Chapter 1: Introduction to Spatial Databases

1.1 Overview
1.2 Application domains
1.3 Compare a SDBMS with a GIS
1.4 Categories of users
1.5 An example of an SDBMS application
1.6 A stroll though a spatial database
  1.6.1 Data models,
  1.6.2 Query language,
  1.6.3 Query processing,
  1.6.4 File organization and indices,
  1.6.5 Query optimization,
  1.6.6 Data mining
**Value of SDBMS**

- Traditional (non-spatial) database management systems provide:
  - Persistence across failures
  - Allows concurrent access to data
  - Scalability to search queries on very large datasets which do not fit inside main memories of computers
  - Efficient for non-spatial queries, but not for spatial queries

- Non-spatial queries:
  - List the names of all bookstore with more than ten thousand titles.
  - List the names of ten customers, in terms of sales, in the year 2001

- Spatial queries:
  - List the names of all bookstores with ten miles of Minneapolis
  - List all customers who live in Tennessee and its adjoining states
Value of SDBMS – Spatial Data Examples

- Examples of non-spatial data
  - Names, phone numbers, email addresses of people

- Examples of Spatial data
  - Census Data
  - NASA satellites imagery - terabytes of data per day
  - Weather and Climate Data
  - Rivers, Farms, ecological impact
  - Medical Imaging

- Exercise: Identify spatial and non-spatial data items in
  - A phone book
  - A cookbook with recipes
Value of SDBMS – Users, Application Domains

Many important application domains have spatial data and queries. Some examples follow:

- **Army Field Commander**: Has there been any significant enemy troop movement since last night?
- **Insurance Risk Manager**: Which homes are most likely to be affected in the next great flood on the Mississippi?
- **Medical Doctor**: Based on this patient's MRI, have we treated somebody with a similar condition?
- **Molecular Biologist**: Is the topology of the amino acid biosynthesis gene in the genome found in any other sequence feature map in the database?
- **Astronomer**: Find all blue galaxies within 2 arcmin of quasars

Exercise: List two ways you have used spatial data. Which software did you use to manipulate spatial data?
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
**SDBMS Example**

- 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:

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

Figure 1.2
Modeling Spatial Data in Traditional DBMS

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

![Diagram of Census_blocks](image)

Figure 1.3
Spatial Data Types and Traditional Databases

- Traditional relational DBMS
  - Support simple data types, e.g. number, strings, date
  - Modeling spatial data types is tedious
- Example: Figure 1.4 shows modeling of polygon using numbers
  - Three new tables: polygon, edge, points
    - Note: Polygon is a polyline where last point and first point are same
  - A simple unit square represented as 16 rows across 3 tables
  - Simple spatial operators, e.g. area(), require joining tables
    - Tedious and computationally inefficient
- Question: Name post-relational database management systems which facilitate modeling of spatial data types, e.g. polygon
Mapping “census table” into a Relational Database

<table>
<thead>
<tr>
<th>Census_blocks</th>
<th>Polygon</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Name</strong></td>
<td><strong>boundary-ID</strong></td>
</tr>
<tr>
<td>340</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>boundary-ID</strong></th>
<th><strong>Area</strong></th>
<th><strong>Population</strong></th>
<th><strong>boundary-ID</strong></th>
<th><strong>edge-name</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>1050</td>
<td>1</td>
<td>1839</td>
<td>1050</td>
<td>A</td>
</tr>
<tr>
<td>1050</td>
<td>1</td>
<td>1839</td>
<td>1050</td>
<td>B</td>
</tr>
<tr>
<td>1050</td>
<td>1</td>
<td>1839</td>
<td>1050</td>
<td>C</td>
</tr>
<tr>
<td>1050</td>
<td>1</td>
<td>1839</td>
<td>1050</td>
<td>D</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>edge-name</strong></th>
<th><strong>endpoint</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1</td>
</tr>
<tr>
<td>A</td>
<td>2</td>
</tr>
<tr>
<td>B</td>
<td>2</td>
</tr>
<tr>
<td>B</td>
<td>3</td>
</tr>
<tr>
<td>C</td>
<td>3</td>
</tr>
<tr>
<td>C</td>
<td>4</td>
</tr>
<tr>
<td>D</td>
<td>4</td>
</tr>
<tr>
<td>D</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Point</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>endpoint</strong></td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>4</td>
</tr>
</tbody>
</table>

Figure 1.4
Evolution of DBMS Technology

Figure 1.5
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.
How is a SDBMS Different from a GIS?

GIS is a 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

GIS uses SDBMS
to store, search, query, share large spatial data sets
How is a SDBMS Different from a GIS?

- SDBMS focusses 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 speedup 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)

Figure 1.1
Three Meanings of the Acronym GIS

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

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

Exercise: Which meaning of the term GIS is closest to the focus of the book titled “Spatial Databases: A Tour”?
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.
  - Figure 1.6 shows these components
  - We discuss each component briefly in chapter 1.6 and in more detail in later chapters.
# Three Layer Architecture

<table>
<thead>
<tr>
<th>Spatial Application</th>
<th>Spatial Database</th>
<th>DBMS</th>
</tr>
</thead>
<tbody>
<tr>
<td>GIS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MMIS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CAD</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

## Spatial Application
- GIS
- MMIS
- CAD

## Spatial Database
- Interface to Spatial Application
  - Abstract Data Types
  - Data Model
  - Interpretation, Discretization, Scale/Resolution Consistency
- Core
  - Space Taxonomy
  - Spatial Data Types and Operations
  - Spatial Query Languages
  - Algorithms for Spatial Operations with Cost Models
- Interface to DBMS
  - Spatial Index Access Methods (with Concurrency Control)
  - Spatial Join
  - Cost Functions Selectivity Evaluation
  - Bulk Loading Concurrency Control Recovery/Backup
  - Views Derived Data

## Figure 1.6
- Visualization
- Interpretation, Discretization, Scale/Resolution Consistency
- Networks
- Data Volume
- Visualization

- Spatial Application
- GIS
- MMIS
- CAD

- Spatial Database
- DBMS

- Object-Relational Database Servers
**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 (networks)** 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.

- More details in chapter 2
**Spatial Query Language**

- **Spatial query language**
  - Multitude of descriptions available to organize space.
  - 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
    
    ```sql
    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
  - More details in chapters 2 and 3
**Spatial Query Language**

- **Spatial join example**
  
  ```sql
  SELECT S.name FROM Senator S, Business B
  WHERE S.district.Area() > 300 AND Within(B.location, S.district)
  ```

- **Non-spatial join example**
  
  ```sql
  SELECT S.name FROM Senator S, Business B
  WHERE S.soc-sec = B.soc-sec AND S.gender = 'Female'
  ```

![Figure 1.7](image.png)
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

Figure 1.8
**Query Processing of Join Queries**

- Example - Determining pairs of intersecting rectangles
  - (a): Two sets R and S of rectangles, (b): A rectangle with 2 opposite corners marked, (c): Rectangles sorted by smallest X coordinate value
  - Plane sweep filter identifies 5 pairs out of 12 for refinement step
  - Details of plane sweep algorithm on page 15

![Diagram](image)

**Figure 1.9**

<table>
<thead>
<tr>
<th>R₄</th>
<th>S₂</th>
<th>S₁</th>
<th>R₁</th>
<th>S₃</th>
<th>R₂</th>
<th>R₃</th>
</tr>
</thead>
</table>

(Tₓᵤ, Tᵧᵤ)  
(Tₓₗ, Tᵧₗ)
A difference between GIS and SDBMS assumptions

- GIS algorithms: dataset is loaded in main memory (Fig. 1.10(a))
- SDBMS: dataset is on secondary storage e.g. disk (Fig. 1.10(b))
- SDBMS uses space filling curves and spatial indices
  - to efficiently search disk resident large spatial datasets

**Figure 1.10**
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 (Fig. 1.11(a), z-order (Fig 1.11(b))
- Allow use of traditional efficient search methods on spatial data

Figure 1.11
Spatial Indexing: Search Data-Structures

- Choice for spatial indexing:
  - B-tree is a hierarchical collection of ranges of linear keys, e.g. numbers
  - B-tree index is used for efficient search of traditional data
  - B-tree can be used with space filling curve on spatial data
  - R-tree provides better search performance yet!
  - R-tree is a hierarchical collection of rectangles
  - More details in chapter 4

Figure 1.12: B-tree

Figure 1.13: 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 Query:**
  ```
  SELECT S.name 
  FROM Senator S, Business B 
  WHERE S.soc-sec = B.soc-sec AND S.gender = 'Female'
  ```

- **Optimization decision examples**
  - Process (S.gender = ‘Female’) before (S.soc-sec = B.soc-sec )
  - Do not use index for processing (S.gender = ‘Female’)

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