MTIGT1104

SPATIAL DECISION SUPPORT SYSTEM

(SDSS)

Dr. Palanivel K
Assistant Professor
Centre for Remote Sensing
Bharathidasan University
Tiruchirappalli - 620023

MTIGT1104: SPATIAL DECISION SUPPORT SYSTEM - 4 Credits

1. **Introduction to Spatial Decision Support System:** Definition – Concepts – Multicriterian Approach – Usefulness. **7 Hrs.**

2. **Designing of Spatial Database:** Identification of Geographic features – attributes & data layer – Defining the storage parameters for each attribute – ensuring of co-ordinate registration – map projection – Transformation. **12 Hrs.**

3. **Designing of Non spatial Database:** Creation of data table file to hold the attributes – Adding up of description attribute values to table – Different types of sources of data entry – Checking for errors. **12 Hrs.**

4. **Linking of Spatial Database & with Non spatial Database:** Verifying of common item, availability and joining of attribute – table with existing spatial records – spatial display of non spatial data. **15 Hrs.**

5. **Designing & Coding of QUBIS:** Planning for the user requirement – preparation of spatial & Non spatial relational databases – QUBIS Designing – QUBIS Coding (Testing, Error handling, Monitoring, User interface

References:

1. Jeremiah Lindemann, Lisa Markham, Robert Burke, Janis Davis, Thad Tilton, Introduction to Programming ArcObjects with VBA, ESRI , USA. 2004.
2. Kang-Tsung Chang, Programming ArcObjects with VBA, A Task Oriented Approach, CRC Press.
3. ArcObjects Developer's Guide – ArcInfo 8, ESRI INC., California, 1999.
4. Andrew Macdonald, Buiding a Geodatabase – ArcInfo 8, ESRI INC., California, 1999.
5. Michael Zeiler, Modelling Our World – The ESRI Guide to Geodatabase Design, ESRI INC., California, 1999.
6. Kang-tsung chang, Introduction to Geographic Information Systems, McGraw Hill, 2002.

UNIT – I

Introduction to Spatial Decision Support System

Definition – Concepts – Multicriterian Approach – Usefulness.

- 7 Hrs.

Virtues & Trends of GIS

- GIS has got a large number of vistas and advanced virtues
- Apart from many, GIS can perform a lot of sophisticated spatial operations.
- Integrate data from various sources – RS, DP, DIP, GPS, CAD, AM/FM, GPR, Secondary Data from Government / Quasi Govt. / NGO (Records of Geophysical and other Surveys, etc.)
- Provides support for making decisions with a set of procedures in a problem solving environment.
- Simplifies tedious time consuming procedures such as, digital spatial data base generation (which includes plotting / digitization, manipulation, data conversion, labeling, etc.), entry of attribute data and link with spatial data, data classification, spatial analysis, modeling, map display, etc.

CUSTOMIZATION & AUTOMATION IN GIS (Advancement : 1)

- Consist of easy Graphical User Interface (GUI)
- Provides user guidance / support and friendly environment
- Reduces multiple time-taking steps in generating the basic databases used in models, model making, map display, etc.
- Hiding operations (unnecessary for the user)
- Easy to modify within simple steps when needed
- No need of rigorous technical training for the users.
- User can get the final results in any form as per the requirement (table, report, graphs, maps, etc.) within a short period (in few minutes / seconds).

SPATIAL INFORMATION RETRIEVAL SYSTEM (Advancement : 2)

- Customized GIS package
- Provides user guidance / support and friendly environment
- Assists the user to access / retrieve the readily available spatial data as well as the existing action plan maps and details easily and quickly in any mode
- User oriented built-in queries for easy retrieval
- Displays attribute data spatially by creating a link (Relate)
- Query Based Information Retrieval Systems (QUBIS)
- Provides authenticated updation of the existing database, addition of new database into the system and allows editing of the same in an easy and user friendly manner for multiple users in GIS networks.

SPATIAL DECISION SUPPORT SYSTEM (SDSS) (Adv. 3)

- GIS based system, useful for quick / strategic decision making by the planners and administrators
- Enhances the accuracy and higher effectiveness of decision making
- Uses recent / latest data into its pre-devised models and gives chances to opt for user specified methodologies while solving a semi structured spatial problems
- Provides the action plan maps and other outputs in a pragmatic way which can be directly implemented by the decision makers, planners and administrators.
- Assists in further post-implementation aspects such as monitoring, maintenance, etc.

SPATIAL DECISION SUPPORT SYSTEM (SDSS) (Contd…)

USER SIDE INPUTS

- User need to help the system to
  - Locate the data sets
  - Select the model / ready made methodology
  - Give the conditions, if necessary, then

The SDSS does the job of

- Self extrapolating the data sets
- Generating the databases
- Self analyzing the databases,
- Generating action plan maps, tables, reports, etc.
- Provides the results in any format that the user is interested.

WHAT IS SDSS?

SDSS- Spatial Decision Support System

DEFINITION

SDSS is an interactive, computer-based system designed to support user or a group of users in achieving a higher effectiveness of decision making while solving a semi-structured spatial problem.

An effective SDSS requires the addition of a range of specific techniques and functionalities, used especially to manage spatial data, to conventional DSS.

(i) Provide mechanisms for the input of spatial data,

(ii) Allow representation of spatial relations and structures,

(iii) Include the analytical techniques of spatial analysis, and

(iv) Provide output in a variety of spatial forms, including maps.

Motivation

- Is there a demand for spatial decision support for environmental resource management?
- Are we ready for spatial decision support for environmental resource management?

Objective

Combination of latest developments in:

Geographic Information Systems	-	(GIS)
Remote Sensing	-	(RS)
Software Engineering	-	(SE)
Model Integration	-	(MI)

The Spatial Decision-Making Process

According to Gao et al. (2004) nine steps are involved in Spatial Decision-Making Process :

(1) Problem identification

(2) Problem modeling

(3) Model instantiation

(4) Model execution

(5) Model integration / scenario modeling

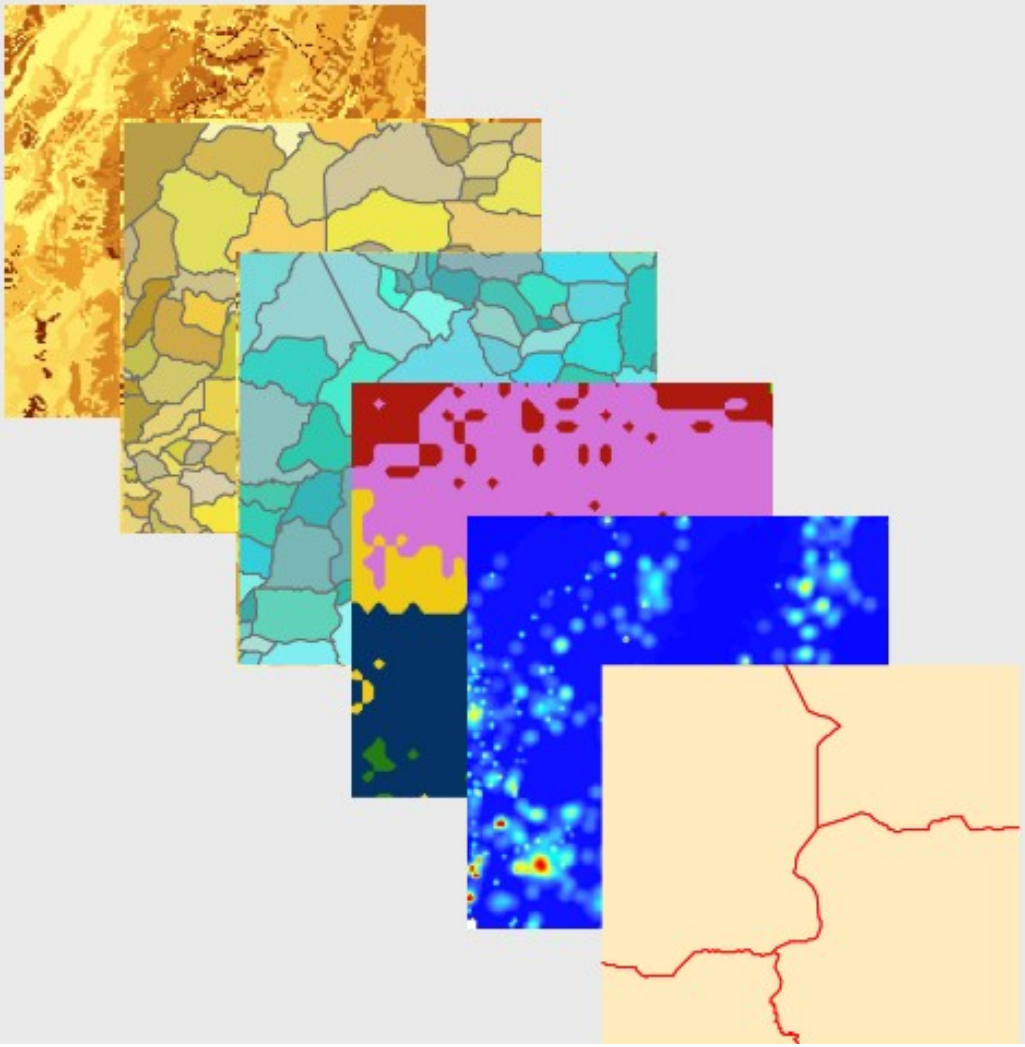
(6) Scenario instantiation

(7) Scenario execution

(8) Scenario evaluation

(9) Decision making.

Required Data



- soil
- land use
- weather
- management
- socio-economic
- administrative
- etc.

Multi-Criteria Spatial Decision Support Systems (MC-SDSS)

- Multicriteria Spatial Decision Support Systems (MC-SDSS) can be viewed as part of the broader fields of SDSS.
- The specificity of MC-SDSS is that it supports spatial multicriteria decision making.
- Spatial multicriteria decision making refers to the use of multicriteria analysis (MCA) to spatial decision problems.
- MCA operations research tools that have experienced very successful applications in different domains since the 1960. It has been coupled with geographical information systems (GIS) since the early 1990s for an enhanced decision making.

General structure of SDSS/MC-SDSS

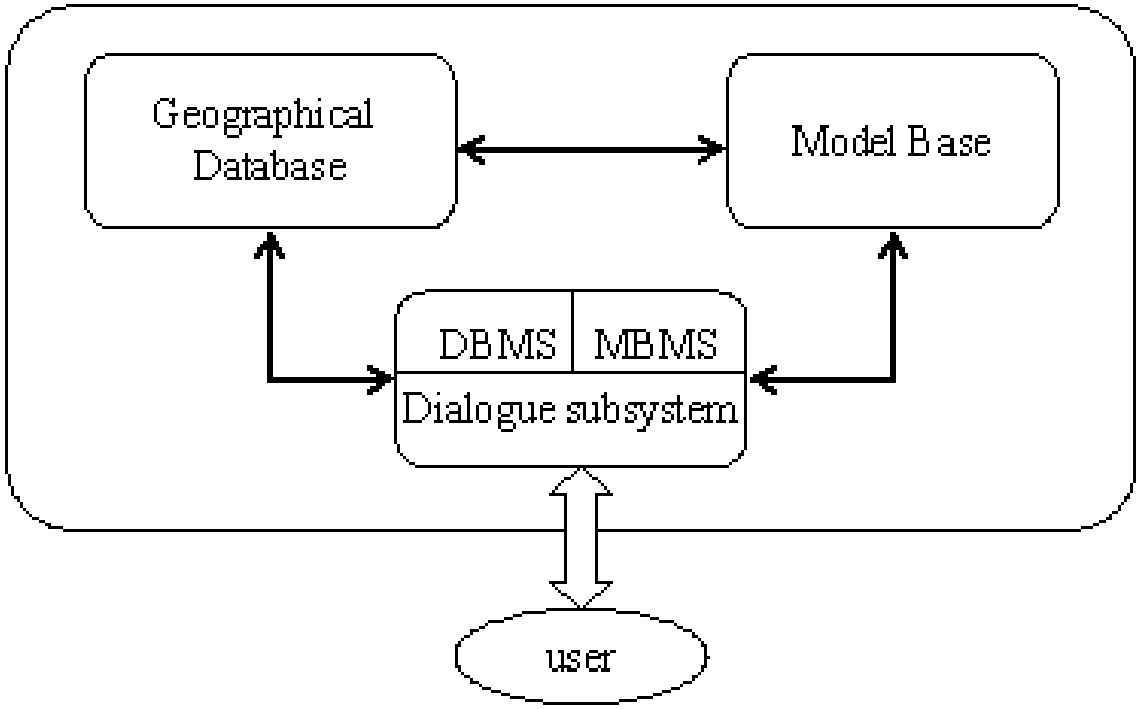
A typical SDSS contains three generic components :

- **A database management system and geographical database,**
- **A model-based management system and model base, and**
- **A dialogue generation system.**

- The **data management subsystem** performs all data-related tasks; that is, it **stores, maintains, and retrieves data from the database, extracts data from various sources, and so on.** it provides access to data as well as all of the control programs necessary to get those data in the form appropriate for a particular decision making problem.

- The **model base management system** component **provides links between different models so that the output of one model can be the input into another model.**

- These three components constitute the software portion of the an SDSS. A fourth important component of any decision support system is the user which may be simple **users, technical specialists, decision makers, and so on.**



General structure of SDSS

- MC-SDSS can be viewed as a part of a broader field of SDSS.
- Accordingly, the general structure of a MC-SDSS is the same that the one of a SDSS.
- However, the model-based management system is enhanced to support multicriteria spatial modelling and the model base is enriched with different multicriteria analysis techniques.

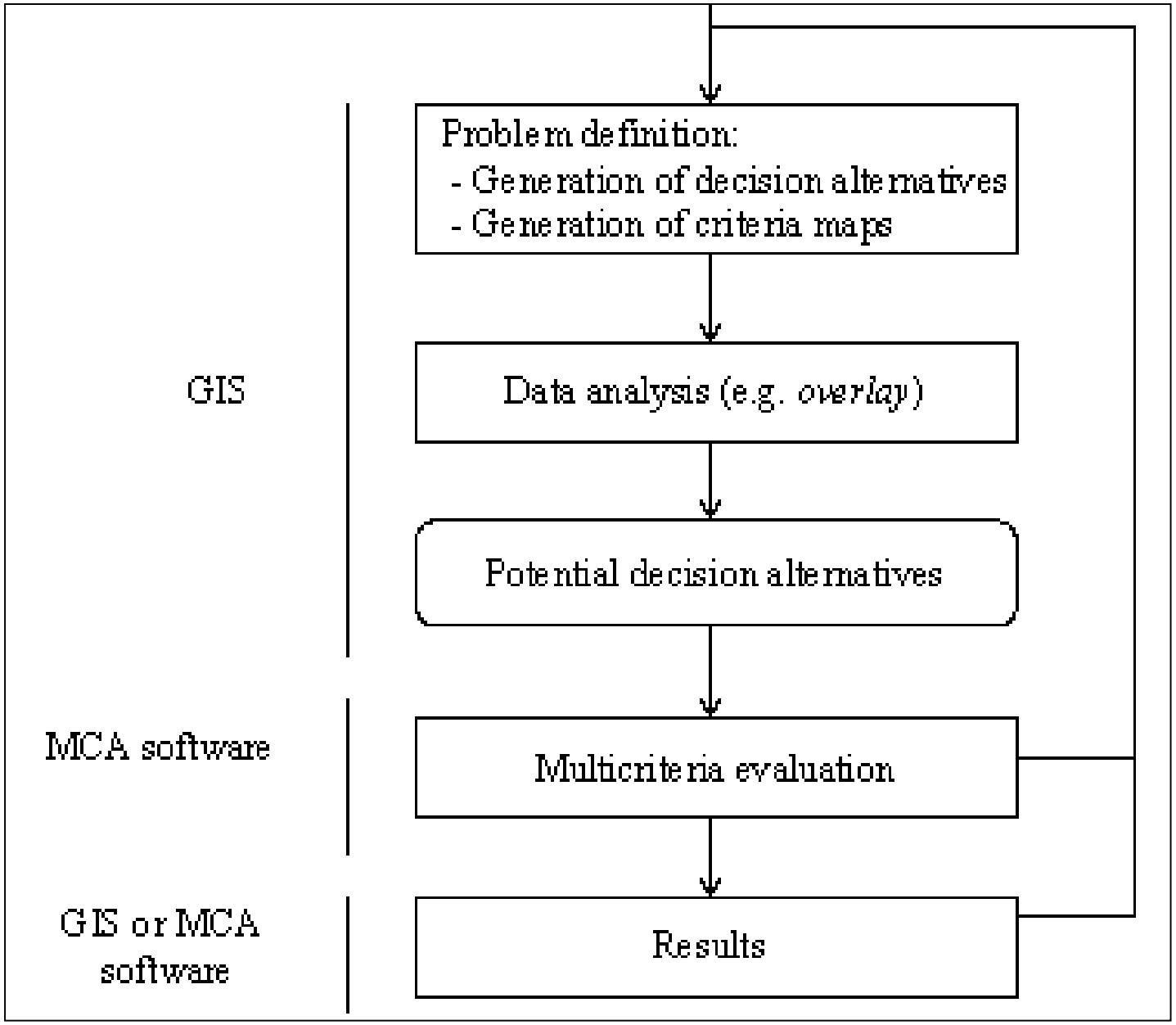
GIS and multicriteria analysis integration modes

- The conceptual idea on which most of GIS-based multicriteria analysis rely is to use the GIS capabilities to prepare an adequate platform for using multicriteria methods.

- **The GIS-based multicriteria analysis starts with the problem identification, where the capabilities of the GIS are used to define the set of feasible alternatives and the set of criteria.**

- Then, the overlay procedures are used in order to reduce an initially rich set of alternatives into a small number of alternatives which are easily evaluated by using a multicriteria method.

- Finally, the drawing and presenting capabilities of the GIS are used to present results.



Conceptual schema for GIS and multicriteria analysis integration

- Physically, there are four possible modes to integrate GIS and multicriteria analysis tools:

  (i) **No integration,**

  (ii) Loose integration,

  (iii) Tight integration, and

  (iv) Full integration.

- The first mode corresponds to the situation dominating until late 1980 where the GIS and multicriteria analysis are used independently without any integration to deal with **spatial problems**.

- The next three modes correspond to increasing levels of **complexity and efficiency**.

Loose integration mode:

- The integration of GIS software and a **stand-alone multicriteria analysis software is made possible by the use of an intermediate system**.

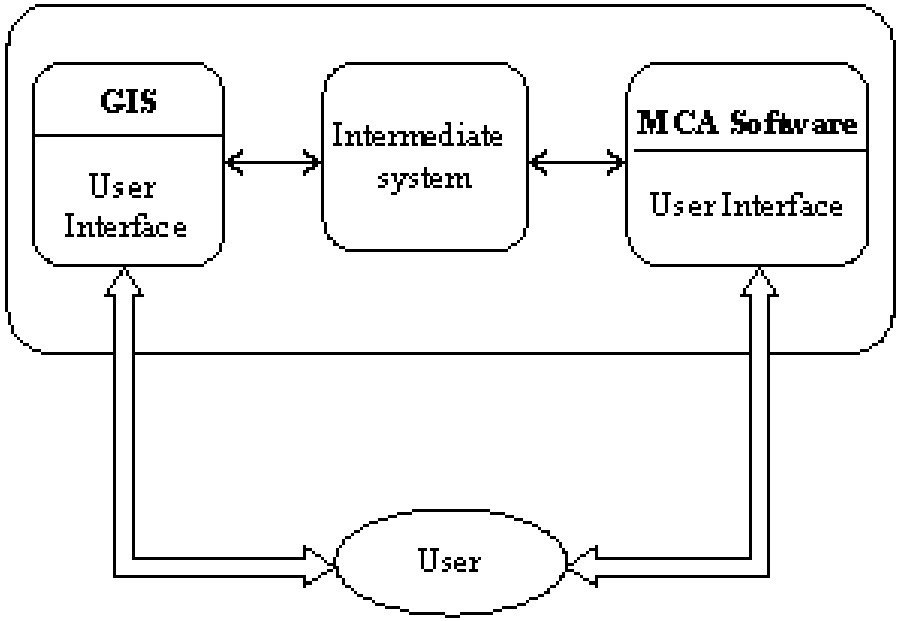
- The intermediate system permits to reformulate and restructure the data obtained from the overlapping analysis which is performed through the GIS into a form that is convenient to the multicriteria analysis software.

Tight integration mode .

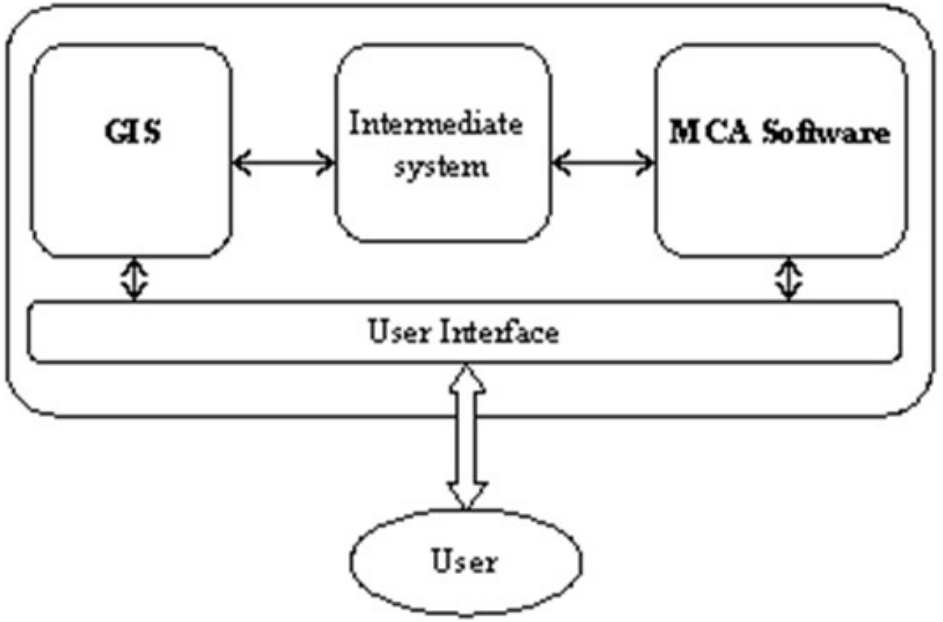
- In this mode, a particular multicriteria analysis method is **directly added to the GIS software.**

- The multicriteria analysis method constitutes an integrated but autonomous part with its own database. The use of the interface of the GIS part alone increases the interactivity of the system.

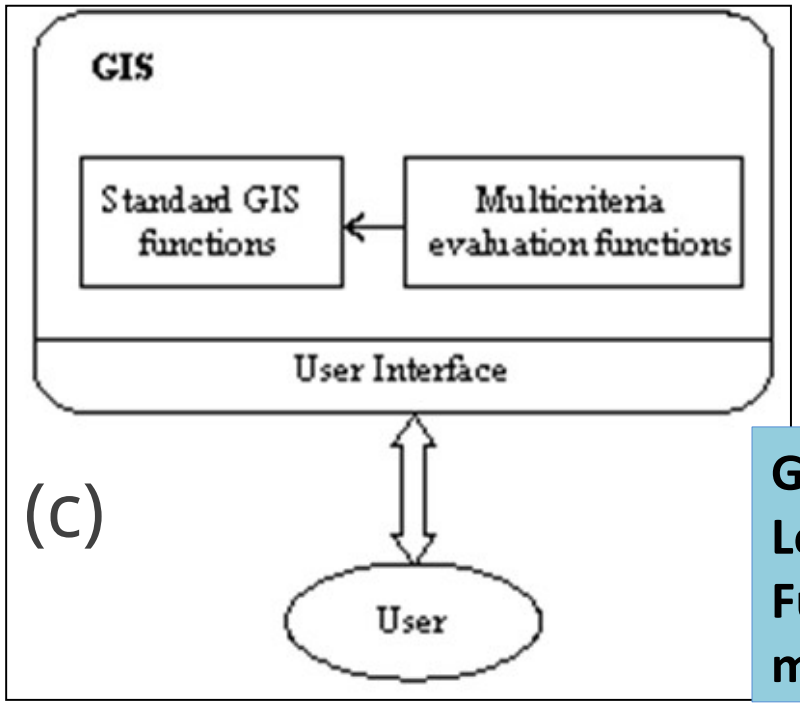
- This mode is the first step towards a complete GIS-multicriteria analysis integrated system.



(a)



(b)



(C)

GIS and multicriteria Loose (a), Tight (b) and Full (c) integration modes

Full integration mode .

- The third mode yields itself to a complete GIS-multicriteria analysis integrated system that has a **unique interface and a unique database.**

- The multicriteria analysis method is activated directly from the **GIS interface as any GIS basic function**.

- The GIS database is extended so as to support both the **geographical and descriptive data**, on the one hand, and the **parameters required for the multicriteria evaluation techniques**, on the other hand.

- **The common graphical interface enhances the user-friendless of global system**

GIS and multicriteria analysis interaction directions

The five directions of interaction :

(i) No interaction,

(ii) One-direction interaction with the GIS as the main software

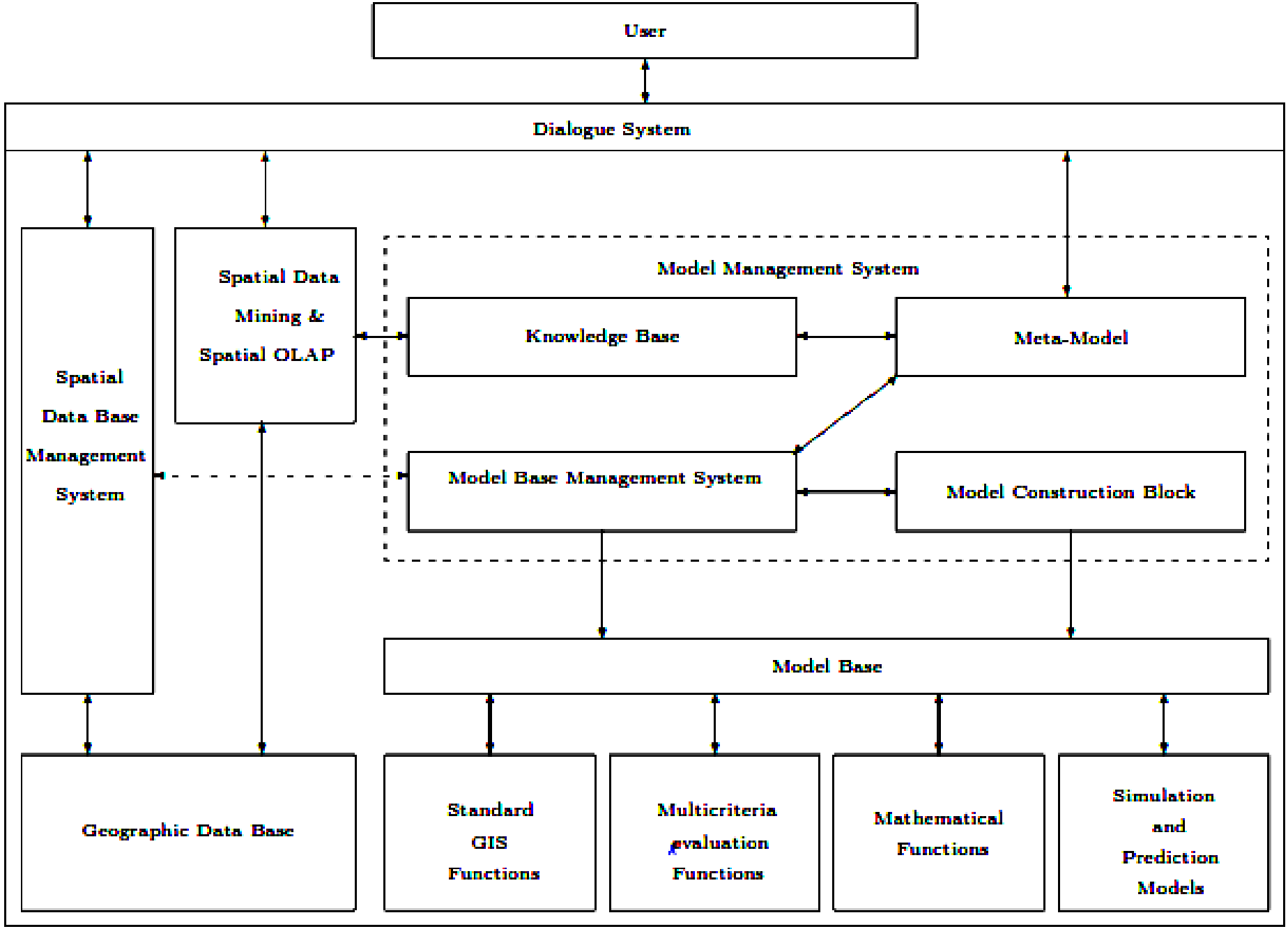
(iii) One-direction interaction with multi-criteria tool as the main software,

(iv) Bi-directional interaction, and

(v) Dynamic interaction.

- One-direction interaction provides a mechanism for importing/exporting information via a single flow that originates either in the GIS or multicriteria software.
- This type of interaction can be based on GIS or multicriteria as the principle software.
- In the bi-directional interaction approach the flow of data/information can originate and end in the GIS and multicriteria modules.
- Dynamic integration allows for a flexible moving of information back and forth between the GIS and multicriteria modules according to the user's needs.

Design of a MC-SDSS



1. Spatial data base Management System
2. Geographic Database
3. Model Base
4. Model Management System
5. Meta-model
6. Knowledge base
7. Model base Management System
8. Model Construction Block
9. Spatial data Mining and Spatial Online Analytical Processing
10. Dialogue System

1.Spatial data base Management System

The spatial data base management system is an extension of the conventional database base management system. It is used specially to manage spatial data.

2.Geographic database

The geographic data base is an extended GIS database. It constitutes the repository for both

(i) The spatial and descriptive data, and

(ii) the parameters required for the different OR/MS tools.

3. Model base

The model base is the repository of different analytical models and functions. Among these functions, there are surely the basic GIS ones *(e.g. statistical analysis, overlaying, spatial interaction analysis, network analysis, etc.)*.

4 Model management system

The role of this component is to manage the different analysis models and functions. The model management system contains four elements:

1. **The meta-model**
2. **The model base management system**
3. **The model construction block and**
4. **The knowledge base.**

5 Meta-model

This element is normally an Expert System used by the decision maker to explore the model base. This exploration enables the decision maker to perform a "what-if" analysis and/or to apply different analytical functions. The meta-model uses a base of rules and a base of facts incorporated into the knowledge base.

6 Knowledge base

Knowledge base is the repository for different pieces of knowledge used by the meta-model to explore the model base. Practically, the knowledge base is divided into a **base of facts** and a **base of rules**.

The base of facts contains the facts generated from the model base. It also contains other information concerning the uses of different models, the number and the problems to which each model is applied, etc. The base of rules contains different rules of decision which are obtained from different experts, or automatically derived, by the system, from past experiences. This base may, for instance, contains: If the problem under study is the concern of many parties having different objective functions then the more appropriate tool is that of MCA.

7 Model base management system

The role of the model base management system is to manage, execute and integrate different models that have been previously selected by the decision maker through the use of the Meta-Model.

8 Model Construction block

- This component gives the user the possibility to develop different ad hoc analysis models for some specific problems.
- The developed ad hoc model is directly added to the model base and its characteristics are introduced into the base of rules of the KB.

9 Spatial data Mining and Spatial on line analytical processing

- Data mining and **On Line Analytical Processing (OLAP)** have been used successfully to extract relevant knowledge from huge traditional databases.
- Recently, several authors have been interested in the extension of these tools in order to deal with huge and complex spatial databases.
- The spatial OLAP technology uses multidimensional views of aggregated, pre-packaged and structured spatial data to give quick access to information.
- Incorporating spatial data mining and spatial OLAP into the MC-SDSS will undoubtedly ameliorate the quality of data and, consequently, add value to the decision-making process.

10 Dialogue system

The dialogue system represents the interface and the equipments used to achieve the dialogue between the user and the MC-SDSS.

It permits the decision maker to enter his/her queries and to retrieve the results.

SDSS Usefulness or Importance

- The SDSS have evolved greatly over the last few decades based on advances in underlying technologies such as computer hardware and software, networking and communication technologies.

- After early development in the 1970s and 1980s, the concept of SDSS gained traction in the 1990s.

- The development of SDSS became much more common in the late 1990s when grater amounts of digital spatial data were becoming available and personal computers were becoming widely used.

- The growth has continued into 2000a with diversification based on technological developments.

- The development of SDSS generally followed developments in Geographic Information Systems(GIS), with many concepts and techniques of the science taken from decision support systems research and advances. Due to these advances, practitioners from a wider range of domains began utilizing computers and GIS software.

- The 1990s saw tremendous growth in the applications of SDSS to a variety of problem domains, including urban, transportation, environmental, natural resources, business, agricultural, emergency planning and others.

- Knowledge-based and artificial intelligence techniques were also introduced into SDSS in the 1990s. The great advances in networking technology and the use of the Web led to the increased use of Web-based technologies in SDSS. With improving wireless communication technologies, the ubiquity of GPS-enabled devices and distributed software techniques, SDSS that operate with mobile computers are becoming feasible.

- The combination of all these technologies is leading to an increase in Web-based and Mobile-based software components in SDSS, which provided greater flexibility for use of real-time data as well as the inclusion of non-expert users in participatory systems.

- The rapid growth in SDSS development that began in the 1990s and has lasted until now should be expected to continue with a greater number of Web and Mobile based SDSS being developed for a variety of disciplines.

Spatial Database

Definition

- A spatial database is a database that is optimized to store and query data related to objects in space, including points, lines and polygons.

- While typical databases can understand various numeric and character types of data, additional functionality needs to be added for databases to process spatial data types. These are typically called geometry or feature.

Spatial Databases Background

- Spatial databases provide structures for storage and analysis of spatial data
- Spatial data is comprised of objects in multi-dimensional space
- Storing spatial data in a standard database would require excessive amounts of space

Spatial Databases Background (Cont.)

- Queries to retrieve and analyze spatial data from a standard database would be long and cumbersome leaving a lot of room for error

- Spatial databases provide much more efficient storage, retrieval, and analysis of spatial data

Types of Data Stored in Spatial Databases

- Two-dimensional data examples
  - Geographical
  - Cartesian coordinates (2-D)
  - Networks
  - Direction

Types of Data Stored in Spatial Databases (Cont.)

- Three-dimensional data examples
  - Weather
  - Cartesian coordinates (3-D)
  - Topological
  - Satellite images

Spatial Databases Uses and Users

- Three types of uses
  - Manage spatial data
  - Analyze spatial data
  - High level utilization

Spatial Databases Uses and Users (Cont.)

- A few examples of users
  - Transportation agency tracking projects
  - Insurance risk manager considering location risk profiles
  - Doctor comparing Magnetic Resonance Images (MRIs)
  - Emergency response determining quickest route to victim
  - Mobile phone companies tracking phone usage

Spatial Database Management System

- Spatial Database Management System (SDBMS) provides the capabilities of a traditional database management system (DBMS) while allowing special storage and handling of spatial data.

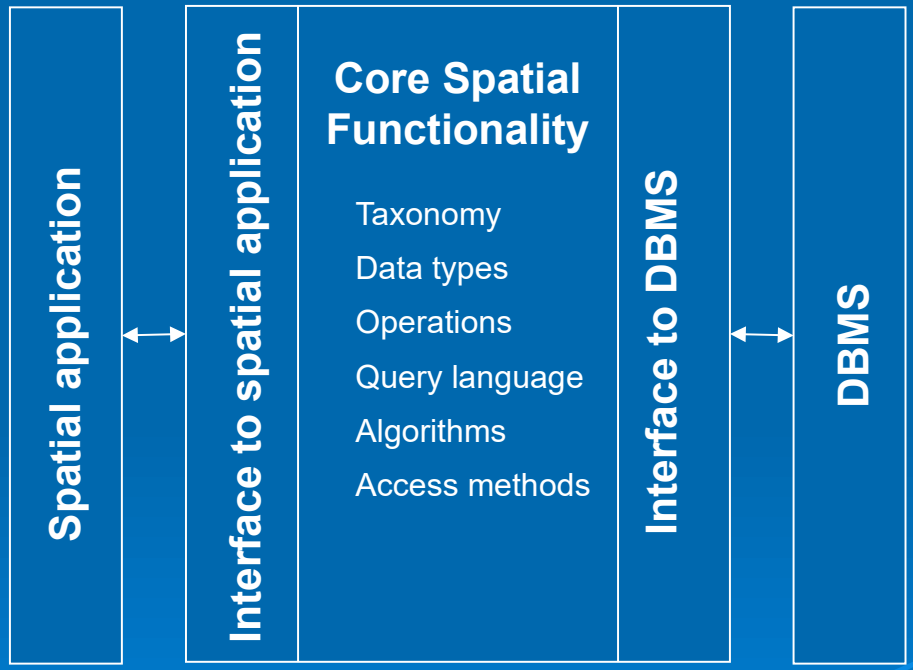
Spatial Database Management System (Cont.)

- SDBMS:
  - Works with an underlying DBMS
  - Allows spatial data models and types
  - Supports querying language specific to spatial data types
  - Provides handling of spatial data and operations

SDBMS Three-layer Structure

- SDBMS works with a spatial application at the front end and a DBMS at the back end
- SDBMS has three layers:
  - Interface to spatial application
  - Core spatial functionality
  - Interface to DBMS

SDBMS Three-layer Structure (Cont.)



Spatial Query Language

- Number of specialized adaptations of SQL
  - Spatial query language
  - Temporal query language (TSQL2)
  - Object query language (OQL)
  - Object oriented structured query language (O_2SQL)

Spatial Query Language (Cont.)

- Spatial query language provides tools and structures specifically for working with spatial data
- SQL3 provides 2D geospatial types and functions

Spatial Query Language Operations

- Three types of queries:
  - Basic operations on all data types (e.g. IsEmpty, Envelope, Boundary)
  - Topological/set operators (e.g. Disjoint, Touch, Contains)
  - Spatial analysis (e.g. Distance, Intersection, SymmDiff)

Spatial Data Entity Creation

- Form an entity to hold county names, states, populations, and geographies

```
)CREATE TABLE County
,varchar(30)Name
,       varchar(30)State
,Integer    Pop
;(PolygonShape
```

Spatial Data Entity Creation (Cont.)

- Form an entity to hold river names, sources, lengths, and geographies

```
)CREATE TABLE River
,varchar(30)Name
,varchar(30)Source
,IntegerDistance
;(LineStringShape
```

Example Spatial Query

- Find all the counties that border on Contra Costa county

C1.NameSELECT

County C1, County C2FROM

Touch(C1.Shape, C2.Shape) = 1WHERE ;'AND C2.Name = 'Contra Costa

Example Spatial Query (Cont.)

- Find all the counties through which the Merced river runs

C.Name, R.NameSELECT

County C, River RFROM

Intersect(C.Shape, R.Shape) = 1WHERE ;'AND R.Name = 'Merced

Features of Spatial Databases

- Database systems use indexes to quickly look up values and the way that most databases index data is not optimal for spatial queries. Instead, spatial databases use a spatial index to speed up database operations.
- In addition to typical SQL queries such as SELECT statements, spatial databases can perform a wide variety of spatial operations.

Features of Spatial Databases (Cont.) - query types

The following query types and many more are supported by the Open Geospatial Consortium:

- Spatial Measurements: Finds the distance between points, polygon area, etc.

Features of Spatial Databases (Cont.) - query types

- **Spatial Functions:** Modify existing features to create new ones, for example by providing a buffer around them, intersecting features, etc.
- **Spatial Predicates:** Allows true/false queries such as 'is there a residence located within a mile of the area we are planning to build the landfill?'

Features of Spatial Databases (Cont.) - query types

- Constructor Functions: Creates new features with an SQL query specifying the vertices (points of nodes) which can make up lines. If the first and last vertex of a line are identical the feature can also be of the type polygon (a closed line).
- Observer Functions: Queries which return specific information about a feature such as the location of the center of a circle

Types of queries - PostGIS

The function names for queries differ across geodatabases. The following list contains commonly used functions built into PostGIS, a free geodatabase which is a PostgreSQL extension (the term 'geometry' refers to a point, line, box or other two or three dimensional shape):

Types of queries - PostGIS (Cont.)

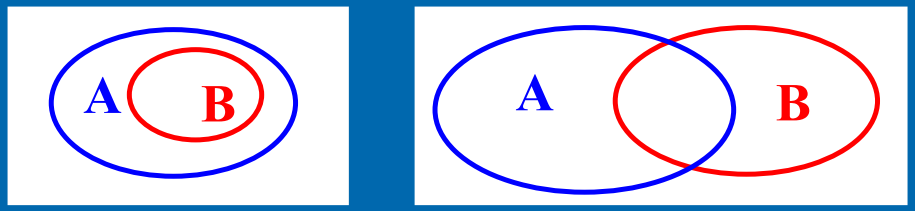
1. Distance(geometry, geometry) : number
2. Equals(geometry, geometry) : boolean
3. Disjoint(geometry, geometry) : boolean
4. Intersects(geometry, geometry) : boolean
5. Touches(geometry, geometry) : boolean
6. Crosses(geometry, geometry) : boolean

Types of queries - PostGIS (Cont.)

7. Overlaps(geometry, geometry) : boolean
8. Contains(geometry, geometry) : boolean
9. Intersects(geometry, geometry) : boolean
10. Length(geometry) : number
11. Area(geometry) : number
12. Centroid(geometry) : geometry

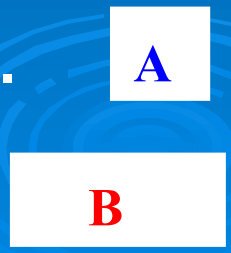
Spatial Relations

- **Topological Relations: containment, overlapping, etc. [Egenhofer et al. 1991]**



- **Metric Relations: distance between objects, etc. [Gold and Roos 1994]**

- **Direction Relations: north of, south of, etc. [Hernandez et al. 1990; Frank et al. 1991]**

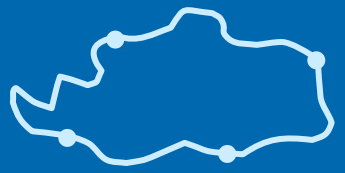


Topological Relations

- **Topological relations are defined using *point-set topology* concepts, such as *boundary* and *interior.***

Topological Relations (Cont.)

- **For example:**
  - **the *boundary* of a region consists of a set of curves that separate the region from the rest of the coordinate space**



  - **The *interior* of a region consists of all points in the region that are not on its boundary**

- **Given this, two regions are said to be *adjacent* if they share part of a boundary but do not share any points in their interior**

Spatial Relations Model

- **An abstract model** (or conceptual model): is a theoretical construct that represents something, with a set of variables and a set of logical and quantitative relationships between them.
- Models in this sense are constructed to enable reasoning within an idealized logical framework about these processes and are an important component of scientific theories.

Geographical Database

Introduction

- **Data** represents the second key component of GIS technology
- The GIS stores and manages the data not as a map but as a series of layers or, as they are sometimes called, themes.
- In GIS, attributes are stored with the geographic data .
- SDBMS is a technology used by GIS application (gegraphic/georeferenced data).

⇒ The database should be viewed as a representation or model of the world developed for a very specific application

- Spatial DB vs. Image/pictorial DB (90s)
  - Spatial DB contains objects **in** the space
  - Image DB contains representations **of** a space (images, pictures...: raster data)

What is a geographical Database?

- Spatial databases provide structures for storage and analysis of spatial data
- Spatial data is composed of objects in multi-dimensional space
- Storing spatial data in a standard database would require **excessive** amounts of space
- Queries to retrieve and analyze spatial data from a standard database would be long.
- Spatial databases provide much more efficient storage, retrieval, and analysis of spatial data

What is a geographical Database? –cont.

- A spatial database is an ORDBMS that has the ability to store, query, manipulate and analyze spatial data as well as traditional data formats
- It offers spatial data types/data models/ query language
  - Structure in space: e.g., POINT, LINE, REGION
  - Relationships among them: (ex:*intersects*)
- It provides spatial indexing (retrieving objects in particular area without scanning the whole space).
- It provides efficient algorithms for spatial joins .
- SDBMS is mainly used for vector format.
- It defines data types for points, lines, polygons, multipoint, multiline, and multipoloygon

What is a geographical Database –cont.

- These databases have built in functions for manipulating spatial data anywhere from 100 to 300 functions
- Most common are functions for querying data such as overlap, intersect, touch, etc.
- Also including are geoprocessing functions such as union, merge, buffer, etc.

Spatial Database Elements

- Entity: a phenomenon of interest in reality that is not further subdivided into phenomena of the same kind eg.city
- Object: a digital representation of all or part of an entity. (city represented by a point or a region)
- Entity types: similar phenomena to be stored in a database are identified as entity types. (road, river…)
- Attribute: an attribute is a characteristic of an entity selected for representation.
- Layers: spatial objects can be grouped into layers, also called overlays, coverage or themes
- Metadata
- Spatial Reference System table

Spatial Database Types

There are two types of spatial databases used

- **A Spatial Warehouse**
  - A spatial warehouse only stores spatial data.
  - It has data types defined for vector data but few or no functions to manipulate the data.
- **GIS Spatial Database**
  - Has data types for vector data
  - Some have data types for raster data
  - Have functions for manipulation and analysis of the data

Data types and models

Spatial Data types:

- Point : object represented only by its location in space
- Line : representation of moving through or connections in space
- Region : representation of an extent in 2d-space

Data types and models

- Partition: set of region objects that are required to be disjoint (adjacency or region objects with common boundaries).
- Networks: embedded graph in plane consisting of set of points (vertices) and lines (edges) objects, e.g. highways, power supply lines, rivers

Data types and models - spatial type system

EXT={lines, regions}, GEO={points, lines, regions}

- Spatial predicates for topological relationships:
  - **inside**: geo x regions → bool
  - **intersect**, **meets**: ext1 x ext2 → bool
  - **adjacent, encloses**: regions x regions → bool
- Operations returning atomic spatial data types:
  - **intersection**: lines x lines → points
  - **intersection**: regions x regions → regions
  - **plus**, **minus**: geo x geo → geo
  - **contour**: regions → lines

Spatial Relationship

- **Topological relationships**: adjacent, inside, disjoint.
- **Direction relationships**: e.g. above, below, or north_of, sothwest_of, …
- **Metric relationships**: e.g. distance

6 valid topological relationships between two simple regions: disjoint, in, touch, equal, cover, overlap

Data types and models

Data for a GIS comes in three basic forms :

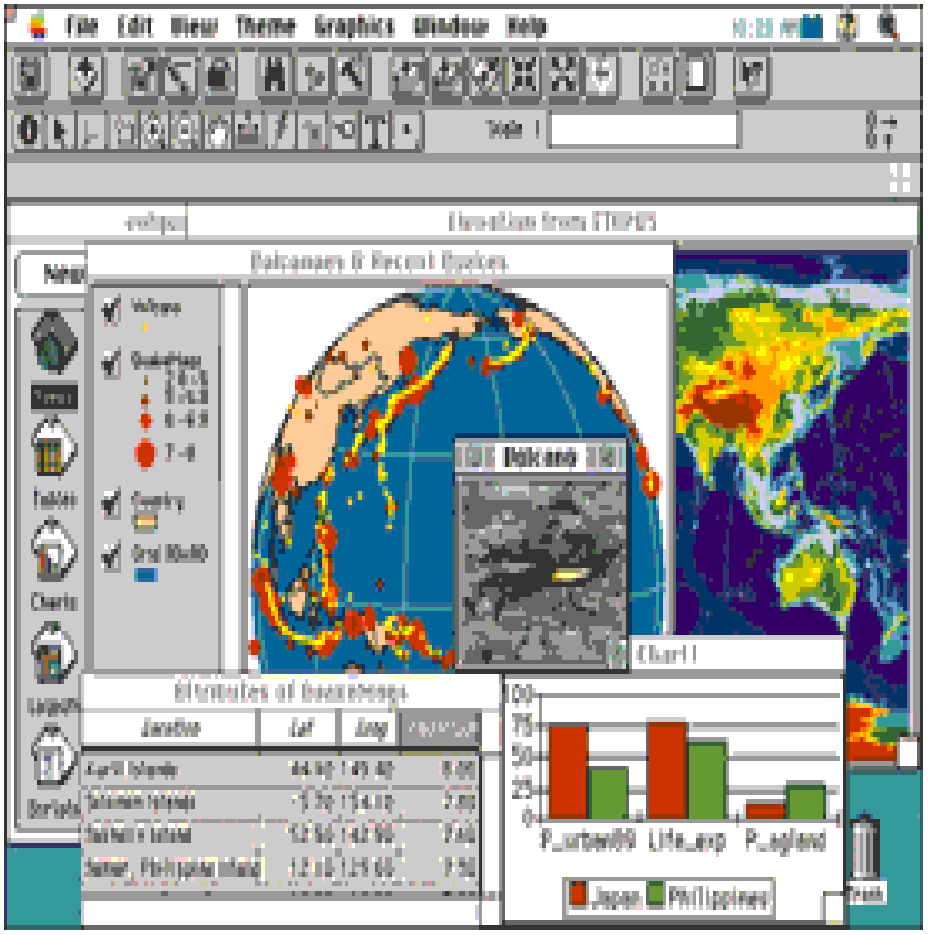
Spatial data:

- maps are made of spatial data, made up of points, lines, and areas.
- Spatial data forms the locations and shapes of map features such as buildings, streets, or cities.

  Spatial data are derived from existing maps or aerial photographs.

Image data—using images to build maps

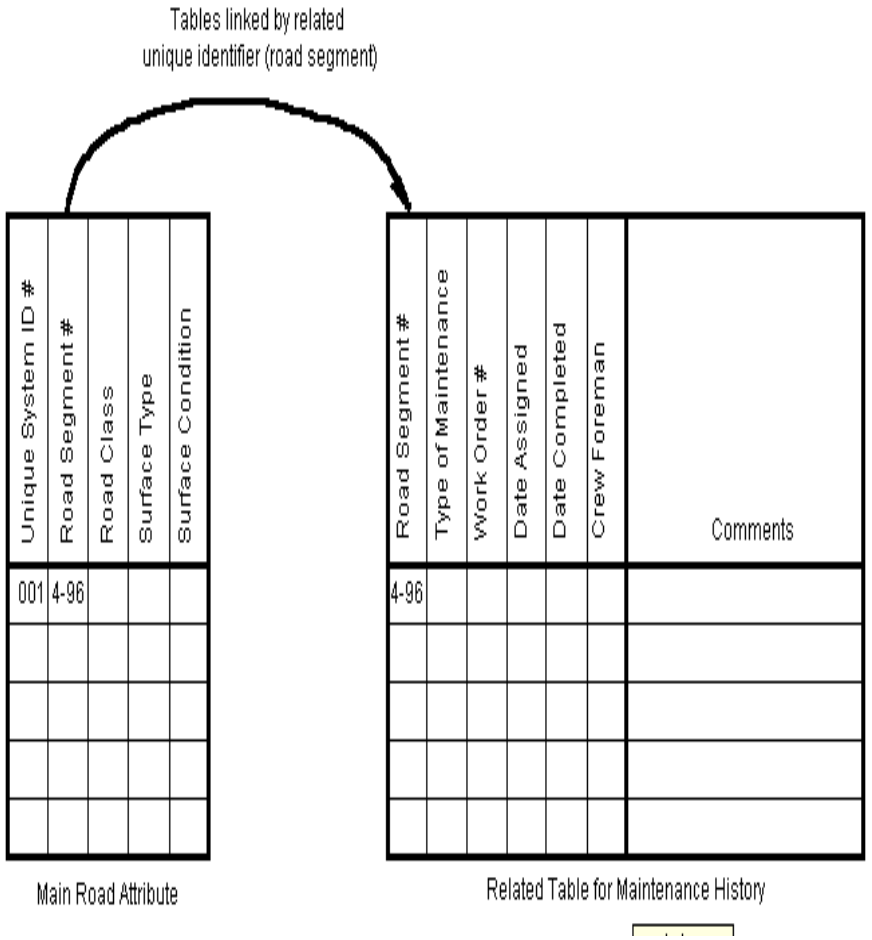
- Image data includes such diverse elements as satellite images, aerial photographs, and scanned data=> data that has been converted from paper to digital format



Data types and models

Tabular data:

- Tabular data is information describing a map feature. For example, a map of customer locations may be linked to demographic information about those customers.
- Use of a database management system (DBMS) to allow the user to define the specific data element types and formats and to store attribute.



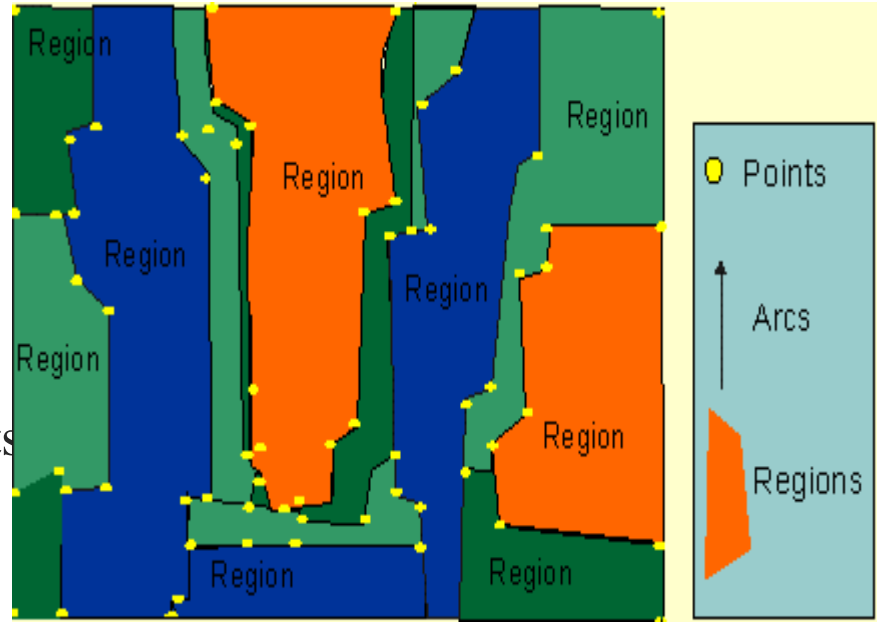
Data types and models

Data can be classified into two types of data models:

Vector model:

- displays graphical data as points, lines or curves, or areas with attributes.
- Cartesian coordinates and computational algorithms of the coordinates define points in a vector system.
- Lines or arcs are a series of ordered points. Areas or polygons are also stored as ordered lists of points

=> **Vector data requires less computer storage space and maintaining topological relationships is easier in this system.**

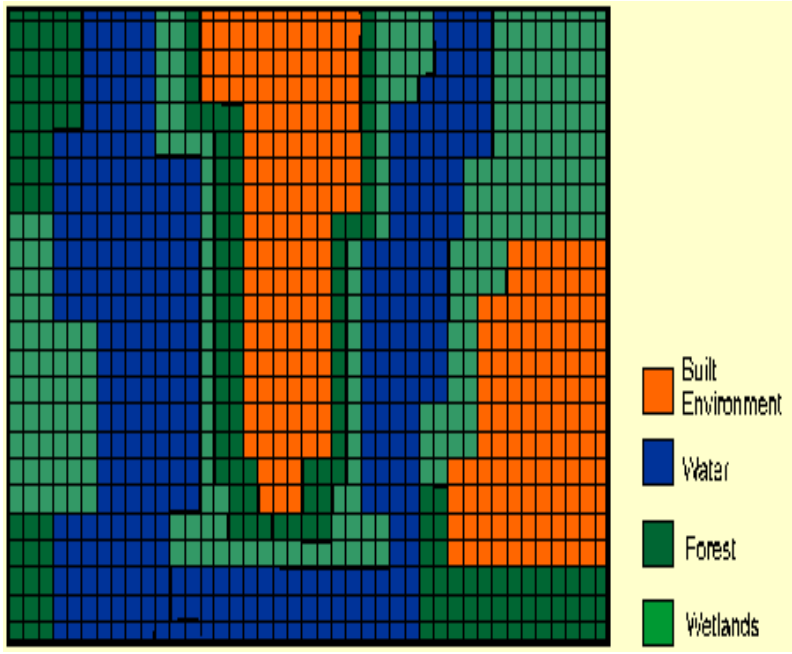


Data types and models

The raster view of the world:

- A raster based system displays, locates, and stores graphical data by using a matrix or grid of cells.
- these data are two-dimensional, GIS store various information such as forest cover, soil type, land use, wetland habitat, or other data in different layers

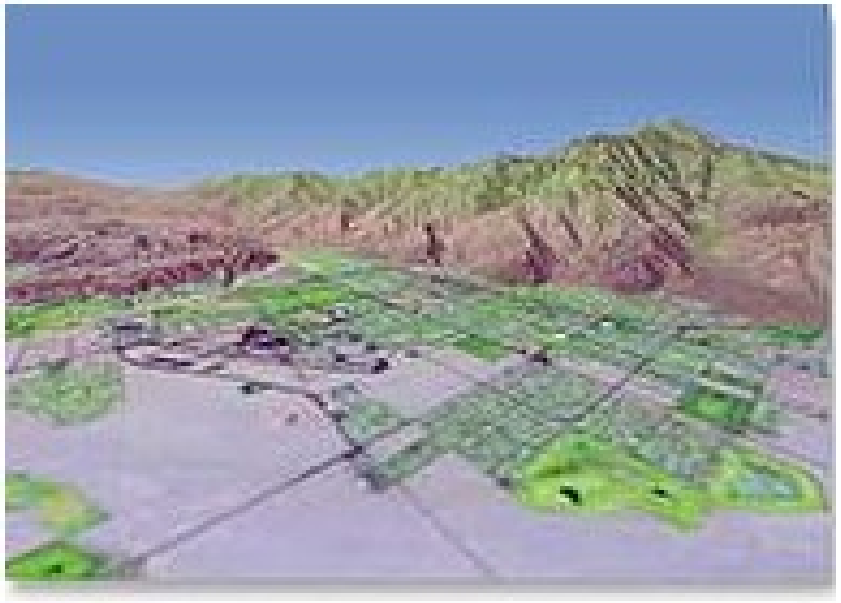
=> **Raster data requires less processing than vector data, but it consumes more computer storage space.**



Data types and models

Raster data model (cont) :

- Continuous numeric values, such as elevation, and continuous categories, such as vegetation types, are represented using the raster model.

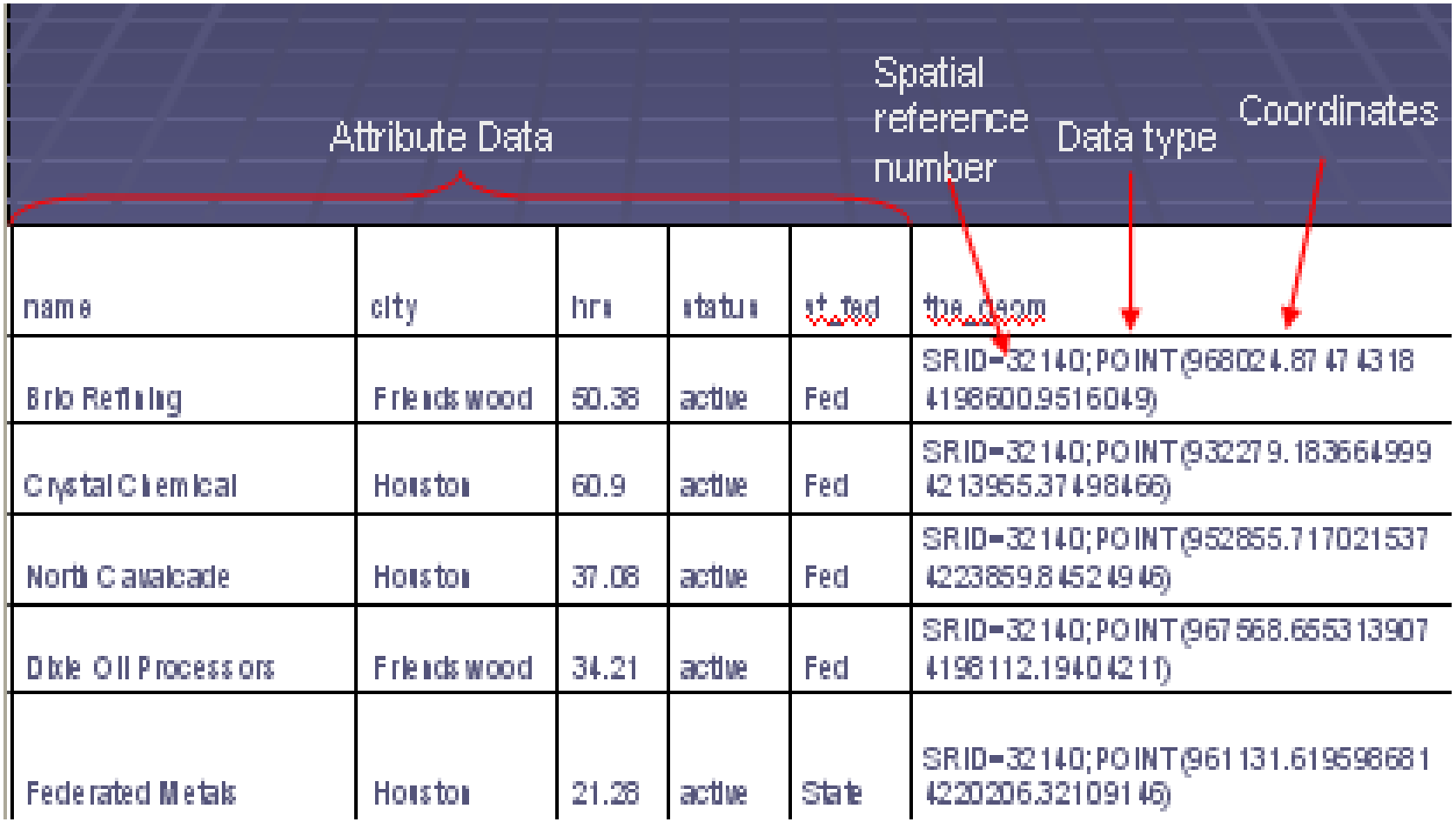


This map shows vector data laid on top of raster data.

How does it work?

- Spatial data is stored using the coordinate system of a particular projection
- That projection is referenced with a Spatial Reference Identification Number (SRID)
- This number corresponds to another table in the database with all of the spatial reference systems used.
- This allows the database to know what projection each table is in, and if need be, re-project those tables for calculations

Example of a table in Spatial Database



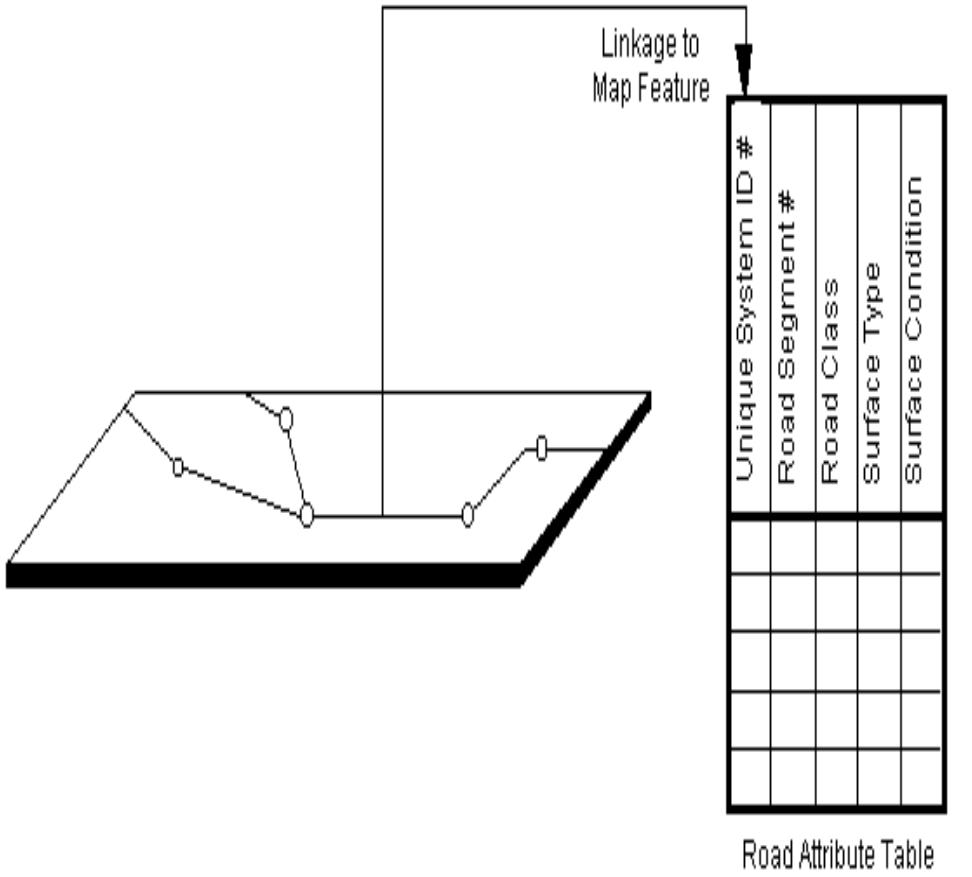
name	city	hrs	status	st_fed	the_geom
Brio Refining	Friendswood	50.38	active	Fed	SRID=32140;POINT(968024.87474318 4198600.9516049)
Crystal Chemical	Houston	60.9	active	Fed	SRID=32140;POINT(932279.183664999 4213955.37498466)
North Cavalcade	Houston	37.08	active	Fed	SRID=32140;POINT(952855.717021537 4223859.845249 45)
Dixie Oil Processors	Friendswood	34.21	active	Fed	SRID=32140;POINT(967568.6553139 07 4198112.19404211)
Federated Metals	Houston	21.28	active	State	SRID=32140;POINT(961131.619598681 4220206.32109145)

Linkage of Tabular Attributes to Map Feature

How does GIS link between spatial and non spatial data?

How does it link a Location Symbol with Its Meaning ?

- Every geographic feature has at least one unique means of identification: a name or number usually just called its ID.
- In other words, **locational information** is linked to specific information in a database



Models, Algebras, Languages

- Model: Extent relational model, or use Object-relational model.
- Spatial algebra: ex. ROSE algebra
- Query languages:
  - Extend SQL : GEOSQL, PSQL
  - New graphical languages: GEO-SAL

The extended Relational Model

- DBMS data model is extended by SDTs at the level of atomic data types (such as integer, string), or is open for user-defined types (OR-DBMS approach).

- **Examples**:

relation provinces (pname: STRING; area: REGION;
ppop: INTEGER)
relation cities (cname: STRING; center: POINT;
ext: REGION; cpop: INTEGER)
relation rivers (rname: STRING; route: LINE)

- Tables have unique values in at least one attribute: key

Rose Algebra

- An operation has the form:

Binary: (type1) op (type2) = type3

Unary: op (type2) = type3

Example of operations:

{lines, regions} inside regions = bool

{points, lines, regions} intersects, meets

lines intersection lines = points

regions intersection regions = regions

- Used to compute the distance, surface.
- **closest** : determines, among a set of objects, which one is closest to another object according to an attribute value

Querying

Two kinds of queries:

- Connecting the operations of a spatial algebra to the facilities of a DBMS query language. Fundamental spatial algebra operator are:
  - Spatial selection
  - Spatial join
- Providing graphical presentation of spatial data or results of queries, and graphical input of SDT values used in queries.

Querying – contd.

- **Spatial selection**: returning objects satisfying a spatial predicate with the query object

- **Spatial join**: A join which compares any two joined objects based on a predicate on their spatial attribute values.

Examples

Spatial selection:

"All cities in Morocco"

SELECT sname FROM cities c WHERE c.center inside Morocco.area.

"All big cities no more than 100 Kms from Ifrane"

SELECT cname FROM cities c
WHERE dist(c.center, Ifrane.center) < 100 and c.pop > 500k
(conjunction with other predicates and query optimization)

Spatial join:

SELECT cname, sname FROM cities, provinces WHERE center inside area

Create a table to hold county names, states, populations, and geographies

```
CREATE TABLE County(
    Name    varchar(30),
    State   varchar(30),
    Pop     Integer,
    Shape   Polygon);
```

Querying – Contd.

- how to find Morocco for ex. Or how to show all cities in a map?
- Requirements for spatial querying
  - Spatial data types
  - Graphical display of query results
  - Graphical combination (overlay) of several query results
  - Display of context (e.g., show background such as a raster image (satellite image) or boundary of states)
  - Legend: clarify the assignment of graphical representations to object classes

Data Structure and Algorithms

1. The implementation of Spatial Algebra should be integrated with DBMS querying.
2. The operations are not simply implemented using Computational geometry, rather it should consider the query processing access method and the spatial join.

Data Structure

- Representation of a spatial data type should be compatible with:

1. DBMS (Can have varying and possibly large size, reside permanently on disk page, can efficiently be loaded into memory)
2. Spatial algebra implementation(Supports efficient computational geometry algorithms for spatial algebra operations)

It should support:

- Approximations: stores some approximations (e.g. MBR – Minimum Boundary Rectangle) to speed up operations
- Stored unary function values: such as perimeter or area be stored once the object is constructed to eliminate future expensive computations.

Access Methods

- **Point Access Methods** (PAMs), A data structure and associated algorithms primarily to search for points defined in multidimensional space. (eg: cities, where each city is represented by three co-ordinates: longitude, latitude and altitude).
- **Spatial Access Methods** (SAMs), data structure to search for lines, polygons, etc. (eg: street segments, land plots).

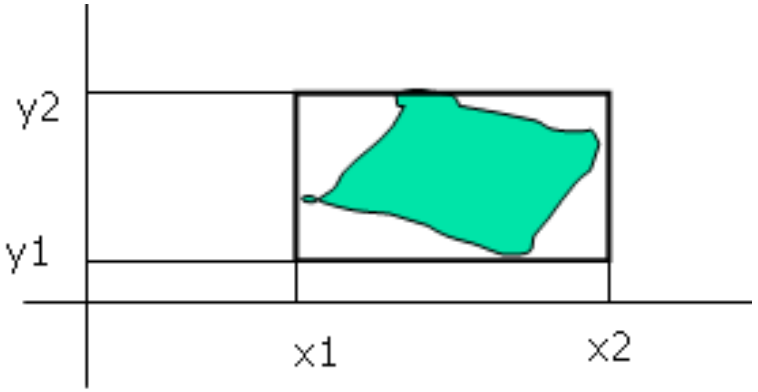
Indexing

- It organizes space and the objects in it in some way so that only parts of the objects need to be considered to answer a query.
- Spatial data store *points* or *rectangles* (for line or region)
- **Query types for points:**
  - Range query: all points within a query rectangle
  - Nearest neighbor: point closest to a query point
  - Distance scan: enumerate points in increasing distance from a query point.
- **Query types for rectangles:**
  - Intersection query
  - Containment query

Indexing

- I - Indexing using SAM:

  Approximate each region with a simple shape: usually **Minimum Bounding Rectangle** (MBR) = [(x1, x2), (y1, y2)]
- II- Use of Grid as an other approximation key (a geometric entity as a set of cells).



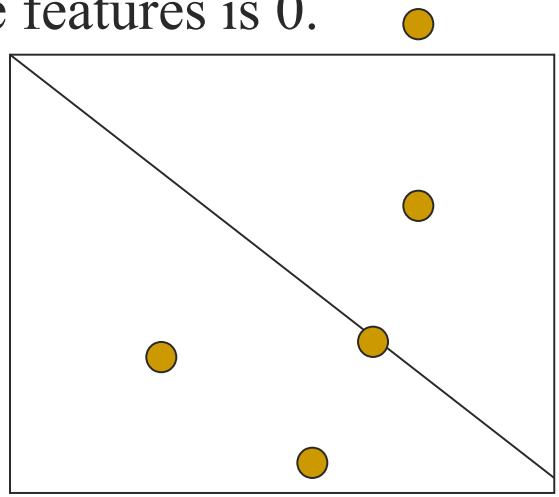
The indexing using SAM happens two steps:

Filtering step: Find all the MBRs that satisfy the query

Refinement step: For each qualified MBR, check the original object against the query

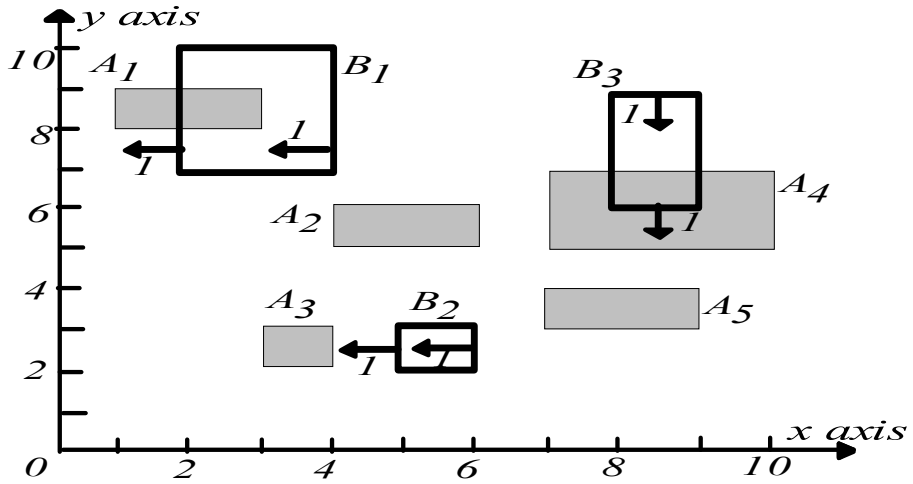
Indexing – Contd.

- In this example, 4 points are within the bounding box of the line
- If you perform a query to find out if any points intersect the line, you will have 4 points returned. **Wrong!!**
- This is where you have to be creative with creating queries.
- Instead of using intersect, query to find where distance between the features is 0.



Spatial join

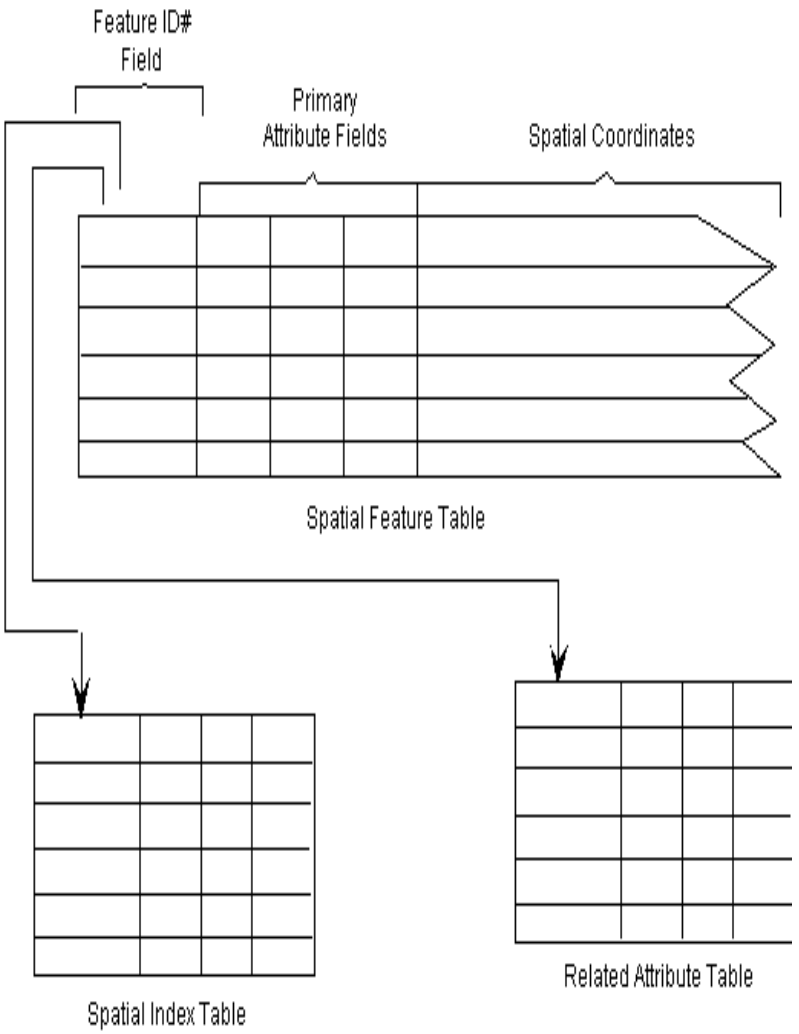
- Unlike equi-join in relational DBMS, intersection join!



Result:
R={(A1, B1), (B3, A4)}

Spatial Data Repositories

- allow the storage and management of all spatial data (map features, attributes, and even raster data) in a relational database.
- move away from the traditional concept of a GIS database in which spatial data is stored in a graphic format and tabular attributes are stored in a separate database table.
- These packages provide a more open architecture for managing GIS data and allow access with multiple client software for data viewing, update and analysis.



format and quality of the GIS database

Coordinate Systems:

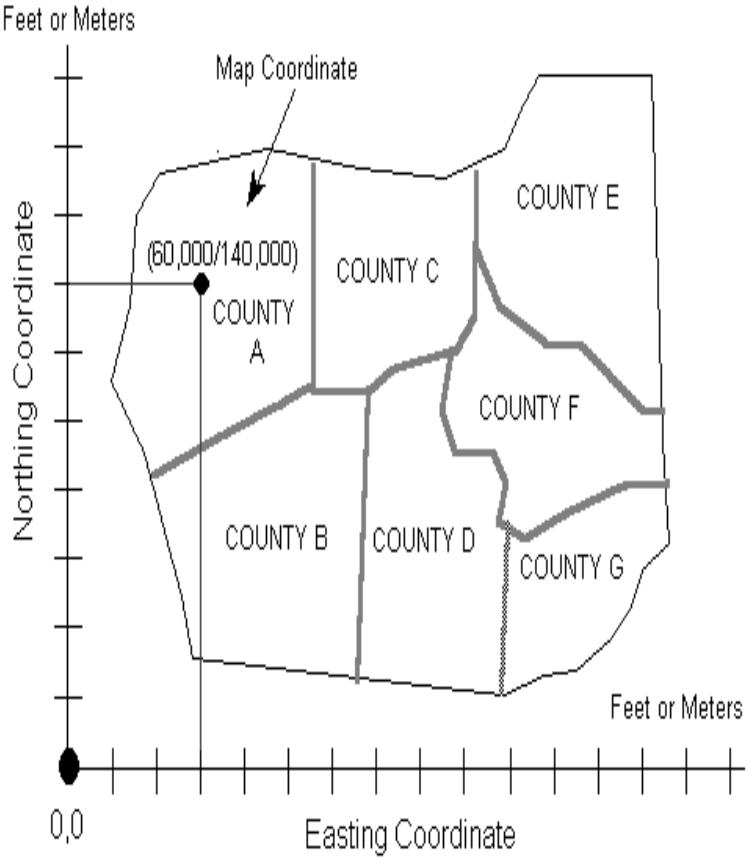
- Projects the curved Earth surface onto a flat surface
- State Plane Coordinate System (SPCS) is used as an origin for a specific geographic region or "State Plane Zone" in which (x,y) coordinates are defined from the origin of the zone.

Spatial Accuracy: Spatial accuracy specifies how well the position of features or boundaries, as plotted on a map, conforms to their actual position on the ground.

Scale: relationship between mapped size and actual size.

Attribute Accuracy: the integrity (and frequency of errors) in the values of attributes entered in the GIS database.

Currency: Database components change frequently and update routine should be available and easy to use.



Building a Spatial Database

- A good geodatabase has rules and constraints
- Quality Control components
  - used to protect the integrity of the data
  - prevent human error

Rules:

- Rules help prevent human error when modifying a data set
- Rules are user defined

Constraints:

- Constraints are similar to rules, but are less assertive.
- Constraints are provided by the DBMS and are applied by the user

Dynamic and Static Data

- Static data is usually maintained in the table with the geometry
- Dynamic data is maintained in a separate table
- Some dynamic tables are generated by computer automation
- A good example of this is weather data
- Permissions for these different tables are independent

Tools

- Oracle Spatial data cartridge, ESRI SDE
  - It can work with Oracle 8i DBMS
  - It Has spatial data types (e.g. polygon), operations (e.g. overlap) callable from SQL3 query language
  - It Has spatial indices.

Standard

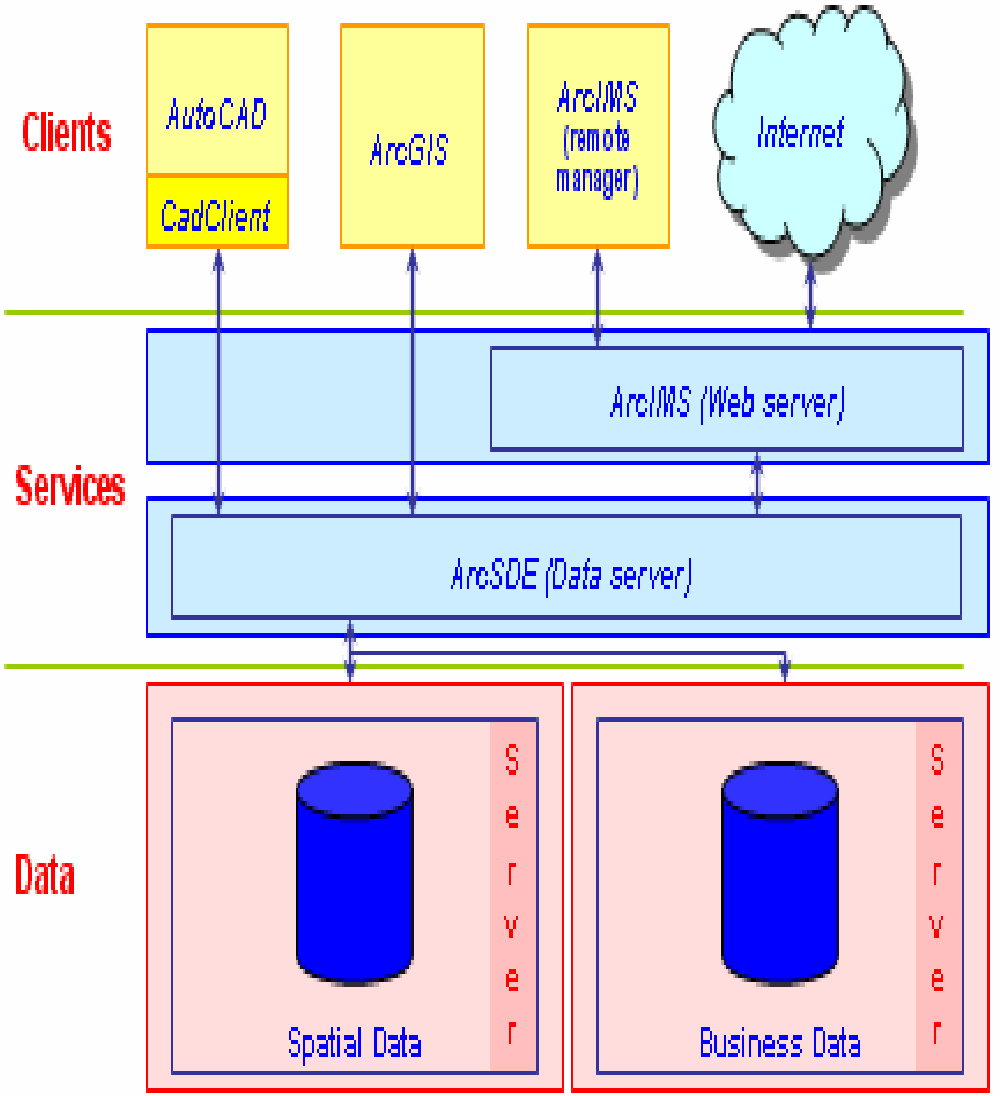
- Follow the OpenGIS.
- Any contract requiring internet deployment of spatial data in the form of maps should be undertaken using Open GIS standards.
- These standards for geospatial information are developed and administered by the OpenGIS Consortium (OGC).
- It support interoperable geospatial data integration and services over the web.

Applications of Spatial Databases

- Web Hosting
- Data Hosting
- Data management for Desktop GIS
- Custom Applications

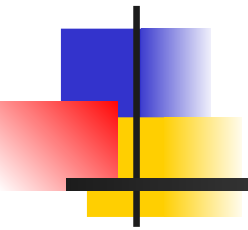
System design & architecture

- Spatial data stored in a SQL Server 7.0 database.
- ArcGIS to update spatial data.
- ArcGIS to perform spatial analysis, develop and plot maps as needed.
- CadClient to facilitate exchange of data between geodatabase and AutoCAD.
- ArcIMS to deploy spatial information over the web.
- ArcSDE to hold all the pieces together.



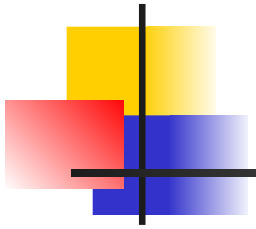
Summary

- Without spatial database you can not design GIS application.
- SDBMS is a software module
  - works with an underlying DBMS
  - provides spatial data typess callable from a query language
  - provides methods for efficient processing of spatial queries
- Components of SDBMS include
  - spatial data model, spatial data types and operators,
  - spatial query language, processing
- SDBMS is used to store, query and share spatial data for GIS as well as other applications



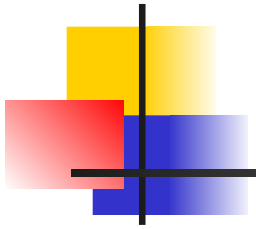
Data Mining – Intro

Course Overview



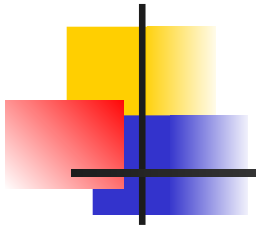
- Spatial Databases
- Temporal Databases
- Spatio-Temporal Databases
- Data Mining 

Data Mining Overview



- Data Mining
  - Data warehouses and OLAP (On Line Analytical Processing.)
  - Association Rules Mining
  - Clustering: Hierarchical and Partitional approaches
  - Classification: Decision Trees and Bayesian classifiers
  - Sequential Patterns Mining
  - Advanced topics: outlier detection, web mining

What is Data Mining?

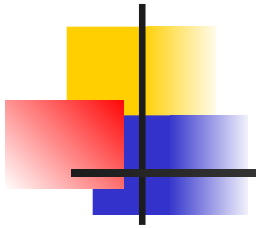


- Data Mining is:

  (1) The efficient discovery of previously unknown, valid, potentially useful, understandable patterns in large datasets

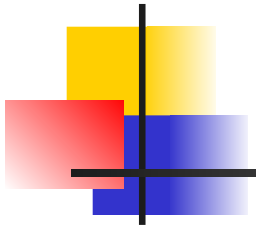
  (2) The analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner

What is Data Mining?



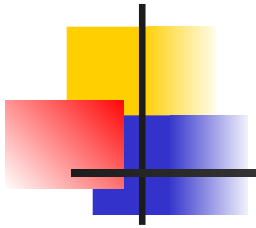
- Very little functionality in database systems to support mining applications
- Beyond SQL Querying:
  - SQL (OLAP) Query:
    - How many widgets did we sell in the 1st Qtr of 1999 in California vs New York?
  - Data Mining Queries:
    - Which sales region had anomalous sales in the 1st Qtr of 1999
    - How do the buyers of widgets in California and New York differ?
    - What else do the buyers of widgets in Cal buy along with widgets

Overview of terms



- Data: a set of facts (items) D, usually stored in a database
- Pattern: an expression E in a language L, that describes a subset of facts
- Attribute: a field in an item *i* in D.
- Interestingness: a function $I_{D,L}$ that maps an expression E in L into a measure space M

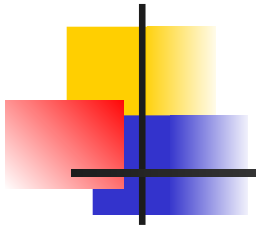
Overview of terms



- The **Data Mining Task:**

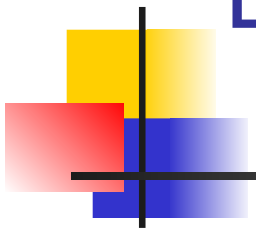
  For a given dataset D, language of facts L, interestingness function $I_{D,L}$ and threshold c, find the expression E such that $I_{D,L}(E) > c$ efficiently.

Examples of Large Datasets



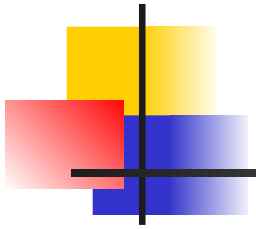
- Government: IRS, …
- Large corporations
  - WALMART: 20M transactions per day
  - MOBIL: 100 TB geological databases
  - AT&T 300 M calls per day
- Scientific
  - NASA, EOS project: 50 GB per hour
  - Environmental datasets

Examples of Data mining Applications



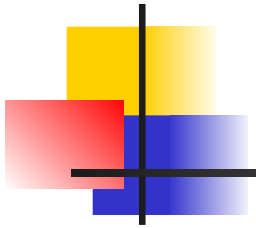
1. Fraud detection: credit cards, phone cards
2. Marketing: customer targeting
3. Data Warehousing: Walmart
4. Astronomy
5. Molecular biology

How Data Mining is used



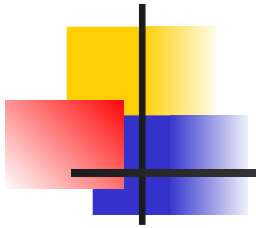
1. Identify the problem
2. Use data mining techniques to transform the data into information
3. Act on the information
4. Measure the results

The Data Mining Process



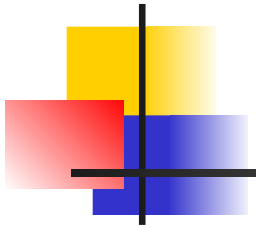
1. Understand the domain
2. Create a dataset:
   - Select the interesting attributes
   - Data cleaning and preprocessing
3. Choose the data mining task and the specific algorithm
4. Interpret the results, and possibly return to 2

Data Mining Tasks



1. Classification: learning a function that maps an item into one of a set of predefined classes
2. Regression: learning a function that maps an item to a real value
3. Clustering: identify a set of groups of similar items

Data Mining Tasks

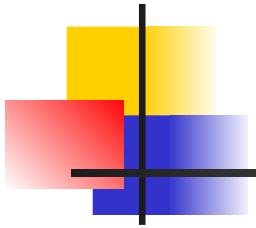


4. Dependencies and associations:

   identify significant dependencies between data attributes

5. Summarization: find a compact description of the dataset or a subset of the dataset

Data Mining Methods

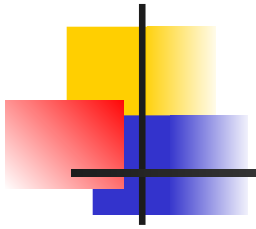


1. Decision Tree Classifiers:
   Used for modeling, classification
2. Association Rules:
   Used to find associations between sets of attributes
3. Sequential patterns:
   Used to find temporal associations in time series
4. Hierarchical clustering:
   used to group customers, web users, etc

Are All the "Discovered" Patterns Interesting?

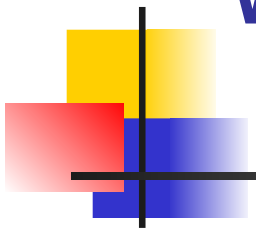
- **Interestingness measures**: A pattern is interesting if it is easily understood by humans, valid on new or test data with some degree of certainty, potentially useful, novel, or validates some hypothesis that a user seeks to confirm
- **Objective vs. subjective interestingness measures:**
  - Objective: based on statistics and structures of patterns, e.g., support, confidence, etc.
  - Subjective: based on user's belief in the data, e.g., unexpectedness, novelty, actionability, etc.

Can We Find All and Only Interesting Patterns?



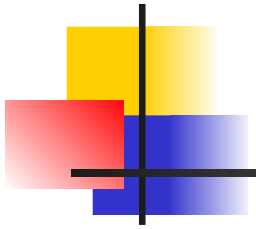
- Find all the interesting patterns: Completeness
  - Can a data mining system find all the interesting patterns?
  - Association vs. classification vs. clustering
- Search for only interesting patterns: Optimization
  - Can a data mining system find only the interesting patterns?
  - Approaches
    - First general all the patterns and then filter out the uninteresting ones.
    - Generate only the interesting patterns—mining query optimization

Why Data Preprocessing?



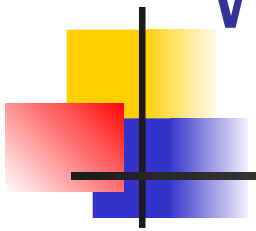
- Data in the real world is dirty
  - incomplete: lacking *attribute values*, lacking certain *attributes of interest*, or containing only aggregate data
  - noisy: containing errors or outliers
  - inconsistent: containing discrepancies in codes or names
- No quality data, no quality mining results!
  - Quality decisions must be based on quality data
  - Data warehouse needs consistent integration of quality data
  - Required for both OLAP and Data Mining!

Why can Data be Incomplete?



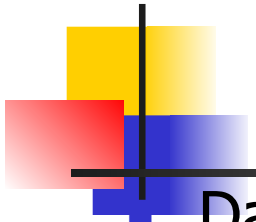
- Attributes of interest are not available (e.g., customer information for sales transaction data)
- Data were not considered important at the time of transactions, so they were not recorded!
- Data not recorder because of misunderstanding or malfunctions
- Data may have been recorded and later deleted!
- Missing/unknown values for some data

Why can Data be Noisy/Inconsistent?



- Faulty instruments for data collection
- Human or computer errors
- Errors in data transmission
- Technology limitations (e.g., sensor data come at a faster rate than they can be processed)
- Inconsistencies in naming conventions or data codes (e.g., 2/5/2002 could be 2 May 2002 or 5 Feb 2002)
- Duplicate tuples, which were received twice should also be removed

Major Tasks in Data Preprocessing

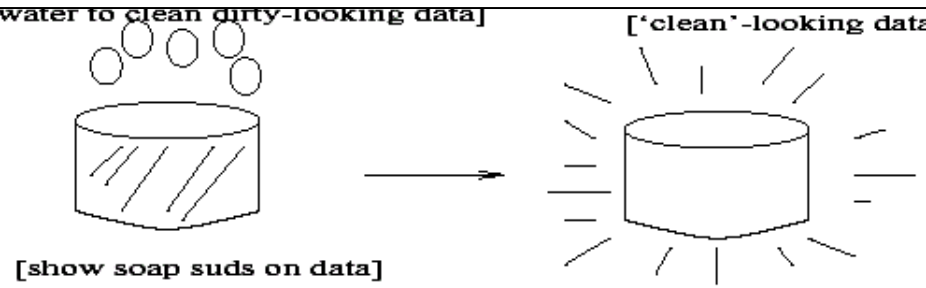


outliers=exceptions!

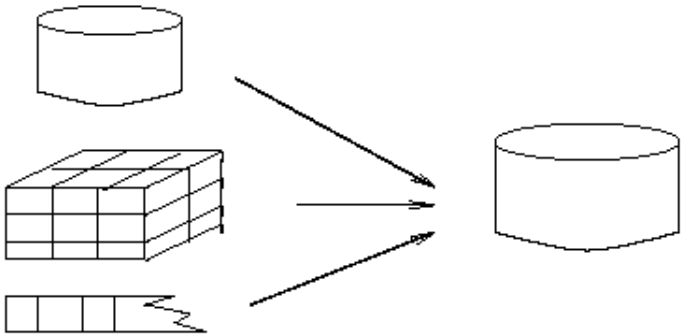
- Data cleaning
  - Fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies
- Data integration
  - Integration of multiple databases or files
- Data transformation
  - Normalization and aggregation
- Data reduction
  - Obtains reduced representation in volume but produces the same or similar analytical results
- Data discretization
  - Part of data reduction but with particular importance, especially for numerical data

Forms of data preprocessing

Data Cleaning



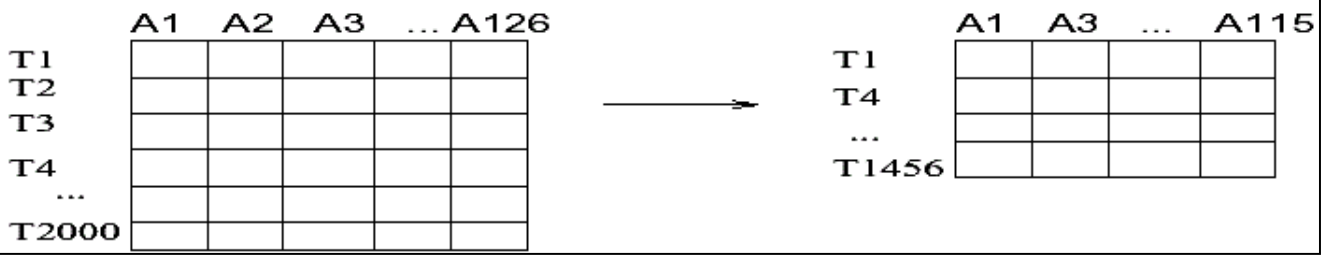
Data Integration



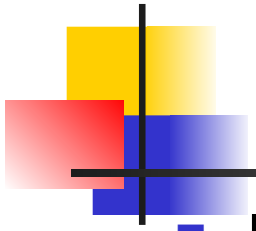
Data Transformation

-2, 32, 100, 59, 48 → -0.02, 0.32, 1.00, 0.59, 0.48

Data Reduction

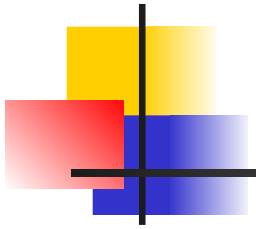


Data Cleaning



- Data cleaning tasks
  - Fill in missing values
  - Identify outliers and smooth out noisy data
  - Correct inconsistent data

How to Handle Missing Data?



- Ignore the tuple: usually done when class label is missing (assuming the tasks in classification)—not effective when the percentage of missing values per attribute varies considerably.
- Fill in the missing value manually: tedious + infeasible?
- Use a global constant to fill in the missing value: e.g., "unknown", a new class?!
- Use the attribute mean to fill in the missing value
- Use the attribute mean for all samples belonging to the same class to fill in the missing value: smarter
- Use the most probable value to fill in the missing value: inference-based such as Bayesian formula or decision tree

How to Handle Missing Data?

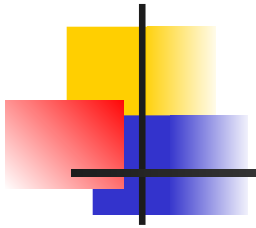
Age	Income	Team	Gender
23	24,200	Red Sox	M
39	?	Yankees	F
45	45,390	?	F

Fill missing values using aggregate functions (e.g., average) or probabilistic estimates on global value distribution

E.g., put the average income here, or put the most probable income based on the fact that the person is 39 years old

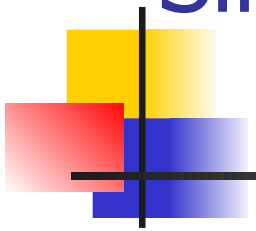
E.g., put the most frequent team here

How to Handle Noisy Data? Smoothing techniques



- Binning method:
  - first sort data and partition into (equi-depth) bins
  - then one can smooth by bin means, smooth by bin median, smooth by bin boundaries, etc.
- Clustering
  - detect and remove outliers
- Combined computer and human inspection
  - computer detects suspicious values, which are then checked by humans
- Regression
  - smooth by fitting the data into regression functions

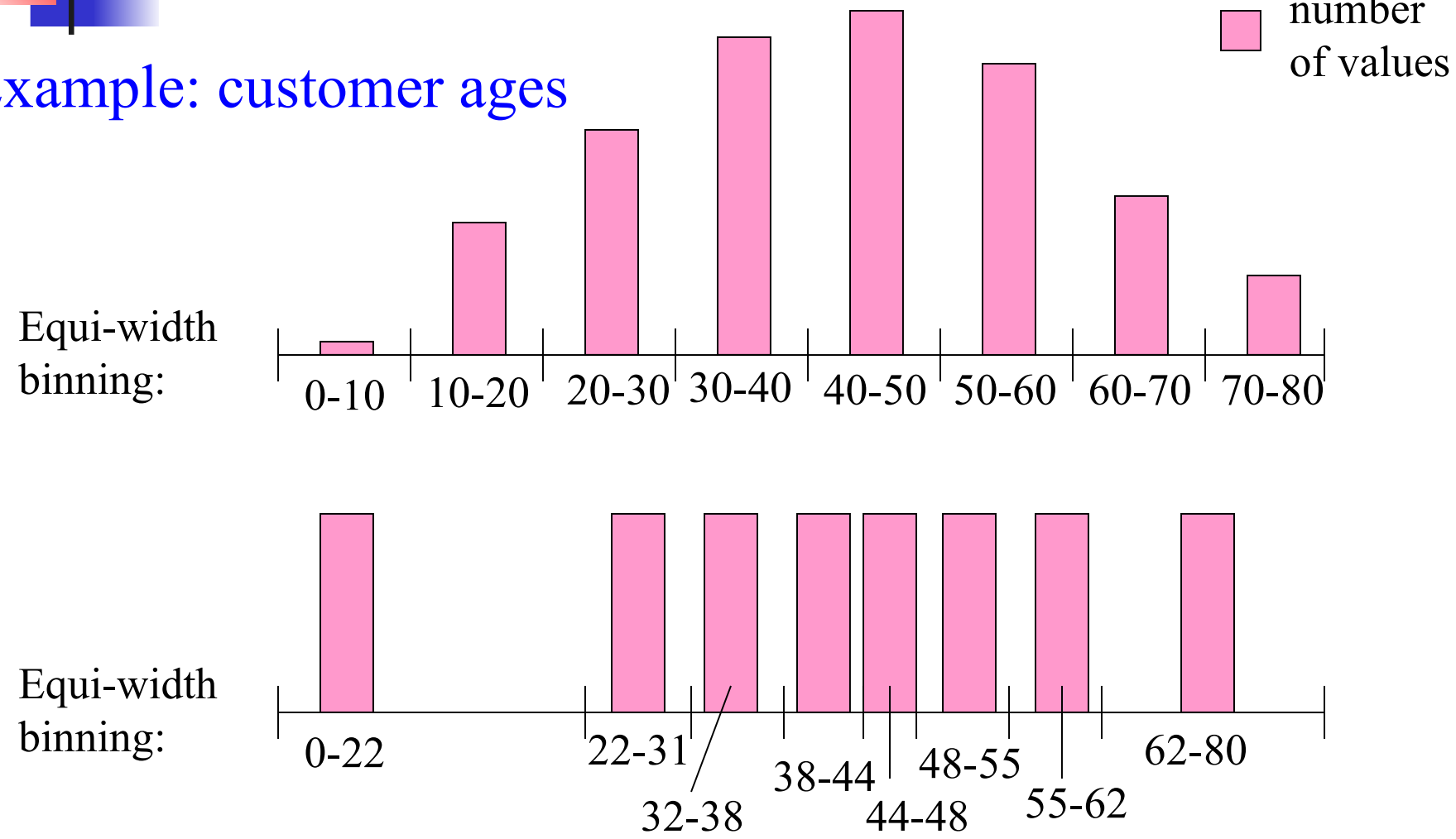
Simple Discretization Methods: Binning



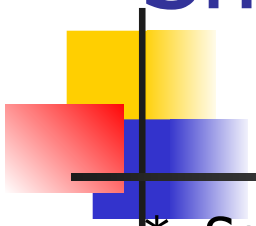
- Equal-width (distance) partitioning:
  - It divides the range into *N* intervals of equal size: uniform grid
  - if *A* and *B* are the lowest and highest values of the attribute, the width of intervals will be: $W = (B-A)/N$.
  - The most straightforward
  - But outliers may dominate presentation
  - Skewed data is not handled well.
- Equal-depth (frequency) partitioning:
  - It divides the range into *N* intervals, each containing approximately same number of samples
  - Good data scaling – good handing of skewed data

Simple Discretization Methods: Binning

Example: customer ages

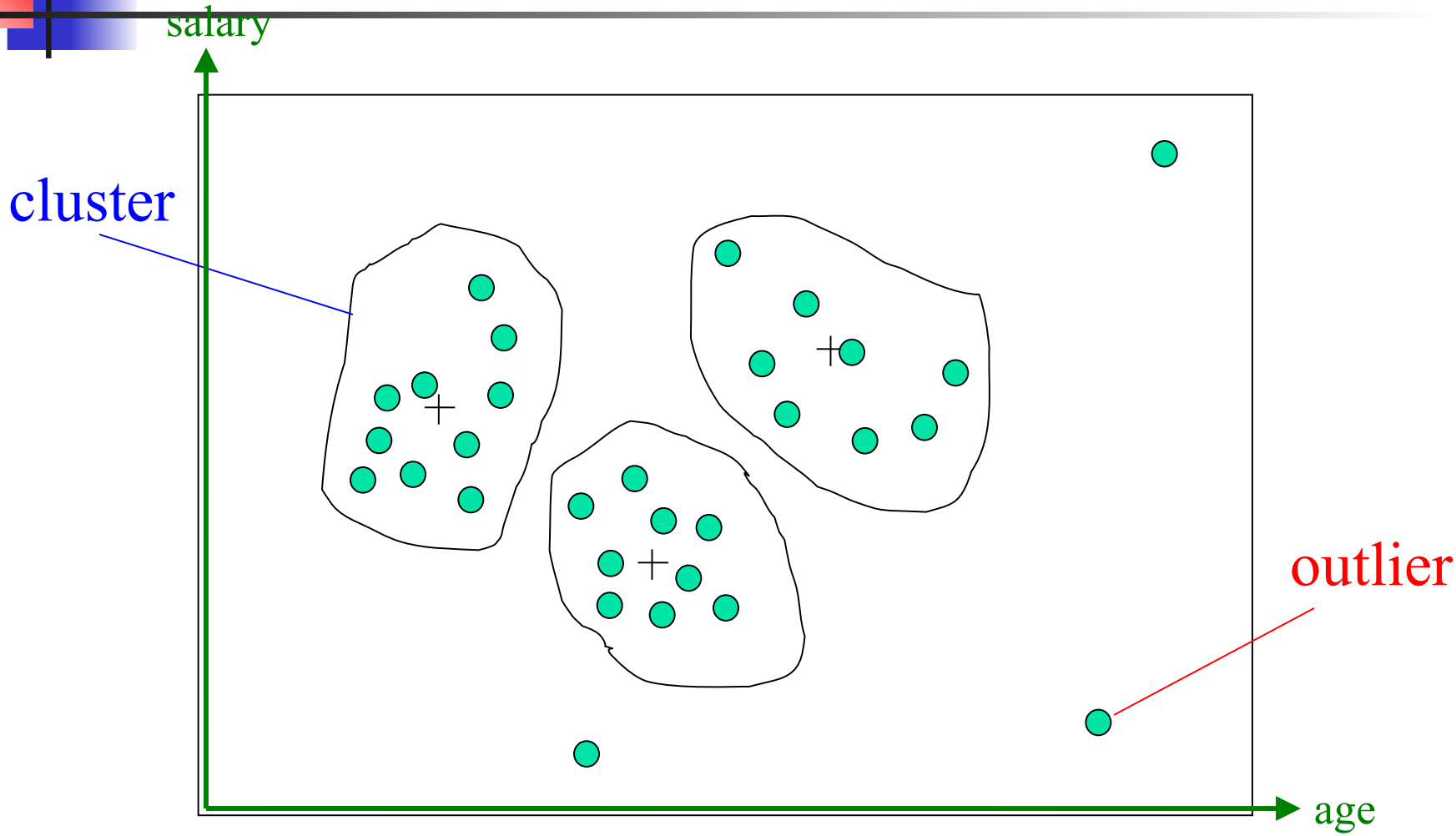


Smoothing using Binning Methods



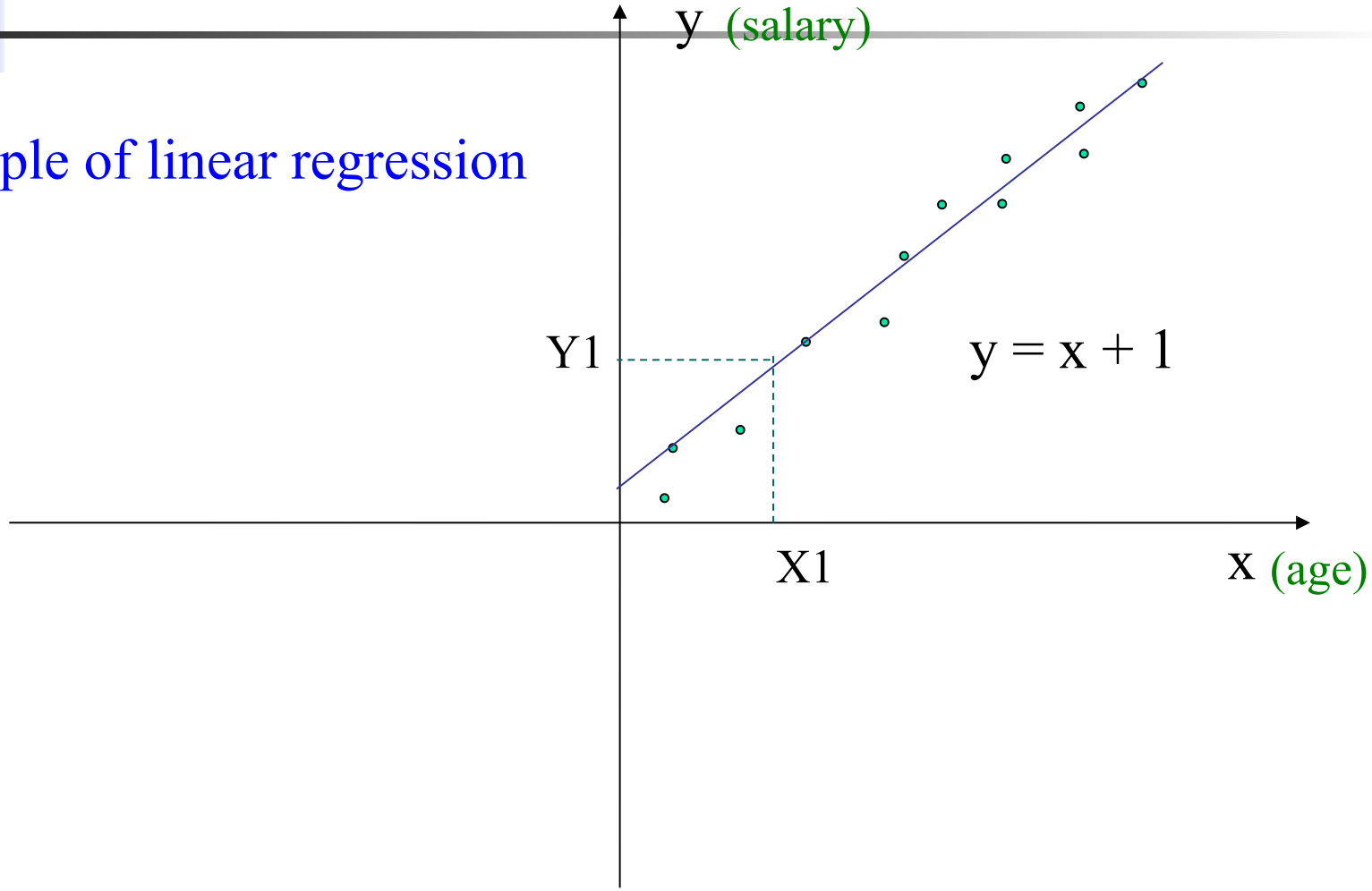
* Sorted data for price (in dollars): 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34
* Partition into (equi-depth) bins:
  - Bin 1: 4, 8, 9, 15
  - Bin 2: 21, 21, 24, 25
  - Bin 3: 26, 28, 29, 34
* Smoothing by bin means:
  - Bin 1: 9, 9, 9, 9
  - Bin 2: 23, 23, 23, 23
  - Bin 3: 29, 29, 29, 29
* Smoothing by bin boundaries: [4,15],[21,25],[26,34]
  - Bin 1: 4, 4, 4, 15
  - Bin 2: 21, 21, 25, 25
  - Bin 3: 26, 26, 26, 34

Cluster Analysis

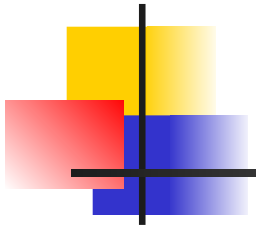


Regression

Example of linear regression

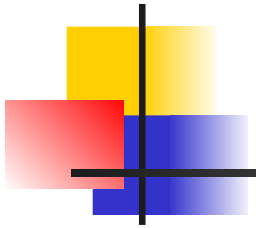


Data Integration



- Data integration:
  - combines data from multiple sources into a coherent store
- Schema integration
  - integrate metadata from different sources
    - metadata: data about the data (i.e., data descriptors)
  - Entity identification problem: identify real world entities from multiple data sources, e.g., A.cust-id ≡ B.cust-#
- Detecting and resolving data value conflicts
  - for the same real world entity, attribute values from different sources are different (e.g., J.D.Smith and Jonh Smith may refer to the same person)
  - possible reasons: different representations, different scales, e.g., metric vs. British units (inches vs. cm)

Data Transformation



- Smoothing: remove noise from data
- Aggregation: summarization, data cube construction
- Generalization: concept hierarchy climbing
- Normalization: scaled to fall within a small, specified range
  - min-max normalization
  - z-score normalization
  - normalization by decimal scaling
- Attribute/feature construction
  - New attributes constructed from the given ones

Normalization: Why normalization?

- Speeds-up learning, e.g., neural networks
- Helps prevent attributes with large ranges outweigh ones with small ranges
  - Example:
    - income has range 3000-200000
    - age has range 10-80
    - gender has domain M/F

Data Transformation: Normalization

- min-max normalization

$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$$

  - e.g. convert age=30 to range 0-1, when min=10,max=80. new_age=(30-10)/(80-10)=2/7

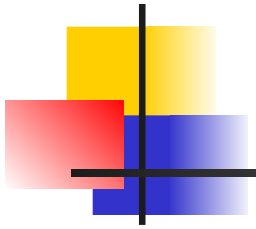
- z-score normalization

$$v' = \frac{v - mean_A}{stand\_dev_A}$$

- normalization by decimal scaling

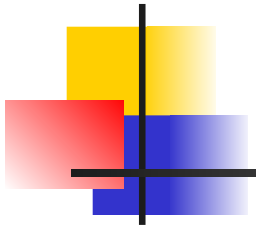
$$v' = \frac{v}{10^j}$$ Where j is the smallest integer such that $\text{Max}(|v'|)<1$

Data Reduction Strategies



- Warehouse may store terabytes of data: Complex data analysis/mining may take a very long time to run on the complete data set
- Data reduction
  - Obtains a reduced representation of the data set that is much smaller in volume but yet produces the same (or almost the same) analytical results

Dimensionality Reduction



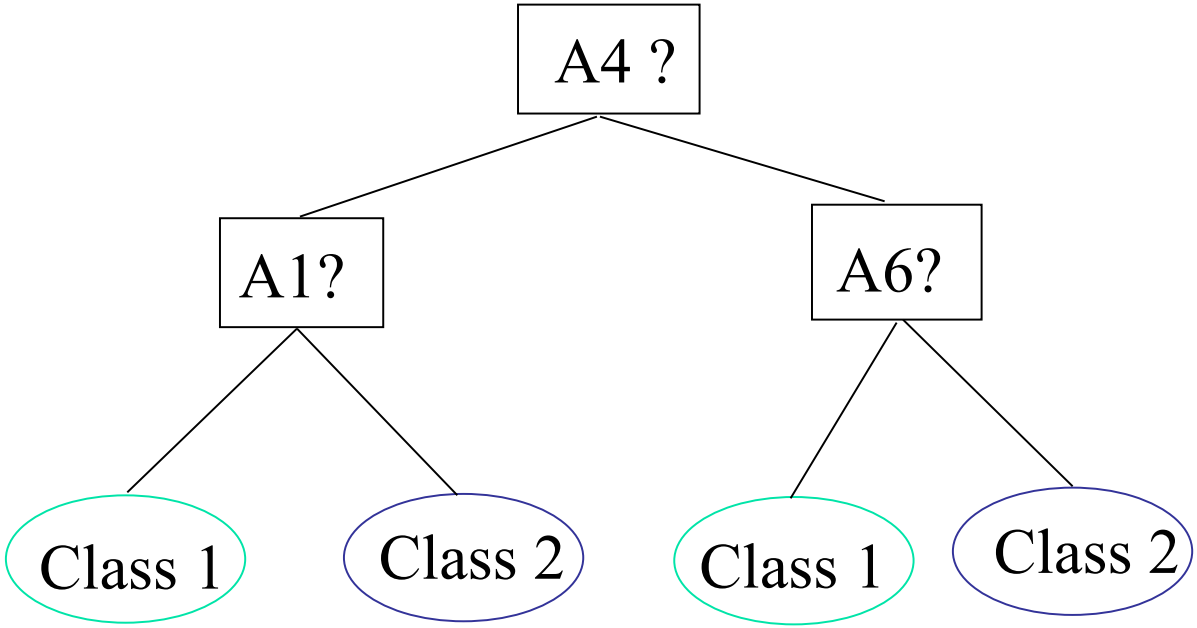
- Feature selection (i.e., attribute subset selection):
  - Select a minimum set of features such that the probability distribution of different classes given the values for those features is as close as possible to the original distribution given the values of all features
  - reduce # of patterns in the patterns, easier to understand
- Heuristic methods (due to exponential # of choices):
  - step-wise forward selection
  - step-wise backward elimination
  - combining forward selection and backward elimination
  - decision-tree induction

Heuristic Feature Selection Methods

- There are 2^d possible sub-features of d features
- Several heuristic feature selection methods:
  - Best single features under the feature independence assumption: choose by significance tests.
  - Best step-wise feature selection:
    - The best single-feature is picked first
    - Then next best feature condition to the first, ...
  - Step-wise feature elimination:
    - Repeatedly eliminate the worst feature
  - Best combined feature selection and elimination:
  - Optimal branch and bound:
    - Use feature elimination and backtracking

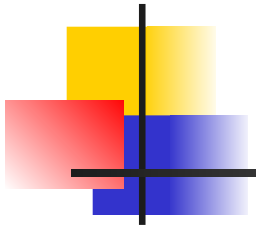
Example of Decision Tree Induction

Initial attribute set:
{A1, A2, A3, A4, A5, A6}



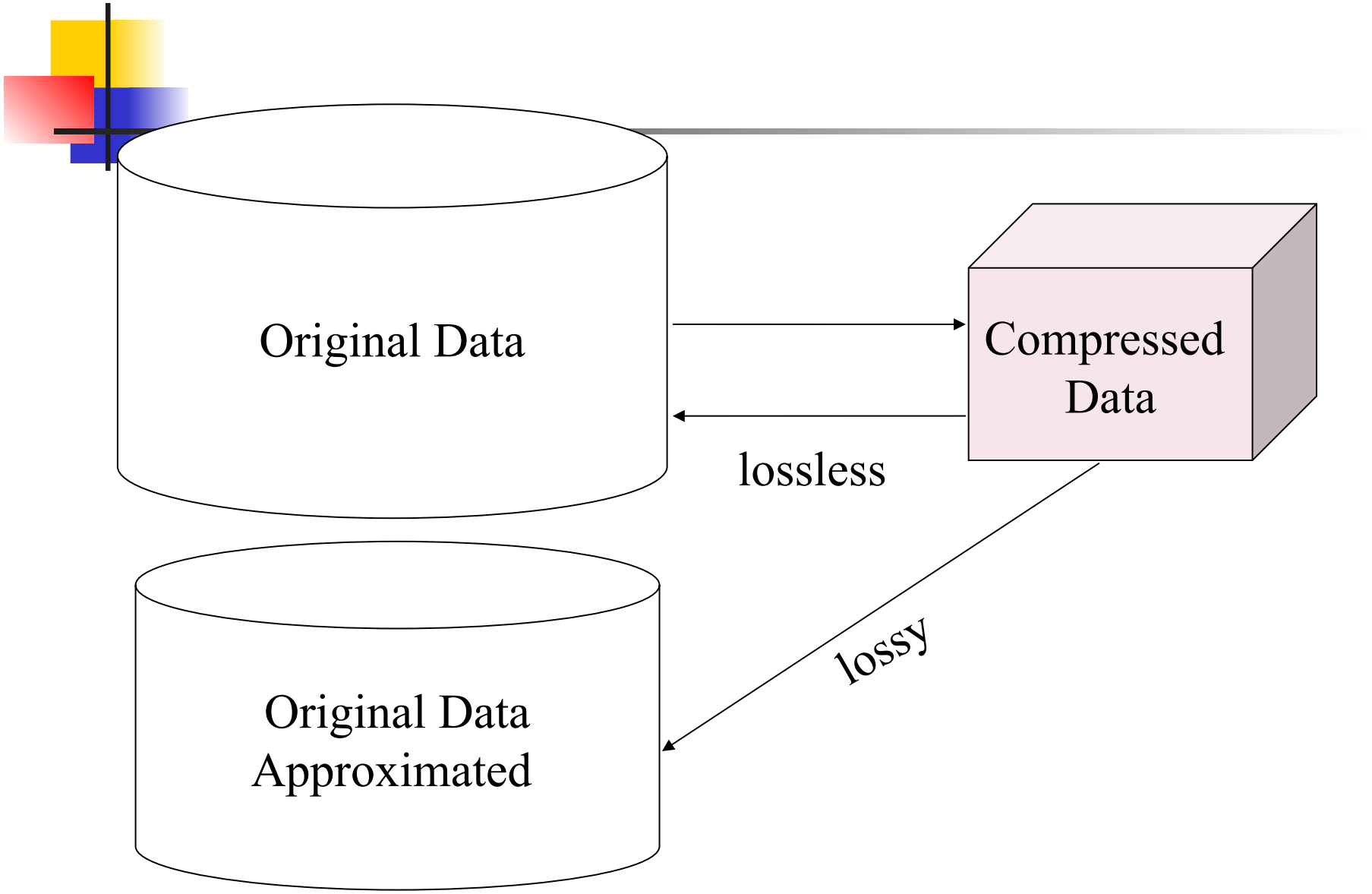
------> Reduced attribute set: {A1, A4, A6}

Data Compression

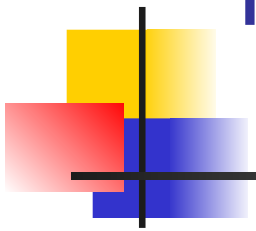


- String compression
  - There are extensive theories and well-tuned algorithms
  - Typically lossless
  - But only limited manipulation is possible without expansion
- Audio/video compression
  - Typically lossy compression, with progressive refinement
  - Sometimes small fragments of signal can be reconstructed without reconstructing the whole
- Time sequence is not audio
  - Typically short and varies slowly with time

Data Compression



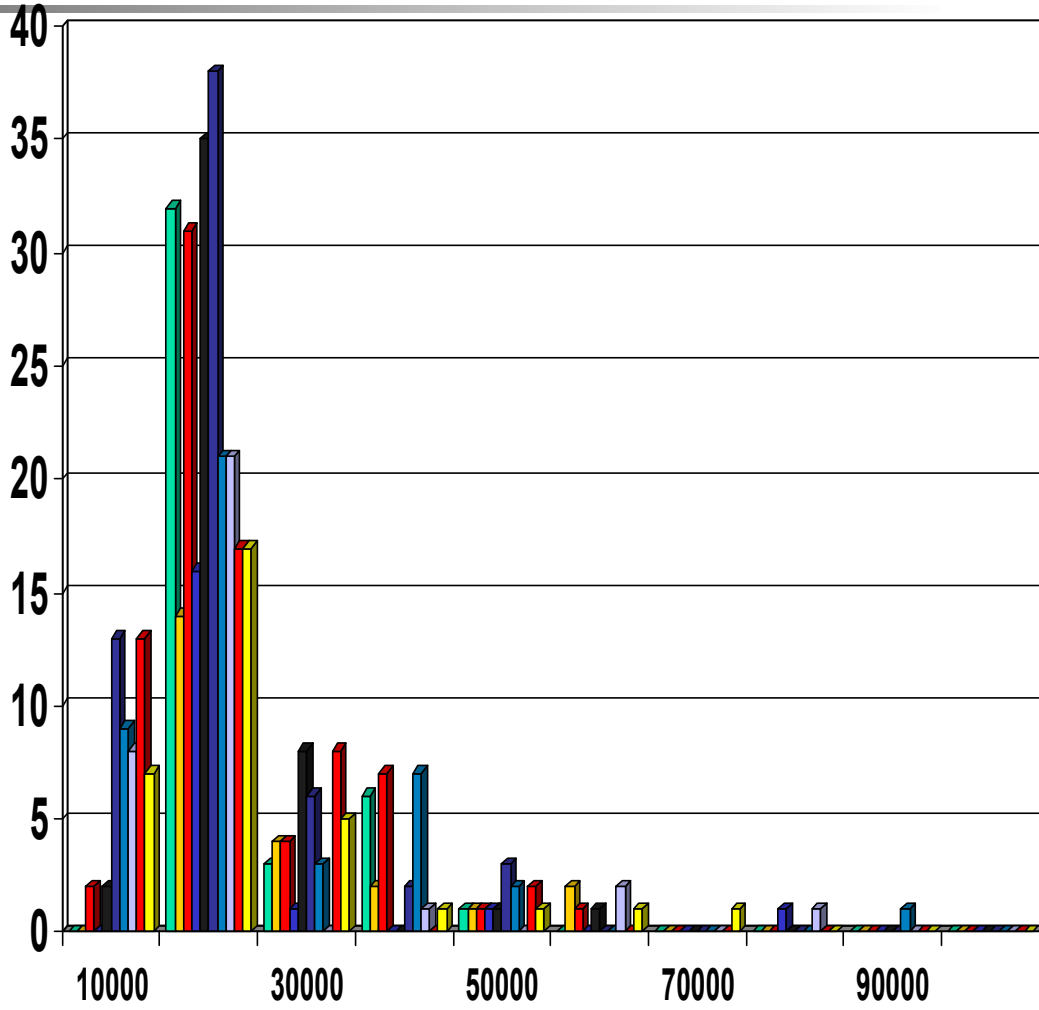
Numerosity Reduction: Reduce the **volume** of data



- Parametric methods
  - Assume the data fits some model, estimate model parameters, store only the parameters, and discard the data (except possible outliers)
  - Log-linear models: obtain value at a point in m-D space as the product on appropriate marginal subspaces
- Non-parametric methods
  - Do not assume models
  - Major families: histograms, clustering, sampling

Histograms

- A popular data reduction technique
- Divide data into buckets and store average (or sum) for each bucket
- Can be constructed optimally in one dimension using dynamic programming
- Related to quantization problems.

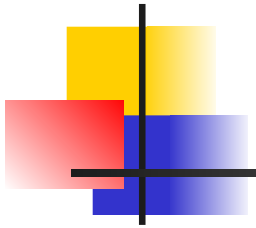


Histogram types

- Equal-width histograms:
  - It divides the range into *N* intervals of equal size
- Equal-depth (frequency) partitioning:
  - It divides the range into *N* intervals, each containing approximately same number of samples
- V-optimal:
  - It considers all histogram types for a given number of buckets and chooses the one with the least variance.
- MaxDiff:
  - After sorting the data to be approximated, it defines the borders of the buckets at points where the adjacent values have the maximum difference
    - Example: split 1,1,4,5,5,7,9, 14,16,18, 27,30,30,32 to three buckets

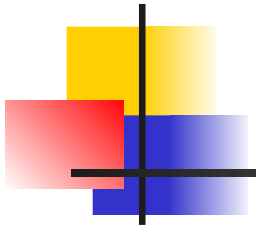
MaxDiff 27-18 and 14-9

Clustering



- Partitions data set into clusters, and models it by one representative from each cluster
- Can be very effective if data is clustered but not if data is “smeared”
- There are many choices of clustering definitions and clustering algorithms, more later!

Hierarchical Reduction

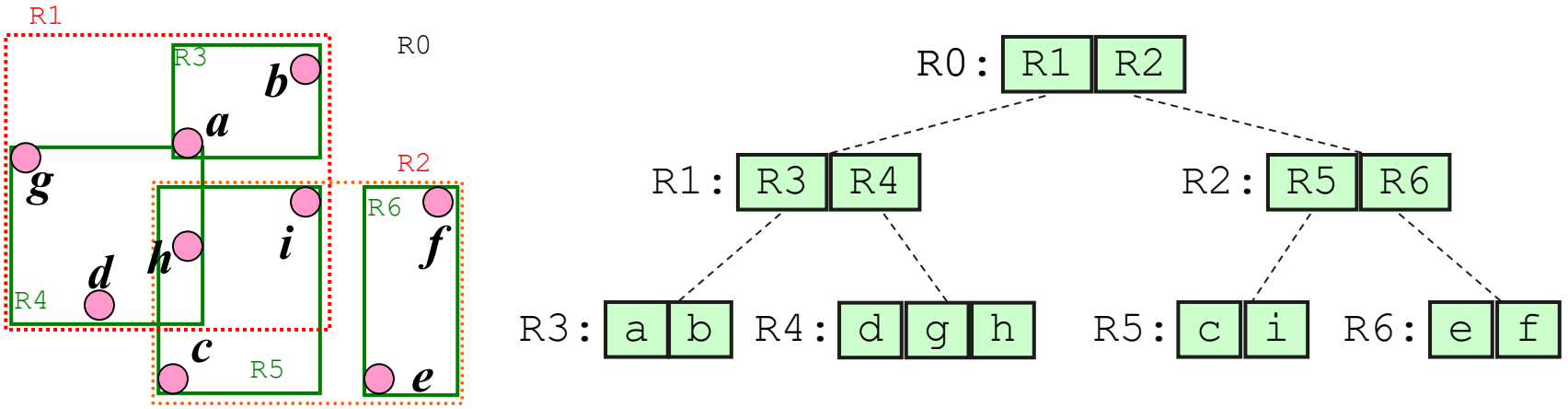


- Use multi-resolution structure with different degrees of reduction
- Hierarchical clustering is often performed but tends to define partitions of data sets rather than "clusters"

- Hierarchical aggregation
  - An index tree hierarchically divides a data set into partitions by value range of some attributes
  - Each partition can be considered as a bucket
  - Thus an index tree with aggregates stored at each node is a hierarchical histogram

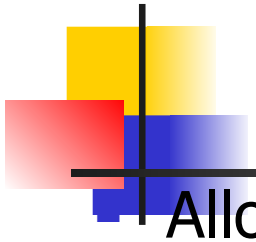
Multidimensional Index Structures can be used for data reduction

Example: an R-tree



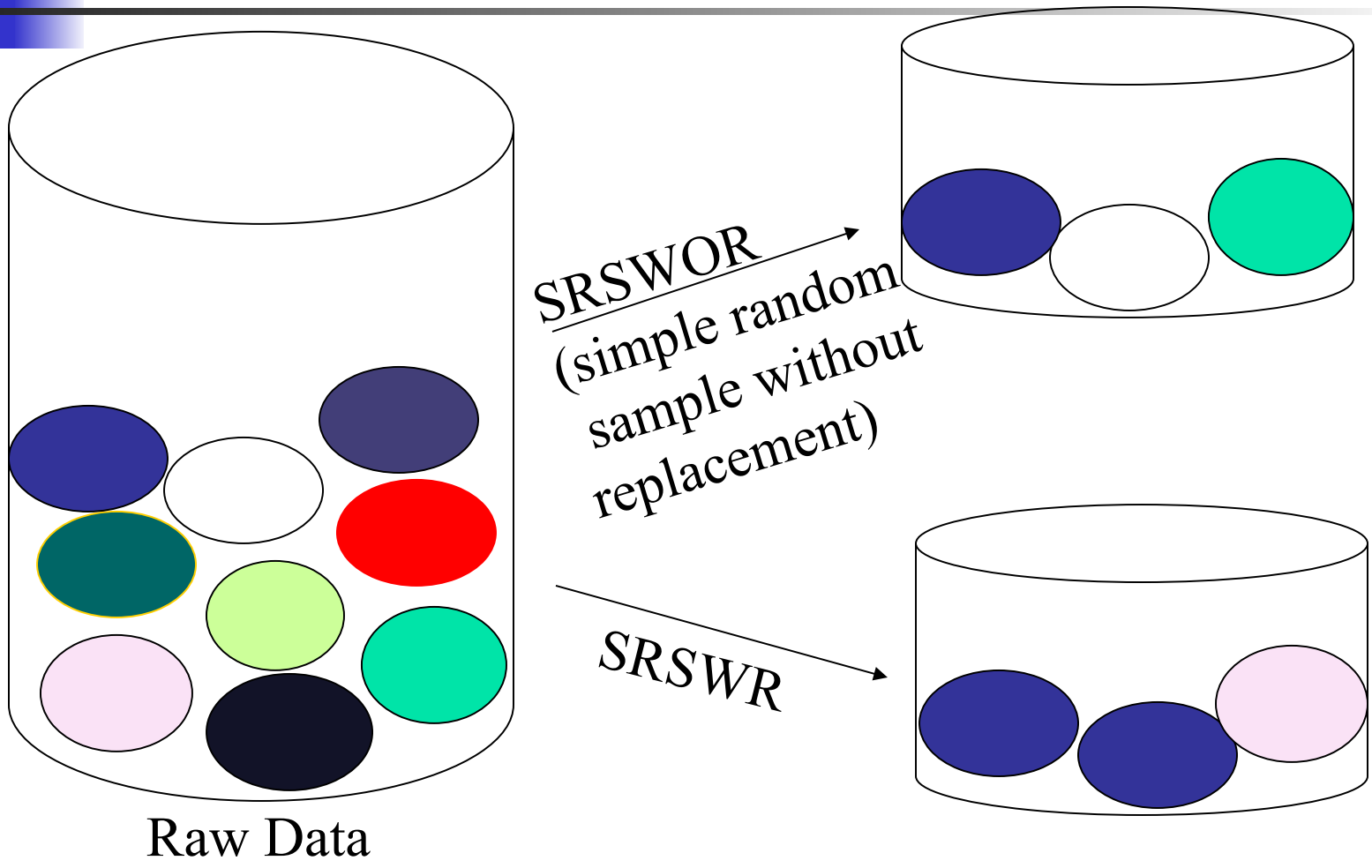
- Each level of the tree can be used to define a milti-dimensional equi-depth histogram
- E.g., R3,R4,R5,R6 define multidimensional buckets which approximate the points

Sampling



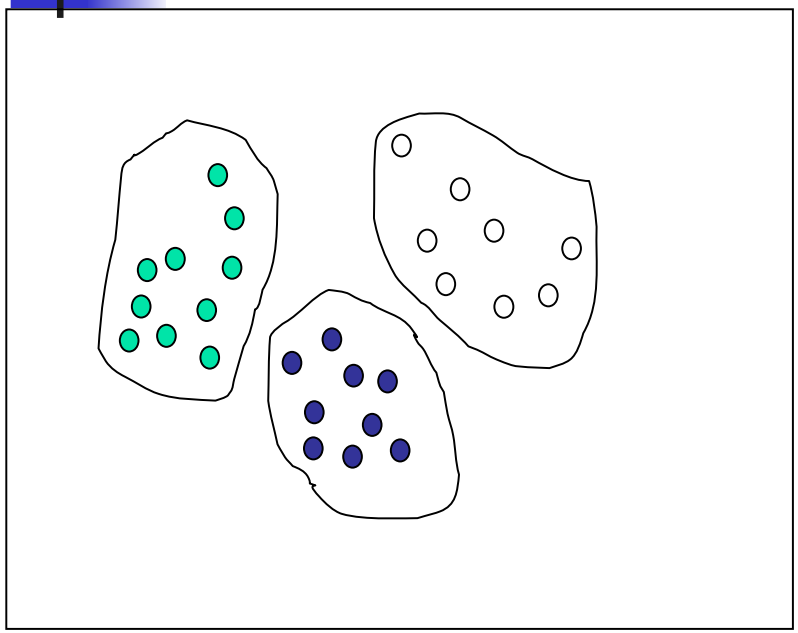
- Allow a mining algorithm to run in complexity that is potentially sub-linear to the size of the data
- Choose a representative subset of the data
  - Simple random sampling may have very poor performance in the presence of skew
- Develop adaptive sampling methods
  - Stratified sampling:
    - Approximate the percentage of each class (or subpopulation of interest) in the overall database
    - Used in conjunction with skewed data
- Sampling may not reduce database I/Os (page at a time).

Sampling

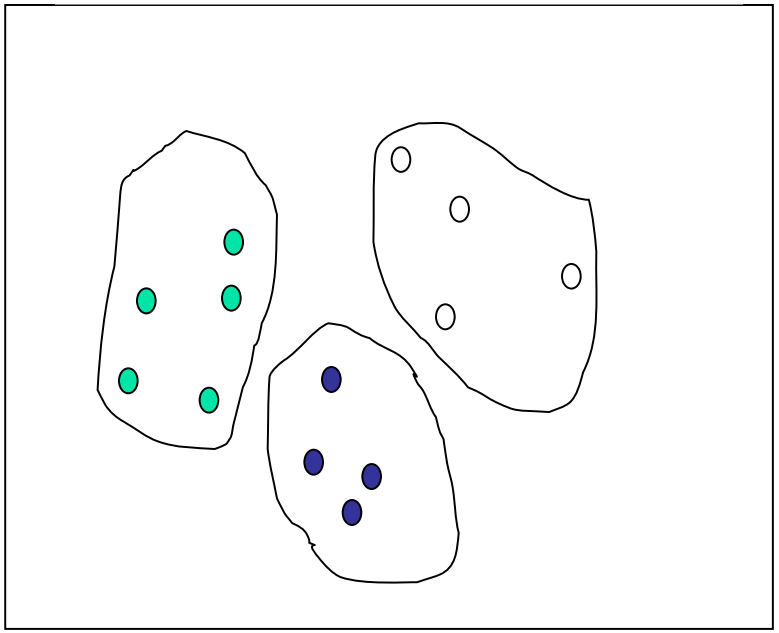

SRSWOR
(simple random sample without replacement)
SRSWR
Raw Data

Sampling

Raw Data

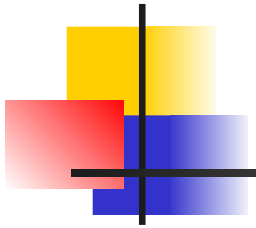


Cluster/Stratified Sample



- The number of samples drawn from each cluster/stratum is analogous to its size
- Thus, the samples represent better the data and outliers are avoided

Summary



- Data preparation is a big issue for both warehousing and mining
- Data preparation includes
  - Data cleaning and data integration
  - Data reduction and feature selection
  - Discretization
- A lot a methods have been developed but still an active area of research