Programme Overview

The MSDS programme structure aims to comprehensively cover all major components of a data science ecosystem. There are five core courses and an optional capstone project. The elective courses are broadly categorized into three categories, namely, foundation stack, exploration stack, and application domain stack.

In the foundation stack, we provide courses that a student may need to take if he/she does not have sufficient background knowledge. In the exploration stack, we have a suite of courses that allow students to explore deeper into various data-driven techniques and issues relevant to data science. Lastly, the application domain stack has an unrestrictive elective course which focuses on a specific application domain of interest. It provides a student necessary background on the target application domain of his/her capstone project.

The Capstone Project shall give students opportunities to realize the data science ecosystem for domain-specific applications in industrial settings or academia. A project can be either group-based or individual and can be undertaken in an industrial setting (i.e., in an industry outside NTU but in Singapore) or in an academic/joint-lab (e.g., NTU Corp labs) setting inside NTU. The former enables a student to address industry-specific data science problem whereas the latter allows students opportunity to address such problem for non-industrial end users

Programme Content

Course Code	Core Courses
SD6101	Data Science Thinking
AI6102	Machine Learning Methodologies and Applications
SD6103	Data Systems
SD6104	Data Preparation
SD6105	Data Visualisation
	Bridging Course
AI6120	Python Programming
	Elective Courses
AI6104	Mathematics for Artificial Intelligence
SD6124	Probability and Statistics for Data Science
SD6128	Introduction to Economics
SD6129	Introduction to Psychology
AI6103	Deep Learning and Applications
AI6122	Text Data Management
AI6123	Time Series And Prediction
AI6130	Large Language Models
SD6106	Capstone Project
SD6122	Modern Hardware for Data Science
SD6123	Data Privacy in Data Science
SD6125	Data Mining
SD6126	Scalable Data Systems
SD6127	Network Science

*Note: Not all courses listed in the curriculum will be offered in a semester. Courses offered are subject to availability of instructors and resources.

Courses

Core

SD6101 Data Science Thinking

Given that data is generated by humans or machines, data science is deeply entangled within the human and social world. Consequently, data science as an academic discipline must move away from purely computing-inspired curricular models and integrate the analytic lenses supplied by social science theories and methodologies. For instance, one of the key goals of many data science projects is to discover patterns in data sets. However, to explain patterns and correlations requires social theory and deep contextual knowledge. Hence, a better understanding of human psychology, power, and the incentive structures in society can enable us to address limitations of purely computing-inspired solutions to address data science problems. This naturally entails a paradigm shift in the way we think about addressing data science problems.

This course introduces a new way to think about addressing data science problems by integrating classical data science ecosystem with relevant methods, theories, and perspectives of the social sciences. The teaching content of this course consists of three parts:

(a) introduction to relevant methods, theories and perspectives from social sciences;
(b) introduction to data science; and
(c) high-level data science solutions design by integrating social sciences theories and methods with computing methods.

This course is the foundation for subsequent courses in the MSDS program. It does *not* focus on the implementation of data science solutions using a programming language. Such endeavor shall be undertaken by subsequent courses in the program. This course mainly teaches students on how to “think” about data science.

AI6102 Machine Learning: Methodologies and Applications

This course aims to provide an introductory but broad perspective of machine learning fundamental methodologies and show how to apply machine learning techniques to real-world applications. It is relevant for anyone pursuing a career in AI or Data Science.

The teaching content of this course includes different machine learning methodologies in various machine learning paradigms, such as supervised learning, semi-supervised learning, unsupervised learning, etc., as well as applications of machine learning.

SD6103 Data Systems

This course is a comprehensive introduction to modern centralized data systems. The teaching content of this course includes data models, data storage systems, data processing techniques, principles of query processing, and transactions management. A data system is effectively a large collection of data structures and algorithms that work together to solve data storage and access requests.

The primary focus is on key issues that are shaping the centralized data systems industry: relational and non- relational data models, relational algebra, normalization and constraints, SQL for data science, SQL and data warehousing, query processing, storage and indexing, transactions management.

Note that the goal of the course is not to teach students how to build a database system (e.g., a traditional database system course in a computer science degree program) but rather to be able to control and use one effectively for data science applications.

SD6104 Data Preparation

Raw data appears in many situations: logs, sensor output, government data, medical research data, climate data, etc. Such data can accumulate in many places and can be messy. In a typical data science pipeline, raw data needs to be transformed structurally and semantically and cleaned before it can be used for analytics. Hence, it is paramount to understand these transformation and cleaning techniques that are realized in the early stages of data science pipeline.

This course systematically introduces these techniques. The primary focus is on key issues that are shaping the data science industry: data discovery, data validation, data structuring, data enrichment, data filtering, and data cleaning.

SD6105 Data Visualisation

This course provides you with the principles and practice to design and present effective data visualisation solutions for different datasets and visualisation objectives. The course focuses on applying design considerations that take into account the psychological principles of human visual perception, with the goal of designing effective and ethical visualisations. You will practice using a variety of techniques and tools for visualising, exploring and interacting with abstract, scientific and geographical datasets, with the purpose of instilling an audience- oriented mindset and the technical competency to explore and explain different types of datasets that you may deal within the future.

Electives

Foundation Stack

AI6104 Mathematics for Artificial Intelligence

This course aims to provide the appropriate mathematical background to students who will study other courses in the Master of Science in Artificial Intelligence programme. Upon completion of this subject, the student should be able to:

- Solve systems of linear equations
- Use linear independence basis to decompose vectors.
- Compute partial derivatives
- Perform approximation based on Taylor series
- Search maximum and minimum in multivariable functions.

AI6120 Python Programming

This course aims to provide appropriate computing background to students who will study other courses in Data Science and AI. Upon completion of this subject, the student should be able to:

- Use the Jupyter Notebook environment for Python.
- Solve standard algorithmic problems in Python.
- Perform standard numeric computations in Python.
- Handle data in Pandas, and graphics in Matplotlib.

SD6124 Probability and Statistics for Data Science

This course provides an introduction to probability and statistics for pursuing higher level courses in data science. In this course, students will learn and get familiar with probability axioms, discrete random variables, continuous random variables, bivariate distributions, marginal and conditional distributions, independence, covariance, correlation, Bayesian inference, regression models, hypothesis testing, weak law of large numbers, central limit theorem, Markov chains and probability transition matrices.

SD6128 Introduction to Economics

This course is designed to provide students with a comprehensive understanding of the core principles of microeconomics and macroeconomics, encompassing the behaviors of individuals, firms, and governments within the economic landscape. Students will learn the foundational theories of economics and acquire the skills needed to employ economic tools for analyzing diverse economic challenges.

SD6129 Introduction to Psychology

This course builds upon Data Science Thinking and provides a comprehensive overview of contemporary psychology. It aims to provide a scientific understanding of the mind, brain, behaviour, and experience, and how these interact with the complex environments in which they exist. It also focuses on developing an understanding of the role of empirical evidence in the creation and constraint of theory. Finally, it aims to develop an understanding of how psychological theory applies to a wide range of real-world questions related to data science.

Exploration Stack

AI6103 Deep Learning and Applications

Deep learning has recently introduced a paradigm shift from human-design features to end-to-end systems, and has revolutionized several fields including computer vision, speech recognition, and natural language processing. Top IT companies like Google, Facebook, Microsoft, Apple, Amazon have been actively redesigned their products with deep learning techniques, and the impacts in the coming decades will go beyond self-driving cars, strategic games like Go, and MRI cancer detection.

The main objective of this course is to introduce the mathematical foundations, the state-of-the-art architectures, and a professional library of deep learning architectures. Students will learn how to design their own artificial neural network to solve their data analysis task. They will also learn how to code efficiently these new algorithms using PyTorch, one of the most powerful libraries in this field.

AI6122 Text Data Management and Processing

This course covers fundamental techniques to manage and process text data. The main topics include:

1) text indexing and search: inverted index, query processing, ranking, and evaluation,
2) word-level, sentence-level, document-level, and collection-level processing: morphological analysis, part-of-speech tagging, parsing, summarization, classification and clustering, and topic modeling, and
3) case studies and applications: social media text, sentiment analysis, and information extraction.

AI6123 Time Series Analysis

Many of the complex systems are dynamic systems in which their states change over time. This course introduces time series models and the corresponding methods for data analysis and inference. Topics include regression models, autoregressive (AR), moving average (MA), ARMA, and ARIMA processes, stationary and non- stationary processes, seasonal processes, identification of models, estimation of parameters, diagnostic checking of fitted models, rare event detection, forecasting, spectral analysis and time series models of heteroscedasticity.

Real world applications for understanding characteristics, modelling and evaluating forecasts of time series data in economics, finance and industries are elaborated with lab on using R.

AI6130 Large Language Models

The course on Large Language Models (LLMs) aims to equip students with a comprehensive understanding of the principles, architectures, and applications of state-of-the-art LLMs like GPT-4. This course is designed for graduate students in computer science, data science, and related fields who have a foundational knowledge of machine learning and artificial intelligence. By taking this course, students will gain valuable skills in developing, fine-tuning, and deploying LLMs, which are increasingly integral to advancements in natural language processing, automated content creation, and AI-driven decision-making. This expertise will not only enhance their academic and research capabilities but also significantly boost their employability in tech industries, research institutions, and innovative startups focused on AI and machine learning technologies.

SD6106 Capstone Project

This project, to be performed over two consecutive semesters, provides students an opportunity to work with faculty members in CCDS and other schools (e.g., NBS, SBS, SPMS, EEE, MAE, SSS, ASE) as well as industry participants to apply Data Science techniques for a particular domain-specific problem. The project is either group-based or individual. It can either be carried out in an industrial setting outside NTU or in an academic/joint lab setting within NTU. Scope of project may be based on interests of the stakeholders. Students are expected to document their work and report their findings in formal reports, and give oral presentations together with demonstration (if any) as the conclusion of their projects.

SD6122 Modern Hardware for Data Science

This course sets out to bridge the gap between cutting-edge hardware advancements and database processing efficiency, aiming to equip students with the knowledge and skills to design and implement data processing algorithms and structures optimized for modern hardware. Targeted at Master-level students with a solid foundation in systems programming and prior exposure to databases and computer architecture, it offers deep insights into leveraging memory hierarchies, parallelizing tasks on multi-core CPUs, and utilizing hardware accelerators like SIMD and GPU for enhanced data processing.

SD6123 Data Privacy in Data Science

Our societies have increasingly become data-driven, creating, capturing, and harnessing more data with time. With that, there is a vastly increasing possibility that some of the data may carry information, which may expose people or entities to serious risks. As such, mechanisms are needed to ensure proper privacy guarantees. To do so, there is foremost a need to define and determine privacy requirements systematically; in addition to creating and applying proper techniques to achieve the requirements.

Consequently, in this course we explore a spectrum of aspects: from regulatory, standardization and governance issues to privacy enhancing technologies and industry best practices, delving on a subset of these most relevant in the data science context. This course is thus relevant for data scientists, data engineers, data, and privacy architects, who, on one hand need to understand the compliance requirements and regulatory landscape, on the other, need to identify and apply suitable techniques to achieve the same in their data science driven applications and products.

SD6125 Data Mining

In the era of big data, large quantities of data are being accumulated. The amount of data collected is said to double every nine months. Seeking knowledge from massive data is one of the most desired attributes of Data Mining.

In general, there is a huge gap from the stored data to the knowledge that could be construed from the data. This transition will not occur automatically, that is where Data Mining comes into picture. In Exploratory Data Analysis, some initial knowledge is known about the data, but Data Mining could help in a more in-depth knowledge about the data. Courses on Database systems give methods to extract information, but they fail to extract knowledge that is actionable.

Manual data analysis has been around for some time now, but it creates a bottleneck for large data analysis. Fast developing computer science and engineering techniques and methodology generates new demands. Data mining techniques are now being applied to all kinds of domains, which are rich in data. Although data mining is partly based on statistical methods, data mining methods give a lot more than the statistical methods. Data mining methods are to a large extent based on machine learning methods. The difference is data mining is meant for huge data whereas machine learning is usually done over relatively small-sized data. Huge data brings completely a new set of problems to be solved. This course aims to introduce you to the exciting and ever-evolving world of data analytics and mining.

SD6126 Scalable Data Systems

An extremely large amount of data is created every day, bringing us to the era of big data. The world of data management has dramatically changed in the “Big Data” era. This is primarily driven by multiple factors including cheaper computing and storage costs, increasing availability of sensors, smart devices, and social media platforms, and stronger cloud computing infrastructure and data systems. Therefore, building scalable data systems is of utter importance toward supporting real-world applications. This course builds upon the Data Systems course and aims to provide abroad understanding of big data and current technologies in managing and processing them. Key topics covered in this course include big data properties, big data and cache conscious designs, distributed processing over big data, MapReduce and Hadoop for big data applications, and NoSQL databases. Upon completion of this course, you will learn to evaluate issues associated with big data management and relevant data analytics in data science.

SD6127 Network Science

Many real-world data in data science applications can be represented as networks. Consequently, it is important understand the models, properties, and analysis techniques associated with such data. Network science is a discipline that investigates the topology and dynamics of such networks, aiming to better understand the behaviour, function and properties of the underlying systems. The key topics covered in this course include network metrics, network properties, network models, querying and analytics on networks, cognitive psychology of network visualization, network dynamics, network robustness, and spreading in networks. The students will be able to apply the insights gained in this course to implement a real-world data science application centered around complex networks.