Expectations for Data Engineers

加納学

Manabu Kano
Department of Systems Science
Graduate School of Informatics
Kyoto University

You do not need to care about whether your data is big or not. Creating value from data matters, and in this respect small data is very attractive. If you have a diamond in the rough, you should keep it and polish it. You do not have to go to a vast desert for gold dust.

An attempt to extract valuable information from a massive amount of data is called data mining, and the term "big data" has long been used in this field. Miners who unearth gold nuggets from a pile of huge data are called data scientists. "Data, data everywhere," an article in *The Economist* (February 25, 2010), probably made many businessmen turn their attention to big data. It defines data scientists as follows:

> *"... a new kind of professional has emerged, the data scientist, who combines the skills of software programmer, statistician and storyteller/artist to extract the nuggets of gold hidden under mountains of data."*

In the midst of the California Gold Rush, those who accumulated wealth were not miners. Samuel Brannan sold shovels, pickaxes, pans, and other essential items for mining, Levi Strauss made work pants with heavy-duty canvas, and Henry Wells and William Fargo provided services to help miners turn their gold into money, deposit and remit the money, and send and receive letters and packages. People want to unearth precious information from big data and expect data scientists as miners to accomplish this task. Who will succeed in accumulating wealth?

From what I have heard, those who had not shown interest in data analysis suddenly ask to know how to make profits with big data. They must have found some pioneers' wonderful achievements. There are many companies that use big data analysis for marketing. They are not limited to web-based enterprises and social networking companies. Some organizations use analyses to prevent crime. The manufacturing industry is no exception. Concepts such as Industrie 4.0 and Internet of Things (IoT) are advocated to improve and expand manufacturing. A key to success is how to handle the vast quantity of data that are exchanged among things via networks. For example, General Electric (GE) offers predictive maintenance of turbines and power generators all over the world, by analyzing data obtained from sensors attached to these instruments. Losses estimated at 70 million dollars are averted by this service every year.

Gathering and analyzing data from many instruments could make it possible to predict the appropriate timing for maintenance. There is a similar process in medicine. Gathering and analyzing data from many subjects (humans) may prevent diseases. To achieve advanced health care, data on physical checkups and medical statements held by medical insurers are being gathered and analyzed. Furthermore, genetic information as well as heartbeats, daily activities, and other data from wearable devices are also the target of analysis. The dawn of medical big data has come.

In the process industry, can we obtain results similar to those expected in health care? Or can we perform predictive maintenance similar to the services provided by GE? To answer these questions, we must understand the difference in the premise. The number of people who undergo medical examinations is in the thousands, or tens of thousands. For predictive maintenance, GE monitors more than 1,500 instruments. Since data are collected from many humans and instruments in various conditions, analyzing the data makes it possible to build models that detect signs of diseases and failures. Meanwhile, each plant is unique in many cases, and it is difficult to obtain data during various kinds of abnormal conditions and build models that can describe such conditions. This is why fault detection methods that define normal conditions, such as multivariate statistical process control (MSPC), have been utilized. More data does not necessarily help. Even if the data volume increases exponentially, knowledge obtained from the data does not increase if meaningful information is not included. What is needed to compensate for the insufficiency of such data is technical knowledge (domain knowledge) about target plants. The integral use of domain knowledge and data analysis is the key to success.

Skillful operators have plenty of knowledge about target plants. However, most of their knowledge is tacit knowledge, often called “know-how”. Therefore, as many veteran workers in the manufacturing industry reach retirement age, which is called the 2007 or 2012 workforce crunch in Japan, how to hand down such knowledge to future generations has become a serious problem. To convert tacit knowledge into explicit knowledge, Exapilot and other operation support systems are used, and they yield remarkable results in automatic start-up and other operations. Before drawing a flow chart, it is necessary to clarify when and what to do and what conditions need to be fulfilled to proceed to the next procedure. Therefore, operation support systems wield their power in the automation of routine work. How should we approach work that does not follow the IF-THEN rule but depends on operators’ expertise? There is still room to use data. You can consider skillful operators as a function that transforms all information about plants into appropriate operations. By identifying the relation between inputs and outputs based on data, you can make a copy of skillful operators. By recording operators’ interventions, the results, and the situations in which they decided to intervene manually, you can make an appropriate intervention when a similar situation occurs. Previously, the operation support system had to be built manually, but it will be automatically built from data.

You can probably build such a system to automatically create an operation support system, but can you leave all tasks to such an operation support system that was generated via a black box from data in a situation that may result in accidents? This is completely different from simple tasks such as displaying ads on a smart phone. Furthermore, how to respond to unexperienced situations remains a serious challenge. Thus, there is a need for domain knowledge about the target plant. Including physical and chemical laws, huge knowledge has been accumulated and systematized. By integrating a huge amount of data and knowledge, futuristic plant operation will be achieved. You can call this integration technology “artificial intelligence (AI).” The crucial point is not how to use artificial intelligence and big data, but what you have to do and what technology is required for this purpose.

In fact, just like plant operation, data analysis depends on tacit knowledge. In the article in *The Economist* introduced above, the word “artist” is used to describe data scientists. I agree that there are many artistic aspects to data analysis. That is, how to analyze data is determined by the experience and inspiration of the analyst, and the results depend on who the analyzer was. This dependency on individuals is a problem in data analysis. Why does data analysis tend to be dependent on individual skills? One reason is that no one knows in advance what results can be obtained when selecting variables, samples, pretreatment methods, and modeling methods. Hence, some organizations standardize data analysis procedures based on their experience. If a huge volume of experience is accumulated, data analysis, in particular the automation of data preprocessing that takes time and effort, will progress. In addition, if objectives are limited to fault detection and virtual sensing, the automation of data analysis may be able to be achieved easily. If automatic data analysis systems appropriately perform the necessary processing even without skilled data analysts, productivity will improve greatly.

In the midst of the third AI boom, the list of occupations that may be taken by AI is attracting attention. Occupations needed in line with the development of technology have changed and will change from time to time, just as narrators for silent movies lost their jobs when movies came with sound. It is you who improve your skills. Although data analysis technology is progressing rapidly, people with advanced skills are required for the time being. For this reason, personnel training for data analysis is an urgent issue. The talented people whom I assume here are not data scientists. Skills that can program with statistical knowledge and can manage artistic analysis are not enough. It is data engineers who have domain knowledge about the target plant, integrate data and knowledge, and create paths by themselves.

For these data engineers to grow and for data analysis technology to take root in organizations, it is important to succeed with a small problem. The object may be small data. Big success will follow it. I hope that many small-data engineers will grow.

* Exapilot is a registered trademark of Yokogawa Electric Corporation.