CS 593: Knowledge Discovery in Databases

Stevens Institute of Technology

Khasha Dehnad

kdehnad@stevens.edu
Khasha.dehnad@aimsinfo.com

Course Requirements

Recommended Prerequisites:

- **Familiarity with the principals of statistics and probabilities and Data Mining; for example, completion of MGT 502 (no credit).**

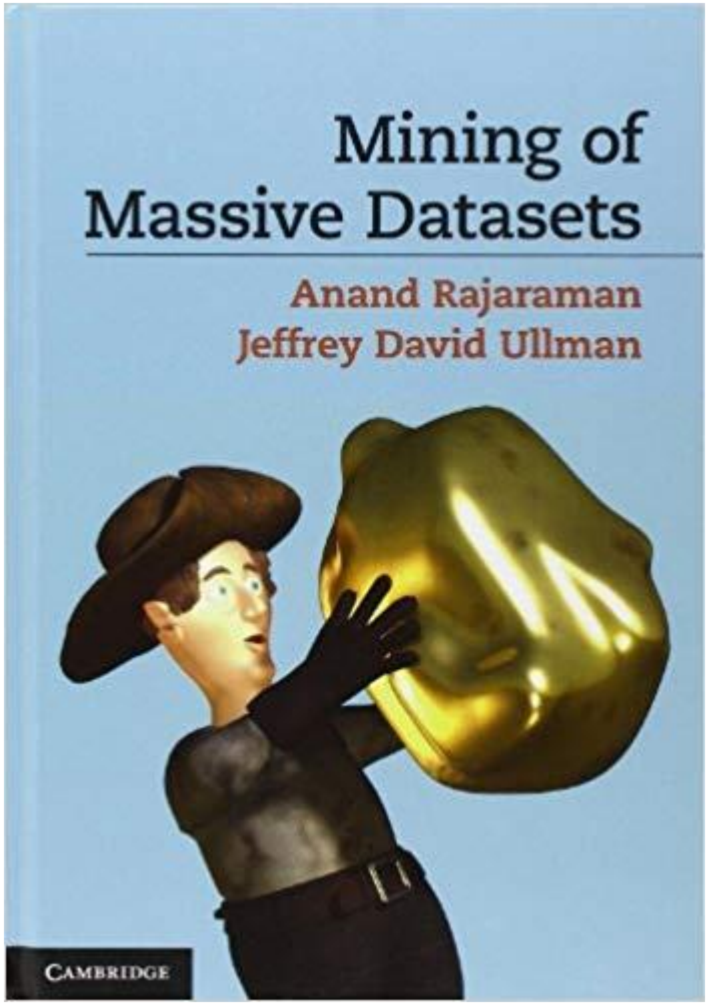
Hardware and Software:

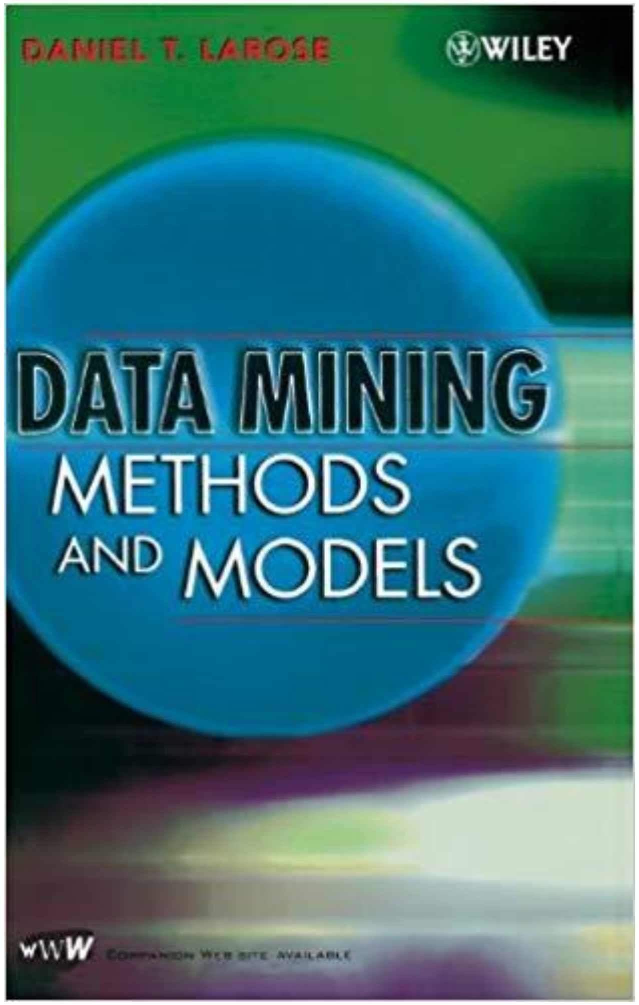
- **Lap top with internet access and ability to install software (admin rights).**
- **Students will be installing SAS on their computers**

Books, Notes, and Manuals:

- Data Mining, Methods and Models, D. T. Larose, Wiley–Interscience, Latest Edition
- Mining of Massive Datasets, A. Rajaraman, J.D Ullman, Stanford University, Cambridge University Press, 2012
- Lecture Notes and Handouts
- Real world projects and case studies

Text books


Mining of
Massive Datasets
Anand Rajaraman
Jeffrey David Ullman
CAMBRIDGE


DANIEL T. LAROSE
WILEY
DATA MINING
METHODS
AND MODELS
WWW
COMPANION WEB SITE AVAILABLE

Course Overview

Big Data refers to data sets whose volume (amount of data collected, number of data sources), velocity (rate at which data is collected) and variety (heterogeneity of data and data sources) are so extreme that advanced Data Mining Algorithms are needed to process and discover useful patterns in data for actionable intelligent decisions, in a reasonable amount of time. The purpose of this course is to introduce theoretical as well as practical aspects of advanced, as well as, well established algorithms for mining massive datasets. Topics include: Naïve Bayes & Bayesian Networks , Stream Data Mining, Big Data Definition,Dimension Reduction techniques e.g. Principal Component Analysis (PCA), and recommendation systems.

Course Schedule

Introduction	**Week 1**
Linear Algebra Review **Intro to SAS**	**Week 2**
Intro to SAS (continued) and **Basic Statistics Review,**	**Week 3**
Introduction to Big Data , Massive Data sets **Map-Reduce,** **Relational Algebra in Big Data environment**	**Week 4**
Big Data , Massive Data sets (continued) **Linear Algebra in Big Data environment** **Recommendation System**	**Week 5**

Course Schedule

Mining Data Streams And Sensor Data Link and Social Network Analysis	Week 6
Affinity and Market Basket Analysis	Week 7
Principal Component Analysis and Factor Analysis	Week 8
Linear Regression	Week 9
Multiple Linear Regression	Week 10
Logistic Regressions	Week 11
Special Topics	Week 12
Student Projects and Final Exam	Week 13 &14

Assignments and Grading

Assignments	**Grade Percent**
Exercises	30%
Mid-term	20%
Final	20%
Final project /research paper	30%
Total Grade	**100%**

Project Case Study

Project:
A real world data mining project (problem statement, data, methodology/algorithm), software, execution and analysis, references, documentation, and presentation). The problem statement, sample data, relevant methodology/algorithm).

Case Study:
A case study from literature/books, prepare and deliver a comprehensive presentation including, problem statement (‘profound question’), data source(s), methodology, data mining, result, suggestions for future work, and references.