1

Introduction to the Big Data Era

Stephan Kudyba and Matthew Kwatinetz

CONTENTS

By now you've heard the phrase "big data" a hundred times and it's intrigued you, scared you, or even bothered you. Whatever your feeling is, one thing that remains a source of interest in the new data age is a clear understanding of just what is meant by the concept and what it means for the realm of commerce. Big data, terabytes of data, mountains of data, no matter how you would like to describe it, there is an ongoing data explosion transpiring all around us that makes previous creations, collections, and storage of data merely trivial. Generally the concept of big data refers

to the sources, variety, velocities, and volumes of this vast resource. Over the next few pages we will describe the meaning of these areas to provide a clearer understanding of the new data age.

The introduction of faster computer processing through Pentium technology in conjunction with enhanced storage capabilities introduced back in the early 1990s helped promote the beginning of the information economy, which made computers faster, better able to run state-of-the-art software devices, and store and analyze vast amounts of data (Kudyba, 2002). The creation, transmitting, processing, and storage capacities of today's enhanced computers, sensors, handheld devices, tablets, and the like, provide the platform for the next stage of the information age. These super electronic devices have the capabilities to run numerous applications, communicate across multiple platforms, and generate, process, and store unimaginable amounts of data. So if you were under the impression that big data was just a function of e-commerce (website) activity, think again. That's only part of the very large and growing pie.

When speaking of big data, one must consider the source of data. This involves the technologies that exist today and the industry applications that are facilitated by them. These industry applications are prevalent across the realm of commerce and continue to proliferate in countless activities:

- Marketing and advertising (online activities, text messaging, social media, new metrics in measuring ad spend and effectiveness, etc.)
- Healthcare (machines that provide treatment to patients, electronic health records (EHRs), digital images, wireless medical devices)
- Transportation (GPS activities)
- Energy (residential and commercial usage metrics)
- Retail (measuring foot traffic patterns at malls, demographics analysis)
- Sensors imbedded in products across industry sectors tracking usage

These are just a few examples of how industries are becoming more data intensive.

DESCRIPTION OF BIG DATA

The source and variety of big data involves new technologies that create, communicate, or are involved with data-generating activities, which produce

different types/formats of data resources. The data we are referring to isn't just numbers that depict amounts, or performance indicators or scale. Data also includes less structured forms, such as the following elements:

- Website links
- Emails
- Twitter responses
- Product reviews
- Pictures/images
- Written text on various platforms

What big data entails is structured and unstructured data that correspond to various activities. Structured data entails data that is categorized and stored in a file according to a particular format description, where unstructured data is free-form text that takes on a number of types, such as those listed above. The cell phones of yesteryear have evolved into smartphones capable of texting, surfing, phoning, and playing a host of software-based applications. All the activities conducted on these phones (every time you respond to a friend, respond to an ad, play a game, use an app, conduct a search) generates a traceable data asset. Computers and tablets connected to Internet-related platforms (social media, website activities, advertising via video platform) all generate data. Scanning technologies that read energy consumption, healthcare-related elements, traffic activity, etc., create data. And finally, good old traditional platforms such as spreadsheets, tables, and decision support platforms still play a role as well.

The next concept to consider when merely attempting to understand the big data age refers to velocities of data, where velocity entails how quickly data is being generated, communicated, and stored. Back in the beginning of the information economy (e.g., mid-1990s), the phrase "real time" was often used to refer to almost instantaneous tracking, updating, or some activities revolving around timely processing of data. This phrase has taken on a new dimension in today's ultra-fast, wireless world. Where real time was the goal of select industries (financial markets, e-commerce), the phrase has become commonplace in many areas of commerce today:

- Real-time communication with consumers via text, social media, email
- Real-time consumer reaction to events, advertisements via Twitter
- Real-time reading of energy consumption of residential households
- Real-time tracking of visitors on a website

Real time involves high-velocity or fast-moving data and fast generation of data that results in vast volumes of the asset. Non-real-time data or sources of more slowly moving data activities also prevail today, where the volumes of data generated refer to the storage and use of more historic data resources that continue to provide value. Non-real time refers to measuring events and time-related processes and operations that are stored in a repository:

- Consumer response to brand advertising
- Sales trends
- Generation of demographic profiles

As was mentioned above, velocity of data directly relates to volumes of data, where some real-time data quickly generate a massive amount in a very short time. When putting an amount on volume, the following statistic explains the recent state of affairs: as of 2012, about 2.5 exabytes of data is created each day. A petabyte of data is 1 quadrillion bytes, which is the equivalent of about 20 million file cabinets' worth of text, and an exabyte is 1000 times that amount. The volume comes from both new data variables and the amount of data records in those variables.

The ultimate result is more data that can provide the building blocks to information generation through analytics. These data sources come in a variety of types that are structured and unstructured that need to be managed to provide decision support for strategists of all walks (McAfee and Brynjolfsson, 2012).

BUILDING BLOCKS TO DECISION SUPPORT

You may ask: Why are there classifications of data? Isn't data simply data? One of the reasons involves the activities required to manage and analyze the resources that are involved in generating value from it. Yes, big data sounds impressive and almost implies that value exists simply in storing it. The reality is, however, that unless data can help decision makers make better decisions, enhance strategic initiatives, help marketers more effectively communicate with consumers, enable healthcare providers to better allocate resources to enhance the treatment and outcomes of their patients, etc., there is little value to this resource, even if it is called big.

Data itself is a record of an event or a transaction:

A purchase of a product
A response to a marketing initiative
A text sent to another individual
A click on a link

In its crude form, data provides little value. However, if data is corrected for errors, aggregated, normalized, calculated, or categorized, its value grows dramatically. In other words, data are the building blocks to information, and information is a vital input to knowledge generation for decision makers (Davenport and Prusak, 2000). Taking this into consideration, the "big" part of big data can actually augment value significantly to those who use it correctly. Ultimately, when data is managed correctly, it provides a vital input for decision makers across industry sectors to make better decisions.

So why does big data imply a significant increase in the value of data? Because big data can provide more descriptive information as to why something has happened:

Why and who responded to my online marketing initiative?
What do people think of my product and potentially why?
What factors are affecting my performance metrics?
Why did my sales increase notably last month?
What led my patient treatment outcomes to improve?

SOURCE OF MORE DESCRIPTIVE VARIABLES

Big data implies not just more records/elements of data, but more data variables and new data variables that possibly describe reasons why actions occur. When performing analytics and constructing models that utilize data to describe processes, an inherent limitation is that the analyst simply doesn't have all the pertinent data that accounts for all the explanatory variance of that process. The resulting analytic report may be missing some very important information. If you're attempting to better understand where to locate your new retail outlet in a mall and you don't have detailed shopper traffic patterns, you may be missing some essential

descriptive information that affects your decision. As a result, you locate your store in what seems to be a strategically appropriate space, but for some reason, the traffic for your business just isn't there. You may want to know what the market thinks of your new product idea, but unfortunately you were only able to obtain 1000 responses to your survey of your target population. The result is you make decisions with the limited data resources you have. However, if you text your question to 50,000 of your target population, your results may be more accurate, or let's say, more of an indication of market sentiment.

As technology continues to evolve and become a natural part of everyone's lives, so too does the generation of new data sources. The last few years have seen the explosion of mobile computing: the smartphone may be the most headlining example, but the trend extends down to your laundry machine, sprinkler system, and the label on the clothing that you bought retail. One of the most unexpected and highest impact trends in this regard is the ability to leverage data variables that describe activities/processes. We all know that technology has provided faster, better computers—but now the trend is for technology to feed in the generation of never before seen data at a scale that is breathtaking. What follows are some brief examples of this.

The following illustrations depict the evolution of big data in various industry sectors and business scenarios. Just think of the new descriptive variables (data resources) that can be analyzed in these contemporary scenarios as opposed to the ancient times of the 1990s!

INDUSTRY EXAMPLES OF BIG DATA

Electioneering

In some recent political campaigns, politicians began to mobilize the electorate in greater proportion than ever before. Previously, campaign managers had relied unduly on door-to-door recruiting, flyering in coffee shops, rallies, and telemarketing calls. Now campaigns can be managed completely on the Internet, using social network data and implied geographic locations to expand connectivity between the like-minded. The focus is not just on generating more votes, but has extended to the

ever-important fund-raising initiatives as well. Campaigners are able to leverage the power of big data and focus on micro-donations and the viral power of the Internet to spread the word—more dollars were raised through this vehicle than had been seen in history. The key function of the use of the big data allowed local supporters to organize other local supporters, using social networking software and self-identified zip code and neighborhood locations. That turned data resources *locational*, adding a new dimension of information to be exploited, polled, and aggregated to help determine where bases of support were stronger/weaker. Where will it go next? It is likely that in the not-so-distant future we will find voter registrations tagged to mobile devices, and the ability to circumvent statistical sampling polls with actual polls of the population, sorted by geography, demography, and psychographics. Democratic campaign managers estimate that they collected 13 million email addresses in the 2008 campaign, communicating directly with about 20% of the total votes needed to win. Eric Schmidt (former CEO of Google) says that since 2008, the game has completely changed: "In 2008 most people didn't operate on [Facebook and Twitter]. The difference now is, first and foremost, the growth of Facebook, which is much, much more deeply penetrated . . . you can run political campaigns on the sum of those tools [Facebook, YouTube and Twitter]" (quotes from *Bloomberg Business Week*, June 18–24, 2012; additional info from Tumulty, 2012).

Investment Diligence and Social Media

"Wall Street analysts are increasingly incorporating data from social media and Internet search trends into their investment strategies" ("What the Experts Say," 2012). The use of social media data is generally called unstructured data. Five years ago, surveys showed that approximately 2% of investment firms used such data—today "that number is closer to 50 percent" (Cha, 2012). The World Economic Forum has now classified this type of data as an economic asset, and this includes monitoring millions of tweets per day, scanning comments on buyer sites such as Amazon, processing job offerings on TheLadders or Monster.com, etc. "Big data is fundamentally changing how we trade," said financial services consultant Adam Honore (adhonore, http://www.advancedtrading.com/Adam-Honore). Utilizing the number and trending features of Twitter, Facebook, and other media platforms, these investors can test how "sticky" certain products, services, or ideas are in the country. From this information, they

can make investment choices on one product vs. another—or on the general investor sentiment. This information does not replace existing investment diligence, but in fact adds to the depth and quality (or lack thereof sometimes!) of analysis.

Real Estate

Investment dollars in the capital markets are split between three main categories, as measured by value: bonds, stocks, and alternative assets, including real estate. Since bonds were traded, an informal network of brokers and market makers has been able to serve as gateways to information, given that many transactions go through centralized clearinghouses. In 1971, NASDAQ was the first stock market to go electronic, and as the information revolution continued, it soon allowed for any person around the world to sit at the hub of cutting-edge news, information, and share prices. After a particular tech-savvy Salomon Brothers trader left that company, he led the further digitization of data and constant updating of news to create a big data empire: Michael Bloomberg. Real estate, however, has been late to the game. To understand real estate prices in any given market has been more challenging, as many transactions are private, and different cities and counties can have significantly different reporting mechanisms and data structures. Through the late 1980s and 1990s, real estate was often tracked in boxes of files, mailed back and forth across the country. As cities began to go digital, a new opportunity was created. In the year 2000, Real Capital Analytics (http://www.rcanalytics.com) was founded by Robert White to utilize data mining techniques to aggregate data worldwide on real estate transactions, and make that data available digitally. Real estate research firms have many techniques to acquire data: programmatically scraping websites, taking feeds from property tax departments, polling brokerage firms, tracking news feeds, licensing and warehousing proprietary data, and more. All of these sources of data can be reviewed on an hourly basis, funneled through analysis, and then displayed in a user-friendly manner: charts, indices, and reports that are sorting hundreds of thousands of daily data points.

Specialized Real Estate: Building Energy Disclosure and Smart Meters

Over 40% of energy use and carbon emissions in the United States come from existing buildings (http://www.eia.gov/consumption/commercial/index.cfm). To put this in perspective, if you combined the energy use and emissions output of all of the SUVs on the road in North America, this would be approximately 3%. So you can see that the use of energy by existing buildings is a very important piece of data. Until recently, this data has been held in many different databases for utilities across the country, with no central repository or easy means for reconciling these data sets. Today, three trends have picked up: (1) energy disclosure ordinances, (2) satellite heat map data, and (3) data warehousing aggregations based on smart meters. The amount of data needed here to control for effective information is staggering: any analysis must account for building size, use, geographic location, seasonality, climactic variation, occupancy, etc. In many of these cases, information is collected on a granularity of 1–15 minutes! That is for every building, in every city, in every state in the country: billions of data points *per day* (http://www.eebhub.org/).

Commerce and Loyalty Data

When you walk into your favorite retail outlet—be it clothing, jewelry, books, or food—there is nothing quite as gratifying as being recognized, your tastes being understood, and receiving personal service ("The usual, please!"). In the distant past, this probably meant a neighborhood shop where you literally were known by the salesperson. In the 1990s this was transformed into a "loyalty program" craze in which large-scale (franchised, national, international) retailers were able to tag your shopping to a digital ID card that they enticed you to use by offering discounts. But Internet commerce, under the thought leadership of Amazon, transformed this experience entirely. Once you are online, not only can a retailer track your purchases, but it can track what products you look at, things you plan to buy (wish lists), items you buy for others (registry), and even what pages you spend time on and for how long. This provides retailers with a competitive advantage: they can tailor your shopping experience and suggest new products. Witness Netflix recommendations, Pandora's preference algorithms, and LinkedIn's suggestion of who you might next want

to connect with or apply for a job from. Moreover, it is not just information from their own site that these online merchants can now pull from—the trend has now reclaimed the point-of-sale data from brick-and-mortar stores as well. Retailers integrate physical data with online point-of-sale data, and can also view what other sites you visit, where else you make purchases, who makes purchases for you, and what "like-minded shoppers" may be in your network.

Crowd-Sourced Crime Fighting

In an effort to aid local policing efforts, policing has found a new ally: you! Over the last decade "hot spot" policing has become the effective leading strategy for reducing crime: take careful record of where crime occurs, measure density regions, and overwhelm the highest density regions with extremely quick and overpowering responses. However, this strategy still relies on actually being able to track all of the crime incidents—no small task, as the force's personnel have limited resources. Enter the crowd sourcing platforms. Some cities have created apps for mobile devices (or other interfaces) that allow individual citizens to upload information that indicates crimes they have witnessed (http://spotcrime.com/ga/augusta)! The upload contains the description of the crime, a geographic location, and a time stamp. As participation increases, so too do "eyes on the street," and the map is filled with the information needed to improve police performance.

Pedestrian Traffic Patterns in Retail

Thanks to some recent controversies, you probably already know that your cell phone allows you to be tracked at nearly any time of day, provided it is powered on. While privacy laws currently still protect you from being identified with this feature (without your opting in), new technologies are available to identify unique movements. Cell tower "repeaters" in strategic locations in malls and downtowns can track "unique cell phones" and their walking patterns. As a result, a mall owner might want to know how many people take the elevator vs. the stairs—and of the ones who take the elevator, do they ever walk by the store on the other side of it? Further, if they find a patron lingering in the leather goods section of the store for more than 12 minutes, but that customer does not stop at the cash register, they will send a text message advertisement

promotion to the customer's phone before he or she leaves the store, offering a discount on—you guessed it—leather goods. This is only the beginning of this technology. Expect to see it deployed in cities to track crime patterns, the safety of certain intersections, and more (http://techcrunch.com/2007/12/14/path-intelligence-monitors-foot-traffic-in-retail-stores-by-pinging-peoples-phones/; http://allthingsd.com/20111103/ex-googlers-raise-5-8-million-to-help-retailers-track-foot-traffic/).

Intelligent Transport Application

New applications being developed for smartphones pool voluntarily offered information from unique sources into a real-time database providing an instant advantage from the use of big data. Uber, a mobile phone-based transportation application, connects drivers (of limousines, taxis) with potential passengers. As each driver "opts in" to uber from his or her phone, the phone sends a GPS signal update to the master Uber map. When a passenger is ready for a ride, the passenger turns on his or her Uber signal and effectively puts out an electronic thumb. Both passenger and driver receive an instant updated map with the potential matches to be found as moving dots across the map, with estimates of congestion (which influence pricing), as well as arrival information. In a similar fashion, Waze is a transport application for local drivers. When drivers get in their car, they turn on Waze, which utilizes the phone's GPS tracker, motion sensors, and built-in geographic road information (speed limits, lights, stop signs) to estimate the level of traffic you are experiencing while driving. Waze then merges your information with all other local drivers' information, creating a real-time picture of road traffic. The application also allows for the reporting of police presence, traffic, accidents, and not-to-miss sights! In essence, this application creates a virtual cloud of self-reported big data.

DESCRIPTIVE POWER AND PREDICTIVE PATTERN MATCHING

As silos are broken down between traditional sources of data, aggregation of big data is allowing astounding predictive capabilities for the data scientist. One example comes from the MIT Media Lab, where a group used

location data from mobile phones to estimate the number of shoppers at a particular department store on the biggest shopping day of the year: Black Friday. By combining this information with historical sales data, demographics of the trade region surrounding the department store, and other relevant factors (macroeconomic, weather, etc.), the team was able to predict retail sales on that day even before the department store itself could (McAfee and Brynjolfsson, 2012)! Another example of the same practice comes from Farecast.com (now owned by Microsoft and accessed through Bing). By aggregating pricing information from all airlines and comparing it to historical information as well as statistically correlated databases that signal pricing, Farecast is able to accurately predict whether the price of a specific airline ticket will go up or down in the near, mid, or short term. At one point it even offered insurance to guarantee the accuracy of its information (http://www.upgradetravelbetter.com/2006/11/13/fare-guarantee-farecast-lets-you-insure-its-fare-predictions/)! Other examples of this approach include predicting housing price changes in the United States with publicly available web information (Wu and Brynjolfsson, 2009) and the Center for Disease Control (CDC) using tweets (twitter.com) to predict the spread of disease, such as cholera in Haiti. In development today is the Square Kilometre Array (SKA), a telescope that is being designed to crunch 300–1500 petabytes of data a year. Just how much data is that? "If you take the current global daily internet traffic and multiply it by two, you are in the range of the data set that the Square Kilometre Array radio telescope will be collecting every day," says IBM researcher Tom Engbersen. "This is big data analytics to the extreme" (Peckham, 2012).

Whatever way you may classify big data, whether it be new variable sources, larger volumes, closer to real-time activity, the mere availability of the resource doesn't necessarily imply greater value to organizations (http://qz.com/81661/most-data-isnt-big-and-businesses-are-wasting-money-pretending-it-is/). A few key elements that have to be present in order for big data to have signification value is that the data must contain relevant information corresponding to a particular process or activity, and the data must have quality. As in the short examples mentioned above, one must realize that simply because new data sources are generated in a particular process, it doesn't imply that it provides descriptive information on the impacts to measuring that process's performance. As far as quality goes, new data variables or more volumes of data must be a reliable and consistent resource to making better decisions. The process of maintaining data quality, variable consistency, and the identification of variables that describe various

activities is a daunting task and requires not only competent analysts, but also the inclusion of subject matter experts and data experts. This book will address the various activities that must be undertaken in order to fully leverage data to create true value for organizations. Remember, analytic techniques of all types are not self-generating methods for decision makers. Skilled professionals are essential to guide the process. Just consider some of the questions below regarding data that potentially describe processes:

- Do Twitter responses reflect accurate consumer sentiment toward events (was the tweet an overreaction or misinterpretation of the reported occurrence?)?
- Were survey questions interpreted correctly by responders?
- Do LinkedIn connections share the same business interests?
- Do Facebook friends share the same product interests?
- Do the demographics generated from credit card purchases truly reflect the profile of the consumer purchasing the product (did younger consumers borrow parents' credit cards?)?

THE VALUE OF DATA

Simply crunching available data elements as they appear and drawing conclusions, whether it's big data or not, can yield suboptimal, even dangerous results to the decision-making process, and end up providing negative value to organizations rather than the assumed positive value. This last statement brings up a vital point to the realm of big data and value. When considering value, probably the most significant add to value that big data brings is the enhancement to the decision-making process to those who access it, manage it appropriately, and utilize it effectively. However, the concept of enhancing the decision-making process by leveraging data involves the widely encompassing realm of analytics and corresponding strategy. We use the phrase "widely encompassing" because the concept of analytics can include a vast variety of applications, depending on what you plan on doing with data. For simplicity's sake, this book will focus primarily on the incorporation of business intelligence and mining applications in leveraging data sources. In the next chapter we will describe a variety of analytic approaches and how they can be used to extract information from data to help decision makers better understand the marketplace with which they are dealing.

CLOSING COMMENTS ON LEVERAGING DATA THROUGH ANALYTICS

Data resources can provide value to organizations from the information that can be extracted from them. This extraction process involves querying data resources for particular variables at particular levels of aggregation in a particular format, and then initiating some type of analytic process. However, before conducting any of these activities, one essential task that underpins the information creation initiative involves the creation of a conceptual model. In other words, whether you have terabytes of data or just a few thousand records, whether you are considering trends over the past few years or focusing on real-time data feeds, decision makers must determine what questions they are looking to answer with data and information. This process can be classified as a conceptual model. Consider using analytics to address the following scenario (e.g., what data variables and level of detail are needed to provide relevant information).

> As a hospital administrator, you are looking to analyze those factors that impact the patients' satisfaction metric that describes their experience while being treated at your hospital.

No matter what industry you operate in, the bottom line to the decision-making process is that individuals must rigorously deliberate over what they are looking to better understand. Once this has been established, the process of leveraging data resources can be undertaken. That process then entails extracting the relevant data variables at corresponding levels of detail and initiating an analytic framework. This concept will be addressed in greater detail in Chapter 5.

ETHICAL CONSIDERATIONS IN THE BIG DATA ERA

Before we go any further in describing the process of leveraging data assets, it is important to stress the adherence to sound ethical practices regarding the various facets of data acquisition, storage, and utilization. Despite this book's focus on describing the various activities involved with extracting value from data, some important concepts should be kept

in mind when dealing with data resources, with a particular emphasis on data that describes individuals.

This book does not promote or support heavy- or underhanded, controversial techniques in acquiring extensive personal data. Individuals should be made aware of how data is generated and gathered regarding their everyday activities, and privacy and security rules should be strictly adhered to. Ultimately, this book adheres to the notion that the management of data resources and analytics should be conducted to yield positive outcomes for processes and individuals who interact with them.

REFERENCES

Cha, A.E. "Big Data" from Social Media, Elsewhere Online Redefines Trend-Watching. *Washington Post*, June 6, 2012.

Davenport, T., and Prusak, L. *Working Knowledge*. Harvard Business Review Press, Boston, Massachusetts, 2000.

Kudyba, S. *Information Technology, Corporate Productivity, and the New Economy.* Westport, Connecticut: Quorum Books. 2002.

McAfee, A., and Brynjolfsson, E. Big Data: The Management Revolution. *Harvard Business Review*, October 2012, pp. 60–62.

Peckham, M. IBM to Help Research and Develop 'Exascale' Supercomputing Telescope. *Time Magazine*, April 2, 2012. http://techland.time.com/2012/04/02/ibm-to-help-research-and-develop-exascale-supercomputing-telescope/.

Tumulty, K. Twitter Becomes a Key Real-Time Tool for Campaigns. *Washington Post*, April 26, 2012.

What the Experts Say: Twitter Guided Trading. *The Week*, June 14, 2012.

Wu, L., and Brynjolfsson, E. The Future of Prediction: How Google Searches Foreshadow Housing Prices and Quantities. In *ICIS 2009 Proceedings,* 2009, paper 147. http://aisel.aisnet.org/icis2009/147.