ADSN2024 Conference Proceedings

Day 1: Monday, 02 December 2024

Keynote Speaker

Reimagining Healthcare: Better Outcomes Through Digital Innovation

Prof. Suzanne Robinson, Deakin University

The digital health revolution offers unprecedented opportunities to transform healthcare delivery and improve patient outcomes. By leveraging cutting-edge technologies, including data science and analytics, we can revolutionize how we diagnose, treat, and prevent diseases. This presentation will explore the factors limiting the translation of research into practice, including knowledge transfer, regulatory barriers, and resource constraints. We will also discuss strategies to bridge this gap, such as fostering collaboration, establishing effective knowledge exchange mechanisms, and addressing regulatory hurdles. Furthermore, we will delve into the importance of a skilled workforce in the digital health sector. Healthcare professionals, data scientists, and technology experts must be equipped with the necessary skills to develop and implement innovative solutions. This requires ongoing education, training, and upskilling programs. By harnessing the power of data science and analytics, we can unlock new insights, improve decision-making, and accelerate the development of innovative digital health solutions. This will ultimately lead to better patient outcomes, enhanced healthcare delivery, and a more patient-centred healthcare system.

Invited Speaker

MYLO: Applying Data Science to a Rule-Based Conversational Agent to Support Problem Clarification for Wellbeing and Mental Health

Prof. Warren Mansell, Curtin University

MYLO is a rules-based system that engages the user in a text-based conversation about a current problem. Its architecture is informed by control systems engineering and research evidence on the active ingredients of psychological change. Through MYLO's curious questioning, the user is supported in problem expression (putting their thoughts and feelings into words), problem clarification (exploring the problem in greater depth and detail), and problem resolution (seeing the problem from a new perspective and getting insights). The Curtin Institute for Data Science implemented two rounds of codesigned improvements to MYLO - a progressive web application for smartphones - and an improvement to its database and decision-making and learning algorithms. We report on the impact of these improvements, and our plans for integrating large language models (LLMs) and natural language processing (NLP) within MYLO to generate a hybrid in the next round of improvements ahead of commercialisation.

Oral Presentations

A user-friendly digital prediction tool for dengue prevention

Dr Wala Draidi, The University of Queensland

Dengue, a viral infection transmitted by Aedes mosquitoes, has surged tenfold in the past 20 years, putting 3.9 billion people at risk worldwide. In 2023 alone, over six million cases and 6,000 deaths were reported globally. This rise is driven by factors such as climate change, urbanization, socio demographic shifts, and increased global travel. Despite advances in prevention, including vaccines and Wolbachia-infected mosquitoes, challenges remain, especially in resource-limited settings.

In Vietnam, particularly the Mekong Delta Region, dengue is a leading cause of hospitalization. The Vietnam National Dengue Control Program primarily uses reactive vector control measures, often delayed and less effective. To improve early intervention, recent studies have developed provincial-level dengue forecasting models, but these do not adequately support district-level decision-making, where interventions are implemented.

This study evaluates models for forecasting dengue two months and 3 months in advance at the district level in the Mekong Delta Region. The goal is to develop an early warning system to empower local health systems to take timely action and reduce the impact of dengue outbreaks.

Investigating EEG markers for responsiveness following surgical noxious stimuli under propofol anaesthesia: A case study

Ms Ilandari Deva Vidushani Jayanga Dhanawansa, Monash University

Leveraging EEG to study the neural correlates of responses to noxious surgical stimulation could drive the development of more sensitive anaesthetic monitors for surgery. Motivated by this fact, we analysed a clinical dataset comprising of EEG of 7 patients undergoing surgery under propofol anaesthesia, where responses to verbal commands were recorded. To address the hypothesis that a change in consciousness would lead to a decrease in the frontal alpha power in responders to the noxious stimulus (NOX), we computed the spectral power and global coherence for selected regions of interest. Logistic regression classifiers were developed using these measures to classify patients between responders and non-responders. To investigate spatial activations during cognitive responses to verbal commands, we correlated the band powers of 64 channels with the response timelines around NOX. The classifiers confirmed no discernible segregation between classes. However, in the patient who responded to the verbal commands after NOX, spatial maps revealed significant activations on scalp regions coinciding with the sensorimotor and auditory cortices. Cortical processing despite a lack of awareness implied by unchanged frontal alpha power, indicates connected consciousness (CC) where patients experience external stimuli while not being completely aware. We identified real-time EEG markers for CC during surgery, which could potentially be leveraged in anaesthetic monitors for responsiveness.

Intuitive exploration of high dimensional single cell transcriptomics data using a graphical representation of Singular Value Decomposition.

Dr Kristen Feher, SAiGENCI, University of Adelaide

Gene expression matrices can have dimension of 10^2-10^6 over the cells and 10^1-10^4 over the genes depending on the measurement technology platform. Standard bioinformatics methods tend to have a 'list of parts' output, making it difficult to understand where they fit in the data's multivariate landscape. UMAPs are in widespread use for visualisation of cells, however they can distort the local neighbourhoods in certain circumstances, and the influence of genes on the local cell neighbourhoods is lost.

I present an informative graph based method of visualising the top K components of a singular value decomposition (SVD) simultaneously. It involves creating a scaffold of genes and cells separately using a KNN analogue of mean shift clustering. These scaffolds can be used to construct graph based visualisations of either the genes or cells which can be used instead of

UMAPs for a more faithful representation of multivariate structure, including easy identification of transitional cell types that don't easily fit in a clustering framework. The gene and cell scaffolds can be merged into a biomarker specificity network which is a bipartite graph of cells and genes, derived from the similarity in the SVD basis and is a high-dimensional analogue of a two dimensional biplot. This pipeline can be recursively applied to subsets of cells to describe local structure. It is a general tool that can be adapted for data exploration in non-biological settings.

The WA Cancer Staging Project

Dr Nancy Tippaya, Curtin Institute for Data Science

Cancer is one of the leading causes of disease burden in Australia, highlighting the need to understand population dynamics for healthcare planning. Staging data at population level is essential for effectively evaluating the impact of screening programs, ensuring compliance with treatment guidelines, and assessing outcomes across various population groups in WA. However, in Australia, the collection of such data is limited due to the significant resources required to manually determine cancer staging information at the individual level. In 2021, Cancer Network WA awarded funding to the WA Cancer Staging Project, establishing a collaboration with Curtin University. This aimed to ensure that WA has a robust, contemporary and sustainable cancer care system, including the ability to collect key metrics such as staging information. The project aims to deliver a state-wide, population-based cancer staging system using advancements in data science, Artificial Intelligence, and Machine Learning. Over the past three years, model development has focused on breast, colorectal, and melanoma cancers - three of the five mode common cancers. The project has created training datasets, developed and validated models to automatically process cancer notification reports, and incorporated an "expert-in-the-loop" approach to support the sustainability of the automated staging models. This presentation will provide an overview of the project and its achievements over the past three years.

Predicting Funding Program Change in Aged Care: A Machine Learning Analysis of Frailty Indicators

Dr Bahar Moezzi, Silverchain

Silverchain is the 3rd largest aged and community care provider in Australia. The majority of Clients are funded via The Commonwealth Home Support Program (CHSP), targeting care needs of low complexity, or Home Care Packages (HCP), which targets those with more comprehensive needs. A Client may initially join the CHSP program, however if their condition deteriorates, they may transfer to HCP. We hypothesized that frailty indicators predict such progression. To assess this hypothesis, the data of 153K clients who are on CHSP or converted from CHSP to HCP were wrangled (8K clients in each class). We included the following features: • Gender (sex) • Age • Total minutes of visits while on CHSP And count of • Hospitalizations • Diagnoses • Diagnoses with Charlson Comorbidity Index • Skin tears • Falls in the past 12 months We used an optimized Random Forest classifier. The results were collated and presented to business SMEs with a Power BI dashboard. Out of 22K clients currently on CHSP, 4K clients were predicted to be progressing. We achieved a prediction accuracy of 75%. We found the main predictors to be the number of hospitalizations and diagnoses, total minutes of visits on CHSP, and age. These were used to rank the clients in order of likelihood to progress. When demonstrating the tool to key stakeholders, feedback was highly positive. Our model can facilitate timely progression to a more suitable program and improve best care and financial sustainability at Silverchain.

Statistical versus generative adversarial network synthetic tabular health data generation: who wins?

Dr Yunwei Zhang, Murdoch University; Macquarie University

There has been growing interest in utilising generative adversarial networks to generate synthetic data. Such datasets are extremely useful in various areas and contexts, including when validating model performance, sharing sensitive data, and protecting privacy. While generating synthetic data for these purposes is not new, statistical data simulation approaches have traditionally been used before the development of generative adversarial networks. Will statistical methods in this context become less relevant? Which of these two approaches is better when learning from health tabular data? With these questions in mind, my talk today aims to shed light on the performance comparison of statistical and generative adversarial network synthetic data generation methods on six different health-related real tabular datasets with continuous, binary, and censored survival outcomes from a practical perspective. Our results reveal that either technique generates synthetic datasets that closely resemble the real data structure. From a practical perspective, either method contributes to statistical modelling and shows strong prediction ability. Data pre-processing steps such as scaling variables are necessary for statistical methods to perform well but are not required for generative adversarial network methods. However, generative adversarial network methods are not reproducible. So, who wins? It depends.

Predicting self-harm among psychiatric inpatients

Mr Ken Bredemeyer, UWA

Suicide is a leading cause of death and has profound and lasting impacts on survivors. Repeated non-suicidal self-injury (NSSI) increases the capability for individuals to attempt and commit suicide, and over the long term reduces the predictability of suicidal behaviour. Predicting when individuals are at heightened risk of NSSI is an important strategy in suicide prevention. Predictive modelling can be employed at psychiatric hospitals to identify who is at most risk of committing self-harm. These predictions can be used to target when and where clinical interventions are most needed. We implement models using a time series embedding of daily patient responses, combining admission data and predicting over an n day window. Challenges in developing optimal models include highly imbalanced data (NSSI events are rare), course granularity time series, and a large proportion of missing data. We find that NSSI history (lifetime NSSI and current visit NSSI) most strongly predict future NSSI. Age, suicidal ideation and quality of life are also found to be important predictors. Around 92% of NSSI events are correctly predicted over a 5 day prediction window. False positive rates are typically high when predicting rare events. For every 4 people predicted by our model to commit NSSI during the following 5 days, 1 person did engage in NSSI (according to our test data). This is a substantial improvement over results reported in the psychology literature.

Identifying Inequality through Urban Health Indicators (the AusUrb-HI project)

Dr Aiden Price, QUT

Extreme heat is increasingly recognized as a significant public health hazard. In Australia, heatwaves are the leading cause of death among all natural hazards. The anticipated rise in extreme heat events due to climate change, coupled with a projected population surge to around 49.2 million by 2066, poses a significant challenge to the nation’s health infrastructure. This challenge is compounded by rapid urbanization, densification, and the escalating demand for new housing in Australian urban areas. These factors highlight the critical need for climate-sensitive urban planning and design strategies to mitigate the health impacts of extreme heat.

This study captures areas-in-need using a Heat Health Vulnerability Index, leveraging demographic factors, environmental conditions, and urban morphology to derive comprehensive spatial layers at a fine-grained level (SA1) in selected regions in New South Wales, Australia. The index methodology utilizes multiple PCA factors scaled by explained variability to produce more complex and data-driven indices from up to 44 indicators of heat exposure, population sensitivity, and adaptive capacity, pinpointing regions particularly susceptible to heat and examining the interplay between human health and the built environment.

This study compares underlying characteristics as well as the resulting vulnerability indices to individual health records received from the New South Wales Ministry of Health, facilitated by the Centre for Health Record Linkage (CHeReL) and the Secured Unified Research Environment (SURE). These new indicators and metrics provide critical data for evidence-based policymaking and planning in health and social infrastructure. The outcomes are expected to lead to more liveable neighbourhoods, enhanced health services, targeted interventions, and a reduction in preventable diseases and their associated economic costs.

Identifying children's postures and movements from a single camera

Professor Andrew Lloyd Rohl, Curtin University

Human Activity Recognition (HAR) is a field of study that focuses on identifying specific activities performed by individuals based on data derived from various sensors or video analysis. In recent years, recognising children's activities has attracted significant attention due to its potential applications in monitoring developmental progress, enhancing interactive learning environments, and ensuring children's safety.

One way to achieve this is via video footage. We have collected video of children (approved by the Curtin University Human Research Ethics Office as HRE2022-0157) in a lab at Curtin undertaking a number of postures (e.g. sitting, standing) and movements (e.g. running and stair climbing).

We have developed a workflow to identify the activities of the children.

1. Segment video footage into 2 minute clips
2. Identify and follow children in the footage using a customised YOLO-v8-based child detector.
3. Extract skeleton data from the video frames using 2D pose estimation techniques. This involves identifying key body joints and constructing a skeleton representation of the children in each frame.
4. Classify the extracted skeleton data into postures and movements using Graph Convolutional Network (GCN)-based models. The GCNs are combined with Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN), including Long Short-Term Memory (LSTM) to leverage the spatial and temporal relationships inherent in the 2D skeleton data.

ComFe: Interpretable Image Classifiers With Foundation Models

Ms Evelyn Mannix, University of Melbourne

Interpretable computer vision models can produce transparent predictions, where the features of an image are compared with prototypes from a training dataset and the similarity between them forms a basis for classification. Nevertheless these methods are introduce additional complexity and may require domain knowledge to adapt hyperparameters to a new dataset. Inspired by developments in object detection, segmentation and large-scale self-supervised foundation vision models, we introduce Component Features (ComFe), a novel explainable-by-design image classification approach using a transformer-decoder head and hierarchical mixture-modelling. With only global image labels and no segmentation or part annotations, ComFe can identify consistent image components, such as the head, body, wings and tail of a bird, and the image background, and determine which of these features are informative in making a prediction. We demonstrate the performance of this approach across numerous challenges without need to change model hyperparameters, from standard computer vision datasets such as ImageNet, to detecting biofouling on commercial shipping vessels in the wild.

Poster Presentations

Forecasting Price Volatility in the Australian Agricultural Sector Using HAR and HAR-X Models

Ms Lily Zhang, Queensland University of Technology

The volatility of Australian agricultural commodity prices is crucial for resource allocation and risk management. The aim of this study is to contribute to the existing literature on volatility forecasting for Australian agricultural price assets specifically by improving the following areas. We start by exploring out-of-sample rolling window forecast performance via statistical evaluation criteria using the standard HAR model and HAR-X model with different sets of exogenous predictors. The results show that models that incorporate realized moments outperform those that include sentiment, suggesting that realized moments are more significant for predicting volatility in these commodities.

Optimal profile limits for childhood Type 1 diabetes incidence in Saudi Arabia

Mrs Ahood Alazwari, RMIT University

This work presented three predictive models that are used to develop for the first time profile limits of childhood T1D in Saudi Arabia using local data. Optimisation procedures were used to develop upper and lower limits for childhood T1D incidence, with the goal of achieving by 2030 the T1D levels recommended by the UN and the Saudi Arabian government. The profile limits for childhood T1D incidence are accompanied by optimal yearly values for three significant factors: the number of children born with birth weight over 3.5kg, the number of advanced maternal age at childbirth, and the number of children exposed early to cow's milk (<6 months). The optimal profile limits provide a quantitative guideline for reducing childhood T1D. They include yearly targets for children with a birth weight over 3.5kg, advanced maternal age at childbirth, and early introduction to cow's milk. The findings can assist authorities in making decisions regarding allocating resources to decrease the incidence of childhood T1D using evidence-based interventions.

COKI Open Science Dashboard Framework

Dr Rebecca Handcock, Curtin University

Scaling up a small data science project requires coordination between data analysis and software development, and aspects such as requirements, meetings, licensing, backups, reporting, and publications that are part of a robust and reproducible research lifecycle. We illustrate best practice for setting up a research data science project with a case study of a biomedical project focussing on open science practices.

Project data storage and data processing (ie SQL) is through Google Cloud Platform (GCP). The JIRA project management tool tracks software development, with work sprints tightly coupled to project meeting cycles. Google Workspace tools are used to establish a central repository of meeting notes and project protocols, as these tools are accessible to project staff. Version control of code (ie GitHub) is essential for collaborative and transparent software development, and clear licencing to support open science practices. Releases of code on GitHub are archived to the Zenodo open repository, with a Digital Object Identifier (DOI) for citing the GitHub code repository in publications. The CRediT (Contributor Roles Taxonomy) is used to assign contributions from project staff. Additional outputs are generated with literate programming (eg R + Mermaid diagramming tool) that populates flowcharts with data from GCP storage.

Rigorous naming and versioning protocols link the project across these various tools and has allowed this project to scale robustly.

PortOCELs: A Simulated Dataset for Studying the Impact of IoT on Logistics Processes

Ms Jia Wei, Queensland University of Technology

With the increasing adoption of IoT technologies, especially in the industrial sector, more IoT data is becoming available. IoT data provides real-time insights into operations, enabling companies to monitor their business processes and make data-driven decisions. Business data capturing process executions are stored in event logs. Therefore, to study the impact of IoT on business processes, event logs integrated with IoT data are essential for analysing IoT-enhanced processes. However, access to business and IoT data is often restricted due to confidentiality, and IoT data is stored in external databases, making integration into event logs complex. To address this challenge, we introduce PortOCELs, a collection of simulated event logs capturing the cargo pickup process at a major port in China, including IoT data involved in the processes. Event logs are commonly stored in formats: XES and OCEL. OCEL is better suited for IoT integration as it captures the relationships between objects and their interactions with events. This work presents two simulated OCEL logs: one for an IoT-enhanced cargo pickup process and the other for a traditional process without IoT. These logs were generated through simulation using CPN, with parameters informed by domain knowledge. PortOCELs contribute valuable datasets to the process mining community, enabling future studies on how IoT benefits business processes. We also showcase analyses using these simulated logs, demonstrating their utility.

Integrating Latin Hypercube Sampling into Composite Designs for Enhanced Response Surface Methodology

Ms Despoina Athanasaki, RMIT University

This study presents a novel method for incorporating Latin Hypercube Sampling (LHS) into composite designs within Design of Experiments (DOE) for Response Surface Methodology (RSM), Jessen [1975]. Composite designs are known for their flexibility and efficiency in probing complex relationships between input variables and output responses, Jones et al. [2021]. By integrating LHS, we significantly enhance the exploration of multidimensional parameter spaces, ensuring a more diverse and representative sampling scheme, Deutsch and Deutsch [2012]. Through practical examples and comprehensive analysis, we demonstrate how LHS can innovate composite designs in RSM applications . Our results highlight LHS's potential to optimize processes, improve product quality, and reduce costs across various fields, offering valuable insights for both practitioners and researchers in experimental design and optimization.

Coherent Noise Modelling for Uncertainty Estimates for Inherent Optical Properties

Dr Daniel Marrable, Curtin Institute for Data Science

The interpretation of remote sensing and optical model data is significantly influenced by uncertainties inherent in algorithm parameters and inputs, and in both in-situ and satellite measurements. These uncertainties pose challenges, especially in semi-analytical models used for estimating the phytoplankton population via ocean colour from space-borne optical sensors, in that there is a lack of comprehensive uncertainty quantification in data products retrieved from inverting remotely sensed ocean-colour data, leading to a major gap in current methodologies.

In particular, uncertainties arise from assumptions about the relationship between Remote Sensing Reflectance (R_{rs}) and Inherent Optical Properties (IOPs). Additionally, anthropogenic actions intensify the complex interaction among water constituents, increasing uncertainties in models for inland waters.

Estimating uncertainties is difficult due to challenges in making repeated measurements and the noise generated from diverse environmental and instrument sources. Existing methods, such as analytically derived solutions from optical models and the Monte Carlo estimation method, have limitations and often assume that a single measurement represents the mean value of all possible realisations or that random perturbation of reflectance will emulate natural variations. Monte Carlo methods often perturb measurements by sampling from an estimated distribution (often a normal distribution) without rigorous justification

Enhancing Content Quality Evaluation Reliability for Generic Knowledge with Semantic Factors

Dr Yingnan Shi, UWA

As real-time content evaluation becomes increasingly critical for enhancing user satisfaction and content quality, especially for new users, this study proposes incorporating semantic-based features such as conciseness, concreteness, and topic integrity to improve evaluation accuracy. With recent rapid advancements in natural language processing (NLP), real-time evaluation of semantic features is more feasible. Leveraging advancements in natural language processing (NLP), the study introduces a model that incorporates semantic features using Named Entity Recognition (NER) to build up new semantic feature. Experiments comparing this enhanced model with a traditional syntax-only approach show that semantic features significantly improve the reliability of content quality assessments. Key contributions include a literature review, validation of semantic feature integration, and an experimental demonstration of NER's effectiveness in evaluating user-generated content. These findings provide insights for content platforms to design more accurate reward and promotion systems. Moreover, our plans involve evaluating various mainstream NER algorithms, such as spaCy, Stanford NER tagger, Stanza, Allennlp, Deeppavlov, FLAIR-fast, etc., to establish a benchmark for performance and improve the semantic evaluation of content. Furthermore, a horizontal comparison can help to establish a standard benchmark for evaluating the performance of NER algorithms, another key future development.

DALL-M: Context-Aware Clinical Data Augmentation with LLMs

Mr Chihcheng Hsieh, Queensland University of Technology

X-ray images are essential in medical diagnostics, but their effectiveness is often limited without clinical context. Radiologists frequently find chest X-rays insufficient for diagnosing diseases, requiring integration with comprehensive clinical data. To enhance this context, we introduce DALL-M, a novel technique for clinical data augmentation using large language models (LLMs) to generate synthetic patient context data. This approach enriches datasets with contextually relevant synthetic features, which are vital for training more robust deep learning models.

DALL-M employs a three-phase process: (i) clinical context storage, (ii) expert query generation, and (iii) context-aware feature augmentation. Applied to 799 cases from the MIMIC-IV dataset, DALL-M generated 91 augmented features from an initial nine by synthesizing chest X-ray images and reports. It is the first method to create contextual values based on patient-specific data like X-ray reports, gender, and age, producing new contextual knowledge during data augmentation.

Empirical validation with models like Decision Trees, Random Forests, XGBoost, and TabNET demonstrated significant performance improvements, including a 16.5% increase in F1 score and approximately 25% boosts in Precision and Recall. DALL-M addresses a critical gap in clinical data augmentation, providing a robust framework for generating contextually enriched datasets and improving AI-driven medical diagnostics.

TabAttackBench: A Benchmark for Adversarial Attacks on Tabular Data

Mr Zhipeng He, Queensland University of Technology

Adversarial attacks pose a significant threat to machine learning models by inducing incorrect predictions through imperceptible perturbations to input data. While these attacks have been extensively studied in unstructured data like images, their application to tabular data presents new challenges. These challenges arise from the inherent heterogeneity and complex feature interdependencies in tabular data, which differ significantly from those in image data. To address these differences, it is crucial to consider imperceptibility as a key criterion specific to tabular data. Most current research focuses primarily on achieving effective adversarial attacks, often overlooking the importance of maintaining imperceptibility. To address this gap, we propose a new benchmark for adversarial attacks on tabular data that evaluates both effectiveness and imperceptibility. In this study, we assess the effectiveness and imperceptibility of five adversarial attacks across four models using ten tabular datasets, including both mixed and numerical-only datasets. Our analysis explores how these factors interact and influence the overall performance of the attacks. We also compare the results across different dataset types to understand the broader implications of these findings. The findings from this benchmark provide valuable insights for improving the design of adversarial attack algorithms, thereby advancing the field of adversarial machine learning on tabular data.

Missing Data Imputation using Vine Copula

Ms Sithara Wijekoon, Queensland University of Technology

Missing data is a frequent issue in observational studies, often leading to biased estimates and inaccurate conclusions. Traditional imputation techniques may not adequately preserve the multivariate dependence structures in datasets with complex, nonlinear relationships. To address this, we introduce novel imputation methods that maintain dependency among continuous variables. Our methods utilize two types of vine copulas: canonical vines (C-vines) and drawable vines (D-vines) which effectively capture complex data features and tail dependencies for more accurate imputation. Missing values are imputed by sampling from C- or D-vine models of the conditional distribution of missing data given the observed multivariate data. This approach has notable advantages over traditional methods: it estimates imputed values using multiple simulated conditional values rather than a single sample, and it adapts to various copula families, enhancing the robustness of the imputation. Notably, our C-vine-based method represents the first application of a C-vine copula model in missing data imputation. The methods are designed to handle non-monotonic missing patterns classified as Missing Completely at Random. We compared our methods with existing imputation techniques:

predictive mean matching and the copula-based imputation method (CoImp), through simulations and found our approaches consistently yield lower errors in Mean Absolute Error and Root Mean Square Error.

Geostatistical Regression as a Designed Experimental Procedure

Associate Professor Ewan Cameron, Curtin University & The Kids Institute (formerly called Telethon Kids)

Confounding is a fundamental challenge to the development of interpretable regression models for scientific inference. When the data are spatially indexed, a geostatistical regression approach may be adopted to protect the linear trend estimates against spatially-structured confounding. The expected bias in this setting has been well characterised through eigenvector decomposition of the spatial precision matrix. In this study we develop complementary insights into the mechanisms through which this protection is achieved by framing geostatistical regression as a designed experimental procedure to adjust non-spatial linear regression estimates for spatial confounding. We demonstrate that the effective character of the experimental procedure created by a given covariance function can be radically different for covariates of differing smoothness. To illustrate these ideas we introduce a design-augmented version of restricted spatial regression in which a partial decoupling is created between the design for linear trend estimation and that for creating spatial shrinkage of the residuals. The application of these ideas to the challenge of malaria risk mapping with health data from routine surveillance systems and high resolution satellite-based covariates will be explored in depth.

Hybrid Statistical Algorithmic Approach for Identifying Labour Market Areas

Ms Jamintha Hashini Nisansala Samarakoon Mudiyanselage, Queensland University of Technology

Understanding the boundaries and characteristics of labour market areas (LMAs) is crucial for effective policymaking and resource allocation. LMAs represent regions where job seekers and employers interact, playing a vital role in employment patterns, regional development, and economic growth. Accurate delineation of LMAs is essential to address regional disparities and guide targeted interventions. However, existing methods for delimiting LMAs often struggle with accuracy and efficiency, as they fail to incorporate key factors and require manual boundary adjustments. This research aims to solve these challenges by developing an automated methodology that integrates multiple influential factors, offering a more precise analysis of labour markets. It begins with an evaluation of contemporary methodologies for identifying LMAs, applying them to simulated datasets to form distinct clusters representing individual labour markets. By assessing the strengths and weaknesses of these methods, the study seeks to improve the accuracy of labour market delimitation. The goal is to design a novel hybrid methodology that combines statistical spatial modelling with simulated annealing, providing a flexible and robust framework better suited to modern labour market complexities and evolving conditions for better policy decisions.

Integrative multi-omics and clinical data analysis to predict outcomes and treatment response in Hepatocellular carcinoma

Dr Saurabh Gupta, Curtin University

Hepatocellular carcinoma (HCC), the most common type of primary liver cancer, is a major cause of cancer-related mortality worldwide, and predicting patient outcomes and treatment responses remains challenging. Multi-omics approaches, which combine data from various biological levels, have the potential to uncover novel biomarkers that could guide personalized treatment for HCC patients. This project integrates bulk RNA sequencing, whole-exome sequencing, and clinical data to identify key biomarkers and develop predictive models for patient stratification.

Day 2: Tuesday, 03 December 2024

Invited Speaker

Leveraging the power of social media analytics

Prof. Mingming Cheng, Curtin University

Social media has become an integral part of daily life, transforming the public’s role from passive consumption ("read-read") to active participation ("read-write"). It not only exerts a significant influence on human decision-making processes but also reshapes our societal structures. In this context, leveraging the power of social media analytics has never been more critical and urgent, particularly in addressing global challenges such as climate change and modern slavery.

Oral Presentations

Does Gender Affect Sentiment Analysis?

Mr Oska Dubsky-Smith, The University of Adelaide

Social media platforms like Twitter, Instagram and TripAdvisor have made it easy for people to express their opinions to a wider audience, providing large amounts of textual data for analysis. Opinions often carry emotional undertones that can be challenging to analyse accurately. These emotional tones may vary depending on the reader's gender, raising the question of whether gender affects sentiment analysis. Sentiment analysis mathematically quantifies emotional tones within text, with one of the more straightforward methods being a lexicon-based approach. Despite the popularity of lexicons like VADER, SentiStrength, and AFINN, research on gender-adjusted lexicons remains limited. Most studies focus on the accuracy of generic lexicons without adjusting for gender. Our research aims to fill this gap by investigating if gender affects lexical sentiment analysis by exploring the need for gender-specific lexicons and models. Using Trustpilot reviews (a platform for consumer feedback) we first analyse whether sentiment varies by reviewer gender and then assess if a gender-specific lexicon provides greater accuracy than the generic AFINN dictionary. Our findings reveal that within our constructed gender-adjusted lexicon, females show a more positive sentiment towards positive words and a more negative sentiment towards negative words when compared to males. When this lexicon is then applied to Trustpilot reviews it outperforms the prediction accuracy of the generic AFINN lexicon.

Liberating Petri Nets in Social Network Data Analysis

Mr Ethan Johnson, The University of Adelaide

The rise of social media offers a unique opportunity to study human behaviour in an unfiltered environment. Its influence in modern society enables political and noteworthy people to influence public discussion. While this can lead to positive change, misinformation, as seen in the 2016 US elections or the 2023 Australian Indigenous Voice referendum, can spread. Traditional methods, such as Markov chains, have been used to model the spread of misinformation and user behaviours. However, these approaches do not explicitly model users' concurrent behaviour. Process mining techniques focus on discovering process models with concurrency and are yet to be applied to social networks. We aim to leverage these methods, specifically Petri nets, to analyse users' behaviours by investigating time distributions of users' activities. Petri nets excel at modelling concurrent behaviour, that is, multiple events occuring independently. This is useful as often users see and interact with posts simultaneously. To effectively capture this behaviour, parallel branches need to be added to the model. This is not achievable with Markov chains which require sequential events. Additionally, Petri nets provide a clear visual representation of the network, displaying the flow of information through tokens and transitions. By comparing time distributions found with Petri nets and Markov chains, we found Petri nets to be a viable option for analysing users' behaviours in social networks.

Big Data and AI in Urban Mobility

Dr Lillian Wu, UWA

This presentation explores the application of big data and AI in urban mobility. By leveraging multi-source data and advanced analytics, we analyze how various factors influence traffic patterns and examine the relationship between the built environment and urban mobility. Additionally, we demonstrate the use of AI techniques to manage traffic and enhance travel efficiency. This talk will provide a broad overview of how data-driven approaches are reshaping urban planning and transportation, making cities smarter and more sustainable.

Advancing Traffic Safety Analysis: Multimodal Data Fusion of Tabular and Textual Data

Mr Shadi Jaradat, Queensland University of Technology

In traffic safety analysis, traditional research typically focuses on either tabular or textual data, such as crash narratives, but often fails to harness the combined potential of these data types. Our study introduces an advanced framework for traffic crash analysis through Multimodal Data Fusion (MDF), integrating tabular data with textual narratives and utilizing the latest advancements in Large Language Models (LLMs), specifically Generative Pre-trained Transformer models GPT-3.5 and GPT-4.5. This research uniquely employs zero-shot and few-shot learning methods, alongside in-context labeling with GPT-4, to analyze traffic crashes. Notably, GPT-4.5's few-shot approach achieved accuracies of 99% in predicting crash severity and 98% in identifying driver fault. GPT-3.5 also demonstrated strong capabilities, with Jaccard scores of 73.1 and 82.9 for extracting driver actions and crash factors, respectively. These findings underscore the effectiveness of our MDF framework in extracting actionable insights from both tabular and textual data, showcasing its transformative potential for traffic crash analysis and other domains lacking extensive training datasets. This study marks a significant advancement in the field, setting a solid foundation for future automated data analysis and multimodal applications using LLMs.

Enabling National Infrastructure for Environmental Science through the Sustainable Futures Portfolio

Dr Sebastian Lopez Marcano, Queensland Cyber Infrastructure Foundation

Significant funding is being invested to develop interconnected research infrastructure, from data collection to data reporting, aiming to enhance earth and environmental research and decision-making at a national scale. The Queensland Cyber Infrastructure Foundation (QCIF) recently launched a new portfolio, QCIF Sustainable Futures, to support the development of computing, software, and data solutions across the environment, conservation, agriculture, climate, and urban planning sectors. This portfolio aims to continue developing well-established data infrastructure in alignment with the objectives of the Planet Research Data Commons and the National Research Infrastructure for Australia. Two of these infrastructures, the Wildlife Observatory of Australia (WildObs) and EcoCommons, integrate all aspects of environmental data collection and reporting into two highly interconnected platforms. From enabling systematic data collection and storage of species occurrences collected with camera traps, to modelling species distribution through temporal and spatial scales, both WildObs and EcoCommons improve environmental monitoring to an unprecedented scale. These processing and analysis pipelines will represent the most comprehensive network in Australia's environmental research sector.

Description Length of Time Series Modelling

Mr Antony Mizzi, UWA

Reservoir computing is an efficient method for modelling time series data using linear combinations of the components of chaotic systems. However, recent work has explored the inclusion of higher order combinations of these components. We demonstrate the usefulness of minimum description length (or MDL) when training reservoir computers, particularly when these higher order terms are included and model sizes are increased. MDL is a powerful method for the comparison of different sized models based on their ability to compress data. We use it to select subsets of reservoir components, to prevent over-fitting and to analyse important structures in the complex systems at the heart of reservoir computing.

Loss-based optimal designs for symbolic data

Mr Ahmad Hakiim Jamaluddin, UNSW Data Science Hub; School of Mathematics and Statistics

The increasing availability of large data sets, driven by advancements in computational power, has highlighted the need for efficient data summarisation techniques. Symbolic Data Analysis

(SDA), an emerging field in statistics, offers a framework for aggregating data into symbolic objects, enabling more efficient analysis by reducing computational complexity. Symbolic data, unlike standard point-wise observations, represent internal variations and can take various forms, with a focus on interval- and histogram-valued data in the literature.

In this presentation, we focus on histogram-valued data and address a relatively unexplored area-symbolic design. Specifically, we propose a statistical methodology to guide the process of aggregating raw data into symbolic forms. The key questions we investigate include determining the optimal number of bins in a histogram, their placement, and the most effective quantiles for defining class boundaries. Our approach builds upon the foundational work of new symbolic models by employing a Bayesian framework and a loss-based method grounded in statistical decision theory. This novel methodology offers a systematic approach to symbolic design, enhancing the accuracy and efficiency of data aggregation for large-scale datasets.

Leveraging Dynamical Systems for Data Driven Characterisation of Time Series

Mr Braden John Thorne, University of Western Australia

Characterising the dynamics - periodic, chaotic, stochastic etc. - underpinning a time series is critical for analysis and model building. There is a long history of methods designed for this purpose within nonlinear time series analysis, but they tend to suffer when presented with noise contamination, low data volume or a lack of information about the underlying system. As a result, there is a demand for robust data driven methods. In this presentation we discuss reservoir time series analysis, a novel field of nonlinear time series analysis that uses a class of dynamical systems called reservoir computers - typically used for machine learning - to characterise time series in an unsupervised manner from data alone. These methods show a distinct robustness to time series length and noise contamination. This has facilitated a number of applications including time series classification and concept drift detection, with the proposed methods offering improved performance to existing nonlinear time series analysis methods on both synthetic and experimental data.

Poster Presentations

Meteor Entry Tracking using the Earth Observation Record

Mr Leigh Tyers, Curtin Institute for Data Science

Fireballs are phenomena of significant interest due to their potential insights into atmospheric dynamics and cosmic processes. In a notable event in 2018, the MODIS and MISR instruments on board the NASA Terra satellite imaged an exceptionally large and bright meteor trail over the Bering Sea, also seen by Himawari (Borovicka et al., 2020). This is one of the few known examples of the dust trail of a meteor captured by a remote sensing instrument.

We investigate other fireballs using the geostationary satellites Himawari 8 and 9, which image the full disk of the Earth every 10 minutes. As the dust trails left behind by material ablated from these meteors can be visible for more than 30 minutes after the event, by employing an analysis of historical Himawari 8 & 9 data, we more accurately geolocate multiple other smaller fireball occurrences documented in NASA's CNEOS' fireball catalogue.

Enabling Practice Knowledge through Intelligent Systems: Co-Developing Socio-Technical Systems Assets for the National Healthcare Research Infrastructure

Dr Gnana Bharathy, Australian Research Data Commons (ARDC)

Recently, we co-developed a framework for advanced analytics in healthcare research infrastructure in collaboration with the Australian Data Science Network. This framework, co-created through an extensive co-design process involving environmental scans, interviews, surveys, and workshops, provides comprehensive guidance by capturing the needs and aspirations of researchers. Participants highlighted challenges, concerns, and opportunities across all stages of the advanced analytics lifecycle, including the need for socio-technical infrastructure. This includes training resources, governance structures, guidelines, communities of practice (CoPs), and facilitation support. While there are numerous courses and a handful of practice guidelines in existence, context specific knowledge, imparted in a user-friendly fashion, is in short supply. In an upcoming initiative, we aim to collaborate with research and pedagogical practitioners to co-develop and deliver relevant socio-technical assets as co-investment projects. These outputs will be embedded into virtual labs and resource hubs through engaging tools such as AI co-pilots and intelligent tutors. We invite the community to join us in this effort, whether through creating new contents, assessing or augmenting existing contents, or through designing and developing delivery vehicles via innovative platforms like chatbots or co-pilots. The resulting open-source materials could also be reused in your class rooms.

Advancing spatio-temporal modelling in ecology with synthetic data: an R Shiny app for generating coral reef monitoring data.

Ms Hina Gluza, Queensland University of technology

Long-term monitoring surveys are essential to assess abundance trends across ecological and spatial levels. Spatio-temporal modelling uses monitoring data to predict trends everywhere and attribute drivers of changes. This approach increases the volume of information to interpret enabling more robust assessments of population trend change across broad geographical areas. Model validation methods are challenging due to the change-of-support between observations and model outputs. In coral reef research, additional issues include lack of long-term observations, sparsity of monitoring data, uneven sampling designs across countries and data sharing concerns.

This presentation explores the use of synthetic data to address these challenges and accelerate the adoption of spatio-temporal modeling techniques for reef management. To do this, we create an interactive dashboard to help ecologists generate synthetic data step by step. In our example, the user generates coral cover trajectories based on information from sampling design represented by the size of spatial domain, fixed or random sampling, number of surveyed locations and replicated years, presence of hierarchical spatial scales, effects of disturbances and coral dynamics. The app will be then integrated into an existing modelling pipeline that incorporates various predictive spatio-temporal models, allowing for comparisons of model performance and sampling designs without the need for data science expertise.

Predictive modelling of concrete corrosion in marine environment

Dr Sam Bakhtiari, Curtin University

Corrosion of reinforcing steel in concrete structures poses a significant challenge, particularly in marine environments. Identifying the onset of corrosion is crucial for maintenance and durability. A comprehensive dataset, encompassing various parameters, including chloride profile data, has been meticulously compiled. This dataset draws from field data available in the scientific literature as well as port authorities worldwide. We propose a Bayesian network (BN) model by using the collected dataset to predict corrosion initiation based on chloride ingress into concrete cover. The BN model identifies key factors influencing corrosion initiation, including chloride concentration, penetration depth, and the presence of additives such as silica fume, slag, and fly ash in the concrete mix. Understanding these factors enhances early detection and informs preventive measures. By integrating Bayesian networks with chloride data and considering concrete mix components, our model provides valuable insights for managing corrosion risks in marine structures.

Unravelling Job Advertisements Through Fine-Tuned Language Models

Mr Calvin Pang, Curtin Institute for Data Science

This collaborative research project involving Curtin University's Future of Work Institute, the University of Saskatchewan, and the University of Calgary aimed to explore how the messaging in job advertisements influences the personality characteristics of attracted candidates through the analysis of over 14,000 job advertisements and the personality profiles of 262,000 associated candidates.

To advance this research, support from the Curtin Institute for Data Science was sought to develop a job advertisement categorisation tool to automate the classification of job advertisement text. This was achieved by using an independent dataset of 1,400 job advertisements coded by 10 psychologists to fine-tune a BERT language model in early 2022. The model was trained to classify texts within job advertisements into high-level and low-level codes covering abilities, employment branding, job characteristics, personality, and work activities. The performance of the model demonstrated promise before the advent of large language models such as GPT-3, highlighting the potential of language models to facilitate research by reducing the time required for manual annotation in the field of job advertisement analysis.

Multi-layer network analysis of deliberation in an online discussion platform: the case of Reddit

Mr Tianshu Gao, Curtin University

Online discussion platforms like Reddit have the potential to enhance citizen participation in social and political discourse, contributing to opinion formation and consensus building. However, such platforms also pose risks such as spreading misinformation, increasing polarization, and facilitating online harassment and hate speech. Political scientists emphasize that for a discussion to be considered "deliberative"-and thus beneficial for consensus building-certain conditions must be met.

We propose a two-layer network model to describe discussions on Reddit, capturing both user-to-user interaction and the structure of post threads. In our model, the discussion layer represents the hierarchical network of the discussion thread, while the actor layer captures user-to-user interactions. The inter-layer links capture comment ownership by users. Additionally, we introduce metrics for measuring deliberation, which provide a more comprehensive understanding of the discussion structure and user interactions, allowing for a nuanced evaluation of the deliberative level of a post.

Using a large dataset of posts from highly active subreddits, we applied the two-layer network model to assess the deliberative potential across various communities. We have several interesting findings, one of which is that subreddits focused on sports or specific geographical regions consistently exhibited high levels of deliberation across all metrics.

The Internet of Behaviour: Cybersecurity Implications

Dr Quazi Mamun, Charle Sturt University

The rapid advancements in technology have led to the rise of the Internet of Behaviour (IoB), which expands on the concepts of the Internet of Things (IoT) by analysing user behaviour and its impact on cybersecurity (Fu et al., 2020). As interconnected devices grow, managing their security and privacy has become a significant challenge (McCallam et al., 2017). The Internet of Things (IoT) has brought about new opportunities and challenges for cybersecurity. While it enables new applications and services, it also exposes vulnerabilities that can be exploited by malicious actors. The growing number of IoT devices has made it more complex to operate them safely and securely. Additionally, the Internet of Behaviour, using data collected from interconnected devices, can lead to security and privacy risks. Analysing data collected from IoT devices can provide valuable insights but also poses the risk of exploitation by cybercriminals for sophisticated attacks. The rapid development of IoT devices has led to security vulnerabilities that can be exploited to gain unauthorised access to sensitive information or disrupt critical infrastructure. The safety, security, and privacy threats posed by the Internet of Behaviour are particularly concerning in smart home and city environments. Addressing these challenges requires the development of secure IoT architectures, robust security protocols, and user-centric privacy-preserving mechanisms.

Deep Autoencoder-like Constrained NMF with adaptive weights for Multi-View Clustering

Mr Sohan Dinusha Liyana Gunawardena, Queensland University of Technology (QUT)

Recent advancements in semi-supervised nonnegative matrix factorization-based techniques for multi-view clustering have garnered significant interest due to their ability to leverage partially labelled data. However, the problem of addressing complex non-linear data is seldom considered in the semi-supervised setting. To address these limitations, this paper introduces a novel method called Deep Autoencoder-like Constrained NMF with Adaptive Weights for Multi-View Clustering (DACNMF). DACNMF derives a robust representation matrix through a deep encoding process and subsequently reconstructs the data through decoding. The method effectively integrates both pointwise and pairwise label information, combined with manifold learning, to achieve a discriminative latent representation for clustering. Additionally, DACNMF employs a regularization term to acquire both complementary and consistent information inherent in multi-view data. An efficient algorithm using multiplicative update rules is proposed to solve this optimisation problem. Comparative evaluations on multiple real-world datasets demonstrate DACNMF's superior performance over existing state-of-the-art multi-view clustering methods.

Ensemble methods for cancer survival prediction

Ms Mildred Mmbone Lumwamu, Queensland University of Technology

The Cox proportional hazards model (CPH) is widely utilized in survival prediction for cancer studies, enabling the exploration of simultaneous covariates and quantification of survival differences while controlling for potential confounding effects. The results are robust under circumstances where all assumptions are satisfied. However, the CPH has inherent limitations, such as the assumption of proportional hazards, independence of survival times, inability to capture interactions, and poor performance when covariates are highly correlated, which might lead to overfitting. Ad hoc techniques such as step-wise regression can be used to manually detect non-linear effects, though this can be challenging when multiple interactions exist. Therefore, novel machine learning methods that automatically detect nonlinearity and interaction effects of the predictors are needed. One such method is Random Survival Forests (RSF). While RSF offers a flexible alternative to traditional regression methods for censored data, it has certain limitations, performance of RSF is influenced by the number of events, performing well with a higher number of events and poorly with fewer events. We explore ways to extend this method, including different splitting rules to accommodate these prediction needs and combining the Cox PH and our improved Survival Random Forest method to maximize the benefits of both methods. We apply these methods to a breast and colorectal cancer dataset to predict survival.

Characterizing Information Flow in Social Networks Using an Information-Theoretic Approach

Dr Siu Wai Ho, University of Adelaide

The dynamics of information flow in social networks is central to understanding user interactions. While information-theoretic tools have been used to investigate predictive information in social media user connections, a comprehensive study using different information measures to characterise information flow is lacking. This poster investigates the suitability of cross-entropy and match-length measures. In special cases, we provide a detailed analysis, including a closed-form expression for cross-entropy and a derived probability distribution for match-lengths between two sequences, which challenges the adequacy of these measures. In addition, this poster demonstrates how information flow can be detected using causal discovery techniques. In some cases, transfer entropy reveals the direction and quantifies the information flow between sequences.

Modelling the relationship between zircon grain shape and rock silica content

Dr Taryn Scharf, Curtin University

Zircon is a mineral found in many common rocks, sediments and derivative materials. It is frequently used to study Earth’s crust formation processes and trace the origins of sediments, given its suitability for age-dating and geochemically fingerprinting deep-time geological processes. In addition, the physical characteristics of zircon are influenced by the physical and chemical conditions of the crystal’s growth environment and thus encode geological information. However, quantitative models that predictively link zircon physical properties to the mineral’s geological history, are lacking. We use multiple linear regression to understand the relationship between the mineral’s source rock chemistry and physical shape. Thereafter, we apply neural networks to predict source rock silica content using data that is readily accessible during routine geochronology analyses; namely, zircon 2D shape, internal growth texture, and zircon U and Th content. This study was conducted on a Western Australian dataset with a broad spatial (>2.5 x106 km2), temporal (>3 Gyr) and compositional (47-69% whole-rock silica) range. Our approach demonstrates potential as an inexpensive, non-destructive tool for extracting geochemical information from historical zircon datasets.

Emulating human expertise in rezoning: a case study of the Australian Statistical Geography Standard

Mr Filip Juricev-Martincev, QUT

For reliable and accurate spatial statistical analyses providing community-level insights, large geographical areas must be rezoned into smaller areas. The rezoning process is influenced by qualitative and subjective agency-specific requirements, leading to challenges such as information loss due to aggregation, and limited results comparability between frameworks. This study aims to quantify and compare common rezoning criteria of compactness, homogeneity, and equality, and explain the decisions spatial scientists made when rezoning. This will allow us to assess their importance in rezoning, thus facilitating greater statistical framework comparability and assist in automating this process. The methodology developed in this study relies on the ordinary and Tikhonov regularised least squares. This approach provides a weighted measure of each criterion's contribution, ensuring stability, accuracy, and consistency of a geographic framework. This study's motivating example is the Australian Statistical Geography Standard (ASGS), and its basic spatial unit of analysis, Mesh Block. Further testing was conducted on simulated data. We combined our methodology with an existing aggregation algorithm, HeLP. Using the rezoning criteria with the addition of context-specific criteria, such as adherence to land use, and natural and urban boundaries, we were able to recreate a rezoning schema that attains the same underlying properties as the ASGS. We also provide a heuristic approach for decision-making in rezoning parcel-based systems such as the ASGS.

Long-read sequencing, 3D culture, and machine learning for epigenetic profiling and drug screening in hepatocellular carcinoma

Ms Danamma Kalavikatte, Curtin University

Hepatocellular carcinoma (HCC), a significant global health burden, accounts for approximately 10% of cancer-related deaths worldwide. It typically arises from chronic liver diseases, such as viral hepatitis, alcoholic liver disease, and, increasingly, non-alcoholic fatty liver disease. The complex pathogenesis of HCC necessitates innovative early diagnosis and treatment approaches. This project proposes a multifaceted approach to advance the understanding and treatment of HCC. The first part focuses on the role of epigenetic modifications in HCC. Epigenetic alterations play a crucial role in HCC initiation and progression, offering potential therapeutic targets and diagnostic markers. However, reliable biomarkers to predict future HCC risk remain elusive. Early detection of epigenetic changes could improve disease prevention and intervention, aiding in treatment strategies and therapy response monitoring. Additionally,

leveraging the extensive high-content imaging data of liver cancer organoids generated by the Organoids group led by Dr. Benjamin Dwyer, the second part of this PhD research involves the application of machine learning techniques to advance drug screening and personalised treatment approaches.

Day 3: Wednesday, 04 December 2024

Invited Speaker

Shared Environmental Analytics Facility - Facilitating research, industry and government translation for cumulative impact assessment

Prof. Owen Nevin, Western Australian Biodiversity Science Institute (WABSI)

The Australian Government's commitment to nature-positive outcomes and recent amendments to the Western Australian Environmental Protection Act emphasize the need for better environmental information systems. With increasing demands for transparency and accountability from both government and corporate sectors, there is a growing need for reliable, accessible, and interoperable environmental data. In 2023, a feasibility study for the Shared Environmental Analytics Facility (SEAF) was conducted to address these challenges. SEAF aims to create a trusted environmental information supply chain by integrating research, government, and industry data to develop robust, repeatable assessment tools for regulators, industry, and communities. The initiative focuses on overcoming barriers to data discovery, accessibility, and reliability, supporting the development of environmental policies and assessments. Piloted in Western Australia, SEAF is funded by the state government, industry partners like BHP and Rio Tinto, and national research infrastructure co-investment. The project's two regional spokes, in Cockburn Sound's Westport development and the Pilbara region, are designed to unlock value from shared data and analytics. SEAF's methodology will fast-track the creation of trusted environmental data supply chains in new priority regions, further enhancing environmental management capabilities across Australia.

Oral Presentations

Recommendations for a National Framework for Advanced Health Analytics Infrastructure

Professor Nicola Armstrong, Curtin University

As part of ARDC's People Research Data Commons, ARDC and ADSN collaborated to evaluate the current state of digital health infrastructure, focusing on computing resources, data and analytics methods, data accessibility, as well as the socio-technical elements. The primary goal was to identify the most critical national research infrastructure (NRI) components needed to support advanced health data analytics and develop an Advanced Analytics Infrastructure Implementation Plan Specification.

A mixed-methods approach was used. A questionnaire was distributed to academics, government officials, and industry representatives to gather comprehensive insights. Additionally, focus groups facilitated in-depth discussions, and an environmental review provided contextual understanding.

The combined findings were synthesised into a series of recommendations to address identified gaps and needs in Australia's advanced health analytics infrastructure. The recommendations are grouped into four main areas: a) Enhance Computational Resources and Data Environments; b) Standardise Data Governance and Curation; c) Promote Collaborative and Ethical Research Initiatives; and d) Support Workforce Development and Practical Implementation.

Our methodological framework and findings underscore the importance of strategic NRI to support advanced health data analytics, ensuring that the ARDC's initiatives align with the needs and priorities of the research community.

Exploring Data and Model Embeddings with EmbedLens: A Plug-and-Play Visualisation Tool

Mr Milan Marocchi, Curtin University

We introduce EmbedLens, an open-source tool for the visualisation of high-dimensional data and model embeddings. EmbedLens supports several dimensionality reduction techniques-such as PaCMAP, UMAP, t-SNE, and PCA-and enables real-time interaction with datasets containing hundreds of thousands of points at 120 frames per second on consumer-grade hardware.

The tool facilitates detailed exploration of embeddings at both data and model levels, offering features such as point-wise investigation and filtering based on user-supplied criteria, including classification performance and label comparison. These capabilities allow users to investigate relationships between data points, explore class separation, and identify potential sources of misclassification, as well as assess the effects of data augmentation on model performance.

EmbedLens is designed as a plug-and-play platform, allowing developers to easily apply it to new datasets and models with minimal setup. The tool has been evaluated in collaboration with Ticking Heart Pty Ltd. for abnormal heart sound detection and explored in preliminary work on speech disorder classification at Telethon Speech and Hearing, where it has shown potential to reveal how noise and data transformations influence model outcomes.

With its Python/FastAPI backend and ReactJS frontend, EmbedLens integrates seamlessly into existing machine learning workflows, offering a robust and accessible solution for both researchers and industry professionals.

Automating Research Evaluation: Development and deployment of an ERA-Like Reporting Impact Evaluation System

Dr Kathryn Napier, Curtin Institute for Data Science, Curtin University

Excellence in Research for Australia (ERA) was a periodic assessment of the research activity of 42 Australian higher education providers (HEPs) across 236 ANZSRC fields of research (FoR). Performance was assessed by comparing research outputs to local and world benchmarks.

ERA reports were usually released every 3-5 years and employed citation-focused methodology in the analysis of research output data, self-reported by the participating HEPs. This was an extremely resource-intensive exercise, prompting its discontinuation in 2022. There has been a long-held interest in automating parts of this process.

The Curtin Open Knowledge Initiative aggregates bibliometric and bibliographic data from open sources such as Crossref, Unpaywall and OpenAlex using automated workflows and cloud-computing. The resultant database contains metadata for over 140 million research publications. We developed an ERA-like on-demand reporting system, the Research Impact Evaluation System (RIES), to demonstrate how open datasets may be used to run an automated analysis.

RIES allows for the comparison of a set of research outputs against a set of calculated benchmarks for fields of research and education, following ERA methodology and automatically apportioning research outputs to FoR and FoE.

RIES is currently being used by an Australian HEP to support their TEQSA accreditation, and has the potential to model and test approaches proposed for future national research assessments.

Bayesian Hierarchical Modeling and Multiclass Priority Queueing for Predicting Emergency Department Wait Times in Australia

Mr Michael Nefiodovas, University of Western Australia

Prolonged and uncertain wait times in Australian emergency departments (EDs) elevate patient stress. We developed a data-driven solution comprising: (1) a patient-facing interface providing real-time wait time estimates; (2) a predictive model using Bayesian hierarchical time-series analysis to forecast patient arrivals across the five Australasian Triage Scale categories; and (3) a dashboard for healthcare providers to optimise resource allocation. Arrival rates are modeled by decomposing daily and yearly periodicities and trend components, with priors informed by historical data. Our model is implemented in Stan which allows for convergence diagnostics and uncertainty quantification. These rates are input into a multiclass priority queue model, treating the ED as a single queue with priority-based patient sorting and service rates dependent on triage levels. Queueing theory yields analytical expressions for expected wait times, propagating uncertainty from the Bayesian model. Validation against synthetic datasets from Perth hospitals assesses performance without breaching patient confidentiality. We propose an integration with existing hospital information systems using HL7 FHIR patient record standards. To mitigate potential anxiety from displayed wait times, we employ empathetic interface design and gather patient feedback. Collaboration with Perth ED physicians and health department officials helped our approach to ensure practical applicability. Our system accurately predicts ED wait times as measured by backtested calibration, aiming to reduce patient stress and enhance operational efficiency. Future work includes pilot implementations in Perth hospitals and trials to evaluate impacts on patient stress and resource optimisation quantitatively.

My Smart Shelf: Cisco MasterTech Hackathon

Mr Alin Ibrahim, Innovation Central Perth

During the National Industry Innovation Network (NIIN) exhibition at the World of Solutions, Cisco Live Melbourne, Innovation Centrals (ICs) from various regions of Australia participated in MasterTech. MasterTech was a five-day hackathon event that brought together outstanding university students from each IC. These students worked collaboratively to develop innovative solutions for Coles supermarket, the sponsor of the hackathon. Innovation Central Perth's proposed solution was My Smart Shelf, a system designed to enhance the customer shopping experience and promote sustainability. My Smart Shelf uses intelligent sensors to guide customers to specific products based on their preferences, including product type, price, carbon footprint, and historical prices. Additionally, the system helps Coles minimize product spoilage and optimize restocking through the use of intelligent algorithms. Furthermore, My Smart Shelf promotes sustainability by offering automatic discounts on products nearing their expiration date and highlighting items with reduced plastic consumption and carbon footprint. We invite individuals to attend the conclusion of this session to observe and test a functional prototype of My Smart Shelf.

Association between team playing styles and individual player performance

Mr Samuel Moffatt, Curtin University

Within team sports, teams implement overarching strategies designed as a team playing style. Implementing these team playing styles depends on the performance of individual players within a match. Identifying the relationship between individual player performance and a team's playing style can lead to better-informed decisions regarding how player performance affects the implementation of team playing styles. In this paper, to model the relationship between individual players and team playing styles in the context of Australian football, Classification Based on Associations (CBA) algorithms are used. Focusing on chains of possession within matches, extracted from transactional match data of 414 matches played during the 2021 and 2022 seasons, 16 team playing styles and 16 player roles are identified. Association rules using the five CBA models quantify the association between the player performance categories and team playing styles. The resulting association rules inform decisions regarding the selection of players required to perform a specific team playing style. The association rules can also identify the void left by an injured player and its impact on the team's ability to perform specific playing styles.

Optimising Machine Failure Detection Through Data Quality Control

Dr Shuixiu Lu, ARC Training Centre for Transforming Maintenance through Data Science at UWA

Data holds significant potential to provide valuable insights for addressing engineering challenges, for example, machine failure detections. However, it is difficult to make something of nothing when it comes to data-driven methods for automatic detections. This study discusses challenges in data-driven machine failure detections and illustrates how effective data labelling can help overcome these obstacles. We focus on time series analysis for rolling element bearing failure detection. Our analysis aims to emphasise the importance of data labelling and quality in data-driven detection while bridging the gap between industry expectations and research objectives.

Multi-class weed classification with Deep learning methods for smart agriculture

Dr Tej Bahadur Shahi, Queensland University of Technology

The world population is estimated to reach 9.7 billion by 2050 and Global food demand is expected to increase significantly from 35% to 56% from 2010 to 2050. However, the expansion of industrialization, desertification and urbanization have reduced the crop production area, impacting productivity. Weeds are among the most undesirable plants that impede crop growth and should be identified early to mitigate their spread effectively (Ahmad et al., 2021). The swiftly evolving deep learning (DL) methods can be utilized to identify the Weeds using the images taken with a standard RGB camera. In this work, we perform the comparative study of various DL methods such as Reset, VGG16, DenseNet21, EfficientNet, and MobileNet and evaluate their performance on two publicly available but challenging weed datasets- DeepWeed (Olsen et al., 3019) and CottonWeed15 (Saini et al. 2024). Our simulation results demonstrate that lightweight models such as MobileNet, achieve comparable accuracy (~92.27%) to that of heavy-weight models such as VGG (~89.05%) and DenseNet (~94.82%). From this study, we propose that lightweight models such as MobileNet and EfficientNet can be deployed in edge devices like smartphones to reasonably identify the weed species, ultimately benefiting the farmers.

International Data Week 2025, Brisbane Queensland

Mr Keith Russell, Australian Research Data Alliance

International Data Week (IDW2025) is being held in Australia for the very first time ever! The conference, to be held in the vibrant city of Brisbane, Queensland, Australia, will be on 13 to 16 October, 2025.

IDW takes place every two years and is the leading international conference on research data. It is jointly organised by the International Science Council's Committee on Data (CODATA) and World Data System (WDS), and the Research Data Alliance (RDA). It combines the RDA Plenary Meeting, the biannual meeting of this international member organisation working to develop and support global infrastructure facilitating data sharing and reuse, with SciDataCon, the scientific conference addressing the frontiers of data in research organised by CODATA and WDS.

This year's theme of 'Data for Positive Change' highlights our commitment to the role of data in instituting change by empowering communities and advancing research. The program features an array of keynote talks, workshops, and interactive sessions led by eminent, as well as emerging experts, in the field. Attendees will have the opportunity to engage in thought-provoking discussions, gain insights from cutting-edge research, and share their own experiences and innovations in using and sharing research data.

We look forward to welcoming you to Brisbane for an unforgettable International Data Week!

Evaluation of multi-dimensional activity label embeddings for visualization and clustering of complex processes traces

Dr Andrzej Janusz, QUT

Process mining is a collection of techniques used to analyse and model processes by extracting knowledge from event logs readily available in today's information systems. It helps organizations understand how their processes are actually performed, identify bottlenecks, and optimize workflows for better efficiency. It can also be used as a tool for analysing logs generated by complex systems and detecting anomalous behaviours.

In my talk, I will discuss the problem of constructing embeddings of activity labels and process traces for visualization and clustering of process trace variants. My main case study will involve data from the cybersecurity domain where on the one hand, machine-generated event logs are abundant, and on the other hand, the corresponding processes are heterogeneous and extremely difficult to analyse using standard process mining techniques. I will explain the design of a benchmark for the quality of activity label embeddings, which is inspired by the compositionality of word representations. Additionally, I will present the results of experiments aiming to evaluate the usefulness of various approaches to constructing process trace embeddings, including algorithms adopted from the NLP domain. These results demonstrate how process mining can augment the extraction of useful knowledge for improving cybersecurity breach detection in practical application scenarios.

A comparative study of chunking strategies for enhanced Retrieval-Augmented Generation in Large Language Model-based system

Ms Dan Dai, Queensland University of Technology / Centre for Data Science

Large Language Models (LLMs) have shown immense potential in natural language generation tasks, but they often struggle in highly specialised domains due to lack of sufficient domain-specific knowledge in their training datasets, leading to issues like hallucination, where models generate content that is nonsensical or unfaithful to the provided source content. Retrieval-Augmented Generation (RAG) addresses this by enhancing LLM performance through external data retrieval. While many frameworks and tools have been developed, one key challenge remains: the choice of an effective chunking strategy.

This study adopts a structured, experimental approach to implement and evaluate various chunking strategies within a Query-Answering RAG system. By implementing the experiment in the context of a specific architecture, engineering and construction (AEC) project, this research provides insights into optimising chunking strategies for domain-specific RAG, and contributes to establish best practices for similar workflows. The findings can significantly inform RAG systems enhancement, particularly in fields similar to AEC, where precise, contextually relevant information is essential. Ultimately, this study can guide better decision-making in designing and implementing strategies to improve LLM performance in handling large, structured documents, while expanding the use of RAG in specialised domains.