10-2016

Big data and data science methods for management research: From the Editors

Gerard GEORGE
Singapore Management University, ggeorge@smu.edu.sg

Ernst C. OSINGA
Singapore Management University, ecosinga@smu.edu.sg

Dovev LAVIE
Technion

Brent A. SCOTT
Michigan State University

DOI: https://doi.org/10.5465/amj.2016.4005

Citation

GEORGE, Gerard; Ernst C. OSINGA; LAVIE, Dovev; and SCOTT, Brent A.. Big data and data science methods for management research: From the Editors. (2016). *Academy of Management Journal*. 59, (5), 1493-1507. Research Collection Lee Kong Chian School Of Business.
Available at: https://ink.library.smu.edu.sg/lkcsb_research/4964

FROM THE EDITORS

BIG DATA AND DATA SCIENCE METHODS FOR MANAGEMENT RESEARCH

Published in Academy of Management Journal, October 2016, 59 (5), pp. 1493-1507.

http://doi.org/10.5465/amj.2016.4005

The recent advent of remote sensing, mobile technologies, novel transaction systems, and high performance computing offers opportunities to understand trends, behaviors, and actions in a manner that has not been previously possible. Researchers can thus leverage 'big data' that are generated from a plurality of sources including mobile transactions, wearable technologies, social media, ambient networks, and business transactions. An earlier *AMJ* editorial explored the potential implications for data science in management research and highlighted questions for management scholarship, and the attendant challenges of data sharing and privacy (George, Haas & Pentland, 2014). This nascent field is evolving rapidly and at a speed that leaves scholars and practitioners alike attempting to make sense of the emergent opportunities that big data holds. With the promise of big data come questions about the analytical value and thus relevance of this data for theory development -- including concerns over the context-specific relevance, its reliability and its validity.

To address this challenge, data science is emerging as an interdisciplinary field that combines statistics, data mining, machine learning, and analytics to understand and explain how we can generate analytical insights and prediction models from structured and unstructured big data. Data science emphasizes the systematic study of the organization, properties, and analysis of data and its role in inference, including our confidence in the inference (Dhar, 2013). Whereas both big data and data science terms are often used interchangeably, big data is about collecting and managing large, varied data while data science develops models that capture, visualize, and analyze the underlying patterns to develop business applications. In this editorial, we address both the collection and handling of big data and the analytical tools provided by data science for management scholars.

At the current time, practitioners suggest that data science applications tackle the three core elements of big data: volume, velocity, and variety (McAfee & Brynjolfsson, 2012; Zikopoulos & Eaton, 2011). Volume represents the sheer size of the dataset due to the aggregation of a large number of variables and an even larger set of observations for each variable. Velocity reflects the speed at which these data are collected and analyzed, whether real-time or near real-time from sensors, sales transactions, social media posts and sentiment data for breaking news and social trends. Variety in big data comes from the plurality of structured and unstructured data sources such as text, videos, networks, and graphics among others. The combinations of volume, velocity and variety reveal the complex task of generating knowledge from big data, which often runs into millions of observations, and deriving theoretical contributions from such data. In this editorial, we provide a primer or a "starter kit" for potential data science applications in management research. We do so with a caveat that emerging fields outdate and improve upon methodologies while often supplanting them with new applications. Nevertheless, this primer can guide management scholars who wish to use data science techniques to reach better answers to existing questions or explore completely new research questions.

BIG DATA, DATA SCIENCE, AND MANAGEMENT THEORY

Big data and data science have potential as new tools for developing management theory, but given the differences from existing data collection and analytical techniques to which scholars are socialized in doctoral training it will take more effort and skill in adapting new practices. The current model of management research is *post hoc* analysis, wherein scholars analyze data collected after the temporal occurrence of the event – a manuscript is drafted months or years after the original data are collected. Therefore, velocity or the real-time applications important for management practice is not a critical concern for management scholars in the current research paradigm. However, data volume and data variety hold potential for scholarly research. Particularly, these two elements of data science can be transposed as data scope and data granularity for management research.

Data Scope. Building on the notion of volume, data scope refers to the comprehensiveness of data by which a phenomenon can be examined. Scope could imply a wide range of variables, holistic populations rather than sampling, or numerous observations on each participant. By increasing the number of observations, a higher data scope can shift the analysis from samples to populations. Thus, instead of focusing on sample selection and biases, researchers could potentially collect data on the complete population. Within organizations, many employers collect data on their employees, and more data is being digitized and made accessible. This includes email communication, office entry and exit, RFID tagging, wearable sociometric sensors, web browsers, and phone calls, which enable researchers to tap into large databases on employee behavior on a continuous basis. Researchers have begun to examine the utility and psychometric properties of such data collection methods, which is critical if they are to be incorporated into and tied to existing theories and literatures. For example, Chaffin et al. (in press) examined the feasibility of using wearable sociometric sensors, which use a Bluetooth sensor to measure physical proximity, an infrared detector to assess face-to-face positioning between actors, and a microphone to capture verbal activity, to detect structure within a social network. As another example, researchers have begun to analyze large samples of language (e.g., individuals' posts on social media) as a non-obtrusive way to assess personality (Park et al., 2015). With changes in workplace design, communication patterns, and performance feedback mechanisms, we have called for research on how businesses are harnessing technologies and data to shape employee experience and talent management systems (Colbert, Yee & George, 2016; Gruber, Leon, Thompson & George, 2015).

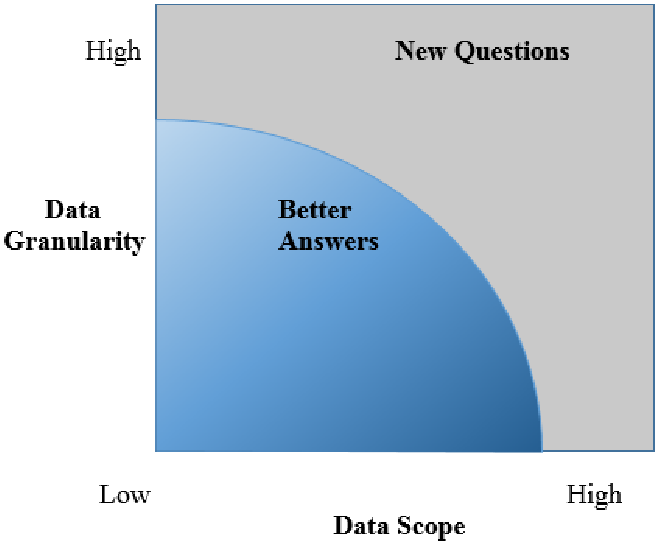
Data Granularity. Following the notion of variety, we refer to data granularity as the most theoretically proximal measurement of a phenomenon or unit of analysis. Granularity implies direct measurement of constituent characteristics of a construct rather than distal inferences from data. For example, in a study of employee stress, granular data would include emotions through facial recognition patterns or biometrics such as elevated heart rates during every minute on the job or task rather than surveys or respondent interviews. In experience-sampling studies on well-being, for example, researchers have begun to incorporate portable blood pressure monitors. For instance, in a 3-week experience-sampling study, Bono, Glomb, Shen, Kim, and Koch (2013) had employees wear ambulatory blood pressure monitors that recorded measurements every 30 minutes for two hours in the morning, afternoon, and evening. Similarly, Ilies, Dimotakis, and DePater (2010) used blood pressure monitors in a field setting to record employees' blood pressure at the end of each workday over a two-week period. Haas, Criscuolo and George (2015) studied message posts and derived meaning in words to predict whether individuals are likely to contribute to problem solving and knowledge sharing across organizational units. Researchers in other areas could also increase granularity in other ways. In network analysis for instance, researchers can monitor communication patterns across employees instead of asking employees with whom they interact or seek advice from retrospectively.

Equivalent data were earlier collected using surveys and indirect observation, but with big data the unit of analysis shifts from individual employees to messages and physical interactions. Though such efforts are already seen in smaller samples of emails or messages posted on a network (e.g., Haas, Criscuolo & George, 2015), organization-wide efforts are likely to provide clearer and holistic representations of networks, communications, friendships, advice-giving and taking, and information flows (van Knippenberg, Dahlander, Haas & George, 2015).

Better Answers and New Questions

Together, data scope and data granularity allow management scholars to develop new questions and new theories, and to potentially generate better answers to established questions. In Figure 1, we portray a stylistic model of how data scope and data granularity could productively inform management research.

Figure 1
Scope and Granularity Implications for Management Theory



Better Answers to Existing Questions. Data science techniques enable researchers to get more immediate and accurate results for testing existing theories. In doing so, we hope to get more accurate estimations of effect sizes and their contingencies. Over the past decade, management theories have begun emphasizing effect sizes. This emphasis on precision is typically observed in strategy research rather than in behavioral studies. With data science techniques, the precision of effect sizes and their confidence intervals will likely be higher and can reveal nuances in moderating effects that have hitherto not been possible to identify or estimate effectively.

Better answers could also come from establishing clearer causal mechanisms. For instance, network studies rely on surveys of informants to assess friendship and advice ties, but in these studies, the temporal dimension is missing, and therefore it is difficult to determine whether network structure drives behavior or vice versa. Instead, collecting email communications or other forms of exchange on a continuous level would enable researchers to measure networks and behavior dynamically, and thus assess more systematically cause and effect.

Although rare event modeling is uncommon in management research, data science techniques could potentially shed more light on, for example, organizational responses to disasters, modeling and estimating probability of failure, at risk behavior, and systemic resilience (van der Vegt, Essens & Wahlstrom, 2015). Research on rare events can use motor car accident data, for instance, to analyze the role of driver experience in seconds leading up to an accident and how previous behaviors could be modeled to predict reaction times and responses. Insurance companies now routinely use such data to price insurance coverage, but this type of data could also be useful for modeling individual-level risk propensity, aggressiveness, or even avoidance behaviors. At an aggregate level, data science approaches such as collecting driver behavior using sensors to gauge actions like speeding and sudden stopping, allow more than observing accidents, and therefore generate a better understanding of their occurrence. Such data allows cities to plan traffic flows, map road rage or accident hot spots, and avert congestion, and researchers to connect such data to timeliness at work, and negative or positive effects of commuting sentiment on workplace behaviors.

Additionally, data science techniques such as monitoring call center calls can enable researchers to identify specific triggers to certain behaviors as opposed to simply measuring those behaviors. This can help better understand phenomena such as employee attrition. Studying misbehavior is problematic due to sensitivity, privacy and availability of data. Yet, banks are now introducing tighter behavioral monitoring and compliance systems that are tracked in real-time to predict and deter misbehaviors. Scholars already examine lawsuits, fraud, and collusion, but by using data science techniques, they can search electronic communication or press data using keywords that characterize misbehavior in order to identify the likelihood of misbehavior before its occurrence. As these techniques become prevalent, it will be important to tie the new measures, and the constructs they purportedly assess, to existing theories and knowledge bases; otherwise, we risk the emergence of separate literatures using “big” and “little” data that have the capacity to inform each other.

New Questions. With higher scope and granularity of data, it becomes possible to explore new questions that have not been examined in the past. This could arise because data science allows us to introduce new constructs, but it could also arise because data science allows us to operationalize existing constructs in a novel way. Web scraping and sentiment data from social media posts are now being seen in the management literature, but they have yet to push scholars to ask new questions. Granular data with high scope could open questions in new areas of mobility and communications, physical space, and collaboration patterns where we could delve deeper into causal mechanisms underlying collaboration and team dynamics, decision-making and the physical environment, workplace design and virtual collaborations. Tracking phone usage and physical proximity cues could provide insight into whether individuals spend too much time on communications technology and attention allocation to social situations at work or at home. Studies suggest that time spent on email increases anger and conflict at work and at home (Butts, Becker & Boswell, 2015). But such work could then be extended to physical and social contingencies, nature of work, work performance outcomes, and their quality of life implications.

Data on customer purchase decisions and social feedback mechanisms can be complemented with digital payments and transaction data to delve deeper into innovation and product adoption as well as behavioral dynamics of specific customer segments. The United Nations' Global Pulse is harnessing data science for humanitarian action. Digital money and transactions through mobile platforms provide a window into social and financial inclusion, such as access to credit, energy and water purchase through phone credits, transfer of money for goods and services, create spending profiles, identify indebtedness or wealth accumulation, and promote entrepreneurship (Dodgson et al., 2015). Data science applications allow the delivery and coordination of public services such as treatment for disease outbreaks, coordination across grassroots agencies for emergency management, and provision of fundamental services such as energy and transport. Data on carbon emissions and mobility can be superimposed for tackling issues of climate change and optimizing transport services or traffic management systems. Such technological advances that promote social wellbeing can also raise new questions for scholars in identifying ways of organizing and ask fundamentally new questions on organization design, social inclusion, and the delivery of services to disenfranchised communities.

New questions could emerge from existing theories. For example, once researchers can observe and analyze email communication or online search data, they can ask questions concerning the processes by which executives make decisions as opposed to studying the individual/TMT characteristics that affect managerial decisions. There is room for using unstructured data such as video and graphic data, and face recognition for emotions. Together, these data could expand conversations beyond roles, experience, and homogeneity to political coalitions, public or corporate sentiment, decision dynamics, message framing, issue selling, negotiations, persuasion, and decision outcomes.

Text mining can be used when seeking to answer questions such as where do ideas or innovations come from -- as opposed to testing whether certain conditions generate ground-breaking innovations. This requires data mining of patent citations that can track the sources of knowledge embedded in a given patent and its relationships with the entire population of patents. In addition, analytics allow inference of meaning, rather than word co-occurrence, which could be helpful in understanding cumulativeness, evolution and emergence of ideas and knowledge.

A new repertoire of capabilities is required for scholars to explore these questions and to handle challenges posed by data scope and granularity. Data are now more easily available from corporates and “Open Data” warehouses such as the London DataStore. These data initiatives encourage citizens to access platforms and develop solutions using big data on public services, mobility and geophysical mapping among others data sources. Hence, as new data sources and analytics become available to researchers, the field of management can evolve by raising questions that have not received attention as a result of data access or analysis constraints.

BIG DATA AND DATA SCIENCE STARTER KIT FOR SCHOLARS

Despite offering exciting new opportunities to management scholars, data science can be intimidating and pose practical challenges. Below, we detail key challenges and provide solutions. We focus on five areas: 1) data collection, 2) data storage, 3) data processing, 4) data analysis, and 5) reporting, where we note that due to automation, the boundaries between these areas become increasingly blurry. Our aim is to provide a data science quick start guide. For detailed information, we refer to relevant references. In Table 1, we present a summary of the data science challenges and solutions.

Insert Table 1 about here

Data Collection

Data science allows collection of data with high scope and granularity. Due to advances in technology, data collection methods are more often limited by the imagination of the researcher than by technological constraints. In fact, one of the key challenges is to think outside the box on how to establish access to detailed data for a large number of observations. Big data collection methods that help to overcome this challenge include sensors, web scraping, and web traffic and communications monitoring.

Using sensors, one can continuously gather large amounts of detailed data. For example, by asking employees to wear activity-tracking wristbands, data can be gathered on their movements, heart rate, and calories burned, and, as mentioned, wearable sensors can be used to collect data on physical proximity (Chaffin et al., in press). Importantly, this data-gathering approach is relatively non-obtrusive and provides information about employees in natural settings. Moreover, this approach allows for the collection of data over a long time period. In contrast, alternative methods for obtaining mentioned data, such as diaries, may influence employees' behavior and it may be difficult to incentivize employees to keep a diary over a long period of time (cf. testing and mortality effects in consumer panels, Aaker, Kumar, Leone & Day, 2013, p. 110). Sensors can also be used to monitor the office environment, e.g. office temperature, humidity, light, and noise, or the movements and gas consumption of vehicles, as illustrated by the monitoring of UPS delivery trucks (UPS, 2016). Similarly, FedEx allows customers to follow not only the location of a package, but also the temperature, humidity, and light exposure (Business Roundtable, 2016). Of course, when using sensors to gather data on human subjects, one should stay within the bounds of privacy laws and research ethics codes. Although this may sound obvious, the ethical implications of collecting such data, coupled with issues such as data ownership, are likely to be complex, and thus call for revisiting formal guidelines for research procedures.

Web scraping allows for the automated extraction of large amounts of data from websites. Web scraping programs are widely available nowadays and often come free of charge, as in the case of plugins for popular web browsers such as Google Chrome. Alternatively, one may use packages that are available for programming languages, e.g., the Beautiful Soup package for the Python programming language.[1] Some websites, e.g., Twitter, offer so-called application program interfaces (APIs) to ease and streamline access to its content.[2] Web scraping can be used to extract numeric data, such as product prices, but it may also be used to extract textual, audio, and video data, or data on the structure of social networks. Examples of textual data that can be scraped from the web include social media content generated by firms and

[1] The beautiful soup package can be obtained from https://www.crummy.com/software/BeautifulSoup/.

[2] The Twitter API can be found on https://dev.twitter.com/rest/public.

individuals, news articles, and product reviews. Importantly, before scraping data from a website, one should carefully check the terms of use of the particular website and the API, if available. Not all websites allow the scraping of data. Lawsuits have been filed against companies that engage in web scraping activities (Reuters, 2013), be it mostly in cases where the scraped data were used commercially. In case of doubt about the legality of scraping a website's data for academic use, it is best to contact the website directly to obtain explicit consent.

Moreover, most firms nowadays monitor traffic on their website, i.e., the data generated by visitors to the firm's webpages. Firms track aggregate metrics such as the total number of visitors and the most visited pages in a given time period (e.g., Wiesel, Pauwels & Arts, 2011). In addition, data for individual visitors are available. Researchers can access data on individual visitors including the page from which they reached the focal firm's website, the pages that they visited and the order at which they were accessed. Additional data include the last page they visited before leaving the website, the time they spent staying on each webpage, the day and time at which they visited the website, and whether they visited the website before. Data at the individual visitor level allow for a detailed analysis of how visitors, such as investors, suppliers, and consumers, navigate through the firm's website and thus to obtain a better understanding of their information needs, their level of interaction with the focal firm, and their path to their transactions with the firm (e.g., Sismeiro & Bucklin, 2004; Xu, Duan & Whinston, 2014). In addition to monitoring incoming traffic to their website, firms may also track outgoing traffic, i.e., the webpages that are visited by their employees. Outgoing web traffic provides information about the external web-based resources used by employees and the amount of time spent browsing the Internet for leisure. Such information may be of use in studies on, for example, product innovation and employee well-being. These data can also shed light on attention allocation, employee engagement, or voluntary turnover.

Data on communications between employees can be gathered as well. For example, phone calls and email traffic are easily monitored. Such data sources may help to uncover intra-firm networks and measure communication flows between organizational divisions, branches, and units at various hierarchy levels. Monitoring external and internal web, phone, and email traffic are non-obtrusive data gathering approaches, i.e., users are not aware that the traffic that they generate is being monitored. With regard to the monitoring of web traffic and other communications, it is important to strictly adhere to privacy laws and research ethics codes that protect employees, consumers, and other agents.

The above methods allow for large-scale, automated, and continuous data collection. Importantly, these features significantly ease the execution of field experiments. In a randomized field experiment, a randomly selected group of agents, e.g. employees, suppliers, customers, etc., is subjected to a different policy than the control group. If the behavior of the agents in both groups is continuously monitored, the effect of the policy can be easily tested (e.g., Lambrecht & Tucker, 2013). Other experimental setups are possible as well. For example, if random selection of agents is difficult, a quasi-experimental approach can be adopted (Aaker et al., 2013, p. 288), where all agents are subjected to a policy change and where temporal variation in the data can be used to assess the effect of this change. In doing so, it is important to rule out alternative explanations for the observed temporal variation in the data. Again, privacy laws and research ethics codes should be strictly obeyed. For example, to protect the privacy of customers, Lewis and Reiley (2014) use a third party to match online browsing and offline purchase data.

Data Storage

Big data typically requires large storage capacity. Often, the required storage capacity exceeds those found in regular desktop and laptop computers. Units of interest are observed at a very granular level and in high detail, by pulling data from various sources. Fortunately, several solutions are available. First, one can tailor the storage approach to the size of the data. Second, one can continuously update and store variables of interest while discarding information that is irrelevant to the study.

How to store data depends on its size. For relatively small datasets, no dedicated storage solutions are required. Excel, SAS, SPSS, and Stata can hold data for many subjects and variables. For example, Excel can hold 1,048,576 rows and 16,384 columns (Microsoft, 2016). Larger datasets require another storage approach. As a first option, a relational database using structured query language (SQL) can be considered. Relational databases store data in multiple tables that can be easily linked with each other. Open source relational databases are available, e.g., MySQL and PostgreSQL. Basic knowledge of SQL is required to use relational databases. For yet larger databases, a NoSQL approach may be required, especially when fast processing of the data is necessary (Madden, 2012). NoSQL stands for not only SQL (e.g., Varian, 2014) to reflect that NoSQL database solutions do support SQL-like syntax. Examples include the open source Apache Cassandra system initially developed by Facebook, and Google's Cloud Bigtable, which offers a pay-as-you-use cloud solution. To effectively store data that is too large to hold on a single computer many firms employ Apache Hadoop, which allows the allocation of data across multiple computers. To build such infrastructure, self-study (see for example Prajapati, 2013) or collaboration with information systems experts is required.

When the research question and data requirements have been clearly defined, it may not be necessary to store all available data. Often it is possible to continuously update variables of interest as new data comes in and then to save only the updated variables and not the new information itself. For example, a sensor may record every single heartbeat of a subject. Typically, we would not be interested in the full series of heartbeat timestamps for each subject but instead would like to know, for instance, a subject's average hourly heart rate. We thus only need to store one hour of heartbeat timestamps for each subject, after which we can create, for each subject, one hourly observation. We can now discard the underlying timestamp data and start recording the next hour of heartbeat timestamps. Particularly, when the number of sensors and number of subjects grows large, this approach will provide an enormous reduction in terms of storage requirements. Also, when interested in consumers' daily behavior on the focal firm's website, storage of consumers' full web traffic data is not required.

In field experiments, situations may occur where group-level instead of subject-level information suffice. For example, one may be interested in the effect of a policy change on whether employees are more likely to adopt a new online self-evaluation tool, where we assume that adoption is a binary variable, i.e., either yes or no. Adoption of the online tool can be measured by monitoring employees' web traffic. It is adequate to know the number of adopters in the control and experimental group and the total number of subjects in each group. Based on these numbers, and the assumption that the dependent variable is binary, we know the exact empirical distribution of adoption. We can thus easily perform significance tests. It must be noted that group-level information does not suffice when adoption would be measured on an interval or ratio scale: to perform significance tests, subject-level data would be required.

Obviously, one can only resort to the second approach, i.e., updating variables of interest when new data becomes available while discarding the new data itself, if the variables of interest are known. Moreover, an important caveat of this approach is that it will not be possible to retrieve the more granular raw data that were discarded to reduce storage requirements.

Data Processing

Variety, one of big data's three Vs, implies that we are likely to encounter new and different types of data that may be non-numeric. An important new source of data is textual data. For example, studies of social media, email conversations, annual report sections, and product reviews require methods for handling textual data (e.g, Archak Ghose & Ipeirotis, 2011; Loughran & McDonald, 2011). Textual data can be used both for theory testing (e.g., we could test the hypothesis that positive firm news makes employees expect a higher annual bonus), and theory development (e.g., we could explore which words in the management section in annual reports are associated with a positive or negative investor response and use these results to develop new theory). Before being able to include non-numeric data in quantitative analyses, these data first need to be processed.

In case of theory testing, we have a clear idea about the information that needs to be extracted from texts. For example, to determine whether news is positive, we could ask independent raters to manually evaluate all news items and to provide a numeric rating. When the number of news items grows substantially, a few independent raters would not be able to complete the task in an acceptable amount of time. One can then scale up the "workforce" by relying on Amazon MTurk, where participants complete tasks for a fixed fee (Archak et al., 2011). Another solution is to automate the rating task. To do so, we typically first remove punctuation marks (Manning, Raghavar & Schűtze, 2009, p. 22) from the news items and convert all letters to lowercase. Words can then be identified as groups of characters separated by spaces. Removal of punctuation marks avoids that commas, semicolons, etc. are viewed as being part of a word. Conversion of characters to lowercase ensures that identical words are treated as such. Manning et al. (2009, p. 30) do note that information is lost by converting words to lowercase, e.g., the distinction between Bush, the former U.S. presidents or British rock band, and bush, a plant or area of land. After these initial steps, a program can be written to determine the number of positive words in each news item, where the collection of positive words needs to be predefined. It is also possible to develop measures based on the number of relevant words that appear within a certain distance from a focal word, such as a firm's name or the name of its CEO.

Several toolboxes are available to help process textual data, e.g., NLTK for Python and tm for R. Corrections for the total number of words in a news item can be made to reflect that a longer text contains more positive words than a shorter text, even when the texts are equally positive (for a discussion on document length normalization, see Manning et al., 2009, p. 129). Also, we can refine the program by making it look at groups of words instead of single words to make it pick up on fragments such as “the financial outlook for firm A is not very positive”, which does not reflect positive news despite the use of the word positive (Das & Chen, 2007).

In theory development, we would be interested in exploring the data. For ease of exposition, we focus on the setting where we want to assess which words in Management Discussion and Analysis (MD&A) sections are associated with a variable of interest. To this end, for each MD&A section, we would first need to indicate the number of times a word occurs. This task is virtually impossible to perform manually. Hence, we would typically automate this process. Again, we would remove punctuation marks and convert all text to lowercase.

Typically, we would also remove non-textual characters such as numbers, as well as small, frequently occurring words such as "and" and "the" as these are non-informative (cf. Manning et al., 2009, p. 27). In addition, words may be stemmed and infrequent words may be removed to reduce the influence of outliers (Tirunillai & Tellis, 2014). As a basic approach, we can determine the set of unique words across all MD&A sections, and, for each MD&A, count the number of times each unique word occurs. The resulting data would look as follows: with each row representing a single MD&A, each column represents a single unique word, and the cells indicate the number of times a word occurs in the particular MD&A. Again, a correction for the length of the MD&A section can be applied. The data would contain many zeroes, i.e., when a particular word is not used in the MD&A. To reduce the storage size of the data, a sparse matrix can be used. To explore which of the many words are associated with the variable of interest, we can make use of the techniques for variable selection that we discuss below.

Other source of non-numeric data include audio, images, and video. Many new and interesting techniques exist for extracting numeric information from such data. For example, image and video data can be used to determine a person's emotions, which may be expressed using numeric scales (Teixeira, Wedel & Pieters, 2012).

Data Analysis

After deciding on the data collection method and having successfully stored and processed the data, a new challenge arises: how to analyze the data? We may encounter one or both of the following scenarios: (1) a (very) large number of potential explanatory variables are available and the aim is to develop theory based on a data-driven selection of the variables; (2) the data are too large to be processed by conventional personal computers. Below, we present solutions to overcome these challenges.

Variable selection. Big data may contain a (very) large number of variables. It is not uncommon to observe hundreds or thousands of variables. In theory testing, the variables of interest, and possible control variables, are known. In exploratory studies where the aim is to develop new theory based on empirical results, we need statistical methods to assist in variable selection. Inclusion of all variables in a model, e.g. a regression model, is typically impeded by high multicollinearity (Hastie, Tibshirani & Friedman, 2009, p. 63). One could try to estimate models using all possible combinations of explanatory variables and then compare model fit using, for example, a likelihood-based criterion. It is easy to see that the number of combinations of the explanatory variables explodes with the number of variables, rendering this approach unfeasible in practice. Methods that can be used in the situation of a large number of (potential) explanatory variables include ridge and lasso regression, principal components regression, partial least squares, Bayesian variable selection, and regression trees (e.g., Varian, 2014).

Ridge and lasso regression are types of penalized regression. Standard regression models are estimated by minimizing the sum of the squared differences between the observed (y_i) and predicted ($\hat{y}_i$) values for observation i, i.e., we use the coefficients that minimize $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ or, substituting $\boldsymbol{x_i\beta}$ for $\hat{y}_i$, $\sum_{i=1}^{n}(y_i - \boldsymbol{x_i\beta})^2$, where $\boldsymbol{x_i}$ is the vector of independent variables for observation i, $\boldsymbol{\beta}$ is the regression coefficient vector, and n indicates the number of observations. In penalized regression, we do not minimize the sum of the squared differences between the observed and predicted values, i.e., the residuals, but instead minimize the sum of the squared residuals plus a penalty term. The difference between ridge and lasso regression lies in the penalty that is added to the squared residuals. The penalty term in ridge regression is $\lambda \sum_{j=1}^{k} \beta_j^2$, where k is the number of slope coefficients, and in lasso regression we use $\lambda \sum_{j=1}^{k} |\beta_j|$ as a penalty term, where $|\beta_j|$ is the absolute of β_j (Hastie et al., 2009, pp. 61-69). The regression coefficients given by ridge regression are thus obtained by minimizing $\sum_{i=1}^{n} (y_i - \boldsymbol{x_i \beta})^2 + \lambda \sum_{j=1}^{k} \beta_j^2$. The expression for lasso regression is obtained by simply replacing the penalty term in this expression. Ridge and lasso regression may also be combined into what is referred to as elastic net regression (Varian, 2014). It must be noted that the intercept, β_0, is not included in the penalty term. To remove the intercept from the model altogether, without affecting the slope coefficients, one can center the dependent and independent variables around their mean. Moreover, researchers typically standardize all variables before estimation, because the coefficient size, and thus also the penalty term, depends on the scaling of the variables (Hastie et al., 2009, p. 63). It is easy to see that by setting λ to zero, one can obtain the least squares estimator. By increasing λ, the penalty term grows in importance, resulting in larger shrinkage of the coefficients until all coefficients are shrunk to zero. We return to the choice of λ below when we discuss cross-validation.

Principal components regression offers another approach to handling a large number of independent variables. In this approach, we first extracts *l* principal components from the independent variables, $\boldsymbol{X}$, and then regress the dependent variable $\boldsymbol{y}$ on the *l* principal components. The principal components, $\boldsymbol{Z}$, are a linear combination of the independent variables, i.e., the *m*th principal component is given by $\boldsymbol{z_m} = \boldsymbol{X\omega_m}$, where $\boldsymbol{\omega_m}$ are the loadings for principal component *m*. Typically, the independent variables are first standardized (Hastie et al., 2009, p. 79). The regression model that is estimated after extraction of the principal components is $\boldsymbol{y} = \theta_0 + \sum_{m=1}^{l} \theta_m \boldsymbol{z_m}$, i.e., the principal components serve as independent variables. Instead of shrinking the coefficients for all variables, as in penalized regression, principal components regression handles the large number of variables by combining correlated variables in one principal component and by discarding variables that do not load onto the first *l* principal components. Below, we return to the choice of *l*, the number of principal components to extract.

Partial least squares provides yet another approach to dealing with a large number of variables. As in principal components regression, the dependent variable is regressed on newly constructed "components" formed from the independent variables. An important difference is that principal components regressions only takes the independent variables into account in constructing these components, whereas partial least squares uses the independent and dependent variables. More specifically, in principal components regression, components are formed based on the correlations between the independent variables, whereas in partial least squares, the level of association between independent and dependent variables serves as input. Several algorithms for implementing partial least squares exist, e.g. NIPALS and SIMPLS. We refer to Frank and Friedman (1993) for a comparison of partial least squares and principal components regression. As in principal components regressions, a decision needs to be made on the number of components to be used in the final solution. Below, we return to this issue.

Variables may also be selected using a Bayesian approach. To this end, George and McCulloch (1993) introduce Bayesian regression with a latent variable, γ_j, that indicates whether variable *j* is selected. This latent variable takes the value one with probability p_j and the value zero with probability $1 - p_j$. The coefficient for variable *j*, β_j, is distributed, conditional on γ_j, as $(1 - \gamma_j)N(0, \tau_j^2) + \gamma_j N(0, c_j^2\tau_j^2)$, where *N* denotes the normal distribution. By setting τ_j to a very small strictly positive number and c_j to a large number, larger than 1, we obtain the following interpretation: if γ_j is zero, the variable is multiplied by a near-zero coefficient β_j (basically not selecting variable $\boldsymbol{x_j}$), if γ_j is one, β_j will likely be non-zero, thus selecting variable $\boldsymbol{x_j}$. For more detail on this method and how to decide on the values for τ_j and c_j, we refer to George and McCullogh (1993).

An interesting method that may assist in variable selection is the regression tree. Regression trees indicate to what extent the average value of the dependent variable differs for those observations for which the values of an independent variable lie above or below a certain value, e.g., the average employee productivity for those employees that are 40 years or older versus those that are younger. Within these two age groups, the same or another variable is used to create subgroups. In every round, that independent variable and that cutoff value are chosen that provide the best fit. The resulting tree, i.e., the overview of selected variables and their cutoff values, indicates which variables are most important in explaining the dependent variable. Also, this method may reveal nonlinear effects of independent variables (Varian, 2014). For example when age is used twice to explain employee productivity, the results may reveal that the most productive employees are younger than 30, or 40 years or older, with those between age 30 and 40 being the least productive. It is important to note that the results from a regression tree, where independent variables are dichotomized, do not necessarily generalize towards a regression method with continuous independent variables. Several interesting techniques for improving the performance of regression trees exist, e.g., bagging, boosting and random forests (Hastie et al., 2009, chapter 15; Varian 2014). Several other methods for variable selection are available, e.g. stepwise and stagewise regression.

When exploring textual data to determine which of the many words in texts are associated with the variable of interest, one can make use of the techniques described above – each word is a single variable. When the data are too sparse and too high-dimensional, more advanced approaches are required. Topic models, such as latent Dirichlet allocation (LDA), may be used to extract latent topics from the MD&A sections (Blei, Ng & Jordan, 2003), thus reducing the dimensionality of the data. The latent topics can be labeled using entropy-based measures (Tirunillai & Tellis, 2014) and then used in subsequent analyses. Newly introduced approaches such as deep learning (LeCun, Bengio & Hinton, 2015) can also help overcome the shortcomings of LDA (Liu, 2015, pp. 169-171).

Tuning variable selection models. We now return to the question of how to choose the λ parameter in ridge and lasso regression, and the number of components in principal components regression and partial least squares. To that aim, researchers can rely on cross-validation (e.g., Hastie et al., 2009, figures 3.8 and 3.10). In its basic form, cross validation requires random allocation of the observations to two mutually exclusive subsets. The first subset, referred to as the training set, is used to estimate the model. The second subset, the validation set, is used to tune the model, i.e., to determine the λ parameter or the number of components. More specifically, one should estimate many models on the training set, each assuming a different λ parameter or number of components. The estimated models can be then used to predict the observations in the validation set. The final step involves choosing the model with the λ parameter or number of components that provides the best fit in the validation set. The model is not tuned on the training set as this would lead to overfitting (Varian, 2014). To assess the predictive validity, i.e., the performance of the model on new data, observations are allocated to three subsets: the aforementioned training and validation set and a third set referred to as the test set. The model is then estimated on the training set, tuned on the validation set, and the predictive validity is assessed by the forecasting performance on the test set. It is important to note that a model that shows a strong relationship between independent variable j and the dependent variable $\boldsymbol{y}$, and that, in addition, shows strong predictive validity, does not demonstrate a causal relationship between $\boldsymbol{x_j}$ and $\boldsymbol{y}$ (Varian, 2014). The association between $\boldsymbol{x_j}$ and $\boldsymbol{y}$ may be due to reverse causality or due to a third variable that drives both $\boldsymbol{x_j}$ and $\boldsymbol{y}$, although the latter explanation requires that the influence of the third variable holds in the training and test set. To test causality, one would ideally rely on a field experiment where $\boldsymbol{x_j}$ is manipulated for a random subset of the observed agents (Lambrecht & Tucker, 2013).

An advantage of working with big data is that typical datasets contain a sufficient number of observations to construct three subsets. In case of a relatively small number of observations, the training set may become too small to reliably estimate the model, and the validation and test set may not be representative of the data, thus giving an incorrect assessment of the model fit (Hastie et al., 2009, p. 241). K-fold cross-validation may be preferred in these situations. K-fold cross-validation partitions the data in K subsets. For each value of the λ parameter or number of components, one should first estimate the model on the combined data from all subsets but the first. The estimated model is then used to predict the observations in the first subset. The same procedure is then repeated for the combined data from all subsets but the second, etc. By computing the average fit across subsets, one can choose the optimal λ parameter or number of components. Common values for K are five and ten (Hastie et al., 2009, p. 242). In case of very small datasets, K may be set to the number of observations minus one, to obtain-leave-one-out cross-validation.

Data too large to analyze. Big data may be too large to analyze, even when storage on a single machine is possible. Typical personal computers and their processors may not be able to complete the commands in a reasonable amount of time and they may not hold sufficient internal memory to handle the analysis of large datasets. Below, we discuss three solutions: parallelization, bags of little bootstrap, and sequential updating. For additional approaches, such as the divide and conquer approach and methods for the application of MCMC to large datasets, we refer to Wang, Chen, Schifano, Wu, and Yan (2015) and Wedel and Kannan (2016).

Parallelization helps to speed up computations by using multiple of the computer's processing units. Nowadays, most computers are equipped with multi-core processors, yet many routines employ only a single processing core. By allocating the task over multiple cores, significant time gains can be obtained, where it must be noted that some time gains are lost due to the costs associated with distributing the tasks. Many statistical programs and programming languages enable parallelization, e.g., Stata's MP version, the parallel package for R, and Matlab's Parallel Computing Toolbox. For example, researchers may consider numerical maximization of the likelihood, which typically relies on finite differencing, i.e., the gradient of the likelihood function is approximated by evaluating the effects of very small changes in the parameters (implementations include the functions optim in R and fmincon and fminunc in Matlab). The large number of likelihood calculations required to obtain the gradient approximation can be sped up dramatically by allocating the likelihood evaluations over the available processing units. It must be noted that in this example, parallelization is possible because the tasks are independent. Dependent tasks are challenging, and often impossible, to parallelize. For instance, different iterations of a likelihood maximization algorithm can only be performed after each other, i.e., one can only decide on the parameter value for evaluation in iteration 1, once iteration 0 is completed; it is impossible to simultaneously perform both iterations on different processing units.

The bags of little bootstrap introduced by Kleiner, Talwalkar, Sarkar, and Jordan (2014) is a combination of bootstrap methods that enhances the tractability of analyses in terms of memory requirements. First, without replacement, subsamples of size *b* are created from the original data of size *n*. Then, within each subsample, bootstrap samples of size *n*, i.e., the total number of observations, with replacement, are drawn. The model of interest can be then estimated on each of the bootstrap samples. Estimation can be done in a computationally rather efficient way because the bootstrap samples contain (many) duplicate observations which can be efficiently handled by using a weight vector. Moreover, the bootstrap samples can be analyzed in parallel. By combining the results from the bootstrap samples, first within subsamples, and then across subsamples, full sample estimates are obtained. Kleiner et al. (2014) provide advice on the number of subsamples and subsample size. Different types of data may require a different bootstrap approach, e.g. time-series (Efron & Tibshirani 1994, pp. 99-101) and network data (Ebbes, Huang & Rangaswamy, 2015).

Finally, sequential methods offer a fast and memory-efficient method that is particularly suited to real-time data analysis (Chung, Rust & Wedel, 2009). These methods sequentially update parameters of interest as new data arrives and require only storage of the current parameter values, not the historical data. Examples of these methods include the Kalman filter (see Osinga, in press, for an introduction to state space models and the Kalman filter) and related Bayesian approaches. In the current time period, the Kalman filter makes a forecast about the next time period. As soon as we reach this next time period and new data arrives, the forecast can be compared to the observed values. Based on the forecasting error, the parameters are updated and used for making a new forecast. This new forecast is compared to new data as soon as they arrive, giving another parameter update and forecast. Thus, historical data need not be kept in memory, which enhances tractability.

Reporting and Visualization

When it comes to reporting, one of the key challenges of big data is to be complete. The variety of big data makes it important to clearly describe the different data sources that are used. Also, steps taken in pre-processing and merging of the data should be carefully discussed. For example, when working with web traffic data, the researcher needs to indicate how long a visitor needs to stay on a webpage for the visit to be counted, whether a re-visit to a webpage within a certain time frame is counted as an additional visit or not, etc. Similarly, when a program is used to rate the degree of positive sentiment in a news item, the researcher needs to be clear as to which words are searched for. Loughran and McDonald (2011) demonstrate the importance of context. For example, the words *cost* and *liability* may express negative sentiment in some settings, whereas they are more neutrally used in financial texts. Also, it should be clear to the reader whether weights are applied, and which other decisions have been made in preparing the data for analysis (e.g., Haas, Criscuolo & George, 2015).

With regard to data analysis, statistical significance becomes less meaningful when working with big data. Even variables that have a small effect on the dependent variable will be significant if the sample size is large enough. Moreover, spurious correlations are likely when considering a large number of variables. One should therefore, in addition to statistical significance, focus on the effect size of a variable and on its out-of-sample performance. Also, it is important to note that traditional statistical concepts apply to situations wherein a sample of the population is analyzed, whereas big data may capture the entire population. Bayesian statistical inference may provide a solution (Wedel & Kannan, 2016) as it assumes the data to be fixed and the parameters to be random, unlike the frequentist approach which assumes that the data can be resampled. That being said, it may also be the case that despite having big data, researchers may ultimately aggregate observations, so that sample size decreases dramatically and statistical significance remains an important issue. This might occur, for example, because one's theory is not about explaining millisecond to millisecond changes in phenomena such as heart rate, but rather about explaining differences between individuals in their overall health and well-being.

When applying methods for variable selection, it is important to describe the method used and, particularly, how the model was tuned. Since different approaches may give different results, it is highly advised to try multiple approaches to show robustness of the findings. In theory testing, it is clear which variables to use. However, these variables may often be operationalized in different ways. Also, the number and way of including control variables may be open for discussion. Simonsohn, Simmons, and Nelson (2015) advise to estimate all theoretically justified model specifications to demonstrate robustness of the results across specifications.

Finally, researchers may consider visualizing patterns in the data to give the audience a better feel for the strength of the effects, to allow easy comparison of effects, and to show that the effects are not an artifact of the complicated model that was used to obtain them. With the rise of data science, many new visualization tools have become available (e.g. Bime, Qlik Sense, and Tableau) that allow for easy application of multiple selections and visualization for large datasets.

CONCLUSION

In this editorial, we discussed the applications of data science in management research. We see opportunities for scholars to develop better answers to existing theories and extend to new questions by embracing the data scope and granularity that big data provide. Our starter kit explains the key challenges of data collection, storage, processing and analysis, and reporting and visualization – which represent departures from existing methods and paradigms. We repeat our caveat that the field is evolving rapidly with business practices and computing technologies. Even so, our editorial provides the novice with the basic elements required to experiment with data science techniques. Few decades ago, the emergence of commercial databases, such as Compustat and SDC, and the availability of advanced analytics packages, such as STATA and UCINET, revolutionized management research by enabling scholars to shift from case studies and simple two-by-two frameworks to complex models that leverage rich archival data. The advent of data science can be the next phase in this evolution, which offers opportunities not only for refining established perspective and enhancing the accuracy of known empirical results, but also for embarking into new research domains, raising new types of research questions, adopting more refined units of analysis, and shedding new light on the mechanisms that drive observed effects. Whereas data science applications are becoming pervasive in marketing and organizational behavior research, strategy scholars are yet to harness these powerful tools and techniques. Data science applications in management will take significant effort to craft, refine, and perfect. With a new generation of researchers evincing interest in these emergent areas, and the doctoral research training being provided, management scholarship is getting ready for its next generational leap forward.

Gerard George
Singapore Management University

Ernst C. Osinga
Singapore Management University

Dovev Lavie
Technion

Brent A. Scott
Michigan State University

REFERENCES

Aaker, D. A., Kumar, V., Leone, R. P., & Day, G. S. 2013. ***Marketing Research***. International student version (11th ed.). Singapore: Wiley.

Archak, N., Ghose, A., & Ipeirotis, P. G. 2011. Deriving the pricing power of product features by mining consumer reviews. ***Management Science,*** 57(8): 1485-1509.

Bono, J. E., Glomb, T. M., Shen, W., Kim, E., & Koch, A. J. 2013. Building positive resources: Effects of positive events and positive reflection on work stress and health. ***Academy of Management Journal,*** 56: 1601-1627.

Business Roundtable. 2016. ***Inventing the future, FedEx.*** http://businessroundtable.org/inventing-the-future/fedex, accessed June 7, 2016.

Butts, M., Becker, W.J., Boswell, W. R. 2015. Hot buttons and time sinks: the effects of electronic communication during nonwork time on emotions and work-nonwork conflict. ***Academy of Management Journal***, 58(3): 763-788.

Chaffin, D., Heidl, R., Hollenbeck, J. R., Howe, M., Yu, A., Voorhees, C., & Calantone, R. in press. The promise and perils of wearable sensors in organizational research. ***Organizational Research Methods.***

Colbert, A., Yee, N., George, G. 2016. The digital workforce and the workplace of the future. ***Academy of Management Journal***, 59(3): 731-739.

Das, S. R. & Chen, M. Y. 2007. Yahoo! for Amazon: Sentiment extraction from small talk on the web. ***Management Science,*** 53(9): 1375-1388.

Dhar, V. 2013. Data science and prediction. ***Communications of the ACM***, 56(12): 64-73.

Dodgson, M., Gann, D.M., Wladwsky-Berger, I., Sultan, N., George, G. 2015. Managing digital money. ***Academy of Management Journal***, 58(2): 325-333.

Ebbes, P., Huang, Z., & Rangaswamy, A., 2015. Sampling designs for recovering local and global characteristics of social networks. ***International Journal of Research in Marketing,*** forthcoming.

Efron, B., & Tibshirani, R. J. 1993. ***An introduction to the bootstrap*****.** New York: Chapman & Hall/CRC.

Frank, L. E., & Friedman, J. H. 1993. A statistical view of some chemometrics regression tools. ***Technometrics,*** 35(2): 109-135.

George, E. I., & McCulloch, R. E. 1993. Variable selection via Gibbs sampling. ***Journal of the American Statistical Association,*** 88(423): 881-889.

George, G. Haas, MR., Pentland, A. 2014. Big data and management. ***Academy of Management Journal***, 57(2): 321-325.

Gruber, M. Leon, N, George, G., Thompson, P. 2015. Managing by design, ***Academy of Management Journal***, 58 (1): 1 – 7.

Haas, M.R., Criscuolo, P., & George, G. 2015. Which problems to solve? Online knowledge sharing and attention allocation in organizations. ***Academy of Management Journal***, 58(3): 680-711.

Hastie, T., Tibshirani, R., & Friedman, J. 2009. ***The elements of statistical learning: data mining, inference and prediction*** (2nd ed.). Berlin: Springer.

Ilies, R., Dimotakis, N., & DePater, I. E. 2010. Psychological and physiological reactions to high workloads: Implications for well-being. ***Personnel Psychology,*** 63: 407-436.

Kleiner, A., Talwalkar, A., Sarkar, P., & Jordan, M. I. 2014. A scalable bootstrap for massive data. ***Journal of the Royal Statistical Society: Series B (Statistical Methodology),*** 76(4): 795-816.

Lambrecht, A. & Tucker, C. 2013. When does retargeting work? Information specificity in online advertising. ***Journal of Marketing Research,*** 50(5): 561-576.

LeCun, Y., Bengio, Y., & Hinton, G. 2015. Deep learning. ***Nature***, 521(7553): 436-444.

Lewis, R. A. & Reiley, D. H. 2014. Online ads and offline sales: Measuring the effect of retail advertising via a controlled experiment on Yahoo!. ***Quantitative Marketing and Economics***, 12(3): 235-266.

Liu, B. 2015. ***Sentiment analysis: Mining opinions, sentiments, and emotions***. New York: Cambridge University Press.

Loughran, T., & McDonald, B. 2011. When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. ***The Journal of Finance,*** 66(1): 35-65.

Madden, S. 2012. From databases to big data. ***IEEE Internet Computing***, 16(3): 4:5.

Manning, C. D., Raghavan, P., & Schütze, H. 2009. ***Introduction to information retrieval,*** Cambridge: Cambridge university press.

McAfee, A., & Brynjolfsson, E. 2012. Big data: The management revolution. ***Harvard Business Review,*** 90(10): 61-67.

Microsoft. 2016, ***Excel specifications and limits.*** https://support.office.com/en-us/article/Excel-specifications-and-limits-ca36e2dc-1f09-4620-b726-67c00b05040f, accessed June 24, 2016.

Osinga, E. C. in press. State space models. In Leeflang, P. S. H., Wieringa, J. E., Bijmolt, T. H. A., & Pauwels, K. E. (Eds.), ***Advanced methods for modeling markets.*** Chapter 5. New York: Springer.

Park, G., Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Kosinski, M., Stillwell, D. J., Ungar, L. H., & Seligman, M. E. P. 2015. Automatic personality assessment through social media language. ***Journal of Personality and Social Psychology,*** 108: 934-952.

Prajapati, V. 2013. ***Big data analytics with R and Hadoop***, Birmingham: Packt Publishing.

Reuters. 2013. ***AP, news aggregator Meltwater end copyright dispute.*** http://www.reuters.com/article/manniap-meltwater-lawsuit-idUSL1N0FZ17920130729, accessed June 7, 2016.

Simonsohn, U., Simmons, J. P., & Nelson, L.D. 2015. ***Specification curve: Descriptive and inferential statistics on all reasonable specifications.*** Available at SSRN.

Sismeiro, C., & Bucklin, R. E. 2004. Modeling purchase behavior at an e-commerce web site: A task-completion approach. ***Journal of marketing research,*** 41(3): 306-323.

Teixeira, T., Wedel, M., & Pieters, R. 2012. Emotion-induced engagement in internet video advertisements. ***Journal of Marketing Research,*** 49(2): 144-159.

Tirunillai, S. & Tellis, G. J. 2012. Does chatter really matter? Dynamics of user-generated content and stock performance. ***Marketing Science***, 31(2): 198-215.

UPS. 2016. ***Leadership matters, telematics.*** https://www.ups.com/content/us/en/bussol/browse/leadership-telematics.html, accessed June 7, 2016.

van der Vegt, G., Essens, P., Wahlstrom, M., George, G. 2015. Managing risk and resilience. ***Academy of Management Journal***, 58(4): 971-980.

van Knippenberg, D., Dahlander, L., Haas, M., George, G. 2015. Information, attention and decision-making. ***Academy of Management Journal***, 58(3): 649-657.

Varian, H. R. 2014. Big data: New tricks for econometrics. ***The Journal of Economic Perspectives,*** 28(2): pp.3-27.

Wang, C., Chen, M.H., Schifano, E., Wu, J., & Yan, J. 2015. ***A survey of statistical methods and computing for big data.*** arXiv preprint arXiv:1502.07989.

Wedel, M., & Kannan, P. K. 2016. Marketing analytics for data-rich environments. ***Journal of Marketing,*** forthcoming.

Wiesel, T., Pauwels, K., & Arts, J. 2011. Practice prize paper - Marketing's profit Impact: Quantifying online and off-line funnel progression. ***Marketing Science,*** 30(4): 604-611.

Xu, L., Duan, J. A., & Whinston, A. 2014. Path to purchase: A mutually exciting point process model for online advertising and conversion. ***Management Science,*** 60(6): 1392-1412.

Zikopoulos, P & Eaton, C. 2011. ***Understanding big data: analytics for enterprise class Hadoop and streaming data***. McGraw Hill Osborne Media: NY.

Bio

Gerry George is dean and Lee Kong Chian Chair Professor of Innovation and Entrepreneurship at the Lee Kong Chian School of Business at Singapore Management University. He also serves as the editor of *AMJ*.

Ernst Osinga is an assistant professor of marketing at the Lee Kong Chian School of Business at Singapore Management University. His research interests include online advertising and retailing, pharmaceutical marketing, and the marketing-finance interface. He received his PhD in marketing from the University of Groningen.

Dovev Lavie is professor and vice dean at the Technion. He earned his PhD at the Wharton School. His research focuses on alliance portfolios, balancing exploration and exploitation, and applications of the resource-based view. He serves as associate editor of *AMJ* handling manuscripts in strategy and organization theory.

Brent Scott is professor of management at the Broad College of Business at Michigan State University. His research focuses on emotions, organizational justice, and employee well-being. He serves as associate editor of *AMJ* handling manuscripts in organizational behavior.

Acknowledgements

We are grateful for the insightful comments of Kevin Boudreau, Avigdor Gal, John Hollenbeck, Mark Kennedy, and Michel Wedel on earlier versions. Gerry George gratefully acknowledges the financial and research support of The Lee Foundation.

Table 1. Big data challenges and solutions

Process	Challenges	Solutions	Key references
Data access and collection	• Easy access to data offered in standardized formats. No practical limit to the size of these data offering unlimited scalability. • Efficiently obtain detailed data for a large number of agents • Protocols on security, privacy and data rights.	• Sensors • Webscraping • Web traffic and communications monitoring	• Chaffin et al. (in press) • Sismeiro and Bucklin (2004)
Data storage	• Tools for data collection, matching and integration of different big datasets • Data Reliability • Warehousing	• SQL, NoSQL, Apache Hadoop • Save essential information only and update in real-time	• Varian (2014) • Prajapati (2013)
Data processing	• Use non-numeric data for quantitative analyses	• Text mining tools to transform text into numbers • Emotion recognition	• Manning, Raghavar, and Schűtze (2009) • Teixeira, Wedel, and Pieters (2012)
Data analysis	• Large number of variables • Causality • Find latent topics and attach meaning • Data too large to process	• Ridge, lasso, principal components regression, partial least squares, regression trees • Topic modeling, LDA, entropy-based measures, and deep learning • Cross validation and holdout samples • Field experiments • Parallelization, bags of little bootstrap, sequential analysis	• Hastie, Tibshirani, and Friedman (2009) • George and McCulloch (1993) • Archak Ghose, and Ipeirotis (2011) • Tirunillai and Tellis (2014) • Blei, Ng, and Jordan (2003) • LeCun, Bengio, and Hinton (2015) • Lambrecht and Tucker (2013) • Wang, Chen, Schifano, Wu, and Yan (2015) • Wedel and Kannan (2016)
Reporting and Visualization	• Facilitate interpretation, representation with external partners and knowledge users. • Difficult to understand complex patterns	• Describe data sources • Describe methods and specifications • Bayesian analysis • Visualization and graphic interpretations	• Loughran and McDonald (2011) • Simonsohn, Simmons, and Nelson (2015)