1

Fostering Smarter Colleges and Universities

Data, Big Data, and Analytics

JASON E. LANE AND B. ALEX FINSEL

ABSTRACT

Many sectors are looking for ways to harness data, particularly Big Data, to improve their operations to build smarter cities, smarter governments, smarter hospitals, and even a smarter planet. Yet, there has been little focus as to how this increasing interest in data actually can help build smarter colleges and universities. This chapter provides academic leaders with an introduction to Big Data and data analytics, exploring definition of terms, examining application to higher education operations, and raising related cautionary concerns.

We have reached a tipping point in the path of human evolution; billions of data points are being generated every minute of every day by humans, computers, and technological devices all around us, creating a real-time, digital footprint of our lives with every credit card swipe and smart phone use. With the availability of this ocean of information, the question becomes how to use the data to better engineer our world to serve our needs. This Big Data movement is transforming everything from healthcare delivery systems to the way cities provide services to citizens. Now is the time to examine how the Big Data movement could help build smarter universities—institutions that can use the huge amounts of data they generate to improve the student learning experience, enhance the research enterprise, support effective community outreach, and advance the campus's infrastructure. While much of the cutting-edge research being done with Big Data is happening at colleges and universities, higher education has yet to turn the digital mirror on itself to innovate the academic enterprise.

Let us put this data explosion in some perspective. From the dawn of civilization to 2003, humans created five exabytes worth of data. As of 2013, humans produced this same amount of information every two days (Miller & Chapin, 2013). The amount of information in the world today is so vast that it can be difficult to visualize, but one way to think about it is that "the world holds twice as many bytes of data as there are liters of water in all its oceans" ("The Big Bang," 2013). If harnessed, these data have the potential to create a feedback loop from human activity, which can be used to enhance institutional productivity and student success. This vast amount of information has become broadly labeled *Big Data*, a term that has come to symbolize the data revolution that we are now experiencing.

For all the hype about Big Data, no data—big or small—are useful unless they can be analyzed to develop meaning. It is similar to crude oil buried deep in the Bakken shale formation in western North Dakota. The formation is one of the largest repositories of oil in the world, but it remained nothing more than an interesting fact until the technological advances were discovered that allowed the oil to be extracted and refined. However, while the oil has become accessible and yielded great opportunities, the success of its extraction has been tarnished by the environmental dangers—many of which were unforeseen in the early years. The immense amount of data that is now being generated is only useful if it can be extracted and refined to be used to make decisions. And, in the same way that the new technologies used to extract the oil from the shale has a cautionary side, the extraction and use of Big Data also generate cautions of which we must be aware.

As the techniques to extract and refine Big Data improve, scientists at universities and corporations are learning how to use it to transform how we shop, work, and play. Analysis of these new data allows people to discover patterns that have been previously overlooked. One of the classic examples comes from when Target discovered that a teen girl was pregnant before her father did. By analyzing the shopping patterns of its customers, Target was able to predict fairly accurately which customers were pregnant and about when they were due. The company then sent coupons and flyers during key times of the pregnancy to encourage women to shop at Target for their pregnancy needs. One day, a father complained to a store manager that it was encouraging his teen daughter to get pregnant by sending her information on baby clothes and cribs. Two days later, the father apologized to the store manager, having just learned that his daughter was indeed pregnant (Hill, 2012).

Target is not alone in its close analysis of customer behavior. Walmart analyzed volumes of transaction data and discovered that consumers purchased a significant amount of Pop-Tarts in certain hurricane-prone areas during storm season. This information enabled the retail giant to increase its supply of Pop-Tarts, which led to increased sales and profits (Fourtané, 2013). Similarly, the Cincinnati Zoo chose to sell ice cream in the afternoon based on the rational expectation that warmer temperatures would lead to a higher demand for ice cream. However, zoo officials used Big Data analytics to determine that visitors significantly preferred ice cream earlier in the day, despite the conventional wisdom surrounding the impact of temperature. Accordingly, this meta-information helped the zoo to change when it offered ice cream to better satisfy customers and increase profits (Callaham, 2013; Vesset, 2013). Netflix analyzes the input of thousands of users to make personalized recommendations for what movie a customer might want to watch next. And, in early 2014, Amazon filed a patent for anticipatory package shipping, a process that would allow predictions of what some customers would buy next and then ship it to them before they actually purchased it (Matyszczky, 2014).

All these new data are being used in the development of smarter cities, smarter governments, and a smarter planet. So, why not smarter colleges and universities?

The intention of this volume is to unpack not just the phenomenon of Big Data but the corresponding renewed interest in how we analyze data of all sizes to build a smarter university. The remainder of this chapter is a primer on Big Data and data analytics and is intended to provide the reader with a basic understanding of related concepts, activities, and cautions. It is important to note from the outset that this volume does not focus on the more technical aspects of Big Data such as how to store and process large amounts of information; rather, it explores how colleges and universities might engage operationally with Big Data to improve student success and better understand the student pipeline.

UNPACKING BIG DATA

In the purest sense, the idea of Big Data is not new. It is actually a moving target. Big Data has always existed in that it is essentially data that exceed current standard abilities to manipulate them. Thus, much of the data that we easily and regularly analyze today was at one time considered Big Data. In fact, what Big Data is can even vary between different organizations as it is essentially data sets so large that they cannot be easily analyzed using available data management programs. What is different today, when the phenomenon is compared to previous eras, is that the amount of data being generated is growing at an unprecedented rate, and the utility of those data for understanding human behavior is unparalleled. Moreover, Big Data has generated a renewed awareness of the importance of using data to systematically improve the work of organizations. Therefore, not all of the data discussed in this volume may meet a purist's definition of Big Data, but it is part of the overall movement toward using data and analytics to improve how we work, play, and live.

Big Data has come to be described by five Vs: *volume*, *velocity*, *variety*, *veracity*, and *visualization*.[1] These terms are discussed in more depth in the following and provide a lens to understand the often ambiguous concept of Big Data, the technological equipment needed to process Big Data, and data themselves as both raw material and finished product. The first three Vs, volume, velocity, and variety, seek to describe the nature of Big Data. The other two Vs, veracity and visualization, are related to the outputs of Big Data.

Volume. There are a lot of data, and these data need to be stored. So how big is big? Although size is relative, the current data explosion creates a staggering amount of information on a scale that is difficult to comprehend. Again, humans create five exabytes of information every two days (Miller & Chapin, 2013). An exabyte is a billion gigabytes. (And a gigabyte is a billion bytes!) Thus, an exabyte is equivalent to a billion billion bytes, which is roughly equivalent to more than 4,000 times the information stored in the Library of Congress (McKinsey Global Institute, 2011). Accordingly, the unfamiliarly large petabyte, exabyte, and zettabyte are poised to become standard units of measure and more commonplace when discussing the size of today's Big Data volume. However, tomorrow's Big Data may extend well beyond the realm of yottabytes and force people to reconsider and redefine how we measure data's size (Foley, 2013). For many colleges and universities, administrative data sets this large may seem like fantasies (or nightmares) compared to the existing databases that track student enrollments and the like. Yet, many institutions now have software that can track how students engage with class materials available through course management software or which buildings they enter and when. These individual data points, when multiplied across an entire student population every day, can add up quickly.

Velocity. Even if one is able to store all of the data created and have room to spare, any storage space can be consistently constrained by the velocity, or the speed, at which data are created, processed, and/or transferred from one point to another. In other words, of importance is the velocity of the feedback loop—how quickly the new data can be harnessed and used to make decisions. At one time, the performance of a student in a given semester was not known until course grades were reported and a GPA calculated at the end of the semester. Now it is possible to track student activity on an almost daily basis and provide interventions in the middle of the semester—with the desire of supporting the success of the student.

Variety. Data can come from many different sources, and it rarely arrives in a form that is simple to process and through which leaders can easily make decisions. In the case of students, information can come from social networks, ID card usage, and the degree audit system. This information can then be married with more traditional data sets that are kept by the registrar or bursar, for example. This variety can make it difficult to realize the full potential of the data.

Identifying ways to get traditional and new data sets to talk with each other is a focus of chapter 4.

Veracity. A significant issue with the use of Big Data is determining its veracity in that the quality of the output depends on the quality of the input and the process used to refine it. Are the data being used meaningful to the analysis being performed? There is always the risk of collecting and analyzing flawed information, which could lead to flawed decision making. Such flaws may not be that significant when recommending a new movie, but they can have devastating effects if the wrong intervention is used to help an at-risk student.

Visualization. It can be difficult for individuals to "see" the patterns in the data. Thus, visualization is a critical part of the refining process that allows people to understand and use the knowledge created from Big Data. Visualization transforms abstract information into a physical image that has dimensions that humans can quickly see and understand, and from which they can extract meaning (Taylor, 2013). Although Big Data scientists have an intimate understanding of the Big Data process, visualization allows a greater number of lay people to access, understand, and use Big Data information to make decisions in day-to-day organizational operations. For example, eBay's online marketplace had almost 108 million users and sold $68 billion worth of merchandise in 2011. This activity generated 52 petabytes of user behavior, shipping, and online transaction data (Lampitt, 2012). Accordingly, visualization software transformed this large, diverse, and complicated data set into simple, insightful, and interactive graphics that employees can use to make decisions in real time.

DATA ANALYTICS, DATA MINING, AND MACHINE LEARNING

Indeed, the existence of data does not alone offer insights. Specific pieces of data need to be extracted and refined. How data are utilized has spurred a new industry of data mining, which is "the process of analyzing data from different perspectives and summarizing it into useful information—information that can be used to increase revenue, cut costs, or both. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases" (Palace, 1996).

Another development is that of predictive analytics (or what is referred to as *machine learning* in the literature), where computer programs use algorithms to analyze volumes of data patterns to make automated and semi-automated decisions. The potential power of predictive analytics is summed up in the title of Siegel's (2013) book: *Predictive Analytics: The Power to Predict Who Will Click, Buy, Lie, or Die*. While readers may not be concerned about buying, lying, and dying, analytics can also predict who will apply, remain, struggle, and complete. Analytics also has the power to determine what interventions may be best to help students to succeed, including providing real-time interventions.

For example, Arizona State University's (ASU) eAdvisor program actively monitors student degree progress, Facebook information, student ID card swipes, and performance patterns to identify at-risk students for immediate intervention and then provides support to keep them on track toward graduation (Parry, 2012). Since ASU launched eAdvisor in 2008, the proportion of freshmen who did not return for their sophomore year decreased from 24% to 16%. Moreover, 42% of ASU's students graduated in four years, which was an increase from 26% in 1997 (Marcus, 2012).

The bulk of the use of data mining and predictive analytics has been within the corporate world, where there may be some inspiration for opportunities in higher education. Both data mining and predictive analytics can help organizations to identify consumer needs and wants to provide custom, tailored services to maximize satisfaction. For example, Netflix used data mining to create a software algorithm that recommends movies and television shows to its consumers based on their interests, often with surprising accuracy (Feinleib, 2012). Taking this concept to another level, Netflix used its data analysis to help create content for its successful series *House of Cards* by identifying and then combining a popular director, actor, and plot premise (Carr, 2013).

BIG DATA AND STUDENT SUCCESS

Today, human activity creates a tremendous amount of information, often in real time. Much of this digital explosion is borne from the technology that people use on a daily basis, such as cell phones, tablets, Internet search engines, online shopping sites, social networks, and GPS navigators. These gadgets and associated activities generate volumes of "sensor-like" information that can be mined for value. Much of this information is passive and reflects traces of human activity in the form of "digital exhaust" or "digital smoke" (Loukides, 2010). Moreover, all these sensory data create a virtual nervous system for our planet that can allow people to know more about themselves, respond in real time (like a reflex action), and predict how people and organizations might respond to a given stimuli.

For decades, researchers have been trying to unpack the black box of the college experience, working to understand how college affects students (Pascarella & Terenzini, 1991, 2005). Data would usually be gathered from surveys about student activities and characteristics and then attempted to be linked to outcomes. Conclusions were drawn about the aggregate, not the individual, and interventions were near impossible to personalize. Interventions are, in many cases, based on data gathered years before, and assumptions about students are often based on group behavior rather than individual activity. These observations are not meant to be critiques of how higher education has operated; they simply describe a reality based on the availability of past data. Information was gathered from surveys and interviews; the need for sampling made it important that data were as pristine as possible; and conclusions were based on groups of people, not individual activity. And, academic leaders should be applauded for their efforts to use the legions of research on this topic to improve the success of their students.

Big Data creates knowledge, but only if it is used with purpose and direction. While health, business, and government have increasingly turned to Big Data to improve their work, higher education has been slow to embrace it. For higher education, the benefits of Big Data extend well beyond Pop-Tarts, ice cream, and presidential elections. Today, we can track the activities of students in real time. Student identification cards can allow us to know when and where students shop, eat, and engage in student activities such as concerts and lectures. Their use of ID cards allows us to know when they access their residence hall, enter a classroom, or go to the recreational center. Course management software records what readings a student accesses and how long he or she engages with the material. We can even track when students sign up for classes, how well they are doing, and whether an intervention is needed—all in real time. The digital footprint of today's college student can be vast, and higher education institutions have not yet realized the full potential of this tidal wave of data. If harnessed correctly, however, all this information can be used to extrapolate models of student success.

The benefits of Big Data for higher education include the potential to provide customized learning experiences, real-time interventions, and a greater awareness of how students progress from cradle to career. For example, companies such as ETS are already capturing student learning data "to develop predetermined learning trees to track certain responses to questions that imply mastery of specific aspects" (Guthrie, 2013). This powerful information can improve student success by providing an individualized course of study, an enhanced student-instructor relationship, and a closer digital community through distance learning.

The digital revolution allows colleges and universities to collect fine-grain detail on a large scale that can enhance operational decisions. For example, in the private sector, businesses may desire to identify their most loyal and profitable customers and "microtarget" them with advertising instead of relying on more costly conventional advertising that appeals to everyone or broader groups. Microtargeting uses data mining and algorithms to predict a person's attitude, sentiment, or behavior concerning a given subject. Each individual is assigned a score, which represents his or her statistical likelihood of engaging in a certain behavior; these scores help the organization prioritize which individuals to focus their attention on (Gavett, 2012). Microtargeting has also transformed political campaign operations by helping political parties to aim "specific ads at potential supporters based on where they live, the Web sites they visit and their voting records" (Vega, 2012). And, such techniques are now coming to higher education. As the pool of high school graduates shrinks in many parts of the nation, college recruiters, many of whom are experienced with using data from groups such as the College Board and ACT to target their marketing, are increasingly augmenting those data sources with information from the web in order to identify the students who are most likely to enroll and those who are most likely to pay the full tuition rate (Rivard, 2013). Similar techniques can also be used to target likely donors.

Beyond recruitment, some higher education institutions have been successful in using data to improve the success of students in the midst of their studies. In 2011, Arizona State University contracted with Knewton to use its Adaptive Learning Platform to facilitate freshman remedial math courses. The platform records various student input data (e.g., mouse clicks, logon times, correct and incorrect answers, selected distracters, time spent logged in, etc.) and uses a computer algorithm to process this information against a larger data set. The program identifies individual input data and then recommends a custom learning path tailored to the student's learning style, strengths, and habits. In short, the Adaptive Learning Platform assesses how individual students learn. This information is shared with the student and the professor, who can make optimal instructional decisions to better suit the individual learning needs of student. This information is also a powerful tool for students to identify their own strengths and learning styles, which can be conscientiously employed in other courses to facilitate academic success. In the 2011–2012 school year, the pass rates for 5,000 freshmen enrolled in remedial math classes using Knewton's platform increased from 66% to 75% (Kolowich, 2013).

Drawing inspiration from Netflix's ability to recommend the next movie that a user may want to watch based on what they have already watched and liked, Austin Peay State University developed an online course advising system called Degree Compass (Young, 2013; see also chapter 6). This program, created by the institution's provost, Tristan Denley, a former professor of mathematics, uses the student's planned major, previous academic performance, and data about the success of similar students to suggest which courses a student should take next. The results so far suggest that the program has led to higher grades and fewer dropouts. By crunching huge amounts of data, something that human advisers are not capable of doing, Degree Compass creates personalized recommendations for students. Human interaction in advising remains important, and Austin Peay still provides advisers for students, but the recommendations from Degree Compass mean that less time is spent on course selection, which frees more time for discussing other issues.

We should note, though, that for higher education, the drive to embrace Big Data is really about more than Big Data. The phenomenon, and this book, raises awareness of the potential benefits (and cautions) of Big Data and encourages a more broad-based drive to better use the data that currently exist—such as linking disparate databases and thereby making all of the data more useful. For example, there are new efforts underway in several states to link higher education data to that from the K–12 sector and the department of labor, using unit record indicators so that one could track the performance of a student from cradle to career (see chapter 11). In many cases, these data sets have existed for decades, siloed from each other due to political or bureaucratic obstacles or a simple lack of awareness. Now, states and at least one group of states are merging these data, which allows the performance of a student to be tracked from first grade through retirement, permitting a more nuanced understanding of the experience at each level, and how that might influence later success.

The full range of possibilities for improving higher education through Big Data is still unknown. This volume provides insights into how some campuses are using Big Data to enhance their work in the areas of student access, completion, and success.

BIG DATA AND RESEARCH

Universities have been at the forefront of the data revolution, developing new analytical techniques to better understand a whole host of important societal issues. The Visualization Center at San Diego State University used thousands of pictures collected from a free app to develop a map of the impact of the BP oil spill on the Gulf of Mexico coastline. The Structural Analysis of Large Amounts of Music Information project, a collaboration between researchers at the University of Illinois, the University of Southampton, and McGill University, is collecting about 23,000 hours of digital music to allow researchers to study the structure of music. In 2005, researchers at the University of Memphis, in partnership with IBM and the City of Memphis, developed Blue CRUSH (Criminal Reduction Utilizing Statistical History) to track patterns of crime and predict when (and where) potential crime would occur to more efficiently direct police resources. The result was a 31% decrease in serious crime in the city of Memphis since 2006 (Henschen, 2010).

All these efforts would not be possible without the infrastructure investment and technological support from universities. While this book does not touch much upon the work of scholars to utilize Big Data in their research, it is important for academic administrators to be aware of their responsibility to support faculty work in this area. Moreover, the data and technology we are dealing with now are unlike much of what have been used by colleges and universities in the past. Inherent in the Big Data movement is the fact that it is difficult, if not impossible, for existing technology to store and analyze the amount of data that is now available. Institutions need to develop or obtain access to data warehouses to store the massive amounts of data that continue to grow rapidly. New techniques are needed to capture the data, whether it is downloading from social networks or uploading from an iPad used in health trials in Africa. New technology platforms are necessary to crunch the data and find meaningful patterns. And staff are needed who can work on infrastructure, analytical models, data sources, and application development.

The data and technology being used are also very different from previous generations, so existing staff members may not have the expertise to support these new initiatives. Academic leaders who are interested in providing a Big Data infrastructure for scholars (as well as building in-house analysis capacities) should be willing to invest in new staff and/or professional development for existing staff.

This new infrastructure can be quite expensive, so some institutions develop partnerships with business and industry to develop Big Data centers, such as the Cloud and Big Data Laboratory at the University of Texas at San Antonio. These centers tend to support cloud computing and provide the processing power necessary to work through a terabyte of data. Such centers can be advantageous as they provide universities with capacity to support the work of their faculty, while drawing on partners, who also have access to the facilities, to help cover the associated costs. Such arrangements are increasing since many organizations cannot fully use the capacity of such a center alone and recognize the value of partnerships. Moreover, colleges and universities tend to be viewed as neutral territory, allowing multiple companies to come together to support such a center without any one of them having the advantage of the investment being at their location.

BIG DATA AND EDUCATION: CREATING DATA SCIENTISTS

The data revolution has also created the need for colleges to prepare students for what the *Harvard Business Review* has called the sexiest new job of the 21st century: the data scientist (Davenport & Patil, 2012). Davenport and Patil (2012) describe this person as "a high-ranking professional with the training and curiosity to make discoveries in the world of big data." This generic description represents the persisting ambiguity in defining the exact nature of the role, but current job data suggest that there will be high demand for individuals with this general skill set. The McKinsey Global Institute (2011) reports that the demand for data scientists is expected to rise and exceed the available supply resulting in a 50–60% talent gap in the United States by 2018. This gap includes the unfulfilled need for "140,000 to 190,000 people with deep analytical skills as well as 1.5 million managers and analysts to analyze big data and make decisions based on their findings" (McKinsey Global Institute, 2011, pp.10–11).

The data scientists have a critical role in leveraging Big Data. First, they possess a broad range of technical skills, which includes computer science, programming, modeling, data management, advanced mathematics, and statistics. Additionally, data scientists possess strong business (or organizational) domain knowledge that enables them to become problem-solving artists. Data scientists know how to prioritize problems, what questions to ask, what to look for during data collection and analysis, and which Big Data tools to utilize. In this process they explore and examine data from a variety of sources and analyze Big Data challenges using multiple perspectives. IBM vice president of Big Data products Anjul Bhambhri remarked, "A data scientist is somebody who is inquisitive, who can stare at data and spot trends. It's almost like a Renaissance individual who really wants to learn and bring change to an organization" (Bhambhri, n.d.).

Data scientists are a combination of computer scientist, statistician, philosopher, and English major. Needless to say, many traditional academic programs do not provide the entirety of the skill set expected of such individuals. One way to overcome this shortcoming is to develop interdisciplinary programs such as the Institute for Advanced Analytics at North Carolina State University, where students can pursue a one-year master of science degree in analytics that includes courses dealing with data management and quality, mathematical and statistical methods for data modeling, and techniques for visualizing data in support of enterprise-wide decision making. More than a simple number cruncher, the data scientist needs to be able to bridge the communications gap between the detailed technical processes of Big Data and the individuals involved in decision-making processes. That is, they must have the ability to simplify, articulate, and present information derived from the data. Consequently, data scientists need to have some entrepreneurial insights as high-ranking professionals with the training and curiosity to make discoveries in the world of Big Data (Davenport & Patil, 2012).

Data scientists have to use creative problem-solving techniques to make enormous and complex problems smaller and simpler. One way to accomplish this is through "data jujitsu" or "the art of using multiple data elements in clever ways to solve iterative problems that, when combined, solve a data problem that might otherwise be intractable" (Patil, 2012). Integrating heterogeneous data from various sources to explore a Big Data problem is also known as "data mashup." This term reflects the data scientist's ability to overcome and embrace the "variety" constraint of Big Data.

After identifying various data sources and determining how to piece them together, data scientists must clean, condition, and prepare raw data for analysis. For example, privacy concerns over certain types of data, such as medical and financial data, require data scientists to "anonymize" or strip away sensitive personal information from collected data. Data conditioning may also require "tagging" certain data elements, which helps to organize and compare information during analysis. Data scientists must also confront the challenge of data veracity by effectively scrutinizing the quality and completeness of collected data.

HIGHER EDUCATION SYSTEMS: TAKING DATA ANALYTICS TO SCALE

Much of this volume focuses on the possibilities for Big Data at the campus level, yet systems and states are well poised to take advantage of the massive amounts of data now being produced in ways no single campus can. Systems and states provide an opportunity both to better understand the academic ecosystem and to provide more refined interventions based on the amount of data available across multiple campuses.

While this data movement will allow colleges and universities to better understand student behavior within their institutions, systems can amalgamate data from across multiple campuses. This practice allows systems to track the mobility (and success) of students as they move across campuses. Working with other agencies, the opportunity exists to follow students from when they begin school through their engagement in the workforce, enhancing understanding of how educational opportunities throughout the pipeline affect persistence in the educational system and performance in the workforce. Moreover, the scale of data available to a system provides for strategic decision making as well as development of more robust predictive models that can be used to improve student success.

As discussed in the foreword to this volume and in chapter 7, the State University of New York (SUNY) is harnessing data from its 64 campuses and 463,000 students to enhance understanding of student mobility. In fact, through some of its initial analysis, the system realized that a large percentage (26.2%) of students who transfer within the system transfer from a four-year institution to a two-year institution. This revelation resulted in retooling of SUNY's transfer policies to provide support for students who engage in this reverse transfer rather than seeing transfer as only an upward or horizontal practice. If all multi-campus systems of higher education in the United States were to pursue this level of multi-institutional analysis, we would have a much greater understanding of how a vast majority of students in public four-year institutions experience higher education.

Similar efforts are being developed through collaborations by higher education institutions, state education departments, and state labor departments to share unit-level data across agencies (see chapter 11). As opposed to looking at patterns of movement across institutions, these sorts of arrangements permit one to see how a student moves from cradle to career. Not only can one better understand how students progress through their educational paths, the data also allow for analyzing how different experiences along individuals' paths may affect their career choices, employment opportunities, and benefits accrued from their workplaces.

Not all the data used in such analyses may be Big Data in the strictest sense, but using data to look at trends across institutions and across lifetimes is an important aspect of the new way in which data are being aggregated to inform decision making.

THE CAUTIONARY SIDE

How does one safely harness the ability to predict who is likely to attend a college, complete a degree, and be successful in the workforce? What if that same ability allowed one to predict who would likely be a successful athlete, a frequent visitor of the campus judicial affairs office, or a defaulter on student loans? What if the information used also allowed one to discover private personal information about a student, such as their sexual orientation or online proclivities? Such questions are important to consider, as they frame important ethical considerations related to the increasing availability of data and their use to predict the future.

Fans of superhero comic books are likely have to have heard the phrase "With great power comes great responsibility." This saying is no less relevant to the power of data being discussed throughout this volume. The analysis of these massive data sets provides the opportunity to peer into individuals' likely futures and make decisions based upon those likelihoods. The analysis may also dictate certain organizational decisions that may have unintended consequences for individuals. For example, what if an institution is able to predict which students are likely to drop a class? Should such a student be dissuaded from enrolling as a way to save the student's money and the faculty member's time? If the faculty member is informed, will he or she ignore the student, thinking that interactions are wasted effort, or will the professor give special attention to the student in the hope of overcoming the predicted outcome? The ability to predict who is likely to drop a class can provide an opportunity to intervene and encourage student success, but such profiling may also create a self-fulfilling prophecy in that a student targeted as being at risk may end up acting that way even if he or she was not really at risk. As discussed in chapter 3, a myriad of concerns exist of which academic leaders need to be aware. We highlight four key areas here: data integrity, privacy, removing choice, and profiling.

Data integrity. There are many potential problems with the integrity of data. We address the primary issues here. In the digital realm, there are two types of data: generated and volunteered. Generated data come from the activities in which one engages, such as the type of course material one accesses via course management software. This type of data tends to be fairly accurate, although it is possible for someone to be acting as another student's digital persona. Volunteered data are actively created by a person (or a different person acting as that person). In this case, people choose which information to volunteer, and self-divulged information may not always be accurate—which may result in an inauthentic digital representation of themselves. Thus, programs that use data from Facebook and other social networks may not be capturing an accurate image of a student. Known as "the big lie," data made available on social media sites may reflect a gap between private reality and public social norms and opinions (DiResta, 2013). Accordingly, those engaging in data analysis need to be aware of this information fallacy since there is an important difference between generated and volunteered data.

Privacy. Most people realize that the digital era has diminished individual privacy. Many of us now make available pictures, opinions, and current locations on a regular basis to friends and strangers through such social networks as Facebook, Twitter, and Instagram. However, many people do not know that the data produced from their activities are being retained by organizations and refined for use in a variety of purposes. Student IDs can tell us what students eat (meal plans) and where they go (concerts, residence hall). Course management software allows us to know which course materials students access and how long they engage with it. It is also possible to discover other private information about students that they may not willingly disclose or want college administrators to know. These types of data give great power to those who possess them, and careful consideration should be given to who has access to such data and how the information is used.

Removing choice. Predictive analytics can help academic leaders make better-informed decisions about how to help students succeed. Many companies now track Internet searches, Facebook "likes," and GPS data to help measure and predict consumer preferences, which allows an organization to present preference-based options to its consumers. As discussed elsewhere in this chapter and in chapter 6, some campuses are doing similar analysis to advise students about which courses to take.

Predictive analytics may simplify the decision-making process and increase the likelihood of student success, but does the predictive process undermine a student's choice? For example, prediction relieves the individual from the full burden of the decision-making process, which includes gathering information, identifying and prioritizing alternatives, and carefully assessing the benefits against the costs. Thus, optimized selections provided by predictive analytics simultaneously simplifies (and optimizes) individual choices, but it also supplants individual critical thinking.

Profiling. Colleges and universities have already begun using predictive analytics to target students who may be at risk of failing a class or dropping out of school (see chapter 6). These predictions have allowed institutions to implement interventions, sometimes in real time, to help improve student success. While such predictions and interventions may seem to be a silver bullet for improving the rate of student completion, there are potential downsides that must be considered. For example, predicting a student's likelihood for success could dissuade institutions from marketing themselves to a student who may not have performed well in high school, thus reducing access. Similarly, in a technological version of what Burton Clark (1960) called the *cooling out* effect, if an institution predicts that a current student has a high likelihood of dropping out before completing his or her degree, the interventions provided may encourage the student to withdraw early rather than helping them persist through the end of their degree program. Moreover, these predictive models incorporate advanced statistical computer algorithms, which may not be accurate, or the outcomes could be used to implement interventions, for example, for which the algorithm was not originally designed. Analytics is a diagnostic tool that must be correctly matched to a given organizational problem.

Using Big Data and data analytics to create a smarter university is filled with ethical issues and measured risks. It is always important for academic leaders to remember that there is a huge difference between trusting Netflix to accurately recommend a movie and a university to accurately recommend a particular course and/or degree path.

THE WAY FORWARD

> "You don't need these big platforms. You don't need all this big fancy stuff. If anyone says 'Big' in front of it, you should look at them very skeptically . . ."
>
> —Harper Reed, CTO Obama for America 2012

Inclusion of this quote may seem strange for a book that incorporates "Big" data in its name. This remark comes from a lecture given at SUNY's 2013 conference on the role of Big Data in transforming higher education.[2] The speaker's message was that issues around Big Data had become distorted as companies and others invoked the term to sell their services and products to those scared by the concept. The reality, Harper Reed said, was that not everything about the current data revolution was "big" and not everything labeled "big" really deserved that description. And, before being scared into buying lots of new equipment and software, institutions should pause and determine what data they want to use and how they want to use those data.

In fact, one of the primary goals of this book is to decode a lot of the mystery that currently surrounds the data revolution, including the hyperfocus on Big Data. There is little doubt that there is something different about the world in which we now live. There is more information today than one could fathom two decades ago; and new information is being produced at an unprecedented rate. This situation makes it very difficult for current technologies to store and process data, but for those who can harness the data, the insights are unparalleled. As discussed throughout the remaining chapters of this volume, colleges and universities can know more than ever before about how their students study, play, and work. They can predict who is likely to drop a course or drop out of school and provide interventions in the real time. Software can now advise students as to which classes are most useful for their course of study and how well they are likely to do, similar to how Amazon uses customers' information to suggest what else they might want to buy.

Such power raises many ethical questions around privacy, profiling, and individual choice. The data that are now available allow, for those who control them, knowledge and insights about individuals that can reveal a great deal of insights about them. How the data are used can provide great benefits to student success, but if the data are inaccurate, or the algorithms used to process them are inappropriate, the interventions provided could do more harm than good. And, by limiting the choices available to students, it may remove personal choice and exploration from the college experience.

The bottom line, however, is that we are too early in this data revolution to fully understand its opportunities and challenges. There is no doubt that colleges and universities are using data to improve what they do, and such efforts will only increase in the future. Thus, academic leaders need to have a basic understanding of the issues surrounding Big Data and its implications for higher education.

NOTES

1. The first three Vs of Big Data (i.e., volume, velocity, and variety) were first reported by Laney (2001). Over the next several years, other V words have been added. We selected to include veracity and visualization as they seemed most applicable to the topic at hand. Some writers have included value as an additional descriptor.
2. SUNY hosts an annual Critical Issues in Higher Education conference. In 2013, the conference—entitled Building a Smarter University: Big Data, Ingenuity, and Innovation—focused on the role of Big Data in advancing student success, institutional infrastructure, and research.

REFERENCES

Bhambhri, A. (n.d.). *What is a data scientist*. Retrieved from IBM website: http://www-01.ibm.com/software/data/infosphere/data-scientist/

Callaham, J. (2013). *Microsoft's super-long infographic gives us the data on Big Data*. Retrieved from Neowin website: http://www.neowin.net/news/microsofts-super-long-infographic-gives-us-the-data-on-big-data

Carr, D. (2013, February 24). Giving viewers what they want. *New York Times*. Retrieved from http://www.nytimes.com/2013/02/25/