Business Intelligence

© Han, J., Kamber, M. and Pei, J., 2011. Data mining concepts and techniques 3rd edition.*The Morgan Kaufmann Series in Data Management Systems*.

Course Outline

- Data Mining Concepts

- An Overview of Business Intelligence Technology

- Getting to Know Your Data

- R Programming Basics

- Data Preprocessing

- Data Warehousing and Online Analytical Processing

- Mining Frequent Patterns, Associations, and Correlations

- Classification: Basic Concepts

- Cluster Analysis: Basic Concepts and Methods

- Outlier Detection

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 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

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

What is (not) Data Mining?
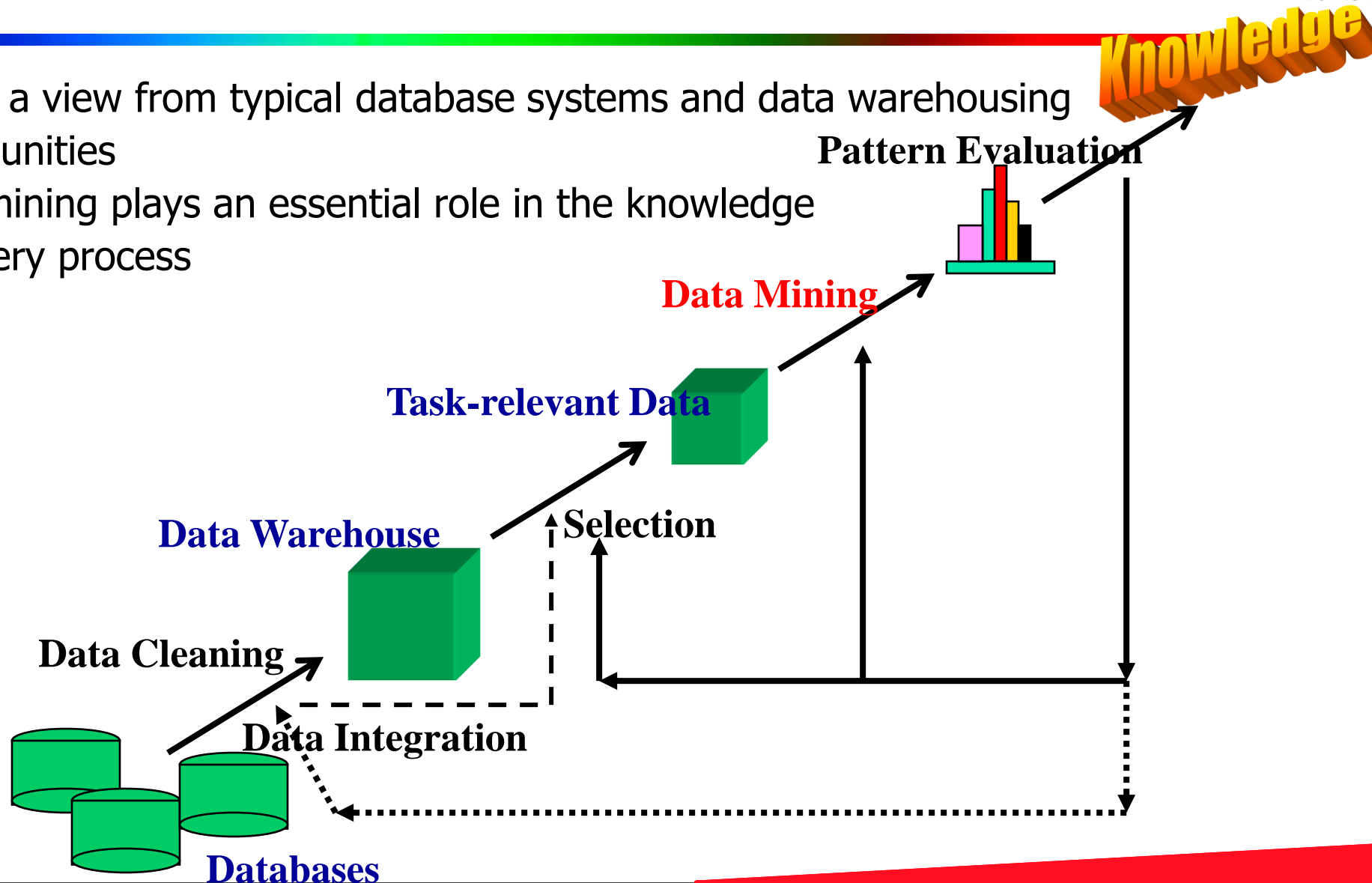
- **What is not Data Mining?**
  - Simple Querying

  - Done with a predetermined hypothesis.

  - Examples
    - Look up phone number in phone directory

    - Query a Web search engine for information about "Amazon"

- **What is Data Mining?**
  - Extraction of patterns and Knowledge

  - Examples:
    - Certain Cars are more prevalent in Colorado Springs (Tesla, Chevy)

    - Group together similar documents returned by search engine according to their context (e.g. Amazon rainforest, Amazon.com,)

Knowledge Discovery (KDD) Process

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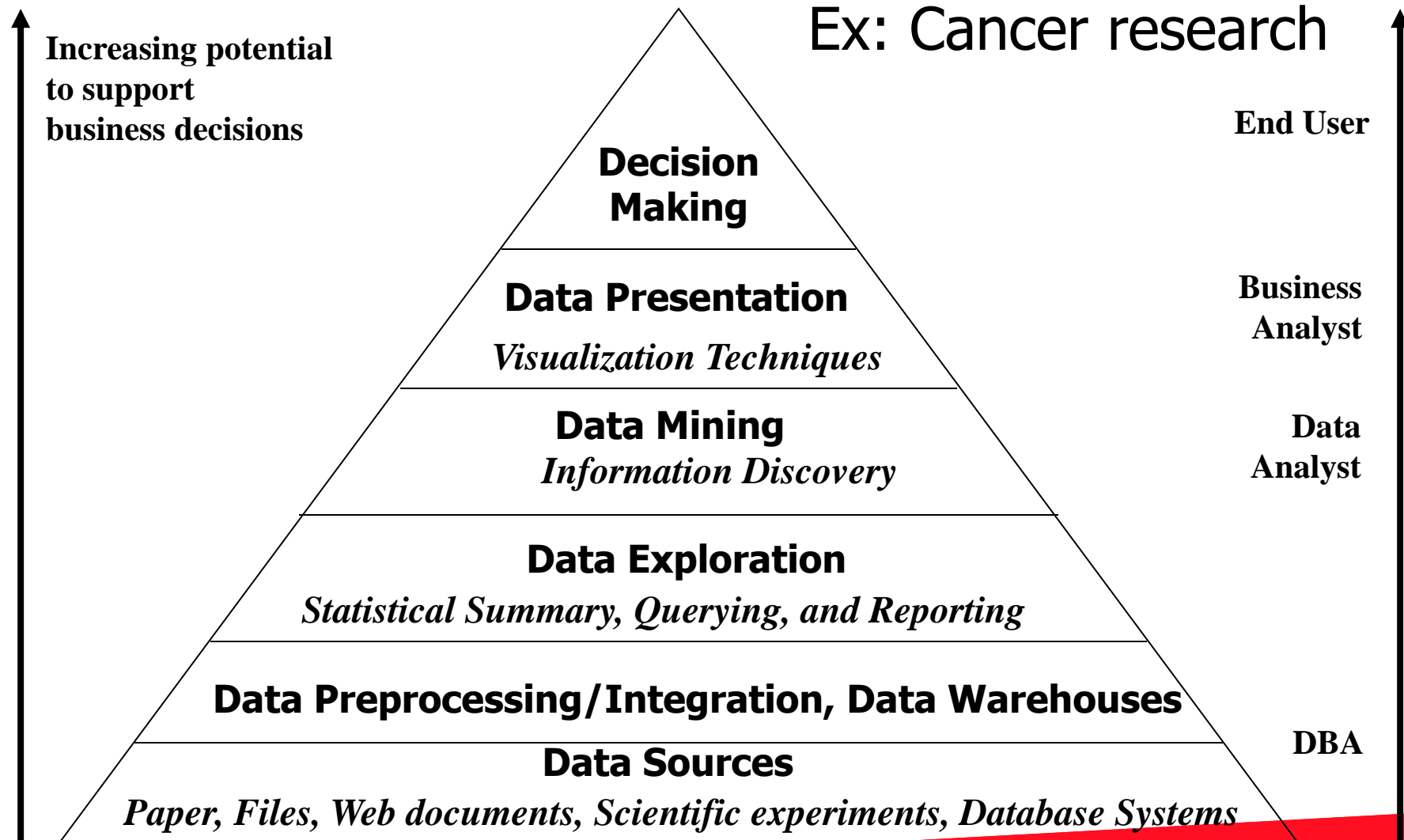- Data mining plays an essential role in the knowledge discovery process

Knowledge

Pattern Evaluation

Data Mining

Task-relevant Data

Selection

Data Warehouse

Data Cleaning

Data Integration

Databases

- 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
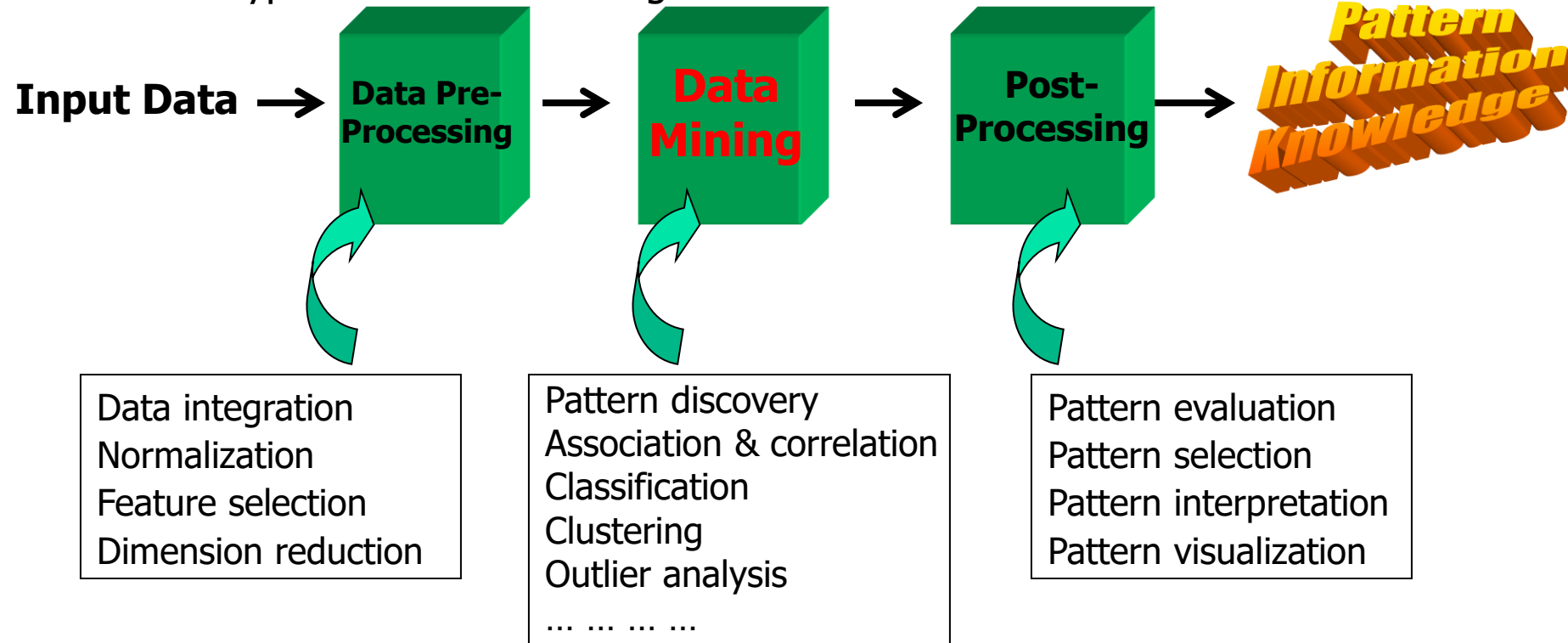
Data Mining Applications

- Two highly successful and popular applications of data mining: *business intelligence* and *search engines*.

- **Business Intelligence (BI)**
  - critical for businesses to acquire a better understanding of the commercial context of their organization, such as their customers, the market, supply and resources, and competitors.
  - BI technologies provide historical, current, and predictive views of business operations.
  - data mining at the core of business intelligence

- **Web Search Engines**
  - specialized computer server that searches for information on the Web.
  - Various data mining techniques are used in all aspects of search engines, ranging from *crawling*, indexing and searching (e.g., ranking , advertising and personalizing results).

Data Mining in Business Intelligence

Ex: Cancer research

Increasing potential to support business decisions

End User

Decision Making

Data Presentation

Visualization Techniques

Business Analyst

Data Mining

Information Discovery

Data Analyst

Data Exploration

Statistical Summary, Querying, and Reporting

Data Preprocessing/Integration, Data Warehouses

Data Sources

DBA

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

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

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

Data integration
Normalization
Feature selection
Dimension reduction

Pattern discovery
Association & correlation
Classification
Clustering
Outlier analysis
… … … …

Pattern evaluation
Pattern selection
Pattern interpretation
Pattern visualization

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.
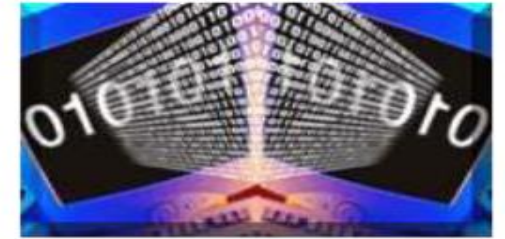
Big Data Applications

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(Web Mining)**

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.

Data Mining Function: (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

- 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?*

- How to mine such patterns and rules efficiently in large datasets

- How to use such patterns for classification, clustering, and other applications
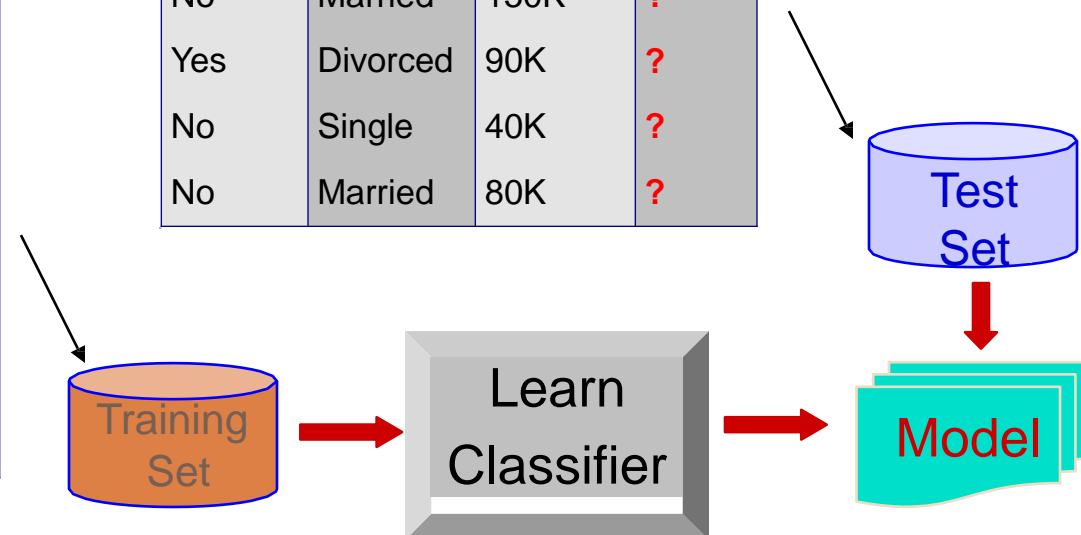
Data Mining Function: (3) Classification

- 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, …

Classification Example

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Refund	Marital Status	Taxable Income	Cheat
No	Single	75K	?
Yes	Married	50K	?
No	Married	150K	?
Yes	Divorced	90K	?
No	Single	40K	?
No	Married	80K	?

Test Set

Training Set → Learn Classifier → Model

www.belgiumcampus.ac.za
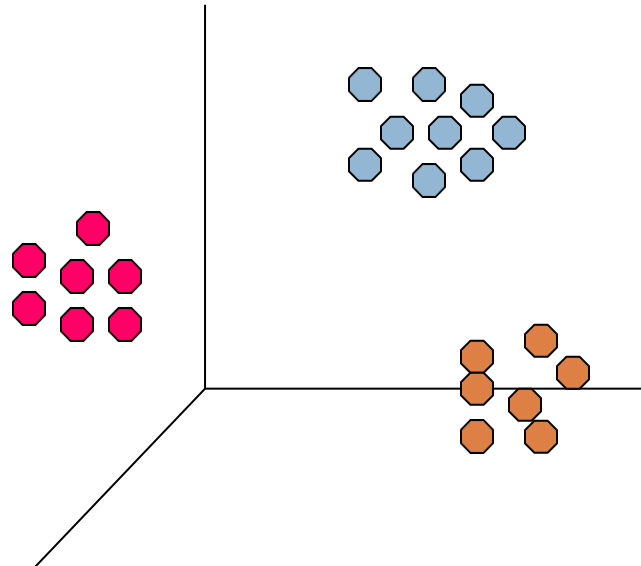
Data Mining Function: (4) Cluster Analysis

- 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

- Euclidean Distance Based Clustering in 3-D space.

Intracluster distances are minimized
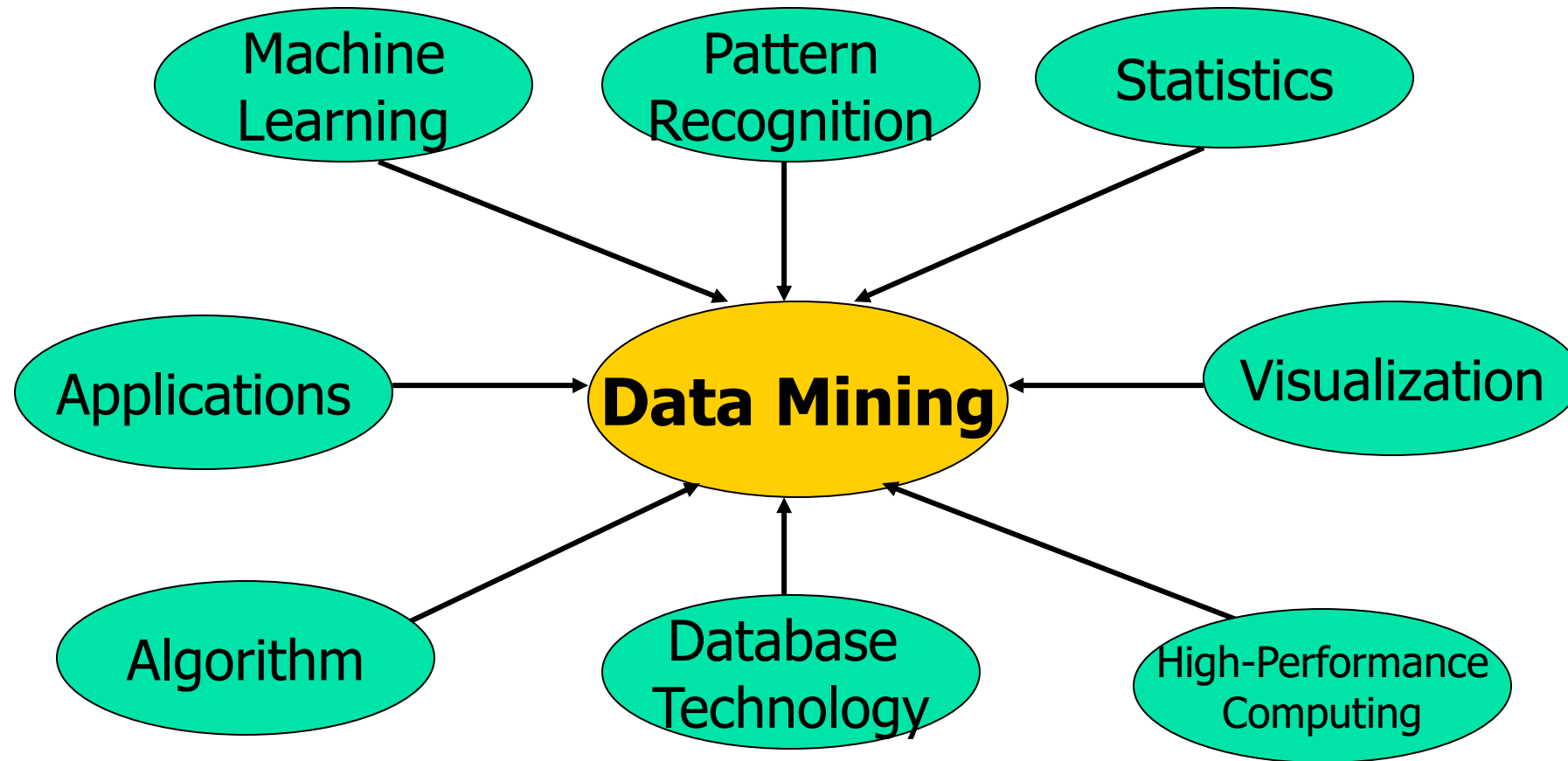
Intercluster distances are maximized

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

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
  - …

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**

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

- **Data analysis and decision support**

  - Market analysis and management

    - Target marketing, customer relationship management (CRM), market basket analysis, cross selling, market segmentation

  - Risk analysis and management

    - Forecasting, customer retention, improved underwriting, quality control, competitive analysis

  - Fraud detection and detection of unusual patterns (outliers)

- **Other Applications**

  - Text mining (news group, email, documents) and Web mining

  - Stream data mining

  - Bioinformatics and bio-data analysis

Ex. 1: Market Analysis and Management

- Where does the data come from?—Credit card transactions, loyalty cards, discount coupons, customer complaint calls, plus (public) lifestyle studies
- Target marketing
  - Find clusters of "model" customers who share the same characteristics: interest, income level, spending habits, etc.
  - Determine customer purchasing patterns over time
- Cross-market analysis—Find associations/co-relations between product sales, & predict based on such association
- Customer profiling—What types of customers buy what products (clustering or classification)
- Customer requirement analysis
  - Identify the best products for different groups of customers
  - Predict what factors will attract new customers
- Provision of summary information
  - Multidimensional summary reports
  - Statistical summary information (data central tendency and variation)

Ex. 2: Corporate Analysis & Risk Management

- Finance planning and asset evaluation

  - cash flow analysis and prediction

  - contingent claim analysis to evaluate assets

  - cross-sectional and time series analysis (financial-ratio, trend analysis, etc.)

- Resource planning

  - summarize and compare the resources and spending

- Competition

  - monitor competitors and market directions

  - group customers into classes and a class-based pricing procedure

  - set pricing strategy in a highly competitive market

- Approaches: Clustering & model construction for frauds, outlier analysis
- Applications: Health care, retail, credit card service, telecomm.
  - <u>Auto insurance</u>: ring of collisions
  - <u>Money laundering:</u> suspicious monetary transactions
  - <u>Medical insurance</u>
    - Professional patients, ring of doctors, and ring of references
    - Unnecessary or correlated screening tests
  - <u>Telecommunications: phone-call fraud</u>
    - Phone call model: destination of the call, duration, time of day or week. Analyze patterns that deviate from an expected norm
  - <u>Retail industry</u>
    - Analysts estimate that 38% of retail shrink is due to dishonest employees
  - <u>Anti-terrorism</u>

Major Issues in Data Mining

- <u>Mining methodology</u>
  - Mining different kinds of knowledge from diverse data types, e.g., bio, stream, Web
  - Performance: efficiency, effectiveness, and scalability
  - Pattern evaluation: the interestingness problem
  - Incorporation of background knowledge
  - Handling noise and incomplete data
  - Parallel, distributed and incremental mining methods
  - Integration of the discovered knowledge with existing one: knowledge fusion
- <u>User interaction</u>
  - Data mining query languages and ad-hoc mining
  - Expression and visualization of data mining results
  - Interactive mining of knowledge at multiple levels of abstraction

Major Issues in Data Mining (2)

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

Where to Find References? DBLP, CiteSeer, Google

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

Recommended Books

Main Book

- **Han, J., Kamber, M. and Pei, J., 2011. Data mining concepts and techniques 3rd edition.** *The Morgan Kaufmann Series in Data Management Systems.*

Reference Books

- Andrea Cirillo (2017) *R Data Mining : Mine Valuable Insights From Your Data Using Popular Tools and Techniques in R.* Birmingham, UK: Packt Publishing. **(available on ebscohost)**

- Zhao, Y., 2012. *R and Data Mining: Examples and Case Studies.* Academic Press.

- Torgo, L., 2011. *Data Mining with R: Learning with Case Studies.* Chapman and Hall/CRC.

- Layton, R., 2017. *Learning Data Mining with Python.* Packt Publishing Ltd.

- Madhavan, S., 2015. *Mastering Python for Data Science.* Packt Publishing Ltd.

- Hastie, T., Tibshirani, R. and Friedman, J., 2009. *The elements of statistical learning: data mining, inference, and prediction.* Springer Science & Business Media.

- Tan, P.N., Steinbach, M. and Kumar, V., 2016. *Introduction to data mining.* Pearson Education India.

- Weiss, S.M. and Indurkhya, N., 1998. *Predictive data mining: a practical guide.* Morgan Kaufmann.

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