Paper Overview: Presentations in Research & Industry Track

Research Track

Speaker	Contribution
Philipp Armbrust (University of Klagenfurt, AT)	Comparison of solution approaches for the propagation of quality requirements of ste
Josh Beal; German Wehinger (Julius Blum GmbH, AT)	Forecast Aggregation and Error Comparison: An Empirical Study
Samuel Saleh Bitrus (V-Research, AT)	CRISP-DM Based Data Mining Methodology for Tribological Optimisation
Pedro Casas; Pavol Mulinka; Juan Vanerio (Austrian Institute of Technology, AT)	NetSEC at High-Speed: Distributed Stream Learning for Security in Big Networking Data
Pedro Casas; Gonzalo Marín; Germán Capdehourat (Austrian Institute of Technology, AT)	DeepMAL - Deep Learning Models for Malware Traffic Detection and Classification
Michele Della Ventura (Music Academy Studio Musica, IT)	Symbolic Music Text Fingerprinting: Automatic Identification of Musical Scores
Stefan Huber (Salzburg University of Applied Sciences, AT)	Persistent Homology in Data Science
Shafi Kamalbasha; Manuel Eugster (Avira Operations GmbH & Co. KG, DE)	Bayesian A/B Testing for Business Decisions
Christina Petschnigg; Jürgen Pilz (University of Klagenfurt, AT)	Uncertainty aware deep point based neural network for 3D object classification
Dino Pitoski; Thomas J. Lampoltshammer; Peter Parycek (Danube University Krems, AT)	Human Migration as a Complex Network: Appropriate Abstraction and the Feasibility o
Alexandra Posekany (TU Wien, AT)	Outlier detection in Bioinformatics with Mixtures of Gaussian and heavy-tailed distribu
Sabrina Rosmann; Thomas Feilhauer; Steffen Finck; Martin Sobotka (Vorarlberg University of Applied Sciences, AT)	An Easy-to-Use Execution Environment for the Parallelisation of Computationally Inter
Johannes Schneider (University of Liechtenstein, LI)	Personalization of Deep Learning
Gunter Spöck; Maximilian Arbeiter (University of Klagenfurt, AT)	Physical-Statistical Modelling of Micro-Meteorology in an Alpine-Valley serving as Inpu

Industry Track

Speaker	Contribution
Maximilian Arbeiter (P.SYS System Creation, AT)	Simulating a cyber-physical system for behavior of elderly persons
Frederick Bednar; Gregor Sieber (EBCONT proconsult GmbH, AT)	Live Quality Validation Criteria for Executing a Test Set: A Heuristic Approach for Text Documents
Frank Blau (Massive Art, AT)	An Introduction to Graph Databases for Business Intelligence
Walter Kollmann (Julius Blum, GmbH, AT)	How to develop and maintain a network of AI champions
Tanja Maier (P.SYS System Creation, AT)	Modelling of Human Behaviour and Detection of Exceptions
Daniil Pozdeev; Andrey Saltan (Okko, Russian Federation; HSE University, St. Petersburg, RU)	Profit-maximizing Approach in Uplift Modelling: Evidence from the Media-service Provider
Lukas Prasol (Hilti AG, AT)	Digitalization in production - From the big picture to a dedicated solution
Dejan Radovanovic (Salzburg University of Applied Sciences, AT)	Introducing Natural Language Interface to Databases for Data-Driven Small and Medium Enterprises
Sinan Tanakz (Kapsch BusinessCom AG, AT)	
Zornica Vaskova Vasileva (Liebherr-Werk Nenzing GmbH, AT)	Multidimensional Sequential Pattern to Find Causes of Problems

Paper Abstracts: Presentations in Research & Industry Track

Research Track

Comparison of solution approaches for the propagation of quality requirements of steering gears

Philipp Armbrust (University of Klagenfurt, AT)

In the supply chain of the automotive industry the propagation of high quality standards is required. In the daily operations of steering system suppliers, the analysis of End of Line (EOL) vibroacoustic measurements encoded as order spectra for ball nut assemblies (BNA) is indispensable. Our goal is to find quality windows for the given BNA order spectra to detect faulty components. Due to the difficult interpretation of heuristic solutions, we use a Mixed Integer Linear Programming (MILP) formulation to analyze the solution quality of a genetic algorithm for the aforementioned problem. We prepare a carefully constructed benchmark set, which reflects the behavior of real-world EOL order spectra. In the provided computational study, we demonstrate the efficiency of the MILP approach on our benchmark instances with up to 945 order spectra, each consisting of 260 spectral orders.

Forecast Aggregation and Error Comparison: An Empirical Study

Josh Beal; German Wehinger (Julius Blum GmbH, AT)

The aim of this paper is to present empirical results associated with forecast performance. It is known that common measures of error fail to be scale invariant, and hence cannot be used to make meaningful error comparisons on forecasts across differing time series. This offers a particular challenge toward forecast improvement when one's intent is to compare error across different units or granularity. Moreover, although it is prudent to test many forecast methods on a time series, one cannot be sure that a single selected method will not lead to complete forecast failure. We address the aforementioned challenges by analyzing a sizable collection of time series inhouse.

CRISP-DM Based Data Mining Methodology for Tribological Optimisation

Samuel Saleh Bitrus (V-Research, AT)

This work provides a guideline for a structural approach towards data mining projects in tribology. Due to the specifics of tribological processes, parts of the DMME methodology need to be refined. The refined data mining methodology is applied to an on-going data mining project in tribology aimed at predicting wear rate and coefficient of friction of nitrocarburised coatings. The applied adapted methodology provides an efficient framework for data generation, preparation and analysis. At the same time, it supports and guides interdisciplinary work between data scientists and tribologists.

NetSEC at High-Speed: Distributed Stream Learning for Security in Big Networking Data

Pedro Casas; Pavol Mulinka; Juan Vanerio (Austrian Institute of Technology, AT)

Continuous, dynamic and short-term learning is an effective learning strategy when operating in very fast and dynamic environments, where concept drift constantly occurs. In this paper, we focus on a particularly challenging problem, that of continually learning detection models capable to recognize network attacks and system intrusions in highly dynamic environments such as communication networks. We consider adaptive learning algorithms for the analysis of continuously evolving network data streams, using a dynamic, variable length system memory which automatically adapts to concept drifts in the underlying data. By continuously learning and detecting concept drifts to adapt memory length, we show that adaptive learning algorithms can continuously realize high detection accuracy over dynamic network data streams. To deal with big network traffic streams, we deploy the proposed models into a big data analytics platform for network traffic monitoring and analysis tasks, and show that high speed up computations (as high as x5) can be achieved by parallelizing off-the-shelf stream learning approaches.

DeepMAL - Deep Learning Models for Malware Traffic Detection and Classification

Pedro Casas; Gonzalo Marín; Germán Capdehourat (Austrian Institute of Technology, AT)

Robust network security systems are essential to prevent and mitigate the harming effects of the ever-growing occurrence of network attacks. In recent years, machine learning-based systems have gain popularity for network security applications, usually considering the application of shallow models, which rely on the careful engineering of expert, handcrafted input features. The main limitation of this approach is that handcrafted features can fail to perform well under different scenarios and types of attacks. Deep Learning (DL) models can solve this limitation using their ability to learn feature representations from raw, non-processed data. In this paper we explore the power of DL models on the specific problem of detection and classification of malware network traffic. As a major advantage with respect to the state of the art, we consider raw measurements coming directly from the stream of monitored bytes as input to the proposed models, and evaluate different raw-traffic feature representations, including packet and flow-level ones. We introduce DeepMAL, a DL model which is able to capture the underlying statistics of malicious traffic, without any sort of expert handcrafted features. Using publicly available traffic traces containing different families of malware traffic, we show that DeepMAL can detect and classify malware flows with high accuracy, outperforming traditional, shallow-like models.

Symbolic Music Text Fingerprinting: Automatic Identification of Musical Scores

Michele Della Ventura (Music Academy Studio Musica, IT)

The explosion of information made available by the Internet requires continuous improvement/ enhancement of the search engines. The heterogeneity of information requires the development of specific tools depending on whether it is text, image, audio, etc. One of the areas considered insufficiently by the researchers concerns the search for musical scores. This paper aims to presents a method able to identifying the fingerprint of a musical score considered in its symbolic level: it is a

compact representation that contains specific information of the score that permits to differentiate it from other scores. A Musical Score Search Engine (MSSE), able to use the fingerprint method to identify a musical score in a repository, has been created. The logic of operation is presented along with the results obtained from the analysis different musical scores.

Persistent Homology in Data Science

Stefan Huber (Salzburg University of Applied Sciences, AT)

Topological data analysis (TDA) applies methods of topology in data analysis and found many applications in data science in the recent decade that go well beyond machine learning. TDA builds upon the observation that data often possesses a certain intrinsic shape such as the shape of a point cloud, the shape of a signal or the shape of a geometric object. Persistent homology is probably the most prominent tool in TDA that gives us the means to describe and quantify topological properties of these shapes.

In this paper, we give an overview of the basic concepts of persistent homology by interweaving intuitive explanations with the formal constructions of persistent homology. In order to illustrate the versatility of TDA and persistent homology we discuss three domains of applications, namely the analysis of signals and images, the analysis of geometric shapes and topological machine learning. With this paper we intend to contribute to the dissemination of TDA and illustrate their application in fields that received little recognition so far, like signal processing or CAD/CAM.

Bayesian A/B Testing for Business Decisions

Shafi Kamalbasha; Manuel Eugster (Avira Operations GmbH & Co. KG, DE)

Controlled experiments (A/B tests or randomized field experiments) are the de facto standard to make data-driven decisions when implementing changes and observing customer responses. The methodology to analyze such experiments should be easily understandable to stakeholders like product and marketing managers. Bayesian inference recently gained a lot of popularity and, in terms of A/B testing, one key argument is the easy interpretability. For stakeholders, “probability to be best” (with corresponding credible intervals) provides a natural metric to make business decisions. In this paper, we motivate the quintessential questions a business owner typically has and how to answer them with a Bayesian approach. We present three experiment scenarios that are common in our company, how they are modeled in a Bayesian fashion, and how to use the models to draw business decisions. For each of the scenario, we showcase a real-world experiment, the results and the final business decisions drawn.

Uncertainty aware deep point based neural network for 3D object classification

Christina Petschnigg; Jürgen Pilz (University of Klagenfurt, AT)

Efforts in various planning scenarios like factory planning, motion and trajectory planning, product design, etc. tend towards full realization in 3D. This makes point clouds an important 3D data type for capturing and assessing different situations. In this paper, we design a Bayesian extension to the classical frequentist PointNet classification network [1] by applying Bayesian convolutions and linear layers with variational inference. This approach allows to estimate the model's uncertainty in its predictions. Further, we are able to describe how each point in the point cloud contributes to the overall predictive uncertainty. Additionally, our network is compared against the state-of-the-art and shows strong performance. We prove the feasibility of our approach using a ModelNet 3D data set. Further, we generate an industrial 3D point data set at a German automotive assembly plant and apply our network. The results show that we can improve the frequentist baseline on ModelNet by about 6.46 %.

Human Migration as a Complex Network: Appropriate Abstraction and the Feasibility of Network Science Tools

Dino Pitoski; Thomas J. Lampoltshammer; Peter Parycek (Danube University Krems, AT)

The number of Network Science studies has risen significantly in recent two decades. Various real phenomena are increasingly analyzed as complex networks. Human migration was seldom analyzed, however, in line with global circumstances, the number of migration-as-network applications has recently grown as well. Those new migration-as-network studies are hands-on implementations of elementary measures and models. Assessments on the right kind of network abstraction for human migration, as well as the feasibility and interpretability of measures on the phenomenon, have not yet been offered. We investigate these aspects, assessing the congruence of network tools used for analyzing migration, and their informative potential for the policy and decision-making domain.

Outlier detection in Bioinformatics with Mixtures of Gaussian and heavy-tailed distributions

Alexandra Posekany (TU Wien, AT)

Starting from approaches in Bioinformatics, we will investigate aspects of Bayesian robustness ideas and compare them to methods from classical robust statistics. Bayesian robustness branches into three aspects, robustifying the prior, the likelihood or the loss function.

Our focus will be the the likelihood itself. For computational convenience, normal likelihoods are the standard for many basic analyses ranging from smiple mean estimation to regression or discriminatory models. However, similar to classical analyses non-normal data cause problems in the estimation process and are often covered with complex models for the overestimated variance or shrinkage. Most prominently, Bayesian non-paramtetrics approach this challenge with infinite mixtures of distributions. However, infinite mixture models do not allow an identification of outlying values in ``near-Gaussian'' scenarios being almost too flexible for such a purpose.

The goal of our works is to allow for a robust estimation of parameters of the ``main part of the data'', while being able to identify the outlying part of the data and providing a posterior probability for not

fitting the main likelihood model. For this purpose, we propose to mix a Gaussian likelihood with heavy-tailed or skewed distributions of a similar structure which can hierarchically be related to the normal distribution in order to allow a consistent estimation of parameters and efficient simulation.

We present an application of this approach in Bioinformatics for the robust estimation of genetic array data by mixing Gaussian and student's t distributions with various degrees of freedom. To this effect, we employ microarray data as a case study for this behaviour, as they are well-known for their complicated, over-dispersed noise behaviour. Our secondary goal is to present a methodology, which helps not only to identify noisy genes but also to recognise whether single arrays are responsible for this behaviour. Although Bioinformatics dropped array technology in favor of sequencing in research, the medical diagnostics has picked up the methodology and thus require appropriate error estimators.

An Easy-to-Use Execution Environment for the Parallelisation of Computationally Intensive Data Science Applications

Sabrina Rosmann; Thomas Feilhauer; Steffen Finck; Martin Sobotka

(Vorarlberg University of Applied Sciences, AT)

With Cloud Computing and multi-core CPUs parallel computing resources are becoming more and more affordable and commonly available. Parallel programming should as well be easily accessible for everyone. Unfortunately, existing frameworks and systems are powerful but often very complex to use for anyone who lacks the knowledge about underlying concepts. This paper introduces a software framework and execution environment whose objective is to provide a system which should be easily usable for everyone who could benefit from parallel computing. Some real-world examples are presented with an explanation of all the steps that are necessary for computing in a parallel and distributed manner.

Personalization of Deep Learning

Johannes Schneider (University of Liechtenstein, LI)

We discuss training techniques, objectives and metrics toward mass personalization of deep learning models. In machine learning, personalization addresses the goal of a trained model to target a particular individual by optimizing one or more performance metrics, while conforming to certain constraints. To personalize, we investigate three methods of ``curriculum learning`` and two approaches for data grouping, i.e., augmenting the data of an individual by adding similar data identified with an auto-encoder. We show that both ``curriculum learning'' and ``personalized'' data augmentation lead to improved performance on data of an individuals'. Mostly, this comes at the cost of reduced performance on a more general, broader dataset.

Physical-Statistical Modelling of Micro-Meteorology in an Alpine-Valley serving as Input to an Online-Pollutant-Dispersion-Simulation

Gunter Spöck (University of Klagenfurt, AT)

The micro-meteorology of Alpine-valleys must follow certain physical laws. We have build-up 13 meteorological measurement stations in an Alpine-valley in Austria. Every five minutes these stations provide data on temperature, air-pressure, air-humidity, luminosity, rain, wind-speed and wind-direction. Additionally, we have one all-sky camera which provides measurements of cloudiness. From these meteorological measurements we can physically define six weather stability classes, based on which different modelling steps must be taken to model spatially and physically coherent the wind-speeds and wind-directions in the valley. For the modelling of all these data combined physical-statistical interpolation techniques are used to predict the different meteorological fields. We make use of generalized additive models, inverse distance weighting, and physical formulas, that the interpolated data must follow to be physically valid predictions of the 3-dimensional weather fields.

In the Alpine-valley there is a strong pollutant emitting source whose minutely emissions of certain pollutants are known. From the predicted weather fields every 5 minutes meteorological model parameters are calculated that serve as inputs to Lagrangian- and Gaussian-plume pollutant dispersion simulations, that calculate the cumulative pollution of the soil and the current pollution of the air.

This talk will concentrate mainly on the physical-statistical modelling of the the micro-meteorology in the valley, give discussions on Bayesian spatio-temporal future work for meteorological field interpolation, and explain only some small hints on pollutant dispersion simulation.

Simulating a cyber-physical system for behavior of elderly persons

Maximilian Arbeiter (P.SYS System Creation, AT)

The validation of a cyber-physical system is an important part of it's development. However, in many cases it is difficult, very expensive or just impossible to do this in a conventional way. Here we consider a monitoring system for elderly lonely living persons which should be installed into their homes. We have an algorithm which learns the behavior of the person based on measurements of various non-invasive sensors. Obviously, the validation of such a system would need a huge amount of test persons due to the large variability of human behavior. To overcome this problem, we develop a simulator to generate sequences of human decisions in a living environment equipped with several virtual sensors. We define a utility function which describes the agent's decision behavior. The utility function, additional with some vital parameters, give us the possibility to parametrize a person. Decisions are made by maximization of the expected value of this utility function. Further, we use Markov decision processes to model such decision sequences and to and optimal strategies for behavioral decision making.

Live Quality Validation Criteria for Executing a Test Set: A Heuristic Approach for Text Documents

Frederick Bednar; Gregor Sieber (EBCONT proconsult GmbH, AT)

Defining reliable KPI for NLP and NER tasks are, besides from the well known quality measures such as the F1 score for validating files, is always a challenge when not knowing in detail and in advance which entities will occur. One would need to hold on expectations if the domain was specified enough.

Several approaches have already been conducted with methods such as using pre-trained machine learning models, including concepts of Data Mining or integrating semantics resp. ontologies. Those ideas dealt with having a reliable gold standard that can be used for checks.

If results were needed a priori resp. immediately right after execution, but one could not refer to an existing gold standard of annotated data (as data quality is available in silver standard only); if the amount of data and the variety of possible entities were both high, one would have to think how much pre-processing is necessary to solve the problem.

We introduce an approach for a simplified quality measure that relies on the combination of statistical and heuristic methods as well as Data Mining and domain knowledge, in order to enable classification of documents with spurious entities rather tending to have false negatives or false positives included. The goal of this paper is to outline a generic method that can be used for all text content that is determined and structured by a specific subject.

An Introduction to Graph Databases for Business Intelligence

Frank Blau (Massive Art, AT)

Graph Databases are the de facto standard for social media and network-data persistence and analysis. As the tools become more readily available and usable, there are new opportunities to implement them for other business intelligence domains as well. This talk will give an overview of the technology and present some interesting use cases for further investigation. A demonstration of a marketing analysis use case using open source graph database technology will be conducted as part of the presentation.

How to develop and maintain a network of AI champions

Walter Kollmann (Julius Blum, GmbH, AT)

Artificial intelligence is a field of technology that has the potential to change the way we work. We believe that it will be a permanent aspect in our business processes. This presentation is about how to set up and establish a team of experts in a company to be well prepared for the future.

Modelling of Human Behaviour and Detection of Exceptions

Tanja Maier (P.SYS System Creation, AT)

Typically, people want to live in their familiar surroundings for as long as possible. However, the risk of accidents, such as falling, or medical incidents, increases with age. If these elderly people live alone, the safety of these persons is affected.

In recent years, many technical systems have been developed to support the privatised or institutional care of elderly people and/or people with special needs. Some of these systems are based on automatic situation and accident detection. All are invasive, expensive and/or not very reliable.

Each person has their own typical habits and activities, which can change over time. Therefore, it is not possible to predefine normal activity sequences or activity sequences, that typify exceptions. We propose algorithms that model a person's normal daily routine in their Living environment to identify exceptions that may occur. These exceptions are represented by deviations from the normal behaviour. The model is built on real-time observations, so it learns the daily routines autonomously. Deviations from the learned models are classified as exceptions.

The system in which these algorithms are implemented are limited to certain restrictions like affordability, non-invasiveness, full privacy, etc. These conditions mandate that the algorithms or models must be extremely sparse and the storage of data is limited.

For the modelling of human behaviour, we use embedded Markov chains. Deviations from normal behaviour are detected by a system of penalty functions. All calculations are recursive. Therefore, the storage of large amounts of data is avoided. The technical system is still in development. The algorithms have been tested and optimised with simulated data and have subsequently been applied to real-life pilot data.

Profit-maximizing Approach in Uplift Modelling: Evidence from the Media-service Provider

Daniil Pozdeev; Andrey Saltan (Okko, Russian Federation; HSE University, St. Petersburg, RU)

Commercial companies allocate huge marketing budgets to retain customers or to attract new clients. Sometimes the use marketing campaigns which may have weak performance. Churn reduction in companies strategy may lead to unnecessary costs for customers not sensitive to marketing actions. To identify the true treatment responders on a marketing action, we examine uplift modeling which could reduce target audience. We use uplift as predictive modeling technique, to estimate casual effect via online controlled experiments in media-services provider. In terms of machine learning, uplift predicting can be outlined as binary classification problem, where positive class of customers are sensitive to intervention customers. However, purpose of marketing strategy may significantly impact on choice of criteria for classifying users as positive. Our empirical study includes the experiment in which we examine profit-driven, revenue-based and conversion targets for classifying users and quantifying model performance in terms of business value. The experiments show that the proposed profit maximization approach can outperform existing uplift modeling algorithms based on conversion targets, which are widely used in contractual settings.

Digitalization in production - From the big picture to a dedicated solution

Lukas Prasol (Hilti AG, AT)

In the light of the omnipresent digitalization, traditional physical processes become digitized with versatile hard- and software that create the potential of making the whole production process visible at different levels in production plants.

Processes as well as the produced parts will become able to sense their own condition as well as the state of their corresponding production environment. Combined with the ability to process and communicate the resulting data, a digital twin of the production process is created.

The consistent application of lean principles on physical processes in production – value streams – describes the first step in a digitalization approach. By reducing or elimination non-value adding activities the starting point is being set. Within the second step an identified business case is mandatory. This shows the added value of digitalization. The third step requires the application of digitized technologies.

Within this presentation a scenario-based approach for digitalization in manufacturing industry is shown. By applying a defined “digitalization process” which is based on four successive process steps a digital twin is created. The consistent and coordinated implementation of this approach in different plant areas creates a digital twin of the entire production.

Introducing Natural Language Interface to Databases for Data-Driven Small and Medium Enterprises

Dejan Radovanovic (Salzburg University of Applied Sciences, AT)

Reading text, identifying key ideas, summarizing, making connections and other tasks that require comprehension and context are easy tasks for humans, but training a computer to perform these tasks is a challenge. Recent advances in deep learning make it possible to interpret the text effectively and achieve high performance results across natural language tasks. Interacting with relational databases trough natural language enables users of any background to query and analyze a huge amount of data in a user-friendly way. The purpose of Natural Language Interface is to allow users to compose questions in Natural Language and receive the response also in Natural Language. The idea of using natural language instead of SQL has promoted the development of new type of processing called Natural Language Interface to Database (NLIDB). This paper is an introduction to Natural Language Processing and Natural Language Interface to Database, significant challenges in this research field and how to construct a company specific dataset. It also gives a brief overview of the major techniques used to develop Natural Language Interface to Databases.

Title pending

Sinan Tanakz (Kapsch BusinessCom AG, AT)

<Abstract pending>

Multidimensional Sequential Pattern to Find Causes of Problems

Zornica Vaskova Vasileva (Liebherr-Werk Nenzing GmbH, AT)

Nowadays machines are like a computer and have many sensors. Thus, each machine generates logs of items. An item contains, for example, information, warning or error data. The target is to find machines with unknown problems or unexpected behaviours comparing the data of a group of similar machines. Then again, if a certain problem or failure has occurred, it is very interesting to find the reason. It should be noted that the number of items increases constantly, so the data changes over time. Therefore, we should use a quick and efficient method to analyse our machine data. The methods for pattern mining, used by the market basket analysis, can be adapted to analyse machine data. A log item can be seen as a product in a supermarket and a machine can be seen as a customer. In machine data analysis, the order of log items is very important. Therefore, we are interested in sequences of items, which occur frequently in machine data. Thus, we search for (maximal) sequential pattern in machine data. In this study, we design a method that finds similar sequential patterns per group of machines. Additionally, we use this method to find the reason for a known error item or a certain failure in a group of machines that share this error. We applied our new method extensively on known problems and certain failures. The method performed remarkably well. It not only found the expected results, in the known cases, but furthermore detected valuable, previously unknown information about the machines. These newly found patterns can now be matched in machine data of different machine types. Analysing the machine data according to the introduced method proved to be very beneficial.
