BRINGING UP DATA SCIENTISTS IN JAPAN: YOKOHAMA CITY UNIVERSITY

Jinfang Wang and Masataka Taguri
School of Data Science, Yokohama City University, Japan
wang@yokohama-cu.ac.jp

To meet the challenge of bringing up data scientists in Japan, the School of Data Science of Yokohama City University(YCU) was established in April 2018. It is the second of its kind after the establishment of the School of Data Science of Shiga University in 2017. In this talk we will give detailed information about YCU's undergraduate program on data science and provide related information on statistics and data science education in Japan as well.

INTRODUCTION

Everything about science is now changing because of the impact of information technology. Nothing is unrelated to data science. However, the shortage of talents in data science is serious across the world, and especially so in Japan. There is no school of statistics in Japanese universities. The School of Data Science of Yokohama City University (YCU) was established in April 1, 2018 under such background. The School of Data Science of YCU is the second of its kind after the establishment of the School of Data Science of Shiga University in 2017.In this talk we will give detailed information about YCU's undergraduate program on data science and provide related information on statistics and data science education in Japan as well.

SCHOOL OF DATA SCIENCE, YOKOHAMA CITY UNIVERSITY

What is data science

Data science (DS), or data-driven science, is regarded by some renowned scientists as the fourth paradigm of science, after the empirical, the theoretical and the computational paradigms (Tansley and Tolle (2009), Bell *et al*. (2009)). Surprisingly, the term *data science* was formally used only very recently. It was Hayashi (1996), a Japanese statistician and former director of the Institute of Statistical Mathematics, who first introduced the term data science in a roundtable discussion in a conference held in Kobe in 1996. Data science aims at extracting knowledge from data in various forms. Data science is sometimes understood as a concept attempting to unify statistics, data analysis, machine learning and related methods. However, the concept of data science is now being used in a much broader sense by many people in academia and industry as well.

Basic skills for a data scientist

We may pin down some basic skills for a successful data scientist. These skills span across many fields. While some skills are much advanced for undergraduate studensts, in designing YCU's DS undergraduate curriculum we however made a great deal of efforts to cover the basic areas of learning.

- Data-Driven Thinking
  Albert Einstein (1879-1955) once said that "*If I had an hour to solve a problem and my life depended on the solution, I would spend the first 55 minutes determining the proper question to ask, for once I know the proper question, I could solve the problem in less than five minutes.*" This perhaps is especially so for a data scientist, when he/she approaches a practical problems. A data scientist is expected to scrutinize the hidden features of complex real problems and be able to ask productive questions for future investigations.
- Programming Skills
  A successful data scientist needs to be familiar with all or most of the dominating programming languages and software packages. These programming languages include, R (R Core Team, 2018), Python (Python Software Foundation, 2018), SQL (structured query language) and Hadoop. While R and Python are two leading programming languages and software environments for statistical data analysis, data visualization, etc., SQL is a programming language for managing data in relational database. On the other hand, Hadoop

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

is a framework that allows for the distributed processing of large data sets across clusters of computers.

- Statistics and Machine Learning
  Basic probability theory, statistics (estimation, testing, regression, etc.), machine learning (e.g., neural networks) are necessary for both implementing and interpreting various data analyses. Behind all these concepts one needs a solid understanding of multivariable calculus and linear algebra.
- Data Visualization
  Data visualization, or sometimes referred to as visual communication, involves the creation and study of the visual representation of data. "Excellence in statistical graphics consists of complex ideas communicated with clarity, precision and efficiency." (Tufte, 2001). A data scientist should not only familiarize himself with data visualization tools but also with the principles of data visualization as well. Some example of data visualization include bar chart, histogram, scatter plot, network, streamgraph, treemap, gantt chart, heat map, etc.
- Communication Skills
  Data scientists must be able not only to communicate with non-technical colleagues in different areas of disciplines, but also be able to elucidate their data-scientific findings in coherent manners.

Curriculum

The following is a bird's-eye view of YCU's curriculum related to learning of data science.

- Freshman
  Besides liberal arts and English courses, Introductory Data Science, Algebra, Calculus and Computer Science are compulsory at the freshman level.
- Sophomore
  At the sophomore level, students learn advanced statistics and computer sciences, along with other applied sciences, such as, econometrics, finance, physics, chemistry, and biology. Some of these courses are optional. At this stage, students are also to be exposed to project-based learning (PBL), which is also part of the requirement for graduation research.
- Junior and Senior
  Throughout the last two years, students continue to study theoretical and applied statistics and computer sciences. A good deal of efforts will also be dedicated to PBL. Students will write their data-scientific reports for their graduation thesis. These theses will be naturally resulted from their PBL.

YCU's School of Data Scienc currently offers the degree of Bachelor of Data Science. We are also being preparing graduate courses in data science in near future.

REFERENCES

Hayashi, C. (1996). What is Data Science? Fundamental Concepts and a Heuristic Example . In *Data Science, Classification, and Related Methods, Proceedings of the Fifth Conference of the International Federation of Classification Societies (IFCS-96)*, Kobe, Japan, March 27-30, 1996.

Hey, G. Bell & A. Szalay, A. (2009). Computer science: beyond the data deluge, *Science, 323* (5919), 1297-1298.

Python Software Foundation (2018). Python Language Reference, version 3.6.5 Available at http://www.python.org

R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Tansley, S. & Tolle, K.M. (2009). *The Fourth Paradigm: Data-intensive Scientific Discovery*, Microsoft Research.

Tufte, E. R. (2001). *The Visual Display of Quantitative Information*, Graphics Pr. 2nd.