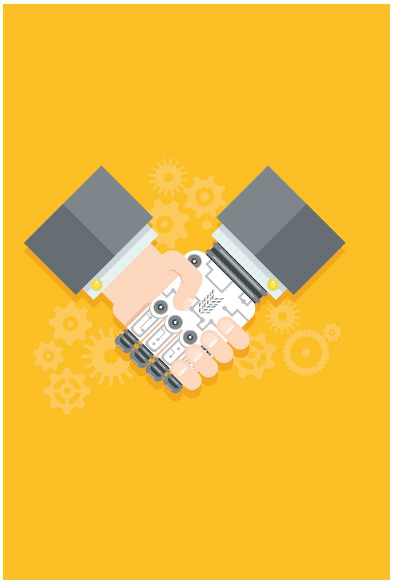
EDITOR: George Strawn, gostrawn@gmail.com



Data Scientist

George Strawn

In the last installment of this column, I dealt with the specter of IT-caused unemployment. Here, I consider a new IT-created employment opportunity—the *data scientist*. First, I look at data, information, and knowledge and current IT job classifications to provide a context for the coming changes. Then, I define big data, which has given impetus to data science. Finally, I define what data science is currently thought to be and what data scientists do.

Data, Information, and Knowledge

In the early days of computing, companies had *data processing* offices. In fact, automated data processing (ADP) began with punch card machines that preceded the electronic digital computer by half a century. In the 1950s, ADP offices began converting to electronic data processing (EDP). *Information management* came next; it reflected the increasing scope and importance of computing to organizations and recognized that whereas computers processed data, they produced information for humans. Later, organizations created offices of *knowledge management*, some directly related to EDP (creating knowledge from data) and some related to HR (capturing the knowledge that humans have in their heads). We now talk about the "data, information, knowledge hierarchy," and it is common to say that a function of computing systems is to turn data into information and knowledge.

Definitions of these three terms vary. The ones that I will use are as follows: Data is the lowest level of the hierarchy, and we can almost see "raw" as an implied adjective. For digital computers, any string of bits can be treated as data. Information, the next level, is to humans what data is to computers. That is, it is the lowest level we normally use. To connect the two levels, computer scientists say that information is data plus *metadata*. Metadata is "data about the data," which helps both computers and humans interpret it in a meaningful way. At the top of the hierarchy is knowledge, which can be defined as *actionable information*. Knowledge or information that enables people or computers to take action is often created by linking together many pieces of information. For example, graph databases, such as those based on the Semantic Web (https://en.m.wikipedia.org/wiki/Semantic_Web), make the links explicit and can be called *knowledge bases*.

After the information word displaced the data word in business parlance—and in light of the expanding importance of computing—a new executive position was created by many organizations, the chief information officer (CIO). At roughly the same time, computing began to be known as information technology, which is a more inclusive term and refers to networked computers as well as stand-alone ones. About this time, a second C-suite officer, the chief technology officer, joined the expanding list. The CTO is supposed to keep an eye on new technology developments and recommend which ones to pursue, whereas the CIO provides information services.

Since the turn of the century, the chief data officer has begun to take

1520-9202/16/$33.00 © 2016 IEEE

a place in the C-suite. One definition of the CDO is "a corporate officer responsible for enterprise-wide governance and utilization of information [sic] as an asset, via data processing, analysis, data mining, information trading, and other means" (https://en.m.wikipedia.org/wiki/Chief_data_officer). Why has this responsibility risen to prominence now?

Big Data

Today's focus on data is partly a result of yesterday's focus on the Internet. That is, the successful effort to interconnect millions of computers has created a cornucopia of interconnected datasets. Also, the exponential increase in disk storage capacity and the similar decrease in cost have resulted in interconnected computers having a huge amount of information to share. "Enterprise-wide governance and utilization of data" doesn't mean only data that the enterprise owns. Google and Yahoo are obvious examples of enterprises whose business plans involve "organizing the world's information and making it universally accessible and useful" (Google's stated mission).

"The world's information" is certainly big and is about to get much bigger as the Internet of Things arrives (https://en.m.wikipedia.org/wiki/Internet_of_Things). More generally, big data can be defined as data that has one or more of the three V attributes: volume, velocity, or variety. That is, big data is either too big, comes at you too fast, or has too much variability to be processed in a reasonable time by today's computer systems. Thus, an active computing research topic is how to process bigger, faster, and more variable data efficiently. Let's look at some of the methods currently being used to process big data.

A prominent example of big volume data is Google's crawling and organizing of the huge amount of information on the Web. The company developed a method called MapReduce to parallel-process on huge server farms the vast treasure of Web information. Subsequently, they published a paper describing how MapReduce worked, and an open source version called Hadoop was made available by Apache, which is now widely used. In addition, as described in last issue's Mastermind article,[1] new database models are supplanting the venerable relational database model, because it doesn't scale to big data.

Examples of successful big velocity data processing are applications such as Google searching and Amazon ordering. A scientific example is the Large Hadron Collider (LHC) in Europe (where the Higgs boson was recently discovered). Even before it was upgraded to a higher power, the LHC produced a petabyte of data per second! Because it is a smart instrument, the LHC had to save "only" a petabyte per month for further processing. Such preprocessing—either by smart instruments or by computers—is an important technique to deal with big velocity data.

Big variety data poses the largest challenge. Without extensive metadata to describe the data, meaningful processing is hopeless. Without interoperable metadata, the situation is not much better. Consider the so-called "long tail" of millions of small scientific datasets. If these datasets could be meaningfully processed together, new scientific results would emerge. Another type of variety in data is *unstructured data*, such as text. If metadata in the form of a thesaurus could solve the synonym problem, and if natural language processing could deduce the forms of the sentences, then computing systems such as the Semantic Web could turn text into knowledge bases. A prototype of such systems—Semantic Medline, a knowledge base created from 22 million biomedical research articles—has a been developed at the US National Library of Medicine (http://youtu.be/GWdyk18RTuA).

The foregoing illustrates some of the research and development that is currently under way to enable more efficient processing of big volume and velocity data. With an eye toward more effective processing, such as that achieved with Semantic Medline, we turn to the subject of data science.

Data Science

My perspective is from that of a computer scientist. However, statisticians also have a claim on this territory. Years ago, a statistician friend told me that a simple definition of statistics was "what to do with the data." If I interpret history correctly, 19th century statistics worked to make sense of all the data in relatively small datasets, and 20th century statistics developed sampling techniques to be able to work with small parts of larger datasets and still infer characteristics of the whole dataset. With today's advanced computing systems, statisticians are also investigating big data processing techniques. That is, they are again working with all the data as they did in the 19th century—but this time, with big data, not small data. With both disciplines in mind, we can next ask: what is a data scientist?

A tongue-in-cheek definition is, a computer scientist who knows more statistics than his or her colleagues, or a statistician who knows more computer science than his or her colleagues. Time will tell if data science will become a new discipline, or if it will remain a cross-disciplinary field between these two (and perhaps other) fields.

The statistician David Donoho published a paper in 2015 with the provocative title "Fifty Years of Data Science" (https://dl.dropboxusercontent.com/u/23421017/50YearsDataScience.pdf). He was referencing the statistician John Tukey's call, more than 50 years ago, for statistics to expand into what we now call data science. Donoho's paper is well worth reading. I'll list the six subfields of data science that he identifies and make several comments about each one:

1. data exploration and preparation;
2. data representation and transformation;
3. computing with data;
4. data modeling;
5. data visualization and preparation; and
6. science about data science.

Subfields 1–3 and 5 have primarily been the province of computer science. Subfield 3 includes the big data processing techniques just discussed. Subfield 4 is the shared subfield, in which computer scientists talk about machine learning and data mining, and statisticians talk about inferences from data. (Tukey complained that the academic statistics of his time dealt only with subfield 4.) Subfield 6 is perhaps the most interesting opportunity. As Donoho points out, data is an empirical entity and, as such, could give rise to a general science about data, not just data as related to a specific discipline.

Data modeling in the sense of machine learning has emerged from the field of artificial intelligence. The original idea of computer programming involved telling the computer what to do. Machine learning expands on this idea by developing programs that learn what to do from the data. As the computer scientist Pedro Domingos states in his book, *The Master Algorithm*,[2] the output of a machine-learning program is another program that can perform a (learned) task. An example of a machine-learning program is one that can be given a number of pictures containing faces and can then pick out other pictures that also contain faces. Domingos identifies five "tribes" of machine learning: symbolists (who use inverse deduction); connectionists (who use backpropagation, neural nets, and deep learning); evolutionaries (who do genetic programming); Bayesians (who deal with uncertainty); and analogizers (who use support vector machines). It's beyond the scope of this article to describe these tribes in detail. If you are interested in learning more about machine learning, you would do well to read Domingos's book.

Educating Data Scientists

A growing demand for people trained in data science has caused the shortage of these people to balloon. Moreover, only limited opportunities to obtain such training exist today. An article in *Bloomberg Business* speaks to this shortage and universities' initial efforts to address it:

> *A new species of techie is in demand these days—not only in Silicon Valley, but also in company headquarters around the world. "Data scientists are the new superheroes." [...] A study by McKinsey projects that "by 2018, the US alone may face a 50 percent to 60 percent gap between supply and requisite demand of deep analytic talent." [...] Accenture's Mulani says he's tallied some 30 new data science programs in North America, either up and running or in the works* (www.bloomberg.com/news/articles/2015-06-04/help-wanted-black-belts-in-data).

Donoho also noted that "a recent and growing phenomenon is the emergence of data science programs at major universities, including UC Berkeley, NYU, MIT, and, most recently, the University of Michigan, which on September 8, 2015 announced a $100 million 'Data Science Initiative' that will hire 35 new faculty" (https://dl.dropboxusercontent.com/u/23421017/50YearsDataScience.pdf).

The next few years will be exciting ones for the emerging field of data science. With application pull and university push, data science is poised to grow rapidly, and people who position themselves to participate in this potentially revolutionary field should have multiple employment opportunities.

References

1. G. Strawn and C. Strawn, "Relational Databases: Codd, Stonebraker, and Ellison," *IT Professional*, vol. 18, no. 2, 2016, pp. 63–65.
2. P. Domingos, *The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World*, Basic Books, 2015.

***George Strawn** is the former director of the National Coordination Office for the Networking and Information Technology Research and Development Program (NITRD). He is now retired. Contact him at gostrawn@gmail.com.*