

## Syllabus Overview

This course will focus on the basic building blocks of data science and introduce you to a range of topics that will enable you to make sense of the huge amounts of data that are accessible in the modern world. In this course, you will learn the basics of data science, from identifying the types of problems that data science can solve, to designing and implementing a complete data science workflow.

You will start by learning what data science is and how it already shapes the world around you. You will be introduced into the steps in a data science workflow, and how to design a data science pipeline. You will also take a deep dive into two important branches of mathematics that form the backbone of data analysis. Next you will learn all about data, including how to work with data that is organized into “dataframes” and how we can create insightful visualisations. Finally, you will learn about machine learning models and how you can use these to turn your data into useful information. In the afternoon sessions you will put your knowledge into practice using Python to perform data analysis and build data science pipelines.

Each topic you will need to enter the following information:

- **Topic Name:** Provide a clear name for the topic (e.g. “Octave for Mathematics and Modern Geometry”) and if possible, an interesting title (e.g. “Global Cooperation’ - Liberalism and the United States”).
- **Topic Description:** Provide a short paragraph outlining the main areas to be covered, any notable activities and tasks planned, and what students can expect from this series of lessons.
- **Topic Key Terms:** For each topic, you need to provide 2-3 Key Terms, complete with definitions. This will provide support for second-language (and sometimes first-language!) learners in preparation for the course and can help build confidence for students ahead of attending Immerse

## Topic 1

### Introduction to Data Science

We will explore what data is and how we can use the methods and tools available to data scientists to use data to understand and solve many real-world problems. You will practice identifying and designing a possible solution to a data science problem. Finally, we will learn about the data science workflow and data science pipelines and practice implementing our solution. In the afternoon session you will be introduced to the Python programming language, set up a workspace on your laptop, and practice some programming fundamentals in Python.

**Data:** Any collection of observations that measure something of interest, or that convey information about a question at hand.

**Data science:** A field of study that uses scientific methods, processes, and systems to extract knowledge and insights from data.

## Topic 2

### The Mathematics of Data Science

In this topic you will be introduced to two branches of mathematics that are fundamental to Data Science. First you will learn about the mathematics of probability and the concepts of a random variable and probability distribution. Secondly, we will look at the field of linear algebra by focussing on a single algorithm called Principal Component Analysis (PCA) and how it can be efficiently performed using matrix decomposition.

**Probability Mathematics:** The branch of mathematics that studies the possible outcomes of given events together with the outcomes' relative likelihoods and distributions.

**Linear Algebra:** The branch of mathematics that is concerned with mathematical structures closed under the operations of addition and scalar multiplication (linear operations).

## Topic 3

### Working with Data

Data is at the centre of data science and the ability to effectively process data is one of the most important skills for a data scientist to develop. In this topic, you will practice working with data, including loading, saving, filtering, transforming, and grouping data. You will learn the data science steps of data preprocessing and feature extraction. You will also practice creating insightful graphs to visualise relationships and structures in your data.

**pandas:** An open-source data analysis and manipulation tool (Python package), built on top of the Python programming language.

**Graph:** A diagram (such as a series of one or more points, lines, line segments, curves, or areas) that represents the variation of a variable in comparison with that of one or more other variables.

## Topic 4

### Machine Learning Models

In the final part of the course, we will focus on machine learning models. These powerful algorithms can learn from datasets to predict future trends or automatically classify new data. You will learn what we mean by “model” in the context of data science and how to build basic regression and classification models. Next, you will be introduced to a number of traditional machine learning models, and the subset of machine learning called deep learning, including a brief look into modern large language models. Finally, we will learn how to evaluate if our models are working correctly and how well they can perform their tasks.

Machine learning: A sub-field of artificial intelligence that focuses on algorithms that can learn patterns from data.

Classify: Assigning an example to one of a set of predefined categories.

### Personal Project:

For your personal project will have the opportunity to work on building your own data science pipeline, using the theories and concepts learnt throughout the course to develop your understanding. This will culminate in a 15-minute presentation of your findings and work on the final day of the course.