



Big data: Dimensions, evolution, impacts, and challenges

In Lee

School of Computer Sciences, Western Illinois University, 1 University Circle, Macomb, IL 61455-1390, U.S.A.

KEYWORDS

Big data;
Internet of things;
Data analytics;
Sentiment analysis;
Social network analysis;
Web analytics

Abstract Big data represents a new technology paradigm for data that are generated at high velocity and high volume, and with high variety. Big data is envisioned as a game changer capable of revolutionizing the way businesses operate in many industries. This article introduces an integrated view of big data, traces the evolution of big data over the past 20 years, and discusses data analytics essential for processing various structured and unstructured data. This article illustrates the application of data analytics using merchant review data. The impacts of big data on key business performances are then evaluated. Finally, six technical and managerial challenges are discussed.

© 2017 Kelley School of Business, Indiana University. Published by Elsevier Inc. All rights reserved.

1. The day of big data

The emerging technological development of big data is recognized as one of the most important areas of future information technology and is evolving at a rapid speed, driven in part by social media and the Internet of Things (IoT) phenomenon. The technological developments in big data infrastructure, analytics, and services allow firms to transform themselves into data-driven organizations. Due to the potential of big data becoming a game changer, every firm needs to build capabilities to leverage big data in order to stay competitive.

IDC (2015) forecasted that the big data technology and services market will grow at a compound annual growth rate of 23.1% over the 2014–2019 period, with annual spending reaching \$48.6 billion in 2019.

While structured data is an essential part of big data, more and more data are created in unstructured video and image forms, which traditional data management technologies are inadequate to process. A large portion of data worldwide have been generated by billions of IoT devices such as smart home appliances, wearable devices, and environmental sensors. Gartner (2015) forecasted that 4.9 billion connected objects would be in use in 2015—up 30% from 2014—and will reach 25 billion by 2020. To meet the ever-increasing storage and processing needs of big data, new big data platforms

E-mail address: i-lee@wiu.edu

are emerging, including NoSQL¹ databases as an alternative to traditional relational databases and Hadoop as an open source framework for inexpensive distributed clusters of commodity hardware.

In this article, I start with a discussion of big data dimensions and trace the evolution of big data since 1995. Then, I illustrate the application of data analytics using a scenario involving merchant review data. In the following section, I discuss impacts of big data on various business performances. Finally, I discuss six technical and managerial challenges: data quality, data security, privacy, data management, investment justification, and shortage of qualified data scientists.

2. Dimensions of big data

Laney (2001) suggested that volume, variety, and velocity are the three dimensions of big data. The 3 Vs have been used as a common framework to describe big data (Chen, Chiang, & Storey, 2012; Kwon, Lee, & Shin, 2014). Here, I describe the 3 Vs and additional dimensions of big data proposed in the computing industry.

Volume refers to the amount of data an organization or an individual collects and/or generates. While currently a minimum of 1 terabyte is the threshold of big data, the minimum size to qualify as big data is a function of technology development. Currently, 1 terabyte stores as much data as would fit on 1,500 CDs or 220 DVDs, enough to store around 16 million Facebook photographs (Gandomi & Haider, 2015). E-commerce, social media, and sensors generate high volumes of unstructured data such as audio, images, and video. New data has been added at an increasing rate as more computing devices are connected to the internet.

Velocity refers to the speed at which data are generated and processed. The velocity of data increases over time. Initially, companies analyzed data using batch processing systems because of the slow and expensive nature of data processing. As the speed of data generation and processing increased, real time processing became a norm for computing applications. Gartner (2015) forecasted that 6.4 billion connected devices would be in use worldwide in 2016 and that the number will reach 20.8 billion by 2020. In 2016, 5.5 million new devices were estimated to be connected every day to collect, analyze, and share data. The enhanced data streaming capability of connected devices will continue to accelerate the velocity.

Variety refers to the number of data types. Technological advances allow organizations to generate various types of structured, semi-structured, and unstructured data. Text, photo, audio, video, clickstream data, and sensor data are examples of unstructured data, which lack the standardized structure required for efficient computing. Semi-structured data does not conform to specifications of the relational database, but can be specified to meet certain structural needs of applications. An example of semi-structured data is Extensible Business Reporting Language (XBRL), developed to exchange financial data between organizations and government agencies. Structured data is predefined and can be found in many types of traditional databases. As new analytics techniques are developed, unstructured data are generated at a much faster rate than structured data and the data type becomes less of an impediment for the analysis.

IBM added *veracity* as a fourth dimension, which represents the unreliability and uncertainty latent in data sources. Uncertainty and unreliability arise due to incompleteness, inaccuracy, latency, inconsistency, subjectivity, and deception in data. Managers do not trust data when veracity issues are prevalent. Customer sentiments are unreliable and uncertain due to subjectivity of human opinions. Statistical tools and techniques have been developed to deal with uncertainty and unreliability of big data with specified confidence levels or intervals.

SAS added two additional dimensions to big data: variability and complexity. *Variability* refers to the variation in data flow rates. In addition to the increasing velocity and variety of data, data flows can fluctuate with unpredictable peaks and troughs. Unpredictable event-triggered peak data are challenging to manage with limited computing resources. On the other hand, investment in resources to meet the peak-level computing demand will be costly due to overall underutilization of the resources. *Complexity* refers to the number of data sources. Big data are collected from numerous data sources. Complexity makes it difficult to collect, cleanse, store, and process heterogeneous data. It is necessary to reduce the complexity with open sources, standard platforms, and real-time processing of streaming data.

Oracle introduced *value* as an additional dimension of big data. Firms need to understand the importance of using big data to increase revenue, decrease operational costs, and serve customers better, but at the same time must consider the investment cost of a big data project. Data would be low value in their original form, but data analytics will transform the data into a high-value

¹ Interpreted as Not Only SQL

strategic asset. IT professionals need to assess the benefits and costs of collecting and/or generating big data, choose high-value data sources, and build analytics capable of providing value-added information to managers.

As discussed above, a number of dimensions were presented in the computing industry. These dimensions together help us understand big data. I propose one additional dimension to big data: decay. *Decay* of data refers to the declining value of data over time. In a time of high velocity, the timely processing and acting on analysis is all the more important. IoT devices generate high volumes of streaming data, and immediate processing is often required for time-critical situations such as patient monitoring and environmental safety monitoring. Wearable medical devices such as glucose monitors, pulse oximeters, and blood pressure monitors worn on or close to the body produce a stream of data on patients' physiological conditions. In the era of big data, the decay of data will be an exponential function of time.

3. An integrated view of big data

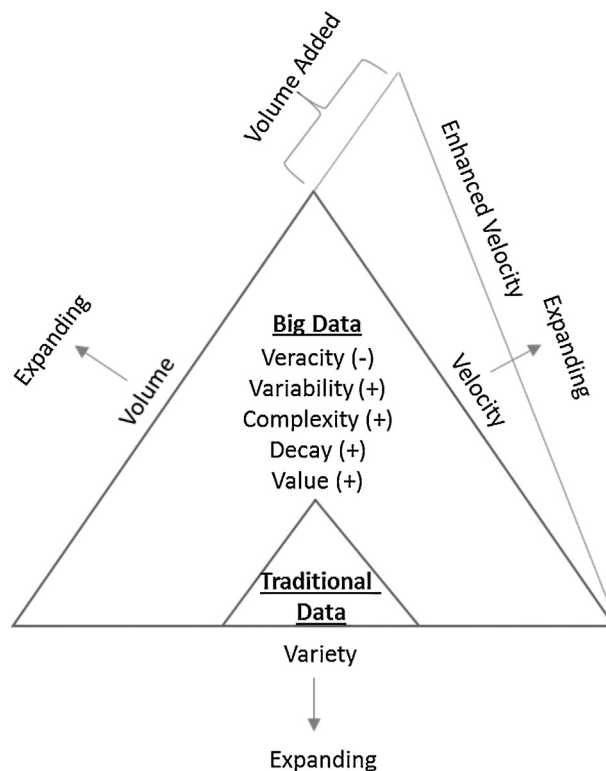
Currently, the dimensions of big data have been proposed separately in the computing industry, but

so far we lack an integrated view of big data. To help managers fully understand the relationships between the dimensions of big data, I present an integrated view of big data in Figure 1. The integrated view shows how these dimensions are inter-related with each other.

The three edges of the integrated view of big data represent three dimensions of big data: volume, velocity, and variety. Inside the triangle are the five dimensions of big data that are affected by the growth of the three triangular dimensions: veracity, variability, complexity, decay, and value. The growth of the three edged dimensions is negatively related to veracity, but positively related to complexity, variability, decay, and value.

The integrated view shows that traditional data is a subset of big data with the same three dimensions, but the scope of each dimension is much smaller than that of big data. Traditional data consist mostly of structured data, and relational database management systems have been widely used to collect, store, and process the traditional data. As the scope of the three dimensions continues to expand, the proportion of the unstructured data increases. The magnitude of big data is expanding with the growth of velocity, volume, and variety. The arrows represent the expansion of each of the three dimensions. The expansion of velocity,

Figure 1. An integrated view of big data



volume, and variety are intertwined with each other. The expansion of each dimension affects the other seven dimensions. For example, [Figure 1](#) shows that the expansion of velocity affects either volume or variety or both, and consequently affects the other five dimensions of big data inside the triangle (i.e., veracity declines, but variability, complexity, decay, and value increase).

4. Evolution of big data and data analytics

While the emergence of big data occurred only recently, the act of gathering and storing large amounts of data dates back to the early 1950s when the first commercial mainframe computers were introduced. During the period between the early 1950s to mid-1990s, data grew relatively slowly due to the high cost of computers, storage, and data networks. Data during this period were highly structured, mainly to support operational and transactional information systems. The advent of the world wide web (WWW) in the early 1990s led to the explosive growth of data and the development of big data analytics. Since the advent of the WWW, big data and data analytics have evolved through three major stages.

4.1. Big Data 1.0 (1994–2004)

Big Data 1.0 coincides with the advent of e-commerce in 1994, during which time online firms were the main contributors of the web content. User-generated content was only a marginal part of web content due to the technical limitation of web applications. In this era, web mining techniques were developed to analyze users' online activities. Web mining can be divided into three different types: web usage mining, web structure mining, and web content mining.

Web usage mining is the application of data mining techniques to discover web users' usage patterns online. Usage data captures the identity or origin of web users along with their browsing behavior. The ability to track individual users' mouse clicks, searches, and browsing patterns makes it possible to provide personalized services to users.

Web structure mining is the process of analyzing the structure of a website or a web page. The structure of a typical website consists of web pages as nodes and hyperlinks as edges connecting related pages. A hyperlink connects a location in a web page to a different location, either within the same web

page or on a different web page. Based on the hyperlink structure, web pages are categorized. Google's PageRank, rooted in social network analysis, analyzes the hyperlink structure of web pages to rank them according to their degree of popularity or importance.

Web content mining is the process of extracting useful information from the content of web pages. A web page may consist of text, images, audio, video, or Extensible Markup Language-based (XML-based) data. Text mining has been applied widely to web content mining. Text mining extracts information from unstructured text and draws heavily on techniques from such disciplines as information retrieval (IR) and natural language processing (NLP). In its simplest form, text mining extracts a certain set of words or terms that are commonly used in the text. Web content mining is concerned about extraction of web page information, clustering of web pages, and classification of web pages into cyberterrorism, email fraud, spam mail filtering, etc. While there existed mining techniques in image processing and computer vision, the application of these techniques to web content mining was limited during the Big Data 1.0 era.

4.2. Big Data 2.0 (2005–2014)

Big Data 2.0 is driven by Web 2.0 and the social media phenomenon. Web 2.0 refers to a web paradigm that evolved from the web technologies of the 1990s and allowed web users to interact with websites and contribute their own content to the websites. Social media embodied the principles of Web 2.0 ([O'Reilly, 2007](#)) and created a paradigm shift in the way organizations operate and collaborate. As social media is tremendously popular among consumers, firms can leverage it to engage in frequent and direct consumer contact with a broad reach at a relatively low cost ([Kaplan & Haenlein, 2010](#)).

Social media analytics support social media content mining, usage mining, and structure mining activities. Social media analytics analyze and interpret human behaviors at social media sites, providing insights and drawing conclusions from a consumer's interests, web browsing patterns, friend lists, sentiments, profession, and opinions. By understanding customers better using social media analytics, firms develop effective relationship marketing campaigns for targeted customer segments and tailor products and services to customers' needs and interests. For example, major U.S. banks analyze clients' comments on social media sites about their service experiences and satisfaction levels. Unlike web analytics used mainly for structured data, social media analytics are used for the analysis of data

likely to be natural language, unstructured, and context-dependent.

The worldwide social media analytics market is growing rapidly from \$1.6 billion in 2015 to an estimated \$5.4 billion by 2020 at a compound annual growth rate of 27.6%. This growth is attributable to advanced analytics and the increase in the number of social media users ([ReportsnReports, 2016](#)). Some social media analytics software programs are provided as cloud-based services with flexible fee options, such as monthly subscription or pay-as-you-go pricing. Social media analytics focus on two types of analysis: sentiment analysis and social network analysis.

Sentiment analysis uses text analysis, natural language processing, and computational linguistics to identify and extract user sentiments or opinions from text materials. Sentiment analysis can be performed at multiple levels, such as entity level, sentence level, and document level. An entity-level analysis identifies and analyzes individual entity's opinions contained in a document. A sentence-level analysis identifies and analyzes sentiments expressed in sentences. A document-level analysis identifies and analyzes an overarching sentiment expressed in the entire document. While a deeper understanding of sentiment and greater accuracy are still to be desired, sentiments extracted from documents have been used successfully by businesses in various ways, including predicting stock market movements, determining market trends, analyzing product defects, and managing crises ([Fan & Gordon, 2014](#)). However, sentiment analysis can be flawed. Sampling biases in the data can skew results as in situations where satisfied customers remain silent while those with more extreme positions express their opinions ([Fan & Gordon, 2014](#)).

Lexical-based methods and machine-learning methods are two widely used methods for sentiment analysis. Lexical-based methods use a predefined set of words in which each word carries a specific sentiment. These methods include:

- simple word or phrase counts;
- the use of emoticons to detect polarity; that is, positive and negative emoticons used in a message ([Park, Barash, Fink, & Cha, 2013](#));
- sentiment lexicons, based on the words in the lexicon that have received specific features marking the positive or negative terms in a message ([Gayo-Avello, 2011](#)); and
- the use of psychometric scales to identify mood-based sentiments.

One of the challenges of the lexical-based methods is to create a lexical-based dictionary to be used for different contexts.

Machine-learning methods often rely on the use of supervised and unsupervised machine-centered techniques. While one advantage of machine-learning methods is the ability to adapt and generate trained models for specific purposes and contexts, the drawback to these methods is the lacking availability of labeled data and hence the low applicability of the methods to new situations ([Gonçalves, Araújo, Benevenuto, & Cha, 2013](#)). In addition, labeling data might be costly or even prohibitive for some tasks. While machine-learning methods are reported to perform better than lexical-based methods, it is hard to conclude whether a single machine-learning method is better than all lexical-based methods across different tasks.

Social network analysis is the process of measuring the social network structure, connections, nodes, and other properties by modeling social network dynamics and growth (e.g., network density, network centrality, network flows). Social network analysis originally was developed before the advent of social media to study relationships among actors in modern sociology. The relationships typically are identified from links directly connecting two actors or inferred indirectly from tagging, social-oriented interactions, content sharing, and voting. Social networking sites such as Facebook, Twitter, and LinkedIn provide a central point of access and bring structure in the process of personal information sharing and online socialization ([Jamali & Abolhassani, 2006](#)).

Social network analysis uses a variety of techniques pertinent to understanding the structure of the network ([Scott, 2012](#)). These techniques range from simpler methods, such as counting the number of edges a node has or computing path lengths, to more sophisticated methods that compute eigenvectors to determine key nodes in a network ([Fan & Gordon, 2014](#)). Social networking sites have been a popular subject of social network analysis. [Marlow \(2004\)](#) employs social network analysis to describe the social structure of blogs. He explores two metrics of authority: popularity measured by bloggers' public affiliations and influence measured by citation of the writing.

4.3. Big Data 3.0 (2015—)

Big Data 3.0 encompasses data from Big Data 1.0 and Big Data 2.0. The main contributors of Big Data 3.0 are the IoT applications that generate data in the form of images, audio, and video. The IoT refers to a technology environment in which

devices and sensors have unique identifiers with the ability to share data and collaborate over the internet even without any human intervention. With the rapid growth of the IoT, connected devices and sensors will surpass social media and e-commerce websites as the primary sources of big data. GE is developing IoT-based sensors that read data from equipment deployed for aviation and healthcare operations. Agribusinesses also use IoT-based sensors to manage resources like water, grain storage, and heavy equipment in an effort to drive down agricultural costs and increase yields.

For many IoT applications, the analysis increasingly is performed by sensors at the source of data gathering. This trend is leading to a new field known as streaming analytics. Streaming analytics continuously extract information from the streaming data. Unlike social media analytics used in a batch mode for the analysis of stored data, streaming analytics involve real-time event analysis to discover patterns of interest as data is being collected or generated. Streaming analytics are used not just for monitoring existing conditions but also for predicting future events.

Streaming analytics have great potential in a number of industries where streaming data are generated through human activities, machine data, or sensor data. For example, streaming analytics embedded in sensors can monitor and interpret patients' physiological and behavioral changes and alert caregivers to urgent medical needs. Streaming analytics also can be useful in the financial industry where electronic transactions need to be monitored under financial regulations and immediate actions are required in the event of suspicious and fraudulent financial activities.

5. An illustrative example: Data analysis of merchant review big data

With the explosive growth of merchant reviews at various vendor/product review sites and social media sites, merchant review big data has drawn the attention of researchers and practitioners. Reviews written by consumers are perceived to be less biased than those provided by advertisers or product experts. The review credibility can be further enhanced by providing a feedback function for viewers to rate the usefulness of the particular reviews. Yelp, TripAdvisor, and Angie's List are popular merchant review sites. These sites enable consumers to rate a particular merchant based on a numerical scale (e.g., 1 to 5). They provide viewers with the entirety of consumers' reviews along with

the ability to vote on the helpfulness of those reviews. It would be challenging for small business merchants to analyze their review data, since the data are large and unstructured. Despite the popularity of the merchant review, it is still the case that merchants fail to fully exploit and translate consumer reviews into business value.

This section illustrates a simple but powerful application of social media analytics to merchant review data. A social media analytics model was developed to discover relationships between consumers' review activities and the viewers' usefulness votes in the context of Groupon users' merchant reviews. Five factors related to consumers' review activities were identified that may influence the usefulness of the review. The five factors include (1) the review score of the reviewer, (2) the number of social network friends of the reviewer, (3) the cumulative number of reviews made by the reviewer, (4) the number of words in the reviewer's comment, and (5) the existence of images or photos in the reviewer's comment. Note that the number of words is derived from the reviewer's comment, which is originally in an unstructured form. The existence of images or photos is a dummy variable in this model. The dependent variable is the number of usefulness votes by viewers.

A multiple regression model is used to identify the factors that are strongly associated with the number of usefulness votes. Merchant reviews were collected in July 2015 from 108 healthcare merchants that launched Groupon promotions between June and July 2011. The term healthcare, as used by Groupon, refers to a variety of businesses and services, from a haircut salon and spa to fitness or yoga training. Groupon users were identified in July 2015 at the Yelp's web pages of the 108 healthcare merchants, and their reviews were transcribed for analysis. Out of 589 reviews, 189 reviews were removed that did not have any response from viewers. 400 reviews were analyzed using a multiple regression model. The descriptive statistics are shown in [Table 1](#). [Table 2](#) shows the beta coefficients of these variables as well as their p-values. The regression was run with a 5% level of significance.

The results show that the review score, the number of social network friends of a reviewer, and the number of words in a review are significant predictors of the number of usefulness votes. The cumulative number of reviews made by a reviewer and the existence of images/photos do not have an impact on the number of usefulness votes. It is interesting to note that as the review score increases, the number of usefulness votes decreases. The review score's negative effect on the number of

Table 1. Descriptive statistics of merchant review

	Mean	Standard Deviation	n
Review Score	3.365	1.5705	400
Number of Friends	72.0675	165.3403	400
Number of Reviews	96.37	168.382	400
Number of Words	221.96250	163.2673	400
Image/Photo	0.0375	0.1899	400

usefulness votes indicates that when review scores are low, viewers feel the review is more useful for their purchase decision making. The reviewers who have more friends in the social network influence the viewers' votes. Therefore, merchants need to pay more attention to network leaders with higher numbers of social network friends. This result is consistent with a finding that consumers are more likely to trust a reviewer who has a higher number of followers (Cheung & Ho, 2015). Finally, the number of words in the reviewer's comment has a significant positive effect on the number of usefulness votes by viewers. This result may be explained by the fact that the higher number of

words brings more helpful information to viewers and leads to the reduced information asymmetry between the merchant and the viewers. While more comprehensive social media analytics might add more value to merchants, this illustration shows that even simple data analytics can deliver highly valuable marketing ideas.

6. Impacts of big data

Big data provides great potential for firms in creating new businesses, developing new products and services, and improving business operations. The

Table 2. Results of the regression

Multiple Linear Regression - Estimated Regression Equation					
useful[t] = +3.18935 – 0.632662score[t] + 0.00557255friend[t] + 0.00135261review[t] + 0.00921948word[t] + 0.717262image[t] + 0.00192933t + e[t]					
Multiple Linear Regression - Ordinary Least Squares					
Variable	Parameter	S.D.	T-STAT H0: parameter = 0	2-tail p-value	1-tail p-value
(Intercept)	+3.189	0.8144	+3.9160e+00	0.0001061	5.306e-05
score	–0.6327	0.1469	–4.3080e+00	2.083e-05	1.041e-05
friend	+0.005573	0.001739	+3.2050e+00	0.001461	0.0007305
review	+0.001353	0.001692	+7.9920e-01	0.4247	0.2123
word	+0.00922	0.001384	+6.6610e+00	9.169e-11	4.585e-11
image	+0.7173	1.203	+5.9620e-01	0.5514	0.2757
t	+0.001929	0.001976	+9.7620e-01	0.3296	0.1648
Multiple Linear Regression - Regression Statistics					
Multiple R					0.4443
R-squared					0.1974
Adjusted R-squared					0.1852
F-TEST (value)					16.11
F-TEST (DF numerator)					6
F-TEST (DF denominator)					393
p-value					1.11e-16
Multiple Linear Regression - Residual Statistics					
Residual Standard Deviation					4.437
Sum Squared Residuals					7737

use of big data analytics can create benefits, such as cost savings, better decision making, and higher product and service quality (Davenport, 2014). Personalized advertising that is finely tuned to what consumers are looking for and news articles related to their interests are some of the impacts of big data (Goetz, 2014). It is also noted that while managers realize that big data has potential impacts on firms, they still face difficulty in exploiting the data. The following discusses the impacts of big data on large firms.

6.1. Personalization marketing

By exploiting big data from multiple sources, firms can deliver personalized product/service recommendations, coupons, and other promotional offers. Major retailers such as Macy's and Target use big data to analyze shoppers' preferences and sentiments and improve their shopping experience. Innovative fintech firms have already started using social media data to assess the credit risk and financing needs of potential clients and provide new types of financial products for them. Banks are analyzing big data to increase revenue, boost retention of clients, and serve clients better. U.S. Bank, a major commercial bank in the U.S., deployed data analytics that integrate data from online and offline channels and provide a unified view of clients to enhance customer relation management. As a result, the bank's lead conversion rate has improved by over 100% and clients have been able to receive more personalized experiences (The Financial Brand, 2014).

6.2. Better pricing

Harnessing big data collected from customer interactions allows firms to price appropriately and reap the rewards (Baker, Kiewell, & Winkler, 2014). Sears uses big data to help set prices and give loyalty shoppers customized coupons. Sears deployed one of the largest Hadoop clusters in the retail industry and now utilizes open source technologies to keep the cost of big data low. Sears analyzes massive amounts of data about product availability in its stores to prices at other retailers to local weather conditions in order to set prices dynamically. eBay also uses open source Hadoop technology and data analytics to optimize prices and customer satisfaction. To achieve the highest price possible for items sellers place for auction, eBay examines all data related to items sold before (e.g., a relationship between video quality of auction items and bidding prices) and suggests ways to maximize results to sellers.

6.3. Cost reduction

Big data reduces operational costs for many firms. According to Accenture (2016), firms that use data analytics in their operations have faster and more effective reaction time to supply chain issues than those that use data analytics on an ad-hoc basis (47% vs. 18%). Big data analytics leads to better demand forecasts, more efficient routing with visualization and real-time tracking during shipments, and highly optimized distribution network management (House, 2014).

GE helps the oil and gas industry improve equipment reliability and availability, resulting in better operations efficiency and higher oil and gas productivity. Real-time monitoring systems transmit massive amounts of data to central facilities where they are processed with data analytics to assess equipment conditions. GE provides Southwest Airlines with proprietary flight efficiency analytics to analyze flight and operational data in order to identify and prioritize fuel savings opportunities (Business Wire, 2015).

Big data has also led to enormous cost reduction in the retail industry. Tesco, a European supermarket store, analyzes refrigerator data to reduce energy cost by about \$25 million annually. Analyses of refrigerator data showed that the temperature of refrigerators were set colder than necessary, wasting electricity. To optimize the temperature of the refrigerators, all Tesco refrigerators in Ireland were equipped with sensors that monitored the temperature every 3 seconds (van Rijmenam, 2016).

6.4. Improved customer service

Big data analytics can integrate data from multiple communication channels (e.g., phone, email, instant message) and assist customer service personnel in understanding the context of customer problems holistically and addressing problems quickly. Big data analytics can also be used to analyze transaction activities in real time, detect fraudulent activities, and notify clients of potential issues promptly. Insurance claim representatives can serve clients proactively based on the correlation analysis of weather data and certain types of claims submitted on stormy or snowy days.

Hertz, a car rental company in the U.S., uses big data to improve customer satisfaction. Hertz gathers data on its customers from emails, text messages, and online surveys to drive operational improvement. For example, Hertz discovered that return delays were occurring during specific hours of a day at an office in Philadelphia and was able to add staff during the peak activity hours to make

sure that any issues were resolved promptly (IBM, 2010). Southwest Airlines uses speech analytics to extract business intelligence from conversations between customers and the company's service personnel. The airline also uses social media analytics to delve into customers' social media data for a better understanding of customer intent and better service offerings (Aspect, 2013).

7. Challenges in big data

Based on the survey of big data practices, in this section I discuss challenges in big data development and management. As with any disruptive innovation, big data presents multiple challenges to adopting firms. For example, SAS (2013) notes that enterprises will face challenges in processing speed, data interpretation, data quality, visualization, and exception handling of big data. I highlight six technical and managerial challenges: data quality, data security, privacy, investment justification, data management, and shortage of qualified data scientists.

7.1. Data quality

Data quality refers to the fitness of data with respect to a specific purpose of usage. Data quality is critical to confidence in decision making. As data are more unstructured and collected from a wider array of sources, the quality of data tends to decline. For firms adopting data analytics for their supply chain, data quality is paramount. If the data are not of high quality, managers will not use the data, let alone want to share the data with their partners. Streaming analytics use data generated by interconnected sensors and communication devices. If a medical monitoring system's sensor generates erroneous data, the streaming analytics may send a wrong signal to the controlling devices that may be fatal to patients. A data quality control process needs to be established to develop quality metrics, evaluate data quality, repair erroneous data, and assess a trade-off between quality assurance costs and gains.

7.2. Data security

Weak security creates user resistance to the adoption of big data. It also leads to financial loss and damage to a firm's reputation. Without installing proper security mechanisms, confidential information could be transmitted inadvertently to unintended parties. This security challenge may be alleviated by establishing strong security management protocol, along

with security solutions such as intrusion prevention and detection systems, encryptions, and firewalls built into big data systems. Blockchain, an underlying technology behind the Bitcoin cryptocurrency, is a promising future technology for big data security management. Saving data in an encrypted form rather than in its original format, blockchain ensures that each data element is unique, time-stamped, and tamper-resistant, and its applications extend beyond financial industries due to the enhanced level of data security.

7.3. Privacy

As big data technologies mature, the extensive collection of personal data raises serious concerns for individuals, firms, and governments. Without addressing these concerns, individuals may find data analytics worrisome and decide not to contribute personal data that can be analyzed later. According to the 2015 TRUSTe Internet of Things Privacy Index, only 20% of online users believe that the benefits of smart devices outweighed any privacy concerns (TRUSTe, 2015). As is the case with smart health equipment and smart car emergency services, sensors can provide a vast amount of data on users' location and movements, health conditions, and purchasing preferences, all of which raise significant privacy concerns. However, protecting privacy is often counterproductive to both firms and customers, as big data is a key to enhanced service quality and cost reduction. Therefore, firms and customers need to strike a balance between the use of personal data for services and privacy concerns. It is noted that there is no one-size-fits-all measure for privacy, but the balance depends on service type, customers served, data type, and regulatory environments.

7.4. Investment justification

According to Accenture (2016), the actual use of big data analytics is limited. Despite the touted benefits of big data, firms face difficulties in proving the value of big data investments. A majority of surveyed executives (67%) expressed concerns about the large investment required to implement and use analytics. Many big data projects have unclear problem definitions and use emerging technologies, thus causing a higher risk of project failure and higher irreversibility of investments than traditional technology projects. In addition, if tangible costs significantly outweigh tangible benefits—despite potentially large intangible benefits—it will be hard to justify investment to senior management due to the calculated negative financial returns. When a

project is highly risky and irreversible, a real option approach may be appropriate (Lee & Lee, 2015). In a real option approach, options such as postponement, expansion, shrinkage, and scrapping of a project are viable, and there is not an obligation to move forward with a big data plan as-is.

7.5. Data management

Social media and streaming sensors generate massive amounts of data that need to be processed. Few firms would be able to invest in data storage for all big data collected from their sources. The current architecture of the data center is not prepared to deal with the heterogeneous nature of personal and enterprise data (Gartner, 2015). Deutsche Bank has been working on big data implementation since the beginning of 2012 in an attempt to analyze all of its unstructured data. However, problems have occurred when trying to make big data applications work with its traditional mainframes and databases. Petabytes of data had been stored across dozens of data warehouses, but extracting these for analysis became an expensive proposition (The Financial Brand, 2014).

A combination of edge computing and Hadoop has the potential to help reduce data management issues for firms. Hadoop is useful for complex transformations and computations of big data in a distributed computing environment. However, Hadoop is not suitable for ad hoc data exploration and streaming analytics. The need for streaming analytics and real time responses is driving the development of edge computing, also known as fog computing. However, edge computing is costlier to develop and maintain than data centers.

7.6. Shortage of qualified data scientists

As the need to manipulate unstructured data such as text, video, and images increases rapidly, the need for more competent data scientists grows. According to an A.T. Kearney survey of 430 senior executives, despite the prediction that firms will need 33% more big data specialists over the next 5 years, roughly 66% of firms with advanced analytics capabilities were not able to obtain enough employees to deliver insights into their big data (Boulton, 2015). The McKinsey Global Institute estimated that the U.S. needs 140,000–190,000 more workers with analytical skills and 1.5 million more managers and analysts with analytical skills to make business decisions based on the analysis of big data (Manyika et al., 2011). IDC (2015) reported that the staff shortage will extend from data scientists to data architects and experts in data

management, and big data-related professional services will grow at a compound annual rate of 23% through 2020 (Vesset et al., 2015). If this shortage continues, firms will have to offer highly competitive salaries to qualified data scientists and may need to develop data analytics training programs in order to groom internal employees to meet the demand.

8. The future of big data

Big data's emergence has not remained isolated to a few sectors or spheres of technology, instead demonstrating broad applications across industries. In light of this reality, companies must first pursue big data capabilities as necessary ground-level developments, which in turn may facilitate competitive advantages. Formidable challenges face firms in pursuit of big data integration, but the potential benefits of big data promise to positively impact company operations, marketing, customer experience, and more. Using tools to assess and understand big data, such as the integrated view of big data dimensions presented here (see Figure 1), will help companies both to realize the individual benefits of big data integration and to position themselves within the wider technological shift as big data becomes a part of mainstream business practices. There is a need for more practical research examining and tackling these challenges to big data within business, as well as a need for industry changes to encourage talent and infrastructure development. This big data reality in which companies find themselves is vast, complex, comprehensive—and here to stay.

References

- Accenture. (2016). *Big data analytics in supply chain: Hype or here to stay?* Available at <https://www.accenture.com>
- Aspect. (2013, August 27). *Southwest Airlines heads to the cloud with Aspect software to provide best-in-class customer experience* [Press Release]. Retrieved from <http://www.aspect.com/company/news/press-releases/southwest-airlines-heads-to-the-cloud-with-aspect-software-to-provide-best-in-class-customer-experience>
- Baker, W., Kiewell, D., & Winkler, G. (2014). Using big data to make better pricing decisions. *McKinsey*. Retrieved from <http://www.mckinsey.com/business-functions/marketing-and-sales/our-insights/using-big-data-to-make-better-pricing-decisions>
- Boulton, C. (2015). Lack of big data talent hampers corporate analytics. *CIO*. Retrieved from <http://www.cio.com/article/3013566/analytics/lack-of-big-data-talent-hampers-corporate-analytics.html>
- Business Wire. (2015, September 30). *Southwest Airlines selects GE's Flight Efficiency Analytics* [Press Release].

- Retrieved from <http://www.businesswire.com/news/home/20150930006279/en/>
- Chen, H., Chiang, R. H. L., & Storey, V. C. (2012). Business intelligence and analytics: From big data to big impact. *MIS Quarterly*, 36(4), 1165–1188.
- Cheung, Y.-H., & Ho, H.-Y. (2015). Social influence's impact on reader perceptions of online reviews. *Journal of Business Research*, 68(4), 883–887.
- Davenport, T. H. (2014). *Big data at work: Dispelling the myths, uncovering the opportunities*. Boston: Harvard Business Review Press.
- Fan, W., & Gordon, M. D. (2014). The power of social media analytics. *Communications of the ACM*, 57(6), 74–81.
- The Financial Brand. (2014). *Big data: Profitability, potential, and problems in banking*. Retrieved from <http://thefinancialbrand.com/38801/big-data-profitability-strategy-analytics-banking/>
- Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, 35(2), 137–144.
- Gartner. (2015, November 10). *Gartner says 6.4 billion connected things will be in use in 2016, up 30 percent from 2015*. Retrieved from <http://www.gartner.com/newsroom/id/3165317>
- Gayo-Avello, D. (2011). Don't turn social media into another 'Literary Digest' poll. *Communications of the ACM*, 54(10), 121–128.
- Goetz, H. (2014, March 26). What Google now can teach enterprises about big data. *Prolifig*. Retrieved from <http://prolifig.com/google-now-big-data/>
- Gonçalves, P., Araújo, M., Benevenuto, F., & Cha, M. (2013). Comparing and combining sentiment analysis methods. In *Proceedings of the 2013 Conference on Online Social Networks* (pp. 27–37). Boston: ACM.
- House, J. (2014, November 20). Big data analytics = Key to successful 2015 supply chain strategy. *ModusLink*. Retrieved from <https://www.moduslink.com/big-data-analytics-key-successful-2015-supply-chain-strategy/>
- IBM. (2010). *How big data is giving Hertz a big advantage*. Retrieved from <https://www-01.ibm.com/software/ebusiness/jstart/portfolio/hertzCaseStudy.pdf>
- IDC. (2015). *New IDC forecast sees worldwide big data technology and services market growing to \$48.6 billion in 2019, driven by wide adoption across industries* [Press Release]. Retrieved from <http://www.idc.com/getdoc.jsp?containerId=prUS40560115>
- Jamali, M., & Abolhassani, H. (2006). Different aspects of social network analysis. In *Proceedings – 2006 IEEE/WIC/ACM International Conference on Web Intelligence* (pp. 66–72). Hong Kong: IEEE.
- Kaplan, A. M., & Haenlein, M. (2010). Users of the world, unite! The challenges and opportunities of social media *Business Horizons*, 53(1), 59–68.
- Kwon, O., Lee, N., & Shin, B. (2014). Data quality management, data usage experience, and acquisition intention of big data analytics. *International Journal of Information Management*, 34(3), 387–394.
- Laney, D. (2001, February 6). 3D data management: Controlling data volume, velocity, and variety. *META Group*. Retrieved from <http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf>
- Lee, I., & Lee, K. (2015). The Internet of Things (IoT): Applications, investments, and challenges for enterprises. *Business Horizons*, 58(4), 431–440.
- Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., et al. (2011, May). Big data: The next frontier for innovation, competition, and productivity. *McKinsey*. Available at <http://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/big-data-the-next-frontier-for-innovation>
- Marlow, C. (2004). Audience, structure, and authority in the weblog community. In *Proceedings of the 54th Annual Conference of the International Communication Association*. Available at <http://alumni.media.mit.edu/~cameron/cv/pubs/04-01.pdf>
- O'Reilly, T. (2007). What is Web 2.0? Design patterns and business models for the next generation of software. *Communications and Strategies*, 65, 17–37.
- Park, J., Barash, V., Fink, C., & Cha, M. (2013). Emoticon style: Interpreting differences in emoticons across cultures. In *Proceedings of the 7th International AAAI Conference on Weblogs and Social Media* (pp. 466–475). Boston: AAAI.
- ReportsnReports. (2016, February 12). Social media analytics market to rise at 27.6% CAGR to 2020. *PRNewswire*. Retrieved from <http://www.prnewswire.com/news-releases/social-media-analytics-market-to-rise-at-276-cagr-to-2020-568584751.html>
- SAS. (2013). *Five big data challenges*. Retrieved from <https://www.sas.com/resources/asset/five-big-data-challenges-article.pdf>
- Scott, J. (2012). *Social network analysis* (3rd ed.). Thousand Oaks, CA: Sage.
- TRUSTe. (2015). *2015 US IoT Privacy Index [Infographic]*. Retrieved from <https://www.truste.com/resources/privacy-research/us-internet-of-things-index-2015/>
- van Rijmenam. (2016). Tesco and big data analytics, a recipe for success? *Datafloq*. Retrieved from <https://datafloq.com/read/tesco-big-data-analytics-recipe-success/665>
- Vesset, D., Olofson, C. W., Nadkorni, A., Zaidi, A., McDonough, B., Schubmehl, D., et al. (2015). Futurescape: Worldwide big data and analytics 2016 predictions. *IDC*. Retrieved from <https://www.idc.com/research/viewtoc.jsp?containerId=259835>