# CS6220: DATA MINING TECHNIQUES

## 1: Introduction

---

**Instructor: Yizhou Sun**

yzsun@ccs.neu.edu

September 8, 2014

# Course Information

- Course homepage: http://www.ccs.neu.edu/home/yzsun/classes/2014Fall_CS6220/index.htm
  - Class schedule
  - Slides
  - Announcement
  - Assignments
  - …

- Prerequisites
  - CS 5800 or CS 7800, or consent of instructor
  - More generally
    - You are expected to have background knowledge in data structures, algorithms, basic linear algebra, and basic statistics.
    - You will also need to be familiar with at least one programming language, and have programming experiences.

# Meeting Time and Location

- When
  - Monday, 6-9pm
- Where
  - Shillman Hall 335

# Instructor and TA Information

- Instructor: Yizhou Sun
  - Homepage: http://www.ccs.neu.edu/home/yzsun/
  - Email: yzsun@ccs.neu.edu
  - Office: 320 WVH
  - Office hour: Wednesdays 1-3pm
- TA:
  - Yupeng Gu
    - Email: ypgu@ccs.neu.edu
    - Office hours: Tuesdays 2:30-4:30pm at 472 WVH
  - Kosha Shah
    - Email: shah.ko@husky.neu.edu
    - Office hours: Thursdays 10:00am-12:00pm at 102 Main Lab WVH

# Grading

- Homework: 40%
- Midterm exam: 25%
- Course project: 30%
- Participation: 5%

# Grading: Homework

- Homework: 40%
  - Four assignments are expected
  - Deadline: 11:59pm of the indicated due date via *Blackboard* or class system
    - *No Late Submission!*
  - No copying or sharing of homework!
    - But you can discuss general challenges and ideas with others

# Grading: Midterm Exam

- Midterm exam: 25%
  - Closed book exam, but you can take a "cheating sheet" of A4 size

# Grading: Course Project

- Course project: 30%
  - Group project (3-4 people for one group)
  - Goal: Compete on the assigned course project
  - You are expected to submit a project report and your code at the end of the semester

# Grading: Participation

- Participation (5%)
  - In-class participation
  - quizzes
  - Online participation (piazza)
    - piazza.com/northeastern/fall2014/cs6220

# Textbook

- Jiawei Han, Micheline Kamber, and Jian Pei. Data Mining: Concepts and Techniques, 3rd edition, Morgan Kaufmann, 2011
- References
  - "Data Mining" by Pang-Ning Tan, Michael Steinbach, and Vipin Kumar (http://www-users.cs.umn.edu/~kumar/dmbook/index.php)
  - "Machine Learning" by Tom Mitchell (http://www.cs.cmu.edu/~tom/mlbook.html)
  - "Introduction to Machine Learning" by Ethem ALPAYDIN (http://www.cmpe.boun.edu.tr/~ethem/i2ml/)
  - "Pattern Classification" by Richard O. Duda, Peter E. Hart, David G. Stork (http://www.wiley.com/WileyCDA/WileyTitle/productCd-0471056693.html)
  - "The Elements of Statistical Learning: Data Mining, Inference, and Prediction" by Trevor Hastie, Robert Tibshirani, and Jerome Friedman (http://www-stat.stanford.edu/~tibs/ElemStatLearn/)
  - "Pattern Recognition and Machine Learning" by Christopher M. Bishop (http://research.microsoft.com/en-us/um/people/cmbishop/prml/)

# Goal of the Course

- Know what is data mining and the basic algorithms
- Know how to apply algorithms to real-world applications
- Provide a starting course for research in data mining

# 1. Introduction

- Why Data Mining?
- What Is Data Mining?
- A Multi-Dimensional View of Data Mining
  - What Kinds of Data Can Be Mined?
  - What Kinds of Patterns Can Be Mined?
  - What Kinds of Technologies Are Used?
  - What Kinds of Applications Are Targeted?
- Content covered by this course

# Why Data Mining?

- The Explosive Growth of Data: from terabytes to petabytes
  - Data collection and data availability
    - Automated data collection tools, database systems, Web, computerized society
  - Major sources of abundant data
    - Business: Web, e-commerce, transactions, stocks, …
    - Science: Remote sensing, bioinformatics, scientific simulation, …
    - Society and everyone: news, digital cameras, YouTube
- We are drowning in data, but starving for knowledge!
- "Necessity is the mother of invention"—Data mining—Automated analysis of massive data sets

# 1. Introduction

- Why Data Mining?
- What Is Data Mining? 
- A Multi-Dimensional View of Data Mining
  - What Kinds of Data Can Be Mined?
  - What Kinds of Patterns Can Be Mined?
  - What Kinds of Technologies Are Used?
  - What Kinds of Applications Are Targeted?
- Content covered by this course

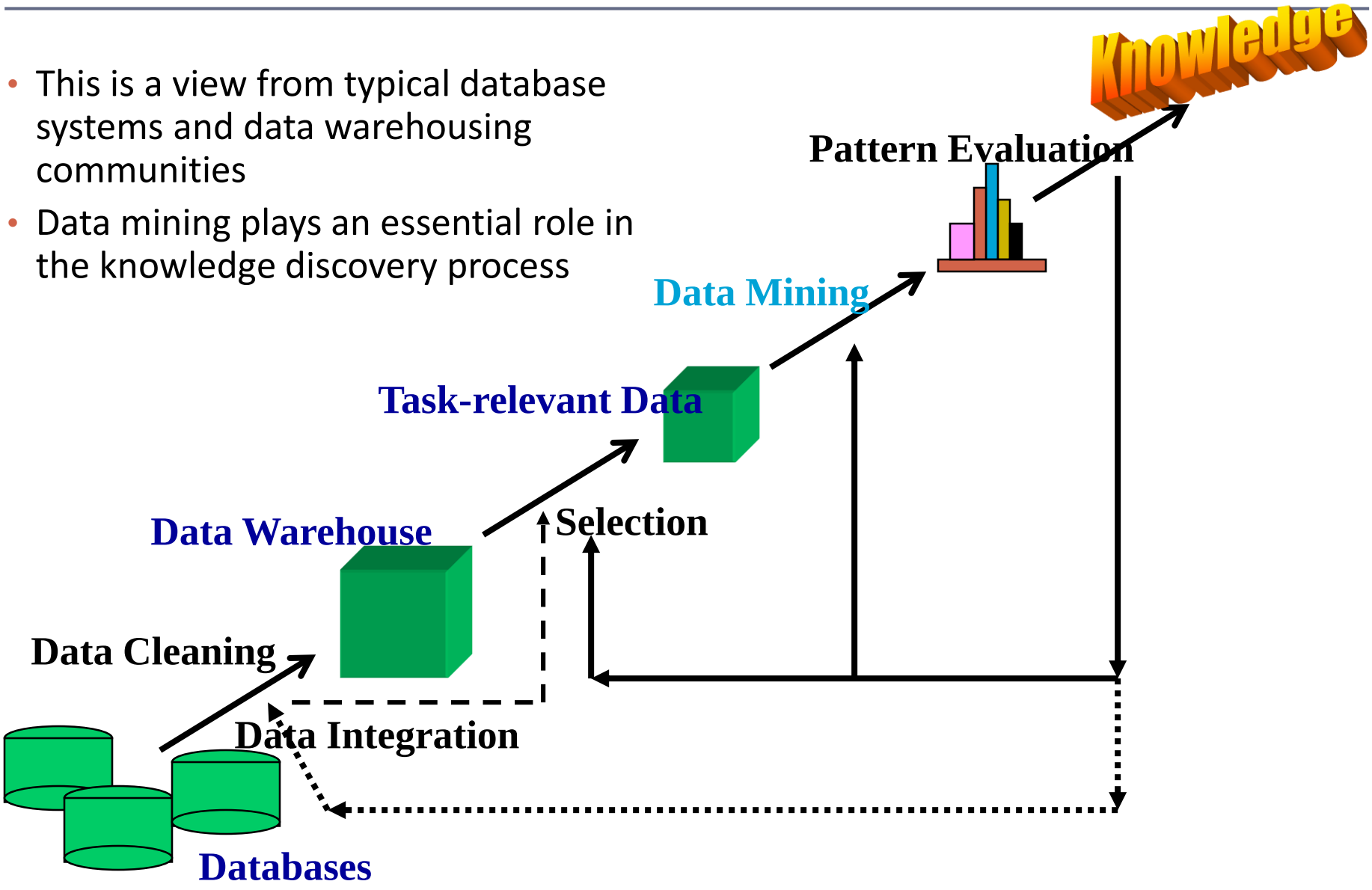
# What Is Data Mining?

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

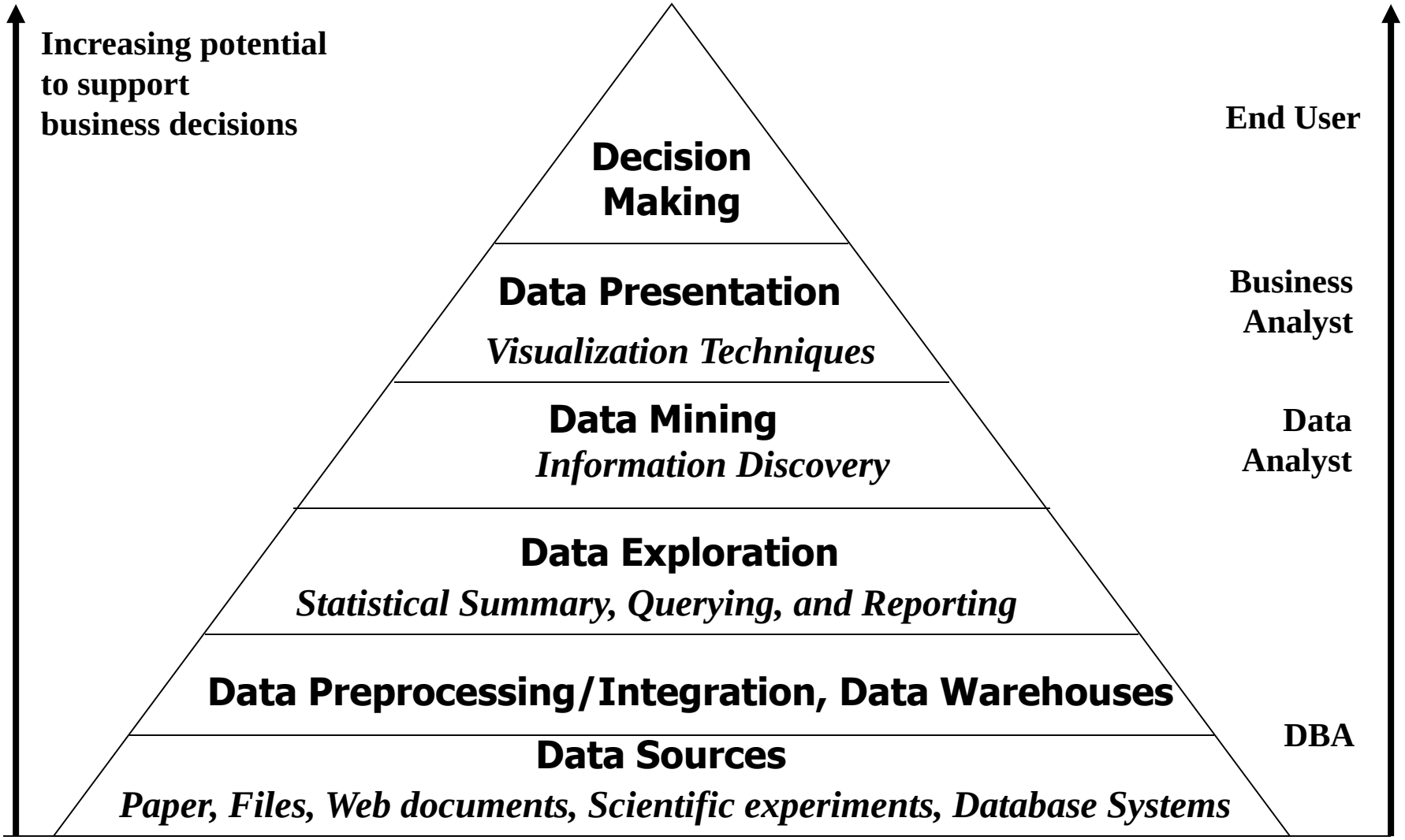
- Alternative names
  - Knowledge discovery (mining) in databases (KDD), knowledge extraction, data/pattern analysis, data archeology, data dredging, information harvesting, business intelligence, etc.

# Knowledge Discovery (KDD) Process

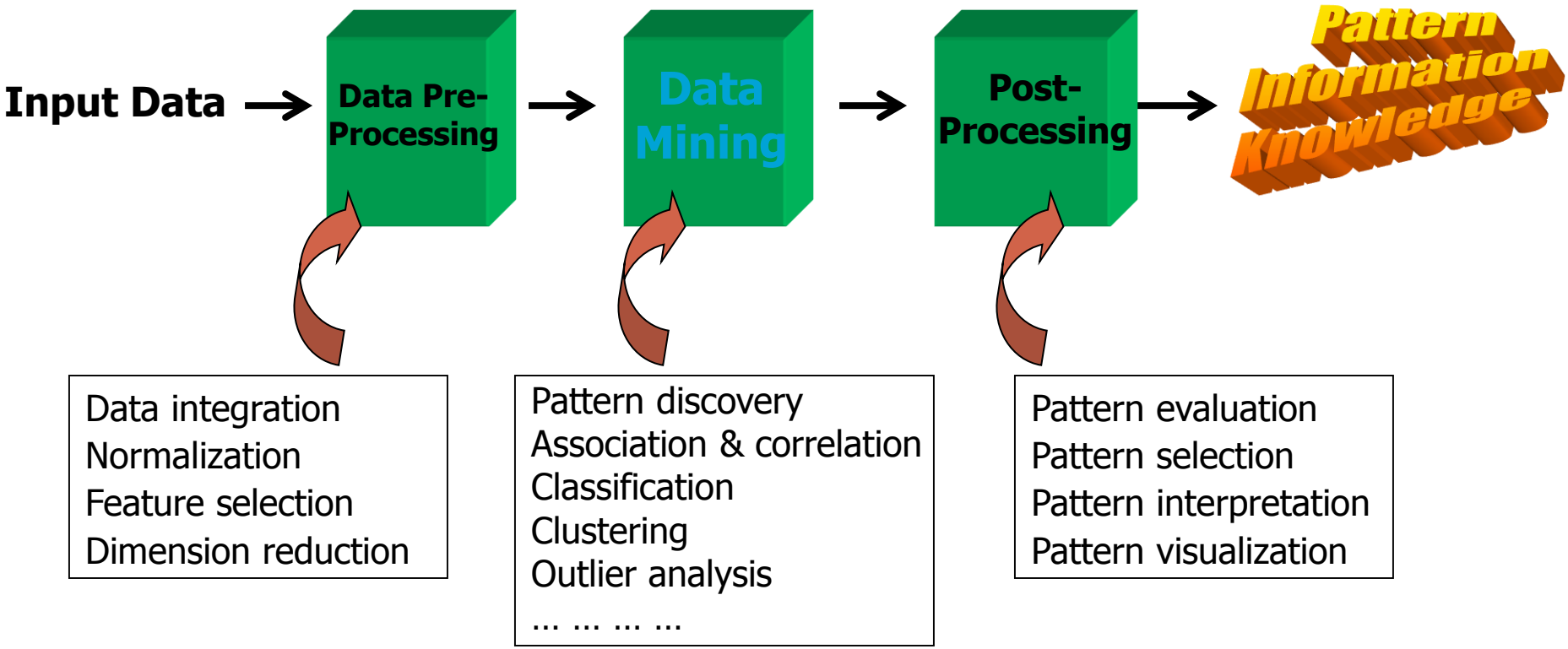
- This is a view from typical database systems and data warehousing communities
- Data mining plays an essential role in the knowledge discovery process



# Data Mining in Business Intelligence



# KDD Process: A Typical View from ML and Statistics




- This is a view from typical machine learning and statistics communities

# 1. Introduction

- Why Data Mining?
- What Is Data Mining?
- A Multi-Dimensional View of Data Mining
  - What Kinds of Data Can Be Mined?
  - What Kinds of Patterns Can Be Mined?
  - What Kinds of Technologies Are Used?
  - What Kinds of Applications Are Targeted?
- Content covered by this course

# Multi-Dimensional View of Data Mining

- **Data to be mined**
  - Database data (extended-relational, object-oriented, heterogeneous, legacy), data warehouse, transactional data, stream, spatiotemporal, time-series, sequence, text and web, multi-media, graphs & social and information networks
- **Knowledge to be mined (or: Data mining functions)**
  - Characterization, discrimination, association, classification, clustering, trend/deviation, outlier analysis, etc.
  - Descriptive vs. predictive data mining
  - Multiple/integrated functions and mining at multiple levels
- **Techniques utilized**
  - Data-intensive, data warehouse (OLAP), machine learning, statistics, pattern recognition, visualization, high-performance, etc.
- **Applications adapted**
  - Retail, telecommunication, banking, fraud analysis, bio-data mining, stock market analysis, text mining, Web mining, etc.

# 1. Introduction

- Why Data Mining?
- What Is Data Mining?
- A Multi-Dimensional View of Data Mining
  - What Kinds of Data Can Be Mined?
  - What Kinds of Patterns Can Be Mined?
  - What Kinds of Technologies Are Used?
  - What Kinds of Applications Are Targeted?
- Content covered by this course

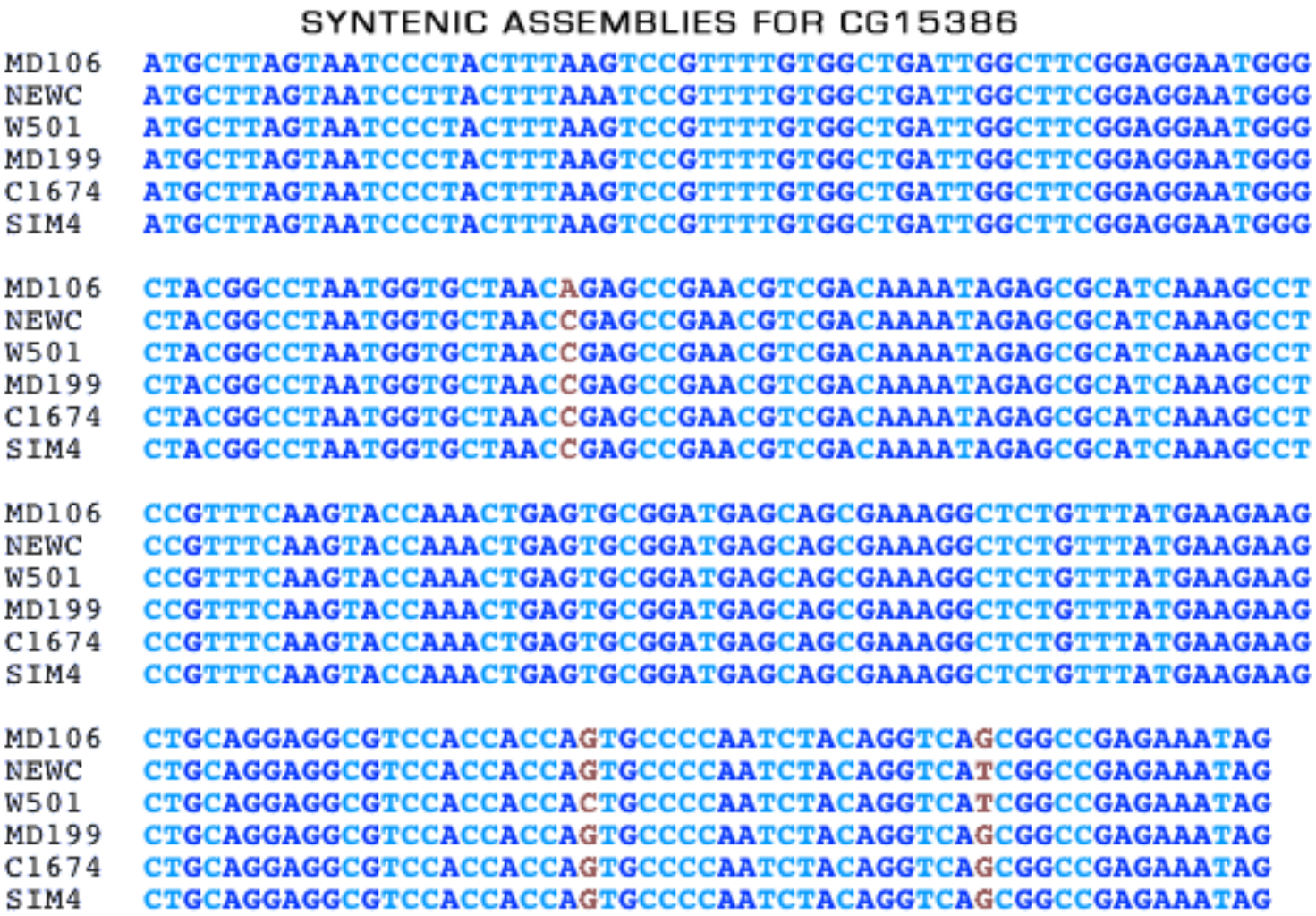
# Matrix Data

| | Sex | Race | Height | Income | Marital Status | Years of Educ. | Liberal-ness |
|---|---|---|---|---|---|---|---|
| R1001 | M | 1 | 70 | 50 | 1 | 12 | 1.73 |
| R1002 | M | 2 | 72 | 100 | 2 | 20 | 4.53 |
| R1003 | F | 1 | 55 | 250 | 1 | 16 | 2.99 |
| R1004 | M | 2 | 65 | 20 | 2 | 16 | 1.13 |
| R1005 | F | 1 | 60 | 10 | 3 | 12 | 3.81 |
| R1006 | M | 1 | 68 | 30 | 1 | 9 | 4.76 |
| R1007 | F | 5 | 66 | 25 | 2 | 21 | 2.01 |
| R1008 | F | 4 | 61 | 43 | 1 | 18 | 1.27 |
| R1009 | M | 1 | 69 | 67 | 1 | 12 | 3.25 |

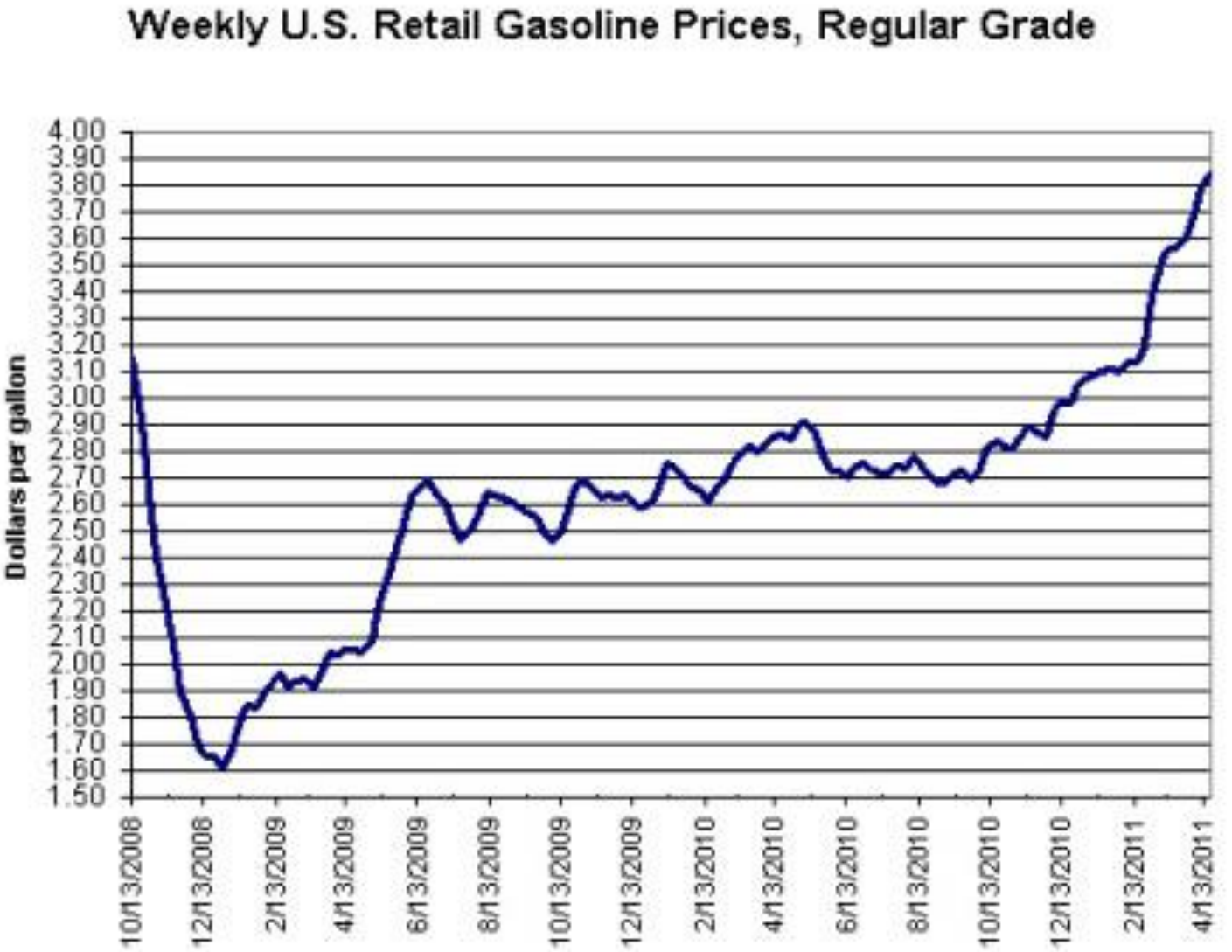
# Set Data

| *TID* | *Items* |
|---|---|
| 1 | Bread, Coke, Milk |
| 2 | Beer, Bread |
| 3 | Beer, Coke, Diaper, Milk |
| 4 | Beer, Bread, Diaper, Milk |
| 5 | Coke, Diaper, Milk |

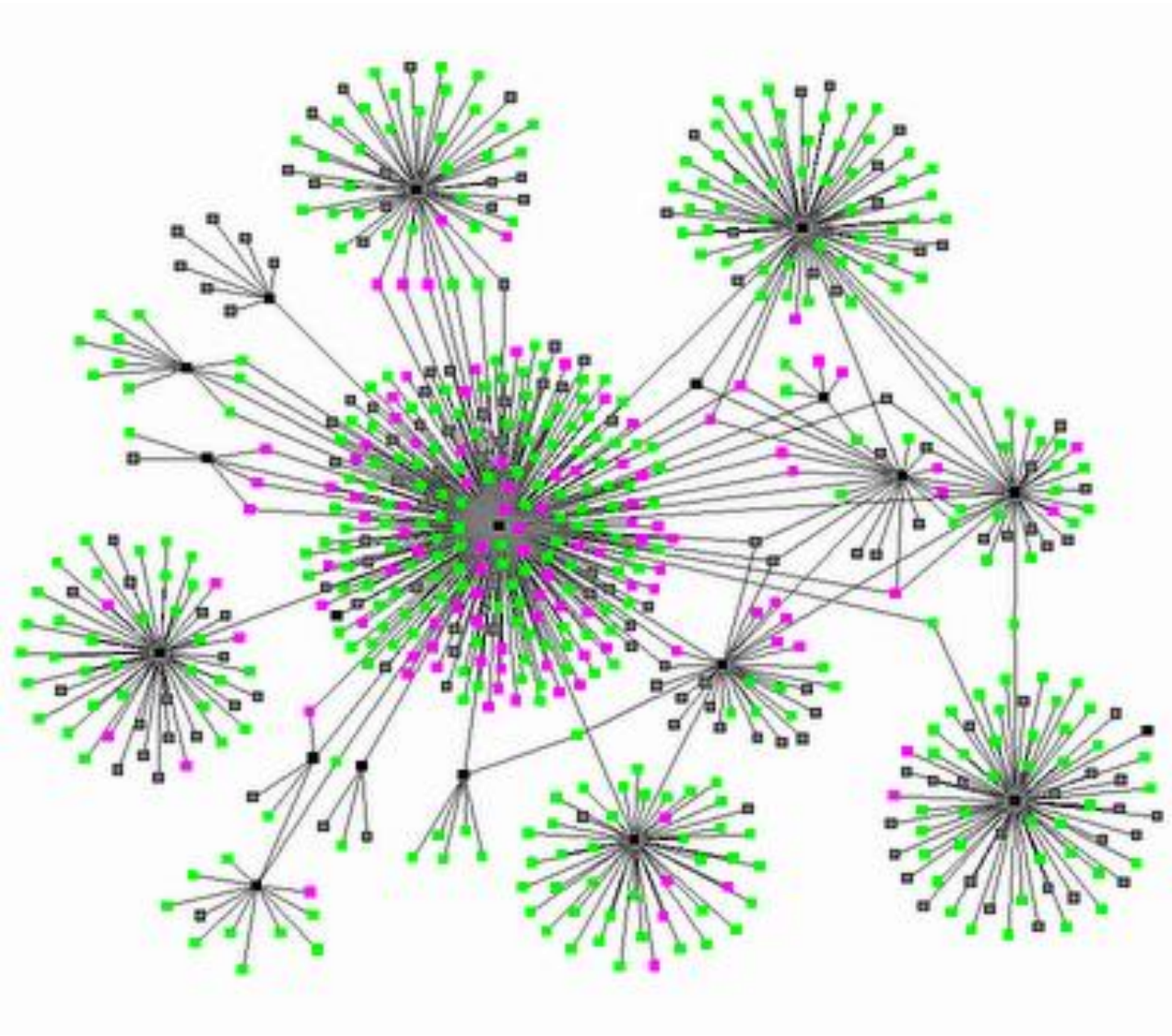
# Sequence Data



# Time Series


Weekly U.S. Retail Gasoline Prices, Regular Grade
Dollars per gallon
4.00
3.90
3.80
3.70
3.60
3.50
3.40
3.30
3.20
3.10
3.00
2.90
2.80
2.70
2.60
2.50
2.40
2.30
2.20
2.10
2.00
1.90
1.80
1.70
1.60
1.50
10/13/2008
12/13/2008
2/13/2009
4/13/2009
6/13/2009
8/13/2009
10/13/2009
12/13/2009
2/13/2010
4/13/2010
6/13/2010
8/13/2010
10/13/2010
12/13/2010
2/13/2011
4/13/2011
Source: Energy Information Administration

# Graph / Network



# 1. Introduction

- Why Data Mining?
- What Is Data Mining?
- A Multi-Dimensional View of Data Mining
  - What Kinds of Data Can Be Mined?
  - What Kinds of Patterns Can Be Mined?
  - What Kinds of Technologies Are Used?
  - What Kinds of Applications Are Targeted?
- Content covered by this course

# Data Mining Function: Association and Correlation Analysis

- Frequent patterns (or frequent itemsets)
  - What items are frequently purchased together in your Walmart?
- Association, correlation vs. causality
  - A typical association rule
    - Diaper → Beer [0.5%, 75%] (support, confidence)
  - Are strongly associated items also strongly correlated?

# Data Mining Function: Classification

- Classification and label prediction
  - Construct models (functions) based on some training examples
  - Describe and distinguish classes or concepts for future prediction
    - E.g., classify countries based on (climate), or classify cars based on (gas mileage)
  - Predict some unknown class labels
- Typical methods
  - Decision trees, naïve Bayesian classification, support vector machines, neural networks, rule-based classification, pattern-based classification, logistic regression, ...
- Typical applications:
  - Credit card fraud detection, direct marketing, classifying stars, diseases, web-pages, ...

# Data Mining Function: Cluster Analysis

- Unsupervised learning (i.e., Class label is unknown)
- Group data to form new categories (i.e., clusters), e.g., cluster houses to find distribution patterns
- Principle: Maximizing intra-class similarity & minimizing interclass similarity
- Many methods and applications

# Data Mining Functions: Others

- Prediction
- Similarity search
- Ranking
- Outlier detection
- …

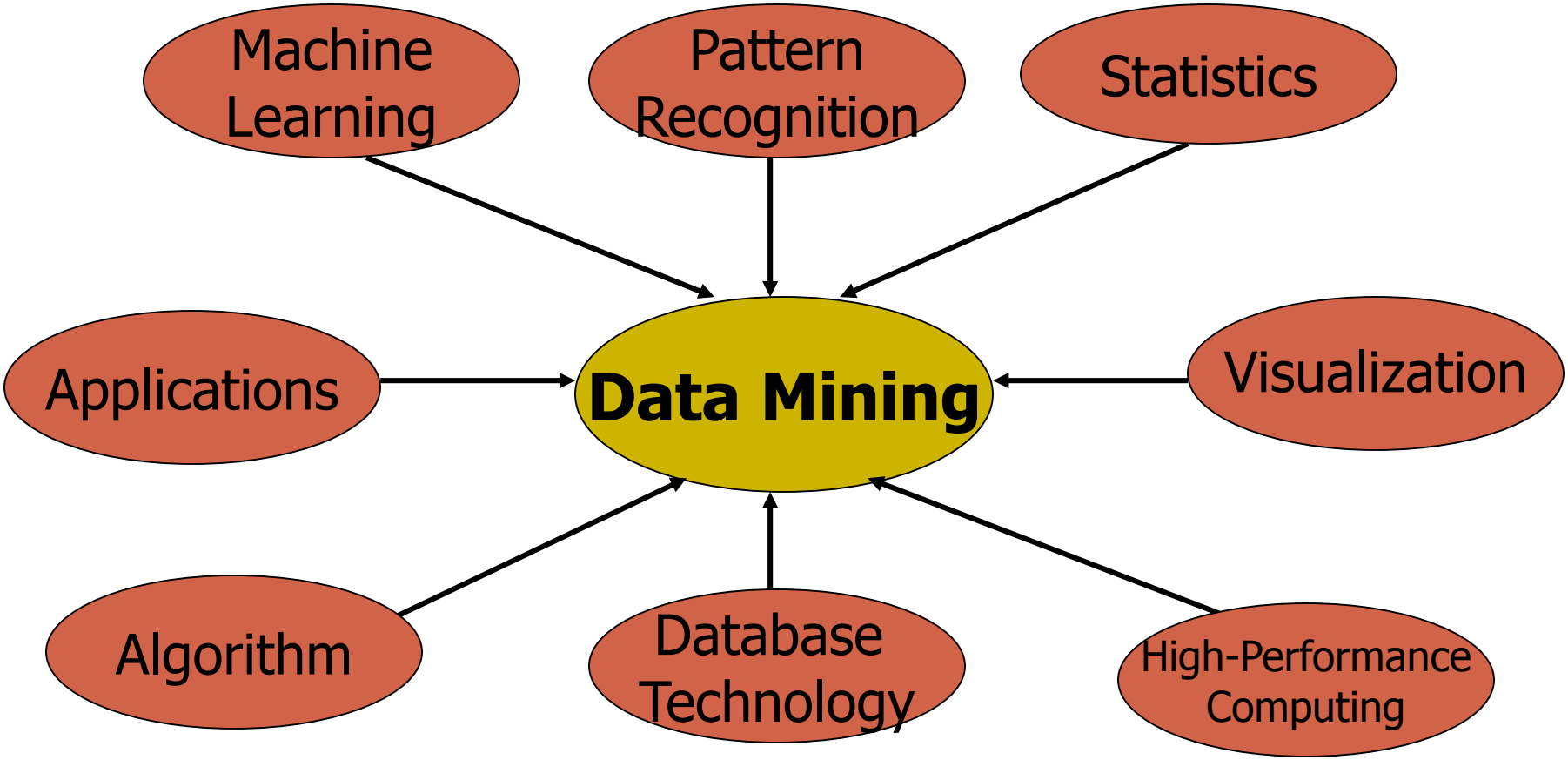
# Evaluation of Knowledge

- Are all mined knowledge interesting?
  - One can mine tremendous amount of "patterns" and knowledge
  - Some may fit only certain dimension space (time, location, …)
  - Some may not be representative, may be transient, …
- Evaluation of mined knowledge → directly mine only interesting knowledge?
  - Descriptive vs. predictive
  - Coverage
  - Typicality vs. novelty
  - Accuracy
  - Timeliness
  - …

# 1. Introduction

- Why Data Mining?
- What Is Data Mining?
- A Multi-Dimensional View of Data Mining
  - What Kinds of Data Can Be Mined?
  - What Kinds of Patterns Can Be Mined?
  - What Kinds of Technologies Are Used? 
  - What Kinds of Applications Are Targeted?
- Content covered by this course

# Data Mining: Confluence of Multiple Disciplines



# 1. Introduction

- Why Data Mining?
- What Is Data Mining?
- A Multi-Dimensional View of Data Mining
  - What Kinds of Data Can Be Mined?
  - What Kinds of Patterns Can Be Mined?
  - What Kinds of Technologies Are Used?
  - What Kinds of Applications Are Targeted?

- Content covered by this course

# Applications of Data Mining

- Web page analysis: from web page classification, clustering to PageRank & HITS algorithms
- Collaborative analysis & recommender systems
- Basket data analysis to targeted marketing
- Biological and medical data analysis: classification, cluster analysis (microarray data analysis), biological sequence analysis, biological network analysis
- Data mining and software engineering (e.g., IEEE Computer, Aug. 2009 issue)
- Social media
- Game

# 1. Introduction

- Why Data Mining?
- What Is Data Mining?
- A Multi-Dimensional View of Data Mining
  - What Kinds of Data Can Be Mined?
  - What Kinds of Patterns Can Be Mined?
  - What Kinds of Technologies Are Used?
  - What Kinds of Applications Are Targeted?
- Content covered by this course 

# Course Content

- By data types:
  - matrix data
  - set data
  - sequence data
  - time series
  - graph and network
- By functions:
  - Classification
  - Clustering
  - Frequent pattern mining
  - Prediction
  - Similarity search
  - Ranking

# Methods to Learn

| | Matrix Data | Set Data | Sequence Data | Time Series | Graph & Network |
|---|---|---|---|---|---|
| **Classification** | Decision Tree; Naïve Bayes; Logistic Regression<br>SVM; kNN | | HMM | | Label Propagation |
| **Clustering** | K-means; hierarchical clustering; DBSCAN; Mixture Models; kernel k-means | | | | SCAN; Spectral Clustering |
| **Frequent Pattern Mining** | | Apriori; FP-growth | GSP; PrefixSpan | | |
| **Prediction** | Linear Regression | | | Autoregression | |
| **Similarity Search** | | | | DTW | P-PageRank |
| **Ranking** | | | | | PageRank |

# Evaluation

- How to determine whether a method is good or not?
  - Effectiveness
  - Efficiency

# Where to Find References? DBLP, CiteSeer, Google

- Data mining and KDD (SIGKDD: CDROM)
  - Conferences: ACM-SIGKDD, IEEE-ICDM, SIAM-DM, PKDD, PAKDD, etc.
  - Journal: Data Mining and Knowledge Discovery, KDD Explorations, ACM TKDD
- Database systems (SIGMOD: ACM SIGMOD Anthology—CD ROM)
  - Conferences: ACM-SIGMOD, ACM-PODS, VLDB, IEEE-ICDE, EDBT, ICDT, DASFAA
  - Journals: IEEE-TKDE, ACM-TODS/TOIS, JIIS, J. ACM, VLDB J., Info. Sys., etc.
- AI & Machine Learning
  - Conferences: Machine learning (ML), AAAI, IJCAI, COLT (Learning Theory), CVPR, NIPS, etc.
  - Journals: Machine Learning, Artificial Intelligence, Knowledge and Information Systems, IEEE-PAMI, etc.
- Web and IR
  - Conferences: SIGIR, WWW, CIKM, etc.
  - Journals: WWW: Internet and Web Information Systems,
- Statistics
  - Conferences: Joint Stat. Meeting, etc.
  - Journals: Annals of statistics, etc.
- Visualization
  - Conference proceedings: CHI, ACM-SIGGraph, etc.
  - Journals: IEEE Trans. visualization and computer graphics, etc.

# Recommended Reference Books


- E. Alpaydin. Introduction to Machine Learning, 2nd ed., MIT Press, 2011
- S. Chakrabarti. Mining the Web: Statistical Analysis of Hypertex and Semi-Structured Data. Morgan Kaufmann, 2002
- R. O. Duda, P. E. Hart, and D. G. Stork, Pattern Classification, 2ed., Wiley-Interscience, 2000
- T. Dasu and T. Johnson. Exploratory Data Mining and Data Cleaning. John Wiley & Sons, 2003
- U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy. Advances in Knowledge Discovery and Data Mining. AAAI/MIT Press, 1996
- U. Fayyad, G. Grinstein, and A. Wierse, Information Visualization in Data Mining and Knowledge Discovery, Morgan Kaufmann, 2001
- J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques. Morgan Kaufmann, 3rd ed. , 2011
- T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd ed., Springer, 2009
- B. Liu, Web Data Mining, Springer 2006
- T. M. Mitchell, Machine Learning, McGraw Hill, 1997
- Y. Sun and J. Han, Mining Heterogeneous Information Networks, Morgan & Claypool, 2012
- P.-N. Tan, M. Steinbach and V. Kumar, Introduction to Data Mining, Wiley, 2005
- S. M. Weiss and N. Indurkhya, Predictive Data Mining, Morgan Kaufmann, 1998
- I. H. Witten and E. Frank, Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations, Morgan Kaufmann, 2nd ed. 2005