Data Mining

- Data mining is a technique that discovers previously unknown relationships in data.
- Data mining is the practice of automatically searching large stores of data to discover patterns and trends that go beyond simple analysis. Data mining uses sophisticated mathematical algorithms to segment the data and to predict the likelihood of future events based on past events. Data mining is also known as Knowledge Discovery in Data (KDD).

The key properties of Data Mining are:

- Automatic discovery of patterns
- Prediction of likely outcomes
- Creation of actionable information
- Focus on large data sets and databases
- Data mining can answer questions that cannot be addressed through simple query and reporting techniques.

Data Mining and Data Warehousing

- Data can be mined whether it is stored in flat files, spreadsheets, database tables, or some other storage format. The important criteria for the data is not the storage format, but its applicability to the problem to be solved.
- Proper data cleansing and preparation are very important for data mining, and a data warehouse can facilitate these activities. However, a data warehouse is of no use if it does not contain the data you need to solve your problem.

Why Data Mining?

- Credit ratings/targeted marketing:
  - Given a database of 100,000 names, which persons are the least likely to default on their credit cards?
  - Identify likely responders to sales promotions
- Fraud detection
  - Which types of transactions are likely to be fraudulent, given the demographics and transactional history of a particular customer?
- Customer relationship management:
  - Which of my customers are likely to be the most loyal, and which are most likely to leave for a competitor? :

Data Mining helps extract such information

Data mining

- Process of semi-automatically analyzing large databases to find patterns that are:
  - valid: hold on new data with some certainty
  - novel: non-obvious to the system
  - useful: should be possible to act on the item
  - understandable: humans should be able to interpret the pattern
- Also known as Knowledge Discovery in Databases (KDD)

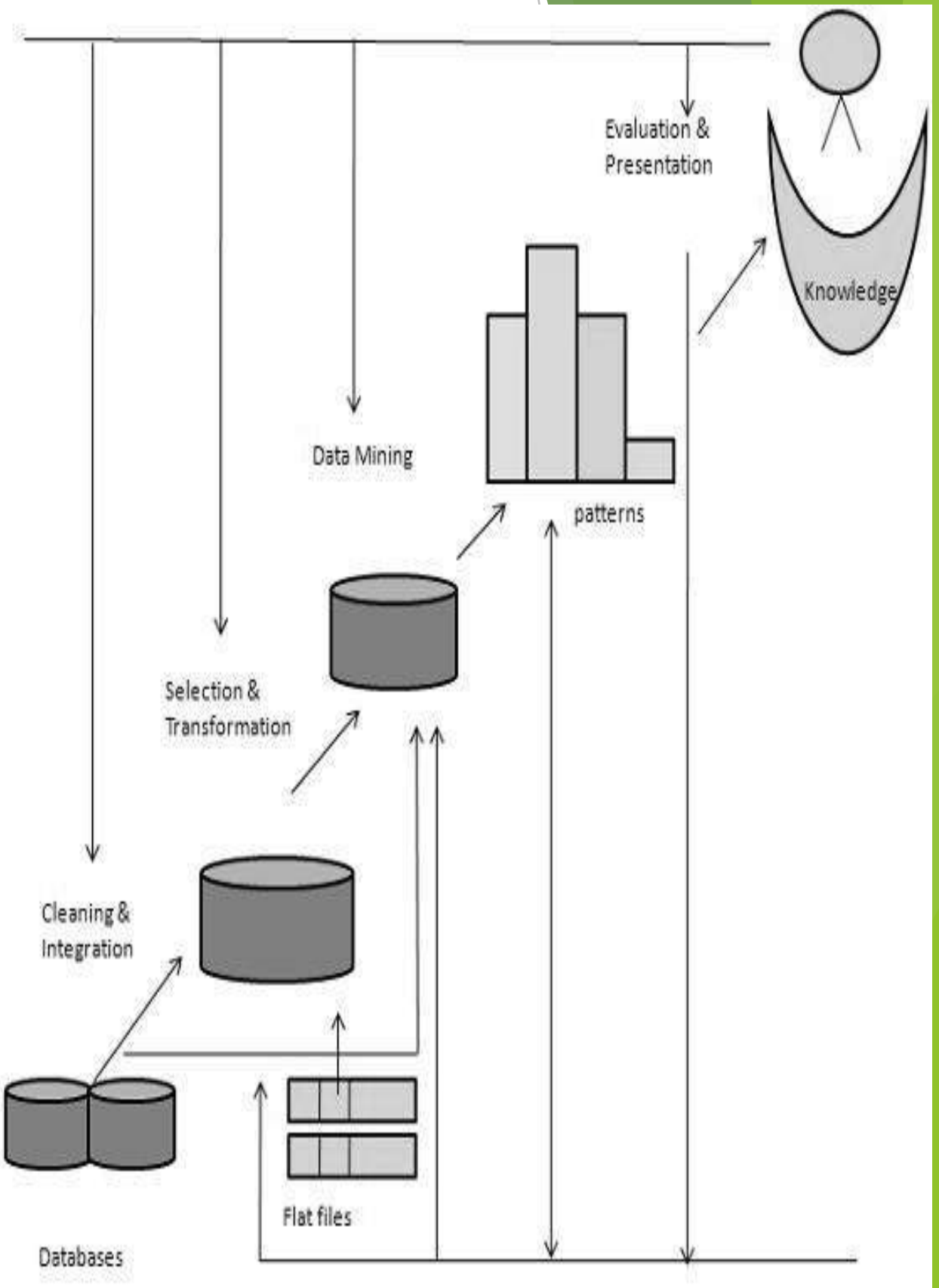
Knowledge discovery in databases

- The main objective of the KDD process is to extract information from data in the context of large databases. It does this by using Data Mining algorithms to identify what is deemed knowledge.
- The Knowledge Discovery in Databases is considered as a programmed, exploratory analysis and modeling of vast data repositories. KDD is the organized procedure of recognizing valid, useful, and understandable patterns from huge and complex data sets.
- Data Mining is the root of the KDD procedure, including the inferring of algorithms that investigate the data, develop the model, and find previously unknown patterns. The model is used for extracting the knowledge from the data, analyze the data, and predict the data.

What is Knowledge Discovery?

Here is the list of steps involved in the knowledge discovery process –

- **Data Cleaning** – In this step, the noise and inconsistent data is removed.
- **Data Integration** – In this step, multiple data sources are combined.
- **Data Selection** – In this step, data relevant to the analysis task are retrieved from the database.
- **Data Transformation** – In this step, data is transformed or consolidated into forms appropriate for mining by performing summary or aggregation operations.
- **Data Mining** – In this step, intelligent methods are applied in order to extract data patterns.
- **Pattern Evaluation** – In this step, data patterns are evaluated.
- **Knowledge Presentation** – In this step, knowledge is represented.



1. Data Cleaning

- Teams need to first clean all process data so it aligns with the industry standard. Dirty or incomplete data leads to poor insights and system failures that cost time and money. Engineers will remove all unclean data from the organization's acquired data.

  They use several different data preprocessing and cleaning methods, depending on the resources of the business. For example, they may manually fill in missing values or utilize the mean of other data to fill in a probable value. Teams will also use binning methods to remove noisy data, identify outliers, and resolve any inconsistencies.

Data Integration

- When data miners combine different data sets and sources to perform analysis, they refer to it as data integration. This is one of the top mining techniques to streamlines the entire extract, transform, and load p r o c e s s .

  Many specialists perform additional data cleaning within different databases during this stage. This further eliminates any inconsistent information and ensures data quality so it meets business requirements. Specialists will use data mining tools such as Microsoft SQL to integrate data.

Data Reduction for Data Quality

- This standard process extracts relevant information for data analysis and pattern evaluation. Engineers take a small size of the data and still maintain its integrity during data reduction. Teams may use neural networks or other forms of machine learning during this mining process. Strategies may include dimensionality reduction, numerosity reduction, or data compression.

  In dimensionality reduction, engineers reduce the quantity of attributes in the analytics data. In numerosity reduction, teams replace the original quantity of data with a smaller quantity of data. In data compression, engineers provide a compressed generalization of the collected data.

- **What to Know About Data Quality:**
- Sales and marketing departments lose 550 hours per week due to inaccurate data
- Companies lose up to 20% of revenue due to poor data quality
- 15% of leads contain duplicate records
- It costs roughly $1 dollar to prevent a duplicate, $10 to correct a duplicate, and $100 to store a duplicate if it is not eliminated

Data Transformation

- In this industry standard process, engineers transform data into an acceptable form to align with mining goals. They consolidate the preparation data to optimize data mining processes and make it easier to discern patterns in the final data set.

  Data transformation encompasses data mapping and other data science techniques. Strategies include smoothing, or eliminating noise from data. Other popular techniques include aggregation, normalization, or discretization.

Data Mining

- Organizations use data mining applications to extract useful trends and optimize knowledge discovery to generate business intelligence. This is only possible if a company takes full advantage of big data and collects the correct type of information.

Engineers apply intelligent patterns to the available data before they extract it. They then represent all information as models. Specialists use clustering, classification, or other modeling techniques to ensure accuracy.

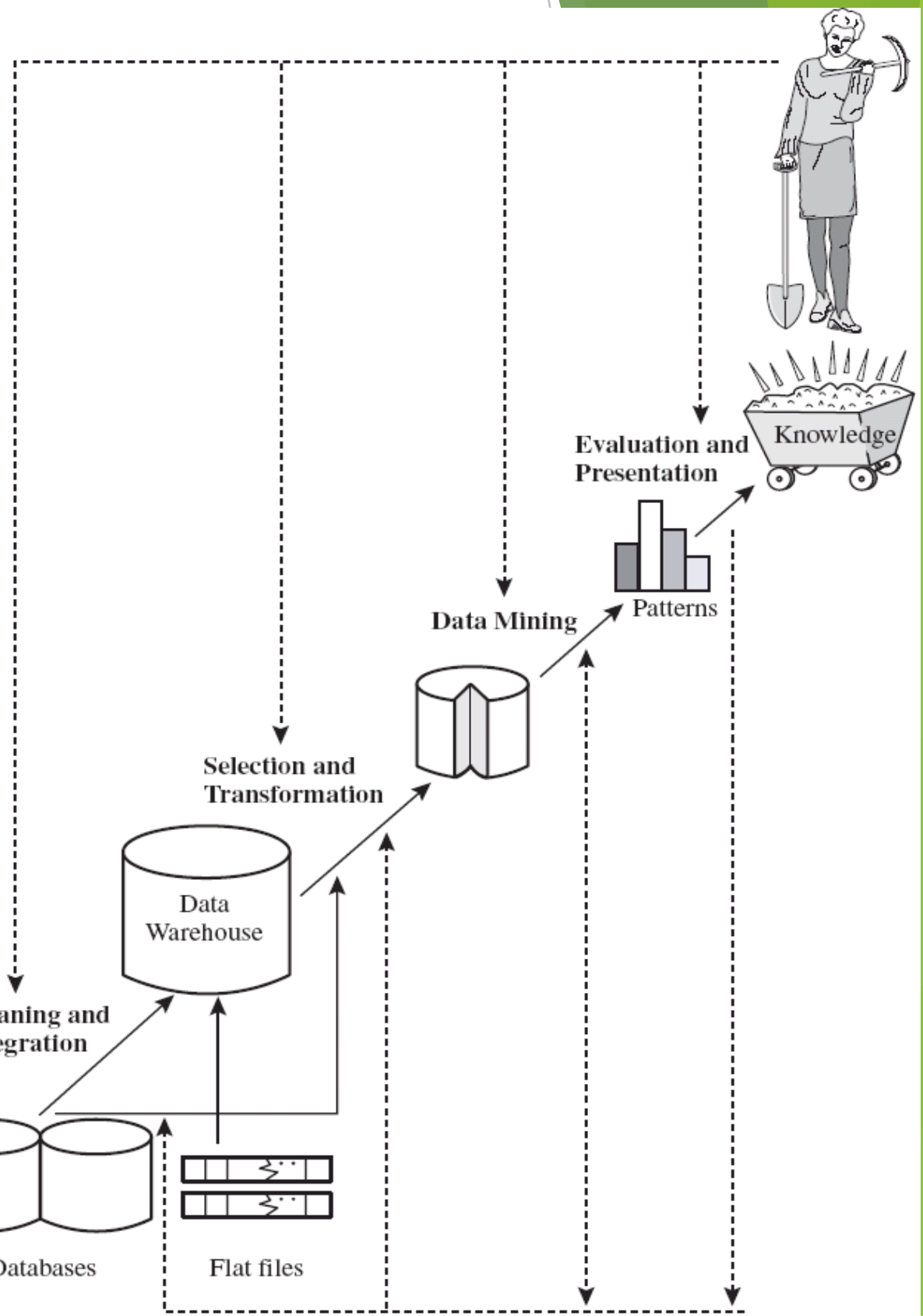
Pattern Evaluation

- This is the stage where engineers stop working behind the scenes and bring insights into the real world. Specialists will pinpoint any useful patterns that can generate business knowledge.

  They will use their models, historical data, and real-time information to find out more about customers, employees, and sales. Teams will also summarize information data or use visualization data mining techniques to make it easier to understand.

Representing Knowledge in Data Mining

- Finally, data analysts use a combination of data visualization, reports, and other mining tools to share the information with others Before the data mining process even started, business leaders communicated data understanding goals and objectives so engineers knew what to look for.



Applications

- Banking: loan/credit card approval
  - predict good customers based on old customers
- Customer relationship management:
  - identify those who are likely to leave for a competitor.
- Targeted marketing:
  - identify likely responders to promotions
- Fraud detection: telecommunications, financial transactions
  - from an online stream of event identify fraudulent events
- Manufacturing and production:
  - automatically adjust knobs when process parameter changes

Applications (continued)

- Medicine: disease outcome, effectiveness of treatments
  - analyze patient disease history: find relationship between diseases
- Molecular/Pharmaceutical: identify new drugs
- Scientific data analysis:
  - identify new galaxies by searching for sub clusters
- Web site/store design and promotion:
  - find affinity of visitor to pages and modify layout

The KDD process

- Problem formulation
- Data collection
  - subset data: sampling might hurt if highly skewed data
  - feature selection: principal component analysis, heuristic search
- Pre-processing: cleaning
  - name/address cleaning, different meanings (annual, yearly), duplicate removal, supplying missing values
- Transformation:
  - map complex objects e.g. time series data to features e.g. frequency
- Choosing mining task and mining method:
- Result evaluation and Visualization:

Knowledge discovery is an iterative process

Relationship with other fields

- Overlaps with machine learning, statistics, artificial intelligence, databases, visualization but more stress on
  - scalability of number of features and instances
  - stress on algorithms and architectures whereas foundations of methods and formulations provided by statistics and machine learning.
  - automation for handling large, heterogeneous data

Some basic operations

- Predictive:
  - Regression
  - Classification
  - Collaborative Filtering
- Descriptive:
  - Clustering / similarity matching
  - Association rules and variants
  - Deviation detection

Data Mining in Practice

Application Areas

Industry	Application
Finance	Credit Card Analysis
Insurance	Claims, Fraud Analysis
Telecommunication	Call record analysis
Transport	Logistics management
Consumer goods	promotion analysis
Data Service providers	Value added data
Utilities	Power usage analysis

Why Now?

- Data is being produced
- Data is being warehoused
- The computing power is available
- The computing power is affordable
- The competitive pressures are strong
- Commercial products are available

Data Mining works with Warehouse Data

- Data Warehousing provides the Enterprise with a memory

ñData Mining provides the Enterprise with intelligence

Usage scenarios

- Data warehouse mining:
  - assimilate data from operational sources
  - mine static data
- Mining log data
- Continuous mining: example in process control
- Stages in mining:
  - data selection → pre-processing: cleaning → transformation → mining → result evaluation → visualization

Mining market

- Around 20 to 30 mining tool vendors
- Major tool players:
  - Clementine,
  - IBM's Intelligent Miner,
  - SGI's MineSet,
  - SAS's Enterprise Miner.
- All pretty much the same set of tools
- Many embedded products:
  - fraud detection:
  - electronic commerce applications,
  - health care,
  - customer relationship management: Epiphany

Vertical integration: Mining on the web

- Web log analysis for site design:
  - what are popular pages,
  - what links are hard to find.
- Electronic stores sales enhancements:
  - recommendations, advertisement:
  - Collaborative filtering: Net perception, Wisewire
  - Inventory control: what was a shopper looking for and could not find..

OLAP Mining integration

- OLAP (On Line Analytical Processing)
  - Fast interactive exploration of multidim. aggregates.
  - Heavy reliance on manual operations for analysis:
  - Tedious and error-prone on large multidimensional data
- Ideal platform for vertical integration of mining but needs to be interactive instead of batch.

State of art in mining OLAP integration

- Decision trees [Information discovery, Cognos]
  - find factors influencing high profits
- Clustering [Pilot software]
  - segment customers to define hierarchy on that dimension
- Time series analysis: [Seagate's Holos]
  - Query for various shapes along time: eg. spikes, outliers
- Multi-level Associations [Han et al.]
  - find association between members of dimensions
- Sarawagi [VLDB2000]

Data Mining in Use

- The US Government uses Data Mining to track fraud
- A Supermarket becomes an information broker
- Basketball teams use it to track game strategy
- Cross Selling
- Target Marketing
- Holding on to Good Customers
- Weeding out Bad Customers

Some success stories

- Network intrusion detection using a combination of sequential rule discovery and classification tree on 4 GB DARPA data
  - Won over (manual) knowledge engineering approach
  - http://www.cs.columbia.edu/~sal/JAM/PROJECT/ provides good detailed description of the entire process
- Major US bank: customer attrition prediction
  - First segment customers based on financial behavior: found 3 segments
  - Build attrition models for each of the 3 segments
  - 40-50% of attritions were predicted == factor of 18 increase
- Targeted credit marketing: major US banks
  - find customer segments based on 13 months credit balances
  - build another response model based on surveys
  - increased response 4 times -- 2%