UNIVERSITY OF CAMBRIDGE

DiS MPHIL

MPhil in Data Intensive Science

PROJECT LIST

Contents

1 Data-driven Computed Tomography Reconstruction

Name	Dr Ander Biguri & Zakhar Shumaylov
Role and Affiliation	Senior Research Associate & PhD student, Department of Applied Mathematics and Theoretical Physics
Contact Email	ab2860@cam.ac.uk & zs334@cam.ac.uk
Key Publication	See list below

Project Description

Computed Tomography (CT) is a widely used imaging technique in medicine and industry, but the process of reconstructing CT images is complex. CT scanners acquire multiple X-ray projections (similar to X-ray images you see when you break a bone) from different angles, but these projections alone don't reveal depth information. To recover that, the tomography machine rotates and acquires several projections. From this set of images (a.k.a. *sinograms*) obtaining the CT reconstructed image is a mathematically very challenging problem, as the problem becomes *ill-posed*, when sinograms are sparsely sampled. As a result, a large number of X-ray projections is required to avoid reconstruction errors, leading to prolonged exposure times. This higher radiation dose is undesirable, especially in medical applications.

Nowadays, the literature on using machine learning to solve ill-posed problems like CT is booming. Most recent methods claim to require a reduced X-ray dosage while producing the same image quality. These methods range from those grounded in mathematical theory to those focused on computationally scalable approaches.

In this project, you will select and reproduce a key paper from the literature on using machine learning for low-dose CT reconstruction.

Project goals

Main goals of the project:

1. Produce a pipeline for loading the Mayo Clinic Low Dose CT Grand Challenge dataset, producing simulated low-dose measurements, and reconstruct it using traditional non-ML algorithms, using tomosipo.
2. Implement one learned reconstruction method from the list below (or propose a new one).
3. Reproduce the experiments, results and plots from the chosen paper.

Method	Math	Code	Physics	ML
Convex Ridge Regularizer	70%	50%	10%	80%
Unrolled Adversarial Regularization	90%	70%	10%	80%
RISING	40%	40%	10%	50%
DeepFBP	20%	30%	10%	30%
SLPD	40%	50%	10%	50%

Table 1: Difficulty estimate for each possible method.

The above goals will get you a high grade in the project. Note that the projects available are of different difficulty level (see Table 1) and we will take that into consideration at grading. More options are available, and we welcome proposals from the students. If you finish your work early and you want to extend your project to research-level, the following directions are possible:

- Transfer your implementation to our research library LION.
- Find ways to improve the method with the help of our research team.

These should be optional extensions for stronger students to attempt if they manage to complete the main project goals. If students complete **one** of these to a high standard then they should be typically awarded a high distinction. Those could be things like updating or changing the methods, adding a new features or functionality to the analysis pipeline, or including new data to the analysis.

Up to 5 students can be selected for this project.

Prerequisites

You need to take the Medical Imaging minor, or have equivalent pre-existing knowledge. Strong coding skills (`pytorch`, `scipy`, `numpy`) are required, but much of this will be taught in the MPhil. For the more "mathy" sub-projects, a strong maths background, particularly in optimization, is desired.

Reading List

A general book on CT that is good to have in hand:

- Hansen, Per Christian, Jakob Jørgensen, and William RB Lionheart, eds. **Computed tomography: algorithms, insight, and just enough theory.** *Society for Industrial and Applied Mathematics,* 2021.

For the particular methods being implemented:

- Convex Ridge Regularization: A. Goujon, *et al.*, **"A Neural-Network-Based Convex Regularizer for Inverse Problems,"** in *IEEE Transactions on Computational Imaging*, vol. 9, pp. 781-795, 2023, DOI: 10.1109/TCI.2023.3306100.
- Unrolled Adversarial Regularization: Mukherjee, Subhadip, *et al.* **"End-to-end reconstruction meets data-driven regularization for inverse problems."** *Advances in Neural Information Processing Systems 34 (2021)*: 21413-21425.
- RISING: Davide Evangelista, *et al.*, **RISING: A new framework for model-based few-view CT image reconstruction with deep learning**, *Computerized Medical Imaging and Graphics*, Volume 103, 2023, 102156, ISSN 0895-6111, doi.org/10.1016/j.compmedimag.2022.102156.
- DeepFBP: X. Tan, *et al.*, **"Deep Filtered Back Projection for CT Reconstruction,"** in *IEEE Access*, vol. 12, pp. 20962-20972, 2024, doi: 10.1109/ACCESS.2024.3357355.
- SLPD Tang, Junqi, *et al.* "**Stochastic primal-dual deep unrolling.**" *arXiv preprint* arXiv:2110.10093 (2021).

More references and reading list will be given once we start the project. Hope you join us!

2 Disentangling the components of the Milky Way with Gaussian Mixture Modelling

Proposer Name	Dr. Anke Ardern-Arentsen
Proposer Role and Affiliation	Postdoctoral Research Fellow, Institute of Astronomy
Proposer Contact Email	anke.arentsen@ast.cam.ac.uk
Key Publication	On the existence of a very metal-poor disc in the Milky Way - Zhang, Ardern-Arentsen & Belokurov 2024

Project Description

There are billions of stars in the Milky Way, our home Galaxy, and we can use their properties to infer our Galaxy's formation history. The stars in the Milky Way are distributed among different components – the main ones being a young thin disc, a slightly older thick disc and an old spherical halo – and the task of Galactic Archaeologists is to figure out how these formed. An active research question is what the very early Milky Way looked like, for example, 'When did a disc start to form?', and we can use the oldest stars in our Galaxy to address this. It is difficult to measure the ages of stars directly, so astronomers often use the chemical composition of stars instead. A star born in the early Universe that still survives today is more "pristine" than stars born later, when the Universe starts being enriched in elements by dying stars. In astronomers' terms: old stars have low metallicities, younger stars have higher metallicities.

In this project, the student will use a large dataset of stellar metallicities and dynamical properties to study the different components of the Milky Way, with a focus on old/low metallicity stars. The data comes from the Gaia mission satellite, which has been collecting detailed positional and colour information for billions of stars in the Galaxy, made available to the community through public data releases. Its most recent data release, DR3, contains the necessary information to homogeneously derive metallicities and orbital properties for millions of bright stars across the entire sky. The student will apply a Gaussian Mixture Model to this data to identify components in the velocity distribution of stars in different metallicity regimes, one of the main goals being to test whether a disc component is present at the lowest metallicities.

Project goals

Main goals of the project:

1. Load the data, cross-match to get distances, familiarise yourself with the data, apply quality cuts and derive velocities from the Gaia data. Reproduce Figures 1, 2 and 3.

2. Apply a Gaussian Mixture Model to the data and use the Bayesian Information Criterion to determine the best number of components in each metallicity range. Reproduce Figures 4 and 5.

3. Test for potential residuals from a disc-like component at low metallicity and place limits on its contribution. Create a figure similar to Figure 7.

Extension directions:

- Extend the analysis to a larger sample with a less strict quality cut on the distances, and discuss the advantages and disadvantages

- Improve the modelling and/or explore other methodologies to identify structure in velocity space
- Add another dimension to the GMM analysis, such as alpha element abundances (public catalogues derived from Gaia DR3 data are available)

Prerequisites

Required: Python familiarity.

Desirable: experience with Python libraries for data manipulation/visualisation (e.g. Pandas, Numpy, Matplotlib). Some background knowledge in physics or astronomy.

Helpful MPhil modules: Statistical Methods, Galactic Archaeology.

Reading List

- *Galactic Archaeology with Gaia* – Belokurov & Deason, 2024, Review article for New Astronomy Reviews – [link]
- *Extreme deconvolution: Inferring complete distribution functions from noisy, heterogeneous and incomplete observations* – Bovy, Hogg & Roweis, 2011, Annals of Applied Statistics, 5, 1657 – [link]

Data Access

- The main dataset to use is the vetted RGB sample from Andrae et al. (2023). This table contains coordinates, metallicities and the necessary Gaia stellar velocity columns, and can be downloaded here: link
- Distances come from [Bailer-Jones et al. (2021)]. Do not download the entire table, cross-match only the necessary subsample of stars (instructions will be provided).

3 Learning Cosmological Observables with Deep Neural Networks

Proposer Name	Dr Boris Bolliet
Proposer Role and Affiliation	Assistant Teaching Professor, Cavendish Astrophysics
Proposer Contact Email	bb667@cam.ac.uk
Key Publication	High-accuracy emulators for observables in ΛCDM, $N_{\rm eff}$, Σm_ν, and w cosmologies

Project Description

This project aims to reproduce results from the paper "High-accuracy emulators for observables in ΛCDM, $N_{\rm eff}$, Σm_ν, and w cosmologies" which presents the most accurate cosmological observable emulators available on the market.

Students will be asked to reproduce the results only for the base ΛCDM cosmology, by training deep neural networks on the Cambridge HPC.

Project goals

Main goals of the project:

1. Train deep neural networks on a large dataset of cosmological observables.
2. Implementation of the swish activation function.
3. Explore different learning aproaches.
4. Evaluate and report on the performance of the networks against test data.

Extension directions

- Experiment with the neural network architecture to improve the performance of the emulators.
- Build emulators for matter power spectrum using a different approach than the naive one presented in the paper (and still being used today).
- Work on a Jax implementation of the emulators.

Prerequisites

Required:

- Familiarity with Python and machine learning libraries.
- Familiarity with working with large datasets and HPC.

Desirable:

- Internest in Cosmology.

Reading List

- High-accuracy emulators for observables in ΛCDM, $N_{\rm eff}$, Σm_ν, and w cosmologies(Key Paper)
- COSMOPOWER: emulating cosmological power spectra for accelerated Bayesian inference from next-generation surveys
- SPECULATOR: Emulating stellar population synthesis for fast and accurate galaxy spectra and photometry
- Understanding Deep Learning
- Andrej Karpathy's Neural Network Training Recipe

Data Access

The training and testing data will be directly provided on the Cambridge HPC by Boris Bolliet. A version of the training code will be provided to the students (see cosmopower), but they will be encouraged to develop their own.

Supervision

The supervision will be provided by Dr Boris Bolliet.

4 Listening to the Noise: Blind Denoising with Gibbs Diffusion

Proposer Name	Dr Boris Bolliet, Dr. Fiona McCarthy
Proposer Role and Affiliation	(BB) Assistant Teaching Professor, Cavendish Astrophysics (FM) Senior Research Associate, DAMTP
Proposer Contact Email	bb667@cam.ac.uk
Key Publication	Listening to the Noise: Blind Denoising with Gibbs Diffusion

Project Description

This project aims to reproduce results from the paper "Listening to the Noise: Blind Denoising with Gibbs Diffusion" which presents an approach to perform diffusion-based denoising without needing the knowledge of the noise level and its covariance. To achieve this, the authors introduce Gibbs Diffusion (GDiff), a general methodology addressing posterior sampling of both the signal and the noise parameters

Students will perform GDiff denoising in two cases:

- blind denoising of natural images involving colored noises with unknown amplitude and spectral index
- a cosmology problem, namely the analysis of cosmic microwave background data, where Bayesian inference of "noise" parameters means constraining models of the evolution of the Universe

and reproduce results from Fig 3 and 5 of the paper.

Project goals

Main goals of the project:

1. Understand and apply the concept of diffusion, denoising and Gibbs sampling.
2. Train a diffusion model on the ImageNet 2012 dataset using the Cambridge HPC GPUs.
3. Construct simulated dust emission and Cosmic Microwave Background maps.

Extension directions

- Test the robustness and optimization of the pipeline by adjusting hyperparameters of the diffusion model and the training data.
- Compare with other denoising methods, e.g., DnCNN and BM3D.

Prerequisites

Required:

- Familiarity with Python and machine learning libraries.
- Interest in Image Analysis and Signal Processing.

Desirable:

- Interest in Cosmology.

Reading List

- Listening to the Noise: Blind Denoising with Gibbs Diffusion (Key Paper)
- Removing Dust from CMB Observations with Diffusion Models
- Denoising Diffusion Probabilistic Models
- Score-Based Generative Modeling through Stochastic Differential Equations
- The Catalogue for Astrophysical Turbulence Simulations (CATS)

Data Access

The codes are publicly available on GitHub and use publicly available datasets. See the GitHub repository for more details.

Supervision

The supervision will be provided by Dr Boris Bolliet and Fiona McCarthy, together, throughout.

5 Parameter Inference with Diffusion Model driven Hamiltonian Monte Carlo

Proposer Name	Dr Boris Bolliet
Proposer Role and Affiliation	Assistant Teaching Professor, Cavendish Astrophysics
Proposer Contact Email	bb667@cam.ac.uk
Key Publication	Diffusion-HMC: Parameter Inference with Diffusion Model driven Hamiltonian Monte Carlo

Project Description

Diffusion generative models have excelled at diverse image generation and reconstruction tasks across fields. A less explored avenue is their application to discriminative tasks involving regression or classification problems. The cornerstone of modern cosmology is the ability to generate predictions for observed astrophysical fields from theory and constrain physical models from observations using these predictions. This work uses a single diffusion generative model to address these interlinked objectives – as a surrogate model or emulator for cold dark matter density fields conditional on input cosmological parameters, and as a parameter inference model that solves the inverse problem of constraining the cosmological parameters of an input field. The model is able to emulate fields with summary statistics consistent with those of the simulated target distribution.

We then leverage the approximate likelihood of the diffusion generative model to derive tight constraints on cosmology by using the Hamiltonian Monte Carlo method to sample the posterior on cosmological parameters for a given test image. Finally, we demonstrate that this parameter inference approach is more robust to the addition of noise than baseline parameter inference networks.

Project goals

Main goals of the project:

1. Understand and apply the concept of diffusion.
2. Train a diffusion model using the Cambridge HPC GPUs.
3. Gain experience with Hamiltonian monte carlo techniques.

Extension directions

- Test the robustness and optimization of the pipeline by adjusting hyperparameters of the diffusion model and the training data.

Prerequisites

Required:

- Familiarity with Python and machine learning libraries.

Desirable:

- Interest in Cosmology.

Reading List

- Diffusion-HMC: Parameter Inference with Diffusion Model driven Hamiltonian Monte Carlo (Key Paper)

Data Access

The codes are publicly available on GitHub and use publicly available datasets. See the GitHub repository for more details.

Supervision

The supervision will be provided by Dr Boris Bolliet.

6 Simulation-based Inference for Stochastic Gravitational Wave Background

Proposer Name	Dr Boris Bolliet and Dr James Alvey
Proposer Role and Affiliation	(BB) Assistant Teaching Professor, Cavendish Astrophysics (JA) Kavli Fellow, KICC
Proposer Contact Email	bb667@cam.ac.uk, jbga2@cam.ac.uk
Key Publication	Simulation-based inference for stochastic gravitational wave background data analysis

Project Description

This project aims to reproduce results from the paper "Simulation-based inference for stochastic gravitational wave background data analysis" which presents an approach to perform simulation based inference (SBI) in order obtain posterior probability distributions for parameters characterizing the stochastic gravitational wave background signals. With SBI, the authors show that they are able to resolve biases that are otherwise present when using traditional likelihood-based inference (e.g., via Monte Carlo Markov Chains).

Students will implement both SBI and MCMC methods and reproduce results from FIG 1 and 2 of the paper. They will use the codes provided by the authors, namely saqqara and swyft, and for the MCMC part the publicly available emcee.

Project goals

Main goals of the project:

1. Gain experience with SBI and MCMC methods, and understand the benefits of SBI over MCMC.
2. Use truncated marginal neural ratio estimation (TMNRE) for posterior estimation.

Extension directions

- Test the robustness of the SBI pipeline when adjusting parameters of the TMNRE and size of training data.
- Implement other sampling techniques (e.g., nested sampling) and compare their performance to SBI/TMNRE and MCMC.
- Use other classes of SBI algorithms, such as Neural Posterior Estimation, Neural Likelihood Estimation, Neural Ratio Estimation available in other packages like sbi.

Prerequisites

Required:

- Familiarity with Python.

Desirable:

- Interest in time series analysis or gravitational wave astronomy.

Reading List

- Simulation-based inference for stochastic gravitational wave background data analysis (Key Paper)
- The frontier of simulation-based inference
- Peregrine: Sequential simulation-based inference for gravitational wave signals

Data Access

The codes are publicly available on GitHub (see saqqara and swyft) and will be used to generate the necessary data.

Supervision

The supervision will be shared between by Dr James Alvey, Dr Boris Bolliet and Dr Christopher Moore, either together or separately.

7 Superhuman Synthesis of Scientific Knowledge with LLM Agents

Proposer Name	Dr Boris Bolliet
Proposer Role and Affiliation	(BB) Assistant Teaching Professor, Cavendish Astrophysics
Proposer Contact Email	bb667@cam.ac.uk
Key Publication	Language Agents Achieve Superhuman Synthesis of Scientific Knowledge

Project Description

Language models are known to "hallucinate" incorrect information, and it is unclear if they are sufficiently accurate and reliable for use in scientific research. The authors developed a rigorous human-AI comparison methodology to evaluate language model agents on real-world literature search tasks covering information retrieval, summarization, and contradiction detection tasks. They show that PaperQA2, a frontier language model agent optimized for improved factuality, matches or exceeds subject matter expert performance on three realistic literature research tasks without any restrictions on humans (i.e., full access to internet, search tools, and time). PaperQA2 writes cited, Wikipedia style summaries of scientific topics that are significantly more accurate than existing, human-written Wikipedia articles. They also introduce a hard benchmark for scientific literature research called LitQA2 that guided design of PaperQA2, leading to it exceeding human performance. Finally, they apply PaperQA2 to identify contradictions within the scientific literature, an important scientific task that is challenging for humans. PaperQA2 identifies 2.34 ± 1.99 (mean $\pm$ SD, N = 93 papers) contradictions per paper in a random subset of biology papers, of which 70These results demonstrate that language model agents are now capable of exceeding domain experts across meaningful tasks on scientific literature.

The students will implement the PaperQA2 agent and reproduce results from Figure 2 of the paper (with the OpenAI and Meta models, i.e., GPT and LLama).

Project goals

Main goals of the project:

1. Apply LLMs to scientific literature.
2. Understand Retrieval Augmented Generation (RAG) and how it can be used to improve the accuracy of LLMs.
3. Understand how to benchmark LLMs and compare their performance to humans.

Extension directions

- Implement PaperQA2 agent in CMBAgent.

Prerequisites

Required:

- Expert in Python.

Desirable:

- Interest in LLMs and AI research.

Reading List

- Language Agents Achieve Superhuman Synthesis of Scientific Knowledge (Key Paper) and references therein.

Data Access

The codes are publicly available on GitHub (see paper-qa and CMBAgent). An Openai API key will be provided to access the OpenAI models.

Supervision

The supervision will be done by Dr. Boris Bolliet.

8 Detecting Laser Points in Antarctic Benthic Imagery to aid Biodiversity Monitoring

Proposer Name	Dr. Cameron Trotter
Proposer Role and Affiliation	Machine Learning Research Scientist, British Antarctic Survey
Proposer Contact Email	cater@bas.ac.uk
Key Publication	Schoening, T., Kuhn, T., Bergmann, M., Nattkemper, T.W., 2015. DELPHI - fast and adaptive computational laser point detection and visual footprint quantification for arbitrary underwater image collections. Front. Mar. Sci. 2. https://doi.org/10.3389/fmars.2015.00020

Project Description

For most of human history our knowledge of what lives in the benthos, or the bottom of a body of water, has relied on nets or devices to bring organisms to the surface. However, utilising these methods for biodiversity monitoring is intrinsically destructive and fails to provide insight into community structure. The development of technologies such as towed camera systems or autonomous underwater vehicles now allow humans to monitor benthic environments in situ, enabling a non-destructive way of understanding community distributions, structure, and function. However, the rapid development of these tools has led to increasing volumes of images being collected. Labelling these images can be extremely resource intensive and require expert knowledge of the monitored ecosystem, resulting in a data bottleneck.

Multiple research groups are now turning to image processing and computer vision to help automate some of this curation and reduce the bottleneck. At the British Antarctic Survey (BAS), we are currently undertaking a project exploring the use of these techniques to automate curation of survey data collected in the Southern Ocean, as manual labelling currently takes around 7-8h per image due to the high level of biodiversity present in the survey area. Data is collected via the Ocean Floor Observation and Bathymetry System (OFOBS; see Reading List item 1) in a top-down view and contains (at most) three laser points, each a set distance apart, allowing ecologists to calculate organism size, analyse community structure, and perform colour correction. However due to environmental conditions and the large scale of the imagery it is often non-trivial to locate each point in the image. This project seeks to automate the detection of these points, allowing for faster data curation. It is hoped that through increased automation, BAS ecologists can spend less time curating data and more time developing mitigation strategies, taking into account a larger volume of data compared to what is presently possible.

Over the course of the project the student will gain experience in techniques such as template matching, (small) object detection/segmentation, binary masking, colour thresholding, use of colour spaces, morphological transformations, as well as how to apply these to high-resolution imagery (patching etc.). Experience with the theoretical and practical considerations of machine learning systems development such as data exploration, model training, tuning, and validation will also be obtained. Outside of this, skills in modular software design and version control will also be developed.

Project goals

Main goals of the project:

1. Implement the underwater laser point detection system as described in the Key Publication.

2. Apply the above system to the OFOBS data. Understand where it is successful, and under what conditions it is likely to fail.
3. Integrate your work into BAS' automated curation system.

Extension directions:

- Implement additional methods as outlined in the Reading List. Compare and contrast the effectiveness of these to the Key Publication implementation.
- Based on the failure cases of the developed implementation(s), propose improvements to the current state of the art.

Prerequisites

Required Skills:

- Experience with the Python programming language.
- Knowledge of image processing techniques and the use of the OpenCV framework.
- Knowledge of machine learning techniques and the use of frameworks for implementing these such as Scikit-learn.

Reading List

1. Purser, A., Marcon, Y., Dreutter, S., Hoge, U., Sablotny, B., Hehemann, L., Lemburg, J., Dorschel, B., Biebow, H., Boetius, A., 2019. Ocean Floor Observation and Bathymetry System (OFOBS): A New Towed Camera/Sonar System for Deep-Sea Habitat Surveys. IEEE J. Oceanic Eng. 44, 87–99. https://doi.org/10.1109/JOE.2018.2794095
2. Mbani, B., Schoening, T., Gazis, I.-Z., Koch, R., Greinert, J., 2022. Implementation of an automated workflow for image-based seafloor classification with examples from manganese-nodule covered seabed areas in the Central Pacific Ocean. Sci Rep 12, 15338. https://doi.org/10.1038/s
022-19070-2
3. Widodo, R.B., Chen, W., Matsumaru, T., 2012. Laser spotlight detection and interpretation of its movement behavior in laser pointer interface, in: 2012 IEEE/SICE International Symposium on System Integration (SII), Fukuoka, Japan, pp. 780–785. https://doi.org/10.1109/SII.2012.6
4. Soetedjo, A., Ashari, M.I., Mahmudi, A., Nakhoda, Y.I., 2014. Raspberry Pi based laser spot detection, in: 2014 International Conference on Electrical Engineering and Computer Science (ICEECS), Kuta, Bali, Indonesia, pp. 7–11. https://doi.org/10.1109/ICEECS.2014.7045210
5. Rzhanov, Y., Mamaenko, A., Yoklavich, M., 2005. UVSD: Software for Detection of Color Underwater Features, in: Proceedings of OCEANS 2005 MTS/IEEE. Presented at the OCEANS 2005 MTS/IEEE, Washington, DC, USA, pp. 1–4. https://doi.org/10.1109/OCEANS.2005.
6. Pilgrim, D.A., Parry, D.M., Jones, M.B., Kendall, M.A., 2000. ROV Image Scaling with Laser Spot Patterns. Underwater Technology 24, 93–103. https://doi.org/10.3723/175605400783259684
7. Istenič, K., Gracias, N., Arnaubec, A., Escartín, J., Garcia, R., 2020. Automatic scale estimation of structure from motion based 3D models using laser scalers in underwater scenarios. ISPRS Journal of Photogrammetry and Remote Sensing 159, 13–25. https://doi.org/10.1016/j.isprsjpr

Data Access

Example data collected by the OFOBS system can be found at https://doi.pangaea.de/10.1594/PANGAEA.911904.

Laser point labels will be provided at the start of the project.

9 Advancing the discovery of Binary Neutron Stars with Gravitational Wave Observations.

Dr P. Canizares	
University of Cambridge	Research Fellow, DAMTP
pc464@cam.ac.uk	

[P.1] Key Publication	A machine-learning classifier for the postmerger remnant of binary neutron stars
[P.2] Key Publication	Improving Early Detection of Gravitational Waves from Binary Neutron Stars Using CNNs and FPGAs

Detection and characterisation of Gravitational Waves from Binary Neutron Star systems

This project focuses on the discovery of Binary Neutron Stars (BNSs) through Gravitational Wave (GW) detections. During LIGO and VIRGO's first two observing runs (O1 and O2), 11 GW signals from compact binary mergers were identified, including the landmark event GW170817—the first detected BNS merger. The detection of GW170817 in both gravitational and electromagnetic spectra marked the advent of Multi-Messenger Astrophysics (MMA), a field that integrates GWs, electromagnetic radiation, cosmic rays, and neutrinos to explore astrophysical phenomena. The goal of project 2 (P.2) is to detect GW signals as early as possible while minimizing false positives to prevent inaccurate alerts from being sent to electromagnetic telescopes.

Following the merger of a BNS system, the remnant can either collapse into a black hole or form a neutron star. Accurately characterizing the properties of the remnant is crucial for advancing our understanding of the equation of state governing the ultra-dense matter inside neutron stars and for uncovering the complex physical processes that occur in the post-merger phase. Furthermore, predicting whether a BNS system will promptly collapse into a black hole or produce a neutron star remnant is essential for developing post-merger GW models and guiding electromagnetic telescopes in follow-up observations. The goal of project 1 (P.1) is to characterise the BNS remnant after collapse.

The student can choose one of the following projects:

[**P.1**] Built a ML classifier to predict the outcome of a BNS merger based on parameters that LIGO can measure from the inspiral GW signal.

[**P.2**] Build a novel Generative Model-based BNS detection algorithm using a curriculum learning strategy.

Project goals

P.1—Main goals of the project:

1. Download, analyse and physically interpret dataset.
2. Implement ML binary classier for prompt collapse to Black Hole (BH) vs a Neutron Star (NS) remnant.

3. Implement ML binary classier for prompt collapse to BH vs a hypermassive NS.
4. Implement ML multi-label classifier for all the cases above.
5. Evaluate performance of the different classifiers to evaluate BNS renmant based on GW observables.

Extension directions P.1:

- Include the effect of tidal deformation.
- Can we distinguish between BNS and a Neutron Star-Black hole binary?
- How do our results change for different masses?

P.2—Main goals of the project:

1. Generate data set using public available software (LALSimInspiralSpinTaylor.)
2. Build a GWaveNet: a CNN exploiting probabilistic and autoregressive models combined with filters and dilated convolutions—this allows to capture larger contexts without increasing the computational load and to learn the time ordering of the data.
3. Use different architectures suitable for CPU and GPU (optionally FPGA).
4. Compare the performance of your GWanet with FindCNN (see P.2 reference).

Extension directions P.2:

- Implement a time-frequency transform module and perform the same study using a wavelet decomposition of the GW data. This will speed-up the computation. Compare accuracy and performance of the time-based and wavelet-based networks.

Prerequisites

The student should be familiar with Pytorch, and ML\DL-based classification techniques. Familiarity with signal analysis and gravitational waves is desirable but not necessary.

Reading List

Beyond the references cited in the main paper, the following sources may also be of interest:

- Long-lived neutron-star remnants from asymmetric binary neutron star mergers: element formation, kilonova signals and gravitational waves(P.1 & P.2)
- Complete phenomenological gravitational waveforms from spinning coalescing binaries(P.2)

Data Access

- Data for project 1 can be found here: data P.1
- Data for *P.*2 has to be generated using LIGO's simulation pipeline in LAL suite data P.2

10 Improving clinical diagnosis with generative modelling and super-resolution

Proposer Name	Dr P. Canizares
Proposer Role and Affiliation	Research Fellow DAMTP
Proposer Contact Email	pc464@cam.ac.uk
Key Publication	InverseSR: 3D Brain MRI Super-Resolution Using a Latent Diffusion Model

Project Description

Magnetic resonance image (MRI) super-resolution has gained significant attention for its ability to enhance low-resolution MRI scans, thereby improving the accuracy of clinical diagnoses. Routine MRI scans often have low resolution (LR) and exhibit considerable variations in contrast and spatial resolution due to differences in scanning parameters.

This project aims to evaluate the potential of novel diffusion model-based AI techniques for enhancing brain MRI clinical images through super-resolution, with the goal of improving their utility in medical diagnosis.

Project goals

The main goals of the project is to build an unsupervised method for MRI super-resolution. The following steps are necessary:

1. Download and analyse the brain latent diffusion model (LDM) from here, which is obtained from the the UK Biobank.
2. Build 3D brain MRI priors: Autoencoder and diffusion model.
3. Obtain the latent representation needed to reconstruct a noisy image into a high-resolution image.
4. Build the *inverseSR* decoder (see main reference).
5. Evaluate the superresloved MRI scans based on a suitable metric.

Extension directions: Replace the slow MCM sampling using a diferent approach. We suggest a score-based approach, similar to this one

Prerequisites

The student should be familiar with Pytorch and probabilistic learning. Prior experience working with large data sets and MCMC methods would be desirable, but not essential.

Reading List

Beyond the references cited in the main paper, the following sources may also be of interest:

1. Diffusion Models, Image Super-Resolution And Everything: A Survey
2. MRI Super-Resolution using Multi-Channel Total Variation

Data Access

a. The UK Biobank

b. The IXI dataset contains $\sim$ 600 high resolution MR images from healthy subjects, of multiple MR contrasts IXI Dataset.

11 Normalising Flows for Accelerating Gravitational Wave Population Inference

Proposer Name	Dr Christopher J. Moore
Proposer Role and Affiliation	Associate Professor, IoA+DAMTP
Proposer Contact Email	cjm96@ast.cam.ac.uk
Key Publication	The key publication for this project isn't a paper. We will work to reproduce and extend the work in this GitHub repository

Project Description

Bayesian inference is used extensively in the new field of gravitational wave (GW) astronomy to measure the properties of sources. The LIGO and Virgo collaborations release posterior samples for each individual event detected; there are now around 100 events, mostly binary black hole (BH) mergers, in the public catalogs and many more expected from the ongoing observing runs. A few thousand posterior samples are generally produced for each individual GW event. These reasonably large and clumbersome data products are not a major issue with only a few hundred events, but will become a problem with the increasing numbers of events expected with next generation of detectors.

Posterior samples are usually obtained using a simple unphysical prior. This is not ideal; we would like to include other astrophysical information in the prior. This would allow us to make improved measurements of the GW sources and to feed the information gained back to learn more about the astrophysical environments and processes that produce the binary BH mergers that LIGO and Virgo are now observing. The process of modelling the distributions of source properties across the entire catalog is called GW population inference. Hierarchical Bayesian models are used to simultaneously infer the properties of individual GW events and the properties of the underlying population. (An example of an individual event parameter might be the mass of a particular BH. An example of a population-level parameter might be the power-law slope of the BH mass function.) GW population inference uses the posterior samples (obtained with unphysical priors) for all the individual events in the catalog and can be computationally expensive.

This project will attemp to address both of the problems mentioned above. Firstly, normalising flows will be used to obtain a compressed (and smooth) representation of the Bayesian posterior for each individual GW event. This compressed GW catalog will be published as a useful resource for the community. Secondly, this compressed GW catalog will be used to perform GW population inference. The expectation is that when using the compressed, normalising flow representations of the individual GW event posteriors (instead of the clumbersome posterior samples) the computational costs of GW population inference will be reduced. If time allows, this project can be extended further exploring ways to accelerate GW population inference that use derivative information (such as Hamiltonian Monte Carlo, or NUTS) and exploit the fact that the normalising flows give smooth representations of the individual GW event posteriors.

Project goals

The main goals of the project:

1. Construct a normalising flow to represent the Bayesian posterior for an individual GW event. This will serve as a smooth, compressed representation of the posterior samples.

2. Construct normalising flows for all the publically available events and publish this data product as a new compressed catalog.

3. Use the compressed representations of the individual GW event posteriors in a Hierarchical Bayesian model to perform GW population inference.

Extension directions:

- Explore different flow architectures and quantify their efficiency at compressing the individual GW event posteriors.
- Explore different sampling methods. Particularly promising are gradient-based algorithm (such as Hamiltonian Monte Carlo sampling, or the NUTS algorithm) which will be able to exploit the fact that the new use of a flow now gives a smooth representation of the the individual event posteriors.

Prerequisites

Necessary: Python 3 familiarity. Desirable: Some prior experience working with large data sets and MCMC methods would be desirable, but not essential.

Reading List

- The LIGO and Virgo collaborations latest "populations paper", "The population of merging compact binaries inferred using gravitational waves through GWTC-3". This is long paper containing a lot of results. It is a very good reference, but probably not the best place to start reading about the subject.
- Mandel, Farr & Gair, "Extracting distribution parameters from multiple uncertain observations with selection biases", Monthly Notices of the Royal Astronomical Society, vol 486, 1 (2019), link. This is the paper that first laid out the formalism for doing population inference in a gravitational wave context.
- Moore, Gerosa, "Population-informed priors in gravitational-wave astronomy", Phys. Rev. D 104, 083008 (2021) link. This paper shows why it is important to use the correct, population-informed prior when doing GW Bayesian inference.

Data Access

All public gravitational wave data products from the LIGO and Virgo observatories may be obtained from the Gravitational Wave Open Science Center (GWOSC).

12 Relativistic Effects in Astrometry with Gaia

Proposer Name	Dr Christopher J. Moore
Proposer Role and Affiliation	Associate Professor, IoA+DAMTP
Proposer Contact Email	cjm96@ast.cam.ac.uk
Key Publication	Gaia Early Data Release 3: Acceleration of the Solar System from Gaia astrometry

Project Description

With the high quality and quantity of the astrometric data in Gaia EDR3 (and now DR3 as well), it is possible to detect small relativistic effects in the astrometric solutions for distant, extragalactic objects. It is well known that the apparent position of a distant star depends on the observer's velocity; this effect is called aberration. The solar system is moving (at a few hundred km/s) with respect to the celestial reference frame. But this velocity is not constant; as the solar system orbits in the potential of the Milky Way, the acceleration towards the galactic centre is $\sim$1 cm/s/yr. Therefore, in the few years Gaia has been flying, the velocity of the solar system has changed by a few cm/s and the aberration from this appears as a dipole ("electric" type) in the velocity field of the proper motions of extragalactic objects directed towards the galactic centre. It's remarkable that a relativistic effect associated with such a small change in velocity can be detected!

This project will use vector spherical harmonics to flexibly model the proper motion vector field of distant quasars using a sum of low-order vector spherical harmonics. The focus will initially be on the Gaia EDR3 data as we aim to reproduce the main results from the key publication.

Project goals

The main goals of the project:

1. Work with the most recent Gaia DR3 data and, possibly, also incorporate VLBI radio data
2. The key result to be reproduced is the measurement of the acceleration (both magnitude and direction) of the solar system barycentre (with respect to the rest frame of the Universe).

Extension directions:

- Bayesian methods for the handling of outliers.
- More ambitiously, the project might also be extended by looking for (or placing upper limits on) other relativistic effects in the astrometric data. Prominent among these are ultra-low-frequency gravitational waves. These effects exist in the quadrupole sector of the VSH decomposition (as opposed to the dipole sector for the acceleration). The main other way of detecting nanohertz GWs is by pulsar timing arrays. This is very timely; in July 2023 all the major international PTA collaborations (NANOGrav, European PTA, Parkes PTA, and Chinese PTA) announced early evidence for what is probably a stochastic background of GWs generated by inspiralling supermassive black holes. For general interest, see https://www.bbc.com/news/science-environment-66039810

Prerequisites

Necessary: Python 3 familiarity Desirable: Some prior experience working with large data sets and MCMC methods would be desirable but not essential.

Reading List

KEY PUBLICATION:

- Gaia collaboration (2021) "Gaia Early Data Release 3: Acceleration of the Solar System from Gaia astrometry", A&A, Volume 649, A19, link.

OTHER USEFUL PAPERS:

- Titov, Lambert & Gontier (2011) "VLBI measurement of the secular aberration drift", A&A Volume 529, A91, link.
- Xu et al. (2012) "Reconsidering the International Celestial Reference System based on the effect of the secular aberration", IAU Joint Discussion 7: Space-Time Reference Systems for Future Research at IAU General Assembly-Beijing, link.
- Titov (2013) "The Secular Aberration Drift and Future Challenges for VLBI Astrometry", link.
- Truebenbach & Darling (2017) "The VLBA Extragalactic Proper Motion Catalog and a Measurement of the Secular Aberration Drift", ApJS 233 3, link.
- Charlot *et al.* (2020) "The third realization of the International Celestial Reference Frame by very long baseline interferometry", A&A, Volume 644, A159, link.

Data Access

The main data set for this project will be Gaia Data Release 3 (DR3).

https://www.cosmos.esa.int/web/gaia/data-release-3

You may need to create a free account in order to download large datasets.

13 Anomaly detection in Galaxy Zoo dataset using machine learning and active learning techniques

Proposer Name	Dr. Sireesha Chamarthi & Dr. Eduardo Gonzalez-Solares
Proposer Role and Affiliation	Data Validation Scientist & Senior Research Associate, Institute of Astronomy
Proposer Contact Email	sc2538@cam.ac.uk
Key Publication	Astronomaly: Personalised Active Anomaly Detection in Astronomical Data

Project Description

Traditional methods of anomaly detection in large astronomical datasets usually require manual inspection making them impractical for enormous data volumes generated by modern surveys. The key publication introduces **Astronomaly**, a personalized active learning framework that combines machine learning with human-in-the-loop interaction to efficiently detect relevant anomalies in large datasets. The Galaxy Zoo data that is a citizen science project, provides image dataset of high signal-to-noise, resolved galaxies to test *Astronomaly* on. The project will focus on using publicly available *Astronomaly* framework on Galaxy Zoo dataset to detect anomalies using machine learning based techniques.

Project goals

The main goal of the project are:

1. Data exploration and understanding anomalies in Galaxy Zoo dataset.
2. Explore machine learning approaches (Isolation Forest, Local outlier Factor) present in *Astronomaly* package.
3. Explore *Astronomaly* visualization interface to use t-SNE plot for anomaly detection and identification of clusters/outliers in the data.
4. Review a subset of anomalies and score them based on relevance to adjust anomaly rankings.
5. Using machine learning techniques, predict relevance scores for the remaining data based on the limited set of above labeled examples and finally predict the final anomaly scores.
6. Discuss the results and the difference in results with the paper. Discuss future work.

The student will be provided with the Galaxy zoo dataset and the model will be trained locally (GPU can also be used if required). The github repository of *Astronomaly* package has sample scripts to run the visualization of the interface. The student will be expected to write a report on the project and present the results.

Extension directions:

- Investigate other machine learning techniques other than the ones described in the paper
- Evaluate the results on various image datasets from other astronomy surveys.
- Explore other kinds of datasets like lightcurves to estimate anomalies using this approach.

Prerequisites

It will be useful for the student to have some knowledge of neural networks, Python stack (numpy, sklearn, matplotlib) and TensorFlow/Pytorch.

Reading List

- Astronomaly: Personalised Active Anomaly Detection in Astronomical Data
- Grad-Galaxy Zoo : Morphologies derived from visual inspection of galaxies from the Sloan Digital Sky Survey
- Anomaly detection: A survey

Data Access

Data access is available from the publicly available data challenge and also in the key publication linked above. It could be made readily available to the student if necessary somewhere else (HPC or local custer).

14 Interpretation of features in chest X-ray images using deep learning

Proposer Name	Dr. Eduardo Gonzalez-Solares
Proposer Role and Affiliation	Senior Research Associate, Institute of Astronomy
Proposer Contact Email	eglez@ast.cam.ac.uk
Key Publication	CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison

Project Description

Chest radiography is the most common imaging examination globally, critical for screening, diagnosis, and management of many life threatening diseases. Automated chest radiograph interpretation at the level of practicing radiologists could provide substantial benefit in many medical settings, from improved workflow prioritization and clinical decision support to large-scale screening and global population health initiatives. CheXpert (Chext eXpert) is a large dataset that contains 224,316 chest radiographs of 65,240 patients. The project will focus in using deep learning with the aim to use the CheXpert dataset to train a deep learning model that is able to predict the probability of 14 observations from X-ray images and compare the accuracy with radiologists.

Project goals

The main goal of the project are:

1. Perform data exploration and preprocessing of the CheXpert dataset.
2. Design and train a model on the X-ray images to predict the probability of each of the observations, i.e., given a X-ray chest image determine the most likely pathology.
3. Produce saliency maps to interpret the model predictions.
4. Compare the accuracy of the model with radiologists.
5. Discuss the results and the difference in results with the paper. Discuss future work.

The student will be provided with the CheXpert dataset and the model will be trained using a GPU. The student will also be provided with a set of Jupyter notebooks that will guide the student through the project but will need to be adapted to the specific goals of the project.

The student will be expected to write a report on the project and present the results. It ie expected that the student will use a Git repository to keep track of the progress of the project and the code used to train the model and produce the results. Code should be readable, documented and tested.

Extension directions:

- Investigate the use of techniques like test time augmentation and ensemble models to improve results.
- Create a web app that allows a user to upload an X-ray image and produce the probability of each observation plus the saliency map.
- Investigate the applicability and performance of other network architetures.

Prerequisites

It will be useful for the student to have some knowledge of neural networks, Python numeric stack (numpy, scipy, matplotlib) and TensorFlow.

Reading List

- CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning:
- Structured dataset documentation: a datasheet for CheXpert
- Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization

Data Access

Data access is available from the key project web page linked above and could be made readily available to the student if necessary somewhere else (HPC or local custer).

15 Remapping WMAP

Proposer Name	Dr Steven Gratton
Proposer Role and Affiliation	Associate Teaching Professor, DAMTP
Proposer Contact Email	stg20@cam.ac.uk
Key Publication	**The WMAP temperature maps and associated papers**

Project Description

NASA's WMAP satellite was launched in June 2001 and observed for nine years, producing maps of the microwave sky at frequencies of 23, 33, 41, 61 and 94 GHz. Much of the radiation observed was emitted when the Universe was only about four hundred thousand years old, a tiny fraction of the fourteen billion or so years old it is now. Such "Cosmic Microwave Background" (CMB) radiation is often described as "the afterglow of the Big Bang" and by studying its statistical properties and comparing them to model predictions we can learn about the Universe's composition, evolution and initial conditions.

In this project you will reproduce the processing of the "time-ordered data" from the WMAP satellite into sky maps of the radiation. You will learn about the application of iterative data-processing techniques to cosmological datasets and the handling of data on spherical surfaces. Such techniques form the basis of the mapmaking processes for current modern CMB telescopes, even with their much larger data volumes than WMAP.

Project goals

Main goals of the project:

1. To reproduce to some level one "individual differencing assembly" first year/DR1 temperature-only map from the calibrated DR1 time-ordered data using the simple map-making scheme presented in the first set of papers (in particular the data processing methods paper) and first explanatory supplement.

2. To present a detailed comparison between your map and the original one.

Extension directions:

- To go on to make maps using more data and/or more advanced algorithms as presented in the later releases. This could include attempting your own calibration, starting say from the DR5 uncalibrated time-ordered data.

- To make polarization maps in addition to temperature ones

- To investigate the performance of advanced map-making schemes, such as the "bilinear" method presented in arXiv:2210:02243, on WMAP data.

- To consider and/or develop an efficient GPU mapmaking implementation and compare its performance to a CPU-based one.

Prerequisites

Required:

- Ability to use python effectively (C1 MPhil module)

Desirable:

- Knowledge of C++ and/or Fortran (C++ MPhil course)
- HPC and GPU programming experience (C2 MPhil module)
- A cosmology course (e.g. MPhil Cosmology module)

Reading List

- See Sec. 2.2 of "First Year Wilkinson Microwave Anisotropy Probe (WMAP) Observations: Data Processing Methods and Systematic Error Limits", available from **LAMBDA here** for a detailed description of iterative mapmaking. Sec. 2.5 gives some refinements it would be nice to incorpate.
- The explanatory supplement, available from **LAMBDA here**, in particular Chapter 4, is important for understanding the time-ordered data.
- The **general LAMBDA page for WMAP** contains links to all relevant papers, data products and target results for this project. Following the evolution of the techniques employed for the different releases may be instructive.
- The full link for the bilinear mapmaking paper mentioned above is **arXiv:2210.02243**.

Data Access

- The resouces for this project will come from the"Legacy Archive for Microwave Background Data" (LAMBDA) website, given above. For example the first-year time-ordered data is found **on LAMBDA here**. You can use the "IDL" code and documentation provided on LAMBDA as a reference for your own work (e.g. for computing the "pointing" of the satellite's detectors for each observation).
- You may find the "healpy" python package for the handling and visualization of data on the sphere using the "healpix" pixelization scheme useful. See **the documentation here**.
- The astropy.io.fits package might be helpful for handing FITS files in python. See **the documentation here**.

16 Transforming Drug Discovery with Generative AI and Foundation Models: Enhancing High-Throughput Screening

Proposer Name	Dr Guang Yang
Proposer Role and Affiliation	Associate Professor, Biomedical Engineering and Imperial-X, Imperial College London
Proposer Contact Email	g.yang@imperial.ac.uk
Key Publication	Can Generative AI Replace Immunofluorescent Staining Processes? A Comparison Study of Synthetically Generated CellPainting Images from Brightfield

Project Description

High-throughput screening (HTS) is a crucial technique in drug discovery that enables the rapid evaluation of extensive compound libraries to assess their therapeutic potential. Cell image assays using fluorescence microscopy are particularly valuable within HTS, as they allow scientists to observe cellular responses by applying fluorescent markers to highlight specific structures, providing detailed insights into drug effects on cell morphology and function. However, HTS through cell imaging presents challenges, particularly in data acquisition and analysis.

The first challenge is the labor-intensive and cost-prohibitive process of biochemical labeling for fluorescence microscopy. This process is not only expensive but also time-consuming, which limits the scalability of HTS. Recent advances in generative models offer potential solutions. These models can convert brightfield images into synthetic fluorescence images, bypassing the need for costly markers. Despite their promise, these models currently lack generalizability, performing well on specific datasets but struggling across diverse experimental conditions. This limitation stems from the insufficient diversity of training datasets and model architectures, which restricts the broader application of these models in HTS.

Another key challenge arises from the growing reliance on phenotypic screening in drug discovery, which focuses on observing changes in single-cell behavior. High failure rates during clinical development have underscored the limitations of existing image analysis methods, which often depend on manual segmentation, traditional cell instance identification processes, and feature extraction methods. These methods are labor-intensive and error-prone, especially when handling high-dimensional data from large-scale screening.

Therefore, this project has two primary objectives. The first is to validate and improve generative models, aiming to develop robust algorithms capable of generalizing across a wide range of experimental datasets. This will expand the application of AI in HTS and reduce the need for expensive fluorescence labeling. The second objective is to leverage deep learning techniques to automate the segmentation and feature extraction of single-cell instances in complex imaging datasets. By applying deep learning, the project seeks to enable more accurate and efficient extraction of detailed morphological features from individual cells, significantly improving the precision of single-cell analysis and providing deeper insights into cellular responses. These advancements aim to enhance the efficiency, scalability, and precision of drug screening, leading to more successful therapeutic discoveries.

By undertaking this project, students will gain a comprehensive skill set at the intersection of generative AI, image analysis, and drug discovery applications, including:

Image-to-Image Generative Models: Developing and optimizing generative models for synthetic image generation in biomedical contexts. Instance Segmentation: Leveraging large foundation

model techniques to automate image segmentation. Feature Engineering and Clustering: Extracting and engineering relevant features from imaging data and applying clustering techniques to analyze cellular phenotypes and patterns.

Project Goals

Main goals of the project:

1. Create and improve image-to-image generative models capable of generating synthetic fluorescence images from brightfield images, with a focus on increasing the models' generalizability across diverse experimental datasets.
2. Utilize deep learning techniques to automate the segmentation of individual cells from complex biomedical images, reducing the need for manual, labor-intensive processes.
3. Implement advanced methods for extracting detailed morphological features from segmented cells and apply clustering algorithms to analyze and categorize cellular phenotypes.

Extension directions:

- Explore advanced synthesis techniques for generating multi-channel fluorescence images from bright field data.
- Explore foundation models such as SegmentAnything, SegmentAnything2 for microscopy image segmentation.
- Design predictive machine learning models to evaluate compound efficacy based on the synthesized and analyzed data.

Prerequisites

Required Skills:

- Basic knowledge with machine learning and deep learning models, such as Convolutional Neural Networks.
- Experience with Python programming and PyTorch.
- Experience in using HPC resources for large-scale data analysis.

Desirable Skills:

- Knowledge of cell isntance segmentation models such as MaskRCNN, CellDist, and CellPose.
- Understanding of single cell feature extraction pipeline with CellProfiler.

Reading List

- Artificial Immunofluorescence in a Flash: Rapid Synthetic Imaging from Brightfield Through Residual Diffusion
- Can Generative AI Replace Immunofluorescent Staining Processes? A Comparison Study of Synthetically Generated CellPainting Images from Brightfield
- CellProfiler 4: improvements in speed, utility and usability

Data Access

Instructions on how to access the data with necessary links, if any.

- CellPainting Dataset
- LightMyCells

17 Chemo-dynamical analysis of Milky Way's stellar populations with unsupervised multi-dimensional clustering

Proposer Name	Dr GyuChul Myeong
Proposer Role and Affiliation	Research Fellow, Institute of Astronomy
Proposer Contact Email	gm564@cam.ac.uk
Key Publication	Milky Way's Eccentric Constitutents with Gaia, APOGEE, and GALAH

Project Description

Large galaxies, such as our Milky Way, go through complex evolutionary phases governed by various different mechanisms. As a result, the present day Galaxy is a composed product of stars from different origins and epochs, although it is not easy to distinguish one star's origin from another individually. Stars formed from the same process and the same source share comparable chemical and dynamical properties which reflects their formation environment. To understand the complex evolutionary history and underlying formation mechanisms of the Galaxy, it is essential to know the chemical and dynamical properties of the stars produced from each process. Using a combination of detailed chemical and dynamical information from Gaia, APOGEE, and GALAH, we attempt to classify the Milky Way stars into an unspecified number of sub-groups with distinguishable chemo-dynamical trends. Gaussian mixture modelling is adopted as an unsupervised clustering method. The chemo-dynamical property of each identified component can help us to trace the Galaxy's evolutionary history. In specific we focus on identifying and studying the Milky Way's known stellar populations, such as GS/E (ex-situ population from a major galactic merger event the Milky Way experienced), Aurora (ancient population from early epoch of Milky Way evolution), Splash (dynamically heated population due to the GS/E merger), and Eos (result of a star-formation as a consequence of GS/E merger).

Project goals

Main goals of the project:

1. Understand the high-dimensional dataset, implement/test unsupervised clustering models
2. Apply model selection criteria based on the goodness-of-fit and complexity
3. Analyse the output result and interpret the physics behind the features uncovered from data

Extension directions:

- Apply dimensionality reduction and visualisation for high-dimensional data
- Explore the effect of data uncertainty in unsupervised clustering method

Prerequisites

Required:

1. Some coding skills in Python or other suitable programming languages

Desirable:

- Some knowledge related to Galactic Archaeology (there will be a minor module course) would be helpful but is not required

Reading List

- From dawn till disc: Milky Way's turbulent youth revealed by the APOGEE+Gaia data, Belokurov et al., 2022, MNRAS, 514, 689B
- Co-formation of the disc and the stellar halo, Belokurov et al, 2018, MNRAS, 478, 611B
- The biggest splash, Belokurov et al, 2020, MNRAS, 494, 3880B
- Extreme deconvolution: Inferring complete distribution functions from noisy, hetrogeneous and incomplete observations, Bovy et al., 2011, AnApS, 3.1657B

Data Access

The main datasets are publicly accessible (Gaia DR3, APOGEE DR17, GALAH DR3), but some of the necessary information requires further computations (e.g., orbital integration, energy calculation) which is beyond the scope of the project. The student can contact the PI (GyuChul Myeong; gm564@cam.ac.uk) for the dataset that can be used for the project. The main software required for the project (Extreme Deconvolution) is publicly available (https://github.com/jobovy/extreme-deconvolution). Alternatively, the student can try scikit-learn's Gaussian Mixture Modelling (https://scikit-learn.org/stable/).

18 Classical simulation of quantum circuits

Proposer Name	Prof Hamza Fawzi
Proposer Role and Affiliation	Professor of Applied Mathematics, DAMTP
Proposer Contact Email	hf323@cam.ac.uk
Key Publication	Classically estimating observables of noiseless quantum circuits

Project Description

Quantum computers offer the perspective of efficiently solving certain computational tasks that are hard on classical computers. The model of computation in quantum computers is one whereby a *quantum state* on n qubits is evolved through a sequence of local unitaries (so-called *quantum gates*). Concretely, a quantum state on n qubits is a vector with 2^n coefficients. Exactly simulating a quantum circuit is thus infeasible for moderately large n.

Despite the exponential dependence in number of qubits, several algorithms have been proposed to approximate quantum computation. In particular, the key publication linked above studies a recent algorithm based on the Pauli algebra. In the paper, some theoretical guarantees about this algorithm are proven, and the numerical efficiency of the algorithm is demonstrated on a circuit with 64 qubits. The goal of the project is to reproduce the numerical experiments in the paper. (The theoretical part of the paper is not relevant for this project.)

Project goals

Main goals of the project:

1. Implement the algorithm described in Section III of the paper to simulate any quantum circuit. We recommend using the Python packages Cirq (and, if needed, the package OpenFermion) for the manipulation of quantum circuits and Pauli operators.
2. Reproduce numerical experiments similar to Figure 2 in the key publication

Extension directions:

- Run the algorithm on the recent IBM experiment (, see also .

Prerequisites

This project requires a certain level of mathematical maturity, i.e., good knowledge of linear algebra and (basic) probability. Knowledge of quantum computation is helpful but not needed.

19 Numerical pology for intractable field theories and gravity

Proposer Name	Haoyang Ye
Proposer Role and Affiliation	Research Associate, Cavendish Laboratory
Proposer Contact Email	hy297@cam.ac.uk
Key Publication	`arXiv:2406.09500 [hep-th]` (under review for Phys. Rev. D)

Project description

The word 'data' doesn't have to be restricted to the numbers that come out of a telescope or a particle detector. In this project, we'll use numbers to explore the analytic properties of different theories of gravity, in a novel approach that doesn't presuppose any knowledge of general relativity (GR).

A quick scan of the `[gr-qc]` and `[hep-th]` arXiv on any day reveals a kind of 'cottage industry' for producing alternatives to Einstein's general theory of relativity. The problem is that new models are produced at a rate far exceeding the capacity of the community to thoroughly test or regulate them. An obvious test is comparison with precision cosmology data, but this is a huge waste of time if the new model is not even theoretically consistent. In fact, most theories one can write down are sick or unstable, as will be explained below, but these diagnoses can be incredibly expensive to make by analytic investigation of the classical or quantum field theory. For this reason, the future of gravitational theory is likely one where the successful researcher proposes a new category of model (based, for example, on gauging a symmetry group), and then a supercomputer searches over a very large parameter space for any healthy instances. One primarily imagines such surveys to be grounded in computer algebra. But, if there were *numerical* approaches instead, then could theoretical physics eventually become a data-intensive science?

In an inconsistent gravity model, the spectrum of quantum particles (gravitons, for example) predicted by the theory may contain modes with negative energy (ghosts) or imaginary mass (tachyons). A computer algebra system was recently developed to diagnose such pathologies, applicable to a very wide class of theories. This includes any theory which (in the limit of weak fields) is associated with a *free* (i.e. quadratic, non-interacting) action

$$S_{\mathrm{F}} = \int \mathrm{d}^4x \, \sum_X \zeta_{\mu_X} \Big[\sum_Y \mathcal{O}^{\mu_X}{}_{\nu_Y} \zeta^{\nu_Y} - j^{\mu_X} \Big], \tag{1}$$

where (1) contains the following ingredients:

1. The (e.g. gravitational) fields ζ_{μ_X} are real tensors. Distinct fields carry the index X, each field has some collection of spacetime indices μ_X, perhaps with some symmetry.

2. The wave operator $\mathcal{O}^{\mu_X}{}_{\nu_Y}$ is a real differential operator constructed from $\eta_{\mu\nu}$ and ∂_μ, linearly parameterised by a collection of coupling coefficients.

3. The (e.g. matter stress-energy tensor) source currents j^{μ_X} are conjugate to the fields ζ_{μ_X}. They encode all external interactions, whilst keeping the external dynamics completely anonymous.

The algebraic version of the consistency algorithm is computationally demanding (see Fig. 1) but numerical versions are expected to be very much faster. Numerics could (eventually) be paired with nested sampling to explore the propagator poles of arbitrarily complicated quantum

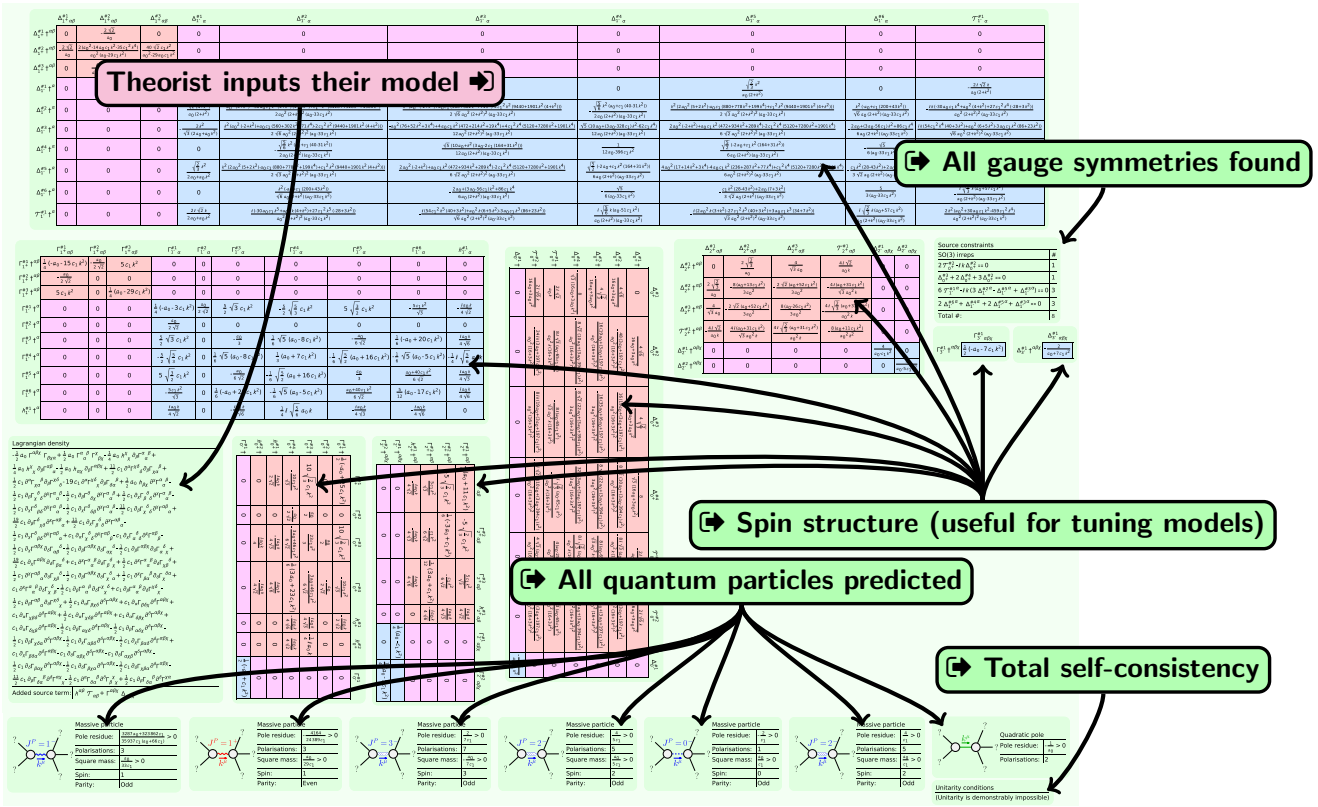

Figure 1: This is a vector graphic: zoom in for details. What we're showing here is the output from the standard computer algebra version of the particle spectrum algorithm. The input theory is a Lagrangian density, referring to the small perturbations of fields $h_{\mu\nu}$ and $\Gamma^{\mu}{}_{\nu\sigma}$ around the Minkowski vacuum — these are examples of ζ_{μ_X} in (1). The fields and their derivatives appear in bilinear operators (i.e. terms), which are parameterised by couplings a_0 and c_1 — these operators define $\mathcal{O}^{\mu_X}{}_{\nu_Y}$ in (1). The sources $T^{\mu\nu}$ and $\Delta_{\mu}{}^{\nu\sigma}$ correspond to the j^{μ_X} in (1). The program obtains the spin-parity matrices, whose elements are functions of a_0, c_1 and the norm of the particle four-momentum k. These matrices determine the gauge symmetries of the theory, and the spectrum of quantum particles (their masses, spins and parities). This particular theory (a candidate alternative to GR) is found analytically to be *inconsistent* for general a_0 and c_1, when all this information is combined with a lot of intermediate algebra. Your job in this project is to start from the matrices, and compute the same information for *numerical* values of a_0 and c_1.

field theories. Such a tool could be very effective in eliminating new models as soon as they are proposed. This project will involve the numerical confirmation of known analytic results, using high-performance computing. Besides numerical fluency, you will have the opportunity to pick up supercomputing skills, and we can introduce you to good programming practices (from version control to copilot and vim). Novel results or especially useful code may lead to inclusion on the first paper in an upcoming series about numerical pology.

Project goals

The aim of the project is to implement a numerical version of the spin-projection/saturated propagator algorithm. This is a new field: in the long term we expect to use these 'quick-and-dirty' numerical techniques to efficiently constrain the space of possible gravity models, and discourage irresponsible model-building in the theoretical physics community. There are broadly two directions to focus on:

- Theoretical development of what it means to make the particle spectrum of a model 'fuzzy' or 'smeared', in the sense that the appearance of new propagator poles or gauge symmetries within certain regions of the parameter space should be made continuous. Such features are of course discrete properties of the theory, but a continuous realisation is needed to drive numerics in the right direction.
- Ghost/tachyon cartography of successively more complicated theories. There is a huge range to chose from, for which we already know the analytic results. Alternatively, with a rudimentary working numerical code, we could explore a theory whose particle spectrum is not known (and, by Galois theory, cannot be known) analytically.

The admixture of these goals will depend, to a certain extent, on the student. With a theoretical focus, we may work more on the algorithm. With a spectrographic focus, we might explore very complicated and popular modified gravity models. Given the two directions stipulated above, the concrete goals of the project are:

- Define a numerical version (sophisticated or otherwise) of the spin-projection/saturated propagator algorithm.
- Model 1: Identify the Maxwell and Proca theories from the three-parameter family of vector interactions.
- Model 2: Identify GR from the five-parameter family of metric interactions.
- Model 3: Identify Fierz–Pauli massive gravity from the five-plus-two-parameter extension of Model 2.

Extra goals include:

- Model 4: Perform ghost/tachyon cartography of the (up to) 28-parameter metric affine theory of gravity.
- Any novel developments in the theory of fuzzy particle spectra.

Prerequisites

The following skills are required for this project:

- Fluency with either Python or C++.
- Fluency with UNIX-like operating systems.
- Undergraduate-level understanding of rudimentary classical field theory (tensor calculus Lorentz invariance, variational principles, Noether theorems, etc.).

The following skills are desirable for this project:

- Fluency with the Wolfram Language.
- Fluency with SLURM or other job scheduling systems.
- Any knowledge of GR is, of course, a bonus, but it is not actually necessary. You can afford to learn about GR 'on the fly' during the project.

Reading list

Don't worry if you don't understand the following papers much: this is primarily a numerical project. It is expected that the cartography will primarily be implemented using basic Monte Carlo techniques. Very ambitious students may wish to explore the more sophisticated approach of *nested sampling*, an introduction to which can be found at [1]. Unlike in the Monte Carlo case, we cannot guarantee a successful project with this latter approach, since it is an area of active research in our group. The computer algebra version of the particle spectrum algorithm is introduced at [2]. As stated above, it is important to have some notion of classical field theory, as it is usually introduced at the undergraduate level. Examples of classical field theories include Maxwell's electromagnetism. However, quantum field theory and GR are not actually requirements for the project. For those who wish to 'go the extra mile' and learn about these concepts, the following papers are recommended:

- For the theoretical background, the papers [3–12] provide lots of context.
- For examples of (very clever) researchers performing the particle spectrum algorithm by hand, see [13–26] and [6, 10].
- For a recent computer algebra implementation of the algorithm, which predates the main paper for this project, see [27–30].

Data access

The project is associated with the following public repositories:

- The PSALTer software.
- Pre-computed spin-parity matrices for four models.

[1] G. Ashton *et al.*, Nested sampling for physical scientists, Nature **2**, 10.1038/s43586-022-00121-x (2022), arXiv:2205.15570 [stat.CO].

[2] W. Barker, C. Marzo, and C. Rigouzzo, PSALTer: Particle Spectrum for Any Tensor Lagrangian, (2024), arXiv:2406.09500 [hep-th].

[3] R. E. Behrends and C. Fronsdal, Fermi Decay of Higher Spin Particles, Phys. Rev. **106**, 345 (1957).

[4] C. Fronsdal, On the theory of higher spin fields, Il Nuovo Cimento **9**, 416 (1958), https://doi.org/10.1007/BF02747684.

[5] S.-J. Chang, Lagrange Formulation for Systems with Higher Spin, Phys. Rev. **161**, 1308 (1967).

[6] A. Aurilia and H. Umezawa, Theory of high-spin fields, Phys. Rev. **182**, 1682 (1969).

[7] L. P. S. Singh, COVARIANT PROPAGATORS FOR MASSIVE ARBITRARY SPIN FIELDS, Phys. Rev. D **23**, 2236 (1981).

[8] E. S. Fradkin and A. A. Tseytlin, CONFORMAL SUPERGRAVITY, Phys. Rept. **119**, 233 (1985).

[9] A. Y. Segal, Conformal higher spin theory, Nucl. Phys. B **664**, 59 (2003), arXiv:hep-th/0207212.

[10] L. Buoninfante, Ghost and singularity free theories of gravity, (2016), arXiv:1610.08744 [gr-qc].

[11] A. P. Isaev and M. A. Podoinitsyn, Two-spinor description of massive particles and relativistic spin projection operators, Nucl. Phys. B **929**, 452 (2018), arXiv:1712.00833 [hep-th].

[12] D. Hutchings and M. Ponds, Spin-(s, j) projectors and gauge-invariant spin-s actions in maximally symmetric backgrounds, JHEP **07**, 292, arXiv:2401.04523 [hep-th].

[13] P. Van Nieuwenhuizen, On ghost-free tensor lagrangians and linearized gravitation, Nucl. Phys. B **60**, 478 (1973).

[14] D. E. Neville, A Gravity Lagrangian With Ghost Free Curvature**2 Terms, Phys. Rev. D **18**, 3535 (1978).

[15] D. E. Neville, Gravity Theories With Propagating Torsion, Phys. Rev. D **21**, 867 (1980).

[16] E. Sezgin, Class of Ghost Free Gravity Lagrangians With Massive or Massless Propagating Torsion, Phys. Rev. D **24**, 1677 (1981).

[17] E. Sezgin and P. van Nieuwenhuizen, New Ghost Free Gravity Lagrangians with Propagating Torsion, Phys. Rev. D **21**, 3269 (1980).

[18] R. Kuhfuss and J. Nitsch, Propagating Modes in Gauge Field Theories of Gravity, Gen. Rel. Grav. **18**, 1207 (1986).

[19] G. K. Karananas, The particle spectrum of parity-violating Poincaré gravitational theory, Class. Quant. Grav. **32**, 055012 (2015), arXiv:1411.5613 [gr-qc].

[20] G. K. Karananas, *Poincaré, Scale and Conformal Symmetries Gauge Perspective and Cosmological Ramifications*, Ph.D. thesis, Ecole Polytechnique, Lausanne (2016), arXiv:1608.08451 [hep-th].

[21] E. L. Mendonça and R. Schimidt Bittencourt, Unitarity of Singh-Hagen model in D dimensions, Adv. High Energy Phys. **2020**, 8425745 (2020), arXiv:1902.05118 [hep-th].

[22] R. Percacci and E. Sezgin, New class of ghost- and tachyon-free metric affine gravities, Phys. Rev. D **101**, 084040 (2020), arXiv:1912.01023 [hep-th].

[23] C. Marzo, Ghost and tachyon free propagation up to spin 3 in Lorentz invariant field theories, Phys. Rev. D **105**, 065017 (2022), arXiv:2108.11982 [hep-ph].

[24] C. Marzo, Radiatively stable ghost and tachyon freedom in metric affine gravity, Phys. Rev. D **106**, 024045 (2022), arXiv:2110.14788 [hep-th].

[25] Y. Mikura, V. Naso, and R. Percacci, Some simple theories of gravity with propagating torsion, Phys. Rev. D **109**, 104071 (2024), arXiv:2312.10249 [gr-qc].

[26] Y. Mikura and R. Percacci, Some simple theories of gravity with propagating nonmetricity, (2024), arXiv:2401.10097 [gr-qc].

[27] Y.-C. Lin, M. P. Hobson, and A. N. Lasenby, Ghost and tachyon free Poincaré gauge theories: A systematic approach, Phys. Rev. D **99**, 064001 (2019), arXiv:1812.02675 [gr-qc].

[28] Y.-C. Lin, M. P. Hobson, and A. N. Lasenby, Power-counting renormalizable, ghost-and-tachyon-free Poincaré gauge theories, Phys. Rev. D **101**, 064038 (2020), arXiv:1910.14197 [gr-qc].

[29] Y.-C. Lin, M. P. Hobson, and A. N. Lasenby, Ghost- and tachyon-free Weyl gauge theories: A systematic approach, Phys. Rev. D **104**, 024034 (2021), arXiv:2005.02228 [gr-qc].

[30] Y.-C. Lin, *Ghost and tachyon free gauge theories of gravity: A systematic approach*, Ph.D. thesis, Cambridge U. (2020).

20 Determining uncertainties in parton densities

Proposer Name	Dr James Moore
Proposer Role and Affiliation	Director of Studies in Physics, Lucy Cavendish College
Proposer Contact Email	jmm232@cam.ac.uk
Key Publication	On the determination of uncertainties in parton densities

Project Description

Parton distributions are the functions which parametrise the structure of protons in terms of their elementary constituents (quarks and gluons); they are usually determined using precise data from collider experiments like the LHC at CERN. Fits of these functions use a range of methods, including various techniques for uncertainty propagation, and various assumed functional forms for the functions themselves (including neural network parametrisations). This project will study, in a toy scenario, the faithfulness of these different uncertainty propagation methods, together with the effect of using different functional forms.

Project goals

A successful project will demonstrate a basic understanding of parton distributions and their use in collider physics (a purely intuitive understanding of their use is fine - a detailed understanding of the quantum chromodynamics background is unnecessary).

Main goals of the project:

1. Investigate, in a toy model, various methods used for uncertainty propagation from experimental data onto parton distribution functions, as described in 2206.10782. This will include a reproduction of Figure 2 from this paper.
2. Investigate, in the same toy model, the effect of using a power law functional form for the parton distributions vs using a neural network functional form for the parton distributions, as described in the same reference. This will include a reproduction of Figures 5 and 8 in the given reference.

Extension directions:

- Further investigation may be carried out beyond the scope of the paper, for example, by including the effect of using toy data modelled by parton distributions entering quadratically in Eq. (28) rather than linearly.

Prerequisites

No prerequisites.

Reading List

- The primary reference is On the determination of uncertainties in parton densities. This focusses specifically on toy models for parton distribution functions, and the successful candidate need not read beyond this paper.
- For interested candidates, a recent realistic fit of parton distributions using neural networks is given in The path to proton structure at one-percent accuracy.

Data Access

This project will rely on the candidate generating their own pseudo-data according to Eq.(28) of 2206.10782 (and possibly extending this pseudo-data generation to the quadratic case).

21 Federated Learning and Class Imbalances

Proposer Name	Joshua Kaggie
Proposer Role and Affiliation	Senior Research Associate, Department of Radiology
Proposer Contact Email	jk636@cam.ac.uk
Main paper	https://github.com/litian96/FedProx Li, Tian, et al. Proceedings of Machine learning and systems 2 (2020): 429-450

Project Description

Federated Learning (FL) has emerged as a powerful approach that enables collaborative distributed model training without the need for data sharing. However, FL grapples with inherent heterogeneity challenges leading to issues such as stragglers, dropouts, and performance variations. Selection of clients to run an FL instance is crucial, but existing strategies introduce biases and participation issues and do not consider resource efficiency. Communication and training acceleration solutions proposed to increase client participation also fall short due to the dynamic nature of system resources.

FedProx and FLOAT optimise resource utilisation dynamically for meeting training deadlines, and mitigates stragglers and dropouts through various optimisation techniques. leading to enhanced model convergence and improved performance. FLOAT, in particular, leverages multi-objective Reinforcement Learning with Human Feedback (RLHF) to automate the selection of the optimisation techniques and their configurations, tailoring them to individual client resource conditions. Moreover, FLOAT seamlessly integrates into existing FL systems, maintaining non-intrusiveness and versatility for both asynchronous and synchronous FL settings.

Project goals

Main goals of the project:

1. Implement a DL classification or denoising algorithm on images within a publicly available database, such as MNIST or CIFAR, and a medical imaging dataset of your choice and within appropriate educational data standards/licencing (e.g., breast MRIs within the TCIA).
2. Implement the DL models in a federated learning fashion, such that they could be performed across two or more computers, or simulated computers.
3. Address class imperfections or imbalances within the federated learning environments, replicating one of the works listed in the Reading List below. These imbalances could be in the amount of noise changing between centres (simulated computers), or in the number of class types available.
4. Implement

Extension directions:

- Demonstrate FL across a larger number of simulated computers than two.
- Implement FL work with a large language model, and address class imbalances or implement a multi-client method for RLHF.
- Train starting with an existing foundational model or more advanced DL architectures, and demonstrate improved results.

Prerequisites

Required: Python proficiency, including standard toolkits (Pandas, NumPy, MatPlotLib); ML proficiency, as gained through this course (Scikit-learn, PyTorch); HPC access, or access to other high end CUDA enabled environment.
Desirable: Experience or ability to learn networking libraries.

Reading List

"Python TCP/IP libraries: A Review"

The papers associated with these: https://github.com/FangXiuwen/Robust_FL
https://github.com/litian96/FedProx
https://github.com/AFKD98/FLOAT/

Other FL examples are:
https://github.com/TiiCHE/fedhadamard
https://github.com/qiu-zheng/FedNova
https://github.com/zheng-xiao/fedamp

Data Access

The MNIST database of handwritten digits

CIFAR-10

The Cancer Imaging Archive

22 Monitoring Vegetation Trends as a Result of Climate Change.

Proposer Name	Dr Hugo Lepage & Prof Crispin Barnes
Proposer Role and Affiliation	Postdoctoral Research Associate
Proposer Contact Email	hl407@cam.ac.uk
Key Publication	Greening and Browning Trends on the Pacific Slope of Peru and Northern Chile

Project Description

Climate change and its effects are among the most critical global challenges today. One significant response to changing climate drivers is the alteration of vegetation patterns, which impacts natural ecosystems.

Recent research by the Environmental Physics Group at the Cavendish Laboratory has highlighted significant greening (increased plant activity) on the western slope of the Peruvian Andes. Using satellite multispectral data, this study analysed vegetation indices across the South American Andes over the past 20 years, confirming statistically significant trends through time-series analysis.

This project will involve rewriting and reproducing the code used to download and process large satellite imagery datasets. Special focus will be placed on handling non-Euclidean geometry, as the Earth's spherical nature must be accounted for when examining data on a continental scale.

Potential extensions to the project include comparing vegetation trends in other regions exhibiting unusual greening or browning, as well as correlating these changes with planetary climate factors such as ocean currents, wind patterns, and precipitation.

Project goals

Main goals of the project:

1. Set up a python script to download process and analyse satellite (MODIS) multispectral data.
2. Calculate a statistically significant greening or browning trend for each pixel time series.
3. Calculate vegetation greening and browning correlations with climate drivers such as precipitation, land surface temperature, ocean temperature, etc.

Extension directions:

- Generalise the methodology to investigate global, rather than local, vegetation response.
- Adapt the methodology to use different satellites (Landsat, Sentinel-2, etc.)

Prerequisites

Familiarity with Python3 is necessary. Familiarity with high performance computing or GPU parallelisation is desirable, but not necessary.

Reading List

- https://www.mdpi.com/2072-4292/15/14/3628
- https://www.mdpi.com/2072-4292/12/15/2418

Data Access

Satellite Datasets are accessible through the Google Earth Engine Catalog https://developers.google.com/ear
engine/datasets

23 Machine Learning and Feature Selection of Imaging-Based Biomarkers for Tumour Classification

Name	Dr Lorena Escudero Sánchez
Role and Affiliation	Senior Research Associate, Department of Radiology
Contact Email	les44@cam.ac.uk
Key Publication	Feasibility and sensitivity study of radiomic features in photoacoustic imaging of patient-derived xenografts

Project Description

Radiomics[1] is an active field of research describing the extraction of mineable features from radiological images such as Computed Tomography (CT) or Magnetic Resonance (MRI). Unlike qualitative image evaluation, which requires a trained reader to assess the images (i.e. presence of disease), radiomics allows for a large number of mathematical measurements to be extracted from standard-of-care scans of different imaging modalities. These features provide quantitative measurements of tissue characteristics, related to shape, intensities and heterogeneity or texture, providing objective, reader-independent non-invasive biomarkers. Machine Learning (ML)-based models using such features are becoming a promising tool for tumour classification and therapy response assessment in oncological research. However, there is still a need for standardisation criteria and further validation of feature robustness with respect to imaging acquisition and reconstruction parameters, as well as interpretability, before they can be implemented in the clinical setting.

Project goals

In this project, students will perform data analyses of radiomic features pre-computed using a standard Python library (PyRadiomics) to reproduce the Key Publication above, which uses a small dataset of preclinical images obtained with a novel imaging modality: photoacoustic. It involves:

1. Background research: including comprehensive review of the original paper and related literature.
2. Data acquisition: available on GitHub, for features and tumour model (images will not be available).
3. Statistical analysis: reproducing the steps of sensitivity analysis using classical statistical methods (ANOVA).
4. Feature selection and machine learning analysis: reproducing the model discrimination analyses detailed in the paper (using Random Forest Classifier, Gradient Boosting Classifier and Support Vector Machines).
5. Interpretability of results: reproduce the paper studies using SHAP[2] plots.
6. Coding: write their own code for the various analyses.
7. Data visualisation: reproduce the relevant figures in the paper.
8. Scientific writing: explain methodology, results and conclusions in the final essay.

The main objective of the project is to reproduce the paper above and the focus is to harness statistical and classical data science and machine learning skills. For students who want to work on an extension after successfully completing this part, the aim will be to develop similar approaches (feature selection and nodule classification) using a publicly available clinical dataset, which will involve working with Computed Tomography (CT) images of real lung cancer patients [3]. If students want to focus on the second part only after reviewing the material and identifying a suitable paper to reproduce, this can be discussed at the start of the project.

Prerequisites

Students should take the Medical Imaging minor, or have equivalent pre-existing knowledge. Strong coding skills (`scipy`, `numpy`) are required. Notice that some of the results of the paper in the first part were implemented in MATLAB, but students can replicate the results with existing Python libraries.

Reading List

[1] Gillies RJ, Kinahan PE, Hricak H. Radiomics: Images Are More than Pictures, They Are Data. Radiology. 2016 Feb;278(2):563-77. doi: 10.1148/radiol.2015151169. Epub 2015 Nov 18. PMID: 26579733; PMCID: PMC4734157.

[2] An introduction to explainable AI with Shapley values

[3] Armato SG 3rd, McLennan G, Bidaut L, et al. The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI): a completed reference database of lung nodules on CT scans. Med Phys. 2011 Feb;38(2):915-31. doi: 10.1118/1.3528204. PMID: 21452728; PMCID: PMC3041807.

In addition, students should add to their reading list the Key Publication above and references within.

Data Access

The necessary data (features already extracted from the photoacoustic images) is available on GitHub.

24 UNet-based Segmentation of Kidney Tumours in Computed Tomography Images

Name	Dr Lorena Escudero Sánchez
Role and Affiliation	Senior Research Associate, Department of Radiology
Contact Email	les44@cam.ac.uk
Key Publication	An attempt at beating the 3D U-Net

Project Description

The automated segmentation of organs and tumours is a key step in radiological image analysis for which many Deep Learning (DL)-based methods are being developed. Amongst the most successful ones are methods based on the UNet architecture or variants thereof. KiTS is a grand challenge to accelerate the development of reliable kidney and kidney tumour semantic segmentation methodologies. There have been three versions so far of this challenge, with increasing number of computed tomography (CT) cases released in 2019, 2021 and 2023. This project will use the released training dataset in the 2019 version, both images and manual segmentations/delineations for 210 cases provided in NIfTI format. The aim of the project is to test three different versions of 3D UNet (plain, with residual connections, and with pre-activation) and reproduce similar Dice scores (measuring overlap of the true and predicted segmentations) to those in the paper linked above, which was the winner of the KiTS19 challenge. The project involves handling and manipulating a dataset of medical imaging of real kidney cancer patients, basic data augmentation, network training (which involves a significant amount of GPU hours) and evaluation of appropriate metrics.

Project goals

This project is a Deep Learning focused project in which students will write their own UNet architecture using the `PyTorch` framework, and they will learn to use the High Performance Computing (HPC) service of the University for training. It involves:

1. Background research: including comprehensive review of the original paper and of literature related to CT and UNets.
2. Data acquisition and medical image format handling: available on GitHub, for images and annotations, in NIfTI format.
3. Coding: write their own code for the various implementations of the UNet-based architectures, training, validation and analysis.
4. Training and validation: involves hyper-parameter tuning etc. Results should be approximately the same than those presented in the paper.
5. Data visualisation: create figures showing predictions and ground truth, as well as summaries of comparisons of implementations.
6. Scientific writing: explain methodology, results and conclusions in the final essay.

The main objective of the project is to reproduce the paper above and the focus is to harness deep learning skills. If time allows, further work can be done for example by extending the test of such algorithms to a larger dataset with the training sets for KiTS21 and KiTS23 or by exploring other algorithms. Based on the student's interest and feasibility, other automated segmentation methods can be discussed for lung nodules in CT as an alternative to KiTS.

Prerequisites

Students should take the Medical Imaging minor, or have equivalent pre-existing knowledge. Strong coding skills (`PyTorch`, `scipy`, `numpy`) are required, as well as experience with scheduling tools for job submission at the HPC.

Reading List

Students should start their reading with the Key Publication above and references within. In addition:

[1] O. Ronneberger, P. Fischer and T. Brox, U-Net: Convolutional Networks for Biomed- ical Image Segmentation, MICCAI 2015, doi: 10.1007/978-3-319-24574-4 28

[2] Cicek, O., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3d u-net: learning dense volumetric segmentation from sparse annotation. In: International conference on medical image computing and computer-assisted intervention. pp. 424-432. Springer (2016)

Data Access

The necessary data is available on GitHub.

25 A deep-NN based tool for the simultaneous fit of Standard Model EFT Wilson coefficients and the proton's subnuclear structure

Proposer Name	Prof. Maria Ubiali
Proposer Role and Affiliation	Professor at DAMTP
Proposer Contact Email	mu227@cam.ac.uk
Key Publication	SIMUnet: an open-source tool for simultaneous global fits of EFT Wilson coefficients and PDFs

Project Description

The success of the ambitious programme of the Large Hadron Collider (LHC) relies on achieving the highest possible accuracy in the experimental measurements and in the corresponding theoretical predictions. At the same time the availability of statistically robust tools capable of yielding global interpretations of all subtle deviations from the Standard Model (SM) that the data might indicate is paramount. The lack of evidence for additional particles at the LHC or at other colliders so far suggests that any particles beyond the Standard Model (BSM) may be significantly heavier than the energy scale probed by the LHC. Hence, the effects of heavy BSM particles may be approximated by integrating them out to obtain higher-dimensional interactions between SM quantum fields. The SM may then be seen as an effective field theory (EFT) and can be supplemented by higher-dimensional operators that are suppressed by inverse powers of new particles' mass/energy scale.

The Standard Model Effective Field Theory (SMEFT) is a powerful framework to constrain, identify, and parametrise potential deviations with respect to SM predictions. It allows for the interpretation of experimental measurements in the context of BSM scenarios featuring heavy new particles while minimising assumptions on the nature of the underlying UV-complete theory. The determination of SMEFT Wilson coefficients from a fit of LHC data, like the determination of SM precision parameters from LHC data, might display a non-negligible interplay with the input set of Parton Distribution Functions (PDFs) used to compute theory predictions. PDFs are the functions parametrising the subnuclear structure of the proton in terms of quarks and gluons. These functions are also extracted from experimental data.

This project will use a deep neural network (NN) to flexibly model the proton structure and to simultaneously fit it alongside the Wilson coefficients of the SMEFT expansion that are relevant for the observables included in the simultaneous fit done with the public SimuNET code. The focus will initially be on the data already included in the analysis presented in the the key publication, with the scope of reproducing those displayed in Section 4 of the key publication, and can optionally be extended to include more datasets and check the stability of the results upon the addition of those.

Project goals

Main goals of the project:

1. Investigate the method used in SimuNET for the simultaneous fit of PDFs and SMEFT coefficients.

2. Reproduce the results of the simultaneous fit of PDFs and SMEFT coefficients in a global analysis of Higgs, top, diboson and electroweak sectors.

3. Explore the hyper-parameters of the deep learning model and explicitly check the dependence of the results on the simultaneous fit of PDFs and SMEFT Wilson coefficients on those.

Extension directions:

- Extend the analysis presented in the paper by adding new experimental data in the fit.
- Check the stability of the results.

Prerequisites

Required:

- Familiarity with Python and machine learning libraries.

Desirable:

- Some background on the Standard Model of particle physics.
- Some understanding of Effective Field Theories.

Reading List

- "A new generation of simultaneous fits to LHC data using deep learning" by S. Iranipour and M. Ubiali, https://arxiv.org/abs/2201.07240. The first paper where the idea of the methodology outlined in the key paper was brought forward.
- "SMEFiT: a flexible toolbox for global interpretations of particle physics data with effective field theories" by T. Giani, G. Magni and J. Rojo, https://arxiv.org/pdf/2302.06660. The paper gives a good account of the methodologies used in the fits of SMEFT Wilson coefficients.
- "The path to proton structure at 1% accuracy" by R. D. Ball et al, https://arxiv.org/pdf/2109.02653. The paper describes in many details the deep-NN methodology behind the NNPDF fits of Parton Distribution Functions.

Data Access

- Public repository of the SimuNET code
- SimuNET documentation
- NNPDF open-source code

26 Symbolic Regression with Learned Concept Libraries

Proposer Name	Miles Cranmer
Proposer Role and Affiliation	Assistant Professor, DAMTP & Institute of Astronomy
Proposer Contact Email	mc2473@cam.ac.uk
Key Publication	Symbolic Regression with a Learned Concept Library

Project Description

This project aims to reproduce results from the paper "Symbolic Regression with a Learned Concept Library," which presents an approach to combining symbolic regression with large language models (LLMs) to create interpretable symbolic models. Instead of using predefined datasets like the Feynman equations, students will generate **synthetic equations** using random symbolic expressions and apply the LaSR method to discover these symbolic relations.

Students will also use a local LLM such as **LLaMA-3.2** for concept generation and experiment with modifying the concept library for symbolic discovery. The use of a local model allows the project to be developed and run entirely on a laptop.

Project goals

Main goals of the project:

1. Use the provided LaSR framework with PySR to discover synthetic symbolic relations generated randomly.
2. Experiment with a local LLM (LLaMA-3.2) to explore how modifying the concept library impacts the model's performance in symbolic discovery.
3. Evaluate and report on how well LaSR performs in comparison to traditional symbolic regression methods.
4. Explore the effect of different prompting strategies on the performance of LaSR.

Extension directions

- Experiment with modifications to the concept generation process.
- Apply LaSR to new discoveries in a dataset of your choice.

Prerequisites

Required:

- Familiarity with Python and machine learning libraries.

Desirable:

- Experience with Julia.
- Experience with open-source LLMs and frameworks for using them.
- Understanding of symbolic regression.

Reading List

- Symbolic Regression with a Learned Concept Library (Key Paper)
- PySR paper
- LLaMA: Open and Efficient Foundation Models for Language

Data Access

The students will generate synthetic equations using random symbolic expressions, so no external dataset is required. The PySR and LaSR framework will assist with symbolic regression tasks.

27 Symbolic Distillation of Neural Networks

Proposer Name	Miles Cranmer
Proposer Role and Affiliation	Assistant Professor, DAMTP & Institute of Astronomy
Proposer Contact Email	mc2473@cam.ac.uk
Key Publication	Discovering Symbolic Models from Deep Learning with Inductive Biases

Project Description

This project aims to reproduce the main results from the paper "Discovering Symbolic Models from Deep Learning with Inductive Biases." The paper demonstrates a technique to translate neural networks into analytic equations, focusing on toy and complex physical systems for test cases, particularly using graph neural networks as the architecture due to their functional similarity to physics.

The project is important as it explores how neural networks can be interpreted to rediscover physical laws, with practical applications in model discovery and physics-informed machine learning. By completing this project, students will gain an understanding of how to impose structure on neural networks and combine them with symbolic methods, contributing to advances in interpretable AI.

Project goals

Main goals of the project:

1. Implement a graph neural network with a single message passing step, and train and evaluate it on toy n-body systems.
2. Measure the latent representations inside the neural network, specifically the activations of the message functions.
3. Apply symbolic regression using PySR to approximate the message functions with analytic equations.
4. Explore the effect of a bottleneck versus sparse regularization on a wide message passing space.
5. Evaluate if meaningful physical laws can be recovered and whether those laws generalize better.

Extension directions

- Experiment with Hamiltonian Neural Networks or Cosmology-related tasks as a next step after reproducing the main experiment.
- Explore a different direction, such as applying the methods to a different type of physical system or architecture.

Prerequisites

Required:

- Experience with neural networks, particularly graph neural networks.
- Familiarity with Python and deep learning frameworks (e.g., PyTorch, TensorFlow).
- Some knowledge of classical mechanics or physics would be helpful.

Desirable:

- Familiarity with symbolic regression techniques (e.g., PySR).
- Understanding of inductive biases and their importance in model architecture.
- Background in machine learning in a physics context.

Reading List

- Relational Inductive Biases, Deep Learning, and Graph Networks
- Andrej Karpathy's Neural Network Training Recipe
- PySR paper: Symbolic Regression for Scientific Discovery

Data Access

Instructions on how to access the data:

- Classical Mechanics Simulation Data

The data is available in the paper's code repository. Students are expected to re-implement the model from scratch, rather than using the existing implementations, to fully engage with the problem.

28 Radial profiles of debris discs

Proposer Name	Dr Roman Rafikov
Proposer Role and Affiliation	Professor, DAMTP
Proposer Contact Email	rrr@damtp.cam.ac.uk
Key Publication	Radial profiles of surface density in debris discs

Project Description

Many young stars harbor debris discs — discs of small particles which reflect stellar light and emit their own thermal radiation that we can observe from the ground. The particles are believed to be produced in collisions of numerous asteroid- or comet-like bodies orbiting around the stars, continuously grinding themselves down into dust on very long timescales (hundreds of millions of years). These discs are typically quite large (tens to hundreds of astronomical units) and often show a lot of interesting substructures in them. These substructures — features in the spatial distribution of particles on the sky — can ultimately provide important information about the perturbers that create them, for example massive planets that we do not observe directly. Thus, studying debris discs can eventually improve our understanding of the composition and dynamical architectures of planetary systems around other stars.

Since the debris particles are very tenuous, they rarely collide with each other and their motion can be described to first order as a Keplerian motion around the central star. It turns out that in this case one can predict analytically the radial distribution of mass in these discs if the distribution of eccentricities of the debris particles is known. This paper developed a formalism relating the two quantities and allowing one to create profiles of radial mass distribution in debris discs, something that is of great interest to observers. The gist of this formalism is in evaluating numerically a particular integral that uses eccentricity distribution of debris particles as an input. A number of illustrative calculations have also been carried out using a simple set of Python modules.

This framework is rather simple conceptually, but the robust, numerically converged evaluation of the underlying integral may be tricky in some cases, e.g. when eccentricity distribution is sharply peaked. Also, in many astrophysically-relevant problems particle eccentricity distribution features multiple radial oscillations, which ideally would require automatic determination of (multiple) integration intervals before carrying out the integration itself. Doing this robustly, efficiently and in a flexible manner is a computational mathematics challenge.

Project goals

Main purpose of this project is to develop a suite of software tools to carry out the tasks listed in the Project Description. The ideal outcome would be to develop a code reproducing the known results that could then also be released to the debris disc community for use in interpreting observations. Main goals of the project:

1. Explore the efficient integration methods suitable for the problen at hand, in light of various pathological possibilities.

2. Develop a Monte Carlo sampler of particle locations to reconstruct debris surface density distribution alternatively and to compare with the analytical framework.

3. Integrate these methods into a software suite with a user-friendly interface allowing for different user inputs, i.e. eccentricity distributions.

4. Reproduce the main results (i.e. plots and key metrics) of the key publication.

5. Release the code to the debris disc community.

This project can be naturally extended by looking at not just radial (one-dimensional), but also the full spatial distributions of particles in debris discs. Extension directions:

- Extend the numerical infrastructure to two dimensions, allowing one to create two-dimensional maps of debris discs.
- Extend the numerical infrastructure to three dimensions, allowing one to create fully three-dimensional distributions of particle density in debris discs and to study them in projection onto the sky plane in different directions.

Prerequisites

Knowledge of Python programming with applications to mathematics and physics. Familiarity with the methods of computational mathematics, in particular function integration. Prior knowledge of the relevant physical/astrophysical phenomenology is desirable.

Reading List

- Wyatt, M. C. 2008, Evolution of Debris Discs, Annual Reviews of Astronomy and Astrophysics., Vol. 46, p. 339-383; Review paper
- Press, W. H. , Teukolsky, S. A. , Vetterling, W. T. , Flannery, B. P. 1992, Numerical recipes in C. The art of scientific computing, Cambridge: University Press

Data Access

- Illustrative examples in this paper

29 Why Should I Trust You: Explainable AI in Cancer Imaging

Proposer Name	Dr Shangqi Gao
Proposer Role and Affiliation	Research Associate, Department of Oncology
Proposer Contact Email	sg2162@cam.ac.uk
Key Publication	SHAP (SHapley Additive exPlanations)

Project Description

While AI models have demonstrated significant potential in clinical diagnosis and treatment, their opaque decision-making processes pose challenges in understanding predictions, identifying biases, and justifying decisions. This is promoting the development of explainable AI in medicine. Recents works in explainable AI, e.g., SHAP, have shown remarkable outcomes in explaining the predictions of AI models. However, the effectiveness of SHPA in explaining the predictions in cancer diagnosis is under-explored.

This project is aimed to use the SHAP to explain the predictions of cutting-edge AI models in cancer diagnosis. By undertaking this project, students will not only learn skills of explainable AI and data visualization, but also know how to use them to interpret the predictions and identify the biases of AI models, especially foundation models.

Project goals

Main goals of the project:

1. Conduct a comprehensive review and understanding of the related works on explainable AI and foundation models, and establish the environment for SHAP by installing all required software libraries and dependencies.

2. Download the MINIST digit dataset, train a deep classifier on the MNIST, and visualize the SHAP values using the DeepExplainer of SHAP.

3. Collect a foundation model for computational pathology and select a classification task for a specific cancer in TCGA, then perform logistic regression on top of the pre-extracted features by the foundation model, and finally visualize the SHAP values as well as measure correlation of the features using the LinearExplainer of SHAP.

4. Select a classification task for several datasets from TCGA which consist of subjects from different race groups, then perform logistic regression on top of the pre-extracted features by the foundation model, show the biases of the classifier to race, and visualize the SHAP values using the LinearExplainer of SHAP to understand where the race-level biases come from.

Extension directions:

- Select a classification task for pan-cancer analysis and perform logistic regression on top of the pre-extracted features by the foundation model for computational pathology, test the linear classifier to show its biases to the cancer types, visualize the SHAP values as well as measure correlation of the features using the LinearExplainer of SHAP to understand where the cancer-level biases come from.

- Collect a foundation model trained using multimodal data, such as radiological images and histopathological images, and select a classification task for pan-cancer analysis, perform logistic regression on top of the pre-extracted multi-omics features by the foundation model for moltimodal data analysis, test the linear classifier to show its bias to the data modalities, visualize the SHAP values as well as measure correlation of the multi-omics features using the LinearExplainer of SHAP to understand where the modality-level biases come from.

Prerequisites

Required: Proficiency in Python programming language; Experience with libraries for data manipulation/visualization (e.g., Pandas, NumPy, Matplotlib) and machine learning (e.g., Scikit-learn, Tensorflow, PyTorch); Experience in using HPC resources for data analysis.

Desirable: Familiarity with relevant deep learning models (e.g., Multi-Layer Perceptron, Convolutional Neural Network, Transformer); Understanding of the data formats for radiological imaging data and whole slide images.

Reading List

- A Unified Approach to Interpreting Model Predictions, NeurIPS 2017.
- From local explanations to global understanding with explainable AI for trees, Nature Machine Intelligence 2020.
- Explainable machine-learning predictions for the prevention of hypoxaemia during surgery, Nature Biomedical Engineering 2018.
- A whole-slide foundation model for digital pathology from real-world data, Nature 2024.
- Towards a general-purpose foundation model for computational pathology, Nature Medicine 2024.
- Segment anything in medical images, Nature Communications 2024.

Data Access

- The MNIST database of handwritten digits
- TCGA Database (wholes slides and classification-related data)
- TCIA Database (matching subjects from TCGA)

30 Inferring the Hubble Constant using the Cepheid calibrated distance ladder

Proposer Name	Dr Suhail Dhawan
Proposer Role and Affiliation	ERC research group PI, IoA
Proposer Contact Email	sd919@ast.cam.ac.uk
Key Publication	Main Sh0es 2022 release paper

Project Description

Measuring the present-day expansion rate of the universe, i.e. the Hubble Constant, has been a key question in cosmology for nearly a century. Currently, the local distance ladder measurement is in ¿5σ tension with the global inference from the early universe. The aim of this project is to use the distance ladder data to reproduce the inferred value of the Hubble Constant. This is a diverse dataset which includes primary calibrator distances to the large magellanic cloud, nearby maser galaxy, and parallaxes to Milky Way Cepheids as well as Cepheids in Type Ia supernova host galaxies and local Hubble flow supernovae. The student will setup the linear model used by the SH0ES collaboration and setup the MCMC for obtaining the multidimensional posterior distribution.

Project goals

Main goals of the project:

1. Setup the linear model with a readable python script
2. Compare the H0 distribution to the value presented in the paper
3. Check how the H0 changes with different sample cuts

Extension directions:

- Adding variation in dust extinction as a free parameter
- Testing different selections to the permutations in the Riess et al. 2022 paper

Prerequisites

Necessary: Python 3 familiarity Desirable: Some prior data analysis with large matrices, MCMC methods Some knowledge of astrophysical distance indicators would be helpful (at the early undergraduate level) however not necessary.

Reading List

https://iopscience.iop.org/article/10.3847/1538-4357/ac756e/pdf

Data Access

https://github.com/PantheonPlusSH0ES/DataRelease

31 Simulation Based Inference for X-ray clusters

Proposer Name	Dr Sven Krippendorf
Proposer Role and Affiliation	Assistant Teaching Professor, Cavendish and DAMTP
Proposer Contact Email	slk38@cam.ac.uk
Key Publication	The eROSITA Final Equatorial-Depth Survey (eFEDS): A Machine Learning Approach to Infer Galaxy Cluster Masses from eROSITA X-ray Images

Project description

Galaxy clusters are the largest gravitationally bound objects in our Universe and provide us with vital information about how structures have formed in our Universe, i.e. the observation of this ensemble allows us to constrain cosmological parameters in the Standard Model of Cosmology. In this project you extract information about galaxy clusters from X-ray observations focusing on the eROSITA all-sky survey.

Traditionally, scaling relations (i.e. simple linear models) were used to estimate the masses of those galaxy clusters, only taking into account a subset of information. In recent years, it has been shown that taking into account all available information in neural network inference models improves performance.

You will work with simulated data to recover these literature results using computer vision architectures, similar to the ones you encounter in M2 using loss functions you also encounter in M1.

You will also look at implementing simulation based inference methods for these tasks. Simulation based inference allows for fast estimation of the posterior associated to a model task and compare them to Gaussian mixture models.

Depending on progress this project can involve actual eROSITA data.

Project goals

Detailed project goals will be communicated at the beginning of the project. Below is a rough indication of what this contains:

- Review basics of galaxy clusters and basic models of their X-ray spectrum.
- Review basics of simulation based inference and Gaussian mixture models.
- Implement exisiting computer vision pipelines from the literature on simulated galaxy cluster data.
- Implement simulation based inference for cluster mass inference.
- You will learn how to work with the standard format in astrophysics of `.fits` files and pandas dataframes.

Possible extensions:

- Upon discussion and progress extend the inference pipeline to other cluster parameters.
- You could implement different vision networks to understand the inference performance.
- Upon discussion and progress discuss the behaviour in different data limits (e.g. shorter observations).

Prerequisites

- Basic understanding of physics and astrophysics to follow the discussion about galaxy clusters.
- Neural networks as discussed in M1 and M2.
- Inference using neural network as discussed in M1.

Further literature

- Simulation based inference paper
- Software packages: SBI, Bayes-Flow

Data

The datasets will be made available at the beginning of the project in the associated gitlab repository.

32 Scaling laws of neural networks

Proposer Name	Dr Sven Krippendorf
Proposer Role and Affiliation	Assistant Teaching Professor, Cavendish and DAMTP
Proposer Contact Email	slk38@cam.ac.uk
Key Publication	Explaining Neural Scaling Laws

Project description

Large neural networks show scaling behaviour, for instance scaling their size results in better generalization behaviour across a variety of tasks. Understanding these scaling behaviours allows for the design of appropriate hyperparameter choices for large scale neural networks.

The aim of this project is to review some of the key findings and to verify this behaviour in some tasks accessible with your compute resources.

Project goals

The aim of this project is to review this behaviour in the first instance. In the second part you will demonstrate this scaling behaviour in practice using some neural networks. Depending on your interest the discussion can also focus on the scaling behaviour observed in linear networks as described in this paper or the $\mu-$transfer framework proposed here.

Prerequisites

M1 and M2 should provide you with sufficient background about neural networks. To compare the scaling behaviour in neural networks with scaling behaviour in other dynamical systems a background in undergraduate physics is helpful. You need to be familiar with resource estimates and HPC compute as covered in C1.

33 RG and sampling

Proposer Name	Dr Sven Krippendorf
Proposer Role and Affiliation	Assistant Teaching Professor, Cavendish and DAMTP
Proposer Contact Email	slk38@cam.ac.uk
Key Publication	RG-Flow: A hierarchical and explainable flow model based on renormalization group and sparse prior

Project description

Generative models have a deep connection with coarse-grained statistical physics models. One of the first such examples is the restricted Boltzmann machine which can be seen as a mean field theory.

Building in this physics knowledge into generative models proves to be a very powerful design paradigm.

In this project you will look at one of the most powerful techniques in theoretical physics with wide applications in particle physics and many-body physics, namely the renormalisation group which systematically allows to remove irrelevant information. This method is then implemented in customised flow-models which learns what the relevant information is for the respective dataset.

Implementing this physics information into our generative models appears to be necessary as *simple* realisations of state of the art generative models do not scale well to simulate large system sizes.

Project goals

A rough summary of the project goals involves:

- Familiarise yourself with the foundations of the renormalisation group.
- Familiarise yourself with flow models and their implementation.
- Re-implement the RG-normalising-flow setup as described in the key publication.
- Run some illustrative experiments. This gives you a hands-on exposure to the renormalisation group.

Potential extensions:

- Apply this method on a statistical physics dataset where the RG-behaviour is known.
- Discuss the connection to functional renormalisation group and learning following these papers.

Prerequisites

Interest in generative modelling and some theoretical physics background. It is advised that students take the ML for particle physics minor which covers some basics about renormalisation and quantum field theory.

34 Neural networks at infinite width in theory and practice

Proposer Name	Dr Sven Krippendorf
Proposer Role and Affiliation	Assistant Teaching Professor, Cavendish and DAMTP
Proposer Contact Email	slk38@cam.ac.uk
Key Publication	Wide Neural Networks of Any Depth Evolve as Linear Models Under Gradient Descent

Project description

Neural networks simplify in the limit of infinite parameters. In particular in the infinite width limit neural networks can be described by Gaussian processes. The dynamics of those neural networks at small learning rate turn out to be described by an ordinary differential equation.

In this project you review the theory behind these approximations and explore them in practice using this existing package. You compare the features learned between the actual NN training and the linear model.

Project goals

You should review the NNGP and NTK descriptions for neural networks in the first instance. You then explore these descriptions in practice using the NTK package. Finally you compare the representations learned by the approximations and the actual NN training.

Depending on progress, a potential line of extension can be to capture the time evolution of the empircial neural tangent kernel.

Prerequisites

The basics of neural networks as covered in the major module. The NTK package is written in JAX and you are expected to work with this package and JAX in this project. This project can involve the training of many neural networks and so you should be prepared to think about how to use your available HPC resources wisely. This project is more on the theoretical side of machine learning so you should be prepared to engage with the relevant mathematics involved.

35 Numerical Calabi-Yau metrics and symbolic regression

Proposer Name	Dr Sven Krippendorf
Proposer Role and Affiliation	Assistant Teaching Professor, Cavendish and DAMTP
Proposer Contact Email	slk38@cam.ac.uk
Key Publication	CYJAX: A package for Calabi-Yau metrics with JAX

Project description:

Solving Einsteins equations can be a non-trivial task. One such examples is in the context of compact Calabi-Yau manifolds. In Yau's field theory work it has been proven that such metrics exist but their explicit construction is outstanding. In recent years, we have made progress in constructing these metrics numerically using neural networks as efficient function approximators (see here). In this project you will dive into this optimisation problem of identifying the metric.

The applications of these metrics are in mathematics and physics. Can on find appropriate analytical representations of these metrics, probably requiring to define novel special functions. In physics, these spaces do appear in the context of string theory where they describe additional spacetime dimensions. The metric ultimately determines properties of the low-energy physics such as the Yukawa couplings, i.e. parameters which can be confronted with data from experimental searches.

Project goals:

- Familiarise yourself with the optimisation problem for Calabi-Yau manifolds.
- Obtain metrics for the quintic hypersurface, the simplest example in three dimensions, using our package cyjax.
- Optimise the performance of these metrics with respect to their deviation from Ricci flatness ($R_{i\bar{j}} = 0$).
- Using these metrics, you can then move on to searching for analytical expressions using symbolic regression (in particular PYSR). To overcome the complexity you need to incorporate appropriate functional biases into your search for analytic expressions. In particular it promises to be fruitful to explore the symmetries of the problem.

Depending on progress and interest we could extend in the following directions:

- We can increase the complexity of the manifold where we want to learn the metric.
- You can review phenomenological applications of Calabi-Yau metrics.
- We can explore non-compact Calabi-Yau metrics where analytic expressions are known.

Project prerequisites

You should be familiar with general relativity in the sense that you have seen Einstein's equations. The machine learning aspects will be covered in M1 and M2. The package is written in JAX, so you should be happy to work with JAX throughout this project. To modify the basis functions in PYSR you most likely use Julia.

36 Clustering in string theory datasets

Proposer Name	Dr Sven Krippendorf
Proposer Role and Affiliation	Assistant Teaching Professor, Cavendish and DAMTP
Proposer Contact Email	slk38@cam.ac.uk
Key Publications	Probing the Structure of String Theory Vacua with Genetic Algorithms and Reinforcement Learning, JAXVacua – A Framework for Sampling String Vacua

Project description

At the fundamental level the Standard Model of Particle Physics falls short to describe all known interactions as a consistent quantum theory, in particular to consistently quantise gravity. String theory is one of the leading candidates to answer this question. In our current understanding string theory provides a plethora of vacuum solutions, featuring an enormous dataset which is waiting for investigation using state of the art numerical methods. In its simplest form this dataset provides a map from high-energy parameters to low-energy observables such as the scales in the string theory realisation of standard model physics. We want to understand the shared properties of the high-energy parameters that lead to interesting low-energy physics. To tackle the high-dimensionality in the high-energy variables this project seeks to explore dimensional reduction techniques which allow the visualisation of such datasets. This will include standard clustering techniques and other methods which we cover in M1. On a wider perspective, this type of analysis allows for the discovery of new mathematical structures which guarantee the relevant phenomenological properties. Although studied here in the context of string theory, the scope is broader as the application to different datasets, i.e. other beyond the Standard Model proposals, is not obstructed.

Project goals

- You should familiarise yourself with this inverse problem as described for instance here.
- You should describe the dataset and summarise how it was generated at a rough level.
- On the provided datasets you should apply a variety of dimensional reduction methods (e.g. clustering). In our meeting, we should discuss the selection of these methods. You should report on your findings.

Possible extensions depending upon progress:

- We can easily expand upon the datasets being used and analysed, i.e. we include further geometries in our analysis.
- We expand upon the probabilistic models studied to describe these datasets.

Project prerequisites

This project requires some familiarity with particle physics and an interest in theories for beyond the Standard Model physics, i.e. it provides an opportunity to familarise yourself on how to connect string theory with observable physics.

37 Multimodal Prototyping for cancer survival prediction

Proposer Name	Dr Zeyu Gao & Dr Ines Machado
Proposer Role and Affiliation	Research Associate, Department of Oncology
Proposer Contact Email	zg323@cam.ac.uk & im549@cam.ac.uk
Key Publication	Multimodal Prototyping for cancer survival prediction

Project Description

Cancer prognostication utilizing multimodal data, such as clinical records, histology images and transcriptomic profiles, is gaining tremendous attention in the current research landscape. The ability to integrate diverse data types to predict patient outcomes more accurately can significantly impact personalized medicine, enabling tailored treatment plans and improving survival rates. The proposed project aims to replicate the findings of a novel approach in the paper "Multimodal Prototyping for cancer survival prediction (Andrew, ICML 2024)", which integrates gigapixel histology whole-slide images (WSIs) and transcriptomic profiles for patient prognostication and stratification. Current methodologies in this domain involve tokenizing WSIs into a vast number of smaller patches (over 10,000 patches) and dividing transcriptomics into gene groups. These tokens are then integrated using a Transformer model to predict patient outcomes. However, such an approach generates a large number of tokens, resulting in high memory requirements for computing attention and complicating post-hoc interpretability analyses, thus hindering their practical application in clinical settings. To address these challenges, this project will explore a more efficient and interpretable deep learning methodology. The central hypothesizes are: first, that the morphological content of a WSI can be effectively summarized by condensing its constituting tokens using morphological prototypes, achieving significant compression; second, that cellular functions can be accurately characterized by encoding the transcriptomic profile with biological pathway prototypes, all achieved in an unsupervised manner. The resulting multimodal tokens will then be processed by a fusion network, either with a Transformer or through an optimal transport cross-alignment, operating with a small and fixed number of tokens without approximations. At the end, this project will evaluate the proposed framework on several different cancer types to demonstrate its superiority over state-of-the-art methods in terms of computational efficiency and interpretability.

By undertaking this project, students will gain a comprehensive set of skills at the intersection of machine learning, bioinformatics, and medical image analysis. (1) Data Preprocessing and Tokenization: Techniques for handling and preprocessing gigapixel histology images and high-dimensional transcriptomic data. (2) Unsupervised Learning: Methods for unsupervised learning, including clustering and prototype extraction for data compression and summarization. (3) Multimodal Data Integration: Approaches for integrating heterogeneous data sources using advanced machine learning models such as Transformers and optimal transport algorithms. (4) Interpretability and Evaluation: Strategies for conducting post-hoc interpretability analyses and evaluating model performance across multiple cancer types.

Project goals

Main goals of the project:

1. Conduct a comprehensive review and understanding of the original paper and related works, with a focus on the proposed methods, data representation, and fusion mechanisms.

2. Establish the computational environment for training the models by installing all required software libraries and dependencies. Select any 3 of the 6 datasets used in the original study and download them from the Cancer Genome Atlas (TCGA) database.

3. Reproduce the cancer survival prediction and risk stratification results by following the same settings as the original paper.

4. Implement the interpretability framework and demonstrate the cross-modal interaction visualization.

Extension directions:

- Integrate DNA mutation information, copy number variation, or DNA methylation data (all available in the TCGA database) into the multimodal framework.
- Adapt this framework for a more clinically relevant task, specifically Homologous Recombination Deficiency (HRD), which is crucial for predicting the drug response (PARP inhibitors) in cancer patients.

Prerequisites

Required: Proficiency in Python programming language; Experience with libraries for data manipulation (e.g., Pandas, NumPy) and machine learning (e.g., Scikit-learn, PyTorch); Experience in using HPC resources for large-scale data analysis.

Desirable: Familiarity with relevant deep learning models (e.g., Multi-Layer Perceptron, Convolutional Neural Network, Transformer); Knowledge of related learning paradigms (e.g., self-supervised learning, prototype learning); Understanding of the data formats for whole slide images and gene expression.

Reading List

- Multimodal Prototyping for cancer survival prediction
- Modeling Dense Multimodal Interactions Between Biological Pathways and Histology for Survival Prediction
- Prototypical Information Bottlenecking and Disentangling for Multimodal Cancer Survival Prediction
- Towards a general-purpose foundation model for computational pathology

Data Access

Instructions on how to access the data with necessary links, if any.

- TCGA Database (Download original WSIs)
- Transcriptomics Data
- Pathway Compositions
- Patient Labels

Supervising Arrangement

The first presentation will involve both supervisors together. Then, for each term (Lent and Easter), there will be 2-3 one-on-one supervisions: 1 or 2 with Dr. Zeyu Gao and 1 with Dr. Ines Machado.