Data Input and its representation

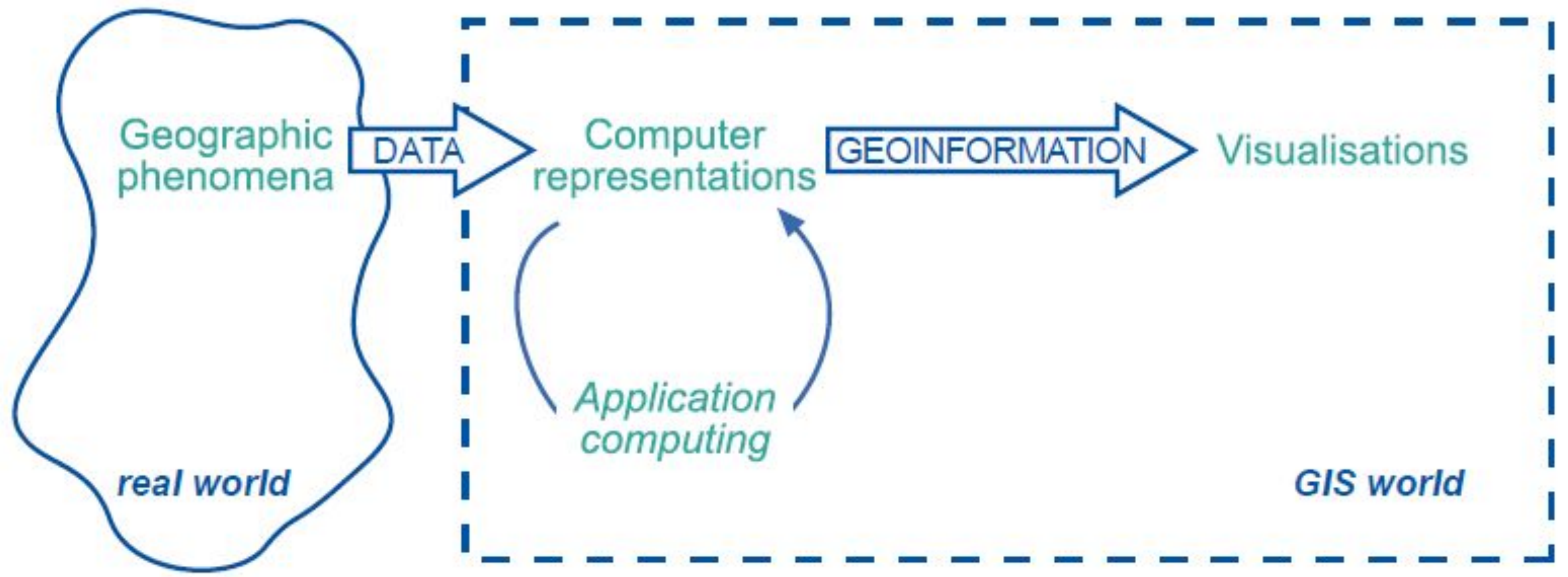
Unit 2

Topics to Be Covered

- Models and representations of the real world
- Geographic Phenomena
- Computer representation of geographic phenomena
- Topology
- Organizing and managing spatial data
- Temporal Dimension
- Data Types
- Raster data structures

Models and representations of the real world

- GIS helps us to understand more about processes and phenomena in real world.
- **Modelling** is a process of representing objects which has certain characteristics in common with real world.
- In other words, **Modelling** is a process of producing an abstraction of the real world so that some part of it can be more easily handled.


Geographic phenomena
DATA
Computer representations
GEOINFORMATION
Visualisations
Application computing
real world
GIS world

- In practical terms, this refers to the process of representing key aspects of the <u>real world digitally in a computer</u>.
- These representation are made up of spatial data, stored in memory, in the form of bits and bytes, on media such as hard drive of the computer.
- This digital representation can then be subjected to various analytical functions in GIS, and the output can be visualized.
- In order to understand both our representation of phenomena and our eventual output from any analysis, we can use the GIS to create **visualisations** from the computer representation either on screen or printed on paper

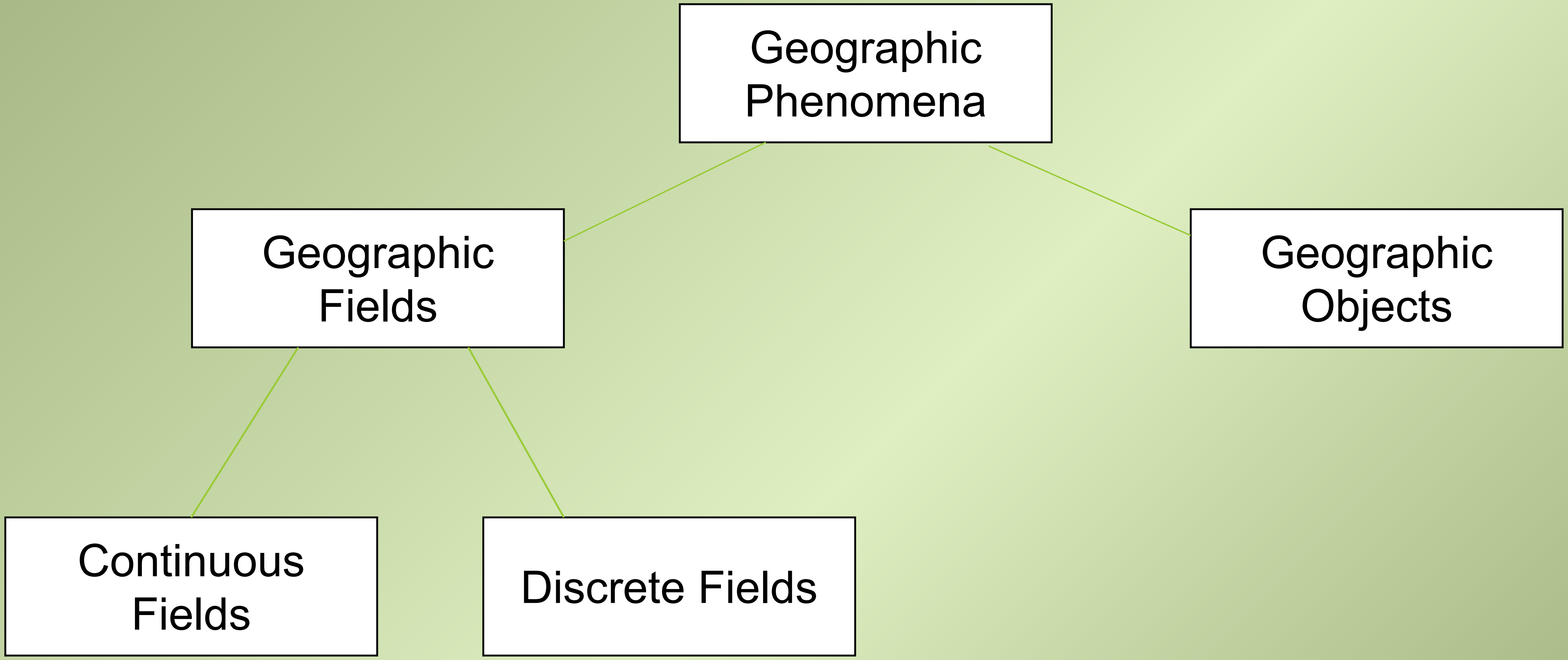
Geographic Phenomena

A geographic phenomena can be defined as manifestation of an entity or process of interest that:

- Can be named or described
- Can be georeferenced (location)
- Can be assigned a time interval.

Geographic phenomena can be defined using triplet: *(description, georeferenced, time interval)*

Types of Geographic Phenomena



Geographic Fields

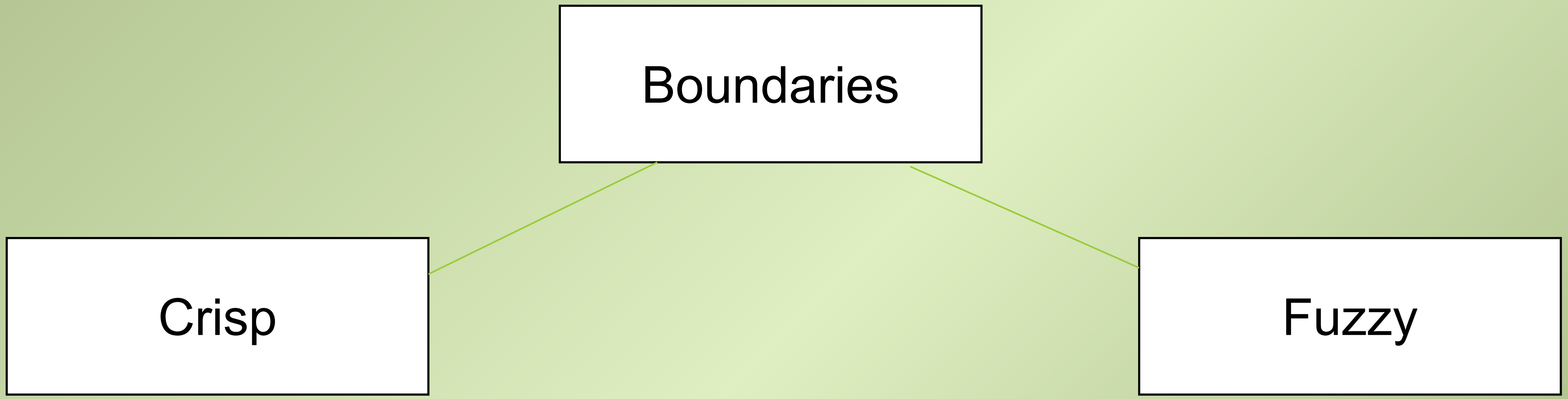
- It is a geographic phenomena for which every point in the study area, a value can be determined.
- It can be considered as a mathematical function “f” that associates a specific value with a particular position in study area.
- Types:
  - Continuous Field
  - Discrete Field

Geographic Objects

- When a geographic phenomena is not present everywhere in the study area but it is sparsely distributed, we call it as Geographic objects.
- An object can be defined using 4 parameters:
  - Location
  - Shape
  - Size
  - Orientation

Boundary

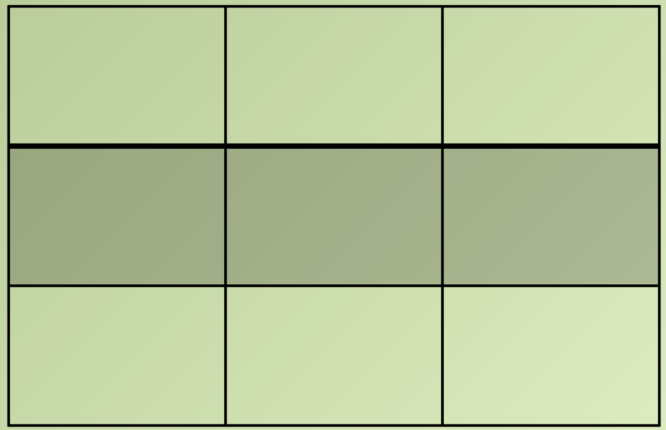
- Boundary is used to separate or define a particular shape or size of an object or field.

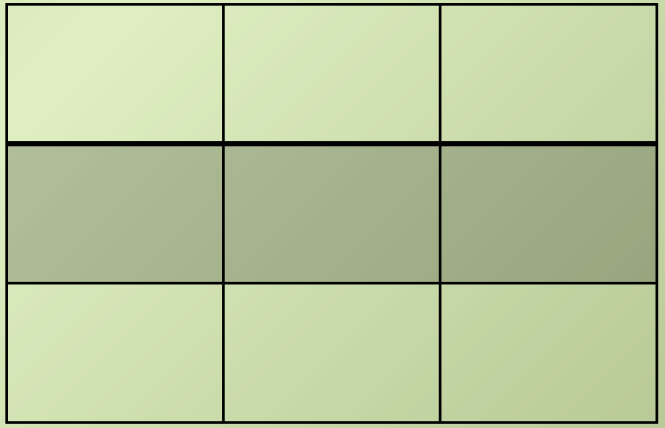


- Crisp Boundaries: A crisp boundary is one that can be determined with almost arbitrary precision, dependent only on data acquisition technique applied.
- Fuzzy boundary: In this, boundary is not a crisp or a precise line, but it is rather an area of transition itself.

Computer representation of geographic phenomena

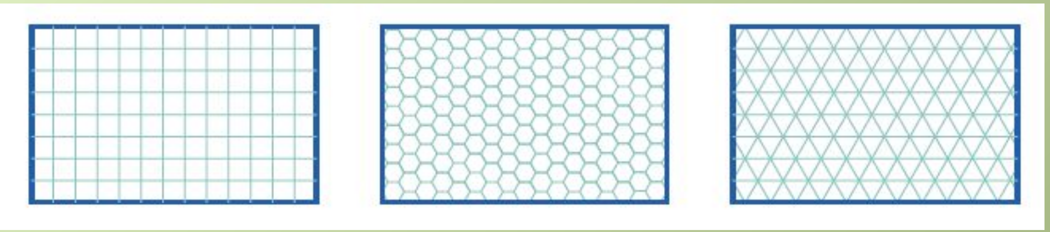
- Fields – Tessellation Approach
- Objects – Vector (Topological) Approach



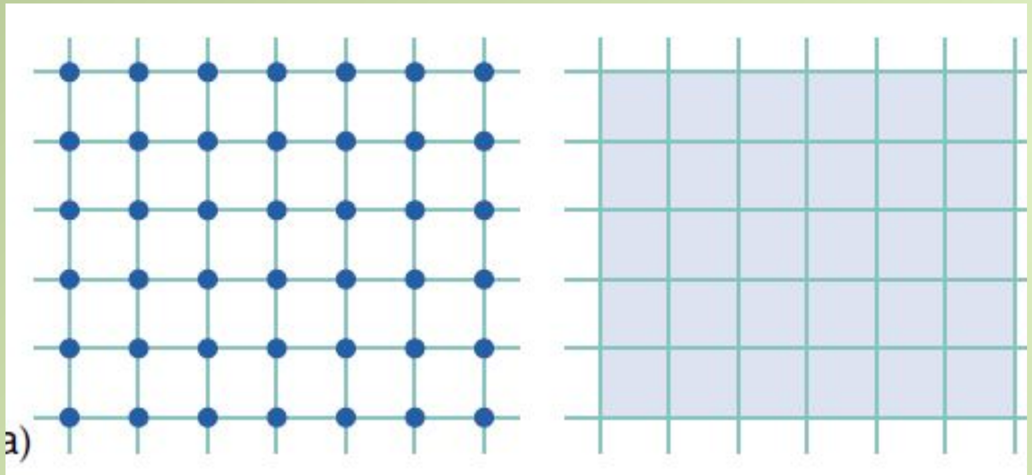


Tessellation Approach

- It is a process of partitioning space into mutually exclusive cells that together make up the complete study space.
- With each cell, some value is associated to characterize that part of space.
- Two types – Regular and Irregular Tessellation Approach
- **Regular Tessellation**: The cells are of same size and shape
  - Square cell
  - Hexagonal cell
  - Triangular cell



- A raster is set of regularly spaced cells with associated field values.
- <u>The associated values are called as the cell values and NOT point values</u>



- The size of the area that a single raster cell can represent is called as raster’s resolution.
- **<u>Irregular Tessellations</u>**: QuadTree

Vector (Topological) Approach

- Tessellations do not store explicitly store georeferences of the phenomena they represent
- In vector representation, an attempt is made to store explicitly the georeferences of the phenomena
- A georeference is a coordinate pair from geographical space, and is also known as “Vector”.
  - Point
  - Line
  - Polygon (Area)

Point representation

- Points are defined as a single coordinate pair (x,y) when we work in 2D, and (x,y,z) when we work in 3D.
- Points are used to describe only the LOCATION of a place along with its characteristics, generally considered as ZERO dimensional.

Line representation

- Line data are used to represent ONE dimensional objects such as roads, railways, etc.
- Lines are used to describe the LOCATION of a place along with SIZE (Length)

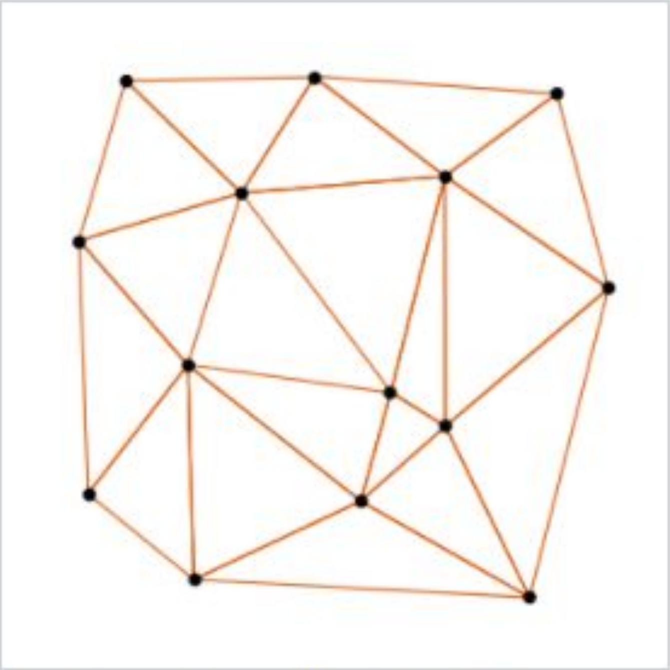
Polygon (Area) representation

- Polygon data are used to represent TWO dimensional objects such as grounds, parks etc.
- Polygons are used to describe the LOCATION of a place along with SIZE (Area) and SHAPE.

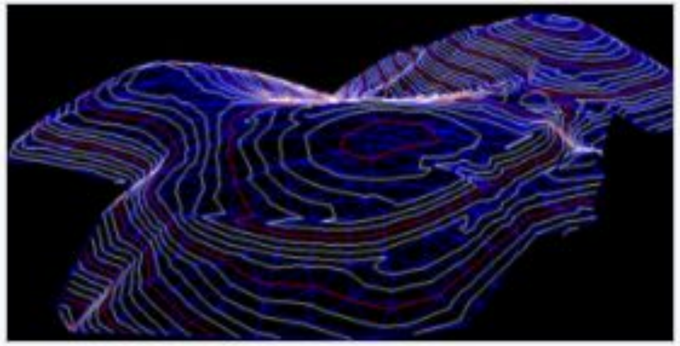
Point	0D	Location	Bus stop, hospital, hotels, restaurant, school, colleges, police stn, fire stn, temple, malls, showroom, airports, shops, post office, banks, telephone exchange, gyms, atm, etc.
Line	1D	Location and Size	Railway, Roads, river, highway, drainage, subways, etc
Polygon	2D	Location, Size and Shape	Garden, building, Area, lake, forest, Play ground, Dumping ground, stadium, parking zone, farms, police training ground, military,

TIN MODEL

- Triangulated Irregular Networks (TIN) is a representation of geographic fields that can be considered as hybrid between tessellations and vector representations
- It is one of the most standard used implementation techniques for digital terrain(land) models, but it can be used to represent any continuous field.
- It is built from a set of locations for which we have a value, for example, an elevation.
- Any location together can be with its elevation value can be viewed as a point in 3D space, it can be scattered in space.



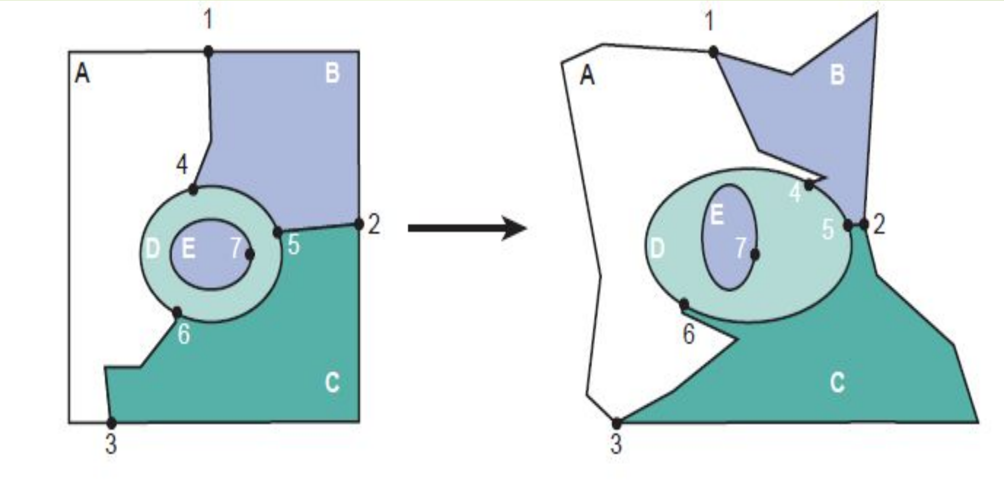
Triangulated irregular network



TIN overlaid with contour lines

Topology

- Topology deals with spatial properties that do not change under certain transformations

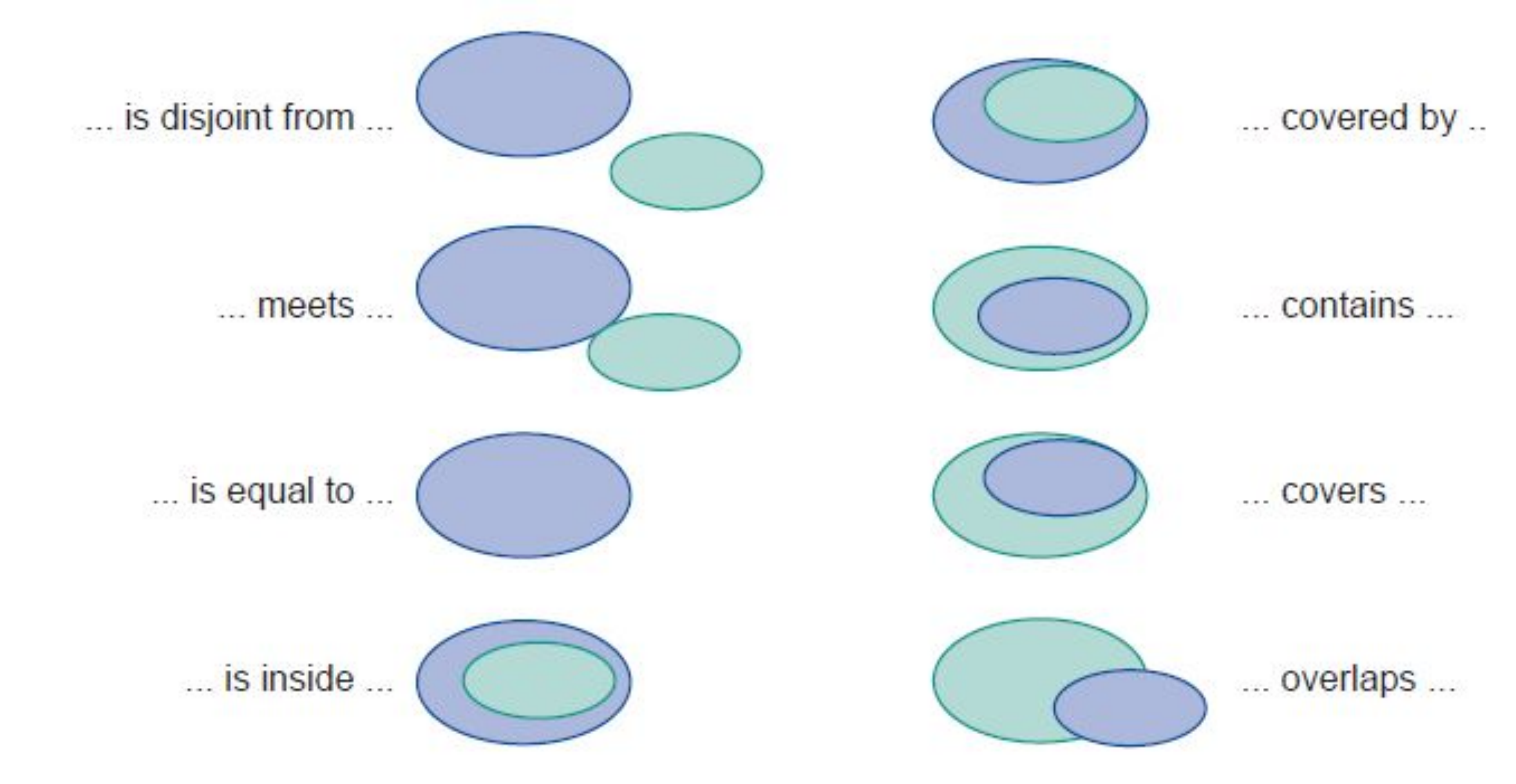


- For example, features drawn on a sheet of rubber can be made to change in shape and size by stretching and pulling the sheet.
- However, some properties of it does not change:
  - Area E is still inside Area D
  - Neighborhood relationship between A, B, C, D stay intact.
  - Area are still bounded by same boundary

Topological relationship

- They are built from simple elements into more complex elements: points define line segments, and line segments connect to define lines, which in turn define polygons.
- The space is a three-dimensional “Euclidean Space” where for every point we define its three-dimensional coordinates as a triple (x, y, z) of real numbers.
- The space is a “Metric Space”, which means that we can always compute the distance between two points
- The space is a “Topological Space”, which means, for every point in the space, we can find neighbourhood for it.

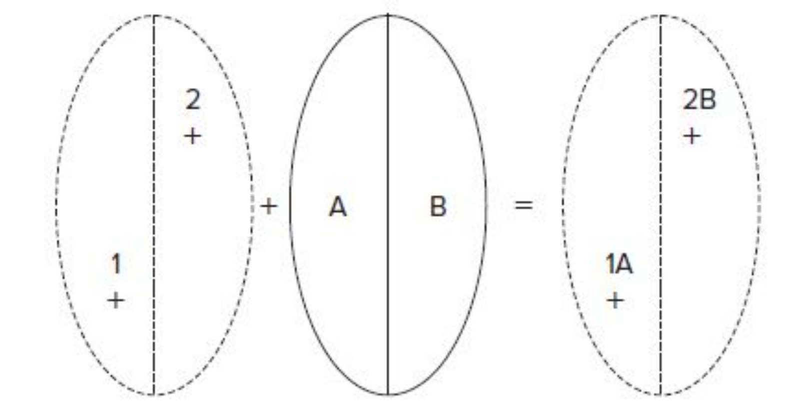
Spatial Relationship



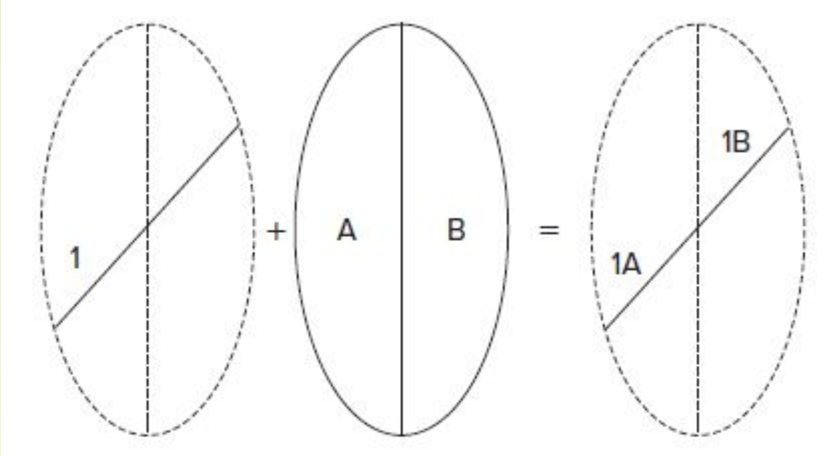
Organizing and Managing Spatial Data

- The main principle of data organization applied in GIS systems is that of a spatial data layer
- A <u>Spatial Data Layer</u> is either a representation of continuous field or discrete field or a collection of objects of the same kind.
- Usually, data is organized so that similar data is in same layer.
- For example, all telephone booth points are in same data layer, all roads are in a layer and so on.

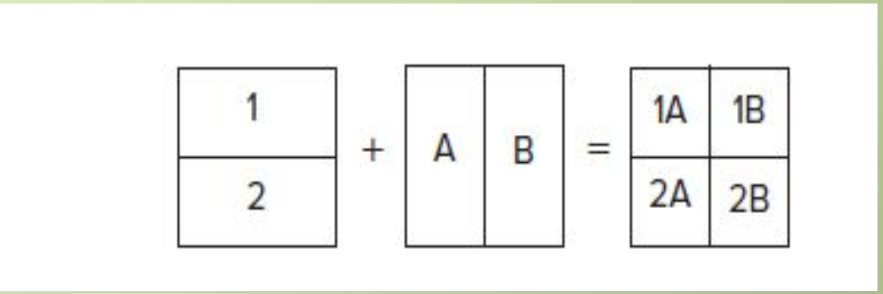
- An **<u>overlay operation</u>** combines the geometries and attributes of two feature layers to create the output.
- The geometry of the output represents the geometric intersection of features from the input layers.
- Overlay Operations:
  - point-in-polygon
  - line-in-polygon, and
  - polygon-on-polygon
- In a <u>point-in-polygon overlay operation</u>, the same point features in the input layer are included in the output but each point is assigned with attributes of the polygon within which it falls.



- In a line-in-polygon overlay operation, the output contains the same line features as in the input layer but each line feature is dissected by the polygon boundaries on the overlay layer. Thus the output has more line segments than does the input layer.



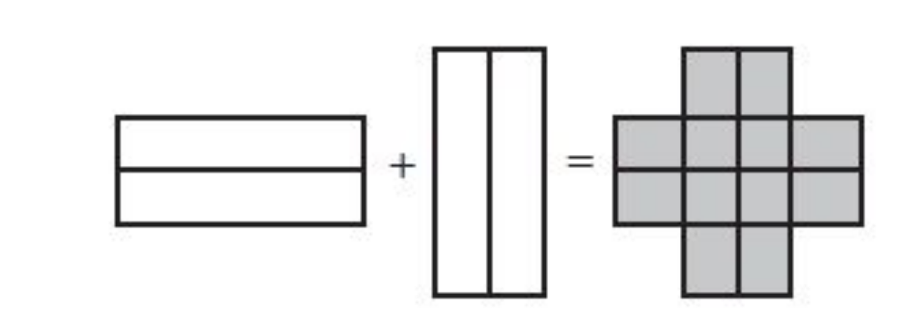
- The most common overlay operation is polygon-on-polygon, involving two polygon layers. The output combines the polygon boundaries from the input and overlay layers to create a new set of polygons



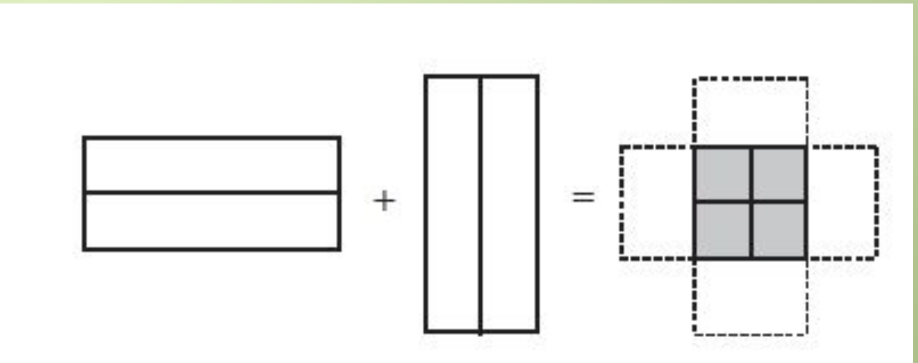
Overlay Methods

Union
Intersect
Symmetric Difference
identity

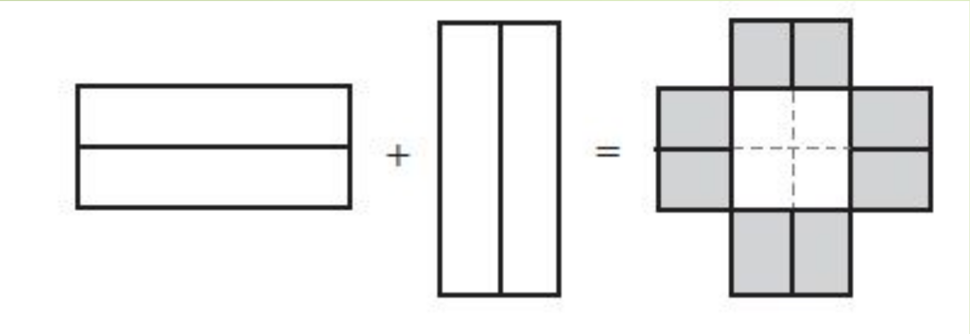
- **<u>Union</u>** preserves all features from the inputs. The area extent of the output combines the area extents of both input layers.



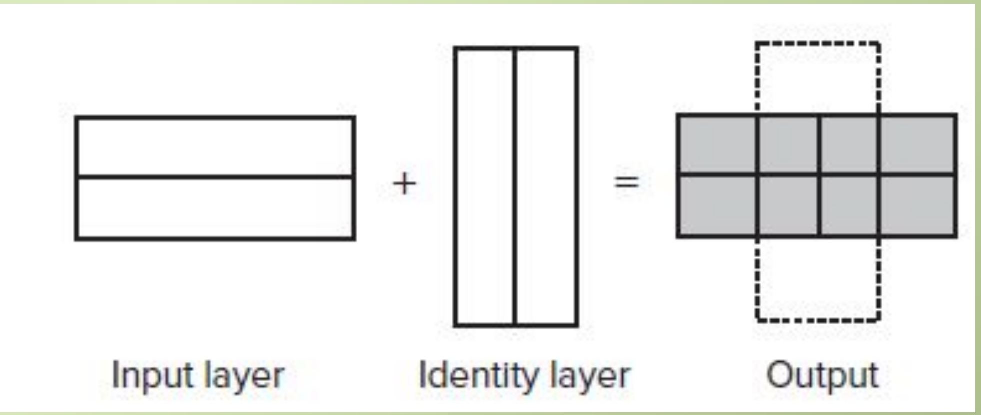
- **<u>Intersect</u>** preserves only those features that fall within the area extent common to the inputs. Intersect is often a preferred method of overlay because any feature on its output has attribute data from both of its inputs



- **Symmetrical Difference** preserves features that fall within the area extent that is common to only one of the inputs. Symmetrical Difference is opposite to Intersect in terms of the output's area extent.



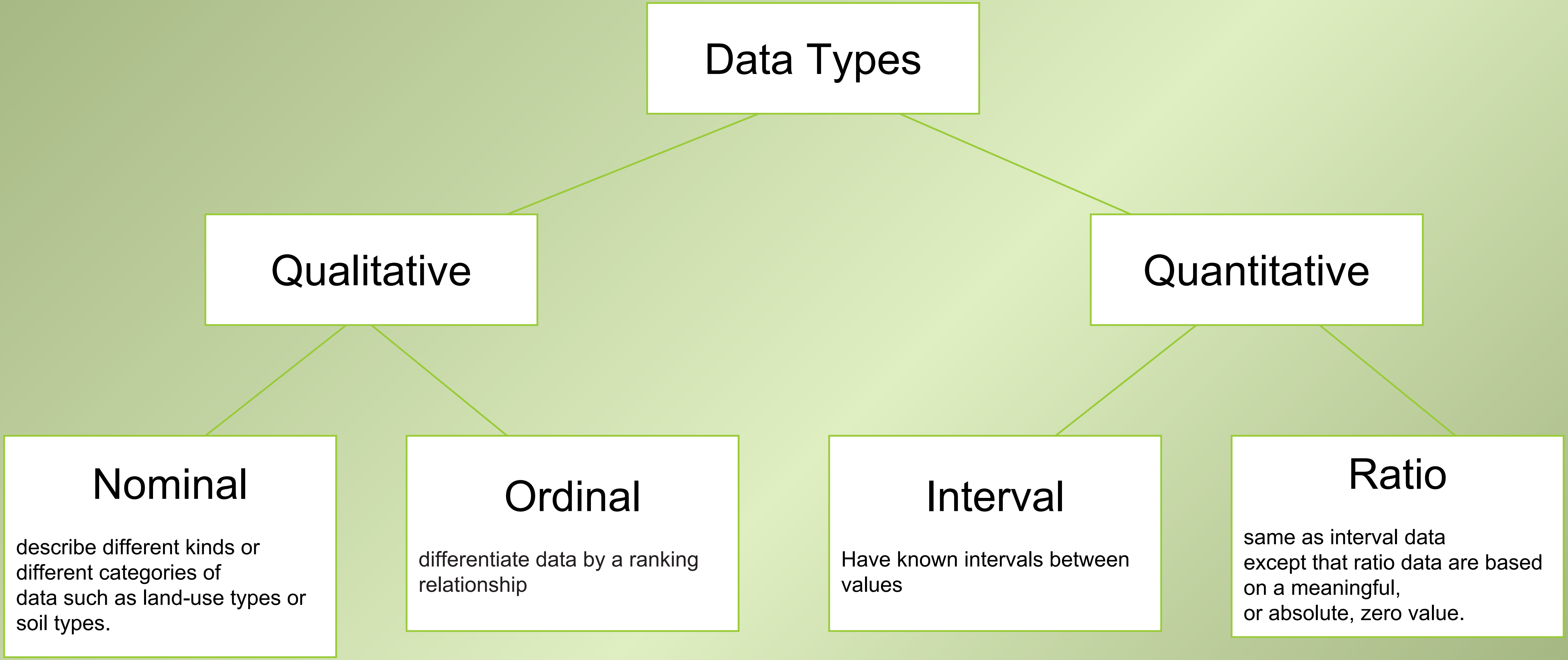
- **Identity** preserves only features that fall within the area extent of the layer defined as the input layer. The other layer is called the identity layer.



Temporal Dimension

- Besides having geometric and topological properties, geographic phenomena are also “dynamic” in nature.
- In other words, geographic phenomena “change” over time.
- The different types of time are:
  - Discrete and Continuous: Discrete time is composed of discrete element such as minutes, seconds, and Continuous time consists of before, overlap and after
  - Valid time and Transaction time: Valid time is the time when an event really happened, and Transaction time is when the event was stored in the database

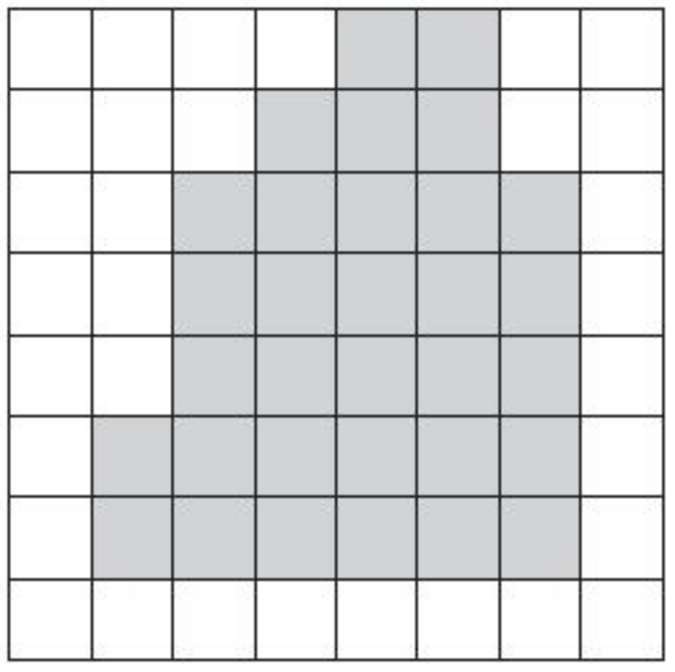
Data Types


Data Types
Qualitative
Quantitative
Nominal
describe different kinds or different categories of data such as land-use types or soil types.
Ordinal
differentiate data by a ranking relationship
Interval
Have known intervals between values
Ratio
same as interval data except that ratio data are based on a meaningful, or absolute, zero value.

Raster Data Structure

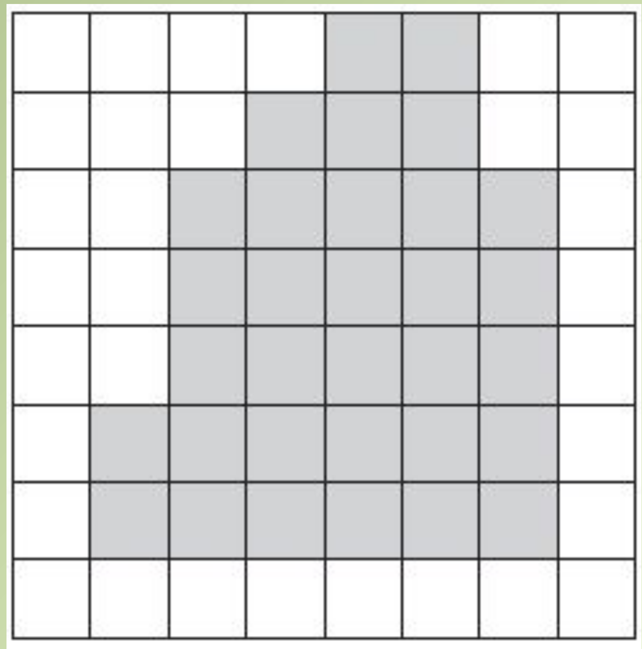
- Raster data structure refers to the method by which raster data are encoded and stored in the computer.
- Three common methods are examined here:
  - cell-by-cell encoding,
  - run-length encoding, and
  - quadtree.

- Cell-by-Cell Encoding
  - The cell-by-cell encoding method provides the simplest raster data structure.
  - A raster is stored as a matrix, and its cell values are written into a file by row and column.



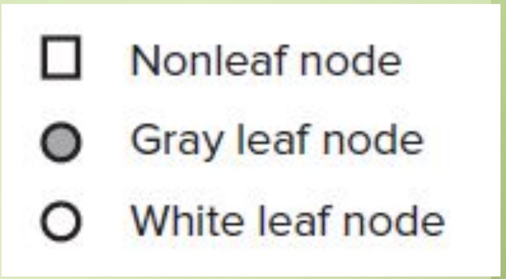
Row 1: 0 0 0 0 1 1 0 0
Row 2: 0 0 0 1 1 1 0 0
Row 3: 0 0 1 1 1 1 1 0
Row 4: 0 0 1 1 1 1 1 0
Row 5: 0 0 1 1 1 1 1 0
Row 6: 0 1 1 1 1 1 1 0
Row 7: 0 1 1 1 1 1 1 0
Row 8: 0 0 0 0 0 0 0 0

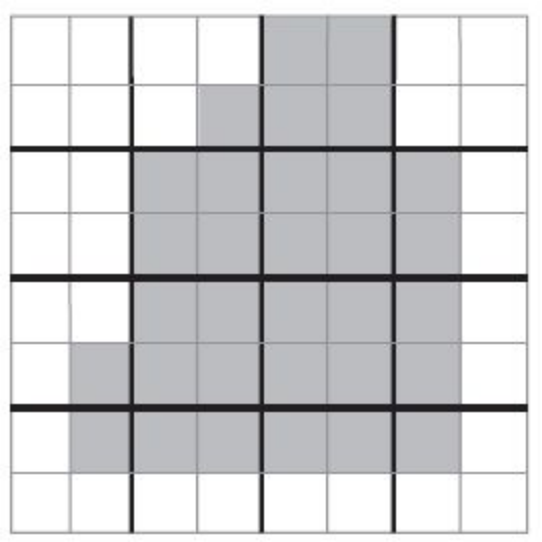
- Run-Length Encoding
  - Raster data with many repetitive cell values can be more efficiently stored using the run-length encoding (RLE) method, which records the cell values by row and by group.
  - A group refers to a series of adjacent cells with the same cell value.

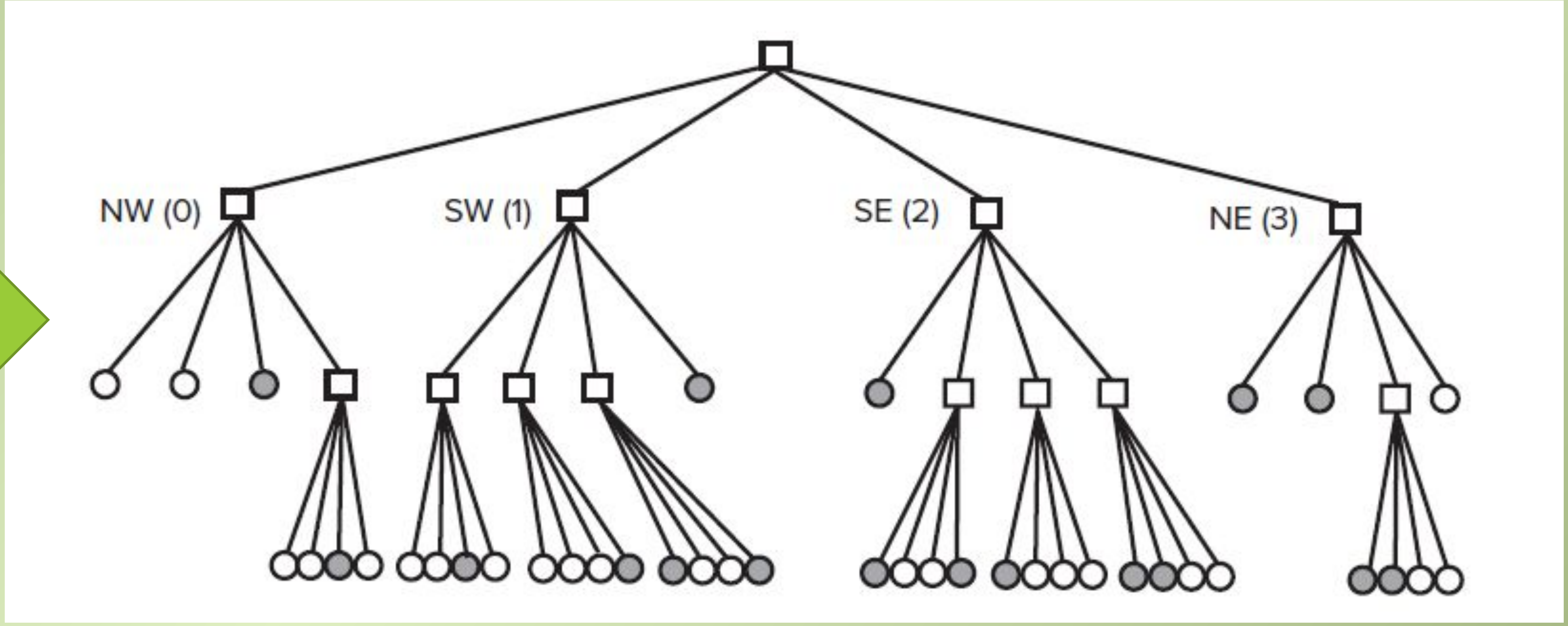


Row 1 : 5 6
Row 2: 4 6
Row 3: 3 7
Row 4: 3 7
Row 5: 3 7
Row 6: 2 7
Row 7: 2 7

- Quadtree
  - Instead of working along one row at a time, quadtree uses recursive decomposition to divide a raster into a hierarchy of quadrants.
  - Recursive decomposition refers to a process of continuous subdivision until every quadrant in a quadtree contains only one cell value.







(02, 032), (102, 113, 120, 123, 13), (20, 210, 213, 220, 230, 231), (30, 31, 320, 321)