



Data Mining: Concepts and Techniques

Prof.Dr. Songül Varlı
Department of Computer Engineering
Yildiz Technical University
svarli@yildiz.edu.tr



Introduction

- Motivation: Why data mining?
- What is data mining?
- Data Mining: On what kind of data?
- Data mining functionality
- Are all the patterns interesting?
- Classification of data mining systems
- Data Mining Task Primitives
- Integration of data mining system with a DB and DW System
- 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, digital cameras,
- We are drowning in data, but starving for knowledge!
- “Necessity is the mother of invention”—Data mining—Automated analysis of massive data sets



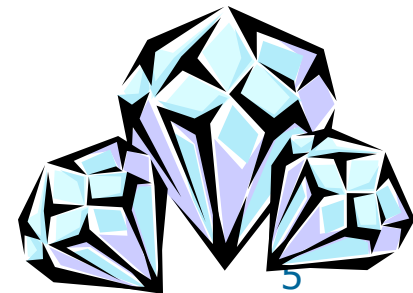
Evolution of Database Technology

- 1960s:
 - Data collection, database creation, IMS and network DBMS
- 1970s:
 - Relational data model, relational DBMS implementation
- 1980s:
 - RDBMS, advanced data models (extended-relational, 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



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 customers
 - Predict what factors will attract new customers
- Provision of summary information
 - Multidimensional summary reports
 - Statistical summary information (data central tendency and variation)

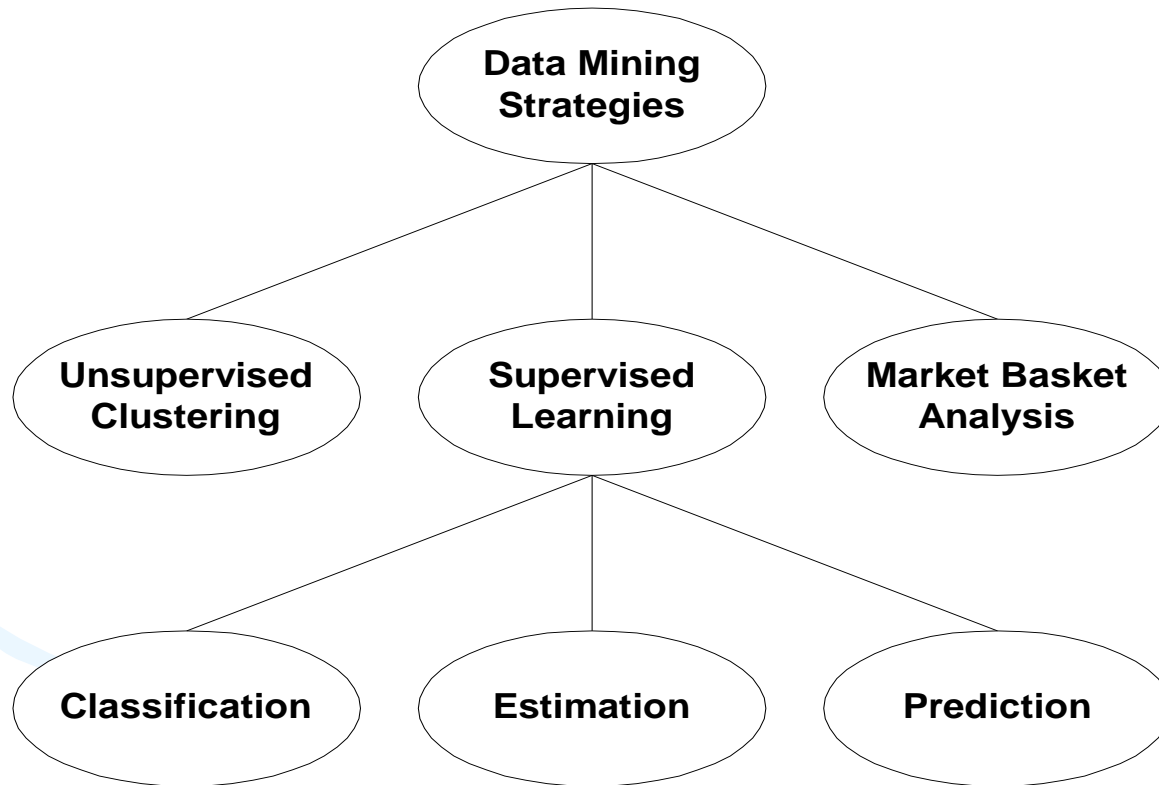
Ex. 2: Corporate Analysis & Risk Management

- 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

Ex. 3: Fraud Detection & Mining Unusual Patterns

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

A hierarchy of data mining strategies





Supervised Learning

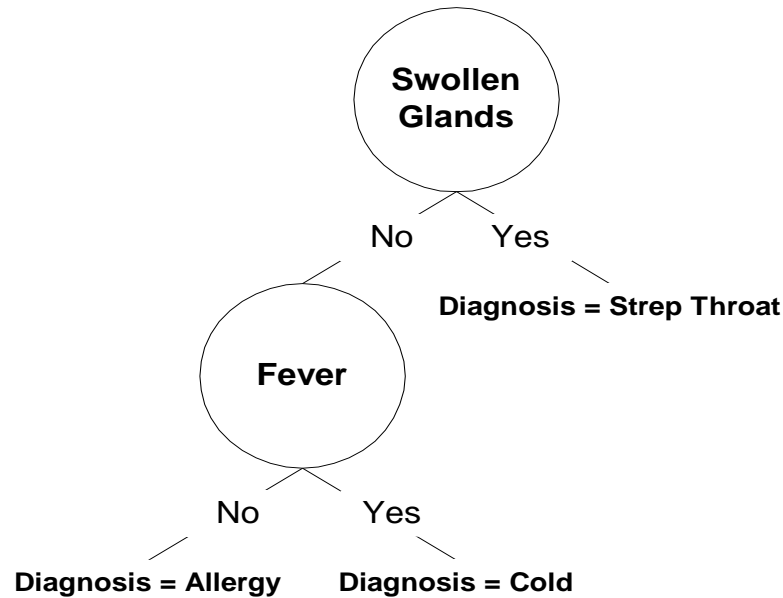
- Build a learner model using data instances of known origin.
- Use the model to determine the outcome new instances of unknown origin.

Supervised Learning: A Decision tree Example

Table 1.1 • Hypothetical Training Data for Disease Diagnosis

Patient ID#	Sore Throat	Fever	Swollen Glands	Congestion	Headache	Diagnosis
1	Yes	Yes	Yes	Yes	Yes	Strep throat
2	No	No	No	Yes	Yes	Allergy
3	Yes	Yes	No	Yes	No	Cold
4	Yes	No	Yes	No	No	Strep throat
5	No	Yes	No	Yes	No	Cold
6	No	No	No	Yes	No	Allergy
7	No	No	Yes	No	No	Strep throat
8	Yes	No	No	Yes	Yes	Allergy
9	No	Yes	No	Yes	Yes	Cold
10	Yes	Yes	No	Yes	Yes	Cold

Supervised Learning: A Decision tree Example



A tree structure where non-terminal nodes represent tests on one or more attributes and terminal nodes reflect decision outcomes.

Decision Tree:

Table 1.2 • **Data Instances with an Unknown Classification**

Patient ID#	Sore Throat	Fever	Swollen Glands	Congestion	Headache	Diagnosis
11	No	No	Yes	Yes	Yes	?
12	Yes	Yes	No	No	Yes	?
13	No	No	No	No	Yes	?



Unsupervised Clustering

- A data mining method that builds models from data without predefined classes.

The Acme Investors Dataset

Table 1.3 • Acme Investors Incorporated

Customer ID	Account Type	Margin Account	Transaction Method	Trades/ Month	Sex	Age	Favorite Recreation	Annual Income
1005	Joint	No	Online	12.5	F	30–39	Tennis	40–59K
1013	Custodial	No	Broker	0.5	F	50–59	Skiing	80–99K
1245	Joint	No	Online	3.6	M	20–29	Golf	20–39K
2110	Individual	Yes	Broker	22.3	M	30–39	Fishing	40–59K
1001	Individual	Yes	Online	5.0	M	40–49	Golf	60–79K

The Acme Investors Dataset & Supervised Learning

- Can I develop a general profile of an online investor?
- Can I determine if a new customer is likely to open a margin account?
- Can I build a model predict the average number of trades per month for a new investor?
- What characteristics differentiate female and male investors?

The Acme Investors Dataset & Unsupervised Clustering

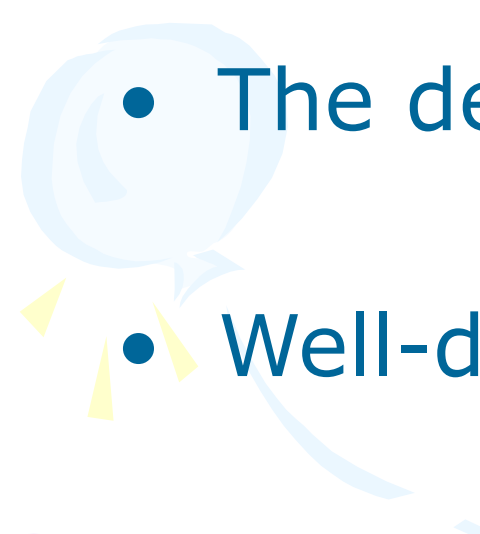

1. What attribute similarities group customers of Acme Investors together?
2. What differences in attribute values segment the customer database?

Data Mining Strategies: Classification

- Learning is supervised.
- The dependent variable is categorical.
- Well-defined classes.
- Current rather than future behavior.



Data Mining Strategies: Estimation

- Learning is supervised.
 - The dependent variable is numeric.
 - Well-defined classes.
 - Current rather than future behavior.
- 
- 

Data Mining Strategies: Prediction

- The emphasis is on predicting future rather than current outcomes.
- The output attribute may be categorical or numeric.

Data Mining Strategies: Market Basket Analysis

A decorative background featuring several balloons in light green, light blue, and light purple colors, along with yellow streamers and triangular flags. A horizontal rainbow-colored bar is positioned below the title.

- Find interesting relationships among retail products.
- Uses association rule algorithms.



Data Mining Strategies: Unsupervised Clustering

- Unsupervised Clustering can be used to:
 - determine if relationships can be found in the data.
 - evaluate the likely performance of a supervised model.
 - find a best set of input attributes for supervised learning.
 - detect Outliers.

The Credit Card Promotion Database

Table 2.3 • The Credit Card Promotion Database

Income Range (\$)	Magazine Promotion	Watch Promotion	Life Insurance Promotion	Credit Card Insurance	Sex	Age
40–50K	Yes	No	No	No	Male	45
30–40K	Yes	Yes	Yes	No	Female	40
40–50K	No	No	No	No	Male	42
30–40K	Yes	Yes	Yes	Yes	Male	43
50–60K	Yes	No	Yes	No	Female	38
20–30K	No	No	No	No	Female	55
30–40K	Yes	No	Yes	Yes	Male	35
20–30K	No	Yes	No	No	Male	27
30–40K	Yes	No	No	No	Male	43
30–40K	Yes	Yes	Yes	No	Female	41
40–50K	No	Yes	Yes	No	Female	43
20–30K	No	Yes	Yes	No	Male	29
50–60K	Yes	Yes	Yes	No	Female	39
40–50K	No	Yes	No	No	Male	55
20–30K	No	No	Yes	Yes	Female	19

A Hypothesis for the Credit Card Promotion Database

- A combination of one or more of the dataset attributes differentiate Acme Credit Card Company card holders who have taken advantage of the life insurance promotion and those card holders who have chosen not to participate in the promotional offer.

Supervised Data Mining Techniques: Production Rules

- IF *Sex = Female & 19 <= Age <= 43*
THEN *Life Insurance Promotion = Yes*

Rule Accuracy: 100.00%

Rule Coverage: 66.67%

- Rule accuracy is a between-class measure.
- Rule coverage is a within-class measure.

Supervised Data Mining Techniques: Production Rules

- IF *Sex = Male & Income Range=40-50*
THEN *Life Insurance Promotion = No*

Rule Accuracy: ? %

Rule Coverage: ? %

Supervised Data Mining Techniques: Production Rules

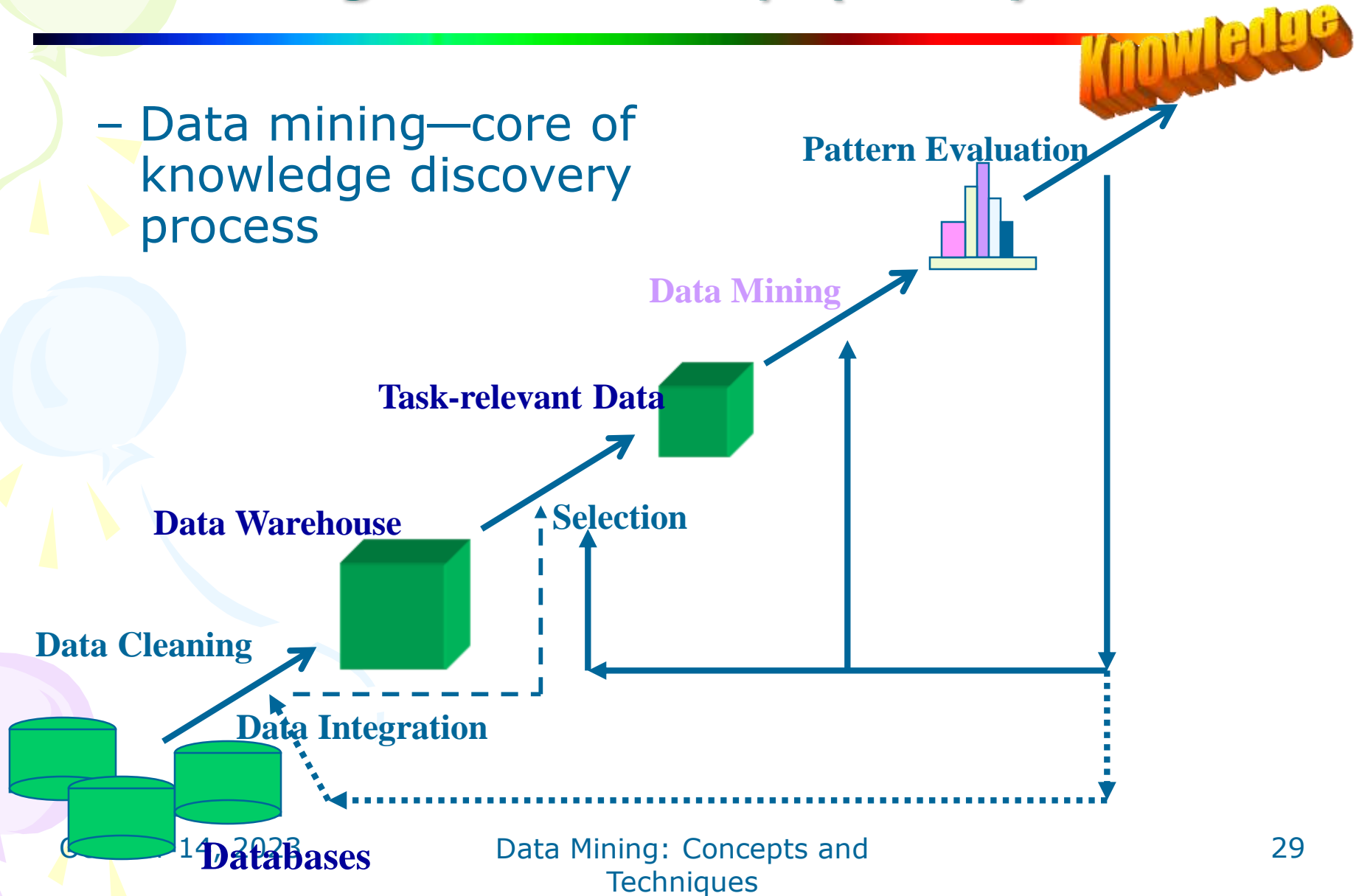
- *IF Sex = Male & Income Range=40-50
THEN Life Insurance Promotion = No*

Rule Accuracy: 100 %

Rule Coverage: 50 %

Knowledge Discovery (KDD) Process

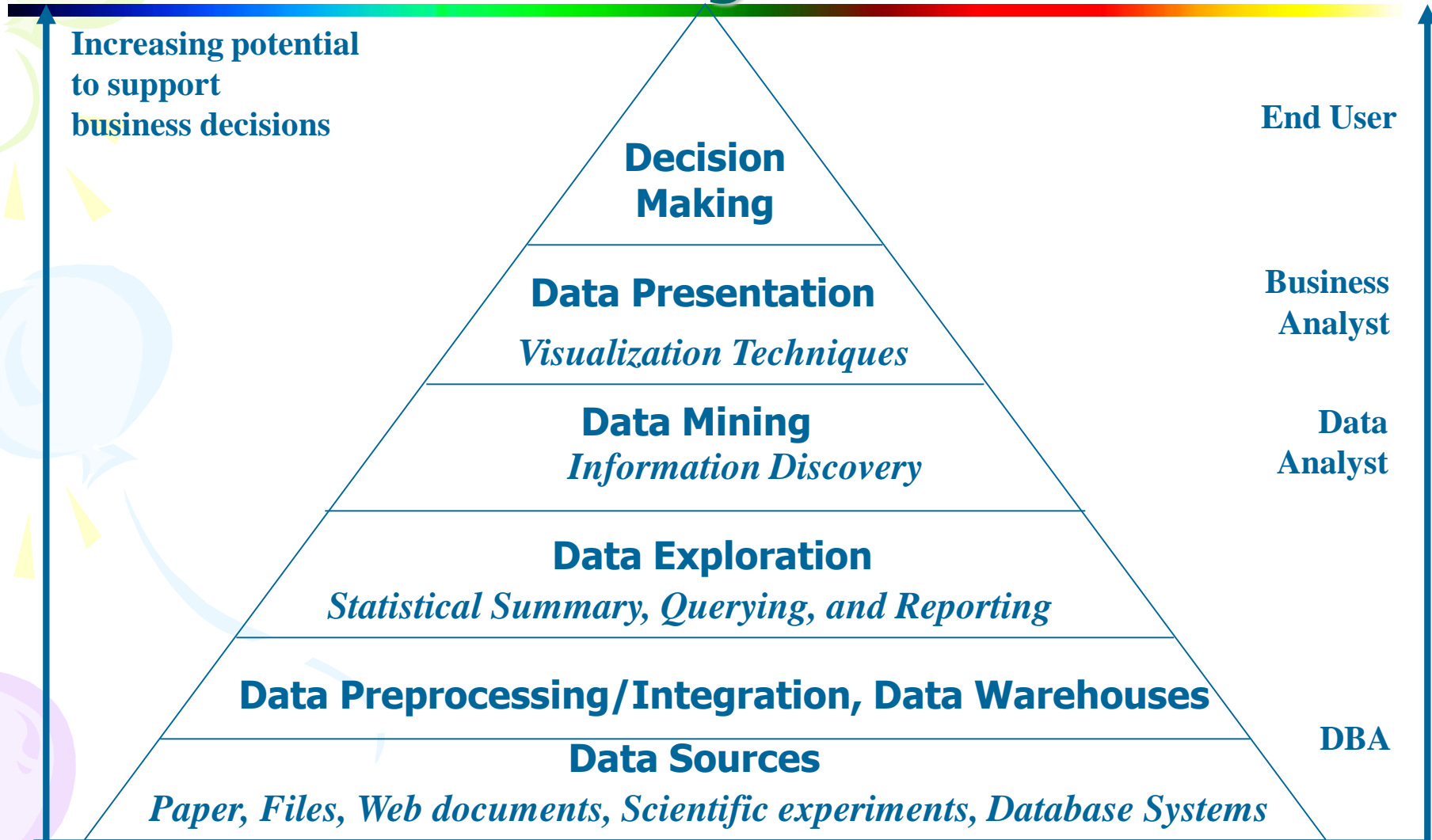
- Data mining—core of knowledge discovery process



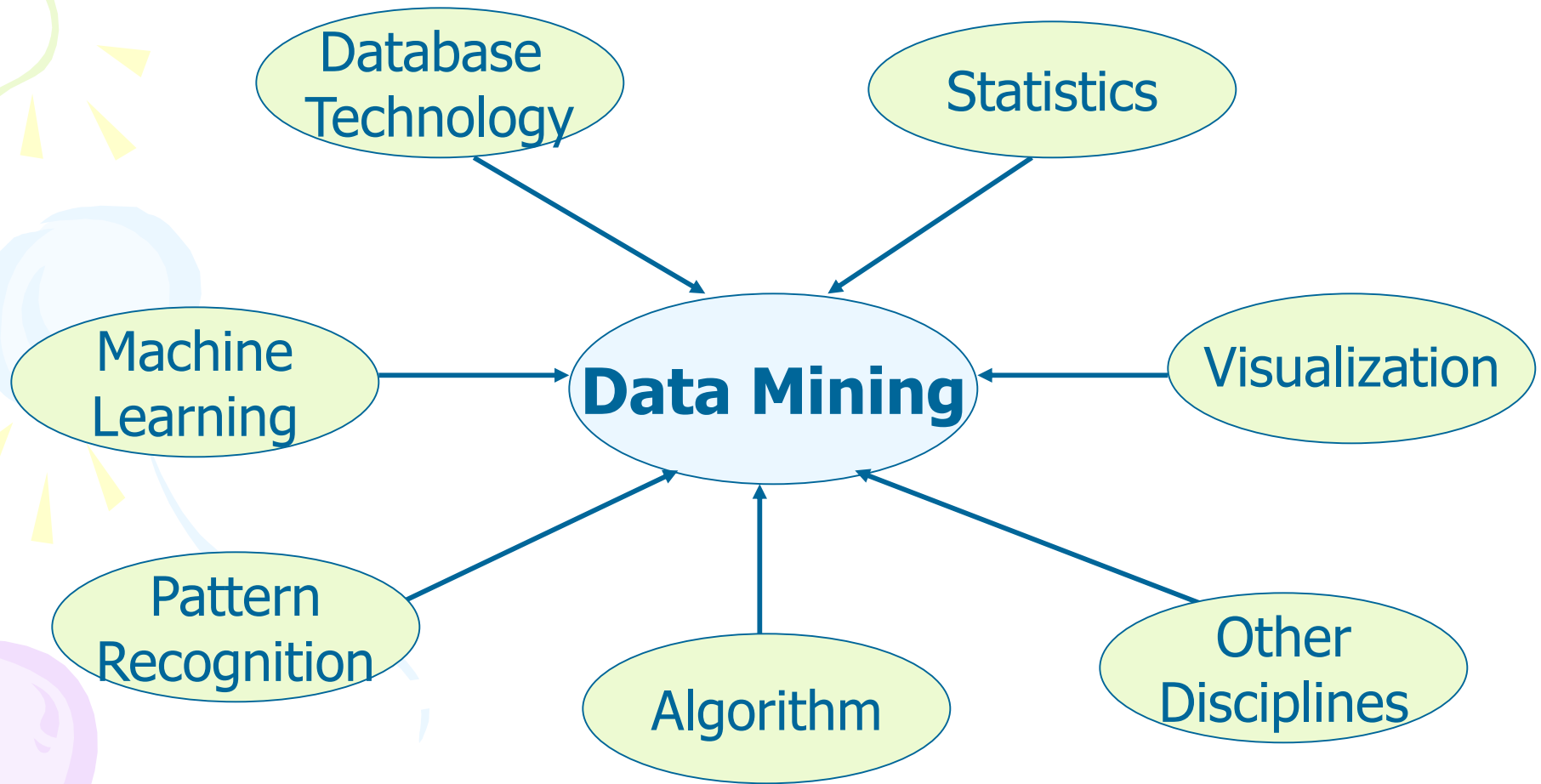
KDD Process: Several Key Steps

- 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

Data Mining and Business Intelligence



Data Mining: Confluence of Multiple Disciplines



Why Not Traditional Data Analysis?

- 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

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, discrimination, 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**
 - Retail, telecommunication, banking, fraud 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 Functionalities

- Multidimensional concept description: Characterization and discrimination
 - Generalize, summarize, and contrast data characteristics, e.g., dry vs. wet regions
- 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

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

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 generate 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
- Type of knowledge to be mined
- Background knowledge
- Pattern interestingness measurements
- Visualization/presentation of discovered patterns

Primitive 1: 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

Primitive 2: Types of Knowledge to Be Mined

- Characterization
- Discrimination
- Association
- Classification/prediction
- Clustering
- Outlier analysis
- Other data mining tasks

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
- Rule-based hierarchy
 - low_profit_margin (X) <= price(X, P₁) and cost (X, P₂) and (P₁ - P₂) < \$50

Primitive 4: Pattern Interestingness Measure

- Simplicity

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

- Certainty

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

- Utility

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

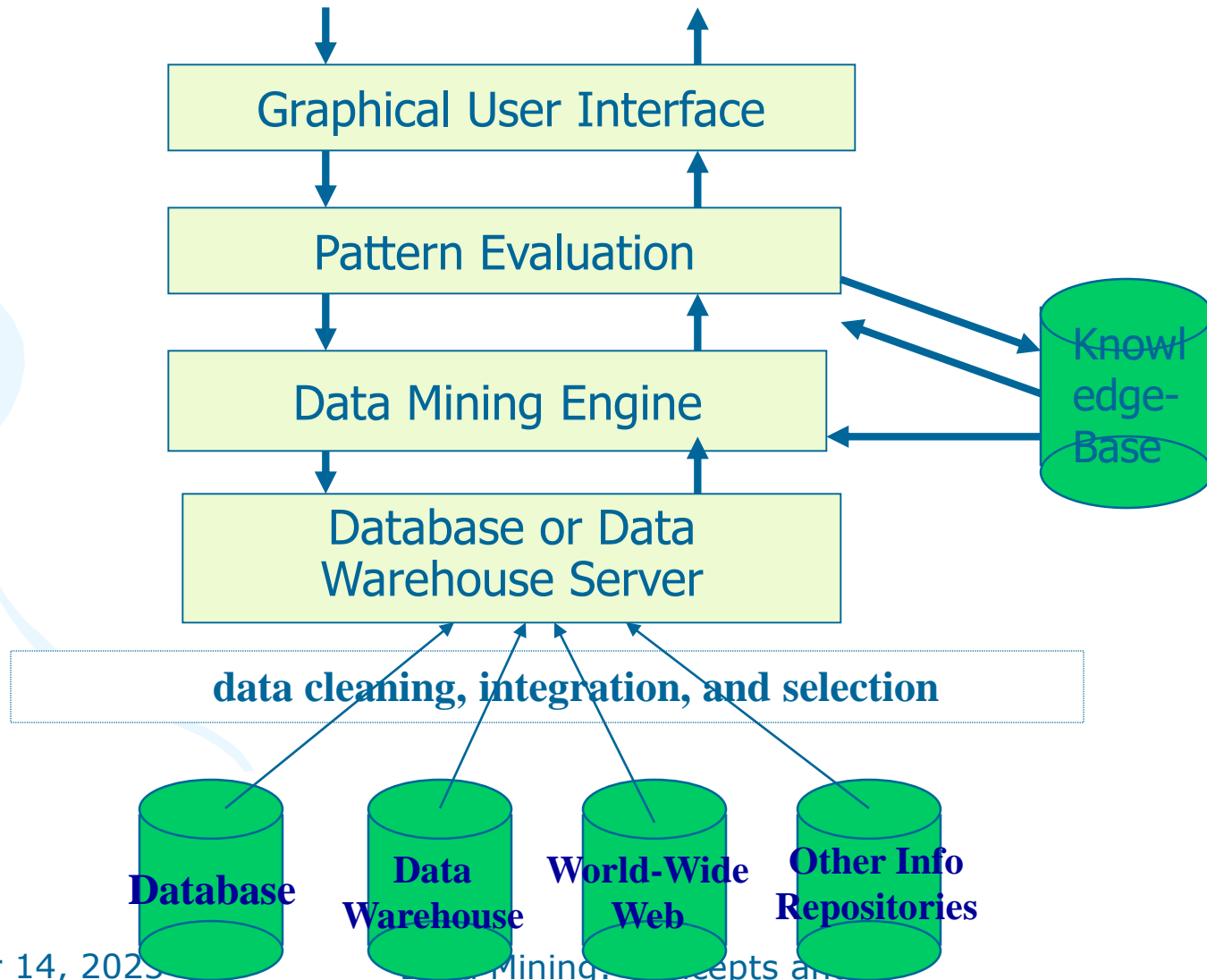
- Novelty

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





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.
- Data mining systems and architectures
- Major issues in data mining

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

Where to Find References? DBLP, CiteSeer, Google

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

Recommended Reference Books

- S. Chakrabarti. Mining the Web: Statistical Analysis of Hypertext 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**
- 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. Indurkha, 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**