UNIVERSITY OF CANBERRA

International Research Workshop
Data Science and AI & Robotics (DSAIR24)

University of Canberra, Australia
19 July 2024
Venue: Building 6 Room B45, University of Canberra

Artificial intelligence will reach human levels by around 2029. Follow that out further to, say, 2045, and we will have multiplied the intelligence - the human biological machine intelligence of our civilization - a billion-fold. Ray Kurzweil

Data science is the extraction of knowledge from data, using techniques drawn from statistics, machine learning, and database management. Jim Gray

Schedule:

Fiona Dyer, Associate Dean Research, Faculty of Science & Technology, UC

09:30-9:40 Welcome: Prof. Janine Deakin, DVC Research and Innovation (Acting), UC

Shuangzhe Liu **Session 1**

09:40-10:20 Dietrich von Rosen, Swedish University of Agricultural Sciences, Sweden
Safety Belt Regression applied to the Growth Curve Model.
10:20-10:50 Kenichi Satoh, Shiga University, Japan
Applying Non-negative Matrix Factorization with Covariates to the Longitudinal Data as Growth Curve Model
10:50-11:20 Georgy Sofronov, Macquarie University
Change-point Detection Problems

11:20-11:40 Morning Tea

Dharmendra Sharma **Session 2**

11:40-12:10 Graham Williams, The Australian National University
Data Science with Privacy through Data Pods
12:10-12:30 Matthew Andrews and Priya Rajgarhia, Xaana.Ai
Transforming Finance – Data and AI in Finance
12:30-12:50 Ghazal Bargshady, University of Canberra
Multimodal Sensing Intelligent Computational Approach to Assess Awareness of Drivers
12:50-13:10 Min Wang, University of Canberra
Privacy-Preserving Brain-Computer Interfaces

13:10-14:00 Working Lunch

Maryam Ghahramani **Session 3**

14:00-14:20 Quanling Deng, The Australian National University
A Mathematical Perspective on Neural Networks
14:20-14:40 Maleen Jayasuriya, University of Canberra
A Journey from Probabilistic Robotics to Deep Learning
14:40-15:00 Liang Zheng, The Australian National University
Data-centric Computer Vision

15:00-15:20 Afternoon Tea

Ghazal Bargshady **Session 4**

15:20-16:00 Richard Duncan, University of Canberra
Modelling Species Range Contractions
15:40-16:20 Susan Hartono and Theo Niyonsenga, University of Canberra
Analysis of Dyadic Data: The Actor-Partner Interdependence Model within the Structural Equations Modelling Approach
16:00-16:30 Andrew Grant, University of Sydney
Noise-augmented Directional Clustering with Application to Genetic Association Data
16:30-17:10 Alan Welsh and Insha Ullah, The Australian National University
Exploring the Counterintuitive Benefits of Overfitting in Linear Models with Noise

17:10-17:15 **Concluding remarks:** Kumudu Munasinghe, Deputy Dean, Faculty of Science & Technology, UC

https://www.canberra.edu.au/about-uc/faculties/SciTech

UNIVERSITY OF CANBERRA

Abstracts

Dietrich von Rosen Dietrich.von.Rosen@slu.se

Title: Safety Belt Regression Applied to the Growth Curve model.

Abstract: Safety belt regression has been established. In this talk ideas are extended to also hold for the GMANOVA model (Growth Curve model). It will be shown how the GMANOVA model can be transformed to a MANOVA model with additional information. The approach works when the restrictions in the model will not be affected by the additional information.

Kenichi Satoh Kenichi-Satoh@biwako.shiga-u.ac.jp

Title: Applying Non-negative Matrix Factorization with Covariates to the Longitudinal Data as Growth Curve Model

Abstract: Using Non-negative Matrix Factorization (NMF), the observed matrix can be approximated by the product of the basis and coefficient matrices. Moreover, if the coefficient vectors are explained by the covariates for each individual, the coefficient matrix can be written as the product of the parameter matrix and the covariate matrix, and additionally described in the framework of Non-negative Matrix tri-Factorization (tri-NMF) with covariates. Consequently, this is equal to the mean structure of the Growth Curve Model (GCM). The difference is that the basis matrix for GCM is given by the analyst, whereas that for NMF with covariates is unknown and optimized. In this study, we applied NMF with covariance to longitudinal data and compared it with GCM. We have also published an R package that implements this method, and we show how to use it through examples of data analyses including longitudinal measurement, spatiotemporal data and text data. In particular, we demonstrate the usefulness of Gaussian kernel functions as covariates.

Georgy Sofronov Georgy.Sofronov@mq.edu.au

Title: Change-point Detection Problems

Abstract: Change-point problems (or break point problems, disorder problems) can be considered one of the central issues of statistical data science, connecting asymptotic statistical theory and Monte Carlo methods, frequentist and Bayesian approaches, fixed and sequential procedures. In many real applications, observations are taken sequentially over time, or can be ordered with respect to some other criterion. The basic question, therefore, is whether the obtained data can be divided in several homogeneous segments that have different statistical models. The change-point problem arises in a wide variety of fields, including bioinformatics, biomedical signal processing, speech and image processing, seismology, industry (e.g. fault detection) and financial mathematics. In this talk, I will give an overview of various approaches to change-point detection methods.

Graham Williams Graham.Williams@anu.edu.au

Title: Data Science with Privacy through Data Pods

Abstract: Over the past decades the World Wide Web has taken a direction not envisaged by its creators. As data scientists and now with the emergence of large language models, we have driven the concept of centralising the collection of data and then deriving considerable value from that data. Often we have had little regard for the owners of that data and the data owners had little insight to what the data has been used for. In recent years we have begun to see a swing back to concerns for privacy and personal governance of our data. In this talk I will present the resurgent vision that our individual data remain under our individual control. I will illustrate this through recent projects, one initiated by a Far North Queensland indigenous health clinic to improve health care delivery, and another supporting clinical trials for a medical device helping patients with depression. These projects deploy the Solid Pods protocol with a software app ecosystem developed and delivered using Flutter. For a glimpse visit https://solidcommunity.au.

Matthew Andrews and Priya Rajgarhia, Matt.Andrews@xaana.ai

Title: Transforming Finance – Data and AI in Finance

Abstract: Finance teams are typically home to tasks that are time consuming, complex and data intensive. AI is poised to play a leading role in transforming finance operations in organisations of all sizes. Priya Rajgarhia, Data & AI Scientist from Canberra firm Xaana.Ai along with Matthew Andrews, Partner at Xaana, will talk about how AI will transform how AI teams operate & the technology driving the transformation.

Ghazal.Bargshady Ghazal.Bargshady@canberra.edu.au

Title: Multimodal Sensing Intelligent Computational Approach to Assess Awareness of Drivers.

Abstract: Human error is one of the primary causes of car crashes. Fatigue, stress, sleeplessness, brain workload, or secondary tasks (e.g., conversation with a co-passenger, using handheld devices) can cause distractions during driving, leading to significant economic and injury losses. Drivers' situational awareness is affected by fatigue and distraction, leading to unsafe driving. In Australia, the fatality rate of road accidents per 100,000 population in 2022 was ranked 18th out of the 31 OECD nations. ACT Road Safety currently recognises distracted and dangerous driving as one of the five strategic priority areas in 2024. Over five years, road fatalities in the ACT have risen by 50%, mainly attributed to driver distraction and human error. A significant issue highlighted in a report is the use of mobile devices while driving, with over 39,000 instances detected, alongside other distractions like adjusting radios and engaging in conversation despite efforts by traffic cameras. On the other hand, the world will soon see self-driving cars. Current self-driving cars are conditionally automated vehicles (L3-SAE) and have not yet become fully autonomous (L5-SAE). In current L3-SAE automated vehicles, when the vehicle cannot handle a situation, it asks the driver to take control through a Take-Over Request (TOR). To do this, autonomous cars need to detect driver distraction and situational awareness (SA) in real-time before sending a Take-Over Request.

This project aims to develop a novel multimodal deep learning model to recognize driver situational awareness in real-time by collecting and analysing features from drivers' behavioural and physiological cues as well as vehicle kinematics data. The data will be collected in driving simulation scenarios. The developed model could be applied in both non-self-driving vehicles and self-driving vehicles to detect driver awareness in real-time when distraction occurs, thereby contributing to enhanced road safety.

Min Wang Min.Wang@canberra.edu.au

Title: Privacy-Preserving Brain-Computer Interfaces

Abstract: In this talk, we explore the rapidly evolving field of pattern recognition and privacy preservation for brain signals. The presentation will delve into advanced techniques for analysing and decoding neural activity, highlighting the potential for brain signal-based identification and authentication systems. I will introduce our work on feature extraction, machine learning and pattern recognition in learning brain biometrics, alongside the critical challenges of safeguarding the sensitive neural data. Emphasizing the importance of privacy, the talk will also introduce our recent work on protecting against data breaches to ensure the confidentiality of brain signal information.

Quanling Deng Quanling.Deng@anu.edu.au

Title: A Mathematical Perspective on Neural Networks

Abstract: Understanding the intricacies of machine learning models has become paramount, especially within the realm of explainable artificial intelligence (XAI), which has witnessed an array of methodologies. Despite the advancements, many existing XAI techniques tend to offer localized perspectives, capturing only fragments of understanding. In this talk, I will introduce a series of computational mathematical tools designed to systematically analyze variable importance and their interactions within a set of equally good neural networks, colloquially known as the Rashomon set.

Maleen Jayasuriya Maleen.Jayasuriya@canberra.edu.au

Title: A Journey from Probabilistic Robotics to Deep Learning

Abstract: This presentation explores the transformative shift in robotics from traditional probabilistic methods to deep learning-based approaches, with a focus on embodied AI and end-to-end manipulation. The advent of transformer networks and their application in robotics is highlighted, showcasing their profound impact on the field. This presentation delves into how this transition affects human-robot collaboration and teaming, particularly in the domains of motion prediction and manipulation, as well as its application in industries such as the construction sector. The implications of these advancements on the future of robotics and humanity are discussed, emphasizing the potential for more intuitive and human-centered robot interactions.

Liang Zheng Liang.Zheng@anu.edu.au

Title: Data-centric Computer Vision

Abstract: Computer vision research depends heavily on data and model. While the latter has been extensive designed and studied, we still lack definition and insights of problems associated data. In this talk, I will introduce a few attempts from my group analyzing and improving data. I will discuss how to improve the quality of training data, such that better models can be trained under distribution shifts. I will also talk about how to evaluate the difficulty of the test data, or in other words, the model accuracy, in an unsupervised way. Finally, I will introduce a new way of formatting videos so that motions can be efficiently captured by existing action recognition networks. I will conclude with perspectives and unaddressed challenges in data-centric problems.

Richard Duncan Richard.Duncan@canberra.edu.au

Title: Modelling Species Range Contractions

Abstract: In Australia, many native species have suffered significant population declines and range contractions due to human impacts, climate change, and invasive species. Quantifying range contractions is difficult because we usually lack the large-scale, long-term data required to accurately track changes in species geographical distributions over time. Instead, species occurrence data are often patchy and biased by strong human sampling preferences. Using tools from spatio-temporal disease mapping, I describe a method to model shifts in species distributions over time from patchy occurrence data. The method is implemented in a hierarchical Bayesian framework using autocorrelation to borrow strength from data distributed in space and time to locally smooth outcomes while accounting for biases in sampling effort. The result is reconstructions of species range contractions making full use of the available data to quantify changes.

Susan Hartono and Theo Niyonsenga Theo.Niyonsenga@canberra.edu.au

Title: Analysis of Dyadic Data: The Actor-Partner Interdependence Model within the Structural Equations Modelling Approach

Abstract: Both Structural equation modelling (SEM) and Path analysis (PA) methods are preferred by researchers in social sciences and epidemiology because they estimate the multiple and interrelated dependence in a single analysis. The researcher will have *priori* hypotheses about causal relationships amongst variables (i.e., an *a priori* theoretical model with pathways for causal relationships or path diagram). The Actor-Partner Interdependence Model (APIM) is a particular example of statistical framework for non-independent dyadic data. It focuses on the interrelationship between two individuals within a pair or dyad. APIM is useful in understanding the mutual influence and interactions between dyad members, such as romantic partners, friends, family members, or therapist-client pairs. As in SEM, through simultaneous statistical modelling in APIM, researchers can investigate how each subject's behaviour, characteristics, or conditions within a dyad influence their own outcomes (actor effects) and the outcomes of their partner (partner effects). While multilevel modelling can estimate the APIM coefficients, SEM approach is often preferred due to its flexibility in handling multiple pathways.

Andrew Grant Andrew.Grant1@sydney.edu.au

Title: Noise-augmented Directional Clustering with Application to Genetic Association Data

Abstract: Many commonly used clustering approaches group observations based on Euclidean distances between vectors. I will present an alternative approach for clustering which is based on grouping observed vectors according to their direction from the origin. The proposed method, NAvMix, is based on a mixture model approach and includes a noise cluster that provides robustness to outliers. A motivating application is clustering genetic variants based on their associations with phenotypic traits. Identifying groups of genetic variants with similar phenotypic association patterns can provide insight into underlying biological mechanisms. I will illustrate this by applying NAvMix to genetic variants which are associated with body mass index.

Alan Welsh and Insha Ullah Alan.Welsh@anu.edu.au

Title: Exploring the Counterintuitive Benefits of Overfitting in Linear Models with Noise

Abstract: Traditionally, the bias-variance trade-off has guided model selection in the under-parameterized regime, with the belief that over-parameterisation leads to overfitting and poor generalisation. Yet, emerging studies have revealed the "double descent" curve, where the test error unexpectedly declines in over-parameterised models, thus challenging the traditional bias-variance framework.

Our research offers a distinct perspective by examining the counterintuitive benefits of overfitting in linear models. We study how incorporating noise—arising either from predictors or observations—affects the prediction accuracy of models. This exploration is essential, as irrelevant variables are an unavoidable aspect of practical applications. Yet, their impact is often overlooked or inadequately distinguished from important variables in theoretical discussions on phenomena like the double descent curve and model regularisation techniques, such as ridge regularisation. Our findings explain the mechanics behind the "double descent" curve and propose that overfitting, under certain conditions, may indeed enhance prediction accuracy.

Recently, it has been shown that minimum norm least squares estimation performs shrinkage in the presence of irrelevant predictors and tends to work better in terms of prediction accuracy than ridge regularisation with a positive ridge penalty. Furthermore, empirical evidence suggests that the optimal value for the ridge penalty in ridge regression models could be zero or negative. This paradox challenges standard practice and prompts further investigation. Our analysis shows why employing a negative ridge penalty might be advantageous. This perspective deciphers the previously observed empirical mystery and underscores the role of noise in model performance.