Data Ware Housing and Mining

Unit-1

Data Mining

Dr. K. Raghava Rao
Professor of CSE,
Dept. of MCA, KL University

Data Mining

Introduction: 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 from

-Business: Web, e-commerce, transactions, stocks, …

-Science: Remote sensing, bioinformatics, scientific simulation, …

-Society and everyone: news, digital cameras, YouTube

>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

Introduction: What is Data Mining?

>Definition: Data mining (knowledge discovery from data) Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data.

>Alternative names for Data Mining:

-Knowledge discovery (mining) in databases (KDD)
-knowledge extraction
-data/pattern analysis
- data archeology
- data dredging
- information harvesting
- business intelligence, etc.

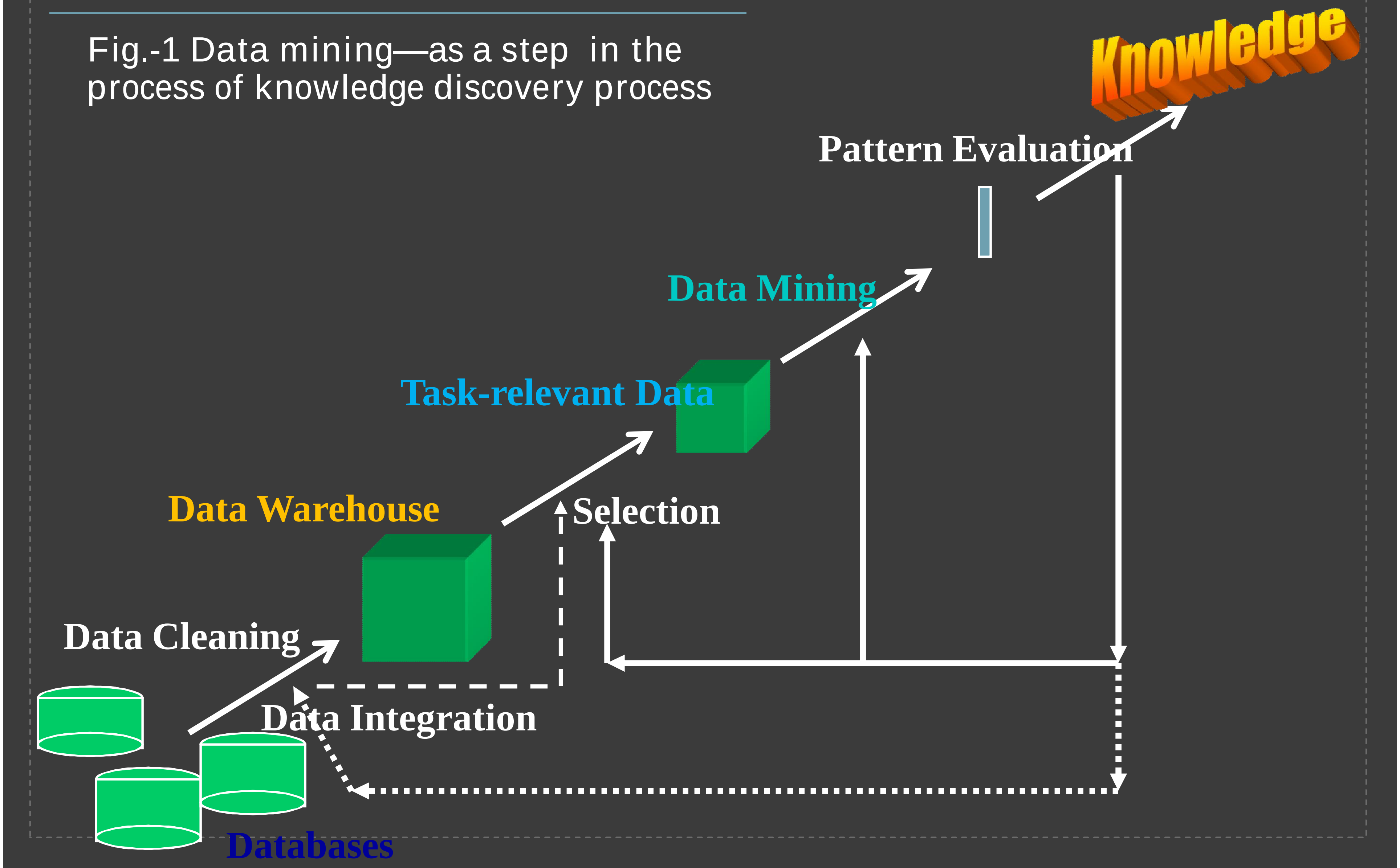
>Data Mining is **NOT**

-Simple search and query processing
-(Deductive) expert systems

Data Mining

Introduction: What is Data Mining?

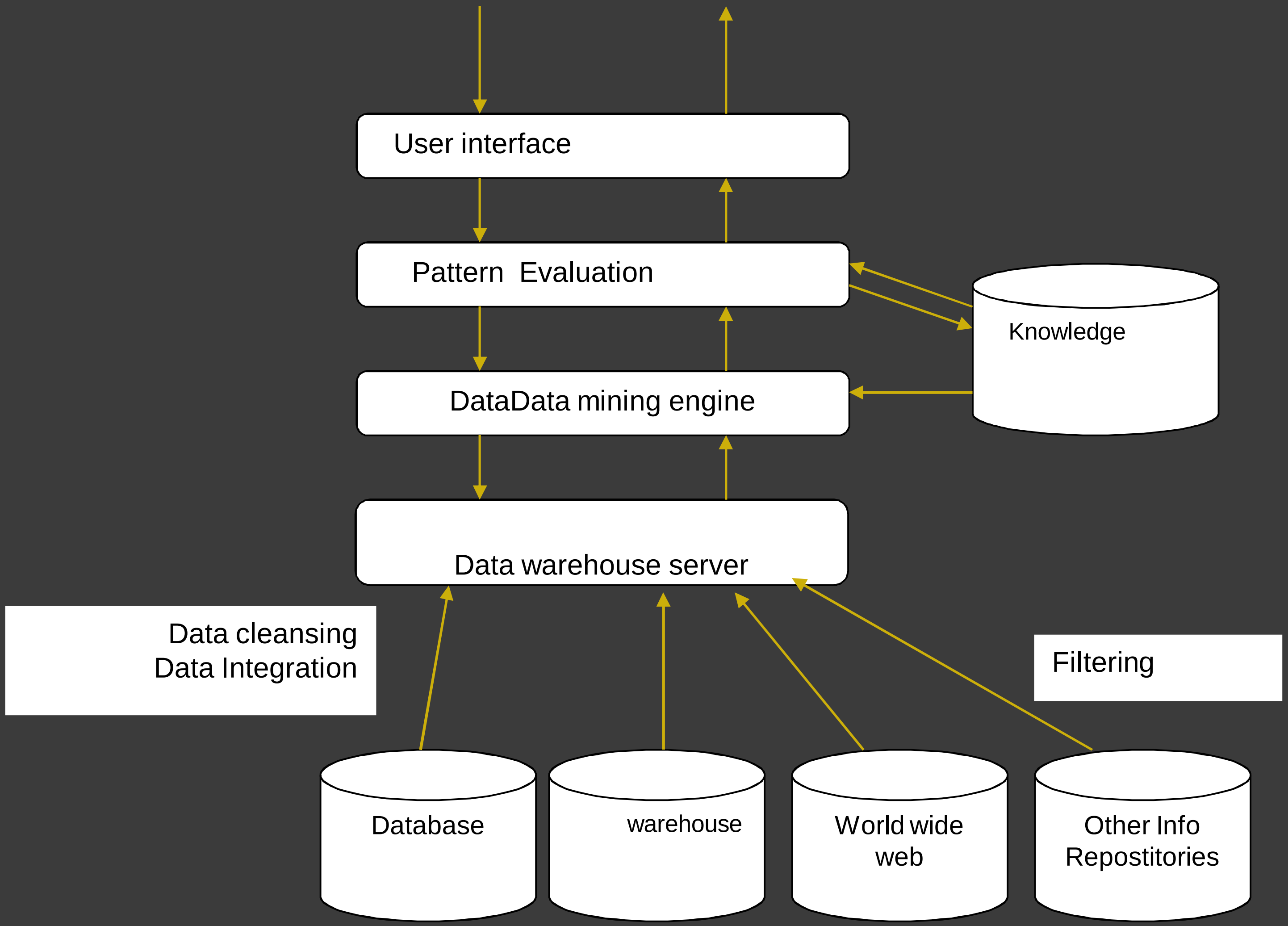
Fig.-1 Data mining—as a step in the process of knowledge discovery process



Data Mining

Introduction: What is Data Mining?

Fig-2. Architecture of Typical Data Mining System.



Data Mining

Introduction: Data Mining on What Kind of Data?

>Database-oriented data sets and applications

- -Relational database, data warehouse, transactional database

>Advanced data sets and advanced applications

- -Data streams and sensor data
- -Time-series data, temporal data, sequence data (incl. bio-sequences)
- -Structure data, graphs, social networks and multi-linked data
- -Object-relational databases
- -Heterogeneous databases and legacy databases
- -Spatial data and spatiotemporal data
- -Multimedia database
- -Text databases
- -The World-Wide Web

Data Mining

Introduction: Data Mining Functionalities

Definition: Data Mining functionalities are used to specify the kind of patterns to be found in data mining tasks.

1)Concept/Class description:

Characterization and discrimination: Generalize, summarize, and contrast data characteristics, e.g., dry vs. wet regions.

2) Mining Frequent patterns, association, correlations vs. causality

Diaper → Beer [0.5%, 75%] (Correlation or causality?)

3)Classification and prediction

Construct models (functions) that describe and distinguish classes or concepts for future prediction

E.g., classify countries based on (climate), or classify cars based on (gas mileage)

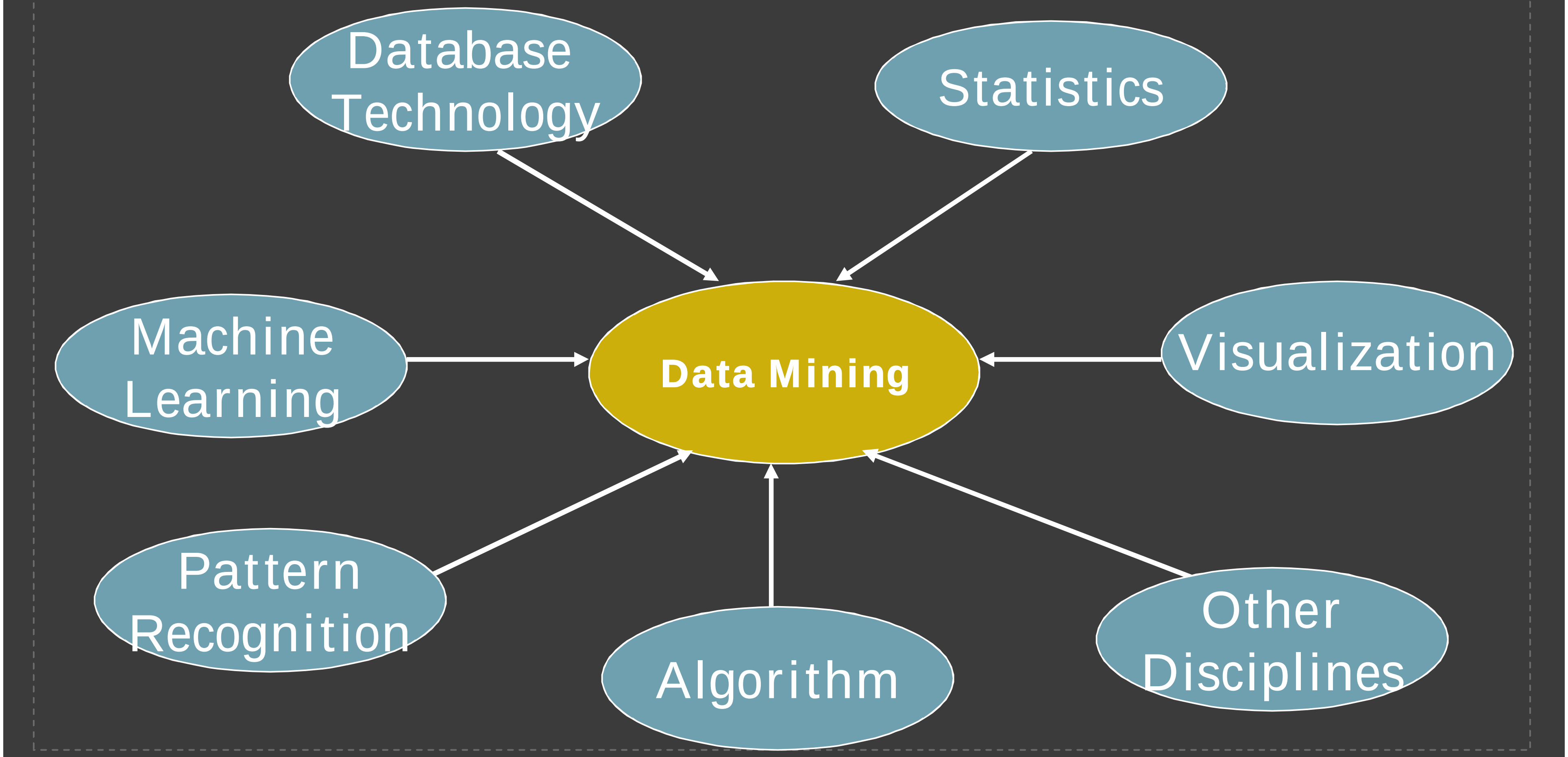
Predict some unknown or missing numerical values

4) Evolution Analysis describes and models regularities or trends for objects whose behavior changes over time.

Data Mining

Introduction Classification of Data Mining Systems

Fig. Classification of Data Mining Systems



Data Mining

Introduction: Classification of Data Mining Systems

A data mining system can be classified according to kinds of databases mined.

1) classification according to the kinds of knowledge mined.

2) Classification according to the kinds of techniques utilized

3) Classification according to the application adapted

4) Classification according to the data model i.e relational, transactional, object-relational or data warehousing mining system.

5) Classification according special data type handled i.e spatial, time-series , text stream data ,multimedia data or world wide web data

Data Mining

Introduction: Data Mining Task Primitives

Data mining primitives define a data mining task, which can be specified in the form of a data mining query.

-Task Relevant Data

-Kinds of knowledge to be mined

-Background knowledge

-Interestingness measure

-Presentation and visualization of discovered patterns

Data Mining

Introduction: Data Mining Task Primitives

Task relevant data

-Data portion to be investigated.

-Attributes of interest (relevant attributes) can be specified.

-Initial data relation

-Minable view

Task relevant data: Example

>-If a data mining task is to study associations between items frequently purchased at *AllElectronics* by customers in Canada, the task relevant data can be specified by providing the following information:

- Name of the *database or data warehouse* to be used (e.g., *AllElectronics_db*)
- Names of the *tables or data cubes* containing relevant data (e.g., *item, customer, purchases* and *items_sold*)
- *Conditions* for selecting the relevant data (e.g., retrieve data pertaining to purchases made in Canada for the current year)
- The *relevant attributes or dimensions* (e.g., *name* and *price* from the *item* table and income and age from the customer table)

Introduction: Data Mining Task Primitives

Kinds of Knowledge to be Mined

->t is important to specify the knowledge to be mined, as this determines the data mining function to be performed.

->Kinds of knowledge include concept description, association, classification, prediction and clustering.

->User can also provide pattern templates. Also called metapatterns or metarules or metaqueries.

Data Mining

Introduction: Data Mining Task Primitives

Kinds of Knowledge to be Mined; Example

>-A user studying the buying habits of *allelectronics* customers may choose to mine *association rules* of the form:

P (X:customer,W) ^ Q (X,Y) => buys (X,Z)

-Meta rules such as the following can be specified:

age (X, “30.....39”) ^ income (X, “40k....49K”) => buys (X, “VCR”)

[2.2%, 60%]

occupation (X, “student ”) ^ age (X, “20.....29”)=> buys (X, “computer”)

[1.4%, 70%]

Data Mining

Introduction: Data Mining Task Primitives

Background Knowledge

-It is the information about the domain to be mined

-Concept hierarchy: is a powerful form of background knowledge.

-Four major types of concept hierarchies:

schema hierarchies
set-grouping hierarchies
operation-derived hierarchies
rule-based hierarchies

Data Mining

Introduction: Data Mining Task Primitives

Concept Hierarchies(1)

-Defines a sequence of mappings from a set of low-level concepts to higher-level (more general) concepts.

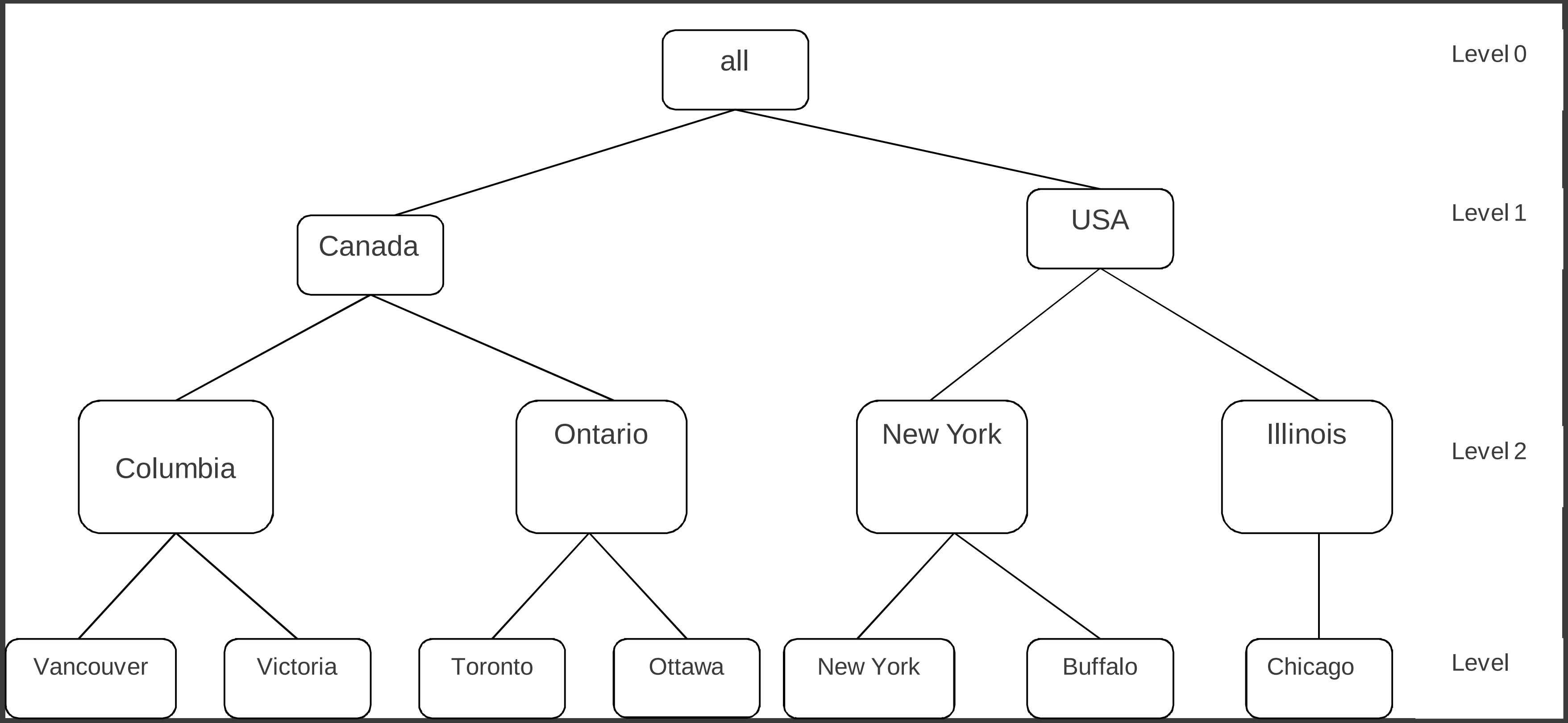
-Allows data to be mined at multiple levels of abstraction.

-These allow users to view data from different perspectives, allowing further insight into the relationships. (location)

Data Mining

Introduction: Data Mining Task Primitives

Concept Hierarchies(1): Example



Data Mining

Introduction: Data Mining Task Primitives

Concept Hierarchies(2):

>-Rolling Up - Generalization of data

- -Allows to view data at more meaningful and explicit abstractions.
- -Makes it easier to understand
- -Compresses the data
- -Would require fewer input/output operations

>-Drilling Down - Specialization of data

- -Concept values replaced by lower level concepts

>-There may be more than concept hierarchy for a given attribute or dimension based on different user viewpoints

Example:

Regional sales manager may prefer the previous concept hierarchy but marketing manager might prefer to see location with respect to linguistic lines in order to facilitate the distribution of commercial ads.

Data Mining
Introduction: Data Mining Task Primitives

Interestingness Measure(1)

>-Used to confine the number of uninteresting patterns returned by the process.

>-Based on the structure of patterns and statistics underlying them.

>-Associate a threshold which can be controlled by the user.

>-patterns not meeting the threshold are not presented to the user.

>-Objective measures of pattern interestingness:

simplicity
certainty (confidence)
utility (support)
novelty

Data Mining
Introduction: Data Mining Task Primitives

Interestingness Measure(2)

>-Simplicity

-a patterns interestingness is based on its overall simplicity for human comprehension.

Example: Rule length is a simplicity measure

>-Certainty (confidence)

Assesses the validity or trustworthiness of a pattern.

confidence is a certainty measure

confidence (A=>B) = # tuples containing both A and B
# tuples containing A

>-A confidence of 85% for the rule buys(X, “computer”)=>buys(X,“software”) means that 85% of all customers who purchased a computer also bought software

Data Mining

Introduction: Data Mining Task Primitives

Presentation and Visualization:

>-For data mining to be effective, data mining systems should be able to display the discovered patterns in multiple forms, such as rules, tables, crosstabs (cross-tabulations), pie or bar charts, decision trees, cubes, or other visual representations.

>-User must be able to specify the forms of presentation to be used for displaying the discovered patterns.

Data Mining

Introduction: Data Mining Task Primitives

Data Mining Query Language:

>-Data mining language must be designed to facilitate flexible and effective knowledge discovery.

>-Having a query language for data mining may help standardize the development of platforms for data mining systems.

>-But designed a language is challenging because data mining covers a wide spectrum of tasks and each task has different requirement.

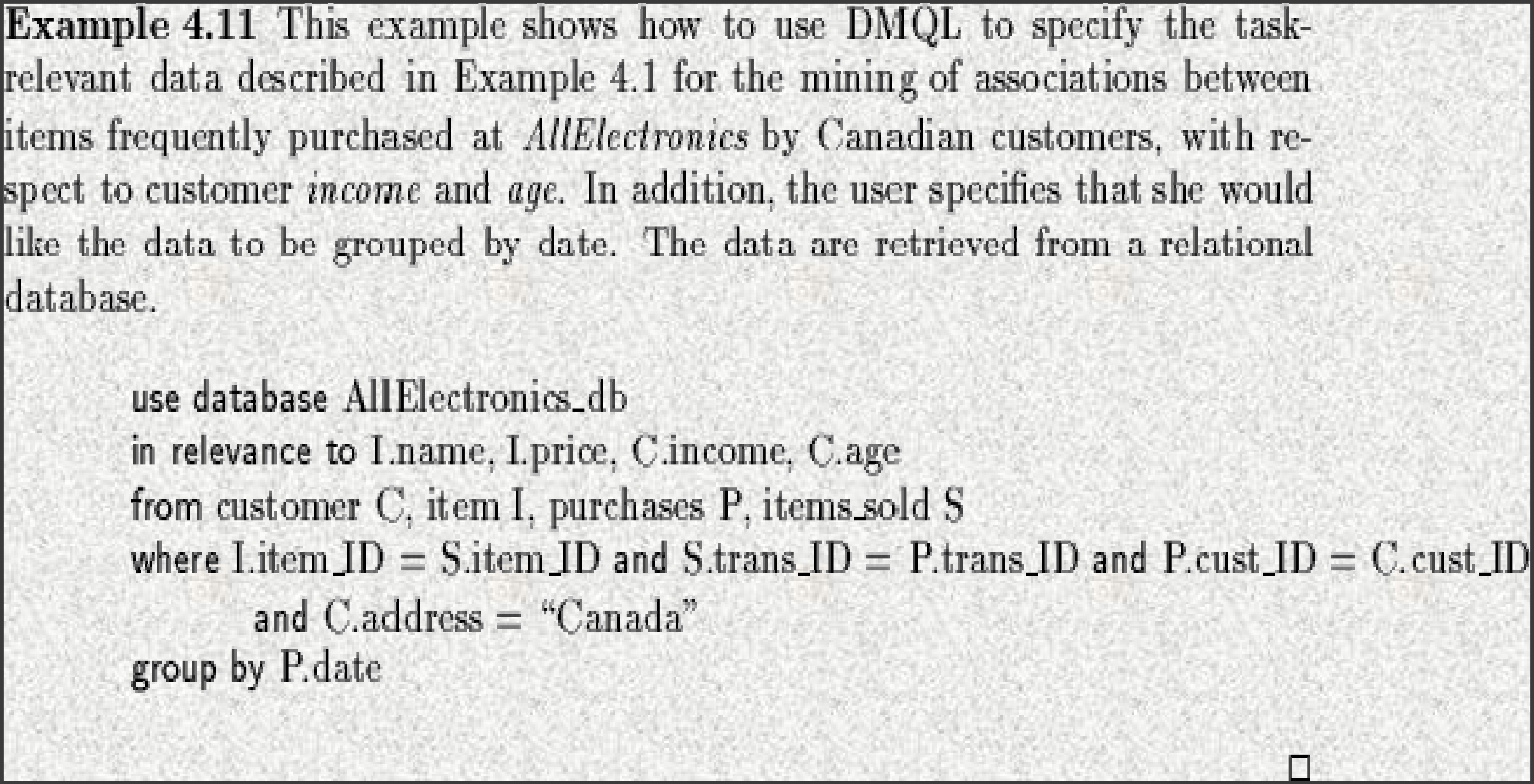
>-Hence, the design of a language requires deep understanding of the limitations and underlying mechanism of the various kinds of tasks.

>-So...how would you design an efficient query language???

>-Based on the primitives discussed earlier.

>-DMQL allows mining of different kinds of knowledge from relational databases and data warehouses at multiple levels of abstraction

EXAMPLE



Example 4.11 This example shows how to use DMQL to specify the task-relevant data described in Example 4.1 for the mining of associations between items frequently purchased at *AllElectronics* by Canadian customers, with respect to customer *income* and *age*. In addition, the user specifies that she would like the data to be grouped by date. The data are retrieved from a relational database.

```
use database AllElectronics_db
in relevance to I.name, I.price, C.income, C.age
from customer C, item I, purchases P, items_sold S
where I.item_ID = S.item_ID and S.trans_ID = P.trans_ID and P.cust_ID = C.cust_ID
        and C.address = "Canada"
group by P.date
```

□

Data Mining

Introduction: Major Issues 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

>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