THE UNIVERSITY OF SYDNEY

FACULTY OF ENGINEERING

Engineering Vacation Research Internship Program

COMPUTER SCIENCE RESEARCH PROJECTS FOR WINTER 2021

CS2021/1 Distributed streaming processing across a large-scale setting made of Raspberry Pi Pico

Supervisors: Prof Albert Zomaya & Dr M.Reza HoseinyF.
Eligibility:

- Good knowledge of C/C++ and Python programming languages
- Writing reusable, testable, and efficient code
- Familiarity with Linux, SSH
- Understanding of the 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 applications
- Interest in and commitment to learn and work with other streaming processing engines such as Timely Dataflow

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 found in most today's applications.
In this project, we need to develop a C/C++ software framework, a computational model, and a set of synchronization mechanisms that enable developers to write more sophisticated streaming computations (e.g. with iterative control flow) to be running across a large scale deployment of several Pico microcontrollers when performing distributed real-time data streams analytics especially when the application contains many pipelined transformations with narrow dependencies (such as the case for Data-Flow applications). 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 software layer 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.
The layer also needs to effectively allocate the available limited computing resources of the microcontroller into the software components.
In the next step we will measure the run-time performance of running such algorithms and the developed platform that includes I/O, computer resources, and latency of computation for handling real-time data in such the design architecture.
Reading material:
1- Raspberry Pi Pico C/C++ SDK documentation
2- Timely Dataflow architecture and its Progress Tracking documentation
Requirement to be on campus: Yes (dependent on government's health advice)

CS2021/2 Self-supervised Learning for Traffic Fingerprinting

Supervisor: Dr Suranga Seneviratne
Eligibility: Machine Learning, Deep Learning, Python Programming
Project Description:
Despite the increasing adoption of end-to-end encryption, our Internet communications are not as private as we think and even a weak adversary such as a passive observer can infer what we do over the Internet. Statistical properties of data traffic such as packet sizes, packet numbers, inter-packet times remain visible to passive eavesdroppers, opening the door to traffic fingerprinting attacks. Many of such attacks are based on machine learning models, especially deep neural networks that require large volumes of labelled training data. As of

now, for a successful attack, several hundreds of training samples are required from each target traffic class. The aim of this project is to use recent ideas of self-supervised learning in the context of traffic classification and reduce the training data requirements of traffic fingerprinting attacks.
Requirement to be on campus: No

CS2021/3 Data privacy in a distributed model

Supervisor: Dr Clement Canonne
Eligibility Criteria: Some mathematical maturity; basics of discrete statistics and probability; having taken classes on algorithms/data structures.
Project Description:
There are several principled approaches to guarantee data privacy while still being able to run algorithms. Most of such current techniques fall under differential privacy (DP). One promising variant of DP is shuffle privacy, which has 4 components:

- the users: each holds part of the data
- a "shuffler:" randomly permutes the users' messages (anonymization)
- the center: runs an algorithm on the messages
- the outside world: observes the algorithm's output
- The idea is that shuffling greatly improves the privacy guarantees.

In this project, the student will consider scenarios where the shuffler does not randomly shuffle the messages but does something weaker (simpler to implement.) E.g., the dataset comes from groups of users, where each group has its own shuffler: the data is randomly permuted within each group, but not between groups. The center knows which group a given data point comes from, just not from which user in it. How does that affect privacy?
Requirement to be on campus: No

CS2021/4 Building future planet-scale VR and AR systems via software-hardware co-design

Supervisors: Dr Shuaiwen Leon Song, Qi Sun (NYU), &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 (dependent on government's health advice)

CS2021/5 Bayesian neural network design and optimizations on leadership supercomputers

Supervisors: Dr Shuaiwen Leon Song, Murali Emani (Argonne 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

CS2021/6 Tiny ML in a big world

Supervisor: Dr Shuaiwen Leon 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 platform)
(2) Be the first to understand the challenges of designing
1. 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
(5) Exploring neural architecture search (building your own Tiny ML network)

Requirement to be on campus: No

CS2021/7 Compiler and architecture support for future quantum systems

Supervisors: Dr Shuaiwen Leon Song, Alan Robertson
Eligibility: Students with the following skill sets are preferred:

- Basic understanding of quantum computing and current NISQ devices
- Good at maths
- Some familiarity with quantum physics
- Knows how to use simple IBM Qiskit programming environment

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 in us!
Requirement to be on campus: No

CS2021/8 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 initiative

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

CS2021/9 Secure distributed machine learning

Supervisors: Dr Qiang Tang & Dr Clement Canonne
Eligibility: Students from CS, Math, Physics departments are all welcome to apply. Strong math background and fast learning capabilities would be required.
Having background in machine learning or computing theory would be a big plus.
Project Description:
Distributed learning environment may involve parties that are malicious, and may influence the final result that will be biased towards their preference. We will look, in various distributed models, at how the "attackers" could corrupt the final model, and how to design robust learning algorithms to be resilient to those attacks.
Requirement to be on campus: No

CS2021/10 Backdoors in AI, attacks and defenses

Supervisors: Dr Qiang Tang & Dr Chang Xu
Eligibility: CS, Math, Physics Majors are all welcome to apply, having good math background, programming skills and fast learning capabilities are required. Having experience on deep learning or cybersecurity would be a big plus.
Project Description:
Imagine an AI model in a smart gate for secure face recognition, the manufacturer of the smart gate may leave a backdoor that for a particular people, it always opens the door, while on all other times, it is as usual. Can we understand such kind of attacks in different AI systems, and propose robust defense mechanisms?
Also, in other settings, we can design certain sample inputs to fool the AI models, or polluting training data set. Can we understand such kind of attacks in different AI systems, and propose robust defense mechanisms?
Requirement to be on campus: No

CS2021/11 Crack-resistant and portable cryptocurrency wallet

Supervisor: Dr Qiang Tang
Eligibility: CS, Math, Physics Majors are all welcome to apply, having good math and programing background and quick learning capability are required. Having basic understanding on blockchain or cryptography would be a big plus.
Project Description:
Wallet for cryptocurrency needs to securely store the secret keys for the accounts of each coins. The secret keys are long, thus cannot be remembered.
To increase usability, users may have multiple devices, each of which needs to be able to access the wallet.
We will investigate potential solutions of using password or biometrics to encrypt the "wallet", while at the same time can be resistant to offline brute-force attacks, leveraging cryptographic tools such as honey encryption.
Requirement to be on campus: No

CS2021/12 Privacy preserving and regulation complied central bank digital currency

Supervisor: Dr Qiang Tang
Eligibility: CS, Math, Physics, Finance Majors are all welcome to apply, having good math background and quick learning capability are required. Having basic understanding on cryptography would be a big plus.
Project Description:
CBDCs promise to realize a broad range of new capabilities, including direct government disbursements to citizens, frictionless consumer payment and money-transfer systems, and many more.
At the same time, CDBCs need to address an innate tension between privacy and transparency, protecting user data from abuse while selectively permitting data mining for end-user services, policymakers, and law enforcement investigations and interventions.
We would do a systematic study of the issue.
Requirement to be on campus: No

CS2021/13 Investigating privacy and scalability issues in De-Fi & NFT applications

Supervisor: Dr Qiang Tang
Eligibility: CS, Math, Physics, Business Majors are all welcome to apply, having good math background and quick learning capability. Having basic understanding on blockchain or cryptography would be a big plus.
Project Description:
Decentralized finance and non-fungible token have been the most popular decentralized applications deployed nowadays.
However, given that all of them are deployed in public blockchain platforms such ass Ethereum, there are serious privacy and scalability issues.
This project is aimed to identify concrete issues and to propose potential fixes.
Requirement to be on campus: No

CS2021/14 AI-enhanced super-resolution and loss recovery for real-time video streaming

Supervisor: Wei Bao
Eligibility: Knowledge in computer networks. Knowledge in deep neural networks.
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 nevertheless suffers from packet loss. Another issue is the limited bandwidth in the network. The sender sends low-resolution video, 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 DNNs fix distorted 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, with small data size and small inference delay.
Requirement to be on campus: No

CS2021/15 Durability properties of MongoDB

Supervisor: Prof Alan Fekete
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 document store, providing database-like functionality to store and access collections of JSON data. The data in MongoDB is usually replicated across several nodes, and the system provides alternative ways to update the data; these choices differ in

how thoroughly the change is protected against loss from system crashes. This project will extend previous work (from an Honours thesis and a summer project) to measure the properties of these choices; the focus will be on identifying conditions which make the system more vulnerable to data loss.
Requirement to be on campus: No

CS2021/16 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

CS2021/17 Extreme-scale visual analytics of big complex data

Supervisor: Prof Seokhee Hong & Amyra Meidiana
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 new visualisation techniques for 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 and many application domains.
Requirement to be on campus: No

CS2021/18 Machine learning-based approach to estimate occupancy 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.
- 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:

Occupancy estimation is important for smart building applications such as controlling heating, ventilation, lighting, air conditioning systems and determining occupants' behaviors. This project aims to detect and estimate (i.e., count) occupancy in a room from thermal imaging data using state-of-the-art machine learning algorithms. 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 context sensing devices. 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 who decide to work on this project 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 occupancy detection and estimation model.

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

CS2021/19 Swarm-based drone delivery dataset

Supervisor: Prof Athman Bouguettaya

Eligibility: Experience in python programming

Project Description:

Drones are unmanned aircrafts that operate with various degrees of autonomy. 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 aforementioned 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 set up a drone swarm using available drones (DJI Edu Tello) and collect a dataset using different formations of swarms and varying wind speeds and directions.

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

CS2021/20 Swarm-based drone delivery demonstration

Supervisor: Prof Athman Bouguettaya

Eligibility: Experience in python programming

Project Description:

Drones are unmanned aircrafts that operate with various degrees of autonomy. 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 aforementioned 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 set up a drone swarm using available drones (DJI Edu Tello) and implement a swarm-based drone delivery demonstration given the delivery environment constraints.

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

CS2021/21 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 (dependent on government's health advice)

CS2021/22 Wireless energy charging

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:
The world is rapidly going wireless. Within a span of a few decades, phones and internet became wireless, and now charging has become wireless. Wireless charging refers to the transfer of energy among devices without the need for a connecting cable. This offers the promise of increased mobility, spatial freedom and advances that could allow devices to get power many feet away from a charger. Even though wireless charging is still pretty much in its early stages, the technology is anticipated to evolve dramatically over the next few years. For example, Xiaomi recently released the development of "Mi Air Charge Technology" which will enable users to remotely charge electronic devices without any cables or wireless charging stands. This project aims to develop a wireless energy sharing platform. The platform will include a one- function app that monitors the process of wireless energy charging from one device to another.
Requirement to be on campus: Yes (dependent on government's health advice)

CS2021/23 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

CS2021/24 Efficient 360 degrees video streaming

Supervisor: Dr Kanchana Thilakarathna
Eligibility: Prospective students should have a basic understanding of network data handling and proper programming skills with python.
Project Description:
360° / VR video is getting popular among the users providing an immersive streaming experience. Due to its spherical view, 360° video demands high bandwidth in the network. Recently proposed viewport aware streaming, which divides 360 video frames into tiles and streams only selected tiles within the user Field of View (FoV), has given a great promise in reducing the amount of data transmission while increasing video quality. Handling and selection of the tiles can be challenging in 360 video streaming. To overcome this challenge, in this project, we propose a novel caching optimization mechanism for 360 video streaming with dynamic tiling shifting the content towards the edge of the network [3]. First, we explore how this dynamic tiling can satisfy viewport aware streaming considering the strong spatial relationship between the tiles. Secondly, we aim to propose efficient cache replacement policies. Finally, we validate our proposed methodology by simulations taking 360 videos in high content diversity with sufficient head movement data.
Requirement to be on campus: No

CS2021/25 Unravelling the nascent privacy risks of 3D spatial mixed reality data

Supervisor: Dr 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

CS2021/26 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

CS2021/27 Understanding the assimilation of electronic medication management system in an Australian hospital

Supervisors: Dr Kevin Kuan
Eligibility: Basic understanding in regression analysis

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

CS2021/28 Efficient algorithms for large-scale graph data processing

Supervisor: Dr Lijun Chang
Eligibility: Comfortable with graph algorithm design and implementation.
Project Description:
Graph model has been widely used to represent the relationships among entities in a wide spectrum of applications such as social networks, communication networks, and biological networks. 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 efficient techniques to speed up graph analytics on the ever-growing large graph datasets. We will consider two settings: (1) one single large graph (with billions of edges), and (2) a graph database consisting of millions of small graphs (each with tens or hundreds of edges). The typical problems we will be investigating are (1) dense subgraph (e.g., densest subgraph, clique, k-plex) computations over a large sparse graph, and (2) finding the graphs that are similar to a user-specified query graph from a graph database.
Requirement to be on campus: No

CS2021/29 DB4ML – In-Database Machine Learning

Supervisors: A/Prof Uwe Roehm
Eligibility: Strong academic background in Computer Science, especially databases, and C programming skills.
Project Description:
Machine Learning algorithms are not well supported by existing database systems as they typically iterate over the dataset until a convergence criterion is met. This is not supported, e.g., by SQL. Hence extensions with user-defined functions (UDF) are needed, which however don't parallelise very well. 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.
Some database and programming skills will be needed for installing and preparing both approaches for benchmarking.
Requirement to be on campus: No

CS2021/30 Analysing SARS-CoV-2 virus sequences with BioSeqDB

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).
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 – and extension of SQL Server to support the management and analysis of DNA sequence data. In this project, you shall load the official SARS-CoV-2 sequence dataset into BioSeqDB to evaluate the storage and performance

overhead when doing a sequence similarity search inside a BioSeqDB database 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

CS2021/31 Progressive Jupyter: Evaluation of interactive data analysis with Jupyter notebooks

Supervisors: A/Prof 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 bitof 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 to interact with a data visualisations 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 this first prototype to demonstrate the capabilities of Progressive DB 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

CS2021/32 Performance Impact of MongoDB Logical Schema Decisions

Supervisor: Dr Alan Fekete
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 document store, providing database-like functionality to store and access collections of JSON data. There are many choices one can make about how to represent the same information in JSON, for example fields can be nested, or not. This project will develop a benchmark and use it to measure the properties of these choices.
Requirement to be on campus: No

CS2021/33 Evidence synthesis of health digitalization impacts on sustainable development goals

Supervisor: Dr Simon Poon
Eligibility: Preference will be given to students with strong interests in multi-disciplinary research. Combined degree students are encouraged to apply. Background in Information Systems, Public Health Informatics, Statistics, Social Sciences and are advantageous.
Project Description:
The notion of complementarities has been used to explain why nations with similar levels of technological progress have translated to varying levels of development goals. This research aims to apply a novel logical evidence synthesis to explore the complex causality of recent health digitalization progress on achieving the sustainable development goals (SDGs).

The intent of using the approaches in evidence synthesis (like systematic reviews, meta-analysis and set-theoretic approach) is to uncover that factors in complex configurations may play different roles as core and periphery factors in achieving SDGs. Certain important factors being overlooked in traditional empirical studies, but they could play a non-trivial role in achieving the development targets.

The set-theoretic based framework may serve as both an analytical tool for understanding causality from complex interdependencies among provisions of digitalization of health, as well as for facilitating the abstraction of those complexities through means of pinpointing core and periphery within the bundle of seven indicator-categories captured in the Global Digital Health Index.
Requirement to be on campus: No

CS2021/34 Benchmarking a Partitioning Advisor for MongoDB

Supervisor: A/Prof Uwe Roehm
Eligibility: Strong academic background in Computer Science, especially databases and ideally also machine learning (the latter can be picked up during the project too). This project also requires good Python programming skills and some knowledge of SQL. Experience in benchmarking and tuning of systems would be desirable.
Project Description:
MongoDB is a popular NoSQL database system which allows to partition data distributed over multiple nodes for fast performance and scalability. An important design decision is how data gets partitioned and on what attributes. While this is currently left to the users, we are working on automating it using deep reinforcement learning (DRL). In this project, we want to adapt an existing partitioning advisor, which our research group has built for different distributed database, 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 DRL 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.
Requirement to be on campus: No

CS2021/35 ShoeSense: Non-optical methods for tracking in-shoe foot motion

Supervisors: Dr Anusha Withana, Dr Alycia Fong Yan & Prof Alistair McEwan
Eligibility: You will work with three supervisors (computer science/health and medicine/bio med), 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:
Understanding the effect of footwear on foot motion has many health benefits for all populations. It is also important to inform footwear designers and manufacturers. The current gold standard optical method of motion capture uses retro-reflective or active markers captured with multiple infrared cameras. This requires cutting holes in the shoe to place the markers and cannot be used outside lab settings. In this project we plan to develop a novel, magnet based wearable sensor to track foot movements inside the shoe without any modification to the shoe. Currently we have developed initial prototypes using sock worn magnets. The goal of this project is to improve the sensor and collect data and analyse them.
Requirement to be on campus: Yes (dependent on government's health advice)

CS2021/36 Neural tactile interfaces for visually impaired as enabling technology

Supervisors: Dr Anusha Withana
Eligibility: You will work with the supervisor and 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:
Modern computers frequently use visual and auditory interfaces. For example, we can see the visual information on our screens and hear the audio computers generate. However, one important modality is less explored, that is, what we feel through our skin. We feel vibrations, temperature, and pressure through sensory receptors in our skin. This modality is particularly important to create enabling interfaces, for instance computer interfaces used by visually impaired people. In this project, we will explore how we can create enabling technologies

using novel tactile interfaces, particularly an interface directly communicates with our neural system.
Requirement to be on campus: Yes (dependent on government's health advice)

CS2021/37 Robust Geometric Map Matching with Sequences of Shortest Paths

Supervisors: Prof Joachim Gudmundsson, Martin Seybold
Eligibility: Interest in algorithms and programming
Project Description:
For a given sequence of location measurements, the goal of the geometric map matching problem is to compute a sequence of movements along edges of a spatially embedded graph which provides a 'good explanation' for the measurements. The problem gets challenging as real-world data, like GPS traces or graphs from the OpenStreetMap project, does not exhibit homogeneous data quality. Graph details and errors vary in areas and each trace has changing noise and precisions. Hence formalizing what a 'good explanation' is, becomes quite difficult.
In [1], shortest paths have been used to improve robustness of geometric map matching objectives based on generic trajectory similarity measures, without the need of incorporating additional parameters in the objective formalization. Independently [2] explores the same direction for the Fréchet distance by introducing additional model parameters.

The goal of this project is to improve the practical performance of geometric, shortest path map matching algorithms by exploiting simple properties of trajectory distance measures.

Reading:
[1] Martin P. Seybold. Robust Map Matching for Heterogeneous Data via Dominance Decompositions. SDM 2017: 813-821
https://doi.org/10.1137/1.9781611974973.91
[2] Erin Chambers, Brittany Terese Fasy, Yusu Wang, and Carola Wenk. 2020. Map-Matching Using Shortest Paths. ACM Trans. Spatial Algorithms Syst. 6, 1, Article 6 (February 2020), 17 pages.
https://doi.org/10.1145/3368617
Requirement to be on campus: No

CS2021/38 Computing Multiplicative Weighted Voronoi Diagrams

Supervisors: Prof Joachim Gudmundsson, Martin Seybold
Eligibility: Interest in algorithms and programming
Project Description:
Multiplicative weighted Voronoi Diagrams (MWVD) naturally appear in models of real world applications – e.g. radio coverage of towers with different transmit powers.

The goal of this project is to provide algorithms to compute MWVDs that are numerically robust and applicable for large instances. Moreover, the project seeks to analyze the practicability and scaleability of the implementation by means of experiments.

Reading:
[1] F. Aurenhammer and H. Edelsbrunner. An optimal algorithm for constructing the weighted voronoi diagram in the plane. Pattern Recognition, Volume 17, Issue 2, 1984, Pages 251-257.
[2] P. Hachenberger and L. Kettner. Boolean Operations on 3D Selective Nef Complexes: Optimized Implementation and Experiments. In: Proc. of
2005 ACM Symposium on Solid and Physical Modeling (SPM), 2005
Requirement to be on campus: No