MIS Quarterly

BIG DATA & ANALYTICS IN NETWORKED BUSINESS

TRANSFORMATIONAL ISSUES OF BIG DATA AND ANALYTICS IN NETWORKED BUSINESS

Bart Baesens
KU Leuven/University of Southampton, Leuven, BELGIUM {Bart.Baesens@kuleuven.be}

Ravi Bapna
Carlson School of Management, University of Minnesota, Minneapolis, MN U.S.A. {rbapna@umn.edu}

James R. Marsden
School of Business, University of Connecticut, Storrs, CT U.S.A. {jmarsden@business.uconn.edu}

Jan Vanthienen
KU Leuven, Leuven, BELGIUM {Jan.Vanthienen@kuleuven.be}

J. Leon Zhao
City University of Hong Kong, Hong Kong, CHINA {jlzhao@cityu.edu.hk}

The era of big data and analytics is upon us and is changing the world dramatically. The field of Information Systems should be at the forefront of understanding and interpreting the impact of both technologies and management so as to lead the efforts of business research in the big data era. We need to prepare ourselves and our students for this changing world of business. In this discussion, we focus on exploring the technical and managerial issues of business transformation resulting from the insightful adoption and innovative applications of data sciences in business. We end by providing an overview of the papers included in this special issue and outline future research directions.

Introduction

The term *big data* may seem omnipresent, but our use of the term entails more than just a large amount of data. Goes (2014), in his editor's comments on big data and IS research, discussed the relationships among the four V's of big data:

> 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 (p. iv).

The 4V definition is but a starting point that outlines the perimeters. The definition does not help us to determine what to do inside the perimeters, how to innovatively investigate and analyze big data to enhance decision making quality, how to anticipate and leverage the transformational impacts of big data, or how best to consider scope as well as scale impacts of big data. We argue for a fifth "V," namely *value*, to complement the 4V framework from a business perspective. In order to extract the fifth V, the value from big data, one needs to go beyond the 4Vs and (1) understand the four major sources of

big data, and (2) appreciate the dual nature of prediction and causality as major methodological paradigms to harness big data.

Alternatively, big data can be defined by its origin in terms of transactions versus non-transactions and internal versus external (Zhao et al. 2014). In the era of data management, most data are captured via business transactions involving stakeholders, processes, products, and services. In recent years, data beyond the corporate boundaries are becoming increasingly available, including significant data from social networks and the Internet-of-things. Further, data can be acquired either internally or externally. In this regard, we use big data to refer to data that includes information acquired externally together with data gathered internally.

We begin this special issue with a detailed consideration that includes our own perspectives on these issues, knowing full well that we are but one segment of many concerned with issues surrounding the transformational impacts of big data. In fact, in the summer of 2015, SCECR (Statistical Challenges in e-Commerce Research), labeled the "original big data conference," held its 15th annual conference. The interested reader might consider Press's (2013) recent *Forbes*' article, which provides a history and brief commentary of the evolution of big data.

Big data in today's globally connected networked economies arises from five major sources:

(1) Large-scale enterprise systems: these comprise the alphabet soup of systems (enterprise resource planning (ERP), customer relationship management (CRM), supply chain management (SCM), and others) that companies have been deploying for close to two decades now.

(2) Online social graphs: if one takes the combination of the major social networks such as Facebook, Twitter, Weibo, and WeChat, we have close to two billion people doing what they like to do, such as interacting with friends, accessing media, and socially networking, and leaving a digital trail in the process, a trail that can be tracked, graphed, and analyzed.

(3) Mobile devices: with close to 5 billion handsets worldwide and with the mobile channel serving as the primary gateway to the Internet in large swaths of India and China, this is another source of big data as every action taken by a user can be tracked and potentially geotagged.

(4) Internet-of-things: the emerging sensor enabled ecosystem to connect objects with each other and with humans. This is the foundation of tomorrow's smarter physical ecosystems, using sensors to connect physical objects (homes, automobiles, even garbage bins and street lights) while generating big data in the process.

(5) Open data/public data: data about topics including weather, traffic, maps, environment, and housing are increasingly becoming available.

Many of the issues in big data may not be new, but there is an evolving positive view of big data and business analytics that is resulting in real business transformations. The use of the term *business analytics* is now becoming standard to communicate the full life cycle of enhanced data-driven business decision making. Analytics goes beyond business intelligence, in that it is not simply more advanced reporting or visualization of existing data to gain better insights. Instead, analytics encompasses the notion of going behind the surface of the data to link a set of explanatory variables to a business response or outcome. This linkage can be in a predictive sense or toward making causal inference. Business analytics, as we see it, encompasses all aspects of the data process to facilitate predictive and/or causal inference-based business decision making. The range of inference techniques includes randomized field experiments, harvesting of observational data from multiple sources, simulation, and laboratory experimentation. Business analytics borrows from statistics, econometrics, machine learning, and distributed computing paradigms. Business analytics differs from data science in that analytics focuses on better data-driven decision making in an organizational context.

As we write, businesses are hiring and integrating specialists across the gamut from data warehousing to high-level analytics. Academic institutions have been and continue to rapidly develop educational programs and research initiatives to meet the demand for individuals at the forefront of analytics expertise. The website "Master's in Data Science" (http://www.mastersindatascience.org) indicates that 112 universities in their directory offer Business Analytics programs. A 2011 McKinsey & Company report suggested that

> The United States alone faces a shortage of 140,000 to 190,000 people with deep analytical skills as well as 1.5 million managers and analysts to analyze big data and make decisions based on their findings (p. 3).

As universities move to deliver programs in data analytics and companies seek to fill their needs, it is important to recognize

the impact of transformational eruptions related to big data. For businesses, big data is (1) impacting business processes, (2) rebalancing the power of relationships in decision making in the commercial world, and (3) altering the scope as well as the scale of optimization challenges. For academics, big data is (1) shifting the IT research environment to become more integrated with other fields rather than "adopting paradigms" from those fields; (2) delivering the opportunity for a triangulation of data using an integrated set of market observations, empirical data, and focused survey data with the ability to cross validate; and (3) challenging the researcher to develop new methods and techniques to effectively leverage big data.

Access to big data and the tools to perform deep analytics suggests that power now equals ***information (data) + trust***. Our concern is that the former part of this equation, the data, has gotten the attention while the latter, trust, has not. In what follows, we set forth our consideration, our examination, of the transformational impacts of big data including the expanded importance of trust. We begin with a basic and time-honored concept, data quality, which continues to be far too frequently ignored. Although data quality concerns are not new, the transformational impacts and growing role of big data add significance to the importance of data quality. We then consider how big data impacts the appropriateness of methodological paradigms and the challenges posed, followed by an illustration of a set of innovative big data applications. We then focus on the disruptive impacts of big data, provide our views on the challenges and research opportunities resulting from the emergence of big data and analytics, and present an overview of the papers in this special issue. Finally, we offer our concluding remarks and a look to the future.

Data Quality

Big data is the key ingredient in the estimation of many analytical models and the well-known GIGO (garbage in, garbage out) principle continues to apply. In the big data era, the requirement that data must be accurate remains a critical mantra. Data quality has been defined as fitness for use (Pipino et al. 2002), suggesting the relative nature of the concept. That is, the quality of data is evaluated differently when working in different problem domains such as sales forecasting, risk management, or medical diagnostics. We note that data will never be of perfect quality, but any acceptable margin of error must be clearly identified, including the impacts of accepted error levels on the transformational impact of the resulting analytical models. Furthermore, data quality is a multidimensional concept. There exists a large body of literature studying data accuracy, data completeness, data latency, data security, data interpretability, and data traceability (Moges 2014).

In prior research, Moges et al. (2013) surveyed more than 50 financial institutions about their investments in data quality in a credit risk-modeling context. One of the surprising findings was that these firms primarily invested in data quality not because of competitive advantage or belief in its added value, but because of the regulatory pressure as imposed by the Basel II/III guidelines for credit risk modeling (www.bis.org). This poses a quandary given the potential impact and economic return of high quality data in today's digitalized world. The importance of data quality has increased, not declined. Far too often, companies (and universities) seem to consider investments in data quality as too expensive and/or too difficult, leading to a failure to identify and pursue potential long-term gains and benefits.

To emphasize the problem, we refer to the many studies which have been conducted to benchmark state-of-the-art analytical techniques (such as regression, decision trees, neural networks, support vector machines, and ensemble methods) across diverse settings (Baesens et al. 2003; Lessmann et al. 2015). A common finding has been that many analytical techniques yield similar predictive performances and the best way to boost the performance of an analytical model in terms of discrimination or calibration is by working on its key ingredient, data (Baesens 2014). It should be noted that a small improvement in analytical performance can result in substantial profit gains or cost savings.

There also is the issue of trust in data and in analytics as a key factor in implementing decision-linked inferences from data. Even with the highest quality data, a firm will fail to enhance its decision making if the firm's leaders do not trust and choose not to rely upon either the data or the relevant analytic techniques. As more and more firms embed analytics, the trust factor will be enhanced *if* the models we build enhance decision making and add value. Quite simply, performance builds trust. But underlying any progress is the absolute requirement of high quality data that will facilitate (but, of course, cannot guarantee) the development of value-added predictive models.

The big data environment does provide an important opportunity for data validation through triangulation. Researchers have analyzed ways to combine data from various sources and to compare and contrast consistent elements of the differing data sets as a means of data validation (Bapna et al. 2006; Hoffman et al. 1987, 1990). But the big data environment expands the playing field along with the ability to consider

data quality through consistency analyses. We term these analyses *data triangulation* since we focus on three broad data source types: (1) direct market observations, (2) indirect empirical data, and (3) focused survey data.

Consider firm A, which can directly observe all of its individual market transactions. Firm A also, through association agreements and/or data purchase agreements, has access to (indirect) empirical data. We use the term *indirect* to signal that the firm is not directly capturing the data and has no direct control over the quality of the data. Finally, through its own website or possibly a third party contractor, firm A can acquire access to focused survey data. The triangulation validation process centers on using the characteristics of available transaction data to set the conditions for selecting relevant and appropriate indirect empirical data and focused survey data. The question involves data correspondence or consistency over the same range of values. The presence of triangular consistency or cross-validation then provides support for employing indirect empirical data and/or structured survey data to perform model evaluations over ranges where the firm does not have its own direct market observations.

One truth permeates all attempts at estimation and inference: bad data is bad data. Despite widespread homage to the truth of this assertion, few companies have adopted a data czar, a chief data officer, charged with ensuring high quality data throughout the firm. Quality data requires careful procedures and oversight through all steps: (1) initial collection, (2) storage and updating, (3) retrieval, and (4) processing and preparation for analysis. We cannot overstate the importance of corporate governance and management oversight in this respect. Investing in quality data is a long term, demanding process with high accompanying costs, the investment will be unlikely unless senior management trusts and understands the importance of quality data to the development of high value-added analytic models.

Methodological Paradigms and Challenges of Big Data

Methodologically speaking, big data allows us to leverage both prediction and causal analysis. State-of-the-art big data research and practice draws on a variety of techniques from machine learning, classical statistics, and econometrics, to design of experiments (industry calls this A/B or multivariate testing) to test existing theories and hypotheses, developing new theories, and creating large-scale business value. From the perspective of scientific inquiry, the entire globally connected networked economy now can be envisaged as a large-scale real-world laboratory where researchers can design and conduct experiments and collect the data needed to obtain answers to a variety of questions including (1) effects of peer influence, (2) impacts of the influence of dynamic ties, (3) impacts of anonymity on online relationships, (4) results from alternative pricing strategies for digital media, (5) the sway of carefully designed next-generation recommender systems, and (6) the changing preference structures of Generation Y and Z consumers. Analyzing these issues certainly present methodological challenges, but we view the challenges as opportunities for talented researchers.

The big data methods mentioned above have also raised controversies. Recent revelations from two popular websites, Facebook and OkCupid, both of which experimented with their users, have sparked interest in the question of what, if any, appropriate applications exist for use of the online social space as a laboratory for furthering our understanding of human behavior (see, for instance, Goel 2014). The key question is whether such experimentation by companies and/or by academics in collaboration with companies provides benefits to society at large, and, if so, what moral scrutiny is appropriate and implementable? We argue that there is a strong case in favor of such experimentation, not just to avoid costly bad decisions, but also for pursuit of better understanding of what drives human social interactions. Bad decisions not only hurt the bottom line of companies but they cost society since bad decisions result in a misallocation of critical resources. Ever since Pierre Louis' first ever randomized trial dispelled the notion that bloodletting was a helpful medical practice, carefully designed experimentation has driven scientific analysis and knowledge discovery. Now, with the availability of large-scale experimentation yielding big data, the potential for investigation, testing, and new knowledge development are intriguing.

Beyond enhanced decision making in commercial settings, we argue great societal benefit from such experimentation. We now have unprecedented digitization of social processes. While interacting with friends and searching for romantic partners date back to the offspring of Adam and Eve, the fact is that a significant fraction of this activity is now conducted online (a recent PNAS study indicates that 33% of recent marriages in the United States start from an online dating site) in heavily engineered digital platforms with features that can be manipulated. The vast amount of micro-level big data about human interactions offers opportunities never available in the physical world due to cost or infeasibility of comparable data collection. We now have opportunities to generate new causal insights, to test the efficacy of age-old rules-of-thumb and norms that govern social interactions. We can rigorously test existing theories and build and test new

theories (Bapna et al. 2015; Bapna and Umyarov 2015). In short, we can go where social scientists could only dream of going in the past.

This brings us to the need to carefully address the issue of whether such large-scale field experiments would cause more societal harm than good. This net consideration of societal benefit is certainly of key interest for our respective universities' ethics and human subjects committees (also known as institutional review boards). Companies need to take a leaf out of, or perhaps collaborate with, academia and set up similar ethics committees dealing with human subjects. If, as a society, we believe in curiosity about us as a species, then we need to learn from what experimenting on the online social graph tells us. As always, we need our institutions to catch up to technology.

Innovative Big Data Applications

The following discussion illustrates what we view as innovative uses of big data. The list is hardly exhaustive, but we hope that it illustrates noteworthy applications.

Online-to-Offline Commerce

As an accompanying phenomenon of big data, online-to-offline (O2O) commerce has become a booming business trend in China (Lau et al. 2015). O2O is more than "clicks and mortar" because the latter simply recognizes that a company can have both an online component and an offline component. However, the former focuses on how to create a closed loop for customers in many aspects of online and offline activities (Bernstein et al. 2008). O2O differs from B2B, B2C, C2B, and C2C because the latter do not necessarily include an offline component (Xiao and Dong 2015). That is, O2O is a new generation of eCommerce. Further, O2O typically requires a major change of business model in order to fuse online and offline components successfully. O2O is an accompanying phenomenon of big data because it generates and uses an enormous amount of data via its location-based services, mobile computing activities, and Internet-of-things techniques.

Consider QB House, a Japanese barbershop with close to 600 shops in Japan, Hong Kong, Singapore, and Taiwan. Sensors in each of its barber chairs detect available seats and, for each haircut, collect and transmit the total processing time. The information is stored and integrated with the firm's online appointment system along with employee and location performance analysis and resource reallocation modeling.

Networks of Smart Vehicles

As an example of Internet-of-things, smart vehicles, or Internet-of-cars, are equipped with an on-board diagnostics (OBD) device so that the behaviors of drivers and car components can be monitored in real time (Stankovic 2014). The recorded data is then used by vehicle owners, drivers, vehicle managing companies, and insurance companies to make better decisions using analytics and value-driven data modeling. Insurance companies could, for example, employ the models to encourage better driving behavior through rate incentives.

The more interesting impact of this Internet-of-cars big data involves the formation of new business models in vehicle management and traffic management. Results of such applications of big data should include better traffic flow, fairer insurance premiums, and better fleet management.

As a derivative of Internet-of-cars, usage-based insurance (UBI) is leading to new business models for insurance companies. While usage-based insurance is not a new concept, practical and large-scale implementation of UBI are now becoming profitable with big data. Companies are experimenting with UBI to combat insurance fraud and poorly managed insurance schemes.

As more and more vehicles are equipped with OBD devices, insurance companies are working to determine profitable models that allow them to attract safe drivers and enhance profit. Another strategy involves the creation of new (and perhaps very profitable) alliances of companies that include insurance companies (provide insurance), fleet-owners (set operational rules for fleet), and banks (finance the fleets). Given the interaction effects of decisions made in each business area, individual decision making is not likely to successfully rival the ability of consortiums or alliances in structuring overall profit maximization strategies.

Proactive Customer Care

Location-based services offered by mobile communication companies offer travelers detailed information on where to buy things as the need arises. Companies can use location-based services to offer eCoupons to attract people to high value products or into their (or their advertisers') stores and enhance overall profits. Location-based services using real-time big data and analytics enable companies to understand how their customers' needs and interests change as the customers move locations.

Location-based services have become a major business as indicated by the formation of the Location Based Marketing Association, an international group consisting of retailers, agencies, advertisers, media buyers, software and services providers, and wireless companies. Under location-based services, businesses become more proactive toward customers and markets so that customers are serviced where and when they might need services. The nature of the business model broadens the scope and depth of big data analytics needed for businesses to succeed.

Research in proactive customer care will be disruptive since the conventional concept of "location is king" in sales is likely to change. The "line of sight" rule in store location selection will fade as location-based services and mobile map services will direct customers. Location-based services plus real time relevant eCoupons will further help in directing customer flows to stores that excel at location-based service delivery. The shifting business model provides intriguing research questions that challenge conventional business wisdom.

In the next section, we touch on the disruptive aspects of big data.

Disruptive Impacts of Big Data

The various changes for businesses due to or enabled by big data are causing or will cause dramatic shifts in various professions. Certain existing job categories will become obsolete while other job categories emerge and gain strength. Below, we discuss a few examples of disruptive impacts of big data.

Business Analysts Retooling

Since new business models rely on the understanding and usage of big data and analytics, business analysts must be trained in big data capture, big data analysis, big data modeling, and big data based decision-making (Waller and Fawcett 2013). Even business analysts who graduated a few years ago are unlikely to have had the appropriate training in big data and analytics. This poses a major challenge for businesses and existing employees. But developing and retooling of business analysts is inevitable for companies that want to get on the big data wagon and draw on the value-added potential. Many universities are not yet ready to offer mature programs and classes on big data and analytics; even worse, unfortunately, is that many university professors are not ready themselves to deliver effective big data education.

Perhaps the most appropriate approach for retooling business analysts is to create alliances among universities and companies to develop big data academic and professional education programs. In the process, progressive departments in universities can develop or play new roles in educational territories much like the early days of information systems (IS) when the failure of computer science departments to adapt and occupy the IS space led to the rapid development and growth of IS departments and programs in business schools. The pattern in European universities was different, with the IS programs frequently emerging in engineering colleges.

Integration of Data and Social Sciences

The availability of micro-level behavioral data creates collaboration opportunities for company researchers, social scientists, and data scientists. Big data makes innovative projects possible and opens opportunities for investigations that can yield deeper insights into and understanding of human motivation, consumer choices, social phenomena, and the micro-level impact of business activities (Lin 2015). For instance, many psychological theories are simple models with merely two dimensions such as the regulatory focus theory (promotion focus versus prevention focus) or the elaboration likelihood framework (strong elaboration versus weak elaboration) that have been used in information systems research in recent years. These simple theoretical models are largely based on observations and abstractions from small data. Careful analysis of micro-level big data offers the opportunity for social scientists to develop more complex (and realistic) psychological models with useful implications of behavior. If success is demonstrated, this will likely lead to a major disruption to social sciences and for social scientists, who will need to upgrade their toolsets and mindsets.

Breakdown of Traditional Business Boundaries

The opportunities provided to companies through big data analytics have begun to break down a variety of traditional business boundaries. eCommerce giants such as Alibaba and Tencent have moved into banking and now offer savings funds and online insurance services, causing uncertainties and major disruptions to China's banks and multinational banks active in that country (*Financial Times*, January 5, 2015). The eCommerce companies possess new business features that are not available to traditional banks, including existing relationships with hundreds of millions of eCommerce customers. With their scale, the eCommerce companies were able to create a savings fund offering higher interest rates than

banks. In its first year of inception, Alibaba's Yu'ErBao (meaning leftover treasure) fund became a worldwide top-ten fund (http://www.wsj.com) and the number one fund in China (Yu and Shen 2015). This fund allows investors to deposit any amount of money, update gains in real time, and freely transfer between the fund and other accounts. The level of this unprecedented financial service stormed the financial industry, leading banks to loudly cry out claiming they were poor victims, a claim that emphasizes the disruptive change that big data can bring, turning dominant players into entities struggling to catch-up in the era of Internet finance, demonstrating the economics of big data (Yan et al. 2015).

Part of the rapid rise and expansion of eCommerce companies into banking can be linked to their much lower operating margins than traditional banks. The eCommerce banking operations do not need local branches, they simply use their eCommerce network computing power. The key point of this disruption is not that banks cannot fight back but rather that banks never had to face competition from companies outside the traditional banking industry. This disruption rapidly emerged and seems to have caught banks by surprise. The takeaway from this example is the importance of rapidly adapting to the value-added potential by having the tools to understand and utilize big data. Big data and analytics disrupt. If existing or new competitors move faster and more effectively into the big data arena, those lagging behind will face a very desolate landscape.

In the next section, we examine some of the key challenges facing those seeking to successfully leverage big data and analytics and highlight related research opportunities.

Challenges and Research Opportunities of Big Data and Analytics

Leveraging big data and analytics comes with a host of challenges, many of which are fertile ground for future research. Big data is more complex to manage than customary corporate data. In the past, companies mainly managed well-structured data. However, companies now need to manage large amounts of internal and external data that frequently will be unstructured or loosely structured. Adding in appropriate and opportunistic survey data from any of the almost ubiquitous survey options, firms now can face a massive array of internal, external, and survey data.

Ubiquitous Informing

Today, individuals and businesses record what they find interesting, store this information for themselves or others, and share the data for personal and/or business purposes. For simplicity, we refer to this phenomenon as *ubiquitous informing* since the ultimate goal is to inform someone about something ranging from a colorful dish, someone's pulse, video captures, social interactions, vehicle operations, and street surveillance—virtually any bit of information. Ubiquitous informing is possible because of technological advances in mobile computing, video streaming, social networking, smart vehicles, and the Internet-of-things.

Information grabbing and exchanging are indeed easy; the challenge is to find value in taking the time to complete such tasks. The issue is whether one can use big data and data analytics to obtain real value. Can a company get to know its customers better? Can the company identify its most valuable clients and enhance profits through providing these customers with more personalized customer relationships or better customer services? Can the company leverage information to gain stronger market position? The challenges from ubiquitous informing include developing even better big data engineering and analytics to manage and leverage big data to deliver business value.

Implementation Environments

From a pure storage perspective, stack environments such as Hadoop have been introduced to manage big data cubes. As with any new technology, such developing environments should be approached with clear vision and sound criticism. As Winston Churchill noted, however beautiful the strategy, one should occasionally look at the results. Often times, stack environments remain heavily underutilized with the risk of introducing yet another costly legacy burden and jeopardizing future big data driven innovations. More affordable in-memory setups could present an interesting alternative in the journey towards the mature big data organization.

Integration Issues

A key big data challenge for any firm is to identify and decipher the interrelationships of the big data cube and draw out the value implications. By linking the various data streams using appropriately defined unique identifiers, it may become possible to get a more complete picture of customer behavior. In an insurance context, a network of linked entities such as claims, claimants, policyholders, cars, onboard diagnostics devices, car repair shops, credit cards, and mobile numbers might be constructed to unveil a unique perspective on complex collusion practices or fraud patterns.

Value Assessment

The disruptive impact and innovative applications of big data cause serious challenges for the analytical techniques and models that will be built. These models primarily originate from classical statistics, econometrics, machine learning, or artificial intelligence. A key characteristic of all of these analytical techniques is that they focus on optimizing a specific accuracy criterion or a statistics-based objective function (e.g., minimizing a mean-squared error or maximizing likelihood). Typically, performance is summarized using corresponding statistical measures that can be difficult to understand for end-users or nonexperts. As analytical models gain more and more influence in the strategic decisions of a firm, it is important to bridge this communication gap in order to generate the necessary trust. Specifically, to gain trust in an analytical model, both data scientists and decision makers should adopt a *lingua franca* in which the concept of value plays a key role. This brings about a whole new perspective on the construction, performance, and evaluation of an analytical model. In other words, besides pure statistical performance (e.g., measured using misclassification rates, mean squared errors, gini curves, top decile), real value-based criteria become the dominant factors. The appropriate criteria are highly dependent upon the specific business setting.

Analytic models should be understandable to decision makers. Obviously, this has a subjective element to it and depends on both the representation and formal complexity of the analytical model as well as the education or background of the end user. Black box analytical models based on highly complex mathematical formulas are unlikely to be trusted to support key strategic business processes such as credit risk measurement, fraud detection, or even medical diagnosis. Yet, if a black-box method delivers accurate medical diagnoses time after time, would the patient-centric doctor opt for less accurate but easily explainable alternatives? (*Authors' side note: By a unanimous vote, the authors indicate that if we were patients, we would want the doctor to use the black box!*)

Another value-based performance criterion concerns operational efficiency including model evaluation, model monitoring, and model updating. The first of these refers to the resources that are needed to gather the necessary data inputs, preprocess them, run them through the model, and act upon the obtained output. In real-time decision-making settings, fast model evaluation is a key requirement. Consider the example of credit card fraud detection, where typically a decision needs to be made in less than five seconds after the transaction is initiated. Recommender systems are another example where any user action or event (for example, recorded using location-based services) might trigger new recommendations. In addition to model evaluation efforts, operational efficiency also entails the resources needed to monitor, backtest, and, where relevant, stress test the analytical models. Finally, the model must be refreshed or updated as new data evolves or business conditions change. The company that can identify changing conditions and quickly adapt its models is the company that will succeed.

Regulatory Compliance

Since both the strategic and societal impact of analytics is now bigger than ever, we see more and more regulatory guidelines being introduced by authorities at different levels to encourage regulation-compliant analytical models to be built. Consider the Basel III Capital Requirements Accord introduced by the Bank of International Settlements for credit risk modeling as an example. Essentially, this accord unambiguously stipulates the various inputs and outputs which lenders must use in building and evaluating their analytical credit risk models. Unfortunately, many of these analytical regulatory guidelines are too geographically fragmented and/or subsequently overruled, often impeding the creation of a world-wide analytically level playing field.

On the other hand, analytic techniques also offer new ways to enhance compliance checking. Process analytics enables the analysis of huge event logs from process-aware information systems and the contrasting of the observed process with the carefully designed process. These techniques can successfully be used in comprehensive compliance checking and risk management (Caron et al. 2013).

Managing Analytic Decisions

Drawing out decision-ready inferences from big data analytics influences and enhances the firm's decision making. With analytics, we will see more decisions being automated, thus impacting the decision processes and responsibilities throughout the organization. With decisions being automated (potentially based on big data analytics), managing and modeling business decisions is an emerging challenge.

Organizations are heavily involved in optimizing their business processes and enabling quick and effective reaction to new challenges, opportunities, or regulations. By explicitly modeling decisions and the logic behind them, decisions can be managed separately from the processes, dramatically increasing business agility. This requires comprehensible decision analysis techniques for business (Baesens et al. 2003) as well as methods and standards to describe, model,

and manage business decision making. The decision model and notation is such a standard for decision modeling, adopted by the Object Management Group, to bridge the gap between business process design and business decisions, enabling intelligent BPM (Taylor et al. 2013).

Return on Investment and Trust

Finally, an analytical model should add economic value by either generating profits or cutting costs or both. A profit-driven evaluation of an analytical model is key to generate trust across various levels and business units in any organization. Managerial decisions are typically based on economic return, rather than a statistically significant analytical model, and that is where analytics can generate mistrust and thus fall short. It is our firm belief that this should be catalyzed by more research in at least two areas. First, innovative approaches should be developed to carefully quantify the return on investment of an analytical model taking into account the total cost of model ownership, including indirect and opportunity costs, and covering a sufficiently long and appropriate time horizon. As a next topic on the research agenda, the resulting economic insights and measures should be directly embedded into an analytical model building process, rather than merely used as *ex post* evaluation measures. Specifically, analytical models should no longer blindly focus on optimizing a likelihood function or minimizing a misclassification rate, but aim at adding business value where it matters, taking into account all the aforementioned criteria. Only then will the necessary trust be obtained across all decision levels and business units in a firm.

Overview of Papers in the Special Issue

Not unlike the 5 Vs defining big data, the creation of this special issue showed volume, velocity, variety, veracity, and value. The editors had a challenging job in selecting the most valuable papers. The number of submissions and the variety of application fields has been striking, but we believe that the very variety of submissions helped lead to a final set of papers that readers will enjoy and find valuable.

We received 82 first submissions in total. Each of these was assigned to an associate editor who invited two or three reviewers to evaluate the paper. Based on the reports received, 16 papers were invited to resubmit a revised version. The senior editors handled the second round of reviews, with each revised paper evaluated by two senior editors. This resulted into 11 finalist papers whose authors were invited to a workshop that took place in Leuven (Belgium) on August 13–14, 2015. The workshop was meant to be constructive rather than competitive. Each paper was presented by one or more authors in 30 minutes. This was followed by a 30 minute question and answer session with the senior editors and other authors asking questions. The informal feedback we received was positive with participants indicating that the discussion at the workshop contributed to the final quality of the final version of their paper. After the workshop was finished, the editors met and decided to accept all 11 papers subject to minor revisions. In what follows, we give a brief paper by paper overview of the special issue.

The paper by Yahav et al. introduces a data-driven tree-based analytic method for assessing interventions in the presence of self-selection bias. The method offers an automated, stand-alone, computationally efficient alternative to traditional propensity score matching. Intending to provide useful analyses for researchers and decision makers, the paper presents easily understandable results and highlights unbalanced variables even in big data. The approach allows assessment of interventions that are difficult to theoretically specify *a priori*, is capable of identifying heterogeneous effects, and avoids data dredging. Multiple real-world cases and a simulation illustrate the usefulness of this new method in studying firm and government interventions from "standard data" to big data scenarios.

The key contribution of the paper by Zhang et al. is an innovative framework for personalized social brand advertising using network methods combined with sentiment and textual analyses. The authors extract implicit brand–brand networks from large-scale social media data on users' interaction with brands and investigate properties of these networks. Since the sizes of their networks are very large, they develop and implement scalable MapReduce-based algorithms for network construction and processing using a Hadoop environment. They empirically evaluate their audience targeting method using a large dataset collected from Facebook.

The paper by Martens et al. provides an in-depth study of the use of a particular type of big data—massive, fine-grained data on consumer behavior—to improve targeted marketing. Specifically, the paper examines how fine-grained payment data can be used (data on which customers made payments to which merchants) to predict which customers are likely to be interested in a given financial product. The results on a real-life dataset with 21 million transactions reveal that there is indeed much predictive value in such fine-grained behavior data. Furthermore, it is shown that bigger data actually lead to better predictive results, suggesting that larger banks may have substantially more valuable data assets than smaller banks.

The paper by Ghose and Todri addresses the challenging problem of digital channel attribution and measures the causal effectiveness of display advertising. The authors' identification strategy leverages exogenous shocks to the firms' targeting mechanisms, including the viewability of advertisements and granular-level instrumental variables. The authors empirically demonstrate that mere exposure to display advertising can increase users' propensity to search for the brand and the corresponding product as well as users' propensity to make a purchase. That is, consumers engage both in active search, exerting effort to gather information, and in passive search, through information sources that arrive exogenously.

Saboo et al. suggest that instead of relying on rules of thumb or historical performance for allocating resources, firms should recognize that relationships between variables (e.g., influence of advertising on sales) evolve over time and explicitly incorporate these changes in their resource allocation decisions. Although firms collect volumes of data, existing estimation approaches do not readily lend themselves to modeling such temporal variations and provide little guidance to managers for such decisions. The authors propose a time-varying effects model (TVEM) that relies on nonparametric assumptions, and hence is ideal for the big data context, to account for and recover the temporal variations in relationships between variables and to enable firms to adjust decisions on a real-time basis. The proposed approach is easy to implement within an existing organizational infrastructure and provides novel insights, helping firms in analyzing and exploiting big data to increase firm value.

Selecting the relevant part of the data that should be used as input in the modeling process is an initial and crucial step in using observational big data for predictive purposes. In their paper, Brynjolfsson et al. focus on this challenging phase in the context of search trend data selection, a well-known application of big data. They suggest a novel, structured, crowd-based method to address the data selection problem and empirically test its effectiveness in two different domains and relative to various benchmarks. Their results demonstrate the usefulness of the approach for selecting search trend data and emphasize the importance of using a structured data selection method in the prediction process.

Scalability and privacy form two critical dimensions that will eventually determine the extent of the success of big data analytics. In the context of sharing transactional data, sensitive information is typically based on relationships derived from frequently occurring item sets. Menon and Sarkar develop optimal and scalable heuristic procedures leveraging intuition from linear programming based column generation to maximize the accuracy of shared databases, while hiding all sensitive item sets. The authors go on to identify an underlying hierarchical structure to the problem: the column generation based approach is particularly efficient when this structure is exploited, with optimal or near-optimal solutions found quickly. The data qualities of the sanitized databases and the recommendation qualities from using such databases outperform use of the original data.

Han et al. develop a utility theory-based structural model for mobile app analytics. They use the theoretical concepts of utility and satiation along with a factor analytic approach in simultaneously modeling the complex relationships among choice, consumption, and utility maximization for consumers of various mobile apps. Using a unique panel dataset detailing individual user-level mobile app time consumption, the authors quantify the baseline utility and satiation levels of diverse mobile apps and delineate how app preferences and consumption patterns vary across demographic groups and are affected by persistent use and time trends. Their modeling approaches and computational methods can unlock new perspectives and opportunities for handling large-scale, micro-level data, while serving as important resources for big data and mobile app analytics.

Breuker et al. design a predictive modeling technique for business process event data that facilitates big data use cases in operational business processes such as early warning systems or anomaly detection. Through the novel application of techniques originating from grammatical inference and process mining, the approach taps into large-scale enterprise systems' data and predicts likely future behavior of currently running business process instances based on event data of past business process executions. The probabilistic model generated by the technique is visualized using a process modeling notation, so that end-users and nonexperts will likely gain trust in the models because they can better understand and inspect the results and can thus decide if the model structure reflects or contradicts their domain knowledge. The authors created prototypical software implementations for their technique's predictive modeling segment, for visualization, and for model analysis. A series of experiments is conducted using synthetic and real-world data to demonstrate the technique's effectiveness and identify for which circumstances the approach outperforms various benchmarks.

The advent of the digital economy is creating a business environment that is characterized by the unprecedented complexity of technology and connectedness between firms and people. With the goal of reducing the difficulty to understand and depict the business landscape, the paper by Shi et al. develops a big data and analytics framework for quantifying firms' positions in the spaces of product, market, and technology and for measuring firms' dyadic business

proximity using the statistical learning technique of topic modeling. The analytic approach is validated in the context of industry intelligence by constructing a network of U.S. high-tech companies, and the proximity measure's effectiveness is demonstrated in an mergers and acquisitions analysis using random graph models. To put the research into action, the authors built a system prototype that provides a convenient search engine for entrepreneurs, venture capitalists, and analysts to navigate the constantly changing landscape of the networked business environment.

Ketter et al. characterize the difficulties that wicked problems (Rittel and Webber 1973) of societal scale pose to IS researchers. They contend that several obstacles limit the ability of current research methods to tackle such problems. They propose competitive benchmarking (CB) to address these obstacles. CB emphasizes the importance of rich problem representations that are jointly developed among stakeholders and researchers and leads to actionable research results with comprehensive supporting data. CB supports analytical and behavioral IS research (insights) and design science research (solutions). Finally, the authors demonstrate CB in the Power Trading Agent Competition that tackles sustainable energy systems (Power TAC; see Ketter et al. 2013).

Conclusions

Editing a special issue is a combination of joy and grief for authors, reviewers, and associate editors. Editors have (hopefully) limited the grief and enhanced the joy. In this paper accompanying the special issue on Transformational Issues of big data and analytics in Networked Business, we provided our own perspective of the emerging landscape and research opportunities therein. We believe that the special issue represents the current state-of-the-art thinking in the realm of big data and analytics. The issue's contents cover a wide range of topical areas and methods, suggesting that the issue might be useful in Ph.D. seminars seeking to provide a broad and well-rounded perspective on big data.

We would like to end by thanking Paulo Goes, Editor-in-Chief of *MIS Quarterly* for giving us this opportunity, as well as the authors, reviewers, and associate editors, without whose valuable contributions, the special issue would not have seen the light of day.

References

Baesens B. 2014. *Analytics in a Big Data World: The Essential Guide to Data Science and its Applications*, Hoboken, NJ: John Wiley & Sons, Inc.

Baesens, B., Setiono, R., Mues, C., and Vanthienen, J. 2003. "Using Neural Network Rule Extraction and Decision Tables for Credit-Risk Evaluation," *Management Science* (49:3), pp. 312-329.

Baesens, B., Van Gestel, T., Viaene, S., Stepanova, M., Suykens, J., and Vanthienen J. 2003. "Benchmarking State of the Art Classification Algorithms for Credit Scoring," *Journal of the Operational Research Society* (54:6), pp. 627-635.

Bapna, R., Goes, P., Gopal, R., and Marsden, J. R. 2006. "Moving From Data-Constrained to Data-Enabled Research: Experiences and Challenges in Collecting, Validating and Analyzing Large-Scale e-Commerce Data," *Statistical Science* (21:2), pp. 116-130.

Bapna, R., Ramaprasad, J., Shmueli, G., and Umyarov, A. 2015. "One Way Mirrors in Online Dating: Evidence from a Randomized Field Experiment," *Management Science*, forthcoming.

Bapna, R., and Umyarov, A. 2015. "Do Your Online Friends Make You Pay? A Randomized Field Experiment in an Online Music Social Network," *Management Science* (61:8), pp. 1902-1920.

Bernstein, F., Song, J. S., and Zheng X. 2008. "'Bricks-and-Mortar' vs. 'Clicks-and-Mortar': An Equilibrium analysis," *European Journal of Operational Research* (187:3), pp. 671-690.

Breuker, D., Matzner, M., Delfmann, P., and Becker, J. 2016. "Comprehensible Predictive Models for Business Processes," *MIS Quarterly* (40:4), pp. 1009-1034.

Brynjolfsson, E., Geva, T., and Reichman, S. 2016. "Crowd-Squared: Amplifying the Predictive Power of Search Trend Data," *MIS Quarterly* (40:4), pp. 941-961.

Caron, F., Vanthienen, J., and Baesens, B. 2013. "Comprehensive Rule-Based Compliance Checking and Risk Management with Process Mining Decision Support Systems," *Decision Support Systems* (54:3), pp. 1357-1369.

Ghose, A., and Todri, V. 2016. "Toward a Digital Attribution Model: Measuring the Impact of Display Advertising on Online Consumer Behavior," *MIS Quarterly* (40:4), pp. 889-910.

Goel, V. 2014. "As Data Overflows Online, Researchers Grapple with Ethics," *New York Times*, Technology Section, August 12 (http://www.nytimes.com/2014/08/13/technology/the-boon-of-online-data-puts-social-science-in-a-quandary.html?).

Goes P. 2014. "Editor's Comments: Big Data and IS Research," *MIS Quarterly* (38:3), pp. 3-8.

Han, S. P., Park, S., and Oh, W. 2016. "Mobile App Analytics: A Multiple Discrete-Continuous Choice Framework," *MIS Quarterly* (40:4), pp. 983-1009.

Hoffman, E.., Marsden, J. R., and Whinston, A. 1987. "Using Different Economic Data Forms," *Journal of Behavioral Economics* (15:4), pp. 67-84.

Hoffman, E., Marsden, J. R., and Whinston, A. 1990. "Laboratory Experiments and Computer Simulation: An Introduction to the Use of Experimental and Process Model Data in Economic Analysis," *Advances in Behavioral Economics* (2), pp. 1-30.

Ketter, W., Collins, J., and Reddy, P. 2013. "Power TAC: A Competitive Economic Simulation of the Smart Grid," *Energy Economics* (39), pp. 262-270.

Ketter, W., Peters, M., Collins, J., and Gupta, A. 2016. "Competitive Benchmarking: An IS Research Approach to Address Wicked Problems with Big Data and Analytics," *MIS Quarterly* (40:4), pp. 1057-1080.

Lau, A., Chi, J., Gong, F., Li, L., and Liao N. 2015. "China's iConsumer 2015: A Growing Appetite for Choice and Change," McKinsey & Company (http://www.mckinseychina.com/chinas-iconsumer-2015-a-growing-appetite-for-change/).

Lessmann, S., Baesens, B., Seow, H. V., and Thomas, L. C. 2015. "Benchmarking State-of-the-Art Classification Algorithms for Credit Scoring: An Update of Research," *European Journal of Operational Research* (247:1), pp. 124-136.

Lin J. 2015. "On Building Better Mousetraps and Understanding the Human Condition Reflections on Big Data in the Social Sciences," *The Annals of the American Academy of Political and Social Science* (659:1), pp. 33-47.

Martens, D., Provost, F., Clark, J., and Junqué de Fortuny, E. 2016. "Mining Massive Fine-Grained Behavior Data to Improve Predictive Analytics," *MIS Quarterly* (40:4), pp. 869-888.

McKinsey & Company. 2011. "Big Data: The Next Frontier for Innovation, Competition, and Productivity," McKinsey Global Institute (http://www.mckinsey.com/business-functions/business-technology/our-insights/big-data-the-next-frontier-for-innovation).

Menon, S., and Sarkar, S. 2016. "Privacy and Big Data: Scalable Approaches to Sanitize Large Transactional Databases for Sharing," *MIS Quarterly* (40:4), pp. 963-981.

Moges, H. 2014. "A Contextual Data Quality Analysis for Credit Risk Management in Financial Institutions," unpublished Ph.D. thesis, Katholieke Universiteit Leuven.

Moges, H. T., Dejaeger, K., Lemahieu, W., and Baesens, B. 2013. "A Multidimensional Analysis of Data Quality for Credit Risk Management: New Insights and Challenges," *Information and Management* (50:1), pp. 43-58.

Pipino, L. L., Lee, Y. W., and Wang, R. Y. 2002. "Data Quality Assessment," *Communications of the ACM* (45:4), pp. 211-218.

Press, G. 2013. "A Very Short History of Big Data," *Forbes*, Technology Section, May 9 (http://www.forbes.com/sites/gilpress/2013/05/09/a-very-short-history-of-big-data/#30b80ddf55da).

Rittel, H. W., and Webber, M. M. 1973. "Dilemmas in a General Theory of Planning," *Policy Sciences* (4:2), pp. 155-169.

Saboo, A. R., Kumar, V., and Park, I. 2016. "Using Big Data to Model Time-Varying Effects for Marketing Resource (Re)Allocation," *MIS Quarterly* (40:4), pp. 911-939.

Shi, Z., Lee, G. M., and Whinston, A. B. 2016. "Toward a Better Measure of Business Proximity: Topic Modeling for Industry Intelligence," *MIS Quarterly* (40:4), pp. 1035-1056.

Stankovic, J. 2014. "Research Directions for the Internet of Things," *Internet of Things Journal* (1:1), pp. 3-9.

Taylor, J., Fish, A., Vanthienen, J., and Vincent, P. 2013. "Emerging Standards in Decision Modeling: An Introduction to Decision Model & Notation," in *iBPMS: Intelligent BPM Systems: Impact and Opportunity*, Lighthouse Point, FL: Future Strategies Inc., pp. 133-146.

Waller, M. A., and Fawcett, S. E., "Data Science, Predictive Analytics, and Big Data: A Revolution that Will Transform Supply Chain Design and Management," *Journal of Business Logistics* (34:2), pp. 77-84.

Xiao, S., and Dong, M. 2015. "Hidden Semi-Markov Model-Based Reputation Management System for Online to Offline (O2O) e-Commerce Markets," *Decision Support Systems* (77), pp. 87-99.

Yadav, I., Shmueli, G., and Mani, D. 2016. "A Tree-Based Approach for Addressing Self-Selection in Impact Studies with Big Data," *MIS Quarterly* (40:4), pp. 819-848.

Yan, J., Yu, W., and Zhao, J. L. 2015. "How Signaling and Search Costs Affect Information Asymmetry in P2P Lending: The Economics of Big Data," *Financial Innovation* (1:19).

Yu, Y., and Shen, M. 2015. "Consumer Protection as the 'Open Sesame' that Allows Alibaba to Crush the Forty Thieves," *Journal of Antitrust Enforcement*, forthcoming, pp. 228-241.

Zhang, K., Bhattacharyya, S., and Ram, S. 2016. "Large-Scale Network Analysis for Online Social Brand Advertising," *MIS Quarterly* (40:4), pp. 849-868.

Zhao, J. L., Fan, S., and Hu, D. 2014. "Business Challenges and Research Directions of Management Analytics in the Big Data Era," *Journal of Management Analytics* (1:3), pp. 169-174.