USI/ICS Group „Computational Time Series Analysis“

Prof. Dr. Illia Horenko
Institute of Computational Science
Faculty of Informatics
Via Giuseppe Buffi 13
6900 Lugano
Switzerland

Dr. Lukas Pospisil

Scalable Probabilistic Approximation (SPA) framework: theory and applications

Typical components of data-processing pipelines in the machine learning are characterized by computational costs that grow quickly with the increasing data complexity and size – accompanied by a very moderate improvement of performance for the increasing model complexity. We introduce a low-cost improved-quality Scalable Probabilistic Approximation (SPA) framework, allowing for a simultaneous joint optimal solution of data-driven feature selection, discretization and Baysian/Markovian model inference problems. After introducing the theoretical foundations, we will show applications to a broad range of application problems from natural sciences – as well as comparisons with common tools from statistics and deep learning. We will show cross-validated applications of SPA to a range of large realistic data classification and prediction problems - revealing drastic cost and performance improvements. For example, we will show that SPA allows the unsupervised next-day surface temperature predictions for Europe with the mean cross-validated one-day prediction error of 0.75°C on a common PC (being around 40% better in terms of errors and five to six orders-of-magnitude cheaper than the next-day surface temperature predictions calculated on supercomputers and provided by the weather services). Theoretical lectures will be followed by the practical hands-on sessions with the parallel MATLAB implementation of the SPA toolbox.

Biomolecular simulations at the Max Planck Institute for Polymer Research

Dr. Tristan Bereau
Project Leader Emmy Noether Group
(MPIP Mainz)
Ackermannweg 10,
55128 Mainz,
Germany

Kernel-based machine learning for molecular simulations

Compared to deep learning, kernel machines operate remarkably well at small regimes of training data. This makes kernels well suited to many problems in the natural sciences, where data is often scarce. In this overview, I will first motivate kernels from Bayesian inference. The explicit definition of the representation will be emphasized, as well as the inclusion of physical symmetries and conservation laws. Examples in the context of molecular simulations will be brought forward. The lecture will be followed by a hands-on session.

Machine learning and molecular dynamics

Univ.-Prov. Dr. Dr. hc. mult.
Michele Parrinello

Department of Chemistry and Applied Biosciences, ETH Zurich, and Facoltà di Informatica, Istituto di Scienze Computazionali, Università della Svizzera italiana, 6900 Lugano, Switzerland

Atom based computer simulation is one of the most important tools of contemporary physical chemistry. In spite of its many successes, it suffers from severe limitations. Here we show how machine-learning techniques can help in solving at least two different problems. The first one is the accuracy of current interatomic potential models; the second is the limited time scale that simulations can explore. In order to solve the first problem we train a neural network on a set of accurate but expensive quantum chemical calculations. In this way, it is possible to obtain an accurate description of the system at a relatively low computational cost. Crucial for the success of this program has been the design of the neural work and the selection of the training set. We apply this approach to study a metal non-metal transition and to chemical reactions in condensed phases. These applications would not have been possible without the use of efficient sampling methods capable of lifting the time scale barrier. To this effect, we have developed two very efficient sampling methods, metadynamics and variationally enhanced sampling. Both methods are based on the identification of appropriate collective variables, or slow modes, whose sampling needs to be accelerated. Machine learning can be used also for the construction of efficient collective variables based on a modification of the well-known linear discriminant analysis classification method. Finally, we use the variational enhanced sampling approach and a deep neural network to further increase our sampling ability.

Training Topic Models on Streaming Text Data

Zahra Ahmadi
Johannes Gutenberg - Universität Mainz
Institut für Informatik
Staudingerweg 9
55128 Mainz, Germany

Dr. Sophie Burkhardt
Johannes Gutenberg – Universität Mainz
Institut für Informatik
Staudingerweg 9
55128 Mainz, Germany

Text data in the form of news articles, reports, emails and many more is created daily. To get an overview of large collections of texts, topic models are a popular method. Latent Dirichlet Allocation (LDA), a generative Bayesian model for text documents, is among the most successful methods, that has found many different applications and extensions to date. We will introduce a simple but effective method for training these models online to process the data in a stream, thus, enabling to observe changing topics over time or detect new trends. In this workshop, we will demonstrate the effectiveness of the method on a sample corpus of news data. Furthermore, we introduce different training methods for LDA based on sampling and variational Bayesian inference. We will also introduce different ways to train these models online on streaming data, discussing challenges such as concept drifts and the emergence of new trends. The talk will be followed by a practical session using python.

Introduction to machine learning

Univ.-Prof. Dr. Michael Wand

Institut für Informatik
Staudinger Weg 9
55128 Mainz

This talk gives as a short overview and recap of the basics of statistical and machine learning. It is intended as an introductory session to get everybody on the same page. It will recap some basic probability theory, summarize the key ingredients of machine learning algorithms, and discuss some fundamental results and algorithms in this area. We will also take a brief look at recent methods based on "deep" neural networks for representation learning.

Agent-based modelling

Agent-based modelling is a technique for describing and simulating social behaviour of populations in order to advance the understanding of the dynamics of interest. The general approach is to describe the individual behaviour of the so-called agents (discrete units such as humans, cars, organizations) in a rule-based fashion instead of the direct modelling of variables of interest on the global population scale, or in other words: bottom-up modelling instead of top-down modelling. Agents act according to certain rules such that their behaviour mimics the behaviour of their counterparts in the real-world system of interest. The agents' rules can be given by very different, potentially complex and even stochastic mathematical equations.

The resulting simulations of the agents' behaviour often showcase emergent patterns on the global population scale, i.e. patterns resulting from the local-scale interactions among the agents and their environment.

We will give an introduction as to why, where and how agent-based modelling is used. An example where we will show this in detail is the modelling of the spreading of the woolly sheep around 6000BC which enables us to get an idea of how the woolly sheep spread over Europa and West Asia and hence how people got in contact with the benefit of the use of textiles.

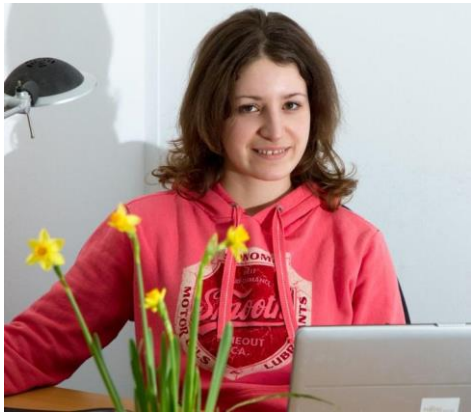
Luzie Helfmann
Department of Mathematics and Computer Science, Mathematics Institute, Biocomputing Group, Berlin

Niklas Wulkow
Department of Mathematics and Computer Science, Mathematics Institute, Biocomputing Group, Berlin

Computational Proteomics

Prof. Dr. Tim Conrad

Mathematics Institute
AG "Computational Proteomics"
Assistant Professor,
Freie Universität Berlin



Dr. Nataša Djurdjevac Conrad

Mathematics for Life and Materials Sciences,
Visual Data Analysis,
Bioinformatics in Medicine,
Zuse Institute Berlin

High throughput techniques are of great significance in many areas of biology, such as genomics or proteomics. The amount of available very large high-dimensional data-sets is increasing steadily, thanks to technological advances. However, analysing this type of data is still a challanging task, due to various reasons. In this session we will present some of these challenges and typical approaches for the analysis of omics data, based on current machine learning and networks analysis tools.

Message-passing neural networks for modeling many-particle systems

Dr. Jan Hermann

Freie Universität Berlin
Arnimallee 6,
Room 209,
14195 Berlin

Machine learning has been recently applied to many problems in quantum chemistry and condensed-matter physics to model systems consisting of many particles, ranging in size from electrons, to atoms, to coarse-grained models. The so-called message-passing neural networks [1], a special case of graph neural networks, are especially fitting to represent the interactions between particles in complex quantum systems. I will introduce the general message-passing approach, its particular implementation in the form of the SchNet architecture [2], and two practical applications of SchNet: (1) an atomic force field for organic molecules, and (2) an electronic molecular wave function. The lecture will be followed by a hands-on tutorial in which we will reproduce the two applications in a Jupyter notebook.

[1] Gilmer et al., "Neural Message Passing for Quantum Chemistry", https://arxiv.org/abs/1704.01212
[2] Schütt et al., "SchNet - a deep learning architecture for molecules and materials", https://arxiv.org/abs/1712.06113