University of Colorado
Boulder

CSCI 4502/5502

Data Mining

Fall 2019
Lecture 01 (Aug 27)

Agenda

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

Instructor (1)

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

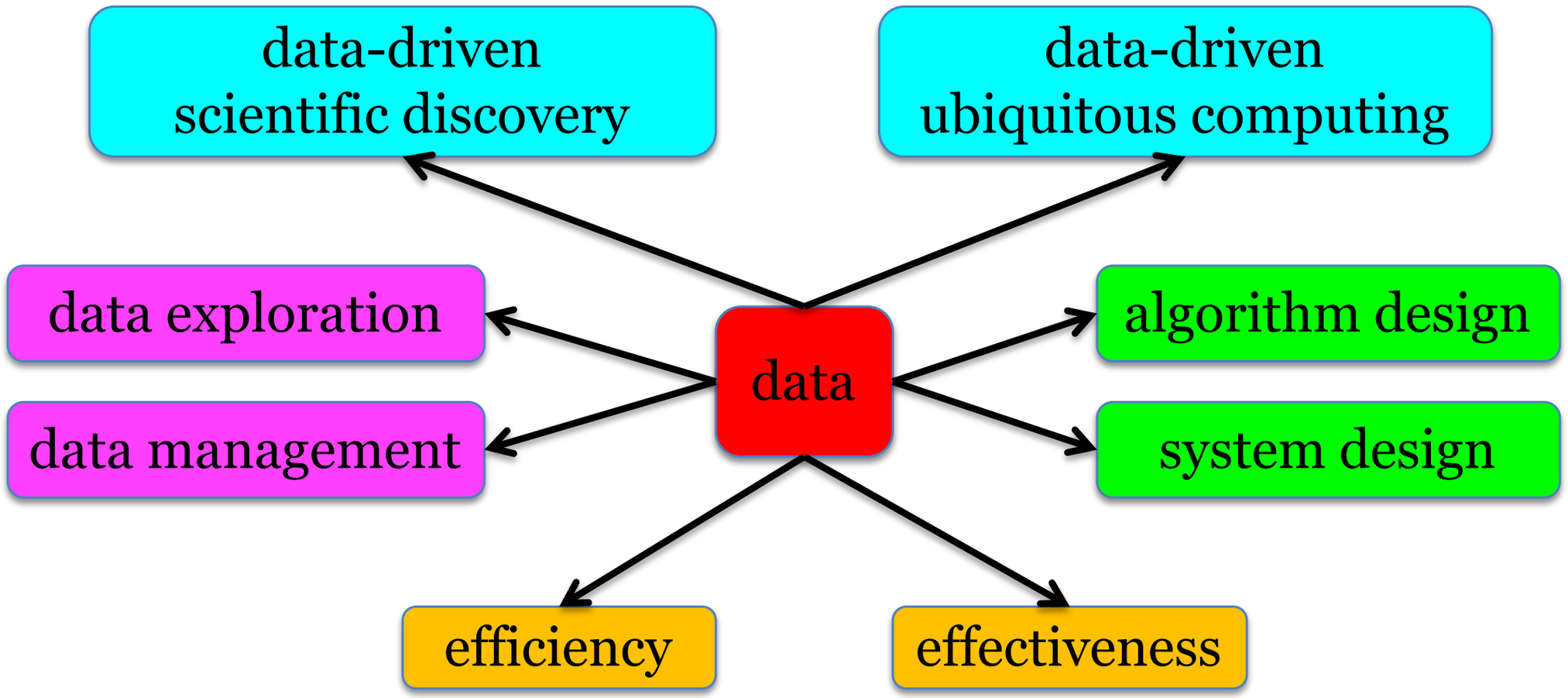
Instructor (2)

- 2000: Tsinghua University, Beijing, China
  - B.E. in Computer Science & Technology
- 2006: Princeton University
  - Ph.D. in Computer Science
- 2006-2007: Princeton University, Postdoc
- 2007-2008: SUNY Stony Brook, Asst. Prof.
- 2008-present: CU Boulder

Instructor (3)

Research Overview

interdisciplinary



Instructor (4)

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

Instructor (5)

- Sample research projects
  - air quality sensing & analysis
  - remote sensing, cryospheric data
  - PHEV, transportation electrification
  - solar farm, wind farm, smart grid
  - Twitter: event analysis, user location
  - recommendation: group, social, stability
  - cybersafety: flashers, cyberbullying

About the Class

- CSCI 4502 / 5502? on campus? distance?
- Undergrad? Master's? PhD?
- Linear algebra/Statistics/Probability? ML?
- C/C++? Java? Python? R? MATLAB? DB? Cloud services (AWS, GCP, Azure)?
- Facebook? Twitter? LinkedIn? Reddit? YouTube? Instagram? Snapchat?
- Web Data? Business Data? Bioinformatics? Scientific Data? Other?

Agenda

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

Administrative Information (1)

- CSCI 4502/ 5502: Data Mining
  - 001: TTh 9:30-10:45am, HUMN1B80
  - 001B: distance section

- Textbook
  - Jiawei Han, Micheline Kamber, Jian Pei
    Data Mining: Concepts and Techniques,
    3rd Edition, Morgan Kaufmann, 2011.

Administrative Information (2)

- Course website
  - http://moodle.cs.colorado.edu
  - enrollment key: SIGKDD2019
  - all lecture slides
  - homework submission
  - forums: news, discussion, project
  - FOR COURSE USE ONLY, DO NOT DISTRIBUTE

Administrative Information (3)

- Zoom meeting ID: 555-163-425
  - https://cuboulder.zoom.us/j/555163425
  - zoom app
  - phone: +1 669-900-6833 or 646-558-8656

- Lecture Videos
  - Access via canvas.colorado.edu

Administrative Information (4)

- Instructor: Qin (Christine) Lv
  - office: ECCR 1B24
  - tel: (303)492-8821 fax: (303)492-2844
  - e-mail: *qin.lv@colorado.edu*
    - office hours
      - Tu 11am-12pm, Fr 1pm-2pm, or by appt
- TA: Yichen Wang *yichen.wang@colorado.edu*
- CM, 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
  - discovering interesting patterns from large amounts of data
    - quality vs. efficiency
- Topics covered
  - data preprocessing, data warehouse, frequent patterns, classification, clustering
  - time-series, social networks, Web data, ...

Course Schedule (tentative)

- Week 1 (8/27, 8/29): Introduction
- Week 2 (9/3, 9/5): Data Preprocessing
- Week 3 (9/10, 9/12): Data Warehouse
- Week 4 (9/17, 9/19): Frequent Patterns
- Week 5 (9/24, 9/26): Classification
- Week 6 (10/1, 10/3): Classification
- Week 7 (10/8, 10/10): Project Proposal
- Week 8 (10/15, 10/17): Clustering

Course Schedule (tentative)

- Week 9 (10/22, 10/24): Clustering
- Week 10 (10/29, 10/31): Midterm Review & Exam
- Week 11 (11/5, 11/7): Outlier Detection
- Week 12(11/12, 11/14): Project Checkpoint
- Week 13 (11/19, 11/21): Data Streams, Time-Series
- Week 14 (11/26, 11/28): Fall Break
- Week 15 (12/3, 12/5): Graphs, Social Networks
- Week 16 (12/10, 12/12): Project Final Report

Policies

- http://www.cs.colorado.edu/~lv/teach/policy.html
- Academic integrity
  - Honor Code Pledge
- Classroom behavior
  - Disability
  - Discrimination and harassment
  - Religious observance

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)
  - submit online at moodle
- Midterm exam (25%) (work alone)
  - in-class, closed book exam
- Course project (40%)
- Late submission
  - at most 2-day delay, 20-point penalty/day

Course Project (40%)

- A self-contained project related to this course's topics
- Team of 3-4 students
  - other group size requires instructor's permission
  - can mix students in different sections
- Pick your own project idea
- Discuss project ideas with the instructor and other students

Project Proposal

- Week 7 (10/8, 10/10)
- Project announcement, peer discussion
- Submit a project proposal (~3 pages)
  - motivation
  - literature survey
  - proposed work
  - how to evaluate
  - milestones

Project Checkpoint

- Week 12 (11/12, 11/14)
- Submit a progress report (~6 pages)
  - updated, extended version of initial proposal, highlight progresses
  - proposal review: motivation, proposed work, evaluation, milestones
  - what you have achieved so far
  - what remains to be done

Final Project Report

- Week 16 (12/10, 12/12)
- 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

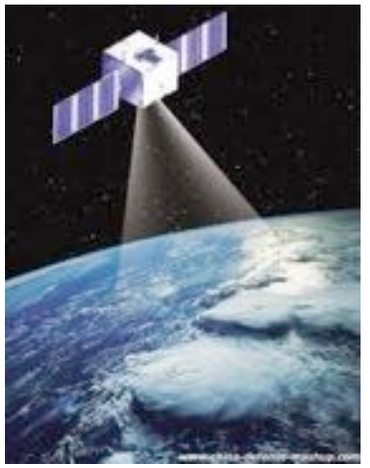
- Week 16 (12/10, 12/12)
- A 10-minute presentation
  - motivation, literature survey, your work, evaluation, conclusions, future work
  - technical depth, evaluations, clarity, style
- Also submit source code & key results
- Contributions by individual team members

Chapter 1:
Introduction to Data Mining

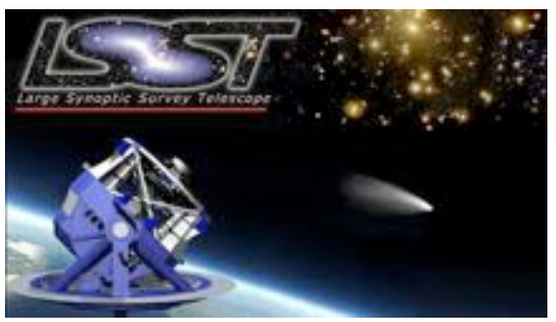
Into the Digital Era

- People's daily lives
  - 4 billion Internet users
  - 500 million tweets/day
- Scientific discovery
  - LHC: 15 PB annually
  - LSST: 20 TB nightly
- IDC Digital Universe Report
  - 0.8ZB (2009) => 35ZB (2020)









flickr

twitter

YouTube
Broadcast Yourself™

facebook

Slashdot

Why Data Mining?

- Data explosion: KB, MB, GB, TB, PB, EB, ...
  - data collection and data availability: automated data collection tools, database systems, Web, computerized society
    - major sources of abundant data
      - business, science, society and everyone
- We are drowning in data, but starving for knowledge!
- Need automated analysis of massive data

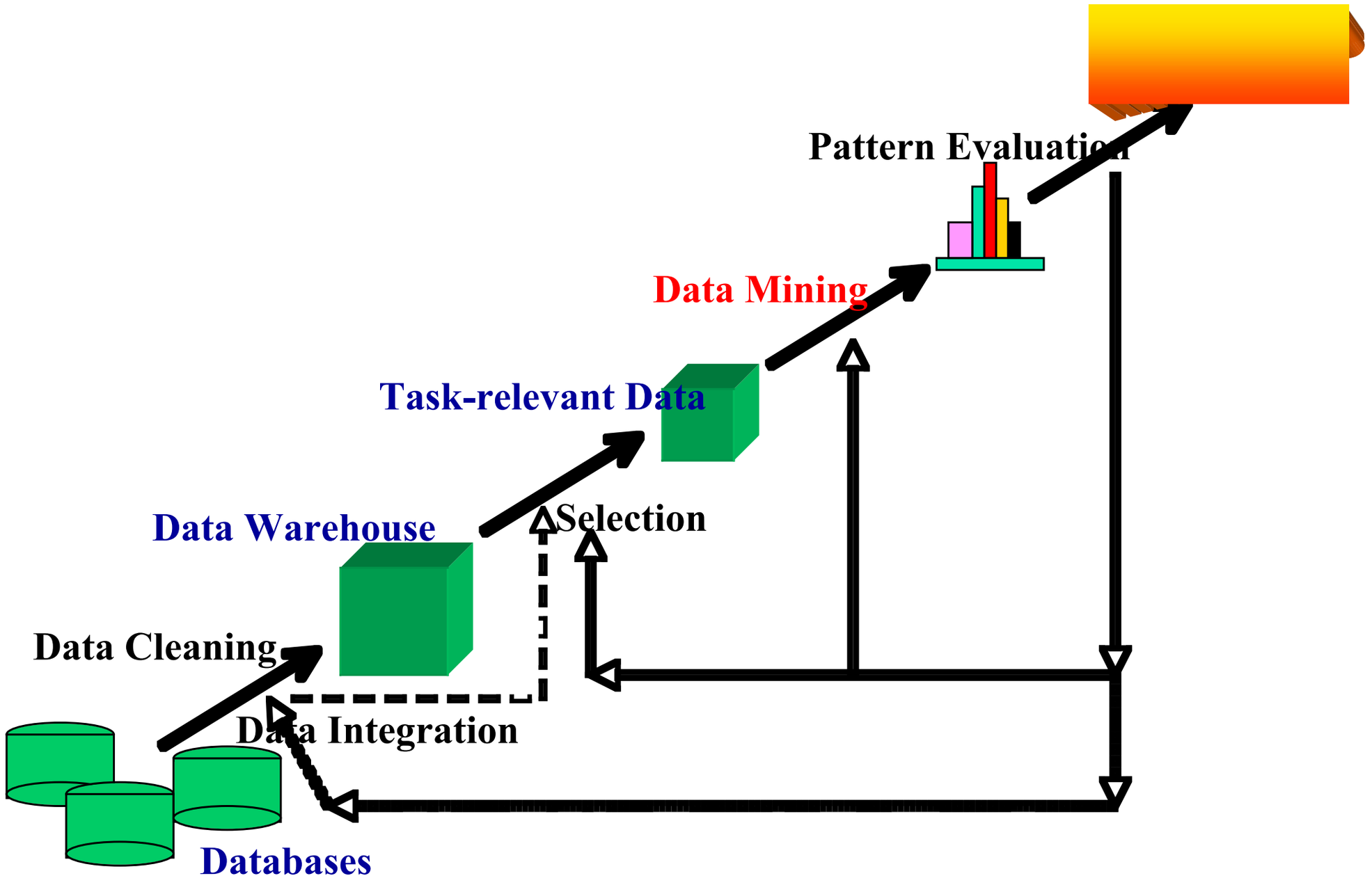
What Is Data Mining?

- Data mining (knowledge discovery from data)
  - extraction of interesting patterns or knowledge from huge amount of data
  - interesting
    - non-trivial, implicit, previously unknown, and potentially useful
  - huge amount of data
    - scalability, efficiency

DM Application Areas

- Science
  - astronomy, bioinformatics, drug discovery, ...
- Business
  - fraud detection, targeted marketing, ...
- Web
  - search engines, advertising, ...
- Government
  - surveillance, crime detection, ...
- ...

Knowledge Discovery



Data Mining: Various Views

- Data view
  - kinds of data to be mined
- Knowledge view
  - kinds of knowledge to be discovered
- Method view
  - kinds of techniques utilized
- Application view
  - kinds of applications adapted

Data View

- The 3Vs, 4Vs, and 5Vs
  - Volume
  - Variety
  - Velocity
  - Veracity
  - Value

Todo List

- Sign for our moodle course
  - http://moodle.cs.colorado.edu
  - enrollment key: SIGKDD2019

- Read Chapter 1

- Start thinking about your project