



CS 412 Intro. to Data Mining

Chapter 1. Introduction

Jiawei Han, Computer Science, Univ. Illinois at Urbana-Champaign, 2017





Data and Information Systems (DAIS)

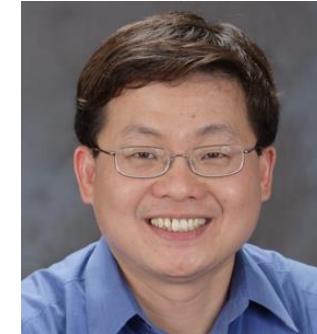
- Database Systems



Jiawei Han



Aditya
Parameswaran



Kevin Chang

- Data Mining



Hari
Sundaram

- Text Information Systems



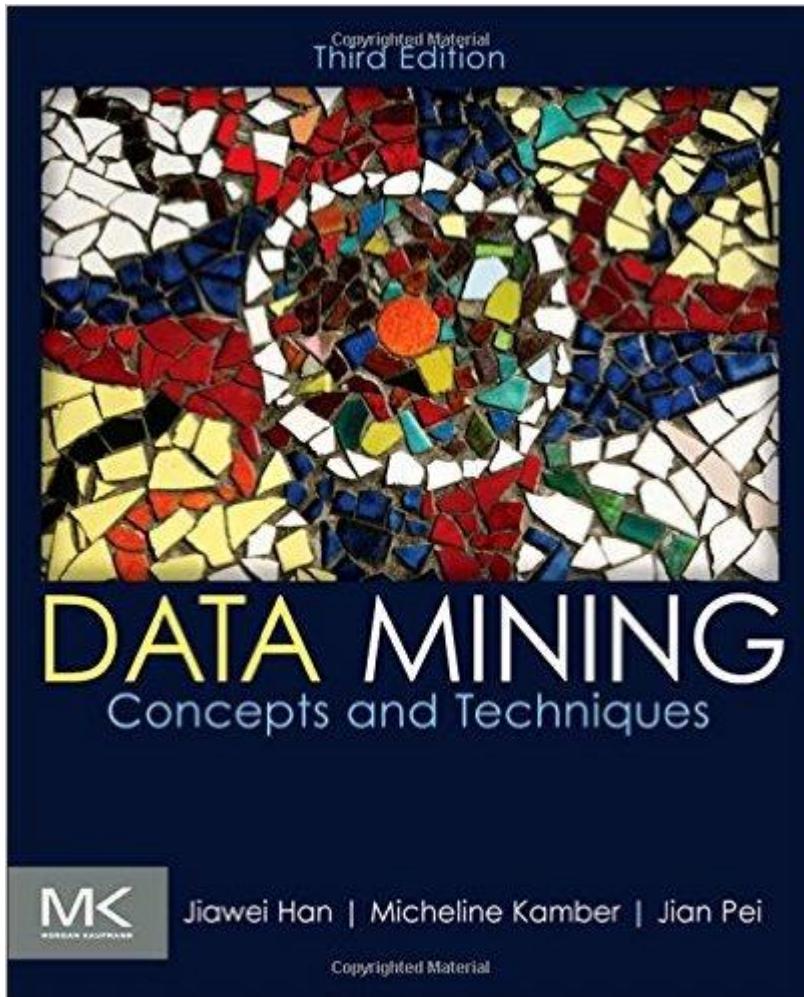
ChengXiang
Zhai

- Networks

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 warehousing and 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)

CS 412. Course Page & Class Schedule

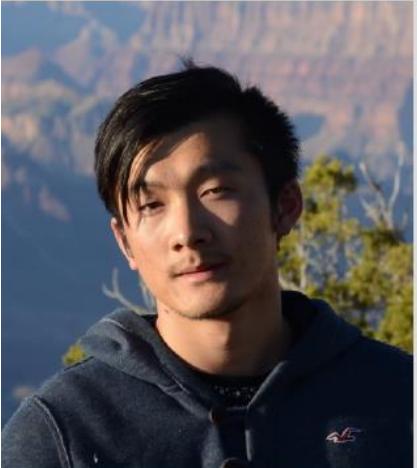


- Textbook
 - Jiawei Han, Micheline Kamber and Jian Pei, *Data Mining: Concepts and Techniques* (3rd ed), Morgan Kaufmann, 2011
- Class Homepage:
<https://wiki.engr.illinois.edu/display/cs412>
- Bookmark on course schedule page
- **Class Schedule: 9:30-10:45 am Tues./Thurs. @1404 SC**
- Office hours: 10:45-11:30am Tues./Thurs. @2132 SC
- Lecture media: recorded; but class attendance is critical

CS 412. Fall 2017. Teach Assistants

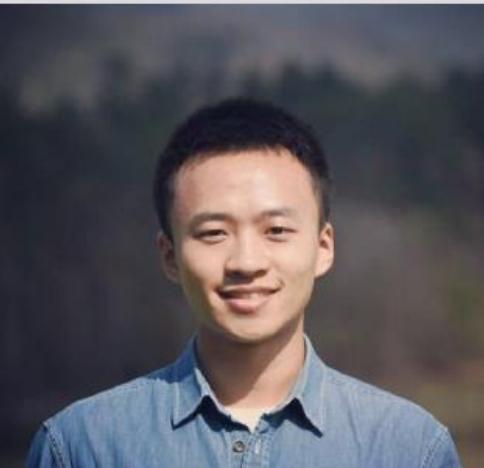


Dongming Lei



Carl Yang

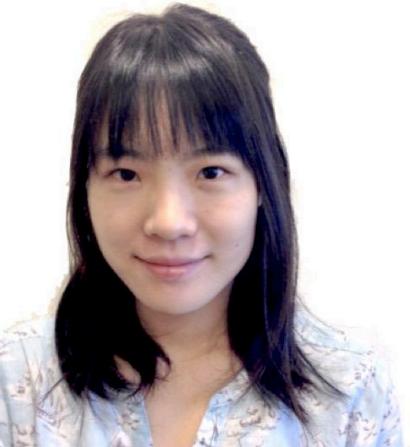
(Online Session)



Yu Shi



Chao Zhang



Shi Zhi

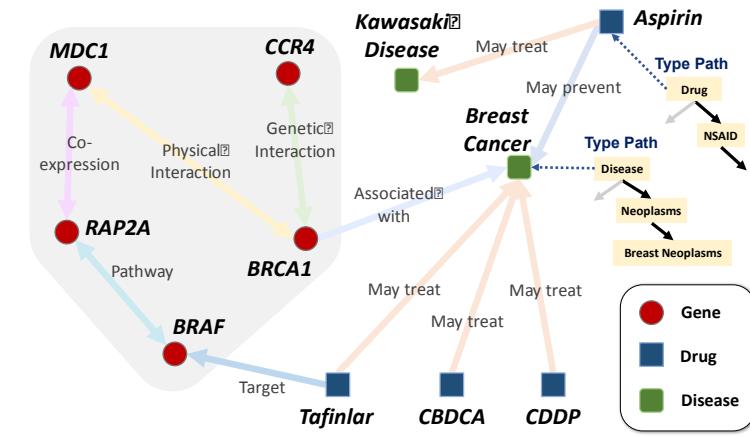
- TA office hours: **4-5pm (Mon.), 11-12pm (Wed.)@0207SC**. Additional hours before due date will be announced at Piazza
- Wait list (No wait list at this time, keep attending class, see if there is space available or there is overflow section opening)
- If you cannot register but still desperately want to get in, please sign on when there is “potential opening”: Explain why you have to take the course This Fall!

CS 412. Course Work and Grading

- ❑ Assignments, Programming Assignments, and Exams
 - ❑ Written Assignments: 15% (three homework assignments expected)
 - ❑ Programming assignments: 20% (two programming assignments expected)
 - ❑ Midterm exam: 30%
 - ❑ Final exam: 35%
- ❑ For students taking 4th credit (TA will provide concrete instructions on the 4th credit project)
 - ❑ For students registering 4 credits: 25%. The overall scores will be scaled proportionally
- ❑ Need help and/or discussions?
 - ❑ Sign on: Piazza (<https://piazza.com/illinois/cs412>)
- ❑ Check your homework/exam scores:
 - ❑ Compass

Help Needed: LifeNet—A Structured Network-Based Knowledge Exploration and Analytics System for Life Sciences

- ❑ What we are doing?
 - ❑ A scalable system that transforms biomedical papers into a knowledge graph & supports various search/analytics functions
- ❑ What we already have?
 - ❑ A working prototype system & an ACL demo paper
- ❑ What we are looking for?
 - ❑ Students with expertise on HTML/CSS & JavaScript
 - ❑ Experiences on web frameworks and databases
 - ❑ System design experience will be a big plus
- ❑ What you will gain?
 - ❑ Hourly pay (\$12-\$15 per hour, 6-20 hours per week)
 - ❑ Possible research publications & a good thesis topic



LifeNet Network Exploration Distinctive Summarization

Argument 1: Cardiomyopathies Argument 2: Gene Relation: GeneDiseaseAssociation

Show Relationships Show predicted relationships

+

ACTC1 Endocardial Fibroelastoses

TAZ ABCC9

TTN TNNI3

Familial dilated cardiomyopathy

Familial Restrictive Cardiomyopathy

Carvaljal syndrome

Dmd-Related Dilated Cardiomyopathy

Centronuclear Myopathy

Recessive truncating titin gene, TTN, mutations presenting as centronuclear myopathy
2013 Oct 1;81(14):1205-14. Epub 2013 Aug 23.
To identify causative genes for centronuclear myopathies (CNM), a heterogeneous

Titin and centronuclear myopathy: The tip of the iceberg for TTN-ic mutations?
2013 Oct 1;81(14):1189-90. Epub 2013 Aug 23.
RESULTS: Autosomal recessive compound heterozygous truncating mutations of the titin gene, TTN, were identified

Send us your resume if interested: Jiaming Shen (mickeysjm@gmail.com)

Chapter 1. Introduction

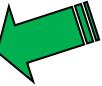
- Why Data Mining? 
- What Is Data Mining?
- A Multi-Dimensional View of Data Mining
- What Kinds of Data Can Be Mined?
- What Kinds of Patterns Can Be Mined?
- What Kinds of Technologies Are Used?
- What Kinds of Applications Are Targeted?
- Major Issues in Data Mining
- A Brief History of Data Mining and Data Mining Society
- Summary

Why Data Mining?

ឧប្បជ្ជកម្ម

- 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

Chapter 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?
- Major Issues in Data Mining
- A Brief History of Data Mining and Data Mining Society
- Summary

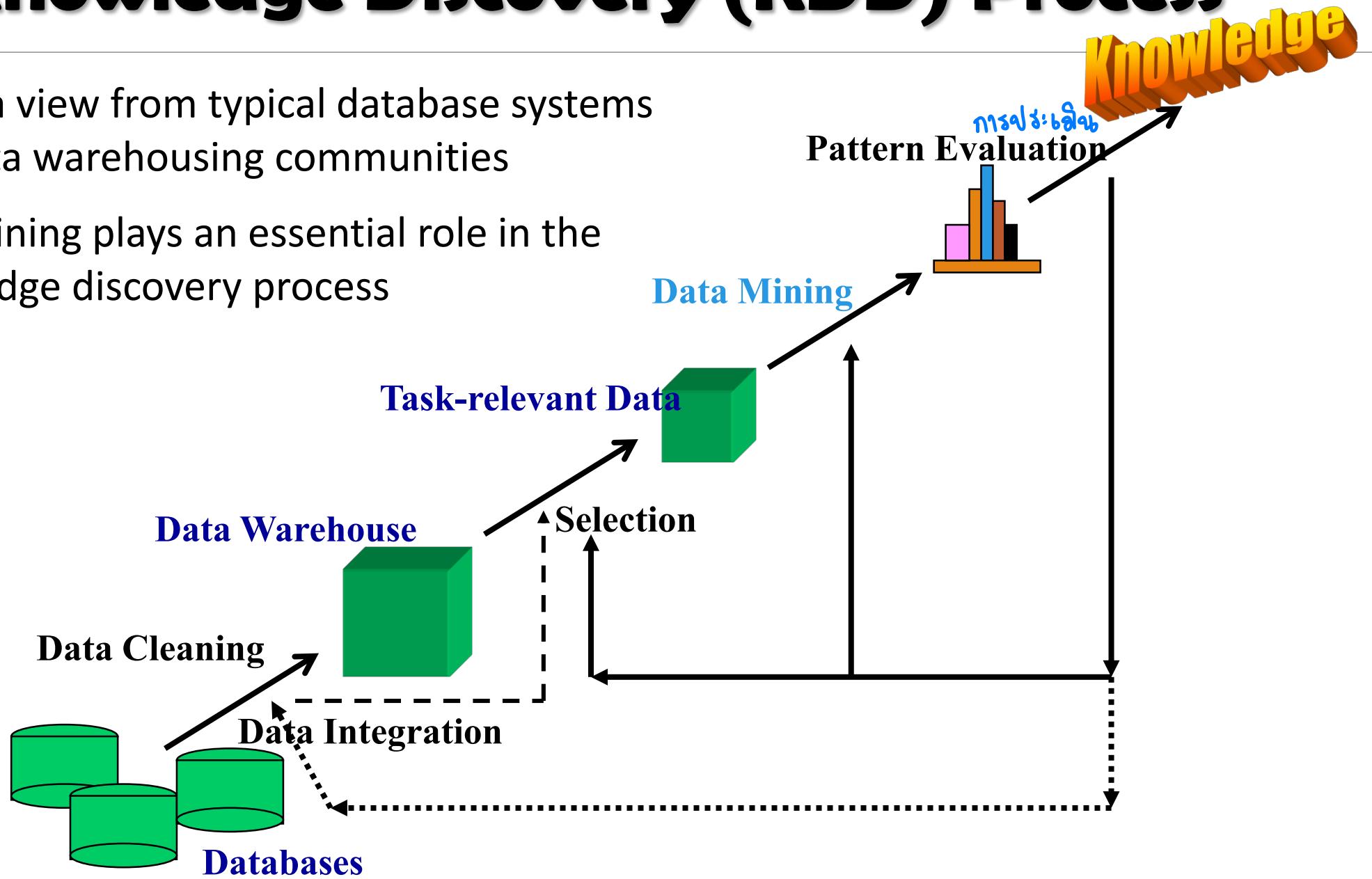
What Is Data Mining?

- Data mining (knowledge discovery from data)
ការបញ្ចូនពេលបច្ចុប្បន្នគោលការណ៍ដែលមិនមែនស្ថាប់គឺជាការបញ្ចូនពេលបច្ចុប្បន្ន
- Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data
ការបញ្ចូនពេលបច្ចុប្បន្នដែលមិនមែនស្ថាប់គឺជាការបញ្ចូនពេលបច្ចុប្បន្ន
- Data mining: a misnomer?
- Alternative names
 - Knowledge discovery (mining) in databases (KDD), knowledge extraction, data/pattern analysis, data archeology, data dredging, information harvesting, business intelligence, etc.
សក់តគោលរំលែក
- Watch out: Is everything “data mining”?
 - Simple search and query processing
 - (Deductive) expert systems



Knowledge Discovery (KDD) Process

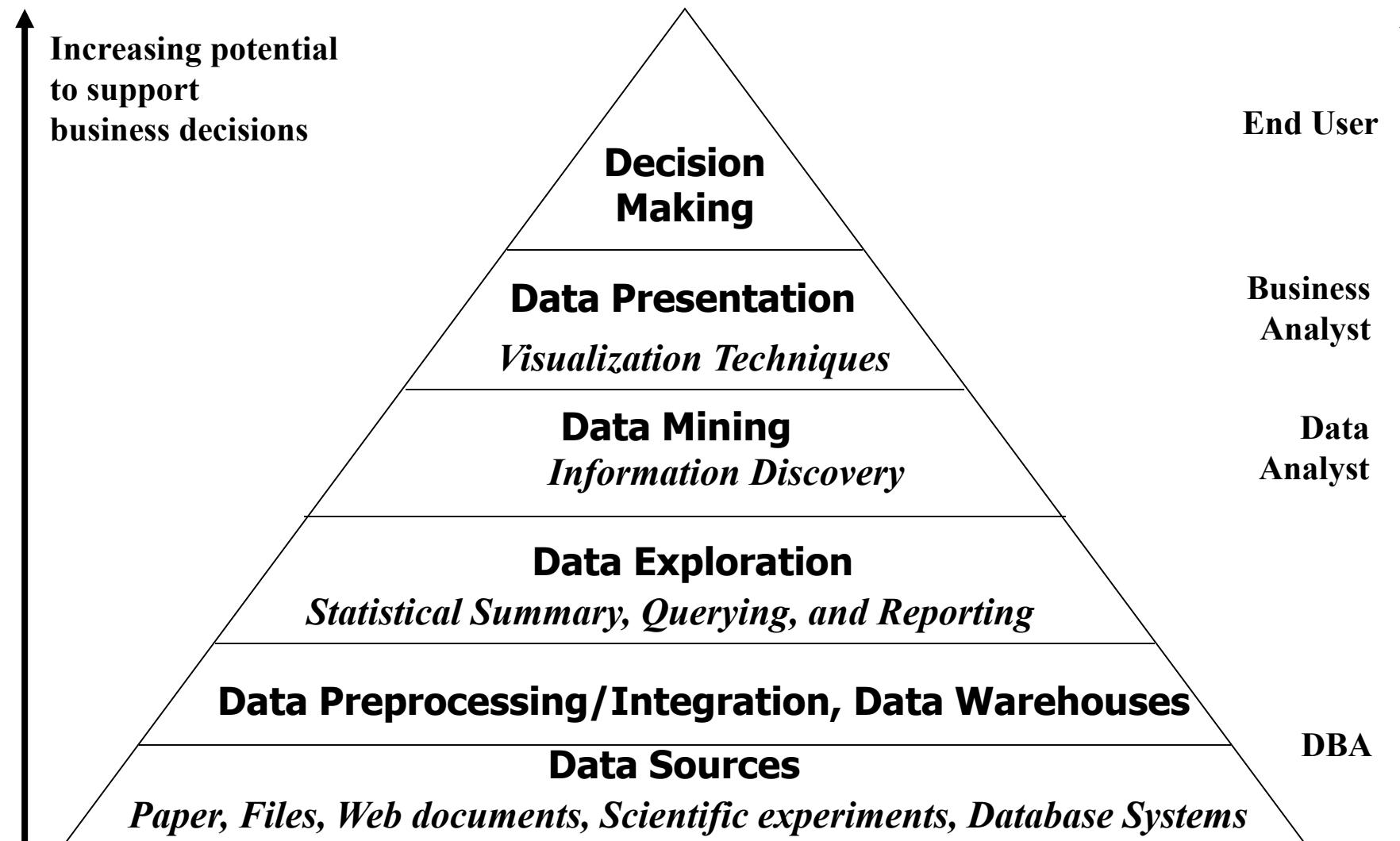
- This is a view from typical database systems and data warehousing communities
- Data mining plays an essential role in the knowledge discovery process



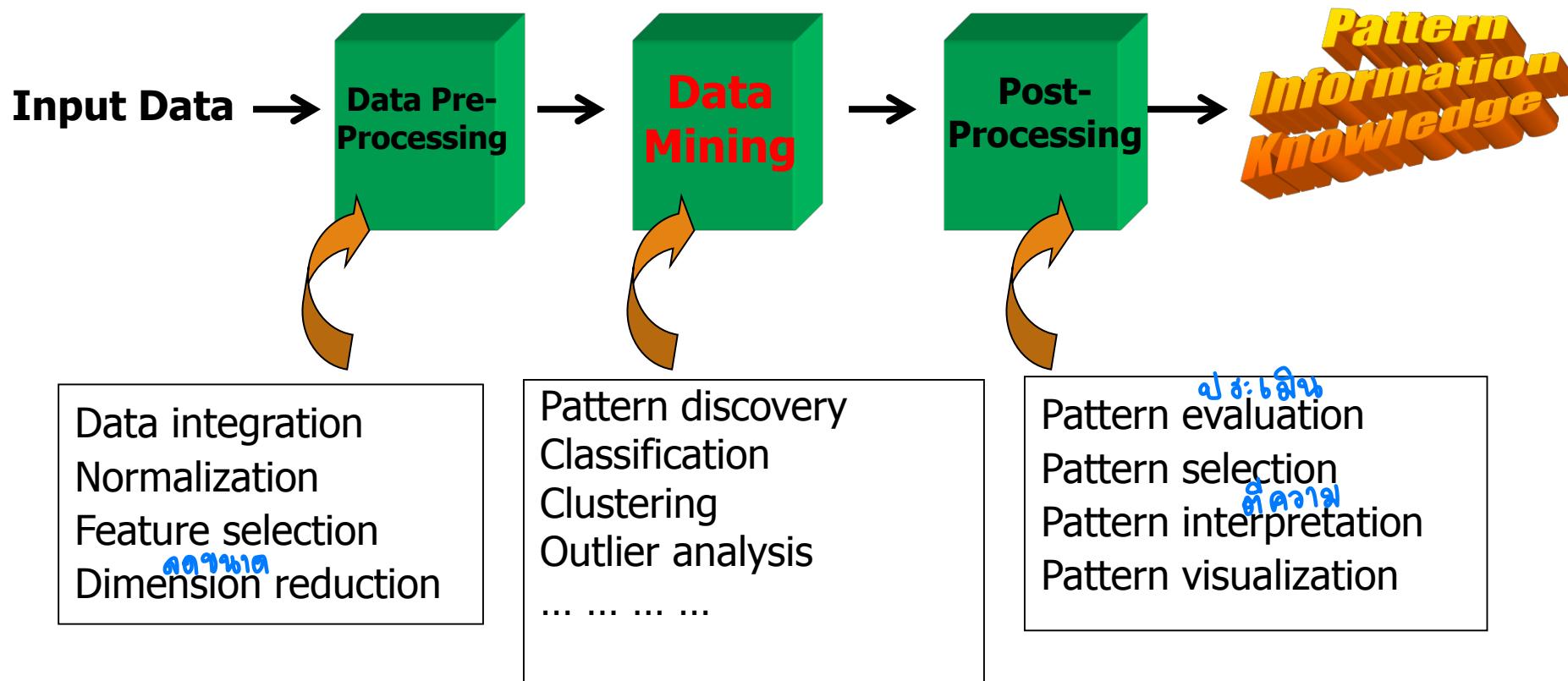
Example: A Web Mining Framework

- Web mining usually involves
 - Data cleaning
 - Data integration from multiple sources
 - Warehousing the data
 - Data cube construction
 - Data selection for data mining
 - Data mining
 - Presentation of the mining results
 - Patterns and knowledge to be used or stored into knowledge-base

Data Mining in Business Intelligence



KDD Process: A View from ML and Statistics



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

Data Mining vs. Data Exploration

- Which view do you prefer?
 - KDD vs. ML/Stat. vs. Business Intelligence
 - Depending on the data, applications, and your focus

- Data Mining vs. Data Exploration
 - Business intelligence view
 - Warehouse, data cube, reporting but not much mining
 - Business objects vs. data mining tools
 - Supply chain example: mining vs. OLAP vs. presentation tools
 - Data presentation vs. data exploration

Chapter 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?
- Major Issues in Data Mining
- A Brief History of Data Mining and Data Mining Society
- Summary



Multi-Dimensional View of Data Mining

Data to be mined

- Database data (extended-relational, object-oriented, heterogeneous), 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, ...
- 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.

Chapter 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?
- Major Issues in Data Mining
- A Brief History of Data Mining and Data Mining Society
- Summary

Data Mining: On What Kinds of Data?

- ❑ Database-oriented data sets and applications
 - ❑ Relational database, data warehouse, transactional database
 - ❑ Object-relational databases, Heterogeneous databases and legacy databases
- ❑ Advanced data sets and advanced applications
 - ❑ Data streams and sensor data
 - ❑ Time-series data, temporal data, sequence data (incl. bio-sequences)
 - ❑ Structure data, graphs, social networks and information networks
 - ❑ Spatial data and spatiotemporal data
 - ❑ Multimedia database
 - ❑ Text databases
 - ❑ The World-Wide Web

Chapter 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?
- Major Issues in Data Mining
- A Brief History of Data Mining and Data Mining Society
- Summary



Data Mining Functions: (1) Generalization

- ❑ Information integration and data warehouse construction

- ❑ Data cleaning, transformation, integration, and multidimensional data model

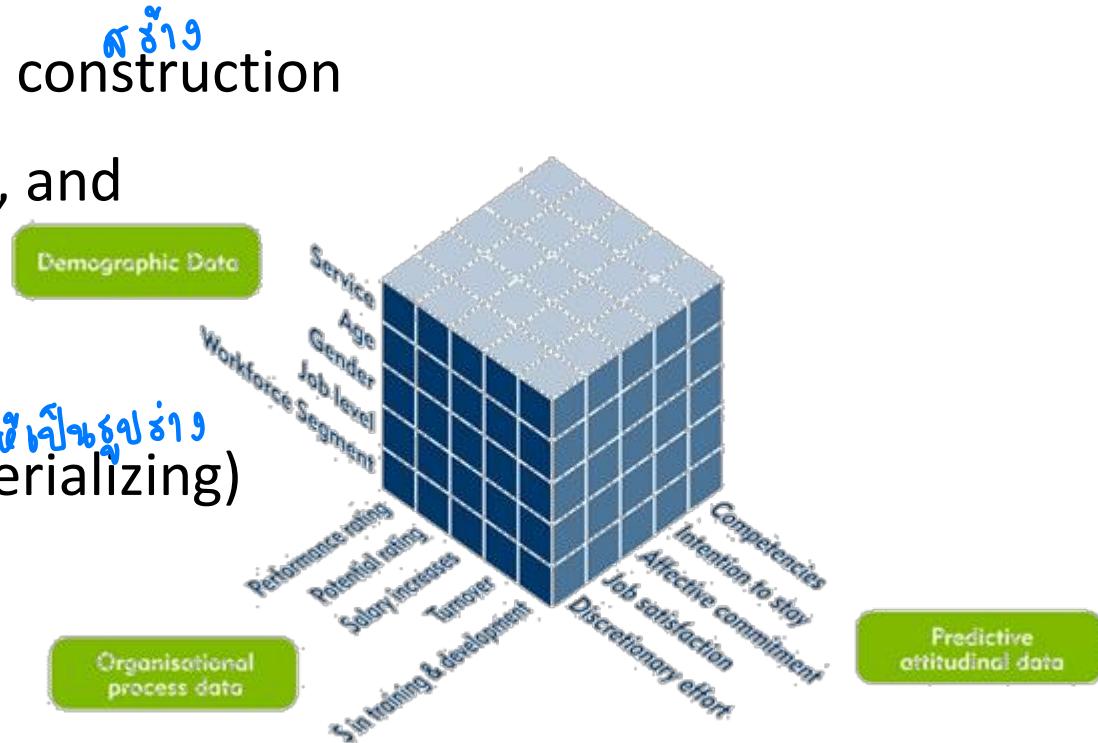
- ❑ Data cube technology

- ❑ Scalable methods for computing (i.e., materializing) multidimensional aggregates

- ❑ OLAP (online analytical processing)

- ❑ Multidimensional concept description: Characterization and discrimination

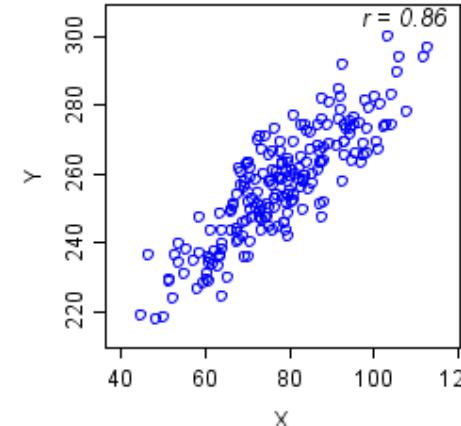
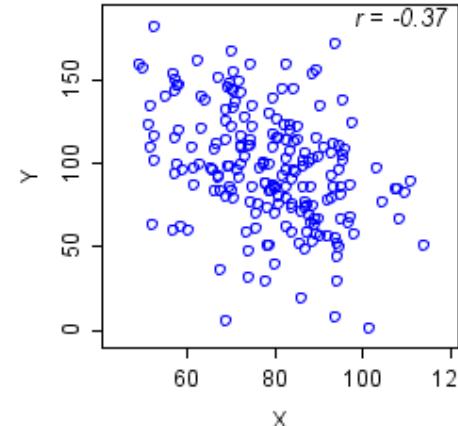
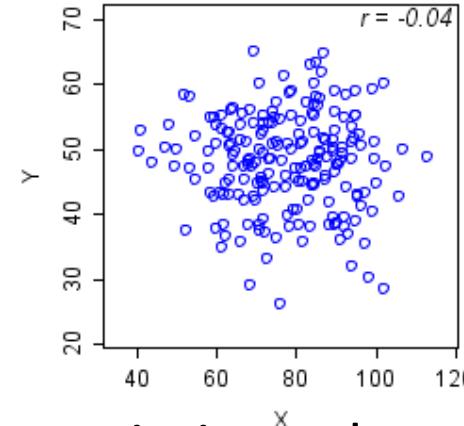
- ❑ Generalize, summarize, and contrast data characteristics, e.g., dry vs. wet region



Data Mining Functions: (2) Pattern Discovery

ក្នុងលេខាតីគឺជាក្នុងរបៀបណា

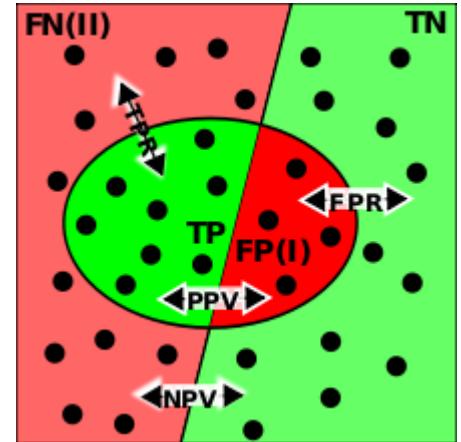
- Frequent patterns (or frequent itemsets)
 - What items are frequently purchased together in your Walmart?
- Association and Correlation Analysis



- A typical association rule
 - Diaper \rightarrow Beer [0.5%, 75%] (support, confidence)
 - Are strongly associated items also strongly correlated?
- How to mine such patterns and rules efficiently in large datasets?
- How to use such patterns for classification, clustering, and other applications?

Data Mining Functions: (3) Classification

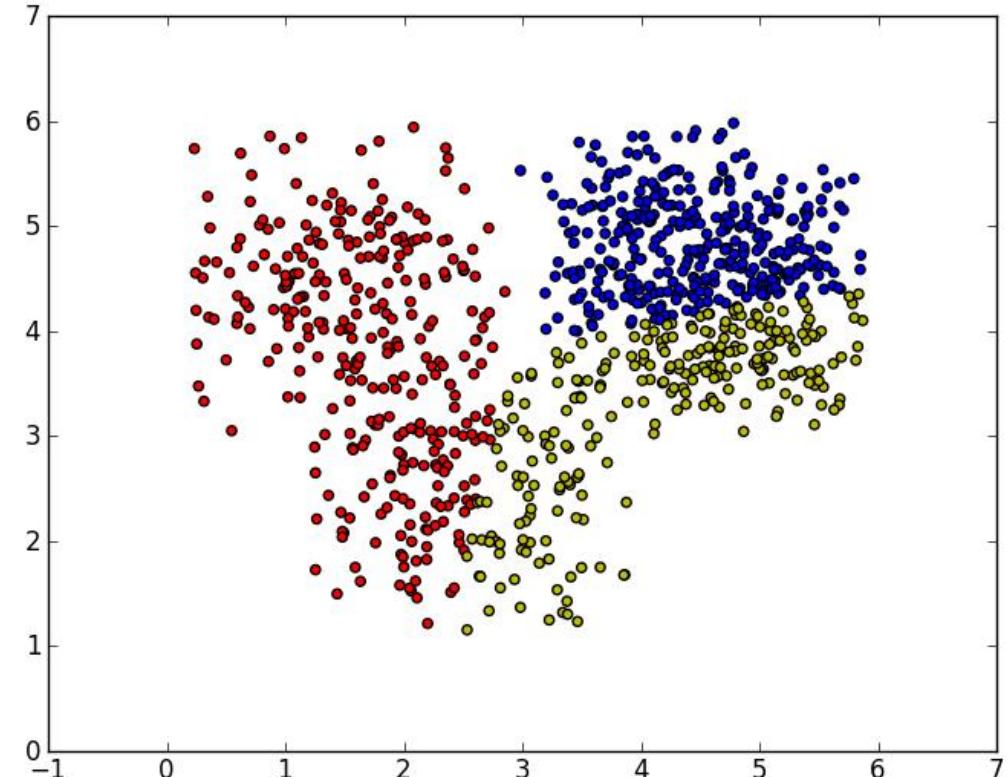
- ❑ Classification and label prediction
 - ❑ Construct models (functions) based on some training examples
 - ❑ Describe and distinguish classes or concepts for future prediction
 - ❑ Ex. 1. Classify countries based on (climate)
 - ❑ Ex. 2. Classify cars based on (gas mileage)
 - ❑ Predict some unknown 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 Functions: (4) Cluster Analysis

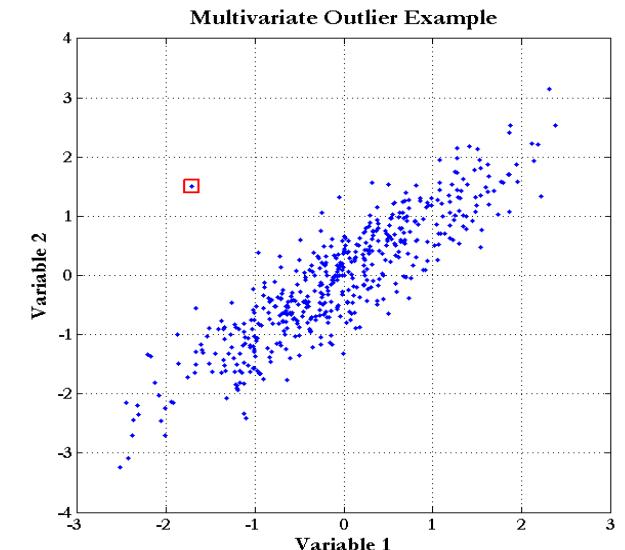
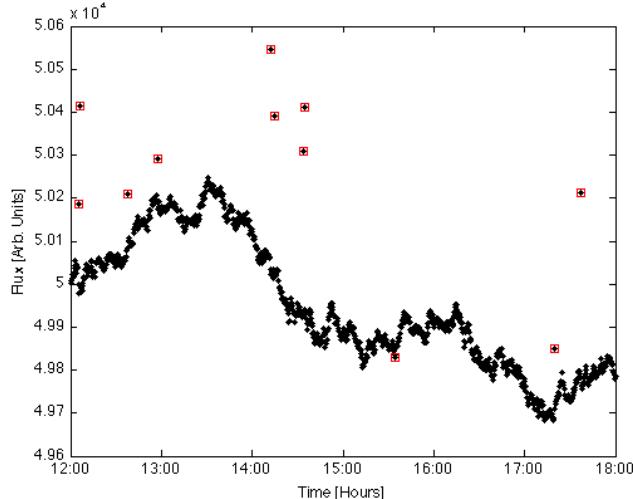
* ແນ່ງກລົມຈຳຂອດ

- 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



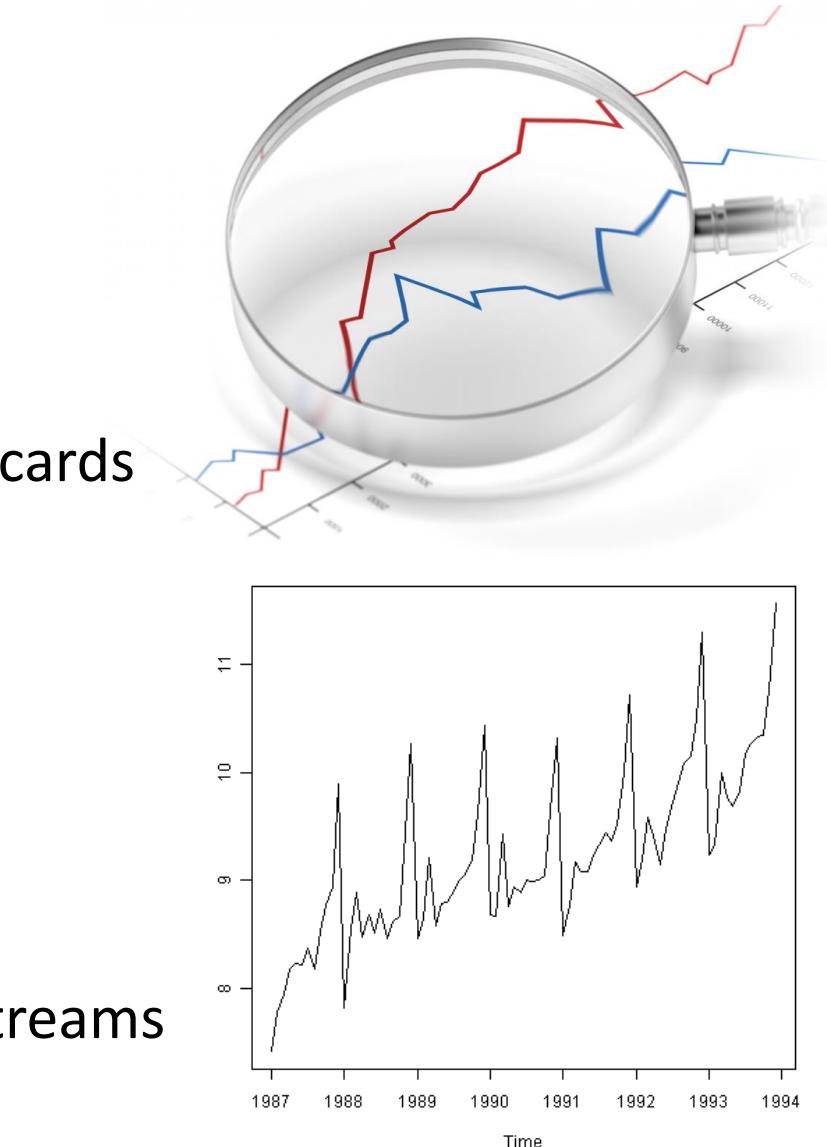
Data Mining Functions: (5) Outlier Analysis

- Outlier analysis
 - Outlier: A data object that does not comply with the general behavior of the data
 - Noise or exception?—One person's garbage could be another person's treasure
 - Methods: by product of clustering or regression analysis, ...
 - Useful in fraud detection, rare events analysis



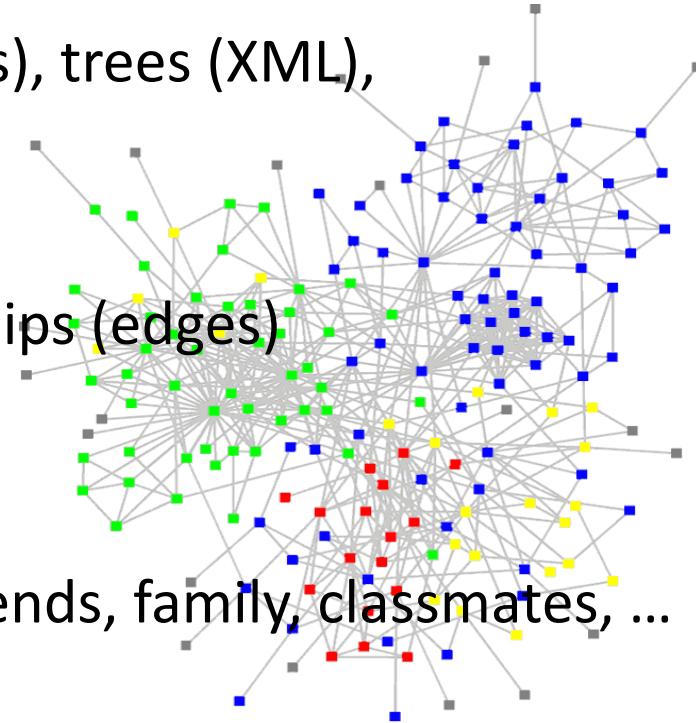
Data Mining Functions: (6) Time and Ordering: Sequential Pattern, Trend and Evolution Analysis

- Sequence, trend and evolution analysis
 - Trend, time-series, and deviation analysis
 - e.g., regression and value prediction
 - Sequential pattern mining
 - e.g., buy digital camera, then buy large memory cards
 - Periodicity analysis
 - Motifs and biological sequence analysis
 - Approximate and consecutive motifs
 - Similarity-based analysis
- Mining data streams
 - Ordered, time-varying, potentially infinite, data streams



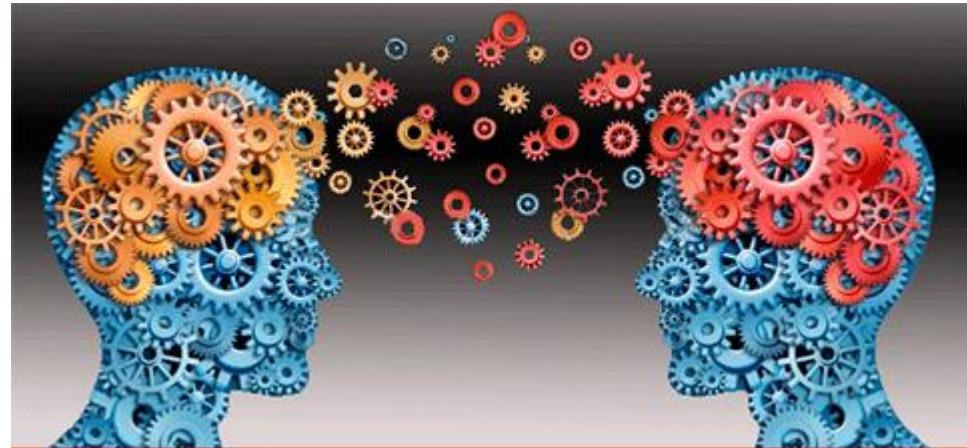
Data Mining Functions: (7) Structure and Network Analysis

- ❑ Graph mining
 - ❑ Finding frequent subgraphs (e.g., chemical compounds), trees (XML), substructures (web fragments)
- ❑ Information network analysis
 - ❑ Social networks: actors (objects, nodes) and relationships (edges)
 - ❑ e.g., author networks in CS, terrorist networks
 - ❑ Multiple heterogeneous networks
 - ❑ A person could be multiple information networks: friends, family, classmates, ...
 - ❑ Links carry a lot of semantic information: Link mining
- ❑ Web mining
 - ❑ Web is a big information network: from PageRank to Google
 - ❑ Analysis of Web information networks
 - ❑ Web community discovery, opinion mining, usage mining, ...



Evaluation of Knowledge

- Are all mined knowledge interesting?
 - One can mine tremendous amount of “patterns”
 - 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
 - ...

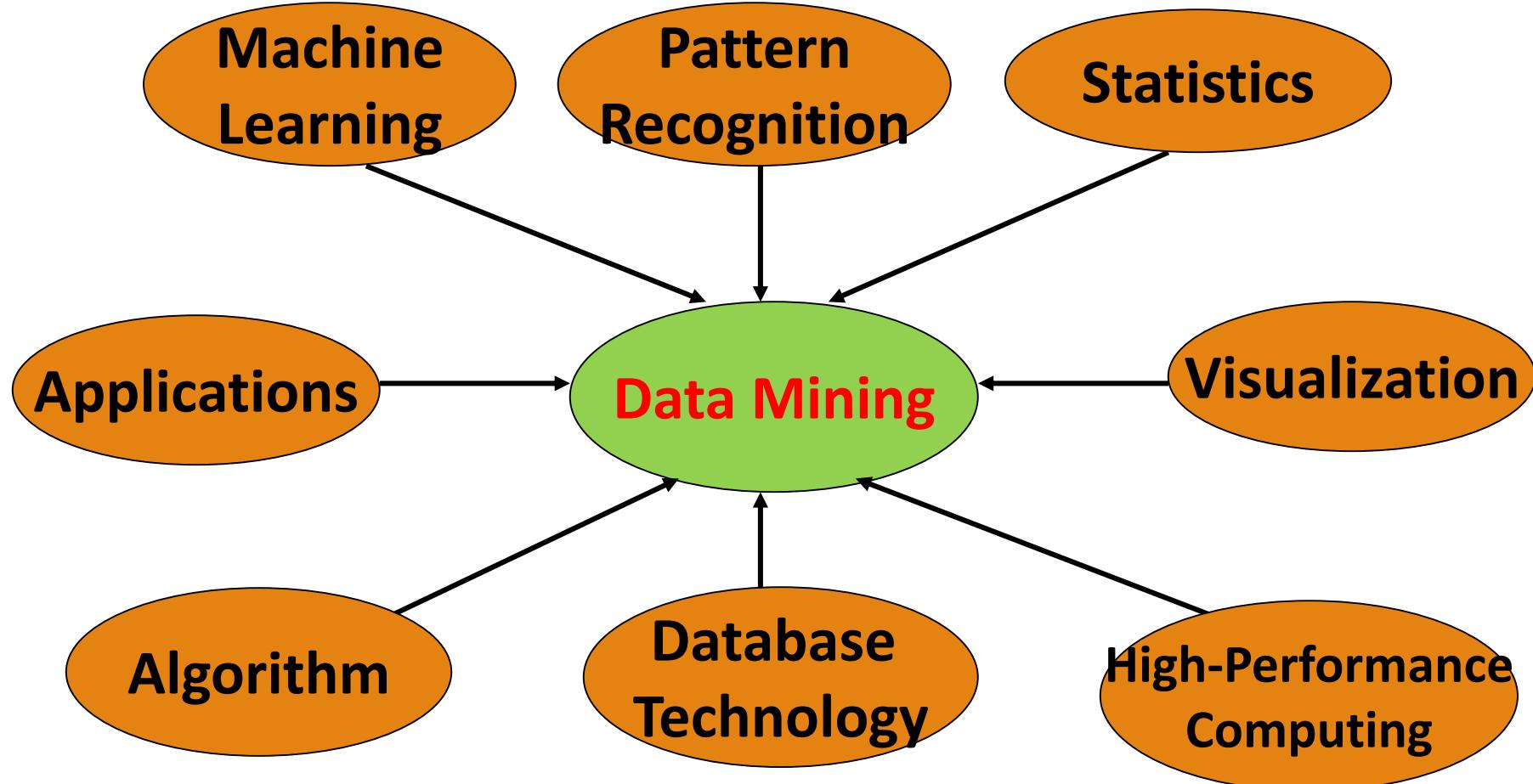


Chapter 1. Introduction

- 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?
- Major Issues in Data Mining
- A Brief History of Data Mining and Data Mining Society
- Summary



Data Mining: Confluence of Multiple Disciplines



Why Confluence of Multiple Disciplines?

- Tremendous amount of data
 - Algorithms must be scalable to handle big data
- High-dimensionality of data
 - Micro-array may have tens of thousands of dimensions
- High complexity of data
 - Data streams and sensor data
 - Time-series data, temporal data, sequence data
 - Structure data, graphs, social and information networks
 - Spatial, spatiotemporal, multimedia, text and Web data
 - Software programs, scientific simulations
- New and sophisticated applications

Chapter 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? 
- Major Issues in Data Mining
- A Brief History of Data Mining and Data Mining Society
- Summary

Applications of Data Mining

- Web page analysis: classification, clustering, ranking
- Collaborative analysis & recommender systems
- Basket data analysis to targeted marketing
- Biological and medical data analysis
- Data mining and software engineering
- Data mining and text analysis
- Data mining and social and information network analysis
- Built-in (invisible data mining) functions in Google, MS, Yahoo!, Linked, Facebook, ...
- Major dedicated data mining systems/tools
- SAS, MS SQL-Server Analysis Manager, Oracle Data Mining Tools)



Chapter 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?
- Major Issues in Data Mining
- A Brief History of Data Mining and Data Mining Society
- Summary



Major Issues in Data Mining (1)

- ❑ Mining Methodology
 - ❑ Mining various and new kinds of knowledge
 - ❑ Mining knowledge in multi-dimensional space
 - ❑ Data mining: An interdisciplinary effort
 - ❑ Boosting the power of discovery in a networked environment
 - ❑ Handling noise, uncertainty, and incompleteness of data
 - ❑ Pattern evaluation and pattern- or constraint-guided mining
- ❑ User Interaction
 - ❑ Interactive mining
 - ❑ Incorporation of background knowledge
 - ❑ Presentation and visualization of data mining results

Major Issues in Data Mining (2)

- Efficiency and Scalability
 - Efficiency and scalability of data mining algorithms
 - Parallel, distributed, stream, and incremental mining methods
- Diversity of data types
 - Handling complex types of data
 - Mining dynamic, networked, and global data repositories
- Data mining and society
 - Social impacts of data mining
 - Privacy-preserving data mining
 - Invisible data mining

Chapter 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?
- Major Issues in Data Mining
- A Brief History of Data Mining and Data Mining Society
- Summary



A Brief History of Data Mining Society

- 1989 IJCAI Workshop on Knowledge Discovery in Databases
 - Knowledge Discovery in Databases (G. Piatetsky-Shapiro and W. Frawley, 1991)
- 1991-1994 Workshops on Knowledge Discovery in Databases
 - Advances in Knowledge Discovery and Data Mining (U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, 1996)
- 1995-1998 International Conferences on Knowledge Discovery in Databases and Data Mining (KDD'95-98)
 - Journal of Data Mining and Knowledge Discovery (1997)
- ACM SIGKDD conferences since 1998 and SIGKDD Explorations
- More conferences on data mining
 - PAKDD (1997), PKDD (1997), SIAM-Data Mining (2001), (IEEE) ICDM (2001), WSDM (2008), etc.
- ACM Transactions on KDD (2007)

Conferences and Journals on Data Mining

- ❑ KDD Conferences
 - ❑ ACM SIGKDD Int. Conf. on Knowledge Discovery in Databases and Data Mining ([KDD](#))
 - ❑ SIAM Data Mining Conf. ([SDM](#))
 - ❑ (IEEE) Int. Conf. on Data Mining ([ICDM](#))
 - ❑ European Conf. on Machine Learning and Principles and practices of Knowledge Discovery and Data Mining ([ECML-PKDD](#))
 - ❑ Pacific-Asia Conf. on Knowledge Discovery and Data Mining ([PAKDD](#))
 - ❑ Int. Conf. on Web Search and Data Mining ([WSDM](#))
- Other related conferences
 - DB conferences: ACM SIGMOD, VLDB, ICDE, EDBT, ICDT, ...
 - Web and IR conferences: WWW, SIGIR, WSDM
 - ML conferences: ICML, NIPS
 - PR conferences: CVPR,
- Journals
 - Data Mining and Knowledge Discovery (DAMI or DMKD)
 - IEEE Trans. On Knowledge and Data Eng. (TKDE)
 - KDD Explorations
 - ACM Trans. on KDD

Where to Find References? DBLP, CiteSeer, Google

- Data mining and KDD (SIGKDD)
 - Conferences: ACM-SIGKDD, IEEE-ICDM, SIAM-DM, PKDD, PAKDD, etc.
 - Journal: Data Mining and Knowledge Discovery, KDD Explorations, ACM TKDD
- Database systems (SIGMOD)
 - 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.

Chapter 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?
- Major Issues in Data Mining
- A Brief History of Data Mining and Data Mining Society
- Summary



Summary

- Data mining: Discovering interesting patterns and knowledge from massive amount of data
- A natural evolution of science and information technology, in great demand, with wide applications
- A KDD process includes data cleaning, data integration, data selection, transformation, data mining, pattern evaluation, and knowledge presentation
- Mining can be performed in a variety of data
- Data mining functionalities: characterization, discrimination, association, classification, clustering, trend and outlier analysis, etc.
- Data mining technologies and applications
- Major issues in data mining

Recommended Reference Books

- Charu C. Aggarwal, Data Mining: The Textbook, Springer, 2015
- E. Alpaydin. Introduction to Machine Learning, 2nd ed., MIT Press, 2011
- R. O. Duda, P. E. Hart, and D. G. Stork, Pattern Classification, 2ed., Wiley-Interscience, 2000
- 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
- T. M. Mitchell, Machine Learning, McGraw Hill, 1997
- P.-N. Tan, M. Steinbach and V. Kumar, Introduction to Data Mining, Wiley, 2005 (2nd ed. 2016)
- I. H. Witten and E. Frank, Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations, Morgan Kaufmann, 2nd ed. 2005
- Mohammed J. Zaki and Wagner Meira Jr., Data Mining and Analysis: Fundamental Concepts and Algorithms 2014

