“Learn to Code”[1][2]: Data Science Education Optimization Report

MIE1624 – Introduction to Data Science and Analytics

Adam Parker, Bassel Tarabay, Jad Zalzal, Lydia Jeong, Nadine Alzaghrini, Yifei Ai

November 30, 2020

UNIVERSITY OF
TORONTO

TABLE OF CONTENTS

Executive Summary

The report begins with an introduction laying out the various goals and education initiatives proposed by the group. An explanation is then provided outlining the development of a new course curriculum for MIE1624, followed by a breakdown of the new content provided in the course. Next, the report explores the data collection and analysis methods used to create a new Master's program in data science. The Master's program is explained and the curricula for various courses included in the program are provided. Finally, a new startup is proposed seeking to address the issue of unemployment in North America. The startup queries an individual's skills and connects them with the courses necessary to find gainful employment in the field of data science.

Introduction

Over centuries, technological transformations have shaped economies and the capacity of individuals to make a living. Most recently, technological developments in artificial intelligence, operating under the wing of globalization, have transformed the landscape of jobs as we know it. In particular, machine learning and data analytics have had a considerable impact [3]. Machine learning is a field of artificial intelligence focused on building applications that learn from data and improve their accuracy over time without being programmed to do so.

The purpose of this report is to promote and enhance education in the fields of data analytics and machine learning with the aim of increasing individuals' employability and chances of success in the workplace. The report advances evidence-based data analysis to determine what the most relevant skills are for individuals presently in data science occupations, and uses this analysis to build curriculums that will be most relevant to the current data science industry. The guiding philosophy of this approach sees education as vocational training, placing the responsibility on course and program designers to be attuned to the realities of the working world.

Along the same lines of promoting employability in the fields of Artificial intelligence, this report introduces "Learn to Code" [1][2], an EdTech startup aimed at making transitions into data science as painless and as cost effective as possible, namely for unemployed individuals equipped with skills that are no longer relevant to the job market.

MIE1624 Course Curriculum

Analysis of Relevant Skills

The group first desired to optimize the curriculum of the MIE1624 course in order to enhance employment outcomes for students. Given that the course is at the graduate level, it was desired to see which skills are most commonly employed by engineering graduates working in industry. Specifically, those with a Master's degree who work in lucrative data science and analytics positions. To do this, data from the 2018 Kaggle Survey [4] was consulted. The data was filtered to reflect only individuals holding a Master's degree, who have less than 5 years experience and who are highly paid (earning more than $80,000 per year) [4]. Experience was set to less than 5 years to see what skills are likely to be important immediately upon starting one's career. At certain other levels of experience, skills may be assumed to have been acquired while working. By contrast, for those with less than 5 years experience, more skills will have to have been earned directly from one's education. Thus, it is imperative that the skills taught in MIE1624 reflect what graduates will need early in their careers. The 20 most common skills from the Kaggle data according to the criteria listed above are visualized in Appendix A.

Based on an analysis of these skills, it is clear that relatively basic programming skills such as Python, SQL and Linear/Logistic Regression feature most prominently in the data [4]. Alternative languages to Python such as R are less common, although they still appear fairly frequently in the data [4].

Next, the job website Indeed was scraped such that job postings at the entry, mid and senior levels for Data Scientist, Data Analyst and Data Manager positions could be studied [5]. Skills were subsequently collected from the educational website Coursera [6] and combined with skills from Kaggle [4] and those that were manually inserted by the group. The postings were then analyzed such that skills could be differentiated for each of the positions. The graphs for the 3 careers are available in Appendix A.

Recommendations for Final Design of the Course

The final design of the MIE1624 course was developed by finding the most common skills from the Coursera website [6] and comparing them with the original course syllabus [7]. Web scraping was performed to extract all the skills from data science courses. Relevant skills such as data science in business, Azure ML Studio, Github, and R programming were not present in the original syllabus [7], therefore they were included in the redesigned curriculum. Figure 3 represents the designed course curriculum, indicating the list of lectures, tutorials, and assignments covered for each week. Bar graph and word cloud visualization of the most common skills are displayed in Appendix B (figure 14-15).

Master's Program Creation

Important Skills

A new curriculum will be proposed that seeks to teach the most relevant skills from industry and that are currently available online. In order to ascertain what skills are most common in industry, data from the 2018 Kaggle survey was consulted. This survey is useful as it considers a wide range of diverse data points, covering over 12,000 respondents in over 50 countries [4]. This international data will be especially useful for the design of the Master's program. For an institution with students from varying backgrounds and of multiple nationalities, the ability to acquire skills that are useful worldwide, rather than just in Canada, is highly advantageous. Coursera was also investigated in order to provide a representation of a typical curriculum in data science across varying institutions [6], which ensures that the new Master's program is competitive and teaches many of the skills that are offered elsewhere.

A web-scraping program was used in order to determine the most common skills from Coursera [6], across courses requiring varying degrees of proficiency, and also from Kaggle [4]. Skills were selected from courses pertaining to data science, given that the Master's program would focus on the same areas. The rationale for this approach is that if a skill appears in a large number of courses, it must be relevant pedagogically or otherwise be important to the development of a student as a data scientist [6], and that if it appears often in Kaggle it would be a skill that is often used in the workplace [4]. Given that both employability and education are a major focus of the program, these two sources were considered to provide the most useful representation of common skills. The results of the web-scraping, a list of the most commonly used skills, would form the basis of the program's curriculum, and would represent a starting point for roughly what skills the program would seek to impart onto its students.

Recognizing Difficulty Levels

The Master's program seeks to accept students who lack proficiency in the field of data science or data analysis. As such, it is important that courses are appropriately ranked according to their difficulty and required level of experience such that students can acquire new skills at a reasonable pace. The program would be structured into semesters with courses divided up accordingly. During the first semester, skills considered beginner skills would be taught, while after that the focus would shift to skills that are intermediate or advanced level. These labels correspond to those used on Coursera [6], and provide a convenient way of classifying difficulty.

Various classification methods were used to determine what skills corresponded with what levels of difficulty. Models were configured such that the dependent variable was the Coursera difficulty level [6] and the independent variables were the various skills. The 100 most common skills were tested according to the models used in this analysis. The testing included 7 different models: a Gaussian Naive Bayes, a logistic regression model, a random forest classifier, a gradient boosting classifier, a support vector classifier, a decision tree and a linear support vector classifier, using functions from sklearn [8]. Each model was then evaluated for its accuracy to narrow down the models to the 4 most accurate. At this point, hyperparameter tuning was further utilized on the remaining models in order to increase performance. Accuracy was evaluated again, and a Gaussian Naive Bayes model [8] with a training accuracy of 78% and a testing accuracy of 72% was selected.

Skill Clustering

With the skills from the previous analysis of Coursera [6] and Kaggle [4], the program was further developed to be more in tune with the nature of work as a data scientist. Rather than linking skills through which courses they are most likely to occur in, it was opted to take a more work-based approach and group based on the number of times they occur in a job. To this end, the 250 most common skills in the aforementioned sources were used to scrape the job posting site Indeed [5]. This returned data allowing for the number of skill co-occurrences in job postings to be recorded.

From here, various algorithms were tested in order to find the most effective for clustering the jobs together: DBSCAN, K-Means and Hierarchical Clustering. The results of the DBSCAN algorithm were found to be highly sensitive to changes in the epsilon value - a mere 0.1 difference would significantly alter the number of clusters - and it was rejected outright. The remaining two clustering algorithms were then tested, and the results of Hierarchical Clustering are shown below (Figure 1).

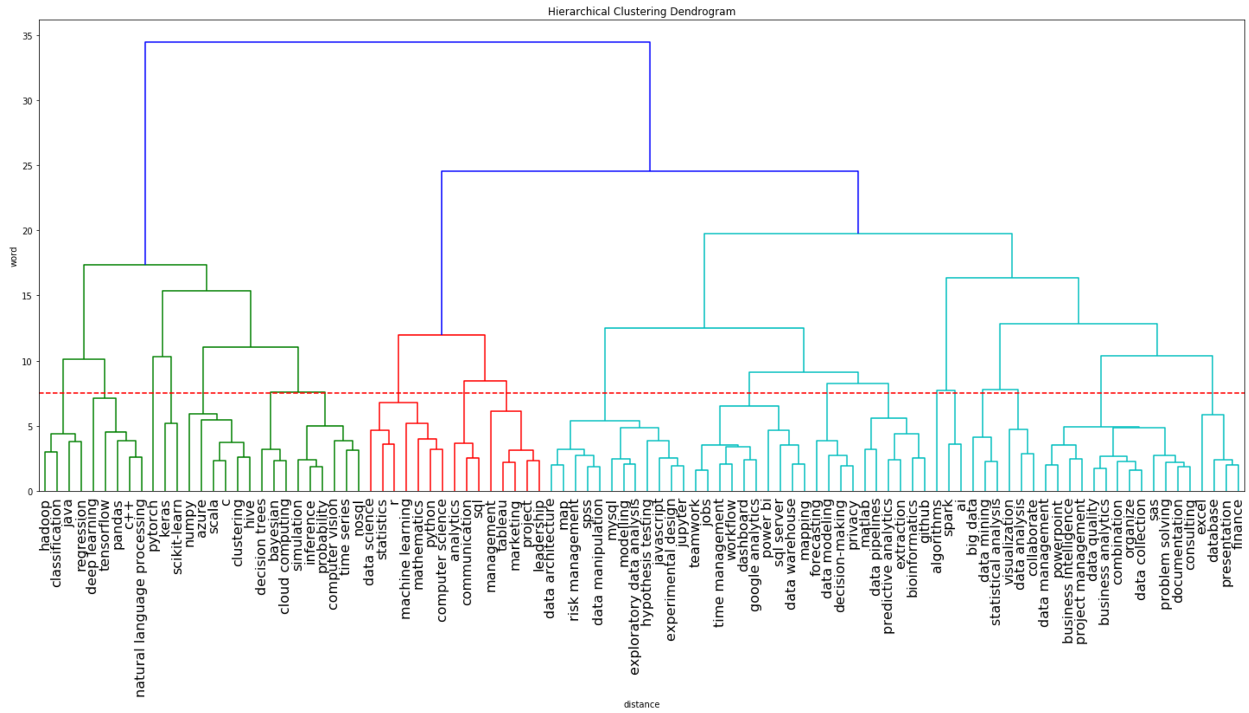


Figure 1: Hierarchical Clustering of Relevant Skills based on Job Co-Occurrences

Ultimately, hierarchical clustering was preferred to KMeans, as the results of Hierarchical Clustering were deemed to be more reasonable. The results of Hierarchical Clustering yielded a more even distribution of skills per course, indicating that the resultant courses would be reasonably balanced in terms of what was being taught in each one. By contrast, the results for KMeans grouped certain individual skills, such as data quality, on their own, while grouping multiple computational skills such as C, C++, Numpy and Pandas, among 6 others, into their own course. This is despite the fact that both clustering methods yielded approximately the same number of courses (17 v. 19). As a result of this visual inspection of the results, the clusterings from Hierarchical Clustering were chosen to guide course creation for the curriculum. Each course would thus cover topics that are likely to be used together in the workplace, allowing for an education program that closely mimics work in the industry.

Course Finalizing

As explained earlier, results of the Hierarchical Clusterings were used as a basis for the creation/selection of courses. In order to obtain relevant and practical courses, these clusters were read in parallel to the courses commonly taught in the Top 10 Masters programs in Data Science and Artificial Intelligence, as ranked by Forbes [9]. The ten programs along with the courses taught in each, the program duration, mission statement and link were listed in an excel sheet. The following elaborates on the method of extraction of the commonly taught programs across all programs: First, the data on the list of courses was cleaned. Stopwords were then removed, and the strings containing the cleaned list of courses were transformed using lemmatization. The top 40 features belonging to an n-gram (2,3) were then visually inspected, and the most significant terms obtained were saved. The universities teaching courses with the same terms were then extracted for later use in the course curriculums.

In accordance with the previous, clusters were divided into core and elective courses which were then ranked in order to ascertain their difficulty. The same model used to assign a difficulty rating from earlier was called on the clusters. Five clusters/courses were qualified as beginner courses (see Appendix B, Figure 13) whereas the remaining thirteen were classified as intermediate/advanced courses. Figure 17 in Appendix B includes a summary of the results obtained. To note, the ranking classification as obtained by the model was not adopted, as all courses deemed beginner courses were classified as elective ones.

Students should apply for an internship during their first semester and make their final decision on where they will do their internship before the start of the second semester. They will have a list of potential companies that the university already has a partnership with (to guarantee an internship for all the students at some of the top companies in the field). During the second semester, the students would formulate their project with the company that they will be interning with and under supervision of a professor. Then, the students would implement their project with the company during their 4 month internship (May - August).

The example syllabus including detailed learning outcomes were prepared for the following core courses and are available in Appendix E:

- Introduction to Machine Learning and Data Analysis;
- Advanced Machine Learning;
- Advanced Statistics;
- Deep Learning and Neural Networks; and
- Big Data

For the remainder of the courses, a course description was added.
Each five course curriculum was visualized as a timeline. Figure 2 illustrates the curriculum for the Advanced Machine Learning course, specifying the list of topics covered in a sequence. The remaining four courses are attached in the Appendix C (Figure 22-24).

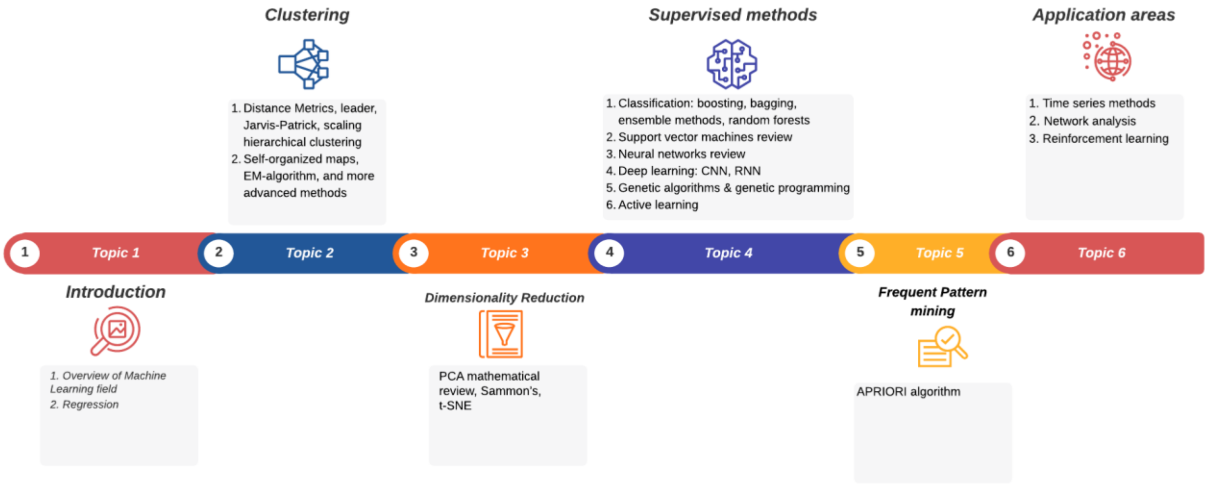

Clustering
1. Distance Metrics, leader, Jarvis-Patrick, scaling hierarchical clustering
2. Self-organized maps, EM-algorithm, and more advanced methods
Supervised methods
1. Classification: boosting, bagging, ensemble methods, random forests
2. Support vector machines review
3. Neural networks review
4. Deep learning: CNN, RNN
5. Genetic algorithms & genetic programming
6. Active learning
Application areas
1. Time series methods
2. Network analysis
3. Reinforcement learning
1 Topic 1
2 Topic 2
3 Topic 3
4 Topic 4
5 Topic 5
6 Topic 6
Introduction
1. Overview of Machine Learning field
2. Regression
Dimensionality Reduction
PCA mathematical review, Sammon's, t-SNE
Frequent Pattern mining
APRIORI algorithm

Figure 2: Course curriculum for Advanced Machine Learning [10]

Internships

With maximizing employability and job related skills in mind, it was considered whether or not the Master's program should feature an internship. Theoretically, an internship could provide job experience which could put graduates in a better position to find gainful employment than were they to focus only on courses. Here, the Kaggle data [4] was once again consulted. The purpose of this was to see how an additional 3-6 months of experience early in one's career can affect salaries. The investigation would look at both the United States and at India, for both a Bachelor's Degree and a Master's Degree level. The USA and India were both chosen as reasonably useful examples as they possessed a large number of respondents each. The comparison then focused on the salary differences between individuals possessing <1 year of experience, individuals possessing 1-2 years, and individuals possessing 3-5 years, to gauge what a modest improvement in experience would mean for salaries.

The samples were bootstrapped to produce normally distributed data, and then an ANOVA test was used to determine whether the differences were statistically significant. The histograms of the results are shown in Appendix B (Figure 18-21).

Based on a visual inspection alone, it is clear that there are large differences in the ranges of the distributions of the data. An initial impression suggests that adding modest increases in experience can have a large impact on salary, and the ANOVA test that was performed confirmed that these differences were statistically significant. Therefore, a 4-month internship was included in the Master's program.

Consequently, the design of the 12-month Master's program curriculum incorporated the following:

1. Optional **boot camp** taking place before the start of the program. This is a one-month intensive training that equips students with essential knowledge necessary to prepare for the master's program. Bootcamp includes introduction to the master's program and comprehensive training for basic statistics, linear algebra, Python, and R programming.

2. **Term 1**: students take five core courses in this 4-month term (Introduction to Machine Learning and Data Analysis; Advanced Machine Learning; Advanced Statistics; Deep Learning and Neural Networks; and Big Data)
3. **Term 2**: students take two core courses in this 4-month term (Advanced Machine Learning, and Methods for Decision Making) and select three from the six optional courses (Cloud computing, Time Series, Database systems, Marketing Analytics, Visualization, and Data Analytics for Business Strategy).
4. **Internship**: Students apply for an internship to get interviewed by potential employers. Interview begins during the start of Term 2 and the company is secured in March. Internship begins starting in May and ends in August.

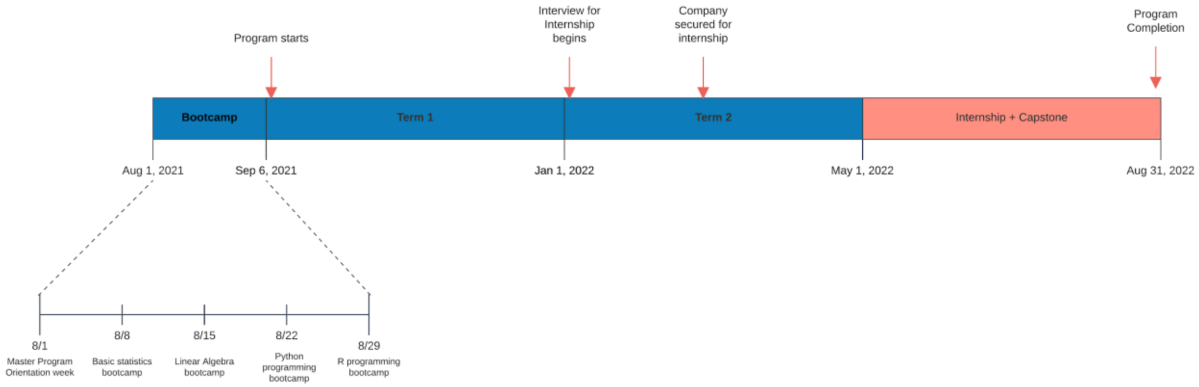


Figure 3. Visualization of Master’s program development [10]

“Learn to Code” [1][2] Initiative

In December of 2019, then former Vice President Joe Biden delivered a controversial remark to a group of New Hampshirites. Biden was attempting to address the financial woes affecting a community of miners when he declared that “Anybody who can go down 3,000 feet in a mine can sure as hell learn to program as well” [1]. The comments were not particularly well received, and joined in a broader trend of programming being offered up as a solution to unemployment [1][2]. Earlier that year, Twitter was accused of proscribing the use of the phrase “learn to code” on its platform, under the justification that they were preventing harassment [2].

These incidents point to perceptions and sensitivities existing within society. Namely, that certain industries are dying, individuals in those industries no longer have useful skills, and the best course of action for members of those industries is to embrace the novel. Given this treatment of coding as a kind of panacea for economic distress, it's worth investigating what the rationale is behind such a sentiment and whether it has any merit. The title of this initiative, “Learn to Code” [1][2], reflects a phrase common to both controversies found in the news.

Unemployment in North America

The recent COVID-19 pandemic has exposed a lot of volatility in employment among the North American workforce. As Figure 4 shows, unemployment skyrocketed in the first months of 2020, more than doubling, and Figure 5 further indicates that industries such as hospitality and retail were hit particularly hard, with many employers making drastic layoffs.

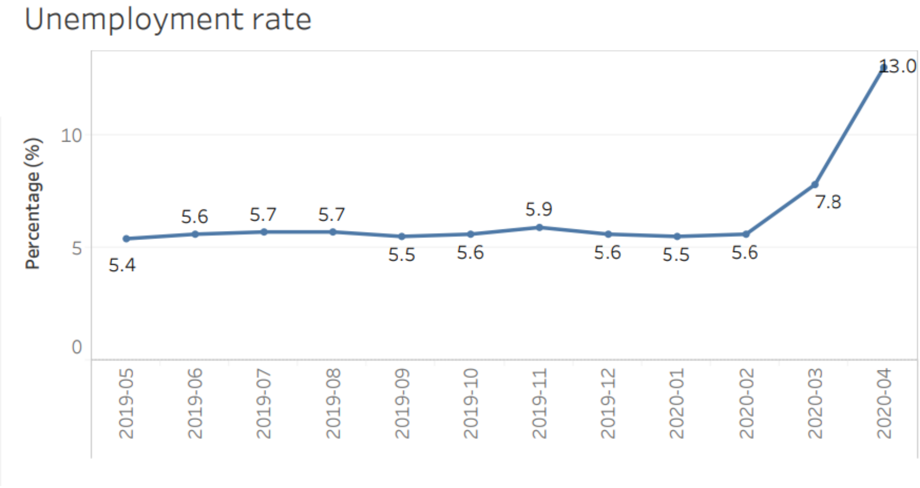
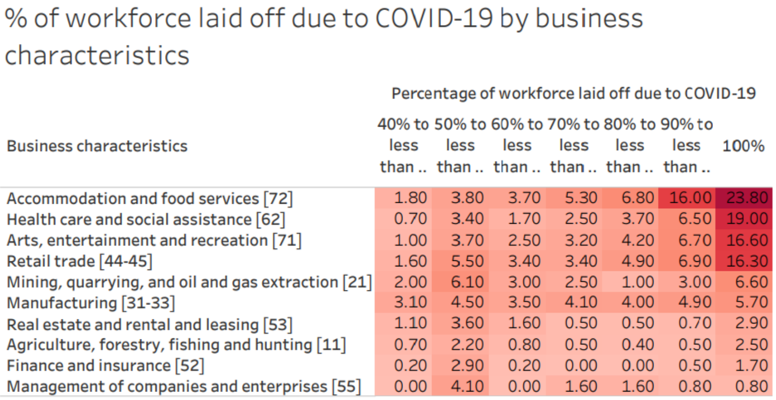


Figure 4: Canadian Unemployment Rate [11]



% of workforce laid off due to COVID-19 by business characteristics

Business characteristics	Percentage of workforce laid off due to COVID-19 40% to less than ..	50% to less than ..	60% to less than ..	70% to less than ..	80% to less than ..	90% to less than ..	100%
Accommodation and food services [72]	1.80	3.80	3.70	5.30	6.80	16.00	23.80
Health care and social assistance [62]	0.70	3.40	1.70	2.50	3.70	6.50	19.00
Arts, entertainment and recreation [71]	1.00	3.70	2.50	3.20	4.20	6.70	16.60
Retail trade [44-45]	1.60	5.50	3.40	3.40	4.90	6.90	16.30
Mining, quarrying, and oil and gas extraction [21]	2.00	6.10	3.00	2.50	1.00	3.00	6.60
Manufacturing [31-33]	3.10	4.50	3.50	4.10	4.00	4.90	5.70
Real estate and rental and leasing [53]	1.10	3.60	1.60	0.50	0.50	0.70	2.90
Agriculture, forestry, fishing and hunting [11]	0.70	2.20	0.80	0.50	0.40	0.50	2.50
Finance and insurance [52]	0.20	2.90	0.20	0.00	0.00	0.50	1.70
Management of companies and enterprises [55]	0.00	4.10	0.00	1.60	1.60	0.80	0.80

Figure 5: Canadian Unemployment Levels per Sector [12]

Many industries already possessed large numbers of unemployed individuals as well, retail and hospitality most especially (Appendix D), and among many industries, a disparity existed wherein women were more severely affected by unemployment than men were before the pandemic.

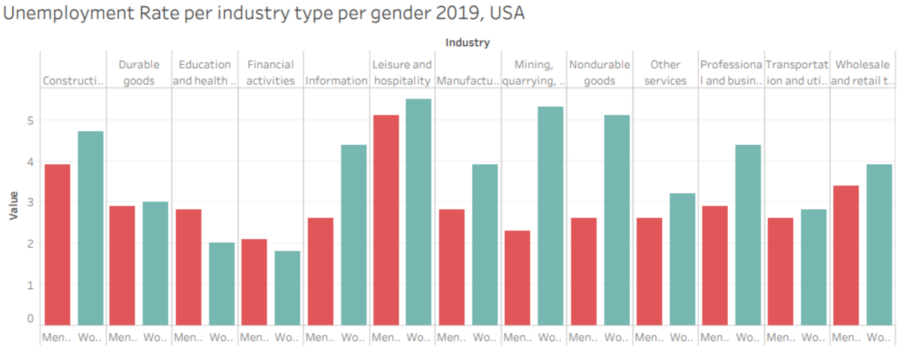


Figure 6: American Unemployment Rate by Gender, 2019 [13]

These issues may naturally create a desire among many of North America's unemployed to shift into more economically prosperous fields, both to provide better crisis job security and more likely employment overall. Here we can see that software related jobs stand out significantly as being desirable (see Appendix D), and as well it can be seen that information services experienced only a modest increase in unemployment between 2019 and 2020 (Appendix D).

However, for an individual from an outside industry or possessing a low level of education, such a transition may be particularly difficult. An individual may not know where to look to find a job in a field such as software or data science, and even if a job is found they may be lacking in the necessary skills to apply for and succeed in such a role. Seeing as the goals for the graduate level MIE1624 course design and for the design of the Data Science Master's program were to tailor courses to teach employable skills, the challenge now is how to provide enough skills to what will be assumed to be an unskilled worker, potentially advanced in age, such that they can apply for new jobs in the data science field.

Recommendations for a Startup

The startup will function along similar lines that the course curricula were designed earlier in the report. Once an individual signs up for the service, they will register what skills they have (including none), and then indicate their location. The program will then query what jobs are available within a selected radius of the individual that do not require formal education (as to keep opportunities open to those without one) and that are offering entry level jobs. Using web scraping of online job websites, the most essential skills from this pool of jobs will be analyzed and ranked according to their importance and their difficulty, much as skills were evaluated for the design of MIE1624.

At this point, the disparity between the skills the applicant currently possesses and the skills required for a set of jobs will be evaluated. It is the resolution of this disparity that is essential to the success of the startup. By eliminating it altogether, an individual will be emancipated to apply for jobs in data science and begin acquiring the experience necessary to establish themselves as career data scientists.

The startup will not seek to provide those skills per se, but rather to connect an individual with the online resources that can allow them to acquire them. Using a wide variety of educational websites such as Coursera [6], courses will be web scraped in order to find what skills they offer, and a set of courses from various sources will be presented to the individual that allows them to comprehensively address their skill deficit. Here, sensitivity must be observed, as an employed individual will not be able to spend large amounts of money applying for courses. The program will thus test a combination of courses to find the most cost effective program for skill acquisition. Additionally, courses will be presented in an order according to difficulty in order to lessen the learning curve and not scare away someone who is new to these fields.

This startup may not only prove lucrative for tapping into a large sector of the unemployed population, but it also functions as a social service, helping to ease the economic suffering of countless people. Additionally, if it succeeds in reducing unemployment, it may receive significant positive attention as a force for economic good and attract investment from outside.

As for the possible future considerations of our startup, the team can investigate analyzing risks, especially from cyber-attacks, to make sure that the collected data is properly secured. Another aspect to investigate are the competitors in the market and the way the market share is distributed among them. The legal aspect should also be further investigated to verify what data the startup can collect from individuals. Finally, our group can look into forming partnerships or merging and acquiring an already established organization in the field. To do so, the group needs to verify the financial capabilities of the startup at the beginning (from funds, affiliate marketing, or selling advertisement space). Thus, the team has to verify if it would be more profitable to operate independently or under an already established organization, where our service is launched as an independent route in it.

Conclusion

Based on the proposals above, the group believes it is possible to greatly improve education outcomes through the use of data. By redesigning the MIE1624 course and a Data Science Master’s program to reflect the most useful skills in industry, it can be ensured that students are more apt to be competitive as employees, reflecting well on the University's reputation. The effect of such changes may drive students to the University of Toronto seeking an education that will prove more lucrative after graduating. Additionally, the startup proposed by the group seeks to take a more non-traditional approach to education and reach into new markets. This startup turns previously low skilled individuals into potential data scientists, offering many a means to escape unemployment while generating a positive social outcome that will reflect well on the group.

References

[1] A. Kelley, “Biden tells coal miners to “learn to code””, *The Hill*, 2019. [Online]. Available: https://thehill.com/changing-america/enrichment/education/476391-biden-tells-coal-miners-to-learn-to-code. [Accessed: 28- Nov- 2020].

[2] R. Soave, “Yes, You Can Get Kicked Off Twitter for Saying ‘Learn To Code’ - Even If It’s Not Harassment”, *reason*, 2019. [Online]. Available: https://reason.com/2019/03/11/learn-to-code-twitter-harassment-ross/. [Accessed: 28- Nov- 2020].

[3] World Economic Forum, “The Future of Jobs Report 2020,” World Economic Forum, Cologny/Geneva, Geneva, CHE, 2020.

[4] Kaggle, “2018 Kaggle ML & DS Survey.” (Nov. 3, 2018). Distributed by Kaggle. https://www.kaggle.com/kaggle/kaggle-survey-2018 (accessed October 25th, 2020).

[5] Indeed, “indeed.” indeed. https://ca.indeed.com/ (accessed October 25th, 2020).

[6] Coursera Inc. “coursera.” coursera.org. https://www.coursera.org/ (accessed October 25th, 2020).

[7] O. Romanko, “MIE1624HF - Introduction to Data Science and Analytics.” Toronto, ON, Canada, 2020.

[8] scikit-learn developers, “API Reference.” scikit learn. https://scikit-learn.org/stable/modules/classes.html (accessed October 25th, 2020).

[9] B. Marr, "The 10 Best AI And Data Science Master’s Courses For 2021", *Forbes*, 2020. [Online]. Available: https://www.forbes.com/sites/bernardmarr/2020/07/20/the-10-best-ai-and-data-science-masters-courses-for-2021/?sh=5446891049a3. [Accessed: 29- Nov- 2020].

[10] Lucid Software Inc., “Lucidchart.” Lucidchart. https://www.lucidchart.com (accessed November 28th, 2020).

[11] Statistics Canada, “Labour force characteristics by province, monthly, seasonally adjusted.” (n.d.). Distributed by Statistics Canada. https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=1410028703 (accessed November 28th, 2020).

[12] Statistics Canada, “Percentage of workforce laid off because of COVID-19, by business characteristics.” (n.d.). Distributed by the Government of Canada. https://open.canada.ca/data/en/dataset/4c6d8b07-af8b-46fb-8445-55f4dea10d36 (accessed November 28th, 2020).

[13] U.S. Bureau of Labor Statistics, “Labor Force Statistics from the Current Population Survey.” (Nov., 6th, 2020). Distributed by U.S. Bureau of Labor Statistics. https://www.bls.gov/web/empsit/cpseea31.htm (accessed November 28th, 2020).

[14] Coggle, “coggle.” coggle. https://coggle.it/ (accessed November 28th, 2020).

[15] Urban Institute, “Estimated Low Income Jobs Lost to COVID-19: sum_job_loss_cbsa.csv.” (n.d.). Distributed by Urban Institute. https://datacatalog.urban.org/dataset/estimated-low-income-jobs-lost-covid-19/resource/faa4c555-d22f-48da-a1d2-cb1c62e8d527 (accessed November 28th, 2020).

[16] Linkedin, “Insight into a rapidly changing economy.” (Sept., 2020). Distributed by Linkedin’s Economic Graph. https://graph.linkedin.com/insights/labor-market (accessed November 28th, 2020).

[17] Statistics Canada, “Staffing actions taken by businesses during the COVID-19 pandemic, by business characteristics.” (n.d.). Distributed by the Government of Canada. https://open.canada.ca/data/en/dataset/ea96aff3-0d4f-4a7d-82cb-ab06dc11f1b2 (accessed November 28th, 2020).

Appendix A: Visualizations of MIE1624 Course Development

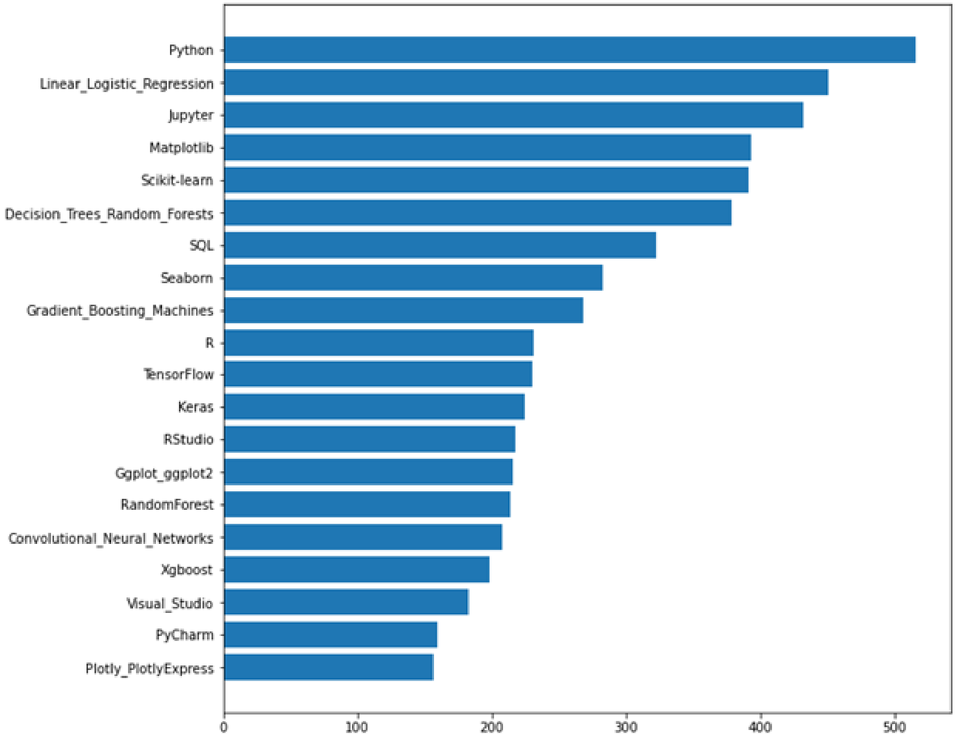


Figure 7: Bar plot of Most Common 20 Skills for Recent Master's Graduates earning over $80,000 [4]

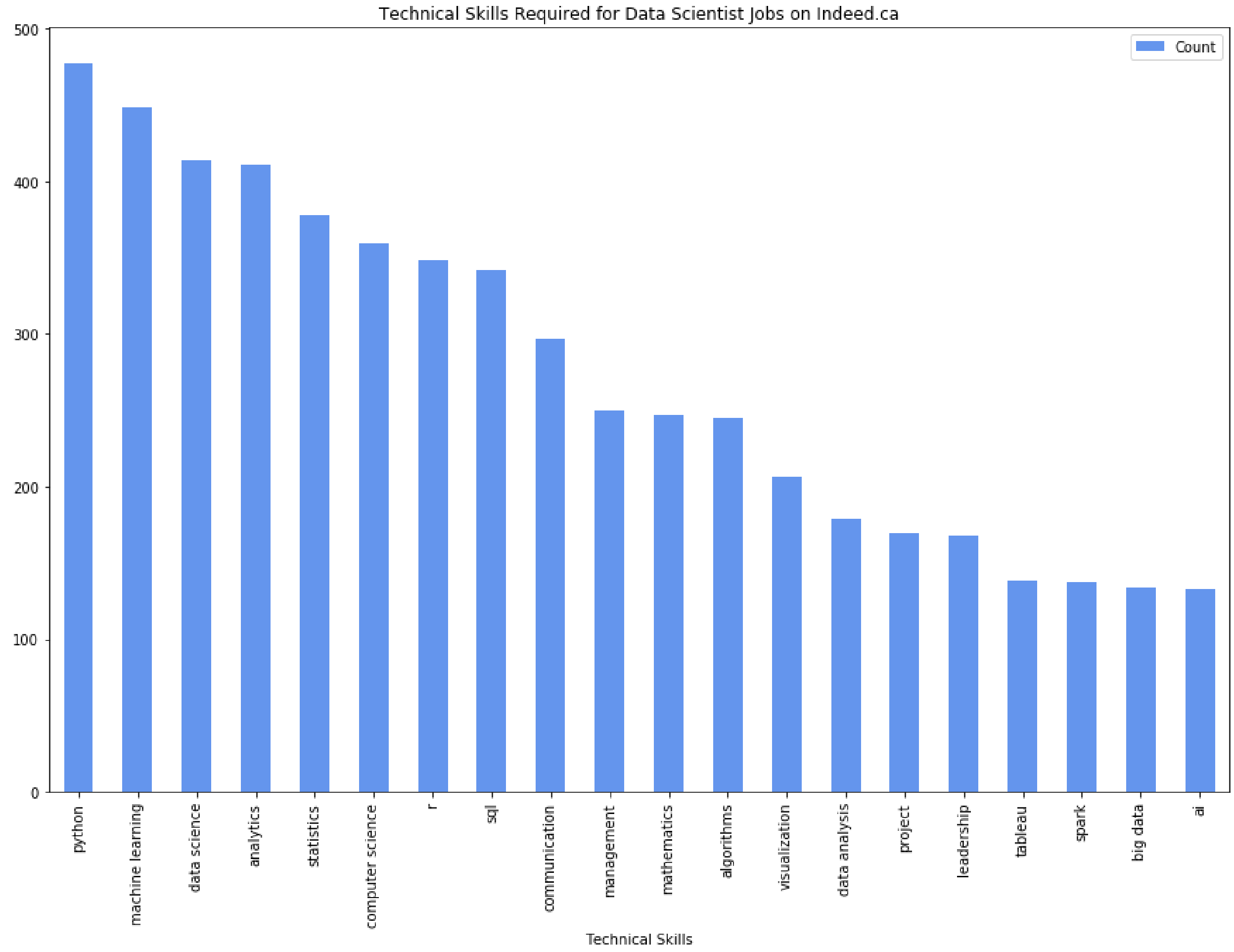


Figure 8: Technical Skills Required for Data Scientist Jobs on Indeed.ca [5]

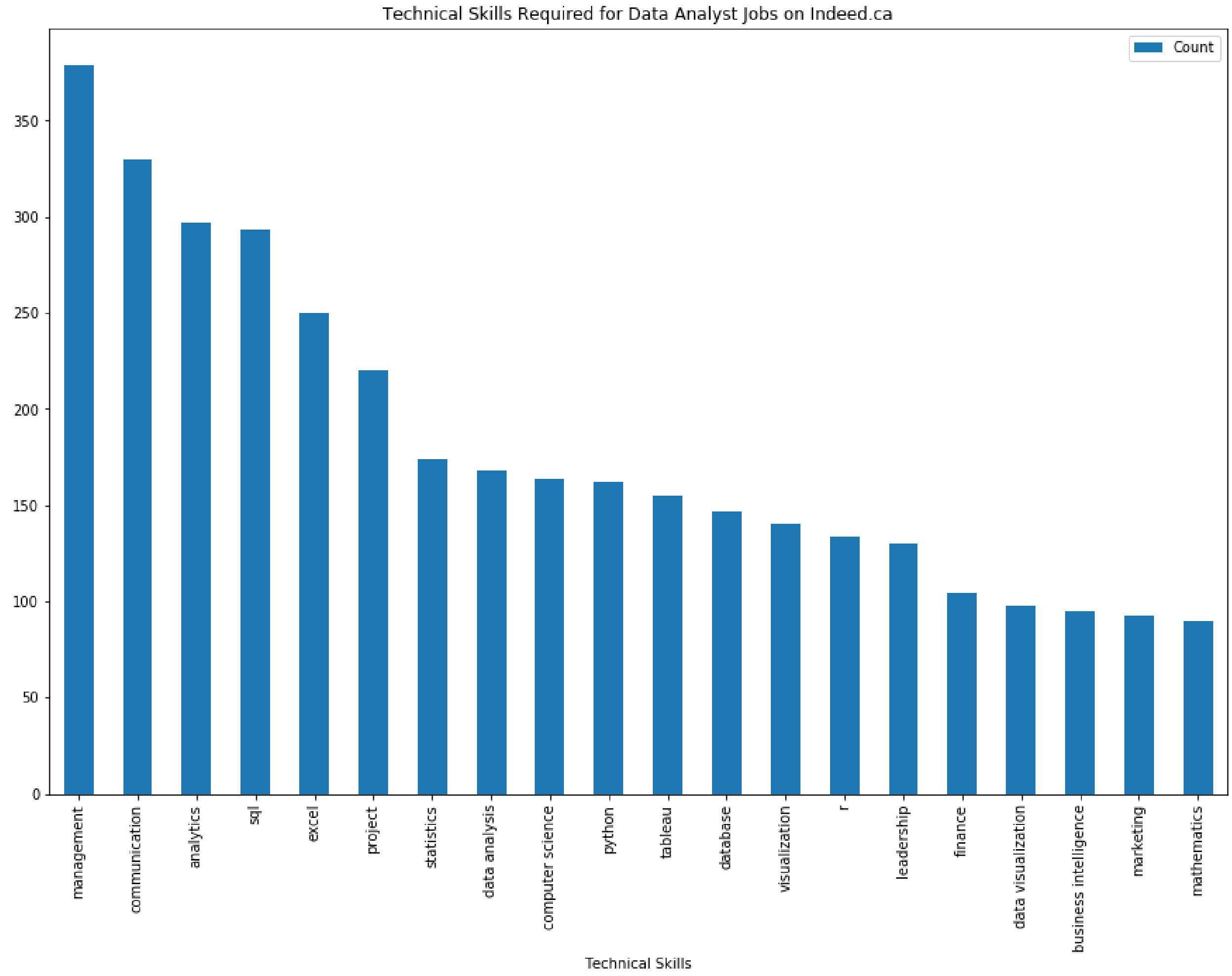


Figure 9: Technical Skills Required for Data Analyst Jobs on Indeed.ca [5]

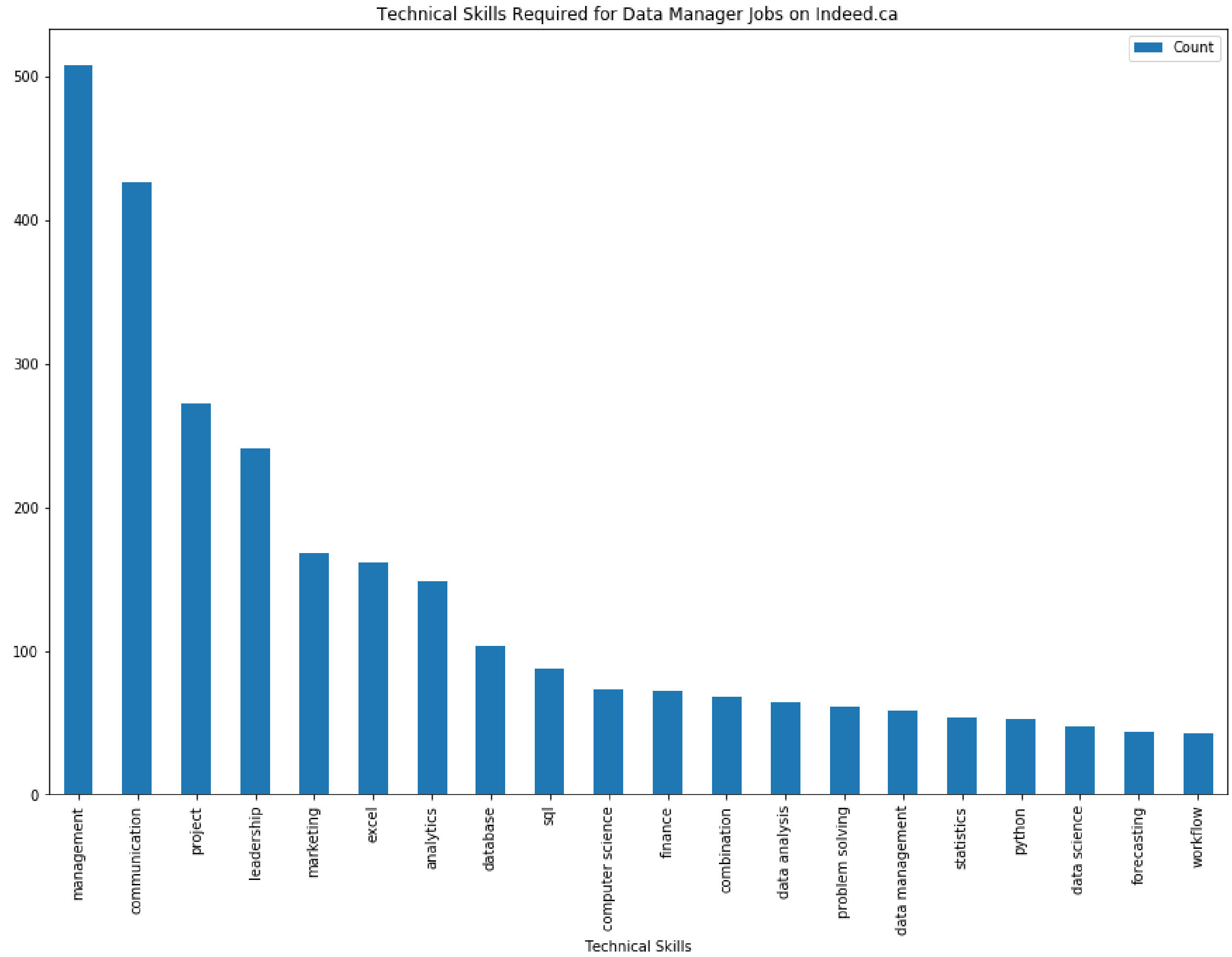


Figure 10: Technical Skills Required for Data Manager Jobs on Indeed.ca [5]

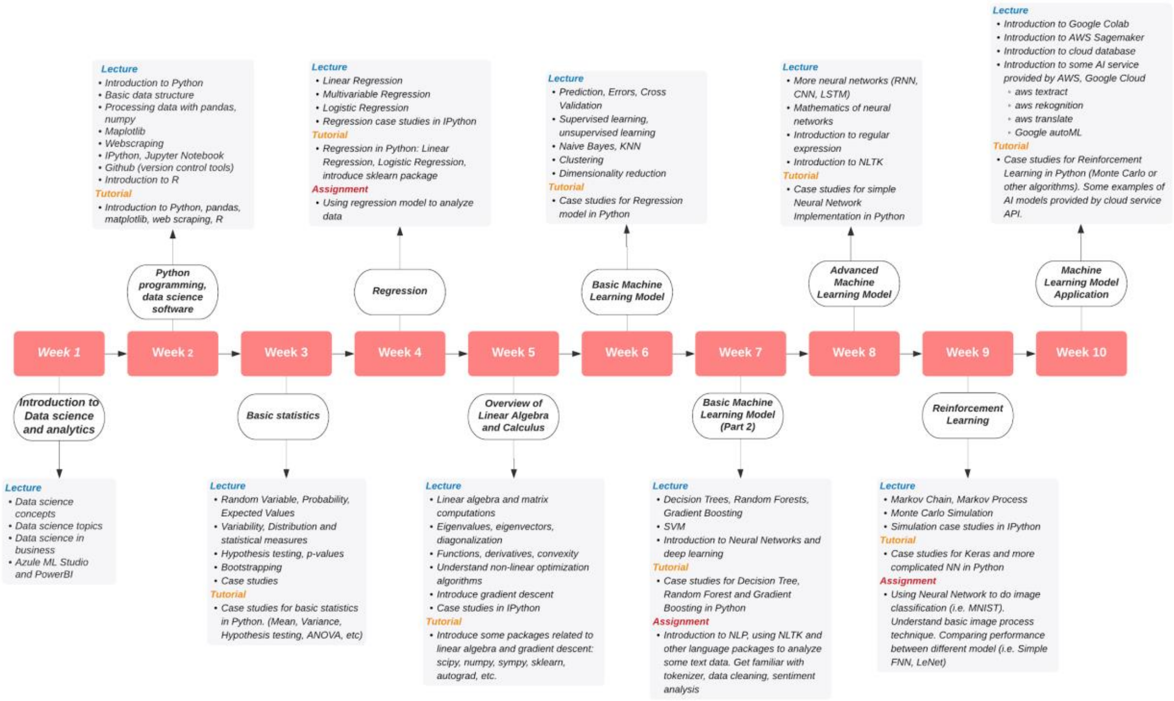

Week 1
Introduction to Data science and analytics
Lecture
• Data science concepts
• Data science topics
• Data science in business
• Azule ML Studio and PowerBI
Week 2
Python programming, data science software
Lecture
• Introduction to Python
• Basic data structure
• Processing data with pandas, numpy
• Maplotlib
• Webscraping
• IPython, Jupyter Notebook
• Github (version control tools)
• Introduction to R
Tutorial
• Introduction to Python, pandas, matplotlib, web scraping, R
Week 3
Basic statistics
Lecture
• Random Variable, Probability, Expected Values
• Variability, Distribution and statistical measures
• Hypothesis testing, p-values
• Bootstrapping
• Case studies
Tutorial
• Case studies for basic statistics in Python. (Mean, Variance, Hypothesis testing, ANOVA, etc)
Week 4
Regression
Lecture
• Linear Regression
• Multivariable Regression
• Logistic Regression
• Regression case studies in IPython
Tutorial
• Regression in Python: Linear Regression, Logistic Regression, introduce sklearn package
Assignment
• Using regression model to analyze data
Week 5
Overview of Linear Algebra and Calculus
Lecture
• Linear algebra and matrix computations
• Eigenvalues, eigenvectors, diagonalization
• Functions, derivatives, convexity
• Understand non-linear optimization algorithms
• Introduce gradient descent
• Case studies in IPython
Tutorial
• Introduce some packages related to linear algebra and gradient descent: scipy, numpy, sympy, sklearn, autograd, etc.
Week 6
Basic Machine Learning Model
Lecture
• Prediction, Errors, Cross Validation
• Supervised learning, unsupervised learning
• Naive Bayes, KNN
• Clustering
• Dimensionality reduction
Tutorial
• Case studies for Regression model in Python
Week 7
Basic Machine Learning Model (Part 2)
Lecture
• Decision Trees, Random Forests, Gradient Boosting
• SVM
• Introduction to Neural Networks and deep learning
Tutorial
• Case studies for Decision Tree, Random Forest and Gradient Boosting in Python
Assignment
• Introduction to NLP, using NLTK and other language packages to analyze some text data. Get familiar with tokenizer, data cleaning, sentiment analysis
Week 8
Advanced Machine Learning Model
Lecture
• More neural networks (RNN, CNN, LSTM)
• Mathematics of neural networks
• Introduction to regular expression
• Introduction to NLTK
Tutorial
• Case studies for simple Neural Network Implementation in Python
Week 9
Reinforcement Learning
Lecture
• Markov Chain, Markov Process
• Monte Carlo Simulation
• Simulation case studies in IPython
Tutorial
• Case studies for Keras and more complicated NN in Python
Assignment
• Using Neural Network to do image classification (i.e. MNIST). Understand basic image process technique. Comparing performance between different model (i.e. Simple FNN, LeNet)
Week 10
Machine Learning Model Application
Lecture
• Introduction to Google Colab
• Introduction to AWS Sagemaker
• Introduction to cloud database
• Introduction to some AI service provided by AWS, Google Cloud
◦ aws textract
◦ aws rekognition
◦ aws translate
◦ Google autoML
Tutorial
• Case studies for Reinforcement Learning in Python (Monte Carlo or other algorithms). Some examples of AI models provided by cloud service API.

Figure 11: Course curriculum for Introduction to Data Science and Analytics [10]

Appendix B: Visualizations of Master's Program Development

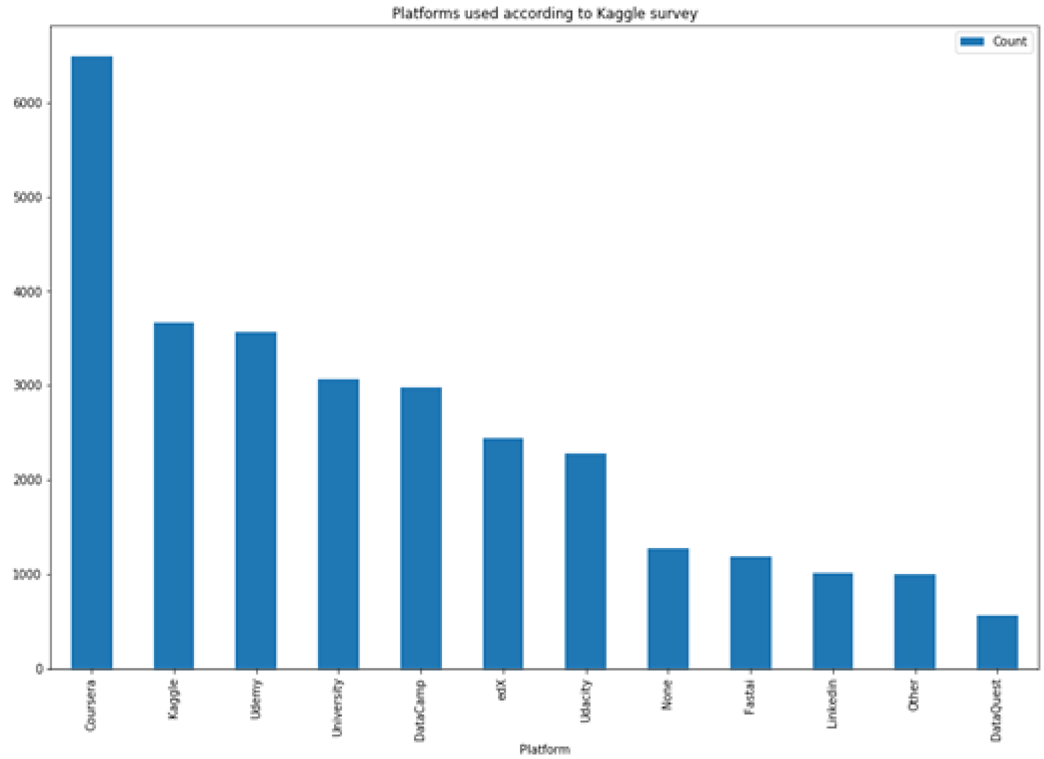


Figure 12: Platforms used according to Kaggle survey [4]

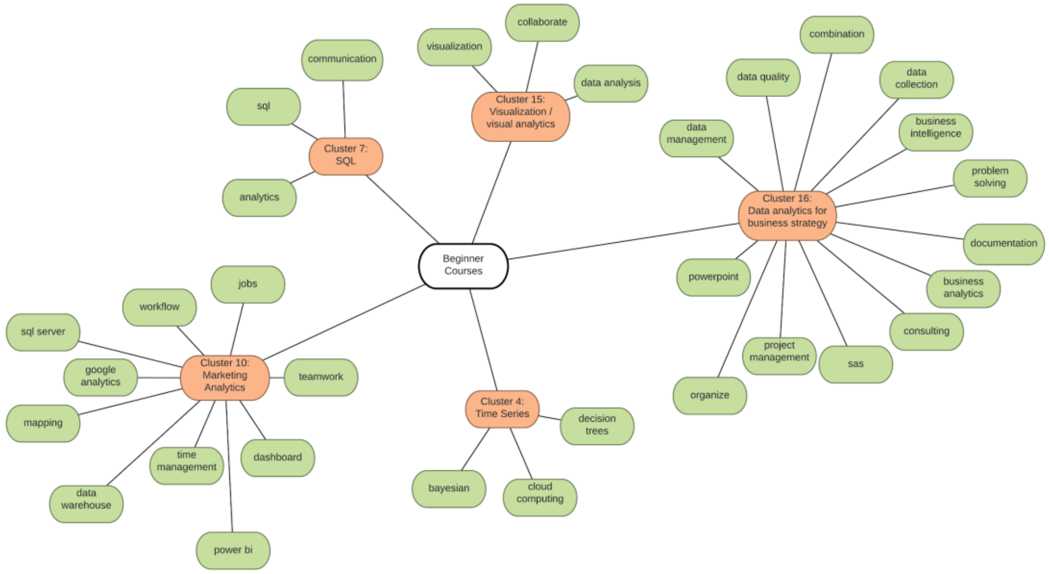


Figure 13: Word net showing the five beginner courses and clustered skills obtained from hierarchical clustering.

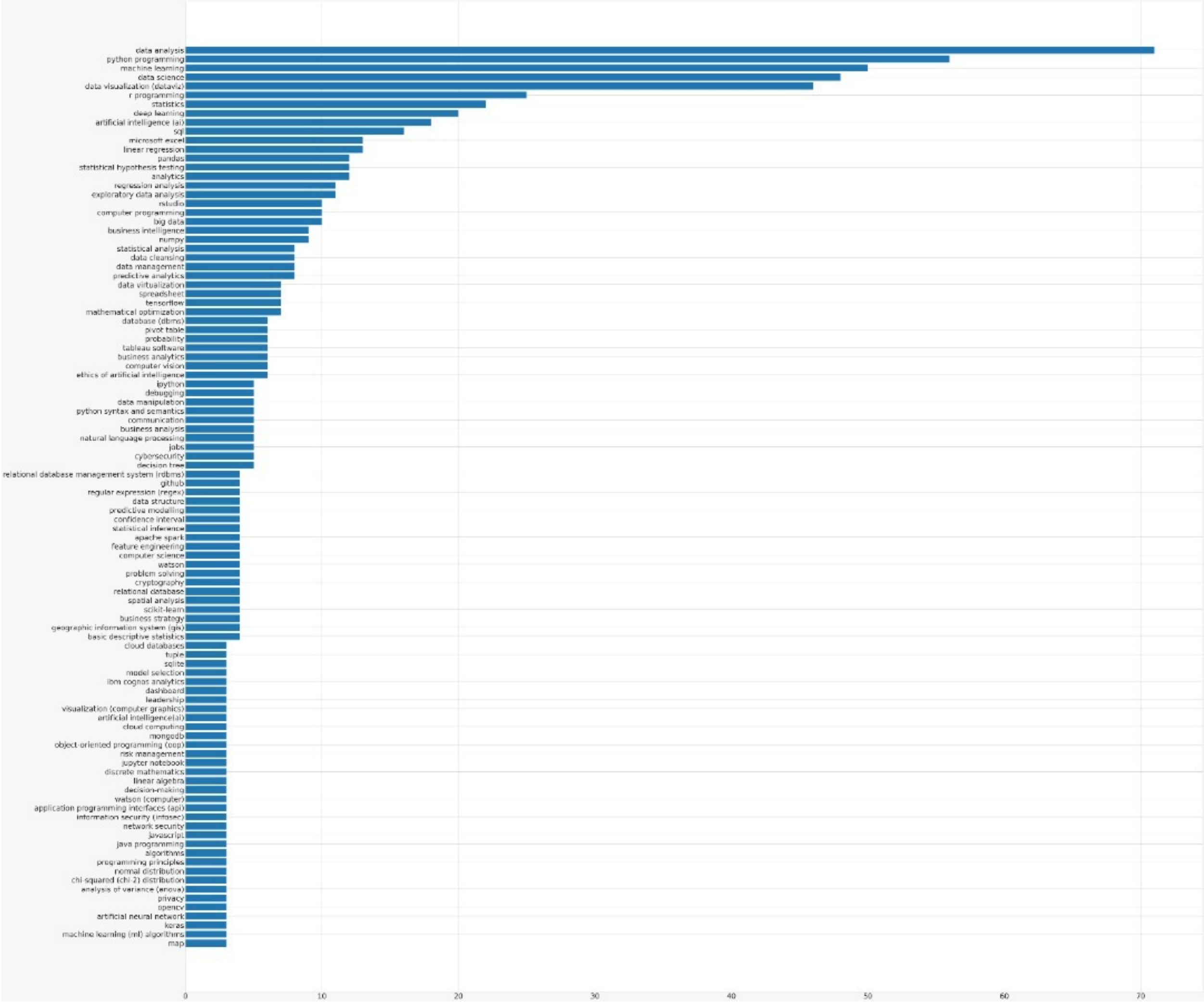


Figure 14: Bar graph of the most common skills in Coursera [6]

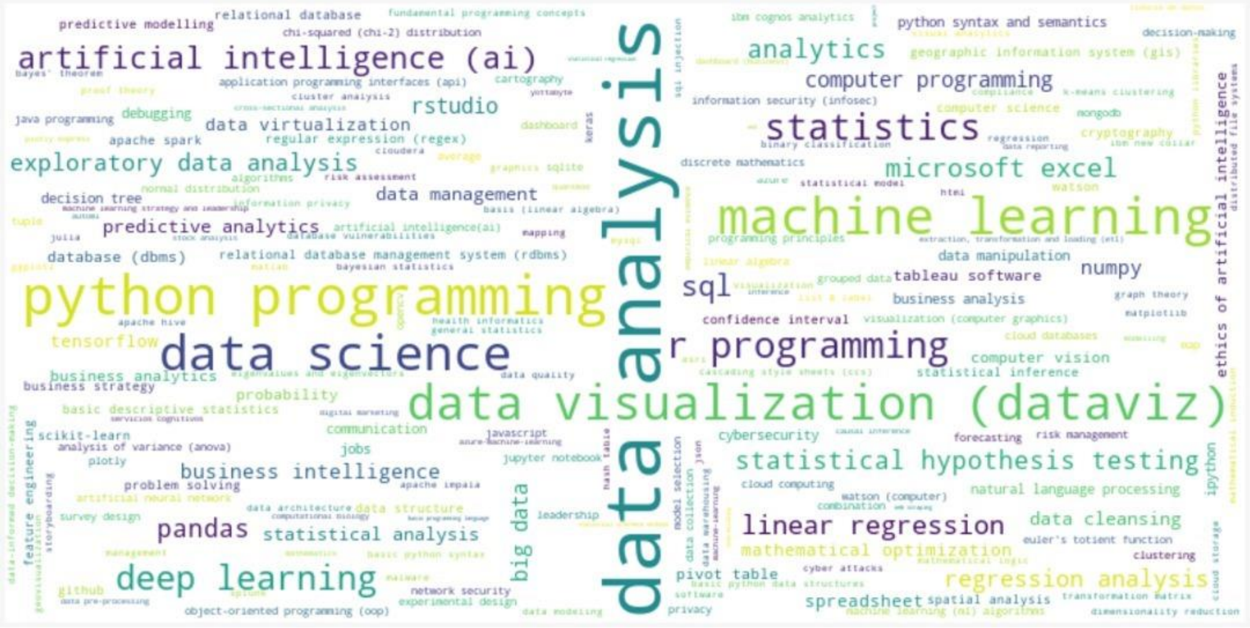


Figure 15: Word cloud of most common words in Coursera [6]

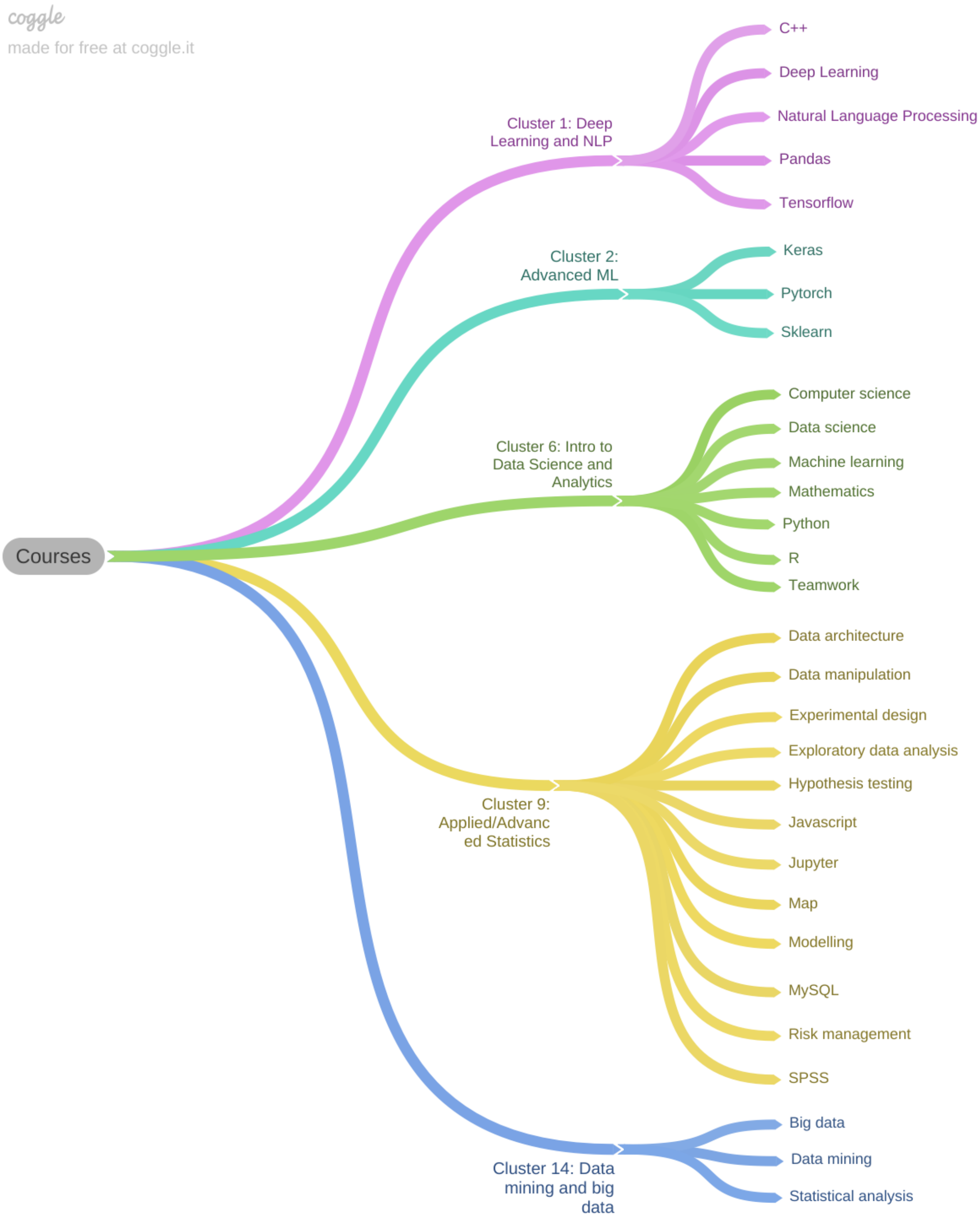
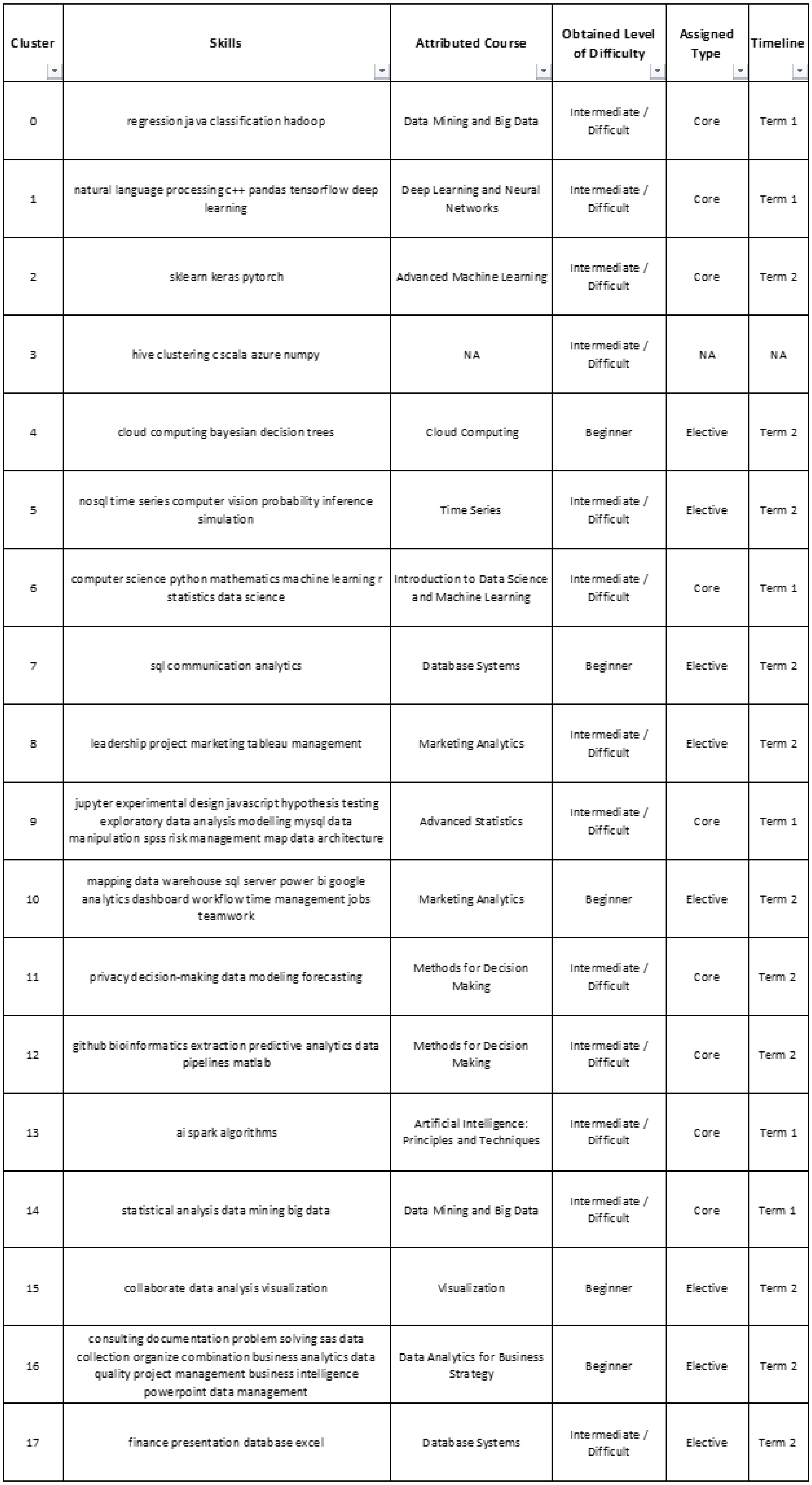


Figure 16: Visualization of five courses/clusters [14]



Cluster	Skills	Attributed Course	Obtained Level of Difficulty	Assigned Type	Timeline
0	regression java classification hadoop	Data Mining and Big Data	Intermediate / Difficult	Core	Term 1
1	natural language processing c++ pandas tensorflow deep learning	Deep Learning and Neural Networks	Intermediate / Difficult	Core	Term 1
2	sklearn keras pytorch	Advanced Machine Learning	Intermediate / Difficult	Core	Term 2
3	hive clustering c scala azure numpy	NA	Intermediate / Difficult	NA	NA
4	cloud computing bayesian decision trees	Cloud Computing	Beginner	Elective	Term 2
5	nosql time series computer vision probability inference simulation	Time Series	Intermediate / Difficult	Elective	Term 2
6	computer science python mathematics machine learning r statistics data science	Introduction to Data Science and Machine Learning	Intermediate / Difficult	Core	Term 1
7	sql communication analytics	Database Systems	Beginner	Elective	Term 2
8	leadership project marketing tableau management	Marketing Analytics	Intermediate / Difficult	Elective	Term 2
9	jupyter experimental design javascript hypothesis testing exploratory data analysis modelling mysql data manipulation spss risk management map data architecture	Advanced Statistics	Intermediate / Difficult	Core	Term 1
10	mapping data warehouse sql server power bi google analytics dashboard workflow time management jobs teamwork	Marketing Analytics	Beginner	Elective	Term 2
11	privacy decision-making data modeling forecasting	Methods for Decision Making	Intermediate / Difficult	Core	Term 2
12	github bioinformatics extraction predictive analytics data pipelines matlab	Methods for Decision Making	Intermediate / Difficult	Core	Term 2
13	ai spark algorithms	Artificial Intelligence: Principles and Techniques	Intermediate / Difficult	Core	Term 1
14	statistical analysis data mining big data	Data Mining and Big Data	Intermediate / Difficult	Core	Term 1
15	collaborate data analysis visualization	Visualization	Beginner	Elective	Term 2
16	consulting documentation problem solving sas data collection organize combination business analytics data quality project management business intelligence powerpoint data management	Data Analytics for Business Strategy	Beginner	Elective	Term 2
17	finance presentation database excel	Database Systems	Intermediate / Difficult	Elective	Term 2

Figure 17: Course Clusterings and Attributed Skills

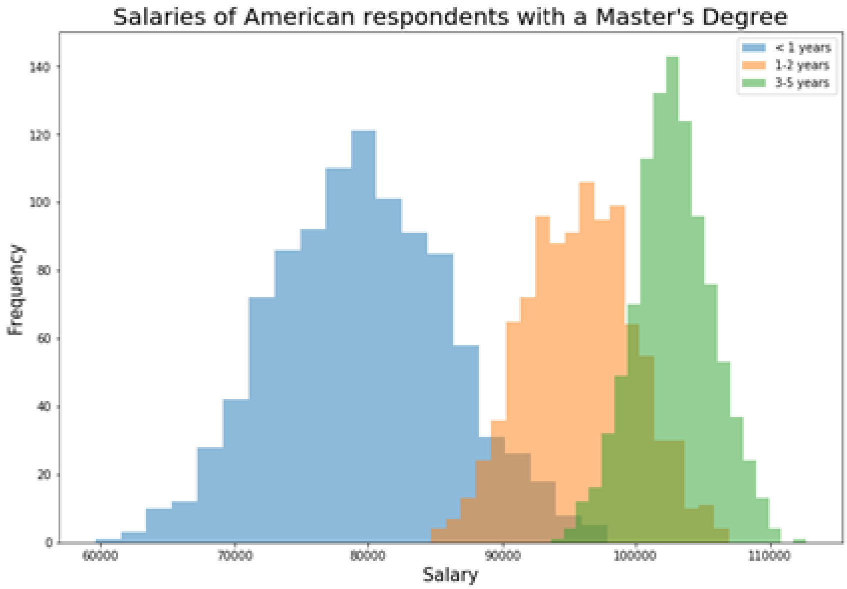


Figure 18: Histogram of Bootstrapped Mean Salaries for American Master's Degree Holders According to Experience [4]

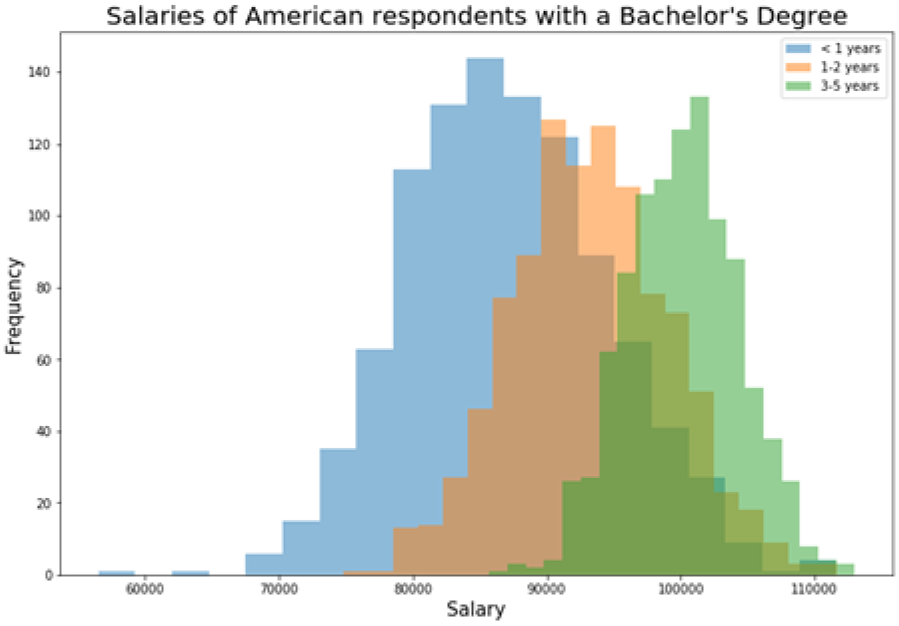


Figure 19: Histogram of Bootstrapped Mean Salaries for American Bachelor's Degree Holders According to Experience [4]

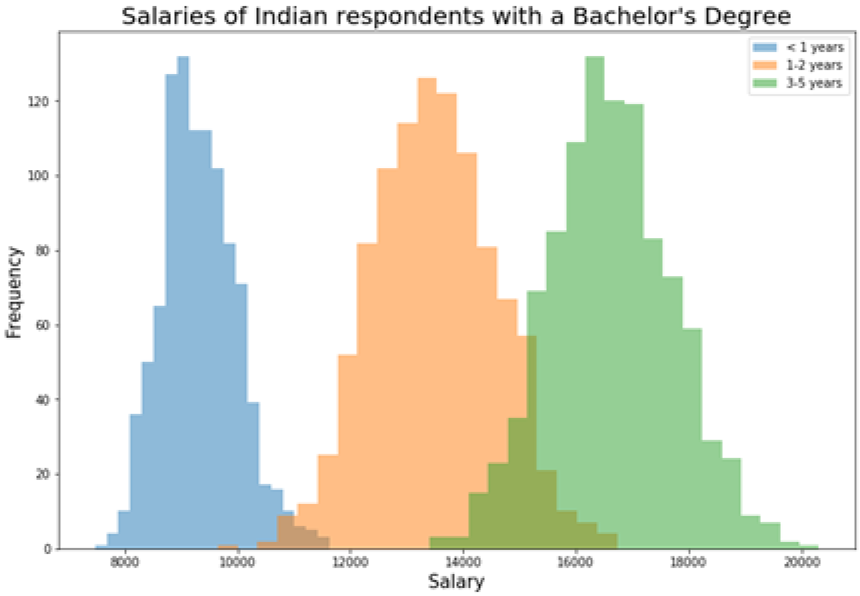


Figure 20: Histogram of Bootstrapped Mean Salaries for Indian Bachelor's Degree Holders According to Experience [4]

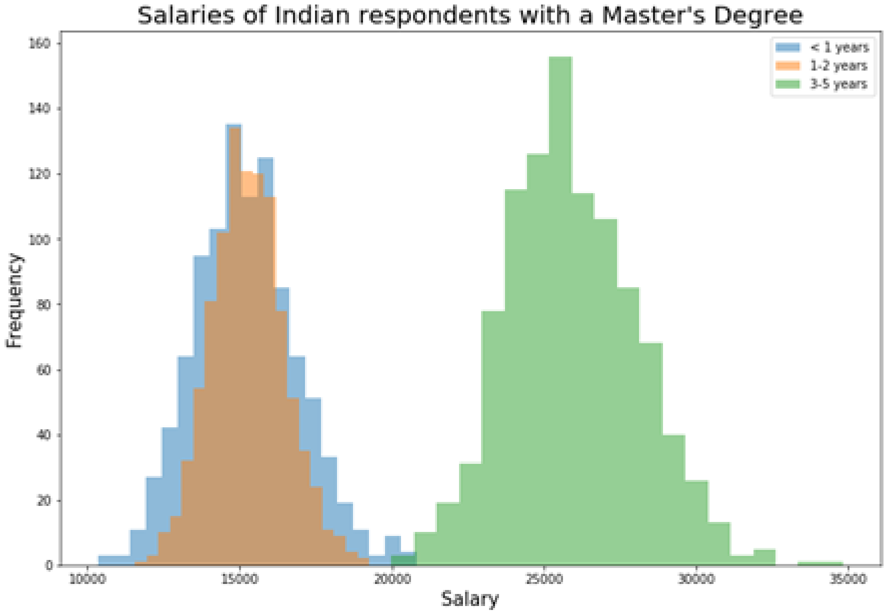


Figure 21: Histogram of Bootstrapped Mean Salaries for Indian Master's Degree Holders According to Experience [4]

Appendix C: Master’s Program Course Descriptions

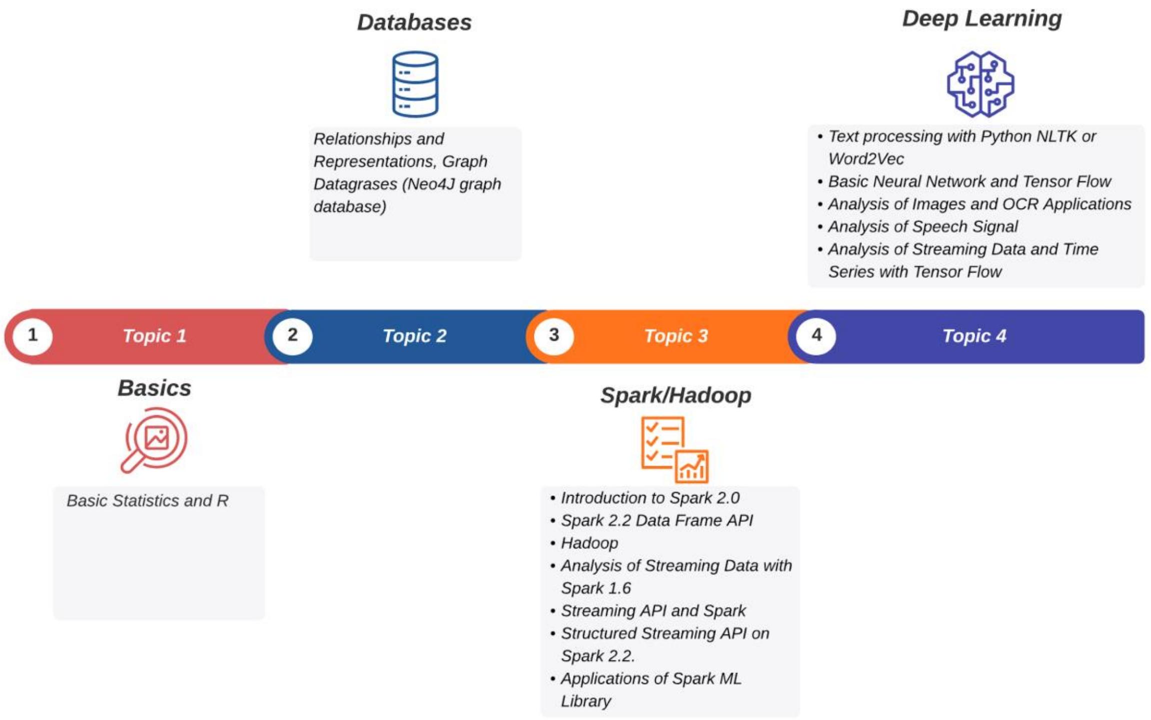

Databases
Relationships and Representations, Graph Datagrases (Neo4J graph database)
Deep Learning
• Text processing with Python NLTK or Word2Vec
• Basic Neural Network and Tensor Flow
• Analysis of Images and OCR Applications
• Analysis of Speech Signal
• Analysis of Streaming Data and Time Series with Tensor Flow
1 Topic 1
2 Topic 2
3 Topic 3
4 Topic 4
Basics
Basic Statistics and R
Spark/Hadoop
• Introduction to Spark 2.0
• Spark 2.2 Data Frame API
• Hadoop
• Analysis of Streaming Data with Spark 1.6
• Streaming API and Spark
• Structured Streaming API on Spark 2.2.
• Applications of Spark ML Library

Figure 22: Course curriculum for Big Data [10]

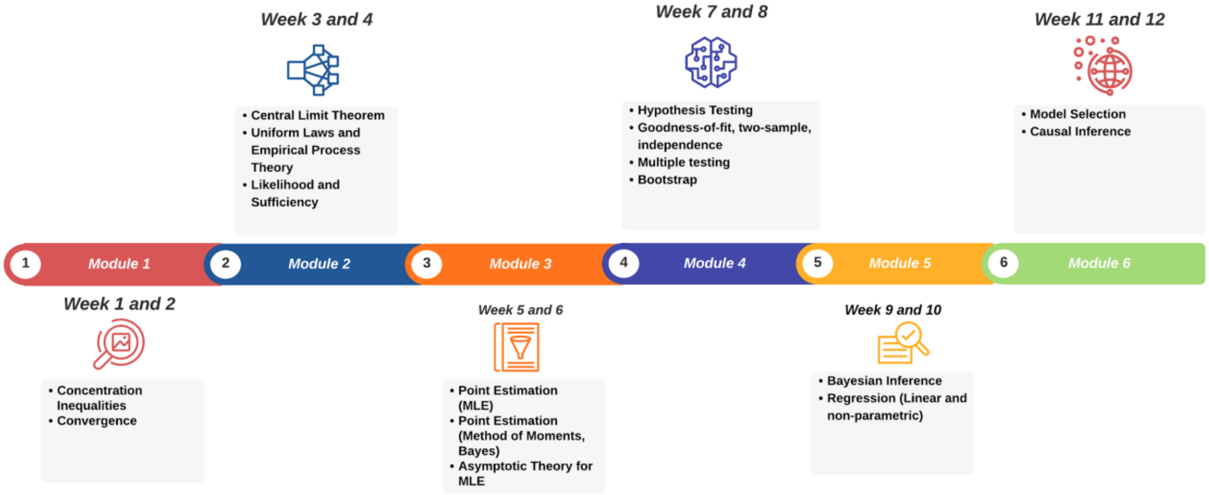
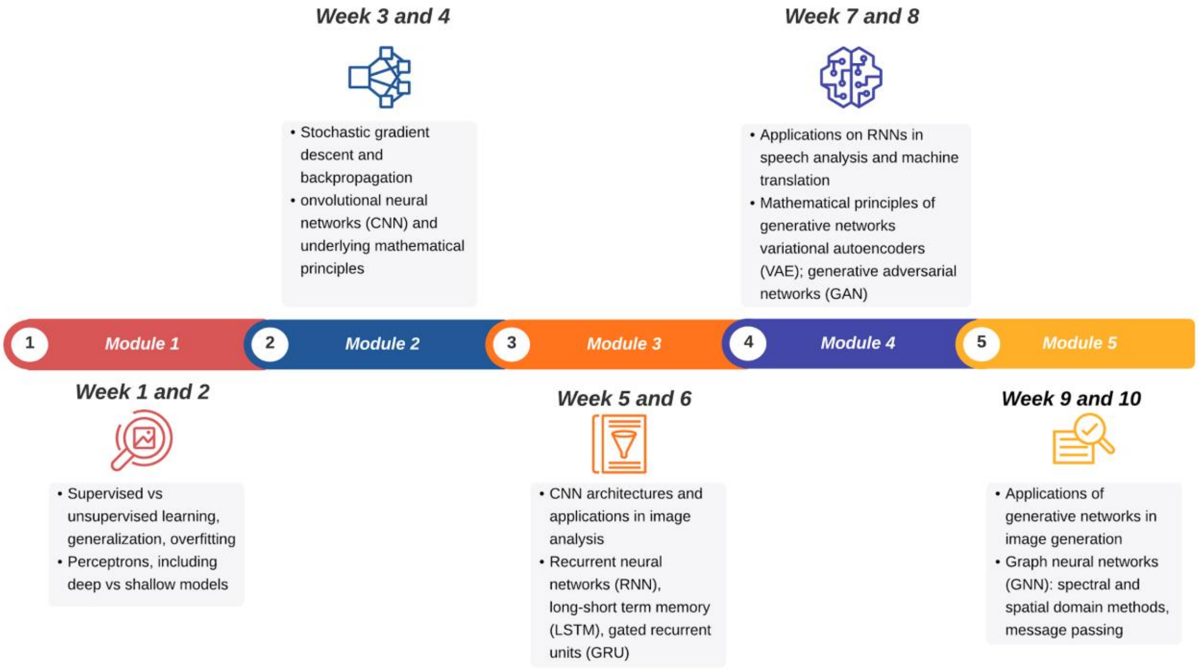


Figure 23: Course curriculum for Advanced Statistics [10]



1 Module 1

Week 1 and 2

- Supervised vs unsupervised learning, generalization, overfitting
- Perceptrons, including deep vs shallow models

2 Module 2

Week 3 and 4

- Stochastic gradient descent and backpropagation
- onvolutional neural networks (CNN) and underlying mathematical principles

3 Module 3

Week 5 and 6

- CNN architectures and applications in image analysis
- Recurrent neural networks (RNN), long-short term memory (LSTM), gated recurrent units (GRU)

4 Module 4

Week 7 and 8

- Applications on RNNs in speech analysis and machine translation
- Mathematical principles of generative networks variational autoencoders (VAE); generative adversarial networks (GAN)

5 Module 5

Week 9 and 10

- Applications of generative networks in image generation
- Graph neural networks (GNN): spectral and spatial domain methods, message passing

Figure 24: Course curriculum for Deep Learning [10]

Appendix D: Visualizations of “Learn to Code” [1][2] Program Development

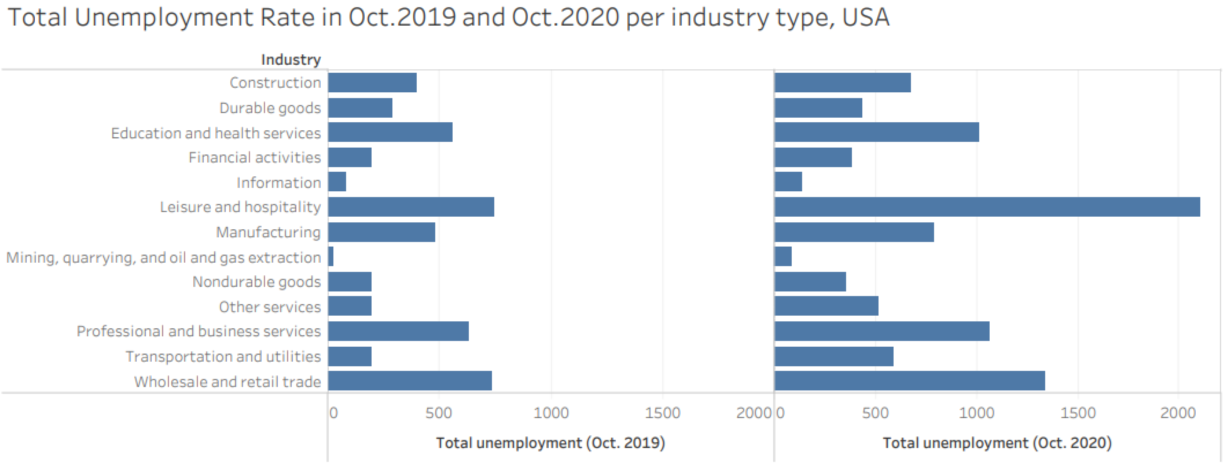


Figure 25: Total Unemployment Rate in Oct. 2019 and Oct. 2020 per industry type, USA [13]

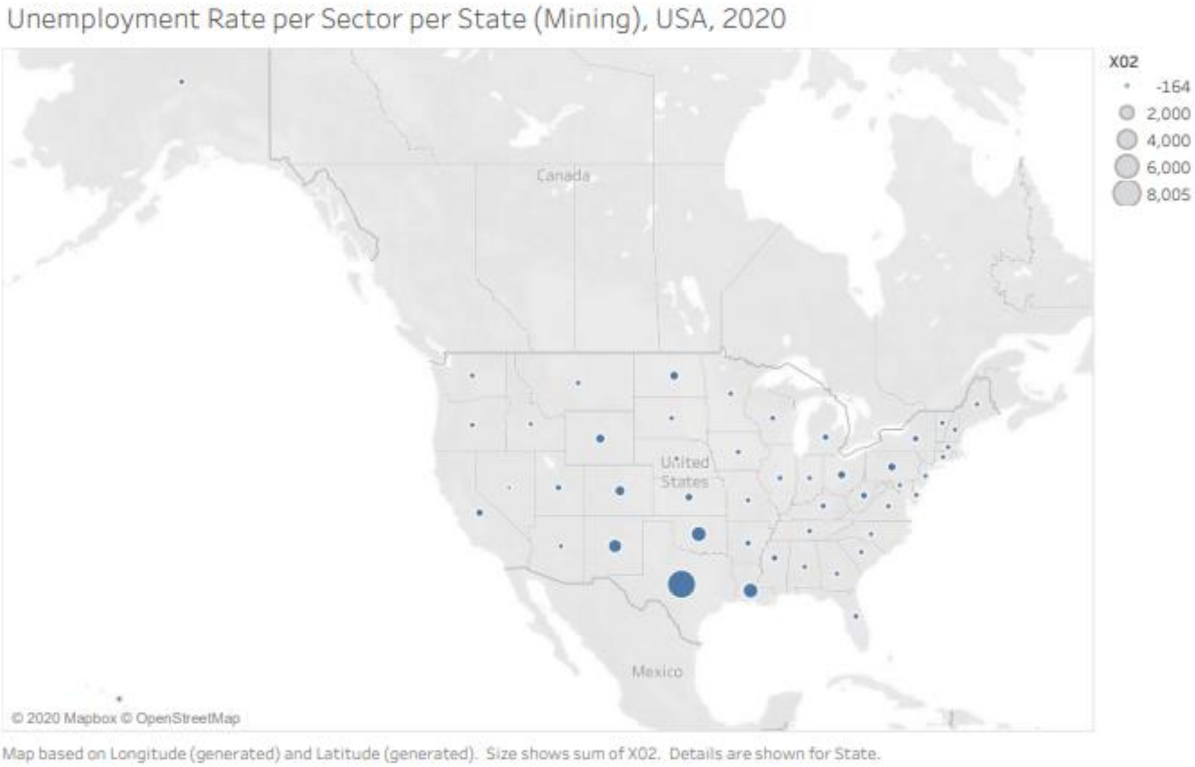

Unemployment Rate per Sector per State (Mining), USA, 2020
X02
-164
2,000
4,000
6,000
8,005
Canada
United States
Mexico
© 2020 Mapbox © OpenStreetMap
Map based on Longitude (generated) and Latitude (generated). Size shows sum of X02. Details are shown for State.

Figure 26: Unemployment Rate per Sector per State (Mining), USA, 2020 [15]

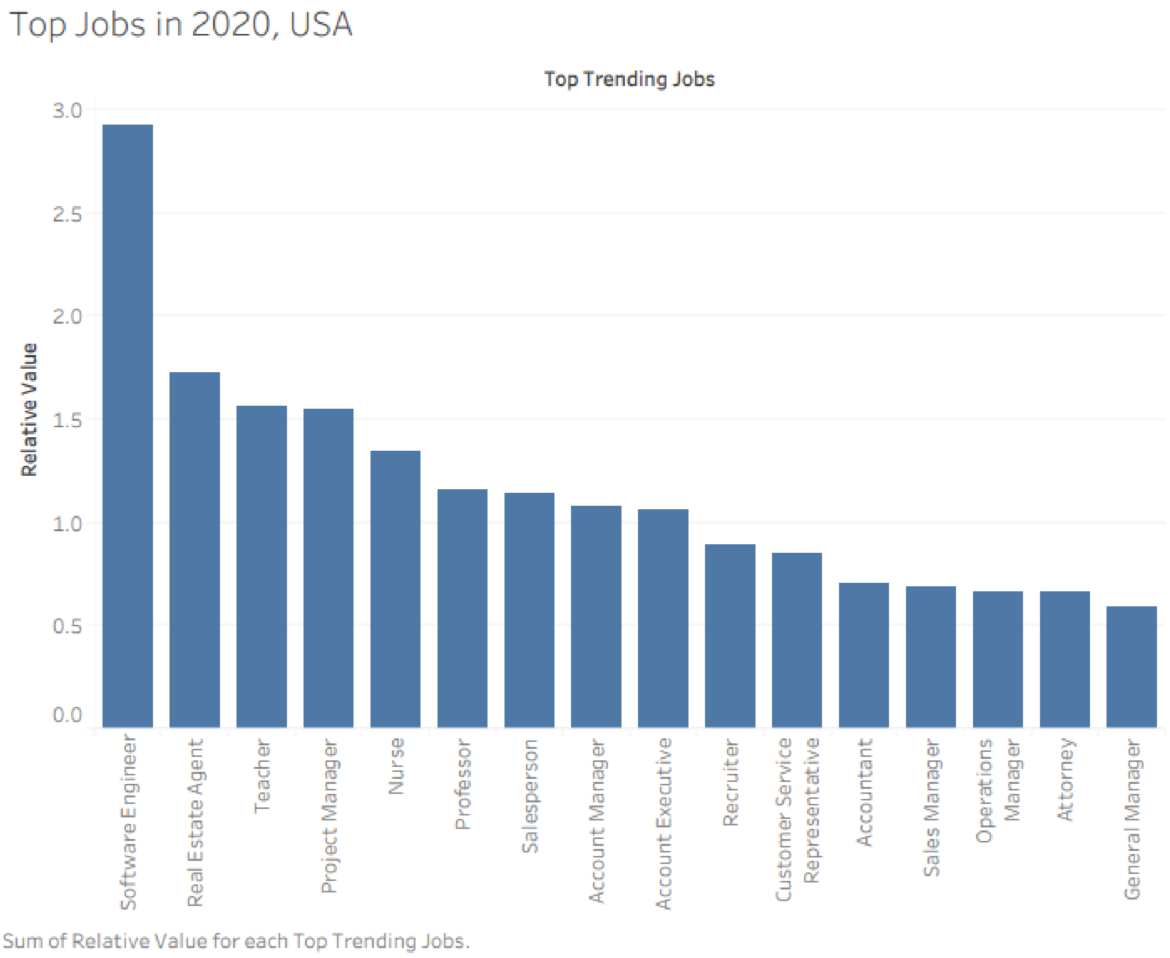


Figure 27: Top Jobs in 2020, USA [16]

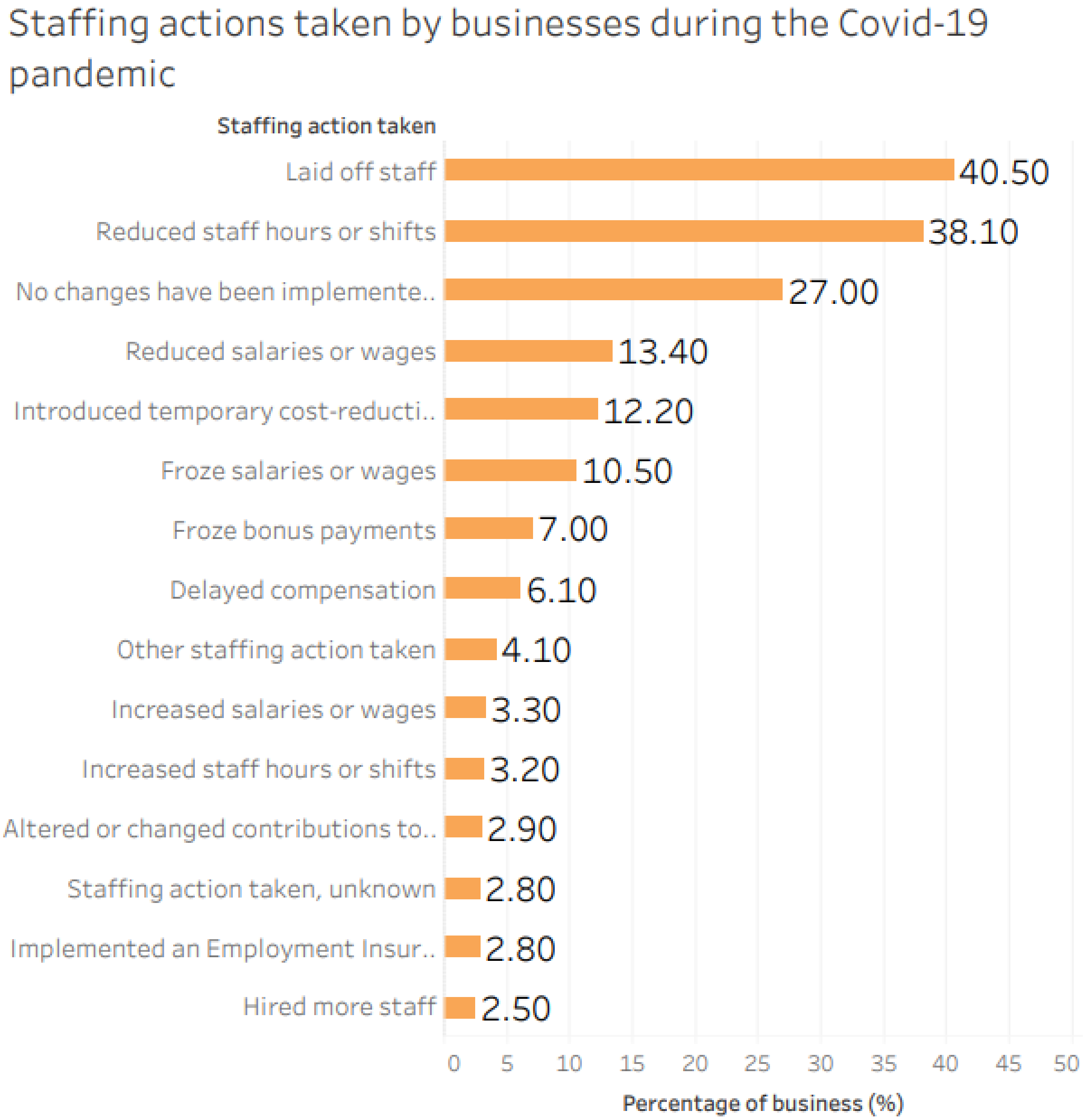


Figure 28: Staffing actions taken by businesses during the Covid-19 pandemic [17]

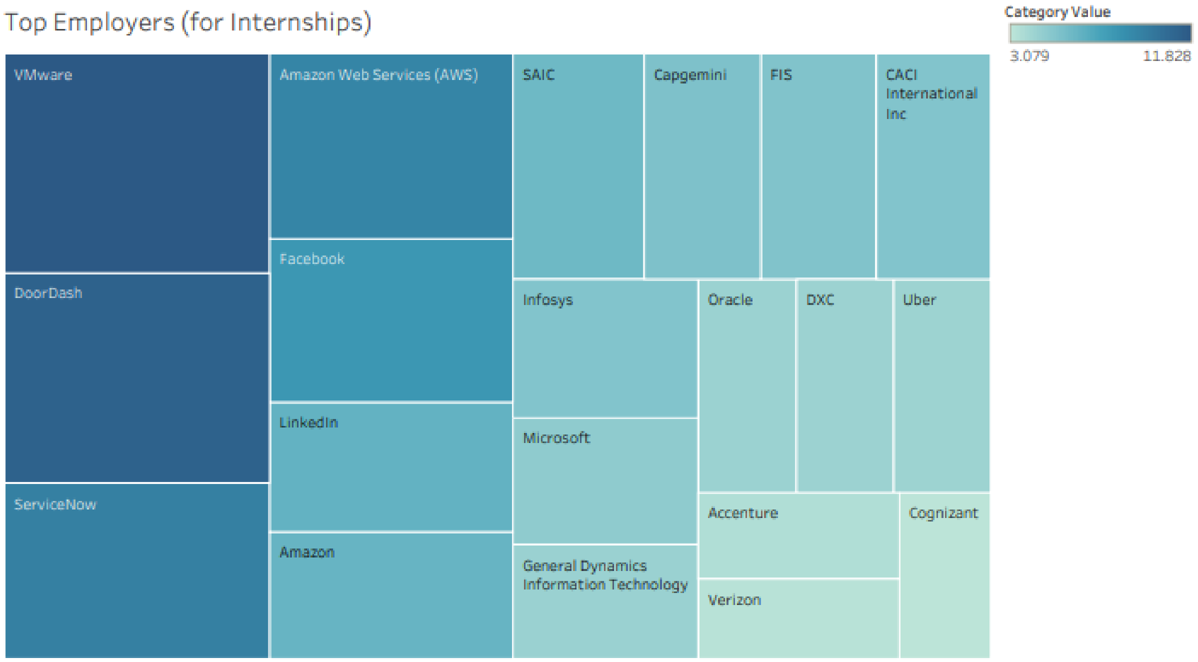


Figure 29: Top Employers (for Internships) [16]

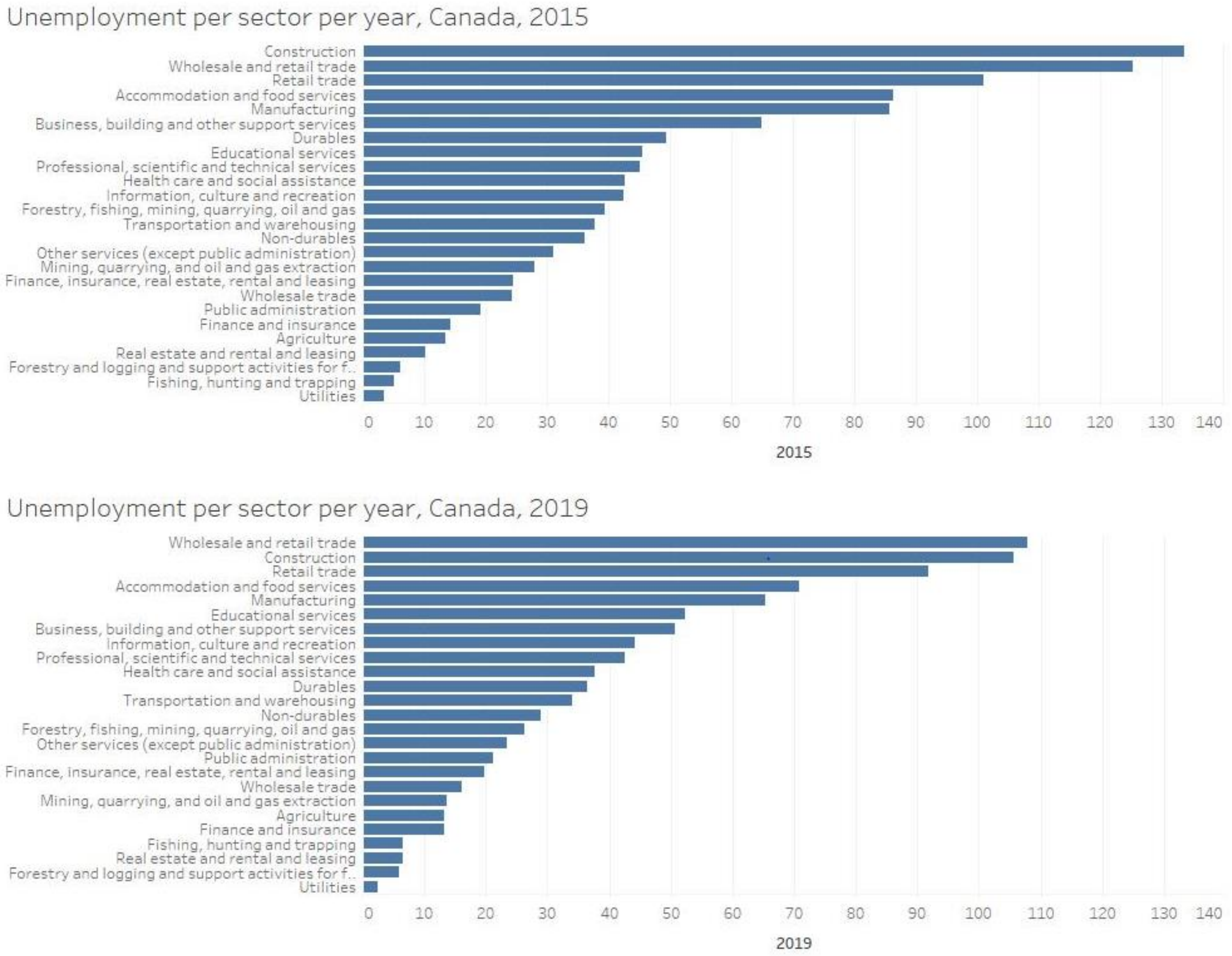
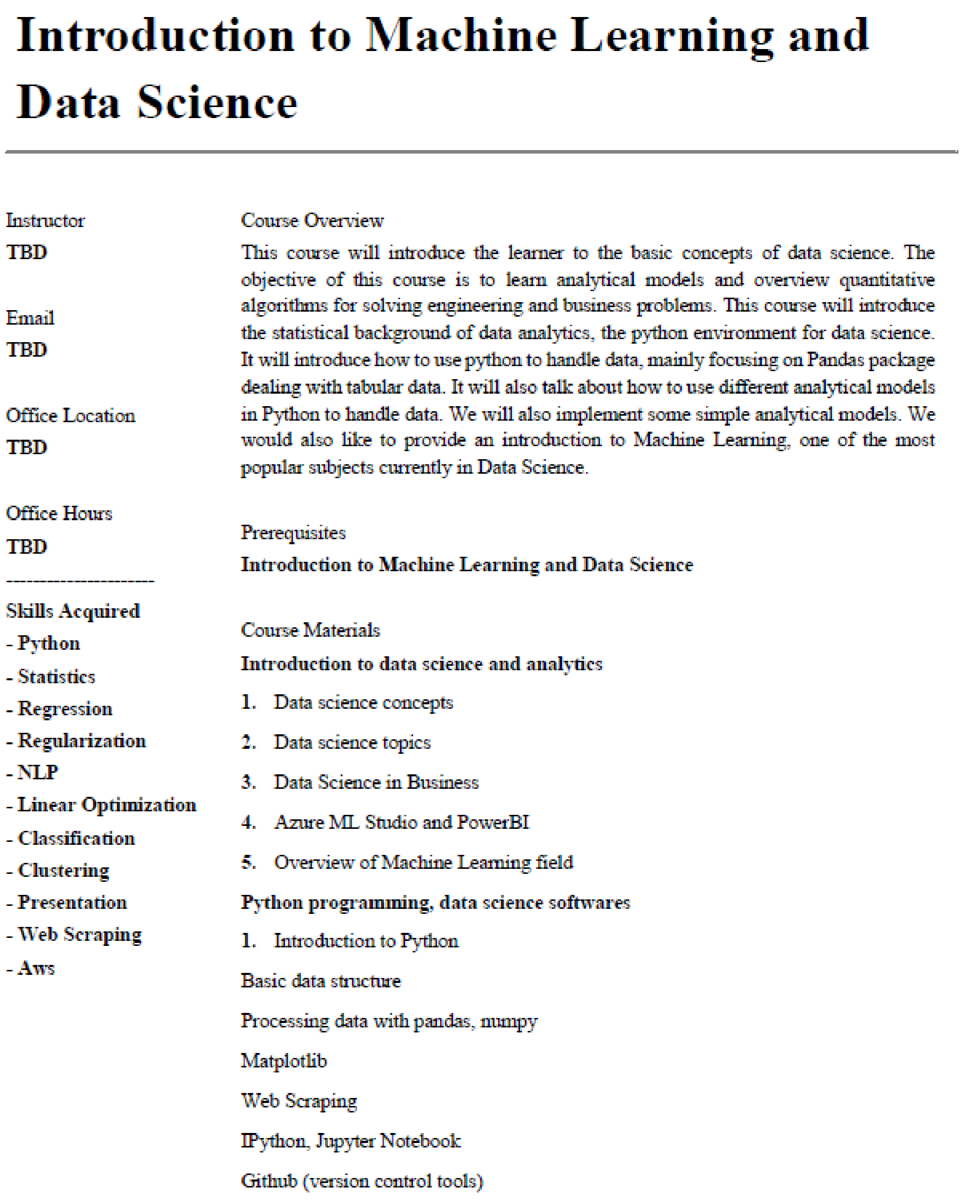


Figure 30: Unemployment per sector per year, Canada, 2015 and 2019 [11]

Appendix E: Example syllabus for the five core courses



Introduction to Machine Learning and Data Science

Instructor

TBD

Email

TBD

Office Location

TBD

Office Hours

TBD

Skills Acquired

- Python

- Statistics

- Regression

- Regularization

- NLP

- Linear Optimization

- Classification

- Clustering

- Presentation

- Web Scraping

- Aws

Course Overview

This course will introduce the learner to the basic concepts of data science. The objective of this course is to learn analytical models and overview quantitative algorithms for solving engineering and business problems. This course will introduce the statistical background of data analytics, the python environment for data science. It will introduce how to use python to handle data, mainly focusing on Pandas package dealing with tabular data. It will also talk about how to use different analytical models in Python to handle data. We will also implement some simple analytical models. We would also like to provide an introduction to Machine Learning, one of the most popular subjects currently in Data Science.

Prerequisites

Introduction to Machine Learning and Data Science

Course Materials

Introduction to data science and analytics

1. Data science concepts
2. Data science topics
3. Data Science in Business
4. Azure ML Studio and PowerBI
5. Overview of Machine Learning field

Python programming, data science softwares

1. Introduction to Python

Basic data structure

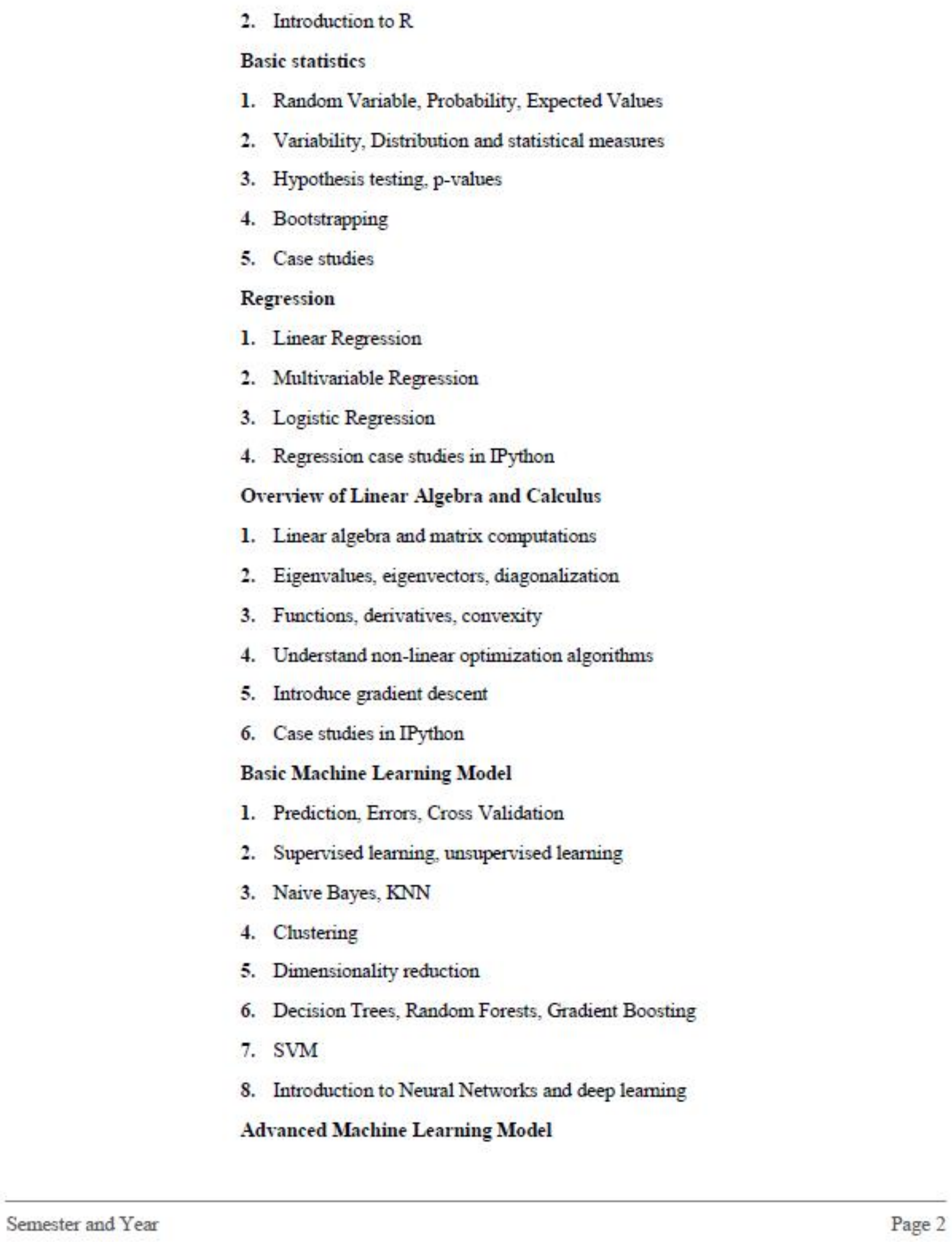
Processing data with pandas, numpy

Matplotlib

Web Scraping

IPython, Jupyter Notebook

Github (version control tools)



2. Introduction to R

Basic statistics

1. Random Variable, Probability, Expected Values
2. Variability, Distribution and statistical measures
3. Hypothesis testing, p-values
4. Bootstrapping
5. Case studies

Regression

1. Linear Regression
2. Multivariable Regression
3. Logistic Regression
4. Regression case studies in IPython

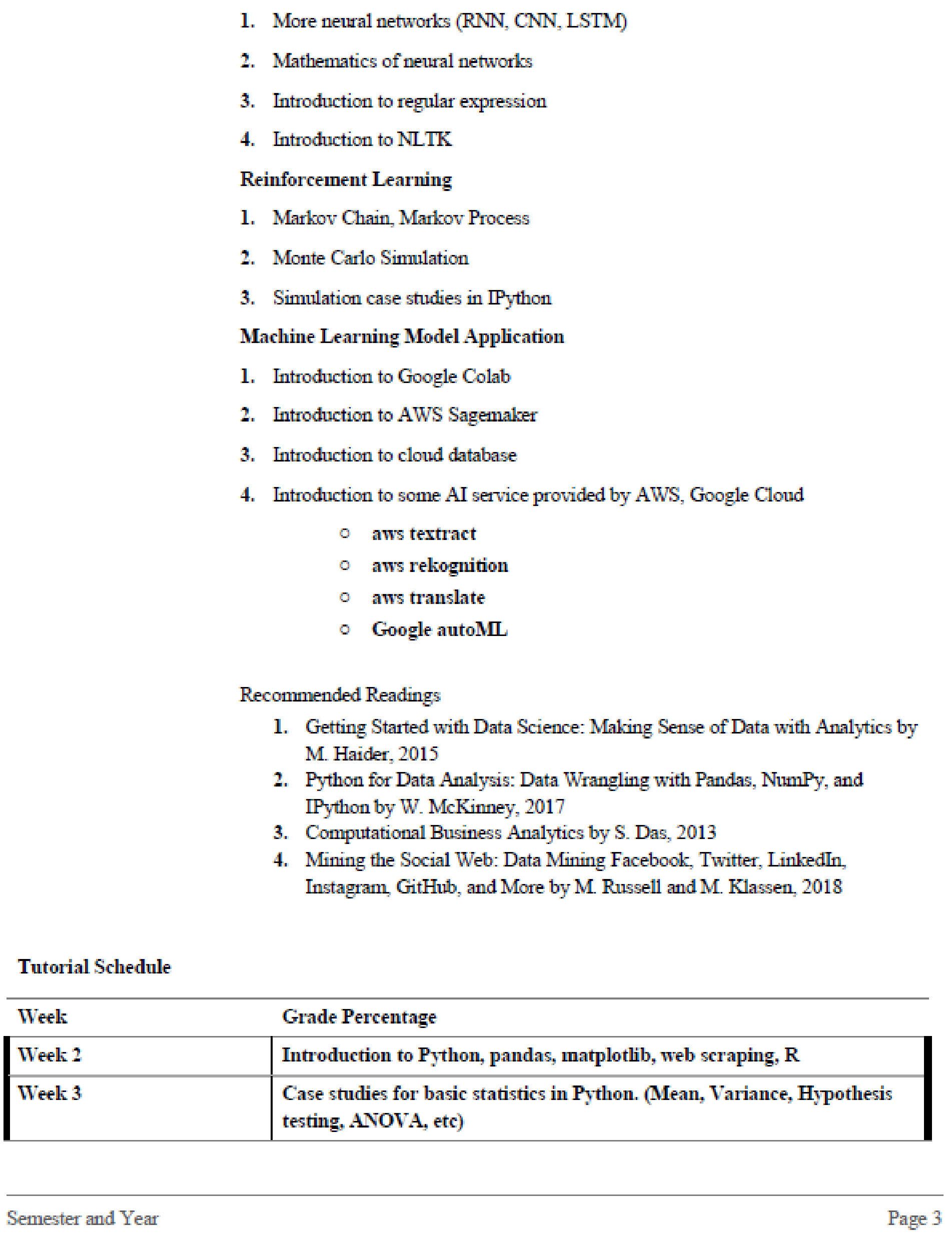
Overview of Linear Algebra and Calculus

1. Linear algebra and matrix computations
2. Eigenvalues, eigenvectors, diagonalization
3. Functions, derivatives, convexity
4. Understand non-linear optimization algorithms
5. Introduce gradient descent
6. Case studies in IPython

Basic Machine Learning Model

1. Prediction, Errors, Cross Validation
2. Supervised learning, unsupervised learning
3. Naive Bayes, KNN
4. Clustering
5. Dimensionality reduction
6. Decision Trees, Random Forests, Gradient Boosting
7. SVM
8. Introduction to Neural Networks and deep learning

Advanced Machine Learning Model



1. More neural networks (RNN, CNN, LSTM)
2. Mathematics of neural networks
3. Introduction to regular expression
4. Introduction to NLTK

Reinforcement Learning

1. Markov Chain, Markov Process
2. Monte Carlo Simulation
3. Simulation case studies in IPython

Machine Learning Model Application

1. Introduction to Google Colab
2. Introduction to AWS Sagemaker
3. Introduction to cloud database
4. Introduction to some AI service provided by AWS, Google Cloud
    - **aws textract**
    - **aws rekognition**
    - **aws translate**
    - **Google autoML**

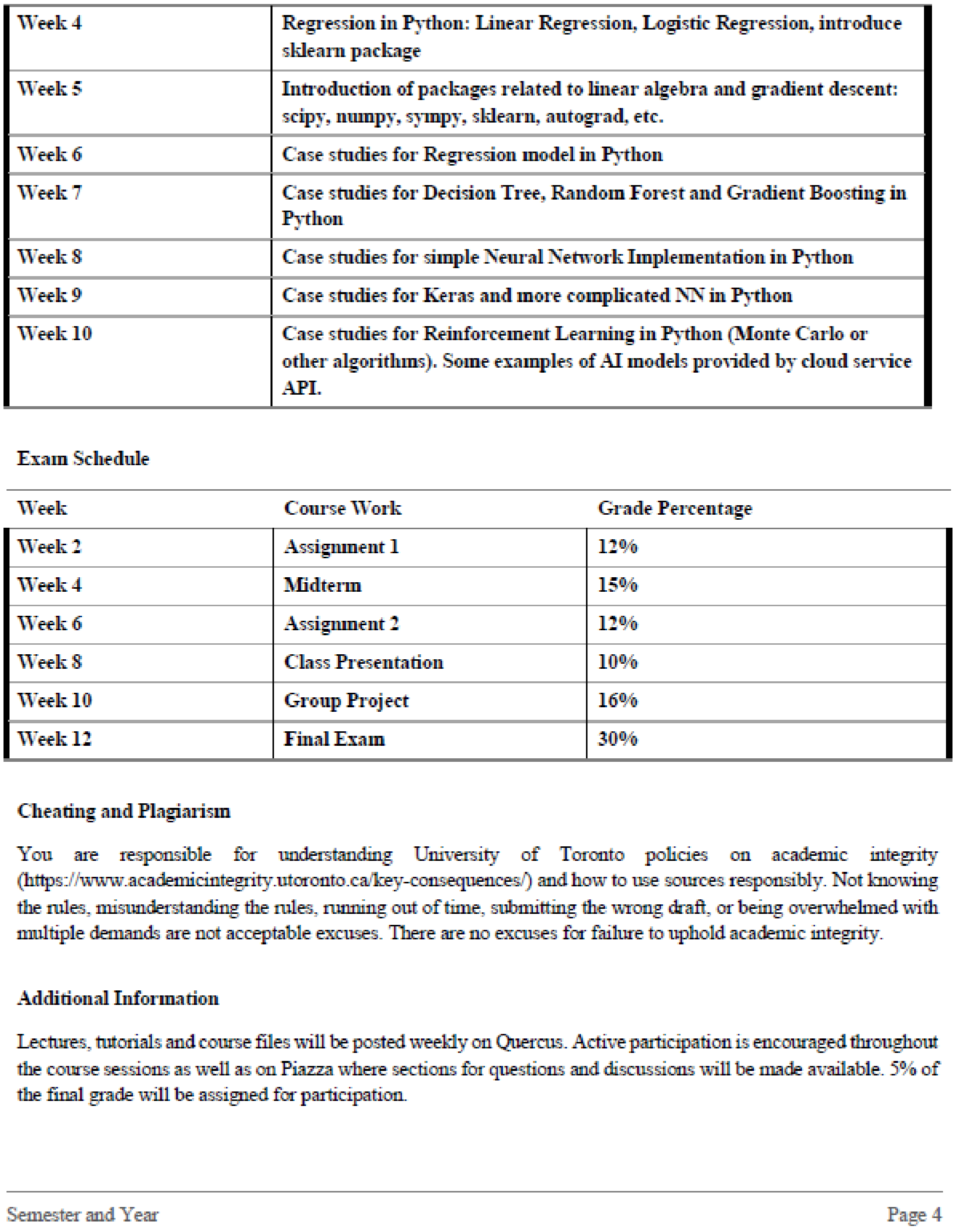
Recommended Readings

1. Getting Started with Data Science: Making Sense of Data with Analytics by M. Haider, 2015
2. Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython by W. McKinney, 2017
3. Computational Business Analytics by S. Das, 2013
4. Mining the Social Web: Data Mining Facebook, Twitter, LinkedIn, Instagram, GitHub, and More by M. Russell and M. Klassen, 2018

Tutorial Schedule

Week	**Grade Percentage**
Week 2	**Introduction to Python, pandas, matplotlib, web scraping, R**
Week 3	**Case studies for basic statistics in Python. (Mean, Variance, Hypothesis testing, ANOVA, etc)**

Semester and Year Page 3



Week 4	Regression in Python: Linear Regression, Logistic Regression, introduce sklearn package
Week 5	Introduction of packages related to linear algebra and gradient descent: scipy, numpy, sympy, sklearn, autograd, etc.
Week 6	Case studies for Regression model in Python
Week 7	Case studies for Decision Tree, Random Forest and Gradient Boosting in Python
Week 8	Case studies for simple Neural Network Implementation in Python
Week 9	Case studies for Keras and more complicated NN in Python
Week 10	Case studies for Reinforcement Learning in Python (Monte Carlo or other algorithms). Some examples of AI models provided by cloud service API.

Exam Schedule

Week	Course Work	Grade Percentage
Week 2	Assignment 1	12%
Week 4	Midterm	15%
Week 6	Assignment 2	12%
Week 8	Class Presentation	10%
Week 10	Group Project	16%
Week 12	Final Exam	30%

Cheating and Plagiarism

You are responsible for understanding University of Toronto policies on academic integrity (https://www.academicintegrity.utoronto.ca/key-consequences/) and how to use sources responsibly. Not knowing the rules, misunderstanding the rules, running out of time, submitting the wrong draft, or being overwhelmed with multiple demands are not acceptable excuses. There are no excuses for failure to uphold academic integrity.

Additional Information

Lectures, tutorials and course files will be posted weekly on Quercus. Active participation is encouraged throughout the course sessions as well as on Piazza where sections for questions and discussions will be made available. 5% of the final grade will be assigned for participation.

Advanced Machine Learning

Instructor

Email

Office Location

Office Hours

Skills Acquired

Course Overview[1]

With the rapid rise of the interdisciplinary data science and big data fields, there has been a push for increased extraction of knowledge and insight from all types, forms, and shapes of data. Both of these new fields are built on algorithms that construct models of data and facilitate decision-making. Machine learning draws upon approaches found in the computer science, math and statistics fields making it less accessible for those without any technical training. Consequently, many new practitioners use these algorithms as black boxes without understanding their nuances or limitations. This course focuses on understanding how the primary machine learning algorithms work so students will be able to select appropriate methods, adapt the methods to solve specific problems, and work to overcome the limitations of the standard algorithms. This can provide a competitive advantage professionally for practitioners. The course will cover a range of current research fields within the topic and provide experience working on different types of data.

Prerequisites

Introduction to Machine Learning and Data Science

Course Materials

Introduction

1. Overview of Machine Learning field ;
2. Regression;

Clustering

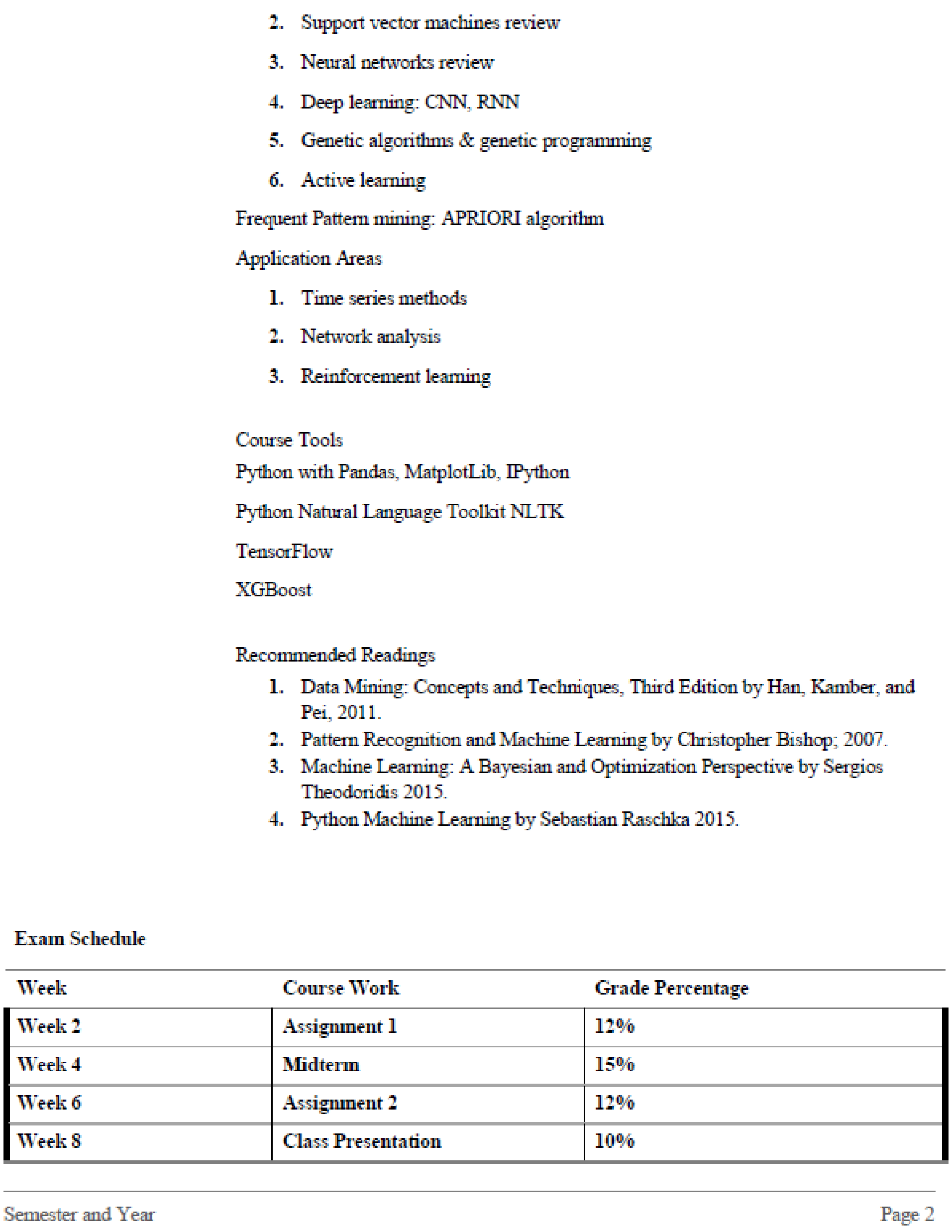
1. Clustering: Distance Metrics, leader, Jarvis-Patrick, scaling hierarchical clustering
2. Clustering: Self-organized maps, EM-algorithm, and more advanced methods

Dimensionality Reduction: PCA mathematical review, Sammon's, t-SNE

Supervised Methods

1. Classification: boosting, bagging, ensemble methods, random forests

[1] Based on syllabus for the course titled "Advanced Machine Learning, Data Mining, and Artificial Intelligence" by Peter V. Henstock, Ph.D., Harvard University;



2. Support vector machines review
3. Neural networks review
4. Deep learning: CNN, RNN
5. Genetic algorithms & genetic programming
6. Active learning

Frequent Pattern mining: APRIORI algorithm

Application Areas

1. Time series methods
2. Network analysis
3. Reinforcement learning

Course Tools

Python with Pandas, MatplotLib, IPython

Python Natural Language Toolkit NLTK

TensorFlow

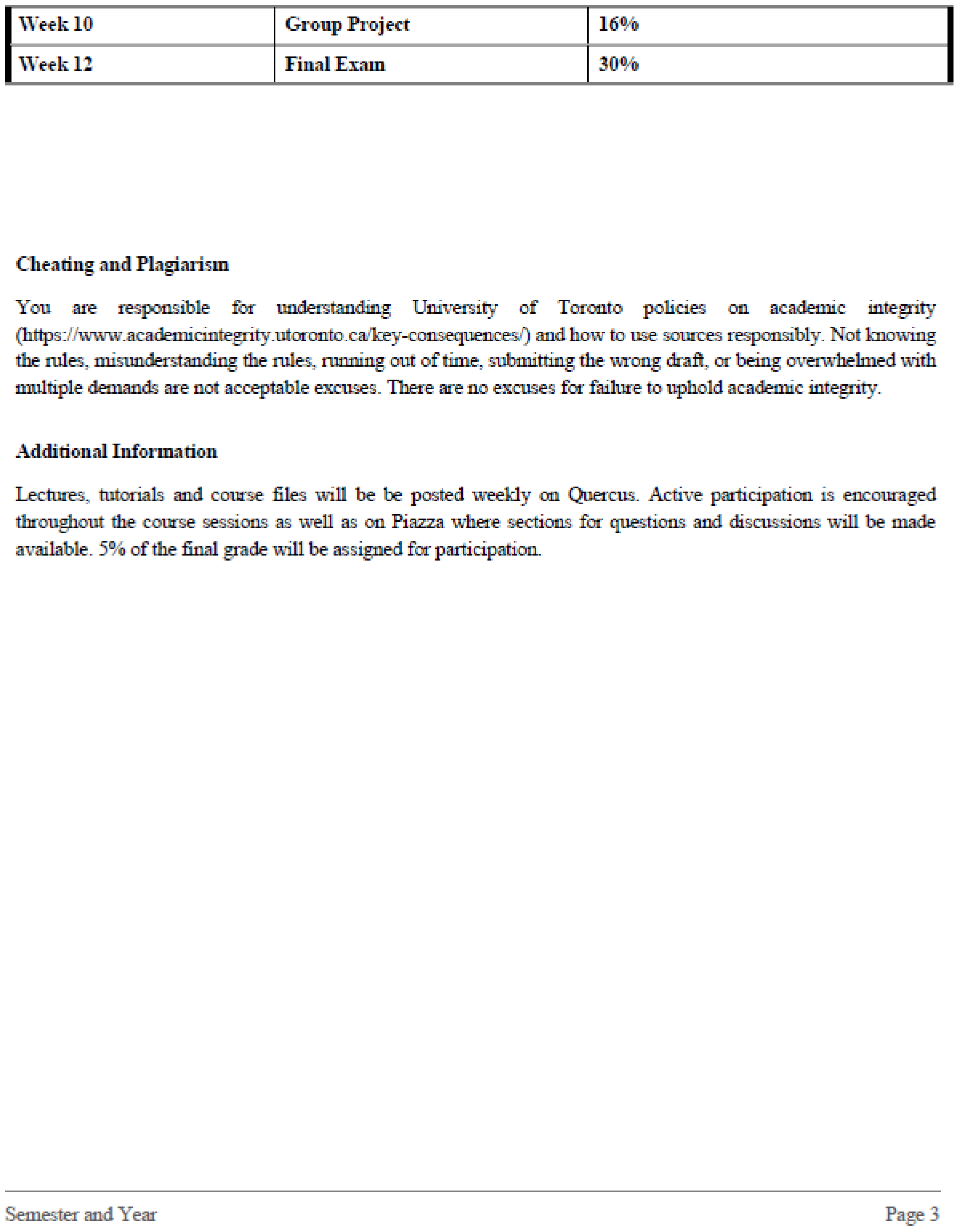
XGBoost

Recommended Readings

1. Data Mining: Concepts and Techniques, Third Edition by Han, Kamber, and Pei, 2011.
2. Pattern Recognition and Machine Learning by Christopher Bishop; 2007.
3. Machine Learning: A Bayesian and Optimization Perspective by Sergios Theodoridis 2015.
4. Python Machine Learning by Sebastian Raschka 2015.

Exam Schedule

Week	**Course Work**	**Grade Percentage**
Week 2	**Assignment 1**	**12%**
Week 4	**Midterm**	**15%**
Week 6	**Assignment 2**	**12%**
Week 8	**Class Presentation**	**10%**



Week 10	**Group Project**	**16%**
Week 12	**Final Exam**	**30%**

Cheating and Plagiarism

You are responsible for understanding University of Toronto policies on academic integrity (https://www.academicintegrity.utoronto.ca/key-consequences/) and how to use sources responsibly. Not knowing the rules, misunderstanding the rules, running out of time, submitting the wrong draft, or being overwhelmed with multiple demands are not acceptable excuses. There are no excuses for failure to uphold academic integrity.

Additional Information

Lectures, tutorials and course files will be be posted weekly on Quercus. Active participation is encouraged throughout the course sessions as well as on Piazza where sections for questions and discussions will be made available. 5% of the final grade will be assigned for participation.

Deep Learning

Instructor

TBD

Email

TBD

Office Location

TBD

Office Hours

TBD

Skills Acquired

- Deep Learning

- NLP

- Tensorflow

Course Overview

This course addresses the fundamental concepts and advanced methodologies of deep learning and relates them to real-world problems in a variety of domains. The aim is to provide an overview of different approaches, both classical and emerging. The module will equip you with the necessary knowledge and skills to work in the field of deep learning and to contribute to ongoing research in the area.

Prerequisites

Linear Algebra

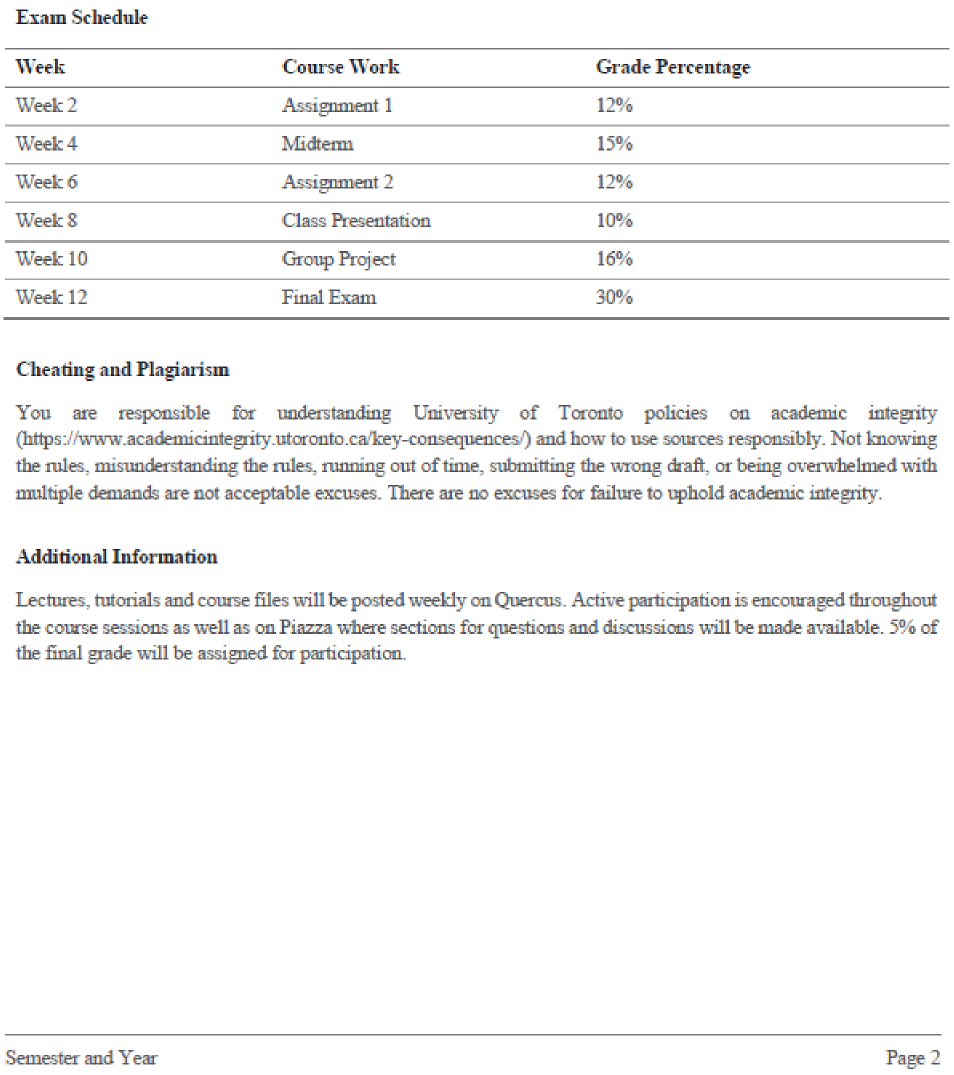
Introduction to Machine Learning

Course Materials

1. Supervised vs unsupervised learning, generalization, overfitting;
2. Perceptrons, including deep vs shallow models;
3. Stochastic gradient descent and backpropagation;
4. Convolutional neural networks (CNN) and underlying mathematical principles;
5. CNN architectures and applications in image analysis;
6. Recurrent neural networks (RNN), long-short term memory (LSTM), gated recurrent units (GRU);
7. Applications on RNNs in speech analysis and machine translation;
8. Mathematical principles of generative networks; variational autoencoders (VAE); generative adversarial networks (GAN);
9. Applications of generative networks in image generation;
10. Graph neural networks (GNN): spectral and spatial domain methods, message passing;

Recommended Readings

1. Deep Learning, a textbook by Yoshua Bengio, Ian Goodfellow, and Aaron Courville.



Exam Schedule

Week	Course Work	Grade Percentage
Week 2	Assignment 1	12%
Week 4	Midterm	15%
Week 6	Assignment 2	12%
Week 8	Class Presentation	10%
Week 10	Group Project	16%
Week 12	Final Exam	30%

Cheating and Plagiarism

You are responsible for understanding University of Toronto policies on academic integrity (https://www.academicintegrity.utoronto.ca/key-consequences/) and how to use sources responsibly. Not knowing the rules, misunderstanding the rules, running out of time, submitting the wrong draft, or being overwhelmed with multiple demands are not acceptable excuses. There are no excuses for failure to uphold academic integrity.

Additional Information

Lectures, tutorials and course files will be posted weekly on Quercus. Active participation is encouraged throughout the course sessions as well as on Piazza where sections for questions and discussions will be made available. 5% of the final grade will be assigned for participation.

Advanced Statistics

Instructor

TBD

Email

TBD

Office Location

TBD

Office Hours

TBD

Skills Acquired

- Statistics

- Hypothesis Testing

- Exploratory Data Analysis

- Regression

- Bayesian Inference

Course Overview[1]

This course covers the fundamentals of theoretical statistics. Topics include: concentration of measure, basic empirical process theory, convergence, point and interval estimation, maximum likelihood, hypothesis testing, Bayesian inference, nonparametric statistics and bootstrap resampling.

Some course objectives for students in machine learning include: (1) Predict which kinds of existing machine learning algorithms will be most suitable for which sorts of tasks, based on formal properties and experimental results. (2) Evaluate and analyze existing learning algorithms.

Course Materials

Concentration Inequalities

Convergence

Central Limit Theorem

Uniform Laws and Empirical Process Theory

Likelihood and Sufficiency

Point Estimation (MLE)

Point Estimation (Method of Moments, Bayes)

Asymptotic Theory for MLE

Hypothesis Testing

Goodness-of-fit, two-sample, independence

Multiple testing

Bootstrap

Bayesian Inference

Regression (Linear and non-parametric)

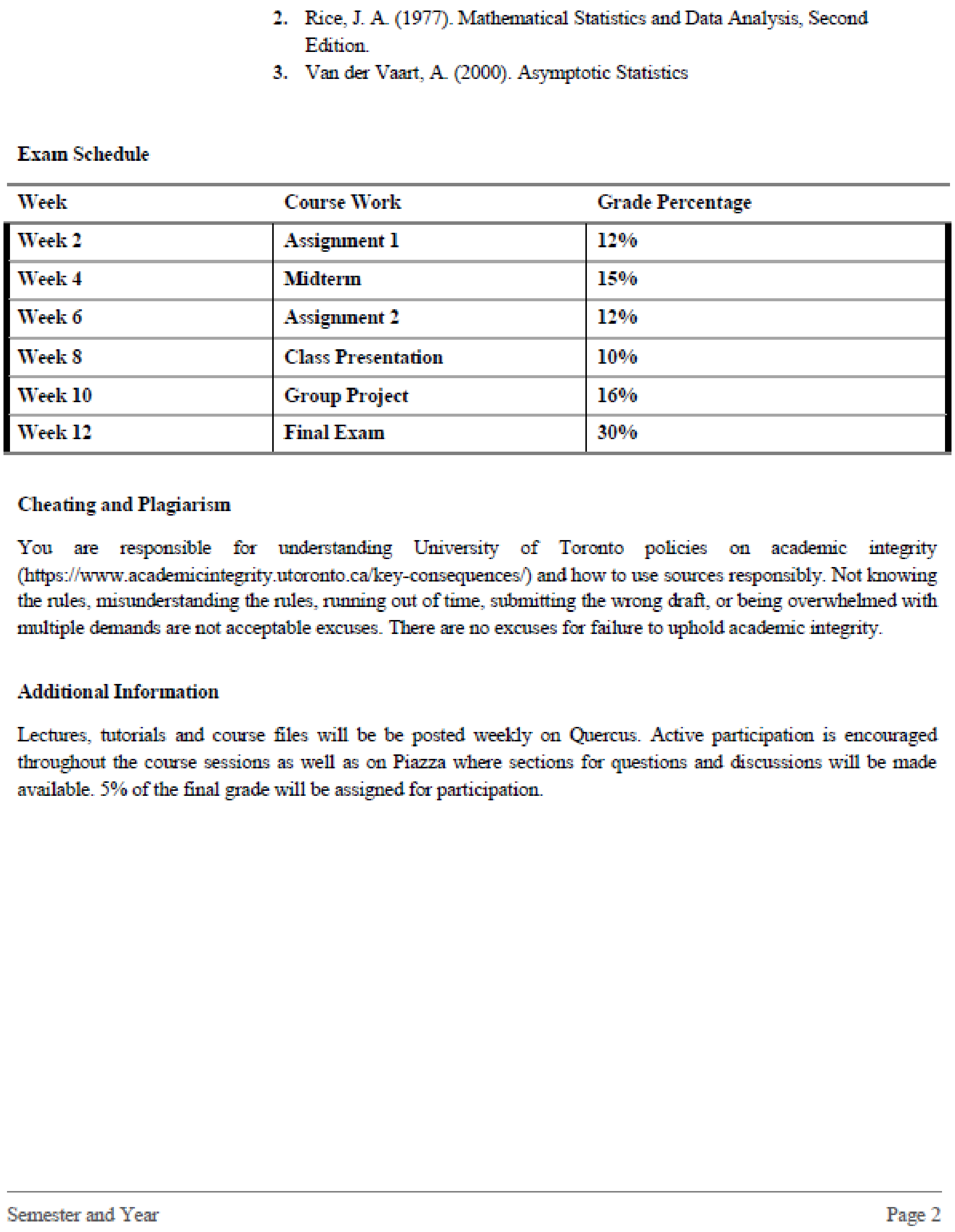
Model Selection

Causal Inference

Recommended Readings

1. Casella, G. and Berger, R. L. (2002). Statistical Inference, 2nd ed.

[1] Based on the syllabus for the course titled "Intermediate Statistics" offered by Carnegie Melon University



2. Rice, J. A. (1977). Mathematical Statistics and Data Analysis, Second Edition.
3. Van der Vaart, A. (2000). Asymptotic Statistics

Exam Schedule

Week	Course Work	Grade Percentage
Week 2	**Assignment 1**	**12%**
Week 4	**Midterm**	**15%**
Week 6	**Assignment 2**	**12%**
Week 8	**Class Presentation**	**10%**
Week 10	**Group Project**	**16%**
Week 12	**Final Exam**	**30%**

Cheating and Plagiarism

You are responsible for understanding University of Toronto policies on academic integrity (https://www.academicintegrity.utoronto.ca/key-consequences/) and how to use sources responsibly. Not knowing the rules, misunderstanding the rules, running out of time, submitting the wrong draft, or being overwhelmed with multiple demands are not acceptable excuses. There are no excuses for failure to uphold academic integrity.

Additional Information

Lectures, tutorials and course files will be be posted weekly on Quercus. Active participation is encouraged throughout the course sessions as well as on Piazza where sections for questions and discussions will be made available. 5% of the final grade will be assigned for participation.

Semester and Year

Page 2

Big Data Analytics

Instructor
TBD

Email
TBD

Office Location
TBD

Office Hours
TBD

Skills Acquired

- Python

- Big Data

- EDA

- Classification

- Prediction

- Hadoop

- Spark

- Aws

- Deep Learning

Course Overview[1]

The emphasis of this course is on mastering two most important big data technologies: Spark 2 and Deep Learning with TensorFlow. Spark is an evolution of Hadoop and Map/Reduce but with massive speedup and scalability improvements. TensorFlow is Google's open-source framework for distributed neural networks-based machine learning. The explosion of social media and the computerization of every aspect of social and economic activity results in the creation of large volumes of semi-structured data: web logs, videos, speech recordings, photographs, e-mails, Tweets, and similar data. In a parallel development, computers keep getting ever more powerful and storage ever cheaper. Today, we can reliably and cheaply store huge volumes of data, efficiently analyze them, and extract business and socially relevant information. This course familiarizes the students with the most important information technologies used in manipulating, storing, and analyzing big data.

Prerequisites

Introduction to Machine Learning and Data Science

Course Materials

Basic Statistics and R

Relationships and Representations, Graph Databases (Neo4J graph database)

Introduction to Spark 2.0

Spark 2.2 Data Frame API

Hadoop

Analysis of Streaming Data with Spark 1.6 Streaming API and Spark Structured Streaming API on Spark 2.2.

Applications of Spark ML Library

Text processing with Python NLTK or Word2Vec

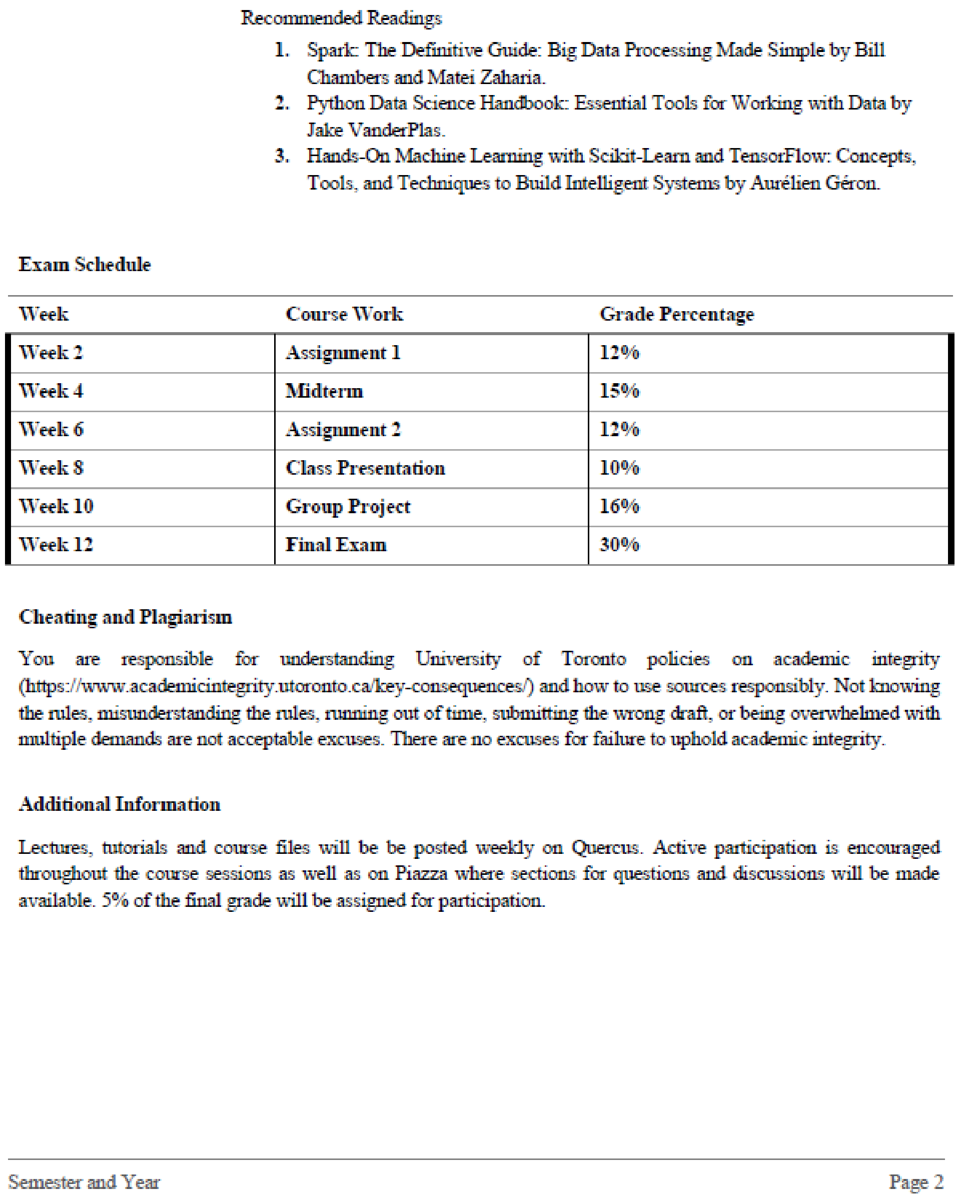
Basic Neural Network and Tensor Flow

Analysis of Images and OCR Applications

Analysis of Speech Signal

Analysis of Streaming Data and Time Series with Tensor Flow

[1] Based on syllabus for course titled "Big Data Analytics" offered by Harvard University



Recommended Readings

1. Spark: The Definitive Guide: Big Data Processing Made Simple by Bill Chambers and Matei Zaharia.
2. Python Data Science Handbook: Essential Tools for Working with Data by Jake VanderPlas.
3. Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems by Aurélien Géron.

Exam Schedule

Week	Course Work	Grade Percentage
Week 2	**Assignment 1**	**12%**
Week 4	**Midterm**	**15%**
Week 6	**Assignment 2**	**12%**
Week 8	**Class Presentation**	**10%**
Week 10	**Group Project**	**16%**
Week 12	**Final Exam**	**30%**

Cheating and Plagiarism

You are responsible for understanding University of Toronto policies on academic integrity (https://www.academicintegrity.utoronto.ca/key-consequences/) and how to use sources responsibly. Not knowing the rules, misunderstanding the rules, running out of time, submitting the wrong draft, or being overwhelmed with multiple demands are not acceptable excuses. There are no excuses for failure to uphold academic integrity.

Additional Information

Lectures, tutorials and course files will be be posted weekly on Quercus. Active participation is encouraged throughout the course sessions as well as on Piazza where sections for questions and discussions will be made available. 5% of the final grade will be assigned for participation.