MIE 1624 – Introduction to Data Science and Analytics

Project Report – Winter 2020 – Group 11

Smart Ed – AI Education Startup Project

Submitted To
Prof. Oleksandr Romanko

Group 11 Members

Akshat Mathur -
Gautam Dawar -
Rami Al-Sahar -
Simon Faux -

Executive Summary

The demand for data science skills is surging due to a rise in the use of AI and Machine Learning across industries. To meet this demand, the University of Toronto is looking to customize its courses to better suit the skills being demanded by industry.

By leveraging the power of data analytics to determine exactly what employers are looking for, the MIE1624 curriculum was redesigned, a new master's program was developed, and a utility to help engineering students select courses was produced.

The redesigned MIE1624 incorporates big data science and tools, as well as practical experience with Tableau. This improves its generalizability while maintaining a reasonable degree of depth for each subject that it engages with. The redesigned course better suits more jobs than any other single course surveyed from world leading analytics programs. With this course, engineering students will be prepared to acquire and excel in a data science position in any industry.

Using industry data, clusters of in demand skills were identified and directly matched to course curriculums to build a new program: the Masters of Business Analytics and Engineering. This program will produce industry leaders as students will graduate with practical experience in all sought after technical skills, with strong experience building the in-demand soft skills.

For students seeking an emphasis in analytics as part of the Masters of Engineering, a new utility has been developed that will help them select courses to enter the career of their choice. By inputting a job posting from Indeed, skills are extracted and matched to courses. The courses provided by the utility will meet the top skills of the position, and the algorithm has been equipped with a feature to ensure diversification of the skills.

Table of Contents

Table of Figures

Table of Tables

1. Introduction

The demand for data science skills is surging due to a rise in the use of AI and Machine Learning across industries. This creates the need for continuous improvement to the data science education, which has become a topic of interest for instructors and businesspeople alike. The University of Toronto is looking to customize its courses to better suit the ever-changing skillset needed to succeed in a data science career. This report provides the justification for and design of a new Masters program to meet the University of Toronto's desire for an invigorated program that will be a world-leader in preparing students for industry. Additionally, a redesigned *MIE1624: Introduction to Data Science and Analytics* course is presented that will better meet industry needs than the current offering. Finally, to support engineering students outside of the new Masters stream, a tool is presented to help select courses that will prepare them to enter their desired career.

This report first briefly discusses the methodology used to develop the redesigned MIE1624 and new Masters Curriculum. Detail is then provided on the redesigned course and curriculum, followed by a presentation of the new course selection utility to assist students pursuing an emphasis in analytics within the Master of Engineering program.

2. Curriculum and Course Design Methodology

The primary objective of the redesigned courses is to best prepare students for industry. There is no better method for determining the skills in demand by industry than surveying the current job market. Over 300 data scientist jobs, posted on Indeed and LinkedIn, were scraped for their content and analyzed. To support course curriculum development, and to understand how the new courses perform relative to existing offerings, course details from existing world-leading education programs, including from University of California at Berkeley (UCB), University of Toronto's Rotman Business School (UOFTR), and Coursera, were scraped.

The following steps were performed:

1. Details from over 300 data scientist jobs were analyzed to form a condensed list of critical, reoccurring, skills. Skills were grouped into clusters.
2. Course descriptions from UCB, UOFTR and Coursera were collected and analyzed. Top courses for each job skill cluster were identified.
3. Existing courses were evaluated against the set of jobs, scoring higher with the more skills they matched.
4. The highest performing individual courses were reviewed and used to rebuild the MIE1624 curriculum.
5. A set of courses was made for the new Masters program by satisfying each of the skill clusters identified from the job data.
6. Independent of the previous steps, a tool was created that receives an Indeed job posting, and recommends courses for Master of Engineering students pursuing an emphasis in analytics.

3. Redesigned MIE1624

MIE1624 is a class taken outside of a Master of Data Science stream with the objective of providing students with the skills necessary to enter a data science position in their field of choice. With any redesign, maintaining generalizability is important. Subjects need not be taught in excessive detail, but with sufficient depth and practical use to be applied in a job. All the courses that were collected from the world leading universities (UCB, UOFTR, Coursera), including MIE1624, were evaluated against the jobs. The more skills that matched, the higher the course scored. By this method, a more general class will rank higher as it matches a wider variety of jobs well. Note that ties between classes for a given job are allowed. Table 1 below shows the top 5 courses.

Table 1. Top 5 Courses

Number of Jobs for Which Course is Best Suited	University	Course Name	Skills
186	UC Berkeley	Machine Learning at Scale	Python, machine learning, financial planning, big data, hadoop, spark, aws
112	University of Toronto	MIE1624: Introduction to Data Science and Analytics	Python, machine learning, project management, financial planning, artificial intelligence, data visualization, predictive modeling
63	UC Berkeley	Fundamentals of Data Engineering	Python, hadoop, spark, aws, cloud computing, database
51	UC Berkeley	Data Visualization	Data visualization, communication, tableau, java
44	UC Berkeley	Applied Machine Learning	Python, machine learning, probability, collaborative

MIE1624 performs quite well. However, other high performing classes involve a big data science and tools component. Another class, Data Visualization at UC Berkeley, performed highly indicating that many employers are looking for experience with Tableau. By incorporating these two subjects into MIE1624, the generalizability was improved significantly. A detailed comparison and visualization of the original and redesigned curriculums are shown below**.**

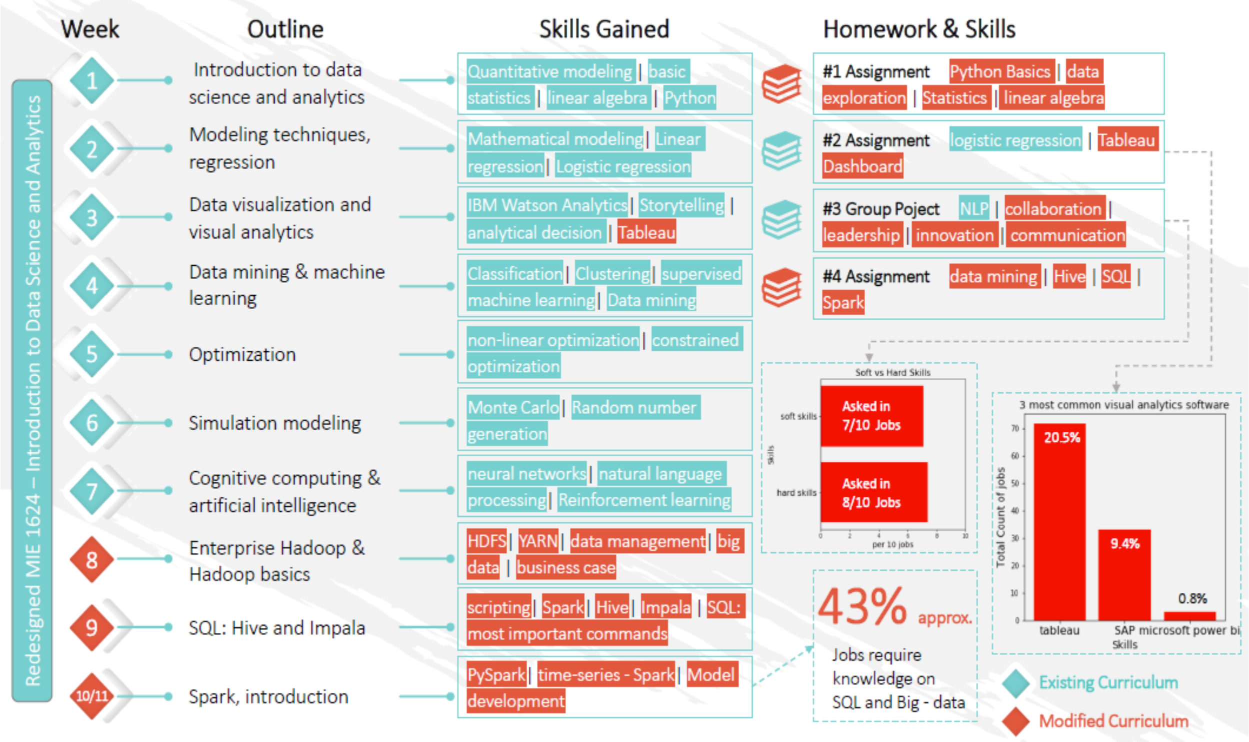

Week
Outline
Skills Gained
Homework & Skills
Redesigned MIE 1624 – Introduction to Data Science and Analytics
1
Introduction to data science and analytics
Quantitative modeling | basic statistics | linear algebra | Python
#1 Assignment Python Basics | data exploration | Statistics | linear algebra
2
Modeling techniques, regression
Mathematical modeling | Linear regression | Logistic regression
#2 Assignment logistic regression | Tableau Dashboard
3
Data visualization and visual analytics
IBM Watson Analytics | Storytelling | analytical decision | Tableau
#3 Group Poject NLP | collaboration | leadership | innovation | communication
4
Data mining & machine learning
Classification | Clustering | supervised machine learning | Data mining
#4 Assignment data mining | Hive | SQL | Spark
5
Optimization
non-linear optimization | constrained optimization
6
Simulation modeling
Monte Carlo | Random number generation
7
Cognitive computing & artificial intelligence
neural networks | natural language processing | Reinforcement learning
8
Enterprise Hadoop & Hadoop basics
HDFS | YARN | data management | big data | business case
9
SQL: Hive and Impala
scripting | Spark | Hive | Impala | SQL: most important commands
10/11
Spark, introduction
PySpark | time-series - Spark | Model development
Soft vs Hard Skills
soft skills
hard skills
Skills
Asked in 7/10 Jobs
Asked in 8/10 Jobs
0
2
4
6
8
10
per 10 jobs
43% approx.
Jobs require knowledge on SQL and Big - data
3 most common visual analytics software
Total Count of jobs
0
10
20
30
40
50
60
70
20.5%
9.4%
0.8%
tableau
SAP
microsoft power bi
Skills
Existing Curriculum
Modified Curriculum

Figure 9. New Re-designed MIE1624 Course Curriculum Visual

The following key changes were made:

- The introductory weeks of MIE1628: Big Data Science, were added to MIE1624 to give students exposure to big data science and tools.
- Space was made available in the curriculum by:
  - Assuming basic statistics and linear algebra to be understood from pre-requisites. Readings and online material will be provided to students needing a refresher.
  - Replacing the introduction to Python lecture with online materials and a short assignment
- Tableau visualizations were added as a requirement for course assignments.
- The course project was dropped in favour of an additional assignment applying big data tools, however, the existing course assignment two was converted to a group project.

Courses were rescored, including the redesigned MIE1624. The results are shown in Table 2**.** The redesigned MIE1624 is now the highest performing course. If students are taking just one data science course, this will prepare them for industry better than any other course available.

Table 2. Top Ranking of courses including redesigned MIE1624

Number of Jobs for Which Course is Best Suited	University	Course Name	Skills
255	University of Toronto	Revised MIE1624	Python, machine learning, project management, financial planning, artificial

			intelligence, data visualization, predictive modeling, tableau, hadoop, spark, big data
115	UC Berkeley	Machine Learning at Scale	Python, machine learning, financial planning, big data, hadoop, spark, aws
52	University of Toronto	Original MIE1624: Introduction to Data Science and Analytics	Python, machine learning, project management, financial planning, artificial intelligence, data visualization, predictive modeling
49	UC Berkeley	Fundamentals of Data Engineering	Python, hadoop, spark, aws, cloud computing, database
38	UC Berkeley	Data Visualization	Data visualization, communication, tableau, java
32	UC Berkeley	Applied Machine Learning	Python, machine learning, probability, collaborative

4. Master of Data Science and Artificial Intelligence Curriculum

The Master of Data Science and Artificial Intelligence degree program is committed to educating the next generation of world-class innovators. In a bid to sustain a culture of empowerment and innovation, it is the program's mission to develop critical human capital for the knowledge economy as well as forge broader University of Toronto academic-industry partnerships.

Job skills were grouped into hierarchical clusters, each cluster is distinct and the objects within each cluster are broadly similar to each other. The full set of courses from the different Universities was evaluated against each cluster to determine the best matching pairs. These courses were then included in the new Masters curriculum to ensure students will be equipped to succeed in a data science position in any industry.

Nine clusters of skills were identified, visualized in Figure 1**,** for which nine unique courses best satisfied each cluster. The most sought after skill clusters, measured by the number of job postings seeking those skills, were included in the curriculum as core courses, while the remaining clusters were provided as optional specializations. The number of jobs seeking skills in each cluster are shown in Figure 2**.** Analyzing the job postings also made clear that employers are demanding practical experience, and strong soft skills like communication, leadership and teamwork. The curriculum includes project experience and an internship to complement the students' technical skills with practical and workplace experience.

In addition to skills matching, the following common characteristics of Masters program were noted and considered when producing the new curriculum:

- Typically, 8-10 courses, with a small number of electives relative to core courses
- Involve some form of practical experience, either in the form of a major project or internship
- Fully completed in three semesters

The new Master of Data Science and Artificial Intelligence curriculum is visualized in Figure 3. Course descriptions, as provided by the authoring institution, are provided in Appendix B.

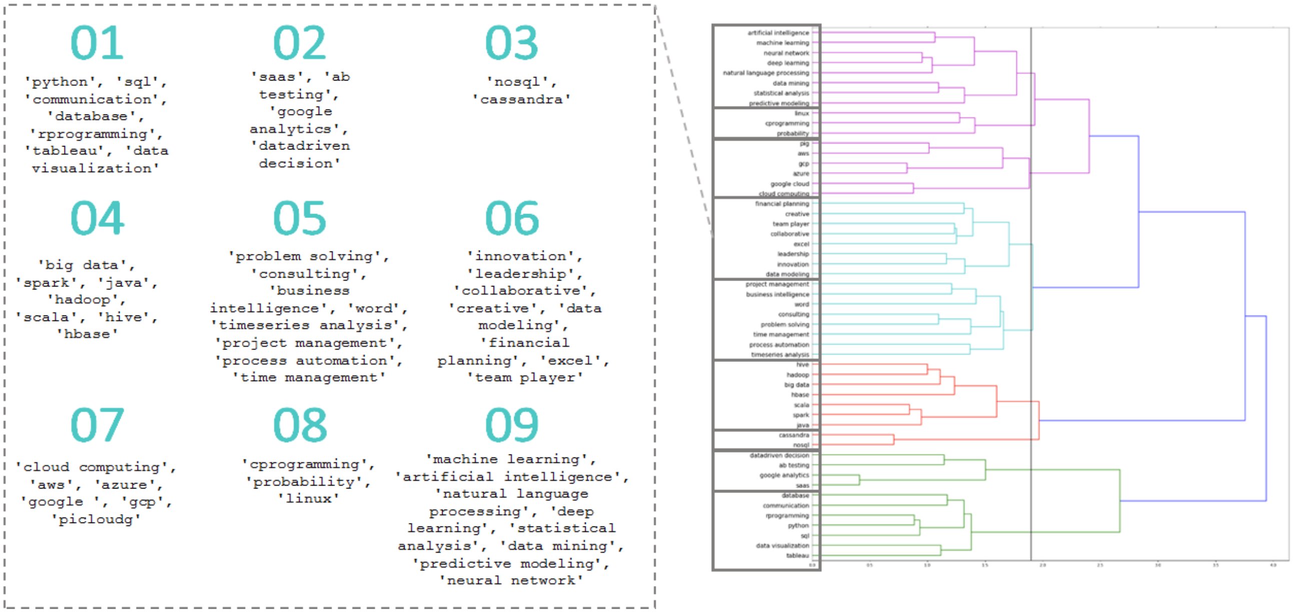

01
'python', 'sql', 'communication', 'database', 'rprogramming', 'tableau', 'data visualization'
02
'saas', 'ab testing', 'google analytics', 'datadriven decision'
03
'nosql', 'cassandra'
04
'big data', 'spark', 'java', 'hadoop', 'scala', 'hive', 'hbase'
05
'problem solving', 'consulting', 'business intelligence', 'word', 'timeseries analysis', 'project management', 'process automation', 'time management'
06
'innovation', 'leadership', 'collaborative', 'creative', 'data modeling', 'financial planning', 'excel', 'team player'
07
'cloud computing', 'aws', 'azure', 'google ', 'gcp', 'picloudg'
08
'cprogramming', 'probability', 'linux'
09
'machine learning', 'artificial intelligence', 'natural language processing', 'deep learning', 'statistical analysis', 'data mining', 'predictive modeling', 'neural network'
artificial intelligence
machine learning
neural network
deep learning
natural language processing
data mining
statistical analysis
predictive modeling
linux
cprogramming
probability
pig
aws
gcp
azure
google cloud
cloud computing
financial planning
creative
team player
collaborative
excel
leadership
innovation
data modeling
project management
business intelligence
word
consulting
problem solving
time management
process automation
timeseries analysis
hive
hadoop
big data
hbase
scala
spark
java
cassandra
nosql
datadriven decision
ab testing
google analytics
saas
database
communication
rprogramming
python
sql
data visualization
tableau

Figure 1. Skill Cluster Identification

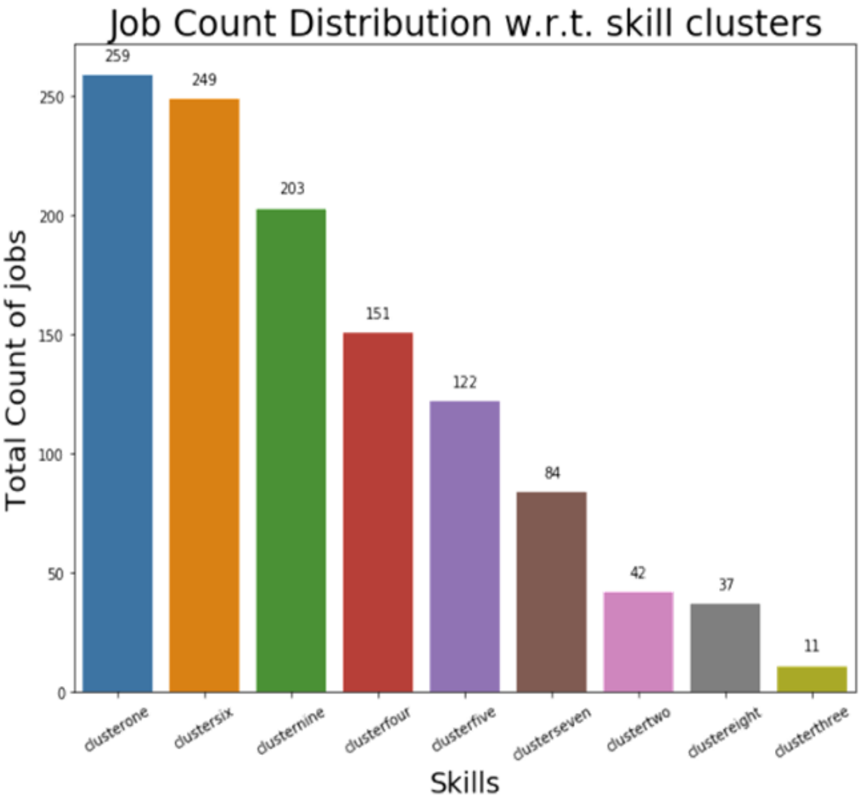
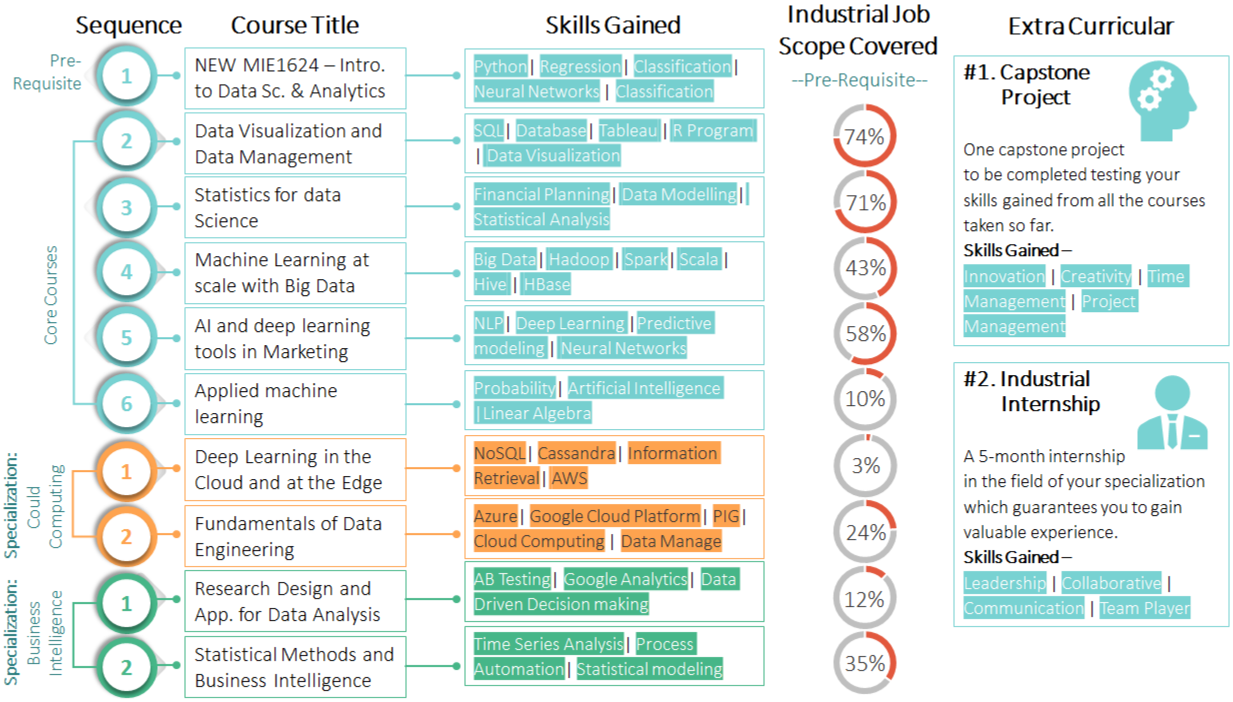


Figure 2. Job Count Distribution w.r.t Skill Clusters

s



	Sequence	Course Title	Skills Gained	Industrial Job Scope Covered
Pre-Requisite	1	NEW MIE1624 – Intro. to Data Sc. & Analytics	Python| Regression| Classification| Neural Networks | Classification	--Pre-Requisite--
Core Courses	2	Data Visualization and Data Management	SQL| Database| Tableau | R Program | Data Visualization	74%
	3	Statistics for data Science	Financial Planning| Data Modelling| Statistical Analysis	71%
	4	Machine Learning at scale with Big Data	Big Data| Hadoop | Spark| Scala | Hive | HBase	43%
	5	AI and deep learning tools in Marketing	NLP| Deep Learning | Predictive modeling | Neural Networks	58%
	6	Applied machine learning	Probability| Artificial Intelligence |Linear Algebra	10%
Specialization: Could Computing	1	Deep Learning in the Cloud and at the Edge	NoSQL| Cassandra| Information Retrieval| AWS	3%
	2	Fundamentals of Data Engineering	Azure| Google Cloud Platform| PIG| Cloud Computing | Data Manage	24%
Specialization: Business Intelligence	1	Research Design and App. for Data Analysis	AB Testing| Google Analytics| Data Driven Decision making	12%
	2	Statistical Methods and Business Intelligence	Time Series Analysis| Process Automation| Statistical modeling	35%

Extra Curricular

#1. Capstone Project

One capstone project to be completed testing your skills gained from all the courses taken so far.

Skills Gained – Innovation | Creativity | Time Management | Project Management

#2. Industrial Internship

A 5-month internship in the field of your specialization which guarantees you to gain valuable experience.

Skills Gained – Leadership | Collaborative | Communication | Team Player

Figure 3. New Master of Data Science and Artificial Intelligence Curriculum

Signature courses will include Data Visualization and Data Management, Statistic for Data Science, Machine Learning at Scale with Big Data, AI and Deep Learning Tools in Marketing and Applied Machine Learning. These courses match the skills clusters and will prepare students to excel. Figures 4 through 8 below visualize these classes.

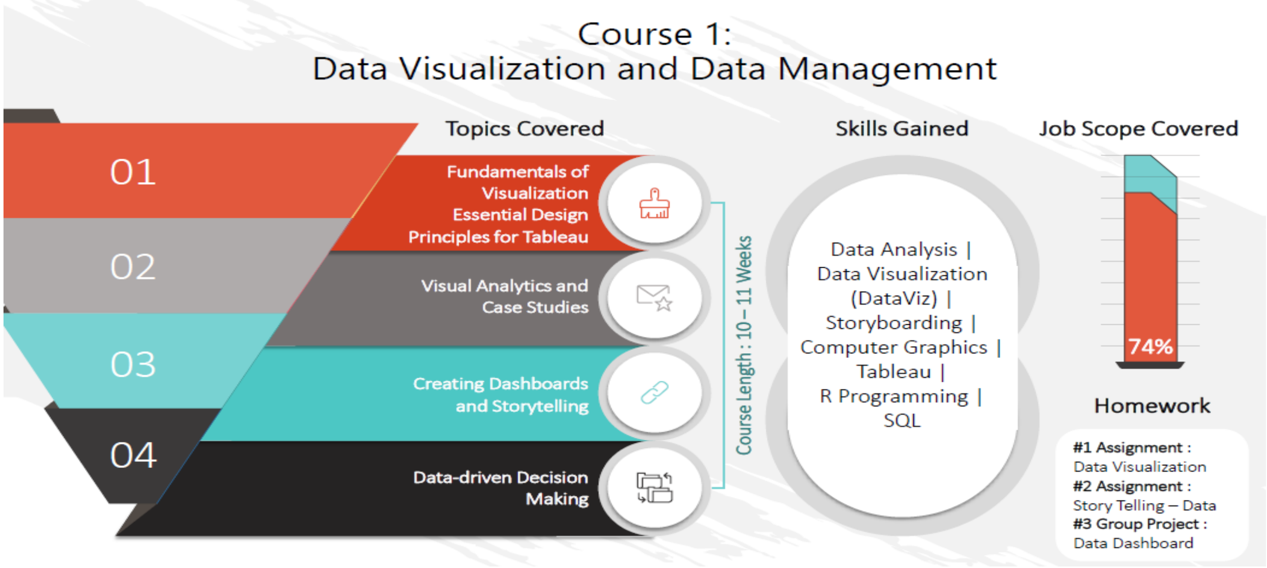
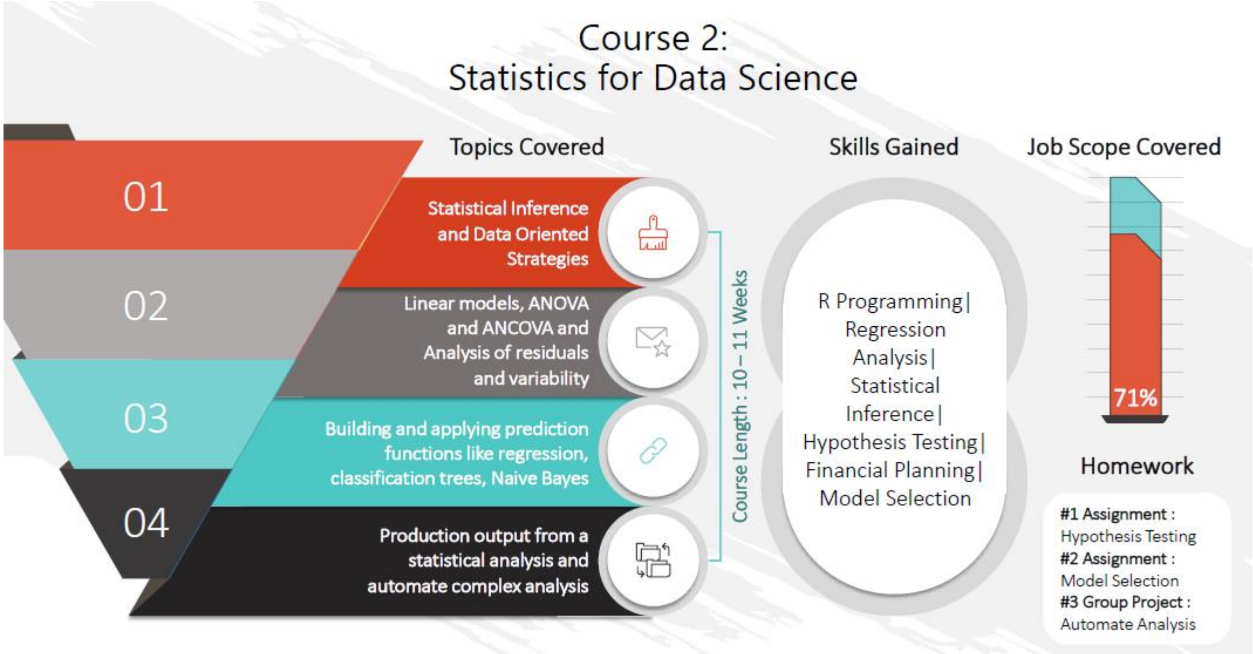

Course 1:
Data Visualization and Data Management
Topics Covered
Skills Gained
Job Scope Covered
01
Fundamentals of Visualization Essential Design Principles for Tableau
02
Visual Analytics and Case Studies
03
Creating Dashboards and Storytelling
04
Data-driven Decision Making
Course Length : 10 – 11 Weeks
Data Analysis | Data Visualization (DataViz) | Storyboarding | Computer Graphics | Tableau | R Programming | SQL
74%
Homework
#1 Assignment : Data Visualization
#2 Assignment : Story Telling – Data
#3 Group Project : Data Dashboard

Figure 4. Course 1: Data Visualization and Data Management Visual



Course 2:
Statistics for Data Science

Topics Covered

01 Statistical Inference and Data Oriented Strategies

02 Linear models, ANOVA and ANCOVA and Analysis of residuals and variability

03 Building and applying prediction functions like regression, classification trees, Naive Bayes

04 Production output from a statistical analysis and automate complex analysis

Course Length : 10 – 11 Weeks

Skills Gained

R Programming| Regression Analysis| Statistical Inference| Hypothesis Testing| Financial Planning| Model Selection

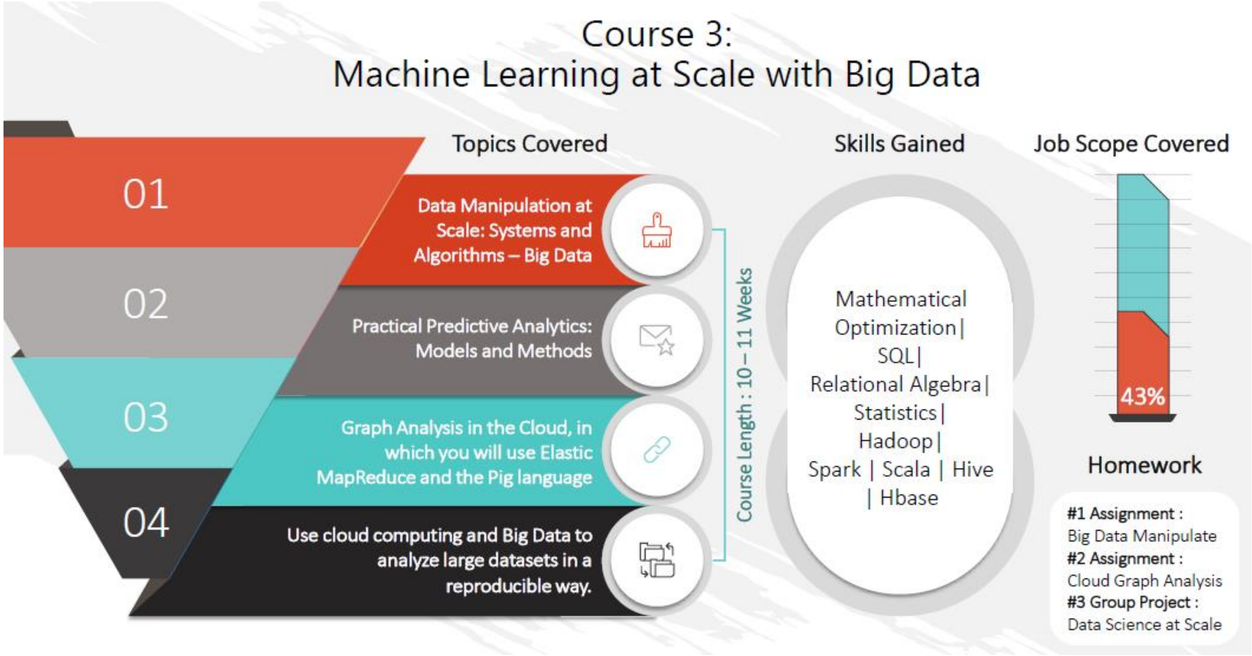
Job Scope Covered

71%

Homework

#1 Assignment : Hypothesis Testing
#2 Assignment : Model Selection
#3 Group Project : Automate Analysis

Figure 5. Course 2: Statistics for Data Science Visual



Course 3:
Machine Learning at Scale with Big Data

Topics Covered

01 Data Manipulation at Scale: Systems and Algorithms – Big Data

02 Practical Predictive Analytics: Models and Methods

03 Graph Analysis in the Cloud, in which you will use Elastic MapReduce and the Pig language

04 Use cloud computing and Big Data to analyze large datasets in a reproducible way.

Course Length : 10 – 11 Weeks

Skills Gained

Mathematical Optimization| SQL| Relational Algebra| Statistics| Hadoop| Spark | Scala | Hive | Hbase

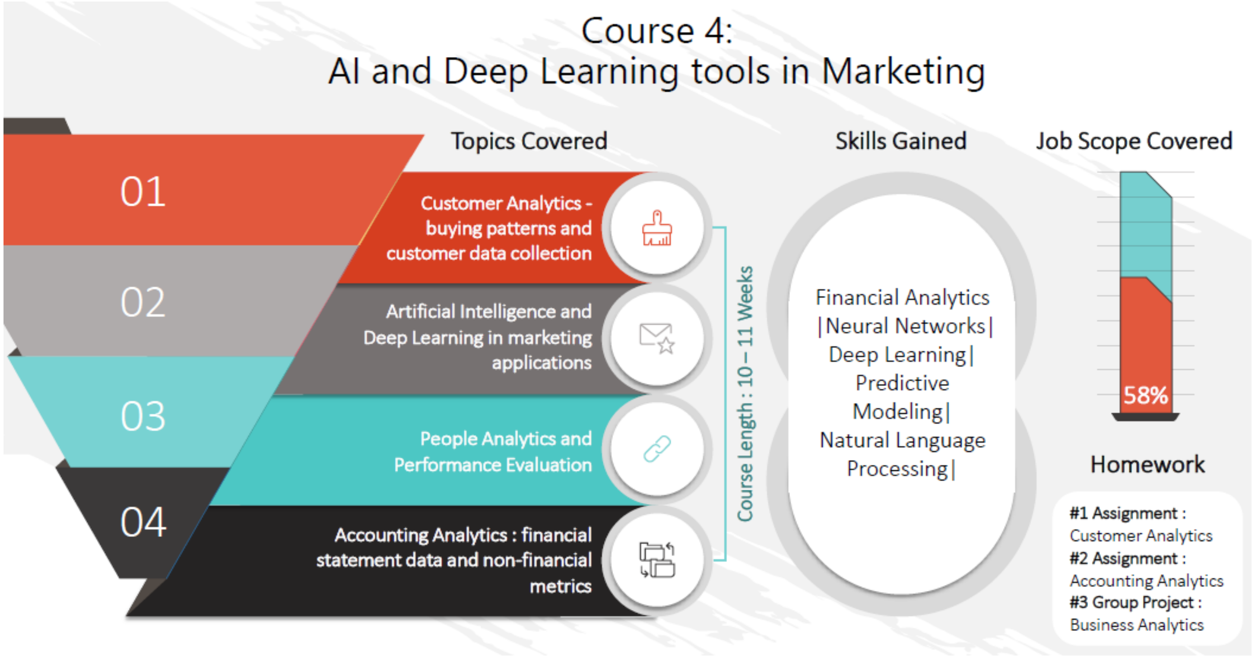
Job Scope Covered

43%

Homework

#1 Assignment : Big Data Manipulate
#2 Assignment : Cloud Graph Analysis
#3 Group Project : Data Science at Scale

Figure 6. Course 3: Machine Learning at Scale with Big Data Visual



Course 4:
AI and Deep Learning tools in Marketing

Topics Covered

01 Customer Analytics - buying patterns and customer data collection

02 Artificial Intelligence and Deep Learning in marketing applications

03 People Analytics and Performance Evaluation

04 Accounting Analytics : financial statement data and non-financial metrics

Course Length : 10 – 11 Weeks

Skills Gained

Financial Analytics |Neural Networks| Deep Learning| Predictive Modeling| Natural Language Processing|

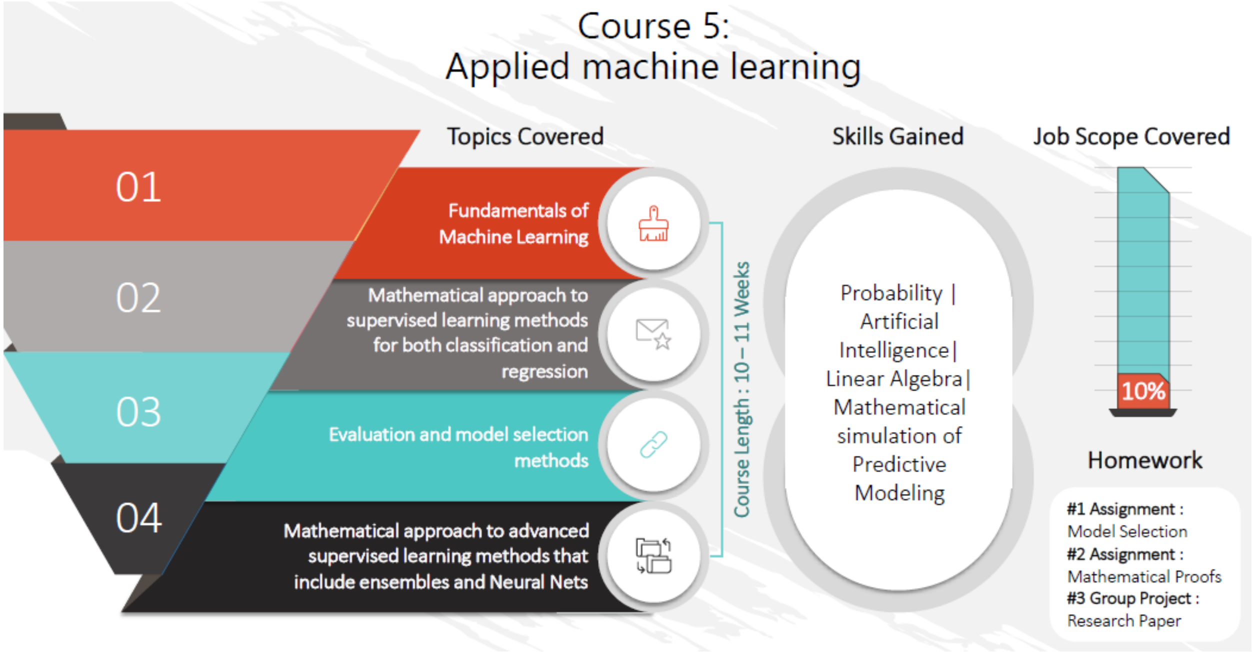
Job Scope Covered

58%

Homework

#1 Assignment : Customer Analytics
#2 Assignment : Accounting Analytics
#3 Group Project : Business Analytics

Figure 7. Course 4: AI and Deep Learning Tools in Marketing Visual



Course 5:
Applied machine learning

Topics Covered

01 Fundamentals of Machine Learning

02 Mathematical approach to supervised learning methods for both classification and regression

03 Evaluation and model selection methods

04 Mathematical approach to advanced supervised learning methods that include ensembles and Neural Nets

Course Length : 10 – 11 Weeks

Skills Gained

Probability | Artificial Intelligence| Linear Algebra| Mathematical simulation of Predictive Modeling

Job Scope Covered

10%

Homework

#1 Assignment : Model Selection
#2 Assignment : Mathematical Proofs
#3 Group Project : Research Paper

Figure 8. Course 5: Applied Machine Learning Visual

5. Course Selection Utility Recommender System

The University of Toronto faculty of engineering offers an emphasis in analytics to the Master of Engineering students. Students taking this degree will not have access to the premier courses developed as part of the new Master of Business Analytics and Engineering program, however it is important to maximize the utility of the existing courses. To assist students in selecting courses for their emphasis in analytics, a tool was developed that receives a job posting from indeed and returns the courses students should take. The courses are evaluated against the job posting similarly to how courses were scored when developing the redesign of MIE1624. The more skills that match, the higher the course rating. This tool is considered highly useful, as often students know the job or field they want to enter, but are unsure of how to get there.

Five courses are recommended by the utility. To satisfy the emphasis requirements, APS1070 and one of either MIE1624 or ECE1513 are always selected. Each time a course is selected, the importance of the skills it satisfies in the job posting is reduced before selecting the next course. This helps to ensure a diversification of skills from the courses. Table 3 shows the output of two different jobs with a data science component.

Table 3. Output of Two Different Jobs with a Data Science Component

Transportation Planner with Entuitive	Information Systems Analyst with Humber River Hospital	Data Scientist I with TD Bank
APS1070: Foundations of Data Science and Machine Learning (Required)	APS1070: Foundations of Data Science and Machine Learning (Required)	APS1070: Foundations of Data Science and Machine Learning (Required)
MIE1624: Introduction to Data Science and Analytics (Core)	MIE1624: Introduction to Data Science and Analytics (Core)	MIE1624: Introduction to Data Science and Analytics (Core)
CIV1507: Public Transport (Elective)	MIE1623: Introduction to Healthcare Engineering (Elective)	MIE1628: Big Data Science (Elective)
APS1070: Supply Chain Management and Logistics (Elective)	APS502: Financial Engineering (Elective)	APS1022: Financial Engineering II (Elective)
CIV1532: Fundamentals of ITS and Traffic Management (Elective)	APS1005: Operations Research for Engineering Management (Elective)	APS1040: Quality Control For Engineering Management (Elective)

6. Conclusions and Recommendations

By leveraging the power of data analytics to determine exactly what employers are looking for now, the MIE1624 curriculum was redesigned, a new Masters program was developed, and a utility to help engineering students select courses was produced.

The redesigned MIE1624 incorporates big data science and tools, as well as practical experience with Tableau, to improve its generalizability while maintaining a reasonable degree of depth for each subject that it engages with. The redesigned course better suits more jobs than any other course surveyed from world leading analytics programs. With this course, engineering students will be prepared to acquire and excel in a data science position in any industry.

The clusters of skills identified from job postings were matched and used to build a new program: the Masters of Business Analytics and Engineering. This program will produce industry leaders, and allow graduates to select any data science in business job as they will have graduated with all skillsets that employers are seeking.

Finally, for students outside of the new program and taking an emphasis in analytics as part of the Master of Engineering, a new utility has been developed that will help them select courses to enter the career of their choice. By inputting a job posting from Indeed, skills are extracted and matched to courses. The courses provided by the utility will meet the top skills of the position, and the algorithm has been equipped with a feature to ensure diversification of the skills.

Appendix A: Original and Revised MIE1624 Curriculum

Table 4. Existing MIE1624 Course Curriculum

Existing MIE1624 Curriculum	
Course title:	Introduction to Data Science and Analytics (MIE1624H)
Course description:	The objective of the course is to learn analytical models and overview quantitative algorithms for solving engineering and business problems. Data science or analytics is the process of deriving insights from data in order to make optimal decisions. It allows hundreds of companies and governments to save lives, increase profits and minimize resource usage. Considerable attention in the course is devoted to applications of computational and modeling algorithms to finance, risk management, marketing, health care, smart city projects, crime prevention, predictive maintenance, web and social media analytics, personal analytics, etc. We will show how various data science and analytics techniques such as basic statistics, regressions, uncertainty modeling, simulation and optimization modeling, data mining and machine learning, text analytics, artificial intelligence and visualizations can be implemented and applied using Python. Python and IBM Watson Analytics are modeling and visualization software used in this course. Practical aspects of computational models and case studies in Interactive Python are emphasized.
Course Outline	
Week 1: Introduction to data science and analytics	1. Data science concepts
	2. Application areas of quantitative modeling
Week 2: Python programming, data science software	1. Introduction to Python
	2. Comparison of Python, R and Matlab usage in data science
Week 3: Basic statistics	1. Random variables, sampling
	2. Distributions and statistical measures
	3. Hypothesis testing
	4. Statistics case studies in IPython Overview of linear algebra

	5. Linear algebra and matrix computations
	6. Functions, derivatives, convexity Modeling techniques, regression
	7. Mathematical modeling process
	8. Linear regression
	9. Logistic regression
	10. Regression case studies in IPython
Week 4: Data visualization and visual analytics	1. Visual analytics
	2. Visualizations in Python and visual analytics in IBM Watson Analytics
Week 5: Data mining and machine learning	1. Classification (decision trees)
	2. Clustering (K-means, Fuzzy C-means, Hierarchical Clustering, DBSCAN)
	3. Association rules
	4. Advanced supervised machine learning algorithms (Naive Bayes, k-NN, SVM)
	5. Intro to ensemble learning algorithms (Random Forest, Gradient Boosting)
	6. Data mining case studies in IPython
Week 6: Optimization	1. Unconstrained non-linear optimization algorithms
	2. Overview of constrained optimization algorithms
	3. Optimization case studies in IPython
Week 7: Simulation modeling	1. Random number generation
	2. Monte Carlo simulations
	3. Simulation case studies in IPython
Week 8: Cognitive computing and artificial intelligence	1. Intro to neural networks and deep learning
	2. Text analytics and natural language processing
	3. Reinforcement learning

	4. Spatio-temporal analytics
	5. Cognitive computing case studies in IPython
Week 9: Storytelling based on analytics, analytical decision making	1. Validating analytics
	2. Storytelling based on analytics
	3. Decision-making based on analytics

Table 5. MIE1628 Curriculum

MIE1628 Curriculum	
Course title:	Big Data Science (MIE1628H)
Course description:	This course is designed to provide students with fundamental understanding of Big Data and help develop skills necessary to handle and implement various aspects of big data projects. The course has an additional focus on Machine Leaning & Data Science developed to advance data science skills that are often required to implement big data projects. Students, while taking the course, will have a unique exposure to both data science and big data technologies that will provide them with skills necessary to implement data science projects in Hadoop as well as support teams executing big data projects.
Course Outline	
Lecture #1 (May 6th): Introduction to Hadoop	1. Definition of big data
	2. Big data industry review
	3. History of Hadoop and distributed computing
	4. Brief overview of HDFS, Yarn, MapReduce, Spark
	5. Review of the big data resources that will be used for this course
	6. Final project: objective and requirements
	7. Review of additional resources to help with big data skills
	1. HDFS, YARN in detail

Lecture 2 (May 8th): Enterprise Hadoop and Hadoop basics	2. Launch of the cluster environment and practice of HDFS and YARN commands
	3. Invited lecture by an industry practitioner responsible for establishing an enterprise Hadoop architecture: Melissa Singh a. Implementation of a big data project: review of a real business case with focus on understanding of the complexity of interactions and convergence of DevOps, legal, data management and data science teams
Lecture 3 (May 13th): SQL: Hive and Impala	1. Programming languages: overview, object-oriented vs functional, scripting languages
	2. Big data languages: overview of Spark, Hive & Impala
	3. SQL: most important commands
	4. Hive vs Impala: comparison between two SQL languages and environments
	5. Assignment 1: data mining using Hive
Lecture 4 (May 15th): Spark, introduction	1. Introduction into Spark a. History of Spark, its evolution, libraries b. Review of Spark APIs: PySpark, Scala, R and Java c. Spark backend transformations d. Practice sessions with DataBricks, Scala
Lecture 5 (May 22nd): Spark, syntax Introduction to Hadoop	1. Spark session
	2. Import/export of data
	3. Spark SQL: functional and scripting a. 1 student paper review: Spark SQL original paper
	4. Working with SQL images
	5. Most important Scala functions
	6. Assignment #2 (Scala)
Lecture 6 (May 27th): Spark: Data Preparation and Feature Transformation for ML	1. Feature preparation: overview of main steps (in Spark) a. Handling nulls/NaNs, missing values, outliers, normalization
	2. Spark libraries for feature transformation
	3. Feature transformation: dimensionality reduction, clustering, conversion from categorical to numerical

	4. Principle component analysis
	5. Regularization: L1, L2
Lecture 7 (May 29th): Spark, model development, pipelines, evaluation metrics	1. Model development process and necessary steps
	2. Spark pipelines
	3. TrainTestSplit and CrossValidation
	4. Regularization: L1, L2
	5. Model evaluation metrics for different classes of models
Lecture 8 (June 3rd): Spark ML: Multivariate analysis and time-series analysis	1. Multivariate ML models in Spark: regression trees
	2. Time-series analysis
	3. Invited speaker: Rogelio Cuevas from TD Securities (H2O?)
Lecture 9 (June 10th): Advanced data science topics	1. Graph theory
	2. Deep learning libraries
Lecture 10 (June 12th): Cloud for Big Data, two invited speakers	

Table 6. Modifications of Existing Course Curriculum

Modifications of Existing Course Curriculum	
Week	**Changes**
1	Assign reading if refresher on pre-requisite content is required - Introduce Python programming, basic statistics & linear algebra
2	Replace lecture with mini-assignment that requires use of basic python skills, learn python by applying, not in lecture
3	Combine the previous three subjects - with a brief refresher on stats/linear algebra
4	Add a mini-assignment using Tableau, it is a fairly easy interface and the library has it
5	No change (very important)

6	No change (very important)
7	No change (very important)
8	No change (very important)
9	Include tableau visualizations in project
Note: Three weeks were saved by condensing intro lectures - assign these to introduction to big data science.	

Table 7. Additional Content from MIE1628 -Big Data Science

Additional Content from MIE1628 -Big Data Science	
Week	**Additions**
8	Add week 2 of MIE1628
9	Add week 3 of MIE1628
10/11/ 12	Combine Lectures 4 through 8 from MIE1628 providing introductory material, and discussing the workflow at a higher level

Table 8. New Re-designed MIE1624 Course Curriculum

New MIE1624 Curriculum	
Course title:	Introduction to Data Science and Analytics (MIE1624H)

Course description:	The objective of the course is to learn analytical models and overview quantitative algorithms for solving engineering and business problems. Data science or analytics is the process of deriving insights from data in order to make optimal decisions. It allows hundreds of companies and governments to save lives, increase profits and minimize resource usage. Considerable attention in the course is devoted to applications of computational and modeling algorithms to finance, risk management, marketing, health care, smart city projects, crime prevention, predictive maintenance, web and social media analytics, personal analytics, etc. We will show how various data science and analytics techniques such as basic statistics, regressions, uncertainty modeling, simulation and optimization modeling, data mining and machine learning, text analytics, artificial intelligence and visualizations can be implemented and applied using Python. Python and IBM Watson Analytics are modeling and visualization software used in this course. Practical aspects of computational models and case studies in Interactive Python are emphasized. Visualization packages such as Tableau are introduced and used in assignments. The course includes an introduction to tools used for machine learning at an industry scale, including Hadoop and Hadoop basics, introduction to big data languages (SQL, Hive, Impala) and machine learning workflows in Spark.
Course Outline	
Week 1: Introduction to data science and analytics	1. Data science concepts
	2. Application areas of quantitative modeling
	3. Review of basic statistics
	4. Review of linear algebra
	5. Introduction to Python and comparison of languages in data science
Instructor Notes: Statistics and linear algebra not lectured in detail, readings assigned for refresher and preparation for coming weeks. Assignment 1: Python Basics. 5% of grade. Small project working through and gaining familiarity with the basics of python programming.	
Week 2: Modeling techniques, regression	1. Mathematical modeling process
	2. Linear regression
	3. Logistic regression
	4. Regression case studies in IPython

Week 3: Data visualization and visual analytics	1. Visual analytics
	2. Visualizations in Python and visual analytics in IBM Watson Analytics Storytelling based on analytics, analytical decision making
	3. Validating analytics
	4. Storytelling based on analytics
	5. Decision-making based on analytics
Instructor Notes: Assignment 1 Due, Assignment 2: Regression and Tableau Dashboard Creation, 25% of grade. Performing a data cleaning and logistic regression project similar to assignment 1 in original course, but with the add-on of a creation of a dashboard in Tableau.	
Week 4: Data mining and machine learning	1. Classification (decision trees)
	2. Clustering (K-means, Fuzzy C-means, Hierarchical Clustering, DBSCAN)
	3. Association rules
	4. Advanced supervised machine learning algorithms (Naive Bayes, k-NN, SVM)
	5. Intro to ensemble learning algorithms (Random Forest, Gradient Boosting)
	6. Data mining case studies in IPython
Week 5: Optimization	1. Unconstrained non-linear optimization algorithms
	2. Overview of constrained optimization algorithms
	3. Optimization case studies in IPython
Week 6: Simulation modeling	1. Random number generation
	2. Monte Carlo simulations
	3. Simulation case studies in IPython
Instructor Notes: Assignment 2 Due. Assignment 3: ML Algorithms and Text Processing, 30% of grade. Performing data cleaning, text analysis and application of ML models similar to existing MIE1624 assignment 2.	
	1. Intro to neural networks and deep learning

Week 7: Cognitive computing and artificial intelligence	2. Text analytics and natural language processing
	3. Reinforcement learning
	4. Spatio-temporal analytics
	5. Cognitive computing case studies in IPython
Week 8: Enterprise Hadoop and Hadoop basics	1. HDFS, YARN in detail
	2. Launch of the cluster environment and practice of HDFS and YARN commands
	3. Invited lecture by an industry practitioner responsible for establishing an enterprise Hadoop architecture: Melissa Singh a. Implementation of a big data project: review of a real business case with focus on understanding of the complexity of interactions and convergence of DevOps, legal, data management and data science teams
Week 9: SQL (Hive and Impala)	1. Programming languages: overview, object-oriented vs functional, scripting languages
	2. Big data languages: overview of Spark, Hive & Impala
	3. SQL: most important commands
	4. Hive vs Impala: comparison between two SQL languages and environments
	5. Assignment 1: data mining using Hive
Instructor Notes: Assignment 4: multi-part assignment consisting of smaller tasks including data mining using Hive, exploring SQL, and Spark, 20% of grade.	
Week 10: Spark, introduction	1. Introduction into Spark a. History of Spark, its evolution, libraries b. Review of Spark APIs: PySpark, Scala, R and Java c. Spark backend transformations d. Practice sessions with DataBricks, Scala
Instructor Notes: Assignment 3 Due	
Week 11/12 Spark: Overview of workflow - Data Preparation, Feature Transformation for ML, model development, multivariate/timeseries, pipelines and evaluation metrics	1. Feature preparation: overview of main steps (in Spark)
	2. Spark libraries for feature transformation
	3. Review of Feature transformation
	4. Multivariate and time-series ML models in Spark

	5. Model development process and necessary steps
	6. Spark pipelines
	7. Review of TrainTestSplit and CrossValidation
	8. Model evaluation metrics for different classes of models
Instructor Notes: Assignment 4 Due during Exam Period. Final exam - 20% of grade	

Appendix B: Master of Business Analytics and Engineering Course Descriptions

Program Requirements:

Completion of 8 graduate level courses.

- 5 core courses selected and using the model + 1 mandatory redesigned MIE 1624 course.
- Any 2 of the 4 specialized tracks.
- A course project
- A 5-6 month industrial internship.

Core Courses:

1. **Data Visualization and Database Management**

    - Fundamentals of Visualization, essential Design Principles for Tableau
    - Visual Analytics and Case Studies
    - Creating Dashboards and Storytelling
    - Data-driven Decision Making

    Assignments:

    - Assignment 1 – Data visualization
    - Assignment 2 – Storytelling - Data
    - Group project – Data dashboard

    Skills gained: Data Analysis, Data Visualization (DataViz), Storyboarding, Computer Graphics, Tableau, Programming, and SQL.

    Source: https://datascience.berkeley.edu/academics/curriculum/data-visualization/

2. **Statistics for Data Science**

    - Statistical Inference and Data Oriented Strategies

- Linear models, ANOVA and ANCOVA and Analysis of residuals and variability
- Building and applying prediction functions like regression, classification trees, Naive Bayes
- Production output from a statistical analysis and automate complex analysis

Assignments:

- Assignment 1 – Hypothesis testing
- Assignment 2 – Model selection
- Group project – Automate analysis

Skills gained: R Programming, Regression Analysis, Statistical Inference, Hypothesis Testing, Financial Planning, and Model Selection

Source: https://datascience.berkeley.edu/academics/curriculum/statistics-for-data-science/

3. Machine Learning at Scale with Big Data

- Data Manipulation at Scale: Systems and Algorithms – Big Data
- Practical Predictive Analytics: Models and Methods
- Graph Analysis in the Cloud, in which you will use Elastic MapReduce and the Pig language
- Use cloud computing and Big Data to analyze large datasets in a reproducible way

Assignments:

- Assignment 1 - Big Data Manipulate
- Assignment 2 - Cloud Graph Analysis
- Group project - Data Science at Scale

Skills gained: Mathematical Optimization, SQL, Relational Algebra, Statistics, Hadoop, Spark, Scala, Hive, Hbase

Source: https://datascience.berkeley.edu/academics/curriculum/machine-learning-at-scale/

4. Leveraging AI and Deep Learning Tools in Marketing

- Customer Analytics - buying patterns and customer data collection
- Artificial Intelligence and Deep Learning in marketing applications
- People Analytics and Performance Evaluation
- Accounting Analytics : financial statement data and non-financial metrics

Assignments:

- Assignment 1 - Customer Analytics
- Assignment 2 - Accounting Analytics
- Group project - Business Analytics

Skills gained: Financial Analytics, Neural Networks, Deep Learning, Predictive Modeling, and Natural Language Processing

Sourcehttps://www.rotman.utoronto.ca/Degrees/MastersPrograms/MMA/TheProgram/CourseDescriptions

5. **Applied Machine Learning**

- Fundamentals of Machine Learning
- Mathematical approach to supervised learning methods for both classification and regression
- Evaluation and model selection methods
- Mathematical approach to advanced supervised learning methods that include ensembles and Neural Nets

Assignments:

- Assignment 1 – Model selection
- Assignment 2 - Mathematical Proofs
- Group project - Research Paper

Skills gained: Probability, Artificial Intelligence, Linear Algebra, Mathematical simulation of Predictive Modeling

Source: https://datascience.berkeley.edu/academics/curriculum/applied-machine-learning/

Specialization courses:

Business Intelligence track -

1. **Research Design and Application for Data and Analysis**

Source:https://datascience.berkeley.edu/academics/curriculum/research-design-application-data-analysis/

This course introduces students to the burgeoning data sciences landscape, with a particular focus on learning how to apply data science reasoning techniques to uncover, enrich, and answer questions facing decision makers across a variety of industries and organizations today. The emphasis throughout is on making practical contributions to decisions that organizations will and should make. Industries explored include sports management, finance, energy, journalism, intelligence, healthcare, and media/entertainment.

Skills gained: SaaS, Ab Testing, Google Analytics, Data driven Decision

2. **Statistical Methods**
   Source: https://datascience.berkeley.edu/academics/curriculum/statistical-methods/

   Classical linear regression and time series models are workhorses of modern statistics, with applications in nearly all areas of data science. This course takes a more advanced look at both classical linear and linear regression models, including techniques for studying causality, and introduces the fundamental techniques of time series modeling. Mathematical formulation of statistical models, assumptions underlying these models, the consequence when one or more of these assumptions are violated, and the potential remedies when assumptions are violated are emphasized throughout.

   Skills gained: Problem Solving, Consulting, Business Intelligence, Word, Time series Analysis, Project Management, Process Automation, and Time Management

Cloud Computing track –

1. **Deep Learning in the Cloud and at the Edge**

   Source: https://datascience.berkeley.edu/academics/curriculum/deep-learning-in-the-cloud/

   This course provides a hands-on introduction to very large-scale data and the practical issues surrounding how the data is stored, processed, and analyzed, both in the Cloud and on the Edge. Students will work with cloud computing systems, edge devices, large data collections, and high-velocity data streams. Hands-on activities will enable the students to learn the practical toolkit required to work with data at scale. Deep Learning applications (image / video processing) will serve as the major use case throughout the class.

   Skills gained: NoSQL, Cassandra

2. **Fundamentals of Data Engineering**
   Source:https://datascience.berkeley.edu/academics/curriculum/fundamentals-of-data-engineering/

   Cloud computing is the on-demand availability of computer system resources, especially data storage and computing power, without direct active management by the user. Storing, managing, and processing datasets are foundational to both applied computer science and data science. Indeed, successful deployment of data science in any organization is closely tied to how data is stored and processed. This course introduces the fundamentals of data storage, retrieval, and processing systems in the context of common data analytics processing needs. This course aims to provide a set of "building blocks" by which one can construct a complete architecture for storing and processing data.

   Skills gained: Cloud Computing, Aws, Azure, Google Cloud, Pig

Mandatory Course:

1. **MIE 1624: Introduction to Data Science and Analytics**

   The objective of the course is to learn analytical models and overview quantitative algorithms for solving engineering and business problems. Data science or analytics is the process of deriving insights from data in order to make optimal decisions. It allows hundreds of companies and governments to save lives, increase profits and minimize resource usage. Considerable attention in the course is devoted to applications of computational and modeling algorithms to finance, risk management, marketing, health care, smart city projects, crime prevention, predictive maintenance, web and social media analytics, personal analytics, etc. We will show how various data science and analytics techniques such as basic statistics, regressions, uncertainty modeling, simulation and optimization modeling, data mining and machine learning, text analytics, artificial intelligence and visualizations can be implemented and applied using Python. Python and IBM Watson Analytics are modeling and visualization software used in this course. Practical aspects of computational models and case studies in Interactive Python are emphasized.

Course Project:

One capstone project to be completed testing your skills gained from all the courses taken so far.

Industrial Internship:

A 5-6 month internship in the field of your specialization which guarantees you to gain valuable experience and soft skills and prepare yourself for full time jobs.