Constructing an EA-level Database for the Census

Amor Laaribi
UN-GGIM Secretariat
UN Statistics Division
New York

Positioning geospatial information to address global challenges

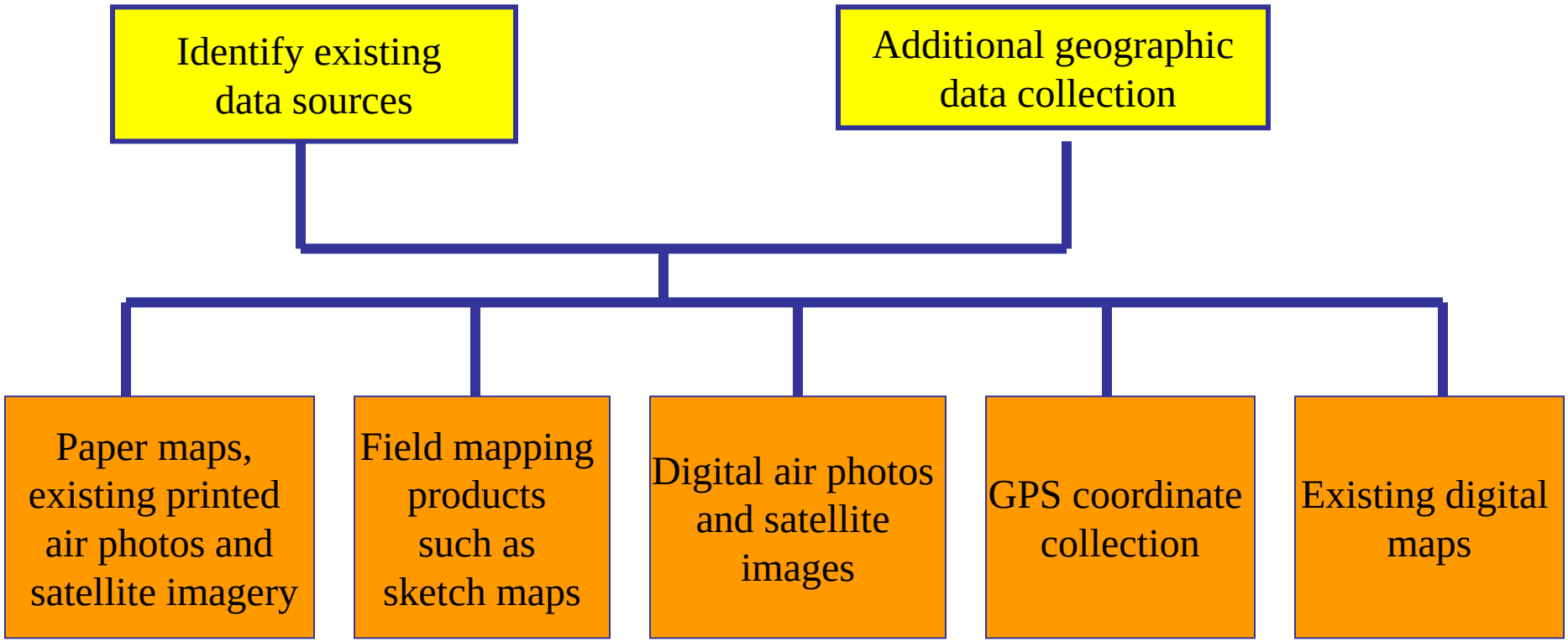
Overview

- **Stages in the Geographic Database Development**
  - Sources of geographic information
  - Data conversion
  - Data integration
- **Implementation of the Database**
  - Data Modelling
  - Relational Data Model
  - Example
- **Conclusion**

Stages in the geographic database development

- **Geographic data sources for EA delineation**
  - Inventory of existing data sources
  - Additional geographic data collection
- **Geographic data conversion**
  - Digitizing/Scanning + raster-to-vector conversion
  - Editing Geographic features
  - Constructing and maintaining topology for geographic features
- **Data integration**
  - Geo-referencing/Coding
  - Combining and integrating/Additional delineation of EA boundaries
- **Parallel activity**
  - Develop geographic attribute database
  - Metadata development

Sources of geographic information



Inventory of existing sources

- National mapping agency (often the lead agency in the country);
- Military mapping services;
- Province, district and municipal governments. (transportation, social services, utility services and planning relevant information);
- Various government/private organizations dealing with spatial data;
  - Geological or hydrological survey, Environmental protection authority, Utility and communication sector companies;
- Donor activities

Why Data Inventory?

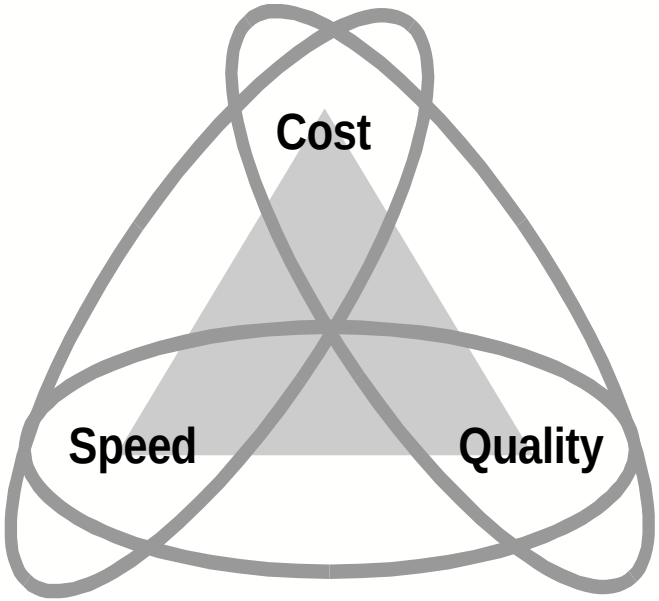
- Geographic data: Labor intensive, tedious and error-prone
- Up to 70% of GIS projects
- Identify existing data sources

Geographic data conversion

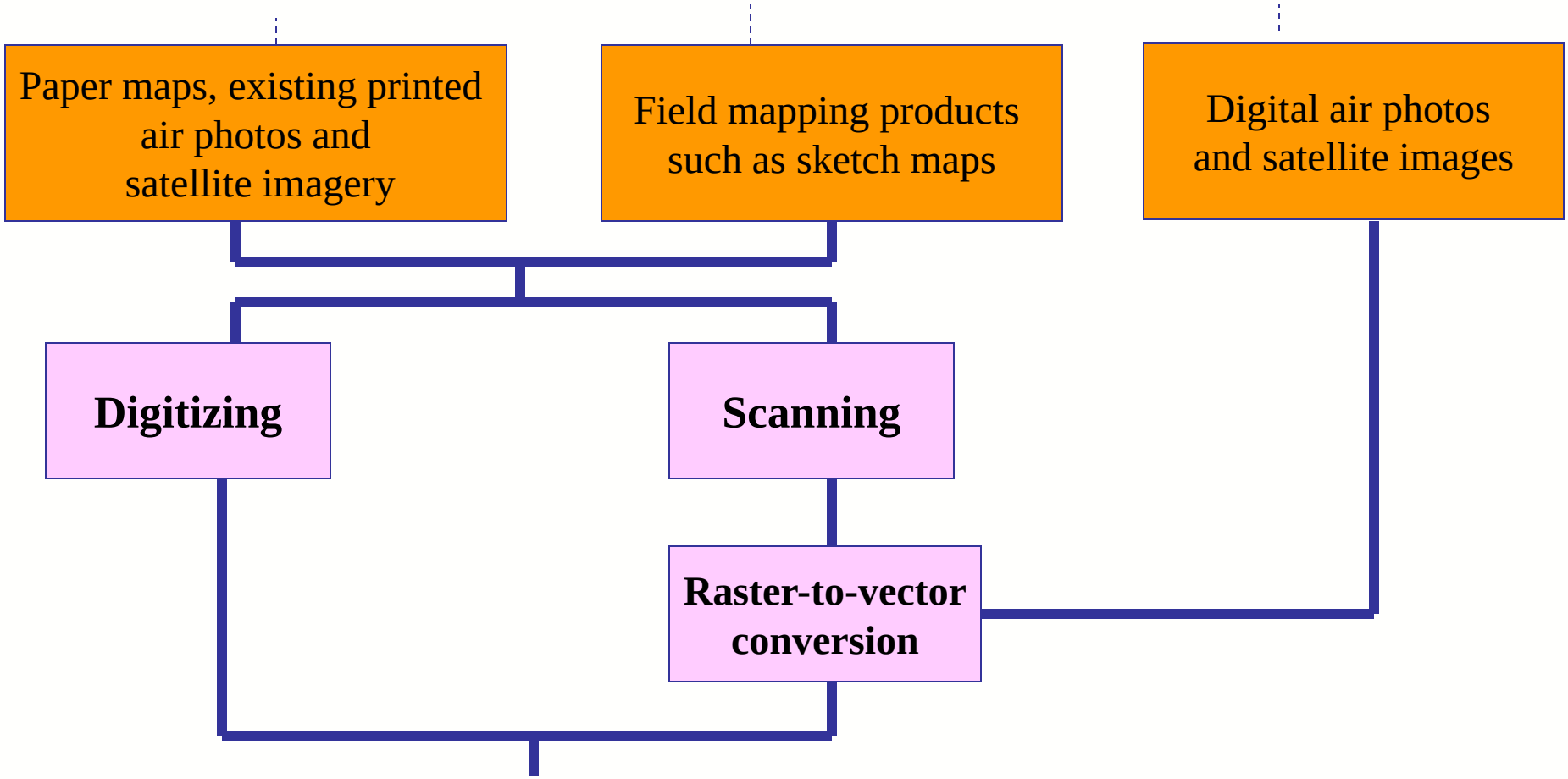
- **Data conversion:**
  - The process of converting features that are visible on a hardcopy map into digital point, line, polygon and attribute information is called data automation or **data conversion**.

- The best strategy for data conversion depends on many factors including data availability, time and resource constraints

- Trade-off between the cost of a project, the amount of time required to complete data conversion, and the quality of the final product.



Data Conversion



Geographic data conversion

- **2 main approaches for converting information on hardcopy maps to digital data:**
  - Scanning
  - Digitizing

Scanning

- Scanning has arguably bypassed digitizing as the main method of spatial data input, mainly because of the potential to automate some tedious data-input steps using large-format feed scanners and interactive “vectorization” software.

- The result of the scanning process is a **raster image** of the original map which can be stored in a standard image format such as GIF or TIFF.

- After geo-referencing it can be displayed in GIS packages as a backdrop to existing vector data.

Advantages and Disadvantages of Scanning

Advantages

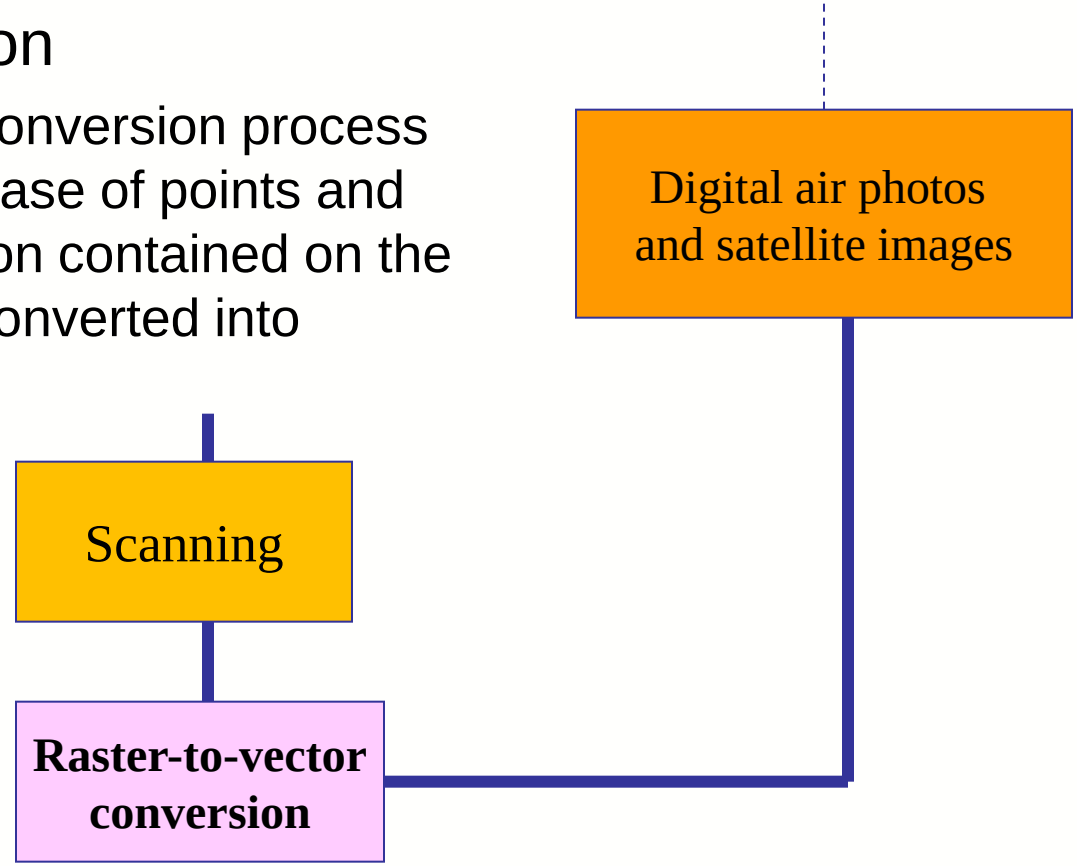
- Scanned maps can be used as image backdrops for vector information;
- Clear base maps or original color separations can be "vectorized" relatively easily using raster-to-vector conversion software; and
- Small-format scanners are relatively inexpensive and provide quick data capture.

Disadvantages

- Converting large maps with small format scanners requires tedious re-assembly of the individual parts;
- Scanning large volumes of hard-copy maps will present challenges for file storage on many desktop computer systems;
- Despite recent advances in "vectorization" software, considerable manual editing and attribute labeling may still be required.

Raster to Vector Conversion

- Raster to Vector Conversion
  - Since the end result of the conversion process is a digital geographic database of points and lines, the scanned information contained on the raster images needs to be converted into coordinate information.

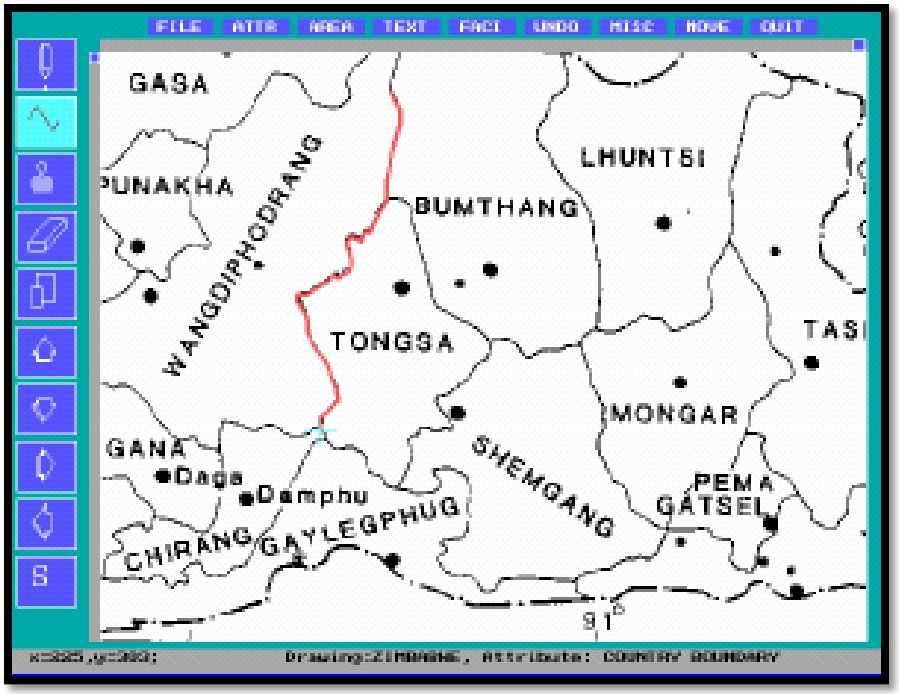


Digitizing

- Manual Digitizing
  - Digitizing is often tedious and tiring to the operators
- Heads up Digitizing
  - In the heads-up digitizing, a scanned map image is used digitally to trace the outlines into a GIS layer

Heads-Up Digitizing II

- Operator uses a Raster-scanned image on the computer screen (a scanned map, air photo or satellite image) as a backdrop.
- Operator follows lines on-screen in vector mode



Advantages and Disadvantages of Digitizing

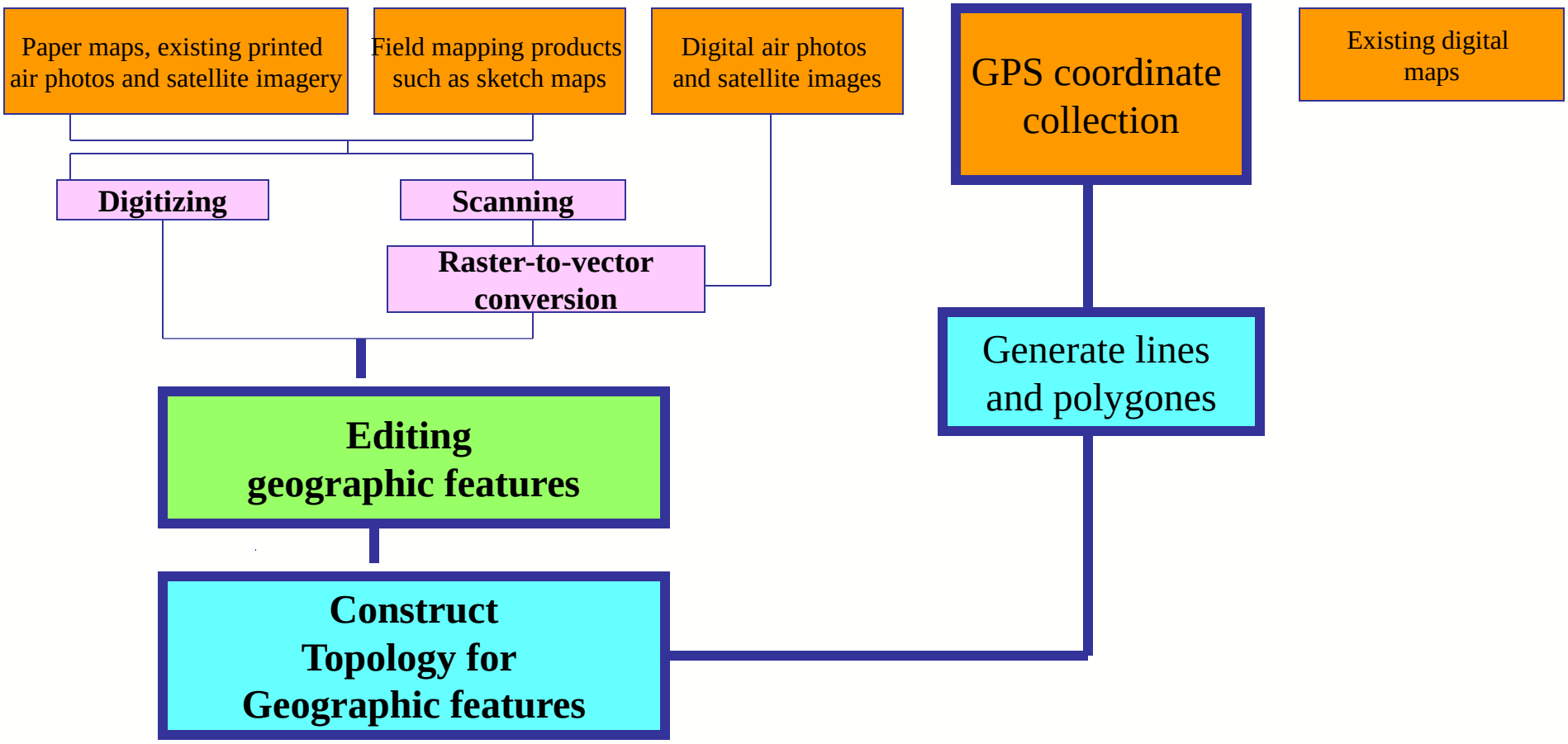
Advantages

- Digitizing is easy to learn and thus does not require expensive skilled labor;
- Attribute information can be added during the digitizing process;
- High accuracy can be achieved through manual digitizing; i.e., there is usually no loss of accuracy compared to the source map.

Disadvantages

- Digitizing is tedious possibly leading to operator fatigue and resulting quality problems which may require considerable post-processing;
- Manual digitizing is quite slow;
- In contrast to primary data collection using GPS or aerial photography, the accuracy of digitized maps is limited by the quality of the source material.

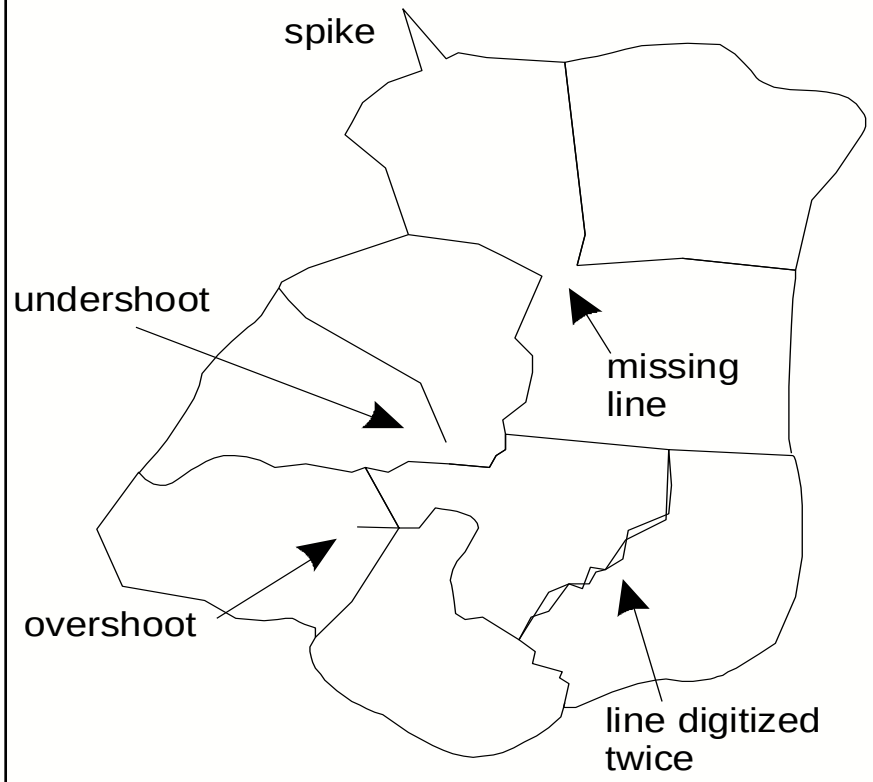
Editing and Building topology



Editing

- Manual digitizing is error prone
- Objective is to produce an accurate representation of the original map data
- This means that all lines that connect on the map must also connect in the digital database
- There should be no missing features and no duplicate lines
- The most common types of **errors**
  - Reconnect disconnected line segments, etc.

Some common digitizing errors



Fixing Errors

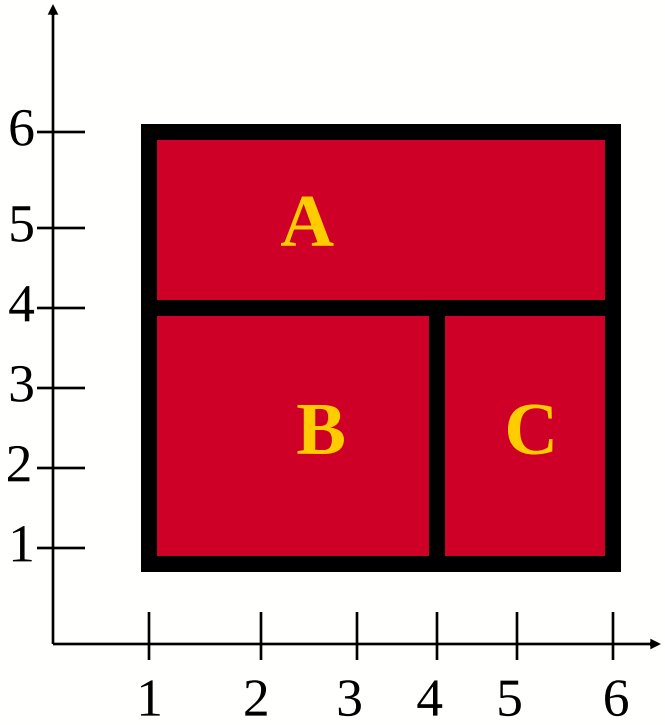
- Some of the common digitizing errors shown in the figure can be avoided by using the digitizing software's snap tolerances that are defined by the user.
- For example, the user might specify that all endpoints of a line that are closer than 1 mm from another line will automatically be connected (snapped) to that line.
- Small sliver polygons that are created when a line is digitized twice can also be automatically removed.

Topology

- Data structure in which each point, line and polygon :
  - “knows” where it is
  - “knows” what is around it
  - “understands” its environment
  - “knows” how to get around

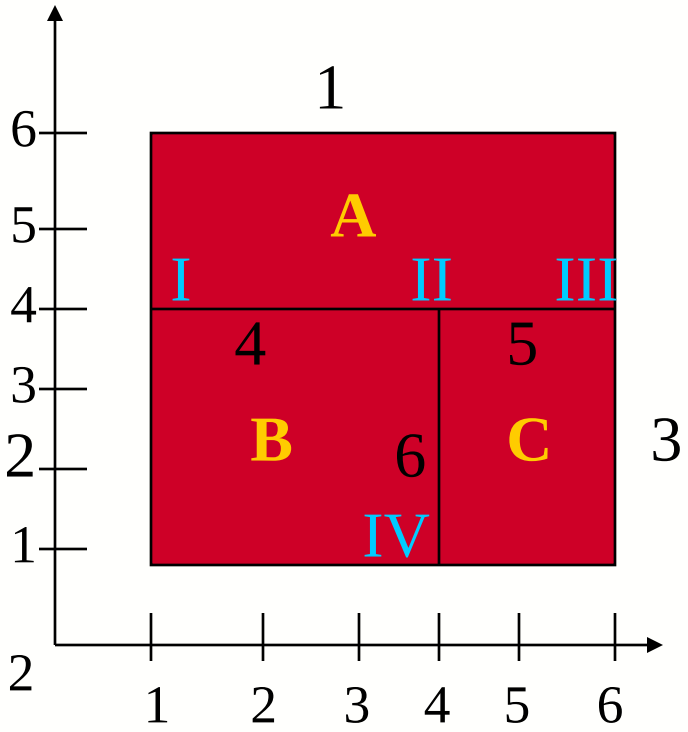
Helps answer the question what is where?

Example of “Spaghetti” data structure



Poly	coordinates
A	(1,4), (1,6), (6,6), (6,4), (4,4), (1,4)
B	(1,4), (4,4), (4,1), (1,1), (1,4)
C	(4,4), (6,4), (6,1), (4,1), (4,4)

Example of Topological data structure



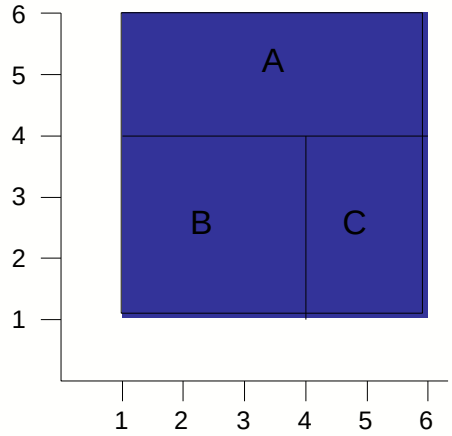
O = "outside" polygon

Node	X	Y	Lines
I	1	4	1,2,4
II	4	4	4,5,6
III	6	4	1,3,5
IV	4	1	2,3,6

Poly	Lines
A	1,4,5
B	2,4,6
C	3,5,6

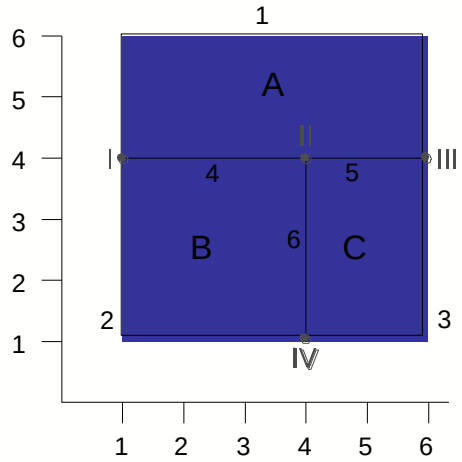
Line	From Node	To Node	Left Poly	Right Poly
1	I	III	O	A
2	I	IV	B	O
3	III	IV	O	C
4	I	II	A	B
5	II	III	A	C
6	II	IV	C	B

“Spaghetti” data structure



Poly	Coordinates
A	(1,4), (1,6), (6,6), (6,4), (4,4), (1,4)
B	(1,4), (4,4), (4,1), (1,1), (1,4)
C	(4,4), (6,4), (6,1), (4,1), (4,4)

Topological data structure



O = “outside” polygon

Node	X	Y	Lines
I	1	4	1,2,4
II	4	4	4,5,6
III	6	4	1,3,5
IV	4	1	2,3,6

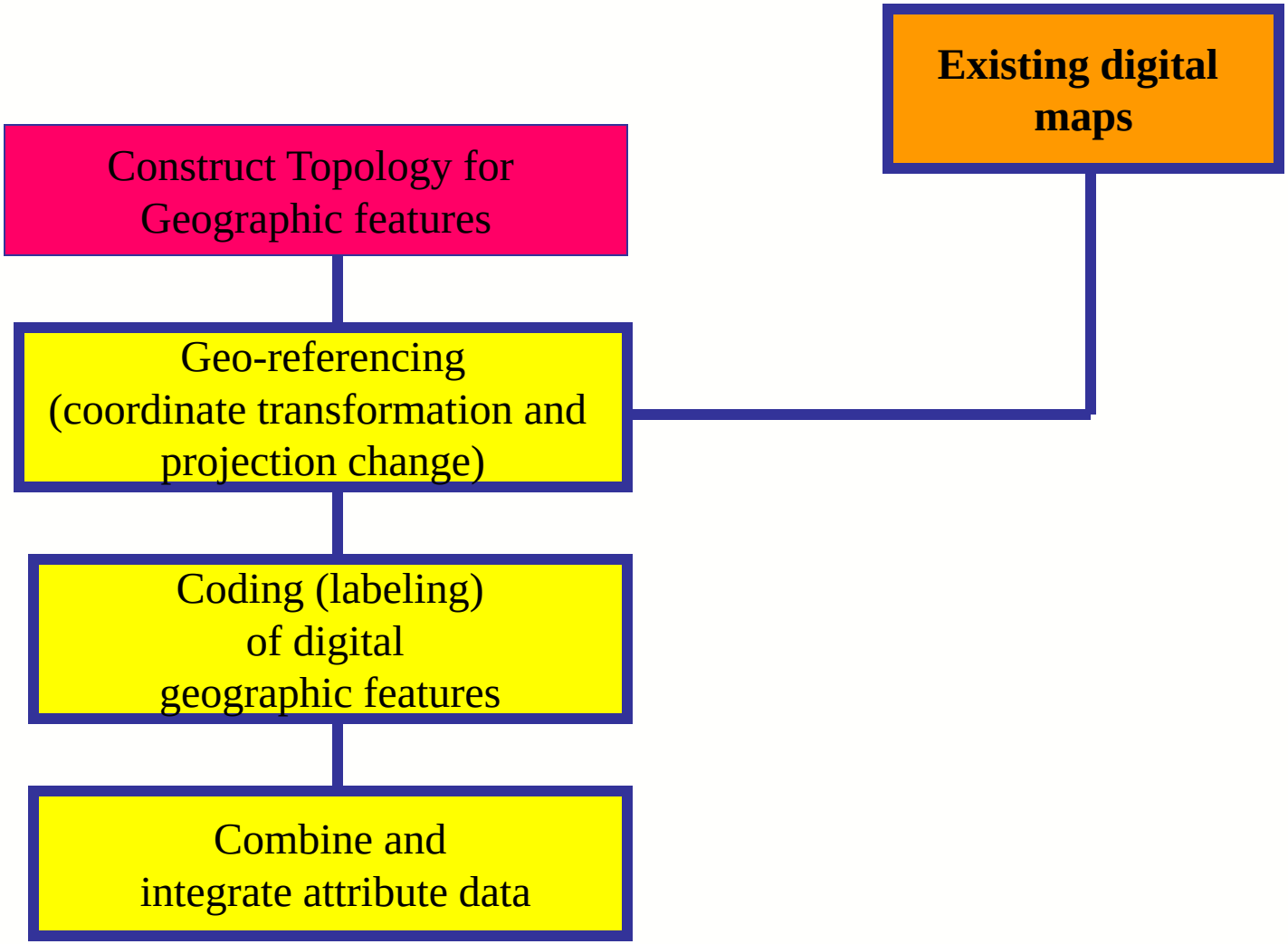
Poly	Lines
A	1,4,5
B	2,4,6
C	3,5,6

Line	From Node	To Node	Left Poly	Right Poly
1	I	III	O	A
2	I	IV	B	O
3	III	IV	O	C
4	I	II	A	B
5	II	III	A	C
6	II	IV	C	B

Constructing and maintaining topology (cont.)

- Storing the topological information facilitates analysis, since many GIS operations do not actually require coordinate information, but are based only on topology
- The user typically does not have to worry about how the GIS stores topological information. How this is actually done is software-specific.
- Building topology thus also acts as a test of database integrity

Digital data integration



Integrating data

- Geo-referencing
  - Converting map coordinates to the **real world coordinates** corresponding to the source map's cartographic projection (or at digitizing stage).
  - Attaching codes to the digitized features
- Integrating attribute data
  - Spreadsheets
  - links to external database

Integrating attribute data

- After the completed digital database has been verified to be error-free, the final step is to add additional attributes
- These can be linked to the database permanently, or the additional information about each database feature can be stored in separate files which are linked to the geographic database as needed

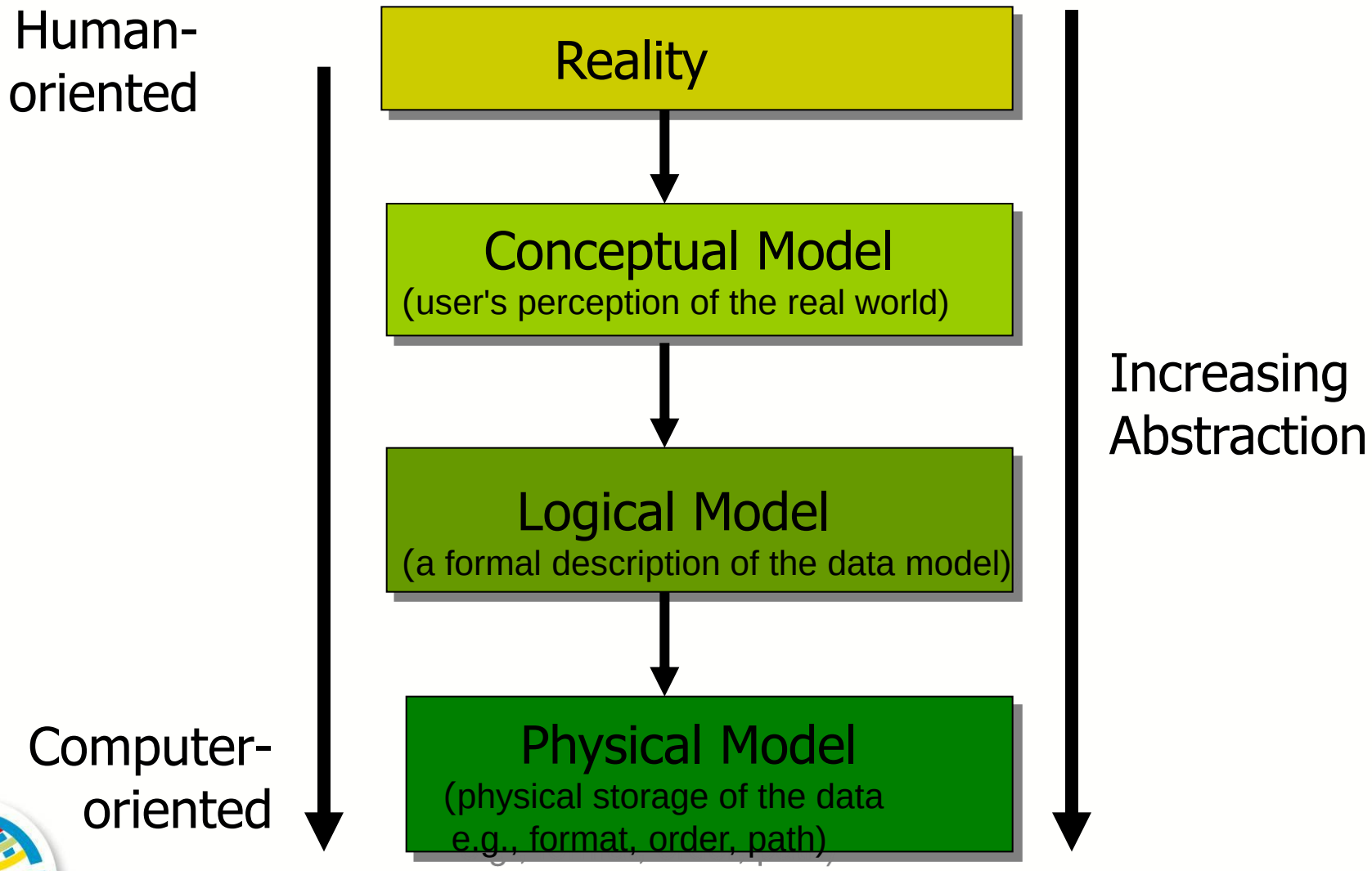
Implementation of an EA database

- Geographic databases (hereafter referred to as geodatabases) are more than spreadsheets
- Entity types can be defined as having specific properties that govern behavior in the real world.
- The EA as a geographic unit is a kind of object whose function is to delineate territory for the census canvassing operation.
- Morphologically, the EA is contiguous, it nests within administrative units, and it is composed of population-based units.

Implementation of an EA database (cont.)

- All large operational GISs are built on geodatabases;
- Arguably the most important part of the GIS
- Geodatabases form the basis for all queries, analysis, and decision-making.
- A DBMS, or database management system, is where databases are stored.

Levels of Data Abstraction

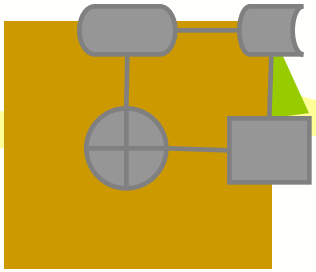


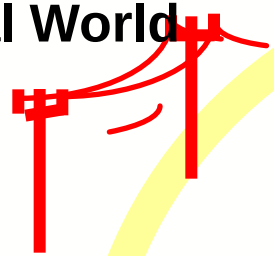
Levels of Data Abstraction

Conceptual Model

- **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, polygons, raster cells
- requires decisions about how real-world dimensionality and relationships will be represented:
  - based on the processing that will be done on these objects
  - e.g. should a building be represented as an area or a point?

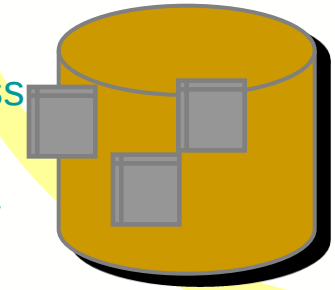
Real World

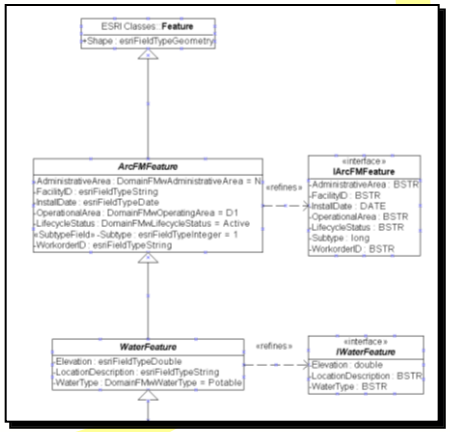




Physical Model

- both hardware and software specific
- how files will be structured for access from the disk - Database Schema



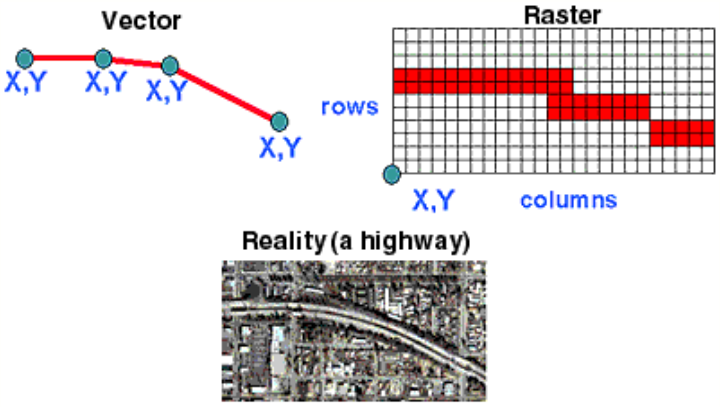


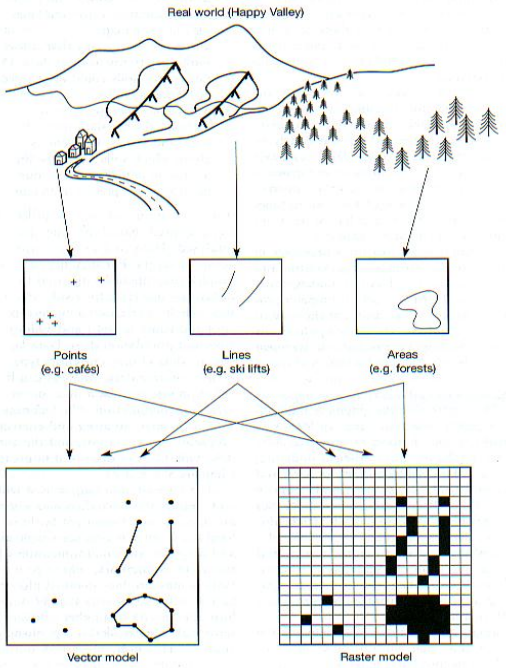
Logical Model

- software specific but hardware independent
- sets out the logical structure of the database elements, determined by the database management system used by the software

GIS Data Models

- Vector data model
- Raster data model





courtesy: Mary Ruvane, http://ils.unc.edu/

Several types of data organization

- Database management systems (DBMSs) can be divided into various types, including:
  - Relational
  - Object
  - Object-relational
- Relational (RDBMS)
  - RDMS is the most popular type of DBMS
    - Over 95% of data in DBMS is in RDBMS
    - DB2; SQL Server, Access; Oracle; Informix

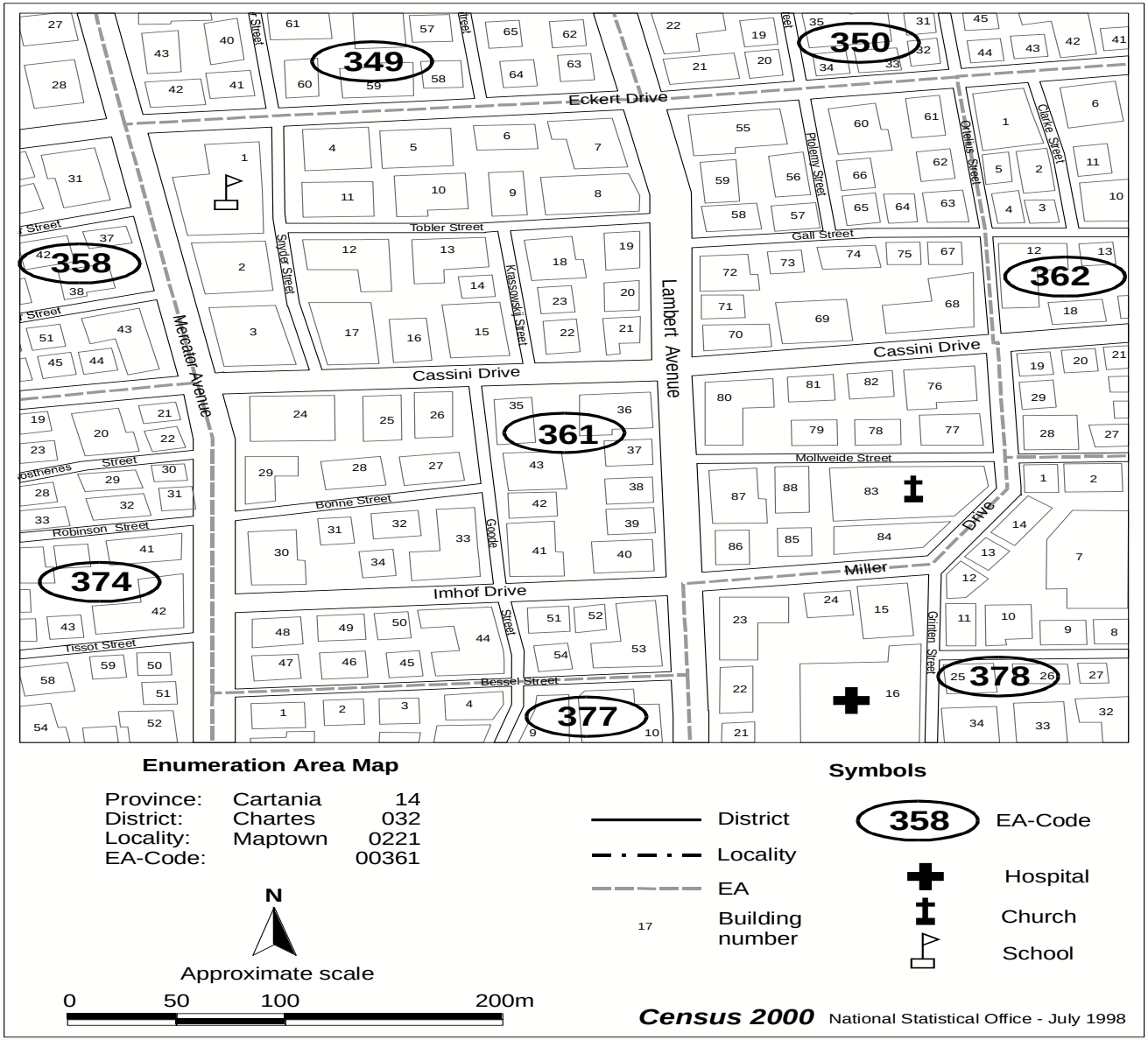
Example: the Relational Database Model

- The relational database model is used to store, retrieve and manipulate tables of data that refer to the geographic features in the coordinate database.
- It is based on the **entity-relationship** model
- In a geographic context, an *entity* can be administrative or census units, or any other spatial feature for which characteristics will be compiled.

Example of an urban EA map

Main components are:

Street network,
Buildings
EA boundaries layer
Annotation,
Symbols,
Labels
Building numbers
Neatlines
Legend



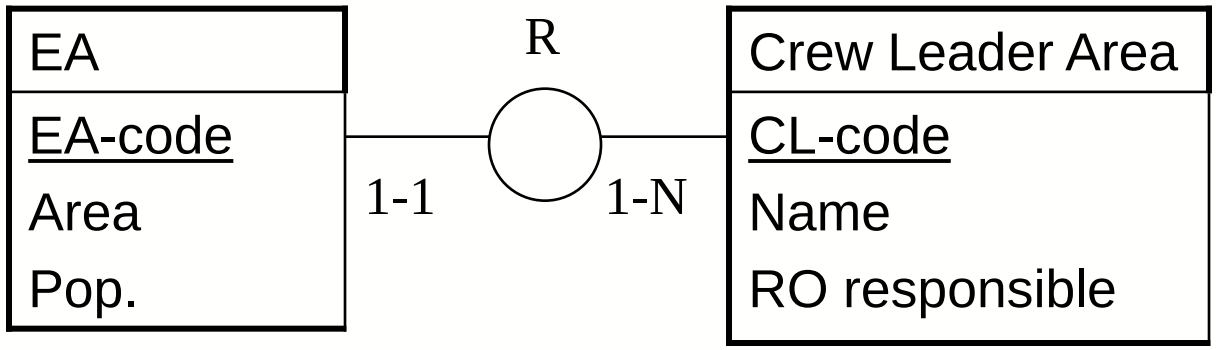
Definition of EA database content

- A **spatial data model:**
  - description of geographic features, such as administrative units, streets, houses, buildings, hydrological features/rivers, etc., and the relationships between those entities.
  - A data model is independent from any specific technology/software package

Entity-Relationship Example:

EA entity can be linked to the entity **Crew Leader Area**.

The table for this entity could have attributes such as the **name of the crew leader**, the **regional office responsible**, **contact information**, and **the crew leader code** (CL code) as **primary code**, which is also present in the EA entity.



Implementation of an EA database

- Example of an entity table – enumeration area

Entity: Enumeration areas

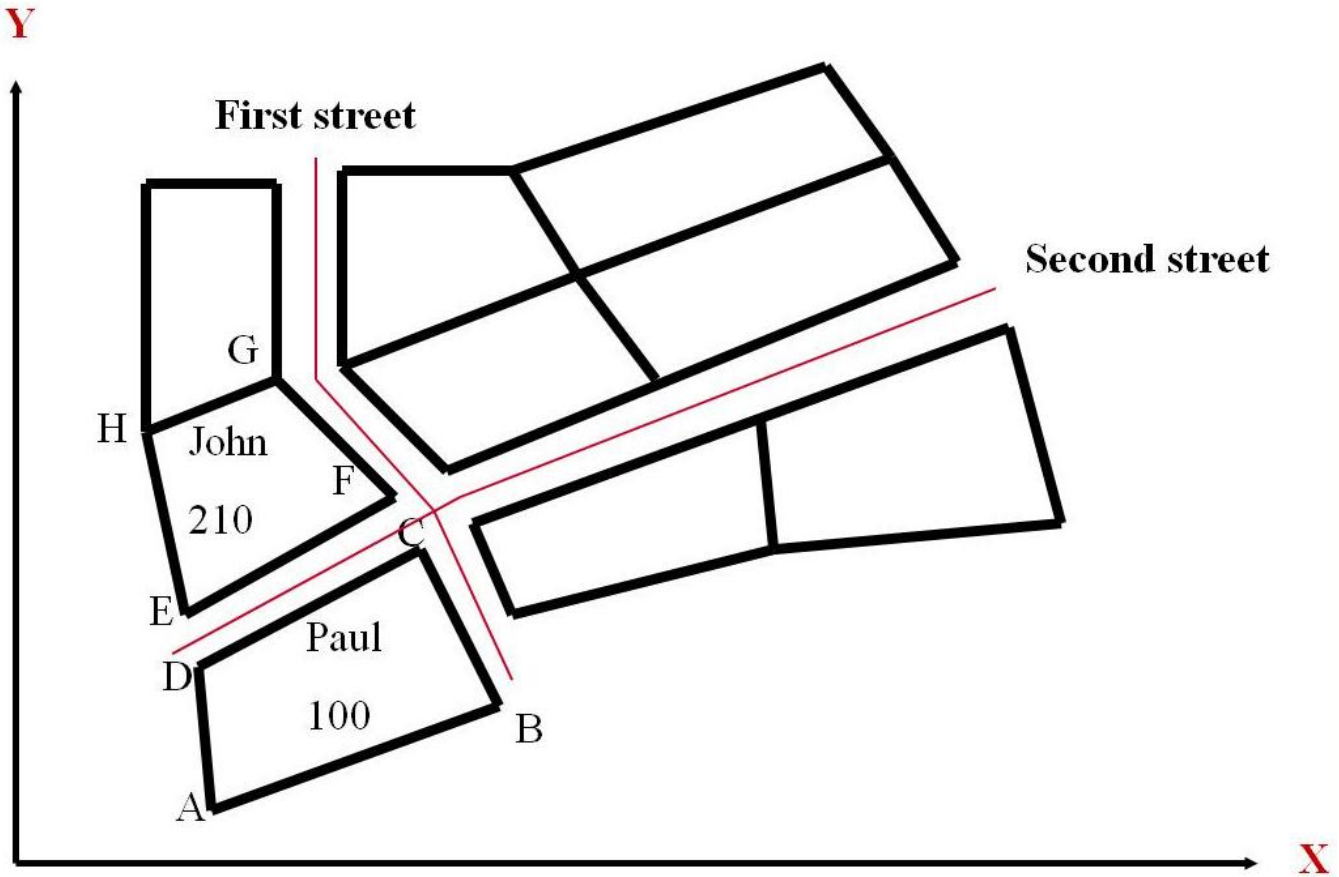
Type (attributes)

Instances

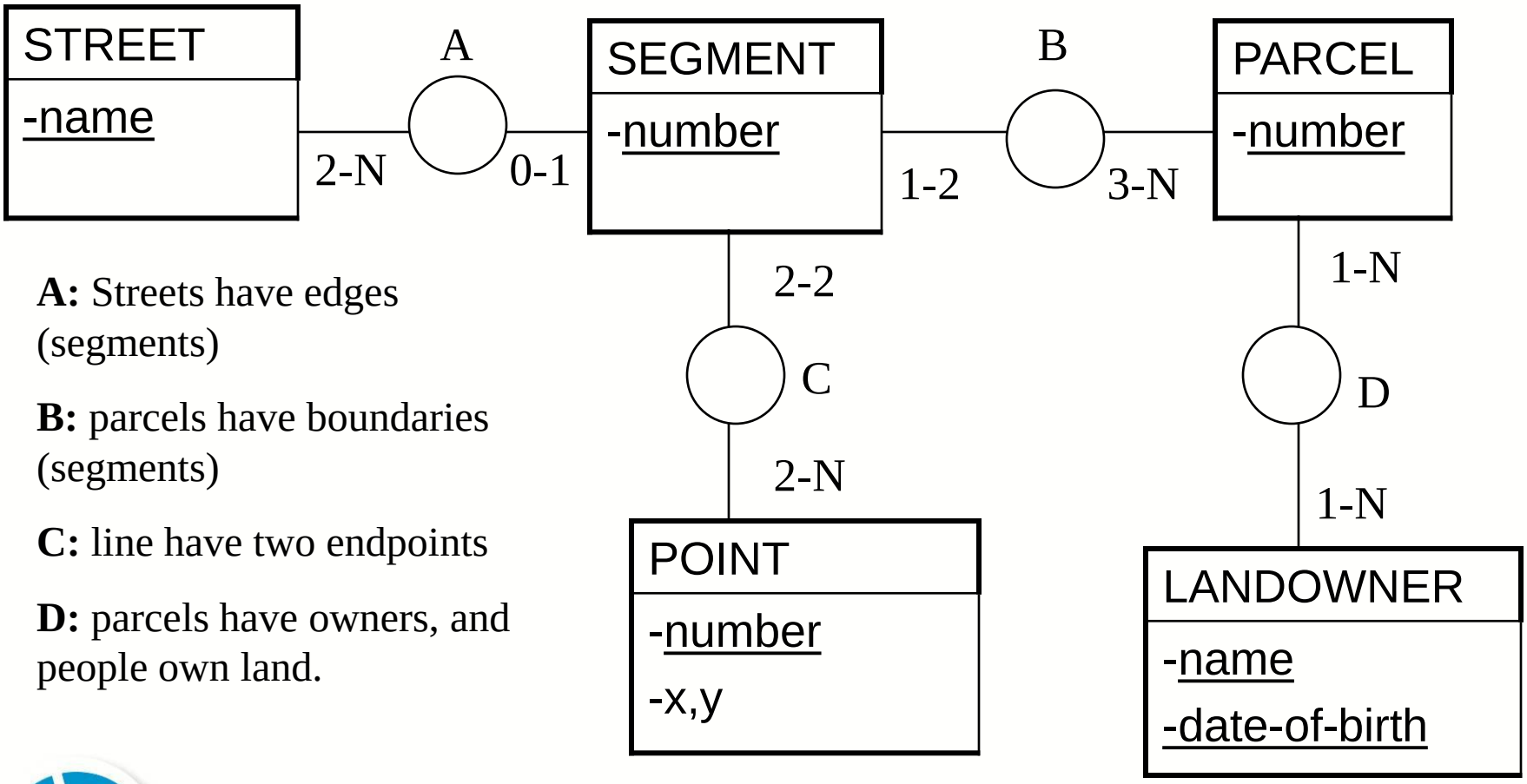
EA-Code	Area	Pop	CL-Code
723101	32.1	763	88
723102	28.4	593	88
723103	19.1	838	88
723201	34.6	832	88
723202	25.7	632	89
723203	28.3	839	89
723204	12.4	388	89
…	…	…	…

Primary key

An example of land parcels



The E/R diagram for land parcels



A: Streets have edges (segments)

B: parcels have boundaries (segments)

C: line have two endpoints

D: parcels have owners, and people own land.

Data Tables

STREET

Street-ID	Street_name
101	First street
102	Second Street

STREET- SEGMENT

Street-ID	Segment-ID
101	g
101	h
…	…
102	i
102	j
…	…

Segment-ID	Point1-ID	Point2-ID
a	A	B
b	B	C
c	C	D
…	…	…

POINT

Point-ID	X	Y
A	101.11	70.12
B	100.67	
145.33		
…	…	

Object-Oriented and Object-Relational GIS DBMS

- Object-oriented (OODBMS)
  - Based on OO concept to store state and behavior of GIS objects in databases
  - Provide OO query tools
  - Commercially not successful

- Object-Relational (ORDBMS)
  - Extend RDMS to handle GIS objects
  - Current Geographic Databases are ORDBMS

Data Dictionary

- Definition:

  A data catalog that describes the contents of a database. Information is listed about each field in the attribute table and about the format, definitions and structures of the attribute tables. A data dictionary is an essential component of metadata information.

Spatial Analysis: Query

- Select features by their attributes:
  - "find all districts with literacy rates < 60%"
- Select features by geographic relationships
  - "find all family planning clinics within this district"
- Combined attributes/geographic queries
  - "find all villages within 10km of a health facility that have high child mortality"

Query operations are based on the SQL (Structured Query Language) concept

Spatial Analysis (cont.)

- **Buffer:** find all settlements that are more than 10km from a health clinic
- **Point-in-polygon operations**: identify for all villages into which vegetation zone they fall
- **Polygon overlay**: combine administrative records with health district data
- **Network operations**: find the shortest route from village to hospital
- Advanced spatial analysis – Multi-criteria Analysis

Summary

- Data conversion
  - Digitizing/Scanning
  - Editing
  - Building Topology
- Data integration
  - Geo-referencing; Projection change
  - Coding
  - Integration of attribute data
- Building EA Database
  - Data Modelling: A spatial data model
  - Database Implementation

Thank You!

Illustration