**Syllabus – Data Analytics and Big Data Certificate Program**

**Fall 2015**

Applications and Benefits:

Nearly all businesses collect data about their operations and examine this data for insight into how to improve their operations. As the amount of data that businesses collect becomes increasingly large, insights from the data can no longer be effectively derived manually. There is a growing trend among companies, organizations, and individuals to exploit data mining’s potential to help them discover and act on the most important patterns contained within the data they collect. Data mining has a myriad of business applications and is used increasingly to drive decisions about all aspects of business including spotting sales trends, developing smarter marketing campaigns, accurately predicting customer loyalty, and predicting and protecting against fraud. In fact, data mining can be applied anywhere in a business or organization where a company is interested in identifying and exploiting predictable outcomes. The skills you will learn are applicable to a wide variety of data analytics and big data projects, and will enable you to start working with business problems that require “supervised learning.”

Even if you are not responsible for data mining in your current or future career, the ever-increasing prevalence of data mining in the business world means that you will likely be working with others who are involved in its use. The skills practiced in this course will also help you interact with data mining professionals effectively to collect inputs to make business decisions.

How you will work:

The Data Analytics and Big Data Certificate Program is delivered 100% online, but you will not work alone. You will work (synchronously) with a small team of fellow students supported by an expert mentor on a series of seven projects distributed across four courses. You and your teammates will collaborate on the conceptual and design aspects of each project, but then you must do the hands-on implementation individually.

**Course 1: Understanding Customers**

Overview:

In this course, you’ll learn how to use statistical machine learning techniques to understand the relationship between customer demographics and purchasing behavior and then to develop a model for recommending products to specific customers. You will use the Weka open source machine learning package.

Details:

As a participant in this Story-Centered Curriculum, you will be working under fictional Blackwell Electronics’ Chief Technology Officer Danielle Sherman, as a member of the Blackwell Data Analytics Team. Blackwell Electronics has been a successful consumer electronics retailer in the southeastern United States for over 40 years. Last year, the company launched an eCommerce website. Your job is to use data mining and machine-learning techniques to investigate the patterns in customer sales data and provide insight into customer buying trends and preferences. The inferences you draw from the patterns in the data will help the business make data-driven decisions about sales and marketing activities.

Student Tasks:

1. Investigate customer buying patterns
2. Make cross-selling recommendations
3. Report to the CEO on data mining activities and lessons learned

Performance Objectives:

* Use data mining tools to investigate patterns in complex data sets
* Preprocess data for data mining (e.g., transforming numeric values to nominal values, discretizing data)
* Use decision tree classifiers to investigate classification and regression problems
* Apply cross-validation methods
* Interpret and draw inferences from the results of data mining
* Assess the predictive performance of classifiers by examining key error metrics
* Identify where learning methods fail and gain insight into why with error analysis
* Draw relationships between a machine learning model’s performance and measured features to help understand model performance
* Conduct feature selection to investigate the correlation between different features in a dataset
* Prepare and present data mining results to nontechnical stakeholders.

**Course 2: Predicting Profitability and Customer Preferences**

Overview:

In this course, you will learn to use statistical machine learning techniques to analyze potential new products to recommend the ones a merchant should offer, and also to build a model to predict brand preferences based on customer characteristics. At your option, you can continue to use the Weka machine learning package or switch to the R statistical programming language.

Details:

As a participant in this Story-Centered Curriculum, you will continue to work with Danielle Sherman, the Chief Technology Officer at Blackwell Electronics. Blackwell Electronics is a successful consumer electronics retailer with both bricks & mortar stores in the southeastern United States and an eCommerce site. They have recently begun to leverage the data collected from online and in-store transactions to gain insight into their customers’ purchasing behavior. Your job is to extend their application of data mining methods to develop predictive models. You will use machine learning methods to predict which of the potential new products the sales team is considering adding to the current product mix will be the most profitable. Next, you will create a model to predict which brand of computer products Blackwell customers prefer based on customer demographics collected from a marketing survey. Finally, you will present to management, explaining your insights and suggestions for data mining process improvements.

Student Tasks:

1. Recommend new products
2. Predict which brands of products customers prefer
3. Report to the CEO on data mining activities and lessons learned

Performance Objectives:

* Explore the patterns in the data to create models to predict new events (e.g., the sales volume of a new product)
* Perform similarity analyses
* Deepen experience with decision tree classifiers
* Employ classifiers like k-Nearest Neighbor algorithms and Support Vector Machines to perform regression analyses
* Apply cross-validation methods
* Assess the predictive performance of classifiers by examining key error metrics
* Optimize the initial performance of a classifier by adjusting its parameters
* Interpret the output of a classifier and use that interpretation to select among different classifiers based on their performance characteristics
* Preprocess data for data mining (e.g., applying filters, addressing missing data)
* Use data mining tools and different classifiers to develop predictive models
* Apply machine learning techniques to classification and regression problems
* Apply data mining in ecommerce (e.g. profitability prediction, customer segmentation, product selection strategy)
* Present data mining results to management

**Course 3: Big Data: Web Mining**

Overview:

In this course, you will learn how to mine and analyze extremely large data sets to provide insight to real-world business problems. You will conduct sentiment analysis utilizing cloud-based computing, machine learning tools, and the Common Crawl of the World Wide Web, and interpret the results to make and communicate predictions of vital interest to business stakeholders.

Details:

As a participant in this Story-Centered Curriculum, you will be working as a data analyst for Alert Analytics, a data analytics consulting firm. You will be asked by Alert’s founding partner and SVP Michael Ortiz to take over for a recently-transferred analyst who has been working on a big data project for Helio, a smart phone and tablet app developer. Helio is working with a government health agency to create a suite of smart phone medical apps for use by aid workers in developing countries. The government agency will be providing workers with technical support services, but they need to limit the support to a single model of smart phone and operating system. This will also help to limit purchase costs and ensure uniformity when training aid workers to use the device. After completing an initial investigation, Helio has created a short list of devices that are all capable of executing the app suite’s functions.

To narrow this list down to one device, Helio has engaged Alert Analytics to conduct a broad-based web sentiment analysis to gain insight into users’ attitudes towards the devices. Your job is to conduct this analysis.

Applications and Benefits:

You will learn the algorithmic and organizational skills required to scale data analysis to large server farms, computing clouds, and the web, including an understanding of the design and implementation differences between single-computer and cloud-scale programs, analytics, and data processing. You will also gain a deep knowledge of predictive data analysis, ranging from discovering patterns and correlations in data to making predictions and estimating their accuracy. In this course, you will discern what algorithms will be able to learn from the data most efficiently to answer the information needs, deploy the learning algorithms, interpret the results, and effectively communicate the results of the analysis to non-technical stakeholders. As part of this process, you will master fundamentals of scaling up data analysis to a large cloud computing platform where you will become proficient in working with Map-Reduce-based systems and leverage the computing power of the cloud to prepare very large data sets for deep analysis, as well as learning how to train and apply modern machine learning algorithms to large processed datasets.

Student Tasks:

1. Set up cloud computing environment
2. Prepare the data
3. Develop models to predict sentiment
4. Apply models to data and summarize findings
5. Evaluate information needs for a new project

Performance Objectives:

* Translate business objectives into data mining opportunities
* Acquire, process, and analyze extremely large data sets using data mining methods to discover  patterns or do data exploration
* Install, run, and apply machine learning tools to different kinds of data
* Operate the Amazon Web Services (AWS) cloud computing platform for data analysis
* Discover and mount very large public datasets on the AWS platform
* Setup and run Elastic Map-Reduce (EMR) instances for data analysis
* Formulate a machine learning process to address domain-specific problems
* Perform lexical analysis to extract features from web pages
* Develop and apply machine learning models for sentiment analysis
* Interpret the results of data analysis and data mining to make predictions and to establish the reliability of those predictions
* Avoid pitfalls and mistakes commonly made when applying machine learning methods
* Communicate results to management and other non-technical audiences

**Course 4: Deep Analytics and Visualization**

Overview:

In this course, you will learn how to use the R statistical programming language and a variety of add-on “packages” to visualize data relationships and to implement classification and regression models for emerging engineering applications, such as understanding behavior in the “Internet of Things.”

Details:

In this course, you’ll be working for an "Internet of Things" analytics firm that wants to use Data Analytics to solve two difficult problems in the physical world:

* Smart energy usage: Modeling patterns of energy usage by time of day and day of the year in a typical residence whose electrical system is monitored by multiple sub-meters.
* Indoor locationing: Determining a person’s physical position in a multi-building indoor space using wifi fingerprinting.

You will begin by installing the R statistical programming language and the R Studio work environment. You’ll perform visualizations, and then you will generate descriptive statistics and predictive models using both statistical classifiers and linear regression techniques. Finally, you’ll present the results to the company’s management, explaining strengths and weaknesses of the approaches you implemented and making suggestions for further improvement.

Applications and Benefits:

Increasingly, technology companies are applying data analytics techniques to the masses of data generated by devices such as smart phones, appliances, vehicles, electric meters, *et cetera*. The ability to deal with data of these types will prove to be a high-demand skill for data analysts as applications of commercial interest increasingly go beyond business intelligence. The skills you will learn are applicable to a wide variety of data analytics projects and will enable you to start working on problems that benefit from the application of machine learning and statistical programming techniques to sensor (and other) data.

Student Tasks:

1. Get started with R
2. Define a data science process
3. Visualize and analyze energy consumption data
4. Evaluate techniques for wifi locationing

Performance Objectives:

* Defining the business purpose of a data analytics project and making a principled, realistic analysis plan
* Installing and using the R statistical programming language and R Studio
* Managing data in R
* Exploring data using visualization techniques and descriptive statistics in R
* Choosing and evaluating classification modeling techniques in R
* Choosing and evaluating regression modeling techniques in R
* Analyzing time series data
* Performing error analyses
* Interpreting performance metrics
* Presenting highly technical, engineering-oriented data mining results to a business audience.

**Program Technical Requirements**

* PC or Mac desktop or laptop computer
* Intel or compatible processor
* 100 GB hard drive (minimum)
* Operating System: Linux, Mac OS (Intel-based), or Windows
* High-speed internet connection
* Current version of Microsoft Office or Microsoft Office for Mac

In addition, you will need to open an Amazon Web Services account for the Big Data course.