Data Mining

Session 1 – Main Theme
Introduction to Data Mining

Dr. Jean-Claude Franchitti

New York University
Computer Science Department
Courant Institute of Mathematical Sciences

Adapted from course textbook resources
Data Mining Concepts and Techniques (2nd Edition)
Jiawei Han and Micheline Kamber

Agenda

1 **Instructor and Course Introduction**

2 **Introduction to Data Mining**

3 **Summary and Conclusion**

Who am I?

- Profile -

- 27 years of experience in the Information Technology Industry, including thirteen years of experience working for leading IT consulting firms such as Computer Sciences Corporation
- PhD in Computer Science from University of Colorado at Boulder
- Past CEO and CTO
- Held senior management and technical leadership roles in many large IT Strategy and Modernization projects for fortune 500 corporations in the insurance, banking, investment banking, pharmaceutical, retail, and information management industries
- Contributed to several high-profile ARPA and NSF research projects
- Played an active role as a member of the OMG, ODMG, and X3H2 standards committees and as a Professor of Computer Science at Columbia initially and New York University since 1997
- Proven record of delivering business solutions on time and on budget
- Original designer and developer of jcrew.com and the suite of products now known as IBM InfoSphere DataStage
- Creator of the Enterprise Architecture Management Framework (EAMF) and main contributor to the creation of various maturity assessment methodology
- Developed partnerships between several companies and New York University to incubate new methodologies (e.g., EA maturity assessment methodology developed in Fall 2008), develop proof of concept software, recruit skilled graduates, and increase the companies' visibility

How to reach me?

☎	Cell	(212) 203-5004
@	Email	jcf@cs.nyu.edu
Y!	AIM, Y! IM, ICQ	jcf2_2003
	MSN IM	jcf2_2003@yahoo.com
Linkedin	LinkedIn	http://www.linkedin.com/in/jcfranchitti
twitter	Twitter	http://twitter.com/jcfranchitti
skype	Skype	jcf2_2003@yahoo.com

What is the class about?

- Course description and syllabus:
  - » http://www.nyu.edu/classes/jcf/g22.3033-002/
  - » http://www.cs.nyu.edu/courses/spring10/G22.3033-002/index.html

- Textbooks:
  - » ***Data Mining: Concepts and Techniques (2nd Edition)***

Jiawei Han, Micheline Kamber
Morgan Kaufmann
ISBN-10: 1-55860-901-6, ISBN-13: 978-1-55860-901-3, (2006)

  - » ***Microsoft SQL Server 2008 Analysis Services Step by Step***

Scott Cameron
Microsoft Press
ISBN-10: 0-73562-620-0, ISBN-13: 978-0-73562-620-31 1st Edition (04/15/09)

Icons / Metaphors

Information

Common Realization

Knowledge/Competency Pattern

Governance

Alignment

Solution Approach

Agenda

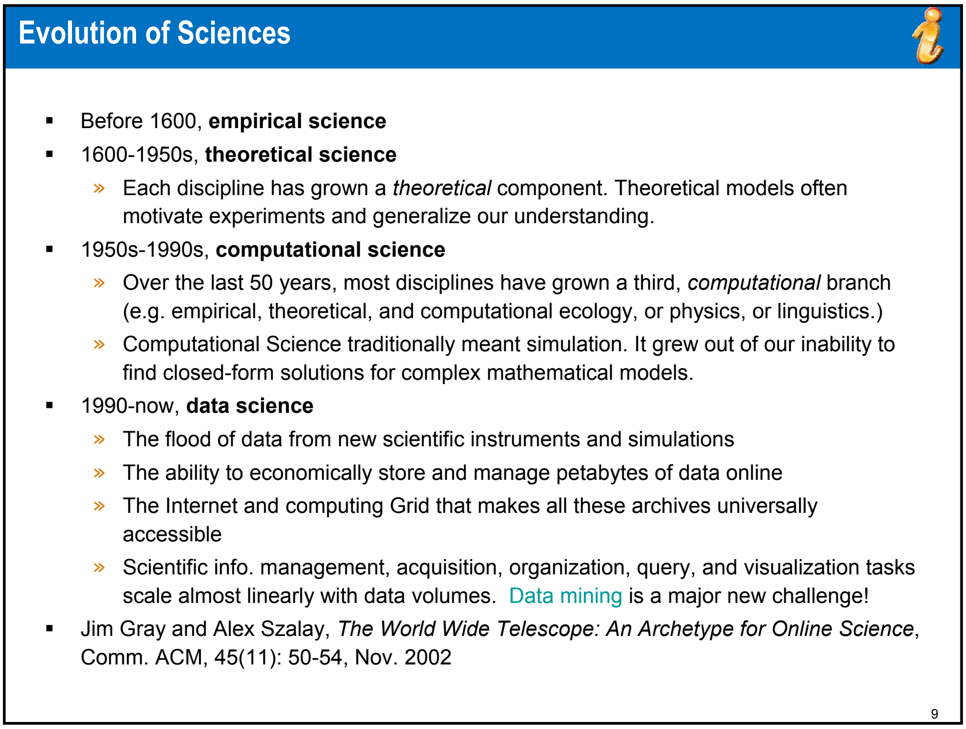
1 Instructor and Course Introduction

2 Introduction to Data Mining

3 Summary and Conclusion

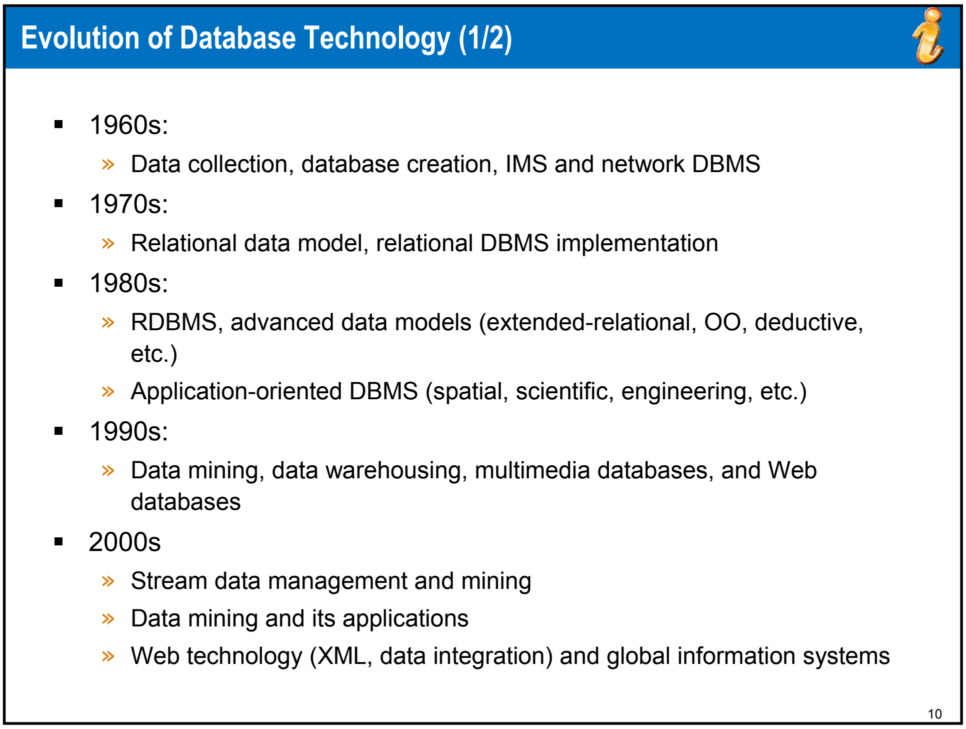
Introduction to Data Mining - Sub-Topics

- Why Data Mining?
  - » Data Mining: A Natural Evolution of Science and Technology
- What Is Data Mining?
  - » Data Mining: Essential in a Knowledge Discovery Process
  - » Data Mining: A Confluence of Multiple Disciplines
- A Multi-Dimensional View of Data Mining
  - » Knowledge to Be Mined
  - » Data to Be Mined
  - » Technology Utilized
  - » Applications Adapted
- Data Mining Functionalities: What Kinds of Patterns Can Be Mined?
  - » Generalization
  - » Mining Frequent Patterns, Associations, and Correlations
  - » Classification
  - » Cluster Analysis
  - » Outlier Analysis
- Data mining: On What Kinds of Data?
- Time and Ordering: Sequential Pattern, Trend and Evolution Analysis
- Structure and Network Analysis
- Evaluation of knowledge
- Applications of Data Mining
- Major Challenges in Data Mining
- A Brief History of Data Mining and Data Mining Society



Evolution of Sciences

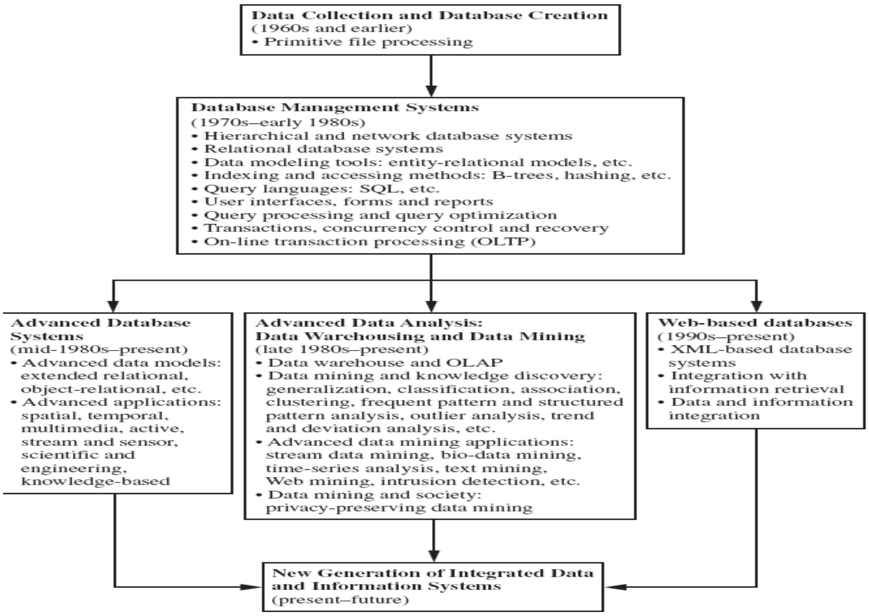
- Before 1600, **empirical science**
- 1600-1950s, **theoretical science**
  - » Each discipline has grown a *theoretical* component. Theoretical models often motivate experiments and generalize our understanding.
- 1950s-1990s, **computational science**
  - » Over the last 50 years, most disciplines 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 (1/2)

- 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

Evolution of Database Technology (2/2)



Data Collection and Database Creation
(1960s and earlier)
- Primitive file processing

Database Management Systems
(1970s–early 1980s)
- Hierarchical and network database systems
- Relational database systems
- Data modeling tools: entity-relational models, etc.
- Indexing and accessing methods: B-trees, hashing, etc.
- Query languages: SQL, etc.
- User interfaces, forms and reports
- Query processing and query optimization
- Transactions, concurrency control and recovery
- On-line transaction processing (OLTP)

Advanced Database Systems
(mid-1980s–present)
- Advanced data models: extended relational, object-relational, etc.
- Advanced applications: spatial, temporal, multimedia, active, stream and sensor, scientific and engineering, knowledge-based

Advanced Data Analysis: Data Warehousing and Data Mining
(late 1980s–present)
- Data warehouse and OLAP
- Data mining and knowledge discovery: generalization, classification, association, clustering, frequent pattern and structured pattern analysis, outlier analysis, trend and deviation analysis, etc.
- Advanced data mining applications: stream data mining, bio-data mining, time-series analysis, text mining, Web mining, intrusion detection, etc.
- Data mining and society: privacy-preserving data mining

Web-based databases
(1990s–present)
- XML-based database systems
- Integration with information retrieval
- Data and information integration

New Generation of Integrated Data and Information Systems
(present–future)

Why Data Mining? (1/2)

- 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
- <u>We are drowning in data, but starving for knowledge!</u>
- “Necessity is the mother of invention”—Data mining—Automated analysis of massive data sets

Why Data Mining? (2/2)

•Associations (e.g. linking purchase of pizza with beer)

•Sequences (e.g. tying events together: marriage and purchase of furniture)

•Classifications (e.g. recognizing patterns such as the attributes of employees that are most likely to quit)

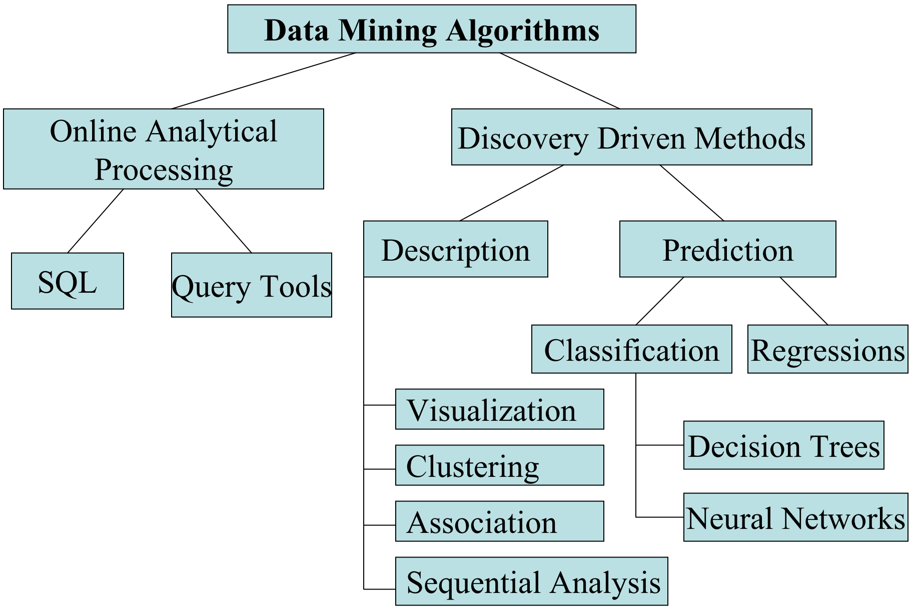
•Forecasting (e.g. predicting buying habits of customers based on past patterns) Expert systems or small ML/statistical programs

DOGBERT CONSULTS
MY DATA-MINING SOFTWARE HAS FOUND ANOTHER MESSAGE FROM GOD.
scottadams@aol.com
www.dilbert.com
IT SAYS YOU'VE BEEN STEALING LUNCHES FROM THE REFRIGERATOR IN THE BREAK ROOM.
© 1999 United Feature Syndicate, Inc.
THEN IT SAYS, "HA HA, THAT WASN'T PUDDING!"

What Can Data Mining Do?

- Classification
  - » Classify credit applicants as low, medium, high risk
  - » Classify insurance claims as normal, suspicious
- Estimation
  - » Estimate the probability of a direct mailing response
  - » Estimate the lifetime value of a customer
- Prediction
  - » Predict which customers will leave within six months
  - » Predict the size of the balance that will be transferred by a credit card prospect
- Association
  - » Find out items customers are likely to buy together
  - » Find out what books to recommend to Amazon.com users
- Clustering
  - » Difference from classification: classes are unknown!

Sample Data Mining Algorithms



Data Mining Algorithms

Online Analytical Processing

SQL

Query Tools

Discovery Driven Methods

Description

Visualization

Clustering

Association

Sequential Analysis

Prediction

Classification

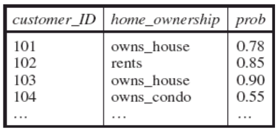
Decision Trees

Neural Networks

Regressions

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



customer_ID	*home_ownership*	*prob*
101	owns_house	0.78
102	rents	0.85
103	owns_house	0.90
104	owns_condo	0.55
...	...	...

Example 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
- Direct Marketing
  - » Identify which prospects should be included in a mailing list
- Market segmentation
  - » identify common characteristics of customers who buy same products
- Market Basket Analysis
  - » Identify what products are likely to be bought together
- 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)

Sample Market Basket Analysis

- Association and sequence discovery
- Principal concepts
  - Support or Prevalence: frequency that a particular
  - association appears in the database
  - Confidence: conditional predictability of B, given A
- Example:
  - Total daily transactions: 1,000
  - Number which include “soda”: 500
  - Number which include “orange juice”: 800
  - Number which include “soda” and “orange juice”: 450
  - SUPPORT for “soda and orange juice” = 45% (450/1,000)
  - CONFIDENCE of “soda à orange juice” = 90% (450/500)
  - CONFIDENCE of “orange juice à soda” = 56% (450/800)

Example 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

Example 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, insurance Claims Analysis
    - Discover patterns of fraudulent transactions
    - Compare current transactions against those patterns
  - » 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

Other Applications (1/2)

- Sports
  - IBM Advanced Scout analyzed NBA game statistics (shots blocked, assists, and fouls) to gain competitive advantage for New York Knicks and Miami Heat
- Astronomy
  - JPL and the Palomar Observatory discovered 22 quasars with the help of data mining
- Internet Web Surf-Aid
  - IBM Surf-Aid applies data mining algorithms to Web access logs for market-related pages to discover customer preference and behavior pages, analyzing effectiveness of Web marketing, improving Web site organization, etc.

Example: Amazon.com book recommendations

- Example: Identify books to recommend to customers
  - Company keeps log of past customer purchases
  - Represent each customer as a vector whose components are the past purchases
  - Define a "distance" function for comparing customers
  - Based on this distance function, identify the customer's nearest neighbor set (NNS)
  - Identify books that have been purchased by a large percentage of the nearest neighbor set but not by the customer
  - Recommend these books to the customer as possible next purchases

Introduction to Data Mining – Sub-Topics

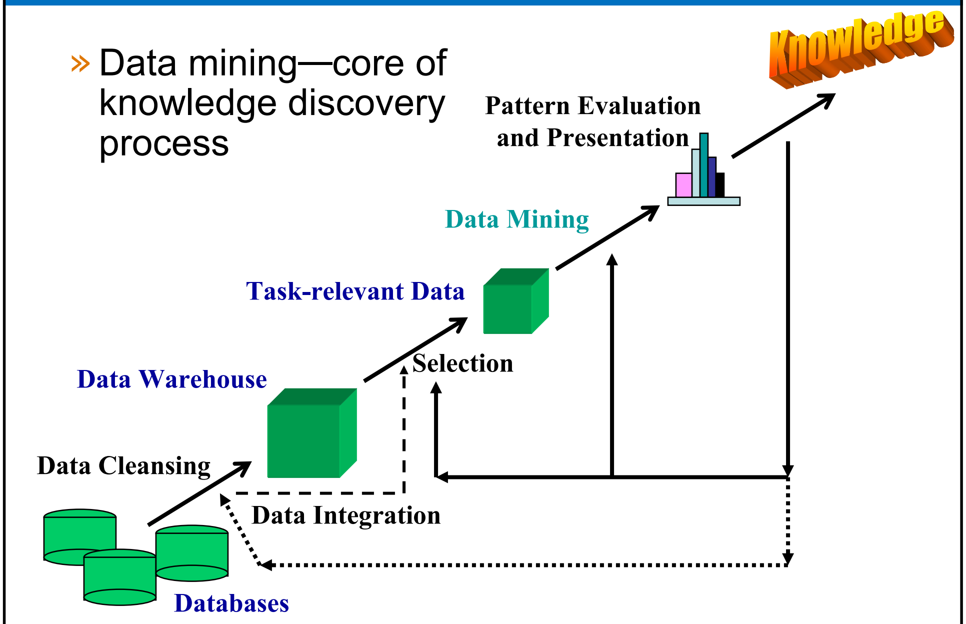
- Why Data Mining?
  - Data Mining: A Natural Evolution of Science and Technology
- What Is Data Mining?
  - Data Mining: Essential in a Knowledge Discovery Process
  - Data Mining: A Confluence of Multiple Disciplines
- A Multi-Dimensional View of Data Mining
  - Knowledge to Be Mined
  - Data to Be Mined
  - Technology Utilized
  - Applications Adapted
- Data Mining Functionalities: What Kinds of Patterns Can Be Mined?
  - Generalization
  - Mining Frequent Patterns, Associations, and Correlations
  - Classification
  - Cluster Analysis
  - Outlier Analysis
- Data mining: On What Kinds of Data?
- Time and Ordering: Sequential Pattern, Trend and Evolution Analysis
- Structure and Network Analysis
- Evaluation of knowledge
- Applications of Data Mining
- Major Challenges in Data Mining
- A Brief History of Data Mining and Data Mining Society

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

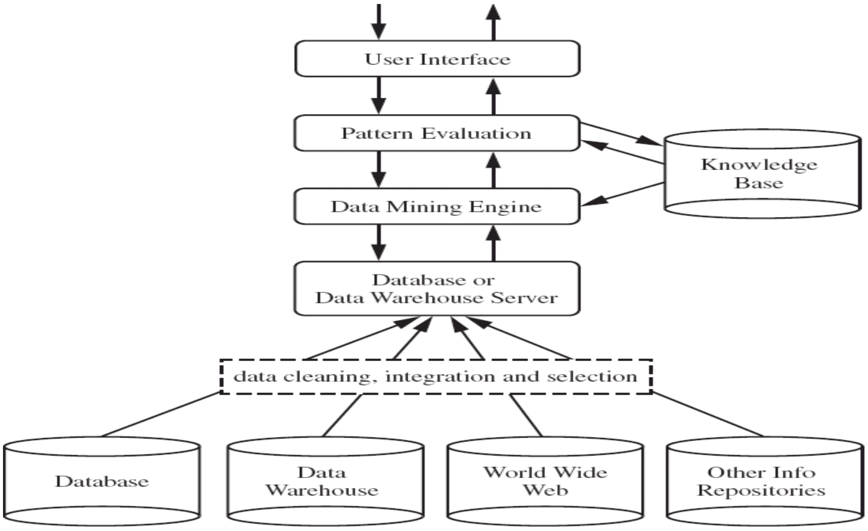
- Data mining—core of knowledge discovery process



Example: A Web Mining Framework (1/2)

- 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

Example: A Web Mining Framework (2/2)



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.

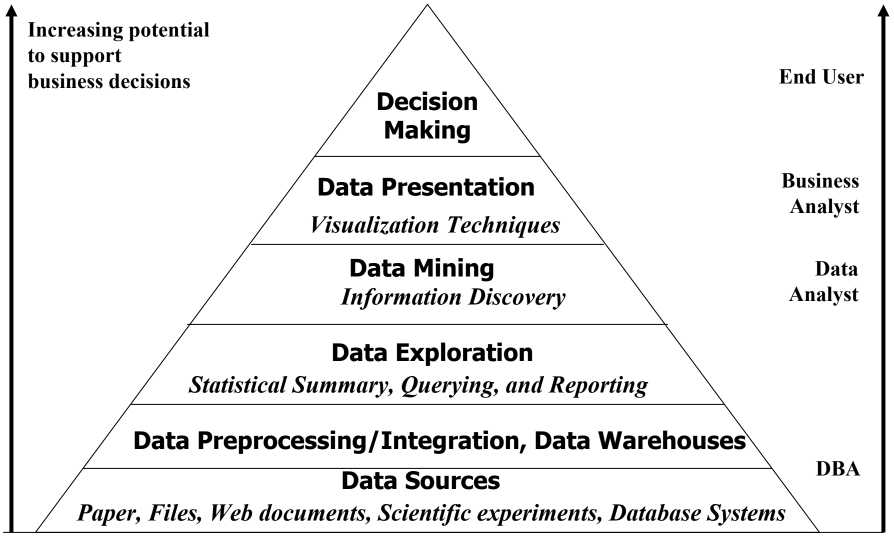
Find All and Only Interesting Patterns?

- 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

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?

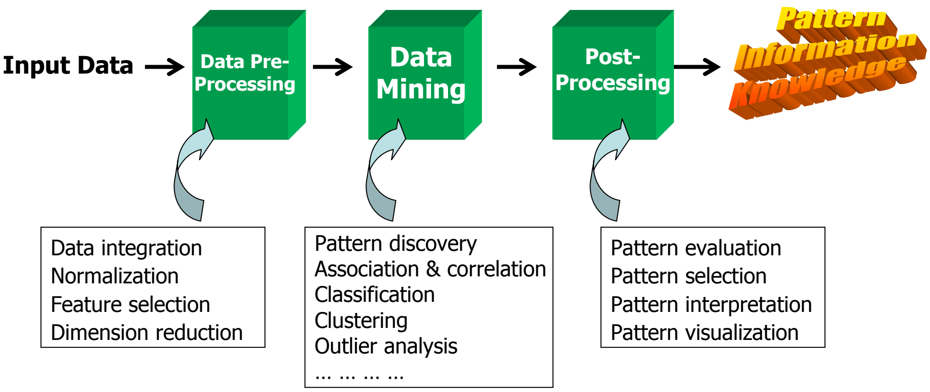
Data Mining and Business Intelligence



Example: Mining vs. Data Exploration

- Business intelligence view
  - Warehouse, data cube, reporting but not much mining
- Business objects vs. data mining tools
- Supply chain example: tools
- Data presentation
- Exploration

KDD Process: A Typical View from ML and Statistics



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
… … … …

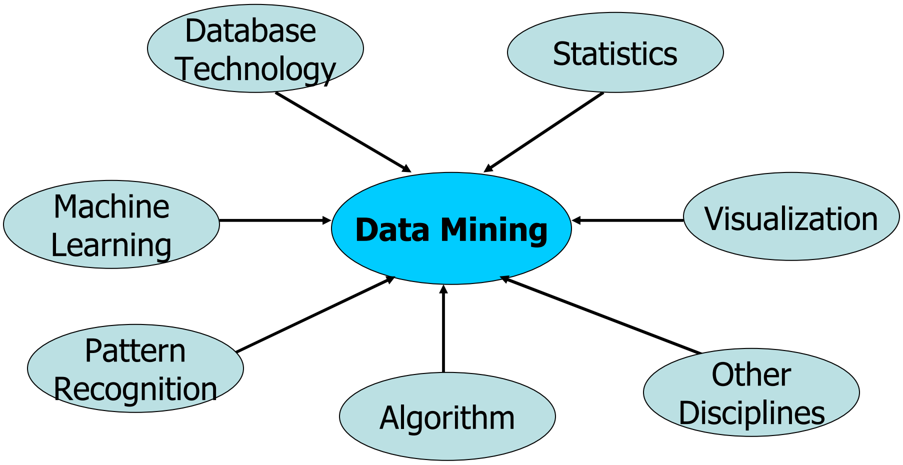
Pattern evaluation
Pattern selection
Pattern interpretation
Pattern visualization

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

Example: Medical Data Mining

- Health care & medical data mining – often adopted such a view in statistics and machine learning
- Preprocessing of the data (including feature extraction and dimension reduction)
- Classification or/and clustering processes
- Post-processing for presentation

Data Mining: Confluence of Multiple Disciplines

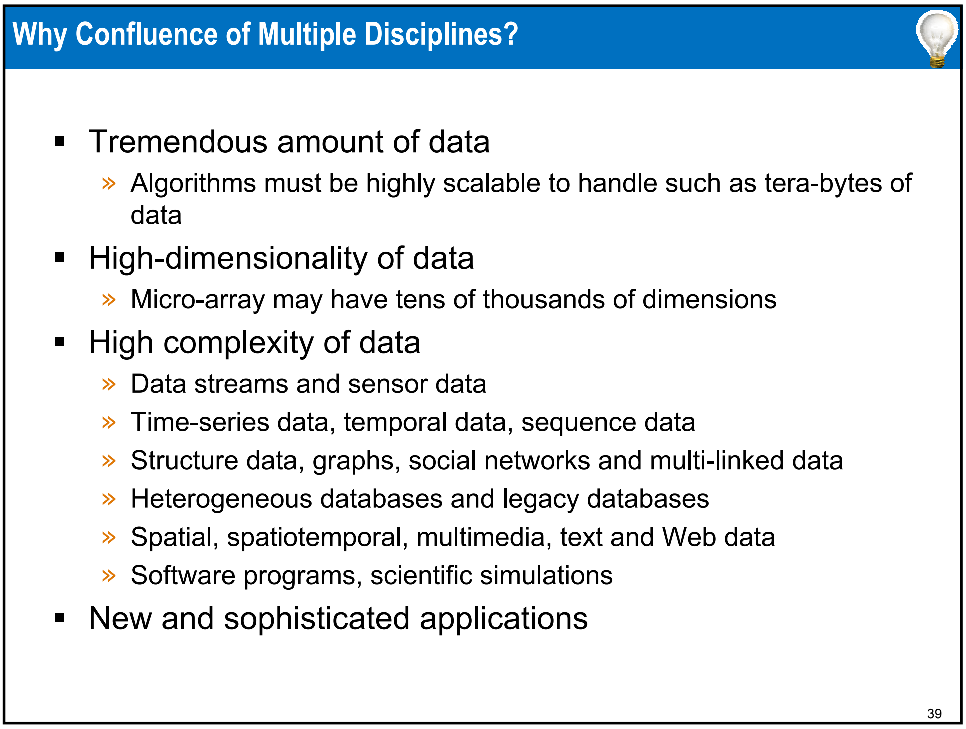


Introduction to Data Mining - Sub-Topics

- Why Data Mining?
  - » Data Mining: A Natural Evolution of Science and Technology
- What Is Data Mining?
  - » Data Mining: Essential in a Knowledge Discovery Process
  - » Data Mining: A Confluence of Multiple Disciplines
- A Multi-Dimensional View of Data Mining
  - » Knowledge to Be Mined
  - » Data to Be Mined
  - » Technology Utilized
  - » Applications Adapted
- Data Mining Functionalities: What Kinds of Patterns Can Be Mined?
  - » Generalization
  - » Mining Frequent Patterns, Associations, and Correlations
  - » Classification
  - » Cluster Analysis
  - » Outlier Analysis
- Data mining: On What Kinds of Data?
- Time and Ordering: Sequential Pattern, Trend and Evolution Analysis
- Structure and Network Analysis
- Evaluation of knowledge
- Applications of Data Mining
- Major Challenges in Data Mining
- A Brief History of Data Mining and Data Mining Society

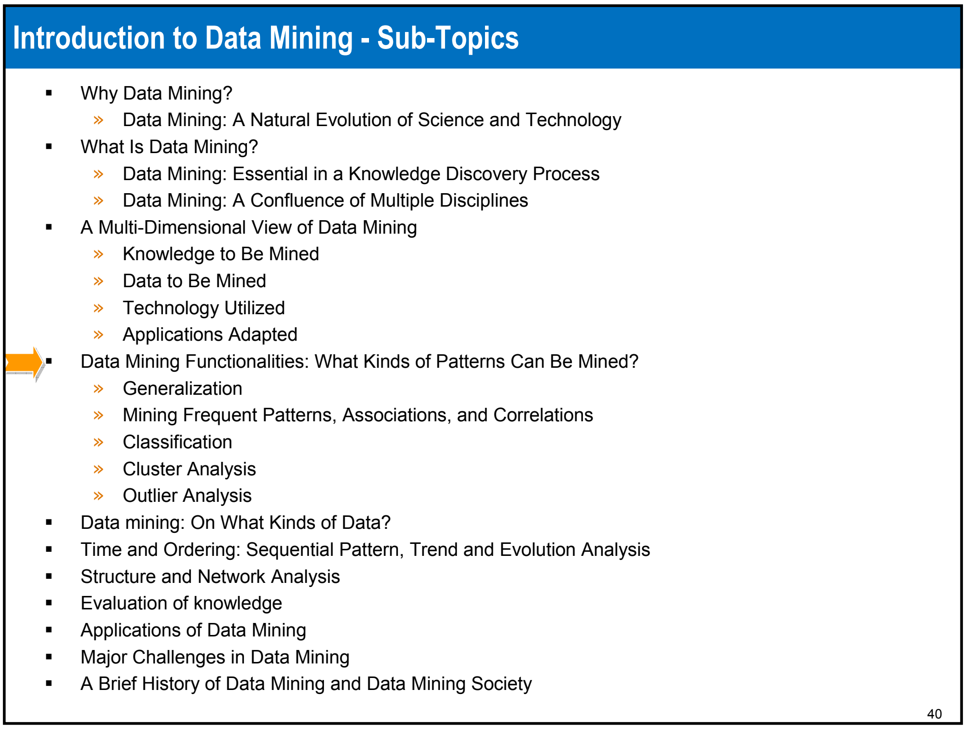
Multi-Dimensional View of Data Mining

- **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.



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



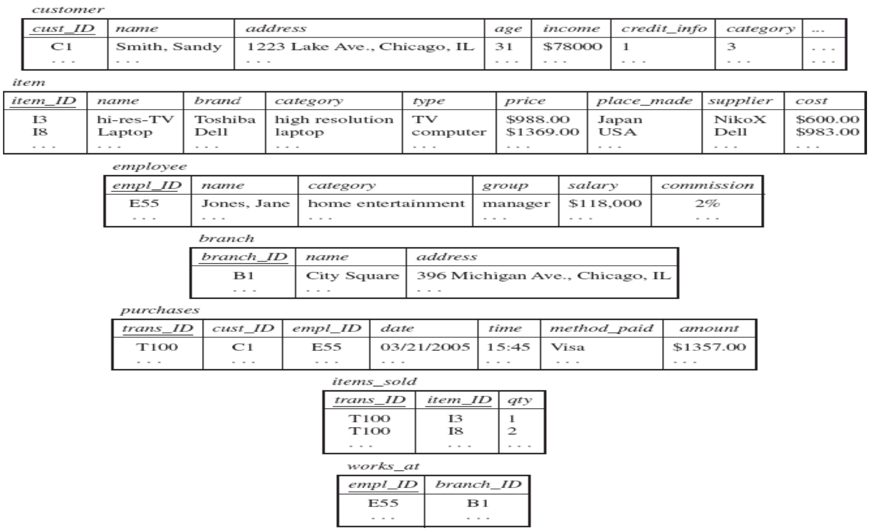
Introduction to Data Mining - Sub-Topics

- Why Data Mining?
  - » Data Mining: A Natural Evolution of Science and Technology
- What Is Data Mining?
  - » Data Mining: Essential in a Knowledge Discovery Process
  - » Data Mining: A Confluence of Multiple Disciplines
- A Multi-Dimensional View of Data Mining
  - » Knowledge to Be Mined
  - » Data to Be Mined
  - » Technology Utilized
  - » Applications Adapted
- Data Mining Functionalities: What Kinds of Patterns Can Be Mined?
  - » Generalization
  - » Mining Frequent Patterns, Associations, and Correlations
  - » Classification
  - » Cluster Analysis
  - » Outlier Analysis
- Data mining: On What Kinds of Data?
- Time and Ordering: Sequential Pattern, Trend and Evolution Analysis
- Structure and Network Analysis
- Evaluation of knowledge
- Applications of Data Mining
- Major Challenges in Data Mining
- A Brief History of Data Mining and Data Mining Society

Data Mining Function: (1) Generalization (1/4)

- 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

Data Mining Function: (1) Generalization (2/4)



customer

cust_ID	*name*	*address*	*age*	*income*	*credit_info*	*category*	...
C1	Smith, Sandy	1223 Lake Ave., Chicago, IL	31	$78000	1	3	. . .
. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .

item

item_ID	*name*	*brand*	*category*	*type*	*price*	*place_made*	*supplier*	*cost*
I3	hi-res-TV	Toshiba	high resolution	TV	$988.00	Japan	NikoX	$600.00
I8	Laptop	Dell	laptop	computer	$1369.00	USA	Dell	$983.00
. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .

employee

empl_ID	*name*	*category*	*group*	*salary*	*commission*
E55	Jones, Jane	home entertainment	manager	$118,000	2%
. . .	. . .	. . .	. . .	. . .	. . .

branch

branch_ID	*name*	*address*
B1	City Square	396 Michigan Ave., Chicago, IL
. . .	. . .	. . .

purchases

trans_ID	*cust_ID*	*empl_ID*	*date*	*time*	*method_paid*	*amount*
T100	C1	E55	03/21/2005	15:45	Visa	$1357.00
. . .	. . .	. . .	. . .	. . .	. . .	. . .

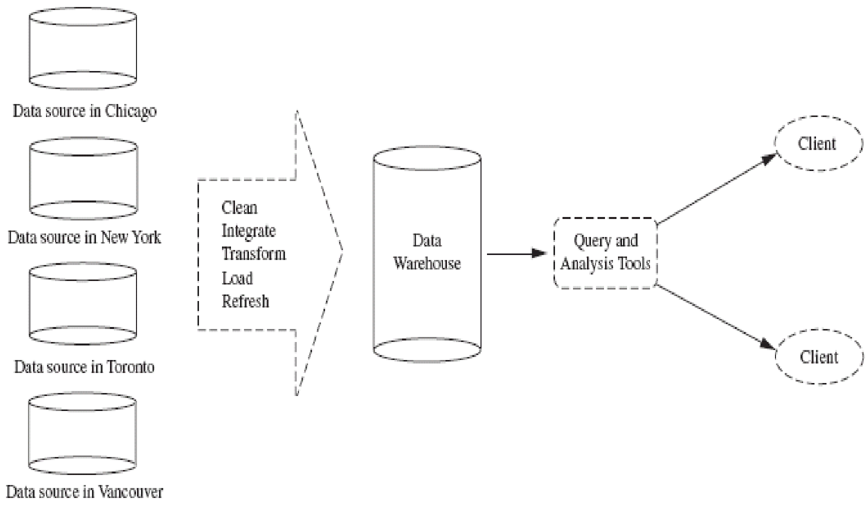
items_sold

trans_ID	*item_ID*	*qty*
T100	I3	1
T100	I8	2
. . .	. . .	. . .

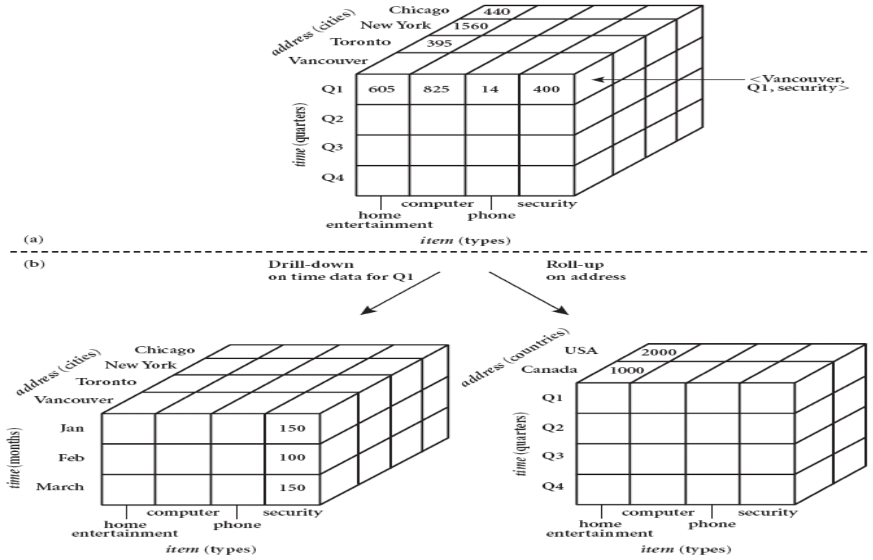
works_at

empl_ID	*branch_ID*
E55	B1
. . .	. . .

Data Mining Function: (1) Generalization (3/4)



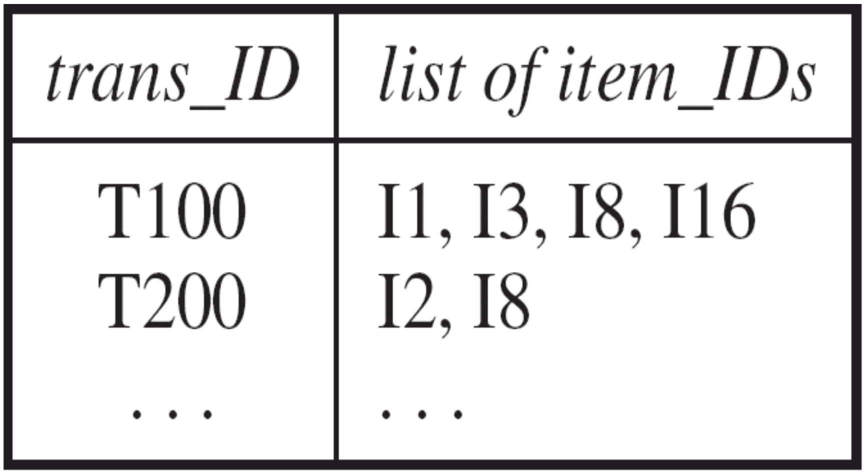
Data Mining Function: (1) Generalization (4/4)



Data Mining Function: (2) Association and Correlation Analysis (1/2)

- 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: (2) Association and Correlation Analysis (2/2)

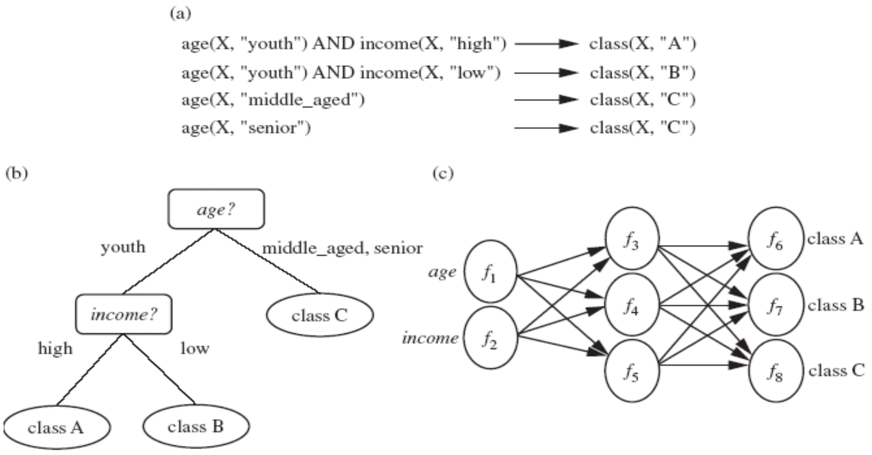


trans_ID	*list of item_IDs*
T100	I1, I3, I8, I16
T200	I2, I8
. . .	. . .

Data Mining Function: (3) Classification (1/2)

- Classification and label prediction
  - » Construct models (functions) based on some training examples
  - » Describe and distinguish classes or concepts for future prediction
    - E.g., classify countries based on (climate), or classify cars based on (gas mileage)
  - » Predict some unknown class labels
- Typical methods
  - » Decision trees, naïve Bayesian classification, support vector machines, neural networks, rule-based classification, pattern-based classification, logistic regression, …
- Typical applications:
  - » Credit card fraud detection, direct marketing, classifying stars, diseases, web-pages, …

Data Mining Function: (3) Classification (2/2)



(a)

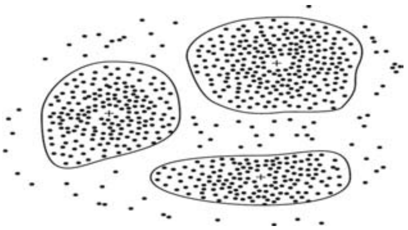
age(X, "youth") AND income(X, "high") ⟶ class(X, "A")
age(X, "youth") AND income(X, "low") ⟶ class(X, "B")
age(X, "middle_aged") ⟶ class(X, "C")
age(X, "senior") ⟶ class(X, "C")

Data Mining Function: (4) Cluster Analysis

- Unsupervised learning (i.e., Class label is unknown)
- Group data to form new categories (i.e., clusters), e.g., cluster houses to find distribution patterns
- Principle: Maximizing intra-class similarity & minimizing interclass similarity
- Many methods and applications

Data Mining Function: (5) Outlier Analysis

- Outlier analysis
  - » Outlier: A data object that does not comply with the general behavior of the data
  - » Noise or exception? — One person’s garbage could be another person’s treasure
  - » Methods: by product of clustering or regression analysis, …
  - » Useful in fraud detection, rare events analysis



Top-10 Most Popular DM Algorithms: 18 Identified Candidates (1/3)

- 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 (2/3)

- 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 (3/3)

- 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)**

Introduction to Data Mining - Sub-Topics

- Why Data Mining?
  - » Data Mining: A Natural Evolution of Science and Technology
- What Is Data Mining?
  - » Data Mining: Essential in a Knowledge Discovery Process
  - » Data Mining: A Confluence of Multiple Disciplines
- A Multi-Dimensional View of Data Mining
  - » Knowledge to Be Mined
  - » Data to Be Mined
  - » Technology Utilized
  - » Applications Adapted
- Data Mining Functionalities: What Kinds of Patterns Can Be Mined?
  - » Generalization
  - » Mining Frequent Patterns, Associations, and Correlations
  - » Classification
  - » Cluster Analysis
  - » Outlier Analysis
- Data mining: On What Kinds of Data?
- Time and Ordering: Sequential Pattern, Trend and Evolution Analysis
- Structure and Network Analysis
- Evaluation of knowledge
- Applications of Data Mining
- Major Challenges in Data Mining
- A Brief History of Data Mining and Data Mining Society

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

Introduction to Data Mining - Sub-Topics

- Why Data Mining?
  - » Data Mining: A Natural Evolution of Science and Technology
- What Is Data Mining?
  - » Data Mining: Essential in a Knowledge Discovery Process
  - » Data Mining: A Confluence of Multiple Disciplines
- A Multi-Dimensional View of Data Mining
  - » Knowledge to Be Mined
  - » Data to Be Mined
  - » Technology Utilized
  - » Applications Adapted
- Data Mining Functionalities: What Kinds of Patterns Can Be Mined?
  - » Generalization
  - » Mining Frequent Patterns, Associations, and Correlations
  - » Classification
  - » Cluster Analysis
  - » Outlier Analysis
- Data mining: On What Kinds of Data?
- Time and Ordering: Sequential Pattern, Trend and Evolution Analysis
- Structure and Network Analysis
- Evaluation of knowledge
- Applications of Data Mining
- Major Challenges in Data Mining
- A Brief History of Data Mining and Data Mining Society

Time and Ordering: Sequential Pattern, Trend and Evolution Analysis

- Sequence, trend and evolution analysis
  - » Trend, time-series, and deviation analysis: e.g., regression and value prediction
  - » Sequential pattern mining
    - e.g., first buy digital camera, then buy large SD memory cards
  - » Periodicity analysis
  - » Motifs and biological sequence analysis
    - Approximate and consecutive motifs
  - » Similarity-based analysis
- Mining data streams
  - » Ordered, time-varying, potentially infinite, data streams

Introduction to Data Mining - Sub-Topics

- Why Data Mining?
  - Data Mining: A Natural Evolution of Science and Technology
- What Is Data Mining?
  - Data Mining: Essential in a Knowledge Discovery Process
  - Data Mining: A Confluence of Multiple Disciplines
- A Multi-Dimensional View of Data Mining
  - Knowledge to Be Mined
  - Data to Be Mined
  - Technology Utilized
  - Applications Adapted
- Data Mining Functionalities: What Kinds of Patterns Can Be Mined?
  - Generalization
  - Mining Frequent Patterns, Associations, and Correlations
  - Classification
  - Cluster Analysis
  - Outlier Analysis
- Data mining: On What Kinds of Data?
- Time and Ordering: Sequential Pattern, Trend and Evolution Analysis
- Structure and Network Analysis
- Evaluation of knowledge
- Applications of Data Mining
- Major Challenges in Data Mining
- A Brief History of Data Mining and Data Mining Society

Structure and Network Analysis

- Graph mining
  - Finding frequent subgraphs (e.g., chemical compounds), trees (XML), substructures (web fragments)
- Information network analysis
  - Social networks: actors (objects, nodes) and relationships (edges)
    - e.g., author networks in CS, terrorist networks
  - Multiple heterogeneous networks
    - A person could be multiple information networks: friends, family, classmates, …
  - Links carry a lot of semantic information: Link mining
- Web mining
  - Web is a big information network: from PageRank to Google
  - Analysis of Web information networks
    - Web community discovery, opinion mining, usage mining, …

Introduction to Data Mining - Sub-Topics

- Why Data Mining?
  - Data Mining: A Natural Evolution of Science and Technology
- What Is Data Mining?
  - Data Mining: Essential in a Knowledge Discovery Process
  - Data Mining: A Confluence of Multiple Disciplines
- A Multi-Dimensional View of Data Mining
  - Knowledge to Be Mined
  - Data to Be Mined
  - Technology Utilized
  - Applications Adapted
- Data Mining Functionalities: What Kinds of Patterns Can Be Mined?
  - Generalization
  - Mining Frequent Patterns, Associations, and Correlations
  - Classification
  - Cluster Analysis
  - Outlier Analysis
- Data mining: On What Kinds of Data?
- Time and Ordering: Sequential Pattern, Trend and Evolution Analysis
- Structure and Network Analysis
- Evaluation of knowledge
- Applications of Data Mining
- Major Challenges in Data Mining
- A Brief History of Data Mining and Data Mining Society

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
  - …

Introduction to Data Mining - Sub-Topics

- Why Data Mining?
  - » Data Mining: A Natural Evolution of Science and Technology
- What Is Data Mining?
  - » Data Mining: Essential in a Knowledge Discovery Process
  - » Data Mining: A Confluence of Multiple Disciplines
- A Multi-Dimensional View of Data Mining
  - » Knowledge to Be Mined
  - » Data to Be Mined
  - » Technology Utilized
  - » Applications Adapted
- Data Mining Functionalities: What Kinds of Patterns Can Be Mined?
  - » Generalization
  - » Mining Frequent Patterns, Associations, and Correlations
  - » Classification
  - » Cluster Analysis
  - » Outlier Analysis
- Data mining: On What Kinds of Data?
- Time and Ordering: Sequential Pattern, Trend and Evolution Analysis
- Structure and Network Analysis
- Evaluation of knowledge
- Applications of Data Mining
- Major Challenges in Data Mining
- A Brief History of Data Mining and Data Mining Society

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

Introduction to Data Mining - Sub-Topics

- Why Data Mining?
  - » Data Mining: A Natural Evolution of Science and Technology
- What Is Data Mining?
  - » Data Mining: Essential in a Knowledge Discovery Process
  - » Data Mining: A Confluence of Multiple Disciplines
- A Multi-Dimensional View of Data Mining
  - » Knowledge to Be Mined
  - » Data to Be Mined
  - » Technology Utilized
  - » Applications Adapted
- Data Mining Functionalities: What Kinds of Patterns Can Be Mined?
  - » Generalization
  - » Mining Frequent Patterns, Associations, and Correlations
  - » Classification
  - » Cluster Analysis
  - » Outlier Analysis
- Data mining: On What Kinds of Data?
- Time and Ordering: Sequential Pattern, Trend and Evolution Analysis
- Structure and Network Analysis
- Evaluation of knowledge
- Applications of Data Mining
- ➡ Major Challenges in Data Mining – Additional Topics
- A Brief History of Data Mining and Data Mining Society

Major Challenges in Data Mining

- Efficiency and scalability of data mining algorithms
- Parallel, distributed, stream, and incremental mining methods
- Handling high-dimensionality
- Handling noise, uncertainty, and incompleteness of data
- Incorporation of constraints, expert knowledge, and background knowledge in data mining
- Pattern evaluation and knowledge integration
- Mining diverse and heterogeneous kinds of data: e.g., bioinformatics, Web, software/system engineering, information networks
- Application-oriented and domain-specific data mining
- Invisible data mining (embedded in other functional modules)
- Protection of security, integrity, and privacy in data mining

Focus Areas 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

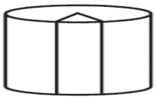
Why Data Mining Query Language?

- 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 (1/2)

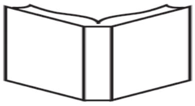
- Task-relevant data
  - Database or data warehouse name
  - Database tables or data warehouse cubes
  - Condition for data selection
  - Relevant attributes or dimensions
  - Data grouping criteria
- Type of knowledge to be mined
  - Characterization, discrimination, association, classification, prediction,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

Primitives that Define a Data Mining Task (2/2)



Task-relevant data
Database or data warehouse name
Database tables or data warehouse cubes
Conditions for data selection
Relevant attributes or dimensions
Data grouping criteria

Knowledge type to be mined
Characterization
Discrimination
Association/correlation
Classification/prediction
Clustering



Background knowledge
Concept hierarchies
User beliefs about relationships in the data

Pattern interestingness measures
Simplicity
Certainty (e.g., confidence)
Utility (e.g., support)
Novelty

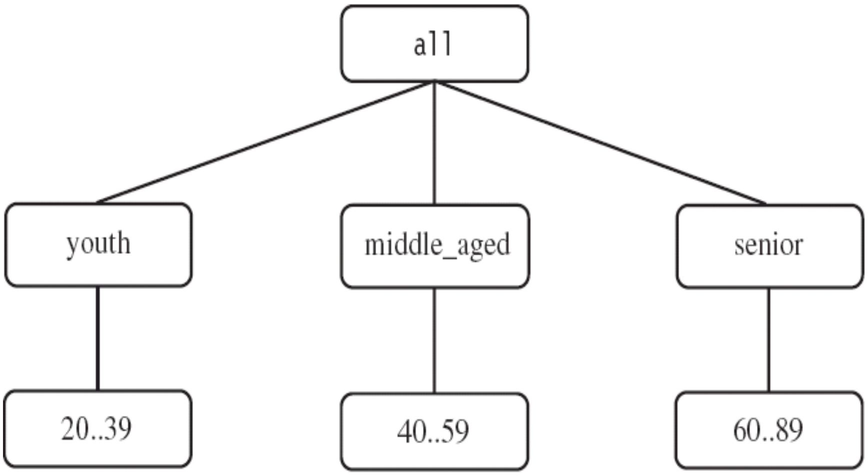
Visualization of discovered patterns
Rules, tables, reports, charts, graphs, decision trees, and cubes
Drill-down and roll-up

Primitive 3: Background Knowledge (1/2)

- 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: xyz@cs.nyu.edu

    login-name < department < university < country
- Rule-based hierarchy
  - » low_profit_margin (X) <= price(X, P_1) and cost (X, P_2) and (P_1 - P_2) < $50

Primitive 3: Background Knowledge (2/2)



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)

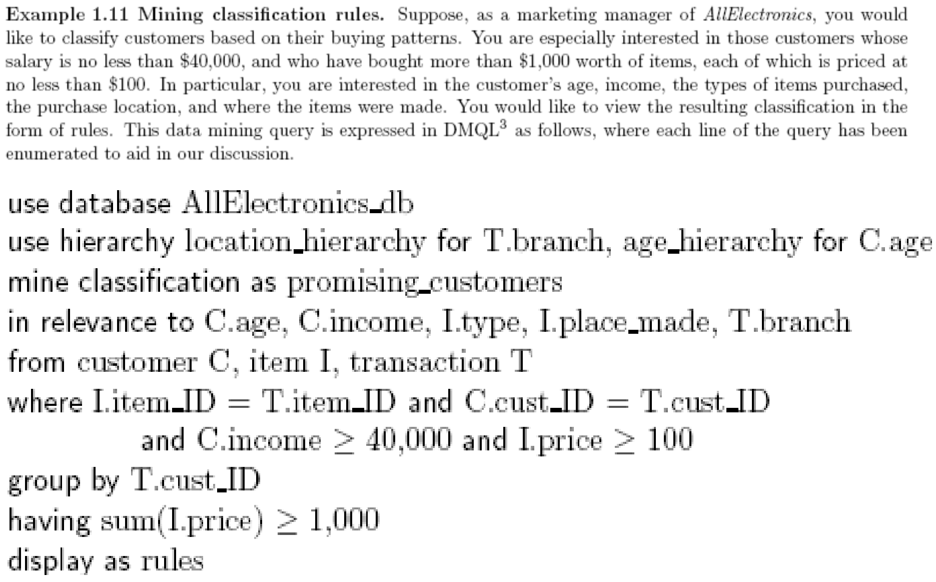
Primitive 5: Presentation of Discovered Patterns

- 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.

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 DMQL[3] 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 ≥ 40,000 and I.price ≥ 100
group by T.cust_ID
having sum(I.price) ≥ 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

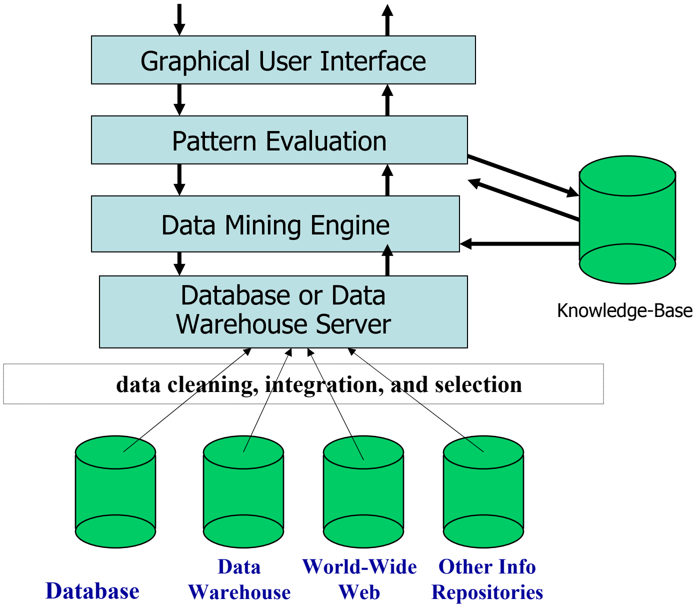
Integration of Data Mining and Data Warehousing

- **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



Graphical User Interface

Pattern Evaluation

Data Mining Engine

Database or Data Warehouse Server

Knowledge-Base

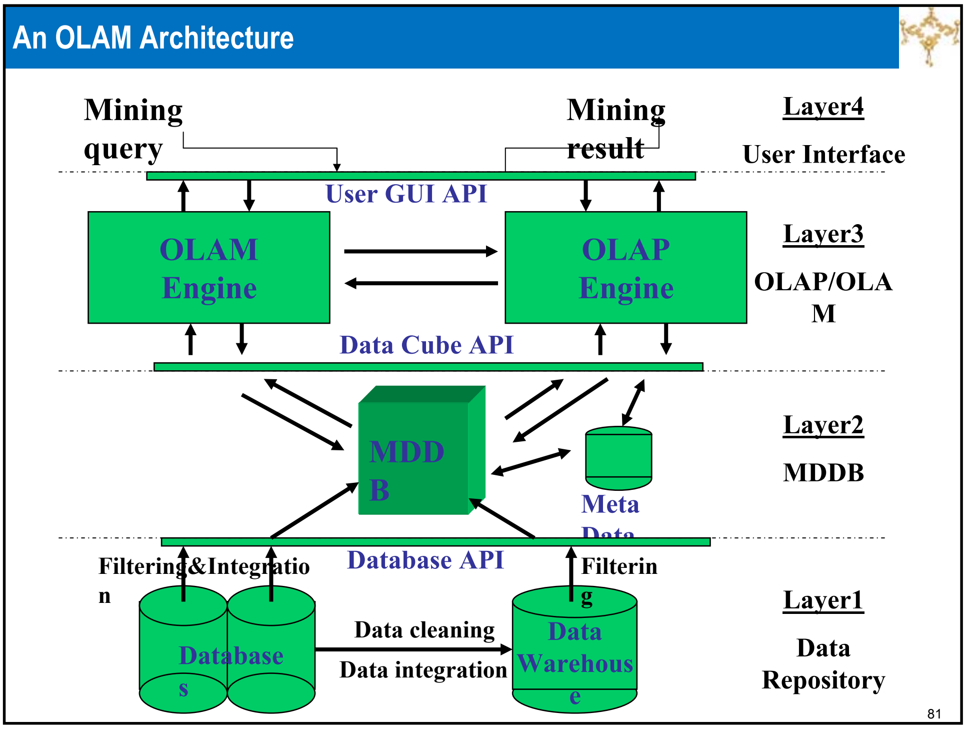
data cleaning, integration, and selection

Database

Data Warehouse

World-Wide Web

Other Info Repositories



An OLAM Architecture

Mining query

Mining result

Layer4

User Interface

User GUI API

OLAM Engine

OLAP Engine

Layer3

OLAP/OLAM

Data Cube API

MDDB

Meta Data

Layer2

MDDB

Database API

Filtering&Integration

Filtering

Databases

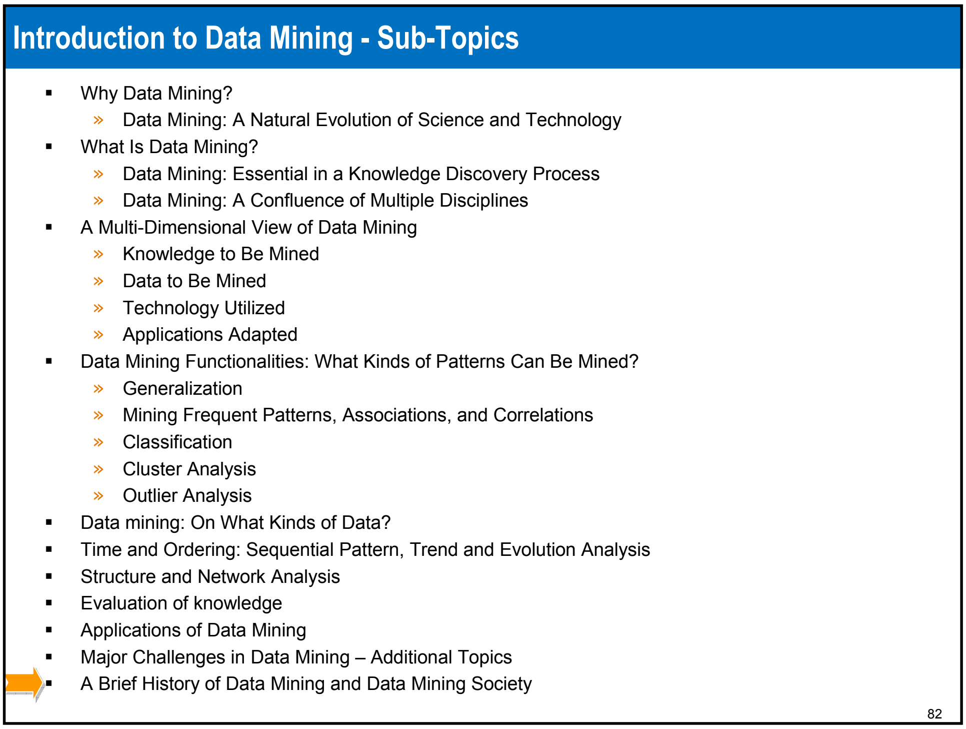
Data cleaning

Data integration

Data Warehouse

Layer1

Data Repository



Introduction to Data Mining - Sub-Topics

- Why Data Mining?
  - Data Mining: A Natural Evolution of Science and Technology
- What Is Data Mining?
  - Data Mining: Essential in a Knowledge Discovery Process
  - Data Mining: A Confluence of Multiple Disciplines
- A Multi-Dimensional View of Data Mining
  - Knowledge to Be Mined
  - Data to Be Mined
  - Technology Utilized
  - Applications Adapted
- Data Mining Functionalities: What Kinds of Patterns Can Be Mined?
  - Generalization
  - Mining Frequent Patterns, Associations, and Correlations
  - Classification
  - Cluster Analysis
  - Outlier Analysis
- Data mining: On What Kinds of Data?
- Time and Ordering: Sequential Pattern, Trend and Evolution Analysis
- Structure and Network Analysis
- Evaluation of knowledge
- Applications of Data Mining
- Major Challenges in Data Mining – Additional Topics
- ➡ A Brief History of Data Mining and Data Mining Society

A Brief History of Data Mining Society

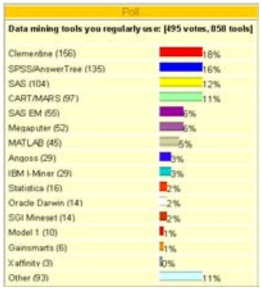
- 1989 IJCAI Workshop on Knowledge Discovery in Databases
  - » Knowledge Discovery in Databases (G. Piatetsky-Shapiro and W. Frawley, 1991)
- 1991-1994 Workshops on Knowledge Discovery in Databases
  - » Advances in Knowledge Discovery and Data Mining (U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and 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

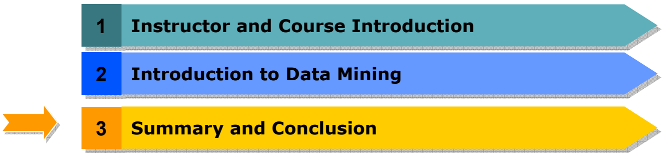
- Data mining and KDD (SIGKDD: CDROM)
  - Conferences: ACM-SIGKDD, IEEE-ICDM, SIAM-DM, PKDD, PAKDD, etc.
  - Journal: Data Mining and Knowledge Discovery, KDD Explorations, ACM TKDD
- Database systems (SIGMOD: ACM SIGMOD Anthology—CD ROM)
  - Conferences: ACM-SIGMOD, ACM-PODS, VLDB, IEEE-ICDE, EDBT,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

Agenda



1 Instructor and Course Introduction

2 Introduction to Data Mining

3 Summary and Conclusion

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
- Focus areas in data mining

Assignments & Readings

- Readings

  - Foreword/Preface and Chapter 1
- Assignment #1
  - Textbook Exercises 1.5, 1.7, 1.10, 1.11, 1.12, 1.15

Next Session: Data Preprocessing