012, from http://www.brainjuicer.com/xtra/BrainJuicer_DigiViduals_Research_Robots_Paper.pdf

King, G., & Lowe, W. (2003). An automated information extraction tool for international conflict data with performance as good as human coders: A rare events evaluation design. *International Organization*, *57*(03), 617–642. doi: 10.1017/S0020818303573064

Krippendorff, K. (2004). *Content analysis. An introduction to its methodology* (2nd ed.). Thousand Oaks, CA: Sage.

Lally, E. (2009). Response to Annette Markham. In A. N. Markham & N. K. Baym (Eds.), *Internet inquiry. Conversations about method* (pp. 156–164). Los Angeles, CA: Sage.

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A.-L., Brewer, D., … van Alstyne, M. (2009). Computational social science. *Science, 323*(5915), 721–723. doi: 10.1126/science.1167742

Manovich, L. (2012). Trending: The promises and the challenges of big social data. In M. K. Gold (Ed.), *Debates in the digital humanites* (pp. 460–475). Minneapolis: University of Minnesota Press.

Markham, A., & Buchanan, E. (2012). Ethical decision-making and Internet research: Version 2.0. Recommendations from the AoIR Ethics Working Committee. Retrieved from http://www.aoir.org/reports/ethics2.pdf

Mazur, E. (2010). Collecting data from social networking Web sites and blogs. In S. D. Gosling & J. A. Johnson (Eds.), *Advanced methods for conducting online behavioral research* (pp. 77–90). Washington, DC: American Psychological Association.

McMillan, S. J. (2000). The microscope and the moving target: The challenge of applying content analysis to the World Wide Web. *Journalism & Mass Communication Quarterly, 77*(1), 80–98.

Mehl, M. R., & Gill, A. J. (2010). Automatic text analysis. In S. D. Gosling & J. A. Johnson (Eds.), *Advanced methods for conducting online behavioral research* (pp. 109–127). Washington, DC: American Psychological Association.

Mitra, A., & Cohen, E. (1999). Analyzing the Web. Directions and challenges. In S. Jones (Ed.), *Doing Internet research. Critical issues and methods for examining the Net* (pp. 179–202). Thousand Oaks, CA: Sage.

Murthy, D. (2008). Digital ethnography. An examination of the use of new technologies for social research. *Sociology, 42*(5), 837–855. doi: 10.1177/0038038508094565

Narayanan, A., & Shmatikov, V. (2008). *Robust de-anonymization of large sparse datasets*. Paper presented at the IEEE Symposium on Security and Privacy, 2008. Retrieved from http://www.cs.utexas.edu/~shmat/shmat_oak08netflix.pdf

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York, NY: McGraw-Hill.

Orgad, S. (2009). How can researchers make sense of the issues involved in collecting and interpreting online and offline data? In A. N. Markham & N. K. Baym (Eds.), *Internet inquiry. Conversations about method* (pp. 33–53). Los Angeles, CA: Sage.

Park, H. W., & Thelwall, M. (2005). The network approach to Web hyperlink research and its utility for science communication. In C. Hine (Ed.), *Virtual methods. Issues in social research on the Internet* (pp. 171–181). Oxford, UK: Berg.

Parker, C., Saundage, D., & Lee, C. Y. (2011). *Can qualitative content analysis be adapted for use by social informaticians to study social media discourse? A position paper*. Paper presented at the ACIS 2011: Proceedings of the 22nd Australasian Conference on Information Systems: Identifying the Information Systems Discipline, Sydney, Australia. Retrieved from http://dro.deakin.edu.au/eserv/DU:30041098/parker-canqualitative-2011.pdf

Romero, D. M., Meeder, B., & Kleinberg, J. (2011). Differences in the mechanics of information diffusion across topics: Idioms, political hashtags, and complex contagion on Twitter. *Proceedings of the 20th international conference on World Wide Web*, WWW '11 (pp. 695–704). New York, NY: ACM. doi: 10.1145/1963405.1963503

Russell, M. A. (2011). *Mining the social Web. Analyzing data from Facebook, Twitter, LinkedIn, and other social media sites*. Beijing, China: O'Reilly.

Russom, P. (2011). Big data analytics. Retrieved November 30, 2012, from http://www.cloudtalk.it/wp-content/uploads/2012/03/1_17959_TDWIBigDataAnalytics.pdf

Savage, M., & Burrows, R. (2007). The coming crisis of empirical sociology. *Sociology, 41*(5), 885–899. doi: 10.1177/0038038507080443

Scharkow, M. (2013). Thematic content analysis using supervised machine learning. An empirical evaluation using German online news. *Quality & Quantity, 47*(2), 761–773. doi: 10.1007/s11135-011-9545-7

Scharkow, M., & Vogelgesang, J. (2011). Measuring the public agenda using search engine queries. *International Journal of Public Opinion Research, 23*(1), 104–113. doi: 10.1093/ijpor/edq048

Schrodt, P. A. (2010). *Automated production of high-volume, near-real-time political event data*. Paper presented at the 2010 American Political Science Association meeting, Washington, DC. Retrieved from http://polmeth.wustl.edu/media/Paper/SchrodtEventDataAPSA10.pdf

Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., & Kappas, A. (2010). Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology*, *61*(12), 2544–2558. doi: 10.1002/asi.21416

Utz, S. (2010). Using automated ''field notes'' to observe the behavior of online subjects. In S. D. Gosling & J. A. Johnson (Eds.), *Advanced methods for conducting online behavioral research* (pp. 91–108). Washington, DC: American Psychological Association.

Vogt, W. P., Gardner, D. C., & Haeffele, L. M. (2012). *When to use what research design*. New York, NY: Guilford.

Welker, M., Wünsch, C., Böcking, S., Bock, A., Friedemann, A., Herbers, M., ... Schweitzer, E. J. (2010). Die Online-Inhaltsanalyse: methodische Herausforderung, aber ohne Alternative [Online content analysis: Methodological challenge, but without alternative]. In M. Welker & C. Wünsch (Eds.), *Die Online-Inhaltsanalyse* (pp. 9–30). Cologne, Germany: Halem.

Xifra, J., & Grau, F. (2010). Nanoblogging PR: The discourse on public relations in Twitter. *Public Relations Review*, *36*(2), 171–174. doi: 10.1016/j.pubrev.2010.02.005

Yanovitzky, I., & Greene, K. (2009). Quantitative methods and causal inference in media effects research. In R. L. Nabi & M. B. Oliver (Eds.), *The Sage handbook of media processes and effects* (pp. 35–52). Los Angeles, CA: Sage.

Zheleva, E., & Getoor, L. (2009). To join or not to join: The illusion of privacy in social networks with mixed public and private user profiles. *Proceedings of the 18th international conference on World Wide Web*, WWW '09 (pp. 531–540). New York, NY: ACM. doi: 10.1145/1526709.1526781

Zimmer, M. (2010). ''But the data is already public'': On the ethics of research in Facebook. *Ethics and Information Technology*, *12*(4), 313–325. doi: 10.1007/s10676-010-9227-5