Applying Big Data in Innovative Ways as the Leader of a New Interdisciplinary Research Center

Naonori Ueda
Director of Machine Learning
Data Science Center,
Senior Distinguished Scientist,
NTT Communication Science Laboratories

Big data has been attracting a great deal of attention in recent years, but there is still no real sense of the role it can play in society. NTT is now taking a giant step toward the widespread, practical application of big data analysis. Senior Distinguished Scientist Dr. Naonori Ueda is an active pioneer in the research of machine learning. We asked him about his career in research and about the Machine Learning Data Science Center that he directs.

Keywords: big data, machine learning, statistics

From an early interest in statistical learning theory to becoming an expert in machine learning

—Dr. Ueda, please tell us about your research career to date.

After receiving my M.S. degree in communication engineering from Osaka University in 1984, I entered the Electrical Communication Laboratories of Nippon Telegraph and Telephone Public Corporation in Yokosuka, Japan. At that time, I was assigned to a laboratory involved in the research and development (R&D) of image-processing and pattern-recognition technologies, but since the group was right in the middle of using their research results in practical applications, and as I was just a newcomer, I was relegated to doing tasks outside research such as taking the minutes of meetings and carrying magnetic tapes to other research sites. During whatever free time I had, I read books and papers on specialized topics to learn as much as I could, and I found myself yearning to become more involved in pure research.

One year later, Nippon Telegraph and Telephone Public Corporation was taken private to become NTT. I was given the task of covering up labels saying "Nippon Telegraph and Telephone Public Corporation" on a huge number of devices and fixtures with stickers having the name "NTT." I remember with some fondness how I was buried under with this work for days on end and how I even surprised, if not angered, some female employees when I crawled under their desks to affix these stickers. NTT R&D was also targeted for restructuring, and with the establishment of NTT Human Interface Laboratories in 1987, some personnel including myself were transferred there, and I was assigned to do basic research

Technology for giving learning capabilities to machines
(information processing system)

General-purpose data analysis technology based on mathematical statistics and optimization theory
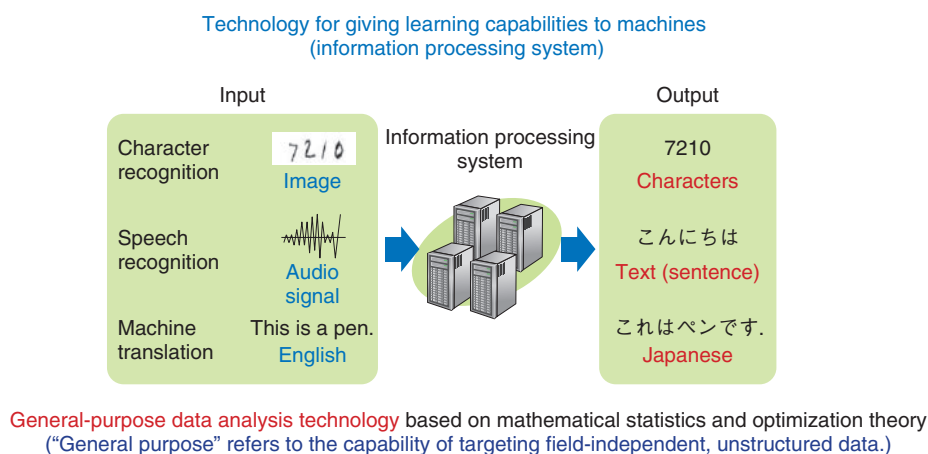("General purpose" refers to the capability of targeting field-independent, unstructured data.)

Fig. 1.   Outline of machine learning technology.

in the field of computer vision. Specifically, I took up the theme of image understanding from line drawings, and it was at this time that I began to develop an interest in statistical learning theory.

Then, in 1991, I moved over to the newly established NTT Communication Science Laboratories in Keihanna and began doing basic research in the field of statistical machine learning (**Fig. 1**). The next ten years proved to be a very fulfilling time for me as I was able to immerse myself in research with stays at Purdue University in the United States as a visiting scholar, the University College London in the United Kingdom for short-term research, and elsewhere. Then, beginning in 2010, I served for about three years as director of NTT Communication Science Laboratories. Finally, in April 2013, I became a Senior Distinguished Scientist and in July, Director of the newly established Machine Learning Data Science Center (MLC). I should mention here that MLC was initially founded as a virtual organization in April 2013 but was launched as a formal organization in July 2013. My superiors said to me, "There are great expectations that MLC will bring machine learning and big data together and make a huge contribution to the R&D of big data applications at NTT." As director, I take my position very seriously and realize that this is a heavy responsibility.

Machine learning as an optimal tool for creating new value from the analysis of human behavior

—*What exactly is machine learning?*

Machine learning was developed in around 1970 as a research field within the area of artificial intelligence. Neural networks, by the way, which boomed in the 1980s, are a type of machine-learning technology that has come to be used in a variety of application fields. This boom died down some years later, but machine learning, perhaps as a result of the fact that the boom died down, went on to evolve into a somewhat simple form of mathematics-oriented research called learning theory.

Learning methods can be broadly divided into supervised learning and unsupervised learning. An example of supervised learning would be a mother (as teacher) instructing her child on how to discriminate between a hat and a ball by saying, "this is a hat but that is a ball." This type of learning has come to be used in pattern recognition technology as in the recognition of hand-written characters. An example of unsupervised learning would be a child observing what is common and what is different between a hat and a ball on his or her own. Clustering, or grouping of similar data, which is a basic process in data analysis, can be thought of as unsupervised learning. Various types of cluster analysis have recently come into use for such applications as determining what sorts of communities have formed on a social networking site or what sorts of shoppers are using an e-commerce site.

What is the difference between machine learning and conventional data analysis? Predicting the weather, for example, is a *science* that derives answers based on the way in which the atmosphere changes, for example, from the formation of certain types of
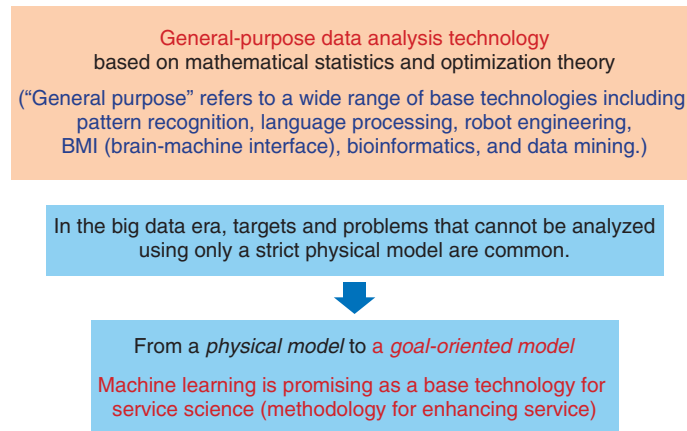
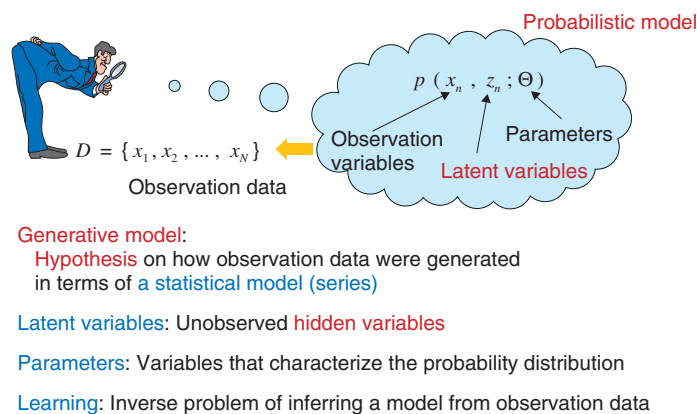Fig. 2.   Usefulness of machine-learning technology in the big data era.

Fig, 3.   Statistical machine learning based on a statistical model.

clouds. Machine learning, however, is different in that it obtains knowledge from many observations much like a local fisherman would say, "If these kinds of clouds appear, it will rain tomorrow."

Although some basic principles may be used, machine learning predicts the future by establishing a model for the mechanism that generates the data behind the observation data. This model is not necessarily true, which lies in contrast with science, in which the theories that it derives must hold true.

For example, if the process of modeling human purchasing behavior came down to the modeling of humans themselves, it could very well take more than 100 years to come up with a good model. A much faster approach would be to use observation data on what kinds of people purchase what types of products

under what conditions to realistically determine what actions an individual can be expected to take. This is called a data-driven approach.

Modeling in machine learning can consequently be called a *service science* (science whose purpose is to enhance service), and in this sense, it has come to be applied to big data applications that deal with a huge and diverse amount of real-world data. In short, machine-learning technology has come to be viewed as a useful tool in big data analysis (**Fig. 2**). There are various types of research approaches in machine learning, but I myself am researching a statistical machine-learning approach based on statistical models (**Fig. 3**). With a statistical model, it becomes possible to make use of even vague information and to make predictions in a statistical manner. I believe that

Key application fields for NTT business:
Expanded research in specialized fields

Database technology:
Diverse data-storage, compiling,
and search techniques

Parallel distributed computing technology:
High-speed computing techniques for
large-scale data

Machine learning technology:
Analysis technology for creating new
value from latent information
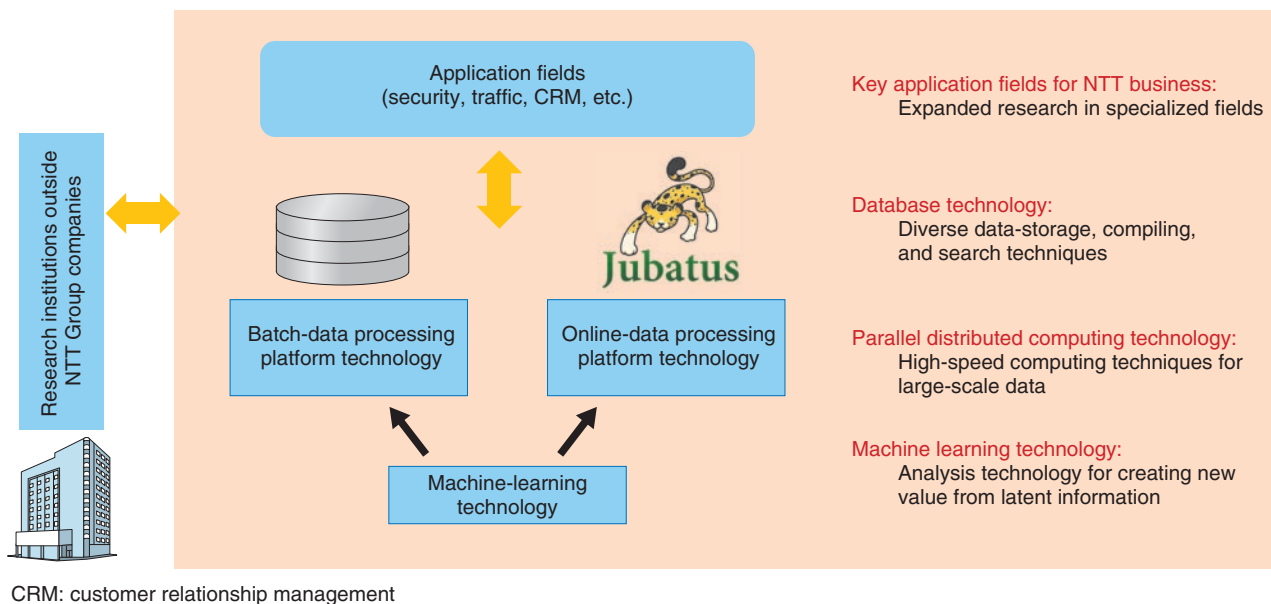
CRM: customer relationship management

Fig. 4.   Machine Learning & Data Science Center (MLC).

this is a useful technology for carrying out decision-making using big data.

Research in the field of machine learning predates the big data era, and machine-learning technology has, in fact, come to be used in a variety of fields such as machine translation. Machine-learning technology has also come to be used as an elemental technology in voice-recognition systems such as NTT DOCO-MO's Shabette Concier (Talking Concierge) voice agent and in portal-site recommendation systems. It would be no exaggeration to say that machine-learning technology is a general-purpose technology for information processing systems.

Aiming toward a new world merging Jubatus and machine learning as Director of the new Machine Learning Data Science Center launched this spring

—*Please tell us about the Machine Learning Data Science Center.*

The term "big data" has come to be somewhat of a buzzword today, but MLC was established as an R&D base within the NTT laboratories to create innovative services from big data. It is an interdisciplinary organization bringing together researchers from related fields that have up to now been dispersed across a number of research laboratories. There are

various dimensions to big data such as data collection and management and data security, but R&D at MLC focuses on the analysis of big data. It is for this reason that machine learning by itself is insufficient. To process a massive and diverse volume of data in an efficient manner, we are researching and developing, for example, the combination of machine-learning technology with the Jubatus parallel distributed computing environment developed jointly by NTT laboratories and Preferred Infrastructure (PFI), a venture company (**Fig. 4**).

Our mission is twofold. First, we will brainstorm with NTT Group companies as a technology consultant to determine how to analyze big data held inside and outside the NTT Group to create new value, and we will develop new technologies for doing so. Second, we will create innovative general-purpose data-analysis technologies in the context of big data. These two parts of our mission have a mutually dependent relationship. In the second part, the presentation of an academic paper on a new data-analysis technology does not mark the end of our work. The technology is meaningless if it is not applied to a real-world problem and evaluated accordingly. In the big data era, field trials take on even more importance.

We currently have about 20 researchers working at MLC, which is hardly a great number, but I think we will become a substantially larger organization as we broaden our interaction with NTT Group companies

going forward.

—*The Machine Learning Data Science Center seems to have received much attention at Open House 2013.*

Yes, I received questions from reporters on the purpose of launching MLC at the NTT Communication Science Laboratories Open House 2013 held in June. There was also much interest in what types of analysis would be targeted by machine learning. To give a specific example, we are collaborating with the medical-care producer of the NTT Research and Development Planning Department to develop a system for analyzing the health-checkup data of NTT employees. We feel that such analysis will prove useful in health promotion and disease prevention for all NTT employees.

I also received questions from reporters on the key feature of MLC. As I mentioned earlier, the MLC is made up of researchers from diverse technology fields such as machine learning, parallel distributed computing, and database technologies as well as applied technologies like customer relationship management (CRM) and security, all specialties of NTT laboratories. Big data analysis through inter-field collaboration is the strength of MLC.

Creating scenarios with a focus on social problems so that big data does not die out as a temporary boom

—*What is the future outlook for big data analysis?*

One difficulty in big data analysis is that the outcome of such analysis is unknown. For example, when research is done on a new device, the specifications for that device may be set beforehand as target values, and work to meet those values can then be carried out, but in big data analysis, my colleagues and I must from the start create a basic analysis scenario that describes what it is that we want to do. The big question here is whether such a scenario can help deal with social issues such as disaster prevention. Simply refining an elemental technology will not bring forth a scenario with a sense of innovation. I would like to hold discussions with group companies that have abundant channels to customers and society with the aim of contributing to the solutions of serious problems and creating value through big data analysis.

Up to now, researchers in the field of machine learning have led a somewhat behind-the-scenes existence, and they have had a comparatively high degree of freedom in their work. Now, in the big data era, they suddenly find themselves in the spotlight and thrust onto the front stage. This is essentially a boost for researchers, but it also means that they now have to deal with questions like "What can you do with your research?" and "What have you accomplished?" If researchers produce no truly useful results, big data will come to an end as just a temporary boom. I believe it is necessary for us to use the current flow in big data as a tailwind and work aggressively to refine our technology with the aim of making important social contributions.

Up to now, I myself have been engaged in basic research in the field of machine learning without any specific plans for taking on big data. Perhaps I am therefore an amateur of sorts, but in this era of big data, my aim is to determine how technology can be used to advance society as a major issue in MLC.

Playing golf—a time to forget about research matters and enjoy one's favorite hobby

—*Can you tell us how you spend your free time? Do you have any hobbies?*

For researchers, the work of research makes it difficult to differentiate clearly between weekdays and holidays and to keep research-related matters out of one's mind. I have said that scenarios are essential to creating new value from big data, but that requires imagination. Since technical books are hardly a source of imagination, I try to read a wide variety of books to expand my interests and broaden my values. I'm not sure whether this can be called a "hobby" as such.

My true hobby is golf. Although I only get out to a golf course about once a month, I also enjoy researching golf theory by reading good books about it and finding information on the Internet. I also played tennis during my student days, but I developed an interest in golf under the influence of my father and older brother. Playing golf was always enjoyable if only for the opportunities it provided for socializing, but about seven or eight years ago, I began in earnest to think about and study how to become a better golfer.

When playing golf, you come to realize that "patience is the key." When starting out, it only took one bad shot for me to feel disappointed in my game, but recently, I have learned that I can recover even if I totally botch one hole. This knowledge is a gift that

comes from actually playing the game; there is a limit to what you can learn from golf theory alone. I said earlier that it is difficult for a researcher to keep research-related matters out of his or her mind, but as for me, it's amazing that it's only when I'm playing golf that I am not thinking about research. Golf is a great way to relieve stress!

Don't become a "jack of all trades, master of none" when young! First, sharpen your skills in your own research field.

—*Dr. Ueda, can you leave us with some advice for young researchers?*

I would be happy to. I have said that imagination is important in creating scenarios in big data analysis and that there is a need to broaden one's outlook and knowledge. However, there is an appropriate time for that depending on one's age, since a young person who tries to do too much at the same time runs the risk of becoming a "jack of all trades, master of none." When you are young, I believe it's important to sharpen your skills in your own research field without worrying too much about results or the usefulness of those results. Today, the research environment is experiencing some difficult conditions, and even universities must compete for funds, which means that even young researchers must be adept at emphasizing the social value of their work. Just the same, research of a superficial nature can hardly lead to success in the end. Since young researchers are blessed with time and physical strength and are at the peak of their mental abilities, doesn't it make sense that they should sharpen their skills in their own research fields while they are still young? The way to succeed is to stay focused and to "seize the bull by the horns" without getting discouraged by criticism such as being told their kind of research is not useful. Of course,

there is also a need to maintain a sense of balance in terms of research beneficial to society, but if you are not ready to make an all-out effort to build an appreciable foundation of basic skills in an enthusiastic manner, working only with ideas will soon reach its limit. "Sharpen up your basic skills" is first and foremost a message for young staff involved in basic research, but I believe that skill sharpening could also be a beneficial approach for young staff in other fields too, such as application and development work.

Naonori Ueda

Director of Machine Learning Data Science Center, Senior Distinguished Scientist, NTT Communication Science Laboratories.

He received the B.S., M.S., and Ph.D. degrees in communication engineering from Osaka University in 1982, 1984, and 1992, respectively. He joined the Yokosuka Electrical Communication Laboratories of Nippon Telegraph and Telephone Public Corporation (now NTT) in 1984. In 1994, he moved to NTT CS Labs in Kyoto, where he has been researching statistical machine learning, Bayesian statistics, and their applications to web data mining. From 1993 to 1994, he was a visiting scholar at Purdue University, Indiana, USA. He is a guest professor at the National Institute of Informatics and a visiting professor at Kyoto University. He is a Fellow of the Institute of Electronics, Information and Communication Engineers and a member of the Information Processing Society of Japan and IEEE. He became Senior Distinguished Scientist in April 2013 and was appointed Director of Machine Learning Data Science Center in July the same year, after serving as Director of NTT Communication Science Laboratories for three years.