CS 504 Principles of Data Management and Mining

Course Description (From Catalog)

Techniques to store, manage, and use data including databases, relational model, schemas, queries and transactions. On Line Transaction Processing, Data Warehousing, star schema, On Line Analytical Processing. MOLAP, HOLAP, and hybrid systems. Overview of Data Mining principles, models, supervised and unsupervised learning, pattern finding. Massively parallel architectures and Hadoop.

Instructor: James J. Nolan, Ph.D.
Contact: jnolan5@gmu.edu
Day/Time: Tuesday, 4:30-7:10pm
Location: Arlington: Founders Hall 118
Office Hours: By appointment

Prerequisites

Graduate Standing
Note: This course cannot be taken for credit by students of the MS CS, MS ISA, MS SWE, MS IS, CS PhD or IT PhD programs.

Honor Code Statement

Please be familiar with the GMU Honor Code. In addition, the CS department has its own Honor Code policies. Any deviation from this is considered an Honor Code violation.

Disability Accommodations

If you are a student with a disability and you need academic accommodations, please see me and contact the Office of Disability Services (ODS) at 993-2474, http://ods.gmu.edu. All academic accommodations must be arranged through the ODS.

Textbooks

Required (available in Safari Books):
Data Science for Business: What You Need To Know About Data Mining and Data-Analytic Thinking (Foster Provost and Tom Fawcett)
NoSQL Distilled: A Brief Guide to the Emerging World of Polyglot Persistence (Pramod J. Sadalage; Martin Fowler)

Optional (available in Safari Books)
Beginning Database Design: From Novice to Professional, Second Edition (Clare Churcher)
R for Data Science (Hadley Wickham; Garrett Grolemund)
Advanced Analytics with R and Tableau (Jen Stirrup; Ruben Oliva Ramos)
Data Analytics with Hadoop (Benjamin Bengfort; Jenny Kim)

R Resources
http://dataservices.gmu.edu/software/r

Grading Policies
Homework and Quizzes: 30%
Midterm: 35%
Final Project: 35%

Grading Scale

A	100-93
A-	92-90
B+	89-87
B	86-83
B-	82-80
C	79-70
F	69-0

LATE ASSIGNMENTS, HOMEWORKS, PROJECTS OR OTHERWISE WILL NOT BE ACCEPTED.

Topics to be covered (subject to change at Instructor's discretion)

- Introduction to Data Analytic Thinking
- Databases and the Relational Model
- Data Visualization
- Data Warehousing
- Big Data- No SQL/MapReduce/Hadoop
- Advanced Data Mining techniques- text, imagery, and video analytics

Important Dates (subject to change at Instructor's discretion)

- October 10 – Columbus Day recess- no class
- October 17- Project proposal and presentations due
- October 24 – Midterm Exam
- December 5 - Project Results Presentations

Class Project
35% of final grade - where you solve a data-science problem from data preparation to data product. Form groups of 3 students.

- Project Proposal Paper- 2 pages maximum plus 10-minute in-class pitch -- due on 10/17.
    - Should include answers to the following questions:
        - What is the problem?

        - Why is it interesting and important?
        - Why is it hard? Why have previous approaches failed?
        - What are the key components of your approach?
        - What data sets and metrics will be used to validate the approach?
- Project Results Presentation - 12-minute presentation -- due on 12/5
- Final report - 6 pages maximum -- due on 12/5.
    - For guidance on writing the final report, see slide 70 of Eamonn Keogh's KDD'09 Tutorial on How to do good research, get it published in SIGKDD and get it cited!
    - Follow ACM formatting guidelines