

## **CSE 347/447 Data Mining**

### **Fall 2009**

Instructor: Brad Askins – MWF 1:10 – 2:00 – Packard Lab 112 (M), Coppee Hall 105 (WF)  
Office: Packard Lab 316  
x86216  
email bja206 at lehigh.edu  
www.cse.lehigh.edu/~bj

Text: Data Mining Concepts and Techniques, 2<sup>nd</sup> Ed.,  
Jiawei Han and Micheline Kamber,  
2006 Morgan-Kaufmann

Data mining is shorthand for discovering patterns and finding meaning in large datasets.

Given that oceans of data and significant computing technology are such a part of our daily lives now, the course will start out by giving extra attention to the fundamentals of exploratory statistics and data visualization. Not only is this the core of data mining technology needed to understand more specialized concepts, but this is technology that everyone can use all the time.

Next covered will be approaches to dealing with imperfect data, and database systems technology for managing and manipulating large datasets.

The course will then cover the traditional data mining techniques dealing with patterns, classifications, predictions, and clusters, and extend those techniques into streams and sequences of data.

Finally, in addition to the usual data to be mined, there now also is a wealth of unstructured textual data such as SEC filings, news repositories, blogs, and twitter streams that can be mined for corporate intelligence or cultural insights. We will cover the application of mining techniques to this type of data as well.

Commercial and open-source software for analysis and visualization, as well as the powerful hardware to run it is now commonplace. We will be using Mathematica, R, and SQLServer in class and for homework. Working with software like this we will be able to focus primarily on the concepts and to a lesser degree on lower-level programming issues.

Monday class will be in the PL 112 (the Sandbox). It will be a lab format class where we will work on data mining examples and experiments. Wednesday and Friday class will be lecture and discussion format.

Graduate students taking 447 will be expected to complete a paper or project in addition to the other course work.