

**Syllabus**  
**MSCS 228: Data Mining**  
**Fall 2003**  
**Call #27049**

<http://www.mscs.mu.edu/~cstruble/class/mscs228/fall2003>

**Instructor: Craig A. Struble, Ph.D.**

**Office:** 369 Cudahy Hall  
**Office Hours:** 11:00 a.m.–1:00 p.m. TuWe, and by appointment  
**Phone:** (414)288-3783  
**Email:** [craig.struble@marquette.edu](mailto:craig.struble@marquette.edu)

**Class Meets: 4:20–5:35 p.m., MW, Cudahy 128**

## Overview

“Data mining” refers to a collection of techniques for extracting “interesting” relationships and knowledge hidden in data. We will study a variety of these techniques and carry out practical exercises to understand what is and what is not “interesting.” In the process, we will identify strengths and weaknesses of each technique.

## Prerequisites

Programming instruction up through data structures (e.g., COSC055) is required. Additionally one course in an area of expertise related to data mining is required. Acceptable courses include database systems (e.g., COSC153), artificial intelligence (e.g., COSC159), or statistics (e.g., MATH164). Students not meeting these requirements must obtain instructor permission.

## Topics Covered

Knowledge discovery in databases, classification and prediction, clustering, association rule mining, feature selection, discretization, data cleansing, decision trees, neural networks, regression, Bayesian statistics, etc.

## Textbook and References

### Required

- Margaret H. Dunham, *Data Mining: Introductory and Advanced Topics*, 2003, Pearson Education, Inc. (Prentice Hall), ISBN 0-13-088892-3.

## Recommended

- Ian H. Witten and Eibe Frank, *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*, 2000, Morgan Kaufmann Publishers, ISBN 1-55860-552-5.
- *Communications of the ACM*, Special Issue on Data Mining, November 1996. **Available online through the library.**
- *Communications of the ACM*, Special Issue on Knowledge Discovery, November 1999. **Available online through the library.**
- *Communications of the ACM*, Special Issue on Data Mining, August 2002. **Available online through the library.**

Additional material will be on reserve in the library and on the course web site.

## Grading

Your grade will consist of the following components, weighted as shown:

Lab Notebook	20%
Midterm Exam	20%
Final Exam	25%
Term Project	30%
Intangibles	5%

This semester, I will be using the following grading scale to assign letter grades. I recommend reading Dr. Struble's *grading philosophy*, which is available on his web site, to understand why he has chosen the following grading scale. Grades for each assignment, exam, etc. will be curved to fit this grading scale.

Range	Letter Grade
[90–100]	A
[80–90)	AB
[70–80)	B
[60–70)	BC
[50–60)	C
[40–50)	CD
[30–40)	D
[0–30)	F

## Writing Expectations

Good writing skills are essential for effective communication of your ideas. All work submitted in this class is expected to be in well written English. Poorly written work may have points deducted, worth no more than 5% of the assignment grade. I will correct or point out grammatical and spelling errors in your work. In extreme cases, the work will be returned to be rewritten and resubmitted, with a 10% penalty.

All written work should contain citations to referenced papers, web sites, books, etc. For examples of acceptable citation style, look at articles from journals and magazines such as Communications of the ACM, SIGKDD, SIGMOD, etc.

If you are not confident in your ability to write English well, you should seek help from the Ott Memorial Writing Center. Visit <http://www.marquette.edu/writingcenter/> for more information.

## Lab Notebooks

During the semester, approximately nine (9) labs exploring the major topics discussed in class will be assigned for students to complete as the semester progresses. Each student is required to maintain a lab notebook, which will include raw results, notes, answers to lab questions, and final lab reports for each lab assigned. The instructor(s) will collect and review the lab notebooks at three times during the semester, shown below.

Labs	Date Due
1-3	Wednesday, September 24
4-6	Wednesday, October 22
7-9	Monday, November 24

Guidelines for the expected organization and quality of your lab notebooks are available through the course web site.

## Term Project

The term project is a group project intended to tie together concepts and techniques discussed in class. Your group will write a project proposal, maintain a project web site, carry out the proposed project, and submit a final project report of no more than 15 pages in length. The goal of the project is to produce a report that could be published in a regional conference about data mining or some other trade publication (e.g., Dr. Dobbs). More details about the project will be made available in lecture and on the course web site. The major milestones for the project are listed below.

Date	Event
September 10, 2003	Project Groups Formed
October 1, 2003	Draft Proposal Due
October 29, 2003	Final Proposal Due
November 19, 2003	Progress Report
December 1 and 3, 2003	Class Presentations
December 5, 2003	Project Due

## Exams

There will be one midterm and one final exam. Questions may be posed in any form, such as short answer, multiple choice, or computational problems. The final exam will be comprehensive, but may emphasize the material covered after the midterm exam. The dates and times for the midterm and final exam are shown below.

Exam	Date and time
Midterm	Wednesday, October 15
Final	Monday, December 8, 3:30 p.m. – 5:30 p.m.

## Intangibles

A small portion of your grade consists of items not easily measured and categorized. These include things like class participation, meetings with the instructor, keeping up with the reading, using the course message boards/ mailing list, etc.

## Late Policy

Assigned work in this course must be turned in by the specified due date. Late work will **not** be accepted.

## Attendance Policy

While attendance at lecture is optional, it has been my experience that there is a direct correlation between attendance and the overall grade received in this course. If you miss class, you are responsible for finding out what you missed from a classmate, including notes and assignments. Requests from absent students for notes or for meetings to discuss what was missed will be ignored. Absences on days when an assignment is due or an exam is scheduled must be accompanied by official documentation in order to make up the work or exam.

## Academic Honesty

All students are expected to adhere to the standards of student conduct as described in the *Community Expectations* section of the student handbook.

Lab assignments are intended to reflect effort by an individual in the course. Students may discuss lab assignments in a general way; i.e., discussing the *nature* of the assignment or providing clarifications for the lab instructions. Students may show each other how to use the software tools in the course, but they may not share results or use the tools for another student (e.g., sitting down and using a tool during a student's session). Sharing solutions to lab assignment questions in any form is strictly prohibited, unless otherwise stated.

You are **ENCOURAGED** to refer to outside material such as journals, web pages, and books. Do not feel guilty about using outside material; just make sure that you cite your references. Furthermore, you must write your solutions in your own words. It is not acceptable to directly copy material from another source. **Failure to properly cite your references may result in a charge of plagiarism. Give proper credit where credit is due!**