CS 512 Data Mining Principles

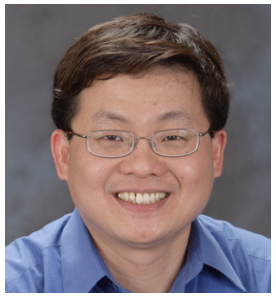
Lecture 1. Introduction

Hanghang Tong, Computer Science, Univ. Illinois at Urbana-Champaign, 2021

Data and Information Systems (DAIS)

- Database Systems
- Data Mining
- Networks
- Text Information Systems
- Healthcare


Jiawei Han


Kevin Chang


Yongjoo Park


Arindam Banerjee


Hari Sundaram


Hanghang Tong


ChengXiang Zhai


Jimeng Sun

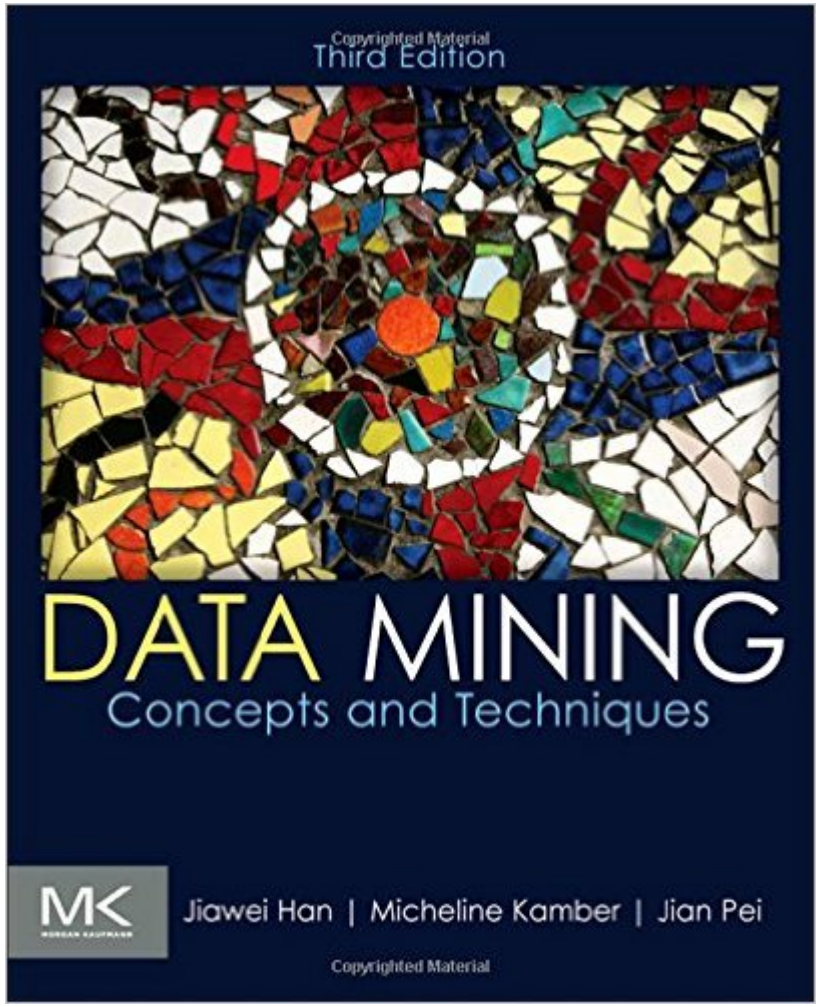
Data and Information Systems (DAIS:) Course Structures at CS/UIUC

- Coverage: Database, data mining, text information systems, Web and bioinformatics
- Data mining
  - Intro. to data mining (CS412)
  - Data mining: Principles and algorithms (CS512)
- Database Systems:
  - Intro. to database systems (CS411)
  - Advanced database systems (CS511)
- Text information systems
  - Text information system (CS410)
  - Advanced text information systems (CS510)

Personnel

- Personnel
  - Instructor: Hanghang Tong (htong@illinois.edu)
  - TAs:
    - Qinghai Zhou (qinghai2@illinois.edu)
    - Zhe Xu (zhexu3@illinois.edu)
- Office hours:
  - Hanghang Tong: 9:00-10:00 am Monday @ zoom
  - Qinghai Zhou: 9:00-9:40am Tuesday; 9:00-9:40pm Thursday
  - Zhe Xu: 8:00-8:40pm Wednesday; 2:00-2:40pm Friday
- All office hours are in Central time zone, on Zooms.
  - Hanghang Tong: https://illinois.zoom.us/j/9635963030; s2020
  - Qinghai Zhou: https://illinois.zoom.us/j/8525819930?pwd=TXBsQWJGdkI0aVQ5a2NGaVBLVWtWZz09; cs512s21
  - Zhe Xu: https://illinois.zoom.us/j/6107341709?pwd=MmRFSXNXN3VDTVJveUlNUXk3aURPQT09; zhe

Course Page & Class Schedule



- Textbook
  - Jiawei Han, Micheline Kamber and Jian Pei, *Data Mining: Concepts and Techniques (3rd ed)*, Morgan Kaufmann, 2011
  - 4th *ed* is coming soon. Pdfs are expected on Compass 2g
- Research Papers + Slides
- Class Homepage:
  - Compass 2g
  - piazza.com/illinois/spring2021/cs512 (pls sign up)
- **Class Schedule: online this semester**

Lecture Schedule (subject to change)

- Logistics & Intro
- Frequent Pattern Mining
- Classification
- Clustering
- Deep Learning
- Outlier & Anomaly Detection
- Graph Connectivity Optimization
- Network Science of Teams
- NoX: Network-as-a-Context
- Network Alignment
- Fair Network Mining

Course Work, Grading & Key Dates

- Two midterm exams: 40% in total (equal weights)
  - First mid-term: 3/15/2021
  - Second mid-term: 4/30/2021
- Two assignments: 30% in total (equal weights)
  - First assignment: 2/1/2021 out; 3/5/2021 due
  - Second assignment: 3/20/2021 out; 4/25/2021 due
- Class project: 30%
  - Proposal (2%): due on 2/25/2021
  - Mid-term report (8%): due on 3/30/2021
  - Final report (20%): due on 5/5/2021
  - Individual project or group (3 members at most per group) project
- Need help and/or discussions?
  - Piazza
- Check course material:
  - lecture notes/grades: Compass 2g

- You can choose Python or Java or C++
- This is NOT a programing class

- We cannot make any changes to these key dates
- We cannot accept any late submissions or arrange make-up exam, except for genuine, verifiable emergence.

- 2% extra credit will be given to the most active users on Piazza, in answering other users' questions.
- Will be given to top 2-5% users based on Piazza statistics
- Posting meaningless content might lead to disqualification of extra credit

Letter Grade Cut-offs

- The following cutoffs represent what will be *likely* used to generate the letter grades:

- **A+ >= 98% A >=94% & < 98% A- >=90% & < 94%**
- **B+ >= 85% & < 90% B >=80% & < 85% B- >=77% & < 80%**
- **C+ >= 74% & < 77% C >=70% & < 74% C- >=67% & <70%**
- **D >= 60%& <67% F < 60%**

- The above cutoffs are tentatively and may be adjusted *slightly*; However, there will be *no general curve-fitting* in assigning the final grades.
- If there is any adjustment of the above cutoffs, we will NOT curve down your letter grades.

CS512 vs. CS412 vs. CS598 (Network Mining)

- Logistics & Intro
- Frequent Pattern Mining
- Classification
- Clustering
- Deep Learning
- Outlier & Anomaly
- Graph Connectivity Optimization
- Network Science of Teams
- NoX: Network-as-a-Context
- Network Alignment
- Fair Network Mining

CS512: this course

- Introduction
- Know Your Data
- Data Preprocessing
- Data Warehousing & OLAP
- Data Cube Technology
- Frequent Patterns: Basics
- Frequent Patterns: Advanced Methods
- Classification: Basic Concepts
- Classification: Advanced Methods
- Cluster Analysis: Basic Concepts
- Deep Learning

CS412: fall 2020

- Introduction
- Patterns, Laws and Generators
- Ranking
- Matrix and tensor
- Anomaly detection
- Embedding
- Network connectivity
- Network science of teams
- Network alignment
- Network of networks
- Network mining auditing

CS598 (Network Mining) spring 2020

Class Policies

- Assignments:
  - The homework is due at **11:59 PM CT** on the due date.
  - We will be using Compass (http://compass2g.illinois.edu) for collecting homework assignments.
  - Please do not hand in a scan of your handwritten solution, only the typed solution (e.g., Microsoft Word, Latex, etc.) will be graded.
  - Contact the TAs if you are having technical difficulties in submitting the assignment. We do NOT accept late homework.
  - The homework should be submitted as a single pdf file using the name convention: yourFirstName-yourLastName.pdf.
  - Feel free to discuss with other members of the class when doing the homework. You should, however, write down your own solution **independently**.

Class Policies (cont.)

- **Academic Integrity Policy**:
  - We have **zero tolerance** on any violation
- **Assuring Non-Hostile Work Environment**
  - In order to assure a non-hostile work environment for course staff, we will strictly enforce the following policy for the future assessment, including exams, assignments and course project. Any assessment containing language that conventionally would be judged as obscene, threatening violence, or of a clearly derogatory nature will be given a 0 without further grading.

A note on (Asynchronous) Online Course

- Many many challenges
- Communication will be the key
- Together, we will make (survive and thrive) it.