Projects

The PhD project proposals listed below will be considered for 2020/21 studentships in the Department of Informatics to start in October 2020 (but an earlier start in April 2020 is possible). This list is not inclusive and the potential applicants can alternatively identify and contact the appropriate supervisors outlining their background and research interests or proposing their own project ideas. The PhD projects are listed in two groups. In the first group are the projects with allocated studentships, that is, each project in the group has one allocated studentship. The remaining studentships will be considered for the projects listed in the second group. The number of these studentships is smaller than the number of the projects in this group.

Contents

Projects with allocated PhD studentships

Algorithms and Data Analysis

Supervisors: Dr Dimitrios Letsios

One PhD studentship will be allocated to the following project which lies at the intersection of algorithms, computational optimization, and data science.

Models and Algorithms for Resource Allocation Problems with Machine Learning Predictions

This project aims to design and analyze optimization models and algorithms for temporal resource allocation problems, e.g. electric power distribution, logistics, and production scheduling problems, arising in different application domains, including the energy sector, manufacturing and process engineering [Letsios et al. 2020]. The goal is to effectively assign resources, e.g. machine time and energy, to activities, so as to optimize performance. Solving instances of such problems may result in substantial economic benefits. Typically, future resource requirements and customer demand are not precisely known in advance, but can be predicted using data science and machine learning capabilities [Bertsimas et al. 2018]. However, these predictions are subject to errors. In this context, determining efficient algorithms for supporting and automating the resource allocation process is a challenge. To this end, prior work develops efficient algorithms and optimization models accounting for the time-varying nature and uncertainty of temporal resource allocation problems [Antoniadis et al. 2020, Letsios et al. 2021,Manish et al. 2018].

This project aims to (i) develop novel discrete optimization methods for temporal resource allocation methods and analyze their performance theoretically, (ii) suggest ways to mitigate the effect of prediction errors in the quality of the obtained solutions, and (iii) evaluate the performance of the proposed approaches numerically using real data. Prior experience on discrete optimization, approximation/online algorithms and/or integer programming will be useful.

- [Antoniadis et al. 2020] Antonios Antoniadis, Christian Coester, Marek Eliás, Adam Polak and Bertrand Simon. Online Metric Algorithms with Untrusted Predictions. International Conference on Machine Learning (ICML), 2020.
- [Bertsimas et al. 2018] Dimitris Bertsimas, Vishal Gupta, Nathan Kallus. Data-Driven Robust Optimization. Mathematical Programming, p. 235-292, 2018.
- [Letsios et al. 2020] Dimitrios Letsios, Radu Baltean-Lugojan, Francesco Ceccon, Miten Mistry, Johannes Wiebe, Ruth Misener. Approximation Algorithms for Process Systems Engineering. Computers and Chemical Engineering 132, 2020.
- [Letsios et al. 2021] Dimitrios Letsios, Miten Mistry, Ruth Misener. Exact Lexicographic Scheduling and Approximate Rescheduling. European Journal of Operational Research, 2021.
- [Manish et al. 2018] Manish Purohit and Zoya Svitkina and Ravi Kumar. Improving Online Algorithms via ML Predictions. Advances in Neural Information Processing Systems (NeurIPS), p. 9661--9670, 2018.

Data architectures with humans-in-the-loop

Supervisor: Professor Elena Simperl

Fundamental data-centric tasks such as conceptual modelling, content labelling, entity extraction and query processing are routinely realised as hybrid processes, which consist of human and algorithmic elements. Examples include any AI system that depends on large amounts of labelled data, interactive machine learning systems, but also knowledge graphs such as Yago, Wikidata, or DBpedia, which are created by people alongside a range of more or less sophisticated bots.

The projects in this category explore methodologies, computational methods and tools that go beyond the capabilities of existing AI and machine learning stacks in terms of tasks, performance and user experience. For example topics will include:

- Novel methodologies and tools to create knowledge graphs, offering advanced user experiences, accessible to non-experts and using the latest tech (audio and video processing, intelligent assistants, AR and VR etc.)
- Methodologies and techniques to acquire and encode common sense knowledge at scale
- Quality of knowledge graphs, including frameworks to define it, methods to assess and repair it, and the link between process, provenance and outcomes
- New interfaces and experiences e.g. conversational agents to collect and curate knowledge and improve algorithmic performance.
- Managing discussions, collaborative decision making and conflicts.

Efficient Mechanism Design for Markets and Reallocation of Goods

Supervisor: Dr Bart de Keijzer

This project focuses on the designing computationally and economically efficient mechanisms for market and exchange platforms.

On such platforms, a number of agents are present who have the intention of selling or buying items from other agents. A mechanism interacts with these agents and determines, based on this interaction, how the agents should trade and against which payments. The agents are assumed to act rationally, in the sense that they have a certain utility function which they want to optimise: For example, a natural setting would be one where agents want to maximise the total value of the items received, plus the payment that they potentially receive in compensation for losing some of their goods. The agents will interact with the mechanism in such a way that their utility is optimised, and mechanisms for such scenarios need to be designed in such a way that trade happens in an optimal way, while agents are not able to "cheat" the mechanism for their own benefit. Moreover, these mechanisms should perform their computations reasonably (and provably) fast. How to design the trading mechanism in such a way that these requirements are satisfied?

As there are many details that need to be specified in the above sketch to yield a very concrete model, this gives rise to a wide range of interesting mechanism design challenges. Different properties of the market require different mechanisms, where one can think of e.g. a static "one-shot" trading scenario versus a scenario where agents can dynamically enter and exit the market, or indivisible versus divisible goods, shareable vs unshareable goods, etc. In this project we will work on trying to solve various challenging variants of this design problem.

This is a project in algorithmic game theory, which means that it lies in the intersection of theoretical computer science and economics. This project relates strongly to computational complexity theory, approximation algorithms, matching theory, auction theory, and (clearly) mechanism design. This is a theoretical research field which is in the lucky position of also being relevant in practice: As examples of where this field is applied, one may think of ad-auctions in search engines, and various automated market platforms where goods are exchanged, or where clients are assigned to service providers (think of various popular platforms for taxi drivers, finding holiday accommodation, food delivery, and transportation services for goods).

Imperfect Rationality and Computation

Supervisor: Professor Carmine Ventre

Algorithmic Game Theory is a research field that provides a set of tools to account for strategic reasoning in computer science. One assumption underlying much of the work in the area is, however, pretty limiting: agents need to be fully rational. This is unrealistic in many real-life scenarios; we, in fact, have empirical evidence that people often misunderstand the incentives and try to game the system even when misbehaving is against their own interest.

This project will look at novel approaches to deal with imperfect rationality, including the analysis of known systems and the design of novel ones. This will involve theoretical work (such as, mechanism design) as well as more applied approaches (such as, agent-based modelling) to get a better understanding of the strategic interactions within a population of agents with imperfect rationality.

Understanding the Complexity of Negotiations

Supervisors: Dr Alfie Abdul-Rahman & Dr Rita Borgo

A negotiated text is the product of a formal decision-making process where a text has been negotiated and drafted over a period of time. Many of the foundational texts of the modern world have not been written by individuals, by negotiated by groups of people in formal settings. For example, treaties between states such as the Universal Declaration of Human Rights or the Treaty of Versailles; or constitutions, such as the one negotiated by the American states in the Constitutional Convention of 1787.

During such negotiations, it is important for us to keep track of the delegations and their involvements in order to grasp their influence on the negotiation process either using techniques such as close reading, distance reading, or machine learning. Even relatively short historical documents written collectively in this way have been the product of thousands of specific proposals and decisions.

This project will apply a visual analytics approach towards the understanding of the complexity of a negotiation and the influence of the delegations during a negotiation process.

Possible research questions:

- Developing new static and interactive visualization to assist with data discovery and insight generation in large datasets of events within interacting timelines.
- Developing new approaches to show the evolution of complicated, technical documents over the period of months or years.
- Developing new approaches for indexing the datasets related to the negotiation of documents, and more intuitive displays of the results.
- Developing natural-language-based approaches to relate information captured in 'informal' archives (such as private diaries, letters, social media feeds etc.) to the formal records of a negotiation.

This project will work closely with the Quill Project, based at Oxford University: https://www.quillproject.net/

Visual recognition with minimal supervision in deep learning context

Supervisors: Dr Miaojing Shi & Dr Michael Spratling

The goal of this PhD is to study object detection/segmentation in images or video with minimal supervision. This task will be placed into a setting where only image-level annotation is provided. To begin, additional supervision such as clicks, strokes, or bounding boxes may also be assumed. Towards the end of the PhD, the student is expected to work with datasets of mixed levels of supervision, including a harder, semi-supervised setting where there are only a few image-level labels as well as a large amount of unlabeled images.

Several ideas can be investigated in the context of deep learning. For instance, generative adversarial learning can be employed to either augment the dataset or bridge the predicted detections with their ground truth. Recurrent neural networks can be applied to video segmentation in particular to localize and segment semantic parts across nearby frames. On unstructured image datasets, ideas like deep metric learning and random-walk label propagation can be extended across pairs or groups of images. Cross-category transfer learning can be a further extension.

Few-shot learning is another challenging direction to explore. After learning on a set of base classes with abundant examples, new tasks are given with only few examples of novel (unseen) classes. For such cases, the learning strategy of multi-million parameters architectures in deep learning needs to be rethought in order to allow the networks to squeeze out the maximum amount of information from the few available samples.

Wearable, Discreet Augmentative and Alternative Communication

Supervisor: Dr Timothy Neate

Please note: applicants to this proposal are welcome to self-fund, or apply for the studentship or K-CSC Scholarship.

Approximately 2.2 million people in the UK experience a form of communication impairment [1], including a third of stroke survivors and 2/3 children in each classroom.

Communication impairments might mean that people find it hard to convey or understand information when they need it most. This is different for everyone, but communication impairments can affect one’s reading, writing, speaking and/or listening. People with communication impairments often use AAC (Augmentative and Alternative Communication) to support them in communication; generally, via a laptop, tablet or smartphone. These, assistive devices are not always quick to access and often carry with them a stigma [2].

Wearables, such as smartwatches and smart glasses, have the potential to provide a range of sensors and modes of input/output within an unobtrusive, commonplace form factor. Wearables are discreet. Their always-available nature, coupled with instant access to the internet and processing (e.g., recognition models) on a companion device (e.g., a smartphone), have the potential to support people with communication impairments in accessing and expressing information in a subtle and less obtrusive manner.

Building upon work supporting access with wearables [3], this PhD project will conduct co-design of wearable applications and models which can support people with communication impairments in everyday life. Using established co-design approaches with users with communication impairments (e.g. [4]) this work will develop a range of input and output approaches with consumer and potentially custom form factor wearables, working closely with end-users and evaluate them in real-world contexts.

REFERENCES

[1] UKGov, “Disability prevalence estimates,” 2012.
[2] Phil Parette and Marcia Scherer. Assistive Technology Use and Stigma. Education and Training

in Developmental Disabilities
Vol. 39, No. 3
[3] Dhruv Jain, Hung Ngo, Pratyush Patel, Steven Goodman, Leah Findlater, and Jon Froehlich. 2020. SoundWatch: Exploring Smartwatch-based Deep Learning Approaches to Support Sound Awareness for Deaf and Hard of Hearing Users. In The 22nd International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '20).
[4] Timothy Neate, Aikaterini Bourazeri, Abi Roper, Simone Stumpf, and Stephanie Wilson. 2019. Co-Created Personas: Engaging and Empowering Users with Diverse Needs Within the Design Process. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19).

Supporting more Accessible Remote Communication

Supervisor: Dr Timothy Neate

Please note: applicants to this proposal are welcome to self-fund, or apply for the studentship or K-CSC Scholarship.

Remote communication platforms such as Zoom and Microsoft Teams make use of video and audio but lack the physical affordances we take for granted when talking in-person. In 'real-world' discussions we frequently use gestures, sketches on pieces of paper and make use of props within the physical environment. While we might make do without these capabilities, non-language communication modalities are particularly important for those with communication impairments and mean that there is an access barrier in remote compared to co-located communication.

Access to remote communication platforms has been spotlighted by the Covid-19 pandemic. Without equal access, people with communication impairments face the risk of further isolation, wherein access to social interaction, civic engagement and even vital speech and language therapy might be challenging.

Building upon prior work in gestural communication [1], this project will design novel interaction techniques which leverage commercially available and bespoke technologies to support communication in remote settings. Platforms will explore how the capabilities of commercial devices beyond video and audio (such as Lidar sensing, augmented reality and external devices such as wearables) might support 'real-world' affordances e.g. through custom recognisers and sensor fusion.

This project will involve a PhD student working with people with diverse language impairments and speech and language therapists to co-develop a fully functional platform aimed specifically at supporting remote communication, to complement existing video conferencing platforms.

REFERENCES

[1] Roper, A., Marshall, J. and Wilson, S. (2016). Benefits and Limitations of Computer Gesture Therapy for the Rehabilitation of Severe Aphasia. Frontiers in Human Neuroscience, 10.

Object-Based Access: Enhancing Accessibility with Data-Driven Media

Supervisor: Dr Timothy Neate

Please note: applicants to this proposal are welcome to self-fund, or apply for the studentship or K-CSC Scholarship.

The tools by which we use to create and consume media-rich digital content such as video streaming, podcasts, TV and radio, are undergoing substantial change. The core idea behind the future media ecosystem is Object-Based Media (OBM). OBM is the practice of linking media assets, such as individually recorded audio and video, to metadata. These metadata might include anything from time-stamped information about what happened in a video, who shot the video, or details about how the media should be played on different devices [1]. Digital media are sent as a collection of objects, arranged on a user's device according to their exact needs. Vitally, this means each user's experience of creation and consumption can be different.

While OBM principals might be used to automatically change a programme's duration to meet our time requirements [2] or make a story more relatable to us by including local information based on our location [3], there are also substantial implications for accessibility. As the content creation and consumption process differs for everyone, they have the potential to be accessible to everyone.

This PhD project will work with a range of stakeholders to develop accessible alternative media formats, workflows and interaction techniques for the creation of novel interactive media experiences. Building upon prior work accessible content workflows (e.g. [4, 5]) this project will consider how the future of content creation might be made more accessible through extreme customisation of content that is different to everyone and bespoke to their needs.

REFERENCES

[1] D. Varghese, P. Olivier, and T. Bartindale, "Towards participatory video 2.0," in proc ACM CHI, 2020.
[2] M. Armstrong, M. Brooks, A. Churnside, M. Evans, F. Melchior, and M. Shotton, "Object-based broadcasting-curation, responsiveness and user experience," 2014.
[3] S. Concannon, N. Rajan, P. Shah, D. Smith, M. Ursu, and J. D. Hook, "Brooke leave home: Designing a personalized film to support public engagement with open data," in proc. CHI, 2020.
[4] T. Neate, A. Roper, S. Wilson, and J. Marshall, "Empowering expression for users with aphasia through constrained creativity," in proc. ACM CHI, pp. 1–12, 2019.
[5] T. Neate, A. Roper, S. Wilson, J. Marshall, and M. Cruice, "Creatable content and tangible interaction in aphasia," in proc. ACM CHI, 2020.

Projects for the remaining 2020/21 studentships

ActOML: Active and Online Learning for Delaying Time Decay in ML Security Tasks

Supervisor: Professor Lorenzo Cavallaro

Concept drift strongly affects malware classifiers [1]: when the statistical properties of testing malware objects change with respect to those the classifiers were trained against, the assumption that training and testing datasets are independent and identically distributed does not hold anymore. As such, the performance of malware classifiers decay over time.

We have recently explored space and time evaluation bias and the understanding of time performance decays of ML classifiers in security tasks under realistic settings due to concept drift [2] . This PhD proposal aims to build on such work and improve the robustness of malware detection algorithms against time decay (safe AI). To this end, we will develop novel active and online learning strategies for malware classifiers with the idea of exploring robustness against adversarial ML [3] as well. In this context, the proposal may venture on exploring traditional as well as deep machine learning to understand whether recent advances, such as Trusted Neural Network [4], represent the ideal framework to encode additional logical constraints to harden classifiers against such threats.

The proposals requires working at the intersection of program analysis and machine learning for systems security and the successful candidate is required to have a solid programming and CS/CEng background with a particular passion and knowledge in the foundations of systems, program analysis and machine learning. Given the nature of the research, the proposal links perfectly with the aims of the Security Hub at King's.

[1] Transcend: Detecting Concept Drift in Malware Classification Models Roberto Jordaney, Kumar Sharad, Santanu K. Dash, Zhi Wang, Davide Papini, Ilia Nouretdinov, and Lorenzo Cavallaro USENIX Security Symposium, 2017, https://s2lab.kcl.ac.uk/papers/files/usenixsec2017.pdf

[2] TESSERACT: Eliminating Experimental Bias in Malware Classification across Space and Time Feargus Pendlebury*, Fabio Pierazzi*, Roberto Jordaney, Johannes Kinder, and Lorenzo Cavallaro USENIX Security Symposium, 2019, https://s2lab.kcl.ac.uk/papers/files/usesec19final.pdf

[3] Evasion Attacks against Machine Learning at Test Time. Battista Biggio, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim Srndic, Pavel Laskov, Giorgio Giacinto, and Fabio Roli. ECML-PKDD, 2013

[4] Trusted Neural Networks for Safety-Constrained Autonomous Control. Shalini Ghosh, Amaury Mercier, Dheeraj Pichapati, Susmit Jha, Vinod Yegneswaran, and Patrick Lincoln. arXiv 2018

Administrative Access Control Policies

Supervisors: Professor Maribel Fernandez & Dr Jose Such

Administrative access control policies specify the rights that security administrators have in the system (e.g. to add or remove users, or change users' rights). These policies are critical to ensure the overall security of the system, but not much work has been done on the development of general models for administrative access control. In this project we aim to define formal generic models of administrative access control, based on the Category-Based Meta Model of access control (CBAC), which can be used to analyse access control systems and help identify the impact of changes made by administrators (impact change) on the overall security of the system.

“Alexa, cover your ears!”: Privacy-Aware AI Personal Assistants

Supervisor: Dr Jose Such

AI personal assistants are becoming mainstream in practice, with the widespread introduction of desktop, phone and home assistants. For instance, over 70 million users utilise smartphone assistants like Siri, Google Assistant, and Cortana every day; and smarthome assistants have been sold in massive numbers, like the five million units of Amazon Echo with the Alexa personal assistant sold in less than two years. However, recent incidents involving AI personal assistants like Alexa recording a private conversation and sending it to a random contact, have increased users' privacy concerns, with some users trashing their assistants all together and companies like Mattel cancelling assistant projects. It is therefore paramount to consider and respect users' privacy to realise the benefits of AI personal assistants and foster trust from users. In this project, you will formalise the social norms that govern information sharing, management, collection, processing and learning in AI personal assistants. Based on this, you will design novel methods to personalise privacy in AI assistants based on the social norms but also on the users' contextual, group, and individual preferences with an optimal accuracy-intervention trade-off.

Analysis of Contingent Liabilities in Debt Sustainability Analysis (DSA)

Supervisors: Dr Kevin Lano

Debt sustainability analysis is of critical importance to national economic management, particularly for developing economies. However, existing DSA models are often simplistic and fail to take into account contingent liabilities or the probabilistic nature of economic variables (Haughton, Commonwealth Secretariat, 2016). The research involves identifying and evaluating alternative approaches for DSA that provide stochastic modelling of debt and contingent liabilities, and comparing to standard models as used by the World Bank and IMF.
Analysis of Novel Interaction Methods in Programming Education
Supervisors: Professor Michael Rolling & Dr Neil Brown

The goals of the design of Frame-based editing [1] was to achieve improvements in various aspects of learning to program. Fewer syntax errors, better motivation, faster achievements, better retention, improved general satisfaction. Does this design achieve these goals? We don't know.

This project would study the effectiveness of Stride compared to other systems with real users.
We envisage that multiple user-studies with pupils in classrooms would be part of these investigations. Studies would employ methodologies from educational and computer science research.

[1] https://www.greenfoot.org/frames/

Automated Signature Generation for Network Intrusion Detection Systems (NIDS)

Supervisor:Dr Fabio Pierazzi

A Network Intrusion Detection System (NIDS) is a probe that passively monitors network traffic and triggers a “security alert” whenever a signature matching a particular pattern is found. However, signatures are still mostly generated manually, a process which is error-prone and time consuming.

This project will explore how AI and ML can support automated generations of signatures which are effective, efficient and interpretable. Evaluations will include adaptivity to different network environments, and performance speedup in terms of DR and FPR with respect to manually-defined signatures and existing statistical approaches.

Backbone Guided Local Search Methods for MAX-SAT

Supervisors: Dr Kathleen Steinhofel & Dr Dimitrios Letsios

Satisfiability (SAT) is a key problem in combinatorial optimisation and has a huge range of real-life applications. It seeks for a given Boolean Formula (conjunctive normal form) an assignment of variables such that the formula returns True. In case such an assignment does not exist, we seek an assignment that satisfies a maximum number of clauses (MAXSAT). As backbone structure, we denote the set of variables that have the same assignment in all optimal solutions.

Knowledge about the backbone structure can be used to guide heuristic methods which aim to find near optimal solutions. For instance, the size of the backbone can give indications of how many optimal solutions exists and consequently how hard it is for the search method to converge to an optimal solution. The guidance can be provided in two different types:

1. Deriving instance dependent methods by using pre-processing to approximate the backbone structure and to derive parameter settings for local search.
2. Estimating the backbone structure based on configurations visited by the local search method.

The findings will lead to faster convergence to optimum solutions and more importantly can produce methods which adapt to instance dependent properties. At the same time, methods to derive and analyse the backbone structure can be used to classify candidate solutions and to model additional, sought after properties such as robustness of candidate solutions.

Behavioral Modeling of Process Memory for Real-Time Detection of Attacks

Supervisor:Dr Fabio Pierazzi

Memory vulnerabilities such as buffer overflow, heap spraying, heap vulnerabilities, are still one of the major threats in all modern systems. Most modern approaches to detect memory attacks are based on heavyweight monitoring and analysis which causes significant overhead and prevents realtime application. This project will explore how AI and ML can be used to create a behavioral model of process memory for real-time anomaly detection of attacks occurring in memory.

Better Error Help Using Large Scale Programmer Data

Supervisors: Professor Michael Kolling & Dr Neil Brown

Could large scale beginning programmer data be used to give useful hints and help to beginners stuck on an error? For example, if a novice had problems with a task, could perhaps useful hints be automatically generated by analysing previous users who had similar problems, what they did, and whether their actions led to solving the problem?

Making use of the Blackbox data set [1] is one option to automatically generate helpful hints and tips for novice programmers.

[1] https://bluej.org/blackbox/

Big Data in Programming Education
Supervisors: Professor Michael Rolling & Dr Neil Brown

The Blackbox project has collected a large amount of data about the behaviour of novice programmers. We have data about hundreds of millions of programming sessions. So far, this data has been analysed only very superficially. An interesting project would be to use a big data approach for deeper analysis of this data set, and to work out what we could learn from this.

Characterization of Immunoglobulins

Supervisors: Professor Costas Iliopoulos, Dr Sophia Karagiannis & Dr Grigorios Loukides

Antibodies, or immunoglobulins, belong to the 'gamma globulin' protein group and can be found mainly in the blood of vertebrates [1]. Antibodies constitute the major serological line of defense of the vertebrates with jaws (gnathostomata) by which the immune system identifies and neutralizes threatening invaders, such as viruses, fungi, parasites, bacteria. The contrivance underlying the reaction efficiency of our immune system to specifically recognize and fight invading organisms or to trigger an autoimmune response and disease still remains to be elucidated. The efficient reaction of our immune system against all kinds of intruders is highly dependent on the number, condition and availability of antibodies, as reaction times are 'key' to the successful elimination of the foreign pathogen.

The importance of antibodies in health care and the biotechnology industry demands knowledge of their structures at high resolution. This information can be used for antibody engineering, modification of the antigens binding affinity and epitope identification of a given antibody. Computational approaches provide a cheaper and faster alternative to the commonly used, albeit laborious and time consuming, X-ray crystallography. Available immunogenetics data can be utilized for computational modelling of antibody variable domains. Standardized amino acid positions and properties can assist in optimizing the relative orientation of light and heavy chains as well as in designing homology models that predict successful docking of antibodies with their unique antigen. As a result, it comes down to identifying conserved motifs or patterns that are implicated and mediate antibody-antigen interactions.

Detection of such motifs by simple sequence comparison is impossible. Consequently, our research is fixated on the investigation of alternative approaches to efficiently study antibodies, mainly by the multimodal fusion of information from genetic, structural and physicochemical analysis.

All in all, herein we propose a holistic approach in the realm of immunoinformatics that will focus on elucidating the mechanism of antibody-antigen recognition. The results and the final tool (in the form of either an online service or a downloadable tool) will be made freely available to the scientific community. We are confident that many fellow researchers from all walks of immunology, bioinformatics and antibody-related sciences will benefit from such a tool, both in terms of applied research and basic understanding of the function of CDRs.

Nowadays, it is certain that such specialized and specific recognition properties cannot be based on random and hypervariable sequences. It is just that using the 20 amino acid code is not a suitable approach to explain the phenomenon. Therefore, herein, we will calculate more than 430 different physicochemical properties to represent each residue of all antibodies, in an effort to identify what is the right dictionary (or indeed the right combination of dictionaries) required to decrypt the antibody-antigen interaction puzzle. The calculation of the physicochemical properties will be done using the QSAR module as it is implemented into the Molecular Operating Suite (CCG).

A brief overview of our main objectives includes the following:
- C ollecting and building the working dataset
- Collect and curate antibody structural data from numerous databases
- Deep learning for feature extraction and prediction • Predict reliable classification markers through

the use of convolutional neural networks (CNNs)

Computational Analysis of Ageing Brains

Supervisors: Dr Kathleen Steinhofel & Professor Zoran Cvetkovic

The ability to acquire and store information is a key function of the brain. This ability is affected by ageing and in various age-related disease, including dementia. In old age the acquisition of new information is more difficult than in young age. Moreover, updating of acquired information is also affected by ageing. The mechanistic basis of the age-related decline is not well understood. It is known that changes at synapses, the connections between nerve cells, are the basis of information storage. But it remains unknown how the synaptic basis of information storage changes with age.

Recently, ultrastructural changes at synapses were discovered and analysed after training in a memory task in young and aged mice. In the research programme, we want to investigate the impact of these changes by using computational approaches based on models of these biological observations. In addition to the modelling of ultrastructural changes, the regulatory function and expression level of microRNA in the neuro cell will be analysed towards the impact to the ability to store information. The findings will not only advance insights into mechanisms of information storage, but also support the analysis of age-related diseases, such as dementia, that affect cognition.

The supervisor team will include Prof Peter Giese (IoPPN) and Anna Zampetaki (Cardiovascular Division).

Contextualising Big Programming Data

Supervisors: Professor Michael Kolling & Dr Neil Brown

The Blackbox project has been running for over five years. It collects data from noviceprogrammers: source code that is written, compilation errors displayed, and various other data about a programmer's interaction with the BlueJ IDE. The data is collected without any further context: we do not know the age or experience of the programmer, whether they are on a course or not, whether they are doing well on their assignments, and so on. This allows for a large data set, but one that is stripped of useful context. This project would investigate collecting useful data (e.g. experience, course grades) for a subset of Blackbox participants, to provide a richer subset for other researchers, and to be used in the project to investigate associations between programming activity and success on a course.

Data Science Strategies for Cancer Immunotherapy Application

Supervisors: Dr Sophia Tsoka & Dr Grigorios Loukides

Computational analysis of biomedical datasets can lead to understanding of disease systems and therapeutic interventions. We propose a project that will target the computational analysis of experimental data on immune activation against cancer using antibodies. Integration of experiments with publicly available data on known cellular interactions will establish a resource for data mining. Such a resource will be used to implement machine learning algorithms to link gene features to cancer response, network analyses to represent molecular interactions and logical modelling to explore regulatory effects from proteomic experiments. The combination of these Data Science frameworks will elucidate signalling networks related to the control of tumor growth by antibody- enhanced human immune cells and identify key altered pathways and their regulation state. The long-term prospect is to improve understanding of disease mechanisms and cell signalling, so as to improve the design of novel drugs and therapies.

Development of Scalable General Artificial Intelligence (AI) Problem Solving Systems

Supervisor: Dr Amanda Coles

This project aims to develop scalable general Artificial Intelligence (AI) problem solving systems, capable of reasoning with the large combinatorial problems that arise in effectively managing the oversubscribed infrastructure of densely populated cities. This project builds on a study, supervised by Dr Amanda Coles (KCL Informatics) an expert in AI Planning and Professor Christopher Beck (University of Toronto) an expert in Constraint Programming (CP), exploring the application of CP and AI planning to disruption recovery in the UK rail network. The PhD project aims to significantly increase the solution quality and scalability of AI problem solving technologies, based on our new understanding of the strengths these approaches, by automatically decomposing problems so CP solvers and AI Planners solve the parts best suited to their strengths. The successful candidate will extend the state-of-the-art in AI research and have the opportunity to apply this to real-world UK rail network problems.

Distributed Computing by Population Protocols

Supervisor: Professor Tomasz Radzik

Population protocols are a simple model of distributed computing, with applications extending to other areas, including processes in chemical network and online social networks. This model assumes that the computing system consists of a large number of identical devices, called agents or nodes, which communicate with each other in pairwise interactions. The pattern of interactions depends on external factors and interacting nodes follow a simple protocol, which should ensure that all nodes gradually learn some global property of the system. This project is a study of the computational potential and limitations of this model and an investigation of applications.

Formal verification of smart contracts

Supervisor: Dr Hana Chockler

Formal verification of software is gaining popularity for verifying increasingly complex and safety-critical software. While the full verification task is unsolvable (the problem is easily reduced to the halting problem, which is undecidable), numerous existing solutions to subproblems are general enough to provide thorough verification and correctness assurance for real-life systems. There is a number of teams currently working on tools for software verification, with the Formal Verification Team at the University of Lugano (USI), led by Prof. Sharygina, being one of the most established ones. Prof. Sharygina recently received Swiss government funding for a large project titled “Beyond Symbolic Model Checking through Deep Modelling”, in which Dr. Chockler (the first supervisor) is a named collaborator. The proposed Ph.D. project will be done in collaboration with the team at USI. The student will be able to travel to work face-to-face with the team in Lugano, and close collaboration via skype and emails is expected when the student is in London.

Current model-checkers (automated formal verification tools) are mostly suitable to verify programs in C and C++. In this project, we will research the direction of formally verifying smart contracts. The student will research different options of extending the verification platform to smart contracts written in Solidity (or other languages) and will analyse whether the verification should be done on the source code level or on the bytecode level.

Smart contracts are typically small. However, they interact with other contracts and are being called in a loop or recursively, thus leading to a number of subtle bugs (see, for example, the exploit of the DAO bug, leading to loss of $50 Millions). It is then reasonable to expect that the best way to formally verify smart contracts is by using *modular* reasoning: for each smart contract, the other contracts with which it interacts can be considered an environment. This environment can be overapproximated using learning techniques in combination with sampling and traditional model checking approaches. After verification of a single contract passes successfully, some symbolic representations of the contracts with respect to the correctness properties will be combined to prove correctness of the overall system.

The project will include a significant implementation component. The implementation is done using the software verification platform developed at USI. The main development task is the new front-end, so that the verification platform is able to analyse programs in Solidity (or EVM bytecode). As model-checkers require writing a large and complex software, the advantage of having such a software available already and being in contact with the team that develops and maintains it is hard to overestimate. In addition, being a part of a very active and experienced research team guarantees discussions and collaborations that further aid the research, especially in the initial stages.

Human data interaction

Supervisor: Professor Elena Simperl

Technology can play an important role in improving people's experiences with data, whether in a professional context, or in everyday life. Projects in this space look at human factors that affect our ability to find, make sense and communicate with data, including topics such as:

- Dataset search and discovery, including Google's dataset search engine
- Data portals: how are they used and how can they be improved
- Communicating and presenting data, metadata and data-related activities
- User experience in data science and data engagement
- Tools and experiences to increase accessibility of data and data science work
- Collaboration in data science
- Data storytelling tools with narrative support
- Data science communities: where are they, how do they work, how can we make them better?

Example project: New interfaces and experiences to data engagement

The project will explore novel ways to present a dataset, for example a CSV file, using speech, audio or video technologies. The aim is to propose an algorithm that given a dataset produces a media summary of the content and context of the data and evaluate the results in a user study. The algorithm could use a range of techniques, including machine learning, computer vision and speech generation. This will also require capabilities to generate text from data, as text is more accessible that metadata to convey what a dataset is about and how it should be used. In previous studies we used a manual approach to create summaries, which does not scale. The aim here would be to use natural language generation to automatically create short text summaries for a given tabular dataset, formatted, for example, in CSV. The project could use machine learning or rule-based techniques.

Example project: Personalising dataset search

In previous studies we explored different ways to present datasets in the context of search, including structured metadata, text descriptions, data previews and visualisations. In this project, the aim is to develop an information retrieval algorithm that tests the impact of these different result presentations and personalizes them based on user preferences and feedback.

This could include, among other things, personalised analogies for numerical data. Research has shown that using familiar concepts to describe numbers and numerical datasets can improve engagement. The aim of this project is to explore the same approach for a wider range of datasets (beyond spatial data such as distances and areas) and to develop an approach that for a given numerical dataset learns to recommend relevant analogies.

Example project: communicating data quality

Most work in data visualisation has focused on choosing and customising charts and stories to communicate data. The aim of this project is to look into contextual aspects of data use, including sources, uncertainty, missing or incorrect values, timeliness and the way this additional information could be embedded into visual design. The project will first undertake a survey of existing approaches for numerical datasets and then propose and test ways to communicate less explored quality aspects.

Human and social factors in information systems

Supervisor: Professor Elena Simperl

Some of most remarkable online platforms and tools we are using today bring together human and social intelligence with data and algorithms in ingenious ways. Underlying them, there is a huge, interdisciplinary research space concerned with the design principles, methods and tools that allow us to build such systems and understand and predict their evolution. The most successful of them seem to share a core set of principles:

- They are decentralised and self-organizing, and can mobilize a critical mass of resources effectively, whether that's people, data or computational devices.
- They make extensive use of mobile, sensor and web platforms, alongside openly available data and software to enable communication, knowledge exchange and coordinated action.
- They know how to bring crowd and machine capabilities together to achieve their aims in a sustainable way.
- They empower individuals to self-organise and commit to being fair, transparent and accountable about the data and resources these contribute.

Relevant topics include applications of crowdsourcing and social computing to AI systems, as well as fundamental crowdsourcing research around task and workflow design, crowd learning, quality assurance, and ethical crowdsourcing. The research would potentially focus on a class of social machines, including peer-production systems, human computation platforms and participatory sensing networks.

Example project: Improving task design

There is a large body of literature exploring how to achieve a particular goal via crowdsourcing and proposing workflows and improvements. The aim of this project is to derive such task design guidelines from a new source: discussion forums used by the crowd, for example on Mechanical Turk or in citizen science projects on the Zooniverse platform. The project will collect a sample of relevant discussions and extract comments pertinent to design guidelines, using, for instance, quantitative (NLP) or qualitative techniques.

Example project: Crowd self-assessment

In crowdsourcing, asking participants to self-assess their skills and performance helps designers understand the feasibility of the task and identify areas of improvement. Previous research has looked at the ability of crowd participants to self-assess. The aim here would be to carry out a follow-up study to understand whether the initial conclusions apply to other tasks, domains and workflows.

Learning of Software Design Patterns

Supervisor: Dr Hana Chockler

A software engineer joining a development team typically does not start writing software immediately; first, she needs to understand the large existing body of code and recognise the key components and how they interact with each other. Documentation is typically sparse and not updated regularly. The software, on the other hand, is large and difficult to understand. “Software is like entropy: It is difficult to grasp, weighs nothing, and always increases.” (Norman Augustine). Project development tools aid understanding the software by identifying the participating classes, and the static dependencies
between them.

The next step is to identify a set of design patterns common to this project, e.g.: when are resources allocated and freed, in what manner are certain components of a class visited? In this area, the static analysis tools are insufficient. The proposed project is to learn the design patterns in the given code automatically by applying grammatical inference (learning) algorithms.

The benefits for automatically learning design patterns go beyond helping the software engineer getting a clear representation of the design patterns. An automatic analysis can discover areas where the same goal is achieved by utilising different patterns: one of the patterns can be erroneous or

obsolete, or the multitude of patterns can point to the lack of precise development guidelines for a certain task, indicating a need for a guiding design pattern. These challenges require developing learning algorithms that can learn several automata simultaneously, such that the resulting automata correctly capture the main abstractions in the given corpus of code.

The project will build upon previous research results of Dr Hana Chockler in collaboration with the University of Oxford.

Machine Learning Augmented Algorithms

Supervisor: Dr Frederik Mallmann-Trenn

Machine learning and in particular deep learning has gained much attention over the past several years, yet theoretical understanding is still very limited. As a remedy, a recent line of research emerged in which neural networks are used as a black box in many online problems, where the data arrives over time. The idea is to use a neural network to give predictions of the data that will arrive in the stream. The goal is to design algorithms that perform much better than previously if the prediction is good and on the other side, to show that even if the prediction is bad, the solution found by the algorithms is still reasonably good.

A toy example is the ski rental problem, where each day a skier on vacation has to make a decision: either rent skis for 10$ or buy skis for 100$. We assume that the ski trip can end abruptly (chosen adversarially). See https://en.wikipedia.org/wiki/Ski_rental_problem for classical algorithms. Now if we assume that a neural network makes a prediction on when the ski trip will end, how much better can we do? The field is very young and great problems of practical and theoretical importance await!

Related literature: http://www14.in.tum.de/personen/albers/papers/inter.pdf for a technical survey on online algorithms.

Model Driven Engineering in Finance

Supervisors: Dr Kevin Lano

In the finance industry there is a strong emphasis on the rapid time-to-market of new financial software products and financial models, which can conflict with the achievement of software quality and correctness. The proposed research will investigate how these conflicting aspects can be managed and partly resolved through, for example, the reuse of trusted components, and the use of model-based rapid application development and iterative (agile) development.

Modular and Hierarchical Learning and Representation of Large Software

Supervisor: Dr Hana Chockler

A software engineer joining a development team typically does not start writing software immediately; first, she needs to understand the large existing body of code and recognise the key components and how they interact with each other. Documentation is typically sparse and not updated regularly. The software, on the other hand, is large and difficult to understand. “Software is like entropy: It is difficult to grasp, weighs nothing, and always increases.” (Norman Augustine).

Being able to represent large software in a graphical way with the ability to zoom in and out of components would help tremendously towards understanding of the software structure and its functionality. The proposed project is to learn a hierarchical compositional structure that will be used for such a graphical representation. The most likely candidate for such a hierarchical structure is state charts which have been used for software design for many years.

There are no existing learning algorithms for learning state charts. There is, however, a number of algorithms for learning similar structures, such as finite automata, transducers, etc. The first part of the proposed project consists of constructing a new learning algorithm for learning state charts. The second part of the proposed project is using this algorithm to learn complex software in a hierarchical way, allowing the user to zoom in and out of components (composite states).

The project will build upon previous research results of Dr Hana Chockler in collaboration with the University of Oxford.

Monitoring Compliance with Dynamic Security Policies under Uncertainty

Supervisors: Dr Natalia Criado & Dr Jose Such

This project will develop the first monitor capable of checking compliance with dynamic and adaptable security policies on the basis of incomplete and uncertain observations. Most existing proposals on security policy compliance monitoring assume that monitors have perfect information and observation capabilities, or that security policies are fixed and known at design time. However, this assumption is too strong for modern hyper-connected, socio-technical, and cyber-physical systems due to their inherent uncertainty, incompleteness and dynamism.

For example, in a business environment it is: unfeasible to observe all files uploaded/downloaded to/from public cloud services, since employees can perform these actions using non-corporate network/devices; impossible to detect all sensitive information contained in files with complete certainty; and security policies controlling access to public cloud services can change as a result of new legislation (GDPR) or threats. This project, will propose a novel security policy monitor for hyper-connected, socio- technical, and cyber-physical systems.

Multiple Robots Performing Random Walks

Supervisor: Dr Frederik Mallmann-Trenn

The goal of the project is to study distributed algorithms for dynamic and noisy settings robot swarms, and biological systems. We will seek new algorithms to solve fundamental problems of communication, construction, reaching agreement, estimation, data processing, searching, shape formation, task allocation, and more.

One example is the setting of [1], where robots have to estimate the fraction of black tiles in a grid. Each of the robots is very simple and performs a random walk. Whenever, two or more robots are close to each other, they can communicate with each other. In the end, the robots have to agree on a joint estimate of the fraction of black tiles. The arising questions here are:

- How much can multiple random walks speed up the process?
- How many samples have to be taken?
- What happens if the communication is noisy?

The goal is also to collaborate with researchers in the robotics community by modelling and analyzing systems theoretically. In addition to a solid understanding of Markov chains, students should be interested in collaborating with researchers across different disciplines.

[1] https://dl.acm.org/citation.cfm?id=3237953

Network Optimisation Algorithms

Supervisors: Professor Tomasz Radzik & Dr Kathleen Steinhofel

Network Optimisation problems are computational problems with input data referring to a network structure. Such problems occur in computer science, operations research, engineering, and applied mathematics. From the computer science point of view, the general objective of studying network optimisation problems is to develop efficient algorithms, which provide strict performance guarantees. This project will focus on algorithms for network optimisation problems with the dynamic network structure, which changes over time. One of the applications is to provide efficient routing in networks where individual node-to-node links are not always available.

The Nexus between Crime, Mental Wellbeing and the Built Environment in Urban Areas

Supervisors: Dr Nishanth Sastry, Dr Rita Borgo & Dr Andrea Mechelli

This project will explore the nexus between crime, mental wellbeing and the built environment in urban areas, using London as a case study. Using a data-driven approach, the student will develop a holistic understanding of the spatial and temporal dynamics of crime. For instance, objective notions of crime such as real-time crime reports from Metropolitan Police can be compared with more subjective notions of how safe a place "feels", measured using crowdsourcing using the UrbanMind App. Spatial variation in crime levels and Temporal Dynamics (night vs. day or weekday vs. weekend) will be mapped. Machine learning on images from Google Street View, Flickr etc can shed light on how the built environment affects perceived notions of safety and whether it has an effect on actual crime, as hypothesised by the "Broken Window" theory. The results will be used to inform future work on urban wellbeing, as well as urban planning.

A Novel Model-driven AI Paradigm for Intrusion Detection

Supervisor: Dr Fabio Pierazzi

Intrusion Detection Systems (IDSs) are commonly deployed in networks and hosts to identify malicious activities representing misuse of computer systems. The numbers and types of attacks have been constantly increasing, and detection based on manually-defined signature is no longer a viable option. Hence, AI-powered IDS solutions have been explored to keep up the arms race and scale to new threats, but they are not yet deployed at scale in companies; this is mostly because such AI-powered systems cannot be trusted and are not interpretable [1], and they suffer from a lot of false positives preventing their applicability in real-world scenarios. In particular, a major limitation is that most existing solutions for AI-powered IDSs are data-driven, where the relationships learned from the data are often artifacts or domain-agnostic, and thus harder to trust and interpret even for network administrators.

This project aims to explore the design of a novel model-driven AI paradigm for intrusion detection, where expert knowledge is embedded in a model to characterize user behaviors (e.g., through formal logic [2]), with the purpose of identifying malicious activities with trust, interpretability and verifiability of the IDS decisions, in particular when deployed to real-world contexts. In other words, this project aims to advance the state-of-the-art in AI-powered IDSs by integrating expert knowledge in the models to achieve trust, interpretability and verifiability of decisions. This will increase the overall safety of protected users by making IDS systems more effective and reliable, and

progress towards industry-wide deployment of AI-based solutions for intrusion detection.

[1] R. Sommer and V. Paxson, "Outside the closed world: On using machine learning for network intrusion detection." *IEEE Symp. Security and* Privacy, 2010.

[2] S. Jajodia, N. Park, F. Pierazzi, A. Pugliese, E.Serra, G.I. Simari, V.S. Subrahmanian. A probabilistic logic of cyber deception. *IEEE Transactions on Information Forensics and Security*, 2017.

Participatory Agent-Based Modelling of Emergency Department Patient Flow

Supervisors: Dr Steffen Zschaler & Dr Simon Miles

Emergency health care is in crisis; the core "4-hour" KPI has not been met since 2015. Emergency

Departments (EDs) are socio-technical systems with complex interactions between a wide range of actors and with their urban environment. To help predict how changes in practice will affect the 4-hour KPI while ensuring patient safety and quality of care, we have been developing agent-based models (ABMs) of EDs, which can provide explainable analyses of behaviour of complex systems emerging from the lower-level interaction of large numbers of agents. However, ABMs currently are implemented in Java or C++, making them too technical to be understood and manipulated by clinical decision makers. Hence, findings from ABM-based analyses are often not translated into interventions. In this PhD project, you will explore how using domain-specific languages (DSLs) closely aligned with clinical staffs' conceptualisation of the ED environment will affect acceptance of ABM. We collaborate with King's College Hospital ED and Westminster City Council.

Personalised Medicine

Supervisors: Professor Costas Iliopoulos & Dr Sophia Tsoka

The explosion of human genomic data is a key driver of the current transition in healthcare to an era of personalised medicine. The correct assembly and subsequent analysis of this data is, therefore, crucial.

Algorithms can be designed to provide answers for the various and vast number of specific questions that collectively elucidate the (dis)functioning of a biological entity, at the most fundamental level. These algorithms can be categorised based on their purposes. For example, pattern matching/discovery algorithms can find biologically significant motifs in sequences; alignment algorithms can identify the similarity between sequences; and compression algorithms can allow the latter two problems to be solved in a more time- and space-efficient manner.

Predictive Visual Analytics for Urban Contingency Planning

Supervisors: Dr Rita Borgo & Dr Grigorios Loukides

The aim of this project is to investigate the power of integrating predictive analytics and data visualization to address the challenge of generation, validation and deployment of contingency plans in the context of urban related scenarios.

Based on initial work already conducted in this area by both supervisors, the project will investigate development and evaluation of:

- algorithms for mining city rich data, both static (stored by the London Councils) and realtime (retrieved from mobile platforms, remote sensors, and social media), in order to predict emergencies efficiently and accurately;
- novel visual encodings to enhance the diagnostic and predictive capabilities of mining algorithms, through their integration within a flexible visual analytics system capable of supporting and leveraging domain expert knowledge.

Application domain: contingency plans are employed across different organizations, from government to businesses, to minimise risk of catastrophic impacts of unexpected events. Research will have impact beyond the city remit.

PRISM: Adversarial Program Generation

Supervisor: Professor Lorenzo Cavallaro

Machine learning algorithms have a major weakness: they are susceptible to adversarial attacks [1], where the classifier is induced into making wrong decisions by an attacker that alters data at training-time (poisoning) or at test-time (evasion). In malware detection, this implies that an adversarial malware object is recognized (wrongly) as goodware by the classifier, and bypasses antimalware solutions. Most research efforts on adversarial ML have focused on the image, speech, and audio domains, where it is trivial to change raw data (e.g. pixels); the malware domain is much more challenging because it needs to change programs code while preserving their functionality. Hence, there is no clear understanding of adversarial attacks in the malware domain.

This proposal aims to study which adversarial attacks for malware are realistic, i.e., to determine which adversarially-generated feature vectors can be transformed into real and functioning malware applications. Understanding what malware adversarial ML attack is realistic becomes fundamental to design appropriate defenses in this domain, and building more robust classifiers.

The proposal requires working at the intersection of program analysis and machine learning for systems security and the successful candidate is required to have a solid programming and CS/CEng background with a particular passion and knowledge in the foundations of systems, program analysis and machine learning. Given the nature of the research, the proposal links perfectly with the aims of the Security Hub at King's.

[1] N. Carlini and D. Wagner. Towards evaluating the robustness of neural networks. IEEE S&P, 2017.

[2] A. Demontis, M. Melis, B. Biggio, D. Maiorca, D. Arp, K. Rieck, I. Corona, G. Giacinto, and F. Roli. Yes, machine learning can be more secure! A case study on android malware detection. IEEE TDSC, 2017.

[3] K. Grosse, N. Papernot, P. Manoharan, M. Backes, and P. McDaniel. Adversarial examples for malware detection. In ESORICS. Springer, 2017.

Privacy in the Internet of Things

Supervisor: Professor Maribel Fernandez

Data Collection policies are used to restrict the kind of data transmitted by devices in the Internet of Things (e.g., health trackers, smart electricity meters, etc.) according to the privacy preferences of the user. The goal of this project is to develop cloud/IoT architectures with integrated data collection and data sharing models, to allow users to specify their own policies and trade data for services. For this, new data collection and data sharing models will have to be developed, with appropriate user interfaces, policy languages, and policy enforcement mechanisms. An important aspect of the project is the development of policy recommendation systems that can suggest/create policies based on user profiles, making privacy an integral part of the system (according to the

privacy-by-design” IoT paradigm).

Programming as an HCI Challenge - IDE Interaction Design

Supervisors: Professor Michael Rolling & Dr Neil Brown

Frame-based editing with Stride [1] was a first attempt to revisit the design of program editing from an HCI perspective, in the context of novice programmers. What would it look like to approach professional IDEs from this perspective?

This project would take a Stride-like approach to professional tools and design, build and evaluate a new, better editor.

[1] https://www.greenfoot.org/frames/

Program editor design for accessibility

Supervisors: Professor Michael Kolling & Dr Neil Brown

Most program editors use text for editing. Screen readers can be used with text-based editing by visually-impaired programmers, but the syntax can often be confusing. Whitespace and punctuation are highly significant in program text but often omitted or have poor interaction with screen readers. Block-based editors rely less on syntax, so are potentially more suitable for accessible programming - but blocks are often manipulated only through drag-and-drop interactions which are ill-suited to visually-impaired users. Our existing Stride editor combines keyboard interactions with structural programming, but does not yet have support for accessibility tools. This project would look at improving the Stride editor to work well with accessibility, especially for vision-impaired users, including the design, implementation and evaluation of the editor with actual users.

Programming history for learning and reflection

Supervisors: Professor Michael Kolling & Dr Neil Brown

Version control provides a way to store and view the history of program code. This is generally considered an advanced tool, used for collaborating or once a programmer is working on a large code base. This project would investigate the implications of using built-in automatic version control. Can this help during novice program development, can it help students in reflecting on their learning progress, and could it be used to provide more accurate programming assessment. This would involve the design, development and multiple evaluations of automatic version-control in a beginner's IDE.

Security and Safety of Cyber-Physical Systems

Supervisor: Professor Luca Vigano

Cyber-Physical Systems (CPSs) are integrations of networking and distributed computing systems with physical processes and associated instrumentation that monitor and control entities in the physical environment, with feedback loops where physical processes affect computations and vice versa. Emerging applications of CPS include all the essential pieces of our social infrastructure: telecoms, banking, manufacturing, health energy, transportation, government smart cities. CPSs have effectively become one of the driving factors of the so-called fourth industrial revolution (Industry 4.0), but all the new opportunities opened by CPSs will only materialize if we can ensure their security and safety.

However, this need is often not addressed in current practice because of the major challenges that are posed by the heterogeneous and distributed nature of the systems and their interaction with the physical world and with the human users. As a consequence, there has been a dramatic increase in the number of attacks, e.g., influencing physical processes to bring the system into an undesired state. System failure can be extremely costly and threaten not only the system's environment but also human life.

The main aim of this PhD project will be to develop model-based AI techniques for representing, analysing and reasoning about the security and safety of both the technical components of a CPS (control, computation, communication) and its social components (e.g., user interaction processes and user behavior) together and at the same time. The goal will be to overcome the limits of the state-of-the-art to devise methodologies and technologies for the formal validation of properties of CPSs to include the human element together with the technical in a holistic, socio-technical approach for security and safety, and to rebound the findings over the users through behavior change techniques. This will greatly simplify the design, development, deployment, and management of socio-technically secure CPSs, and thus have a disruptive and lasting impact.

Software Verification and Nominal Dependent Type Theory

Supervisor: Professor Maribel Fernandez

Dependent Type Theory is a mathematical tool to write formal specifications and prove the correctness of software implementations. The proof assistants used to certify the correctness of programs (such as Coq), are based on dependently-typed higher-order abstract syntax. The goal of this project is to explore alternative foundations for proof assistants using nominal techniques. The nominal approach has roots in set theory and has been successfully used to specify programming languages. This project will focus on the combination of dependent types and nominal syntax, and explore the connections between the nominal approach and the higher-order syntax approach used in current proof assistants.

String Sanitisation with Applications to Internet of Things Data

Supervisors: Dr Grigorios Loukides, Professor Costas Iliopoulos, Professor Luca Vigano

The overall aim of the project is to develop and evaluate a robust and efficient approach that allows organisations and businesses to protect the privacy of data represented as strings. The project will consider the protection of aggregated data (event sequences), as well as string databases, and it will also address the interrelated issues of usefulness, security, and scalability. It aims to develop a methodology (model, algorithms, protocols) for sanitising (i.e., transforming) data that is: (I) privacy-preserving, by designing and applying a privacy model along with algorithms for sanitising string data. (II) Utility-preserving, by designing measures and tools for quantifying the level of usefulness of data that must be traded-off for achieving privacy. (III) Secure and scalable, by designing efficient protocols that allow multiple parties to securely and jointly protect their data. The methodology will be evaluated on data from the Internet of Things (IoT) domain.

Temporal and Resource Controllability of Workflows of Autonomous Systems

Supervisor: Professor Luca Vigano

Workflow technology has long been employed for the modeling, validation and execution of business processes, and will play a crucial role in the design, development and maintenance of future autonomous systems. A workflow is a formal description of a business process in which single atomic work units (tasks), organized in a partial order, are assigned to processing entities (agents) in order to achieve some business goal(s). Workflows can also employ workflow paths in order (not) to execute a subset of tasks. A workflow management system coordinates the execution of tasks that are part of workflow instances such that all relevant constraints are eventually satisfied.

Temporal workflows specify business processes subject to temporal constraints such as controllable or uncontrollable durations, delays and deadlines. The choice of a workflow path may be controllable or not, considered either in isolation or in combination with uncontrollable durations. Access controlled workflows specify workflows in which users are authorized for task executions and authorization constraints say which users remain authorized to execute which tasks depending on who did what. Access controlled workflows may consider workflow paths too other than the uncertain availability of resources. When either a task duration or the choice of the workflow path to take or the availability of a user is out of control, we need to verify that the workflow can be executed by verifying all constraints for any possible combination of behaviors arising from the uncontrollable parts. Indeed, users might be absent before starting the execution (static resiliency), they can also become so during execution (decremental resiliency) or they can come and go throughout the execution (dynamic resiliency).

Temporal access controlled workflows merge the two previous formalisms by considering several kinds of uncontrollable parts simultaneously. Authorization constraints may be extended to support conditional and temporal features.
This PhD project will aim to ensure the safety and trust of autonomous systems by reasoning about the temporal and resource controllability under uncertainty of the workflows that govern them.

Towards Protection of Users in Online Social Networks

Supervisor: Dr Guillermo Suarez de Tangil

Methods currently used to detect unwanted content in Online Social Networks (OSN) suffer from a number of limitations. First, they are prone to produce unfair decision dominated by the skewed population they are modelled with. Second, algorithms are unable to explain why certain content has been flagged as unwanted.

Active adversaries are currently exploiting this flaws to evade the detection mechanisms placed in current OSN. This applies to several application domains such as: i) the use of spear phishing or malware attacks to enable cyber-dependent crime, ii) the use of coordinated harassment campaign to deliver harmful or deceiving content (e.g., politically-biased memes), or iii) the use of commercially-driven sensational content to deliver fake news.

The purpose of this PhD is to devise new disciplines aiming at protecting users, and specially minors, from malicious actions in OSN and understanding novel threats. The scope of the project will focus on studying the problem of online aggression from a broad perspective.

Tracing Trust - Visual Frameworks for Explainable AI

Supervisors: Dr Rita Borgo, Dr Daniele Magazzeni, Dr Alfie Abdul-Rahman

Explainable Artificial Intelligence (XAI) is a topic receiving close review and increasing interest across different fields. Crucial to explainability is understanding of cause-effect relationships which in complex intelligent systems are anything but clear. Lack of ability to present the rationale behind a decision making process inevitably mines trust and introduces uncertainty with respect to accountability of consequences.

The proposed research program will focus on the creation of a theoretical and applied framework to support the creation of systems to help people interpret the reasoning behind decisions made by AI systems. The project will entail design, implementation, and testing of visualization interfaces connecting to and integrating with explainable intelligent systems designed by partners.

This project places itself across three different fields: visual analytics, human-computer interaction, and artificial intelligence.

Hub relevance: Autonomous Systems

The Undergound Economy: Understanding and Modelling Misuse in the Darkest Corners of the Web

Supervisor: Dr Guillermo Suarez de Tangil

Underground markets play a key role in the proliferation of cybercrime. Users with few technical skills can easily acquire services and tools in the darknet to set up their own criminal operations. An example is illicit crypto-mining campaigns, that use botnets rented in such markets to leverage stolen resources and covertly mine 4.4% of the entire Monero in circulation (circa 58 million USD) [1]. Profits generated by these campaigns introduce massive incomes to cyber-criminals. These incomes fuel the underground economy and gear other cyber-criminal activities. More importantly, these threats generally cause important economical loses to victims.

The purpose of this PhD is to develop data driven approaches to better understand how these communities are structured and the type of crimes they support.

[1] A First Look at the Crypto-Mining Malware Ecosystem: A Decade of Unrestricted Wealth. Sergio Pastrana and Guillermo Suarez-Tangil. ACM Internet Measurement Conference (IMC). Oct 21-23, 2019. Amsterdam, Netherlands

Understanding Cyber-Dependent Crimes that are enabled by Malware from a Software Development Perspective

Supervisors: Dr Guillermo Suarez de Tangil & Professor Luca Vigano

The goal of this thesis is to better understand cyber-dependent crimes that are enabled by malware from a software development perspective. The purpose is threefold: a) to profile malware developers, b) to understand their business model, and c) to measure the impact of malware trading in underground markets and surface forums. Throughout this Thesis, the PhD candidate will learn how to reverse engineer malicious code and feed this information to different machine learning algorithms. The candidate will also be conducting malware-related measurements in underground markets and darknet forums. The qualifications obtained by the candidate will be relevant to different stakeholders such as: i) low-enforcement and anti-crime agencies when designing strategies to prosecute these actors, ii) incident response teams and forensic analysts to make informed decisions when a malware is discovered, and iii) national advisory centres to understand novel infection vectors or CERTs to design both mitigation and early detection strategies.

Unstructured Big Data

Supervisors: Professor Costas Iliopoulos & Dr Grigorios Loukides

A major challenge in today's society is the explosive growth of unstructured data such as text, images, videos and speech data. These forms of data exhibit the three characteristics of velocity, volume and variety that make processing and comprehending them a challenging task.

The initial processing of this data is invariably done using automated methods, as manual processing would be prohibitively expensive. The output of this automated processing is uncertain, either due to inaccuracies or inconsistencies in the raw data, or due to the automated processing. The database community has recognised this phenomenon in recent years, and several probabilistic formulations of uncertain data have been proposed, with a focus on processing SQL-like or ranking queries on such data. However, the science of mining, pattern analysis and pattern discovery on uncertain data expressed in probabilistic terms is very much in its infancy.

Mining probabilistic uncertain data to obtain reliable and actionable information is a critical challenge. Since the proliferation of "data science pipelines" uncertainties in one stage can propagate and magnify in later stages. It is essential both that uncertainty is processed appropriately by the system and that the data is not artificially made certain by, for example, choosing the most probable outcome at each stage.

The central hypothesis of this proposal is that the new field of algorithms on uncertain sequences that we propose is an important and broad foundation for representing and mining uncertain data arising in a wide variety of contexts. In addition, novel algorithmic techniques and ideas will be needed and could be useful for other high throughput data processing. The breadth of the proposed area investigation can be illustrated by the three abstract models given below:

A) Probabilistic sequences: they model a number of real-world data, as - DNA sequences, either to represent single nucleotide polymorphisms, or errors introduced by wet-lab sequencing platforms during the process of DNA sequencing.

- Converting sensor readings into meaningful human actions (e.g., accelerometer readings into kinds of human activity, using blood pressure/voice pitch to infer emotions) due to the process' intrinsically uncertainty.
- Software behaviour is often characterised in terms of sequences of events, such as the order of user interactions with a GUI or a web-page, the order of function invocations within a program, or the order in which network packets are sent to a server. A common assumption is that system behaviour is deterministic. It is however easy to envisage situations in which this assumption is violated. Network packets might arrive in a different order, depending on their route through a network, a unit of code might include stochastic behaviour to arise from random number generators, or different interleaving of concurrent processes.

B) Uncertain Event Sequences: arise from a number of sources including measurement error, randomness in the underlying phenomenon, and due to distributed and asynchronous data gathering. They are used in a number of real-world scenarios to model and analyse spatial or temporal data, which is of interest in diverse disciplines as computational neuroscience, earth science and telecommunications. Marked event sequences are even more general and can be applied to computer and economic systems for examples.
Uncertain Time Series: are most naturally associated with measurement errors, but can directly represent a range of variation (e.g. high/low stock prices in a day's trading, confidence intervals for predictions) or deliberate obfuscation for reasons of privacy preservation. They can be seen as special cases of event sequences, but while in uncertain time series the uncertainty lies in the value, in uncertain event sequences, the uncertainty is in the time that the event occurred.

Research Programme and Methodology:

We will focus on highly-scalable methods discovering repetitive structures in uncertain sequences. Given the uncertainty in the underlying data, these repetitions will of necessity be approximate, rather than exact.

There are two major technical obstacles to overcome: firstly, classical measures of approximation (edit or Hamming distance) are inadequate to measure similarity between uncertain sequences. One objective of this project is to define new, alternative, well-founded and powerful approaches for measuring similarity between sequences. Secondly, we need to develop novel algorithmic techniques for solving problems in the context of uncertain sequences.

These problems are directly motivated by bioinformatics applications, such as studying genetic mutations; DNA sequence analysis of antibodies and identification of "hairpins" that occur in DNA sequences in Tuberculosis and HIV virus strains, respectively. However, they are also closely related to pattern discovery tasks that arise in other problem domains. Furthermore, they are also the most intensively studied problems in mining time series data and, to the best of our knowledge, these problems have not been considered in the uncertain time series framework, and it is not at all clear how to extend the known methods to this case. A solution is to build upon the experience we have in musical and biological computation pattern analysis, which share some characteristics with uncertain sequence processing, to suggest lines of attack.
Objectives:

1. Devise appropriate uncertain / probabilistic sequence formulations for modelling large-scale complex heterogeneous data.

2. Develop highly-scalable algorithms for pattern / motif discovery and sequential pattern mining in uncertain sequence data.

3. Build a theoretical framework for pattern discovery in dynamic, streaming and high-throughput uncertain sequence data.

4. Develop robust and well-founded methods for inferring actionable models of uncertain sequence data.

5. Devise appropriate and tractable formal frameworks for modelling stochastic dependencies in uncertain sequence data.