

# **IMPEL Course Recommendation System**

**MEIE 4702**

**Technical Report**

**Capstone II Midterm/Final Report**

**Advisor: Prof. X. Jin**

**Design Team**

**Daniel Cone, Christian Etherton  
Selim Umit, Jason Kumar**

**April 27, 2022**

**Department of Mechanical and Industrial Engineering**

**College of Engineering, Northeastern University**

**Boston, MA 02115**

# **IMPEL Course Recommendation System**

## **Design Team**

**Daniel Cone, Christian Etherton  
Selim Umit, Jason Kumar**

## **Design Advisor**

**Xiaoning Jin**

## **Abstract**

The gap between the current level of knowledge in the new hire pool and the knowledge required for data science jobs in the production and manufacturing industry is an issue that has been growing in the past decade. The IMPEL Course Recommendation system serves to close this growing gap in knowledge. Content analysis is the first process that pools information such as key words and subject matter gathered from online job postings into domains which feed into large databases to later be sorted and matched with the course content. Following this, a user interface was developed to connect the experiences and knowledge of the user to potential course plans. This course plan is based on user input and is tailored to teach the user the knowledge they need to pursue their desired career path. This system is rooted in industry experience and expertise. After the completion of the Capstone project, the system was handed off to the IMPEL team to be linked to their curricula within the Canvas platform. It will continue to undergo iterative improvements to more closely match students situations and desires to the content that will help them succeed and play a bigger role in closing the employment gap.

## Table of Contents

1	Acknowledgements	5
2	Copyright	5
3	Introduction	6
3.1	Problem Statement	6
4	Background	6
4.1	Context of the Problem	6
4.2	Phases of the Project	7
5	Initial Concepts	11
5.1	Solutions Approach	11
5.2	Data Collection	11
5.3	Data and Demand	13
5.4	Data Interpretation / Analysis	14
5.5	KPIs & Constraints	15
5.6	Initial Simulations & Feedback	16
5.7	Criteria for Choosing the Final Solution	17
6	Final Solution	17
6.1	Overview	17
6.2	Results, Interpretation & Impact of Solution	20
7	Verification and Testing	21
8	Future Work	21
9	Ethics and Societal and Global Impact	22
9.1	Societal and Global Impacts	22
9.2	Diversity, Equity, and Inclusion Considerations	23
10	Validated Conclusion	23
11	Project Management	23
11.1	Timeline	23
12	Intellectual Property	24
12.1	Intellectual Property Used	24
12.2	Intellectual Property Created	24
12.3	Disposition of Intellectual Property	25
13	References	26

## List of Figures

Figure 1: Frequency gap analysis.....	8
Figure 2: Breakdown of course-modules and difficulty .....	9
Figure 3: First prototype of user interface .....	10
Figure 4: Skill-to-Job Field and Skill-to-LDA-Generated Topics matrix.....	12
Figure 5: System Process Diagram.....	16
Figure 6: Final version of User Interface with inputs .....	17
Figure 7: Populated Skills & Domain column .....	18
Figure 8: Skill-to-Job Field and Skill-to-LDA-Generated Topics matrix with mapping.....	18
Figure 9: Topics corresponding to selected Skills and Domains .....	19
Figure 10: LDA-Generated Topic-Course Module matrix .....	19
Figure 11: Final display of system output.....	20
Figure 12: Final output with recommended courses.....	20
Figure 13: Capstone 2 Gantt chart .....	24

## List of Tables

Table 1: Domain Frequency.....	<b>Error! Bookmark not defined.</b>
Table 2: Skill Frequency .....	8

# 1 Acknowledgements

The team draws special attention to Guoyan Li and Dr. Huanyi Shui. Though they are mentioned later in this report, their contributions were crucial to the team's success. The final solution is ultimately built on the decision-making process made possible by Guoyan Li's research and use of the LDA algorithm to derive topics out of data science job market data that connect skills to the IMPEL team's course modules. Without his work, the final product simply would not function. Dr. Huanyi Shui's direction during the design review process guided the team toward creating a system output for users that was more holistic, flexible, and professional. The team cannot thank these two individuals enough for their support and guidance.

# 2 Copyright

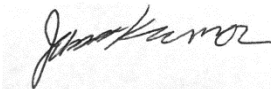
We the team members,



Selim Umit



Dan Cone



Jason Kumar



Christian Etherton



Xiaoning Jin, Faculty Advisor

hereby grant the Mechanical and Industrial Engineering (MIE) Department of Northeastern University unlimited non-exclusive license to use, modify or distribute this report and the corresponding Executive Summary and Poster. We also hereby agree that the video or other digital recordings of our Oral Presentations and Demonstrations are the full property of the MIE Department.

The publication of this report does not constitute approval by Northeastern University, the MIE Department or its faculty members of the findings or conclusions contained herein. It is published for the exchange and stimulation of ideas.

## 3 Introduction

### 3.1 *Problem Statement*

The labor markets, especially for data science and manufacturing, are filled with job postings that contain requirements that are impossible to meet when first entering the labor market. Students and those new to the labor market are often intimidated and can feel inadequate when looking at these postings. Recently job postings have become increasingly specific with requirements that the hire has an array of skills only someone in the labor force for years could have. Teaching industry-specific skills is a solution to this ever-growing problem of elevated expectations. What is needed is a system that can provide students with the necessary information to close any gap they have with their desired job. With this new system, students and those new to the labor market will be able to use their desired job to identify what they need to learn and be provided with a customized curriculum to efficiently prepare them for the job. Currently, there are very few solutions to this issue. People are left looking for ways to learn these skills from various in-person and online sources. This method often provides no legitimate qualifications on material learned. This is inefficient and time-consuming, causing hires to miss the opportunity to apply for jobs. This gap in actual knowledge and knowledge desired by employers is a huge drain on the job market causing many qualified hires to pass on opportunities while also causing employers to skip over hires because their resumes do not fulfil 100% of the desired skills. This leads to greater unemployment, less effective workers, and considerable time and energy loss for those looking for work and those looking for workers.

The IMPEL course recommendation system has been built to solve this issue affecting today's data science labor market. The solution approach involves identifying available educational resources and creating a system that recommends resources to best prepare individuals for their desired jobs, based on their competency and the workplace needs. In this instance, courses and modules created by the NSF-funded IMPEL program will be considered. The recommendation algorithm will be based on expert knowledge to map course topics to learners' needs. This will be refined over time with user feedback and market data to connect learners more precisely with the resources they need to get the jobs that they want. A database will connect a user interface to learners' course and module choices. The system will work because the labor market is constantly growing and ever changing, allowing the expansion of this system into other markets. A growing job market creates higher demand for a system that can aid people in making decisions regarding their professional development.

## 4 Background

### 4.1 *Context of the Problem*

Labor markets in the tech and data science industry have grown rapidly in recent years. However, a specific set of skills are required for a prospect to be hired or even considered by most employers. While having degrees such as a B.S. or even an M.S. have been the standard for years, there has been a recent shift toward considering hard skills [1]. Recently, graduated students are having difficulty getting hired by employers because they do not possess the exact skills required. These skills are typically those learned

while being in the labor force. The course recommendation system is designed to filter through thousands of job listings online, sort for specific keywords, and collect data on the most common employer-requested skills. After data was collected on this issue, it was algorithmically-grouped into general topic areas which were then directly mapped to course modules available to users. With this matrix and U.I. (user interface) developed, the user could then input their experiences and desired learnings into the U.I. to receive a curated list of modules connected to the skills required for the desired job field.

## ***4.2 Phases of the Project***

The first phase of the IMPEL Course Recommendation system required the collection of market data to analyze and assign to the course modules. The preliminary market data has been provided by the EMSI market analytics report on current and past skills and domain knowledge demand [2]. The report provides a baseline level of market data which defines what skills and domain knowledge are most frequently found in job postings and public resumes. The report also provides an insight into what skills and domain knowledge are currently rising in demand from year to year. A secondary section of data was also provided to the IMPEL project group. This data included the current gap between demand and supply for each skill or domain [3]. This gap is called the frequency gap and is calculated by taking the frequency from the demand side and subtracting the frequency from the supply side. The frequency gap reveals the skills and domains left unfilled in the job market. The data trends almost always show a greater demand than supply. The data collected from this report is what the courses are based on. The data shows what technical skills have the most demand compared to the supply. This lets the team decide what sort of courses are most needed, and the calculations made can be used to create a matrix that can link course modules to users.

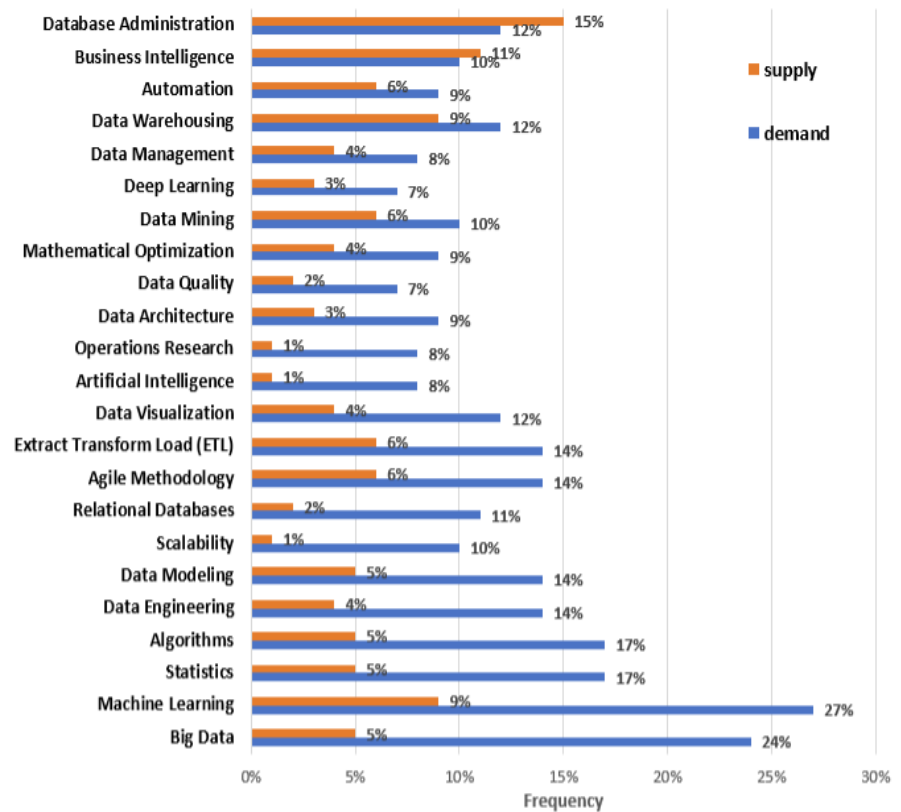
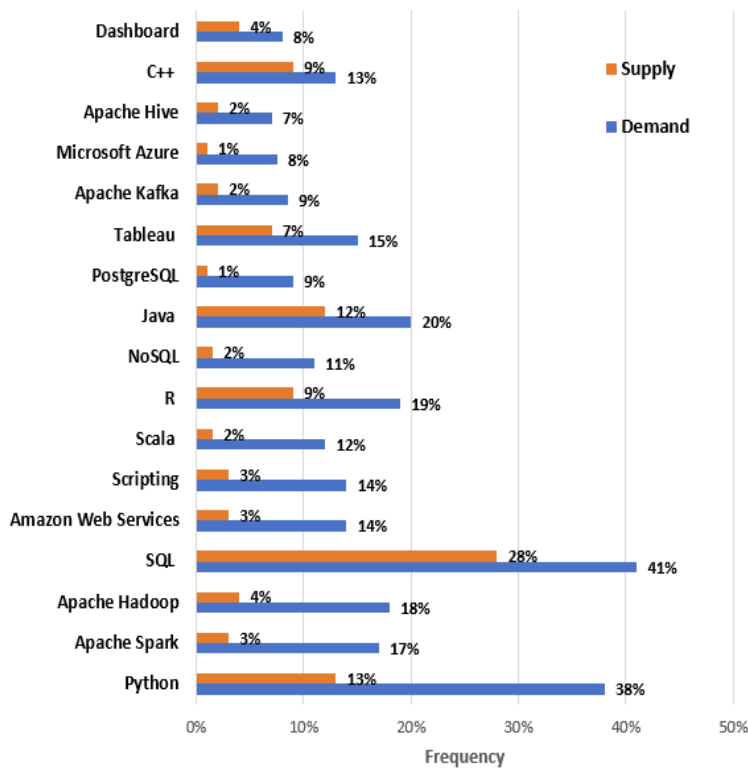


Figure 1: Frequency gap analysis

The IMPEL course recommendation system currently uses a set of course modules provided by three of the professors associated with the IMPEL team. These modules were further broken down into distinct categories that represent what can be learned by taking that module. Since the modules are not all the same level of difficulty, they were again broken down into beginner and advanced. With this breakdown it became easier to use the users' input, especially the input on experience, to help assign the modules. This is necessary because a user who begins with a lot of experience does not need to take the beginner level class and can then be assigned to the advanced level modules. Below is a breakdown of these separate categories. Note that our original categories included an intermediate category that ultimately was not used.



Class List							
	Data Analytics	Algorithms in Engineering Applications	Sensor Analytics	Robotics in Manufacturing	*might get syllabus Data Management for Analytics	Cyber-Manufacturing Systems	*just being started Data Visualization
<b>Beginner &gt;&gt;</b>	Introduction to Data Mining	Foundations of Algorithm	Measurement Systems, I/O,	Systems and components Current practices in R&A for Parts Handling and Bin Palletizing Cutting, grinding and	Database and Schema	Basic Modeling Concepts	Introduction to Data
<b>Intermediate &gt;&gt;</b>	Dimension Reduction Evaluating Predictive Multiple Linear Regression	Graphs and Graph Greedy Algorithms Divide and Conquer	Sensor Signal Processing Time series Analysis Feature Extraction & Regression and Correlation	Collaborative Robots Automated Ground Vehicles Efficiency, Safety and	Query Processing and Object-Relational Databases, XML Databases NoSQL Databases	Architectures (Design) Sensors and Sensor Communication Protocols Cyber-Secured IoT Networking and	Visualization Workflow Data Representation for Interactivity in Visualization Annotation and Color for Composition in Visualization Network Visualization Text Visualization
<b>Advanced &gt;&gt;</b>	k-Nearest Neighbors Naive Bayes Classifier Classification and Logistic Regression Neural Networks Discriminant Analysis Association Rules & Cluster Analysis	Dynamic Programming Network Flow Algorithms Interactability Metaheuristic Algorithms	Classification Methods with Clustering methods with Sensor Fusion for	Programmable LC Mechatronic	Big Data Data Integration, Quality, Database in Mfg. IT Sys.	Supervisory Control and MTCConnect, AutomationML,	Visualization Application of Application of Tableau
							*For future Selected topics in Data Science Probability Machine learning Machine Learning Linear programming Mathematical Python 1 Python 2 Python 3 Deep Learning 1 Deep Learning 2 Deep Learning 3 Data Visualization 1 Data Visualization 2 Data Visualization 3 Introduction to

**Figure 2: Breakdown of course-modules and difficulty**

While the end goal of the IMPEL project remains the same, the scope of the undergraduate project has changed since its conception in Capstone I. Originally, the scope involved the Capstone team collecting all the data, creating an algorithm, and creating the U.I. After the changes implemented by the graduate team and the professors, the scope has been narrowed to the creation of the U.I. and the matrices that will link the user with the desired course modules. Initially, excel was used to generate a skeleton system that could successfully suggest a set of modules based off the users selected input. This prototype involved a set of 60+ key words that will activate links to the IMPEL topics and modules when the user's input is collected. The U.I. involves a main excel sheet with input and output sections. The user inputs their desired skills and domain knowledge, which is then used to display an array of skills and domain knowledge. If the skills map to a particular topic, and that topic maps to particular course modules, they are identified in the displayed output. The prototype featured in Figure 3 also contains a preliminary screening U.I. that lets the user input their desired job field, which narrows down the number of key words the system must sort through.



## **5 Initial Concepts**

### ***5.1 Solutions Approach***

Some I.E. (industrial engineering) concepts that have been used in the implementation of this project are as follows: operations management, project management, job design, management engineering, process/systems engineering, ergonomics, cost/value engineering, quality engineering, facility management, and design review.

These concepts are a few that have been used to create the current form of the IMPEL project. In the final stages of the project, those pertaining to the design and design review process are what are being used the most. With a strong project direction, design review is an extremely important concept because while there is a base system in place, it is nowhere close to the final product. The design review is used to take the prototype and continue to refine it, making it more dynamic and user friendly. This was accomplished by creating an initial design and then having a project member run through the process and make comments on efficiency and effectiveness. The comments and recommended changes were then implemented into the system. This process was continually repeated. Project management is also an extremely important concept being used by the Capstone team. The project is at a critical stage in its development life. From this point forward, without project management and design review, forward progress can stall due to complacency. With proper management, the team can focus on which member needs to accomplish what by when so progress is still being made. With proper design review being practiced, there can be effective discussions on where to take the prototype and what it will begin to look like in its final form.

### ***5.2 Data Collection***

The IMPEL course recommendation project requires data to be collected from the current job market to determine what skills are in demand for new workforce employees. The team was given a preliminary set of data collected by EMSI market analytics that provided a clear view of the skills and domain knowledge that is currently in demand for the manufacturing market [2]. The skills and domain knowledge collected by EMSI market analytics was provided to the subject matter experts. These subject matter experts then used the excel interface in Figure 4, provided by the Capstone team, to assign sets of skills and domain knowledge to the proper topic that will eventually be linked to the course model that would teach the skills and domain knowledge.

#	Skill / Domain Knowledge	LDA - Generated Topics																		Job Fields						
		Topic 0	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10	Topic 11	Topic 12	Topic 13	Topic 14	Topic 15	Topic 16	Topic 17	Topic 18	Topic 19	Data Science	Automation	Cyber	Undecided	
		<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
1	SQL																						1	1	1	1
2	Python																						1	1	1	1
3	Java																						1	1	1	1
4	R																						1			
5	Hadoop																						1			
6	Spark																						1			
7	AWS																						1		1	
8	Tableau																						1			
9	Scripting																						1	1	1	1
10	C++																						1	1	1	1

**Figure 4: Skill-to-Job Field and Skill-to-LDA-Generated Topics matrix**

While this project is based off skills and domain knowledge, it is also dependent on the preferences of the users who will be taking the modules. To do this there will be a survey created to collect data on students' preferred methods of learning, whether that be through online classes, live in-person lectures, prerecorded lessons, etc. This survey will be completed alongside the main system. Once this data is collected, it can then be used to find the most preferred methods of learning and create certain sections dedicated to each method. After creating the input system, there will be a preference selection where the learner can put this information which will then be considered when assigning the learner courses and modules. It will also be used to put users together with those who have the same preference.

### 5.3 Data and Demand

The data currently supplied from the EMSI market analytics which involves job posting analytics taken from hundreds of millions of job postings from employers [2]. It also includes profile analytics from the public and self-reported information on profiles and resumes. The data in Table 1 shows these samples ends up with a distribution of skills and knowledge that is frequent in job postings. Using this data, it will be possible to form it into a usable collection that will help assign learners courses that will best help them in learning the required knowledge for their desired career [4]. Jobs in data and analytics, cyber and cloud, automation and robotics, and signal and sensor have all been fields that have had steady growth in the job market the past several years. Below is a shot of the frequency of domain knowledge found in data and analytics.

**Table 1: Domain Frequency**

Rank	Data & Analysis	
	Domain Knowledge	Frequency
1	Machine Learning	27%
2	Big Data	24%
3	Statistics	17%
4	Algorithms	17%
5	Data Engineering	14%
6	Agile Methodology	14%
7	Extract Transform Load	14%
8	Data Modeling	14%
9	Data Warehousing	12%
10	Data Visualization	12%

All of these are involved in manufacturing positions which require workers to be knowledgeable in systems in the field as well as smaller and minor topics. SQL, Python, Linux, and C++ have had the highest frequency of appearance in job postings in the previously mentioned fields, respectively. With this data there will be the ability to rank and prioritize which skills need to be taught more than others. This data can also be used to tailor a course program for our learners to guide them through the most efficient path to learn what they need.

## 5.4 Data Interpretation / Analysis

The data provided is in a format that reveals insight into what skills are the most frequent in job postings. The data is given in percentages, and the skills data is given in frequency of appearance on resumes and job postings. This means that the higher the percentage, the more often it is found in required skills in job applications. This data shows which skills are the most sought after. Looking at Table 2, C++ has on average the highest frequency across all job categories. With this information, it is important to provide a system that can accurately provide users who desire knowledge in C++ the specific modules that give an effective lesson on C++. Providing users with targeted module plans is an extremely important part of the IMPEL system. If the system is not able to accurately assign modules to the users, then it will need to undergo more development to get to that point. The data also provides insight on which modules will need to be prioritized in material. Having modules that provide up-to-date information on key topics is crucial to its success. This is something that will need to be the focus of the professors creating the course content, ensuring that it does not become out of date.

**Table 2: Skill Frequency**

Skills requirement statistics.

Data Science		Automation		Cyber		Sensor	
Skill	Frequency	Skill	Frequency	Skill	Frequency	Skill	Frequency
SQL	41.0 %	Python	7 %	Linux	27 %	C++	35 %
Python	38.0 %	C++	7 %	C++	22 %	C	32 %
Java	20.0 %	Linux	7 %	Java	21 %	Debugging	24 %
R	19.0 %	Java	6 %	Python	19 %	Firmware	23 %
Hadoop	18.0 %	Debugging	6 %	OS	18 %	Python	21 %
Spark	17.0 %	C	5 %	C	17 %	MATLAB	17 %
AWS	14.0 %	AutoCAD	5 %	AWS	14 %	Linux	15 %
Tableau	15.0 %	SolidWorks	5 %	Debugging	13 %	Oscilloscope	15 %
Scripting	14.0 %	C#	4 %	C#	13 %	Real-TimeOS	11 %
C++	13.0 %	SQL	4 %	Scripting	12 %	OS	10 %
Scala	12.0 %	Scripting	4 %	JavaScript	12 %	Scripting	8%
NoSQL	11.0 %	OS <sup>a</sup>	4 %	Unix	11 %	Git	6%
Dashboard	8.0 %			SQL	11 %	Java	6%
Kafka	8.5 %			Azure	9 %		
Azure	7.5 %			Git	9 %		
PostgreSQL	9.0 %			Scrum	8 %		
Apache Hive	7.0 %			OOP <sup>a</sup>	8 %		
				Jenkins	8 %		
				Docker	7 %		

## 5.5 *KPIs & Constraints*

The success of this project will be measured in several ways. The user interface of the course recommendation system must be user-friendly and be able to provide suggestions and context suitable for recent high school graduates and currently working professionals. The system should be easy to implement and not constrained by data. Despite the small number of modules that it will be recommending, the system should be scalable for larger sets of courses with the ability to process substantial amounts of feedback when it becomes automated. Because the system will initially be made using heuristics and expert faculty input, the recommendations should clearly resemble their recommendations. Finally, the surveys used to evaluate the effectiveness of the content and the learners' individual curricula should be developed to get consistent and reliable feedback. None of these KPIs (key performance indicators) are easily quantifiable but can and will be evaluated.

The biggest foreseeable constraint will be the flexibility of transferring the recommendation system from an excel-based U.I. to one run on HTML and Python. The team will be building this application – the U.I. and the feedback database. The team's communication virtually and across time zones was no longer as much of an issue as it was in Capstone 1. With Capstone 2 being in-person, there is better communication within the team and many more opportunities to develop a dynamic relationship amongst the team members. Feedback, initially from faculty and then from users, was necessary to fine tune the system. The amount and quality of this feedback is directly correlated to the success of the system. Criteria to be considered include the following:

- a) User-friendly and Scalable U.I.
- b) Algorithmic Output – Expert Suggestion Integrity
- c) Programming Proficiency
- d) Faculty Feedback
- e) Number of Feedback Subjects

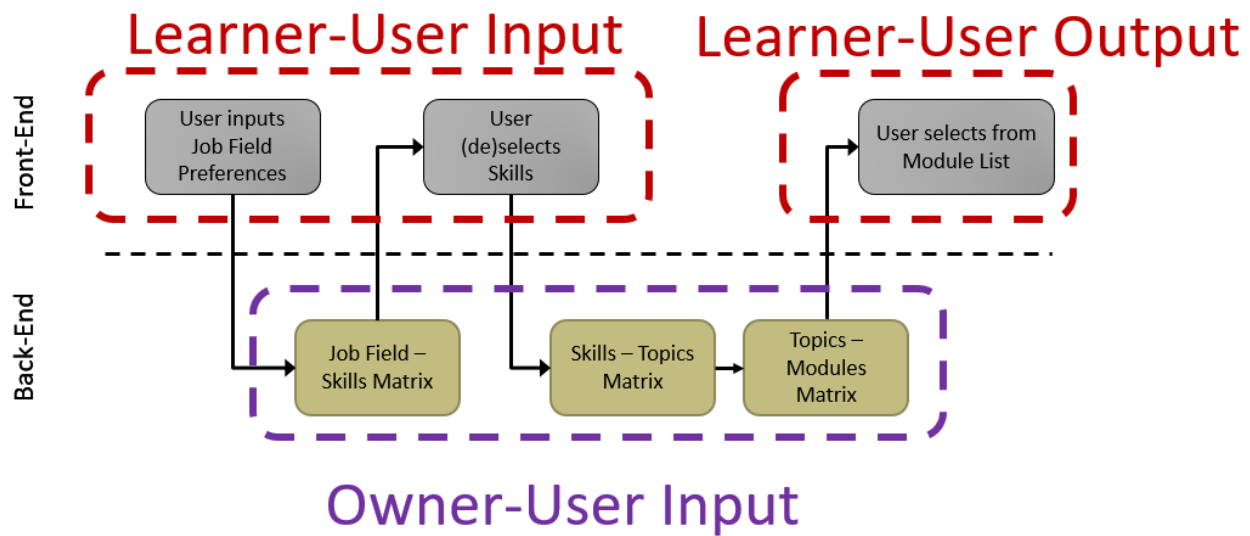
An initial survey was conducted in which the team considered how possible undergraduate students would react to the introduction of a course recommendation system in overall student life. Students surveyed demonstrated a keen interest in this tool, as it would allow them to be provided with courses tailored to them eventually. This would lead to a more beneficial academic career. The survey demonstrated a prevalent interest in a course recommendation system that is tailored to the skillsets, interests, and experience of the student in question. The survey also indicated that students believed certain issues and inconsistencies would arise related to the accuracy of the recommendations. This survey was conducted and then analyzed. It was particularly useful in the creation of the recommendation system because it was created with the benefit of the students in mind. Therefore, it was imperative to get adequate feedback from a sample of students themselves.

It is important to note that the criteria for implementing the course recommendation system will be derived from student input. Students will be responsible for entering their preferences and interests, as stated above. Recommendations will be displayed from those choices in the form of a list of recommended courses and a module map or sequence. This system will not derive data from outside sources such as the internet. This is a liability, as most data is derived from outside sources and not from

the students themselves. Since undergraduate students represent the main target audience, any results should be derived from students.

## 5.6 Initial Simulations & Feedback

With the help of an outside design reviewer, Dr. Huanyi Shui, a machine learning and artificial intelligence researcher at Ford Motor Company, additional updates were made to the user interface. Dr. Shui strongly recommended that the output of the user interface include additional courses the user could select that might not align perfectly with their inputs but might still align with the ultimate goals.



**Figure 5: System Process Diagram**

Initial configurations involved the user selecting their desired job which were mapped to algorithmically-derived topics. The user would then select from the correlated topics which were then mapped to the course modules. This did not work due to the lack of data and subject matter expertise available to connect specific jobs to the topics. Ultimately, the system requires users to select a broader job field which is mapped to specific skills based on research completed by the IMPEL team. After the user selects correlated skills, they are mapped through the same algorithmically derived topics to the modules, as seen in Figure 5 above.

The next phase of prototype development involved conferring with roommates and close friends as a preliminary way of gathering information on the effectiveness of the recommendation system. With this collected data, changes were made to the system. Once this revision was completed, the team planned for an official testable prototype to be sent out in mass to receive more thorough feedback data. From there the team would have additional information to further refine the system. To accomplish this, the team filled out the IRB application form to earn consent to test with human subjects. The team also acquired Amazon gift cards from the MIE department using funds from the Capstone course for a raffle for those who test the recommendation system. One of the ways that this system is going to be tested will be



sending out the software to those who signed up. With this data it will be easier to plan the next steps of development.

## 5.7 Criteria for Choosing the Final Solution

The IMPEL system is dependent on the accuracy of the modules provided to the users. With a system that can provide users with the correct modules that supply them with the knowledge they need to fill the gap, then the system will be considered a success. At the current stage of development, the IMPEL system has work that is required before taking further steps. To get to that stage there must be insight and feedback into the accuracy of assigned modules and the platform language that is preferred by most. Accurate module assignment is the most important because if users are assigned incorrect modules and must take them, they will find the system to be a waste of time. The platform is also important because having a system on a platform that users enjoy will encourage its usage to be continued. If the platform is undesirable, then first-time users may not return or continue to use it. Once a prototype platform is found that test users find enjoyable and productive to their time, then development on the final form can begin.

## 6 Final Solution

### 6.1 Overview

The final version of the IMPEL system that was chosen was focused on simplicity and ease of use for the user. Figure 6 shows the final user input section of the system. Example jobs have been narrowed down to 10 under each field making it easier for a user to look and find which job field their future career may lay under. The system is limited to 4 inputs. The first 3 are a part of the first round of user input. The 1<sup>st</sup> being the selection of the users' current level of expertise, the 2<sup>nd</sup> being the users' desired level of expertise, and the 3<sup>rd</sup> being a selection of their desired Job field.

User Input

Current Level of Expertise	Desired Level of Expertise	Desired Job Field

>>

Example Jobs in specified Job Field			
Data Science	Automation	Cyber	Sensor
Data Analysts	Manufacturing Engineers	Software Engineers	Software Engineers
Data Scientists	Mechanical Engineers	Security Counselors	Electrical Engineers
Data Engineers	Electrical Engineers	Electrical Engineers	Algorithm Engineers
Database Administrators	Automation Controls Engineers	Cloud Security Architects	Hardware Engineers
Big Data Engineers	Systems Engineers	Cybersecurity Systems Engineers	Embedded Software Engineers
Data Architects	Test Engineers	Cybersecurity Analysts	Digital Design Engineers
Data Visualization Engineers	Controls Engineers	Information Systems Security Officers	Electronics Technicians
Data Science Engineers	Project Engineers	Cybersecurity Architects	Signal Processing Engineers
Machine Learning Engineers	Maintenance Technicians	Network Engineers	Firmware Engineers
Business Analysts	Software Engineers	Solutions Architects	Radar Systems Engineers

Skills or Domains	Interest
	<input type="checkbox"/>
	<input type="checkbox"/>
	<input type="checkbox"/>
	<input type="checkbox"/>
	<input type="checkbox"/>
	<input type="checkbox"/>
	<input type="checkbox"/>
	<input type="checkbox"/>
	<input type="checkbox"/>
	<input type="checkbox"/>
	<input type="checkbox"/>
	<input type="checkbox"/>
	<input type="checkbox"/>
	<input type="checkbox"/>
	<input type="checkbox"/>
	<input type="checkbox"/>
	<input type="checkbox"/>
	<input type="checkbox"/>
	<input type="checkbox"/>

**Figure 6: Final version of User Interface with inputs**

The 4<sup>th</sup> user input is located after the selection of job field is filled out. This display or input is seen in Figure 7. After a selection of job field has been made, the “Skills or Domains” column is populated with all the skills or domains that correspond to that job field. Once this column has been populated, the user has the option to go through and select exactly which skills and/or domains they wish to learn. After those selections are made the system then has all the necessary information to populate the final output display.

Skills or Domains	Interest
Big Data	<input type="checkbox"/>
Statistics	<input type="checkbox"/>
Algorithms	<input type="checkbox"/>
Data Engineering	<input type="checkbox"/>
Data Modeling	<input type="checkbox"/>
Data Warehousing	<input type="checkbox"/>
Data Visualization	<input type="checkbox"/>
Database Administration	<input type="checkbox"/>
Data Mining	<input type="checkbox"/>
Mathematical Optimization	<input type="checkbox"/>
Data Architecture	<input type="checkbox"/>
Automation	<input type="checkbox"/>
Artificial Intelligence	<input type="checkbox"/>
Data Management	<input type="checkbox"/>
Operations Research	<input type="checkbox"/>
Deep Learning	<input type="checkbox"/>
Data Quality	<input type="checkbox"/>

**Figure 7: Populated Skills & Domain column**

The system uses the Skill-Topic matrix found in Figure 8 to find and collect all the skills and knowledge domains that correspond to the selected job field. These values were filled in by the team’s subject matter experts. If a value of 1 was filled in a cell that meant that the skill or domain was related to that job field as seen in the right 4 columns. With this information, it was possible to create the selection system in Figure 7.

#	Skill / Domain Knowledge	LDA - Generated Topics																		Job Fields									
		Topic-0	Topic-1	Topic-2	Topic-3	Topic-4	Topic-5	Topic-6	Topic-7	Topic-8	Topic-9	Topic-10	Topic-11	Topic-12	Topic-13	Topic-14	Topic-15	Topic-16	Topic-17	Topic-18	Robotics/AI/ML	Data Mining	Cyber-AI/ML Systems	Data Visualization	Data Science	Automation	Cyber	Senior	
		▼	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼	▼
1	Machine Learning		1									1	1	1									1	1	1	1			
2	Big Data		1			1	1	1	1	1		1	1	1									1	1	1	1			
3	Statistics				1								1										1	1	1	1			
4	Algorithms				1								1										1	1	1	1			
5	Data Engineering			1	1	1	1			1	1	1	1	1	1		1						1	1	1	1		1	1
6	Data Modeling				1	1	1	1	1	1		1	1	1	1		1	1	1				1	1	1	1			
7	Data Warehousing																						1	1	1	1			
8	Data Visualization		1																				1	1	1	1			
9	Database Administration																						1	1	1	1			
10	Data Mining				1	1	1	1	1				1										1	1	1	1			

**Figure 8: Skill-to-Job Field and Skill-to-LDA-Generated Topics matrix with mapping**

Following the selection of skills and domains in the right-hand column in Figure 7 the system then looks at which skills or domains have been selected. The system takes the selected values, travels back to the matrix in Figure 8, then looks for SME populated values and delivers a dynamic backend table of topics. This table of topics provides a visual as to which topics cover the selected skills and domains. Figure 9 shows rows corresponding to each selected skill and domain under the Data Science job selection dropdown and the topics required to teach each respective skills/domain.

Topic 0	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 8	Topic 9	Topic 10							Topic 16				
Topic 0	Topic 2		Topic 4					Topic 10											
	Topic 2	Topic 3			Topic 6	Topic 7	Topic 8		Topic 11		Topic 13		Topic 15						
	Topic 2	Topic 3	Topic 4					Topic 9	Topic 10		Topic 12			Topic 16		Topic 19			
	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6		Topic 8	Topic 9	Topic 10		Topic 12			Topic 16	Topic 17	Topic 19			
Topic 0		Topic 3	Topic 4																
	Topic 2	Topic 3	Topic 4	Topic 5				Topic 10						Topic 16					
			Topic 4				Topic 7		Topic 11		Topic 13	Topic 14	Topic 15						
			Topic 4																
	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10						Topic 17				
Topic 0	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10						Topic 16	Topic 17			
			Topic 4																
	Topic 2	Topic 3				Topic 7			Topic 11		Topic 13	Topic 14	Topic 15						
	Topic 3	Topic 4					Topic 9	Topic 10											
Topic 0	Topic 3	Topic 4						Topic 8											
	Topic 3	Topic 4																	
														Topic 17					

**Figure 9: Topics corresponding to selected Skills and Domains**

With a table collecting the necessary topics, the system then needed a way to relate the chosen topics to the proper course modules. To do this the Topic-Module matrix comes into play. The system takes a single value from each column of Figure 9, so there are no repeated values, then moves to the Topic-Module matrix finding all the modules that correspond to the topic above a certain correlation value. The correlation value is used to narrow or widen scope of modules that will be recommended. A higher correlation value means that the courses recommended will be strictly like what the user is looking for. A lower correlation value will include those courses as well as others that may extend past what is exactly desired.

		Topics																			
#	Module Name	Topic 0	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10	Topic 11	Topic 12	Topic 13	Topic 14	Topic 15	Topic 16	Topic 17	Topic 18	Topic 19
1	Overview of the Data Mining Process	0.003	0.077	0.047	0.025	0.435	0.05	0.010	0.005	0.015	0.018	0.018	0.014	0.011	0.008	0.045	0.002	0.103	0.005	0.025	0.056
2	Basic Data Visualization	0.014	0.042	0.018	0.017	0.069	0.035	0.011	0.012	0.006	0.006	0.022	0.007	0.032	0.005	0.011	0.009	0.012	0.022	0.006	0.045
3	Advanced Data Visualization	0.014	0.074	0.007	0.620	0.038	0.026	0.007	0.006	0.005	0.006	0.004	0.012	0.005	0.005	0.018	0.003	0.014	0.026	0.003	0.041
4	Dimension Reduction	0.012	0.083	0.107	0.015	0.079	0.032	0.004	0.004	0.006	0.012	0.013	0.015	0.030	0.004	0.006	0.003	0.042	0.008	0.004	0.058
5	Measures and Metrics for Performance Evaluation	0.002	0.065	0.126	0.003	0.043	0.004	0.004	0.003	0.004	0.020	0.476	0.003	0.007	0.003	0.011	0.003	0.186	0.005	0.011	0.005
6	Predictive and Classification Performance Evaluation Metrics	0.004	0.084	0.012	0.012	0.014	0.010	0.020	0.005	0.002	0.021	0.726	0.005	0.003	0.003	0.011	0.003	0.044	0.009	0.005	0.009
7	Multiple Linear Regression	0.003	0.058	0.443	0.006	0.014	0.004	0.005	0.006	0.007	0.013	0.020	0.014	0.040	0.004	0.014	0.005	0.283	0.012	0.003	0.033
8	k-Nearest Neighbors	0.005	0.088	0.280	0.035	0.033	0.035	0.021	0.004	0.005	0.017	0.147	0.021	0.005	0.006	0.008	0.004	0.056	0.006	0.016	0.085
9	The Naive Bayes Classifier	0.006	0.104	0.033	0.007	0.050	0.007	0.006	0.005	0.005	0.353	0.044	0.026	0.035	0.010	0.014	0.012	0.052	0.007	0.007	0.013
10	Classification Trees	0.003	0.126	0.022	0.005	0.044	0.011	0.545	0.005	0.007	0.081	0.047	0.005	0.006	0.006	0.023	0.005	0.042	0.005	0.007	0.005

**Figure 10: LDA-Generated Topic-Course Module matrix**

After this process is complete the system provides a complete list of courses that will teach the user exactly what they desire. After the team's design review with Dr. Shui, the final output was changed from a simple list of courses to a table of every course available. Figure 11 shows this table with all the courses, Figure 12 shows this table with all the courses and those recommended by the system. Dr. Shui encouraged this display method to allow the user to get a chance to see other courses that could grab their attention and not be limited to what is being recommended.

Class List			
	Data Analytics	Algorithms in Engineering Applications	Sensor Analytics
<b>Beginner &gt;&gt;</b>	Introduction to Data Mining Process Dimension Reduction Evaluating Predictive Performance Multiple Linear Regression	Foundations of Algorithm Graphs and Graph Algorithms Greedy Algorithms Dynamic Programming Network Flow Algorithms	Measurement Systems, I/O, Sensor Signal Processing Time Series Analysis Feature Extraction & Dimension Regression and Correlation
<b>Advanced &gt;&gt;</b>	k-Nearest Neighbors Naïve Bayes Classifier Classification and Regression Trees Logistic Regression Neural Networks Discriminant Analysis Association Rules & Collaborative Cluster Analysis	Interactability Metaheuristic Algorithms	Classification Methods with Clustering Methods with Sensor Fusion for Inferencing

Figure 11: Final display of system output

## 6.2 Results, Interpretation & Impact of Solution

Class List			
	Data Analytics	Algorithms in Engineering Applications	Sensor Analytics
<b>Beginner &gt;&gt;</b>	Introduction to Data Mining Process <b>Dimension Reduction</b> Evaluating Predictive Performance <b>Multiple Linear Regression</b>	Foundations of Algorithm Graphs and Graph Algorithms Greedy Algorithms Dynamic Programming Network Flow Algorithms	<b>Measurement Systems, I/O,</b> Sensor Signal Processing <b>Time Series Analysis</b> <b>Feature Extraction &amp; Dimension</b> <b>Regression and Correlation</b>
<b>Advanced &gt;&gt;</b>	<b>k-Nearest Neighbors</b> Naïve Bayes Classifier Classification and Regression Trees <b>Logistic Regression</b> Neural Networks <b>Discriminant Analysis</b> Association Rules & Collaborative <b>Cluster Analysis</b>	Interactability Metaheuristic Algorithms	<b>Classification Methods with</b> <b>Clustering Methods with</b> Sensor Fusion for Inferencing

Figure 12: Final output with recommended courses

The system successfully suggested modules, as shown in Figure 12 above, that do correspond to the chosen skills. In the above case, the selected job field was *Data Science*, and the selected skills were *Big Data* and *Data Engineering*. However, the results are not exactly as expected.

A few of the topics seem to offer a bridge for a group of correlated skills and a group of correlated modules. There are also either several skills or several topics which are not highly correlated with their next step – topics and modules, respectively. These relationships result in minimal differences for the output despite changing the inputs.

In the same way that errors propagate, minimal correlation likely propagates as the relationship is diluted using multiple matrices. The use of two matrices is helpful, however, in diluting too great an influence from bias within both the subject matter experts' knowledge and the algorithm's structure.

The content has been completed for three of the courses, but none of the courses are being actively used or tested. The success of the course recommendation system is inextricably tied to the availability, use, and evaluation of the courses and modules. The impact of this solution cannot be judged until the system is fully operational and connected to the courses and until the entire program is scaled or distributed and evaluated.

## **7 Verification and Testing**

To confirm the functionality of the system, the Capstone group shared the course recommendation system to others in two batches. The team initially sent the system out to close acquaintances and friends and had them test it out. These students were provided with a survey to collect their input and initial feedback based on their run-through of the system. Initially the pool of students surveyed varied in major and year.

Based on the results of the survey, the team was able to analyze certain common themes and improvements to be carried out. Feedback suggested that the recommendation system functioned as advertised for the most part, with few bottlenecks. Some students ran into issues in utilizing the excel system itself, which the team was later able to fix. The feedback also revealed some inconsistencies and areas of improvement regarding the matching between the skills and domains with the keyword matrices in the back end of the system. As the system was based out of excel, the design is basic, yet practical to begin with. The feedback indicated a need for an improvement in design, and more instruction for utilizing the system, specifically in labelling the fields of user input. Most of the feedback regarding the fidelity of the system was positive as most of the students found the course output helpful and saw further potential in the system itself.

Based on the initial feedback received, the system's user interface and back-end were refined. Instructions regarding the user input were added, and the overall design was also improved, providing the user with a more visually appealing experience. The initial survey also questioned students on the quality of the survey itself, as a refined survey would be a very useful tool to hand-off to the IMPEL team at the culmination of the semester.

## **8 Future Work**

According to the results of the survey, alterations were made to optimize the interface of the system. They helped make the U.I. more user friendly and with the hand-off to the IMPEL team, including this system being integrated into Canvas, there is potential for this course recommendation system to be adopted by other universities as a far more efficient course selection system.

Thanks to a more personalized approach to course selection, this would enable earlier course planning for students. This could reduce the anxiety of selecting courses upon arrival at university. Further funding and testing would also allow this system to expand and be utilized throughout the country.

More importantly, this system would reduce the volume of predictable approaches to university course plans, allowing students to plan for their desired career. This solution offers greater flexibility in degree plans and more classes relating to the career goals of an individual.

The system could also be scaled beyond IMPEL to be incorporated into several departments within Northeastern, such as the registrar's office. If it works for Northeastern, the system could indeed be scaled to other universities and educational programs. On top of all that, two-stage input could generally impact internet search optimization.

In the future, surveys of people who use IMPEL and can relate their professional experience to the courses and modules they decide to take could be analyzed collectively. This process could be modified so that a portion of the testers receive instructions on how to properly use it and some do not in order to analyze how usable and intuitive the recommendation system is. The results could then be incorporated into ML algorithms that would eventually replace the back-end matrices.

## **9 Ethics and Societal and Global Impact**

### ***9.1 Societal and Global Impacts***

The project's primary stakeholder is the IMPEL team, comprised of several members of the Northeastern faculty. The Capstone team has been working closely with them since its inception and has received crucial pointers and guidance in order to assist them in their goal. One of the primary objectives of the IMPEL team was to implement the recommendation system with Canvas, Northeastern University's main course dashboard. Although the scope of this project remains basic at this moment, course dashboard and recommendation system websites such as Canvas remain viable stakeholders in the future. This would be considered as the project's main impact, as it could completely alter Northeastern University's course selection process for students, as well as the faculty's method of recommending courses to students.

A successful release of a recommendation system could also potentially benefit students and professionals in reaching their desired educational goals. The versatility of this system allows it to be easily accessible and integrated to various similar dashboards, allowing for a potentially broader impact. The primary focus of the recommendation system are individuals interested in a career in data science within the manufacturing industry. It could expand into various other career paths and professions. It could also have an impact on large education companies that use recommendation systems. Outside of Canvas, examples include LinkedIn Learning, Coursera, and more. One notable, yet indirect impact could be the implementation of Canvas and a course recommendation system in high school education. This would allow high schoolers to develop academic goals and plans before entering college, giving them a head start on course planning and preventing their being overwhelmed with course selection upon arrival at university. Finally, another indirect impact of the IMPEL recommendation system would be the

integration of the system for job-seeking. A modified version of the system could map professional goals, skills, and level of experience to viable job fields and potentially postings.

## ***9.2 Diversity, Equity, and Inclusion Considerations***

The goal is to create a fully functional, well-rounded recommendation system which does not have any unintended consequences in terms of inclusion, poverty, etc. This means that the system will be usable by every possible demographic that has interest. A major possible positive consequence would be for the system to help marginalized communities and demographics connect with course content that could help them professionally. A possible negative consequence would be the recommendation system is in English which has the potential to widen the skills gap for non-English speakers. To avoid this, development of the system for use in multiple languages would need to be considered when moving forward into the commercial industry.

## **10 Validated Conclusion**

The team will follow the progress of the IMPEL program and the curriculum development, specifically. More importantly, the team will continue validating the U.I. structure and back-end matrices for the course recommendation system and begin testing it on users. The team looks forward to their role in closing this employment gap in the service of the greater manufacturing labor market.

## **11 Project Management**

### ***11.1 Timeline***

The project implementation plan was created in spring 2020. Then, during the spring, summer, and fall of the same year, Production 4.0 data science curriculum for manufacturing employees and students was designed, and modular courses that support active and project-based learning were developed. The development of these modular courses continues. As of April 2022, three of the eight courses have been completed.

Meanwhile, the development of the course-module recommendation system was initiated in spring 2021 at the start of Capstone 1. After researching alternative online MOOCs and their recommendation systems, the first prototype was created in November 2021 on excel. Initial feedback from the IMPEL team co-PIs in February 2022 contributed to the development of the second prototype. Subject matter experts then mapped the user inputs to the topics that are algorithmically connected to the modules. This is within the back end of the recommendation system and enables real output. After completing the matrices, the latest prototype was tested on close friends of the team. The Canvas team at Northeastern University will then implement the solution into the IMPEL program on the Canvas platform and update it with added content through Fall of 2022, until the end date of the IMPEL program at Northeastern

University and the public dissemination of IMPEL courses and modules. Figure 13 shows an updated version of the Gantt chart up to the end of the semester to reference the Capstone team's progress:

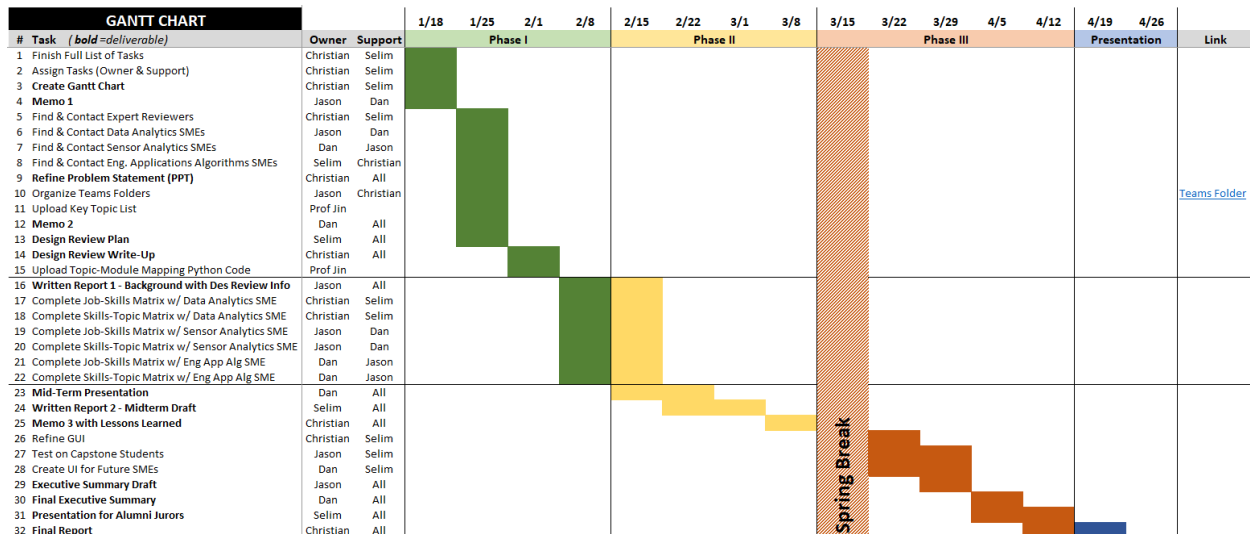


Figure 13: Capstone 2 Gantt chart

## 12 Intellectual Property

### 12.1 Intellectual Property Used

The current list of intellectual property that has been used thus far is listed as follows:

- IMPEL Team
  - Course Content
  - Research
  - Graduate Student's (Guoyan Li's) code & algorithm

Platforms and programs that have been used so far involve Microsoft Excel and Canvas, with Python potentially being added to this list in the future.

### 12.2 Intellectual Property Created

Intellectual property that has been created for this project is as follows:

- Recommendation System
  - Excel
  - UI
  - Matrices



### ***12.3 Disposition of Intellectual Property***

The faculty advisor and the IMPEL team will own the recommendation system. Their work and the recommendation system will eventually become open source.

## 13 References

- [1] George, Steve, and Kumar, Ravi, S. “Why Skills - and Not Degrees - Will Shape the Future of Work.” *World Economic Forum*, 21 Sept. 2020, <https://www.weforum.org/agenda/2020/09/reckoning-for-skills/>.
- [2] IMPEL Working Group, "Job Market Data Analysis-Key Results," presented to Capstone 1, Northeastern University, Boston, Massachusetts, United States, June 23rd, 2021. [*PowerPoint slides*]. Available: [Powerpoint link](#) , Accessed on: June 23rd, 2021.
- [3] M. Salehi and I. N. Kmalabadi, “A Hybrid Attribute-based Recommender System for E-learning Material Recommendation,” *IERI Procedia*, vol. 2, pp. 565–570, 2012.
- [4] J. Buncle, R. Anane, and M. Nakayama, “A Recommendation Cascade for e-Learning,” *2013 IEEE 27th International Conference on Advanced Information Networking and Applications*, pp. 740–747, 2013.
- [5] E. Mangina and J. Kilbride, "Evaluation of key phrase extraction algorithm and tiling process for a document/resource recommender within e-learning environments", *Computers & Education*, vol. 50, no. 3, pp. 807820, 2008. Available: 10.1016/j.compedu.2006.08.012.