Master of Science in Data Mining
2013 – 2014 Assessment Report

Prepared by Daniel Larose, PhD
Program Coordinator
Department of Mathematical Sciences
School of Engineering, Science, and Technology

Program Rationale

- The Master of Science in Data Mining prepares students to find interesting and useful patterns and trends in large data sets.

- Students are provided with expertise in state-of-the-art data modeling methodologies to prepare them for information-age careers.

Section 1 – Learning Outcomes

Learning Outcomes for Program Graduates

Students in the program will be expected to:

1. approach data mining as a process, by demonstrating competency in the use of CRISP-DM (the Cross-Industry Standard Process for Data Mining), including the business understanding phase, the data understanding phase, the exploratory data analysis phase, the modeling phase, the evaluation phase, and the deployment phase;

2. be proficient with leading data mining software, including IBM/SPSS *Modeler* (formerly *Clementine*), WEKA, Perl, and the R language;

3. understand and apply a wide range of clustering, estimation, prediction, and classification algorithms, including k-means clustering, BIRCH clustering, Kohonen clustering, classification and regression trees, the C4.5 algorithm, logistic Regression, k-nearest neighbor, multiple regression, and neural networks; and

4. understand and apply the most current data mining techniques and applications, such as text mining, mining genomics data, and other current issues.

Program learning outcomes are available online at:
http://web.ccsu.edu/datamining/learning_outcomes.html

Section 2 – Findings

Findings from the Evaluation of Student Learning in the Program

The primary program assessment instrument is the thesis capstone, required of all students. Thus far, 31 students have completed the thesis, as follows.

- *Using Predictive Analysis to Enhance the Efficiency of* C*ommuter Transportation Networks,* by John Ryan Almeida, May, 2014.
- *Implementation of Customer Lifetime Value Model in the Context of Financial Services*, by Alex Bitiukov, May, 2014.
- *Predicting Change in County-Level Presidential Election Voter Turnout Using Data Mining Methods*, by Richard Aceves, May, 2014.
- *Assessment of Similarity-Based Cluster Validation Methods*, by Jill Willie, May 2014.
- *The Application of Decision Trees for Diagnosing Liver Disease*, by Sairam Tadigadapa, May, 2014.
- *Distributor Price Optimization Using Market Segmentation*, by Paolo Carbone, May, 2014.
- *Predictive Modeling of Pay-per-Click Keywords Bid Value*, by Abe Weston, May, 2014.
- *An Application of Gradient Boosted Decision Trees and Random Forests to Prospect Direct Marketing Response Modeling*, by Jeffrey Allard, December, 2013.
- *Improving Workplace Accident Fatality Classification Models with Text Mining and Ensemble Methods*, by Thomas Wilk, Jr., December, 2013.
- *Applying Cost Benefit Analysis to a Trinary Classification Model*, by George DeVarennes, December, 2013.
- *Using Crime Prediction Models to Aid Law Enforcement in Resource Allocation and Decision Making*, by Daniel Aloi, December, 2013.
- *Classifying Web Pages by Image Attributes*, by William E. Rowe, December, 2013.
- Martin Couture (2013), *Applying Data Mining Techniques in Classifying Personal Automobile Insurance Risk*, by Martin Couture, May, 2013.
- Steven Cultrera (2013), *Analysis of the Impact of Weather on Runs Scored in Baseball Games at Fenway Park*, by Steven Cultrera, May, 2013.
- *Applying Misclassification Costs to Ameliorate the False Positive Rate in Bioassay Screening*, by Kay Batta, May, 2013.

- *Measuring Serial Emotional Content in the Enron Email Corpus*, by Scott W. Burk, PhD. December, 2012.

- *Mining for Profitable Low-Risk Delta-Neutral Long Straddle Option Strategies,* by Senthil Murugan, December, 2012.

- *Modeling of Flight Delays*, by Rajiv Sambisavan, November, 2012.

- *Topical Discovery of Web Content*, by Giancarlo Crocetti, October, 2012.

- *Using Data Mining to Model Market Reaction to Corporate Earnings Announcements*, by Judith Gu, July, 2012.

- *Applying Natural Language Processing and Document Classification to Text Mining RSS Feeds in Order to Classify Documents as Interesting or Not, to an Analyst at the Company, Alliant*, by Malcolm Houtz, April, 2012.

- *Anti-Money Laundering Behavior: Reducing the Number of Non-Productive Alerts in Structuring through Effective Data Mining*, by Edwin Rivera, April, 2012.

- *Comparing Classification Algorithms in Data Mining*, by Sampson Adu-Poku, April, 2012.

- *Estimating Potential Customer Value Using Customer Data Using a Classification Technique to Determine Customer Value*, by Thierry Vallaud. April 2009

- *Latent Semantic Analysis and Classification Modeling in Applications for Social Movement Theory*, by Judith E. Spomer. March 2009.

- *Extending the Data Mining Software Packages SAS Enterprise Miner and SPSS Clementine to Handle Fuzzy Cluster Membership: Implementation with Examples*, by Donald K. Wedding, PhD. March 2009.

- *The Discovery by Data Mining of Rogue Equipment in the Manufacture of Semiconductor Devices*, by Steven G. Barbee. April 2007.

- *Identifying Patterns of Potentially Preventable Emergency Department Utilization by American Children*, by Kathleen M. Alber. January 2007.

- *Multivariate Normal Finite Mixture Clustering - An Approach to Distributive Computing and Overcoming Local Optimum Solutions Using Stratified Datasets*, by Eric W. Taylor. April 2005.

- *Netpix: A Method of Feature Selection Leading to Accurate Sentiment-Based Classification Models*, by James B. Steck. April 2005.

- *Knowledge Discovery in Microarray Data*, by Rafiqul Islam. December 2004.

We invite the reader to choose whichever of these theses interests them, download it, and read it. All theses are available from the CCSU Library.

The Learning Outcomes.

During thesis preparation, and before approving the draft, the thesis advisor and the two other faculty members of the thesis advisory committee measure the thesis against the four learning outcomes, as follows:

1. **Adherence to the CRISP-DM standard process**. Data analysis of this level of sophistication requires strict adherence to the industry standard process. In this way, data analysts across the world know what to expect next when reading the thesis. Our students are trained in crystal clear report-writing using CRISP-DM in the core courses Stat 521, Stat 522, and Stat 523.

2. **Proficiency with data mining software**. Each thesis is based on a huge database of thousands of records and tens or hundreds of variables. Students demonstrate in their thesis their ability to use leading data mining software packages, such as IBM/SPSS Modeler (formerly Clementine), Weka, Perl, and R. They have learned one or more of these software packages in Stat 521, Stat 522, Stat 523, Stat 525, Stat 526, Stat 527, and Stat 520.

3. **Expertise in the use of data mining algorithms**. The task of the data miner is to extract useful information from large databases. To accomplish this, the data miner brings to bear a large variety of tools. The thesis demonstrates the student's familiarity with algorithms for clustering, estimation, prediction, and classification. The heart of the student's thesis is the unique modeling techniques he or she applies to his or her challenging research problem. Students become familiar with the panoply of modeling algorithms in Stat 521, Stat 522, Stat 523, Stat 525, Stat 526, Stat 527, and Stat 570.

4. **Familiarity with state-of-the-art methodology**. Not all theses apply the most current data mining techniques and applications, such as text mining, and mining genomics data. However, all students must do well in Stat 526 and Stat 527, which are dedicated to these topics.

Section 3 – Analysis

The data mining faculty has found that the quality of the thesis work has been just as high for those well-prepared in mathematical statistics as for those who are less well-prepared. Further, that many otherwise highly qualified students are doing very poorly in Stat 416 Mathematical Statistics. This finding suggested a mismatch between our target audience, and the program requirements.

Further, prior to 2012, there were only 8 graduates over 11 years of the program's existence. This suggested a need to streamline the program to make it easier for students to complete the program in a timely fashion.

Section 4 –Use of Results

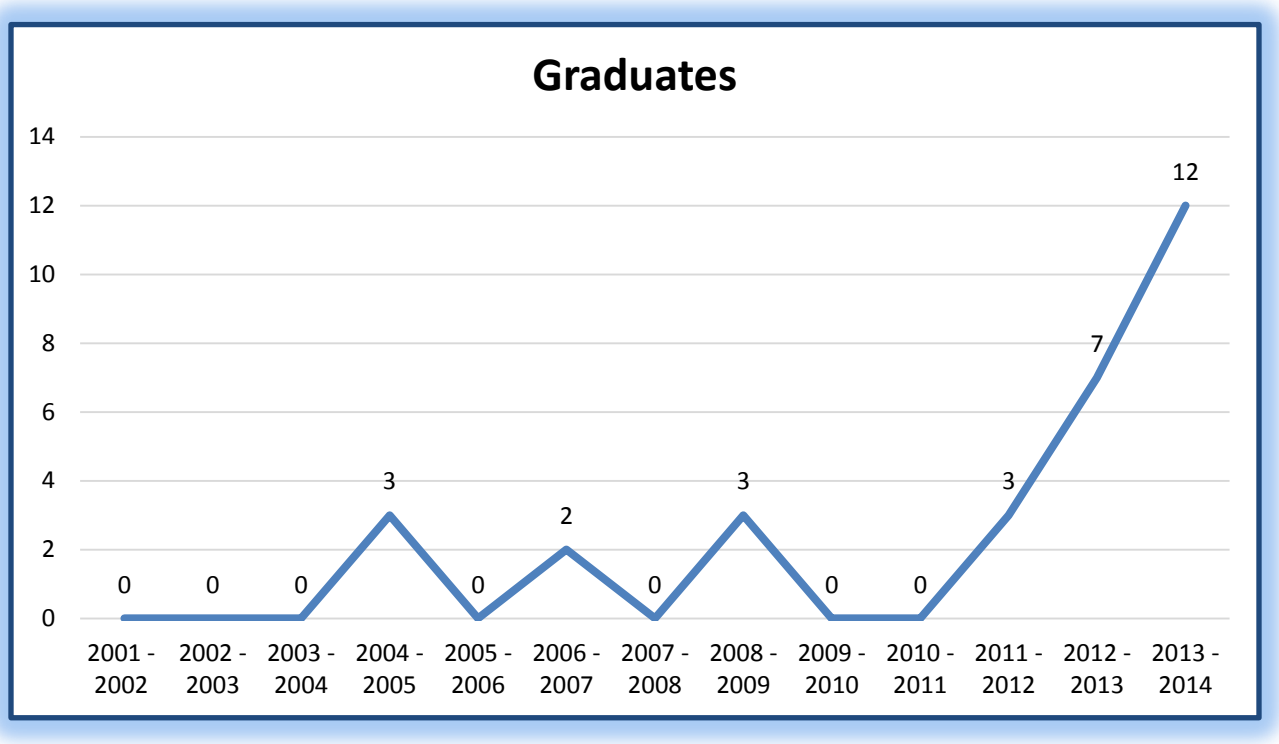
Therefore, we carefully crafted an extensive revision to the Master of Science in data mining. The curricular revision eliminated the mathematical statistics prerequisite and core course, as well as the calculus prerequisite.

Toward the goal of streamlining the program, the revision also:

- Increased the number of credits for Stat 526 *Data Mining for Genomics and Proteomics* and Stat 527 *Text Mining* from 3 credits to 4 credits, to reflect the increased coverage of these growing fields.
- Reduced the number of core courses required, from 8 to 6. Stat 416 *Mathematical Statistics 2* and Stat 525 *Web Mining* were removed from the core.
- Since Stat 416 is no longer required, we eliminated the program prerequisites of Stat 315 and Math 221.
- Kept the two-course elective requirement, to (a) enhance enrollment in some non-data mining courses, and (b) allow for faculty creative development of new courses, such as Stat 534 *Applied Categorical Data Analysis* (Krishna Saha).
- Added a new course to the core: Stat 520 *Multivariate Analysis for Data Mining*. Because Stat 570 *Applied Multivariate Analysis* had the Stat 416 prerequisite, and because these topics are of value and interest to the data mining curriculum, we decided to develop a course which would cover these topics from a conceptual, software-based, assumptions-checking standpoint, rather than from the more formal Stat 570.
- Redesigned Stat 522 (new name: *Clustering and Affinity Analysis*) and Stat 523 (new name: *Predictive Analytics*) so that either may be taken immediately after Stat 521.

The results of these program modifications have been very encouraging. The positive results shown in Figure 1 speak for themselves. The number of Master of Science in Data Mining Graduates has increased over each of the last three years, from zero, to three, to seven to twelve.

Figure 1. Number of Graduates, MS in Data Mining, by Academic Year.



Scholarships

Graduate Academic Award ($750)

2013 - 2104	Andrew Hendrickson
2012 - 2013	Frederick Rountree
2011 - 2012	Rajiv Sambisavan
2010 - 2011	Jeffrey Allard
2009 - 2010	Alexander Bitiukov
2008 - 2009	Thomas Wilk
2007 - 2008	Donald Wedding
2006 - 2007	Senthil Murugan
2005 - 2006	Kathleen Alber
2004 - 2005	James Steck

Larose / Fuller Data Mining Scholarship ($1000 each)

Spring 2014	James Cunningham
Fall 2013	Tarek Abdel-Azim and Ramin Reybod
Spring 2013	George DeVarennes and Philip Hickey
Fall 2012	Eric Flores-Acosta and Jill Willie
Spring 2012	Jeffrey Richardson and Rick Rountree
Fall 2011	Rajiv Sambisavan and Thomas Wilk
Spring 2011	Jeffrey Allard and Malcolm Houtz

Section 5 –General Education

Not Applicable.

Section 6 –Assessment Plan

Future Directions for Assessment Activities

1. With a greater number of students completing their capstone theses, we look forward to an increased opportunity to apply program assessment to implement continuous quality improvement for the Master of Science in data mining.

2. We will monitor the success of our students in the new courses, such as Stat 520, *Multivariate Analysis for Data Mining*, to confirm that the new arrangement is working for them.

3. We will monitor the number of graduates, to make sure that the increase in throughput is not an aberration.