EdTech Consulting Report

MIE1624 December 2020

Group 14:

Samuel Atkins: 1002951754

Marie Ishimwe: 1002608555

Rohith Sothilingam: 1005028355

Dijia Zhang: 1002162269

Yifan Zhang

Table of Contents:

Introduction

There are three main parts to this project. First, students are required to re-design the course curriculum for “MIE1624: Introduction to Data Science and Analytics”. Instead of using personal experience or authoritative opinions, we will make data-driven decisions to formulate this curriculum. We will perform an analysis using a variety of data sources to capture the skills that a successful data scientist requires. The second part of this project requires students to design a sequence of courses and curriculum for a technically and business-oriented program: “Master of Data Science and Artificial Intelligence” at the University of Toronto. The curriculum for this program will once again be based on the information extracted from the sources analyzed while re-designing the course curriculum for “MIE1624: Introduction to Data Science and Analytics”. The final part of the project prompts students to advance AI education by establishing an EdTech startup or through other means.

1. Course Curriculum Design

To formulate a curriculum for a re-designed MIE1624 our team extracted useful skill indicator data from a handful of data sources. Using this data, we visualized the skills most commonly sought after by employers as well as the most used skills by data scientists. Using this information, we were able to determine the content that needed to be present in our redesigned MIE1624. To group and order the content according to similarity, we implemented a clustering algorithm. Using this information, we designed a structured course that aims to introduce the various concepts required to succeed in the world of data science.

1.1 Data Analysis

1.1.1 Data Sources

Our group utilized three data sources for our analysis. We started with some open-source datasets available on Kaggle, specifically the 2018 Kaggle ML and Data Science Survey (Kaggle et al.) and the 2019 Kaggle ML and Data Science Survey (Kaggle et al.). Then, we performed web-scraping Indeed.ca, specifically the data science job postings, to obtain useful information regarding job descriptions and requirements in the field of data science.

1.1.2 Data Visualization: 2018 Kaggle Dataset

To understand the skills required and program masteries needed to succeed as a data scientist, our group first considered the previously mentioned Kaggle surveys. Tens of thousands of data scientists from a wide variety of backgrounds participated in each of the surveys. The questions asked in this survey, like “What data visualization libraries or tools have you used in the past 5 years?” and “Which of the following machine learning products have you used at work or school in the last 5 years?” shed light on some of the tools used and skills maintained by practicing data scientists.

Our group hand-picked questions in each of the datasets that would enhance our understanding of common data science experiences, thus allowing us to formulate a curriculum for the re-designed MIE1624. After parsing the 2018 and 2019 datasets, we visualized the most common skills required, tools and platforms used, and environments utilized in the field of data science. Note that soft skills were not included in our analysis because designing a curriculum around soft skills is not realistic. It makes more sense to require students to take communication and design courses. These soft skills are considered in-depth in Section 2. Figure 1 below illustrates the 30 most sought after skills according to the 2018 Kaggle dataset:

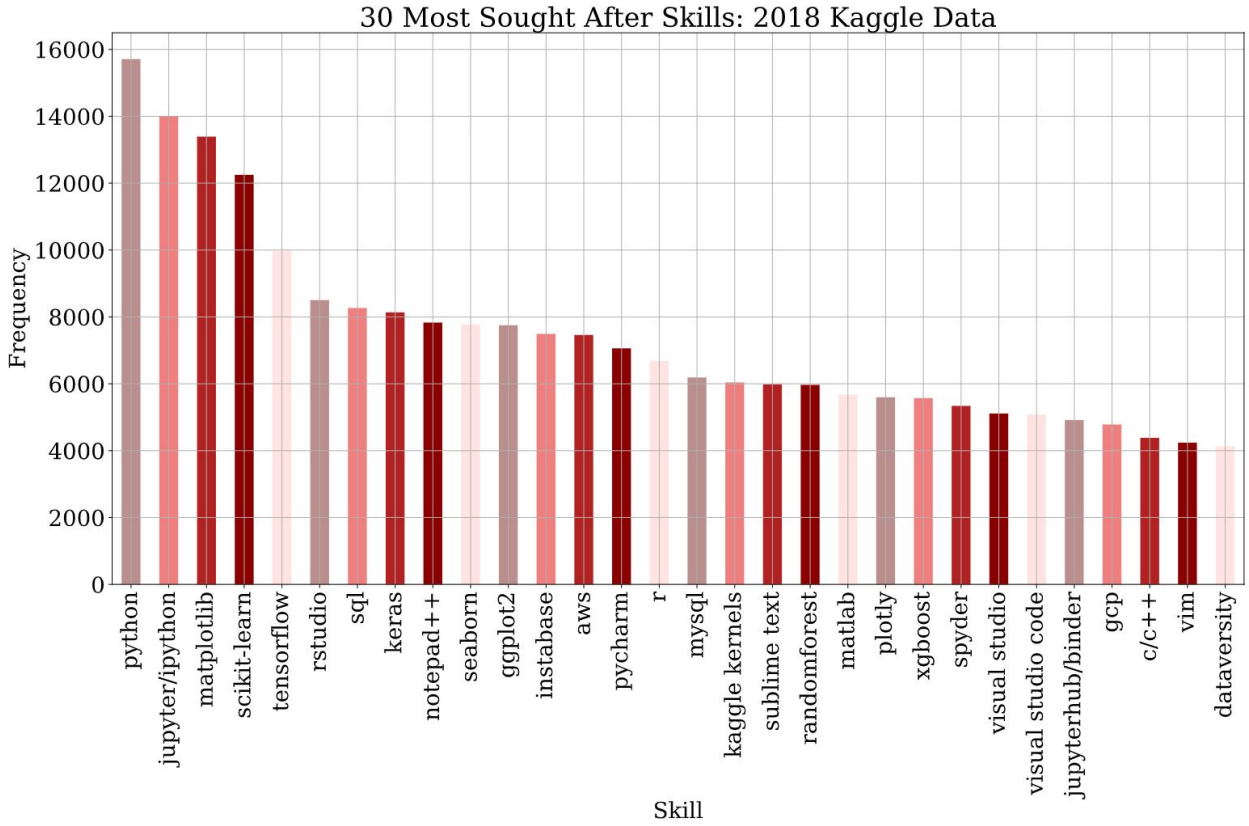

Figure 1: 30 most sought after skills according to the 2018 Kaggle dataset

From the above Figure we can see that the Python programming language reigns supreme. Many Python libraries, utilities, and environments are also present in the top 30 like iPython, matplotlib, scikit-learn, TensorFlow, Keras, Seaborn, and PyCharm. Other programming languages like SQL, R, and C/C++ are mixed in with these Python utilities. Clearly, in the world of data science, Python and applications/tools associated with Python are vital to the success of a data scientist, at least in 2018.

Given the plethora of skills outlined by Figure 1, we have a good starting point for which topics we should introduce in our MIE1624 introductory data science course. The next logical step is to group these topics so that we can begin to design modules and lectures for our curriculum. To group the topics outlined by Figure 1, we implemented an agglomerative clustering algorithm using Euclidean distance as our affinity metric and ward linkage. We chose to use 6 clusters because this was the minimum number of

clusters such that each cluster had at least a size of 2. The results of our clustering algorithm are displayed below:

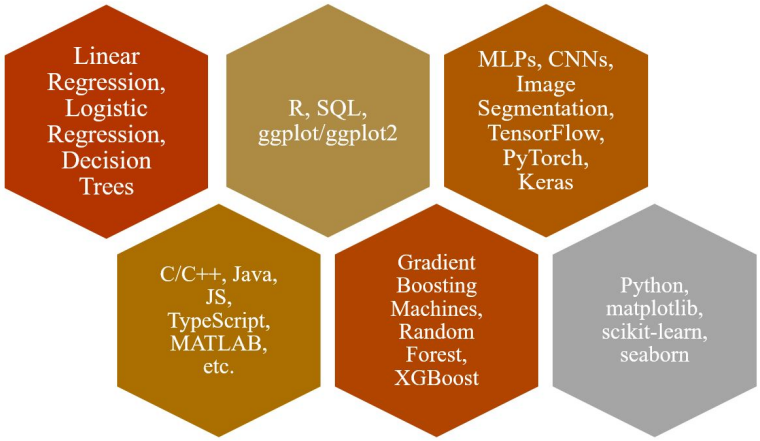

Figure 2: Results of agglomerative clustering algorithm on 2018 Kaggle dataset

These clusters provide insight into potential modules that could be implemented in an introductory data science course. A module could be dedicated to learning the basics of databases by introducing students to SQL and then to R and ggplot, as delineated by the cluster in the middle of the first row. Another module could be dedicated to data visualization in Python, in reference to the cluster containing Python, matplotlib, and seaborn. Other modules could be dedicated to learning fundamental modelling techniques like linear regression, logistic regression, and decision trees. The results of this clustering algorithm will inform our curriculum design in the later parts of this report.

1.1.3 Data Visualization: 2019 Kaggle Dataset

After analyzing the 2018 dataset, we moved on to the 2019 dataset. This dataset had a similar number of records and questions. By hand-picking questions expected to yield useful information, we were able to visualize highly sought after skills in 2019. Figure 3 in Appendix 1.1 visualizes the most sought after data science masteries, according to the 2019 Kaggle dataset. Similar themes are present. Python, and libraries, environments, and tools created for Python analysis like matplotlib, seaborn, TensorFlow, and PyTorch, are the most prevalent. Given that the questions asked in this dataset are different, however, some new skills have cropped up. Modelling architectures like “linear or logistic regression”, “decision trees or random forests”, “convolutional neural networks”, "random forest", “xgboost”, and “dense neural networks” are common replies to one or more of the questions in the survey. Furthermore, general data science tools like “amazon web services (aws)”, “image/video tools”, and “google cloud platform (gcp)” also appear frequently as responses to the questions posed by this survey.

Just as with the 2018 dataset, we used the exact same agglomerative clustering algorithm to cluster the skills illustrated by Figure 3. Interestingly enough, by using the same number of clusters we obtained the exact same clusters. These clusters are included in Figure 4 of Appendix 1.2. The results of Figure 4 cultivate credibility for the potential modules suggested by the clusters illustrated in Figure 2.

1.1.4 Data Visualization: Indeed Web-Scraping

After dealing with the 2018 and 2019 Kaggle datasets, we decided to analyze the job market to gain some insight into which skills employers, as opposed to employees, value the most. To do this, we scraped Indeed.ca. Specifically, we extracted all of the text data in the job description field of all of the postings that were returned when searching for "data analyst, data scientist" in the Indeed search bar. Then, the job description text data was further parsed to extract the skills candidates were expected to possess. Figure 5 below illustrates the 30 most sought after skills according to Indeed.ca:

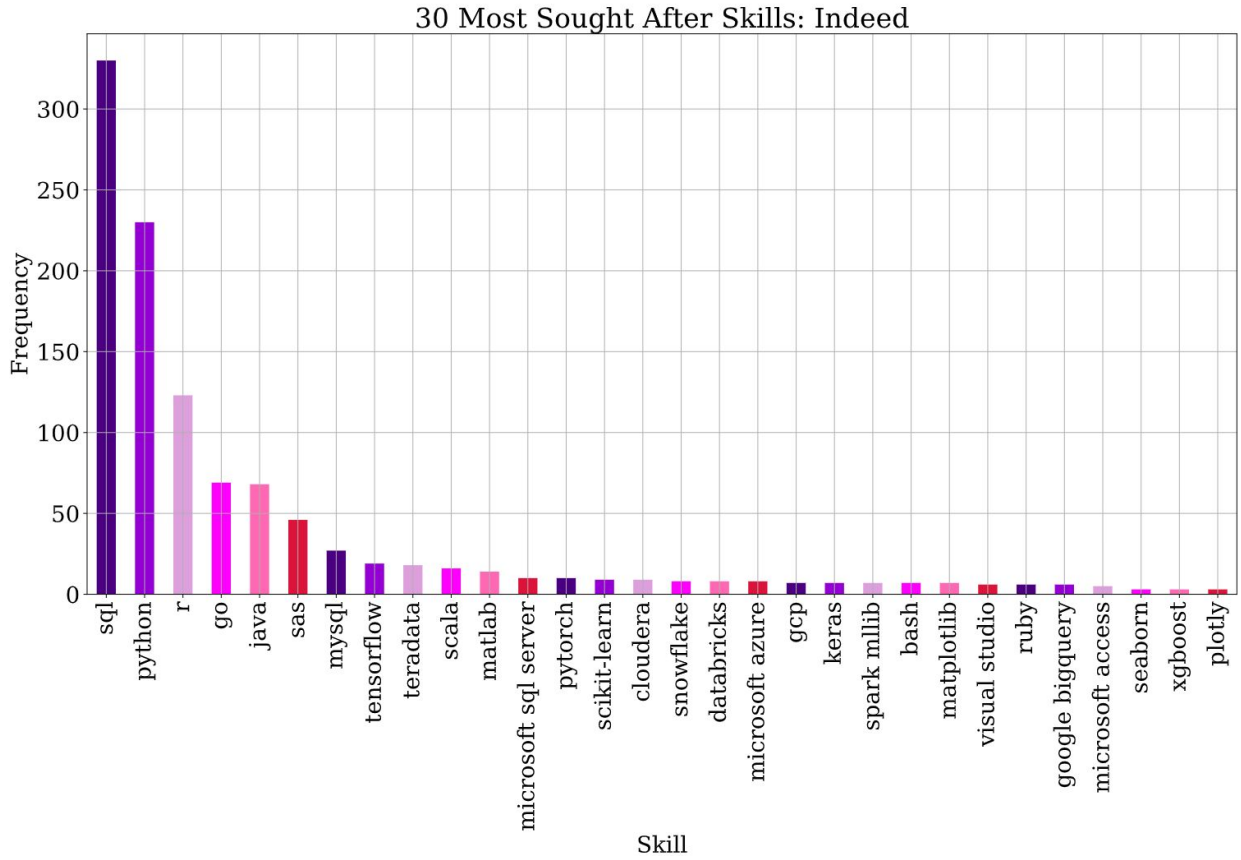

Figure 5: 30 most sought after data analyst/scientist skills according to Indeed.ca

From the plot above we can see that the employers on this website don't specify which model architectures candidates should be familiar with. The most commonly mentioned skills in the job description section are programming languages. Python and Python libraries, however, are once again very prevalent in this context. Agglomerative clustering was once again applied, this time to the web-scraped Indeed data. Since there is significantly less data in this dataset, the clusters obtained through this algorithm did not assist our module delineation. Nonetheless, the visualization illustrated in Figure 5 still provides tremendous insight into the skills that employers desire.

1.2 Course Curriculum Design

1.2.1 Course Logistics

Assuming that lectures start on September 8th and end on December 1st, there are approximately 12 weeks of class. To ensure that our course is operating under the same constraints that the current MIE1624 is operating under, we will assume that our redesigned MIE1624 has been allotted 10 3-hour lectures and 12 tutorials. Moreover, we will assume that we can hold in-person lectures, tutorials, mid-terms, and finals because factoring in the COVID-19 pandemic introduces additional difficulties.

1.2.2 Course Modules Based on Clustering

Our course design will heavily rely on the data gathered in the previous part. The clustered skills defined by Figures 2 and 4 can be segregated into concepts that students must understand and tools that students must be familiar with. The segregation of the text into these categories is illustrated by Table 1 in Appendix 1.1. Some dependencies are becoming apparent. Students cannot, for example, design text classification algorithms if they do not first understand how to appropriately clean data. Further, students cannot be expected to implement linear/logistic regression models without first understanding underlying statistical assumptions that these models are based on. These relationships can be best expressed using a directed graph. Figure 6 concisely illustrates the relationships present in Table 1:

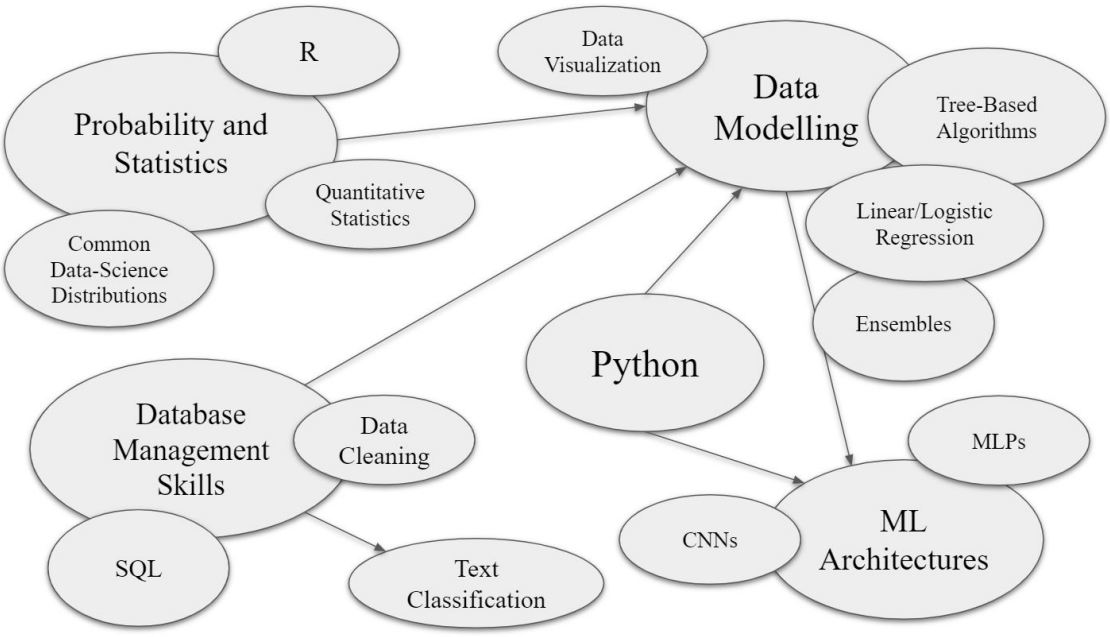

Figure 6: Directed graph illustrating the relationships between the clusters in Table 1

The above graph enforces an order in which these topics must be taught. Assuming that we wish to at least introduce all of the above topics, we must make some assumptions about the students that are taking MIE1624. Many of these topics require strong mathematical and statistical fundamentals. We cannot dedicate a significant amount of time to teaching students basic statistics, optimization, and Python

programming. We will assume that the students in this course already have a reasonable understanding of statistics and probability, calculus, and Python programming.

1.2.3 Course Curriculum Design

Using the relationships illustrated in Figure 6, we can now design our curriculum. Table 2 illustrates the lecture schedule, Table 3 illustrates the tutorial schedule, and Table 4 defines the grading scheme. These Tables are included in Appendix 1.1. The course begins with a crash course in database management, statistics, and data visualization so that students can get a good grasp of the tools they need to succeed. Then, prior to diving into linear and logistic regression, loss functions, performance metrics, and modelling ethics are explored. After exploring linear and logistic regression, students are introduced to decision trees, tree-based modelling techniques, and then to various ML architectures.

The students will be required to complete 4 assignments, each worth 12.5%. There will be a midterm covering lectures 1-5 with 25% and a final exam covering lectures 6-10 also worth 25%. This grading scheme was selected to maintain a balance between in-class learning and application. Typically, courses have greater weighting on assessments. It made sense, however, to put extra weighting on the assignments to provide students with projects that they can refer to during the application process. This curriculum will provide students with a thorough introduction to the field of data science. By modelling the concepts and curriculum on the most commonly possessed skills by data scientists and the most sought after skills by employers, this curriculum will prepare students for their careers in data science.

2. Data Science Program Curriculum Design

After redesigning MIE1624, our team used the previously gathered data to design a curriculum for a new “Master of Data Science and Artificial Intelligence” at the University of Toronto. Given that this master’s degree is meant to cover not only technical skills, but business skills as well, we decided to modify the data used in the first section such that soft and business skills were included. Skills like “time-management”, “communication”, and “excel” were considered in this section.

2.1 Course Clustering

The results displayed in Figures 2 and 4 define sections that do not include the soft skills that will enable students to succeed. To obtain new clusters we re-applied our clustering algorithm to Indeed’s job posting data. If two skills tend to be present together in a job-posting, they are considered as more similar and are more likely to be in the same cluster. After defining the clusters using a clustering algorithm, these clusters can further be broken down into courses. The results of the clustering algorithm applied to the Indeed technical and soft skills are displayed in Figure 7 in Appendix 2.1. As seen in the Figure, there are 8 groups obtained from 5 clusters produced by the clustering algorithm. Almost every topic requires strong mathematical and statistical fundamentals, and basic programming skills. Therefore, there should be a foundational course which enables students to have a reasonable understanding of statistics and probability, calculus, and programming.

2.2 Curriculum Design

Based on the data above, we designed a program that provides students with strong foundational skills, hands-on coursework, and work experience to ensure that participants obtain the skills necessary to succeed. Each student is required to complete 8 courses to graduate: 1 foundational course, 6 core courses, and 1 elective. Students are also required to complete at least one work-term in an approved field to obtain degree related work-experience.

2.2.1 Curriculum Design: Foundation Courses

Incoming students must complete 1 of the 2 foundational courses defined below:

1. Fundamentals of Computer Science for Data Science (designed for non-CS major background students)
2. Statistical Concepts for Data Science (designed for non-STAT major background students)

During course selection, students will choose one of the above courses depending on their academic background. These courses will have significant overlap to ensure that regardless of a student's selected foundational course, he or she will obtain the necessary prerequisite skills to succeed in his or her core courses.

2.2.2 Curriculum Design: Core Courses

Six core courses and their detailed curriculum were designed based on the clustering results detailed in Section 2.1. These core courses, their term schedule, and topics covered are illustrated in Table 5 in Appendix 2.1. Since some of the skills students learn in certain courses are foundational in others, the term-ordering of courses delineated by Table 5 is important. Further, we expect special care to be taken with respect to the ordering of the modules. We cannot require students to build advanced statistical models using R in "Statistical Learning" without first gaining a strong understanding of the fundamentals of R in "Data Mining and Exploratory Analysis". "Distributed Computing & Big Data" and "Data Mining and Exploratory Analysis" are explained in detail in the following section. The remainder of the courses are also explained in detail in Appendix 2.3.

2.2.3 Core Course Visualization: Distributed Computing & Big Data

Figure 12 in Appendix 2.2.5 delineates a timeline of the modules and topics covered. The content and the ordering of the content is shown in Figure 13, below:

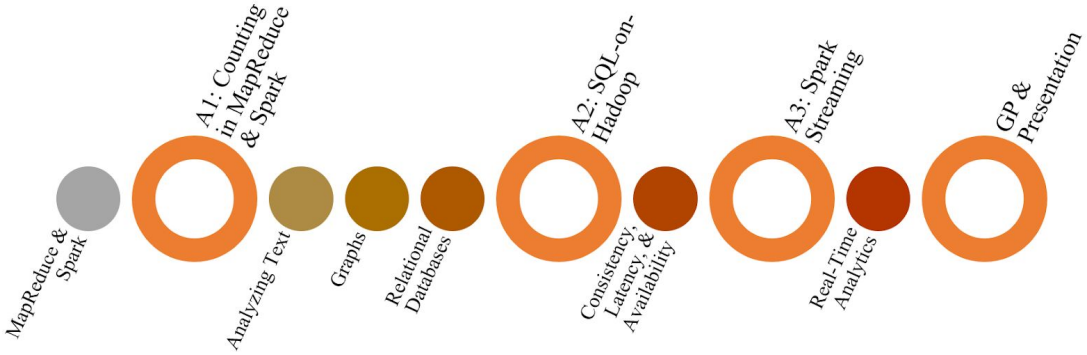

Figure 13: Visualization of the content order in distributed computing & big data

To mandate hands-on learning, this course is assignment based. For a course like this one, this format makes the most sense because the topics covered by the course are fundamentally practical.

2.2.4 Core Course Visualization: Statistical Learning

This course will be the majority of each student's workload. A wide variety of topics are covered. Some of these topics are optimzation-based learning (loss minimization and regularization), statistical learning (maximum likelihood and Bayesian learning), and modelling (classification, regression, and clustering). A breakdown of the modules covered as well as the lecture schedule, tentative assignment timelines, and course project schedule are displayed in Table 6 and Figure 9, respectively, in Appendix 2.2.2. A visualization of the ordering of the content is shown in Figure 14:

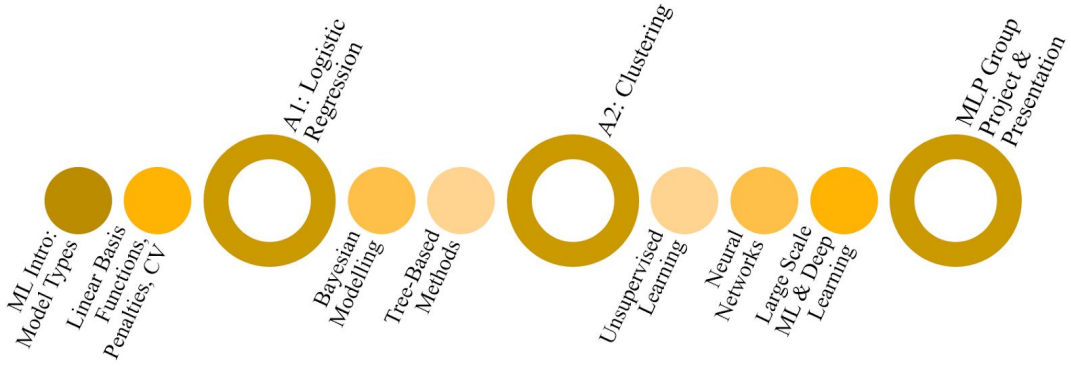

Figure 14: Visualization of the content order in distributed computing & big data

A course project was included to teach students how to communicate in a high-stakes environment. Students will present the results of their group project to the class. This will force students to not only understand various data analysis techniques, but to also use the results of their analysis to make data-driven decisions.

2.2.3 Electives and Internship

By allowing students to choose an elective, we create a more personalized learning experience. The elective courses are "Deep Learning", "Large-Scale Optimization for Data Science", "Advanced Topics in Artificial Intelligence", and "Database Systems Implementation". Students can choose which elective course to take based on their personal interests and career goals. These courses are explained in detail in Appendix 2.3.

Internship and extra-curricular activities must also be part of the program because soft skills are as important as technical skills as seen in Section 2.1. Although students are expected to develop through their coursework, professional experience is vital. As such, students are required to acquire at least 600 hours (1 term) of paid work-experience. In the following section we identify the most attractive full and part time data science opportunities in Canada.

3. Identifying Domestic Data Science Opportunities

We have identified internships as a requirement for our proposed "Master of Data Science and Artificial Intelligence" program at the University of Toronto. Viable internship opportunities can be hard to find. The purpose of this section is to take a data-driven approach to identifying internship opportunities in Canada. Given that company rating data reflects a company's environment and employee salary data reflects compensation, we will use the rating and salary data available on Glassdoor and Workopolis to determine which companies have positive work environments and competitive compensation.

3.1 Data Analysis

3.1.1 Rating Data

Glassdoor provides company ratings that depend on reviews by its employees. In addition to overall company ratings, Glassdoor provides ratings on career opportunities, work/life balance, culture & values, senior management, and benefits. These factors accurately reflect the experience of the employees at each company. Our analysis will focus on four of the ratings: career opportunities, work/life balance, culture & value ratings, and overall ratings. Figure 15 illustrates the companies with the highest Glassdoor ratings. To obtain these companies, companies with an overall rating less than 4 or a career & opportunities rating less than 3.5 were removed.

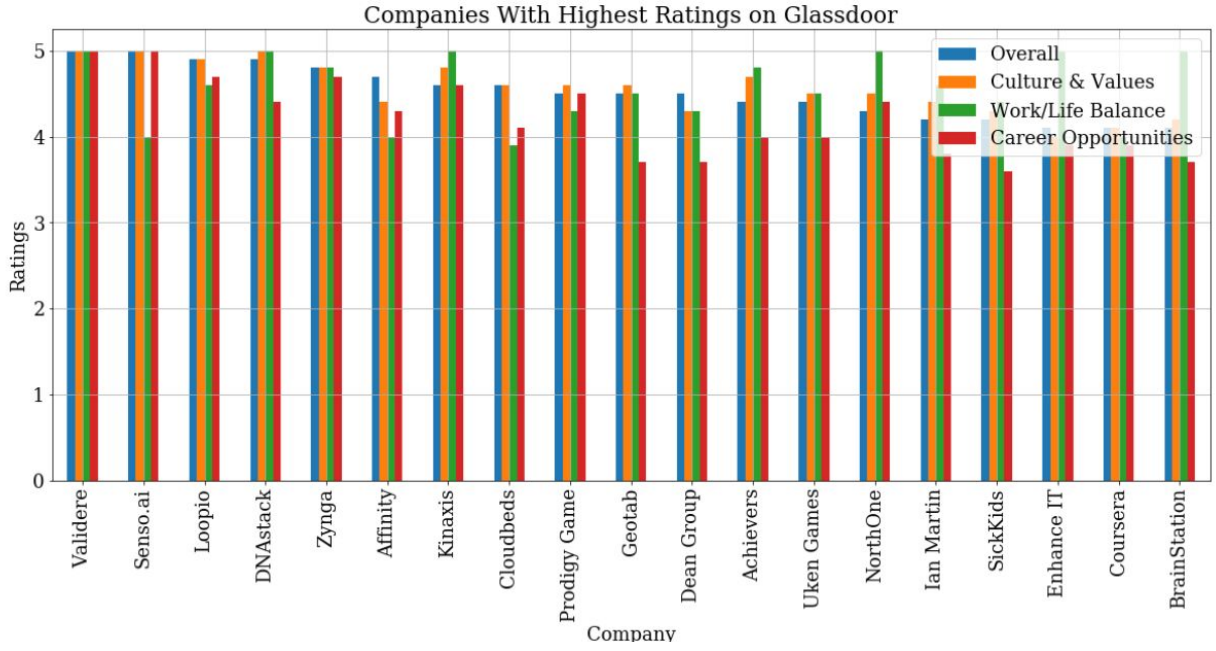

Figure 15: Glassdoor highest-rated companies

Many of these companies are not big-name companies. Many companies present in Figure 15 likely do not aggressively compensate. Given that this is a part of our criteria, we will use salary data from Workopolis to identify the companies that compensate well.

3.1.2 Salary Data

By scraping the salary data from Workopolis, we identified the top 30 companies with the highest salaries. Figure 16 in Appendix 3.1 illustrates these results. Using the insight gained from the rating and salary data, we can combine these metrics to make some recommendations. Note that there are many holes in our analysis that are definitely worth talking about. One thing to consider is that most people don't review the companies that they worked for. This will definitely affect our analysis as our data may not accurately capture the average employee at each of these companies. Furthermore, our analysis is quite shallow in the sense that the exact occupations of the people leaving reviews at these companies was not considered. What if the only employees leaving reviews at IMS Group were Junior Analysts while the only employees leaving salary reviews at BRP were Senior Analysts? This would result in a huge pay discrepancy that we did not account for. Nonetheless, our method is a reasonable way to determine which companies are in the top tier with respect to compensation assuming that most of the reviews are coming from employees of similar backgrounds.

3.2 Recommendations

Figure 17 in Appendix 3.1 illustrates our domestic data science opportunity recommendations. To reiterate, these recommendations are based on the review data from Glassdoor and the salary data from Workopolis. These recommendations were obtained by taking the highest-rated companies displayed in Figure 15 and thresholding them based on our salary metrics. The resulting companies have excellent work environments and compensate their employees well. A few examples of these companies are illustrated in Figure 18:

Figure 18: Examples of high-paying companies with positive work environments (Achievers, Dean Group, Coursera, Kinaxis)

These companies all have excellent review data and high-paying salaries. Using these recommendations, future students enrolled in our proposed "Master of Data Science and Artificial Intelligence" will thrive during their internship semester and in their professional careers.

Citations:

CognitiveClass. *Courses offered by CognitiveClass*, https://cognitiveclass.ai/courses.

Glassdoor. “Data Science Jobs.” *Glassdoor*, 2020, https://www.glassdoor.ca/Job/data-science-jobs-SRCH_KO0,12.htm. Accessed 28 11 2020.

Kaggle. “2019 Kaggle ML & DS Survey.” *Kaggle*, 2019, https://www.kaggle.com/c/kaggle-survey-2019. Accessed 23 11 20.

Kaggle, et al. “2018 Kaggle ML & DS Survey.” *Kaggle*, 03 11 2018, https://www.kaggle.com/kaggle/kaggle-survey-2018. Accessed 23 11 2020.

Princeton University. “Large-scale Optimization for Data Science.” http://www.princeton.edu/~yc5/ele522_optimization/index.html.

Rotman. *Master of Management Analytics*, https://www.rotman.utoronto.ca/Degrees/MastersPrograms/MMA.

University of Waterloo. *Master of Data Science and Artificial Intelligence (MDSAI)*, https://uwaterloo.ca/graduate-studies-academic-calendar/mathematics/data-science-and-artificial-intelligence/master-data-science-and-artificial-intelligence-mdsai.

Workopolis. “Top 20 Data Science Jobs.” *Workopolis*, 2020, https://www.workopolis.com/jobsearch/data-science-jobs. Accessed 28 11 2020.

Appendices:

Appendix 1: Course Curriculum Design

Appendix 1.1: Tables and Figures

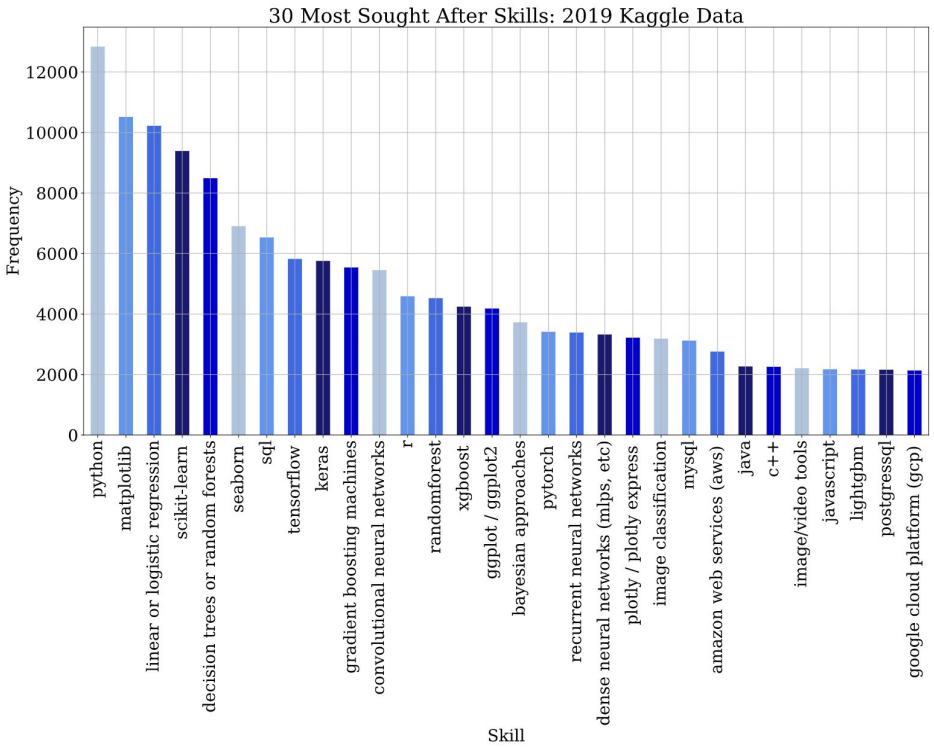

Figure 3: 30 most sought after skills according to the 2019 Kaggle dataset

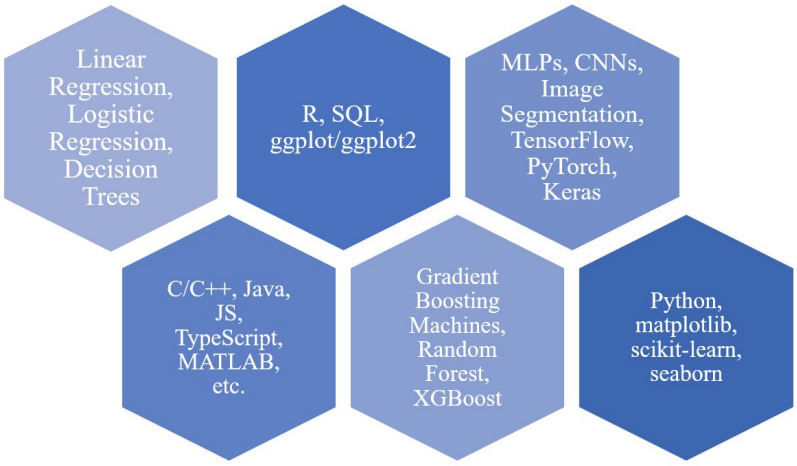

Figure 4: Results of agglomerative clustering algorithm on 2019 Kaggle dataset

Table 1: Clusters from Kaggle 2018/2019 data segregated by concepts and tools

	Clusters:	Concepts:	Tools:
Cluster 1:	Linear Regression, Logistic Regression, Decision Trees	Fundamental modelling architectures and ensembles	Python, scikit-learn, PyTorch, other Python modelling libraries
Cluster 2:	R, SQL, ggplot/ggplot2	Database management, statistical analysis	R, SQL, ggplot/ggplot2
Cluster 3:	MLPs, CNNs, Image Segmentation, TensorFlow, PyTorch, Keras	Machine learning architectures	TensorFlow, PyTorch, Keras
Cluster 4:	C/C++, Java, JS, TypeScript, MATLAB, etc.	Web-dev, low-level programming, mathematical modelling	C/C++, Java, JS, TS, MATLAB
Cluster 5:	Gradient Boosting Machines, Random Forest, XGBoost	Advanced tree-based algorithms and learned architectures	Python, scikit-learn, PyTorch, other Python modelling libraries
Cluster 6:	Python, matplotlib, scikit-learn, seaborn	Python programming and data visualization	Python, matplotlib, scikit-learn, seaborn

Table 2: Re-designed MIE1624 lecture schedule

	Lecture Focus:	Tools Used:
L1	Database management skills, data cleaning	SQL
L2	Statistics, common distributions, sampling techniques, visualizing data using R	R, ggplot/ggplot2
L3	Data visualization and exploratory data analysis using Python	Python, matplotlib, scikit-learn, seaborn
L4	Loss functions, performance metrics, ethical modelling	-
L5	Fundamental modelling architectures (linear regression, logistic regression)	Python, Pandas, scikit-learn
L6	Decision trees and ensembles	Python, scikit-learn
L7	Advanced tree-based algorithms, introduction to machine learning	Python, PyTorch, scikit-learn
L8	Dense neural network architectures	Python, PyTorch, Keras

L9	Image classification using CNNs	Python, PyTorch, Keras
L10	Text analysis: BoW, TF-IDF, introduction to sentiment analysis using learned architectures	Python, scikit-learn, PyTorch

Table 3: Re-designed MIE1624 tutorial schedule

	Tutorial Assignment/Goals:		Tutorial Assignment/Goals:
T1	SQL Tutorial: merging datasets, designing simple datasets, manipulating data	T7	Assignment 3: Comparing Models - Using the same dataset from assignment 2, implement a handful of new models - Compare and visualize the results
T2	Assignment 1: Data Manipulation Using SQL & R - Import a set of datasets - Join the datasets using SQL - Analyze the distributions present in the data - Compute and display various quantitative statistics using R	T8	PyTorch Tutorial: - Preparing data for training - Training/testing MLPs - Training/testing CNNs
T3	Python library tutorial: scikit-learn, matplotlib, seaborn	T9	Assignment 4: Image Classification Using PyTorch - Using a provided dataset (MNIST or other), design and implement a CNN for the purpose of image classification
T4	Assignment 2: Logistic Regression - Using sk-learn and Pandas, perform logistic regression on a supplied dataset - Visualize the results using matplotlib & seaborn	T10	BoW/TF-IDF text representation tutorial: performing sentiment analysis using various unlearned models and BoW/TF-IDF features
T5	Midterm Help Session	T11	Sentiment analysis using learned models and GloVe/Word2Vec features
T6	Complex modelling tutorial (decision trees, tree-based algorithms)	T12	Final Exam Help Session

Table 4: Re-designed MIE1624 grading scheme

	Evaluation Breakdown:	Percentage of Final Grade:
Assignments	4 5-hour assignments each worth 12.5%	50%

Midterm	Covers Lectures 1-5	25%
Final Exam	Covers Lectures 6-10 (not cumulative)	25%

Appendix 2: Program Curriculum Design

Appendix 2.1: Tables and Figures

Table 5: Master's in DS/AI core courses

	Term:	Topics Covered:
Data Mining and Exploratory Analysis	1	Data collection, data preparation, data cleaning, programming using R, data visualization
Statistical Learning	1	Optimization-based learning: loss minimization, regularization; statistical learning: maximum likelihood, Bayesian learning; foundational and advanced modelling skills: classification, regression, and clustering
Data Visualization	1	Tableau, Power BI, seaborn, other visualization tools
Introduction to Data Management	2	Data models: ER, relational, others; query languages: relational algebra, SQL; database management systems: index structures, concurrency control, recovery, and query processing
Distributed Computing and Big Data	2	Data mining; machine learning techniques applied to text, graphs, and relational data; MapReduce; Spark
Introduction to Artificial Intelligence	2	Representation and reasoning, learning, and NLP; AI-based applications

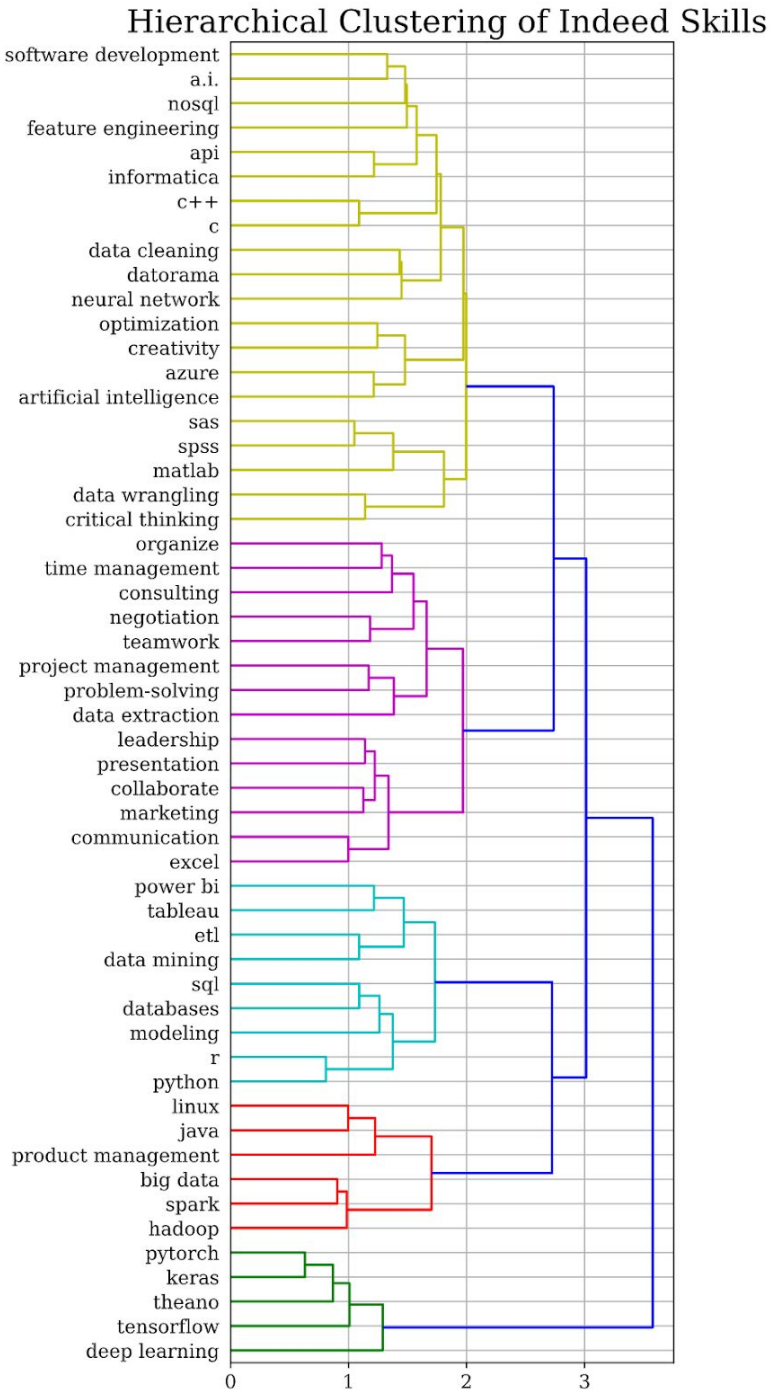

Figure 7: Hierarchical clustering of technical and soft skills according to Indeed data

Appendix 2.2: Core Course Breakdown

Appendix 2.2.1: Data Mining and Exploratory Analysis Curriculum Design

In this course, we will analyze all stages of data collection, preparation and cleaning. Further, we will teach students how to import data into a statistical programming environment, manipulate data, as well as analyze and visualize data. Material will be presented via case studies which will involve hands-on programming and analysis. These cases are designed to provide students with a breadth of understanding of the applications of the visualization techniques they are taught.

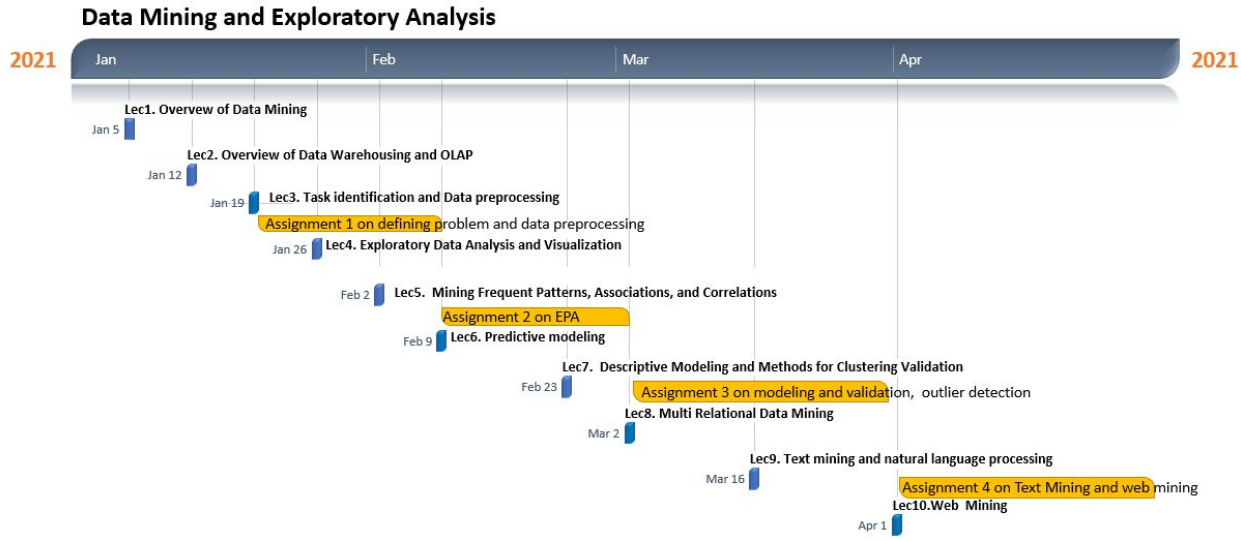

Figure 8: Data Mining and EPA curriculum design

Appendix 2.2.2: Statistical Learning

Table 6: Statistical learning module break-down

Lecture:	Summary:
L1	Supervised/Unsupervised Learning Overview
L2	Linear Basis Functions, Penalties, CV
L3	Bayesian Linear Basis Function Models
L4	Tree-Based Methods
L5	Unsup. Learning: Clustering & Dim. Reduc.
L6	Neural Networks
L7	Large Scale ML: SGD, Deep Learning

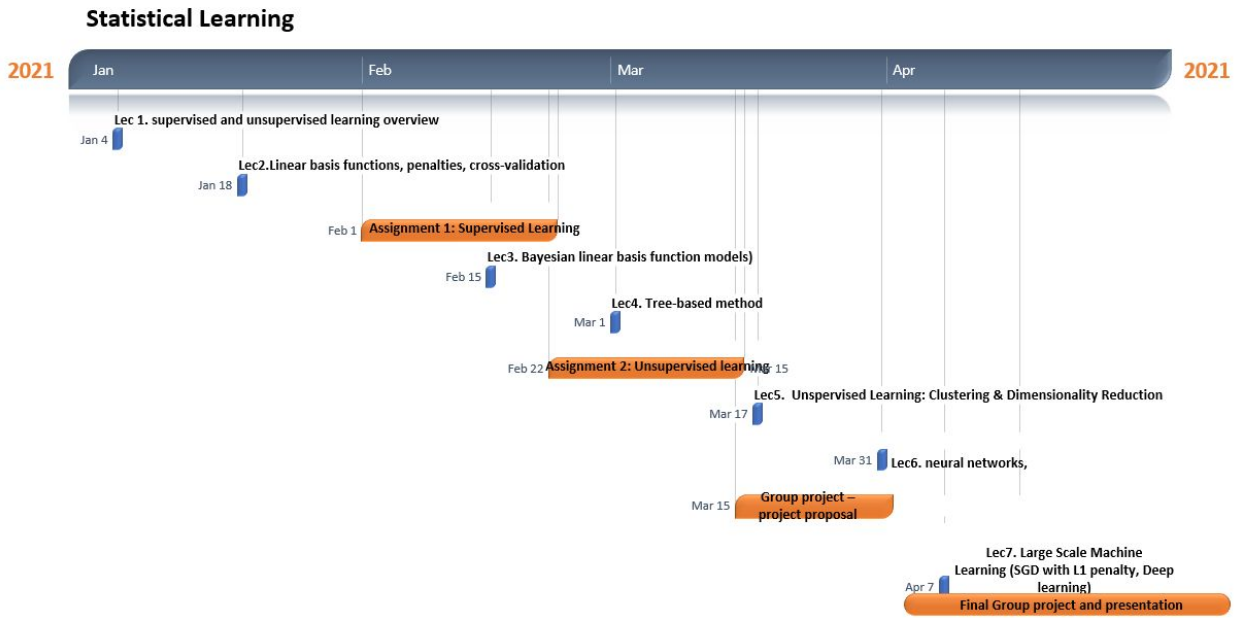

Figure 9: Statistical Learning curriculum design

Appendix 2.2.3: Data Visualization Curriculum Design

In this course, we will explore how to design and create data visualizations based on data available and tasks to be achieved, including visualization of high dimensional data and interactive methods directed at exploration and assessment of structure and dependencies in data.
This course shows students how to better understand data, present clear evidence of findings, and tell engaging data stories that clearly depict the main points all through data graphics. Tableau, PowerBI, and various data visualization tools will be used in this course.

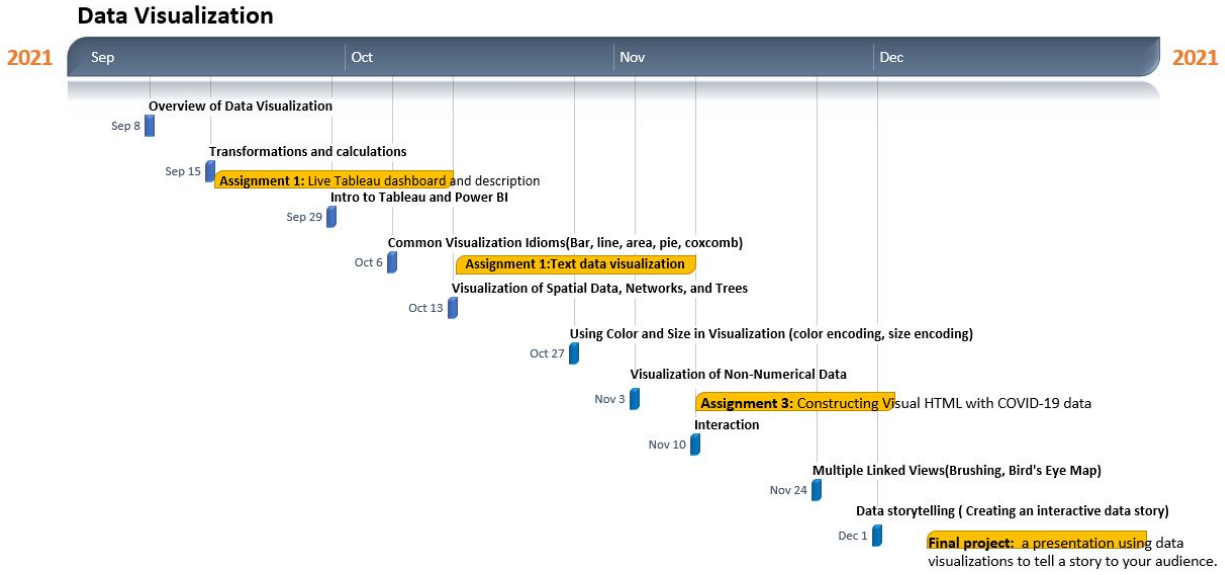

Figure 10: Data Visualization curriculum design

Appendix 2.2.4: Introduction to Data Management Curriculum Design

This course is about the management of large collections of data. It covers the fundamental concepts of database systems. Topics include data models (ER, relational, and others); query languages (relational algebra, SQL, and others); implementation techniques of database management systems (index structures, concurrency control, recovery, and query processing); distributed and noSQL databases. It also introduce students to topics in database research, such as warehousing, data mining, managing data streams, data cleaning, data integration, and distributed/parallel databases.

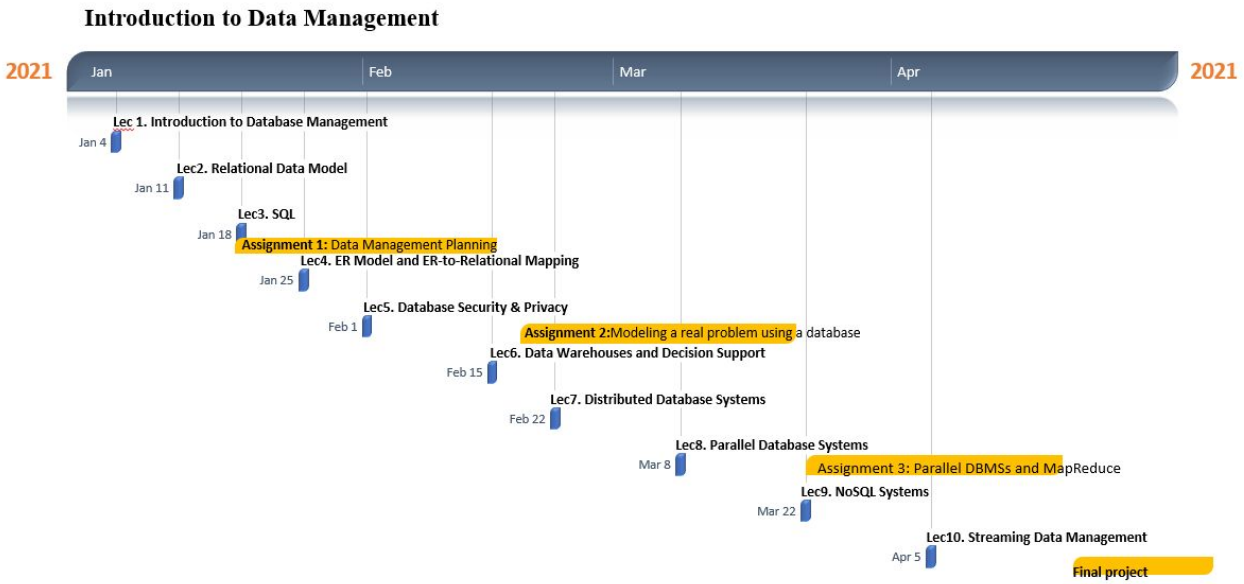

Figure 11: Introduction Data Management curriculum design

Appendix 2.2.5: Distributed Computing and Big Data

This course provides an introduction to data-intensive distributed computing. Our focus is algorithm design and thinking at scale: we will cover data mining and machine learning techniques as applied to text, graphs, and relational data. Most of the course will be taught in a combination of MapReduce and Spark, two representative dataflow abstractions for large-scale data analysis.

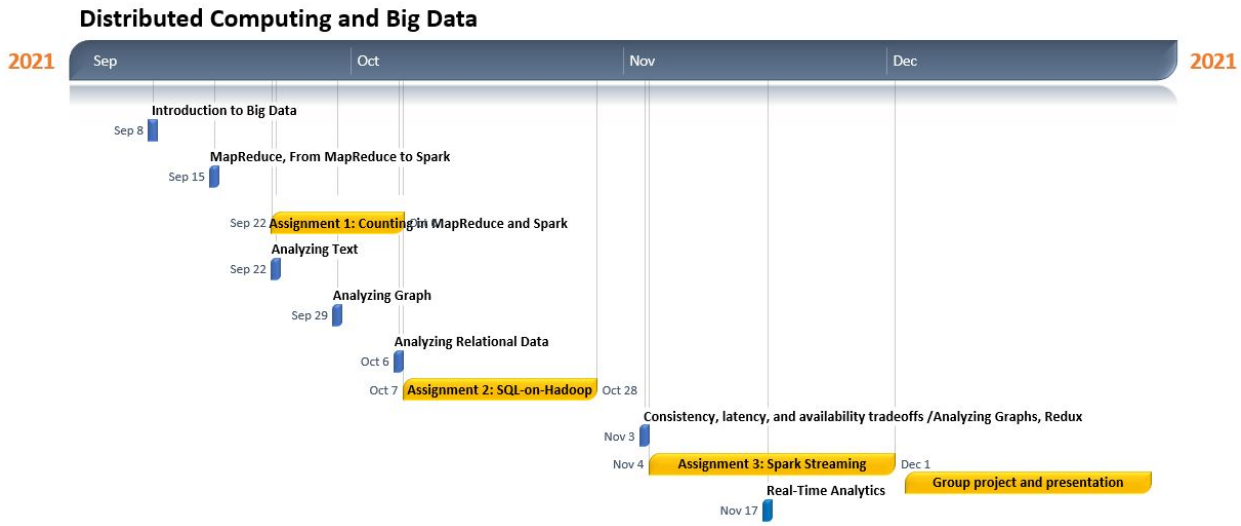

Figure 12: Distributed Computing and Big Data curriculum design

Appendix 2.2.6: Introduction to Artificial Intelligence

In this course, we start by describing what the latest generation of artificial intelligence techniques can actually do. After an introduction of some basic concepts, the course illustrates both the potential and current limitations of these techniques with examples from a variety of applications. We focus on three central areas in AI: representation and reasoning, learning, and natural language processing. Exercises will include hands-on application of basic AI techniques as well as selection of appropriate technologies for a given problem and anticipation of design implications. In a final project, groups of students will participate in the creation of an AI-based application.

Appendix 2.3: Elective Course Breakdown

Appendix 2.3.1: Deep Learning

This course is an introduction to deep learning, a branch of machine learning concerned with the development and application of neural networks. Deep learning algorithms extract layered high-level representations of data in a way that maximizes performance on a given task. We will cover a range of topics from basic neural networks, convolutional and recurrent network structures, deep unsupervised and reinforcement learning, and applications to problem domains like speech recognition and computer vision. Major deep learning frameworks including TensorFlow, Keras, and Theano will be used.

Appendix 2.3.2: Large-Scale Optimization for Data Science

This course further explores optimization methods that are suitable for large-scale problems arising in data science and machine learning applications. Firstly, algorithms such as gradient methods, proximal methods, mirror descent, ADMM, quasi-Newton methods, stochastic optimization, and distributed optimization are revived. Then, we will discuss the efficacy of these methods in data science problems under appropriate statistical models. Finally, we will introduce a global geometric analysis to

characterize the nonconvex landscape of the empirical risks in several high-dimensional estimation and learning problems.

Appendix 2.3.3: Advanced Topics in Artificial Intelligence

This course goes in depth on selected topics and methods within artificial intelligence (AI), machine learning (ML) and their applications. Examples include computational intelligence algorithms in search, optimization and classification, which to a large extent consist of bio-inspired mechanisms. Examples of relevant applications include robotics, music, health and medicine.

Appendix 2.3.4: Database Systems Implementation

This course is a hands-on systems-focused course on the implementation of a database management system (DBMS), especially, a relational DBMS (RDBMS). This course will cover key systems topics in implementing an RDBMS: data storage, buffer management, indexing, sorting, relational operator implementations, query optimization, and transaction management and concurrency control. The implementation of newer Big Data systems such as Spark and MapReduce/Hadoop, as well as distributed NoSQL/key-value stores, in-memory RDBMSs, and streaming DBMSs will also be covered. The major component of this course is hands-on C++ programming to implement two key components of an RDBMS, a buffer manager and a B+ Tree index, on top of a basic RDBMS skeleton.

Appendix 3: Identifying Domestic Data Science Opportunities

Appendix 3.1: Tables and Figures

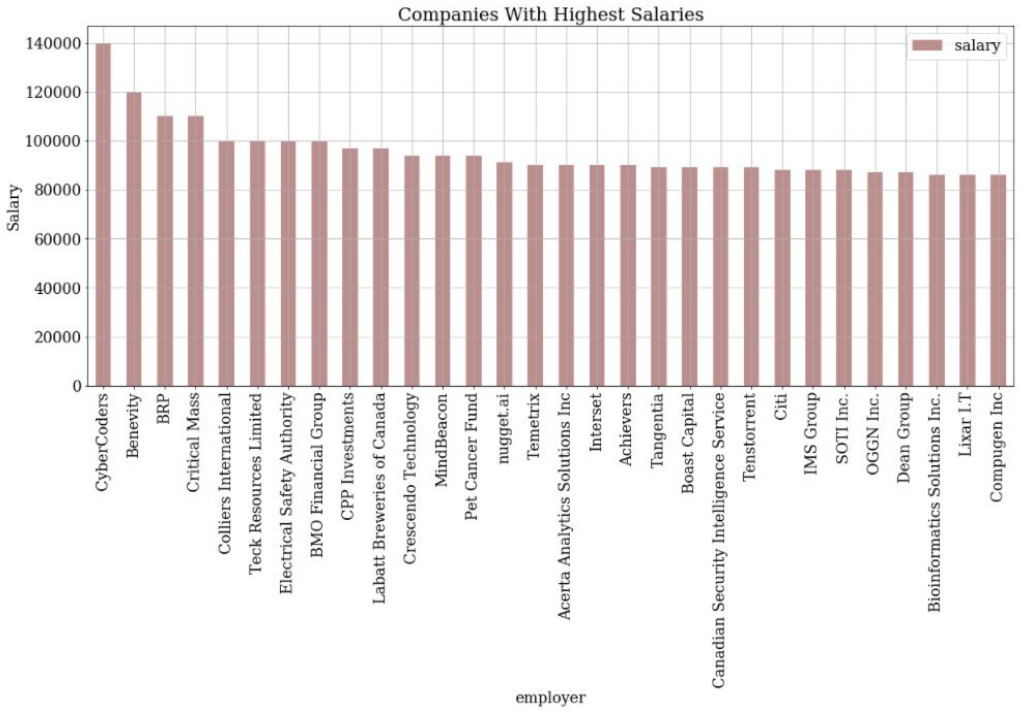

Figure 16: 30 Companies with the highest salaries

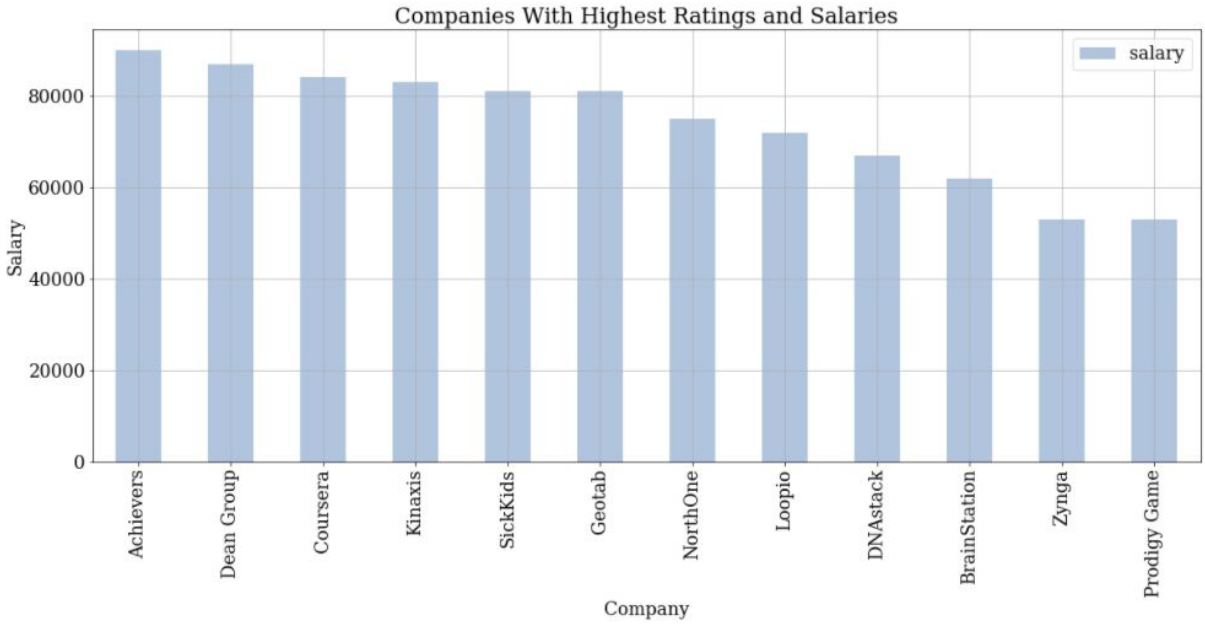

Figure 17: Companies with the highest salaries and ratings