Big Data, Digital Media, and Computational Social Science: Possibilities and Perils

By
DHAVAN V. SHAH,
JOSEPH N. CAPPELLA,
and
W. RUSSELL NEUMAN

We live life in the network. We check our e-mails regularly, make mobile phone calls from almost any location … make purchases with credit cards … [and] maintain friendships through online social networks. … These transactions leave digital traces that can be compiled into comprehensive pictures of both individual and group behavior, with the potential to transform our understanding of our lives, organizations, and societies.

—Lazer et al. (2009, 721).

Powerful computational resources combined with the availability of massive social media datasets has given rise to a growing body of work that uses a combination of machine learning, natural language processing, network analysis, and statistics for the measurement of population structure and human behavior at unprecedented scale. However, mounting evidence suggests that many of the forecasts and analyses being produced misrepresent the real world.

—Ruths and Pfeffer (2014, 1063)

The exponential growth in "the volume, velocity and variability" (Dumbill 2012, 2) of structured and unstructured social data has confronted fields such as political science, sociology, psychology, information systems, public health, public policy, and communication with a unique challenge: how can scientists best use computational tools to analyze such data, problematical as they may be, with the goal of understanding individuals and their interactions within social systems? The unprecedented availability of information on discrete behav-

Dhavan V. Shah is the Louis A. & Mary E. Maier-Bascom Professor at the University of Wisconsin–Madison, where he is director of the Mass Communication Research Center. His work focuses on framing effects on social judgments, digital media influence on civic engagement, and the impact of health ICTs.

DOI: 10.1177/0002716215572084

Downloaded from ann.sagepub.com at UNIV OF WISCONSIN-MADISON on April 10, 2015

iors, social expressions, personal connections, and social alignments provides insights on a range of phenomena and influence processes—from personality traits to political behaviors; from public opinion to relationship formation—despite issues of representativeness and uniformity. That is, even though data from social media may not represent the entirety of a population, that does not mean they are without research value for understanding that population. And the challenges of interpreting these sorts of social data are not limited to population biases and tailored content (Pariser 2011); they extend to the ethics of research practice and personal privacy, the value of theory and reasoning in relation to prediction and engineering, and, of course, the application of appropriate and rigorous modes of inference. This introduction, to a volume considering these possibilities and perils, explores some of the key issues confronting researchers who pursue computational social science in the age of big data.

At the outset, we should explain what we mean by *computational social science* as a specific subcategory of work on big data. It is an approach to social inquiry defined by (1) the use of large, complex datasets, often—though not always—measured in terabytes or petabytes; (2) the frequent involvement of "naturally occurring" social and digital media sources and other electronic databases; (3) the use of computational or algorithmic solutions to generate patterns and inferences from these data; and (4) the applicability to social theory in a variety of domains from the study of mass opinion to public health, from examinations of political events to social movements.

We emphasize that the phrase "naturally occurring" in the above definition is of special importance. Surveys and experiments, the traditional work horses of the social sciences, by their nature involve the intervention of researchers into social processes, engaging unavoidably in various types of experimenter effects (e.g., unintentionally treating experimental and control group subjects differently in ways that shape their responses) and self-report/social desirability biases (e.g., survey respondents tendency to overreport good qualities and behaviors, while underreporting less desirable ones). Computational analyses of big data offer a welcome counterpoint and potential triangulation of multimethod confirmation of key findings in concert with experiments and surveys (Campbell and Fiske 1959). Nonetheless, working with such data remains challenging, not least because of the issues of generalizability, ethics, and theory noted above, but also because of the acquisition, archiving, and analysis of these types of data, which are not easily processed using conventional database applications.

Yet recent increases in storage capacity, boosts in processing power, and the availability of analytic systems have fundamentally expanded the ability of social

Joseph N. Cappella is the Gerald R. Miller Professor of Communication at the Annenberg School for Communication at the University of Pennsylvania. His research focuses on health and political communication and the ways that effective and ineffective messages function in these arenas.

W. Russell Neuman is a professor of media technology at New York University and Evans Professor of Media Technology, Emeritus at the University of Michigan. His research focuses on media effects in the evolving media environment, higher education, and technology policy.

Downloaded from ann.sagepub.com at UNIV OF WISCONSIN-MADISON on April 10, 2015

scientists to collect and utilize these sorts of data. What previously required access to networked computing cores in a dedicated facility can now be handled by a small server cluster housed in a corner of an office or, alternatively, "in the cloud," through a distributed computing system. In this volume, social and electronic media sources are being used by psychologists, epidemiologists, and political and communication scientists to (1) code content and sentiment to infer subjective well-being and personality traits (Schwartz and Ungar, this volume), score the emotionality of news content (Soroka et al., this volume), and trace signals of public opinion (González-Bailón and Paltoglou, this volume); (2) cluster and map networks to understand political alignments (Bode et al., this volume; Freelon et al., this volume), and predict the emergence of online relationships (Welles and Contractor, this volume); and (3) examine dynamics between conventional and social media during presidential debates (Shah et al., this volume) and school shootings (Guggenheim et al., this volume).

Commercial ventures and academic researchers are also deploying large-scale data interventions in social media. This approach introduces changes in the online environments of social media users and observes the consequences in online behavior. For example, Kramer, Guillory, and Hancock (forthcoming) obtained evidence of text-based emotional contagion resulting from small linguistic changes in news stories forwarded to Facebook users. Work of this sort has produced significant and robust, though small, changes in observed behaviors but also provoked outrage from those who have been (or might have been) the targets of the intervention without prior consent. High-profile events such as this may have long-term consequences for how computational social science is undertaken.

Equally ambitious, these sorts of tools and techniques are also deployed to examine entire social networks on a longitudinal basis (Resnick et al., this volume; Han et al. 2011), connect natural language processing with neuroimaging to understand message transmission (O'Donnell and Falk, this volume), and generate more effective health messages through the use of automated filtering systems (Cappella et al., this volume). Such holistic and personalized information collection also speaks to the growing set of conceptual and ethical questions concerning data use and its limits. The acquisition and archiving of complex data systems—let alone their manipulation—often involve collecting personally identifiable information. This forces some reflection on issues of data privacy and de-identification, especially in an era of increased tracking of expression and action, especially regarding physical and mental health (Crosas et al., this volume). Such concerns must be weighed against the value of scholarly understanding, with appropriate steps taken to protect individual privacy and honor the principle of informed consent.

Computational Research and Data Science

On a fundamental level, scholars across the social sciences are questioning the role of big data in relation to conventional methods, theory building, and formal

Downloaded from ann.sagepub.com at UNIV OF WISCONSIN-MADISON on April 10, 2015

scientific reasoning. Hybrid methods that combine or compare established approaches, such as manual content coding or conventional survey research, with computational systems, like machine learning or network mapping, are gaining ground (Burscher et al., this volume; Park et al., this volume; Zamith and Lewis, this volume). Some scholars integrate while others contrast methods to highlight the strengths of each approach, though both camps tend to stress their complementarity. This suggests the need for scholars who can "employ an interdisciplinary skill set that draws from traditional social sciences, statistics, and computer science" (Miller 2011, 1815), and is in sharp contrast to those who suggest abandoning established methods in favor of data science.

Indeed, some have gone so far as to suggest that "with massive data, this approach to science—hypothesize, model, test—is becoming obsolete" (Anderson 2008, 108). Rich data and algorithmic approaches such as machine learning permit more accurate forecasts in many areas (Hindman, this volume), though often absent theoretical justification. Some laud these approaches for allowing the engineering of "useful computational artifacts," even though they may not serve the goal of producing deeper social scientific understanding (Lin, this volume). This is not a perspective we fully share, nor do we advocate the notion that big data will replace and make surveys, lab experiments, clinical trials, and content analyses irrelevant (Anderson 2008).

Nonetheless, from a policy perspective, these systematic predictive and analytic techniques can provide insight into, if not directly solve, significant social problems. The availability of large amounts of social communication about daily life—real-time reactions to media, political, environmental, and social events—and evaluation of those data by the groups producing the content raises the possibility of immediate access to cultural (and subgroup) discourse. For example, we can mine data for insight into environmental pollution patterns by "sniffing social media" (Mei et al. 2014) and understand the spread of contagious diseases as traced through symptom posting (Khoury and Ioannidis 2014). Mining social data can also be used to enable health interventions to better target affected subpopulations (Barrett et al. 2013). In addition, the push toward electronic health records creates opportunities to cull available data to test "the effectiveness of the intervention among real patients in real settings, the safety and side effects of the intervention, and determining for whom the intervention may be most effective" (Hesse et al., this volume).

Computational approaches, in other words, have the capacity to gather and process large quantities of information quickly to serve the public good and examine the public agenda. In doing so, researchers must contend with content intended to mislead citizens and consumers through the deployment of "spam bots" generating comments on everything from political candidates' policy briefs to hotel accommodations' service quality. It is not surprising, then, that in expressing concerns about representativeness of social data, Ruths and Pfeffer (2014) also note the flaws and distortions that emerge as a consequence of nonhuman accounts. Others have sought to find indicators of real and planted commentary in the online environment (Ott et al. 2011). These features can introduce serious distortions into research and must be addressed to ensure data validity.

Downloaded from ann.sagepub.com at UNIV OF WISCONSIN-MADISON on April 10, 2015

While we recognize that a sample of tweets or even a universe of tweets, for example, is not representative of or projectable to a universe of individuals as implied by the term public opinion (Hargittai, this volume), that is not our primary concern here. Indeed, such collections are not even a reliable sample of content circulating within social media more broadly defined (Driscoll and Thorson, this volume). But as an indicator of sentiment or behavior or diffusion in relation to a particular topic at a particular time, it is an increasingly prominent and important indicator of what is occurring in the public sphere, now accessible to systematic and real-time analysis. To manage and analyze these complex datasets, social scientists are forging connections with specialists in mathematics, statistics, engineering, computer science, information systems, and high-throughput computing and are using tools often developed by industry and government, a step toward transdisciplinary work.

Organizing Themes and Vexing Issues

It is with these issues in mind that we organized this volume around five main themes that represent frictions in this field space and reflect the cutting edge of research:

(1) Reflections on tools for collaborative research and computational modeling, with particular attention to prediction, privacy, and sampling biases;
(2) Examinations of language and discourse as indictors of traits, cognitions, and behaviors, as tapped through automated text coding and machine learning;
(3) Studies about social connections in the form of network ties, information flows, and social clustering that define interpersonal connections and political action;
(4) Research considering influences of and on social media as understood in relation to overtime changes in external factors such as traditional media content; and
(5) Advances into complementary procedures, including large-scale data management, recommendation systems, neuroimaging, and hybrid approaches.

As noted above, a number of additional themes cut across these sections and studies. In particular, several articles consider (1) the role of collection systems in computational social science, (2) the need to attend to multiple platforms when sampling content, (3) the tension between conventional and computational methods, (4) the need for team science and transdisciplinary research, and (5) the relation between theory and big data. The richness of this volume rests on these intersections and the core themes that we have selected, which are critical issues for social scientists to confront.

Yet there are other, less obvious themes and issues that arise across these entries. One such emergent issue is the use of inferential statistics in a de facto

Downloaded from ann.sagepub.com at UNIV OF WISCONSIN-MADISON on April 10, 2015

census, albeit one that involves a draw from a larger corpus, or a massive social media sample. A census of every tweet generated last Tuesday could be seen as a "sample" of weekly data, but not likely a meaningful random sample if it is intended to define a broader universe. Given the extraordinarily large sample sizes in much of this research, nearly everything is statistically significant in big data analytics (Lohr 2012). As such, researchers must use inferential statistics carefully, recognizing the risk of "false discoveries"—Type I errors, or the assertion of a relationship that is not present. Indeed, the issue of huge samples making insignificant findings seem meaningful because they achieve conventional thresholds of statistical significance may be a more problematic issue than the question of what constitutes a true census of social data.

At times when the risk of a Type I error looms large, theory can provide essential guidance. In the work by Kramer, Guillory, and Hancock (forthcoming) on emotional contagion, for example, the effects observed in that study are tiny but statistically significant, partly a function of the size of the sample involved. Nonetheless, the authors argue that the effect is meaningful and consequential in part because extant theories driving and explaining processes of emotional contagion are well established. So rather than seeing their results as capitalizing on chance or simply being inconsequential, they argue that they have another manifestation of a core social interactive process—emotional contagion—in a totally unique social media context. In this respect, we see psychological and social theory as critical to interpreting computational findings.

A related concern centers on representativeness of social data relative to population parameters. We contend that the discussion of bias begs the question about what the population of interest is and what can be done to deal with bias. The latter is especially important in an arena where sentiment analysis from social media is seeking to replace or supplement more representative public opinion work. For example, the Twitter "fire hose" is the actual universe of tweets, so there is no bias there if the scholarly aim is to speak to the dynamics in that social space. In contrast, a low-response-rate survey is a biased sample of public opinion, despite the fact that it aims to represent the population. Even a high-response-rate survey represents a biased sample of the public to an extent. In both cases, there are methods to correct for distortion. Scholars must incorporate these, or they must at least acknowledge the ways their data limit inference.

Two additional issues reoccur across the articles collected here and emanate from these smaller issues. One is big data versus theory. Some have argued that big data mean the end of theory, while others assert that theory will be generated through a combination of inductive and deductive approaches (Anderson 2008; boyd and Crawford 2012). There is a sharp tension between these positions, with one camp advancing the view that data science and algorithmic systems can produce faster, deeper, and more accurate and actionable results than scientific specialists incrementally building knowledge, while the other camp argues for the centrality of the interpreter of data and the essential role of theory in big data analytics. Given that some big data collections render statistical hypothesis testing essentially irrelevant because of the overwhelming levels of power, we contend that making sense of findings requires good theory to offer

Downloaded from ann.sagepub.com at UNIV OF WISCONSIN-MADISON on April 10, 2015

clear a priori predictions and sensible explanations of what are otherwise uninterpretable statistical tests (Kramer, Guillory, and Hancock, forthcoming). That is, we tend to fall on the side of the latter position but can see value in the former, especially the inductive approach. Others in this volume have addressed this same issue (e.g., Lin, this volume).

Tied to this is the issue of prediction versus explanation. Many approaches in the world of big data are primarily oriented toward building predictive models that solve a problem, whether commercial or social or political. Is there a role for explanation in heavily prediction-oriented modeling associated with some approaches to big data? A number of our contributors talk about prediction-oriented approaches and seem to suggest that prediction for its own sake is just fine and that explanation—specifically causal explanation—can catch up later, if necessary. We are of the opinion that the outcomes of some prediction-only approaches can provide the grist for explanatory approaches, melding two of the key components of computational social science—successful predictions and explanatory models. For example, as the magnitude of data available allows prediction models on smaller and smaller sets of cases—the individual case in the extreme—then variation in the parameters of those prediction models themselves become interesting objects of explanation and theorizing. The two orientations do not need to be seen as either-or choices but rather as complementary at least or even as opening new avenues of research and theory previously unavailable because of a surfeit of data at the individual level.

Computational Communication Science

It is clear that the era of data science is reshaping the fields of communication, political science, psychology, sociology, and public health. Computational social science, with its focus on large-scale data and social media data, will precipitate other shifts in the commitments and training of researchers, some obvious and some less so. Much of the data of computational social science are and will be textual, and require honing skills in natural language processing. Quantitative social scientists have been accustomed to numerical data, collected either through self-reported responses on scales or the assessment of formal instruments (e.g., skin conductance).

With much of the core social data now in textual form, changing in central ways how data are acquired and reduced, scholars will need to come to new agreements on what constitutes reliable and valid descriptions of the data; the categories used to organize those data; and the tools necessary to access, process, and structure those data. Although communication researchers are well positioned to move into these domains because of their long history of careful assessment of the content of communication, some retooling will be inevitable as will the need for collaborative research with computer scientists and engineers, and the redirection of training for graduate students into the latest and most powerful techniques for analyzing textual materials in electronic form. Visual materials still

Downloaded from ann.sagepub.com at UNIV OF WISCONSIN-MADISON on April 10, 2015

remain a challenge for computational analysis by researchers, but one that is receiving increasing attention (Shah et al., this volume).

The focus on "text-as-data" in much of the work in computational social science places the field of communication at the center of this evolving domain, suggesting the rise of *computational communication science*. Attention to the content of communications—how they are produced and how they are responded to—is central to work in the field. As the articles in this volume show, communication researchers are at the forefront of confronting the challenges of computational social science and integrating insights and approaches from different disciplines to answer the questions posed.

References

Anderson, Chris. 2008. The end of theory: The data deluge makes the scientific method obsolete. *Wired Magazine* 16 (7): 108–9.

Barrett, Meredith A., Olivier Humblet, Robert A. Hiatt, and Nancy E. Adler. 2013. Big data and disease prevention: From quantified self to quantified communities. *Big Data* 1 (3): 168–75.

boyd, danah, and Kate Crawford. 2012. Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society* 15 (5): 662–79.

Campbell, Donald T., and Donald W. Fiske. 1959. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin* 56:81–105.

Dumbill, Edd. 2012. *Planning for big data*. Sebastopol, CA: O'Reilly Media, Inc.

Han, Jeong Yeob, Dhavan V. Shah, Eunkyung Kim, Kang Namkoong, Sun-Young Lee, Tae Joon Moon, Rich Cleland, Q. Lisa Bu, Fiona M. McTavish, and David H. Gustafson. 2011. Empathic exchanges in online cancer support groups: Distinguishing message expression and reception effects. *Health Communication* 26 (2): 185–97.

Khoury, Muin J., and John P. A. Ioannidis. 2014. Big data meets public health. *Science* 346 (6213): 1054–55.

Kramer, Adam D. I., Jamie E. Guillory, and Jeffrey T. Hancock. Forthcoming. Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Science*.

Lazer, David, Alex Sandy Pentland, Lada Adamic, Sinan Aral, Albert Laszlo Barabasi, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, and Myron Gutmann, et al. 2009. Life in the network: The coming age of computational social science. *Science* 323 (5915): 721–23.

Lohr, Steve. 11 February 2012. The age of big data. *New York Times*.

Mei, Shike, Han Li, Jing Fan, Xiaojin Zhu, and Charles R. Dyer. 2014. Inferring air pollution by sniffing social media. In *Advances in social networks analysis and mining*, 534–39. Washington, DC: IEEE Computer Society.

Miller, Greg. 2011. Social scientists wade into the tweet stream. *Science* 333 (6051): 1814–15.

Ott, Myle, Yejin Choi, Claire Cardie, and Jeffrey T. Hancock. 2011. Finding deceptive opinion spam by any stretch of the imagination. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL 2011)*, 309–19. Stroudsburg, PA: Association for Computational Linguistics.

Pariser, Eli. 2011. *The filter bubble: What the Internet is hiding from you*. New York, NY: Penguin Press.

Ruths, Derek, and Jürgen Pfeffer. 2014. Social media for large studies of behavior. *Science* 346 (6213): 1063–64.

Downloaded from ann.sagepub.com at UNIV OF WISCONSIN-MADISON on April 10, 2015