Data exploration, query and vector data analysis

Module 4

Topics to be covered

Chap: Data exploration, query

- Exploration
- Attribute data query
- Spatial data query
- Raster data query
- Geographic visualization.

Chap: Vector data analysis

- Introduction
- Buffering
- Map overlay
- Distance measurement and
- Map manipulation.

Data Exploration

- Data exploration has its origin in statistics.
- Statisticians have traditionally used graphic techniques and descriptive statistics to examine data prior to more formal and structured data analysis
- This kind of data exploration has been a component of data visualization, the discipline of using a variety of exploratory techniques and graphics to understand and gain insight into the data.

Descriptive Statistics

Graphs

Descriptive Statistics

- Descriptive statistics summarize the values of a data set.
- Assuming the data set is arranged in the ascending order, the different statistics that can be performed are:
  - The range is the difference between the minimum and maximum values.
  - The median is the midpoint value, or the 50th percentile.
  - The first quartile is the 25th percentile.
  - The third quartile is the 75th percentile.
  - The mean is the average of data values.
  - The variance is a measure of the spread of the data about the mean.

Graphs

- Different types of graphs are used for data exploration.
- A graph may involve a single variable or multiple variables, and it may display individual values or classes of values.
- A **line graph** displays data as a line.
- A **bar chart**, also called a histogram, group data into equal intervals and uses bars to show the number or frequency of values falling within each class. A bar chart may have vertical bars or horizontal bars.
- A **cumulative distribution graph** is one type of line graph that plots the ordered data values against the cumulative distribution values

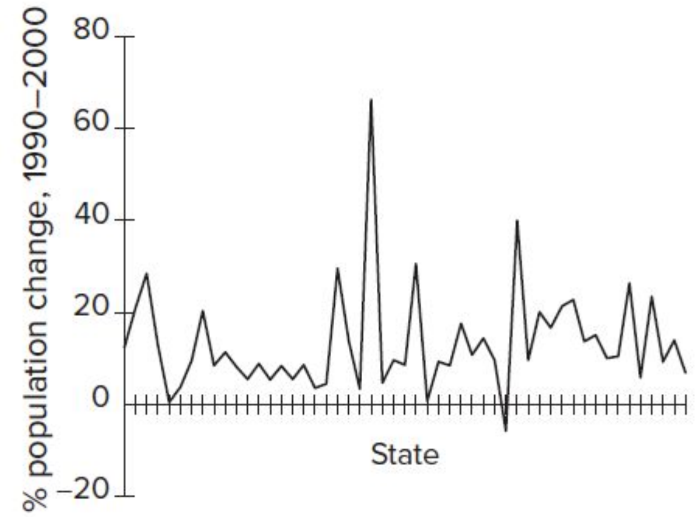


Figure 10.1
A line graph.

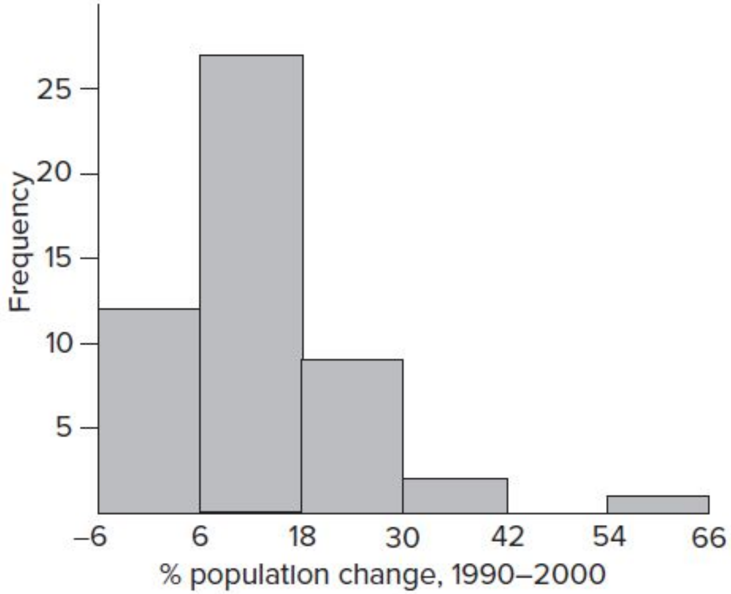


Figure 10.2
A histogram (bar chart).

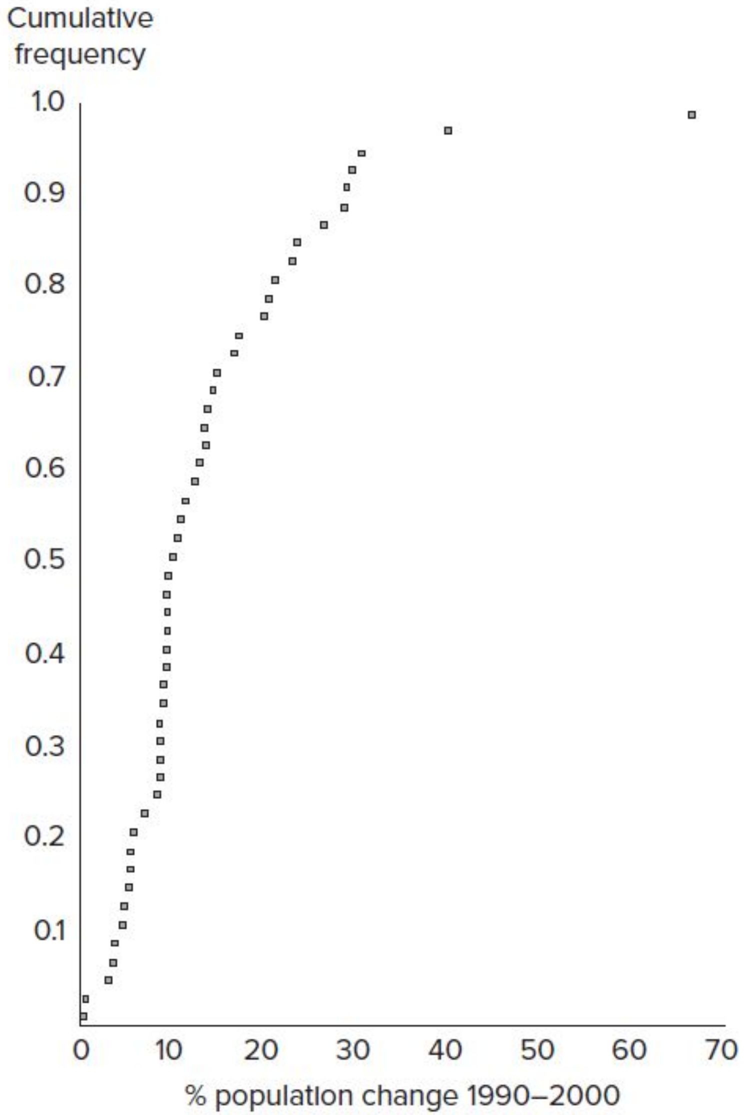
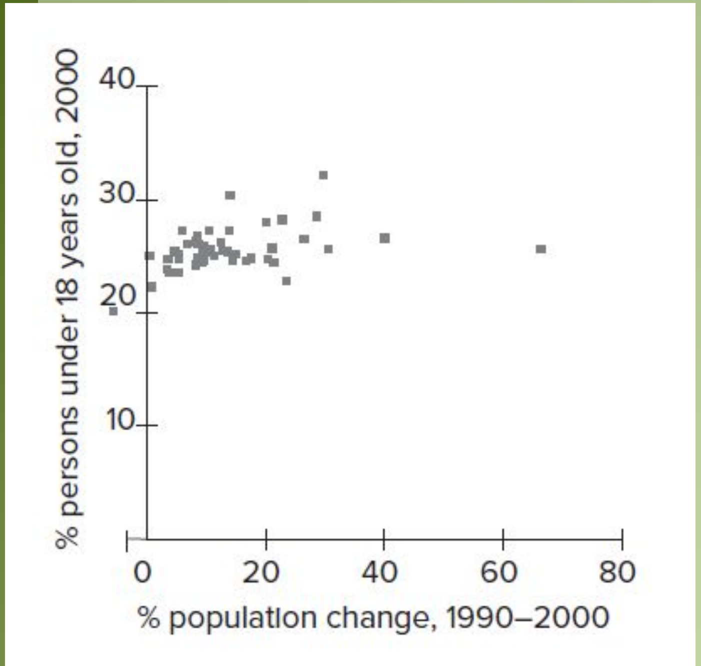
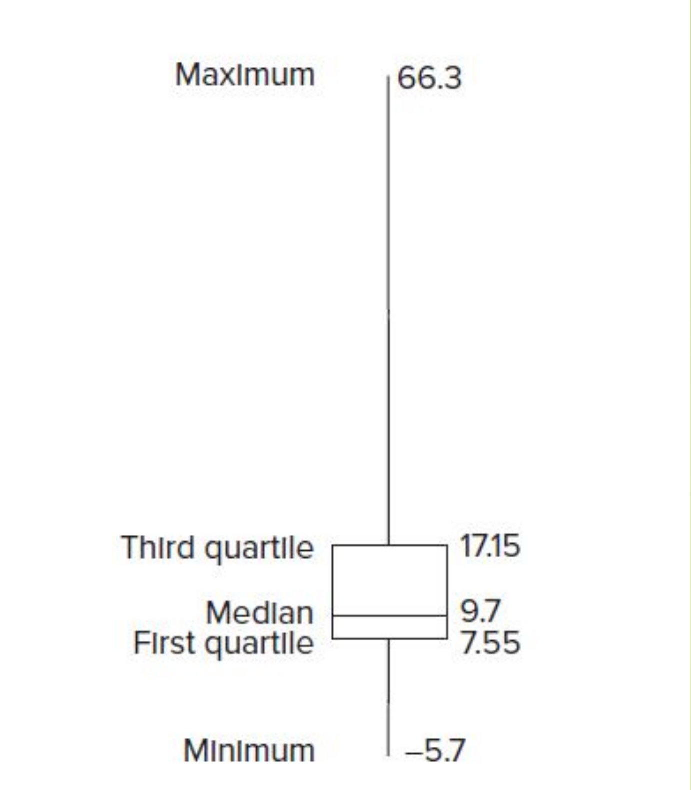
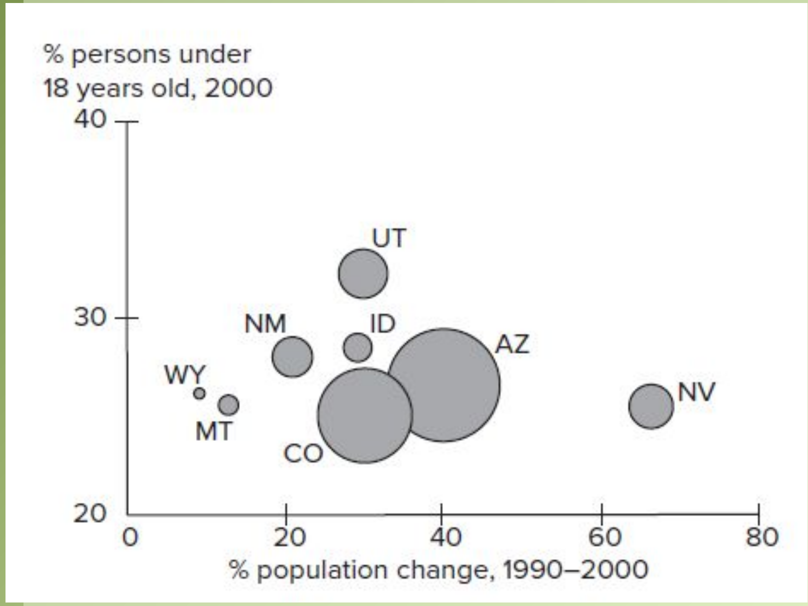


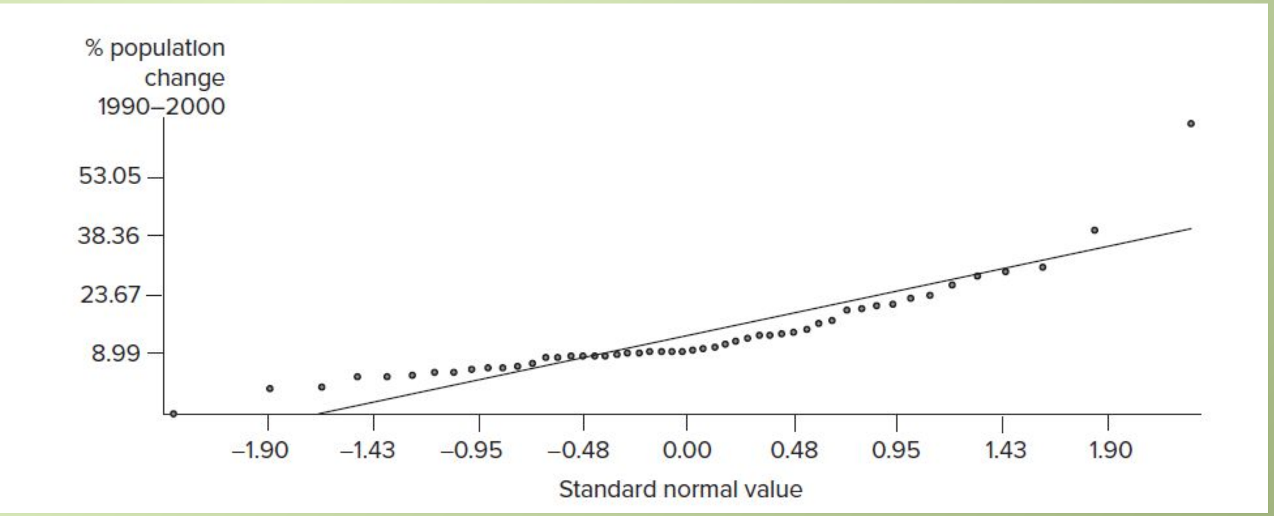
Figure 10.3
A cumulative distribution graph.

- A **scatterplot** uses markings to plot the values of two variables along the x- and y-axes.
- **Bubble plots** are a variation of scatterplots. Instead of using constant symbols as in a scatterplot, a bubble plot has varying-sized bubbles that are made proportional to the value of a third variable.
- **Boxplots**, also called “box and whisker” plots, summarize the distribution of five statistics from a data set: the minimum, first quartile, median, third quartile, and maximum.
- **Quantile–quantile plots**, also called QQ plots, compare the cumulative distribution of a data set with that of some theoretical distribution such as the normal distribution, a bell-shaped frequency distribution

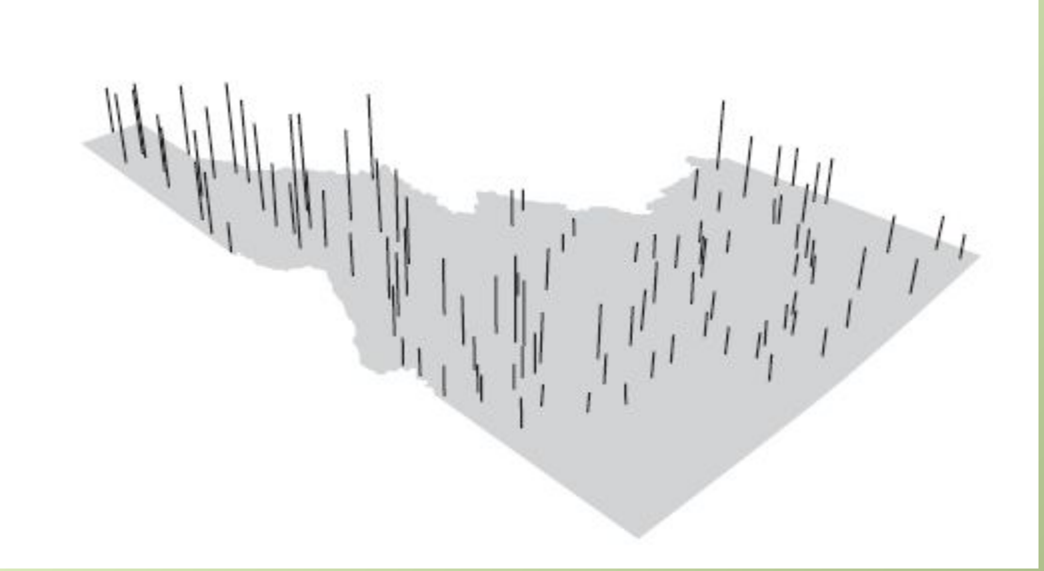

% persons under 18 years old, 2000
40
30
20
10
0
20
40
60
80
% population change, 1990–2000


Maximum
66.3
Third quartile
17.15
Median
9.7
First quartile
7.55
Minimum
–5.7


% persons under
18 years old, 2000
40
30
20
UT
NM
ID
AZ
WY
MT
CO
NV
0
20
40
60
80
% population change, 1990–2000


% population
change
1990–2000
53.05
38.36
23.67
8.99
–1.90
–1.43
–0.95
–0.48
0.00
0.48
0.95
1.43
1.90
Standard normal value

- Some graphs are designed for spatial data.
- Following figure, for example, shows a plot of spatial data values by raising a bar at each point location so that the height of the bar is proportionate to its value.
- This kind of plot allows the user to see the general trends among the data values in both the x-dimension (east–west) and y-dimension (north–south).

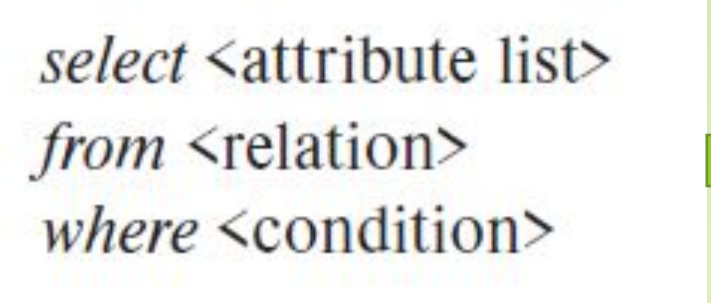


Attribute data query

- Attribute data query retrieves a data subset by working with attribute data.
- The selected data subset can be simultaneously examined in the table, displayed in charts, and linked to the highlighted features in the map.
- The selected data subset can also be printed or saved for further processing.
- Attribute data query requires the use of expressions, which must be interpretable by a database management system.

SQL

- SQL is a data query language designed for manipulating relational databases.
- For GIS applications, SQL is a command language for a GIS (e.g., QGIS) to communicate with a database.
- To use SQL to access a database, we must follow the structure (i.e., syntax) of the query language. The basic syntax of SQL, with the keywords in italic, is:



```
select <attribute list>
from <relation>
where <condition>
```

The **select** keyword selects field(s) from the database,
the **from** keyword selects table(s) from the database, and
the **where** keyword specifies the condition or criterion for data query.

Query Expressions

- Query expressions, or the where conditions, consist of Boolean expressions and connectors.
- A simple Boolean expression contains two operands and a logical operator.
- For example, Parcel.PIN = ‘P101’ is an expression in which PIN and P101 are operands and = is a logical operator.
- Operands may be a field, a number, or a text.
- Logical operators may be equal to (=), greater than (>), less than (<), greater than or equal to (>=), less than or equal to (<=), or not equal to (<>).
- Boolean expressions may contain calculations that involve operands and the arithmetic operators +, −, ×, and /.
- Boolean connectors are AND, OR, XOR, and NOT, which are used to connect two or more expressions in a query statement

Type of Operation

- Attribute data query begins with a complete data set.
- A basic query operation selects a subset and divides the data set into two groups: one containing selected records and the other unselected records.
- Given a selected data subset, three types of operations can act on it:
  - add more records to the subset,
  - remove records from the subset, and
  - select a smaller subset

Spatial data query

- Spatial data query refers to the process of retrieving a data subset by working directly with the geometries of spatial features.
- Feature geometries are stored in a spatial subsystem in the georelational model (e.g., the shapefile) and integrated with attribute data in the object-based model (e.g., the geodatabase)
- Based on a spatial index structure, spatial query can be performed **using graphics or spatial relationships between features.**
- The result of a query can be simultaneously inspected in the map, linked to the highlighted records in the table

Feature Selection by Graphic

- The simplest spatial query is to select a feature by pointing at it or to select features of a layer by dragging a box around them.
- Alternatively, we can use a graphic such as a circle, a box, a line, or a polygon to select features that fall inside, or are intersected by, the graphic object.
- These graphics can be made using the drawing tools or converted from a selected spatial feature.

Feature Selection by Spatial Relationship

- This query method selects features based on their spatial relationships to other features.
- Containment—selects join features that fall within, or are contained by, target features. Examples include finding schools within each county, and national parks within each state.
- Intersect—selects join features that intersect, or are crossed by, target features. Examples include finding urban places that intersect an active fault line and land parcels that are crossed by a proposed road.
- Proximity—selects join features that are close, or adjacent, to target features. Examples include state parks that are within a distance of 10 miles of an interstate highway, land parcels that are adjacent to a flood zone

Raster data query

- Query by Cell Value
- Query by Select Features

Query by Cell Value

- The cell value in a raster represents the value of a spatial feature (e.g., elevation) at the cell location.
- Therefore, to query the feature, we can use the raster itself, rather than a field, in the operand.
- One type of raster data query uses a Boolean statement to separate cells that satisfy the query statement from cells that do not.
- The expression, [road] = 1, queries a road raster that has the cell value of 1.
- The operand [road] refers to the raster and the operand 1 refers to a cell value, which may represent the interstate category

- Raster data query can also use the Boolean connectors of AND, OR, and NOT to string together separate expressions.
- A compound statement with separate expressions usually applies to multiple rasters, which may be integer, or floating point, or a mix of both types.
- For example, the statement, ([slope] = 2) AND ([aspect] = 1), selects cells that have the value of 2 in the slope raster, and 1 in the aspect raster

Query by Select Features

- Features such as points, circles, boxes, or polygons can be used directly to query a raster.
- The query returns an output raster with values for cells that correspond to the point locations or fall within the features for selection.
- Other cells on the output raster carry no data.

Geographic visualization - GIS and Maps

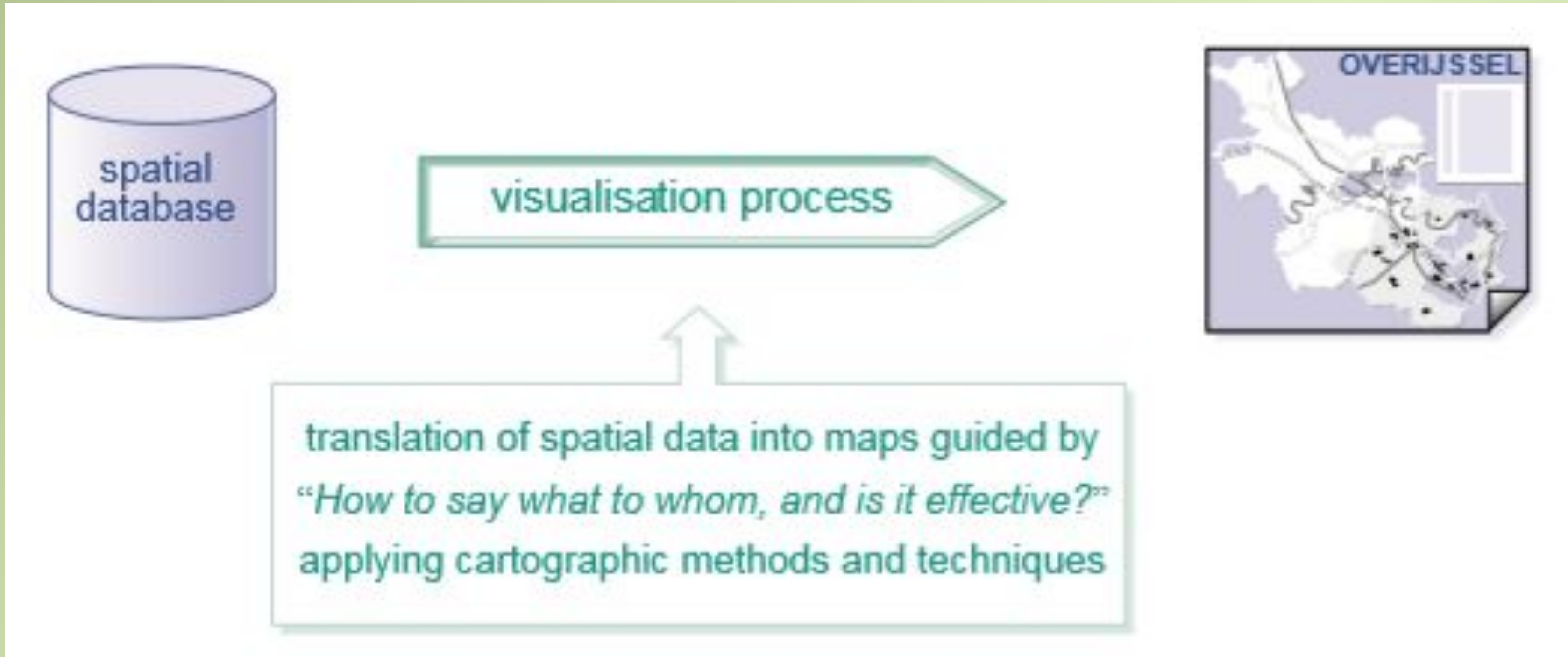
- There is a strong relationship between maps and GIS. More specifically, maps can be used as INPUT to GIS
- They play a key role in relation to all FUNCTIONAL COMPONENT of GIS
- As soon as the question arises of “**WHERE**”, a map can often be the most suitable tool to solve the questions and provide the answers.
- Apart from location, map also contains additional information of a particular location, answering the question, **“WHAT”**
- Maps can also answer the third type of question, **“WHEN”.**
- To summarize, maps can deal with questions related to basic components of geographic data: location, characteristics and time.

MAP SCALE

- **MAP SCALE** is a ratio between distance on the map and the corresponding distance in reality.
- Maps that show much detail of a small area (detailed map) are called as **LARGE-SCALE** maps and vice-versa is called as **SMALL-SCALE** maps
- DEFINITION of Map-" **a representation or abstraction of geographic reality. A tool for presenting geographic information in VISUAL or in DIGITAL FORM**"

Visualization Process

- TRANSLATION OF SPATIAL DATA INTO MAP



translation of spatial data into maps guided by
"How to say what to whom, and is it effective?"
applying cartographic methods and techniques

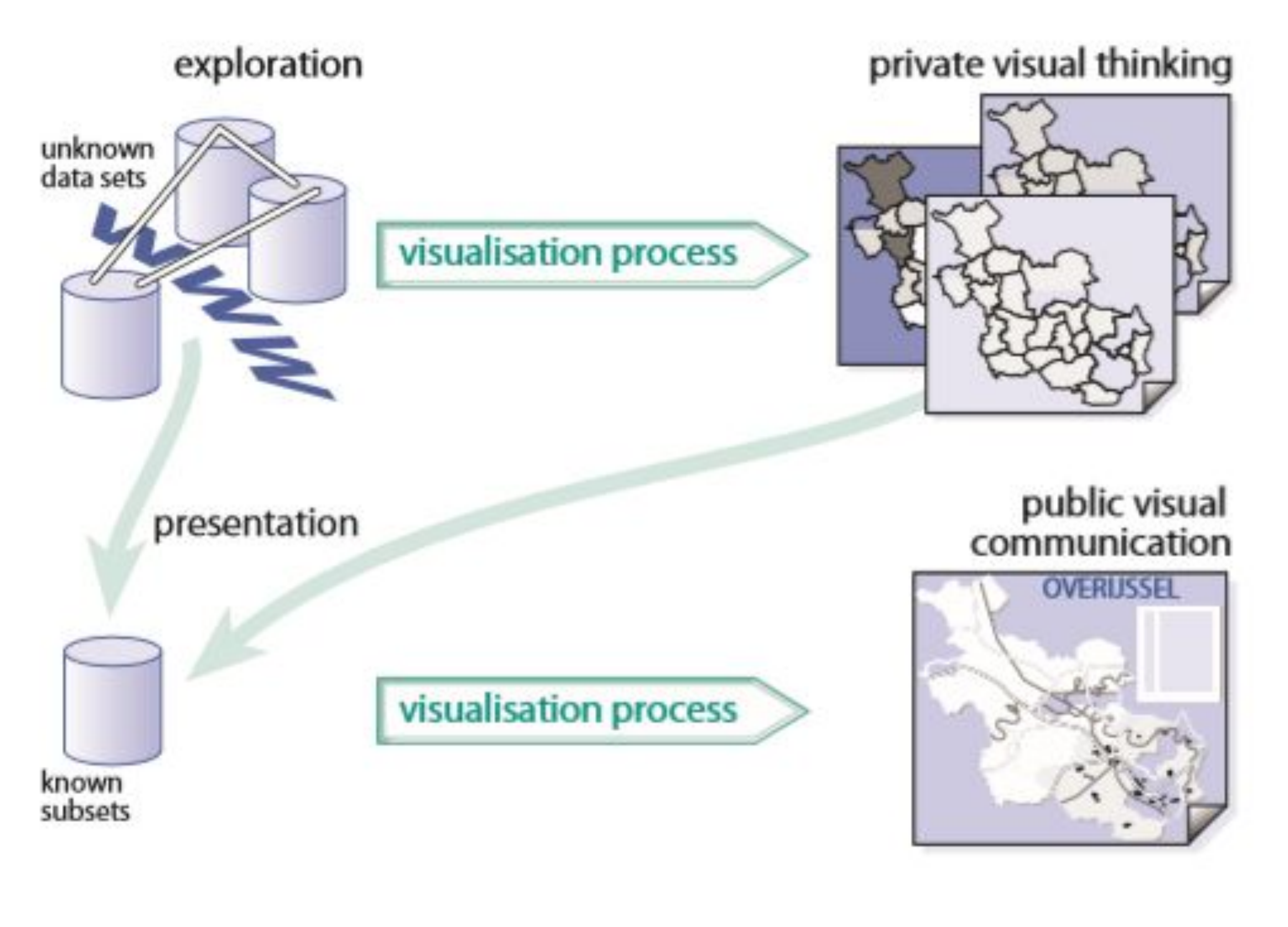
- Influenced by several factors:

1. What will be scale of map?
2. Type of data we are dealing with? (Topographic data or thematic data)
3. Qualitative data or quantitative data?

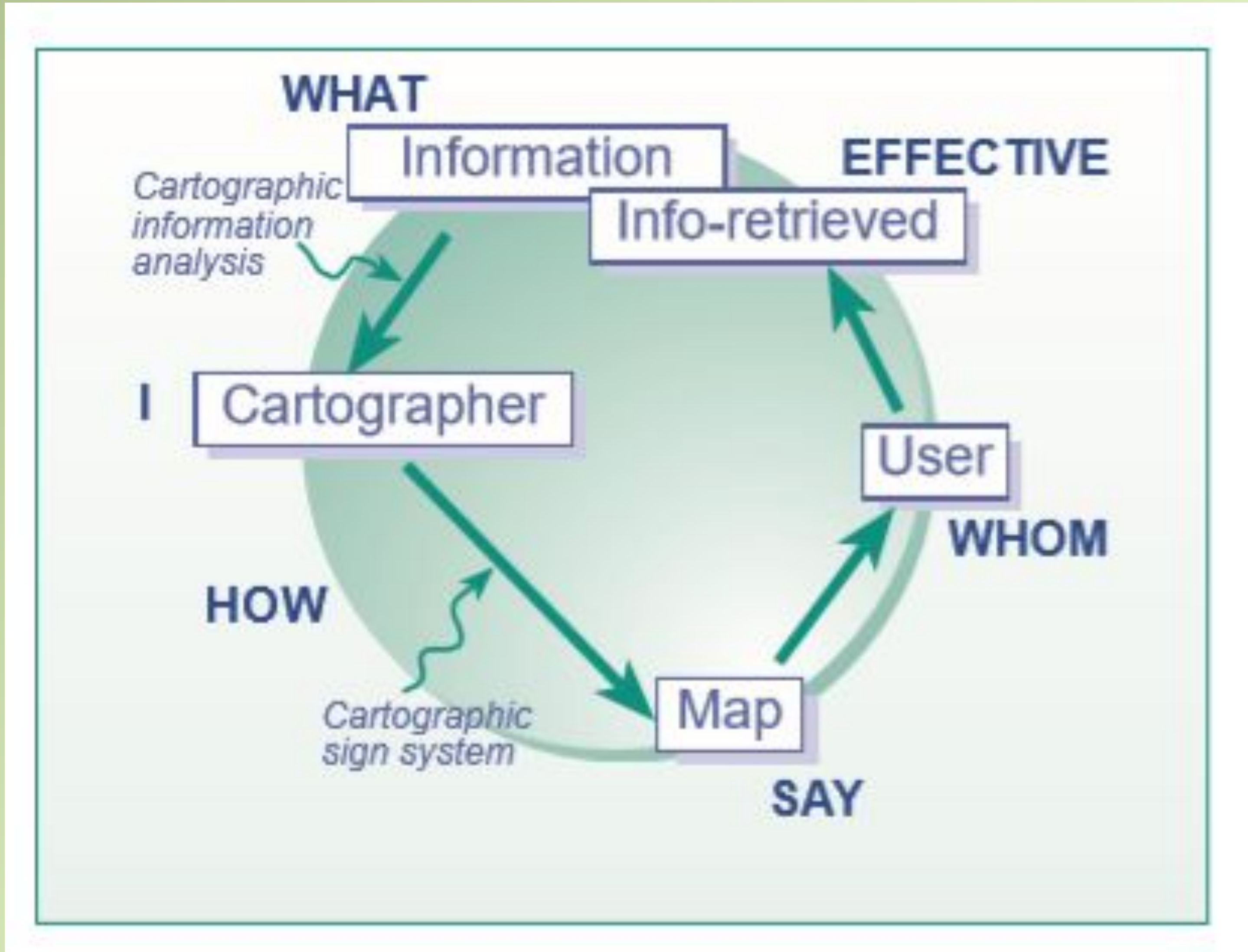
Visualization Strategy

1. Visual Communication: The main function of map is to communicate geographic information i.e. to inform user about location and nature of data.
2. Visual thinking process: Since data is digitally visible too, nowadays, in collaboration with IT, Thinking process as begun with respect to GIS
3. Visual data mining: Many datasets are available on WEB which needs to filtered and retrieve the required information

DEMOCRATIZATION OF CARTOGRAPHY



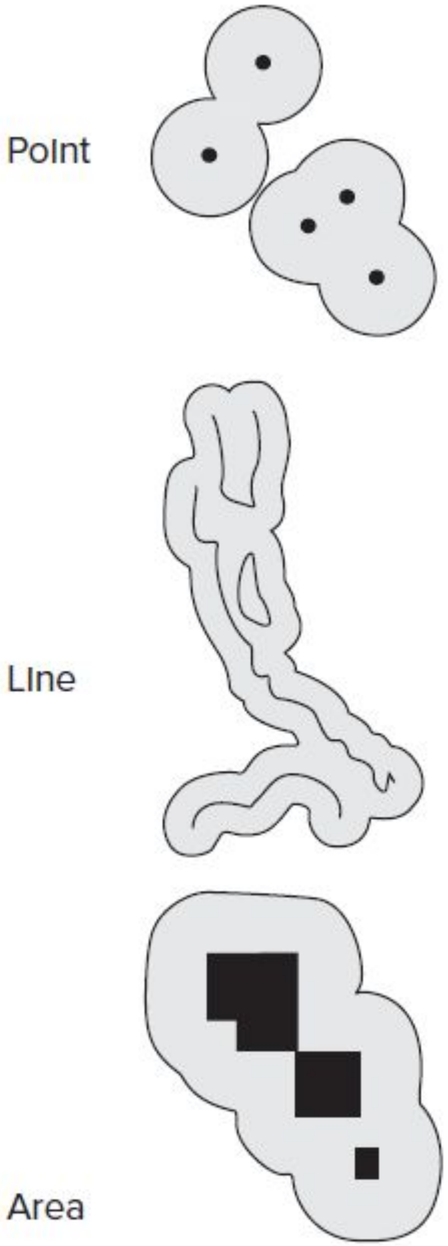
CARTOGRAPHIC / VISUAL COMMUNICATION PROCESS



Chap: Vector data analysis

Buffering

- Based on the concept of proximity, buffering creates two areas: one area that is within a specified distance of select features and the other area that is beyond.
- The area within the specified distance is the **buffer zone**.
- Features for buffering may be points, lines, or polygons.
- Buffering around points creates circular buffer zones.
- Buffering around lines creates a series of elongated buffer zones around each line segment.
- And buffering around polygons creates buffer zones that extend outward from the polygon boundaries.


Point
Line
Area

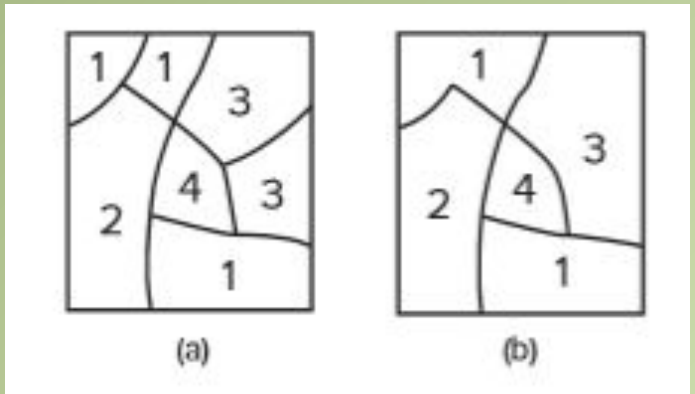
Distance measurement

- Distance measurement refers to measuring straight line distances between features.
- Measurements can be made from points in a layer to points in another layer, or from each point in a layer to its nearest point or line in another layer.
- In both cases, distance measures are stored in a field.
- Distance measures can be used directly for data analysis.

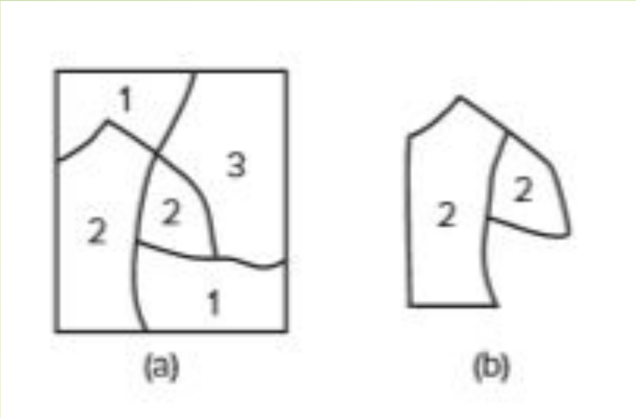
Map manipulation

- Tools are available in a GIS package for manipulating and managing features in one or more feature layers.
- When a tool involves two layers, the layers must be based on the same coordinate system.
- Like overlay, these feature tools are often needed for data preprocessing and data analysis.
- Different Map manipulation tools are as follows:

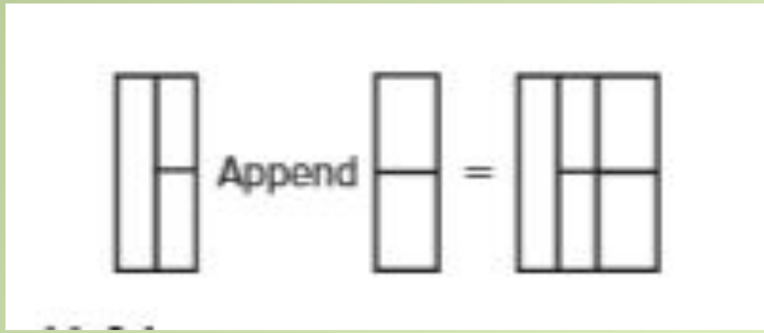
Dissolve



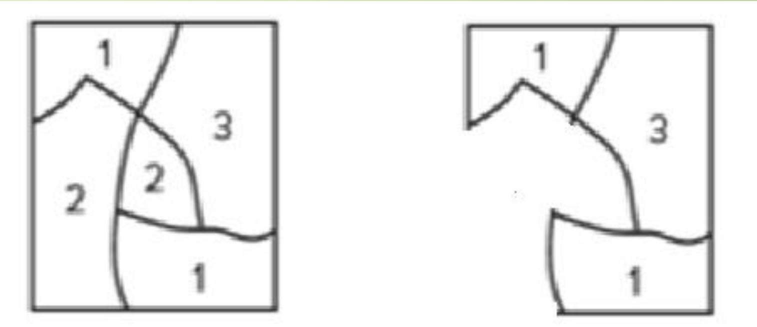
Select



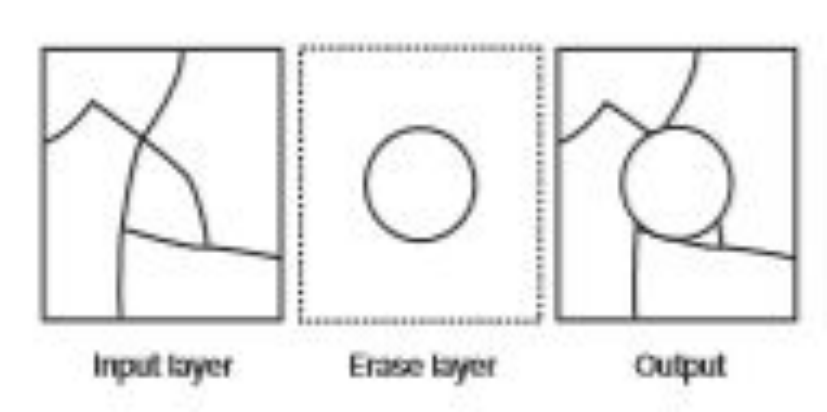
Append



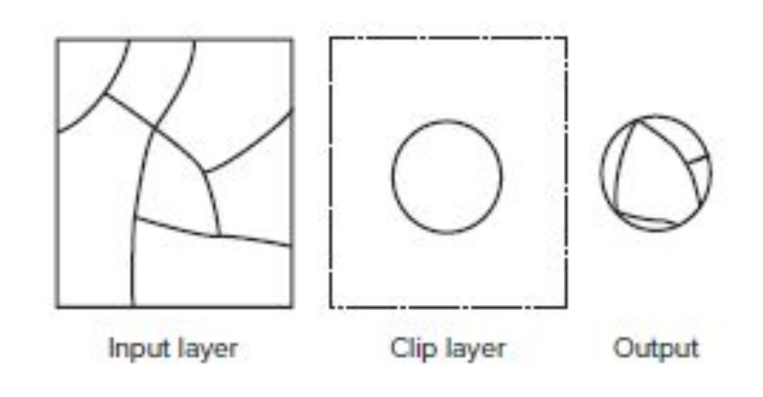
Eliminate



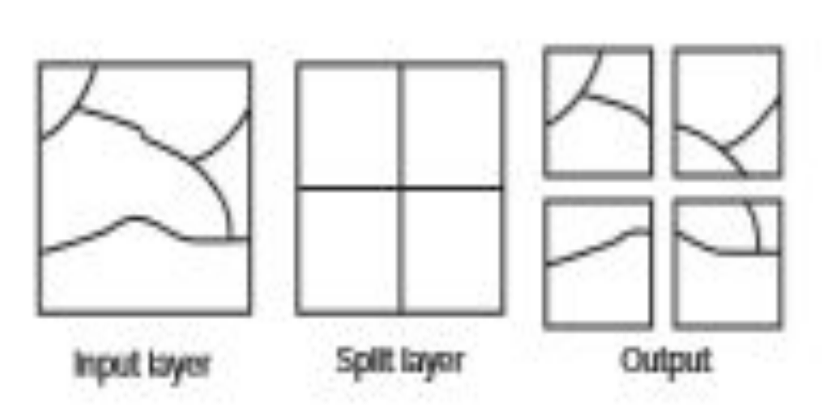
Erase



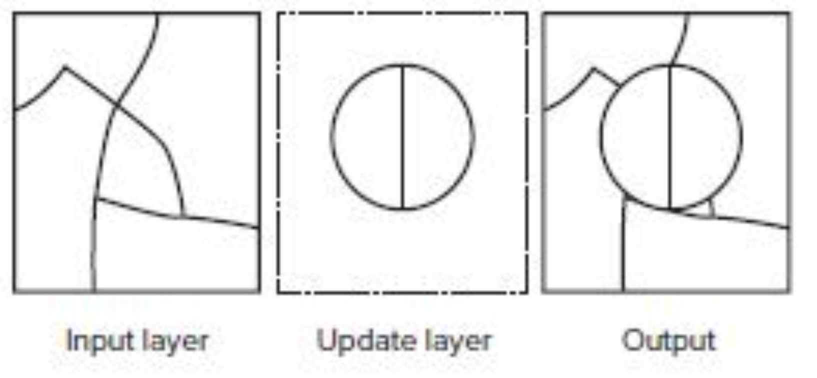
Clipping



Split



Overwrite (update)


Input layer
Update layer
Output