# Big Data and Business Intelligence: Debunking the Myths

CHRIS KIMBLE
AND GIANNIS MILOLIDAKIS

*Offering unprecedented levels of intelligence concerning the habits of consumers and rivals, big data promises to revolutionize the way enterprises are run. And yet, the concept of big data is one of the most poorly understood terms in business today. The implications for big data analytics are not as straightforward as they might seem—particularly when it comes to the so-called dark data from social media. Increases in the volume of data, the velocity with which they are generated and captured, and the variety of formats in which they are delivered all must be taken into account. To make the best use of the ever-burgeoning store of knowledge and insight at their fingertips, organizational leaders must confront two commonly held fallacies: that methodological issues no longer matter and that big data offers a complete and unbiased source of information on which to base their decisions. © 2015 Wiley Periodicals, Inc.*

Big data—the vast quantities of data that flow relentlessly from web sites, databases, information systems, mobile devices, social networks, and sensors—is one of the most hyped, and one of the most confusing, business terms in use today (Laney, LeHong, & Lapkin, 2013). While some say that big data will not only give businesses unprecedented insights into their customers' buying habits but also into their own internal processes, others contend that it heralds nothing less than a management revolution whereby technology replaces human judgment, enabling businesses to make better decisions more quickly and provide value for their customers in previously unimagined ways (McAfee & Brynjolfsson, 2012).

Dark data (Laney et al., 2013)—that is, data present in social media sites, such as Facebook, YouTube, and Twitter, which business fails to use effectively—holds a special promise in this respect. Firms use social media to interact with their customers and to build their brand's identity as well as to monitor their rivals. Social media sites can attract hundreds of millions of visitors and grow so quickly that statistics about their use becomes outdated before they reach the page. The growth of mobile communications, including cellphones and tablets, combined with "the Internet of things"—where Internet-enabled devices exchange data without human intervention—has further contributed to the mushrooming of such data.

These large datasets seem to offer the prospect of access to forms of knowledge and insight that were previously thought to be unobtainable. The value of such data to business appears to be unquestionable, and few, if any, would argue that it should be ignored. The real problem appears to be how to make the best use of it.

Businesses have always sought to glean intelligence from data and use it to gain competitive advantage. Today business intelligence has developed into a wide range of activities that organizational leaders undertake to understand their internal and external environment. The claims of unparalleled


Published online in Wiley Online Library (wileyonlinelibrary.com)

accuracy and objectivity made by the advocates of big data (Anderson, 2008) may not always be what they seem, however. Among the issues that must be addressed are:

- the tremendous volume of data,
- the ever-increasing speed with which data are produced,
- the growing variety of formats that are used to store and transmit data,
- the lack of transparency behind the methods with which data are collected,
- the complexity of subsequent data processing, and
- the human element, particularly when it comes to data obtained through social media.

### Business Intelligence and the Internet

The impact of big data is undeniable. Newspapers and academic journals are full of anecdotes and case studies that illustrate the value of such data to businesses. For example, McAfee and Brynjolfsson (2012) contrast a physical book shop—which can keep track of which books are sold and, through a loyalty program, can link some of those sales to individual customers—with an online store, such as Amazon. Online stores, can not only track, with almost total accuracy, what was sold to whom and when, but they can also track what else those customers looked at, how they navigated their way through the website, and how they were influenced by promotions and special offers. Furthermore, they can then use these data to predict what a customer might like to buy next.

Some, however, argue that the impact of big data is even more far reaching. For instance, McAfee and Brynjolfsson (2012) contend that big data even changes the way that businesses are managed and managers are rewarded. They argue that when data are scarce, highly placed people make decisions based on their intuition: the experience they have built up and the patterns they have internalized over their careers. Big data, they argue, will spell the end for HiPPOs—highest-paid person's opinions—as executive decisions become truly data driven. Some go further still and say that big data will make entire sectors of human knowledge obsolete. For example, Anderson writes: "Out with every theory of human behavior . . . Who knows why people do what they do? The point is they do it, and we can track and measure it with unprecedented fidelity" (2008).

> Businesses have always sought to glean intelligence from data and use it to gain competitive advantage.

The ability of a business to make use of the data that are available to it is sometimes termed *business intelligence*. Luhn (1958) first popularized the term when he used it to describe the abstracting, encoding, and archiving of internal documents and their dissemination using "data-processing machines." Later, the emphasis changed, and by the 1980s the ability to convert raw data into useful information for decision making was more highly stressed. Today, the term *business intelligence* is used to cover a broad range of intelligence regarding competitors, customers, markets, products, strategy, and technology and even business counterintelligence. Consequently, Gartner, a leading information technology research and advisory firm, now describes business intelligence as simply "an umbrella term that includes the applications, infrastructure and tools, and best practices that enable access to and analysis of information to improve and optimize decisions and performance" (Gartner, 2013).

The growth of the Internet at the turn of the last century provided businesses with a wealth of new data that could be used for business intelligence. Indeed, the web is considered the largest publicly accessible data source in the world; Google's search engine alone has indexed more than 45 billion websites (worldwidewebsize.com, 2015).

Twitter posts in excess of 500 million tweets a day (internetlivestats.com, 2015). Facebook claims to have more than 936 million active users a day (Facebook.com, 2015), while YouTube claims to have more than 1 billion unique visitors each month and more than 6 billion hours of video: almost an hour of video for every person on earth (YouTube.com, 2015).

Initially, however, the Internet was simply seen as a way to increase operational and financial efficiency by providing firms with a new channel to deal with their customers and suppliers. Consequently, organizations began to invest in Internet technologies merely as a means of communicating with suppliers and increasing their customer base. Nevertheless, as online markets began to grow, customers began to use the Internet in new ways: to express their opinions, or to seek the opinions of others, about the products and services that were on offer. According to Nielsen (2012), of the customers who engage with companies through social media channels such as Facebook or YouTube, 70 percent do so to learn of others' experiences; 65 percent do so to learn more about brands, products, or services; and 50 percent do so to express concerns or make complaints.

Online reputation now makes a clear impact on the bottom line, and the demand for consumer review and comparison vehicles have spawned the development of such popular sites as TripAdvisor.com. A study by The Kelsey Group and comScore (Kelsey, 2007) showed that consumers were willing to pay up to 20 percent more for services rated by other customers as 5-star in online reviews. Firms now actively encourage the users of social media to create reviews, initiate discussions, and make comments. A report by Burson-Marsteller Research on the Fortune Global Top 100 corporations' use of social media (Burson-Marsteller, 2012) showed that, in 2012, 87 of the firms used at least one social media platform, an increase of 8 percent from 2010. Recent research indicates that 93 percent of the Fortune Top 500 corporations used social media tools (Barnes, Lescault, & Augusto, 2014).

## Business Analytics and Business Intelligence: The Impact of the "Three Vs"

Clearly, business intelligence generated from big data could be of immense value; however, the current generation of analytics—the analysis of web-based data—is unable to keep up. A look at the "Three Vs"—volume, velocity, and variety—describing the changes related to the growth of e-commerce shows why. Although some have attempted to add extra "Vs," such as value, veracity, and viability, taken as a whole, the original three are sufficient to provide a comprehensive picture of the implications of big data for the gathering of business intelligence.

Recent research indicates that 93 percent of the Fortune Top 500 corporations used social media tools.

### Volume

Most intuitive definitions of big data focus on the volume of data that is being produced, often measured in terms of tera ($10^{12}$), peta ($10^{15}$), or exa ($10^{18}$) bytes or, more colloquially, by making comparisons to a more tangible repository of data, such as "X number of Libraries of Congress" (Johnston, 2012). Some say we are entering the Petabyte Age (Anderson, 2008), while others prefer to talk of how many exabytes of data are produced each day (McAfee & Brynjolfsson, 2012). Although volume is undoubtedly one aspect of big data, it is probably the least troublesome. As technology develops, what was big in the past will be normal tomorrow and probably thought of as quite small in the future.

According to Hendler (2013), volume originally referred to the amount of data held in large organizational databases. As businesses go about their

work, they inevitably generate data. As long ago as 1988, Zuboff noted that as information systems automate organizational processes, they also produce new information, making visible activities and events that were previously unseen. For example, data from supply chain applications have the potential to make each stage in a product's journey visible, no matter where the product is physically located. The volume of data that is generated within organizations will continue to grow inexorably as long as businesses use computers to manage their daily operations and engage in data gathering to support these activities.

More recently, however, the discussion has shifted from internal to external data, such as that found in web platforms. The volume of data available from the web has increased dramatically, thanks to technologies like data streaming, as well as everyday activities like sending videos, photos, and text messages. More recent developments—for example, context-aware applications that provide data about what users are doing, where they are located, whom they are with, and even, in the case of such devices as activity trackers, physiological data—have contributed to this trend. Much of these data are available to businesses through widely adopted application programming interfaces (APIs), so that businesses are now able to access an enormous volume of data about their customers, potential new customers, the market, and their competitors.

### Velocity

Whereas volume refers to what might be thought of as a "stock" of data, velocity refers to the rate at which that stock changes—for example, the speed at which data are generated, the frequency at which they are updated, or the rate at which they are delivered. Examples of high-velocity data include financial data from stock markets, real-time data from sensors and video cameras, and clickstream data generated by visitors to online stores. In extreme cases, such as streamed data, both the generation and delivery of data are continuous.

Those who are more agile and are the first to observe and exploit opportunities can gain significant competitive advantages (McAfee & Brynjolfsson, 2012); however, dealing with data velocity involves more than simply having sufficient bandwidth. An area of particular interest to business is being able to reduce the latency between the time when the data are created and when they are available to decision makers. In the case of Internet users in particular, real-time or close to real-time data can provide knowledge about incipient market trends, as well as highlight emergent issues concerning a brand's reputation.

Although a great deal of high-speed data, such as Twitter's infamous firehose, is in theory available to business through streaming APIs, deciding on which data to save is a challenge. At present, most businesses are able to view this type of data only through a 2- to 10-minute sliding window (ScaleDB, 2015).

### Variety

Although perhaps not as immediately obvious as volume or velocity, variety often poses the biggest problem for the analysis of big data. Variety refers to the number of different sources that data can come from and the formats, structures, and semantics that are associated with them (see **Exhibit 1**). Problems can occur because each different data source needs to be processed in a different way; therefore, although the data exist, they may not be structured in a way that makes them usable.

Structure refers to both the format in which the data are stored, such as the number and length of fields, and, more crucially, the semantics that need to be associated with those fields. For a computer to be able to process data in a way that is valid and meaningful for human beings, the data first need to be codified—that is a semantic value (a meaning) has to be allocated to each item of data (Kimble, 2013).

In many ways, this challenge is similar to the one faced in the early days of information systems, when

**Exhibit 1. Examples of Data Variety**

| | Structured Data | Unstructured Data |
|---|---|---|
| **Machine generated** | *Sensor data*: Data from RFID tags, smart meters, medical devices, GPS, or any sensors that automatically record data in a predefined way. | *Satellite images*: Including weather data or movement of tectonic plates, etc. |
| | *Web log data*: Operational data from servers, applications, network routers, etc., which collect data about their activity. | *Photographs and video*: Including security, surveillance, traffic video, etc. |
| | *Financial data*: Financial systems that generate data for stocks, bonds, etc., on a daily, hourly, or real-time basis. | *Radar or sonar data*: Including vehicular, meteorological, and oceanographic seismic profiles. |
| **Human generated** | *Input data*: Any kind of data that humans input into a computer. For example, forms, CRM systems, surveys, and questionnaires. | *Internal textual data of an organization*: Including e-mails, logs, survey results, and reports. |
| | *Click-stream data*: Generated by human interactions with websites. | *Social media data*: From social media platforms, such as Facebook, YouTube, Twitter, LinkedIn, and Flickr. |
| | *Data related to virtual environments*: Movement and actions of users in virtual worlds, such as SecondLife. | *Mobile data*: Including videos, pictures, text messages, and location. |

Adapted from Hurwitz, Nugent, Halper, and Kaufman, 2013.

businesses had to deal with data that had been generated by isolated pieces of software that had been built, in an uncoordinated way, to solve a variety of problems. The solution then was to develop relational databases in which the different formats and semantics could be combined under one master data schema. This, however, is not the solution to the problems associated with big data.

Although the term *unstructured*, as used in Exhibit 1, often characterizes data coming from certain sources, strictly speaking this is inaccurate. Data cannot be truly unstructured; some sort of structure must exist, as a result of either the way the data were produced or the way they are consumed. Something that is easy for a human to understand may pose severe difficulties for a machine. *Unstructured data*, therefore, is a term that is usually used to describe data whose information content is not readily amenable to automated analysis.

## Extracting Useful Intelligence From Big Data

Although analytics and business intelligence are clearly related, extracting business intelligence from big data is not as straightforward as it might seem. A review of some of the issues associated with big data analytics follows. It adopts the categorization of big data analytics into three approaches: BI&A 1.0 (Business Intelligence and Analytics 1.0), BI&A 2.0, and BI&A 3.0 (Chen, Chiang, & Storey, 2012). In addition, it distinguishes between the "unstructured" data that is principally derived from social media and other more structured forms of big data.

### A Simple Typology of Business Intelligence and Analytics

BI&A 1.0 has its roots in relational databases, statistical techniques, and data-mining techniques developed in the 1970s and 1980s. The data it deals with are mostly structured, internal data that have been collected by companies and stored in

commercial, relational database systems. Chen et al. (2012) note that most of the data-processing and analytical technologies for BI&A 1.0 have already been incorporated into the commercial business intelligence packages offered by major IT vendors.

> Data cannot be truly unstructured; some sort of structure must exist, as a result of either the way the data were produced or the way they are consumed.

BI&A 2.0 began to emerge at the turn of the twenty-first century as businesses began to move online and interact with their customers directly. A vast amount of company, industry, product, and customer information can be gathered by using various text and web-mining techniques. In addition to information held in traditional databases, detailed, user-specific logs can be collected through cookies and server logs that can be used to guide website design and product placement (Ting, Clark, & Kimble, 2009). Similarly, the analysis of customer transactions can be used to help understand market structure and generate product recommendations. Although proprietary solutions exist, Chen et al. (2012) note that at present, apart from basic query and search capabilities, no advanced analytics for unstructured data exist in commercially available business intelligence packages.

Finally, Chen et al. (2012) frame their discussion of BI&A 3.0 around the increasing use of mobile devices, such as the iPad, iPhone, and smartphones, and the development of ubiquitous computing, where such devices as televisions and automobiles can contain embedded processors. Such mobile, Internet-enabled devices, they argue, will soon be used to support location-aware, person-centered, context-sensitive services. They also note that no commercial BI&A 3.0 systems currently exist, and academic research on them is still in an embryonic state.

**Analytics and More Structured Forms of Data**

Dealing with the volume, velocity, and variety of big data is a major problem for those taking the BI&A 2.0 approach. Although technological solutions may be in sight, the ability to process large amounts of data does not mean that the data will be either relevant or useful.

Boyd and Crawford (2012) point out that Internet sources are prone to outages and losses and that any gaps and errors that result from them tend to be magnified when several datasets are merged. Corruption and loss of data are almost inevitable when dealing with large volumes of high-velocity data. Big data are not delivered into the hands of analysts pristine and ready for use but must first be cleaned and conditioned to make them suitable for processing. The opaque and under-documented way in which data are gathered also raises doubts about the supposed completeness and accuracy of big data (Ekbia et al., 2015).

From a slightly different viewpoint, Boyd and Crawford (2012) question the validity of the statistical techniques that are often used to analyze big data. To be able to use a statistical test to make claims about data, we need to know the properties of the data: where they come from, their distribution, and their weaknesses and biases. Simply because a dataset contains billions of items does not mean that it is either random or representative. Without knowing how the data were collected and processed, it is not possible to know whether the assumptions upon which the tests are based have been violated. Ekbia et al. (2015) go further claiming that because many of the tests that are used were designed to overcome the problems associated with small samples their use with big data leads to apophenia: seeing patterns where none exist. They conclude that rather than removing the traditional dilemmas faced by analysts about what can legitimately be claimed from data, dealing with big data has actually made them worse.

### Analytics and Less Structured Data From Social Media

Despite the recent emergence of the notion of the Internet of things, the content of the Internet is still primarily created by people. Often credited with the invention of the World Wide Web, Tim Berners-Lee said, "The web is more a social creation than a technical one. I designed it for a social effect—to help people work together—not as a technical toy" (Berners-Lee & Fischetti, 1999, p. 123). The early web was characterized by static webpages providing one-way communication, often termed Web 1.0. The technologies that dominate the Internet today, known as Web 2.0, allow the creation and modification of content by groups of people, as well as the combination and reuse of data from different applications. The most prominent of these are social networking sites created to serve groups of people who share common interests.

Web 2.0 has changed the way people interact online and has led to the formation of what have become known as virtual communities. The growth in the use of social media is an almost inevitable consequence of this. Effectively, the web has become a medium for human communication, with all the subjectivity, confusion, misunderstandings, misinterpretations, and deliberate deception this entails. These communities form to share knowledge, opinions, and experiences about products and services. Despite the potential value of this information to business, Patterson (2012) notes that existing analytics tend to be limited to quantitative assessments, such as how many times a brand is mentioned (see **Exhibit 2**).

Metrics based on simple counts of activities are unlikely to provide any deeper understanding of the interactions that take place. Such measures treat social interactions as unproblematical quantitative data and risk oversimplifying the rich and dynamic nature of the communication that takes place. As an example, Ekbia et al. (2015) cite the experience of the British Broadcasting Corporation (BBC) when it unveiled an initiative to "map the mood of the nation" by classifying Twitter feeds according to eight basic human emotions. They ask, "even if we assume that human emotions can be meaningfully reduced to eight basic categories (what of complex emotions such as grief, annoyance, contentment, etc.?) . . . how does one differentiate the 'happiness' of the fans of Manchester United after a winning game from the expression of the 'same' emotion by the admirers of the Royal family on the occasion of the birth of the heir to the throne?" (p. 8).

**Exhibit 2. Common Social Media Metrics**

| Metric | Description |
|---|---|
| Channel distribution | Calculated across several platforms to see which brands are the subjects of discussion, and on which platforms. |
| Engagement | Indicates the level of involvement of users in the brand, usually measured by the number of likes, followers, shares, tweets, etc. |
| Geography | Indicates the geographical origin of comments based on information provided by users or via IP addresses/GPS sensors. |
| Influencer ranking | A measurement of the popularity of users that create content referring to a specified brand; calculated on the number of connections that user has. |
| Sentiment | Indicates the attitude toward the brand using linguistic algorithms that identify positive and negative words. |
| Topic and theme detection | Information relating to a specific brand concerning the nature of a topic that was discussed; allows popular topics to be identified. |
| Volume of posts | Indicates the number of items of user-generated content (e.g., blog posts, articles, or videos) that contain the name of the brand. |

In addition, these measures are blind to the playful, creative, unusual, and sometimes eccentric ways in which people use social media. The content of social media should give businesses access to information about their customers' opinions, ideas, thoughts, and feelings. Moving beyond the generation of simple quantitative measures, however, poses some difficult practical and philosophical questions.

### Social Media and the Nature of Human Communication

The Austrian mathematician and philosopher Ludwig Wittgenstein spent the first part of his life searching for stable, ideal meanings for words in an attempt to define the principles of language using logic. He later rejected the notion of a language in which words had meanings that were unique, identifiable, and stable and instead claimed that language could be understood only in the context in which it was used. He argued that linguistic terms arise from social conventions created by people rather than by reference to some objective external reality. He saw language as a game in which the rules were created as it was played; consequently, the only way to understand the rules was to participate in the game. Thus, "linguistic meaning is never complete and final . . . it is unstable and open to potentially infinite interpretation and reinterpretation in an unending play of substitution" (Marshall & Brady, 2001, p. 101).

Viewed in this way, the players of Wittgenstein's language game can be seen as communities whose practices provide the only fixed point against which the meanings given to words can be anchored. His observation—that the semantics of words and images are inextricably rooted in the life experience of the people who use them—poses the greatest problem for the analysis of data from social media.

Different communities will view the same thing in different ways and use different words or images to describe it. Similarly, the same words or images may have quite different meanings in different communities. For example, the meaning attached to the logos of established brands can be parodied to give them a quite different meaning. As Petty (2012) notes, while regular searches for the use of a brand name will uncover uses that spell the name correctly, they will not uncover misspellings that seek to poke fun at or create a negative image of the brand.

The content of social media should give businesses access to information about their customers' opinions, ideas, thoughts, and feelings.

### Analyzing Social Media

As Chen et al. (2012) observed, research on BI&A 3.0 is still in an embryonic stage. This is particularly the case when it comes to social media. Although it is possible to use targeted image recognition and text analysis software to highlight the misuse of brand names and registered logos, the real value of social media data lies not in protecting established trademarks but in discovering new ideas and identifying emerging trends. The question remains, however: If these trends and ideas are truly novel, and they are expressed using a new or unknown terminology, how could they even be identified using conventional approaches?

Currently, there are a plethora of methods used for analyzing social media, including social network analysis, text and web mining, natural language processing, and sentiment analysis. The results of such analyses, however, are often limited to the constraints of the particular analytical tool that was used or to a particular source of data (Milolidakis, Akoumianakis, Kimble, & Karadimitriou, 2014b). Thus, although social media users rarely restrict their activities to one platform, an analysis of Facebook using social network analysis may not recognize that the same user is cross-posting similar material to Twitter. This points up the need for some form of methodological protocol that

combines data from a range of analytical tools and data sources so that changing patterns of meanings can be tracked across time and across media (Milolidakis, Akoumianakis, & Kimble, 2014a).

Jones (2003) offers one such approach, based on traditional archeology, that he terms *cyber-archeology*. Jones set out to develop a methodology to study online public interactions that was neither culture- nor time-specific. He and Rafaeli (2000) note that traditional archaeology provides a perspective for studying the process of cultural change and identifying the gradual impact of changes in a community's behavior over time. They argue that cyber-archeology has a similar potential and that, like the excavation of archaeological tells (the mounds of debris that accumulate around human settlements), the excavation of virtual tells (the digital traces left by virtual communities) can reveal what has taken place in those communities.

This approach has been adopted in a number of studies on such diverse topics as support groups for people with various types of cancer (Akoumianakis, Karadimitriou, Vlachakis, Milolidakis, & Bessis, 2012), fan pages of Greek telecommunication companies on Facebook (Milolidakis et al., 2014a), and cross-platform Facebook and YouTube use by telecommunication companies (Milolidakis et al., 2014b). It is important to stress, however, that, like traditional archaeology, this approach tends to be slow and labor intensive, as much of the work involves interpretation rather than the automated processing of data. Nevertheless, as Boyd and Crawford (2012) point out, as soon as an analyst starts to ask what the data mean, regardless of the source, the process of interpretation by human beings begins.

Cyber-archeology is not a panacea, however. In the same way that modern archaeological techniques developed from the work of a few individuals in the eighteenth and ninteenth centuries, more work is needed to develop protocols for the excavation of data from virtual settlements in the twenty-first century. In addition, technological factors also limit the usefulness of this approach. For example, the weak and incomplete archiving associated with some social media and the limitations of the associated APIs limit the number of layers of context and meaning that can be uncovered. Thus, while Wikipedia maintains meticulous records of changes to its content, it offers only basic APIs for exporting that content. On the other hand, Twitter offers a range of ways to access content but has only weak archiving facilities.

### Debunking Big Data

Big data appears to offer businesses the possibility of obtaining unparalleled insights into customers' needs and competitors' strategies; it also seems set to transform the way in which businesses are run, with hard data rather than intuition-driving decisions. Big data, analytics, business intelligence, and the Internet have aligned to usher in a brave new world.

Dealing with big data is not a simple matter not only because of the sheer quantity of data that are now being generated but also because of the speed and the variety of formats with which they are delivered. Meanwhile, dark data from social media, whose patterns are invisible to the human eye, pose particular problems because of the interpretive flexibility of words and images and the mischievous tendency of human language to morph and change over time.

There is no doubt that advances in technology will help overcome some of these problems, particularly those associated with the handling of large volumes of high-velocity data. It is also probable that some of the problems associated with the range of formats in which data are supplied will be overcome in due course. Like human language, however, standards and formats change over time and "standards wars" are natural features of competition as

companies struggle to establish the preeminence of one standard over another to gain or maintain a dominant position in the market.

What, then, does this say about the use and value of big data to business? Having reviewed the limits of big data, there is clearly a need to de-bunk some of the myths that surround it.

> Dealing with big data is not a simple matter not only because of the sheer quantity of data that are now being generated but also because of the speed and the variety of formats with which they are delivered.

First, big data does not provide easy answers. Boyd and Crawford (2012) note that Anderson's sweeping dismissal of all other theories reflects an undercurrent present in many discussions of big data, in which all other forms of analysis are scorned. Commenting on managerial judgment, Spender (2014) observes that whereas managers were once expected to manage under determined situations, now with big data and the trend toward IT-intensive practices, they are expected to deal rationally with determinable situations. Such an approach is viable, he maintains, only if we believe that what was under-determined in the past has now become fully determined, calculable, and forecastable.

Ekbia et al. (2015) argue that big data has led to a shift from causal explanations toward predictive modeling and simulation. Echoing the words of microbiologist Carl Woese, they warn that while this might show us how to get there, it will not tell us where "there" is. To those, such as Anderson (2008), who argue, "Who knows why people do what they do? The point is they do it," this may seem to be of little importance. Experience has shown, however, that data taken out of context lose their meaning and value, and when large datasets are turned into mathematical models, they are inevitably decontextualized and reduced to what will fit into such models. The risk is that big data will provide accurate, but essentially meaningless, answers.

Second, there is a need to de-bunk the belief in the supposed objectivity of big data. There are several technological and methodological reasons why big data may not be as complete and objective as it seems, which should be apparent to even a casual user of social media. For example, the use of Facebook's "Like" button, which is taken as an indication of approval, can easily be manipulated by such offers as "Like our product and enter a draw to win a luxury holiday." Similarly, Boyd and Crawford (2012) note that "Twitter does not represent 'all people,' and it is an error to assume 'people' and 'Twitter users' are synonymous" (p. 669). Some people have multiple Twitter accounts, while some Twitter accounts are used by multiple people; and some "people" are not people at all, but automated bots that generate content without direct human intervention. Unless we believe that big data sweeps away the need for methodology, observations such as these should give us cause to question the objectivity of any data we receive.

Similarly, one of the arguments for using big data from social media is that the data have all been made freely and publicly available, so there is no need to ask permission to use it. Leaving aside the issues of anonymity and who actually has access to the data, potential legal and ethical concerns arise. Laney (2012) calls Facebook's users "the largest unpaid workforce in history," indicating that the average value of the data posted on Facebook comes to approximately $81 per person. It is reasonable to assume that many of Facebook's users are unaware of how the information that they have posted will be used or of the profits and other gains that will flow from it. There is no doubt that legal issues concerning the use of social media data, as well as commercial judgments, will affect what is made available and to whom in the future, further

undermining the supposed completeness and objectivity of big data.

Thanks to big data, business leaders can now make predictions that are faster and more accurate than before and possibly use that information to make better-informed decisions. It is equally clear, however, that the blind enthusiasm with which some have taken up the cause of big data risks undermining any potential gains. Neither big data nor technological wizardry alone will solve the challenges of capturing and getting the most from information. Rather, leaders will need to bring both human wisdom and technological prowess to bear on the complex of data-driven issues they face.

## References


Akoumianakis, D., Karadimitriou, N., Vlachakis, G., Milolidakis, G., & Bessis, N. (2012). Internet of things as virtual settlements: Insights from excavating social media sites. In Proceedings of the 2012 Fourth International Conference on Intelligent Networking and Collaborative Systems (pp. 132–39). Bucharest, Romania: IEEE Computer Society.

Anderson, C. (2008). The end of theory: The data deluge makes the scientific method obsolete. Retrieved July 2015, from http://archive.wired.com/science/discoveries/magazine/16-07/pb_theory

Barnes, N. G., Lescault, A. M., & Augusto, K. D. (2014). LinkedIn dominates, Twitter trends and Facebook falls: The 2014 Inc. 500 and social media. Retrieved August 2015, from http://www.umassd.edu/cmr/socialmediaresearch/2015fortune500andsocialmedia/

Berners-Lee, T., & Fischetti, M. (1999). Weaving the web: The original design and ultimate destiny of the World Wide Web by its inventor. San Francisco, CA: Harper.

Boyd, D., & Crawford, K. (2012). Critical questions for big data. Information, Communication & Society, 15(5), 662–679.

Burson-Marsteller. (2012). Global social media check-up 2012. Retrieved July 2015, from http://sites.burson-marsteller.com/social/Presentation.aspx

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.

Ekbia, H., Mattioli, M., Kouper, I., Arave, G., Ghazinejad, A., Bowman, T., Suri, V. R., Tsou, A., Weingart, S., & Sugimoto, C. R. (2015). Big data, bigger dilemmas: A critical review. Journal of the Association for Information Science and Technology, 66(8), 1523–1545.

Facebook.com. (2015). Stats. Retrieved July 2015, from http://newsroom.fb.com/company-info/

Gartner. (2013). Business intelligence. Gartner IT Glossary, Retrieved from http://www.gartner.com/it-glossary/business-intelligence-bi/

Hendler, J. (2013). Broad data: Exploring the emerging web of data. Big Data, 1(1), 18–20.

Hurwitz, J., Nugent, A., Halper, F., & Kaufman, M. (2013). Big data for dummies. Hoboken, NJ: Wiley.

internetlivestats.com. (2015). Twitter statistics. Retrieved July 2015, from http://www.internetlivestats.com/twitter-statistics/

Johnston, L. (2012). How many Libraries of Congress does it take? Digital preservation. Retrieved July 7, 2015, from http://blogs.loc.gov/digitalpreservation/2012/03/how-many-libraries-of-congress-does-it-take/

Jones, Q. (2003). Applying cyber-archaeology. In K. Kuutti, E. H. Karsten, G. Fitzpatrick, P. Dourish & K. Schmidt (Eds.), ECSCW 2003: In Proceedings of the Eighth European Conference on Computer Supported Cooperative Work (pp. 41–60). Helsinki, Finland: Springer.

Jones, Q., & Rafaeli, S. (2000). What do virtual "tells" tell? Placing cybersociety research into a hierarchy of social explanation. In Proceedings of the 33rd Annual Hawaii International Conference on System Sciences, 2000, IEEE Press.

Kelsey. (2007). Online consumer-generated reviews have significant impact on offline purchase behavior. Retrieved July 2015, from http://www.comscore.com/Insights/Press_Releases/2007/11/Online_Consumer_Reviews_Impact_Offline_Purchasing_Behavior

Kimble, C. (2013). Knowledge management, codification and tacit knowledge. Retrieved 19 June 2013, from http://informationr.net/ir/18-2/paper577.html

Laney, D. (2012). To Facebook you're worth $80.95. Wall Street Journal, CIO report. Retrieved July 2015, from http://blogs.wsj.com/cio/2012/05/03/to-facebook-youre-worth-80-95/

Laney, D., LeHong, H., & Lapkin, A. (2013, May 6). What big data means for business. Financial Times. Retrieved September 2015, from http://www.ft.com/intl/cms/s/0/b1dec7f4-b686-11e2-93ba-00144feabdc0.html

Luhn, H. P. (1958). A business intelligence system. IBM Journal of Research and Development, 2(4), 314–319.

Marshall, N., & Brady, T. (2001). Knowledge management and the politics of knowledge: illustrations from complex products and systems. European Journal of Information Systems, 10, 99–112.

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

Milolidakis, G., Akoumianakis, D., & Kimble, C. (2014a). Digital traces for business intelligence: A case study of mobile telecoms service brands in Greece. Journal of Enterprise Information Management, 27(1), 66–98.

Milolidakis, G., Akoumianakis, D., Kimble, C., & Karadimitriou, N. (2014b). Excavating business intelligence from social media. In W. John (Ed.), Encyclopedia of business analytics and optimization (pp. 897–908). Hershey, PA: IGI Global.

Nielsen. (2012). State of the media: The social media report. New York: Nielsen Holdings.

Patterson, A. (2012). Social-networkers of the world, unite and take over: A meta-introspective perspective on the Facebook brand. Journal of Business Research, 65(4), 527–534.

Petty, R. D. (2012). Using the law to protect the brand on social media sites. Management Research Review, 35(9), 758–769.

ScaleDB. (2015). High-velocity data—The data fire hose. Retrieved July 2015, from http://www.scaledb.com/high-velocity-data.php

Spender, J. (2014). A rumination on managerial judgment. Revue française de gestion, 40(238), 19–32.

Ting, I.-H., Clark, L., & Kimble, C. (2009). Identifying web navigation behaviour and patterns automatically from clickstream data. International Journal of Web Engineering and Technology, 5(4), 398–426.

worldwidewebsize.com. (2015). The size of the World Wide Web (the Internet). Retrieved July 2015, from http://www.worldwidewebsize.com/

YouTube.com. (2015). Statistics—YouTube. Retrieved July 2015, from http://www.youtube.com/yt/press/statistics.html

Zuboff, S. (1988). In the age of the smart machine. New York: Basic Books.

---

***Chris Kimble**, senior academic editor of* Global Business and Organizational Excellence*, is an associate professor of strategy and technology management at KEDGE Business School and is affiliated to the MRM Laboratory at Université Montpellier. Before moving to France, he lectured in the United Kingdom on information systems and management at the University of York and on information technology at the University of Newcastle. He was a researcher with both the Business School and the Department of Informatics at the University of Northumbria. His broad fields of research are business strategy and the management of the fit between the digital and social worlds. He can be reached at chris.kimble@kedgebs.com.*

***Giannis Milolidakis** is a PhD student at KEDGE Business School in Marseille, France. He holds a BSc degree in applied information technology and multimedia from the Technological Education Institution of Crete. His research focuses on social media and business intelligence, and their implications for information systems.*

---