Programs, Talks & Poster Abstracts Bundle

bilateral network of
Heidelberg University
Kyoto University
Karlsruhe Institute of Tech.
Tohoku University
Goettingen University
Osaka University

/Description

The German-Japanese University Network (HeKKSaGOn) was established in 2010. Member university president conferences have been alternately held every year and a half to discuss student/researcher exchange, joint programs, and more.

presented by

HeKKSaGOn
Heidelberg | Kyoto | Karlsruhe | Sendai | Göttingen | Osaka

as Satellite Event of The 8th German-Japan Annual Presidential Conference

Data Science Workshop

7-8 SEP 2021

via ZOOM

Talks & Poster Presentations

/Hosted by
Graduate School of Information Sciences
TOHOKU UNIVERSITY, Japan

HeKKSaGOn
Heidelberg | Kyoto | Karlsruhe | Sendai | Göttingen | Osaka

Data Science Workshop

Talks & Poster Presentations

7-8 SEP **2021**

Schedule Overview

	🇩🇪	🇯🇵	
SEP 7	08:30 08:40	15:30 15:40	**Opening Remarks** Prof. Mitsuyuki Nakao
		TOPIC	**Data Science and Computing Resource Integration** Chair: Prof. Shinji Shimojo
	08:40 09:10	15:40 16:10	**How to Integrate HPC Resources in Different Workflow Environments** Prof. Sven Bingert
	09:10 09:40	16:10 16:40	**Dynamic Compute Resource Integration for Collaborative Scientific Analyses** Prof. Max Fischer
	09:40 10:10	16:40 17:10	**Learning and Evidence Analytics Framework (LEAF)** **Educational Data Science with Multimodal Learning Traces.** Prof. Rwitajit Majumdar
	10:10 11:10	17:10 18:10	**Poster Session** Presenters of odd-numbered posters are required to attend
SEP 8		TOPIC	**Data Science for Studies using Large Scale Science Infrastructure** Chair: Prof. Mitsuyuki Nakao
	08:30 09:10	15:30 16:10	**Machine Learning Applications for Particle Accelerators** Prof. Andrea Santamaria Garcia
	09:10 09:50	16:10 16:50	**Application of the Machine Learning to the Collider Experiments** Prof. Masako Iwasaki
	09:50 10:30	16:50 17:30	**Collaboration between X-ray Ptychography and Data Science** **or the Use of Next-generation Synchrotron Radiation** Prof. Yukio Takahashi
	10:30 11:30	17:30 18:30	**Poster Session** Presenters of even-numbered posters are required to attend.
	11:30 11:40	18:30 18:40	**Closing Remarks** Prof. Ramin Yahyapour

🇩🇪 = CEST (GMT +02.00) 🇯🇵 = JST (GMT +09.00)

Talk Abstracts

Day 1

Talk 01

How to integrate HPC resources in different workflow environments

Sven Bingert[1], Christian Köhler[1], Hendrik Nolte[1]

[1]Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen, Germany

Keywords data analysis, workflows, high-performance computing, data science

The demand for computational resources to conduct data analysis or data science pipelines is constantly increasing. Not only the amount of data but also the complexity of the algorithms used require to have access to a more powerful infrastructure. High-Performance Computation (HPC) systems do offer the required resources but are often limited in the access possibilities, and thus are difficult to be included in workflows. In our presentation, we would like to present a generic API that allows connecting different workflow systems, e.g., git or flowable, to HPC systems. HPCSera is an implementation of this API and is running as a service at GWDG. HPC jobs can be submitted via the REST interface and are eventually submitted to the specific queuing system of the HPC environment. In the presentation, we will show examples of a productive environment.

Day
1

Talk
02

Dynamic integration of opportunistic resources into large scale scientific computing infrastructure

Max Fischer[1], Eileen Kuehn[1], Manuel Giffels[1], Matthias Schnepf[1]

[1]Steinbuch Centre for Computing, Karlsruhe Institute of Technology

Keywords grid, cloud, opportunistic computing

Modern scientific research fields increasingly rely on massive computing infrastructure to analyze ever-increasing data sets. As a scientific field that naturally combines large data sets and international collaborations, the field of High Energy Physics has historically been a pioneer in both discovering and approaching challenges of massive large-scale scientific computing. While this has given rise to many innovative solutions for large-scale collaborative data analysis, much of these have been specific to the infrastructure and protocols employed by High Energy Physics only. The COBalD/TARDIS project aims at making one of the key pillars of High EnergyPhysics available to other fields as well: the usage of many distinct resource providers via so-called Overlay Batch Systems. By expanding on the strengths of the approach as well as reworking the key shortcomings of existing frameworks, COBalD/TARDIS is relevant both for High Energy Physics and any other scientific field or group desiring to do large-scale scientific computing. The presentation visits the key insights that High Energy Physics has gained for large scale scientific computing, shows how this knowledge can be applied for scientific computing in general and provides a practical overview of how COBalD/TARDIS combines the two

Day 1

Talk 03

Learning and Evidence Analytics Framework (LEAF) - Educational Data Science with Multimodal Learning Traces

Rwitajit Majumdar[1], Brendan Flanagan[1], Hiroaki Ogata[1]

[1]Learning and Educational Technologies Research Unit, Academic Center for Computing and Media Studies, Kyoto University, Kyoto, Japan

Keywords Learning and Evidence Analytics Framework (LEAF), Educational data science, Learning Analytics, BookRoll, GOAL

Learning and Evidence Analytics Framework (LEAF), is an integrated framework that is co-designed with the stakeholders from schools and universities in Japan, led by Kyoto University's Learning Research Unit. The platform links learner's data from multiple sources such as LMS, e-book readers, and wearable devices to provide an educational big data pool and design various AI-driven services for the end-users. In this talk, I shall demonstrate the data-driven infrastructure and the learning analytics research that this framework enables. Current large-scale implementation of the system spans multiple institutions, including public schools at a compulsory education level and universities in Japan. We will share our research findings, learnings, and challenges faced from the past 4years of the project and also highlight the advantage of LEAF during this emergency remote teaching period of the pandemic at Kyoto University.

Day 2

Talk 01

Machine learning applications for particle accelerators

Andrea Santamaria Garcia[1], Erik Bründermann[2]

[1]Laboratory for Applications of Synchrotron Radiation (LAS), Karlsruhe Institute of Technology,
[2]Institute for Beam Physics and Technology, Karlsruhe Institute of Technology

Keywords particle accelerators, machine learning

The Institute for Beam Physics and Technology (IBPT) at the Karlsruhe Institute of Technology (KIT) hosts two research accelerator facilities, KARA and FLUTE, that serve as platforms for the development and testing of new beam acceleration technologies and new cutting-edge accelerator concepts, including Machine Learning (ML) methods. In this talk I will present three ML activities in accelerator physics carried out at KIT:

- Real-Time Control of the Micro-Bunching Instability with Reinforcement Learning
- Bayesian Optimization of the Injection Efficiency
- Machine Learning Towards Autonomous Accelerators: Control of the Bunch Profile with Reinforcement Learning

Additionally, we will introduce the action plan "ErUM-Data - from Big Data to Smart Data", a funding scheme of the German Federal Ministry of Education and Research to support the digital transformation in basic research and natural sciences, and more specifically the handling of large data volumes in large-scale infrastructures.

Day 2

Talk 02

Application of the Machine Learning to the Collider Experiments

Masako Iwasaki[1234], Hajime Nagahara[4], Yuta Nakashima[4], Noriko Takemura[4], Takashi Nakano[34], Taikan Suehara[5]

[1]Osaka-City University Graduate School of Science,
[2]Nambu Yoichiro Institute of Theoretical and Experimental Physics (NITEP),
[3]Research Center for Nuclear Physics, Osaka University (RCNP)
[4]Osaka University Institute for Datability Science (IDS)
[5]Kyushu University Graduate School of Science

Keywords Elementary Particle Physics Experiment, Machine Learning Application, Accelerator control, Data Analysis

In this talk, we show the results of our R&D works on the Machine Learning application to the collider experiments. One of the important approaches of the recent elementary particle physics experiments is the precise measurement based on the high statistics data to probe the new physics phenomena beyond the standard model. “Big-data” processing becomes the important key for such high statistics colliding experiments. The modern machine learning techniques developed in information science, are expected to be powerful tools to provide precise and efficient data processing in colliding experiments. As research projects in IDS and RCNP, Osaka University, we form a group that consists of about 20 researchers on both information science and collider physics (experiment and theory) to work on the R&D of machine learning application to the collider experiments. The R&D works are related to data analysis, detector calibration, accelerator control, and lattice QCD, etc. In this talk, we'll introduce the recent activities on the machine learning applications based on the low-level feature data (physics analysis, and detector calibration)and accelerator control using the machine learning.

Day 2

Talk 03

Collaboration between X-ray ptychography and data science for the use of next-generation synchrotron radiation

Yukio Takahashi

International Center for Synchrotron Radiation Innovation Smart (SRIS), TohokuUniversity

Keywords Synchrotron radiation, X-ray ptychography, Visualization, Datascience

X-ray ptychography is a rapidly emerging technique at synchrotron facilities, which provides three-dimensional imaging big data of the structure and chemical state of materials. So far, we have developed the techniques for high-resolution X-ray ptychography at SPring-8, which is the third-generation synchrotron radiation facility in Japan, and applied them to the observation of the various samples. In this talk, I will introduce the results of visualizing the chemical state of three-way catalyst particles using X-ray ptychography and revealing the oxygen-diffusion-driven oxidation behavior using unsupervised learning. Finally, I will present a perspective on the collaboration between X-ray ptychography and data science for the use of the next-generation 3GeV synchrotron radiation facility in Japan, which is scheduled to start operation in 2023.

Poster Abstracts

P01

Large-scale Analysis of The Honey Bee Brain Using Micro-CT Imaging and Deep Learning

Philipp D. Lösel[123], Coline Monchanin[23], Mathieu Lihoreau[23], Vincent Heuveline[1]

[1]Engineering Mathematics and Computing Lab (EMCL), Interdisciplinary Center for Scientific Computing (IWR), Heidelberg University,
[2]Research Center on Animal Cognition (CRCA), Center for Integrative Biology (CBI); CNRS, University Paul Sabatier – Toulouse III, France,
[3]Department of Biological Sciences, Macquarie University, NSW, Australia

Keywords image segmentation, large-scale analyses, honey bee brain, deep learning

The analysis of volumetric medical and biological imaging data often requires isolating individual structures from the 3D volume by segmentation. In insects, analysis of large numbers of samples can reveal minor but statistically and biologically relevant variations in brain morphology and lateralization addressing major issues related to behaviour, ecology and evolution.
However, the manual effort of conventional methods (e.g. histology, scanning electron or confocal laser scanning microscopy in combination with manual segmentation) limits the number of samples required for a large-scale analysis. Here we use micro-CT combined with automated 3D reconstruction to analyse the neuro-architecture of 110 honey bees. The reconstruction is achieved with Biomedisa, an online segmentation platform that utilizes deep neural networks for automated image segmentation. We analyse the inter-individual variability of brain morphologies and lateralization in honey bees and describe architectural asymmetries. In particular, we found a subsequent lateralization of antennal lobes and lobulae that may explain behavioral lateralizations previously reported for olfactory and visual learning. Overall, Biomedisa enables easy access to large-scale quantitative comparative analyses. The platform is accessible via a web browser and does not require complex and tedious configuration of software and model parameters, thus addressing the needs of scientists without substantial computational expertise.

P02

Understanding and Visualizing Deep Face Recognition

Hiroya Kawai*, Takashi Kozu*, Koichi Ito*, Hwann-Tzong Chen**, Takafumi Aoki*

*Tohoku University, Graduate School of Information Sciences
**National Tsing Hua University, Department of Computer Science

Keywords Face Recognition, Face Parsing, CNN, Biometrics

The performance of face recognition has been extremely improved by deep learning techniques such as Convolutional Neural Networks (CNNs), although there is a problem of difficulty in interpreting the results. The general visualization methods used to interpret recognition results are specialized for object recognition models and are not necessarily effective for face recognition models. We propose a novel visualization method for face recognition models using face parsing. Images containing only one facial part are generated from the semantic labels obtained from the facial part segmentation, and the importance of each part in inference is visualized. The visualization results provide more intuitively interpretation than the results using the general visualization method for CNN.

P03

A Provenance Aware Data Lake for Sciences

Hendrik Nolte[1], Piotr Kasprzak[1], Sven Bingert[1], Julian Kunkel[1], Philipp Wieder[1],Ramin Yahyapour[1]

[1]Gesellschaft für Wissenschaftliche Datenverarbeitung mbH Göttingen

Keywords Data Lake, Provenance Auditing, FAIR Digital Objects

Across various domains, data lakes are successfully utilized to centrally store all data of an organization in their raw format. This promises high reusability of the stored data since a schema is implied on reading, which prevents an information loss due to ETL processes. Despite this schema-on-read approach, some modeling is mandatory to ensure proper data integration, comprehensibility, and quality. These data models are maintained within a central data catalog which can be queried. To further organize the data in the data lake, different architectures have been proposed, like the most widely-known zone architecture. Here, data is assigned to different zones according to the processing they were subjected to. In this work, we present a novel data lake architecture based on FAIR Digital Objects (FDO) with (high-performance)processing capabilities. The FAIR Digital Objects are connected by a provenance-centered graph. Users can define generic workflows, which are reproducible by design, making this data lake implementation ideally suited for science

P04 System Architecture for the integration, processing, and publishing of many heterogenous text resources

Triet Ho Anh Doan[1], Sven Bingert[1], Ramin Yahyapour[1]

[1]Gesellschaft für Wissenschaftliche Datenverarbeitung mbH Göttingen

Keywords workflow; text analysis; knowledge graph; persistent identifier; high-performance computing

Göttingen University Library, founded in 1734, is one of the five largest libraries in Germany. The library has a huge amount of text resources stored in different repositories, such as GDZ, Goescholar, and Textgrid. Since the resources are scattered in many repositories, it is challenging for scientists to access them. In addition, the access to raw data is still managed manually. These problems make it impossible to build a complete automatic text analysis workflow because the data gathering process cannot be automated. For that reason, a service is developed to overcome those difficulties. It offers various features to users, such as full-text search, raw data acquisition, knowledge graph exploration, and data analysis on High-Performance Cluster. Behind the scene, this service collects data from various repositories, processes them, and offer them to users. Additionally, it helps users to easily run their analyses on the HPC. Currently, there are three ways to interact with the service: using the webinterface, the REST API, and the Python client.

P05 Group Learning Orchestration Based on Evidence (GLOBE) framework and its implementations

Changhao Liang[12], Rwitajit Majumdar[2], Hiroaki Ogata[2]

[1]Graduate School of Informatics, Kyoto University
[2]Ogata Laboratory, AcademicCenter for Computing and Media Studies, Kyoto University

Keywords group learning, collaborative learning, group formation, peer evaluation, CSCL

Collaborative and interpersonal skills are vital in modern society and hence group learning receives increasing attention in various pedagogical contexts. Considering group learning in the digital environment, learning logs created there provides immense opportunities to apply Learning Analytics (LA)approaches and support such educational activities. We put forward the GroupLearning Orchestration Based on Evidence (GLOBE) framework for such a data-rich environment and focus on using data and developing algorithms for four stages of group learning: formation, orchestration, evaluation, and reflection. Agenetic algorithm-based group formation function and a peer evaluation module are introduced as implementations of GLOBE. We conducted empirical studies at school and university classes and report some of our current findings regarding the impact of the use of the system.

P06 Look Wide and Interpret Twice: Improving Performance on Interactive Instruction-following Tasks

Van-Quang Nguyen[1], Masanori Suganuma[1], Takayuki Okatani[1]

[1]Graduate School of Information Sciences, Tohoku University
[2]RIKEN Center for Advanced Intelligence Project,

Keywords Vision-language, AFRED, Interactive Instruction-following Tasks

There is a growing interest in the community in making an embodied AI agent perform a complicated task while interacting with an environment following natural language directives. Recent studies have tackled the problem using ALFRED, a well-designed dataset for the task, but achieved only very low accuracy. This paper proposes a new method, which outperforms the previous methods by a large margin. It is based on a combination of several new ideas. One is a two-stage interpretation of the provided instructions. The method first selects and interprets an instruction without using visual information, yielding a tentative action sequence prediction. It then integrates the prediction with the visual information., yielding the final prediction of action and an object. As the object's class to interact is identified in the first stage, it can accurately select the correct object from the input image. Moreover, our method considers multiple egocentric views of the environment and extracts essential information by applying hierarchical attention conditioned on the current instruction. This contributes to the accurate prediction of actions for navigation. A preliminary version of the method won the ALFRED Challenge 2020.

P07 A Method for Reducing Time-to-Solution in Quantum Annealing Through Pausing

Michael R. Zielewski[1], Hiroyuki Takizawa[1]

[1]Graduate School of Information Sciences, Tohoku University, Sendai, Japan

Keywords quantum annealing, D-Wave, annealing schedule, pausing

Recent research has shown that alternative annealing schedules provide the means for improving performance in modern quantum annealing devices. One such type of schedule is forward annealing with a pause, in which there is a period of time when system evolution is paused. While the results from using this type of schedule have been promising, effectively using a pause is not a trivial task. One challenge associated with introducing a pause into the schedule is determining the point in the anneal at which the pause will start. Additionally, tuning the schedule in real-time requires a significant amount of time. A second challenge is that while a pause may increase the number of correct solutions returned from the annealer, the time-to-solution, a standard metric for measuring performance in quantum annealing, will not necessarily be improved. We propose a method for constructing annealing schedules containing a pause that avoids the costly process of determining the optimal pause location in an online manner. We also evaluate our method on the subset sum problem, a problem of practical significance, and show that our method is able to achieve a 70% reduction in time-to-solution from a standard schedule containing no pause.

P09

Effects of Habitual Sleep/Wake Pattern and Menstrual Cycle on Subjective SleepQuality and Heart Rate Variability Dynamics

Siwalee Choilek[1], Akihiro Karashima[2], Ikuko Motoike[13], Norihiro Katayama[24], Kengo Kinoshita[13], Mitsuyuki Nakao[1]

[1]Graduate School of Information Sciences, Tohoku University [2]Tohoku Institute of Technology
[3]Tohoku University Tohoku Medical Megabank Organization
[4]Department of Humanities and Social Studies, Shokei Gakuin University

Keywords Menstrual cycle, Sleep quality, Sleep/wake pattern, Heart rate variability

It has been widely known that the menstrual cycle modulates various hormones and reproductive systems which possibly affect various body functions such as temperature regulations, cardiovascular system, and sleep. Meanwhile, individuals also have a preference in sleep/wake timing. When such a preference was constrained by social schedule, sleep quantities changed, and sleep qualities significantly degraded. This study, by combining 2 aspects, aims to clarify whether a menstrual phase-modulated biological signal (through heart rate and its variabilities), hormonal variation, and sleep, were further amplified (or suppressed) by habitual sleep/wake patterns (HSWP) preference.
In this experiment, 82 Japanese female college students (age: 22.28±1.89 years, BMI: 20.55 ± 2.07) with no severe sleep complaints were requested to record their physical activities and heart rate by using wearable sensors. Subjects were also requested to evaluate daily subjective sleep quality (SSQ) and recorded their basal body temperature and collected saliva samples for evaluating the menstrual cycle phase. HSWP, which generally reflects sleep timing and regularity was estimated. Associations between SSQ and level of hormones, followed by evaluation of the causal relationship between HSWP, menstrual cycle, heart rate variability (HRV), and SSQ were obtained. Aside from the well-known menstrual phase-dependency of SSQ, the result also suggested that HRV parameters during sleep onset were modulated by HSWP. In the model predicting SSQ based on HR obtained during the initial 3 hours of sleep and HSWP, significant effects were both parameters played an important role in modulating the subsequent SSQ.

P10

Predicting Children's Behavior Problems with Toy Block Play Actions and Patterns

Xiyue Wang[1], Kazuki Takashima[1], Tomoaki Adachi[2], Yoshifumi Kitamura[1]

[1]Research Institute of Electrical Communication, Tohoku University,
[2]Department of Education, Miyagi Gakuin Women's University

Keywords Free play, CBCL, Motion data, Well-being

Although children's behavioral and mental problems are generally diagnosed in clinical settings, the prediction and awareness of children's mental wellness in daily settings are getting increased attention. Toy blocks are both accessible in most children's daily lives and provide physicality as a unique non-verbal channel to express their inner world. In this paper, we propose a toy block approach for predicting a range of behavior problems in young children (4-6 years old) measured by the Child Behavior Checklist (CBCL). We defined and classified a set of quantitative play actions from IMU-embedded toy blocks. Play data collected from 78 preschoolers revealed that specific play actions and patterns indicate total problems, internalizing problems, and aggressive behavior in children. The results align with our qualitative observations and suggest the potential of predicting the clinical behavior problems of children based on short free-play sessions with sensor-embedded toy blocks.