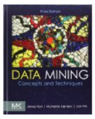
Veri Madenciliği - Giriş

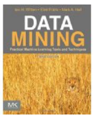
Şadi Evren ŞEKER

Syllabus – Ders İzlencesi

- www.SadiEvrenSEKER.com -> Courses -> Data Mining, Istanbul Commerce University
- http://sadievrenseker.com/wp/?p=558
- Slide'lar : http://web.engr.illinois.edu/~hanj/bk3/bk3_slidesindex.htm

Kaynaklar





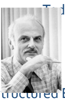
- Data Mining: Concepts and Techniques, Third Edition, Jiawei Han, Micheline Kamber, Jian Pei
- Data Mining: Practical Machine Learning Tools and Techniques, Third...Ian H. Witten, Eibe Frank, Mark A. Hall

Necessity Is the Mother of Invention

- Data explosion problem
  - Automated data collection tools, widely used database systems, computerized society, and the Internet lead to tremendous amounts of data accumulated and/or to be analyzed in databases, data warehouses, WWW, and other information repositories
- We are drowning in data, but starving for knowledge!
- Solution: Data warehousing and data mining
  - Data warehousing and on-line analytical processing (**OLAP**)
  - Mining interesting knowledge (rules, regularities, patterns, constraints) from data in large databases

Evolution of Database Technology

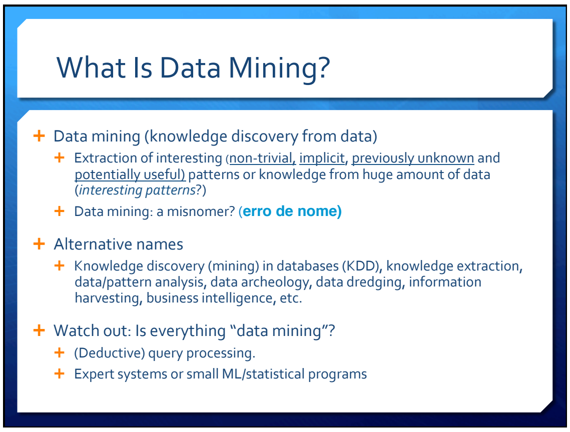
- 1960s:
  - Data collection, database creation, IMS and network DBMS
- 1970s:
  - **Relational data model**, relational DBMS implementation
  - Ted Codd (1923-2003)



  - Structured English Query Language (SEQUEL), **SQL**
- 1980s:
  - Advanced data models (extended-relational, OO, deductive, etc.)
  - Application-oriented DBMS (spatial, scientific, engineering, etc.)

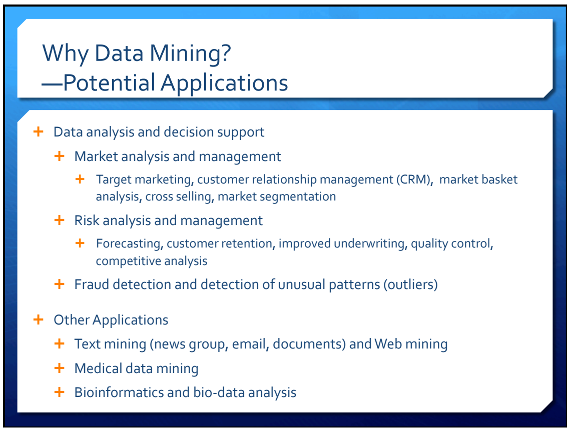
Evolution of Database Technology

- 1990s:
  - Data mining, data warehousing, multimedia databases
  - **Web databases (..,Amazon)**
- 2000s
  - Stream data management and mining
  - Data mining and its applications
  - Web technology (XML, data integration) and global information systems



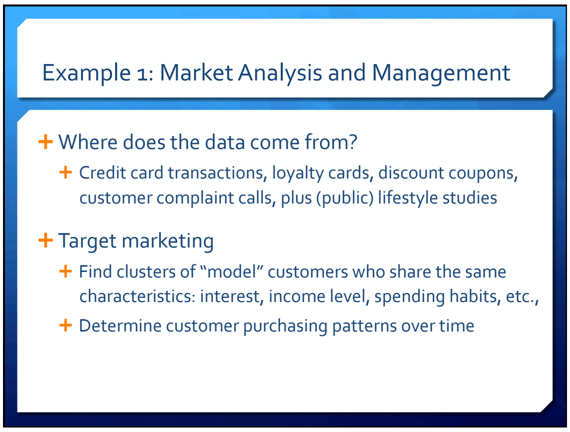
What Is Data Mining?

- 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 (*interesting patterns?*)
  - Data mining: a misnomer? **(erro de nome)**
- 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"?
  - (Deductive) query processing.
  - Expert systems or small ML/statistical programs



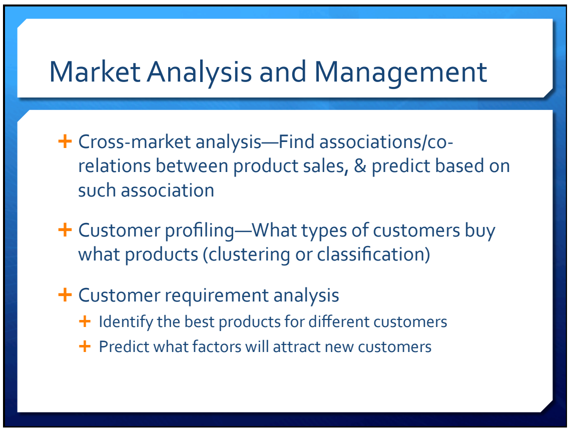
Why Data Mining?
—Potential Applications

- 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
  - Medical data mining
  - Bioinformatics and bio-data analysis



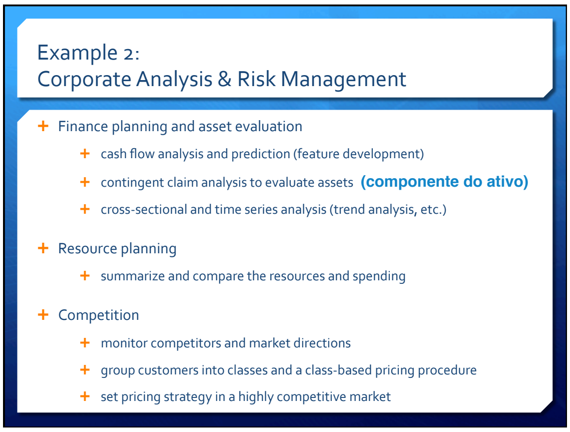
Example 1: Market Analysis and Management

- 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



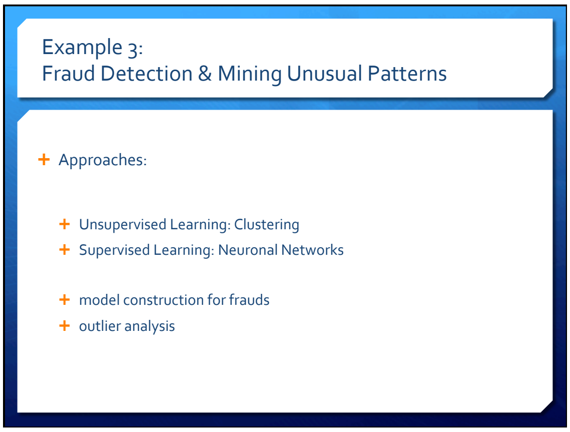
Market Analysis and Management

- 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



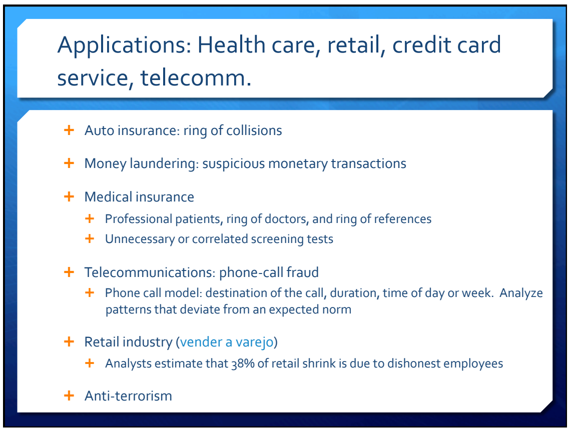
Example 2:
Corporate Analysis & Risk Management

- Finance planning and asset evaluation
  - cash flow analysis and prediction (feature development)
  - contingent claim analysis to evaluate assets **(componente do ativo)**
  - cross-sectional and time series analysis (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



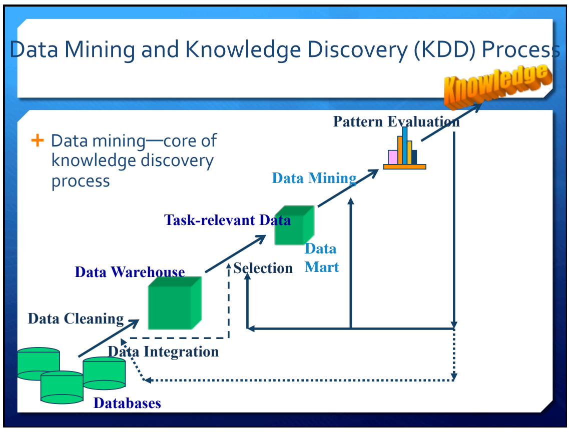
Example 3:
Fraud Detection & Mining Unusual Patterns

- Approaches:
  - Unsupervised Learning: Clustering
  - Supervised Learning: Neuronal Networks
  - model construction for frauds
  - outlier analysis



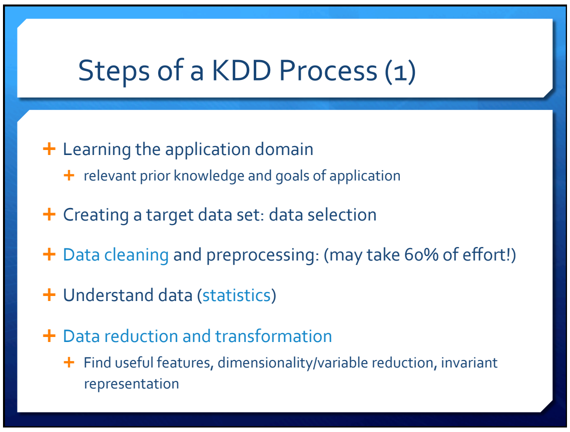
Applications: Health care, retail, credit card service, telecomm.

- 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 (vender a varejo)
  - Analysts estimate that 38% of retail shrink is due to dishonest employees
- Anti-terrorism



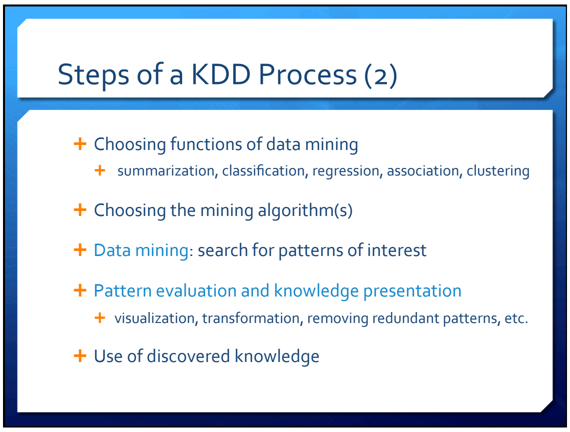
Data Mining and Knowledge Discovery (KDD) Process

- Data mining—core of knowledge discovery process



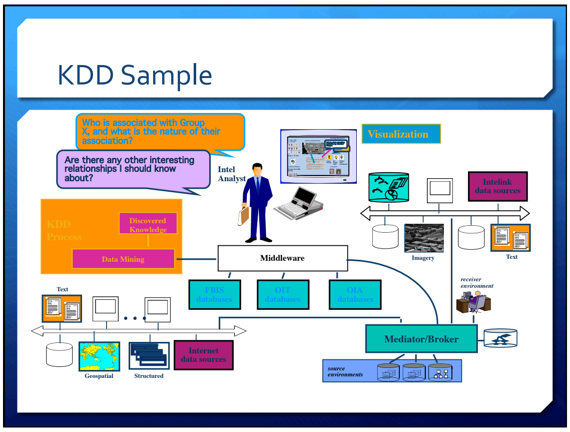
Steps of a KDD Process (1)

- 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!)
- Understand data (statistics)
- Data reduction and transformation
  - Find useful features, dimensionality/variable reduction, invariant representation

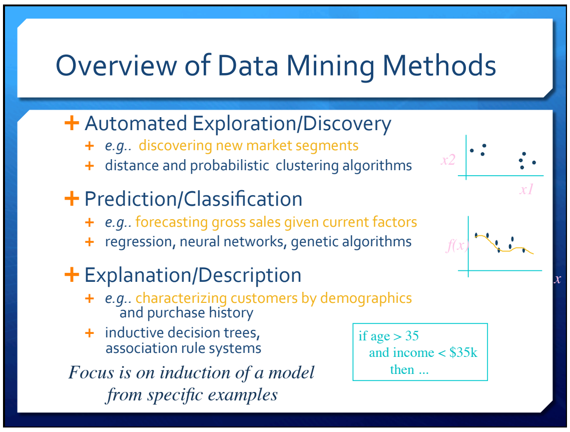


Steps of a KDD Process (2)

- 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



KDD Sample

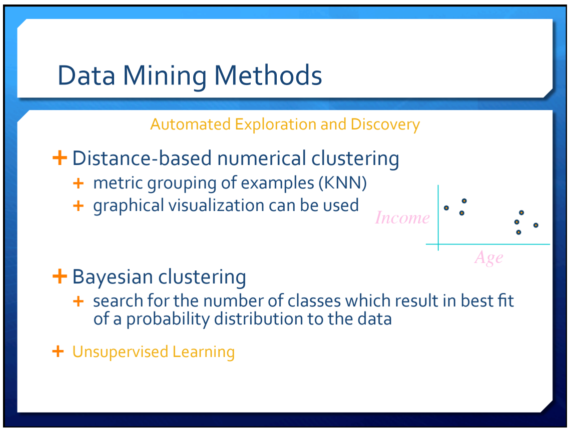


Overview of Data Mining Methods

- Automated Exploration/Discovery
  - *e.g.*.. discovering new market segments
  - distance and probabilistic clustering algorithms
- Prediction/Classification
  - *e.g.*.. forecasting gross sales given current factors
  - regression, neural networks, genetic algorithms
- Explanation/Description
  - *e.g.*.. characterizing customers by demographics and purchase history
  - inductive decision trees, association rule systems

Focus is on induction of a model from specific examples

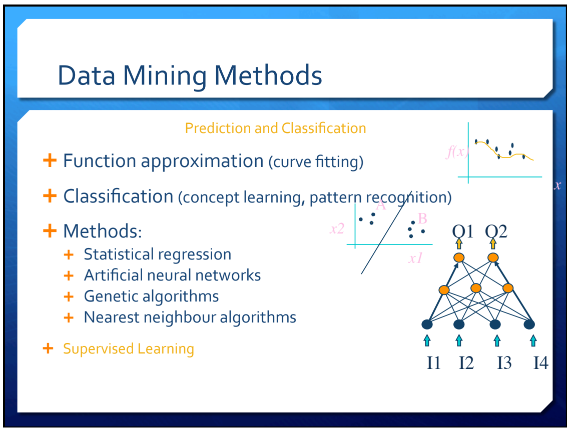
if age > 35
and income < $35k
then ...



Data Mining Methods

Automated Exploration and Discovery

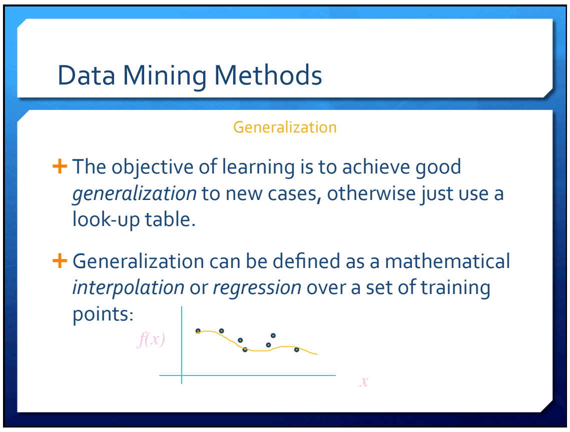
- Distance-based numerical clustering
  - metric grouping of examples (KNN)
  - graphical visualization can be used
- Bayesian clustering
  - search for the number of classes which result in best fit of a probability distribution to the data
- Unsupervised Learning



Data Mining Methods

Prediction and Classification

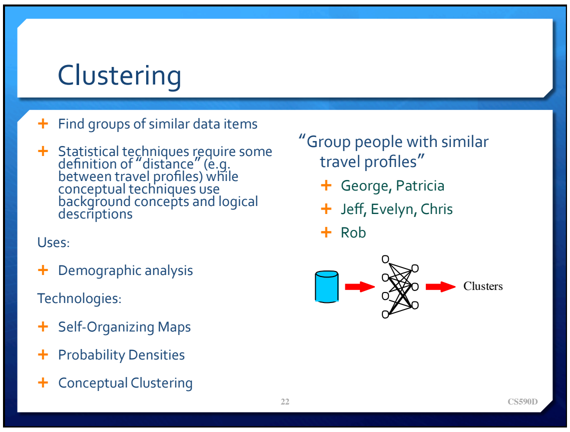
- Function approximation (curve fitting)
- Classification (concept learning, pattern recognition)
- Methods:
  - Statistical regression
  - Artificial neural networks
  - Genetic algorithms
  - Nearest neighbour algorithms
- Supervised Learning



Data Mining Methods

Generalization

- The objective of learning is to achieve good *generalization* to new cases, otherwise just use a look-up table.
- Generalization can be defined as a mathematical *interpolation* or *regression* over a set of training points:



Clustering

- Find groups of similar data items
- Statistical techniques require some definition of "distance" (e.g. between travel profiles) while conceptual techniques use background concepts and logical descriptions

Uses:

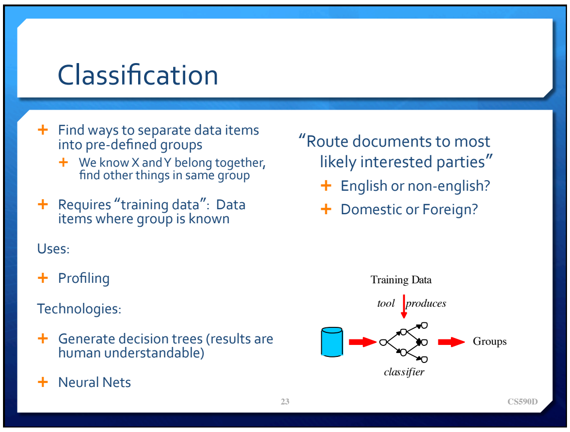
- Demographic analysis

Technologies:

- Self-Organizing Maps
- Probability Densities
- Conceptual Clustering

"Group people with similar travel profiles"

- George, Patricia
- Jeff, Evelyn, Chris
- Rob



Classification

- Find ways to separate data items into pre-defined groups
  - We know X and Y belong together, find other things in same group
- Requires "training data": Data items where group is known

Uses:

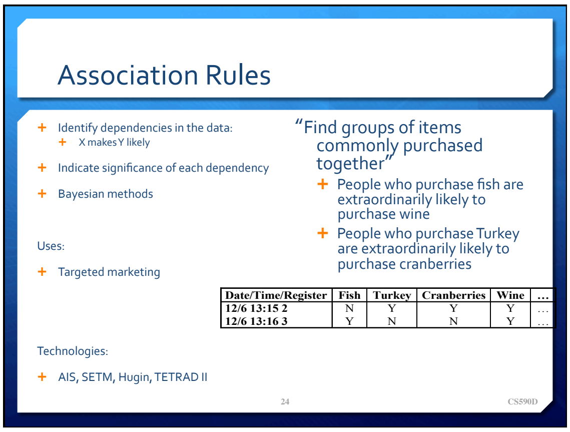
- Profiling

Technologies:

- Generate decision trees (results are human understandable)
- Neural Nets

"Route documents to most likely interested parties"

- English or non-english?
- Domestic or Foreign?



Association Rules

- Identify dependencies in the data:
  - X makes Y likely
- Indicate significance of each dependency
- Bayesian methods

Uses:

- Targeted marketing

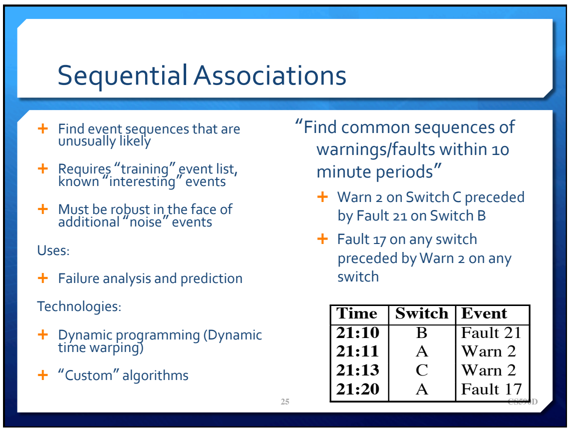
Technologies:

- AIS, SETM, Hugin, TETRAD II

"Find groups of items commonly purchased together"

- People who purchase fish are extraordinarily likely to purchase wine
- People who purchase Turkey are extraordinarily likely to purchase cranberries

Date/Time/Register	Fish	Turkey	Cranberries	Wine	…
12/6 13:15 2	N	Y	Y	Y	…
12/6 13:16 3	Y	N	N	Y	…



Sequential Associations

- Find event sequences that are unusually likely
- Requires "training" event list, known "interesting" events
- Must be robust in the face of additional "noise" events

Uses:

- Failure analysis and prediction

Technologies:

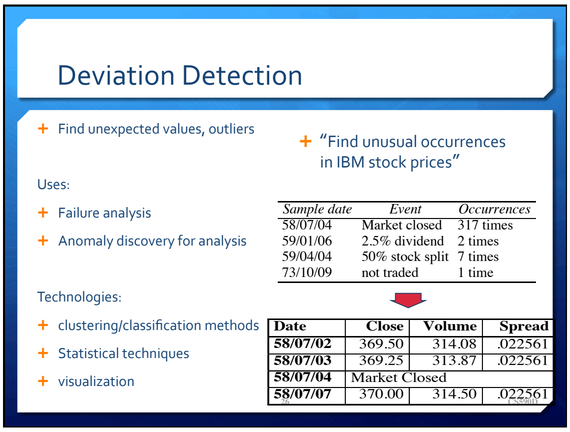
- Dynamic programming (Dynamic time warping)
- "Custom" algorithms

"Find common sequences of warnings/faults within 10 minute periods"

- Warn 2 on Switch C preceded by Fault 21 on Switch B
- Fault 17 on any switch preceded by Warn 2 on any switch

Time	Switch	Event
21:10	B	Fault 21
21:11	A	Warn 2
21:13	C	Warn 2
21:20	A	Fault 17

25 CS590D



Deviation Detection

- Find unexpected values, outliers

Uses:

- Failure analysis
- Anomaly discovery for analysis

Technologies:

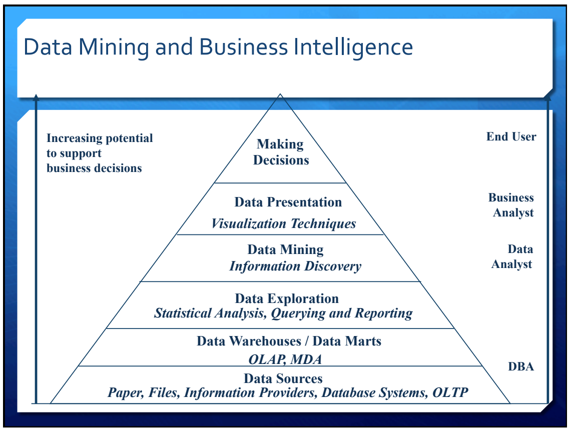
- clustering/classification methods
- Statistical techniques
- visualization

- "Find unusual occurrences in IBM stock prices"

Sample date	*Event*	*Occurrences*
58/07/04	Market closed	317 times
59/01/06	2.5% dividend	2 times
59/04/04	50% stock split	7 times
73/10/09	not traded	1 time

Date	Close	Volume	Spread
58/07/02	369.50	314.08	.022561
58/07/03	369.25	313.87	.022561
58/07/04	Market Closed		
58/07/07	370.00	314.50	.022561

26 CS590D



Data Mining and Business Intelligence

Increasing potential to support business decisions

Making Decisions — End User

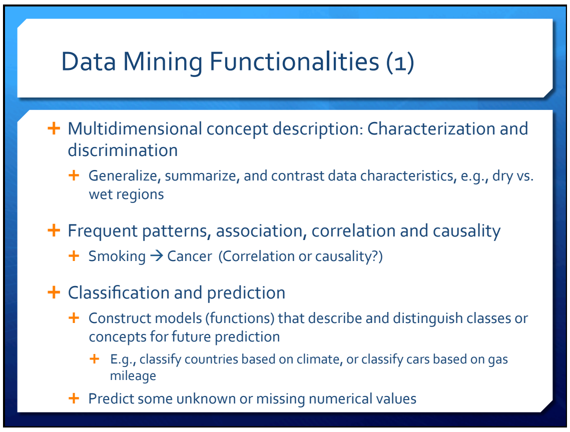
Data Presentation *Visualization Techniques* — Business Analyst

Data Mining *Information Discovery* — Data Analyst

Data Exploration *Statistical Analysis, Querying and Reporting*

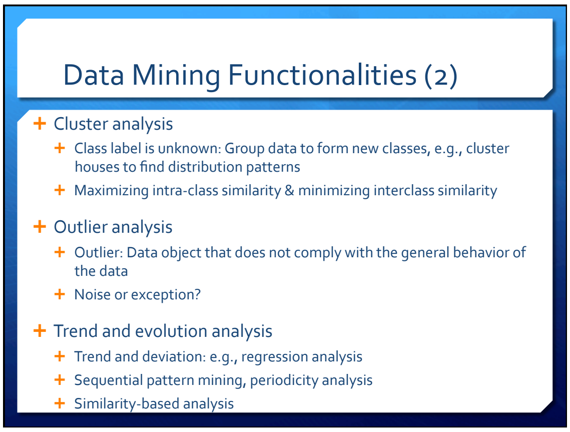
Data Warehouses / Data Marts *OLAP, MDA* — DBA

Data Sources *Paper, Files, Information Providers, Database Systems, OLTP*



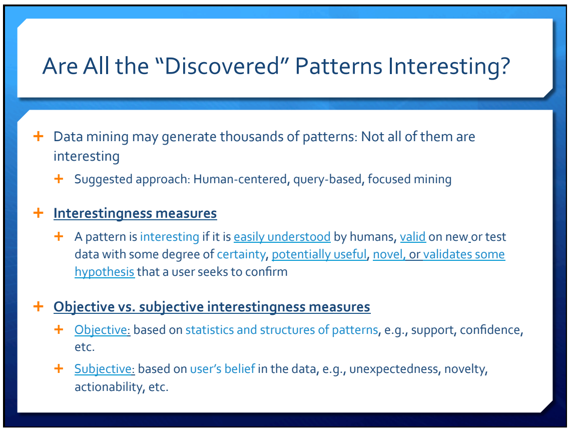
Data Mining Functionalities (1)

- Multidimensional concept description: Characterization and discrimination
  - Generalize, summarize, and contrast data characteristics, e.g., dry vs. wet regions
- Frequent patterns, association, correlation and causality
  - Smoking → Cancer (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



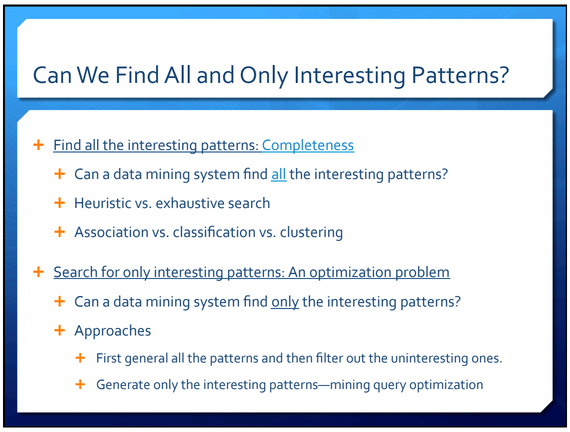
Data Mining Functionalities (2)

- 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?
- Trend and evolution analysis
  - Trend and deviation: e.g., regression analysis
  - Sequential pattern mining, periodicity analysis
  - Similarity-based analysis



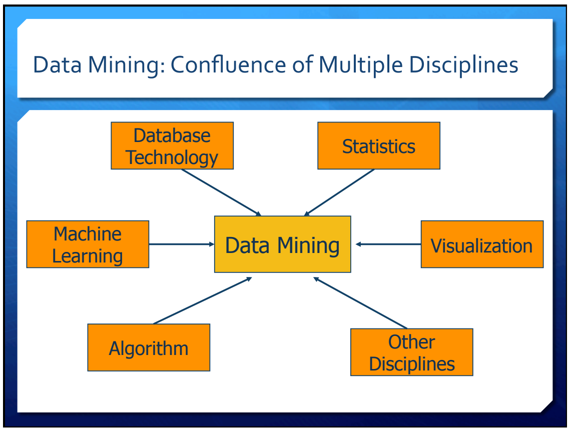
Are All the "Discovered" Patterns Interesting?

- 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.

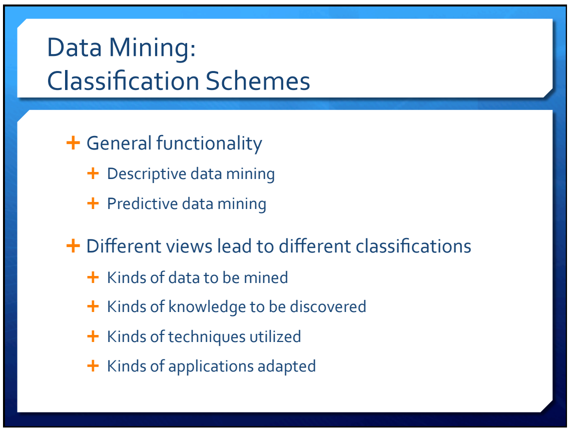


Can We Find All and Only Interesting Patterns?

- Find all the interesting patterns: Completeness
  - Can a data mining system find all 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

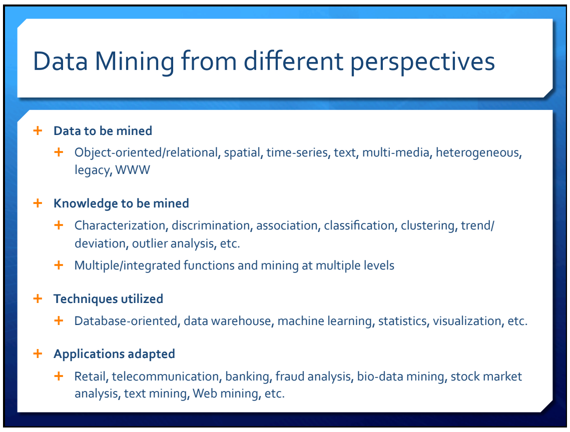


Data Mining: Confluence of Multiple Disciplines



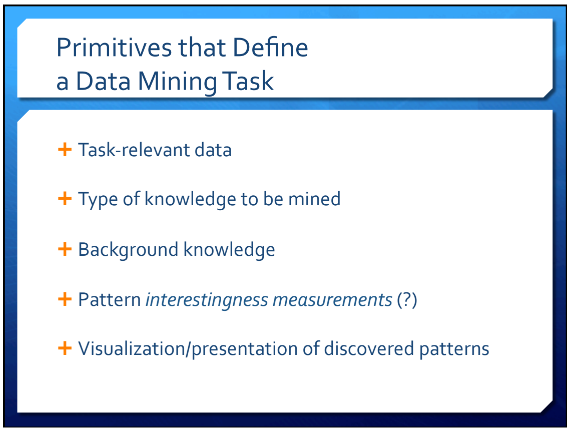
Data Mining: Classification Schemes

- General functionality
  - Descriptive data mining
  - Predictive data mining
- Different views lead to different classifications
  - Kinds of data to be mined
  - Kinds of knowledge to be discovered
  - Kinds of techniques utilized
  - Kinds of applications adapted



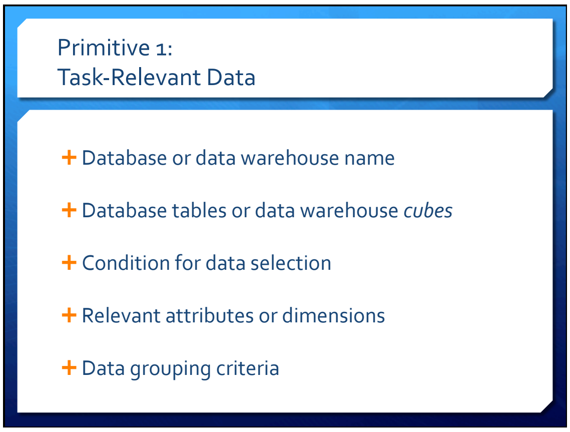
Data Mining from different perspectives

- **Data to be mined**
  - Object-oriented/relational, 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, machine learning, statistics, visualization, etc.
- **Applications adapted**
  - Retail, telecommunication, banking, fraud analysis, bio-data mining, stock market analysis, text mining, Web mining, etc.



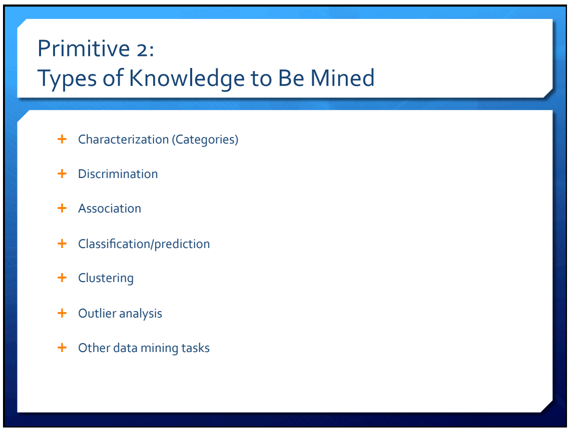
Primitives that Define a Data Mining Task

- Task-relevant data
- Type of knowledge to be mined
- Background knowledge
- Pattern *interestingness measurements* (?)
- Visualization/presentation of discovered patterns



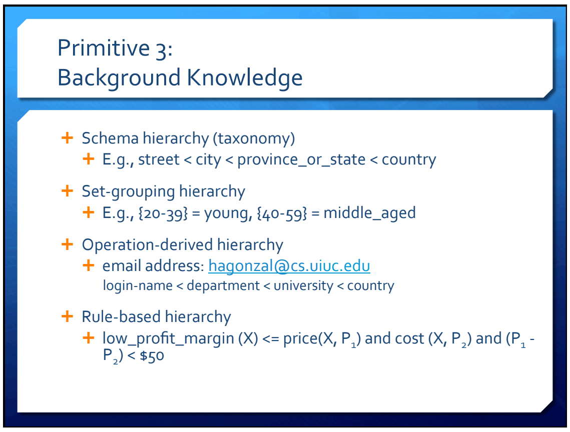
Primitive 1: Task-Relevant Data

- Database or data warehouse name
- Database tables or data warehouse *cubes*
- Condition for data selection
- Relevant attributes or dimensions
- Data grouping criteria



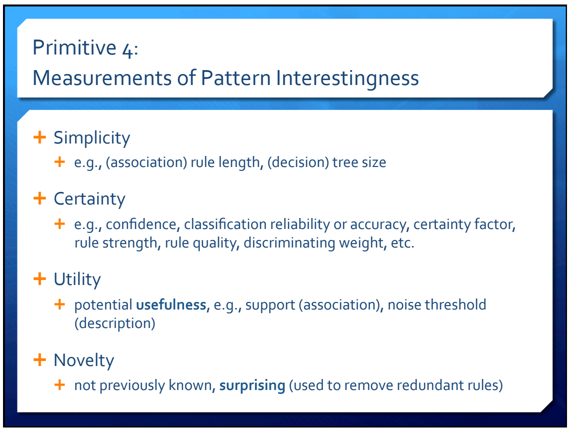
Primitive 2: Types of Knowledge to Be Mined

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



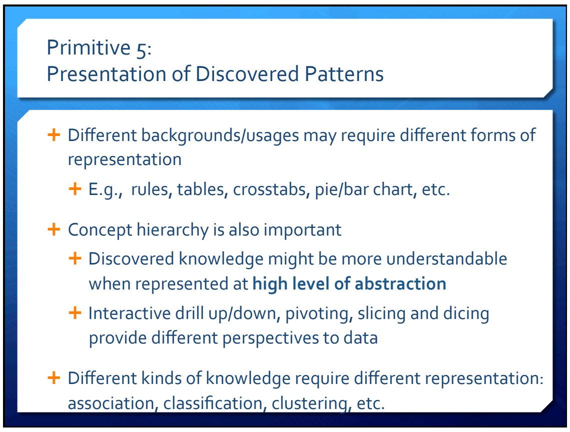
Primitive 3: Background Knowledge

- Schema hierarchy (taxonomy)
  - 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, P_1) and cost (X, P_2) and (P_1 - P_2) < $50



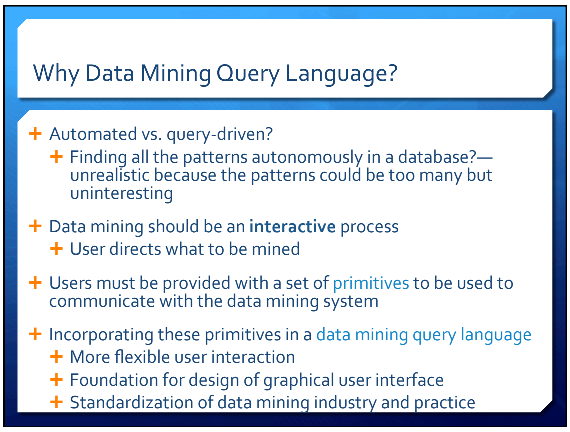
Primitive 4: Measurements of Pattern Interestingness

- Simplicity
  - e.g., (association) rule length, (decision) tree size
- Certainty
  - e.g., confidence, 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)



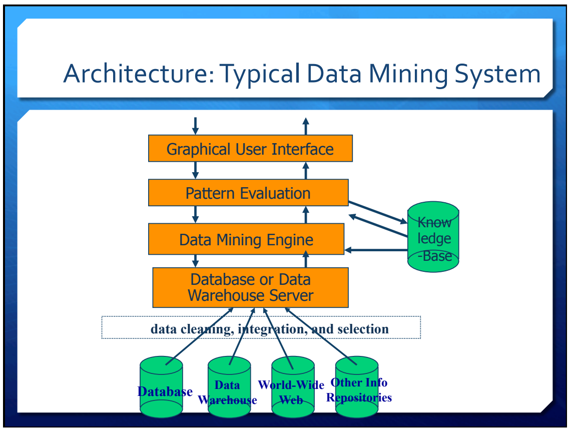
Primitive 5: Presentation of Discovered Patterns

- 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.

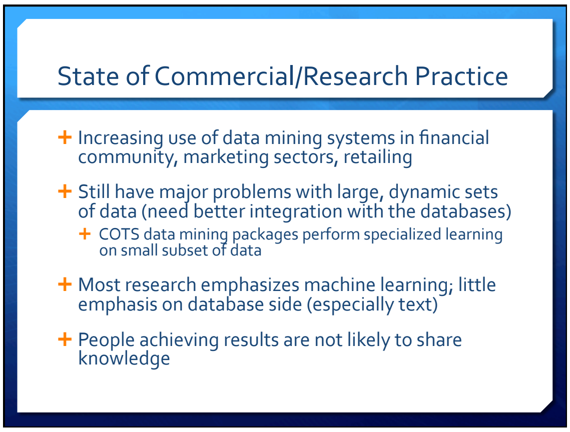


Why Data Mining Query Language?

- 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

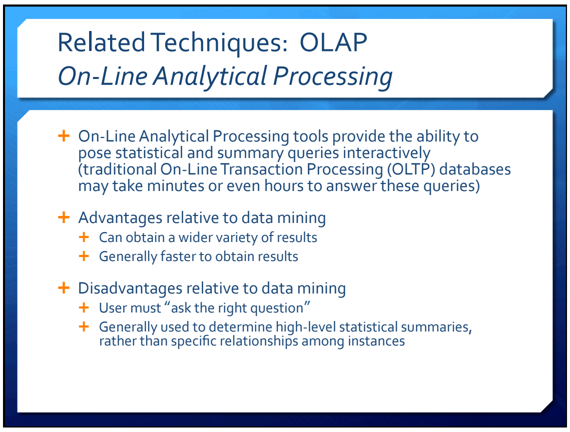


Architecture: Typical Data Mining System



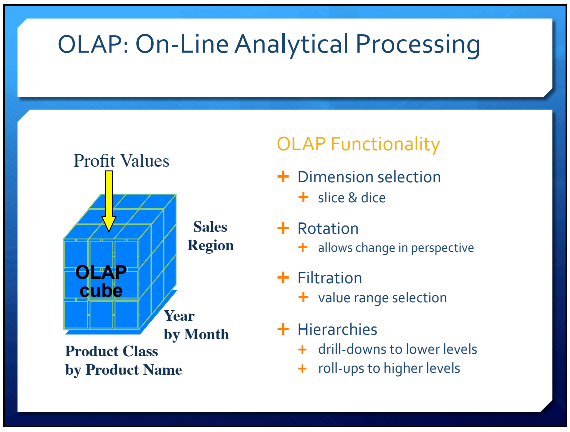
State of Commercial/Research Practice

- Increasing use of data mining systems in financial community, marketing sectors, retailing
- Still have major problems with large, dynamic sets of data (need better integration with the databases)
  - COTS data mining packages perform specialized learning on small subset of data
- Most research emphasizes machine learning; little emphasis on database side (especially text)
- People achieving results are not likely to share knowledge



Related Techniques: OLAP *On-Line Analytical Processing*

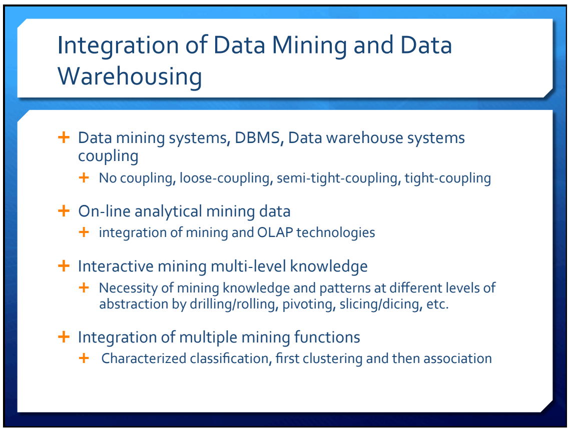
- On-Line Analytical Processing tools provide the ability to pose statistical and summary queries interactively (traditional On-Line Transaction Processing (OLTP) databases may take minutes or even hours to answer these queries)
- Advantages relative to data mining
  - Can obtain a wider variety of results
  - Generally faster to obtain results
- Disadvantages relative to data mining
  - User must "ask the right question"
  - Generally used to determine high-level statistical summaries, rather than specific relationships among instances



OLAP: On-Line Analytical Processing

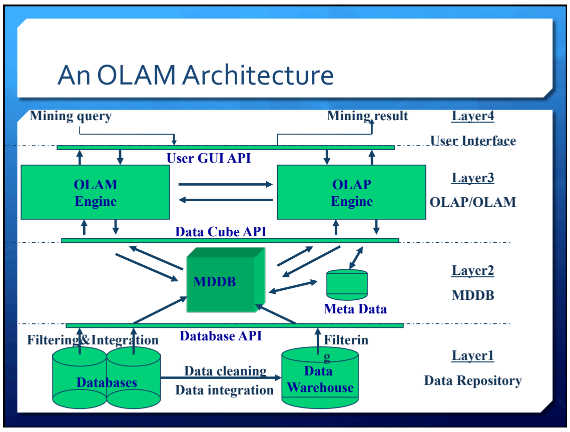
OLAP Functionality

- Dimension selection
  - slice & dice
- Rotation
  - allows change in perspective
- Filtration
  - value range selection
- Hierarchies
  - drill-downs to lower levels
  - roll-ups to higher levels

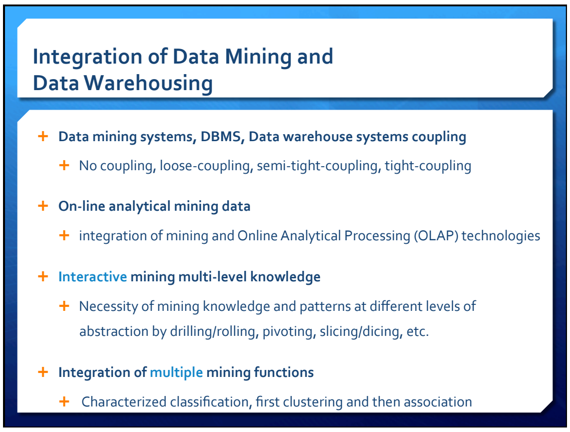


Integration of Data Mining and Data Warehousing

- 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

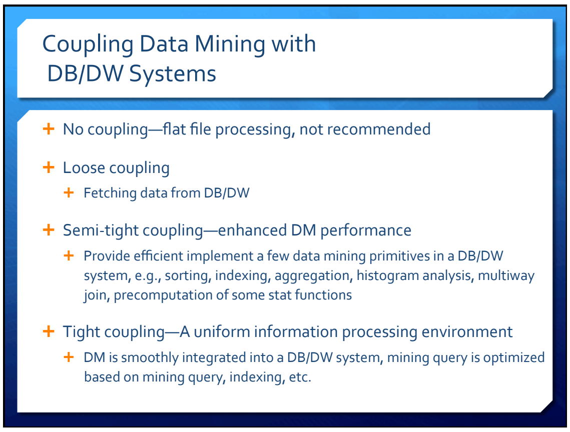


An OLAM Architecture



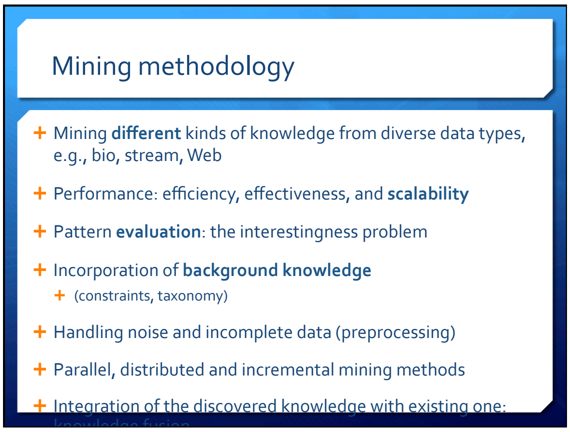
Integration of Data Mining and Data Warehousing

- **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 Online Analytical Processing (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



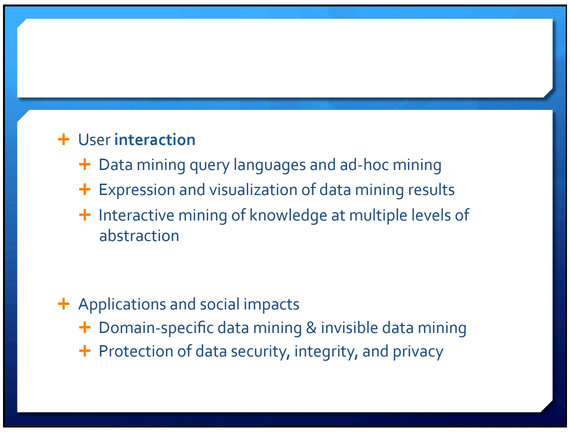
Coupling Data Mining with DB/DW Systems

- 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, etc.

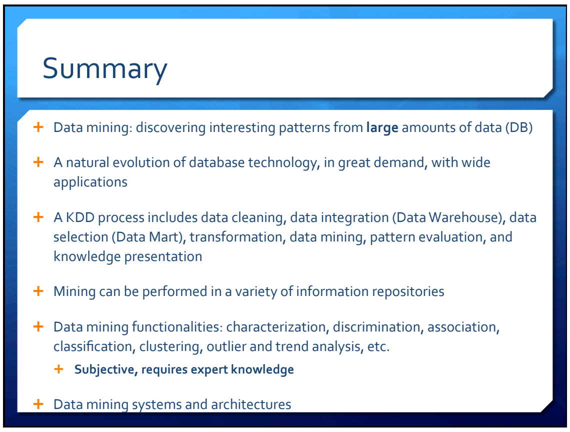


Mining methodology

- 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**
  - (constraints, taxonomy)
- Handling noise and incomplete data (preprocessing)
- Parallel, distributed and incremental mining methods
- Integration of the discovered knowledge with existing one: knowledge fusion

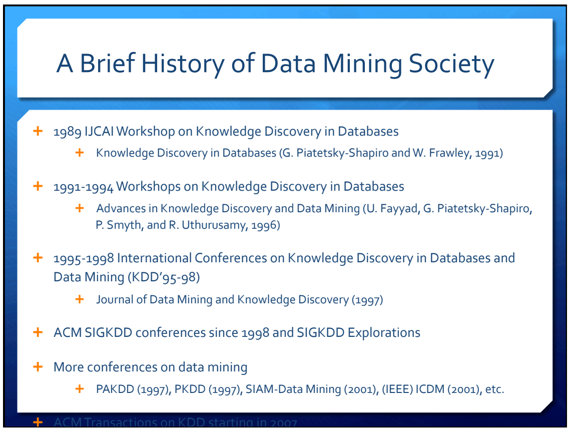


- User **interaction**
  - Data mining query languages and ad-hoc mining
  - Expression and visualization of data mining results
  - Interactive mining of knowledge at multiple levels of abstraction
- Applications and social impacts
  - Domain-specific data mining & invisible data mining
  - Protection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



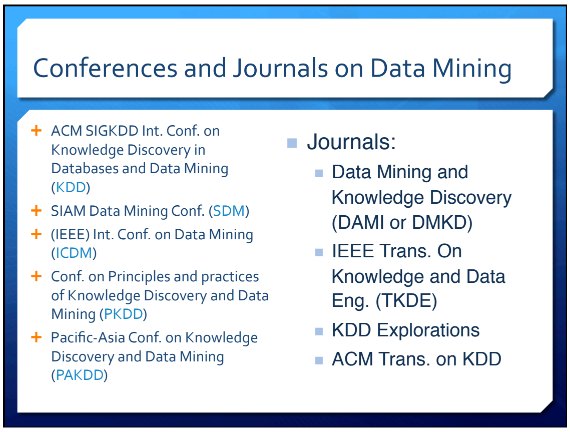
Summary

- Data mining: discovering interesting patterns from **large** amounts of data (DB)
- A natural evolution of database technology, in great demand, with wide applications
- A KDD process includes data cleaning, data integration (Data Warehouse), data selection (Data Mart), 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.
  - **Subjective, requires expert knowledge**
- Data mining systems and architectures



A Brief History of Data Mining Society

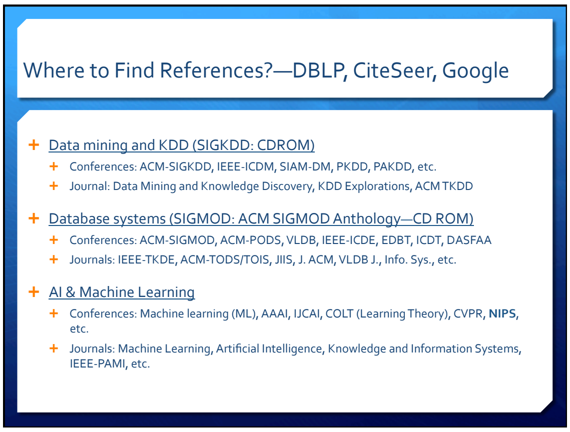
- 1989 IJCAI Workshop on Knowledge Discovery in Databases
  - Knowledge Discovery in Databases (G. Piatetsky-Shapiro and W. Frawley, 1991)
- 1991-1994 Workshops on Knowledge Discovery in Databases
  - Advances in Knowledge Discovery and Data Mining (U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy,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

- 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)

- 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.

Data Warehousing and OLAP Technology

www.SadiEvrenSEKER.com