University of Colorado
Boulder

CSCI 4502/5502
Data Mining

Fall 2020
Lecture 01 (Aug 25)

Agenda

- Introduction: Instructor, class
- Administrative information
- Course overview
- Policies
- Chapter 1: Introduction to Data Mining

Instructor (1)

- Qin (Christine) Lv
- Associate Professor, Associate Co-Chair for Graduate Education
- Department of Computer Science
- Contact information
  - Office: ECCR 1B24 Phone: (303)492-8821
  - Email: Qin.Lv@Colorado.EDU https://www.cs.colorado.edu/~lv

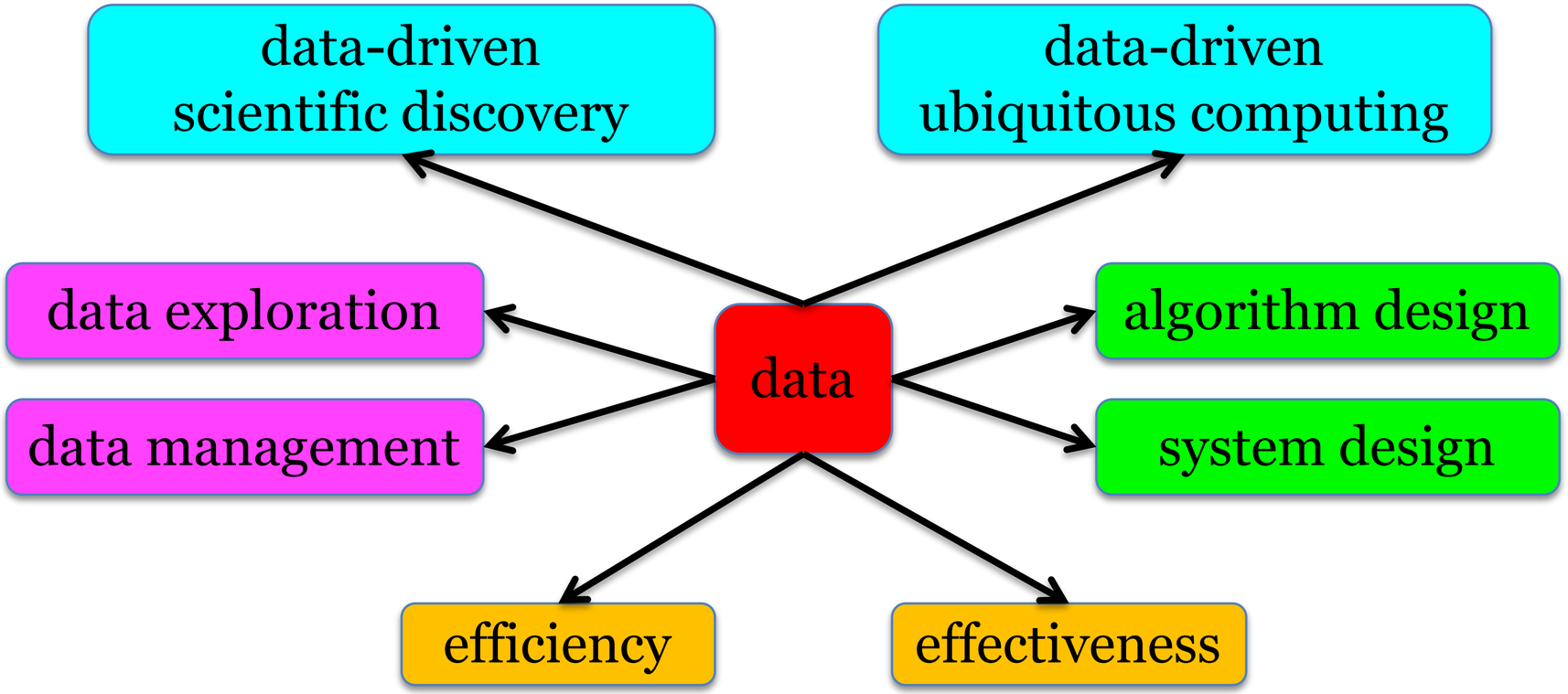
Instructor (2)

- B.E. in Computer Science & Technology
  - Tsinghua University
- MA & Ph.D. in Computer Science
  - Princeton University
- 2008-present: CU Boulder

Instructor (3)

Research Overview

interdisciplinary



Instructor (4)

- Data-driven scientific discovery
  - Earth sciences, environment, transportation, sustainable energy, materials science, seismology, oceanography, …
- Data-driven ubiquitous computing
  - mobile/wearable/IoT sensing/analytics, online social networks, recommender systems, cyber safety, …
- Key research topics: data sensing, data fusion, spatial-temporal data analysis, anomaly detection,…

Instructor (5)

- Sample research projects
  - air quality sensing & analysis, Wi-Fi sensing
  - remote sensing data, cryospheric data, water column sonar data
  - PHEV, transportation
  - solar/wind farm, smart grid
  - Twitter: event analysis, IRA interference, earthquakes
  - recommendation: group, social, stability
  - cybersafety: flashers, cyberbullying, fake news

About the Class

- Class section
- Degree program
- Programming language
- Cloud computing platform
- Social networks
- Types of data

Agenda

- Introduction: Instructor, class
- Administrative information
- Course overview
- Policies
- Chapter 1: Introduction to Data Mining

Administrative Information (1)

- CSCI 4502/ 5502: Data Mining
  - 001: TuTh 9:35-10:50am, ECCR 1B12 (assigned cohort only)
  - 001B: distance section
- Textbook *Data Mining: Concepts and Techniques*
  - Jiawei Han, Micheline Kamber, Jian Pei. 3rd Edition, Morgan Kaufmann, 2011.

Administrative Information (2)

- Course website
  - https://canvas.colorado.edu/courses/65838
  - all lecture slides, videos, homework submission
  - announcements, discussion, project
  - FOR COURSE USE ONLY, DO NOT DISTRIBUTE

Administrative Information (3)

- Zoom meeting ID: 990-9964-0528
  - https://cuboulder.zoom.us/j/99099640528
  - CU Boulder zoom sign in, wait room
  - mute/unmute; video on/off; share screen
  - chat, raise hand, polling, breakout room

Administrative Information (4)

- Instructor: Qin (Christine) Lv *qin.lv@colorado.edu*
  - office hours: Tu 11-12pm, Th 7-8pm, or by appointment
- TA: Tao Ruan *tao.ruan@colorado.edu*
  - Office hours: M 8-9am, W 4-5pm, or by appointment
- GSS: TBD

Agenda

- Introduction: Instructor, class
- Administrative information
- Course overview
- Policies
- Chapter 1: Introduction to Data Mining

Why Data Mining?

- Data, lots of data, and fast increasing
- Discover interesting patterns from data
- Example 1: Market basket analysis: Walmart, Amazon, NetFlix, ...
- Example 2: Cluster analysis, classification: loan application, medical diagnosis, ...
- Example 3: Time series analysis: social network, wind speed, stock, ...
- And a lot more!

Course Summary

- Data mining
  - concepts and techniques; quality vs. efficiency
  - discovering interesting patterns from large amounts of data
- Topics covered
  - data preprocessing, data warehouse, frequent patterns, classification, clustering, outliers
  - complex data, data mining trends

Course Schedule (tentative)

- Week 1 (8/25, 8/27): Introduction
- Week 2 (9/1, 9/3): Data Preprocessing
- Week 3 (9/8, 9/10): Data Warehouse
- Week 4 (9/15, 9/17): Frequent Patterns
- Week 5 (9/22, 9/24): Classification
- Week 6 (9/29, 10/1): Project Proposal
- Week 7 (10/6, 10/8): Classification
- Week 8 (10/13, 10/15): Clustering

Course Schedule (tentative)

- Week 9 (10/20, 10/22): Clustering
- Week 10 (10/27, 10/29): Midterm Review & Exam
- Week 11 (11/3, 11/5): Outlier Detection
- Week 12(11/10, 11/12): Project Checkpoint
- Week 13 (11/17, 11/19): Complex Data
- Week 14 (11/24, 11/26): Exam review, no class on Thanksgiving
- Week 15 (12/1, 12/3): Data Mining Trends
- Week 16 (12/8, 12/10): Project Final Report

Policies

- Classroom Behavior
- Requirements for COVID-19
- Accommodation for Disabilities
- Preferred Student Names and Pronouns
- Honor Code
- Sexual Misconduct, Discrimination, Harassment and/or Related Retaliation
- Religious Holidays

Academic Integrity

- WORK ALONE, unless instructed explicitly as a group assignment
- All submitted work should include the Honor Code Pledge
- Properly acknowledge other people's work, including information you find on the Web
- Cheating or plagiarism will NOT be tolerated!

Grading

- Homework assignments (35%) (work alone)
- Midterm exam (25%) (work alone)
- Course project (40%)
- Late submission
  - at most 2-day delay, 20-point penalty each day

Course Project (40%)

- A self-contained project related to this course's topics
- Team of 3-4 students
  - check with instructor for smaller or larger groups
  - can mix students in different sections
- Pick your own project idea
- Discuss project ideas with instructor, TA, and others

Project Proposal

- Week 6 (9/29, 10/1)
- Submit a project proposal (~3 pages)
  - motivation: why this problem?
  - literature survey: what has been done before?
  - proposed work: what do you plan to do?
  - evaluation: what metrics? how to claim success?
  - milestones: when to accomplish what?

Project Checkpoint

- Week 12 (11/10, 11/12)
- Submit a progress report (~6 pages)
  - updated, extended version of initial proposal, highlight progresses
  - proposal review: motivation, proposed work, evaluation, milestones
  - what have been accomplished so far?
  - what remains to be done?

Final Project Report

- Week 16 (12/8, 12/10)
- Follow the format of regular research papers (10-12 pages)
  - title, authors' information, abstract
  - introduction, related work
  - main technique, evaluation
  - conclusions, future work, references

Final Project Presentation

- Week 16 (12/8, 12/10)
- A 10-minute presentation
  - motivation, literature survey, your work, evaluation, conclusions, future work
  - technical depth, evaluations, clarity, style
- Also submit source code & key results
- Contributions by individual team members

University of Colorado
Boulder

Chapter 1
Introduction to Data Mining