An Introduction

SPATIAL DATABASES AND GEOGRAPHIC INFORMATION SYSTEMS

SPATIAL DATA

SPATIAL DATA

Data related to space

1D: time, time intervals (scheduling)

2D: VLSI design, plane geography

3D: Earth, the brain, the universe

4D: space-time …

SPATIAL AND NON-SPATIAL DATA

Non-spatial:

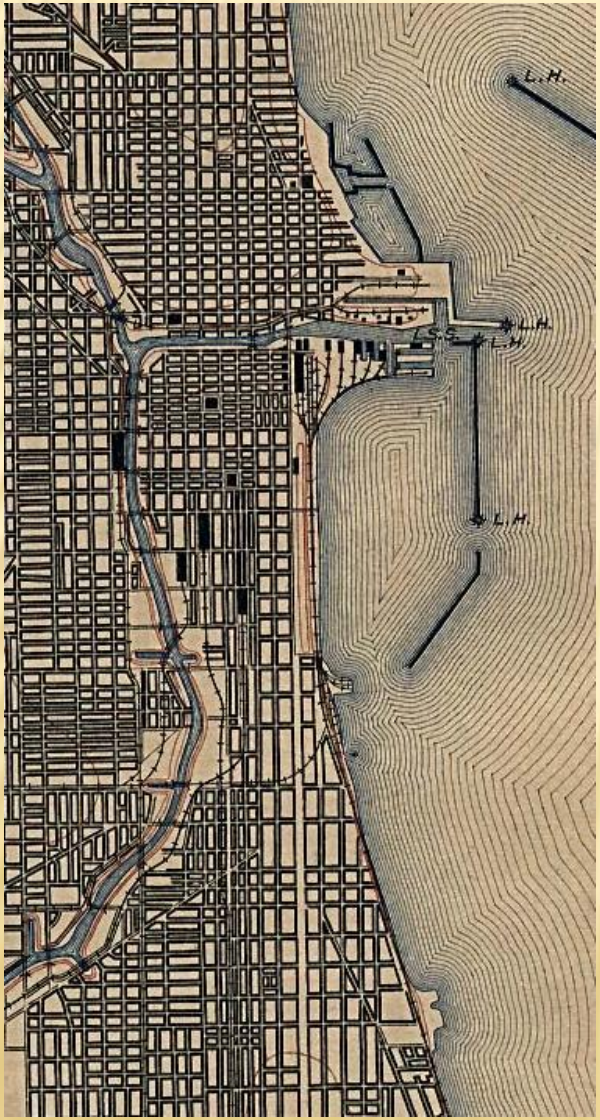
- SSN, telephone number, email, name

Spatial:

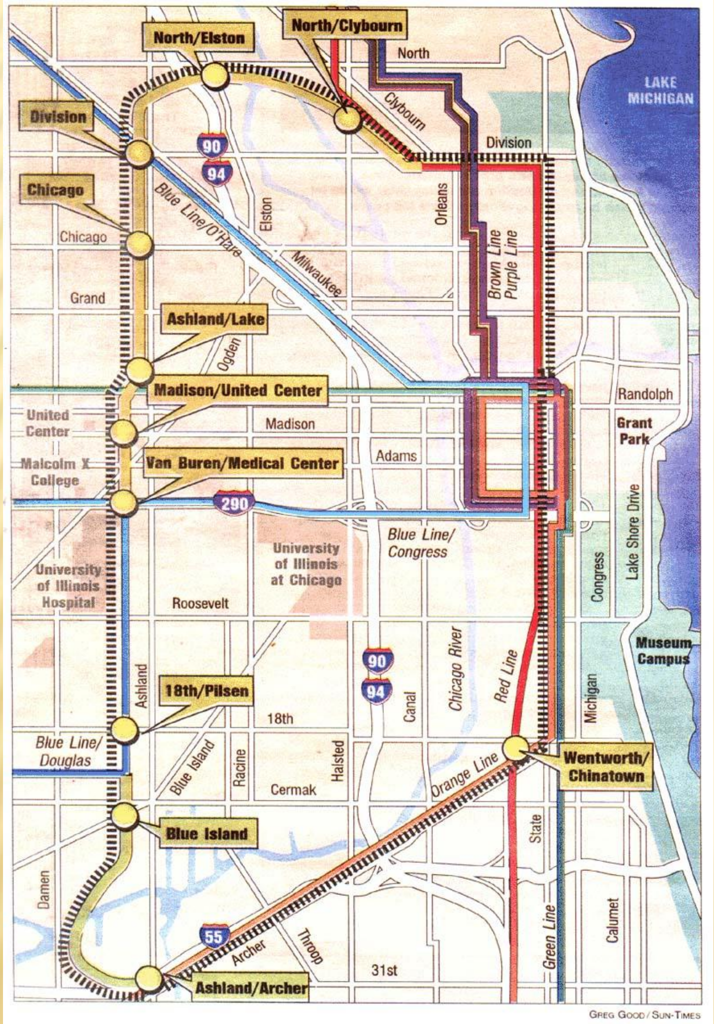
- satellite images
- census data
- Climate data
- maps
- Medical Imaging

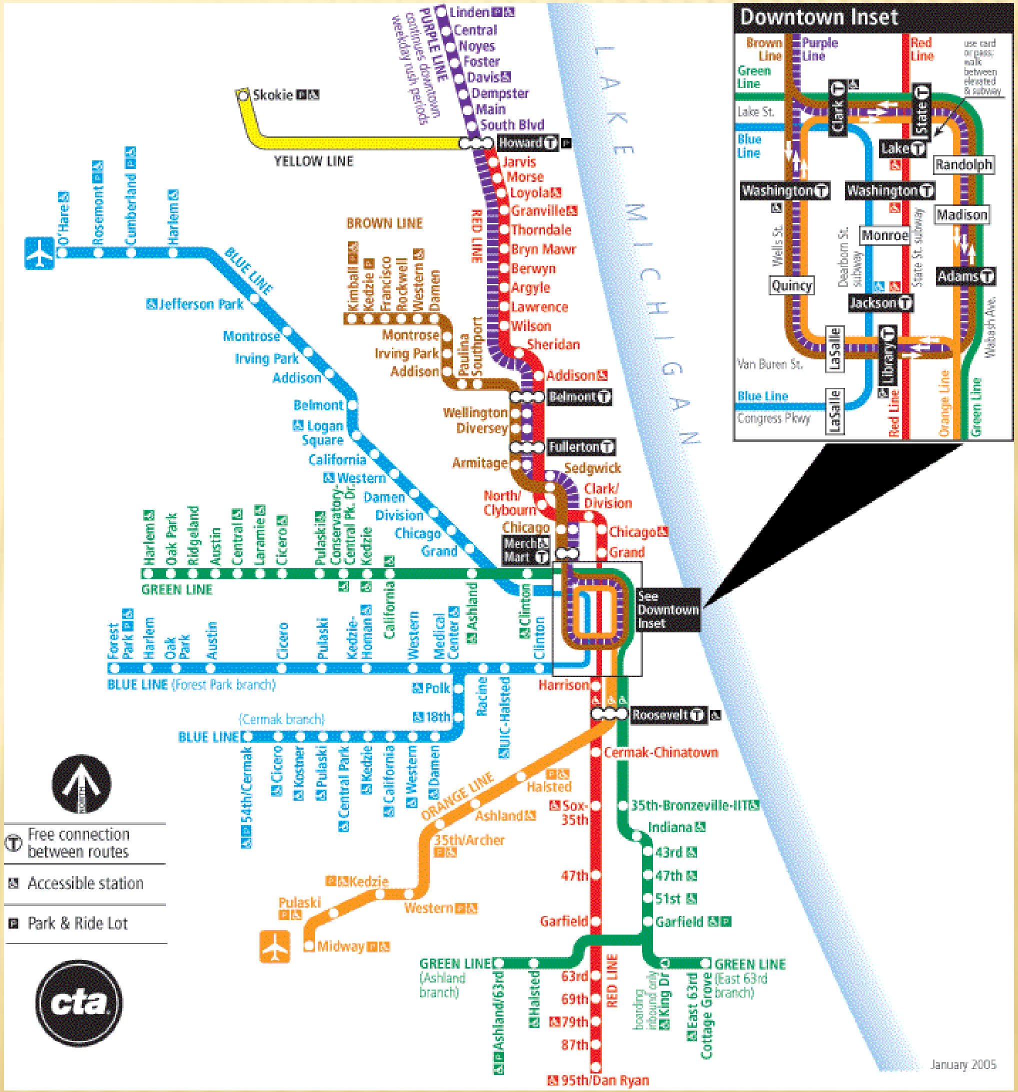
MAPS

MAPS



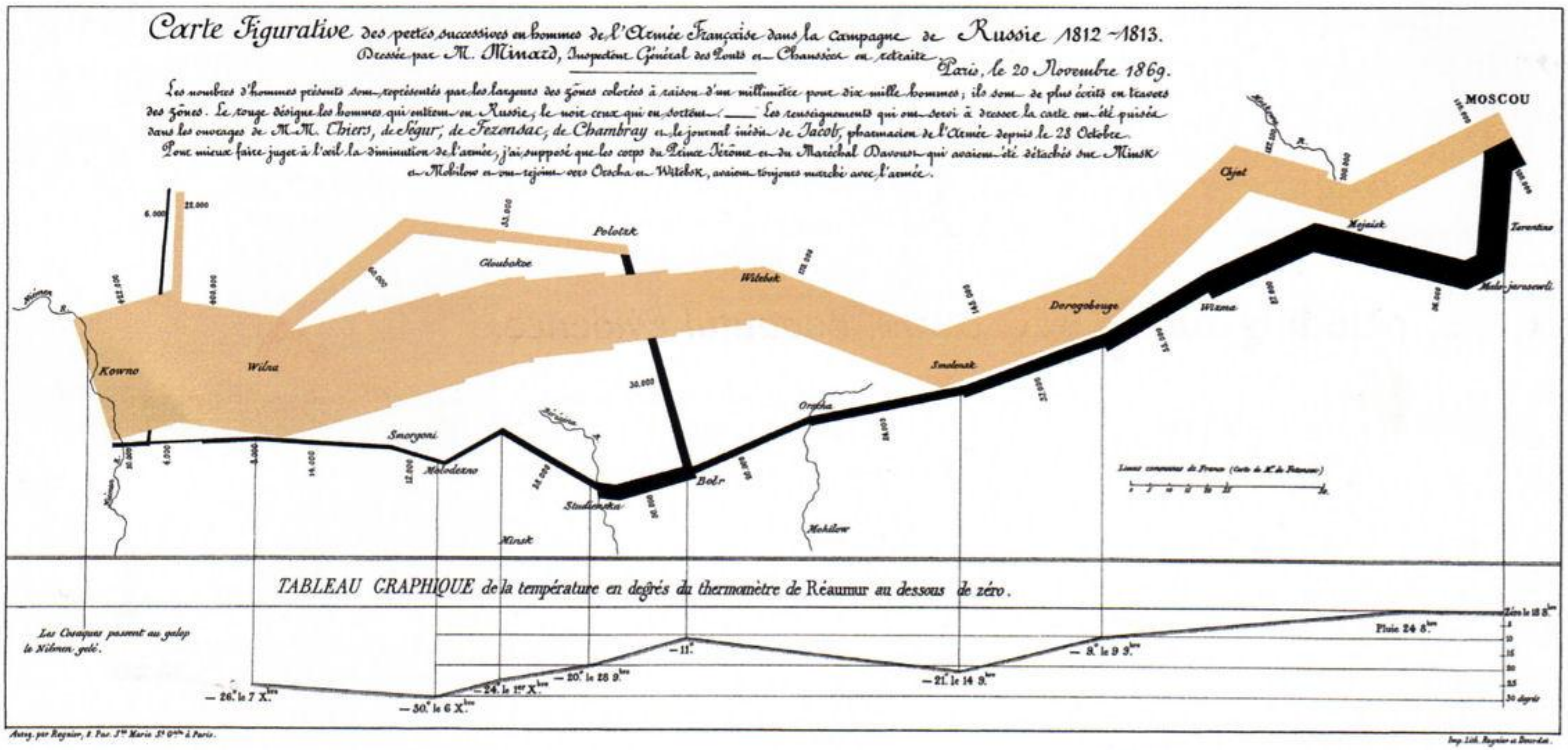
DIAGRAMS





VISUALIZATION

VISUALIZATIONS



This map drawn by Charles Joseph Minard portrays the losses suffered by Napoleon's army in the Russian campaign of 1812. Beginning at the left on the Polish-Russian border near the Niemen, the thick band shows the size of the army (422,000 men) as it invaded Russia. The width of the band indicates the size of the army at each position. In September, the army reached Moscow with 100,000 men. The path of Napoleon's retreat from Moscow in the bitterly cold winter is depicted by the dark lower band, which is tied to temperature and time scales. The remains of the Grande Armée struggled out of Russia with 10,000 men. Minard's graphic tells a rich, coherent story with its multivariate data, far more enlightening than just a single number bouncing along over time. Six variables are plotted: the size of the army, its location on a two-dimensional surface, direction of the army's movement, and temperature on various dates during the retreat from Moscow. It may well be the best statistical graphic ever drawn. Napoleon's March poster $14 postpaid; English/French version $18 postpaid.

GEOGRAPHIC INFORMATION SYSTEMS

GIS

"software to visualize and analyze spatial data using spatial analysis functions such as

- **Search** Thematic search, search by region, (re-)classification
- **Location analysis** Buffer, corridor, overlay
- **Terrain analysis** Slope/aspect, catchment, drainage network
- **Flow analysis** Connectivity, shortest path
- **Distribution** Change detection, proximity, nearest neighbor
- **Spatial analysis/Statistics** Pattern, centrality, autocorrelation, indices of similarity, topology: hole description
- **Measurements** Distance, perimeter, shape, adjacency, direction"

Shekhar, Chawla, Spatial Databases (based on Albrecht)

Examples?

WHY?

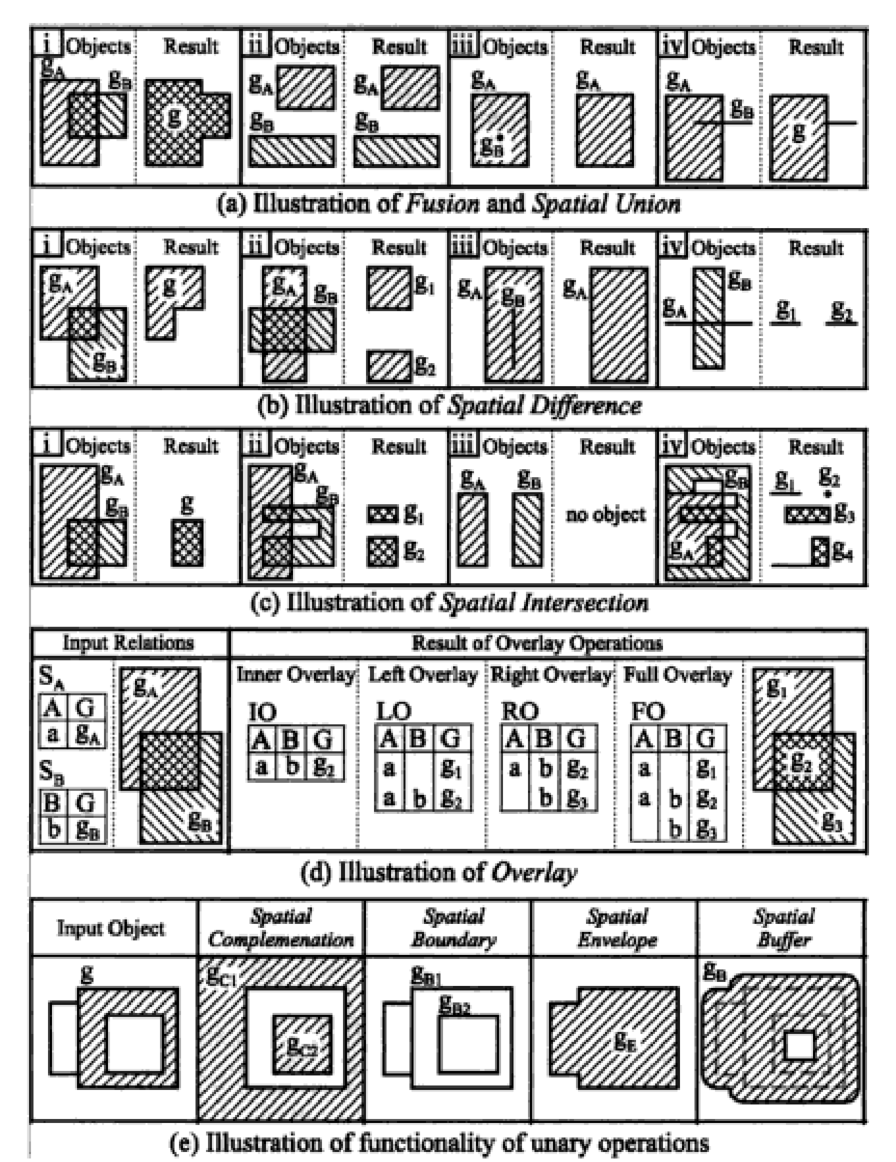
- Mobile phones (nearest coffeeshop)
- Cars (navigation systems)
- Climatology
- Emergency services
- Bus-tracker
- Epidemics

How have you used spatial data?
What software did you use?

GIS LANGUAGE

- Theme
  - a spatial “relation”; e.g. roads, bus routes, countries
- Geographic Objects
  - (description, spatial component)

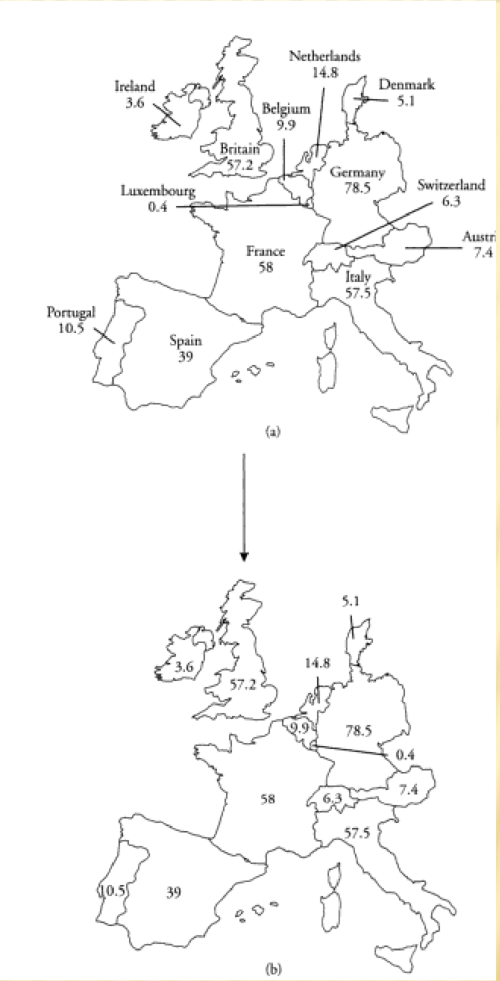
OPERATIONS ON GEOGRAPHIC OBJECTS



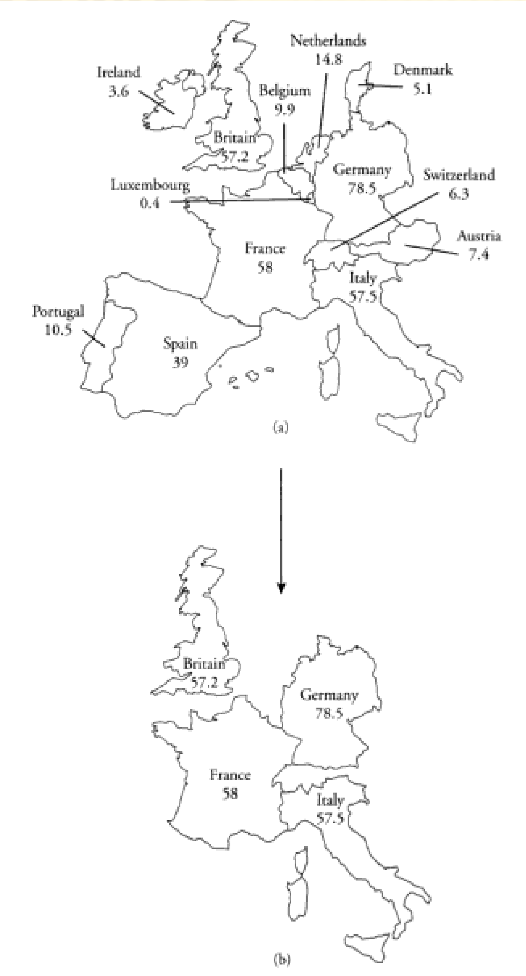
(a) Illustration of *Fusion* and *Spatial Union*
(b) Illustration of *Spatial Difference*
(c) Illustration of *Spatial Intersection*
(d) Illustration of *Overlay*
(e) Illustration of functionality of unary operations

from Spatial Databases: Technologies, Techniques and Trends
by Vassilakopoulos, Papadopoulos

OPERATIONS ON THEMES

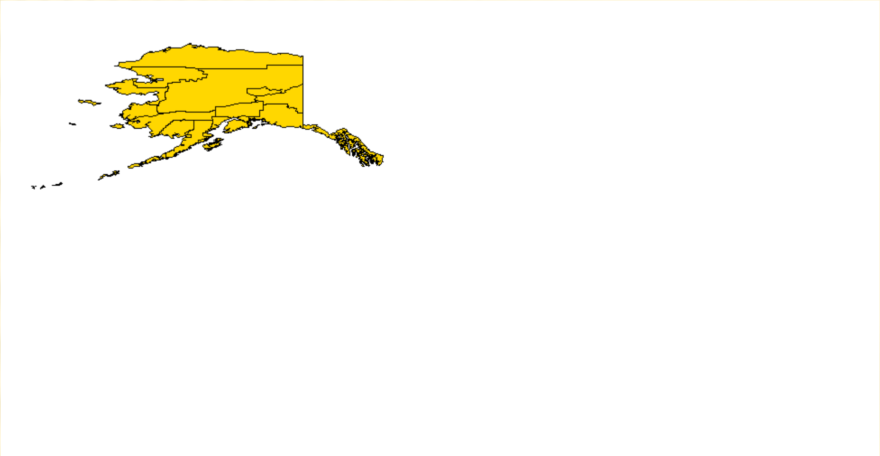


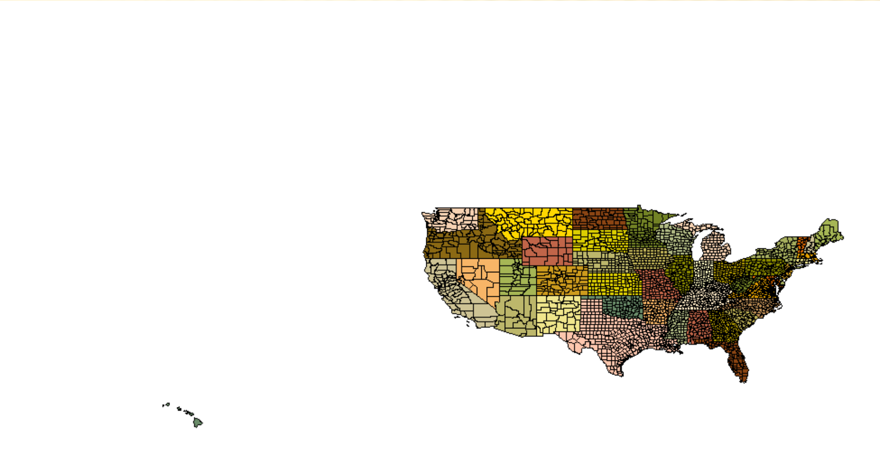
projection



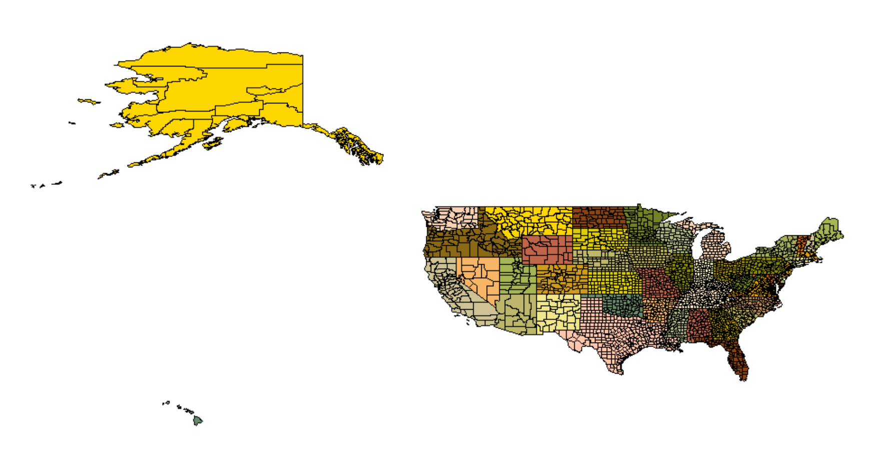
selection

OPERATIONS ON THEMES

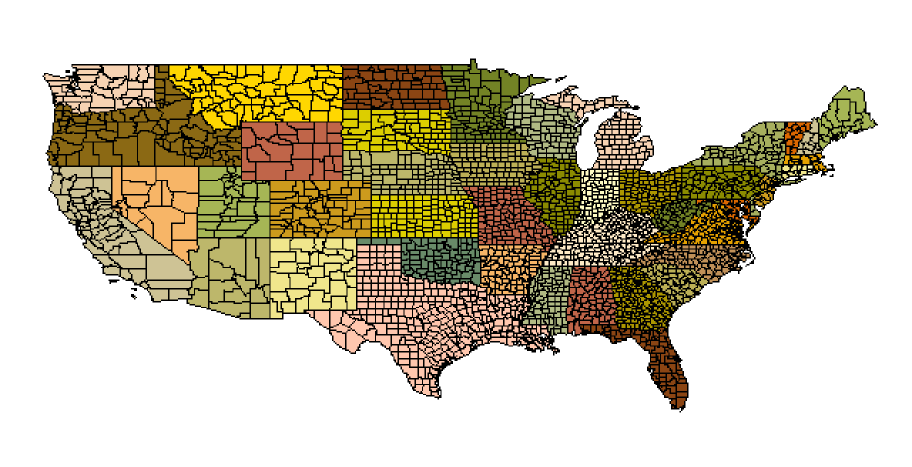




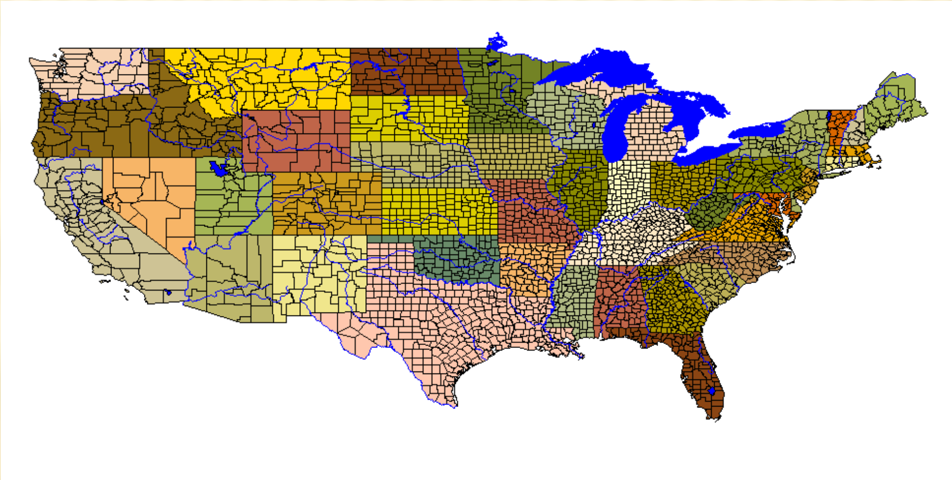
union



OPERATIONS ON THEMES



overlay



OPERATIONS ON THEMES

- geometric selection
  - windowing (only objects in window)
  - point query (objects containing a point)
  - clipping (changes geometry of objects)
- merge

MORE OPERATIONS ON THEMES

- metric operations
  - distance, area
- topological operations
  - adjacent, within, connected
- interpolation/extrapolation
- location
- allocation

GIS AND DATABASES

Four possible approaches to GIS:

- avoid databases
- relational DBMS
- loosely coupled (ArcView, TiGRis)
- integrated (SDBMS) (Oracle Spatial, PostGIS)

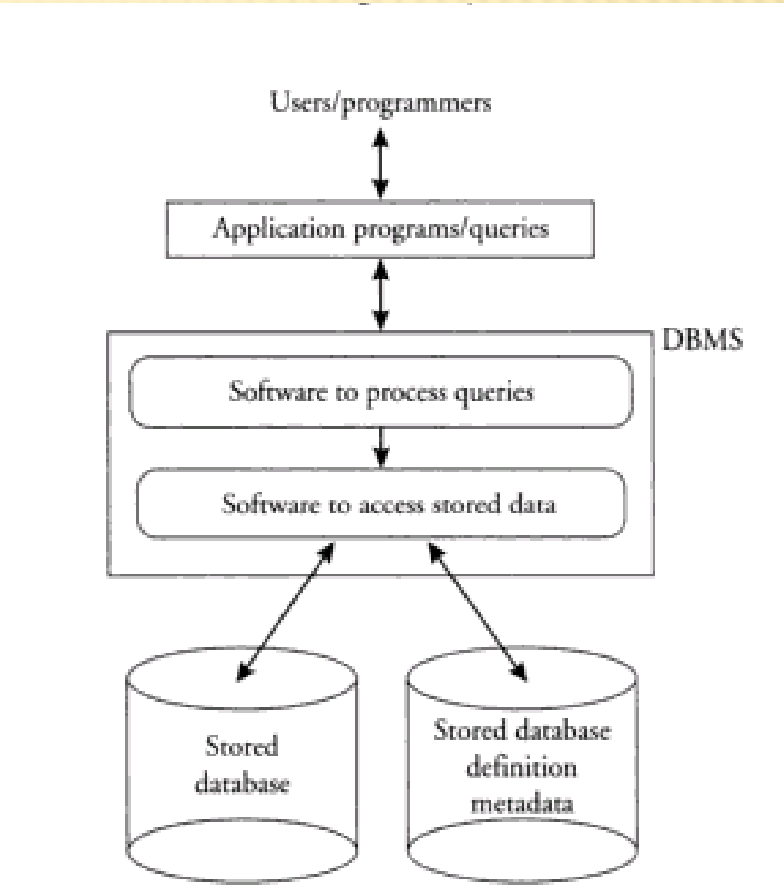
USING A RELATIONAL DBMS (AN EXAMPLE FROM SHEKHAR)

DBMS

- data/metadata
- data independence

physical-level tasks

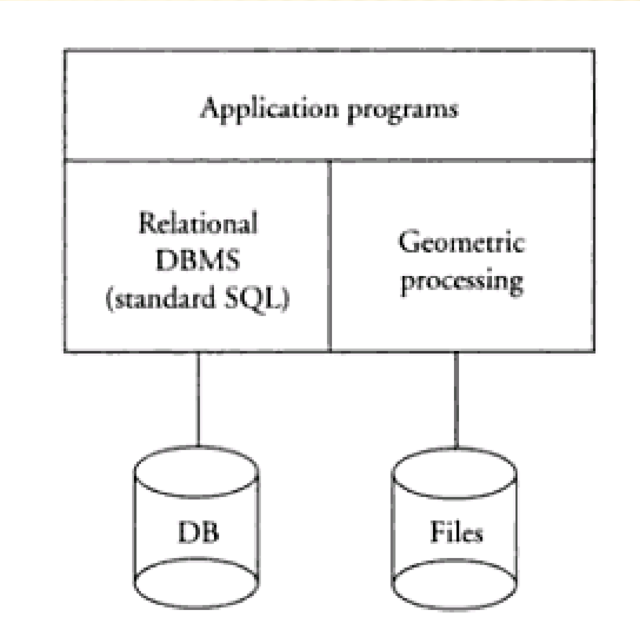
- storage
- access paths
- query processing
- query optimization
- concurrency/recovery



AVOID DATABASES/LOOSELY COUPLED

drawbacks

- heterogeneous data models
- loss of DBMS functionality



SDBMS EXAMPLE (SHEKHAR/CHAWLA)

- Consider a spatial dataset with:
  - County boundary (dashed white line)
  - Census block - name, area, population, boundary (dark line)
  - Water bodies (dark polygons)
  - Satellite Imagery (gray scale pixels)

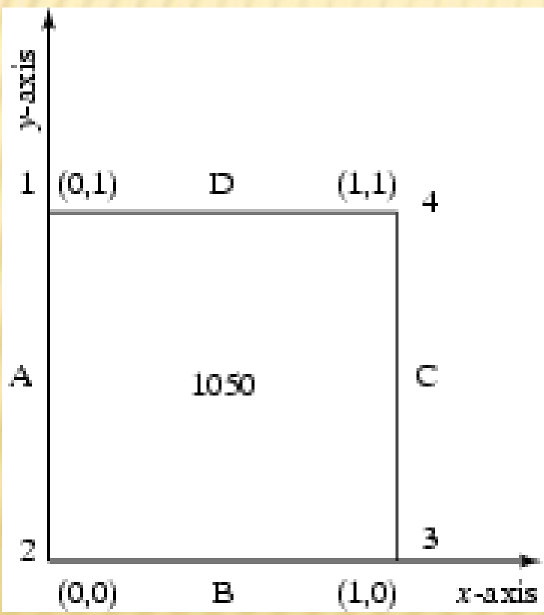
- Storage in a SDBMS table:

```
create table census_blocks (
   name              string,
   area              float,
   population        number,
   boundary          polygon );
```

Fig 1.2

MODELING SPATIAL DATA IN TRADITIONAL DBMS (SHEKHAR/CHAWLA)

- A row in the table census_blocks (Figure 1.3)
- Question: Is Polyline datatype supported in DBMS?



Census_blocks

Name	Area	Population	Boundary
1050	1	1839	Polyline((0,0),(0,1),(1,1),(1,0))

Figure 1.3

SPATIAL DATA TYPES AND TRADITIONAL DATABASES (SHEKHAR/CHAWLA)

Traditional relational DBMS

- Support simple data types, e.g. number, strings, date
- Modeling Spatial data types is tedious, but can be done: polygons are modeled using three tables: polygon, edge, points

MAPPING "CENSUS_TABLE" INTO A RELATIONAL DATABASE (SHEKHAR/CHAWLA)

Census_blocks

Name	Area	Population	boundary-ID
340	1	1839	1050

Polygon

boundary-ID	edge-name
1050	A
1050	B
1050	C
1050	D

Edge

edge-name	endpoint
A	1
A	2
B	2
B	3
C	3
C	4
D	4
D	1

Point

endpoint	x-coor	y-coor
1	0	1
2	0	0
3	1	0
4	1	1

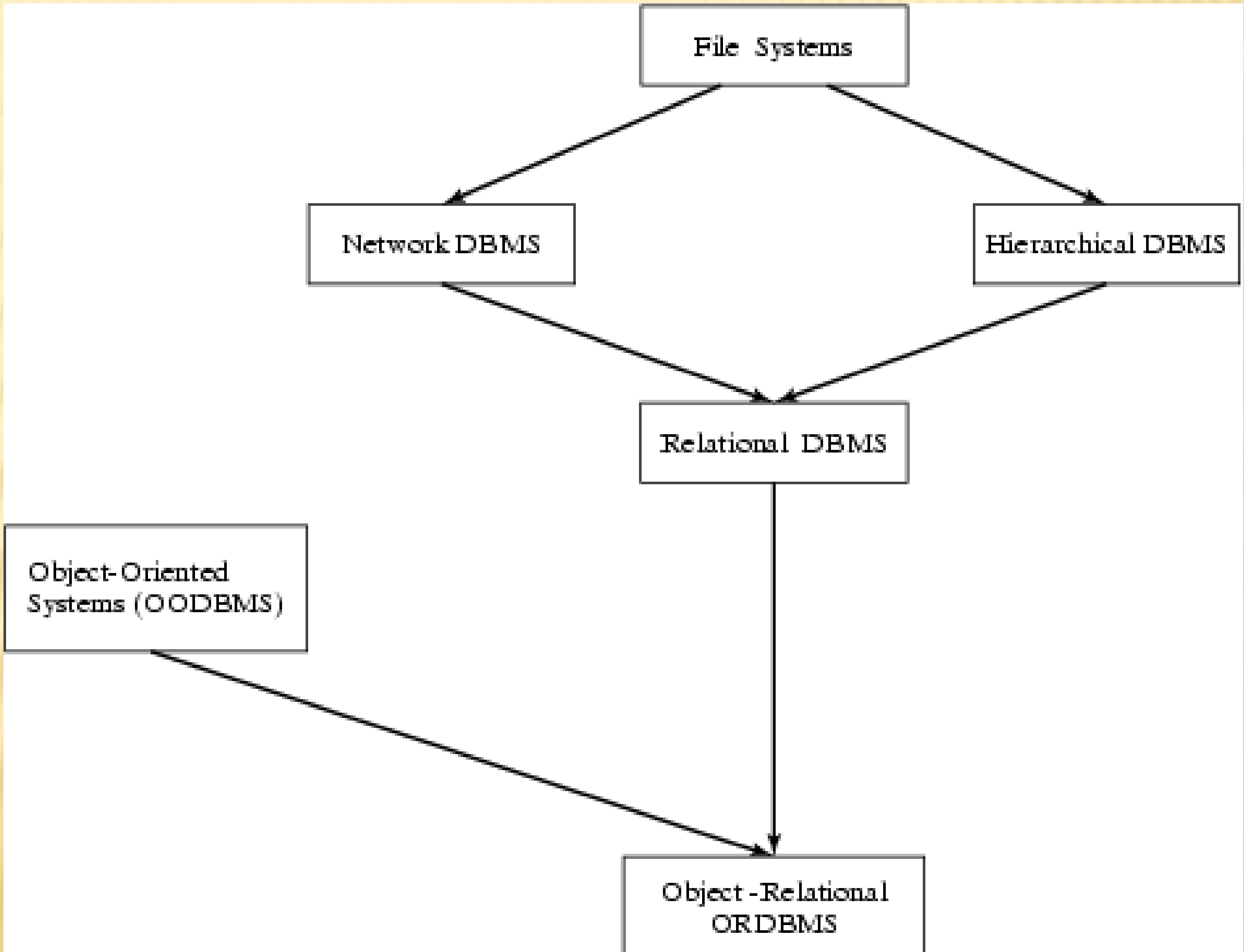
Problems ?

SPATIAL DATA TYPES AND TRADITIONAL DATABASES (SHEKHAR/CHAWLA)

Drawbacks:

- A simple unit square represented as 16 rows across 3 tables
- Simple spatial operators, e.g. area(), require joining tables, some cannot be expressed easily or at all
- violates data independence
- lacks flexibility
- Tedious and computationally inefficient

EVOLUTION OF DBMS TECHNOLOGY



SPATIAL DATA TYPES AND POST-RELATIONAL DATABASES

- Post-relational DBMS
  - Support user defined abstract data types
  - Spatial data types (e.g. polygon) can be added

- Choice of post-relational DBMS
  - Object oriented (OO) DBMS
  - Object relational (OR) DBMS

- A spatial database is a collection of spatial data types, operators, indices, processing strategies, etc. and can work with many post-relational DBMS as well as programming languages like Java, Visual Basic etc.

WHAT IS AN SDBMS

SPATIAL DATABASE SYSTEM

According to Güting (1994):

- database system
- supports spatial data types
- supports spatial operations

  indexing, joins, …

WHAT IS A SDBMS ?

- A SDBMS is a software module that
  - can work with an underlying DBMS
  - supports spatial data models, spatial abstract data types (ADTs) and a query language from which these ADTs are callable
  - supports spatial indexing, efficient algorithms for processing spatial operations, and domain specific rules for query optimization

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

HOW IS A SDBMS DIFFERENT FROM A GIS ?

- GIS is a software to visualize and analyze spatial data using spatial analysis functions

- GIS uses SDBMS
  - to store, search, query, share large spatial data sets

HOW IS A SDBMS DIFFERENT FROM A GIS ?

- SDBMS focuses on
  - Efficient storage, querying, sharing of large spatial datasets
  - Provides simpler set based query operations
  - Example operations: search by region, overlay, nearest neighbor, distance, adjacency, perimeter etc.
  - Uses spatial indices and query optimization to speed up queries over large spatial datasets.
- SDBMS may be used by applications other than GIS
  - Astronomy, Genomics, Multimedia information systems, ...
- Will one use a GIS or a SDBM to answer the following:
  - How many neighboring countries does USA have?
  - Which country has highest number of neighbors?

EVOLUTION OF ACRONYM "GIS"

- Geographic Information Systems (1980s)
- Geographic Information Science (1990s)
- Geographic Information Services (2000s)

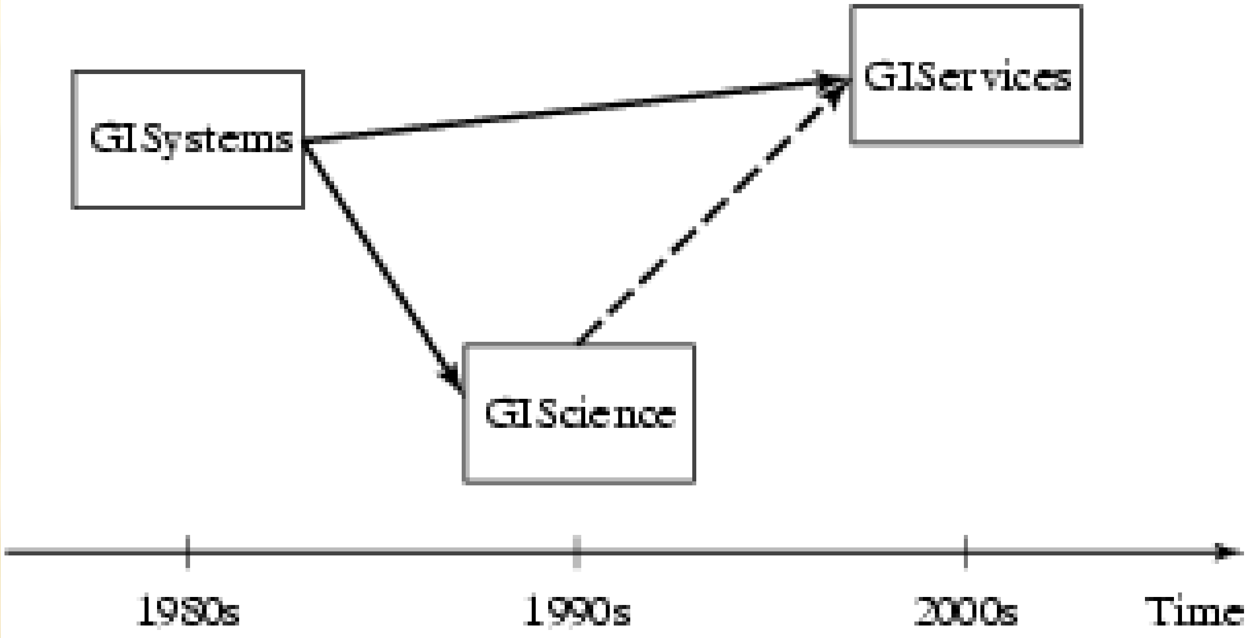


Fig 1.1

THREE MEANINGS OF THE ACRONYM GIS

- Geographic Information Services
  - Web-sites and service centers for casual users, e.g. travelers
  - Example: Service (e.g. AAA, mapquest, google) for route planning

- Geographic Information Systems
  - Software for professional users, e.g. cartographers
  - Example: ESRI Arc/View software

- Geographic Information Science
  - Concepts, frameworks, theories to formalize use and development of geographic information systems and services
  - Example: design spatial data types and operations for querying

A TOUR OF SDBMS

COMPONENTS OF A SDBMS

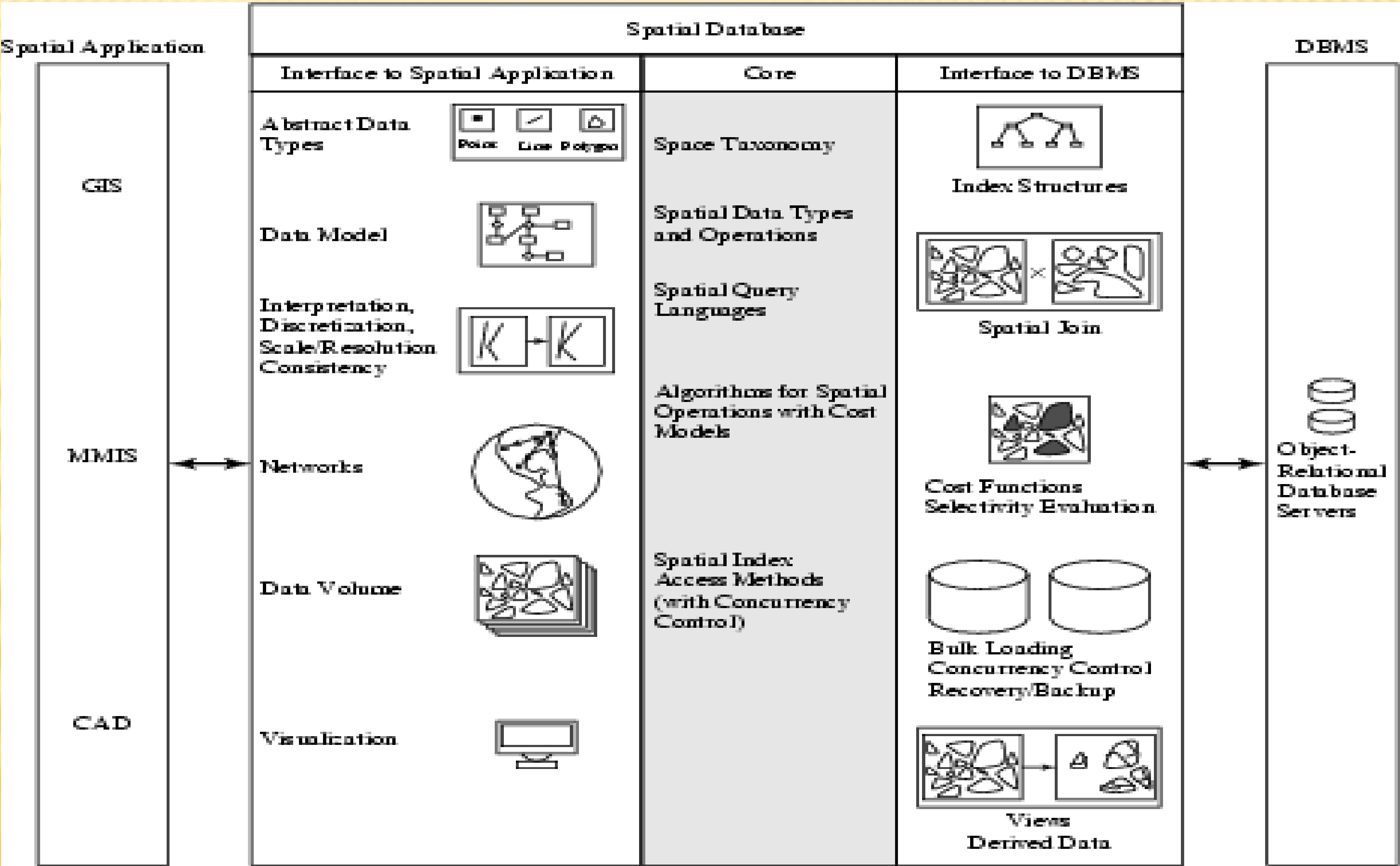
Recall: a SDBMS is a software module that

- can work with an underlying DBMS
- supports spatial data models, spatial ADTs and a query language from which these ADTs are callable
- supports spatial indexing, algorithms for processing spatial operations, and domain specific rules for query optimization

Components include

- spatial data model, query language, query processing, file organization and indices, query optimization, etc.

THREE LAYER ARCHITECTURE



SPATIAL TAXONOMY, DATA MODELS

Spatial Taxonomy:

- multitude of descriptions available to organize space.
- Topology models homeomorphic relationships, e.g. overlap
- Euclidean space models distance and direction in a plane
- Graphs models connectivity, Shortest-Path

Spatial data models

- rules to identify identifiable objects and properties of space
- *Object model* help manage identifiable things, e.g. mountains, cities, land-parcels etc.
- *Field model* help manage continuous and amorphous phenomenon, e.g. wetlands, satellite imagery, snowfall etc.

SPATIAL QUERY LANGUAGE

Spatial data types, e.g. point, linestring, polygon, …
Spatial operations, e.g. overlap, distance, nearest neighbor, …
Callable from a query language (e.g. SQL3) of underlying DBMS

```
SELECT S.name
FROM   Senator S
WHERE  S.district.Area() > 300
```

Standards

- SQL3 (a.k.a. SQL 1999) is a standard for query languages
- OGIS is a standard for spatial data types and operators
- Both standards enjoy wide support in industry

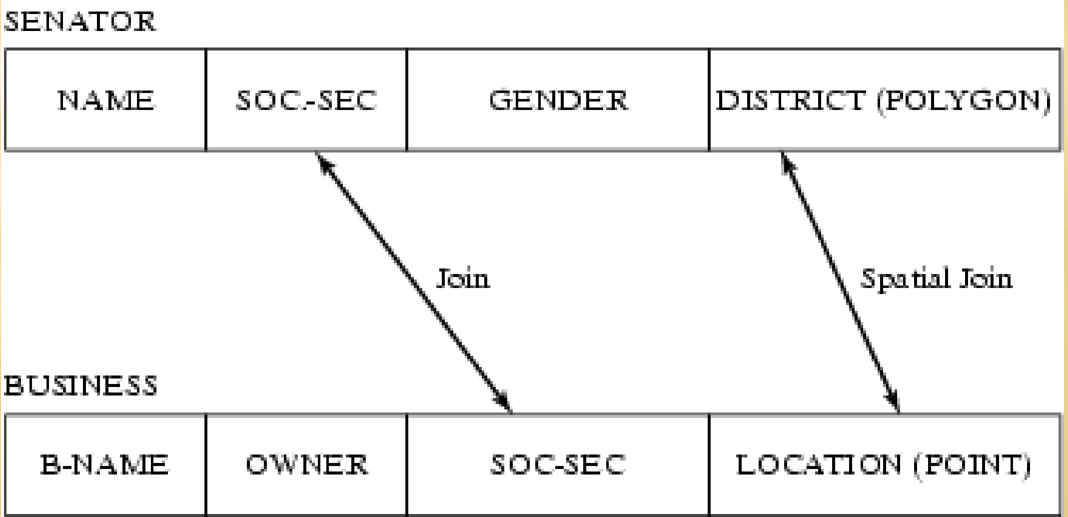
MULTI-SCAN QUERY EXAMPLE

Non-Spatial Join example

```
SELECT S.name
FROM Senator S, Business B
WHERE S.soc-sec = B.soc-sec AND S.gender = 'Female'
```

Spatial join example

```
SELECT S.name
FROM  Senator S, Business B
WHERE S.district.Area() > 300 AND
         Within(B.location, S.district)
```

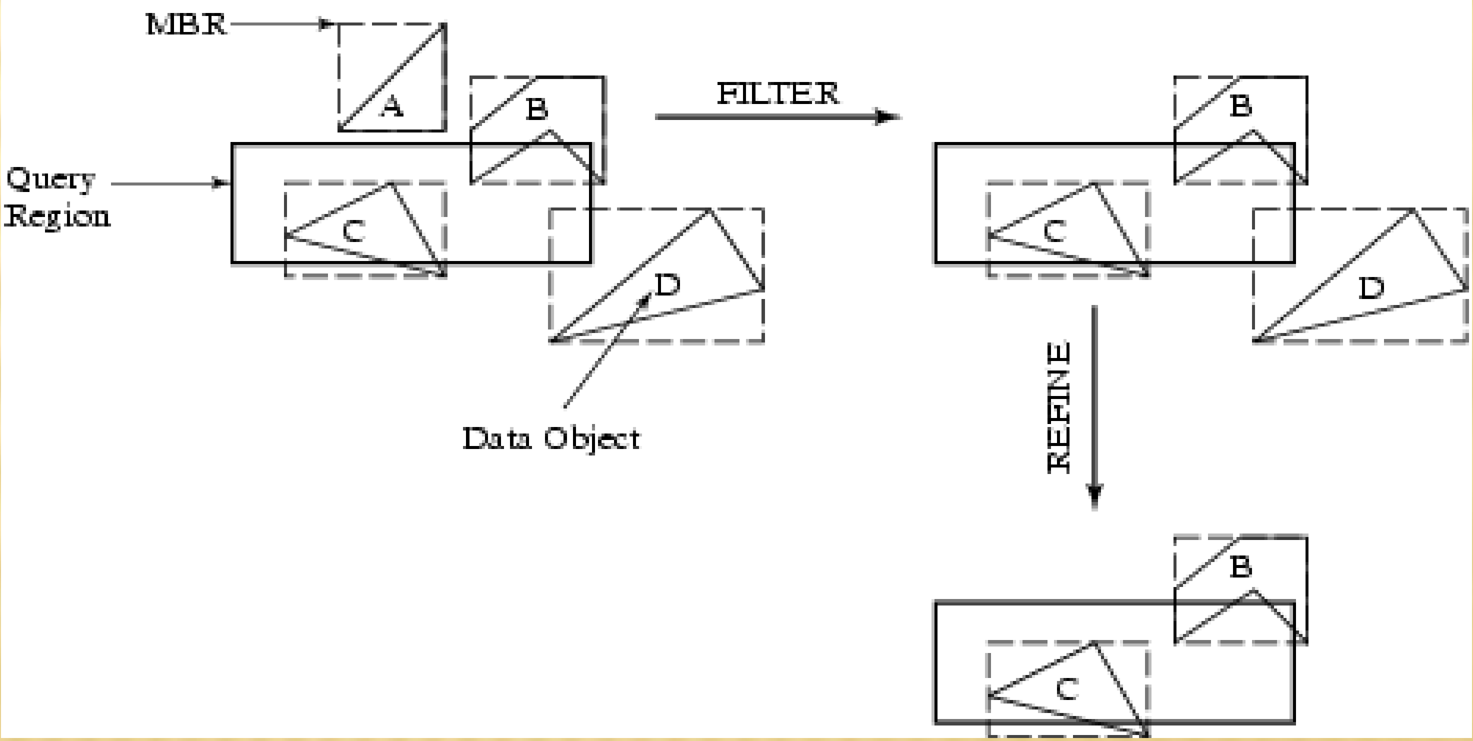


QUERY PROCESSING

Efficient algorithms to answer spatial queries

Common Strategy - filter and refine

- Filter Step:Query Region overlaps with MBRs of B,C and D
- Refine Step: Query Region overlaps with B and C



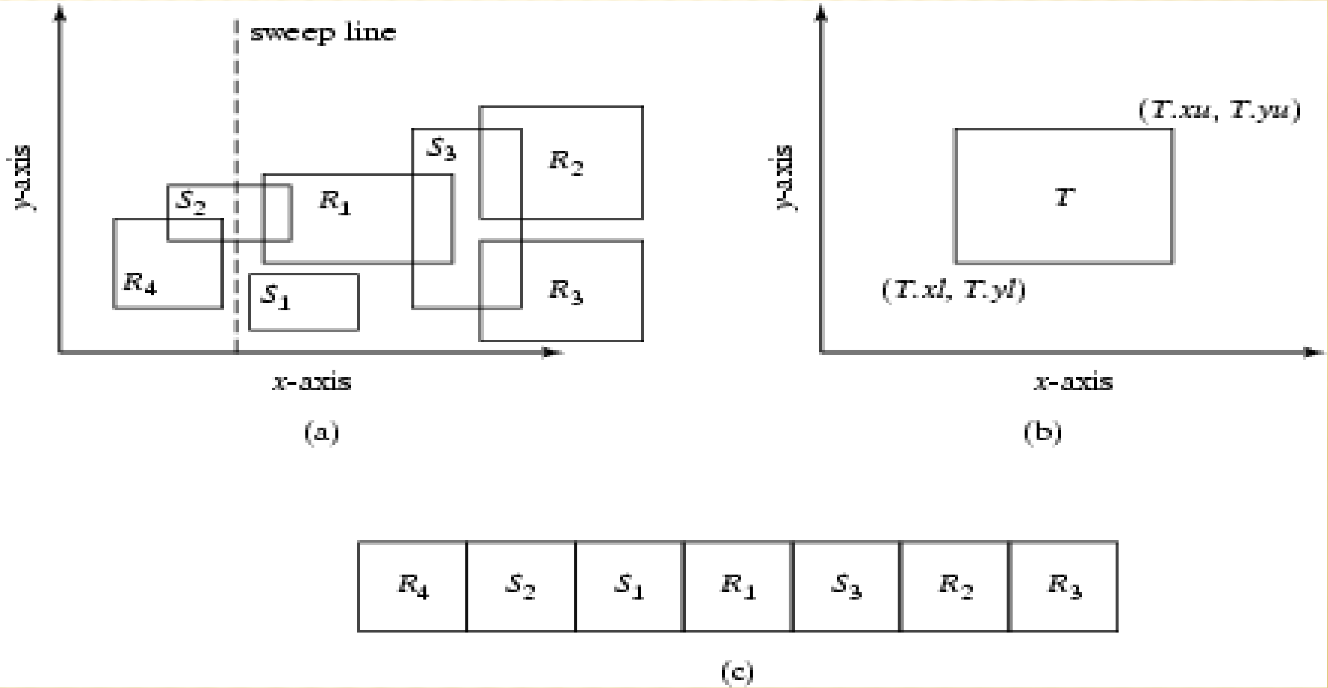
EXAMPLE: QUERY PROCESSING OF JOIN

Task: find all intersections between R and S rectangles as shown in (a)

1) sort rectangles by left x-value as in (c)

process using a line sweep:

1) find first rectangle, T
2) find first rectangle U from other set with T.xu < U.xl
   while doing so check rectangles from other set for overlap

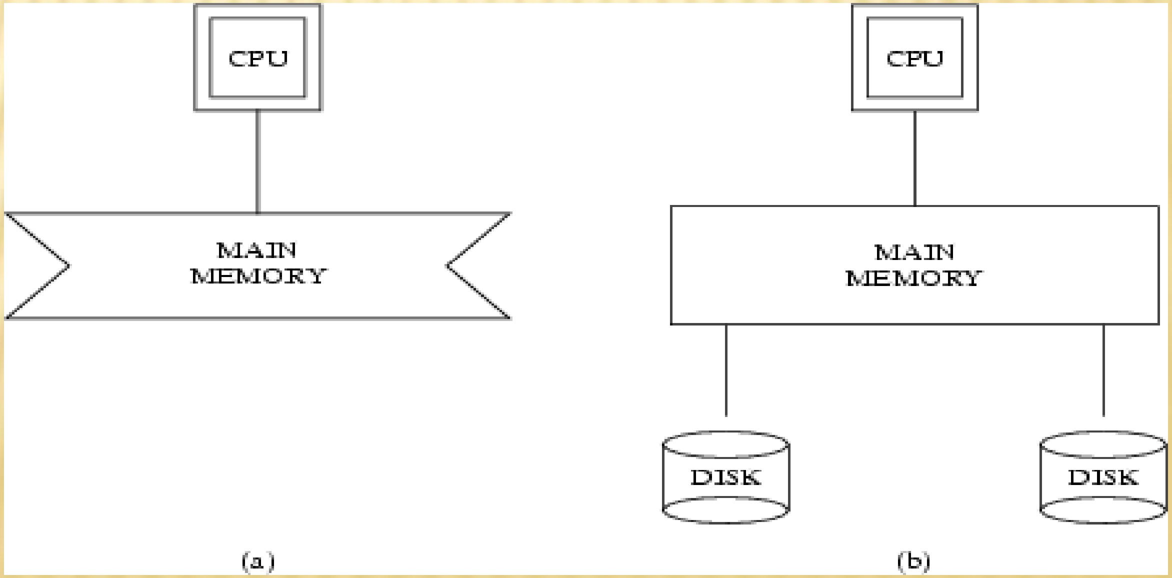


FILE ORGANIZATION AND INDICES

A difference between GIS and SDBMS assumptions

- GIS algorithms: dataset is loaded in main memory (a)
  access time: nanoseconds
- SDBMS: dataset is on secondary storage e.g disk (b)
  access time: split seconds

SDBMS uses space filling curves and spatial indices
to efficiently search disk resident large spatial datasets



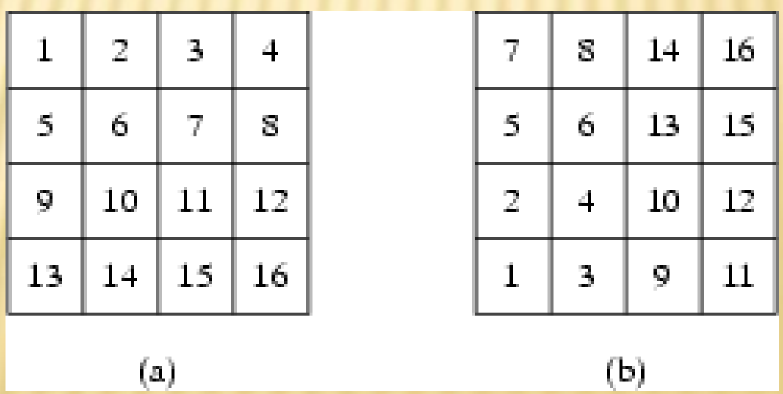
ORGANIZING SPATIAL DATA WITH SPACE FILLING CURVES

Issue:

- Sorting is not naturally defined on spatial data
- Many efficient search methods are based on sorting datasets

Space filling curves

- Impose an ordering on the locations in a multi-dimensional space
- Examples: row-order (a), z-order (b)
- Allow use of traditional efficient search methods on spatial data



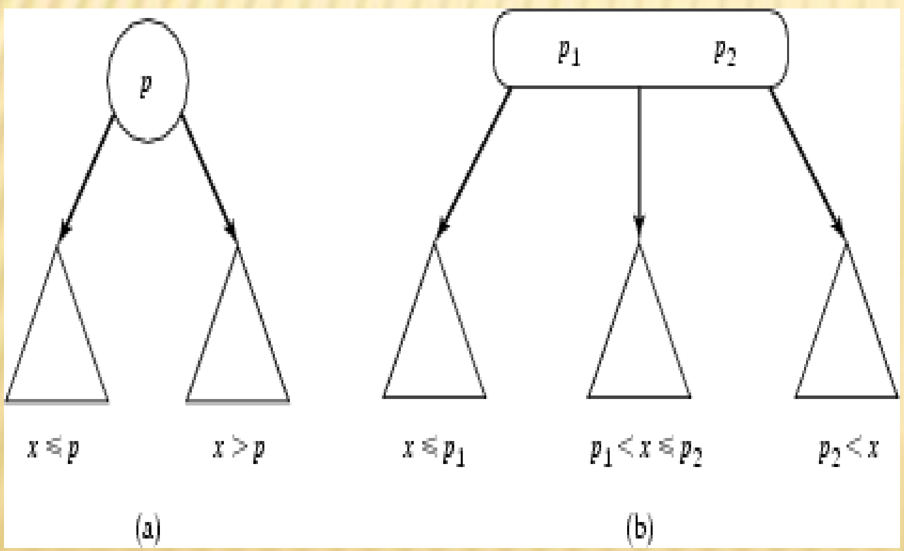
SPATIAL INDEXING: SEARCH DATA-STRUCTURES

Reminder: B-tree

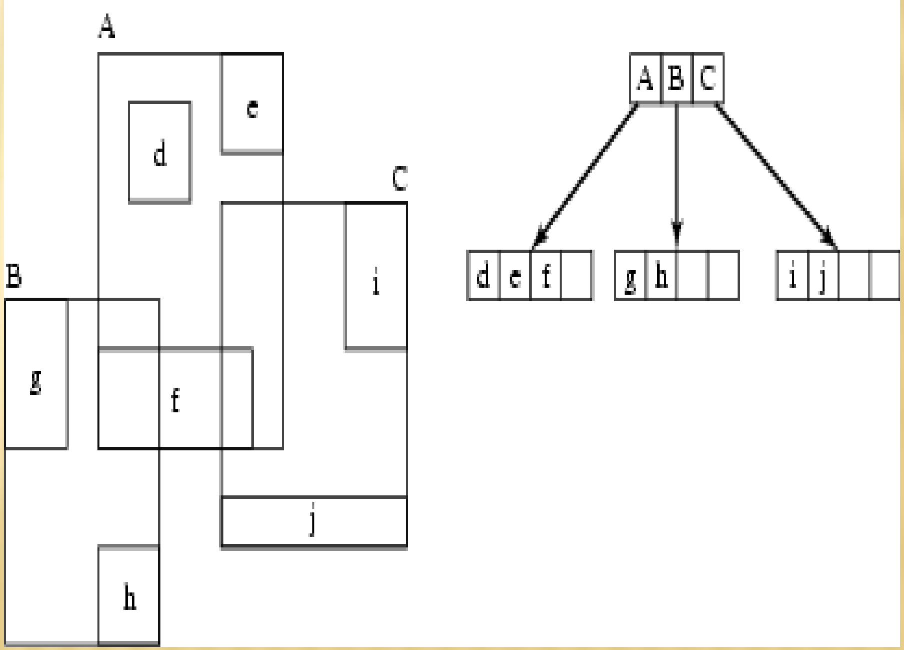
- hierarchical collection of ranges of linear keys, e.g. numbers
- B-tree index is used for efficient search of traditional data

Spatial Index:

- use B-tree with space filling curve on spatial data
- R-tree has better search performance
- R-tree is a hierarchical collection of rectangles



binary versus B-tree



R- tree

QUERY OPTIMIZATION

Query Optimization

- A spatial operation can be processed using different strategies
- Computation cost of each strategy depends on many parameters
- Query optimization is the process of
  - ordering operations in a query and
  - selecting efficient strategy for each operation
  - based on the details of a given dataset

Example 1)

```
SELECT S.name
FROM Senator S, Business B
WHERE S.soc-sec = B.soc-sec AND S.gender = 'Female'
```

- which condition should be evaluated first:

  (S.soc-sec = B.soc-sec) or (S.gender = 'Female')

- do we use index for S.gender = 'Female', for S.soc-sec = B.soc-sec ?

QUERY OPTIMIZATION

Query Optimization

- A spatial operation can be processed using different strategies
- Computation cost of each strategy depends on many parameters
- Query optimization is the process of
  - ordering operations in a query and
  - selecting efficient strategy for each operation
  - based on the details of a given dataset

Example 2)

```
SELECT S.name
FROM Senator S, Business B, District D
WHERE S.soc-sec = B.soc-sec AND
      S.district.Area() > 300;
```

which condition should be evaluated first:

(S.soc-sec = B.soc-sec) or (S.district.Area() > 300)

DATA MINING

Analysis of spatial data is of many types

- Deductive Querying, e.g. searching, sorting, overlays
- Inductive Mining, e.g. statistics, correlation, clustering,classification, …

Data mining is a systematic and semi-automated search for interesting non-trivial patterns in large spatial databases

Example applications include

- Infer land-use classification from satellite imagery
- Identify cancer clusters and geographic factors with high correlation
- Identify crime hotspots to assign police patrols and social workers

SUMMARY

- SDBMS is valuable to many important applications
- SDBMS is a software module
  - works with an underlying DBMS
  - provides spatial ADTs 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 and optimization
  - spatial data mining
- SDBMS is used to store, query and share spatial data for GIS as well as other applications

CLASS OUTLINE

Week	Topic
1	Introduction to Spatial Databases
	ER modeling, pictograms
2	Representation of Geometric Data
3-4	Logical Models and Query Languages
	Spatial Networks
5	Constraint Data Model
6	Midterm
7	Computational Geometry
8	Spatial Access Methods
9	Query Processing
10	Commercial Systems Overview