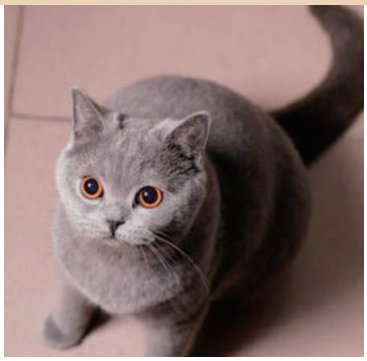
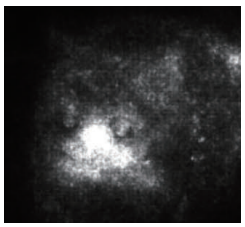
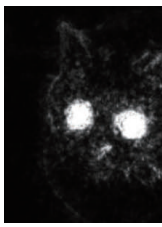
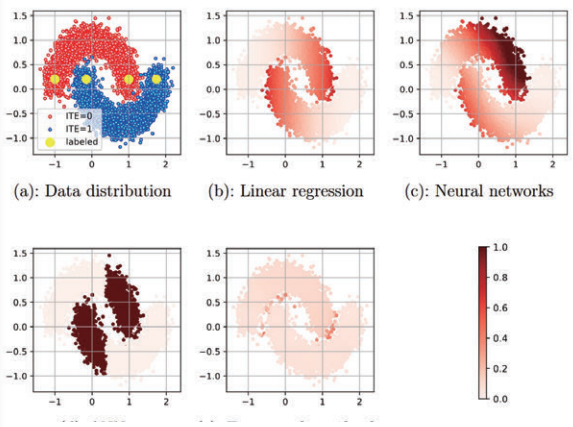
Data Science and Beyond

Data science has become very important in recent years as a methodology for inductively constructing models and hypotheses from data.
This trend has also spilled over into other academic disciplines, such as medicine, education, disaster prevention, agriculture, economics, and linguistics, leading to a surge in demand for data scientists in industries spanning agriculture and manufacturing, distribution and retailing, and finance and entertainment. The new Data Science Course, established in April 2023 as a center for education and research in mathematics, data science, and machine learning, also serves to foster top-class data science experts.
By actively welcoming students and researchers from adjacent disciplines, we also contribute to the creation of new interdisciplinary academic fields.

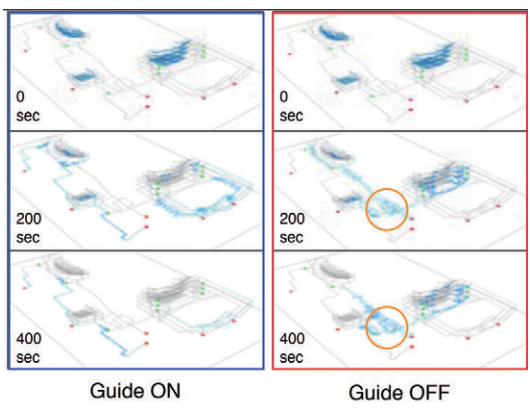




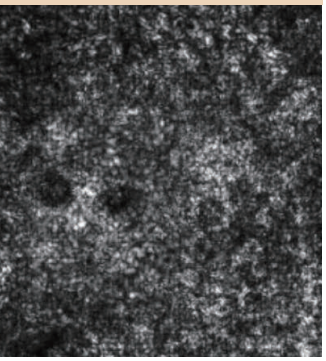


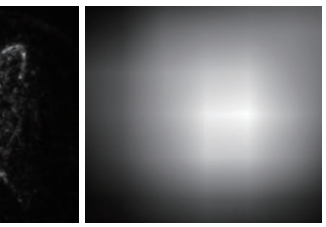


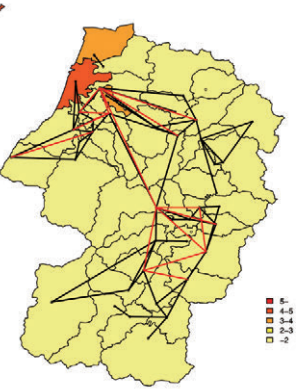
(a): Data distribution (b): Linear regression (c): Neural networks



Guide ON Guide OFF







Data Science is Drawing a Lot of Attention

Thanks to advances in information and communication technologies, a wide variety of data are now being collected and stored on an unprecedented scale. A lot of efforts are also being made in trying to inductively draw out knowledge from such data. Various new developments are also emerging in statistical and machine learning theories, which serve as frameworks to support the process of inductive inference from data. Such advances are making it possible to systematically deal with types of data that were previously difficult to handle, as well as to perform new types of data analysis. Then there is deep learning. Starting with images and text, varieties of data that are being accumulated in massive quantities, models with superior capabilities that can be mistaken as the artificial "intelligence" have now been constructed for different tasks, by appropriately training multilayer neural networks. These striking achievements in deep learning have in turn been raising new questions about frameworks for making inductive inferences from data. Why does deep learning work so well? According to optimization theory, it should be difficult to optimize the kind of non-convex objective functions that typically appear in deep learning. Statistical learning theory also states that complex models, such as multilayer neural networks, cannot be expected to have a good generalization ability. In the light of these findings, it is hard to explain why deep learning works so effectively. In the search for an answer to this conundrum, new systems and methodologies of deep learning, and by extension, data science, are likely to emerge.

We invite you to dive with us into this exciting world of data science.

Toshiyuki Tanaka

Data Science Course

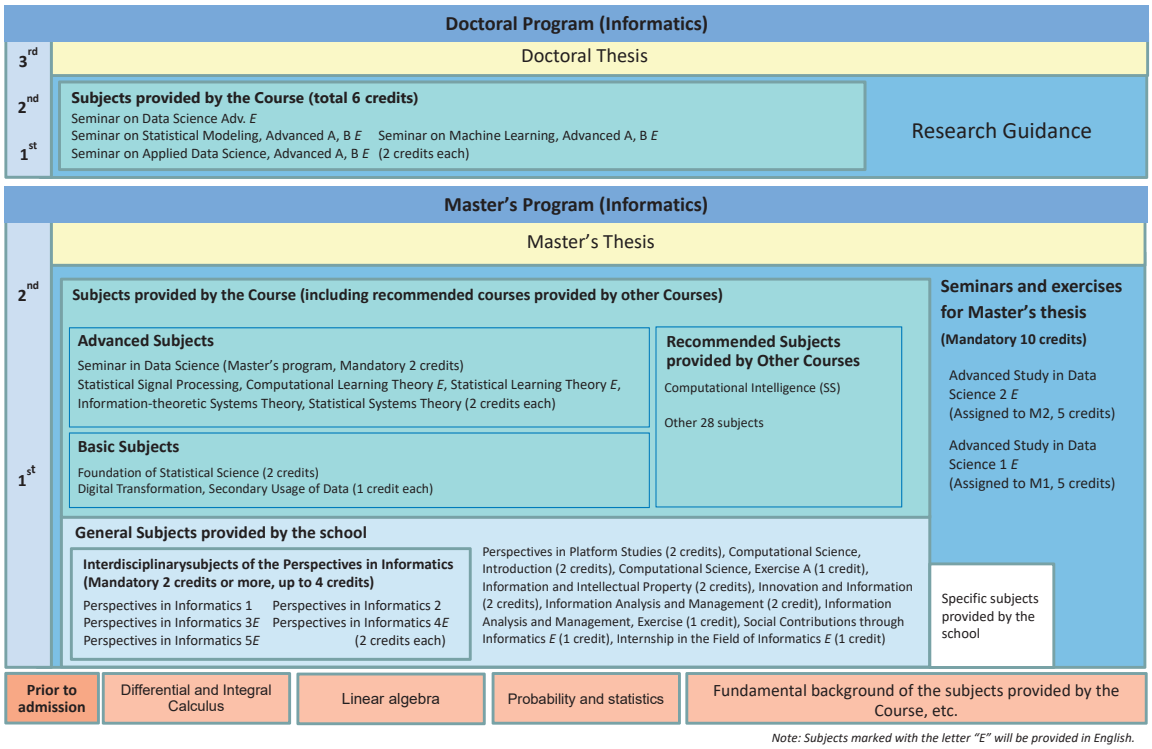
March 1993: Received a doctoral degree from the University of Tokyo. April 1993: Assistant, Department of Electronics Engineering, Faculty of Technology, Tokyo Metropolitan University. April 2000: Lecturer, Department of Electrical Engineering, Graduate School of Engineering, Tokyo Metropolitan University. January 2002: Associate Professor, Department of Electrical Engineering, Graduate School of Engineering, Tokyo Metropolitan University. April 2005: Associate Professor, Department of System Design, Faculty of System Design, Tokyo Metropolitan University. October 2005: Professor, Department of Systems Science, Graduate School of Informatics, Kyoto University.

Outline

Group and Teaching Staff

Group	Teaching Staff
Statistical Mathematics	Hidetoshi Shimodaira/Professor　Junya Honda/Associate Professor
Statistical Inference	Hisayuki Hara/Professor (Secondary appointment : Institute of Liberal Arts and Sciences)
Information-Theoretic Learning	Toshiyuki Tanaka/Professor　Tomoyuki Obuchi/Associate Professor
Signal and Information Processing	Kazunori Hayashi/Professor (Secondary appointment : Institute of Liberal Arts and Sciences)
Applied Machine Learning (Computational Intelligence)	Akihiro Yamamoto/Professor　Ichigaku Takigawa/Program-Specific Professor Brendan John Flanagan /Program-Specific Associate Professor(Institute of Liberal Arts and Sciences)
Applied Machine Learning (Collective Intelligence)	Hisashi Kashima/Professor　Koh Takeuchi/Assistant Professor
Medical and Healthcare Data Science	Hiroshi Tamura/Professor (Secondary appointment : Institute of Liberal Arts and Sciences)
Computational Intelligence Systems (Adjunct unit)	Toshiyuki Tanaka/Professor　Nonori Ueda/Adjunct Professor Tomoharu Iwata/Adjunct Teaching Staff (NTT Communication Science Laboratories)

Curriculum of Data Science Course



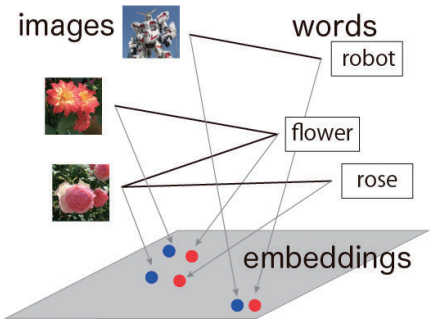
Note: Subjects marked with the letter "E" will be provided in English.

Statistical Mathematics

Theory and Application of Statistics and Machine Learning

Statistics plays a vital role as the theoretical foundation of the highly influential trends of big data, data mining, and AI. A key feature of statistics is that it provides a methodology for making inductive inferences from data, in a way that accounts for randomness. On this foundation, machine learning has developed very rapidly in recent years, from extracting information from large volumes of data to making decisions utilizing only small amounts of data. Our aim at this turning point is to develop new methodologies driven by mathematics and programming to use with real-world data.

[Hidetoshi Shimodaira , Junya Honda]



Graph embedding using neural nets to reduce the dimensionality of data and perform image-text mutual search.

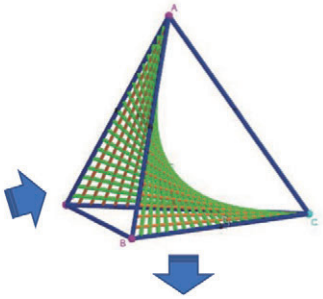
Statistical Inference

Mathematics of Statistical Inference in the Age of Big Data and Its Applications

With the rise of big data, machine learning, and AI, statistical inference theory and algorithms for statistical inference in high-dimensional complex systems are increasingly important. Although massive quantities of data are now readily accessible, the complexity of the phenomena we want to understand clearly remains daunting. It is therefore necessary to develop robust inference methods that allow us to make reliable inferences using even small sample sizes. Our research and education are focused on the development of high-dimensional inference methods and their application to fields as diverse as archaeology, anthropology, and cognitive science, by making use not only of classical inference methods but also various mathematical techniques such as combinatorics, algebraic geometry, and optimization methods.

[Hisayuki Hara]

	1	2
1	x_{11}	x_{12}
2	x_{21}	x_{22}



$$p_{11}p_{22} - p_{12}p_{21} = 0 \qquad p_{11}^{x_{11}}p_{22}^{x_{22}} - p_{12}^{x_{12}}p_{21}^{x_{22}} = 0$$

Many statistical models can be interpreted as algebraic varieties.
Algebraic quantities can offer valuable insights for statistical inference.

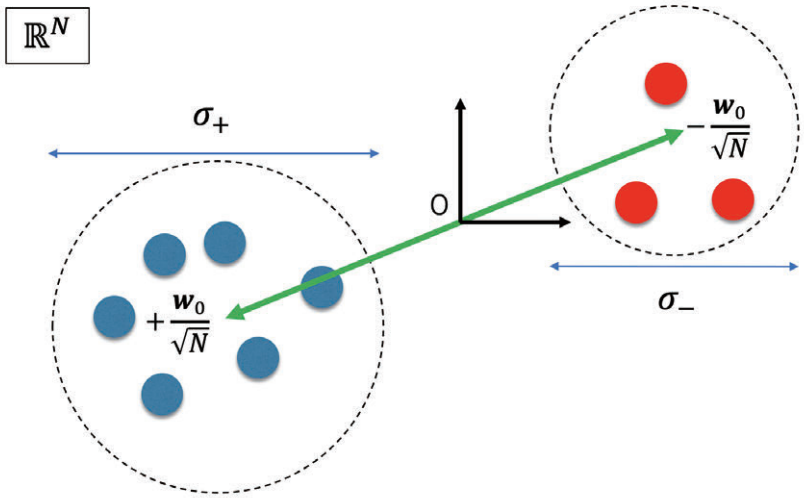
Outline

Information-Theoretic Learning

Information Mathematics for Machine Learning

Over recent years, deep learning has advanced enormously, with a rich variety of ideas and models generated and proposed. In addition to deep learning, other methods of machine learning have also been developed for dealing with various challenges confronted in data science. In some cases, the reasons why such methods are effective or the properties of such methods are not even understood. Our aim is to shed light on the mechanisms of these methods via utilizing knowledge from information theory and the statistical mechanics of disordered systems.

[Toshiyuki Tanaka, Tomoyuki Obuchi]

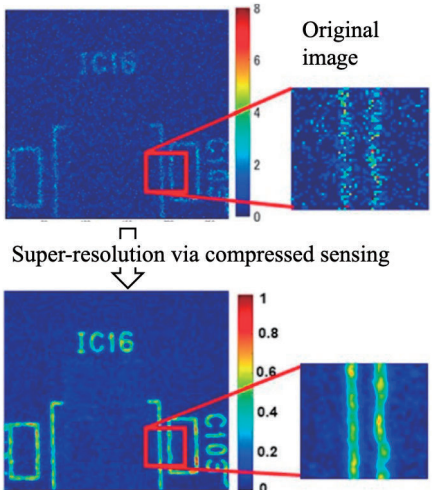


Signal and Information Processing

Stochastic and Statistical Approaches for Understanding Mathematical Systems

Our focus in this field of signal information processing is on the construction and the analysis of mathematical models for stochastic and statistical problems that emerge in various kinds of systems, and on the development of effective algorithms for practical application of such models. More specifically, we are making use of statistical signal processing, an approach to systematizing methodologies for extracting useful information from observed raw signals and data, as a tool for tackling problems mainly in communications systems, such as mobile communications, optical fiber communications, and IoT.

[Kazunori Hayashi]



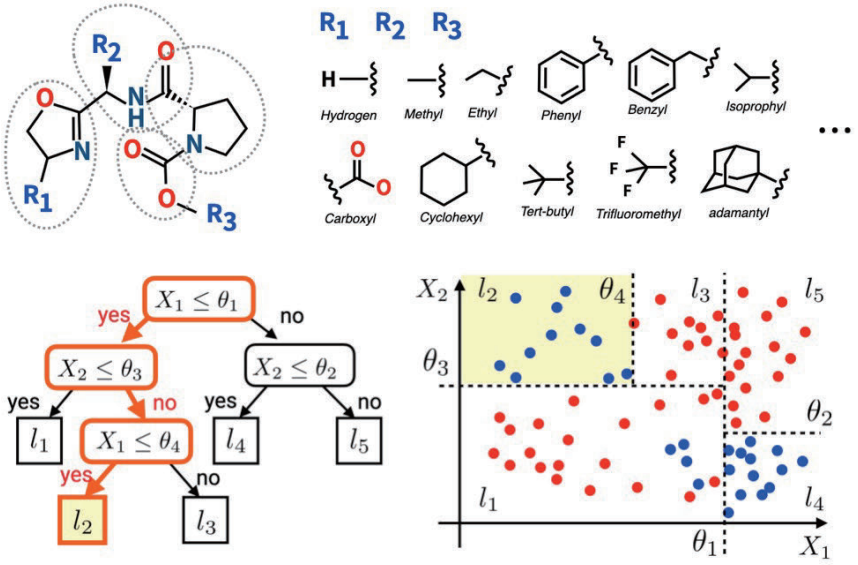
Example of super-resolution image by compressed sensing for X-ray fluorescence measurement

Applied Machine Learning (Computational Intelligence)

Machine Learning and Data Science for Solving Real-world Problems

In our work on applied machine learning (computational intelligence), we conduct practical research on machine learning and data science with a view to developing solutions to real-world problems and challenges in the fields of natural sciences and education. In life sciences and chemistry, discrete combinatorial structures, such as genome sequences, molecular structures, and intermolecular interactions, need to be reflected in machine learning. We are developing technology to address these need. In education, we are working on the development of an educational information platform and methodology to improve the effectiveness and efficiency of our own learning. Our goal is to pave the way for a new foundation for computational intelligence, by making concrete implementation through scientific discovery and understanding; through integration with general discrete structure processing using symbolic logic, knowledge representation, and algorithm design; and by comparison between human and machine learning.

[Akihiro Yamamoto, Ichigaku Takigawa, Brendan John Flanagan]

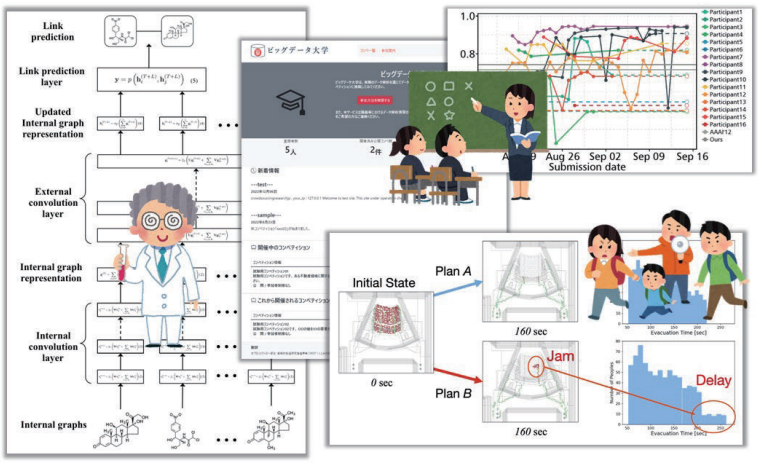


Applied Machine Learning (Collective Intelligence)

Pioneering Innovative Applications of Data Analytics

We focus on research and development in statistical machine learning and data mining techniques, and design data-driven solutions that address critical challenges in various domains, such as healthcare, biotechnology, pharmaceuticals, education, transportation, and marketing. By collaborating with industry leaders, governmental agencies, and other organizations, our goal is to break new ground in the applications of data analytics and generate impactful solutions for real-world problems.

[Hisashi Kashima , Koh Takeuchi]



Leveraging advanced analytics to address challenges across various domains.

Outline

Medical and Healthcare Data Science

The medical research outlined below is conducted mainly in Kyoto University Hospital.

Research Themes

■ **Epidemiological research utilizing clinical data including imaging and genetic data**
Clinical research consisting mainly of observational studies, in combination with multimodal imaging and genetic data, elucidating pathophysiology or predicting prognosis.

■ **Epidemiological research utilizing claims data, including the National Database of Health Insurance Claims and Specific Health Checkups of Japan (NDB)**
Epidemiological studies based on the secondary use of claims data, which plays a central role in the Data Health Plan featured in the Japan Revitalization Strategy

■ **Research on implementation and improvement of hospital information systems for improving hospital management**
Research aimed at improving hospital management by organizing and verifying implementation of issues in the renovation, improvement, and secondary use of hospital information systems involving new technologies such as IoT and AI

■ **Research on the development of a gaze analyzing visual function evaluation device**
Study on using a "gaze analyzing visual function evaluation device" featuring eye tracking and AI for analysis, to assess multiple visual functions, such as visual acuity and eye movements, with a focus on the visual field. A further goal is to develop more advanced eye tracking technologies suitable for screening of cognitive impairment.

[Hiroshi Tamura]

Computational Intelligence Systems (Adjunct Unit)

Toward Knowledge Creation from Diverse Data

With advances in deep learning, machine learning methods have achieved a high level of performance in image and language processing. However, when a large quantity of good training data is not available, low performance becomes a problem. To enable the effective use of machine learning in a broader range of fields, we are working on machine learning methods that can achieve high performance even under undesirable conditions, e.g., when there are only a small amount of data available or when high-quality data cannot be obtained. More specifically, we are working on meta-learning, i.e., the process of learning how to learn, and on machine learning that takes advantage of inductive bias.

[NTT Communication Science Laboratories: Naonori Ueda, Tomoharu Iwata]