Special Contribution
Discovery Science Revisited

Setsuo Arikawa
Fujitsu Laboratories Fellow
Arikawa Discovery Science Center Director

1. Introduction

The term “discovery science” (hereafter, DS) was initially coined as shorthand in the research project Foundations of Knowledge Discovery from Science and Business Information, which was led by the author between 1998 and 2001 and was funded by the Ministry of Education, Science and Culture’s Grants-in-Aid for Scientific Research. The DS Project was composed of approximately 60 leading researchers in Japan, who pursued diverse studies related to the emerging academic field of discovery science. As part of the initiative, the International Conference on Discovery Science was implemented and has been held annually since the first year of the project. This year will mark its 20th anniversary.

As a witness to the twenty-year benchmark since discovery science was first promoted, this paper revisits the arguments presented during its conception period and considers its future directions as well as the challenges to be addressed in modern-day society, where machine learning (ML) and artificial intelligence (AI) have made significant progress and become widely accepted. The paper also briefly presents the research policies of the Arikawa Discovery Science Center at Fujitsu Laboratories.

2. Basic research themes in field of DS

Discovery science was initially conceived as a new type of science that would continue to have value as long as the world has scientists, engineers, and vast volumes of data. It was developed upon a structural body of research outcomes[1)–9)] covering the basic theories of ML, analogical reasoning and other inferential processes, proposal and development of an integral framework for theoretical constructions, basic theories advanced for solving specific problems and their application, and basic theoretical constructions of machine discovery.

Initially, DS research mainly aimed to (1) provide new methodologies for making scientific discoveries, (2) foster a computer network environment for knowledge discovery, and (3) establish a new fundamental domain of information science. Research activities were pursued by research groups/teams that were allocated to each research topic established on the basis of the following themes.

1) Logic of knowledge discovery

The ability to discover knowledge using computers (i.e., machine discovery) would address not only individual hypotheses and theories but also the refutability of the hypothetical space itself. Similarly, discovery research studies are evidence-based and require a theoretical foundation with precise idioms and deductive structures.

2) Inferential knowledge discovery

Discovery of (scientific) knowledge relies not only on deduction, as mentioned above, but also on inductive and abductive inference. Abduction is used in contexts where explanation is attempted by means of hypotheses about a certain phenomenon that cannot be explained with existing knowledge. In this way, it serves as an opportunity that leads to (scientific) discoveries.

3) Knowledge discovery based on computational learning theory

ML is an effective means of discovering knowledge from experimental and observational data as computers proceed by learning from the data. The underpinning theory, mathematical learning theory, is the basis for knowledge discovery methodologies, so it is necessary to conduct studies with consideration for the computational complexity required by the learning.

4) Knowledge discovery from big databases

Research concerning knowledge discovery should ultimately aim for system integration, and it should be applied in practice to evaluate ways to handle big databases in science and business, which would then serve as a valuable resource for future studies. There is also a need to establish methodologies and specific methods of data mining.

5) Knowledge discovery in the network environment

Knowledge discovery emphasizes interactive collaboration between humans and machines as well as between humans and between machines. Therefore, there is a need for research and development in computer network environments, including network agents, in order to efficiently pursue and support the collaboration related to congruity, competition, compromise, and arbitration.

The group for Theme 1) included philosophers and mathematicians. Each research topic was pursued by its respective leaders and was underpinned by outcomes of many internationally acclaimed and innovative studies. Thus, the group was structurally organized to form a new academic domain.

These research activities were followed by a project funded by the Grant-in-Aid for Scientific Research on Priority Areas, "Implementation of Active Mining in the Era of Information Flood" (abbreviated as "Active Mining") (2001–2005), led by Professor Hiroshi Motoda of Osaka University and then by Professor Masaru Kitsuregawa of The University of Tokyo, who conducted Research on New IT Infrastructure for the Information-Explosion Era, a study supported by a Grant-in-Aid for Scientific Research on Priority Areas (abbreviated as "Information Explosion"), for five years from 2005. Then, in 2012, the U.S. administration under then-President Barak Obama launched the Big Data Research and Development Initiative, which heralded the global expansion of big data involving the general public sector.

3. New perspectives for DS

In its early days, DS handled a data volume of up to several dozen terabytes. Although this was considered enormous, it is worlds apart from the state of affairs today. In Japan, meanwhile, the aforementioned studies among others played a major role in leading the world in DS research before the arrival of the big data era, through DS and new concepts such as information flood, active mining, and information explosion. Subsequent DS research continued on its unique path, with international conferences supported by the diligent efforts of people working in the DS domain. However, DS needs to develop even further in accordance with the following new perspectives.

3.1 Machine learning comprehensible to humans

If we are witnessing the third AI boom today, then the DS Project started as a new academic field based on the technologies and knowledge accumulated during the second boom. The AI of that time focused on the development of expert systems for knowledge acquisition, such as the MYCIN medical diagnostic system, so it revolved around the themes of knowledge expression and acquisition. The knowledge acquisition was broadly divided into two categories: machine-readable human knowledge, such as knowledge gleaned from interviews with experts, and knowledge acquired from vast databases.

Knowledge expression included collective knowledge, such as rules, definite clauses, and other logical expressions, and knowledge with a tree structure, such as decision trees. Both were comprehensible to machines and humans. Diverse research projects on knowledge acquisition and ML were conducted in order to explain big data.

Given this background, there was a close relationship between ML and knowledge discovery, and, levels of complexity aside, the outcomes of the learning were comprehensible to humans. Therefore, expressing the learning outcomes in such a way that people could understand them was never an issue. In recent years, however, ML and machine recognition are backed by deep learning, and they yield outcomes of learning and recognition that are difficult to understand in familiar language. This presents a significant challenge.

Thus, DS research, which will be based on future ML, must address this important theme of expressing the results of learning/recognition so that it can be read and understood by humans. It is also important to pursue research and development to discover new methods of ML that match the level of the current big data era, with its capability of leveraging super-massive active data, in order to find clear knowledge expressions, as we did during the second AI boom. If this becomes a reality, the results of ML can be translated from data to discoveries.

3.2 Review of various kinds of reasoning from big data perspective

In promoting DS, we have pursued theoretical approaches to various kinds of reasoning, such as induction and analogy, in order to propose and construct formulas and uniform frameworks for discussion,[2] and we have considered the possibilities of reasoning. For example, Professor Takeshi Shinohara of the Kyushu Institute of Technology used a uniform framework to prove that inductive reasoning from positive data is far more powerful than previously believed.[3] His study was a milestone in the field of inductive reasoning.

In the field of analogy, Professor Makoto Haraguchi of Hokkaido University formalized an extension of ordinary deductive reasoning and put forward a basic theory.[1] Analogy is a form of reasoning regularly used to find solutions and identify challenges. This is a reasoning style in which similarities are identified among given objects, and the facts and/or knowledge that apply to one are transformed to the other. Clues for finding solutions are thereby elucidated, and unknown facts or knowledge are predicted. Similar facts/knowledge thus play a key role. Although this was only theorized in the past, today we can find many possible answers online. It is thought that leveraging these answers makes it possible to apply and deploy analogical reasoning effectively.

The same can be said for other fields. For example, it is expected that the fact/knowledge-centered learning and reasoning of the second AI boom, such as inductive logic programming, can develop DS further by supporting it with modern computers and network capabilities or by reviewing it from the perspective of big data.

3.3 Data science and open science

Today, data science is drawing attention as a domain closely related to DS. Data science deals with data engineering and ML as an extension of mathematical statistics, data analysis, and other traditional fields, aiming to discover or create new value from big data. Data science and discovery science have more than their initials in common—these domains share closely related interests. As specialists who master the science and leverage it to find solutions and challenges, data scientists have been in greater demand in and of greater interest to society in recent years, and many organizations invest in nurturing human resources, including the Ministry of Internal Affairs and Communications as well as the Ministry of Education, Culture, Sports, Science and Technology.[10]

Furthermore, data is being opened progressively,[11] including not only the public-domain data provided by national and/or local governments but also data from the private sector, such as that generated by public transport operators and other industries, data obtained from large-scale experiments and observational or measurement devices, as well as data used in publicly funded studies and their outcomes. There is a growing trend in research and business alike for operators to utilize open-source data rather than collect data themselves. Academically speaking, this is a trend toward open source for science, but its scope includes open science in humanities, or citizen science, which is brought about as more and more classic literature is converted into digitized open sources.

3.4 Discovery involving human interaction

While telecommunication network systems have become faster and more prevalent, DS, like many AI applications, does not necessarily require that machines autonomously complete entire processes automatically. For example, in the conventional Japanese language input method for word processors or in the latest high-end medical diagnostic systems based on ML, candidate solutions are presented to users/medical professionals, so the user simply has to complete the selection from, or judgments based on, the given options. This way of using machines is considered sound and open to many possibilities. As far as DS is concerned, communication between machines and humans should not be limited to one time. Machines may communicate and interact repeatedly with humans as well as with the informational and social environments. Furthermore, ML or other AI capabilities do not have to be present at the other end of the network. It can just as well be human beings, who judge, react, and respond through the network. A realistic approach would be to start from an arrangement like this and work toward mechanized systems over time, taking into consideration the cost, experience, and accumulated data.

3.5 Discovery informatics

The international conference on DS, mentioned above, was the only DS-related initiative outside Japan for some time, but recently, the U.S. National Science Foundation (NSF) has been organizing Discovery Informatics (DI) workshops since 2012. They are keen to make scientific discoveries leveraging information science and technology and have published a report on systematic dialogs.[12)]

Through their initiative, they aim to explore the processes of scientific discoveries that require knowledge assimilation and inferential reasoning and apply intelligent computation and information systems in order to understand, automate, improve, or further innovate aspects of such processes. The ten important challenges they face include extraction of information and text understanding, model development, inferences from a diversity of scientific knowledge, planning/execution/pursuit of experiments, and model-guided data collection. In the basic research area, they focus on the fields of information and knowledge, interaction, and autonomy; and the major themes for DI are computational support for discovery processes, data, and models, as well as social computing for discoveries.

4. Activities at Fujitsu Laboratories

In December 2015, the Arikawa Discovery Science Center (hereafter, DS Center) was launched in tandem with the author joining the Fujitsu Laboratories Fellows. The DS Center aims to develop this ambitious study area (which originated in Japan and is unprecedented in the world as explained above) from a corporate laboratory with a new, strategic perspective.

Its members include world-leading professors from institutions such as Hokkaido University, Tohoku University, The University of Tokyo, Kyoto University, Kyushu University, and Kyushu Institute of Technology. We commission DS-related studies assigned to professors on topics that are closely related to their respective fields. Within Fujitsu Laboratories, most work is pursued in liaison with researchers in the AI laboratory. In addition to joint seminars held at the AI laboratory, close communication with the professors is maintained through visits by the Director of the DS Center and AI lab researchers (site visits) during which seminars are organized for them and their colleagues. In this way, we practice a new style of industry-academic collaboration, with goals always in focus. We also engage in Fujitsu Laboratories' unique initiative to generate solutions to social challenges, which the members pursue outside their ordinary area of work. Furthermore, we offer DS perspectives on possible solutions.

Our goals for FY2017 are to develop universal platform technology for problem-finding in order to enable conventional AI, which only solves given problems, to identify problems to be solved and to apply the problems thus discovered to social system design and open science. Our planned research activities are as follows (**Figure 1**).

1) Multi-agent: to apply game theory, matching theory, and other basic multi-agent theories in order to identify problems with and perform design of social systems.
2) Learning and discovery: to examine learning outcomes through explainable ML and develop supporting technology for problem identification by humans and/or machines.
3) Data archiver: to develop a problem-identification platform that helps to extract, compress, and store data on features and structures from big data generated by the Internet of Things (IoT).

Note that the social system design involving local communities and specialist communities, depicted in the figure, is modeled on external activities carried out by members of Fujitsu Laboratories.

5. Conclusion

This paper has presented an overview of DS from its birth to the present and described recent trends in the US and at Fujitsu Laboratories. DS should evolve with the clear objectives of discovering problems and pursuing discoveries for finding their solutions, as the DS Center also aspires to do. This pursuit involves the elucidation of the system limits through theoretical research as an important component because knowing the limits will facilitate the decision to dismiss the limited systems and move on to explore new ones. Today, there is a growing interest in the Science of Science Policy from an administrative point of view, and we expect that future DS studies will play a role in this

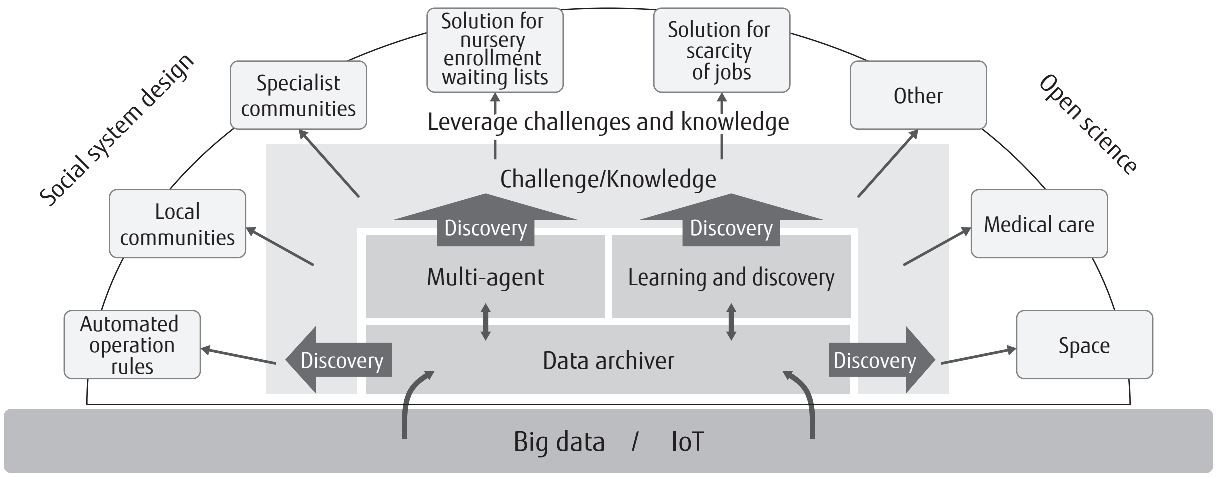

Figure 1
DS Center: Research objectives and vision for study plans.

interdisciplinary research area.

References

1) M. Haraguchi and S. Arikawa: A Formulation of Analogical Reasoning and Its Realization. J. of Japanese Society for Artificial Intelligence, Vol. 1-1, pp. 132–139, 1986.
2) S. Arikawa et al.: Learning Elementary Formal Systems. Theoretical Computer Science, Vol. 95, No. 1, pp. 97–113, 1992.
3) T. Shinohara: Rich Classes Inferable from Positive Data: Length-Bounded Elementary Formal Systems. Information and Computation, Vol. 108, Issue 2, pp. 175–186, 1994.
4) S. Shimozono et al.: Knowledge Acquisition from Amino Acid Sequences by Machine Learning System BONSAI. Trans. Inf. Proc. Soc. Japan, Vol. 35, No. 10, pp. 2009–2018, 1994.
5) Y. Mukouchi et al.: Towards a Mathematical Theory of Machine Discovery from Facts. Theoretical Computer Science, Vol. 137, No. 1, pp. 53–84, 1995.
6) S. Arikawa: Basic Research on Knowledge Discovery from Massive Academic and Social Data. Application document for FY1998 Grant-in-Aid for Scientific Research on Priority Areas, 1997.
7) S. Arikawa et al.: The Discovery Science Project. J. of Japanese Society for Artificial Intelligence, Vol. 15-4, pp. 595–607, 2000.
8) S. Morishita and S. Miyano (eds): Discovery Science and Data Mining. Kyoritsu Shuppan, 2000.
9) S. Arikawa et al.: Progress in Discovery Science, Lecture Notes in Artificial Intelligence, LNAI 2281, Springer-Verlag, 2002.
10) Ministry of Education, Culture, Sports, Science and Technology: Strengthening university education on mathematical science and data science. Report on discussion for strengthening education on mathematical science and data science, 2016.
11) Cabinet Office, Government of Japan: Promoting Open Science in Japan. Report by the Expert Panel on Open Science, based on Global Perspectives, 2015.
12) Y. Gil et al.: Final Report on the 2012 NSF Workshop on Discovery Informatics. National Science Foundation Workshop Report, 2012.
http://www.discoveryinformaticsinitiative.org/diw2012

