CM3005 Data Science

Course description

Data science is a significant subfield in computer science. Data science has many application areas ranging from medicine to climate science and business analytics. This module builds on several topics covered in earlier parts of the computer science programme including mathematics, databases, programming and graphics. It provides the skillset required to gather, analyse and present data.

Target Audience
Students enrolled on BSc in Computer Science course

Course goals and objectives

Upon successful completion of this course, you will be able to:

1. Understand the scope and impact of Data Science
2. Apply acquired knowledge and skills in solving Data Science related tasks
3. Extract insights from data and draw informative conclusions
4. Generate informative and impactful data visualisation tools
5. Build and modify machine learning models by using various ML libraries and algorithms

Textbook and Readings

Specific essential readings for each week are included in the Readings page for each week. Refer to the separate Reading list in lesson 1 in this course.

Course outline

The course consists of ten topics that focus on key areas of the fundamentals of data science:

Topic 1.	**Key concepts:** 1. Data Science in academia and industry 2. Data types, data points and datasets 3. Python and Jupyter **Learning outcomes:** 1. Understand the scope and impact of Data Science 2. Familiarise with different types of data 3. Recognise structured and unstructured data 4. Familiarise with the Python programming language 5. Understand and use key features of Python syntax 6. Familiarise with the Jupyter IDE
Topic 2.	**Key concepts:** 1. Data points and datasets, qualitative and quantitative data 2. Exploratory and Explanatory Data Visualisation **Learning outcomes:** 1. Manipulate data with NumPy 2. Use indices to extract sub-tables 3. Use NumPy functionality to obtain statistical information from datasets 4. Perform basic Linear Algebra operations involving vectors and matrices 5. Evaluate determinants, ranks and traces of matrices 6. Use the pandas library to manipulate datasets

	7. Process and visualise time series data
Topic 3.	**Key concepts:** 1. Data points and datasets, qualitative and quantitative data 2. Exploratory and Explanatory Data Visualisation **Learning outcomes:** 1. Define data visualisation 2. Articulate the importance and value of visualising data 3. Describe the similarities and differences between information and scientific data visualisation 4. Explain how data visualisation can be used at different stages in a data science investigation 5. Identify practical applications of data visualisation in a range of different contexts 6. Articulate core principles of good visualisation design
Topic 4.	**Key concepts:** 1. Data points and datasets, qualitative and quantitative data 2. Measures of central tendency, measures of spread **Learning outcomes:** 1. Define population and sample and explain how these concepts are crucial in making valid visual representations of data 2. Distinguish descriptive and inferential statistics 3. Apply appropriate visualisation techniques to individual variables of each variable type

	4. Cross-tabulate and group data to combine variables of different types 5. Visualise relationships between categorical variables 6. Visualise relationships between numerical and categorical variables 7. Visualise relationship between numerical variables 8. Analyse and quantify correlations between numerical variables
Topic 5.	**Key concepts:** 1. Supervised and unsupervised learning 2. Bias-variance tradeoff 3. Evaluation and validation **Learning outcomes:** 1. Understand ML fundamentals 2. Use scikit-learn to apply ML techniques 3. Understand feature engineering and model validation
Topic 6.	**Key concepts:** 1. Text processing principles & applications 2. Text processing use cases 3. Symbolic & statistical methods **Learning outcomes:** 1. Understand text processing fundamentals 2. Apply text processing techniques 3. Manipulate unstructured data
Topic 7.	**Key concepts:** 1. NLP principles & applications 2. NLP use cases 3. Symbolic & statistical methods **Learning outcomes:** 1. Understand NLP fundamentals 2. Use nltk to apply NLP techniques 3. Manipulate and analyse language data

Topic 8.	**Key concepts:** 1. Connected data, graph as a data structure, graph components: nodes, edges, attributes, types of graphs, graph properties 2. Building graphs and networks with Python and NetworkX, graph/network traversal algorithms 3. RGB colour model, heat maps, parallel coordinates **Learning outcomes:** 1. Understand the underpinning theory of graphs and networks 2. Use algorithms to traverse graphs and to identify shortest and optimal routes 3. Use the Python programming language and libraries such as Matplotlib and NetworkX to build and visualise networks 4. Use the acquired knowledge and skills to solve real-world network-related problems and to develop and visualise large and complex networks 5. Design and develop visualisation tools presenting large and multidimensional datasets 6. Evaluate visualisation approaches based on their aesthetic, usefulness and the ability to convey information 7. Choose appropriate visualisation approaches based on the types of the data points 8. Work with datasets with various dimensions and sizes
Topic 9.	**Key concepts:** 1. ML concepts & principles 2. ML algorithms 3. Evaluation and validation

	Learning outcomes: 1. Understand ML algorithms 2. Apply ML techniques to real-world data 3. Evaluate ML solutions
Topic 10.	**Key concepts:** 1. Applying data science in a production context 2. Evaluating data science solutions 3. Bias and ethics in data science **Learning outcomes:** 1. Understand how data science concepts and principles are applied in industry 2. Gain insight into the challenges faced by data science practitioners 3. Compare and contrast different contexts for data science practice

Activities of this course

The course is comprised of the following elements:

- Lecture videos introduce the main concepts of the topics and illustrate them with examples
- Practice quizzes will be used to reinforce your learning and understanding
- Activities drive the work that you do for each topic, where you are asked to solve challenges of different types
- Graded assignments include a practical coursework assignment and a written exam.
- Discussions with your peers will help to guide your work and encourage you to explore different types of solutions to problems
- Readings will help to reinforce your learning of concepts

How to pass this course

The course has two major assessments each worth 50% of your grade:

- Coursework: this will be assessed mid way through course. The coursework comprises a variety of exercises which in total will take up to 25 hours of study time to complete.
- The examination will be two hours long, and consist of multiple choice questions and longer written answers.

Activity	Required?	Deadline week	Estimated time per course	% of final grade
Written, staff graded coursework	Yes	12	Approximately 25 hours	50%
Written examination	Yes	22	2 hours	50%