Lecture series: Data Science in Action, fall 2023

Zoom, Thursdays, 12 – 1.30 pm

Please note that the syllabus will be updated if necessary

14 September Heiner Stuckenschmidt (Chair of Artificial Intelligence, University of Mannheim, Germany)

Lecture topic: Organizational matters (only relevant for students who are taking the exam)

21 September no lecture

28 September Evelina Gavrilova-Zoutman (Associate Professor at NHH Norwegian School of Economics, Norway)

Lecture topic: Forensic Data Science

Open Source Intelligence (OSINT) involves collecting intelligence from publicly available data. In this lecture I would like to focus on the application of Benford's law, which can be used both as data verification tool, in for e.g. election forensics, and as a fraud detection tool. In the lecture, I would show students how to implement Benford's law to detect presence of cartel behavior in fixing the LIBOR exchange rate around the 2008 financial crisis. We would use a software tool like R. We would start with a dataset with real LIBOR rates, we would extract the second digit and run a chi-square test. Then, we would build a loop around all available time periods.

5 October Kevin Bauer (Assistant Professorship of E-Business and E-Government, University of Mannheim, Germany)

Lecture topic: eXplainable Artificial Intelligence

Artificial Intelligence (AI) has undeniably reshaped numerous facets of our world, but its opacity often inhibits its potential, stirring concerns about ethics, fairness, and accountability. This lecture introduces the compelling field of eXplainable AI (XAI), a critical response to this 'black box' problem. We will initially delve into why explainability is crucial in today's AI-driven era. Next, we will chart the landscape of various XAI methods, from interpretable models that inherently provide explanations to post-hoc techniques that elucidate complex, black-box models. The exploration will extend to local and global explanations, demonstrating how they cater to different explanatory needs, followed by a brief foray into surrogate models, and counterfactual explanations. Finally, we will discuss the implications of these XAI methods on human-AI interaction based on recent research.

12 October Jessica Eynard (Assistant Professor at the University of Toulouse Capitole, France)

Lecture topic: Digital identity

Traditionally, identity has been understood as the set of traits or characteristics that make it possible to recognize a person and establish his or her individuality in the eyes of the law. It encompasses identifiers such as surname, first name, date of birth, etc. Technological developments have led to the multiplication of these elements. A Facebook profile, a video game avatar, a virtual double created from browsing tracks, etc., are all data that have led to the emergence in doctrine of an electronic, digital or even biometric identity. Some authors even refer to "identities", to cover every aspect of an individual's life. But do these realities really come under the heading of identity in the legal sense of the term? Is there even such a thing as a digital identity?

Answers to the questions surrounding digital identity are essential at a time when the European Commission intends to enable 80% of EU citizens to use such an identity. These answers are all the more eagerly awaited as all entities, whether public or private, are now looking for simple, effective solutions to identify their users, customers, patients, etc.

Several players have already positioned themselves in the identity supply market. Is this trend towards the privatization of identification desirable, and what role should these players and the state play in establishing a digital identity?

19 October Margret Keuper (Professor for Visual Computing, University of Siegen, Germany)

Lecture topic: Neural Architecture Search and Model Robustness

Recently, quite some effort has been dedicated to optimize neural architectures to facilitate ever improved model accuracy. The resulting models are often highly accurate in practice when the data used at test time is sufficiently similar to the training data. However, many state-of-the-art models are highly susceptible to even small domain shifts such as slight data corruptions, which can hamper their deployment in practical settings. As a result, the joint optimization of neural architectures for high accuracy on clean data as well as on data that has undergone some unknown domain shift (usually referred to as robust accuracy) becomes an important research question. It is different from established multi-objective neural architecture search (NAS) approaches such as Hardware aware NAS, (i) because the objective “clean accuracy” and “robust accuracy” are highly entangled. The robust accuracy of a model is usually bounded by its clean accuracy, or, in other words, its high accuracy on clean data is a necessary condition for a reasonable robust accuracy. (ii) A model’s robust accuracy depends on its architecture as well as on the training procedure, such that an architecture that can be trained to perform well under diverse conditions does not necessarily show this behavior when trained conventionally. In this talk, we will discuss multi-objective NAS and its relationship to improving model robustness, and the expected pitfalls when searching for best models to be deployed in practice.

26 October Margeret Hall (Associate Professor of Information Systems and Quantitative Analysis, University of Nebraska at Omaha, United States)

Lecture topic: Questions of accuracy and fairness in radicalization research

The Internet eases the broadcasting of data, information, and propaganda. This massive information-sharing platform is occasionally abused for propagating extreme and radical ideologies and that can pose threats to national security and citizens. Detecting the so called dark material has gained more

impetus following the recent outbreak of extremist groups and radical ideologies across the Web. The goal of this project, being the first of its own, is to surveil online social networks (OSN) and Web for real-time detection of visual propaganda by violent extremist organizations. Automated image classification for this content is a highly sought-after goal, yet raises the question of potential bias and discrimination in case of incorrect classification. This talk highlights the case of the group ISIS due to their prolific digital content creation. We will address aspects of ahe developed training dataset is used to train a convolutional neural network and implications on detecting and classifying based on the type of VEO and focus or intent of the image. Over 1.2 million images were automatically collected from suspicious OSN accounts and Web pages over a course of four years.

2 November Mikael Laakso (Associate Professor at Hanken School of Economics, Finland)

Lecture topic: Researching Research

The role of scientific information on the web is constantly evolving, appearing among general web search results, and being linked to within news articles and social media. There are millions of new scientific articles published every year, with an increasing degree of them being available openly on the web. During the last decade there has been a giant leap forward in what kind of tools and datasets are available for conducting research on both the content and metadata of these publications. Data science can help with uncovering interesting phenomena and trends that would be hard to discern without such approaches. The main focus of this lecture is to teach how research outputs on the web and metrics can be studied in different ways, giving examples from my own research as well as that of others. Despite advances in recent years there is still a lot of untapped potential for data science in this space, both from an academic and commercial perspective.

9 November Raik Stolletz & Seyed Mohammad Zenouzzadeh (Chair of Production Management, University of Mannheim, Germany)

Lecture topic: Performance Analysis of Queues with Machine Learning

Stochastic variability is prevalent in many operational systems, where random factors play a significant role. For instance, the number of arriving orders or the time to process an order are not constant and exhibit uncertain characteristics. Queueing systems are used to model and analyze such stochastic operations systems. The allow for predicting important performance measures to support decisions on how to set or how to use the capacities. Performance measures to evaluate the efficiency and effectiveness of the system are for example the resource utilization, expected waiting times, or specific service levels. Under certain assumptions, analytical methods are available to derive main performance measures. However, due to the system's structure and complexity, especially when the system parameters are time-dependent, no closed-form solutions are available. We present a flexible machine learning approach that can be used to approximate the performance measures of such complex queueing systems that are affected by time-dependent parameter changes.

16 November Marc Ratkovic (Professor for Social Data Science, University of Mannheim, Germany)

Lecture topic: Estimation and Inference on Nonlinear and Heterogeneous Effects

While multiple regression offers transparency, interpretability, and desirable theoretical properties, the method’s simplicity precludes the discovery of complex heterogeneities in the data. We introduce

the Method of Direct Estimation and Inference (MDEI) that embraces these potential complexities, is interpretable, has desirable theoretical guarantees, and, unlike some ex- isting methods, returns appropriate uncertainty estimates. The proposed method uses a machine learning regression methodology to estimate the observation-level partial effect, or "slope," of a treatment variable on an outcome, and allows this value to vary with background covariates. Importantly, we introduce a robust approach to uncertainty estimates. Specifically, we combine a split-sample and conformal strategy to fit a confidence band around the partial effect curve that will contain the true partial effect curve at some controlled proportion of the data, say 90% or 95%, even in the presence of model misspecification. Simulation evidence and an application illustrate the method's performance.

23 November Elisabeth Huis in 't Veld (Assistant Professor at Tilburg University, The Netherlands)

Lecture topic: Data Science for Health Innovations and Entrepreneurship

Data Science innovations can make a huge contribution to benefit individual and public health. However, knowing how to code might only get you so far. To maximise the chances of you actually making impact with your Data Science solutions, especially in the Healthcare sector, you need to design with the end in mind. There are certain entrepreneurial techniques and models that you can apply early in your research or development project, which will help guide, test and validate whether what you are developing will actually make sense in real life. By talking you through our journey in founding AINAR, from idea to research to an AI based game to conquer needle fear that we are commercializing through a start-up, I will discuss models, tips and tricks to make you a better data scientist.

30 November Marijn van Wingerden (Assistant Professor at Tilburg University, The Netherlands)

Lecture topic: Machine learning approaches to health data

In many care settings, a lot of health data is routinely collected and stored. However, the application of machine learning models to these rich datasets has been proven difficult due to unusual data formats, large amounts of missing data and implicit clinical knowledge that is essential for understanding the patterns in the data. In this presentation, several machine learning modelling projects, run with clinical partners involved in the design of the experiments, will be discussed to showcase the promise and the pitfalls of working with realistic health datasets. Examples include pattern mining of patient-reported outcomes (PROMs) from longitudinal recovery data, external validation of predictive models for the risk of heart failure, and machine learning models as an early warning system for patient deterioration.

7 December Beatrice Rammstedt (Professor for Psychological Assessment, Survey Design and Methodology, University of Mannheim, Germany)

Lecture topic: Quality of survey responses – the biasing effect of acquiescence