Introduction

Hiding within those mounds of data is knowledge that could change the life of a patient, or change the world.

—Atul Butte, Stanford University

Cancer" is the term given for a class of diseases in which abnormal cells divide in an uncontrolled fashion and invade body tissues. There are more than 100 unique types of cancer. Most are named after the location (usually an organ) where they begin. Cancer begins in the cells of the body. Under normal circumstances, the human body controls the production of new cells to replace cells that are old or have become damaged. Cancer is not normal. In patients with cancer, cells do not die when they are supposed to and new cells form when they are not needed (like when I ask my kids to use the copy machine and I get back ten copies instead of the one I asked for). The extra cells may form a mass of tissue; this is referred to as a tumor. Tumors come in two varieties: benign tumors, which are not cancerous, and malignant tumors, which are cancerous. Malignant tumors spread through the body and invade the tissue. My family, like most I know, has lost a family member to the disease. There were an estimated 1.6 million new cases of cancer in the United States in 2013 and more than 580,000 deaths as a result of the disease.

An estimated 235,000 people in the United States were diagnosed with breast cancer in 2014, and about 40,000 people will die in 2014 as a result of the disease. The most common type of breast cancer is ductal carcinoma, which begins in the lining of the milk ducts. The next most common type of breast cancer is lobular carcinoma. There are a number of treatment options for breast cancer including surgery, chemotherapy, radiation therapy, immunotherapy, and vaccine therapy. Often one or more of the treatment options is used to help ensure the best outcome for patients. About 60 different drugs are approved by the Food and Drug Administration (FDA) for the treatment of breast cancer. The course of treatment and which drug protocols should be used is decided based on consultation between the doctor and patient, and a number of factors go into those decisions.

One of the FDA-approved drug treatments for breast cancer is tamoxifen citrate. It is sold under the brand name of Nolvadex and was first prescribed in 1969 in England but approved by the FDA in 1998. Tamoxifen is normally taken as a daily tablet with doses of 10 mg, 20 mg, or 40 mg. It carries a number of side effects including nausea, indigestion, and leg cramps. Tamoxifen has been used to treat millions of women and men diagnosed with hormone-receptor-positive breast cancer. Tamoxifen is often one of the first drugs prescribed for treating breast cancer because it has a high success rate of around 80%.

Learning that a drug is 80% successful gives us hope that tamoxifen will provide good patient outcomes, but there is one important detail about the drug that was not known until the big data era. It is that tamoxifen is not 80% effective in patients but 100% effective in 80% of patients and ineffective in the rest. That is a life-changing finding for thousands of people each year. Using techniques and ideas discussed in this book, scientists were able to identify genetic markers that can identify, in advance, if tamoxifen will effectively treat a person diagnosed with breast cancer. This type of analysis was not possible before the era of big data. Why was it not possible? Because the volume and granularity of the data was missing; volume came from pooling patient results and granularity came from DNA sequencing. In addition to the data, the computational resources needed to solve a problem like this were not readily available to most scientists outside of the super computing lab. Finally the third component, the algorithms or modeling techniques needed to understand this relationship, have matured greatly in recent years.

The story of Tamoxifen highlights the exciting opportunities that are available to us as we have more and more data along with computing resources and algorithms that aid in classification and prediction. With knowledge like that was gained by the scientists studying tamoxifen, we can begin to reshape the treatment of disease and disrupt positively many other areas of our lives. With these advances we can avoid giving the average treatment to everyone but instead determine which people will be helped by a particular drug. No longer will a drug be 5% effective; now we can identify which 5% of patients the drug will help. The concept of personalized medicine has been discussed for many years. With advances in working with big data and improved predictive analytics, it is more of a reality than ever. A drug with a 2% success rate will never be pursued by a drug manufacturer or approved by the FDA unless it can be determined which patients it will help. If that information exists, then lives can be saved. Tamoxifen is one of many examples that show us the potential that exists if we can take advantage of the computational resources and are patient enough to find value in the data that surrounds us.

We are currently living in the big data era. That term "big data" was first coined around the time the big data era began. While I consider the big data era to have begun in 2001, the date is the source of some debate and impassioned discussion on blogs—and even the *New York Times*. The term "big data" appears to have been first used, with its currently understood context, in the late 1990s. The first academic paper was presented in 2000, and published in 2003, by Francis X. Diebolt— "Big Data Dynamic Factor Models for Macroeconomic Measurement and Forecasting"—but credit is largely given to John Mashey, the chief scientist for SGI, as the first person to use the term "big data." In the late 1990s, Mashey gave a series of talks to small groups about this big data tidal wave that was coming. The big data era is an era described by rapidly expanding data volumes, far beyond what most people imagined would ever occur.

The large data volume does not solely classify this as the big data era, because there have always been data volumes larger than our ability to effectively work with the data have existed. What sets the current time apart as the big data era is that companies, governments, and nonprofit organizations have experienced a shift in behavior. In this era, they want to start using all the data that it is possible for them to collect, for a current or future unknown purpose, to improve their business. It is widely believed, along with significant support through research and case studies, that

organizations that use data to make decisions over time in fact do make better decisions, which leads to a stronger, more viable business. With the velocity at which data is created increasing at such a rapid rate, companies have responded by keeping every piece of data they could possibly capture and valuing the future potential of that data higher than they had in the past. How much personal data do we generate? The first question is: What is personal data? In 1995, the European Union in privacy legislation defined it as any information that could identify a person, directly or indirectly. International Data Corporation (IDC) estimated that 2.8 zettabytes[1] of data were created in 2012 and that the amount of data generated each year will double by 2015. With such a large figure, it is hard to understand how much of that data is actually about you. It breaks down to about 5 gigabytes of data per day for the average American office worker. This data consists of email, downloaded movies, streamed audio, Excel spreadsheets, and so on. In this data also includes the data that is generated as information moves throughout the Internet. Much of this generated data is not seen directly by you or me but is stored about us. Some examples of nondirect data are things like traffic camera footage, GPS coordinates from our phones, or toll transactions as we speed through automated E-ZPass lanes.

Before the big data era began, businesses assigned relatively low value to the data they were collecting that did not have immediate value. When the big data era began, this investment in collecting and storing data for its potential future value changed, and organizations made a conscious effort to keep every potential bit of data. This shift in behavior created a virtuous circle where data was stored and then, because data was available, people were assigned to find value in it for the organization. The success in finding value led to more data being gathered and so on. Some of the data stored was a dead end, but many times the results were confirmed that the more data you have, the better off you are likely to be. The other major change in the beginning of the big data era was the rapid development, creation, and maturity of technologies to store, manipulate, and analyze this data in new and efficient ways.

Now that we are in the big data era, our challenge is not getting data but getting the right data and using computers to augment our domain knowledge and identify patterns that we did not see or could not find previously.

Some key technologies and market disruptions have led us to this point in time where the amount of data being collected, stored, and considered in analytical activities has grown at a tremendous rate. This is due to many factors including Internet Protocol version 6 (IPv6), improved telecommunications equipment, technologies like RFID, telematics sensors, the reduced per unit cost of manufacturing electronics, social media, and the Internet.

Here is a timeline that highlights some of the key events leading up to the big data era and events that continue to shape the usage of big data and the future of analytics.

BIG DATA TIMELINE

Here are a number of items that show influential events that prepared the way for the big data era and significant milestones during the era.

1991

- The Internet, or World Wide Web as we know it, is born. The protocol Hypertext Transfer Protocol (HTTP) becomes the standard means for sharing information in this new medium.

1995

- Sun releases the Java platform. Java, invented in 1991, has become the second most popular language behind C. It dominates the Web applications space and is the de facto standard for middle-tier applications. These applications are the source for recording and storing web traffic.
- Global Positioning System (GPS) becomes fully operational. GPS was originally developed by DARPA (Defense Advanced Research Projects Agency) for military applications in the early 1970s. This technology has become omnipresent in applications for car and airline navigation and finding a missing iPhone.

1998

- Carlo Strozzi develops an open-source relational database and calls it NoSQL. Ten years later, a movement to develop NoSQL databases to work with large, unstructured data sets gains momentum.
- Google is founded by Larry Page and Sergey Brin, who worked for about a year on a Stanford search engine project called BackRub.

1999

- Kevin Ashton, cofounder of the Auto-ID Center at the Massachusetts Institute of Technology (MIT), invents the term “the Internet of Things.”

2001

- Wikipedia is launched. The crowd-sourced encyclopedia revolutionized the way people reference information.

2002

- Version 1.1 of the Bluetooth specification is released by the Institute of Electrical and Electronics Engineers (IEEE). Bluetooth is a wireless technology standard for the transfer of data over short distances. The advancement of this specification and its adoption lead to a whole host of wearable devices that communicate between the device and another computer. Today nearly every portable device has a Bluetooth receiver.

2003

- According to studies by IDC and EMC, the amount of data created in 2003 surpasses the amount of data created in all of human history before then. It is estimated that 1.8 zettabytes (ZB) was created in 2011 alone (1.8 ZB is the equivalent of 200 billion high-definition movies, each two hours long, or 47 million years of footage with no bathroom breaks).
- LinkedIn, the popular social networking website for professionals, launches. In 2013, the site had about 260 million users.

2004

- Wikipedia reaches 500,000 articles in February; seven months later it tops 1 million articles.
- Facebook, the social networking service, is founded by Mark Zuckerberg and others in Cambridge, Massachusetts. In 2013, the site had more than 1.15 billion users.

2005

- The Apache Hadoop project is created by Doug Cutting and Mike Caferella. The name for the project came from the toy elephant of Cutting's young son. The now-famous yellow elephant becomes a household word just a few years later and a foundational part of almost all big data strategies.
- The National Science Board recommends that the National Science Foundation (NSF) create a career path for "a sufficient number of high-quality data scientists" to manage the growing collection of digital information.

2007

- Apple releases the iPhone and creates a strong consumer market for smartphones.

2008

- The number of devices connected to the Internet exceeds the world's population.

2011

- IBM's Watson computer scans and analyzes 4 terabytes (200 million pages) of data in seconds to defeat two human players on the television show *Jeopardy!* (There is more about the show in Part Two.)
- Work begins in UnQL, a query language for NoSQL databases.
- The available pools in the IPv4 address space have all been assigned. IPv4 is a standard for assigning an Internet protocol (IP) address. The IPv4 protocol was based on a 32-bit number, meaning there are 2^{32} or 4.5 billion unique addresses available. This event shows the real demand and quantity of Internet-connected devices.

2012

- The Obama administration announces the Big Data Research and Development Initiative, consisting of 84 programs in six departments. The NSF publishes "Core Techniques and Technologies for Advancing Big Data Science & Engineering."
- IDC and EMC estimate that 2.8 ZB of data will be created in 2012 but that only 3% of what could be usable for big data is tagged and less is analyzed. The report predicts that the digital world will by 2020 hold 40 ZB, 57 times the number of grains of sand on all the beaches in the world.
- The *Harvard Business Review* calls the job of data scientist "the sexiest job of the 21st century."

2013

- The democratization of data begins. With smartphones, tablets, and Wi-Fi, everyone generates data at prodigious rates. More individuals access large volumes of public data and put data to creative use.

The events of the last 20 years have fundamentally changed the way data is treated. We create more of it each day; it is not a waste product but a buried treasure waiting to be discovered by curious, motivated researchers and practitioners who see these trends and are reaching out to meet the current challenges.

WHY THIS TOPIC IS RELEVANT NOW

You've read this far in the book because I expect you are looking for ideas and information to help you turn data into information and knowledge. What I hope you learn in the subsequent pages are strategies and concrete ideas for accomplishing your business objective or personal edification regarding how you can harness the data to better your situation, whether in the office, the home, or a fantasy football league.

You should also understand that this is not a new problem—data has always been "too big" to work with effectively. This problem has only been exacerbated as now individuals are generating so much more data than ever before as they go through their daily lives. This increase in data, however, has caused the information management industry to provide better solutions than ever on how to store, manage, and analyze the data we are producing.

In addition, we also have more opportunity to engage with data. A simple example that is discussed in more detail in Part Two is the recommendations you get from Amazon. That small application at the bottom of its web pages

illustrates this point very well. In order to make these recommendations, Amazon can use a few different techniques that mostly center on three pieces of information; how you are similar to other shoppers, similar shoppers' opinions of the product you are viewing, and what product similar shoppers ultimately purchased. Alternatively, Amazon could make recommendations from an item point of view. Take, for example, my recent purchase of a baseball glove. The recommendations included items like baseball bats, baseballs, baseball glove oil, and other baseball-related equipment. These recommendations are based on item-to-item recommendations. Baseball gloves are usually sold with baseballs, bats, and glove oil so Amazon recommends them to me. The other method is to look at my profile and find users who have purchased similar items or have similar details as they relate to Amazon and then recommend to me what they purchased. To be effective at making recommendations requires a real commitment to recording, storing, and analyzing extremely large volumes of data.

I think you only need to look up from this book to see a device that is generating data at this very moment. That data will soon be used to inform some business process, recommendation, or public safety issue. This not a future or theoretical problem, this is now.

IS BIG DATA A FAD?

Data! Data! Data! he cried impatiently. I can't make bricks without clay!

—Sherlock Holmes, "The Adventure in the Copper Beeches"

Many informed individuals in the analytics and information technology (IT) communities are becoming sensitive to the actual term "big data." It has no doubt been co-opted for self-promotion by many people and organizations with little or no ties to storing and processing large amounts of data or data that requires large amounts of computation. Aside from marketing and promotional mischaracterizations, the term has become vague to the point of near meaninglessness even in some technical situations. "Big data" once meant petabyte scale, unstructured chunks of data mined or generated from the Internet.

I submit that "big data" has expanded to mean a situation where the logistics of storing, processing, or analyzing data have surpassed traditional operational abilities of organizations; said another way, you now have too much data to store effectively or compute efficiently using traditional methods. This may also include having too little time to process the data to use the information to make decisions. In addition, big data means using all available data in your models, not just samples. Big data indicates the capability of using entire data sets instead of just segments as in years past. In the purview of wider popular usage, the definition of this term will likely continue to broaden.

For certain leaders in the field, "big data" is fast becoming, or perhaps always was, just data. These people are long accustomed to dealing with the large amounts of data that other fields are just beginning to mine for information. For evidence, look to the larger web companies and certain entities within the U.S. government that have been using extremely large amounts of unstructured data operationally for years, long before anyone ever coined the term "big data." To them, it was just "data." In addition, the big banks and insurance companies have been pushing up against the limits of commercial, column-oriented data storage technologies for decades, and to them this was just "data" too. Consider whether the scale at which Google is indexing all available information, or which the National Security Agency is recording, has really changed since before the term "big data" entered the popular lexicon. It was difficult to comprehend how much data this was before, and it is still is just as hard to comprehend. However, to these leaders in the field, dealing with it is just a day's work. The rest of world is now joining these industries in storing, computing, and analyzing these immense amounts of data, and now we have a word to describe it and a time period to reference. Figure I.1 shows the popularity of the term "big data" as it came into common usage beginning in 2011. Since the 1940s when computer was a job title or in the 1960s when file transfer involved moving a carton of punch cards from one location to another and hoping you did not trip, organizations have had data challenges. Today those challenges are just on a larger scale. Nearly every company must deal with these challenges or accept the idea that the company itself may become irrelevant. *Forbes* in 2013 published an article that said that companies without a big data strategy miss out on $71.2 million per year. If you could raise revenue over $70 million this year and each subsequent year, I am sure your future would be very bright in your organization. The key to capitalize on this opportunity is to have a well-thought-out strategy on big data and execute to the strategy.

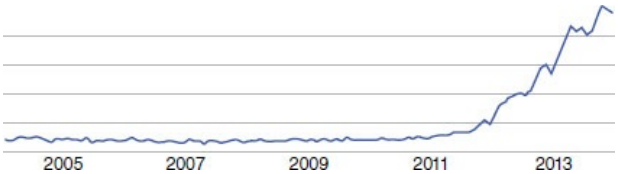

Figure I.1 Trend of Google Searches of "Big Data" over Time Showing the Popularity of the Term

Source: Google Trends

To be clear, the solutions surrounding the storage, processing, and analyzing "big data" are not a fad even if the term turns out to be one. Although some are overhyped right now, they are extremely valuable and in some cases actually revolutionary technologies. They have drawn such attention for this reason. Data is not magic—it's just a valuable raw material that can be refined and distilled into valuable specific insights. These insights are then converted to information that eventually creates knowledge.

The bigger the data, the more resource-intensive it is to work with, the better the value of the information must be to make the trade-off a wise business decision. While there is simply no underlying principle stating that the size of

data is positively correlated with its value, the size of a data set is positively correlated with its cost to maintain. The value of using big data is defined by how valuable the information gleaned from its process is compared to the time and resources it took to process that information.

This being said, there is a great deal of evidence that prudent data analysis can create value, whether knowledge or monetary, for most organizations. This is why the technologies that allow us to store, process, and analyze large amounts of data will continue to receive increased usage and are anything but a passing fad. For the leading analytic companies and entities, using large amounts of data has been a favorable value proposition for some time, and there is no reason for that trend to decelerate.

Data is the new oil. It has all of the same challenges in that it is plentiful but difficult and sometimes messy to extract. There are a few entities that control most of it, and a vast infrastructure has been built to transport, refine, and distribute it. With the performance of the required technologies increasing and their prices decreasing, if organizations currently struggle to deal efficiently with the medium-size data, they will be ill prepared and at a strategic disadvantage against their competition when data sizes increase, which they inevitably will do. If nothing else, companies, hospitals, and universities will face competition that will drive them to adopt technology to handle the growing volume of data. Other organizations, such as nongovernmental agencies, may be slower to invest in the new generation of data technologies and personnel. This is due to planning and implementation costs as well as the shortage of analytical professionals needed to produce value from this significant investment in hardware human capital; and some smaller organizations will not need sophisticated analysis to understand their operational environment.

WHERE USING BIG DATA MAKES A BIG DIFFERENCE

There have been so many news stories and hype about big data, and how it can transform your business, that it begins to sound like something you would find in Shangri-La. It is often portrayed as the answer to all things that cause problems for organizations. There are promises that it will identify the right customers for marketing campaigns and help academic institutions select the perfect students for their admissions process. Don't let skepticism turn you into a cynic.

It is indeed true that having more data, especially historical data, often will help model predictions be more accurate. Using additional sources of data, such as social networks, will help organizations make better predictions about customer choices and preferences, because all of us are influenced to some degree by those in our social network, either the physical or the virtual one.

Consider a situation where I have a poor customer experience with my cable provider; so poor that I cancel all of my services and look for another provider. Does this situation make my friends, family, and associates more likely or less likely purchase new services? How does that knowledge of my cancellation because of poor customer service, along with knowing my close friends, family, and associates, affect the cable provider's action? This is a prime example of big data in action; five to ten years ago, this type of analysis would not have been possible because the data sources just did not exist. The answer to my question of how this affects those who know is that my family and friends are less likely to add new services and potentially may follow suit and cancel their cable service as well. Having more data about your customers, products, and processes allows you to consider these types of effects in predicting customers' future behavior. It also needs to be pointed out that having the data and doing the analysis are vital steps to taking advantage of the opportunity in the big data era, but unless the organization is equipped to use new data sources and methods in their processes and act on this information, all of the data and the best analysis in the world will not help it improve.

There is danger in not taking the time to know how much weight to give to this large amount of newly available information, when that information is compared to all of the other attributes that affect a person's decision-making process. Returning to the poor customer service and my social network example, it is clear that the people in my social network are now more likely to cancel services, but how much more likely? One company that takes many different metrics and creates a single aggregate score is Klout; the higher your Klout score, the more influential you are online. Klout uses comments, mentions, retweets, likes and so on to create this score. The company is able to measure online influence only because that is all the data to which it has access.

The question of sampling and big data is a hotly debated topic, and I have read and heard on several occasions that sampling is dead. As a trained statistician and former employee of the U.S. Census Bureau, I would never say that sampling is dead or that it has no place in business today. Sampling is useful and valid. For certain types of problems, sampling from the population yields just as good a result as performing the same analysis using the entire population (all the data).

However, sampling cannot meet the objectives of many critical high-value projects, such as finding outliers. Companies that cling to sampling will miss out on opportunities to learn insights that can be found only by considering all the data. Outliers are one such example. In a statistical case, the term "outliers" usually has a negative connotation. But in many business problems, the outliers are your most profitable customers or the new market segments that can be exploited.

The best advice is to be an informed data user. Having seven years of customer history instead of three in August 2008 would not have helped in any way to predict people's upcoming spending habits in the United States. In just a few weeks, the financial markets were going to collapse and several large investment firms were going to declare bankruptcy or be purchased at fire sale prices. The U.S. government would begin a massive bailout of the financial markets that would, in some way, affect everyone in the industrialized world. No amount of data would have helped.

No amount of data analysis modeling of the preceding months' spending could forecast what the months after the bailout would look like for the average consumer. In order to build useful models in that time, you needed competent practitioners who understood how to simulate and adjust for economic conditions that no one in the workforce had seen before.

There are, however, two major advantages of using all the available data in solving your analytical business problems. The first is technical, and the other is productivity through an improved workflow for analysts.

Technical Issue

Many statistical and machine learning techniques are averaging processes. An averaging process is an algorithm or methodology that seeks to minimize or maximize the overall error and therefore make the best prediction for the average situation. Two examples of averaging algorithms are linear regression and neural networks. Both of these methods are explained in more detail in Part Two, but for now understand that regression seeks to minimize the overall error by fitting a line that minimizes the squared distance from the line to the data points. The square is used because the distances from the line to the data points will be both negative and positive. A neural network works by connecting all the input, or dependent variables, to a hidden set of variables and iteratively reweighting the connections between them until the classification of the holdout sample cannot be improved.

These averaging methods can be used on big data, and they will work very well. It is also very common for these methods to be very efficient in processing the data due to clever and persistent developers who have organized the computation in a way that takes advantage of modern computing systems. These systems, which have multiple cores, can each be working on a part of the problem; to get even greater speedups, multiple computers can be used in a distributed computing environment. Nonparametric techniques, such as a rank sign test, also fall into this category of an averaging model technique. I call this type of algorithm an averaging model process. Because we are seeking to average, sampling is a potential substitute that can provide comparable answers.

However, a second class of modeling problem is not averaging but is an extremity-based model or tail-based modeling process. These model types are used for a different objective and seek to find extreme or unusual values. These are sometimes referred to as outliers, but not in the strict statistical scenes. Instead, there are notable and unusual points or segments that are often the problems that are the most challenging to companies and carry the biggest return on investment in the present business climate. Examples of tail-based processes are fraud detection, offer optimization, manufacturing quality control, or microsegmentation in marketing.

Next I show why it is imperative to use complete data sets in these types of problems from several different industries and domains. In these cases and many others, these problems cannot be solved effectively without all the data, which is large and complex.

Fraud

If the rate of credit card fraud is 1 in 1,000 transactions,[2] and you sample 20% of your data to build your fraud model, it is likely that you will not have a single fraudulent activity. In fact, if you take anything less than all the data, you likely will never include the fraudulent activity. Predictive models work by using past performance to identify current behavior that has the same characteristics. Without having those fraudulent transactions, you will lack enough information to create an equation that can recognize fraud when it is happening. It is necessary to capture and analyze past fraudulent activity to create effective models for predicting future fraudulent activity.

In addition, if you are sampling the current incoming credit card transactions looking for fraud, you will miss those transactions that could have been flagged had they been present in the data being analyzed.

Optimization

For the optimization family of problems, consider the example of U.S. commercial airline traffic. If we want to understand the congestion points in the air traffic network, or model the effect of a tornado in Dallas, or a tropical storm in Miami, we need to use the entire data set. This is necessary to measure and accurately describe the scope of the problem and its effects on overall flight delays or cancellations and measure the costs to the industry. Imagine if a sample was taken in this case. First we would have to design a sampling scheme. Do you take a sequential sample of every tenth flight across the system? Do you weigh each airport by the number of commercial flights that occur there on the average day? Do you use a random sample? All of these schemes have shortfalls in seeing the complete picture, and those shortfalls will be most prominent in the calculation of the standard error and the confidence intervals of the prediction. To do these calculations correctly will take time to measure, calculate, and verify. If it is not done correctly, then your answer is wrong, but you will never know. The sampling also requires a front-loaded investment; before I can work with the sampled data, I have to invest the time to create the sample and validate that it is accurate. I also cannot see any type of result until a significant amount of work is already completed.

Relationship Management

Retention of telecommunications customers is a key component to revenue stream. Consider the challenge of predicting which customers are likely to cancel their contract and move to a different carrier. (This is referred to as attrition or churn.) It would be useful to know the typical behavior patterns of customers in the period of time before they cancel, including whom they call or from whom they receive calls. If you use only 10% of the calls each customer sends or receives, use only 10% of the customers, or look at only 10% of the numbers they call, you could be misled in predicting the true likelihood that a particular customer will cancel their contract. This is true for two reasons; first, since only a small percentage of customers leave each month, it would be probable that not a single

dissatisfied costumer (or even a whole segment of customers) would be included in the sample. It would also be possible that some of a dissatisfied customer's calls are not included in the sample. However, without the complete set of calls for a given customer, it is much more difficult to identify the pattern that you are looking for. (This is like working on a puzzle that is missing most of the pieces.) With the inability to identify those customers likely to cancel their contract, the problem will grow over time. Given the significant costs to acquire new customers in the telecommunications market, implementing an effective strategy to keep existing customers is worth millions and millions of dollars in annual revenue.

Work Flow Productivity

The second consideration is ensuring that the productivity of the analyst stays as high as possible. Analytics has become a very hot topic in the past few years, and predictions from McKinsey & Company project a shortfall of 140,000 to 190,000 people with the analytical expertise and 1.5 million managers needed to evaluate and make decisions based on big data. This translates to a deficit of 50% to 60% of the required personnel by the year 2018 in the United States alone. With this significant shortfall in capable people, the human capital you have already made in your organization needs to be preserved and improved. The analytical talent you already have in your organization will become more scarce as other organizations work to make better use of their big data through better analytics and governance. It will be critical to keep analytical talent engaged and productive.

From the same report:

> "Several issues will have to be addressed to capture the full potential of big data. Policies related to privacy, security, intellectual property, and even liability will need to be addressed in a big data world. Organizations need not only to put the right talent and technology in place but also structure workflows and incentives to optimize the use of big data. Access to data is critical—companies will increasingly need to integrate information from multiple data sources, often from third parties, and the incentives have to be in place to enable this." (McKinsey) www .mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation

As I mentioned in the prior section, the work to sample is a very front-loaded task; the majority of the work is done before any results can be created or exploration can begin. This is really backward from the optimum work flow. The best thing is to make the access to data exploration and quick modeling against the data simple and readily available. Organizations should enable the "failing fast" paradigm. Fail has a negative connotation, but failing fast is a useful strategy for determining which projects have merit and which do not. When people have the ability to work with the entire set of data, they can explore and prototype in a more efficient and natural way that does not require a great deal of up-front work to access the data. To enable this type of environment for your organization, ongoing commitments to capital and technological investments are required.

Making effective work flow and data computing resources for employees translates to large productivity gains and short timelines to pay back the return on investment. I have seen this transformation firsthand when I was working on a credit card modeling project for a large U.S. bank. Using the traditional methods (hardware and software), it was taking many hours to solve the problem. When I switched to a new distributed computing environment, I was able to solve the same problem in two to three minutes. I no longer had to multitask across so many projects because each one had significant downtime while models were being built. I was able to try a number of algorithms and tune each one to a degree that would not have been possible before. The work flow was reminiscent of class projects in school where data volumes were small and software ran nearly instantaneously. This was the method I had been trained in, and it felt more natural. I saw immediate benefits in the form of better model lift, which the customer saw as millions of dollars in revenue increases.

The Complexities When Data Gets Large

Big data is not inherently harder to analyze than small data. The computation of a mean is still just the sum of the values divided by the number of observations, and computing a frequency table still requires reading the data and storing the number of times a distinct value occurs in the data. Both of these situations can be done by reading the data only one time. However, when data volumes gets large or when the data complexity increases, analytics run times can grow to the point that they take longer to compute than the operational constraints will allow. This can result in misleading results or a failure to find a result at all.

Nonlinear Relationships

In real data, there are often nonlinear relationships between variables. Sometimes these relationships can be described "well enough" using linear relationships, but sometimes they cannot. A linear relationship is sometimes hard to imagine, so let us use exiting a parking lot as an example. My family and a few others attended a symphony performance and fireworks show for the Fourth of July (Independence Day in the United States). We parked near each other in the same section of the parking lot, which was about a seven-minute walk from the venue. After the fireworks concluded, our group made our way to the exit, but one of the families became separated. Instead of taking the closest exit to our seats that is open only after the event, they took a longer route through the venue to the main entrance where we had entered. This alternate route added about three minutes to their departure time. A parking lot after a major event is always something I, as a quantitative person, dread. I easily grow frustrated over inefficiency, and this exit situation is known to be poor and bordering on terrible. My family arrived at the car, loaded our cooler, chairs, and blankets, and began to drive to the exit. Traffic inside the parking lot was quite slow, because of poor visibility and all the pedestrian traffic leaving the venue. We proceeded to move with the traffic and following police direction made it home in about 40 minutes.[3] As we were arriving home, my wife received a text

message from our friends who had taken the other exit, asking if we were stuck in the parking lot like they were. So, while our drive took twice as long as it does on a normal day, those three extra minutes added not three minutes (which is what we would expect from a linear relationship between time of departure and time of arrival) to their drive but almost another 45 minutes to their drive home (in addition to the 40 minutes it took my family) from the event. This example is one many can relate to, and it illustrates an important point: Knowing about your data can be a huge asset in applying analytics to your data.

A second example of nonlinear relationship is that of the space shuttle Challenger disaster in 1986. Even though it has been almost 30 years, I still remember sitting in Mrs. Goodman's class in my elementary school, with eager anticipation as we were going to see the Challenger liftoff and take a teacher, Sharon McAuliffe, into space. Many of you know the tragic events of that day and the findings of the NASA Commission. To review the details, 73 seconds after liftoff, the primary and secondary O-rings on the solid-state boosters failed and caused an explosion due to excess hydrogen gas and premature fuel ignition. This resulted in the Challenger being torn apart. The reason the O-rings failed is blamed primarily on the weather. That January day was only about 30 degrees at launch time,[4] much colder than any space shuttle launch NASA had attempted before. The cold weather created a problem, because NASA personnel planned assuming a linear relationship between the air temperature and O-ring performance, but instead that relationship was nonlinear and the O-ring was actually much more brittle and ineffective than preplanning had anticipated. This was a tragic lesson to learn as it cost the lives of many remarkable people. After this incident, NASA changed a number of procedures in an effort to make space flight safer.[5]

In statistics, there are several terms of art to describe the shape or distribution of your data. The terms are: mean, standard deviation, skewness, and kurtosis. At this point, the important facts to understand and keep in mind are that: (1) there are often nonlinear relationships in real-world data; (2) as the data size increases, you are able to see those relationships more clearly; and, more frequently (3) nonlinear relationships can have a very significant effect on your results if you do not understand and control for them.

In Part One of this book, the focus is on the technology aspects of creating an analytical environment for data mining, machine learning, and working with big data and the trade-offs that result from certain technology choices. In Part Two, the focus is on algorithms and methods that can be used to gain information from your data. In Part Three, case studies show how, by utilizing these new technology advances and algorithms, organizations were able to make big impacts. Part Three also illustrates that using high-performance computing, analytical staff productivity went up in meaningful ways.

NOTES

[1] A zettabyte is 1 billion terabytes.

[2] The incidence rate is actually much smaller, but this makes for easier math.

[3] The drive between my home and the venue takes about 20 minutes on the average day.

[4] The launch control parameters for the Space Shuttle were a 24-]hour average temperature of 41 degrees F and not warmer than 99 degrees F for 30 consecutive minutes.

[5] I got to witness the safety protocol in action. I was at Cape Canaveral for the launch of STS-111, but it was scrubbed with less than an hour before liftoff.