

GIS Data Sources and Structures

Contents

3.1 Capturing GIS data

3.2 Remote Sensing

3.3 Sources: Maps, GPS, Images and Databases

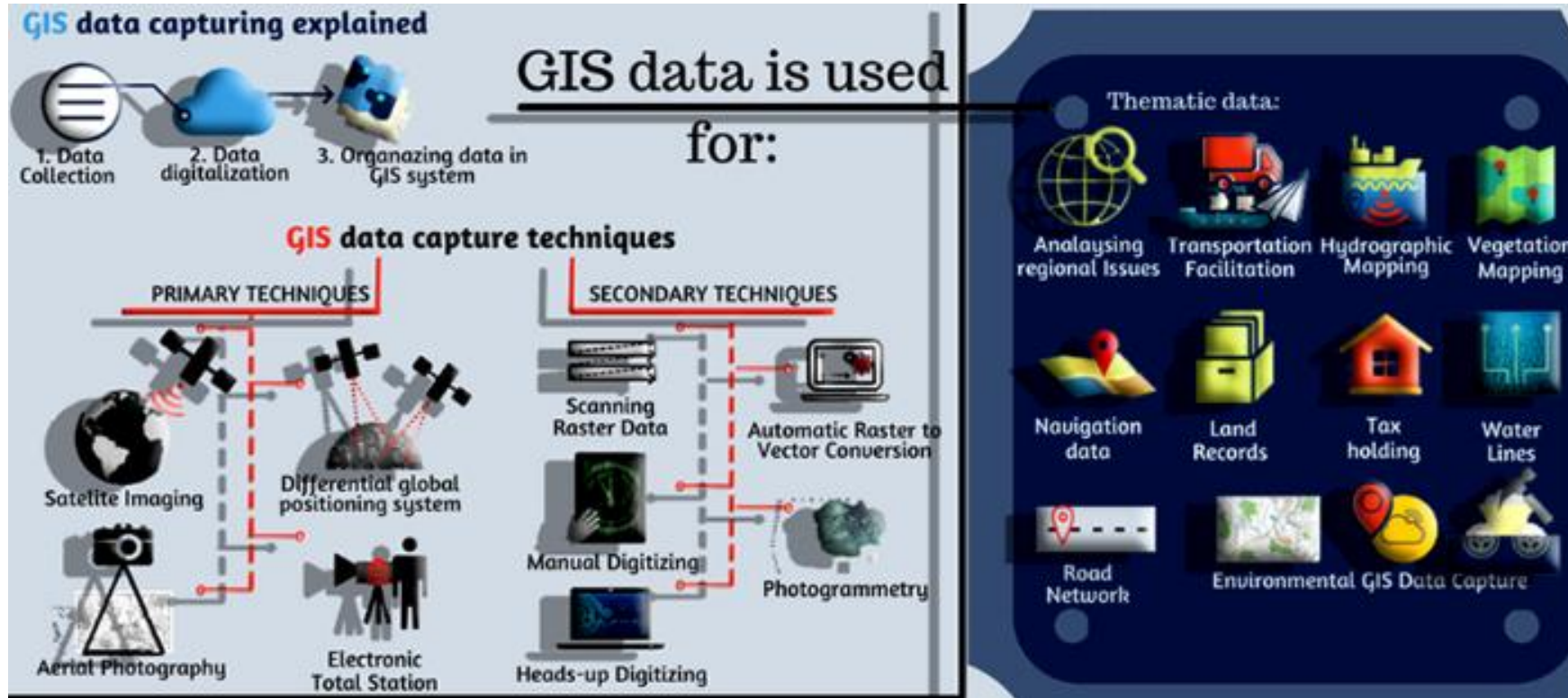
3.4 Structure: Vector, Raster and TIN data Structures

3.5 Spatial relationships and topology

3.6 GIS data modeling

3.7 GIS database design

Different methods of data capture



1. Primary data capture techniques

Raster Data Capture:

- Capturing of attributes without physical contact.
- This is usually done with the help of satellite imaging techniques and aerial photography.
- The advantage of this GIS data capture method is that there is a consistency in the data generated and the whole process can be regularly and systematically organized to get the most accurate data in a very cost-effective manner.

Vector Data Capture:

- Capturing of data-sets through physical surveying techniques such as Differential Global Positioning System (DGPS-which provide improved location accuracy) and Electronic Total Station (ETS).
- Although this technique is the most effective process in order to obtain accurate results on the target GIS system, it is more time consuming and costly.

2. Secondary data capture techniques

- **Scanning Raster Data:** High resolution scanners give very accurate raster images from the hard copies, which then can be georeferenced and digitized to get the vector output.
- **Manual Digitizing:** Digitization is done directly over the raster by the use of a digitizing tablet.
- **Heads-up Digitizing:** The raster scanned data is imported and laid below the vector data to be traced on the computer screen itself.
- **Automatic Raster to Vector Conversion:** UIZ(Umwelt und Information Zentrum) uses special software with intelligent algorithms to recognize the patterns of the points, lines and polygon features and capture them automatically to generate vector GIS data.
- **Photogrammetry:** UIZ uses digital stereo-plotters to capture the vector data from the aerial photographs(taking of photographs from an aircraft or other flying object). This is comparatively the most effective method for accurately capturing GIS data.

Some capturing methods in detail

- **Scanning from paper maps:**
- The scanner will take any printed image and take a picture of it. By capturing the image in digital form, it can be stored on the computer and displayed on screen.
- Scanning a map is a straightforward process and generally fast, but it does not provide for the capture of attribute information for features, such as the address of a building.



- **Digitizing from paper maps:**

- Digitizing requires the use of special equipment. The source map is laid flat on a table (tablet) and an electronic cursor is passed over the features of the map.
- For digitizing to work, the tablet must have a magnetic field embedded in the flat surface, so as the cursor is moved around the map, its location can be identified.
- Digitizing can be very time consuming because every single point or vertex must be captured individually.



- **Heads-up digitizing and vectorization:**

- Vectorization is the process of converting raster data into vector data.
- The simplest way to create vectors from raster layers is to digitize vector objects manually straight off a computer screen using a mouse or digitizing cursor.
- Describes how automated vectorization is performed

- **Surveying:**

- Ground surveying is based on the principle that the 3-D location of any point can be determined by measuring angles and distances from other known points.
- Traditional equipment like transits and theodolites have been replaced by total stations that can measure both angles and distances to an accuracy of 1 mm
- Ground survey is a very time-consuming and expensive activity, but it is still the best way to obtain highly accurate point locations.
- Typically used for capturing buildings, land and property boundaries, manholes, and other objects that need to be located accurately.
- Also employed to obtain reference marks for use in other data capture projects.

- **Photogrammetry:**

- Is the science and technology of making measurements from pictures, aerial photographs, and images.
- Measurements are captured from overlapping pairs of photographs using stereo plotters.

- **Satellite Data :**
- The data obtained from the Satellites are in digital form, which can be directly imported to GIS. There are numerous satellite data sources such as LANDSAT or SPOT.
- A new generation of high-resolution satellite data that will increase opportunities and options for GIS database development is becoming available from private sources and national governments.
- These satellite systems will provide panchromatic (black and white) or multi-spectral data in the 1- to 3-meter ranges as compared to the 10- to 30-meter range available from traditional remote sensing satellites.



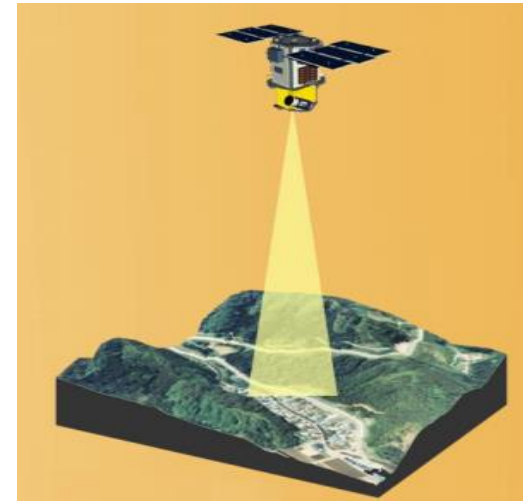
- **Global Positioning System (GPS):**
- GPS can be used almost anywhere in the world, 24 hours a day, in all weathers. A constellation of 24 satellites orbit the earth and send signals that can be picked up by GPS receivers.
- GPS measurements are taken by computing the distance between the receiver and the satellite. If a receiver picks up signals from four or more satellites, a 3-dimensional position can be calculated.
- GPS measurements are obtained in the GPS coordinate system.
- **Tabular Data Entry:**
- The tabular attribute data that is normally in a GIS database exists on maps as annotation and or can be found in paper files
- **Document Scanning :**
- Smaller-format scanners can also be used to create raster files of documents such as permit forms, service cards, site photographs, etc.
- These documents can be indexed in a relational database by number, type, date, engineering drawings, etc., and queried and displayed by users.

GIS Data Capture Methods Used in Various Fields

- Thematic data is used for analyzing regional issues, transportation facilitation, hydrographic mapping, vegetation and other types of related features.
- Navigation data is captured and used for easy navigation purposes.
- Land records and survey data are captured for property, land, water and holding tax, etc.
- Utility infrastructure GIS data capture for water lines, road network, pavements, sewerage network, and other related features.
- Environmental GIS Data capture is done from geological maps, weather maps, mining and mineral exploration maps, etc.

Remote Sensing

- The science of acquiring information about the earth using instruments which are remote to the earth's surface, usually from aircraft or satellites. Instruments may use visible light, infrared or radar to obtain data.



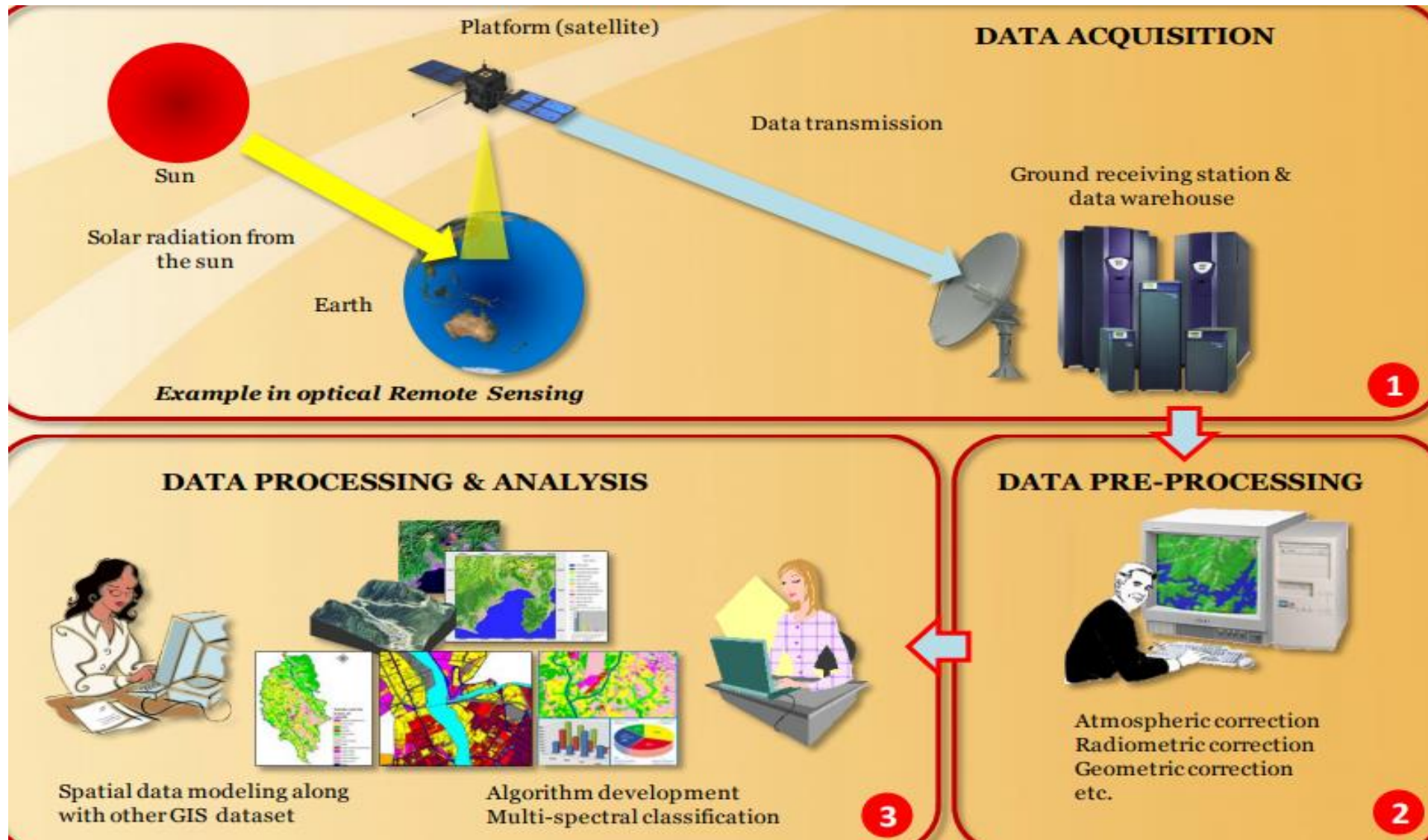
How does remote sensing work?

Components in Remote Sensing:

Platform: The vehicle which carries a sensor. i.e. satellite, aircraft, balloon, One platform can carry more than one sensor. For example:

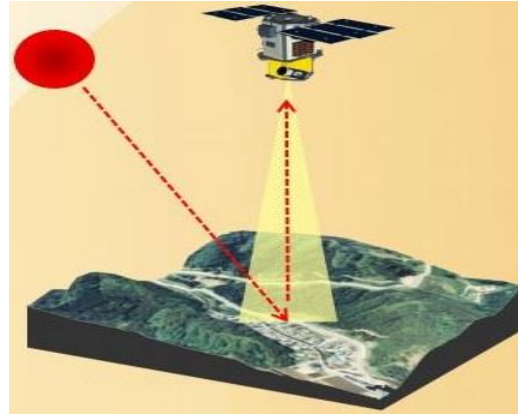
etc...

Sensors: Device that receives electromagnetic radiation and converts it into a signal that can be recorded and displayed as either numerical data or an image.

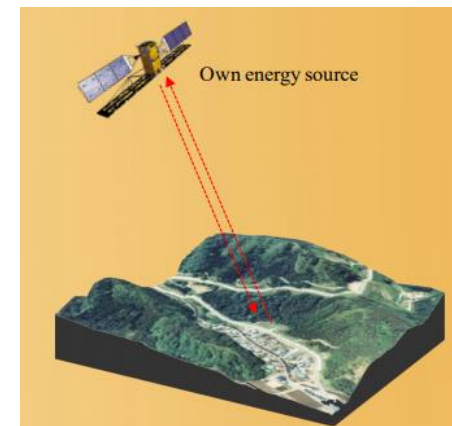


Types of Remote Sensing:

- **Passive Remote Sensing** Remote sensing of energy naturally reflected or radiated from the terrain.



- **Active Remote Sensing** Remote sensing methods that provide their own source of electromagnetic radiation to illuminate the terrain. Radar is one example.



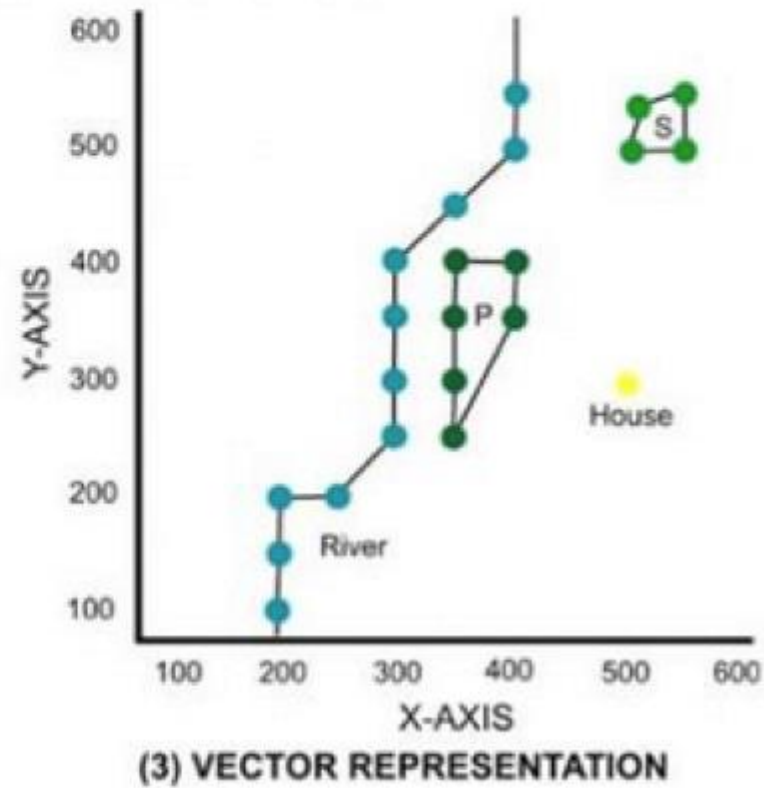
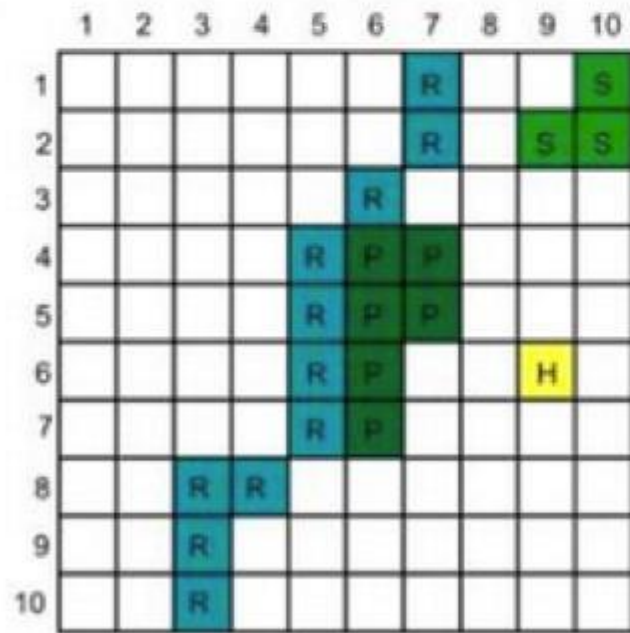
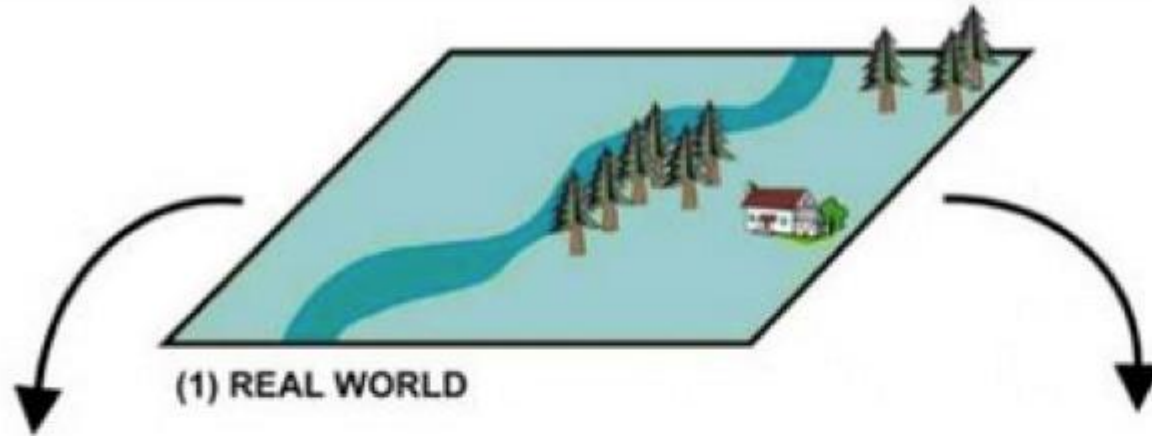
Data Types in GIS

The data in a GIS can be classified into two main categories:

1. **Spatial data:** Describes the absolute and relative location of geographic features.
2. **Attribute data or Non-spatial data:** Describes characteristics of the spatial features. These characteristics can be quantitative and/or qualitative in nature.

The Data Model

- Data model is a conceptual description (mental model) of how spatial data are organized for use by the GIS.
- The data model represents a set of guidelines to convert the real world (called entity) to the digitally and logically represented spatial objects consisting of the attributes and geometry.
- The attributes are managed by thematic or semantic structure while the geometry is represented by geometric-topological structure.
- There are two major types of geometric data model.
 - a. Vector Model:** Vector model uses discrete points, lines and/or areas corresponding to discrete objects with name or code number of attributes.
 - b. Raster Model :**Raster model uses regularly spaced grid cells in specific sequence. An element of the grid cell is called a pixel (picture cell). The conventional sequence is row by row from the left to the right and then line by line from the top to bottom. Every location is given in two dimensional image coordinates ; pixel number and line number, which contains a single value of attributes.s

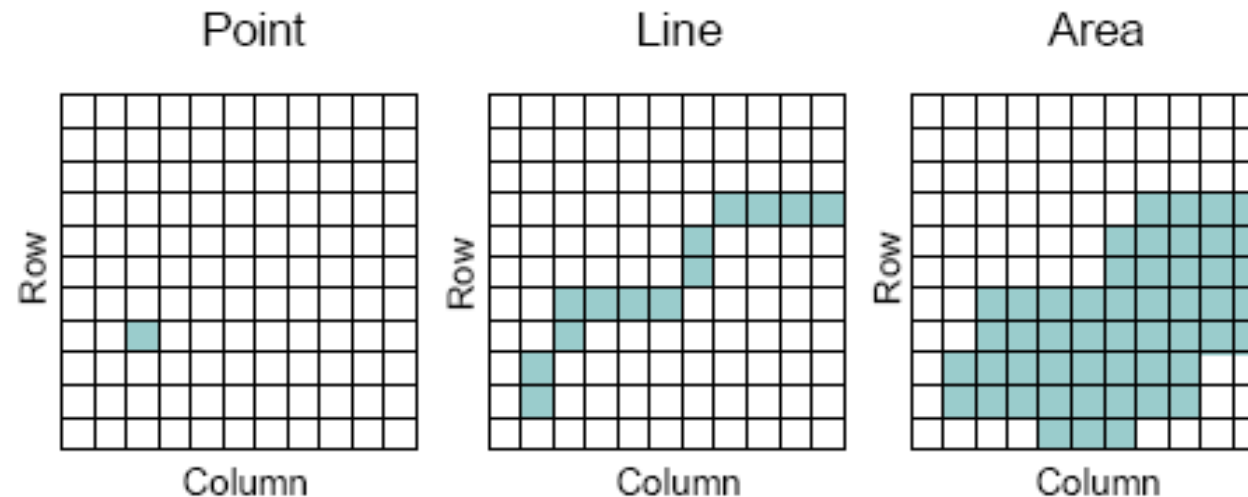


Data models in GIS

- Raster data model
- Vector data model
- Triangulated irregular network model(TIN)
- Digital elevation model (DEM)
- Network models

Raster Data Model

- The term raster implies a regularly spaced grid . Raster data consists of rows and columns of cells (or pixels). In this format a single value is stored against each cell. Raster data can represent a multiplicity of things including:
 - Visual images (that is colour and/or hue)
 - Discrete value, such as land use
 - Continuous value, such as rainfall
 - Null values if no data is available.



Cell Size of Raster Data

- The level of detail represented by a raster is often dependent on the cell (pixel) size or spatial resolution of the raster. The cell must be small enough to capture the required detail but large enough so computer storage and analysis can be performed efficiently.

Smaller cell size

- Higher resolution
- Higher feature spatial accuracy.
- Slower display
- Slower processing
- Large file size

Larger cell size

- Lower resolution
- Lower feature spatial accuracy
- Faster display
- Faster processing
- Smaller file size

Advantages of Raster

- It is a simple data structure.
- It has the ability to represent continuous surfaces and perform surface analysis.
- The ability to uniformly store points, lines, polygons and surfaces.
- The ability to perform fast overlays (than vector datasets) with complex datasets.













Disadvantages of Raster

- Inherent spatial inaccuracies due to the cell-based feature representation.
- Raster datasets are potentially very large. Resolution increases as the size of cells decreases. Accordingly cost and disk space used also increases.
- There is also a loss of precision that accompanies restructuring data to a regularly spaced raster cell boundary.
- Difficult in representation topology connections.

Vector Data Model

- Vectors are graphical objects that have geometrical primitives such as points, lines and polygons to represent geographical entities in the computer graphics.
- A vector refers to a geometrical space which has a precise direction, length and shape.
- Points, Lines and Polygons can be defined by the coordinate geometry.
- A vector spatial data model uses two-dimensional Cartesian (x, y) coordinate system to store the shape of a spatial entity.

- In vector world the point is the basic building block from which all spatial entities are constructed.
- The simplest spatial entity, the point, is represented by a single (x, y) coordinate pair.
- Line and area entities are constructed by connecting a series of points into chains and polygons.

	Qualitative Distinction	
POINT		Town
		Church
		Triangulation pillar
		Wind pump
LINE		River
		Road
		Railway
		Boundary
AREA		Marsh
		Desert
		Forest
		Political units

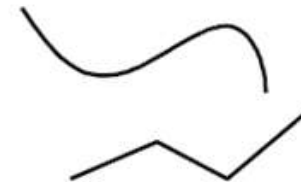
Point

- A point is a 0 dimensional object and has only the property of location (x,y).
- Points can be used to Model features such as a well, building, power pole, sample location etc.
- Other names for a point are vertex, node, 0-cell.



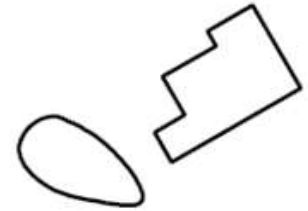
Line

- A line is a one-dimensional object that has the property of
- Lines can be used to represent road, streams, faults, dikes, marker beds, boundary, contacts etc.
- Lines are also called an edge, link, chain, arc, 1-cell.
- Connected multiple lines are called **polylines**.



Polygon

- Polygon features are made of one or more lines that encloses an area.
- A polygon is a two dimensional object with properties of area and perimeter represented by a closed sequence of lines.
- A polygon can represent a city, geologic formation, dike, lake, river, etc.



Advantages of Vector

- Requires less disk storage space.
- Efficient for topological relationship
- Graphical output more closely resembles hand-drawn maps.
- Easy to edit
- Accurate map output
- Efficient projection transformation

Disadvantages of Vector

- Complex data structure.
- Less compatibility with remotely sensed data.
- Expensive software and hardware.
- Not appropriate to represent continuous data
- Overlaying multiple vector are often time consuming.

Difference between Raster and Vector

Raster

- It is a simple data structure.
- Overlay operations are easily and efficiently implemented.
- High spatial variability is efficiently represented in a raster format.
- The raster format is more or less required for efficient
- manipulation and enhancement of digital images.
- The raster data structure is less compact.
- Topological relationships are more difficult to represent.
- The output of graphics is less aesthetically pleasing because boundaries tend to have a blocky appearance rather than the smooth lines of hand drawn maps. This can be overcome by using very large number of cells, but it may result in unacceptably large files.

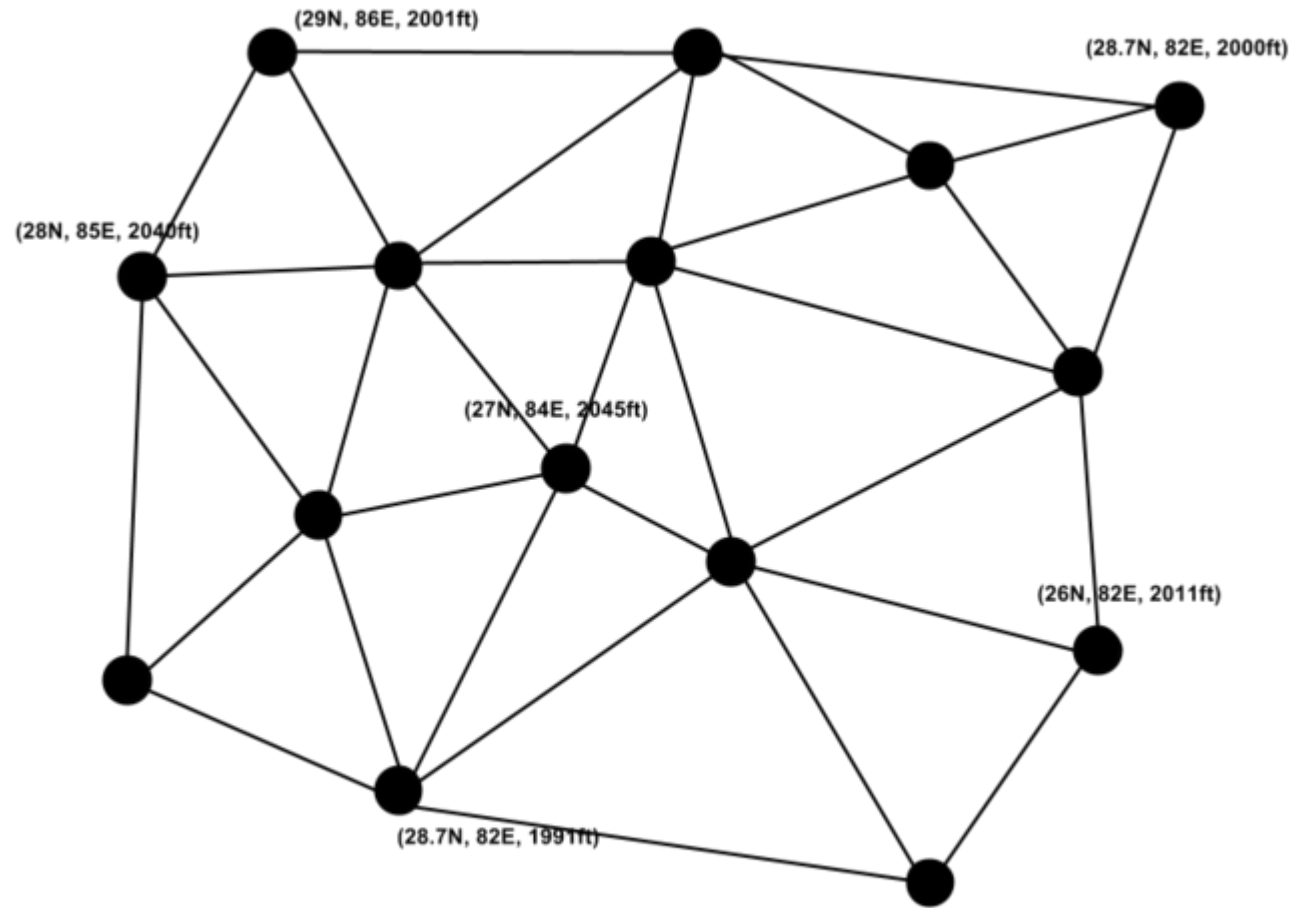
Vector

- More complex data structure.
- Overlay operations are more difficult to implement.
- The representation of high spatial variability is inefficient.
- Manipulation and enhancement of digital images cannot be effectively done in the vector domain.
- Vector provides a more compact data structure.
- Provides efficient encoding of topology.
- The vector data model is better suited to supporting graphics that closely approximate hand-drawn maps.

Triangulated Irregular Network(TIN)

- A triangulated irregular network (TIN) approximates the terrain with a set of non overlapping triangles.
- Each triangle in the TIN assumes a constant gradient, Flat areas of the land surface have fewer but larger triangles, whereas areas with higher variability in elevation have denser but smaller triangles.
- A TIN is a vector based representation of the physical land surface or sea bottom, made up of irregularly distributed nodes and lines with three dimensional coordinates (x,y, and z) that are arranged in a network of non overlapping triangles.
- The TIN is commonly used for terrain mapping and analysis, especially for 3-D display. The final result gives users a TIN surface.

Triangulated Irregular Network (TIN)



Advantages of TIN

- TIN's give researchers the ability to view 2.5D and 3D at an area that was interpolated from minimal data collection.
- Users can describe a surface at different levels of resolution based on the points that were collected.
- TIN interpolation gives GIS users greater analytical capabilities. TIN models are easy to create and use.
- They provide users a simplified model that represents collected data points.
- Using a TIN surface in conjunction with ArcMap extensions such as Spatial Analysis and 3D Analyst, TIN users can also derive slope, aspect, elevation, contour lines, hillshades, etc.

Digital Elevation Model (DEM)

- A Digital Elevation Model (DEM) is a representation of the bare ground (bare earth) topographic surface of the Earth excluding trees, buildings, and any other surface objects.
- Digital Elevation Model is a data model which represents the surface of a terrain in 3 dimension.
 - DEM can be represented as a raster or as TIN.
 - The TIN DEM dataset is also referred to as a primary DEM or measured DEM.
 - Raster DEM is referred to as secondary DEM or computed DEM.

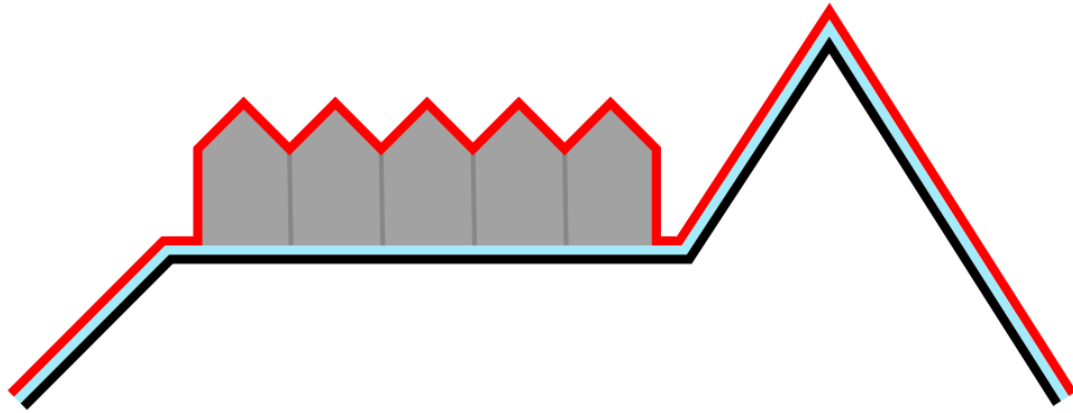
Digital terrain model (DTM)

- “A digital elevation model (DEM) represents the elevation of Earth’s surface, including features (vegetation, buildings, etc.). A digital terrain model (DTM) provides a bare earth representation of terrain ((also topographical relief) involves the vertical and horizontal dimensions of land surface.) or surface topography. Both are highly useful data sets for visualizing our planet for scientific and commercial landscape study. With each technological advancement, the digital elevation models have improved in accuracy, resulting in a much more useful model of the Earth.”
- DTM can be stored in a GIS database in several ways:
 - 1) a set of contour vectors;
 - 2) a rectangular grid of equal-spaced corner/point heights; or,
 - 3) an irregularly spaced set of points connected as triangles (TIN - Triangular Irregular Network).

- The DTM data sets are extremely useful for the generation of 3D renderings of any location in the area described.
- 3D models rendered from DTM data can be extremely useful and versatile for a variety of applications.
- DTMs are used especially in civil engineering, geodesy & surveying, geophysics, and geography

The main applications are:

1. Visualization of the terrain
2. Reduction (terrain correction) of gravity measurements (gravimetry, physical geodesy)
3. Terrain analyses in Cartography and Morphology
4. Rectification of airborne or satellite photos
5. Extraction of terrain parameters, model water flow or mass movement

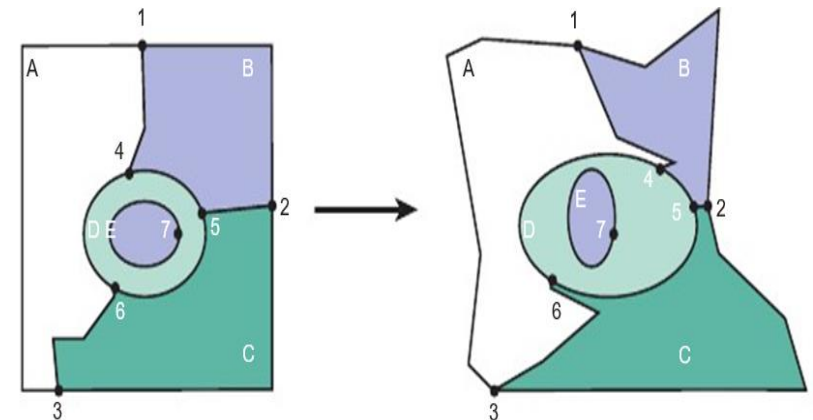


	Digital Surface Model
	Digital Terrain Model

Spatial relationships and topology

General spatial topology

- Topology refers to the spatial relationships between geographical elements in a data set that do not change under a continuous transformation.
- Example: Assume you have some features that are drawn on the sheet of rubber (as in figure). Now take the sheet and pull on its edges, but do not tear or break it. The features will change in shape and size. But some properties, however, do not change.
 - Area E is still inside area D,
 - The neighborhood relationships between A, B, C, D, and E stay intact, and their boundaries have the same start and end nodes, and
 - The areas are still bounded by the same boundaries, only the shapes and lengths of their perimeters have changed.



Topological relationships

- The space is a three-dimensional Euclidean space where for every point we can determine its three-dimensional coordinates as a triple (x, y, z) of real numbers. In this space, we can define features like points, lines, polygons, and volumes as geometric primitives of the respective dimension. A point is zero-dimensional, a line one-dimensional, a polygon two-dimensional, and a volume is a three-dimensional primitive.
- The space is a metric space (a metric space is a set for which distances between all members of the set are defined.), which means that we can always compute the distance between two points according to a given distance function. Such a function is also known as a metric

- The space is a topological space, of which the definition is a bit complicated. In essence, for every point in the space we can find a neighborhood around it that fully belongs to that space as well.
- Interior and boundary are properties of spatial features that remain invariant under topological mappings. This means, that under any topological mapping, the interior and the boundary of a feature remains unbroken and intact.

Topology has three main advantages

- First, it ensures data quality and integrity.
- This was in fact polygons in a topology-based data set, specific habitat types (e.g., old growth and clear-cuts) along edges can easily be tabulated and analyzed.
- Third, topological relationships between spatial features allow GIS users to perform spatial data query.
- As examples, we can ask how many schools are contained within a county and which land parcels are intersected by a fault line.

Data Models for Spatial Data

There are presently three types of representations for geographic data: raster vector, and objects.

- **Raster:** set of cells on a grid that represents an entity (entity --> symbol/color --> cells).
- **Vector:** an entity is represented by nodes and their connecting arc or line segment (entity --> points, lines or areas --> connectivity)
- **Object:** an entity is represented by an object which has as one of its attributes spatial information.

Raster Data model

- Realization of the external model which sees the world as a continuously varying surface (field) through the use of 2-D Cartesian arrays forming sets of thematic layers. Space is discretized into a set of connected two dimensional units called a tessellation.

a.Map overlays: separate set of Cartesian arrays or "overlays" for each entity.

b.Logical data models: 2-D array, vertical array, and Map file

- Each overlay is a 2-D matrix of points carrying the value of a single attribute.
- Each point is represented by a vertical array in which each array position carries a value of the attribute associated with the overlay.
- Map file- each mapping unit has the coordinates for cell in which it occurs (greater structure, many to one relationship).

c.Compact methods for coding: Vertical array not conducive to compact data coding because it references different entities in sequence and it lacks many to one relationship. The third structure references a set of points for a region (or mapping unit) and allows for compaction.

Vector data model

- Realization of the discrete model of real world using structures for storing and relating points, lines and polygons in sets of thematic layers.
- Represents an entity as exact as possible.
- Coordinate space continuous (not quantized like raster).
- Structured as a set of thematic layers.

Representation

1. **Point entities:** geographic entities that are positioned by a single x,y coordinate. (historic site, wells, rare flora. The data record consists for x,y - attribute.
2. **Line Entity:** (rivers, roads, rail)
 - a. all linear feature are made up of line segments.
 - b. a simple line 2 (x,y) coordinates.
 - c. an arc or chain or string is a set of n (x,y) coordinate pairs that describe a continuous line. The shorter the line segments the closer the chain will approximate a continuous curve. Data record n(x,y).

d. a line network gives information about connectivity between line segments in the form of pointers or relations contained in the data structure. Often build into nodes pointers to define connections and angles indicating orientation of connections (fully defines topology).

3. Area Entity: data structures for storing regions. Data types, land cover, soils, geology, land tenure, census tract, etc.

a. Cartographic spaghetti or "connect the dots". Early development in automated cartography, a substitute for mechanical drawing. Numerical storage, spatial structure evident only after plotting, not in file.

b. Location list: describe each entity by specifying coordinates around its perimeter.

- shared lines between polygons.
- polygon sliver problems.
- no topology (neighbor and island problems).
- error checking a problem.

c. Point dictionary:

- unique points for entire file, no sharing of lines as in location lists (eliminate sliver problem) but still has other problems.
- expensive searches to construct polygons.

d. Dime Files (Dual Independent Mapping and Encoding):

- designed to represent points lines and areas that form a city though a complete representation of network of streets and other linear features.
- allowed for topologically based verification.
- no systems of directories linking segments together (maintenance problem).

e. Arc/node:

- same topological principles as the DIME system.
- DIME defined by line segments, chains based on records of uncrossed boundary lines (curved roads a problem for DIME).

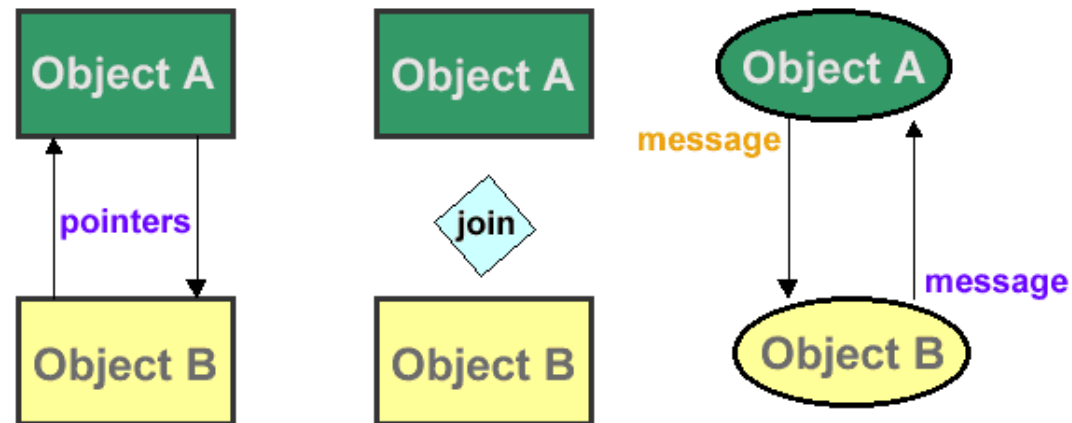
- chains or boundaries serve the topological function of connecting two end points called a node and separating two zones.
- points between zones cartographically not topologically required (generalization possible).
- solves problems discussed above (neighbor, dead ends, weird polygons).
- can treat data input and structure independently.

Object-oriented data model

Realization of the discrete model of real world using an object centered approach in which an object has both physical (attribute) and geometric characteristics. Different types of objects can interact because they are not confined to separate layers.

Object-oriented Data Model

- * Simplified, less constrained interface between objects.
- * Relationships between entities is via **messages** not pointers or join fields.



- The biggest single difference between the object-oriented conceptual model and the vector-layered based conceptual model, for representing geographic information, is that in the object model, the real world object is the basis for abstraction, not its geometry. In other words, the objects not the geometric components of layers are the "units" for modeling and interactions

Spatial database design with the concepts of geodatabase

Spatial DBMS:

- A spatial database system may be defined as a database system that offers spatial data types in its data model and query language, and supports spatial data types in its implementation, providing at least spatial indexing and spatial join methods.
- Spatial database systems offer the underlying database technology for geographic information systems and other applications.
- We survey data modeling, querying, data structures and algorithms, and system architecture for such systems.
- A spatial database therefore has the following characteristics:
 - (1) A spatial database system is a database system.
 - (2) It offers spatial data types (SDTs) in its data model and query language.
 - (3) It supports spatial data types in its implementation, providing at least spatial indexing and efficient algorithms for spatial join.

- Three main categories of spatial modeling functions that can be applied to geographic features within a GIS are:
 - (1) Geometric models, such as calculating the Euclidean distance between features, generating buffers, calculating areas and perimeters, and so on;
 - (2) Coincidence models, such as topological overlay; and
 - (3) Adjacency models (path finding, redistricting, and allocation).

All three model categories support operations on spatial data such as points, lines, polygons, tins, and grids.

Steps in database design

1. Conceptual

- software and hardware independent
- describes and defines included entities
- identifies how entities will be represented in the database

i.e. selection of spatial objects - points, lines, areas, raster cells

- requires decisions about how real-world dimensionality and relationships will be represented

these can be based on the processing that will be done on these objects

e.g. should a building be represented as an area or a point?

e.g. should highway segments be explicitly linked in the database?

2. Logical

- Software specific but hardware independent
- Sets out the logical structure of the database elements, determined by the data base management system used by the software

3. Physical

- Both hardware and software specific
- Requires consideration of how files will be structured for access from the disk

Characteristics of a good database design

- The data should be updated regularly
- Flexible and extensible so that additional datasets may be added as necessary for the intended applications
- The categories of information and subcategories within them should contain all of the data needed to analyze or model the behavior of the resource using conventional methods and models.

- Positionally accurate – if for example the boundary between the residential and agricultural land has changed, this may be incorporated with ease.
- Exactly compatible with other information that may be overlain with it
- Internally accurate, portraying the nature of phenomena without error -requires clear definitions of phenomena that are included
- Readily updated on a regular schedule

Spatial Database Management

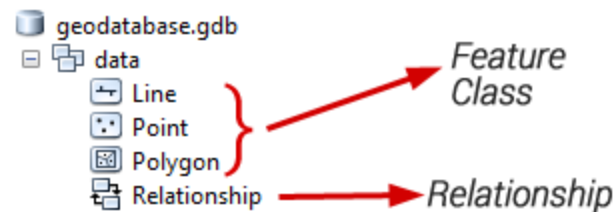
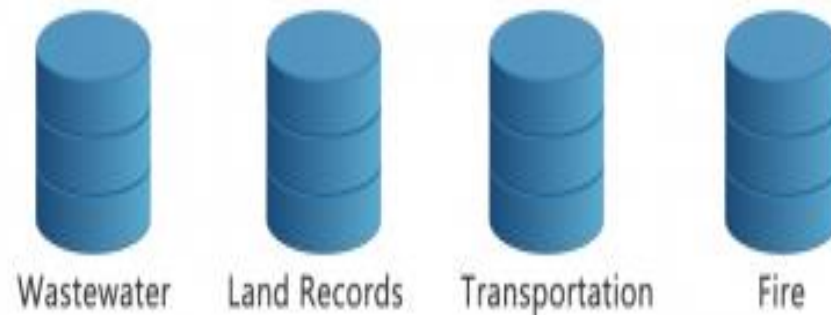
A good spatial database management software package should be able to:

1. Scale and rotate coordinate values for "best fit" projection overlays and changes.
2. Convert (interchange) between polygon and grid formats.
3. Permit rapid updating, allowing data changes with relative ease.
4. Allow for multiple users and multiple interactions between compatible data bases.
5. Retrieve, transform, and combine data elements efficiently.
6. Search, identify, and route a variety of different data items and score these values with assigned weighted values, to facilitate proximity and routing analysis.
7. Perform statistical analysis, such as multivariate regression, correlations, etc.
8. Overlay one file variable onto another, i.e., map super positioning.
9. Measure area, distance, and association between points and fields.
10. Model and simulate, and formulate predictive scenarios, in a fashion that allows for direct interactions between the user group and the computer program.

Geodatabases

- Geodatabases organize geographic data into a hierarchy of data objects. These data objects are stored in feature classes, object classes, and feature datasets.
- An object class is a table in the geodatabase that stores non-spatial data. A feature class is a collection of features with the same type of geometry and the same attributes.
- A feature dataset is a collection of feature classes that share the same spatial reference. Feature classes that store simple features can be organized either inside or outside a feature dataset.
- Simple feature classes that are outside a feature dataset are called standalone feature classes. Feature classes that store topological features must be contained within a feature dataset to ensure a common spatial reference.

- When you can add coded value domains, raster catalogs, relationship classes and geometric networks, geodatabases truly are the multi-functional engine an organization needs.
- Geodatabases also excel in performance. Spatial functions run quicker in a database such as Performance Querying Indexing.



Representation

Individual geographic entities can be represented as

- Feature classes (sets of points, lines, and polygons).
- Imagery and raster.
- Continuous surfaces that can be represented using features (such as contours), rasters(digital elevation models [DEM]), or triangulated irregular networks (TINs) using terrain datasets.
- Attribute tables for descriptive data.