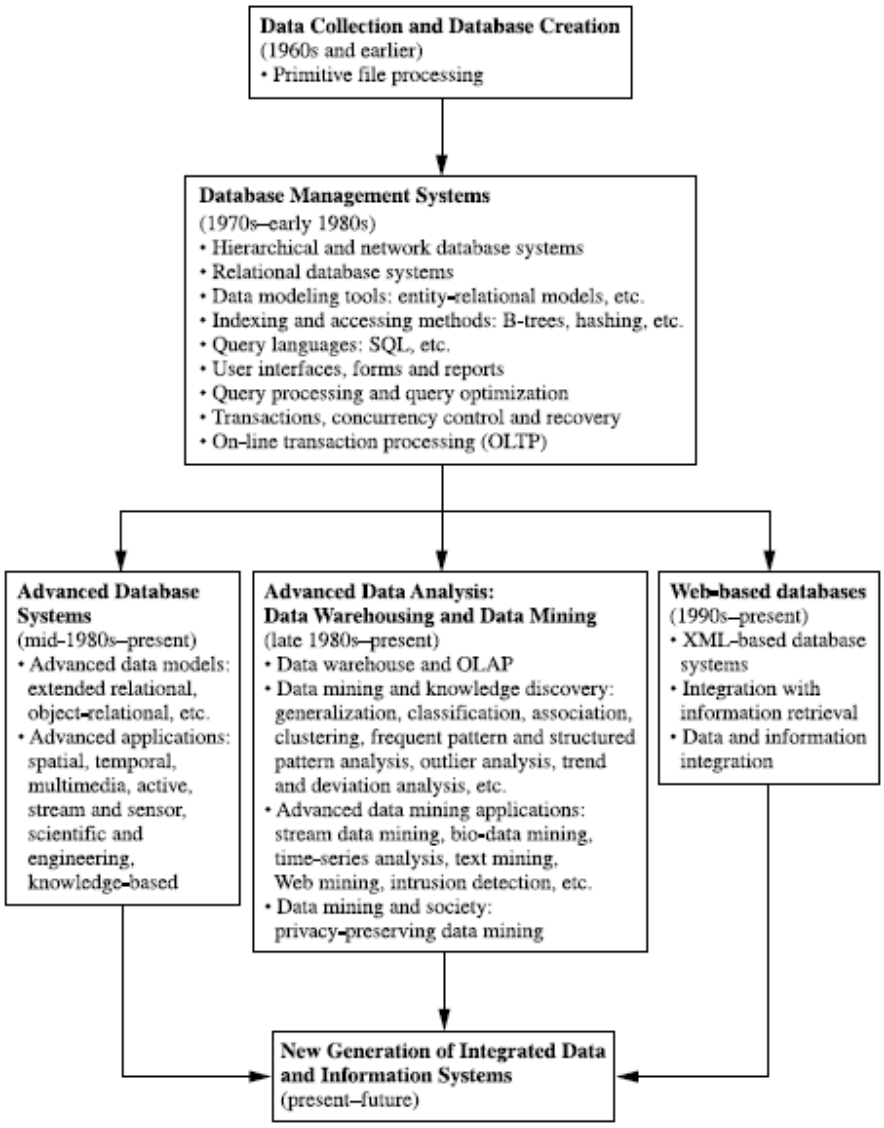
WHAT MOTIVATED DATA MINING? WHY IS IT IMPORTANT?



Data mining is mainly used for decision making in business. The abundance of data, coupled with the need for powerful data analysis tools, has been described as a *data rich but information poor* situation. The fast-growing, tremendous amount of data, collected and stored in large and numerous data repositories (storage), has far exceeded our human ability for comprehension without powerful tools. As a result, data collected in large data repositories become "data tombs"—data archives that are seldom visited. Consequently, important decisions are often made based not on the information-rich data stored in data repositories, but rather on a decision maker's intuition, simply because the decision maker does not have the tools to extract the valuable knowledge embedded in the vast amounts of data. The widening gap between data and information calls for a systematic development of *data mining tools* that will turn data tombs into "golden nuggets" of knowledge.(explain the diagram in your words).

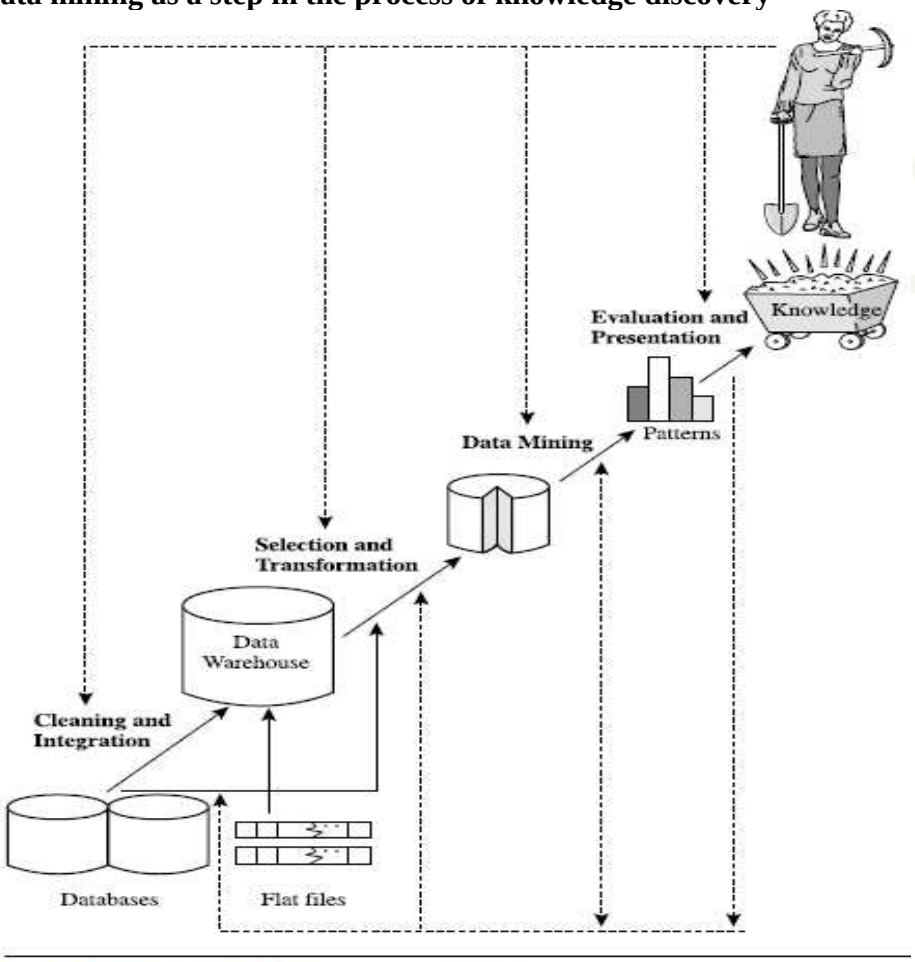
WHAT IS DATA MINING?

EXTRACTING OR "MINING" KNOWLEDGE FROM LARGE AMOUNTS OF DATA IS REFERRED TO AS DATA MINING.

Other names for data mining:

- *Knowledge Discovery (Mining) In Databases (KDD),*
- *Knowledge Extraction*
- *Data/Pattern Analysis,*
- *Data Archeology*
- *Data Dredging,*
- *Information Harvesting,*
- *Business Intelligence*

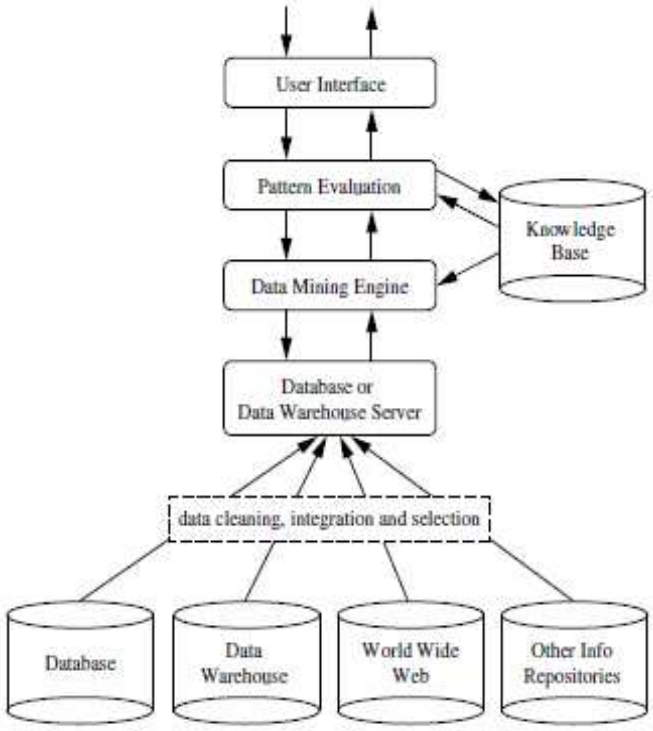
Data mining as a step in the process of knowledge discovery



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

1. **Data cleaning**: It is used to remove noise and inconsistent data from raw data.
2. Data integration: Data from multiple sources are combined.
3. Data selection: Data relevant to the analysis task are retrieved from the database.
4. Data transformation: Data are transformed or consolidated into forms appropriate (understandable) for mining by performing summary or aggregation operations.
5. Data mining: It is an essential process where intelligent methods are applied in order to extract data patterns (like characterization discrimination, classification association analysis, prediction, clustering etc...).
6. Pattern evaluation it is used to identify the truly interesting patterns representing knowledge based on some interestingness measures.
7. Knowledge presentation: visualization and knowledge representation techniques are used to present the mined knowledge to the user such as charts, graphs, etc...
Note: 1 to 4 steps are known as preprocessing steps.

ARCHITECTURE OF A TYPICAL DATA MINING SYSTEM.



Database, data warehouse, World Wide Web, or other information repository:
This is one or a set of databases, data warehouses, spreadsheets, or other kinds of information repositories. Data cleaning and data integration techniques may be performed on the data.

Database or data warehouse server: The database or data warehouse server is responsible for fetching the relevant data, based on the user's data mining request.

Knowledge base: This is the domain knowledge (background knowledge often given by user) that is used to guide the search or evaluate the interestingness of resulting patterns.
Such knowledge can include:

- Concept hierarchies: used to organize attributes or attribute values into different levels of abstraction (attribute city can be organized into district, state, country etc..)
- user beliefs
- Thresholds, and metadata (e.g., describing data from multiple heterogeneous sources).

Data mining engine: This is essential to the data mining system and ideally consists of a set of functional modules for tasks such as

- Characterization
- association and correlation analysis
- classification & prediction
- cluster & outlier analysis
- Evolution analysis.

Pattern evaluation module: This component interacts with the data mining modules so as to *focus* the search toward interesting patterns. These tools are used to filter the patterns (raw information) generated by data mining module. For efficient data mining, it is highly recommended to push the evaluation of pattern interestingness as deep as possible into the mining process

User interface:

- Helps for communication between users and the data mining system.
- Helps to specify a data mining query
- Allows the user to browse database and data warehouse schemas or data structures.
- Evaluate mined patterns
- Visualize the patterns in different forms (graph, tree, chart, tables).

DATA MINING—ON WHAT KIND OF DATA MINING CAN BE PERFORMED?

Data mining can be performed on different data repositories such as:

- Relational databases
- Data warehouses
- Transactional databases,
- Advanced Data and Information Systems and Advanced Applications
  - object-relational databases
  - object oriented databases
  - Spatial databases
  - Time-series databases and temporal database
  - Text databases and multimedia databases
  - Heterogeneous Databases and Legacy Databases
  - World Wide Web

1) RELATIONAL DATABASES

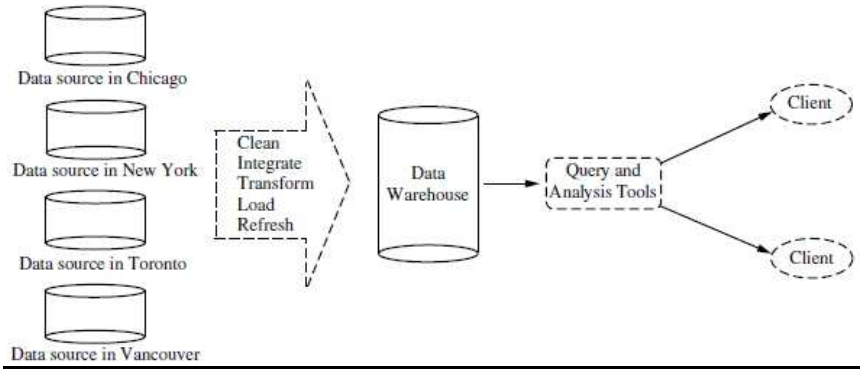
A database system, also called a database management system (DBMS), consists of a collection of interrelated data, known as a database, and a set of software programs to manage and access the data. A relational database is a collection of tables, each of which is assigned a unique name. Each table consists of a set of attributes (*columns* or *fields*) and usually stores a large set of tuples (*records* or *rows*).

customer

cust_ID	*name*	*address*	*age*	*income*	*credit_info*	*category*	...
C1	Smith, Sandy	1223 Lake Ave., Chicago, IL	31	$78000	1	3	...
...	...	...	...	...	...	...	...

2) DATA WAREHOUSE -

A Data warehouse is a repository of information collected from multiple sources, stored under a unified schema, and that usually resides at a single site. Data warehouses are constructed via a process of data cleaning, data integration, data transformation, data loading, and periodic data refreshing.



3) TRANSACTIONAL DATABASE -

In general, a transactional database consists of a file where each record represents a transaction. A transaction typically includes a unique transaction identity number (*trans ID)* and a list of the items making up the transaction (such as items purchased in a store).

trans_ID	*list of item_IDs*
T100	I1, I3, I8, I16
T200	I2, I8
...	...

4) Advanced Data and Information Systems and Advanced Applications

a) Object-oriented databases,

- Each entity is considered as an object
- Data and code relating to an object are *encapsulated* into a single unit.
- Each object has associated with it the following:
    - A set of variables that describe the objects. These correspond to attributes in the entity-relationship and relational models.
    - A set of messages that the object can use to communicate with other objects, or with the rest of the database system.
    - A set of methods, where each method holds the code to implement a message. Upon receiving a message, the method returns a value in response.
- Objects that share a common set of properties can be grouped into an object class.

b) Object-relational databases

Object-relational databases are constructed based on an object-relational data model.
This model extends the relational model by providing a rich data type for handling complex objects and object orientation. Because most sophisticated database applications need to handle complex objects and structures, object-relational databases are becoming increasingly popular in industry and applications. Conceptually, the object-relational data model inherits the essential concepts of object-oriented databases.

c) Temporal Databases, Sequence Databases, and Time-Series Databases

- A **Temporal database** typically stores relational data that include time-related attributes. These attributes may involve several timestamps, each having different semantics.
- A **Sequence database** stores sequences of ordered events, with or without a concrete notion of time. Examples include customer shopping sequences, Web click streams, and biological sequences.
- A **Time-series database** stores sequences of values or events obtained over repeated measurements of time (e.g., hourly, daily, weekly). Examples include data collected from the stock exchange, inventory control, and the observation of natural phenomena (like temperature and wind).

d) Spatial Databases

- Spatial data may be represented in **raster format**, consisting of *n*-dimensional bit maps or pixel maps. For example, a 2-D satellite image may be represented as raster data.
- Maps can be represented in **vector format**, where roads, bridges, buildings, and lakes are represented as unions or overlays of basic geometric constructs, such as points, lines, polygonsetc…
- Applications:
    - vehicle navigation
    - forestry and ecology planning to providing public service information regarding the location of telephone and electric cables, pipes, and sewage systems.

e) Text Databases and Multimedia Databases

Text databases are databases that contain word descriptions for objects. These word descriptions are usually not simple keywords but rather long sentences or paragraphs.
Text databases may be:

- Highly unstructured: such as some Web pages on the World Wide Web.
- Semi structured: such as e-mail messages and many HTML/XML Web pages.
- Structured: such as library catalogue databases.

Multimedia databases store image, audio, and video data. They are used in applications such as picture content-based retrieval, voice-mail systems, video-on-demand systems, the World Wide Web, and speech-based.

f) Heterogeneous Databases and Legacy Databases

- A **Heterogeneous database** consists of a set of interconnected, autonomous component databases. The components communicate in order to exchange information and answer queries.
- **Legacy Database** formed as a result of long history of IT Development. A legacy database is a group of *heterogeneous databases that combines* different kinds of data systems, such as relational or object-oriented databases, hierarchical databases, network databases, spreadsheets, multimedia databases, or file systems.

g) WWW

The **World Wide Web** and its associated distributed information services, such as Yahoo!, Google, America Online, and AltaVista, provide rich, worldwide, on-line information services, where data objects are linked together to facilitate interactive access Users seeking information of interest traverse from one object via links to another.Such systems provide ample opportunities and challenges for data mining.

DATA MINING FUNCTIONALITIES—WHAT KINDS OF PATTERNS CAN BE MINED?

Data mining functionalities:

- Characterization
- Discrimination
- association analysis
- classification
- prediction
- cluster analysis
- Outlier analysis
- Evolution analysis.

1) Data characterization

- Data characterization is a summarization of the general characteristics or features of a target class of data.
- The data corresponding to the user-specified class are typically collected by a database query.
- For example, to study the characteristics of software products whose sales increased by 10% in the last year, the data related to such products can be collected by executing an SQL query**.**
- **Methods for attribute characterization:**
  Attribute-oriented induction
  OLAP roll-up operation

2) Data discrimination:

- Data discrimination is a comparison of the general features of target class data objects with the general features of objects from one or a set of contrasting classes.
- The target and contrasting classes can be specified by the user.
- For example, the user may like to compare the general features of software products whose sales increased by 10% in the last year with those whose sales decreased by at least 30% during the same period.

3) Association analysis:

- It is the discovery of associations rules showing attribute value conditions that occur frequently together in a given set of data.
- For Example:
  buys(X; “computer”)=>buys(X; “software”)
  [support = 1%; confidence = 50%]

It means that out of all transactions 1% of transactions contain computer and software together. And 50% of people who bought computer also bought software.

4) Classification:

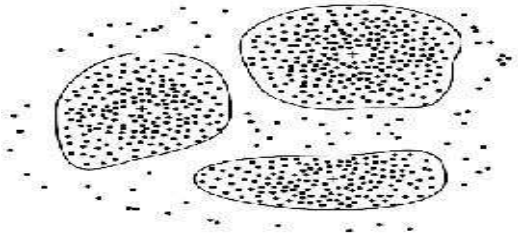
- Classification is the process of finding a model (or function) that describes and distinguishes data classes or concepts, for the purpose of being able to use the model to predict the class of objects whose class label is unknown.
- The derived model may be represented in various forms, such as
  - Classification (IF-THEN) rules,
  - Decision trees,
  - Bayesian classification,
  - Back propagation,
  - Neural network etc.

5) PREDICTION:

- Prediction is used to predict some missing values or unavailable values rather than class labels.

6) CLUSTER ANALYSIS:

- Unlike classification and prediction, which analyze class-labeled data objects, clustering analyzes data objects without consulting a known class label.
- The objects are clustered or grouped based on the principle of maximizing the intra class similarity and minimizing the interclass similarity.



7) OUTLIER ANALYSIS:

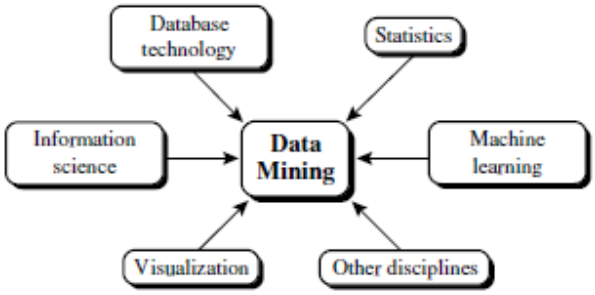
- A database may contain data objects that do not comply with the general behavior or model of the data. These data objects are outliers.
- However, in some applications such as fraud detection, the rare events can be more interesting than the more regularly occurring ones.

8) EVOLUTION ANALYSIS:

Data evolution analysis describes and models regularities or trends for objects whose behavior changes over time. Although this may include characterization, discrimination, association and correlation analysis, classification, prediction, or clustering of *time related data*, distinct features of such an analysis include time-series data analysis, sequence or periodicity pattern matching, and similarity-based data analysis.

CLASSIFICATION OF DATA MINING SYSTEMS

Data mining is an interdisciplinary field, the confluence of a set of disciplines, including database systems, statistics, machine learning, visualization, and information science Because of the diversity of disciplines contributing to data mining, there are large variety of data mining systems. Therefore, it is necessary to provide a clear classification of data mining systems



1) Classification according to the kinds of databases mined:

The data may in any type; the data mining algorithms can be classified according to the type of data they use as input or/and output such as

- Relational
- Data warehouse
- Transactional
- Stream data
- Object oriented/relational,
- Spatial,
- Time-series,
- Text or multi-media
- Heterogeneous or legacy
- WWW

2) Classification according to the *kinds of knowledge* mined:

Type of information that can be extracted from a data depends on the functionality used. The data mining algorithms are classified according to the functionality they use.

- Characterization
- Discrimination,
- Association,
- Classification,
- Clustering,
- Outlier analysis, etc

3) Classification according to the *kinds of techniques* utilized:

The data mining algorithms can be classified according to the underlying techniques they use.

- Autonomous systems, interactive exploratory systems, query-driven systems.
- Database-oriented, data warehouse (OLAP), machine learning, statistics, visualization, information science etc.

4) Classification according to the *applications adapted*:

They can be classified according to the purpose for which they are designed.

- Retail, telecommunication, banking, fraud analysis, bio-data mining, stock market analysis, text mining, Web mining, etc.

MAJOR ISSUES IN DATA MINING

- Mining methodology and user interaction issues
  - *Mining different kinds of knowledge in databases*
  - *Interactive mining of knowledge at multiple levels of abstraction*
  - *Incorporation of background knowledge*
  - *Data mining query languages and ad hoc data mining*
  - *Presentation and visualization of data mining results*
  - *Handling noisy or incomplete data*
  - *Pattern evaluation—the interestingness problem.*
- Performance issues.
  - *Efficiency and scalability of data mining algorithms*
  - *Parallel, distributed, and incremental mining algorithms*
- Issues relating to the diversity of database types
  - *Handling of relational and complex types of data*
  - *Mining information from heterogeneous databases and*
  - *global information systems*