TOHOKU UNIVERSITY

Graduate Program in Data Science

1st Data Science Cafe

"Data Science Workshop" held by the Graduate School of Information Sciences Tohoku University

2023 March 13 MON

10:00 - 10:05	Opening Remarks: Dr Hafumi Nishi
10:05 - 11:20	Invited Talk 1: Dr Pawan Tripathi（Case Western Reserve University）
11:20 - 13:00	Lunch break
13:00 - 14:00	Invited Talk 2: Dr Naho Orita（Waseda University）
14:00 - 14:30	Discussion: Dr Naho Orita（Waseda University）
14:30 - 15:10	Invited Talk 3: Dr Junyue Wu（Tohoku University）
15:10 - 16:25	Invited Talk 4: Dr Anthony Poole（University of Auckland）
16:25 - 16:30	Closing Remarks: Dr Mitsuyuki Nakao

＊Refreshments and snacks will be served during the discussion time.

MAP

Venue 会 場

Large Meeting Room (3F), Innovation Center for Creation of a Resilient Society

東北大学青葉山キャンパスレジリエント社会構築イノベーションセンター 3 階大会議室

Contact

JP

東北大学大学院情報科学研究科

山田和範

Phone: 022-752-2206

Email: yamada@tohoku.ac.jp

EN

Siwalee Choilek

Graduate School of Information Sciences

Phone: +81-22-795-7691

Email: choilek.siwalee.a5@tohoku.ac.jp

English

Japanese

Data-driven Automated Analysis Pipeline for High-energy Synchrotron Beamline 2-D XRD Datasets

Pawan K. Tripathi

Department of Materials Science and Engineering, Case Western Reserve University, Cleveland OH, USA. Materials Data Science for Stockpile Stewardship: Center of Excellence, Cleveland, OH, USA

High energy 2-D X-ray diffraction (XRD) diffractograms obtained at synchrotron facilities play an important role in determining materials properties associated with crystal structure such as crystalline phases, texture, chemistry, etc. However, the data produced at third-generation synchrotron sources is enormously massive and difficult to analyze. In the current work, we are developing a comprehensive pipeline for automated analysis of beamline 2-D XRD patterns using state-of-the-art data science approaches. The pipeline involves developing standard practices for the FAIRification of the data to make them findable, usable, interoperable, and reusable. It further involves the use of distributed computing environment, image processing, and deep learning techniques to help create an efficient and accurate solution capable of processing large amounts of image data. Herein, we analyzed patterns of a Ti-6Al-4V (Ti-64) alloy that was heat treated throughout the capture of the diffraction patterns at Advanced Photon Source (APS) facility. A convolutional neural network (CNN) was designed to predict the titanium beta phase volume percentage during temperature fluctuations. Images were pre-processed to remove bias in the XRD patterns and prevent loss of information. The 2-D XRD patterns were input directly into the CNN model instead of traditional 1-D XRD peak patterns. The model was trained using an experimental dataset that contains 3,012 XRD images and was tested with another 1,102 experimental XRD images, achieving a mean square error (MSE) of 0.076%.

What artificial intelligence can learn from first language acquisition

Naho Orita

Faculty of Science and Engineering, Waseda University

This talk will show a series of research in the field of first language acquisition and discuss whether and how these findings could help improve "artificial intelligence" that could potentially learn a natural language like a human child.

Machine learning applications in 3D modeling

Junye Wu

Graduate Program in Data Science, GSIS, Tohoku University

This talk will introduce new machine learning applications in 3D modeling with live demonstrations using open-source software Blender. 4 add-ons will be focused:

1. AI render: Feed rendering and text prompt to generate image using Stable Diffusion.
2. DMT-mesh: Generate 3D mesh with text or image based on Pint-E.
3. Dream Texture: Generate/Project Texture with Stable Diffusion.
4. DeepBump: Generate physics-based rendering (PBR) material.

The origin and longevity of information on Earth.

Anthony M. Poole

Professor, School of Biological Sciences, University of Auckland, New Zealand

All life on Earth uses a common system for storing genetic information. Information is stored in DNA as genes. Genes can be read to generate two products: RNA and proteins. Proteins are encoded by a universal genetic code, indicating that all life shares a common evolutionary origin, and some RNAs are central to the machinery for reading the genetic code. Scientists have developed methods to read information in genes, and, more recently, methods for synthesising artifical genes have become common. It is even possible to store digital data in DNA polymers.

In this seminar, I will address three broad questions.

1. How did DNA originate?
I will examine how DNA came to become the primary material for genetic storage, and I will present some of my group's work on this topic which shows that DNA carries a faint but discernible signal of its own origins. I will also present ongoing experimental work aimed at understanding how DNA systems originated.

2. How has evolution changed information processing in biological systems?
One fascinating feature of some systems is that not all the information stored in DNA is encoded simply within DNA. Some must be processed to remove unnecessary packets and some even require editing and correction before the genetic code can be deciphered. I will examine what we know about how such systems evolve, and present results showing that we can evolve such systems in the laboratory.

3. How might DNA storage systems evolve in the digital era?
Humans are now generating enormous amounts of information that may soon rival the information stored as DNA in the biosphere. However, our approach to digital data storage has been more functional than archival. Some have noted that, because of digital obsolescence, future historians will know more about the beginning of the 20th century than the beginning of the 21st century. Thus, there are major challenges for the long-term storage of digital data. In contrast, genetic information is incredibly long-lived. In the last part of my talk, I will consider how we can create biology-inspired DNA storage systems that could be readable over vastly longer timescales than our present stems.