# cse521
# DATA MINING

Professor Anita Wasilewska

Spring 2023

# COURSE SYLLABUS

## Course Web Page
## www.cs.stonybrook.edu/~ cse521

The webpage contains:

Detailed Lectures Notes slides

Some course Book slides

Slides of some previous Data Mining Presentations

Course Syllabus

Course Project Description and data

Course Final PRESENTATIONS Description

Course Final Presentations Evaluation Description

Please **check it** often

## Course Text Book

**DATA MINING**

**Concepts and Techniques**

Jiawei Han, Micheline Kamber, and Jian Pei

**Morgan Kaufman Publishers, 2011**

Second or Third Edition

My Lectures follow the Both Editions

There is no essential differences between the editions

We will follow the book very closely

## Course Description

Data Mining is a **multidisciplinary** field

It brings together research and ideas from
database technology,
machine learning, statistics, pattern recognition,
knowledge based systems, information retrieval,
high-performance computing, and
data visualization to name **the few**

## Course Description

Data Mining main focus is the **automated extraction** of **patterns** representing knowledge implicitly stored in large databases, data warehouses, and other **massive** information repositories

The course will closely follow the book

Course **Lectures** are designed to explain in details the material from book chapters

## Course Description

The course is designed to give a broad, yet in-depth **overview** of the Data Mining field

We will examine **slowly** and in rigorous detail the most **basic** and **important** algorithms and techniques

We will also will explore the newest trends and developments

# GRADING GENERAL PRINCIPLES

1. **Test 1** and **Test 2** are conducted in person in CLASS

2. Project, Final Report Presentation, and Final Report Paper are to be conducted in **Teams** of 4 - 5 students

3. All members of the **Team** receive the same grade

4. **NONE of the grades will be curved**

# TEAMS FORMATION

Please email **TA** ( to be defined) names, IDs, and emails of your

**Team** members denoting who is the Team Leader

**TA** will assign a Team number to each Team to be used for future **identification**

**Contact** the TA if you **do not have** a team partner and he will help you to form a **Team**

# COURSE STRUCTURE

The course **Lecture Slides** are written by me, except when I say "Book Slide" or give other credentials

We list here chapters numbers from 2nd edition

We follow 2nd edition chapter numbers by chapters numbers from 3rd edition put between parenthesis

We will follow the **course structure** below

**Part 1 Introduction**

Data Preprocessing, Data Warehouse

Book chapters 1- 3 (1 - 4 ) and Lectures 1- 3

# COURSE STRUCTURE

**Part 2 Classification**

Decision Tree Induction and Neural Networks

Book chapter 6 (8- 9) and Lectures 4 - 8

Test 1 Review Lecture

**TEST 1**

**Classification Project**

Project Description is posted on the course webpage

# COURSE STRUCTURE

**Part 3**
**Association Analysis**

Apriori Algorithm

Classification by Association

Book chapters 5, 6 (6, 9) and Lectures 9, 10

**Part 4 Other Classification Methods**

Genetic Algorithms

Bayesian Classification

Book chapter 6 (9) Lectures 11, 12, 13

Test 2 Review Lecture

**TEST 2**

# COURSE STRUCTURE

**Part 5 Clustering, Statistical Prediction**

Book chapter 7 (10, 11) and Lectures 14, 15

**Part 6 Other DM Areas** and **Foundations of DM**

Chapters 9 - 10 (13) and Lectures 16, 17

We will also cover, if time allows, in some level of detail the following subjects

Types of Neural Networks

Protein Secondary Structure Prediction - Multiclassifiers

Descriptive Granularity - a Data Mining Model

# GRADING COMPONENTS

During the semester students are responsible for the following (in order as listed)

1. Test1 (70pts)
2. Classification Project (30pts)
3. Test 2 (70pts)
4. Final Report Presentation (20pts)
5. Final Report Paper(10pts)

# TESTS SCHEDULE

**Preliminary Test Schedule**

**TEST 1** March 7

Spring Break March 13 - 19

**Project** due March 27 - submit to Blackboard

**TEST 2** April 13

Final Report **Presentation** April 18 - May 2

Final Report **Paper** - due May 5 - submit to Blackboard

# FINAL GRADE COMPUTATION

**NONE of GRADES will be CURVED**

During the semester you can earn 200pts or more (in the case of extra points)

The **% grade** will be determine in the following way:

# of earned points divided by 2 = % grade

The % grade is **translated** into a letter grade in a standard way as follows

100 - 90 % is A range

A (100 - 96%), A- (95- 90%)

89 - 80 % is B range

B- (80 - 82%), B (83 -85%), B+ (86 -89%)

79 - 70 % is C range:

C- (70- 72%), C (73-75%), C+ (76-79%)

69 - 60 % is D range

F is below 60%

## Course Contents

The course will follow the book very closely and in particular we will cover **all** or **parts**of the following chapters and subjects

**The order does not need to be sequential**

### Chapter 1

Introduction and General overview

What is Data Mining, which data, what kinds of patterns can be mined

# Course Contents

**Chapter 2**

Data preprocessing

Data cleaning, data integration and transformation, data reduction, discretization and concept hierarchy generation

**Chapters 3**

Data Warehouse

**Chapter 5**

Mining Association Rules in Large Databases

Transactional databases and Apriori Algorithm

# Course Contents

**Chapter 6**

Classification and Prediction

**1.** Decision Tree Induction ID3, C4.5

**2.** Neural Networks

**3.** Bayesian Classification

**4.** Classification by Association rule mining

**5.** Genetic algorithms

**6.** Statistical Prediction

# Course Contents

**Chapter 7**

Cluster Analysis

A Categorization of major Clustering methods

**Chapter 8**

Mining Sequential Patters in Biological Data

**Chapter 10]**

Text Mining

**Chapter 11**

Foundations of Data Mining