Introduction to Spatial Database Systems

by Cyrus Shahabi

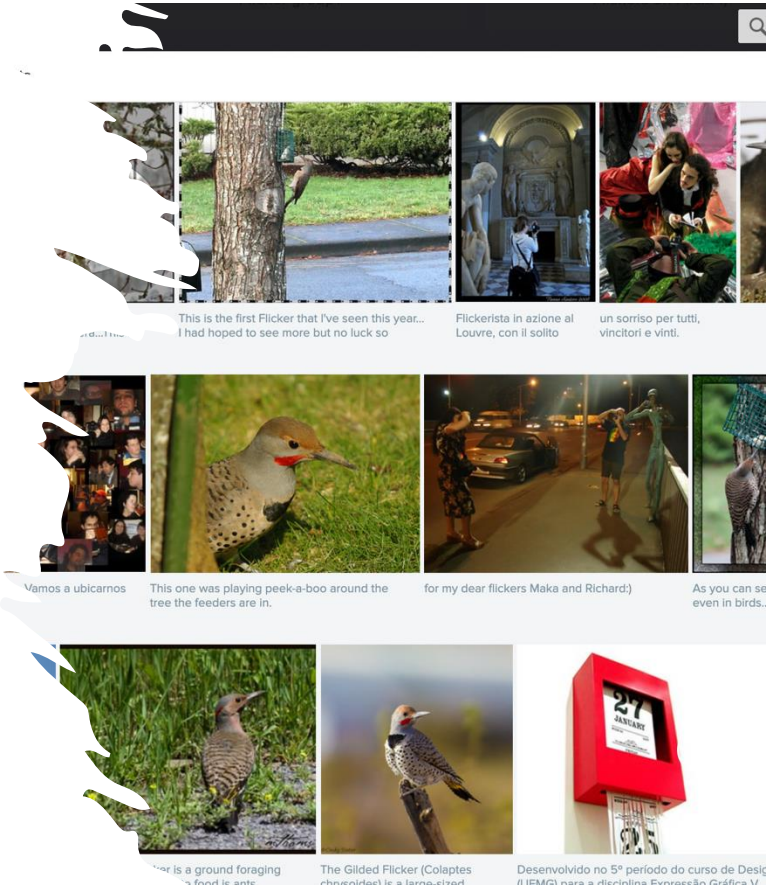
Spatial Databases: A Tour, Shashi Shekhar and Sanjay Chawla

* Hart Hartmut Guting's VLDB Journal v3, n4, October 1994

What is Data Management?

How do you manage your photos?

- Most cellphones take nice photos
  - Taking 3 photos a day will give you ~1,000 photos a year
  - Taking a 5-day vacation would give you 200 photos
- Ways to managing photos
  - Leave them on the phone?
  - Organize them into folders?
  - Upload them to some cloud services?
- Which method is the best?



Considerations for Managing Photos

Find photos by time

Find photos by subjects

Find photos by locations

Searching must be fast!

Photos need to be secure

Available resources

Data Management (Oracle)

- Data management is the practice of collecting, keeping, and using data securely, efficiently, and cost-effectively.
- help people, organizations, and connected things
  - optimize the use of data within the bounds of policy and regulation
  - (use data to) make decisions and take actions that maximize the benefit to the organization

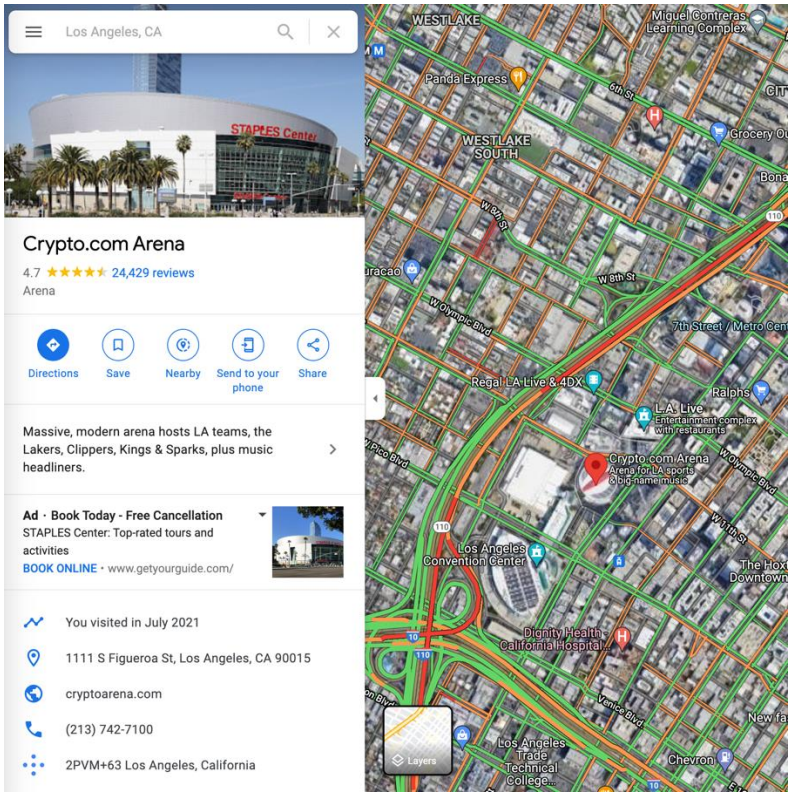
https://www.oracle.com/database/what-is-data-management/

What are Spatial Data?

Spatial Data

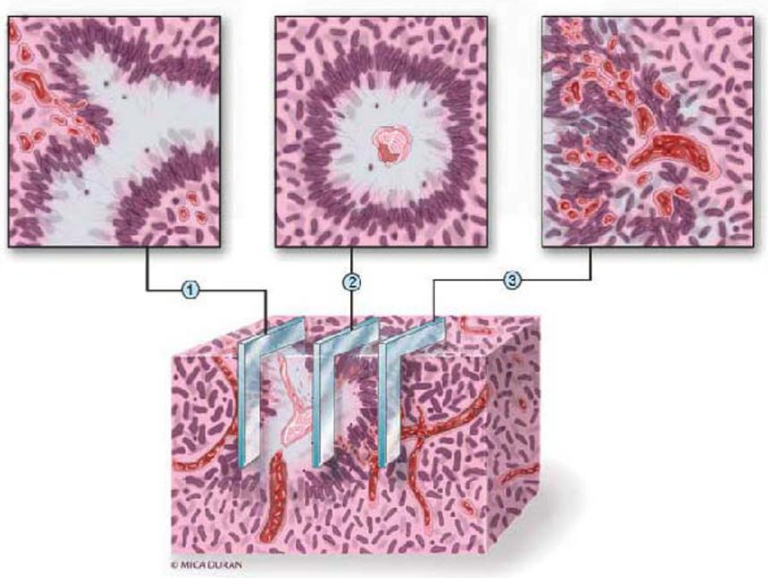
What is Spatial Data?

- Data that can be spatially referenced, e.g.,
  - Time series from fixed-site sensors (e.g., traffic, air quality)
  - Remotely sensed data (e.g., satellite imagery)
  - Geotagged photos and tweets
  - Documents mentioning location entities



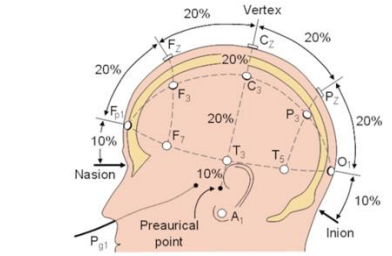
Spatial Data Does Not Have to be Geo Data

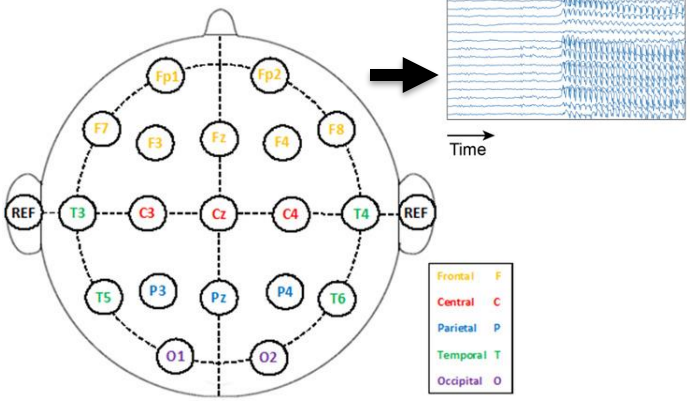
- **Digital Pathology**



Rong, Y., Durden, D. L., Van Meir, E. G., & Brat, D. J. (2006). 'Pseudopalisading'necrosis in glioblastoma: a familiar morphologic feature that links vascular pathology, hypoxia, and angiogenesis. *Journal of Neuropathology & Experimental Neurology*, *65*(6), 529-539.

- **EEG Data**

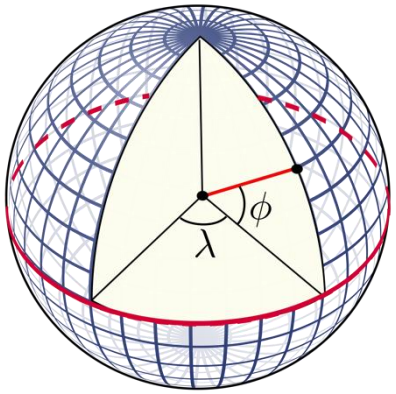


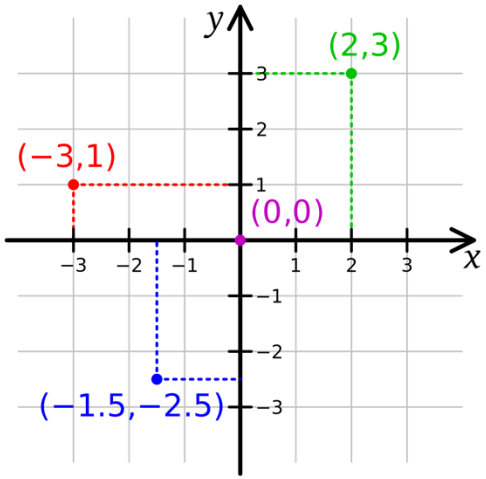


What does "Spatially Referenced" Mean?

Spatial Coordinates

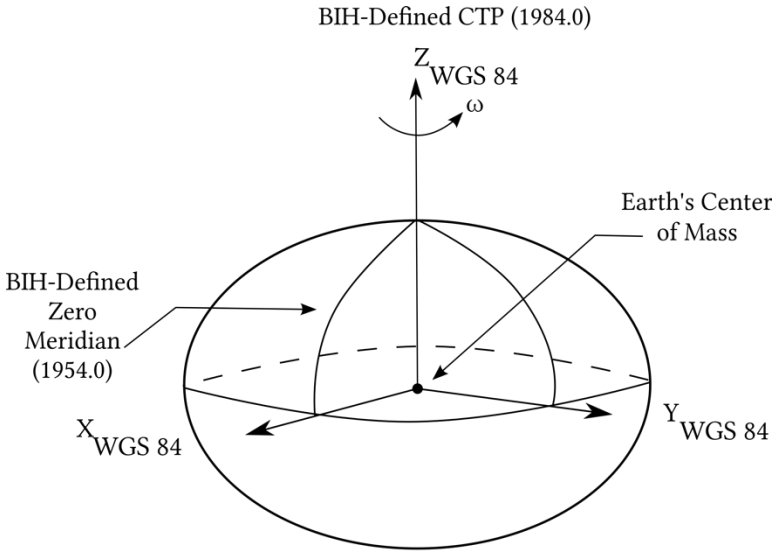
e.g., latitude and longitude, X and Y





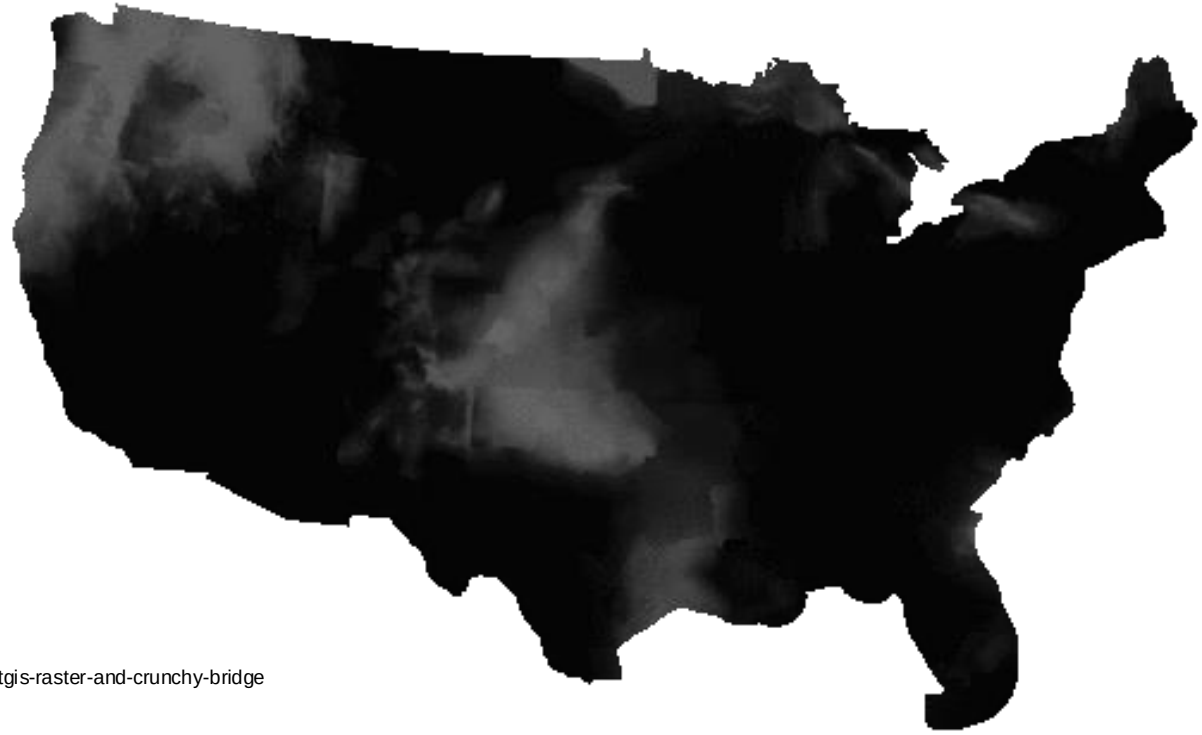
Spatial Reference System

e.g., WGS84, Cartesian System



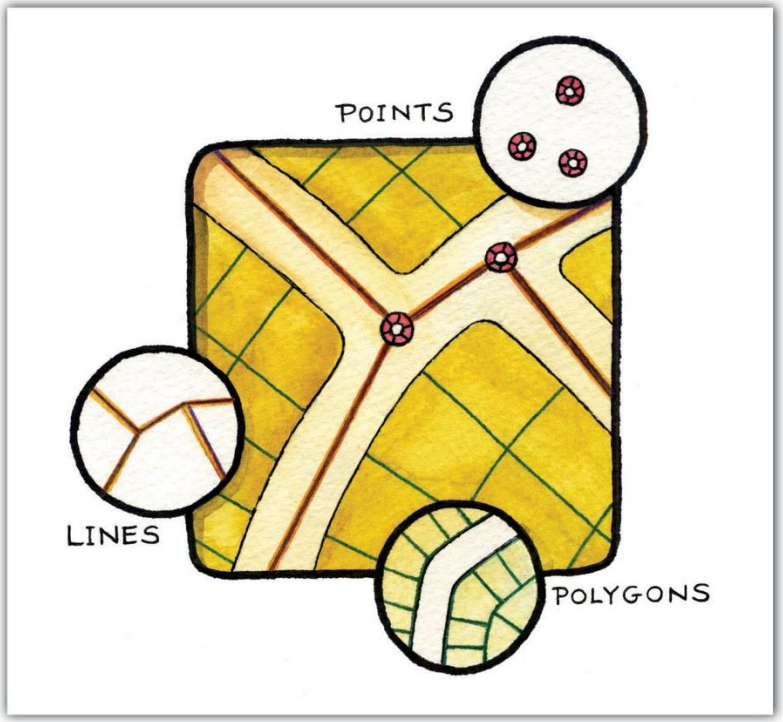
Spatial Data Representations – Raster Data

Probability of Precipitation



https://blog.crunchydata.com/blog/postgis-raster-and-crunchy-bridge

Spatial Data Representations – Vector Data



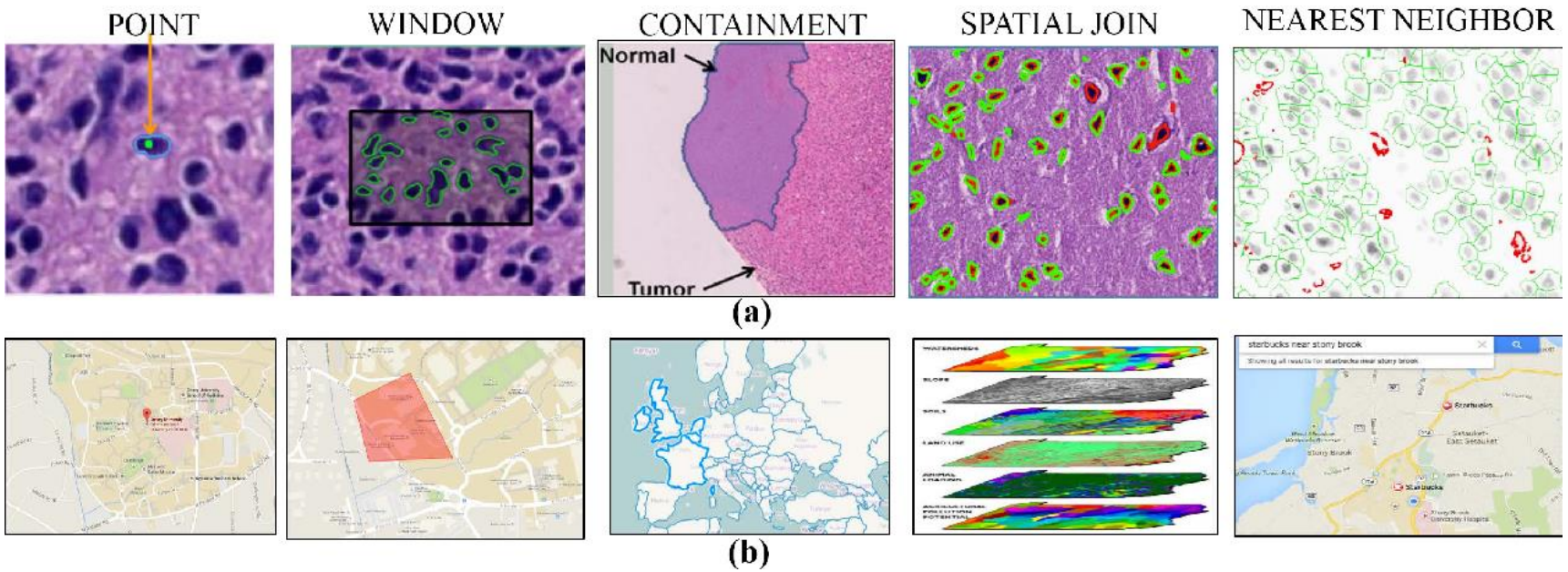
https://saylordotorg.github.io/text_essentials-of-geographic-information-systems/s08-02-vector-data-models.html

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 Query Examples

- Example of spatial query use cases in (a) pathology imaging; (b) GIS



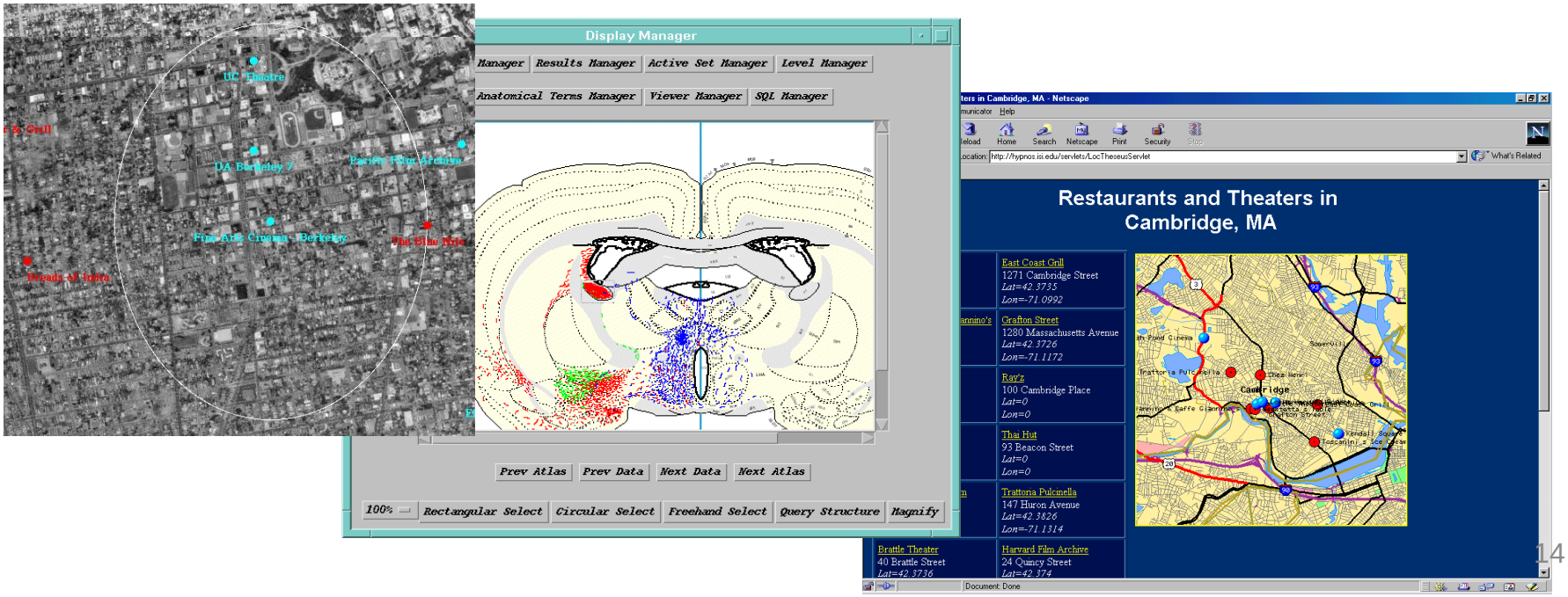
Wang, Fusheng, Ablimit Aji, and Hoang Vo. "High performance spatial queries for spatial big data: from medical imaging to GIS." Sigspatial Special 6, no. 3 (2015): 11-18.

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.

Applications+

- Various fields/applications require management of geometric, geographic or *spatial* data:
  - A geographic space: surface of the earth
  - Man-made space: layout of VLSI design
  - Model of rat brain



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 (since Oracle-8)
  - Has spatial data types (e.g. polygon), operations (e.g. overlap) callable from SQL3 query language
  - Has spatial indices, e.g. R-trees
  - More on available spatial-db's later

What is an SDBMS?*

- Common challenge: dealing with large collections of relatively simple geometric objects
- Different from *image* and *pictorial* database systems:
  - Containing sets of objects in space rather than images or pictures of a space

SDBMS Example

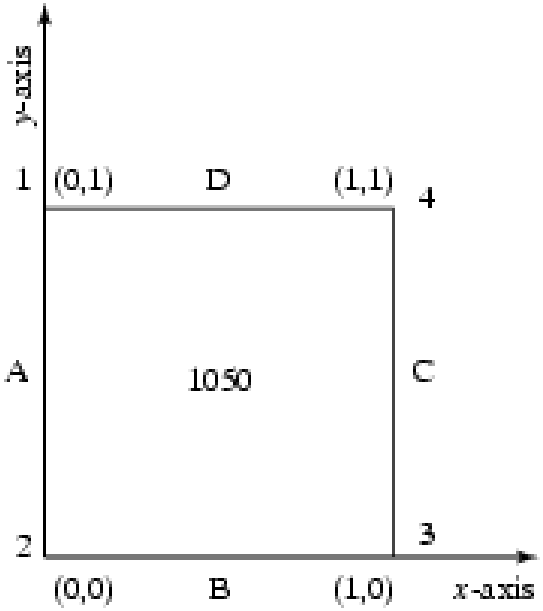
- Consider a spatial dataset with:
  - County boundary (dashed white line)
  - Census block - name, area, population, boundary (dark line)
  - Water bodies (dark polygons)
  - Satellite Imagery (gray scale pixels)

- Storage in a SDBMS table:

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

Modeling Spatial Data in Traditional DBMS

- A row in the table census_blocks
- 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))

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: next slide 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 sqaure represented as 16 rows across 3 tables
  - Simple spatial operators, e.g. area(), require joining tables
  - Tedious and computationally inefficient

Mapping "census_table" into a Relational Database

Census_blocks

Name	Area	Population	boundary-ID
340	1	1839	1050

Polygon

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

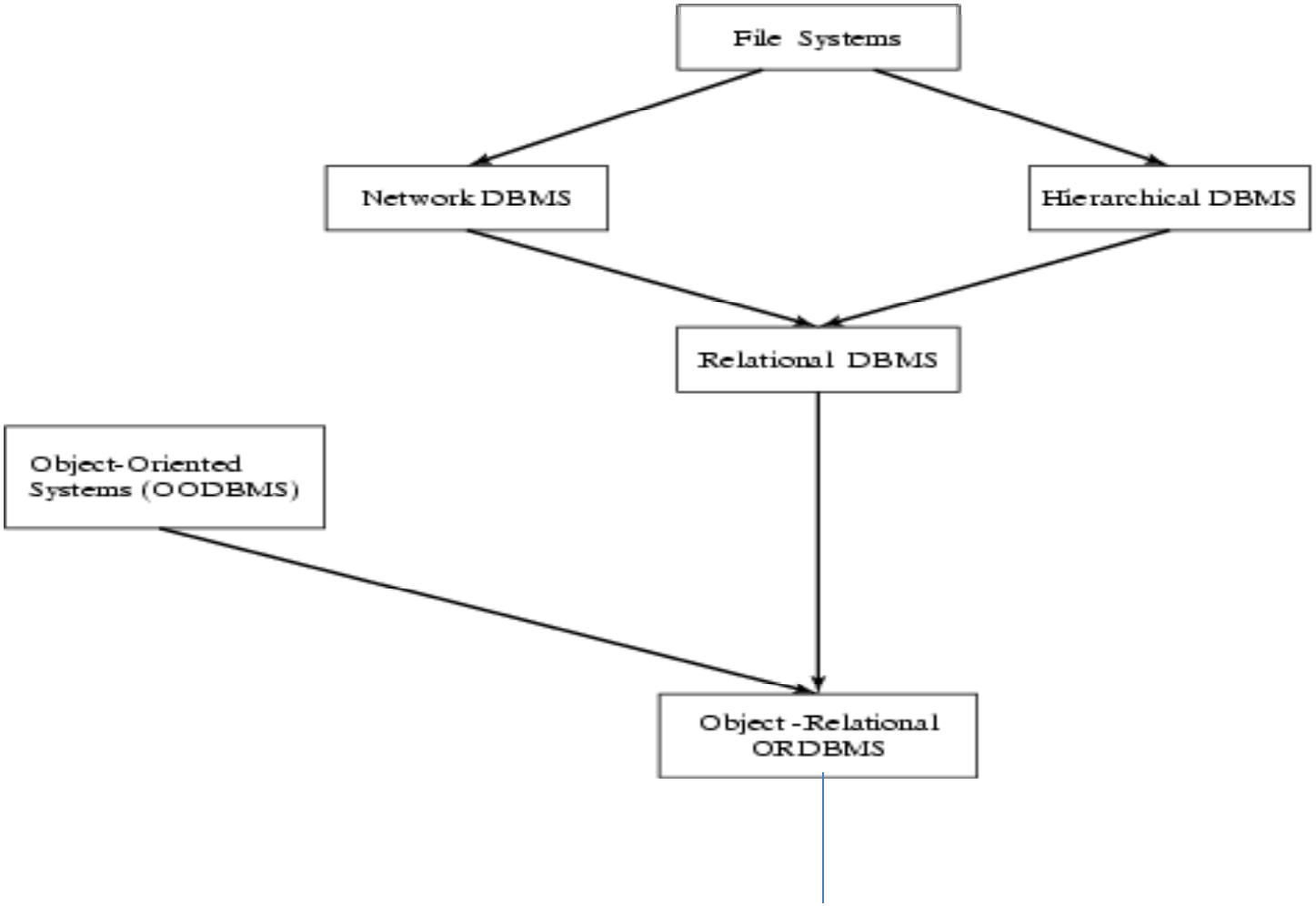
Edge

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

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.

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 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 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?

Three meanings of the acronym GIS

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

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.

Spatial Taxonomy, Data Models *

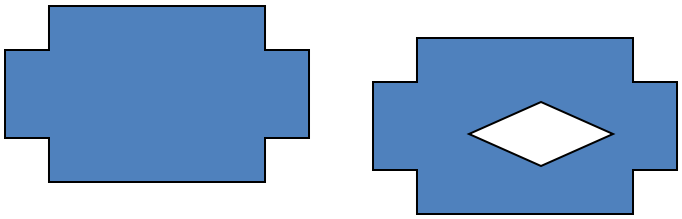
- Spatial Taxonomy:
  - multitude of descriptions available to organize space.
  - Topology models homeomorphic relationships, e.g. overlap
  - Euclidean space models distance and direction in a plane
  - Graphs models connectivity, Shortest-Path
- Spatial data models
  - rules to identify identifiable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

Modeling*

- WLOG assume 2-D and GIS application, two basic things need to be represented:
  - Objects in space: cities, forests, or rivers
  - ➔modeling *single objects*
  - Space: say something about every point in space (e.g., partition of a country into districts)
  - ➔modeling *spatially related collections of objects*

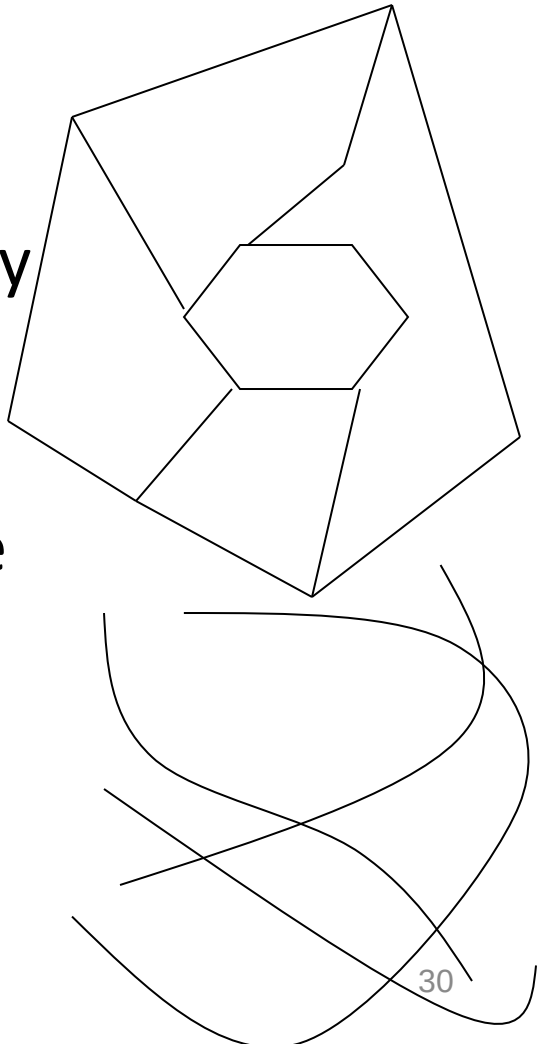
Modeling …

- Fundamental abstractions for modeling single objects:
  - Point: object represented only by its location in space, e.g., center of a state
  - Line (actually a curve or ploy line): representation of moving through or connections in space, e.g., road, river
  - Region: representation of an extent in 2d-space, e.g., lake, city



Modeling* …

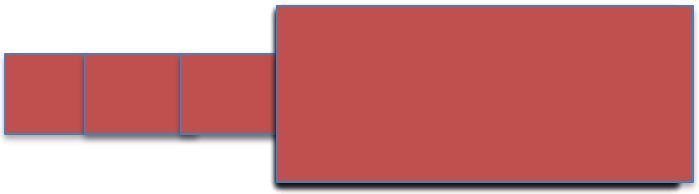
- Instances of spatially related collections of objects:
  - Partition: set of **region** objects that are required to be disjoint (adjacency or region objects with common boundaries), e.g., thematic maps
  - Networks: embedded graph in plane consisting of set of points (vertices) and lines (edges) objects,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



Modeling …

- Spatial relationships:
  - *Topological* relationships: e.g., adjacent, inside, disjoint. Are invariant under topological transformations like translation, scaling, rotation
  - *Direction* relationships: e.g., above, below, or north_of, southwest_of, …
  - *Metric* relationships: e.g., distance
- Enumeration of all possible topological relationships between two simple regions (no holes, connected):
  - Based on comparing two objects boundaries (δA) and interiors (A^o), there are 4 sets each of which be empty or not = 2^4=16.

    8 of these are not valid and 2 symmetric so:
- 6 valid topological relationships:
  - disjoint, touch, overlap, cover, in, equal



Modeling* …

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

relation states (sname: STRING; area: REGION; spop: INTEGER)

relation cities (cname: STRING; center: POINT; ext: REGION; cpop: INTEGER);

relation rivers (rname: STRING; route: LINE)

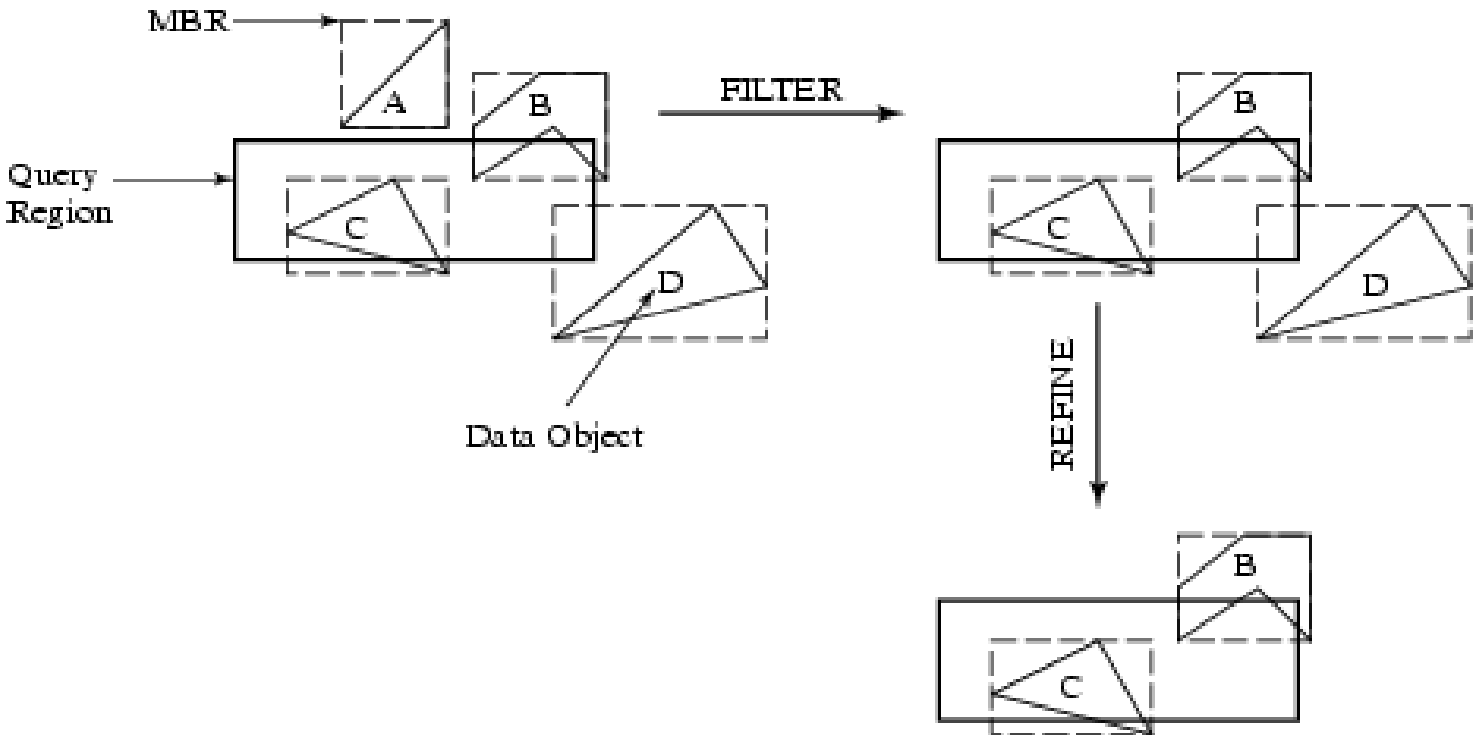
Spatial Query Language

- 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
```

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



Querying* …

Fundamental spatial algebra operations:

- *Spatial selection*: returning those objects satisfying a spatial predicate with the query object
  - "All cities in Bavaria"
    SELECT sname FROM cities c WHERE c.center inside Bavaria.area
  - "All rivers intersecting a query window"
    SELECT * FROM rivers r WHERE r.route intersects Window
  - "All big cities no more than 100 Kms from Hagen"
    SELECT cname FROM cities c WHERE dist(c.center, Hagen.center) < 100 and c.pop > 500k
    (conjunction with other predicates and query optimization)

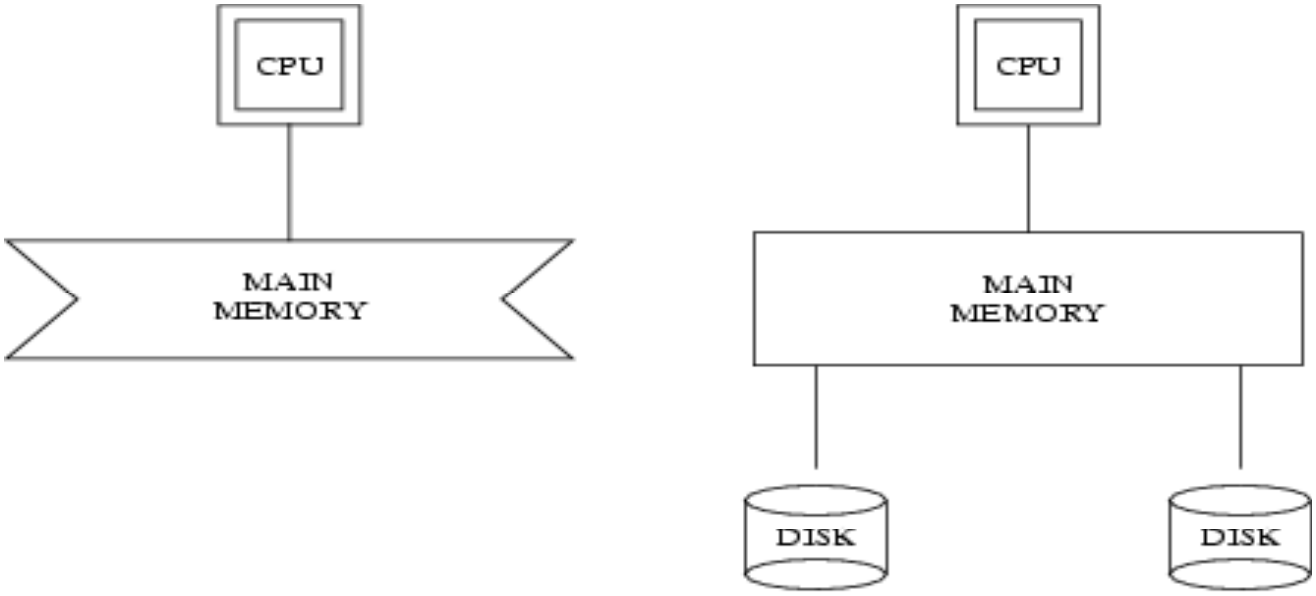
Querying* …

- *Spatial join:* A join which compares any two joined objects based on a predicate on their spatial attribute values.
  - "For each river pass through Bavaria, find all cities within less than 50 Kms."

```
SELECT r.rname, c.cname, length(intersection(r.route,
c.area))
FROM rivers r, cities c
WHERE r.route intersects Bavaria.area and
        dist(r.route,c.area) < 50 Km
```

File Organization and Indices

- A difference between GIS and SDBMS assumptions
  - •GIS algorithms: dataset is loaded in main memory (a)
  - •SDBMS: dataset is on secondary storage e.g disk (b)
  - •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
  - More details on next sessions

1	2	3	4
5	6	7	8
9	10	11	12
13	14	15	16

(a)

7	8	14	16
5	6	13	15
2	4	10	12
1	3	9	11

(b)

Spatial Indexing

- To expedite spatial selection (as well as other operations such as spatial joins, …)
- It organizes space and the objects in it in some way so that only parts of the space and a subset of the objects need to be considered to answer a query.
- Two main approaches:
  - Dedicated spatial data structures (e.g., R-tree)
  - Spatial objects mapped to a 1-D space to utilize standard indexing techniques (e.g., B-tree)

Available Spatial Databases

- Early 90's
  - Illustra (1st OR-DBMS) with some spatial capabilities
  - ESRI's proprietary ArcInfo → SDE (now can use Oracle)
- Late 90's
  - Informix, first commercial database with spatial "blade" (OR-DBMS); sold to IBM
- 2000
  - IBM DB2 with spatial "extender" (still available)
  - Oracle 10g with spatial "extension" → default from Oracle 11g
  - Teradata – TOR with spatial → Acquired Wisconsin's paradise
- Until 2013
  - Domination by Oracle → Oracle 12c (standalone graph and spatial)
- Now…

Centralized Spatial RDBMS

Name	PostGIS (Open-Source)	MySQL (Open-Source)	SpatiaLite (Open-Source)	Oracle Spatial
Description	PostgreSQL with spatial database extensions. [1]	MySQL implements the datatype geometry, plus some spatial functions implemented according to the OpenGIS specifications. [2]	SQLite with spatial datatypes, functions, and utilities.	Oracle with spatial database extensions.
Authority Website	http://postgis.net/	http://www.mysql.com	https://www.gaia-gis.it/fossil/libspatialite/home	https://www.oracle.com/database/technologies/spatialandgraph.html
Operating System	Windows XP, Windows Vista, (haven't tested on 2008), Linux, Unix, Mac OS[8]	Windows 2000+ (including Vista and 2003, haven't tested on 2008), Linux, Unix, Mac OS [8]	MS-Windows, GNU/Linux, Mac OS X, POSIX compliant systems [11]	Windows, Linux [12]
Spatial Type	Points LineStrings Polygons MultiPoints MultiLineStrings MultipPolygons GeometryCollections [3]	Geometry Point LineString Polygon MultiPoint MultiLineString MultiPolygon GeometryCollection [4]	Point LineString Polygon MultiPoint MultiLineString MultiPolygon [7]	Geometry Point LineString Polygon MultiPoint MultiLineString MultiPolygon GeometryCollection [13]
Spatial Index	R-tree-over-GiST (Generalized Search Tree) spatial indexes for high speed spatial querying [3]	R-Tree quadratic splitting - indexes only exist for MyISAM [8]	R-Tree variants [6]	R-Tree variants [13]
Spatial Functions	Over 300 functions and operators, no geodetic support except for point-2-point non-indexed distance functions, custom PostGIS for 2D and some 3D, some MM support of circular strings and compound curves	OGC mostly only MBR (bounding box functions) few true spatial relation functions, 2D only [8]	Basic functions for Point, LineString and Polygon [9]	Basic functions for Point, LineString and Polygon [13]

References

1. PostGIS, http://postgis.net/
2. Spatial database (Wikipedia), https://en.wikipedia.org/wiki/Spatial_database
3. PostGIS (wikipedia), https://en.wikipedia.org/wiki/PostGIS
4. MySQL spatial data types, http://dev.mysql.com/doc/refman/5.7/en/spatial-datatypes.html
5. MySQL create Index Syntax, http://dev.mysql.com/doc/refman/5.7/en/create-index.html
6. A quick tutorial to SpatiaLite - a Spatial extension for SQLite, https://www.gaia-gis.it/gaia-sins/spatialite-tutorial-2.3.1.html
7. SpatiaLite - spatial extensions for SQLite, https://www.gaia-gis.it/spatialite-2.1/SpatiaLite-manual.html
8. Cross Compare SQL Server 2008 Spatial, PostgreSQL/PostGIS 1.3-1.4, MySQL 5-6, http://www.bostongis.com/PrinterFriendly.aspx?content_name=sqlserver2008_postgis_mysql_compare
9. SpatiaLite 4.2.0 SQL functions reference list, http://www.gaia-gis.it/gaia-sins/spatialite-sql-4.2.0.html#p5
10. GIS: PostGIS/PostgreSQL vs. MySql vs. SQL Server?,http://stackoverflow.com/questions/3743632/gis-postgis-postgresql-vs-mysql-vs-sql-server
11. SpatiaLite, https://en.wikipedia.org/wiki/SpatiaLite
12. Oracle Spatial: https://www.oracle.com/technetwork/database/enterprise-edition/downloads/index.html
13. Oracle Spatial Features: https://docs.oracle.com/cd/B28359_01/appdev.111/b28400/sdo_objrelschema.htm#SPATL020

Key-Value Store Spatial DB (NoSQL)

- Couchbase: http://developer.couchbase.com/documentation/server/4.5/indexes/querying-using-spatial-views.html
  - It uses spatial views. Easy-to-write small apps with JavaScript that run on the Couchbase server.
  - The real intent behind the spatial views was to support multidimensional indexes.
  - You can learn more about Spatial Couchbase by watching the video here - there is a small demo: http://www.couchbase.com/nosql-resources/presentations/introducing-spatial-views-for-location-aware-applications-with-couchbase-server-4.0.html
- MongoDB: https://docs.mongodb.com/manual/applications/geospatial-indexes
  - Built-in geospatial operators (e.g. near, within, intersect, ...)
  - It supports geospatial indexing using either 2d or sphere2d
  - MongoDB is one of the fastest, there is also an experiment to compare with Spatial MySQL: https://www.percona.com/blog/2016/04/15/creating-geo-enabled-applications-with-mongodb-geojson-and-mysql/
- CouchDB and ElasticSearch: support indexing and searching point data type (minimum spatial support), not as popular

Cloud-based Spatial DB

- GeoMesa, is an **open-source**, **distributed, spatiotemporal** database
  - Download: http://www.geomesa.org/
  - Document: http://www.geomesa.org/documentation/
  - Support as much as PostGIS adds to Postgres: **indexing of points, polygons and linestrings**, etc.
  - GeoMesa's Index creates a **three-dimensional space filling curve (Z-curve)** from three dimensions of longitude, latitude, and time. The values of the points along this curve are the key.
  - Support integrations with key-value stores: Accumulo, HBase, Cassandra, Kafka, **Spark analytics,** and Google Cloud Bigtable.
  - **Scale horizontally easily** (add more servers to add more capacity)
  - Can integrate with GeoServer - an open source server for sharing geospatial data
  - Store gigabytes to petabytes of spatial data (tens of billions of points)
  - Serve up tens of millions of points in seconds
  - Ingest data faster than 10,000 records per second per node

Cloud-based Spatial DB …

- Hadoop with spatial extensions (developed by ESRI)
  - Overview: https://esri.github.io/gis-tools-for-hadoop/
  - Download link: https://github.com/Esri/spatial-framework-for-hadoop
  - Run a filter and aggregate operations on billions of spatial data records based on location.
  - Define new areas represented as polygons and run a point in polygon analysis on billions of spatial data records inside Hadoop.
  - Visualize analysis results on a map and apply informative symbology.
  - Integrate your maps in reports or publish them as map applications online.
- Others: Hadoop-GIS, SpatialHadoop , MD-HBase, Parallel Secondo

Spark-based Spatial DB

- Apache Spark™ with spatial extensions
  - GeoSpark: https://datasystemslab.github.io/GeoSpark/
    - GeoSpark is implemented on top of Java API for Spark. It seems to have the most succinct and complete implementation of spatial operations.
    - Defines the notion of SpatialRDD (SRDD)
    - Data types: Point, Polygon, Linestring, Multi-point, multi-polygon, multi-linestring, GeometryCollection
    - Indexing: R-Tree, Quad-Tree
    - Partitioning: KDB-Tree, Quad-Tree, R-Tree, Voronoi diagram, Hilbert curve, Uniform grids
    - Operators: Spatial Range Query, Join Query and KNN Query
- Others: SpatialSpark, STARK, Magellan
- More to come…

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