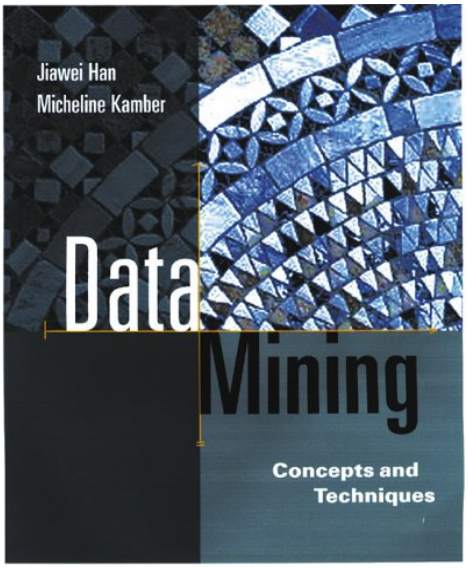
Data Mining: Concepts and Techniques

Introduction

Introduction

- Motivation: Why data mining?
- What is data mining?
- Data Mining: On what kind of data?
- Data mining functionality
- Are all the patterns interesting?
- Classification of data mining systems
- Major issues in data mining

Data Mining: Concepts and Techniques



Data Mining

Data Mining -------- What ?

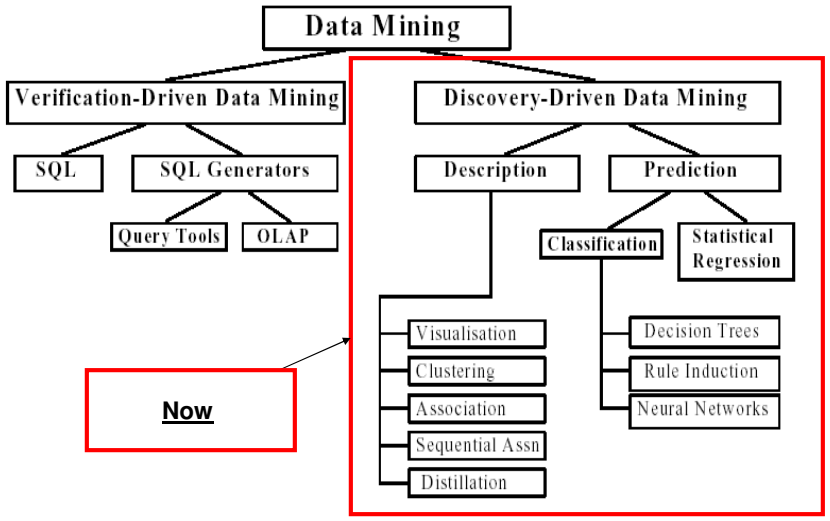
- Google returns around 0.5 million hits for "Data Mining" + Definition

- Some common terms shared by almost all definitions

  Discovery – Hidden Patterns – Non trivial – Process

- What's in a Name
  - Information Extraction, Information Retrieval (IR), Knowledge Discovery, Knowledge Discovery in Databases (KDD), Pattern Detection etc.

Data Mining -------- What ?



Data Mining -------- When?

- Sporadic examples can be traced back to as early as 1854 when famous English physician John Snow found the cause of cholera outbreak by "mining" patient demographics.

Majority cases found at intersections where community water pumps are situated.

- History of Data Mining can better be traced from the evolution of
  - Statitistics
  - Artificial Intelligence
  - Machine Learning

Data Mining -------- Why?

- Datasets Mounting both Horizontally and vertically

	Fields (d)	Records (n)
Retail (WalMart scale)	~100-150	~100 m
Web (Amazon/Google scale)	~ 40-50	~ 1000 m
Medical (Gene expression scale)	~1k - ~10K	~ 100 - ~1000 m

- An RDBMS solution to info retrieval
  - Summarization and Aggregation
- Wont work in many cases
  - Multiple formats, interesting patterns lost in summarization, outliers mix-up with noise

Data Mining ------- Where?

- Numerous groups around the world, here we review some of these:
- Microsoft Research: Data Management, Exploration and Mining Group (DMX)
  - Integration of data mining with database systems (OLE DB-DM)
  - OS Integrated Search (Remember OS Integrated Browser)
- Google
  - Highest number of PhDs under one roof!!
  - Web-Scale search; Currently the top-dog in Google Wars.
- IBM
- Stanford DB group
- CMU DB group
- University of Wisconsin DB group
- University of Helsinki
- .....

Necessity Is the Mother of Invention

- Data explosion problem
  - Automated data collection tools and mature database technology lead to tremendous amounts of data accumulated and/or to be analyzed in databases, data warehouses, 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
  - Mining interesting knowledge (rules, regularities, patterns, constraints) from data in large databases

Evolution of Database Technology

- 1960s:
  - Data collection, database creation 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 with a variety of applications
  - Web technology and global information systems

What Is Data Mining?

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

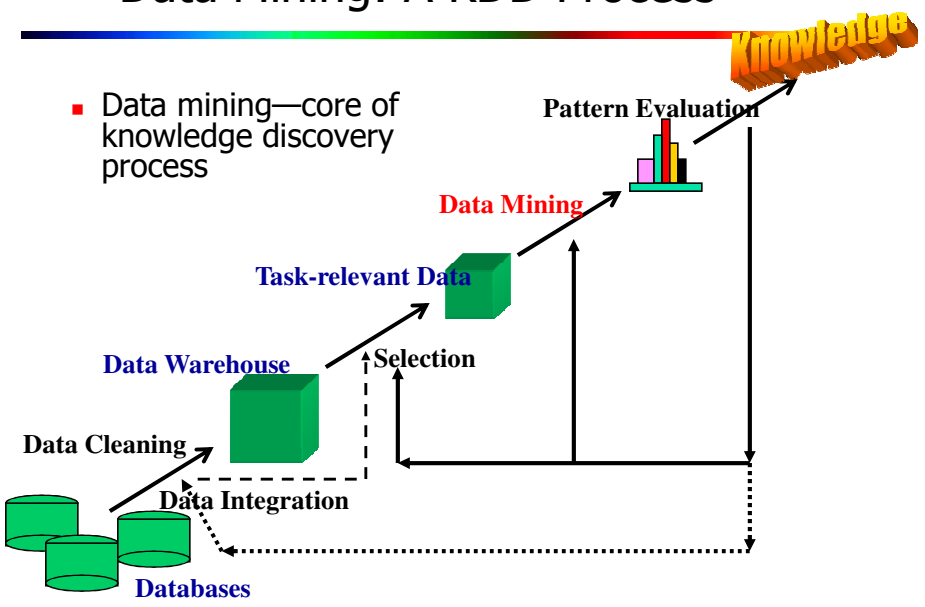
- Approaches: Clustering & model construction for frauds, outlier analysis
- Applications: Health care, retail, credit card service, telecomm.
  - Auto insurance: ring of collisions
  - Money laundering: suspicious monetary transactions
  - Medical insurance
    - Professional patients, ring of doctors, 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
- 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.

Data Mining: A KDD Process

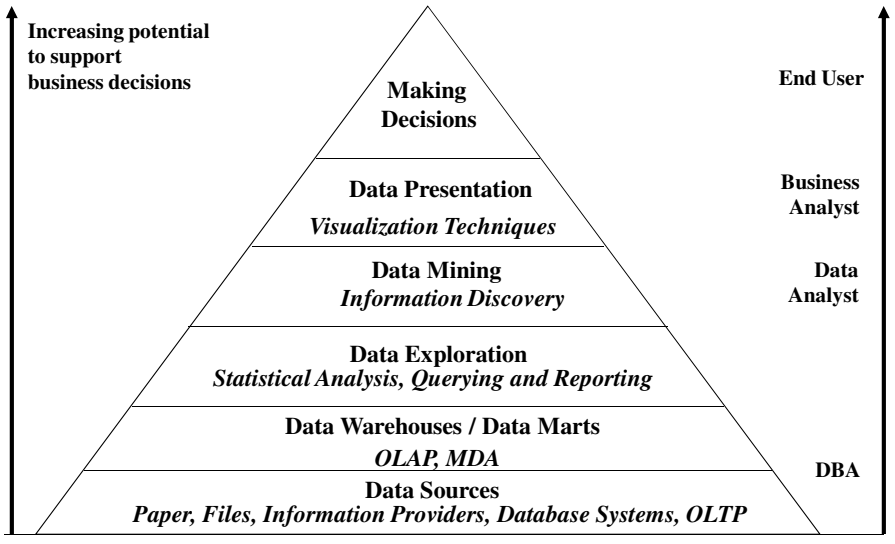
- Data mining—core of knowledge discovery process



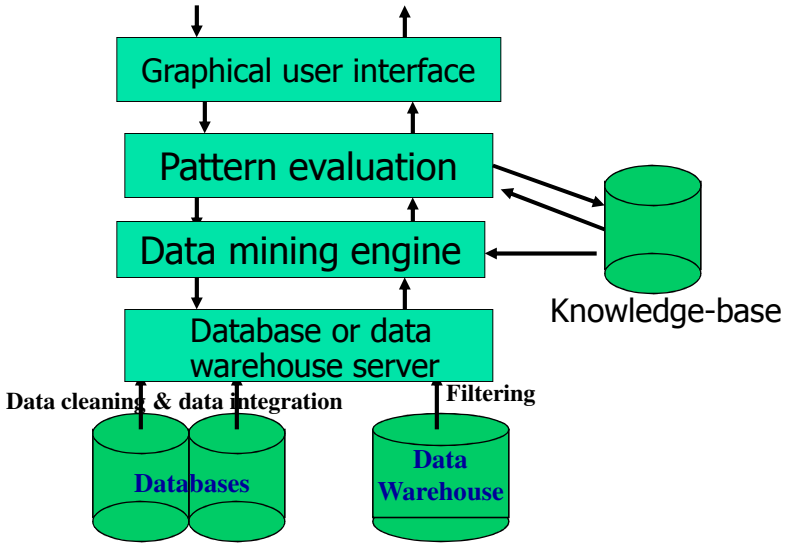
Steps of a KDD Process

- 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



Data Mining: On What Kinds of Data?

- Relational database
- Data warehouse
- Transactional database
- Advanced database and information repository
  - Object-relational database
  - Spatial and temporal data
  - Time-series data
  - Stream data
  - Multimedia database
  - Heterogeneous and legacy database
  - Text databases & WWW

Data Mining Functionalities

- Concept description: Characterization and discrimination
  - Generalize, summarize, and contrast data characteristics, e.g., dry vs. wet regions
- Association (correlation and causality)
  - Diaper → Beer [0.5%, 75%]
- 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
  - Presentation: decision-tree, classification rule, neural network
  - 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: a data object that does not comply with the general behavior of the data
  - Noise or exception? No! useful in fraud detection, rare events analysis
- Trend and evolution analysis
  - Trend and deviation: regression analysis
  - Sequential pattern mining, periodicity analysis
  - Similarity-based analysis
- Other pattern-directed or statistical analyses

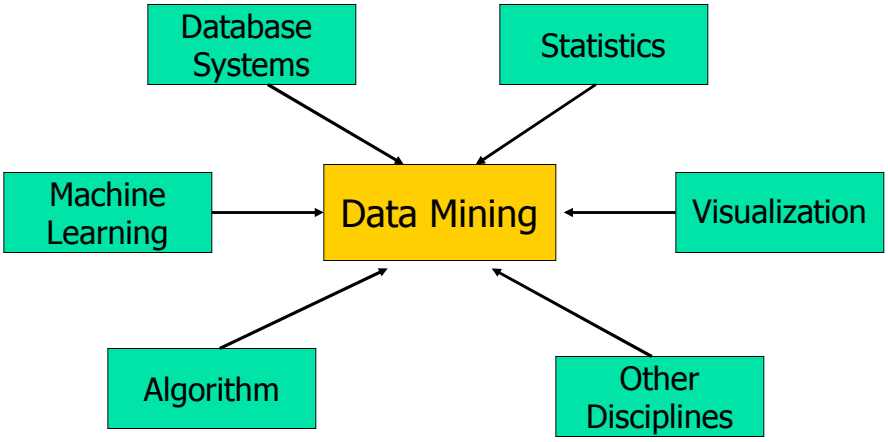
Are All the "Discovered" Patterns Interesting?

- Data mining may generate thousands of patterns: Not all of them are interesting
  - Suggested approach: Human-centered, query-based, focused mining
- **Interestingness measures**
  - A pattern is interesting if it is easily understood by humans, valid on new or test data with some degree of certainty, potentially useful, novel, or validates some hypothesis that a user seeks to confirm
- **Objective vs. subjective interestingness measures**
  - Objective: based on statistics and structures of patterns, e.g., support, confidence, etc.
  - Subjective: based on user's belief in the data, e.g., unexpectedness, novelty, actionability, etc.

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, different classifications
  - Kinds of data to be mined
  - Kinds of knowledge to be discovered
  - Kinds of techniques utilized
  - Kinds of applications adapted

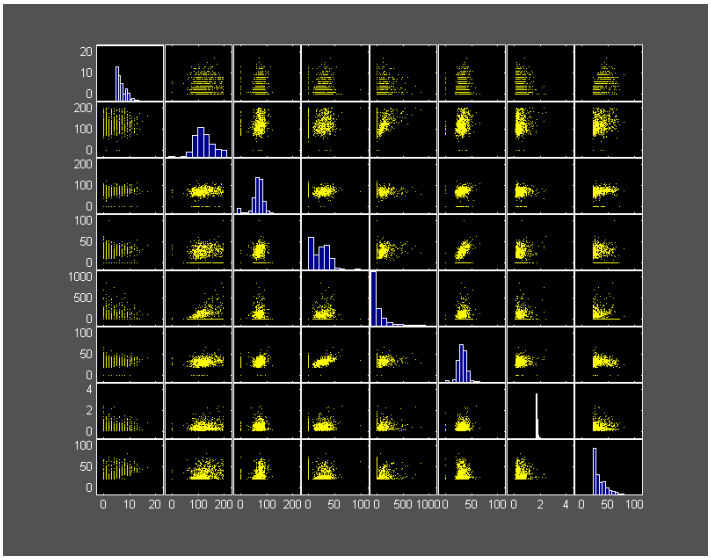
Different Data Mining Tasks

- Exploratory Data Analysis
- Descriptive Modeling
- Predictive Modeling
- Discovering Patterns and Rules
- + others….

Exploratory Data Analysis

- Getting an overall sense of the data set
  - Computing summary statistics:
    - Number of distinct values, max, min, mean, median, variance, skewness,..
- Visualization is widely used
  - 1d histograms
  - 2d scatter plots
  - Higher-dimensional methods
- Useful for data checking
  - E.g., finding that a variable is always integer valued or positive
  - Finding the some variables are highly skewed
- Simple exploratory analysis can be extremely valuable
  - You should always "look" at your data before applying any data mining algorithms

Example of Exploratory Data Analysis
(Pima Indians data, scatter plot matrix)



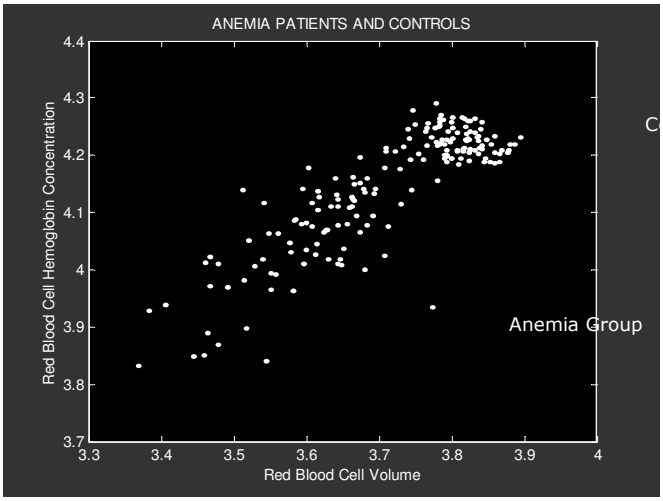
Different Data Mining Tasks

- Exploratory Data Analysis
- Descriptive Modeling
- Predictive Modeling
- Discovering Patterns and Rules
- + others….

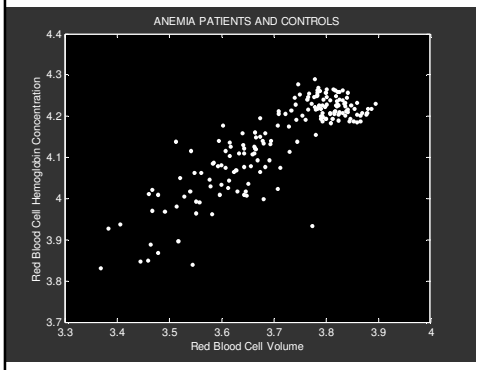
Descriptive Modeling

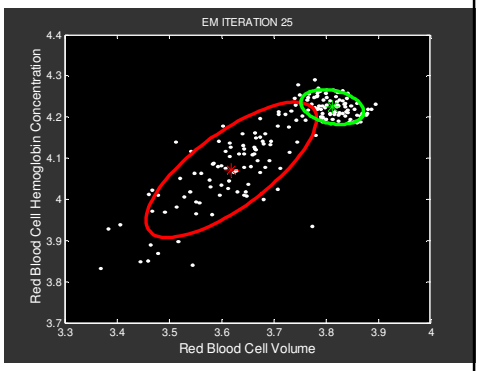
- Goal is to build a "descriptive" model
  - e.g., a model that could simulate the data if needed
  - models the underlying process
- Examples:
  - Density estimation:
    - estimate the joint distribution $P(x_1,\ldots\ldots x_p)$
  - Cluster analysis:
    - Find natural groups in the data
  - Dependency models among the p variables
    - Learning a Bayesian network for the data

Example of Descriptive Modeling



Example of Descriptive Modeling





Learning User Navigation Patterns from Web Logs

128.195.36.195, -, 3/22/00, 10:35:11, W3SVC, SRVR1, 128.200.39.181, 781, 363, 875, 200, 0, GET, /top.html, -,
128.195.36.195, -, 3/22/00, 10:35:16, W3SVC, SRVR1, 128.200.39.181, 5288, 524, 414, 200, 0, POST, /spt/main.html, -,
128.195.36.195, -, 3/22/00, 10:35:17, W3SVC, SRVR1, 128.200.39.181, 30, 280, 111, 404, 3, GET, /spt/images/bk1.jpg, -,
128.195.36.101, -, 3/22/00, 16:18:50, W3SVC, SRVR1, 128.200.39.181, 60, 425, 72, 304, 0, GET, /top.html, -,
128.195.36.101, -, 3/22/00, 16:18:58, W3SVC, SRVR1, 128.200.39.181, 8322, 527, 414, 200, 0, POST, /spt/main.html, -,
128.195.36.101, -, 3/22/00, 16:18:59, W3SVC, SRVR1, 128.200.39.181, 0, 280, 111, 404, 3, GET, /spt/images/bk1.jpg, -,
128.200.39.17, -, 3/22/00, 20:54:37, W3SVC, SRVR1, 128.200.39.181, 140, 199, 875, 200, 0, GET, /top.html, -,
128.200.39.17, -, 3/22/00, 20:54:55, W3SVC, SRVR1, 128.200.39.181, 17766, 365, 414, 200, 0, POST, /spt/main.html, -,
128.200.39.17, -, 3/22/00, 20:54:55, W3SVC, SRVR1, 128.200.39.181, 0, 258, 111, 404, 3, GET, /spt/images/bk1.jpg, -,
128.200.39.17, -, 3/22/00, 20:55:07, W3SVC, SRVR1, 128.200.39.181, 0, 258, 111, 404, 3, GET, /spt/images/bk1.jpg, -,
128.200.39.17, -, 3/22/00, 20:55:36, W3SVC, SRVR1, 128.200.39.181, 1061, 382, 414, 200, 0, POST, /spt/main.html, -,
128.200.39.17, -, 3/22/00, 20:55:36, W3SVC, SRVR1, 128.200.39.181, 0, 258, 111, 404, 3, GET, /spt/images/bk1.jpg, -,
128.200.39.17, -, 3/22/00, 20:55:39, W3SVC, SRVR1, 128.200.39.181, 0, 258, 111, 404, 3, GET, /spt/images/bk1.jpg, -,
128.200.39.17, -, 3/22/00, 20:56:03, W3SVC, SRVR1, 128.200.39.181, 1081, 382, 414, 200, 0, POST, /spt/main.html, -,
128.200.39.17, -, 3/22/00, 20:56:04, W3SVC, SRVR1, 128.200.39.181, 0, 258, 111, 404, 3, GET, /spt/images/bk1.jpg, -,
128.200.39.17, -, 3/22/00, 20:56:33, W3SVC, SRVR1, 128.200.39.181, 0, 262, 72, 304, 0, GET, /top.html, -,
128.200.39.17, -, 3/22/00, 20:56:52, W3SVC, SRVR1, 128.200.39.181, 19598, 382, 414, 200, 0, POST, /spt/main.html, -,

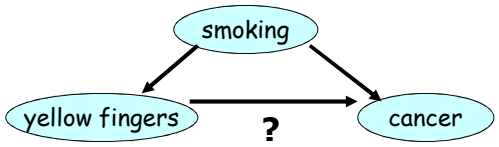
User 1	2	3	2	2	3	3	3	1	1	1	3	1	3	3	3	3
User 2	3	3	3	1	1	1										
User 3	7	7	7	7	7	7	7	7								
User 4	1	5	1	1	1	5	1	5	1	1	1	1	1	1		
User 5	5	1	1	5												
...		...														

Another Example of Descriptive Modeling

- Learning Directed Graphical Models (aka Bayes Nets)
  - goal: learn directed relationships among p variables
  - techniques: directed (causal) graphs
  - challenge: distinguishing between correlation and causation

  – example: Do yellow fingers cause lung cancer?

hidden cause: smoking



Different Data Mining Tasks

- Exploratory Data Analysis
- Descriptive Modeling
- Predictive Modeling
- Discovering Patterns and Rules
- + others….

Predictive Modeling

- Predict one variable Y given a set of other variables X
  - Here X could be a p-dimensional vector
  - Classification: Y is categorical
  - Regression: Y is real-valued
- In effect this is function approximation, learning the relationship between Y and X
- Many, many algorithms for predictive modeling in statistics and machine learning
- Often the emphasis is on predictive accuracy, less emphasis on understanding the model

Predictive Modeling: Fraud Detection

- Credit card fraud detection
  - Credit card losses in the US are over 1 billion $ per year
  - Roughly 1 in 50k transactions are fraudulent
- Approach
  - For each transaction estimate p(fraudulent | transaction)
  - Model is built on historical data of known fraud/non-fraud
  - High probability transactions investigated by fraud police
- Example:
  - Fair-Isaac/HNC's fraud detection software based on neural networks, led to reported fraud decreases of 30 to 50%
  - http://www.fairisaac.com/fairisaac
- Issues
  - Significant feature engineering/preprocessing
  - false alarm rate vs missed detection – what is the tradeoff?

Predictive Modeling: Customer Scoring

- Example: a bank has a database of 1 million past customers, 10% of whom took out mortgages
- Use machine learning to rank new customers as a function of p(mortgage | customer data)
- Customer data
  - History of transactions with the bank
  - Other credit data (obtained from Experian, etc)
  - Demographic data on the customer or where they live
- Techniques
  - Binary classification: logistic regression, decision trees, etc
  - Many, many applications of this nature

Predictive Modeling: Telephone Call Modeling

- Background
  - AT&T has about 100 million customers
  - It logs 200 million calls per day, 40 attributes each
  - 250 million unique telephone numbers
  - Which are business and which are residential?
- Approach (Pregibon and Cortes, AT&T,1997)
  - Proprietary model, using a few attributes, trained on known business customers to adaptively track p(business|data)
  - Significant systems engineering: data are downloaded nightly, model updated (20 processors, 6Gb RAM, terabyte disk farm)
- Status:
  - running daily at AT&T
  - HTML interface used by AT&T marketing

Different Data Mining Tasks

- Exploratory Data Analysis
- Descriptive Modeling
- Predictive Modeling
- Discovering Patterns and Rules
- + others….

Pattern Discovery

- Goal is to discover interesting "local" patterns in the data rather than to characterize the data globally
- given market basket data we might discover that
  - If customers buy wine and bread then they buy cheese with probability 0.9
  - These are known as "association rules"
- Given multivariate data on astronomical objects
  - We might find a small group of previously undiscovered objects that are very self-similar in our feature space, but are very far away in feature space from all other objects

Example of Pattern Discovery

ADACABDABAABBDDBCADDDDBCDDBCCBBCCDADADAADABDBBDABABBCDDDCD
DABDCBBDBDBCBBABBBCBBABCBBACBBDBAACCADDADBDBBCBBCCBBBDCABD
DBBADDBBBBCCACDABBABDDCDDBBABDBDDBDDBCACDBBCCBBACDCADCBACC
ADCCCACCDDADCBCADADBAACCDDDCBDBDCCCCACACACCDABDDBCADADBCBD
DADABCCABDAACABCABACBDDDCBADCBDADDDDCDDCADCCBBADABBAAADAAA
BCCBCABDBAADCBCDACBCABABCCBACBDABDDDADAABADCDCCDBBCDBDADDC
CBBCDBAADADBCAAAADBDCADBDBBBCDCCBCCCDCCADAADACABDABAABBDDB
CADDDDBCDDBCCBBCCDADADACCCDABAABBCBDBDBADBBBBCDADABABBDACD
CDDDBBCDBBCBBCCDABCADDADBACBBBCCDBAAADDDBDDCABACBCADCDCBAA
ADCADDADAABBACCBB

December 20, 2010 45

Example of Pattern Discovery

ADACABDABAABBDDBCADDDDBCDDBCCBBCCDADADAADABDBBDABABBCDDDCD
DABDCBBDBDBCBBABBBCBBABCBBACBBDBAACCADDADBDBBCBBCCBBBDCABD
DBBADDBBBBCCACDABBABDDCDDBBABDBDDBDDBCACDBBCCBBACDCADCBACC
ADCCCACCDDADCBCADADBAACCDDDCBDBDCCCCACACACCDABDDBCADADBCBD
DADABCCABDAACABCABACBDDDCBADCBDADDDDCDDCADCCBBADABBAAADAAA
BCCBCABDBAADCBCDACBCABABCCBACBDABDDDADAABADCDCCDBBCDBDADDC
CBBCDBAADADBCAAAADBDCADBDBBBCDCCBCCCDCCADAADACABDABAABBDDB
CADDDDBCDDBCCBBCCDADADACCCDABAABBCBDBDBADBBBBCDADABABBDACD
CDDDBBCDBBCBBCCDABCADDADBACBBBCCDBAAADDDBDDCABACBCADCDCBAA
ADCADDADAABBACCBB

December 20, 2010 46

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

Summary

- Data mining: discovering interesting patterns from large amounts of data
- A natural evolution of database technology, in great demand, with wide applications
- A KDD process includes data cleaning, data integration, data selection, transformation, data mining, pattern evaluation, and knowledge presentation
- Mining can be performed in a variety of information repositories
- Data mining functionalities: characterization, discrimination, association, classification, clustering, outlier and trend analysis, etc.
- Data mining systems and architectures
- Major issues in data mining