Pillars of Analytics Applied in MS Degree in Information Sciences and Technologies

Jai W. Kang
Rochester Institute of Technology
152 Lomb Memorial Drive
Rochester, NY 14623
585-475-5362
jai.kang@rit.edu

Edward P. Holden
Rochester Institute of Technology
152 Lomb Memorial Drive
Rochester, NY 14623
585-475-5361
edward.holden@rit.edu

Qi Yu
Rochester Institute of Technology
152 Lomb Memorial Drive
Rochester, NY 14623
585-475-6929
qi.yu@rit.edu

ABSTRACT
The Master of Science (MS) program in Information Sciences and Technologies (IST) at Rochester Institute of Technology conducted a significant upgrade of its curriculum in 2013, aiming to better prepare its graduates for the new trends and challenges in the fast evolving IT computing industry. In particular, the upgraded MS program places a strong emphasis on data analytics, where all students in the program get an intensive training in data analytics foundation in our core courses. Students can then continue with advanced work in the Analytics Track to receive deeper theoretical knowledge in the field. In this paper, we report our experience of offering this analytics-centric curriculum over the past two years. We first formally define four pillars of analytics and trace the skills needed to support each pillar and the courses that provide those skills. We then describe the course experiences through a sampling of the projects completed by students in their course work. We also provide some student feedback on the course experience. We conclude with a discussion of the capstone experience and a sampling of capstone projects. We show the movement toward analytics in the capstone experiences, particularly since the program began in 2013. The positive course experience and the fast increasing number of capstone projects in the analytics area show strong evidence about the initial success of the analytics-centric curriculum.

Categories and Subject Descriptors
K.3.2 [Computers and Education]: Computer and Information Science Education – curriculum, information systems education.

Keywords
Information sciences and technologies, curriculum, data analytics, database, web technologies.

SIGITE'15, September 30–October 3, 2015, Chicago, IL, USA. © 2015 ACM. ISBN 978-1-4503-3835-6/15/09…$15.00.
DOI: http://dx.doi.org/10.1145/2808006.2808028

1. INTRODUCTION
As mentioned in our last paper [5], the MS in Information Sciences and Technologies (MS/IST) degree continues to move toward being analytics-centric and is an upgrade from our older MS in Information Technology (MS/IT) degree. The paper mentions the number of Google hits on the topics of "Big Data", "Data Science" and "Analytics". This trend continues.

A recent report in Computerworld [8] talks about the 10 hottest technology skills for 2015. Based on Computerworld's 2015 Forecast Survey, the article lists Business Intelligence (BI) / Analytics as number 7 on the list of highly sought after skills with 24% of respondents planning to hire people with these skills within the next 12 months. Respondents expect BI/Analytics skills to be difficult to find. This year's rank is up one from number 8 last year.

In the same article, the closely related skill of Big Data ranks 10th on the list on hot jobs, up from number 11 last year. 20% of survey respondents indicated that they would be hiring Big Data skills in the next 12 months. Career website dice.com reports similar findings in that job listings for Big Data increased by 56% since last year [3]. This places it at number 3 on its list of the fastest growing technology skills.

Dice also reports that other Big Data related job listings have increased since last year: NoSQL 49%, Hadoop 38%, and cloud 34%. These skills rank 4th, 6th and 8th, respectively, on the list of the job skills with the fastest growing demand [3].

The dice article points out that Big Data skills are needed across industries from pharmaceuticals to defense and video games as more data is collected and needs to be analyzed. In the NoSQL area, "professionals who know when to use (and when not to use) these new approaches will bring much-needed flexibility, efficiency and agility to their companies' operations." [3] Many companies are looking for professionals who can combine NoSQL databases with the Hadoop framework to analyze the increasing amounts of data coming into companies.

This paper begins by introducing four pillars of analytics and discusses the skills needed with each one in Section 2. Section 3 reports our experience of offering the analytics-centric MS/IST curriculum. Section 4 describes our students' choice of their capstone project topics before concluding the paper in Section 5.

2. PILLARS OF ANALYTICS
Entry into the MS/IST program requires that the student have prerequisite skills in object-oriented programming, database theory, web concepts and statistics. If these skills are lacking then there are bridge courses available. The key to our analytics-

centric curriculum is what we call the Pillars of Analytics: 1) Data Preprocessing, Storage & Retrieval, 2) Analytical Models & Algorithms, 3) Data Exploration and 4) Data Product as shown in Figure 1. Table 1 shows the Pillars with the skills needed with each one.

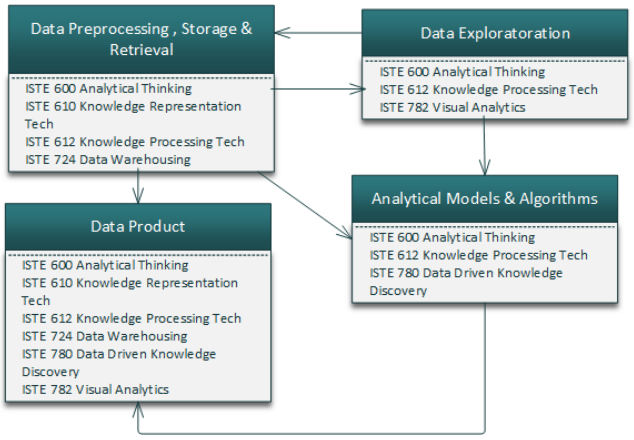

Figure 1. Courses By Pillar of Analytics

Table 1. Skills Required By Pillar of Analytics

Pillars of Analytics	Skills
Data Preprocessing, Storage & Retrieval	NoSQL, Data Modeling, Data Warehousing & Distribution/Parallel Computing
Data Exploration	Statistical Analysis & Visualization
Analytical Models & Algorithms	Machine Learning/Data Mining (ML/DM), Natural Language Processing (NLP) & Information Retrieval (IR)
Data Product	Data and information Organization, Knowledge Representation & Application Development

2.1 Data Preprocessing, Storage & Retrieval

This pillar is covered in ISTE-610 Knowledge Representation Technologies which focuses on non-relational methods of organizing and storing data; and is a major topic studied in ISTE-612 Knowledge Processing Technologies, which primarily focuses on dealing with unstructured data (e.g., text, images, and videos). Key techniques covered in these courses that fall into this pillar include:

- The use non-relational databases to store unstructured data and the organization of that data
- Basic text processing technologies (e.g., tokenization, stop word removal, and stemming/lemmatization) [6]
- Unstructured data modeling (e.g., Boolean model, probabilistic model, and vector space model) [1]
- Efficient storage and fast retrieval of unstructured data (e.g., inverted index, skip pointers, and positional index) [6]

2.2 Analytical Models & Algorithms

These are covered in ISTE-612 with a focus to extract high-level knowledge from large-scale unstructured text. Students are exposed to the widely adopted text analytics techniques and learn how to implement simple yet effective text classification and clustering algorithms, which include

- Naïve Bayes (NB), K-Nearest Neighbor (KNN), and Support Vector Machine (SVM) for classification [6, 2]
- Flat (e.g., K-means) and hierarchical (e.g., single-link, complete-link, and group-average) clustering [6]
- Evaluation of classification and clustering algorithms using standard metrics [2]

ISTE-780 Data Driven Knowledge Discovery, as a key course in the analytics track of the MS-IST program, makes a significant extension of ISTE-600 and ISTE-612 by providing a much deeper coverage on analytical models and algorithms. It provides students with an exposure to the theoretical foundation of key data analytic methods, especially statistical learning approaches, within the context of the data-driven knowledge discovery process. With the explosion of "Big Data" problems, statistical learning has become a very hot field in many scientific areas as well as marketing, finance, and other business disciplines. People with statistical learning skills are in high demand [8]. The main objectives of this course is to

- Model and understand complex datasets using statistical machine learning tools that discover useful information and knowledge from large-scale datasets by conducting both supervised and unsupervised learning [1, 4].
- Scale statistical machine learning algorithms with powerful, distributed, and cloud-based systems (e.g., Apache Hadoop and Mahout) to handle large-scale datasets [9].
- Learn data analytics languages (R and Python) and apply statistical packages (R and scikit-learn) to tackle real-world data analytics problems [4].

2.3 Data Exploration

Students will learn to use their analytical thinking skills developed in ISTE-600 and their data organization skill developed in ISTE-610 combined with software development skills to compute simple statistics (e.g., mean and variance) to perform preliminary data analysis in ISTE-612. They will be trained to use specialized statistical packages (e.g., R and scikit-learn) in advanced courses (e.g., ISTE-780) to perform more sophisticated exploration of data using their visualization and statistical analysis functionalities.

2.4 Data Product

The ISTE-610 and ISTE-612 course projects require students to work in teams to combine their learned knowledge and skills in data organization, information storage, processing, and retrieval as well as text analytics with software development skills over different platforms (e.g., desktop, web, or mobile) to develop a data product. The team should identify the users of this data product, implement its functionalities, justify its unique features, and demonstrate the usefulness through a well-designed user interface.

3. COURSE EXPERIENCES

In this section, we report our experience of offering the analytics centric MS/IST curriculum over the past two years. We will focus on the three core courses of the program, which are required to be

taken by all the students enrolled in the program. We first describe some of the course projects completed by the students from these three courses. As these course projects usually require students to apply what they learn in the class to real-world problems of practical significance, we can use them as an important metric to evaluate how well the students grasp the key skills covered in these courses. Meanwhile, we collected feedback from students to help us better understand whether they are interested in learning the skills covered these courses and whether these skills will benefit their future career.

3.1 Course Project Summary

In the three core course projects, students are organized to work in teams on a problem of their choosing that is interesting, significant, and relevant to apply the eight elements and standards of analytical thinking approaches [5] to solve a data mining problem (ISTE-600); to organize data from an unstructured data source of their choice using XML and a NoSQL database and develop a search application in a language of their choice (ISTE-610); to building an information storage, processing, and retrieval system and/or developing data analytics algorithms for knowledge extraction from unstructured data (ISTE-612). While having great latitude in what a team may choose to work on, the project needs to fulfill a number of requirements:

- It must use some non-trivial data, which could be some sample social media data (e.g., from one of the APIs listed over at Programmable Web), downloaded from an existing collection (e.g., Wikipedia, IMDB), collected using a simple crawler from pre-organized data (e.g., CIA docs, The Simpsons, or other sources)
- It must implement at least one core algorithm that is presented in class or is closely related to the course topic (e.g., supervised & unsupervised classifications, inverted index, vector space model, probabilistic model, text classification, and clustering).
- It is preferred that the project is deployed as either a web app or mobile app (ISTE-612).

Some sample course projects resulted from these classes are presented in the Table 2. For each project, we also present a short description on what this project is about along with the key skills used.

Table 2. Course Projects

Title	Brief Description	Skills Used
WIKI Voyage	An information retrieval system that will allow the user to search Holiday Destinations based on a specified criteria.	NLP & IR
Search and Cook	A system to search various recipes, which are mostly relevant to the user's request.	NLP & IR
Trend Spotter	A web app that provides real-time trends to users based on the location and tweets of other users.	NLP & IR
Spam Rating Detector	A tool to identify star based ratings on BestBuy.com that are inconsistent with the text based reviews for a particular product.	ML/DM, NLP & IR
TV Tweet Analytics	An application that analyzes the comments said on Twitter about a show and provides a rating about this show.	ML/DM, NLP & IR
Recipe Project	Organize recipes for cooking from unstructured data in a form that they can be presented to a user in an organized fashion.	Data Modeling & IR
Customer Complaint Database	Organize data from a customer complaint database and develop an app that will allow users to retrieve and comment on complaints.	Data and information Organization, IR & Application Development
Camera reviews	Organize data from a Amazon camera reviews and develop an app that will allow users to retrieve and comment on products.	Data and information Organization, IR & Application Development
Stack Overflow Questions	Load a database from Stack Overflow information and develop an app that allows the user to select topics and look at further detain on selected entries	Data and information Organization, IR & Application Development
Search application using Yelp dataset	A search application using Yelp data where a user can retrieve Yelp information and add additional comments.	Data and information Organization, IR & Application Development

Table 2. Course Projects (Continued)		
Flight Delay and Cancellation Predictor	This project is to develop a classifier to predict flight delays and cancellations.	ML/DM & Visualization
Prediction of purchasing insurance policies	To predict which insurance policy a customer will choose based on the customer's previous shopping history.	ML/DM & Visualization
Road Accidents in UK	Understanding the factors and other reasons that causes road accidents prevents accidents from happening and thus many injuries can be prevented and many lives can also be saved.	ML/DM & Visualization
Movie reviews sentiment analysis	To analyze the sentiments that can be found in movies reviews and provide a weighted rating to movies.	ML/DM & Visualization
Forest cover type prediction	Cartographic geological data shows a good alternative to identify human disturbances to identify forest cover type.	ML/DM & Visualization

3.2 Student Feedback

We group student feedback into two categories: (1) interestingness: whether the topics are interesting, and (2) usefulness: whether the skills they learn are useful (or can benefit their future career). We provide some representative student feedback for each of these two categories in the following table. The positive feedback clearly indicated that students are keen on the topics covered in the courses. Many of them appreciate the practical aspects and feel that they can build a career on top of the skills they learned.

4. CAPSTONE PROJECTS

After completing required foundation and concentration coursework, students enrolled in the MS/IST at RIT pursue a culminating experience, which allows them to apply their mastery of coursework in either producing a project, involving more in depth research in a thesis, or taking a capstone based course.

As the MS/IST has been evolving to reflect both the computing technological advances and IT industry demands over the past 15 years, students' choice of their capstone topics have also exhibited an evolving pattern, accordingly. The types of capstone projects encompass a wide range of computing areas including Analytics, Application development, Bioinformatics, Database, E-Commerce, Gaming, Geographic Information Systems (GIS), Human Computer Interfaces, Informatics, Management, Multi-Media, Networking, Project Management, Programming, Security, Software Development, Systems Administrations and Web among others.

Table 3. Student Feedback

Interestingness	Usefulness
Applying the concepts learned through a project helped me understand in why they were used. There's nothing like figuring out a problem with a newly acquired set of tools. In addition, it is cool to see how a search engine really works with VSM and the use of an Inverted Index.	The overall topics covered in this course were very knowledgeable and gave me a clear understanding of how a software system can be implemented, evaluated and fine tuned based on the user needs. It gave clarity of how a system functioning can be implemented. It gave a confirmation of what all factors have to be taken in to consideration while handling unstructured data.
I enjoyed learning the common search query algorithm practices. It's an interesting topic	The project was quite interesting and provided us with a lot of exposure to text analytics and recommender systems. This would be a good starting point to build a career in the field of data science.
It was a great learning experience. I was able to witness the power of vector space model and learned various mathematical complexities involved in an information retrieval system. I was also able to study, how to integrate disparate systems via an API.	Learning vector space model, inverted index and handling wildcard queries was really good since it can be used in many different fields too. Understanding it though was a difficult task. Also, the last part about classification and clustering are such useful that they can be later used in our Capstone project. This project can be continued later to develop in many other ways.

There have been over 700 capstone projects produced since the year of 2001. Figure 2 and Table 4 categorize 160 projects by concentrations/tracks that are offered in the current MS/IST degree program.

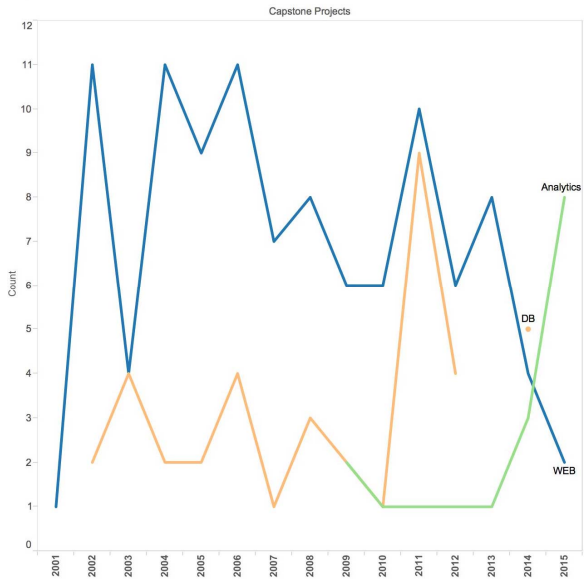

Figure 2. Annual Capstone Projects By Concentration

Table 4. Annual Capstone Projects by Track

	Concentration Tracks			
Year	**Analytics**	**DB**	**WEB**	**Total**
2001			1	1
2002		2	11	13
2003		4	4	8
2004		2	11	13
2005		2	9	11
2006		4	11	15
2007		1	7	8
2008		3	8	11
2009	2	2	6	10
2010	1	1	6	8
2011	1	9	10	20
2012	1	4	6	11
2013	1		8	9
2014	3	5	4	12
2015	8		2	10
Total	17	39	104	160

The three resultant categories include Analytics, Information Management & Database Technology (DB), and Web. As reported in [5], the current master degree program has been designed to be analytic centric and has been offered as of the 2013 academic year.

Table 5. Selective Analytics Capstone Projects

Capstone Projects	Skills Used
Similarity Thesaurus based intelligent search for a job search website	NLP, IR & ML/DM
Evolutionary User Selection to Maximize the Influence of Viral Marketing	NLP & IR
Twitter Data Warehouse for Movie Analytics	NLP, IR, ML/DM & Visualization
Earthquake Tracking & Crowd Sourcing with Geo-Tagged Tweets	NLP & IR
Classifying Forms of Dementia Through the Use of Machine Learning	ML/DM & Visualization
Exploring News Content for Popularity Prediction	ML/DM
Use of Social Media in Promoting Democracy Through Political Campaign and Election Monitoring in Nigeria	NoSQL, NLP, IR & ML/DM
Effective Mining of Twitter Data for Analyzing the Indian General Election	NLP, IR & ML/DM
Mining Unstructured Data to Extract Meaningful Keywords for Large-Scale Data Analysis	NLP, IR, ML/DM & Distribution/Parallel Computing

Even though there were two analytics projects in 2009 and one in each year from 2010 to 2013, the new analytics centric curriculum demonstrates a sharp increase to three and eight analytics projects in 2014 and 2015 respectively. It is also worth noting only five months of data have been collected in 2015. Table 5 lists titles of selective capstone projects in Analytics.

5. CONCLUSIONS

In this paper we present our two years of experience with the new analytics-centric MS Degree Program in Information Sciences and Technologies at Rochester Institute of Technology (RIT). We formally define four pillars of analytics, which include 1) Data Preprocessing, Storage & Retrieval, 2) Data Exploration, 3) Analytical Models & Algorithms, and 4) Data Products. We identify needed skills for each pillar, and use analytics course projects to evaluate how well students grasp the key skills. The positive course feedback demonstrates that students are interested in learning the skills and believe that they can build a career on top of them. When students engage in capstone projects after completing required coursework, this paper discusses the distinct movement toward analytics, particularly since the program began in 2013. The evidence in collected through both course offering and student capstone projects help demonstrate the initial success of the analytics-centric curriculum.

6. ACKNOWLEDGMENTS

The authors would like to thank Ms. Theresa Pozzi, who is Senior Staff Assistant in the Information Sciences and Technologies Department at RIT. Theresa contributed to the paper by helping the authors collect information of students' capstone projects over the past 15 years.

7. REFERENCES

[1] Baeza-Yates, R. and Ribeiro-Neto, B. 2011. *Modern Information Retrieval: The Concepts and Technology behind Search (2nd Edition)*, ACM Press Books.

[2] Bishop, C. 2007. *Pattern Recognition and Machine Learning* (Information Science and Statistics), Springer (October 1, 2007)

[3] Goli, S. *Need for Tech Pros with Analytics Skills Keeps Growing*. http://insights.dice.com/2014/05/05/need-tech-pros-analytics-skills-keeps-growing/, retrieved 5/7/2015

[4] James, G., Witten, D., Hastie, T. and Tibshirani, R. 2014. *An Introduction to Statistical Learning with Applications in R*, Springer.

[5] Kang, J., Holden, E. and Yu, Q. 2014. *Design of an Analytic Centric MS Degree in Information Sciences and Technologies*. In Proceedings of the SIGITE Conference on Information Technology Education (Atlanta, Georgia, USA, October 16-18, 2014). *SIGITE'14*. ACM, New York, NY, 147-152. DOI= http://dx.doi.org/10.1145/2656450.2656460

[6] Manning, C., Raghavan, P. and Schütze, H. 2008. *Introduction to Information Retrieval*, Cambridge University Press.

[7] MS in Information Sciences & Technologies (IST) at RIT Capstone Guide: http://www.ist.rit.edu/assets/pdf/IST%20MS%20Capstone%20Guide.pdf, retrieved 5/7/2015.

[8] Pratt, M. *10 Hottest IT Skills for 2015.* http://www.computerworld.com/article/2844020/10-hottest-it-skills-for-2015.html, retrieved 5/7/2015.

[9] Rajaraman, A. and Ullman, J. 2011. *Mining of Massive Datasets*, Cambridge University Press.