246 Part Two Information Technology Infrastructure

How Reliable Is Big Data? CASE STUDY

oday’s companies are dealing with an ava-lanche of data from social media, search, and sensors, as well as from traditional sources.

According to one estimate, 2.5 quintillion bytes of data per day are generated around the world. Making sense of “big data” to improve decision making and business performance has become one of the pri-mary opportunities for organizations of all shapes and sizes, but it also represents big challenges.

Businesses such as Amazon, YouTube, and Spotify have flourished by analyzing the big data they collect about customer interests and purchases to create mil-lions of personalized recommendations. A number

of online services analyze big data to help consum-ers, including services for finding the lowest price on autos, computers, mobile phone plans, clothing, air-fare, hotel rooms, and many other types of goods and services. Big data is also providing benefits in sports (see the chapter-opening case), education, science, health care, and law enforcement.

Analyzing billions of data points collected on pa-tients, healthcare providers, and the effectiveness of prescriptions and treatments has helped the UK National Health Service (NHS) save about 581 mil-

lion pounds (U.S. $784 million). The data are housed in an Oracle Exadata Database Machine, which can quickly analyze very large volumes of data (review this chapter’s discussion of analytic platforms). NHS has used its findings from big data analysis to create dashboards identifying patients taking 10 or more medications at once, and which patients are taking too many antibiotics. Compiling very large amounts of data about drugs and treatments given to cancer patients and correlating that information with pa-tient outcomes has helped NHS identify more effec-tive treatment protocols.

New York City analyzes all the crime-related data it collects to lower the crime rate. Its CompStat

crime-mapping program uses a comprehensive city-wide database of all reported crimes or complaints, arrests, and summonses in each of the city’s 76 precincts to report weekly on crime complaint and arrest activity at the precinct, patrol borough, and citywide levels. CompStat data can be displayed on maps showing crime and arrest locations, crime

hot spots, and other relevant information to help

precinct commanders quickly identify patterns and trends and deploy police personnel where they are most needed. Big data on criminal activity also pow-ers New York City’s Crime Strategies Unit, which tar-gets the worst offenders for aggressive prosecution. Healthcare companies are currently analyzing big data to determine the most effective and economi-cal treatments for chronic illnesses and common diseases and provide personalized care recommenda-tions to patients.

There are limits to using big data. A number of companies have rushed to start big data projects without first establishing a business goal for this new information or key performance metrics to measure success. Swimming in numbers doesn’t necessarily mean that the right information is being collected

or that people will make smarter decisions. Experts in big data analysis believe too many companies, se-duced by the promise of big data, jump into big data projects with nothing to show for their efforts. They start amassing mountains of data with no clear objec-tive or understanding of exactly how analyzing big data will achieve their goal or what questions they are trying to answer. Organizations also won’t benefit from big data that has not been properly cleansed, organized, and managed—think data quality.

Just because something can be measured doesn’t mean it should be measured. Suppose, for instance, that a large company wants to measure its web-

site traffic in relation to the number of mentions on Twitter. It builds a digital dashboard to display the results continuously. In the past, the company had generated most of its sales leads and eventual sales from trade shows and conferences. Switching to Twitter mentions as the key metric to measure changes the sales department’s focus. The depart-

ment pours its energy and resources into monitoring website clicks and social media traffic, which pro-duce many unqualified leads that never lead to sales.

Although big data is very good at detecting corre-lations, especially subtle correlations that an analysis of smaller data sets might miss, big data analysis doesn’t necessarily show causation or which correla-tions are meaningful. For example, examining big data might show that from 2006 to 2011 the United States murder rate was highly correlated with the

Chapter 6 Foundations of Business Intelligence: Databases and Information Management 247

market share of Internet Explorer, since both de-clined sharply. But that doesn’t necessarily mean there is any meaningful connection between the two phenomena. Data analysts need some business knowledge of the problem they are trying to solve with big data.

Big data predictive models don’t necessarily give you a better idea of what will happen in the future. Meridian Energy Ltd., an electricity genera-tor and distributor operating in New Zealand and Australia, moved away from using an aging predic-tive equipment maintenance system. The software was supposed to predict the maintenance needs

of all the large equipment the company owns and operates, including generators, wind turbines, trans-formers, circuit breakers, and industrial batteries. However, the system used outdated modeling techniques and could not actually predict equipment failures. It ran simulations of different scenarios and predicted when assets would fail the simulated tests. The recommendations of the software were use-

less because they did not accurately predict which pieces of equipment actually failed in the real world. Meridian eventually replaced the old system with IBM’s Predictive Maintenance and Quality software, which bases predictions on more real-time data from equipment.

All data sets and data-driven forecasting models reflect the biases of the people selecting the data and performing the analysis. Several years ago, Google developed what it thought was a leading-edge algo-rithm using data it collected from web searches to de-termine exactly how many people had influenza and how the disease was spreading. It tried to calculate the number of people with flu in the United States by relating people’s location to flu-related search que-ries on Google. Google consistently overestimated flu rates, when compared to conventional data collected afterward by the U.S. Centers for Disease Control (CDC). Several scientists suggested that Google was “tricked” by widespread media coverage of that year’s severe flu season in the United States, which was fur-ther amplified by social media coverage. The model developed for forecasting flu trends was based on a flawed assumption—that the incidence of flu-related searches on Googles was a precise indicator of the number of people who actually came down with the flu. Google’s algorithm only looked at numbers, not the context of the search results.

In addition to election tampering by hostile na-tions, insufficient attention to context and flawed

assumptions may have played a role in the failure of most political experts to predict Donald Trump’s victory over Hillary Clinton in the 2016 presiden-tial election. Trump’s victory ran counter to almost every major forecast, which had predicted Clinton’s chances of winning to be between 70 to 99 percent.

Tons of data had been analyzed by political ex-perts and the candidates’ campaign teams. Clinton ran an overwhelmingly data-driven campaign, and big data had played a large role in Obama’s victo-ries in 2008 and 2012. Clinton’s team added to the database the Obama campaigns had built, which connected personal data from traditional sources, such as reports from pollsters and field workers, with other data from social media posts and other online behavior as well as data used to predict con-sumer behavior. The Clinton team assumed that the same voters who supported Obama would turn out for their candidate, and focused on identifying vot-ers in areas with a likelihood of high voter turnout. However, turnout for Clinton among the key groups

who had supported Obama—women, minorities, col-lege graduates, and blue-collar workers—fell short of expectations. (Trump had turned to big data as well, but put more emphasis on tailoring campaign mes-sages to targeted voter groups.)

Political experts were misled into thinking Clinton’s victory was assured because some predic-tive models lacked context in explaining potentially wide margins of error. There were shortcomings in polling, analysis, and interpretation, and analysts did not spend enough time examining how the data used in the predictive models were created. Many polls used in election forecasts underestimated the strength of Trump’s support. State polls were inac-curate, perhaps failing to capture Republicans who initially refused to vote for Trump and then changed

their minds at the last moment. Polls from Wisconsin shortly before the election had put Clinton well ahead of Trump. Polls are important for election predictions, but they are only one of many sources of data that should be consulted. Predictive models were unable to fully determine who would actually turn out to vote as opposed to how people thought they would vote. Analysts overlooked signs that Trump was forging ahead in the battleground states. Britain had a similar surprise when polls mistakenly predicted the nation would vote in June 2016 to stay in the European Union.

And let’s not forget that big data poses some challenges to information security and privacy.

248 Part Two Information Technology Infrastructure

As Chapter 4 pointed out, companies are now ag-gressively collecting and mining massive data sets on people’s shopping habits, incomes, hobbies, residences, and (via mobile devices) movements from place to place. They are using such big data to discover new facts about people, to classify them based on subtle patterns, to flag them as “risks” (for

example, loan default risks or health risks), to predict their behavior, and to manipulate them for maxi-mum profit.

When you combine someone’s personal infor-mation with pieces of data from many different sources, you can infer new facts about that person (such as the fact that they are showing early signs of Parkinson’s disease, or are unconsciously drawn toward products that are colored blue or green). If asked, most people might not want to disclose such information, but they might not even know such in-formation about them exists. Privacy experts worry that people will be tagged and suffer adverse con-sequences without due process, the ability to fight back, or even knowledge that they have been dis-criminated against.

*Sources:* Linda Currey Post, “Big Data Helps UK National Health Service Lower Costs, Improve Treatments,” *Forbes*, February 7, 2018; Michael Jude, “Data Preparation Is the Key to Big Data Success,” *InfoWorld*, February 8, 2018; Rajkumar Venkatesan and

Christina Black, “Using Big Data: 3 Reasons It Fails and 4 Ways to Make It Work,” University of Virginia Darden School of Business Press Release, February 8, 3018; Ed Burns, “When Predictive Models Are Less Than Presidential,” *Business Information*, February 2017; Aaron Timms, “Is Donald Trump’s Surprise Win a Failure of Big Data? Not Really,” *Fortune*, November 14, 2016; Steve Lohr and Natasha Singer, “The Data Said Clinton Would Win. Why You Shouldn’t Have Believed It,” *New York Times*, November 10, 2016; Nicole Laskowski and Niel Nikolaisen: “Seven Big Data Problems and How to Avoid Them,” *TechTarget Inc.*, 2016; Joseph Stromberg, “Why Google Flu Trends Can’t Track the Flu (Yet),” [smithsonianmag.com](http://smithsonianmag.com), March 13, 2014; and Gary Marcus and Ernest Davis, “Eight (No, Nine!) Problems With Big Data,” *New York Times*, April 6, 2014.

CASE STUDY QUESTIONS

6-13 What business benefits did the organizations described in this case achieve by analyzing and using big data?

6-14 Identify two decisions at the organizations described in this case that were improved by using big data and two decisions that big data did not improve.

6-15 List and describe the limitations to using big data.

6-16 Should all organizations try to collect and analyze big data? Why or why not? What man-agement, organization, and technology issues should be addressed before a company decides to work with big data?

**MyLab MIS**

Go to the Assignments section of MyLab MISto complete these writing exercises.

6-17 Identify the five problems of a traditional file environment and explain how a database management system solves them.

6-18 Discuss how the following facilitate the management of big data: Hadoop, in-memory computing, ana-lytic platforms.

Chapter 6 References

Aiken, Peter, Mark Gillenson, Xihui Zhang, and David Rafner. “Data Management and Data Administration: Assessing

25 Years of Practice.” *Journal of Database Management* (July-September 2011).

Beath, Cynthia, Irma Becerra-Fernandez, Jeanne Ross, and James Short. “Finding Value in the Information Explosion.” *MIT Sloan Management Review* 53, No. 4 (Summer 2012).

Bessens, Bart. “Improving Data Quality Using Data Governance.” *Big Data Quarterly* (Spring 2018).

Buff, Anne. “Adapting Governance to the Changing Data Land-scape.” *Big Data Quarterly* 3, No. 4 (Winter 2017).

Bughin, Jacques, John Livingston, and Sam Marwaha. “Seizing the Potential for Big Data.” *McKinsey Quarterly* (October 2011).

Caserta, Joe, and Elliott Cordo. “Data Warehousing in the Era of Big Data.” *Big Data Quarterly* (January 19, 2016).

Chai, Sen, and Willy Shih. “Why Big Data Isn’t Enough.” *MIT Sloan Management Review* (Winter 2017).

Clifford, James, Albert Croker, and Alex Tuzhilin. “On Data Repre-sentation and Use in a Temporal Relational DBMS.” *Informa-tion Systems Research* 7, No. 3 (September 1996).

DalleMule, Landro, and Thomas H. Davenport. “What’s Your Data Strategy?” *MIT Sloan Management Review* (Winter 2017).