BIG DATA, DATA SCIENCE, AND ANALYTICS: THE OPPORTUNITY AND CHALLENGE FOR IS RESEARCH

ABDUL KHADEER, PAMBALA NAGESWARA RAO, J RANJITH
4.KALYAN KUMAR, 5.TAMAKANTH ROOHI
Assistant Professor, Department of Computer Engineering, Ellenki college of Engineering and Technonlogy, patelguda (vi), near BHEL ameenpur (m), Sangareddy Dist. Telangana 502319

.

ABSTRACT

It is difficult, nay, impossible to open a popular pub- lication today, online or in the physical world and not run into a reference to data science, analytics, big data, or some combination thereof. To use a Twitter-esque phrase, that's what's trending now. A search on Google in the middle of August 2014 for the phrases "Big data," "Analytics," and "Data science" yielded 822 million, 154 million, and 461 million results, respectively. Our major journals (Management Science, MIS Quarterly) have commissioned special issues on these topics, and a large number of position announce- ments in the information systems field are specifying knowledge of, and skills in, one or more of these areas as desirable, if not a requirement for the job. A new journal called Big Data,1 launched just over a year ago, is already seeing thousands of downloads of articles. It would not be hyperbole to claim that big data is possi- bly the most significant "tech" disruption in business and academic ecosystems since the meteoric rise of the Internet and the digital economy.

What does this tsunami mean for information sys- tems researchers? Rather than simply rely on our own view of the world, we invited other accomplished scholars with a history of publication in this arena to participate in the conversation. We posed five ques- tions to frame our thinking about the domain:

1. Are big data, analytics, and data science, as being described in the popular outlets, old wine in new bot- tles or is it something new?
2. What are the strengths that the information systems (IS) community brings to the discourse on business analytics? In other words, what is our com- petitive advantage?
3. What are important and interesting research questions and domains that may "fit" with on-going research in our community? How might we push the envelope by extending or modifying our existing research agendas? What about new areas of inquiry?
4. To what extent should robust prediction prowess be used as a criterion in evaluating data-driven mod- els versus current criteria that favor "explanatory" models without subjecting them to rigorous tests of future predictability?
5. As editors and reviewers, how should we evalu- ate research in this domain? What constitutes a "sig- nificant" contribution?

We would like to acknowledge the contributions of Galit Shmueli and Martin Bichler whose thought- ful responses to these questions gave us much food for contemplation. In the rest of this commentary we offer a synthesis of our collective reflection on the five questions.

On the Novelty of Big Data, Analytics and Data Science

We believe that some components of data science and business analytics have been around for a long time, but there are significant new questions and oppor- tunities created by the availability of big data and major advancements in machine intelligence.2 While the notion that

ISSN (PRINT): 2393-8374, (ONLINE): 2394-0697, VOLUME-3, ISSUE-10, 2016
DOI: 10.21276/ijcesr.2016.3.10.5

analytical techniques can be used to make sense of and derive insights from data is as old as the field of statistics, and dates back to the 18th century, one obvious difference today is the rapid pace at which economic and social transactions are moving online, allowing for the digital capture of big data. The ability to understand the structure and con- tent of human discourse has considerably expanded the dimensionality of data sets available. As a result, the set of opportunities for inquiry has exploded exponentially with readily available large and com- plex data sets related to any type of phenomenon researchers want to study, ranging from deconstruct- ing the human genome, to understanding the pathol- ogy of Alzheimer's disease across millions of patients, to observing consumer response to different market- ing offers in large scale field experiments. And, easy (and relatively inexpensive) access to computational capacity and user-friendly analytical software have democratized the field of data science allowing many more scholars (and practitioners) to participate in the opportunities enabled by big data.

In some ways in could be argued that the nature of inquiry has also changed, turbocharged by machines becoming a lot smarter through better algorithms, and by information technologies that enable people and things to be inherently instrumented for observation and interaction that feeds the algorithms. Increasingly, data are collected not with the aim of solely testing a human-generated hypotheses or essential record- keeping, but to the extent that data torrents are cap- tured inexpensively, often for the possibility of testing hypotheses that have not yet been envisioned at the time of collection. When such data are gathered on a scale that observes every part of the joint distributions of the observed variables (behaviors, demographics, etc.), the computer becomes an active question ask- ing machine as opposed to a pure analytic servant. By initiating interesting questions and refining them without active human intervention, it becomes capa- ble of creating new knowledge and making discover- ies on its own (Dhar 2013). It can, for example, dis- cover automatically from a large swath of healthcare system data that younger people in a specific region of the world are becoming increasingly diabetic and then conjecture and test whether the trend is due to specific habits, diet, specific types of drugs, and a

2 Arguably, for the first time in history, a machine passed the famous Turing test by defeating human champions at Jeopardy where topics are not known in advance and questions are posed in natural language of considerable complexity and nuance.

range of factors we may not have hypothesized as humans. This is powerful. As scientists, we have not seriously entertained the possibility of theory origi- nating in the computer, and as science-fiction-like as that may sound, we are in principle already there.

New challenging problems and inquiry also lead to research on better algorithms and systems. Since the torrent of data being generated is increasingly unstructured and coming from networks of people or devices, we are seeing the emergence of more pow- erful algorithms and better knowledge representation schemes for making sense of all of this heteroge- neous and fragmented information. Text and image processing capability are one frontier of research, with systems such as IBM's Watson being on the cutting edge in natural language processing, albeit with a long way to go in terms of their capability for ingest- ing and interpreting big data across the Internet.

Networks, such as those created by connections between individuals and/or products, further create significant and unique challenges at a fundamental level such as how we sample them or infer treat- ment effects. For example, in A/B testing, a "standard approach" for estimating the average treatment effect of a new feature or condition by exposing a sam- ple of the overall population to it, the treatment of individuals can spill over to neighboring individu- als along the structure of the underlying network. To address this type of "social interference," newer algorithms are required that support valid sampling and estimation of treatment effects (Ugander et al. 2013). This is but one example of how "relational" and "networked" data necessitates new development in algorithms. Developments may emerge not only from computer science

ISSN (PRINT): 2393-8374, (ONLINE): 2394-0697, VOLUME-3, ISSUE-10, 2016
DOI: 10.21276/ijcesr.2016.3.10.5

but also from IS or other disciplines where researchers are "closer" to the problem being studied than pure methods researchers tend to be.

Finally, the Internet has fueled our ability to conduct large scale experiments on social phenomena. As of this writing, Facebook researchers conducted a massive study to determine whether the mood of users could be manipulated and found that it could (Kramer et al. 2014). By conducting controlled experiments in large numbers of people such studies can extract the causal structure among variables. While the study raised important questions about privacy and the ethical implications of conducting experi- ments without informed consent, the broader point is that researchers now have a medium for theory development through massive experimentation in the social, health, urban, and other sciences.

A Comparative Advantage for IS?

Arguably, this is the golden age for IS researchers.

Data science and big data research from IS has attracted attention at the level of widely read scien- tific outlets such as PNAS and Science (Aral et al. 2009, Aral and Walker 2012) because of the impor- tance and generic nature of the questions asked, such as "are choices in social networks a result of influ- ence or homophily?" Such a question has profound implications for phenomena too numerous to mention that involve how we communicate, persuade, moni- tor, and more. In other words, perhaps it is time to set our sights higher, beyond our traditional journals, to communicate with the larger community of scien- tists and businesses. It is a time of opportunity for social scientists that have heretofore been hamstrung by the lack of data. For the first time we are able to observe and measure human behavior on a global scale. IS researchers are, quite incredibly, at the center of the digital world unfolding before us.

To put things in historical perspective, one could reasonably assert that the IS discipline has the longest history of conducting research at the nexus of com- puting technology and data in business and society. In fact, it is widely recognized that the discipline of "MIS" emerged when computers enabled the automation of business processes and the digital capture of business transactions. Understanding how to design and implement systems to provide the "right infor- mation to the right person at the right time" was the raison d'être of IS programs and defined much of early IS research. These endeavors entailed understanding what individuals' and executives data needs are, designing structures to capture and manage data, and conceptualizing interfaces that made the data accessible and usable. So, it is not unreasonable to claim that IS scholars started off with a comparative advantage with respect to big data: we knew how to store, manage, process data; and about the complexity of data structures very early on in the history of computing. We also understand the challenges associ- ated with the infrastructure needed for handling the volume of data being generated today.

Armed with these skills and tools, IS scholars have been catalyzed to focus on problems and outcomes. As has been suggested elsewhere (Agarwal and Lucas 2005) among all functional areas of business, IS re- searchers perhaps have the broadest perspective on the enterprise as a whole, and how different pieces fit together. This focus creates a tighter linkage between data and business models: we care deeply about business transformation and value creation through data, and less for algorithms or frameworks without a linkage to business value. Our research has been alternately praised and criticized for being too cross- disciplinary, but we believe this is strength and not a weakness in today's data rich environment. IS scholars have invoked theories from the fields of economics,

sociology, psychology, and political science to name a few, and studied phenomena such as electronic mar- kets, consumer behavior, crowdsourcing, information security, and online retailing from a diversity of per- spectives. This cross- and transdisciplinary nature of IS research to date positions us uniquely to exploit the big data opportunity. We can do so by addressing the same types of questions as we have in the past but with significantly richer data sets, or we can begin to initiate new inquiries that represent questions that were not possible to ask and answer previously.

ISSN (PRINT): 2393-8374, (ONLINE): 2394-0697, VOLUME-3, ISSUE-10, 2016
DOI: 10.21276/ijcesr.2016.3.10.5

On Research Questions

The wealth of possibilities enabled by big data is too many to enumerate in a brief commentary. Nonethe- less, we offer some illustrations of fruitful research projects that IS scholars could begin to explore (and, evidence suggests that many already are). First, the ability to observe and measure micro, individual-level data on a comprehensive scale enables us to address grand problems on a societal level with deep pol- icy implications that go beyond the confines of a single organization. Examples include using micro- level technology usage data to ask if ubiquitous dig- itization exacerbates or diminishes social inequities, exploring how labor markets are evolving by exam- ining the actions of workers on social networking and employment sites, and investigating if the quan- tified self-movement with sensors tracking exercise and nutrition information during daily living is, in fact, producing a measurable effect on the incidence of disease or health problems. Indeed, the entire field of personalized medicine (and, the heretofore under- studied opportunity of personalized technology inter- ventions) is enabled by big data.

Second, following the call to focus on the transformational aspects of IT (Lucas et al. 2013), big data allows for the design and execution of studies related to the profound changes our own profession is expe- riencing. We could extend our existing research agen- das that have explored the effects of technology medi- ation on learning outcomes, learner satisfaction, etc., to examine individual-level effects of MOOCs, online courses, blended learning, and the like on an unprece- dented scale.

Third, and this is an area where perhaps IS research has a robust set of studies to build on, is big data research in the context of social networks and marketing outcomes. Geo-coded social media interac- tions coupled with extensive demographic and socioe- conomic data allow us to quantify how networks affect micro (individual behavior), meso (organizational value), and societal outcomes (economic and social value). Pervasive mobile devices and the rapid rise in commercial transactions (banking, purchase, etc.) on mobile platforms that can be captured and recorded enable novel insights into the classic market- ing problems of advertising and promotions, and their effects on sales. And of course, the ability to run field experiments in these settings, varying interventions on a grand scale, substantially enhances the scope of causal relationships we can extract from the data.

Big Data, and Prediction vs. Explanation

At the time of writing this editorial, two big data stories were receiving substantial media attention. One, the scathing review of Google flu trends (Lazer et al. 2014) (that uses search terms to predict the inci- dence of flu) with respect to its accuracy as com- pared with estimates produced by the Centers for Disease Control and Prevention, and two, the ethi- cal debate ignited by experiments conducted on Face- book. In some ways both stories are related to the issue of whether big data are useful solely for pre- diction or also for an understanding of the causal processes that are yielding the observed outcomes. In the case of Google flu trends, the algorithm has been criticized for overfitting a small number of cases and masking a simple question, namely, does it pre- dict flu or is it merely reflecting the incidence of winter, and for not taking into account the fact that technologies like Google's search engine are profit- driven and changing and therefore limited for sci- entific inquiry. While the Facebook study raises the specter of widespread experimentation on the Inter- net without adequate protection of individual rights and privacy, ironically, the experiment is a great example of where causal claims can be made with some confidence. This tension between correlational analysis and causal testing of hypotheses represents a fundamental dilemma in the use of big data for expla- nation versus prediction.

But we should not be too hasty in dismissing prediction. Indeed, as has been argued by the philoso- pher of science Karl Popper, prediction is a key epis- temic criterion for assessing how seriously we should entertain a theory or a new insight: a good theory makes "bold" predictions that stand repeated effects as falsification

ISSN (PRINT): 2393-8374, (ONLINE): 2394-0697, VOLUME-3, ISSUE-10, 2016
DOI: 10.21276/ijcesr.2016.3.10.5

(Popper 1963). Popper goes on to criti- cize certain social science theories like those of Adler or Marx that can conveniently "bend" the theory to accommodate even contradictory data without any onus on prediction whatsoever.3 In this regard, we

3 Popper used opposite cases of a man who pushes a child into theshould encourage additional rigorous testing of mod- els on data that were collected later in time than those used to construct the models.

Prediction as an initial basis for theory building also has particular value in a world where patterns often emerge before the reasons for them become apparent (Dhar 2013). While the scientific process is predicated on a cycle of hypothesis generation, experimentation, hypothesis testing, and inference, there are multiple starting points. Big data are as, and increasingly more so, useful at the hypothesis generation stage as they are at the hypothesis testing phase. A study that seeks to predict rather than explain may reveal associa- tions between variables that form the foundation for the development of theory that can be subsequently subject to rigorous testing. In this context, Shmueli (2010) provides an elegant summary of when predic- tive modeling can be particularly useful in scientific endeavors, including newly available rich data sets with newly measured concepts for which theory is yet to be developed (as can be the case with big data), and for the discovery of new measures. Scientific progress does not rely on a single study, and big data-based studies offer the promise of novel discoveries.

Some domains may often view prediction to be as, if not more, valuable than explanation. A compelling example of this is in healthcare, where the cost of delaying action based on a good predictive model until the construction and testing of explanatory mod- els is complete, is measured in lives that may be lost. This is not to say that clinical and biomedi- cal researchers do not seek to build causal models, quite the contrary. The far-reaching human genome project and recently launched efforts to understand the human brain in greater detail aim to unravel the underlying causal structures of disease. But in many instances prediction with big data in and of itself is of immense value such as determining the probability of hospital re-admittances, or the risk of development of hepatocellular carcinoma among patients with cirrho- sis (Waljee et al. 2014). The biomedical community has begun to acknowledge that big data provides a criti- cal complement to the gold standard of randomized controlled trials by supporting massive observational studies that were not feasible before Weil (2014).

Evaluating Knowledge Claims Based on Big Data

There is little doubt that the IS community will increasingly look for ways to conduct innovative re- search that leverages the power of big data. Some of water with the intention of drowning the child and that of a man who sacrifices his life in an attempt to save the child. In Adler's view, the first man suffered from feelings of inferiority (produc- ing perhaps the need to prove to himself that he dared to commit the crime), and so did the second man (whose need was to prove to himself that he dared to rescue the child at the expense of his own life). this research may be "non-traditional" in that it devel- ops predictive models of phenomena for their own sake or as a first step towards theory building. Some of it may use new variables and measures enabled by the digital capture of social and economic activ- ity, and user generated context creatively combined with external data sets such as that obtained from the Census Bureau or the Bureau of Labor Statistics. How should we, as editors and reviewers, evaluate such research?

There is no simple answer to this question and, to a large extent, our standards and methods of eval- uation will inevitably evolve as our experience with such research grows. With respect to ISR in particu- lar, we revert to the general criteria used by editors and reviewers when assessing research: Fit, Interest- ingness, Rigor, Story, Theory (Agarwal 2012). From the standpoint of fit, certainly novel and original research that uses big data to address a phenomenon of interest to the IS community "fits" with the mis- sion of ISR, and is welcomed. Among the other cri- teria, while it is difficult to provide a rank ordering, we suspect that interestingness and rigor will be, at the margin, more important in evaluating research based on big data, at least in these

ISSN (PRINT): 2393-8374, (ONLINE): 2394-0697, VOLUME-3, ISSUE-10, 2016
DOI: 10.21276/ijcesr.2016.3.10.5

early stages. Above all, the research must pose an interesting and rel- evant question. We expect that questions that chal- lenge conventional wisdom and intuition or those that pose a hitherto unexplored puzzle, i.e., those that are novel, will be received more positively by review- ers than those that have been examined extensively in prior work. But we should also encourage testing and replication of prior results since such research is key to scientific advancement, as argued eloquently by Popper (1963). Unfortunately, such work is often given short-shrift relative to research that claims new results based on data sets that are not shared and do not provide opportunity for falsification or con- firmation. Data sharing and transparency must be encouraged.

Interesting is hard to define explicitly since it could arise in many different forms. Part of the interesting- ness of a question might well derive solely from the power of big data, where it becomes possible to inves- tigate a previously formulated research question with a novel and extensive data set that integrates observa- tions from diverse sources. Indeed, it is entirely pos- sible that the contribution of a study lies primarily in the uniqueness of the data set and the rigor of the empirical methods used to analyze the data.

With respect to rigor, we do not expect the stan- dards for evaluating knowledge claims in studies using big data to be substantially different from those using more traditional data sets. Reviewers would expect the usual threats to inferential validity and causal claims including self-selection and identifica- tion to be adequately accounted for. Researchers must also pay attention to the special challenges of working with very large data sets and reflect thoughtfully on the economic significance of their findings. In contrast to studies that utilize archival data collected by "trust- worthy" entities such as businesses, governmental and other agencies, or studies based on primary data collection by researchers, big data could originate from unknown sources with questionable at best or unknown at worst credentials. It is incumbent upon authors to convince readers that their "big" data set was validly generated from appropriate foundations. Second, and this challenge is not unique to big data but may be compounded by sheer size and variety in variables, data that the researcher does not generate herself is often an imperfect observation for the real world concept that is being referred to. For exam- ple, if we are not able to capture individuals' income in a social network, is the zip code in which they reside an appropriate proxy for socioeconomic sta- tus? The answer, of course, depends on a variety of interrelated factors including the nature of the study, its research questions, other variables the researcher has, etc. Nonetheless, this example serves to illustrate the special difficulty that may be amplified as a func- tion of the number of variables in the data set, espe- cially when they are integrated from multiple sources.

Closing Thoughts

As a community of scholars we would be remiss not

to take full advantage of the scientific possibilities cre- ated by the availability of big data, sophisticated ana- lytical tools, and powerful computing infrastructures. Indeed, for reasons mentioned above, this is an excit- ing time to be an IS researcher and to think beyond IS to science in general. Big data still aims in large part to deliver the right information to the right person at the right time in the right form, but is now able to do so in a significantly more sophisticated form. The IS discipline has been thinking and researching ques- tions at the intersection of technology, data, business, and society for five decades and should leverage its thought leadership to become a centerpiece of educa- tion, business, and policy.

References

Agarwal R(2012) Editorialnotes. Inform. Systems Res. 23(4):1087–1092.

Agarwal R, Lucas HC (2005) The IS identity crisis: Focusing on high visibility and high impact research. MIS Quart. 29(3):381–398.

Aral S, Walker D (2012) Identifying influential and susceptible members of social networks. Science 20:337–341.

Aral S, Muchnik L, Sundararajan A (2009) Distinguishing influence

based contagion from homophily driven diffusion in dynamic networks. Proc. Natl. Acad. Sci. USA 106(51):21544–21549.

Dhar V (2013) Data science and prediction. Comm. ACM 56(12):

64–73.
Kramer ADI, Guillory JE, Hancock JT (2014) Experimental evidence of massive-scale emotional contagion through social networks. Proc. Natl. Acad. Sci. USA 111(24):8788–8790.
Lazer LD, Kennedy R, King G, Vespignani A (2014) The parable of Google flu: Traps in big data analysis. Science 343:1203–1205.
Lucas HC, Agarwal R, Clemons E, El-Sawy O, Weber B (2013) Impactful research on transformational IT: An opportunity to inform new audiences. MIS Quart. 27(2):371–382.
Popper K (1963) Conjectures and Refutations (Routledge, London).
Shmueli G (2010) To explain or to predict? Statist. Sci. 25(3):289–310. Ugander U, Karrer B, Backstrom L, Kleinberg J (2013) Graph cluster randomization: Network exposure to multiple universes.
Proc. 19th ACM SIGKDD Internat. Conf. Knowledge, Discovery, and Data Mining (ACM, New York).
Waljee AK, Higgins PDR, Singal AG (2014) A primer on predictive models. Clin. Translational Gastroenterology 5(1):e44.
Weil AR (2014) Big data in health: A new era for research and patient care. Health Affairs 33(7):1110.