MIE 1624 Introduction to Data Science and Analytics

Course/Program Curriculum Design Project Report

Group 11: Tiago Fernandes Lins, Ramin Mardani, Xinxiu Tian, Shengbo Zhang, Qingqing Zhou, Yisi Zou

1. Background and Objective

Data science education has become increasingly popular among educators and entrepreneurs, leading to increasing demands in data analytics programs and courses. The objective of this project is to design the course curriculum for MIE 1624 course, Technical Data Science Program, Managerial and Business Data Science Program, and an Edtech startup.

2. Course Curriculum Design

In this section, the course curriculum for MIE 1624: Introduction to Data Science and Analytics was re-designed based on the skills required for data analyst/scientist jobs. The Kaggle ML and Data Science Survey data was used because the team decided that salary is an important predictor to determine which data science skills need to be included in the course. Questions 4, 9, 10, 11, 12, 13, 14, 17, 18, 19, 20, 21, 22, 28, 30, 37, and 47 were selected as they provide the most relevant information for this analysis. The team performed preliminary data cleaning to prepare the Kaggle Salary data set for model implementation in the following parts. Irrelevant rows and irrelevant columns were dropped such that only numerical values were considered. Then, missing values in categorical questions were filled with mode and missing values in multiple choices questions were filled with zero. Finally, categorical values were converted to numerical values by using the `get_dummies` module in `pandas` library. The data cleaning resulted in the data set containing 15429 records with 222 features. Then, the correlation function was used to determine the correlation between each feature and the annual compensation. The top 50 features were selected and they were clustered using hierarchical clustering, which is an unsupervised machine learning algorithm that groups similar features. The result is a tree-based representation of the observations which is called a dendrogram. The team experimented with several calculation methods and metrics for determining the distance between clusters when forming the tree, and it has been identified that the "Euclidean" distance metric and "ward" distance calculation method result in the optimal visual observations. The dendrogram is presented in Figure 2.1.

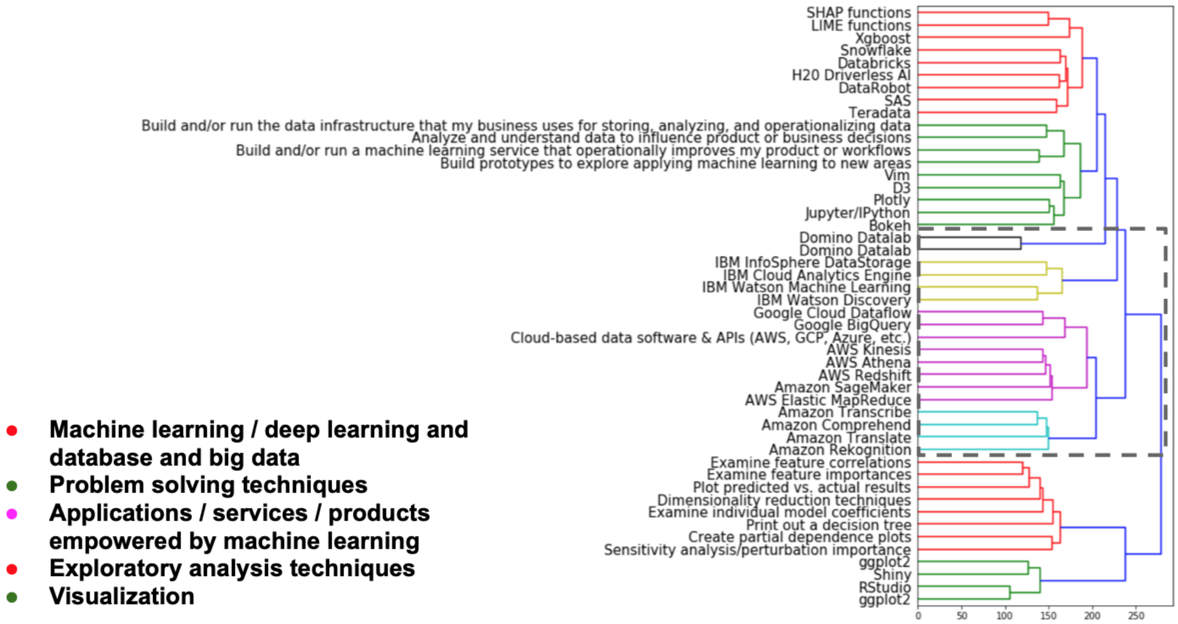


__Figure 2.1__ Dendrogram created with hierarchical clustering of top 50 features that are most correlated with annual compensation.

From the dendrogram, it can be seen that MIE 1624 should cover the following parts: 1) Machine Learning/Deep Learning, 2) Database and Big Data, 3) Problem Solving Techniques, 4) Applications/Services/Products Empowered by Machine Learning Algorithms, 5) Exploratory Analysis Techniques, and 6) Visualization Techniques.

3. Technical Data Science Program Curriculum Design

3.1 Problem Definition

The Technical Data Science Program Curriculum concerns with the technical knowledge required for data scientists. In designing this curriculum, the team analyzed two datasets: 1) the Kaggle survey from "2018 Kaggle ML & DS Survey Challenge" competition and 2) Indeed job postings from "data scientist" search query.

3.2 Data Cleaning

The first dataset used is the Kaggle Salary dataset. Six most informative questions were selected for the curriculum design, including: 1) Q20: ML library most used, 2) Q27: Cloud computing products, 3) Q28: Machine learning products, 4) Q29: Relational database products, 5) Q30: Big data and analysis products, and 6) Q47: Methods preferred for explaining and/or interpreting decisions that are made by ML models. These questions were selected based on not only the top features concluded from Course Curriculum Design, but also the team's insightful discussions on what the most demanding skills in data science industry are. The second dataset was obtained from the Indeed website with a search query of "data scientist". For each job posting, skills were parsed out using keyword matching of skills using a dictionary that contains 131 skill entries. The resultant dataset contains 800 job postings and 112 features (skills) after data cleaning.

3.3 Visualization and Results

Below are the dendrograms for above two datasets and get the following result. For each dendrogram, the features with the same color and under the same hierarchy are close to each other.

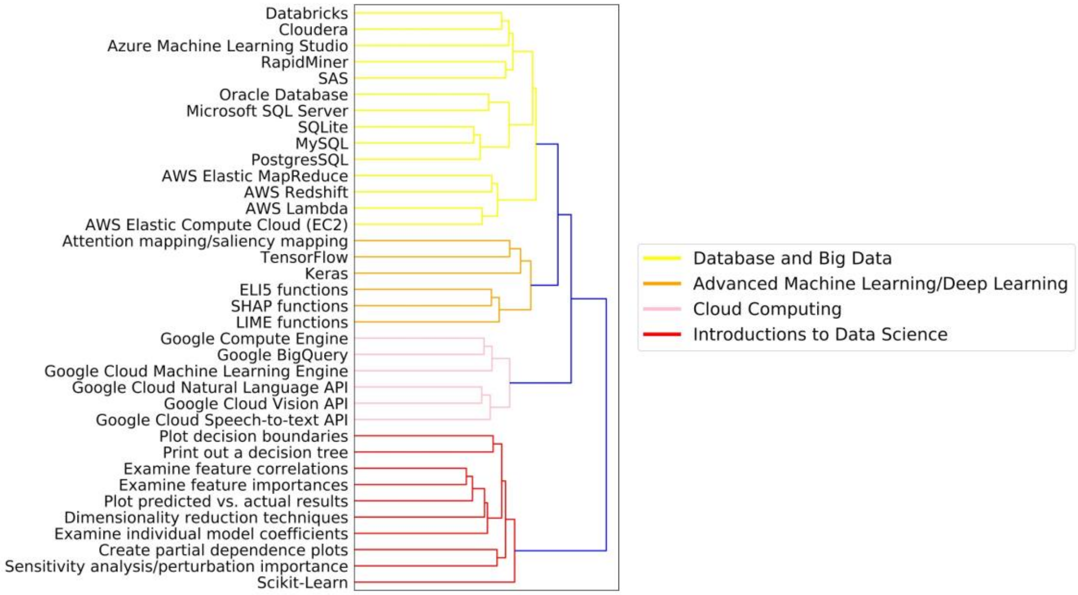


Figure 3.1 *Dendrogram and its classification for technical data science program curriculum design using Kaggle Salary dataset.*

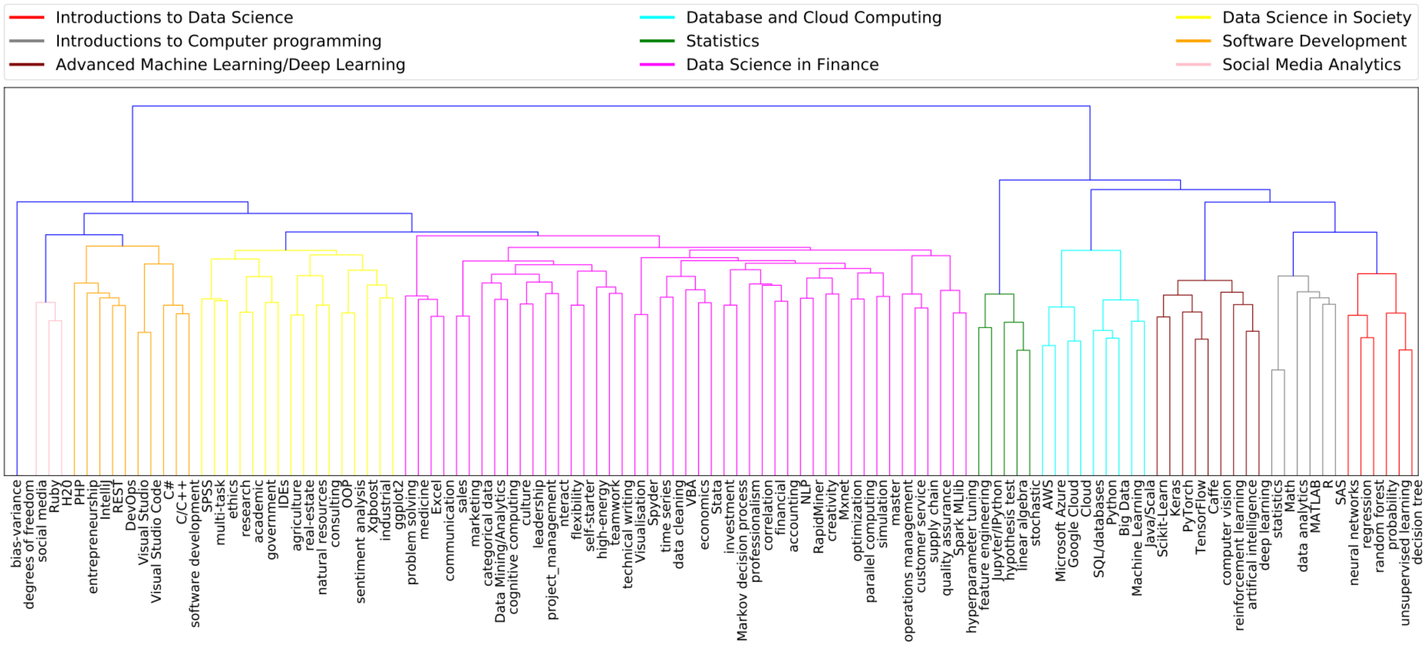


Figure 3.2 *Dendrogram and its classification for technical data science program curriculum design using Indeed web scrapped dataset*

Figure 3.1 shows that the Data Science Program should cover four basic courses, including: 1) Database and Big data, 2) Advanced Machine Learning/Deep Learning, 3) Cloud Computing, and 4) Introduction to Data Science.
Figure 3.2 shows that this program should cover 9 courses, including: 1) Introduction to Data Science, 2) Introduction to computer programming, 3) Advanced Machine Learning/Deep Learning, 4) Database and Cloud Computing, 5) Statistics, 6) Data Science in Finance, 7) Data Science in Society, 8) Software Development, and 9) Social media Analysis.

3.4 Discussion

Overlapping features from both datasets highlights the required courses to cover in this curriculum, and these courses should be offered as required courses for this program, including: 1) Introduction to Data Science, 2) Cloud Computing, 3) Advanced Machine Learning/Deep Learning, and 4) Database. Other courses from the Indeed dataset are tailored for specific industries, thus they can be offered as elective courses. Social media analysis, however, does not have sufficient contents to be covered as a course, thus it can be offered as a design project.

4. Managerial and Business Data Science Program Curriculum Design

4.1 Problem Definition

The managerial and business data science curriculum concerns with data science as well as project management and business skills. In obtaining the dataset, web scraping from Indeed job postings was used with the search query of "project management and business".

4.2 Data Cleaning

Skills were parsed out using keyword matching of skill sets. Skills that appear less than 40 times in our dataset after parsing were removed because of their low prevalence. The dataset contains 848 job postings and 46 features after cleaning.

4.3 Visualization and Results

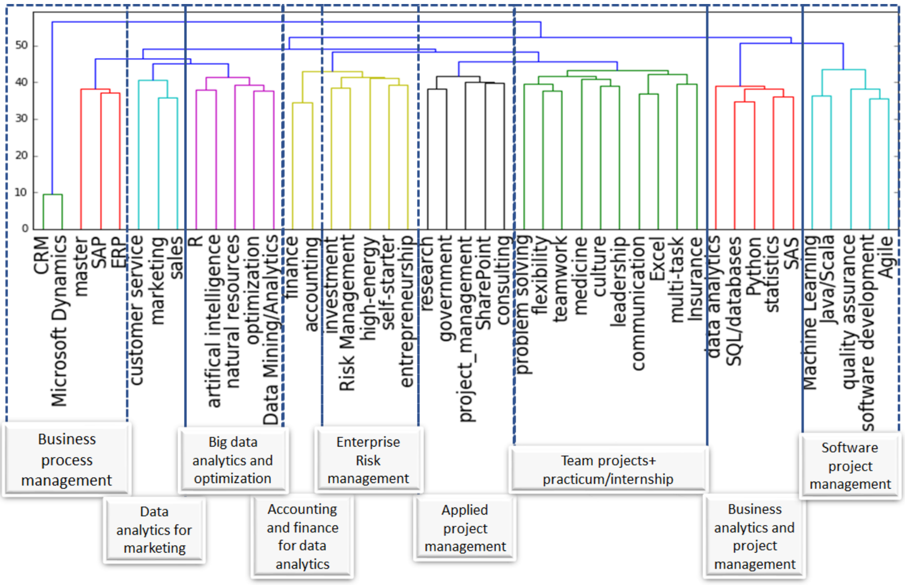


Figure 4.1 *Dendrogram for managerial and business data science program curriculum design*

The skills were clustered into 9 groups using hierarchical clustering and each group is designed as an individual course in this program. Figure 4.2 demonstrates the 9 courses that could be offered in this program, including: 1) Business Process Management, 2) Data Analytics for Marketing, 3) Big Data Analytics and Optimization, 4) Accounting and Finance for Data Analytics, 5)Enterprise Risk Management, 6)Applied Project Management, 7) Team Project and practicum/Internship, 8)Business Analytics and Project Management, and 9) Software Project Management.

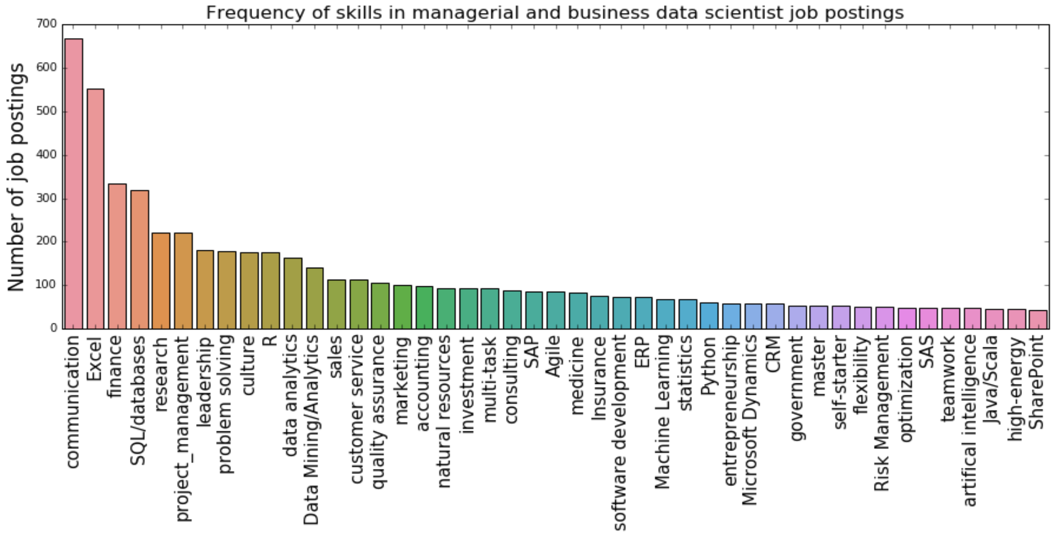


Figure 4.2 *Frequency of skills in managerial and business data scientist job postings*

4.4 Discussion

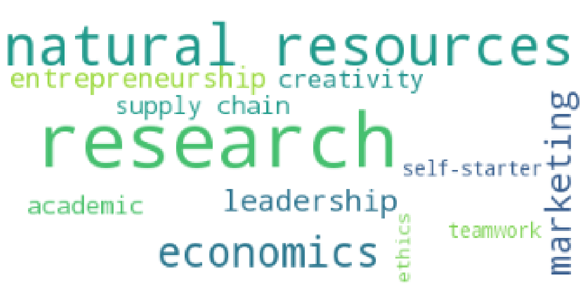
Comparing top 20 frequent skills in Figure 4.2 with courses offered based on Figure 4.1, it can be concluded that only a portion of these courses are relevant to these 20 skills. These courses are thus more important to cover and are offered as required courses, including: 1) Data Analytics for Marketing, 2) Big Data Analytics and Optimization, 3) Accounting and Finance for Data Analytics, 4) Applied Project Management. Team Project/Practicum in Data Analysis and Business Management is offered as a project-oriented course and the other courses can be offered as elective courses.

5. Data Science Education EdTech Effort

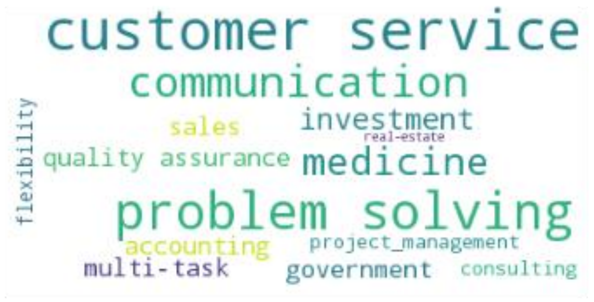
Due to the increasing importance of data science in many niches, it is crucial to design an education institution to prepare students for a smooth entry into the field of data science and analytics. There are various obstacles a person might encounter when they first begin to learn data science skills; For example, a steep learning curve. Moreover, topics often contain many jargons that may confuse students as they do not possess required background knowledge. Above all, examples or projects provided in data science courses are often irrelevant to the area of interest or career goals of the student. Given that data science is becoming increasingly subdivided into different branches, each with its own methods and preferred algorithms, one-size fits all training is no longer appropriate. Therefore, it is crucial to build a matching system that offers courses and projects to a student based on his/her self-described skills and interests. The matching system would simply gather information from the student, and direct him/her into the most appropriate path.

Using data obtained from Indeed web scrapping and Random Forest Regression algorithm, industries that have high demand in data science and data analytics skills can be identified by word clouds:

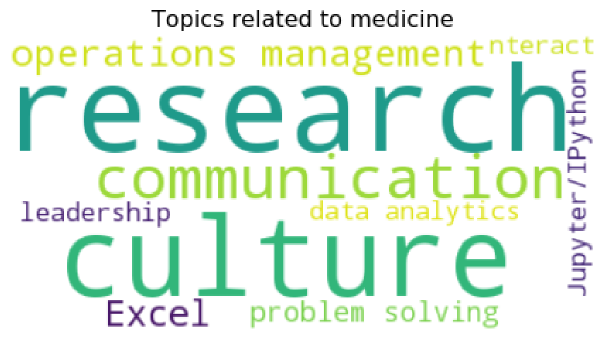
Data Scientist: essential topics

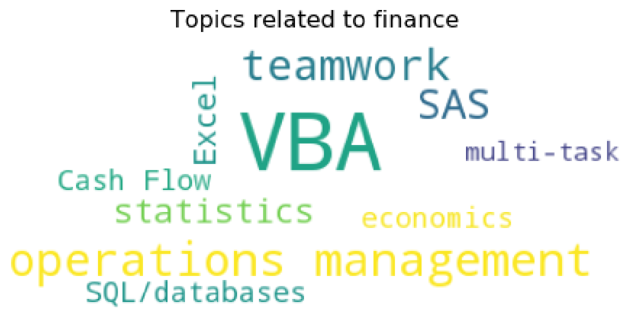


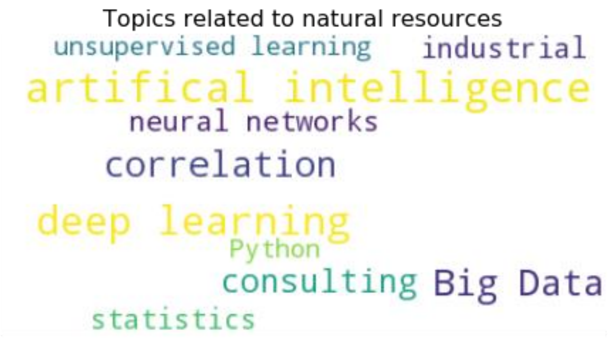
Data Analytics: essential topics

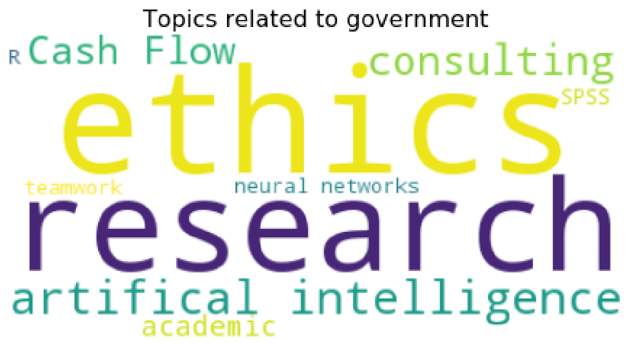


Concluded from above, natural resources and financial industries show high demands for data scientists whereas government agencies and medical industries demonstrate much need for data analysts. Insights in these industries are shown in the following word clouds:









Based on these trends, a curriculum was designed that could better suit the current data science education market. The curriculum contains a series of introductory courses, which a student would take depending on what was recommended to them based on the matching system. Afterwards, the student could also take courses in specialized areas, depending on their interests. Last but not least, the EdTech Enterprise also offers projects from industrial partners, and would proactively seek partnership with companies that are hiring data scientists to understand their workforce needs. Career advisers from companies would be connected to dedicated students, as a way of providing connection and networking opportunities.

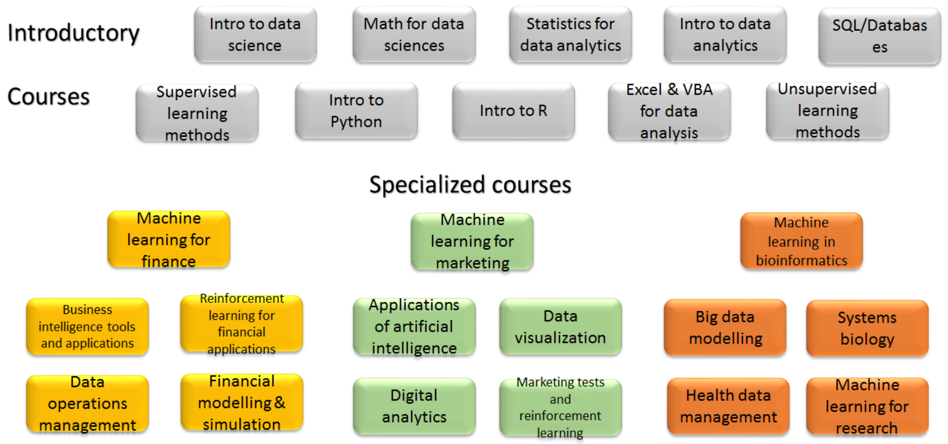

Introductory
Courses
Intro to data science
Math for data sciences
Statistics for data analytics
Intro to data analytics
SQL/Databas es
Supervised learning methods
Intro to Python
Intro to R
Excel & VBA for data analysis
Unsupervised learning methods
Specialized courses
Machine learning for finance
Machine learning for marketing
Machine learning in bioinformatics
Business intelligence tools and applications
Reinforcement learning for financial applications
Applications of artificial intelligence
Data visualization
Big data modelling
Systems biology
Data operations management
Financial modelling & simulation
Digital analytics
Marketing tests and reinforcement learning
Health data management
Machine learning for research

Figure 5.1 *Specialized curriculum offered at EdTech*