IOM 528 – DATA WAREHOUSING, BUSINESS INTELLIGENCE AND DATA MINING -16274
(Wednesday 6:00-10:00 p.m. in JKP112) and three Saturdays (6/9/12, 6/16/12 and 6/30/12 - 10:00-2:00 p.m.)

Instructor: Dr. Arif Ansari
Office: BRI 401 R (Bridge Hall 401R)
Office Hours: Wednesday 5:00-6:00 p.m. and by appointment
Office phone: (213) 821-5521
Email: aansari@usc.edu
Emergency Contact number: 213-740-0172

http://www.youtube.com/watch?v=PD8w24vSaYA&list=PL0FD8B7FC060A7708&index=10&feature=plpp_video

COURSE OBJECTIVES

- To develop an understanding of the various concepts and tools behind data warehousing and mining data for business intelligence (BUSINESS ANALYTICS) -
- To develop quantitative skills pertinent to the analysis of data from huge corporate data warehouses
- To develop industry level data mining skills using SAS enterprise Miner (optional) and desktop level data mining skills using SAS JMP software
- To develop Data warehouse architecture from Business point of view
- To develop new business strategies based on DW, BI and DM
- To develop Critical thinking and Modeling Skill Set.
  - **This year I will emphasize on New Direct Marketing methods using Data Mining.**
  - Introduce Hadoop, MOBI, BIG DATA and SEM

COURSE STRUCTURE

- 60% of the class will be focused on Data Mining
- 20% on Business Intelligence
- 20% on Data warehousing.

Overview:

This course is about how companies apply two new technologies, data warehousing (DW) and data mining (DM, including business intelligence, BI) to empower their employees, and build and manage a customer-centric business model. Besides learning the strategic role DW and DM plays in an enterprise, you will also get a close-up look at
DW and DM by working on cases and gaining hands-on experience using software tools. Students taking this class will get an overview of the technologies of DW and BI/DM from a managerial perspective.

Fortune 500 companies such as American Express, Wells Fargo and Wal-mart have accumulated a great deal of data from their day to day business. Data warehouse is the technology that integrates the data collected from various sources that include transaction processing systems and e-commerce data collecting systems. Collecting and integrating data is just the first step. What are really critical are information, knowledge and insight. So the questions are what is the utility of the data? How can one use data in managing customer relationship and empowering employees? How can one uncover patterns and relationships hidden in organizational databases? These issues are addressed by a fast growing body of research and applications, broadly known as business intelligence and Data Mining (BI/DM). These technologies draw their strengths from the fields of information technology, statistics, machine learning and artificial intelligence.

In summary, managers need to understand the strategic values of their company's information assets. DW, BI and DM are cornerstones

**COURSE GOALS:**

After taking this class, students should be able to:

- Understand the basic terms that are used in DW, On Line Analytical Process (OLAP), BI and DM
- Communicate to Information Technology workers their business perspective in terms of the language of DW and DM
- Choose appropriate tools for specific purposes of storing, integrating and analyzing data (business consideration, and technical consideration).
- Use tools provided in class to perform simulated tasks in warehousing data.
- Use Enterprise Miner (optional) and JMP to perform DM activities on moderately large data sets.
- Articulate and present the results of their analyses and the business implications of these results
- Gain inference from your analysis, from Business and Statistical point of view.

1) In Data Mining you will develop in-depth skill set to do desktop Data Mining.
2) In Data Warehousing part you will learn, why companies need DW, advantages of DW and how to create a DW schema that an executive will understand, I will not teach the hands-on programming part, programming part is made available through Teradata student network and you can learn on your own.
3) In Business Intelligence you will learn what current BI can do, how to develop the requirements of a BI system for a company. I will not teach the hands-on programming part, programming part is made available through Teradata student network and you can learn on your own.
4) You will learn how the 3 parts are interconnected and integrated to form the basis of corporate knowledge system. How to leverage them to convert your company to near real-time corporation.

**Structure of lectures:**

IOM 528 will be organized in a way that includes some combination of the following: lectures, case-based class discussion, group project, computer lab work, and guest lectures.

This class is designed in such a way that only limited mathematical and statistical (Descriptive Statistics, Hypothesis testing and Regression) background is required. I will give a brief review on the above mentioned topics. Learning and understanding underlying DW concepts, studying cases, applying DM ideas and methods to business data, and communicating ideas and solutions will be our main theme. Technical details of selected DM methods will be discussed. Students are expected to use Data Mining software for various business applications.

**COURSE REQUIREMENTS**

1. **Class Attendance & Participation.** I strongly suggest that you attend all classes. I strongly encourage, as well as expect, questions during the lectures. I am always accessible by e-mail, and will be more than happy to speak with you before or after class or during office hours. I will keep you updated about a course TA/grader who will have office hours to assist you.

2. **Mini Cases and Business Insight Journal.** We will analyze mini cases during the semester. The mini cases will be evaluated and will be part of the class work points. You are required to maintain a Business insights Journal for your mini cases and group project, I will ask you to submit the business insight journal at least twice during the semester and it will be graded.

3. **Group Project.** I strongly believe the students learn the most during the project. The Group will consist of 3, 4 or 5 students. Learning to work in teams is essential and to get different perspective and will greatly enhance your learning. The project points will be based on the following criteria:

   a) **Selection of the project and submitting the proposal - 10%**
   b) **Submission of the Data set and descriptive statistics - 10%**.
   c) **Preliminary report with Analysis and further direction of the project – 40%**
   d) **Final Project report and Presentation – 40%**
A word document of the Final project report is required as well as a hard copy of the Final project report. The groups will also do peer evaluation of the group. The final report will include an Executive Summary write-up that translates the quantitative findings into a real-world analysis. You will be expected to participate in the discussion of your project during the semester to share your methodologies and interesting findings.

4. Midterm and Final Exam. The midterm will also take place at the beginning of class approximately one hour and 45 minutes. You may bring two sheets (four pages) containing formulas, definitions etc., to the midterm except solved problems and solved multiple choice questions. For the final, you may bring four sheets (eight pages) containing formulas, definitions etc., except solved problems and solved multiple choice questions. No make-ups of mid-term or the final will be given. You will receive a grade of zero for each missed exam unless you have a written excuse from your doctor or the professor. In case of emergency or approved absence, the professor may decide to give a make-up exam or redistribute the points.

There will be 1 midterm and 1 final exam. They are close-book.

<table>
<thead>
<tr>
<th>Grading</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class Work (Mini Cases)</td>
<td>20</td>
</tr>
<tr>
<td>Midterm</td>
<td>30</td>
</tr>
<tr>
<td>Final</td>
<td>20</td>
</tr>
<tr>
<td>Project</td>
<td>30</td>
</tr>
<tr>
<td>Total</td>
<td>100</td>
</tr>
</tbody>
</table>

Review Session. There may be a review session before the final exam.

Course Materials. The following items will be necessary for completion of reading assignments and homework and successful completion of the course.

1. Online Resources

➢ Sign up with Teradata University Network

Teradata University Network (www.TeradataUniversityNetwork.com) is a free learning portal designed to help faculty to teach, learn, and connect with others in the fields of data warehousing, DSS/Business Intelligence, and database.

Teradata University offers web-based courses and related web sites on data warehousing, DSS/BI and database. They have a library of Teradata white papers. Students can
become Teradata certified. We will use their material and software in the class particularly for the Business Intelligence and Data warehousing part of the class.

Students register for www.TeradataStudentNetwork.com and login using the current password: Big Data

➢ Sign up with DMreview.com

DMreview is another portal with many resources for DM/BI/DW professionals.

2. Text Books and Class notes

➢ The first book is a standard book for Data Mining, the book talks about the various techniques and it is written from computer science perspective. (Recommended)


    Note: The book is written from a Computer Science and it will help you to understand the data mining techniques but it is does not have real world business application – Buy the book if you want to understand Data Mining Algorithms. My PowerPoint slides will cover the data mining topics but not in depth.

➢ The second book is a standard book for Direct Marketing, the book talks about the various techniques and it is written from Marketing perspective. (Recommended)

    The New Direct Marketing: How to Implement A Profit-Driven Database Marketing Strategy
    Hardcover: 736 pages
    Publisher: McGraw-Hill; 3 edition
    ISBN-10: 0070580561

➢ The third book is from SAS – The world’s leading Data mining software company. This book introduces you to industry level Data mining software – SAS Enterprise Miner. (Recommended)

    website: [www.sas.com](http://www.sas.com)
The fourth book is from SAS (I am currently reviewing this detailed technical book). This book introduces you to industry level Data mining software – SAS Enterprise Miner. (Recommended)

Data Mining Using SAS Enterprise Miner by Randall Matignon, John Wiley and Sons publishing

website: www.sasenterpriseminer.com

Class notes. Class notes for this class will be available on blackboard. You should familiarize yourself with these notes before they are covered in class.

Recommended (If you want to concentrate on Data Warehousing)


Recommended (If you want to concentrate on Analytics, Business Intelligence and Forecasting)

1. Information Dashboard Design: The Effective Visual Communication of Data (Paperback) –by Stephen Few (Author)
   http://www.amazon.com/Information-Dashboard-Design-Effective-Communication/dp/0596100167/ref=sr_1_1?ie=UTF8&s=books&qid=1240440281&sr=1-1

2. Balanced Scorecards & Operational Dashboards with Microsoft Excel (Paperback) by Ron Person (Author)
   http://www.amazon.com/Balanced-Scorecards-Operational-Dashboards-Microsoft/dp/0470386819/ref=sr_1_1?ie=UTF8&s=books&qid=1240440389&sr=1-1

3. Competing on Analytics: The New Science of Winning (Hardcover) by Thomas H. Davenport (Author), Jeanne G. Harris (Author)
   http://www.amazon.com/Competing-Analytics-New-Science-Winning/dp/1422103323/ref=sr_1_1?ie=UTF8&s=books&qid=1240442477&sr=1-1

4. Introduction to Time Series and Forecasting (Hardcover) by Peter J. Brockwell (Author), Richard A. Davis (Author)
   http://www.amazon.com/Introduction-Time-Forecasting-Peter-Brockwell/dp/0387953515/ref=sr_1_1?ie=UTF8&s=books&qid=1240442944&sr=1-1

Important dates: (Refer to Schedule of classes for up-to-date information)

Midterm exam:
June 13, 2012, Wednesday 6:00-10:00 P.M.

*Final Exams:*

**July 11, 2012, Wednesday 6:00 P.M.-10:00 P.M.**

**Academic Integrity.** Academic dishonesty of any type will not be tolerated in this class. Students who find this statement ambiguous should consult the Student Conduct Code, page 83, of the USC *SCampus* handbook.

A comment about writing the assignments up individually and working in teams: You can work together in teams to discuss the problems and concepts. However, you are required to write up the assignments individually. This means that all the words in you assignments are your own, and you generate all of your own computer output and graphs.

Now, while correct solutions will have very similar or even the same computer output, no two answers should be phrased the same way. If I find two or more assignments that are highly similar, I will at a minimum give the homework a zero, and may refer the incident to the Dean. *Do not test me on this policy.*

**STUDENTS WITH DISABILITIES**

Any student requesting academic accommodations based on a disability is required to register with Disability Services and Programs (DSP) each semester. A letter of verification for approved accommodations can be obtained from DSP. Please be sure the letter is delivered to me as early in the semester as possible. DSP is located in STU 301 and is open 8:30 am - 5:00 pm, Monday through Friday. The phone number for DSP is 213 740-0776.

**Tentative Schedule:**

---

I. The course will start with Data Mining. The Data Mining part of the class will be quantitative and the following topics will be covered in it.

1. Standard Data Mining techniques:
   a. Classification
   b. Clustering
   c. Association
   d. Prediction
   e. Text Mining, Link Analysis, Visual Data mining (if time permits)

Using various appropriate techniques,
   i) Bayesian Estimation
   ii) Neural Networks
   iii) Decision Tree
   iv) Similarity Measures
v) Other techniques like Boosting, Bagging (if time permits)

2. Statistical Model Building using Regression and Logistic Regression. Depending on the project other topics may be covered.

II. The second part of the course will be Business Intelligence Software. You will be introduced to software used as Business Intelligence software.

- We will discuss Forecasting methods to identify trends, seasonality etc.,

III. The third part of the course will be Data Warehousing. You will be introduced to Data Warehousing from business perspective, how to create Data Warehouse Architecture.

IV. In addition to the above, if time permits we will discuss the following,

a. Hadoop
b. Big Data
c. SEM – Search Engine Marketing
d. MOBI
V. **Approximate Schedule of class – I will add the reading pages from the New Direct Marketing book.**

<table>
<thead>
<tr>
<th>Date</th>
<th>Topic</th>
<th>Reading from textbooks</th>
<th>Reading from Class notes</th>
<th>Due/Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>16 May</td>
<td><strong>LEC &amp; LAB</strong> 1 Introduction to Classification - Distance Based Algorithms / Evaluating Classification/ Google Analytics and Adwords</td>
<td>JM 1-26, JM 36-40, JM 359-362, JM 372-375, JM 285-290, JM 310-318, JM 347-350, SAS 39-67</td>
<td>Dr. Ansari Notes</td>
<td>Bring in Laptop with JMP software loaded</td>
</tr>
<tr>
<td>23 May</td>
<td><strong>LAB &amp; LEC</strong> 2 Classification Methods – Decision Tree Based Methods and Bayesian Classification</td>
<td>JM 291-306, SAS 19-36</td>
<td>Dr. Ansari Notes</td>
<td>Bring in Laptop</td>
</tr>
<tr>
<td>30 May</td>
<td><strong>LAB &amp; LEC</strong> 3 Review of Statistical Model Building and Introduction to Enterprise Miner</td>
<td>SAS 39-67, SAS 67-81</td>
<td>Dr. Ansari Notes</td>
<td>Bring in Laptop</td>
</tr>
<tr>
<td>6 June</td>
<td><strong>LAB &amp; LEC</strong> 4 Visual Data Mining/Logistic Regression/New methods</td>
<td>JM 358-359, JM 327-336</td>
<td>Dr. Ansari Notes</td>
<td>Bring in Laptop</td>
</tr>
<tr>
<td>9 June</td>
<td><strong>LAB &amp; LEC</strong> 5 Logistic Regression / Neural Network/Clustering and Association</td>
<td>JM 384-414, JM 227-234, JM 384-414, JM 227-234, SAS 91-104, SAS 105-109</td>
<td>Dr. Ansari Notes</td>
<td>Bring in Laptop</td>
</tr>
<tr>
<td>13 June</td>
<td><strong>Midterm/Business Intelligence</strong></td>
<td>TUN: Microstrategy BI Information/Continental Airlines Takes Off with Real-time Business Intelligence</td>
<td>Dr. Ansari Notes</td>
<td>Midterm</td>
</tr>
<tr>
<td>16 June</td>
<td><strong>LAB &amp; LEC</strong> 6 Business Intelligence/ Dash Boards</td>
<td>TUN relevant information</td>
<td>Dr. Ansari Notes</td>
<td>Bring in Laptop</td>
</tr>
<tr>
<td>20 June</td>
<td><strong>LAB &amp; LEC</strong> 7 Lecture DW1: Data Warehousing(I): Strategic View</td>
<td>JM 105-114, JM 127-134 and TUN relevant information</td>
<td>Dr. Ansari Notes</td>
<td>Mini Case 4</td>
</tr>
<tr>
<td></td>
<td>Lecture DW2: A Tactical View</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>27 June</td>
<td><strong>LAB &amp; LEC</strong> 8 Lecture DW3 : Dimensionally Designed DW (I)</td>
<td>JM 114-123 and TUN relevant information</td>
<td>Dr. Ansari Notes</td>
<td>Turn in your Preliminary report</td>
</tr>
<tr>
<td></td>
<td>Lecture DW4: Dimensionally Designed DW (II)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>30 June</td>
<td><strong>LAB &amp; LEC</strong> 9 Lecture DW5: OLAP and Business Intelligence Lecture DW6: Web-Based OLAP and Business Reporting/Presentation/Review</td>
<td>JM 135-137, JM 144-152 and TUN relevant information</td>
<td>Dr. Ansari Notes</td>
<td>Mini Case 5</td>
</tr>
<tr>
<td>11 July</td>
<td><strong>Final Exam/Project Presentation</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

TUN – Teradata University Network, SAS – Enterprise Miner Text book

JM – Data Mining textbook by Jiawei Han and Micheline Kamber