



CS 412 Intro. to Data Mining

Chapter 1. Introduction

Hanghang Tong, Computer Science, Univ. Illinois at Urbana-Champaign, 2023



Data and Information Systems (DAIS)

□ Database Systems

□ Data Mining

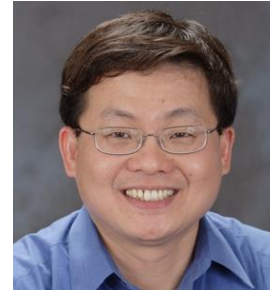
□ Networks

□ Text Information Systems

□ Healthcare



Jiawei Han



Kevin Chang



Yongjoo Park



Arindam Banerjee



Hari Sundaram



Hanghang Tong



ChengXiang Zhai

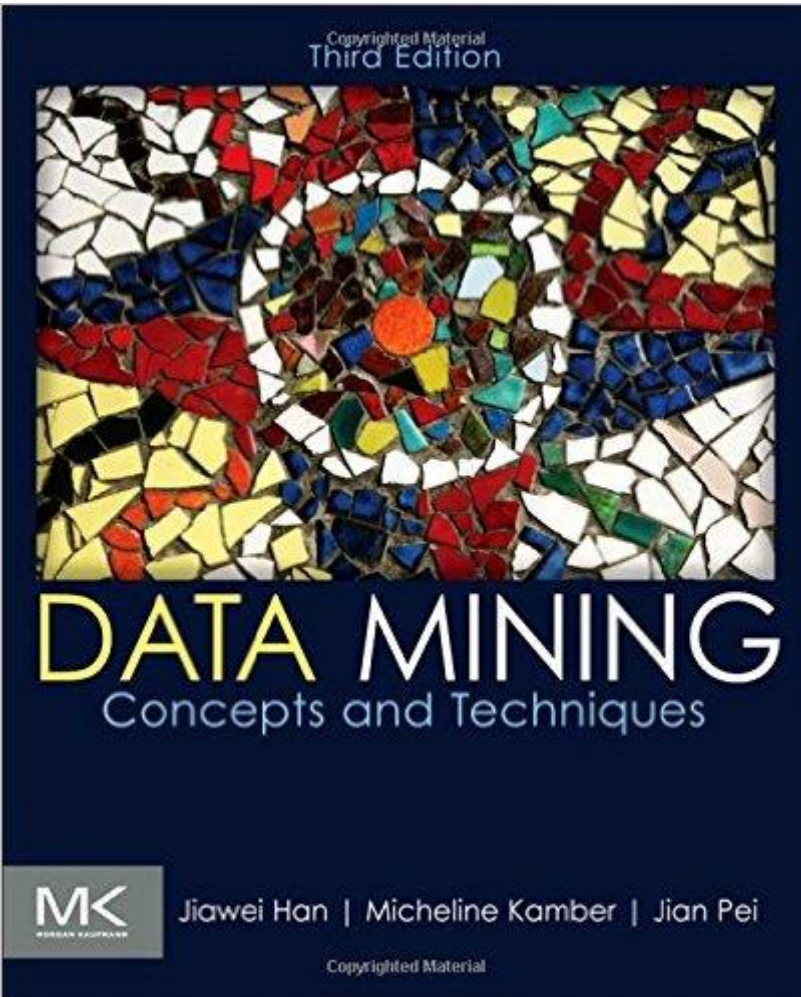


Jimeng Sun

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 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



- ❑ Textbook
 - ❑ Jiawei Han, Jian Pei and Hanghang Tong, *Data Mining: Concepts and Techniques (4rd ed)*, Morgan Kaufmann
- ❑ Class Homepage:
 - ❑ Canvas
 - ❑ <https://piazza.com/illinois/spring2023/cs412>

CS 412. Class Schedule

- ❑ **Class meetings: 2:00-3:15pm T/R, 3039 CIF**
- ❑ Take the best usage of class meeting
- ❑ No recorded videos (there are no recording facility in this classroom😞)

Teach Assistants



Wenxuan Bao

wbao4@



Mukesh Chugani

chugani2@



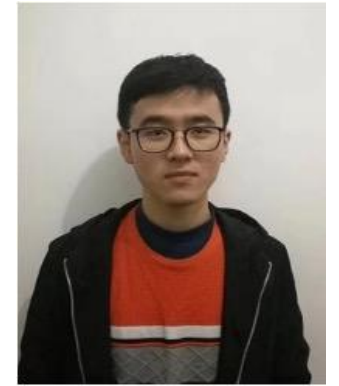
Derek Wang

dingsuw2@



Yian Wang

yian3@



Yuchen Yan

yucheny5@

- ❑ TA office hours:
 - ❑ **See next page**
- ❑ Wait list (No wait list at this time, keep attending class, see if there is space available)

CS412 Office Hours

- ❑ All office hours are in Central time zone, on Zooms.
- ❑ Each of us has a different zoom link & pwd
 - ❑ Hanghang Tong: <https://illinois.zoom.us/j/9635963030?pwd=OHg3RzBCU2ZrQXgxZmhDU3M2OTBLUT09>
 - ❑ Wenxuan Bao: <https://illinois.zoom.us/j/84990995262?pwd=MHcwQXJybXFDV3FnRGRQa213R08yQT09>
 - ❑ Mukesh Chugani: <https://illinois.zoom.us/j/83753824117?pwd=Vk1YSkcrMnp4bmFPNU54Z0FYZGRRUT09>
 - ❑ Derek Wang: <https://illinois.zoom.us/j/4112927402?pwd=ZmsxVzRWVm4raEt0WWRoSVhlUFNTQT09>
 - ❑ Yian Wang: <https://illinois.zoom.us/j/82532946827?pwd=TXptNk92ZFI1bWNGU2pORUVVYV1N5QT09>
 - ❑ Yuchen Yan: <https://illinois.zoom.us/j/9109085092?pwd=ZERodHBveTdBRDZ5MFhhVHV6OVV4Zz09>
- ❑ Schedule
 - ❑ Hanghang Tong: 9-10am, Tuesday
 - ❑ Wenxuan Bao: 2:30-3:15pm, Wednesday & 10:15-11:00am, Thursday
 - ❑ Mukesh Chugani: 6:15-7pm, Monday & Friday
 - ❑ Derek Wang: 3:30– 4:15pm, Tuesday & 10–10:45am, Friday
 - ❑ Yian Wang: 10–10:45am Wednesday & 3:30-4:15pm, Thursday
 - ❑ Yuchen Yan: 10-10:45am Monday & Wednesday

CS 412. Course Work and Grading

- ❑ Assignments, Programming Assignments, and Exams
 - ❑ Assignments: 50% (five homework assignments, 10% each, mixture of written and programming questions)
 - ❑ Midterm exams: 30% (two midterms, 15% each)
 - ❑ Final exam: 20%
 - ❑ For students taking 4th credit
 - ❑ For students registering 4 credits: 25%. The overall scores will be scaled proportionally
 - ❑ Group project: 2-3 members
 - ❑ Need help and/or discussions?
 - ❑ Sign on: Piazza (<https://piazza.com/illinois/spring2023/cs412>)
 - ❑ Check course material: lecture notes/homework/exam scores:
 - ❑ Canvas
- You can choose Python or Java or C++
 - This is NOT a programming class
- 2% extra credit will be given to the most active users on Piazza, in answering other users' questions.
 - Will be given to top 1-2% users based on Piazza statistics
 - Posting meaningless content might lead to disqualification of extra credit

4-Credit Projects

☐ Survey

- ☐ TA will release survey topics on Piazza
- ☐ Paper length: 10+reference

☐ Research project: Join a competition or propose your own project

- ☐ Example: <https://www.kaggle.com/competitions>
- ☐ Paper length: 6+reference

Choose
ONLY
one of
them

☐ Important Dates

- ☐ Project proposal due: Feb. 16th
- ☐ Mid-point report due: Mar. 21st
- ☐ Research project paper submission due: May 4th
- ☐ Survey submission due: May 4th

Lecture Schedule (subject to change)

- ❑ Class Outline / Chapter 1: Introduction (week 1)
- ❑ Chapter 2: Data, Measurements, and Data Preprocessing (weeks 1 & 2)
- ❑ Chapter 3: Data Warehousing and Online Analytical Processing (weeks 3 & 4)
- ❑ Chapter 4: Mining Frequent Patterns, Associations, and Correlations: Basic Concepts and Methods (weeks 4 & 5 & 6)
- ❑ Chapter 5: Advanced Pattern Mining (weeks 7 & 8)
- ❑ Chapter 6: Classification: Basic Concepts (weeks 10 & 11, 12)
- ❑ Chapter 8: Cluster Analysis: Basic Concepts (weeks 13 & 14)
- ❑ Chapter 10: Deep Learning (week 15)

Key Dates

☐ Assignments

- ☐ A1: Jan. 17th out, Feb. 2nd due
- ☐ A2: Feb. 2nd, Feb. 14th due
- ☐ A3: Feb. 14th, Mar. 23rd due
- ☐ A4: Mar. 23rd out, Apr. 18th due
- ☐ A5: Apr. 18th out, May. 4th due

☐ Exams

- ☐ Mid-term 1: 2:00-3:15pm, Feb. 21st, (in class)
- ☐ Mid-term 2: 2:00-3:15pm, Mar. 30th, (in class)
- ☐ Final: TBD

☐ Project/survey (for students taking 4th credit)

- ☐ Project/survey proposal due: Feb. 16th
- ☐ Mid-point report due: Mar. 21st
- ☐ Paper/survey submission due: May 4th

- We cannot make any changes to these key dates
- We cannot accept any late submissions or arrange make-up exam, except for genuine, verifiable emergence.
- Please mark them in your calendar

Letter Grade Cut-offs

- ❑ The following cutoffs represent what will be *likely* used to generate the letter grades:
- ❑

A+ $\geq 98\%$	A $\geq 94\% \ \& \ < 98\%$	A- $\geq 90\% \ \& \ < 94\%$
B+ $\geq 85\% \ \& \ < 90\%$	B $\geq 80\% \ \& \ < 85\%$	B- $\geq 77\% \ \& \ < 80\%$
C+ $\geq 74\% \ \& \ < 77\%$	C $\geq 70\% \ \& \ < 74\%$	C- $\geq 67\% \ \& \ < 70\%$
D $\geq 60\% \ \& \ < 67\%$	F $< 60\%$	
- ❑ The above cutoffs are tentatively and may be adjusted *slightly*; However, there will be *no general curve-fitting* in assigning the final grades.
- ❑ If there is any adjustment of the above cutoffs, we will NOT curve down your letter grades.

Assignment Policies

- ❑ The homework is due at **11:59 PM CT** on the due date.
- ❑ We will be using Canvas: <https://canvas.illinois.edu/> for collecting homework assignments.
- ❑ Please do not hand in a scan of your handwritten solution, only the typed solution (e.g., Microsoft Word, Latex, etc.) will be graded.
- ❑ Contact the TAs if you are having technical difficulties in submitting the assignment. We do NOT accept late homework. Any late submission will receive a zero grade.
- ❑ The homework should be submitted as a single pdf file using the name convention: yourFirstName-yourLastName.pdf.

Class Policies (cont.)

❑ Academic Integrity Policy:

- ❑ We have **zero tolerance** on any violation
- ❑ Feel free to discuss with other members of the class when doing the homework. You should, however, write down your own solution **independently**. Please note
 - ❑ there is a fine line between collaboration and completing the assignment by yourself.
 - ❑ Aiding others to cheat would have the same consequence as the cheating itself.

❑ Assuring Non-Hostile Work Environment

- ❑ In order to assure a non-hostile work environment for course staff, we will strictly enforce the following policy for assessment, including exams, assignments and course project. Any assessment containing language that conventionally would be judged as obscene, threatening violence, or of a clearly derogatory nature will be given a zero grade without further grading.


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?

- ❑ 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

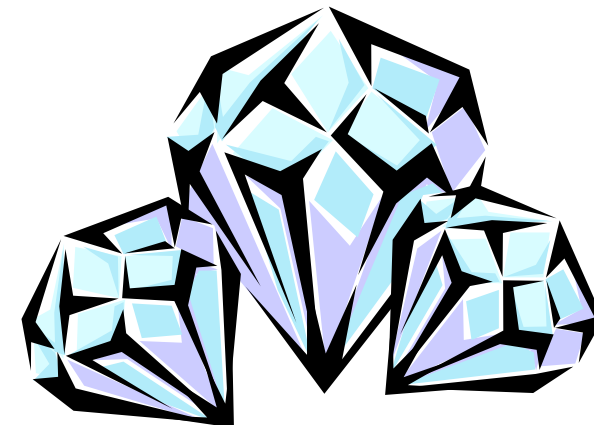
Chapter 1. Introduction

- ❑ 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?

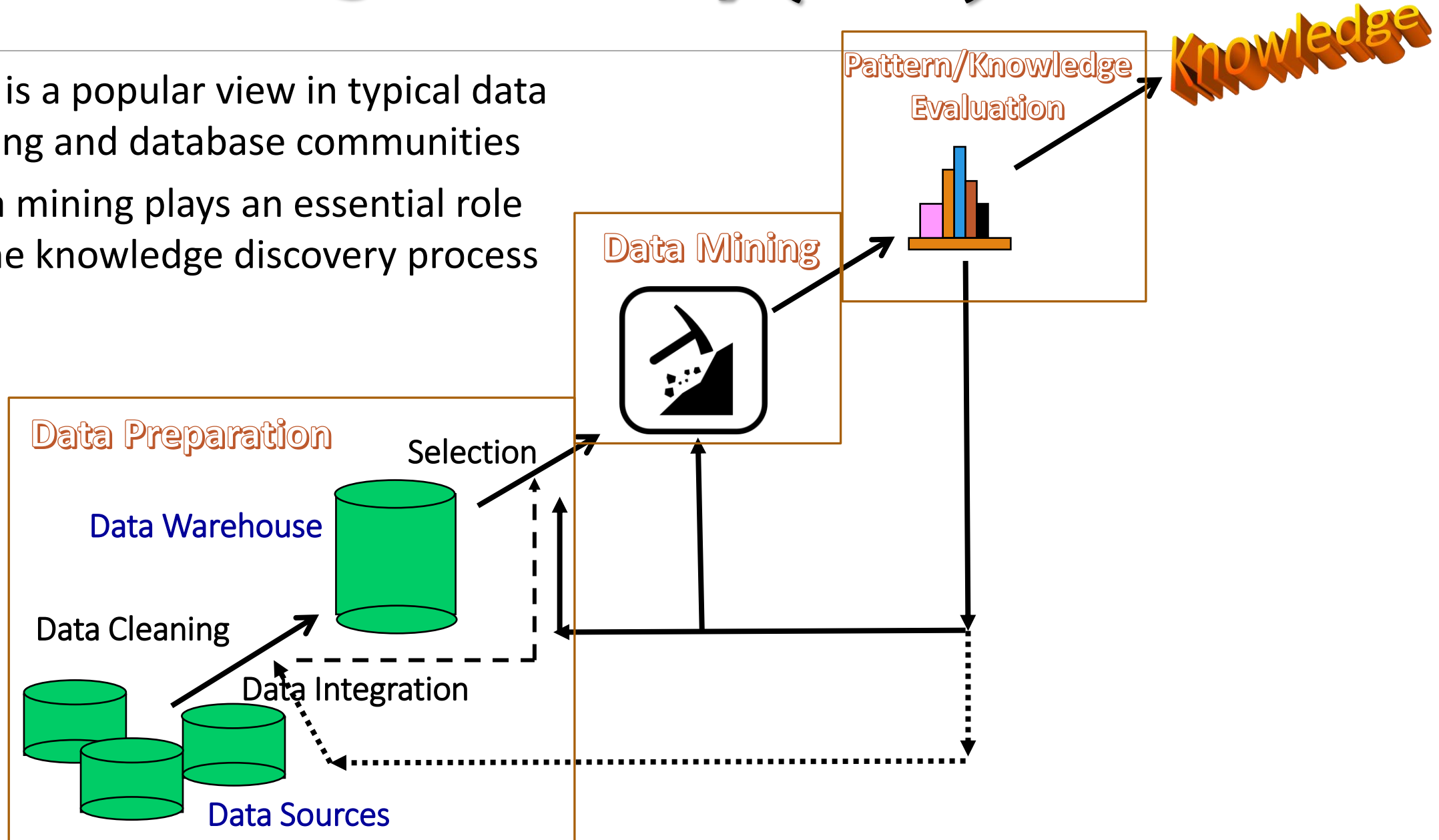


- ❑ Data mining (knowledge discovery from data)
 - ❑ Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data
 - ❑ Data mining: a misnomer?
- ❑ Alternative names
 - ❑ Knowledge discovery (mining) in databases (KDD), knowledge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



Knowledge Discovery (KDD) Process

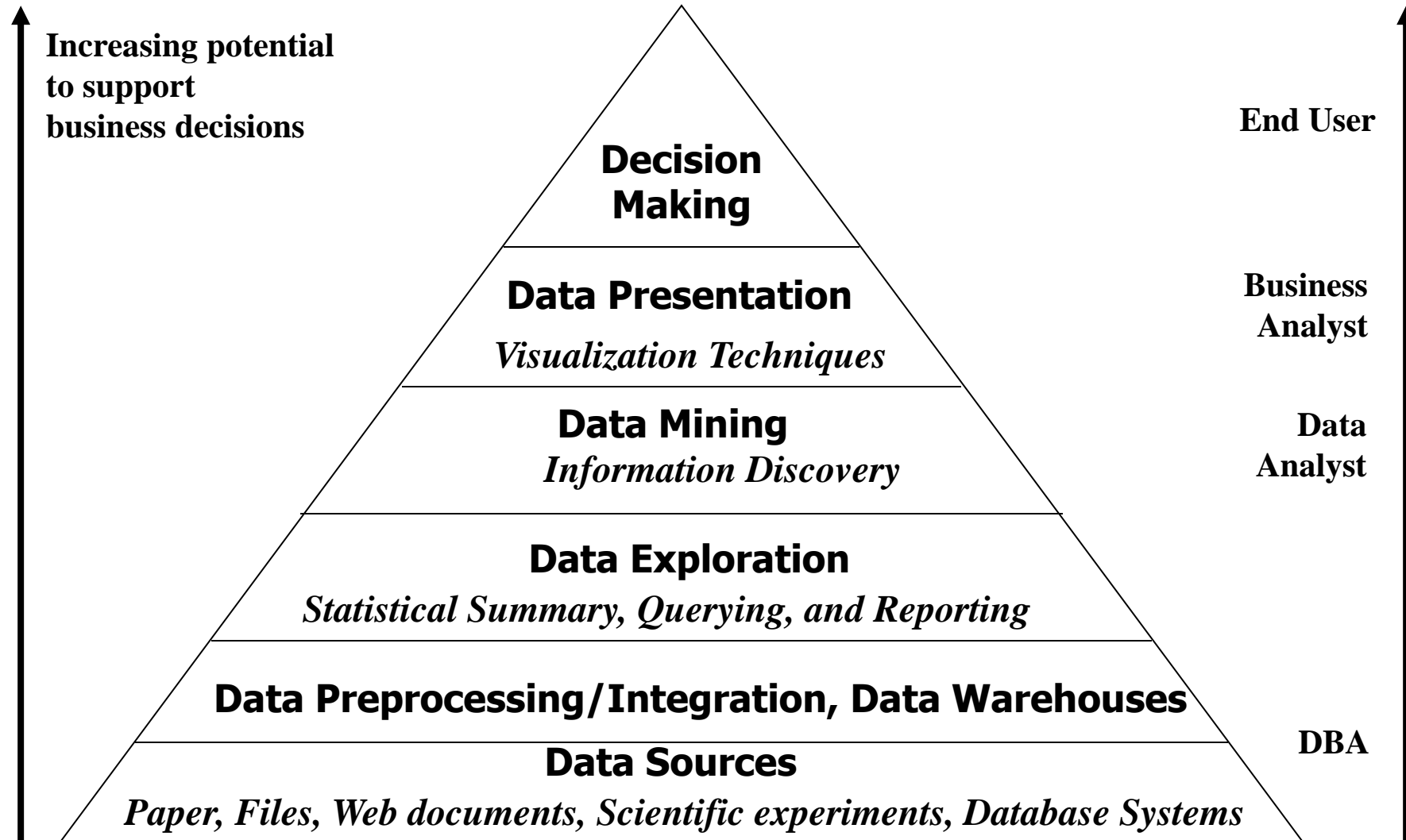
- ❑ This is a popular view in typical data mining and database communities
- ❑ Data mining plays an essential role in the knowledge discovery process



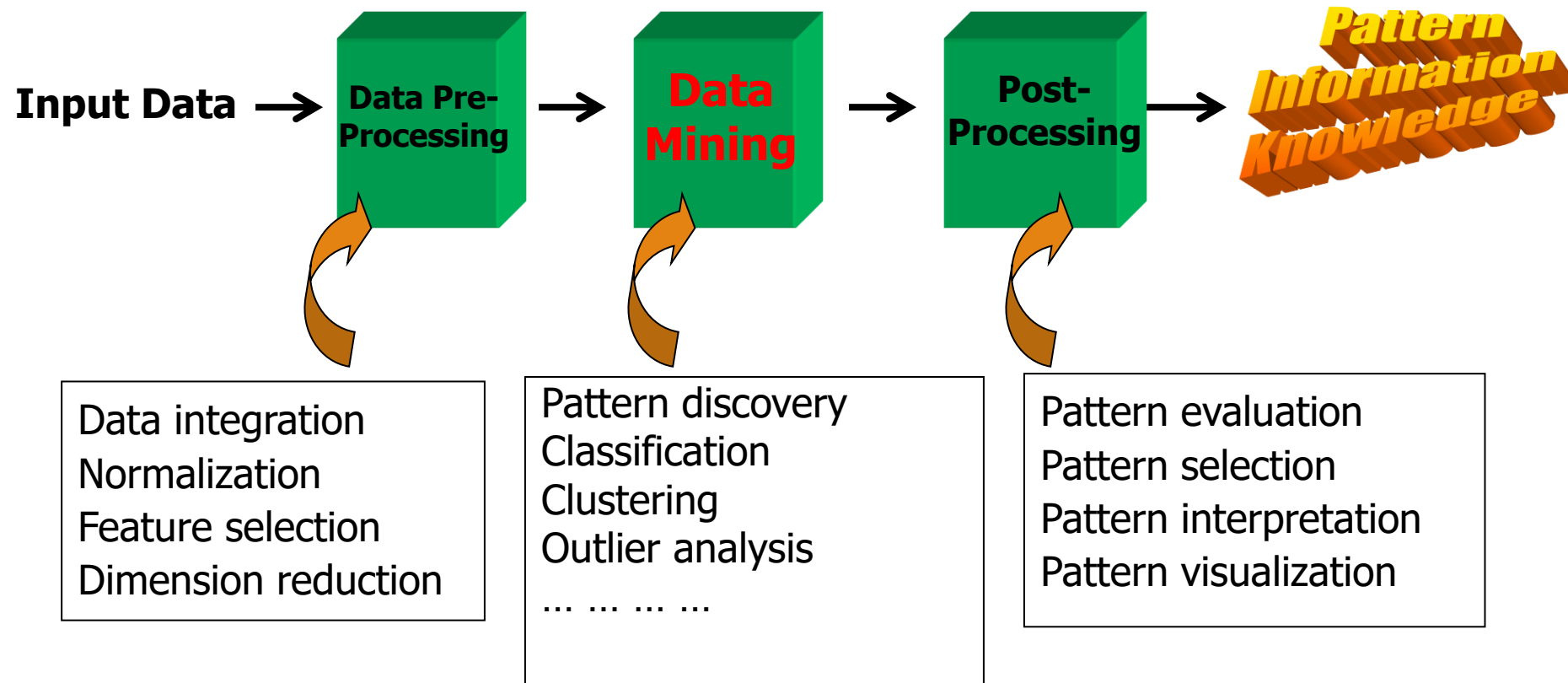
Example: A Web Mining Framework

- ❑ 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

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.

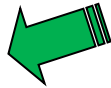
Chapter 1. Introduction

- ❑ 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

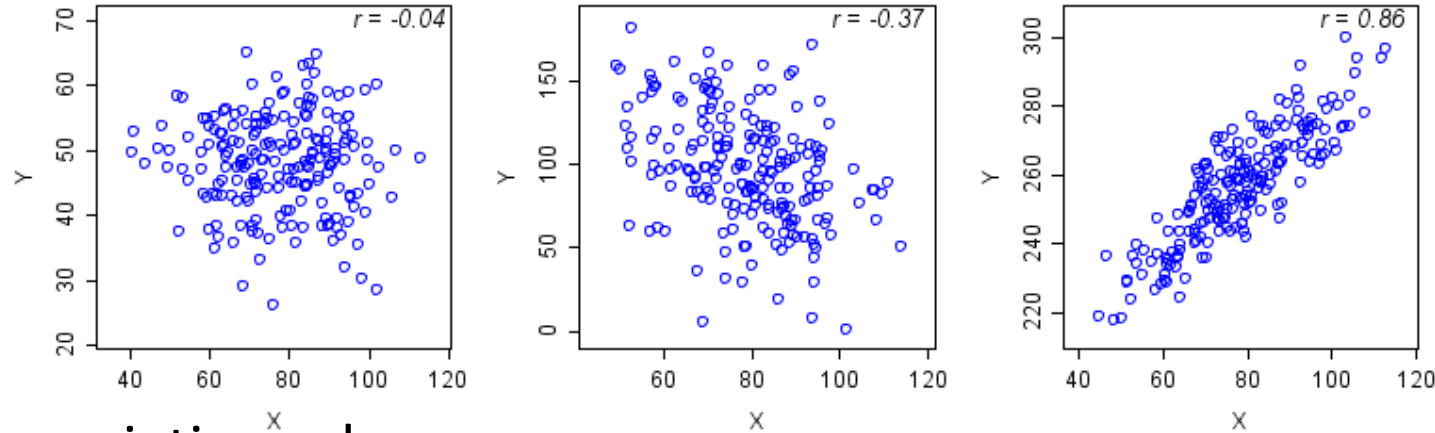
Data Mining Functions: (1) Generalization

- ❑ 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

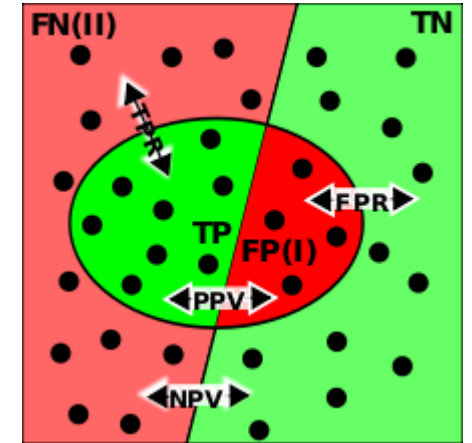
- Frequent patterns (or frequent itemsets)
 - What items are frequently purchased together in your Walmart?
- Association and Correlation Analysis



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

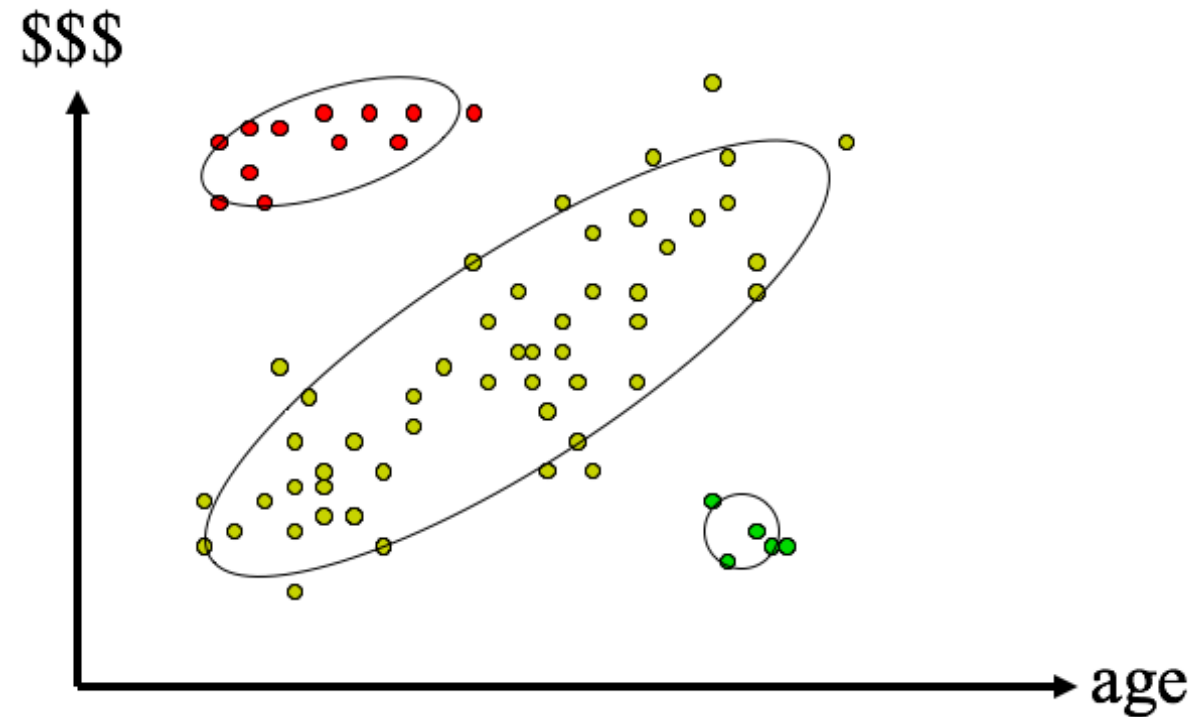
Data Mining Functions: (3) Classification

- ❑ 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

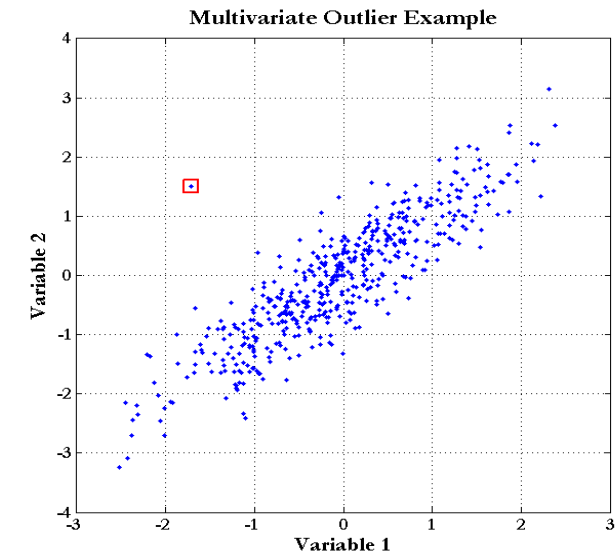
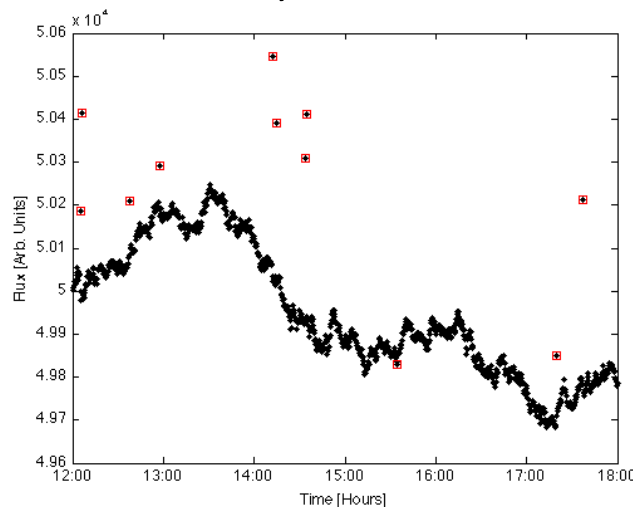
- ❑ Unsupervised learning (i.e., Class label is unknown)
- ❑ Group data to form new categories (i.e., clusters), e.g., cluster houses to find distribution patterns
- ❑ Principle: Maximizing intra-class similarity & minimizing interclass similarity
- ❑ Many methods and applications



Data Mining Functions: (5) Outlier Analysis

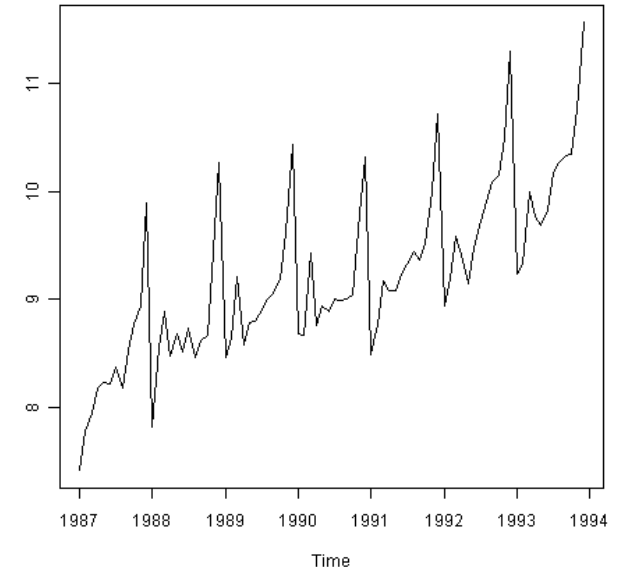
❑ Outlier analysis

- ❑ 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



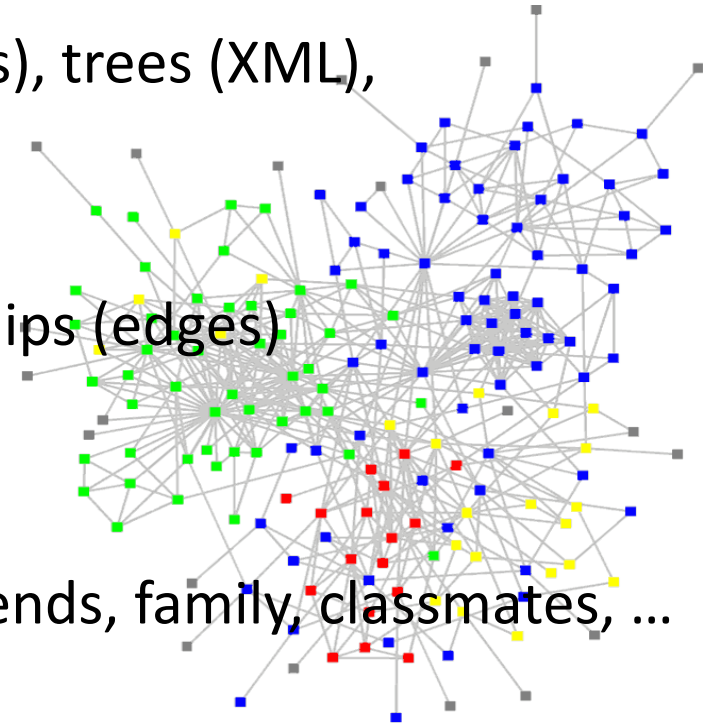
Data Mining Functions: (6) Time and Ordering: Sequential Pattern, Trend and Evolution Analysis

- 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, ...

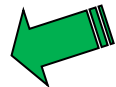


Evaluation of Knowledge

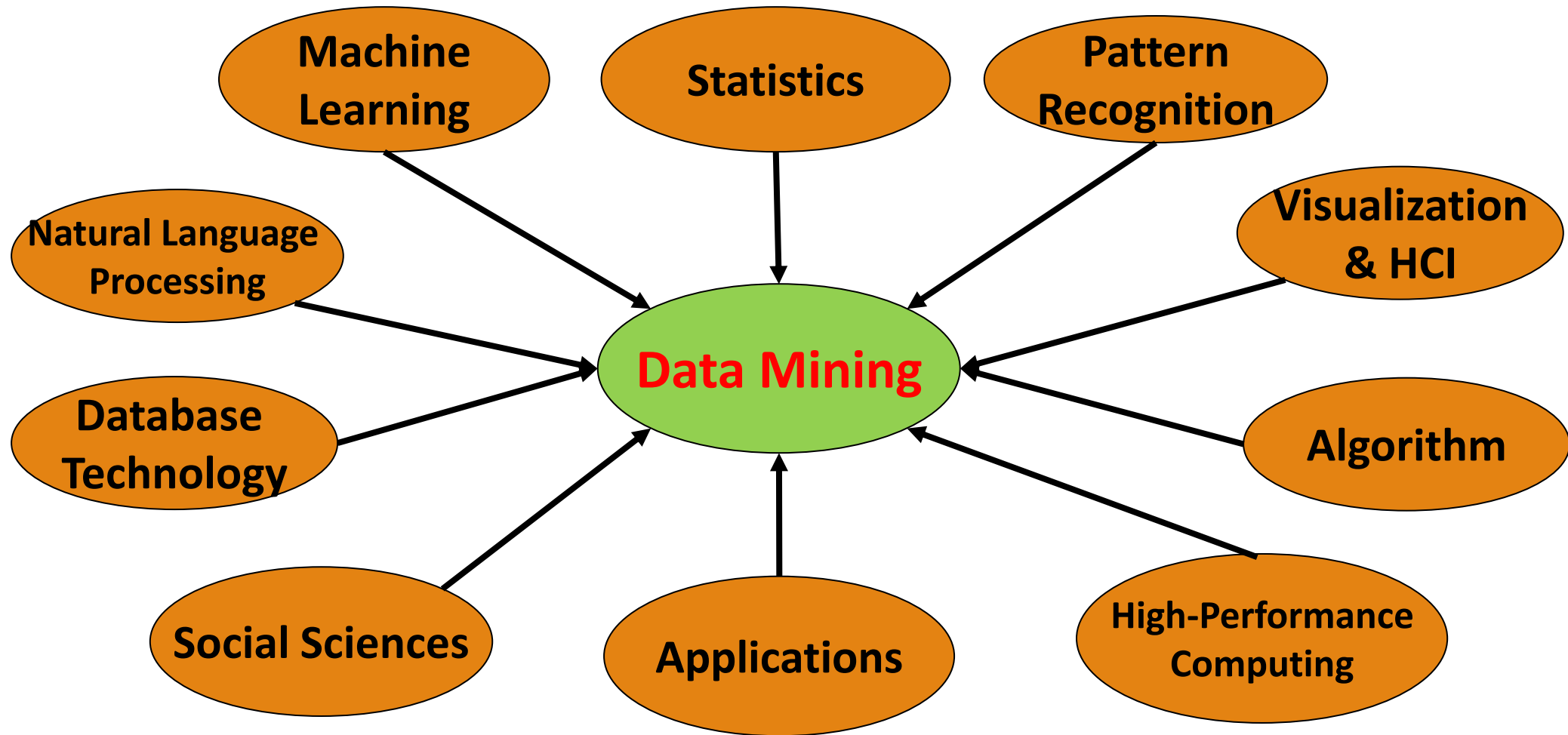
- ❑ Are all mined knowledge interesting?
 - ❑ One can mine tremendous amount of “patterns”
 - ❑ Some may fit only certain dimension space (time, location, ...)
 - ❑ Some may not be representative, may be transient, ...
- ❑ Evaluation of mined knowledge → directly mine only interesting knowledge?
 - ❑ 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

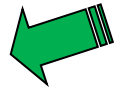
- ❑ Why Data Mining?
- ❑ What Is Data Mining?
- ❑ A Multi-Dimensional View of Data Mining
- ❑ What Kinds of Data Can Be Mined?
- ❑ What Kinds of Patterns Can Be Mined?
- ❑ What Kinds of Technologies Are Used?
- ❑ What Kinds of Applications Are Targeted? 
- ❑ Major Issues in Data Mining
- ❑ A Brief History 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 In, 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)

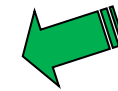
- ❑ Mining Methodology
 - ❑ Mining various and new kinds of knowledge
 - ❑ Mining knowledge in multi-dimensional space
 - ❑ Data mining: An interdisciplinary effort
 - ❑ Boosting the power of discovery in a networked environment
 - ❑ Handling noise, uncertainty, and incompleteness of data
 - ❑ Pattern 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
 - ❑ Fairness of 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, NeuIPS
- PR conferences: CVPR,

■ Journals

- Data Mining and Knowledge Discovery (DAMI or DMKD)
- IEEE Trans. On Knowledge and Data Eng. (TKDE)
- SIGKDD 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, J. Pei, and H. Tong, Data Mining: Concepts and Techniques. Morgan Kaufmann, 4th ed. , 2022
- ❑ 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, Cambridge University Press, 2014

