
Working with big data

Simon Sheather, head of the Department of Statistics at Texas A&M University in College Station, Texas, is looking through row after row of airfare data—nearly 8 million of them. But he isn't planning a vacation. He's using the huge dataset to create a model that predicts ticket prices to help customers save money, based on the route they fly.

The increased amount of data in the world has created many opportunities for the kind of analysis Sheather does. Recent advances in technology, such as e-commerce, smart phones, and social networking, are generating new types of data on a scale never seen before—a phenomenon known as "big data." According to some data experts, 90 percent of the data that exists in the world today was created in the last 2 years. And society increasingly relies on data to tell us things about the world.

This year, 2013, is The International Year of Statistics. It's a designation intended to highlight the role that data and statistical analysis have in society. To further that goal, this article describes work with big data. The first section outlines what big data is. The second section provides an overview of big data work. The third section explains some of the challenges that big data work entails. The fourth section describes how to prepare for this work. Sources of information are provided at the end.

What is big data?

Big data generally is defined as a collection of large datasets that cannot be analyzed with normal statistical methods. The datasets are so big, they are measured in exabytes—one quintillion (1 followed by 18 zeroes) bytes. By comparison, an mp3 song is typically less than 10 megabytes (1 followed by 6 zeroes).

The data do not have to be just numbers; they can be videos, pictures, maps, words and phrases, and so on. Examples of big data include customer reviews on commercial websites, comments on social networking websites, photos and videos posted online, electronic medical records, and bank records.

There are two types of big data: structured and unstructured.

Structured data are numbers and words that can be easily categorized and analyzed. These data are generated by things like network sensors embedded in electronic devices, smartphones, and global positioning system (GPS) devices. Structured data also include things like sales figures, account balances, and transaction data.

Unstructured data include more complex information, such as customer reviews from commercial websites, photos and other multimedia, and comments on social networking sites. These data cannot easily be separated into categories or analyzed numerically. "Unstructured big data is the things that humans are saying," says big data consulting firm vice president Tony Jewitt of Plano, Texas. "It uses natural language."

Analysis of unstructured data relies on keywords, which allow users to filter the data based on searchable terms. The explosive growth of the Internet in recent years means that the variety and amount of big data continue to grow. Much of that growth comes from unstructured data.

Sara Royster

Sara Royster is an economist in the Office of Occupational Statistics and Employment Projections, BLS. She can be reached at (202) 691-5645 or at royster.sara@bls.gov.

Working with big data

Some of the work of big data is automated, but not all of it. Workers are still involved in the collection, processing, and analysis of big data. For example, software developers write specialized computer programs that help analyze big data. (For more information about occupations that use big data, see the box below.)

Workers who use big data

Most workers who deal with big data are known as data scientists, although they may also be called data analysts or have some other designation. "The term 'data scientist' is so new, we don't yet have it in our job descriptions at Fermilab," says physicist Robert Roser, head of the Scientific Computing Division at this national laboratory in Batavia, Illinois. The U.S. Bureau of Labor Statistics (BLS) classifies these workers as statisticians, computer programmers, or in other occupations, depending on their tasks.

Whatever their title, these workers study big data using both conventional and newly developed statistical methods. Many of the new methods were developed specifically for use with big data. These workers run computer programs or algorithms, often hundreds of times, to detect patterns or to find usable information. The data are so complex that the workers use software that has been specifically designed to analyze large, unstructured datasets.

But big data work may begin before analysis. Workers may confer with computer specialists to design ways to collect and aggregate the data from sources. After the data are collected, workers devise methods of storing and organizing the huge amounts of data. "Often, traditional means of storage are not enough for big data," says Jewitt.

Workers also process and clean the data, a task known as scrubbing. To make analysis easier, they organize and remove errors or excess information from the data. By consulting with an organization's managers, these workers determine which data should be saved and analyzed, and which are not relevant. "We are looking for that handful of data that is meaningful for what we are doing," says Roser.

After they complete the analysis, workers create graphs, charts, tables, or other tools to

Occupations in big-data work

There are lots of occupations that work with big data in one way or another. The job tasks of these workers are evolving, as are their job titles. Several occupations that might work with big data are described below, along with their relevant job tasks.

Managers who work with big data are known as chief data officers or chief information officers. They create the policy for how their organization will use data, as well as supervise the analysts, computer programmers, and other workers.

Postsecondary teachers who use big data usually instruct students in statistical analysis and computer science. These teachers may have a lot of expertise and experience working with big data but choose to help new generations of workers develop their skills to enter the workforce.

Software developers have an important role in working with big data. They write the computer programs that aggregate, process, analyze, and visualize the data, along with the trends and other useful information that can be found in those data.

Software developers generally are not associated with a single industry but create computer programs for use across industries for lots of different data. They may explore alternative sources of data and alter their programs to work with specific kinds.

summarize the results. They may also present the results to managers and clients.

Big data work also may include developing computer software programs that are used with it. Workers on big data projects consult with computer programmers to write the code that is used to analyze the data. Often, the code is specific to each project and must be written almost entirely from scratch.

Big data by field

Data analysts' job tasks differ, based on the source of the big data with which they work. Sometimes the work that is done in one field is applicable to another, but a lot of the work is specific to that field or organization. Big data collected by one organization can often be of use to another, especially combined with its own proprietary data. For example, traffic pattern data collected by a package delivery service might be useful to urban planners. The exchanging and sharing of data in this manner has created a secondary market for big data.

The following are examples of specific kinds of big data and how workers are involved with them.

Business. Increasingly, businesses base their decisions on data. Businesses need workers to collect relevant product data and analyze that data in the context of the industry. Analysts look at purchase data and customer reviews to decide what kinds of improvements or new products they should make to meet their customers' needs. For example, workers may study transaction data from store loyalty cards to see what types of products customers buy and when they buy them.

Big data can also help businesses run more efficiently. Analysts use supply chain data to manage inventories. They also detect errors by studying real time production data.

E-commerce. Purchase-transaction data from commercial websites have long been collected, but now new kinds of big data are generated by commercial websites. Data analysts help a company improve customer service by studying how consumers feel about its products through customer reviews, comments, and suggestions. Many commercial websites use predictive modeling techniques to suggest similar options when users browse products.

Analysts also search data to find trends in purchasing or website traffic.

Finance. Account data, credit and debit card transactions, and financial market data are examples of financial big data. Analysts study transaction data to look for fraud and other security breaches. They also monitor investment portfolios and alter them to compensate for increased risk and unexpected price changes.

Government. Governments collect a lot of data about their constituents, but policies and security concerns may keep them from sharing or using that data. However, use of big data can help governments serve their constituents better and improve policy decisions. For example, some governments use data to pre-fill tax or other forms for constituents. Analysts also study constituent satisfaction levels by monitoring social networking sites.

Healthcare. The move toward electronic health records generates even more new uses for patient medical data. Patient data can include video feeds from surgeries and other medical procedures. In addition, remote patient monitoring is becoming increasingly popular, and a way to organize and evaluate data from all these video feeds is now necessary.

Analysts use data gathered from drug trials for evidence-based drug therapy and to estimate the cost effectiveness of new drugs. Using social networking, analysts also have created software to track disease outbreaks in real time.

Through the Human Genome Project, scientists have mapped electronically the entirety of the structure of human DNA. Analysts work with scientists to devise uses for the vast amounts of data collected by the project. These uses include the development of drugs tailored specifically to an individual's genetic makeup and the creation of lifesaving medical treatments.

Science. Many different fields of science produce huge datasets. For example, physicists study the properties of particles by colliding them with other particles in high-tech experiments. Data analysts record the location, velocity, and other information about every particle in the experiment. "Particle physics has been dealing with big data since its inception," says Roser. "We just had to wait for the technology to catch up."

Analysts collect data from the experiment data on site, and then ship them off to another lab to be analyzed. "Big data is underlying all

that we do," says Roser. For example, Fermilab hosts all the data for the Large Hadron Collider. Experiments are done in Europe, and then data are transported to Fermilab for analysts to determine how to house and study them.

Other areas of science, such as climatology, chemistry, and biology, also are using these workers to help with the logistics and analysis of large datasets.

Social networking. Data analysts who specialize in social networking study how big data is used after it is generated. Analysts gather huge volumes of comments, pictures, and videos from social networking sites. By sorting these data, the analysts study user preferencesthat can help create more targeted advertising and better customer services. And as social networking continues to grow, analysts search for new ways to use the rich amounts of data that can be found there.

A large portion of the data from social networking websites and online maps and GPS services is personal location data. Even nonhuman objects, such as packages or shipping containers, have location data that are collected and tracked. Analysts use this data to help businesses make better products or advertise more effectively.

Telecommunications. With the proliferation of smart phones, the amount of telecommunications big data has increased rapidly. Smart phones can learn their users' preferences through their actions and can track user location through GPS data. This allows data analysts who work for telecommunications companies to better tailor their services to their customers' preferences, based on their phone use. Analysts also study huge amounts of data from phone records to try and minimize dropped calls and other problems.

Other. Other areas where big data increasingly is used include politics, utilities, and smart meters on appliances.

Politicians rely on polling data and approval ratings, which were traditionally numerical. Now, however, analysts gather public sentiment data from comments on social networking and other websites.

Utility data include power generation and usage information from homes and businesses. Analysts study the data to reduce costs by determining which parts of the system are working at full capacity and where future investments should be made. They can also detect patterns that lead to equipment failure, allowing them to fix outages more quickly.

Smart meters are installed on different kinds of equipment, such as cars and electric meters. The meters transmit data about the equipment's performance. Data analysts examine this data to determine the cause of any malfunctions and help prevent future ones. (For more information about the smart grid, see "Powering the nation: Smart grid careers," elsewhere in this issue of the *Quarterly,* at **www.bls.gov/ooq/2013/fall/art03.pdf**.)

Employment, wages, and outlook

The growth in big data will continue to expand the kinds of work that use this information. As mentioned previously, BLS does not collect data specifically about data

scientists. Instead, BLS classifies these workers as statisticians or computer programmers or in other occupations.

In May 2012, BLS data for wage and salary workers show that there were 25,570 statisticians and 316,790 computer programmers. These occupations had median annual wages of $75,560 and $74,280, respectively—more than double the median annual wage of $34,750 for all workers in May 2012.

In fact, wages in mathematics- and computer-related occupations continue to outpace wages in other occupations. According to BLS Occupational Employment Statistics data, median annual wages in these occupations were $76,270 in May 2012, more than double the median wage for all occupations.

BLS projects both statisticians and computer programmers to have average employment growth between 2010 and 2020. Statistician, a relatively small occupation, is projected to add about 3,500 new jobs over the decade. The larger occupation of computer programmer is projected to add about 43,700 new jobs during the same period.

Workers who use big data are employed by many kinds of institutions and in many different industries: government, businesses, financial institutions, healthcare, scientific research facilities, colleges and universities, and others. The collection and use of big data continues to expand in all of these.

Challenges presented by big data

The growth of big data has provided new insights but also has presented new challenges to those who work with it. A large part of big data work involves not only dealing with these challenges but trying to overcome them as well.

One of the biggest challenges facing those who work with big data is the availability of funding. At a time when the economy has been in recession and governments are trying to cut costs, new investments are often eliminated. Because big data is a recent phenomenon, funds that were meant for big data analysis software, improved computing equipment, or hiring new data analysts often are targeted. "Right now, we have many more questions than we do resources," Roser says of the scientific community.

Another challenge in working with big data is its storage. The volume of data can require many—perhaps hundreds—of servers

to both store and process all the information for even one user or organization. Data need to be saved and made easily accessible to many users for an indefinite length of time. "Nowadays, the common wisdom says to keep everything," says Jewitt. "Make it all electronically available for people to use and study it."

Finding the usable data among the unusable information is also a challenge, given the large volume of data that exists. And after the relevant data have been isolated, useful information needs to be extracted from them. Analyzing not only numerical data but words, pictures, videos, and other information, then combining everything into a meaningful result, takes different and unconventional methods. Isolating usable data also can be very time consuming. Time-sensitive issues, such as fraud detection and epidemic monitoring, require data to be aggregated and analyzed as quickly as possible.

Another challenge is to ensure that big data is accurately measuring what it is meant to measure. With the amount and variety of unstructured data—which can include videos, comments, and other non-numerical data that are not easily analyzed—it is often unclear how the data should be interpreted. If major business and policy decisions are based on the data, analysts need to make sure they are as accurate as possible.

The question of ownership is a challenge unique to big data. Data, unlike a physical asset, are used by many people at once. This raises the question of who owns the data. Do they belong only to the entity that collected them? Is that entity permitted to use the data any way it likes? Because the question of data ownership is new, few laws exist to resolve it.

A related challenge is how to protect and control the data once they have been collected. There is controversy about what data may be collected and how those data can be used or shared without violating people's privacy. Analysts may be responsible for devising methods of keeping big data secure. Big data can contain highly sensitive information, such as location data, financial and medical records, and telecommunications data. Security will be a major concern as data collection continues to grow.

Preparing to work with big data

A major impediment to the widespread use of big data is the lack of workers with the appropriate training and skills. Big data work can require not only knowledge of statistical analysis and computer systems, but experience in the relevant field or industry, such as healthcare or physics.

Education and training

In addition to having a bachelor's degree, most analysts who work with big data have a master's or higher degree. Common specialties include mathematics, statistics, or computer science.

Courses. Coursework in math, statistics, and computer programming prepares students to work with big data. Math helps students develop the logical thinking and problem-solving skills they need. Statistics provides the analytical knowledge that they need to properly study the data and to interpret the results in a meaningful way.

And computer programming courses are a must for those who want to be involved in software development. "A background in computer science is very important," says Sriram Mohan, a computer science professor at Rose-Hulman Institute of Technology in Terre Haute, Indiana. "You need the ability to program and also to think logically."

Workers who use big data come from various fields of study, including engineering. "We look for people with engineering backgrounds, because they think a certain way," says Jewitt. "They know how to break down a problem."

Other training. Workers who use big data may also need education in the industry in which they work, especially in highly technical industries, such as physics and healthcare. The education they need is more specialized than what they can learn on the job—a

degree or work experience usually is necessary. When it comes to hiring data analysts at Fermilab, for example, Roser looks for people with a background in particle physics. "Understanding the basics is most important," he says. For workers who have little formal training, companies may offer classes or instruction to help these workers gain experience with large datasets.

But even workers who have a statistics or data analysis background need to stay current with the fast-changing world of big data. "I spend a lot of time reading books and blogs to try and keep up with new developments," says Mohan. "There are a lot of supportive communities online."

Skills

Problem-solving skills are important for working with big data. Analysts have to create new ways of doing things that account for the different kinds of data and the large scope of the datasets. "We need creative thinkers and problem-solvers," says Jewitt.

Communication skills. Working with big data is highly technical, but workers need to be able to clearly explain their results to other workers. "We need workers who can communicate with others who have nontechnical backgrounds," explains Jewitt. And many times, those other workers may not be as data-savvy as the analysts.

Teamwork. The ability to collaborate and work well with others also is helpful in big data jobs. Work is usually spread among teams of analysts because the data are so complex. Each member of the team has a different responsibility: organizing the data, using software for analysis, or making graphics of the results, for example. It takes the entire team to complete a project. Jewitt uses a sports analogy to explain why coordination and sharing the work are a must. "No one person can do all that analysis," he says. "It takes a baseball team."

Curiosity. But most importantly, data analysts must possess intellectual curiosity. "A very important skill is the ability to learn new things, because the technology is always changing," says Mohan. "Big data is not static."

Roser agrees. "You need a willingness to learn," he says. "If you have that, the rest we can teach."

For more information

To learn more about some of the occupations related to big data, as well as many others, see the *Occupational Outlook Handbook* (*OOH*). The *OOH* is available online at **www.bls.gov/ooh**.

For more information about careers in math-related occupations, see "Math at work: Using numbers on the job" in the fall 2012 issue, available online at **www.bls.gov/ooq/2012/fall/art01.pdf**.

Information about occupations that work with big data is available from professional associations and other sources. Contact information for some of these is listed below:

American Mathematical Society
201 Charles St.
Providence, RI 02904
Toll free: 1 (800) 321-4AMS (321-4267)
(401) 455-4000
www.ams.org

American Statistical Society
732 N. Washington St.
Alexandria, VA 22314
Toll free: 1 (888) 231-3473
(703) 684-1221
www.amstat.org

International Year of Statistics
www.statistics2013.org

National Science Foundation
Big Data Initiative
www.nsf.gov/funding/pgm_summ.jsp?pims_id=504767

Open Government Initiative
www.whitehouse.gov/open


Considering careers? Dare to dream.
www.bls.gov/ooq