EDUCATION OF DATA SCIENCE IN JAPAN -SHIGA UNIVERSITY MODEL-

Seiji Takata, Shizue Izumi and Akimichi Takemura
Faculty of Data Science, Shiga University, Japan
seiji-takata@biwako.shiga-u.ac.jp

In 2017, Shiga University established the Faculty of Data Science, the first university faculty specialized to data science in Japan. The goal of our faculty is "extracting value from data." We intend to teach mathematical statistics, computer skills and practical business experience in a balanced manner, so that students can solve real-world problems using various data. Various practices of data science are treated, including business applications, information technology, medical science, disaster prevention, public policy, etc. In particular, we place special importance on project-based learnings (PBL), in which students actively tackle real-world problems using real data.We extensively collaborate with various private sectors, so that students can experience real-world problems. We also collaborate with central/local governments to contribute to better decision-makings in public policies.

INTRODUCTION

Japan has long traditions in statistical theory and applied statistics (e.g., quality control at industry.) However, such traditions were, in a sense, isolated. For example, there was no faculty of statistics at universities. Statisticians belonged to faculties of science, economics, medicine, etc., and they had little relationship in each other. Statistical quality control was limited to manufacturing plants, and statistical methods were rarely applied in other business sectors.

Such situation has changed. Many Japanese recognize the power of statistical methods, and the needs for statisticians (or, data scientists) are quite large in many business sectors. And it was not until 2017 that Japan has the first faculty of statistics (or, faculty of data science) at Shiga University.

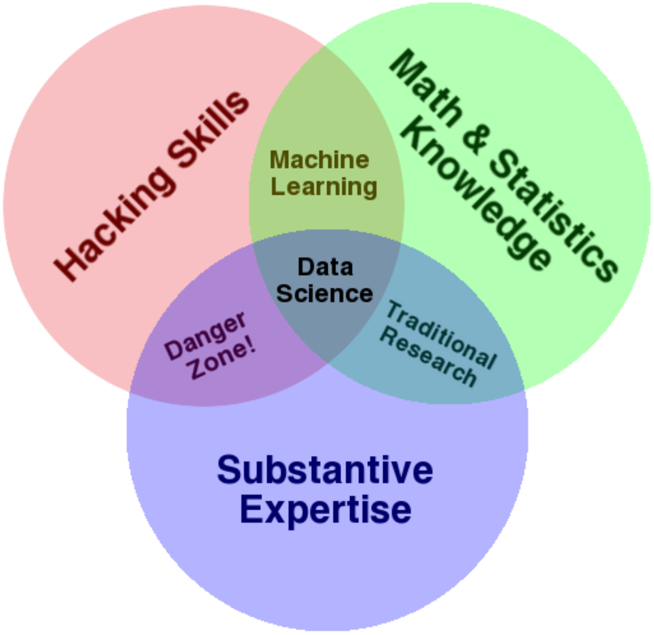
In this paper, we explain the aim, methods and the curriculum of data science education at Shiga University. We hope that our experience would be useful for universities, public and private sectors.

OUR GOAL AT SHIGA UNIVERSITY

We believe that data science is a study of "extracting value from data". Data science has to be pragmatic. Statistical theory is a basis for data science, but pure theory is not enough for data science. Therefore, our education for data science at Shiga University is different from that of traditional education for statistics.

Figure 1 is a famous diagram for data scientists. For data science, "math & statistics knowledge", "hacking skills" and "substantive expertise" are equally important. Math & statistics knowledge is the basis for data science. For example, if we want to predict future events, uncertainty is unavoidable. Probability theory is an essential tool for analyzing uncertainty. Statistical inference theory is necessary for decision-making. In the real world, data contain some errors. So, if we want to extract meaningful results from data, statistical inference is necessary. Unfortunately, math & statistics knowledge is not enough for data science. Today, the volume of data we treat is huge, that is, "big data". Everyday new data come in via internet, mobile phones, sensors, etc. Excel may be good for small data analysis. But, in order to treat big data, we have to learn computer programming skills, such as R, Python, etc. Substantive expertise, or, business skill, is also necessary for data science. When we apply statistical methods to real data (for research, business problem, etc.), we have to understand the nature of the data. How these data are collected, accuracy of the data, etc. We also have to understand the aim of the analysis. Is the conclusion meaningful for business purpose? Is it feasible in the real world? Data science must have such substantive expertise, in order to be useful in the real world.

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



©Drew Conway Data Consulting

Figure 1. Skills for Data Science

In addition to them, we also stress the importance of communication skill. Data scientists have to communicate with various persons, e.g., business partners, clients, data providers, etc. And a data scientist cannot be almighty. One may have to consult various experts. In order to fully utilize the power of data, effective communications are crucial. Presentation / data visualization skills may be included in this category. In order data science to be useful in the real world, we have to communicate our results effectively and impressively. Presentation skill and data visualization technique are necessary for data scientists.

Therefore, the pillars of our data science education are
-Mathematics and statistics theory,
-Computer skills,
-Practical experience at various fields,
-Communication and presentation skills.

Data science is applied to a wide variety of fields, public policy is one of them. In public policy, the notion of evidence-based policy making (EBPM) is increasing its importance. Data sources for EBPM are, not only traditional official statistics, but also various big data, such as administrative data, social-network data, etc. To deal with big data, advanced computer skill is necessary. Also, EBPM needs, beyond the traditional tabulated statistics, statistical modelling and inference. Practical experience in policy making is, of course, important for EBPM. That is, skills for data science are also necessary for EBPM.

OUR METHODS -REAL DATA AND PROJECT-BASED LEARNING-

We believe that using real data is essential for data science education. Beautiful artificial data may be good for studying pure theory. However, the real world is filled with the slaying of beautiful theories by tiny ugly facts. Or, seemingly ugly data may lead to a new theory. Therefore, using real data is important in the education of data science.

In addition to that, we will extensively employ the project-based learning (PBL) method. In the world of data science, the answer is not always unique. Various techniques may be applied to one problem. It is the art of data science how to find the appropriate technique. Students have to find out which method is fit for the problem by trial and error. To be a good data scientist, such training is essential.

“Experience of Success” is important for encouraging students. If the result of the project-based learning is successful, students will be positive and actively tackle more advanced study.

Failure may be a stepping-stone to success, but, especially for beginners, experience of success is necessary to keep motivation for study. So, careful instruction is necessary.

We also provide experiences from experts in various fields, such as business, medicine, disaster prevention, meteorology, public policy etc. Students will understand how data science is useful in various fields, and these success stories will stimulate students.

Group work is also important in the classroom. In data science, there are several methods to be applied, and usually the answer is not unique. In group works, students teach each other and exchange their ideas. For computer programming, group work is especially effective. Teaching each other would be more effective than learning from an instructor.

OUR CURRICULUM OF DATA SCIENCE

In this section, we will briefly describe our curriculum of data science in Shiga University, from freshman to senior in order.

- Freshman

In the first semester, several introductory courses are given. In “introduction to data science”, students will learn how data science is useful in the real world. Examples of data science applications in various fields (business, industry, medicine, environment, meteorology, social survey, etc.) are treated. In “introduction to computer science”, students learn basic concepts of computers, software, and ethics in computer science. “Presentation skills training” is a course for presentation / communication / data-visualization. In this course, students will prepare presentation using power-point files. “Theory of data science” is a course for experiences of data science experts from various fields. “Basics of computer application” treats how to use computer software (Microsoft-office, etc.) and internet, and computer security. “Basics of data analysis” gives basic techniques of data analysis, such as descriptive statistics, tabulation, drawing statistical graphs.

In the second semester, we start a project-based learning (PBL) course, “introduction to data science analysis”. Students conduct PPDAC (Problem – Plan – Data – Analysis - Conclusion) cycle using real data (consumer data, social-network data, official statistics or several open data), and they present their results using presentation skills already learned in the first semester. Several basic mathematics / mathematical statistics courses are given in the second semester. Students will learn basic calculus and liner algebra necessary for data science. In “fundamentals of statistical inference”, students will learn basic mathematical statistics theory, including population and sample, statistical inference, and analysis of contingency table. In “data structure and algorithm”, students will learn basic concepts of data structure and algorithm which are needed in computer programming. In “programming I”, students will learn the computer programming of Python, associated with a practice session. In “basics of computational data analysis”, students will use Microsoft-Excel for tabulation, computing some descriptive statistics, statistical inference, statistical test, regression and data visualization.

- Sophomore

In the third semester, “data science ethics” treats ethics and legal requirements which are necessary for using personal microdata. Concepts of informed consents and intellectual property are also treated in this course. In “introduction to practical data analysis”, students will learn, following “theory of data science A” in the first semester and “theory of data science B” in the second semester, experiences of experts from various fields. Not only their results of the analysis, but also the methods and techniques of their analysis will be treated in detail. “Programming II” treats, following “programming I” in the second semester, advanced topics in computer programming. “Basics of computational data analysis” is a course in which students learn the programming of R, a free software widely used in data analysis. Several courses of mathematical statistics are given, such as linear regression models and multivariate analysis. “Sample survey” treats basic concepts of sample surveys, methodologies of sample design, questionnaire design, data collection, compilation and tabulation. In “social research I”, students will learn practical methods of how to conduct social research surveys.

Figure 2. Curriculum tree of data science in Shiga University

https://www.ds.shiga-u.ac.jp/en/

In the fourth semester, a PBL course, “data science fieldwork practice” is given. In this course, students will, in a group work, collect data (by a statistical survey, via internet, sensor data in a manufacturing factory, etc.) and analyze them. In this course, students will experience how to “extract value from data”, that is, data science. Advanced courses in computer science and mathematical statistics are given, such as “database”, “programming design”, “applied mathematics (discrete mathematics)”, “introduction to time series analysis” and “multivariate analysis”.

- Junior and Senior

In the fifth semester, courses for graduate thesis start, “data science advanced practice A/B” (junior) and “data science graduation practice A/B” (senior). In these courses, students will conduct research under the instruction of professors. Students will use real data and extract value from data. Staff in Shiga University has wide variety of research area, such as mathematical statistics, time series analysis, artificial intelligence, machine learning, business statistics, finance, statistical quality control, medicine, etc. Students can select their interested topics.

Various “methodology of creating values from data” courses are given, such as marketing, financial statement analysis, environment policy, official statistics, etc. Also we give several advanced courses in various fields, such as “statistical quality control”, “time series analysis”, “survival analysis”, “qualitative data analysis”, “machine learning”, etc.

We assume three types of course series, “data engineering”, “data analytics/analysis” and “data consulting”. Students specialized to data engineering will take courses such as “information theory”, “pattern recognition”, “artificial intelligence” and “data mining”. Students specialized to data analytics/analysis will take courses such as “Bayesian approach”, “optimization theory”, “probability theory (stochastic process)” and “advanced course in statistics (statistical estimation, maximum likelihood, information criterion, MCMC, etc.)”. Students specialized to data consulting have to know various methods and examples of data science application, so they will take several advanced courses already stated above, such as “methodology of creating value from data”.

COLLABORATION WITH VARIOUS INSTITUTIONS

We believe that data science is a practical science. Using real data in data science education is essential. On the other hand, there is a strong demand for data science in business sector. Therefore, we extensively collaborate with various businesses. They provide us real data they use, we use these data for education, students analyze data, and we give back the result of the analysis.

In the public sector, there is a strong demand for EBPM. We also collaborate with central/local governments, to conduct research on EBPM with local governments, on how to use data science for policy makings. We also collaborate with governments to train their staff on data science.

CONCLUSION -SHIGA UNIVERSITY MODEL OF DATA SCIENCE EDUCATION-

We believe that data science is a study of “extracting value from data”. For this purpose, data scientists have to obtain the following skills;

-Mathematics and statistics theory,
-Computer skills,
-Practical experience at various fields,
-Communication and presentation skills.

Figure 3 summarizes Shiga University Model. Our curriculum at Shiga University is designed to study these skills in a balanced manner.

In addition, the project-based learning (PBL) using real data is essential for data science education. We extensively employ PBL courses in our curriculum.

We call these overall structures “Shiga University model of data science education”.

In the public sector, there is a strong demand for EBPM. Data science is a key factor to promote EBPM. We actually collaborate with central/local governments to train their staff, and we hope our model of data science education would help to promote EBPM.

Figure 3. Summary of Shiga University Model

ACKNOWLEDGEMENTS

Partially supported by national university reform improvement promotion subsidies from the Ministry of Education, Culture, Sports, Science, and Technology of Japan and by the Institute of Statistical Mathematics (ISM) Cooperative Research Program (2017 ISM CRP -4306 and CRP-4308).

REFERENCES

Shiga University. (2017). *Website of Faculty of Data Science*. https://www.ds.shiga-u.ac.jp/

Takemura, A., Izumi, S., Saito, K., Himeno, T., Matsui, H., & Date, H. (2018). Shiga-University Model of Data Science Education (in Japanese). to appear in *Proceedings of the Institute of Statistical Mathematics, 66*(1).

Drew Conway Data Consulting, LLC. (2015). *The Data Science Venn Diagram*. Retrieved from http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram