PURDUE
UNIVERSITY

CS490D: Introduction to Data Mining

Chris Clifton

January 12, 2004

Course Overview

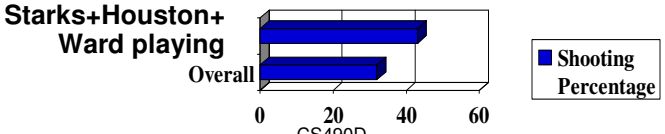
Indiana Center for Database Systems

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"?
  - (Deductive) query processing.
  - Expert systems or small ML/statistical programs

What is Data Mining?
Real Example from the NBA

- Play-by-play information recorded by teams
  - Who is on the court
  - Who shoots
  - Results
- Coaches want to know what works best
  - Plays that work well against a given team
  - Good/bad player matchups
- Advanced Scout (from IBM Research) is a data mining tool to answer these questions



http://www.nba.com/news_feat/beyond/0126.html

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
  - DNA and bio-data analysis

Course Outline

www.cs.purdue.edu/~clifton/cs490d

1. Introduction: What is data mining?
   - What makes it a new and unique discipline?
   - Relationship between Data Warehousing, On-line Analytical Processing, and Data Mining
2. Data mining tasks - Clustering, Classification, Rule learning, etc.
3. Data mining process: Data preparation/cleansing, task identification
   - Introduction to WEKA
4. Association Rule mining
5. Association rules - different algorithm types
6. Classification/Prediction
7. Classification - tree-based approaches
8. Classification - Neural Networks
   Midterm
9. Clustering basics
10. Clustering - statistical approaches
11. Clustering - Neural-net and other approaches
12. More on process - CRISP-DM
    - Preparation for final project
13. Text Mining
14. Multi-Relational Data Mining
15. Future trends
    Final

Text: Jiawei Han and Micheline Kamber, *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers, August 2000.

First: Academic Integrity

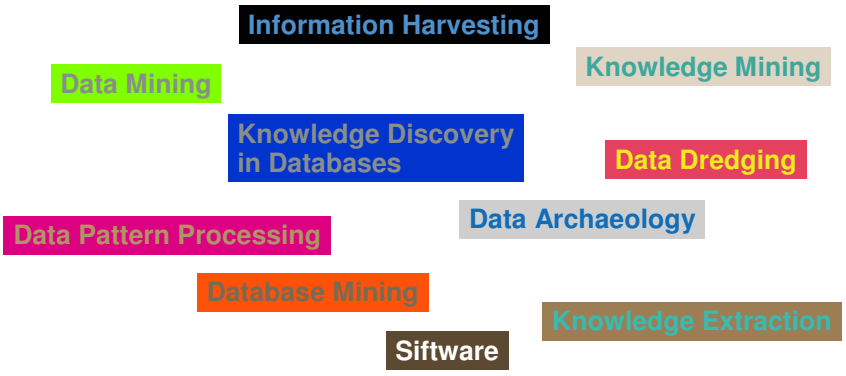
- Department of Computer Sciences has a new Academic Integrity Policy
  - https://portals.cs.purdue.edu/student/academic
  - Please read and sign
- Unless otherwise noted, worked turned in should reflect your independent capabilities
  - If unsure, note / cite sources and help
- Late work penalized 10%/day
  - No penalty for document emergency (e.g., medical) or by prior arrangement in special circumstances

Acknowledgements

Some of the material used in this course is drawn from other sources:

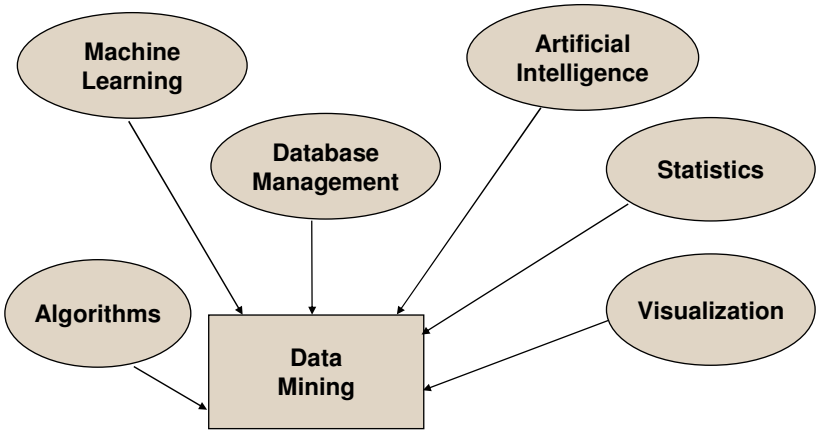
- Prof. Jiawei Han at UIUC
  - Started with Han's tutorial for UCLA Extension course in February 1998
  - Other subsequent contributors:
    - Dr. Hongjun Lu (Hong Kong Univ. of Science and Technology)
    - Graduate students from Simon Fraser Univ., Canada, notably Eugene Belchev, Jian Pei, and Osmar R. Zaiane
    - Graduate students from Univ. of Illinois at Urbana-Champaign
- Profs. Goharian and Grossman at IIT
  - NSF-funded course development
- Dr. Bhavani Thuraisingham (MITRE Corp. and NSF)

Data Mining—What's in a Name?



The process of discovering meaningful new correlations, patterns, and trends by sifting through large amounts of stored data, using pattern recognition technologies and statistical and mathematical techniques

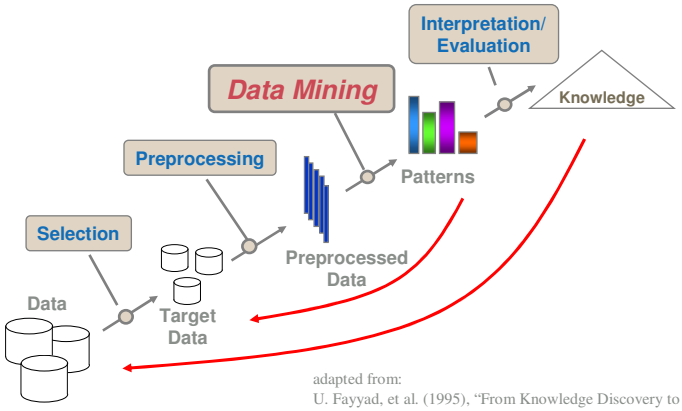
Integration of Multiple Technologies



Data Mining: Classification Schemes

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

Knowledge Discovery in Databases: Process

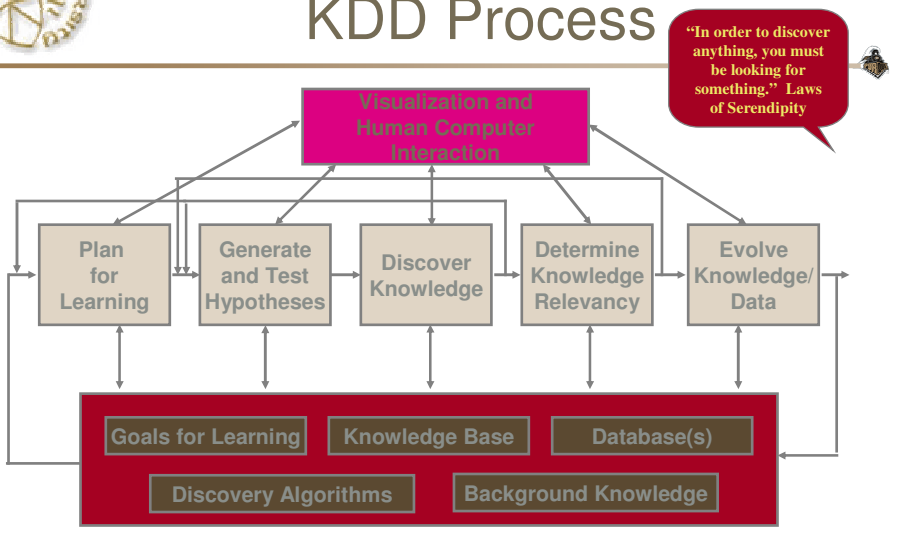


adapted from:
U. Fayyad, et al. (1995), "From Knowledge Discovery to Data Mining: An Overview," Advances in Knowledge Discovery and Data Mining, U. Fayyad et al. (Eds.), AAAI/MIT Press

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, Web mining, etc.

Ingredients of an Effective KDD Process



Data Mining: History of the Field

- Knowledge Discovery in Databases workshops started '89
  - Now a conference under the auspices of ACM SIGKDD
  - IEEE conference series started 2001
- Key founders / technology contributors:
  - Usama Fayyad, JPL (then Microsoft, now has his own company, Digimine)
  - Gregory Piatetsky-Shapiro (then GTE, now his own data mining consulting company, Knowledge Stream Partners)
  - Rakesh Agrawal (IBM Research)

The term "data mining" has been around since at least 1983 – as a pejorative term in the statistics community

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
  - DNA and bio-data analysis

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
  - Associations/co-relations between product sales, & prediction based on such association
- Customer profiling
  - What types of customers buy what products (clustering or classification)
- Customer requirement analysis
  - identifying the best products for different customers
  - predict what factors will attract new customers
- Provision of summary information
  - multidimensional summary reports
  - statistical summary information (data central tendency and variation)

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

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

Other Applications

- 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.

PURDUE
UNIVERSITY

CS490D: Introduction to Data Mining

Chris Clifton

January 14, 2004

Examples

Data Mining Tasks/Outcomes

Example: Use in retailing

- Goal: Improved business efficiency
  - Improve marketing (advertise to the most likely buyers)
  - Inventory reduction (stock only needed quantities)
- Information source: Historical business data
  - Example: Supermarket sales records

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

  - Size ranges from 50k records (research studies) to terabytes (years of data from chains)
  - Data is already being warehoused
- Sample question – what products are generally purchased together?
- The answers are in the data, if only we could see them

Data Mining applied to Aviation Safety Records (*Eric Bloedorn*)

- Many groups record data regarding aviation safety including the National Transportation Safety Board (NTSB) and the Federal Aviation Administration (FAA)
- Integrating data from different sources as well as mining for patterns from a mix of both structured fields and free text is a difficult task
- The goal of our initial analysis is to determine how data mining can be used to improve airline safety by finding patterns that predict safety problems

Aircraft Accident Report

- This data mining effort is an extension of the FAA Office of System Safety's Flight Crew Accident and Incident Human Factors Project
- In this previous approach two database-specific human error models were developed based on general research into human factors
  - FAA's Pilot Deviation database (PDS)
  - NTSB's accident and incident database
- These error models check for certain values in specific fields
- Result
  - Classification of some accidents caused by human mistakes and slips.

Problem

- Current model cannot classify a large number of records
- A large percentage of cases are labeled 'unclassified' by current model
  - ~58,000 in the NTSB database (90% of the events identified as involving people)
  - ~5,400 in the PDS database (93% of the events)
- Approximately 80,000 NTSB events are currently labeled 'unknown'
- Classification into meaningful human error classes is low because the explicit fields and values required for the models to fire are not being used
- Models must be adjusted to better describe data

Data mining Approach

- Use information from text fields to supplement current structured fields by extracting features from text in accident reports
- Build a human-error classifier directly from data
  - Use expert to provide class labels for events of interest such as ‘slips’, ‘mistakes’ and ‘other’
  - Use data-mining tools to build comprehensible rules describing each of these classes

Example Rule

- Sample Decision rule using current model features and text features

  If (person_code_1b= 5150,4105,5100,4100) and
  ((crew-subject-of-intentional-verb = true) or
  (modifier_code_1b = 3114))

  Then
  mistake
- “If pilot or copilot is involved and either the narrative, or the modifier code for 1b describes the crew as intentionally performing some action then this is a mistake”

Data Mining Ideas: Logistics

- Delivery delays
  - Debatable what data mining will do here; best match would be related to "quality analysis": given lots of data about deliveries, try to find common threads in "problem" deliveries
- Predicting item needs
  - Seasonal
    - Looking for cycles, related to similarity search in time series data
    - Look for similar cycles between products, even if not repeated
  - Event-related
    - Sequential association between event and product order (probably weak)

What Can Data Mining Do?

- Cluster
- Classify
  - Categorical, Regression
- Summarize
  - Summary statistics, Summary rules
- Link Analysis / Model Dependencies
  - Association rules
- Sequence analysis
  - Time-series analysis, Sequential associations
- Detect Deviations

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:

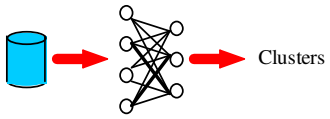
- Demographic analysis

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:

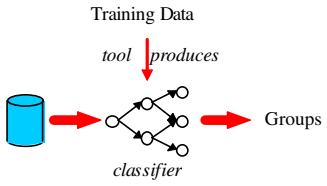
- Profiling

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

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

War Stories: Warehouse Product Allocation

The second project, identified as "Warehouse Product Allocation," was also initiated in late 1995 by RS Components' IS and Operations Departments. In addition to their warehouse in Corby, the company was in the process of opening another 500,000-square-foot site in the Midlands region of the U.K. To efficiently ship product from these two locations, it was essential that RS Components know in advance what products should be allocated to which warehouse. For this project, the team used IBM Intelligent Miner and additional optimization logic to split RS Components' product sets between these two sites so that the number of partial orders and split shipments would be minimized.

Parker says that the Warehouse Product Allocation project has directly contributed to a significant savings in the number of parcels shipped, and therefore in shipping costs. In addition, he says that the Opportunity Selling project not only increased the level of service, but also made it easier to provide new subsidiaries with the value-added knowledge that enables them to quickly ramp-up sales.

"By using the data mining tools and some additional optimization logic, IBM helped us produce a solution which heavily outperformed the best solution that we could have arrived at by conventional techniques," said Parker. "The IBM group tracked historical order data and conclusively demonstrated that data mining produced increased revenue that will give us a return on investment 10 times greater than the amount we spent on the first project."

http://direct.boulder.ibm.com/dss/customer/rscomp.html

War Stories: Inventory Forecasting

American Entertainment Company

Forecasting demand for inventory is a central problem for any distributor. Ship too much and the distributor incurs the cost of restocking unsold products; ship too little and sales opportunities are lost.

IBM Data Mining Solutions assisted this customer by providing an inventory forecasting model, using segmentation and predictive modeling. This new model has proven to be considerably more accurate than any prior forecasting model.

More war stories (many humorous) starting with slide 21 of:
http://robotics.stanford.edu/~ronnyk/chasm.pdf

Reading Literature you Might Consider

- R. Agrawal, J. Han, and H. Mannila, Readings in Data Mining: A Database Perspective, Morgan Kaufmann (in preparation)
- 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, 2001
- D. J. Hand, H. Mannila, and P. Smyth, Principles of Data Mining, MIT Press, 2001
- T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Springer-Verlag, 2001
- T. M. Mitchell, Machine Learning, McGraw Hill, 1997
- G. Piatetsky-Shapiro and W. J. Frawley. Knowledge Discovery in Databases. AAAI/MIT Press, 1991
- 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, 2001

Necessity for Data Mining

- Large amounts of current and historical data being stored
  - Only small portion (~5-10%) of collected data is analyzed
  - Data that may never be analyzed is collected in the fear that something that may prove important will be missed
- As databases grow larger, decision-making from the data is not possible; need knowledge derived from the stored data
- Data sources
  - Health-related services, e.g., benefits, medical analyses
  - Commercial, e.g., marketing and sales
  - Financial
  - Scientific, e.g., NASA, Genome
  - DOD and Intelligence
- Desired analyses
  - Support for planning (historical supply and demand trends)
  - Yield management (scanning airline seat reservation data to maximize yield per seat)
  - System performance (detect abnormal behavior in a system)
  - Mature database analysis (clean up the data sources)

Data Mining Complications

- Volume of Data
  - Clever algorithms needed for reasonable performance
- Interest measures
  - How do we ensure algorithms select “interesting” results?
- “Knowledge Discovery Process” skill required
  - How to select tool, prepare data?
- Data Quality
  - How do we interpret results in light of low quality data?
- Data Source Heterogeneity
  - How do we combine data from multiple sources?

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

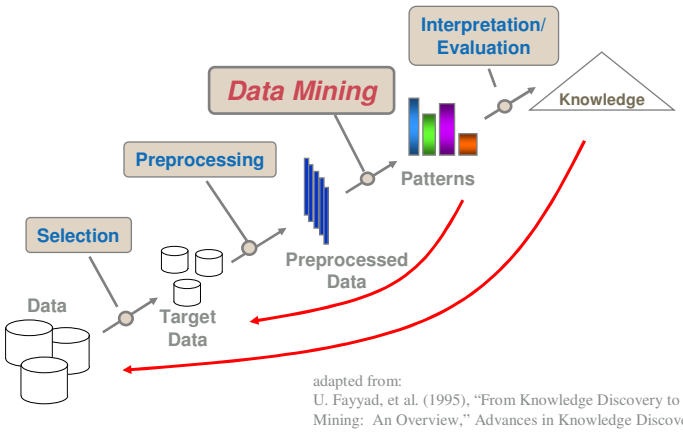
Are All the "Discovered" Patterns Interesting?

- 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

Knowledge Discovery in Databases: Process

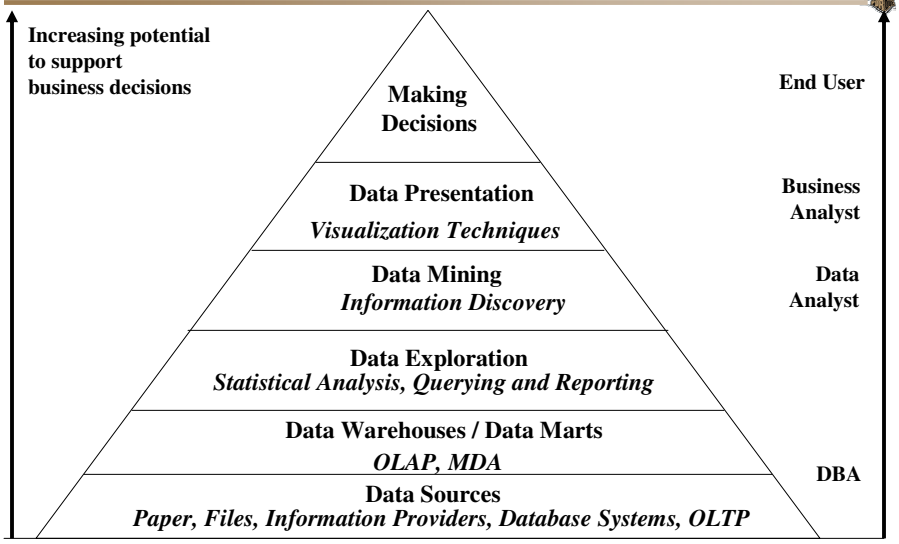


adapted from:
U. Fayyad, et al. (1995), "From Knowledge Discovery to Data Mining: An Overview," Advances in Knowledge Discovery and Data Mining, U. Fayyad et al. (Eds.), AAAI/MIT Press

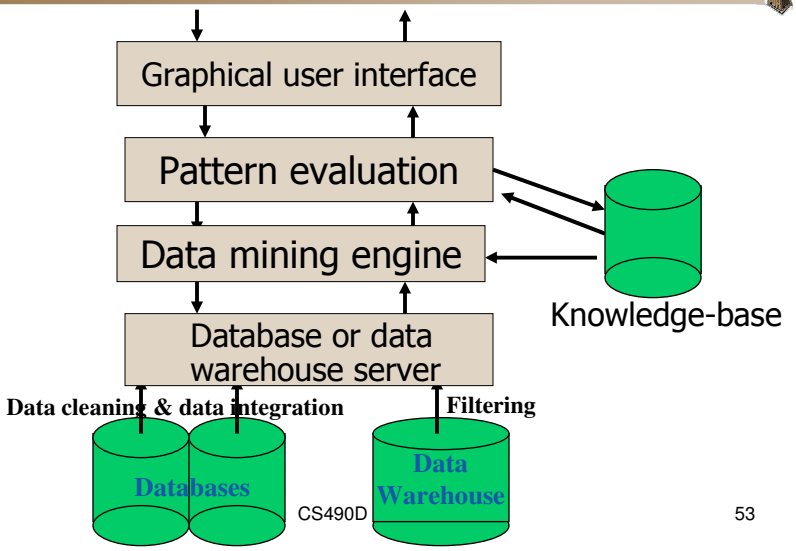
Steps of a KDD Process

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

Data Mining and Business Intelligence



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

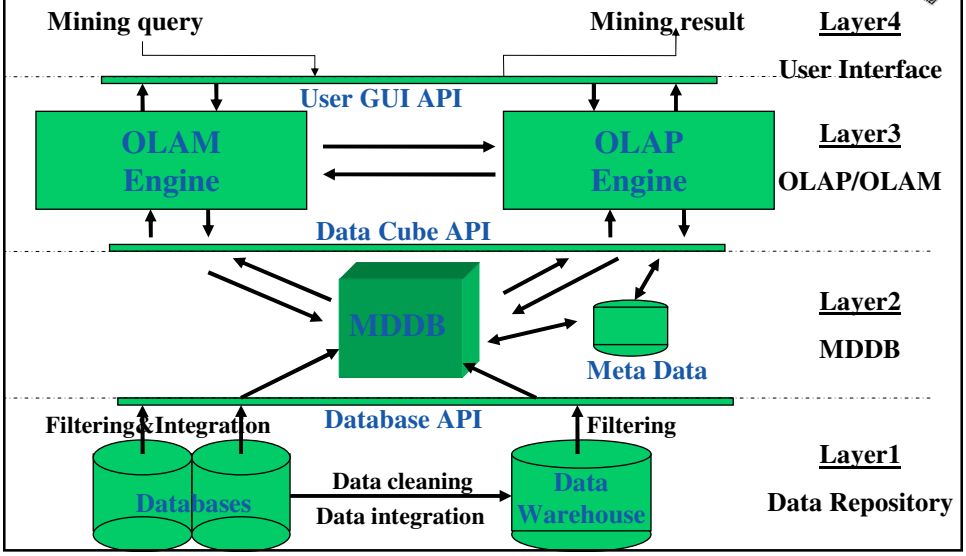
Related Techniques: OLAP
On-Line Analytical Processing

- 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

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

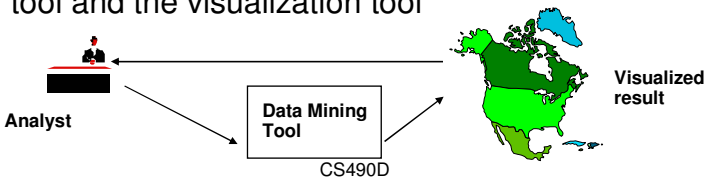


Related Techniques: Visualization

- Visualization uses human perception to recognize patterns in large data sets
- Advantages relative to data mining
  - Perceive "unconsidered" patterns
  - Recognize non-linear relationships
- Disadvantages relative to data mining
  - Data set size limited by resolution constraints
  - Hard to recognize "small" patterns
  - Difficult to quantify results

Data Mining and Visualization

- Approaches
  - Visualization to display results of data mining
    - Help analyst to better understand the results of the data mining tool
  - Visualization to aid the data mining process
    - Interactive control over the data exploration process
    - Interactive steering of analytic approaches ("grand tour")
- Interactive data mining issues
  - Relationships between the analyst, the data mining tool and the visualization tool



Large-scale Endeavors

Products

	Clustering	Classification	Association	Sequence	Deviation
SAS		Decision Trees			
SPSS		√	√		
Oracle (Darwin)	√	ANN			
IBM	Time Series	Decision Trees	√	√	√
DBMiner (Simon Fraser)		√	√		

Research