DISCUS Interdisciplinary MSc Projects - 21-22 Academic Year

Supervisor	Department	Proposed project title	Project Description	Co-Supervisor
Spyros Skarvelis-Kazakos	Engineering and Design	Impact Of The COVID-19 Epidemic On Critical Infrastructure Resilience	The COVID-19 coronavirus pandemic is causing staff shortages throughout the industry. Critical infrastructure such as energy, water and telecommunications depends on various staff levels, from higher level management to field operators. Companies have continuity plans, but there is a need to perform detailed modelling of the company's organisational structure and the role of different staff, in order to understand the potential impacts. More importantly, this understanding will allow preventative and mitigating actions to be planned. This project is about building a model of an energy company's organisational structure. This will involve gathering of information on different staff levels / roles and business continuity plans, then building a simple model using the concept of agent-based modelling (Artificial Intelligence). More than one type of company can be modelled (e.g. Distribution Network Operator, Transmission System Operator, Energy Supplier, etc), so multiple students can take on this project.	N/A
Spyros Skarvelis-Kazakos	Engineering and Design	Adaptation Of Water And Energy Networks To Climate Change	Electricity, gas and water networks are separate infrastructure systems, but they depend on each other for stable operation. Water networks have pumps that require electricity to run, many electricity generators burn gas to generate electricity and require water for cooling. In addition, the resilience and reliability of energy and water supply relies on their combined response during rare high impact events. Climate change increases the likelihood of high-impact-low-probability (HILP) events such as a 100-year flood or "the Beast from the East" in 2018, thus reducing resilience. Electricity supply disruption can cascade to water supply disruption, or water supply disruption could ensue directly. The aim of this project is to perform an assessment of climate change expectations, which might bring insight for the future of supply resilience and potential measures that can be taken to avoid disruption. Especially important are ways to utilise the massive amount of data available by climate scientists.	N/A
Luc Berthouze	Informatics	Exploring the science behind cancer diagnostics through network modelling	This project aims at exploring the scientific field behind three types of cancers (lung, prostate and colorectal) through the use of publication data from 1990 to 2015. It builds on a publication dataset of ~370 000 publications. A large number of variables are available, such as co-authorship (between individuals and organisations), publication year, cited references, discipline of the paper/health classification ... The two main areas of interest are to (i) explore how the specific field of diagnostics has evolved within these three cancers and (ii) whether there is any specific area in cancer which is more attractive to companies. Analysis of the data will involve the deployment of text mining/processing techniques, citation and co-citation analysis as well as state of the art methods from network science such as centrality measures and modularity-based community structure detection.	Dr Frederique Bone
Helfrid Hochegger	Life Sciences	Mining cancer gene dependencies to identify and validate novel cell cycle regulators.	Large scale gene dependency screens have delivered a dataset for over 900 cell lines and 14,000 genes. This is a rich resource for data mining and machine learning to identify new functional associations and dependencies. Dr. Hochegger and Dr. Istvan Kiss have been collaborating over the past year to identify new cell cycle regulators using correlation- and cluster-analysis of gene dependency data in this data set. This project aims to further develop this approach. In parallel we aim to functionally validate already predicted cell cycle regulators in various cancer cell lines. The project will have two components, complex network analysis of gene dependent correlations, and analysis of high throughput microscopy data using Python image processing and quantification work flows. The student will be exposed to both data science and cancer cell biology in an interdisciplinary environment. They will become fluent in applying state of the art programming languages (Python, Matlab), set up machine learning algorithms and develop new code to query cancer cell line and tumour genomics data bases. In parallel, the student will learn to analyse high throughput gene depletion screens and will set up automated image segmentation and analysis work flows to detect cell cycle phenotypes. This will be done on an already identified set of approximately 30 novel cell cycle regulators, while novel hits will be generated using more refined ML algorithms. The interdisciplinary outlook of this project provides a good opportunity for the student to learn a broad scope of data science skills for a successful career as a cancer biologist.	Istvan Kiss

Evi Soutoglou	Genome Damage and Stability Center	Analysis of 3D gene positioning during differentiation of Stem cells	Every day the DNA in our cells are subjected to thousands of breaks, which must be repaired to maintain the health the organism as unrepaired or incorrectly repaired DNA breaks can lead to disease, such as cancer. DNA double strand break (DSB) are amongst the most deleterious DNA lesions, since they can lead to chromosomal translocation, which are a leading cause of cancer. To preserve genomic stability, cells have evolved various DNA repair pathways, including homologous recombination (HR), non-homologous (NHEJ) or microhomology-mediated end joining (MMEJ) pathways. The NHEJ and MMEJ pathways have been implicated in the formation of chromosomal translocations. HR and NHEJ/MMEJ repair are compartmentalized into different parts of a mammalian nucleus, suggesting repair pathway choice regulation may in part be based on where in the nucleus the broken DNA is located (Lemaitre et al., 2014, Schep et al., 2021). It remains unknown how the nuclear position of a DNA break impacts the frequency and the nature of chromosomal translocations. To tackle this question, we induce DSBs in specific gene loci in two different cell types, and will identify chromosomal translocation partners (which genomic loci have been incorrectly joined together) using a dedicated sequencing assay (LAM-HTGTS Hu et al., 2016). Specifically, we induce DSBs at loci that are located at the nuclear periphery in mouse embryonic stem cells (ESC), but which are relocate in to the nuclear centre during differentiation into neural precursor cells (NPC). We will then compare the frequency of translocations and any changes in translocation partners between the two cell types to assess the effect nuclear positioning has on chromosomal translocations. We have selected candidate genes (n=20) that have previously been shown to have a differential location in ESC and NPC (Peric-Hupkes et al., 2010, Therizols et al., 2014). As an important first step, the position of these candidate genes must be validated in both cell types, before and after break induction. To assess a genes position, we use 3D fluorescence in situ hybridization to fluorescently label specific individual genes of interest, and immunostain the lamin B protein to demark the nuclear edge. We currently perform manual imaging analysis, on approx. 200 nuclei per gene, per cell type, using ImageJ. We measure the position of the gene loci relative to the nuclear periphery by measuring the distance between the centre of the FISH signal and the lamin B staining. (continues in next row ...)	Anotida Madzvamuse
		Analysis of 3D gene positioning during differentiation of Stem cells (cont)	We would like to develop an automatic image analysis tool for this project, which is capable of: accurate nuclei segmentation, distance measurements between the gene locus and the nuclear periphery, and perform segmentation of nuclei into shells of equal areas (*), and position the gene loci relative to these shells. Overall, the question we aim to answer has significant implications for understanding the mechanisms controlling the formation of chromosomal translocations within the nuclear environment and will aid the understanding of why certain translocations are recurrent in cancer. (*) Useful references associated with FISH analysis: Zaki et al., 2020 DOI: 10.1002/cyto.a.24257 Meaburn and Misteli 2008 The Journal of Cell Biology, Vol. 180, No. 1, January 14, 2008 39–50 http://www.jcb.org/cgi/doi/ JCB 39 10.1083/jcb.200708204 Shachar et al., 2015 doi: 10.1101/sqb.2015.80.027417 Nandy et al., 2009 doi:10.1109/IEMBS.2009.5332922. Nandy et al., 2011 doi:10.1109/IEMBS.2011.6091480. Gudla et al., 2008 DOI: 10.1002/cyto.a.20550 Therizols et al., 2014 DOI: 10.1126/science.1259587	
Sylvia Schroeder	School of Life Sciences	The neural encoding of visual stimuli during different behavioural states	We have recorded the activity of hundreds of neurons in the early visual system of the mouse (retina and superior colliculus) [1]. The data are continuous traces (collected using two-photon imaging), which reflect the activity of each neuron at a temporal resolution of tens to hundreds of milliseconds. During the experiments, the mouse saw gratings moving in different directions, while we recorded the mouse’s behaviour (running speed, eye movements, and pupil size). We want to use adevelop a machine learning model that can predicts the visual stimulus from the activity of the neural population and then ask the following questions:. (a) Is the prediction better when the mouse was running? (b) If we train the model only with data when the mouse was not running, how well accurate is the prediction during periods when the mouse was running? (c) How do the models differ when trained with data during running versus not running differ? [1] Schröder S, Steinmetz NA, Krumin M, et al. Arousal Modulates Retinal Output. Neuron. 2020;107(3):487-495.e9. doi:10.1016/j.neuron.2020.04.026	Luc Berthouze (Data Scientist)

Peter Overbury	Engineering (Phil), Life Science (Kieran) and Industry (Peter)	Automatic identification of bats from video data.	The problem of automating observations of animals is of great importance to almost all fields of ecology (see https://www.microsoft.com/en-us/ai/ai-lab-snow-leopard). Here, there is a need for the processing of 700 hours of footage of bats crossing country roads in the UK in order to determine their flight height. This will allow us to better understand the collision risk posed to bats from vehicles, particularly for the rare British woodland bat, the barbastelle. Further, the automation of this task could allow for wider citizen science projects into this bats life cycle and behaviors, which are still not fully understood. As such, this project has good scope for publication and could even be expanded into identification of other flying object problems such as drones and or publication of this as a data set for other ML/computer vision problems.	Kieran O'Malley (domain expert), Dr Phil Birch, Dr Peter Overbury
Peter Overbury	Informatics & Iproov (industry)	Biometric application of variational diffusion models	Problem: •Investigate the impact of diffusion-based models on biometrics Objectives: •Use variational diffusion models to aid biometric attack detection •Explore the implementation of novel diffusion model techniques to generate •synthetic imagery •Improve model inference times by optimising ML operations Reading •https://arxiv.org/pdf/2107.00630.pdf •https://arxiv.org/pdf/2105.05233.pdf Supervisors: Dr Peter Overbury & Dr Julie Weeds About Iproov Iproov is a World leader in face authentication technology for online identity verification. We are based in London and are happy to support students in person or remote. We are looking for a few students to take up projects along these lines (project above is one example). If you are interested in doing a project with us please send a copy of your CV to edward.crookenden@iproov.com or peter.overbury@iproov.com , along with the projects that you are most interested in. If there’s something else you’re interested in that you think would fit with the work we do at Iproov, please contact us and let us know.	Julie Weeds

Peter Overbury	Informatics & Iproov (industry)	Unsupervised video feature disentanglement	Problem: •How to efficiently reconstruct a 3D scene from a series of sparse-view images Objectives: •Explore a number of new deep learning methods including Neural reflectance surfaces, neural radiance fields and video autoencoders to learn 3D structure and camera trajectory •Current 3D reconstruction methods can be extremely computationally expensive. •Improve model training and inference times without loss of precision Reading •https://zlai0.github.io/VideoAutoencoder/resources/video_autoencoder.pdf •https://arxiv.org/pdf/2110.07604.pdf •https://arxiv.org/pdf/2003.08934.pdf Supervisors: Dr Peter Overbury & Dr Julie Weeds About Iproov Iproov is a World leader in face authentication technology for online identity verification. We are based in London and are happy to support students in person or remote. We are looking for a few students to take up projects along these lines (project above is one example). If you are interested in doing a project with us please send a copy of your CV to edward.crookenden@iproov.com or peter.overbury@iproov.com , along with the projects that you are most interested in. If there's something else you're interested in that you think would fit with the work we do at Iproov, please contact us and let us know.	Julie Weeds
Peter Overbury	Informatics & Iproov (industry)	Unsupervised representation learning of structured data	Problem: • Detecting anomalous samples of structured data Objectives: • Research effectiveness of unsupervised representation learning with our large structured dataset • Explore learnt embedding space for useful features • Anomaly detection strategies using the generated feature embeddings Reading ● https://assets.amazon.science/60/53/7b0e54fb4ee0bbcba20dc0c5348a/record2vec-unsupervised-representation-learning-for-structured-records.pdf Supervisors: Dr Peter Overbury & Dr Julie Weeds About Iproov Iproov is a World leader in face authentication technology for online identity verification. We are based in London and are happy to support students in person or remote. We are looking for a few students to take up projects along these lines (project above is one example). If you are interested in doing a project with us please send a copy of your CV to edward.crookenden@iproov.com or peter.overbury@iproov.com , along with the projects that you are most interested in. If there's something else you're interested in that you think would fit with the work we do at Iproov, please contact us and let us know.	Julie Weeds

Peter Overbury	Informatics & Iproov (industry)	Generative facial modelling	Problem: • Investigating how machine learning could be used to commit identity fraud Objectives: • Explore state-of-the-art methods in deep learning for reconstructing high-fidelity facial imagery • Investigate computer vision techniques and develop statistical algorithms for creating and mimicking short video captures of real humans Reading https://arxiv.org/abs/2106.12423 https://arxiv.org/pdf/2112.00532.pdf https://arxiv.org/pdf/2007.03898.pdf Supervisors: Dr Peter Overbury & Dr Julie Weeds About Iproov Iproov is a World leader in face authentication technology for online identity verification. We are based in London and are happy to support students in person or remote. We are looking for a few students to take up projects along these lines (project above is one example). If you are interested in doing a project with us please send a copy of your CV to edward.crookenden@iproov.com or peter.overbury@iproov.com , along with the projects that you are most interested in. If there's something else you're interested in that you think would fit with the work we do at Iproov, please contact us and let us know.	Julie Weeds
Peter Overbury	Informatics & Iproov (industry)	Time series forecasting of video data	Problem: •Explore using spatio-temporal machine learning methods to generate unseen video frame(s) based on seen video frames. This could be: ogiven a sequence of previous frames predict the next frame ogiven a single frame generate a realistic video of many frames ogiven a small subset of frames from an original video, predict the missing frames Objectives: •Explore the field and get a good understanding of the current solutions to the above problem. Understand the challenges of video prediction with our dataset vs those used in papers •Implement one of these solutions yourself and improve upon it or come up with your own solution altogether •Train on our dataset with the aim of achieving accurate frame prediction / convincing video generation Reading •https://arxiv.org/pdf/2004.05214.pdf Supervisors: Dr Peter Overbury & Dr Julie Weeds About Iproov Iproov is a World leader in face authentication technology for online identity verification. We are based in London and are happy to support students in person or remote. We are looking for a few students to take up projects along these lines (project above is one example). If you are interested in doing a project with us please send a copy of your CV to edward.crookenden@iproov.com or peter.overbury@iproov.com , along with the projects that you are most interested in. If there's something else you're interested in that you think would fit with the work we do at Iproov, please contact us and let us know.	

Peter Overbury	Informatics & Iproov (industry)	Gabor filters for improving DNN robustness to adversarial attacks	Deep Convolutional Neural Networks (DNN) are often vulnerable to adversarial attacks and this is a growing threat in the real world of face authentication. Research has shown that bio-inspired Gabor filters can be used to improve DNN's robustness to these kinds of attacks by making them less reliant on overly specific details in the data. Problem: ● Investigate the impact of adding Gabor filters to the training of facial matching DNN both on the overall performance of the DNN and also the effect on their robustness to adversarial attacks Reading https://www.biorxiv.org/content/10.1101/2021.02.18.431827v2 Supervisors: Dr Peter Overbury & Dr Ben Evans About Iproov Iproov is a World leader in face authentication technology for online identity verification. We are based in London and are happy to support students in person or remote. We are looking for a few students to take up projects along these lines (project above is one example). If you are interested in doing a project with us please send a copy of your CV to edward.crookenden@iproov.com or peter.overbury@iproov.com , along with the projects that you are most interested in. If there's something else you're interested in that you think would fit with the work we do at Iproov, please contact us and let us know.	Ben Evans
Kate Shaw	Physics	Machine Learning Visualisation with ATLAS data from the Large Hadron Collider	The Large Hadron Collider (LHC) at CERN [1] accelerates particles near to the speed of light before colliding them millions of times a second inside massive particle detectors such as the ATLAS detector, to study fundamental particles of the Universe. Cutting-edge machine learning techniques are used to analyse the huge amount of collision data collected by ATLAS to study rare fundamental particles such as the Higgs boson and search for new physics such as dark matter [2]. This project, working with ATLAS physicists and data science experts, will develop interactive data visualisation tools for non-specialist users to understand and analyse data using machine learning techniques to search for and discover new rare fundamental particles. This is part of the ATLAS Open Data project [3] led by Sussex researchers [4], which provides ATLAS data, tools and software frameworks to the public [5]. This project will produce the first ever tool allowing the non-specialist to visualise machine learning on LHC data. This project is suitable for students with an interest in data visualisation, machine learning, and big data. Some knowledge of Java, machine learning and Python or C++ will be needed. [1] https://home.cern/ [2] https://iml.web.cern.ch/ [3] https://atlas.cern/resources/opendata [4] https://home.cern/news/news/knowledge-sharing/atlas-releases-13-tev-open-data-science-education [5] http://opendata.atlas.cern/release/2020/documentation/visualization/histogram-analyser-2_13TeV.html	George Parisis

Dr Andrew Penn	Life Sciences	Learning descriptive feature sets for accurate classification of spontaneous synaptic events	step to understanding how neurons encode and process information. Dr Penn’s lab recently developed and released the first version of some open source software, Eventer [1,2], as a machine learning solution for the challenge of classifying candidate spontaneous synaptic events measured by electrophysiology or imaging and detected by deconvolution [3]. The accurate detection and classification of spontaneous synaptic activity is particularly difficult when events occur frequently and their waveforms are overlapping; a problem often associated with in vivo recordings. Currently, training the software requires manual classification and uses a set of hand-crafted features extracted from a training data set as input to a Random Forests algorithm. The software serves as a framework to use machine learning as solution for improving consistency and automating the correct identification of synaptic events. The purpose of this MSc project is to develop a data-driven framework for creating sets of general-purpose descriptive features that enable accurate classification of candidate synaptic events. These features will be used in downstream machine learning models and could be used to make an improved version of the Eventer tool. The types of features to consider could include convolutional filters of the raw signal or spectrogram. These features will be learned from example datasets using either a fully-supervised or semi-supervised learning paradigm. The resulting feature sets will be used to train and evaluate classifiers on a range of challenging data sets from our lab and from online data archives (e.g. DANDI). The existing toolkit is in MATLAB, and developments would need to be compatible with this. However, Python frameworks (e.g. PyTorch and TensorFlow) could still be used for learning the features, which could then be translated into MATLAB. Bibliography: [1] Winchester, G., Liu, S., Steele, O.G., Aziz, W. and Penn, A.C. (2020) Eventer. Software for the detection of spontaneous synaptic events measured by electrophysiology or imaging. http://doi.org/10.5281/zenodo.3991677 [2] Steele, O.G., Liu, S., Winchester, G., Aziz, W., Chagas, A. and Penn, A. Eventer: Software you can train to detect spontaneous synaptic responses for you (TP001324) in BNA 2021 Festival of Neuroscience Poster abstracts. (2021). Brain and Neuroscience Advances. https://doi.org/10.1177/23982128211035062 [3] Pernía-Andrade, A.J., Goswami, S.P., Stickler Y., Fröbe, U., Schlögl, Alois, and Jonas, Peter (2012) A Deconvolution-Based Method with High Sensitivity and Temporal Resolution for Detection of Spontaneous Synaptic Currents In Vitro and In Vivo. Biophys J 103, 1429–1439.	Dr Ivor Simpson
Jamie Ward	Psychology and Sussex Neuroscience	Developing a connectome biomarker for synaesthesia	Magnetic resonance imaging (MRI) has been demonstrated as a powerful and flexible technique for understanding the human brain. In recent years large scale studies, such as the Human Connectome Project (HCP)[1], have identified acquisition protocols to allow consistent high-quality data collection and preprocessing [2][3]. This preprocessing creates a set of measurements, which describe individual anatomical and functional properties. This project seeks to develop a framework for identifying biomarkers [4], patterns of measurements that are indicative of a particular condition, from pre-processed HCP data. Specifically, this project will explore how different machine learning analysis strategies can be used to extract biomarkers, with the added goal of identifying approaches that robustly detect interpretable differences between populations. A student on this project will have access to HCP data from people with synaesthesia (unusual experiences such as music triggering vision) which could be compared to normative samples or clinical data from other sources. The student could either run a pre-planned analysis or search the literature to develop an alternative biomarker. References: [1] https://www.humanconnectome.org/ [2] Glasser, Matthew F., et al. "The minimal preprocessing pipelines for the Human Connectome Project." Neuroimage 80 (2013): 105-124. [3] Glasser, Matthew F., et al. "A multi-modal parcellation of human cerebral cortex." Nature 536.7615 (2016): 171-178 [4] Woo, Choong-Wan, et al. "Building better biomarkers: brain models in translational neuroimaging." Nature neuroscience 20.3 (2017): 365.	Ivor Simpson

Helfrid Hochegger	Life Sciences	Analysing cell microscopy images with machine learning	Quantifying cell properties from microscopy images is an integral part of experimental cell biology. Recent work has demonstrated the potential of multi-channel imaging for measuring progression through the cell-cycle [1]. Such data can be used to quantify differences between different cell-lines, which could be used to measure the effects of genetic changes. This project has two directions that could be investigated: Improvements to cell segmentation models using semi-supervised learning: Recent works have created general purpose machine learning approaches for labelled the pixels of cells [2,3]. However, due to the diversity of appearance and cells overlapping these may fail in some situations. One route this project could take is leveraging a small quantity of labelled images of the target cells and many unlabelled images (or videos) to improve the segmentation performance [4, 5]. Learning the statistical relationships between the different image channel information through a generative model, such as a variational autoencoder [6]. This may lead to a representation that enables more reliable classification of cell-cycle stages. It also enables analysis of how observing the cell in one image channel can predict the other channels [7], which may enable subsequent optimisation of the acquisition process. Skills: Python, Machine Learning, Interests in computer vision and cell biology helpful References: [1] Zerjatke, Thomas, et al. "Quantitative cell cycle analysis based on an endogenous all-in-one reporter for cell tracking and classification." Cell reports 19.9 (2017): 1953-1966. [2] Stringer, Carsen, et al. "Cellpose: a generalist algorithm for cellular segmentation." Nature Methods 18.1 (2021): 100-106. [3] Schmidt, Uwe, et al. "Cell detection with star-convex polygons." MICCAI 2018 [2] Bortsova, Gerda, et al. "Semi-supervised medical image segmentation via learning consistency under transformations." International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, Cham, 2019. [4] https://github.com/xiaomengyc/Few-Shot-Semantic-Segmentation-Papers [5] Kingma, Diederik P., and Max Welling. "Auto-encoding variational bayes." ICLR 2013 [6] Ivanov, Oleg, Michael Figurnov, and Dmitry Vetrov. "Variational autoencoder with arbitrary conditioning." ICLR 2019	Ivor Simpson
Juan Totero Gongora	Physics	Machine Learning for Controlling Ultrafast Lasers	Description: Optical frequency combs are an advanced class of lasers emitting ultra-precise pulses of light [1]. They are ideal candidates to provide the fast-beating “optical heart” required by transformative technologies such as portable atomic clocks, highly-sensitive hazardous chemical detectors, wearable devices for high-precision medical diagnostics, and computer chips operating at photonics speeds. Despite recent technological breakthroughs, frequency combs remain surprisingly hard to control and stabilise at the high emission powers required by real-life applications. This is because existing linear analysis techniques are poorly suited to modeling highly nonlinear states, which limits our access to the extensive range of potential emission regimes. Recent results have demonstrated the vast potential of Machine Learning (ML) in stabilising lasers, delivering improved emission performance (e.g. higher pulse intensity) in a fraction of the time required by standard techniques [2]. In the case of a frequency comb laser, for example, a deep-learning model could iteratively learn to “predict” which combination of input parameters can improve the laser emission. Driving and maintaining a micro-comb laser into an arbitrary high-energy state, however, is elusive and requires a more advanced approach merging AI predictions with a precise understanding of the system’s internal nonlinear dynamics, not necessarily known a priori. This task eludes standard techniques and requires developing an entirely new conceptual approach to tackle real-life laser dynamics. This interdisciplinary project aims to overcome this conceptual gap. You will investigate extending an existing approach [3] for characterising and controlling a real-life ultrafast laser using variational auto encoders and LSTMs trained on simulated laser outputs. This project follows on from a successful MSc dissertation in this area, with several areas of interest to explore depending on the students interests. References: [1] H. Bao et al., ‘Laser cavity-soliton microcombs’, Nature Photonics, p. 1, Mar. 2019, doi: 10.1038/s41566-019-0379-5. [2] G. Genty et al., ‘Machine learning and applications in ultrafast photonics’, Nature Photonics, pp. 1–11, Nov. 2020, doi: 10.1038/s41566-020-00716-4. [3] Baumeister, Thomas, Steven L. Brunton, and J. Nathan Kutz. "Deep learning and model predictive control for self-tuning mode-locked lasers." JOSA B 35.3 (2018): 617-626. Video related to this paper: https://youtu.be/b4wZyAh99wM	Ivor Simpson

Juan Totero Gongora	Physics	Machine Learning acceleration of hyperspectral imaging at terahertz frequencies	Text Description Terahertz is a novel form of electromagnetic radiation with a frequency range lying between microwave and radio waves. In recent years, terahertz science has attracted sizeable research efforts, due to a large number of potential applications across biology, material characterisation, security, and industrial diagnostics . On one side, many common materials (plastic, paper, fabrics) are transparent to Terahertz waves. At the same time, terahertz waves carry only an infinitesimal amount of energy and are therefore much safer than highly-ionizing X-rays when inspecting the interior of a sample. Besides, a wide range of complex materials and compounds (for example, chemical and pharmaceutical substances or explosives) exhibit a unique and very distinguishable response when illuminated with short pulses of terahertz light. The ability to see inside objects while discriminating their material composition is at the heart of current research on the development of advanced imaging devices based on ultrashort terahertz pulses. Despite recent technological breakthroughs, further developments in this research area are limited by the availability of terahertz cameras and imaging sensors. To tackle this limitation, the Emergent Photonics (EPic) Lab at Sussex is focusing on developing single-pixel imaging approaches, also known as computational imaging or ghost-imaging techniques [1]–[3]. In these approaches, rather than employing a sensor composed of a large number of pixels, the sample is illuminated with a series of known spatial patterns, and the transmitted terahertz waves corresponding to each pattern are sequentially acquired by a single-pixel detector. By combining the spatial information of the incident patterns and their corresponding time-dependent outputs, one can reconstruct the properties of the sample through a numerical inversion process. For more information on our research see [4] However, understanding how to extract the sample information from measurements becomes increasingly challenging when dealing with extremely small objects (i.e., much smaller than the incident wavelength) or samples with a complex three-dimensional structure (e.g., a biological cell sample). On the one hand, the experimental measurements for each incident pattern can become extremely long, leading to unpractical overall imaging times. At the same time, when considering objects much smaller than the incident wavelength, the spatial and temporal properties of the sample become entangled. This multidisciplinary MSc research project aims to bring the power of Machine Learning (ML) to overcome these conceptual and technological gaps. ML has emerged as an ideal tool to "disentangle" complex information in standard imaging, allowing, for example, to significantly reduce the number of patterns required in computational imaging techniques, or to retrieve the image of samples concealed by scattering materials (e.g., fog) [5]–[7]. You will investigate and evaluate the most suitable strategies to characterise a real-life sample in the typical experimental conditions of	Ivor Simpson
Miguel Maravall	Life Sciences	Analysing neuronal activity in two-photon calcium imaging	Description: This project will examine imaging data acquired from two-photon calcium microscopy in mice brains; the experimental setup is described in [1]. The data reflect activity of individual neurons recorded while mice perform a trained sensory-guided behaviour. The ultimate aim is to understand how neurons respond to sensory and behavioural variables as the mouse interacts with its environment. This data has two analysis challenges: firstly the observed image data has a low temporal sampling frequency (~10Hz) and its relationship with neuronal activations is not a straightforward linear function; secondly, neuronal activations are likely to correlate both with sensory stimuli under the control of the experimenter, and potentially with actions of the mouse and other factors. This makes disentangling neuronal activations related to the stimulus challenging without incorporating observations of confounding factors. The project could lead to publishable insights into how groups of neurons collectively predict stimuli and other variables. Depending on the students' interest, this project could investigate two directions of study: Develop a multivariate model for predicting stimuli given the simultaneously imaged signal from sets of neurons. This approach would investigate a combination of temporal feature engineering and/or learning, to independently interpret the signal at each neuron, in combination with a regularised linear classifier to provide interpretability in terms of neuronal contribution. Build an implementation of a variational autoencoder for probabilistic spike inference spikes from calcium imaging data following a similar approach to [2]. Subsequently, the dataset can be re-analysed using the predicted spikes. Skills: Programming (preferably Python), knowledge of statistics and/or machine learning 1. Michael R. Bale, Malamati Bitzidou, Elena Giusto, Paul Kinghorn, Miguel Maravall, Sequence Learning Induces Selectivity to Multiple Task Parameters in Mouse Somatosensory Cortex, Current Biology, Volume 31, Issue 3, 2021. 2. Speiser, Artur, et al. "Fast amortized inference of neural activity from calcium imaging data with variational autoencoders." NeurIPS 2017. General reading on analysing calcium imaging data can be found in: Giovannucci, Andrea, et al. "CaImAn an open source tool for scalable calcium imaging data analysis." Elife 8 (2019): e38173.	Ivor Simpson

Darya Gaysina	Psychology and Sussex Neuroscience	Predictive modelling of depression and anxiety across the life course	What are the factors in childhood and adolescence that can predict the onset and persistence of depression and/or anxiety across the life span? This project will tackle this question by analysing the National Child Development Study (British 1958 birth cohort) dataset [1], which has a large amount of longitudinal data on a cohort of over 17,000 people born in the UK in 1958, e.g. on physical and educational development, economic circumstances, health behaviour and family life. [1] https://cls.ucl.ac.uk/cls-studies/1958-national-child-development-study/	Adam Barrett
David Weir	InformaticsInformatics / Sussex Humanities Lab	Active learning methods for low resource document classification	In a wide variety of NLP applications, the need arises to create one or more bespoke document classifiers. For example, suppose that you want analyse the conversation on Twitter concerned with attitudes towards COVID-19 vaccination. A first step would be to identify a set of search terms that could be used to collect a reasonable sample of this conversation. A typical second step would involve the creation of a (relevancy) document classifier that is intended to filter out tweets that contain one the search terms, but which are not relevant to the conversation being targeted (perhaps because the matching search term was ambiguous). Once a (mostly) relevant set of documents (tweets) has been assembled, it is common to create a further document classifier (or collection of classifiers) that attempt to divide up the dataset according to which of a number of identifiable sub-topics the documents (tweets) are concerned with. As this example illustrates, this kind of scenario is likely to require the creation of a number of bespoke document classifiers: i.e. classification problems for which no suitable labelled data exists. State-of-the-art classifiers are currently built using very large pre-trained masked language models (e.g. BERT), and to achieve high performance this typically involves the use of several thousand labelled examples. BERT: https://arxiv.org/abs/1810.04805 RoBERTa: https://ai.facebook.com/blog/roberta-an-optimized-method-for-pretraining-self-supervised-nlp-systems/ In this project you will explore active learning methods that can be used to produce classifiers with reasonable performance, with a 10's, or 100's of labelled examples. Active learning methods are algorithms that iteratively select the most useful documents to labelled at each iteration. The methods under consideration would be evaluated on a wide variety of datasets involving different types of document classification, such as sentiment analysis, and topic-based classification. Good python programming skill and a familiarity with NLP are essential for this project. BERT: https://arxiv.org/abs/1810.04805 RoBERTa: https://ai.facebook.com/blog/roberta-an-optimized-method-for-pretraining-self-supervised-nlp-systems/	Shaun Ring

David Weir	InformaticsInformatics / Sussex Humanities Lab	Data augmentation methods for low resource document classification	In a wide variety of NLP applications, the need arises to create one or more bespoke document classifiers. For example, suppose that you want analyse the conversation on Twitter concerned with attitudes towards COVID-19 vaccination. A first step would be to identify a set of search terms that could be used to collect a reasonable sample of this conversation. A typical second step would involve the creation of a (relevancy) document classifier that is intended to filter out tweets that contain one the search terms, but which are not relevant to the conversation being targeted (perhaps because the matching search term was ambiguous). Once a (mostly) relevant set of documents (tweets) has been assembled, it is common to create a further document classifier (or collection of classifiers) that attempt to divide up the dataset according to which of a number of identifiable sub-topics the documents (tweets) are concerned with. As this example illustrates, this kind of scenario is likely to require the creation of a number of bespoke document classifiers: i.e. classification problems for which no suitable labelled data exists. State-of-the-art classifiers are currently built using very large pre-trained masked language models (e.g. BERT), and to achieve high performance this typically involves the use of several thousand labelled examples. In this project you will explore data augmentation methods that can be used to produce classifiers with reasonable performance, with a tens, or hundreds of labelled examples. Data augmentation methods are algorithms that expand (augment) a (potentially small) set of labelled instances by generating new labelled instances based on various heuristics. For example, rephrasing sentences in a document in ways that are presumed not to change the document's label. Data augmentation techniques that could be explored include simple approaches such as swapping words, to the use of generation model such as AUG-BERT. The methods under consideration would be evaluated on a wide variety of datasets involving different types of document classification, such as sentiment analysis, and topic-based classification. Good python programming skill and a familiarity with NLP are essential for this project. Supervisors: David Weir and Shaun Ring Links BERT: https://arxiv.org/abs/1810.04805	Shaun Ring
David Weir	InformaticsInformatics / Sussex Humanities Lab	Classification of long documents with transformer models	Transformer models such as BERT offer state-of-the-art performance on document classification problems. One drawback of such models, however, is that they are limited to documents of up to 512 tokens. While this is adequate for many scenarios, e.g. the classification of tweets, there are situations where this presents a problem, e.g. the classification of news articles. It is possible to break up a long document into suitably sized chunks and classify each one separately, before producing an overall classification of the whole document based on the class of each of the chunks. There are, however, problems with this approach, in particular: - how best to combine the decisions on each of the chunks to produce a decision for the whole document - how to handle the fact that the labelled data used to train the classifier will not provide chunk level labels, but document level labels. In this project you will explore the effectiveness of approaches to this problem, considering a variety of different datasets and types of classification problems. Supervisors: David Weir and Shaun Ring Links BERT: https://arxiv.org/abs/1810.04805	Shaun Ring

David Weir	InformaticsInformatics / Sussex Humanities Lab	Measuring sentence similarity: tackling the cross-encoders asymmetry problem	This project considers methods for the problem of measuring the similarity of two sentences. There are two dominant approaches both of which involve the use of a BERT (or RoBERTa) masked language model: cross-encoders and bi-encoders. Cross-encoders generally achieved higher performance than bi-encoders improvements on various sentence pair tasks. However, an asymmetry issue arises from the fact that cross-encoder involve sentence concatenation in its input scheme. The issue is that this has the potential to violate the reasonable assumption that sentence order should not have an impact on a measure of sentence similarity. This is an issue that is not discussed in recent work involving the use of cross-encoders In this project, you will explore methods that relax the asymmetry constraint in BERT (RoBERTa) cross-encoders, in particular, attempt to decreases the percentage of inconsistent predictions while maintaining the overall performance. For example, you could try to make modifications to BERT(RoBERTa)'s structure such as changes its positional embedding. Good python programming skill and a familiarity with NLP are essential for this project. Supervisors: David Weir and Qiwei Peng Links BERT: https://arxiv.org/abs/1810.04805 RoBERTa: https://ai.facebook.com/blog/roberta-an-optimized-method-for-pretraining-self-supervised-nlp-systems/	Qiwie Peng
David Weir	Informatics / MAH	Machine Storytelling	Language models have advanced rapidly in recent years, with OpenAI's GPT-3 language model capable of generating natural language text that can be difficult to distinguish from human-authored texts. The student(s) will examine the limitations and opportunities of current language models, and formulate a dissertation project around the development of a new application/tool within this space. Areas that could be explored include: sustaining coherence over longer outputs; creating stories with beginnings, middles and ends; enriching generated stories with knowledge of worlds or settings; multimedia storytelling incorporating imagery and sound; training with smaller text corpora (e.g. languages other than English); location-specific storytelling using geolocation data; distinguishing generated texts from human-authored texts; applying text generation within games, immersive storytelling, or other interactive media contexts; addressing biases embedded in training data (e.g. racism); developing tools to evaluate specific social and ethic risks of new language models; enabling more human oversight and input during text generation; developing machine-human collaboration interfaces incorporating generated text; developing more sustainable and less energy-intensive approaches to text generation; developing countermeasures against possible abuses of language models (e.g. extremism, fraud); improving text generation for specific genres or kinds of texts. Other topics related to NLP can also be considered. Supervisors: This interdisciplinary project will be jointly supervised by Ben Roberts, Jo Walton (MAH) and either David Weir or Julie Weeds. References: Alexander, Anne, Alan Blackwell, Caroline Bassett, Jo Walton. 2021.Ghosts, Robots, Automatic Writing: an AI Level Study Guide. CDH, 2021.https://www.cdh.cam.ac.uk/ghostfictions Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. 'On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? &#x1f99c'; In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 610–23. FAccT '21. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3442188.3445922. Hua, Minh, and Rita Raley. "Playing With Unicorns: AI Dungeon and Citizen NLP."DHQ: Digital Humanities Quarterly 14.4 (2020).https://www.proquest.com/scholarly-journals/playing-with-unicorns-ai-dungeon-citizen-nlp/docview/2553526112/se-2 Weidinger, Laura, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, et al. 2021. 'Ethical and Social Risks of Harm from Language Models'.ArXiv:2112.04359 [Cs], December. http://arxiv.org/abs/2112.04359.	Ben Roberts & Jo Walton (MAH)

David Weir	InformaticsInformatics / Sussex Humanities Lab	Discussion-based Community detection in social networks	Much of the existing work on community detection in social networks is based on networks where the edges in the network are derived from the way in which users (nodes in the network) interact. For example, in the case of Twitter, this could be based on who a user follows, or on who a user mentions, retweets, or replies to. An interesting alternative to this is to base the (strength of) edges in a network on the extent to which users are contributing similar content to an overall conversation taking place on the platform on one or more topics. Communities found on such networks correspond to groups of users who are contributing similar content. To create a network of this sort, it is possible to convert the content that a user posts into a point in high dimensional space, and this can be done with the use of transformers such as BERT. Users that are mapped to points that are close to each other in the space are more likely to be clustered together into the same community. In this project, a suitable dataset would be created and used to explore the potential of this approach. Supervisors: David Weir and Chris Inskip Links BERT: https://arxiv.org/abs/1810.04805	Chris Inskip
David Weir	InformaticsInformatics / Sussex Humanities Lab	A tool for visualizing and exploring network (graph) structured data in 2D or 3D	There is a great deal of interest in the data science community in algorithms that construct and analyse social networks. This project will be concerned with building a tool that can be used to visualise and explore such networks in various ways. The underlying data from which the network is produced may be derived from social-networks, but ideally the tool will be agnostic to any specific domain of data. The tool may be built as a interactive counterpart to existing network processing libraries (e.g. NetworkX in python). Such tooling will aim to facilitate the exploration of “communities” or “clusters” present in network structure, with functionality to, for example: - Interact with nodes and edges (e.g. adding and removing annotations) - Show node and edge annotations (e.g. through the use of textual labels, colours, sizes) - Highlight differences between changing network structures (e.g. across time) - Adjust the layout/position of nodes and edges (e.g. through the use of existing layout algorithms such as ForceAtlas) - Submit edits/annotations made to the network for processing and (re)visualise the result. This would be part of an existing project involving machine learning on network structured data. Supervisors David Weir and Chris Inskip	Chris Inskip

David Weir	Informatics / Sussex Humanities Lab	Tracing the Idea of a Scientific Instrument across Centuries	Objective and Sources Scientific instruments played a key role in the intellectual and industrial endeavours and developments of British history, from the sixteenth century to the twentieth. Over this extended historical period, how these instruments were understood and referred to, in diverse kinds of published texts, changed significantly. This MSc project will set out to identify the appearance of references in historical corpora and to trace some aspect(s) of these changes. Access will be available, for each student, to one of the following: a medium scale corpus of mixed texts from pre-1700 and eighteenth century (English language); a large corpus of nineteenth century medical texts (multiple European languages but primarily English); the Hansard record of parliamentary debate. These will enable such questions as: how the ideas of 'compass' or 'microscope' were used literally or figuratively; in what contexts were scientific measurement referenced in Parliament; how were names invented to refer to new kinds of instruments and what new processes were they associated with in nineteenth century medical texts? Background 'Tools of Knowledge: modelling the communities of scientific instrument makers in Britain, 1550-1914' is a thirty-month-long AHRC-funded project that began in early 2020. It is a collaboration between the University of Sussex (Dr Alex Butterworth), which is leading on the data processing and digital interpretation, the Department of History and Philosophy of Science at the University of Cambridge and the National Museum of Scotland, with the National Maritime Museum, Greenwich as a partner organisation. The project is creating a semantically modelled database of over ten thousand named instrument makers and businesses, their family and professional relationships with other individuals and institutions, together with the thousands of instruments - historical 'tools of knowledge' - made by them that still exist in museums. It is drawing on a wide range of written and visual sources and collections data, and applying many different computational methods - new 'tools of knowledge' - to extract, model, analyse and interpret. It aims to generate transformative insights into an area of activity crucial to the changing intellectual, industrial and commercial life of Britain over several centuries. The project is intended to enable new research at all levels, and offers a number of MSc students of AI/Data Science at Sussex an early opportunity	Alex Butterworth
David Weir	Informatics / Sussex Humanities Lab	Identifying Person and Object References in Online Sources	Objective and Sources The 'Tools of Knowledge' project seeks to expand the information about instrument makers, types of scientific instruments, and associated places and institutions recorded in its database by creating links to Wikidata and by identifying (and scraping) relevant online data. The task for this MSc project is to work with the directories and gazetteers created with reference to the project's database and to discover, link and collect related information discovered on the web. Work on this MSc project will be enabled by access to the Tools of Knowledge 'SENSIM' database of Makers, instruments, dates, etc, and the associated controlled vocabularies and taxonomies. Background 'Tools of Knowledge: modelling the communities of scientific instrument makers in Britain, 1550-1914' is a thirty-month-long AHRC-funded project that began in early 2020. It is a collaboration between the University of Sussex (Dr Alex Butterworth), which is leading on the data processing and digital interpretation, the Department of History and Philosophy of Science at the University of Cambridge and the National Museum of Scotland, with the National Maritime Museum, Greenwich as a partner organisation. The project is creating a semantically modelled database of over ten thousand named instrument makers and businesses, their family and professional relationships with other individuals and institutions, together with the thousands of instruments - historical 'tools of knowledge' - made by them that still exist in museums. It is drawing on a wide range of written and visual sources and collections data, and applying many different computational methods - new 'tools of knowledge' - to extract, model, analyse and interpret. It aims to generate transformative insights into an area of activity crucial to the changing intellectual, industrial and commercial life of Britain over several centuries. The project is intended to enable new research at all levels, and a number of MSc students of AI/Data Science at Sussex an early opportunity to explore some of the materials with which we are working, using data science methods to address real project needs. Methods	Alex Butterworth

David Weir	Informatics / Sussex Humanities Lab	Extracting ‘Event’ data from textual biographies	Objective and Sources The ‘Tools of Knowledge’ project aims to discover events in the lives of the instrument makers and the stories of the instruments they made, and to model and analyse these as data. Examples of these events might be: when and where was a person educated or involved in a legal dispute; or when were their instruments used in particular experiments or bought be a particular collector? The task, therefore, is to Identify likely events (as entity-graph structures) in digitised corpora collected for the current research projects, indexed to controlled vocabularies/taxonomies (names of persons, instrument types, institutions, event types, etc) produced by the project. A student working on this MSc project will be provided with a number of small to medium-sized corpora and may choose to work with any or all of them. These comprise (a) Six digitised texts of encyclopaedia-type works: short (1-3 page) biographies and object entries (English); (b) Automatically transcribed (HCR; handwritten character recognition) text from 10k index card records. Not ‘cleaned’ so fuzzy-matching challenge (English), (c) Free text database entries, short (1-5) sentence summary notes, object history or person biography related (English) Background ‘Tools of Knowledge: modelling the communities of scientific instrument makers in Britain, 1550-1914’ is an thirty-month-long AHRC-funded project that began in early 2020. It is a collaboration between the University of Sussex (Dr Alex Butterworth), which is leading on the data processing and digital interpretation, the Department of History and Philosophy of Science at the University of Cambridge and the National Museum of Scotland, with the National Maritime Museum, Greenwich as a partner organisation. The project is creating a semantically modelled database of over ten thousand named instrument makers and businesses, their family and professional relationships with other individuals and institutions, together with the thousands of instruments - historical ‘tools of knowledge’ - made by them that still exist in museums. It is drawing on a wide range of written and visual sources and collections data, and applying many different computational methods - new ‘tools of knowledge’ - to extract, model, analyse and interpret. It aims to generate transformative insights into an area of activity crucial to the changing intellectual, industrial and commercial life of Britain over several centuries. The project is intended to enable new research at all levels, and offers ta number of MSc students of AI/Data Science at Sussex an early opportunity to explore some of the materials with which we are working, using data science methods to address real project needs.	Alex Butterworth
David Weir	Informatics / Sussex Humanities Lab	Automatically Find Key Phrases In A Corpus - Utilising Dependency Analysis	This project concerns the exploration of approaches to the problem of identifying a set of key phrases that provides a good indication of the content of a corpus. The project would involve the following steps: Select existing datasets from e.g. Kaggle (no ethical review required), or data using a platform’s public API such as Reddit or YouTube (ethical review required). Then apply a keyword finding algorithm to find terms of interest. Implement an algorithm to extract important uses of these terms in coherent phrases using dependency analysis. Design an evaluation to compare strategies. Consider single large documents vs large collections of small documents. Consider how “important” might be defined. Supervisor Andrew Robertson	

David Weir	Informatics / Sussex Humanities Lab	Summarisation of Interview Responses in a Project Success Dataset	This project will involve the analysis of an existing 241,750 word dataset consisting of transcripts of interviews with project professionals on the role of different organisational factors for ensuring successful projects across a range of industrial sectors. The transcribed interviews cover a number of topics including sustainability, technology and data, and interpersonal skills and appropriate keywords will be used to identify which topic each interviewee's response relates to. Using these topic-specific keywords, it will be possible to collect the text of all of the responses relating each of the topics of interest, and the idea underlying this project is to explore Natural Language Processing (NLP) methods that are designed to provide a summary of these collections. The project will involve investigating the effectiveness of state-of-the-art abstractive and extractive summarisation methods on this dataset. Abstractive summarisation involves the generation of a new piece of text that provides a summary of the input text, whereas extractive summarisation selects content from the input texts and uses it to form the summary. There are many methods that could be explored, including those found here: https://paperswithcode.com/area/natural-language-processing/text-summarization Supervisors David Weir (Informatics) David Eggleton (SPRU)	David Eggleton
Julie Weeds	Informatics/SPRU	Mapping AI	This project aims to use AI to understand AI! Using keywords to collect a corpus of articles from different sub-fields of AI, this project will explore variation in terminology across time and sub-fields, using diachronic word embeddings (Kutozov et al. (2018)) and investigate diffusion of concepts from one area to another. Kutozov et al. (2018). Diachronic Word Embeddings and semantic shifts: a survey. https://www.aclweb.org/anthology/C18-1117/	Frederique Bone (SPRU)
Julie Weeds	Informatics & Eastbourne College	Marking Assistant	The aim of this project is to use NLP and ML methods to develop a marking assistant for GCSE-level short-answer questions. A dataset with around 30 student answers per question (4 questions as of 2020) is being developed and will be provided by Eastbourne College. Benomran and Ab Aziz (2013). Automatic essay grading for short answers in English Language. Journal of Computer Science https://www.researchgate.net/publication/269338716_Automatic_essay_grading_system_for_short_answers_in_English_language Taghipour and Ng (2016). A neural approach to automated essay scoring. EMNLP. https://www.aclweb.org/anthology/D16-1193/	Owen Dennis (Eastbourne College)

Julie Weeds	Informatics/LifeSciences	Exploiting automatic machine translation to find wildlife exploitation studies in other languages	Automatic methods for searching for studies on academic databases and on the Internet more generally are being used to identify articles for biodiversity and macroecological datasets (Cornford et al. 2020). These datasets may be used to provide evidence for the sustainability (or not) of various forms of wildlife exploitation including hunting by indigenous populations. However, ignoring non-English-language studies may skew the results (Konno et al. 2020). The aim of this project is to exploit automatic translation methods in order to identify non-English-language studies and assess the effect on conclusions drawn. Cornford et al. 2020. Fast, scalable, and automated identification of articles for biodiversity and macroecological datasets. Global ecology and Biogeography https://onlinelibrary.wiley.com/doi/full/10.1111/geb.13219 Konno et al. 2020. Ignoring non-English-language studies may bias ecological meta analyses https://onlinelibrary.wiley.com/doi/10.1002/ece3.6368	Fiona Matthews (Life Sciences)
Julie Weeds	Informatics/Industry	MeeTwo: Mental health chatroom for young people	MeeTwo Education Ltd provides a safe (and fully-moderated) online space for young people to talk about issues affecting their mental health. Anonymised posts (with meta-data) and a collection of mental health resources are available which have been developed over the past 4 years and provides an opportunity for a student to develop a project in a number of ways including automated moderation of posts, topic tracking, relevance matching / search and recommendation. Wang et al. 2020. Improving mental health using machine learning to assist humans in the moderation of forum posts. Health Informatics. http://users.sussex.ac.uk/~juliewe/wang20.pdf	Kerstyn Comley (MeeTwo)
Julie Weeds	Informatics	Fact Checking, fake news and Confirmation Bias	Being able to automatically check facts on the internet is a hot topic in NLP and machine learning. However, the way in which we check facts often leads to confirmation bias – we will find evidence supporting the fake fact that we are checking. If I search for “vaccines are bad for you”, the top results will be about side effects and adverse effects whereas if I search for “vaccines are good for you” – the top results are about the benefits of vaccines. This project will look at rewording potential facts as queries or neutral statements designed to return more balanced results. •https://www.scientificamerican.com/article/the-psychology-of-fact-checking1/ •https://fever.ai/ •https://misinforeview.hks.harvard.edu/article/the-presence-of-unexpected-biases-in-online-fact-checking/ •https://libguides.reynolds.edu/fakenews/bias	N/A
Julie Weeds	Informatics	Revision Assistant	The aim of this project would be to help students find and learn definitions of key terms and phrases for an area of study. For example, the input would likely be a set of lecture notes. NLE techniques would be applied to identify key terms / phrases and then link them to other occurrences of those key terms - potentially identifying occurrences which are definitional or most useful for a glossary. A game or quiz element could also be introduced to automatically generate questions and answers based on the linked content.	N/A
Julie Weeds	Informatics	Automatic Transcription	Over the past year, we have all become familiar with the use of automatic transcription on zoom and other platforms and also its current inadequacy in appropriate supporting non-native speakers. The aim of this project will be to post-process and improve the output of an automatic transcription service using available information about the topic or discourse. For example, the automatic transcription of lectures could be improved by using a language model based on the notes for that module.	
Julie Weeds	Informatics	Plagiarism Detection	This project will look at plagiarism detection methods for application to jupyter notebooks. The expectation is that the project would focus on the text cells but an alternative project could be developed looking at code cells. A variety of methods could be looked at including most simply word or n-gram overlap. Extensions could include looking for style markers which suggest that the text has been copied from another source (even when that source is unknown).	
Julie Weeds	Informatics	Source Language Detection	When text is translated from one language to another, tell-tale markers of the source language remain which native speakers might detect as disfluencies. These markers vary according to the source language. This project would look at developing a dataset and classifying translated texts according to the source language.	

Julie Weeds	Informatics/Psychology	Impact of Lockdown on Children’s Mental Health	This project would be looking at questions such as “Did lockdown harm children’s mental health?” and would be based on the analysis of surveys of parents carried out by Psychologists at Sussex during lockdown. The dataset contains responses from around 2500 parents over the period of a year with parents responding between 1 and 8 times each. The data is largely quantitative (rather than free text) and this project would likely suit a Human and Social Data Scientist.	Sam Cartwright-Hatton