DATA MINING CONCEPTS AND TECHNIQUES

Marek Maurizio

E-commerce, winter 2011

INTRODUCTION

- Overview of data mining
- Emphasis is placed on basic data mining concepts
- Techniques for uncovering interesting data patterns hidden in *large data sets*

"GETTING INFORMATION OFF THE INTERNET IS LIKE TAKING A DRINK FROM A FIRE HYDRANT"

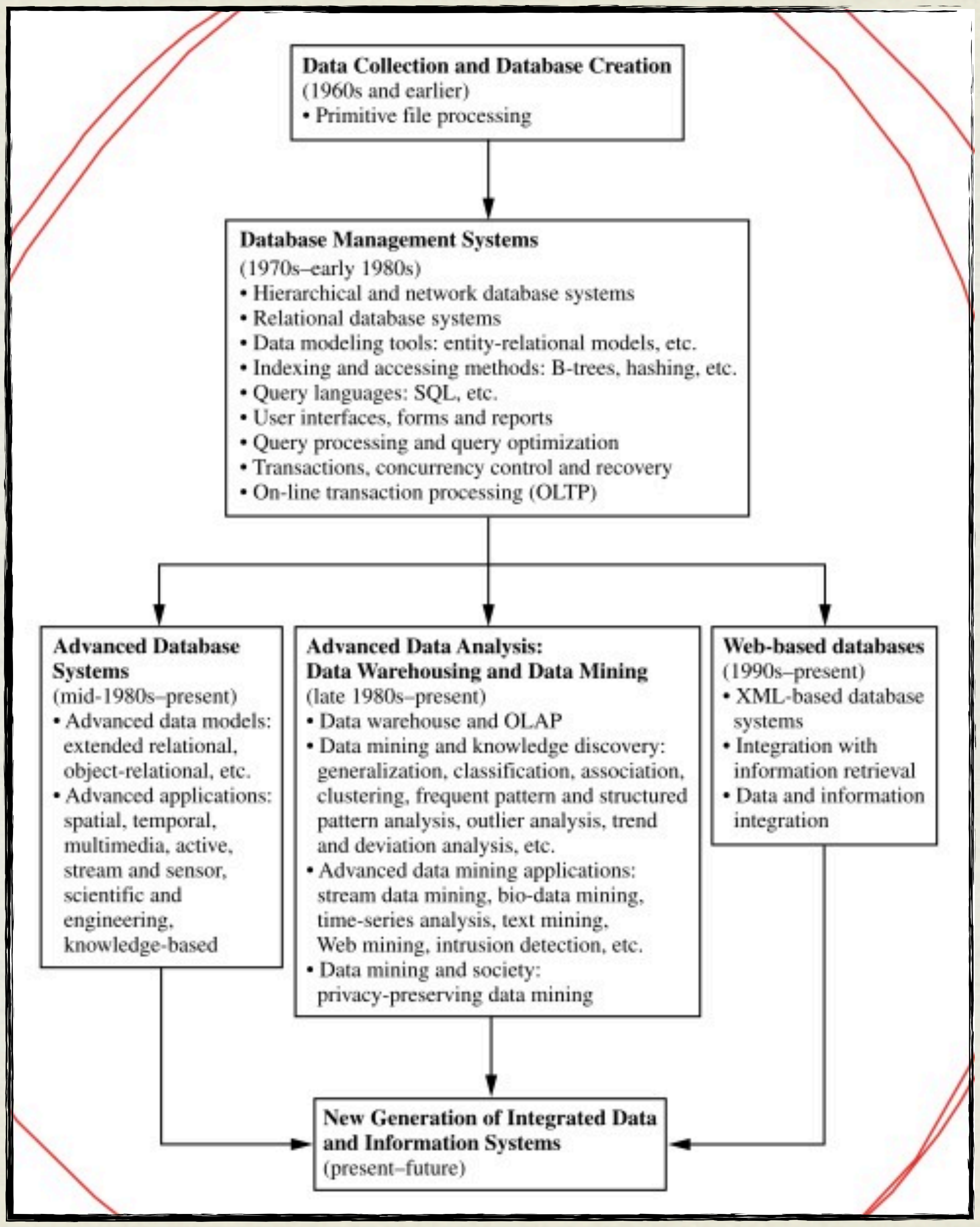
MITCH KAPOR, FOUNDER OF LOTUS DEVELOPMENT

MOTIVATIONS

- Data mining has attracted a great deal of attention in the information industry and in society as a whole in recent years
- Wide availability of huge amounts of data and the imminent need for turning such data into useful information and knowledge
- Market analysis, fraud detection, and customer retention, production control and science exploration

EVOLUTION

- Data mining can be viewed as a result of the natural evolution of information technology
- Since the 1960s, database and information technology has been evolving systematically from primitive file processing systems to sophisticated and powerful database systems



Data Collection and Database Creation
(1960s and earlier)
- Primitive file processing

Database Management Systems
(1970s–early 1980s)
- Hierarchical and network database systems
- Relational database systems
- Data modeling tools: entity-relational models, etc.
- Indexing and accessing methods: B-trees, hashing, etc.
- Query languages: SQL, etc.
- User interfaces, forms and reports
- Query processing and query optimization
- Transactions, concurrency control and recovery
- On-line transaction processing (OLTP)

Advanced Database Systems
(mid-1980s–present)
- Advanced data models: extended relational, object-relational, etc.
- Advanced applications: spatial, temporal, multimedia, active, stream and sensor, scientific and engineering, knowledge-based

Advanced Data Analysis: Data Warehousing and Data Mining
(late 1980s–present)
- Data warehouse and OLAP
- Data mining and knowledge discovery: generalization, classification, association, clustering, frequent pattern and structured pattern analysis, outlier analysis, trend and deviation analysis, etc.
- Advanced data mining applications: stream data mining, bio-data mining, time-series analysis, text mining, Web mining, intrusion detection, etc.
- Data mining and society: privacy-preserving data mining

Web-based databases
(1990s–present)
- XML-based database systems
- Integration with information retrieval
- Data and information integration

New Generation of Integrated Data and Information Systems
(present–future)

EVOLUTION - II

- From early hierarchical and network database systems to the development of relational database systems
- Users gained convenient and flexible data access through query languages, user interfaces, optimized query processing, and transaction management
- Research on advanced data models such as extended-relational, object-oriented, object-relational, and deductive models

DATA WAREHOUSE

- One data repository architecture that has emerged is the data warehouse
- Repository of multiple heterogeneous data sources organized under a unified schema at a single site
- Facilitate management decision making

DATA WAREHOUSE - II

- Data warehouse technology includes:
  - data cleaning
  - data integration
  - on-line analytical processing (OLAP)
    - analysis techniques with functionalities such as summarization, consolidation, and aggregation
    - ability to view information from different angles

We are data rich, but information poor

INFORMATION POORNESS

- The abundance of data, coupled with the need for powerful data analysis tools, has been described as a *data rich but information poor* situation
- Data collected in large data repositories become “data tombs”
  - data archives that are seldom visited

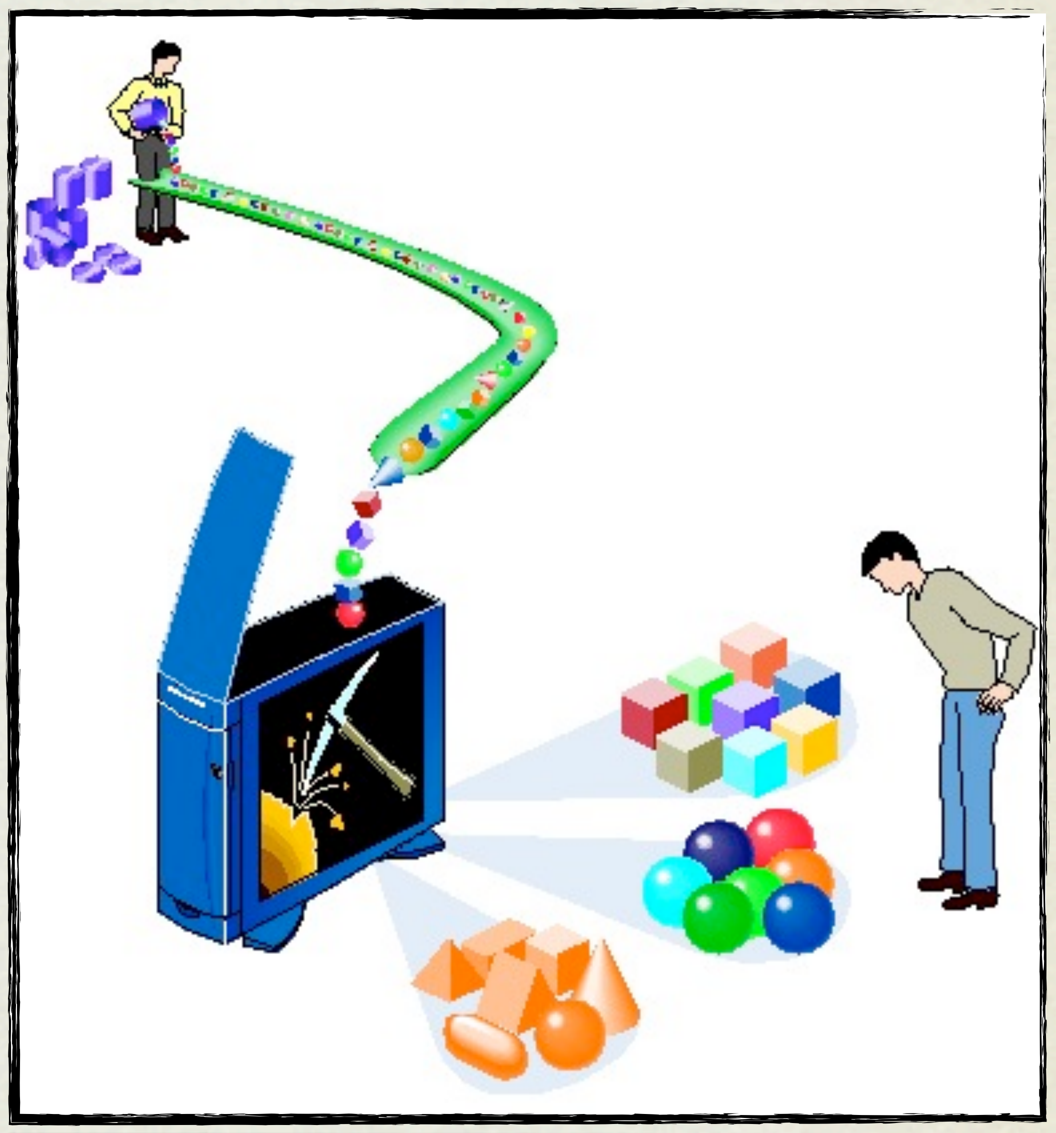
DECISION MAKING

- Important decisions are often made based not on the information-rich data stored in data repositories, but rather on a decision maker’s intuition
- The decision maker does not have the tools to extract the valuable knowledge embedded in the vast amounts of data

DATA ENTRY

- Often systems rely on users or domain experts to manually input knowledge into knowledge bases.
- Unfortunately, this procedure is prone to biases and errors, and is extremely time-consuming and costly

Simply stated, data mining refers to extracting or “mining” knowledge from large amounts of data, usually automatically gathered

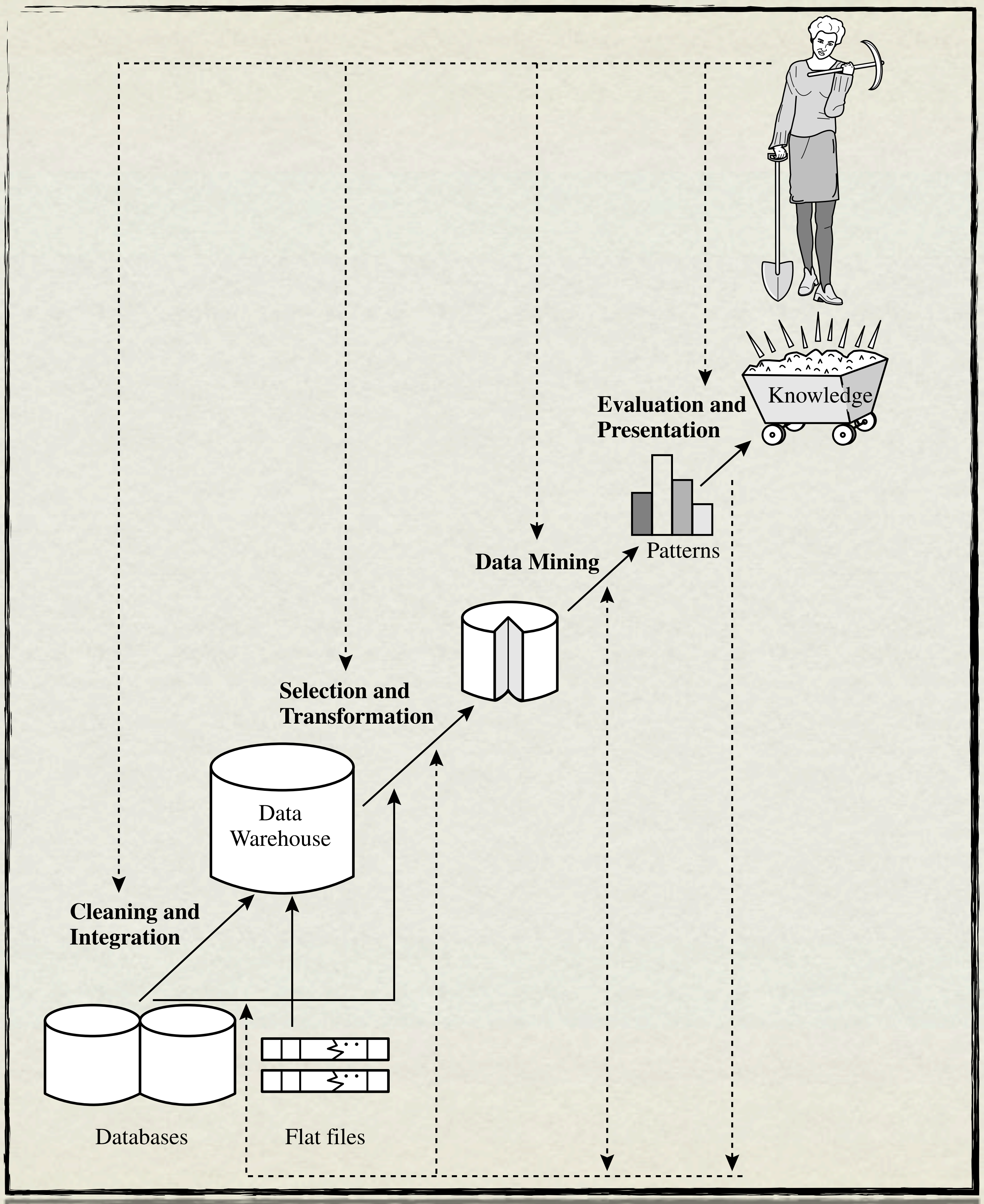


BAD NAME

- The term is actually a misnomer.
- Mining of gold from rocks or sand is referred to as gold mining rather than rock or sand mining.
- Data mining should have been more appropriately named "*knowledge mining from data*"
  - which is unfortunately somewhat long

KDD

- Many people treat data mining as a synonym for another popularly used term, Knowledge Discovery from Data (KDD)
- Data mining is, instead, an (essential) step in the KDD process


Evaluation and Presentation
Knowledge
Patterns
Data Mining
Selection and Transformation
Data Warehouse
Cleaning and Integration
Databases
Flat files

KDD STEPS

1. Data cleaning (to remove noise and inconsistent data)

2. Data integration (where multiple data sources may be combined)

3. Data selection (where data relevant to the analysis task are retrieved from the database)

4. Data transformation (where data are transformed or consolidated into forms appropriate for mining by performing summary or aggregation operations, for instance)

5. Data mining (an essential process where intelligent methods are applied in order to extract data patterns)

6. Pattern evaluation (to identify the truly interesting patterns representing knowledge based on some interestingness measures)

7. Knowledge presentation (where visualization and knowledge representation techniques are used to present the mined knowledge to the user)

MORE ON TERMINOLOGY

- We agree that data mining is a step in the knowledge discovery process

- in industry, in media, and in the database research milieu, the term data mining is becoming more popular than the longer term of knowledge discovery from data

- broad view of data mining functionality: data mining is the process of discovering interesting knowledge from large amounts of data stored in databases, data warehouses, or other information repositories

DATA MINING ON WHAT KIND OF DATA?

a number of different data repositories on which mining can be performed

RELATIONAL DATABASES

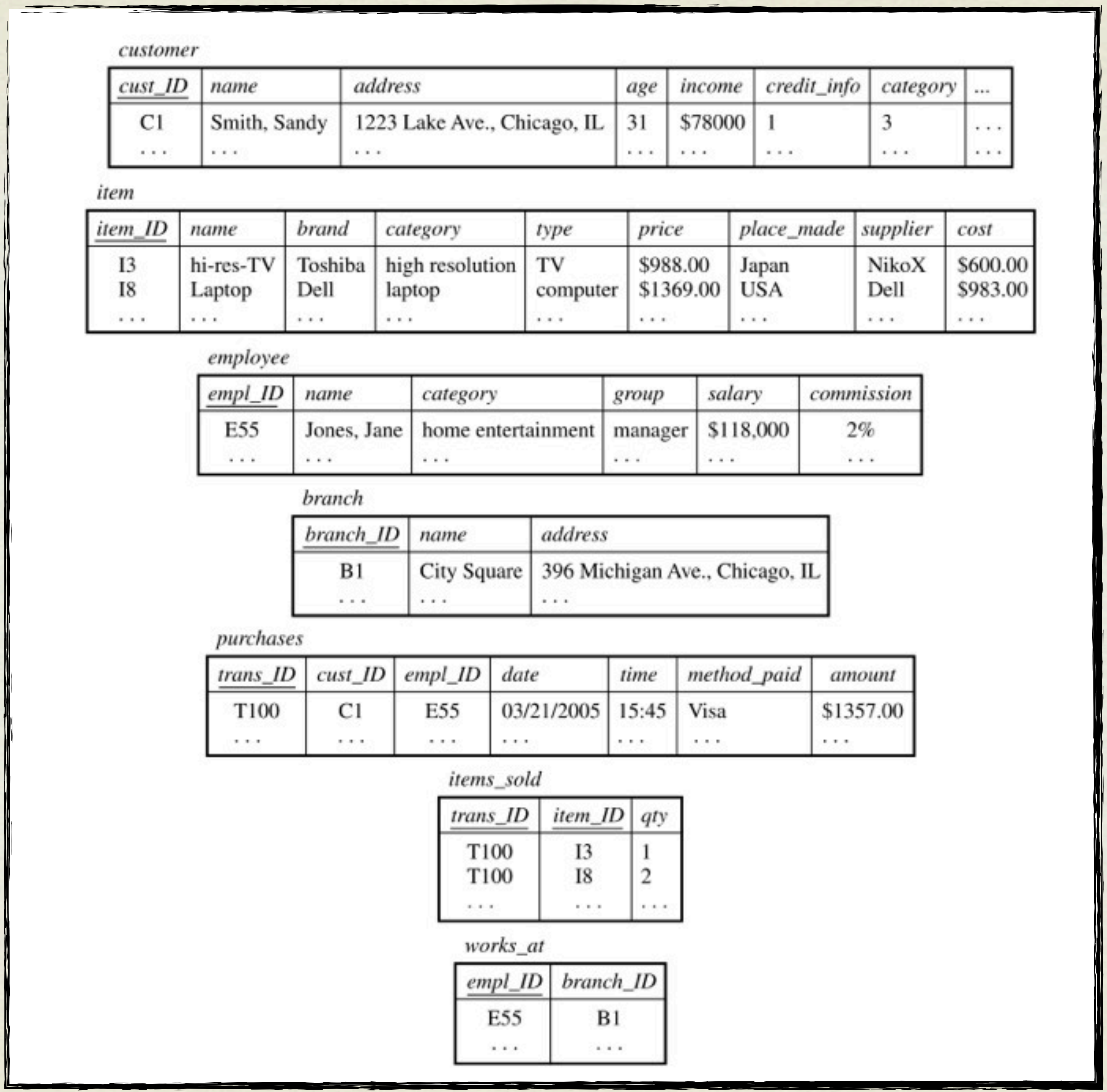
- A database system, also called a database management system (DBMS), consists of a collection of interrelated data, known as a database, and a set of software programs to manage and access the data
- A relational database is a collection of tables, each of which is assigned a unique name. Each table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

RELATIONAL DATABASES - II

- Each tuple in a relational table represents an object identified by a unique key and described by a set of attribute values

- A semantic data model, such as an entity-relationship (ER) data model, is often constructed for relational databases. An ER data model represents the database as a set of entities and their relationships

- Relational data can be accessed by database queries written in a relational query language, such as SQL



customer

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

item

item_ID	*name*	*brand*	*category*	*type*	*price*	*place_made*	*supplier*	*cost*
I3	hi-res-TV	Toshiba	high resolution	TV	$988.00	Japan	NikoX	$600.00
I8	Laptop	Dell	laptop	computer	$1369.00	USA	Dell	$983.00
...	...	...	...	...	...	...	...	...

employee

empl_ID	*name*	*category*	*group*	*salary*	*commission*
E55	Jones, Jane	home entertainment	manager	$118,000	2%
...	...	...	...	...	...

branch

branch_ID	*name*	*address*
B1	City Square	396 Michigan Ave., Chicago, IL
...	...	...

purchases

trans_ID	*cust_ID*	*empl_ID*	*date*	*time*	*method_paid*	*amount*
T100	C1	E55	03/21/2005	15:45	Visa	$1357.00
...	...	...	...	...	...	...

items_sold

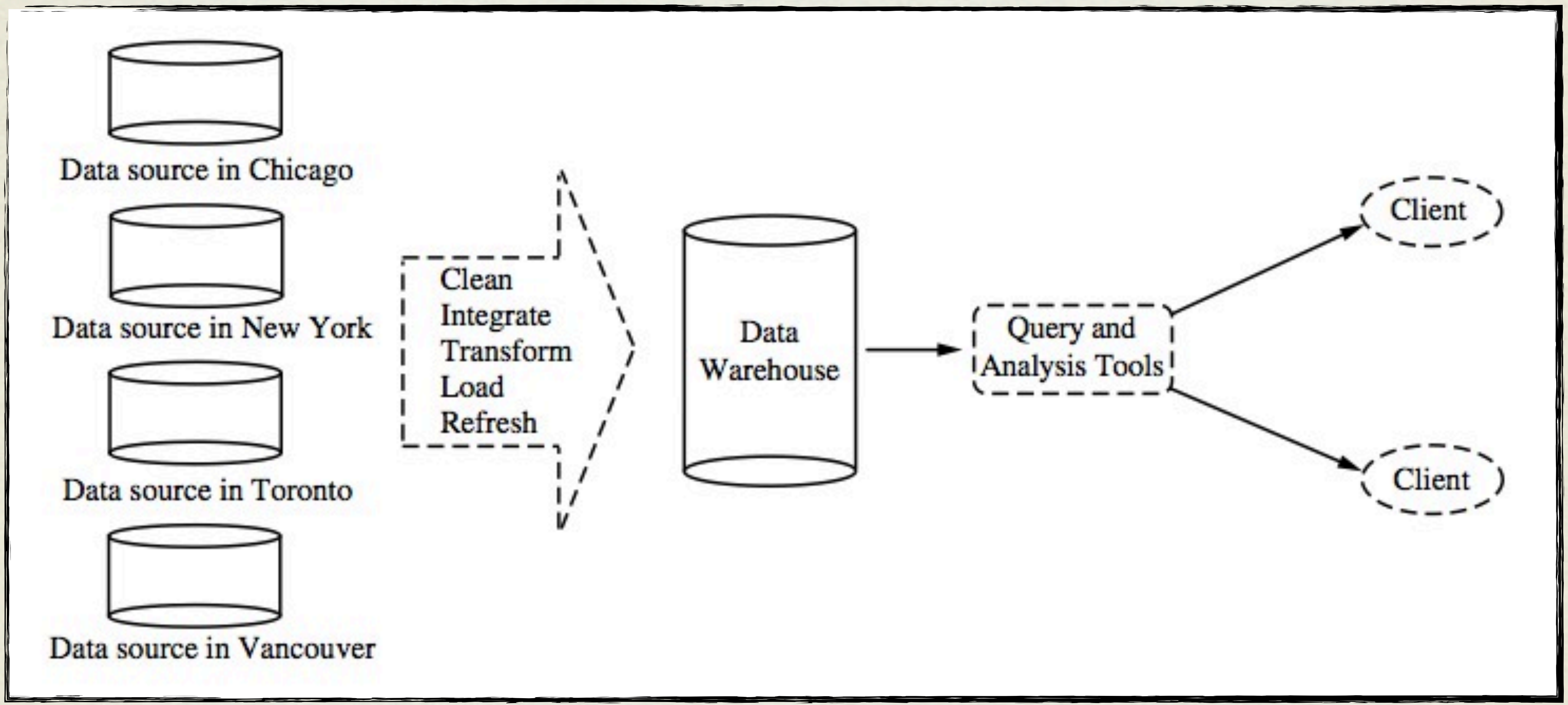
trans_ID	*item_ID*	*qty*
T100	I3	1
T100	I8	2
...	...	...

works_at

empl_ID	*branch_ID*
E55	B1
...	...

DATA WAREHOUSES

- 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



Typical framework of a data warehouse for AllElectronics.

OBJECT-RELATIONAL DATABASES

- Based on an object-relational data model
- Extends the relational model by providing a rich data type for handling complex objects and object orientation
- Objects that share a common set of properties can be grouped into an object class. Each object is an instance of its class. Object classes can be organized into class/subclass hierarchies

ADVANCED DATA AND INFORMATION SYSTEMS

- With the progress of database technology, various kinds of advanced data and information systems have emerged and are undergoing development to address the requirements of new applications
  - handling spatial/temporal data (such as maps)
  - engineering design data (such as the design of buildings, system components, or integrated circuits)
  - hypertext and multimedia data (including text, image, video, and audio data)
  - time-related data (such as historical records or stock exchange data)
  - stream data (such as video surveillance and sensor data, where data flow in and out like streams)
  - the World Wide Web (a huge, widely distributed information repository made available by the Internet)

THE WORLD WIDE WEB

- The World Wide Web and its associated distributed information services, such as Yahoo! and Google provide rich, worldwide, on-line information services, where data objects are linked together to facilitate interactive access
- Capturing user access patterns in such distributed information environments is called *Web usage mining* (or Weblog mining)

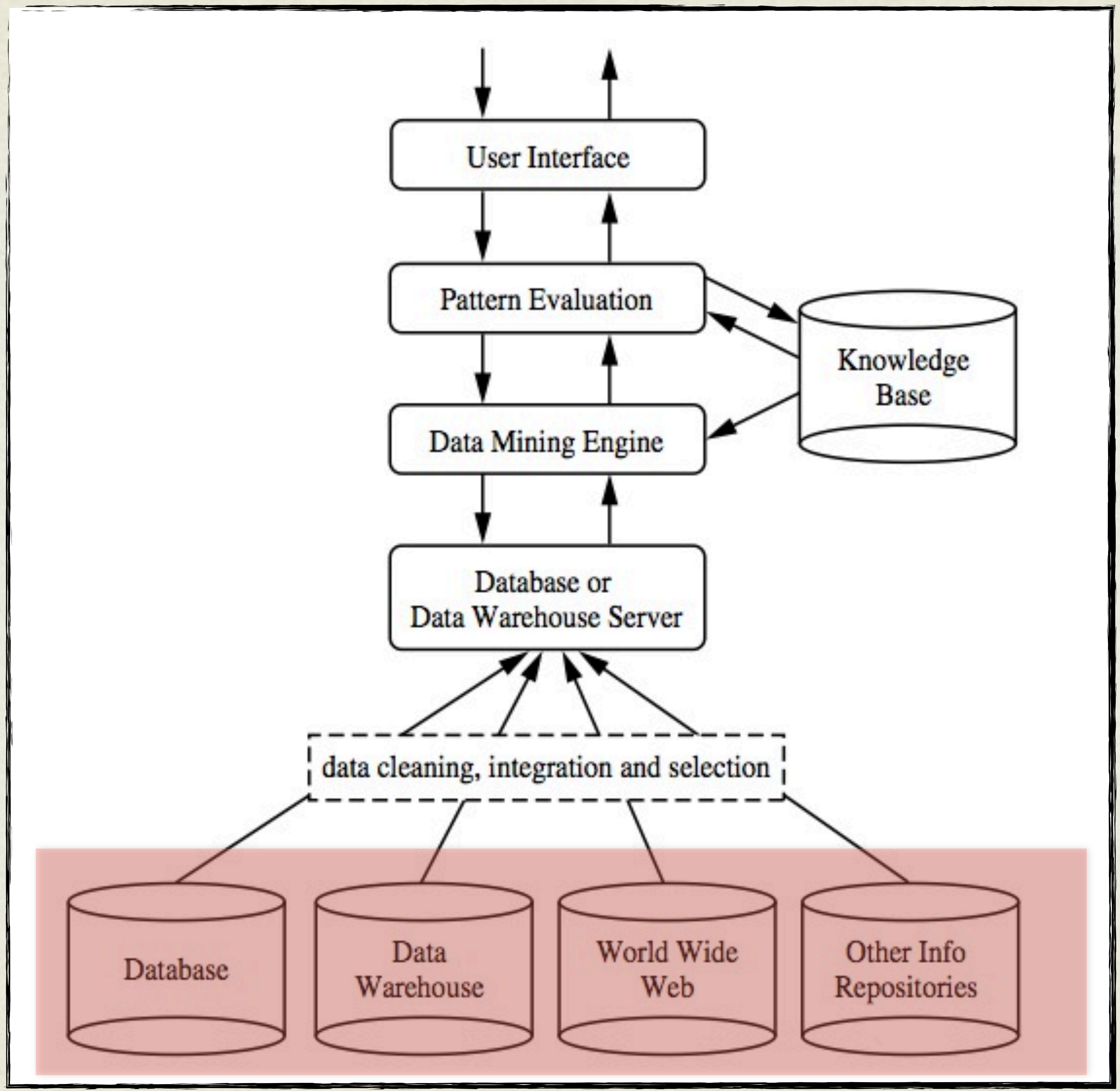
THE WORLD WIDE WEB - II

- Although Web pages may appear fancy and informative to human readers, they can be highly unstructured and lack a predefined schema, type, or pattern. Thus it is difficult for computers to understand the semantic meaning of diverse Web pages and structure them in an organized way for systematic information retrieval and data mining.
- Automated Web page clustering and classification help group and arrange Web pages in a multidimensional manner based on their contents.
- Web community analysis helps identify hidden Web social networks and communities and observe their evolution

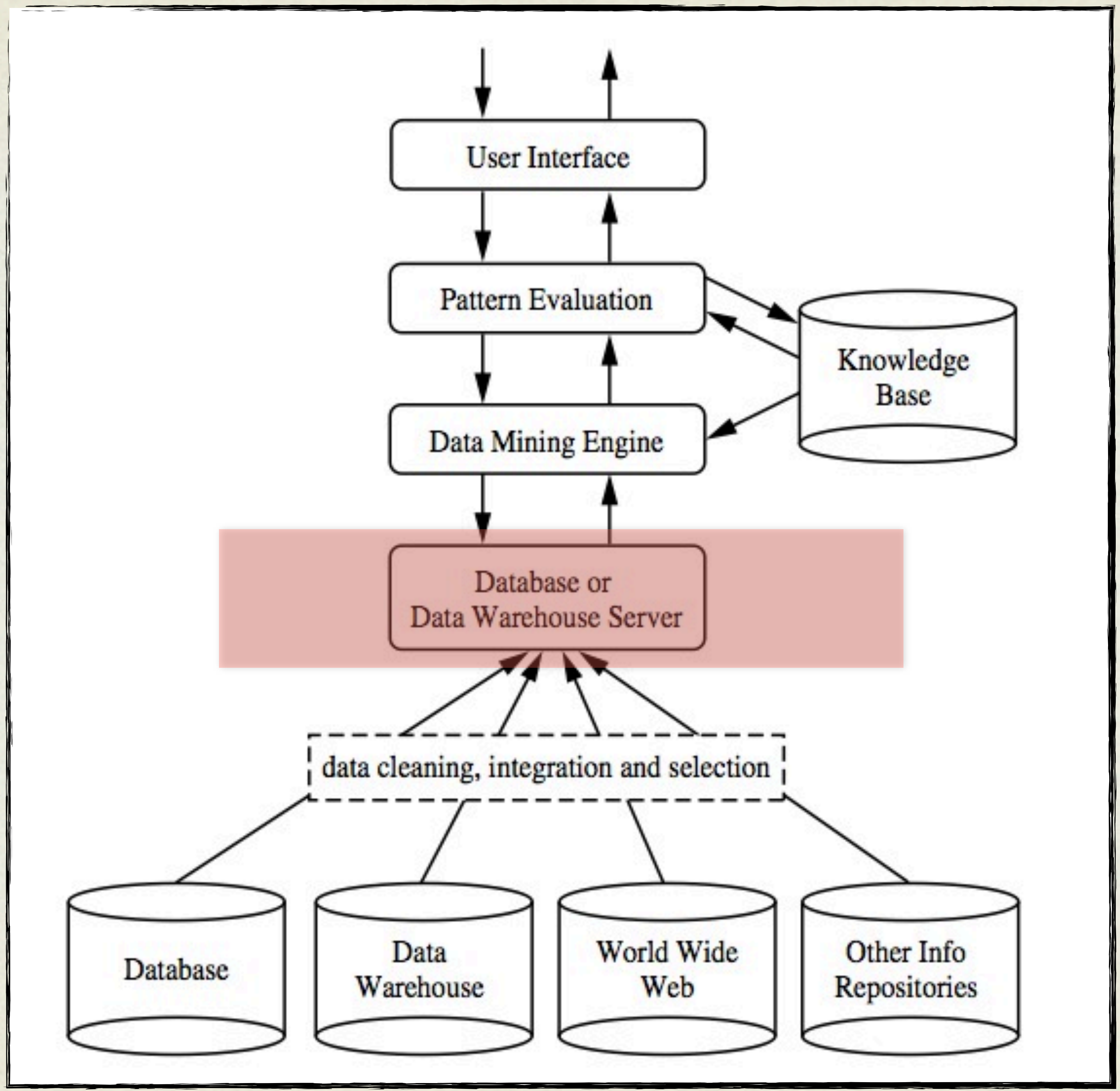
DATA MINING ARCHITECTURE

domenica 20 marzo 2011

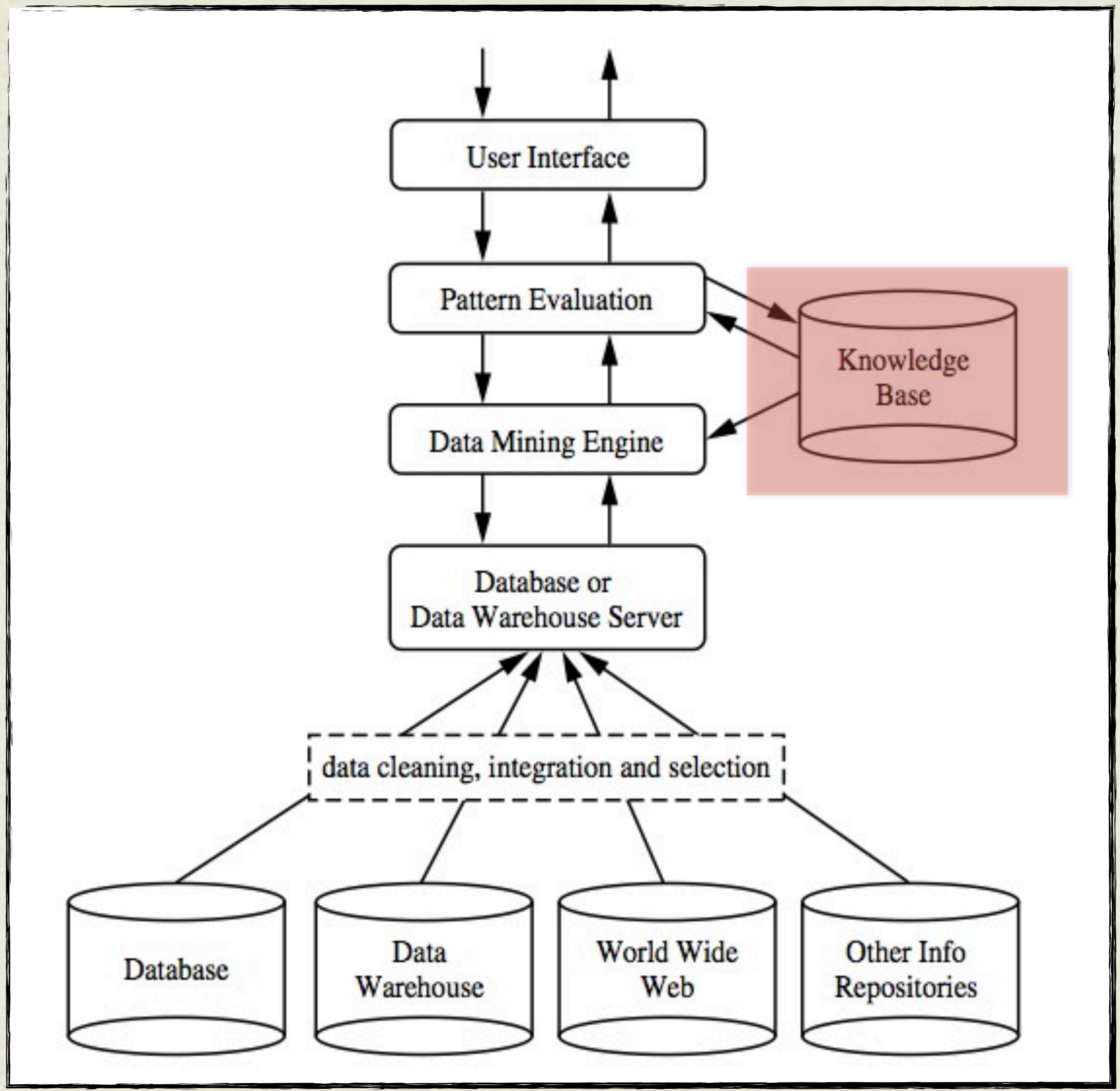
The architecture of a typical data mining system may have the following major components


User Interface
Pattern Evaluation
Knowledge Base
Data Mining Engine
Database or Data Warehouse Server
data cleaning, integration and selection
Database
Data Warehouse
World Wide Web
Other Info Repositories

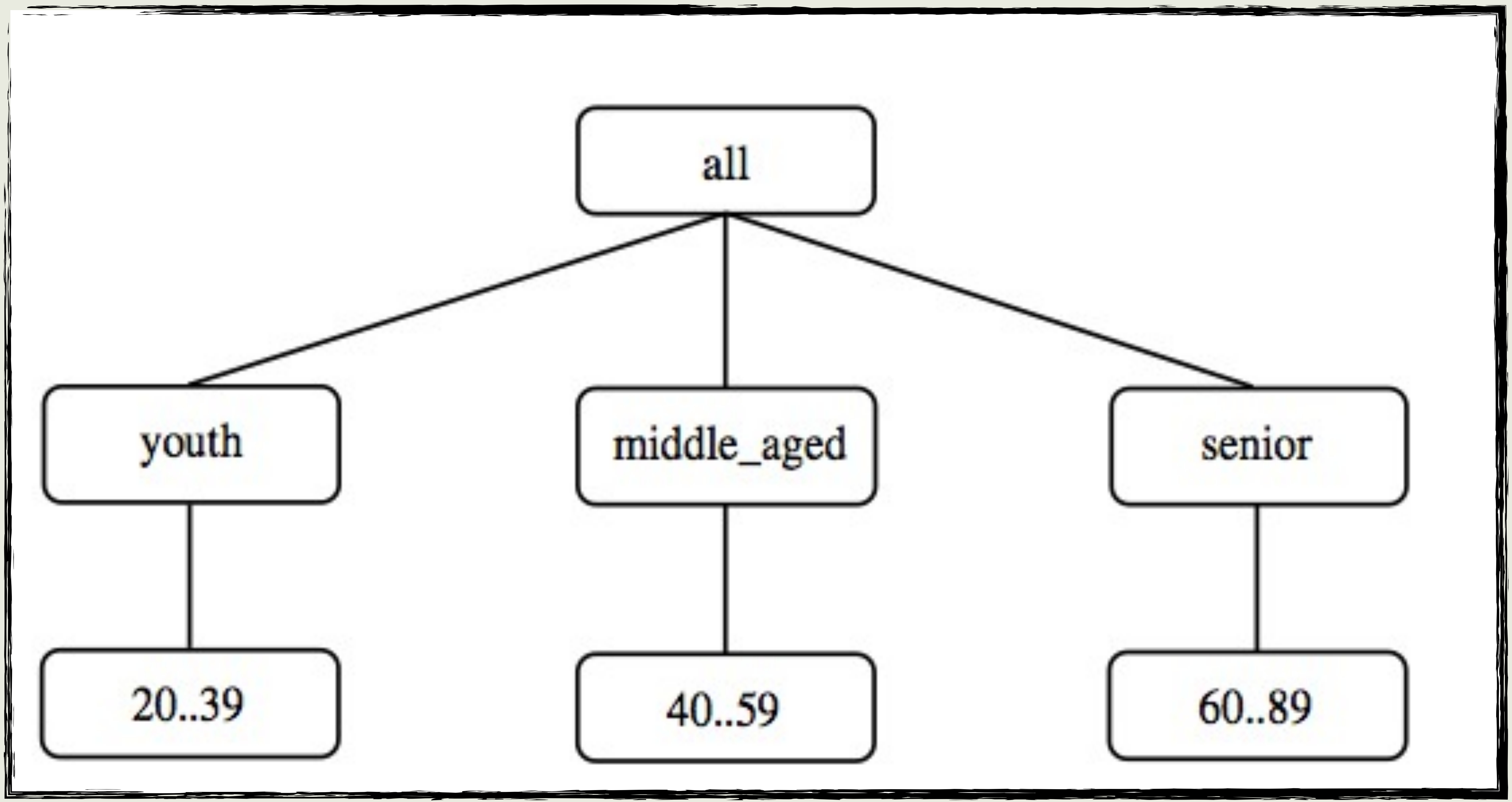
- Database, data warehouse, World Wide Web, or other information repository
    - one or a set of databases, data warehouses, spreadsheets, or other kinds of information repositories
    - data cleaning and data integration techniques may be performed on the data


User Interface
Pattern Evaluation
Knowledge Base
Data Mining Engine
Database or Data Warehouse Server
data cleaning, integration and selection
Database
Data Warehouse
World Wide Web
Other Info Repositories

- Database or data warehouse server
    - responsible for fetching the relevant data, based on the user’s data mining request
    - can be decouples/loose coupled/tightly coupled with the database layer

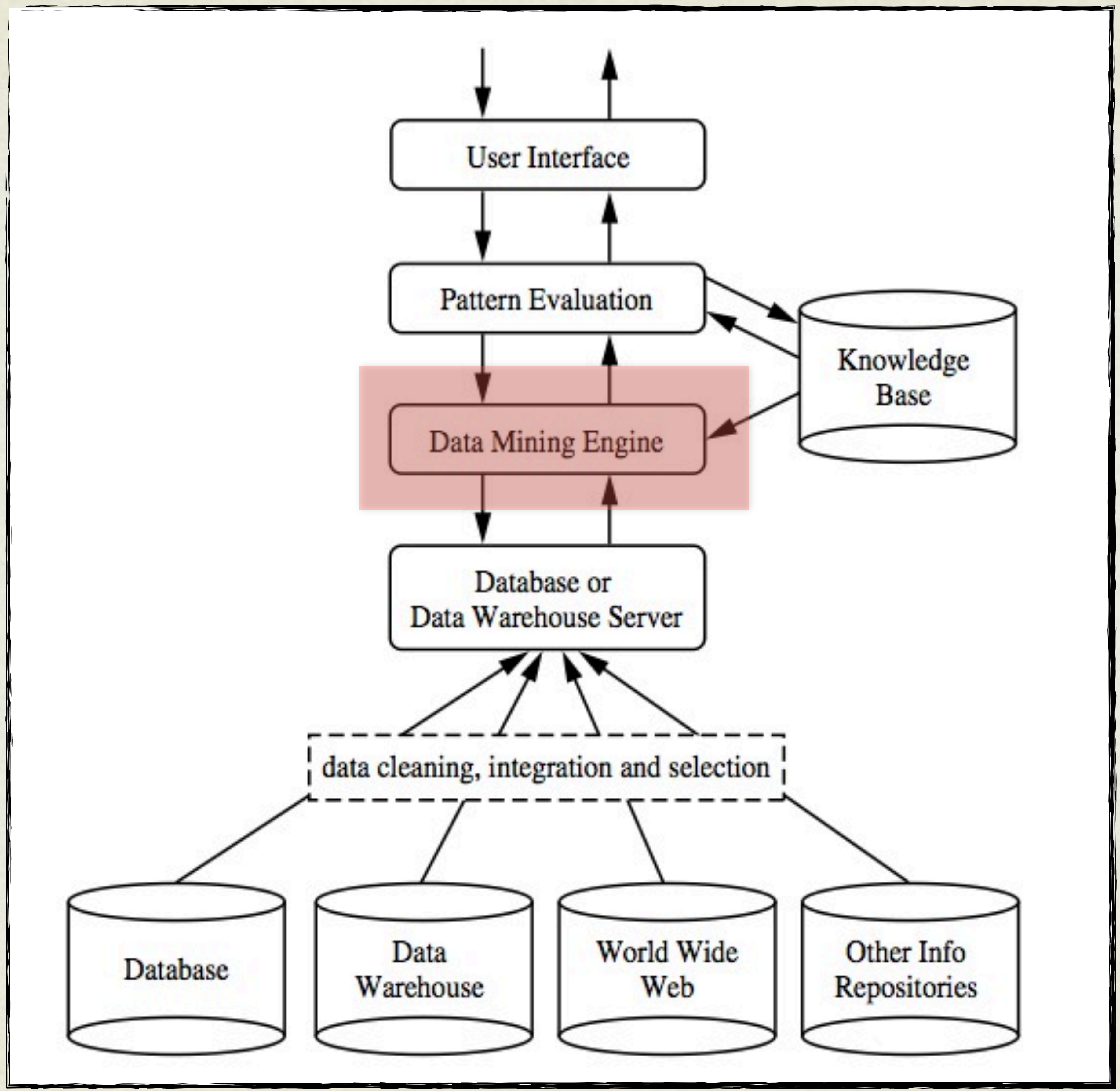

User Interface
Pattern Evaluation
Knowledge Base
Data Mining Engine
Database or Data Warehouse Server
data cleaning, integration and selection
Database
Data Warehouse
World Wide Web
Other Info Repositories

- Knowledge base
  - the domain knowledge that is used to guide the search or evaluate the interestingness of resulting patterns
  - interestingness constraints or thresholds, metadata, concept hierarchies, etc.

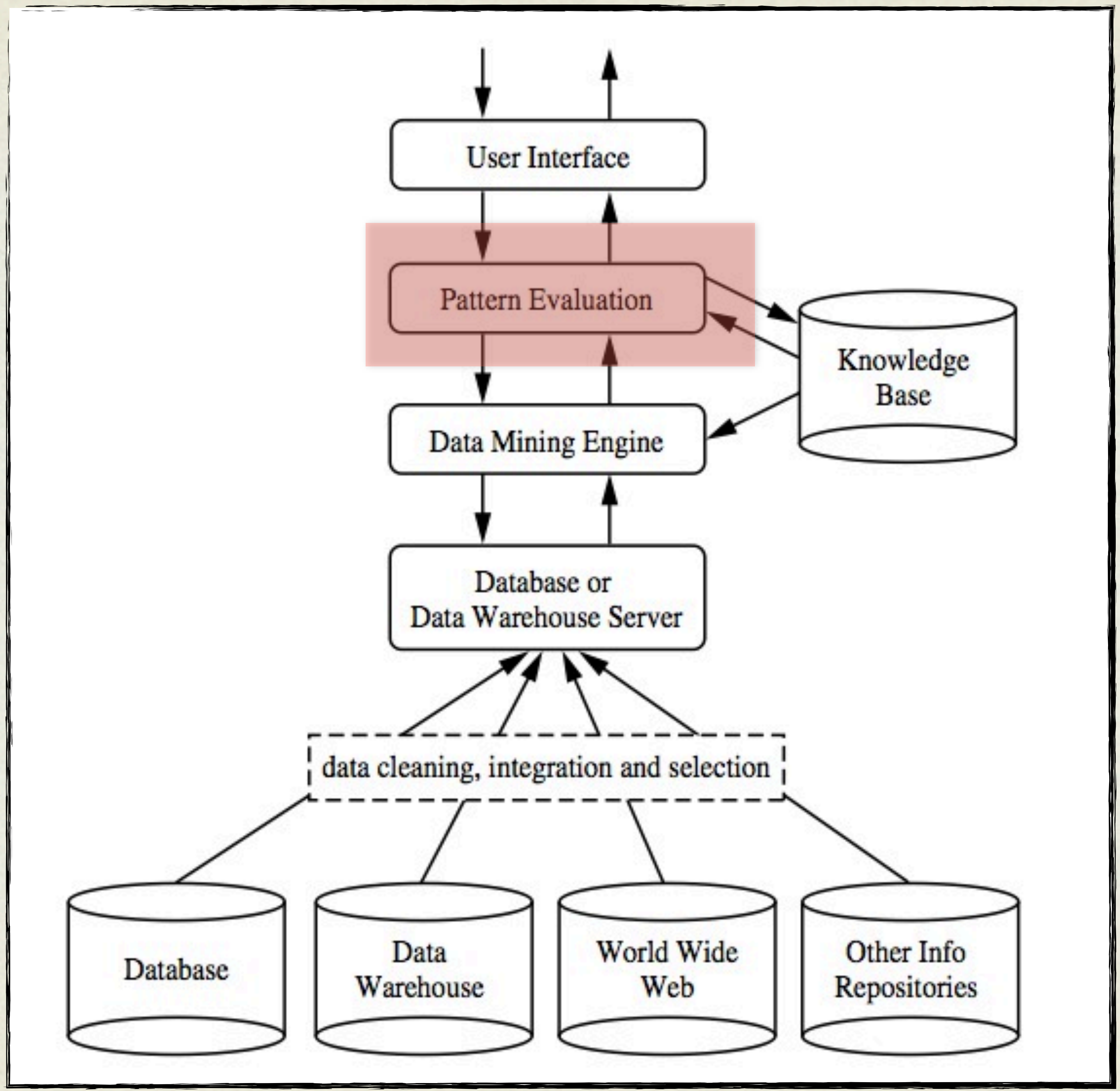


domenica 20 marzo 2011

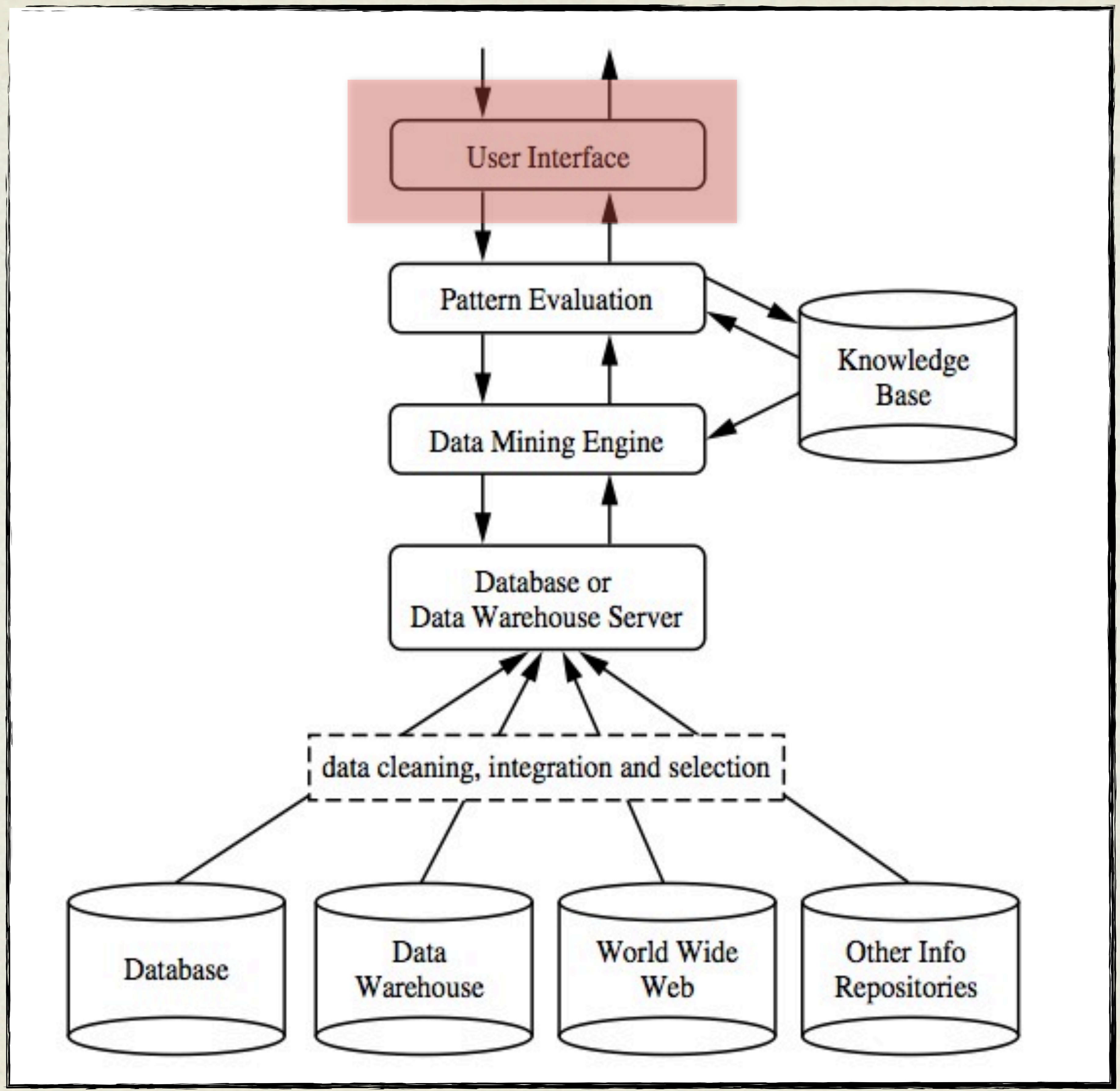
A concept hierarchy for the attribute (or dimension) age. The root node represents the most general abstraction level, denoted as all.


User Interface
Pattern Evaluation
Knowledge Base
Data Mining Engine
Database or Data Warehouse Server
data cleaning, integration and selection
Database
Data Warehouse
World Wide Web
Other Info Repositories

- 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 analysis, outlier analysis, and evolution analysis
  - query languages (DMQL) based on mining primitives to access the data


User Interface
Pattern Evaluation
Knowledge Base
Data Mining Engine
Database or Data Warehouse Server
data cleaning, integration and selection
Database
Data Warehouse
World Wide Web
Other Info Repositories

- Pattern evaluation module
  - interacts with the data mining modules so as to *focus* the search toward interesting patterns
  - may use interestingness *thresholds* to filter out discovered patterns
  - may be *integrated* with the mining module


User Interface
Pattern Evaluation
Knowledge Base
Data Mining Engine
Database or Data Warehouse Server
data cleaning, integration and selection
Database
Data Warehouse
World Wide Web
Other Info Repositories

- User interface
    - *communicates* between users and the data mining system
    - allows the user to *interact* with the system by specifying a data mining query or task, providing information to help focus the search, and performing exploratory data mining based on the intermediate data mining results
    - allows the user to *browse* database and data warehouse schemas or data structures, evaluate mined patterns, and visualize the patterns in different forms

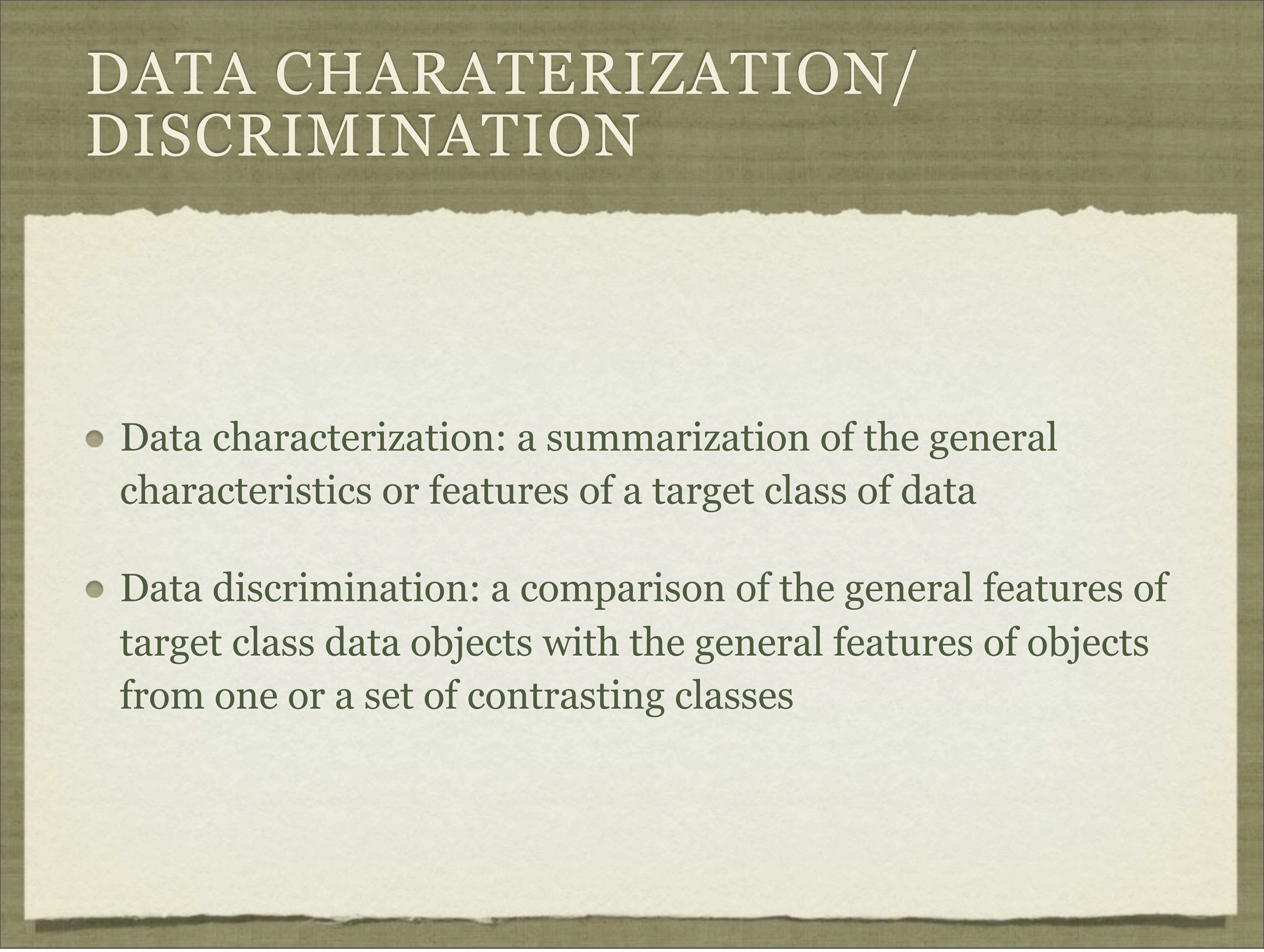
WHAT KIND OF PATTERNS CAN BE MINED?

DESCRIPTIVE & PREDICTIVE

- Data mining tasks can be classified into two categories: *descriptive* and *predictive*:
    - descriptive mining tasks characterize the general properties of the data in the database.
    - predictive mining tasks perform inference on the current data in order to make predictions

CONCEPT/CLASS

- Data can be associated with classes or concepts
- classes of items for sale include *computers* and *printers*, and concepts of customers include *bigSpenders* and *budgetSpenders*



DATA CHARATERIZATION/ DISCRIMINATION

- Data characterization: a summarization of the general characteristics or features of a target class of data
- Data discrimination: 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

domenica 20 marzo 2011

Charaterization: printers, computers
Discrimination: spendono tanto/ spendono poco

MINING FREQUENT PATTERNS

- Frequent patterns, as the name suggests, are patterns that occur frequently in data.
- There are many kinds of frequent patterns, including itemsets, subsequences, and substructures.
- A frequent itemset typically refers to a set of items that frequently appear together in a transactional data set, such as milk and bread
- Mining frequent patterns leads to the discovery of interesting associations and correlations within data

ASSOCIATION ANALYSIS

age(X , “20...29”) ∧ income(X , “20K...29K”) ⇒ buys(X , “CD player”)
[support = 2%, confidence = 60%]

- Frequent itemset mining is the simplest form of frequent pattern mining
- Example: determine which items are frequently purchased together within the same transactions

CONFIDENCE/SUPPORT

- A *confidence*, or certainty, of 50% means that if a customer buys a computer, there is a 50% chance that she will buy software as well.

- A 1% *support* means that 1% of all of the transactions under analysis showed that computer and software were purchased together

MINIMUM SUPPORT/CONFIDENCE

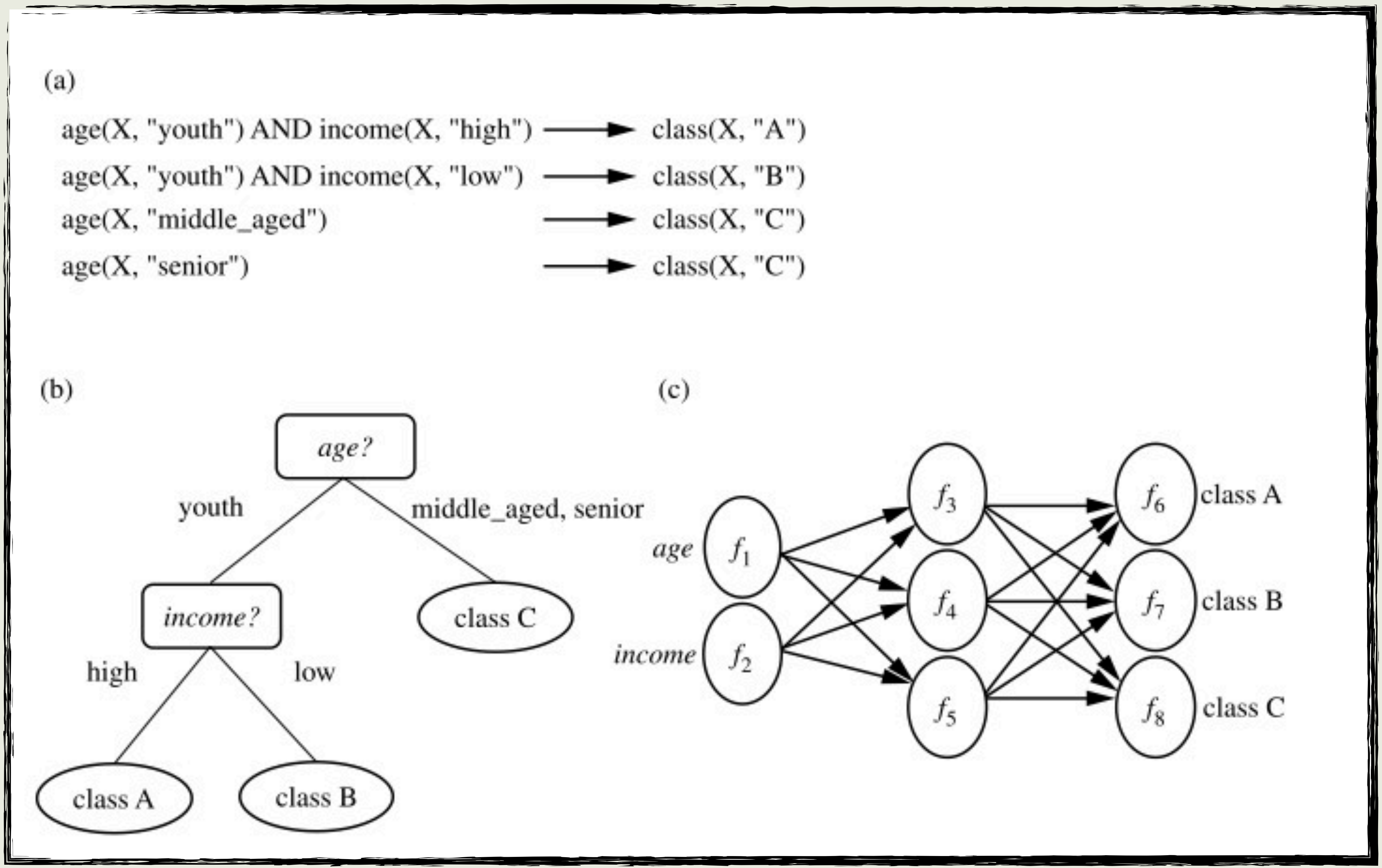
- association rules are discarded as uninteresting if they do not satisfy both a minimum support threshold and a minimum confidence threshold.
- Additional analysis can be performed to uncover interesting statistical correlations

CLASSIFICATION/PREDICTION

- 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 is based on the analysis of a set of training data
- prediction models continuous-valued functions

“How is the derived model presented?”

The derived model may be represented in various forms, such as classification (IF-THEN) rules, decision trees, mathematical formulae, or neural networks



(a)

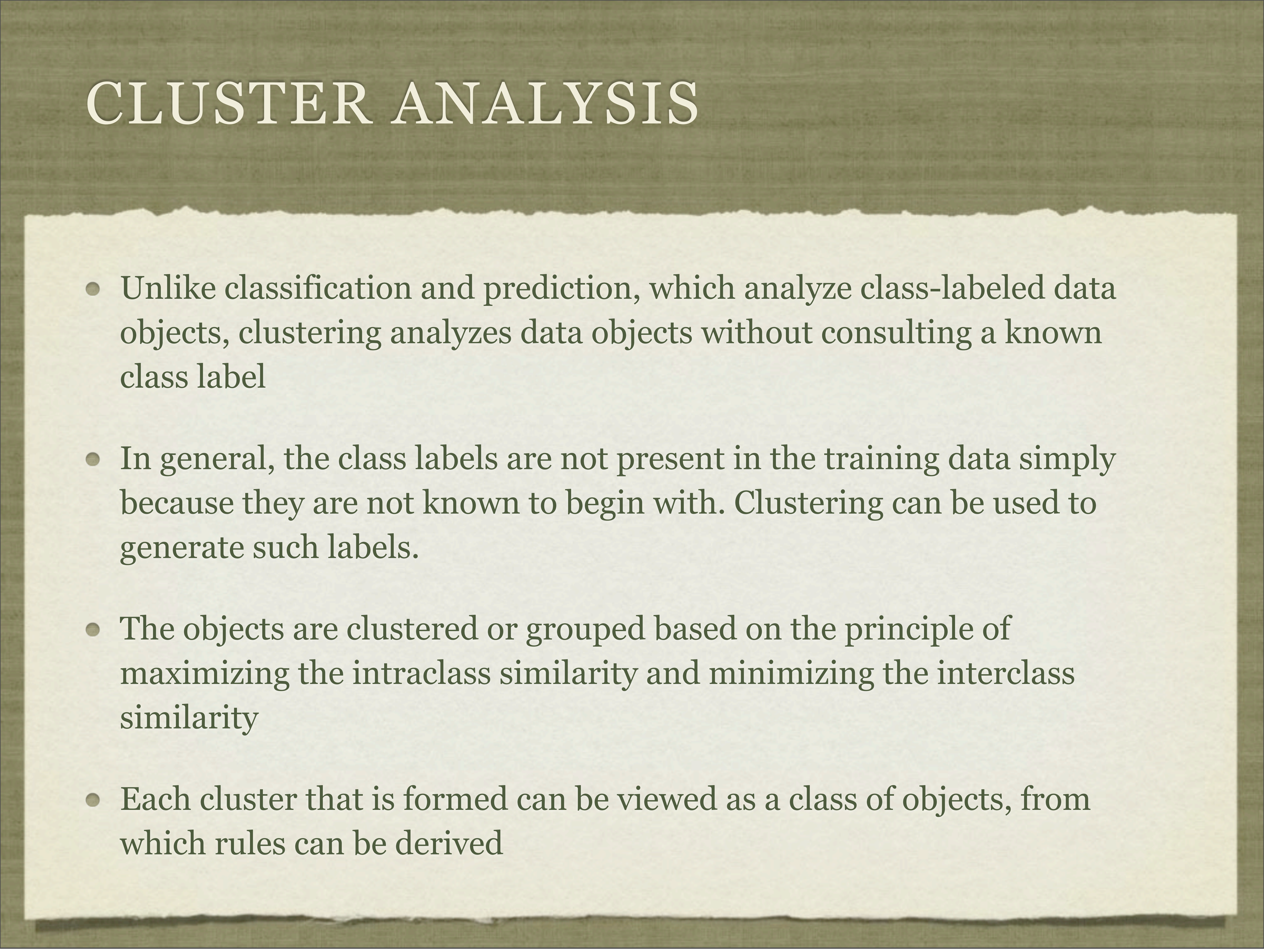
age(X, "youth") AND income(X, "high") ⟶ class(X, "A")

age(X, "youth") AND income(X, "low") ⟶ class(X, "B")

age(X, "middle_aged") ⟶ class(X, "C")

age(X, "senior") ⟶ class(X, "C")

A classification model can be represented in various forms, such as (a) IF-THEN rules, (b) a decision tree, or a (c) neural network

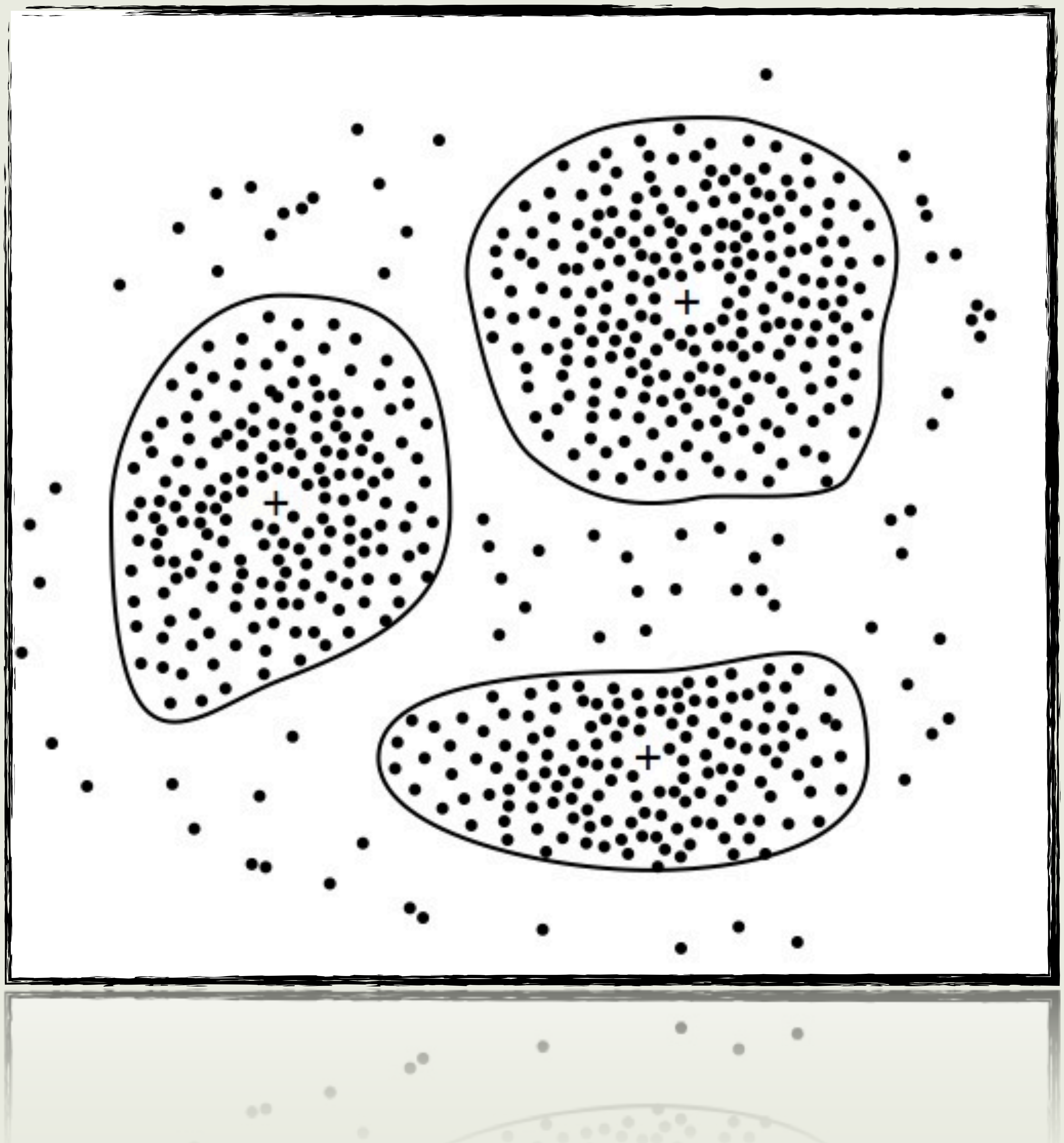


CLUSTER ANALYSIS

- Unlike classification and prediction, which analyze class-labeled data objects, clustering analyzes data objects without consulting a known class label
- In general, the class labels are not present in the training data simply because they are not known to begin with. Clustering can be used to generate such labels.
- The objects are clustered or grouped based on the principle of maximizing the intraclass similarity and minimizing the interclass similarity
- Each cluster that is formed can be viewed as a class of objects, from which rules can be derived

domenica 20 marzo 2011

esempio: cluster analysis dei compratori



A 2-D plot of customer data with respect to customer locations in a city, showing three data clusters. Each cluster “center” is marked with a “+”.

OUTLIERS ANALYSIS

- Data objects that do not comply with the general behavior or model of the data
  - most analysis discard outliers as noise or exceptions
- Outliers may be detected using statistical tests, or using distance measures where objects that are a substantial distance from any other cluster are considered outliers
- Example: outlier analysis may uncover fraudulent usage of credit cards by detecting purchases of extremely large amounts for a given account number in comparison to regular charges incurred by the same account

"ARE ALL PATTERNS INTERESTING?"

INTERESTING PATTERNS

- only a small fraction of the patterns potentially generated would actually be of interest to any given user
- 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

INTERESTING PATTERNS - II

- Pattern is also interesting if it validates a hypothesis that the user sought to confirm.
- An interesting pattern represents knowledge

INTERESTINGNESS MEASURES

- Objective measures of pattern interestingness exist (support, confidence)
- Insufficient unless combined with subjective measures that reflect the needs and interests of a particular user
- Many patterns represent common knowledge (i.e. womens buy most makeups)
- A pattern is interesting if it is *unexcepted* or if they confirm an hypothesis

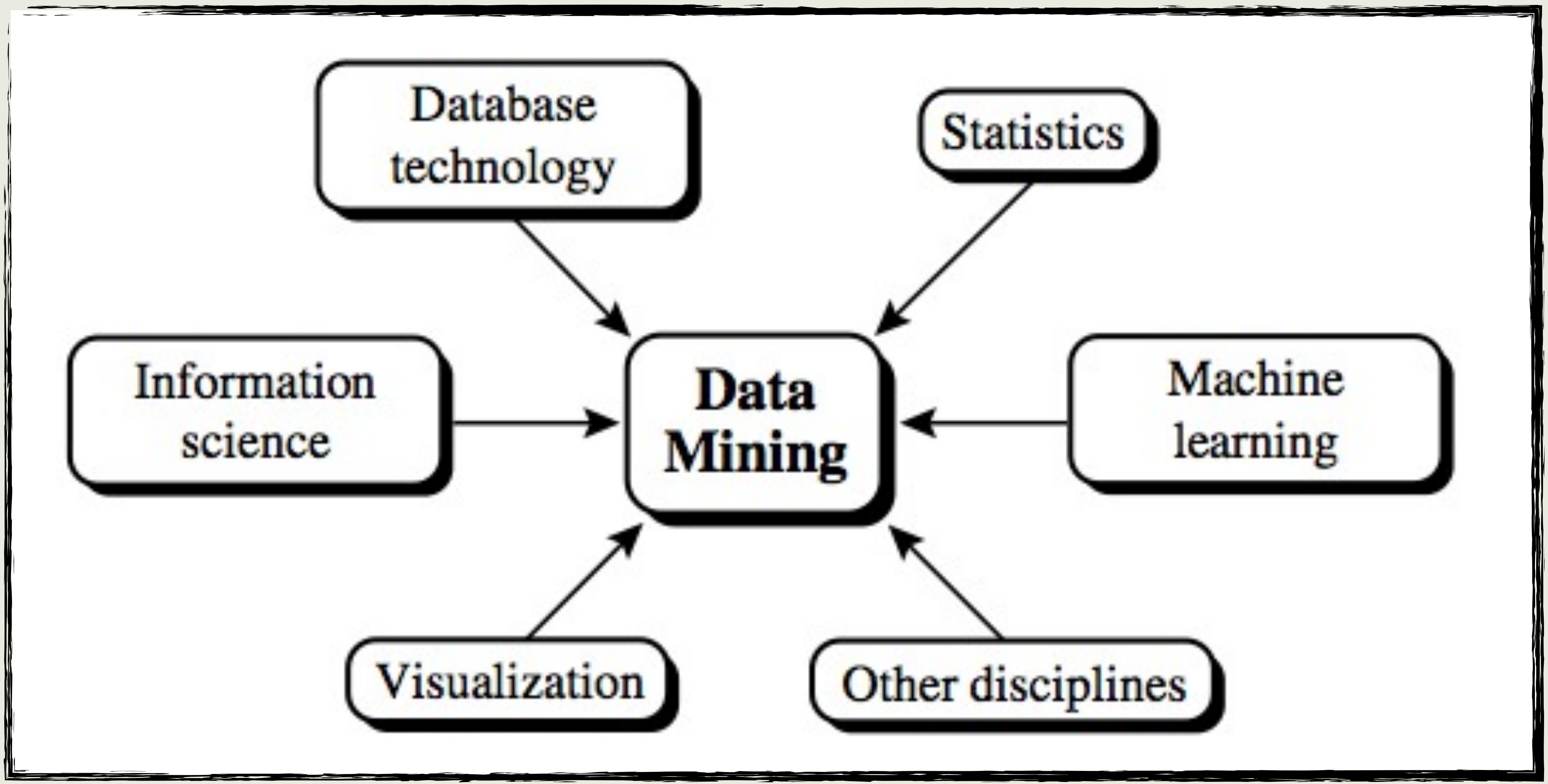
“Can a data mining system generate all of the interesting patterns?”

It is often unrealistic and inefficient for data mining systems to generate all of the possible patterns. Instead, user-provided constraints and interestingness measures should be used to focus the search.

“Can a data mining system generate only interesting patterns?”

It is highly desirable for data mining systems to generate only interesting patterns. It’s an optimization problem.

CLASSIFICATION OF DATA MINING SYSTEMS



domenica 20 marzo 2011

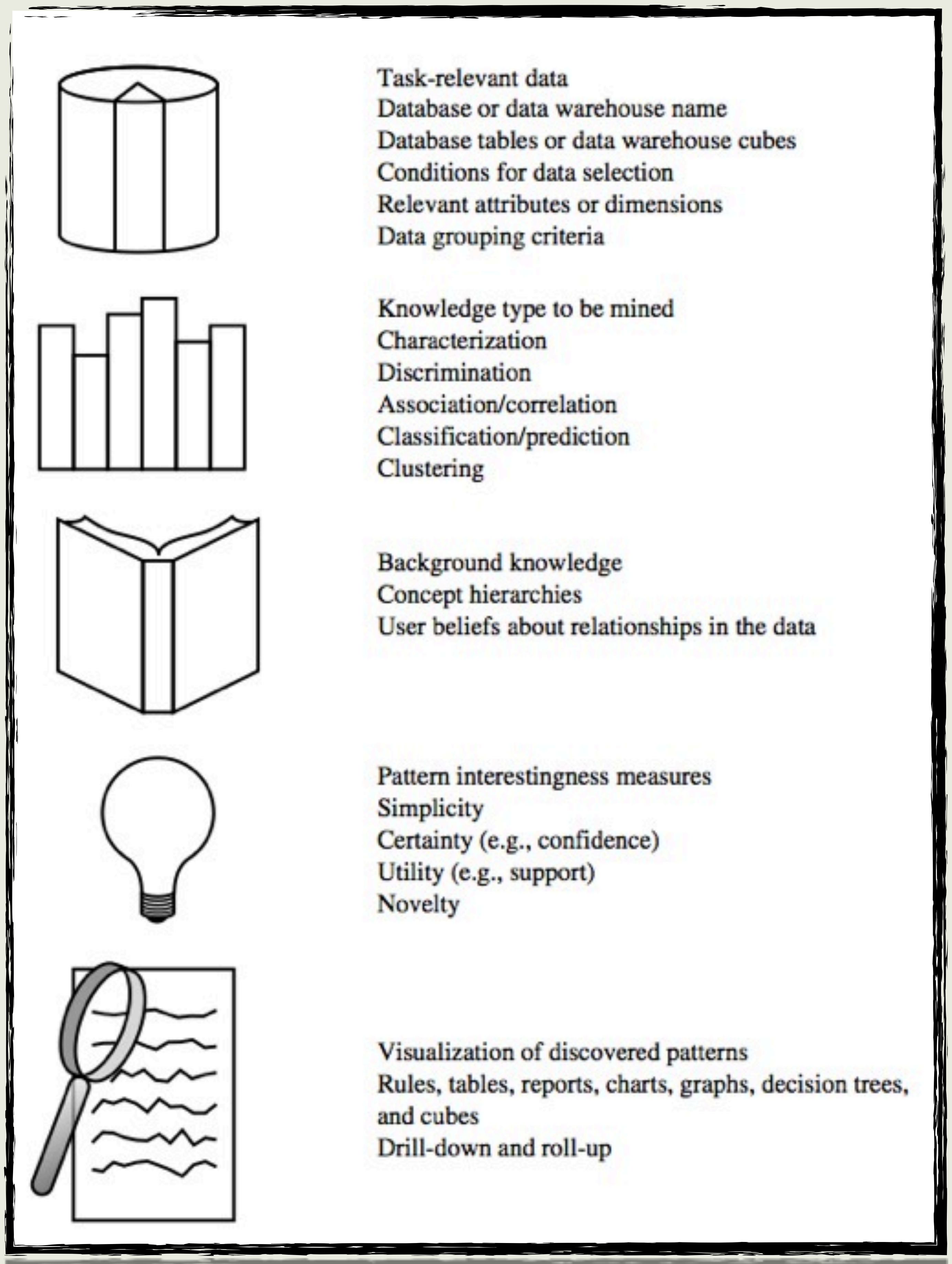
Data mining as a confluence of multiple disciplines. interdisciplinary field, the confluence of a set of disciplines, including database systems, statistics, machine learning, visualization, and information science

CLASSIFICATIONS

- Kinds of databases mined
- Kinds of knowledge mined
- Kinds of techniques utilized
- Applications adopted

DATA MINING TASK

- Each user will have a data mining task in mind, that is, some form of data analysis that he or she would like to have performed.
- 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 to interactively communicate with the mining system



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

Primitives for specifying a data mining task.

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 performed, such as characterization, discrimination, association or correlation analysis, classification, prediction, clustering, outlier analysis, or evolution analysis.

background knowledge: Concept hierarchies are a popular form of back- ground knowledge. knowledge about the domain to be mined is useful for guiding the knowledge discovery process and for evaluating the patterns found

he **interestingness measures** and thresholds for pattern evaluation: They may be used to guide the mining process or, after discovery, to evaluate the discovered patterns. Different kinds of knowledge may have different interestingness measures. For exam- ple, interestingness measures for association rules include support and confidence.

The expected representation for visualizing the discovered patterns

QUERY LANGUAGES

- A data mining query language can be designed to incorporate these primitives, allowing users to flexibly interact with data mining systems
- DMQL (Data Mining Query Language), which was designed as a teaching tool, based on the above primitives

DMQL EXAMPLE

```
use database AllElectronics db

use hierarchy location hierarchy for T.branch, age hierarchy
for C.age

mine classification as promising customers

in relevance to C.age, C.income, I.type, I.place made, T.branch

from customer C, item I, transaction T

where I.item ID = T.item ID and C.cust ID = T.cust ID

and C.income ≥ 40,000 and I.price ≥ 100

group by T.cust ID

having sum(I.price) ≥ 1,000

display as rules
```

CONCLUSIONS

- Data mining is the task of discovering interesting patterns from large amounts of data, where the data can be stored in databases, data warehouses, or other information repositories.

- It is a young interdisciplinary field, drawing from areas such as database systems, data warehousing, statistics, machine learning, data visualization, information retrieval, and high-performance computing.

- Other contributing areas include neural networks, pattern recognition, spatial data analysis, image databases, signal processing, and many application fields, such as business, economics, and bioinformatics.

CHALLENGES

- Efficient and effective data mining in large databases poses numerous requirements and great challenges to researchers and developers.
- The issues involved include data mining methodology, user interaction, performance and scalability, and the processing of a large variety of data types.
- Other issues include the exploration of data mining applications and their social impacts.