Can We Trust Big Data?

CASE STUDY

Today's companies are dealing with an avalanche 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 primary opportunities for organizations of all shapes and sizes, but it also represents big challenges.

Big data helps streaming music service Spotify create a service that feels personal to each of its 75 million global users. Spotify uses the big data it collects on user listening habits (more than 600 gigabytes daily) to design highly individualized products that captivate its users around a particular mood or moment in time rather than offering the same tired genres. Users can constantly enhance their listening experience with data-driven features such as the Discovery tool for new music, a Running tool that curates music timed to the beat of their workout, and Taste Rewind—which tells what they would have listened to in the past by analyzing what they listen to now. By constantly using big data to fine-tune its services, Spotify hopes to create the perfect user experience.

A number of services have emerged to analyze big data to help consumers. There are now online services to enable consumers to check thousands of different flight and hotel options and book their own reservations, tasks previously handled by travel agents. For instance, a mobile app from Skyscanner shows deals from all over the web in one list—sorted by price, duration, or airline—so travelers don't have to scour multiple sites to book within their budget. Skyscanner uses information from more than 300 airlines, travel agents, and timetables and shapes the data into at-a-glance formats with algorithms to keep pricing current and make predictions about who will have the best deal for a given market.

Big data is also providing benefits in law enforcement (see this chapter's Interactive Session on Organizations), sports, education, science, and health care. A recent McKinsey Global Institute report estimated that the U.S. healthcare system could save $300 billion each year—$1,000 per American—through better integration and analysis of the data produced by everything from clinical trials to health insurance transactions to "smart" running shoes. Healthcare companies are currently analyzing big data to determine the most effective and economical treatments for chronic illnesses and common diseases and provide personalized care recommendations 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. Swimming in numbers and other data 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, seduced by the promise of big data, jump into big data projects with nothing to show for their efforts. They start amassing and analyzing mountains of data without no clear objective or understanding of exactly how analyzing big data will achieve their goal or what questions they are trying to answer. Darian Shirzai, founder of Radius Intelligence Inc., likens this to haystacks without needles. Companies don't know what they're looking for because they think big data alone will solve their problem.

According to Michael Walker of Rose Business Technologies, which helps companies build big data systems, a significant majority of big data projects aren't producing any valuable, actionable results. A recent report from Gartner, Inc. stated that through 2017, 60 percent of big data projects will fail to go beyond piloting and experimentation and will eventually be abandoned. This is especially true for very large-scale big data projects. Companies are often better off starting with smaller projects with narrower goals.

Hadoop has emerged as a major technology for handling big data because it allows distributed processing of large unstructured as well as structured data sets across clusters of inexpensive computers. However, Hadoop is not easy to use, requires a considerable learning curve, and does not always work well for all corporate big data tasks. For example, when Bank of New York Mellon used Hadoop to locate glitches in a trading system, Hadoop worked well on a small scale, but it slowed to a crawl when many employees tried to access it at once. Very few of the company's 13,000 IT specialists had the expertise to troubleshoot this problem. David Gleason, the bank's chief data officer at the time, said he liked Hadoop but felt it still wasn't ready for prime time. According to Gartner, Inc. research director for information management Neil Heudecker, technology originally built to index the web may not be sufficient for corporate big data tasks.

It often takes a lot of work for a company to combine data stored in legacy systems with data stored in Hadoop. Although Hadoop can be much faster than traditional databases for some tasks, it often isn't fast enough to respond to queries immediately or to process incoming data in real time (such as using smartphone location data to generate just-in-time offers).

Hadoop vendors are responding with improvements and enhancements. For example, Hortonworks produced a tool that lets other applications run on top of Hadoop. Other companies are offering tools as Hadoop substitutes. Databricks developed Spark open source software that is more adept than Hadoop at handling real-time data, and the Google spinoff Metanautix is trying to supplant Hadoop entirely.

It is difficult to find enough technical IT specialists with expertise in big data analytical tools, including Hive, Pig, Cassandra, MongoDB, or Hadoop. On top of that, many business managers lack numerical and statistical skills required for finding, manipulating, managing, and interpreting data.

Even with big data expertise, data analysts need some business knowledge of the problem they are trying to solve with big data. For example, if a pharmaceutical company monitoring point-of-sale data in real time sees a spike in aspirin sales in January, it might think that the flu season is intensifying. However, before pouring sales resources into a big campaign and increasing flu medication production, the company would do well to compare sales patterns to past years. People might also be buying aspirin to nurse their hangovers following New Year's Eve parties. In other words, analysts need to know the business and the right questions to ask of the data.

Just because something can be measured doesn't mean it should be measured. Suppose, for instance, that a large company wants to measure its website 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 department pours its energy and resources into monitoring website clicks and social media traffic, which produce many unqualified leads that never lead to sales.

Although big data is very good at detecting correlations, especially subtle correlations that an analysis of smaller data sets might miss, big data analysis doesn't necessarily show causation or which correlations are meaningful. For example, examining big data might show that from 2006 to 2011 the U.S. murder rate was highly correlated with the market share of Internet Explorer, since both declined sharply. But that doesn't necessarily mean there is any meaningful connection between the two phenomena.

Several years ago, Google developed what it thought was a leading-edge algorithm using data it collected from web searches to determine 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 queries on Google. The service has consistently overestimated flu rates when compared to conventional data collected afterward by the U.S. Centers for Disease Control and Prevention (CDC). According to Google Flu Trends, nearly 11 percent of the U.S. population was supposed to have had influenza at the flu season's peak in mid-January 2013. However, an article in the science journal Nature stated that Google's results were nearly twice the actual amount estimated by the CDC, which had 6 percent of the population coming down with the disease. Why did this happen? Several scientists suggested that Google was "tricked" by widespread media coverage of that year's severe flu season in the United States, which was further amplified by social media coverage. Google's algorithm only looked at number of flu search requests, not the context of the searches.

Big data can also provide a distorted picture of the problem. Boston's Street Bump app uses a smartphone's accelerometer to detect potholes without the need for city workers to patrol the streets. Users of this mobile app collect road condition data while they drive and automatically provide city government with real-time information to fix problems and plan long-term investments. However, what Street Bump actually produces is a map of potholes that favors young, affluent areas where more people own smartphones. The capability to record every road bump or pothole from every enabled phone is not the same as recording every pothole. Data often contain systematic biases, and it takes careful thought to spot and correct for those biases.

And let's not forget that big data poses some challenges to information security and privacy. As Chapter 4 pointed out, companies are now aggressively 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 maximum profit.

When you combine someone's personal information 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 information about them exists. Privacy experts worry that people will be tagged and suffer adverse consequences without due process, the ability to fight back, or even knowledge that they have been discriminated against or manipulated in the marketplace.

Sources: Nicole Laskowski and Niel Nikolaisen, "Seven Big Data Problems and How to Avoid Them," TechTarget Inc., 2016; "The Most Innovative Companies of 2016: Top Companies by Sector," www.fastcompany.com, accessed March 4. 2016; Ed Burns, "Big Data Analytics Not Just a Grab and Go Process," *Business Information*, February 2015; Elizabeth Dwoskin, "The Joys and Hype of Software Called Hadoop," *Wall Street Journal*, December 16, 2014; Tim Harford, "Big Data: Are We Making a Big Mistake?" *Financial Times Magazine*, March 28, 2014; Laura Kolodny, "How Consumers Can Use Big Data," *Wall Street Journal*, March 23, 2014; Joseph Stromberg, "Why Google Flu Trends Can't Track the Flu (Yet)," smithsonianmag.com, March 13, 2014; Gary Marcus and Ernest Davis, "Eight (No, Nine!) Problems with Big Data," *New York Times*, April 6, 2014; and Shira Ovide, "Big Data, Big Blunders," *Wall Street Journal*, March 11, 2013.

CASE STUDY QUESTIONS

6-13 What business benefits did the companies and services 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 were not improved by using big data.

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

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