

Danish Data Science 2022

Abstract book

November 7-8th, 2022
Legoland Hotel & Conference



Danish
Data Science
Academy



Table of Contents

Table of Contents	1
A1. Unlocking to power of our data.....	3
A2. NLP Danske Bank.....	4
A3. Large language models via automated machine learning for epidemiological research: a case study and method comparison from the British National Child Development Study.....	5
A4. Representation learning for image-based characterization of industrial flocculation processes.....	6
A5. Improving image understanding with deep multimodal fusion.....	7
A6. Optimal Control of Residential Energy Storage Systems.....	8
A7. The National Health Data Science Sandbox for Training and Research.....	9
A8. Towards a unified framework for prediction of extremely imbalanced big data	10
A9. NextGP: A Julia package for next generation genomic prediction tools	11
A10. Machine learning methods for dynamic risk prediction of perianal fistulas in Crohn's disease.....	12
A11. Data Science as a part of Pharmaceutical Product Design.....	13
A12. Drug target family databases enabling data science and data-driven drug design.....	14
A13. The Privatization of AI Research(-ers): Causes and Potential Consequences – From university-industry interaction to public research brain-drain?.....	15
A14. A new approach for Topic Modeling based on Transformer Models and NER.....	16
A15. Discovery of human signaling systems v2.0 – pairing peptides to G protein-coupled receptors	17
A16. A neural network alternative to non-negative matrix factorization for mutational signature extraction in cancer genomics	18
A17. Is the USA a Shangri-la for European scientists? Inferring the causal effect of transatlantic mobility using quasi-experimental methods in repurposed data	19
A18. DELPHI – Data Environment for Life science, sensitive Personal data, and Health Investigations	20
A19. Speech and Multi-Sensory Data Modeling for Child and Youth Psychiatry	21
A20. Scholia - displaying a knowledge graph of science.....	22
A21. Challenges in the Energy Sector	23
A22. Changes in pupil size during listening as potential biomarkers of the hearing status – A classification study.....	24
A23. Using Deep Generative Models for Atomic Structure Solution of Metal Oxide Nanoparticles from Pair Distribution Function Data.....	25
A24. Residential Mobility and Segregation in Denmark	26

A25. Genomic prediction of winter-survival in perennial ryegrass (<i>Lolium perenne</i> L.), and the effect of genotype x environment interactions at the margin of the species distribution	27
A26. Prediction of Type 2 Diabetes using Machine Learning on Electronic Health Records ..	28
A27. Clinical Proteomics Data Science Tutorial	29
A28. The PSYchiatric clinical outcome prediction (PSYCOP) cohort: leveraging the potential of electronic health records in the treatment of mental disorders	30
A29. Combining chemistry knowledge and Graph Neural Networks Towards Interpretable Molecular Property Models	31
A30. Uncovering coherence networks in the human brain using directional statistics	32
A31. Utilizing Domain Knowledge and Data Science in Chemical Process Modelling.....	33
A32. Properties and transitions of mesoscale convective organisation during EUREC4A using unsupervised learning	34
A33. Vocal markers of neuropsychiatric conditions: assessing the generalizability of machine learning models and their clinical applicability.....	35
A34. Robust spatiotemporal actin filament disentanglement using a network theoretic framework	36
A35. Publicly Available Privacy-preserving Benchmarks for Polygenic Prediction	37
A36. Real-time process monitoring with label scarcity	38
A37. First Principal Models and Neural Networks for Defining Metabolic Capacity from Continuous Glucose Measurements (CGM) As a Tool for Personalized Nutrition.....	39
A38. The meaning in the machine - Inductive exploration of group discourse using dynamic word embeddings	40
A39. Causal representation learning for out-of-distribution generalization	41
A40. Health Data Science Sandbox project and interactive supercomputing training.....	42
A41. Prediction of Anxiety and Depression in ICD-Patients using Machine Learning Algorithms	43
A42. An Explainable Machine Learning Approach for the Detection of Lung Cancer in Denmark.....	44
A43. Mapping Complex Technologies via Science-Technology Linkages; The Case of Neuroscience-A transformer based keyword extraction approach	45
A44. Machine learning models application in proteomics.....	46
A45. A self-supervised model of the brain for psychiatric phenotyping - a project outline	47
A46. Efficient Closed Form Updates for Archetypal analysis	48
Poster walks	49

A1. Unlocking to power of our data

Sujit Khune

Novo Nordisk

Across the pharma industry, opportunities exist to improve R&D returns by unlocking the power of data to enhance and accelerate data driven decision-making across the value chain. While many data sources can be purchased or accessed (e.g. RWD), it is RCT data that holds the greatest promise in the search for relevant insights. However, making the most of this goldmine is not trivial.

First hand a structured and connected data lake must be prepared to overcome the historically siloed data in RCT. Secondly any friction, e.g., related to technology or simple governance, currently experienced in accessing data must be reduced. Finally, change management is critical to ensure cultural transformation to enable teams to leverage the data effectively.

We'll present our vision for this democratisation of RCT data that will allow scientist to work together, finding and sharing data and ultimately enhance the pace of innovation for the benefit of the patients.

A2. NLP Danske Bank

Philip Theut Stehr-Nielsen

Danske Bank A/S

Improving customer service and experience with NLP. An application of BERT and huggingface in order to understand message topics and create automatic replies to customers.

Commercial Excellence in Danske Bank is testing out the use of opens-source pre-trained natural language processing models to improve customer experience. This is done with a focus on identifying purpose, tone, and urgency of the message, to act appropriate.

The first test case is to categorize the messages that historically have been answered with a guide to self-service. This will improve response time and solving the customers inquire faster.

The model will be implemented as an API that can be called from the existing routing tool and be incorporated in existing work-flows.

For future purposes, we expect this API to be used in many more applications in the bank. The longer vision is to be able to treat all text data in the bank the same way to get a true cross-channel insight into customer needs. This is also a way to extract the important parts of the data, while removing sensitive information from the text that we don't need for this purpose.

A3. Large language models via automated machine learning for epidemiological research: a case study and method comparison from the British National Child Development Study

Rasmus Wibaek¹, Gregers Stig Andersen¹, Christina C. Dahm², Daniel R. Witte^{2,3}, Adam Hulman³

¹*Steno Diabetes Center Copenhagen, Herlev, Denmark*

²*Aarhus University, Aarhus, Denmark*

³*Steno Diabetes Center Aarhus, Aarhus University Hospital, Aarhus, Denmark*

We designed a study to demonstrate the potential of large language models in epidemiology. Pre-trained models were fine-tuned using AutoTrain (Huggingface) to predict current reading comprehension and future body mass index (BMI) and physical activity based on essays written by 11-year-old children (n=10,567) about how they imagine themselves as 25-year-olds. Then we compared predictive performance with regression approaches including demographic and lifestyle factors as predictors.

We observed different rankings of performance between the deep learning and regression approaches across the three outcomes. The performance of the deep learning approach appeared to depend on how closely the actual task i.e. writing an essay about the future was related to outcome. The language model could have picked up on linguistic features like grammatical correctness, vocabulary, complexity of sentences, etc., which led to the NLP method clearly outperforming linear regression when predicting reading comprehension score. Predictive performance for physical activity was similarly poor for the two methods (AUC-ROC=0.55), but slightly better than random assignment, while linear regression clearly outperformed the NLP approach when predicting BMI.

Our study demonstrated the potential of large language models to utilize textual data in epidemiological studies beyond the analysis of electronic health records and social media data.

A4. Representation learning for image-based characterization of industrial flocculation processes

Andreas Baum

DTU Compute

We introduce a novel type of model consisting of a convolutional neural network (CNN) and a partial least squares (PLS) part. We present parametrisation constraints in order to learn meaningful latent variable representations and show that the presented approach results in more robust models, while at the same time offering enhanced interpretability of the CNN activation patterns.

A5. Improving image understanding with deep multimodal fusion

Galadrielle Humblot-Renaux, Thomas B. Moeslund

Visual Analysis & Perception lab, Aalborg University

As humans, we rarely rely on a single source of information for understanding what's around us & making decisions. Similarly, autonomous systems greatly benefit from having multiple complementary modalities as input. For instance, RGB cameras cannot "see" in the dark, and should therefore be complemented with depth, thermal, and/or audio data for driving around at night. Deep learning from multiple modalities is an interesting problem in computer vision with broad applications. Specifically, how do we fuse modality-specific features into a common representation, while leveraging the unique strengths and properties of each modality? How do we handle conflicting or uninformative features? What happens in the case of sensor failure or domain shift?

A6. Optimal Control of Residential Energy Storage Systems

Maxim Khomiakov

Otovo

Energy prices are at historical highs whilst the accelerated transition towards renewable energy and political unrest unfolds. While many will attribute this a grim outlook, the argument towards residential photovoltaics seems as appealing as it has ever been. The question remains however, if the generated energy is most often generated when it is not readily needed, in what way may we learn an optimal control of the residential solar feedback loop. For a PV installation to be economically viable, self-consumption ratio of the energy produced need to be as high as possible. While recent technological developments made private storage solutions more financially viable, the volatility in the electrical spot market made the timing of consumption, storage and loading an ever more complex and crucial task. Prior works have looked at shorter temporal timespans, or in countries with a substantially higher amount of sun hours than in the Nordics. Most prevalent studies are also based on simulated consumption/production models, while our study spans several years of actual data from a large number of residential households in Norway. We demonstrate results from our method using a simple rule-based baseline, a non-parametric forecasting method as well as a dynamic programming approach for solving the optimal actions of storage, loading or discharging a residential household battery. The results show a substantial improvement of residential energy savings, an increase of self-consumption ratio and a lowering of PV system payback time. Our studies demonstrate the untapped potential of household on-line energy optimisation, which we believe will likely only become more apparent, as we continue the renewable energy transition and electrical appliances, storage and generation, becomes even more readily integrated.

A7. The National Health Data Science Sandbox for Training and Research

Jennifer Bartell^{1*}, Jose Alejandro Herrera Romero¹, Samuele Soraggi², Sander Boisen³, Jesper Roy Christiansen⁴, Jacob Fredegaard Hansen⁵, Conor O'Hare¹, and Anders Krogh^{1*}

¹Center for Health Data Science, Department of Health and Medical Sciences, University of Copenhagen, Denmark., ²Bioinformatics Research Center, Aarhus University, Denmark., ³Clinical Data Science Group, Department of Clinical Medicine, Aalborg University, Denmark., ⁴Computerome, Technical University of Denmark, Denmark, ⁵Institute for Biochemistry and Molecular Biology, University of Southern Denmark, Denmark.

The National Health Data Science Sandbox is a team project spanning five Danish universities with the aim of enhancing and supporting training and research in health data science. Key aims include: 1) providing non-sensitive health datasets that resemble real-world, protected datasets but can be shared with low risk to patient privacy, 2) deploying training materials and resources in user-friendly computational environments that health data scientists will use for their sensitive data projects in Denmark, and 3) supporting training, data exploration, and prototyping by students and researchers using advanced computing resources (HPC). Our team is pairing anonymous public, simulated or synthetic health datasets with modern tools and analysis pipelines for analyzing omics datasets, health records and more. Packaged with user guides, these training modules are also used for courses and workshops as well as self-guided study. The Sandbox platform is being deployed on Danish academic supercomputers Computerome and UCloud, with training modules also available freely on GitHub. We're currently deploying omics-focused modules and investigating the safe generation of synthetic health data from electronic health records. Please contact us if you're interested in using the Sandbox or collaborating on a training module or works.

A8. Towards a unified framework for prediction of extremely imbalanced big data

Sara Nielsen

Technical University of Denmark

The problem of binary prediction of imbalanced data is relevant in many fields, and many strategies and specific algorithms have been developed to overcome it e.g. resampling methods, ensemble techniques and different cost-sensitive approaches. However, these strategies show varying results depending on e.g. the number of observations and the degree of imbalance. The performance seems to differ significantly between datasets, and especially the performance of the classifiers decreases as the amount of imbalance increases.

This calls for a unified framework, which we aim to develop by using more data-centric methods that not only focus on the imbalance in the response variable, y , but on the conditioned probability given the data matrix, X . We focus on problems where the degree of imbalance is extreme ($<1\%$) and aim to address some of the most common issues such as overlapping classes and small disjuncts, which often result in poor predictions.

We focus on density-based resampling which can benefit from a data-centric approach since conditional probabilities may help automatically adjust strategies for imbalance. We experiment with combining density-based resampling with aspects from active learning to find the most informative data samples.

A9. NextGP: A Julia package for next generation genomic prediction tools

Emre Karaman

Center for Quantitative Genetics and Genomics, Aarhus University

NextGP is an open source project aims at providing quantitative geneticists advanced statistical models and necessary analysis tools to reveal complex genetic networks for understanding the genetic architecture of complex traits, and for making better predictions of future phenotypes in animals, humans and plants. One of its key contributions to the field is its automated pipeline for genomic predictions integrating data from different biological layers. Among others, a family of novel graphical methods for inference of complex networks, some novel genomic prediction methods, and traditional linear mixed model analysis have been implemented. It uses Bayesian approach in parameter estimations, in most cases, a Gibbs sampling procedure is run. The NextGP package is written purely in an open source language, Julia.

- The DDS2022 may increase the awareness of researchers from other disciplines the current applications and some unmet needs of data science applications in genomic analysis of complex traits.
- NextGP is constantly being improved, and publicly available.
- Although its early version is being used by some researchers at QGG and in dairy industry, it has not been properly introduced to the scientific community.
- I hope that collaborations can be achieved at DDS2022, and help me to improve functionalities of NextGP.

A10. Machine learning methods for dynamic risk prediction of perianal fistulas in Crohn's disease

Heidi Søgaard Christensen (1,2), Tine Jess (1), and Martin Bøgsted (1,2,3,4)

(1) Center for Molecular Prediction of Inflammatory Bowel Disease, PREDICT, Department of Clinical Medicine, Aalborg University

(2) Clinical Data Science Group, Department of Haematology, Aalborg University Hospital

(3) Clinical Cancer Research Center, Aalborg University Hospital

(4) Department of Clinical Medicine, Aalborg University.

Crohn's disease is a chronic inflammatory bowel disease with a heterogeneous disease course. A common complication of Crohn's is perianal fistulas, causing symptoms like pain and fecal incontinence and entailing an increased risk of specific cancers as well as major surgery. When a patient is diagnosed with Crohn's disease, it is currently not possible to predict whether the patient will develop perianal fistulas. The patient's risk may depend on many factors and plausibly change over time as the disease evolves. In this project we will investigate whether machine learning approaches can be used to provide updated risk estimates of a patient's risk of developing perianal fistulas each time the patient has contact with the hospital. Both static and dynamic predictors, such as genetics, family history, disease history, biomarker values, demographics and lifestyle, will be considered. This work will thus involve a wide variety of heterogenous and longitudinal data, including register data from the Danish nationwide registries, genotype data, and a detailed treatment data base from the North Denmark Region.

A11. Data Science as a part of Pharmaceutical Product Design

Johan Boetker, Anders Madsen, Natalja Genina, Jukka Rantanen

Department of Pharmacy, University of Copenhagen

Data science is becoming omnipresent in all parts of the pharmaceutical product design, with examples ranging from molecular level insight into active compounds to process monitoring and control of full-scale production systems. We give an overview of active research topics in the Manufacturing and Materials group at Department of Pharmacy (University of Copenhagen):

Computer vision for quality control of tablet coatings. Determining amount of active substance in drug products using support vector machine regression models for process analytical technology purposes. Devising data-enriched edible pharmaceuticals where quick response codes provides the possibility for encapsulating information in a digital format on a single dosage unit and deep learning for crystal structure determination.

We envision that data sciences will be an enabling element in designing improved future pharmaceuticals.

A12. Drug target family databases enabling data science and data-driven drug design

David Gloriam and Albert Kooistra

University of Copenhagen



A majority of approved drugs and agents in clinical trials target one of two major drug target families, G protein-coupled receptors (GPCRs) or kinases (1,2). These drug target families have online field hubs, GPCRdb (3) and KLIFS (4) that have become indispensable infrastructures for tens of thousands of researchers. This poster will provide an overview of key reference data, analysis tools, data-driven experiment design, and data deposition.

The latest additions include state-specific structure models using AlphaFold-MultiState (5) and new ligand/drug resources (6). Current development focuses on enabling data-driven drug design on the molecular level by recombining ligand fragments across drug targets using unique residue-level data representations. Both online resources also greatly facilitate data science analyses using AI/ML by providing vast structured and integrated sequence-structure-function data.

References

1. Hauser, A.S., Attwood, M.M., Rask-Andersen, M., Schioth, H.B. and Gloriam, D.E. (2017) Trends in GPCR drug discovery: new agents, targets and indications. *Nat. Rev. Drug Discov.*, **16**, 829-842.
2. Santos, R., Ursu, O., Gaulton, A., Bento, A.P., Donadi, R.S., Bologa, C.G., Karlsson, A., Al-Lazikani, B., Hersey, A., Oprea, T.I. *et al.* (2017) A comprehensive map of molecular drug targets. *Nat. Rev. Drug Discov.*, **16**, 19-34.
3. Kooistra, A.J., Mordalski, S., Pandey-Szekeres, G., Esguerra, M., Mamyrbekov, A., Munk, C., Keseru, G.M. and Gloriam, D.E. (2021) GPCRdb in 2021: integrating GPCR sequence, structure and function. *Nucleic Acids Res.*, **49**, D335-D343.
4. Kanev, G.K., de Graaf, C., Westerman, B.A., de Esch, I.J.P. and Kooistra, A.J. (2021) KLIFS: an overhaul after the first 5 years of supporting kinase research. *Nucleic Acids Res.*, **49**, D562-D569.
5. Heo, L. and Feig, M. (2022) Multi-state modeling of G-protein coupled receptors at experimental accuracy. *Proteins*, 2021.2011.2026.470086.
6. Pándy-Szekeres, G., Caroli, J., Mamyrbekov, A., Kermani, A.A., Keserű, G.M., Kooistra, A.J. and Gloriam, D.E. (2023) GPCRdb in 2023: State-specific structure models using AlphaFold2 and expansion of ligand resources. *Nucleic Acids Res.*

A13. The Privatization of AI Research(-ers): Causes and Potential Consequences – From university-industry interaction to public research brain-drain?

Roman Jurowetzki (AAU), Daniel S. Hain (AAU), Juan Mateos-Garcia (NESTA, UK), Konstantinos Stathoulopoulos (NESTA, UK), Stefano Bianchini (Univ. Strasbourg, FR), Kevin Wirtz (Univ. Strasbourg, FR)

We study the growing transition of AI / ML researchers from academia into industry research labs looking at what drives such transitions as well as the effects it has on the research output. We find that such transitioning researchers are among the most productive and impactful among their peers. Results also indicate that after transitioning into industry, researchers experience a slight over-time decline in their impact - compared to very similar peers remaining in academia.

Our findings highlight the importance of strengthening the public AI research sphere in order to ensure that the future of this powerful technology is not dominated by private interests.

A14. A new approach for Topic Modeling based on Transformer Models and NER

Roman Jurowetzki, Hamid Bekamiri, Daniel S. Hain

Aalborg University Business School, Denmark

In this study, by combining Named-entity recognition (NER) and document embedding-based clustering, we propose a new method for finding and describing topic clusters in extensive document collections. Topic modeling approaches like Latent Dirichlet Allocation (LDA) or more traditional variants based on matrix factorisation (e.g., Latent Semantic Analysis or Cortex) are commonly used to identify latent patterns within text corpora and present the most prominent terminology describing such groupings. Two significant drawbacks of topic modeling approaches are that they focus solely on word-cooccurrence without considering context and the extent and frequency of words within each category and the difficulty of topic interpretation without deep domain expertise. While topic modeling with Transformer models (such as BERT or SBERT) was proposed as a way of addressing the first challenge, word-based interpretation remains problematic. We propose an approach for addressing the second challenge by extracting keywords based on a NER model that detects domain-relevant keywords. We then use c-CF-IDF to find the most relevant and important NER keywords within each cluster. Using research abstracts within the field of neuroscience as a source of textual data, we demonstrate all steps of the proposed topic modeling algorithm.

A15. Discovery of human signaling systems v2.0 – pairing peptides to G protein-coupled receptors

Kay Schaller

University of Copenhagen, Department of Drug Design and Pharmacology

G protein-coupled receptors (GPCRs) are major drug targets, being involved in over a quarter of all FDA-approved drugs. However, there are still over 100 GPCRs with unknown physiological ligands, and these so-called orphan receptors have an immense hidden therapeutic potential. The experimental testing of all orphan receptors against all possible ligands remains unfeasible and computational strategies are needed to limit the set of receptor-ligand pairs. In this work, we develop a comparative genomics and machine learning approach for the identification of novel peptide-protein interactions. Building up from previous work based on gene co-evolution, we will implement a refined approach utilizing residue co-evolution of the receptor-peptide pairs and state-of-the-art structural modeling to evaluate the pairing. To narrow down the human proteome of many thousand proteins to relevant candidates, we will develop a machine learning model, trained on known GPCR peptide ligands and their evolutionary data. We will apply our computational screening approach to all orphan GPCR receptors in search of new physiological peptide ligands. Computationally predicted hits will be sent on to experimental collaborators for testing. We hope that our work will lead to the identification of novel physiological ligands of GPCRs, contributing to the understanding of human disease and opening up new therapeutic potential.

A16. A neural network alternative to non-negative matrix factorization for mutational signature extraction in cancer genomics

Ida Egendal^{1,4,5}, Rasmus F. Brøndum^{1,4,5}, Inge Søkilde Pedersen^{3,4,5}, Karen Dybkær^{2,4,5}, Martin Bøgsted^{1,4,5}

1) Clinical Data Science Group, Department of Haematology, Aalborg University Hospital, 2) Research Laboratory, Department of Haematology, Aalborg University Hospital, 3) Department of Molecular Diagnostics, Aalborg University Hospital, 4) Clinical Cancer Research Center, Aalborg University Hospital, 5) Department of Clinical Medicine, Aalborg University

Knowledge of the mutational signatures present in a cancer genome can reveal which mutagenic processes played an active role in the development of cancer in a specific patient, and thus potentially reveal personalized treatment options.

The composition of mutational signatures is currently determined using non-negative matrix factorization (NMF) which assumes that the mutagenic processes act linearly on the genome and without interaction. However, a recent study revealed that the co-occurrence of two different DNA repair errors in a genome results in a third, distinct signature that cannot be expressed as a linear combination of said processes. This calls for a more flexible extraction model that can account for the non-linearity that occurs when signatures co-occur on a genome. This study compares NMF to a single-hidden-layer and fully connected autoencoder in their ability to rediscover latent signatures on simulated data and their ability to reconstruct unseen data on simulated and real-world-data. This is to our knowledge the first study to do so quantitatively.

Based on this comparison, extensions of the autoencoder allowing a more flexible, non-linear extraction of mutational signatures are discussed.

A17. Is the USA a Shangri-la for European scientists? Inferring the causal effect of transatlantic mobility using quasi-experimental methods in repurposed data

Benjamin C. Holding^{1*}, Claudia Acciai¹, Jesper W. Schneider², Mathias W. Nielsen^{*1}

¹*Department of Sociology, University of Copenhagen, Øster Farimagsgade 6, 1353 Copenhagen, Denmark.*

²*Danish Centre for Studies in Research and Research Policy, Department of Political Science, Aarhus University, Bartholins Allé 7, 8000 Aarhus C, Denmark.*

In this poster, we present our usage of causal inference methods (difference-in-difference with Coarsened Exact Matching) within a repurposed extract of a global bibliometric database. We elucidate whether European researchers who moved to the United States show a benefit in future research performance. Location of individual researchers was determined using publication metadata, and individual level information such as gender was estimated using a gender-inference algorithm. The results show substantial increases in the publishing rates and scientific impact of transatlantic migrants per year post move (Cohen's d range: 0.19 – 0.40). Much of the observed boost in citation- and journal impact associated with mobility was attributable to changes in employment prestige. The study will be of interest for researchers who are interested in repurposing existing data for novel intentions. It will also be of interest for those wanting to know more about quasi-experimental methods within causal inference.

A18. DELPHI – Data Environment for Life science, sensitive Personal data, and Health Investigations

Jesper Roy Christiansen

Computerome

One of the key challenges facing academic and clinical researchers working in the life sciences is access to data. This poster will describe DELPHI, a solution which addresses this challenge by providing a secure and compliant research environment designed for sensitive data.

The purpose of DELPHI is to facilitate easy and transparent access to sensitive data, in addition to enabling data sharing between research projects. The DELPHI infrastructure is hosted at Computerome, which allows DELPHI research projects to leverage the compute, storage, and software resources of a supercomputing cluster.

The poster will provide an overview of the DELPHI solution, with a focus on the data access workflows.

A19. Speech and Multi-Sensory Data Modeling for Child and Youth Psychiatry

Sneha Das,

DTU Applied Mathematics and Computer Science

Advanced machine learning and data analytic techniques for speech and multi-sensory modeling have demonstrated potential for application in healthcare and psychiatry. Given the general dearth of resources in mental health-care and well-being, this is a positive development as psychologist and psychiatrist can better allocate resources towards patients with the aid of technology and artificial intelligence (AI). Furthermore, advancement in sensor technology (camera, microphones, biosensors) has led to efficient data collection mechanisms, that can be leveraged for improved management of mental disorders. We present techniques for speech and biosignal modeling for real-world data under resource constrains. Further, from our work on the associations between obsessive compulsive disorder (OCD) and speech and multi-sensory data modeling, we highlight the approaches available for modeling in psychiatry and insights on the existing challenges.

A20. Scholia - displaying a knowledge graph of science

Finn Nielsen

Technical University of Denmark

As a large knowledge graph Wikidata contains scientific data and metadata about publications. Scholia is a web application that with Linked Data technology displays Wikidata data in dynamically created interactive pages with tables and graphs. The user of Scholia is able to view scholarly profiles of individual researchers, research organization, overview of for instance software, taxons, chemicals and general topics, or combinations of items, e.g., machine learning researchers in Copenhagen. Scholia enables relatively easy creation of new publication items in Wikidata via DOI information or arXiv information. Scholia is created as an Open Source application and available from <https://scholia.toolforge.org/>.

A21. Challenges in the Energy Sector

Niels Kjeldsen (Energinet), Jesper Kronborg Jensen (Energinet), Mette Gamst (Energinet)

The increasing amount of renewable energy in the energy system provides a number of very significant challenges. On our poster we highlight a number of key challenges the energy sector and especially the transmission system operators face. Solving these challenges are key for a successful transition to an energy system based on renewable sources – further the challenges are prime opportunities for the application of data science and in many cases the data to solve the problems are freely available.

A22. Changes in pupil size during listening as potential biomarkers of the hearing status – A classification study

Patrycja Lebiecka-Johansen^{1,2,3}, Adriana A. Zekveld^{1,2}, Dorothea Wendt^{3,4}, Thomas Koelewijn⁵, Afaan I. Muhammad^{1,2}, Sophia E. Kramer^{1,2}

¹Amsterdam UMC, Vrije Universiteit Amsterdam, Otolaryngology – Head and Neck surgery, Ear & Hearing, De Boelelaan 1117, Amsterdam, the Netherlands; ²Amsterdam Public Health, Quality of Care, Amsterdam, the Netherlands; ³Eriksholm Research Centre, Snekkersten, Denmark; ⁴Department of Health Technology, Technical University of Denmark, Denmark; ⁵Department of Otorhinolaryngology/Head and Neck Surgery, University Medical Center Groningen, University of Groningen, Groningen, The Netherlands

Speech understanding in noise may be effortful, especially for people with hearing impairment. Research has shown that hearing impaired (HI) listeners, to compensate for their reduced audibility, may be allocating effort differently than normal hearing (NH) listeners. From the literature we know that effort can be reflected in changes in the pupil size during listening. However, little is known of pupil features specifically sensitive to the listener's hearing status. Here, we investigated whether a set of pre-defined pupil features can be used to classify the hearing status, and whether classification performance depends on the listening condition (e.g., with varying signal-to-noise ratio (SNR), or with specific SNR). We trained the Elastic Net classifier, tested its performance with the Matthew's Correlation Coefficient, as well as identified the relative importance of pupil features.

Reliable classification of the hearing status was possible (MCC range 0.2-0.3, $P < 0.05$). Despite of documented sensitivity of pupil to SNR, specifying this factor worsened classification performance. The analysis of misclassifications and feature distributions revealed that even if not provided directly, the classifier might have encoded and utilized information about the varying SNR from the pupil features. Finally, a combination of features was necessary to classify the hearing status.

A23. Using Deep Generative Models for Atomic Structure Solution of Metal Oxide Nanoparticles from Pair Distribution Function Data

Ulrik Friis-Jensen (Department of Chemistry, Nano-Science Center and Department of Computer Science, University of Copenhagen), Frederik L. Johansen (Department of Computer Science, Department of Chemistry and Nano-Science Center, University of Copenhagen), Andy S. Anker (Department of Chemistry and Nano-Science Center, University of Copenhagen), Emil T. S. Kjær (Department of Chemistry and Nano-Science Center, University of Copenhagen), Raghavendra Selvan (Department of Computer Science and Department of Neuroscience, University of Copenhagen), Kirsten M. Ø. Jensen (Department of Chemistry and Nano-Science Center, University of Copenhagen)

My poster outlines the research I do on the application of machine learning in chemistry, specifically using generative models for solving materials science problems. The project uses a Conditional Variational Autoencoder (CVAE) architecture to do atomic structure solution of metal oxide nanoparticles from Total Scattering and Pair Distribution Function (PDF) data. The nanoparticle dataset is represented as graphs and the CVAE therefore uses a Graph Neural Network (GNN) to encode the chemical information into a latent space. This latent space is then analyzed in depth to help explain what the model has learnt and quantify the similarity of different types of nanoparticles.

I think this poster would be interesting for all DDS2022 attendees who wants to know more about deep generative models and/or explainable machine learning. It would also be interesting for the attendees who wants to know more about the specific challenges related to applying machine learning to chemistry problems. Lastly it would be both interesting and a great experience for me to present my work at a larger conference and to a broader audience than I usually have the chance to.

A24. Residential Mobility and Segregation in Denmark

Louis Boucherie (DTU, Danmarks Statistik), Lasse Mohr (DTU, Danmarks Statistik) and Sune Lehmann (DTU, KU)

Using large scale registry datasets we constructed the network of co-habitation in Denmark since 1980. We identify groups of people using graphs embedding and community detection algorithms. These communities are significant because they coincide with socioeconomic indicators (real estate value, monthly income, education).

The poster will include maps of Denmark that indicate area of segregation as well as plots of the latent space. The maps have a very fine resolution (over 3 million addresses).

This study also uncovers a scaling law that holds over a large order of magnitude. The ratio between the distribution of residential moves distance and the distribution of distance between houses scales as one over the distance (from 0.1km to 1000km). The poster will feature the scaling law and simulations using a toy model.

This research is interesting for DDS2022 attendees because it combines machine learning techniques (graph embedding) with pure data science (registry data). Moreover, I am eager to create a poster because this project has a strong visual component. The poster will be aesthetically attractive due to the interplay between the geographical position of the addresses and their geometry in the latent space (some colleagues already asked me to print the plots to decorate their office).

A25. Genomic prediction of winter-survival in perennial ryegrass (*Lolium perenne* L.), and the effect of genotype x environment interactions at the margin of the species distribution

Natasha H. Johansen, Andrea Bellucci, Pernille B. Hansen, Torben Asp, Guillaume P. Ramstein.

Center for Quantitative Genetics and Genomics, Aarhus University (AU), Denmark

The effect of genotype x environment interactions (GxE) on different agronomic traits has previously been documented for a range of important crops, but few studies on GxE have been carried out on perennial ryegrass (*Lolium perenne* L.), despite its high production potential. Commercial breeders are interested in increasing production of this grass in Northern Europe. However, the forage crop is currently not well-adapted to the colder and more variable environments that some of these regions experience, and thus the grass experiences extensive winter damage. The aim of this study is hence to investigate and compare the performance of 352 perennial ryegrass accessions in a multi-environment trial, in the effort to identify genetic variation that enhances winter survival, which in turn will facilitate the selection of superior breeding material. In this study we will use and compare the traditional reaction norm approach to the more recent enviromic-aided genomic prediction approach. Both approaches allow for the incorporation of GxE interactions, while the more recent enviromic-aided genomic prediction approach includes ecophysiological information that in turn is used to describe the quality of the environment with respect to the given crop's abiotic tolerance and requirements. The GxE models will be tested under different breeding scenarios. Prediction ability and accuracies will be compared across models, wherein both additive and non-additive effects are included. Conclusively, the results of this study should provide more in-depth knowledge about perennial ryegrass's adaptability and level of phenotypic plasticity.

A26. Prediction of Type 2 Diabetes using Machine Learning on Electronic Health Records

Martin Bernstorff^{1,2}, Lasse Hansen^{1,2,3}, Kenneth Enevoldsen^{2,3}, Andreas A. Danielsen^{2,4}, Søren D. Østergaard^{1,2}

1. *Department of Affective Disorders, Aarhus University Hospital – Psychiatry, Aarhus, Denmark*
2. *Department of Clinical Medicine, Aarhus University, Aarhus, Denmark*
3. *Center for Humanities Computing, Department of Culture and Society, Aarhus University, Aarhus, Denmark*
4. *Psychosis Research Unit, Aarhus University Hospital - Psychiatry, Aarhus, Denmark*

Background: Type 2 diabetes (T2D) is common among patients with mental illness and may be overlooked in clinical practice. Early identification of patients at risk of T2D may lead to more timely diagnosis and treatment of T2D. Electronic health records likely contain information on the risk of T2D, which may be leveraged through machine-learning.

Methods: We aimed to predict T2D using electronic health record data on laboratory results, medications, diagnoses etc. A total of 120.000 patients treated for mental illness in the Psychiatric Services of the Central Denmark Region from 2013 to 2021 were included in our data. Time-series were "flattened" to represent each prediction time and a variety of machine learning models were trained to predict incident T2D within the 5 years following each hospital visit.

Results: The area under the receiver operating characteristic curve for the preliminary model (XGBoost) was 0.84 for predicting T2D at the level of a hospital visit. The accuracy decreased as the time interval between prediction time and outcome increased, but remained clinically relevant even with a 5-year time interval.

Conclusion: We trained a model for predicting T2D among patients with mental illness based on data from electronic health records. Model performance was sufficient for clinical utility and we will therefore pursue implementation.

A27. Clinical Proteomics Data Science Tutorial

Jacob Fredegaard Hansen, Ole Nørregaard Jensen, Veit Schwämmle

*Department of Biochemistry and Molecular Biology, University of Southern Denmark, DK-5230 Odense M
University of Southern Denmark*

The National Health Data Science Sandbox, a data science infrastructure project, contains non-sensitive health data and tools to facilitate the education and training of health data scientists in a supercomputing environment. This infrastructure is currently under construction at UCloud and Computerome, and here we describe the proteomics-focused training module in development.

The clinical proteomics module will contain state-of-the art software tools for a complete proteomics pipeline including FragPipe, MaxQuant, PDV, SearchGUI, and PeptideShaker. Annotated proteomics data analysis workflows will be provided to facilitate training in computational clinical proteomics analyses (both self-study and guided courses). An implementation of AlphaFold, used for prediction of the 3D structure of a protein based on its amino acid sequence, is also in progress along with other deep learning approaches. Together, these tools could potentially provide valuable insights of personalized treatment procedures of individuals and are key components of training in proteomics-based research and development.

We plan for this training module to be deployed in early 2023 as the 'Proteomics Sandbox' on UCloud, which provides high-performance computing power in a user-friendly digital research environment with an intuitive graphical interface. We seek feedback as well as proposals/collaborations that could expand the module's clinical proteomics content and applications.

A28. The PSYchiatric clinical outcome prediction (PSYCOP) cohort: leveraging the potential of electronic health records in the treatment of mental disorders

Søren Dinesen Østergaard

Aarhus University

The quality of life and lifespan are greatly reduced among individuals with mental illness. To improve prognosis, the nascent field of precision psychiatry aims to provide personalised predictions for the course of illness and response to treatment. Unfortunately, the results of precision psychiatry studies are rarely externally validated, almost never implemented in clinical practice, and tend to focus on a few selected outcomes. To overcome these challenges, we have established the PSYchiatric Clinical Outcome Prediction (PSYCOP) cohort, which will form the basis for extensive studies in the upcoming years.

PSYCOP is a retrospective cohort study that includes all patients with at least one contact with the psychiatric services of the Central Denmark Region in the period from January 1, 2011, to October 28, 2020 ($n = 119\,291$). All data from the electronic health records (EHR) are included, spanning diagnoses, information on treatments, clinical notes, discharge summaries, laboratory tests, etc. Based on these data, machine learning methods will be used to make prediction models for a range of clinical outcomes, such as diagnostic shifts, treatment response, medical comorbidity, and premature mortality, with an explicit focus on clinical feasibility and implementation.

Discussions: We expect that studies based on the PSYCOP cohort will advance the field of precision psychiatry through the use of state-of-the-art machine learning methods on a large and representative data set. Implementation of prediction models in clinical psychiatry will likely improve treatment and, hopefully, increase the quality of life and lifespan of those with mental illness.

A29. Combining chemistry knowledge and Graph Neural Networks Towards Interpretable Molecular Property Models

Adem R.N. Aouichaoui (DTU), Fan Fan (DTU), Jens Abildskov (DTU), Gürkan Sin (DTU)

QSPR models are models capable of inferring the property molecules through their structural information and are essential in many fields; biotechnology, pharma, thermodynamics, and safety/environmental assessment. Graph Neural Networks are currently revolutionizing the field of molecular science by allowing task-specific feature construction from a geometrical representation of the molecule and correlating this to the desired target property. In these models, molecules are represented as a graph with the nodes representing atoms and the edges representing the chemical bonds. However, such models in their present form are not physics informed and lack the aspect of interpretability. This could lead to accurate prediction but for the wrong reasons (clever Hans's effect) and might reduce the wider application and acceptance of these models, especially in a field dominated by a first-principle understanding of the phenomena. To remedy this, we integrate chemistry knowledge in the form of functional groups with the attention mechanism to provide insights into the substructure of the molecule with the most important to the resulting prediction. We demonstrate that the insights provided by the model are consistent with those from a thermodynamic and chemistry understanding of the properties.

A30. Uncovering coherence networks in the human brain using directional statistics

Anders Stevnhoved Olsen

DTU Compute

The complexity of the human brain arises from segregated areas with distinct function or anatomy integrating their activity through cross-brain connectivity. Conventional methods for uncovering brain networks in functional imaging data are limited due to:

1. Using correlation coefficients containing a mix of phase and (potentially spurious) amplitude information.
2. Preimposing structure by using an atlas to reduce dimensionality of data, thereby restricting results to a poorer resolution.
3. Performing computations on time-series averages, thereby assuming signal stationarity.

Here we assess global synchronization networks in the resting brain by extracting the leading eigenvector of instantaneous phase difference maps, a procedure known as Leading Eigenvector Dynamics Analysis (LEiDA). The leading eigenvectors are conventionally grouped into networks using (Euclidean) k-means clustering in a low-dimensional representation. Here we cluster the leading eigenvectors in voxel-space using a mixture of Watson distributions to account for unit norm and sign ambiguity of eigenvectors, a fundamentally different approach.

Preliminarily, we have achieved centroids consistent with anticorrelated functional brain networks described in the literature on data from the Human Connectome Project. We aim to proceed with experiments for selecting the number of clusters and evaluating cluster expression statistics.

A31. Utilizing Domain Knowledge and Data Science in Chemical Process Modelling

Peter Jul-Rasmussen (DTU), Xiaodong Liang (DTU), Jakob Kjøbsted Huusom (DTU)

Across the Chemical and Biochemical industry, the focus on improving digital infrastructure has increased the accessibility of process data, making data-driven process modelling more appealing. However, truly “big-data” is still often not available for Chemical processes causing problems in the applicability of Machine Learning models. To mitigate these problems hybrid-models are introduced, in which domain knowledge is combined with Machine Learning. In hybrid-models, the well-known phenomena such as physical laws can be included using classical deterministic models, while unknown or uncertain phenomena can be modelled using Machine Learning. Typically, the Machine Learning models are trained to predict parameters for the deterministic model or to estimate the error in the deterministic model by training on the residuals. In this work, different approaches for applying hybrid-models are investigated using selected case studies. The focus in the modelling is on developing predictive models for the case studies, which are robust with regards to process noise, low measurement frequency, outliers, and process disturbances. The hybrid-models can be used both in offline applications for optimization purposes, and in online applications in process forecasting.

A32. Properties and transitions of mesoscale convective organisation during EUREC4A using unsupervised learning

Leif Denby

University of Leeds

The representation of shallow tradewind cumulus clouds in climate models accounts for the majority of inter-model spread in climate projections, highlighting an urgent need to understand these clouds better. In particular, their spatial organisation appears to cause a strong impact of their radiative properties and dynamical evolution. The precise mechanisms driving different forms of convective organisation which arise both in nature and in simulations are, however, currently unknown.

Using unsupervised learning to train a deep neural network to autonomously identifying regimes of convective organisation in satellite observations, we will show results from analysing: a) what the radiative properties of different forms of organisation are, b) what atmospheric characteristics coincide with different forms of organisation and c) what transitions occur when following air-masses along Lagrangian trajectories. Specifically, we find: a) net radiation changes significantly between different forms of organisation, b) agreement with previous studies on the importance of boundary layer wind-speed and to some degree atmospheric stability, and c) we are able to succinctly capture what transitions occur between regimes.

A33. Vocal markers of neuropsychiatric conditions: assessing the generalizability of machine learning models and their clinical applicability

Alberto Parola a, b, c, Astrid Rybner a, Emil Trenckner Jessen a, Marie Damsgaard Mortensen a, Stine Nyhus Larsen a, Arndis Simonsen b, d, Jessica Mary Lin a, b, Yuan Zhou e, Huiling Wang f, Shiho Ubukata g, Katja Koelkebeck h,i, Vibeke Bliksted b, d, Riccardo Fusaroli a, b, I

a Department of Linguistics, Cognitive Science and Semiotics, Aarhus University, Aarhus, Denmark

b The Interacting Minds Center - Institute of Culture and Society, Aarhus University, Aarhus, Denmark

c Department of Psychology, University of Turin, Turin, Italy

d Psychosis Research Unit - Department of Clinical Medicine, Aarhus University, Aarhus, Denmark

e Institute of Psychology, Chinese Academy of Sciences, Beijing, China

f Department of Psychiatry, Renmin Hospital of Wuhan University, Wuhan, China

g Department of Psychiatry, Kyoto University, Kyoto, Japan

h LVR-Hospital Essen, Department of Psychiatry and Psychotherapy, Hospital and Institute of the University of Duisburg-Essen, Essen, Germany

i Center for Translational Neuro- & Behavioral Sciences (C-TNBS), University Duisburg-Essen, Germany

I Linguistic Data Consortium, University of Pennsylvania, USA

Machine learning (ML) approaches are a promising venue for identifying vocal markers of neuropsychiatric disorders, such as schizophrenia. While recent studies have shown that voice-based ML models can reliably predict diagnosis and clinical symptoms of schizophrenia, it is unclear to what extent such ML markers generalize to new speech samples collected using a different task or in a different language: the assessment of generalization performance is however crucial for testing their clinical applicability.

In this research, we systematically assessed the generalizability of ML models across contexts and languages. We trained ML models of vocal markers of schizophrenia on a large cross-linguistic dataset (4 languages: Danish, German, Chinese, Japanese) of patients with schizophrenia and controls. We developed a rigorous pipeline to minimize overfitting, including cross-validated training set and Mixture of Experts models. We tested the generalizability of the models on: (i) different participants, speaking the same language (hold-out test set); (ii) different participants, speaking a different language. Model performance was comparable to state-of-the-art findings when trained and tested on participants speaking the same language (out-of-sample performance). Crucially, however, the models did not generalize well - showing a substantial decrease of performance (close to chance) - when tested on new languages.

A34. Robust spatiotemporal actin filament disentanglement using a network theoretic framework

Isabella Østerlund 1,2, Staffan Persson 1 and Zoran Nikoloski 2,3

1: Department of Plant and Environmental Sciences, University of Copenhagen, 1871 Frederiksberg C, Denmark, 2: Bioinformatics, Institute of Biochemistry and Biology, University of Potsdam, 14476 Postdam, Germany, 3: Systems Biology and Mathematical Modeling, Max Planck Institute of Molecular Plant Physiology, 14476 Postdam, Germany

The cytoskeleton provides and supports essential cellular functions. In plant cells, microtubules guide cellulose synthesis, whereas actin provides the basis for secretion and cytoplasmic streaming. Actin cytoskeleton dynamics are difficult to compute. This is due to the complex dynamics of actin filaments and bundles. Therefore, a new tool to quantify and understand spatiotemporal behavior of the actin cytoskeleton is needed to address cytoskeletal complexity.

Here, we propose an automated image-based framework, GraFT, that: (1) builds a network-based representation of the segmented and enhanced filament-like structures based on pre-processed images; (2) identifies individual filaments by untangling the resulting network. To this end, individual filaments are defined based on the longest paths in a constrained depth first search, taking account of angles along the path; (3) tracks filaments over time by solving a linear assignment problem using the information of spatial location of each individual filament.

We show that GraFT provides robust identification and tracking of actin filaments in plant cells. The tool measures several actin characteristics, including filament length, intensity, angles, and bendiness of the actin cytoskeleton. Importantly, GraFT may be used to determine and compare properties of individual filaments and bundles in different time-resolved imaging in cells of *Arabidopsis thaliana*.

A35. Publicly Available Privacy-preserving Benchmarks for Polygenic Prediction

Meindert Witteveen

Aarhus University

Recently, several different approaches for creating polygenic scores have been developed and this trend shows no sign of abating. However, it has thus far been challenging to determine which approaches are superior, as different studies report seemingly conflicting benchmark results. This heterogeneity in benchmark results is in part due to different outcomes being used, but also due to differences in the used set of genetic variants, data preprocessing, and other quality control steps. As a solution, a publicly available benchmark for polygenic prediction is presented here, which allows researchers to both train and test polygenic prediction methods using only summary-level information, thus preserving individual privacy. We show, using simulations and real data, that with our approach, model performance can be estimated with accuracy, using only linkage disequilibrium (LD) information and genome-wide association summary statistics for target outcomes. Finally, we make this PGS benchmark - consisting of 8 outcomes, including somatic and psychiatric disorders - publicly available for researchers to download. We believe this benchmark can help establish a clear and unbiased standard for future polygenic score methods to compare against.

A36. Real-time process monitoring with label scarcity

Davide Cacciarelli ^{a, b}, Murat Kulahci ^{a, c}, John Sølve Tyssedal ^b

a Department of Applied Mathematics and Computer Science, Technical University of Denmark, Kgs. Lyngby, Denmark

b Department of Mathematical Sciences, Norwegian University of Science and Technology, Trondheim, Norway

c Department of Business Administration, Technology and Social Sciences, Luleå University of Technology, Luleå, Sweden

The amount of unlabeled data collected in production processes has amplified because of the increased availability and usage of sensors. However, acquiring the corresponding quality information, or label, is not straightforward as it often necessitates the use of expensive testing equipment. To this extent, we first present a novel unsupervised learning approach based on autoencoders. It will be shown how statistical process control can be facilitated by regularizing the bottleneck of autoencoder networks. The second part of the work will be focused on active learning. Indeed, in many cases we are interested in fitting a predictive model that relates the easy-to-measure process variables to the hard-to-measure quality information. When the labeling task is very expensive and we can only afford a certain number of labels, it becomes crucial to determine how to efficiently select the most informative data points to be used for training a model. In particular, we analyze the online active learning case, which depicts a situation where the decision about the acquisition of the label for a data point needs to be taken in real time. To this extent, we propose a new approach for the online active linear regression framework inspired by the optimal experimental design theory.

A37. First Principal Models and Neural Networks for Defining Metabolic Capacity from Continuous Glucose Measurements (CGM) As a Tool for Personalized Nutrition.

Zhi Ye

University of Copenhagen

Tailored dietary recommendation holds a promise for promoting health at the individual level. We will leverage the use of wearable devices for continuous glucose measurements (CGM), meal images, and physical activities to achieve this goal. Mathematical modeling of glucose dynamics is an essential tool to understand fundamental aspects of personal life metabolism. The homeostasis of glucose keeps its active stabilization from outside disturbances and helps us to simulate the dynamic system which takes external disturbances including glucose increase via eating and decrease via exercise, inner glucose consumption, and the internal self-regulation mechanism.

A38. The meaning in the machine - Inductive exploration of group discourse using dynamic word embeddings

August Lohse and Thyge Ryom Enggaard

UCPH (SODAS)

We introduce and validate a new approach to explore the polarization of meaning between groups of people, by examining the differences discourse in their texts inductively. We conceptualize this as an issue of (mis)translation of meaning, and thus examine the words that are hardest to translate to themselves between the two corpora using a range of static and dynamic models. We test and validate our approach using a synthetic dataset with word swapping and apply our method to two large corpora from Reddit, comparing the meaning of words amongst political opposites from R/Democrats and R/Republicans. We find that the (mis)translation approach provides both intuitive, but also surprising results, showing that the meaning of some very politically charged words may not have as contested meanings as we may think. We also show that it is possible to track politicized words over time, by examining the evolution of mistranslation between the corpora. Our new approach provides a framework for how to study the polarization of meaning, that importantly not only provides an estimate of how much groups differ, but also allows us to examine the differences inductively, thus giving us access to knowledge about how the different groups understand the world.

A39. Causal representation learning for out-of-distribution generalization

Thea Brusch

DTU

Most modern machine learning algorithms rely on the assumption that training and test data follow the same distribution. This means that the models often fail to generalize when the environments – and thus the distributions - change. This scenario may for instance occur if we are recording medical data at multiple different hospitals and want to train a model for a specific task that generalizes in new, unseen hospital environments. Since available medical data is already sparse, it is essential to solve this issue of out-of-distribution generalization if we wish to enhance the field of medical AI.

We propose to use causal representation learning to solve the issue of out-of-distribution generalization. This approach relies on the assumption that the train and test data are sampled from the same underlying causal model – allowing us to learn representations that are invariant across environments.

The first part of my PhD will focus on learning causal representations for better out-of-distribution generalization – especially for medical time series data. The research is still in the early phase and the poster will therefore be a presentation of the ideas for the first part of my PhD.

A40. Health Data Science Sandbox project and interactive supercomputing training

Samuele Soraggi (Bioinformatics research center, Aarhus University)

Alejandro Herrera (HeaDS, Copenhagen University)

Jennifer Bartell (HeaDS, Copenhagen University)

Jesper Christiansen (Danmarks Tekniske Universitet)

Sander Valentin (Aalborg University)

Jonas Hansen (Syddansk Universitet)

We support health data science training and research in Denmark by supporting the implementation and execution of courses with SDU's interactive supercomputing platform. We show in this poster our interactive app-based approach to distribute training and research courses, tools and datasets. Come by our poster to know how you can benefit from (or contribute to) our FREE and OPEN-SOURCE data science applications!

A41. Prediction of Anxiety and Depression in ICD-Patients using Machine Learning Algorithms

Ali Ebrahimi

University of Southern Denmark

Heart failure is a complex, chronic condition and the end stage of most heart diseases. Many patients with heart failure are implanted with an implantable cardioverter defibrillator (ICD) in order to prevent sudden cardiac death. In the last decade, the management of ICD patients has seen a shift from face-to-face outpatient visits to remote monitoring alternatives. Despite the clinical and economic advantages of remote monitoring of patients, the decrease in patient-nurse and patient-physician interactions has made it more difficult to identify vulnerable patients. Since anxiety and depression have been linked to a higher rate of morbidity, it is crucial to identify those patients to proactively provide them with the correct treatment. The main goal of this project is to develop explainable predictive models based on machine learning (ML) algorithms to track down cardiac patients who are at increased risk of developing anxiety and depression and to help clinical staff to have a better understanding of patients' anxiety and depression risk. Findings indicate promising results of ML models developed based on collected data from 478 patients in the AQUIRE-ICD RCT for a duration of 24 months.

A42. An Explainable Machine Learning Approach for the Detection of Lung Cancer in Denmark

Abdolrahman Peimankar

University of Southern Denmark

Lung cancer (LC) is now the leading type of cancer in Denmark, accounting for 23.4% of cases among all the cancer incidences in 2020. In addition, it has the highest number of deaths compared to other cancer types with around 21% mortality rate. Patients are usually diagnosed at a late stage of the disease, which results in a limited number of patients eligible for curative intended treatment (~35%). Consequently, most of the patients are left with only palliative care and treatment options. It is therefore crucial to refer patients for diagnostics as soon as a suspicion of LC arises, which will result in a substantial and growing economic burden on the healthcare system. The aim of this project is to propose an explainable and interpretable machine learning (ML) model which can be implemented in clinical settings to help general practitioners to make better and timely diagnosis of lung cancer. The obtained results from a cohort of 5587 Danish subjects (i.e., 1372 LC and 4215 non-LC) show a great potential of ML models to detect LC from blood sample analysis.

A43. Mapping Complex Technologies via Science-Technology Linkages; The Case of Neuroscience-A transformer based keyword extraction approach

Daniel Hain (AAU), Roman Jurowetzki (AAU), Mariagrazia Squicciarini (UNESCO)

We present an efficient deep learning based approach to extract technology-related topics and keywords within scientific literature, and identify corresponding technologies within patent applications. Specifically, we utilize transformer based language models, tailored for use with scientific text, to detect coherent topics over time and describe these by relevant keywords that are automatically extracted from a large text corpus. We identify these keywords using Named Entity Recognition, distinguishing between those describing methods, applications and other scientific terminology. We create a large amount of search queries based on combinations of method- and application-keywords, which we use to conduct semantic search and identify related patents. By doing so, we aim at contributing to the growing body of research on text-based technology mapping and forecasting that leverages latest advances in natural language processing and deep learning. We are able to map technologies identified in scientific literature to patent applications, thereby providing an empirical foundation for the study of science-technology linkages. We illustrate the workflow as well as results obtained by mapping publications within the field of neuroscience to related patent applications.

A44. Machine learning models application in proteomics

Pawel Palczynski (SDU), Tobias Rehfeldt (SDU), Richard Röttger (SDU, Veit Schwämmle (SDU).

We are working on applying machine learning models for protein identification to greatly improve the current database search based methods. Within tandem mass spectroscopy area most of the focus is placed upon MS2, i.e. the narrower and more detailed part of the overall data collected. We, however, are using the MS1 spectra which allows for potentially more accurate identification of known peptides but also for discovery of new peptides that generally are being left unlabeled and unknown. There are certain features within the MS1 spectra such as the isotopic pattern distribution that can be picked up by the deep learning models and thus enhance the prediction rate. This approach can further be used to integrate with the spectrometers to improve the MS2 data quality by running the preliminary predictions on MS1 mid-experiment.

A45. A self-supervised model of the brain for psychiatric phenotyping - a project outline

Fabian Mager¹, Lars Kai Hansen¹, Bjørn Ebdrup²

¹DTU Compute, Department of Applied Mathematics and Computer Science, Technical University of Denmark, Kgs. Lyngby, Denmark

²Center for Neuropsychiatric Schizophrenia Research (CNSR) and Center for Clinical Intervention and Neuropsychiatric Schizophrenia Research (CINS), Copenhagen University Hospital, Mental Health Center Glostrup, Glostrup, Denmark

In Denmark, mental disorders make 25% of the total disease burden, with yet increasing prevalence. In psychiatry, little data paired with complex clinical traits and weak pathological signals make brain imaging research challenging. In the field of machine learning, scarcity of labelled data and richness of unlabeled data has engendered self-supervised learning paradigms. We aim to develop a self-supervised model of the brain using structural magnetic resonance imaging data. We want to fine-tune the self-supervised model to address psychiatric phenotypes and use explainable AI to visualize its hidden representations.

A46. Efficient Closed Form Updates for Archetypal analysis

Anna Emilie Jennow Wedenborg ¹ and Morten Mørup¹

¹DTU Compute, Department of Applied Mathematics and Computer Science, Technical University of Denmark, Kgs. Lyngby, Denmark

One of the biggest challenges for modern data scientists is to efficiently and reliably extract information facilitating human understanding of complex and huge data sets. This project will focus on a prominent and easy interpretable data science method called Archetypal analysis (AA) that if appropriately designed is capable of identifying distinct characteristics, so-called archetypes, and how observations can be described in terms of convex combinations of these distinct characteristics forming polytopes in high-dimensional real-world data.

The aim of this project is to develop an efficient scalable algorithm for Archetypal Analysis by exploring convex sub-problems with simple closed form solutions. The implementation will be optimized for large datasets, where fast and stable convergence is a major concern. Specifically, we explore alternating optimization, as originally suggested by Cutler and Breiman[1] equipped efficient fast-nonnegative least squares optimization (fnnls) and sequential minimal optimization (SMO) updates.

The proposed approach is compared to existing prominent algorithms for AA in terms of stability and speed. We further explore how the inference procedure admits efficient uncertainty characterization when appropriately combined with bootstrapping.

This work was supported by the Danish Data Science Academy, which is funded by the Novo Nordisk Foundation (NNF21SA0069429) and VILLUM FONDEN (40516).

[1] A. Cutler and L. Breiman, "ARCHETYPAL ANALYSIS," 1993

Poster walks

During the poster session, participants can join one of the groups that will tour a subset of posters (poster walk). Each walk will visit up to five posters, grouped by common topics, and is expected to last between 20-25 minutes. During the walk, each poster presenter will give a 3-minute poster pitch followed by a couple of minutes for discussion. Please see below the different posters that each group will visit. All groups will be gathered at Multihuset (where the poster session will take place) and will start their walk at 13:00.

Group #1

- A4. Representation learning for image-based characterization of industrial flocculation processes
- A23. Using Deep Generative Models for Atomic Structure Solution of Metal Oxide Nanoparticles from Pair Distribution Function Data
- A24. Residential Mobility and Segregation in Denmark
- A29. Combining chemistry knowledge and Graph Neural Networks Towards Interpretable Molecular Property Models
- A30. Uncovering coherence networks in the human brain using directional statistics

Group #2

- A3. Large language models via automated machine learning for epidemiological research: a case study and method comparison from the British National Child Development Study
- A10. Machine learning methods for dynamic risk prediction of perianal fistulas in Crohn's disease
- A16. A neural network alternative to non-negative matrix factorization for mutational signature extraction in cancer genomics
- A19. Speech and Multi-Sensory Data Modeling for Child and Youth Psychiatry

Group #3

- A22. Changes in pupil size during listening as potential biomarkers of the hearing status – A classification study
- A26. Prediction of Type 2 Diabetes using Machine Learning on Electronic Health Records
- A28. The PSYchiatric clinical outcome prediction (PSYCOP) cohort: leveraging the potential of electronic health records in the treatment of mental disorders
- A37. First Principal Models and Neural Networks for Defining Metabolic Capacity from Continuous Glucose Measurements (CGM) As a Tool for Personalized Nutrition

Group #4

- A33. Vocal markers of neuropsychiatric conditions: assessing the generalizability of machine learning models and their clinical applicability
- A39. Causal representation learning for out-of-distribution generalization
- A41. Prediction of Anxiety and Depression in ICD-Patients using Machine Learning Algorithms
- A42. An Explainable Machine Learning Approach for the Detection of Lung Cancer in Denmark
- A45. A self-supervised model of the brain for psychiatric phenotyping - a project outline

Group #5

- A1. Unlocking to power of our data
- A6. Optimal Control of Residential Energy Storage Systems
- A13. The Privatization of AI Research(-ers): Causes and Potential Consequences – From university-industry interaction to public research brain-drain?

- A17. Is the USA a Shangri-la for European scientists? Inferring the causal effect of transatlantic mobility using quasi-experimental methods in repurposed data

Group #6

- A2. NLP Danske Bank
- A21. Challenges in the Energy Sector
- A38. The meaning in the machine - Inductive exploration of group discourse using dynamic word embeddings

Group #7

- A20. Scholia - displaying a knowledge graph of science
- A27. Clinical Proteomics Data Science Tutorial
- A36. Real-time process monitoring with label scarcity
- A40. Health Data Science Sandbox project and interactive supercomputing training

Group #8

- A18. DELPHI – Data Environment for Life science, sensitive Personal data, and Health Investigations
- A7. The National Health Data Science Sandbox for Training and Research
- A8. Towards a unified framework for prediction of extremely imbalanced big data
- A46. Efficient Closed Form Updates for Archetypal analysis

Group #9

- A12. Drug target family databases enabling data science and data-driven drug design
- A15. Discovery of human signaling systems v2.0 – pairing peptides to G protein-coupled receptors
- A31. Utilizing Domain Knowledge and Data Science in Chemical Process Modelling
- A32. Properties and transitions of mesoscale convective organisation during EUREC4A using unsupervised learning

Group #10

- A34. Robust spatiotemporal actin filament disentanglement using a network theoretic framework
- A44. Machine learning models application in proteomics
- A14. A new approach for Topic Modeling based on Transformer Models and NER
- A43. Mapping Complex Technologies via Science-Technology Linkages; The Case of Neuroscience-A transformer based keyword extraction approach

Group #11

- A5. Improving image understanding with deep multimodal fusion
- A11. Data Science as a part of Pharmaceutical Product Design
- A9. NextGP: A Julia package for next generation genomic prediction tools
- A25. Genomic prediction of winter-survival in perennial ryegrass (*Lolium perenne* L.), and the effect of genotype x environment interactions at the margin of the species distribution
- A35. Publicly Available Privacy-preserving Benchmarks for Polygenic Prediction



Danish
Data Science
Academy