Data Science | Lancaster University

MSc in Data Science

Course Handbook 2019-20

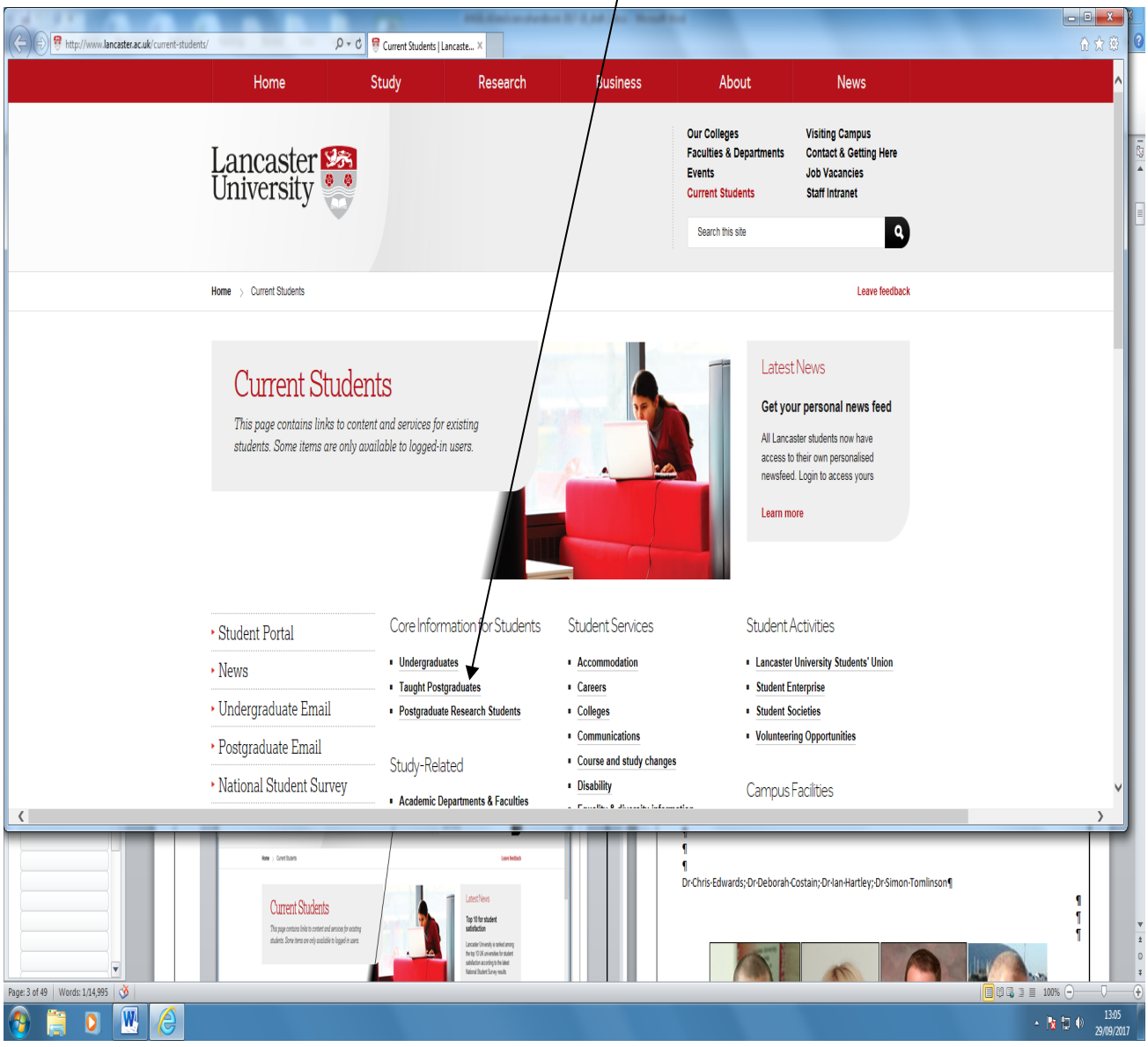
http://www.lancaster.ac.uk/dsi/education/

Section	Contents
Preface	University Core Information for Students
1	Welcome to Data Science at Lancaster
2	Overview of Data Science at Lancaster
3	Data Science: Computing Specialism
4	Data Science: Statistical Inference Specialism
5	Designing your Programme of Study
6	Mode of Attendance
7	Submission of Coursework and Feedback
8	Student Feedback Mechanisms
9	Programme Rules and Requirements for Awards
10	Student Support, Advice and Facilities
Appendix A	Module Descriptions
Appendix B	Term Dates

Preface: Core Information for Lancaster University PG Students

This course handbook has been compiled to assist MSc Data Science students studying at Lancaster University. More generally, all `core' university-level postgraduate information is available from the web page: http://www.lancaster.ac.uk/current-students/ and can be accessed from 'Core Information for Students' categorised under the respective student groupings:

Taught Postgraduate Core Information



1 Welcome to Data Science at Lancaster

Welcome on board the MSc in Data Science!

This is the fifth year that we have run programmes in the expanding field of Data Science. The programmes have been developed and crafted to give students a challenging, yet rewarding, experience that prepares them for the growing job market surrounding Data Science.

In formulating the degree schemes we have drawn on our world-leading research and teaching expertise in Statistics, Computer Science, Data Mining, Business Analytics, Health Informatics and Environmental Science to provide you with a cutting-edge curriculum. This will include the use of state-of-the-art methodologies and big data technologies together with data from a variety of domains (e.g. social media, health, ecology, business etc.). When designing the curriculum, we also sought guidance from businesses and research organisations (so you will be exposed to real-world data science problems) several of whom will provide students placements for masters students.

Ultimately, the degree is about your experience, which we hope will be challenging, yet fulfilling. To enhance this, we have provided virtual learning environments through which you can network with your peers and ask questions of your tutors and lecturers. Should you have any issues related to your studies then the course directors and support staff will be on hand to help you.

This handbook provides an overview of the programmes structure and the modules offered. We are incredibly excited about this coming year, the curriculum, the available pathways and the opportunities that your studies will provide.

Dr Chris Edwards; Dr Emma Eastoe; Dr Simon Tomlinson, Dr Clement Lee, Dr Leandro Soriano Marcolino

Prof. Chris Edwards
c.edwards@lancaster.ac.uk
+44 (0)1524 510329

Dr. Emma Eastoe
e.eastoe@lancaster.ac.uk
+44 (0)1524 593954

Dr. Simon Tomlinson
s.tomlinson2@lancaster.ac.uk
+44 (0)1524 510537

Dr. Clement Lee
clement.lee@lancaster.ac.uk
+44 1524 595190

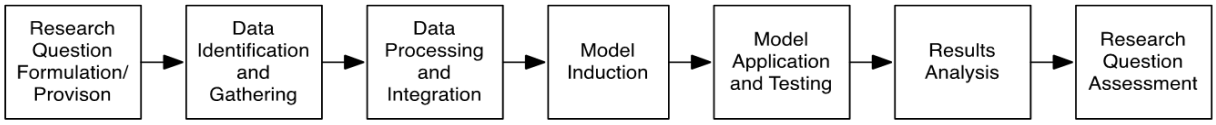
Dr. Lenadro Soriano Marcolino
l.marcolino@lancaster.ac.uk
+44 1524 510369

2 Overview of Data Science at Lancaster

2.1 Data Science at Lancaster

Data are now being generated at an unprecedented rate and scale. Whether derived from environmental sensors, social media streams, businesses, or public sector organisations, there is a need to process these data sets and extract and harness information, to allow business and research questions to be answered. This is the job of the data scientist whose role involves research question formation, gathering and synthesis of data, data mining and visualisation, statistical -modelling,- inference and -validation, forecasting and prediction and dissemination of findings.

A data scientist is thus a highly skilled individual with the ability to: articulate a research question; gather, process and model data at large-scale (developing scalable algorithms for performing inference and modelling complex and heterogeneous data structures) and to disseminate research findings in context. This process is sequential, often cyclic in nature, and flows to form the data science pipeline:



The Lancaster programme combines interdisciplinary teaching from world-leading departments: the School of Computing and Communications (SCC); the Department of Mathematics and Statistics and the Lancaster Environment Centre (LEC); Lancaster University Management School (LUMS) and the CHICAS group in the Faculty of Health and Medicine (FHM).

Building upon a common 'core' set of modules the Data Science Masters schemes allow for subject specialism:

- **MSc: Data Science**
  - Statistical Inference Specialism
  - Computing Specialism.

The *Computing specialism* of the Data Science MSc is aimed at students with a background in computer science who want to develop strong quantitative, data handling and analytical skills, along with skills that support the engineering of systems to support data science. The *Statistical Inference* specialism of the Data Science MSc is aimed at students with a background in mathematics and statistics and who want to develop their statistical, computing and analytical skills for the extraction, synthesis, processing and analysis of large and complex data. The choice of elective modules cover advanced statistical modelling topics and provide the opportunity for students to follow designated; **Societal, Population Health**, **Environmental**, **Bioinformatics** and **Business Intelligence** pathways. Course details, by specialism, are provided in Section 3 and Section 4. All module descriptions are provided in Appendix A.

Note that, subject to module pre-requisites, students may choose their elective modules across the two specialisms (Computing and Statistical Inference) in order to fit their background and career aspirations. See Section 5: `Designing your Programme of Study'.

"This new initiative represents a significant investment in post-graduate teaching at Lancaster, positioning us among the world leaders in the education of highly qualified people with the relevant skills necessary to make a real impact in the world, in fields from big business and engineering to big data and environmental science".

Professor Mark Smith, Vice-Chancellor, 2014

2.2 Admissions Criteria

We expect to recruit students who are interested in a career at the interface of computing, statistics and their application.

For the MSc in Data Science you should hold, or expect to obtain, a BSc in Mathematics and, or, Statistics or a BSc in Computing. Admission to the Master's programme requires a minimum of a 2:1.

Students wishing to follow designated Societal, Population Health, Bioinformatics, Environmental or Business Intelligence pathways should hold a first degree in a highly relevant subject area (minimum 2:i) and have computing skills and mathematics and statistics knowledge to A' level standard as minimum. Entry to the programme will be subject to interview to ensure programme entry requirements are met.

For applicants whose first language is not English a recognised English qualification is required:

- IELTS: 6.5 (with at least 6.0 in each skill).

Overseas students will require a visa to be able to study with us in the UK.

2.3 Making an Application

You should apply online using the My Applications website

Supporting documentation includes:

- Your degree transcripts and certificates, including certified English translations if applicable
- Two references

If English is not your first language, you should also enclose copies of your English language test results

You also need to complete a personal statement to help us understand why you wish to study your chosen degree. You should also state your intended specialism/focus.

If you are a current Lancaster University student, you will not need to provide your Lancaster degree transcript and only need to provide one reference.

The postgraduate section of Lancaster University's website has plenty of Information about applying to Lancaster

3 MSc Data Science: Computing Specialism

Entry Requirements: 2:1 degree in a subject relevant to computer science.

Key Contact: Prof Chris Edwards, School of Computing and Communications: c.edwards@lancaster.ac.uk

3.1 Computing specialism: overview

Underpinning the data scientist role are the technologies that enable the processing of data at large-scale, often using parallel processing paradigms.

The Computing specialism provides the training to understand how these technologies function and how they are implemented within both enterprise and research environments. Students will get hands-on experience of building, from scratch, large-scale systems that enable data science questions to be answered, using technologies.

The Computing specialism consists of a series of **taught modules** (120 credits) followed by the completion a **dissertation** (60 credits). The taught course component consists of 8 modules and can be decomposed as follows:

A '**core'** set of five compulsory modules **(75 credits)** spanning modern data science fundamentals:

- SCC460 Data Science Fundamentals (Michaelmas term & Lent term)
- SCC403 Data Mining (Michaelmas term)
- SCC461 Programming for Data Scientists (Michaelmas term)
- CFAS440 Statistical Methods and Modelling (Michaelmas term)
- CFAS406 Statistical Inference (Michaelmas term)

A set of three '**compulsory'** specialism specific modules **(45 credits):**

- SCC462 Distributed Artificial Intelligence (Lent)
- SCC413 Applied Data Mining (Lent)
- SCC411 Building Big Data Systems (Lent)

All module descriptions are, in alpha-numeric order, contained in Appendix A. Modes of assessment and credit weightings for all modules follow in Table 1, Section 3.2.

The dissertation component consists of a substantial project applying data scientific methods to address a substantive research question. This will be undertaken during the summer term and will typically incorporate a ten-week industrial or research organisation placement.

Dissertation Assessment

The dissertation component carries a weighting of 60 credits at level 7. Each dissertation will be double-marked, and a provisional mark agreed between the two markers. **You will be asked to attend a viva in person or via Skype to defend your work. You will also be required to attend the poster conference, presenting your work, on September 11th 2020.** A copy of each dissertation and a brief report (including a provisional mark) agreed between the two internal markers will then be sent to the external examiner in advance of the final meeting of the Board of Examiners in October. The deadline for submission of the dissertation is Friday 4th September 020.

Note that the above module structure is the designated Computing Specialism. However, subject to module pre-requisites, students can select their non-core modules (45 credits) across the specialisms of the degree and tailor their programme of study to reflect expertise and aspirations

3.2 Computing Specialism: Course Structure; Weightings and Assessment Strategy

The taught courses run during weeks 1 to 20. Core modules are taught in weeks 1 to 10 (Michaelmas term). Compulsory, specialism specific, modules are taught in weeks 11 to 20 (Lent term). Term dates for 2019/20 are given in Appendix B.

Core Module Assessment

All Computing specialism modules are assessed via 100% coursework (Section 3.2, Appendix A). Marks are ratified by the Board of Examiners which meets in October. Credit for a module is given if the overall module mark is 50% or more.

Dissertation Assessment

The dissertation component carries a weighting of 60 credits. Each dissertation will be double-marked, and a provisional mark agreed between the two markers. A copy of each dissertation and a brief report (including a provisional mark) agreed between the two internal markers will then be sent to the External Examiner in advance of the final meeting of the Board of Examiners in October. The deadline for submission of the dissertation is 4th September.

Table 1. Course Structure, module weightings and mode of assessment

1. A set of five 'core' modules (75 credits) in common with Statistical Inference specialism except that the statistics modules studied are tailored according to background: hence core in principle.

Module number	Title	Weeks scheduled	Coursework %	Exam %	Credit Weighting
SCC460	Data Science Fundamentals	1 – 10	100	NA	15
SCC403	Data Mining	1 – 10	100	NA	15
SCC461	Programming for Data Scientists	1 – 10	100	NA	15
CFAS440	Statistical Methods and Modelling	1 – 5	100	NA	15
CFAS406	Statistical Inference	6 – 10	100	NA	15

2. A set of three 'compulsory' (45 credits) computing specialism specific modules:

Module number	Title	Weeks scheduled	Coursework %	Exam %	Credit Weighting
SCC.462	Distributed Artificial Intelligence	11 - 13	100	NA	15
SCC.413	Applied Data Mining	14 - 20	100	NA	15
SCC.411	Building Big Data Systems	14 - 20	100	NA	15

SCC modules: taught by School of Computing and Communications except for SCC461 which is cross-disciplinary; CFAS modules: taught by the Department of Mathematics and Statistics

3. A Dissertation (60 credits) with an associated industrial placement/research organisation.

4 MSc Data Science: Statistical Inference Specialism

Entry Requirements: 2:1 in Mathematics and/or Statistics.

Key Contact: Dr Emma Eastoe, Mathematics and Statistics: e.eastoe@lancaster.ac.uk

4.1 Statistical Inference Specialism Overview

The Statistical Inference specialism is aimed at students with a background in mathematics and statistics and who want to develop their statistical, computing and analytical skills for the extraction, synthesis, processing and analysis of large and complex data.

The programme encompasses data science fundamentals but with an additional focus upon statistical modelling and inference of large and complex data structures and the underpinning theories. More specifically, the course provides a thorough 'core' training in statistical theory; data analysis and computing via a distinctive blend of leading-edge methodology and practical techniques (including Bayesian and computational methods and data mining) and 'optional' modules spanning, for example, genomics, longitudinal data analyses, time-to-event data, extremal analysis, spatial data analyses and forecasting. The modules reflect inter-departmental research expertise and prepare students for particular career options in areas with a growing demand for data scientists. The elective modules can be self-selected to form designated **Societal**, **Population Health, Environmental**, Bioinformatics and **Business Intelligence** pathways (Section 5), or more generally according to interests and aspirations. An overview of the structure is below.

The Statistical Inference specialism of the MSc in Data Science consists of a series of **taught modules** (120 credits) followed by the completion a **dissertation** (60 credits). The taught course component consists of nine modules decomposed as follows:

A '**core'** set of five compulsory modules **(75 credits)** spanning modern data science fundamentals:

- SCC.460 Data Science Fundamentals (Michaelmas & Lent term)
- SCC.403 Data Mining (Michaelmas term)
- SCC.461 Programming for Data Scientists (Michaelmas term)
- MATH552 Generalised Linear Modelling (Michaelmas term)
- MATH551 Likelihood Inference (Michaelmas term)

A `**compulsory specialist'** module **(15 credits)** in Bayesian Inference for Data Science (MATH555) or Statistical Learning (CFAS420);

A set of three `**optional**' modules **(30 credits)** chosen from a number of electives which span a range of specialist/advanced statistical methods relevant to the study design, analysis and interpretation of observational and experimental data:

MATH562	Extreme Value Theory	CHIC571	Infectious Disease Modelling
MATH563	Design and Analysis of Clinical Trials	CHIC581	Statistical Genetics and Genomics
MATH564	Principles of Epidemiology	BIOL445	Bioinformatics
MATH566	Longitudinal Data Analysis	LEC402	Geoinformatics
MATH573	Survival and Event History Analysis	LEC468	Modelling Environmental Processes
CFAS414	Methods for Missing Data	MSCI523	Forecasting
CFAS415	Structural Equation Modelling	MSCI526	Introduction to Intelligent Data Analysis
CHIC565	Environmental Epidemiology	MSCI534	Optimisation & Heuristics

All module descriptions are, in alpha-numeric order, contained in Appendix A. Modes of assessment and credit weightings follow in Section 4.2, Table 2.

A **dissertation** component consisting of a substantial project (60 credits) applying data scientific methods to address a substantive research question. This will be undertaken during the summer term and will typically incorporate a ten-week industrial or research organisation placement.

4.2 Statistical Inference: Course Structure; Weightings and Assessment Strategy

Course Scheduling and Assessment and Resit Examinations

The taught courses run during weeks 1 to 20. Core modules are taught in weeks 1 to 10. Optional and specialism specific modules are taught in weeks 11 to 20 (Lent term). **Exams are held in May/June 2020.** Term dates for 2019/20 are provided in Appendix B. All r**esit examinations are held in the last two weeks of September each year**.

Assessment

Testing of knowledge and understanding is achieved through a range of assessment methods. The credit weighting and balance between coursework and examination varies between modules, as shown in Section 4.3. Marks are ratified by the Board of Examiners which meets in June and October. Credit for a module is given if the overall module mark is 50% or more.

Exam Timetable

Examinations for modules studied in academic year 19/20 will be timetabled during the period May 2020 and June 2020. Provisional exam results will be made available to students in late June 2019. Final results will be communicated by post by late November 2020.

Dissertation Assessment

The dissertation component carries a weighting of 60 credits at level 7. Each dissertation will be double-marked, and a provisional mark agreed between the two markers.

You will be required to attend a viva in person or via Skype to defend your work and to present a poster at the poster session (mini conference) in September 2020.

A copy of each dissertation and a brief report (including a provisional mark) agreed between the two internal markers will then be sent to the external examiner in advance of the final meeting of the Board of Examiners in October. The deadline for submission of the dissertation is Friday 4th September 2020.

Table 2. Course Structure, module weightings and mode of assessment

2.1 **A set of five 'core' modules (75 credits)**

Module	Title	Weeks Scheduled	Coursework %	Exam %	Credit Weight
SCC.460	Data Science Fundamentals	1 – 10	100	NA	15
SCC.403	Data Mining	1 – 10	100	NA	15
SCC.461	Programming for Data Scientists	1 – 10	100	NA	15
MATH551	Likelihood Inference	1 – 5	30	70	15
MATH552	Generalised Linear Models	1 – 5	50	50	15

2.2 **A 'compulsory' fundamentals module in Bayesian Inference module (15 credits) OR statistical earning (15 credits).**

Module	Title	Weeks Scheduled	Coursework %	Exam %	Credit Weight
MATH555	Bayesian inference for Data Science	11 – 20	50	50	15
CFAS420	Statistical Learning	16 – 19	100	NA	15

2.3 **Three 'self-selected' optional modules (30 credits) from:**

Module	Title	Weeks Scheduled	Coursework %	Exam %	Credit Weight
MATH562	Extreme Value Theory	11 – 12	50	50	10
MATH563	Clinical Trials	11 – 12	50	50	10
MATH564	Principles of Epidemiology	13 – 14	50	50	10
MATH566	Longitudinal Data Analysis	15 – 16	50	50	10
MATH573	Survival and Event History Analysis	17 – 18	50	50	10
CHIC565	Environmental Epidemiology	19 – 20	50	50	10
CHIC571	Infectious Disease Modelling	14 – 15	100	NA	10
CHIC581	Statistical Genetics and Genomics	18 & 20	100	NA	10
BIOL445	Bioinformatics	17 – 18	50	50	10
MSCI523	Forecasting	11 – 20	100	NA	10
MSCI526	Introduction to Intelligent Data Analysis	11 – 20	100	NA	10
MSCI534	Optimisation & Heuristics	11 – 20	30	70	10
CFAS414	Methods for Missing Data	12 – 13	100	NA	10
CFAS411	Multilevel Modelling	14 – 15	100	NA	10
CFAS415	Structural Equation Modelling	20 – 20b	100	NA	10
LEC402	Geoinformatics	11 – 20	100	NA	15
LEC468	Modelling Environmental Processes	11 – 15	50	50	15

This is an interdisciplinary programme. SCC modules taught by School of Computing and Communications; MATH/CFAS modules: taught by the Department of Mathematics and Statistics; MSCI modules taught by Lancaster University Management School: CHIC taught by CHICAS, FHM. *NOTE THE SCHEDULE MAY CHANGE*

3. A Dissertation (60 credits) with an associated industrial / research organisation placement.

5 Designing your Programme of Study

The optional modules in Table 2: (Section 2.3) are eclectic (reflecting university research expertise) and are in areas in which there is a high demand for data scientists. Students are required to choose modules to the value of 30 credits; building upon the `core' modules (75 credits) and the additional foundations module chosen from Applied Data Mining, Bayesian inference for Data Science or Statistical Learning (15 credits).

Elective modules may then be chosen to follow designated **Business Intelligence, Societal, Population Health, Bioinformatics and Environmental pathways**. The computing specialism also forms a designated route, with a systems and technologies focus. Table 3 provides details of the modules forming each of the five, above named, Statistical Inference Specialism pathways and also the Computing specialism (path 6 below)

Note that subject to module pre-requisites and timetabling, modules can be selected across specialisms and pathways to form a bespoke programme of study, according to interests and aspirations.

Table 3: Designated Pathways: students choose modules to the value of 30 credits.

1: Business intelligence Forecasting (10 credits) Optimisation & Heuristics (10 credits) Introduction to Intelligent Data Analysis (10 credits)	**2: Societal** Multilevel Modelling (10 credits) Methods for Missing Data (10 credits) Structural Equation Modelling (10 credits)
3: Population Health* Principles of Epidemiology (10 credits) Longitudinal Data Analysis (10 credits) Infectious Disease Modelling (10 credits) Survival and Event History Analysis (10 credits) Environmental Epidemiology (10 credits)	**4: Bioinformatics*** Clinical Trials (10 credits) Principles of Epidemiology (10 credits) Bioinformatics (10 credits) Statistical Genetics and Genomics (10 credits)
5: Environment* Geoinformatics (15 credits) Modelling Environmental Processes (15 credits)	**6: Computing** Distributed Artificial Intelligence (15 credits) Building Big Data Systems (15 credits)

* to follow this designated pathway: choose modules totalling 30 credits from those available

Note that Extreme Value, Longitudinal Data Analysis, Survival and Event History Analysis and Environmental Epidemiology require MATH551 and MATH552. This subset form a set of study modules for those students wanting an increased mathematical focus.

6 Mode of Attendance

6.1 Full-time: One Year

The academic year runs from October to September and consists of Michaelmas, Lent and summer terms**.** Lectures take place during the Michaelmas (October to December: weeks 1 - 10) and Lent (January to March: weeks 11 – 20) University terms.

Exams are held each year in May and June, after which the dissertation is prepared (June to September). Details of the course structure, assessment arrangements and module descriptions are provided in Sections 3, 4, 5 and Appendix A. Appendix B provides term dates.

Modules delivered during Michaelmas term (weeks 1 to 10) are typically presented over a five or ten week period. Modules delivered in Lent term run variously, with many running intensively over two weeks comprising of an intensive period of lectures and labs (typically 20 hours over four days) followed by a period for a module specific project.

Full-time students are expected to be available for attendance at the University throughout the year**.**

6.2 Part-time: Two Years/Three Years

Students can study part-time over two or three years. The two-year arrangement is designed for students who are able to attend classes at the University on a regular basis. Such students will typically follow half of the taught modules that are available during each of the Michaelmas (weeks 1-10) and Lent (weeks 11-21) terms during each of their two years. Alternatively, students may study full time in term 1 in year 1 and full time in term II in year 2. The three-year arrangement is to allow for increased flexibility. Subject to timetabling and module pre-requisites, a suitable split of the modules will be designed for individual students, recognising their other time constraints and previous knowledge. *Such individualised study plans must be discussed with and approved by the Course Directors prior to commencement of the degree programme.*

In both cases, assessment is undertaken during the respective year of study alongside full-time students studying the module. Where appropriate, and only if necessary, the Course Directors may approve alternative submission deadlines for part-time students.

Part-time students will typically undertake the dissertation in the final year of study*, regardless of whether they take the two- or three-year option*. Subject to skills and knowledge students may begin their dissertation earlier subject to approval with the Course Directors.

6.3 Attendance for Examinations, Dissertation Submission Dates, Assessment and Awards

Students have to be present at the University for all written examinations that take place in May/June each year. Part-time students undertake examinations for modules studied during the same examination period as full-time students. They also submit the dissertation by the same deadline stipulated for full-time students in their second year of study. The scheme of credits, assessment and awards apply to both full- and part-time students.

All students are required to attend the university in September for the Dissertation Poster Session and to defend their work in a viva voce. The viva may be conducted remotely via Skype, for example.

Resit examinations take place in the last two weeks of September each year.

The Graduation Ceremony takes place Mid-December each year.

7 Submission of Coursework and Feedback

Modules are offered across various departments. Each module has its own procedure for handling coursework and the module tutor will provide information to you regarding the submission of coursework. One paper copy is submitted and also an electronic copy is uploaded online through Moodle (https://modules.lancs.ac.uk/). This allows us to check all coursework electronically for plagiarism. Please ensure that you check and are aware of the coursework deadlines and submission procedures. Coursework MUST NOT be handed directly to the module tutors since all coursework submissions need to be recorded by the administrating PG Office.

All hard-copies of coursework assignments are to be submitted to the **relevant** Masters Submission Box located in the administrating department for the module:

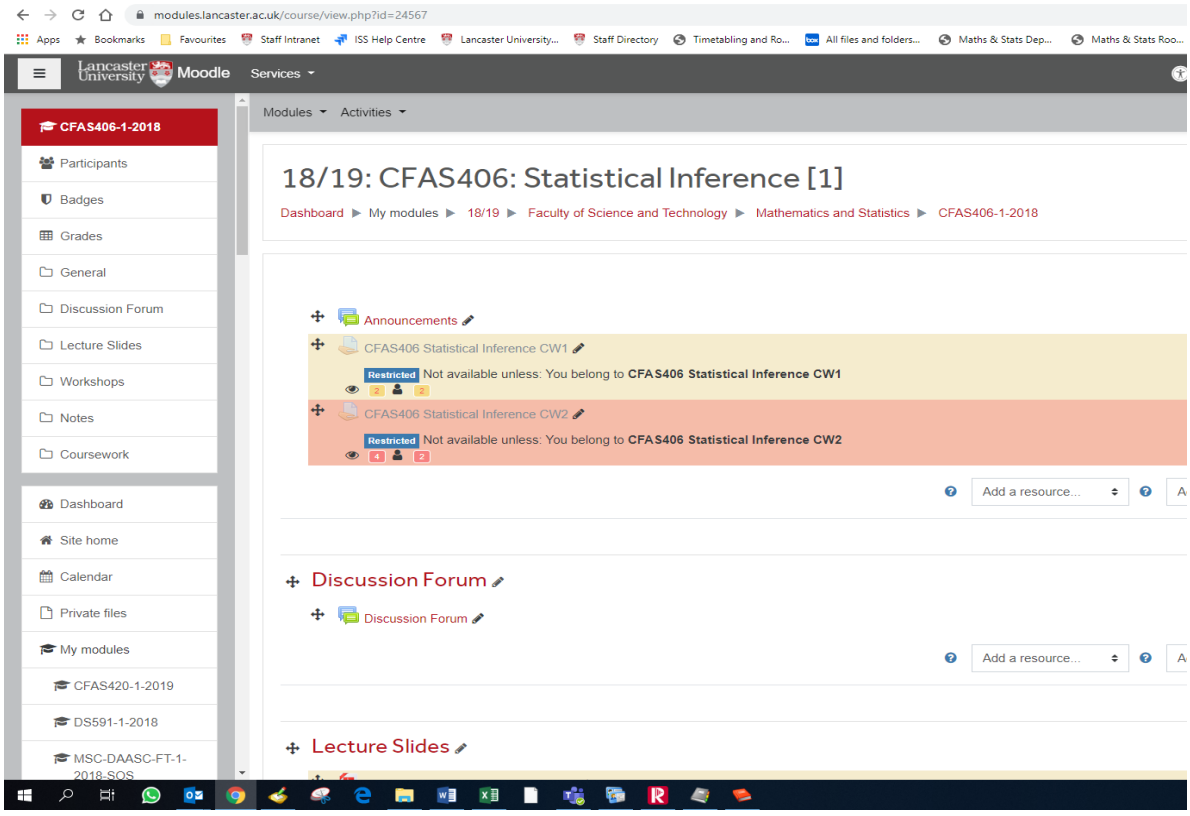
1. **SCCXXX modules**: **School of Computing and Communications -** Data Science submission box (contact Heather Hilton: h.hilton3@lancaster.ac.uk)
2. **MATHXXX/CFASXXX/CHIC565/581 modules: Mathematics and Statistics Dept. -** MSc Data Science submission box (contact: Roger Marsden: mathsteaching@lancaster.ac.uk)
3. **LECXXX: Lancaster Environment Centre -** PGT submission box (contact: Suzanne Stelling: lec-pgt@lancaster.ac.uk)
4. **MSCIXXX: Lancaster University Management School** (contact Jackie Clifton: j.clifton@lancaster.ac.uk)

Confirmation of the submission process for each module will be provided by the module lecturer.

Note: a plagiarism declaration form must signed and submitted at the start of each academic term to declare that the submitted work for the period is your own.

Submitting your assignments online

Each module you study will have its own Moodle page. You can view all of you Moodle pages by accessing https://modules.lancs.ac.uk/ or logging into your student portal https://portal.lancs.ac.uk/. Your assignments should be submitted online via Moodle. The example below illustrates this.



7.1 Late submission of assignments and penalties

There are strict university-wide rules in place for the late submission of assignments:

`Work submitted up to three days late without an agreed extension will receive a penalty of 10 percentage points (for example a mark of 62% would become 52%) and zero (non-submission) thereafter'.

Saturdays and Sundays are included as days in this regulation. However, when the third day falls on a Saturday or Sunday, students will have until 10.00 a.m. on Monday to hand in work without receiving further penalty.

Work submitted more than three days late without an agreed extension will be awarded a zero and considered a non-submission, and treated according to the standard procedures for failed work.

For regular small components of work when it is necessary to provide feedback quickly, the three-day period within which a graded penalty would be applied to late work may be shortened. Work handed in after the deadline but before [specified time and date when answers are to be released] will be subject to a percentage drop/reduction in letter grade in accordance with the standard regulations.

For pieces of work given a set number of marks, e.g. out of 30, the penalty for late submission will be department to complete in accordance with the General Assessment Regulations, section GR 2.3.7.

Where exceptional circumstances may have led a student to miss a stipulated deadline (which may be an already extended deadline), the student should make those circumstances known to the department. They should inform the department within 48 hours of the missed deadline unless prevented from doing so by acceptable circumstances in which case students should inform the department as soon as possible. Students should provide the department with evidence of the exceptional circumstances as soon as they are able. In this context, exceptional circumstances are defined by University regulations as actions or events outside the control of the student which result in any circumstances which are thought reasonably to have caused an individual student to fail to complete all the required assessment for a programme or contributing module by a stipulated deadline (e.g. missed exam or coursework deadline).

7.2 Plagiarism

The University has established an institutional framework for dealing with plagiarism: Plagiarism Framework and is a member of the JISC Plagiarism Detection Service (Turnitin), which searches for matching text between a paper and available material on the Internet. The administrating PG Office reserves the right to assess some or all of your work submitted electronically using this service.

You commit plagiarism if you try to pass off the work of any other person, whether a published author, an internet source or a fellow-student, as your own. To copy passages, source code, illustrations or even sentences from a book or on-line resource without acknowledgement is gross plagiarism; to copy from another student's essay or dissertation is a particularly grave offence. Never use another's actual words (or a close paraphrase of them) without putting them in quotation marks and giving an exact reference to your source.

Serious plagiarism may result in a zero mark, and if repeated - or if particularly severe - may result in your being excluded from further courses within the University (this matter would be decided by the Standing Academic Committee). These strict rules are in place to ensure that the work you undertake as part of your course has real value to you both educationally and in giving you a sense of personal achievement.

7.3 Feedback and Notification of Marks

Feedback on assessed work will typically be provided within four weeks of submission (excluding vacations and unforeseen staff absences). Once coursework has been marked and the marks recorded by the PG Office, students will be informed and can then collect coursework from the administrating PG Office.

Exam papers are marked in line with university timescales. It is University policy not to return examination scripts to students.

Students may also view their coursework marks via the 'Student Portal' once they have been processed by the administrating PG Office.

It should be remembered that until the External Exam Board has met, (October each year) any marks given to students are provisional and may be subject to change. The External Exam Board does not usually meet until around 6 weeks after the end of the programme.

As per the University regulations there is no appeal against academic judgement.

8 Student Feedback Mechanisms

8.1 Module Evaluation

You will be contacted by email at the end of each module and asked to complete a module evaluation form online. Evaluation is more than feedback of the good and bad elements of a module. It provides continual information for us to improve the modules we offer. Consequently it is very important that all students actively participate in the evaluation process. We ask that you respond to the request to submit your feedback as soon as possible. **Please note that your feedback is anonymous.**
Student Representatives

Student representatives are invited to sit on the departmental Staff-Student Consultative Committees. This is the official mechanism for communication between students, staff the Data Science team. The committees are formed in accordance with University Regulations, which require staff-student consultation prior to programme changes.

Two representatives will be required, one from each of the Data Science specialisms. Volunteers will be invited to express an interest at the beginning of the academic year.

The committees meet at least once per term as a consultative body to discuss current programmes and future proposed changes. Dates for the Committees and the minutes of each meeting will be made available on Moodle by the departmental Course Coordinators.

9 Programme Rules and Requirements for Awards

9.1 Module Rules:

The pass mark for taught Masters programmes is 50% per contributing module, with credit for a module being awarded when the overall mark for the module is 50% or greater. The mark for each module is given by a stipulated combination of written examination and/or coursework (Table 1, Table 2).

Awards:

The overall mark is the weighted average of the marks gained in the taught modules and the dissertation for the award of MSc in Data Science. The following criteria apply for the degree awards.

MSc in Data Science

To qualify for the Masters' degree in Data Science you must achieve 180 credits: 120 credits from the taught course component and 60 credits from the Dissertation. Credit for a module is given if the overall module mark is 50% or more.

Condonation:

Notwithstanding this requirement, candidates shall be eligible for an award by compensation/ condonation in respect of up to a maximum of 45 credits of a taught Masters programme provided that:

a) no single module mark falls below 40%;
b) the candidate's weighted mean is 50% or greater.

9.2 Higher awards:

MSc in Data Science with 'Merit'

In addition to the requirements for the MSc in Data Science, an 'MSc in Data Science with Merit' shall be awarded when the weighted average mark over the programme is 60% or more.

MSc in Data Science with 'Distinction'

In addition to the requirements for the award of the MSc in Data Science, an `MSc in Data Science with Distinction' is awarded when the weighted average mark over the programme is 70% or more.

Borderline Cases

Subject to the standard condonation rules above, where the overall weighted average falls within two percentage points of the Merit / Distinction degree award range (i.e. 58% / 68%, respectively) the following rules for degree awards will apply:

58.0% to 59.9%:

If more than 50% of module credits (i.e. > 90 credits) are at `Merit level' (i.e. are at 60% or above) the MSc in Data Science will be awarded with Merit.

68.0% to 69.9%:

If more than 50% of module credits (i.e. > 90 credits) are at Distinction level (i.e. at 70% or above) the MSc in Data Science will be awarded with Distinction.

Postgraduate Diploma in Data Science

To qualify for the Postgraduate Diploma in Data Science you must achieve a total of 120 from the taught courses. Credit for a module is given if the overall module mark is 50% or more.

Condonation

Notwithstanding this requirement, candidates shall be eligible for an award by compensation/ condonation in respect of up to a maximum of 30 credits provided that:

a) no single module mark falls below 40%;
b) the candidate's weighted mean is 50% or greater.

Higher awards:
Postgraduate Diploma in Data Science with `Merit'

In addition to the requirements for the Postgraduate Diploma in Data Science, a 'Postgraduate Diploma in Data Science with Merit' shall be awarded when the weighted average mark over the programme is 60% or more.

Postgraduate Diploma in Data Science with `Distinction'

In addition to the requirements for the award of the Postgraduate Diploma in Data Science, a `Postgraduate Diploma in Data Science with Distinction' shall be awarded when the weighted average mark over the programme is 70% or more.

Postgraduate Certificate in Data Science

To qualify for the Postgraduate Diploma in Data Science you must achieve a total of 60 from the taught courses. Credit for a module is given if the overall module mark is 50% or more.

Condonation

Notwithstanding this requirement, candidates shall be eligible for an award by compensation/ condonation in respect of up to a maximum of 20 credits provided that:

a) no single module mark falls below 40%;
b) the candidate's weighted mean is 50% or greater.

Higher awards:
Postgraduate Certificate in Data Science with `Merit'

In addition to the requirements for the Postgraduate Certificate in Data Science, a 'Postgraduate Certificate in Data Science with Merit' shall be awarded when the weighted average mark over the programme is 60% or more.
Postgraduate Certificate in Data Science with `Distinction'

In addition to the requirements for the award of the Postgraduate Diploma in Data Science, a `Postgraduate Certificate in Data Science with Distinction' shall be awarded when the weighted average mark over the programme is 70% or more.

Re-sits

A student who fails to achieve a mark of 50% for a module/element is entitled to one opportunity for reassessment in each failed module/element.

A mark of not more than 50% can be given for modules re-taken.

The form of the reassessment is at the absolute discretion of the Examination Board, save that the form of reassessment must allow the student a realistic chance of achieving 50% in the re-sit.

Resit examinations take place in the last two weeks of September each year. Students undertaking re-sits will need to be present at the university.

Notification of Final Degree Marks

The External Exam Board meets in late October to recommend awards. Final marks are released to students as soon as possible, thereafter. **Please note that the University Regulations state that written confirmation of results, provisional and final, may not be released to students who are in debt to the University.**

The Board of Examiners

The Board of Examiners consists of the:

- External Examiner for the MSc.
- Heads of Department;
- Course Directors;
- Course Tutor;
- Placement Officer;
- Examinations Officers;
- Postgraduate Coordinator.

The External Examiner for the period 2019-2020 is Professor Peter Triantafillou, Department of Computer Science, University of Warwick,

Graduation

The Postgraduate Graduation Ceremony will be in December 2020. Information regarding Graduation will be sent to you from the University Ceremonies Office.

Please note that it is essential that you keep your contact details address up-to-date in order to receive the relevant graduation mailings.

10. Student Support, Advice and Facilities

Who to go to for support & advice

If you are having problems with the course or have any queries please contact us as soon as possible. Support can come from a number of sources. Here's who to contact if you have problems with:

Course content: The Module Lecturer, your fellow students or the Course Tutor/Directors.

Assignment submission:

- **SCC modules**: contact Heather Hilton: h.hilton3@lancaster.ac.uk in the School of Computing and Communications teaching office
- **MATH/CFAS/CHIC565/581 modules:** contact Roger Marsden: mathsteaching@lancaster.ac.uk in the Department of Mathematics and Statistics teaching office
- **MSCI modules**: contact Jackie Clifton: j.clifton@lancaster.ac.uk in the Lancaster University Management School
- **CHIC571, CHIC581, BIOL445**: Nicola Caldwell: n.caldwell@lancaster.ac.uk in the Faculty of Health and Medicine
- **LEC modules**: contact Suzanne Stelling: lec-pgt@lancaster.ac.uk in the in the Lancaster Environment Centre

Some useful contact numbers are:

Name	Role	Telephone No. (01524)
SCC		
Chris Edwards	Programme Co-Director	510329
Simon Tomlinson	Business Engagement Manager	510537
Leandro Soriano Marcolino	Well-Being and Experience Officer	510369
Heather Hilton	SCC Module Coordinator	TBC
Maths & Statistics		
Emma Eastoe	Programme Co-Director	593954
Amy Pearson	Student Programmes Officer	593798
Roger Marsden	Teaching Coordinator	593067
General contacts		
Library	Library Enquiry desk	592516
LUSU	Lancaster University Students' Union	593765
Sports' Centre	Enquiry Desk	510600
The Base	Student Support	592525

10.1 Learning Development for the Faculty of Science and Technology

The faculty Learning Developers offer additional writing and study support to all students to help you to enhance your academic progress.

SCIENTIFIC WRITING & STUDY DEVELOPMENT

Dr Robert Blake & Dr Louise Innes (job share)

Robert and Louise are the Learning Developers for this Faculty. They can provide one-to-one consultations, for example, about scientific writing, academic reading, time management and preparation for exams. They also run scientific writing courses for international students and scientific writing workshops in departments.

Additional study consultations are available if you think you have a disability, such as dyslexia, that affects your academic work.

Study & Scientific Writing Consultations: Monday and Thursday afternoons, bookable through LibCal here.

Please note that we are unable to proofread completed dissertations, reports, etc. but offer diagnostic feedback on excerpts from your writing in progress.

Science Writing & Exam Revision Clinic January-May

FST 601 Scientific Writing for International students (optional module) weeks 2-7. Covers writing better reports, scientific essays and citation.

FST602 MSC International Students' Dissertation Writing Group, July-August- provide peer feedback (text organisation & grammar on your dissertation extracts).

Email: learningdevelopmentFST@lancaster.ac.uk

Moodle: https://modules.lancaster.ac.uk/course/view.php?id=282. For up to date information about consultation times and classes, please check on the Moodle.

Room: A33 Science & Technology Building

Availability: Louise Monday-Tuesday; Robert: Tuesday-Thursday

ENGLISH LANGUAGE

Dr Helen Hargreaves is the Learning Developer for English for Academic Purposes across the university. Her provision includes:

- Weekly writing workshops in the Library
- One-to-one English language advice for academic writing
- Culture Exchange (social events for PG students to develop English in an informal setting) Enrol on the English Language Development Moodle for up-to-date information and how to book.

Email Helen at englishlanguagedevelopment@lancaster.ac.uk with any questions

10.2 University systems, Resources and Support

Moodle

Each postgraduate programme is supported by its own online page which uses Moodle as its base. This system is used by students from departments all across the University so you are not alone. You should familiarise yourself with this as much as possible as it is a key resource for students on the course. To find the page for your course, go to the main Moodle site from your Student Portal. Go to https://portal.lancs.ac.uk/. You will need your University username and password to get access to the site. Moodle is used for:

- Gaining access to your Module pages,
- Viewing your timetable for the year and receiving notifications about changes to the timetable,
- Asking questions to tutors or fellow students,
- Getting access to additional materials and resources,
- Accessing and submitting your coursework,
- Checking course deadlines etc.,
- Giving us feedback about the course,
- Discussing issues about work

Email

All students will be given a Lancaster University email address, in the form yourname@lancaster.ac.uk. Please note that any contact we make with you will be through your Lancaster email address and therefore **you must access this e-mail account on a daily basis.** Failure to check your Lancaster account regularly does not constitute an excuse for missing important information, dates etc.

Computing Facilities

There are numerous open access PC labs located around campus. The PC labs provide a wide range of software, printers (colour and monochrome) and scanning facilities. All lab PCs are connected to the campus network and internet.

Information Systems Services (ISS) also provides other IT services to students, including IT workshops and courses.

It is also possible to access University services remotely e.g. from home, or via a smart phone.

The ISS Service Desk can be contacted if you require any general computing-related assistance.

Learning Zone

The Learning Zone is located centrally on Alexandra Square and is accessible 24-7. It provides relaxed surroundings for students to work within and bookable 'pods' for meetings, presentations and group work.

University Library facilities

Lancaster University Library is a valuable reference resource. Many of the main texts for the module on your programmes are available from here. Your registration with the library should have been completed when you registered with the University.

Using the Sports Facilities on Campus

As student of the University, you are entitled to student membership of the Sports Centre for the duration of your course. Please note however, that membership runs annually according to the academic year (October – September) so, depending on when you are attending your course, it may/may not be cost-effective to apply for membership. Membership details are available from the Sports Centre itself http://sportscentre.lancs.ac.uk/. Details of opening times / classes etc. can be found in the leaflet in your induction pack (or ring the Sports Centre itself on 01524 510600).

The Students' Union & Student Support Office

LUSU is a body that represents all student views to the University, providing professional, academic and other advice for students. Students registering at Lancaster automatically become members of the Students' Union. There are no financial obligations associated with membership, though you can withdraw from the union if you wish, by completing an opt-out form.

You can also apply online for your NUS "Purple Card" at http://card.lusu.co.uk/members and then collect this from LUSU in Bowland College, next to Alexandra Square.

Careers Advice

The University's Careers Service, offers an extensive service tailored to your needs. Their professional staff includes specialists in careers information, employer liaison, event management and careers guidance. They work closely with other staff within the university, the Students Union, professional bodies and a broad range of national and international employers to provide a variety of opportunities to help you progress your career goals. Careers are located in the Base, just off Alexander Square.

TARGETconnect is an online system administered by Careers and provides students with access to student and graduate vacancies, details of careers events, an appointment booking system to see a careers adviser and the online careers query system. Careers information including online psychometric testing and video resources are available online.

10.3 Student Based Services

Student Based Services provide information, advice and guidance covering different areas.

We hope you have an enjoyable and productive time at Lancaster, but we recognise that sometimes problems can affect your ability to study. Please do not forget that it is your responsibility to seek help if you are experiencing difficulties. The University will do whatever is possible to assist you, provided that we are aware of your problems. These may be personal, financial or academic. If you find yourself getting into difficulties we strongly urge you to consult the PG Administrator in the first instance.

In addition, Student Based Services provide information, advice and guidance covering different areas of student welfare.

'The Base' is situated on A-Floor of University House in Alexandra Square and is a one-stop enquiry desk for all Student Based Services. Staff there will be able to make appointments with specialist staff where needed. Details on the various student-based services can be found on the links below.

Student Registry – responsible for all regulations, policies and procedures governing your award. The Student Registry is also responsible for managing your official record, including personal details. They can provide information on many aspects of student administration.	**International Student Advisory Service** - advice on visa extension, rules on working in the UK, budgeting, general welfare and cultural orientation. They are the designated point of advice for immigration issues.
Disabilities Service – the University has been developing services for students in this area for over 20 years and aims to help all prospective and current students who have a disability.	**Assessment Centre** – Lancaster University has its own assessment Centre to identify study aids and strategies required to provide equal access to the curriculum.
Counselling and Mental Health Service - staff provide confidential and professional support on issues such as personal, family, social or academic matters over the short term, to more complex or difficult longer-term problems. The service offers both appointment and drop-in sessions.	**Student Funding & Financial Aid** - provide information, advice and guidance on student funding and financial aid. This includes student living cost loans/grants, tuition fee loans and living costs/budgeting.

Appendix A

Module Descriptions

Module Mnemonic:	BIOL445
Module Title:	Bioinformatics
Module Convenor:	Dr Derek Gatherer
Assessment:	Exam (50%) and coursework (50%)
Duration:	25 hours
Credits:	15
Term:	L1/L2 TBC
Prerequisites:	(MATH551 and MATH552) OR (CFAS406 and CFAS440)
Software used:	MEGA, DNASp, BEAST, FigTree, Tracer, SimPlot, SwissModel, Galaxy, Artemis, EMBOSS

This course will equip students with a working knowledge of the main themes in bioinformatics. On successful completion, students should be confident and competent in all aspects of bioinformatics that can be executed via the web or on software running on Windows/Mac systems. They will have an understanding of the theoretical algorithms that underpin the various software applications that they use, and be able to perform bioinformatics within their own biological sub-field. More generally, this module also aims to encourage students to access and evaluate information from a variety of sources and to communicate the principles in a way that is well-organised, topical and recognises the limits of current hypotheses. It also aims to equip students with practical techniques including data collection, analysis and interpretation.

Topics covered will include:

- Reading lists and how to manage reading. Doing a PubMed search;
- The foundations of Bioinformatics;
- Advanced bioinformatics I: Going deeper into algorithms;
- Advanced bioinformatics II: Structural bioinformatics;
- Advanced bioinformatics III: Phylogenetic : How do we use sequences to investigate evolution?
- Advanced bioinformatics IV: Detecting natural selection;
- Advanced bioinformatics V: Processing deep sequencing data.

On successful completion students will be able to:

- Perform bioinformatics via the web GenBank, Pfam, Uniprot, PDB, SCOP. Use Artemis for genome visualization.;
- Download and align sequences, curate sequences, derive statistics on alignments. Use DNASp for sliding window analysis;
- Building phylogenetic trees in MEGA. Use SimPlot for recombination analysis;
- Structural bioinformatics – do homology modelling via SwissModel. Use a protein sequence viewer. Use Galaxy for deep sequence assembly;
- Build a Bayesian phylogenetic tree with BEAST.

Bibliography:

- Michael Agostino. *Practical Bioinformatics*. Garland Science. ISBN 978-0-8153-4456-8
- Arthur M Lesk. *Introduction to Bioinformatics* 4th ed. Oxford Univ Press ISBN 978-0-19-965156-6
- Paul H Dear (ed) *Bioinformatics*. Scion ISBN 978-1-90-484216-3
- Masatoshi Nei & Sudhir Kumar *Molecular evolution and phylogenetics* (Available on http://lib.myilibrary.com/Open.aspx?id=83437)
- Drummond AJ & and Bouckaert RR. *Bayesian Evolutionary Analysis with BEAST*. Cambridge University Press ISBN 978-1107019652

Module Mnemonic:	CFAS406
Module Title:	Statistical Inference
Module Convenor:	Dr Clement Lee
Assessment:	Coursework (100%)
Duration:	25 hours
Credits:	15
Term:	M2
Prerequisites:	(MATH551 and MATH552) OR (CFAS406 and CFAS440)
Software used:	R

This modules aims to provide an in-depth understanding of statistics as a general approach to the problem of making valid inferences about relationships from observational and experimental studies. The emphasis will be on the principle of Maximum Likelihood as a unifying theory for estimating parameters. The module is delivered as a combination of lectures and practical's over four week.

Topics covered will include:

- Revision of probability theory and parametric statistical models.
- The properties of statistical hypothesis tests, statistical estimation and sampling distributions.
- Maximum Likelihood Estimation of model parameters.
- Asymptotic distributions of the maximum likelihood estimator and associated statistics for use in hypothesis testing.
- Application of likelihood inference to simple statistical analyses including linear regression and contingency tables.

Learning: Students will learn through the application of concepts and techniques covered in the module by application to real data sets. Students will be encouraged to examine issues of substantive interest in these studies. Students will acquire knowledge of:

- Application of likelihood inference to simple statistical analyses including linear regression
- The basic principles of probability theory.
- Maximum Likelihood as a theory for estimation and inference.
- The application of the methodology to hypothesis testing for model.

Students will, more generally, develop skills to:

- apply theoretical concepts
- identify and solve problems

Bibliography:

- Dobson, A. J. (1983). *An Introduction to Statistical Modelling.* Chapman and Hall.
- Eliason, S. R. (1993). *Maximum Likelihood Estimation: Logic and Practice.* Sage Publications.
- Pickles, A. (1984). *An introduction to likelihood analysis*. Geo Books.
- Pawitan, Y. (2001). *In all likelihood: statistical modelling and inference using likelihood*. Oxford University Press.

Module Mnemonic:	CFAS411
Module Title:	Multi Level Models
Module Convenor:	Dr Tom Palmer
Assessment:	Coursework (100%)
Duration:	20 hours
Credits:	10
Term:	L2
Pre-requisites:	(MATH551 and MATH552) OR (CFAS406 and CFAS440))
Software used:	R & MLwiN

The aim of this module is to introduce how to analyse data that has a multi-level, hierarchical structure. The mathematical form of multilevel models is described. The models are developed first for continuous outcomes moving from linear regression to the random intercept model to the random coefficient model. Multilevel models are then shown for binary and other outcomes. Software implementation is described with the lme4 package in R. Some use of MLwiN is also made.

Topics covered will include:

- The intra class correlation coefficient.
- Two level random intercept and random coefficient models with continuous outcomes.
- Checking model assumptions and residual diagnostics.
- Models with three or more levels.
- Generalized multilevel models including two-level logistic regression models, multilevel ordinal logistic regression models, and multilevel Poisson regression models.
- Worked examples are shown of fitting such models in statistical software (mainly in R, but also some in MLwiN).
- Students will also gain insight into that there are different estimation algorithms available for multilevel models.

On successful completion students will be able to:

- Comprehend the notation used to describe multilevel models
- Demonstrate knowledge of multilevel models by formulating appropriate models to answer specific questions
- Demonstrate and understand how to use statistical software to fit multilevel models and how to interpret the relevant output
- Demonstrate how to perform model diagnostics for such models
- Be able to interpret the results of fitting multilevel models.

Bibliography:

- Bryk, A. S., and Raudenbush, S. W., (1992) Hierarchical Linear Models, Sage.
- Goldstein, H., (2003) Multilevel Statistical Models. London, Edward Arnold
- Holmes Finch W, Bolin JE, Kelley K. Multilevel Modeling Using R. Chapman & Hall. 2014
- Hox, J., (2002) Multilevel Analysis: Techniques and Applications, Malwah, N.J: Lawrence Erlbourn Associates.
- Longford, N. T., (1993) Random Coefficient Models. Oxford University Press.
- Snijders, T. A. B., and Bosker, R. J., (1999) Multilevel Analysis. An Introduction to Basic and Advanced Multilevel Modelling. London: Sage.

Module Mnemonic:	CFAS414
Module Title:	Methods for Missing Data
Module Convenor:	Dr Robin Mitra
Assessment:	Coursework (100%)
Duration:	20 hours
Credits:	10
Term:	L2r
Prerequisites:	(MATH551 and MATH552) OR (CFAS406 and CFAS440)
Software used:	R Studio with packages: VIM, mice, amelia 2, miss MDA

This module deals with the ubiquitous and often neglected problem of dealing with missing data, common in many types of statistical analysis. We survey some ad-hoc strategies to deal with them and show how the can lead to bias and inefficiencies. We advocate using a principled approach and the formulating of the inherent missing data mechanism. We look at several principled methods of dealing with missing data. First we present a fully Bayesian approach using Winbugs. Secondly we create multiply imputed datasets using chained equation and then apply Rubin's rules for combing the analyses of the models. We then do the same thing as the previous method but use multivariate techniques rather than chained equations as the method of multiple imputation. Finally we look at examples where no imputation is needed at all. All of the methods will be illustrated through good examples using the appropriate tools for exploration and diagnostics. We will also touch on models for imputation for hierarchical models when a mixed effects

Topics covered will include:

- The missing data mechanisms: Illustration using directed graphical models and exploration of the missingness models using appropriate software.
- A survey of Ad-Hoc methods illustrating their drawbacks.
- Missing data in the covariates or explanatory variables.
- Full Bayesian imputation using WinBugs to demonstrate the role of the three models (the model for missingness, the imputation model and the substantive model).
- Multiple imputation using chained equations and multivariate methods
- Rubin's rules for combining the modelling of multiply imputed datasets
- Diagnostics of the imputation process.
- A survey of methods of dealing with missingness in hierarchical datasets.

On successful completion students will be able to:

- To demonstrate mastery of tools for exploring the missingness patterns using VIM and mice software libraries for R
- To formulate a possible missing data mechanism, for a given scenario, and to identify cases where the missing data mechanism is ignorable
- To formulate and differentiate: the model for missingness, the imputation model and the substantive model (model of interest)
- To be able to differentiate between sampling and parameter uncertainty and to recognise that the predictive distribution of the missing data incorporates both types of uncertainty
- To implement some naive methods for dealing with missingness (such as single imputation or list wise deletion), to recognise the limitations of each methods and identify situations where their use may be appropriate
- To be able to explain the differences between a multivariate imputation model and one using chained equations.
- To estimate the between imputation variability and the within imputation variability and to combine in a sensible way to estimate the total variability and the fraction of information lost through missingness

Bibliography:

- Stef van Buuren, 2012 Flexible Imputation of Missing Data, (Chapman & Hall/CRC Interdisciplinary Statistics Series).
- James R. Carpenter and Michael G. Kenward , 2013. Multiple Imputation and Its Application (Statistics in Practice). Wiley.

Module Mnemonic:	CFAS415
Module Title:	Structural Equation Modelling
Module Convenor:	Dr Andrew Titman
Assessment:	Coursework (100%)
Duration:	20 hours
Credits:	10
Term:	L2
Prerequisites:	(MATH551 and MATH552) OR (CFAS406 and CFAS440) & basic familiarity: SPSS
Software used:	SPSS, AMOS 22

This module will introduce participants to latent variables (variables which are not directly measured themselves) and to the use of factor analysis in investigating relationships between latent variables and observed, or measured, variables. These techniques will then be extended into the wider area of structural equation modelling, where complex models involving several latent variables will be introduced.

The module is aimed at researchers and research students who have experience of statistical modelling (up to linear regression) and hypothesis testing, who wish to develop techniques to analyse more complex data involving latent variables. The aim of the module is to provide a background of theory with opportunities to apply the techniques in practice, and each session will consist of a lecture/ demonstration and a practical. The software packages used will be IBM SPSS and AMOS, no prior knowledge of the structural equation modelling package AMOS will be assumed.

Topics covered will include:

- Introduction to latent variables and measurement error
- Exploratory and confirmatory factor analysis; measurement models
- Structural equation models
- Theoretical issues involved in the development and application of structural equation models.

Learning: Students will learn through the application of concepts and techniques covered in the module to real data sets. Students will be encouraged to examine issues of substantive interest in these studies.

On successful completion students will be able to:

- Investigate data using factor analysis
- Build and verify appropriate measurement models for latent constructs
- Confirm hypotheses and develop structural equation models
- Apply theoretical concepts
- Identify and solve problems
- Analyse data using appropriate techniques
- Interpret statistical output

Assessment: One assignment (100%) to be submitted in the form of two reports covering all aspects of the module material. The projects involve investigating datasets that require the student to investigate a substantive issue using appropriate statistical techniques and interpreting the results.

Bibliography:

- Byrne, B.M. (2010) Structural Equation Modelling with AMOS: Basic Concepts, Applications and Programming. New York: Routledge
- Kline, R. B. (2010) Principles and Practices of Structural Equation Modelling London: The Guildford Press.

Module Mnemonic:	CFAS420
Module Title:	Statistical Learning
Module Convenor:	Dr Alex Gibberd
Assessment:	Coursework (100%)
Duration:	25 hours
Credits:	15
Term:	L1
Prerequisites:	(MATH551 and MATH552) OR (CFAS406 and CFAS440) Basic familiarity: SPSS and R
Software used:	R, SPSS, Latent Gold

The module will provide students with the statistical tools needed to understand the analysis of large data sets, and the statistical background to such tools. It will seek to integrate the various methods used in such analysis into a common modelling framework. An important part of the course is on interpretation and on communicating the results to others.

Topics covered will include:

- Introduction to statistical learning; problems of missing data, biased samples and recency.
- Statistical significance and big data.
- Sample splitting. Calibration, training and validation samples. Entropy and likelihood.
- Unsupervised learning: K-means, PAM and CLARA for big data. Mixture models. Latent class analysis.
- Variable reduction methods and variable selection. The Lasso.
- Classification methods: logistic and multinomial logistic models. Probability cutoffs; the ROC curve; sensitivity and specificity.
- Classification methods Regression trees, random forests and boosted trees
- Classification methods Neural networks as generalised linear modelling extensions.
- Classification methods: other methods (PRIM)
- Smoothing models through GAMs.
- Bayesian networks

Learning: Students will learn through the application of concepts and techniques covered in the module to real data sets. Students will be encouraged to examine issues of substantive interest in these studies.

On successful completion students will be able to:

- Understand the need for a statistical basis for data analytics.
- Appreciate that different terminologies used in different data analytic technologies can be integrated through statistical modelling concepts and the idea of likelihood.
- Understand the tradeoff between interpretability and predictive performance.
- Have gained skills about the appropriate choice of statistical learning methods for various forms of real-life problem.
- Build predictions for logistic and multinomial logistic models.
- Choose an appropriate clustering method which has a statistical basis.
- Split big datasets appropriately, and understand the predictive performance should be based on the validation sample.
- Carry out a regression tree analysis including pruning, assessing its performance
- Carry out more complex forms of regression tree ensemble techniques, including random forests.
- Carry out simple neural network analyses, while understanding the need for construction of a start value strategy.

Assessment: Two assignments (100%) to be submitted in the form of reports covering all aspects of the module material. The projects involve the analysis of datasets that require the student to investigate a substantive issue using appropriate statistical learning techniques and interpreting the results.

Bibliography:

- Bramer, M. (2016) Principles of Data Mining Third Edition. New York: Springer
- Efron, B. and Hastie, T. (2017) Computer Age Statistical Inference: Algorithms, Evidence, and Data Science. Cambridge: Cambridge University Press.
- Hastie, T., Tibshirani, R. and Friedman, J. (2009) The Elements of Statistical Learning, Second edition. New York, Springer.
- James, G., Witten, D., Hastie, T. and Tibshirani, R. (2013) An Introduction to Statistical Leaming with Applications in R. New York: Springer
- Williams, G. (2011) Data Mining with Rattle and R. New York: Springer

Module Mnemonic:	CFAS440
Module Title:	Statistical Methods and Modelling
Module Convenor:	Dr Kanchan Mukherjee & Dr Gareth Ridall
Assessment:	Coursework (100%)
Duration:	45 hours
Credits:	15
Term:	M1, M2
Software used:	R

The aim of this module will be to address the fundamentals of statistics for those who do not have a mathematics and statistics undergraduate degree. Building upon the pre-learning 'mathematics for statistics' module is delivered over five weeks via a series of lectures and practical's. Students will develop an understanding of the theory behind core statistical topics; sampling, hypothesis testing, and modelling. They will also be putting this knowledge into practice, by applying it to real data to address research questions. The module is an amalgamation of three short courses and is taught in weeks 1 - 5:

The module is an amalgamation of three short courses and is taught in weeks 1-5:

Topics covered will include:

- Statistical Methods; commonly used probability distributions, parameter estimation, sampling variability, hypothesis testing, basic measures of bivariate relationships.
- Generalised Linear Models; the general linear model and the least-squares method, logistic regression for binary responses, Poisson regression for count data. More broadly, how to build a flexible linear predictor to capture relationships of interest.

These short courses are supported by tutorial sessions and office hours.

On Successful completion students will be able to:

- Comprehend the mathematical notation used in explaining probability and statistics.
- Demonstrate knowledge of basic principles in probability, statistical distributions, sampling and estimation.
- Make decisions on the appropriate way to test hypothesis, carry out the test and interpret the results.
- Demonstrate knowledge of the general linear model, the least-squares method of estimation, and the linear predictor. As well the extensions to generalised linear models for discrete data.
- Decide on the appropriate way to statistically address a research question.
- Carry out said statistical analyses, assessing model results and performance.
- Report their findings in context

Assessment: There will be three pieces of coursework:

- One assessment for Statistical Methods; assessing understand and application of statistical concepts, and interpretation of results from hypothesis testing.
- Two independently produced reports for Generalized Linear Models; centred on in-depth statistical analyses.

Bibliography:

- Upton, G., & Cook, I. (1996). *Understanding statistics*. Oxford University Press.
- Rice, J. (2006). *Mathematical statistics and data analysis*. Cengage Learning.
- Dobson, A. J., & Barnett, A. G. (2008). *An Introduction to Generalized Linear Models*. CRC Press.
- Fox, J. (2008). *Applied regression analysis and generalized linear models*. Sage Publications.

Module Mnemonic:	CHIC565
Module Title:	Environmental Epidemiology
Module Convenor:	Dr Ben Taylor
Assessment:	Coursework (50%) and written exam (50%)
Duration:	20 hours (intensively teaching in week 19)
Credits:	10
Term:	L2 (weeks 19/20)
Prerequisites:	MATH551; MATH552; R
Software used:	R

This course aims to introduce students to statistical methods commonly used by epidemiologists and statisticians to investigate the relationship between risk of disease and environmental factors. Specifically the course concentrates on studies with a spatial component. A number of published studies will be used to illustrate the methods described, and students will learn how to perform similar analyses using the R statistical package. By the end of the course students have an awareness of methodology used in environmental epidemiology, including an appreciation of their limitations, and should be capable of a number of these analyses themselves.

Topics covered will include:

- Introduction: Motivating examples for methods in course
- Spatial Point Processes: theory and methods for the analysis of point patterns in two - dimensional space
- Clustering of disease: case-control point-based methods and methods based on counts
- Spatial variation in risk: case-control and point-based methods; generalized additive models
- Disease mapping: investigating variation in risk with count data
- Geographical correlation studies: the ecological fallacy; relation with disease mapping
- Point source methods: Investigation of risk associated with distance from a point or line source, for point and count data
- Geostatistics: introduction to the analysis of geostatistical data. Kriging and spatial prediction

On successful completion students will be able to:

- Define and give examples of spatial point processes; describe the first and second moments of a point process
- Define, estimate and calculate theoretical K functions for a spatial point process
- Test for spatial clustering of a point pattern using the K function
- Use generalised additive models to construct smooth maps of spatial variation in disease risk and interpret key model outputs
- Use Poisson regression to analyse area-level count data and interpret key model outputs
- Describe what is meant by the ecological fallacy
- Carry out simple analyses of case-control data in relation to a point source
- Gaussian geostatistical models including a Gaussian process random effect term
- Perform basic analyses of geostatistical data, define and interpret the variogram
- Recognise the difference between point process data, area-level data and geostatistical data
- Describe some practical issues involved in undertaking environmental epidemiology studies.

Bibliography:

- P.J. Diggle. Statistical Analysis of Spatial Point Patterns (2nd edition). London: Edward Arnold. 2003.
- P.Elliott, M.Martuzzi and G. Shaddick, Spatial statistical methods in environmental epidemiology: a critique. Statistical methods in Medical Research, 4, 137-159, 1995.
- P.Elliott, J. Wakefield, N. Best and D. Briggs (eds), Disease and Exposure Mapping. Oxford University Press, Oxford, 1999.
- L. Waller and C.A. Gotway. Applied Spatial Statistics for Public Health Data. New York: Wiley, 2004

Module Mnemonic:	CHIC571
Module Title:	Modelling of infectious diseases
Module Convenor:	Dr Johnathan Read
Assessment:	Presentation (20%) and project (80%)
Duration:	20 hours intensive teaching in week 17, 80 hours private study week 18
Credits:	10
Term:	L2 (weeks 14 - 15)
Prerequisites:	(MATH551 and MATH552) OR (CFAS406 and CFAS440), R
Software used:	R

This module aims to provide students with the necessary knowledge, and analytical and modelling skills to develop and fit mathematical transmission models to understand infection dynamics, explore interventions, and to inform control policy. It will also provide students with the ability to analyse outbreak information, and to implement transmission models using the R programming language. Students will gain experience of handling and linking epidemiological data relevant to infectious disease outbreaks. They will gain hands-on experience of developing transmission models, appropriate to a specific research question or epidemiological application, and of using those models for scenario exploration. Students will also gain experience in communicating and presenting epidemic models and their outputs.

Topics covered will include:

- Construction of mathematical disease models appropriate to their purpose;
- The differences between deterministic and stochastic infectious disease modelling frameworks;
- The dynamical behavior of infectious disease models;
- Statistical inference using infectious disease models;
- Analysis and critical interpretation of infectious disease data for outbreak analysis;
- Communication of disease models and interpretation of their output.

On successful completion students will be able to:

- Demonstrate a deep understanding of the role of mathematical modelling in epidemiology;
- Take a critical approach to linking sources of epidemiological data required for infectious disease models;
- Understand epidemiology of infectious disease and modelling literature
- Interpret modelling studies critically;
- Take a responsible approach towards the use of mathematical modeling, and appreciate and the ethical and social impacts of research and practice within this subject area.

Bibliography:

- Keeling MJ, Rohani P. Modelling Infectious Diseases in Humans and Animals. Princeton University Press. 2007.
- Andersson H, Britton T. Stochastic Epidemic Models and their Statistical Analysis. Lecture Notes in Statistics. Springer. 2000.

Module Mnemonic:	CHIC581
Module Title:	Statistical Genetics and Genomics
Module Convenor:	Dr Joanne Knight & Dr Frank Dondelinger
Assessment:	70% Practical and 30% Essay.
Duration:	21 hours (weeks 18 and 20).
Credits:	10
Term:	L2
Pre-requisites	(MATH551 and MATH552) or (CFAS440 and CFAS406)
Software	R

This module will give the students a working knowledge of recent statistical approaches for analyzing modern genomic and genetic datasets. The students will learn about significance testing for genetic variants using logistic regression, multiple testing correction using strategies such as Bonferroni and False discovery rate control, quantification of gene expression in RNA-seq data using expectation-maximization to determine ambiguous isoforms, differential expression testing using a negative binomial model, and Bayesian network models for gene regulation.

Topics covered will include:

- Introduction to Molecular Biology
- Introduction to Human Genetics Studies
- Genome wide associations studies (QC, analysis, multiple testing correction, population stratification)
- RNA-Seq gene expression analysis
- Differential Gene Expression
- Statistical Models for gene regulation

On successful completion students will be able to:

- Discuss the key aspects of genetics and genomics
- Define the statistical challenges in the analysis of genetics and genomics data
- Explain Genome-Wide Association Studies (GWAS) and how to find trait markers
- Perform a GWAS analysis and assess the significance of identified risk variants
- Identify differentially expressed genes in RNA-seq gene expression data
- Sketch the process of gene regulation and model it using statistical tools
- Understand the kinds of methods used in statistical genomics and genetics, including their limitations
- Analyse complex genetic and genomic datasets using statistical programming packages
- Perform a literature survey of statistical applications to a novel scientific field

Bibliography:

- Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. (2007) The Wellcome Trust Case Control Consortium, Nature. 447, 661-678 http://www.nature.com/nature/journal/v447/n7145/full/nature05911.html
- *RNA-seq Data Analysis: A Practical Approach* (2014) Eija Korpelainen, Jarno Tuimala, Panu Somervuo, Mikael Huss and Garry Wong, CRC Press.
- *Data Analysis for the Life Sciences* (2015) Rafael A. Irizarry and Michael I. Love. Leanpub. https://leanpub.com/dataanalysisforthelifesciences/

Module Mnemonic:	LEC402
Module Title:	Geoinformatics
Module Convenor:	Prof Alan Blackburn
Assessment:	100% Coursework
Duration:	30 hours
Credits:	15
Term:	L1/L2
Prerequisites:	(MATH551 and MATH552) or (CFAS440 and CFAS406)
Software used:	ArcGIS & ERDAS Imagine

This module introduces students to the fundamental principles of Geographical Information Systems (GIS) and Remote Sensing (RS) and shows how these complimentary technologies may be used to capture/derive, manipulate, integrate, analyse and display different forms of spatially-referenced environmental data. The module is highly vocational with theory-based lectures complimented by hands-on practical sessions using state-of-the-art software (ArcGIS & ERDAS Imagine). In addition to the subject specific aims, the module provides students with a range of generic skills to synthesise geographical data, develop suitable approaches to problem solving, undertake independent learning (including time management) and present the results of analysis in novel graphical formats.

Topics covered will include:

- Geoinformatics: definitions, components and the nature of spatial data
- Principles of RS: physical basis, sensors, platforms and systems
- Applications of RS
- Principles of GIS
- Vector GIS
- Raster GIS and spatial modelling
- Geoinformatics project design
- Data Integration and Metadata

On successful completion of this module students will be able to

- Recognise fundamental principles and applications of GIS and Remote Sensing;
- Appreciate the strong linkages between these disciplines and their fusion to create meaningful spatially-referenced environmental information;
- Appraise current and future potential applications;
- Use state-of-the-art software packages such as ArcGIS and ERDAS Imagine;
- Demonstrate project management skills through completion of a geoinformatics project.
- Identify and retrieve spatial data from a variety of different sources
- Visualise analyse and interpret spatial data using simple and advanced approaches
- Conduct an independent piece of research

Bibliography:

- Demers. M.N., 2009. *GIS for Dummies*.
- Heywood, I, Cornelius, S and Carver, S, 2011. *An Introduction to Geographical Information Systems* (4e). Pearson.
- Lillesand, T.M., Kiefer, R.W. and Chipman, J.W, 2008. *Remote Sensing and Image Interpretation* (6e). Wiley.
- Longley, P.A, Goodchild, M.F, Maguire, D.J and Rhind, D.W, 2011. *Geographic Information Systems & Science* (3e). Wiley.

Module Mnemonic:	LEC468
Module Title:	Modelling Environmental Processes
Module Convenor:	Dr Wlodek Tych
Assessment:	Coursework (50%) and written exam (50%)
Duration:	30 hours
Credits:	15
Term:	L1
Prerequisites:	(MATH551 and MATH552) or (CFAS440 and CFAS406)
Software used:	Matlab/Simulink package

This module provides an introduction to the basic principles and approaches to computer-aided modelling of environmental processes with applications to real environmental problems such as catchment modelling, pollutant dispersal in rivers and estuaries, population dynamics etc. More general, the module provides an introduction to general aspects of dynamic systems modelling including the role of uncertainty and data in the modelling process.

Topics covered include:

- Introduction to modelling as a process and as evaluation of scientific hypotheses: approaches to modelling: the role of data and perceptions in the modelling process; the problems of badly defined systems in the context of modelling environmental processes; problems of scale (temporal and spatial) and uncertainty in quantifying environmental systems.
- The concept of dynamic system. First order linear systems, with the Nicholson blowfly dynamics and the Aggregated Dead Zone (ADZ) model of dispersion in a river used as practical case studies. Transfer function models, steady state gain and time constant; serial, parallel and feedback connections of first order systems. Block diagram analysis.
- Muskingum-Cunge, Lag and Route, and General Transfer Function models of flow in a river system
- Second order linear systems with the predator-prey equations and a climate model as practical examples; natural frequency and damping ratio; higher order systems.
- Linear vs. Nonlinear systems – basic introduction.

Throughout the course case studies and examples will be used to illustrate the material. Guest lecturers may be invited to contribute depending on availability.

On successful completion of this module students will be able to:

- Evaluate the principles and problems of computer aided modelling of environmental systems.
- Use contemporary industry standard numerical software for basic analysis and simulation of environmental systems.
- Communicate with mathematicians and numerical analysts in joint projects involving modelling.
- Understand the way in which simple mathematical concepts can be used to build models of environmental systems
- Undertake some simple modelling tasks, to analyse experimental data and interpret the modelling outcomes.

Bibliography:

The following texts may be useful if read with discretion.

- Young, P.C. (1993) Concise Encyclopaedia of Environmental Systems. Pergamon: Oxford (selected articles)*
- Young, P.C., Parkinson, S. and Lees, M.J. (1996) Simplicity out of complexity: Occam's Razor revisited* Journal of Applied Statistics, 23, 165-210
- Young, P.C. Recursive Estimation and Time Series Analysis. An Introduction, Springer, 1984
- Bennett, R.J., Chorley, R.J. Environmental Systems, Philosophy, Analysis and Control, Methuen 1980*
- Hardisty, J. et al. Computerised Environmental Modelling, A practical introduction using Excel, Wiley, 1993

Module Mnemonic:	MATH551
Module Title:	Likelihood Inference
Module Convenor:	Dr Alex Gibberd
Assessment:	Coursework (30%) and written exam (70%)
Duration:	25 hours (weeks 1 to 5)
Credits:	15
Term:	M1
Prerequisites:	UG Mathematics/Statistics (probability theory; calculus; matrices etc)
Software	R

This course considers the idea of statistical models and how the likelihood function, defined to be the probability of the observed data viewed as a function of unknown model parameters, can be used to make inference about those parameters. This inference includes both estimates of the values of these parameters, and measures of the uncertainty surrounding these estimates. We consider single and multi-parameter models, and models which do not assume the data are independent and identically distributed. We also cover computational aspects of likelihood inference that are required in many practical applications, including numerical optimization of likelihood functions and bootstrap methods to estimate uncertainty.

Topics covered will include:

- Definition of the likelihood function for single and multi-parameter models, and how it is used to calculate point estimates (maximum likelihood estimates)
- Asymptotic distribution of the maximum likelihood estimator, and the profile deviance, and how these are used to quantify uncertainty in estimates
- Inter-relationships between parameters, and the definition and use of orthogonality
- Generalised likelihood ratio statistics, and their use for hypothesis tests
- Calculating likelihood functions for non-IID models
- Use of computational methods in R to calculate maximum likelihood estimates and confidence intervals; perform hypothesis tests and calculate bootstrap confidence intervals

On successful completion students will be able to:

- Understand how to construct statistical models for simple applications
- Appreciate how information about the unknown parameters is obtained and summarized via the likelihood function
- Calculate the likelihood function for basic statistical models
- Evaluate point estimates and make statements about the variability of these estimates
- Understand the inter-relationships between parameters, and the concept of orthogonality
- Perform hypothesis tests using the generalised likelihood ratio statistic
- Use computational methods to calculate maximum likelihood estimates
- Use computational methods to construct both likelihood-based and bootstrapped confidence intervals, and perform hypothesis tests

Bibliography:

- A Azzalini. Statistical Inference: Based on the Likelihood. Chapman and Hall. 1996.
- D R Cox and D V Hinkley. Theoretical Statistics. Chapman and Hall. 1974.
- Y Pawitan. In All Likelihood: Statistical Modeling and Inference Using Likelihood. OUP. 2001.

Module Mnemonic:	MATH552
Module Title:	Generalised Linear Modellling
Module Convenor:	Dr Clement Lee
Assessment:	Coursework (50%) and written exam (50%)
Duration:	25 hours (weeks 1 to 5)
Credits:	15
Term:	M1
Prerequisites:	UG Mathematics/Statistics (probability theory; calculus; matrices etc)
Software	R

Generalised linear models are now one of the most frequently used statistical tools of the applied statistician. They extend the ideas of regression analysis to a wider class of problems that involves exploring the relationship between a response and one or more explanatory variables. In this course we aim to discuss applications of the generalised linear models to diverse range of practical problems involving data from the area of biology, social sciences and time series to name a few and to explore the theoretical basis of these models.

Topics covered will include:

- We introduce a large family of models, called the generalised linear models (GLMs), that includes the standard linear regression model as a special case and we discuss the theory and application of these models
- We learn an algorithm called iteratively reweighted least squares algorithm for the estimation of parameters
- Formulation of sensible models for relationship between a response and one or more explanatory variables, taking into account of the motivation for data collection
- We fit and check these models with the statistical package R; produce confidence intervals and tests corresponding to questions of interest; and state conclusions in everyday language

On successful completion students will be able to:

- Define the components of GLM
- Express standard models (normal, poisson,…) in GLM form
- Derive relationships between mean and variance and parameters of an exponential family distribution
- Specify design matrices for given problems
- Define and interpret model deviance and degrees of freedom
- Use model deviances to assist in model selection
- Define deviance and Pearson residuals, and understand how to use them for checking model quality
- Use R to fit standard (and appropriate) GLM's to data
- Understand and interpret R output for model selection and diagnosis, and draw appropriate scientific conclusions

Bibliography:

- P. McCullagh and J. Nelder. Generalized Linear Models, Chapman and Hall, 1999.
- A.J. Dobson, An Introduction to Generalised Linear Models, Chapman and Hall, 1990.

Module Mnemonic:	MATH555
Module Title:	Bayesian Inference for Data Science (weeks 11 to 20)
Module Convenor:	Dr Marco Battison
Assessment:	Coursework (50%) and written exam (50%)
Duration:	25 hours
Credits:	15
Term:	L1, L2
Prerequisites:	MATH551, MATH552
Software used:	R

This module aims to introduce the Bayesian view of statistics, stressing its philosophical contrasts with classical statistics, its facility for including information other than the data into the analysis and its coherent approach towards inference and model selection. The module will also introduce students to MCMC (Markov chain Monte Carlo), a computationally intensive method for efficiently applying Bayesian methods to complex models. By the end of the course the students should be able to formulate an appropriate prior for a variety of problems, calculate, simulate from and interpret the posterior and the predictive distribution, with or without MCMC as appropriate and to carry out Bayesian model selection using the marginal likelihood. Students should be able to carry out all of this using the programming language R.

Topics covered will include:

- inference by updating belief
- The ingredients of Bayesian inference: the prior, the likelihood, the posterior, the predictive and the marginal distribution
- Methods for formulating the prior
- Conjugate priors for single parameter models
- Normal distribution, known and unknown variance, regression
- Sampling for the posterior and predictive distributions
- Model checking and model selection
- Gibbs sampling, data augmentation, hierarchical models
- The Metropolis-Hastings algorithm, random walk Metropolis, independence sampler

On Successful completion students will be able to:

- Understand the Bayesian statistical framework and its philosophy
- Demonstrate knowledge of key concepts: the prior, the likelihood, the posterior, the predictive and the marginal distribution
- Calculate, simulate from and interpret the posterior and the predictive distribution
- Construct an MCMC algorithm for a variety of statistical models and implement them in R

Bibliography:

- Hoff, P. (2008) A first course in Bayesian statistics. Springer
- Gamerman, D. and Lopez, H. (2006) MCMC statistical simulation for Bayesian inference. Chapman and Hall 2nd Edition.
- Gilks, W.R., Richardson, S. and Spiegelhalter, D. (1996) Markov chain Monte Carlo in Practice. Chapman and Hall.

Module Mnemonic:	MATH562
Module Title:	Extreme Value Theory
Module Convenor:	Dr Jenny Wadsworth
Assessment:	Coursework (50%) and written exam (50%)
Duration:	20 hours (intensive teaching mode in week 11)
Credits:	10
Term:	L1 (weeks 11/12)
Prerequisites:	MATH551, MATH552
Software used:	R

This module develops the asymptotic theory, and associated techniques for modelling and inference, required for the analysis of extreme values of random processes. The course will focus on the mathematical basis of the models, the statistical principles for implementation and the computational aspects of data modelling. Students are expected to acquire the following: an appreciation of, and facility in, the various asymptotic arguments and models; an ability to fit appropriate models to data using specially developed R software; the ability to understand and interpret fitted models.

For many physical processes, especially environmental processes, it is extremes of the process that are of greatest concern; the highest sea-levels cause floods; the fastest wind-speeds destroy buildings, etc. Most of the statistical theory is concerned with modelling typical behaviour; in contrast, the analysis of extremes requires us to model the unusual. This means that we have very little data with which we can either develop or estimate models. In the absence of alternatives, asymptotic theory is used as the basis for model development, but the issue of data scarcity leads to interesting challenges for creating models that optimise such data as are available.

Topics covered will include:

- Asymptotic theory for maxima of univariate independent and identically distributed (iid) random variables: limit distributions, Generalised Extreme Value distribution, and domains of attraction.
- Extension of asymptotic theory for univariate iid variables to cover top order statistics and threshold exceedances: Generalised Pareto distribution.
- Statistical modelling and inference using maxima and threshold methods.
- Statistical modelling of extremes of non-identically distributed random variables.
- Asymptotic theory and statistical methods for extreme values of stationary sequences: clustering, extremal index.

On successful completion students will be able to :

- Understand the asymptotic distribution of maxima and threshold exceedances, and the asymptotic theory that leads to these distributions
- Model extreme value data in the statistical program R, and interpret the model fit correctly
- Understand the concept of return levels
- Fit and interpret models with covariates
- Understand the role that dependence plays in altering the distribution of maxima and threshold exceedances
- Account for dependence in a practical data analysis

Bibliography:

- S G Coles, An Introduction to the Statistical Modelling of Extreme Values, Springer-Verlag, London, 2001.

Module Mnemonic:	MATH563
Module Title:	Design and Analysis of Clinical Trials
Module Convenor:	Dr Andrew Titman
Assessment:	Coursework (50%) and written exam (50%)
Duration:	20 hours (intensive teaching in week 11).
Credits:	10
Term:	L1 (weeks 11/12)
Prerequisites	(MATH551 and MATH552) or (CFAS440 and CFAS406)
Software	R

This course aims to introduce students to aspects of statistics, which are important in the design and analysis of clinical trials.

Clinical trials are planned experiments on human beings designed to assess the relative benefits of one or more forms of treatment. For instance, we might be interested in studying whether aspirin reduces the incidence of pregnancy-induced hypertension; or we may wish to assess whether a new immunosuppressive drug improves the survival rate of transplant recipients. On completion of the module students should understand the basic elements of clinical trials, be able to recognise and use principles of good study design, and be able to analyse and interpret study results to make correct scientific inferences.

Topics covered will include:

- Clinical trials fundamentals: trial terminology, principles of sound study design and ethics
- Defining and estimating treatment effects: continuous and binary data
- Crossover trials: motivation, design issues and analyses
- Sample size determination; continuous and binary data
- Equivalence and Non-inferiority trials
- Systematic reviews and Meta Analysis

On Successful completion students will be able to:

- Understand the basic elements of clinical trials
- Recognise and use principles of good study design, and be able to analyse and interpret study results to make correct scientific inferences
- Determine the different approaches that can be taken in addressing clinical questions related to the effectiveness of treatments and other types of interventions

Bibliography:

- D.G. Altman, Practical Statistics for Medical Research, Chapman and Hall, 1991.
- S. Senn, Cross-over trials in clinical research, Wiley, 1993.
- S. Piantadosi, Clinical Trials: A Methodologic Perspective, John Wiley & Sons, 1997.
- ICH Harmonised Tripartite Guidelines.
- J.N.S. Matthews, Introduction to Randomised Controlled Clinical Trials, Arnold, 2000.

Module Mnemonic:	MATH564
Module Title:	Principles of Epidemiology
Module Convenor:	Dr Tom Palmer
Assessment:	Coursework (50%) and written exam (50%)
Duration:	20 hours (intensive teaching in week 13)
Credits:	10
Term:	L1 (weeks 13/14)
Prerequisites	(MATH551 and MATH552) or (CFAS440 and CFAS406)
Software	R

This course introduces the principles of epidemiology and the statistical methods applied in epidemiological studies. It also introduces important concepts related to study design and statistical modelling concepts such as confounding and mediation.

Topics covered will include:

- The history of epidemiology and the role of statistics therein;
- Measures of health and disease including incidence and prevalence;
- Traditional approaches to controlling for confounding including matching and stratification;
- Epidemiological study design including cohort studies, case-control studies, cross-sectional studies, ecological studies;
- Making causal inferences in epidemiology including the use of directed acyclic graphs to describe confounding, collider bias, and mediation;
- Properties of parameters such as odds ratios and risk ratio including collapsibility;
- Critical appraisal of published epidemiological journal articles including an appreciation of their structure, and strengths and weaknesses.

On successful completion students will be able to:

- Define and calculate appropriate measures of disease incidence and prevalence;
- Describe the key statistical issues in the design of ecological studies, case-control studies, cohort studies, and cross-sectional studies;
- Discuss and implement strategies for dealing with confounding and mediation;
- Define and estimate important parameters such as the risk difference, risk ratio, and odds ratio;
- Discuss the strengths and weaknesses of a published epidemiological paper and summarise these for different audiences.

Bibliography:

- Clayton D. and Hills M. (1993) Statistical models in epidemiology. Oxford, Oxford University Press.
- Rothman K.J., Greenland S. and Lash T.L. (2008) Modern Epidemiology. Lippincott, Williams and Wilkins, US.

Module Mnemonic:	MATH566
Module Title:	Longitudinal Data Analysis
Module Convenor:	Dr Emma Eastoe
Assessment:	Coursework (50%) and written exam (50%)
Duration:	20 hours (intensive teaching in week 15)
Credits:	10
Term:	L2 (weeks 15/16)
Prerequisites	MATH551; MATH552
Software used:	R

Longitudinal data arise when a time-sequence of measurements is made on a response variable for each of a number of subjects in an experiment or observational study. For example, a patient's blood pressure may be measured daily following administration of one of several medical treatments for hypertension. The practical objective of many longitudinal studies is to find out how the average value of the response varies over time, and how this average response profile is affected by different experimental treatments. This module presents an approach to the analysis of longitudinal data, based on statistical modelling and likelihood methods of parameter estimation and hypothesis testing.

The specific aim of this module is to teach students a modern approach to the analysis of longitudinal data. Upon completion of this course the students should have acquired, from lectures and practical classes, the ability to build statistical models for longitudinal data, and to draw valid conclusions from their models.

Topics covered will include:

- What are longitudinal data?
- Exploratory and simple analysis strategies
- Normal linear model with correlated errors
- Linear mixed effects models
- Non-normal responses with GLMs
- Dealing with dropout

On Successful completion students will be able to:

- Explain the differences between longitudinal studies and cross-sectional studies
- Select appropriate techniques to explore data
- Compare different approaches to estimation and their usage in the analysis
- Build statistical models for longitudinal data and to draw valid conclusions from their models
- Express the problems arising in longitudinal studies in mathematical language
- Use computer packages in statistical modeling and analysis of longitudinal data
- Summarise the findings in writing and present to wider audience

Bibliography:

- H. Brown and R. Prescott, Applied Mixed Models in Medicine, Wiley, 1999.
- P.J. Diggle, P. Heagerty, K.Y. Liang and S.L. Zeger, Analysis of Longitudinal Data (second edition), Oxford University Press, 2002.
- G.M. Fitzmaurice, N. M. Laird and J. H. Ware, Applied Longitudinal Analysis, Wiley Series in Probability and Statistics, 2004.
- G. Verbeke and G. Molenberghs, Linear Mixed Models for Longitudinal Data, Springer, 2000.
- R. E. Weiss, Modelling longitudinal data, Springer, 2005.

Module Mnemonic:	MATH573
Module Title:	Survival and Event History Analysis
Module Convenor:	Dr Kanchan Mukherjee
Assessment:	Coursework (50%) and written exam (50%)
Duration:	20 hours (intensive teaching in week 17)
Credits:	10
Term:	L2 (weeks 17 - 18)
Pre-requisites:	MATH551; MATH552
Software	R

This course aims to describe the theory and to develop the practical skills required for the analysis of medical studies leading to the observation of survival times or multiple failure times. By the end of the course students should be able to carry out sophisticated analyses of this type, should be aware of the variety of statistical models and methods now available, and understand the nature and importance of the underlying model assumptions.

In many medical applications interest lies in times to or between events. Examples include time from diagnosis of cancer to death, or times between epileptic seizures. This advanced course begins with a review of standard approaches to the analysis of possibly censored survival data. Survival models and estimation procedures are reviewed, and emphasis is placed on the underlying assumptions, how these might be evaluated through diagnostic methods and how robust the primary conclusions might be to their violation.

The course closes with a description of models and methods for the treatment of multivariate survival data, such as repeated failures, the lifetimes of family members or competing risks. Stratified models, marginal models and frailty models are discussed.

Topics covered will include:

- Survival data. Censoring. Survival, hazard and cumulative hazard functions. Kaplan-Meier plots. Parametric models and likelihood construction. Cox's proportional hazards model, partial likelihood, Nelson-Aalen estimators. Survival time prediction
- Diagnostic methods. Schoenfeld and other residuals. Testing the proportional hazards assumption. Detecting changes in covariate effects
- Frailty models and effects. Identifiability and estimation. Competing risks. Marginal models for clustered survival data

On successful completion students will be able to:

- Apply a range of appropriate statistical techniques to survival and event history data using statistical software
- Accurately interpret the output of statistical analyses using survival models fitted using standard software
- Construct and manipulate likelihood functions from parametric models for censored data
- Identify when particular models are appropriate through the application of diagnostic checks and model building strategies
-

Bibliography:

rsen, O. Borgan, R.D Gill, and N.Keiding, Statistical Models Based on Counting Processes. Springer, 1993.

- P. Hougaard, Analysis of Multivariate Survival Data. Springer, 2000.
- T.M. Therneau and P.M. Grambsch, Modelling Survival Data: Extending the Cox Model. Springer, 2000.
- T.H. Fleming, and D.P. Harrington, Counting processes and survival analysis. Wiley, 1991.

Module Mnemonic:	MSCI523
Module Title:	Forecasting
Module Convenor:	Dr Sven Crone
Assessment:	Coursework (100%)
Credits:	10
Term:	L1, L2
Prerequisites:	(MATH551 and MATH552) OR (CFAS406 and CFAS440)
Software	TBC

The module introduces time series and causal forecasting methods so that passing students will be able to prepare methodologically competent, understandable and concisely presented reports for clients. By the end of the course, students should be able to model causal and time series models, assess their accuracy and robustness and apply them in a real world problem domain.

Topics covered will include:

- Introduction to Forecasting in Organisations: Extrapolative vs. Causal Forecasting; History & academic research in Forecasting; Forecasting case studies.
- Data Exploration: Time Series Patterns; Univariate & Multivariate Visualisation; Naïve Forecasting Methods & Averages.
- Exponential Smoothing Methods: Single, Seasonal & Trended Exponential Smoothing; Model Selection; Parameter Selection.
- ARIMA Methods: AR-, MA-, ARMA and ARIMA Models; ARIMA Model specification & estimation; Automatic selection.
- Time Series Regression : Simple & multiple regression on time series; Hypothesis testing; Model evaluation; Diagnostics
- Time Series Regression: Model specification and constraints; Dummy Variables, Lag, Non-linearities; Stationarity; Building regression models.
- Applications in operations and marketing.
- Judgmental Forecasting: Judgmental methods for forecasting; Biases and heuristics.

Bibliography:

- Ord K. & Fildes R. (2013), Principles of Business Forecasting, South-Western Cengage Learning.
- G. James, D. Witten, T. Hastie and R. Tibshirani (2013) An Introduction to Statistical Learning: with Applications in R,
- Springer M.R. Berthold, C. Borgelt, F. Höppner and F. Klawonn (2010) Guide to Intelligent Data Analysis,

SpringerP.-N. Tan, M. Steinbach and V. Kumar (2005). Introduction to data mining. Boston, Pearson Addison Wesley

Module Mnemonic:	MSCI526
Module Title:	Introduction to Intelligent Data Analysis
Module Convenor:	Dr Nicos Pavlidis
Assessment:	Coursework (100%)
Credits:	10
Term:	L1, L2
Prerequisites:	(MATH551 and MATH552) OR (CFAS406 and CFAS440)
Software	R

At The heart of many real-world industrial and scientific problems are increasingly large data sets that need to be analysed efficiently in order to gain novel and useful insights. The field of intelligent data analysis, also known as data mining brings together real large-scale datasets and algorithms from statistics, machine learning and computational intelligence that can work efficiently with real-world datasets.

The course provides an introduction to the fundamental methods and approaches from the interrelated areas of data mining, statistical/ machine learning, and intelligent data analysis. The course covers the entire data analysis process, starting from the formulation of a project objective, developing an understanding of the available data and other resources, up to the point of statistical modelling and performance assessment. The focus of the course is classification.

The course uses the R programming language and more specifically the RStudio integrated programming environment. The course makes extensive use of online video lectures from top scientists in the field, and has been and previously (and hopefully will continue to be) supported by DataCamp.

Topics covered will include:

- Exploratory data analysis including visualisation and simple dimensionality reduction methods
- 2. Classification methods like: Logistic Regression, Decision trees (Random forests), k-Nearest Neighbours, and Naive Bayes
- Performance assessment and model selection

On successful completion students will be able to:

- Understanding of how to approach a data mining/ statistical modelling problem in a real world application
- Understand how to use simple visualisation methods to obtain important insights into data
- Understand classification algorithms and their advantages / limitations
- Understand how to assess the performance of these algorithsm
- Use R for visualisation and to estimate statistical models

Bibliography:

- G. James, D. Witten, T. Hastie and R. Tibshirani (2013) An Introduction to Statistical Learning: with Applications in R,
- Springer M.R. Berthold, C. Borgelt, F. Höppner and F. Klawonn (2010) Guide to Intelligent Data Analysis,
- SpringerP.-N. Tan, M. Steinbach and V. Kumar (2005). Introduction to data mining. Boston, Pearson Addison Wesley

Module Mnemonic:	MSCI534
Module Title:	Optimisation and Heuristics
Module Convenor:	Prof Adam Letchford
Assessment:	Coursework (30%), Exam (70%)
Credits:	10
Term:	L1, L2
Prerequisites:	(MATH551 and MATH552) OR (CFAS406 and CFAS440)
Software	MPL, LINDO and EXCEL SOLVER etc

Optimisation, sometimes called *mathematical programming*, has applications in many fields, including Operational Research, Computer Science, Statistics, Finance, Engineering and the Physical Sciences. Commercial optimisation software is now capable of solving many industrial-scale problems to proven optimality. On the other hand, there are still many practical applications where finding a provably-optimal solution is not computationally viable. In such cases, *heuristic* methods can allow good solutions to be found within a reasonable computation time.

The course is designed to enable students to apply optimisation techniques to business problems. Building on the introduction to optimisation in MSCI502 and/or MSCI519, students will be introduced to different problem formulations and algorithmic methods to guide decision making in business and other organisations.

Topics covered will include

- Linear Programming.
- Non-Linear Programming.
- Integer and Mixed-Integer Programming.
- Dynamic Programming.
- Heuristics.

On successful completion students will be able to:

- know how to formulate problems as mathematical programs and solve them.
- be aware of the power, and the limitations, of optimisation methods.
- be able to carry out sensitivity analysis to see how robust the recommendation is.
- be familiar with commercial software such as MPL, LINDO and EXCEL SOLVER.
- be aware of major heuristic techniques and know when and how to apply them.

Bibliography:

- HP Williams (2013) Model Building in Mathematical Programming (5th edition). Wiley. ISBN: 978-1-118-44333-0 (pbk).
- J Kallrath & JM Wilson (1997) Business Optimisation Using Mathematical Programming. Macmillan. ISBN: 0-333-67623-8.
- WL Winston (2004) Operations Research - Applications and Algorithms (4th edition). Thompson. ISBN: 978-0534380588.
- DR Anderson, DJ Sweeney, TA Williams & M. Wisniewski (2008) An Introduction to Management Science. Cengage Learning. ISBN: 978-1844805952.
- E.K. Burke & G. Kendall (eds.) (2005) Search Methodologies: Introductory Tutorials in Optimization and Decision Support Techniques. Springer.

Module Mnemonic:	SCC.403
Module Title:	Data Mining
Module Convenor:	Prof Plamen Angelov, Dr Leandro Soriano Marcolino
Assessment:	100% Coursework
Credits:	15
Term:	M1, M2
Software	Python

This module will provide a comprehensive coverage of the problems related to Data representation, manipulation and processing aiming for extracting information from the data. It covers in detail different data pre-processing, anomaly detection, data partitioning, clustering and classifiers.

The module has been designed to provide a fundamental theoretical level of knowledge and skills (at the related laboratory sessions) to this specific aspect of Data Science, which plays an important role in any system and application. In this way, it prepares students for the second module on the topic of Data as well as for their projects.

Topics to be covered will include:

- Data Primer: Setting the scene: Data-rich environments, Data Streams;
- Nature of the Data: Randomness vs Determinism, frequentist and belief based approaches, probability density, averages and moments.
- Proximity/distance measures, linkages
- Data Pre-processing: Standardisation, Normlaisation, Principle Component Analysis (PCA)
- Anomaly Detection, Recursive Density Estimation
- Data Partitioning and Clustering, including:
- Hierarchical, k means, Fuzzy C Means (FCM) method
- Clustering Data Streams, evolving clustering
- Classification principles, Classifiers: Linear Discriminant Analysis (LDA)
- Artificial Neural Networks (ANN), Multi-layer Perceptron, Radial Basis Functions (RBF)
- Fuzzy Logic Classifiers
- Decision Trees based classifiers
- Support Vector Machines, Naive Bayes
- Dealing with Missing Data, Imbalanced Data

On successful completion of this module students will:

- Demonstrate understanding of the concepts and specific methodologies for data representation and processing and their applications to practical problems
- Analyse and synthesise effective methods and algorithms for data representation and processing
- Develop software scripts that implement advanced data representation and processing and demonstrate their impact on the performance
- List, explain and generalise the trade-offs of performance and complexity in designing practical solutions for problems of data representation and processing in terms of storage, time and computing power

Module Mnemonic:	SCC.411
Module Title:	Building Big Data Systems
Module Convenor:	Dr Yehia Elkhatib, Dr Peter Garragan
Assessment:	100% Coursework
Credits:	15
Term:	L2
Software	

In this module we explore the architectural approaches, techniques and technologies that underpin today's global Big Data infrastructures and particularly large-scale enterprise systems. It is one of two complementary modules that comprise the Systems stream of the Data Science MSc, which together provide a broad knowledge and context of systems architecture enabling students to assess new systems technologies, to know where technologies fit in the larger scheme of enterprise systems and state of the art research thinking, and to know what to read to go deeper.

The principal ethos of the module is to focus on the principles, emergent properties and application of systems elements as used in large-scale and high performance systems. Detailed case studies and invited industrial speakers will be used to provide supporting real-world context and a basis for interactive seminar discussions.

Topics to be covered will include:

- System architecture design and evaluation
- Scalability and performance concerns
- Systems integration/interoperability
- Problem-driven end-to-end data science project management

In addition to the discussion and seminar led aspects of the course, we envisage 'hands-on' measurement-based coursework whereby students will design and create a complete data science infrastructure pipeline.

On successful completion of this module students will:

Subject specific

- Acquire deep understanding of Big Data system architecture through a combination of fundamental principles and exposure to real world use cases of Big Data systems within industry.
- Identify and appraise optimal Big Data systems to tackle specific business and scientific objectives imposed by organisations and stakeholder.
- Architect and engineer Big Data systems, spanning stakeholder request down to the underlying infrastructure.
- Apply knowledge on concrete Big Data scenarios informed by industry (transport, energy, healthcare, etc).
- Understand and evaluate trade-offs, strengths and limitations of Big Data system architectures in principle and practice in modern IT systems.

General

- Managing and leading group work relating to specific Big Data problems.
- Identify and build skills relating to the different data science are and auxiliary roles.
- Develop improved technical analysis and scientific presentation skills.

Module Mnemonic:	SCC.413
Module Title:	Applied Data Mining
Module Convenor:	Dr. Alistair Baron
Assessment:	100% Coursework
Credits:	15
Term:	L1, L2
Software	Python

This module will provide students with up-to-date information on current applications of data in both industry and research, with a focus on how online data, particularly textual data, can be used to gain insights.

Topics to be covered will include:

- Collecting online data: practical steps needed to collect usable and useful data from the web to provide meaningful insights and answer research questions.
- Natural Language Processing pipeline: the steps required to take textual data and extract meaningful information, features, and numbers that can be used with machine learning (e.g. classification and clustering).
- Natural language processing tasks: various applications of natural language processing currently used will be covered, e.g. topic modelling, sentiment analysis, automatic summarisation, etc.
- Online social networks: connecting the myriad of user generated content available in online social networks, community and geographic analyses, making inferences, recommendations and predictions.
- Visualisation: presenting results in a meaningful way, to assist the user of the data in performing their task.
- Applications to cyber security: how data mining can be applied to current cyber security challenges, from analysis to active investigations.

On successful completion of this module students will be able to:

- Collect and process online data to address business needs and research questions.
- Perform various analyses to extract meaningful information and gain insights.
- Understand the current trends of research analysing online activity.
- Visualise results of analysis to assist the user.

Module Mnemonic:	SCC.460
Module Title:	Data Science Fundamentals
Module Convenor:	Dr Ioannis Chatzigeorgiou, Dr Keith Cheverst
Assessment:	100% Coursework
Duration:	30 hours
Credits:	15
Term:	M1, M2

This module teaches students about how data science is performed within academic and industry (via invited talks), research methods and how different research strategies are applied across different disciplines, and data science techniques for processing and analysing data. Students will engage in group project work, based on project briefs provided by industrial speakers, within multi-skilled teams (e.g. computing students, statistics students, environmental science students) in order to apply their data science skills to researching and solving an industrial data science problem.

Topics covered will include:

- The role of the data scientist and the evolving epistemology of data science.
- The language of research, how to form research questions, writing literature reviews, and variance of research strategies across disciplines.
- Ethics surrounding data collection and re-sharing, and unwanted inferences.
- Identifying potential data sources and the data acquisition processes.
- Defining and quantifying biases, and data preparation (e.g. cleaning, standardisation, etc.).
- Choosing a potential model for data, understanding model requirements and constraints, specifying model properties a priori, and fitting models.
- Inspection of data and results using plots, and hypothesis and significance tests.
- Writing up and presenting findings.

Learning: Students will learn through a series of group exercises around research studies and projects related to data science topics. Invited talks from industry tackling data science problems will be given to teach the students about the application of data science skills in industry and academia. Students will gain knowledge of:

- Defining a research question and a hypothesis to be tested, and choosing an appropriate research strategy to test that hypothesis.
- Analysing datasets provided in heterogeneous forms using a range of statistical techniques
- How to relate potential data sources to a given research question, acquire such data and integrate it together.
- Designing and performing appropriate experiments given a research question.
- Implementing appropriate models for experiments and ensuring that the model is tested in the correct manner.
- Analysing experimental findings and relating these findings back to the original research goal.

Recommended texts and other learning resources:

- O'Neil. C., and Schutt. R. (2013) *Doing Data Science: Straight Talk from the Frontline*. O'Reilly.
- Trochim. W. and Donnelly J. (2006) *The Research Methods Knowledge Base*. Atomic Dog Publishing Inc; 3rd edition

Module Mnemonic:	SCC.461
Module Title:	Programming for Data Science
Module Convenor:	Dr Leandro Soriano Marcolino (SCC) and Dr Tom Palmer (M&S)
Assessment:	Coursework (50%), End of module report (50%)
Duration:	30 hours
Credits:	15
Term:	M1, M2, L1
Software	R, Python

This module is designed for students that are completely new to programming, and for experienced programmers, bringing both to a skills level to handle complex data science problems. Beginner students will learn the fundamentals of programming, while experienced students will have the opportunity to sharpen and further develop their programming skills. The students are going to learn data-processing techniques, including visualisation and statistical data analysis. For a broad formation, in order to handle the most complex data science tasks, we will also cover problem solving, and the development of graphical applications.

In particular students will gain experience with two very important open source languages: R and Python. R is the best language for statistical analysis, being widely applied in academia and industry to handle a variety of different problems. Being able to program in R gives the data scientists access to the best and most updated libraries for handling a variety of classical and state of the art statistical methods. Python, on the other hand, is a general-purpose programming language, also widely used for three main reasons: it is easy to learn, being recommended as a "first" programming language; it allows easy and quick development of applications; it has a great variety of useful and open libraries. For those reasons, Python has also been widely applied for scientific computing and data analysis. Additionally, Python enables the data scientist to easily develop other kinds of useful applications: for example, searching for optimal decisions given a data-set, graphical applications for data gathering, or even programming Raspberry Pi devices in order to create sensors or robots for data collection. Therefore, learning these two languages will not only enable the students to develop programming skills, but it will also give them direct access to two fundamental languages for contemporary data analysis, scientific computing, and general programming.

Additionally, students will gain experience by working through exercise tasks and discussing their work with their peers; thereby fostering interpersonal communications skills. Students that are new to programming will find help in their experience peers, and experienced programmers will learn how to assist and explain the fundamental concepts to beginners.

Topics covered will include:

- Fundamental programming concepts (statements, variables, functions, loops, etc)
- Data abstraction (modules, classes, objects, etc)
- Problem-solving
- Using libraries for developing applications (e.g., SciPy, PyGames)
- Performing statistical analysis and data visualisation

On successful completion of this module students will be able to:

- Solve data science problems in an automatic fashion
- Handle complex data-sets, which cannot be easily analysed "by hand"
- Use existing libraries and/or develop their own libraries
- Learn new programming languages, given the background knowledge of two important ones

Bibliography:

- Introductory statistics with R. Dalgaard, Peter. Springer, 2008. ISBN-13: 978-0387954752
- R Cookbook. Paul Teetor. O'Reilly Media; 1 edition. 2011. ISBN-13: 978-0596809157.
- Python Documentation: https://www.python.org/doc/
- SciPy Documentation: https://www.scipy.org/docs.html
- PyGames Documentation: https://www.pygame.org/docs/

Module Mnemonic:	SCC.462
Module Title:	Distributed Artificial Intelligence
Module Convenor:	Dr Leandro Soriano Marcolino
Assessment:	100% Coursework (2 written assignments: 25% each, one project: 50%)
Duration:	30 hours
Credits:	15
Term:	L1, L2
Pre-requisites:	(MATH551 and MATH552) or (CFAS440 and CFAS406)
Software Used:	Optional. A Linux environment will be used with training in labs

Distributed artificial intelligence is fundamental in contemporary data analysis. Large volumes of data and computation call for multiple computers in problem solving. Being able to understand and use those resources efficiently is an important skill for a data scientist. A distributed approach is also important for fault-tolerance and robustness, as the loss of a single component must not significantly compromise the whole system. Additionally, contemporary and future distributed systems go beyond computer clusters and networks. Distributed systems are often comprised of multiple agents -- multiple software, humans and/or robots that all interact in problem solving. As a data scientist, we may have control of the full distributed system, or we may have control of only one piece, and we have to decide how it must behave in the face of others in order to accomplish our goals.

Therefore, a strong data scientist must go beyond "passive" data analysis. Even a very accurate classification may become useless if it does not lead to high-performing decisions in actual problems. It is fundamental to use data to create systems that are able to behave in an intelligent manner, considering the presence of multiple actors, which may or may not be cooperative with our system. The "data" may be historical information stored in files or data-bases, as in classical machine learning; or it might be arriving continuously in an "on-line" way; or it might even be the system's own experience. All that must be used in the creation of intelligent systems.

Hence, in this module we will study how to use multiple agents for creating powerful machine learning systems. Furthermore, we will go beyond data classification, and will study how to take intelligent decisions autonomously given the presence of multiple actors, whether they are cooperative or not. Therefore, topics to be covered will include:

- Fundamental concepts of multi-agent systems
- Local coordination rules and emergence
- Ensemble Systems
- Decision Theory and Game Theory
- On-line learning
- Multi-agent Reinforcement Learning

On successful completion of this module students will be able to:

- Understand the difference between single and multi-agent artificial intelligence; including the advantages and challenges of distribution
- Use a computer cluster for experimental work and data analysis
- Solve problems by using loose control, where local individual behaviour leads to complex self-organised systems
- Design systems that intelligently interact with others -- including those outside their control
- Design systems that learn from their own experience in an on-line way
- Improve classification/prediction performance by intelligently using multiple algorithms
- Read and critique research papers

The bibliography consists of research papers and course notes, which will be available during the course.

Appendix B

Academic Year 2019-2020

Welcome Week: (week 0)

30 September 2019 to 4th October

Michaelmas Term: (weeks 1 to 10)

4 October 2019 to 13 December 2019

Lent Term: (weeks 11 to 20)

10 January 2020 to 20 March 2020

Summer Term: (weeks 21 to 30)

17 April 2020 to 26 June 2020