University of Colorado
Boulder

CSCI 4502/5502

Data Mining

Fall 2019
Lecture 02 (Aug 29)

Announcements

- http://moodle.cs.colorado.edu
  - enrollment key: SIGKDD2019

- Course project
  - start early
  - any language, any platform
  - team: may include ugrad, grad, distance

Chapter 1:
Introduction to Data Mining

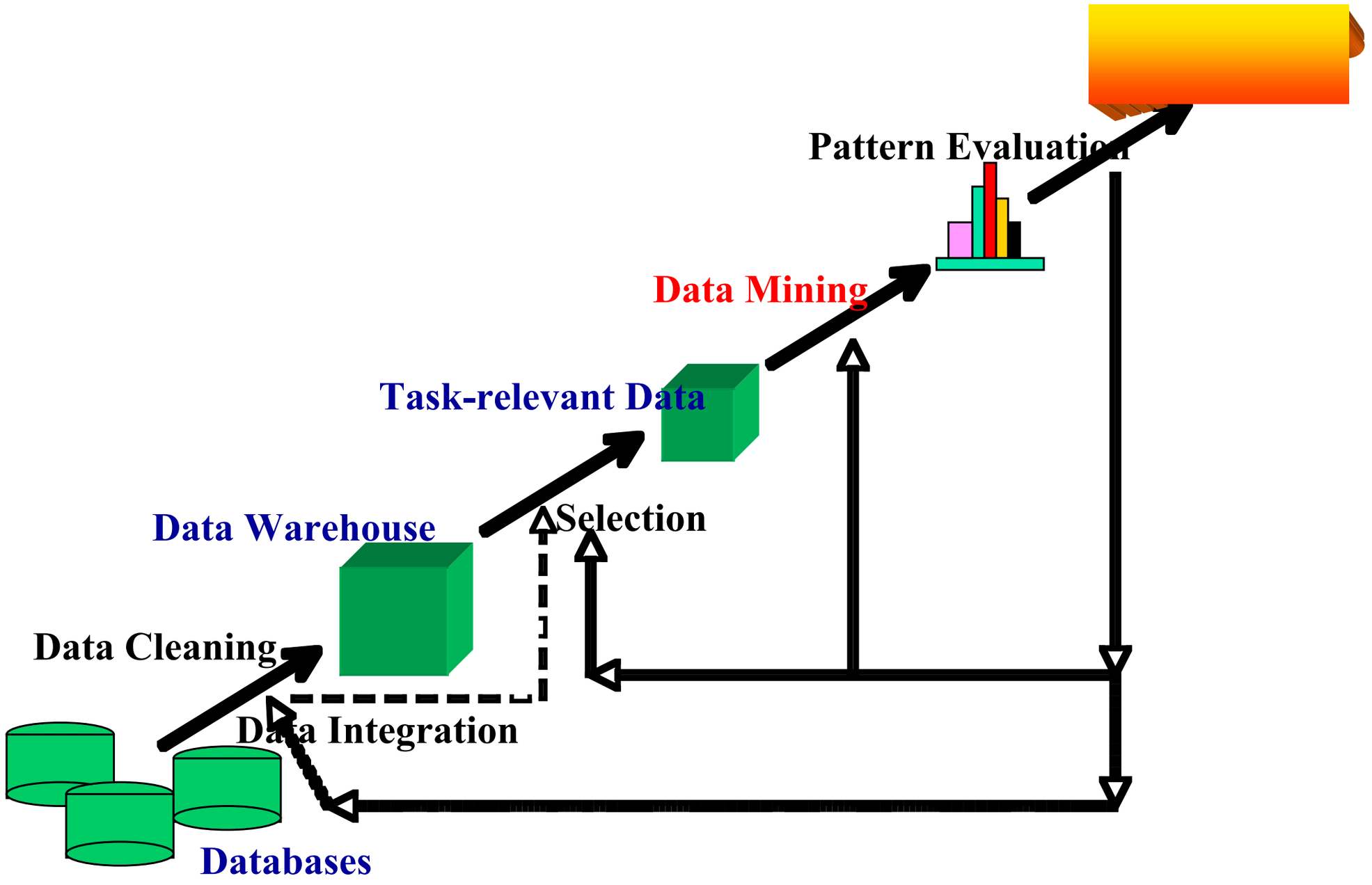
Introduction

- Why data mining
  - data explosion (generation & sharing)
  - data rich but information poor
- What is data mining
  - discovering interesting patterns
  - in huge amounts of data

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 (1)

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

Data View (2)

- Database-oriented
  - relational database
    - (studentID, name, age, gender, major, ...)
  - data warehouse
    - CU: Boulder, Colorado Springs, Denver
  - transactional database
    - (transID, total, item1, item2, item3, ...)

Data View (3)

- Sequence, stream, temporal, time-series data
  - trend analysis, anomaly
- Spatial, spatial-temporal data
- Text, multimedia, Web data
  - topic detection, similarity, popularity
- Graph, social networks data
  - substructures, shared interests

Knowledge View

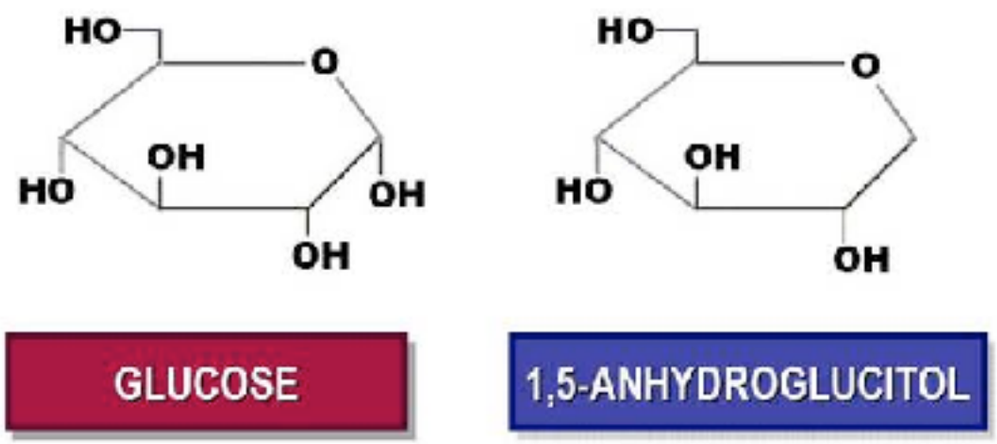
- Concept/class description
- Frequent patterns, associations, correlations
- Classification and prediction
- Cluster analysis
- Outlier analysis
- Evolution analysis

Concept/Class Description

- Data characterization (summarization)
  - customers who spent $1000 a year
  - age 40-50, employed, good credit ratings
- Data discrimination (contrast)
  - frequent vs. infrequent customers
    - age, education, employed,
  - dry vs. wet regions
    - precipitation, temperature, humidity

Frequent Patterns

- Frequent itemsets
  - e.g., (milk, bread), (beer, diaper)
- Frequent sequences
  - e.g., <printer, toner>, <dinner, movie>
- Frequent structures



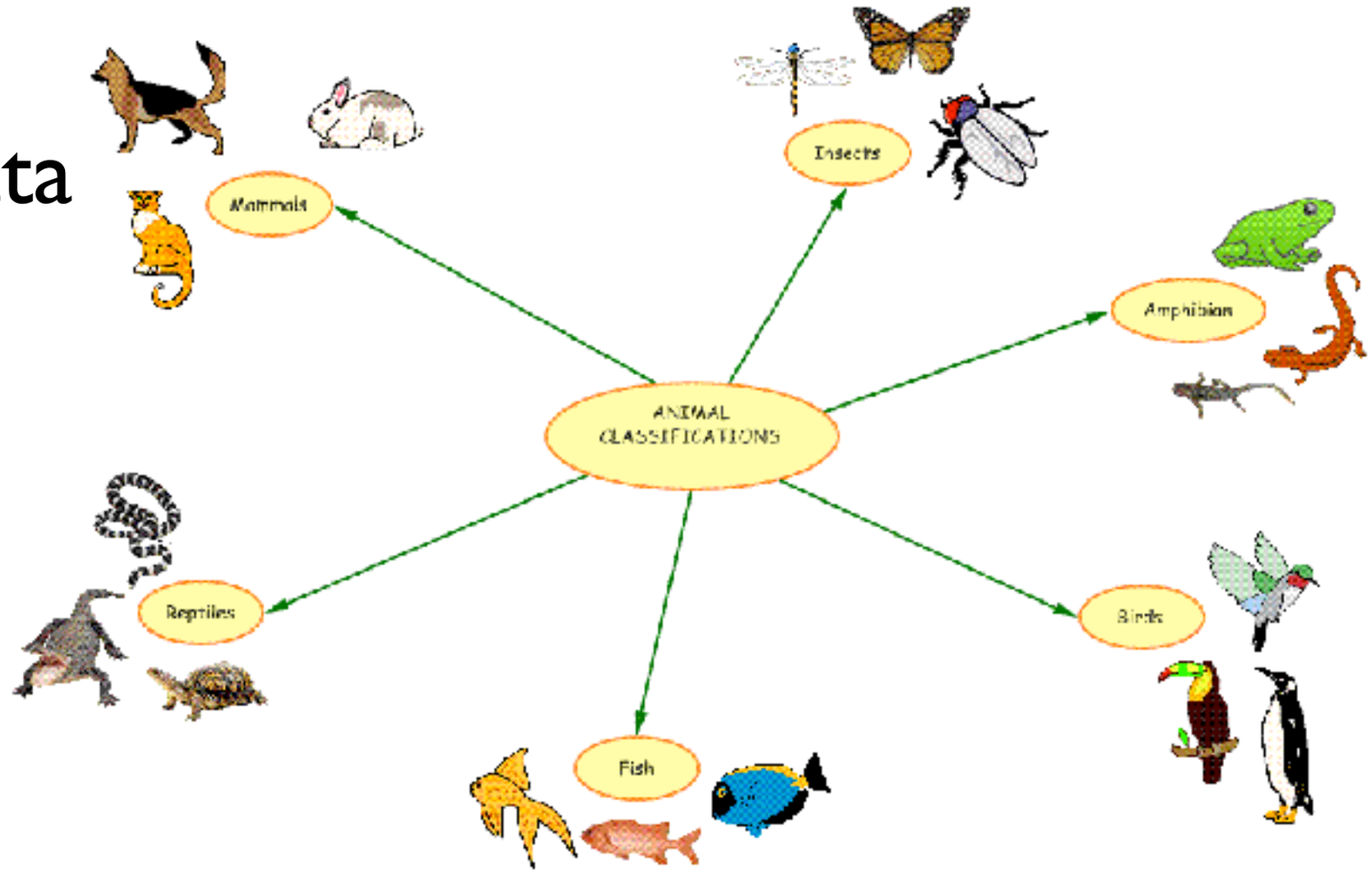
http://www.endotext.org/diabetes/diabetes12/figures/figure12.jpg

Associations

- Association analysis
  - buys (X, milk) => buys (X, bread)
  - [support = 0.5%, confidence = 75%]
- Support
  - chance of A and B appearing together
- Confidence
  - if A appears, chance of B appears
- Minimum support (or confidence) threshold

Classification and Prediction

- Finding a model (or function) that describes and distinguishes data classes or concepts
- Training data

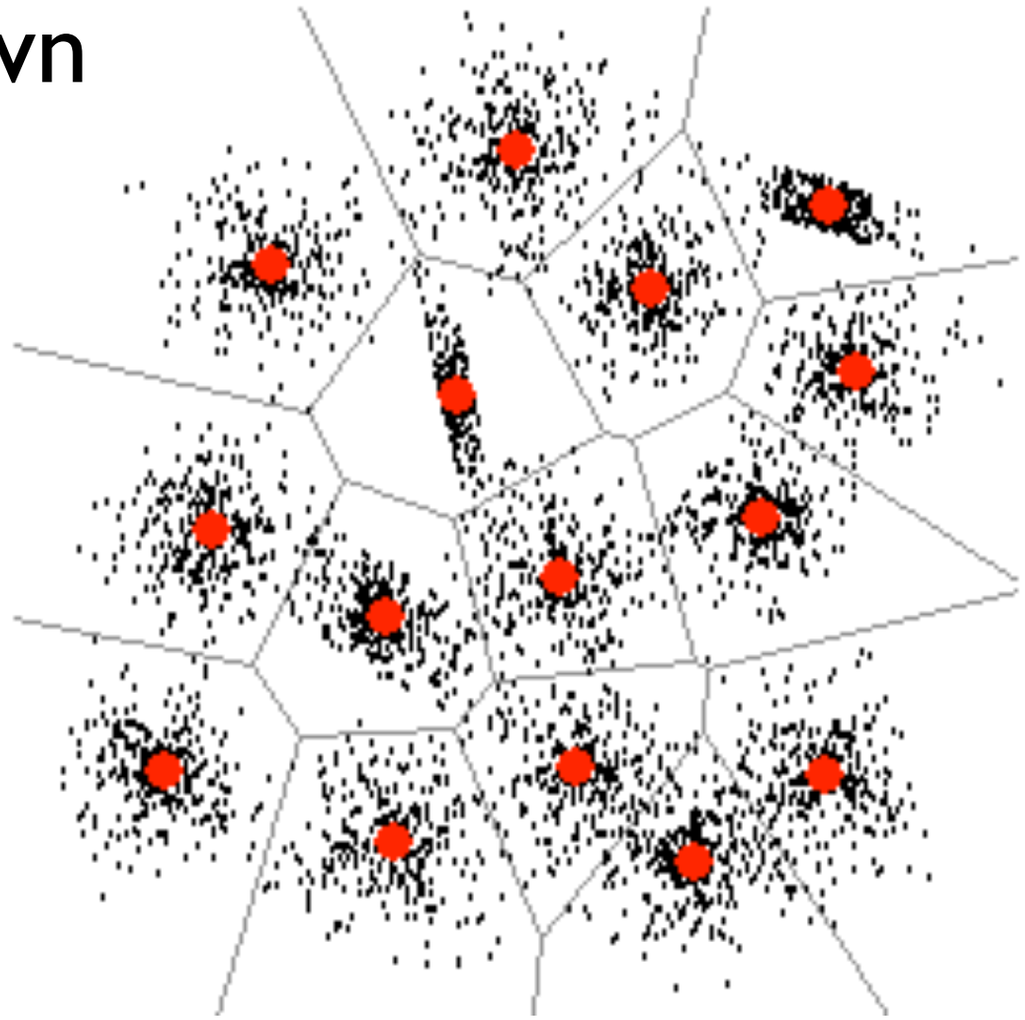


Classification and Prediction

- IF-THEN rules, decision tree, neural network
- Numeric prediction
  - continuous-valued instead of class labels
  - age (young, older), income (low, high)
  - price, brand, place_made => revenue

Cluster Analysis

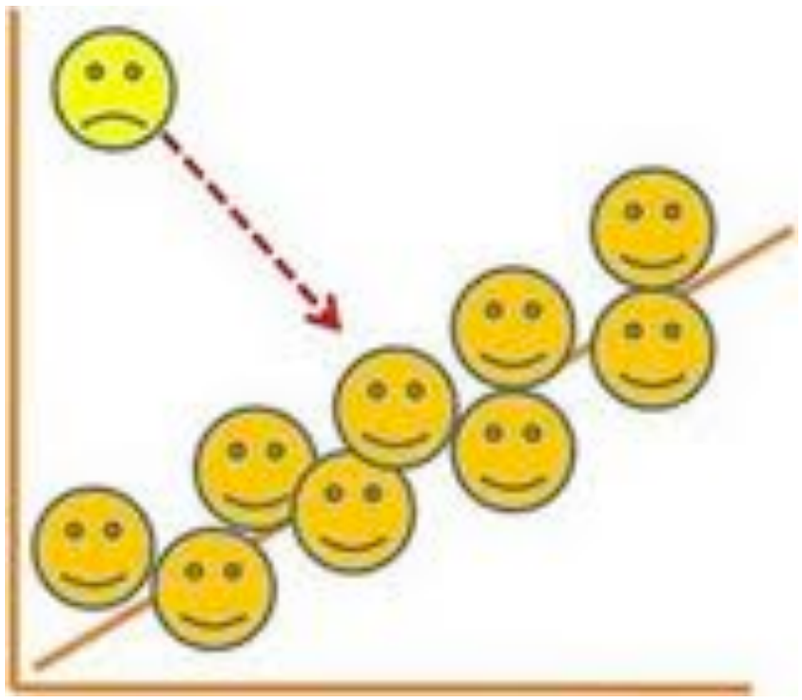
- Class labels unknown
- Intraclass similarity
  - maximize
  - closeness
- Interclass similarity
  - minimize
  - separation
- Hierarchical



www.cs.joensuu.fi/ pages/franti/vq/lkm15.gif

Outlier Analysis

- Outliers
  - do not comply with the general model
- Noise or exception
- Fraud detection, rare event analysis
- E.g. credit fraud analysis



http://dinamehta.com/blog/wp-content/uploads/2007/10/outlier.jpg

Trend and Evolution Analysis

- Trends, deviations
- Sequential pattern mining
  - e.g. digital camera => large SD memory
- Periodicity analysis
- Similarity-based analysis
- E.g. stock data

Market Analysis/Management

- Data sources: credit card transactions, club cards, customer calls, lifestyle studies
- Target marketing
  - find customers with similar patterns: interest, income level, spending habits, etc.
  - determine customer purchasing patterns
- What types of customers buy what products
- What factors attract new customers
- Giving out free samples, discounted products

Fraud Detection, Rare Events

- Approaches: clustering & model construction for frauds, outlier analysis
- Applications: health care, retail, credit card service, telecommunications
  - auto insurance, medical insurance
  - money laundering
  - phone-call fraud
  - retail industry
    - 38% shrink due to dishonest employees

Are the Patterns Interesting?

- Interesting pattern
  - valid on new/test data with some certainty
  - novel
  - potentially useful
    - ultimately understandable by humans
- Objective measures (e.g. support, confidence)
- Subjective measures
- Completeness, exclusiveness

Major Issues in Data Mining

- Mining technology
  - mining different knowledge from diverse data types (maybe noisy or incomplete)
  - pattern evaluation: interestingness
  - efficiency, effectiveness, scalability
  - parallel, distributed, incremental mining
  - incorporation of background knowledge
  - integration of discovered knowledge with existing knowledge: knowledge fusion

Major Issues in Data Mining

- User interaction
  - data mining query languages, ad-hoc mining
  - expression and visualization of results
  - interactive mining at multiple granularities
- Applications and social impacts
  - Domain-specific data mining
  - Applications of data mining results
  - Protect data security, integrity, privacy

Data Mining: A Brief History

- 1989 IJCAI Workshop on Knowledge Discovery in Databases
- 1991-1994 Workshops on Knowledge Discovery in Databases
- 1995-1998 International Conferences on Knowledge Discovery in Databases and Data Mining (KDD'95-98)
  - Journal of Data Mining and Knowledge Discovery (1997)

Data Mining: A Brief History

- ACM SIGKDD conferences since 1998 and SIGKDD Explorations
- More conferences on data mining
  - PAKDD (1997), PKDD (1997), SIAM-Data Mining (2001), IEEE ICDM (2001), etc.
- ACM Transactions on KDD starting in 2007

Confs. and Journals in DM

- KDD conferences
  - KDD, SDM, ICDM, PKDD, PAKDD
- Other related conferences
  - WSDM, CIKM, SIGMOD, VLDB, ICDE, WWW, SIGIR, ICML, CVPR, NIPS
- Journals
  - TKDE, TKDD, DMKD, TPAMI

Summary

- Chapter 1: Introduction to data mining
  - Data mining: discovering interesting patterns in huge amounts of data
  - Knowledge discovery process
  - Different views of data mining
    - data, knowledge, method, application
  - Measure of pattern interestingness
  - Major issues in data mining