EDITOR'S COMMENTS

Big Data and IS Research

By: **Paulo B. Goes**
Editor-in-Chief, *MIS Quarterly*
Salter Professor of Technology and Management
Head, Management Information Systems
Eller College of Management
University of Arizona
pgoes@eller.arizona.edu

For the last two or three years, the field of "big data" has emerged as the new frontier in the wide spectrum of IT-enabled innovations and opportunities allowed by the information revolution. The ever-increasing creation of massive amounts of data through an extensive array of several new data generating sources has prompted organizations, consultants, scientists, and academics to direct their attention to how to harness and analyze big data. Businesses are looking for a technology-based competitive advantage, while the public, academic, and scientific sectors look for opportunities to understand the world in unprecedented ways. The expectations are high that big data will propel our society into an exciting era of across the board innovations.

As with any new emerging technology wave, quotes and wild forecasts abound:

> *"Information is the oil of the 21st century, and analytics is the combustion engine."* (Peter Sondergaard, Gartner Group)
>
> *"Data is the new science. Big Data holds the answers."* (Pat Gelsinger, EMC)
>
> *"Data are becoming the new raw material of business."* (The Economist 2010)

Gartner's hype curve predicts big data will start making deep transformational impact in two to five years (Heudecker 2013).

Aside from splashy quotes and predictions by consultants, company executives are actually trying to do something about the new wave, or at least thinking about it (IBM 2011), and even midmarket companies are gearing up for it (Dell Midmarket Research 2014).

In my role as department head of a prominent IS department, which has always been active in interdisciplinary collaboration, sponsored research, and business partnerships, I have been part of extensive discussions about big data in the last two years both with potential industry partners as well as academic collaborators in my university. I find that in industry, there is a considerable gap in the understanding of the area, its challenges, and its potential. Except for a few large companies, especially information-intensive ones like Linkedin, Facebook, and Google, executives of most corporations and midmarket companies are struggling with understanding and deciding what to do. The confusion is exacerbated by the highly fragmented environment of solutions and applications that are intended to work in the big data realm.

Universities have been moving to address the industry gaps. IS groups have been responding especially to the opportunity of delivering academic programs that specialize in data and business analytics, to form data scientists. Such programs are proliferating fast.

As far as academic research, big data opportunities lure. Through its funding agencies, the federal government has identified big data as a strategic priority for funding. A quick search of grant opportunities in the U.S. government's site *grants.gov* using "big

data" as the key term will return dozens of opportunities, with the National Institute of Health and the National Science Foundation leading in number of opportunities. However, efforts in most academic institutions are still scattered, which I believe is normal for a fast-evolving area, in which most of the innovation leadership is coming from a few sectors of industry.

I give as example my institution, the University of Arizona, a top research state university. In my opinion, its big data research efforts represent the status of big data research in many research universities. For the last two years, I have been involved in several big data initiatives, with the aim of developing a coherent strategy for the University to be able to go after funding, as well as to foster interdisciplinary research among its several units. I have cochaired a University-wide big data research task force and currently chair the big data research committee, which is charged to inform the University's upper administration on a plan for big data infrastructure resources and development to support big data research.

I generally see three types of academic units that come to discussions about big data research:

(1) *The hard core science units*. Research in astronomy, climate science, and genomics, for example, have always relied on large datasets. The data streams generated by today's sophisticated equipment, such as telescopes or weather satellites, are really large. These research units are also the ones on campus that consume most of the high-performance computing resources available. In my opinion, yes, they do work with very large datasets and require HPC, and they know how to conduct this research already. However, I find this research only marginally related to the big data paradigm.

(2) *The information sciences units*. In any modern university, information sciences are scattered and fragmented around campus. This situation is a consequence of the advances of the information revolution, which permeates all facets of society and impacts science and academics in so many ways. At the University of Arizona, we have Computer Science in the College of Science, a "traditional" computer science department that focuses on important issues of network protocols, security, data storage, software engineering, etc. There is also information sciences in the School of Information Retrieval and Library Sciences, the Electrical Engineering Department, and a new unit that was formed recently, the School of Information Systems Technology and the Arts, which has been trying to become the seed of an i-School. And, of course, there is my own department, the Management Information Systems (MIS) Department in the College of Management. Each unit has a different focus. In my opinion, each unit can contribute to the big data paradigm, but at present the approach resembles that well-known cartoon of making sense of an elephant by grabbing isolated parts of the animal. Because of its proximity to business, where big data industry leadership is today, and because of our background in data and systems integration, I think we are better equipped to act as coordinators of big data interdisciplinary efforts. I'll elaborate more on this issue later.

(3) *All units that have realized the potential of working with diverse and large datasets*. There is a growing number of such units, among them those related to health care, public health, education, public policy, government studies, marketing and retail, and finance, just to name the more obvious ones. They bring to the discussion research questions that simultaneously need behavioral, social, and technical approaches anchored on data analysis. These units are close to their domain and research questions, but typically lack the technical capabilities to work in the big data environment. They understand the power to relate diverse data sources, but lack the infrastructure skills, the integration capabilities and the analytics acumen.

Big data has been defined by the 4 V's: volume, velocity, variety, and veracity. The new paradigm comes by combining these dimensions. The hard core science disciplines have been working on volume and perhaps velocity, but it's the 4 V's together, the integration of the several data sources, different data types, making sure we work with valid data, that are what this paradigm is all about.

I personally find variety the most interesting dimension of big data from an IS perspective. Putting together data from sensors and the "Internet of things," the vast repository that we call the Web, user-generated content, social media, data generated and consumed on mobile platforms, and data from enterprise systems, allows researchers to ask and answer questions that explain and predict individual behavior and detect population trends. These are the interesting interdisciplinary questions. And we IS researchers have been working on related questions since the Internet and e-commerce came about. These forces have allowed the discipline to look outward.

I see opportunities for IS research in the big data environment at three levels: (1) big data infrastructure, (2) big data analytics, and (3) transformation and impact. To help with the discussion about where IS research fits, I present below a taxonomy that we

have been using at the University of Arizona to characterize big data research projects along the dimensions of infrastructure and analytics. The aim of this taxonomy was to create an inventory of the various big data initiatives on campus, find the common threads, challenges and opportunities to help create campus wide strategies. Here I use it to highlight opportunities for IS research and the leadership role IS research must play.

Big Data Infrastructure

Many of the research topics that can be derived from Figure 1 are not typically the realm of IS research. The technical issues related to capture, streaming, archiving, parallel computing, etc. are generally Computer Science and Electrical Engineering topics. Also, many of the technical infrastructure innovations are being driven and led by industry. For example, the Hadoop family of technologies and services originated from Google. The same goes for cloud services and some other technologies. However, there are several IS topics related to big data infrastructure management, governance, and control that need research contribution. For example:

- Data integration is key to the big data infrastructure. A good deal of IS research has looked at entity reconciliation and record matching. See, for example, Dey et ak, (1998) and Zhao and Ram (2005). These are critical issues in big data both at the semantic and syntactic levels.

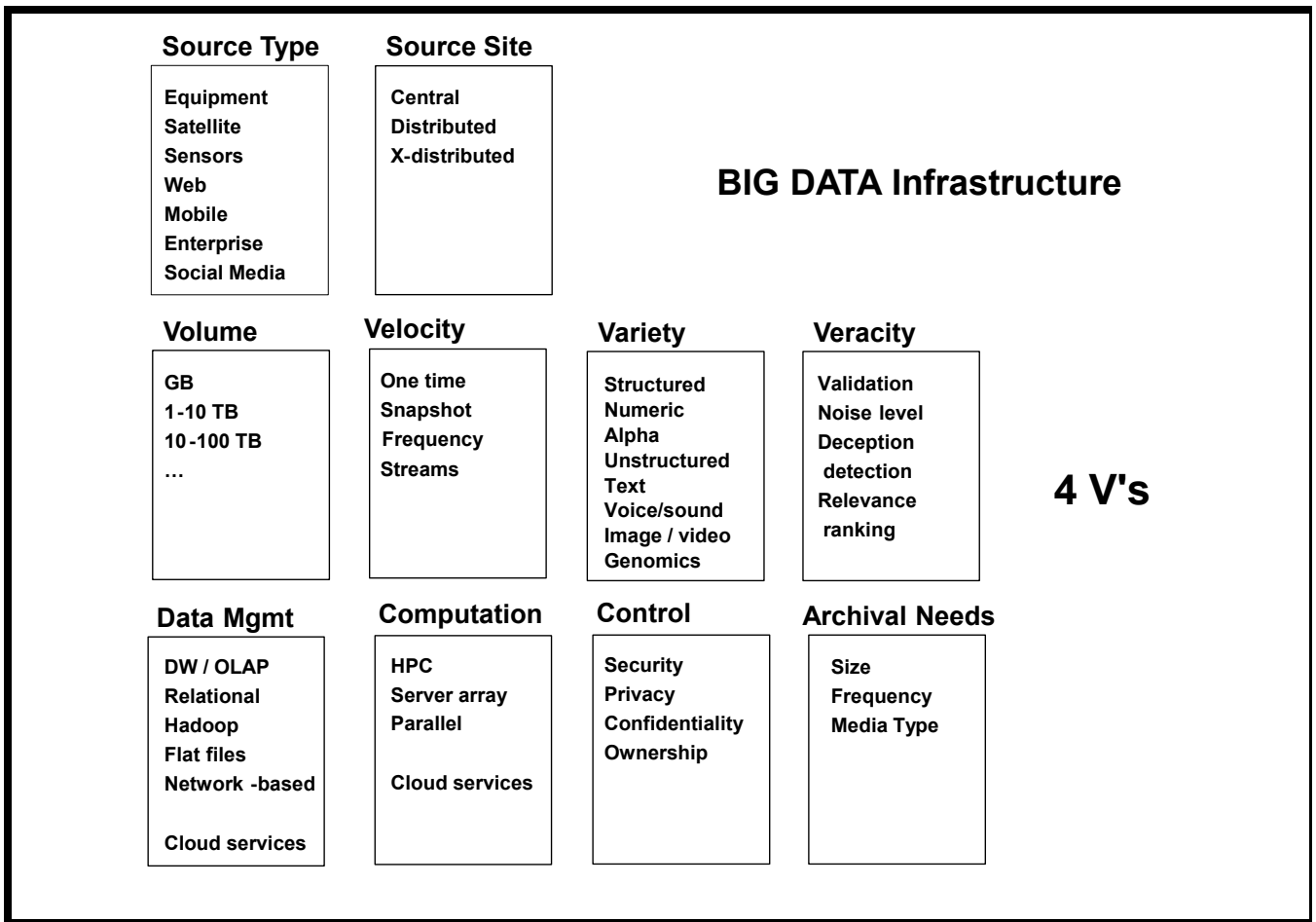

Figure 1. Big data Infrastructure Taxonomy

- Validating the veracity of the data, sorting out the noise and the malicious information from valid, actual information has been the subject of IS research, and will continue to be extremely important with big data. Deception detection in unstructured data, relevance-based rankings, etc. are topics in which IS researchers can contribute in a substantial way (Fuller et al. 2011; Jiang et al. 2005; Nunamaker et al. 2011). And of course we as a field have worked extensively on the topic of trust in the digital era; see, for example, the 2010 "Special Issue on Trust Research" in *MIS Quarterly*.

- Designing and managing a platform of services, which will often be a patchwork of several components, is also an exciting IS research area. Researchers can build on the vast existing literature that looks at what drives different infrastructure implementation choices. There are several works on optimization of cost, benefits and service levels. Provision through the cloud adds another dimension. Economic modeling based on game theoretic or principal–agent models may be useful here too, and let us not forget about contract theories, vendor management, etc. We have a solid base upon which to build.

- Security has in recent years become a mainstream IS research, as evidenced by the increasing number of submissions to *MIS Quarterly*, and the formation of vibrant workshops and conferences such as IFIP Technical Committee 11 (Security and Privacy Protection in Information Processing Systems). IS security research encompasses behavioral, technical, economic, and organizational approaches. In the big data realm, addressing the various security and control issues, including privacy and confidentiality, are of fundamental current and future importance.

Studying the IT governance issues that come with big data provides a great window for IS researchers. We have contributed a lot to this area, which we own as researchers. Issues related to business alignment, provision of services, ownership, etc., take very interesting forms in the inherent multidisciplinary, multientity and fast moving environment of big data.

Analytics and Decision Support Systems

Figure 2 presents a taxonomy to describe big data Analytics research projects. Analytics is that layer of services that are directly tied to decision making. It is an extension of our own well-established IS research area of decision support systems enhanced by very sophisticated techniques and wide and deep extensive data sources.

The IS field has been moving along the hierarchy Data → Information → Knowledge → Intelligence. Analytics refers to the upper stages of the hierarchy: the generation of knowledge and intelligence to support decision making and strategic objectives.

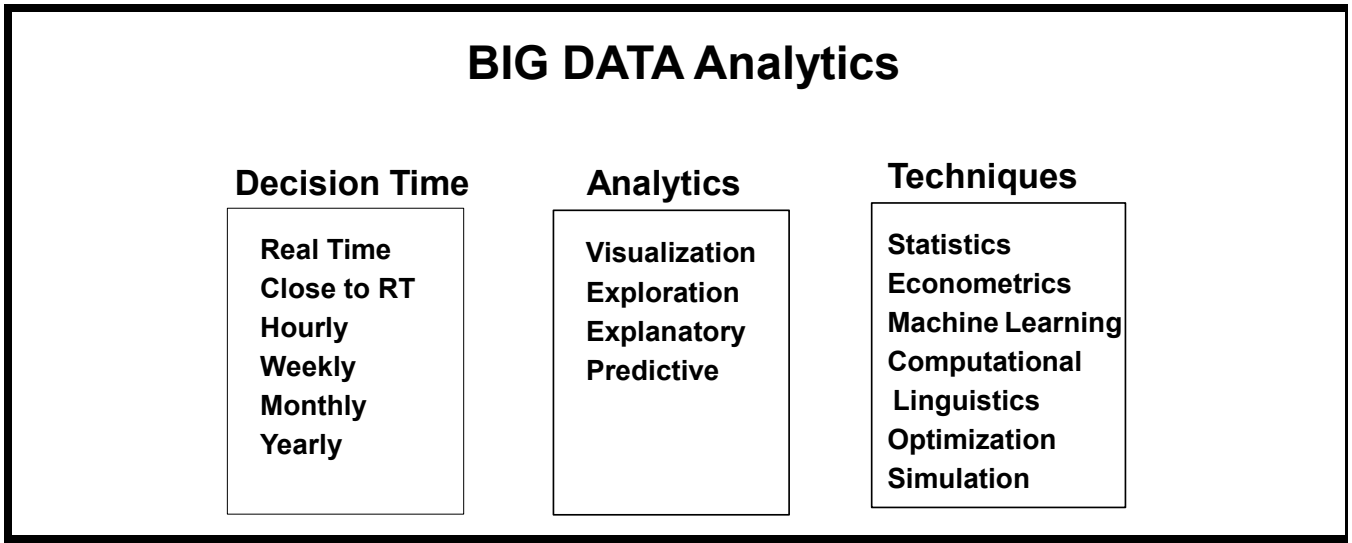

Figure 2. Simplified Taxonomy of Big Data Analytics

IS Research in E-Commerce and Digital Marketing: Springboard to Big Data Analytics

In decision making, context is key, therefore knowledge of the domain is fundamental. Interdisciplinarity is the name of the game for analytics research. The Internet and e-commerce boom of the late 1990s and early 2000s allowed the IS field to thrive working together with other business disciplines in an interdisciplinary way. Marketing, in particular, has been the object of intense interdisciplinary work by IS researchers published in top journals in both IS and Marketing (for example, Dellarocas 2012; Ghose et al. 2013; Yang and Ghose 2010), and many other interdisciplinary IS–Marketing research projects at New York University, Massachusetts Institute of Technology, University of Minnesota, University of Arizona, Carnegie-Mellon University, and their forward-looking analytics research centers). As a discipline, Marketing relies more and more on information technology. From channel choice to personalization and recommendation systems, user-generated content, online reviews, and social influence in social networks, IS researchers have led impactful interdisciplinary analytics research. IS researchers have utilized various data sources, mastered collection and analysis of web data and user-generated content, and have become very proficient in the utilization of advanced econometric and machine learning techniques. The seeds for working with the 4V's of big data addressing interdisciplinary research have been planted. An important forum to advance this research has been SCECR—the Statistical Challenges on eCommerce Research, an annual workshop (http://scecr.org/scecr2014/).

IS research in e-commerce and digital marketing has shown us the way to work with large, disparate data sources. IS researchers have also started to work in Finance, Operations, Strategy, and other business disciplines using the big data paradigm. The *MIS Quarterly* "Special Issue in Business Intelligence and Analytics" published in December 2012 carries six excellent examples of this research.

Big data is about massive amounts of observational data, of different types, supporting different types of decisions and decision time frames. In terms of methodologies for observational data analytics, the IS field has evolved tremendously, especially with econometric modeling. In the last eight years, the field has achieved maturity in employing advanced econometric models, including hierarchical longitudinal models, latent models and structured models. These models have been employed primarily in explanatory contexts, studying causation effects.

The big data world is moving toward real-time or close to real-time decision making. The IS field needs to embrace and develop context-dependent methodologies that strengthen prediction and co-occurrence.

Large-scale network analysis is extremely important in the big data environment. Modeling explicit and implicit interactions derived from the vast amounts of data is critical. With big data, these interactions are now visible. A few IS researchers (Aral et al. 2014; Zhang et al. 2013) have started to delve into large-scale network analytics, but we need much more. The models are extremely complex, starting from sampling techniques, inference, and identifying influence. The networks are complex, multirelational (links can mean different relationships), dynamic, and evolve very fast.

As mentioned above, due to the availability of very fine micro data from different sources, big data allows modeling individuals at a very detailed level with a rich representation of the environment surrounding them with their social, economic, and cognitive dimensions. The implications for disciplines such as healthcare, education, and marketing/retail are limitless. Barriers between online and offline representations of human behavior are disappearing fast.

Precise capture of individual behavior and surrounding events also allows for spotting population trends and the impact of events such as emergency situations, disease outbreak, and severe climate impact. With high instrumentation of everything, including people, with geospacial tracking capabilities, the notion of smart cities should become a reality, with impact on e-government, public safety, public health, and public administration.

In their introductory article to the "Special Issue on Business Intelligence and Analytics" in MISQ, Chen et al. (2012) identify the following five areas as "big impact" areas of big data research: (1) e-commerce and market intelligence, (2) e-government and politics, (3) science and technology, (4) smart health and well being, and (5) security and public safety.

Transformation and Impact

I now go back the point I made above about interdisciplinary collaboration with other units on campus and the current status of big data research in a typical university today. Huge opportunities exist for collaboration in big data. No other "information sciences" group has what IS groups have to engage in these interdisciplinary collaborations. Our experience with topics such as data integration, infrastructure design and evaluation, and service platforms are unique to IS. Our track record with interdisciplinary collaboration in the fields of e-commerce and digital marketing can catapult us into interdisciplinary collaboration with big data. We do have to be open-minded as a discipline, to identify and collaborate on non-business problems. The information revolution has pushed the boundaries of our discipline outward and we should embrace the changes and provide leadership in the new environment.

To adapt to the data-driven knowledge economy, organizations are shifting. Particularly, IS groups need to provide new leadership. CIOs are being converted into Chief Innovation Officers, and significant changes in the IT function are taking place. Analytics are taking hold in organizations who demand "data science" professionals. Chief Analytics Officers are being created to head data analytics groups. IS leadership has to be involved and lead these changes. There is much room for our research to inform organizations about the issues we know best: business alignment, IT and knowledge value, business process transformation, strategic analytics, deployment, and utilization of the new systems.

It is a great time to be in IS doing IS research. We are uniquely positioned in this new big data analytics environment. Let's embrace the challenges and claim our territory. In closing, I would like to promote the *MIS Quarterly* "Special Issue Transformational Issues in Big Data Analytics" and invite innovative submissions.

References

Aral, S., and Walker, D. 2014. "Tie Strength, Embeddedness & Social Influence: A Large-Scale Networked Experiment," *Management Science* (60:6), pp. 1352-1370.

Chen, H., Chiang, R., and Storey, V. 2012. "Business Intelligence and Analytics: From Big Data to Big Impact," *MIS Quarterly* (36:4), pp. 1165-1188.

Dell Midmarket Research. 2013. "Dell Survey: Midmarket Companies Aggressively Embrace Big Data Projects," Press Release (http://www.dell.com/learn/us/en/uscorp1/press-releases/2014-04-28-dell-software-big-data-midmarket-survey).

Dellarocas, C. 2012. "Double Marginalization in Performance-Based Advertising: Implications and Solutions," *Management Science* (58:6), pp. 1178-1195

Dey, D., Sarkar, S. ,and De, P. 1998. "A Probabilistic Decision Model for Entity Matching in Heterogeneous Databases," *Management Science* (44:10), pp. 1379-1395.

Economist. 2010. "Data, Data Everywhere," February 25 (http://www.economist.com/node/15557443).

Fuller, C. M., Biros, D. P., Burgoon, J. K., and Nunamaker, J. F. 2011. "An Examination and Validation of Linguistic Constructs for Studying High-Stakes Deception," *Group Decision and Negotiation* (22), pp. 17-134.

Ghose, A., Ipeirotis, P., and Li, B. 2012. "Designing Ranking Systems for Hotels on Travel Search Engines by Mining User-Generated and Crowd-Sourced Content," *Marketing Science* (31:3), pp. 493-520.

IBM. 2011. "Insights from the 2011 IBM Global CIO Study" (http://www-935.ibm.com/services/c-suite/cio/study/).

Heudecker, N. 2013. "Hype Cycle for Big Data, 2013," July 31 (https://www.gartner.com/doc/2574616).

Jiang, Z., Mookerjee V. S., and Sarkar, S. 2005. "Lying on the Web: Implications for Expert Systems Redesign," *Information Systems Research* (16:2), pp. 131-148.

Nunamaker, J. F., Derrick, D. C., Elkins, A. C., Burgoon, J. K., and Patton, M. 2011. "Embodied Conversational Agent (ECA) Based Kiosk for Automated Interviewing," *Journal of Management Information Systems* (28:1), pp. 17-48.

Yang, S., and Ghose, A. 2010. "Analyzing the Relationship Between Organic and Paid Search Advertising: Positive, Negative, or Zero Interdependence?," *Marketing Science* (29:4), pp. 602-623.

Zhang, B., Thomas, A. C., Doreian, P., Krackhardt, D., and Kerishnan, R. 2013. "Constrasting Multiple Social Network Autocorrelations for Binary Outcomes, with Appilcations to Technology Adoption," *ACM Transactions on Management Information Systems* (3:4), pp. 1-18.

Zhao, H., and Ram, S. 2005. "Entity Identification for Heterogeneous Database Integration: A Multiple Classifier System Approach and Empirical Evaluation," *Information Systems* (30:2), pp. 119-132.