Data Mining:

Concepts and Techniques

(3rd ed.)

— Chapter 1 —

Jiawei Han, Micheline Kamber, and Jian Pei

University of Illinois at Urbana-Champaign &

Simon Fraser University

Chapter 1. Introduction

- Why Data Mining?

- What Is Data Mining?

- A Multi-Dimensional View of Data Mining

- What Kind of Data Can Be Mined?

- What Kinds of Patterns Can Be Mined?

- What Technology Are Used?

- What Kind 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

Evolution of Sciences

- Before 1600, **empirical science**
- 1600-1950s, **theoretical science**
    - Each discipline has grown a *theoretical* component. Theoretical models often motivate experiments and generalize our understanding.
- 1950s-1990s, **computational science**
    - Over the last 50 years, most disciplines have grown a third, *computational* branch (e.g. empirical, theoretical, and computational ecology, or physics, or linguistics.)
    - Computational Science traditionally meant simulation. It grew out of our inability to find closed-form solutions for complex mathematical models.
- 1990-now, **data science**
    - The flood of data from new scientific instruments and simulations
    - The ability to economically store and manage petabytes of data online
    - The Internet and computing Grid that makes all these archives universally accessible
    - Scientific info. management, acquisition, organization, query, and visualization tasks scale almost linearly with data volumes. Data mining is a major new challenge!
- Jim Gray and Alex Szalay, *The World Wide Telescope: An Archetype for Online Science*, Comm. ACM, 45(11): 50-54, Nov. 2002

Evolution of Database Technology

- 1960s:
    - Data collection, database creation, IMS and network DBMS
- 1970s:
    - Relational data model, relational DBMS implementation
- 1980s:
    - RDBMS, advanced data models (extended-relational, OO, deductive, etc.)
    - Application-oriented DBMS (spatial, scientific, engineering, etc.)
- 1990s:
    - Data mining, data warehousing, multimedia databases, and Web databases
- 2000s
    - Stream data management and mining
    - Data mining and its applications
    - Web technology (XML, data integration) and global information systems

Chapter 1. Introduction

- Why Data Mining?

- What Is Data Mining?

- A Multi-Dimensional View of Data Mining

- What Kind of Data Can Be Mined?

- What Kinds of Patterns Can Be Mined?

- What Technology Are Used?

- What Kind of Applications Are Targeted?

- Major Issues in Data Mining

- A Brief History of Data Mining and Data Mining Society

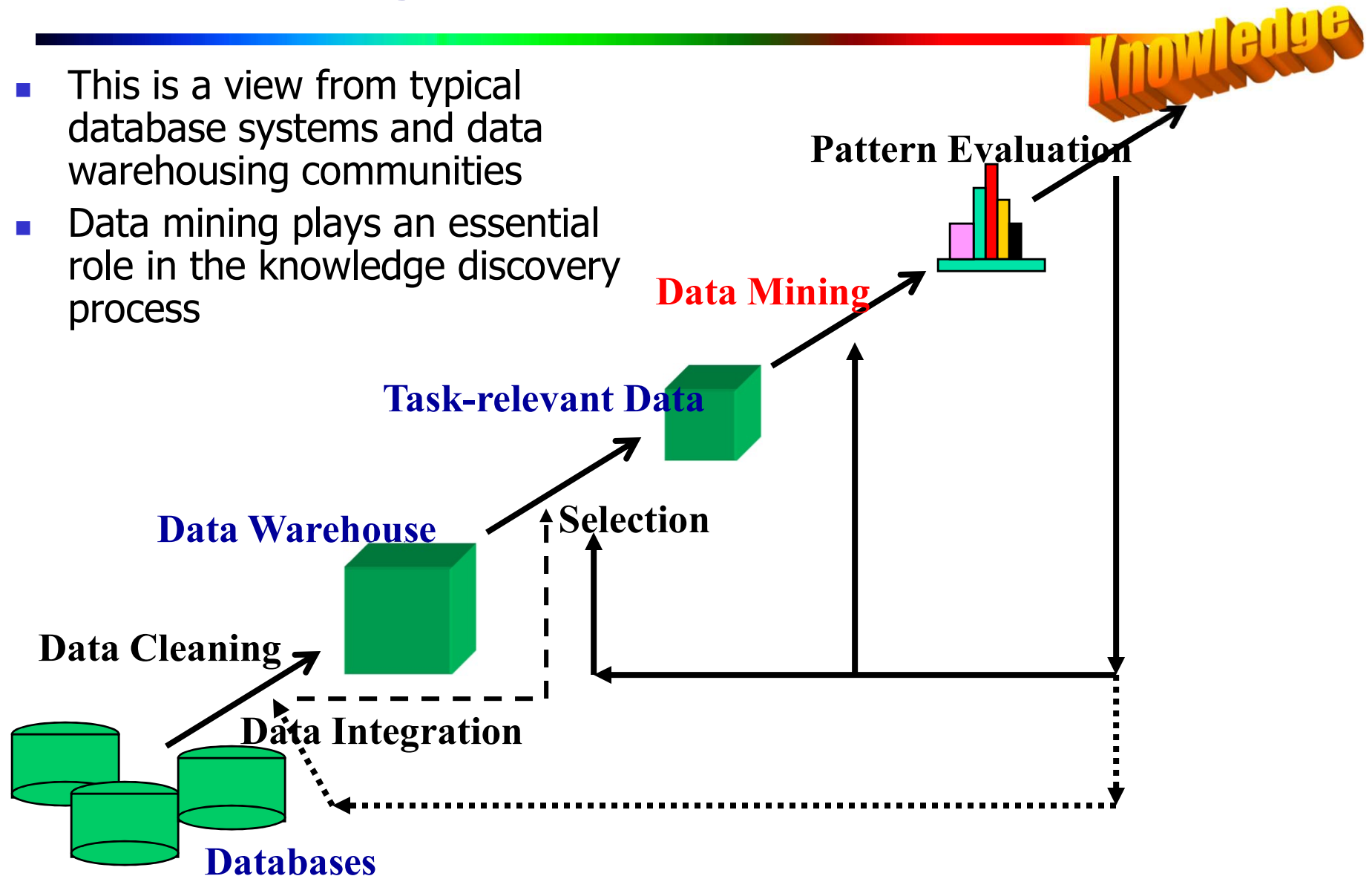- Summary

What Is Data Mining?

- Data mining (knowledge discovery from data)
  - Extraction of interesting (<u>non-trivial,</u> <u>implicit</u>, <u>previously</u> <u>unknown</u> and <u>potentially useful)</u> 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
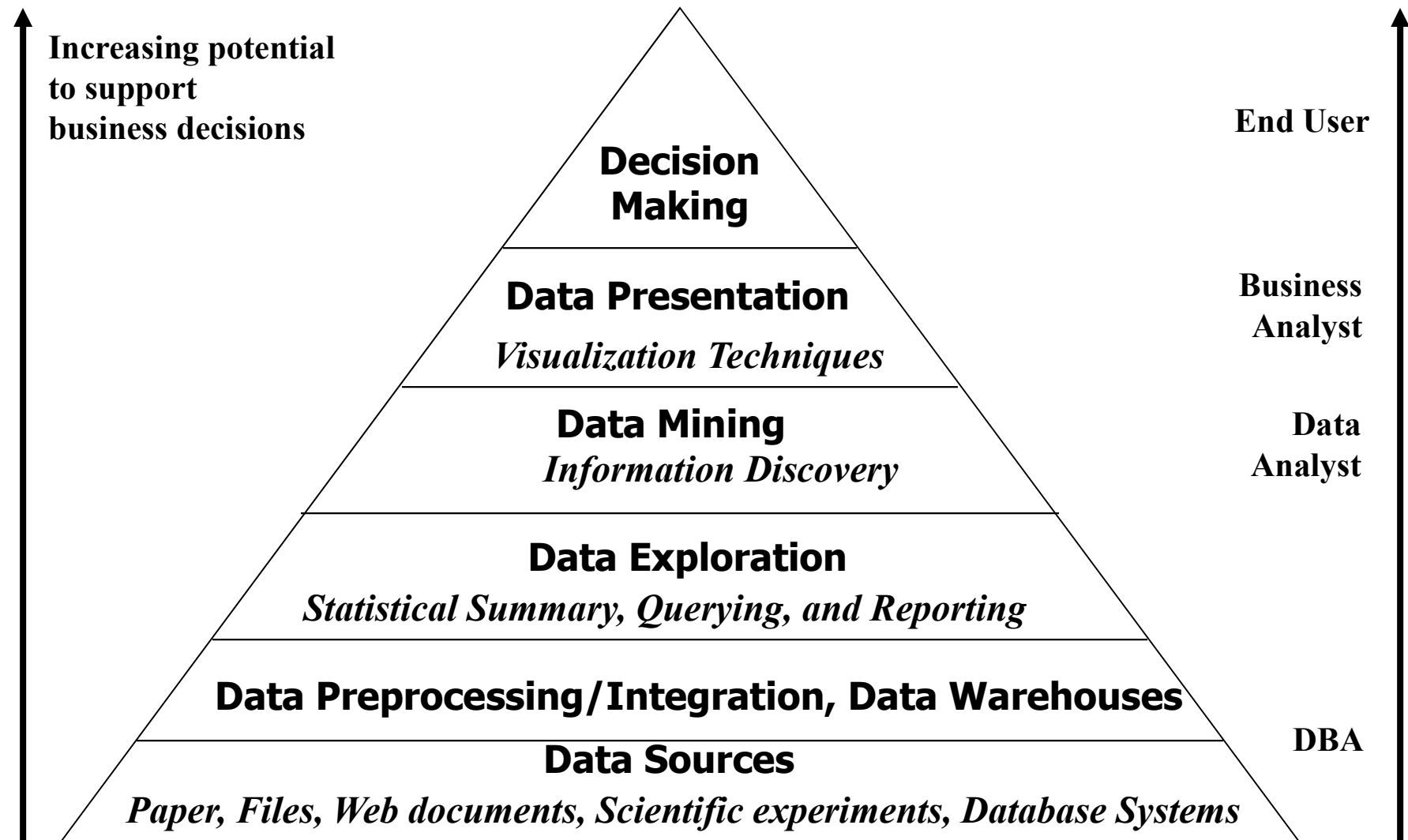
Knowledge Discovery (KDD) Process

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

Knowledge

Pattern Evaluation

Data Mining

Task-relevant Data

Data Warehouse

Selection

Data Cleaning
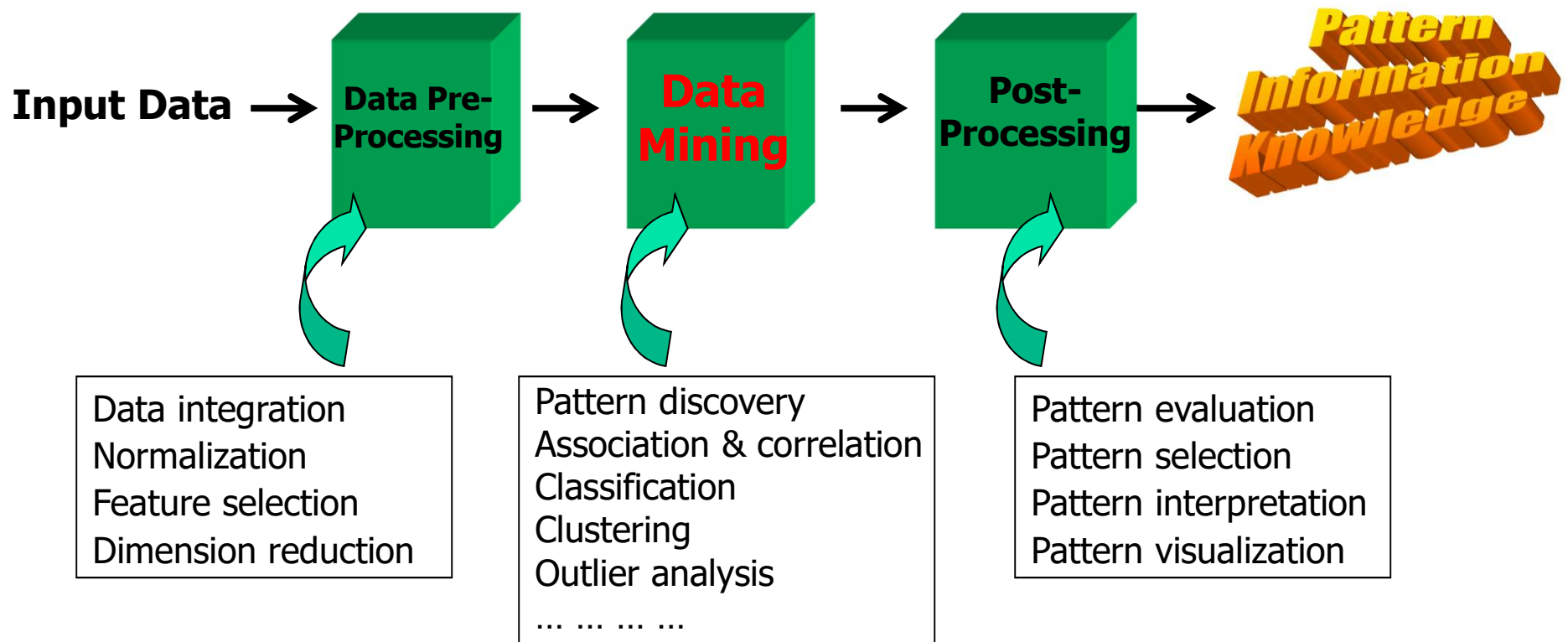
Data Integration

Databases

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

Increasing potential
to support
business decisions

Decision
Making

Data Presentation
Visualization Techniques

Data Mining
Information Discovery

Data Exploration
Statistical Summary, Querying, and Reporting

Data Preprocessing/Integration, Data Warehouses

Data Sources
Paper, Files, Web documents, Scientific experiments, Database Systems

End User

Business
Analyst

Data
Analyst

DBA

KDD Process: A Typical View from ML and Statistics

Input Data → **Data Pre-Processing** → **Data Mining** → **Post-Processing** → *Pattern Information Knowledge*

Data integration	Pattern discovery	Pattern evaluation
Normalization	Association & correlation	Pattern selection
Feature selection	Classification	Pattern interpretation
Dimension reduction	Clustering	Pattern visualization
	Outlier analysis	
	… … … …	

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

Example: Medical Data Mining

- Health care & medical data mining – often adopted such a view in statistics and machine learning

- Preprocessing of the data (including feature extraction and dimension reduction)

- Classification or/and clustering processes

- Post-processing for presentation

Chapter 1. Introduction

- Why Data Mining?

- What Is Data Mining?

- A Multi-Dimensional View of Data Mining

- What Kind of Data Can Be Mined?

- What Kinds of Patterns Can Be Mined?

- What Technology Are Used?

- What Kind 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

- **<u>Data to be mined</u>**
  - 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
- **<u>Knowledge to be mined (or: Data mining functions)</u>**
  - Characterization, discrimination, association, classification, clustering, trend/deviation, outlier analysis, etc.
  - Descriptive vs. predictive data mining
  - Multiple/integrated functions and mining at multiple levels
- **<u>Techniques utilized</u>**
  - Data-intensive, data warehouse (OLAP), machine learning, statistics, pattern recognition, visualization, high-performance, etc.
- **<u>Applications adapted</u>**
  - 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 Kind of Data Can Be Mined? ⬅

- What Kinds of Patterns Can Be Mined?

- What Technology Are Used?

- What Kind 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
- 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 multi-linked data
  - Object-relational databases
  - Heterogeneous databases and legacy databases
  - Spatial data and spatiotemporal data
  - Multimedia database
  - Text databases
  - The World-Wide Web

Relational Database: Example

- The fictitious AllElectronics is described by the following relation tables:
  - customer, item, employee, and branch

- The relation customer consists of a set of attributes describing the customer information, including a unique customer identity number (cust ID), customer name, address, age, occupation, annual income, credit information, and category.

- Similarly, each of the relations item, employee, and branch consists of a set of attributes describing the properties of these entities.

- Tables can also be used to represent the relationships between or among multiple entities:
  - purchases (customer purchases items, creating a sales transaction handled by an employee)
  - items sold (lists items sold in a given transaction)
  - works at (employee works at a branch of AllElectronics)

AllElectronics database

- Relational schema for the AllElectronics relational database, :

customer	(cust_ID, name, address, age, occupation, annual_income, credit_information, category, ...)
item	(item_ID, brand, category, type, price, place_made, supplier, cost, ...)
employee	(empl_ID, name, category, group, salary, commission, ...)
branch	(branch_ID, name, address, ...)
purchases	(trans_ID, cust_ID, empl_ID, date, time, method_paid, amount)
items_sold	(trans_ID, item_ID, qty)
works_at	(empl_ID, branch_ID)

Query vs. mining

- Relational data can be accessed by database queries written in a relational query language (e.g., SQL) .

- A given query is transformed into a set of relational operations, such as join, selection, and projection. A query allows retrieval of specified subsets of the data.
- Example: "Show me a list of all items that were sold in the last quarter"

- Relational languages also use aggregate functions such as sum, avg, count, max, and min. Using aggregates allows you to ask:

- "Show me the total sales of the last month, grouped by branch,"
- "How many sales transactions occurred in the month of December?"
- "Which salesperson had the highest sales?"
- ...

Query vs. mining (cont.)

- When **mining** relational databases, we can go further by searching for trends or data patterns.

- For example, data mining systems can analyze customer data to predict the credit risk of new customers based on their income, age, and previous credit information.

- Data mining systems may also detect deviations—that is, items with sales that are far from those expected in comparison with the previous year.

- Such deviations can then be further investigated. For example, data mining may discover that there has been a change in packaging of an item or a significant increase in price.

Data Warehouse

- Suppose that AllElectronics is a successful international company with branches around the world. Each branch has its own set of databases.

- The president of AllElectronics has asked you to provide an analysis of the company's sales per item type per branch for the third quarter. This is a difficult task, particularly since the relevant data are spread out over several databases physically located at numerous sites.

- A data warehouse is a repository of information collected from multiple sources, stored under a unified schema, and usually residing at a single site.

- Data warehouses are constructed via a process of data cleaning, data integration, data transformation, data loading, and periodic data refreshing.

Data Warehouse (cont.)

- To facilitate decision making, the data in a data warehouse are organized around major subjects (e.g., customer, item, supplier, and activity).

- The data are stored to provide information from a historical perspective, such as in the past 6 to 12 months, and are typically **summarized**. For example, rather than storing the details of each sales transaction, the data warehouse may store a summary of the transactions per item type for each store.
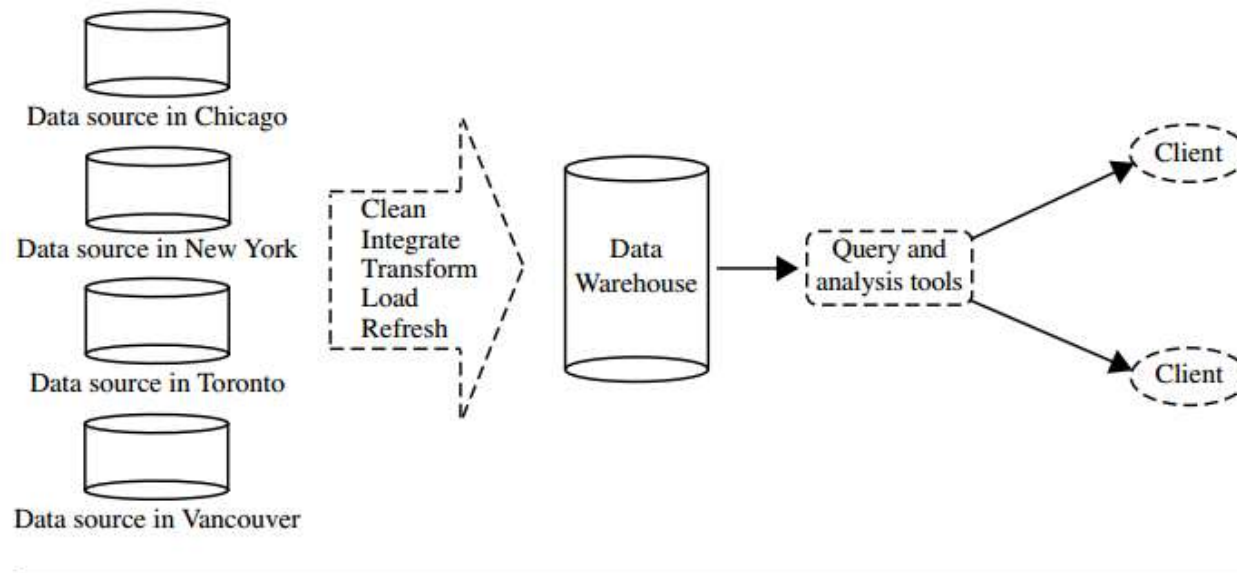
Extract, Transform, and Load

- **ETL** in data mining consists of the construction of new data subsets derived from existing data sources.

- **ETL** stands for the whole process of taking data from various sources and combining it, transforming it, and loading big data using database tools.

- **Extract** is to get data out of different data sources.
- **Transform** means to change the data format in order to better support querying and analysis.
- **Load** is to get this data into a target storage.

- **data warehouse** is a system that actually performs some ETL operations: extract, clean, conform and deliver source data into a dimensional data store and then support and implement querying and analysis for the purpose of decision making.

Chapter 1. Introduction

- Why Data Mining?

- What Is Data Mining?

- A Multi-Dimensional View of Data Mining

- What Kind of Data Can Be Mined?

- What Kinds of Patterns Can Be Mined?

- What Technology Are Used?

- What Kind of Applications Are Targeted?

- Major Issues in Data Mining

- A Brief History of Data Mining and Data Mining Society

- Summary

Data Mining Function: (1) Generalization

- Information integration and data warehouse construction
  - Data cleaning, transformation, integration, and multidimensional data model
- Multidimensional concept description: Characterization and discrimination
  - Generalize, summarize, and contrast data characteristics
  - For example, in the *AllElectronics* store, classes of items for sale include *computers* and *printers*, and concepts of customers include *bigSpenders* and *budgetSpenders*.
  - It can be useful to describe individual classes and concepts in summarized, concise, and yet precise terms.

Data Mining Function: (2) 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
    - Bread → Butter [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 Function: (3) 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: (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 Function: (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

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., first buy digital camera, then buy large SD 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

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" 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
    - …

Chapter 1. Introduction

- Why Data Mining?

- What Is Data Mining?

- A Multi-Dimensional View of Data Mining

- What Kind of Data Can Be Mined?

- What Kinds of Patterns Can Be Mined?

- What Technology Are Used?

- What Kind 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
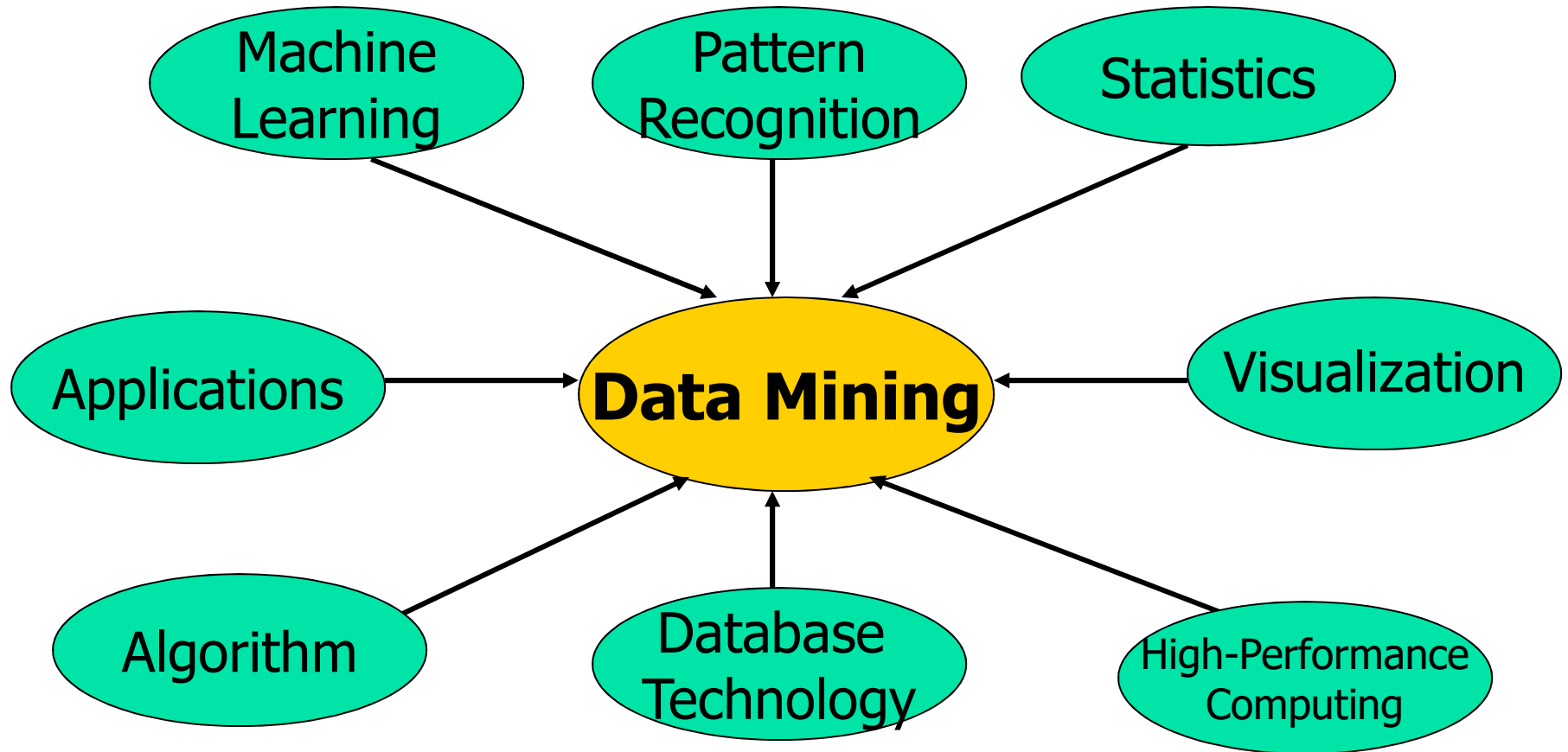
Why Confluence of Multiple Disciplines?

- **Tremendous amount of data**
  - Algorithms must be highly scalable to handle such as tera-bytes of 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 networks and multi-linked data
  - Heterogeneous databases and legacy databases
  - 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 Kind of Data Can Be Mined?

- What Kinds of Patterns Can Be Mined?

- What Technology Are Used?

- What Kind 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: 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)

- From major dedicated data mining systems/tools (e.g., SAS, MS SQL-Server Analysis Manager, Oracle Data Mining Tools) to invisible data mining

Chapter 1. Introduction

- Why Data Mining?

- What Is Data Mining?

- A Multi-Dimensional View of Data Mining

- What Kind of Data Can Be Mined?

- What Kinds of Patterns Can Be Mined?

- What Technology Are Used?

- What Kind 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 Kind of Data Can Be Mined?

- What Kinds of Patterns Can Be Mined?

- What Technology Are Used?

- What Kind 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), etc.
- ACM Transactions on KDD starting in 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

Chapter 1. Introduction

- Why Data Mining?

- What Is Data Mining?

- A Multi-Dimensional View of Data Mining

- What Kind of Data Can Be Mined?

- What Kinds of Patterns Can Be Mined?

- What Technology Are Used?

- What Kind 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 database 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, outlier and trend analysis, etc.

- Data mining technologies and applications

- Major issues in data mining