Engineering Vacation Research Internship Program

COMPUTER SCIENCE RESEARCH PROJECTS FOR SUMMER 2020/21

CS2020/1 Fake Image Detection................................................................. 3
CS2020/2 Geometric Networks in the Presence of Obstacles ..................... 3
CS2020/3 Routing in Geometric Networks............................................. 3
CS2020/4 Consumer-view measurement and tuning for MongoDB Atlas ........... 3
CS2020/5 Accelerated Distributed Optimization for Federated Edge Learning .... 4
CS2020/6 Extreme-scale Visual Analytics of Big Complex Data.................... 4
CS2020/7 Service Mining in IoT-based Smart Home.................................. 4
CS2020/8 Perturbation Attacks against Image Retrieval Systems.................. 4
<table>
<thead>
<tr>
<th>Paper ID</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>CS2020/9</td>
<td>Assessing the Portability of Encrypted DNS Traffic Signatures</td>
</tr>
<tr>
<td>CS2020/10</td>
<td>Speeding Up and Scaling Large Graph Analytics</td>
</tr>
<tr>
<td>CS2020/11</td>
<td>Lightweight transactions for in-memory databases</td>
</tr>
<tr>
<td>CS2020/12</td>
<td>Reasoning about Communication Quality in WSNs using Rewriting Logic</td>
</tr>
<tr>
<td>CS2020/13</td>
<td>Using Rewriting Systems to Reason about Security Conflict in Networks</td>
</tr>
<tr>
<td>CS2020/14</td>
<td>Bayesian Neural Network design and optimization in critical missions</td>
</tr>
<tr>
<td>CS2020/15</td>
<td>Tiny ML in a big world</td>
</tr>
<tr>
<td>CS2020/16</td>
<td>Can machine understand better and faster with scalable Reinforcement Learning + Meta Learning?</td>
</tr>
<tr>
<td>CS2020/17</td>
<td>Studying the impact of latency and optical designs on current and future virtual reality systems</td>
</tr>
<tr>
<td>CS2020/18</td>
<td>Maintaining perfect matchings under dynamic inputs</td>
</tr>
<tr>
<td>CS2020/19</td>
<td>Scalability Issues of Real-time Differential Computation over Streaming Data Processing Frameworks (Apache Spark, Flink and Timely Dataflow)</td>
</tr>
<tr>
<td>CS2020/20</td>
<td>Healthcare Monitoring and Prediction Using Wearable Activity Sensors</td>
</tr>
<tr>
<td>CS2020/21</td>
<td>Machine Intelligent Forex Trading</td>
</tr>
<tr>
<td>CS2020/22</td>
<td>Sleep apnea event prediction using machine learning</td>
</tr>
<tr>
<td>CS2020/23</td>
<td>Unravelling the Nascent Privacy Risks of 3D Spatial Mixed Reality Data.</td>
</tr>
<tr>
<td>CS2020/24</td>
<td>Real-time Encrypted Network Traffic Profiling with Deep Learning</td>
</tr>
<tr>
<td>CS2020/25</td>
<td>Characterising Newcomers in Social Media Networks</td>
</tr>
<tr>
<td>CS2020/26</td>
<td>Stance Prediction using Latent Subscribed Values</td>
</tr>
<tr>
<td>CS2020/27</td>
<td>Applied Data Science Projects on Electronic Medical Records</td>
</tr>
<tr>
<td>CS2020/28</td>
<td>Development of Social-Technical Measurement Model for Improvement in Telehealth</td>
</tr>
<tr>
<td>CS2020/29</td>
<td>Validating drug-drug interaction alerts in electronic medication management systems</td>
</tr>
<tr>
<td>CS2020/30</td>
<td>Development and evaluation of mobile app: eGuardian Angel</td>
</tr>
<tr>
<td>CS2020/31</td>
<td>Health Data Analytics</td>
</tr>
<tr>
<td>CS2020/32</td>
<td>Supporting Consumer Decision Using Online Reviews</td>
</tr>
<tr>
<td>CS2020/33</td>
<td>Efficient Streaming of 360 Degree Videos by Deep Video Content Analysis</td>
</tr>
<tr>
<td>CS2020/34</td>
<td>Creating the learning dashboard that gives the user control</td>
</tr>
<tr>
<td>CS2020/35</td>
<td>Enabling elderly and frail people to easily keep in touch</td>
</tr>
<tr>
<td>CS2020/36</td>
<td>Improving blockchain performance by using alternative storage platforms</td>
</tr>
<tr>
<td>CS2020/37</td>
<td>Progressive Jupyter: Interactive Data Analysis with Jupyter Notebooks</td>
</tr>
<tr>
<td>CS2020/38</td>
<td>Generating personalised hints in an online programming tutor using machine learning</td>
</tr>
<tr>
<td>CS2020/39</td>
<td>Intelligent Two-way Peer Review System for Critical Learning</td>
</tr>
<tr>
<td>CS2020/40</td>
<td>NLU2CMD: Translate a Natural Language Utterance into a Command (Bash terminal Shell)</td>
</tr>
<tr>
<td>CS2020/41</td>
<td>Toxic Language Detection in multiplayer online game</td>
</tr>
<tr>
<td>CS2020/42</td>
<td>Learned Partitioning Advisor for MongoDB</td>
</tr>
</tbody>
</table>
DATA ANALYTICS FOR RESOURCES & ENVIRONMENTS (DARE) PROJECTS..........................16

DARE2020/1 Emulator Models for Accelerating Inverse Problem Solutions ..................16
DARE2020/2 Improving Anomaly Detection with Transfer Learning ............................16
DARE2020/3 Sequential inference for large Bayesian networks ................................17
CS2020/1  Fake Image Detection
Supervisor: Zhiyong Wang
Eligibility: Strong programming skills and deep learning
Project Description:
Advances in digital image processing have produced a large variety of intelligent tools for manipulating and even generate new visual content, such as enhancing visual quality, removing or adding an object from an image, and generating images what do not physically exist. However, the processed visual content could be used to convey false or misleading information, which has been an emerging threat to our society trust, as the saying “seeing is believing” is not always true anymore. This project aims to develop advanced visual content analysis and machine learning techniques to establish the trustworthiness of digital images.
Requirement to be on campus: No

CS2020/2 Geometric Networks in the Presence of Obstacles
Supervisor: Andre 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

CS2020/3  Routing in Geometric Networks
Supervisor: Andre 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 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

CS2020/4 Consumer-view measurement and tuning for MongoDB Atlas
Supervisors: Alan Fekete and Uwe Roehm
Eligibility Criteria: Strong academic background in Computer Science, especially Database and/or Operating Systems. Experience in benchmarking or tuning systems would be desirable.
Project Description:
This project aims first to find ways to get good information about the performance of instances of MongoDB Atlas (a cloud-hosted document-store-as-a-service). Some performance monitoring is available from the cloud provider, and other metrics are reported by the instance; but both these are limited in details (eg not updated frequently enough). An alternative is to run microbenchmarks to determine how the instance is running; this project will try to determine suitable measurement practices, and combine them with the existing sources
for a more precise overall view. If time allows, we will aim to use this to tune the instance to improve its performance.

Requirement to be on campus: No

**CS2020/5  Accelerated Distributed Optimization for Federated Edge Learning**  
**Supervisor:** Nguyen Tran  
**Eligibility:** Knowing Python for Machine Learning, e.g. PyTorch, Tensorflow  
**Project Description:**  
Federated Learning is a decentralized machine learning paradigm that promises to achieve highly effective and learning performance, while preserving the data privacy of participating clients in a distributed network. There is growing interest in applying this technique to edge computing, forming federated edge learning. Compared with conventional distributed machine learning in a data center, federated edge learning faces non-independent and identically distributed (non-i.i.d.) and heterogeneous data, and the communications between edge workers, possibly through distant locations with unstable wireless networks, are more costly than their local computational overhead. However, edge workers are generally more computationally capable than devices in a cross-device federated learning network, making it possible to further exploit local computation to reduce the need for communication. This project will examine the current adoption of second-order optimization methods in federated edge learning and devise distributed approximate Newton-type algorithms according to that objective.  

Requirement to be on campus: yes (dependent on government’s health advice)

**CS2020/6  Extreme-scale Visual Analytics of Big Complex Data**  
**Supervisor:** Seokhee Hong  
**Eligibility:** Background in Data structure/Algorithm, Programming skills (Python, Java, C++)  
**Project Description:**  
This project focuses on improve convenience for the occupants in the smart home. Intuitively, convenience can be considered as the benefits brought by the intelligent system which is able to recognize the occupant’s ongoing activities and respond proactively at the right time and in the right place. This reduces the occupant’s interactions with the IoT services. Let us image a convenient life scenario in the occupant’s smart home. When the occupant wakes up in the morning, the heater in the bathroom starts to preheat. So, when the occupant goes into the bathroom to take a shower, it is already warm, and he does not need to wait. While he is taking a shower, the music player is turned on automatically to create a more relaxing showering experience. To fulfill aforementioned convenient life scenarios, our project coordinates multiple key tasks including activity discovery, activity recognition, activity prediction, service personalization, and movement detection.

Requirement to be on campus: No

**CS2020/7  Service Mining in IoT-based Smart Home**  
**Supervisor:** Prof Athman Bouguettaya  
**Eligibility:** Postgraduate coursework  
**Project Description:**  
This project focuses on improve convenience for the occupants in the smart home. Intuitively, convenience can be considered as the benefits brought by the intelligent system which is able to recognize the occupant’s ongoing activities and respond proactively at the right time and in the right place. This reduces the occupant’s interactions with the IoT services. Let us image a convenient life scenario in the occupant’s smart home. When the occupant wakes up in the morning, the heater in the bathroom starts to preheat. So, when the occupant goes into the bathroom to take a shower, it is already warm, and he does not need to wait. While he is taking a shower, the music player is turned on automatically to create a more relaxing showering experience. To fulfill aforementioned convenient life scenarios, our project coordinates multiple key tasks including activity discovery, activity recognition, activity prediction, service personalization, and movement detection.

Requirement to be on campus: No

**CS2020/8  Perturbation Attacks against Image Retrieval Systems**  
**Supervisor:** Suranga Seneviratne
Eligibility: Python programming, Knowledge in Deep Learning and Image processing.

Project Description:
Convolutional neural network (CNN) based image retrieval systems [1] are becoming popular and they have started to produce similar performances as traditional feature-based image retrieval methods. While it is known that CNNs are prone to adversarial examples in image classification tasks, their vulnerabilities when used in image retrieval systems remains less explored. A few studies [2] highlighted that carefully crafted perturbations can make the retrieval system return irrelevant images to the query at the top-ranking list. This project will evaluate some of the existing perturbation attacks on the USYD mobile app icon dataset [3] to understand how attackers can use such methods to publish mobile counterfeit apps (i.e. apps with icons very similar to popular apps) in app stores. We will also explore some ideas on how to fine-tune CNN based image retrieval systems so that they are more resilient to perturbation attacks.


Requirement to be on campus: No

CS2020/9 Assessing the Portability of Encrypted DNS Traffic Signatures
Supervisor: Suranga Seneviratne
Eligibility: Knowledge in Computer Networks and Python programming.

Project Description:
Domain Name System (DNS) translates human-readable domain names to numerical IP addresses so that packets can be routed through the Internet. For a very long time, DNS is considered as a privacy threat (allowing censoring and content filtering) as well as a bottleneck in the modern internet infrastructure due to the vulnerabilities of centralised DNS servers to DDoS attacks. As a result, encrypted DNS is getting popular with the availability of public resolvers from the likes of Google and Cloudflare. Recent works [1,2] showed that encrypted DNS traffic is prone to side-channel attacks and using the statistical properties of the DNS packet flows the websites that are being visited by the users can be predicted with high accuracy. This project aims to evaluate the consistency of encrypted DNS traffic signatures between different providers and try to assess the training data requirements of the attackers.


Requirement to be on campus: No

CS2020/10 Speeding Up and Scaling Large Graph Analytics
Supervisor: Lijun Chang
Eligibility: Strong programming skills in C or C++, Familiar with algorithm design and implementation

Project Description:
Graph is ubiquitous in modelling real-word data, as it captures the relationship information. With the rapid development of information technology, we are nowadays facing large graph data with billions of nodes and edges where the traditional algorithms/techniques are inefficient in processing these large graph data. In this project, our aim is to design efficient algorithms/techniques to speed up and scaling large graph analytics (e.g., finding dense subgraphs, computing shortest paths, identifying communities, etc).

Requirement to be on campus: No

CS2020/11 Lightweight transactions for in-memory databases
Supervisor: Baptiste Lepers and Prof Willy Zwenepoel
Eligibility: Eligibility: Very strong academic record; outstanding programming skills (preferably in C, C++); desirable: knowledge of operating systems, database internals, or programming language implementation

Project Description:
To provide a consistent view of the data to transactions, databases rely on Snapshot Isolation. Under Snapshot Isolation, every update generates a new version of the data, and previous versions are kept for as long as they belong to a transaction's snapshot. Long transactions
problematic in this context because they force the database to accumulate many old versions. Eventually the database may crash because it runs out of memory. This project will implement a new transaction model that was developed at the University of Sydney and offers the same consistency guarantees as Snapshot Isolation with low memory overhead.

**Requirement to be on campus:** No

**CS2020/12**  
**Reasoning about Communication Quality in WSNs using Rewriting Logic**  
**Supervisor:** Xi Wu  
**Eligibility:** Strong skills in mathematics and functional programming; basic knowledge of networks  
**Project Description:**  
Unreliability of communication links is a very common problem in wireless sensor networks (WSNs), and it may result in abnormalities and decrease the quality of service provided by a wireless system. The unreliability may be caused by many reasons, among which, node mobility as one of the essential and interesting features of WSNs, may significantly affect the reliability of communication links. A new process calculus was recently proposed, focusing on the communication quality issues in WSNs caused by node mobility. However, without a trustworthy semantics foundation, it is very hard to prove the correctness of the calculus and to demonstrate its effectiveness regarding the improvement of network communication quality. The objective of this project is to reason about the correctness of the behaviors of this formal framework using rewriting logic and to apply this formal framework on case studies.  

**Requirement to be on campus:** No

**CS2020/13**  
**Using Rewriting Systems to Reason about Security Conflict in Networks**  
**Supervisor:** Xi Wu  
**Eligibility:** Strong skills in mathematics and functional programming; basic knowledge of networks  
**Project Description:**  
With the explosive growth of networks and network applications, Software-Defined Networking (SDN) as a new paradigm in networking, has been proposed to use a programmable network infrastructure to separate the control plane and the data plane. The OpenFlow protocol is designed as the most common southbound interface between the control and forwarding layers of a SDN network architecture. Some approaches are recently proposed, paying attention to the security issues in OpenFlow-based networks. However, without a trustworthy semantics foundation, it is very hard to prove the correctness of these approaches, which may cause the false alarm. The objective of this project is to reason about the correctness of the behaviors of flow rules based on a formal framework using rewriting systems, with a focus on detecting potential conflicts between flow rules imposed by dynamic OpenFlow applications.  

**Requirement to be on campus:** No

**CS2020/14**  
**Bayesian Neural Network design and optimization in critical missions**  
**Supervisor:** Shuaiwen Song  
**Eligibility:** Students with the following skill sets are preferred:  
1. understanding workload characterization and performance analysis methods  
2. familiar with FPGA implementation and RTL level design  
3. experience with Verilog or VHDL or high-level synthesis language.  
**Project Description:**  
In this project, we will investigate the training and inference challenges of Bayesian NN and analyse the best design option in the current accelerator landscape. The project will result in a FPGA or FPGA+RISC-V end-to-end prototype. We also want to test it in critical missions such as autonomous driving.  

**Requirement to be on campus:** No

**CS2020/15**  
**Tiny ML in a big world**  
**Supervisor:** Shuaiwen Song  
**Eligibility:** Students with the following skill sets are preferred:  
1. Experience of writing neural networks on Tensorflow or Pytorch or other frameworks
Experience with embedded system deployment
3. System programming in C is required
4. Eager to learn and explore
5. Read and understand TVM
6. 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 platform)
2. Be the first to understand the challenges of designing TinyML models and their performance, power and accuracy trade-offs.
3. Exploring runtime and compiler-level optimizations on tiny devices
4. Using Meta-learning to make tinyML design possible. Exploring neural architecture search (building your own tinyml network)

Requirement to be on campus: No

CS2020/16  Can machine understand better and faster with scalable Reinforcement Learning + Meta Learning?
Supervisor: Shuaiwen Song and Xu Chang
Eligibility: Students with the following skill sets are preferred:
1. Experience of writing neural networks on Tensorflow or Pytorch or other frameworks
2. Interested in understanding new neural networks and customized designs.
3. Understand the basic concept and algorithmic features of meta-learning

Project Description:
In recent years, deep RL (reinforcement learning) systems have been widely-adopted in areas like robotics, games, and natural language processing. However, the major limitations of deep RL are their demand for massive training data and time, and their poor ability to be adapted in new environments (distribution). On the other hand, Meta-learning tries to train a reasonable model with only a few training examples. This summer vacation project aims to extend the novel meta-learning approaches to be applied in deep RL. We will consider prospects for extending and scaling up deep RL plus meta-learning and investigate their important implications (e.g., neuroscience).

Requirement to be on campus: No

CS2020/17  Studying the impact of latency and optical designs on current and future virtual reality systems
Supervisor: Shuaiwen Song and Juno Kim(UNSW)
Eligibility: Students with the following skill sets are preferred:
1. Understanding of system architecture, OS, compiler
2. Good knowledge with C system programming
3. Experience with using Unity or Unreal engine (not required)
4. Experience with writing simple neural networks tested on Tensorflow or Pytorch
5. Curious of learning new techniques and designs

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.

Requirement to be on campus: No

CS2020/18  Maintaining perfect matchings under dynamic inputs
Supervisor: William Umboh
Eligibility: Ability to reason about algorithms mathematically, e.g. a grade of D+ in COMP3027/COMP3927 or equivalent course on algorithms.
Project Description:
Bipartite matchings have important applications such as job scheduling and content delivery. In these applications, the two sides of the bipartite graph model servers and clients and an edge between a client and a server model which servers can serve each client. A critical aspect of these applications is that clients arrive over time and each client needs to be matched immediately on arrival to a free server without any knowledge about future clients. We are also allowed to re-match previous clients, but re-matchings can be expensive: in the context of job scheduling, re-matching a client to a different server requires interrupting its job. Thus, we would like to minimise the number of re-matchings.
The aim of this project is to develop and analyse algorithms for the dynamic matching problem with few re-matchings

Requirement to be on campus: No

CS2020/19 Scalability Issues of Real-time Differential Computation over Streaming Data Processing Frameworks (Apache Spark, Flink and Timely Dataflow)

Supervisor: Prof. Albert Zomaya, Reza HoseinvF

Eligibility:
- Good knowledge of at least one programming language, preferably Python, Rust, Scala, Java, C++
- Writing reusable, testable, and efficient code
- Familiarity with Linux, SSH
- Understanding of distributed and multi-threading programming paradigm
- Experience working with Git
- Work closely with the rest of team to solve problems, and transfer knowledge
- Interest in and commitment to learn Design and implementation of low-latency and performant computer systems
- Interest in and commitment to learn and work with Timely Dataflow, Apache Flink, Storm or Apache Spark platforms
- ML/AI/Data mining algorithms: Basic Concepts

Project Description:
Today, more companies are facing with a huge amount of streaming data that needs to be quickly processed in a real time fashion to extract meaningful information (e.g. by applying sophisticated machine learning algorithms). Apache Spark, Flink, Strom and Timely Dataflow are recent stream-data processing platforms which are developed as a mean for increasing the speed of conventional distributed computation over large-scale clusters.

Differential computation, on the other hand, can be defined as a data processing paradigm to efficiently process large volumes of data and to quickly respond to "arbitrary changes" in input collections.

A differential computation responds to arbitrary changes to its initially empty input collections, reporting the corresponding changes to each of its output collections (in either a batch or streaming mode).

There is not comprehensive study yet to show the effectiveness of modern processing platforms for performing differential computation, particularly in presence of high arrival rate of streaming data within short periods (e.g. if such incidents can cause serious degradation to the overall performance of underlying system).

In this project, we are looking for conducting a systematic study to answer the fundamental question what is the core bottlenecks of modern data processing systems (Spark/Flink/Storm/TDF) when performing differential computation across large scale cluster, and if we can develop advanced optimization mechanisms to allocate effectively available computing resources for such execution as the number of nodes in such a cluster grows. We also need to develop a comprehensive monitoring systems to get insight into overall Spark/Flink/Storm/TDF cluster-wide utilization and resource bottlenecks (such as CPU, memory and disk usage) when running compute and I/O intensive differential computation over different machine learning (ML), Data mining (DM) and Artificial Intelligence (AI) algorithms.

Requirement to be on campus: No
CS2020/20  Healthcare Monitoring and Prediction Using Wearable Activity Sensors
Supervisor: Anusha Withana
Eligibility: You will work with a PhD student, and we expect you are a fast learner. Excellent skills in programming, skills in embedded systems, machine learning, knowledge in design and fabrication, and human computer interaction are added benefits.
Project Description:
Preventive measures for many health risks depend on the continuous measurement of physical and physiological activities. For instance, cardiovascular diseases are the leading contributor of death and acquired disability. Risk pools are often identified but researchers lack accessible sensing technologies to monitor them over long periods of time. In this project we aim to develop affordable technologies to monitor physical and physiological activities. Particularly, we are interested in creating technologies that blend with users’ body, cloths and accessories that allow us to collect long-term data. This project not only involve data collection and analysis, but also computational design and fabrication of attractive and fashionable wearable sensors.
Requirement to be on campus: No

CS2020/21  Machine Intelligent Forex Trading
Supervisor: Matloob Khushi
Eligibility: Machine/Deep Learning Knowledge
Project Description:
Every change of trend in the forex market presents a great opportunity as well as a risk for investors. Accurate forecasting of forex prices is a crucial element in any effective hedging or speculation strategy. However, the complex nature of the forex market makes the predicting problem challenging, which has prompted extensive research from various academic disciplines. Data mining methods including machine and deep learning have been implemented successfully to predict the future prices. Therefore in this project students will apply and innovate price prediction pipeline using deep learning models. See prior project videos https://bit.ly/2DFPI60.
Requirement to be on campus: No

CS2020/22  Sleep apnea event prediction using machine learning
Supervisor: Irena Koprinska
Eligibility: Machine learning skills, e.g. completed COMP3308/COMP3608 or COMP5318, Good programming skills
Project Description:
Sleep apnea is a breathing disorder characterised by periods of reduced breathing or no breathing at all during sleep. It affects 4% of the adult population, causing daytime sleepiness, and can also lead to heart attacks, diabetes and even death. The goal of this project is to predict sleep apnea events in advance using machine learning techniques. This will allow for medical devices to pre-emptively intervene. The project will use a large dataset containing data from multiple signals recorded overnight, e.g. respiratory, brain activity and heart rhythm.
Requirement to be on campus: No

CS2020/23 Unravelling the Nascent Privacy Risks of 3D Spatial Mixed Reality Data
Supervisor: Kanchana Thilakarathna
Eligibility: The ability and desire to experiment with real devices, e.g. Oculus and HoloLens; and knowledge in applied machine learning.
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. In light of that, this project focuses on holistic 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

### CS2020/24 Real-time Encrypted Network Traffic Profiling with Deep Learning

**Supervisor:** Kanchana Thilakarathna  
**Eligibility:** Knowledge on applied machine learning and computer networking basics are desirable.

**Project Description:**
Providers of large, enterprise-class networks find it hard to track hosts, servers and other vulnerable assets in their networks. Network profiling systems provide valuable insights to the assets on a network and their purpose. A network profile enables providers to better consider how configuration changes will impact networks, and security administrators to identify suspicious activity. However, effective network profiling under real world conditions is increasingly challenging. The primary focus of this research is to develop means to address issues in traffic profiling imposed by real-time constraints such as high-speed networking and ubiquitous encryption. The project aims to develop a network profiling method based on deep learning operating at high real-time speed. This project is a collaboration with Data61.**

**Requirement to be on campus:** No

### CS2020/25 Characterising Newcomers in Social Media Networks

**Supervisor:** Dr. Basem Suleiman

**Eligibility:** Very good knowledge of algorithms, data structure, database, and graph network analysis. Very good programming skills; (preferably Python, R

**Project Description:**
A community structure is a very common feature for many real-world networks in many fields including communication, social, collaboration, and biological networks. Finding and connecting communities in large scale networks has many applications, e.g., discovering users with a common location, finding friends sharing similar interests or occupations in an online social network, how information is cascaded on social media etc. In various types of structured communities newcomers choose their interaction partners by selecting a role-model and copying their social networks. A role model could simply be defined by the centrality of the node. In this project, we focus on social networks. The goal of this project is to:

(a) design an algorithm to characterize new users in a social network and assign them to their corresponding communities only based on the topology of the network.  
(b) Evaluate the effectiveness of the designed algorithm through conducting laboratory experiments

**Requirement to be on campus:** No

### CS2020/26 Stance Prediction using Latent Subscribed Values

**Supervisor:** Dr. Basem Suleiman

**Eligibility:** Experience in programming languages such as JavaScript, Python  
- Good knowledge of machine learning, data science/analytics  
- Working knowledge of web programming, MongoDB, MySQL/SQL

**Good to have (but not essential):** Knowledge of AWS and GitHub

**Project Description:**
Stance prediction in social media helps to identify opinions of very large numbers of users on different issues including marketing and political campaigning. People’s stance about certain topics is generally determined as a predefined explicit set of values the individual is subscribed to. Often people’s stance changes over time and this temporal activity that is implicit tells a lot about the future stance of the individual about certain topics. This project focuses on finding implicit common factors in people that are better predictive of their stance rather than their explicit predefined values. The goal of this project is to develop algorithm and UI for predicting stance on certain topics from social media. This includes:

(1) Extract related tweets based on a predefined set of topics.
(2) Classify the users’ stances based on their tweets using AI-based sentiment analysis
(3) Backtrace the users’ activity to discover the common traits between the same stance users.

**Requirement to be on campus:** No

**CS2020/27  Applied Data Science Projects on electronic medical records**

**Supervisor:** A/Prof Simon Poon, Dr Charmaine Tam, Centre of Translational Data Science

**Eligibility:** Depending on the project, the candidate should have very strong programming and analytical skills. They will have a basic knowledge of statistics and machine learning and a willingness to learn and contribute to open source software projects that requires documentation. They will have a high attention to detail and an enthusiasm for working with real-world data that has potential to make significant impacts on the health system. Students interested in pursuing Honours or Research programs are encouraged to apply.

**Project Description:**
The widespread adoption of electronic medical records (eMR) has created unprecedented opportunities to harness information stored within eMRs to ascertain clinical outcomes and improve healthcare efficiency at large scale and low cost. eMRs contain vast and heterogeneous data sources ranging from structured data to text, reports and images. **Data sources:** Existing publicly available datasets as well as electronic medical records extracted in a 5 year dataset from two local health districts in Sydney.

**Potential projects include:**
1) Information extraction from clinical text to risk stratify patient presentations
2) Applying algorithms to identify a patient’s risk of returning to hospital
3) Data visualisation of safety and quality indicators to better assist clinicians, hospital and government stakeholders with decision making

**Requirement to be on campus:** No

**CS2020/28  Development of Social-Technical Measurement Model for Improvement in Telehealth**

**Supervisors:** A/Prof Simon Poon, Dr Neysa Petraña, A/Prof Georgina Luscombe, School of Rural Health

**Eligibility:** Basic knowledge of statistics, Good knowledge of qualitative research method, interest in application of technology in clinical context. Students interested in pursuing Honours or Research programs are encouraged to apply.

**Project Description:**
In the time of COVID-19, the process of integrating telehealth into general practice service delivery had to be done rapidly. The standard face-to-face consultation has now been partially or fully replaced by the use of telehealth technology so that continuity of care can be provided for patients, while minimizing the transmission of COVID-19. Telehealth is the use of information and communication technologies (both voice and video calls modality) by GP for diagnosis, treatment, monitoring and prevention of disease and injuries. This project aims to conduct an exploration into the use and acceptance of telehealth by General Practitioners (GP) in NSW. Using a mixed method approach, we will design, develop and implement a measurement model to capture the level of GP telehealth acceptance.

**Requirement to be on campus:** No

**CS2020/29  Validating drug-drug interaction alerts in electronic medication management systems**

**Supervisors:** A/Prof Simon Poon, Dr Jonathan Penm, School of Pharmacy Health, Dr Angus Ritchie Sydney Local Health District, Professor Sid Patanwala, School of Pharmacy

**Eligibility:** Familiarity with drug-drug interactions is preferred. Knowledge of database operations is advantageous. Students interested in pursuing Honours or Research programs are encouraged to apply.

**Project Description:**
Introduction: Drug–drug interaction alerts are displayed every day from computerized provider order entry (CPOE) systems in hospitals and general practice. However, many of these alerts are considered irrelevant, with low clinical value and result in alert fatigue. In this study, we aim to identify and evaluate the clinical relevance of drug-drug interaction alerts
commonly identified in electronic medication management systems in Sydney Local Health District and General Practice.

Methods Commonly identified drug-drug interactions from the most common computerized provider order entry (CPOE) systems will be investigated in this study. Drug-drug interactions data will be extracted from eMeds (Cerner Millennium, Kansas), Best Practice (Best Practice Software, Mascot) and Medical Director (Health Communication Network Pty Limited, Haymarket).

Commonly identified drug-drug interactions alerts will be presented to health care professionals (prescribers and pharmacists) to identify their clinical relevance and severity. Computerized presentation of DDI alerts will also be evaluated.

**Requirement to be on campus:** Yes (dependent on government’s health advice)

---

**CS2020/30 Development and evaluation of mobile app: eGuardian Angel**

**Supervisors:** Associate Professor Simon Poon, Professor Julie Redfern (Westmead Applied Research Centre)

**Eligibility:** Relevant Skills: Mobile app design and development, user evaluation methodology; understanding/interest in social determinants of health. Communication skill is critical for user requirements gathering & evaluation. Students interested in pursuing Honours or Research programs are encouraged to apply.

**Project Description:**

The Guardian Angel concept proposes the idea that a large, decentralised social support network can more effectively motivate individual in a connected community to collectively improve a population’s health and lifestyle habits than a traditional centralised system of a few localized hubs, e.g., health care professionals (e.g. clinicians), monitoring a large number of spokes (e.g. patients). The Guardian Angel idea is to organise patients, so that each individual has a guardian angel (or more than 1 guardians) from among the other patients (defined as children) and is assigned to be a guardian angel for someone else. The angel has the ability to engage with the other person’s activities and then encourage and motivate the person to continue on a positive trajectory (to accomplish health goal). Every person tries to motivate someone and is motivated by someone else so that the health and lifestyle of the whole community improves.

This summer project is a part of larger project, consisting two sub-parts:

- Mobile application development: This part of the project would involve developing a mobile application prototype to enable the Guardian Angel Network. One aspect is to identify functionalities to encourage and motivate, and the other aspect is about user interface design.

**Evaluation:** This part of the project would involve designing and implementing a suitable clinical trial to test and evaluation the usability & usefulness of the mobile Guardian Angel application.

**Requirement to be on campus:** No

---

**CS2020/31 Health Data Analytics**

**Supervisor:** Josiah Poon

**Eligibility:** Good programming, data analytics and statistical skill; knowledge in data mining & machine learning preferred.

**Project Description:**

Multiple Sclerosis (MS) is a common chronic neurological disease that is found among young adults in the age range of 20 to 40 and. Its effect among women is three times more than men. MS not only has a devastating and profound effect on the quality of life, and the worst is that it is no cure. Hence, anything to improve the condition is a welcome. We have our collaborators who have carried out an integrative treatment to a recruited group of patients. Some preliminary works have been done. The aim is to extend the analysis and creative visualisation to inform the clinicians the best clinical approach and decision.

**Requirement to be on campus:** No

---

**CS2020/32 Supporting Consumer Decision Using Online Reviews**

**Supervisors:** Kevin Kuan

**Eligibility:** Basic understanding in Python for data mining and R for statistical analysis.
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

CS2020/33 Efficient Streaming of 360 Degree Videos by Deep Video Content Analysis
Supervisor: Kanchana Thilakarathne
Eligibility: Knowledge on applied machine learning and computer networking basics are desirable. Mobile programming (iOS or Android) experience will be an added advantage.
Project Description:
360° videos are a popular application of virtual reality. However, streaming 360-videos requires high bandwidth consumption. Tile-based streaming, which partitions a video frame into tiles and sends selected tiles based on user field-of-view (FoV) can fail if user FoVs are not available in real-time. This project aims to predict future user FoVs by analysing content features and using these predictions for efficient tile partitioning.
Firstly, you will investigate different psychological factors that affect visual attention such as the contextual relationships between objects. Existing research shows that humans tend to be attracted to faces and text. However, there is plenty of untapped psychological research such as semantic guidance which you will put into practice. You will then focus on developing a novel content-based tile-distribution that allocates different quality levels for tiles leveraging methods such as DNNs. Finally, you will evaluate these approaches by developing an end-to-end 360-video streaming platform.

Requirement to be on campus: No

CS2020/34 Creating the learning dashboard that gives the user control.
Supervisor: Judy Kay
Eligibility: These projects have different facets. Students with strong technical programming knowledge can build on those aspects. Students with HCI, Usability and Design skills can focus on user aspects. Students with strong skills in synthesising literature can work in that area.
Project Description:
For effective learning, feedback on progress is very important. This is true for formal learning, such as university subjects. It also holds for lifelong and life-wide learning, such as individuals who want to use personal tracking data to be healthier and clinicians who want to use clinical data to become more effective at their work. One challenge is that in a world where there are huge amounts of personal data that could help people learn, it is simply too hard for people to get the information they need at an interface that helps them harness that data to improve their learning.
The projects build on decades of education research on metacognition, self-regulated learning bringing this together with computer science research on artificial intelligence in education (AIED), human computer interaction (HCI) and information visualisation (InfoViz). The overall goal is to create a set of personalised learning dashboard interfaces (called Open Learner Models).

Requirement to be on campus: No

CS2020/35 Enabling elderly and frail people to easily keep in touch
Supervisor: Judy Kay
Eligibility: These projects require strong programming skills and will involve building prototype systems. Experience with mobile platforms would be an advantage.
This project has different facets. Students with strong technical programming knowledge can build on those aspects. Students with HCI, Usability and Design skills can focus on user aspects.
Students with strong skills in synthesising literature can work in that area.
Project Description:
Social connections are important for everyone but are critically important for older people. There is considerable research that shows that regular social interaction helps to delay onset of
cognitive decline and frailty. People in care-homes are particularly at risk of loss of social connection since they lose their regular contact with family and friends. This project involves the design and implementation of novel devices that make it easier for elderly people to "keep in touch" with their family and carers.

**Requirement to be on campus:** No

**CS2020/36 Improving blockchain performance by using alternative storage platforms**

**Supervisors:** A/Prof Bernhard Scholz, Dr Shuaiwen Song, Prof Alan Fekete

**Eligibility:** Very strong academic record; outstanding programming skills (preferably in at least one of Rust, C++ and/or Go); desirable: knowledge of operating systems, database internals, or programming language implementation.

**Project Description:**

Modern blockchain clients support "smart contracts" which are executed against a storage layer which maintains the world-state. Several important clients (eg Parity, Geth) for the Ethereum blockchain use state-of-art database management platforms, such as RocksDB or LevelDB, to store the world-state. Experiments have shown that these platforms can become a major bottleneck impeding performance of the smart contracts. This project will port clients to use a new storage layer, which was developed in the University of Sydney, and can take advantage of the latest hardware advances in storage such as NVMe.

**Requirement to be on campus:** No

**CS2020/37 Progressive Jupyter: Interactive Data Analysis with Jupyter Notebooks**

**Supervisor:** Uwe Roehm

**Eligibility:** Strong academic background in Computer Science, especially databases and Python programming. This is a rather technical project which requires good programming skills in Python, a bit of Java, and SQL

**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 can 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 which we have implemented already which splits an analytical SQL query into a series of smaller queries on an underlying database, and then produces a continuous stream of approximate query results to the client with guaranteed interactive response times. In this project, we will now extend an existing Jupyter kernel extension for databases, so that we can demonstrate the capabilities of ProgressiveDB with interactive data analysis via a Jupyter notebook on-top of a multi-GB data set about the on-time performance of airlines.

This is a rather technical project which requires good programming skills in Python, a bit of Java, and SQL.

**Requirement to be on campus:** No

**CS2020/38 Generating personalised hints in an online programming tutor using machine learning**

**Supervisors:** Benjamin Passen, Irena Koprinska and Kalina Yacef

**Eligibility:** Good programming skills, and foundations of machine learning, neural networks (eg COMP3308 or COMP5318). Interest in computer programming education is very welcome.

**Project Description:**

This project is using machine learning to generate automated next-step hints in an online computer programming tutor by taking the student’s current code as input and returning a suitable hint. Such techniques have been shown to improve learning gains and speed, but they are more efficient if they are personalised to the student. Using our existing variational autoencoder that transforms the program into a point in space and vice-versa, this project will consist in integrating features from the student data into the last layer of the encoding/decoding stage to personalise the neural network to that student and therefore derive better hints.

**Requirement to be on campus:** No
CS2020/39 Intelligent Two-way Peer Review System for Critical Learning

**Supervisors:** Caren Han and Josiah Poon

**Eligibility:** Student has done any of following courses:
- COMP5046 (Natural Language Processing)
- COMP5329 (Deep Learning)

**Project Description:**
Peer assessment, or self-assessment, is a process whereby students or their peers grade assignments or tests based on a teacher’s benchmarks. The practice is employed to save teachers time and improve students’ understanding of course materials as well as improve their metacognitive skills. However, some academic staff hesitate to use a peer assessment approach for their classes due to concerns about its credibility and reliability. The focus of our research is to detect the credibility level of each assessment performed by students during peer assessment. We will apply machine learning and deep learning models in order to make intelligent peer assessment system. This will enhance the credibility and reliability.

**Requirement to be on campus:** No

CS2020/40 NLU2CMD: Translate a Natural Language Utterance into a Command (Bash terminal Shell)

**Supervisors:** Caren Han and Josiah Poon

**Eligibility:** Student has done any of following courses:
- COMP5046 (Natural Language Processing)
- COMP5329 (Deep Learning)

**Project Description:**
The project aims to build an algorithm that can translate an English description (nlc) of a command line task to its corresponding command line syntax. It solicits entries that can translate a given natural language utterance into a command to be executed on the Bash terminal shell. For example, “show me a list of all files” should produce something similar to “ls” as the predicted command. The project features two tasks: the first is the accuracy track, which is measured in terms of whether the right utility (for example, “ls”) is predicted, along with the correct flags required for it to complete the required task. The second track is the efficiency track — energy efficiency is increasingly an important consideration for AI and Machine Learning models, and the aim of this track is to encourage systems that are judicious in their energy consumption. As part of a previous survey on the Command Line AI (CLAI) open source offering from IBM, we had a chance to collect some community feedback on features that are important in an AI assistant for the Bash terminal.

**Readings and resources:**
https://github.com/IBM/clai
https://github.com//TellinaTool//nl2bash

**Requirement to be on campus:** No

CS2020/41 Toxic Language Detection in multiplayer online game

**Supervisors:** Caren Han and Josiah Poon

**Eligibility:** Student has done any of following courses:
- COMP5046 (Natural Language Processing)
- COMP5329 (Deep Learning)

**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 behavior, 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://www.kaggle.com/romovpa/gosuai-dota-2-game-chats
DATA ANALYTICS FOR RESOURCES & ENVIRONMENTS (DARE) PROJECTS

DARE2020/1 Emulator Models for Accelerating Inverse Problem Solutions

Supervisor: Richard Scalzo

Eligibility: Programming in Python including standard scientific computing libraries (numpy, scipy, matplotlib). Familiarity with mathematical foundations for machine learning (multivariable calculus and linear algebra). Familiarity with contemporary machine learning frameworks such as scikit-learn, Google Jax, PyTorch or Keras desirable.

Project Description:
Inverse problems in geology and geophysics -- inferring the history and structure of a region of the Earth's crust from observations -- occur in a range of applications including mining, groundwater, and natural hazard assessment. These problems are by their nature ill-posed and uncertain, and adequately exploring and characterizing the space of possibilities consistent with the data involves repeated evaluations of numerical models for different geophysical sensors and their derivatives with respect to parameters. These models can be computationally intensive and/or black boxes, making it difficult to scale their use to high-dimensional problems.
This project will investigate emulators, or proxies for the numerical likelihood, that can be learned to create adaptive, scalable methods for quantifying uncertainty in high-dimensional inverse problems. Possible emulator methods might include Gaussian processes, Bayesian neural networks, and/or probabilistic graphical models.

Requirement to be on campus: No

DARE2020/2 Improving Anomaly Detection with Transfer Learning

Supervisors: Simon Luo and Lamiae Azizi

Eligibility: The candidate should have experience implementing statistical machine learning models in python (not just using packages)

Project Description:
Anomaly detection is a challenging problem in machine learning because there are very few training examples of the anomaly to fit a model. Often times there could be related datasets which can be used to improve the learning process. However, currently there are no good ways of combining the models we have trained on different datasets. We are interested in developing new transfer learning techniques to combine anomaly detection models that have been trained on different datasets. The research project involves a literature review of the
techniques currently used and implementing the state-of-the-art techniques in python. For more details please contact s.luo@sydney.edu.au

Requirement to be on campus: No

DARE2020/3 Sequential inference for large Bayesian networks

Supervisors: Wanchuang Zhu, Sally Cripps

Eligibility: Background from Statistics, Mathematics or computer science. Knowledge of Bayesian inference. Experience with R

Project Description:
Bayesian network is a powerful framework to model hierarchical structure of multiple variables whose independencies are represented by the underlying directed acyclic graph (DAG). The classic inference of a Bayesian network includes structure learning and parameter learning, among which the structure learning poses challenges to researchers. The challenge originates from the discontinuity and huge size of the parameter space. Many methods have been proposed to alleviate the challenge, including structure MCMC, order MCMC and partition MCMC. However, performance of these approaches are not satisfactory in the following perspectives: (a) the approaches are not robust with respect to the starting point of DAG structure; (b) the approaches are incapable to explore full parameter space especially for large Bayesian networks. The goal of this project is: (a) to measure and evaluate robustness of existing methods; (b) to reduce the size of parameter space by using sequential inference techniques.

Requirement to be on campus: No