Appendix D

Big Data

CHAPTER OBJECTIVES

1. Differentiate among the terms *Big Data*, *machine learning*, *data science*, and *data analytics*.
2. Explain Doug Laney's three V's of Big Data.
3. Explain Mark van Rijmenam's additional four V's of Big Data.
4. Differentiate among methods researchers can utilize to evaluate Big Data.
5. Critique various issues related to the ethical use of Big Data.

Imagine that you could live in a world where all of the following can happen:

- Your local department store predicts when you are pregnant based on your buying patterns.
- Your watch lets you know when you are having a heart attack.
- A website determines what fashion trends, colors, and styles are going to be "in" to help clothiers prepare for the season.
- Your video game knows when you are attempting to cheat.
- Your phone knows your frequent locations and remembers them.

Sound like a far-fetched world? All of these things are happening right now as a result of a new quantitative tool called Big Data.

First, Target has created a predictive model based on buying patterns that can fairly accurately predict when a woman is pregnant. Second, one of the eventual goals of Apple's smart watch is to predict and tell when someone is having a heart attack. This predictive ability will be based on large studies examining precursors to heart attacks. There is already a computer program that can diagnose a heart attack up to four hours before a cardiologist. Third, a company called Edited has won awards for the real-time analytics it has been using to help fashion houses, designers, and corporate buyers prepare for changing trends. In fact, their corporate slogan is "how the world's best apparel retailers, brands & suppliers have the right product, at the right price, at the right time." Fourth, if you play video games, you probably know that one of the most popular games on the planet is *Call of Duty*. Quantitative researchers who work for Activision, the company that owns *Call of Duty*, have been working on predictive analytics to determine when someone is boosting. Their analytics are based on a similar premise that credit card companies use to predict credit card fraud. So yes, Activision is working to ensure that everyone who plays *Call of Duty* will be on equal footing in the near future. Finally, your smartphone (i.e., Android, iPhone, etc.) already keeps track of the places you go and how often you visit each place. Your phone then "learns" which places are important to you and provides you with personalized services such as predictive traffic routing to help you arrive more efficiently to your frequently visited locations.

All of these examples demonstrate the new and powerful world of quantitative research in the 21st century. More specifically, these are all part of a particular type of quantitative research called Big Data. The notion of Big Data has graced the covers of magazines from the traditionally technical (e.g., *Nature*, *Popular Science*, and *Cosmos*) to the more general (e.g., *Time*, *The Economist*, and *Harvard Business Review*). This chapter explores the new and fascinating world of Big Data.

The term *Big Data* has been used by a range of different people in slightly different ways. And it is often misconstrued with other, closely related concepts that can confuse the picture. Specifically, **Big Data** refers to data that are simply too large to store on a single computer and are beyond the scope of traditional statistical research software. Saying that something is big is not enough to really make it Big Data, but we will revisit this idea in a moment.

Chapters 9, 10, and 11 showed you how to conduct surveys, content analyses, and experiments; however, this appendix is less about showing you how to conduct Big Data studies and more about what these studies are. Because Big Data is a fairly recent concept, most of the writing on the subject is extremely technical and not at the introductory level. The goal here is to help you understand how Big Data is being used all around you, every day, to make a wide range of decisions that impact your life. Many of the examples we use in this chapter have helped researchers predict phenomena, and many were designed by corporate entities and not for the sake of academic research.

As we explore Big Data, we will define what Big Data is, explain the intersection of cloud computing and Big Data, explore common analytical techniques used to examine Big Data, examine some recent research studies using Big Data in communication, and discuss some real ethical issues related to Big Data research today. Before we go any further, however, we really do need to define exactly what we mean by Big Data.

WHAT ARE THE DATA IN BIG DATA?

Before we break down the definition of Big Data, let's have a brief conversation of what is meant by "data." In this book, we defined **data** as the collected measures of independent and dependent variables that can be used for statistical calculations. However, in the traditional research process, data are something that is often specifically collected for the purposes of research. For example, when you conduct an experiment and have a participant fill out a survey, the survey is being collected explicitly for that experiment. The data used in Big Data are not generally collected in the same way. To help us understand what we mean by this, we must talk about two important types of data: data generated by humans and data generated by machines.

Human-Generated Data

Humans produce data all the time without thinking twice about it. You go on Facebook and click the iconic "thumbs-up" symbol: you just generated data. You make a purchase from iTunes: you just generated data. You text your best friend where to meet you: you just generated data. Humans generate a ton of data in the 21st century. Every purchase, like webpage view, and so on is a piece of data that we are creating. We may not always think of these behaviors as generating data, but they are. We call this intentional data because the data result from our direct behavior. We may not realize that the data are being stored in any fashion, but our behavior allows the generation of that data. Think about your average day. You wake up in the morning and check your e-mail: data generated. You then listen to Pandora or Spotify while in the shower: data generated. You plug your iPhone into your car and listen to iTunes: data generated. And so on and so on. We go through our days amassing tons of data about ourselves, our habits, our personalities, and more.

Machine-Generated Data

Besides humans, machines also create tons of data in what is called machine-to-machine data. Think about your house right now. How many devices do you currently have connected to your Internet? In the home of one of our authors, he has a cloud storage drive, an iPad, an iPhone, a laptop, a Surface, and a desktop, which are all on the more normal side. However, he also has a DVR, television, and stereo system linked to the Internet. At any given moment, these devices are interacting with each other and with the Internet, where in turn they are interacting with large computer mainframes that are

Metadata

Not all data generated by humans are intentional. In fact, much of the data we generate on a daily basis is called metadata, or information that provides context for other pieces of data. For example, you take a photo with your iPhone and post it to Instagram. Unbeknown to you, that picture contains a range of data that have nothing to do with the image itself: GPS coordinates of where the picture was taken (altitude, latitude, longitude, etc.), the date the picture was taken, the model of phone used to take the picture, and a range of factors related to photography (aperture, brightness, focal length, lens model, shutter speed, etc.). There are even other apps out there that allow you to take the metadata from an image file and pinpoint the exact location where the photo was taken. Thankfully, other apps will allow you to clean or alter the metadata from an image before you post or share that image. It is even possible to control some metadata in the systems settings of your phone by controlling location services.

Do not be fooled, however; images are not the only pieces of data we create that contain metadata. In fact, every 140-character tweet someone sends on Twitter contains more than 2,150 pieces of metadata, including tweet content, geographic location, author's biography, author's URL, creation date of account, number of followers, number of tweets ever sent, and so on (Dwoskin, 2014). That is more pieces of metadata than you have characters in your tweet.

collecting data. His stereo system is the Amazon Echo, which connects to the Internet and has access to his digital music archive, but Echo also can tell him the weather and time and even keep him up to date on breaking news. His "stereo" can even tap into his household lights with an adaptor that gives it control over specific light fixtures. Furthermore, each of his rooms now contains an Alexa device that is constantly listening and connected to the Internet. We do not tell you all of this to sell you an Echo, but rather to explain that each of these functions requires the device to constantly interact with the world outside of his house without him paying attention and telling it what to do. With each of these activities, the device is communicating with other devices and amassing massive quantities of data in the process.

This process of connecting a wide variety of our lives together through the Internet is commonly called the **Internet of Things** (IoT). In fact, increasingly more of our basic devices are now being embedded with microchips that help them collect data in an effort to perform better. There are more devices connected to the Internet today than there are people connected to the Internet. In a 2014 report, Hewlett-Packard examined 10 of the most common IoT devices: televisions, webcams, home thermostats, remote power outlets, sprinkler controllers, hubs for controlling multiple devices, door locks, home alarms, scales, and garage door openers. Of these different devices, 90% collected at least one piece of personal information that was transmitted. Admittedly, you probably only have one or two of these devices (if any) right now. But in the future, you can expect everything to be connected and collecting data. Imagine that your printer can

detect when it is running low on ink, so it orders more ink on its own. Imagine a car that can drive itself while you are able to enjoy the ride and catch up on sleep or finish getting ready in the morning. Imagine a situation where a major car accident causes traffic delays on your normal route, so your alarm clock wakes you early and even gets your coffee started earlier as well. These may seem far-fetched right now, but they are just a few of the technologies researchers are looking to create in the near future. In fact, it is predicted that the overwhelming majority of data created in the near future will be machine-to-machine data that are never witnessed by the machine's users.

BIG DATA?

When people start talking about this idea of Big Data, most of the time it is in reference to the world of business. Early on, business realized that the massive amounts of data they were collecting could help them be more productive and efficient, which of course leads to higher profit margins. However, businesses are not the only people who are using Big Data. In fact, Big Data can be used for a wide range of different purposes, as we saw in the introduction of this appendix.

Big Data Explained

As mentioned at the beginning of this appendix, a common definition of Big Data involves data that are too large for traditional statistical software packages. However, this definition is not overly descriptive. As such, it is important to differentiate Big Data from a number of closely related terms: machine learning, data science, and data analytics. Without going into too much detail on each of these terms, we will differentiate each of them from Big Data.

MACHINE LEARNING

The first commonly heard term involving Big Data is machine learning. **Machine learning** is an interesting branch of science involving the creation of algorithms that enable a computer to learn and make decisions when exposed to new data. In essence, machine learning is a type of artificial intelligence. Let's say that you are shopping at Amazon.com. You've been purchasing DVDs, .mp3 files, clothing, food, and so on for a while now, and Amazon recommends what it thinks you will enjoy and may want to buy. No one has to explain to Amazon's computer network what your recommendations will be. Instead, Amazon's computers learn over time about your varied interests and buying patterns and then use that information to make individualized recommendations to you. The computer has learned about you and is using that information to market specific products tailored for your tastes. Is this process always perfect? Of course not. However, the more interaction you have with their system, the more specific and accurate these predictions can become over time.

DATA SCIENCE

The second common term used in conjunction with Big Data is data science. For our purposes, we define **data science** as the emerging field that attempts to extract knowledge from data through advanced mathematical analyses, computers, and databases. Obviously, data scientists are the individuals who are working to combine these three fields into unique processes for analyzing data using the intersection of these three fields.

According to data scientist Drew Conway (2010), data science is the intersection of three primary skill sets: hacking skills, substantive expertise, and math and statistics knowledge (Figure D.1). Regarding hacking skills, we are not talking about the illegal types of hacking commonly associated with "black hat" activities. Instead, data hackers must have the ability to manipulate text files using an array of programing skills. As for substantive expertise, Conway argues that data scientists must understand the scientific research process. Everything we have discussed in this book is information that a good data scientist should know. Finally, data scientists must understand how to extract information from data using appropriate statistical methods and a variety of different statistical software packages when necessary.

DATA ANALYTICS

Data analytics is a set of tools one can use to analyze Big Data. More specifically, **data analytics** are a set of tools used to make predictions about the future based on information from the past. There are three types of data analytics to discuss here: predictive, descriptive, and prescriptive.

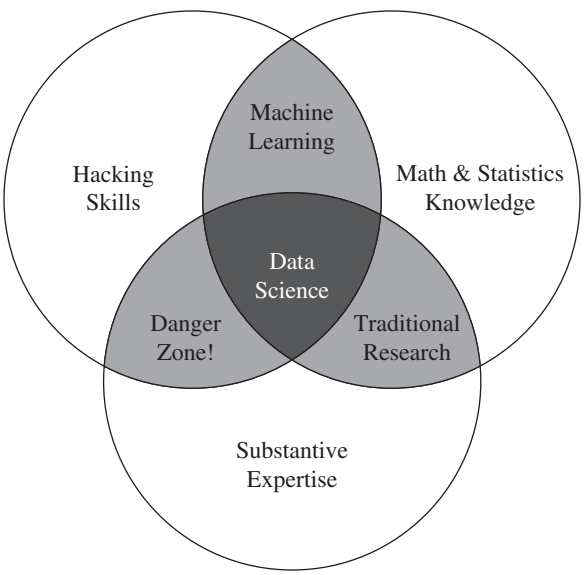

FIGURE D.1
Data Science (Used with permission by Drew Conway, http://drewconway.com/)

PREDICTIVE ANALYTICS When you rely on past data to make predictions, you are engaging in **predictive analytics**. For example, if you want to predict how much your organization will spend on paper this year, you can analyze how much paper has been purchased over the past five years and make a decent prediction about what will probably happen in the future.

DESCRIPTIVE ANALYTICS **Descriptive analytics** involve describing the data—for example, how many people are viewing a website, how many people make a purchase, or how many likes you have on Facebook. In fact, a great deal of the information generated from Big Data is ultimately descriptive.

PRESCRIPTIVE ANALYTICS **Prescriptive analytics** are the newer of the three types. The goal of prescriptive analytics is to use data to determine possible courses of action and what the ramifications of these different courses of action will be.

In the past few pages, we have introduced a number of terms that are often clustered together with the notion of Big Data. With all of this in mind, we will turn our attention to trying to understand what data scientists mean when they talk about Big Data.

Laney's Three V's

In 2001, Doug Laney wrote a white paper (i.e., an authoritative report) examining the concept of data and the future it would play for businesses. Specifically, Laney predicted that businesses were on the precipice of a period when data would become an increasingly larger part of day-to-day operations. To help explain the forthcoming new

Algorithms in Big Data

Data scientists often use algorithms to investigate and make predictions. An algorithm can be simply described as a series of necessary steps to accomplish a task. In the case of Big Data, we are dealing with necessary mathematical computations to help us analyze data. Imagine that you are a nonverbal researcher examining facial expressions. If you got hold of a large database of faces (with facial expressions), how would you create a program that could analyze millions of faces and place names with those faces? AT&T started doing this back in the early 1990s and have made such a database openly available for people: http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html/. The AT&T facial database contains images for 40 different people. Each person has 10 different facial expression pictures, for a total of 400 images in the database. Two researchers, Alex Rodriguez and Alessandro Laio (2014), created a new algorithm that enabled a computer to look for similarities among pictures and determine how many people existed when given the 400 different images without any other information. As you can imagine, the algorithm involved here was complex and involved many different mathematical equations and models all running together to help correctly identify that there were only 40 people represented by the 400 pictures.

world of data science, Laney differentiated these new data from previous data. Laney recommended three specific characteristics to describe this new type of data: volume, velocity, and variety.

VOLUME

In our basic definition of Big Data, we said that it comprises data that are too large for a traditional computer or statistical software package. With that said, calling something "too large" is a problem in its own right. When one of our coauthors first bought a laptop computer, for example, the internal hard drive at that time seemed ridiculously large, and there was no way our coauthor would ever need something larger. That amazing internal hard drive had six gigabytes of space. Now our coauthor's desktop computer has a two-terabyte SSD drive, a 256-gigabyte SSD drive, a six-terabyte expansion hard drive, and an eight-terabyte backup hard drive in the computer. The idea of "big" can change rapidly.

In 1965, Gordon E. Moore wrote an article specifically focusing on this issue of size. In that article, he discussed what is now commonly called Moore's Law—basically, that the physical capacity and performance of a computer doubles every two years. This doubling has been amazingly predictive of the state of computing since Moore's Law originally gained traction. As such, what we call Big Data today might not be that big tomorrow. For example, look at Microsoft Excel. In Excel 2003, the maximum-sized spreadsheet was 65,536 rows by 256 columns. With Excel 2013, however, the spreadsheet size was 1,048,576 rows by 16,384 columns.

We mention all of this to show you that the idea of "volume," or the pure amount of data that we are dealing with, changes over time. If you look at Figure D.2, you will see a history of data. From the good old days of the 1.44-MB floppy disk to the petabyte hard drives that all of us will see fairly soon, our idea of data volume has changed. Today's Big Data may be tomorrow's small data as our computer processors and hard drives keep on doubling every two years. As such, just looking at the volume of data is not enough for something to be truly considered Big Data.

In terms of Big Data, keeping accurate records of how much data are being generated already is almost impossible on a day-to-day basis. However, IBM has been widely cited as saying that in 2012 approximately 2.5 exabytes (or 2.5 *billion* gigabytes) of data were generated daily (Wall, 2014; Ward & Barker, 2013).

VELOCITY

The second V in Laney's three V's of Big Data is velocity, or the speed of the continual onslaught of data activity. In traditional research, we often collect what are called "static" datasets, or datasets that once they are collected stay the same. For example, Dawson (1995) put together a simple dataset examining the sinking of the *Titanic*. In this dataset, he had four variables: class (crew, first, second, or third), age (adult or child), sex (male or female), and survived (yes or no). Is this dataset going to evolve over

Bit (b)	1 or 0	A big stands for a binary digit and is always represented as a 1 or 0. Bits are the basic foundation of all computers.
Byte (B)	8 bits	In computer code, a byte is just enough to create a single letter or number.
Kilobyte (KB)	1,0001 bytes	Roughly ½ of a page of typed text.
Megabyte (MB)	1,000 kilobytes	One of the old 3 ½ floppy disks was 1.44 MBs. Your average .mp3 song is approximately 4 MBs.
Gigabyte (GB)	1,000 megabytes	1 GB is about a 90 minute long standard definition movie.
Terabyte (TB)	1,000 gigabytes	1 TB is the equivalent of 1,500 CD-ROMS. 1 TB can hold 1 million minutes of .mp3s.
Petabyte (PB)	1,000 terabytes	1 PB is enough to play .mp3 files continuously for 2,000 years. 1 PB is the equivalent of 20 million 4-drawer filing cabinets of text.
Exabyte (EB)	1,000 petabytes	It is estimated that 5 EBs would be the equivalent to every spoken word from a human throughout history. 1 EB is roughly the equivalent of 36,000 years of nonstop streaming HD movies.
Zettabyte (ZB)	1,000 exabytes	Global internet traffic is approximately 1.1 ZBs per month. The entire amount of information available on the internet is expected to reach over 400 ZBs by 2018.

[1] Some argue that a KB is actually 1,024 bytes, a MB is 1,024 kilobytes, etc…

FIGURE D.2
Understanding Volume

time? No. These data are set in stone, and how they exist today is how they will exist 100 years from now (assuming some radical information about the sinking of the *Titanic* and its survivors is not discovered). We have known this information for a long time, and we will continue to know this information. Of course, you can still use this information to make some interesting assessments about who survived and who did not. Survey data, content analysis data, and experimental data are generally examples of static datasets.

Velocity, on the other hand, refers to datasets that are not as easy to capture because the data are continuously being recorded over time and/or space. If a pristine lake is an example of a static dataset, then Niagara Falls is an example of Big Data. Where a 7.5-acre lake has 2.5 million gallons of water standing still and can be watched from day to day, Niagara Falls has 150,000 gallons of water rushing over its ledges every second. Niagara Falls reaches the same amount of water contained in that static lake in about 17 seconds. So, you may be wondering, what does this have to do with quantitative research? Well, your average research may collect data from 200 participants over the

course of a one-year study, and then the researcher analyzes this static data. Big Data researchers, on the other hand, look at data that are coming in at much higher volumes and much faster.

Imagine you want to examine how people in the world are using Twitter. Twitter admits to tracking who is tweeting, the GPS location of those tweets, the networks used to access Twitter, and many other pieces of metadata. For one user, this would not seem like much information; however, when you have more than 8,000 tweets per second, that is a lot of data being collected quickly. Currently, the Twitterverse sends more than 1 billion tweets in under 2 days. Want to see how much data are being generated via social media per second, today, or this year? Check out http://www.internetlivestats.com/ to see this information live. The speed at which these data are generated causes problems for most of our traditional statistical software (e.g., Excel and Statistical Package for the Social Sciences [SPSS]). These software programs are designed for fairly large static datasets. As such, new data-management techniques have been created to help us analyze these data in real time. Thankfully, there are a number of both proprietary and free software packages that will help you collect Twitter and other social networking site data. The Social Media Lab at Ted Rogers School of Management, Ryerson University, has put together a comprehensive list of tools that have been used by researchers who study social networking sites (http://socialmediadata.org/social-media-research-toolkit/). One of our coauthors recently collected data about a large march happening in Washington, DC. Our coauthor collected over 500,000 tweets that were sent about the march during a 24-hour period. As you can imagine, that's a lot of data to examine.

VARIETY

The last V in Laney's three V's of Big Data is variety. In the other research methods we have discussed in this book, all of these data can be easily input into a traditional spreadsheet for analytic purposes (e.g., Excel or SPSS). And other traditional data can be used in a traditional relational database (e.g., Access or FileMaker). Whether you want to create an employee database containing contact information, photos, and summaries of annual reports or you want to handle the reservation system for a multimillion-dollar hotel chain, a database is the most effective tool to help you.

Although most people just call them databases, there are a number of different types of databases that can exist. The most common one is a relational database. Edgar Codd (1970) coined the term *relational database* while working at IBM in the late 1960s. The basic part of a relational database is a single table that consists of rows (records) and columns (fields). Figure D.3 is an example of a basic relational database. These databases are called relational because of the ability to establish mathematical relationships across multiple tables; the relationships between rows in these databases are sometimes described as parent and child. Your average database may have 10 different tables that it creates various relationships among. Advanced databases could have up to 1,000 different tables.

Fields

Employee Name	Tenure at Company	Department	Salary
Rebecca Smith	8 Years	Human Resources	$75,000
Donald Praeger	17 Years	Management	$225,000
Candice Flayhan	3 Years	Marketing	$100,000
Joan Johnson	17 Years	Management	$225,000
Jessie Attias	8 Months	Public Relations	$60,000

Records

FIGURE D.3
Relational Database

The "variety" of Big Data comes into play because the statistical techniques associated with Big Data can analyze traditional structured data (e.g., spreadsheets and databases), but it can also handle semistructured and nonstructured data that come in the forms of audio, images, text, video, and so on. For example, most social media sites enable people to post some kind of text in addition to photos and videos that are shared, along with links to other websites. So to understand how people are truly using social media, you cannot ignore these other forms of data just because they do not easily fit into the more traditional data management box. Some information scientists have argued that 80% to 85% of the data being analyzed today are unstructured data, especially in the form of text; however, these numbers do not always have actual data to support them (Grimes, 2008). And in the end, whether you believe that 80% to 85% of data being analyzed are unstructured is not as relevant as the reality that unstructured data comprise a huge part of Big Data today.

Four More V's

Besides the original three V's proposed by Laney (2001), van Rijmenam (2014) has proposed an additional four V's: veracity, variability, visualization, and value.

VERACITY

Data are only good when you can trust the data. When your data are bad, you will make decisions based on those data that are simply invalid. This is especially true when much of the decision making is automated by computers. For example, on Tuesday, April 23, 2013, a hacker used the Associated Press's Twitter account to send out the following tweet at 12:07 PM: "Breaking: Two Explosions in the White House and Barack Obama Is injured." The U.S. stock market took a $200 billion nose dive after this tweet was

sent. Of course, the tweet was fake, and it was corrected within minutes. So how did the stock market dive so quickly? One of the main reasons is that many stock trading companies have supercomputers constantly analyzing data across the Internet (including Twitter). When those computers received the tweet from a predetermined trusted source like the Associated Press, the computers went into a selling frenzy automatically, without any human actually telling them to do so. As you can see, when you have bad data, you will have bad decisions.

VARIABILITY

Variability refers to data in which meaning is always changing. For example, imagine that you tweet "Lacey's Coffee shop is great," and another customer responds to your tweet by saying, "Lacey's Coffee Shop is great if you like bad customer service." Although data analysts who are studying tweets about Lacey's Coffee Shop might search the data for the word "great," they would need to have sophisticated algorithms that look at the surrounding pattern of words to truly understand the meaning of great in this tweet. Thus, great is the piece of data under analysis, but the meaning of great is variable. Context does matter. As van Rijmenam (2014) explains, "Variability is often confused with variety. Say you have a bakery that sells 10 different breads. That is variety. Now imagine you go to that bakery three days in a row and every day you buy the same type of bread but each day it tastes and smells different. That is variability" (p. 11).

VISUALIZATION

The third additional V is visualization, or taking a large amount of data and putting it into some kind of easy-to-read format. Van Rijmenam (2014) warns against thinking of visualization as simple barographs and pie charts. Visualization is designed to help people quickly see and understand what the data are saying. Although visualization may seem simple, it can be one of the hardest parts of the Big Data process. For example, imagine you want to display a history of the use of various terms in printed books going back to the 1800s. You could spend the rest of your natural life going through books trying to accomplish this task, or you could rely on the Big Data set created through the Google Books Library Project. Google has been scanning massive numbers of books dating back to the 1800s in an effort to create a digital knowledge base. One of the fascinating parts of this process involves being able to analyze the text within these books in a way that could never have been dreamed of before.

Getting back to our example, when we analyze the use of the words "elocution," "speech," "communication," "media," and "nonverbal" in books dating back to the 1800s, an interesting trend emerges. When you look at Figure D.4, you can see how the inclusion of these words has changed over time. Although "speech" was the preferred term for a long time, "communication" has usurped "speech" as the more commonly written word. The other three words ("elocution," "media," and "nonverbal") hovered fairly close

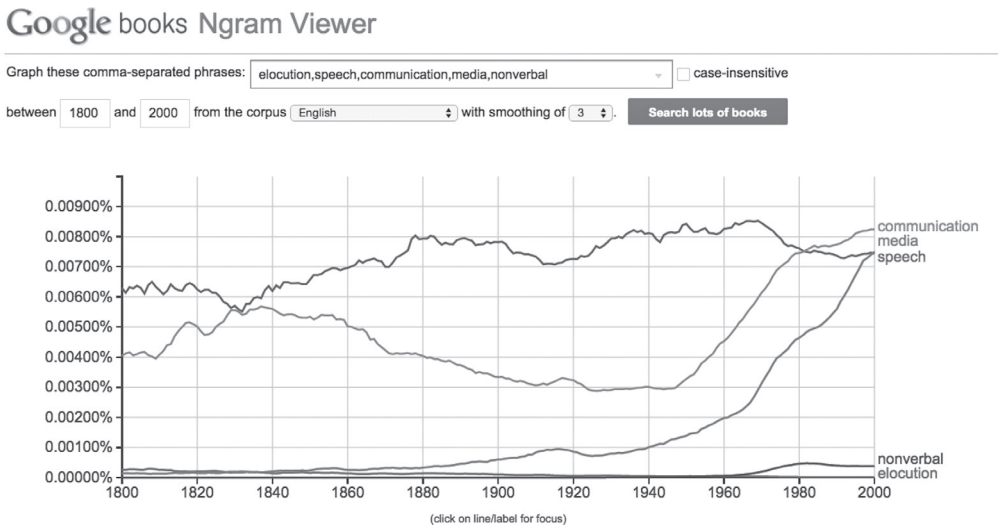

FIGURE D.4
Google Books Library Project

together toward the lower end of the spectrum. The first recorded inclusion of the word "nonverbal" by Google happened in 1806, but you really do not see its use take off until the 1960s. As for the word "media," we saw steadily increasing use from the 1800s into the 1900s, but the real takeoff of the term did not happen until the 1930s. Now, the word "media" is more common in books than the word "speech." As for the word "elocution," it was never as common as "speech" and "communication," but it was used more commonly at the beginning of the 1800s than it was by the 1900s.

The one simple chart in Figure D.4 was created using massive amounts of data and could not have been completed even back at the turn of the 21st century. Although you can make some interesting conjectures about the use of those terms over time based on the data, our purpose here is for you to look at how the data from massive quantities of books for more than 200 years are visualized in a simple chart. If you want to play with this tool on your own, you can go to their website at https://books.google.com/ngrams/.

VALUE

The last of van Rijmenam's (2014) four additional V's is value, or the ability to turn Big Data into information that helps generate knowledge and wisdom. For the professional world, this knowledge and wisdom means ensuring that people using the data can make decisions based on the data that will help the organization competitively. In academia, we want these data to tell us something about how humans behave, interact, and communicate with one another. If you are going to take the time and expense to use Big Data, you want to ensure that the end result has value.

BIG DATA AND THE CLOUD

As discussed in the previous section, one factor that makes Big Data "big" is that the data cannot be stored on a single personal computer, which Laney (2001) referred to as the volume of the data. In fact, even adding an extra external hard drive to your computer will not help you handle the amount of data being produced. Imagine you want to analyze the data that a company like Facebook produces in a single day. You are looking at sifting through around 700 terabytes of data. Your average computer may have a 1-terabyte hard drive, so you would need 700 average PCs to store all of the data necessary. And that 700 terabytes is just storage space. It does not include any of the software that you need for your computer or statistical analyses. In this section, we explore what is meant by "the Cloud," the relationship between the Cloud and data, and the intersection of Big Data and the Cloud.

Understanding the Cloud

Needing access to data at ever-increasing sizes is not new. As mentioned earlier in this chapter, in the good old days, when you needed your data to go with you, you put it on a floppy disk that contained about 1.44 MB of memory. These quickly became obsolete, however, and the next major player in this space game was the zip drive. Most of you reading this text probably have never seen a zip drive, which was introduced in 1994 and gone by 2000. These drives originally stored around 250 MB and eventually grew to 750 MB before the writeable CD-ROM became the norm. The CD-ROM came in a few different sizes as well, ranging from 650 to 900 MB. Then came the DVD-ROM, which could hold 4.7 GB of data. One interesting technology that changed storage capacities was the USB flash drive, which was first introduced by IBM and Trek Technology. The original USB flash drive (also called a thumb drive because of their size) only contained 8 MB of data. Unlike some of the other technologies discussed here, the USB flash drive is the only one that has grown with the ever-increasing need to handle larger amounts of data. Today, you can purchase a reliable two-terabyte USB drive. Admittedly, the current cost is close to $1,400, so it is priced out of the range of most people for a simple thumb drive, but you can be assured that this price will be dropping.

You may wonder why this brief history of the size of portable data-storage devices is important to the world of Big Data, but it is essential for understanding where we are today. In recent years, instead of taking a USB drive with you when you must carry data, many of us rely on Internet-based storage. Internet-based storage is like a giant filing cabinet of your data kept on the Internet so that you can access it from anywhere you need. We call this giant filing cabinet on the Internet "the Cloud." This does make it sound like our data are stored in some magical place, but in actuality, our data get stored on servers that are hosted by cloud-storage companies. These servers are actually housed in extremely large data-storage centers (or data farms). It is not uncommon to see many of

the top names in social networking or cloud storage having servers located within a single data-storage center just rows away from each other. Some companies, like Facebook, have opted to stop leasing space in a data-storage center and open multiple facilities of their own. The ultimate purpose of these large data-storage facilities is to allow these massive servers to have easy, direct access to the Internet so that users can access their information without ever knowing the necessary computing power behind what is happening.

We will use one basic company as an example (we are not promoting this company). Dropbox was founded in 2007 by Drew Houston and Arash Ferdowsi, two former MIT computer science graduates. The idea behind Dropbox was simple. Houston was tired of having to save work to a flash drive that was taken between home, school, and work. Flash drives, although great devices, can fail. Houston started creating a way to access his files through the Internet in an effort to ditch the flash drive. So how does Dropbox work? When you sign up for a free 2-GB membership, you save files into a Dropbox folder on your computer that is automatically synced to a storage server at a data-storage center. When you want to access those folders on a different computer or device, the Dropbox application is able to locate your content on the server and give you access to those files. Voilà! You have now used cloud computing.

The Cloud and Data

The cloud has enabled a wide range of businesses to have almost unlimited amounts of storage capacity online because of massive data-storage centers. One of the biggest names in cloud storage is actually Amazon Web Services. Amazon's S3 data storage has more than 50 regional storage centers around the world. In fact, if you have ever watched Netflix, then you have accessed data that were streamed from an Amazon storage facility. For partnering with a storage center, there are three important characteristics to consider: scalability, redundancy, and speed. Obviously, as scalability, redundancy, and speed increase, so does the cost of the cloud service.

SCALABILITY

Scalability refers to the ability of the data-storage center to grow as you grow. Can it meet the increasing demands that you have when you have them? As organizations amass more and more data, they often must increase the amount of data storage they have available. In the past, organizations had their own data-storage servers. Often, however, these quickly became obsolete because more storage capacity was needed, or the company purchased too much storage capacity that was never utilized.

REDUNDANCY

Redundancy involves how many backup copies of your data can be made. Data can even have problems in the Cloud, so having redundancies built into data storage prevents any data from getting corrupted or lost. Dropbox is a great example of redundancy. When

someone uses Dropbox, copies of the files are kept on that person's hard drive, any other hard drive where that individual installs Dropbox and syncs files, and by Dropbox itself on multiple servers. In fact, Dropbox stores copies of every file uploaded to Dropbox (including multiple versions of files and deleted files) for up to 30 days. So if someone accidentally deletes a file, Dropbox's redundancy is right there to help her or him get the data back.

SPEED

Speed refers to how quickly you want to access your data. If you are working with a program like Dropbox, you want fast access to the files that you have stored with that company. However, there are other instances when immediate access to your data may not be as important. Imagine you are a college or university that has just digitized all of its student records since the school opened. This amount of digitization will create a lot of data, but do you need quick, immediate access to all of these data? Probably not. Many data-storage centers have repositories for what is called cold data, or data that are not needed often but must be stored somewhere. Amazon's cold data program is called Glacier. It is cheap (about $0.01 per gigabyte) and can grow as an organization needs more data space; however, getting access to these data could take a few hours when a request is made.

Big Data and the Cloud

Now that we have explained how the Cloud works, it is also important to discuss the notion of cloud computing, or computer services that are delivered to your computer via the Cloud. Four basic types of common services are delivered: infrastructure as a service (IaaS), platform as a service (PaaS), software as a service (SaaS), and data as a service (DaaS).

As the names suggest, the first three are more about how we interact with cloud computer rather than the data themselves. IaaS is really what we have been talking about with regard to data storage centers. IaaS involves access to storage, hard drives, servers, and so on. PaaS is a service that allows application creators to develop, test, and deploy their apps at a much faster rate. As an end user of all these tools, SaaS is where most of us really experience cloud computing. When you interact with a website software like Google Docs, Microsoft OneDrive, or iWork for iCloud, you are experiencing SaaS.

Finally, we have DaaS, which enables individuals to buy access to data. According to Pringle (2014), DaaS can be defined "as the sourcing, management, and provision of data delivered in an immediately consumable format to organizations' business users as a service" (p. 3). DaaS providers realize that in the 21st century, data are a commodity, so collecting and selling that data to others is a profitable business. Maybe you want access to every tweet that has ever been sent. You could scroll through all of that

information, or you could buy access to that information from a DaaS provider who has already done the legwork. Some more common sources of this type of data are Factual (http://www.factual.com/), Gnip (https://gnip.com/), Infochimps (http://www.infochimps.com/), Opera Solutions (http://operasolutions.com/), and Oracle (https://www.oracle.com/cloud/daas.html/).

Thankfully, not all accessible data must be gathered by yourself or purchased. There are a number of great sources of data that can be freely used. Figure D.5 lists some good, publicly available sources of Big Data.

Type of Data	Data Source
Amazon Web Services Public Data	
List of various publicly available datasets.	http://aws.amazon.com/datasets
Data.gov	
Source of open data created by the U.S. government.	http://www.data.gov/
Data.gov.UK	
Open data compiled and available from the United Kingdom Government.	http://data.gov.uk/
European Union Open Data Portal	
Range of data collected by various institutions within the European Union.	http://open-data.europa.eu/en/data/
HealthData.gov	
125 years of public health data collected within the U.S.	https://www.healthdata.gov/
Million Song Database	
This database contains metadata for over one million songs.	http://aws.amazon.com/datasets/6468931156960467
Pew Research Center	
Downloadable datasets from PRC research projects. Not all are necessarily Big Data.	http://www.pewinternet.org/datasets
WikiData	
Centralized, structured database of content from other Wikimedia sites.	https://www.wikidata.org/
Yelp's Academic Dataset	
Data and reviews of the 250 closest businesses for 30 universities (maybe your university is on the list).	https://www.yelp.com/academic_dataset

FIGURE D.5
Free Sources of Big Data

BIG DATA ANALYSIS

Before we begin our discussion of Big Data analysis, we want to clarify that we are introducing you to the basic ideas and not showing you the detailed ins and outs of Big Data analysis. It would take many volumes of information to even begin that discussion. So, in this section, we discuss some of the common analytic methods used with Big Data: data mining and monitoring and anomaly detection.

Data Mining

Data mining is the process of examining data for new, useful information. The concept of data mining became popularized in the 1990s and has been a commonly used set of techniques to look for information in large datasets. For our purposes, data mining is a technique that data scientists can use to analyze Big Data. In fact, data mining techniques can lead to all kinds of interesting findings related to business and health care. In a number of cases, prescription drugs have been pulled from the market after they were found to pose significant risks to users that had not been discovered using the traditional research process.

Unfortunately, data mining can also lead to some interesting and spurious relationships. A researcher can overexamine a dataset looking to find anything that may exist, which is often called data dredging or data fishing (looking for relationships between variables that happen only by chance). Data dredging can lead to all kinds of statistical outcomes, but these statistics have no place in actual reality because they occur by chance. According to Milloy (1995), data dredging is akin to a sharpshooter who fires his gun at the broadside of the barn and then goes up to the barn and draws his target on the side, ensuring that all of his shots are right in the bullseye.

One of the most notorious cases of data dredging was conducted by Peters et al. (1994). In this study, data-mining techniques were used to fish for relationships between one's diet and childhood leukemia. This group of researchers tested everything from fruits and vegetables to processed meats. The only finding that indicated any kind of statistical relationship, however, was hot dogs. Yes, Peters et al. found that if a child ate 12 or more hot dogs a month, he or she had an increased risk of childhood leukemia. Some newspapers took the risk seriously, reporting it as sacrosanct scientific law, but many others scoffed at the study and criticized the data-mining methods used. Marian Burros (1994) in the *New York Times* wrote, "HO-HUM. Another month, another scare. The latest is about a possible connection between the high consumption of hot dogs and childhood leukemia" (para. 1). Eventually, a more empirically structured study examined this issue and determined that the finding was not accurate (Kwan, Block, Selvin, Month, & Buffler, 2004). However, many websites still tout the original Peters et al. (1994) study as evidence of the harm from hot dogs in general.

Although data dredging or data fishing can be highly problematic, data mining is a useful and valid set of statistical techniques if someone follows the scientific method and has a clear purpose for analyzing data. Most data-mining experts will argue that there can be times when exploring one's data can be meaningful (LaRose & LaRose, 2014).

Monitoring and Anomaly Detection

Two common analyses that Big Data are used for are monitoring and detection of anomalies.

Monitoring is a form of Big Data analysis where individuals have a general idea of what they are looking for in the first place. In the beginning of this appendix, we discussed two examples of monitoring: Target, being able to predict when someone is pregnant and *Call of Duty,* knowing when people are cheating. In each of these cases, data scientists have created predictive models designed to monitor an individual's behavior. When a certain number of variables are triggered, then the data scientist is notified that the event has potentially occurred. We say that the data scientist is notified, but more often than not, the positive monitoring automatically triggers a response to occur (e.g., an individual receives coupons for baby furniture, toys, and products in the mail, or a user has her or his account suspended for cheating).

Big Data analysis can also be used to help people find anomalies. Anomalies occur when something within the data falls outside of normal activity. When looking for anomalies, you often have no idea what these anomalies may look like, so there is no way to monitor for them. Instead, anomaly detection attempts to look for patterns within the data that are outside of the norm. When these boundaries are crossed, a data scientist is notified that an anomaly has occurred. The data scientist is informed because you still need a human set of eyes to look at the results and attempt to make sense of them. Although not necessarily used for research purposes, anomaly detection is often useful. For example, if you work for a stock brokerage firm, you may want to know when stock trading patterns suddenly jump outside of the traditional trading patterns. Although this anomaly may mean nothing, it could also be a huge red flag that something drastic and unexpected is occurring. For example, a colleague worked on a research study where an automated phone system monitored the symptoms of heart-failure patients. When a patient reported gaining more than three pounds in three days, this was considered to be an anomaly, and a nurse would receive an e-mail. This e-mail triggered the nurse to call the patient to determine whether the weight gain was caused by eating a bag of chips or a worsening of the patient's heart failure.

Both monitoring and anomaly detection are not necessarily new ideas, but Big Data has made the analysis of both stronger. For example, imagine you had data from 1,000 people and you would have one anomaly. For traditional research purposes, one anomaly of 1,000 cases is not interesting. However, when you can analyze 1 million cases, now you have 1,000 anomalies occurring, so understanding what is causing the

anomaly becomes a fascinating research study. Furthermore, with more than 1,000 cases, researchers have the ability to make better predictive models to start predicting these anomalies and manage them when they occur.

COMMUNICATION AND BIG DATA

Overall, there has not been a huge amount of research conducted within the field of communication using Big Data. In 2014, the *Journal of Communication* published a special issue to highlight some of the ways that researchers are using Big Data within the field. The issue's editor had this to say: "Big Data research is still in its infancy in communication. Relatively little of the work done in this early stage will stand the test of time, but all of it will likely be critical in the ongoing process of conceptual and methodological advance" (Parks, 2014, p. 355). Researchers in communication have focused on a number of interesting areas using Big Data to investigate health communication, mass communication, political communication, and social networking. We expect to see an increase in Big Data research in the near future, but we are writing this appendix while this specific research method is still "in its infancy." In this section, we are not going to look at each of the articles published in this special issue, but we are going to highlight a few of the techniques these researchers utilized.

One of the classic areas of research within media studies has consistently been agenda setting. Agenda-setting theory purports that the media has the ability to set the political agenda discussed within society. If the media finds a topic important and discusses it enough, then the public will likewise also find the topic important. Neuman, Guggenheim, Jang, and Bae (2014) conducted a comparison analysis of political discussions via Twitter versus what the mainstream media was determining to be important. The researchers used a third-party vendor to get access to the "Twitter firehose," which is the enormous, real-time stream of tweets. The firehose is also recorded, so it enables researchers to examine past phenomena as well as current ones. Overall, the researchers were able to compare what was important to social media users directly with what the traditional media outlets were broadcasting as news at the same time. Not surprising, social media and modern news outlets were not really in sync with one another. And without this massive amount of Twitter data, this type of research project would have been impossible. When looking at the various articles in the special issue of the *Journal of Communication*, a number of studies used Twitter as a source of information (Colleoni, Rozza, & Arvidsson, 2014; Emery, Szczypkal, Abrill, Kim, & Vera, 2014; Giglietto & Selva, 2014; Jungherr, 2014; Park, Baek, & Chal, 2014; Vargo, Guo, McCombs, & Shaw, 2014).

Although Twitter is the most popular source of Big Data by communication scholars thus far, it is not the only source of Big Data that communication researchers have utilized to understand human communication. Balmas and Sheafer (2014) used 800,000 news items over three decades to examine how political leaders in six countries were

represented in the international media. These scholars used the news archiving service LexisNexis to locate the articles and then performed a giant content analysis of the 800,000 news items they found. In another study, Shaw and Hill (2014) used data from Wikipedia to examine how people collaborate on wikis. The researchers examined 683 different wikis to see how people go about creating and interacting with one another in the collaborative environment. According to Shaw and Hill, "This dataset consists of rich longitudinal records, which include every contribution made to each wiki, recorded with timestamps accurate to within 1 second. It includes 33,278,993 distinct contributions by 469,524 different contributors to 6,167,797 different wiki pages—more than 264 gigabytes of raw data" (p. 222).

As you can see, whether using Twitter, Lexis Nexis, or Wikipedia, communication researchers are making inroads in the field of Big Data. However, there is still more work to be done. And one of the biggest hurdles many communication scholars face regarding Big Data is access (Parks, 2014). Many of the most dominant social networking sites of our time do not make data available to the public without a price. Even as this chapter was being written, the Twitter hose that Neuman et al. (2014) used is being turned off to third-party vendors (Goel, 2015). Instead, the Twitter hose will only be accessible through Gnip (https://gnip.com/), which Twitter recently purchased. Twitter sees these data as a commodity, and people who want access to the data are going to be forced to pay a lot of money to have that access. Unfortunately, without funds to get access to the data, many scholars will simply avoid delving into the Big Data world. However, as discussed earlier in this chapter, there are tools that will allow you to capture tweets as they are happening live. These tools do not have access to all of the metadata a single tweet possesses.

BIG DATA ETHICS

Hopefully by this point, you have realized that Big Data is an amazing analytic approach/paradigm, but there are definitely points where some of it seems a little creepy. Do you really want Target to know when you are pregnant before anyone else (including yourself)? Clearly, Big Data raises some big ethical questions. Unlike traditional academic research, much of the research conducted using Big Data simply does not fall under the traditional ethical oversight that academic research does. As such, we rely on individuals and organizations to treat these data ethically. According to Big Data ethicists, Big Data ethics involves four primary issues: privacy, identity, ownership, and reputation (Davis & Patterson, 2012; Richards & King, 2014).

Privacy

Probably the single biggest concern regarding ethics and Big Data involves an individual's privacy. We live in a world where information is being collected about us all of the time without our knowledge. Most data collectors will argue that the collection of data

is completed in a way to maintain an individual's anonymity or confidentiality. However, a number of studies have shown that this simply is not the case.

For example, Latanya Arvette Sweeney (2000) was a graduate student when the state of Massachusetts decided to release the medical data for all public employees. Massachusetts did due diligence in its attempt at de-identifying the dataset. In fact, the governor at the time, William Weld, promised that the data were de-identified and that employees had no reason to worry anyone could go about identifying their individual medical data. Sweeney took this challenge and set out to re-identify the governor's data. First, she knew the governor resided in Cambridge, Massachusetts, so for $20, she purchased the voter ID data. The voter ID database included the names, ages, addresses, zip codes, and birthdates of everyone in Cambridge. Sweeney quickly compared the information she learned about the governor to that in the patient database and found that only six people had the governor's birthdate, only three were male, and only one lived in his zip code. With that information, Sweeney was able to correctly identify the governor's health records, including diagnoses and prescriptions that he was taking. Sweeney would later determine that with just three pieces of information (date of birth, sex, and zip code) she could correctly identify 87% of people.

In 2002, the implementation of Health Insurance Portability and Accountability Act (HIPAA) did put a few new protections in place with regard to health information. However, de-identifying data in other venues has also become a concern. In 2006, Netflix released a dataset to the general public containing the de-identified movie rating history of 480,189 Netflix customers for 17,770 movies. Overall, the dataset contained 100,480,507 ratings from Netflix users. Netflix released these data in an effort to allow individuals and teams to create new algorithms that could predict an individual's rating of a movie. Researchers Arvind Narayanan and Vitaly Shmatikov (2008) took this dataset and decided to see whether they could re-identify the individuals within the de-identified dataset. Comparing the Netflix data with information taken from ratings on the Internet Movie Database (IMDB), the researchers successfully re-identified a Netflix user's movie watching and rating history.

As you can see from both of these examples, privacy is a serious problem in the world of Big Data. When it takes only three pieces of information to correctly identify 87% of people, one must always question the privacy of these data.

Identity

In Walt Whitman's (1871) poem "Song of Myself," he writes, "I am large, I contain multitudes." In this short phrase, Whitman is acknowledging that many of us have varying parts to our individual identities. Social psychologists have argued for years that an individual's identity is a complex and ever-evolving idea. More specifically, a person's identity is her or his own conception of who he or she is based on unique individuality and group affiliations. Cohen (2013) asserts, "[P]eople are born into networks of

relationships, practices, and beliefs, and over time encounter and experiment with others, engaging in a diverse and ad hoc mix of practices that defies neat theoretical simplification" (p. 1910).

Unfortunately, not everyone in the networked world sees people as having this right to create their own identities. One of the biggest concerns related to Big Data is who gets to determine your identity in the 21st century. Richards and King (2014) noted that "[Big Data] analytics can compromise identity by allowing institutional surveillance to moderate and even determine who we are before we make up our own minds" (p. 422). Big Data can create a picture of who you are to others without you having any real say in the matter. Instead of you determining your own identity, the numbers determine that identity for you and place you in a clearly labeled box.

A secondary identity concern involves the creation of multiple identities in life. We often have differing identities in school, with our friends, with our families, and in the workplace. However, some argue that this kind of pluralistic understanding of identities has no place in the online world. Mark Zuckerberg, founder and chief executive officer of Facebook, was once quoted as saying, "Having two identities for yourself is an example of a lack of integrity" (Helft, 2011, para. 3). Clearly, Zuckerberg sees the world of identity as purely 100% authentic or 100% inauthentic, with no room in between.

We should mention that there is a highly symbiotic relationship between issues of privacy and identity. As everything we do online is collected and fed into giant storehouses of information, a serious question arises about how this information is used by others. Calo (2014, 2015) has argued that the next wave of Big Data analytics will be targeted less at understanding who we are as people and more at how Big Data analytics can shape us as people. Basically, Calo (2014) argues that Big Data is getting to the point where the information that businesses learn about us as individuals will be used to directly influence who we are as individuals. Marketers will attempt to shape our identities based on what they have learned about our identity. Behavioral economists call this practice **digital nudging**, or "the use of user-interface design elements to guide people's behavior in digital choice environments" (Weinmann, Schneider, & vom Brocke, 2016, p. 433). Weinmann et al. (2016) go on to explain, "Humans face choices every day, but the outcome of any choice is influenced not only by rational deliberations of the available options but also by the design of the choice environment in which information is presented, which can exert a subconscious influence on the outcomes" (p. 433). In other words, humans can only choose between the choices they are given and how those choices are presented. Calo (2014) then takes this one step further and asks to what extent this same technology could be used against the populace. That politicians could gather "information about individual citizens to better persuade them comes very close to the sort of Orwellian propaganda society has collectively rejected. A related critique of nudging is that it tends to infantilize the citizen by removing the habit of choice. Again, the constant mediation of the citizen by technology could accelerate this effect" (p. 1049).

Ownership

The third ethical concern related to Big Data involves ownership. To what extent (if any) do we own the information gathered about ourselves? Do we own basic demographic information about ourselves (e.g., height, weight, eye/hair color, ethnicity/race, date of birth, and zip code)? What about family information (e.g., genetic history, family lineage, and family health histories)? Do we own data about our hobbies (e.g., video editing, basketball, car mechanic, or crafter)? Do we own information about the skills we have (e.g., making a free throw, kicking a field goal, sewing clothing, or arranging large events)? What about our own individual personal tastes (e.g., Coke vs. Pepsi, broccoli vs. asparagus, McDonald's vs. Burger King, and plastic vs. paper)? A lot of information can (and is) collected about us all the time, but do we have any actual ownership of these data?

Europe has long had much more stringent rules for how companies can collect and use data. In fact, the European Union's website devoted to data protection states, "Protecting your personal data—a fundamental right!" (http://ec.europa.eu/justice/data-protection/). The European Union has a set of fairly straightforward policies that argue an individual's data can only be collected for legitimate purposes. You cannot just collect data on or about people simply because you want to collect that data for unspecified reasons to be determined at a later time. Furthermore, the new European Union General Data Protection Regulation (GDPR) rules have reshaped data collection and privacy rules within the EU. However, these same rules do not apply in the United States. Although there have been pushes to create a Consumer Privacy Bill of Rights that tackles issues of data ownership, these legislative pushes have not gone far.

Reputation

The last major ethical question related to Big Data involves an individual's reputation, or the extent that an individual has control over how others think about her or him. According to Davis and Patterson (2012), "Unless we were famous for some other reason, the vast majority of us managed our reputation by acting well (or poorly) in relation to those directly around us. In some cases, a second-degree perception—that is, what the people who knew you said about you to the people who they knew—might influence one's reputation" (p. 18). Today, however, Big Data allows people the ability to infer your identity and form opinions about who you are as a person based purely on the data that are collected about you.

The tech industry has not always been understanding of issues related to reputation. Google's chief executive officer, Eric Schmidt, once told a reporter, "If you have something that you don't want anyone to know, maybe you shouldn't be doing it in the first place" (Newman, 2009, para. 3). To put this quotation in context, Schmidt was discussing the amount of data that are kept and tracked for each individual under the USA Patriot Act. However, there is an inherent idea behind Schmidt's statement that impacts one's deeper reputation. Most of us have probably seen, even if only accidentally,

something on the Internet that we wish we could simply unsee. Should your reputation be based on you having seen whatever that was? In fact, online reputation management is becoming a huge business. Just ask the many companies that now profit by helping individuals, groups, and organizations with this task. BrandYourself says it all right on their website (https//brandyourself.com): "Look great when employers, clients, and even dates Google you."

Manage Your Online Reputation

1. *Google yourself.* It is important to know what information exists on the Internet about you. You may be surprised by some of what you see.
2. *Buy your domain name.* Part of brand management is purchasing the domain name that corresponds with your name. For example, the first author of this book has his domain name at http://www.JasonSWrench.com/. The last thing you want is someone else to own your domain and have information on that website that is deemed inappropriate.
3. *Join social networks.* One mistake that some people make is thinking that if they are not on social networks, that is a good thing. Actually, not being on social networks can be a red flag. Instead, join social networks like LinkedIn, Facebook, and Twitter, but be mindful of the content you post and support.
4. *Optimize your presence.* You want people to find you, so you want to make it easy for them to do this. For example, the first author of this book can be found on Facebook, Twitter, LinkedIn, YouTube, Vimeo, SlideShare, and other social networking sites all under the name JasonSWrench. Why does he do this? When you are consistent across your social networking platforms, you can create a brand that tells the story of who you are. And when you link these various accounts together, they help you rise in listings of Google's and other search engines. You can also use the free reputation report available at http://brandyourself.com/to help manage this.
5. *Keep your private things private.* If you do not want your grandmother, boss, or future boss to learn it about you, then do not put it online. Admittedly, some younger people have been on social networking sites since they were children and did not think or care about their reputations. Unfortunately, anything that exists on the Internet about you can come back to harm you.

GLOSSARY

Big Data: Data that are simply too large to store on a single computer and are beyond the scope of traditional statistical research software.

Data: Collected measures of independent and dependent variables that can be used for statistical calculations.

Data Analytics: Set of tools used to make predictions about the future based on information from the past.

Data Mining: The process of examining data for new, useful information.

Data Science: The emerging field that attempts to extract knowledge from data through advanced mathematical analyses, computers, and databases.

Digital Nudging: The purposeful use of user-interface designs of various forms of technology that help guide consumers' decision-making processes and behaviors in digital environments.

Descriptive Analytics: Data analytics associated with describing the data.

Internet of Things: The process of connecting a wide variety of devices that we use in our lives together through the Internet.

Machine Learning: Branch of science involving the creation of algorithms that enable a computer to learn and make decisions when exposed to new data.

Predictive Analytics: Data analytics designed to use existing data to make predictions about the future.

Prescriptive Analytics: Data analytics designed to determine possible courses of action and what the ramifications of these different courses of action would be.

REFERENCES

Balmas, M., & Sheafer, T. (2014). Charismatic leaders and mediated personalization in the international arena. *Communication Research, 41*, 991–1015. doi:10.1177/0093650213510936

Burros, M. (1994, June 22). Eating well. *New York Times*. Retrieved from http://www.nytimes.com/1994/06/22/garden/eating-well.html/

Calo, R. (2014). Digital market manipulation. *The George Washington Law Review, 82*(4), 995–1051.

Calo, R. (2015). Robot-sized gaps in surveillance law. In M. Rotenberg, J. Horwitz, and J. Scott (Eds.), *Privacy in the modern age: The search for solutions* (pp. 41–45). New York, NY: New Press.

Codd, E. F. (1970). A relational model of data for large shared data banks. *Communications of the ACM, 13*(6), 337–387.

Cohen, J. E. (2013). What is privacy for? *Harvard Law Review, 126*(6), 1904–1933.

Colleoni, E., Rozza, A., & Arvidsson, A. (2014). Echo chamber or public sphere? Predicting political orientation and measuring political homophily in Twitter using Big Data. *Journal of Communication, 64*, 317–332. doi:10.1111/jcom.12084

Conway, D. (2010, September 30). *The data science Venn diagram*. Retrieved from http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram/

Davis, K., & Patterson, D. (2012). *Ethics of big data: Balancing risk and innovation*. Sebastopol, CA: O'Reilly.

Dawson, R. J. M. (1995). The "unusual episode" data revisited. *Journal of Statistics Education, 3*(3). Retrieved from http://www.amstat.org/publications/jse/v3n3/datasets.dawson.html/

Dwoskin, E. (2014, June 6). In a single tweet, as many pieces of metadata as there are characters. *Wall Street Journal*. Retrieved from http://blogs.wsj.com/digits/2014/06/06/in-a-single-tweet-as-manypieces-of-metadata-as-there-are-characters/

Emery, S. L., Szczypkal, G., Abrill, E. P., Kim, Y., & Vera, L. (2014). Are you scared yet? Evaluating fear appeal messages in tweets about the Tips campaign. *Journal of Communication, 64*, 278–295. doi:10.1111/jcom.12083

Giglietto, F., & Selva, D. (2014). Second screen and participation: A content analysis on a full season dataset of tweets. *Journal of Communication, 64*, 260–277. doi:10.1111/jcom.12085

Goel, V. (2015, April 11). Twitter's evolving plans to make money from its data stream. *New York Times*. Retrieved from http://bits.blogs.nytimes.com/2015/04/11/twitters-evolving-plans-to-makemoney-from-its-data-stream/?_r=0/

Grimes, S. (2008, August 1). Unstructured data and the 80 percent rule. *Breakthrough Analysis*.Retrievedfromhttp://www.lakeandpondsolutions.com/helpful-info/acreage-and-volumecalculations/

Helft, M. (2011, May 13). Facebook, foe of anonymity, is forced to explain a secret. *New York Times*. Retrieved from http://www.nytimes.com/2011/05/14/technology/14facebook.html?_r=0/

Hewlett-Packard. (2014, July 29). *HP study reveals 70 percent of internet of things devices vulnerable to attack: IoT devices averaged 25 vulnerabilities per product, indicating expanding attack surface for adversaries* [Press release]. Retrieved from http://www8.hp.com/us/en/hp-news/press-release.html?id=1744676#.Wy5YQ6dKiUk

Jungherr, A. (2014). The logic of political coverage on twitter: Temporal dynamics and content. *Journal of Communication, 64*, 239–259. doi:10.1111/jcom.12087

Kwan, M. L., Block, G., Selvin, S., Month, S., & Buffler, P. A. (2004). Food consumption by children and the risk of childhood acute leukemia. *American Journal of Epidemiology, 160*, 1098–1107.

Laney, D. (2001, February 6). *3D data management: Controlling data volume, velocity, and variety* (File 949). Stamford, CT: META Group.

Larose, D. T., & Larose, C. D. (2014). *Discovering knowledge in data: An introduction to data mining*. Hoboken, NJ: Wiley.

Milloy, S. (1995). *Science without sense: The risky business of public health research*. Washington, DC: Cato.

Moore, G. E. (1965, April 19). Cramming more components onto integrated circuits. *Electronics, 38*(8), 114–117.

Narayanan, A., & Shmatikov, V. (2008). Robust de-anonymization of large sparse datasets. *Proceedings of the IEEE Symposium on Security and Privacy, USA*, 111–125. doi:10.1109/SP.2008.33

Neuman, W. R., Guggenheim, L., Jang, S. M., & Bae, S. Y. (2014). The dynamics of public attention: Agenda-setting theory meets big data. *Journal of Communication, 64*, 193–214. doi:10.1111/jcom.12088

Newman, J. (2009, December 11). Google's Schmidt roasted for privacy comments. *PCWorld*. http://www.pcworld.com/article/184446/googles_schmidt_roasted_for_privacy_comments.html/

Park, J., Baek, Y. M., & Chal, M. (2014). Cross-cultural comparison of nonverbal cues in emoticons on Twitter: Evidence from Big Data analysis. *Journal of Communication, 64*, 333–354. doi:10.1111/jcom.12086

Parks, M. (2014). Big Data in communication research: Its contents and discontents. *Journal of Communication, 64*, 355–360. doi:10.1111/jcom.12090

Peters, J. M., Preston-Martin, S., London, S. J. Bowman, J. D., Buckley, J. D., & Thomas, D. C. (1994). Processed meats and risk of childhood leukemia (California, USA). *Cancer Causes Control, 5*, 195–202.

Pringle, T. (2014, July 18). *Data-as-a-service: The next step in the as-a-service journey*. London, UK: Ovum.

Richards, N. M., & King, J. H. (2014). Big Data ethics. *Wake Forest Law Review, 49*, 393–432.

Rodriguez, A., & Laio, A. (2014). Clustering by fast search and find of density peaks. *Science, 344*(6191), 1492–1496. doi:10.1126/science.1242072

Shaw, A., & Hill, B. M. (2014). Laboratories of oligarchy? How the iron law extends to peer production. *Journal of Communication, 64*, 215–238. doi:10.1111/jcom.12082

Sweeney, L. (2000). Foundations of privacy protection from a computer science perspective. In the *Proceedings of the Joint Statistical Meeting, AAAS*, Indianapolis, IN. Retrieved from https://dataprivacylab.org/projects/disclosurecontrol/paper1.pdf

van Rijmenam, M. (2014). *Think bigger: Developing a successful Big Data strategy for your business*. New York, NY: AMACOM.

Vargo, C. J., Guo, L., McCombs, M., & Shaw, D. L. (2014). Network issue agendas on twitter during the 2012 U.S. presidential election. *Journal of Communication, 64*, 296–316. doi:10.1111/jcom.12089

Wall, M. (2014, March 4). Big Data: Are you ready for blast-off? *BBC News*. Retrieved from http://www.bbc.com/news/business-26383058/

Ward, J. S., & Barker, A. (2013, September 20). *Undefined by data: A survey of big data definitions*. Retrieved from arXiv:1309.5821/[cs.DB]

Weinmann, M., Schneider, C., & vom Brocke, J. (2016). Digital nudging. *Business & Information Systems Engineering, 58*, 433-436. doi: 10.1007/s12599-016-0453-1.

Whitman, W. (1871). *Leaves of grass* (5th ed.). New York, NY: Redfield.

FURTHER READING

Dean, J. (2014). *Big Data, data mining, and machine learning: Value creation for business leaders and practitioners*. Hoboken, NJ: Wiley.

Granville, V. (2014). *Developing analytic talent: Becoming a data scientist*. Indianapolis, IN: Wiley.

Kitchen, R. (2014). *The data revolution: Big Data, open data, data infrastructures, & their consequences*. Thousand Oaks, CA: Sage.

McArdle, J. G., & Ritschard, G. (Eds). (2014). *Contemporary issues in exploratory data mining in the behavioral sciences*. New York, NY: Routledge.

Weiss, S. M., Indurkhya, N., Zhang, T., & Damerau, F. (2005). *Text mining: Predictive methods for analyzing unstructured information*. New York, NY: Springer.