Course Content

The MSDS programme structure aims to comprehensively cover all major components of a data science ecosystem. There are five **core courses** and a compulsory **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-driventechniques and issues relevant to data science. Lastly, the application domain stack has an unrestrictive elective course which focuses on a specific application domain ofinterest. It provides a student necessary background on the target application domainof his/her capstone project. A student is required to take three courses from these stacks.

The **compulsory *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 a student opportunity to address such problem for non-industrial end users.

Course Details

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

in the data science ecosystem. 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 provides an introductory and broad perspective of machine learning fundamental algorithms, and is relevant for anyone pursuing a career in Data Science. It aims to provide students with the essential concepts and principles of traditional algorithms in machine learning so that students can use various machine learning techniques to solve real-world data science problems. The course also introduces the impact of social science theories such as human bias on machine learning techniques.

SD6103 Data Systems

Constructing large and complex data science pipelines to address needs in a wide variety of domains is a critical challenge in today's world. A critical ingredient for such pipelines is storing and accessing large amounts of data. In fact, this issue is central to any data science pipeline as it is well known that moving data is significantly more expensive than doing any kind of computation. Data systems provide the means to store and analyze a massive amount of data.

This course is a comprehensive introduction to modern centralized data systems. The teaching content of this course include 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 with in the future.

SD6106 Capstone Project

This project, to be performed over two consecutive semesters, provides students an opportunity to work with faculty members in SCSE 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.

AI6120 Python Programming

This course aims to provide an introductory but broad perspective of fundamental computing tools and techniques in Python. It also prepares you to apply the knowledge in Python towards Data Science (DS) applications. In particular, this is a 0 AU course and is designed for students who do not have background knowledge in programming. It gives students an opportunity to pick of Python skills before delving into core data science modules.

Electives

Foundation Stack

AI6104 Mathematics for Artificial Intelligence

This course ensures that students have essential mathematical knowledge to pursue MSDS program. This is a foundation course primarily designed for students who do

not have sufficient mathematical knowledge that is necessary for some data science courses. Specifically, it introduces linear systems, matrices, vector spaces and related concepts, eigenvectors and eigenvalues, multivariable functions, partial derivatives and their applications in data science, and optimization techniques.

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.

Exploration Stack

AI6122 Text Data Management

A large portion of real-world data are in the form of free text. Because of the unstructured nature and the ambiguities in natural languages, it is challenging to manage and sense making from text data. This course covers techniques to support information access from large amount of unstructured text data (e.g., web pages, customer reviews, medical reports, and emails), and also fundamental techniques to mine meaningful information from the text data.

AI6123 Time Series And Prediction

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. It aims at developing basic knowledge and introducing methods used in time series analysis. An important focus of the course is to understand the mathematical root and statistics techniques so that one may be able to apply and generalize them to different real-world data analysis tasks.

AI6103 Deep Learning and Applications

The purpose of this course is to systematically introduce fundamental concepts, theories, and best engineering practices of modern deep learning, so that they can be applied to discover patterns in data, make data-driven predictions, and provide visualization and interpretation of the deep learning models. The students will learn to apply appropriate technologies to address the specific needs of end users while being cognizant of the potential negative impact of them on their psychology and ways to mitigate them. The course is geared toward students who want to pursue professional data science careers based on deep neural networks.

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 a broad 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.

SD6128 Introduction to Economics

Economics impacts behaviours of humans and institutions, which in turn impact data involving them or generated by them. This course builds upon the Data Science Thinking course and aims educate participants on the fundamental concepts of microeconomics and macroeconomics involving individuals, firms, governments and economies. It also introduces analysis of basic economic theories and guide participants to apply economic tools to analyze economic issues in the context of data science problems.

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.

* Students may take one 3AU or two 1.5AU courses from other graduate programmes.