GE2215 Lecture 8
Spatial Database and Attribute Data Management

Dr. Yan Yingwei
Department of Geography
National University of Singapore

Recap: Data Quality and Spatial Data Editing

- Spatial data quality
  - The propose of spatial data quality problem
  - Reflection of spatial data quality
  - Causes of spatial data quality problem
  - Characteristics of spatial data quality problem
- Spatial data editing
  - Topological errors and editing
  - Non-topological editing

Recap: Reflection of spatial data quality

- Location error
- Time error
- Attribute error

Recap: Causes of spatial data quality problem

1. Multiple data sources
2. Data entry is not strict
3. Too much data
4. Data standards by different users
5. Different encoding methods
6. Different data access limit

Recap: Characteristics of spatial data quality

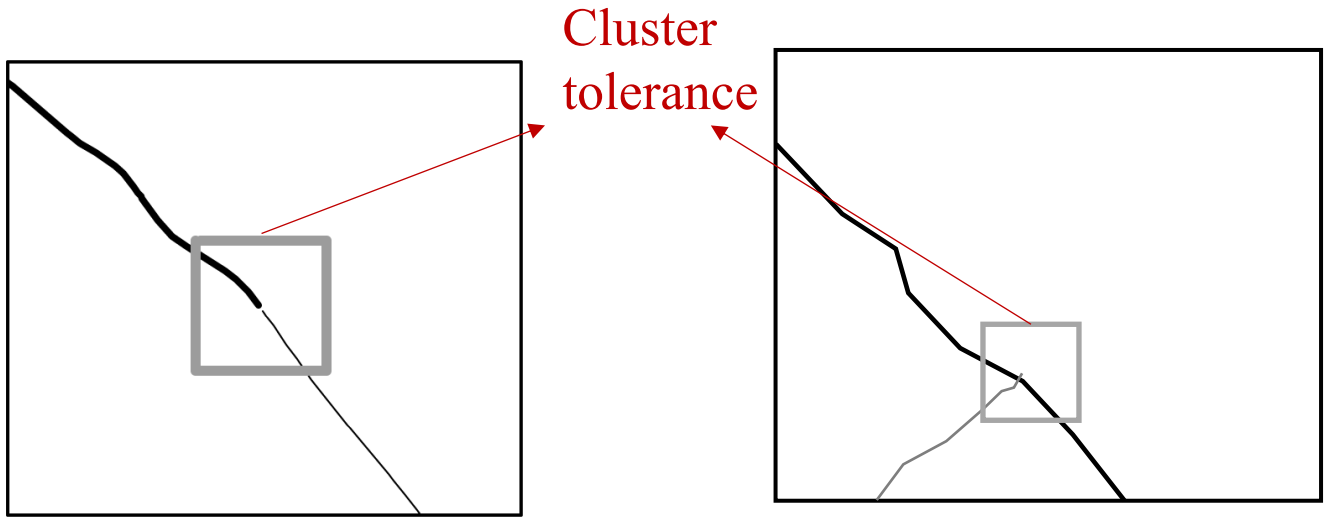
- Data accuracy and precision
  - What is data accuracy and data precision?
- Data uncertainty
  - Uncertainty of the real world itself
  - Uncertainty of the human cognition
- Data compatibility

Recap: Topological errors

- **Topological errors** violate topological relationship rules
- **Topological relationship rules** can be defined by:
  - The data model (e.g., a polygon must be closed)
  - The users (e.g., highways across two states must connect perfectly)
- **Topological relationship rules** can be defined:
  - Within a feature class (point, polyline or polygon)
  - Between feature layers

Recap: Topological editing – Cluster Tolerance

- Cluster tolerance is powerful for topological editing
- Cluster tolerance, also called XY tolerance, is used to snap vertices if they fall within a square area specified by the tolerance



Cluster tolerance is powerful in handling undershoot and overshoot

Recap: Topological editing – Topology Rules

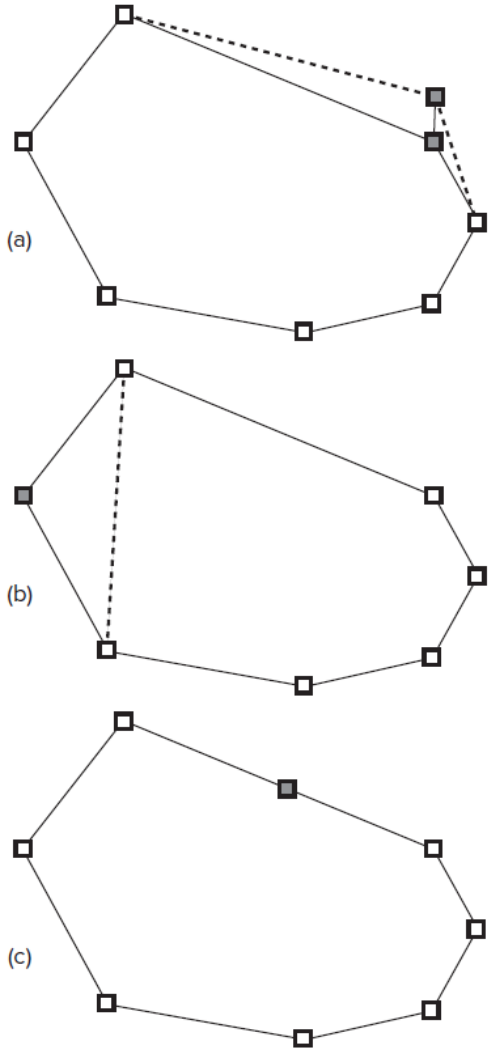
- **Topology rules** can be defined based on only one feature class or between spatial layers
- The geodatabase has more than 30 topology rules
- How to use topology rules to correct topological errors
  1. **Create** a new topology by defining the participating feature classes, the ranks of each feature class, the topology rules, and a cluster tolerance
  2. **Validation** of topology. Identify errors that have violated the topology rules
  3. **Fix** topological errors or accept them as exceptions (e.g., acceptable dangling nodes)

Recap: Non-topological editing

- Non-topological editing
  - Does not involve **topology** as defined in a **map topology** or a **topology rule**
  - Modify existing features
  - Create new features from existing features

Recap: Non-topological editing

- Modify existing features
  - Extend/trim lines
  - Delete/move features
  - Reshape features
  - Split lines and polygons



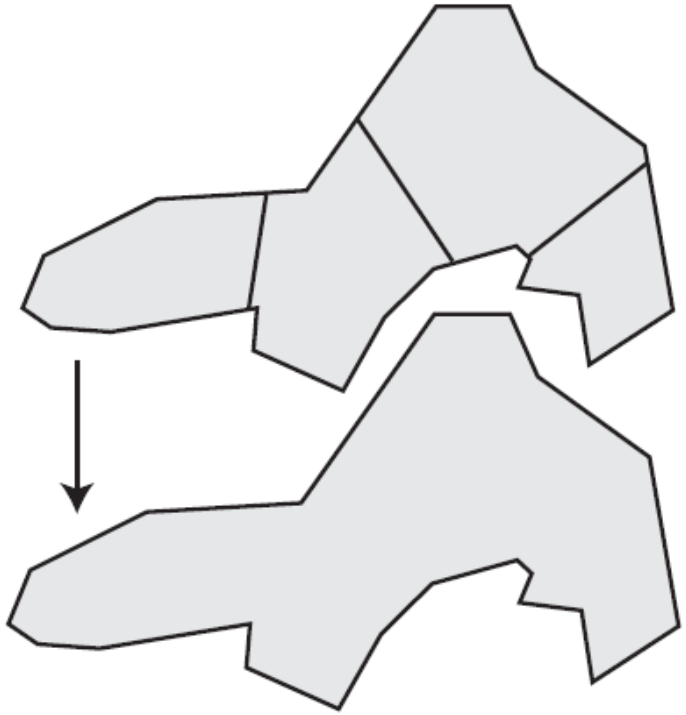
Reshape a polygon by **moving** a vertex

Reshape a polygon by **deleting** a vertex

Reshape a polygon by **adding** a vertex

Recap: Non-topological editing

- Create features from existing features
  - Merge features into one feature
  - Buffer features to create a buffer feature
  - Union features to combine features from the same layer
  - Intersect features to create a new feature



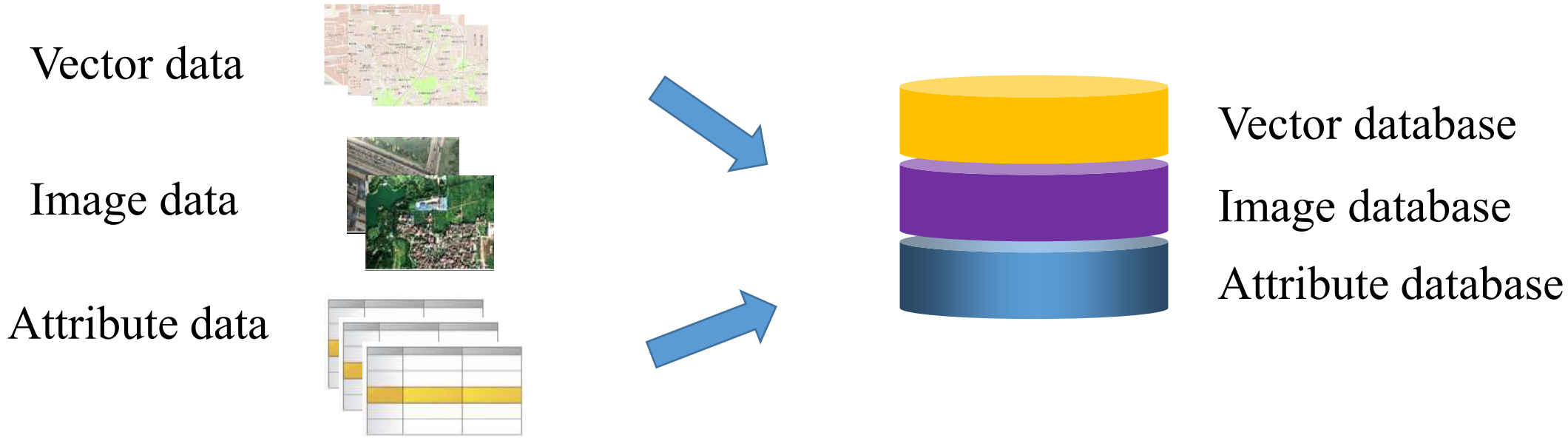
Merge four polygons into one

Outline of this lecture

- Basics of spatial database
- Basics of attributes
  - Types of attribute tables
  - Database management systems (DBMS)
  - Types of attribute data
- Insights into relational model

What is spatial database

- Database: an integrated set of data on a particular subject, which is often used to store, and organize data
- Spatial (Geographic) database: database containing geographic data of a particular subject for a particular area



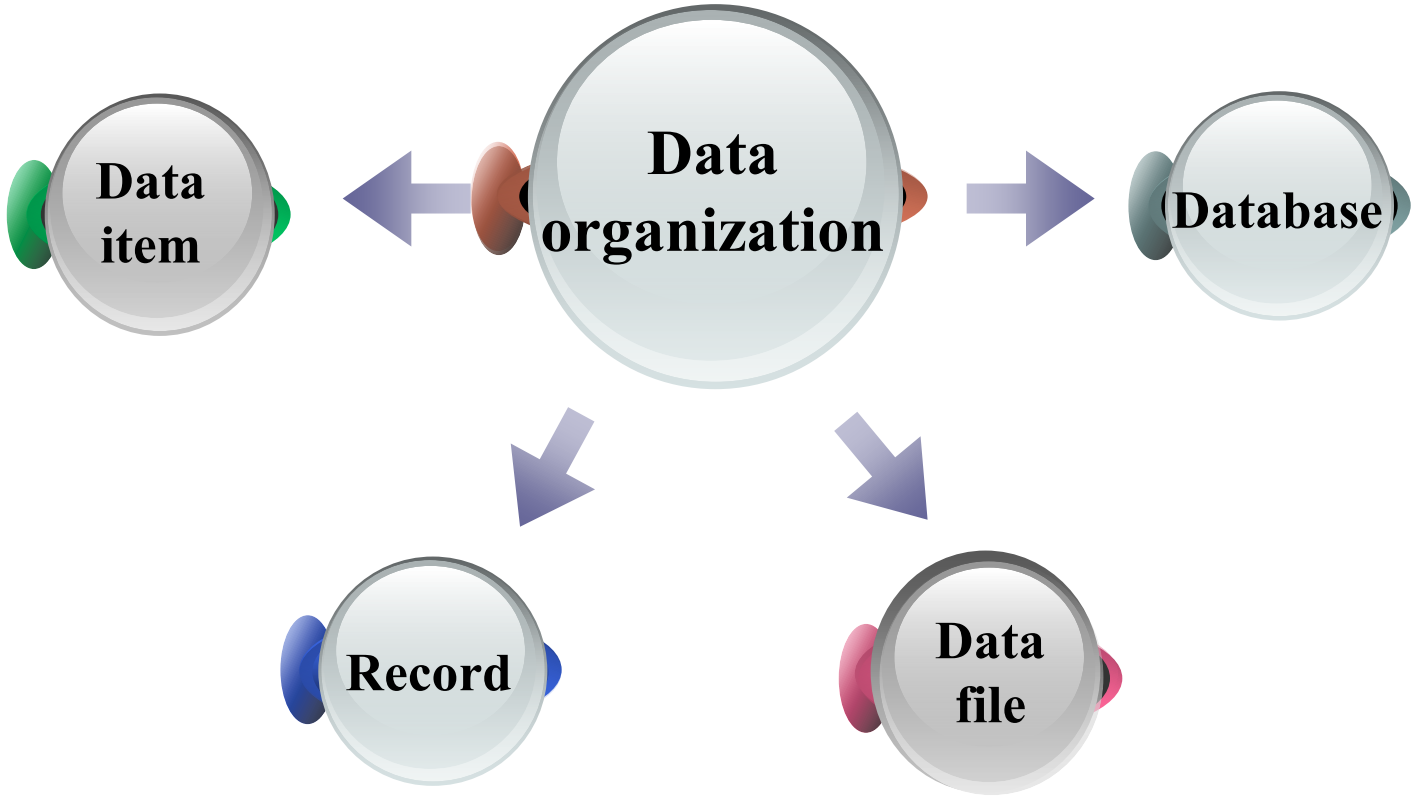
Characteristics of spatial database

- Data is under centralized control
  - Can guarantee data sharing among different users and applications
  - Different from file management in which files are dispersed
- Data are independent
  - Database is independent of the application systems, and thus can be called by various application systems
- Data redundancy is small
  - Avoid repetitive data storage
  - Improve data usage efficiency

Characteristics of spatial database

- Database has complex data model structure
  - The complex data model structure is used for data organization and data management
  - Vital difference from file management
- Database has the function of data protection
  - A password and permission for access must be set

Data organization



Data item: the smallest unit in defining data

Table

Eldercare_Centers

FID	Shape *	Id	Name	Weightage	mx	my
0	Point	0	NTUC Health Silver Circle (Jurong Central)	20	15663.848054	37273.653125
1	Point	0	NTUC Health Silver Circle (Jurong West)	20	12419.923633	35967.877208
2	Point	0	St Luke's ElderCare Jurong East Centre	20	16378.854313	36854.357065
3	Point	0	NTUC Health SilverACE	20	15722.954129	35294.181831
4	Point	0	THK Seniors Services@ Taman Jurong	20	15665.841685	35582.224371
5	Point	0	Silver circle/NTUC Health (Taman Jurong)	20	15266.487269	35724.816455
6	Point	0	NTUC Health/Cluster Support	20	15451.099331	35350.043489
7	Point	0	Lakeside FSC Jurong West	0	15467.095803	36413.090119
8	Point	0	Lakeside FSC Jurong East	0	17006.024963	36309.095375
9	Point	0	The Agape	0	15895.402695	34347.735379
10	Point	0	Loving Heart MSC	0	16735.385903	36513.027974
11	Point	0	Boon Lay Wellness centre	0	14850.530315	36492.748002
12	Point	0	Yuhua SAC	0	16617.00105	36639.215635
13	Point	0	Adventist Active Centre@Golden Peony	0	14898.918294	36676.519133

Data organization is classified into four levels

Data organization

- **Record**: composed by several relevant data items about one entity
- Each row is a record

Table

Eldercare_Centers

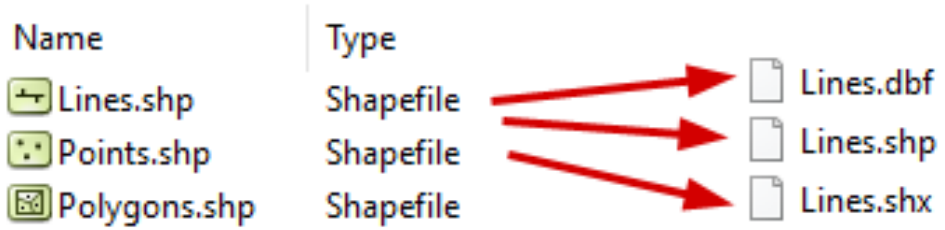
FID	Shape *	Id	Name	Weightage	mx	my
0	Point	0	NTUC Health Silver Circle (Jurong Central)	20	15663.848054	37273.653125
1	Point	0	NTUC Health Silver Circle (Jurong West)	20	12419.923633	35967.877208
2	Point	0	St Luke's ElderCare Jurong East Centre	20	16378.854313	36854.357065
3	Point	0	NTUC Health SilverACE	20	15722.954129	35294.181831
4	Point	0	THK Seniors Services@ Taman Jurong	20	15665.841685	35582.224371
5	Point	0	Silver circle/NTUC Health (Taman Jurong)	20	15266.487269	35724.816455
6	Point	0	NTUC Health/Cluster Support	20	15451.099331	35350.043489
7	Point	0	Lakeside FSC Jurong West	0	15467.095803	36413.090119
8	Point	0	Lakeside FSC Jurong East	0	17006.024963	36309.095375
9	Point	0	The Agape	0	15895.402695	34347.735379
10	Point	0	Loving Heart MSC	0	16735.385903	36513.027974
11	Point	0	Boon Lay Wellness centre	0	14850.530315	36492.748002
12	Point	0	Yuhua SAC	0	16617.00105	36639.215635
13	Point	0	Adventist Active Centre@Golden Peony	0	14898.918294	36676.519133

A column is also called a field

- **Keyword**
  - A data item that can be used to identify and differentiate the current record from other records

Which data item in the table on the left can be used as the **keyword**?

Data organization

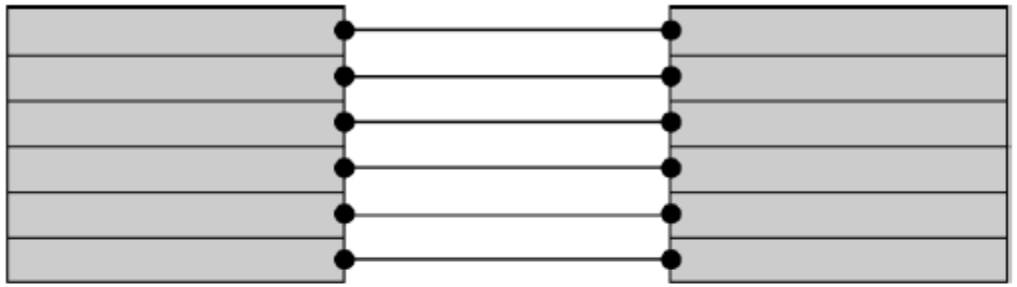


- File
  - A collection of the entire records
  - DBF file is the commonly used file format to record data tables

- Database
  - A collection of data with certain relevance
  - It is a collection of data files, which are dependent on each other and can not exist independently
  - .gdb is the geodatabase format in ArcGIS
  - PostgreSQL + PostGIS is used by QGIS

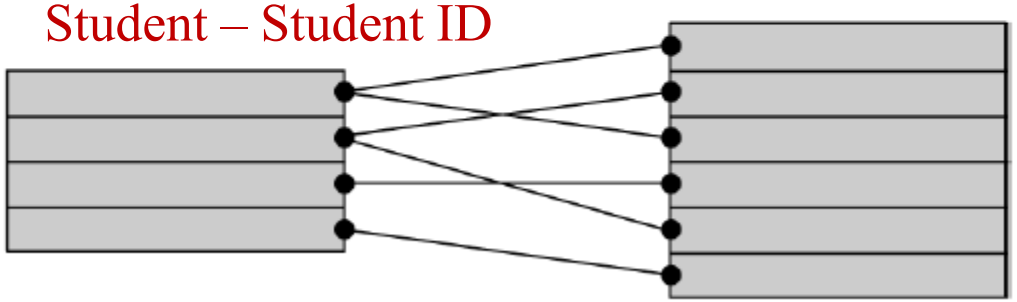
Logical relation types

- Logical relations among data refer to the connection between records



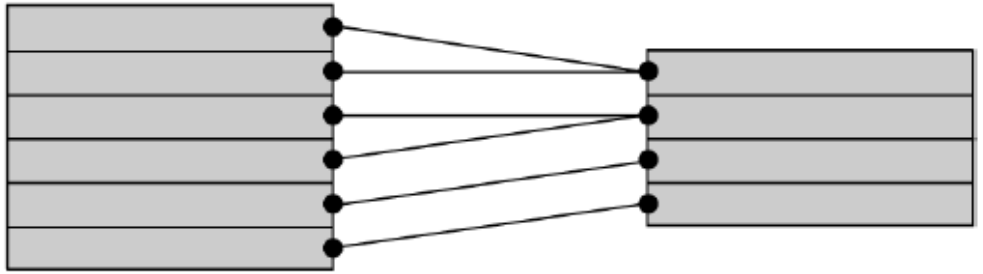
One-to-one relationship

Student – Student ID



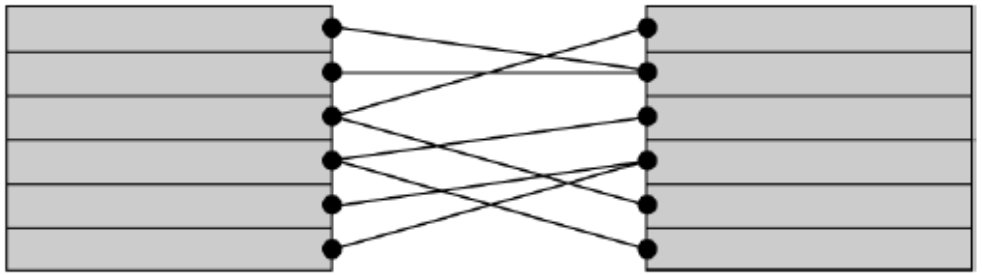
One-to-many relationship

State – County



Many-to-one relationship

Student – Department



Many-to-many relationship

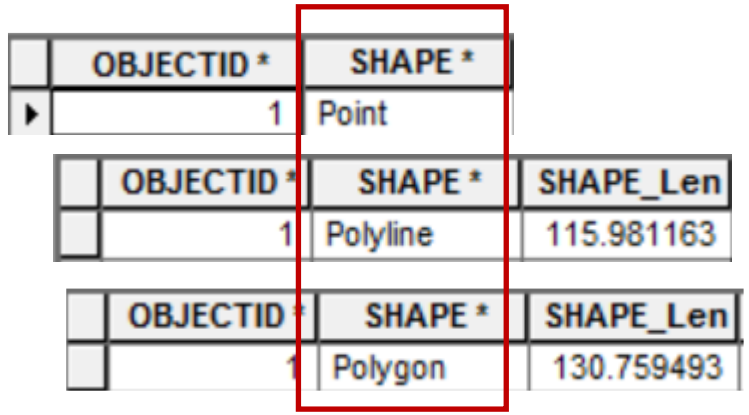
Student – Course

Outline of this lecture

- Basics of spatial database
- Basics of attributes
  - Types of attribute tables
  - Database management systems (DBMS)
  - Types of attribute data
- Insights into relational model

Types of attribute tables

- Feature attribute table
  - Has access to the feature geometry
  - Exist in every vector data set



OBJECTID *	SHAPE *
1	Point

OBJECTID *	SHAPE *	SHAPE_Len
1	Polyline	115.981163

OBJECTID *	SHAPE *	SHAPE_Len
1	Polygon	130.759493

- How is the feature geometry stored in feature attribute table?
  - Stored in a field in the feature attribute table (Object-based data model)
  - Linked by the feature ID (Georelational data model)

Types of attribute tables

- Non-spatial attribute table
  - Does not have direct access to the feature geometry
  - Has a field linking the table to the feature attribute table when necessary
  - Contains general information
  - Delimited txt, dBASE files, Excel, Access…

	OBJECTID *	CNTRY_NA	SOVEREIGN	ISO_3_COD	y_2004
	8	Singapore	Singapore	SGP	6.601
	4	Lao People's	Lao People's	LAO	23.1
	6	Myanmar	Myanmar	MMR	29.57
	1	Brunei Darus	Brunei Darus	BRN	60.226
	2	Cambodia	Cambodia	KHM	70.42
	10	Viet Nam	Viet Nam	VNM	241
	5	Malaysia	Malaysia	MYS	1128.543
▶	9	Thailand	Thailand	THA	1358.32
	7	Philippines	Philippines	PHL	1389.81
	3	Indonesia	Indonesia	IDN	3925.47

Database management systems (DBMS)

- Database management system (DBMS)
  - A system to manage tables
  - A software package that enables people to build and manipulate a database
  - Most GIS packages include DBMS tools for local databases
    - Microsoft Access is used by ArcGIS 
    - PostgreSQL + PostGIS is used by QGIS 

  - Not only used in GIS applications but also used in other information system

Database management systems (DBMS)

- Functions of DBMS
  - File handling and file management
  - Adding/deleting/updating records
  - Provides tools for data input, search, retrieval, manipulation, output
  - Maintaining data security

Types of attribute data

- Method I – Classifying by **data type**
  - Number
    - Integer (short int or long int) (e.g., 1234)
    - Float (e.g., 1.234)
    - Double (e.g., 1.7976931348623 2E308)
  - Text (or string) (e.g., abcd)
  - Date (e.g., 03/01/2016)
  - Binary large object (blob): A collection of binary data stored as a single entity
    - images, audio, multimedia, and feature geometries as long sequences of binary numbers

Types of attribute data

- Method II – Classifying by measurement scale
  - **Nominal data**
    - Describes different categories of data, e.g., land-use types, soil types
  - **Ordinal data**
    - Differentiate data by a ranking relationship
    - E.g., Severe – moderate – light soil erosion, low – moderate – high risk

Types of attribute data

- Method II – Classifying by measurement scale
  - **Interval data**
    - Have known intervals between values **(can represent values below zero)**
    - E.g., temperature, elevation
  - **Ratio data**
    - Similar with interval data, but are based on an absolute **zero value (never fall below zero)**
    - E.g., population density, crime rate

Could you give me another example of ratio data?

Types of attribute data

- Cell values of **raster** data
  - **Categorical**
    - Include nominal and ordinal data
  - **Numeric**
    - Include interval and ratio data

Outline of this lecture

- Basics of spatial database
- Basics of attributes
  - Types of attribute tables
  - Database management systems (DBMS)
  - Types of attribute data
- Insights into relational model

Four types of databases

- A flat file
- Hierarchical database
- Network database
- Relational database

A flat file

- A flat file
  - A flat file contains all data in a large table
  - A feature attribute table is like a flat file
  - The spreadsheet with attribute data only

(a) Flat file

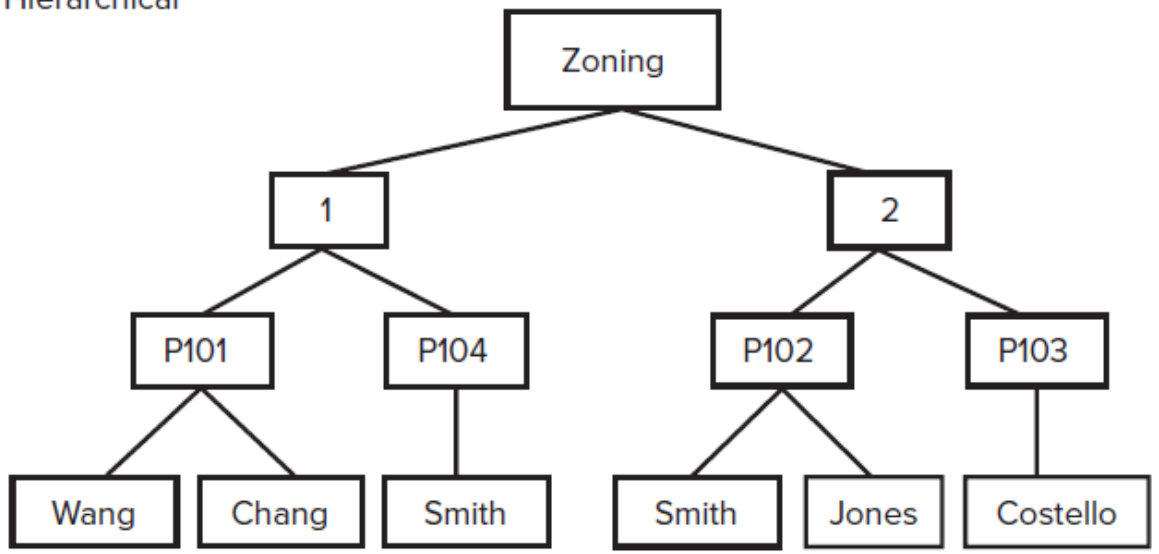
PIN	Owner	Zoning
P101	Wang	Residential (1)
P101	Chang	Residential (1)
P102	Smith	Commercial (2)
P102	Jones	Commercial (2)
P103	Costello	Commercial (2)
P104	Smith	Residential (1)

Hierarchical database

- A hierarchical database
  - It organizes its data at different levels
  - It uses one-to-many association between levels
  - Each level is divided into different branches

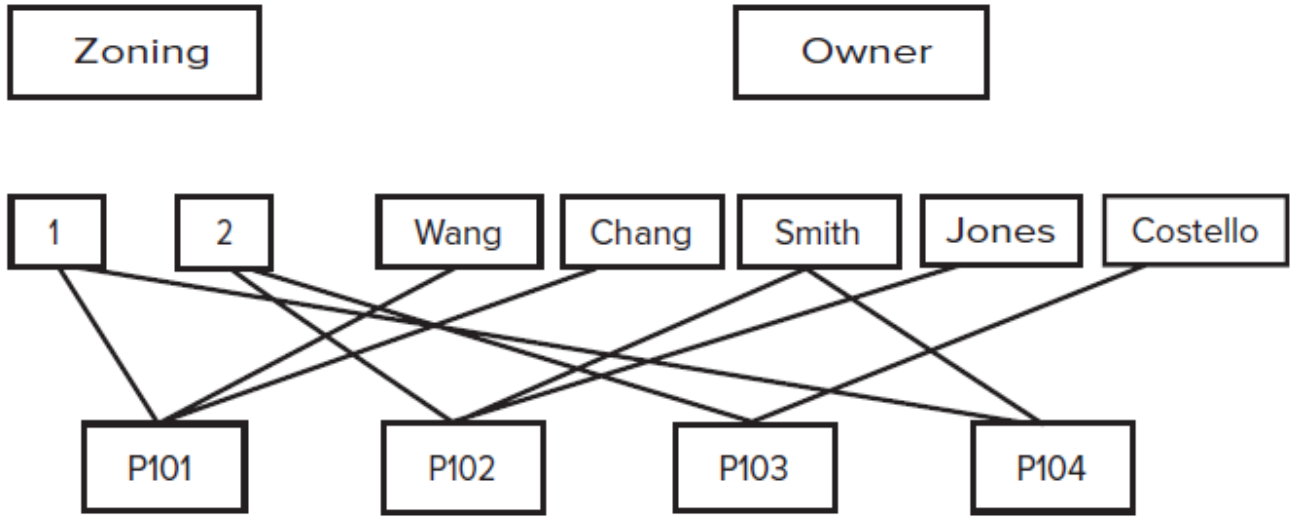
PIN	Owner	Zoning
P101	Wang	Residential (1)
P101	Chang	Residential (1)
P102	Smith	Commercial (2)
P102	Jones	Commercial (2)
P103	Costello	Commercial (2)
P104	Smith	Residential (1)

(b) Hierarchical



Network database

- A network database
  - It builds connections across tables
  - Many-to-many association between levels

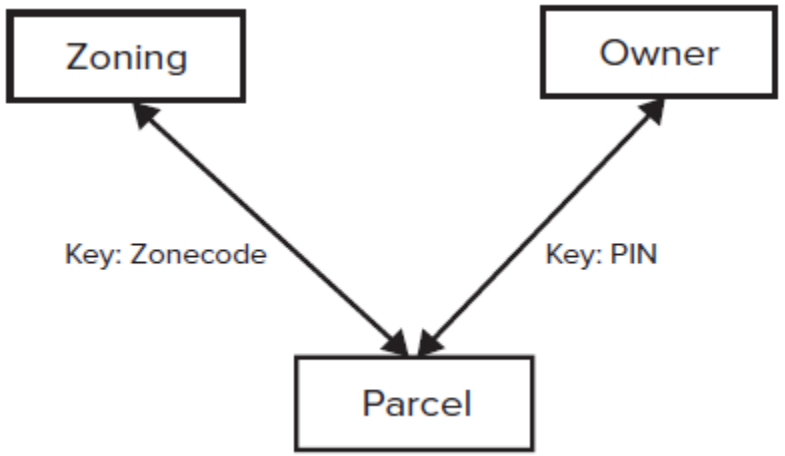


PIN	Owner	Zoning
P101	Wang	Residential (1)
P101	Chang	Residential (1)
P102	Smith	Commercial (2)
P102	Jones	Commercial (2)
P103	Costello	Commercial (2)
P104	Smith	Residential (1)

Relational database

- A relational database
  - It is a collection of tables, also called relations
  - The tables are connected to each other by keys
  - A **primary key**: represents one or more attributes whose values can uniquely identify a record in a table
  - A **foreign key**: is one or more attributes that refer to a primary key in another table

Relational database



PIN	Owner	Zoning
P101	Wang	Residential (1)
P101	Chang	Residential (1)
P102	Smith	Commercial (2)
P102	Jones	Commercial (2)
P103	Costello	Commercial (2)
P104	Smith	Residential (1)

Primary key Foreign key

Parcel table

PIN	Zone code
P101	1
P102	2
P103	2
P104	1

Foreign key

Owner table

PIN	Owner name
P101	Wang
P101	Chang
P102	Smith
P102	Jones
P103	Costello
P104	Smith

Primary key

Zone table

Zone code	Zoning
1	Residential
2	Commercial

Relational database

- Advantages of relational database
  1. Each table in the database can be prepared, maintained, and edited separately from other tables
     - This is important as more GIS data are being recorded and added
  2. The tables can remain separate until a query or an analysis requires that attribute data from different tables be linked together, which is favorable to both data management and data processing

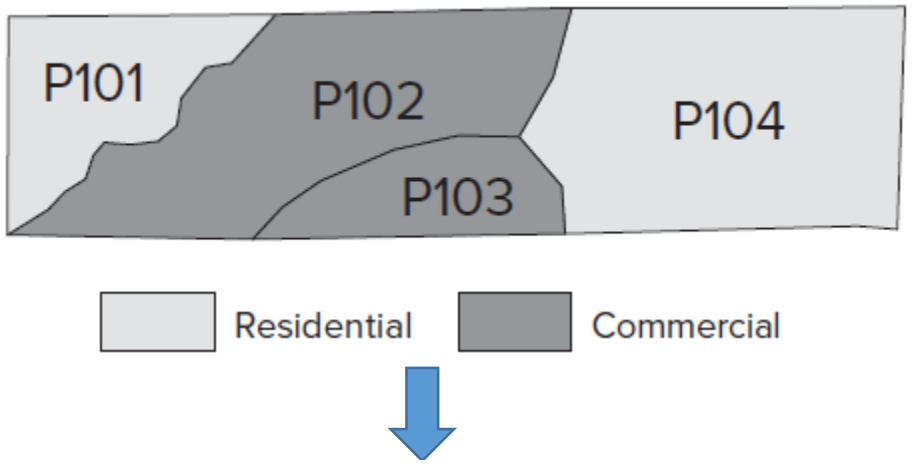
ArcGIS Geodatabase & PostgreSQL + PostGIS

- GIS data
  - Spatial data
  - Attribute data
- Adopt the relational database
- Link the two components and integrates both spatial and attribute data into a single database
- The linkage ensures GIS to be capable of handling feature geometries and the spatial relationships between features

Normalization: Preparing a relational database (optional – for your interest)

- **Normalization** is the process of decomposition, taking a table with all the attribute data, and breaking it down into small tables while maintaining the necessary linkages between them
- Objectives of normalization:
  - To avoid redundant data
  - To ensure that attribute data in separate tables can be maintained and updated separately and can be linked whenever necessary
  - To facilitate a distributed database

An unnormalized table



- The map shows four land parcels with the PINs of P101, P102, P103, P104
- Two parcels are zoned residential, and two others commercial

PIN	Owner	Owner Address	Sale Date	Acres	Zone Code	Zoning
P101	Wang	101 Oak St	1-10-98	1.0	1	Residential
	Chang	200 Maple St				
P102	Smith	300 Spruce Rd	10-6-68	3.0	2	Commercial
	Jones	105 Ash St				
P103	Costello	206 Elm St	3-7-97	2.5	2	Commercial
P104	Smith	300 Spruce Rd	7-30-78	1.0	1	Residential

Normalization: Step 1 – fill empty cells

PIN	Owner	Owner Address	Sale Date	Acres	Zone Code	Zoning
P101	Wang	101 Oak St	1-10-98	1.0	1	Residential
P101	Chang	200 Maple St	1-10-98	1.0	1	Residential
P102	Smith	300 Spruce Rd	10-6-68	3.0	2	Commercial
P102	Jones	105 Ash St	10-6-68	3.0	2	Commercial
P103	Costello	206 Elm St	3-7-97	2.5	2	Commercial
P104	Smith	300 Spruce Rd	7-30-78	1.0	1	Residential

- Step 1 fills the empty cells, and each cell has one value
- But the problem of data redundancy has increased

Step 2 – decompose the larger table

Primary key

Parcel table

PIN	Sale date	Acres	Zone code	Zoning
P101	1-10-98	1.0	1	Residential
P102	10-6-68	3.0	2	Commercial
P103	3-7-97	2.5	2	Commercial
P104	7-30-78	1.0	1	Residential

Foreign key

Foreign key

Owner table

PIN	Owner name
P101	Wang
P101	Chang
P102	Smith
P102	Jones
P103	Costello
P104	Smith

Primary key

Address table

Owner name	Owner address
Wang	101 Oak St
Chang	200 Maple St
Jones	105 Ash St
Smith	300 Spruce Rd
Costello	206 Elm St

- Step 2 decomposes the larger table into three small tables
- There is data redundancy with the fields of zone code and zoning

Step 3 – further decomposition

Parcel table

PIN	Sale date	Acres	Zone code	Zoning
P101	1-10-98	1.0	1	Residential
P102	10-6-68	3.0	2	Commercial
P103	3-7-97	2.5	2	Commercial
P104	7-30-78	1.0	1	Residential

- Step 3 further decomposes the parcel table into smaller tables

Parcel table

PIN	Sale date	Acres	Zone code
P101	1-10-98	1.0	1
P102	10-6-68	3.0	2
P103	3-7-97	2.5	2
P104	7-30-78	1.0	1

Zone table

Zone code	Zoning
1	Residential
2	Commercial

The process repeats until none of the tables can be further decomposed

Table join

- A join operation brings together two tables by using a common field or a primary key and a foreign key

Origin table

PIN	Sale date	Acres	Zone code
P101	1-10-98	1.0	1
P102	10-6-68	3.0	2
P103	3-7-97	2.5	2
P104	7-30-78	1.0	1

Destination table

Zone code	Zoning
1	Residential
2	Commercial

Table join: 1:1 or M:1

Table relate

- A relate operation temporarily connects two tables but keeps the tables physically separate
- Does not append the date from one table to another
- Three or more tables can be simultaneously connected
- Support all relationships

Parcel table

PIN	Sale date	Acres	Zone code	Zoning
P101	1-10-98	1.0	1	Residential
P102	10-6-68	3.0	2	Commercial
P103	3-7-97	2.5	2	Commercial
P104	7-30-78	1.0	1	Residential

Owner table

PIN	Owner name
P101	Wang
P101	Chang
P102	Smith
P102	Jones
P103	Costello
P104	Smith

Address table

Owner name	Owner address
Wang	101 Oak St
Chang	200 Maple St
Jones	105 Ash St
Smith	300 Spruce Rd
Costello	206 Elm St

Spatial join

Read this:
https://storymaps.arcgis.com/stories/85f6170907de460ea7bec930a1b3f748

- A spatial join uses a spatial relationship to join two sets of spatial features and their attribute data

- Join a school to a county in which the school is located

Containment | Intersect | Proximity

- Join a highway to a forest area by which the highway is intersected
- Join a villages to a fault line which the village is closest to

Summary

- Basics of spatial database
  - What is database and what is spatial database?
  - Five characteristics of spatial database
  - Four levels of data organization
  - Four types of relation types

Summary

- Basics of attributes
  - Types of attribute tables
    - Feature attribute table and non-spatial attribute table
  - Database management systems (DBMS)
    - Functions of DBMS
  - Types of attribute data
    - Classifying by data type: number, text, date, BLOB
    - Classifying by measurement scale: nominal, ordinal, interval, ratio

Summary

- Insights into relation model
  - Four types of database
  - Relational database and its advantages
  - Table join, table relate, spatial join

THANK YOU