Date: 8/25/2015

Institutional action: Graduate Certificate

Graduate Certificate Title: Graduate Certificate in Data Analytics

CIP Code Number & Title: 11.0301 ("Data Processing and Data Processing Technology/Technician)

VT Offer Code: DAC

Starting Term, Year: Spring 2016 First Award Term, Year: Spring 2016

Description:

Description: The purpose of this certificate is to prepare students for technical careers in big data analytics and data science. Students will acquire in-depth technical skills that will enable them to understand the underlying technical fundamentals of data analytics, to develop new analytical methods, and to engineer new analytical tools. Students will acquire skills that integrate computational, statistical, and engineering techniques that form the heart of big data analytics. The certificate will provide students with formal recognition of their skills to better support their career prospects.

There is a growing **need** for technically trained engineers and scientists to lead the rapidly evolving field of big data analytics. The U.S. presidential administration has identified big data analytics as a core area of national need. Data science is one of the fastest growing career paths, and demand for technical expertise is out-pacing supply. Technical expertise is needed to develop new methods, tools, and infrastructures required to support novel big data analytics operations in industry, government, and academia. The technical expertise required involves a combination of computation, statistics, and engineering, such that training in any one of these individual disciplines alone does not suffice. This certificate will serve to train technical students with a broader view across these disciplines to support the data analytics field.

The **learning outcomes** of this certificate program are as follows:

- Students will have technical depth in the fundamentals of data analytics, in terms of understanding the underlying principles and implementations of analytical methods.
- Students will have broad understanding of multi-disciplinary perspectives on technical methods in data analytics, including computational, statistical, and engineering perspectives.

Target Audience and Time to Complete:

The target audience of this certificate is technically oriented students in engineering and science. In particular, the certificate is ideally suited to complement the technical training of students enrolled in Virginia Tech's graduate programs in Computer Science, Statistics, and Electrical and Computer Engineering. Since the certificate requirements fit well with these existing degree program requirements, it is expected that the time to completion of the certificate will not substantially increase their time to completion for their degree program. Per university

requirements, at most 6 of the required 12 credits for the certificate can be double counted towards their degree program, meaning that students will need to take at least two additional courses beyond their existing degree requirements. However, students in other graduate programs at VT are not precluded. The estimated time to completion for students in other degree programs and for non-degree seeking participants is one year.

Admission:
Admission to the Graduate School and completing a Graduate Certificate Application are required for both degree- and non-degree seeking students.

Degree-seeking applicants:
The Graduate School requires completion of a bachelor's degree from an accredited institution with a GPA of 3.0 or better for admission to Certificate Status. Applicants with an undergraduate GPA < 3.0 may qualify for Commonwealth Campus admission. Students pursuing a degree and a certificate simultaneously are classified within their degree program. Certificate credits may be used to meet degree requirements if they are appropriate for inclusion on the degree Plan of Study.

Non-degree seeking applicants:
A qualified person who wishes to enter Virginia Tech to obtain a graduate certificate, without being enrolled in a degree program, may apply for graduate admission to Graduate Certificate status. Such applicants submit an Application for Admission and a Graduate Certificate Application http://graduateschool.vt.edu/forms/academics/certificate_application.pdf, and must meet the following criteria:

- GPA of 3.0 for admission for the last half of the credits earned for the undergraduate (bachelors) degree*
- official transcripts must be submitted.
- academic background meets the requirements of the admitting academic unit.
- International applicants must submit scores from the Test of English as a Foreign Language (TOEFL) or the International English Language Testing System (IELTS). A minimum TOEFL score of 550 paper-based (PBT) or 80 internet-based test (iBT) is required for consideration of the application. On the iBT, subscores of at least 20 on each subtest (Listening, Speaking, Reading, and Writing) are required for admission. A minimum IELTS score of 6.5 is required for admission. Some departments have higher TOEFL or IELTS score requirements than those set by the Graduate School.

Curriculum Requirements and Descriptions:

Number of Credit Hours:
Students should complete at least 2 courses from the core list (see below) and 2 courses from the elective list, for a total of 12 credits. For all students, courses taken must span all three departments; Computer Science, Statistics and Electrical and Computer Engineering. All

courses must be graded A-F, and students must attain a minimum 3.0 GPA in the designated courses. Transfer credits are not permitted.

Core Courses: (Choose 2)

CS/STAT 5525 Data Analytics I
Basic techniques in data analytics including the preparation and manipulation of data for analysis and the creation of data files from multiple and dissimilar sources. The data mining and knowledge discovery process. Overview of data mining algorithms in classification, clustering, association analysis, probabilistic modeling, and matrix decompositions. Detailed study of classification methods including tree-based methods, Bayesian methods, logistic regression, ensemble, bagging and boosting methods, neural network methods, use of support vectors and Bayesian networks. Detailed study of clustering methods including k-means, hierarchical and self-organizing map methods. Prerequisites: none. (3H, 3C)

CS/STAT 5526 Data Analytics II
Techniques in unsupervised and visualized learning in high dimension spaces. Theoretical, probabilistic, and applied aspects of data analytics. Methods include generalized linear models in high dimensional spaces, regularization, lasso and related methods, principal component regression (pca), tree methods, and random forests. Clustering methods including k-means, hierarchical clustering, biclustering, and model-based clustering will be throughly examined. Distance-based learning methods include multi dimensional scaling, the self organizing map, graphical/network models, and isomap. Supervised learning will consist of discriminant analyses, supervised pca, support vector machines, and kernel methods. Prerequisites: CS/STAT 5525. (3H, 3C)

CS 5824/ECE 5424G: Advanced Machine Learning
Algorithms and principles involved in machine learning; focus on perception problems arising in computer vision, natural language processing and robotics; fundamentals of representing uncertainty, learning from data, supervised learning, ensemble methods, unsupervised learning, structured models, learning theory and reinforcement learning; design and analysis of machine perception systems; design and implementation of a technical project applied to real-world datasets (images, text, robotics). Prerequisites: none. (3H, 3C)

Restricted Elective Courses: (Choose 2)

CS 5234 Advanced Parallel Computation
Survey of leading high-end computing systems and their programming environments. Advanced models of parallel computation. Mapping of parallel algorithms to architectures. Performance programming and tools for performance optimization on parallel systems. Execution environments and system software for large-scale parallel computing. Case studies of parallel applications. Prerequisites: none. (3H, 3C)

CS 5604 Information Storage and Retrieval

Analyzing, indexing, representing, storing, searching, retrieving, processing and presenting information and documents using fully automatic systems. The information may be in the form of text, hypertext, multimedia, or hypermedia. The systems are based on various models, e.g., Boolean logic, fuzzy logic, probability theory, etc., and they are implemented using inverted files, relational thesauri, special hardware, and other approaches. Evaluation of the systems' efficiency and effectiveness. Prerequisites: none. (3H, 3C)

CS 5614 Database Management Systems
Emphasizes concepts, data models, mechanisms, and language aspects concerned with the definition, organization, and manipulation of data at a logical level. Concentrates on relational model, along with introduction to design of relational systems using Entity-relationship modeling. Functional dependencies and normalization of relations. Query languages, relational algebra, Datalog, and SQL. Query processing, logic and databases, physical database tuning. Concurrency control, OLTP, active and rule-based elements. Data Warehousing, OLAP. Prerequisites: none. (3H, 3C)

CS 5764 Information Visualization
Examine computer-based strategies for interactive visual presentation of information that enable people to explore, discover, and learn from vast quantities of data. Learn to analyze, design, develop, and evaluate new visualizations and tools. Discuss design principles, interaction strategies, information types, and experimental results. Research-oriented course surveys current literature, and group projects contribute to the state of the art. Prerequisites: none. (3H, 3C)

CS 5804 Introduction to Artificial Intelligence
A graduate level overview of the areas of knowledge representation, machine vision, natural language processing, search, logic and deduction, problem solving, planning, and robotics. Prerequisites: none. (3H, 3C)

CS 6604 Advanced Topics in Data and Information
This course treats a specific advanced topic of current research interest in the area of data and information. Papers from the current literature or research monographs are likely to be used instead of a textbook. Student participation in a seminar style format may be expected. Prerequisites: 5604 or 5614. (3H, 3C)

STAT 5114 Statistical Inference
Decision theoretic formulation of statistical inference, concept and methods of point and confidence set estimation, notion and theory of hypothesis testing, relation between confidence set estimation and hypothesis testing. Prerequisites: none. Co: 5104. (3H, 3C)

STAT 5314 Monte Carlo Methods in Statistics
Theoretical and applied aspects of simulation-based sampling methodology. Monte Carlo integration, importance sampling, Markov chain Monte Carlo, particle methods, Kalman filtering. Programming in Matlab, R, or SAS. Prerequisites: none. (3H, 3C)

STAT 5414 Time Series Analysis I
Analysis of serially dependent data -, including stationary and nonstationary time series, Box-Jenkins modeling, trend elimination, prediction, unit root testing, intervention analysis, transfer function models, and applications in economics and engineering. Prerequisites: STAT 5114. (3H, 3C)

STAT 5444 Bayesian Statistics
Introductory course of Bayesian statistics on basic concepts of probability, Bayesian inference of Normal, Binomial, Poisson, Uniform and other common distributions, selections of prior information, Bayesian decision theory, Bayesian analysis of regression and analysis of variance and Bayesian foundation. Prerequisites: STAT 5114. (3H, 3C)

STAT 5444G Advanced Applied Bayesian Statistics
Bayesian methodology with emphasis on applied statistical problems: data displaying, prior distribution elicitation, posterior analysis, models for proportions, means and regression. Prerequisites: none. (3H, 3C)

STAT 5504 Multivariate Statistical Methods
Methods useful for description and inference for multivariate data. Multivariate distributions, location and dispersion problems for one and two samples, multivariate analysis of variance, linear models, repeated measurements, principal components, factor analysis, biplots, discriminant and canonical analysis, cluster analysis, multidimensional scaling and correspondence analysis. Uses SAS or R. Prerequisites: (5104 or 5616), MATH 5524. (3H, 3C)

STAT 5544 Spatial Statistics
Spatial data structures: geostatistical data, lattices, and point patterns. Stationary and isotropic random fields. Autocorrelated data structures. Semivariogram estimation and spatial prediction for geostatistical data. Mapped and sampled point patterns. Regular, completely random, and clustered point processes. Spatial regression and neighborhood analyses for data on lattices. Prerequisites: STAT 5124. (3H, 3C)

ECE 5524 Pattern Recognition
Computational methods for the identification and classification of objects. Feature extraction, feature-space representation, distance and similarity measures, decision rules. Supervised and unsupervised learning. Statistical pattern recognition: multivariate random variables; Bayes and minimum-risk decision theory; probability or error; feature reduction and principal components analysis; parametric and nonparametric methods; clustering; hierarchical systems. Syntactic pattern recognition: review of automata and language theory; shape descriptors; syntactic recognition systems; grammatical inference and learning. Artificial neural networks as recognition systems. Prerequisites: none. (3H, 3C)

ECE 5554 Computer Vision
Techniques for automated analysis of images and videos. Image formation, feature detection, segmentation, multiple view geometry, recognition, and video processing. Prerequisites: none. (3H, 3C)

ECE 5606 Stochastic Signals and Systems
Response of continuous and discrete time, linear and nonlinear systems to Gaussian and non-Gaussian random processes. Signal to noise power ratio computations (SNR) of systems. Introduction to signal detection theory. Optimal filtering (estimation) techniques of Wiener and Kalman to both open and closed loop systems. Prerequisites: none. (3H, 3C)

ECE 5734 Convex Optimization
Recognizing and solving convex optimization problems. Convex sets, functions, and optimization problems. Least-squares, linear, and quadratic optimization. Geometric and semidefinite programming. Vector optimization. Duality theory. Convex relaxations. Approximation, fitting, and statistical estimation. Geometric problems. Control and trajectory planning. Prerequisites: none. (3H, 3C)

ECE 6504 Deep Learning for Perception
Advanced topics of current interest in computer engineering which are taken from current research topics and/or technical publications. Prerequisites: none. (3H, 3C)

ECE 6554 Advanced Computer Vision
Current and state-of-the-art trends in computer vision, particularly in object recognition and scene understanding. Application of approaches in computer vision to various automatic perception problems. Strengths and weaknesses of computer vision techniques. Open questions and future research directions. Prerequisites: ECE 5554. (3H, 3C)

CS 6424/ECE 6424 Probabilistic Graphical Models and Structured Prediction
Advanced concepts in machine learning; focus on probabilistic graphical models and structured output prediction. topics include directed models (Bayes Nets), undirected models (Markov/Conditional Random Fields), exact inference (junction tree), approximate inference (belief propagation, dual decomposition), parameter learning (MLE, MAP< EM, max-margin), structure learning. Prerequisites: 5824 or ECE 5424G. (3H, 3C)

Faculty Credentialing:
The graduate certificate will be managed primarily within the Department of Computer Science, with cooperation of the Department of Statistics and Department of Electrical and Computer Engineering, at Virginia Tech. All involved instructional faculty have doctoral degrees in related fields. The certificate will be administered by the Discovery Analytics Center (DAC). Each academic year, a DAC faculty member from the list below will be designated as the certificate administrator for the year.

Affiliated Faculty:
- Dr. Naren Ramakrishnan, Professor, Department of Computer Science
- Dr. Chris North, Professor, Department of Computer Science
- Chang-Tien Lu, Associate Professor, Department of Computer Science
- Dr. Aditya Prakash, Assistant Professor, Department of Computer Science

- Dr. Bert Huang, Assistant Professor, Department of Computer Science
- Dr. Scotland Leman, Associate Professor, Department of Statistics
- Dr. Leanna House, Associate Professor, Department of Statistics
- Dr. Dhruv Batra, Assistant Professor, Department of Electrical and Computer Engineering
- Dr. Devi Parikh, Assistant Professor, Department of Electrical and Computer Engineering

Course Delivery Format:
Most courses are classroom-based, located on the Virginia Tech campus in Blacksburg and in the National Capital Region.

Some courses are delivered via distance learning. Virginia Tech has advanced infrastructure and active support for online curricular delivery though Technology-enhanced Learning and Online Strategies (TLOS; http://tlos.vt.edu/).

Some courses are delivered online. Virginia Tech has advanced infrastructure and active support for online curricular delivery though Technology-enhanced Learning and Online Strategies (TLOS; http://tlos.vt.edu/).

Resources:
Virginia Tech has the resources required to offer and sustain this certificate program. These include such resources as student support services (e.g., enrollment, help desk, library, etc.); faculty support services (e.g., copying, contracts, etc.); and general administration (e.g., budgeting and forecasting, etc.). All courses in the certificate program are already existing courses that are taught regularly. A faculty member from the Discovery Analytics Center will administer the program requirements, which is not expected to be a significant burden.