Master's Programme in

Data Intensive Science

Course Management

Course Director: **Dr James Fergusson**
Email: director.mphildis@phy.cam.ac.uk
Department of Applied Maths and Theoretical Physics
Centre for Mathematical Science
B1.12
Wilberforce Road, Cambridge CB3 0WA, UK
Tel: +44 (0)1223 765247

Course Coordinator: **Mrs Sri Aitken**
Email: admin.mphildis@phy.cam.ac.uk
Maxwell Centre,
Department of Physics, Cavendish Laboratory,
J J Thomson Avenue, Cambridge, CB3 0HE, UK
Tel: +44 (0)1223 746634

Admissions: **Email: admissions@phy.cam.ac.uk**
Department of Physics, Cavendish Laboratory,
J J Thomson Avenue, Cambridge, CB3 0HE, UK
Tel: +44 (0)1223 337420

Table of Contents

Introduction

We are witnessing a revolution in the way we process information driven by the exponential growth in both computing power and data volume. The development of modern statistical methods and machine learning tools allow us to harness both of these resources to generate knowledge significantly faster, and in areas which were previously thought too complex for traditional analysis. It is clear that this will be a defining movement of our time and will completely transform how science, and business, is done in the future.

This MPhil programme is our response to this changing world and has been designed to provide the skills required for those who wish to lead the transformation. We are delighted to be able to share it with you and to welcome you to Cambridge.

The MPhil programme in **Data Intensive Science** is offered by the University of Cambridge as a full-time course which aims to provide education of the highest quality at master's level.

Covering topics of Machine Learning (ML) and Artificial Intelligence (AI), Statistical Data Analysis, Research Computing and Software Development and High-Performance Computing, the course will produce graduates with all the skills required for modern scientific data analysis, enabling them to participate in large experimental or observational programmes using the latest statistical and machine learning tools deployed on leading-edge computer architectures.

The programme draws on multiple disciplines for core material – including mathematics, computer science, natural sciences and technology and is applied in multiple scientific domains in the broad field of Physics, specifically: Astronomy, Cosmology, High Energy Physics, and Imaging

The MPhil is administered by the Department of Physics, in strong collaboration with the departments of the Department of Applied Mathematics and Theoretical Physics and the Institute of Astronomy. It will provide students with the theoretical knowledge, practical experience and transferrable skills required to undertake world-leading data intensive scientific research.

Course objectives

By the end of this course, students will have had the opportunity to develop:

1. A thorough knowledge of statistical analysis including its application to research and how it underpins modern machine learning methods

2. A comprehensive understanding of data science and machine learning techniques and packages and their application to several practical research domains

3. Developed advanced skills in computer programming utilising modern software development best practice and created in accordance with Open Science standards.

4. Demonstrated abilities in the critical evaluation of data science tools and methodologies for their real-world application to scientific research problems.

Overview

The MPhil in Data Intensive Science is a 10-month cross-departmental programme in the School of the Physical Sciences which aims to provide education of the highest quality at the master's level. The programme covers a broad range of skills required for modern data-driven science from the fields of machine learning and artificial intelligence (AI), statistical data analysis, and research computing.

The course structure has been designed in collaboration with our leading researchers and industrial partners to provide students with the theoretical knowledge, practical experience, and transferable skills required to undertake world-leading data-intensive scientific research.

Students will gain the broad set of skills required for scientific data analysis, covering traditional statistical techniques as well as modern machine learning approaches. Both the theoretical underpinnings and practical implementation of these techniques will be taught, with the later aspect including training on software development best practice and the principles of Open Science. The course also aims to provide students with direct experience applying these methods to current research problems in specific scientific fields. Students who have completed the course will be equipped to undertake research on data-intensive scientific projects. Beyond academic disciplines, students will be well prepared for a career as a data science professional in a broad range of commercial sectors.

The structure of the course	
Length	10 months
Course structure	Major taught modules: Michaelmas and Lent Terms. Minor taught modules: Lent and Easter Terms. Data Analysis Project: Course long.
Teaching methods	Lectures, demonstrations, and supervisions
Forms of assessment and weighting	Project report (Data Analysis project) 25% credit. Taught modules examination (mix of written assignment, written examination, and oral presentation) 75% credit. • Each major module will count for 12% of the final grade. • Each minor module will count for 7.5% of the final grade.

Course Elements

The course consists of both **taught elements** and a **data analysis project**.

The **taught elements** comprise both major and minor modules, with the former covering the essential skills for data science in research computing, statistical data analysis, and machine learning and AI and the later covering the application of these skills to specific scientific research frontiers.

The **data analysis project** will be focused on investigating the reproducibility of a key scientific data analysis in a specific research area.

All aspects feature theoretical and practical elements and have been designed to help students develop additional transferable skills, including written and oral communication skills. There are also several non-assessed courses and programmes designed to support the student which may include: seminar series, transferable skills workshops, industry led modules, and short courses on advanced topics.

Students will be required to take seven modules in total for assessment comprising **five** major and **two** minor modules. These will be chosen from:

Major modules list:

- S1 - Principles of Data Science (Michaelmas)
- M1 - Applied Data Science (Michaelmas)
- C1 - Research Computing (Michaelmas)
- S2 - Statistical Methods for Data Science (Lent)
- M2 - Applications of Machine Learning (Lent)
- C2 - Advanced Research Computing (Lent)

The dependency of major modules taken in Michaelmas and Lent terms is illustrated below:

- S1 - Principles of Data Science ⇒ S2 - Statistical Methods for Data Science
- M1 - Applied Data Science ⇒ M2 - Applications of Machine Learning
- C2 - Research Computing ⇒ C2 - Advanced Research Computing

Minor advanced topic list:

- Data Driven Astronomy in the Square Kilometre Array (SKA) era (Lent)
- Applications of Data Science to High Energy Physics (Lent)
- Applications of Data Science to Medical Imaging (Lent)
- Galactic Archaeology for Near Field Cosmology (Lent)
- Applications of Data Science to Cosmology (Easter)
- Applications of Data Science to Gravitational Waves (Easter)
- Image analysis (Easter)
- Exoplanets (Easter)

Each student will be required to submit their initial choice of modules for examination to the course administrator in writing (using the relevant form) at the start of Michaelmas Term, for approval by the Course Director. This selection is non-binding but is used to demonstrate that the student has a clear and achievable plan for completing the MPhil.

Each Student will be assigned a progress supervisor, who they will meet with them termly to discuss their progress on the MPhil.

Each student will be required to confirm the modules that they wish to be examined on, in agreement with their progress supervisor, by the middle of the term in which the module is taught, and to update their plan accordingly. This will be required each term and will again be subject to approval by the Course Director.

The meeting with the progress supervisor in Michaelmas term will also discuss the choice of projects that the student will nominate. A list of approved projects will be announced in Michaelmas term before the 27th of October. Students must submit a ranked list of project titles as part of this meeting which will be used to determine project assignment. Students are strongly advised to discuss their preferred projects with the project supervisors in advance of this meeting.

If the Course Director judges that any submitted plan represents a risk to the student's successful completion of the MPhil programme the student will be advised that this is the case and invited to discuss alternative plans with the Course Director. Subsequent to this, if the student choses to continue with the original proposed plan they may do so on the understanding that they acknowledge and accept the associated risks of doing so.

Assessments

The MPhil in Data Intensive Science is assessed via a summative assessment on five major modules, two minor modules, and one data analysis project.

Specifically, the elements will be assessed by the following items:

(i) The form of the **module assessment** will depend on each module and will be a mix of Coursework, Written examinations, and Oral assessments.

(ii) The **data analysis project** assessment will be summative assessment of:

- Project Report (maximum of 7,000 words) on a subject from the list approved by the Degree Committee.
- An Executive Summary (maximum of 1,000 words) of the project.
- The Data Analysis Pipeline used for the analysis presented in the reports.
- An Oral Presentation of the work.

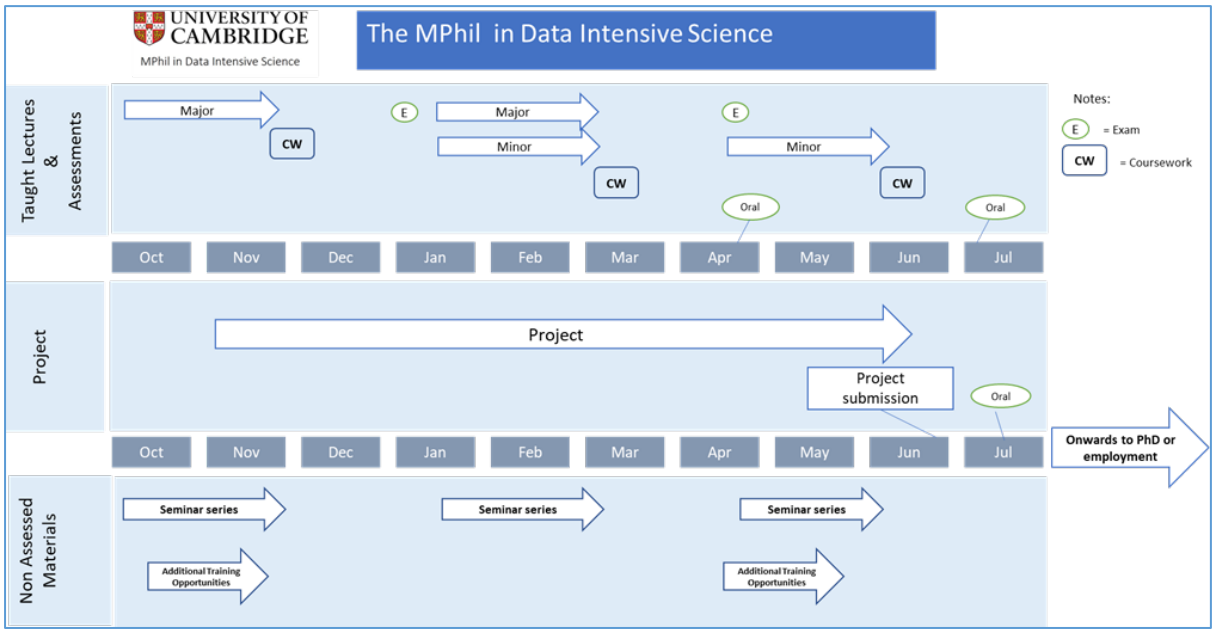
The major modules will count for 12% of the final grade each.

The minor modules will count for 7.5% of the final grade each.

The data analysis project element will count as 25% of the final grade.

Timeline

A timeline of the year is shown below:



Student representatives

Two course representatives will be selected from the student cohort whose role is to monitor all aspects of the student experience on the MPhil. Course representatives will be confirmed before the 27th of October, and they will meet regularly with the course director to discuss issues and to make suggestions for improvements to the programme. The student representatives will also represent the student cohort on the course management committee, which will meet termly to discuss the strategic direction and operational implementation of the programme.

Program Details

Lectures

MPhil students take lecture courses, or modules, consisting either of 24 lectures for Major modules or of 16 lectures for minor modules. Major modules are taught in Michaelmas and Lent terms and minor modules are taught in Lent and Easter

Lectures are given over an eight-week period in each of the Michaelmas and Lent terms and over a six-week period in the Easter term. Lectures are usually timetabled between 9am and 5pm for Monday to Friday starting at 9.00 am on Friday 6th October. The lecture timetable is published online.

All lecture courses will have their own site on Moodle, the University's Virtual Learning Environment. Further information on how to access lecture course resources on Moodle will

be provided to students at the introduction session. Lectures will be recorded; recordings are expected to be available on Panopto which will be available via the Moodle site.

Students are expected to attend lectures in person whenever possible. The recordings should be primarily used as a study aid or for those who are unable to attend lectures due to exceptional circumstances. In-person learning provides students with the opportunity to ask questions and to interact with the lecturer and their peers. Lectures may also feature practical demonstrations with live support in ways that cannot be easily replicated online.

There is no requirement that students restrict their choice of courses to those on which they wish to be examined. Courses may be selected freely from those available, within the constraints of the lecture timetable, with an expectation that the student is able to keep up with the material.

Lectures proceed at a brisk rate, and a complete understanding of the material during a lecture is not expected. You should try to appreciate the general outline of the material during lectures, and then work through the details afterwards. Given the practical nature of the programme students are expected to spend a significant time outside of lectures revising and practicing the taught skills, reading around topics, and discussing material with their peers. One feature of this research area is that there are often many valid approaches, both to solving any problem and to implementing any solution, so students are strongly encouraged to discuss problems widely to gain a full appraisal of the subject.

Further, you should take care to work through your lecture notes very soon after each lecture and complete any practical examples in them. At the beginning of each new lecture a good understanding of previous lectures will be assumed. Going through the lecture notes in detail shortly after each lecture will make it much easier to keep on top of the new material and to keep up with the course. Most lecturers are happy to answer brief questions on the lectured material.

Towards the end of each lecture course, you will be asked to complete a questionnaire. Please fill it in even if you feel that you have little to say, as the fact that you have little to say is important. The completed questionnaires are read by the lecturer and subsequently considered by the MPhil Committee.

Examples Classes

Examples sheets will be made available by the lecturer at regular intervals and there are examples classes to help you understand the material.

Attending examples classes is an important part of studying the course and valuable preparation for the assessment. You will gain much more from a class if you:

(a) prepare work in advance
(b) think about the course in general before the class and
(c) take an active part in the class
(d) actively work to support peer to peer learning by creating a supportive and non-judgemental environment. Most problems in this course have multiple ways to solve them and everyone gains from seeing the different approaches used by your diverse and interesting cohort.

Please note that although examples sheets and examples classes are very useful in helping you to learn the material in the course, they do not contribute to your formal assessment or to your final classification for the MPhil. Their sole purpose is to help you learn, so it is a good strategy to make the most of them and to participate actively.

Most MPhil courses are challenging, so the person giving the class may not be able to answer all your questions on the spot. You should also note that debugging your code is not the focus of these sessions, but the demonstrator may be able to give general guidance. For more specific support please use office hours.

Examples classes are given either by the lecturer, a post-doctoral researcher or by a senior PhD student. In the paragraphs below, the word “instructor” refers to whoever is giving the class.

If the examples classes for a course are given in more than one group, you will be assigned to a specific examples class group at the time that the examples class timetable is drawn up. You must attend the examples class group to which you have been assigned. Except for exceptional circumstances, it is not possible to change to a different examples class group. If you cannot attend a particular class, you must email the instructor at least 48h in advance of the class.

Self-assessment forms

The lecturer/instructor may ask you to complete a self-assessment form before the examples class, which will be available on the lecture course Moodle. This is invaluable in letting the lecturer/instructor know which questions have proved difficult and need more explanation during the class.

Feedback

There is an opportunity to leave feedback on examples classes on the lecture course questionnaires at the end of each term.

Progress meetings

Each student will be assigned a Progress Supervisor, who will typically be a member of the MPhil teaching team, to discuss their progress on the MPhil, to guide their choices of modules for examination, and proposed project topics, to be approved by the course director**.** The programme requires student to have three interviews during the academic year with their department contact, one in each term. **The three interviews are mandatory.**

Allocation of your Progress Supervisor

Students registered with the MPhil are assigned a Progress Supervisor by the Course Director using the subject interests identified by the students during the registration process. Every attempt is made to assign a Progress Supervisor whose scientific interests match those of the student. Students are notified of their Progress Supervisors by email during the first half of the Michaelmas term.

Progress Interviews

In preparation for progress interviews, the student will be asked by email to complete an interview form which asks about the modules they are planning to take for examination, what projects they plan to examine, their future plans, and any feedback on the programme. The completed form and any other questions about the MPhil are discussed during the interview.

The progress supervisor completes their section of the interview form by writing comments about the interview itself. The completed forms are considered by the Course Director and are used to formally register students for assessment for the MPhil modules. The interview completed in Michaelmas will be used to register students ranked preference for projects.

As well as providing an opportunity for individual discussion of progress on the MPhil, the interview process and forms are helpful in allowing the Faculty and College to identify problems so that any necessary help and support can be put in place. If a student does not attend an interview, then a reminder email is sent. If the student does not attend after this, then the College and Course Director are notified of their failure to attend, with the recommendation that the College should follow this up as a matter of urgency.

The dates and deadlines for interviews are given in the MPhil Calendar. Students are sent an email by the MPhil Administration team to let them know when and how to complete the interview forms and when to arrange the interview with their Progress Supervisor.

Students may also consult their Progress Supervisor for advice at other times outside the interviews or may ask their Progress Supervisor to write a reference for them.

Managing Your Workload

This is an intensive 10-months course where, due to the practical nature of the topic, students should expect to spend a significant fraction of their time consolidating and revising their learning. Students should work to implement the taught material in specific problems related to their problem sets, coursework and project elements. Student should also expect to spend a significant fraction of time reviewing literature related to their modules and project to gain a greater understanding of the field. Students should expect this to continue outside of normal Cambridge term times, when they will also need to complete coursework and presentations for assessment, and to prepare for examinations on the major modules. As a result, students should be careful when planning holidays to ensure they leave sufficient time for these activities.

To balance your workload throughout the year, you are recommended to carefully consider your choice of modules for examination. Students should ensure that they are on top of the material in their examined modules before considering other modules or participating in non-examined content. Student should start researching their project topic in Michaelmas and be working actively on it from the beginning of Lent term until submission.

In particular:

Students who wish to take 2 major modules for examination in Michaelmas and 3 from Lent should carefully consider both the prerequisite requirements for the Lent modules and the effect this will have on their workload balance during the year. Students considering this option should discuss it with a member of the MPhil teaching team in their progress meeting before submitting their choices for approval.

Students who wish to take 2 minor modules for examination in Lent should consider how this will affect their workload balance during that term and time for completion of the project. Students who wish to take 2 minor modules in Easter should consider how they will adapt their programme if one or more of their choices prove more challenging than expected and how it will affect their time for completion of the project. Students considering either of these options should discuss them with a member of the MPhil teaching team in their progress meeting before submitting their choices for approval.

Major Modules

The major modules supported by the MPhil cover the essential skills required for Data Intensive Science.

Each module will typically involve 24hrs of lecturer contact time in the form of lectures and supervised practical sessions. In addition, you will have demonstration classes. Students must select exactly 5 Major modules for examination. Students may follow, but not be examined on, the Major module not selected for examination if they are sure that it will not adversely affect their achievement on the programme.

Major modules list:

- S1 - Principles of Data Science (Michaelmas)
- M1 - Applied Data Science (Michaelmas)
- C1 - Research Computing (Michaelmas)
- S2 - Statistical Methods for Data Science (Lent)
- M2 - Applications of Machine Learning (Lent)
- C2 - Advanced Research Computing (Lent)

It is considered typical for student to choose 3 major modules for examination in Michaelmas term and 2 major modules for examination in Lent. This is because the major modules offered in Lent term build on knowledge obtained in Michaelmas term. The dependency of modules is illustrated below:

- S1 - Principles of Data Science $\Rightarrow$ S2 - Statistical Methods for Data Science
- M1 - Applied Data Science $\Rightarrow$ M2 - Applications of Machine Learning
- C1 - Research Computing $\Rightarrow$ C2 - Advanced Research Computing

As stated in the previous section on “managing your workload”, students who wish to take 2 major modules for examination in Michaelmas and 3 from Lent should carefully consider both the prerequisite requirements for the Lent modules and the effect this will have on their

workload balance during the year. Students considering this option should discuss it with a member of the MPhil teaching team in their review meeting before submitting their choices for approval.

Each Major module will count for 12% of the final grade and will be examined via a mix of coursework and written exams.

Minor Modules

The minor modules supported by the MPhil will cover the application of the core techniques taught in the major modules to specific scientific research problems. Beyond illustrating useful data science methodology that may have broader applications, the minor modules are also intended to give insight into the goals and status of the relevant field and hence can help in preparing interested students for research in one of these areas.

Each module will typically involve 16hrs of contact time in the form of lectures and supervised practical sessions. Details of the supervision arrangements for the minor modules will be announced in lectures.

Students must select exactly 2 Minor Advanced Topic Modules for examination. Students may audit, but not be examined on, any Minor Advanced Topic Modules not selected for examination if they are sure that it will not affect their achievement on the programme.

The final list of minor modules, and their respective schedules, will be announced at the beginning of Michaelmas term. The selection of minor modules offered will be updated annually to respond to changes in the research landscape so candidates should be aware that modules offered in previous years may not continue to be available in current or future years.

Minor advanced topic list:

- Data Driven Astronomy in the Square Kilometre Array era (Lent)
- Applications of Data Science to High Energy Physics (Lent)
- Applications of Data Science to Medical Imaging (Lent)
- Galactic Archaeology for Near Field Cosmology (Lent)
- Applications of Data Science to Cosmology (Easter)
- Applications of Data Science to Gravitational Waves (Easter)
- Image analysis (Easter)
- Exoplanets (Easter)

It is considered typical for student to choose one topic for examination in Lent term and one topic for examination in Easter. This is to balance the workload of the student over the year and to leave ample time for completion of the project in Easter term. This approach also leaves space for the selection of alternatives if the students chosen modules prove more challenging than expected.

As stated in the previous section on “managing your workload”, Students who wish to take two minor modules for examination in Lent should consider how this will affect their workload

balance during that term and time for completion of the project. Students who wish to take two minor modules in Easter should consider how they will adapt their programme if one or more of their choices prove more challenging than expected and how it will affect their time for completion of the project. Students considering either of these options should discuss them with their progress supervisor in their progress interview before submitting their choices for approval.

Each Minor Advanced Topic module will count for 7.5 % of the final grade and will be examined via a mix of coursework and oral presentation.

Project

The data analysis project will be focused on investigating the reproducibility of a key scientific data analysis in the literature. The choice to focus on reproducibility is to reinforce the great importance of validation of other's work as part of the scientific process, which is the key driver of the open science movement. The projects will be chosen to be open-ended so students can improve or extend the published work. It is expected that the very best projects would reach the standard for publication themselves.

The projects have been designed for each student to demonstrate their independence and research aptitude. The project is an opportunity for students to demonstrate several core research skills:

- Review and comprehend published scientific literature.
- Understand and manipulate scientific data sets.
- Reproduce publication level data analyses.
- Critique, improve, or extend on published work.

Project topics will be announced in Michaelmas term and students will be required to indicate ordered preferences for the projects they would like to work on to the Course Coordinator in writing (using the relevant form) for approval by the Course Director. Several students will be able to be assigned to each project topic, but they will work independently on them. Students will not be able to nominate topics for projects but must select one from the approved list published by the course.

If their first preference has high demand, students may be offered their lower ranked choices to balance the distribution of projects between supervisors. This is required to ensure that supervisors have sufficient time both; to support each student completing their project, and for assessment of the projects once submitted. The number of students who will be able to do each topic will be determined by the supervisor for that topic based on the above considerations.

Allocation of projects will be decided by the following method:

- Students will provide up to 4 ranked choices for projects.
- Each choice will be given a score, where the n^{th} choice will have a score of 2^{n-1}.
- All un-ranked project topics will be assigned an equal score, which will be equal to the score that would be assigned to the next unspecified choice.
- Projects will be allocated to minimise the total score across all students.

What this scoring equates to, is specified in the table below for four students who each submit between one and four ranked choices:

Scoring matrix	1st Choice	2nd Choice	3rd Choice	4th Choice	Remaining un-ranked projects
1 ranked project	1	-	-	-	2
2 ranked projects	1	2	-	-	4
3 ranked projects	1	2	4	-	8
4 ranked projects	1	2	4	8	16

Students should note that they minimise the probability of being assigned a project they have not ranked by **submitting the maximum of 4 ranked choices**, as this maximises the score for all un-ranked projects.

This scheme is designed to ensure "**minimum unhappiness**", as it will strive to allocate the fewest number of low ranked choices. For example:

- Allocating two students their 2nd choices, which has a score of 4 (2+2), is preferred over allocating one student their 1st choice and the other their 3rd choice, which has a score of 5 (1+4).
- Allocating two students their 3rd choices, which has a score of 8 (4+4), is preferred over allocating one student their 1st choice and the other their 4th choice, which has a score of 9 (1+8).

The focus on minimising low ranked choices, (over, for example, a method which maximises the number of first preference allocations) is to minimise cases where students are allocated unsuitable projects, which they may struggle to complete.

We realise that it can be very disappointing when you are not allocated your preferred project, especially if it is one that you are particularly passionate about or you feel suits your skill set particularly well. If you are assigned a lower ranked choice you should try to remain open minded. All projects will provide a unique and valuable learning experience and working in an unfamiliar area can often provide the greatest opportunity for growth.

Students who are particularly unhappy with their project allocation should contact the Course Director to discuss the situation. Students should note that it will often not be possible to change their project allocation and the Course Director's decision on project allocations is final.

The student will typically be provided with a key publication, a reading list of supporting material and access to the data set (if not already public). The student will be offered an initial meeting with the project proposer to introduce the topic and to answer questions regarding it. After projects are assigned at the beginning of Lent, students will be offered individual meetings with the project proposer, typically 2 per term in Lent and Easter, as well as further advice and support from the teaching team.

Typically, students would use the time from when the projects are announced until the start of Lent term to perform background reading and preliminary analysis on their preferred topics to help with their selection. Students would then work on the project from the start of Lent, through to the end of Easter term when it will be submitted for examination.

Each Data Analysis Project will count for 25% of the final grade and will be examined via a mix of coursework and oral presentation.

Non-Assessed Elements

The MPhil will contain a mix of additional non-examined material designed to expand the students training in various aspects of Data Intensive Science. These can take the form of seminars, workshops, limited lecture series, hackathons or by attending a part of other master's level courses in the University with the agreement of the relevant lecturers.

Examples of specialist non-examinable training that may be offered on this MPhil DIS are:

- Weekly seminar series with speakers from industry and academia on topics in Data Intensive Science
- Scientific Communication Workshop
- Auditing of Scientific Computing in C++ lecture course
- Data intensive Science Enterprise module (Michaelmas) covering how to commercialise your research.
- Industry-led seminar-workshops on specific topics in Data Intensive Science in Easter term.

Non-Examinable training will be announced appropriately during the year as opportunities are confirmed. Attendance for all non-assessed elements is optional. It is expected that this material will change year to year so events offered in previous years may not continue to be available in current or future years.

Due to the non-examinable nature of the non-assessed elements, they do not form part of the core delivery of the MPhil but are extra-curricular and only provided in addition to the main programme at the discretion of the Course Director. As such, while every effort will be made to ensure all advertised events are provided, specific events may be withdrawn at any time without warning.

Assessments

The MPhil in Data Intensive Science is assessed via a summative assessment on five major modules, two minor modules, and one data analysis project.

Specifically, the elements will be assessed by the following items:

(i) The form of the **module** assessment will depend on each module and will be a mix of Coursework, Written examinations, and Oral presentations.

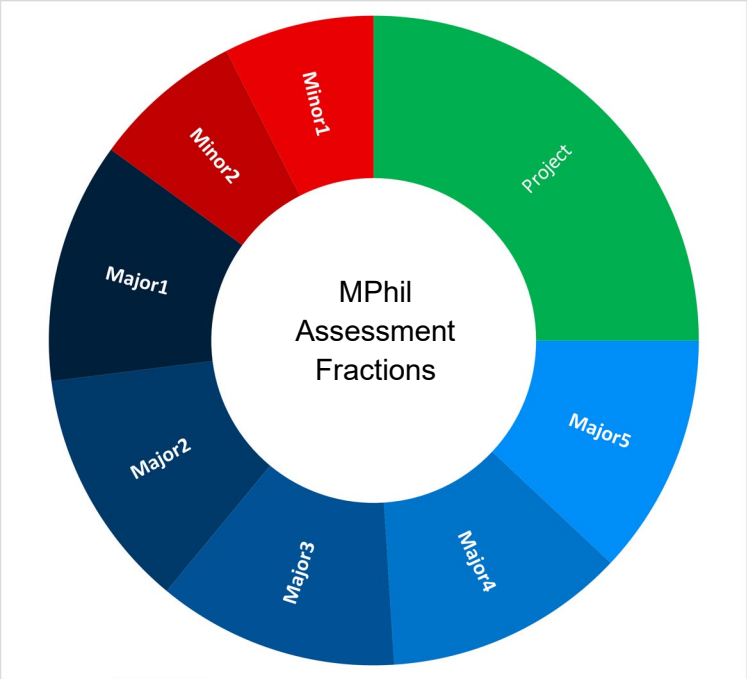
(ii) The **data analysis project** assessment will be summative assessment of:

- Project Report (maximum of 7,000 words)
- An Executive Summary (maximum of 1,000 words)
- The Data Analysis Pipeline used for the analysis presented in the reports.
- An Oral Presentation of the work.

The major modules will count for 12% of the final grade each.

The minor modules will count for 7.5% of the final grade each.

The data analysis project element will count as 25% of the final grade.



The Coursework

Coursework will typically be in the form of a report describing the development and implementation of specific data analytic methods, typically of not more than 3,000 words in length, in conjunction with the data analytic pipeline itself. However, the exact form will be module dependent and will be announced as part of each module.

The reports will be expected to be concise, with deliberately tight word limits to enforce this, and will be judged on the quality of argumentation, the clarity of presentation, and the insightfulness of interpretation. The pipeline itself will be judged on conformity to software development best practice as taught in the “Research Computing” major module, and the quality of the pipeline in terms of its accuracy, range of application, ease of use, and robustness and stability.

Each piece of submitted work must be accompanied by a standard cover sheet, including a signed declaration to the effect that the work is the student’s own unaided effort and meets the University’s guidelines and regulations on plagiarism. These are outlined in the relevant page of this site.

The Written Exams

Written exams will be closed book and will primarily test candidates' theoretical knowledge via calculations, short answer questions, and essays.

The written exams are marked by suitable assessors, but the marks are subject to moderation by the Board of Examiners of the course.

The Oral Assessments

Oral assessments will typically consist of a short presentation by the student followed by questioning by the examiners. They will be used to confirm the candidate's understanding of the associated coursework or project, and to clarify any points which were unclear in the associated report or analysis pipeline. Assessors may ask questions of the candidate during the presentation to further explore any aspect of the project, the submitted materials, the presentation, or other background knowledge relevant to the project.

Judgements are based on:

a) How well the presentation is organised and communicated. This includes the quality of slides or other supporting media, whether the discussion is clearly and logically presented with cohesion and flow.

b) The depth of understanding demonstrated. This includes demonstrating knowledge of the general area, the methods used and their applicability, providing clear rational for design choices both for algorithms and software design.

Assessment of the Major Modules

The Major modules will be graded by a summative assessment of:

- Coursework
- Written examination

The relative weighting of the two is specified below:

Principles of Data Science (Michaelmas) - 50% Coursework + 50% 3hr Exam
Applied Data Science (Michaelmas) - 67% Coursework + 33% 2hr Exam
Research Computing (Michaelmas) - 67% Coursework + 33% 2hr Exam
Statistical Methods for Data Science (Lent) - 50% Coursework + 50% 3hr Exam
Applications of Machine Learning (Lent) - 67% Coursework + 33% 2hr Exam
Advanced Research Computing (Lent) - 67% Coursework + 33% 2hr Exam

Coursework will typically be set no later than one week before the end of term and due no earlier than one week after term end with exact dates to be announced in the term in which the module falls. Exams will be taken either just before, or at the start of, the following term. This will allow time for exam preparation after coursework submission.

Assessment for any major module may include, at the discretion of the Examiners, an oral examination on the work submitted by the candidate and on the general field of knowledge within which such work falls.

Assessment of the Minor Advanced Topic Modules

The Minor Advanced Topic modules will be graded by a summative assessment of:

- Coursework
- Oral Assessment

The relative weighting will be 75% Coursework and 25% Oral Assessment. The assessor will mark the two elements collectively, using their judgement of each element to inform their judgement of the other. Candidates will only be supplied with a single mark for the minor module covering both aspects, and informal feedback on their work.

Each module will set coursework topics no later than one week before the end of term and due no earlier than one week after term end with exact dates to be announced in the term in which the module falls. Minor modules which cover multiple sub-domains may offer several coursework topics from which students may choose. In this case students will nominate coursework in preference order. Students will then be asked to complete one of their nominated coursework's (which may not be their first choice, to ensure a realistic distribution of topics and assessors) and to provide an oral presentation on it.

Assessment of the Project

The data analysis project assessment will be summative assessment of:

- Project Report (maximum of 7,000 words)
- An Executive Summary (maximum of 1,000 words)
- The Data Analysis Pipeline used for the analysis presented in the reports.
- An Oral Presentation of the work.

The relative weighting will be 50% Project Report and executive summary, 25% Data Analysis Pipeline and 25% Oral Presentation. The assessor will mark the three elements collectively, using their judgement of each element to inform their judgement of the other two. Candidates will only be supplied with a single mark for the project covering all three aspects, and informal feedback on their work.

The Final Assessment

The final grade will be the sum of the weighted grades for each component.

- A final grade of 75% or more will be considered a "Distinction".
- A final grade of 60%-74.9% will be considered a "Pass".
- A final grade of 55%-59.9% will be considered a "Marginal Fail".
- A final grade of <55% will be considered a "Fail".

A candidate who achieves a “Marginal Fail” will be examined in a viva which will be chaired by the Course Director; the External Examiner will be invited to be present to ensure appropriate academic judgements are reached. If such a candidate satisfies the examiners in a viva that they pass the degree overall, they will be recommended to be allowed the minimum mark required to satisfy conditions above. Any such candidate who does not so satisfy the examiners in a viva shall fail the degree.

Candidates who wish to appeal their final assessment should do so via the processes detailed here.

Plagiarism

Plagiarism is treated by the University with the utmost seriousness, and severe penalties are imposed whenever it is detected. This may result in a candidate failing the degree for which they are entered.

Plagiarism is presenting the work of others as if it were one’s own. If discovered by the Examiners, it will be treated as an attempt to gain credit under false pretences and may be referred to the University Court of Discipline.

Each candidate who submits a project report, essay, dissertation, or any other work for examination will be required to sign a declaration that the submission is their own work, unaided, except as may be specified in the declaration, that all sources are fully acknowledged and referenced, and that the submission does not contain material that has already been used to any substantial extent for a comparable purpose. If two or more candidates submit work in collaboration, they will each be required to sign the declaration and will be held jointly responsible for adhering to it.

The Examiners will normally consider as plagiarism any instance in which the work/ideas of another person have been included in the submission of examinable work, whether or in paraphrase, without full acknowledgement to their author. This acknowledgement must include detailed bibliographic references (including Internet addresses where appropriate) to any sources from which information or ideas have been derived.

It is appreciated that candidates will often perform practical exercises together, and that they may wish to study in groups to learn from each other and to solve problems together. However, it is essential that any material finally submitted for marking is the work of the candidate making the submission, written in their own words, and presented in their own way, with proper acknowledgement of all sources from which information has been derived, and a clear indication of the extent to which use has been made of the work of others.

Any marks awarded will be conditional on the above requirements having been met. Coursework marks contribute significantly to your overall mark. Because this work is not carried under examination conditions the distinction between beneficial co-operation and deliberate cheating should be clear in everyone’s mind.

ChatGPT and other text generating tools

New tools for code and text generation, like ChatGPT, have the potential for transforming the way we generate academic work. We expect these tools to become a regular feature for accelerating scientific research and, as such, rather than prohibiting their use we will provide guidelines for how to use them ethically and transparently in this programme.

These guidelines should not be taken as an endorsement or recommendation for use of these tools in academic work. Students must remain mindful that the use of AI tools, like ChatGPT, for new or unfamiliar tasks may inhibit their learning. Students should also remember that they will not be able to use these tools in either written or oral exams; therefore, over-reliance on them could lead to poor performance.

Candidates should carefully consider the following guidelines when considering their use in submitted work:

1. Generation tools must only be used **supportively**:
   - These tools should not be used as a substitute for your own informed work, and they should never be used for any task you could not confidently complete without them.
   - You should never use these tools to generate code in a programming language, or for a task, where you are not already proficient as you will not be able to review and debug the code effectively.
   - You should never use them to generate text on a topic you do not comprehensively understand as you will not be able to effectively proofread any output and generative AI is known to make many factual errors.
   - Where these tools are used in unfamiliar areas, or for unfamiliar tasks, e.g. by asking it to provide an example of how to use an unfamiliar API, the output should not be directly used in your work.

2. Generation tools must be used **transparently**:
   - All use of auto-generation tools must be explicitly cited in every instance of their use.
   - This applies to generating code, whether used for prototyping, creation, reformatting, or any other purpose. Students should add the citations to the README in home repository, and in any accompanying reports stating the prompts submitted, where the output was used, and how it was modified.
   - When used in conjunction with submitted reports for drafting, proofreading, suggesting alternative wordings, or for any other task it should be explicitly noted in an appendix to the report with the prompts submitted, where the output was used, and how it was modified.
   - Failure to adequately cite use of these tools is considered academic misconduct.

3. Generation tools' output must be **understood**:
   - Assessment for any module may include, at the discretion of the Examiners, an oral examination on the work submitted by the candidate and on the general field of knowledge within which such work falls.
   - If, because of the use of using auto-generation, candidates are not able to demonstrate sufficient understanding of any submitted work then they will receive zero marks for this component.
   - Students are wholly responsible for all work they submit, whether generated or otherwise, and are responsible for validating and reviewing all outputs to ensure accuracy and correctness.

Co-operation and teamwork

It is perfectly acceptable to discuss assessed work with other students or supervisors. Such discussions are beneficial, and we wish to encourage them. It is appropriate that effective use of such discussions can lead to higher marks, if it is the student who has made the main contribution to the work submitted and understands all of it.

Cooperation can go too far, however, especially if one student is effectively carried by another. Thus, while it may well be beneficial for students to discuss a problem, it is unacceptable for two students to submit effectively identical essays or other assignment work. The named author must have made the main contribution to the work submitted and the report must be in their own words. Any attempt to pass off the work of others as being produced by the named author is cheating.

Web-based plagiarism

With the proliferation of easily accessible information on the internet it is feasible for a student to use cut and paste techniques to import non-attributed material into their own work. Under no circumstances is this practice allowed and it is expressly forbidden. Sophisticated search engines are now available to staff to match passages suspected as having been plagiarised with the original source material. In circumstances where this confirms plagiarism from the internet the offending student will be immediately reported to the University authorities for disciplinary action.

The course team treats the issue of plagiarism very seriously. Integrity and responsibility in fulfilment of all course requirements is expected from all course participants.

More guidelines on plagiarism

In the UK, words and ideas are intellectual property, owned by the individual who created them, in the same way they might own land or a laptop computer. A person's intellectual property must not be used without permission. Deliberate and conscious copying is unethical and against the high standards set by scientific researchers, academic authors, and professional engineers.

When writing academic & scientific work, it is essential that the reader is clearly informed where the source material has been derived from and identify any ideas or forms of expression that are not your own. This means all sources must be accurately cited so that the person owning the intellectual property is given proper acknowledgement for the work they have done. These are the high standards which are strictly adhered to at Cambridge University.

Citing a source

This means including a reference in your text to show that material such as words, data, ideas, diagrams, software, etc. has been extracted from another source. This can be done easily by including in parenthesis the author's last name and date of publication e.g. (Smith, 2002). This reference is cross-referenced to a complete list at the end of your paper or report in the form of a Bibliography, which directs the reader to the location of the material (book, Journal, web-site page etc.). This information must be complete and accurately presented so

the reader can find the source for themselves. Not only does this approach properly acknowledge the work of others but it also allows the reader to judge how much you are relying on information from perhaps just one or two, as opposed to many, authors and how recent and up to date this information is.

In general, any specific information, which is not common knowledge, must be cited. If in any doubt whether a fact or other information is common knowledge, then a source must be cited. Other people's ideas can be included in two ways: either by quoting the source directly within quotation marks, or by paraphrasing in your own words the idea. In both cases, the reference to the source material must be cited. However direct quotes should not be overused, and it is best to only include them in your work if the author has made a point in a particularly insightful way. These quotations can complement, but cannot be a substitute for, your own line of reasoning.

A piece of work, which merely cites the ideas and results of other authors' endeavours, is not transformed into "original" work simply using extensive referencing and footnotes. It is vital that your work adds a critical dimension to this material through your own judgement and analysis.

If in any doubt make it clear to the reader by citation and references where the original idea, material or data has come from. If you don't, it will be considered as lying, cheating, stealing and an insult to the original author.

For up-to-date advice on plagiarism please see the University's website, but note that in the case of any discrepancy between the information on the website and the handbook, the handbook has precedence:

https://www.plagiarism.admin.cam.ac.uk/

Provisional Calendar / Important Dates

Although most lecture courses take place during the University term dates (see table below), coursework, exams and project-related work are carried out outside these dates. Students are required to be resident in Cambridge for the duration of the course and are expected to participate in all mandatory course activities outside the periods of Cambridge Terms.

Event Timetable	
Induction:	*2 Oct 2023*
Communication Training	*2 - 5 Oct 2023*
Written exams (Major modules)	*8 -12 Jan 2024 (MT)* *15 - 19 Apr 2024 (LT)*
Progress review and module selection sign off:	*(MT2023) before Friday 3rd Nov 2023* *(LT2023) before Friday 16th Feb 2024* *(ET2023) before Friday 17th May 2024*
Coursework submission:	*Typically no earlier than the second week after term finishes. To be announced in each module.*
Project submission:	*5pm Friday 28th June 2024*
Project Oral presentations:	*15th -19th July 2024*
Final Examiners meeting:	*Friday 26th July 2024*
Course end:	*Wednesday 31st July 2024*

University Full Terms		
	from	*to*
Michaelmas	*Tues 3 Oct 2023*	*Fri 1 Dec 2023*
Lent	*Tues 16 Jan 2024*	*Fri 15 March 2024*
Easter	*Tues 23 Apr 2023*	*Fri 14 June 2024*

Resources

Student Mental Health Support

Whilst choosing to continue your studies is an exciting opportunity to develop and gain new experiences, it is normal to experience challenges or setbacks which might impact our physical or mental wellbeing. Whether you would like to engage with mental health support or want to work on strategies to improve your wellbeing at Cambridge, please remember you are not alone.

Support is available through your college, department, university, and the wider Cambridge community. If you'd like to know more about these services and how to access them, please see the link below:

Student Support | University of Cambridge

Student Wellbeing Team and Data Intensive Science (MPhil)

Tailored specifically to taught and research postgraduates, a newly established University Student Wellbeing Service provides wellbeing and mental health support and advice. You can find out more about the service, including how to access support, through the link above.

This academic year the Student Wellbeing Service will connect with your course as part of a new pilot initiative. Our aim is to raise awareness of university support services available, to promote sustainable practices and wellbeing in education, and to encourage reaching out where help may be needed. This academic year we will offer an induction session and termly sessions in partnership with your course and department. Dates and topics covered in these termly sessions will be announced in due course.

Student Feedback Procedures:

The MPhil Academic Committee values and very strongly encourages feedback from students on the performance of its academic and administrative staff and other aspects of the MPhil programme.

Unless they get feedback in a standardised form and from a statistically significant sample of the class the staff cannot determine how the quality of their provision is changing from year

to year and from module to module and are severely hampered both in addressing problem areas and in meeting their objective of continuing quality improvement. The Course Administrator will contact the students with a university- approved polling system.

In addition to the formal mechanisms, informal feedback is welcome at any time and through any route (through student representatives, directly to the coordinator of the course or to other staff members). Any serious or potentially serious problems should be communicated as quickly as possible so that action can be taken to correct them.

Other Resources

Please visit the resources page on our website to access further resources, including the current Student's Code of Practice, Equality and Diversity information and Safety policies.

Here are some examples:

Student Information

- MPhil DIS Moodle page: Course: MPhil in Data Intensive Science (cam.ac.uk)
- Code of Practice for Master's students: Code of Practice for Master's students | Cambridge students
- Mitigation and allowances: Mitigation | Cambridge students
- Departments of Physics Intranet: Departmental Information | Department of Physics (cam.ac.uk)
- Centre for Mathematical Sciences: Index | Centre for Mathematical Sciences (cam.ac.uk)
- Institute of Astronomy: Institute of Astronomy (cam.ac.uk)
- Cambridge Libraries Services for Master's: https://libguides.cam.ac.uk/mcamguides/index
-

Equality and Diversity

- Equality and Diversity
- Disability resource centre

Health and Safety

- Health, Safety and Environment
- Health and safety training
- Safeguarding work away
- Policies and codes of practice
- Risk management
- Advice for computer users and DSE self-assessment

Student Support & Wellbeing

- Student support: Student Support | University of Cambridge
- Mental health and wellbeing: Mental health and wellbeing | Student Support (cam.ac.uk)
- Crisis support and information: Crisis support and information | Student Support (cam.ac.uk)

Career

- Careers Service**:** Careers Service | (cam.ac.uk)