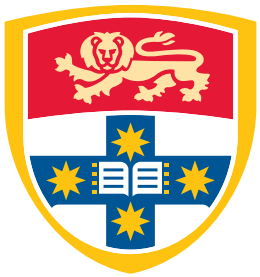
Honours in Data Science

Detailed Guide for the 2023 academic year



THE UNIVERSITY OF SYDNEY

School of Mathematics and Statistics

Contents

1 Entry requirements

Preliminary entrance into the honours program is through the Faculty of Science application portal. The Faculty requirements which must be met include:

- qualifying for a degree in a major which is cognate to the proposed honours stream, that is, a major which provides a suitable background for the honours stream. Note that a major in Statistics, Data Science or Financial Mathematics & Statistics are inherently cognate to our honours program while in borderline cases the decision of whether a major is cognate is in the hands of the Honours coordinator and the faculty);
- having a WAM of at least 65;
- *securing the agreement of a supervisor.*

In addition, the School of Mathematics and Statistics requires that the student has a total of at least 18CP of relevant 3XXX or 4XXX courses in which

- the average mark of Advanced level courses is at least 65;
- the average mark of Mainstream level courses is at least 75

If you have a mix of advanced and mainstream courses, where some are above and some below the thresholds, if you are not sure which of your courses are relevant, or if your average is just on the wrong side of the threshold you can seek further advice from the relevant program's honours coordinator.

1.1 Formally...

The faculty offers three main Honours pathways and it can be confusing:

- Combined Bachelor of Science/Bachelor of Advanced Studies is an option if you commenced your studies after 2018 and it allows completing Honours as an embedded pathway in the final year of the program. Requires two majors.
- Standalone Bachelor of Advanced Studies (Honours) is the same as above, including the two-majors requirement, except technically this is an appended, standalone Honours year.
- The Bachelor of Science (Honours) is a standalone (appended) Honours requiring an additional year of study. It is for students who
  - are not on track to complete two majors in the Bachelor of Science, or
  - are external students with only one major, or
  - commenced before 2018 and did not choose to transfer to the new curriculum version of their degree.

1.2 It's important to note that:

- All acceptances into Honours (including in cases where the School's requirements are not met) are ultimately at the discretion of the School. However, a student meeting all of the above criteria (or the equivalent from another institution) should be confident of acceptance.
- The Faculty of Science Honours **application deadline** (for Honours commencement in Semester 1, 2023) is 15 January 2023 and for Semester 2, 2023 it is 25 June 2023.

2 Structure of Honours

An Honours year in Data Science involves

- four 6CP courses (worth 50% of the final mark) and
- and a project (worth 50%).

2.1 The Honours project (or thesis) (50%)

The Honours project centres around an essay/thesis consisting of roughly 60 pages[1] written on a particular topic from your chosen area. It does not need to contain original research (although it might) but it should clearly demonstrate that you have understood and mastered the material. The assessment of the Honours thesis is based on the scientific and quantitative content and its exposition, including the written english. The thesis is due at the end of your second semester, specifically at 5pm on Monday of Week 13.

Toward the end of the second semester (Friday weeks 9-10), each student gives a 25 minutes talk on their thesis project. The aim of the talk is to explain to a broader audience the purpose and nature of the project. The talk is followed by 5 minutes dedicated to questions from the audience which includes staff members and fellow students.

2.1.1 Writing proficiency

As mentioned above your essay is also assessed based on the quality of the writing. This does not mean we look for the next Shakespeare however you should make sure you express your ideas in an organized manner using a clear and grammatically correct English. The University of Sydney offers several resources that can help you achieve this goal. The Learning Centre offers workshops for students that need help with extended written work, and a trove of online resources for improving your writing skills is also available. Make sure you make use of these resources as early as possible as writing skills develop slowly over time and with much practice.

2.2 Course work (50%)

The Honours program in *Data Science* specifies different combinations of courses that can be taken including courses offered by the School of Mathematics and Statistics, the School of Computer Science and the School of Physics.

A list of courses that will be offered in 2023 is available online. However students should carefully consult the *Data Science* degree structure document which outlines the combinations of courses that can be taken for credit.

[1] This page number is a very rough guideline and should not be taken as binding.

3 Important course work information for all students

3.1 Selecting your courses

Please make sure you **select your courses after consulting the Honours supervisor and the Honours coordinator!**

3.2 AMSI courses

Students are welcomed to check the courses offered in January at the AMSI Summer School and also courses available via the Advanced Collaborative Environment (ACE). These courses can possibly be taken for credit (by enrolling in the unit AMSI4001), but this can only be done in consultation with the student's supervisor and with the approvals of the specific honours coordinator as well as the School's Honours coordinator, Prof. Laurentiu Paunescu.

4 Program Administration

The Data Science Honours coordinator is

Dr. Clara Grazian,
Carslaw Building, Room 822
Email: clara.grazian@sydney.edu.au

The Co-director of Teaching (Statistics & Data Science) is

A/Prof. Jennifer Chan,
Carslaw Building, Room 817, Phone 9351 4873,
Email: jennifer.chan@sydney.edu.au

The Program Coordinator is the person that students should consult on all matters regarding the Honours program. In particular, students wishing to substitute a course from another Department, School or University must get prior written approval from the Program Coordinator. Matters of ill-health or misadventure should also be referred to the Program Coordinator.

5 Potential Supervisors and their Research Interests

See the individual staff member webpages for more detail about their research and their contact information. Supervisors are listed in alphabetical order.

5.1 The Statistics Group

- A/Prof Uri Keich

  false discoveries in multiple hypotheses testing; statistical analysis of proteomics data; computational statistics; statistical methods for bioinformatics

- Dr. Shila Ghazanfar

  data science; R Shiny; interactive data visualisation; single-cell RNA-sequencing; spatially resolved genomics data

- Dr Clara Grazian

  Bayesian statistics; mixture models; copula models; spatio-temporal data; genomic data; approximate Bayesian algorithms

- Dr Linh Nghiem

  measurement error modelling; dimension reduction; graphical models; human perception of music

- Dr Ellis Patrick

  applied statistics; bioinformatics; machine learning; image analysis; focus on method development for high-dimensional biomedical assays including; high-parameter imaging cytometry data

- Dr Michael Stewart

  mixture model; selection; extremes of stochastic processes; empirical process approximations; semiparametric theory and applications

- Dr Garth Tarr

  applied statistics; robust methods; model selection; data visualisation; biometrics

- Prof Jean Yang

  statistical bioinformatics; applied statistics; analysis of multi-omics data; biomedical data science; single-cell data analytics; statistical learning in precision medicine

- A/Prof Pengyi Yang

  machine learning; deep learning; statistical modelling; single-cell omics; multi-omics; systems stem cell biology

5.2 The Applied Mathematics Group

- Dr Lindon Roberts

  numerical analysis; data science; nonconvex and derivative-free optimization

- Prof Dingxuan Zhou

  theory of deep learning; statistical learning; approximation theory; wavelet analysis

5.3 The Computer Science Group

- Dr Clément Cannone (Computer Science)

  algorithms; high dimensional data; information constraints

- Dr Michael Harre

  minimisation; commerce; decision making; evolutionary computation; game theory; graph theory; modelling; network theory (graphs); phase transformations; social networking (online); social sciences; multi-agent systems; data mining; data visualisation; knowledge based systems; learning (artificial intelligence); stochastic games; very large databases

- Prof Judy Kay (Computer Science)

  smart technology; metacognitive interface; open learner models

- Dr Nguyen Tran (Computer Science)

  data privacy; edge network performance; stochastic modelling; game theory

- A/Prof Zhiyong Wang (Computer Science)

  multimedia data; multimedia information retrieval; computer graphics and animation; surveillance; medical diagnostics; health care; environmental monitoring; astronomy

6 Honours courses in Data Science

The Bachelor of Advanced Studies (Honours) (Data Scence) requires 48 credit points from this table including:

- 12 credit points of 4000-level and above Honours coursework selective units from List 1, and
- 12 credit points of 4000-level and above Honours coursework selective units from List 1, List 2, List 4 or List 5 with a maximum of 6 credit points of units from List 5, and
- 24 credit points of 4000-level Honours research project units

Note: not all courses are offered every year. Moreover, some courses may have pre-requisites and exclusions. Please, check the Honours page for udpate information about offerings, pre-requisites, and exclusions.

6.1 List 1

- **COMP5046 Natural Language Processing (6 credits, Semester 1)**
  This unit introduces computational linguistics and the statistical techniques and algorithms used to automatically process natural languages (such as English or Chinese). It will review the core statistics and information theory, and the basic linguistics, required to understand statistical natural language processing (NLP). Statistical NLP is used in a wide range of applications, including information retrieval and extraction; question answering; machine translation; and classifying and clustering of documents. This unit will explore the key challenges of natural language to computational modelling, and the state of the art approaches to the key NLP sub-tasks, including tokenisation, morphological analysis, word sense representation, part-of-speech tagging, named entity recognition and other information extraction, text categorisation, phrase structure parsing and dependency parsing. You will implement many of these sub-tasks in labs and assignments. The unit will also investigate the annotation process that is central to creating training data for statistical NLP systems. You will annotate data as part of completing a real-world NLP task.

- **COMP5048 Visual Analytics (6 credits, Semester 1 and 2)**
  Visual Analytics aims to facilitate the data analytics process through Information Visualisation. Information Visualisation aims to make good pictures of abstract information, such as stock prices, family trees, and software design diagrams. Well designed pictures can convey this information rapidly and effectively. The challenge for Visual Analytics is to design and implement effective Visualisation methods that produce pictorial representation of complex data so that data analysts from various fields (bioinformatics, social network, software visualisation and network) can visually inspect complex data and carry out critical decision making. This unit will provide basic HCI concepts, visualisation techniques and fundamental algorithms to achieve good visualisation of abstract information. Further, it will also provide opportunities for academic research and developing new methods for Visual Analytic methods.

- **COMP5328 Advanced Machine Learning (6 credits, Semester 2)**
  Machine learning models explain and generalise data. This course introduces some fundamental machine learning concepts, learning problems and algorithms to provide understanding

and simple answers to many questions arising from data explanation and generalisation. For example, why do different machine learning models work? How to further improve them? How to adapt them to different purposes?

- **COMP5329 Deep Learning (6 credits, Semester 1)**
  This course provides an introduction to deep machine learning, which is rapidly emerging as one of the most successful and widely applicable set of techniques across a range of applications. Students taking this course will be exposed to cutting-edge research in machine learning, starting from theories, models, and algorithms, to implementation and recent progress of deep learning. Specific topics include: classical architectures of deep neural network, optimization techniques for training deep neural networks, theoretical understanding of deep learning, and diverse applications of deep learning in computer vision.

- **COMP5338 Advanced Data Models (6 credits, Semester 2)**
  This unit of study gives a comprehensive overview of post-relational data models and of latest developments in data storage technology. Particular emphasis is put on spatial, temporal, and NoSQL data storage. This unit extensively covers the advanced features of SQL:2003, as well as a few dominant NoSQL storage technologies. Besides in lectures, the advanced topics will be also studied with prescribed readings of database research publications.

- **COMP5349 Cloud Computing (6 credits, Semester 1)**
  This unit covers topics of active and cutting-edge research within IT in the area of 'Cloud Computing'. Cloud Computing is an emerging paradigm of utilising large-scale computing services over the Internet that will affect individual and organization's computing needs from small to large. Over the last decade, many cloud computing platforms have been set up by companies like Google, Yahoo!, Amazon, Microsoft, Salesforce, Ebay and Facebook. Some of the platforms are open to public via various pricing models. They operate at different levels and enable business to harness different computing power from the cloud. In this course, we will describe the important enabling technologies of cloud computing, explore the state-of-the art platforms and the existing services, and examine the challenges and opportunities of adopting cloud computing. The unit will be organized as a series of presentations and discussions of seminal and timely research papers and articles. Students are expected to read all papers, to lead discussions on some of the papers and to complete a hands-on cloud-programming project.

- **STAT4025 Time Series (6 credits, Semester 1)**
  This unit will study basic concepts and methods of time series analysis applicable in many real world problems in numerous fields, including economics, finance, insurance, physics, ecology, chemistry, computer science and engineering. This unit will investigate the basic methods of modelling and analyzing of time series data (i.e. data containing serially dependence structure). This can be achieved through learning standard time series procedures on identification of components, autocorrelations, partial autocorrelations and their sampling properties. After setting up these basics, students will learn the theory of stationary univariate time series models including ARMA, ARIMA and SARIMA and their properties. Then the identification, estimation, diagnostic model checking, decision making and forecasting methods based on these models will be developed with applications. The spectral theory of time series, estimation of spectra using periodogram and consistent estimation of spectra using lag-windows will be studied in detail. Further, the methods of analyzing long memory

and time series and heteroscedastic time series models including ARCH, GARCH, ACD, SCD and SV models from financial econometrics and the analysis of vector ARIMA models will be developed with applications. By completing this unit, students will develop the essential basis for further studies, such as financial econometrics and financial time series. The skills gained through this unit of study will form a strong foundation to work in a financial industry or in a related research organization.

- **STAT4026 Statistical Consulting (6 credits, Semester 1)**
  In our ever-changing world, we are facing a new data-driven era where the capability to efficiently combine and analyse large data collections is essential for informed decision making in business and government, and for scientific research. Statistics and data analytics consulting provide an important framework for many individuals to seek assistance with statistics and data-driven problems. This unit of study will provide students with an opportunity to gain real-life experience in statistical consulting or work with collaborative (interdisciplinary) research. In this unit, you will have an opportunity to have practical experience in a consultation setting with real clients. You will also apply your statistical knowledge in a diverse collection of consulting projects while learning project and time management skills. In this unit you will need to identify and place the client's problem into an analytical framework, provide a solution within a given time frame and communicate your findings back to the client. All such skills are highly valued by employers. This unit will foster the expertise needed to work in a statistical consulting firm or data analytical team which will be essential for data-driven professional and research pathways in the future.

- **STAT4027 Advanced Statistical Modelling (6 credits, Semester 2)**
  Applied Statistics fundamentally brings statistical learning to the wider world. Some data sets are complex due to the nature of their responses or predictors or have high dimensionality. These types of data pose theoretical, methodological and computational challenges that require knowledge of advanced modelling techniques, estimation methodologies and model selection skills. In this unit you will investigate contemporary model building, estimation and selection approaches for linear and generalised linear regression models. You will learn about two scenarios in model building: when an extensive search of the model space is possible; and when the dimension is large and either stepwise algorithms or regularisation techniques have to be employed to identify good models. These particular data analysis skills have been foundational in developing modern ideas about science, medicine, economics and society and in the development of new technology and should be in the toolkit of all applied statisticians. This unit will provide you with a strong foundation of critical thinking about statistical modelling and technology and give you the opportunity to engage with applications of these methods across a wide scope of applications and for research or further study.

6.2 List 2

- **AMSI4001 AMSI Summer School (6 credits, Intensive February)**
  A Completed a first degree with a major in Mathematics, Statistics, Financial Mathematics and Statistics, Data Science or equivalent Note: Department permission required for enrolment. This unit has been designed to enable University of Sydney students to continue to take advantage of the premier Mathematics and Statistics summer school held in Australia. Intensive February

- **MATH4061 Metric Spaces (6 credits, Semester 1)**
  Topology, developed at the end of the 19th Century to investigate the subtle interaction of analysis and geometry, is now one of the basic disciplines of mathematics. A working knowledge of the language and concepts of topology is essential in fields as diverse as algebraic number theory and non-linear analysis. This unit develops the basic ideas of topology using the example of metric spaces to illustrate and motivate the general theory. Topics covered include: Metric spaces, convergence, completeness and the Contraction Mapping Theorem; Metric topology, open and closed subsets; Topological spaces, subspaces, product spaces; Continuous mappings and homeomorphisms; Compactness Connectedness Hausdorff spaces and normal spaces. You will learn methods and techniques of proving basic theorems in point-set topology and apply them to other areas of mathematics including basic Hilbert space theory and abstract Fourier series. By doing this unit you will develop solid foundations in the more formal aspects of topology, including knowledge of abstract concepts and how to apply them. Applications include the use of the Contraction Mapping Theorem to solve integral and differential equations.

- **MATH4062 Rings, Fields and Galois Theory (6 credits, Semester 1)**
  This unit of study lies at the heart of modern algebra. In the unit we investigate the mathematical theory that was originally developed for the purpose of studying polynomial equations. In a nutshell, the philosophy is that it should be possible to completely factorise any polynomial into a product of linear factors by working over a large enough field (such as the field of all complex numbers). Viewed like this, the problem of solving polynomial equations leads naturally to the problem of understanding extensions of fields. This in turn leads into the area of mathematics known as Galois theory. The basic theoretical tool needed for this program is the concept of a ring, which generalises the concept of a field. The course begins with examples of rings, and associated concepts such as subrings, ring homomorphisms, ideals and quotient rings. These tools are then applied to study quotient rings of polynomial rings. The final part of the course deals with the basics of Galois theory, which gives a way of understanding field extensions. Along the way you will see some beautiful gems of mathematics, including Fermat's Theorem on primes expressible as a sum of two squares, solutions to the ancient Greek problems of trisecting the angle, squaring the circle, and doubling the cube, and the crown of the course: Galois' proof that there is no analogue of the quadratic formula for the general quintic equation. On completing this unit of study you will have obtained a deep understanding of modern abstract algebra.

- **MATH4063 Dynamical Systems and Applications (6 credits, Semester 1)**
  The theory of ordinary differential equations is a classical topic going back to Newton and Leibniz. It comprises a vast number of ideas and methods. The theory has many applications and stimulates new developments in almost all areas of mathematics. The emphasis is on qualitative analysis including phase-plane methods, bifurcation theory and the study of limit cycles. The more theoretical part includes existence and uniqueness theorems, linearisation, and analysis of asymptotic behaviour. The applications in this unit will be drawn from predator-prey systems, population models, chemical reactions, and other equations and systems from mathematical biology. You will learn how to use ordinary differential equations to model biological, chemical, physical and/or economic systems and how to use different methods from dynamical systems theory and the theory of nonlinear ordinary differential equations to find the qualitative outcome of the models. By doing this unit you will develop skills in using and analysing nonlinear differential equations which will prepare you for

further studies in mathematics, systems biology or physics or for careers in mathematical modelling.

- **MATH4068 Differential Geometry (6 credits, Semester 2)**
  This unit is an introduction to Differential Geometry, one of the core pillars of modern mathematics. Using ideas from calculus of several variables, we develop the mathematical theory of geometrical objects such as curves, surfaces and their higher-dimensional analogues. For students, this provides the first taste of the investigation on the deep relation between geometry and topology of mathematical objects, highlighted in the classic Gauss-Bonnet Theorem. Differential geometry also plays an important part in both classical and modern theoretical physics. The unit aims to develop geometrical ideas such as curvature in the context of curves and surfaces in space, leading to the famous Gauss-Bonnet formula relating the curvature and topology of a surface. A second aim is to remind the students about all the content covered in the mathematical units for previous years, most importantly the key ideas in vector calculus, along with some applications. It also helps to prepare the students for honours courses like Riemannian Geometry. By doing this unit you will further appreciate the beauty of mathematics which originated from the need to solve practical problems, develop skills in understanding the geometry of the surrounding environment, prepare yourself for future study or the workplace by developing advanced critical thinking skills and gain a deep understanding of the underlying rules of the Universe.

- **MATH4069 Measure Theory and Fourier Analysis (6 credits, Semester 2)**
  Measure theory is the study of fundamental ideas as length, area, volume, arc length and surface area. It is the basis for Lebesgue integration theory used in advanced mathematics ever since its development in about 1900. Measure theory is also a key foundation for modern probability theory. The course starts by establishing the basics of measure theory and the theory of Lebesgue integration, including important results such as Fubini's Theorem and the Dominated Convergence Theorem which allow us to manipulate integrals. These ideas are applied to Fourier Analysis which leads to results such as the Inversion Formula and Plancherel's Theorem. The Radon-Nikodyn Theorem provides a representation of measures in terms of a density. Key ideas of this theory are applied in detail to probability theory to provide a rigorous framework for probability which takes in and generalizes familiar ideas such as distributions and conditional expectation. When you complete this unit you will have acquired a new generalized way of thinking about key mathematical concepts such as length, area, integration and probability. This will give you a powerful set of intellectual tools and equip you for further study in mathematics and probability.

- **MATH4071 Convex Analysis and Optimal Control (6 credits, Semester 1)**
  The questions how to maximise your gain (or to minimise the cost) and how to determine the optimal strategy/policy are fundamental for an engineer, an economist, a doctor designing a cancer therapy, or a government planning some social policies. Many problems in mechanics, physics, neuroscience and biology can be formulated as optimisation problems. Therefore, optimisation theory is an indispensable tool for an applied mathematician. Optimisation theory has many diverse applications and requires a wide range of tools but there are only a few ideas underpinning all this diversity of methods and applications. This course will focus on two of them. We will learn how the concept of convexity and the concept of dynamic programming provide a unified approach to a large number of seemingly unrelated problems. By completing this unit you will learn how to formulate optimisation problems

that arise in science, economics and engineering and to use the concepts of convexity and the dynamic programming principle to solve straightforward examples of such problems. You will also learn about important classes of optimisation problems arising in finance, economics, engineering and insurance.

- **MATH4074 Fluid Dynamics (6 credits, Semester 1)**
  Fluid Dynamics is the study of systems which allow for a macroscopic description in some continuum limit. It is not limited to the study of liquids such as water but includes our atmosphere and even car traffic. Whether a system can be treated as a fluid, depends on the spatial scales involved. Fluid dynamics presents a cornerstone of applied mathematics and comprises a whole gamut of different mathematical techniques, depending on the question we ask of the system under consideration. The course will discuss applications from engineering, physics and mathematics: How and in what situations a system which is not necessarily liquid can be described as a fluid? The link between an Eulerian description of a fluid and a Lagrangian description of a fluid, the basic variables used to describe flows, the need for continuity, momentum and energy equations, simple forms of these equations, geometric and physical simplifying assumptions, streamlines and stream functions, incompressibility and irrotationality and simple examples of irrotational flows. By the end of this unit, students will have received a basic understanding into fluid mechanics and have acquired general methodology which they can apply in their further studies in mathematics and/or in their chosen discipline.

- **MATH4076 Computational Mathematics (6 credits, Semester 1)**
  Sophisticated mathematics and numerical programming underlie many computer applications, including weather forecasting, computer security, video games, and computer aided design. This unit of study provides a strong foundational introduction to modern interactive programming, computational algorithms, and numerical analysis. Topics covered include: (I) basics ingredients of programming languages such as syntax, data structures, control structures, memory management and visualisation; (II) basic algorithmic concepts including binary and decimal representations, iteration, linear operations, sources of error, divide-and-concur, algorithmic complexity; and (III) basic numerical schemes for rootfinding, integration/differentiation, differential equations, fast Fourier transforms, Monte Carlo methods, data fitting, discrete and continuous optimisation. You will also learn about the philosophical underpinning of computational mathematics including the emergence of complex behaviour from simple rules, undecidability, modelling the physical world, and the joys of experimental mathematics. When you complete this unit you will have a clear and comprehensive understanding of the building blocks of modern computational methods and the ability to start combining them together in different ways. Mathematics and computing are like cooking. Fundamentally, all you have is sugar, fat, salt, heat, stirring, chopping. But becoming a good chef requires knowing just how to put things together in creative ways that work. In a previous study, you should have learned to cook. Now you're going to learn how to make something someone else might want to pay for more than one time.

- **MATH4077 Lagrangian and Hamiltonian Dynamics (6 credits, Semester 2)**
  Lagrangian and Hamiltonian dynamics are reformulations of classical Newtonian mechanics into a mathematically sophisticated framework using arbitrary coordinate systems. This formulation of classical mechanics generalises elegantly to modern theories of relativity and quantum mechanics. The unit develops dynamics from the Principle of Least Action us-

ing the calculus of variations. Emphasis is placed on the relation between the symmetry and invariance properties of the Lagrangian and Hamiltonian functions and conservation laws. Coordinate and canonical transformations are introduced to simplify apparently complicated dynamical problems. Connections between geometry and different physical theories beyond classical mechanics are explored. Students will be expected to describe and solve mechanical systems of some complexity including planetary motion and to investigate stability. Hamilton-Jacobi theory will be used to solve problems ranging from geodesic motion (shortest path between two points) on curved surfaces to relativistic motion in the vicinity of black holes. Students will study an application of Lagrangian and Hamiltonian dynamics described in a modern research paper.

- **MATH4078 PDEs and Applications (6 credits, Semester 2)**
  The aim of this unit is to introduce some fundamental concepts of the theory of partial differential equations (PDEs) arising in Physics, Chemistry, Biology and Mathematical Finance. The focus is mainly on linear equations but some important examples of nonlinear equations and related phenomena re introduced as well. After an introductory lecture, we proceed with first-order PDEs and the method of characteristics. Here, we also nonlinear transport equations and shock waves are discussed. Then the theory of the elliptic equations is presented with an emphasis on eigenvalue problems and their application to solve parabolic and hyperbolic initial boundary-value problems. The Maximum principle and Harnack's inequality will be discussed and the theory of Green's functions.

- **MATH4079 Complex Analysis (6 credits, Semester 1)**
  The unit will begin with a revision of properties of complex numbers and complex functions. This will be followed by material on conformal mappings, Riemann surfaces, complex integration, entire and analytic functions, the Riemann mapping theorem, analytic continuation, and Gamma and Zeta functions. Finally, special topics chosen by the lecturer will be presented, which may include elliptic functions, normal families, Julia sets, functions of several complex variables, or complex manifolds.

- **MATH4311 Algebraic Topology (6 credits, Semester 2)**
  One of the most important aims of algebraic topology is to distinguish or classify topological spaces and maps between them up to homeomorphism. Invariants and obstructions are key to achieve this aim. A familiar invariant is the Euler characteristic of a topological space, which was initially discovered via combinatorial methods and has been rediscovered in many different guises. Modern algebraic topology allows the solution of complicated geometric problems with algebraic methods. Imagine a closed loop of string that looks knotted in space. How would you tell if you can wiggle it about to form an unknotted loop without cutting the string? The space of all deformations of the loop is an intractable set. The key idea is to associate algebraic structures, such as groups or vector spaces, with topological objects such as knots, in such a way that complicated topological questions can be phrased as simpler questions about the algebraic structures. In particular, this turns questions about an intractable set into a conceptual or finite, computational framework that allows us to answer these questions with certainty. In this unit you will learn about fundamental group and covering spaces, homology and cohomology theory. These form the basis for applications in other domains within mathematics and other disciplines, such as physics or biology. At the end of this unit you will have a broad and coherent knowledge of Algebraic Topology,

and you will have developed the skills to determine whether seemingly intractable problems can be solved with topological methods.

- **MATH4312 Commutative Algebra (6 credits, Semester 2)**
  Commutative Algebra provides the foundation to study modern uses of Algebra in a wide array of settings, from within Mathematics and beyond. The techniques of Commutative Algebra underpin some of the most important advances of mathematics in the last century, most notably in Algebraic Geometry and Algebraic Topology. This unit will teach students the core ideas, theorems, and techniques from Commutative Algebra, and provide examples of their basic applications. Topics covered include affine varieties, Noetherian rings, Hilbert basis theorem, localisation, the Nullstellansatz, ring specta, homological algebra, and dimension theory. Applications may include topics in scheme theory, intersection theory, and algebraic number theory. On completion of this unit students will be thoroughly prepared to undertake further study in algebraic geometry, algebraic number theory, and other areas of mathematics. Students will also gain facility with important examples of abstract ideas with far-reaching consequences.

- **MATH4313 Functional Analysis (6 credits, Semester 1)**
  Functional analysis is one of the major areas of modern mathematics. It can be thought of as an infinite-dimensional generalisation of linear algebra and involves the study of various properties of linear continuous transformations on normed infinite-dimensional spaces. Functional analysis plays a fundamental role in the theory of differential equations, particularly partial differential equations, representation theory, and probability. In this unit you will cover topics that include normed vector spaces, completions and Banach spaces; linear operators and operator norms; Hilbert spaces and the Stone-Weierstrass theorem; uniform boundedness and the open mapping theorem; dual spaces and the Hahn-Banach theorem; and spectral theory of compact self-adjoint operators. A thorough mechanistic grounding in these topics will lead to the development of your compositional skills in the formulation of solutions to multifaceted problems. By completing this unit you will become proficient in using a set of standard tools that are foundational in modern mathematics and will be equipped to proceed to research projects in PDEs, applied dynamics, representation theory, probability, and ergodic theory.

- **MATH4314 Representation Theory (6 credits, Semester 1)**
  Representation theory is the abstract study of the possible types of symmetry in all dimensions. It is a fundamental area of algebra with applications throughout mathematics and physics: the methods of representation theory lead to conceptual and practical simplification of any problem in linear algebra where symmetry is present. This unit will introduce you to the basic notions of modules over associative algebras and representations of groups, and the ways in which these objects can be classified. You will learn the special properties that distinguish the representation theory of finite groups over the complex numbers, and also the unifying principles which are common to the representation theory of a wider range of algebraic structures. By learning the key concepts of representation theory you will also start to appreciate the power of category-theoretic approaches to mathematics. The mental framework you will acquire from this unit of study will enable you both to solve computational problems in linear algebra and to create new mathematical theory.

- **MATH4315 Variational Methods (6 credits, Semester 2)**

Variational and spectral methods are foundational in mathematical models that govern the configurations of many physical systems. They have wide-ranging applications in areas such as physics, engineering, economics, differential geometry, optimal control and numerical analysis. In addition they provide the framework for many important questions in modern geometric analysis. This unit will introduce you to a suite of methods and techniques that have been developed to handle these problems. You will learn the important theoretical advances, along with their applications to areas of contemporary research. Special emphasis will be placed on Sobolev spaces and their embedding theorems, which lie at the heart of the modern theory of partial differential equations. Besides engaging with functional analytic methods such as energy methods on Hilbert spaces, you will also develop a broad knowledge of other variational and spectral approaches. These will be selected from areas such as phase space methods, minimax theorems, the Mountain Pass theorem or other tools in the critical point theory. This unit will equip you with a powerful arsenal of methods applicable to many linear and nonlinear problems, setting a strong foundation for understanding the equilibrium or steady state solutions for fundamental models of applied mathematics.

- **MATH4411 Applied Computational Mathematics (6 credits, Semester 1)**
  Computational mathematics fulfils two distinct purposes within Mathematics. On the one hand the computer is a mathematician's laboratory in which to model problems too hard for analytical treatment and to test existing theories; on the other hand, computational needs both require and inspire the development of new mathematics. Computational methods are an essential part of the tool box of any mathematician. This unit will introduce you to a suite of computational methods and highlight the fruitful interplay between analytical understanding and computational practice. In particular, you will learn both the theory and use of numerical methods to simulate partial differential equations, how numerical schemes determine the stability of your method and how to assure stability when simulating Hamiltonian systems, how to simulate stochastic differential equations, as well as modern approaches to distilling relevant information from data using machine learning. By doing this unit you will develop a broad knowledge of advanced methods and techniques in computational applied mathematics and know how to use these in practice. This will provide a strong foundation for research or further study.

- **MATH4412 Advanced Methods in Applied Mathematics (6 credits, Semester 2)**
  Mathematical approaches to many real-world problems are underpinned by powerful and wide ranging mathematical methods and techniques that have become standard in the field and should be in the toolbag of all applied mathematicians. This unit will introduce you to a suite of those methods and give you the opportunity to engage with applications of these methods to well-known problems. In particular, you will learn both the theory and use of asymptotic methods which are ubiquitous in applications requiring differential equations or other continuous models. You will also engage with methods for probabilistic models including information theory and stochastic models. By doing this unit you will develop a broad knowledge of advanced methods and techniques in applied mathematics and know how to use these in practice. This will provide a strong foundation for using mathematics in a broad sweep of practical applications in research, in industry or in further study.

- **MATH4413 Applied Mathematical Modelling (6 credits, Semester 1)**
  Applied Mathematics harnesses the power of mathematics to give insight into phenomena in the wider world and to solve practical problems. Modelling is the key process that translates

a scientific or other phenomenon into a mathematical framework through applying suitable assumptions, identifying important variables and deriving a well-defined mathematical problem. Mathematicians then use this model to explore the real-world phenomenon, including making predictions. Good mathematical modelling is something of an art and is best learnt by example and by writing, refining and analysing your own models. This unit will introduce you to some classic mathematical models and give you the opportunity to analyse, explore and extend these models to make predictions and gain insights into the underlying phenomena. You will also engage with modelling in depth in at least one area of application. By doing this unit you will develop a broad knowledge of advanced mathematical modelling methods and techniques and know how to use these in practice. This will provide a strong foundation for applying mathematics and modelling to many diverse applications and for research or further study.

- **MATH4414 Advanced Dynamical Systems (6 credits, Semester 2)**
  In applied mathematics, dynamical systems are systems whose state is changing with time. Examples include the motion of a pendulum, the change in the population of insects in a field or fluid flow in a river. These systems are typically represented mathematically by differential equations or difference equations. Dynamical systems theory reveals universal mechanisms behind disparate natural phenomena. This area of mathematics brings together sophisticated theory from many areas of pure and applied mathematics to create powerful methods that are used to understand and control the dynamical building blocks which make up physical, biological, chemical, engineered and even sociological systems. By doing this unit you will develop a broad knowledge of methods and techniques in dynamical systems, and know how to use these to analyse systems in nature and in technology. This will provide a strong foundation for using mathematics in a broad sweep of applications and for research or further study.

- **MATH4511 Arbitrage Pricing in Continuous Time (6 credits, Semester 1)**
  The aim of Financial Mathematics is to establish a theoretical background for building models of securities markets and provides computational techniques for pricing financial derivatives and risk assessment and mitigation. Specialists in Financial Mathematics are widely sought after by major investment banks, hedge funds and other, government and private, financial institutions worldwide. This course is foundational for honours and masters programs in Financial Mathematics. Its aim is to introduce the basic concepts and problems of securities markets and to develop theoretical frameworks for pricing financial products and hedging the risk associated with them. This unit will focus on two ideas that are fundamental for Financial Mathematics. You will learn how the concept of arbitrage and the concept of martingale measure provide a unified approach to a large variety of seemingly unrelated problems arising in practice. You will also learn how to use the wide range of tools required by Financial Mathematics, including stochastic calculus, partial differential equations, optimisation and statistics. By doing this unit, you will learn how to formulate problems that arise in finance as mathematical problems and how to solve them using the concepts of arbitrage and martingale measure.

- **MATH4512 Stochastic Analysis (6 credits, Semester 2)**
  Capturing random phenomena is a challenging problem in many disciplines from biology, chemistry and physics through engineering to economics and finance. There is a wide spectrum of problems in these fields, which are described using random processes that evolve with

time. Hence it is of crucial importance that applied mathematicians are equipped with tools used to analyse and quantify random phenomena. This unit will introduce an important class of stochastic processes, using the theory of martingales. You will study concepts such as the Ito stochastic integral with respect to a continuous martingale and related stochastic differential equations. Special attention will be given to the classical notion of the Brownian motion, which is the most celebrated and widely used example of a continuous martingale. By completing this unit, you will learn how to rigorously describe and tackle the evolution of random phenomena using continuous time stochastic processes. You will also gain a deep knowledge about stochastic integration, which is an indispensable tool to study problems arising, for example, in Financial Mathematics.

- **MATH4513 Topics in Financial Mathematics (6 credits, Semester 2)**
  Securities and derivatives are the foundation of modern financial markets. The fixed-income market, for example, is the dominant sector of the global financial market where various interest-rate linked securities are traded, such as zero-coupon and coupon bonds, interest rate swaps and swaptions. This unit will investigate short-term interest rate models, the Heath-Jarrow-Morton approach to instantaneous forward rates and recently developed models of forward London Interbank Offered Rates (LIBORs) and forward swap rates. You will learn about pricing and hedging of credit derivatives, another challenging and practically important problem and become familiar with stochastic models for credit events, dependent default times and credit ratings. You will learn how to value and hedge single-name and multi-name credit derivatives such as vulnerable options, corporate bonds, credit default swaps and collateralized debt obligations. You will also learn about the most recent developments in Financial Mathematics, such as robust pricing and nonlinear evaluations. By doing this unit, you will get a solid grasp of mathematical tools used in valuation and hedging of fixed income securities, develop a broad knowledge of advanced quantitative methods related to interest rates and credit risk and you will learn to use powerful mathematical tools to address important real-world quantitative problems in the finance industry.

- **STAT4021 Stochastic Processes and Applications (6 credits, Semester 1)**
  A stochastic process is a mathematical model of time-dependent random phenomena and is employed in numerous fields of application, including economics, finance, insurance, physics, biology, chemistry and computer science. In this unit you will rigorously establish the basic properties and limit theory of discrete-time Markov chains and branching processes and then, building on this foundation, derive key results for the Poisson process and continuous-time Markov chains, stopping times and martingales. You will learn about various illustrative examples throughout the unit to demonstrate how stochastic processes can be applied in modelling and analysing problems of practical interest, such as queuing, inventory, population, financial asset price dynamics and image processing. By completing this unit, you will develop a solid mathematical foundation in stochastic processes which will become the platform for further studies in advanced areas such as stochastic analysis, stochastic differential equations, stochastic control and financial mathematics.

- **STAT4022 Linear and Mixed Models (6 credits, Semester 1)**
  Classical linear models are widely used in science, business, economics and technology. This unit will introduce the fundamental concepts of analysis of data from both observational studies and experimental designs using linear methods, together with concepts of collection of data and design of experiments. You will first consider linear models and regression methods

with diagnostics for checking appropriateness of models, looking briefly at robust regression methods. Then you will consider the design and analysis of experiments considering notions of replication, randomisation and ideas of factorial designs. Throughout the course you will use the R statistical package to give analyses and graphical displays. This unit includes material in STAT3022 Applied Linear Models, but has an additional component on the mathematical techniques underlying applied linear models together with proofs of distribution theory based on vector space methods.

- **STAT4023 Theory and Methods of Statistical Inference (6 credits, Semester 2)**
  In today's data-rich world, more and more people from diverse fields need to perform statistical analyses, and indeed there are more and more tools to do this becoming available. It is relatively easy to "point and click" and obtain some statistical analysis of your data. But how do you know if any particular analysis is indeed appropriate? Is there another procedure or workflow which would be more suitable? Is there such a thing as a "best possible" approach in a given situation? All of these questions (and more) are addressed in this unit. You will study the foundational core of modern statistical inference, including classical and cutting-edge theory and methods of mathematical statistics with a particular focus on various notions of optimality. The first part of the unit covers aspects of distribution theory which are applied in the second part which deals with optimal procedures in estimation and testing. The framework of statistical decision theory is used to unify many of the concepts that are introduced in this unit. You will rigorously prove key results and apply these to real-world problems in laboratory sessions. By completing this unit, you will develop the necessary skills to confidently choose the best statistical analysis to use in many situations.

- **STAT4028 Probability and Mathematical Statistics (6 credits, Semester 1)**
  Probability Theory lays the theoretical foundations that underpin the models we use when analysing phenomena that involve chance. This unit introduces the students to modern probability theory and applies it to problems in mathematical statistics. You will be introduced to the fundamental concept of a measure as a generalisation of the notion of length and Lebesgue integration which is a generalisation of the Riemann integral. This theory provides a powerful unifying structure that bring together both the theory of discrete random variables and the theory of continuous random variables that were introduce to earlier in your studies. You will see how measure theory is used to put other important probabilistic ideas into a rigorous mathematical framework. These include various notions of convergence of random variables, 0-1 laws, and the characteristic function. You will then synthesise all these concepts to establish the Central Limit Theorem and also verify important results in Mathematical Statistics. These involve exponential families, efficient estimation, large-sample testing and Bayesian methods. Finally you will verify important convergence properties of the expectation-maximisation (EM) algorithm. By doing this unit you will become familiar with many of the theoretical building blocks that are required for any in-depth study in probability or mathematical statistics.

6.3 List 4

5000-level units available in the School of Mathematics and Statistics except STAT5002, STAT5003, DATA5810 or DATA5811.

6.4 List 5

4000-level or 5000-level units at other Schools at the University

7 Project

7.1 General information on projects

Each student is expected to have made a choice of a project and supervisor well before the beginning of the first semester (or the beginning of the second semester for students starting in July). Students are welcome to consult on this matter with the Honours coordinator. The Honours coordinator must be informed of the choice of supervisor before the start of the program.

Work on the project should start as soon as possible but no later than the start of the semester. The break between the semesters is often an excellent time to concentrate on your research but you should make sure you make continuous progress on your research throughout the year. To ensure that, students should consult their appointed supervisor regularly, in both the researching and writing of the work.

Lists of suggested project topics for Data Science Honours are provided in Section 7.2 below. Prospective students interested in any of these topics are encouraged to discuss them with the named supervisors as early as possible. **Keep in mind that this list is not exhaustive.** Students can work on a project of their own topic provided they secure in advance the supervision of a member of the School of Mathematics and Statistics or the School of Computer Science (including emeritus staff) and provided they receive the approval of the Honours Coordinator.

Three copies of the essay typed and bound, as well an electronic copy must be submitted to the Honours coordinator before the beginning of the study vacation at the end of your last semester. The exact date will be made known.

It is recommended that you go through the following checklist before submitting your thesis:

- Is there an adequate introduction?
- Have the chapters been linked so that there is overall continuity?
- Is the account self-contained?
- Are the results clearly formulated?
- Are the proofs correct? Are the proofs complete?
- Have you cited all the references?

7.2 Proposed project topics

The projects are divided into two categories: Data Science and Statistics. If you are a Data Science student but you prefer a project that is listed under Statistics or the other way around then feel free to talk the relevant supervisor about it. Projects are given in alphabetical order depending on the name of the supervisor.

- **Title**: The cost of (differential) privacy in data analysis.

  Supervisor: Dr Clément Canonne (School of Computer Science)

  The interest in privacy-preserving approaches to data analysis has sharply increased over the past decade. One leading notion of privacy, "differential privacy" (DP), provides a mathematical framework to design and analyse algorithms, with strong and provable privacy guarantees. This project focuses on exploring the trade-offs between accuracy and privacy achievable under DP for some standard (but yet not fully understood in the private regime) statistical tasks.

- **Title**: Novel construction of decoys for controlling the FDR in mass spectrometry.

  Supervisor: A/Prof Uri Keich (School of Mathematics and Statistics)

  Project: In a shotgun proteomics experiment tandem mass spectrometry is used to identify the proteins in a sample. The identification begins with associating with each of the thousands of the generated peptide fragmentation spectra an optimal matching peptide among all peptides in a candidate (target) database. Unfortunately, the resulting list of optimal peptide-spectrum matches contains many incorrect, random matches. The canonical way to control the associated type I error is by controlling the false discovery rate (FDR) through target-decoy competition. Specifically, the matches to peptides in the target database are contrasted with matches to pseudo-peptides in a decoy database. Invariably, the decoy peptides are generated by shuffling or by reversing the target peptides - methods that have limitations that have recently came to light. In this project we will look at alternative methods for constructing decoys. No prior understanding of proteomics is required.

- **Title**: Interactive platforms for communicating high-dimensional and complex data.

  Supervisor: Dr Shila Ghazanfar (School of Mathematics and Statistics)

  Project: Recent technological developments in single cell RNA-sequencing and spatially resolved genomics have resulted in immense datasets corresponding to hundreds of thousands of observed cells and thousands of measured features. Effectively displaying these data for interactive use poses a challenge to data scientists to i) effectively communicate these data, ii) ensure fast generation of visualisations, and iii) facilitate analytical tasks on the fly. This project will examine the use and efficacy of interactive Shiny apps in R for describing high-dimensional and complex single cell data. While there are comprehensive resources for creating Shiny apps [1], it is unclear to what extent this technology is being used in the single cell research community, and in what ways. This project will comprise a literature survey of existing Shiny apps and their characteristics, an assessment of the nature and efficiencies of components of these apps, and result in a set of guiding principles and strategies for creating an interactive app for a given single cell dataset.

  – https://mastering-shiny.org/

– https://github.com/federicomarini/awesome-expression-browser

- **Title**: Incorporating subcellular, cellular and tissue level features for imaging-based spatial genomics.

  Supervisor: Dr Shila Ghazanfar (School of Mathematics and Statistics)

  Project: Recent technological developments in single cell RNA-sequencing and spatially resolved genomics have resulted in immense datasets corresponding to hundreds of thousands of observed cells and thousands of measured features. Current spatial omics approaches use gene measurements associated with each cell/spot to integrate data between spatial omics and dissociated single cell technologies. The use of multiscale features, covering subcellular and supercellular resolutions, has the potential to significantly stabilize mapping between datasets and extract new insights from the data. However, the use and design of such features is challenging due to the induced mosaic data integration problem. Algorithms exist for horizontal and vertical integration, but mosaic data integration is an area of active development, including StabMap. These new methods promise additional features not captured in all modalities to nevertheless be incorporated in downstream analysis. This project aims to address the above challenge by developing tools for generating such multiscale features, assessing their explanatory power, and exploring their ability to be integrated across multiple distinct datasets.

- **Title**: Assessing continuous changes in correlation structure for replicated single cell data.

  Supervisor: Dr Shila Ghazanfar (School of Mathematics and Statistics)

  Project: Single-cell genomics has transformed our ability to examine cell fate choice. Examining cells along a computationally ordered 'pseudotime' or across spatial coordinates offers the potential to unpick subtle changes in variability and covariation among key genes. A key challenge now is how to perform such continuous differential variation or covariation testing within a multi-sample and potentially multi-condition experiment. This project will examine the use of generalised additive models (GAMs) towards assessing changes in covariation patterns along pseudotime and spatial coordinates using a mixture of real-world and simulated data.

- **Title**: Assessing error propagation in mosaic data integration.

  Supervisor: Dr Shila Ghazanfar (School of Mathematics and Statistics)

  Project: This project aims to perform error estimation for gene expression imputation and spatial position inference from mosaic data integration of single cell spatial genomics. This proposed framework will try to extract measures of ambiguity for cells' joint integration. These noise estimates will be obtained by identifying, for each cell, the relative size of the cell's neighbourhoods in the joint network, repeatedly estimated under varying bootstrapping. This will result in three key outputs, first, the estimated gene expression and associated variances for each gene for imputation of each spatial cell; second, the estimated spatial position and associated posterior distribution (represented as a field) for each non-spatial cell; and third, an ambiguity metric, represented by the relative proportion of plausible cells an either spatial or non-spatial cell is most similar to. The variance estimates will be used as feature weights for downstream analysis (e.g. differential expression) to enable propagation of error. This investigation will go towards further understanding the underlying data structure and information content of these distinct data modalities.

- **Title**: Feature normalisation for barcoded spatial gene expression.

  Supervisor: Dr Shila Ghazanfar (School of Mathematics and Statistics)

  Project: This project will develop an effective framework for analysis of data from imaging-based spatial genomics at single cell resolution by performing feature normalisation, correcting for optical field of view batches, and enabling spatial analyses to understand cell interactions. Compared to traditional single molecule RNA FISH technologies, where a single colour corresponds directly to a single gene, spatial genomics techniques like seqFISH use barcoding strategies to encode hundreds of genes into very few chemical hybridisation rounds. This presents a statistical challenge since there are differing levels of similarity in the pseudocolour barcode between genes, creating an especially complex statistical relationship between the features (genes), resulting in the potential for induced positive correlation among genes with similar pseudocolour barcodes. This project will devise strategies to correcting for such induced positive correlation through the use of factor models on the observed gene-gene correlation matrix.

- **Title**: Novel approaches for diagonal data integration of single cell -omics data.

  Supervisor: Dr Shila Ghazanfar (School of Mathematics and Statistics)

  Project: The problem of data integration can be classified into four main groups, horizontal, vertical, mosaic and diagonal [ref]. Diagonal data integration refers to the task of placing observations from distinct datasets which capture disparate features into a common low-dimensional embedding. This project will examine current approaches for performing such integration tasks for single cell -omics data including what additional information or assumptions are required to achieve this, as well as how to effectively assess its quality. Further, this project will aim to devise new approaches for diagonal data integration, for example by assessing the relationships of feature sets that explain most variation in each dataset.

- **Title**: A Bayesian Copula Directional Dependence method for causal inference.

  Supervisor: Dr Clara Grazian (School of Mathematics and Statistics)

  Project: Modelling and understanding directional gene networks is a major challenge in biology as they play an important role in the architecture and function of genetic systems. Copula directional Dependence (CDD) can measure the directed connectivity among genes without any strict requirements of distributional and linearity assumptions. Furthermore, copulas can achieve that by isolating the dependence structure of a joint distribution. In this work, a novel extension of the frequentist CDD in the Bayesian setting is introduced. The new method is compared against the frequentist CDD and validated on six gene interactions, three coming from a mouse scRNA-seq dataset and three coming from a bulk epigenome dataset. The results illustrate that the novel proposed Bayesian CDD was able to identify four out of six true interactions with increased robustness compared to the frequentist method. Therefore, the Bayesian CDD can be considered as an alternative way for modeling the information flow in gene networks.

- **Title**: vPET-ABC: Voxel-wise approximate Bayesian inference for parametric imaging.

  Supervisor: Dr Clara Grazian (School of Mathematics and Statistics)

  Project: We recently developed a method, called PET-ABC, that provides complete Bayesian statistical analysis of ROI-based dynamic PET data. The aim of this work was to extend the

method to voxel-based analyses (vPET-ABC), with an initial focus on parametric imaging of neurotransmitter (NT) release in PET activation studies. In this study, the kinetic model used in conjunction with vPET-ABC was the linear-parametric neurotransmitter PET (lp-ntPET) model. This model, which incorporates time-varying kinetic parameters, describes the effect of neurotransmitter changes on dynamic receptor-ligand data during PET activation studies. The vPET-ABC pipeline produces reliable estimates of voxel-based parameters and their associated posterior probability distributions. It also computes model probability at the voxel level, where the two alternative models under consideration are lp-ntPET, which allows for an activation, and the more parsimonious Multilinear Reference Tissue Model (MRTM) which does not. Initials results, obtained from a simulated GATE 4D [11C]raclopride scan with realistic noise, demonstrated that vPET-ABC can provide insightful information about NT release at the voxel level, including the reliability of parameter estimates, which is important for reliably identifying subtle activations in noisy PET data. We believe the technique will be equally useful in the analysis of total body dynamic PET data, where the applicability of a single kinetic model to all tissues cannot be assumed.

- **Title**: The logical foundations and empirical hurdles to an artificial intelligence with a 'Theory of Mind'.

  Supervisor: Dr Michael Harre (School of Computer Science)

  Project: One of the most important psychological properties of human intelligence is how we infer the internal mental states of other people, this is called our Theory of Mind (ToM). We use this mental apparatus every day all day when we are dealing with other people and it makes our interactions with others easier and more efficient as we know from many cognitive studies that show how difficult it is to navigate our social and professional environment for those people who don't have this capacity. It's also the case that some other animals may have a limited version of a ToM, but this is itself a highly contested area of research because of how difficult it is to convincingly show or prove that another animal is thinking about our internal mental states. So given its usefulness in humans and the difficulty in empirically demonstrating its existence in other animals, can we even theoretically build a ToM into a modern artificial intelligence? This project will explore the theoretical limitations of 'proving' that another animal or AI has a ToM and to what extent our ToM can be directly derived from observable data. This will suit someone with a theoretical, mathematical, or logical approach to problems.

- **Title**: Information theory and criticality in the analysis of 'bio-inspired' and 'machine' intelligence.

  Supervisor: Dr Michael Harre (School of Computer Science)

  Project:How do we know that a collection of interacting elements in a system, like neurons in the brain, insects in a hive, or logic gates on a circuit board, are carrying out some sort of computation? Even more challenging is how do we know when one system is doing more computation, or computing more effectively, than another system? Finding objective measures for the occurrence and effectiveness of computation in many different classes of systems, many of which don't appear to be related to one another at all, is a challenging task. This project will evaluate a number of different measures from information theory to test how well a computing system performs depending on how the system is structured and whether or not there are universal measures of computation that allow us to test a system's computational power before we implement it in a real system.

- **Title**: Sufficient dimension reduction for genomics data.

Supervisor: Dr Linh Nghiem and Dr Shila Ghazanfar (School of Mathematics and Statistics)

Project: Dimension reduction is an essential step for processing large genomics datasets where tens of thousands of features are measured simultaneously. At the same time, samples can be labelled with biologically relevant outcomes (e.g. disease or treatment status, or cell type labels of single cells). Among many dimension reduction methods that have been developed, the class of sufficient dimension reduction methods has a unique advantage in that they both reduce the dimensionality of the feature matrix and ensure no information to predict the target outcome from the original features is lost in the reduced features. This project focuses on reviewing the most popular sufficient dimension reduction methods and comparing their performance with other supervised and unsupervised dimension reduction techniques in single cell, spatial and bulk genomics datasets.

- Rodrigues, Sabrina A., Richard Huggins, & Benoit Liquet (2022). Central Subspaces Review: Methods and Applications. Statistics Surveys 16 (none): 210–37.

- **Title**: Data-intensive science to understand the molecular aetiology of disease.

  Supervisor: Dr Ellis Patrick (School of Mathematics and Statistics)

  Project: Biotechnological advances have made it possible to monitor the expression levels of thousands of genes and proteins simultaneously promising exciting, groundbreaking discoveries in complex diseases. This project will focus on the application and/or development of statistical and machine learning methodology to analyse a high-dimensional biomedical experiment. My group works on projects spanning multiple diseases including melanoma, acute myeloid leukemia, Alzheimer's disease, multiple sclerosis and HIV. We also work with various high-throughput technologies including single-cell RNA-Seq, flow cytometry, CyTOF, CODEX imaging and imaging mass cytometry.

- **Title**: Identifying changes in network structure to identify complex cellular interactions.

  Supervisor: Dr Ellis Patrick (School of Mathematics and Statistics)

  Project: You will develop a novel network based hypothesis testing framework to detect if cells are collocating in high-dimensional cellular imaging data. This framework will inherently overcome some complications that arise in concordance based tests due to image noise and tissue inhomogeneity while also identifying the relationships that are most descriptive of the biology. The Pearson correlation coefficient approach (Manders et al. 1992) is the simplest and hence most widely used method for assessing cell-type colocalisation. We will generalise the Pearson correlation coefficient and Manders overlap coefficient methods for use with multiple markers by using partial correlation matrices, an approach I have applied to gene expression datasets (Patrick et al. 2017) for decomposing gene regulatory networks. By conceptualising colocalisation in terms of partial correlation matrices, you will test for colocalisation and changes in colocalisation in three ways:

  - You will use a sparse graphical lasso to identify cell-type markers that are colocalised accounting for the behaviour of all other markers. Following the sparsity constraints you will use post-selective inference for Gaussian graphical models (GSell et al. 2013) to assign significance to each cell-cell interaction.

- Next, you will adapt a two-sample network inference approach typically used for brain connectivity analysis (Xia et al. 2017) to detect if colocalisation between two-cells, after accounting for the interactions between all other cells, is changing.
- Finally, two-sample network inference methods (Ghoshdastidar et al. 2018) can be adjusted to detect global changes in colocalisation between two conditions. This will produce a novel hypothesis testing framework to detect if whole systems of cells are interacting in distinct ways under different conditions.

- **Title**: Stochastic gradient descent with randomised reshuffling.

  Supervisor: Dr Lindon Roberts (School of Mathematics and Statistics)

  Project: Many problems in data science can be reduced to optimising a very large sum of functions, and methods based on stochastic gradient descent (SGD) are the most popular choices. Standard analysis of SGD assumes that the stochastic gradients (formed by selecting a random data point) are unbiased and sampled independently at each iteration. However better performance is usually observed if we shuffle the data points and use each one sequentially. This is standard practice despite it violating the usual assumptions for SGD. In this project we will investigate the convergence properties of randomised reshuffling for finite sum optimisation.

  - K. Mishchenko, A. Khaled & P. Richtarik. Random Reshuffling: Simple Analysis with Vast Improvements. 34th Conference on Neural Information Processing Systems (NeurIPS 2020).
  - L. M. Nguyen, Q. Tran-Dinh, D. T. Phan, P. H. Nguyen & M. van Dijk. A Unified Convergence Analysis for Shuffling-Type Gradient Methods. Journal of Machine Learning Research, 22 (2021), p. 1-44.
  - M. Gürbüzbalaban, A. Ozdaglar & P. A. Parrilo. Why random reshuffling beats stochastic gradient descent. Mathematical Programming, 186 (2021), p. 49-84.
  - K. Mishchenko, A. Khaled & P. Richtarik. Proximal and Federated Random Reshuffling. Proceedings of the 39th International Conference on Machine Learning (2022).

- **Title**: Stochastic bilevel optimisation.

  Supervisor: Dr Lindon Roberts (School of Mathematics and Statistics)

  Project: Several data science problems, most notably hyperparameter tuning, is an example of bilevel optimisation - optimising a function depends on the solution of a different optimisation problem. This is a complicated problem, particularly when the inner problem is difficult to solve. In this project we will look at extensions of stochastic optimisation algorithms to bilevel optimisation (where both the outer and inner problems are solved using stochastic methods).

  - C. Crockett & J. A. Fessler. Bilevel methods for image reconstruction. arXiv preprint arXiv:2109.09610 (2021).
  - K. Ji, J. Yang & Y. Liang. Bilevel Optimization: Convergence Analysis and Enhanced Design. Proceedings of the 38 th International Conference on Machine Learning (2021).

- T. Chen, Y. Sun, Q. Xiao & W. Yin. A Single-Timescale Method for Stochastic Bilevel Optimization. Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS) 2022.
- P. Khanduri, S. Zeng, M. Hong, H.-T. Wai, Z. Wang & Z. Yang. A Near-Optimal Algorithm for Stochastic Bilevel Optimization via Double-Momentum. 35th Conference on Neural Information Processing Systems (NeurIPS 2021).

- **Title**: Logspline-based estimation of a centre of symmetry.

  Supervisor: Dr Michael Stewart (School of Mathematics and Statistics)

  Project: The R package `logspline` implements a density estimation method due to Stone (1990); Kooperberg and Stone (1992); Stone *et al.* (1997) which uses cubic splines to approximate the logarithm of the density. The choice of number and location of knots is made in a data-driven manner, designed to make the estimation of the density itself as good as possible. Efficient semiparametric estimation of the centre of symmetry of a density involves first estimating the location score function (derivative of the log-density) and then solving the resultant "estimated" score equation. The logspline method can be used for this, but its performance can possibly be improved for this task by "tweaking" the algorithm used to choose the number and location of spline knots. This project will have both a theoretical and computational component. The theory will be studied to determine how the algorithm may be adjusted to improve estimation of the derivative of the log-density under symmetry. A second aim of the project is to implement this improved algorithm and produce an efficient R package, possibly implementing the main routine in a fast low-level language, e.g. C, C++ or Fortran.

  - C. Kooperberg & C. J. Stone. Logspline density estimation for censored data. Journal of Computational and Graphical Statistics, 301–328, (1992).
  - C. J. Stone. Large-sample inference for log-spline models. Annals of Statistics, 18(2): 717–741 (1990).
  - C. J. Stone, M. H. Hansen, C. Kooperberg, & Y. K. Truong.Polynomial splines and their tensor products in extended linear modeling `with discussion and a rejoinder by the authors and J.-Z. Huang.`. Annals of Statistics, 25(4): 1371–1470 (1997).

- **Title**: Fitting nested normal mixtures.

  Supervisor: Dr Michael Stewart (School of Mathematics and Statistics)

  Project: A difficult problem in unsupervised learning/clustering is to answer the question: "how many clusters?". Various approaches exist for tackling this, one of which is model-based clustering where we posit that the population consists of a mixture of (multivariate) normal "subpopulations", each one roughly corresponding to a "cluster". Many methods involve fitting k-component mixtures, often by maximum likelihood, and then comparing fits over different values of k, often by incorporating a penalty (which increases with k in some way). A non-trivial task in this context, particularly in high dimensions, is to obtain such fits for a sequence of k values of interest. This project will involve surveying different approaches currently used and comparing them to a novel approach involving a "hybrid" approach combining the EM-algorithm (Dempster *et al.*, 1977) with a variant of the intra-simplex direction method (ISDM) of Lesperance and Kalbfleisch (1992).

- A. P. Dempster, N. M. Laird, & D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm `with discussion`. Journal of the Royal Statistical Society - Series B, 39(1): 1–38, (1977).
- M. L. Lesperance & J. D. Kalbfleisch. An algorithm for computing the nonparametric mle of a mixing distribution. Journal of the American Statistical Association, 87(417): 120–126 (1992).
- C. J. Stone, M. H. Hansen, C. Kooperberg, & Y. K. Truong.Polynomial splines and their tensor products in extended linear modeling. Annals of Statistics, 25(4): 1371–1470 (1997).

- **Title**: What's most important in experimental design?

  Supervisor: Dr Garth Tarr (School of Mathematics and Statistics)

  Project: When fitting models, we should try to account for as many sources of variation as possible. To what extent should we let this drive the design of the experiment? This project will explore issues around experimental design in multilevel models. For example, how important is the pen effect relative to individual animal variation when predicting outcomes at the animal level. Another example is to assess the impact the quality of the first sample has on consumer sensory trials when consumers are asked to evaluate a sequence of samples. The student will develop recommendations, best practice guidelines and a power calculation app specific to the common types of experiments that are run by Meat and Livestock Australia.

- **Title**: Finite sample performance of robust location estimators.

  Supervisor: Dr Garth Tarr (School of Mathematics and Statistics)

  Project: Consumer sensory scores are typically constrained within bounded intervals, for example when asked to give a score out of 100, however the measurements often exhibit outliers within that bounded interval. This project will investigate finding an "optimal" robust location estimator for bounded data with a focus on small sample performance. This project will consider various univariate and multivariate robust location estimators and assess their small sample performance. This work may be extended to compare robust multilevel models against simpler robustly summarised models. You will have access to an extensive sensory database with which to evaluate various techniques and put forward recommendations that will help shape the future of consumer sensory evaluation of lamb and beef. Development of more efficient processes and protocols for conducting and summarising consumer sensory scores will lead to substantial savings for future research experiments and/or enable more research to be done with the same amount of funding.

- **Title**: Generative modelling and optimisation in federated learning.

  Supervisor: Dr Nguyen Tran (School of Computer Science)

  Project: This project will explore several aspects of distributed optimisation and generative modelling in the context of federated learning. Specifically, we will investigate how data stored on local devices in a network can be leveraged to generate new, high-quality examples (e.g., images, texts, audio, etc.), all while preserving the privacy of these devices. We will explore different sampling techniques and analyse them through the lens of optimisation in order to design novel generative models. The nature of decentralised learning presents several major challenges such as limited communication and statistical heterogeneity, which

complicates the design and analysis of these models. The project will aim to address these challenges both from theoretical and an empirical perspective.

- **Title**: FeDEQ (Federated Deep Equilibrium Models).

  Supervisor: Dr Nguyen Tran (School of Computer Science)

  Project: Federated Learning (FL) has emerged as a cutting-edge distributed machine learning paradigm in which a network of clients collaborates to build a shared global model without revealing their private data. Unlike traditional centralized machine learning, FL obviates the need of sending raw data enabling clients to train and send only local models to the server, thereby introducing a reliable infrastructure which not only retains the trust and privacy for machine learning but also reduces the computation and communication costs. However, the inherent statistical diversity in clients' devices restricts the global model from delivering good performance on each client's task. Additionally, the resource-constrained nature of clients' devices complicates the training of FL models. In this project, we address these challenges by proposing a new scheme for personalizing FL based on representation learning, namely FeDEQ. In particular, we leverage the memory efficiency and representation power of Deep Equilibrium Models (DEQ), one of the emerging implicit deep learning frameworks, to design a FL algorithm that learns a common representation in the context of heterogeneous distributed data. By sharing the representation, clients then efficiently personalize their small local model to adapt with the local data. Furthermore, we show the inference of the proposed framework in many FL applications such as computer vision, sequence modelling, etc.

- **Title**: Multimedia information retrieval.

  Supervisor: A/Prof Zhiyong Wang (School of Computer Science)

  Project: Multimedia information retrieval (MIR) aims to empower humans to effectively and efficiently access big multimedia data (e.g., audio, image, video, and text) being increasingly acquired in almost every domain. It underpins many digital services we are using every day, such as search engines and virtual assistant. This project aims to develop novel underlying techniques in the field, such as content based retrieval, cross-modal retrieval, video summarization, text summarization, object recognition, action recognition, visual captioning, question and answering, dialogue systems, recommender systems, deepfake detection, and multimedia data mining. Students will gain comprehensive knowledge in multimedia content analysis, information retrieval, natural language processing, knowledge discovery, data mining, pattern recognition, and machine learning.

  Technical skills: strong programming skills and math, knowledge on deep learning preferred

- **Title**: Multimedia content understanding.

  Supervisor: A/Prof Zhiyong Wang (School of Computer Science)

  Project: Multimedia data has been increasingly acquired in almost every domain, ranging from our daily experiences through smart phones, lecturing videos, footage of security surveillance, music performances, and sports videos, earth observation, to medical images and surgical videos. This has created great opportunities for knowledge discovery from massive multimedia data and for enabling new interactions such as VR and AR among humans, physical worlds, and virtual worlds with intelligent understanding of physical environments. This project aims to address the challenges in the field, including object detection, tracking, human action recognition and prediction, even detection, affective analysis, multimedia

forensics, multimodal understanding, and 3D reconstruction and synthesis. Students will gain comprehensive knowledge in multimedia data processing, computer vision, 3D vision, pattern recognition, and machine learning.

Technical skills: strong programming skills and math, knowledge on deep learning preferred

- **Title**: Multimedia content creation and synthesis.

  Supervisor: A/Prof Zhiyong Wang (School of Computer Science)

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

  Technical skills: strong programming skills and math, knowledge on deep learning and fluid dynamics preferred

- **Title**: Human motion analysis, modeling, animation, and synthesis.

  Supervisor: A/Prof Zhiyong Wang (School of Computer Science)

  Project: Humans are the focus in most activities; hence investigating human motion has been driven by a wide range of applications such as visual surveillance, action recognition, 3D human motion capture, character animation, novel human computer interaction in VR and AR, sports innovation, and advanced medical diagnosis and treatment. This project aims to address the challenging issues of this area in realistic scenarios, including human tracking, motion detection, recognition, modelling, 3D reconstruction, animation, and synthesis to advance human behaviour analysis. Students will gain comprehensive knowledge in computer vision, 3D human motion modelling, computer graphics, and machine learning.

  Technical skills: strong programming skills and math, knowledge on deep learning preferred

- **Title**: Multimedia computing for medicine and health.

  Supervisor: A/Prof Zhiyong Wang (School of Computer Science)

  Project: With the advances of various sensing techniques, various multi-modal data such as different types of medical images, surgical videos, and brain signals have been widely used and health and medicine domains for disease prevention, diagnosis and personalized treatment. This project aims to investigate novel multimedia computing techniques, including identifying regions of interest such as tumorous tissues and surgical tools, surgical video analysis, and detecting and predicting disease progression with multi-modal data (e.g., medical images and EEG signals). The outcomes of this project could assist doctors and physicians in clinical assessments and treatments.

  Technical skills: strong programming skills and math, knowledge on deep learning preferred

- **Title**: Methods towards precision medicine.

  Supervisor: Prof Jean Yang (School of Mathematics and Statistics)

Project: Over the past decade, new and more powerful -omic tools have been applied to studying complex diseases such as cancer and generated a myriad of complex data. However, our general ability to analyse this data lags far behind our ability to produce it. This project is to develop a statistical method that delivers better prediction toward healthy aging by identifying a risk prediction framework that is interpretable. Students have a choice of creating a risk prediction method using one of the four case studies described below:

- *Case study I - Melanoma:* The Melanoma Institute of Australia (MIA) Stage III data collection is a unique set of multi-layered omics data (Mann et al. 2013; Jayawardana et al. 2016) with measured gene, protein, and microRNA expression and linked clinical and mutation data for subjects with Stage III lymph node metastases.
- *Case study II – Cardiovascular disease:* This case study is in collaboration with Prof Gemma Figtree's team who will provide access to the [BioHEART] dataset. This is a unique large scale multi-omics data with over 1,000 samples in lipidomics, proteomics and metabolomics and 250 in CyTOF already generated to allow evaluation of scalable models.
- *Case study III – Infectious disease COVID19:* There currently exists over 30 public COVID19 multi-omics datasets, including ten single-cell RNA-seq datasets with over 300 individuals and this work will involve a newly established collaboration with Laboratory of Data Discovery for health, $D^2 4H$) on global health protection.
- *Case study IV – Parkinson disease:* This case study is in collaboration with Prof Carolyn Sue (Kolling Institute). She will provide access to nearly 200 individual microbiome data with matched diet and nutrition information (with possible matched metabolomics data) to study the longitudinal impact on Parkinson disease (PD). Additional public microbiome and omics data on PD have also been curated.

- **Title**: Machine learning for kidney allocation.

  Supervisor: Prof Jean Yang (School of Mathematics and Statistics)

  Project: Kidney transplantation offers improved survival and quality of life for many patients with kidney failure compared to being on dialysis. Non-invasive -omics biomarkers may predict adverse events such as acute rejection after kidney transplantation and may be preferable to existing methods because of superior accuracy, timeliness and convenience. A recent study has shown that a set of gene signatures for acute rejection derived from a single study do not appear to provide adequate prediction in an independent cohort of transplant recipients. This project aims to develop an approach to integrate multiple gene signatures sets that improve the prediction performance of these markers and ensure the biomarkers behave consistently across multiple data platforms. This project will involve the evaluation of the new approach and communicate this information via an interactive web interface. The project will contribute to the widespread use of omics signatures in clinical practice.

- **Title**: Single cell Living benchmark for multi-sample studies.

  Supervisor: Prof Jean Yang (School of Mathematics and Statistics)

  Project: This project aims to develop the Single-Cell LIving Benchmarking (scLIB) platform - a framework for benchmarking data workflow and analytical choice for handling multi-sample and multi-condition cohort data. Globally, scientists are now generating data that

carry critical information for important clinical and public health applications. Many of these studies are designed to perform comparisons between two or more groups/conditions of samples, such as diseased and healthy individuals. Concurrently, methodological advances in single-cell research have risen at an incredible rate.

With the number of multi-sample data expecting to rise in the coming years (Petukhov et al., 2022), we will see the emergence of new analytical workflows and methods for performing comparative analyses between different groups of single-cell data. This has created a new data analytics challenge. Researchers are often overwhelmed with too many options and a steep learning curve to decide on the best collection of analytical approaches for performing their own comparative analysis in multi-conditional studies. In this project, you will address this challenge by developing a living benchmarking platform to systematically evaluate methods for comparative analytics in multi-conditional cohorts and identify the strengths and weaknesses of different methods in relation to the properties of their data. This new platform will provide (i) a gold standard for methodological researchers to develop methods and (ii) a recommendation guide to enable applied researchers to perform best-practice data analytics.

- **Title**: Reference-free signature for cross-species predictive modelling in nutriomics data.

  Supervisor: Prof Jean Yang (School of Mathematics and Statistics)

  Project: Globally, many communities are facing an aging population and an increase in health care prices; hence, big data on food and nutrition holds the promise to boost our well-being and prevent us from experiencing negative health consequences; reflecting the well-known saying "we are what we eat". This project will examine the concept of 'reference-free' signatures that constrain those features of the organismal system that most strongly influence physiological outcomes. If adequately defined, such signatures can be exploited as integration points for multi-omics predictive modelling using in house data and a large amount of publicly available "microbiome x environment x host" datasets. This project aims to construct a transferable model to integrate microbial features in both mouse study and human study. This is done by extending previous work in scJoint (Lin et al., 2022), a semi-supervised transfer learning method to construct a microbiome transfer learning (miJoint) to integrate both mouse and human microbiome communities with the following three steps. The scope of this project contains one or more than one of the following components:

  1. Construct microbial features between humans and mouse such as mapping to common pathways
  2. Joint nonlinear dimension reduction with a semi-supervised transfer learning approach between human and mouse data.
  3. A saliency map finding core microbial communities across humans and mouse.

- **Title**: Constructing network-based biomarkers which are critical for disease prognosis.

  Supervisor: Prof Jean Yang and Dr Ellis Patrick (School of Mathematics and Statistics)

  Project: With the advancements of single-cell sequencing technology, we are now able to explore cell identity at a resolution that was previously out of reach. From a statistical viewpoint, this data-type is very interesting as it has unusual data-structure which means there is an incredible amount of statistical methods development that needs to be completed

to make full use of this amazing technology. We propose a project below that explores state-of-the-art clustering and classification approaches that, while generalizable to other contexts, will be applied to tangible and translationally relevant biomedical questions.

Many classical approaches in classification are primarily based on single features that exhibit effect size difference between classes. Recently, we have demonstrated that approaches which use network-based features can be used to classify alternate subsets of patients as compared to those that use single-features. Building on our previous experience identifying network-based biomarkers (classifiers of disease) we will instead use cell-type specific networks generated from single-cell sequencing data. This process will allow us to construct network biomarkers that are specific for a cell-type of interest, are capable of assigning a score to a single individual and can be integrated with classification approaches such as DLDA, SVM, LASSO and Random Forests. Bootstrap and resampling will be used to ensure stability and robustness of identified features.

- **Title**: A data science approach to investigate human development.

  Supervisor: A/Prof Pengyi Yang (School of Mathematics and Statistics)

  Project: Recent generation of artificial human blastoids that recapitulate the early human embryos is a groundbreaking achievement. These technologies are transforming our understanding of early human development and hold promise to revolutionise regenerative medicine that utilises stem cell-derived tissues and organs. While a panel of experimental protocols have been established for generating human blastoids, the fidelity of these blastoids for modelling early human embryos remains to be assessed. This project aims to use a data analytics approach (e.g. machine learning, feature selection, statistical prediction) to investigate single cell transcriptomics data to assess the quality of artificial human blastoids in recapitulating early human embryos.

- **Title**: Integrative data analysis for understanding transitional cell states.

  Supervisor: A/Prof Pengyi Yang (School of Mathematics and Statistics)

  Project: The study of cell types has been propagated by the recent advancements in single cell technologies that have enabled unprecedented insight into cellular heterogeneity. Yet most studies have investigated discrete cell states, and the investigation of intermediate and transitional cell types has been hampered by challenges associated with identifying these rarer cell states. However, characterising these intermediate cell states is important to enhance our understanding of cell state transitions during embryonic development. The increasing number of single-cell fetal transcriptomic and epigenomic atlases enables us to begin interrogating this question. Extending on our recently developed computational method, Cepo (https://doi.org/10.1038/s43588-021-00172-2), this project will undertake a global approach to identify, investigate, and characterise transitional cell states.

- **Title**: Predict missing data modality from single-cell data.

  Supervisor: A/Prof Pengyi Yang (School of Mathematics and Statistics)

  Project: Recent advancement in single-cell multimodal sequencing technologies such as CITE-seq, SHARE-seq, REAP-seq, and TEA-seq has enabled the profiling of different modalities at the single-cell level. A promising application of this emerging multimodal technology is to provide a way to infer the "missing" modality in a single modality dataset by analysing

the connection between the multi modalities, which has the potential to largely reduce the multimodal sequencing costs in future. This project aims to design a machine learning method to predict the missing modality based on multimodal analysis.

- **Title**: Evaluate label transfer methods in single-cell omics data analysis.

  Supervisor: A/Prof Pengyi Yang (School of Mathematics and Statistics)

  Project: Multimodal single-cell technologies, which simultaneously profile multiple data types in the same cell, represent a new frontier for the discovery and characterization of cell states. In general, unsupervised or supervised training can be very challenging without pairing information across different modalities, with finding a common embedding manifold becoming more difficult as the complexity of the data increases. Recently, various multi-modal label transfer algorithms have been proposed. However, the lack of benchmark studies complicates the choice of the methods. This project aims to benchmark these methods and identify the strengths and weaknesses of each method on multiple criteria, which could give guidance for people to use.

- **Title**: Mathematics of deep structured neural networks.

  Supervisor: Prof Dingxuan Zhou (School of Mathematics and Statistics)

  Project: This project aims at mathematics of deep learning with some structured deep neural networks including deep convolutional neural networks and recurrent neural networks. The objectives include mathematical and statistical analysis of the deep learning algorithms induced by such neural networks and numerical simulations in dealing with some practical data.

- **Title**: Deep learning for text readability.

  Supervisor: Prof Dingxuan Zhou (School of Mathematics and Statistics)

  Project: Deep learning has achieved great success in speech recognition, computer vision, natural language processing, and some other practical domains. This project aims at applying deep learning to text readability and some related mathematical questions for understanding deep learning.

8 Assessment

8.1 The honours grade

The student's honours grade is based on the average mark achieved by each student, over the 4 courses and the project. Courses account for 50% of the assessment and the project for the remaining 50%.

According to the Faculty of Science guidelines, the grade of Honours to be awarded is determined by the honours mark as follows:

Grade of Honours	Faculty-Scale
First Class, with Medal	95–100
First Class (possibly with Medal)	90–94
First Class	80–89
Second Class, First Division	75–79
Second Class, Second Division	70–74
Third Class	65–69
Fail	0–64

The Faculty has also given the following detailed guidelines for assessing of student performance in Honours.

95–100 Outstanding First Class quality of clear Medal standard, demonstrating independent thought throughout, a flair for the subject, comprehensive knowledge of the subject area and a level of achievement similar to that expected by first rate academic journals. This mark reflects an exceptional achievement with a high degree of initiative and self-reliance, considerable student input into the direction of the study, and critical evaluation of the established work in the area.

90-94 Very high standard of work similar to above but overall performance is borderline for award of a Medal. Lower level of performance in certain categories or areas of study above.

Note that in order to qualify for the award of a university medal, it is necessary but not sufficient for a candidate to achieve a SCIWAM of 80 or greater and an honours mark of 90 or greater. Faculty has agreed that more than one medal may be awarded in the subject of an honours course.

The relevant Senate Resolution reads: "A candidate with an outstanding performance in the subject of an honours course shall, if deemed of sufficient merit by the Faculty, receive a bronze medal."

80-89 Clear First Class quality, showing a command of the field both broad and deep, with the presentation of some novel insights. Student will have shown a solid foundation of conceptual thought and a breadth of factual knowledge of the discipline, clear familiarity with and ability to use central methodology and experimental practices of the discipline, and clear evidence of some independence of thought in the subject area.

Some student input into the direction of the study or development of techniques, and critical discussion of the outcomes.

75-79 Second class Honours, first division – student will have shown a command of the theory and practice of the discipline. They will have demonstrated their ability to conduct work at an independent level and complete tasks in a timely manner, and have an adequate understanding of the background factual basis of the subject. Student shows some initiative but is more reliant on other people for ideas and techniques and project is dependent on supervisor's suggestions. Student is dedicated to work and capable of undertaking a higher degree.

70-74 Second class Honours, second division – student is proficient in the theory and practice of their discipline but has not developed complete independence of thought, practical mastery or clarity of presentation. Student shows adequate but limited understanding of the topic and has largely followed the direction of the supervisor.

65-69 Third class Honours – performance indicates that the student has successfully completed the work, but at a standard barely meeting Honours criteria. The student's understanding of the topic is extremely limited and they have shown little or no independence of thought or performance.

0-64 The student's performance in fourth year is not such as to justify the award of Honours.

8.2 The coursework mark

Students are required to attend 4 courses of 6CP during the academic year and the coursework mark is a simple average of the courses they took.

Student performance in each honours course is assessed by a combination of assignments and examinations. The assignment component is determined by the lecturer of each course and the examination component makes up the balance to 100%.

8.3 The project mark

The project's mark is split 90% for the essay and 10% for the student's presentation. The presentation mark is determined by the stats staff attending the presentation.

The essay is assessed by three members of staff (including the supervisor). The overall final mark for the essay is a weighted average of all three marks awarded. A weighting of 50% is attached to the supervisor's original mark, while a weight of 25% is attached to each of the two marks awarded by the other examiners.

The criteria which the essay marks are awarded by each examiner include:

- quality of synthesis of material in view of difficulty and scope of topic, and originality, if any;
- evidence of understanding;
- clarity, style and presentation;
- mathematical and/or modelling expertise and/or computing skills.

The student's supervisor will also consider the following criteria:

- Has the student shown initiative and hard work which are not superficially evident from the written report?

- Has the student coped well with a topic which is too broad or not clearly defined?

8.4 Procedures

All assessable student work (such as assignments and projects) should be completed and submitted by the advertised date. If this is not possible, approval for an extension should be sought in advance from the lecturer concerned or (in the case of honours projects) from the Honours Coordinator. Unless there are compelling circumstances, and approval for an extension has been obtained in advance, late submissions will attract penalties as determined by the Board of Examiners (taking into account any applications for special consideration).

Appeals against the assessment of any component of the course, or against the class of Honours awarded, should be directed to the Head of School.

Note: Students who have worked on their projects as Vacation Scholars are required to make a declaration to that effect in the Preface of their theses.

9 Seminars

Mathematical Statistics and Data Science seminars are usually held every week on Friday afternoons. These seminars are an important forum for communicating ideas, developing critical skills and interacting with your peers and senior colleagues. Seminars are usually given by staff members and invited speakers. All Honours students are encouraged to attend these seminars. Keep in mind that attending these seminars might help develop your presentation skills.

10 Entitlements

Honours students enjoy a number of privileges, which should be regarded as a tradition rather than an absolute right. These include:

- Office space and a desk in the Carslaw building.
- A computer account with access to e-mail and the internet, as well as LaTeX and laser printing facilities for the preparation of projects.
- Photocopy machine for any of your work related material.
- After-hours access to the Carslaw building.
- A pigeon-hole in room 728 — please inspect it regularly as lecturers often use it to hand out relevant material.
- Participation in the School's social events.
- Class representative at School meetings.

11 Scholarships, Prizes and Awards

University of Sydney Honours Scholarships
These $6,000 Honours Scholarships are awarded annually on the basis of academic merit and personal attributes such as leadership and creativity.

The following prizes may be awarded to statistics Honours students of sufficient merit. Students do not need to apply for these prizes, which are awarded automatically. The complete list is available here.

The Joye Prize
Awarded annually to the most outstanding student completing fourth year Honours in Applied Mathematics, Pure Mathematics or Mathematical Statistics (provided the work is of sufficient merit).

George Allen Scholarship
This is awarded to a student proceeding to Honours in Mathematical Statistics who has shown proficiency in all Senior units of study in Mathematical Statistics.

University Medal
Awarded to Honours students who perform outstandingly. The award is subject to Faculty rules, which require a mark of at least 90 in Mathematical Statistics 4 and a SCIWAM of 80 or higher. More than one medal may be awarded in any year.

Ashby Prize
Offered annually for the best essay, submitted by a student in the Faculty of Science, that forms part of the requirements of Honours in Pure Mathematics, Applied Mathematics or Mathematical Statistics.

Barker Prize
Awarded at the fourth (Honours) year examination for proficiency in Pure Mathematics, Applied Mathematics or Mathematical Statistics.

Norbert Quirk Prize No IV
Awarded annually for the best best entry to the SUMS Competition by an Honours student.

Veronica Thomas Prize
Awarded annually for the best honours presentation in statistics.

Australian Federation of University Women (NSW) Prize in Mathematics
Awarded annually, on the recommendation of the Head of the School of Mathematics and Statistics, to the most distinguished woman candidate for the degree of BA or BSc who graduates with first class Honours in Applied Mathematics, Pure Mathematics or Mathematical Statistics.

12 Life after Fourth Year

Students seeking assistance with post-grad opportunities and job applications should feel free to ask lecturers most familiar with their work for advice and written references. The Head of Statistics Programme, the Program Coordinator and the course lecturers may also provide advice and personal references for interested students.

Students thinking of enrolling for a higher degree (MSc or PhD) should direct all enquiries to the Director of Postgraduate Studies:

`pg-director@maths.usyd.edu.au`

Students are also strongly encouraged to discuss potential research topics with individual staff members.

Students who do well in their Honours studies may be eligible for postgraduate scholarships, which provide financial support during subsequent study for higher degrees.

Last but not least, there is a number of jobs for people with good statistical knowledge. Have a look here.