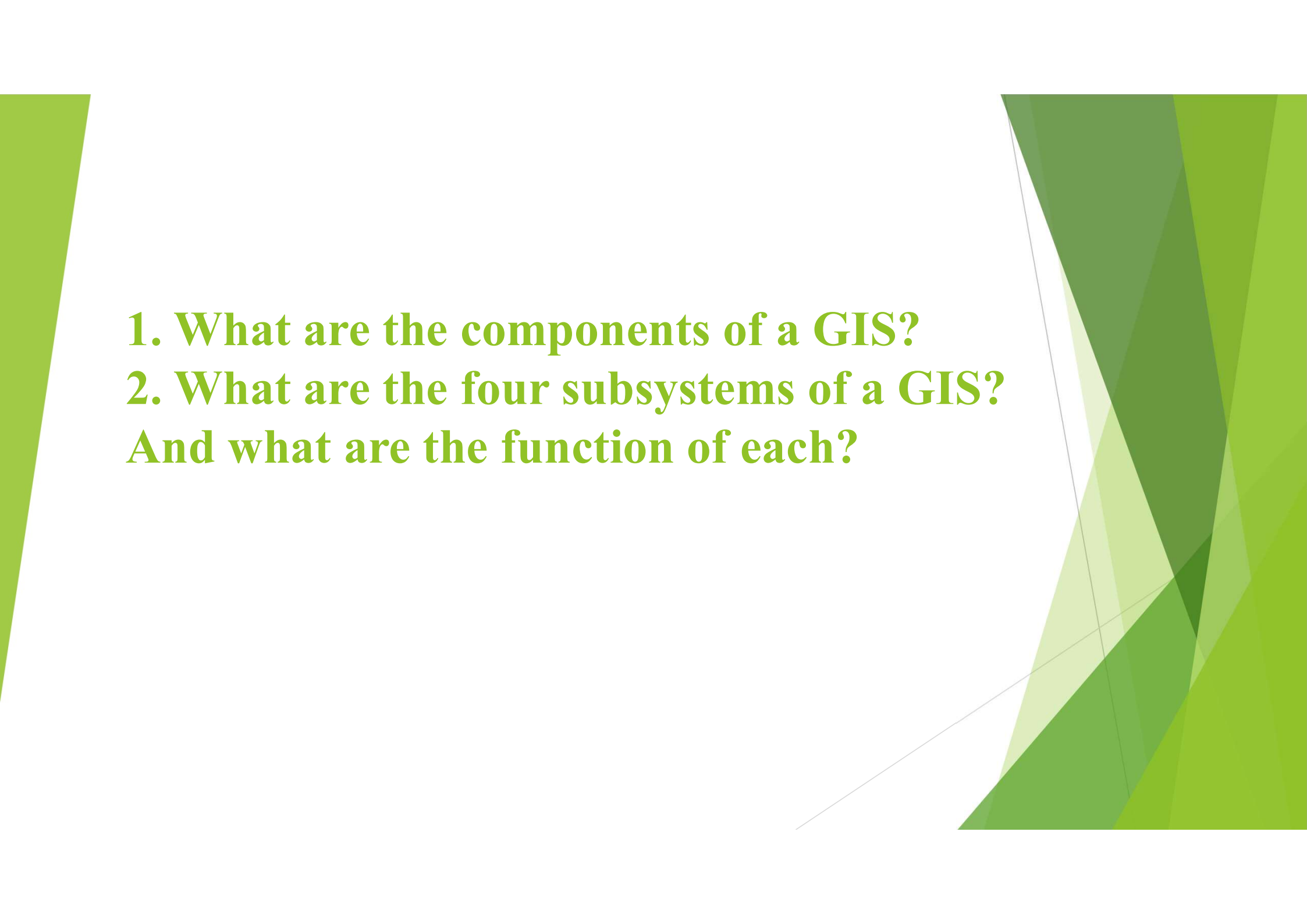


The background features abstract green geometric shapes. On the left, a solid green trapezoid points towards the center. On the right, a complex arrangement of overlapping, semi-transparent green triangles and polygons creates a layered, crystalline effect. A thin, light gray line extends from the bottom left towards the right side of the composition.

Lecture 3: GIS DATA Sources

- 
- The background of the slide features abstract, overlapping green geometric shapes, primarily triangles and polygons, in various shades of green, creating a modern and dynamic visual effect.
- 1. What are the components of a GIS?**
 - 2. What are the four subsystems of a GIS?**
- And what are the function of each?**

GIS DATA Sources

- This lecture reviews different sources, formats, and input techniques for GIS data.
- The focus is on reviewing different data input techniques for spatial data.
- This lecture also describes data input errors, spatial and attribute, and reviews typical procedures to correct input errors.
- This lecture outlines are:
 - Sources of data
 - Data input techniques
 - Data editing and quality assurance

GIS Data Sources

- As previously identified, two types of data are input into a GIS, spatial and attribute.
- The data input process is the operation of encoding both types of data into the GIS database formats.
- The creation of a clean digital database is the most important and time-consuming task upon which the usefulness of the GIS depends.
- The establishment and maintenance of a robust spatial database is the cornerstone of a successful GIS implementation.
- As well, the digital data is the most expensive part of the GIS. Yet often, not enough attention is given to the quality of the data or the processes by which they are prepared for automation.

Spatial Data Sources

- The general consensus among the GIS community is that 60 to 80 % of the cost incurred during implementation of GIS technology lies in data acquisition, data compilation and database development.
- A wide variety of data sources exist for both spatial and attribute data.
- The most common general sources for spatial data are:
 - Hard copy maps
 - Aerial photography
 - Remotely sensed images
 - Point data samples from a survey
 - Existing digital data files

Spatial Data Sources

- ▶ Existing hard copy maps, e.g. sometimes referred to as *analogue maps*, provide the most popular source for any GIS project.
- ▶ Potential users should be aware that while there are many private sector firms specializing in providing digital data, federal, provincial and state government agencies are an excellent source of data.
- ▶ Because of the large costs associated with data capture and input, government departments are often the only agencies with financial resources and manpower funding to invest in data compilation.
- ▶ Federal agencies are also often a good source for base map information. An inherent advantage of digital data from government agencies is its cost.
- ▶ It is typically inexpensive. However, this is often offset by the data's accuracy and quality.
- ▶ Thematic coverages are often not up to date.

Attribute Data Sources

- ▶ Attribute data has an even wider variety of data sources.
- ▶ Any textual or tabular data that can be referenced to a geographic feature, e.g. a point, line, or area, can be input into a GIS.
- ▶ Attribute data is usually input by manual keying or via a bulk loading utility of the DBMS software.
- ▶ ASCII format is a de facto standard for the transfer and conversion of attribute information.

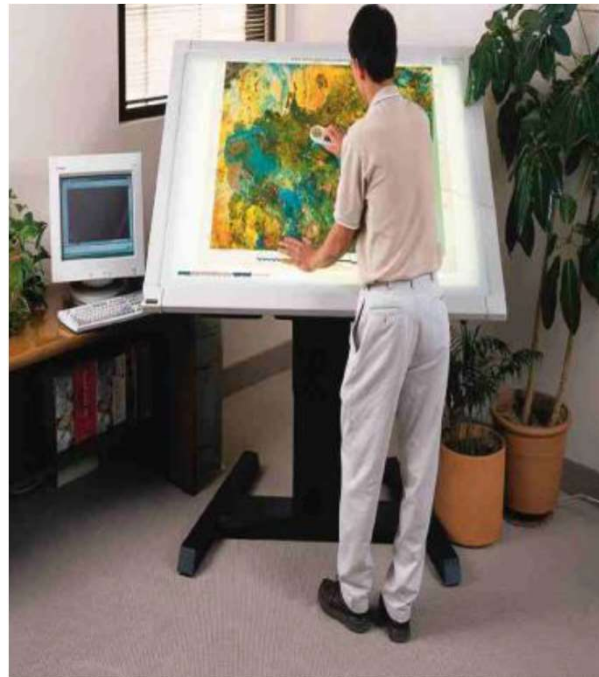
Data Input Techniques

- ▶ Since the input of attribute data is usually quite simple, the discussion of data input techniques will be limited to spatial data only.
- ▶ There is no single method of entering the spatial data into a GIS. Rather, there are several, mutually compatible methods that can be used singly or in combination.
- ▶ The choice of data input method is governed largely by the application, the available budget, and the type and the complexity of data being input.
- ▶ There are at least four basic procedures for inputting spatial data into a GIS. These are:
 - ▶ Manual digitizing
 - ▶ Automatic scanning
 - ▶ Entry of coordinates using coordinate geometry
 - ▶ Conversion of existing digital data

Digitizing

- ▶ While considerable work has been done with newer technologies, the overwhelming majority of GIS spatial data entry is done by manual digitizing.
- ▶ A digitizer is an electronic device consisting of a table upon which the map or drawing is placed.
- ▶ The user traces the spatial features with a hand-held magnetic pen, often called a *mouse* or cursor.
- ▶ While tracing the features the coordinates of selected points, e.g. vertices, are sent to the computer and stored.
- ▶ All points that are recorded are registered against positional control points, usually the map corners, that are keyed in by the user at the beginning of the digitizing session.
- ▶ The coordinates are recorded in a user defined coordinate system or map projection.
- ▶ Latitude and longitude is most often used.

Digitizing



Advantages of Manual Digitizing

Manual digitizing has many advantages. These include:

- Low capital cost, e.g. digitizing tables are cheap;
- Low cost of labour;
- Flexibility and adaptability to different data types and sources;
- Easily taught in a short amount of time - an easily mastered skill
- Generally the quality of data is high;
- Digitizing devices are very reliable and most often offer a greater precision than the data warrants; and
- Ability to easily register and update existing data.

Digitizing Raster Data

- ▶ For raster based GIS software data is still commonly digitized in a vector format and converted to a raster structure.
- ▶ The procedure usually differs minimally from vector based software digitizing, other than some raster systems allow the user to define the resolution size of the grid-cell. Conversion to the raster structure may occur *on-the-fly* or afterwards as a separate conversion process.





Automatic Scanning

- ▶ A variety of scanning devices exist for the automatic capture of spatial data.
- ▶ While several different technical approaches exist in scanning technology, all have the advantage of being able to capture spatial features from a map at a rapid rate of speed.
- ▶ However, as of yet, scanning has not proven to be a viable alternative for most GIS implementation.
- ▶ Scanners are generally expensive to acquire and operate. As well, most scanning devices have limitations with respect to the capture of selected features, e.g. text and symbol recognition.
- ▶ Experience has shown that most scanned data requires a substantial amount of manual editing to create a clean data layer.

Automatic Scanning



Automatic Scanning

- ▶ Given these basic constraints some other practical limitations of scanners should be identified. These include :
 - ▶ hard copy maps are often unable to be removed to where a scanning device is available, e.g. most companies or agencies cannot afford their own scanning device and therefore must send their maps to a private firm for scanning;
 - ▶ hard copy data may not be in a form that is viable for effective scanning, e.g. maps are of poor quality, or are in poor condition;
 - ▶ geographic features may be too few on a single map to make it practical, cost-justifiable, to scan;
 - ▶ often on *busy* maps a scanner may be unable to distinguish the features to be captured from the surrounding graphic information, e.g. dense contours with labels;
 - ▶ scanning is much more expensive than manual digitizing, considering all the cost/performance issues.

Special Data Models

- ▶ Consensus within the GIS community indicates that scanners work best when the information on a map is kept very clean, very simple, and uncluttered with graphic symbology.
- ▶ The sheer cost of scanning usually eliminates the possibility of using scanning methods for data capture in most GIS implementations. Large data capture shops and government agencies are those most likely to be using scanning technology.
- ▶ Currently, general consensus is that the quality of data captured from scanning devices is not substantial enough to justify the cost of using scanning technology. However, major breakthroughs are being made in the field, with scanning techniques and with capabilities to automatically clean and prepare scanned data for topological encoding. These include a variety of *line following* and *text recognition* techniques. Users should be aware that this technology has great potential in the years to come, particularly for larger GIS installations.

Coordinate Geometry

- A third technique for the input of spatial data involves the calculation and entry of coordinates using coordinate geometry (COGO) procedures.
- This involves entering, from survey data, the explicit measurement of features from some known monument.
- This input technique is obviously very costly and labour intensive.
- In fact, it is rarely used for natural resource applications in GIS.
- This method is useful for creating very precise cartographic definitions of property, and accordingly is more appropriate for land records management.

Conversion of Existing Digital Data

- ▶ A fourth technique that is becoming increasingly popular for data input is the conversion of existing digital data.
- ▶ A variety of spatial data, including digital maps, are openly available from a wide range of government and private sources.
- ▶ The most common digital data to be used in a GIS is data from CAD systems.
- ▶ A number of data conversion programs exist, mostly from GIS software vendors, to transform data from CAD formats to a raster or topological GIS data format.
- ▶ Most GIS software vendors also provide an ASCII data exchange format specific to their product, and a programming subroutine library that will allow users to write their own data conversion routines to fulfil their own specific needs.
- ▶ As digital data becomes more readily available this capability becomes a necessity for any GIS. Data conversion from existing digital data is not a problem for most technical persons in the GIS field.
- ▶ However, for smaller GIS installations who have limited access to a *GIS analyst* this can be a major stumbling block in getting a GIS operational.

Data Editing and Quality Assurance

- ▶ Data editing and verification is in response to the errors that arise during the encoding of spatial and non-spatial data.
- ▶ The editing of spatial data is a time consuming, interactive process that can take as long, if not longer, than the data input process itself.
- ▶ Several kinds of errors can occur during data input. They can be classified as:
 - ▶ **Incompleteness of the spatial data.** This includes missing points, line segments, and/or polygons.
 - ▶ **Locational placement errors of spatial data.** These types of errors usually are the result of careless digitizing or poor quality of the original data source.
 - ▶ **Distortion of the spatial data.** This kind of error is usually caused by base maps that are not scale-correct over the whole image, e.g. aerial photographs, or from material stretch, e.g. paper documents.

Data Editing and Quality Assurance

- ▶ **Incorrect linkages between spatial and attribute data.** This type of error is commonly the result of incorrect unique identifiers (labels) being assigned during manual key in or digitizing. This may involve the assigning of an entirely wrong label to a feature, or more than one label being assigned to a feature.
- ▶ **Attribute data is wrong or incomplete.** Often the attribute data does not match exactly with the spatial data. This is because they are frequently from independent sources and often different time periods. Missing data records or too many data records are the most common problems.

Data Editing and Quality Assurance

- ▶ The identification of errors in spatial and attribute data is often difficult.
- ▶ Most spatial errors become evident during the topological building process. The use of *check plots* to clearly determine where spatial errors exist is a common practice. Most topological building functions in GIS software clearly identify the geographic location of the error and indicate the nature of the problem. Comprehensive GIS software allows users to graphically walk through and edit the spatial errors. Others merely identify the type and coordinates of the error. Since this is often a labour intensive and time consuming process, users should consider the error correction capabilities very important during the evaluation of GIS software offerings.

Spatial Data Errors

- ▶ A variety of common data problems occur in converting data into a topological structure.
- ▶ Usually data is input by digitizing. Digitizing allows a user to trace spatial data from a hard copy product, e.g. a map, and have it recorded by the computer software.
- ▶ Most GIS software has utilities to *clean* the data and build a topologic structure.
- ▶ If the data is unclean to start with, for whatever reason, the cleaning process can be very lengthy. Interactive editing of data is a distinct reality in the data input process.
- ▶ Experience indicates that in the course of any GIS project 60 to 80 % of the time required to complete the project is involved in the input, cleaning, linking, and verification of the data.
- ▶ The most common problems that occur in converting data into a topological structure include:
 - ▶ gaps in the line work;
 - ▶ dead ends, e.g. also called dangling arcs, resulting from overshoots and undershoots in the line work; and
 - ▶ bow ties or weird polygons from inappropriate closing of connecting features.

Image Data Type

- ▶ Of course, topological errors only exist with linear and areal features. They become most evident with polygonal features. *Slivers* are the most common problem when cleaning data. Slivers frequently occur when coincident boundaries are digitized separately, e.g. once each for adjacent forest stands, once for a lake and once for the stand boundary, or after polygon overlay. Slivers often appear when combining data from different sources, e.g. forest inventory, soils, and hydrography. It is advisable to digitize data layers with respect to an existing data layer, e.g. hydrography, rather than attempting to match data layers later. A proper plan and definition of priorities for inputting data layers will save many hours of interactive editing and cleaning.
- ▶ *Dead ends* usually occur when data has been digitized in a *spaghetti* mode, or without snapping to existing nodes. Most GIS software will clean up undershoots and overshoots based on a user defined tolerance, e.g. distance. The definition of an inappropriate distance often leads to the formation of *bow ties* or *weird polygons* during topological building. Tolerances that are too large will force arcs to snap one another that should not be connected. The result is small polygons called *bow ties*. The definition of a proper tolerance for cleaning requires an understanding of the scale and accuracy of the data set.
- ▶ The other problem that commonly occurs when building a topologic data structure is *duplicate lines*. These usually occur when data has been digitized or converted from a CAD system. The lack of topology in these type of drafting systems permits the inadvertent creation of elements that are exactly duplicate. However, most GIS packages afford automatic elimination of duplicate elements during the topological building process. Accordingly, it may not be a concern with vector based GIS software. Users should be aware of the duplicate element that retraces itself, e.g. a three vertice line where the first point is also the last point. Some GIS packages do not identify these feature inconsistencies and will build such a feature as a valid polygon. This is because the topological definition is mathematically correct, however it is not geographically correct. Most GIS software will provide the capability to eliminate bow ties and slivers by means of a feature elimination command based on area, e.g. polygons less than 100 square metres. The ability to define custom topological error scenarios and provide for semi-automated correction is a desirable capability for GIS software.
- ▶ The adjoining figure illustrates some typical errors described above. Can you spot them? They include undershoots, overshoots, bow ties, and slivers. Most bow ties occur when inappropriate tolerances are used during the automated cleaning of data that contains many overshoots. This particular set of spatial data is a prime candidate for numerous bow tie polygons.

Attribute data Errors

- ▶ The identification of attribute data errors is usually not as simple as spatial errors.
- ▶ This is especially true if these errors are attributed to the quality or reliability of the data.
- ▶ Errors as such usually do not surface until later on in the GIS processing.
- ▶ Solutions to these type of problems are much more complex and often do not exist entirely.
- ▶ It is much more difficult to spot errors in attribute data when the values are syntactically good, but incorrect.
- ▶ Simple errors of linkage, e.g. missing or duplicate records, become evident during the linking operation between spatial and attribute data.
- ▶ Again, most GIS software contains functions that check for and clearly identify problems of linkage during attempted operations.
- ▶ This is also an area of consideration when evaluating GIS software.

Data Verification

- ▶ **Six clear steps stand out in the data editing and verification process for spatial data. These are:**
 - ▶ **Visual review.** This is usually by check plotting.
 - ▶ **Cleanup of lines and junctions.** This process is usually done by software first and interactive editing second.
 - ▶ **Weeding of excess coordinates.** This process involves the removal of redundant vertices by the software for linear and/or polygonal features.
 - ▶ **Correction for distortion and warping.** Most GIS software has functions for scale correction and *rubber sheeting*. However, the distinct rubber sheet algorithm used will vary depending on the spatial data model, vector or raster, employed by the GIS. Some raster techniques may be more intensive than vector based algorithms.
 - ▶ **Construction of polygons.** Since the majority of data used in GIS is polygonal, the construction of polygon features from lines/arcs is necessary. Usually this is done in conjunction with the topological building process.
 - ▶ **The addition of unique identifiers or labels.** Often this process is manual. However, some systems do provide the capability to automatically build labels for a data layer.

Data Verification

- ▶ These data verification steps occur after the data input stage and prior to or during the linkage of the spatial data to the attributes.
- ▶ Data verification ensures the integrity between the spatial and attribute data.
- ▶ Verification should include some brief querying of attributes and cross checking against known values.

What are the four basic procedures for inputting spatial data ?

