Computer Science and Technology
Research Students' Symposium

19-21 May 2021

Programme

Thursday 20 May

9:00AM–9:15AM · Stage
Dr Amanda Prorok
Keynote talk

9:15AM–9:30AM · Stage
Dr Jamie Vicary
Keynote talk

9:30AM–9:45AM · Stage
Andi Zhang
Uncertainty and Out-of-Distribution (OOD) detection

9:45AM–10:00AM · Stage
Andrei Paleyes
Data Oriented Architectures for Deploying Machine Learning

10:00AM–10:15AM · Stage
Junwei Yang
Multi-Contrast MRI Synthesis and Diagnosis

10:15AM–10:30AM · Stage
Chelsea Edmonds
Formal Proof Techniques for Combinatorial Structures in Isabelle/HOL

10:30AM–10:45AM · Stage
Dimitrije Erdeljan

10:45AM–11:00AM · Stage
Break

11:00AM–11:15AM · Stage
Indigo Orton
Concurrency optimisation via dynamic analysis

11:15AM–11:30AM · Stage
Faustyna Krawiec
Efficient Automatic Differentiation in Functional Programming Languages

11:30AM–11:45AM · Stage
Alexander Rice
Semistrictness for higher categories

11:45AM–12:00PM · Stage
Angus Hammond

12:00PM–12:15PM · Stage
Derek Sorensen

12:15PM–12:30PM · Stage
Daniel Hugenroth
Strong Metadata-Privacy for Mobile Devices and Group Communication

12:30PM–1:15PM · Stage
Lunch

1:15PM–2:15PM · Sessions
- Computer Architectures theme meetings
- Graphics, Vision and Imaging theme meeting
- Mobile Systems, Robotics and Automation theme meeting
- Natural Language Processing theme meeting

2:15PM–2:30PM · Stage
Christine De Kock
Towards understanding constructive disagreements online

2:30PM–2:45PM · Stage
Georgi Karadzhov
DEliBots - Deliberation Enhancing Bots

2:45PM–3:00PM · Stage
Sian Gooding
Textual Complexity is in the Eye of the Beholder

3:00PM–3:15PM · Stage
Wen Hoi Gladys Tyen
Chatbots for language learning

3:15PM–3:30PM · Stage
Tea break

3:30PM–3:45PM · Stage
Ran Zmigrod
Decoding and Inference in Graph-based Dependency Parsing

3:45PM–4:00PM · Stage
Ramon Viñas Torné
Learning the generating mechanisms of multi-omics data

4:00PM–4:15PM · Stage
Dimitrios Los
Balanced Allocations with Incomplete Information: The Power of Two Queries

4:15PM–4:30PM · Stage
Felipe Ferreira Santos
Lower Bounds on Space Constrained Computational Models

4:30PM–4:45PM · Stage
Erika Bondareva
Segmentation-free Heart Murmur Detection Using Deep Learning

4:45PM–5:00PM · Stage
Filip Svoboda
Learned Communication in FL

5:00PM–5:10PM · Stage
End day 2
See you tomorrow

Andi Zhang *supervised by Dr D. Wischik*

Uncertainty and Out-of-Distribution (OOD) detection

Modern machine learning methods, especially deep learning, are achieving great success in predictive accuracy these years. However, for some high-stakes applications, such as self-driving car and medical diagnoses from imaging, the need for robustness and security is harsher. As deep learning uncertainty is a big topic, this talk will concentrate on Out-of-Distribution (OOD) detection in supervised and unsupervised settings.

Andrei Paleyes *supervised by Prof N. D. Lawrence*

Data Oriented Architectures for Deploying Machine Learning

Deploying machine learning in real life is hard. Many businesses and organizations still struggle to apply it in their domains. Recent reports claim that 87% of machine learning projects never make it to production, which means massive waste of work, time and money. Our research group believes the reason for that might lie in the way we build software. Traditionally software services are optimized for reduced latency and big throughput, and not for things like data collection, which are crucial for successful application of machine learning. To solve this problem we aim to create a novel approach to software system design, which we call Data Oriented Architectures. It prioritizes data access, reproducibility, fast experimentation, while still maintaining good performance according to common software metrics. We build upon existing paradigms, such as data streaming, Actor model and dataflow, to define an approach to build software that makes machine learning deployment simpler and more accessible.

Junwei Yang *supervised by Prof. P. Liò*

Multi-Contrast MRI Synthesis and Diagnosis

Multimodal magnetic resonance imaging (MRI) has been commonly adopted as a scanning technique to provide medical images for clinical diagnosis. With the

generated high-fidelity MRI images from a scanner, the doctors can study the human body noninvasively. In addition, because of the diversity of contrasts that MRI can acquire in soft tissues, images of the same anatomy with different contrasts can produce more diagnostic information for clinical or research studies. However, the acquisition is usually time-consuming, especially for multimodal imaging due to the separate scanning of each modality, which could seriously affect the efficiency of clinical diagnosis, and may also lead to motion artifacts in MRI images and uncomfortable experiences for patients. Therefore, the ability to accelerate the acquisition of MRI images and synthesize missing or corrupted contrasts have the potential to improve the overall image quality, and diagnostic tasks such as segmentation. To this end, my studies focus on development of such techniques through better exploitation of information across different modalities, and how such methods can assist other diagnostic tasks.

Chelsea Edmonds *supervised by Prof. L. C. Paulson*

Formal Proof Techniques for Combinatorial Structures in Isabelle/HOL

The formalisation of mathematics is an area of increasing interest, enabling us to verify correctness, gain deeper insights into proofs, and introduce automation to research level mathematics. However, the field of combinatorics remains underrepresented in formal environments, despite its many real-world applications and the interesting challenges it presents to formalisation. This presentation will present an overview of the progress made to overcome some of these challenges by developing general formalisation techniques for combinatorics in the proof assistant Isabelle/HOL. In particular, I'll focus on recent work completed to develop modular techniques using Isabelle's locale system to formalise combinatorial structures. I'll present the first formalisation of combinatorial designs, which demonstrated the flexibility and extendibility of a locale approach for formalising complex structural hierarchies and combinatorial properties. The presentation will conclude with an overview of several other identified formal combinatorial proof techniques and their challenges which the remainder of this PhD project aims to explore.

Dimitrije Erdeljan *supervised by Dr M. G. Kuhn*

Hidden camera detection based on compromising emanations

Compromising electromagnetic emanations have mainly been studied as a source of information leakage in computer displays and side channels in cryptographic devices. We show that the emanations of small "hidden cameras" -- in particular, those caused by storage bus transactions -- can also be used to identify the presence of a camera, using an approach based on comparing the emanations caused by storing frames recorded in different conditions (e.g. with the light on and off).

Indigo Orton *supervised by Prof. A. Mycroft*

Concurrency optimisation via dynamic analysis

Dynamic analysis can identify improvements to programs that cannot feasibly be identified by static analysis; concurrency improvements are a motivating example. Fundamentally, concurrency and performance behaviour is only exhibited at runtime. Moreover, the complex interactions of concurrently executing code, both functional interactions (shared memory) and system interactions (scheduling of threads), mean that isolated analysis of individual components is insufficient. Instead, the whole program execution must be considered at once to identify real improvements. Excitingly, taking this approach allows us to accurately estimate the performance benefits of individual improvements and combinations of improvements, meaning all suggested changes have clear value. In this talk I will cover some of these challenges and my approaches to addressing them, from execution tracing to estimating potential improvements and generating source code patches to implement these improvements.

Faustyna Krawiec *supervised by Dr N. Krishnaswami*

Efficient Automatic Differentiation in Functional Programming Languages

Training complex machine learning models requires accurate and efficient methods of calculating derivatives. Automatic differentiation (AD) is an increasingly ubiquitous solution to this task, and works by transforming a program representing a computable function into a program which calculates that function's derivative. A derivative-calculating program produced using this source-to-source approach has time complexity proportional to that of the input program, which makes AD favourable over the traditional method of symbolic differentiation. AD is also more accurate than numerical differentiation, because it avoids the pitfalls of round-off errors, which can be particularly problematic when dealing with complex functions, such as those used in deep learning. Many machine learning frameworks (such as TensorFlow or PyTorch) now rely on AD as a core component of their gradient-based optimisation algorithms. The functional programming community has been interested in improving the expressivity, ease of use and efficiency of these frameworks, and major functional languages now come with forward-mode AD packages. These functional alternatives are, however, orders of magnitude slower than AD implemented in for instance TensorFlow, which makes their use infeasible on real-world problems. In this presentation I'll explain what reverse-mode AD is, and how we can try to implement it efficiently in functional languages.

Alexander Rice *supervised by Dr J. Vicary*

Semistrictness for higher categories

For some operations, such as addition, we have associativity rules (a + (b + c) = (a + b) + c) and unital rules (a + 0 = 0 + a = a). When writing proofs in a computer checked proof system, we need to insert these rules everywhere, which makes proofs more complex and more difficult to write. Luckily it is known that operations such as these can be "strictified", so that the left and right hand side of these rules are definitionally the same object, removing the need to put in explicit rules. In my research, I am applying these techniques in a higher dimensional setting, where full strictification is no longer possible.

Angus Hammond *supervised by Dr N. Krishnaswami and Prof. P. M. Sewell*

Program logics for Realistic ISAs

Real instruction sets have complex semantics that are often only specified in ambiguous prose documentation. To allow for practical verification of assembly and machine code programs, it is desirable to have program logics which permit reasoning logically about program state. Ideally such logics would be mechanised in a theorem prover, so users can create complex proofs and be confident they have made no mistakes. While a number of such program logics for assembly and machine code programs exist, they are usually hand crafted by researchers from an informal understanding of the ISA spec. I'm working on

generating useable program logics for real ISAs, such as Armv8 and RISC-V, which are verified sound in the Coq theorem prover with respect to authoritative mechanised semantics, using the Iris program logic framework and the Sail ISA specification language. We use symbolically evaluated Sail source code to allow automatic simplification of ISA specifications under arbitrary assumptions so program logic rules can be generated for a specific verification effort that have significantly reduced complexity compared to the general rules. In particular this work is focused on verifying correctness of a real hypervisor.

Derek Sorensen *supervised by Prof. M. P. Fiore*

Synthetic Stable Homotopy Theory

In this talk we'll be talking about some of the fundamental notions of homotopy type theory and stable homotopy theory, what it means to make it synthetic, and some results I've proved in the line of this project.

Daniel Hugenroth *supervised by Prof. A. R. Beresford*

Strong Metadata-Privacy for Mobile Devices and Group Communication

In my PhD I study strong metadata-privacy for modern applications that still work well in practise on mobile devices. In this talk I present my work on an efficient multicast protocol for mix networks which is the main result of the first year of my PhD.

When I examined the feasibility of decentralized collaborative applications of anonymity networks, I discovered that multicast operations were an unaddressed bottleneck. The presented protocol reduces the latency of multicast messages in modern mix network designs from O(m) to O(log m) for groups of size m. It works by asking group members who already received a message to help distribute it further.

I also show that the scheme can be extended to be fault-tolerant to offline group members while maintaining the same asymptotic guarantees. These situations are discovered through timeouts and efficiently redistributed of forwarding duties through a seeded scheduling algorithm.

Christine De Kock *supervised by Dr A. Vlachos*

Towards understanding constructive disagreements online

Conversations online have become an increasingly important means of communication; however, they are also known to give rise to copious disagreements (see eg. Graham, 2006). Existing work on online disagreements mainly focus on the adverse effects of such disagreements: for instance, detecting hate speech, harassment and personal attacks. However, disagreement can also have a positive impact through the introduction and evaluation of different perspectives. We aim to gain an understanding of what makes disagreement constructive in certain cases. So far, we have investigated this in two ways: (1) we introduced a dataset of disagreements on Wikipedia Talk pages, and defined the task of predicting whether a dispute will be escalated to mediation, as a proxy for less successful dispute resolution. (2) We evaluated the usefulness of survival regression for time-to-event prediction in conversations. This is done through two tasks: firstly, having seen a subset of utterances in a conversation, we predict how many more utterances will follow before the conversation terminates. Secondly, we predict how soon a personal attack will

occur, based on a subset of seen utterances. Our current work focuses on representing online collaboration patterns using graphs.

Georgi Karadzhov *supervised by Dr A. Vlachos*

DEliBots - Deliberation Enhancing Bots

Research on dialogue systems focuses mostly on task-oriented and user engagement bots, thus focusing on human-to-bot dialogues. However, conversations with multiple human participants are common in real-world settings, including group problem solving, where it is known that small groups often outperform individuals.

To address that, we propose a novel type of dialogue system - bots that improve deliberation in group discussions - DEliBots. We introduce a novel dataset containing collaborative discussions on solving a cognitive task, consisting of 200 group discussions. Furthermore, we propose a novel annotation schema that captures deliberation cues and release 50 dialogues annotated with it. Finally, in this talk, I will present some of the experiments we've done on the collected dialogues, in terms of the experimental setup, analysis, and modelling.

Sian Gooding *supervised by Prof. E. J. Briscoe*

Textual Complexity is in the Eye of the Beholder

The difficulty of a text is highly subjective, yet this factor is often neglected in text simplification and readability systems which use a “one-size-fits-all” approach. In this talk, I will discuss what contributes to text complexity, emphasising how this is dependent on the intended audience. Furthermore, I will present work showing how on-device reading strategies may differ based on a reader’s first language and proficiency.

Gladys Tyen *supervised by Prof. P. J. Buttery*

Chatbots for language learning

What does it take to build a chatbot for students learning a new language? Teachers, linguists, computer scientists, and engineers, have all tried to answer this question. As the first step in my PhD, I conducted a systematic literature review to identify previous attempts at building a chatbot for language learning. I also built rudimentary versions of my own chatbot, discovering some do's and don'ts in the process. In this presentation, I describe my literature review methodology which maximises coverage over different fields of research, then outline the overall findings in my first year and implications for future work.

Ran Zmigrod *supervised by Dr T. G. Griffin and Dr R. D. Cotterell (external co-supervisor)*

Decoding and Inference in Graph-based Dependency Parsing

Many of the state-of-the-art dependency parsers are graph-based parsers: They model words as nodes of a directed graph, and possible dependency relations as edges. Therefore, dependency trees are seen as directed spanning trees (also known as arborescences) Many dependency parsing schemes such as the Universal Dependencies (UD) have an important constraint that only one word may depend on a special root symbol in the tree. This was first addressed in Koo et al. (2007) but is often left out of many dependency parsers. I will discuss how one can incorporate this root constraint into decoding and inference algorithms for dependency parsing. Furthermore, I will present an efficient algorithm for computing first- and second-order expectations over the distributions of trees which could be used in unsupervised and semi-supervised learning contexts.

Ramon Viñas Torné *supervised by Prof. P. Liò*

Learning the generating mechanisms of multi-omics data

High-dimensional multi-omics data spans multiple molecular layers and originates from an unobserved set of variables. The generative process that gives rise to the observable data is governed by an algorithm that, through a chain of steps, transforms latent causes into measurable effects. A ubiquitous example of this idea is the central dogma of molecular biology - our genetic information encoded in DNA is converted into functional products through a series of complex steps, involving transcription of DNA into messenger RNA (mRNA); translation of mRNA into a chain of amino acids; and folding the chain into functional 3D proteins that determine our phenotype. Modeling multi-omics data from a generative perspective has therefore potential to reveal important biological insights. In this talk, I will describe my work during the first year of the PhD, which explores state-of-the-art generative models for in-silico generation and imputation of multi-tissue transcriptomics data.

Dimitris Los *supervised by Dr T. M. Sauerwald*

Balanced Allocations with Incomplete Information: The Power of Two Queries

We consider the problem of allocating m balls into n bins with incomplete information. In the classical one-choice process, each ball is allocated by choosing a bin uniformly at random. This leads to an m/n+sqrt(m/n log n) maximum load whp. A well-known improvement is the two-choice process, where each ball first queries the load of two randomly chosen bins and is then placed in the lesser loaded. The maximum load decreases to m/n+log2 log n+O(1) whp.

In our setting, each ball also samples two random bins but can only send binary queries such as “Is your load above the median load?" or “Is your load larger than 100?". These are more lightweight than a full comparison of loads. A scheme that achieves O(sqrt(log n/log log n)) maximum load is known for m=n, and it was conjectured that this extends for m>n. We disprove this conjecture by showing that any adaptive binary query scheme whp will reach an Ω(m/n+log n/log log n) maximum load.

We also design a k=O(log log n) quantile process which achieves a maximum load of m/n+k(log n)^1/k, establishing a dichotomy similar to the “power of two

choices". This family of processes provides a smooth trade-off between load information and max load. For k=Θ(log log n), we recover the power of two choices result, which is of independent theoretical interest.

Felipe Ferreira Santos *supervised by Prof Anuj Dawar*

Lower Bounds on Space Constrained Computational Models

In the talk we examine the LOGSPACE (aka L) vs. P problem, one of the biggest open problems in the field of complexity theory. We survey two different computational models researchers have studied in hope of showing L is not equal to P. For each of these models we briefly outline a lower bound argument on a restricted version of the model. This is the component of the talk covering the main results of our research. We finish by demonstrating where these arguments fail when the restrictions are removed.

Erika Bondareva *supervised by Prof. C. Mascolo*

Segmentation-free Heart Murmur Detection Using Deep Learning

Cardiovascular diseases are the leading cause of death in the world, and auscultation is typically an essential part of a cardiovascular examination. The ability to diagnose a patient based on their heart sounds, however, is a rather difficult skill to master. Many approaches for automated heart auscultation have been explored. However, most of the previously proposed methods involve a segmentation step, the performance of which drops significantly for high pulse rates or noisy signals. In this work, we propose a novel segmentation-free heart sound classification method. Specifically, we apply discrete wavelet transform to denoise the signal, followed by feature extraction and feature reduction. Then, Support Vector Machines and Deep Neural Networks are utilised for classification. On the PASCAL heart sound dataset our approach showed superior performance compared to others, achieving 81% and 96% precision on normal and murmur classes, respectively. In addition, for the first time, the data were further explored under user-independent setting, where the proposed

method achieved 92% and 86% precision on normal and murmur, demonstrating the potential of enabling automatic murmur detection for practical use.

Filip Svoboda *supervised by Dr N. D. Lane*

Learned Communication in FL

Federated Learning pushes training to the edge devices themselves – so that data need not leave the full control of its owners. It offers an appealing solution to many of the security, privacy, and bias challenges Deep Learning is experiencing right now. Its major limiting factor is the cost of communication imposed on the network of participating devices. My work focuses on decreasing this cost using learned compression approaches powered by the RL and the NAS.