THE UNIVERSITY OF SYDNEY

FACULTY OF ENGINEERING

Engineering Vacation Research Internship Program

COMPUTER SCIENCE RESEARCH PROJECTS FOR SUMMER 2021-22

CS2021-22/1 Evaluating the Performance-Durability Tradeoff of MongoDB

Supervisors: Prof Alan Fekete and A/Prof Uwe Roehm

Eligibility: Strong academic background in Computer Science. Particularly useful would be good knowledge of Database and/or Operating Systems. Experience in benchmarking or tuning systems would be desirable.

Project Description:
MongoDB is a popular NoSQL database that offers database like functionality to store and access collections of JSON data. The data in MongoDB is usually replicated across several nodes, and for performance reasons, the system provides alternative ways to update the data; these choices differ in how thoroughly the change is protected against loss from system crashes.

The goal of this project is to develop some guidance for programmers to help them decide which MongoDB configuration is best suited for their applications. This project will extend previous work (from an Honours thesis) to measure the properties of these choices; the focus will be on identifying conditions which make the system more vulnerable to data loss, and to evaluate the trade-of between update scalability and write durability.

This summer internship is offered in collaboration with MongoDB and will allow a close insight into the performance tradeoffs with modern distributed NoSQL databases.

Requirement to be on campus: No

CS2021-22/2 Geometric Networks in the Presence of Obstacles

Supervisor: Dr André van Renssen

Eligibility: A strong background in discrete math and/or algorithms is strongly recommended for this project.

Project Description:
Geometric networks are networks where every node has a location associated with it. When two nodes are connected, they know each other's location. These networks are quite well understood, and most networks allow for modifications in order to enhance it, such as guaranteeing a certain maximum degree for each node, bounding the number of hops needed to reach any other node, or having efficient routing algorithms. However, when there are obstacles that block communication between nodes on opposite sides, far less is known about these networks. In this project, we aim to shed some light on these networks in the presence of obstacles in order to design networks that have additional good properties and/or allow for efficient routing algorithms.

Requirement to be on campus: No

CS2021-22/3 Routing in Geometric Networks

Supervisor: Dr André van Renssen

Eligibility Criteria: A strong background in discrete math and/or algorithms is strongly recommended for this project.

Project Description:
Geometric networks are networks where every node has a location associated with it. When two nodes are connected, they know each other's location. These networks are quite well understood, and a number of efficient routing algorithms have been developed for them. Some of these routing algorithms come with theoretical worst-case guarantees on the length of the routing path, but not all do. Hence, we aim to perform a comprehensive study comparing various routing algorithms in order to discover which performs best in practice and which types of geometric graphs the known routing algorithms have trouble with. The latter will also provide a starting point for the development of new routing algorithms specifically designed to handle these problematic situations.

Requirement to be on campus: No

CS2021-22/4 Predictive Gesture Classification in Virtual Reality (VR)

Supervisor: Dr Anusha Withana

Eligibility: You will closely work with the supervisor, and we expect you are a fast learner. Excellent skills in programming, skills in machine learning, knowledge in VR, and human computer interaction are added benefits.

Project Description:
Can we predict a players next move in a VR game? Gestural input, for example using your hand movements as input, has become one of the most popular input technologies for rapidly developing virtual reality (VR) and augmented reality (AR) applications (eg. GearVR, HTC Vive, etc.). Gestures, such as hand movements and poses in space, are an essential part of our daily communication (ie. body language) and thus create an intuitive modality for interacting with these immersive new computer applications. In this project we focus on predicting hand movements for the purpose of pre-recognising user activities in VR. The project will build our existing work on continuous hand movement recognition (See video in the link) and the data we have collected.
https://www.dropbox.com/s/jkgik9mj4bd5e83/uist21a-sub1729-cam-i27.mp4?dl=0

Requirement to be on campus: No

CS2021-22/5 Constraint-aware Trajectory Dataset for Drone Delivery Services

Supervisor: Prof Athman Bouguettaya

Eligibility: Having Python programming and preferably machine learning knowledge

Project Description:
Drones offer cost-effective and fast delivery services. The potential utilization of drones is limited by payload capacity and battery consumption constraints. Drones may need multiple times of recharge for persistent delivery operation. The drone delivery environment is highly constrained because the availability of recharging stations is not guaranteed. We leverage the service paradigm to address the key challenges in delivery by drones. The functional and non-functional properties of drones are abstracted as Drone Services. The drone services operate in a skyway network which is constructed by linking skyway segments. Each node in the network is a recharging station or a delivery target. Given a source and a destination, the objective is to collect a trajectory dataset considering battery limitations and availability of pads at stations. A machine learning based model will be developed to predict the drone's arrival at certain stations that will help in computing best skyway segments leading to the destination.

Requirement to be on campus: Yes

CS2021-22/6 Smart App for Covid19 Contact Tracing

Supervisor: Prof Athman Bouguettaya

Eligibility:

- Experience in programming Android apps (Java or Kotlin).
- Experience with backend developments (preferably NodeJS).
- Working knowledge of database technologies (e.g., MySQL).

Project Description:

Contact tracing refers to the process of notifying people if someone, who they have been in contact with, got infected. The recently developed contact tracing apps rely on using Bluetooth to identify users in contact. However, the usage of Bluetooth has several issues including recording contacts that never happened and miss others that did. Other developed apps, such as the service NSW app, use a check-in/check-out feature. This solution offers a more accurate tracking of users. However, it is likely that fewer people will use it – due to what is known as the "intention-behavior gap". Therefore, there is a need to use incentives to encourage people to use the app. Incentives such as gamification have been used to increase user engagement. This project aims to develop an easy-to-use and yet entertaining game-based contact tracing app for people to check-in/check-out when they enter a confined area such as a classroom, café, etc.

Requirement to be on campus: Yes

CS2021-22/7 Machine Learning-based Approach to Recognize Indoor Activities from Thermal Image Data

Supervisor: Prof Athman Bouguettaya

Eligibility: Required Skills:

- Experience in programming in Python or Matlab.
- Experience in image processing libraries such as OpenCV or PyTorch.
- Able to implement image processing algorithms to separate background infra-red radiation from human and object.
- Able to implement machine learning classification algorithms, including K-Nearest Neighbors (KNN), Multilayer Perceptron Artificial Neural Network (MLP), and Linear Regression.

Good to Have (But Not Essential):

- Knowledge of Github

Project Description:

Activity recognition in the indoor environment such as home, office aims to detect the movement of a human in a sequence of images or video. Activity recognition can be very beneficial to several applications such as healthcare, surveillance, and home automation. The main data source for the task of activity recognition is sensors. Existing approaches either use intrusive sensors that do not ensure people's privacy when obtaining high accuracy estimations, such as RGB cameras or use non-intrusive sensors with lower accuracy, such as motion sensors. In this regard, we use thermal cameras to preserve people's privacy. In addition, thermal cameras have some advantages over the RGB cameras such as they are not light-dependent and can work in dark environments.

Students will have the opportunity to implement image processing algorithms to extract features from the thermal image data and implement machine learning algorithms to build the indoor activity recognition model.

Requirement to be on campus: Yes

CS2021-22/8 Swarm-based Drone Delivery Dataset

Supervisor: Prof Athman Bouguettaya

Eligibility: Experience in python programming and preferably machine learning knowledge.

Project Description:
The wide availability of drones opens opportunities for a wide number of applications including package delivery. Drones for delivery present some unique challenges to fully deliver on their potential. In particular, drones have limited payload and battery capacity. There are instances where there is a need to deliver goods by a deadline and which weigh more than the maximum of a single drone's payload. In this case, the use of drone swarms is an effective alternative to address the constraints for the timely delivery of heavier and/or multiple packages which go beyond the capability of one single drone. Drone swarms are teams of autonomous unmanned aerial vehicles that act as a collective entity. The goal of the project is to collect a dataset using different formations of swarms and varying wind speeds and directions. Machine learning models will be trained on the dataset to learn the best swarm flying conditions.

Requirement to be on campus: Yes

CS2021-22/9 Wi-Fi-based Indoor Positioning

Supervisor: Prof Athman Bouguettaya

Eligibility:
- Experience in programming Android apps (Java or Kotlin)
- Experience with backend developments (preferably NodeJS)
- Working knowledge of database technologies (e.g., mysql)
- Basic knowledge of WiFi access points

Project Description:
Indoor positioning refers to the process of finding out the location of people/objects indoors, e.g., inside malls or offices. One way to determine the position of an object is to leverage WiFi access points. Specifically, the strength of the received signal can reveal information about how far the access point is. This project aims to use the signals from multiple access points to determine the position of objects. The project involves developing a smartphone app that monitors WiFi signals in the background. The app would use the received signal to compute the position of the device. The app then would push the computed position to a server for storage.

Requirement to be on campus: Yes

CS2021-22/10 Anomaly and Pattern Detection using Advanced Machine/Deep Learning

Supervisor: Dr. Basem Suleiman

Eligibility: Required Skills and knowledge:
- Adequate knowledge and skills in Machine learning/deep learning, data analytics
- Excellent programming skills in Python.
- Majoring in computer science and computational Data science
- Very good critical thinking problem solving skills
- Self-motivated and take initiatives

Project Description:
Data can provide invaluable insights for better solutions to challenges we face in various domains and even at the personal level.

The goal of this project is design and implement machine/deep learning model(s) to detect patterns and identify abnormal events, behaviour (anomalous data) in large datasets. The

predictive models will be trained using multiple datasets of interest and tested to evaluate its accuracy in detecting pattens and identifying anomalies. The application areas include cyber-security/security and privacy in smart home/cities. The datasets will be provided in this project.

Interns will get hands on experience including:

1. Analysis of large IoT dataset to discover interesting patterns and insights
2. Applying machine/deep learning algorithms to model and predict patterns, and anomalous data
3. Experimenting with the developed models with different settings

Requirement to be on campus: No

CS2021-22/11 Diffusion of False News in Social Media Networks

Supervisor: Dr Basem Suleiman

Eligibility: Required Skills and knowledge:

- Adequate knowledge and skills in Machine learning/deep learning, data analytics
- Excellent programming skills in Python.
- Majoring in computer science and computational Data science (should be strong in Machine/Deep Learning)
- Very good critical thinking problem solving skills
- Self-motivated and take initiatives

Project Description:
Social media platforms such as Facebook and Twitter have become the predominant medium for information sharing in various fields including marketing, politics, and education. Individuals, governments, and organizations heavily rely on gaining information and news as well as sharing with their interested parties. The popularity of such platforms has been accompanied with new challenges including the spread of false news and misinformation. Such challenges have enforced major social media platforms to prioritize various techniques to reduce the spread of false news and misinformation.

The goal of this project is to investigate and collect the differential diffusion of all of the verified true and false news stories distributed on Social media platforms from before and after the start of pandemic pre 2019 and post 2019. Falsehood and true diffusion analysis should cover different topics such as natural disasters, science, urban legends, or financial information. A key part of this project is to work on developing machine/deep learning that can help to learn and detect the spread of false news in social networks and apply techniques to stop such spread.

Requirement to be on campus: No

CS2021-22/12 Predictive Elasticity for the Cloud

Supervisor: Dr Basem Suleiman

Eligibility: Required Skills and knowledge:
cloud/distributed computing, machine learning; data science/analytics; python programming, deep learning (preferred); AWS/Google cloud (preferred).

Project Description:
Cloud elasticity (or auto-scaling) is one of the key characteristics which attracts cloud consumers due to the provisioning of computing resources “on-demand” and based on “pay-per-use” model. Although cloud providers such as Amazon, Google and Microsoft provide some techniques for enabling cloud elasticity, those techniques are not effective and efficient when one considers different factors such as the type of application and its workload.

The goal of this project is to develop an intelligent elasticity engine to scale cloud resources efficiently and effectively. This engine considers various crucial aspects to make scaling decisions such as application's workload characteristics and cloud resources performance and costs.

Achieving this would require using machine/deep learning algorithms. This project will be an opportunity for students to gain very good knowledge and skills of major cloud computing services (such as AWS/GCP) and how it works. Also, students will work on applying machine/deep learning algorithms to tackle cloud auto-scaling challenges.

Requirement to be on campus: No

CS2021-22/13 Predictive Models in Smart Homes/City Environments

Supervisor: Dr Basem Suleiman

Eligibility: Required Skills and knowledge:

- Adequate knowledge and skills in Machine learning/deep learning, data analytics
- Excellent programming skills in Python.
- Majoring in computer science and computational Data science
- Very good critical thinking problem solving skills
- Self-motivated and take initiatives

Project Description:
The proliferation of IoT devices results in an enormous amount of data being collected every second. However, this data would not be useful without exploration. Data can provide invaluable insights for better solutions to challenges we face in various domains and even at the personal level.

This project aims to leverage IoT data collected from IoT devices (e.g., smartphone/watches, wearables, sensors) to address key challenges in smart home/city environments. Interns will design and implement machine/deep learning models to intelligently predicts patterns and insights from large IoT datasets. The developed models can be applied into personalized services, cyber-security, recommender systems, smart homes, smart cities, using real IoT datasets.

Interns will get hands on experience including:

1. Analysis of large IoT dataset to discover interesting patterns and insights
2. Applying machine/deep learning algorithms to model and predict interesting events, activities, and patterns
3. Experimenting with the developed models with different settings

Requirement to be on campus: No

CS2021-22/14 Touchless Hand Gesture-based Dialog system using Deep Learning Techniques

Supervisors: Dr Caren Han and Dr Josiah Poon

Eligibility: Student has completed any of the following courses:

- COMP5046 (Natural Language Processing)
- COMP5329 - Deep learning

Project Description:
Video-based hand gesture detection and tracking solutions are considered as vital components in various technological domains. The hand tracking and gesture recognition algorithms detect locations of the hand in real-time allowing you to implement touchless dialog interfaces that can be controlled contact-free using gestures. Right from accessing touchless interfaces at fast food centers to understand sign language, these SOFT models help us overlay the digital

content and information on top of the physical world in AR. The primary goal of this project is to leverage deep learning-based object detection method (Mediapipe “Plam/Hand detection”) and bring forth creative dialog use cases.

Readings and resources:
Demo: https://www.youtube.com/watch?v=2AaqUD-Ctko

Resources: Mediapipe "Palm/Hand detection"
https://google.github.io/mediapipe/solutions/hands

Requirement to be on campus: No

CS2021-22/15 Visual Question Answering

Supervisors: Dr Caren Han and Dr Josiah Poon

Eligibility: Student has completed the following the course:
- COMP5046 (Natural Language Processing)

Project Description:
Visual question answering (VQA) is a multimodal task that processes both image and text features. Given an image and a question, a VQA model is expected to predict the correct answer to the question based on the image contents. This requires a model's understanding in both images and questions. Current VQA models use different techniques to integrate the image features and question features together for an answer representation, which is then decoded for the answer prediction. In this project, we aim to work on the state-of-the-art VQA models, making modifications to achieve an improved performance with higher accuracy.

Readings and resources:
https://www.youtube.com/watch?v=6UmKHiGOyVI
https://www.youtube.com/watch?v=XYeo_bOXnc4
(research paper, video, and code)
https://usydnlp.info/ (usydnlp website)

Requirement to be on campus: No

CS2021-22/16 Toxic Language Detection in multiplayer online game

Supervisor: Dr Caren Han and Dr Josiah Poon

Eligibility: Student has completed the following courses:
- COMP5046 (Natural Language Processing)

Project Description:
Abusive Language on the Internet, sometimes referred to as the offensive language, cyberbullying and hate speech has become a severe social problem.

Social interactions in multiplayer online games are an essential feature for a growing number of players world-wide. However, this interaction between the players might lead to the emergence of undesired and unintended behaviour, particularly if the game is designed to be highly competitive. Communication channels might be abused to harass and verbally assault other players, which negates the very purpose of entertainment games by creating a toxic player-community. By using a novel natural language processing framework, we detect profanity in chat-logs of a popular Multiplayer Online Battle Arena (MOBA) game and develop a method to classify toxic remarks.

Readings and resources:
https://github.com/usydnlp/CONDA (research paper and code)

https://usydnlp.info/ (usydnlp website)

Requirement to be on campus: No

CS2021-22/17 Randomness in Learning and Testing Algorithms

Supervisor: Dr. Clément Canonne

Eligibility: Background in discrete mathematics and/or basics of probability and statistics. Programming experience not necessary, but strongly encouraged for aim (2).

Project Description:
Fast and theoretically accurate algorithms for testing or learning properties of data are known, under various types of "constraints" (such as communication constraints, or privacy requirements).

Many of those algorithms, however, are randomised, which can be an issue in settings where generating random seeds is expensive, impossible, or could lead to vulnerabilities.

The aim of this project is to (1) investigate whether one can optimise the amount of randomness required by these algorithms; (2) implement and evaluate in practice their performance (time- and accuracy-wise); and (3) assess whether the type of randomness introduced can lead to security issues or attacks.

[The applicant can choose to focus on a subset of the above 3 goals, depending on their background and preferences.]

Requirement to be on campus: No

CS2021-22/18 Road safety, infrastructure, and design: what are the attributes of unsafe intersections?

Supervisors: Dr. Emily Moylan, Prof. David Levinson, Dr. Mohsen Ramezani, Prof Judy Kay

Eligibility: Experience with or willingness to learn Python or similar language. Experience with or willingness to learn GIS.

Project Description:
The design of transport infrastructure can incentivise positive or negative behaviours in users. Some behaviours are associated with crashes and near misses. When designing road infrastructure, engineers are tasked with numerous decisions including widths, alignments, signalling, signage and paint, speed limits and location of transit stops. This project uses the attributes of road intersections across New South Wales from diverse publicly available data sources to explain safety outcomes focusing on pedestrian safety and the elements of intersection design that are most likely to influence behaviours associated with pedestrian crashes. Due to the scope of the project from civil infrastructure to human-technology interaction, the student will be working with a team of researchers across Civil Engineering and Computer Science. This project would be suitable for continuation as an honours topic.

Requirement to be on campus: No

CS2021-22/19 Video processing techniques for observing pedestrian interactions with intersections

Supervisors: Dr. Emily Moylan, Prof. Judy Kay, Prof. David Levinson, Dr. Mohsen Ramezani

Eligibility: Familiarity with human-technology interaction. Familiarity with common video processing techniques.

Project Description:
The design of transport infrastructure can incentivise positive or negative behaviours in users. Some behaviours are associated with crashes and near misses, and we should choose designs that minimise these. This project focuses on extracting information about pedestrian and driver behaviour from video footage. Existing off-the-shelf products provide trajectory and delay metrics but no insights into hesitations, trajectory adjustment, near-misses or distraction. The student will review the approaches in the literature, assess the unmet needs for the intersection design application and develop customised algorithms to record the relevant behaviours. Due to the scope of the project from civil infrastructure to human-technology interaction, the student will be working with a team of researchers across Civil Engineering and Computer Science. This project would be suitable for continuation as an honours topic.

Requirement to be on campus: No

CS2021-22/20 Sleep disorder analysis using advanced machine learning techniques

Supervisors: A/Prof Irena Koprinska and Dr Bryn Jeffries

Eligibility:

- Machine learning skills, e.g. completed COMP3308/COMP3608 COMP5318 with D/HD
- Excellent programming skills

Project Description:
Insomnia and sleep apnea are common sleep disorders. Insomnia is defined by difficulties falling asleep and staying asleep. Sleep apnea is characterized with periods of reduced breathing or no breathing at all during sleep. Both disorders cause daytime sleepiness, fatigue, reduced energy and motivation, and may lead to depression, heart disease, diabetes and other adverse health effects. The goal of this project is to use machine learning techniques to analyse sleep disorders, e.g. to objectively detect insomnia and normal sleepers based on EEG data and predict sleep apnea events in advance based on respiratory data to allow for medical devices to intervene. The project will use large datasets containing data from multiple signals recorded overnight.

Requirement to be on campus: No

CS2021-22/21 Minimum Partition into Plane Subgraphs [CG Challenge]

Supervisors: Prof. Joachim Gudmundsson, Martin P. Seybold

Eligibility: Interest in algorithms and programming.

Project Description:
Computational Geometry Programming Challenge: Given a geometric graph G=(V,E), with vertices represented by points in the plane, and edges by straight-line connections between vertices.

The task is to partition E into as few subsets E_1,…, E_k as possible, such that each subgraph G_i=(V,E_i) is plane: In the given geometric representation, line segments representing edges my touch at end points if and only if the corresponding edges are incident in G_i; no edges my cross, i.e., share points that are not segment end points. See https://cgshop.ibr.cs.tu-bs.de/competition/cg-shop-2022/#problem-description
Competition Dates:
Sept. 19, 2021 - Jan. 19, 2022, 11:59 a.m. (AoE)

Requirement to be on campus: No

CS2021-22/22 Randomized Block Trees

Supervisors: Prof. Joachim Gudmundsson, Martin P. Seybold

Eligibility: Interest in algorithms, analysis, and programming.

Project Description:
Processing inputs in random order effectively avoids worst-case behaviour of simple, deterministic algorithms on many occasions – often providing optimal expected performance. Unbalanced Binary Search Trees for example attain remarkable properties such as insertions with expected O (1) write operations.

The well-known Treap maintains these properties and allows simple implementations with concurrent access and without storage overhead.

However, exploiting randomness for trees with block layout is still not well understood.
The known block layout for Treaps is very difficult to implement and even less is known for maintaining randomness in the balanced B-Trees.

The goal of this internship is to investigate simpler block layouts for Treaps and ways to maintain randomness in B-Trees.

Requirement to be on campus: No

CS202-22/23 Elastic Edge Computing Overlay with Neural Computing Sticks

Supervisor: Dr. Kanchana Thilakarathna

Eligibility: The student should have a fundamental knowledge of computer networks and applied machine learning skills.

Project Description:
Edge Computing (EC) pushes the frontiers of applications, data and services away from centralised model to closer to end users at the edge of the network. In parallel, there has been significant development in low-cost portable neural computing devices such as Neural Compute Sticks from Intel and Movidius. These devices can essentially turn any device into an edge computing resource. We need to fundamentally rethink how such devices can be effectively leveraged to scalable delivery of latency-sensitive services. For example, if there is a user who lives in a poorly connected area, can the user improve quality of experience in interactive video streaming by simply plugging a neural computing stick to the TV? We aim to answer this question with software defined networking that allows sharing of a physical network a more controlled fashion. The student will develop a proof-of-concept demonstrator integrating Movidius neural computing stick at the client side with a video streaming session.

Requirement to be on campus: No

CS2021-22/24 Unravelling the Nascent Privacy Risks of Augmented Reality

Supervisor: Dr. Kanchana Thilakarathna

Eligibility: Knowledge on applied machine learning.
Mobile programming (iOS or Android) experience will be an added advantage.

Project Description:
Augmented, virtual, and/or mixed reality technology (AR/VR/MR) is increasingly becoming popular. From face filters to virtual pets or monsters that seemingly inhabit the physical-world, various MR applications are now widely accessible to most users.
MR platforms require spatial understanding of objects or surfaces, including their structural and photo-metric (e.g. colour and texture) attributes. Aside from objects being detected, spatial

information also reveals the location of the user with high specificity, e.g. in which part of the house the user is, or even detect user poses, movement, or changes in their environment which poses additional and, potentially, latent risks to user privacy. This project focuses on experimental validation of the existence of privacy risks associated with MR devices, e.g. Oculus, and measures to quantify and detect the extent of the threats. This is a collaborative project with Facebook Reality Labs.

Requirement to be on campus: No

CS2021-22/25 Generative Adversarial Framework for Synthetic Time Series Data Generation

Supervisor: Dr. Kanchana Thilakarathna

Eligibility: Basic knowledge of statistics and thorough knowledge on applied machine learning.

Project Description:
Gathering data for training machine learning model is a challenge in many domains due to reasons such as user privacy, ethics, time, etc. In addition to the probabilistic approaches such as Copula, recent advancements in Generative Adversarial Network (GAN) based approaches have shown outstanding performance due to their capability of generating high-fidelity datasets in many contexts such as images. However, ordinary GAN architectures are only capable of capturing the distribution of continuous and complete data but cannot be used for learning the distribution of discrete variables in time series data. We proposed a data generation framework for encrypted video traffic data realising Wasserstein GANs in the past. In this summer project, we aim to extend this data generation framework to control the reidentification risk of raw attributes, i.e. private information leakage, from the synthesized data.

Requirement to be on campus: No

CS2021-22/26 Supporting Consumer Decision Using Online Reviews

Supervisor: Dr Kevin Kuan

Eligibility: Basic proficiency in Python for data mining.

Project Description:
Consumers are increasingly relying on online reviews in their everyday lives, including shopping, dining, traveling etc. This project aims to better support consumer decisions using online reviews by understanding how different numeric and text features of online reviews affect consumers in their decision-making and behaviour. In the project, students will have the opportunity to review literature in disciplines such as business, psychology, computer science, etc., and to analyse a large data set of online reviews containing both number and text.

Requirement to be on campus: No

CS2021-22/27 Understanding the Assimilation of Electronic Medication Management System in an Australian Hospital

Supervisor: Dr Kevin Kuan

Eligibility: Basic proficiency in Python for statistical modelling.

Project Description:
Electronic Medication Management (eMeds) systems support the improved quality, safety and effectiveness of medication management with NSW hospitals. This includes providing support for doctors, nurses and pharmacists to prescribe, order, check, reconcile, dispense and record the administration of medicines. In the project, students will have the opportunity to review

literature in health IT management, and to perform statistical analyses on a survey data using techniques such as Partial Least Squares Structural Equation Modelling (PLS-SEM) and fuzzy set Qualitative Comparative Analysis (fsQCA).
Requirement to be on campus: No

CS2021-22/28 Implementation and performance evaluation of iterative multi-join algorithm in C/C++ across a large-scale distributed setting made of Raspberry Pico devices

Supervisors: Prof Albert Zomaya and Dr M. Reza Hoseiny

Eligibility:

- Good knowledge of C/C++
- Experience working with Git, Linux, SSH
- Understanding of the multi-threading programming paradigm, low-latency and performant applications
- Work closely with the rest of team to solve problems, and transfer knowledge
- Writing reusable, testable, and efficient code
- Good knowledge of Design patterns, fundamental data structures and algorithms
- Good knowledge of Performance evaluation metrics for computer systems

Project Description:
Raspberry Pi Pico, a recent microcontroller board equipped with RP2040 microcontroller chip, that offers a lot of interesting features. It comes with a special C/C++ SDK for developing extra features. In this project, we try to investigate its capabilities for processing the real-time streaming data and performing Incremental and Iterative computation and delta computation commonly to be running across a large-scale deployment of several Pico microcontrollers when performing distributed iterative multi-join algorithm. The setting here is that there are multiple Pico microcontrollers, each of whom move data through a common dataflow graph in the target application. The data may move between Pico microcontrollers, and as the data are processed by the functions in each Pico microcontroller, we have limited guarantees about the consequence of such data movement. The proposed algorithm will attach a logical timestamp (e.g. sequence numbers) to every data before movement among microcontrollers that provides each microcontroller with an understanding of progress in the computation in the entire distributed setting.

Evaluating the relational join is perhaps the most centric algorithm in the database systems. While there are many ways to perform join over multiple relationships (such as hash-join), Ngo, Re, and Rudra [1] proposed a worst-case optimal join algorithm based on iterative stateful approach for computing relational joins over multiple relationships with common attributes (aka Generic-join). The approach adds in common attributes one at a time to the result of the join from the previous round. In each round of computation, each relation needs to propose some extensions, and then ask the other relations to validate them. In this project, our aim is to provide a distributed implementation of Generic-join algorithm over multiple relationships when the data enters into the system in a continuous streaming manner. The aim is to improve the run-time performance by exploiting the parallel computation, pipelining the functions, minimizing the loss in bandwidth, and reusing the optimal local on-chip memory. We can improve the performance of computations by extending a strategy to pipeline records through relations of a computing session in a set of never-ending data streams of Pico devices.

Scope:
Student(s) may start by understanding the detail of Generic-Join mechanism over multiple streaming data. The student/group can continue by implementing the algorithm using C or C++ standard libraries. The assumption here is that the data is distributed over multiple Raspberry Pi Pico and can have arbitrary data types (hence, the C++ implementation needs to use templates). Next step is to evaluate the run-time performance of Generic-Join approach over other implementations of multi-join such as hash-based join used in conventional database systems (e.g. MySQL), and streaming engine such as Apache Spark/Storm/Flink. Finally, possible performance improvement of implementations (such as parallel computation,

pipelining the kernel functions, and local on-chip memory) can be further explored if time allows.

Expected outcomes/deliverables:
A set of C/C++ libraries/ scripts for implementation and testing of Generic-Join approach over distributed Raspberry Pi Pico. The implementation is expected to use templates as the data types of incoming data may be arbitrary in the run-time) and corresponding make file to store large data input. A detailed report that explains the internal mechanism of the developed system and interactions among different components is expected.

Reading material, Specific required knowledge, skills, and/or technology:
- Raspberry Pi Pico C/C++ SDK documentation
- Timely Dataflow architecture and its Progress Tracking documentation
- Templates and standard libraries in C/C++

References:
[1] Hung Q. Ngo and Christopher Re and Atri Rudra, ``Skew Strikes Back: New Developments in the Theory of Join Algorithms``, https://arxiv.org/abs/1310.3314, 2013
[2] Frank McSherry, Worst-case optimal joins, in dataflow, http://www.frankmcsherry.org/dataflow/relational/join/2015/04/11/genericjoin.html

Requirement to be on campus: No

CS2021-22/29 Secure AMMs in Decentralized Finance (DeFi)

Supervisor: Dr. Qiang Tang

Eligibility: CS, math, and physics majors are welcome. Background of cryptography will be a big plus.

Project Description:
We will study how to rigorously defend new attacks on DeFi protocols, particularly secure automatic market-maker (AMM) via smart contracts.

Example Reference:
https://arxiv.org/pdf/2106.07371.pdf
https://eprint.iacr.org/2021/609.pdf

Requirement to be on campus: No

CS2021-22/30 Attacks in Decentralized Finance (DeFi)

Supervisor: Dr. Qiang Tang

Eligibility: CS, math, and physics majors are welcome.

Project Description:
We will study a set of new attacks (such as front-running) for an adversary to obtain earnings without paying corresponding fees in decentralized finance projects enabled by blockchain. The nice feature transparency of blockchain also brings about new possible attacks.

We will analyze those attacks and propose better attack strategies to have a better understanding of those attacks.

Example Reference:
https://arxiv.org/pdf/2101.08778.pdf
https://arxiv.org/abs/2009.14021
https://arxiv.org/abs/2003.03810
Requirement to be on campus: No

CS2021-22/31 Accountable backdoors

Supervisors: Dr Qiang Tang

Eligibility: CS, math, and physics majors are welcome. Background of cryptography will be a big plus.

Project Description:
The increasing deployment of end-to-end encrypted communications services has ignited a debate between technology firms and law enforcement agencies over the need for lawful access to encrypted communications.
We will investigate the problem of constructing law enforcement access systems that mitigate the possibility of unauthorized surveillance. We may also study other accountability problems.

Example Reference:
https://eprint.iacr.org/2021/321.pdf

Requirement to be on campus: No

CS2021-22/32 Tiny ML in a big world

Supervisor: Dr Shuaiwen Song

Eligibility: Students with the following skill sets are preferred:

- Experience of writing neural networks on Tensorflow or
- Pytorch or other frameworks
- Experience with embedded system deployment
- System programming in C is required
- Eager to learn and explore
- Read and understand TVM
- Like hands-on exercises.

Project Description:
You should pick this project if you are interested in:

1. Tiny ML design, optimization, and deployment in modern edge devices (we will use the edge TPU as the test
2. platform)
3. Be the first to understand the challenges of designing Tiny ML models and their performance, power, and accuracy trade-offs.
4. Exploring runtime and compiler-level optimizations on
5. tiny devices
6. Using Meta-learning to make tinyML design possible
7. Exploring neural architecture search (building your own tiny Ml network)

Requirement to be on campus: No

CS2021-22/33 Compiler and Architecture Support for Future Quantum Systems

Supervisor: Dr. Shuaiwen Song and Alan Robertson

Eligibility: Students with the following skill sets are preferred:

- Some familiarity with quantum physics and quantum computing
- A strong background in maths
- Proficient at programming, an understanding of memory management

Project Description:
We are building quantum programming framework and defining architecture components for non-experts to write real algorithms with loop structures on quantum computers. If you are interested in quantum computing and want to learn more about this project, please join us!

Requirement to be on campus: No

CS2021-22/34 Bayesian Neural Network design and optimizations on leadership supercomputers

Supervisors: Dr Shuaiwen Song and Murali Emani (Argonie National Lab, Chicago)

Eligibility: Understanding C programming, MPI, OpenMP, OpenMPI; Knowledge of optimizing Deep learning codes; Good background in math and modelling experience using GPUs

Project Description:
Bayesian optimization is a methodology for sample-efficient learning and optimization. By leveraging a probabilistic model, it allows practitioners and researchers to explore large design spaces using only a small number of experimental trials. For this project, we will be looking into two real world DoE applications (Cosmic exploration for NASA and Cancer Genome analysis for NIH and DOE) and help them scale on large-scale supercomputers.

Requirement to be on campus: No

CS2021-22/35 Building Future Planet-Scale VR and AR systems via software-hardware co-design

Supervisors: Dr. Shuaiwen Leon Song (USYD), Qi Sun (NYU) and Michael Taylor (UW Seattle)

Eligibility:
- System programming knowledge
- Willing to learn Unity and Unreal and other system tools to build VR and AR environment
- Good programming skills, hands on, very passionate about research and publishing papers

Project Description:
In this project, students will work on experimenting with the state-of-the-art commercial virtual reality head-mounted- displays (HMDs) and identify different software and hardware level latencies and how they impact users' perception and feeling. Then, they will be instructed to analyse the graphics rendering pipeline, eye-tracking accelerator as well as different optical designs to discover the mapping between the motion/perception issues and the fundamental system-level design choices. Students will be using the testing platform we have built from the previous internship to help our current team to further build a cutting-edge collaborative rendering prototype system.

For more information, please read our ASPLOS'21 paper (preprint):
https://arxiv.org/ftp/arxiv/papers/2102/2102.13191.pdf

Requirement to be on campus: Yes

CS2021-22/36 Exploring concept drift detection for process mining

Supervisor: A/Prof Simon Poon

Eligibility: Good knowledge in Machine Learning and Data Mining, programming, and software engineering. Students interested in process management and process mining research are encouraged apply.
Project Description:
Business process management (BPM) is the discipline in which people use various methods to discover, model, analyse, measure, improve, optimize, and automate business processes. Recently, a new subject called process mining was invented which aims at bridging the gap between BPM and data science. Process mining is an integral part of data science which aims at discovering insights and actions directly from business process (event log) data.

An important application of process mining is to discover "changes" from data generated from business processes activities. Business processes are continuously evolving in order to adapt to changes due to various factors such as user behavioural responses to the Information Systems, and such changes could be unintended which may impact significantly to the business. Modern process-oriented businesses can get huge benefits if business changes can be detected and analysed.

In process mining, the event log data contains rare and infrequent behaviours are often treated as noise and ignored in the process mining process. However, such ignored infrequent behaviours could have significantly impact to concept drift detection.

Requirement to be on campus: No

CS2021-22/37 Evaluation of end users' perception of virtual hospital care during COVID-19

Supervisors: A/Prof Simon Poon and Dr. Neysa Petrina (in collaboration with RPA virtual hospital)

Eligibility: Good knowledge of information systems, statistical analysis and applications, appreciation of qualitative research method, interest in application of technology in clinical context. Students interested in pursuing Information Systems and health informatics evaluation research is encouraged to apply.

Project Description:
In the time of COVID-19, we have seen a rapid shift in the way healthcare is being delivered. The pandemic has intensified the need to integrate health technology with remote monitoring capability, so that continuity of care can be provided for patients, while minimizing the transmission of COVID-19.

The project aims to conduct an exploration of user perceptions of remote monitoring technology that are currently integrated into the RPA virtual hospital system. Prospective students will gain skills in:

- Application of health technology evaluation in a real-world context
- Understanding of research design and methodology
- Systematic and scientific procedure of data collection and analysis
- Evaluation of best research evidence to inform Evidence Based Practice in a clinical context.

Prospective student/s will be involved in the area of:

- Design of evaluation measure
- Quantitative and qualitative data collection and analysis.

Requirement to be on campus: No

CS2021-22/38 Evaluation of remote blood oxygenation monitoring device

Supervisor: A/Prof Simon Poon

Eligibility: Good knowledge in Machine Learning and Data Mining, programming, and software engineering.
Students interested in medical informatics are encouraged to apply.

Project Description:
During the ongoing coronavirus disease (COVID-19) pandemic, infected individuals show a wide range of symptoms. Particularly critical are respiratory failures and the accompanying low blood oxygenation levels (SpO2) that have the potential to cause severe hypoxemia in the

absence of dyspnea, called "silent hypoxemia". These imperceptibly falling SpO2 levels can become fatal when not carefully monitored and treated.

COVID-19 may become endemic despite a rising vaccination rate, leading to increased infection case numbers. This situation causes a need for patients with mild symptoms to be monitored remotely for the onset of deteriorating SpO2 levels, which can be measured with commercial pulse oximeters. The critical issue with these small non-invasive electronic devices is that they may suffer from device- or user-related measurement errors.

We have developed a device capable of measuring SpO2 and collected gigabytes of data. This project aims to develop methods, possibly based on deep learning, to improve measurement accuracy compared to a commercial device, identify potential user errors, and improve usability.

Requirement to be on campus: No

CS2021-22/39 Open-set Traffic Classification

Supervisor: Dr. Suranga Seneviratne

Eligibility: Machine Learning, Deep Learning, Python Programming

Project Description:
Despite the excellent performance of Deep Neural Networks in many machine learning tasks, most of these models were developed under the closed-set assumption, i.e. all possible classes are assumed to be present in the training set. Nonetheless, in real-world applications such as traffic classification, this assumption is not realistic due to the high number of available classes and the dynamic nature of systems that frequently introduce new classes. Using a model trained under the closed-set assumption in the real world would result in the model assigning open-set samples (unknown-unknowns) into one of the target classes (knowns), resulting in false positives that can have adverse effects on safety-critical applications. A more realistic scenario would be the open-set classification that aims to correctly classify all samples from a set of target classes while effectively rejecting samples from classes that do not belong to the target list (unknown-unknowns).

The aim of this project is to explore how the recent ideas of class anchor clustering can be used to increase the accuracy of open-set website fingerprinting and traffic classification.
Related Reading - https://arxiv.org/abs/2004.02434

Requirement to be on campus: No

CS2021-22/40 Self-supervised learning for Behavourial Biometrics

Supervisor: Dr. Suranga Seneviratne

Eligibility: Machine Learning, Deep Learning, Python Programming

Project Description:
Commonly used authentication methods in mobile, wearable, and IoT such as passwords and PINs are not observation resistant and less secure as many of us tend to reuse them. Attempts to increase the complexity of passwords by adding more constraints on character, numeral, and symbol combinations, reduce the usability due to high cognitive load. The other alternative of static biometrics such as fingerprints and face IDs are prone to replay attacks and spoofing. As a result, behavioural biometrics are emerging as an exciting new alternative. Behaviour biometrics focus on behaviour modalities that are potentially unique to individuals in the likes of typing patterns, gait, heart rate, and gestures. Behaviour biometrics are more secure as they are difficult to mimic, record, and synthesize and naturally can be used for implicit continuous authentication.

Many of the recent behavioural biometric solutions use deep learning as the underlying machine learning model which require large volumes of labelled training data. As of now, for a successful behavioural biometric solution, require several hundreds of training samples. The aim of this project is to use recent ideas of self-supervised learning in the context of behavioural biometrics and reduce the training data requirements. You will be provided with the EEG dataset used in our MusicID work [2] and your goal is to use self-supervised learning methods and achieve a similar level of performance, yet with less training data.

Related Reading -
[1] https://arxiv.org/abs/1907.11879
[2] https://arxiv.org/pdf/2006.01751.pdf

Requirement to be on campus: No

CS2021-22/41 DB4ML – In-Database Machine Learning

Supervisor: A/Prof Uwe Roehm

Eligibility: Strong academic background in Computer Science, especially databases, and C programming skills.

Project Description:
Modern data analysis requires scalable machine learning (ML) on up-to-date data. However, state-of-the-art is to copy data from operational systems into special ML platforms where data analysis is conducted periodically (e.g. daily). On the other hand, machine Learning algorithms are not well supported by existing database systems, which are the basis of the operational systems, as they typically iterate over the dataset until a convergence criterion is met – which is not supported by query languages such as SQL.

We have developed a novel approach to parallelise machine learning algorithms inside databases with multi-version storage layer, called DB4ML. In this project, we want to benchmark our existing DB4ML prototype against existing UDF-based approaches such as MADlib for PostgreSQL. We will use simple machine learning algorithms, such as clustering or PageRank. This project will give you a great introduction into machine learning with databases and has the potential for being used in a publication.

Requirement to be on campus: No

CS2021-22/42 Analysing SARS-CoV-2 Virus Sequences in BioSeqDB with GPU acceleration

Supervisor: A/Prof Uwe Roehm

Eligibility: Good academic background in Computer Science, especially databases, and some Bioinformatics background is helpful (though the latter can be picked up during the project too). Applicants should also be able to work with Linux and Windows command line tools.

Project Description:
One of the most important tools for controlling the spread of COVID-19 is the analysis of the SARS-CoV-2 virus sequences to identify new virus variants and how they spread across countries. In a pandemic situation as we experience right now, this has to be done fast and on scale.
We are currently working on BioSeqDB – an extension of SQL Server to support the management and analysis of DNA sequence data – which we use as a testbed to study hardware acceleration of complex data analysis tasks inside databases. In this summer project, you shall use the official SARS-CoV-2 sequence dataset to evaluate the storage and performance overhead of BioSeqDB when doing a sequence similarity search inside a BioSeqDB database – with and without GPU acceleration – versus using command-line based tools that run out-of-database. Note that this is in nature more a database performance evaluation project than it is a Bioinformatics project. Nevertheless, we will work with real data.

Requirement to be on campus: No

CS2021-22/43 Progressive Jupyter: Evaluation of Interactive Data Analysis with Jupyter Notebooks and Progressive DB

Supervisor: A/Prof Uwe Roehm

Eligibility: Strong academic background in Computer Science, especially databases and Python programming. Good programming skills in Python, a bit of Java, and SQL would be particularly useful. Experience in benchmarking would be desirable.

Project Description:
Human-centred data analysis requires interactive user interfaces that allow to interact with data visualisations in a natural way. This project aims to integrate a progressive SQL engine with Jupyter Notebooks - the main user interface for Data Scientists - so that users to interact with a data visualisation while the underlying query is still executing in the background. This is based on our existing research prototype called ProgressiveDB that splits an analytical SQL query into a series of smaller queries and produces a continuous stream of approximate query results with guaranteed interactive response times. We also already know how to use ProgressiveDB from Jupyter notebooks via the bokeh visualisation library.

In this project, we will extend and evaluate our existing prototype to demonstrate the capabilities of ProgressiveDB with interactive visual data analysis via a Jupyter notebook on a large dataset about the on-time performance of airlines, and evaluate its performance, latency and usability.

Requirement to be on campus: No

CS2021-22/44 ML-based Auto-Tuning for MongoDB

Supervisor: A/Prof Uwe Roehm

Eligibility: Strong academic background in Computer Science, especially databases, and programming skills. Some experience with machine learning would be of advantage.

Project Description:
MongoDB is a popular NoSQL database system which allows to both index and partition data distributed over multiple nodes for fast performance and scalability. An important design decision is which indexes to create and how data gets partitioned and on what attributes. While this is currently left to the users, we are working on automating it using reinforcement learning (RL).

In this project, we want to adapt an existing database advisor, which our research group has built for relational distributed databases, to MongoDB. This will involve adjusting the advisor's current database interface to MongoDB, and then to evaluate it using a standard benchmark with MongoDB. Most of the required code, especially the RL part, already exists from a previous project, written in Python. This project will give good insight into the usage of machine learning for the auto-tuning of database systems, as well as into MongoDB in particular. It is offered in collaboration with MongoDB.

Requirement to be on campus: No

CS2021-22/45 Deep neural network enhanced super-resolution and loss recovery for real-time video streaming

Supervisor: Wei Bao

Eligibility:

- Knowledge in computer networks (especially in UDP and HTTP DASH).
- Knowledge in deep neural networks.
- Experience in Python programming.

Project Description:
Real-time videos (video conferencing, online teaching, etc.) have experienced tremendous growth during the pandemic. UDP, instead of TCP, is adopted as the underlying transport-layer protocol to avoid delay, but it suffers from packet loss. Another issue is the limited bandwidth in the network. The sender sends low-resolution video when the network bandwidth is low, which drastically lowers users' Quality of Experience (QoE).

With the advances of deep neural networks (DNNs), a variety of DNN models have been designed to improve video quality. Super-resolution DNNs recover high-resolution videos from low-resolution ones, and loss-recovery (inpainting or interpolation) DNNs can fix distorted or missing frames. Without the need of investing on the improvement in network infrastructure, we can still drastically improve the QoE of real-time videos by DNNs.

You are expected to design a new DNN model to realise super-resolution and loss recovery. The model should be adaptive to the system environment (limited bandwidth and computational capacity), with small data size and small inference delay.

Requirement to be on campus: No

CS2021-22/46 Deep Learning based Video Generation

Supervisor: A/Prof Zhiyong Wang

Eligibility:
- 3rd year or higher
- Technique skills: strong programming skills and math

Project Description:
Recent success of deep learning has demonstrated a great potential to generative video content. This has opened a new door for innovation inn many domains, such as media, film, and game. This project aims to address the technical challenges of generating highly realistic video content by developing novel deep learning models. Students will gain comprehensive knowledge in computer vision, 3D modelling, computer graphics, and machine learning.

Requirement to be on campus: No

CS2021-22/47 Motion Capture Data Analysis

Supervisor: A/Prof Zhiyong Wang

Eligibility:
- 3rd year or higher
- Technique skills: strong programming skills and math

Project Description:
Human motion capture data has been widely used for character animation in film industry. However, it is very challenging to utilize such data effectively and efficiently in film production. This project aims to address the challenges in motion capture data driven film industry by developing novel analysis and processing techniques for better utilizing motion capture data. Students will gain comprehensive knowledge in computer vision, 3D modelling, and machine learning.

Requirement to be on campus: No

CS2021-22/48 Extreme-scale Visual Analytics of Big Complex Data

Supervisor: Professor Seokhee Hong

Eligibility: Strong Background in Data Structure/Algorithm, Strong Programming skills (Java, C++, Python)

Project Description:
Technological advances have increased data volumes in last few years, and now we experience a "data deluge". These big complex data have grown in importance due to international terrorism, success of genomics, complex software systems, and widespread fraud detection.

Visual analytics is the science of analytical reasoning facilitated by visualisation. This project aims to design/implement and evaluate new visualisation techniques of big complex data for analysts to find patterns and discover unexpected. Data sets include social networks, telephone call networks, biological networks, computer networks, and stock market networks.

These new visualisation methods are in high demand by industry as well as data scientists and analysts many application domains.

Requirement to be on campus: No

CS2021-22/49 Large-scale Graph Analysis

Supervisor: Dr Lijun Chang

Eligibility: Good at algorithms and programming

Project Description:
Graph model has been widely used to represent the relationships among entities in a wide spectrum of applications. For example, social networks are naturally modelled as graphs. We are nowadays facing a tremendous amount of large real-world graphs. Thus, there is a need of efficient techniques for processing large graphs. In this project, our aim is to design as well as implement algorithms to efficiently process large-scale graphs. The typical problems we will be investigating are dense subgraph (e.g., clique, k-plex, average-degree densest subgraph) computations over a large sparse graphs.

Requirement to be on campus: No