MIE 1624

Group Project Report

Group 9

Baoheng Kuang, Da Da, Ruoling Yang, Sheng Gao, Yaqi Zhang, Yunran Qiu

Contents

Part 1 Course curriculum design 2

Course Description 2

Course Outline 2

Part 2 Data Science program curriculum design 4

Overview 4

Curriculum 4

Course Descriptions 5

Part 3 Visualizations of course curriculum 8

Part 4 Data Science education EdTech effort 10

Part 1 Course curriculum design

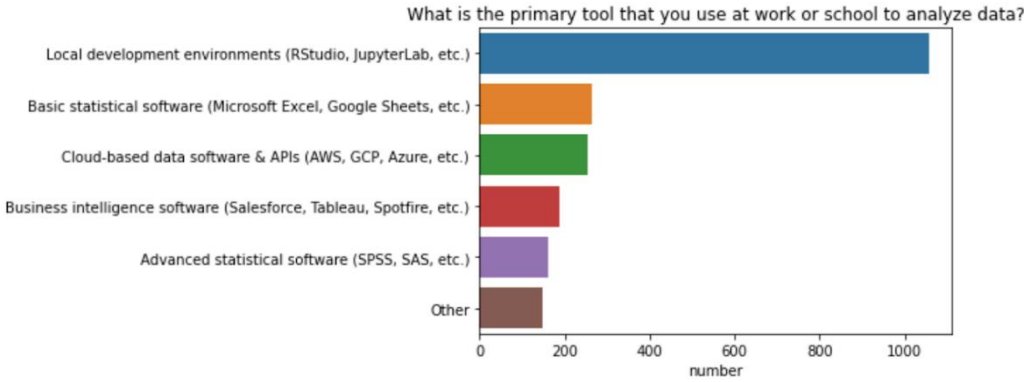
Course Description

Based on some research the team did on some job posting websites, this course targets to teach students to achieve the basic requirement of a data analyst/scientist/manager. Why data science? Data science is about solving business problems: from marketing to product management to finance.

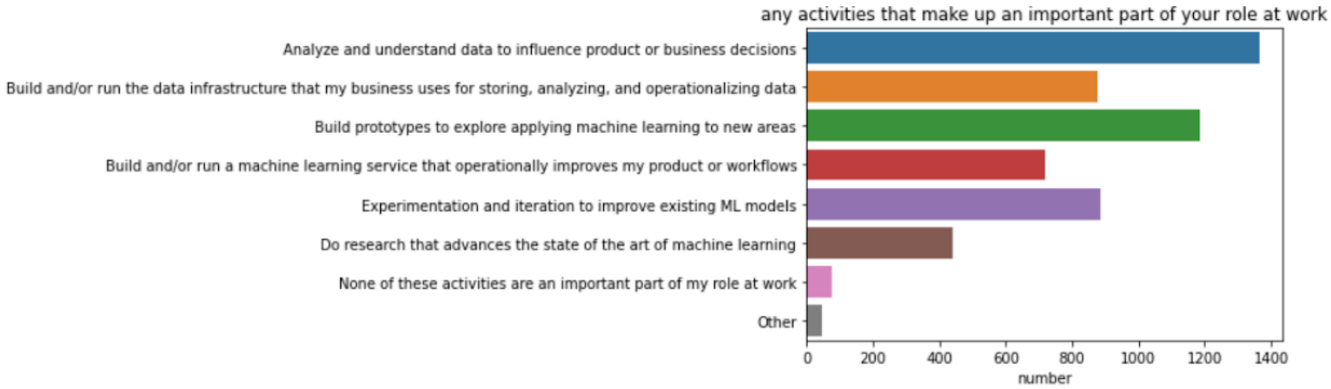
Based on the research, Python is the most popular language that used to analysis, so this course mainly focuses on learning Python analysis skills, algorithms, data visualization, etc.

Course Outline

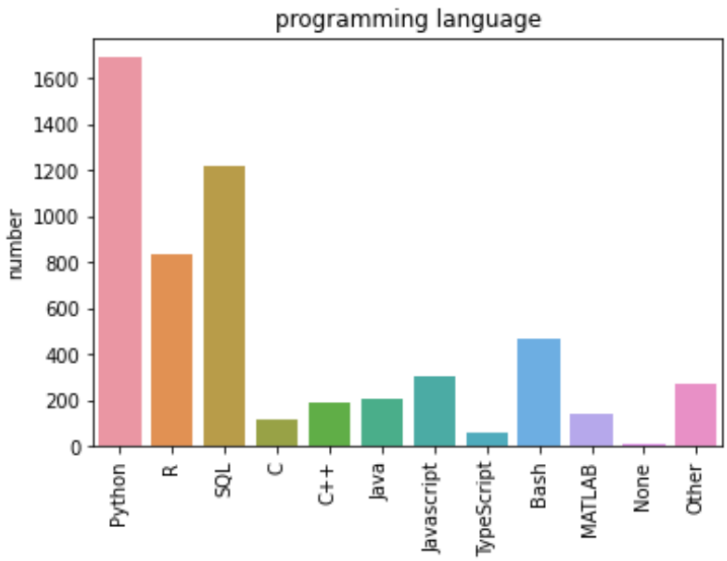
1. **Introduction of data science and analysis**
2. **Introduction of Python and Jupyter notebook**
3. **Mathematic and statistics**
   Linear algebra and calculus
   Probability Distributions and Sampling in statistics
   Other statistics techniques
4. **Exploratory Data Analysis and Visualization**
   Pandas Python package
5. **Machine learning model implementation**
   Supervised/Unsupervised Learning (Linear Regression and Logistic Regression, K-Means)
   Feature Selection, Engineering, and Data Pipelines (PCA/ICA, regularization)
   Advanced Supervised/Unsupervised Learning (SVM, decision trees, and random forest for regression and classification, DBSCAN)
   Advanced Model Evaluation and Data Pipelines (cross-validation and bootstrapping)
6. **Machine learning skills**
   Predictions, estimates and trends from business data
   Ability to visually identify objects
   Detection of unusual data
   Apply ML to new area
   Improve workflow



a. What is the primary tool you use for data analysis?



b. Which activities play an important role in your work?



c. Which programming language is used most in your job?

Figure 1 Research results

Part 2 Data Science program curriculum design

Overview

The Master of Data Science and Artificial Intelligence program provides students advanced professional skills in theory and practice to handle large-scale data management and analysis challenges that arise in today's data-driven organizations.

Curriculum

Table 1 Curriculum for Master of Data Science and Artificial Intelligence program*

	Course Name	Course Type	Credit
Semester 1	Introduction to Machine Learning and Data Analysis I	Core	0.5
	Artificial Intelligence: Principles and Applications	Core	0.5
	Communication and Business Organization	Core	0.5
	Introduction to SQL	Core	0.5
Semester 2	Introduction to Machine Learning and Data Analysis II	Core	0.5
	Fundamental Statistics in Computer Science	Core (Data Scientist)	0.5
	Natural Language Processing	Core (Data Scientist)	0.5
	Introduction to Data Engineering	Core (Data Engineer)	0.5
	Project management	Core (Data Engineer)	0.5
Semester 3	Quantitative Analysis in Finance	Elective	0.5
	Biostatistics and healthcare engineering	Elective	0.5
	Deep Learning and Computer Vision	Elective	0.5
	Decision making in marketing	Elective	0.5
	Advanced Data Visualization and Reporting	Elective	0.5
Semester 4	Graduate Capstone	Core (Pick one of two)	3
	Internship		3

* The technology priority for most courses is Python.

Course Descriptions

Semester 1

- **Introduction to Machine Learning and Data Analysis I**

A foundations course in data science, emphasizing both concepts and techniques. This course will introduce the key elements of machine learning and data analysis. Most highly used machine learning and data analysis techniques are covered in this course, including classification, clustering, association analysis and anomaly detection. Students will complete a series of programming assignments involving broad fields of data science and its various application areas.

- **Artificial Intelligence: Principles and Applications**

Artificial intelligence is the math that helps make the right decisions with incomplete information and limited computation. It has become an increasingly influential science and technology discipline. This course provides students an introduction of basic AI. The evolution of AI and its technologies will be taught. Students will explore why it is important to have an AI strategy, and deduce how to gain strategic advantage using different kinds of intelligence. (Topics may include machine learning, probabilistic reasoning, robotics, computer vision, and natural language processing.)

- **Communication and Business Organization**

Expected learning outcomes include that, in the context of data science and analytics, students should be able to: summarize, report, organize prose, statistics, graphics, and presentations; explain uncertainty, sensitivity/robustness, limitations; describe model generation and representation; discuss interpretations and implications; communicate effectively to diverse audiences within a business organization, and possibly other outcomes.

- **Introduction to SQL**

This course aims to introduce basic SQL programming skills. Topics include relational database fundamentals, database design techniques, and some common administrative skills. Upon completion of this course, students will have a thorough understanding of SQL functions, join techniques, database objects and constraints, and gain the skills of writing useful SELECT, INSERT, UPDATE and DELETE statements. After this course, students will be able to write efficient SQL queries to successfully handle a variety of data analysis tasks.

Semester 2

- **Introduction to Machine Learning and Data Analysis II**

This is an advanced course for students who have finished Introduction to Machine Learning and Data Analysis I. The course has four modules: machine learning for business, machine learning for finance, machine learning for marketing and designing machine learning workflows. Module 1 (Machine learning for business) will focus on the key insights and base practices how to evaluate the appropriateness of a business application for machine learning.

Module 2 (Machine learning for finance) will be conducted by predicting stock data values using linear models, decision trees, random forests, and neural networks. In Module 3 (Machine learning for marketing), students will learn how to use different techniques to predict customer churn and build customer segments based on their product purchase patterns. The last module (designing machine learning workflows) aims to introduce how to build data science pipeline including overview of machine learning procession, digging deep into the cutting edge of sklearn, and dealing with real-life datasets from various areas.

- **Fundamental statistics in computer science**

The statistical tools of modern data analysis can be used in a range of industries to help us guide social and scientific progress. Tools and techniques to achieve this goal will be introduced in this course. Topics covered will include continuous and discrete distributions, descriptive statistics, hypothesis testing, confidence intervals, regression, one-way analysis of variance, and so on. The priority technology of this course is R.

- **Natural Language Processing**

Natural language processing (NLP) is a crucial part of artificial intelligence (AI), modeling how people share information. This course aims to provide a solid understanding of NLP and a broad introduction of its applications in current technologies. In this course, computational properties of natural languages and cutting-edge natural language processing techniques will be taught. Students will also learn how to evaluate the appropriateness of a business application for natural language processing.

- **Introduction to Data Engineering**

In this course, we will explore the primary differences between a data engineer and a data scientist. Students will learn different types of databases data engineers use and get a thorough understanding of how cloud technology plays a role in data engineering. This course is designed for people who want to work at a company with several data sources and don't have a clear idea of how to use all these data sources in a scalable way.

- **Deep Learning and Computer Vision**

This course introduces the constructions and applications of deep neural networks. The focus of this course is an in-depth understanding of the main features of deep learning structure and its application in computer vision. Specific contents include how to reduce the computational cost of training neural nets, research on convolutional neural networks and discussion of various coding tools. The techniques introduced in this course will be illustrated through specific application examples about computer vision. Students will gain the skills of building deep learning models for computer vision and learn about the Specify-Compile-Fit workflow to make predictions and have all the tools necessary to build their deep learning models.

- **Project management**

This course introduces students to the problems and issues in managing large sets of data, focusing on modeling, storing, searching, and transforming large collections of data for analysis. The course will cover database management and information retrieval systems, including relational database systems, massively parallel/distributed computation models and

various NoSQL systems that are designed to handle extremely large-scale and complex data collections. Emphasis is placed on the application of large-scale data management techniques to domains. Programming projects are required.

Semester 3

- **Quantitative Analysis in Finance**

This course introduces some quantitative methods in finance. There are three main topics: asset pricing, non-linear optimization, and financial statement analysis. The emphasis is to explore the development of the key techniques and their application to practical problems in finance.

- **Biostatistics and healthcare engineering**

This course aims the application of biostatistical techniques in the field of healthcare. Common strategic, tactical, and operational decision-making problems arising in healthcare will be approached from a machine learning perspective. Real-world datasets will be provided to illustrate the complexity of applying data science methods to healthcare.

- **Decision making in marketing**

This course covers practical problems in marketing analytics, it provides several quantitative methods to help the decision maker to make strategies in the offline and online marketing campaign and commercials. The objective of the course is to give students a general understanding in how data science delivers to marketing.

- **Advanced Data Visualization and Reporting**

This course discusses business intelligence, reporting skills and popular data visualization software (Tableau, Power BI) in order to discover greater insights within huge datasets. By the end of the course, you'll understand the unique features of each visualization, as well as when to best use them while creating best practice, interactive dashboards and rich reports for complete data insight with higher creativity and quality.

Semester 4

Students have two options in their fourth semester:

- **Graduate Capstone**

This non-class-based experience provides the student with an individual opportunity to explore a project that advances knowledge in an area of data science. Students need to select a problem, conduct background research, develop a research approach, analyze the results, and build a professional report and presentation.

- **Internship**

Students are encouraged to find an internship in their final semester. This internship can be transferred into 6 credits.

Part 3 Visualizations of course curriculum

To help the potential applicants of the Master of Data Science and Artificial Intelligence program clearly grasp program curriculum, program structure, acquired skills, etc., a serious of visualization is designed.

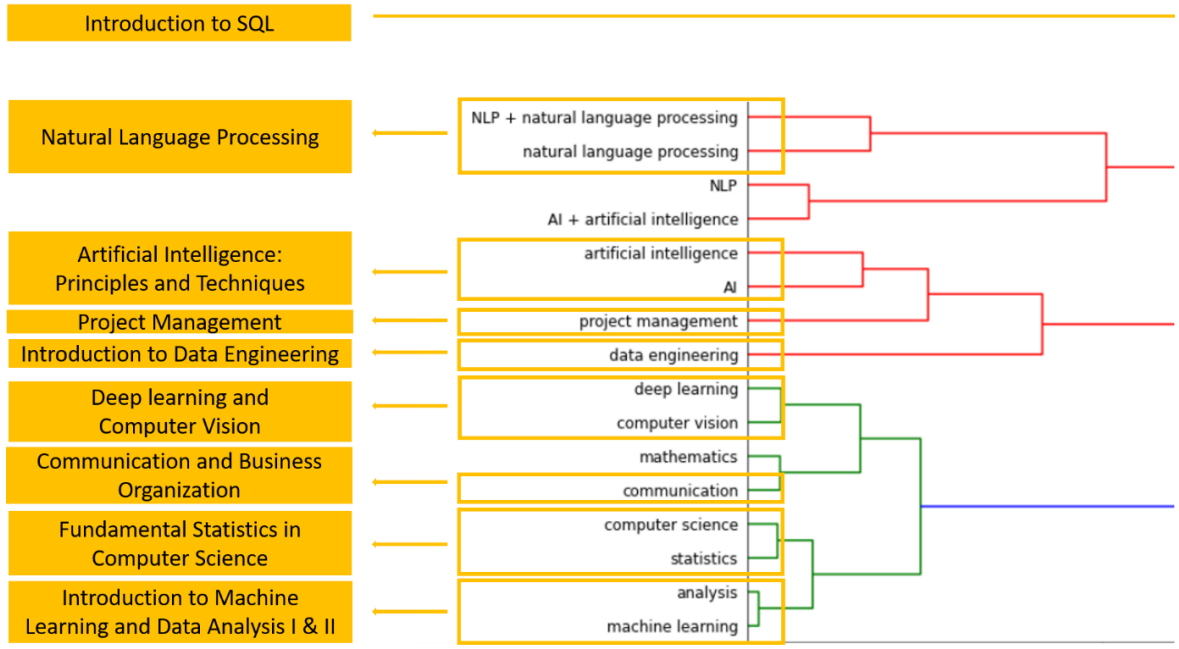


Figure 2 Inspiration for course designing

We designed a series of core courses for the program based on the datasets from Part 1 (Figure 2 shows the basis of our design) and some similar programs our team collect. To identify skills that core course is teaching, a radar chart (Figure 4) is plotted to show the focus of skills for each course. According to the data analysis results in figure 3, in the second semester, we provide different courses for students with different career plan.

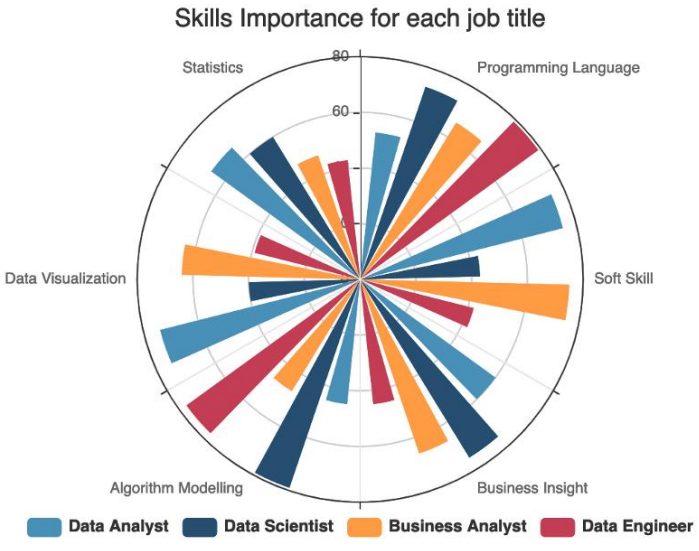


Figure 3 Skills importance for different job

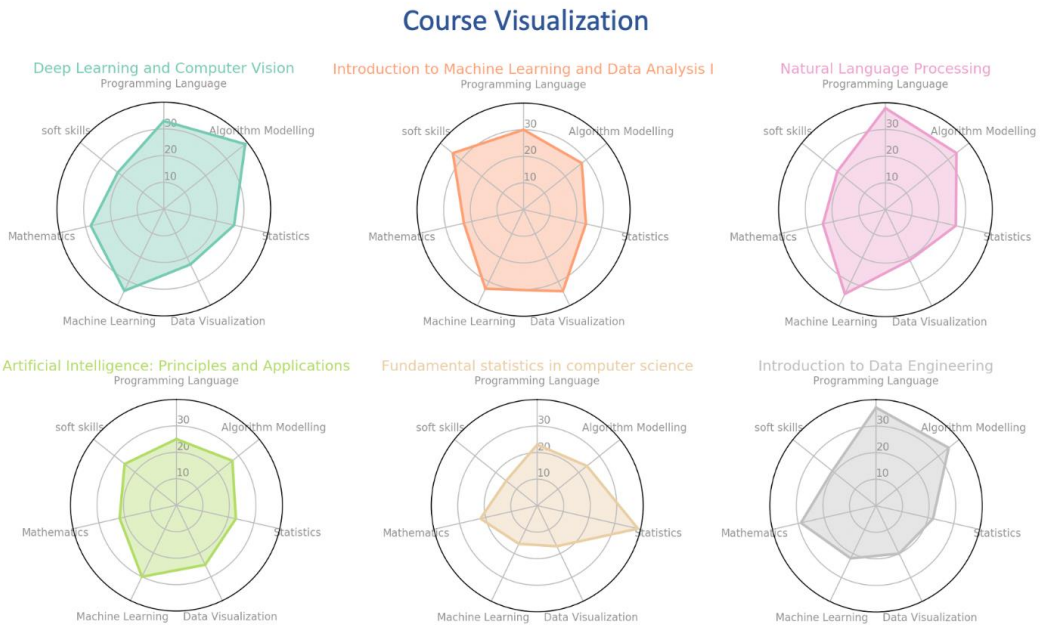


Figure 4 Core course visualization

Considering this is a career-oriented program, five elective courses will be offered in the third semester. Those courses have a clear professional direction and will be very useful for students' future employment. Besides, in the final semester, students are required to finish a final project with professional report and presentation or have an internship that related to data science and artificial intelligence.

Figure 5 shows the courses sequences in this master program.

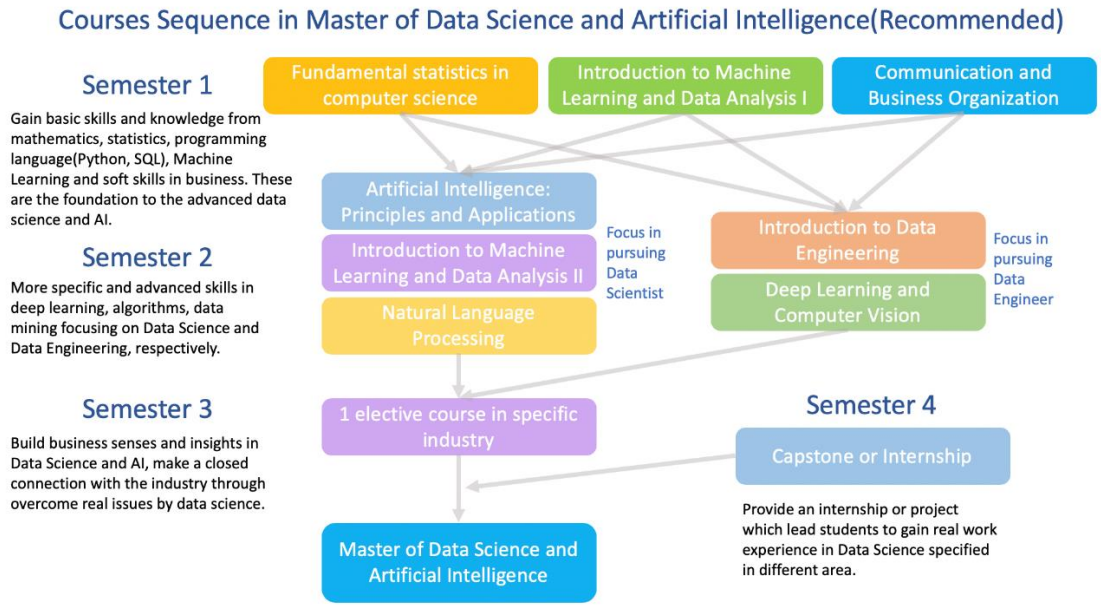


Figure 5 Courses sequence in Master of Data Science and Artificial Intelligence

Part 4 Data Science education EdTech effort

In this part we will design an online course delivery system that automatically adapts to learning patterns of an individual student and automatically adjusts a sequence of slides in online course for her/him.

Choose 3 topics and 15 courses

```
[ ]  1 UI(#From the above topics, please enter what you are interested in (in the list format)
     2
     3   ['data science','machine learning','python'],
     4
     5   #Please enter how many courses you want to take:
     6
     7   15
     8
     9   )
```

```
['IBM Data Science',
 'SAP Data Intelligence for Enterprise AI',
 'Applied Social Network Analysis in Python',
 'Explorez vos données avec des algorithmes non supervisés',
"Évaluez et améliorez les performances d'un modèle de machine learning",
 'Foundations of Data Science',
 'IBM Data Science',
 'Data Science and Machine Learning Capstone Project',
 'Data Science Interview Prep',
 'Developing Intelligent Apps and Bots',
 'Data Analysis and Interpretation',
 'Applied Data Science with Python',
 'How to Win a Data Science Competition: Learn from Top Kagglers',
 'Text Mining & Analytics',
 'Programming with Python for Data Science']
```

Figure 6 course delivery system

Firstly, we allow the users to choose multiple topics they are interested in and each recommended course is likely to cover all the topics. For example, if we search for "statistics" and "machine learning", the top ranked recommendation is "Data Analysis for social Scientist". The course topic doesn't seem like related but if we open the course website (https://www.classcentral.com/course/edx-data-analysis-for-social-scientists-6842) and read through the descriptions, we will find out it covers both machine learning and statistics. Similarly, for "IBM Data Science" (https://www.classcentral.com/course/ibm-data-science-18394), it covers all of data science, python and machine learning. The second key feature is that our system does not make recommendations based on the course titles, but highly depend on the course descriptions. For example, if we enter "nlp", the returned courses will include like "Text Retrieval", "Text Mining", "Computational Social Science" etc.