Data Science and Natural Science

Shiro Ikeda

Professor, The Institute of Statistical Mathematics, The Research Organization of Information and Systems, and Kavli IPMU Visiting Senior Scientist

The word "science" means "the deepening of and distribution of knowledge of a subject." When the subject is a natural phenomenon, we call it natural science; if the subject is society, we call it social science. The subjects studied in the Kavli IPMU are physics and mathematics; therefore, they fall into the categories of formal science and natural science. In my institute, the main subject is the methodology of data analysis. Therefore, we call our science "data science." To be precise, our subject is "to deepen and distribute the knowledge of data analysis."

Recently, the fields of data science, including statistics, machine learning, and artificial intelligence, are attracting a great deal of interest from both academia and industry. The amount of data is increasing rapidly as measurement technology advances and the internet expands, while data science is becoming powerful as applied mathematical methods advance and computational power increases. It is expected that combining these two elements will bring new innovations.

Data science consists of theoretical and practical studies. As for the theoretical side, I have studied information geometry, which applies differential geometry to information theory and statistics. On the practical side, I have been working on astronomical data analysis for several years. Since last year, I have been a visiting scientist at the Kavli IPMU, and I am starting new collaborations. Each project of the Kavli IPMU is interesting from the data scientific viewpoint. Some of them are listed below.

The author reports his research activities in collaboration with Kavli IPMU researchers to the External Advisory Committee of the Kavli IPMU on August 22, 2016.

For the last two years, I have been participating the Subaru/Hyper Suprime-Cam (HSC) survey project through the Japan Science and Technology Agency (JST) CREST Big-Data research program (PI: Naoki Yoshida). The HSC repeatedly takes images of target fields, and one of the tasks is to find new supernovae from image subtraction (see Fig. 1). Although the number of candidate transients detected from the

image subtraction is more than 50,000 per night, the expected number of supernovae is around 50. In order to discover good supernova candidates, an automatic detection system must be implemented. We have developed a type Ia supernova detector using machine learning technique in collaboration with NTT Communication Science Laboratories. The detector is used for real observations.

I have started discussions with researchers working on experiments conducted at Kamioka. In both T2K and XMASS, whose scientific goals are different, it is essential to find the target events from a vast amount of data. A natural strategy for event detection is to compute the probability of the target events based on physical models. We need to develop a method to compute the probability quickly. We will continue discussions for possible contribution.

LiteBIRD is a future satellite project to measure temperature fluctuations in the CMB. The fluctuations are much smaller than the foreground emissions, and the separation based on their statistics would be essential for the measurement. For the Planck data analysis, an Independent Component Analysis (ICA) method, which was developed in the 90s, was applied. I studied ICA methods when I was a post-doctoral fellow, and am expecting to apply them to new data.

When we collaborate with researchers in other fields, just receiving data via e-mails and sending the results back does not work properly. Sometimes their expectations of data science are too high, and sometimes our understanding of the data is not sufficient. In either case, one of the main reasons for misunderstanding is a lack of communication.

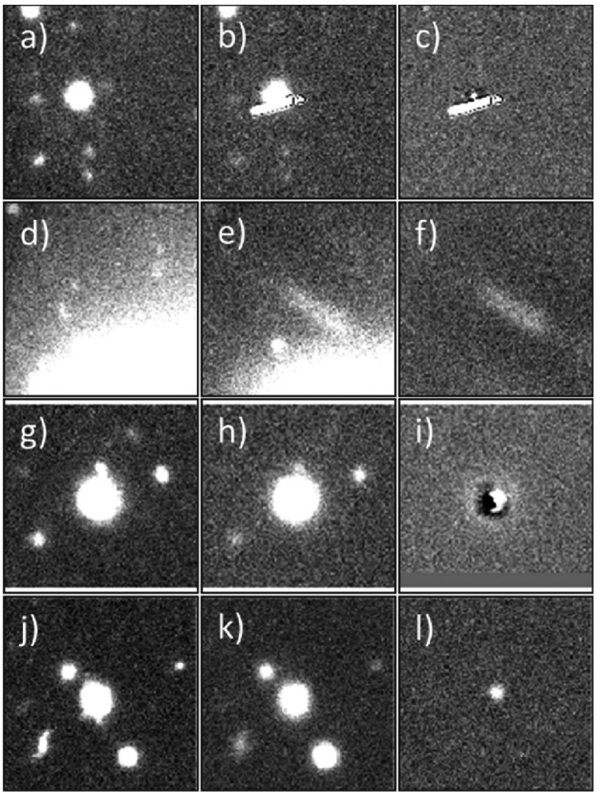

Fig. 1. Examples of real and bogus objects obtained with Subaru-HSC. The left, middle, and right columns show the reference, new, and difference images, respectively. Each row shows the cosmic ray (a–c), ghost near a bright star (d–f), inaccurate image convolution or astrometric alignment (g–i), and a real transient located in a galaxy (j–l).

An important role of a data scientist in a project is the consultation of data analysis. It is important to have discussions to understand the data and what data science can do for it. I have just started collaborations with the Kavli IPMU. I am looking forward to participating in different projects.