EDITOR'S COMMENTS

Synergies Between Big Data and Theory

By: **Arun Rai**
Editor-in-Chief, *MIS Quarterly*
Regents' Professor of the University System of Georgia
Robinson Chair of IT-Enabled Supply Chains and Process Innovation
Harkins Chair of Information Systems
Robinson College of Business
Georgia State University
arunrai@gsu.edu

Because of its potential to generate significant impact across problem domains, one phenomenon—big data[1]—has been generating significant interest in the IS scholarly community as well as in other disciplines and practice in recent years. *MIS Quarterly* has published two editorials related to big data: one outlining opportunities for IS research at three levels—infrastructure, analytics, and transformation and impact (Goes 2014)—and a second specifying criteria for data science contributions in IS research (Saar-Tsechansky 2015). It has also published a special issue on Business Intelligence Research (2012) and will be publishing a special issue on Big Data and Analytics in Networked Business in December 2016. Other leading IS journals have also published editorials on different aspects of the big data phenomenon (e.g., Abassi et al. 2016; Agarwal and Dhar 2014).

In my previous editorial, I shared the trifecta vision for *MIS Quarterly*, which encompasses *impact* of work that is published; *range* in the problems, theories, and methods in published work; and *speed* of editorial processes. A question that I have seen come up in discussions with colleagues and students is the relationship between big data and theory, whether the two are necessarily incompatible, or can be mutually reinforcing, and, if yes, when and how. I use this editorial to share thoughts on (1) changes in the practices to generate and source data for research, (2) certain cautions that arise from these changes, and (3) synergies that can be achieved between big data and the testing, elaboration, and generation of theory in IS through research designs and methods, as summarized in Table 1.

Changing Landscape in the Generation and Sourcing of Data for Research

Innovations in data-generating technologies (e.g., sensors, Internet of things, social media, mobile devices) are elaborating what data are generated, how the data are generated, and who generates the data. First, turning to what and how data are being generated, technological innovations are making it possible to generate not only datasets with a much larger number of observations, but also datasets where each observation is represented by a larger number of attributes, an increasing amount of unstructured data (e.g., text, image, audio, video), and a less clear dependence with other observations. Real-time traces of activities are being captured on processes to create and appropriate value, experiences of individuals, interactions among individuals and systems; use of IT features for different purposes; and so on. And, spaciotemporal data are enabling us to trace the progression of states, networks, and events over space and time.

Accompanying the increasing availability of granular digital traces of activities, researchers have expanding options to source data. For example, we are witnessing automated extraction of data from online sources; sharing of administrative datasets by government agencies; government-sponsored partnerships to establish big data research infrastructures; data-sharing agreements between researchers and private organizations; and providers of information, other than government agencies, generating indexes on aggregate economic and social activities (Einav and Levin 2014), as I now elaborate.

[1] I use the term big data in a general sense to reflect the dramatic changes in the volume, velocity, variety, and veracity of data commonly associated with the big data phenomenon.

Table 1. Synergies Between Big Data and Theory: The Changing Data Landscape, the Cautions, and the Approaches

Changing Landscape in the Generation and Sourcing of Data for Research	Automation of data extraction	Crawling and scraping of web sites and using APIs to extract data
	Government agency datasets	Construction, validation, and sharing of data sets by government agencies
	Government-sponsored big data research infrastructures	Development of regional/national big data infrastructures that can be used to address pressing issues
	Data sharing agreements with private organizations	Negotiated access to corporate data on activities such as transactions, interpersonal interactions, and policy interventions as well as outcomes
	Breaking government monopolies on the reporting of aggregate activities	Private sources and university researchers are generating indexes that are released faster and are more granular than those provided by government agencies
Cautions Accompanying the Changes to the Generation and Sourcing of Data	Avoid the streetlight effect	Caution about chasing easy-to-access datasets at the expense of focusing on important problems
	Address foundational empirical issues	Pay attention to foundational principles of measurement: • Do the derived measures capture the construct of interest? • Are the measures comparable across observations? • Self-interest affecting the data generation process: Blue team and Red team dynamics
	Address the rigidity of data capture schemas	Design innovations address the dual objectives of data quality and a rich representation of the phenomenon
Achieving Big Data–Theory Synergies Through Research Designs and Methods	Theory to provide focus and ensure measure correspondence	• Enable selection of constructs and relationships examined • Specify boundary conditions • Lens to evaluate correspondence between constructs and measures
	Precision in theory testing	• Generate granular measures of activities and outcomes • Create new measures from unstructured data • Solidify robustness of evidence underlying claims by accounting for confounding effects • Conduct cross-validation tests to mitigate risks of overfitting data to models and to evaluate the predictive utility of models
	Contextual elaboration of theory	Role of Theory • Select omnibus and situational contextual characteristics where different explanations and outcomes may manifest • Explain how and why differences arise across contextual characteristics Role of Research Design and Methods • Develop context-dependent estimates that are masked by average treatment effects • Uncover interactions and nonlinear effects through machine learning approaches
	Generating theory	Role of Theory • Use emergent patterns in data to conceptualize phenomenon and generate theoretical conjectures and explanations Role of Research Design and Methods • Uncover patterns in data through computational, statistical, and qualitative theory building approaches

Automation of data extraction: With the expanding range of economic transactions and social activities occurring online, researchers are extracting data by crawling and scraping web sites and using application program interfaces (APIs), as evidenced by several publications in our leading journals. For example, they have used automated processes to capture data on prices and reviews for products from online shopping platforms such as eBay and Amazon.

Construction and sharing of datasets by government agencies: Public sector initiatives are underway worldwide to construct and openly share data that can be used by researchers. For example, the Centers for Medicare and Medicaid Services make available data for hospitals such as reimbursements, healthcare quality outcomes by diagnosis-related group, and meaningful use of IT[2]; the Government of Singapore has established a one-stop portal for administrative datasets[3]; and the Government of India has established a similar one.[4] The "Smart Cities" initiative announced by the U.S. Federal Government in 2015 is directing investments in technologies that will capture and share granular data in key facets of citizen activity, with the objective to catalyze innovative solutions to traffic congestion, crime, economic growth, climate change, and the delivery of city services.[5]

Development of government-sponsored big data research infrastructures: Government-sponsored initiatives are establishing big data infrastructures that can be used by various stakeholders including researchers to address pressing economic and societal issues. For example, the National Science Foundation (NSF) recently sponsored four Big Data Regional Innovation Hubs that will establish the data and infrastructure resources for research to address regional challenges in a variety of science and education domains.[6] The program has generated collaboration commitments from over 250 organizations that include universities, cities, foundations, and Fortune 500 corporations. As another example, Research Councils United Kingdom (RCUK), the strategic partnership of UK's seven research councils, is investing £189 million in its Big Data and Energy Efficient Computing program for researchers and industry to collaborate for scientific discovery and development in a wide variety of problem domains.[7]

Data-sharing agreements with private organizations: Researchers are establishing data-sharing agreements with organizations to source granular, restricted-access data (e.g., online tracking of stocks and flows in supply chains, use of services and features on digital platforms, social media interactions, health monitoring and outcomes).

Breaking government monopolies on the reporting of aggregate activities: Private organizations and researchers are stepping in to generate indexes on aggregate activities in social and economic systems. They are doing so faster and with greater granularity than being done by government agencies in various countries. The Billion Prices Project (BPP) at the Massachusetts Institute of Technology generates daily price indexes by collecting information on daily prices and product attributes from the web sites of hundreds of online retailers worldwide.[8] Premise, an economic data tracking platform which states its mission as "improving people's communities and livelihoods across the world by increasing economic and societal transparency," generates a wide range of indexes (e.g., food prices, percentage of electrified homes). It uses a model that involves (1) boots-on-the-ground daily data collection by 16,000 paid contributors in 200 cities spread over 30 countries, who use a mobile phone app to capture pictures and other contextual information, and (2) machine learning algorithms and human experts that monitor the submitted data to rate contributors and refine sampling design.[9]

Clearly, the expanding availability of granular, real-time data from a variety of sources creates enormous possibilities for knowledge creation, but also requires us to be attentive to some important issues.

[2] Source: http://www.resdac.org/cms-data/request/cms-data-request-center; accessed April 17, 2016.

[3] Source: https://data.gov.sg; accessed April 17, 2016.

[4] Source: https://data.gov.in; accessed April 17, 2016.

[5] Source: https://www.whitehouse.gov/the-press-office/2015/09/14/fact-sheet-administration-announces-new-smart-cities-initiative-help; accessed April 17, 2016.

[6] Source: http://www.nsf.gov/news/news_summ.jsp?preview=y&cntn_id=136784; accessed April 17, 2016.

[7] Source: http://www.rcuk.ac.uk/research/infrastructure/big-data/; accessed April 17, 2016.

[8] Source: http://bpp.mit.edu/; accessed April 17, 2016.

[9] Baker, D., "Photos Are Creating a Real-Time Food Index," http://www.wired.co.uk/magazine/archive/2016/04/features/premise-app-food-tracking-brazil-philippines; accessed April 17, 2016.

Cautions Accompanying the Changes to the Generation and Sourcing of Data

There are three areas that I suggest we be particularly attentive to: (1) the risk of formulating questions around easily available datasets rather than important problems, (2) foundational empirical issues, and (3) the rigidity of data categorization schemas, as I now discuss.

1. **Avoiding the streetlight effect:** The low cost of sourcing certain data introduces the risk of the *streetlight effect*, or the drunkard effect (Freedman 2010). As the allegory goes, a police officer finds a drunk man crawling on his hands and knees searching for his wallet, only to learn that the drunk man has focused his search there because of better lighting than across the street where the drunk man thinks he most likely dropped his wallet. Although the changing landscape of how data are generated and sourced is creating exciting possibilities for research, we need to heed caution not to formulate research around easy-to-access datasets, but rather we need to maintain focus on important problems.

2. **Addressing foundational empirical issues:** Big data still requires the researcher to address foundational issues of empirical research such as construct validity and reliability of measures, dependencies among observations, comparability of measures across observations, and selection bias (Einav and Levin 2014; Lazer et al. 2014; Patty and Penn 2015). The importance about the quality of measures, for instance, is very effectively brought up in Lazer et. al's (2014) discussion on why the Google Flu Tracker (GFT) dramatically overpredicted the proportion of doctor visits for influenza-like illness than the Centers for Disease Control and Prevention (CDC), although the GFT was developed to predict the CDC estimates. Part of the issue was that the initial version of the GFT was "part flu detector, part winter detector" (Lazer et al. 2014, p. 1203). In fact, GFT developers report removing seasonal search terms such as *high school basketball* that were structurally unrelated to the flu but were correlated with the CDC data and the propensity of the flu—a sign that the GFT was overfitting the data. An *ad hoc* approach to deleting seasonal search terms did not solve the problem, with the GFT entirely missing the nonseasonal 2009 influenza A–H1N1 pandemic.

   Self-interest affecting the data-generating process: As data are generated by a variety of sources, we need to consider if the self-interest of the sources may affect the data-generating process. In the case of data-generating processes on digital platforms, there are at least two types of self-interest dynamics to consider: "blue team" dynamics that characterize the influence of the self-interest of the platform owner and "red team" dynamics that capture the influence of the self-interests of platform-service users and parties affected by the signals from the data streams (Lazer et al. 2014).

   > *Blue team dynamics*: Data generated by private companies such as Amazon, Google, Facebook, and Twitter are likely to be affected by changes to their business models and underlying technologies and algorithms to improve services to customers. For example, with 86 reported changes in June and July 2012 to the Google search algorithm, the observed search patterns are affected by the changes made by the company's programmers (Lazer et al. 2014). The rapid pace of change by platform owners to their business models and data-generating processes, and the likely secrecy of these changes for competitive reasons, can make it challenging for researchers to replicate studies conducted on some of these platforms. These confounds also make it challenging to validly analyze and compare data collected longitudinally from these platforms.
   >
   > *Red team dynamics*: As we, as a society, pay greater attention to signals from open platforms such as Twitter, the incentives for certain stakeholders to manipulate these signals increase. For example, in closely contested marketing or political campaigns, stakeholders may manipulate data streams to trend favorably over rivals in the real-time court of public opinion. Indeed, we are seeing an expanding array of tactics such as the use of bots to manipulate these signals that are being countered by solutions to prevent, detect, and correct for such pollution (Freitas et al. 2015).

3. **Addressing the rigidity of data categorization schemas:** As we design big data research infrastructures, we need to address the tension between the stability of preexisting categorization schemas that may have worked well for historical data and the need to challenge and revise ontological assumptions underlying these schemas when anomalies are detected in new observational data. We are witnessing design innovations that capture large-scale data on activities and experiences from diverse sources that were hitherto infeasible and that analyze these data to detect breakdowns in existing categorization schemas. By revising the schemas to avoid overfitting data to predefined categories, these design innovations address the

dual objectives of data quality and a rich representation of the phenomenon. *PatientsLikeMe* (PLM) is a powerful example of an organization that breaks from the traditional processes of conducting medical research, in which data were collected through medical tests and medical professionals within the confines of a medical consultation (Kallinikos and Timpeni 2014). PLM uses a model that is built on networking and computational technologies to collect self-reported data from an open, distributed base of patients. There are two complementary novel characteristics of PLM's model: (1) routinely generating diverse information flow through unsupervised data entry by patients from the contexts of their everyday living in which they experience symptoms and the effects of treatments, and (2) using the expertise of trained medical professionals such as RNs to refine categories to which symptoms are mapped so as to avoid overfitting symptoms into preexisting categories.

IS scholars have an opportunity to lead the scholarly conversation on how systems can be designed in important problem contexts to (1) generate information flow from diverse sources that can provide potentially valuable data for knowledge discovery and (2) refine categorization schemas based on anomalies so to avoid overfitting data to preexisting categories.

Achieving Big Data–Theory Synergies Through Research Designs and Methods

Theory and big data are not mutually exclusive, and can be synergistic in advancing our knowledge about phenomena and how we solve problems. The pathways to achieve synergies between theory and big data are manifold, including (1) theory providing a conceptual framework to work with big data, (2) testing theory with precision, (3) elaborating theory to achieve greater granularity in explanation and accuracy in prediction, and (4) generating theory for emergent phenomena, as I briefly discuss.

1. **Theory to provide focus and ensure measure correspondence:** Theory can help make sense of big data in that theory can inform the selection of constructs (out of a very large number of possibilities), the boundary conditions, and the relationships among constructs that are meaningful to examine. Theory has been referred to as the "heart of measurement" (Patty and Penn 2015) and can provide a lens to evaluate correspondence between constructs with the vast number of measures that can be created from high-dimensional data (Einav and Levin 2014).

2. **Precision in theory testing:** Big data can enhance precision in theory testing in multiple ways. First, we are now able to devise measures for previously hard to measure constructs by leveraging the granular observation of activities and outcomes as well as methods to extract information from increasingly complex unstructured data.

   Second, the expanded coverage and detail of data make it possible to carry out extra investigative work to strengthen the evidence underlying claims (e.g., robustness of identifying assumptions and matching strategies to rule out potential sources of confounding). In addition, large-scale randomized online experiments, typically involving a researcher partnering with a private organization, are making it possible to disambiguate effects that were traditionally hard to differentiate and avoid biases in the identification of causal effects (e.g., Aral and Walker 2012).

   Third, although predictive models that can be developed using machine learning techniques do not help in concluding causality, they can play useful roles in developing better predictive models for counterfactuals that can enable researchers to better estimate the causal effect of a treatment (Varian 2014).

   Fourth, large-scale datasets are enabling researchers to design cross-validation tests so that models are evaluated with holdout samples across contexts. These empirical strategies shift the focus from sample uncertainty (which becomes much less of a consideration with very large datasets) to model uncertainty (Varian 2014), mitigate the risk of overfitting models to the data, make it feasible to assess the generalizability of models across contexts, and enable researchers to refine the model based on contextual considerations, as I now discuss.

3. **Contextual elaboration of theories:** Datasets incorporating contextual information (or those from which such information can be extracted through computational, statistical, or qualitative methods) can enable researchers to detect anomalies where explanations break down across contexts. These anomalies can be useful to develop models that achieve context-dependent predictions and estimates and elaborate theories by integrating the role of context.

Broadly speaking, there are two types of context: (1) omnibus, or broad, context (e.g., who, when, where); and (2) discrete, or particular, situational variables (e.g., individual, technology, or team characteristics) that can directly or indirectly affect outcomes of interest (Johns 2006). Instead of considering only average treatment effects, models can be developed to uncover differences in treatment effects over time and space and across subpopulations (e.g., demographic, geographic, political affiliation, health status, personality type, and so on). By using methods to flexibly model interactions in high dimensions, heterogeneous treatment effects can be estimated (Wager and Athey 2015). As the salient influence of contextual characteristics is discovered, theories can be elaborated to integrate the roles of omnibus and situational contextual characteristics and correspond better to the real-word phenomenon (Van de Ven 2007).

4. **Generating theory for emergent phenomena:** Big data creates opportunities for researchers to generate theoretical insights about important problems and emergent phenomena by analyzing data through a variety of methods (e.g., statistical, computational, qualitative), and their combinations, without starting with a preconceived theory. Indeed, qualitative researchers have used inductive approaches such as grounded theory that require intensive engagement by the researcher in data collection, coding, validation, and interpretation to develop explanations for the phenomenon of interest (Sarker et al. 2013). We are witnessing the application of computational approaches such as Latent Dirichlet Allocation (LDA) topic modeling (Blei et al. 2003) to uncover the distribution of "topics" (or concepts) in text corpus as well as in each document. Strategies such as varying the parameters of algorithms to evaluate the robustness of topics, visualization of topics, and expert interpretation are typically used to assess meaningfulness and retention of topics and accord definitions to them. Developments in computational approaches for the analysis of unstructured data raise opportunities to evaluate when and how these approaches can be synergistically combined with qualitative approaches to generate theory.

Concluding Thoughts

There are significant synergies to be realized between big data and theory. Scholars across IS research traditions can utilize a variety of research designs and methods to achieve these synergies between big data and testing, elaborating, and generating theory in highly consequential problem domains. We need to be cautious not to let easy-to-access data sway us to study dull and piffling problems. In addition, the changes in how data are generated and sourced require us to be vigilant to foundational issues for empirical research such as data quality, measurement, dependence structures among observations, and selection bias. The IS community is also well positioned to contribute to knowledge of how to effectively design big data research infrastructures that will significantly accelerate knowledge discovery in key problem domains. These are exciting times for the IS discipline, and *MIS Quarterly* is especially well positioned to be at the forefront of publishing the best work that advances IS knowledge in high-impact problem domains by synergistically combining theory and big data.

References

Abbasi, A., Sarker, S., and Chiang, R. H. 2016. "Big Data Research in Information Systems: Toward an Inclusive Research Agenda," *Journal of the Association for Information Systems* (17:2), pp. i-xxxii.

Agarwal, R., and Dhar, V. 2014. "Editorial—Big Data, Data Science, and Analytics: The Opportunity and Challenge for IS Research," *Information Systems Research* (25:3), pp. 443-448.

Aral, S., and Walker, D. 2012. "Identifying Influential and Susceptible Members of Social Networks," *Science* (337), pp. 337-341.

Blei, D. M., Ng, A. Y., and Jordan, M. I. 2003. "Latent Dirichlet Allocation," *The Journal of Machine Learning Research* (3), pp. 993-1022.

Einav, L., and Levin, J. 2014. "Economics in the Age of Big Data," *Science* (346:6210), 1243089.

Freedman, D. H. 2010. *Wrong: Why Experts* Keep Failing Us—and How to Know When Not to Trust Them,* New York: Little, Brown and Company.

Freitas, C., Benevenuto, F., Ghosh, S., and Veloso, A. 2015. "Reverse Engineering Socialbot Infiltration Strategies in Twitter," in *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, J. Pei, F. Silvestri, and J. Tang (eds.), New York: ACM, pp. 25-32.

Goes, P. 2014. "Editor's Comments: Big Data and IS Research," *MIS Quarterly* (38:3), pp. iii-viii.

Johns, G. 2006. "The Essential Impact of Context on Organizational Behavior," *Academy of Management Review* (31:2), pp. 386-408.

Kallinikos, J., and Tempini, N. 2014. "Patient Data as Medical Facts: Social Media Practices as a Foundation for Medical Knowledge Creation," *Information Systems Research* (25:4), pp. 817-833.

Lazer, D., Kennedy, R., King, G., and Vespignani, A. 2014. "The Parable of Google Flu: Traps in Big Data Analysis," *Science* (343), pp. 1203-1205.

Patty, J. W., and Penn, E. M. 2015. "Analyzing Big Data: Social Choice and Measurement," *PS: Political Science and Politics* (48:1), pp. 1-14.

Saar-Tsechansky, M. 2015. "Editor's Comments: The Business of Business Data Science in IS Journals," *MIS Quarterly* (39:4), pp. iii-vi.

Sarker, S., Xiao, X., and Beaulieu, T. 2013. "Guest Editorial: Qualitative Research in Information Systems: A Critical Review and Some Guiding Principles," *MIS Quarterly* (37:4) pp. iii-xviii.

Van de Ven, A. H. 2007. *Engaged Scholarship: A Guide for Organizational and Social Research*, New York: Oxford University Press.

Varian, H. R. 2014. "Big Data: New Tricks for Econometrics," *The Journal of Economic Perspectives* (28:2), pp. 3-27.

Wager, S., and Athey, S. 2015. "Estimation and Inference of Heterogeneous Treatment Effects Using Random Forests," *arXiv* preprint arXiv 1510.04342.