

University of Southern California
MARSHALL SCHOOL OF BUSINESS
Fall 2014 – First Draft

Course Guidelines & Syllabus

IOM 528 – DATA WAREHOUSING, BUSINESS INTELLIGENCE AND DATA MINING -16274D
(AKA – Big Data Analytics)

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

COURSE OBJECTIVES

- Introduce Data Warehousing (DW) , Big Data (BD) and Business Intelligence
- To develop data mining skills to monetize data
- To develop desktop level data mining skills using SAS JMP software and introduce industry level “SAS-Enterprise Miner” tool.
- To provide hands-on skills of building predictive models
- To develop Critical thinking and Modeling Skill Set.
- Introduce “MAGIC” framework for building efficient predictive models.

COURSE STRUCTURE

- **70% of the class will be focused on Data Mining**
- **10% on Business Intelligence tools**
- **20% on Data warehousing and Big Data Tools**

Overview:

Big Data has made available a lot of Structured and Unstructured data along with it new Paradigms for Monetizing data. Many of these new paradigms are ill-defined problems and the businesses are struggling to develop methods and models to leverage the “Big Data Opportunity.” Most entry level Analytics professionals are “man-in-the-middle” between data and business requirements they don’t know how to fully utilize all the available data and are not fully aware of the domain expertise needed to build efficient business savvy models. This course is for students who want to be “Standing on the Shoulders of Giants (Big Data Analytics)” and have great vision on the data side and on the business side, understand Big Data - its potential and drawbacks, Statistics - its usefulness and limitations, data mining - its usefulness and limitations, Business needs and available opportunity. In short this class is about how to monetize Data in customer facing applications.

Companies have huge amount of data in their data warehouse and have access to Big Data through 3rd party APIs. Companies want to leverage data for decision making by building “Data Driven Decision Making Models” and they want to monetize big data using data mining (DM) and Business Intelligence.

The access to social, demographic, transactional, click-stream, web usage etc., data has made companies “data rich” and now they want new ways to monetize data as well as enhance the traditional predictive models using Enriched data. For example, 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 now they want to monetize the data by providing value to customers and sell their products and services through Omni channels in an efficient manner.

In the Big Data Analytics space what are really critical are information, knowledge, insight and monetization. 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 databases? How can one creatively find ways to monetize data through analytical models? How can one enhance the performance of existing models?

In summary, managers need to understand the strategic values of their company's information assets, be capable of building analytical models to monetize data, understand the models build by third party companies, be able to extract insights for the models and be able to visualize data and insights.

COURSE GOALS:

After taking this class, students should be able to:

- Understand the basic terms that are used in DW, BD, On Line Analytical Process (OLAP), BI and DM
- Communicate to Information Technology workers their business perspective in terms of the language of DW, BD and DM
- Choose appropriate tools for specific purposes of storing, integrating and analyzing data (business consideration).
- Identify a Business Analytics Project, collect Data, White Board, Story Board the problem, Build Analytical Models, evaluate performance and monetize it.
- Use Enterprise Miner 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 and learn the industry level Data Mining tool.
- 2) In Data Warehousing/Big Data (DW/BD) part you will learn, why companies need DW/BD, advantages of DW/BD and how to create a DW schema that an

- executive will understand, I will not teach the hands-on programming for DW part, DW 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 Tableau) 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. How to monetize data.

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/BD concepts, studying cases, applying DM/BD 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 cases in class.

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.
2. **Class Work (Mini Cases).** We will analyze mini cases during the semester. The mini cases will be evaluated and will be counted towards the class work points. **Note:** I will assign discussion Business cases from TeraDataUniversity for class discussion.
3. **Topic Presentation: Each student (or a team of 2) in the class will be making a 5-minute presentation on a topic related to “Big Data Analytics” – the topic has to be approved by the Professor.**
4. **Group Project.** I strongly believe the students learn the most during the project. The Group will consist of 4 or 5 or 6 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 approval of the proposal - 15% (Recursive Process)**
- b) **Submission of the Data set and descriptive statistics - 15%.**
- c) **Preliminary report with Analysis and further direction of the project – 30%**
- d) **Final Project report, Poster and Presentation – 40%**

A word/PowerPoint 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 in class.

- 5. **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.
- 6. **Analytics Hackathon: This is a new feature that will be tried this year, currently no points are allocated to it.**
- 7. **Industry Mentor: Optional**
- 8. **SAS-Enterprise Miner Training: Optional**
- 9. **IBM Big Data Platform Training: Optional**

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

Grading	Percentage
Class Work	25
Midterm	25
Class Presentation	5
Final	20
Project	25
Total	100

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: *****

➤ **Sign up with IBM → Currently working with IBM to get their Academic Resources.**

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)**

Data Mining: Concepts and Techniques, Second Edition by Jiawei Han and Micheline Kamber, Morgan Kaufmann Publishers, ISBN 13: 978-1-55860-901-3, ISBN-10: 1-55860-901-6, website: www.mkp.com

Note: The book is written from a Computer Science and it will help you to understand the data mining techniques but it 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
ISBN-13: 978-0070580565

- 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)**

Data Mining Using SAS Enterprise Miner – A Case Study Approach, Second Edition. ISBN 1-59047-190-3, SAS publishing
website: 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)

- Building the Data Warehouse 3rd Edition, W.H. Inmon, Wiley, ISBN 0-471-08130-2
- Data Warehouse: Practical Advice from the experts, Joyce Bischoff and Ted Alexander, Prentice hall, ISBN 0-13-577370-9
- Recommended: Data Warehousing: using the Wal-mart model. *Paul Westerman*, Morgan Kauffman publishers.

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: TBA

Final Exams: TBA

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 your 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 (Similar to what is covered in core MBA class)
 - 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 Data Mining methods 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.

- III. The third part of the course will be Data Warehousing/Big Data. You will be introduced to Data Warehousing from business perspective, how to create Data Warehouse Architecture. Big Data Platforms will be discussed.
- IV. In addition to the above, if time permits we will discuss the following,
 - a. Hadoop
 - b. MOBI
 - c. Advanced Data Mining Concepts

Approximate Schedule of class

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

JM – Data Mining textbook by Jiawei Han and Micheline Kamber

Date		Topic	Reading from textbooks	Reading from Class notes	Due/Other
	1	Introduction to Big Data Analytics Classification - Distance Based Algorithms / Evaluating Classification Critical Thinking Exercise / Project Ideas	JM 1-26, JM 36-40, JM 359-362, JM372-375 JM285-290, JM 310-318, JM 347-350, SAS 39-67JM 291-306	Dr. Ansari Notes	Bring in Laptop
	2	Descriptive Statistics/Classification Methods – Decision Tree Based Methods Data Hackathon/Project Idea Discussion	JM 291-306 SAS 19-36SAS 39-67, SAS 67-81	Dr. Ansari Notes	Bring in Laptop Group List Due
	3	Visual Data Mining/Review of Hypothesis Testing/Logistic Regression/ Visual Data Hackathon/	JM 358-359, JM 327-336 JM 358-359, JM 327-336	Dr. Ansari Notes	Bring in Laptop Project Proposal Due Turn in your Class work 1
	4	Logistic Regression / Naïve Bayesian/ Project Refinement	JM 384-414, JM 227-234 JM 384-414, JM 227-234, SAS 91-104, SAS 105-109	Dr. Ansari Notes	Bring in Laptop HW1-optional
	5	Extra Time to Cover Topics			
	6	Business Intelligence and SAS Miner Project	TUN relevant information	Dr. Ansari Notes	Bring in Laptop Turn in your Class work 2 Turn in your Data set and Descriptive Stats
	7	Neural Network/ Analytics Hackathon/ Data Enrichment via Big Data	JM 227-234, SAS 91-104,	Dr. Ansari Notes	Midterm/ Turn in your Class work 3
	8	Midterm/Business Intelligence/ Search Engine Marketing	TUN: Microstrategy BI Information/Continental Airlines Takes Off with Real-time Business Intelligence	Dr. Ansari Notes	Bring in Laptop Turn in your class work 4
	9	Clustering and Association/ Google Analytics and Adwords	JM 114-123 and TUN relevant information JM 123-126 and TUN relevant information	Dr. Ansari Notes	Bring in Laptop Turn in your Preliminary report
	10	Lecture DW1: Data Warehousing(I): Strategic View Lecture DW2/BD: What is Big Data Lecture DW3: Dimensionally Designed DW (I&II)	JM 114-123 and TUN relevant information JM 123-126 and TUN relevant information JM 135-137, JM 144-152 and TUN relevant information	Dr. Ansari Notes	Turn in your class work 5 HW2-optional
	11	Lecture DW5: OLAP,/More on Big Data			
	12	Review for Final Exam/Poster Session			Turn in your Poster

	13	Final Exam/Project Presentation			Oral Presentation of Project and Turn in your Final Report
--	----	--	--	--	--

IOM528 Course Objectives Achieved from Spring Class

Course Objectives achieved and more...

1. Allowed students to take ownership of the class, be involved → Learn from each other, facebook, IBM Watson Competition, write meeting minutes of class, volunteer for class etc.,
2. Allowed students to get maximum exposure at various events and through Industry Mentorship.
3. Develop Critical Thinking and Creative Thinking
4. Gave industry level tools and arranged for training
5. Went with students to Digital Analytics and TeraData Marketing Summit
6. Arranged for two workshops SAS-Enterprise Miner Training and IBM Big Data tools
7. Introduce “MAGIC” framework and trained students in building “Practical Data Mining” Models
8. Helped Students to solve “New Paradigm” problems through semester long project and guided them on the project.
9. Provide theory on Classification, Clustering and Association Methods.
10. Discussed other students’ projects in class to help students to become experts in “OPEN MIND” and “clean SWEEP” methods
11. Instead of being the “man in the middle” of data and executives, helped “students to stand on the shoulder of giants” foresee how data can be utilized and how to communicate effectively with Business Executive.
12. Provided many practitioners tricks to develop better models.
13. Developed “Business Mindset” for various non-business students
14. Enabled students to leverage “Computer Science” enabled data availability
15. Personally learned a lot from students and provided the additional knowledge back to students.
16. Pushed students to commit to “Life Long Learning” and made available my mentorship for the future.