

---

# *MIE 1624 EDUCATION ANALYTICS CONSULTING REPORT*



*Arun Shanmugam*

**University of Toronto  
MIE 1624 Introduction to Data Science and Analytics  
2020-03-30**

Neil Juan  
Tyler Rankin  
Lei Lei  
Ke Ren  
Raghavendran Thiruvengadam  
Arun Shanmugam

## Table of Contents

EXECUTIVE SUMMARY .....	II
EXECUTIVE SUMMARY .....	3
INTRODUCTION .....	1
MOTIVATION .....	2
APPROACH .....	2
MIE 1624: INTRO TO DATA SCIENCE & ANALYTICS COURSE REDESIGN .....	4
MASTER'S IN DATA SCIENCE AND ARTIFICIAL INTELLIGENCE PROGRAM DESIGN .....	6
RECOMMENDER SYSTEM .....	7
CONCLUSION .....	9
REFERENCE .....	10

## Executive Summary

Data science and Machine Learning have quickly risen to prominence as one of the most sought-after professions, with Harvard Business Review famously calling the role of data scientist the ‘sexiest job of the 21<sup>st</sup> century.’ The meteoric buzz around the prospects of a career in Data Science, Machine Learning and AI has given rise to the need for a holistic business-oriented Data Science and Machine Learning course that would be targeted at professionals and students from a wide range of backgrounds. This report outlines the approach and decisions involved in the design of such a course.

The key insights gleaned from the design of the Data Science course inspired the design of a Master of Data Science and Artificial Intelligence program with six mandatory courses as the pillars on which the program is built on. All course materials will start from *ab initio* fundamentals and will help students learn the necessary concepts through practical examples & hands-on assignments. The program is geared towards producing data scientists and ML & AI engineers that are industry-ready the day they step out of campus.

Analytics and visualizations were used to support key course design considerations to ensure that the learning experience matched industry experience and benchmarks in the data science community. Assignments and projects were tailored to provide opportunities for “learning-by-doing” and inculcate aspects of teamwork and leadership as students work in collaborative settings on complex problems to come up with innovative solutions. The program will also place a consistent emphasis on presenting recommendations and making persuasive business cases, so that students can hone their public speaking and communication skills.

All too often, students encounter difficulties when it comes to picking electives since they may not know enough about the course contents or the career prospects that come with it. The start-up’s proprietary recommender system seeks to remediate this problem by recommending courses that will put them on track to securing the job of their dreams.

## Introduction

The past two decades have ushered in an era of technological advancements occurring at an unprecedented velocity in both the manufacturing and services industries powered largely by the wave of automation and Artificial Intelligence (AI). This period of technological revolution has now popularly come to be referred to as the Fourth Industrial Revolution, and central to this period of exponential growth are breakthrough developments in Data Science and Machine Learning [1]. Data is indeed the oil of the 21st century, as evidenced by the massive windfalls in companies such as Google, Facebook and Amazon that have built their businesses on monopolizing data [2].

Businesses want to remain competitive in this period as competitors and industry frontrunners proactively adopt data science and Machine Learning into core verticals (see Figure 1). They are now actively looking at opportunities to collect and put to good use data from several of their systems and processes. As a consequence, they now require consummate engineers who can handle, analyze and derive critical insights from large amounts of data and provide actionable recommendations to top management on business processes and systems. Wharton School of Business reports that the past three years has seen a massive 20x growth in data science jobs in the fields of education, marketing and marketing [3]. This signals the dawn of a new era in the education of Analytics, Data Science and Machine Learning.

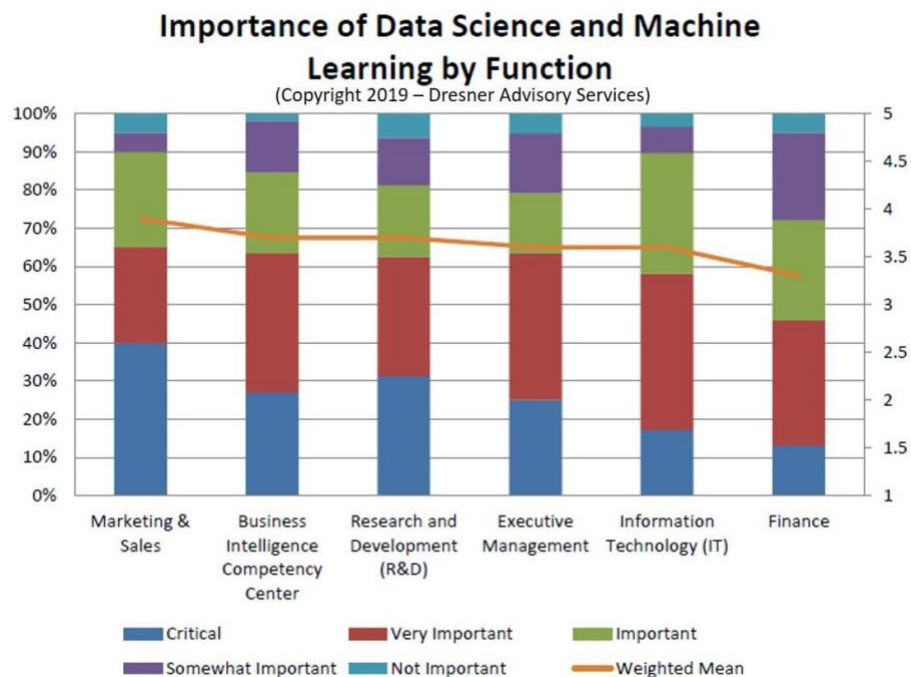


Figure 1: "State Of AI And Machine Learning In 2019", Forbes [4]

## Motivation

The push towards Data Science and Machine Learning has sparked a rise in demand for educational resources that will enable students and professionals alike to hop aboard the data revolution bus (see Figure 2). The target audience can be broadly classified into two large groups: graduates fresh out of school who are actively considering careers in Data Science and professionals who are domain experts in their respective fields but are now looking to incorporate Machine Learning into their business domains or those who want to pivot to a career in Machine Learning.

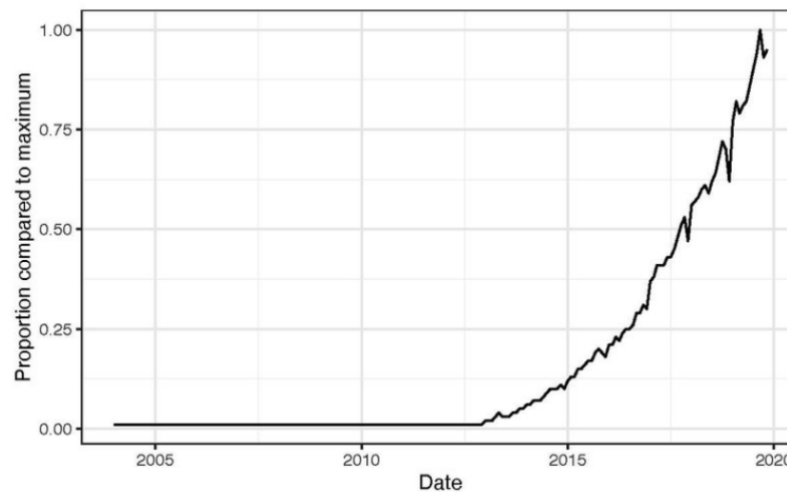


Figure 1. Google Trends monthly data for the term "Data Science" as of November 30, 2019. The y-axis is the proportions of searches compared to the maximum, which was on September 2019.

Figure 2: "The Role of Academia in Data Science Education", Harvard Data Science Review

The need of the hour is clear; an accessible curriculum that provides a good mix of data science, Machine Learning and business acumen. The curriculum must be flexible and take into consideration the wide range of experiences and backgrounds of the participants. The courses offered must tick the following boxes: introduce all core concepts from scratch, help all participants scale a steep learning curve, provide opportunities for application-based learning and most importantly, enable participants to make actionable and practical business recommendations based on insights gleaned from data. The rest of this report will outline the design of such a holistic curriculum.

## Approach

The effort to design a new course curriculum for MIE1624: Intro to Data Science and Analytics was carried out with a consistent emphasis on industry-focussed career requirements, and the knowledge and skills expected of a budding data scientist in Canada. The data required for the course design process was obtained through web scraping from job boards such as Indeed, Glassdoor and Zip Recruiter. The scraped job descriptions would then provide a deep repository

of information regarding the skills and the expectations that potential employers who are hiring for data science are looking for.

Since there was an abundance of text data to be analyzed, Natural Language Processing techniques were used. Firstly, the text was pre-processed by removing html tags, punctuations, special characters and other non-essential stop words to distil the cleaned text. The next step involved determining Word Frequency of each skill to generate a matrix of skills, which was then subjected to unsupervised hierarchical clustering to generate a dendrogram (see Appendix A). This dendrogram was then used as a guide to inform course design since ultimately, human judgement and decision-making is an essential part of any good design.

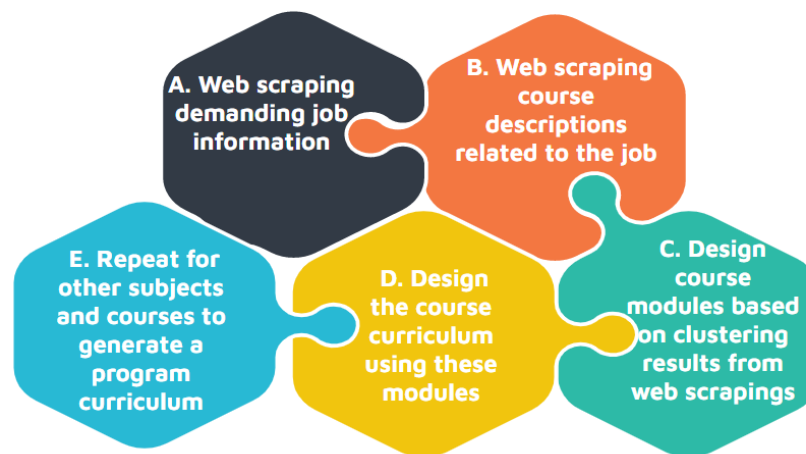


Figure 3: Graphic detailing approach taken to course design

This was followed by the design of the Master of Data Science and Artificial Intelligence program. The input for this process was obtained by the web scraping of online course aggregation & review websites such as Coursera, Course Central & Course Report. These websites aggregate and display courses from a plethora of leading schools, bootcamps and institutions along with user ratings for each course. The program was broken down into six pillar courses: Introduction to Programming, Financial Engineering, Statistics, Intro to Data Science and Analytics, Advanced Machine Learning and Deep Learning and AI in Finance. The approach taken to design the individual courses was similar to MIE1624 redesign with dendrograms guiding clustering of course concepts into modules.

The 2019 Kaggle Survey data also proved to be an invaluable source of information regarding current trends & preferences amongst data scientists in the community and this aided in the construction of concepts that were relevant and topical [5]. Data visualization was primarily but not exclusively carried out using the Kaggle survey data.

## MIE 1624: Intro to Data Science & Analytics Course Redesign

The key factors that influenced the course redesign of MIE 1624: Intro to Data Science and Analytics were the demand for the skills that were to be taught and the employability of candidates with those skills. From the Kaggle dataset, it was determined that the average salary of the employees who were proficient in Machine Learning was a whopping 220% higher than those who didn't know Machine Learning. This justifies that there is a significant market pull for Machine Learning and the skills associated with it. So, a decision was made to include three modules on Machine Learning. The splitting of the course content across three modules allows for the exploration of key concepts such as Classification, Bias Variance Tradeoff and Feature Engineering in greater depth and detail.

Another key departure from the existing course is the emphasis on data management and distributed cloud computing. Most employers use variants of SQL for their data management needs and this was the key reason behind inclusion of SQL in the course curriculum. Cloud solutions such as AWS, Hadoop and Spark have been shown to have higher market value than those who possess cloud computing skills (by almost 80%) and this is something that is definitely worth learning through the course (see Figure 5).

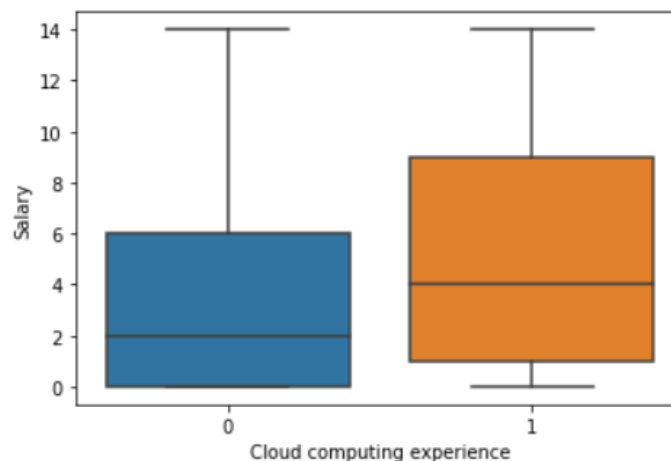


Figure 5: Cloud computing makes a difference in market value



## Introduction to Data Science

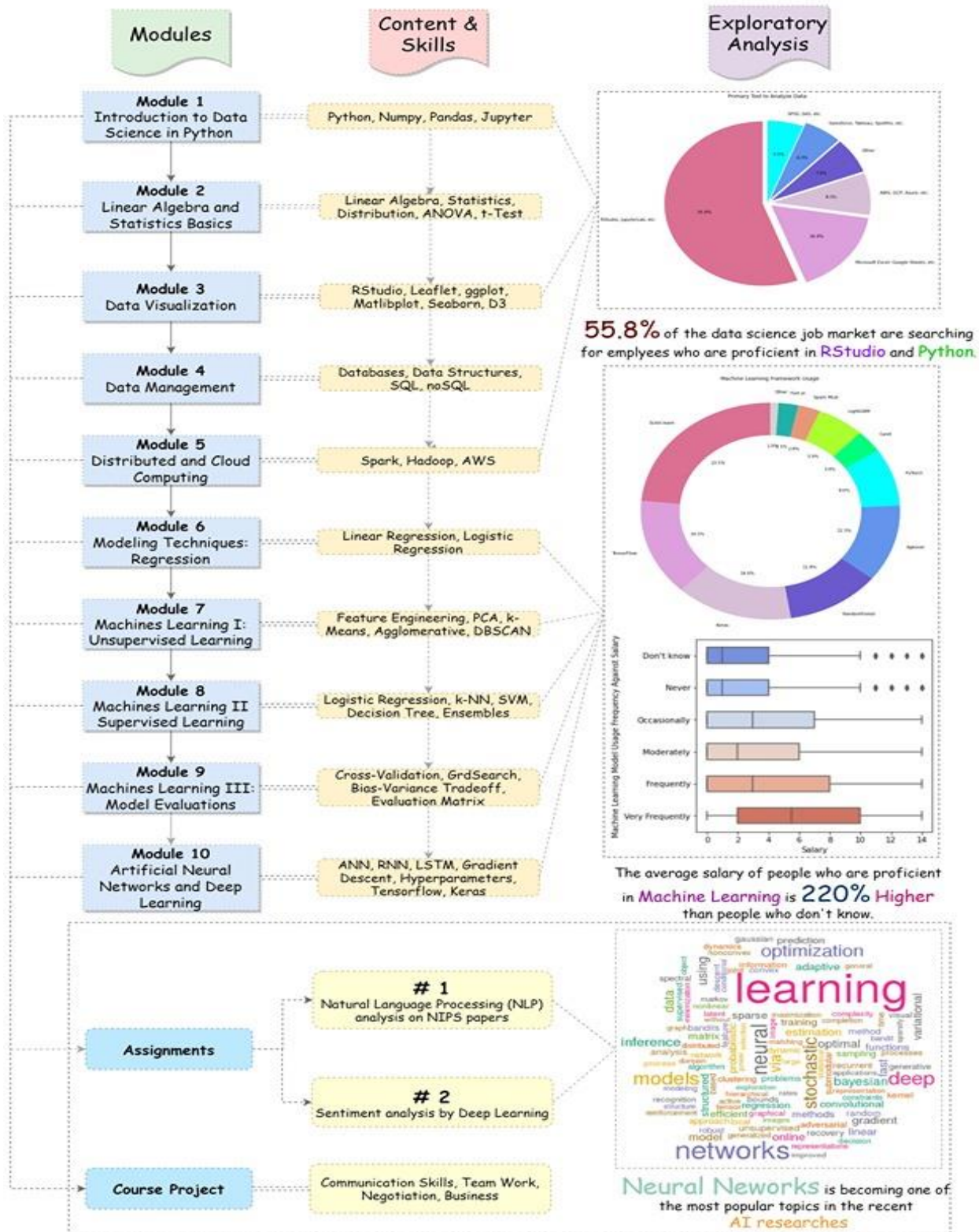


Figure6: Course design for MIE1624: Introduction to Data Science & Analytics



Another key design observation is the inclusion of two NLP-based assignments. Natural Language Processing has quickly grown to become one of the hottest applications in Data Science and Machine Learning due to the practical & ubiquitous nature of use-cases from automated chatbots (speech prediction), autocorrect (text prediction) and language translation. The assignments will allow students to work on practical problems to get acquainted with the data science work ethic from data wrangling, cleaning and visualization to model implementation and interpretation of results.

The course project will typically allow student teams of five to six members work on a relevant or on-going problem in the data science community and this will provide the students to get creative and provide innovative data-driven solutions. Through the assignments and project, students can “learn by doing” and then present their findings through a presentation which will provide them the opportunity to hone their communication skills and confidence which are crucial when it comes to developing business acumen.

## Master’s in Data Science and Artificial Intelligence Program Design

The design of the course was the first step in creating a holistic business-oriented Master’s program. This program is targeted at individuals who are looking at forging a career in data science and Artificial Intelligence and will not assume any background knowledge of the students. All course materials will start from *ab initio* fundamentals and will help students learn the necessary concepts through practical examples & hands-on assignments. The program is built on the following six courses which will serve as the pillars of the program (see Figure 7 and Appendix for course visualizations).



Figure 7: Program Design

The Python programming course is aimed at providing a solid conceptual framework for understanding any Object-Oriented Programming language but with a special spotlight on Python. The Intro to Finance course will seek to make students financially literate and understand the language of entrepreneurs, businessmen and executives. Statistics is the engine that powers data science fundamentals and it is crucial that students have a strong understanding of this course. The Advanced Machine Learning course will allow students to explore Deep Learning and Neural Networks and will serve as a great bridge course to AI. The AI in Finance will enable students to learn and perform crucial operations such as portfolio optimization, market analysis and risk assessment.

The program is geared towards producing data scientists and ML & AI engineers that are industry-ready the day they step out of campus. Towards this, the program's courses are designed with open-ended projects and assignments that will require the students to go beyond tried and tested means to come up with innovative solutions that they would be encouraged to pursue as start-ups. The teams with the best ideas and solutions would be put in touch with the University's entrepreneurship hatchery to get their idea off the ground and put them on course to becoming the next "unicorn" in the tech space.

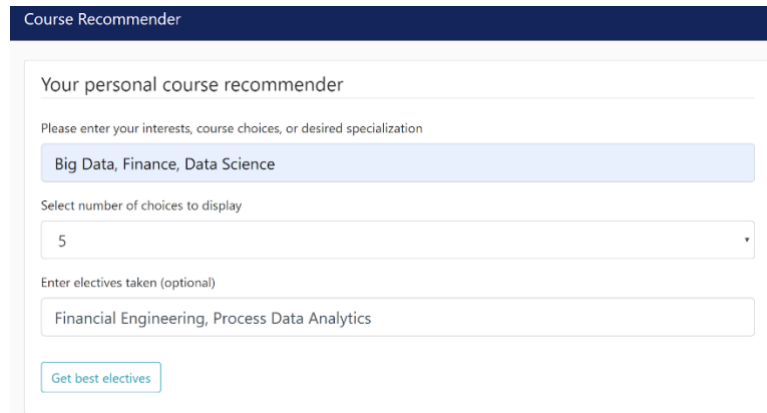
The "pillar" courses are mandatory and will typically receive the highest possible credits. All students are expected to know the nuances in these courses so that they can pick up skills that make them well-rounded data scientists and enhance their chances of employability in the market. The students will have to choose electives from a list of available courses to meet their credit requirements. The list of electives is exhaustive and includes not only technical courses, but courses geared towards honing soft skills of the students such as communication, public speaking, teamwork and leadership. The program will also offer access to established University resources such as Troost I-LEAD that will provide additional help in honing these soft skills wherever necessary.

## Recommender System

It is understandable that choice of electives can be quite a harrowing task when students don't quite know what they are signing up for or what potential career opportunities that their choices may take them closer to or take them further away from. This is a point of contention for many students in schools across the globe and is the motivation behind the start-up's proprietary revolutionary recommender system.

The recommender system takes in the user's interests, career goal, course name, or a combination of each. Then, they will input the number of electives they wish to pick, and specify if any past electives have already been taken. The recommender system will then produces a set of courses as an output. These courses are tailored to the student's needs and will put them on track to securing the job of their dreams. The interface of the recommender system is minimalistic and straight-forward and has been designed as a web-based application and it was

coded on Flask - a Python-based web framework. The minimalistic nature of the UI makes it easy to use for students of all ages, and computer skills. Fig. 8 depicts the screen that the user will interact with. The first text bar will specify the interests, course choices, specialization, and/or dream job that the user is looking for. The next is a dropdown tab, where the student can choose the number of elective courses they want to enrol in. Finally, the last text bar specifies the previous electives already taken if any (electives separated by commas).



The interface is titled "Course Recommender" and contains a form for "Your personal course recommender". It includes a text input field for interests (containing "Big Data, Finance, Data Science"), a dropdown menu for the number of choices to display (set to 5), and another text input field for elective courses taken (containing "Financial Engineering, Process Data Analytics"). A "Get best electives" button is at the bottom.

Fig 8: Course recommender system interface

The course recommender system works on the principle of cosine similarity. Cosine similarity provides a means of measuring the similarity of course descriptions regardless of the length of the descriptions. This is done by vectorizing the course terms and calculating the angle between the word frequency factors using the cosine function rather than the Euclidean distance between the word frequency vectors. This ensures that the courses recommended to the students follow a clear overall narrative that will enable them to leverage the right strengths and skills through the course of their education and get the most “bang for their buck” out of the hefty tuition fees that professional graduate programs usually come with.

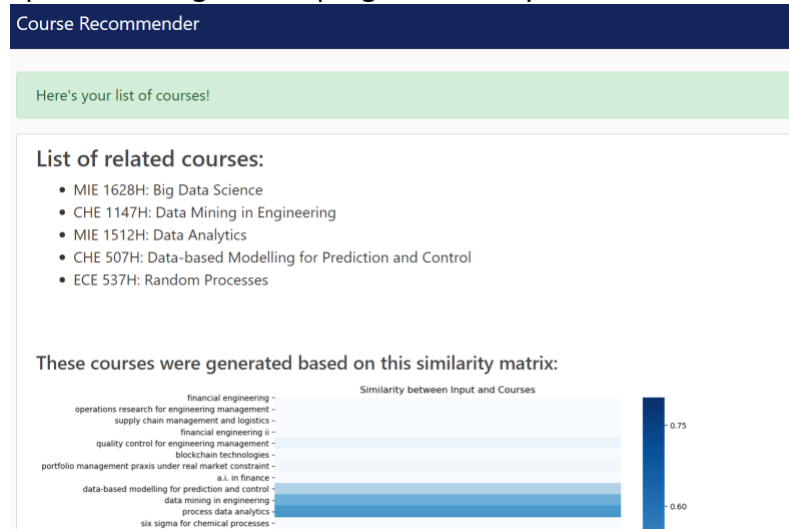


Fig 9: Recommender provides output based on cosine similarity

## Business Overview

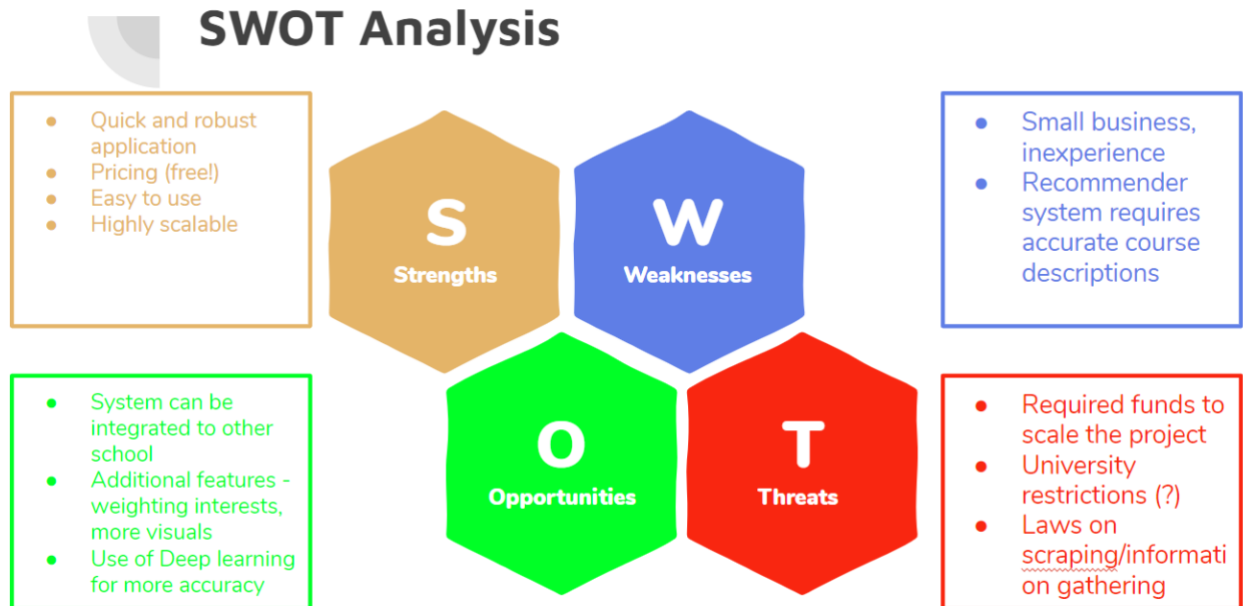


Fig 10: SWOT Analysis

An overview of the business is presented in Fig. 10. The business offers the 'Course Recommender' application to students looking for elective courses at the University of Toronto (UofT). Providing a quick product with easy installation and an intuitive design, the product can be used in all backgrounds of technical skills, and age. The program is also highly scalable. As of now, this recommender system only works for the MEng. Program with an emphasis on analytics, but it could easily be integrated to many other programs within UofT.

The most evident limitation of the business is its small scale. Right now, the application only targets a small user group, and would require more time to make it universal. Moreover, the system algorithm will only be as good as the course descriptions provided by the university. If the description is inadequate, the program might not recommend that as much. This needs to be improved at the university level.

As a new business, there are plenty of opportunities available. As mentioned, this application can be scaled on multiple programs and across several universities. As long as course descriptions are available for use, the algorithm would be able to recommend any course based on the users key words. Further personalization of user input could can be explored as well. The use of weights allows the user to place more importance on one term or another. However, if students are unsure, leaving it without weights would be possible as well. Furthermore, investigating deep learning and selecting an optimal model could be another opportunity to improve precision in the program.

## Conclusion

The design of the course and the program were carried out with the end user (namely the student) at the heart of the process. All design considerations were made based on insights obtained from data to enhance the student learning experience and ensure that they can craft the career of their dreams upon completion of their programs. The proprietary web-based recommender system of the start-up stands testament to the power of Machine Learning tools in providing elegant solutions to real and present problems.

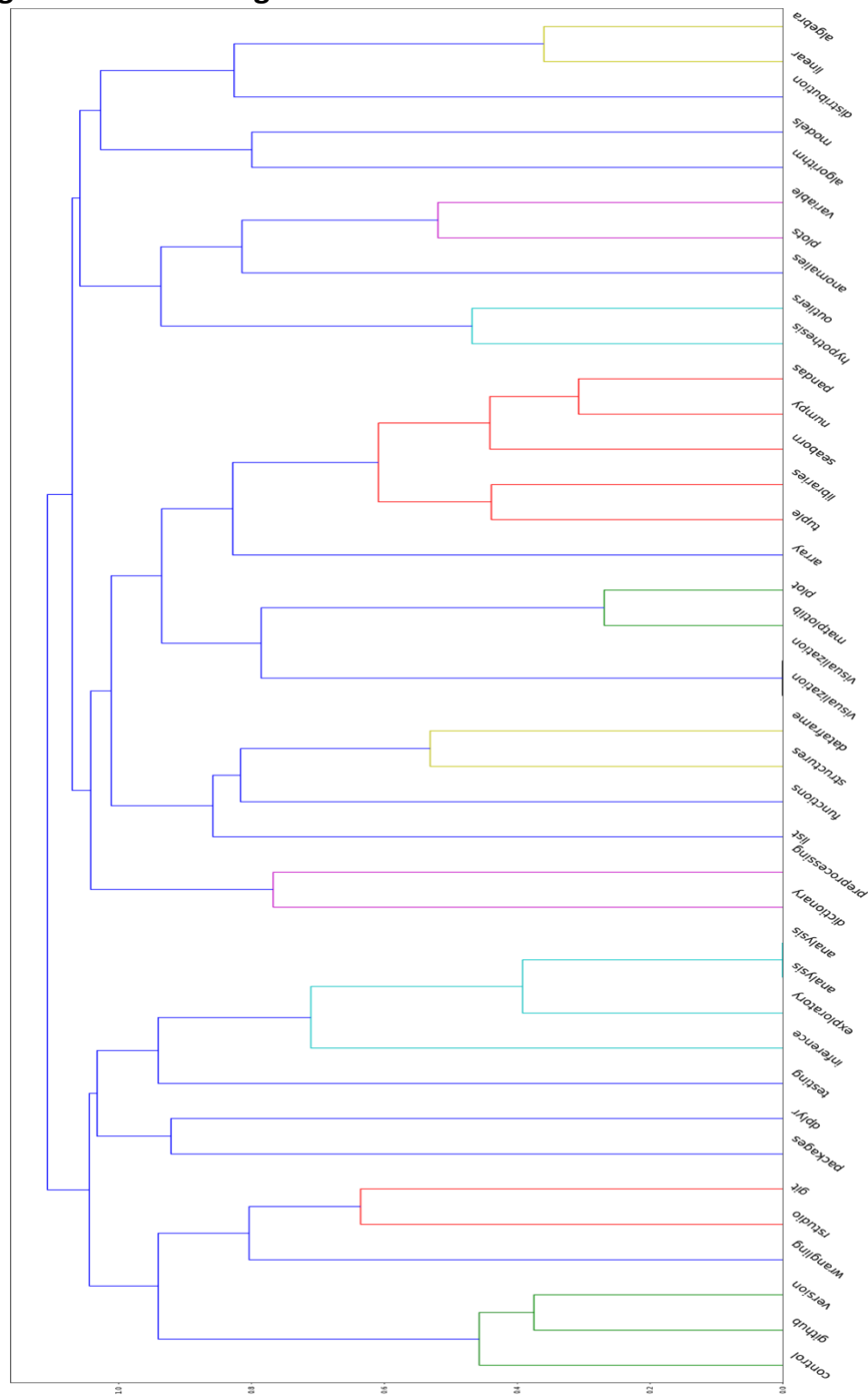
The program will also work to cultivate the entrepreneurial mindset in the minds of the students. Initiatives with the hatchery and I-LEAD will seek to provide a holistic approach to Data Science and Machine Learning. The spirit of the program lies in ultimately enabling students to make persuasive business cases based on data-driven recommendations that can be adopted by top management executives.

## References

- [1] K. Schwab, "The Fourth Industrial Revolution: what it means and how to respond", *World Economic Forum*, 2020. [Online]. Available: <https://www.weforum.org/agenda/2016/01/the-fourth-industrial-revolution-what-it-means-and-how-to-respond/>. [Accessed: 29- Mar- 2020].
- [2] K. Bhageshpur, "Council Post: Data Is The New Oil -- And That's A Good Thing", *Forbes*, 2020. [Online]. Available: <https://www.forbes.com/sites/forbestechcouncil/2019/11/15/data-is-the-new-oil-and-thats-a-good-thing/#2369be537304>. [Accessed: 29- Mar- 2020].
- [3] "What's Driving the Demand for Data Scientists? - Knowledge@Wharton", *Knowledge@Wharton*, 2020. [Online]. Available: <https://knowledge.wharton.upenn.edu/article/whats-driving-demand-data-scientist/>. [Accessed: 29- Mar- 2020].
- [4] L. Columbus, "State Of AI And Machine Learning In 2019", *Forbes*, 2020. [Online]. Available: <https://www.forbes.com/sites/louiscolumbus/2019/09/08/state-of-ai-and-machine-learning-in-2019/#7ac6c8961a8d>. [Accessed: 29- Mar- 2020].
- [5] "2019 Kaggle ML & DS Survey", *Kaggle.com*, 2020. [Online]. Available: <https://www.kaggle.com/c/kaggle-survey-2019>. [Accessed: 30- Mar- 2020].

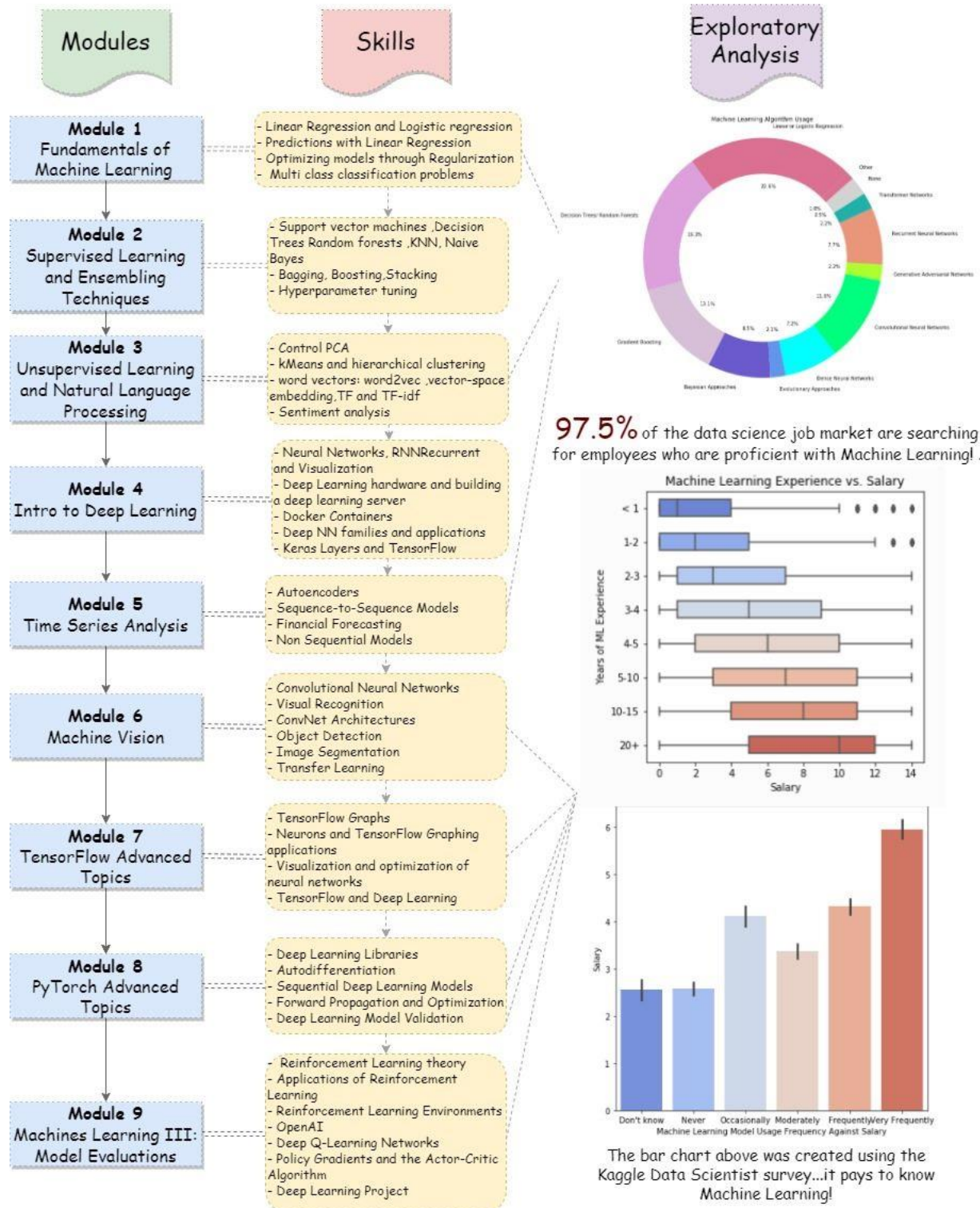
## Appendix

### Clustering results -- dendrogram



## Course curriculum for five other courses

### Advanced Machine Learning





**Modules**

- Module 1**  
Introduction to AI, Modeling, Visualizations, and Optimization
- Module 2**  
Unsupervised Learning, Dimensionality Reduction, Recommender systems
- Module 2**  
ML, Neural Networks and Deep Learning in Finance
- Module 2**  
Applications of AI and ML in Finance

**Skills**

- Python, EDA, Supervised Learning Review (Linear, KNN, tree methods), HPT, CV, Loss Functions
- K-Means, Hierarchical, Probabilistic clustering, PCA, Recommender systems
- Neural Networks, LSTMs, optimization
- MPT, Sharpe Ratios, Risk/Insurance Assessment, Fraud Detection

**Exploratory Analysis**

Recommended Language by Data Scientists

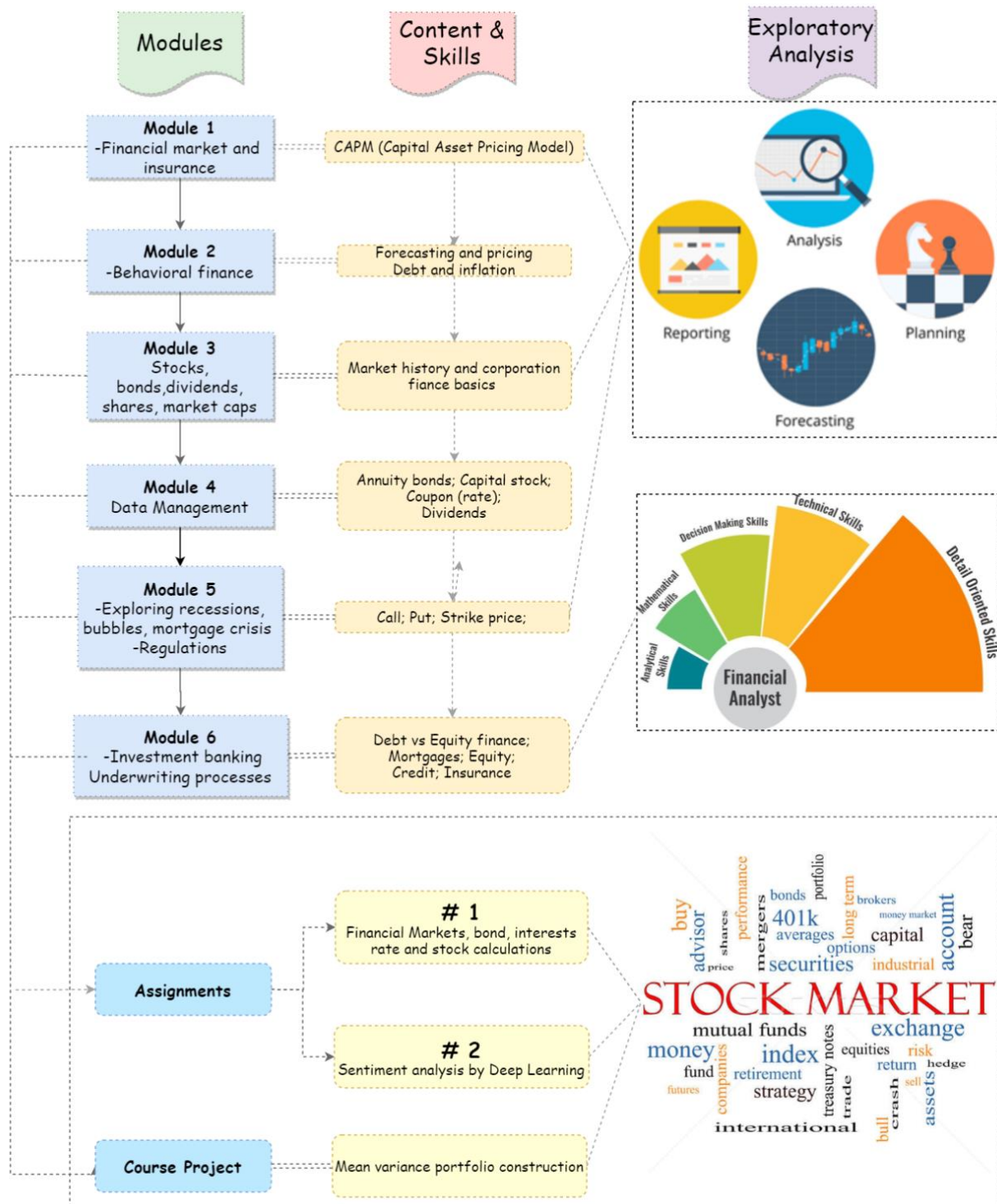
Language	Percentage
Python	80.3%
R	8.8%
SQL	6.1%
Other	0.9%
JavaScript	0.3%
Java	0.2%
C++	0.1%
Bash	0.1%
MATLAB	0.0%

**Grading**

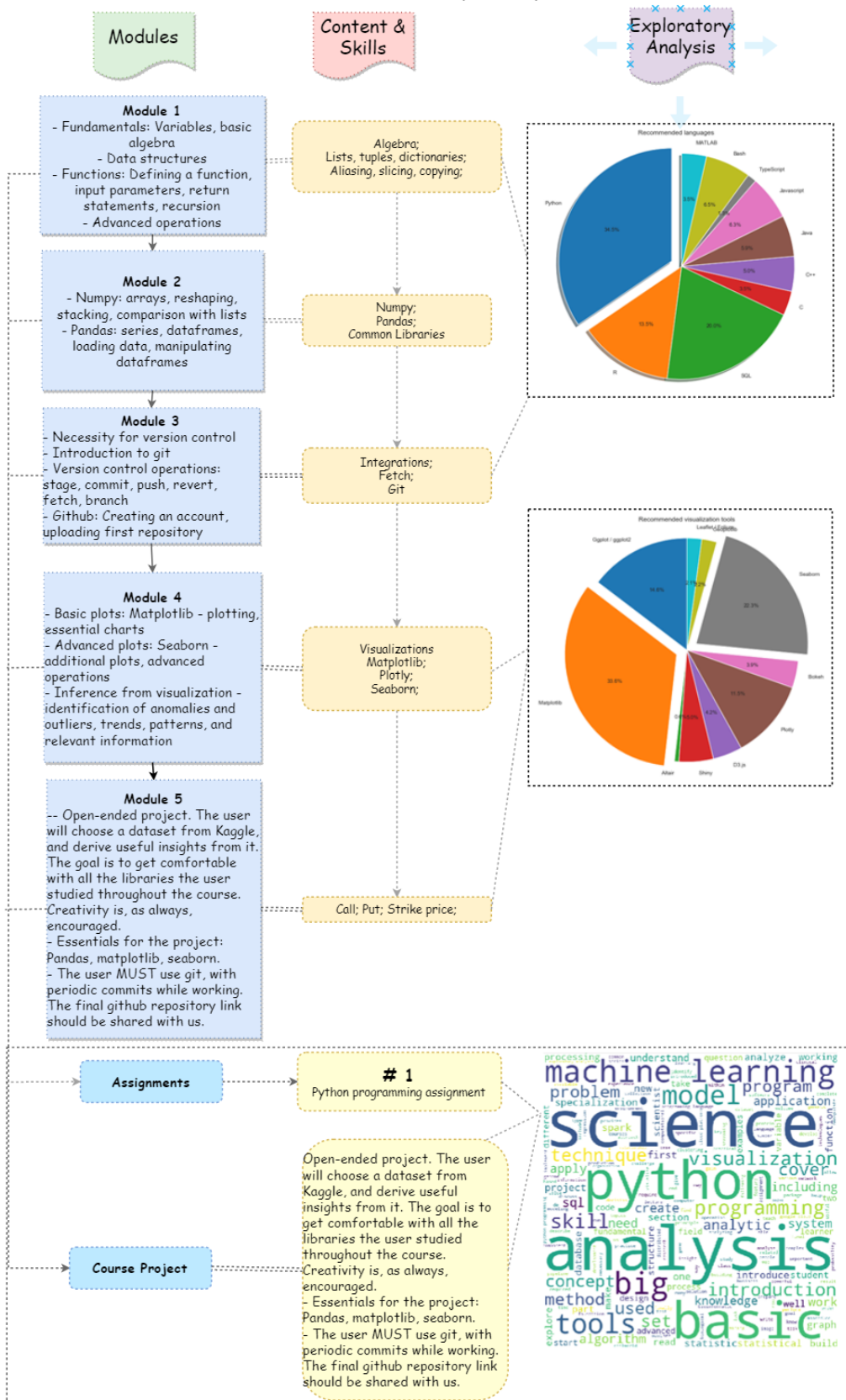
- A1: Market analysis using Supervised Learning
- A2: Portfolio analysis using Unsupervised Learning
- A3: Risk Assessment, Portfolio creation, and optimization

**Exam**

## Introduction to Finance



## Introduction to Python programming



# Statistics for Data Science & Business

