THE UNIVERSITY OF SYDNEY

FACULTY OF ENGINEERING

Engineering Vacation Research Internship Program

COMPUTER SCIENCE RESEARCH PROJECTS – SUMMER 2023-24

THE UNIVERSITY OF SYDNEY

FACULTY OF ENGINEERING

COMPUTER SCIENCE PROJECTS

CS2023-24/1 Faithful Visual Analytics of Big Complex Data

Supervisors: **Prof. Seokhee Hong**

Eligibility: Skills Required: Data Structure and Algorithms and Programming (Java, C++, Python, Javascript)

Project Description:
Technological advances have increased data volumes in the last few years, and now we are experiencing a "data deluge" in which data is produced much faster than it can be understood by humans.
These big complex data sets have grown in importance due to factors such as international terrorism, the success of genomics, increasingly complex software systems, and widespread fraud on stock markets.
Visualisation is a powerful tool to compute good geometric representation of abstract data to support analysts to find insights and patterns in big complex data sets.
This project aims to design, implement, and evaluate new visualisation algorithms for faithful visualisation of big complex data, to enable humans to find ground truth structure in big complex data sets, such as social networks and biological networks.
These new visualisation methods are in high demand by industry for the next generation visual analytic tools.

Requirement to be on campus: Yes * *dependent on government's health advice.*

CS2023-24/2 Identifying Buggy Patterns

Supervisor: Dr. Rahul Gopinath

Eligibility: You will work with the supervisor directly for this project. You should be a fast learner. Excellent skills in programming (Basic Python & Java knowledge is necessary), problem solving, as well as the ability to work independently are required.

Project Description:
Bugs are often induced by similar inputs. If we can identify and isolate such inputs, we can prevent bug inducing code from being exercised, and hence prevent vulnerabilities. The requirement for getting this to work is a large set of test cases. Fortunately, we know how generate a large number of test cases through fuzzing. This project will explore how to learn regular expressions from such bug inducing tests such that any input that matches such regular expressions is guaranteed to trigger the bug. Such regular expressions can be then used to combine different bugs to produce more novel buggy behavior.
This project, if completed successfully, may be extended for a paper in one of the A/A* conferences in software engineering.

Requirement to be on campus: No

CS2023-24/3 Enable Design By Contract in JDK 19

Supervisor: Dr. Rahul Gopinath

Eligibility Criteria: The project requires a quick learner, Furthermore, the project is to update a Java program. Hence, this calls for an expert Java programmer.

Project Description:
Design by contract is a technique which allows one to specify the pre and post conditions as well as invariants for classes and methods. This was popularized by Bertrand Meyer for Eiffel. While this enables programmers to ensure that their code is bug free, the programming language support for design by contract has been lacking. Fortunately, for Java, there has been a limited support with a previous project called cofoja from google. The library supports adding annotations to support pre,post conditions and invariants to Java source files. Unfortunately, the project is no longer compatible with new Java versions and libraries. This project asks to (1) update Cofoja so that it is up to date with new JDK and dependencies (2) Compare the contracts supported in Cofoja with those that are supported by Eiffel (3) Extend Cofoja with some of the unsupported methods (4) Verify the utility of Cofoja by converting simple opensource Java project to use Cofoja.
This project, when finished can spark the attention of industry giants such as Google, Oracle, Parasoft etc. with a significant Java codebase. The project can also serve as a foundation for future research in design by contract, which is a well-regarded software design technique.

Requirement to be on campus: No

CS2023-24/4 Extracting Key Value Access Patterns from Applications

Supervisor: Dr. Rahul Gopinath

Eligibility:

Project Description:
This short project is within a larger funded activity that aims to improve the performance of key-value stores, which provide a simple get/put interface and are used in many applications to store data. It is important to know about the workload pattern, for example whether there are more reads (gets) compared to writes (puts) or whether some keys are accessed very frequently, because different implementations of a key-value store are beneficial depending on the kind of workload.
In this project we will seek to show how one can capture the workload pattern. To begin, we will dynamically instrument an application by wrapping external interfaces to data stores and tracking the workload. We will measure the overhead of doing so. If time allows, we may explore ways to detect changes in the access pattern over time or compare to other ways of capturing workload pattern.

Requirement to be on campus: No

CS2023-24/5 Applying Operator Precedence to Parsed Output

Supervisor: Dr. Rahul Gopinath

Eligibility: This project requires a fast learner, who is comfortable with doing literature search, understanding the algorithms, along with the implementation.

Project Description:
human beings. A general parser takes a specification (typically provided as a grammar) the input string and generates a derivation tree. A key difficulty when using parsers is that in many languages, specific operators have specific precedence. For example, multiplication (*) typically has higher precedence than addition (+). A typical way to account for this is to try to encode this precedence in the grammar itself. However, this is error-prone, and has led to a bad reputation for general context-free parsers. The alternative is to extract the first parse tree and transform the tree to the derivation tree with the correct precedence bracketing. LaLonde et al showed how to do this for simple languages (LaLonde 1981).
This project will involve (1) doing the background literature review (2) understanding and implementing LaLonde's algorithm (3) applying this algorithm on common programming

language patterns that typically are solved with handwritten precedence parsers and evaluating the effectiveness.

Requirement to be on campus: No

CS2023-24/6 Analysing the Impact of Interferences on Drone Delivery Efficiency using Advanced Machine Learning Techniques

Supervisor: Prof. Athman Bouguettaya

Eligibility: Experience in python programming, Ansys Simulation software

Project Description:
A continuous expansion of urban areas is leading to an increased demand for instant package deliveries from warehouses to customers' doorstep. Unmanned Aerial Vehicles (UAVs) or drones have the potential to serve customers with cost effective, green, anywhere anytime and fast deliveries. Drones usually operate in skyway networks consisting of an interconnected set of nodes. The nodes are typically building rooftops which serve as recharging stations or delivery destination or stops for drones. Drones may recharge at nodes as they are constrained by limited battery capacity. These nodes are connected through skyway segments. Multiple drones may share the same skyway segment to transit from one node to other. As drones operate in in these skyway segments simultaneously, they may interfere with each other because of their aerodynamic forces. The aim of the project is to conduct an extensive set of experiments in the drone lab with real drones (DJI Edu Tello) to analyse the impact of interfering aerodynamic forces on the delivery efficiency using a set of advanced machine learning techniques.

Requirement to be on campus: Yes * *dependent on government's health advice.*

CS2023-24/7 Detecting Fake Content on social media

Supervisor: Professor Athman Bouguettaya

Eligibility: Programming in Python, writing scripts to pre-process data, Typical Machine learning skills, Writing and communication skills.

Project Description:
social media has become an essential part of our lives. Social media users upload a lot of content on different social media platforms. The content shared on social media may contain untrustworthy images. Existing approaches relies on image processing and object-oriented techniques to detect fake images on social media. These solutions are costly and computationally intensive. Some recent solutions investigate comments on a post to assess credibility of an image. These solutions may not accurately determine the credibility of an image because the fake posts on social media can still get supportive comments. We leverage the metadata of an image and the related posted information to determine the trustworthiness of a crowdsourced image. This information may reflect some changes in the image. For instance, there may be some discrepancies in the metadata which might be an indication of modifications in the image. Therefore, investigating this meta-information may provide useful insights about the changes introduced in the image. This project will focus on detecting and analyzing subtle tampering within the image meta-information to inform a decision whether an image is fake. We will use a set of text-based machine learning techniques to aid in detecting these changes.

Requirement to be on campus: Yes * *dependent on government's health advice.*

CS2023-24/8 Trust Management Framework for Crowdsourced IoT Services

Supervisor: Professor Athman Bouguettaya

Eligibility: Good programming background in either Java or Python, and good knowledge on Algorithms.

Project Description:
Rapid advancements in IoT-related technologies have enabled many applications, including smart homes and industries. However, IoT devices vary in processing, storage, communication, and energy capabilities. Therefore, some IoT devices may benefit from additional resources shared by other devices. For example, a smartphone may share computational resources as-a-service with a nearby smartwatch. This concept of service crowdsourcing depends on trust between IoT devices. Trust assessment helps to identify malicious IoT devices in the ecosystem. Trust-related data should be stored to assess the trustworthiness of IoT devices. However, the dynamic nature of the IoT ecosystem limits the possibility of having a centralized authority to store and manage trust-related data. This project focuses on the development of a distributed data storage framework (using technologies like blockchain) where the integrity of the trust data can be protected.

Requirement to be on campus: No

CS2023-24/9 Wireless power transfer among IoT devices

Supervisor: Professor Athman Bouguettaya

Eligibility:

a. Knowledge of electromagnetism and circuit design.
b. Experience in popular scripting languages such as Python or R.
c. Data visualisation skills (e.g., matplotlib) to visualise of the collected energy data.

Project Description:
Wireless charging has recently been proposed as a safe, convenient, ubiquitous (anywhere, anytime), efficient, green alternative to charging smartphones and other IoT devices from nearby power sources, e.g., a charging pad. This is part of a smart campus project where IoT devices in confined areas (e.g., coffee shops, restaurants, classrooms, offices, etc) can freely share safely and efficiently crowdshared energy from other devices. Wireless charging relies on the wireless transfer of energy from a power source to the smartphone. A recent system, named MagMiMo, was proposed as a solution to transfer power by beamforming the non-radiated magnetic field and steering it toward a phone [2]. MagMIMO system can charge a phone remotely independently of its orientation. For instance, MagMIMO may charge a smartphone without being removed from the user's pocket. This project aims to develop and extend MagMiMo to demonstrate the wireless transfer of low energy from a smartphone to another over short distances to end the so-called tyranny of power points. The project will use small electric coils to allow the transfer of low-powered energy over short distances among IoT devices, typically in the tens of centimetres. This project will focus on the implementation of a prototype that will enable energy to transfer between smartphones over short distances. It will also focus on the analysis, investigation, and summarization of the characteristics of real energy transfer data collected from the built prototype.

Requirement to be on campus: Yes **dependent on government's health advice.*

CS2023-24/10 Visual Prompting for Generative AI

Supervisor: Daochang Liu

Eligibility:

a. Be familiar with deep learning.
b. Have good coding skills using PyTorch.
c. Be interested in research on generative AI.
d. Preferably plan to pursue a higher degree by research in future.

Project Description:
Recently we have witnessed a surge of generative artificial intelligence advancements. Text prompting techniques, such as chain-of-thoughts and self-consistency prompting etc., are essential to fully unleash the power of large language models. In this project, we will explore a new form of prompting called "visual prompting" to enhance or adapt generative visual models such as Stable Diffusion.
This project aims to cover a full research cycle of literature review, algorithm development, experimental validation, and manuscript preparation. The expected outcome will potentially lead to a paper submission to top-tier conferences in machine learning or computer vision.

Requirement to be on campus: Yes **dependent on government's health advice.*

CS2023-24/11Co-word analysis of literature on emotion tracking

Supervisor: Dr. Zhanna Sarsenbayeva

Eligibility: You will work with the supervisor directly for this project. You should be a fast learner. Excellent skills in programming (Basic Python & R knowledge is necessary), knowledge of stats, problem solving, as well as the ability to work independently are required.

Project Description:
Co-word analysis is widely used to analyse textual content. It focuses on understanding the relationship between terms in a text and is used for mapping patterns and trends of associated words. Given these features, co-word analysis offers a powerful bibliometric approach to map the evolution and assess the structure of scientific disciplines using publication data, including metadata, titles, abstracts, and keywords. Ideally, keywords are used to describe the content of a research article. Thus, co-word analyses of keywords can reveal the conceptual structure and evolution of the research topics within the area or field in focus, based on the interaction of the respective keyword. To define a conceptual structure and the characteristics of a research area or field, we use keyword networks and clusters of keywords. In this work you are required to perform co-word analysis of literature on emotion tracking. You will learn how to program in R and apply statistical analysis and hierarchical clustering in the real-world scenario.
This project, if completed successfully, may be extended for a paper in one of the top venues of the HCI research field.

Requirement to be on campus: No

CS2023-24/12 Media Content Generation and Synthesis

Supervisor: A/Prof Zhiyong Wang

Eligibility:

a. Strong Programming Skills (Python) and Math Skills
b. WAM > 80

Project Description:
Recent success of deep learning has demonstrated a great potential to generate and synthesise media content. This has opened a new door for creativity and innovation in many domains, such as media, film, and game, even metaverse. This project aims to address the technical challenges of creating highly realistic media content by developing novel computing techniques, such as audio/image/video generation and editing, motion retargeting, 3D animation, cross-modal simulation, and 3D physical simulation. Students will gain comprehensive knowledge in multimedia data processing, computer vision, 3D vision, computer graphics, and machine learning.

Requirement to be on campus: Yes **dependent on government's health advice.*

CS2023-24/13 Multimedia Forensics

Supervisor: A/Prof Zhiyong Wang

Eligibility:

a. Strong Programming Skills (Python) and Math Skills
b. WAM > 80

Project Description:
Multimedia data has been widely used to store information in almost every domain, from photos shared on social media platforms and transaction receipts to electronic health records. Meanwhile, advances in digital media processing have produced a large variety of intelligent tools for manipulating media content, such as enhancing visual quality and removing or adding an object from an image. However, the processed media content could be used to convey false, misleading, or hidden information, which has increasingly challenged the saying "seeing is believing". This project aims to develop advanced multimedia computing and machine learning techniques to identify the forensic and security trails and improve security of multimedia data.

Requirement to be on campus: Yes **dependent on government's health advice.*

CS2023-24/14 Generative Adversarial Networks (GANs) for Predicting Alzheimer's Disease (AD) Progression

Supervisor: A/Prof. Xiuying Wang

Eligibility:

a. The project is open to students with a strong interest in medical imaging and deep learning.
b. Students are expected to have a profound understanding of machine learning concepts and strong programming skills in Python.
c. Background knowledge in imaging processing, generative neural networks, and PyTorch framework will be advantageous.

Project Description:
Deep-learning networks are gaining overwhelming attentions and interests in the medical imaging domain. Multi-modality images are now indispensable to accurate diagnosis of various diseases. Alzheimer's disease (AD), as a progressive dementia-related illness, negatively impacts memory and cognitive functions of the elderly patients. While Positron Emission Tomography (PET) images have demonstrated promise in identifying cognitive impairments, high PET scan costs limit availability and access to PET datasets for AD progression.
This project aims to predict AD progression by exploring the feasibility of Generative Adversarial Networks to generate PET images based on corresponding Magnetic Resonance Imaging (MRI) images. Students will investigate the variations of GANs and optimize the GAN-based PET image generation process. Additionally, validation of the

usability of these generated PET images for data augmentation in predicting AD progression will be necessary.
This project's success will yield publishable results and its extendable nature makes it suitable for pursuing higher-level research degrees.

Requirement to be on campus: Yes * *dependent on government's health advice.*

CS2023-24/15 Self-Supervised Learning (SSL) for Multimodal Medical Images

Supervisor: A/Prof. Xiuying Wang

Eligibility:

a. The project is open to self-motivated students with strong analytic thinking skills and passionate in deep learning.
b. Students are expected to have proficient Python programming skills and basic understanding of deep learning/machine learning.
c. Background knowledge in imaging processing and PyTorch framework will be advantageous.

Project Description:
Training of the supervised deep learning models for medical applications requires a large amount of annotated data that is costly yet subjective and operator-dependant. Alternatively, self-supervised learning (SSL) models learn general representations from unlabelled data, which are then fine-tuned with limited labels for different downstream tasks to achieve comparable performance as the supervised counterparts. While multi-modal images provide comprehensive information for disease classification and diagnosis, however, how to leverage multi-modal images for SSL remains an open question.
This project aims to develop an effective multi-modal SSL network for different downstream tasks. The student will explore the cutting edge SOTA models and will investigate the current unsupervised clustering methods for SSL. Further, contrastive learning of both heterogeneous and homogeneous representations from multi-modal images will be incorporated to enhance SSL models.
The success of this project will lead to publishable results. The project can be extended for HDR research.

Requirement to be on campus: Yes * *dependent on government's health advice.*

CS2023/16 Neural Interfaces for Visually Impaired as Tactile Enabling Technology

Supervisor: 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.*

CS2023-24/17 Quantify tumour heterogeneity in cancers through multi-parametric MR imaging

Supervisor: Prof. Jinman Kim

Eligibility: WAM>75 and Undergraduate candidates must have already completed at least 96 credit points towards their undergraduate degree at the time of application.

Project Description:
This project aims to quantify tumour heterogeneity in cancers through multi-parametric MR imaging (MRI). Tumour heterogeneity, the variation in cellular characteristics within a tumour, affects treatment response and patient outcomes. Existing research, however, focuses on quantifying imaging representation of the entire tumour. Multi-parametric MRI offers a non-invasive approach to assess various tumour properties simultaneously, providing valuable and complementary insights. By integrating multiple MRI parameters, this project aims to generate a comprehensive tumour heterogeneity characterization, aiding personalized treatment planning. The project will also determine tumour heterogeneity's clinical relevance and its association with tumour prognosis. By utilizing publicly available mpMRI images of the brain, prostate and breast cancer, this project contributes to advancing cancer care, understanding tumour biology, and the field of radiogenomics.
The project student will collaborate closely with other research members and have the opportunity to work with clinicians.

Requirement to be on campus: Yes * *dependent on government's health advice.*

CS2023-24/18 Large-Scale Social Network Analysis

Supervisor: A/Prof. Lijun Chang

Eligibility: Good algorithm design and C (or C+++) programming skills

Project Description:
We are nowadays facing a tremendous amount of large-scale social networks with millions or billions of edges. Thus, there is a need of designing efficient algorithms for processing large graphs. In this project, our aim is to design efficient algorithms to speed up graph processing on ever-growing large graph datasets. The problems that we will be investigating can be (1) dense subgraph (e.g., clique, near-clique) computation over a large sparse graph which finds one dense subgraph of the maximum size, or (2) dense subgraph enumeration which enumerates all maximal dense subgraphs.

Requirement to be on campus: No

CS2023-24/19 Data-driven analysis of sleep disorders using advanced machine learning

Supervisor: A/Prof Irena Koprinska

Eligibility:

a. Machine learning skills - completed COMP3308/COMP3608 or COMP5318 with D/HD
b. Excellent programming skills

Project Description:
Insomnia and sleep apnea are common sleep disorders. Insomnia is defined by difficulties falling asleep and staying asleep. Sleep apnea is characterized with periods of reduced breathing or no breathing at all during sleep. Both disorders cause daytime sleepiness and fatigue, and may lead to depression, heart disease, diabetes, and other adverse health effects. The goal of this project is to use machine learning techniques to analyse sleep disorders, e.g. to objectively detect insomnia and normal sleepers based on EEG

trajectories and predict sleep apnea events in advance based on respiratory data to allow for medical devices to intervene. The project will use large datasets containing data from multiple signals recorded overnight.

Requirement to be on campus: No

CS2023-24/20 Predicting student progress in online programming courses

Supervisors: A/Prof Irena Koprinska and Dr Bryn Jeffries

Eligibility:

a. Machine learning skills - completed COMP3308/COMP3608 or COMP5318 with D/HD
b. Excellent programming skills

Project Description:
Programming skills are in high demand but mastering them is difficult especially for students without prior programming background. On the other hand, online programming courses generate a vast amount of data and there is an opportunity for applying machine learning techniques to analyse this data, to help teachers understand student progress and take remedial actions. This project aims to develop interpretable machine methods to track student progress in programming courses, predict student grades and drop-out and identify intervention points.

Requirement to be on campus: No

CS2023-24/21 Evidence synthesis of Health Digitalization Progress on Sustainable Development Goals

Supervisor: Associate Professor 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

CS2023-24/22 Causal Complexity: Augmentation of set-theoretics for machine learning and Statistical Analysis

Supervisor: Associate Professor Simon Poon

Eligibility: Good knowledge in data science and statistical techniques. Students with good programming skills and are interested in multi-disciplinary studies (in conjunction social science) are encouraged to apply.

Project Description:
Assessing effects from complex interactions of multiple study factors are common in empirical studies. Changing one factor may have little effect on study outcome if other factors remain unchanged. Furthermore, such interactions may extend to different configurations with complex synergistic relationships amongst many seemingly unrelated factors. From the view of statistical association analysis, especially correlation analysis with the well-known correlation coefficients has been a useful technique for identifying relationships from observational data. However, conventional statistical methods cannot account for situations in which only specific combinations of variables reveal their impact on the outcome (conjunctural causation) or all paths that lead to an outcome need to be simultaneously uncovered (equifinality). These methods also fall short in explaining situations in which a given combination of variables contributes to the presence of an outcome but at the same time is irrelevant for the absence of that outcome (causal asymmetry). In this project, we address these issues by integrating set-theoretic approaches in conjunction with statistical approaches (like regressions) for discovering meaning configurations from observational data with limited diversity.
Reference: A Configurational Analysis of Risk Patterns for Predicting the Outcome After Traumatic Brain Injury **https://pubmed.ncbi.nlm.nih.gov/29854144/**

Requirement to be on campus: No

CS2023-24/23 Distributed Learning with Minimal Communication

Supervisors: Dr Kanchana Thilakarathna and Prof Teng Joon Lim

Eligibility: Knowledge and experience on machine learning is essential. Prior experience with federated learning is desirable.

Project Description:
Edge Artificial Intelligence (AI) is concerned with the algorithms and computational methods used for the accomplishment of AI tasks (such as classification) by leveraging devices at the edge of the network, each with their own computing capabilities, training data and communication resources. In the context of the proposed project, these devices are small UAVs with limited energy and computing resources and the sensors might be cameras or other devices that can be mounted on a UAV. The primary objective of the VRI project is to minimize the communication overhead needed in sharing information between UAVs while solving the distributed AI problem. The initial focus will be on the family of federated learning (FL) algorithms, which rely on local machine learning model training and global aggregation of the local models without explicit sharing of data between clients/agents.

Requirement to be on campus: No

CS2023-24/24 An Automatic Inequality Prover

Supervisors: Dr. Clement Canonne

Eligibility: WAM>75 and Undergraduate candidates must have already completed at least 96 credit points towards their undergraduate degree at the time of application.

Project Description:
Valiant and Valiant developed in 2014 an algorithm to automatically prove or disprove mathematical inequalities of a certain (quite general) form [1]. They proved the correctness of their algorithm and provided a basic Matlab implementation. (Since ported in Python during a Capstone project at U Syd).

The goal of this project is to extend the Valiant-Valiant algorithm to make it easier to use, more flexible, and make it available to the wider mathematical community (ideally by developing a Python, Julia, or R package). Proficiency in Python, Julia or R is required, as well as a solid background in algorithms.

[1]https://theory.stanford.edu/~valiant/papers/instanceOptFull.pdf

Requirement to be on campus: Yes * *dependent on government's health advice.*

CS2023-24/25 User-Centered Development of a Data Visualization Dashboard for a Neonatal Intensive Care Unit

Supervisors: Dr Anusha Withana, Dr Zhanna Sarsenbayeva

Eligibility: Knowledge of user cantered research, thematic analysis is a bonus.

Project Description:
In this thesis project, you will follow a user-centred design approach and data analysis (thematic analysis) to develop an interactive dashboard to be used in Neonatal Intensive Care Units (NICUs) for data visualisation. The project was founded on the requirements presented in a previous thesis; however, the previous thesis did not analyse the data. In this thesis project, you will use thematic analysis and coding of data. Then, you will develop a design framework needs to be implemented.

Requirement to be on campus: Yes * *dependent on government's health advice.*

CS2023-24/26 Benchmarking a DRL-based Query 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. 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 indexed so that only specific partitions need to be contacted for a given query. In previous work, we have shown that MongoDB's own query optimiser sometimes selects a sub-optimal query execution plan by wrongly estimating the benefit of indexes.

Together with MongoDB, we are working on an external query advisor that is using deep reinforcement learning (DRL) to direct MongoDB towards the optimal plan by monitoring past query executions and then adding suitable execution 'hints' to new queries. In this project, the task is to take our existing query advisor and benchmark its performance against the latest version of 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 – the project is a collaboration with MongoDB, and its outcome is planned for being used in a publication about our approach.

Requirement to be on campus: No

CS2023-24/27 DB4ML – In-Database Machine Learning

Supervisor: A/Prof Uwe Roehm

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

Project Description:
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

CS2023-24/28 Logic and Computation

Supervisor: Dr Sasha Rubin

Eligibility: Very strong mathematical ability. Typically, you would have achieved HD in the advanced streams of at least one of the following courses (or similar courses): Discrete Mathematics, Models of Computation, Algorithms & Data Structures.

Project Description:
Formal logic allows humans to precisely express knowledge and facts so that computational systems can use them and reason about the world.

There are a few possible topics, depending on candidate interest and ability:
1. Explainable AI: whose goal is to produce models that enable human users to understand their decisions. This project would involve extending the state-of-the-art in logic-based explainability for ML classifiers or logic-programs and elaborating on the deficiencies of popular heuristic approaches. This will extend **https://dblp.org/rec/conf/aaai/GorjiR22**

2. Graph query languages: GQL (Graph Query Language) is being developed as a new ISO standard for graph query languages to play the same role for graph databases as SQL plays for relational databases. This project will explore using automata theoretic techniques to solve queries. This will draw on https://drops.dagstuhl.de/opus/volltexte/2023/17743/pdf/LIPIcs-ICDT-2023-1.pdf

3. Rational-by-design AI: A grand challenge in AI is to provide algorithms for automatically producing agents that achieve their goal in uncertain environments (think of poker bots). This project will involve designing and implementing algorithms for producing agents that exploit opportunities when they arise (e.g., if an opposing agent in the environment makes a mistake). This will draw on https://dblp.org/rec/conf/ijcai/AminofGRZ22

Requirement to be on campus: Yes **dependent on government's health advice.*