3rd Data Science Doctoral Retreat

within the

DSC 2025

May 28th 2025, Salzburg University of Applied Sciences / Schloss Urstein

The Data Science Doctoral Retreat was initiated by the Lab for Intelligent Data Analytics (IDALab), which is a collaborative initiative involving the Paris Lodron University of Salzburg, the Paracelsus Medical University, the Salzburg Research Forschungsgesellschaft, and the Salzburg University of Applied Sciences.
Funded by the Federal State Government of Salzburg

This year, the retreat is organized by the Salzburg University of Applied Sciences as part of the Interdisciplinary Data Science Conference.

Lab for Intelligent
Data Analytics Salzburg

3rd Data Science Doctoral Retreat

iDSC'25 May 28th Program

Moderation: FH-Prof. DI Dr. Stefan Huber, MSc

09:00–10:30 Stochastic Dependence Modeling

A new coefficient of separation
Carsten Limbach, Patrick Langthaler (University of Salzburg)

Comparison of adaptive Wasserstein distances in two-time step models
Stefanie Steinmaßl (University of Salzburg)

R Package: Directed Dependence Coefficient
Yuping Wang (University of Salzburg)

10:30–11:00 Kaffeepause

11:00–12:30 Reinforcement Learning & Symbolische Repräsentationen

Of Mountain Car Problems and Chebyshev Policies
Hannes Waclawek (Salzburg University of Applied Sciences)

Recent Advancements in Reinforcement Learning for the Quanser Aero 2 System
Georg Schäfer (Salzburg University of Applied Sciences / University of Salzburg)

Multi-Agent Reinforcement Learning to achieve reliable wireless communication
Sabrina Pochaba (Salzburg Research, University of Salzburg)

12:30–13:30 Mittagspause

13:30–15:00 Time-Series

From Load Profiles to Life Patterns
Dejan Radovanovic (Salzburg University of Applied Sciences)

Anomaly Detection for Multivariate Time Series in Industrial Applications
Martin Uray (Salzburg University of Applied Sciences)

Deep Learning-based Time Series Forecasting for Industrial Discrete Process Data
Olaf Sassnick (Salzburg University of Applied Sciences / University of Salzburg)

15:00–15:30 Kaffeepause

15:30–17:00 Sequential Data and Language

LLM-KABOOM: Blasting Unstructured Web Data into Knowledge Graphs by using LLMs
Jasmin Saxer (Zurich University of Applied Sciences)

Sleep staging and apnea detection using low-cost wearables and deep learning
Sebastian Baron (University of Salzburg)

Profile Generators: A Link between the Narrative and the Binary Matrix Representation
Raoul Kutil (Paracelsus Medical University, University of Salzburg, IDA-Lab)

18:00 Gasthof Kirchenwirt

A new coefficient of separation

March 27, 2025

Abstract

A coefficient is introduced that quantifies the extent of separation of a random variable Y relative to a number of variables $\mathbf{X} = (X_1, \dots, X_p)$ by skillfully assessing the sensitivity of the relative effects of the conditional distributions. The coefficient is as simple as classical dependence coefficients such as Kendall's tau, also requires no distributional assumptions, and consistently estimates an intuitive and easily interpretable measure, which is 0 if and only if Y is stochastically comparable relative to $\mathbf{X}$, that is, the values of Y show no location effect relative to $\mathbf{X}$, and 1 if and only if Y is completely separated relative to $\mathbf{X}$. As a true generalization of the classical relative effect, in applications such as medicine and the social sciences the coefficient facilitates comparing the distributions of any number of treatment groups or categories. It hence avoids the sometimes artificial grouping of variable values such as patient's age into just a few categories, which is known to cause inaccuracy and bias in the data analysis. The mentioned benefits are exemplified using synthetic and real data sets.

Keywords: conditional distributions, complete separation, relative effect, sensitivity, stochastic comparability

Authors: Fuchs Sebastian, Limbach Carsten, Langthaler Patrick

Presenters: Limbach Carsten, Langthaler Patrick

Affiliation: Paris Lodron Universität Salzburg, IDA Lab

Comparison of adaptive Wasserstein distances in two-time step models

Stefanie Steinmaßl
Department of Artificial Intelligence Human Interfaces,
University of Salzburg,

This is joint work with Jonathan Ansari. The Wasserstein distance between two distributions is an optimal transport-based metric that quantifies the minimal cost required to transform one probability distribution into another. For stochastic processes, a modified version—the adaptive Wasserstein distance—addresses the time dependency of the underlying random variables by restricting the admissible couplings to those adapted to the processes' current filtration. Under a stochastic monotonicity condition on the underlying processes, the conditionally comonotone Knothe-Rosenblatt rearrangement proves optimal for the adaptive Wasserstein distance. However, aside from the multivariate normal distribution, its behavior has not been studied in the literature. For two-time-step models, the conditionally comonotone dependence structure is described by the recently investigated upper product of bivariate copulas. By incorporating a novel sign-change ordering for upper products, we derive monotonicity results that allow a comparison of bivariate distributions with respect to the adaptive Wasserstein distance. Various diagrams illustrate the key aspects of our findings.

R Package: Directed Dependence Coefficient

Yuping Wang*

Abstract

The Directed Dependence Coefficient (`didec`) is a dependence measure which estimates the degree of directed dependence of a random vector $\boldsymbol{Y}$ on a random vector $\boldsymbol{X}$, based on an i.i.d. sample of $(\boldsymbol{X},\boldsymbol{Y})$. Two applications of `didec` are included in the R package: a forward feature selection algorithm for multiple-outcome data `mfoci` (Multivariate Feature Ordering by Conditional Independence) and a dependence-based hierarchical clustering method for variables `VarClustPartition`. Simulation results and real case studies are used to demonstrate the broad applicability and excellent performance of the proposed data analysis tools.

Keywords: directed dependence; (mutual) perfect dependence; predictability; feature selection; agglomerative hierarchical variable clustering; dissimilarity

References

[1] Y. Wang, S. Fuchs and J. Ansari (2024). didec: Directed Dependence Coefficient. R package version 0.1.0. CRAN.R-project.org/package=didec.

*Department for Artificial Intelligence & Human Interfaces, University of Salzburg, Hellbrunner Strasse 34, 5020, Salzburg, Austria, `yuping.wang@plus.ac.at`

Of Mountain Car Problems and Chebyshev Policies

Polynomial Models in Reinforcement Learning

Hannes Waclawek

Josef Ressel Centre for Intelligent and Secure Industrial Automation
Salzburg University of Applied Sciences, Austria
hannes.waclawek@fh-salzburg.ac.at

Due to its close ties to control theory, Reinforcement Learning (RL) is an interesting candidate for utility in Mechatronics, however, approximate RL solution methods mostly rely on Neural Networks (NNs) as approximators, which lack explainability. When training and utilizing NNs, we arguably pass control onto a black box and receive results without knowing the exact inner workings (exact model) of the network. In Mechatronics, however, non-transparency to this extent is not always a desired approach and often the resulting model description is of interest. This is why, in the context of this doctoral research project, we investigate how optimizers of modern machine learning frameworks can be utilized directly, outside of the scope of NNs, in order to optimize polynomial models. This allows for an optimization using state of the art methods, while at the same time working with an explainable model.

In our current research, we train Chebyshev policies of degree 3 on the mountain car problem, a classical benchmark problem in RL. We derive the optimal solutions of this control problem to answer the question of how well state of the art agents actually perform in comparison to our polynomial policies trained using classical REINFORCE. In this talk, we discuss our findings and show how our approach aligns with the state of the art in this regard.

Keywords: Machine Learning · Reinforcement Learning · Mountain Car Problem · Polynomials · Linear Policies · Gradient Descent Optimizers · Mechatronics

Recent Advancements in Reinforcement Learning for the Quanser Aero 2 System

Georg Schäfer[1,2,3]

[1] Josef Ressel Centre for Intelligent and Secure Industrial Automation
[2] Salzburg University of Applied Sciences
[3] Paris Lodron University of Salzburg
georg.schaefer@fh-salzburg.ac.at

Abstract

Reinforcement learning (RL) offers a promising alternative to classical control strategies in industrial cyber-physical systems. In our previous work, we compared RL to model predictive control (MPC) and linear-quadratic regulators (LQR) on the Quanser Aero 2 testbed, revealing that the performance of RL algorithms is highly sensitive to the design of the problem formulation [1]. Our recent study systematically investigated how modifications in state representation, reward design, and training protocols – such as observation normalization, target randomization, and extended episode durations – can significantly enhance learning stability, sample efficiency, and overall policy performance [2]. These findings underscore that careful problem formulation is a critical first step toward establishing a robust RL engineering pipeline for real-world applications.

In addition, we briefly explore a multi-objective RL framework that extends this work by simultaneously optimizing pitch tracking and energy efficiency. Preliminary results suggest that integrating an energy penalty into the reward structure can effectively balance control performance with power consumption. This integrated approach lays the foundation for future research aimed at deploying RL in practical, energy-constrained industrial environments.

References

1. G. Schäfer, J. Rehrl, S. Huber, and S. Hirlaender, "Comparison of model predictive control and proximal policy optimization for a 1-dof helicopter system," in *2024 IEEE 22nd International Conference on Industrial Informatics (INDIN)*. IEEE, 2024, pp. 1–7.
2. G. Schäfer, T. Krau, J. Rehrl, S. Huber, and S. Hirlaender, "The Crucial Role of Problem Formulation in Real-World Reinforcement Learning," in *2025 IEEE 8th International Conference on Industrial Cyber-Physical Systems (ICPS)*. Emden, Germany: IEEE, May 2025.

Multi-Agent Reinforcement Learning to achieve reliable wireless communication

Sabrina Pochaba[1,2], Peter Dorfinger[1], Roland Kwitt[2], Simon Hirländer[2]
[1]Intelligent Connectivity, Salzburg Research Forschungsgesellschaft mbH, Salzburg, Austria
[2] Department of Artificial Intelligence and Human Interfaces, University of Salzburg, Austria
Corresponding author: Sabrina Pochaba (email: sabrina.pochaba@salzburgresearch.at)

Reinforcement learning (RL) is gaining more and more importance in the field of machine learning (ML). One subfield of RL is Multi-Agent RL (MARL). Here, several agents learn to solve a problem simultaneously rather than a single agent. For this reason, this approach is suitable for many real-world problems. Since learning in a multiple agent scenario is highly complex, further conflicts can arise in addition to the difficulties in single-agent RL. These are, for example, scalability problems, non-stationarity or ambiguous learning goals.

To explore the difficulties of MARL, we have implemented the environment for a wireless network communication problem. There we look at the assignment of frequency resources to guarantee a reliable communication. Thus, different devices in a given area should learn on their own, which frequency they can use without disturbing the communication of other devices. As an overlap of frequencies can lead to a lack of communication, it is important that all devices select their own frequencies.

To accomplish this task, all devices that need to communicate become agents. The given area in which their communication takes place, the so-called communication cell, is the environment of the MARL problem. At each time step, every device chooses a communication channel (a frequency band) that it wants to use for its communication. To ensure that the problem can be solved reliably, each agent receives the following information in its state: the communication channel used in the previous step, the own Quality of Service (QoS) achieved by the last action, a vector of all neighbouring devices and the communication channels the neighbouring devices used in their last action. After all agents have selected a communication channel, they receive their next state and a reward. We choose the reward to be the sum of the achieved QoS of all agents, since a shared reward avoids adversarial behavior and leads to cooperation between the agents.

We train this MARL task using a Q-Learning algorithm. We train our agents as well with a NashQ algorithm, which is adapted for Multi-Agent learning from Game Theory. The results show that the agents learn to communicate in a reliable way. However, the number of agents influences the training, since the number of possible state combinations increases exponentially with the number of agents. By comparison of the two different algorithms, the NashQ algorithm needs slightly less episodes to converge to an optimal policy than the Q-Learning algorithm.

Deep Learning-based Time Series Forecasting for Industrial Discrete Process Data

Olaf Sassnick[1,2,3]

[1] Josef Ressel Centre for Intelligent and Secure Industrial Automation
[2] Salzburg University of Applied Sciences
[3] Paris Lodron University of Salzburg
olaf.sassnick@fh-salzburg.ac.at

Abstract

With the introduction of Industry 4.0, the continuous collection and monitoring of industrial process data have become fundamental aspects of modern operational technology (OT) systems. As a result, multivariate time series data with high resolution and dimensionality are potentially available. One promising approach for utilizing this information is data-driven forecasting, where insights from historical data allow to predict future outcomes. These predictions enable proactive interventions to improve the efficiency of an industrial process by either predicting the impact of changes of process parameters beforehand or detecting process anomalies and malfunctions on-the-fly. For instance, time series forecasting can be used to optimize process parameters by predicting the impact of changes in advance.

While deep learning-based forecasting models have demonstrated strong performance in various domains [1], their effectiveness for discrete manufacturing processes are not yet well studied. In this talk, we present a dataset that captures the key characteristics of discrete manufacturing time series data. Consequently we present results of state-of-the-art deep learning-based forecasting models on this dataset, identifying Crossformer [3] and DUET [2] as the best-performing architectures. We conclude that particularly more research needs to be done for applications requiring long-term time series generation with recursive forecasting due to real-time constraints. While we already improved the performance significantly with DUET when performing a second-pass training based on the model's own forecasts, it is still not stable for long-term forecasts.

References

1. Qiu, X., Hu, J., Zhou, L., Wu, X., Du, J., Zhang, B., Guo, C., Zhou, A., Jensen, C.S., Sheng, Z., Yang, B.: TFB: Towards comprehensive and fair benchmarking of time series forecasting methods. Proc. VLDB Endow. **17**(9), 2363–2377 (2024)
2. Qiu, X., Wu, X., Lin, Y., Guo, C., Hu, J., Yang, B.: DUET: Dual clustering enhanced multivariate time series forecasting. In: SIGKDD (2025), accepted Version
3. Zhang, Y., Yan, J.: Crossformer: Transformer utilizing cross-dimension dependency for multivariate time series forecasting. In: International Conference on Learning Representations (2023)

Anomaly Detection for Multivariate Times Series in Industrial Applications

Martin Uray

Josef Ressel Centre for Intelligent and Secure Industrial Automation
Salzburg University of Applied Sciences, Salzburg, Austria
martin.uray@fh-salzburg.ac.at

Multivariate Time Series (MTS) data occur in diverse fields such as neuroscience, biology, cybersecurity, and industrial manufacturing. Due to the inherent complexity of such data, advanced analytical methods are required to identify meaningful patterns. A key application is anomaly detection, where a model learns the normal, benign behavior of a system, and deviations from this learned distribution are flagged as anomalies. However, state-of-the-art approaches fail to incorporate the underlying structure and dynamics of the data. Moreover, they typically fail to provide insights into the root cause of an anomaly, reducing their effectiveness as assistive systems.

In this talk, we present our research on a novel approach to anomaly detection that explicitly models the latent dynamical processes governing MTS data. We assume that all observed processes originate from an underlying latent space, where the system dynamics can be described using learned stochastic differential equations. By reconstructing the observed data via Bayesian inference, we derive a reconstruction loss – the likelihood – to quantify the degree of abnormality. This approach naturally enhances the ability to capture the cyclic, structured behavior common in industrial data while potentially improving the robustness and interoperability of anomaly detection on existing benchmarks. Furthermore, it enables the precise identification of an anomaly's root cause.

From Load Profiles to Life Patterns: The Privacy Dilemma in Smart Energy Data

Dejan Radovanovic
dejan.radovanovic@fh-salzburg.ac.at
Center for Secure Energy Informatics at the Salzburg University of Applied Sciences, Salzburg, Austria

Smart metering systems offer significant benefits for energy management, but their detailed data collection raises serious privacy concerns. This presentation summarizes ongoing doctoral research investigating privacy risks in smart meter data through load profile analysis. It focuses on two main attack scenarios: (i) re-identification of households among several hundred pseudonymized load profiles, and (ii) re-identification of certain households even within anonymized subsets. Additionally, the research shows that socio-demographic characteristics, such as owning a sauna or swimming pool, can be inferred from weekly load profiles.

The talk also explores the privacy-utility tradeoff (Figure 1), evaluating countermeasures like temporal coarsening and load anonymization. While these reduce privacy risks, they can reduce data utility, impacting applications such as demand response, load forecasting, and personalized energy feedback. The findings highlight the urgent need for privacy-preserving mechanisms that are both robust and utility-aware, ensuring smart meter data can be used safely without losing its value for energy-related applications

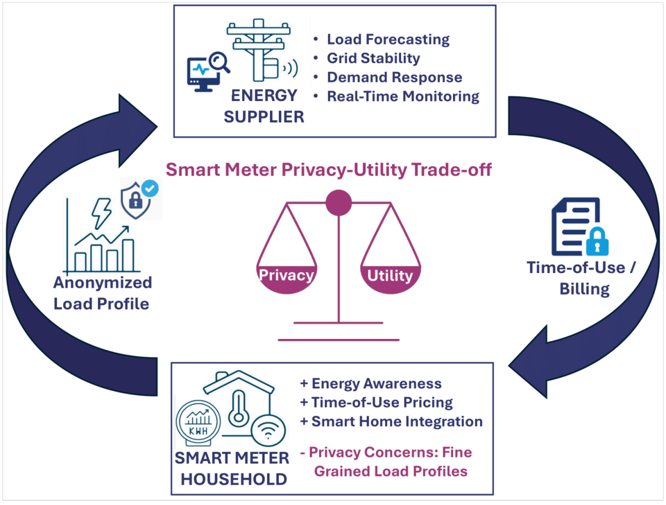

Figure 1: Balancing Privacy and Utility: Anonymizing load profiles to protect privacy while supporting essential applications like load forecasting and demand response.

LLM-KABOOM: Blasting Unstructured Web Data into Knowledge Graphs by using LLMs

Jasmin Saxer[1][0009-0000-1263-5757]

Institute of Computer Science, School of Engineering, Zurich University of Applied Sciences, Winterthur, Switzerland
jasmin.simonasaxer@zhaw.ch
https://www.zhaw.ch/en/about-us/person/saxr

Abstract. As digital content continues to grow, more data than ever is available on the web. However, much of this information is unstructured or semi-structured, making it difficult to extract and organize effectively. As a result, Knowledge Graph Construction (KGC) has emerged as a crucial research area, particularly in extracting structured knowledge from semi-structured websites. While existing datasets (*e.g.*, SWDE[1]), and methods (*e.g.*, OpenCeres, WebKE), have been effective for simple domains like product information or movie data, they haven't been applied on more complex domains and unstructured websites due to the need for extensive training data for these methods and the lack of a zero-shot learning option.

Many research efforts focus on predefined topics, allowing for well-defined relationships and entity types (ontology). However, for new topics without a predefined ontology, Large Language Models (LLMs) such as ChatGPT or Gemini are increasingly being used to generate ontologies and automate data extraction.

Building on these advancements, the research project *LLM-KABOOM* (*L*arge *L*anguage *M*odels for *K*nowledge *A*cquisition, *B*lasted *O*ntology *O*rganization, and *M*odeling) aims to construct knowledge graphs from user queries using webpages as an information source. Our approach follows a structured pipeline: ontology development, webpage classification, information extraction, and graph construction. To evaluate the effectiveness of LLMs, we compare their performance against both traditional and state-of-the-art methods at each step.

As a case study, we first focus on the analysis of the digitalization degrees of municipal services. Specifically, as data source we extract service-related information and data from municipality web pages. In the first step, LLMs, webpage classification, or rule-based methods are used to determine the presence of services. When a service is detected, we compare different information extraction methods to obtain the relevant details of the service.

By integrating LLMs into the KGC process and comparing them with traditional methods, we aim to identify the steps where LLMs offer the most utility and feasibility.

[1] Structured Web Data Extraction Dataset:
https://academictorrents.com/details/411576c7e80787e4b40452360f5f24acba9b5159

Sleep staging and apnea detection using low-cost wearables and deep learning

Sebastian Baron,

Department for Artificial Intelligence & Human Interfaces, University of Salzburg,
Jakob-Haringer-Strasse 2, 5020 Salzburg, Austria

Sebastian.Baron@plus.ac.at,

Sleep staging based on polysomnography (PSG) performed by human experts is the de facto "gold standard" for the objective measurement of sleep. PSG and manual sleep staging is, however, personnel-intensive and time-consuming and is thus impractical for monitoring a person's sleep architecture over extended periods of time. The same holds true for sleep apnea detection, which is also based on PSG-analysis and performed by human experts. Both, sleep architecture as well as sleep apnea play a crucial role for individual health and well-being, presenting the need for a less resource- and time-intensive method to monitor a person's sleep on a regular basis. Over the course of my doctoral research, I investigate in possibilities to perform these tasks with the help of low-cost wearable sensors and deep-learning algorithms.

Sleep apnea is defined as the absence (obstructive or central apnea) or significant reduction (hypopnea) of respiratory effort, paired with a desaturation of blood oxygen. Therefore, it is usually diagnosed using the respiration-related signals of the polysomnography (usually the thermistor signal) as well as blood oxygen saturation (SpO2). Previous machine learning approaches in this area mainly tried to model this procedure directly using signals related to respiration, blood oxygen and sometimes sound, while little attention is paid to the effects of changing blood oxygen levels on cardiac signals.

Similar to my work in the area of sleep staging, my main research question is how and to which level of precision sleep apnea can be detected in a procedure that is as non-invasive for a potential user as possible and can therefore be carried out at home with minimal equipment. Cardiac signals are therefore especially interesting, as they are relatively easy to obtain in a rather non-invasive manner using a single sensor. In my work I'll therefore mainly focus on apnea severity classification using Inter-beat-interval (IBI) time series data, which describe the duration between two consecutive heartbeats. However, I'll also report on the extend to which the consideration of additional signal data (such as respiration-related signals and sleep position), which can also be obtained by at least some of the sensors that were used within the 'Virtual Sleep Lab' project, which represents another part of my thesis, that focused on sleep-stage classification with low-cost wearables.

Profile Generators: A Link between the Narrative and the Binary Matrix Representation

Kutil Raoul
raoulhugo.kutil@plus.ac.at
PMU, PLUS-AIHI, IDA-Lab

Background: Mental health disorders (cognitive disorders) are defined by deficits in the cognitive abilities of the patient. The DSM-V is diagnostic manual containing detailed definitions of mental-health and brain-related conditions with details, examples of the signs and symptoms of conditions. A simple machine-actionable representation of the cognitive disorders was developed to measure their similarity and separability, but this representation is not equipped to deal with the most complex disorders. Neither the generation of the true binary matrix representation, nor the use of it in any similarity calculation is viable because of the size these disorders have.

Objective: The goal of this research is to develop an alternative representation that functions as a link between the narrative form of the DSM-V and the binary matrix representation and allows for an automated generation of valid symptom profiles.

Methods: By using a pre-defined strict format of lists, sets and numbers with a few variations, it is possible to represent the complex diagnostic pathways which manifest in a huge amount of symptom combinations. This format is called symptom profile generator (short: generator) because it depicts the binary matrix in a still readable, adaptable and comprehensive form, while also allowing an easy generation of all the symptom combinations. Most of the time a cognitive disorder consists of multiple diagnostic criteria containing several symptoms, which can now be translated in a list of multiple generators.

Results: Embedding several psychotic disorders in generator form and using these to create all symptom combinations revealed that complex disorders in matrix representation are so large that they are barely workable with. The MPCS algorithm (maximum pairwise cosine similarity) for similarity calculation is not equipped to deal with matrices of this size. This incentivized the development of a profile reduction method through clever generator manipulation when looking for a specific MPCS value of two disorders.

Conclusion: The generators allow for an easier generation of the binary representation and handling of disorders over huge matrices. Additionally, it makes it possible to calculate a specific case of the MPCS between complex disorders by reducing the number of symptom combinations in an approach called conditional generators.