

MIE 1624 Course Project

11.30.2020

**─**  
Furqan Ahmed  
Ahmad Sajedi   
Allen Zhang   
Lisa Li  
Michelle Cheng  
Sai Anirudh Basamsetty

**Table of Contents**

[1 Introduction 2](#_Toc57663867)

[2 Program Curriculum 2](#_Toc57663868)

[2.1 Data Collection and Analysis 2](#_Toc57663869)

[2.2 Defining the Program 3](#_Toc57663870)

[2.3 Building the Program 4](#_Toc57663871)

[2.3.1 Program Curriculum 5](#_Toc57663872)

[2.4 Building the Skill Profiles 5](#_Toc57663873)

[2.5 Detailed Course Curriculum Breakdown – MIE1624 7](#_Toc57663874)

[3 Personalized Career Coach 7](#_Toc57663875)

[3.1 Implementation Details 8](#_Toc57663876)

[3.1.1 Data gathering 8](#_Toc57663877)

[3.1.2 Data Exploration 8](#_Toc57663878)

[3.1.3 Collaborative Filtering 9](#_Toc57663879)

[3.1.4 Course Recommendations 9](#_Toc57663880)

[3.2 Model Results 10](#_Toc57663881)

[4 References 12](#_Toc57663882)

[5 Appendices 13](#_Toc57663883)

[Appendix A – Additional Dendrogram Diagram 13](#_Toc57663884)

[Appendix B – Market Research 13](#_Toc57663885)

# Introduction

Recent advancements in machine learning technology and rapidly increasing speed of data acquisition have caused a surge in demand for Data Scientists. Below are a few statistics on the demand for data scientists obtained from [1]. From Figure 1, we can see that Data Scientists and Data Engineers have amongst the highest projected growth in the next 5 years.

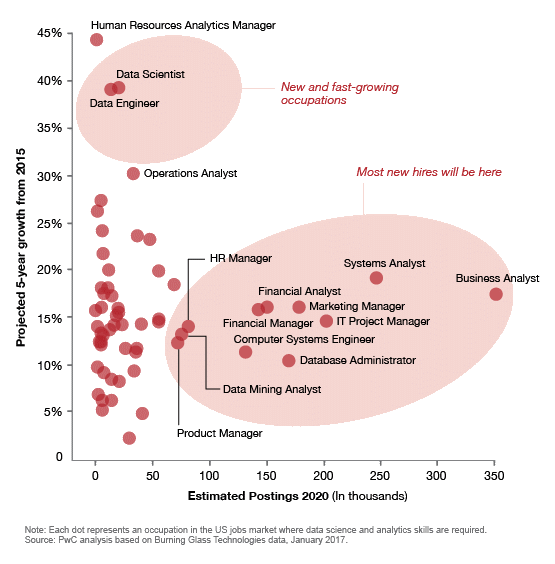
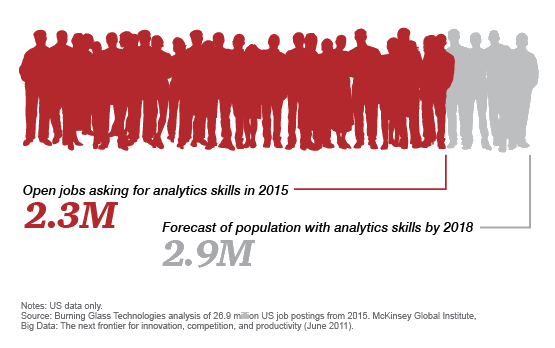


Figure 1: Trends in Data Science

Given this context, we expect a significant increase in demand for educational programs in the data science field. In this project we propose educational solutions focused at addressing this surge in demand. Our proposal is focused around 3 main pillars:

* Redesigning the current course structure of MIE1624 – Introduction to Data Science and Analytics, to better address industry relevant skills
* Proposing an entire program curriculum for a Data Science focused master’s degree
* Designing an EdTech solution that helps working professionals identify and bridge gaps in their skills through dedicated online courses.

# Program Curriculum

In this section, we focused on designing a sequence of courses and a curriculum for each of those to be used by the University of Toronto for a new program that is equal parts technical as it is business-oriented.

The Master of Data Science and Artificial Intelligence program will be very beneficial to students as these skills are necessary for both the tech and business industries. More and more companies are turning towards data science and artificial intelligence to make critical business decisions and these skills are in high demand within today’s society.

## Data Collection and Analysis

In order to prepare students and properly equip them with the necessary knowledge, we performed our analyses on reliable data sources to ensure our curriculum would be relevant and useful in the real world. We scrapped data from Coursera and LinkedIn that was filtered specifically for topics related to data science or artificial intelligence. A Kaggle data science survey was also included for further analysis. The figures below are examples of our data analysis:

Chart, bar chart

Description automatically generated

Figure 2: Top 10 Most Frequent Job Skills from LinkedIn

Chart, funnel chart

Description automatically generated

Figure 3: Most Popular Machine Learning Algorithms from Kaggle Survey

## Defining the Program

From the data mentioned above, the foundation of the course curriculum was built. Hierarchical clustering was used to visualize which skills were most correlated with each other with the purpose of determining which topics should be grouped and taught together. The following diagram visualizes the four core buckets that were extracted and would become the basis for developing the program structure.

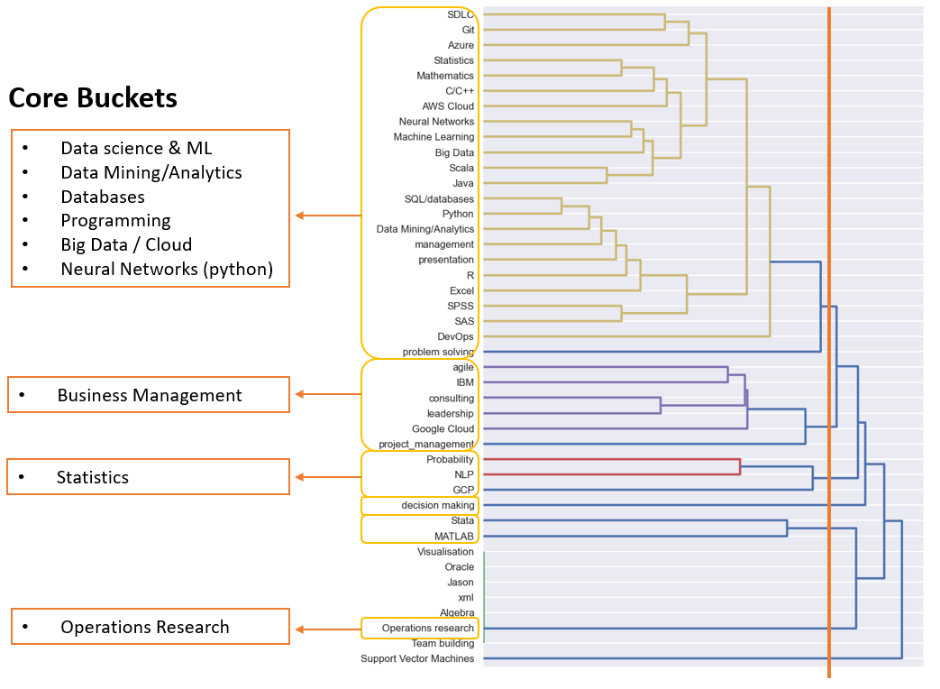


Figure 4: Dendrogram of Skills

The four core buckets can be summarized as: 1) programming related courses, 2) business management related courses, 3) statistic courses, and 4) operations research courses. The programming bucket was the largest and makes sense because many of these skills are fundamental to learning more advanced topics. In the next section, we will discuss how these core buckets translate into courses and then further into skills.

## Building the Program

After determining the core buckets, a high-level program organization skeleton was defined. This skeleton expanded the core buckets further, into more detailed sub-categories and found courses in the current market that aligned with these concepts. Since data science/machine learning and data analytics often walk hand-in-hand, those two core buckets were merged into one in the summary table below.

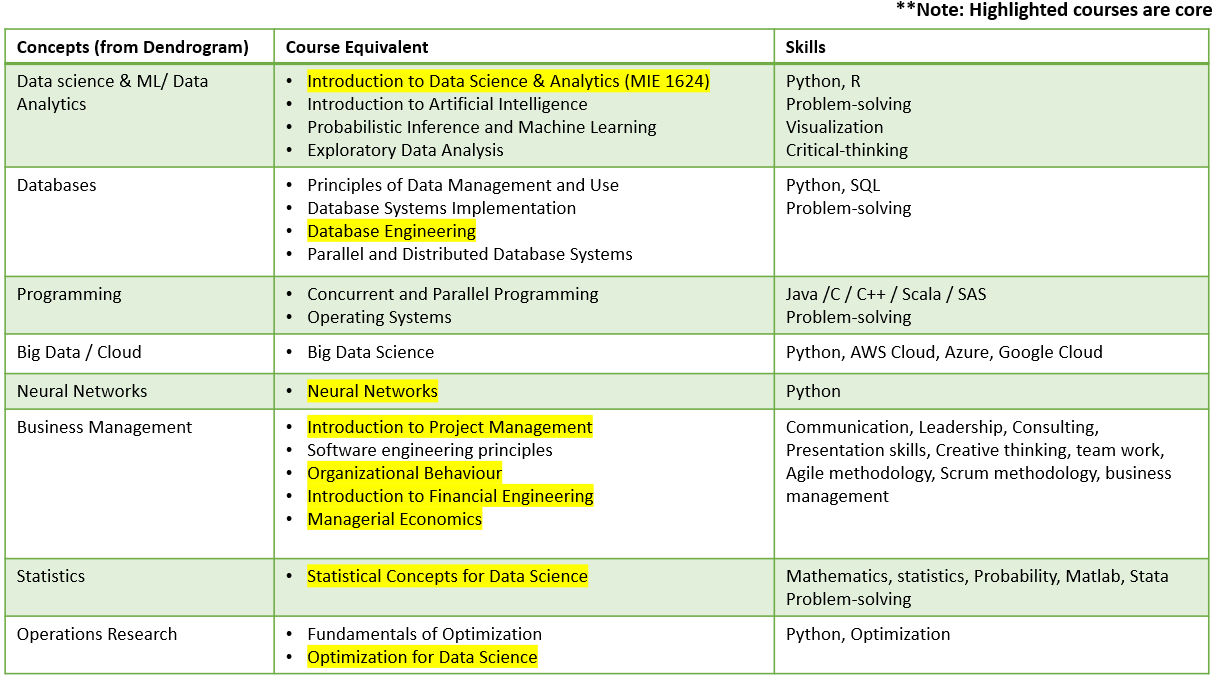


Figure 5: High-level Program Outline

As noted in the figure above, the highlighted courses are deemed as “core” courses and are mandatory for all students to take. Since this program is geared towards equal exposure to both technical and business aspects, we wanted to make sure students had a solid foundation in both areas, so the core courses are split into five technical and four business. This makes our program unique because many existing programs on focused on developing technical skills, but often forget the importance of the business side.

The aim in determining core courses was to try and have at least one core course from every core bucket to ensure students received a diverse skill set. If there was more than one course in the core bucket, the course that was the most common among other universities offering similar programs was selected. Market research was conducted throughout this process to explore what was currently in the market, allowing us to propose a program that is innovative and competitive. For more details on other comparable programs that currently exist, see Appendix A.

### Program Curriculum

The next step was to determine the overall flow of the program using what was found in the section above. Figure 6 shows an overview of what the program will look like. The proposed program is two years long and will consist of 10 courses in total, with equal emphasis placed on developing technical and business-oriented skills. Nine of these courses will be core courses, which are the ones highlighted in Figure 5**,** and one will be an elective that can be selected from any of the courses that are not highlighted.

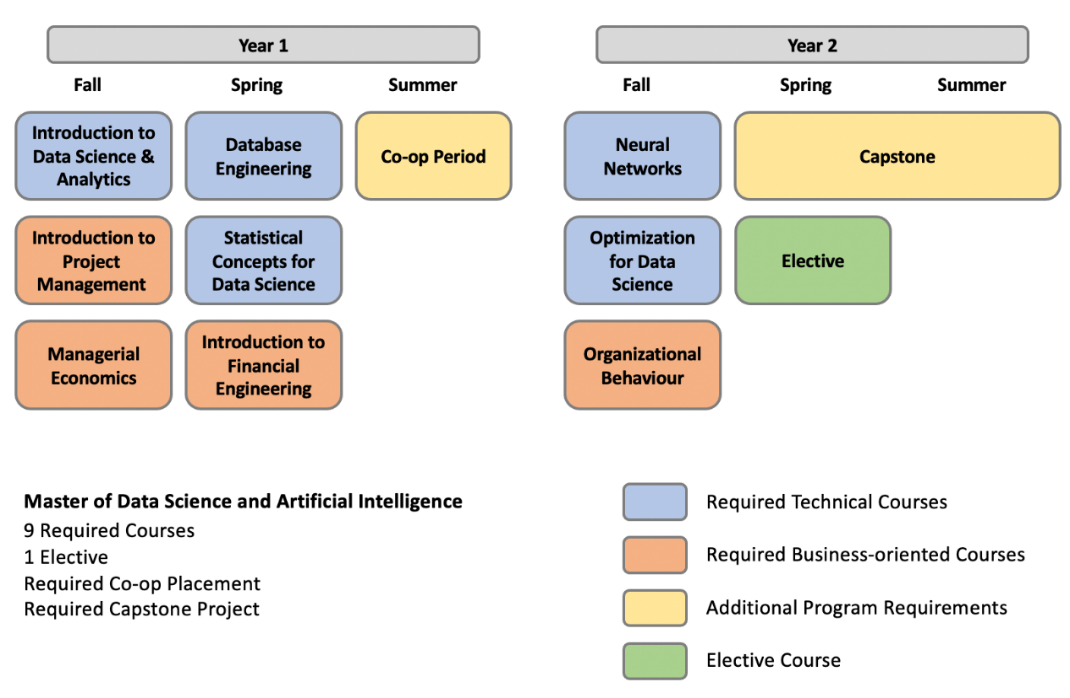


Figure 6: Program Curriculum

The program also includes a mandatory co-op term and a capstone project to allow students to gain practical work experience, develop teamwork skills, and provide students with a head start on career networking.

## Building the Skill Profiles

After obtaining a high-level program curriculum, five separate courses were outlined in more detail with the skills that would be learned and the order that they would be taught. The skills that were chosen for each course were first determined through the exploratory data analysis and then checked against market research.

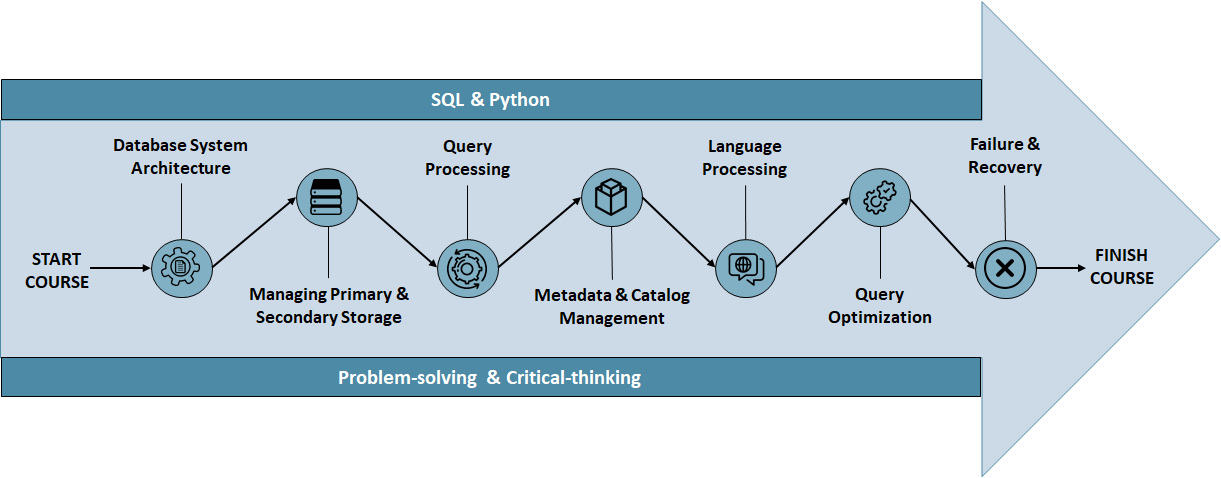


Figure 7: Database Engineering Course

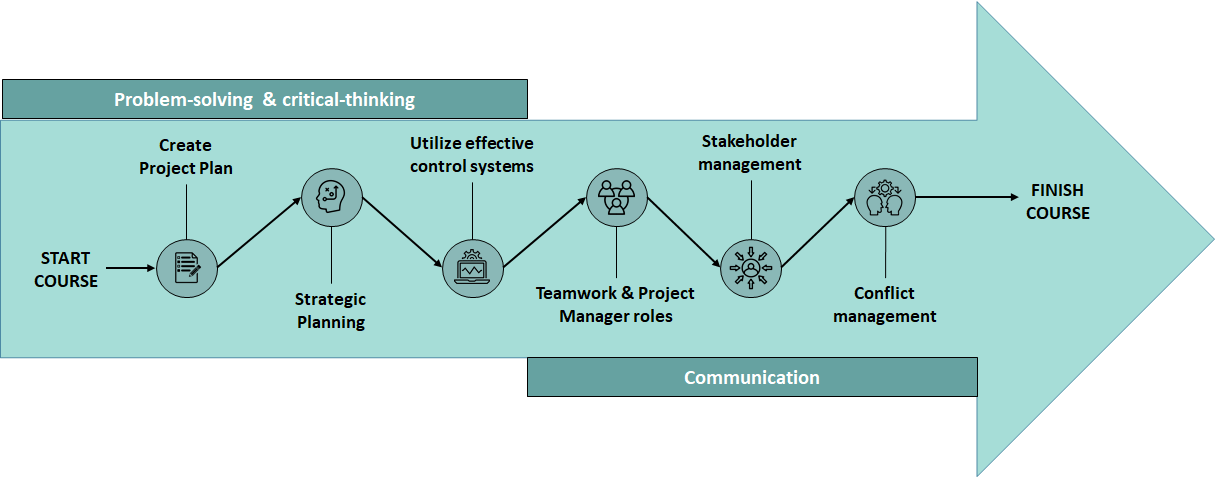


Figure 8: Introduction to Project Management Course

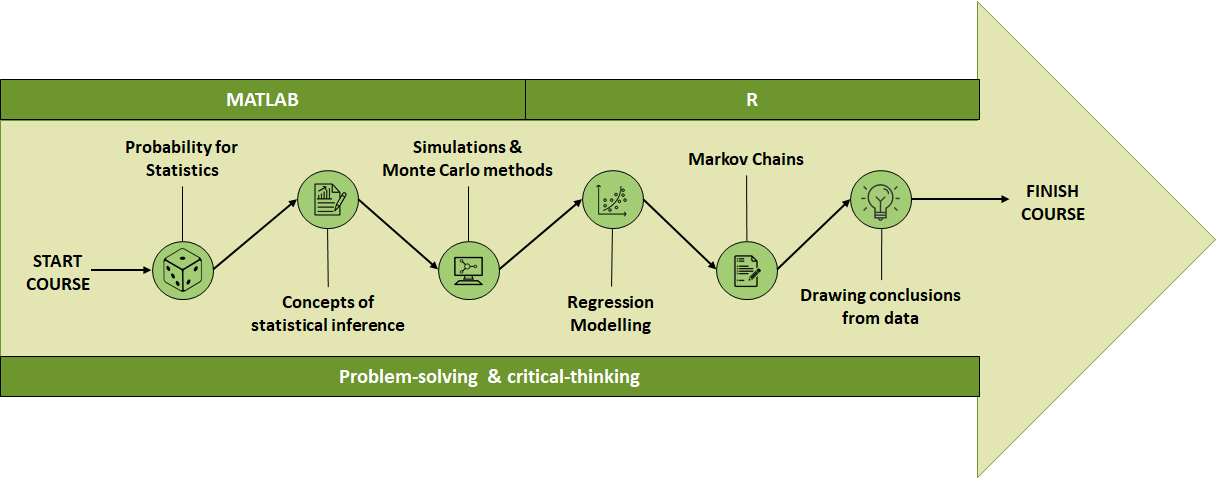


Figure 9: Statistical Concepts for Data Science Course

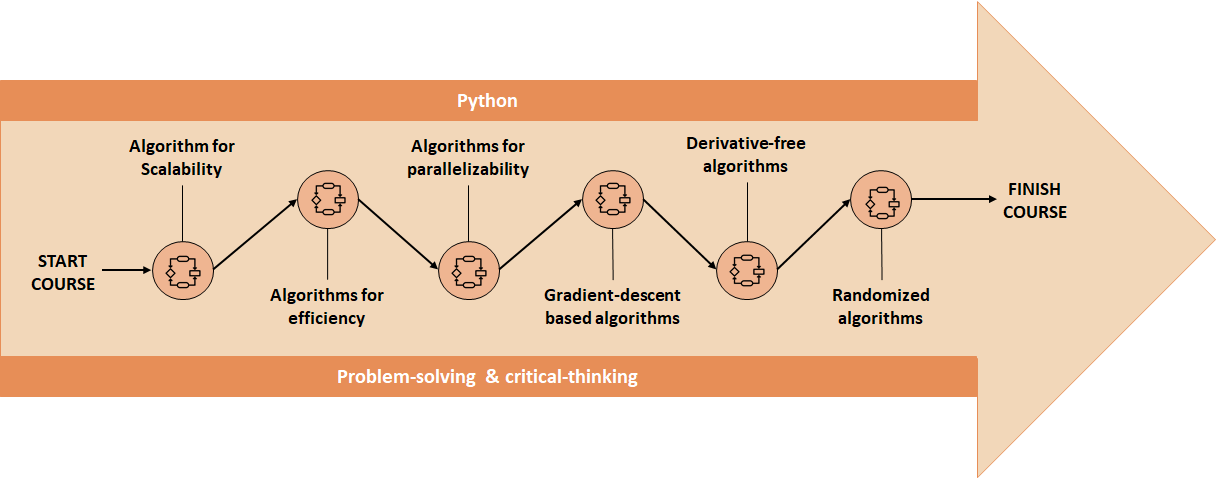


Figure 10: Optimization for Data Science Course

From the figures above, it can be seen that for each of the courses described, there are set skills that should be achieved by the end of the course and there is a specific sequence that the skills should be learned in, with each new skill building upon the last. The five displayed visuals are just a few examples within the overall program curriculum that display how skills are sequentially ordered and what the students can expect to take away from each of these courses.

## Detailed Course Curriculum Breakdown – MIE1624

The following section describe a redesign proposal for MIE1624 *Introduction of Data Science and Analytics*. Based on the data provided in previous sections, technical skill required for data analyst/scientist/manager job posting are used for course topic selection. Compared to current course curriculum the following changes should be considered:

1. Adding SQL/database and R language as part of the project
2. Popular visualization techniques/tools such as Tableau and Power BI should be included
3. Including Excel in the tutorial as Excel is frequently mentioned in the skill requirement

The proposed new curriculum can be found in the flow chart below, where core concepts of the course are displayed in the circles and the highlighted skills are indicated above and below the arrows. Detailed topics are listed under each core concepts:

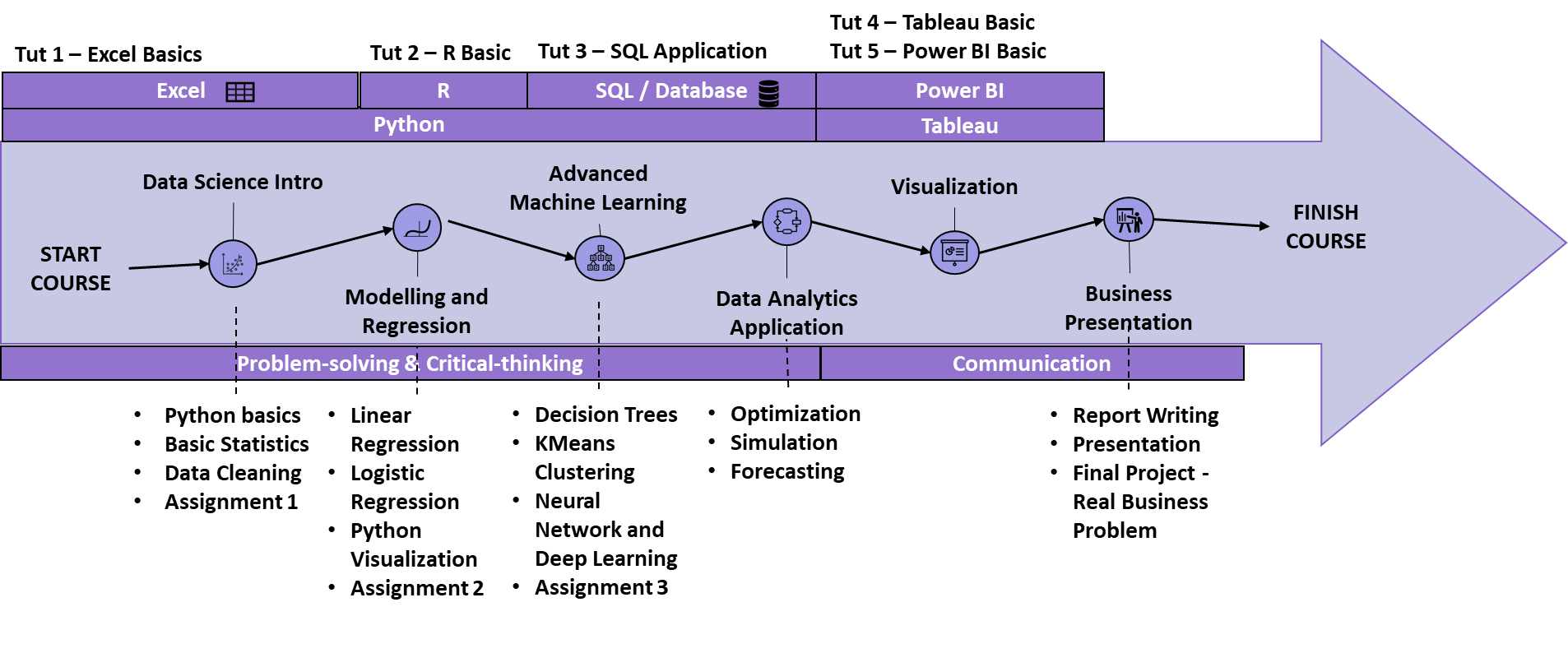


Figure 11: MIE1624 Skill Profile

# Personalized Career Coach

In our project we propose an EdTech solution that tackles the following problem:

“*How to help working professionals identify gaps in their knowledge base and upgrade their skills?*”

Our solution is to build a recommendation system that analyzes the user’s LinkedIn profile and identifies skills that are most relevant for them to learn. This identification is done by using a collaborative filtering approach. The intuition behind this approach is as follows:

1. Look for users who share the same skill patterns with the active user.
2. Use the skills from those like-minded users to calculate a prediction for the active user.

Once the relevant missing skills are predicted, we then also recommend a list of online courses that address the missing skills.

## Implementation Details

### Data gathering

Data for the modelling task was collected by scraping LinkedIn profiles of users. LinkedIn limits the number of profiles that can be accessed through a scraping bot to a few hundred profiles. Therefore, we decided to focus our efforts on only data scientist profiles. We collected a total of 460 profiles with their job titles, total experience, education, and listed skills. We also scraped Coursera and extracted data science related courses along with their associated skills.

### Data Exploration

Having gathered data from user profiles, we proceeded to perform some data exploration to better understand our dataset. Findings from the data exploration are summarized in the graphs below.

#### User Skills

A total of 2050 unique skills were identified in the data set. The plot below shows the distribution of the top 25 skills in the set. Top 100 most common skills were retained for model analysis.

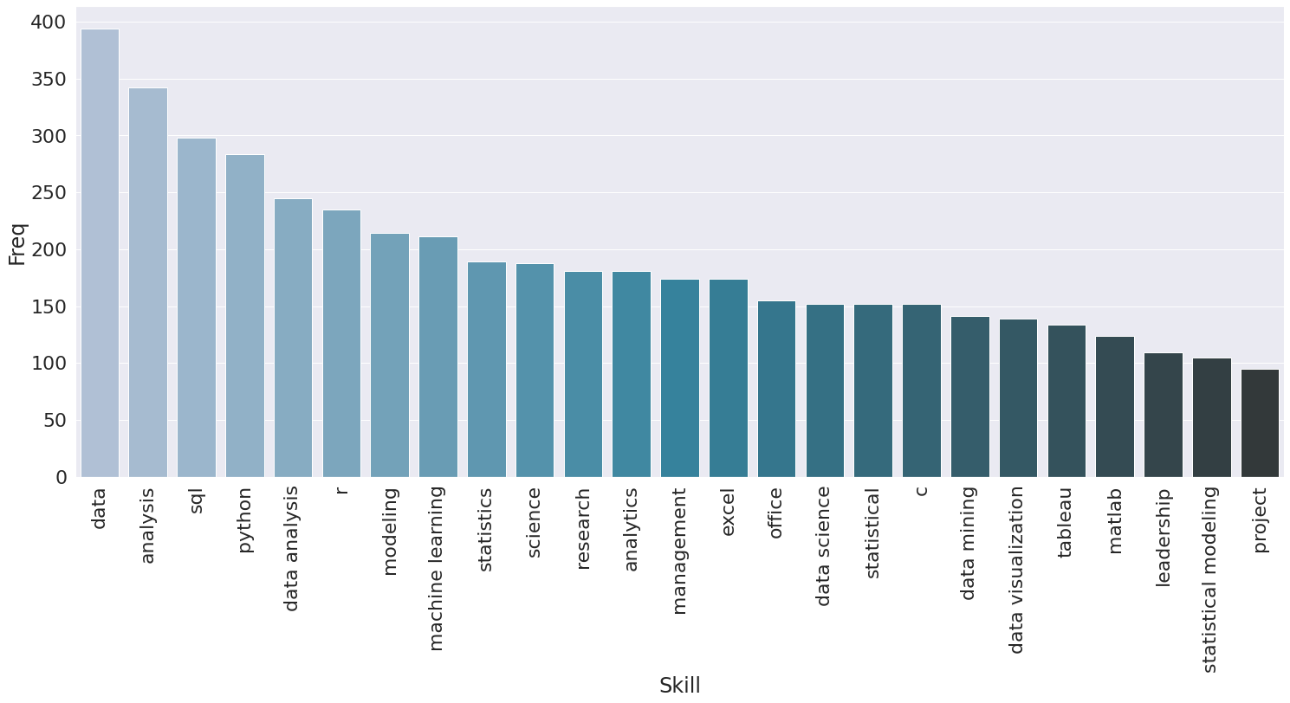


Figure 12: Most common skills amongst Data Scientists on LinkedIn

#### Skill Rating

Each profile was assigned a ‘Skill Rating’. This rating is defined as the total number of skills each profile possesses. The idea here is that the more skills a profile possesses, the more competitive it is. We analyzed the distribution of ‘Skill Rating’ within each experience bucket. Bootstrapping the data resulted in the following distributions.

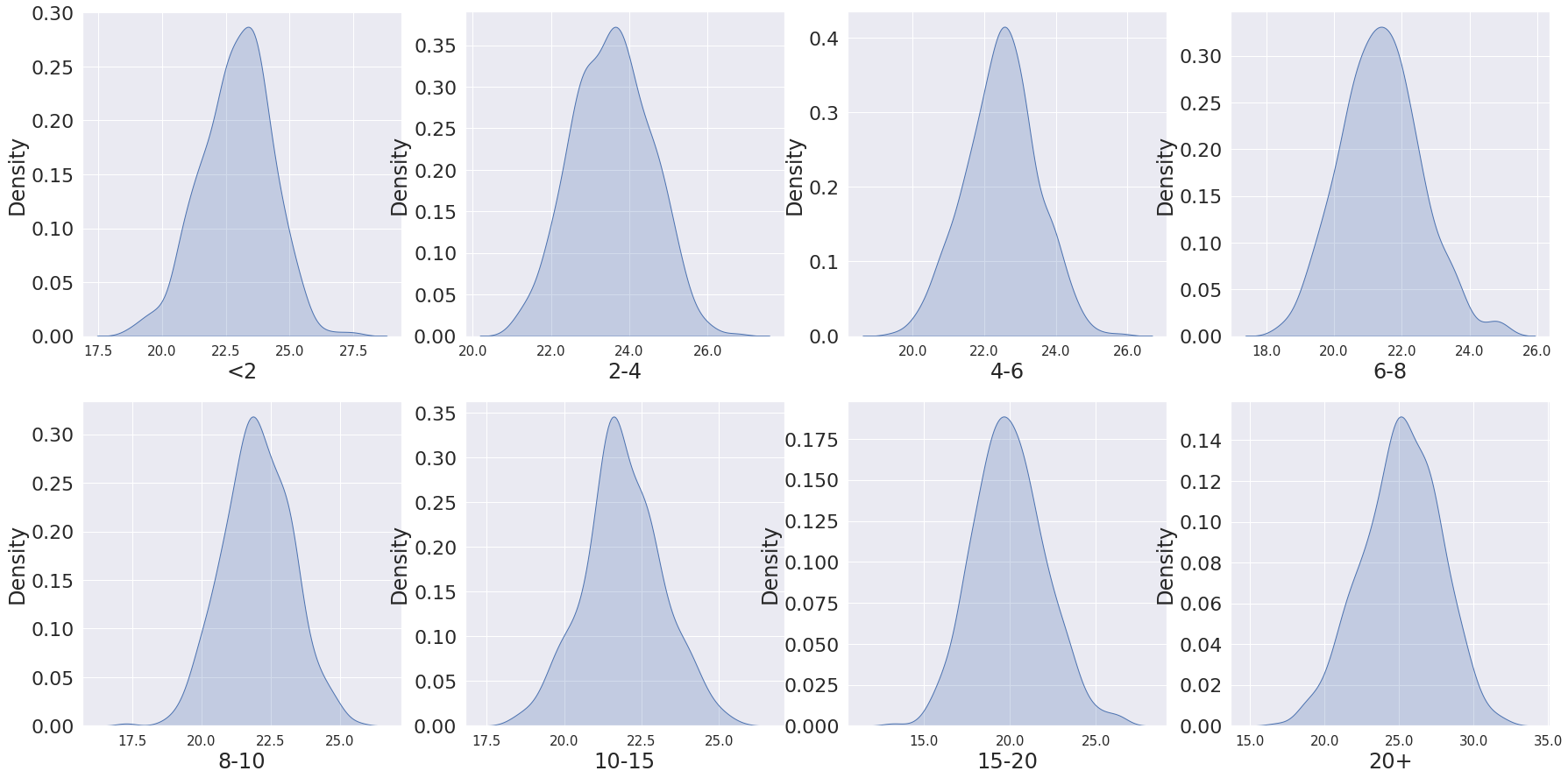


Figure 13: Skill Rating Distribution Per Experience Bucket

### Collaborative Filtering

Collaborative filtering is a technique that is widely used across social media, retail, and streaming services. The concept behind collaborative filtering is straightforward: It is based on the idea that people who share an interest in certain things will probably have similar tastes in other things as well. You experience collaborative filtering first-hand every time you go online and see “Customers Who Bought This Item Also Bought,” or “Users like you also liked…”.

The main difference between collaborative filtering and content-based filtering is conceptual. Where content-based filtering is built around the attributes of a given object, collaborative filtering relies on the behavior of users.

Collaborative Filtering approaches can be divided into two main sections: **user-item filtering** and **item-item filtering**.

* **User-item filtering:** Take a particular user, find users that are like that user based on similarity of ratings, and recommend items that those similar users liked i.e., “Users who liked this item also liked …”.
* **Item-item filtering:** Take an item, find users who liked that item, and find other items that those users or similar users also liked. In a simple word “Users who are similar to you also liked …”.

In this project, we used both user-item and item-item filtering for our users in LinkedIn. Our core for the course recommendation task is using the memory-based collaborative filtering.

### Course Recommendations

In recent years, learning online using the e-learning platforms has become indispensable in the teaching process. Companies and scientific researchers try to find new optimal methods and approaches that can improve online education. In this report, we propose a new recommendation approach for recommending relevant courses to users in LinkedIn based on their skills. Our method is based on social filtering and collaborative filtering for defining the best way in which the learner must learn and recommend courses which better match the learner’s profile and social content.

The Course Recommender System is based on the several different collaborative filtering algorithms like user-based, item-based, and OC1. First, the system can predict the usefulness of skills to a particular user in LinkedIn based on other similar users’ skills. We used both item-item and user-item filtering to recommend the skills. After predicting the selected skills for each user, we can recommend the courses for increasing the perspective skills easily. In other words, we recommend skills and their corresponding courses to each user via a memory-based algorithm like collaborating filtering.

Here is a recommendation example for one selected LinkedIn users:

* We recommend [[‘Machine Learning and Reinforcement learning in Finance Specialization’], [‘Machine Learning’, ‘Machine Learning: Classification’], [‘Python Data Products for predictive Analytics Specialization’]] to user 364 for increasing [‘financial engineering’, ‘logistic regression’, ‘data processing’] skills, respectively.

## Model Results

Having gathered the data and trained our recommendation model, we are now able to suggest skills and course recommendations for a given user profile. Based on these profile details, we can produce the following summary statistics and predictions for the profile.

The following page compares two user examples in detail:

The first user is assessed to have a percentile rank of 7.2% as compared to other users with similar experience and skillset. The system recommended three courses that help user to learn the key skills they are currently missing (Table 1, Figure 14).

On the other hand, the second user has a better skill rating than 82% of the other users. We can find that the system recommended 3 totally different courses compared to the first user (Table 2, Figure 15).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Index | Job1 | Job2 | Experience (yrs) | Skills |
| 32 | Senior Data Scientist | Data Science Instructor | 10.58 | Sql, Data Analysis, Machine learning, C, Matlab, Scripting, +13 more |

Table 1: User Profile 1 from LinkedIn Users Database

Chart, histogram

Description automatically generated

Figure 14: Course Recommendations for User 1 Based on their Profile

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Index | Job1 | Job2 | Experience (yrs) | Skills |
| 52 | Data Scientist | Senior Analyst | 1.83 | Statistics, R, Python, Visual Studio, SPSS, SQL,+18 more |

Table 2: User Profile 2 from LinkedIn Users Database

Chart, histogram

Description automatically generated

Figure 15: Course Recommendations for User 2 Based on their Profile

# References

1. PricewaterhouseCoopers. “What's next for the Data Science and Analytics Job Market?” *PwC*, [www.pwc.com/us/en/library/data-science-and-analytics.html](http://www.pwc.com/us/en/library/data-science-and-analytics.html).
2. David CravenFollowSoftware Acquisition & Data Analyst at EOS IT Management SolutionsLike374Comment65ShareLinkedInFacebookTwitter4, and Follow. *Use Selenium & Python to Scrape LinkedIn Profiles*, [www.linkedin.com/pulse/how-easy-scraping-data-from-linkedin-profiles-david-craven/](http://www.linkedin.com/pulse/how-easy-scraping-data-from-linkedin-profiles-david-craven/).
3. Reinstein, I. “Best Masters in Data Science and Analytics in US/Canada”. *KDnugget*, <https://www.kdnuggets.com/2017/11/best-masters-data-science-analytics-us-canada.html>

# Appendices

## Appendix A – Market Research Findings

A search of the top Master’s in Data Science and Analytics in US/Canada was also conducted prior to designing our program structure. A list of the top Data Science programs was retrieved from an article in KDnuggets, a discussion and learning website for business analytics, data mining, and data science [3]. Below is a table which summarize the structure of other comparable data science programs:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **University** | **Program Name** | **Course Requirements** | **Program Duration** | **Additional Activities** |
| Massachusetts Institute of Technology | Master of Business Analytics | * 10 Courses * Capstone | 24-months | * Communications and data storytelling seminar |
| Carnegie Mellon University | Master of Computational Data Science | * 9 Courses * Capstone | 16-months or 20-months | * Summer internship |
| Harvard University | Master of Engineering in Data Science | * 12 Courses | Minimum 12-months | * Optional summer internship * Option of thesis or capstone |
| University of Toronto | Master of Science in Applied Computing (Data Science Concentration) | * 6 Courses | 24-months | * Mandatory internship |
| University of Washington | Master of Science in Data Science | * 8 Courses * Capstone | 18-months (full-time) | * N/A |
| University of British Columbia | Master of Data Science | * 24 four-week long courses * Capstone | 10-months | * N/A |
| University of Texas | Master of Science in Business Analytics | * 11 Core Courses + 3 Credits from Elective * Capstone | 10-months | * N/A |