UNIVERSITY OF TORONTO

Creating and Designing Data Science Education

MIE1624: Introduction to Data Science and Analytics

Group-17

Team Members:

Harsh Panchal	1005661507
Hajra Begum	1005614015
Kaan Kucukoglu	1005611189
Shreyas Choudhary	1006376217
Shresh Kumar Mathur	1006586883
Aditya Subramaniam Sivasubramaniam	1005675789

Executive Summary

Data Science and Analytics covers a multi-disciplinary area which is expanding ever so rapidly as businesses rely more heavily on data to drive decision-making than ever. Organizations today need professionals who can extract and analyze extremely large amounts of data - Big Data - and present useful insights to business leaders. They require fast solutions for easy model deployment, constant model monitoring, and flexible model management. According to a report, in 2020, the job requirements for data science and analytics are projected to boom to by 364,000 openings to 2,720,000. And according to the U.S. Bureau of Labor Statistics, 11.5 million new jobs will be created by the year 2026. According to a survey by the MIT Sloan Management Review, 43 percent of companies report a lack of appropriate analytic skills as a key challenge with a greater demand for data scientists, therefore the demand for data science education is increasing accordingly [1]. With a widening skills gap in the data science sector, there is a need to optimize data science education so that graduates are equipped with the necessary skills and qualifications to succeed in the workforce.

To redesign and fully encompass the core concepts of Data sciences, Data from Indeed Canada and LinkedIn were web scraped to find the key aspects about the current job scenario and postings, from Coursera and leading universities in data science (University of Waterloo , Queens University, etc.) for a more in-depth analysis on data science-related course curriculums. A Kaggle Machine Learning and Data Science Survey 2019 was also used in this analysis.

The first focus was to re-design the course curriculum for "MIE1624: Introduction to Data Science and Analytics" course at the University of Toronto on data science. The most important requirement in redesigning is to cover the most relevant topics and skills in data science and AI. Based on the content and structure of the top introductory data science courses of 2020 from Coursera data, the latest skills needed for data science Jobs are obtained and have been used to find the course curriculum topics that have to be included.

For students interested in pursuing a master's related to data science, they have an option to choose a more technically- and business-oriented program named Master of Data Science and Artificial Intelligence. The program was developed by performing Natural Language Processing of job descriptions. The course curriculums were then developed using a combination of exploratory Analysis and clustering of skills from the Job data.

The Master of Data Science and Artificial Intelligence Program is designed with an option of a 1 year (Fulltime) or 2 years (Extended Fulltime). The degree completion requirements involve the completion of 10 graduate-level courses (at least 6 technical courses) or 7 courses (at least 4 technical courses) with a project.

On the quest to find an effective alternative to traditional education methods, relevant trends in education and industry were analyzed.

Table of Contents

1. Introduction

"The ability to take data — to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it — that's going to be a hugely important skill in the next decades." - Hal Varian, chief economist at Google and UC Berkeley professor of information sciences, business, and economics [2]

Data scientists have become critical assets over the past decade and are present in virtually all organizations. They need to be well-rounded, data-driven individuals with high-level technical expertise who are able to create sophisticated quantitative algorithms to coordinate and synthesize vast volumes of knowledge used in their company to address quantitative and qualitative based questions and drive strategy. This is combined with the communication and leadership skills required to produce measurable outcomes through a company or sector to different stakeholders.[3]

The fastest-growing roles are Data Scientists and Advanced Analysts, which are projected to see demand spike by 28% by 2020. [1] As growing volumes of data are more available, big tech companies are no longer the only ones that require data scientists. A lack of eligible applicants available for filling the open roles is challenging the rising need for data science talent across sectors, large and small. The demand for data scientists is showing no sign of slowing down in the years to come. In 2017 and 2018, LinkedIn listed data scientists as one of the most exciting jobs, along with multiple data-science-related skills as the most requested by companies [4].

According to a study by IBM, more than 35% of Data Scientists and Advanced Analyst positions out of the 2 million positions every year require a Master's or Ph.D. The most in-demand positions in data science and analytics require advanced schooling, further boosting competition and wages for those qualified professionals.[1,5] With today's enormous need for data scientists and data analyst professionals on the market, the need for a program and curriculum is ever prevalent to jumpstart one's journey in data science. To provide a portfolio of deliverables in data science to give people the learning required to take the plunge and start their data science career.[6]

In response to which several Canadian institutions have recently established revenue-generating master's programs related to data science education. The University of Waterloo's course specializing in Data Science, York University introduced a Master's in Business Analytics and the Desautels Faculty of Management at McGill University and the Rotman School Of Management at the University of Toronto both began a Master of Management Analytics in September 2018. Additionally, University of Toronto also offers a Master of Science in Applied Computing (Data Science Concentration) is offered jointly by their Department of CS and Statistical Sciences. These master's programs are targeting quantitatively strong students who have recently completed undergraduate studies to provide them with advanced data management and communications skills.

2. Problem Statement

Learning online became one of the most popular forms of learning. With great platforms like MIT OCWare, Udemy, Coursera, Data Camp and many others, anyone at any level can find something to learn from. In this digital era, it is very important for top Universities to make their course curriculum that could cover the latest holistic set of skills. As the University of Toronto is one of the best, it's course curriculum must be competitive even at the introductory level for Data science. Therefore, there is a need to redesign "Introduction to Data Science and Analytics" for consistent and up to date learning.

A recent Burning Glass report, "The Quant Crunch", found that 42% of Data Scientist positions require a graduate degree.[1] On the other hand, things move very fast in data science and machine learning. And they'll only move faster in the future so think to give 2 years for education. Lastly with dozens of new Master's in Data Science degree programs from other Universities. This concludes a need to design a master's degree that can help students develop skills in many requisite areas that some self-taught candidates may miss out on.

"What can U of T offer to help their students find more accurate jobs within shorter time frames?" This question has mostly been unanswered for both students and companies related to Data Science. In the urge to find a simple solution to that a brand-new portal named Data Science Warriors is to be proposed.

3. Data Acquisition

Web Scraping (also termed Screen Scraping, Web Data Extraction, Web Harvesting etc.) is a technique employed to extract large amounts of data from websites whereby the data is extracted and saved to a local file in your computer or to a database in table (spreadsheet) format.

This technique is used to acquire data from Indeed and LinkedIn for Job postings for Data scientists in Toronto, Canada and saved as CSV files. Then the curriculum of the latest courses is obtained from the Coursera website along with the skills. These two data files are important to redesign the course curriculum.

Secondly, the Kaggle Machine Learning and Data Science Survey 2019 was taken to find the Job title categories based on Salaries. Then using these Job titles data is scraped from Indeed Canada for High-level job postings and Mid-level Job postings. This data sets the framework for designing the Degree program.

4. Part 1 - Course curriculum redesign [Existing MIE 1624: Intro to Data Science and Analytics]

4.1 Aim:

To redesign the course MIE1624 at the University of Toronto. This introductory course should be able to cover a spectrum of skills needed to obtain a Junior or Middle-level jobs in Data Science.

4.2 Approach

We first scrapped indeed and LinkedIn job portals to get job titles and job descriptions and then to get the necessary skills we scrapped the Coursera learning portal. We considered only junior level data scientist jobs as, with the midlevel course, we won't be able to secure senior-level jobs. Our goal is to get skills that are required to get junior-level jobs.

4.3 Exploratory Analysis

The features (skills) obtained are visualized to check their importance based on occurrences of them in the job description. From the obtained graph, it is evident that almost 50% of the Junior level jobs required SQL as a necessary skill in their job posting. We have visualized the top 15 skills in the below bar graph.

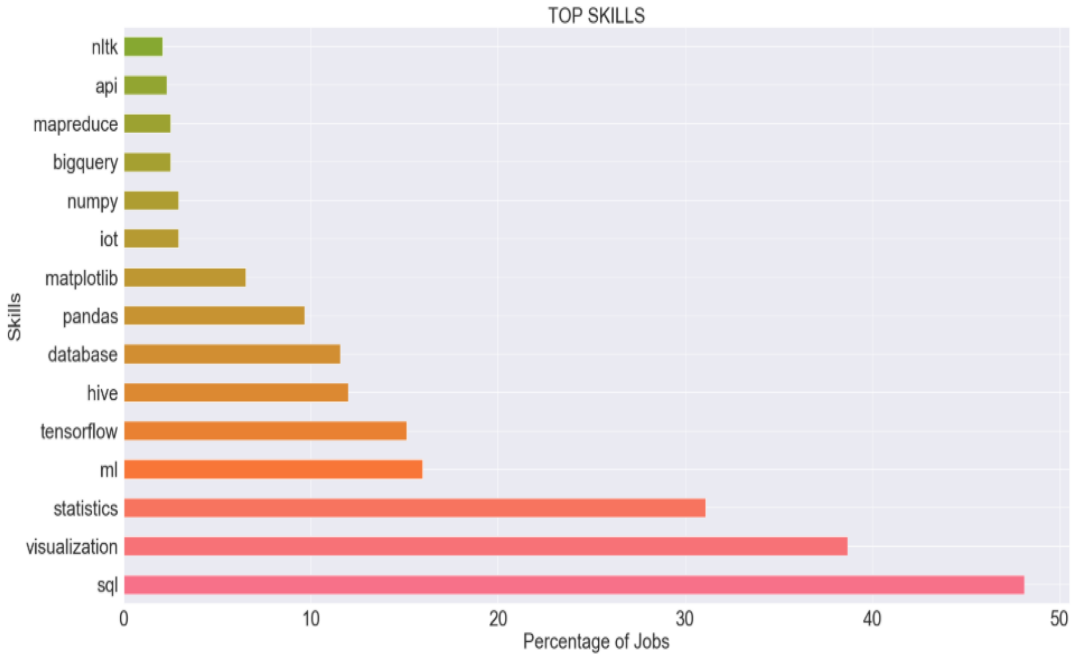

Fig 1- Skillsets identified for the course curriculum redesign

4.4 Course Redesign:

The features obtained from the job description are used to find the topics /courses related to those skills. We selected relevant top skills to redesign the MIE1624 course. The Final implementation part was to compare the current topics with the new topics obtained from the model.

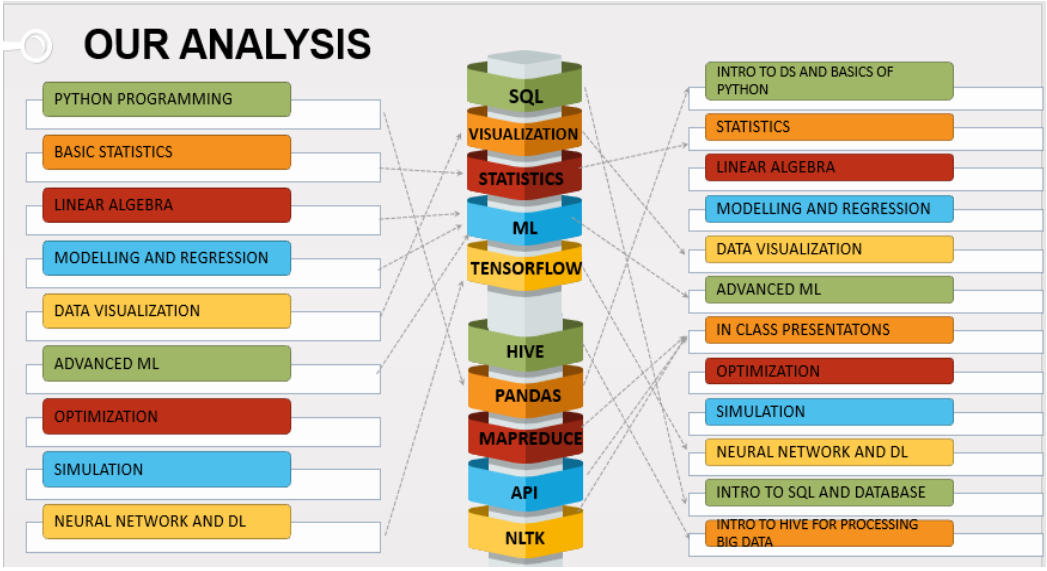

Fig 2- Analysis skillset mapping to the topics in Data Science

4.5 Results

Based on the skills, we have designed the curriculum for the new proposed course.

COURSE NAME: MIE 1624 - "DATA SCIENCE" (PROPOSED COURSE REDESIGN)

LECTURE 1: INTRODUCTION TO DATA SCIENCE AND BASICS OF PYTHON	1. Data science concepts 2. Basics of python programming 3. Application areas of quantitative modeling 4. Comparison of Python, R and MATLAB usage in data science
LECTURE 2: STATISTICS	1. Random variables, sampling 2. Distributions and statistical measures 3. Hypothesis testing 4. Statistics case studies in IPython.

LECTURE 3: LINEAR ALGEBRA	1. Linear algebra and matrix computations 2. Functions, derivatives, convexity
LECTURE 4: MODELING AND REGRESSION	1. Mathematical modeling process 2. Linear regression 3. Logistic regression 4. Regression case studies in IPython
LECTURE 5: DATA VISUALIZATION	1. Visual analytics 2. Visualizations in Python and visual analytics in IBM Watson Analytics
LECTURE 6: ADVANCED MACHINE LEARNING	1. Classification (decision trees) 2. Advanced supervised machine learning algorithms (Naive Bayes, k-NN, SVM) 3. Intro to ensemble learning algorithms (Random Forest, Gradient Boosting) 4. Intro to neural networks and deep learning 5. Text analytics and natural language processing 6. Clustering (K-means, Fuzzy C-means, Hierarchical Clustering, DBSCAN) 7. Dimensionality reduction 8. Association rules 9. Overview of reinforcement learning 10. Machine learning case studies in IPython
LECTURE 7: OPTIMIZATION	1. Unconstrained non-linear optimization algorithms 2. Overview of constrained optimization algorithms 3. Optimization case studies in IPython
LECTURE 8: SIMULATION	1. Random number generation 2. Monte Carlo simulations 3. Simulation case studies in IPython
LECTURE 9: NEURAL NETWORK AND DEEP LEARNING	1. Deep learning - Mathematics of NN 2. Neural Networks in detail 3. convolutional Neural Nets

	4. TensorFlow
LECTURE 10: INTRO TO SQL AND DATABASE	1. SQL basics 2. SQL Joins 3. SQL Aggregations 4. Subqueries and Temp Tables 5. Window Functions
LECTURE 11: INTRO TO HIVE FOR PROCESSING BIG DATA	1. Hive concepts and setup 2. Working with data in Hive 3. Retrieving data from Hive 4. Aggregating data and Manipulation of data

IN-CLASS PRESENTATION

Class presentation as a learning tool has a significant contribution to our in-depth understanding of different topics so we want to continue with this part in the course curriculum. with the current presentation topics, we want to add some more topics in an in-class presentation: (1) MapReduce (2) Hive (3) SQL and database as these topics are the top skills required by junior data scientist jobs.

ASSIGNMENT 1

The current trend of designing assignment 1 based on Kaggle competition is perfect. So, in the proposed curriculum also there will be assignment 1 based on the same idea.

ASSIGNMENT 2

We have two options for assignment 2: 1) based on NLP or 2) based on SQL and database management we are proposing SQL based assignment as nearly 50% of the jobs are having SQL in their requirements.

PROJECT

We understand that the course project is an integral part of the course. We are also including a one-course project in the proposed course as it gives a better idea about the overall subject.

SUPPLEMENTARY MATERIAL

This supplementary material is focused on ethics in the data science field. We are proposing to upload a note on this as an optional or supplementary material for additional read to students. Based on the research of different universities' data science curriculum we have decided to put this in the proposed course. We think that as an engineer, ethics are necessary in this dynamic and cutthroat competitive world.

CONCLUSION: As we can see that with the proposed syllabus, we have covered all the top skills which we are getting from our analysis. So, after completion of the course, we as students can be able to fit **for junior (or entry) level jobs**. We have changed only two lectures in the current curriculum which shows that the current course is nearly perfect. (Nothing is perfect in the world).

5. Part 2 - Data Science program curriculum design [Master of Data Science and Artificial Intelligence]

Now that we have designed a Data Science-based Course to gain an insight into the field of Data Science, we proceeded to create a curriculum for a Master of Data Science and Artificial Intelligence based on the Skills and Qualifications extremely relevant for the Job Market.

In order to find and visualize the essential skills and qualifications for building the Master's Program we applied a 3-Step Process, wherein

Step 1: Classification of Job titles based on Salaries to find the key qualifications.

Kaggle dataset gave separate job titles which were categorized into two different categories "HIGH-LEVEL JOBS" and "MEDIUM LEVEL JOBS". We get a Box plot from the Kaggle dataset, which is present in Appendix A.

We see that "Data Scientist", "Project Manager" and "Data Engineers" have higher salary compared to the rest of the job postings. So, we consider them in High-Level Jobs. Also "Business Analyst" and "Data Analyst" have lower salaries so we consider them in Mid-Level Jobs.

Step 2: Getting the word cloud for the greatest number of occurring skillsets in the job descriptions.

Using the job descriptions in these scrapped data we created a word cloud to see what words are frequently occurring in these job postings. With this word cloud and with our knowledge about data science we define a set of skills to look for in a job posting. Then this skill list we search the job description.

Step 3: Building the hierarchical clusters from the skillsets obtained from the word cloud.

With these skills we perform Hierarchical clustering by assigning distance values to skills. If a skill occurs in a specific job posting, we will assign a distance of 0 and we assign a distance of 1 when the skills are not occurring in the job posting. We get 2 separate hierarchical clusters one for High-level jobs and one for Mid-level jobs.

From Medium level job postings, we get the following clusters

- Analytics; Data mining and SQL; Databases; Machine Learning and Optimization; Statistics; Neural Network and Python; Cloud; Big Data; Algebra; Data Visualizations and Web Scraping

From High-level job postings, we get the following clusters

- Machine Learning; Neural Network and Python; Data mining and Analytics; Database and SQL; Cloud; Big Data and Optimization

Apart from technical subjects we must develop soft skills. We have excluded the following soft skill before implementing hierarchical clustering, co now we will address these topics

- Decision Making; Team Building; Presentation; Project Management; Leadership; Consulting

Out of these 6 topics we can have 4 subjects namely

- Management& Technology Consulting
- Leading for Group organizations
- Project Management
- Soft Skills &Presentations

Thus, the subjects identified to be included in the curriculum of the program are:

Technical:

1. Introduction to Data Analytics: Basic Methods
2. Database and SQL in data science
3. Mathematics in data science
4. Neural network and deep learning
5. Introduction to cloud computing
6. Foundation of big data
7. Advanced SQL for business intelligence and analytics
8. Advanced cloud computing
9. Data Science
10. Modeling and optimization for big data
11. Optimization for machine learning

Apart from this we went through other university website and identified the following subjects 5 subjects

1. Reinforcement learning

2. Trustworthy Machine Learning
3. Business analytics
4. Time series analysis and forecasting
5. Operations research and management for data science

Course duration: 1 year (full time) or 2 years (extended full time)

Degree requirements:

1) 10 courses (6 minimum technical) **OR**

2) 7 courses + project (industrial-based- under a professor) (3 technical and 4 non-technical courses)

Pre-requisites: undergraduate knowledge in mathematics and statistics

Since data science can be applied to a wide variety of domains there is no hard-prerequisite set.

Courses are divided into 3 categories:

- **Mandatory courses M**
- **Core courses C**
- **Elective courses E**

Total 17 courses: 2 mandatory courses; 8 core courses ; 10 elective courses

Graduation requirements: 2M+3C+5E

It is recommended to do 4 soft skill/business elective courses

In case a student is opting for an internship /project option it is compulsory for the student to complete 2 soft skill/business courses before the start of the internship.

6. Part 3- Visualizations of Course Curriculum

Data visualization is the graphic representation of data. It involves producing images that communicate relationships among the represented data to viewers of the images. This communication is achieved using a systematic mapping between graphic marks and data values in the creation of the visualization.

6.1 Master of Data Science and Artificial Intelligence Program

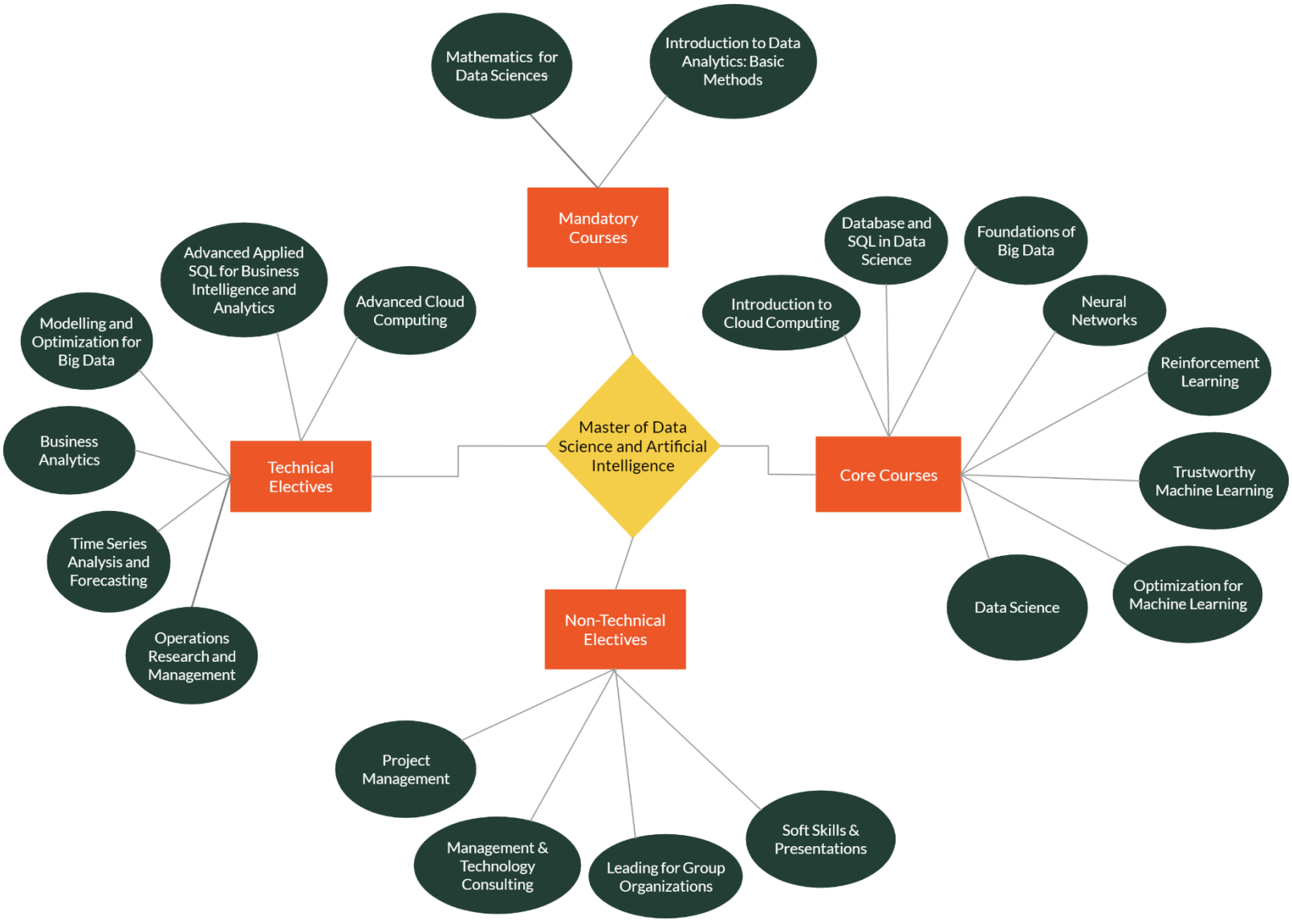

Fig 3- Master of Data Science and Artificial Intelligence Program Overview

6.2: Why "Master of Data Science and Artificial Intelligence"

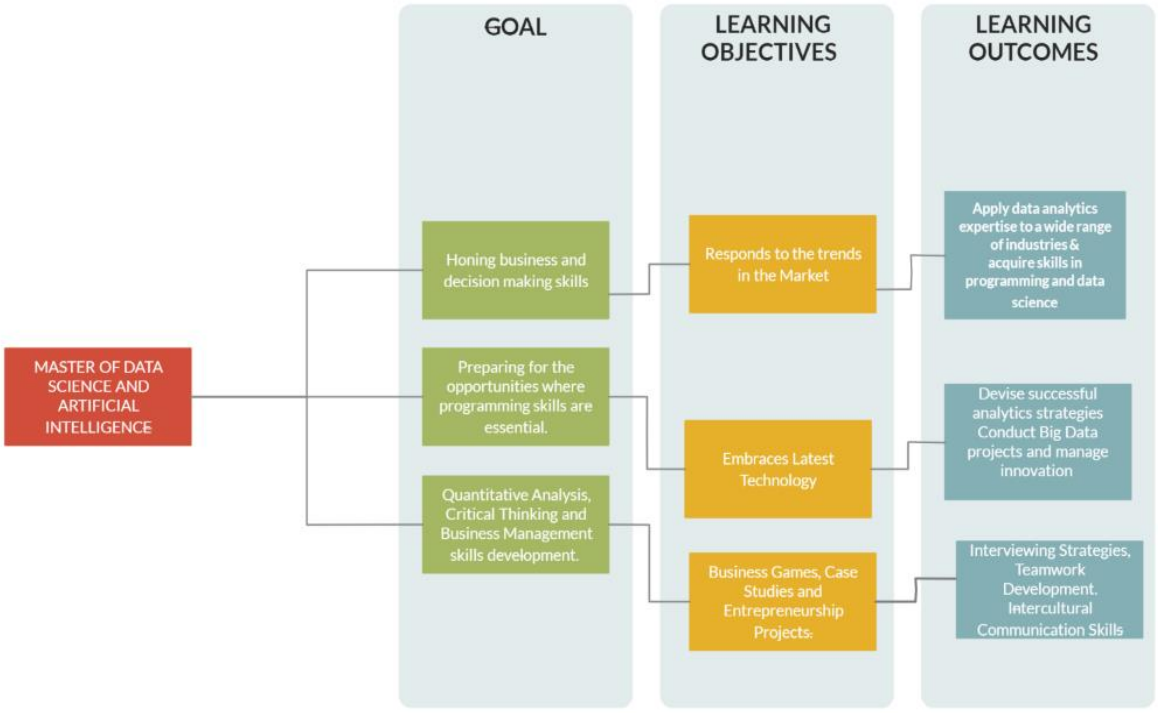

Figure 4- Overview of Degree Program Need

6.3 Course Curriculum Visualizations

MANDATORY COURSES:

1. **Introduction to Data Analytics: Basic Methods**

Course curriculum inspiration: Ryerson University, Data Analytics: Basic Methods

Figure 5- Visualization of the course Introduction to Data Analytics: Basic Methods

CORE COURSES:

1. Database and SQL in Data Science

Course curriculum inspiration: O'Reilly, Advanced SQL for Data Scientists; Udacity -SQL for Data Analysis.

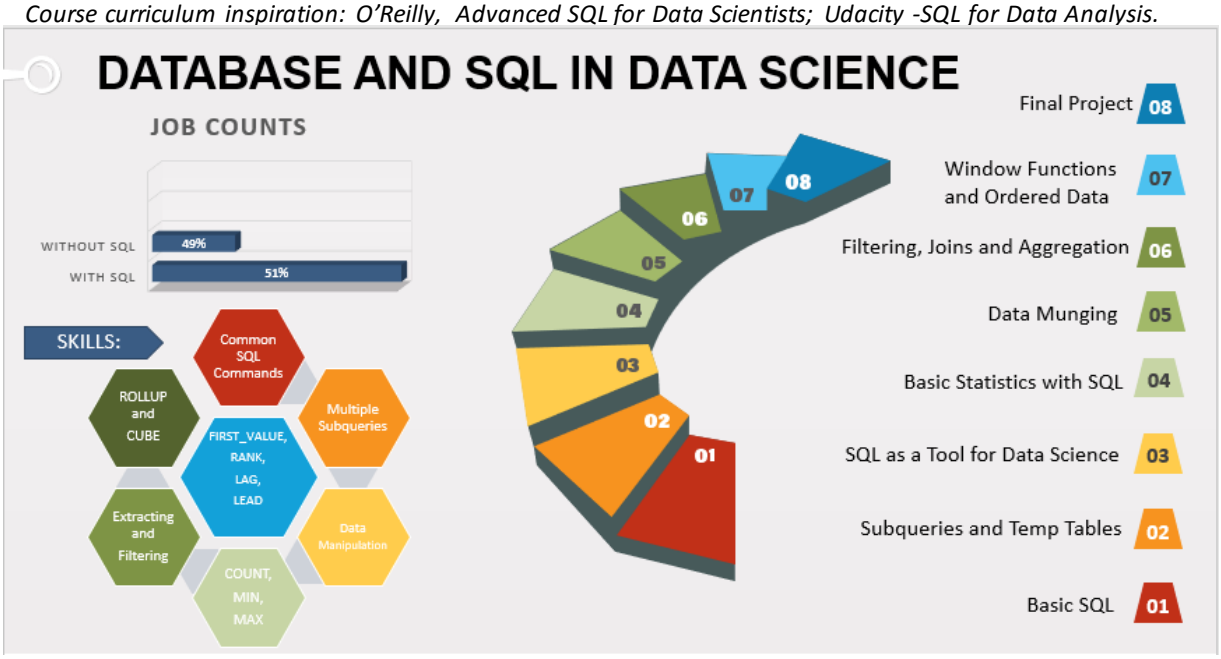

Figure 6- Visualization of the course Database and SQL in Data Science

ELECTIVE COURSES:

1. **Advanced Applied SQL for Business Intelligence and Analytics**

Course curriculum inspiration: O'Reilly, Advanced Applied SQL for Business Intelligence and Analytics; LinkedIn Learning, Advanced SQL for Data Scientists.

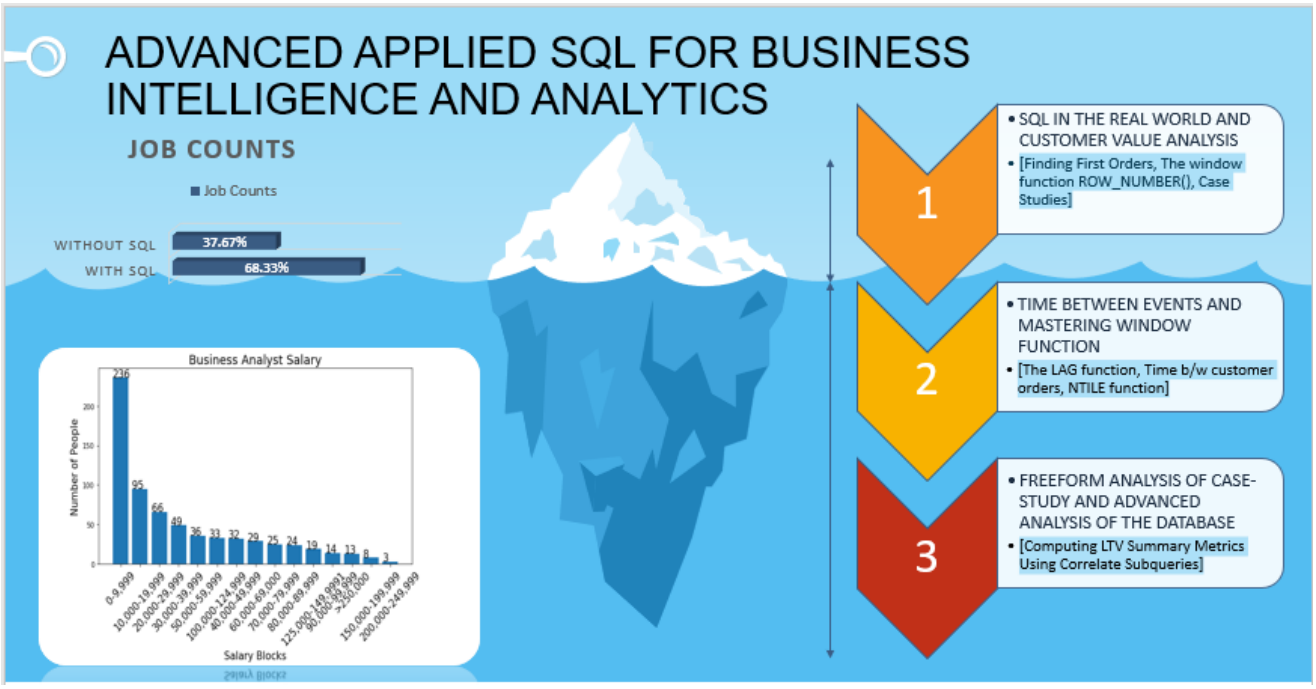

Fig 7- Visualization of the course Advanced Applied SQL for Business Intelligence and Analytics

7. Part 4 - University of Toronto -Start-up Finder for Data Scientists (SFDS)

Data Science Warriors

Even though academic curriculums intend to provide unique qualities in data analytics, it does not have a major contribution in the phase of job search. It may be challenging for students to find the right job position that would require the provided academic qualifications. One of the reasons can be the complexity in big companies, where the number of employees is very high. Tasks may get mixed up amongst coworkers or the position itself may not consist of the desired qualities obtained from the master's program.

In this case, this may bring the following question: **What can U of T offer to help their students find more accurate jobs within shorter time frames?**

In order to minimize such problems about employment opportunities, current students and graduates of Master of Data Science and Artificial Intelligence (MDSAI) will be able to access a brand-new portal named SFDS without paying any additional fees. This portal will allow only companies that are considered start-ups to post job openings for students. Companies will have to create a corporate account and fill out a start-up criteria assessment. They will be entering information that basically consists of company size, number of employees, revenue stream etc. This is then computed by a machine learning algorithm that evaluates responses and returns a conclusion on whether the company falls in the range of being a start-up or not. This assessment will be mandated to fill once every month in order to ensure that the company's start-up status remains unchanged. After passing this initial assessment, companies will then be able to create job postings that require answering more detailed questions. These will consist of factors like required skills, preferred background on specific courses and anticipated date of graduation. As a result, job postings with more specific and relevant criteria will be available for students to view.

From the students' perspective, they will be asked to create a profile that will also sync their academic history from ACORN. This will eliminate the process of entering the classes taken manually, which has the intention to spend more time on more specific information. For the rest of the profile, students will have two options:

1) Uploading Resume – Since most students tend to create a resume of their own, they can use this option to upload their resume and let a machine learning algorithm fill in the information as accurately as possible.
2) Manual Entry – Students can fill in the rest of the information in their desired ways.

Upon completion, students will have a profile with information on some criteria that are listed below:

- Academic History
- Obtained Skills
- Anticipated Graduation Date
- Areas of Expertise
- General Interest of Field/Industry
- Citizenship & Employment Status
- Professional Experience
- Group Projects
- Interested Skills to Practice
- Program's Impact on Student

They will also be given the opportunity to fill an optional survey that simply asks about the level of convenience and satisfaction the portal provides. Additional feedback can be written, especially about what other criteria to consider because the industry is rapidly growing, and students should be able to have an up to date portal that can still maintain its high level of accuracy in the job search. That is why their advice will be taken into deliberate consideration.

For the job browsing section, there will be an option to apply filters automatically, based on the individual's profile. This filtering process can also be applied manually, which is believed to provide students more information about different opportunities that they might be interested in more. Upon applying for a job, there will be a built-in chatting platform that will connect both sides automatically, after the company staff completes reading the application. Companies are also mandated to review all applications within a specific time period that will be determined by a specific machine algorithm. For example, if there are over 100 applications for a job posting, the company is required to finish all of them in 5 business days. In case of having less than 100, the algorithm would determine a shorter time period proportionally. Hence, students will have better information about when companies will respond to their applications, since the communication phase can be significantly spontaneous and disorganized. Last but not the least, company staff will have the option to video-call the applicant on the portal and save a significant amount of time from scheduling in-person interviews.

In addition to U of T students, this portal intends to be open to the public as well. Non-U of T students will be able to use this resource in exchange for a monthly fee. However, this portal will also offer a certificate program created by the Department of Data Science and AI (DSAI). Prospects will be able to take these courses during Fall, Winter or Summer sessions. Attending this online program will require a tuition fee, while eliminating the monthly fee requirement, and will provide all mandatory courses upon payment. Upon completion, they will be able to use the same profile to access the job board database.

In the long run, SFDS has a purpose of providing the right employment opportunities for future data scientists and reduce the level of stress a student can potentially suffer during their job search. Major websites like LinkedIn and Indeed have millions of job postings that are either outdated or inaccurate, which mostly causes issues in the filtering process. SFDS intends to solve potential recruitment-related issues that companies encounter and find the right applicant within a much shorter time frame. Therefore, start-ups can have a higher chance of succeeding in their business, while U of T students go through a much less stressful job search experience. As a result, U of T would also achieve much higher post-graduation employment statistics that can eliminate competition against other academic institutes.

8. References

[1] Columbus, L. (2020). IBM Predicts Demand For Data Scientists Will Soar 28% By 2020. https://www.forbes.com/sites/louiscolumbus/2017/05/13/ibm-predicts-demand-for-data-scientists-will-soar-28-by-2020/#6f41ea977e3b

[2] Hal Varian on how the Web challenges managers. (2020). https://www.mckinsey.com/industries/technology-media-and-telecommunications/our-insights/hal-varian-on-how-the-web-challenges-managers

[3] Analytics, Data Science and Statistics:. (2020). https://towardsdatascience.com/analytics-data-science-and-statistics-f9f140d731ba

[4] LinkedIn Data Reveals the Most Promising Jobs and In-Demand Skills of 2018. (2020). https://blog.linkedin.com/2018/january/11/linkedin-data-reveals-the-most-promising-jobs-and-in-demand-skills-2018

[5] Data Science. (2020). https://www.ibm.com/analytics/data-science

[6] Sharma, H. (2020). What Is Data Science? A Beginner's Guide To Data Science | Edureka. https://www.edureka.co/blog/what-is-data-science/

[7] edX, HarvardX (2020) https://www.edx.org/professional-certificate/ibm-data-science

[8] Zhang, V., & Neimeth, C. (2020). Why data science and machine learning are the fastest-growing jobs in the US. https://www.infoworld.com/article/3259891/why-data-science-and-machine-learning-are-the-fastest-growing-jobs-in-the-us.html

[9] How Much Data Does The World Generate Every Minute?. (2020). https://www.iflscience.com/technology/how-much-data-does-the-world-generate-every-minute/

[10] Data Science Trends in 2020 - DATAVERSITY. (2020).

https://www.dataversity.net/data-science-trends-in-2020/

[11] Demand for data scientists is booming and will only increase. (2020). https://searchbusinessanalytics.techtarget.com/feature/Demand-for-data-scientists-is-booming-and-will-increase

[12] Do you need a graduate degree for data science?. (2020) https://towardsdatascience.com/do-you-need-a-graduate-degree-for-data-science-8e3d0ef39253

[13] Do Data Scientists Need a Master's in Data Science?. (2020). https://quanthub.com/data-science-masters/

APPENDIX A: SUPPLEMENTAL FIGURES

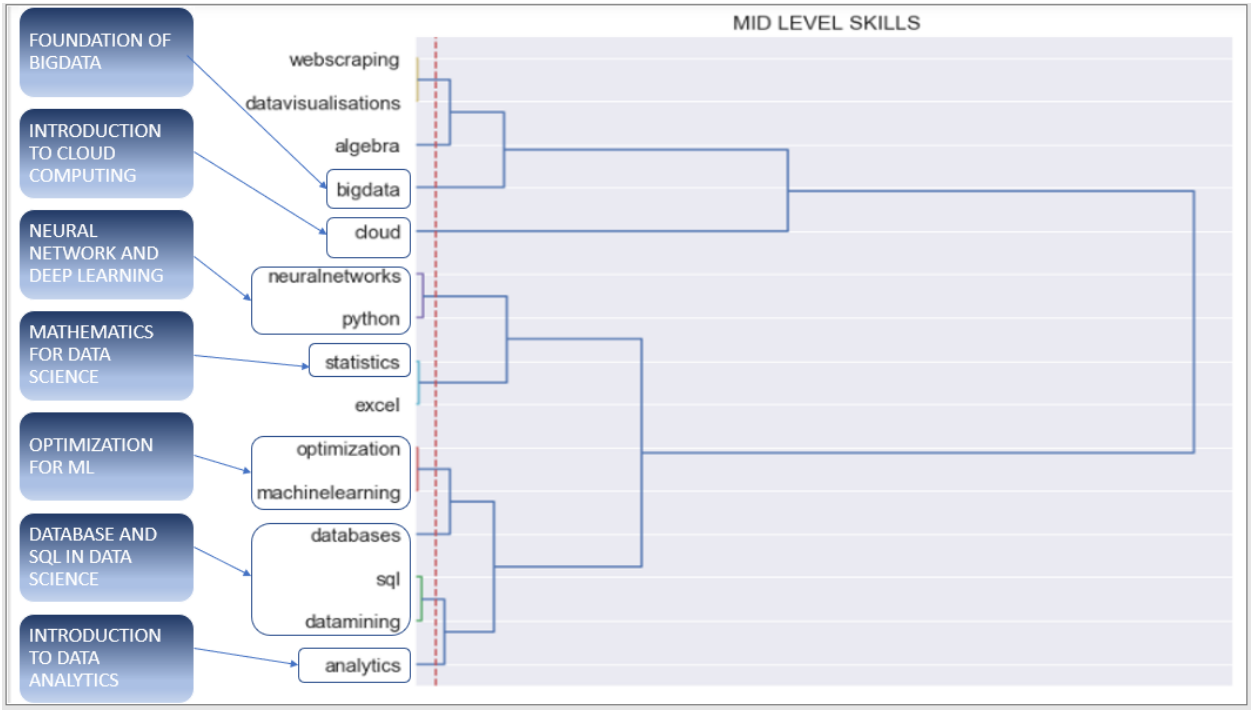

Fig 8 – Hierarchical Clustering for the Mid-level skillset

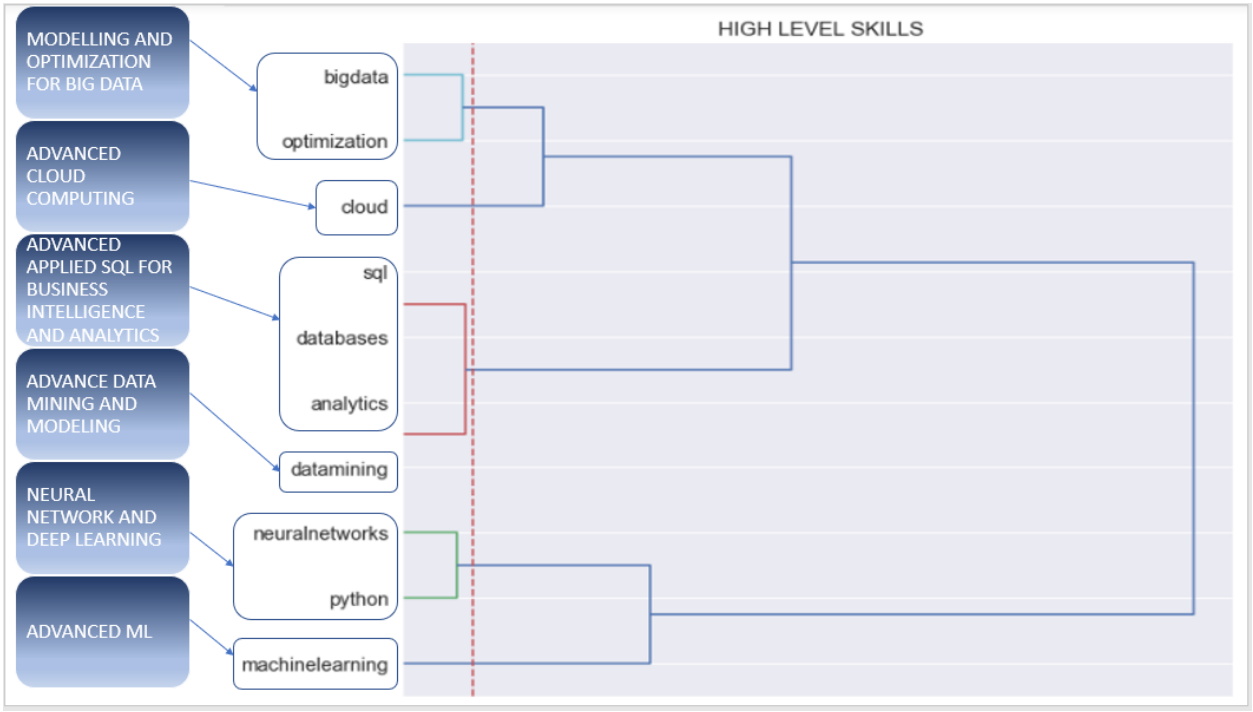

Fig 9 – Hierarchical Clustering for the High-level skillset

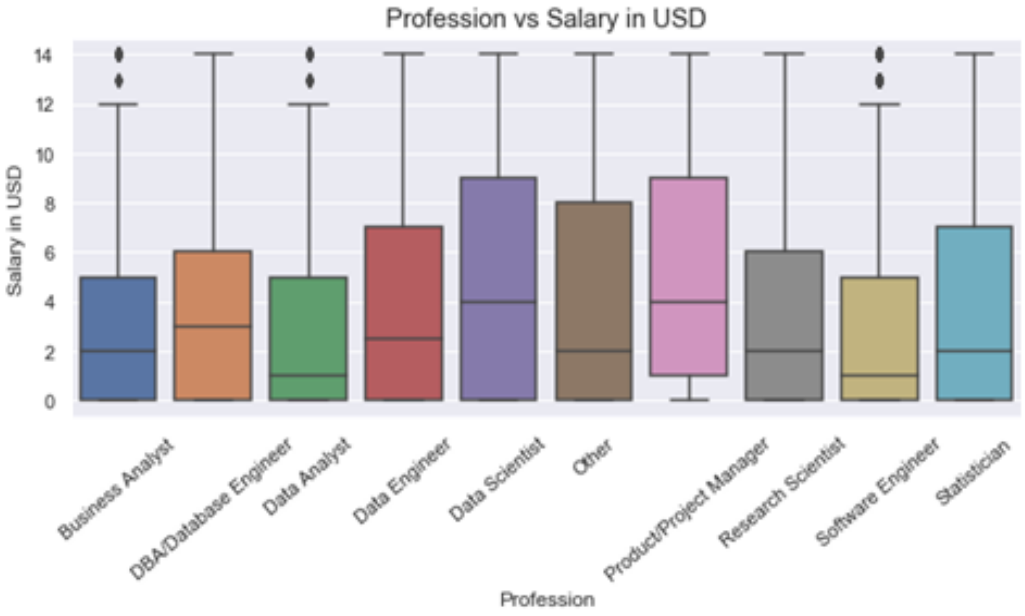

Fig 10 –Boxplot of salary buckets according to the Job Roles

Fig 11 – Word cloud for the skill set to be obtained

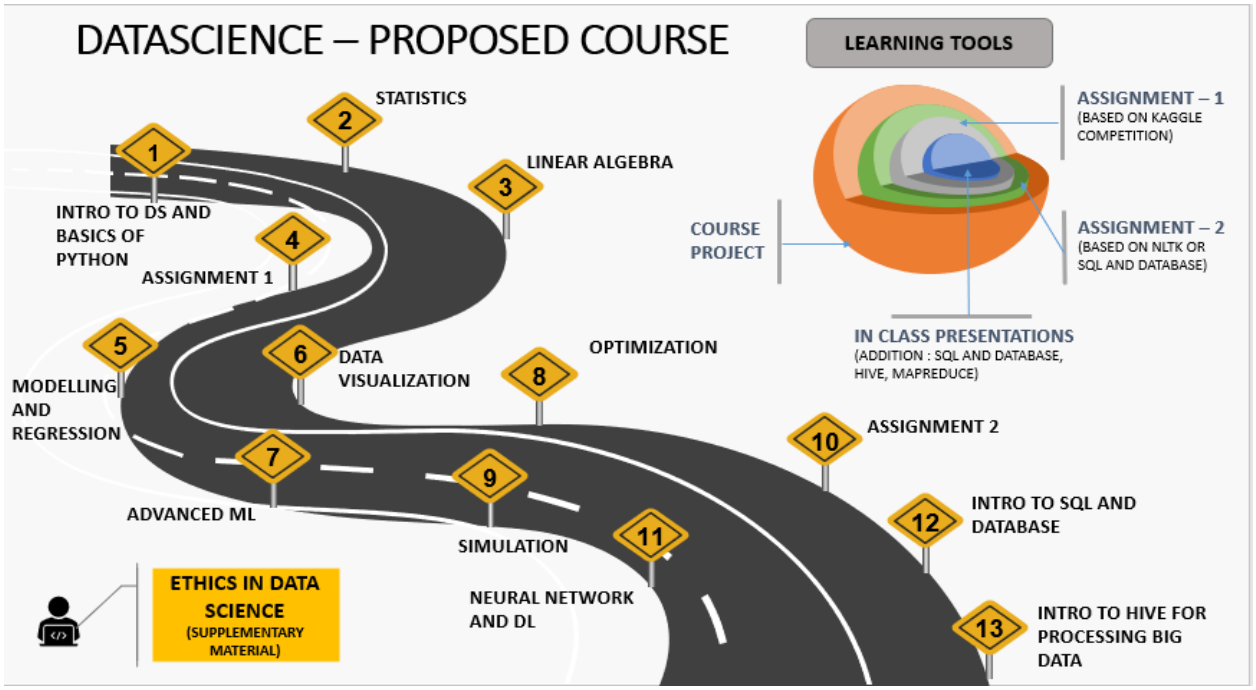

Fig 12 – Roadmap for the proposed Data Science course redesign

Introduction to Cloud Computing

Course curriculum inspiration: The University of Toronto, Introduction to Cloud Computing; Harvard University, Cloud Computing; Carnegie Mellon University, Cloud Computing

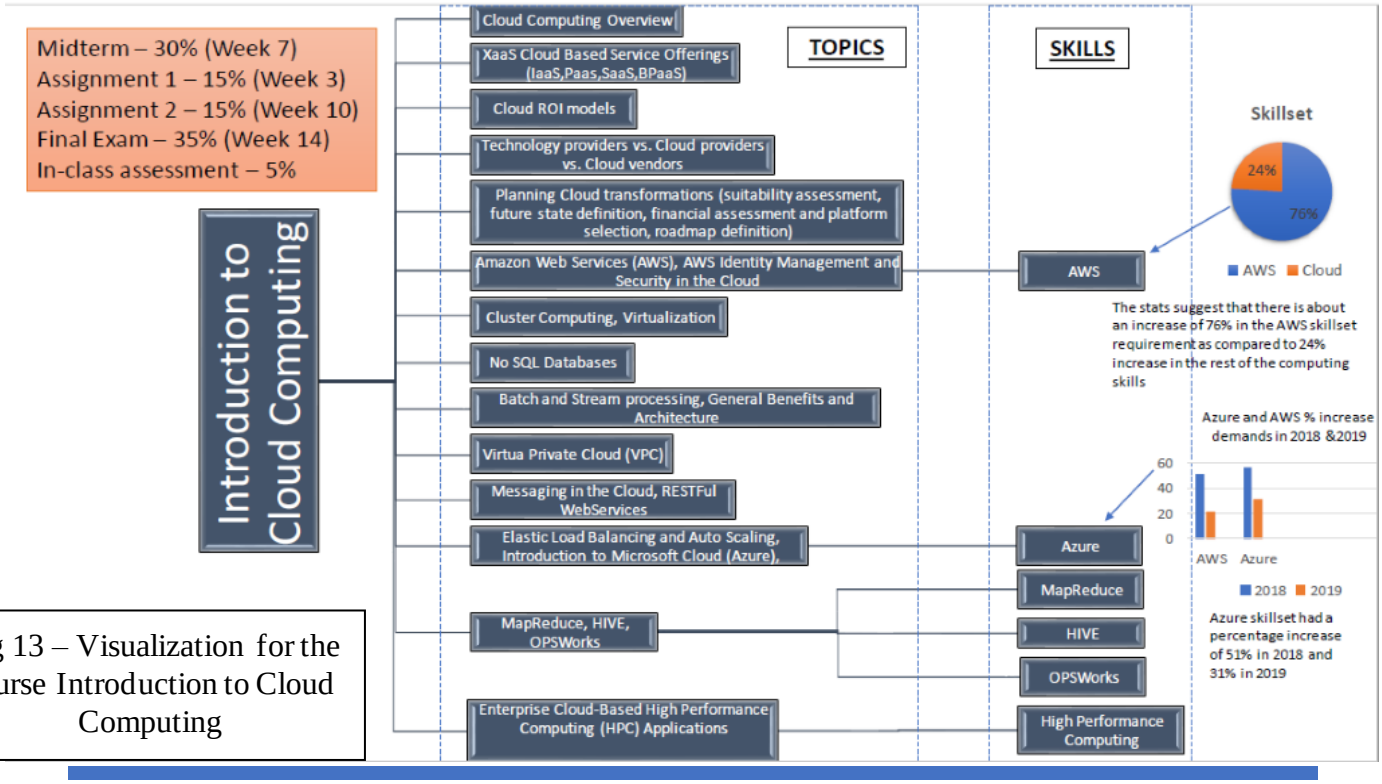

Fig 13 – Visualization for the course Introduction to Cloud Computing

Mathematics for Data Sciences

Course curriculum inspiration: KDnuggets, Essential Math for Data Science 'Why' and 'How'

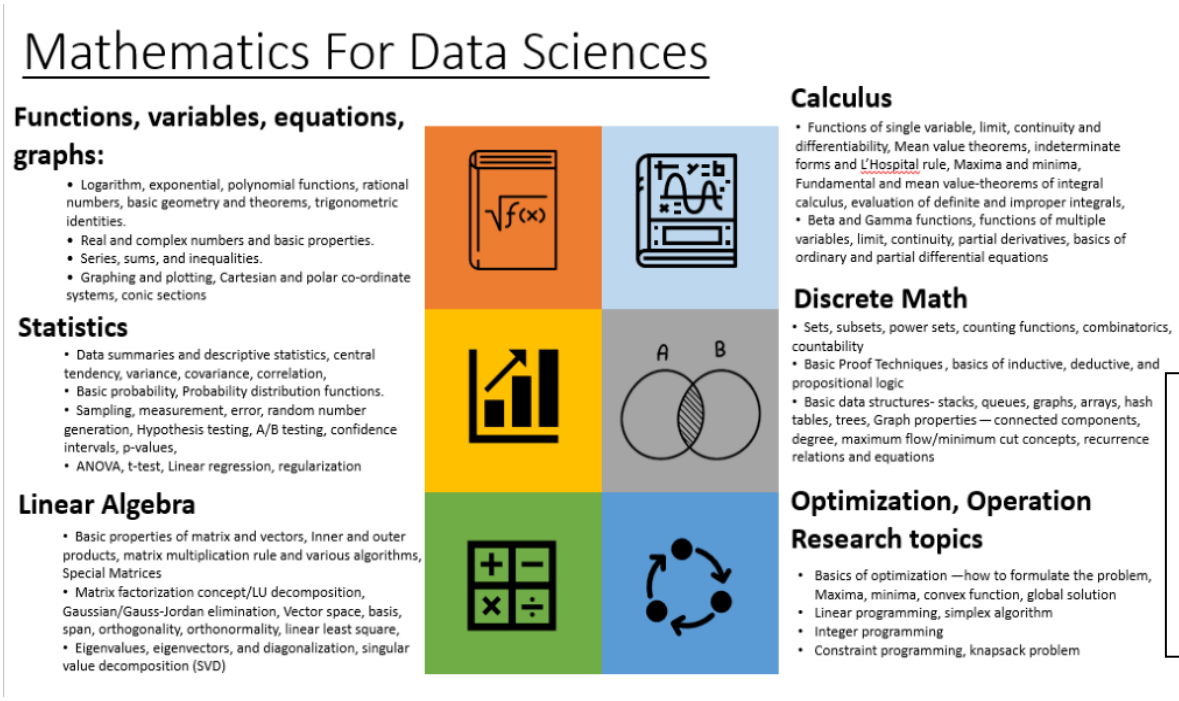

Fig 14 – Visualization for the course Mathematics for Data Science

Neural Networks and Deep Learning

Course curriculum inspiration: University of Toronto, Machine Learning and Data Mining, Introduction to Machine Learning, Inference Algorithms and Machine Learning, Intro to Neural Networks and Machine Learning

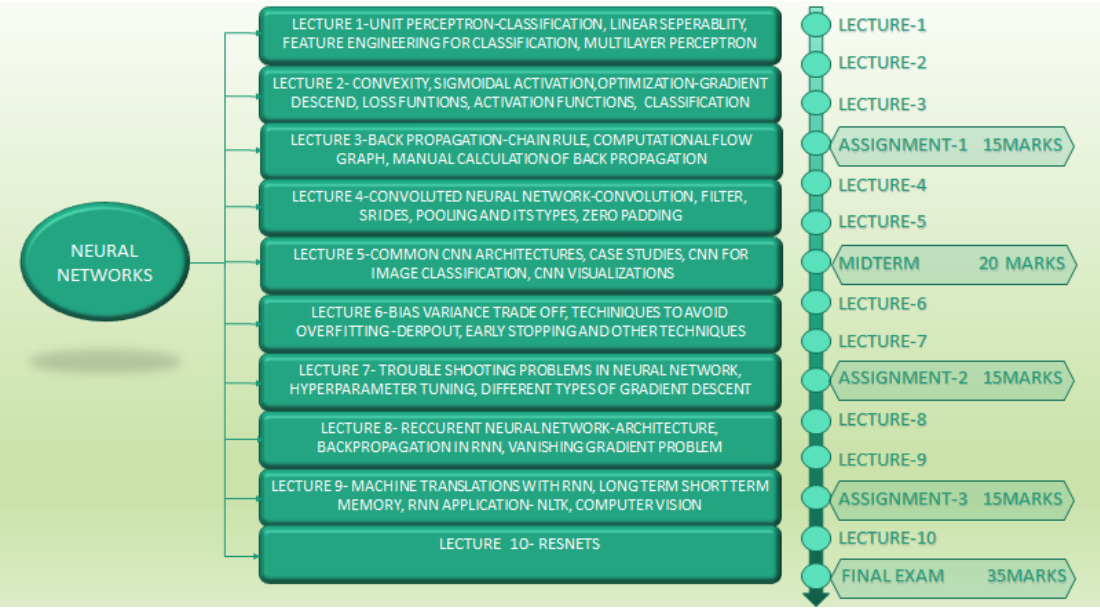

Fig 15 – Visualization for the course Neural Networks and Deep Learning

Foundations of Big Data

Course curriculum inspiration: SimpliLearn, Big Data Hadoop Certification Training course; Ryerson University , Data Science and Analytics; McMaster University, Big Data Analytics

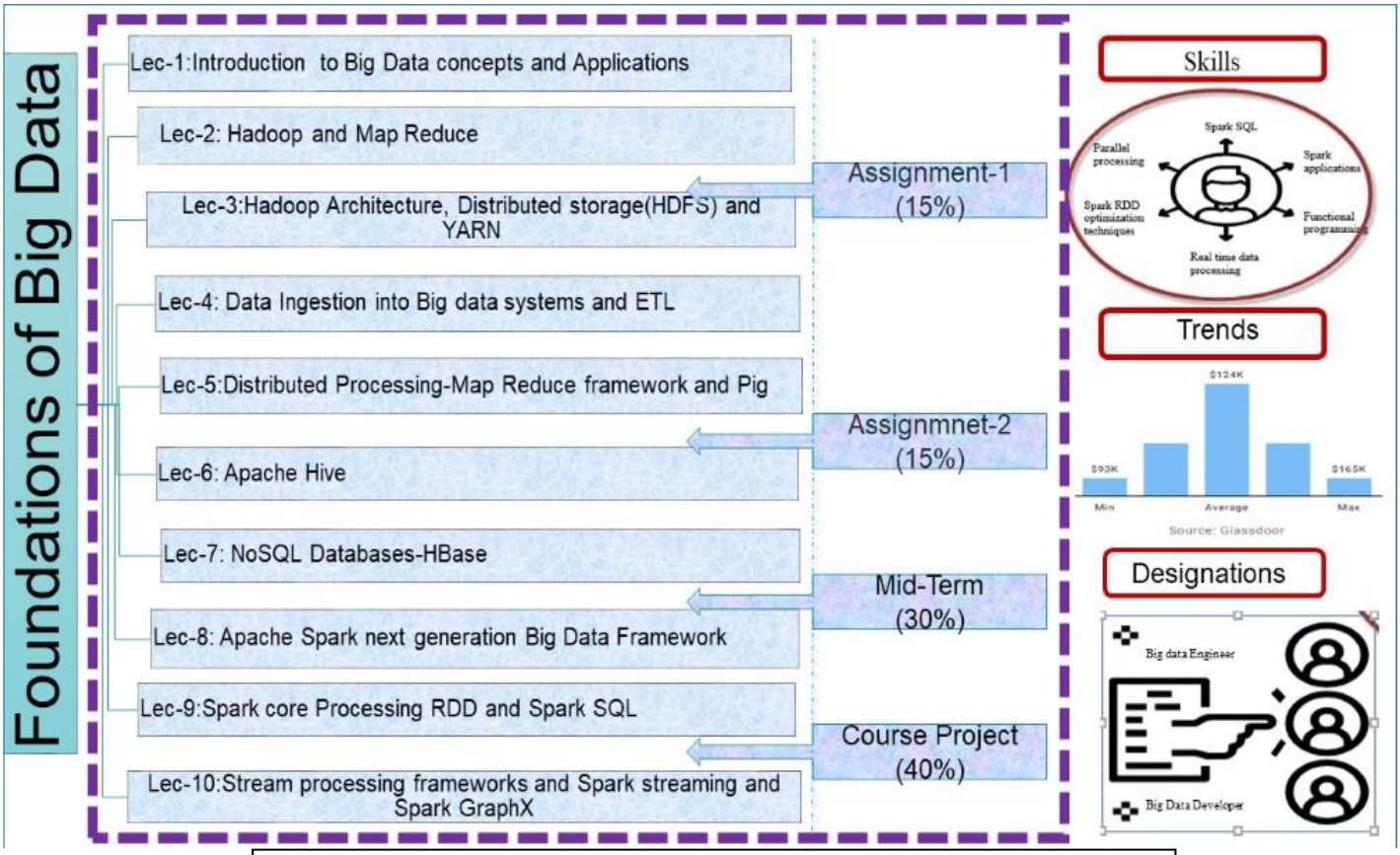

Fig 16 – Visualization for the course Foundations of Big Data

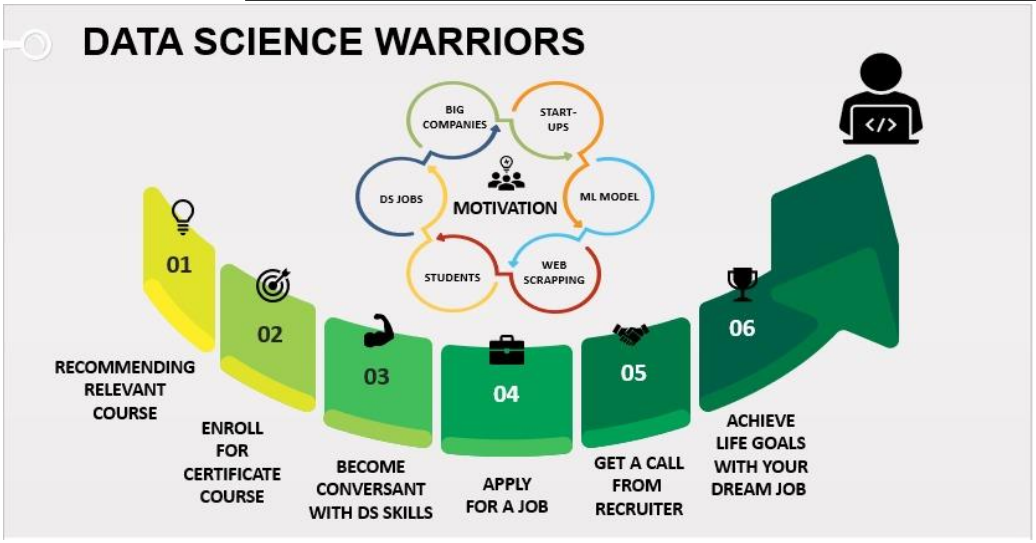

Fig 17 – Map for the proposed EdTech Startup idea

APPENDIX B- Detailed Curriculum of the courses in the MDSAI Program.

NO	Course Name	Topics Covered (Syllabus)	Level
1	Introduction to Cloud Computing	Cloud Computing Overview, XaaS Cloud-Based Service Offerings, Cloud ROI models , Technology providers vs. Cloud providers vs. Cloud vendors , Planning Cloud transformations (suitability assessment, future state definition, financial assessment and platform selection, roadmap definition), Amazon Web Services (AWS), AWS Identity Management and Security in the Cloud, Cluster Computing, Virtualization, No SQL Databases, Batch and Stream processing, General Benefits and Architecture, Virtua Private Cloud (VPC), Messaging in the Cloud, RESTFul WebServices, Elastic Load Balancing and Auto Scaling, Introduction to Microsoft Cloud (Azure), MapReduce, HIVE, OPSWorks, Enterprise Cloud-Based High-Performance Computing (HPC) Applications.	C
2	Advanced Cloud Computing	Cloud 101: Fundamentals, Cloud service models, Data centers, Cloud infrastructure, VM and containers, Cloud infrastructure, Programming Models and Frameworks II, Storage in the cloud I, Storage in the cloud II, Tail latency & interference, Geo-replication, Mobility and the Cloud, Key-Value Stores, Scheduling I, Scheduling II, Reliability & fault tolerance, Diagnosis via monitoring & tracing, Cloud Computing and ML/AI, Cloud IoT and Edge.	E
3	Soft Skills &Presentations	Short Technical Talk, One-minute Introduction, Hand-drawn Illustration or Diagram, XY Graph of Data, A three-minute, three-slide presentation about an important aspect, A PowerPoint Illustration or Diagram, One-minute Technical Talk, A Chalk Talk, One-minute Toast ,Two written reflections on the process of preparing and delivering a presentation to influence others.	E
4	Project Management	Introduction, Defining and scoping a project ,Project customers, Project charter, Purpose vs Objective, Organizational strategy, structure and projects , Strategic project management, Project selection, Building a project team, Leadership and management, Negotiation and organizational culture, Scheduling and Planning,	E

		Budgeting, Resources, Risk management, Performance management ,Project closure Audit, Careers in Project Management.	
5	Leading for Group organizations	Introduction to the foundation for leading effectively, Leading, managing and following, Process of leading, The work of the leaders, Leadership skills, Leader Character, Behavior and Leadership styles, High-Performance work systems, High-Performance business organizations, Vitality and virtue of the organizations, Glance at the course.	E
6	Management& Technology Consulting	The Changing Consulting Industry, Consultants: Types, Skills and Values, Consulting as a Profession, Marketing and Selling of Consulting Services, Discussion of Data Gathering Methods, Skills for Success in Management Consulting, Strategic and Organization Information Technology Consulting, Strategy in Organizations Consulting, Management Consulting in Context, Strategic Marketing Consulting, Analyzing and Framing Problems, Strategy and Operations Management Consulting, Human Resources in Organization Consulting, Managing Engagements, Managing Consulting Firms - The Knowledge Sharing Problem, The Future of Consulting.	E
7	Introduction to Data Analytics: Basic Methods	Introductory Statistics, Introduction to Python Programming, Basics of Linear Algebra, Introduction to Data types, Reconfiguration& Visualization, User-defined functions and Error Analysis, Probability and Distributions, Introduction to Sampling Distributions, Regression, Introduction to Correlation: A Statistical parameter test, Testing of hypotheses.	M
8	Advanced Applied SQL for Business Intelligence and Analytics	SQL in the real world and Customer value Analysis, Time between events and Mastering Window Function, Freeform analysis of case study and advanced analysis of the database.	E
9	Database and SQL in Data Science	Basic SQL, Subqueries and Temp Tables, SQL as a tool for Data Science, Basic Statistics with SQL, Data Munging, Filtering, Joins and Aggregation, Window Functions and Ordered Data.	C
10	Foundations of Big Data	Introduction to Big Data concepts and applications, Hadoop and MapReduce, Hadoop Architecture, Distributed Storage(HDFS) and YARN, Data Ingestion into Big Data systems and ETL, Distributed processing-MapReduce Framework and Pig, Apache Hive, NoSQL Database- HBase, Apache Spark next-generation Big	C

		Data Framework, Spark core processing RDD and Spark SQL, Stream processing frameworks, Spark Streaming and Spark GraphX.	
11	Modeling and Optimization for Big Data	Introduction to Big Data, Big Data Modelling and management systems, Big Data Integration and processing, Machine Learning with big data, Introduction to optimization algorithms, Smooth convex optimization, Non-smooth convex optimization, Stochastic optimization, Optimization perspective in machine learning case studies.	E
12	Neural Network and Deep learning	Unit perception- Classification, Linear separability, Feature Engineering for classification, Multilayer Perceptron, Convexity, Sigmoidal Activation, Optimization – Gradient Descent, Loss Functions, Activation Functions, Classification, Back Propagation-Chain Rule, Computational Flow Graph, Manual Calculation of Back Propagation, Convoluted Neural Network – Convolution, Filter, Strides, Pooling and its types, Zero Padding, Common chain Architectures, Case studies, CNN for image classification, CNN visualizations, Bias-Variance tradeoff, Techniques to avoid Overfitting- Dropout, Early stopping and other techniques, Troubleshooting problems in Neural Networks, Hyperparameter Tuning, Different types of Gradient Descent, Recurrent Neural Network-Architecture, Backpropagation in RNN, Vanishing Gradient Problem, Machine translations with RNN, Long Term Short Term Memory, RNN application-NLTK, computer vision, Resnets.	C
13	Reinforcement Learning	Fundamentals of Reinforcement Learning, The K-Armed Bandit Problem, Markov Decision Processes, Value Functions & Bellman Equations, Dynamic Programming, Monte Carlo Methods for Prediction & Control, Temporal Difference Learning Methods for Prediction, Temporal Difference Learning Methods for Control, Prediction and control with function Approximation	C
14	Trustworthy Machine Learning	Overview & motivation, Training-time integrity, Test-time integrity, Test-time integrity, Confidentiality of the model, Privacy attacks, Differential privacy, Confidentiality, Safety, Fairness & Ethics	C
15	Business Analytics	Customer Analytics-Descriptive Analytics, Predictive Analytics, Prescriptive Analytics, uncertainty, risk, People Analytics-Performance evaluation, staffing,	E

		collaboration, talent management, Accounting analytics-Ratios and Forecasting, Earnings Management, Big Data and Prediction Models, Linking Non-financial Metrics to Financial Performance	
16	Time Series Analysis and Forecasting	Basic Statistics, Stochastic process and its main characteristics , Visualizing Time Series, and Beginning to Model Time Series, Stationarity, MA(q), AR(q) processes, AR(p) processes, Autoregressive-moving average models ARMA (p,q), Yule-Walker equations, PACF, Coefficient estimation in ARMA (p,q) processes. Box-Jenkins' approach, Forecasting in the framework of the Box-Jenkins model. Non-stationary time series, Regressive dynamic models, Vector autoregression model and co-integration, Causality in time series.	E
17	Data Science	Introduction to Data Science and basics of python, Statistics, Linear Algebra, Modelling and Regression, Data Visualization, Advanced Machine Learning, Optimization, Simulation, Neural Network and Deep Learning, Intro to SQL and Database, Intro to Hive for processing Big Data	C
18	Mathematics for Data Science	Functions, variables, equations, Graphs, Statistics, Linear Algebra, Calculus, Discrete Math, Optimization, Operation Research topics	M
19	Operations research and management for data science	Operations Research and Management (ORM) Introduction, Strategy Development & Implementation in OM, Diffusion of Innovation, Work System and Product Development, Organizational Approach in Lean Enterprises, Maintenance Management, Cybersecurity in Data Science, Lean Product Development & Knowledge, Deploying & Tracking Operations Research, Total Quality Management	E
20	Optimization for Machine Learning	Introduction: Numerical Sets, Functions and Limits, Limits and Multivariate Functions, Derivatives and Linear Approximations: Single Variate Functions, Derivatives and Linear Approximations: Multi-Variate Functions, Linear and Integer Programming	C