Data Mining and Data Warehousing

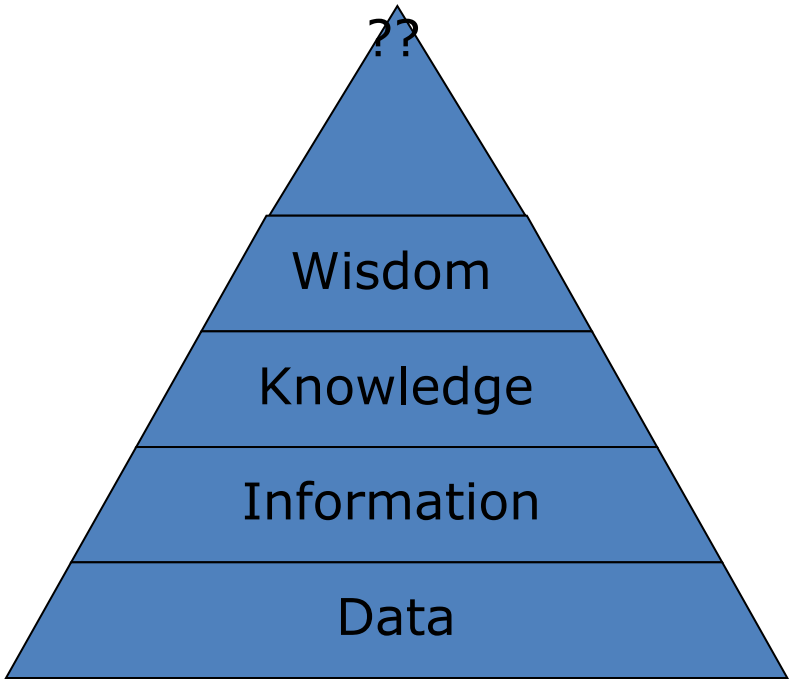
B.E. Computer

8th Semester

Unit 1 : Introduction to Data Mining and Data Warehousing

What is Data?

- A representation of facts, concepts, or instructions in a formal manner suitable for communication, interpretation, or processing by human beings or by computers.



Review of basic concepts of data warehousing and data mining

- The Explosive Growth of Data: from terabytes to petabytes
- Data accumulate and double every 9 months
- High-dimensionality of data
- High complexity of data
- New and sophisticated applications
- There is a big gap from stored data to knowledge; and the transition won't occur automatically.
- Manual data analysis is not new but a bottleneck
- Fast developing Computer Science and Engineering generates new demands

Very Large Databases

- Terabytes -- 10^{12} bytes: Walmart -- 24 Terabytes
- Petabytes -- 10^{15} bytes: Geographic Information Systems
- Exabytes -- 10^{18} bytes: National Medical Records
- Zettabytes -- 10^{21} bytes: Weather images
- Zottabytes -- 10^{24} bytes: Intelligence Agency Videos

Data explosion problem

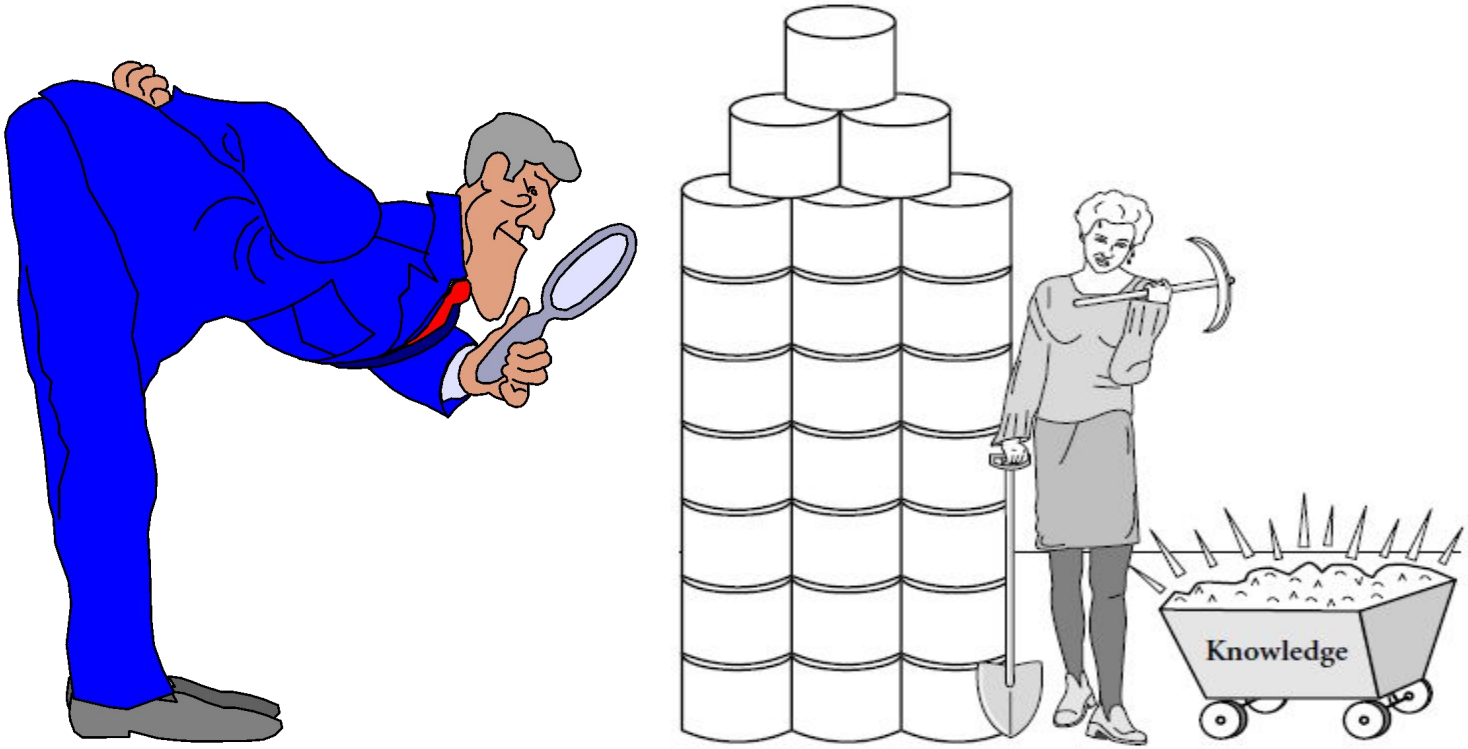
Automated data collection tools and mature database technology lead to tremendous amounts of data accumulated and/or to be analyzed in databases, data warehouses, and other information repositories

We are drowning in data, but starving for knowledge!

Solution:

"Necessity is the mother of invention"—**Data Warehousing and Data Mining**

What is Data Mining?



Art/Science of extracting non-trivial, implicit, previously unknown, valuable, and potentially Useful information from a large database

Data mining is

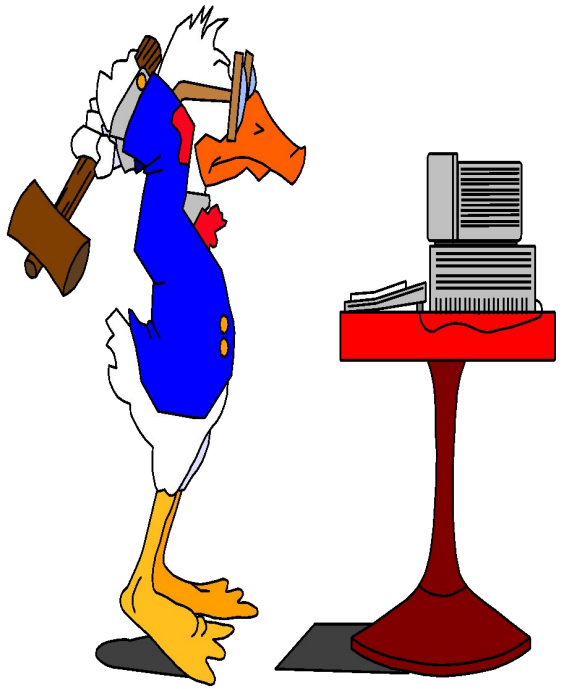
- "Extraction of interesting information or patterns from data in large databases is known as data mining."
- A hot buzzword for a class of techniques that find patterns in data
- A user-centric, interactive process which leverages analysis technologies and computing power
- A group of techniques that find relationships that have not previously been discovered
- A relatively easy task that requires knowledge of the business problem/subject matter expertise

- Data mining is a logical process that is used to search through large amount of data in order to find useful data.
- The goal of this technique is to find patterns that were previously unknown.
- Once these patterns are found they can further be used to make certain decisions for development of their businesses.
- Three steps involved are:
  - Exploration
  - Pattern identification
  - Deployment

- ***Exploration:*** In the first step of data exploration data is cleaned and transformed into another form, and important variables and then nature of data based on the problem are determined.
- ***Pattern identification:*** Once data is explored, refined and defined for the specific variables the second step is to form pattern identification. Identify and choose the patterns which make the best prediction.
- ***Deployment:*** Patterns are deployed for desired outcome.
- Simply, data mining refers to extracting or "mining" knowledge from large amounts of data.
- "Mining" is a vivid term characterizing the process that finds a small set of precious nuggets from a great deal of raw material.

Data mining is not

- Brute-force crunching of bulk data
- “Blind” application of algorithms
- Going to find relationships where none exist
- Presenting data in different ways
- A difficult to understand technology requiring an advanced degree in computer science
- Searching a phone number in a phone book
- Searching a keyword on Google
- Generating histograms of salaries for different age groups
- Issuing SQL query to a database and reading the reply



Data mining is not

- A cybernetic magic that will turn your data into gold. It's the process and result of knowledge production, knowledge discovery and knowledge management.
- Once the patterns are found Data Mining process is finished.
- Queries to the database are not DM.

Why use Data Mining today?

Because it can improve customer service, better target marketing campaigns, identify high-risk clients, and improve production processes.

Data mining has been used to:

- Identify unexpected shopping patterns in supermarkets.
- Optimize website profitability by making appropriate offers to each visitor.
- Predict customer response rates in marketing campaigns.
- Defining new customer groups for marketing purposes.
- Predict customer defections: which customers are likely to switch to an alternative supplier in the near future.
- Distinguish between profitable and unprofitable customers.

- 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)
    - telecommunications, financial, insurance industries
- Other Applications
  - Text mining (news group, email, documents) and Web mining
  - Stream data mining
  - Bioinformatics and bio-data analysis
- Product development - biotechnology, pharmaceutical industry
- Entertainment - digital convergence, sports
- Diagnosis and monitoring - medical, aerospace, automotive.

APPLICATIONS OF DATA MINING

- Data mining has many and varied fields of application some of which are listed below:
- **Sales/Marketing**
  - Identify buying patterns from customers
  - Find associations among customer demographic characteristics
  - Predict response to mailing campaigns
  - Market basket analysis.
- **Banking**
  - Credit card fraudulent detection
  - Identify 'loyal' customers
  - Predict customers likely to change their credit card affiliation
  - Determine credit card spending by customer groups
  - Find hidden correlation's between different financial indicators
  - Identify stock trading rules from historical market data

- **Insurance and Health Care**
  - Claims analysis i.e., which medical procedures are claimed together
  - Predict which customers will buy new policies
  - Identify behavior patterns of risky customers
  - Identify fraudulent behavior
- **Transportation**
  - Determine the distribution schedules among outlets
  - Analyze loading patterns
- **Medicine**
  - Characterize patient behavior to predict office visits
  - Identify successful medical therapies for different illnesses

DISADVANTAGES OF DATA MINING

Privacy issues

- The concerns about the personal privacy have been increasing enormously recently especially when internet is booming with social networks, e-commerce, forums, blogs etc.
- Because of privacy issues, people are afraid of their personal information is collected and used in unethical way that potentially causing them a lot of trouble.

Security issues

- Businesses own information about their employee and customers including social security number, birthday, payroll and etc.

DISADVANTAGES OF DATA MINING

- There have been a lot of cases that hackers were accesses and stole big data of customers from big corporation such as Ford Motor credit company, Sony, etc. with so much personal and financial information available, the credit card stolen and identity theft become a big problem.

Misuse of Information/ inaccurate Information

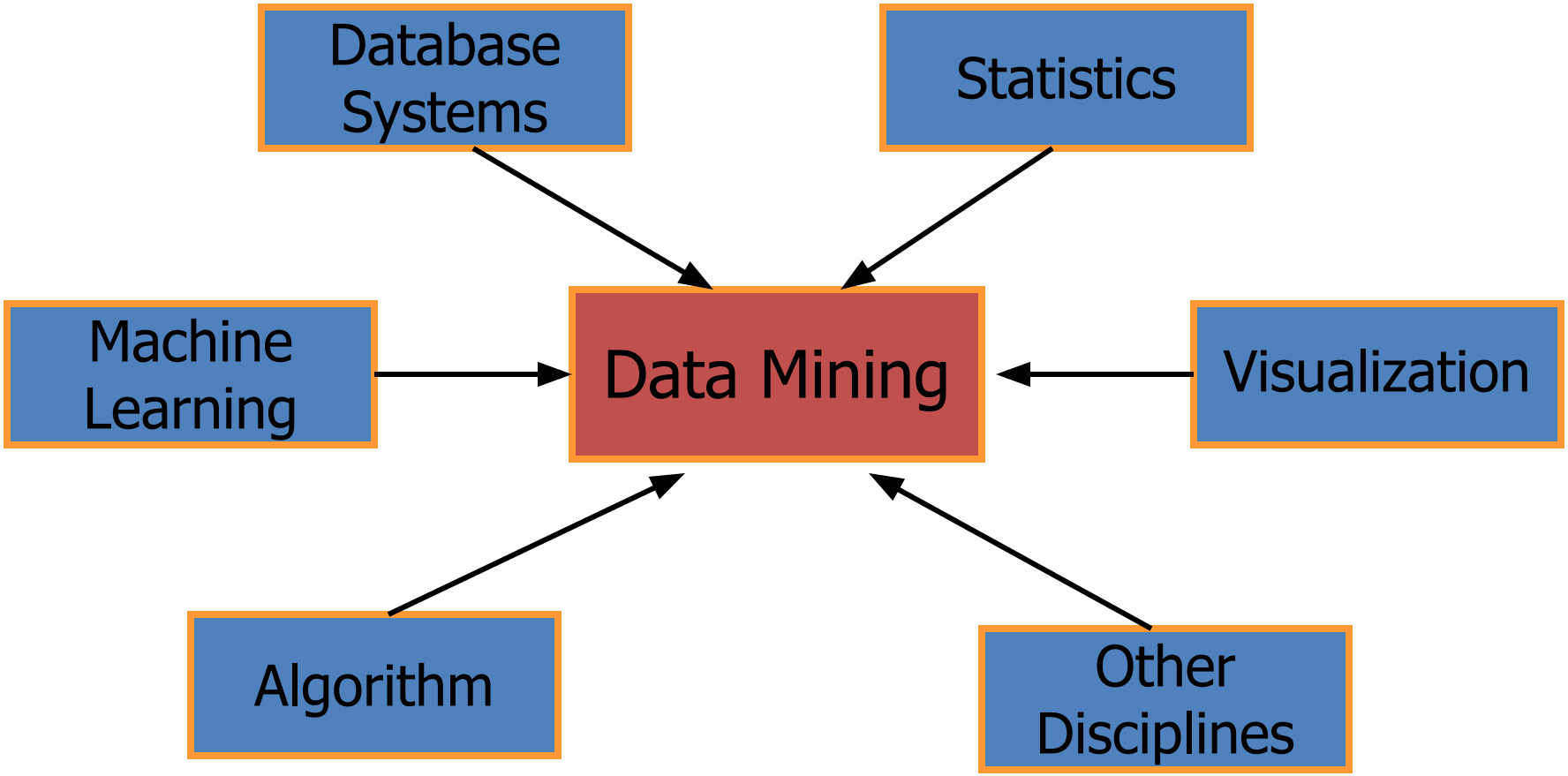
- Information collected through data mining intended for marketing or ethical purposes can be misused.
- This information is exploited by unethical people or business to take benefit of vulnerable people or business to take benefit of vulnerable people or discriminate against a group of people.

Functions of Data mining

- Data mining has five main functions:
- ***Classification:*** infers the defining characteristics of a certain group (such as customers who have been lost to competitors).
- ***Clustering:*** identifies groups of items that share a particular characteristic. (Clustering differs from classification in that no predefining characteristic is given in classification.)
- ***Association:*** identifies relationships between events that occur at one time (such as the contents of a shopping basket).
- ***Sequencing:*** similar to association, except that the relationship exists over a-period of time (such as repeat visits to a supermarket or use of a financial planning product).
- ***Forecasting:*** estimates future values based on patterns within large sets of data (such as demand forecasting).

Data Mining: Confluence of Multiple Disciplines

Data mining is a combination of multidisciplinary field. It can be applied in many fields and can be done using many algorithm and techniques.



Data Mining Vs. Query Tools

- SQL can find normal queries from the database such as what is an average turnover? Whereas data mining tools find interesting patterns and facts such as what are the important trends in sells?
- Data mining is much more faster than SQL in trend and pattern analysis since it uses algorithm like machine learning, genetic algorithm.
- If we know exactly what we are looking for, we use SQL nut if we know only vaguely what we are looking for we use data mining.
- Hybrid information can’t be easily be traced using SQL.

Knowledge Discovery in Databases (KDD) Process

- Many people treat data mining as a synonym for another popularly used term, Knowledge Discovery in Database, or KDD.
- Alternatively, others view data mining as simply an essential step in the process of knowledge discovery.
- Data mining, or knowledge discovery in databases (KDD) as it is also known, is the nontrivial extraction of implicit, previously unknown, and potentially useful information from data.

Some Alternative names to data mining are:

- Knowledge discovery (mining) in databases (KDD)
- Knowledge extraction
- Data/pattern analysis
- Data archeology
- Data Dredging
- Information Harvesting
- Business intelligence, etc.

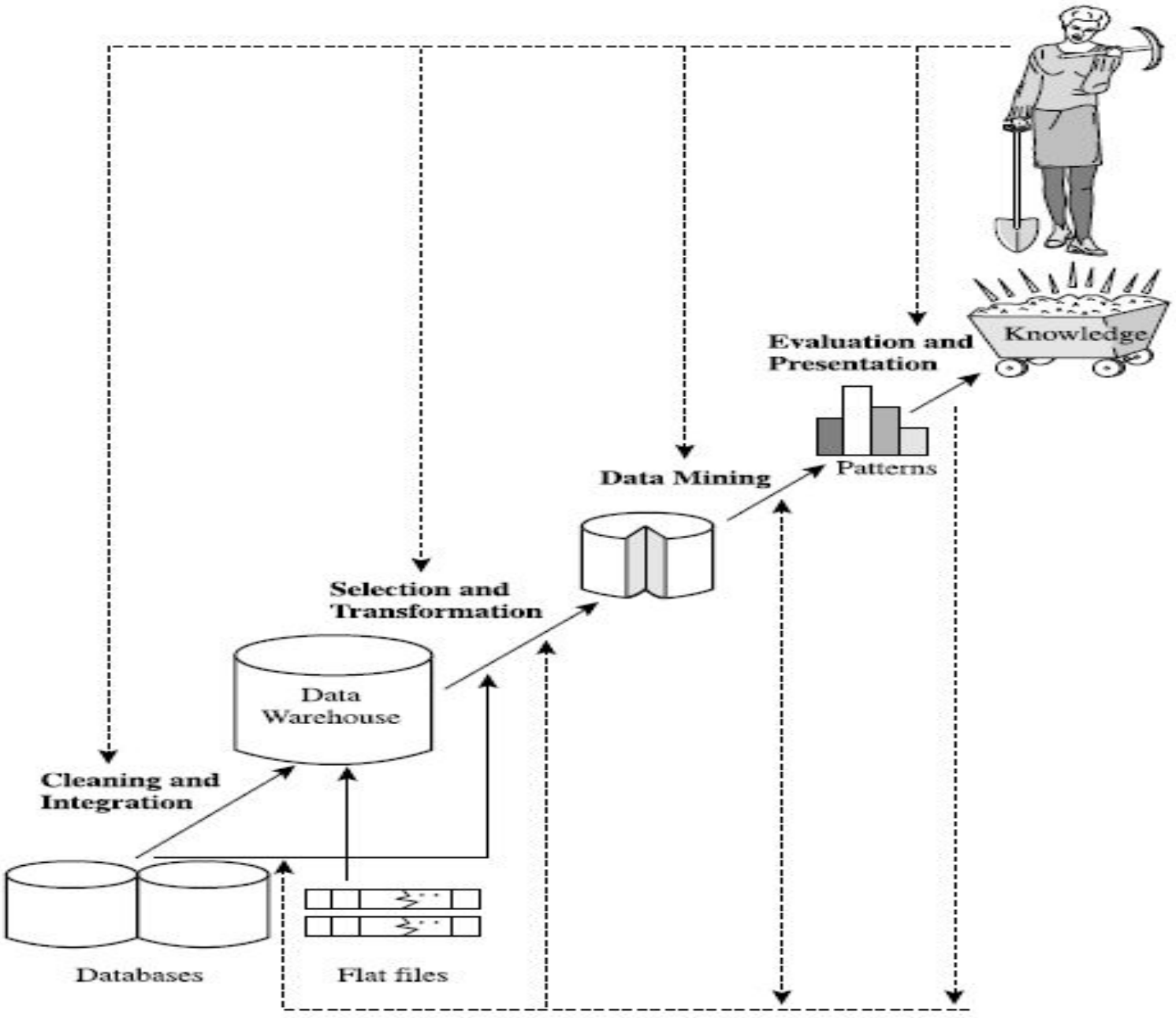


Figure: Data mining as a step in the process of knowledge discovery.

THE PROCESS OF KNOWLEDGE DISCOVERY

- The main steps of knowledge discovery process are:
  - Identify business Problem
  - Data mining
  - Action
  - Evaluation and measurement
  - Deployment and integration into businesses processes.

- **Data cleaning** (to remove noise or irrelevant data):
  - Data cleaning is the process of ensuring that, for data mining purposes, the data is uniform in terms of key and attributes usage.
  - Data cleaning is separate from data enrichment and data transformation because data cleaning attempts to correct misused or incorrect attributes in existing data.
  - An important element in a cleaning operation is the de-duplication of records.
- **Data integration** (where multiple data sources may be combined)
- **Data selection:** There are two parts to selecting data for data mining:
  - *locating data*
  - *identifying data*

- **Data enrichment:**
  - Data enrichment is the process of adding new attributes, such as calculated fields or data from external sources, to existing data.
  - Most references on data mining tend to combine this step with data transformation.
  - Data transformation involves the manipulation of data, but data enrichment involves adding information to existing data.
  - This can include combining internal data with external data, obtained from either different departments or companies or vendors that sell standardized industry-relevant data.

- **Data transformation:**
  - Data transformation, in terms of data mining, is the process of changing the form or structure of existing attributes.
  - Data transformation is separate from data cleansing and data enrichment for data mining purposes because it does not correct existing attribute data or add new attributes, but instead grooms existing attributes for data mining purposes.
- **Data mining:**
  - The data mining step may interact with the user or a knowledge base.
  - The interesting patterns are presented to the user, and may be stored as new knowledge in the knowledge base.
  - Data mining is the process of discovering interesting knowledge from large amounts of data stored either in databases, data warehouses, or other information repositories.

- **Pattern evaluation:**
- Pattern evaluation is used to identify the truly interesting patterns representing knowledge based on some interestingness measures.
- **Knowledge presentation** (where visualization and knowledge representation techniques are used to present the mined knowledge to the user).
- ***Visualization techniques*** are a very useful method of discovering patterns in datasets, and may be used at the beginning of a data mining process to get a rough feeling of the quality of the data set and where patterns are to be found.
- Scatter diagrams can be used to identify interesting subsets of the data sets so that we can focus on the rest of the data mining process.

- The steps involved in data mining when viewed as a process of knowledge discovery are as follows:
- **Data cleaning,** a process that removes or transforms noise and inconsistent data.
- **Data integration,** where multiple data sources may be combined.
- **Data selection,** where data relevant to the analysis task are retrieved from the database.

- **Data transformation,** where data are transformed or consolidated into forms appropriate for mining.
- **Data mining,** an essential process where intelligent and efficient methods are applied in order to extract patterns.
- **Pattern evaluation,** a process that identifies the truly interesting patterns representing knowledge, based on some interestingness measures.
- **Knowledge presentation,** where visualization and knowledge representation techniques are used to present the mined knowledge to the user.

WHAT ARE THE ISSUES IN DATA MINING?

- **Security and social issues:** Security is an important issue with any data collection that is shared and/or is intended to be used for strategic decision-making.
- **User interface issues:** The knowledge discovered by data mining tools is useful as long as it is interesting, and above all understandable by the user. The major issues related to user interfaces and visualization are "screen real-estate", information rendering, and interaction.

- **Mining methodology issues:** These issues are relevant to the data mining approaches applied and their limitations.
- **Performance issues:** raises the issues of scalability and efficiency of the data mining methods when processing considerably large data.
- **Data source issues:** We are storing different types of data in a variety of repositories. It is difficult to expect a data mining system to effectively and efficiently achieve good mining results on all kinds of data and sources.

Thank you !!!