NEW UNDERGRADUATE DEPARTMENTS AND PROGRAMS OF DATA SCIENCE IN JAPAN

Akimichi Takemura
Faculty of Data Science, Shiga University
1-1-1 Banba, Hikone, Shiga 522-8522, Japan
a-takemura@biwako.shiga-u.ac.jp

In April of 2017, Shiga University established the faculty of data science, which is the first in Japan. This faculty considers statistics as an essential part of data science and the faculty is also the first faculty in Japan, where students can systematically learn statistics. In April of 2018, Yokohama City University established the school of data science, following Shiga University. These new data science departments mark an important turning point in statistics education in Japanese universities.

CHANGES BROUGHT ABOUT BY BIG DATA

From around 2010, the media started to use the word "big data" to depict the fact that large and varied data became available everywhere in the society. Often big data is characterized by 3V, which are "Variety", "Volume" and "Velocity". Variety refers to the fact that data are now obtained in many forms, i.e., not only in the form of numerical data, but also as text messages, digital images, sound data, etc. Volume refers to the large amount of the data. Major web sites collects terabytes of data every day. Velocity emphasizes the fact that the stream of data needs to be handled in real time. For example, in the web marketing, when a user accesses a commercial web page, the server has to decide which advertisement to put on the page in a fraction of a second. Big data are typically collected by large companies on Internet, such as Google or Amazon, and used for business purposes. By the progress of ICT (Information and Communication Technology) real-time behavior of consumers is being recorded and this stream of data presents opportunities for new services and products.

There is also big data in science. In scientific research, large volume of data became available. The book by Hey et al. "The Fourth Paradigm; Data-Intensive Scientific Discovery" gives a compelling argument that many scientific fields are now driven by massive data and this kind of research changes the paradigm of scientific research. According to Hey et al. the first paradigm dates back to Greek science, where philosophical interpretation of natural phenomena was given based on various observations. The second paradigm is the traditional science since the Renaissance, where theoretical models are tested by scientific experiments. The third paradigm is the simulation based research which became possible by the rapid progress of computers. In simulation based research, properties of large theoretical models are evaluated by computer experiments and used for prediction. The third paradigm is a deductive approach. Finally in the fourth paradigm, the data comes first in scientific research and theoretical models are constructed based on the data. Sometimes the purpose of the model is for prediction of future data, rather than explanation of the natural phenomenon producing the data. This paradigm is an inductive approach driven by data.

In education, the way students learn is also changing. Traditionally the teaching has been based on text books and teachers taught students based on text books. In books, the knowledge was sequentially organized and students acquired knowledge by reading the books sequentially. Knowledge can be considered as an accumulation of useful facts with systematic and theoretical explanation of relation among various facts. Memorization of facts was helped by systematic explanations. However, in the area of big data and Internet, the role of memorization is changing. With the smartphones which are always connected to Internet, memorization of facts are not as important as before. More important is the skill for retrieving various facts from Internet and the ability for organizing the obtained facts as needed. Obviously the systematic and theoretical explanation of the world is of basic importance even in the era of big data and reading books is still the best way to learn how to understand the world systematically. But memorization itself is not as important as before.

In M. A. Sorto, A. White, & L. Guyot (Eds.), *Looking back, looking forward. Proceedings of the Tenth International Conference on Teaching Statistics (ICOTS10, July, 2018), Kyoto, Japan.* Voorburg, The Netherlands: International Statistical Institute. iase-web.org [© 2018 ISI/IASE]

Thus the availability of big data is changing many fields of the society, such as business, science and education.

DATA SCIENCE AND DATA SCIENTIST FOR EXTRACTING VALUE FROM DATA

The availability of big data is a rather recent phenomenon and those who can collect and utilize the data has a fast competitive advantage. In fact, the Web (World Wide Web) itself started in 1993 and hence it is only about 25 years old. The large companies on Internet, such as Google or Amazon, are based on Web and they achieved their successes in only about 20 years. This is a very fast historical change.

The Economist magazine in its May 6th 2017 issue declared that "the world's most valuable resource is no longer oil, but data." The article explains that "the five Internet giants, Alphabet (Google's parent company), Amazon, Apple, Facebook and Microsoft, are the five most valuable listed firms in the world. Their profits are surging: they collectively racked up over $25bn in net profit in the first quarter of 2017. Amazon captures half of all dollars spent online in America. Google and Facebook accounted for almost all the revenue growth in digital advertising in America last year." The article also discusses the need for regulating internet giants. There is a strong "network effect" for digital services: "By collecting more data, a firm has more scope to improve its products, which attracts more users, generating even more data, and so on." Hence it is not easy to regulate internet giants. The media is now talking about "Amazon effect", where traditional retail business cannot compete with Amazon. Amazon does not operate storefronts and can offer lower prices than traditional retail business with storefronts. In U.S. many shopping malls are disappearing.

At this point we should mention that data does not have to be "big." Before the wide spread use of personal computers, office works and economic transactions were handled in analogue form, such as paper and telephone. It was hard to keep records of these activities and analyze them. Nowadays, even small shops and companies record their activities digitally and keep their data in spreadsheets and word processor documents. Also the storage cost is diminishing due to falling prices of hard disks. In 2018, a hard disk with one terabyte capacity, which can hold a trillion characters, cost much less than 100 U.S. dollars. Hence even small organizations can keep their digital records easily. These small or medium sized data are also useful economic resources. We can say that data, big or small, is now ubiquitous in society and awaiting to be exploited. At the beginning of this article we emphasized big data, because it is a typical recent phenomenon and it had a large impact on the recent changes.

From the above facts, the importance of data as an economic resource is clear. We can think of data science as the field for utilizing data and extracting value from data. With this definition, we can define data scientists as those who can handle data and can extract useful information from data. Data scientists needs computational skills, such as distributed data bases, to handle data and statistical skills for analysis of data. In addition, data scientists have to be able to work and communicate with people in the domain of application of data science.

DATA SCIENTISTS IN DEMAND

Around 2008, Hal Varian told the press as follows:

"I keep saying the sexy job in the next ten years will be statisticians."

His prediction turned out to be correct. An article by Pierson in October 2017 issue of AMSTAT news shows a very strong growth of the number of conferred statistics and biostatistics degrees. In 2016 the number of Master's degrees and Bachelor's degrees in statistics and biostatistics are about 4000 and 3000, respectively, which are about five times of those in 2008. This is the result of a strong demand for statisticians and data scientists in U.S. In fact, careers site Glassdoor shows that data scientist is the No.1 in 50 best jobs in U.S., with 4.8/5 Job Score, 4.2/5 Job Satisfaction and $110,000 Median Base Salary. This trend has been continuing for the last several years, putting data scientists, statisticians or biostatisticians as one of the best jobs in similar business surveys.

The same phenomenon is happening in China. China has their own Internet giants, such as Tencent and Alibaba and they are hiring many statisticians and data scientists. There are now more than 300 statistics departments in universities in China, according to Prof. Wei Yuan of Renmin University of China, who gave a talk on January 19, 2017 at Shiga University.

As I mention below, Japan is very much behind in statistics and data science area. However recently there is a clear tendency of shortage of data scientists in Japan. Many Japanese companies are now trying to hire data scientists, but they can not find any suitable candidate.

DATA SCIENCE AND STATISTICS

Data science is a new word and many statisticians argue that statistics has been data science from long time ago. Donoho (2017) presents a compelling argument that statistics has been data science at least from Tukey (1962) and his emphasis on exploratory data analysis. I totally agree with Donoho and I strongly believe that statistics is an essential part of data science as I argue below. But the word "data science" is currently used to reflect the changes brought by the 3V (Variety, Volume and Velocity) of big data and traditional statistical methods may not be sufficient for analyzing new kinds of data, such as texts or images. Also computer science is another essential part of data science as the technology to handle large data or real-time data.

One of the main reasons why statistics is an essential part of data science is that fundamental statistical ideas, such as population or bias, is needed in extracting correct information from data. For example, the data obtained from smartphone usage is a typical big data, but this data may be biased for the purpose of extracting relevant information from data. In Japan there is a clear tendency from various surveys that older people tend to use smartphones less than younger people. Hence if we analyze data from smartphones, the analysis may not apply to older generations of Japan. Another example is text messages on social networking services. Some people post lots of messages to SNS, but some people avoid to put messages to those services. Some people get tired of SNS services and might quit posting messages. This means that the population of active SNS users might not be well defined or stable.

Another important point to make is that big data is almost always observational data. In observational data, we can observe various correlations but it is hard to derive some causal interpretations from observational data. In the media, correlations and causal relations are often unintentionally or intentionally confused. To assert well-evidenced causal relations, ideally we perform randomized controlled trials. However, usually data from randomized controlled trials is not a "big data." In order to obtain valid causal interpretations from observational data, data scientists need to have good understanding of e.g. partial correlations, confounding factors, latent variables etc. Knowledge of newer statistical techniques for causal inference is also important at looking at big data. We should mention that even in the framework of clinical trials, where randomized controlled trials is essential, big medical data with lots of information is also valuable. For example a new drug might have some infrequent but serious side effects, which might not be observed during clinical trials. Big data analysis is needed to find such cases.

JAPAN IS BEHIND IN STATISTIS AND DATA SCIENCE

We discussed that data scientists are in large demand in U.S. and China. Similar phenomen is also happening in Japan, but the Japanese market of data scientists is very small. One of the reasons for the shortage of data scientists in Japan is that there was no statistics department in Japan before the establishment of data science faculty in Shiga University.

To be precise, I should mention that there is Department of Statistical Science, School of Multidisciplinary Sciences, The Graduate University of Advanced Studies, which which was established in October of 1988 and is closely connected to the Institute of Statistical Mathematics in Tokyo. This department is mainly concerned with Ph.D. program and confers about 5 Ph.D. degrees in statistical science in each year. It has been the only department in Japan giving Ph.D. degrees in statistical science. Also it should be mentioned that for a brief period before 1970, there was a department of statistics in an engineering school of Nihon University. This department was led by Junjiro Ogawa. Details of this department are not well known but the political turmoil around 1970 might have affected this department.

It is natural to ask why there was no statistics department in Japan. In Japanese universities statisticians are scattered into various faculties, such as economics, mathematics, engineering, or education. This is in a sharp contrast to United States, Korea and China, where there are independent statistics departments.

There may be many reasons for this. One reason is that Japanese statisticians tended to emphasize applying statistics to other fields than studying theoretical statistics itself. In fact, there were some very original contributions from Japanese statistics, such as statistical quality control for manufacturing industries and the Akaike Information Criterion for model selection. These works were very much motivated by application of statistics to practical problems. Another reason may be that the voices of Japanese statisticians were not necessarily united. The community of Japanese statisticians is divided into many academic societies. As an umbrella organization of six academic societies on statistical science, we have the Japanese Federation of Statistical Science Associations (JFSSA) only since 2005.

However probably the biggest reason was that there was no "statistics industry" for students to get jobs and statistics teachers were not sure whether graduates from statistics departments could have good employment opportunities in Japan. The faculties and departments of Japanese universities are organized to reflect the industrial organization. In contrast to this "vertical" segmentation of faculties and departments according to those of industries, statistics is a common methodology useful for many fields. We can call statistics a "horizontal field." Horizontal means "trans-disciplinary" or "transversal". Also computer science has a horizontal characteristic, because information technology is useful for many fields. One difference between statistics and computer science is that there is an industry of manufacturing computers and information machinery. Japan used to be strong in this manufacturing area.

In Japan, horizontal fields and techniques were not considered of primary importance. Students were supposed to first learn specific fields, such as economics and mechanical engineering, and they learn statistics only when it becomes necessary in research or product developments. Many good applied statisticians in Japan actually taught themselves statistics, because formal and systematic education of statistics was not available.

Recently in Japan there is a genuine demand for data scientists and the Japanese government itself started to emphasize need for data scientists. Japan revitalization strategy 2016 says "in the big data era, technologies for new business and services are based on utilization of data. They include AI, big data, IoT, etc." The 5th Basic Program For Science and Technology says "Japan is in a very risky position compared to other countries, because of severe lack of people knowledgeable in data analysis and statistical science."

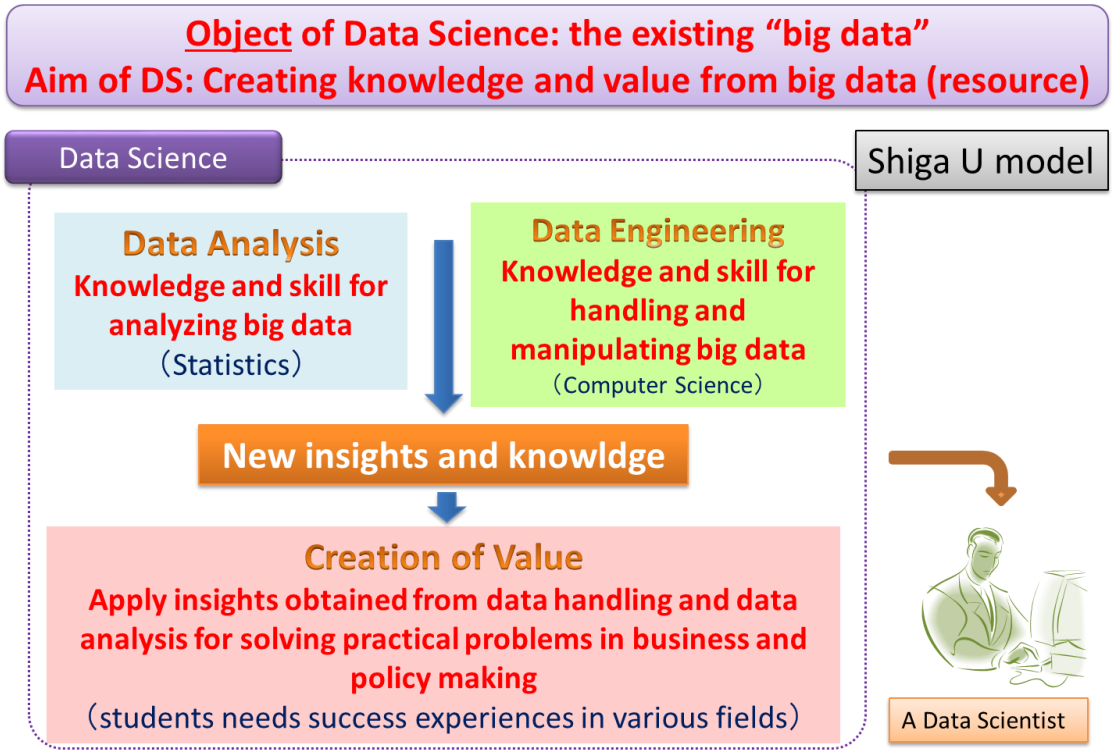
NEW UNDERGRADUATE FACULTIES IN SHIGA UNIVERSITY AND YOKOHAMA CITY UNIVERSITY

Shiga University is one of 86 national universities in Japan. It is a rather small university as a national university with about 800 students in each year. Until the establishment of data science faculty it consisted of only two faculties: economics and education. For a long time, Shiga university tried to establish the third faculty, which is more science and technology oriented than economics and education. In 2014, Prof. Takamitsu Sawa, then the president of Shiga University and known for his econometric research, proposed to form the faculty of data science and quicly got the approval of the executive office and two faculties of the university. The ministry of education was supportive of the new faculty, because they were already aware of the severe shortage of data scientists in Japan. The data science faculty of Shiga university accepts 100 students in each year. It also tries to establish a master course as soon as possible. We probably open our master course in April of 2019.

Following Shiga University, Yokohama City University (YCU) established its own School of Data Science with 60 students in each year. The design of their curriculum is similar to that of Shiga University. Some difference between Shiga and YCU is that YCU emphasizes more medical application of data science, because YCU operates a large hospital. Shiga and YCU maintain good relations, because our purposes are similar. YCU data science is currently led by Prof. Manabu Iwasaki, who was the 32nd president of Japan Statistical Society, where as I was the 30th president of Japan Statistical Society.

Details of curriculum of Data Science Faculty of Shiga University is reported in the paper by Takata et al. for this conference. The following Figure 1 depicts our idea of data scientists graduating from our faculty.

Figure 1. Shiga University Model for Education of Data Scientists



CONCLUSION

I have discussed the background of new undergraduate departments and programs of data science in Japan. The most important factor for these developments is that the big data era has arrived and Japanese government and business became very much aware of the power of data science. Then they found that there is a severe shortage of data scientists in Japan. In my opinion, this shortage and delay was mainly caused by the absence of statistics faculties and departments of statistics in Japanese universities. As is well known, Japan faces a severe problem of aging population and shortage of young people. This puts much pressure on management of Japanese universities. At this moment it is very difficult for major Japanese universities, such as Tokyo University or Kyoto University to form a new department.

Shiga University was lucky to establish the first data science faculty of Japan. Its establishment attracted lots of attention from the media and it started with a good success. Then Yokohama City University followed with its own School of Data Science. They are also very successful. We hope that other universities in Japan will follow Shiga and YCU and systematic education of data science will be widely available.

REFERENCES

Economist, T. (2017). The world's most valuable resource is no longer oil, but data. *The Economist: New York, NY, USA*.

Donoho, D. (2017). 50 years of data science. *Journal of Computational and Graphical Statistics*, *26*(4), 745-766.

Glassdoor (2018). 50 Best Jobs in America. https://www.glassdoor.com/List/Best-Jobs-in-America-LST_KQ0,20.htm. Retrieved on February 17, 2018.

Pierson, S. (2017). Bachelor's, master's statistics and biostatistics degree growth strong through 2016. *AMSTAT news: the membership magazine of the American Statistical Association*, (484), 14-16.

Hey, T., Tansley, S., & Tolle, K. M. (2009). *The fourth paradigm: data-intensive scientific discovery* (Vol. 1). Redmond, WA: Microsoft research.

Takata, S., Izumi, S. and Takemura, A. (2018). Education of Data Science in Japan - Shiga University model -. A talk in this conference (ICOTS10, Kyoto, Japan), July 2018.

Tukey, J. W. (1962). The future of data analysis. *The annals of mathematical statistics*, *33*(1), 1-67.

Varian, H. (2008). Statistics - Dream Job of the next decade. Keynote Presentation at the 2008 Almaden Institute.

Yuan, W. (2017). Big Data Analytics Education in China. A talk at Shiga University. January 19, 2017.