

**COVER SHEET - NEW and REVISED COURSES**

Commission on Undergraduate Studies and Policies/ Commission on Graduate Studies and Policies/ University Core Curriculum Committee  
Effective August 1993

•SEE I - VII for Basic Course Proposal Guidelines•  
•SEE APPENDIX FOR NOTES, EXPLANATIONS AND ADDITIONAL GUIDELINES•  
•PRINT CLEARLY, TYPE or COMPLETE ELECTRONICALLY•

**APPROVED**  
CGC 3-24-11  
CGSP 4-6-11

PROPOSAL DATE: October 15, 2010 DEPARTMENT: Statistics

COURSE DESIGNATOR AND NUMBER: STAT (CS) 5525-5526

TITLE OF COURSE: Data Analytics

TRANSCRIPT (ADP) TITLE (MAX-30 Characters): Data Analytics

INSTRUCTOR and/or DEPARTMENTAL CONTACT: Susan Haymore CONTACT MAILCODE: 0408

CONTACT PHONE: 1-1181 CONTACT E-MAIL: shaymore@vt.edu

CHECK IF GRADUATE CREDIT IS REQUESTED (15 copies required for CGSP)

CHECK ONLY ONE OF THE FOLLOWING BOXES

NEW COURSE  REVISED COURSE [Revision > 20% \_\_\_\_\_ Revision < 20% \_\_\_\_\_]

NEW COURSE & INCLUSION IN THE CORE [Area \_\_\_\_\_]  OTHER: Cross-list course  
Include Attachment, If Needed

REVISED COURSE FOR INCLUSION IN THE CORE OR CORE AREA CHANGE

•Courses routed directly to the University Core Committee MUST be endorsed by the appropriate Department Head or Dean.  
•The Chair of the University Core Committee shall inform the appropriate college curriculum committee of all courses under review by the core committee.

•A Attach Statement from Dean or Departmental Representative as to whether Teaching this Course will Require or Generate the Need for Additional Departmental Resources.

•B Attach Appropriate Letters of Support from Affected Departments and/or Colleges.

•C Effective Semester: Spring 2012

•D Change in Title From: \_\_\_\_\_  
To: \_\_\_\_\_

•E Change in Lecture and/or Lab Hours From: \_\_\_\_\_ To: \_\_\_\_\_

•F Change in Credit Hours From: \_\_\_\_\_ To: \_\_\_\_\_

•G Percentage of Revision from Current Syllabus: \_\_\_\_\_ Revision Summary: \_\_\_\_\_

•H Course Number(s) and Title(s) to be Deleted from the Catalogue with APPROVAL of course:

CS 5624 Introduction to Data Mining

APPROVAL SIGNATURES

Department Representative Sign Williams Date: 10/15/2010

College Curriculum Committee Representative \_\_\_\_\_ Date: 02/09/2011

College Dean Joe C. Chee Date: 2/11/11

To: Graduate Curriculum Committee

Re: Cross-Listing of CS & STAT course on Data Analytics

5525-5526

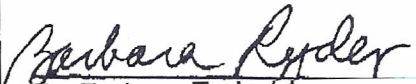
The Computer Science and Statistics departments request joint ownership of the ~~STAT 5505-5506/CS 5625-5626~~ Data Analytics sequence. Per the Graduate Curriculum Committee *Reference Guide to the Graduate Course and Certificate Proposal Development and Review Process (Rev. October 2010)*, it is understood that one department may continue to offer the sequence if the other drops the sequence and that any changes to either course in the sequence via university governance will require the support of both departments.




Dr. Eric P. Smith  
Chair, Department of Statistics



Dr. Jill Sible  
Associate Dean, College of Science



Dr. Barbara Ryder  
Chair, Department of Computer Science



Dr. Don Lee  
Associate Dean, College of Engineering

*Invent the Future*

**DATA ANALYTICS  
STAT(CS) 5525-5526**

***I -- Catalogue Description***

5525: 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. Pre: Graduate Standing. (3H, 3C).

5526: Techniques in supervised, unsupervised, and visualized learning in high dimensional 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 thoroughly 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. Pre: 5525. (3H, 3C).

**Course Number:** 5525-5526

**ADP TITLE:** Data Analytics

***II - Learning Objectives***

Having successfully completed this course, students will be able to:

- Explain the basic framework of data analytics techniques to solve real-life problems.
- Use and interpret various statistical methods to find meaningful structures in high dimensional data.
- Use statistical software and algorithms to write programs to implement different data analytics techniques.

### ***III - Justification***

Recent advances in database technology and the phenomenal growth of the Internet have resulted in an explosion of data collected, stored, and disseminated by various organizations. The massive sizes of many datasets preclude their manual analysis. Furthermore, modern statistical techniques to analyze such datasets tend to be highly algorithmically and computationally oriented. Abilities in numerical methods, computer programming, and algorithm construction and utilization are essential for the contemporary statistician. The focus in this course is on contemporary statistical methodologies that are both algorithmically and computationally oriented and especially useful for analysis of high dimensional data (data with both a large number of observations and a large number of variables).

STAT(CS) 5526 builds upon the methods taught in STAT(CS) 5525 by expanding the base of statistical theory, by providing a greater connection between student-developed programs and current available statistical software, and extending the methods to include those based on clustering, distance learning, and supervised learning.

Graduate standing is required for this course sequence as both courses require a breadth of technical knowledge in algorithmic and statistical methodologies that is found only in students who have completed an undergraduate degree in computer science, statistics, electrical and computer engineering, or mathematics.

### ***IV - Prerequisites and Corequisites***

Graduate standing.

### ***V - Texts and Special Teaching Aids***

#### A. Required Text:

5525: Tan, P.N, Steinbach, M. & Kumar, V., INTRODUCTION TO DATA MINING, Boston MA: Addison Wesley, 2005, 769.

5526: Hastie, T., Tibshirani, R., & Friedman, J. H., THE ELEMENTS OF STATISTICAL LEARNING, second edition, Stanford California: Springer. 2009. 745.

#### B. Recommended Text:

5526: Bishop, C., PATTERN RECOGNITION AND MACHINE LEARNING, (2007), New York NY: Springer, 2007, 740.

## VI - Syllabus

5525:

1. Introduction	5%
2. Data mining and knowledge discovery methodology	10%
3. Classification	
a. Tree based	10%
b. Bayes classification	5%
c. Linear discriminant and Bayes classification	5%
d. Comparing models using ROC	5%
e. Ensembles, bagging and boosting	5%
f. Support vector machines	10%
g. Bayesian networks	5%
4. Clustering methods	
a. k-means	10%
b. hierarchical	10%
5. Association analysis, text analysis	10%
6. Data visualization and management	5%
7. Matrix decompositions	5%
	<hr/>
	100%

5526:

1. Generalized Linear Models	10%
a. Logistic Regression	
b. Probit Regression	
c. Categorical and Ordinal Regression	
2. High Dimensional Analysis	10%
a. Regularization	
b. The Lasso	
3. Principal Components	10%
a. Projections and Projectors	
b. Principal Component Analysis, Clustering, and Regression	
c. Probabilistic PCA	
4. Tree Methods	15%
a. CART	
b. Random Forests	
c. Neighbor Joining Trees	
5. Heuristic Clustering	15%
a. K means	
b. Agglomerative methods	
c. Hierarchical clustering	

- d. Biclustering
- e. Model based clustering
- f. The Generative Topographical Mapping
- 6. Probabilistic Clustering 10%
  - a. Model based clustering
  - b. The Generative Topographical Mapping
- 7. Distance Methods 15%
  - a. Multi Dimensional Scaling
    - i. Relationship to PCA
  - b. The Self Organizing Map
  - c. Graphical Models
  - d. The Isomap
- 8. Supervised Learning 15%
  - a. Discriminant Analysis
  - b. Naïve Bayes
  - c. Supervised PCA
  - d. Support Vector Machines
  - e. Kernel Methods

---

100%