Unit-3
Data Mining

Contents

- Fundamentals of data mining, Data Mining definitions, KDD V/s Data Mining, Data Mining Functionalities, From Data Warehousing to Data Mining, DBMS V/s DM, Issues and challenges in Data Mining.
- Data Mining Primitives, Data Mining Query Languages.
- Data Mining applications-Case studies

What is Data?

Data is a set of facts/observations/ measurements about objects/ events/processes of interest

Higher Standards meijer Lower Prices
Elida Rd.
Lima, OH - # 110
(419)331-5400 meijer.com

The Meijer Team appreciates your business
05/28/11
Your fast and friendly checkout was
provided by Fastlane107

****SAVINGS TODAY****
* TOTAL MEIJER PROMOTIONS 1.00 *
* TOTAL NON-COUPON SAVINGS 5.12 *
* TOTAL COUPON SAVINGS OF 4.49 *
SAVINGS TOTAL 10.61

GROCERY
*2670012911 FRENCH DIP
was 1.69 now .99 F
mPerks Offer
-- 1.00 off -.99 F
* Limit of .99 reached
*2670032200 DEANS DIP
was 1.69 now .99 F
mPerks Offer
-- 1.00 off -.01 F
* Limit of 1.00 reached
*3760028225 SALSA
was 2.49 now .99 F
*4335400750 TORTILLAS
was 2.39 now 1.89 F
*1901401852 DOG FOOD
2 @ 16.99
was 35.76 now 33.98 T

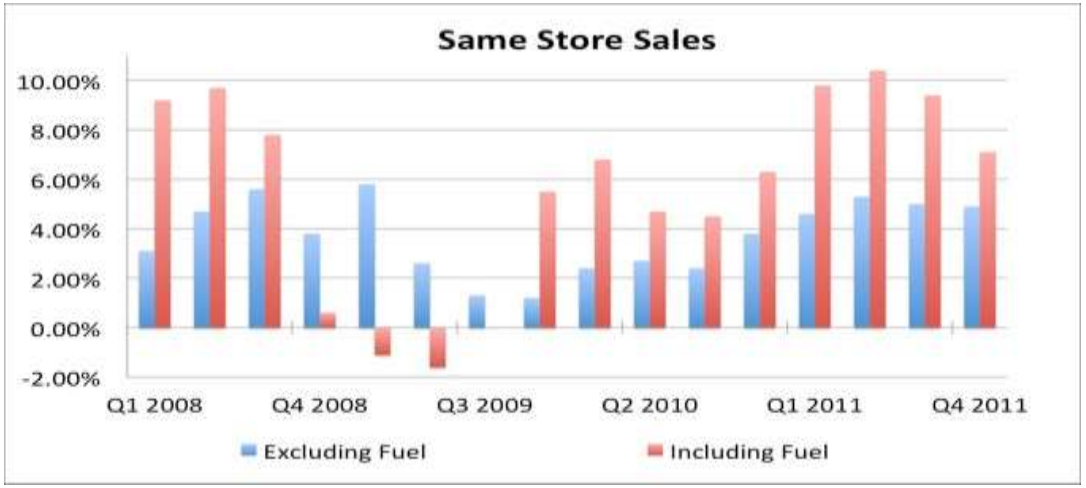
COUPONS
Vendor Coupon -.50 F
EXTRA COUPON -.50 F
81101003760008 Vendor Coupon -.50 F
81101003760008 EXTRA COUPON -.49 F
81101001901403 Vendor Coupon -1.00 N
81101001901403 Vendor Coupon -1.00 N
81101002670002 Vendor Coupon -.500 F

Mperks # ********63
Meijer Card 00900875
Monthly purch (up to 48 hr delay) .00
TOTAL
TOTAL TAX 2.21
TOTAL 35.56

© IKSIN C

What is Information?

Information is processed data that is useful in one way or the other, for example for decision making, communication etc.



While the data is fixed, information from it can differ based on needs

© IKSIN C

What is Knowledge?

Patterns of relationships in data and information that exhibit a high degree of certainty

© IKSIN C

Data Classification

Structured Data

Data consisting of well-defined fields of numeric or alphanumeric values

© IKSIN C

ORDER FILE

Order Number	Order Date	Customer Number	Delivery Address	Concrete Type	Amount	Truck Number	Driver ID
100000	9/1/2004	1234	55 Smith Lane	1	8	111	123456789
100001	9/1/2004	3456	2122 E. Biscayne	1	3	222	785934444
100002	9/2/2004	1234	55 Smith Lane	5	6	222	435296657
100003	9/3/2004	4567	1333 Burr Ridge	2	4	333	435296657
100004	9/4/2004	4567	1333 Burr Ridge	2	8	222	785934444
100005	9/4/2004	5678	1222 Westminster	1	4	222	785934444
100006	9/5/2004	1234	222 East Hampton	1	4	111	123456789
100007	9/6/2004	2345	9 W. Palm Beach	2	5	333	785934444
100008	9/8/2004	6789	4532 Lane Circle	1	8	222	785934444
100009	9/7/2004	1234	987 Furlong	3	8	111	123456789
100010	9/9/2004	6789	4532 Lance Circle	2	7	222	435296657
100011	9/9/2004	4567	3500 Tomahawk	5	6	222	785934444

CUSTOMER FILE

Customer Number	Customer Name	Customer Phone	Customer Primary Contact
1234	Smelding Homes	3333333333	Bill Johnson
2345	Home Builders Superior	3334444444	Marcus Connolly
3456	Mark Akey	3335555555	Mark Akey
4567	Triple A Homes	3336666666	Janielle Smith
5678	Sheryl Williamson	3337777777	Sheryl Williamson
6789	Home Makers	3338888888	John Yu

EMPLOYEE FILE

Employee ID	Employee Last Name	Employee First Name	Date of Hire
123456789	Johnson	Emilio	2/1/1985
435296657	Evaraz	Antonio	3/3/1992
785934444	Robertson	John	6/1/1999
984568756	Smithson	Allison	4/1/1997

TRUCK FILE

Truck Number	Truck Type	Date of Purchase
111	Ford	6/17/1999
222	Ford	12/24/2001
333	Chevy	1/1/2002

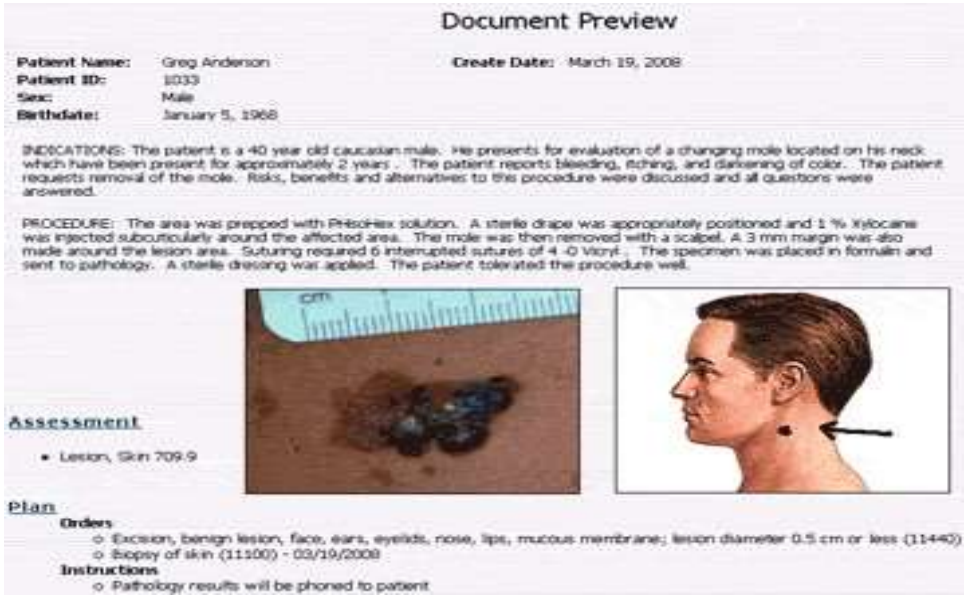
Data Classification

- **Unstructured Data**
  - No well defined fields of information
  - Requires extensive processing to extract content information
  - Examples include blogs, news reports, images, videos, tweets etc.
  - Fastest growing data segment

© IKSIN C

Data Classification

- Semi-Structured Data
  - Data with partial structure (medical reports, executive summaries, interview scripts, web documents etc.)



Document Preview

Patient Name: Greg Anderson **Create Date:** March 19, 2008
Patient ID: 1033
Sex: Male
Birthdate: January 5, 1968

INDICATIONS: The patient is a 40 year old caucasian male. He presents for evaluation of a changing mole located on his neck which have been present for approximately 2 years . The patient reports bleeding, itching, and darkening of color. The patient requests removal of the mole. Risks, benefits and alternatives to this procedure were discussed and all questions were answered.

PROCEDURE: The area was prepped with PhisoHex solution. A sterile drape was appropriately positioned and 1 % Xylocaine was injected subcuticularly around the affected area. The mole was then removed with a scalpel. A 3 mm margin was also made around the lesion area. Suturing required 6 interrupted sutures of 4 -0 Vicryl . The specimen was placed in formalin and sent to pathology. A sterile dressing was applied. The patient tolerated the procedure well.

Assessment

- Lesion, Skin 709.9

Plan

Orders

- Excision, benign lesion, face, ears, eyelids, nose, lips, mucous membrane; lesion diameter 0.5 cm or less (11440)
- Biopsy of skin (11100) - 03/19/2008

Instructions

- Pathology results will be phoned to patient

© IKSIN C

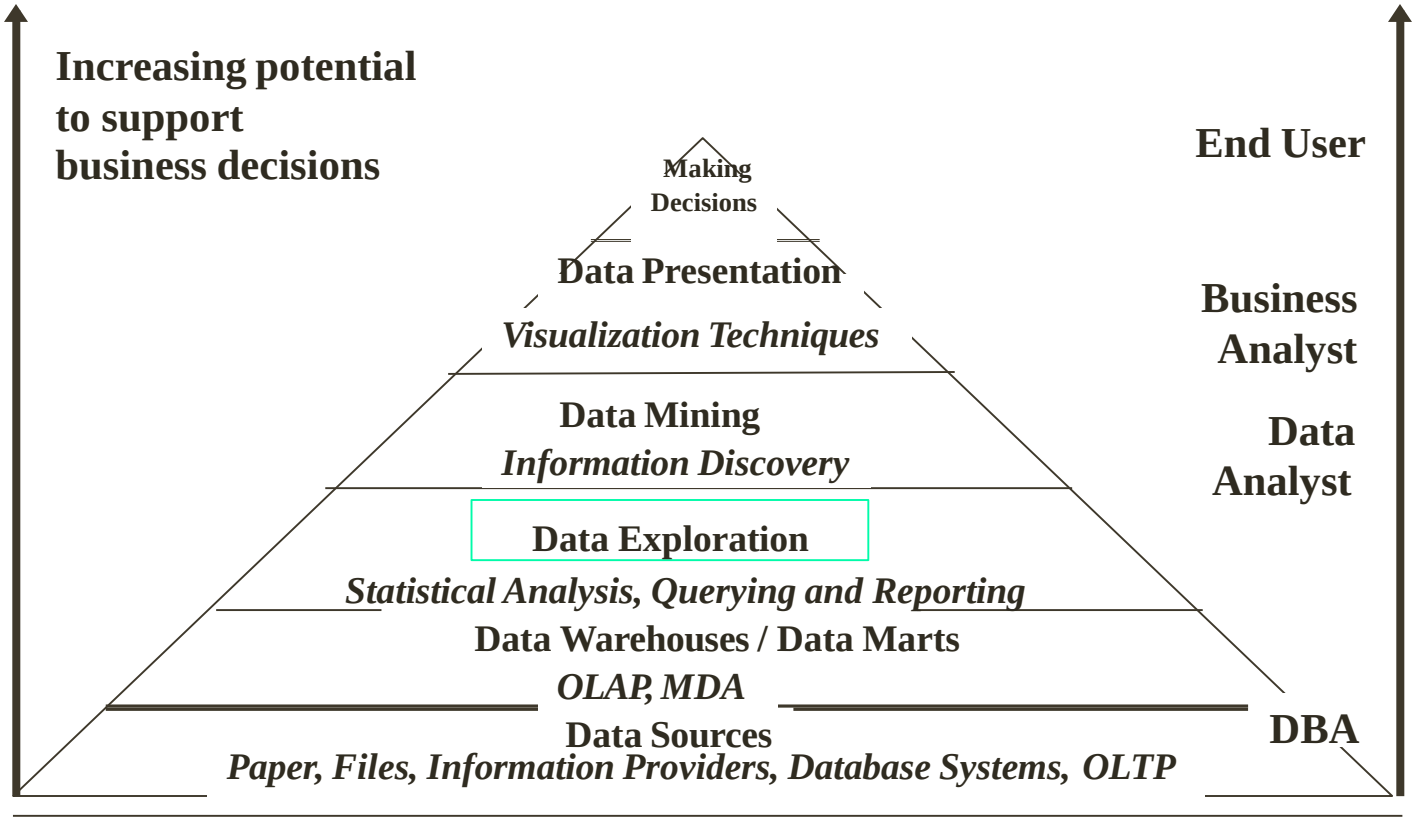
Why Data Mining?

- The Explosive Growth of Data: from terabytes to petabytes
  - Data collection and data availability
    - Automated data collection tools, database systems, Web, computerized society
  - Major sources of abundant data
    - Business: Web, e-commerce, transactions, stocks, …
    - Science: Remote sensing, bioinformatics, scientific simulation, …
    - Society and everyone: news, digital cameras,
- We are drowning in data, but starving for knowledge!
- "Necessity is the mother of invention"—Data mining—Automated analysis of massive data sets

Data Mining: Definitions

- Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data.
- The efficient discovery of previously unknown, valid, potentially useful, understandable patterns in large datasets
- The analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner

Data Mining and Business Intelligence



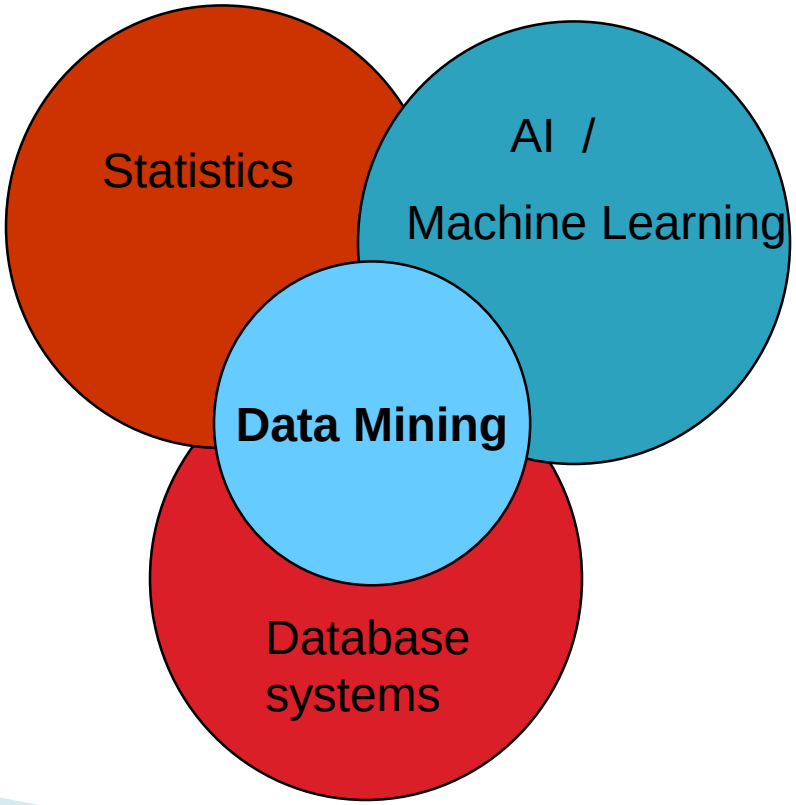
© IKSIN C

DBMS Vs Data Mining

S. No.	Database Management System (DBMS)	Data Mining
1.	DBMS is basically a product or software which is used to manage data.	Data mining is the procedure to extract the information from huge amount of raw data which can be used to take business decisions or to decide future strategies.
2.	DBMS support query languages which are useful for query triggered explorations.	Data mining involve use of various algorithms for automatic exploration of data to generate useful information.
3.	A DBMS query will suffice if it is exactly known what information user is seeking.	Data mining techniques are useful if possible correlations or patterns are vaguely known.
4.	A DBMS is complete system used for managing digital databases that allows organization, storage and retrieval of database content.	Data mining is generally used for hypothesis testing.
5.	DBMS include four important elements – the modeling language, data structures, query language and mechanism for transactions.	Data mining algorithms are based on statistical techniques.

Origins of Data Mining

- Draws ideas from machine learning/AI, pattern recognition, statistics, and database systems
- Must address:
  - Enormity of data
  - High dimensionality of data
  - Heterogeneous, distributed nature of data

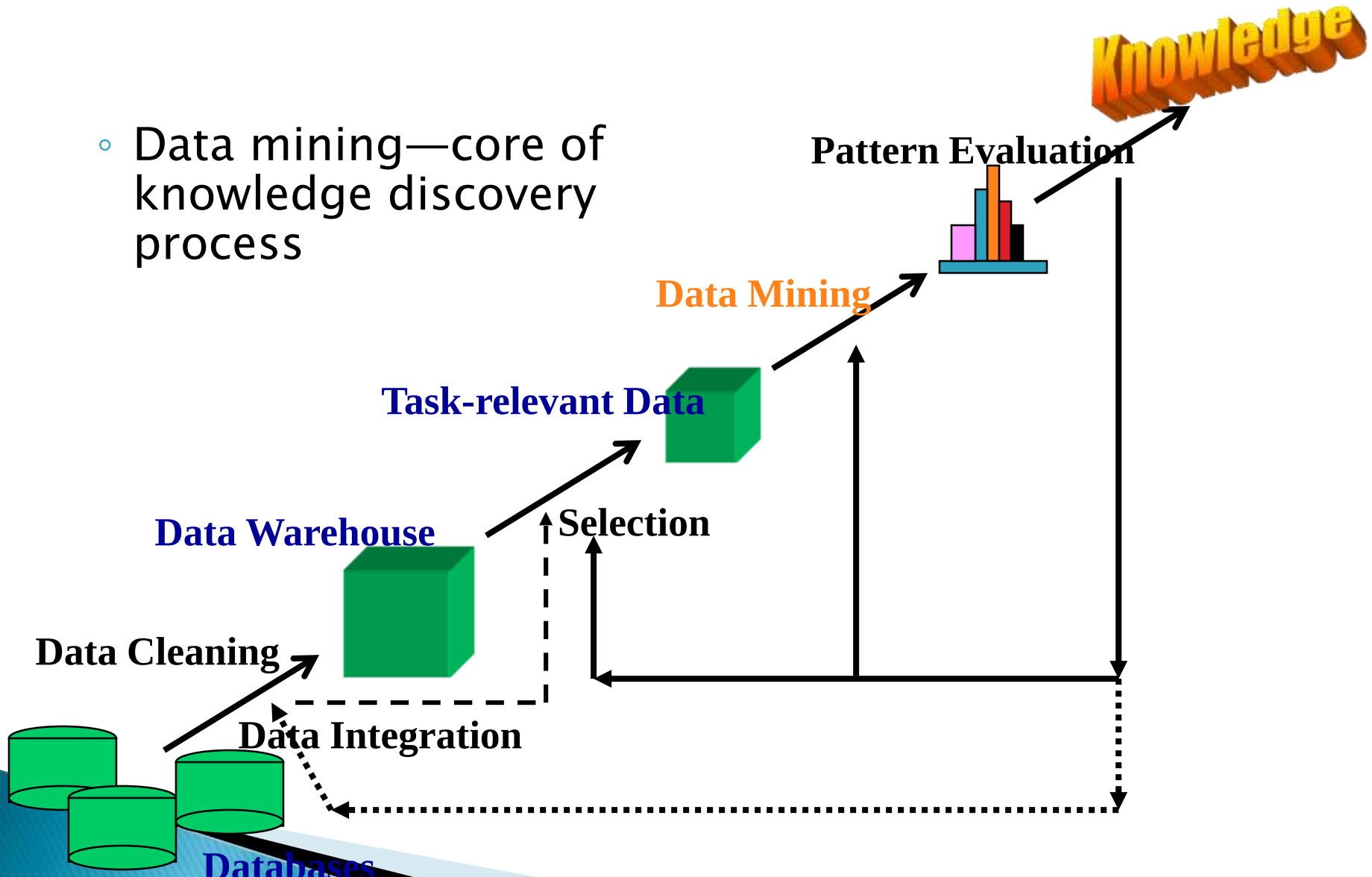


Knowledge Discovery in Databases

- Knowledge Discovery in Databases (KDD) is the process of identifying a valid, potentially useful and ultimately understandable structure in data.
- Multidisciplinary activities.
- Data storage, data cleansing, algorithms to handle massive data sets, and interpreting results.

Knowledge Discovery in Databases

- Data mining—core of knowledge discovery process



Steps of a KDD Process

1. **Goal Identification:** Identify the goal of the KDD process from the customer's perspective.
2. **Data Selection:** Select a target data set or subset of data samples on which knowledge discovery is to be performed.
3. **Data Cleaning:** Cleanse and preprocess data by deciding strategies to handle missing fields and alter the data as per requirements.
4. **Transformation:** Simplify the data sets by removing unwanted variables. Then, analyze useful features that can be used to represent the data, depending on the goal.

Steps of a KDD Process

- **Data Mining**
- Choose functions of data mining
  - Summarization, classification, regression, association, clustering.
- Choose the mining algorithm(s) and decide the parameters.
- **Pattern Evaluation and Interpretation:** Search for patterns of interest in a particular representational form. Essential knowledge from the mined patterns is interpreted.
- **Knowledge Acquisition:** Use the knowledge, document it and make reports for interested parties.

Data Mining: On What Kinds of Data?

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

Integrating a Data Mining System with a DB/DW System

- There are three different ways in which data mining system integrate with data source:

1. No Coupling
2. Loose Coupling
3. Tight Coupling

1. No Coupling

- In this scheme, the data mining system does not utilize any of the database or data warehouse functions.
- It fetches the data from a particular source and processes into its memory and storage system.
- Then data mining is performed on that data.
- The data mining result is stored in another file.
- Merit:
  - The memory management can be optimized specific to the data mining algorithm.
- Limitation:
  - These systems lack the field proven technologies of DBMS such as recovery, concurrency, and so on.

2. Loose Coupling

- DBMS is used only for storage and retrieval of data.
- It fetches the data from the data respiratory managed by DBMS into the application address space.
- Then data mining is performed on that data.
- It then stores the mining result either in a file or in a designated place in a database or in a data warehouse.
- It does not use query capability provided by DBMS.

3. Tight Coupling

- The data mining system is smoothly integrated into the database or data warehouse system.
- The data mining subsystem is treated as one functional component of an information system.
- Merit:
  - The data mining application goes where data naturally resides, this avoid performance degradation and takes full advantage of database technology.

Example Data Sources

- Point – of- sale
- Credit card charge records
- Warranty claims
- Medical insurance claims
- Direct mail response data

© IKSIN C

Example Data Sources

- Telephone call records
- Web activity data
- Economic data
- Utility charges
- Census returns
- Magazine subscription records

© IKSIN C

Why Data Preprocessing?

- Data in the real world is dirty
  - incomplete: lacking *attribute values*, lacking certain *attributes of interest*, or containing only aggregate data
  - noisy: containing errors or outliers
  - inconsistent: containing discrepancies in codes or names
- No quality data, no quality mining results!
  - Quality decisions must be based on quality data
  - Data warehouse needs consistent integration of quality data
  - Required for both OLAP and Data Mining!

Data Cleaning and Preprocessing Stage

- Data comes from many sources -- internal and external
- Data comes in many forms and formats
  - Hierarchical databases, flat files, COBOL data sets
- Data is never clean
- Most important stage. Typically consumes about 60 - 80% of the total data mining effort

© IKSIN C

Business Data Corruption Examples

- Duplication - A common problem with direct mailers and credit card companies
- Missing and Confusing Data Fields
- Outliers -Generally present due to incorrect entry/coding of a data field.

Study above indicates that up to 30% of your CRM data could be wrong.

© IKSIN C

Data Preprocessing

This step is also known as *data transformation*. The aim here is to map data fields into representations suitable for the data discovery stage.

Examples:

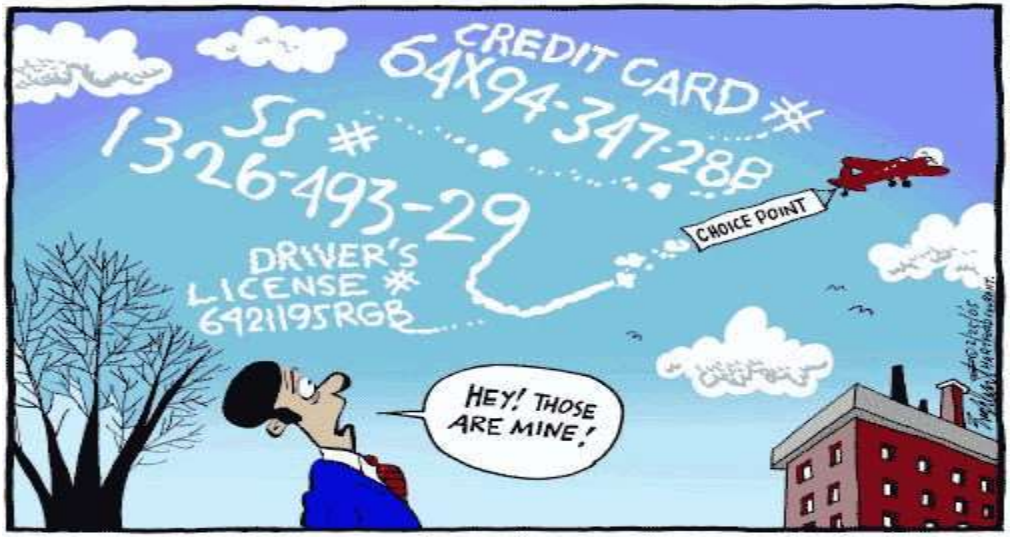
Month/Date/Year ===> Age Groups

Customer Address ===> Geographic Zone Code

© IKSIN C

Pattern Discovery Stage

- Discovery Model?
- Discovery Methodology?



© IKSINC

Discovery Models

- Association Model
- Classification Model
- Clustering Model
- Sequential Model
- Visual Model

"The computer predicted that I would have a sandwich for lunch. I ate cake!"

© IKSIN C

Association Model

90% of customers who subscribe to at least three premium channels also subscribe to pay-per-view events

Also known as Market Basket Analysis

© IKSIN C

Association Model: Application 1

- Marketing and Sales Promotion:
  - Let the rule discovered be

    {Bagels, … } ⇒{Potato Chips}
  - Potato Chips as consequent => Can be used to determine what should be done to boost its sales.
  - Bagels in the antecedent => can be used to see which products would be affected if the store discontinues selling bagels.
  - Bagels in antecedent *and* Potato chips in consequent => Can be used to see what products should be sold with Bagels to promote sale of Potato chips!

Association Model: Application 2

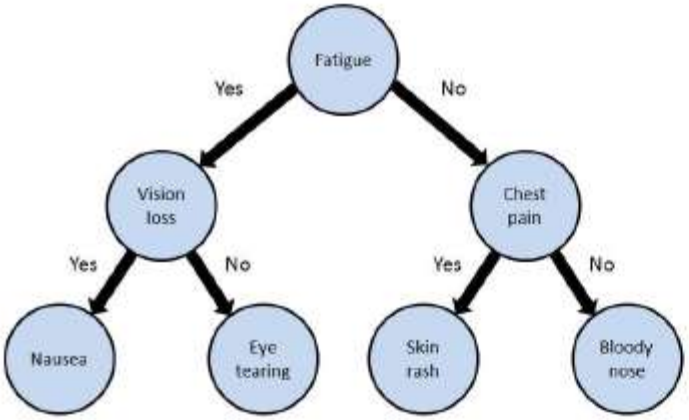
- Supermarket shelf management.
  - Goal: To identify items that are bought together by sufficiently many customers.
  - Approach: Process the point-of-sale data collected with bar-code scanners to find dependencies among items.
  - A classic rule -
    - If a customer buys diaper and milk, then he is very likely to buy beer.
    - So, don’t be surprised if you find six-pack stacked next to diapers!

Classification Model

- If *Annual_Income >40,000* **AND** *Home-Owner*, Then *Credit-Risk* ➔ *Medium*

- If $\frac{(Annual\_Income)^{1.2}}{(Avg\_Monthly\_CreditCardBalance + Mortgage)^{1.5}} > 25$,

  Then *Loan-Approval* ➔ *Yes*



"I'd like to diversify my portfolio. For a change, why don't you get me a stock that's not in a group I like to classify as 'losers?'"

© IKSIN C

Classification: Application 2

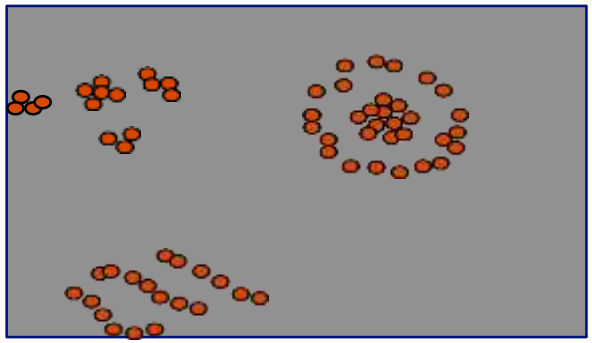
- Fraud Detection
  - Goal: Predict fraudulent cases in credit card transactions.
  - Approach:
    - Use credit card transactions and the information on its account – holders as attributes.
      - When does a customer buy, what does he buy, how often he pays on time, etc
    - Label past transactions as fraud or fair transactions. This forms the class attribute.
    - Learn a model for the class of the transactions.
    - Use this model to detect fraud by observing credit card transactions on an account.

Classification: Application 3

- Customer Attrition/Churn:
  - Goal: To predict whether a customer is likely to be lost to a competitor.
  - Approach:
    - Use detailed record of transactions with each of the past and present customers, to find attributes.
      - How oqen the customer calls, where he calls, what time – of – the day he calls most, his financial status, marital status, etc.
    - Label the customers as loyal or disloyal.
    - Find a model for loyalty.

Clustering Model

Clustering models are similar to classification models except that no a-priori information is available for classes.



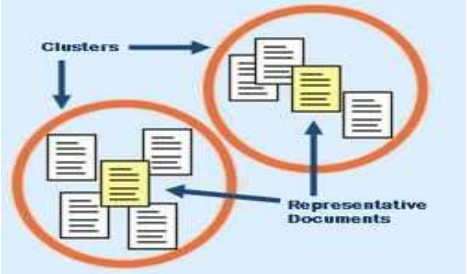
© IKSIN C

Clustering: Application 1

- Market Segmentation:
  - Goal: subdivide a market into distinct subsets of customers where any subset may conceivably be selected as a market target to be reached with a distinct marketing mix.
  - Approach:
    - Collect different attributes of customers based on their geographical and lifestyle related information.
    - Find clusters of similar customers.
    - Measure the clustering quality by observing buying patterns of customers in same cluster vs. those from different clusters.

Clustering: Application 2

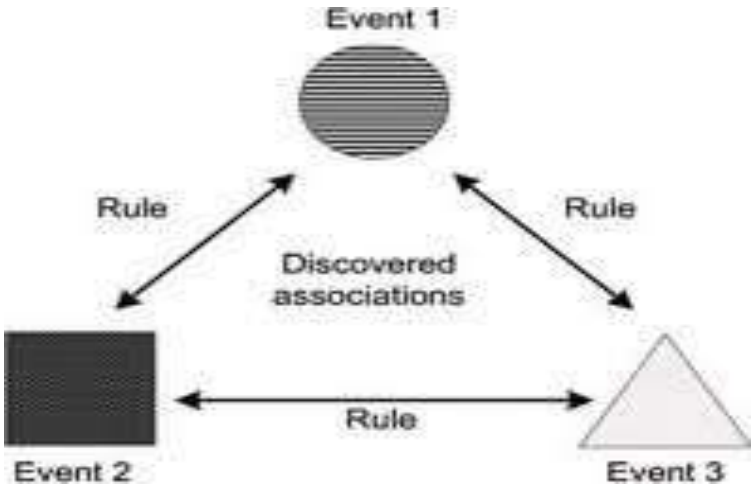
- Document Clustering:
  - Goal: To find groups of documents that are similar to each other based on the important terms appearing in them.
  - Approach: To identify frequently occurring terms in each document. Form a similarity measure based on the frequencies of different terms. Use it to cluster.
  - Gain: Information Retrieval can utilize the clusters to relate a new document or search term to clustered documents.



Sequential Model

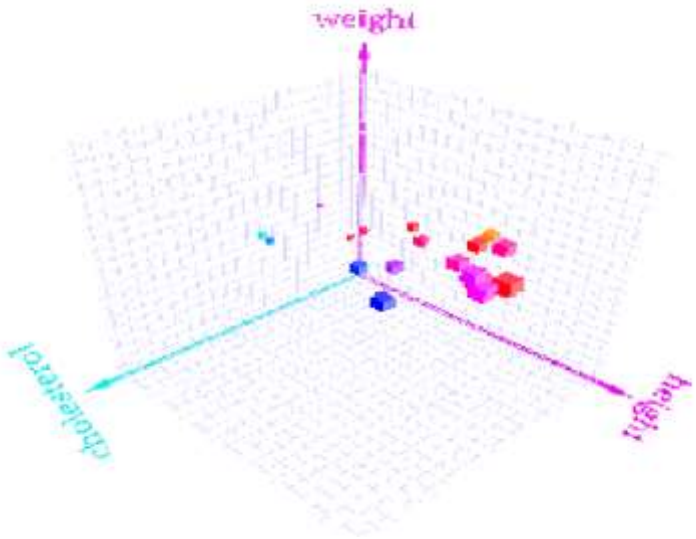
Similar to association models except that sequences of events are considered. For example:

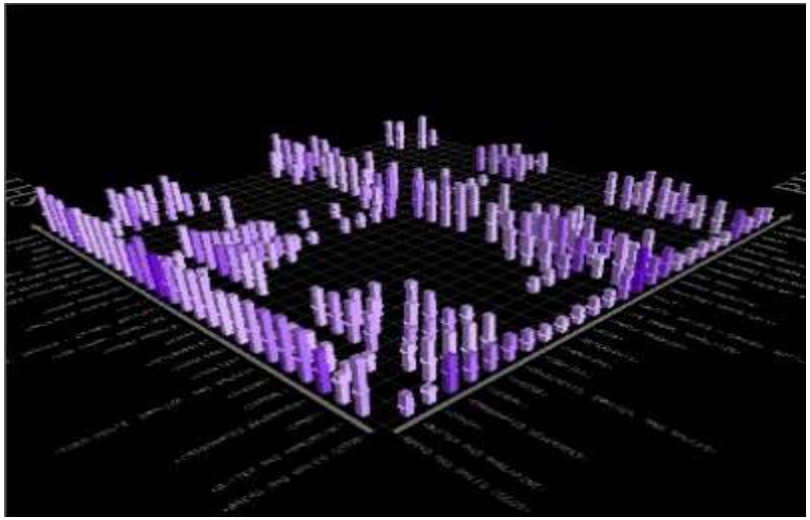
"80% of customers who buy a product X are likely to buy product Y in next six months"



© IKSIN C

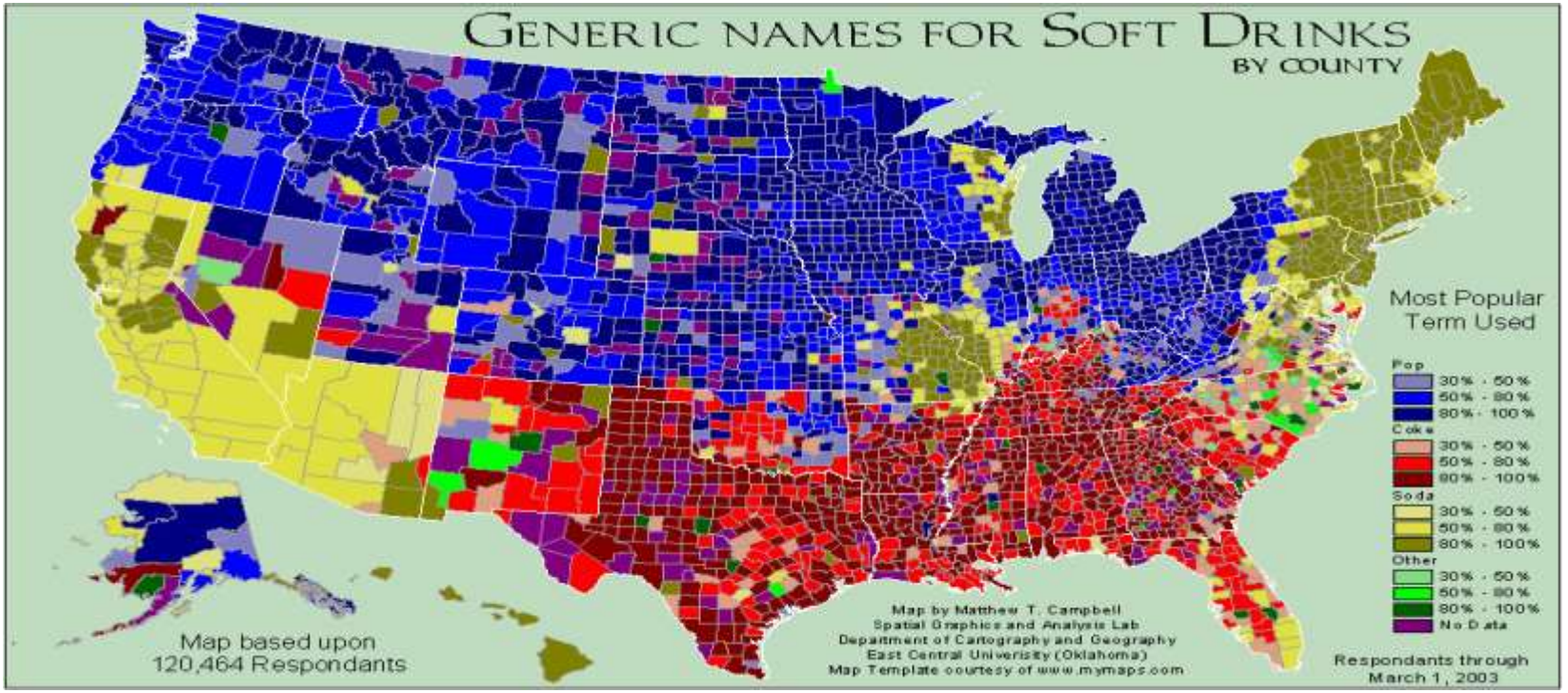
Visual Model





© IKSIN C

Geographical Patterns and Map Visualization



© IKSINC

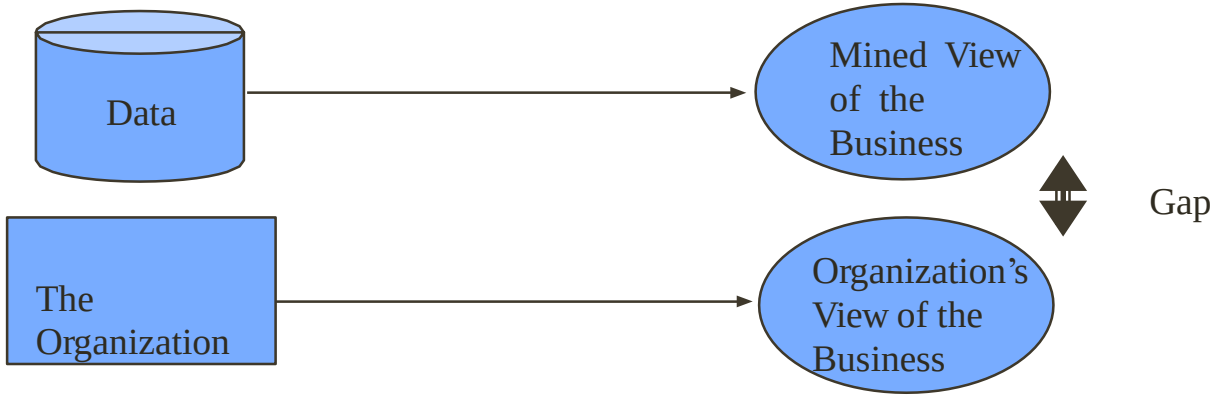
Interpretation Stage

- Evaluate the quality of the discovered patterns
- Determine the value of the discovery to the business

© IKSIN C

Value of Mined Information

- Perceptive gap



© IKSIN C

Reporting Stage

- Reporting the discovery to higher management
- Transforming the discovery to new actions or products

© IKSIN C

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

Interestingness of Patterns Contd..

- **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.

Data Mining Primitives

- A data mining task can be specified in the form of a data mining query, which is input to the data mining system.
- A data mining query is defined in terms of data mining task primitives.
- These primitives allow the user to interactively communicate with the data mining system during discovery in order to direct the mining process, or examine the findings from different angles or depths.

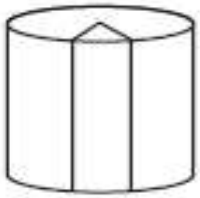
Data Mining Primitives

- The set of *task-relevant data to be mined: This specifies the portions of the database* or the set of data in which the user is interested.
- The *kind of knowledge to be mined: This specifies the data mining functions to be per-* formed, such as characterization, discrimination, association or correlation analysis, classification, prediction, clustering, outlier analysis, or evolution analysis.

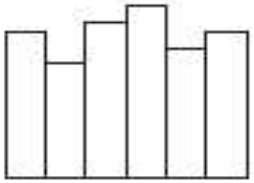
Data Mining Primitives

- The *background knowledge to be used in the discovery process:*
- The *interestingness measures and thresholds for pattern evaluation: They may be used* to guide the mining process or, after discovery, to evaluate the discovered patterns.
- The expected *representation for visualizing the discovered patterns: This refers to the f*orm in which discovered patterns are to be displayed, which may include rules, tables, charts, graphs, decision trees, and cubes.

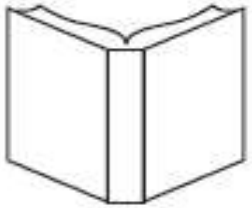
Data Mining Primitive Tasks



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

Applications

In finance, telecom, insurance and retail:

- Loan/credit card approval
- market segmentation
- fraud detection
- better marketing
- trend analysis
- market basket analysis
- customer churn
- Web site design and promotion

Loan/Credit card approvals

In a modern society, a bank does not know its customers. Only knowledge a bank has is their information stored in the computer.

Credit agencies and banks collect a lot of customers' behavioural data from many sources. This information is used to predict the chances of a customer paying back a loan.

Market Segmentation

- Large amounts of data about customers contains valuable information
- The market may be segmented into many subgroups according to variables that are good discriminators
- Not always easy to find variables that will help in market segmentation

Market Analysis and Management

- Where does the data come from?
  - Credit card transactions, discount coupons, customer complaint calls
- 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

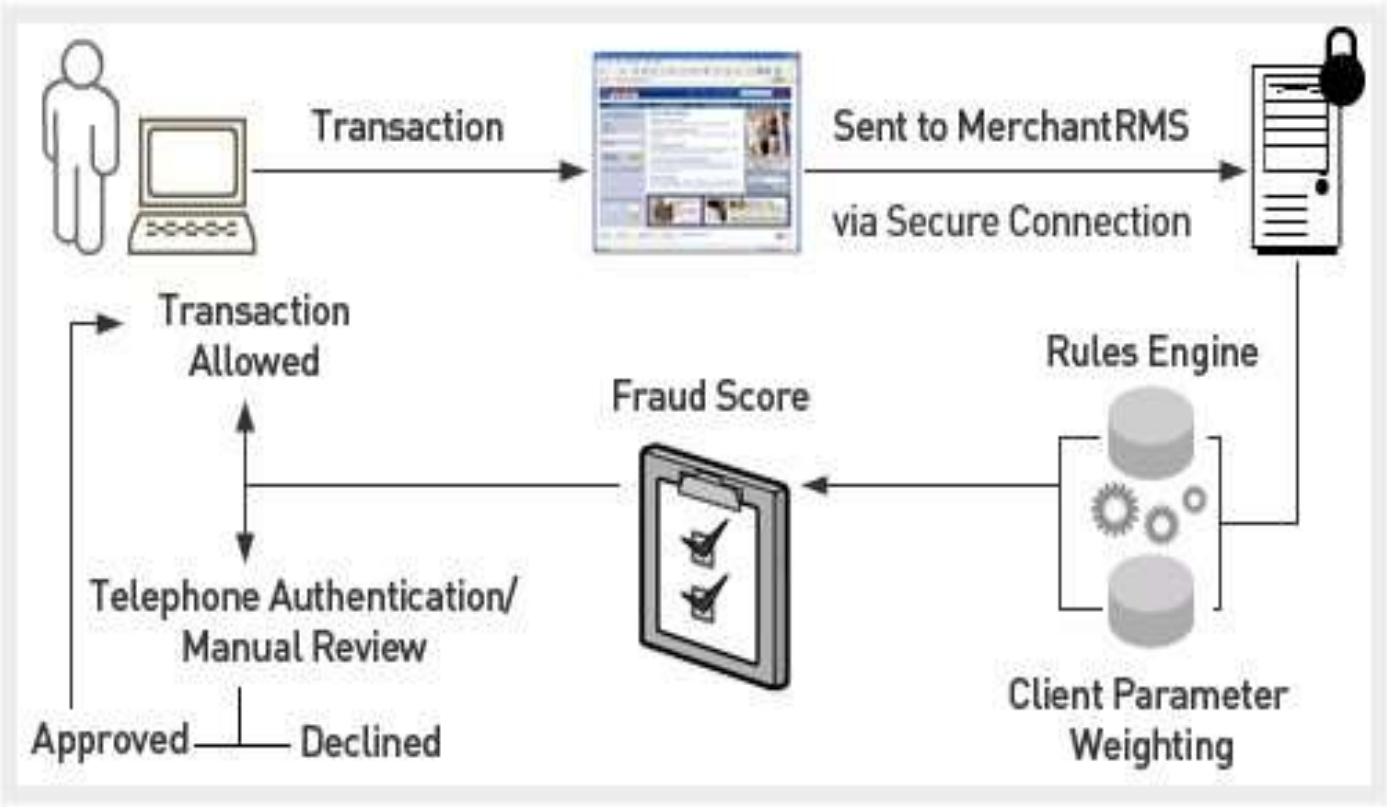
- Cross-market analysis
  - Associations/co-relations between product sales, & prediction based on such association
- Customer profiling
  - What types of customers buy what products
- Customer requirement analysis
  - Identifying the best products for different customers
  - Predict what factors will attract new customers

Fraud Detection

- Very challenging since it is difficult to define characteristics of fraud. Often based on detecting changes from the norm.
- In statistics, it is common to throw out the outliers but in data mining it may be useful to identify them since they could either be due to errors or perhaps fraud.

Fraud Detection & Mining Unusual Patterns

- Approaches: Clustering & model construction for frauds, outlier analysis
- Applications: Health care, retail, credit card service, telecomm.
  - Medical insurance
    - Professional patients, and ring of doctors
    - Unnecessary or correlated screening tests
  - Telecommunications:
    - 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


Transaction
Sent to MerchantRMS
via Secure Connection
Transaction
Allowed
Rules Engine
Fraud Score
Telephone Authentication/
Manual Review
Client Parameter
Weighting
Approved
Declined

Better Marketing

When customers buy new products, other products may be suggested to them when they are ready.

As noted earlier, in mail order marketing for example, one wants to know:

- will the customer respond?
- will the customer buy and how much?
- will the customer return purchase?
- will the customer pay for the purchase?

Better Marketing

It has been reported that more than 1000 variable values on each customer are held by some mail order marketing companies.

The aim is to “lift” the response rate.

Trend analysis

In a large company, not all trends are always visible to the management. It is then useful to use data mining software that will identify trends.

Trends may be long term trends, cyclic trends or seasonal trends.

Market Basket Analysis

- Aims to find what the customers buy and what they buy together
- This may be useful in designing store layouts or in deciding which items to put on sale
- Basket analysis can also be used for applications other than just analysing what items customers buy together

Customer Churn

- In businesses like telecommunications, companies are trying very hard to keep their good customers and to perhaps persuade good customers of their competitors to switch to them.
- In such an environment, businesses want to find which customers are good, why customers switch and what makes customers loyal.
- Cheaper to develop a retention plan and retain an old customer than to bring in a new customer.

Customer Churn

- The aim is to get to know the customers better so you will be able to keep them longer.
- Given the competitive nature of businesses, customers will move if not looked after.
- Also, some businesses may wish to get rid of customers that cost more than they are worth e.g. credit card holders that don't use the card, bank customers with very small amount of money in their accounts.

Web site design

- A Web site is effective only if the visitors easily find what they are looking for.
- Data mining can help discover affinity of visitors to pages and the site layout may be modified based on this information.

Other Applications

- 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.

Challenges of Data Mining

- Scalability
- Dimensionality
- Complex and Heterogeneous Data
- Data Quality
- Data Ownership and Distribution
- Privacy Preservation

Hey, wanna buy some data? Only slightly used! Heap of clicks left in it. Its last owner was a little old lady who only used it for shopping ...

© IKSIN C

Major Issues and Challenges 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

Major Issues and Challenges in Data Mining

- 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

Data Mining Query Language

- A data mining query language can be designed to incorporate data mining primitives, allowing users to flexibly interact with data mining systems.
- Having a data mining query language provides a foundation on which user-friendly graphical interfaces can be built.

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

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                                    (1)
use hierarchy location_hierarchy for T.branch, age_hierarchy for C.age   (3)
mine classification as promising_customers                        (2)
in relevance to C.age, C.income, I.type, I.place_made, T.branch   (1)
from customer C, item I, transaction T                            (1)
where I.item_ID = T.item_ID and C.cust_ID = T.cust_ID
      and C.income ≥ 40,000 and I.price ≥ 100                     (1)
group by T.cust_ID                                                (2)
having sum(I.price) ≥ 1,000                                       (1)
display as rules                                                  (5)
```

OLAP Mining: Integration of Data Mining and Data Warehousing

- Data mining systems, DBMS, Data warehouse systems 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.
- Integration of multiple mining functions
  - Characterized classification, first clustering and then association

DBMS, OLAP, and Data Mining

	DBMS	OLAP	Data Mining
Task	Extraction of detailed and summary data	Summaries, trends and forecasts	Knowledge discovery of hidden patterns and insights
Type of result	Information	Analysis	Insight and Prediction
Method	Deduction (Ask the question, verify with data)	Multidimensional data modeling, Aggregation, Statistics	Induction (Build the model, apply it to new data, get the result)
Example question	Who purchased mutual funds in the last 3 years?	What is the average income of mutual fund buyers by region by year?	Who will buy a mutual fund in the next 6 months and why?

Data Mining Software

- Weka
- KNIME
- RapidMiner
- Spmf
- And many more..

ANY QUESTIONS ?

THANK YOU