Data Mining
Unit 1

Chapter 3. Data
Data Mining - A Knowledge Discovery Approach
Krzysztof J. Cios

Topics

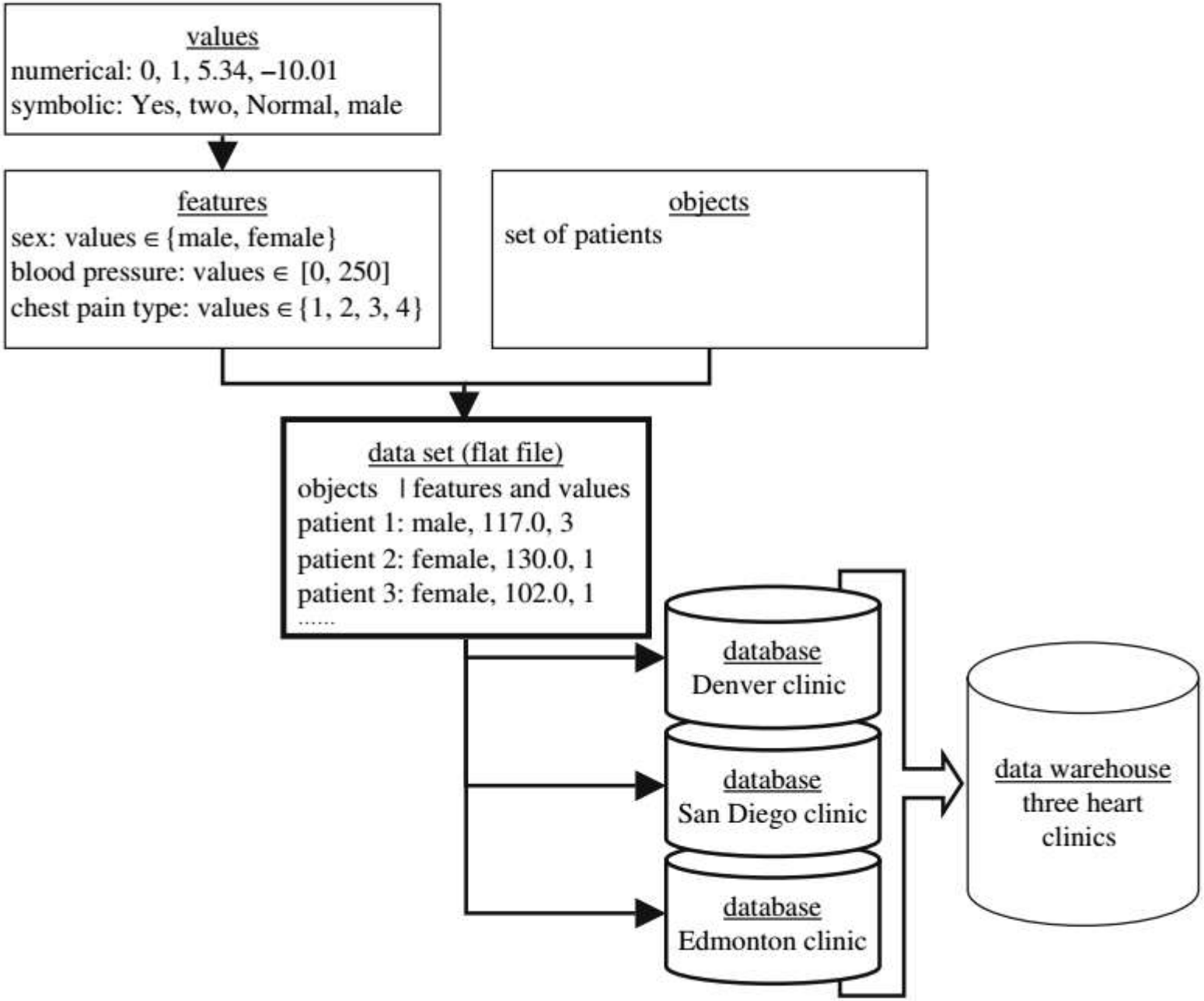
1. Introduction
2. Attributes, Data Sets, and Data Storage
3. Issues Concerning the Amount and Quality of Data

Introduction

- This Chapter focuses on three issues:
  - data types,
  - data storage techniques, and
  - amount and quality of the data.

2. Attributes, Data Sets, and Data Storage

- **Data**
  - Formats : diverse
  - Storage: different models
- A single **unit of information**
  - is a value of a feature/attribute
  - Each feature can take a number of different values.
- **Objects**
  - described by features
  - are combined to form data sets
  - Which are stored as flat (rectangular) files and in other formats
  - Using databases and data warehouses.



Relationships between values, features, objects, data sets, databases and data warehouses.

2.1. Values, Features, and Objects

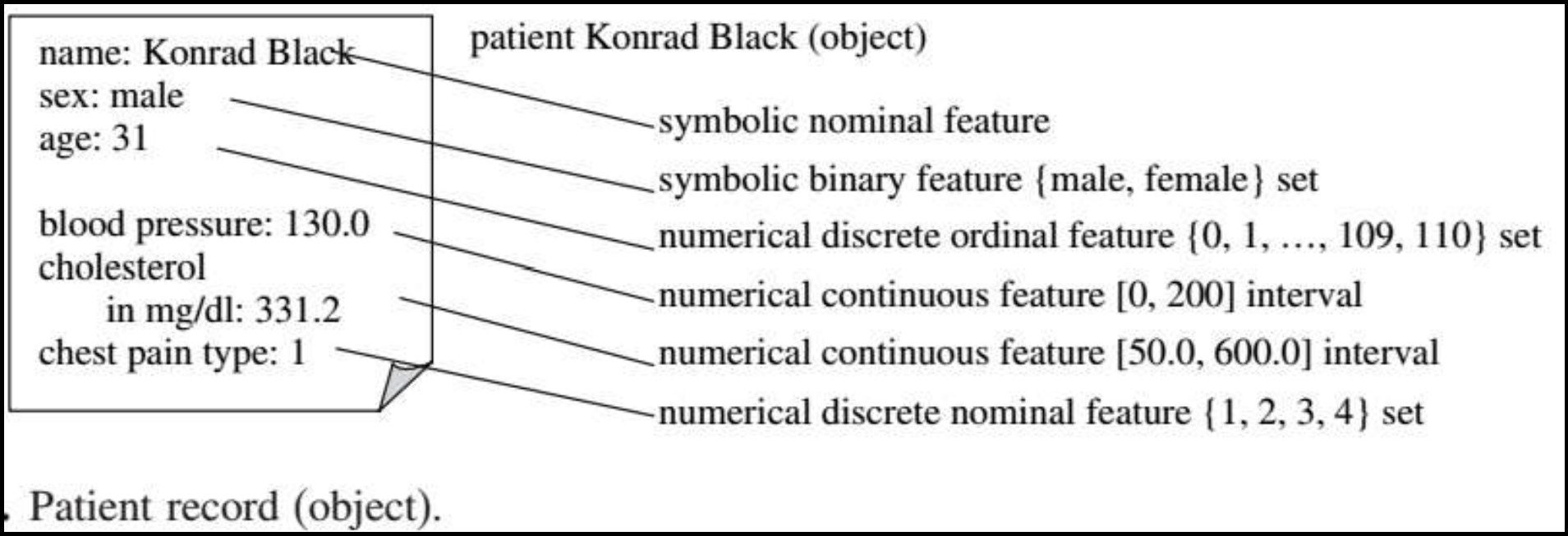
- **Values** - Types
  - numerical
    - expressed by numbers,
    - Ex, real numbers (–1.09, 123.5),
    - integers (1, 44, 125),
    - prime numbers (1, 3, 5),
  - Symbolic
    - describe qualitative concepts
    - such as colors (white, red) or sizes (small, medium, big)

- **Features**
  - also known as attributes
  - usually described by a set of corresponding values.
  - ex, height is usually expressed as a set of real numbers.
  - Features described by both numerical and symbolic values
  - can be either discrete (categorical) or continuous

- **Discrete Features**
  - Total values small in number
  - binary (dichotomous) feature
  - nominal (polytomous) feature
  - ordinal feature
  - **Discretization :** a necessary preprocessing step to transform continuous features into discrete ones

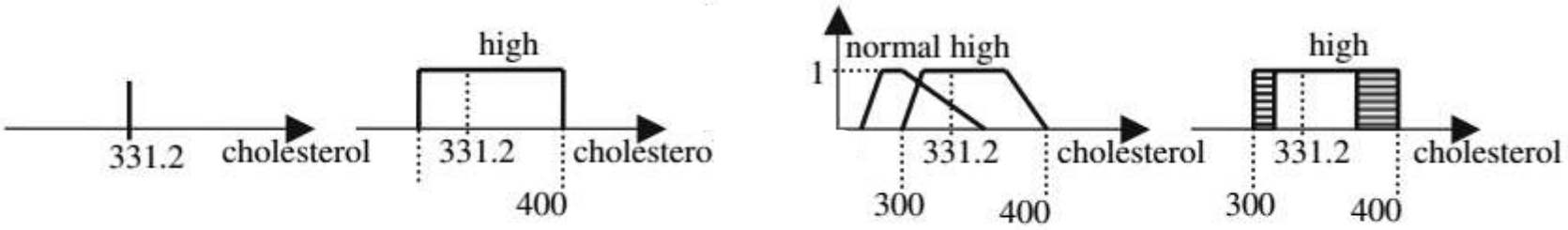
- **Continuous features**
  - total number of values is very large (infinite)
  - and covers a specific interval (range).

- **Objects**
  - also known as records,
  - Ex: units, cases, individuals, data points
  - Represent entities described by one or more features.
  - **multivariate data** refers to situation in which an object is described by many features
  - **univariate data** a single feature describes an object



. Patient record (object).

- **information granulation**
  - means encapsulation of numeric values into single conceptual entities
  - information is often “granulated” and represented at a higher level of abstraction
  - Ex. For a cholesterol value of 331.2, its meaning can easily be understood when this numerical value is expressed in terms of aggregated information such as a “high” or “low” level of cholesterol.



(a) *331.2* value of cholesterol

(b) *high* value of cholesterol

(c) *high* with degree of 1 and *normal* with degree of 0.5 value of cholesterol

(d) *positively high* value of cholesterol

Information granularization methods.

(a) Numerical; (b) interval based;
(c) fuzzy set based;
(d) rough set based.

Data Sets

- Objects described by the same features are grouped to form **data sets.**
- Frequently assumed that –
  - data sets are organized as flat files,
  - in a rectangularly formatted table composed of rows and columns.
  - The rows represent objects and
  - the columns represent features,
  - resulting in a flat file that forms a two-dimensional array.
- Flat files are used to store data in a simple text file format, and they are often generated from data stored in other, more complex formats, such as spreadsheets or databases.

Name	Age	Sex	Blood pressure	Blood pressure test date	Cholesterol in mg/dl	Cholesterol test date	Chest pain type	Defect type	Diagnosis
Konrad Black	31	male	130.0	05/05/2005	NULL	NULL	NULL	NULL	NULL
Konrad Black	31	male	130.0	05/05/2005	331.2	05/21/2005	1	normal	absent
Magda Doe	26	female	115.0	01/03/2002	NULL	NULL	4	fixed	present
Magda Doe	26	female	115.0	01/03/2002	407.5	06/22/2005	NULL	NULL	NULL
Anna White	56	female	120.0	12/30/1999	45.0	12/30/1999	2	normal	absent
.....	.....	.....	.....	.....	.....	.....	.....	.....	.....

Data Set Example:

- A data set of heart patients, shown in Table above.
- Each patient (object) is described by the following set of features:
  - name (symbolic nominal feature)
  - age (numerical discrete ordinal feature from the 0,1,…. 109, 110 set)
  - sex (symbolic binary feature from the {male, female} set)
  - blood pressure (numerical continuous feature from the [0, 200] interval)
  - blood pressure test date (date type feature)
  - cholesterol in mg/dl (numerical continuous feature from the [50.0, 600.0] interval)
  - cholesterol test date (date type feature)
  - chest pain type (numerical discrete nominal feature from the {1, 2, 3, 4} set)
  - defect type (symbolic nominal feature from the {normal, fixed, reversible} set)
  - diagnosis (symbolic binary feature from the {present, absent} set)

DATA STORAGE: DATABASES AND DATA WAREHOUSE

- **Flat files -** files containing plain text are very popular to store data.
- Data mining tools can be applied on other formats(such as tables, graphs, lists etc.) of data for knowledge discovery
- Other formats of data can be found in
  - Databases
  - Data Warehouses
  - Advanced Database Systems
    - Object-oriented
    - Object-Relational
    - Data-specific
      - Transactional
      - Spatial
      - Temporal
      - Text
      - Multimedia

Specialized Database Mgmt Systems

- Four main reasons we use these specialized system.
  - **Memory of computer might not be enough** to store large data sets , so specialized systems are used to fetch data
  - Sometimes required **data might be present at different subsets of data**, so specialized systems are used to retrieve required piece of data
  - Specialized systems are used to add and update data dynamically by different people in various locations and it offers **failure recovery options**
  - Flat files may contain redundant data, which can be avoided if single data set is stored in multiple tables so that **retrieval of data would be easier for user.**

DATABASE

- **Database Management systems (DBMS)** consists of a database that stores data and set of programs for management and fast access of data.

- **Services provided by DBMS**
  - Ability to define structure (schema)
  - Store the data
  - Access the data in concurrent ways
  - Ensure the security and consistency of data

- **Most common database type**
  - relational database which consist of set of tables.
  - Each table is assigned a unique name
  - Table consist of attributes(columns) and tuples(rows)
  - Each tuple in a table is assigned a special attribute called a key, this key is unique for each row or record and it is used to relate the tuple between the tables of the database.
- **Relational database also includes Entity-Relationship model which defines a set of entities (Tables, Attributes, Rows or Tuples etc.) and their relationships.**

patient

Patient ID	Name	Age	Sex	Chest pain type	Defect type	Diagnosis
P1	Konrad Black	31	male	1	normal	absent
P2	Magda Doe	26	female	4	fixed	present
P3	Anna White	56	female	2	normal	absent
...	...	...	...	...	...	...

blood_pressure_test

Blood pressure test ID	Patient ID	Blood pressure	Blood pressure test date
BPT1	P1	130.0	05/05/2005
BPT2	P2	115.0	01/03/2002
BPT3	P3	120.0	12/30/1999
...	...	...	...

cholesterol_test

Cholesterol test ID	Patient ID	Cholesterol in mg/dl	Cholesterol test date
SCT1	P1	331.2	05/21/2005
SCT2	P2	407.5	06/22/2005
SCT3	P3	45.0	12/30/1999
...	...	...	...

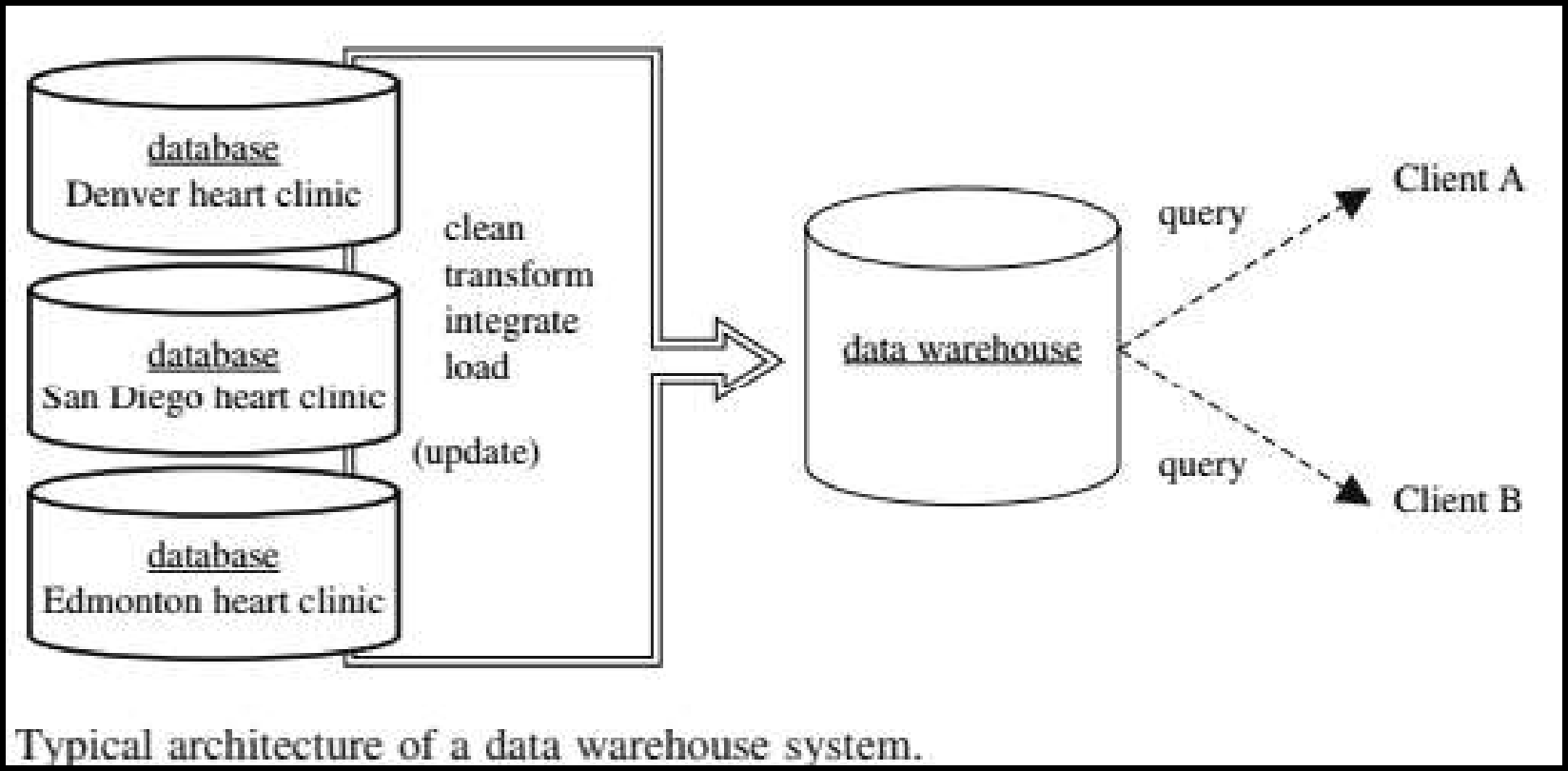
performed_tests

Patient ID	Blood pressure test	Cholesterol test
P1	BPT1	SCT1
P2	BPT2	SCT2
P3	BPT3	SCT3
...	...	...

Figure 3.4. Relational database for heart clinic patients

Data warehouse

- Data warehouse
  - is a repository of data collected in different locations (relational databases)
  - stored using a unified schema
- Data warehouses are usually created by applying a set of processing steps to data coming from multiple databases
- Steps - data cleaning, data transformation, data integration, data loading, and periodical data update



Typical architecture of a data warehouse system.

- **main purpose of database** - storage of the data
- **the main purpose of a data warehouse** – analysis
- **Subject Oriented** - the data in a data warehouse are organized around a set of subjects of interest to the user.
- Ex, in case of the heart clinic, these subjects could be patients, clinical test types, and diagnoses.
- **The analysis** is performed to provide information knowledge) from a historical perspective.
- For instance, we could ask for a breakdown of the most performed clinical tests in the past five years. Such requests (queries) require the availability of summarized information, and therefore data warehouses do not store the same content as the source databases but rather a summary of these data.

- A data warehouse usually uses a **multidimensional database structure**
- Each cell in the database corresponds to some **summarized (aggregated)** measure, such as average, count, minimum, etc.
- The actual implementation of the warehouse can be a **relational database or a multidimensional data cube**.
- It provides a **three-dimensional view** of the data and allows for fast access to the summarized data via precomputation.
- **OLAP** - Multidimensional data allows OLAP
- Two commonly used OLAP operations are **roll-up and drill-down**.
- The first operation merges data at one or more dimensions to provide the user with a higher–level summarization,
- while the latter breaks down data into subr-anges to present the user with more detailed information

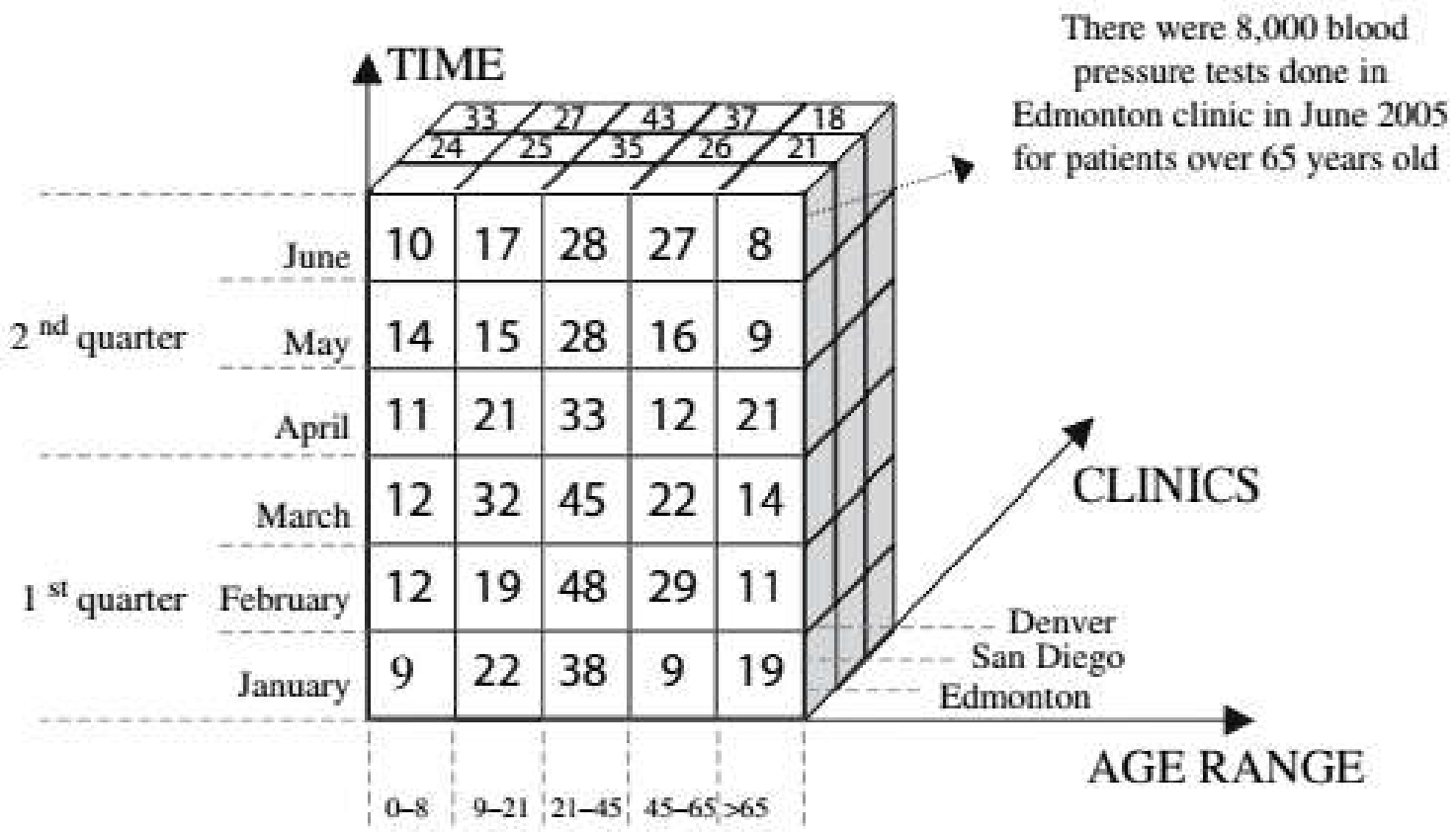


Figure 3.6. A multidimensional data cube. The values are in thousands and show the number of performed blood pressure tests. For readability, only some of the cube cells are shown.

Advanced Data Storage

- Are more specialized and advanced database systems
- The new databases handle :
  - transactional data;
  - spatial data such as maps;
  - Hypertext such as HTML and XML;
  - multimedia data such as combinations of text, image, video and audio;
  - temporal data, such as time–related series concerning stock exchange and historical records;
  - and the WWW, which is enormously large and distributed

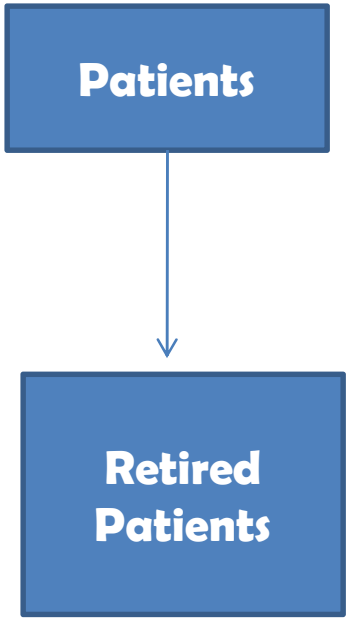
- The special data types require equally specialized databases that utilize efficient data structures and methods for handling operations on such complex data structures.
- challenges for databases are to
  - cope with variable-length objects,
  - structured and semi-structured data, unstructured text in various languages,
  - multimedia data formats,
  - and very large amounts of data.
- Due to these challenges, numerous specialized databases have been developed.
- These include object-oriented and object-relational databases, spatial databases, temporal databases, text databases, multimedia databases, and the WWW

Object-Oriented Databases

- These databases treats each stored entity as an object.
- The object encapsulates three entities:
  1. A set of variables.
  2. A set of messages.
  3. A set of methods.
- It groups similar objects into classes.
- These classes can be organized into hierarchies.
- This helps in better information sharing.

Example:

Super class



Derived class

Variables:
- -Name
- -Age
- -Sex

Variables:
- -Name
- -Age
- -Sex
- -Date of release

Object-Relational Databases

- These databases are based on the **object-relational data model**.
- They are extended by providing a set of complex data types to handle complex objects.
- This extension requires availability of a specialized query language to retrieve the complex data from the database.
- It requires Specialized Query Language(SQL).
- Extensions include ability to handle data types like:
    1. Trees.
    2. Graphs.
    3. Lists.
    4. Class Hierarchies.
    5. Inheritance.

Transactional Databases

- Transactional databases are stored as flat files and consist of records that represent transactions.
- A transaction includes
  - a unique identifier and
  - a set of items that make up the transaction.

sales

Transation ID	Set of item IDs
TR000001	Item1, Item32, Item52, Item71
TR000002	Item2, Item3, Item4, Item57, Item 92, Item93
TR000003	Item11, Item101
...	...

Example transactional database

- **Difference between the relational and the transactional databases**
- **Transactional database store a set of items (values)**
- **It stores information about the presence/absence of an item**
- **A relational database stores information about specific feature values that an example (item) possesses.**
- **It stores a set of values of the related features.**

Spatial Databases

- Spatial databases are designed to handle spatially related data.
- Example data includes geographical maps, satellite images, medical images, and VLSI chip-design data.
- Spatial data can be represented in two ways: in a raster format and in a vector format.
- The raster format concerns using n–dimensional pixel maps, while the vector format requires representing all considered objects as simple geometrical objects, such as lines, triangles, polygons, etc., and using vector based geometry to compute relations between the objects.

Temporal Databases

- Temporal databases (also called time-series databases) are used to store time-related data.
- Similarly, as in the case of object-relational databases, the temporal databases extend the relational databases to handle time-related features.
- Such attributes may be defined using timestamps of different semantics such as days and months, hours and minutes, days of the week, etc.
- The database keeps time-related features by storing sequences of their values that change with time.
- In contrast, a relational database usually stores the most recent value only.

Text databases

- Text databases include features (attributes) that contain word descriptions for objects.
- These features are not simple nominal but hold long sentences or paragraphs of text.
- Examples for the heart clinic would be reports from patient interviews, physician's notes and recommendations, descriptions of how a given drug works for a given patient, etc.
- The text may be either unstructured, like sentences in plain written language (English, Polish, Spanish, etc.);
- Semistructured, where some words or parts of the sentence are annotated, such as descriptions of how a drug works, which may use special annotations for the drug's name and the dose; and
- structured, where all the words are annotated, like a physician's recommendation that may be a fixed form listing only specific drugs and doses.

Multimedia Databases

- Multimedia databases allow **storage, retrieval, and manipulation of image, video, and audio data.**
- The main concern regarding such data sources is their very large size, and therefore, specialized storage and search techniques are required.
- Additionally, both video and audio data are recorded in real time, and thus the database must include mechanisms that assure a steady and predefined rate of acquisition to avoid gaps, system buffer overflows, etc.
- An example application of a multimedia database for the heart clinic would be to find the relation between a video of heart motion and a recording of the heart beats.

World Wide Web

- The World Wide Web (WWW) is an enormous distributed repository of data that is linked together via the use of hyperlinks.
- The hyperlinks link together individual data objects of possibly different types, allowing for interactive access to the information.
- The most specific characteristic of the WWW is that users seek information traversing between objects via links.
- The WWW also provides specialized query engines such as Google, Yahoo!, AltaVista, and Bing

Issues Concerning the Amount and Quality of Data

- Fundamental issues related to data that cause a direct impact on the quality of result which is obtained after the data is being processed are:
  - Huge volume of data(problems related to scalability).
  - Dynamic nature of data.
  - Data related problems such as incompleteness, redundancy , missing values , etc.

High dimensionality

- This issue is related to **scalability**
- The scalability issue is related to the algorithm design and methods which are required to handle massive amount of data.
- While handling with high dimensionality issues three values are considered .
  1- Number of objects
  2- Number of features
  3 -Number of values a features assumes or possesses.

- A number of techniques are available to improve the scalability of data mining algorithms.
- These can be divided into two groups:

1. Techniques that speed up the algorithm.

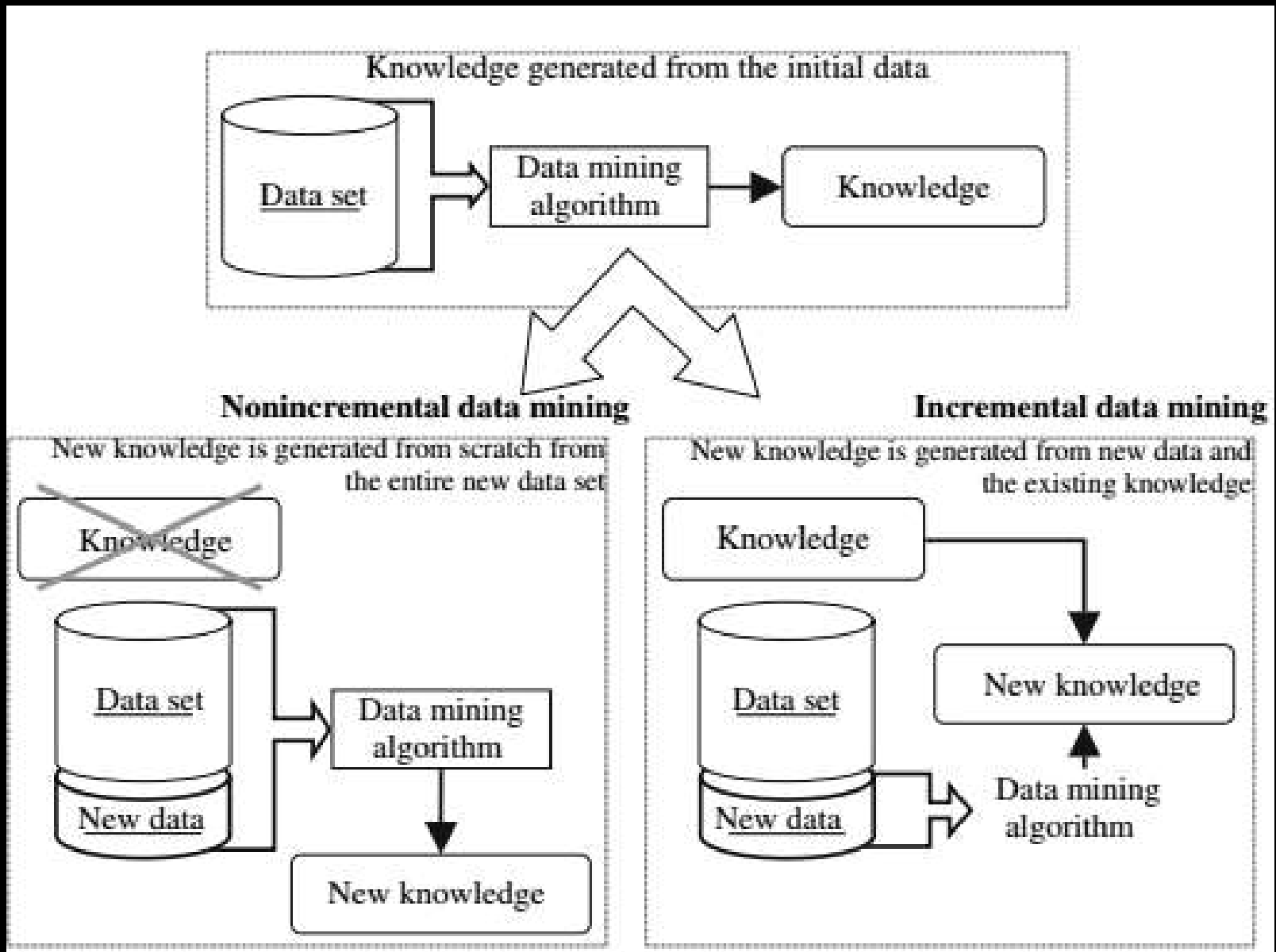
- This outcome can be achieved through use of heuristics, optimization, and parallelization.
- Heuristics simplify the processing of the data.
- Optimization of the algorithm is often achieved by using efficient data structures. parallelization aims to distribute the processing of the data by the algorithm into several processors that work in parallel and thus can speed up the computations.

2. Techniques that partition the data set.

- This outcome can be achieved by reducing the dimensionality of the input data set through reduction of the number of objects, features, number of values per feature, and sequential or parallel processing of data divided into subsets.

Dynamic Data

- Data sets are often dynamic in nature, i.e., new objects and/or features may be added and some objects and/or features may be removed or replaced by new ones.
- In this case, data mining algorithms should also evolve with time, which means that the knowledge derived so far should be also incrementally updated.
- The difference between **incremental and nonincremental DM algorithms** is shown in Figure 3.10.
- The main challenge in incremental data mining methods is merging the newly generated knowledge from new data with the existing, previous knowledge.
- The merger may be as simple as adding the new knowledge to the existing knowledge, but most often it requires modifying the existing knowledge to preserve consistency



. Incremental vs. nonincremental data mining.

Imprecise, Incomplete, and Redundant Data

- Imprecise Data,
- Incomplete Data,
- Redundant Data,
- Missing value,
- Noise

IMPRECISE DATA

- Imprecise data means inaccurate data.
- Real data often include imprecise data objects
- For instance, we may not know the exact value of a given medical test, but instead we know whether the value is high, average, or low.
- In such cases, fuzzy or rough sets can be used to process such imprecise information.

INCOMPLETE DATA

- The term incomplete data refers to the situation in which the available data do not contain enough information to discover new (desired) knowledge.
- Incompleteness may be a result of an insufficient feature, insufficient number of objects, or missing values for a given feature.
- For instance, when analyzing heart patient data, if one wanted to distinguish between sick and healthy patients but only demographic (numeric) information was available, the task could be impossible to complete.
- In dealing with incomplete data, we first need to identify the problem and then take measures to remove it.

REDUNDANT DATA

- The term Redundant Data refers to a problem when there are two or more identical objects.
- Redundant data can increase the load on database/server.
- Redundant data are removed to speed up the processing time.
- For instance, we can expect that a patient's name is irrelevant (inappropriate) with respect to his/her heart condition and thus can be removed.

MISSING VALUES

- Missing Value means there is lack of information or some data are missing.
- Missing Value can result from incomplete data or incorrect data.
- Such values are usually denoted by “NULL”, “ ” and “?” values.
- The methods for dealing with these missing values can be divided into two categories:
  1. Removal of missing data
  2. Imputation (filling in) of missing data.

Removal of missing data

- Features with missing values are simply discarded.
- It is practical only when the data contain small amounts of missing values

Imputation (filling in) of missing data.

- In **mean imputation**, the mean of the values of a feature that contains missing data is used to fill in the missing values.

- In ***hot deck imputation***, for each object that contains missing values, the most similar object is found, and the missing values are imputed from that object. The similarity between objects is usually measured using distance

NOISE

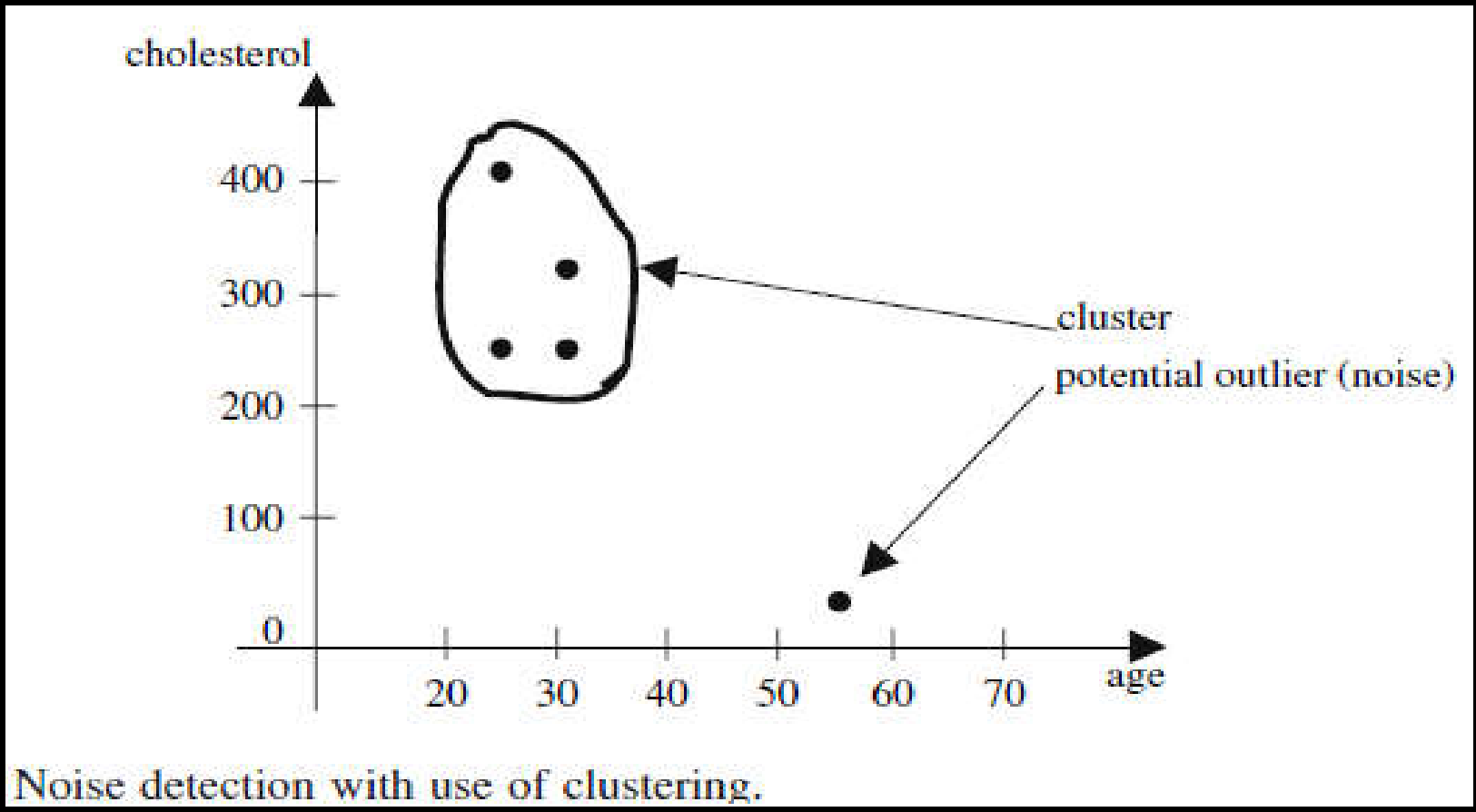
- Noise in the data is defined as a value that is a random error or variance in a measured feature.
- The influence of noise on the data can be prevented by imposing constraints on features in order to detect anomalies when the data are entered.
- When noise is already present, it can be removed by using one of the following methods:
  1. manual inspection with the use of predefined constraints on feature values,
  2. binning, and
  3. clustering.

Manual Inspection & Clustering

- In Manual Inspection, the user checks feature values against predefined constraints and manually removes all values that do not satisfy these constraints.
- Clustering finds groups of similar objects and simply removes all values that fall outside the clusters.

CLUSTERING

- **Noise detection with use of clustering**



Noise detection with use of clustering.