

Data Mining:

Concepts and Techniques

— Chapter 1 —
— Introduction —

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

www.cs.uiuc.edu/~hanj

©2006 Jiawei Han and Micheline Kamber. All rights reserved.

December 26, 2013

Data Mining: Concepts and Techniques



Data and Information Systems (DAIS:) Course Structures at CS/UIUC

- Coverage: Database, data mining, text information systems and bioinformatics
- Data mining
 - Intro. to data warehousing and mining (CS412: Han—Fall)
 - Data mining: Principles and algorithms (CS512: Han—Spring)
 - Seminar: Advanced Topics in Data mining (CS591 Han—Fall and Spring. 1 credit unit)
 - Independent Study: only if you seriously plan to do your Ph.D. on data mining and try to demonstrate your ability
- Database Systems:
 - Database mgmt systems (CS411: Kevin Chang Fall and Spring)
 - Advanced database systems (CS511: Kevin Chang Fall)
- Text information systems
 - Text information system (CS410 ChengXiang Zhai)
- Bioinformatics
 - Introduction to Bioinformatics (Saurabh Sinha)
 - CS591 Seminar on Bioinformatics (Sinha, Zhai, Han, Schatz, Zhong)

CS412 Coverage (Chapters 1-7 of This Book)

- The book will be covered in two courses at CS, UIUC
 - CS412: Introduction to data warehousing and data mining (Fall)
 - CS512: Data mining: Principles and algorithms (Spring)
- CS412 Coverage
 - Introduction
 - Data Preprocessing
 - Data Warehouse and OLAP Technology: An Introduction
 - Advanced Data Cube Technology and Data Generalization
 - Mining Frequent Patterns, Association and Correlations
 - Classification and Prediction
 - Cluster Analysis

CS512 Coverage (Chapters 8-11 of This Book)

- Mining data streams, time-series, and sequence data
- Mining graphs, social networks and multi-relational data
- Mining object, spatial, multimedia, text and Web data
 - Mining complex data objects
 - Spatial and spatiotemporal data mining
 - Multimedia data mining
 - Text mining
 - Web mining
- Applications and trends of data mining
 - Mining business & biological data
 - Visual data mining
 - Data mining and society: Privacy-preserving data mining
- Additional (often current) themes could be added to the course



Jiawei Han

[Professor, Department of Computer Science](#)
[Univ. of Illinois at Urbana-Champaign](#)
[Rm 2132, Siebel Center for Computer Science](#)
201 N. Goodwin Avenue
Urbana, IL 61801, USA
E-mail: [hanj\[at\]cs.uiuc.edu](mailto:hanj[at]cs.uiuc.edu)

Ph.D. (1985), [Computer Science, Univ. Wisconsin-Madison](#)

Data Mining and Database Systems

[Data Mining Research Group](#)
([Data Mining Group Summary Report: Spring 2006](#))
[Database and Information Systems Research Laboratory](#)
([UIUC Academic Calendar](#))

Fax: (217) 265-6494
Web: www.cs.uiuc.edu/~hanj

● Current Research Projects ([Selected Publications](#)) ([Schedule: Meetings and Appointments](#))

- [Mining Sequential and Structured Patterns: Scalability, Flexibility, Extensibility and Applicability \(NSF/IDM\)](#)
- [Mining Dynamics of Data Streams in Multi-Dimensional Space \(NSF/IDM\)](#)
- [Endowing Biological Databases with Analytical Power: Indexing, Querying, and Mining of Complex Biological Structures \(NSF/BIO\)](#)
- [MotionEye: Querying and Mining Large Datasets of Moving Objects \(NSF/SEIII\)](#)
- [Automatic On-the-fly Detection, Characterization, Recovery, and Correction of Software Bugs in Production Runs \(NSF/ITR\)](#)
- [MAIDS: Mining Alarming Incidents in Data Streams \(ONR project, with NCSA/UIUC\)](#)
- [Mining Database Structures and Linkages across Multiple Databases \(IBM\)](#)

● Teaching

- [UIUC CS412: An Introduction to Data Warehousing and Data Mining \(Fall 2006\)](#) 12:30-1:45pm WF 1304 SC
- [UIUC CS512: Data Mining: Principles and Algorithms \(will offer again in Spring 2007\)](#)

Chapter 1. Introduction

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

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.

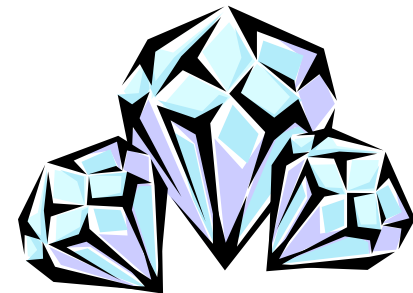
Evolution of Database Technology

- 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

What Is Data Mining?

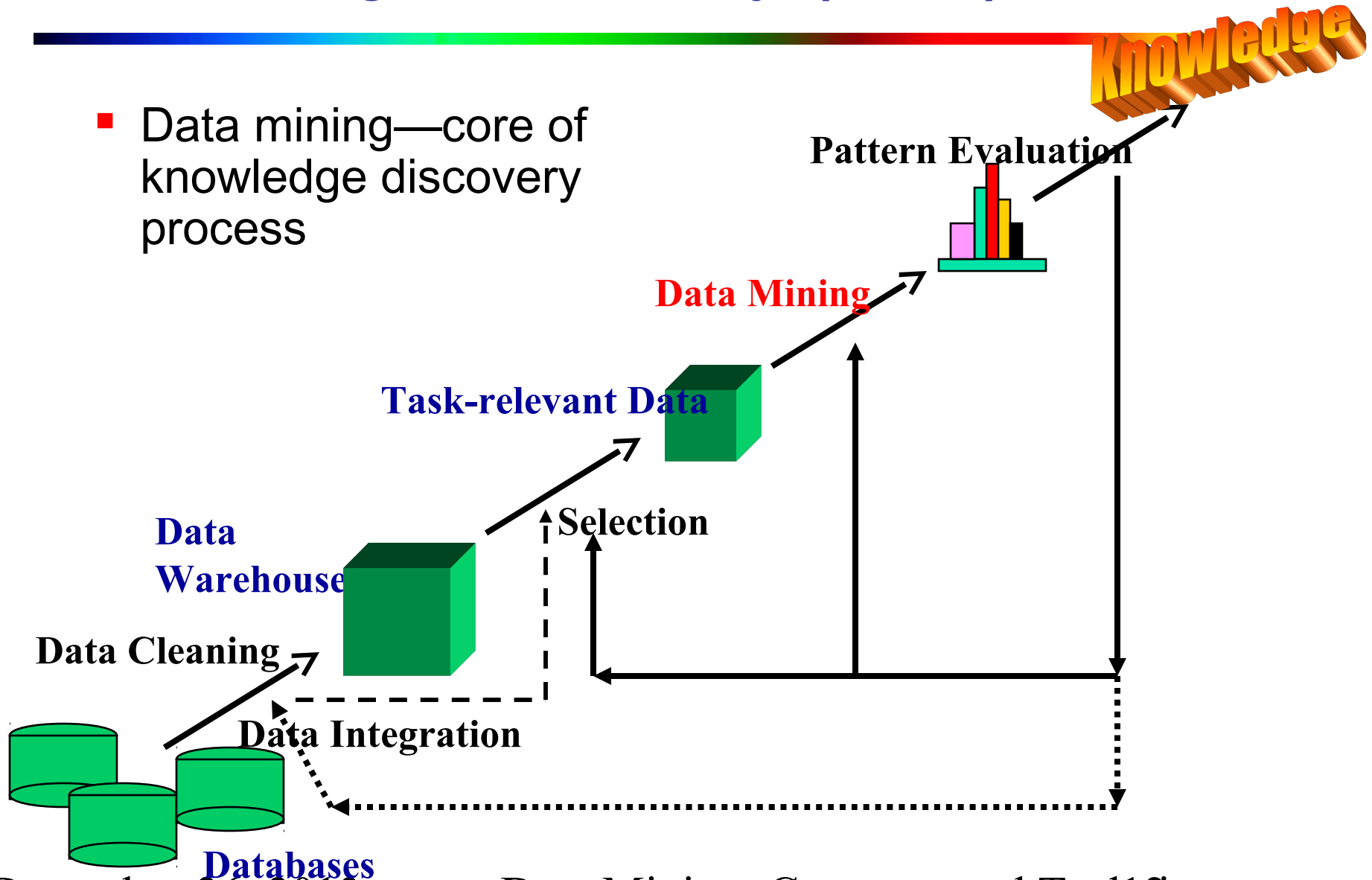


- Data mining (knowledge discovery from data)
 - Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data
 - Data mining: a misnomer?
- Alternative names
 - Knowledge discovery (mining) in databases (KDD), knowledge extraction,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

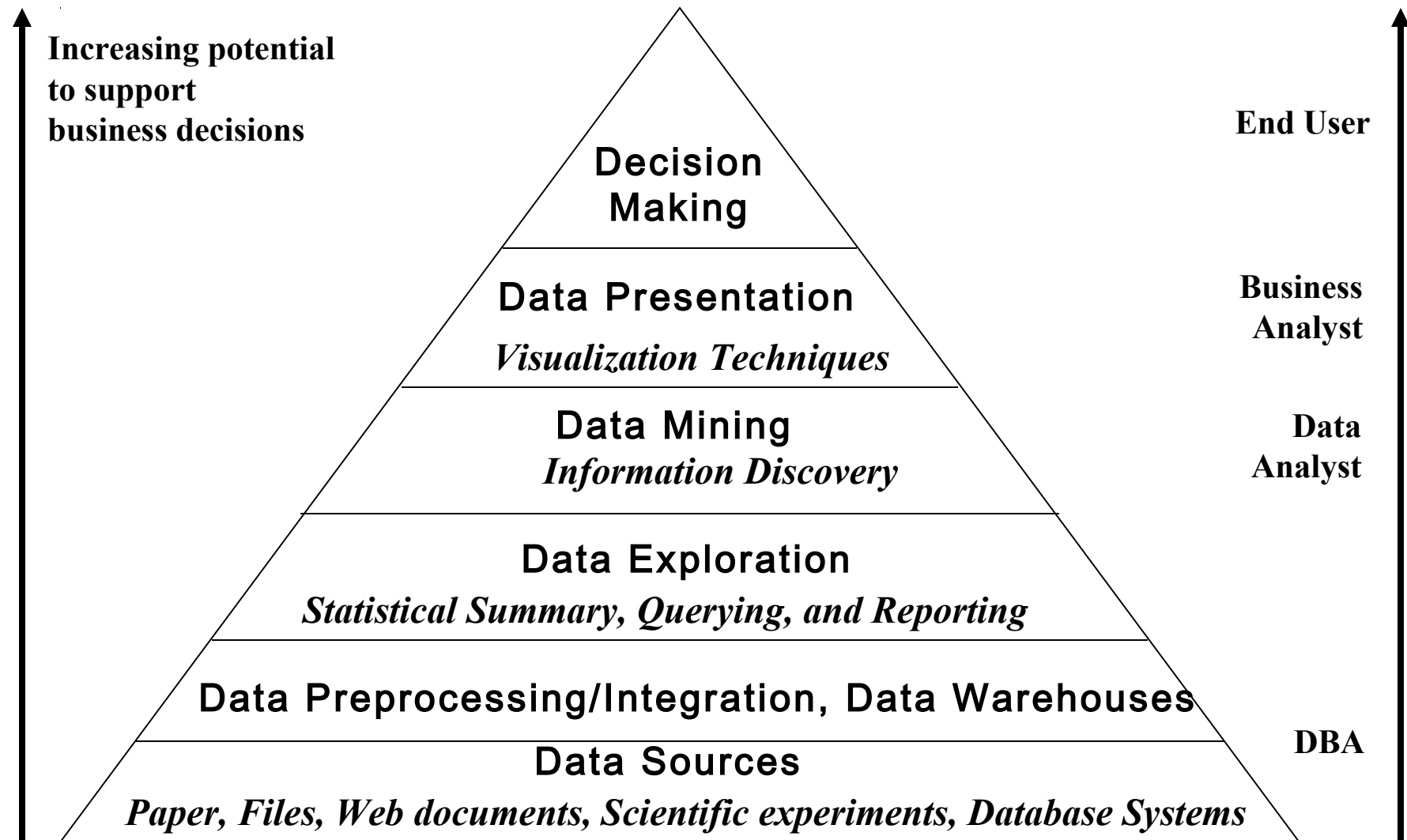


Knowledge Discovery (KDD) Process

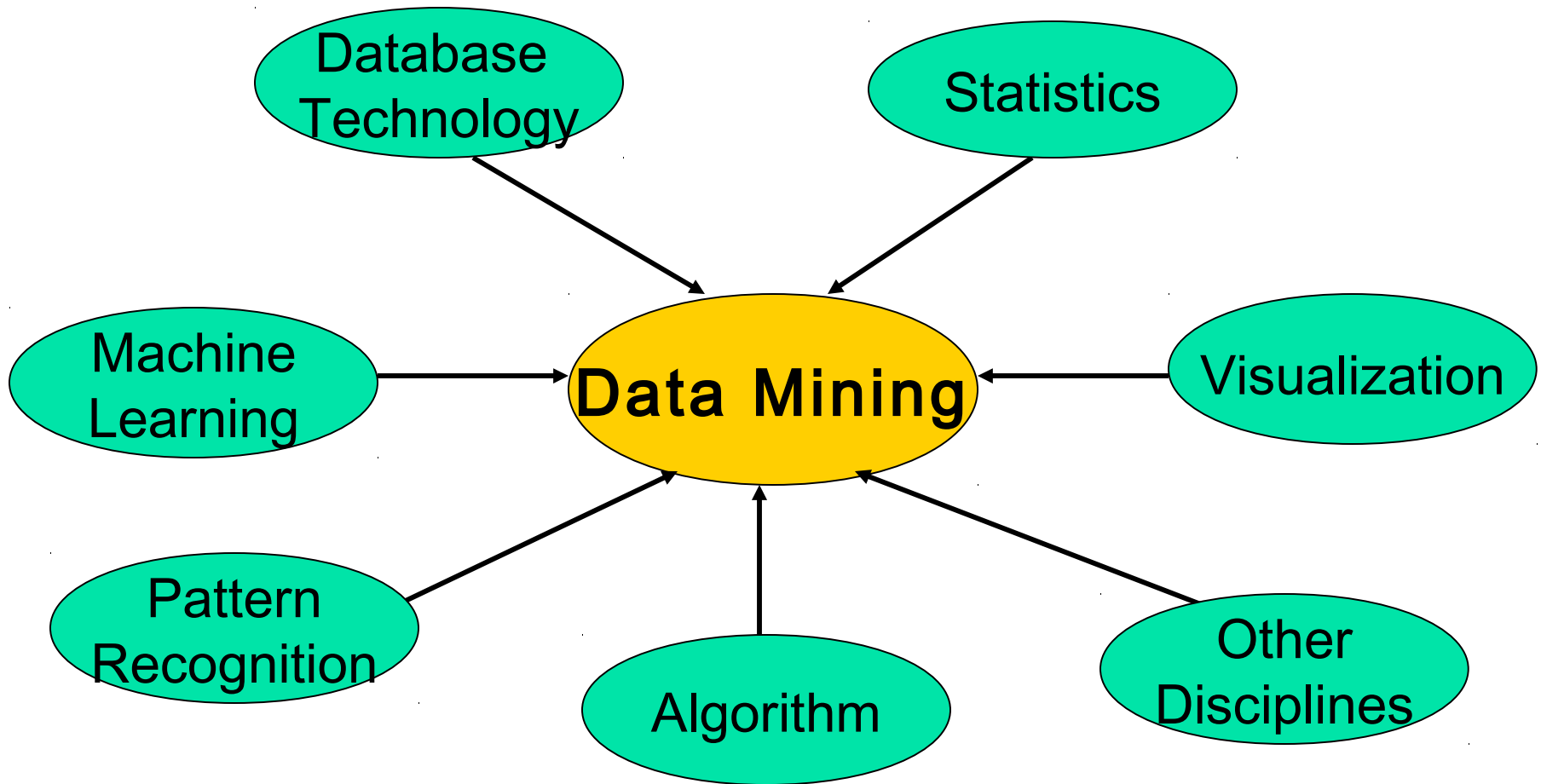
- Data mining—core of knowledge discovery process



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: 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

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

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 & invisible 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

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), AAI, 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
- 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. 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

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



Supplementary Lecture Slides

- Note: The slides following the end of chapter summary are supplementary slides that could be useful for supplementary readings or teaching
- These slides may have its corresponding text contents in the book chapters, but were omitted due to limited time in author's own course lecture
- The slides in other chapters have similar convention and treatment

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)

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

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 generate all the patterns and then filter out the uninteresting ones
 - Generate only the interesting patterns—mining query optimization

Other Pattern Mining Issues

- Precise patterns vs. approximate patterns
 - Association and correlation mining: possible find sets of precise patterns
 - But approximate patterns can be more compact and sufficient
 - How to find high quality approximate patterns??
 - Gene sequence mining: approximate patterns are inherent
 - How to derive efficient approximate pattern mining algorithms??
- Constrained vs. non-constrained patterns
 - Why constraint-based mining?
 - What are the possible kinds of constraints? How to push constraints into the mining process?

A Few Announcements (Sept. 1)

- A new section CS412ADD: CRN 48711 and its rules/arrangements
- 4th Unit for I2CS students
 - Survey report for mining new types of data
- 4th Unit for in-campus students
 - High quality implementation of one selected (to be discussed with TA/Instructor) data mining algorithm in the textbook
 - Or, a research report if you plan to devote your future research thesis on data mining

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
- Rule-based hierarchy
 - $\text{low_profit_margin}(X) \leq \text{price}(X, P_1) \text{ and } \text{cost}(X, P_2) \text{ and } (P_1 - P_2) < \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.

DMQL—A Data Mining Query Language

- Motivation
 - A DMQL can provide the ability to **support ad-hoc and interactive data mining**
 - By providing a **standardized language** like SQL
 - Hope to achieve a similar effect like that SQL has on relational database
 - Foundation for system development and evolution
 - Facilitate information exchange, technology transfer, commercialization and wide acceptance
- Design
 - DMQL is designed with the **primitives** described earlier

An Example Query in DMQL

Example 1.11 Mining classification rules. Suppose, as a marketing manager of *AllElectronics*, you would like to classify customers based on their buying patterns. You are especially interested in those customers whose salary is no less than \$40,000, and who have bought more than \$1,000 worth of items, each of which is priced at no less than \$100. In particular, you are interested in the customer's age, income, the types of items purchased, the purchase location, and where the items were made. You would like to view the resulting classification in the form of rules. This data mining query is expressed in DMQL³ as follows, where each line of the query has been enumerated to aid in our discussion.

```
use database AllElectronics_db
use hierarchy location_hierarchy for T.branch, age_hierarchy for C.age
mine classification as promising_customers
in relevance to C.age, C.income, I.type, I.place_made, T.branch
from customer C, item I, transaction T
where I.item_ID = T.item_ID and C.cust_ID = T.cust_ID
      and C.income  $\geq$  40,000 and I.price  $\geq$  100
group by T.cust_ID
having sum(I.price)  $\geq$  1,000
display as rules
```


Other Data Mining Languages & Standardization Efforts

- Association rule language specifications
 - MSQL (Imielinski & Virmani'99)
 - MineRule (Meo Psaila and Ceri'96)
 - Query flocks based on Datalog syntax (Tsur et al'98)
- OLEDB for DM (Microsoft'2000) and recently DMX (Microsoft SQLServer 2005)
 - Based on OLE, OLE DB, OLE DB for OLAP, C#
 - Integrating DBMS, data warehouse and data mining
- DMML (Data Mining Mark-up Language) by DMG (www.dmg.org)
 - Providing a platform and process structure for effective data mining
 - Emphasizing on deploying data mining technology to solve business problems

Integration of Data Mining and Data Warehousing

- **Data mining systems, DBMS, Data warehouse systems coupling**
 - No coupling, loose-coupling, semi-tight-coupling, tight-coupling
- **On-line analytical mining data**
 - integration of mining and OLAP technologies
- **Interactive mining multi-level knowledge**
 - Necessity of mining knowledge and patterns at different levels of abstraction by drilling/rolling, pivoting, slicing/dicing, etc.
- **Integration of multiple mining functions**
 - Characterized classification, first clustering and then association

Coupling Data Mining with DB/DW Systems

- No coupling—flat file processing, not recommended
- Loose coupling
 - Fetching data from DB/DW
- Semi-tight coupling—enhanced DM performance
 - Provide efficient implement a few data mining primitives in a DB/DW system, e.g., sorting, indexing, aggregation, histogram analysis, multiway join, precomputation of some stat functions
- Tight coupling—A uniform information processing environment
 - DM is smoothly integrated into a DB/DW system, mining query is optimized based on mining query, indexing, query processing methods, etc.

Architecture: Typical Data Mining System

