Data Mining: Concepts and Techniques

— Chapter 1 —

— Introduction —

Jiawei Han and Micheline Kamber
Department of Computer Science
University of Illinois at Urbana-Champaign

Course Structures at SCE

- Intro. to data mining
- Data mining: Principles and algorithms
- Project (40%)
- Final exam (50%)
- Participation in lessons (10%)

Chapter 1. Introduction

- Motivation: Why data mining?
- What is data mining?
- Data Mining: On what kind of data?
- Data mining functionality
- Top-10 most popular data mining algorithms
- Major issues in data mining

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, 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

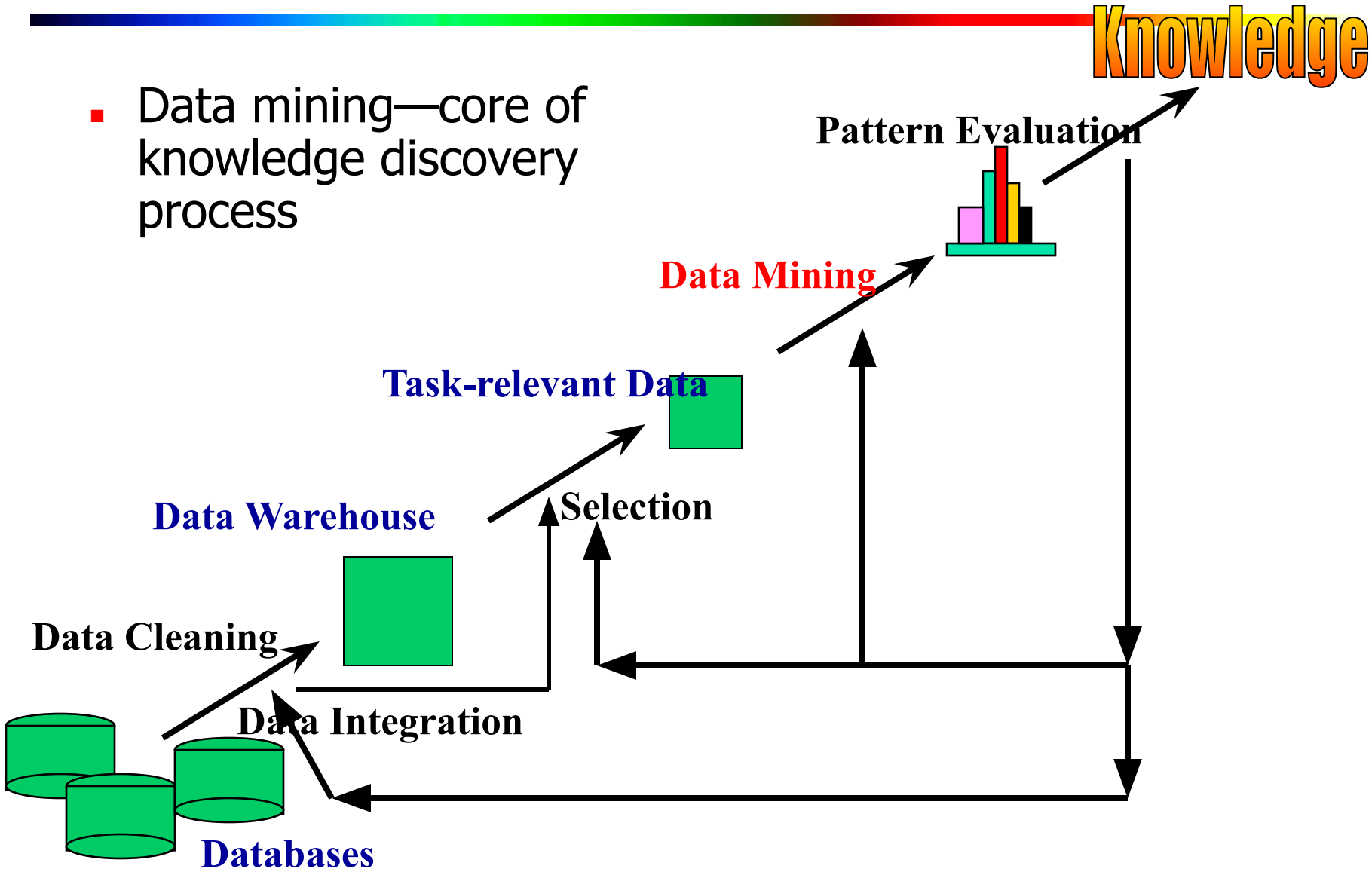
- 1960s:
  - Data collection, database creation, IMS and network DBMS
- 1970s:
  - Relational data model, relational DBMS implementation
- 1980s:
  - RDBMS, advanced data models (extended-relational, OO, deductive, etc.)
  - Application-oriented DBMS (spatial, scientific, engineering, etc.)
- 1990s:
  - 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
  - Big/Small Data
  - Business Intelligence
  - Schema Matching, etc.

What Is Data Mining?

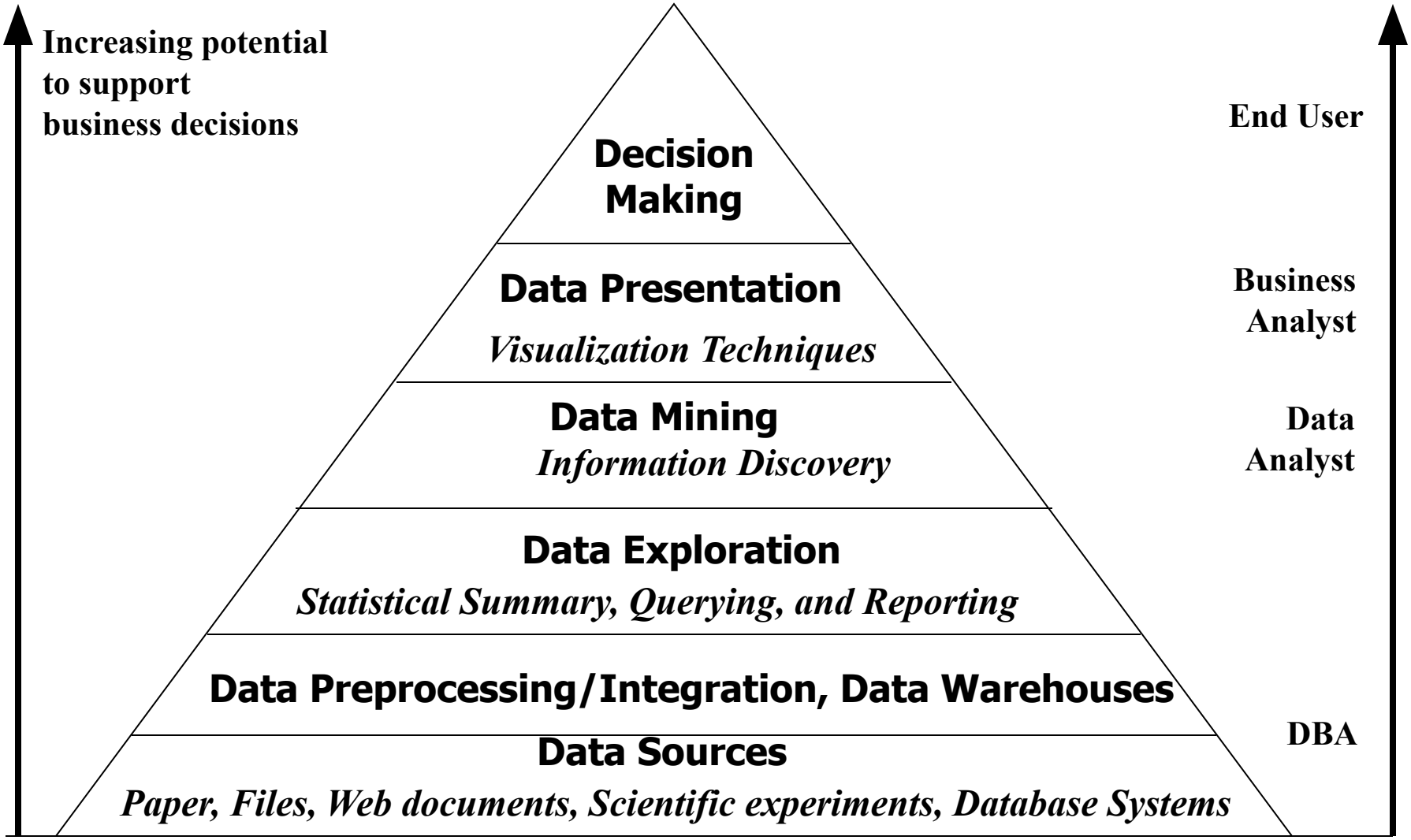
- Data mining (knowledge discovery from data)
  - Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data
  - Data mining: a misnomer?
- Alternative names
  - Knowledge discovery (mining) in databases (KDD), knowledge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

Knowledge Discovery (KDD) Process

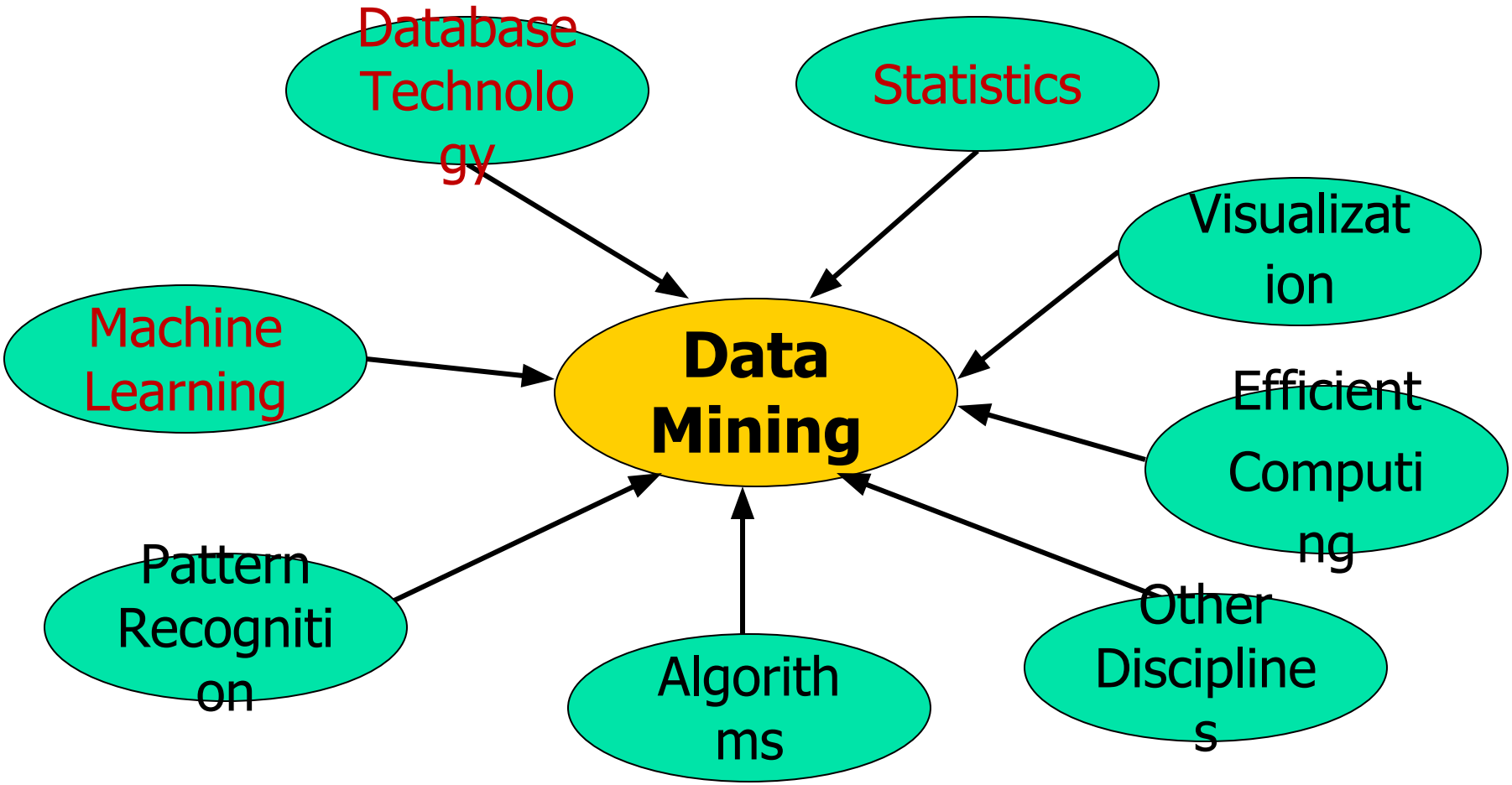
- Data mining—core of knowledge discovery process



Data Mining and Business Intelligence



Data Mining: Confluence of Multiple Disciplines



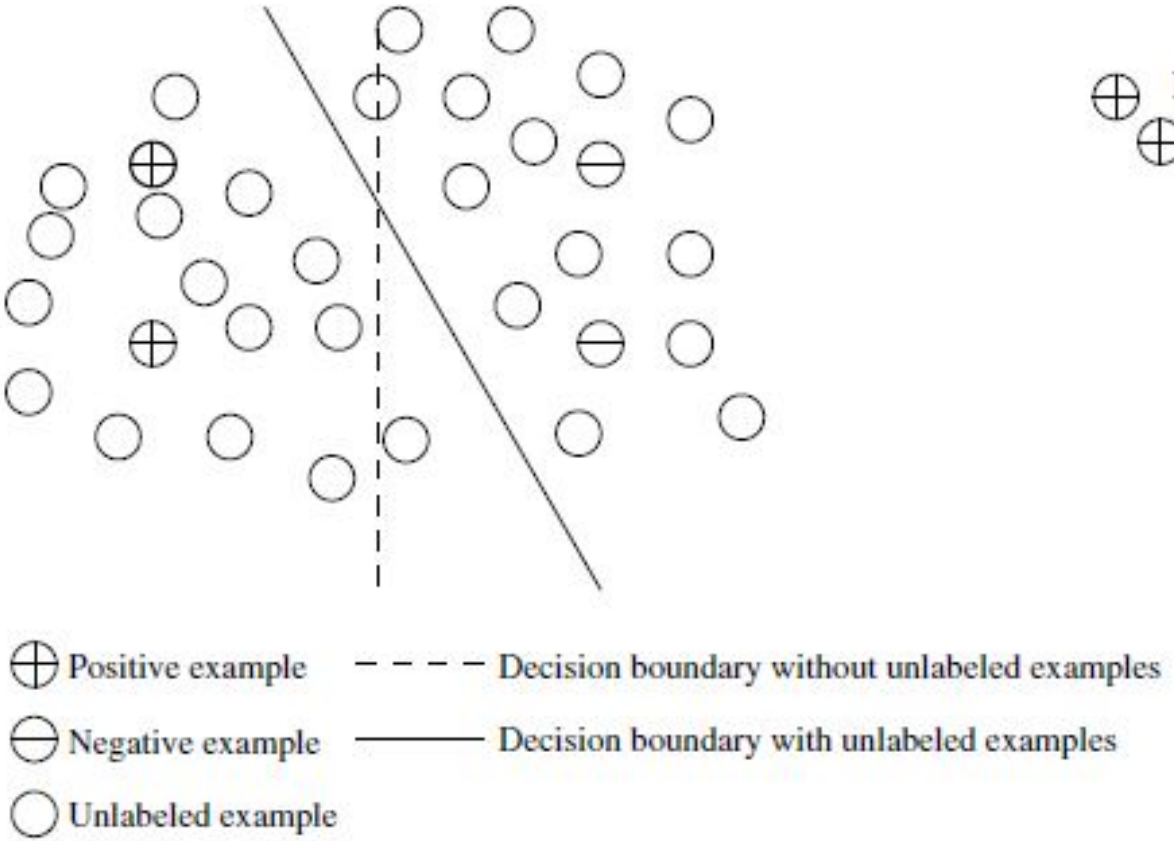
Statistics

- A **statistical model** is a set of mathematical functions that describe the behavior
  - of the objects in a target class
  - in terms of random variables and their associated probability distributions.
- Statistical models are widely used to model data
  - mining various patterns from data
  - understanding the underlying mechanisms generating and affecting the patterns.
- Also is used to verify data mining results (**statistical hypothesis test)**

Machine Learning

- **Machine learning** investigates how computers can learn based on data.
  - **Supervised learning** (classification) – labeled data
  - **Unsupervised learning** (clustering) – unlabeled data
  - **Semi-supervised learning** – both
  - **Active learning** – interactive (with user)

Semi-supervised learning



Semi-supervised learning.

Database systems

- **DB systems research** focuses on the creation, maintenance, and use of databases.
  - query languages
  - query processing
  - optimization methods
  - data storage
  - indexing and accessing methods

- Goal - high scalability in processing very large, relatively structured data sets.

Challenges of DM

- 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 (news, sensor data)
  - Time-series data, temporal data, sequence data
  - Structure data, graphs, social networks and multi-linked data
  - Heterogeneous databases and legacy databases
  - Multimedia, text and Web data
  - Software programs, scientific simulations
- New and **sophisticated applications**

Multi-Dimensional View of Data Mining

- **Data to be mined**
  - Relational, data warehouse, transactional, stream, object-oriented/relational, active, spatial, time-series, text, multi-media, heterogeneous, legacy, WWW
- **Knowledge to be mined**
  - Characterization, association, classification, clustering, trend/deviation, outlier analysis, etc.
  - Multiple/integrated functions and mining at multiple levels
- **Techniques utilized**
  - Database-oriented, data warehouse (OLAP), machine learning, statistics, visualization, etc.
- **Applications adapted**
  - Telecommunication, banking, fraud/cheating analysis, bio-data mining, stock market analysis, text mining, Web mining, etc.

Data Mining: Classification Schemes

- General functionality
  - Descriptive data mining
  - Predictive data mining
- Different views lead to different classifications
  - Data view: Kinds of data to be mined
  - Knowledge view: Kinds of knowledge to be discovered
  - Method view: Kinds of techniques utilized
  - Application view: Kinds of applications adapted

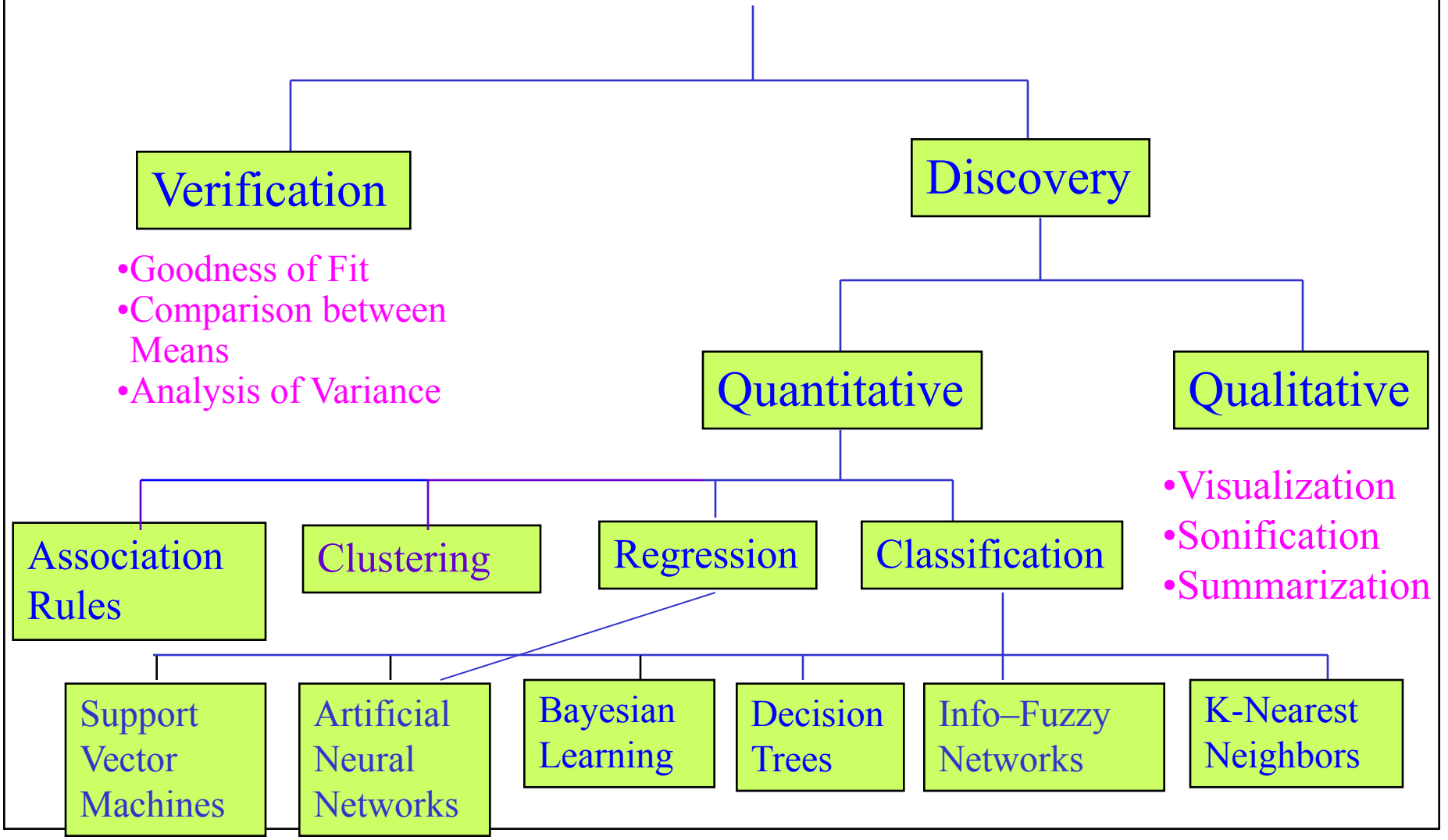
Data Mining: On What Kinds of Data?

- 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
  - Multimedia database
  - Text databases
  - The World-Wide Web

Common Data Mining Tasks

- Classification
  - Credit approval
  - Product recommendation
  - Medical diagnosis
- Prediction / Regression
  - Stock price forecasting
- Clustering
  - Customer segmentation
  - Organization of search results
- Association Rules
  - Retail analysis
- Sequencing Rules
  - Preventive maintenance
  - Analysis of patient records
- Outlier analysis
  - Anomaly detection
- Feature Selection
  - Quality Assurance

Data Mining Methods



What is not "Data Mining"?

- Database Management Systems (DBMS)
- Data Warehouses (DWH)
- Simple search and query processing
- Expert (Rule-based) Systems
- Statistical Hypothesis Testing (e.g., *t-test)*

Data Mining Functionalities

- Multidimensional **concept description**: Characterization and discrimination
  - Generalize, summarize, and contrast data characteristics, e.g., dry vs. wet regions (charts)
- **Frequent patterns**, association, correlation vs. causality
  - Diaper → Beer [0.5%, 75%] (Correlation or causality?)
- **Classification** and **prediction**
  - Construct models (functions) that 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 or missing numerical values

Correlation or causality?

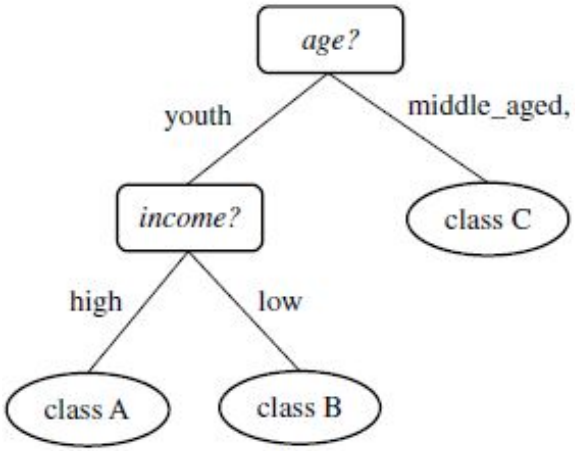
- Assoc. rules:

$$age(X, \text{“20..29”}) \wedge income(X, \text{“40K..49K”}) \Rightarrow buys(X, \text{“laptop”})$$
$$[support = 2\%, confidence = 60\%].$$

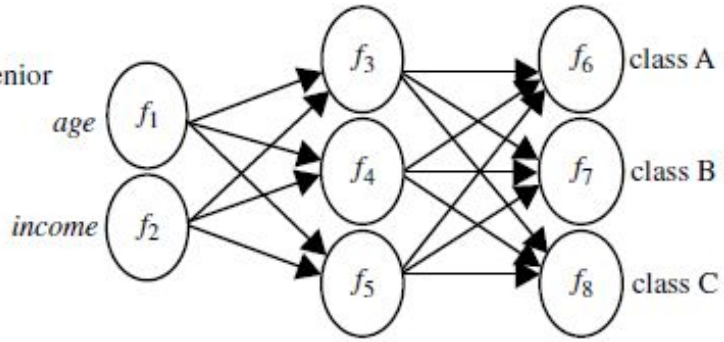
- Classification:

age(*X*, “youth”) *AND income*(*X*, “high”) ⟶ *class*(*X*, “A”)
age(*X*, “youth”) *AND income*(*X*, “low”) ⟶ *class*(*X*, “B”)
age(*X*, “middle_aged”) ⟶ *class*(*X*, “C”)
age(*X*, “senior”) ⟶ *class*(*X*, “C”)

(a)



(b)



(c)

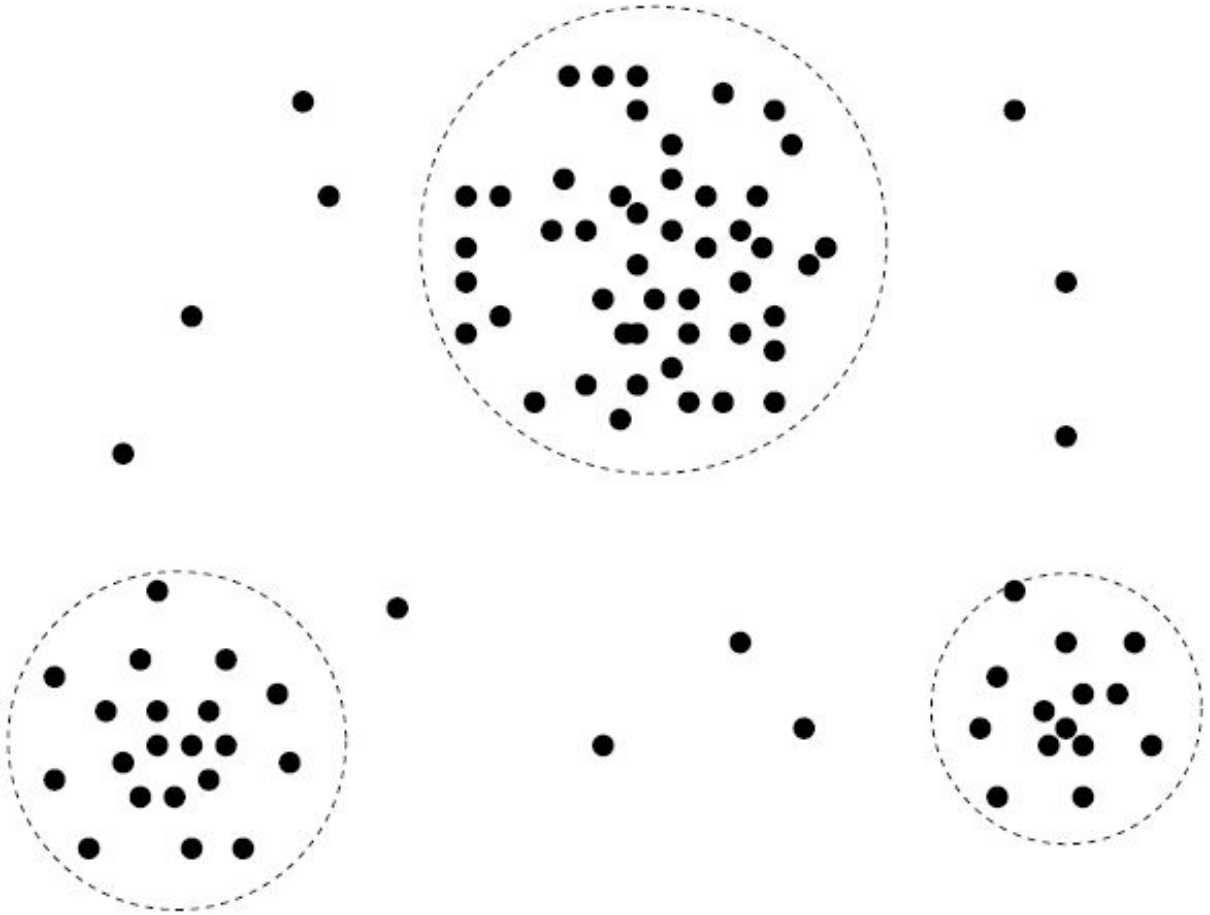
A classification model can be represented in various forms: (a) IF-THEN rules, (b) a decision tree, or (c) a neural network.

*

Data Mining Functionalities (2)

- **Cluster analysis**
  - Class label is unknown: Group data to form new classes, e.g., cluster houses to find distribution patterns
  - Maximizing intra-class similarity & minimizing interclass similarity
- **Outlier analysis**
  - Outlier: Data object that does not comply with the general behavior of the data
  - Noise or exception? Useful in *fraud detection*, rare events analysis
- **Trend** and evolution **analysis**
  - Trend and deviation: e.g., regression analysis
  - Sequential pattern mining: e.g., digital camera → large SD memory
  - Periodicity analysis
  - Similarity-based analysis
- Other **pattern**-directed or **statistical analyses**

Clustering



A 2-D plot of customer data with respect to customer locations in a city, showing three data clusters.

Top-10 Most Popular DM Algorithms: 18 Identified Candidates (I)

- **Classification**
  - #1. **C4.5**: Quinlan, J. R. C4.5: Programs for Machine Learning. Morgan Kaufmann., 1993.
  - #2. **CART**: L. Breiman, J. Friedman, R. Olshen, and C. Stone. Classification and Regression Trees. Wadsworth, 1984.
  - #3. K Nearest Neighbours (**kNN**): Hastie, T. and Tibshirani, R. 1996. Discriminant Adaptive Nearest Neighbor Classification. TPAMI. 18(6)
  - #4. **Naive Bayes** Hand, D.J., Yu, K., 2001. Idiot's Bayes: Not So Stupid After All? Internat. Statist. Rev. 69, 385-398.
- **Statistical Learning**
  - #5. **SVM**: Vapnik, V. N. 1995. The Nature of Statistical Learning Theory. Springer-Verlag.
  - #6. **EM**: McLachlan, G. and Peel, D. (2000). Finite Mixture Models. J. Wiley, New York. Association Analysis
  - #7. **Apriori**: Rakesh Agrawal and Ramakrishnan Srikant. Fast Algorithms for Mining Association Rules. In VLDB '94.
  - #8. **FP-Tree**: Han, J., Pei, J., and Yin, Y. 2000. Mining frequent patterns without candidate generation. In SIGMOD '00.

The 18 Identified Candidates (II)

- **Link Mining**
  - #9. **PageRank**: Brin, S. and Page, L. 1998. The anatomy of a large-scale hypertextual Web search engine. In WWW-7, 1998.
  - #10. **HITS**: Kleinberg, J. M. 1998. Authoritative sources in a hyperlinked environment. SODA, 1998.
- **Clustering**
  - #11. **K-Means**: MacQueen, J. B., Some methods for classification and analysis of multivariate observations, in Proc. 5th Berkeley Symp. Mathematical Statistics and Probability, 1967.
  - #12. **BIRCH**: Zhang, T., Ramakrishnan, R., and Livny, M. 1996. BIRCH: an efficient data clustering method for very large databases. In SIGMOD '96.
- **Bagging** and **Boosting**
  - #13. **AdaBoost**: Freund, Y. and Schapire, R. E. 1997. A decision-theoretic generalization of on-line learning and an application to boosting. J. Comput. Syst. Sci. 55, 1 (Aug. 1997), 119-139.

The 18 Identified Candidates (III)

- **Sequential Patterns**
  - #14. GSP: Srikant, R. and Agrawal, R. 1996. **Mining Sequential Patterns**: Generalizations and Performance Improvements. In Proceedings of the 5th International Conference on Extending Database Technology, 1996.
  - #15. PrefixSpan: J. Pei, J. Han, B. Mortazavi-Asl, H. Pinto, Q. Chen, U. Dayal and M-C. Hsu. PrefixSpan: Mining Sequential Patterns Efficiently by Prefix-Projected Pattern Growth. In ICDE '01.
- **Integrated Mining**
  - #16. **CBA**: Liu, B., Hsu, W. and Ma, Y. M. Integrating classification and association rule mining. KDD-98.
- **Rough Sets**
  - #17. **Finding reduct**: Zdzislaw Pawlak, Rough Sets: Theoretical Aspects of Reasoning about Data, Kluwer Academic Publishers, Norwell, MA, 1992
- **Graph Mining**
  - #18. **gSpan**: Yan, X. and Han, J. 2002. gSpan: Graph-Based Substructure Pattern Mining. In ICDM '02.

Top-10 Algorithm Finally Selected at ICDM'06

- **#1: C4.5 (61 votes)**
- **#2: K-Means (60 votes)**
- **#3: SVM (58 votes)**
- **#4: Apriori (52 votes)**
- **#5: EM (48 votes)**
- **#6: PageRank (46 votes)**
- **#7: AdaBoost (45 votes)**
- **#7: kNN (45 votes)**
- **#7: Naive Bayes (45 votes)**
- **#10: CART (34 votes)**

Major Issues in Data Mining

- Mining methodology
  - 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**
- User interaction
  - Data mining **query languages** and **ad-hoc mining**
  - Expression and **visualization of** data mining **results**
  - **Interactive mining of knowledge** at multiple levels of abstraction
- Applications and social impacts
  - **Domain-specific data mining**
  - Protection of **data security**, integrity, and **privacy**

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**)
  - Conf. on Principles and practices of Knowledge Discovery and Data Mining (**PKDD**)
  - Pacific-Asia Conf. on Knowledge Discovery and Data Mining (PAKDD)
- Other related conferences
  - ACM SIGMOD
  - VLDB
  - (IEEE) ICDE
  - WWW, **SIGIR**
  - **ICML**, CVPR, NIPS
- Journals
  - Data Mining and Knowledge Discovery (DAMI or DMKD)
  - IEEE Trans. On Knowledge and Data Eng. (TKDE)
  - KDD Explorations
  - ACM Trans. on KDD

Recommended Reference Books

- S. Chakrabarti. Mining the Web: Statistical Analysis of Hypertex and Semi-Structured Data. Morgan Kaufmann, 2002
- R. O. Duda, P. E. Hart, and D. G. Stork, Pattern Classification, 2ed., Wiley-Interscience, 2000
- T. Dasu and T. Johnson. Exploratory Data Mining and Data Cleaning. John Wiley & Sons, 2003
- U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy. Advances in Knowledge Discovery and Data Mining. AAAI/MIT Press, 1996
- U. Fayyad, G. Grinstein, and A. Wierse, Information Visualization in Data Mining and Knowledge Discovery, Morgan Kaufmann, 2001
- J. Han and M. Kamber. Data Mining: Concepts and Techniques. Morgan Kaufmann, 2nd ed., 2006
- D. J. Hand, H. Mannila, and P. Smyth, Principles of Data Mining, MIT Press, 2001
- T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Springer-Verlag, 2001
- B. Liu, Web Data Mining, Springer 2006.
- T. M. Mitchell, Machine Learning, McGraw Hill, 1997
- G. Piatetsky-Shapiro and W. J. Frawley. Knowledge Discovery in Databases. AAAI/MIT Press, 1991
- P.-N. Tan, M. Steinbach and V. Kumar, Introduction to Data Mining, Wiley, 2005
- S. M. Weiss and N. Indurkhya, Predictive Data Mining, Morgan Kaufmann, 1998
- I. H. Witten and E. Frank, Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations, Morgan Kaufmann, 2nd ed. 2005

Summary

- **Data mining**: Discovering interesting patterns from large amounts 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 information repositories
- Data mining functionalities: *characterization*, *discrimination*, *association*, *classification*, *clustering*, *outlier* and *trend analysis*, etc.

Why Data Mining?—Potential Applications

- 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)

KDD Process: Several Key Steps

- Learning the application domain
  - relevant prior knowledge and goals of application
- Creating a target data set: data selection
- Data cleaning and preprocessing: (may take 60% of effort!)
- Data reduction and transformation
  - Find useful features, dimensionality/variable reduction, invariant representation
- Choosing functions of data mining
  - summarization, classification, regression, association, clustering
- Choosing the mining algorithm(s)
- Data mining: search for patterns of interest
- Pattern evaluation and knowledge presentation
  - visualization, transformation, removing redundant patterns, etc.
- Use of discovered knowledge

Are All the "Discovered" Patterns Interesting?

- Data mining may generate thousands of patterns: Not all of them are interesting
  - Suggested approach: Human-centered, query-based, focused mining
- **Interestingness measures**
  - A pattern is interesting if it is easily understood by humans, valid on new or test data with some degree of certainty, potentially useful, novel, or validates some hypothesis that a user seeks to confirm
- **Objective vs. subjective interestingness measures**
  - Objective: based on statistics and structures of patterns, e.g., support, confidence, etc.
  - Subjective: based on user's belief in the data, e.g., unexpectedness, novelty, actionability, etc.

Find All and Only Interesting Patterns?

- Find all the interesting patterns: Completeness
  - Can a data mining system find all the interesting patterns? Do we need to find all of the interesting patterns?
  - Heuristic vs. exhaustive search
  - Association vs. classification vs. clustering
- Search for only interesting patterns: An optimization problem
  - Can a data mining system find only the interesting patterns?
  - Approaches
    - First general all the patterns and then filter out the uninteresting ones
    - Generate only the interesting patterns—mining query optimization

Why Data Mining Query Language?

- Automated vs. query-driven?
  - Finding all the patterns autonomously in a database?—unrealistic because the patterns could be too many but uninteresting
- Data mining should be an interactive process
  - User directs what to be mined
- Users must be provided with a set of primitives to be used to communicate with the data mining system
- Incorporating these primitives in a data mining query language
  - More flexible user interaction
  - Foundation for design of graphical user interface
  - Standardization of data mining industry and practice

Primitives that Define a Data Mining Task

- Task-relevant data
  - Database or data warehouse name
  - Database tables or data warehouse cubes
  - Condition for data selection
  - Relevant attributes or dimensions
  - Data grouping criteria
- Type of knowledge to be mined
  - Characterization, discrimination, association, classification, prediction, clustering, outlier analysis, other data mining tasks
- Background knowledge
- Pattern interestingness measurements
- Visualization/presentation of discovered patterns

Primitive 3: Background Knowledge

- A typical kind of background knowledge: Concept hierarchies
- Schema hierarchy
  - E.g., street < city < province_or_state < country
- Set-grouping hierarchy
  - E.g., {20-39} = young, {40-59} = middle_aged
- Operation-derived hierarchy
  - email address: hagonzal@cs.uiuc.edu

    login-name < department < university < country

Primitive 4: Pattern Interestingness Measure

- Simplicity

  e.g., (association) rule length, (decision) tree size

- Certainty

  e.g., confidence, P(A|B) = #(A and B)/ #(B), classification reliability or accuracy, certainty factor, rule strength, rule quality, discriminating weight, etc.

- Utility

  potential usefulness, e.g., support (association), noise threshold (description)

- Novelty

  not previously known, surprising (used to remove redundant rules, e.g., Illinois vs. Champaign rule implication support ratio)

Primitive 5: Presentation of Discovered Patterns

- Different backgrounds/usages may require different forms of representation
  - E.g., rules, tables, crosstabs, pie/bar chart, etc.
- Concept hierarchy is also important
  - Discovered knowledge might be more understandable when represented at high level of abstraction
  - Interactive drill up/down, pivoting, slicing and dicing provide different perspectives to data
- Different kinds of knowledge require different representation: association, classification, clustering, etc.

Architecture: Typical Data Mining System