Data quality and validation

IDEA MAP SUDAN

Part of the material Adapted from presentation prepared by Jon Wang , Monika Kuffer
ITC, University of Twente
And Gis Gate

October 2022

African Population and Health Research Center
ITC UNIVERSITY OF TWENTE.
ULB
Nuffic meet the world

IDEAMAP SUDAN

Overview

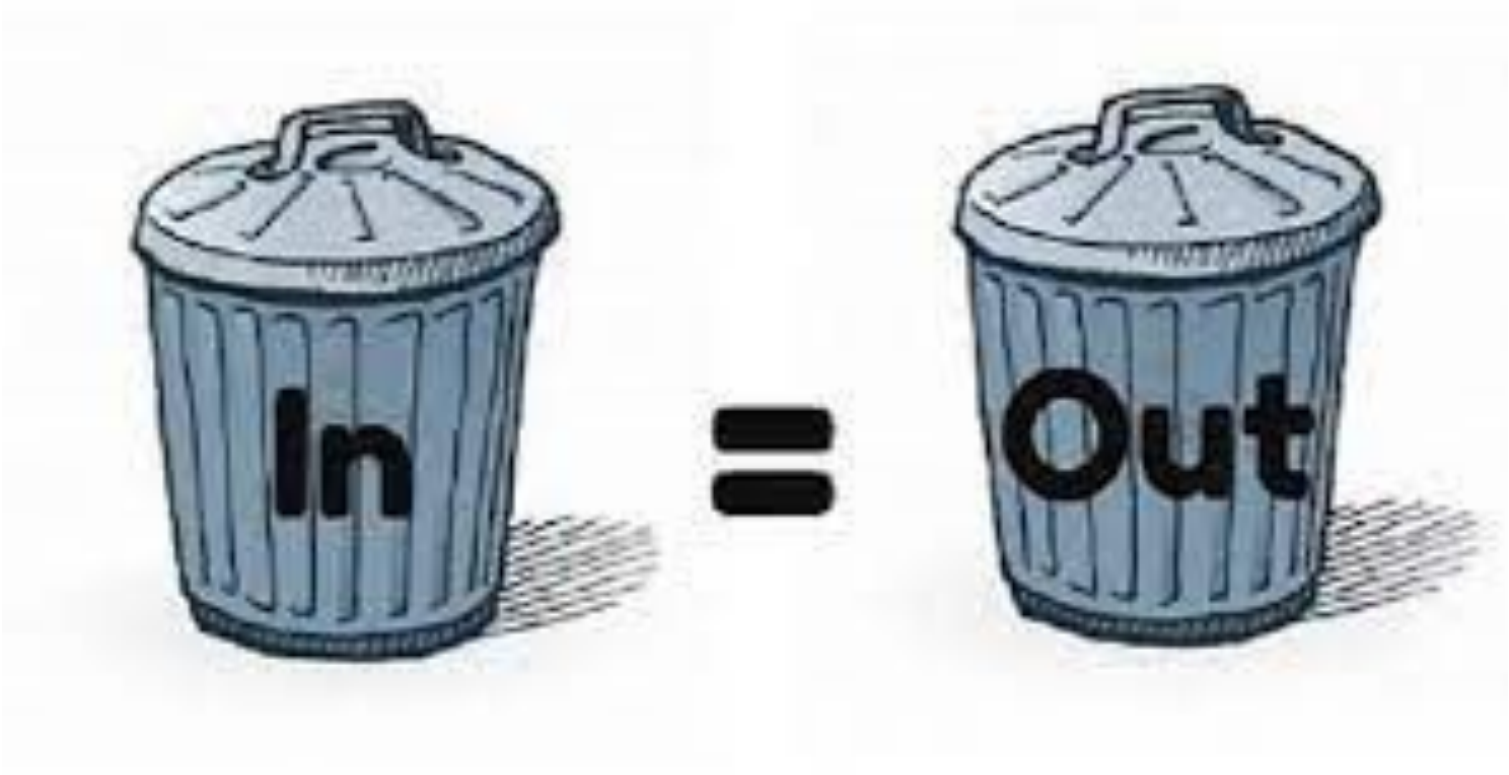
Program of the day

- Importance of data quality
- Definitions
- Topology
- Validation

IDEAMAP SUDAN

Data quality

Why data quality is important issue?



IDEAMAP SUDAN

Why data quality is important issue?

Valid Data

↳ Valid Analysis

↳ Valid Result

IDEAMAP SUDAN

Data quality

Basic Definition and Terminology

Moving from Accuracy precision uncertainty and Errors To Quality.

- **Accuracy :**

  - Closeness of observation to it's true value
  - Accurate measurement has a mean close to the true value

IDEAMAP SUDAN

Data quality

Basic Definition and Terminology

- **Precision**

  - The smallest unit of measurements to which data can be recorded

  - A precise measurements has small variance

Data quality

Basic Definition and Terminology

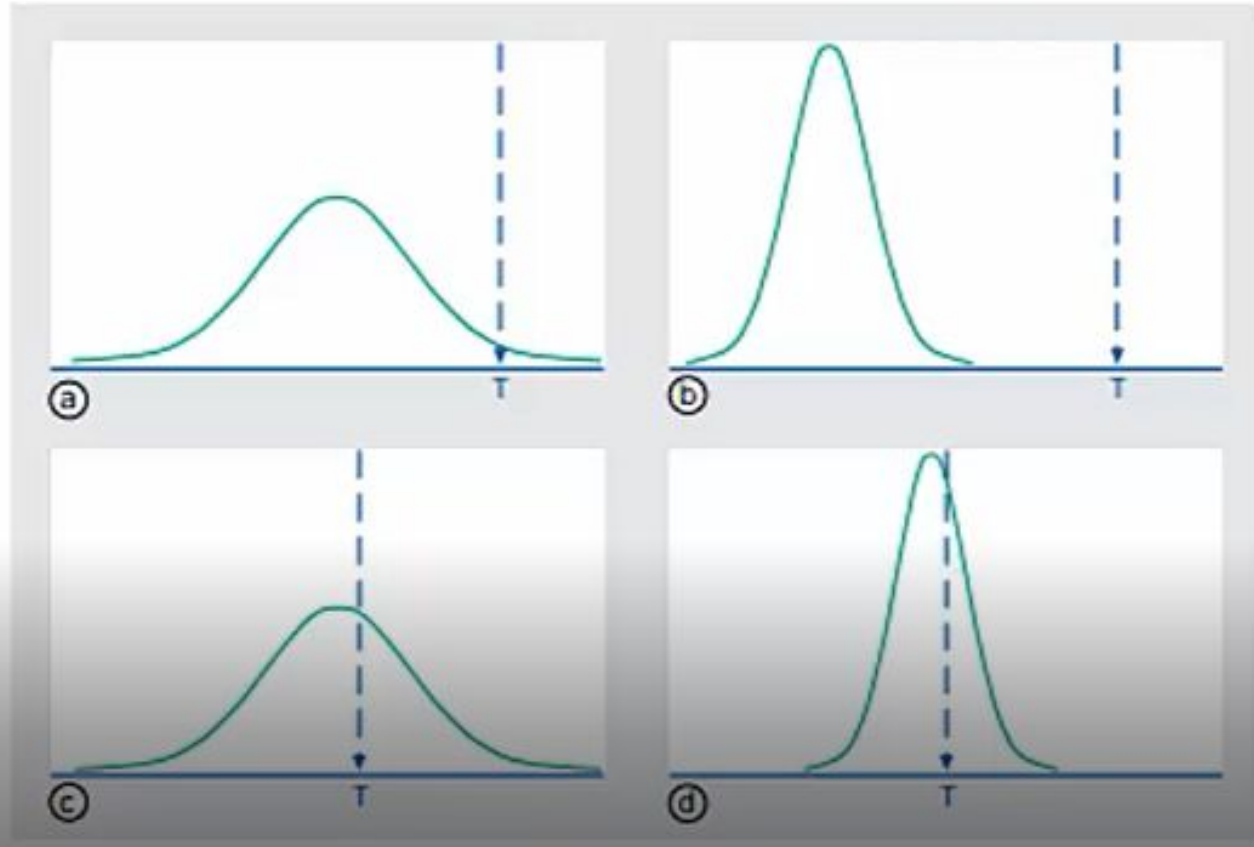


Figure 8.50
A measurement probability function and the underlying true value T: (a) bad accuracy and precision, (b) bad accuracy/good precision, (c) good accuracy/bad precision, and (d) good accuracy and precision.

IDEAMAP SUDAN

Data quality

Basic Definition and Terminology

- Error :

The different between measurement value and a true or theoretically correct value

- Type of errors:
  - Human errors (collecting, digitizing ,interpretation….
  - Instruments error (systematic error not calibrated instrument)
  - Random errors (natural variation , small error)

IDEAMAP SUDAN

Basic Definition and Terminology

- **Uncertainty:**

  - Simply you are not sure about your measurements
  - Uncertainty is not an error
  - Errors usually originate from uncertainty

IDEAMAP SUDAN

Data quality

Basic Definition and Terminology

- **Data Quality :**

  Characteristic of a product or service that bear on it's ability to satisfy stated and implied needs

  - Quality is relative
  - Quality doesn't mean excellence
  - Quality is measured as fitness for purpose

IDEAMAP SUDAN

Data quality

Elements of spatial data quality :

- Accuracy (positional, Thematic)
- Lineage (family tree of the data set)
- Logical consistency (value should not be there)
- Completeness (you have the geometry but not have the names)

These and other data what expected to see in meta data

IDEAMAP SUDAN

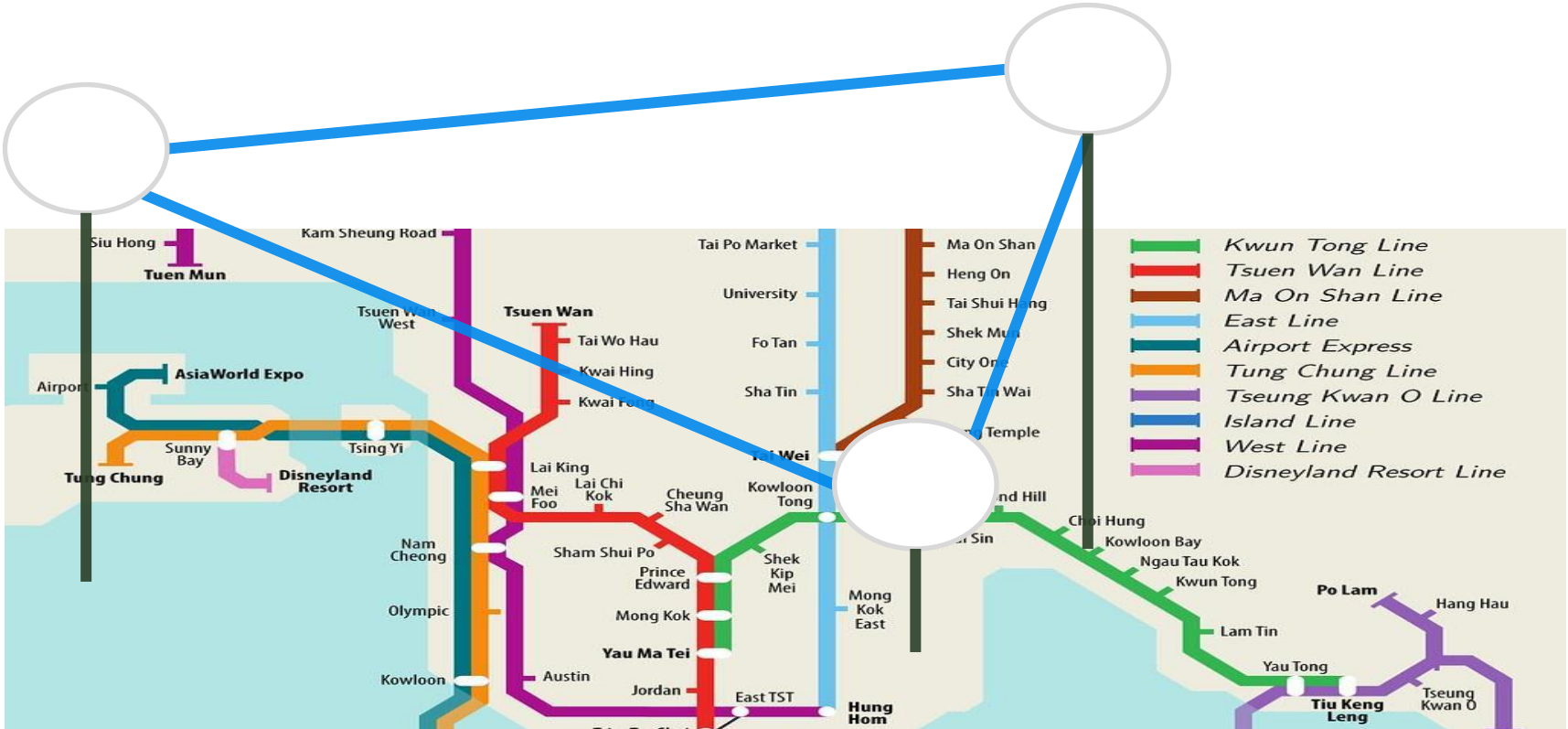
How to deal with incompleteness or erroneous data in general

- Completeness is not always their so managing data gaps is essential in research
- There is no single approach to solve this problem but basic principles to keep in mind:
- Be consistent especially if the goal is to compare
- Depending on the nature of the data (estimating , deleting , ignoring) missing values might be an option
- Always document what you did and be consistence

IDEAMAP SUDAN

What is a Topology?

- Essentially, topology refers to the relationship between things.
- In GIS, topology refers to the relationship between spatial objects.



IDEAMAP SUDAN

Data quality

Topology

- What is topology?

In GIS relation between spatial objects

- Why topology is important ?

Topology can guarantee that the acquired data is valid

⟶ Valid analysis ⟶ Valid results

Spatial data Data quality

Topology

Topology relationship:

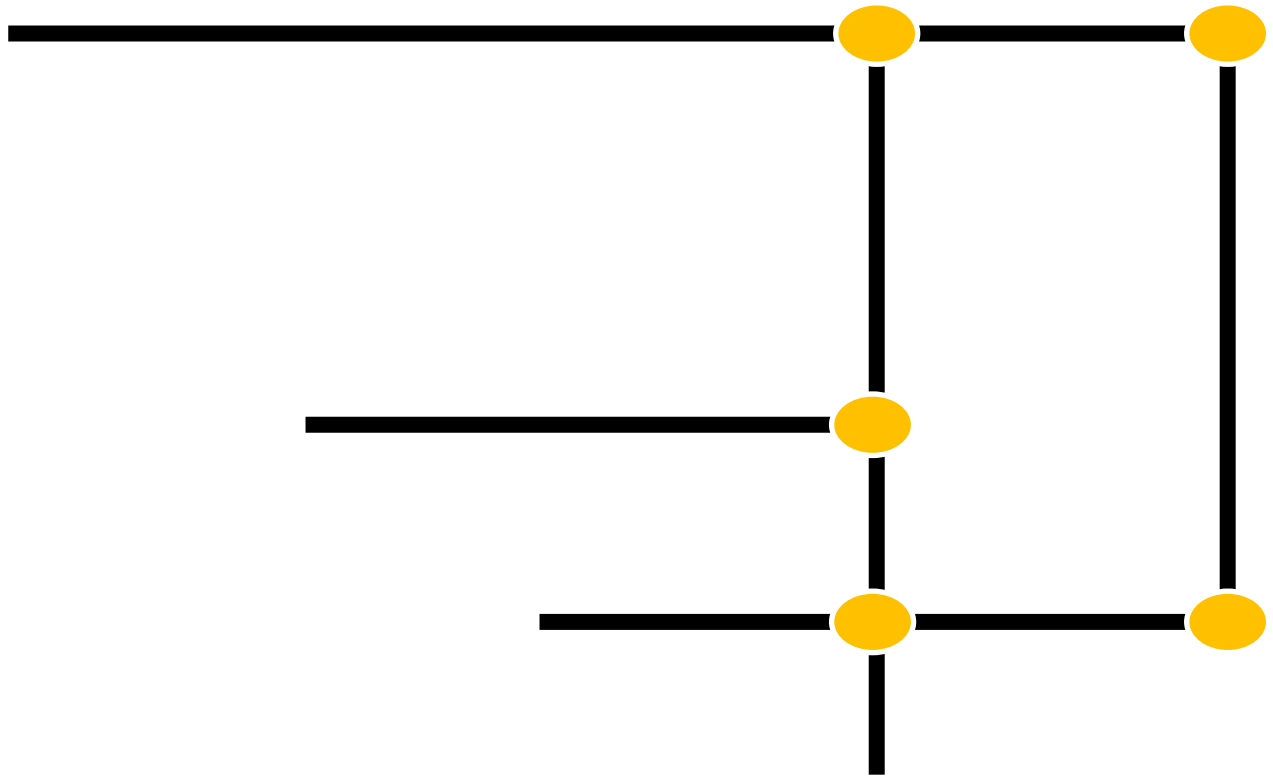
- Connectivity
- Adjacency
- Inclusion

IDEAMAP SUDAN

Topological Relationship

- **Connectivity**

Describes how lines are connected to each other to form a network



IDEAMAP SUDAN

Topological Relationships

- **Adjacency**

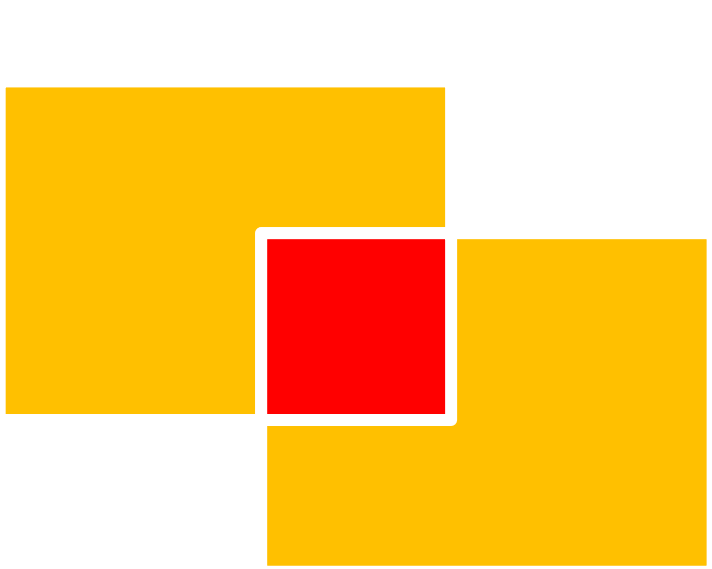
Describes whether two areas are next to each other.

IDEAMAP SUDAN

Topological Relationships

- **Inclusion**

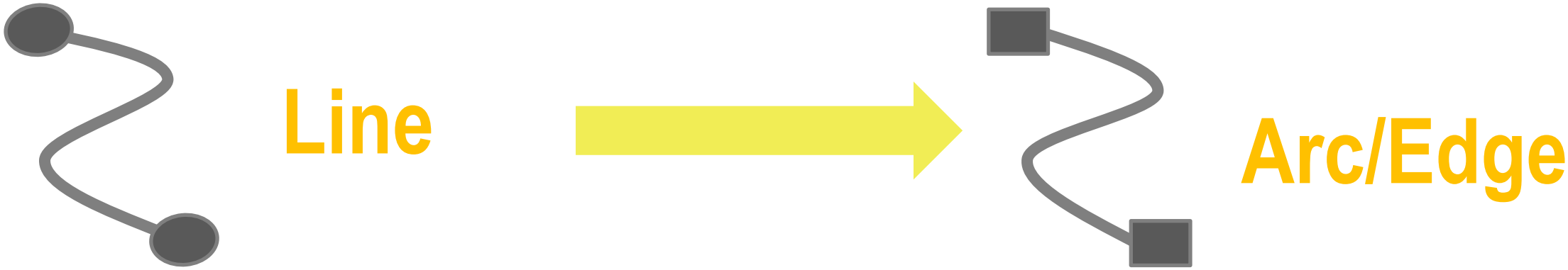
Describes whether two areas are nested.



IDEAMAP SUDAN

Topology Data Structure

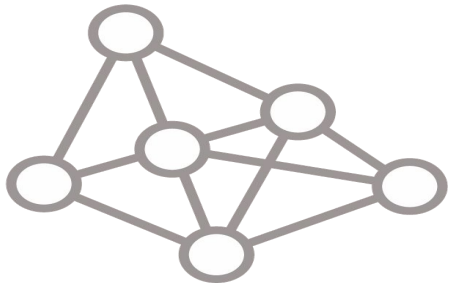
The most common topological data structure is the **Arc/Node** data model



IDEAMAP SUDAN

How does topology work?

- Defining topology rules
- Validating data
- Resolving violation



IDEAMAP SUDAN

Data quality

Topology

Topology Rules:

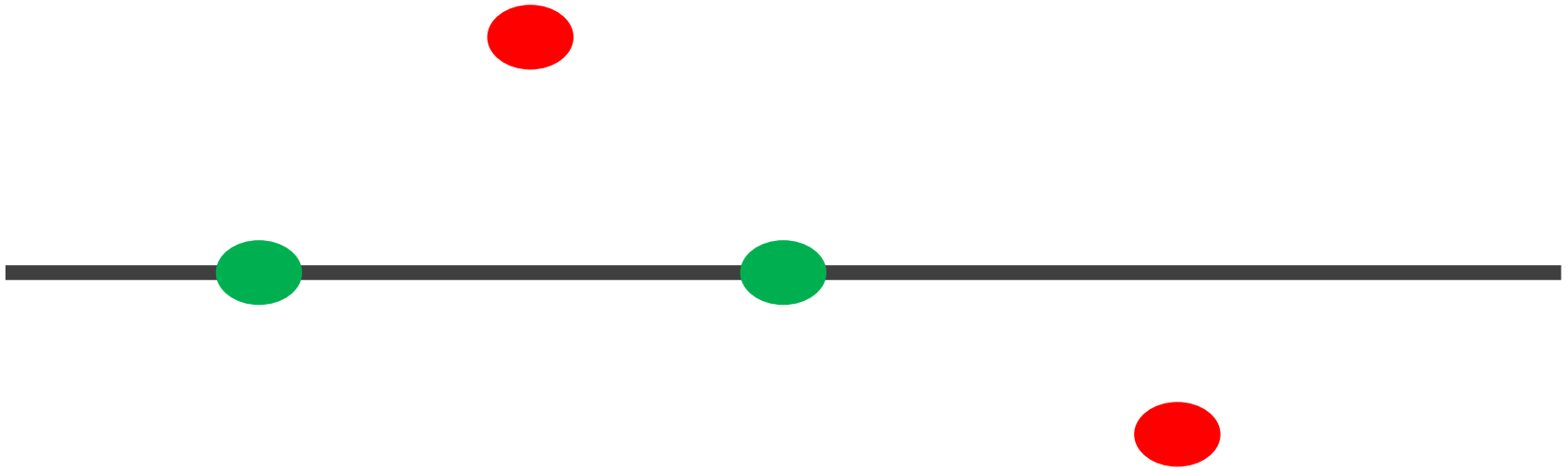
- Define rules and constrains for the feature classes
- +30 different type of topology rules
- Applied in point, line and polygon feature
- Violation expressed by topology errors

IDEAMAP SUDAN

Samples of Topology Rules

- **Point Geometry**

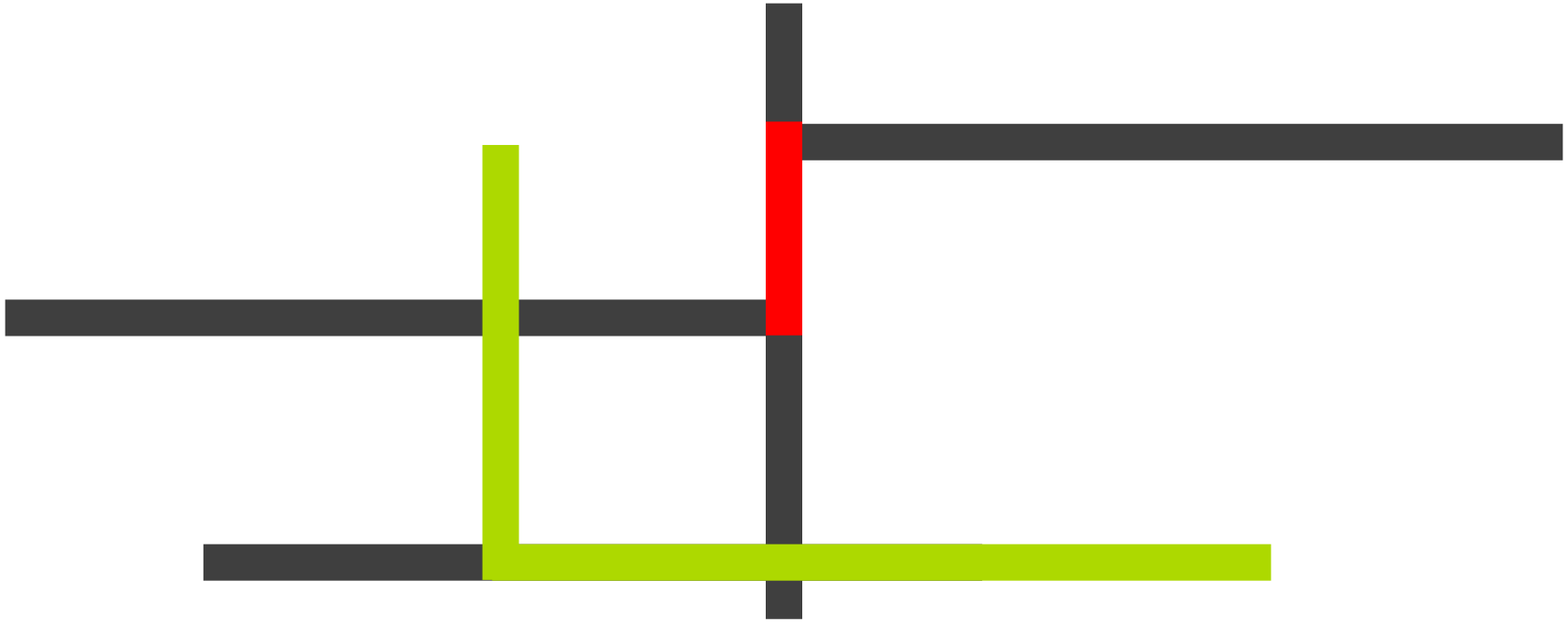
Points Must Be Covered By Line



Samples of Topology Rules

- **Line Geometry**

Lines Must Not Overlap

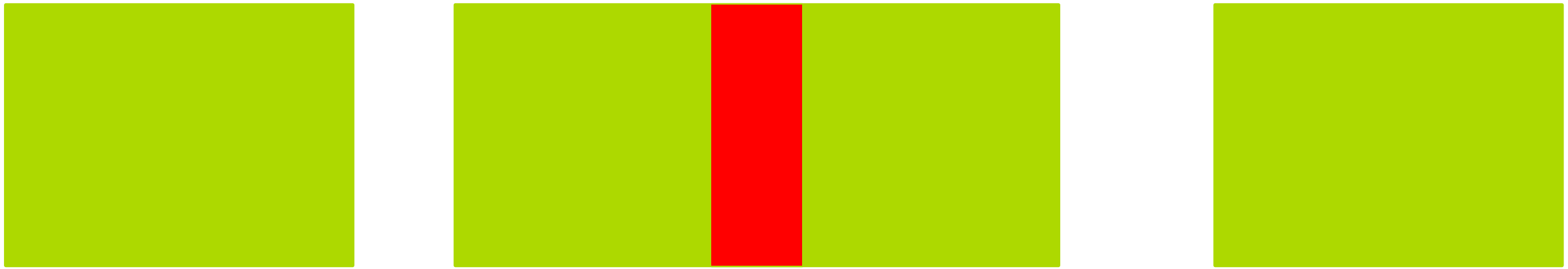


IDEAMAP SUDAN

Samples of Topology Rules

- **Polygon Geometry**

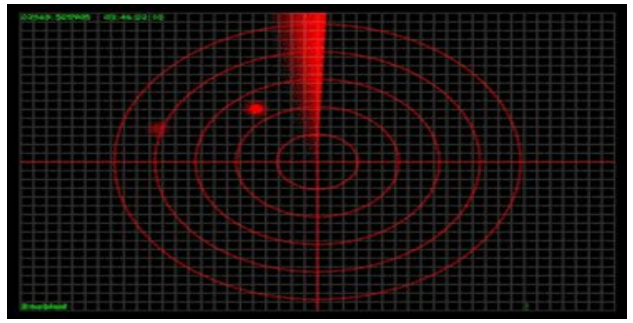
Polygons Must Not Overlap



IDEAMAP SUDAN

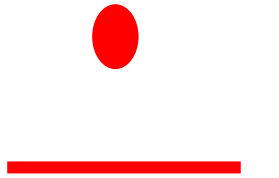
Validation

- It is the process of scanning data for rules violation Scanning the whole data or the visible data



Topology layers:

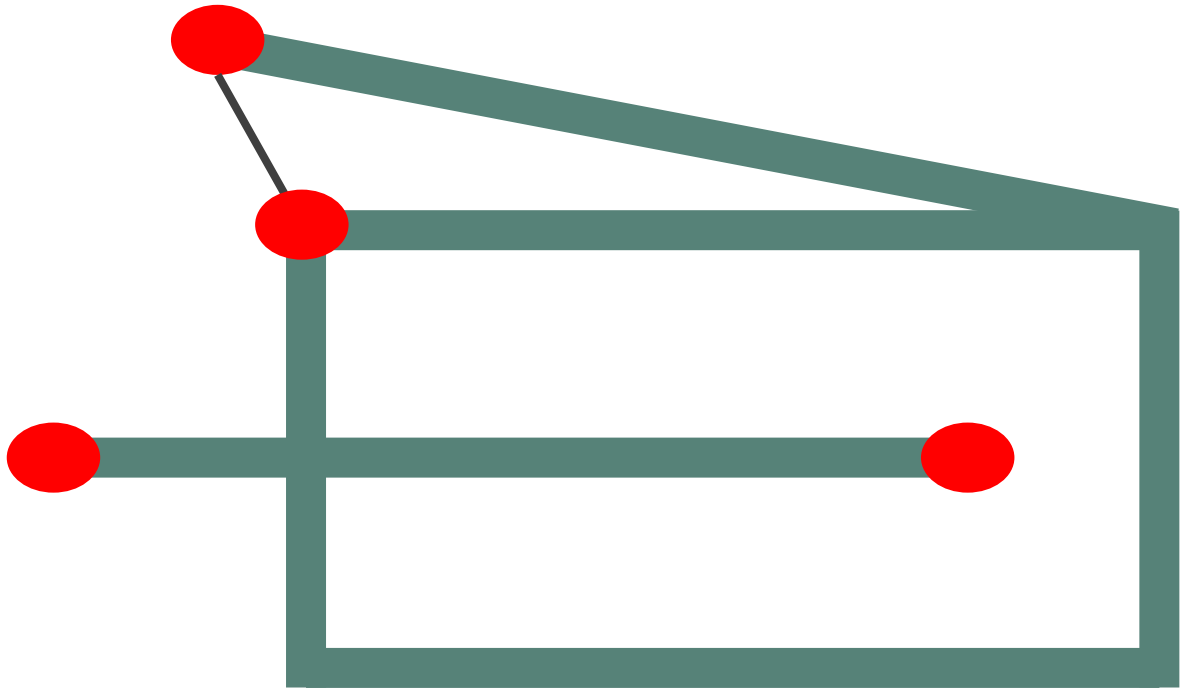
- Point errors
- Line errors
- Area errors



IDEAMAP SUDAN

Topology Solutions for Line Errors

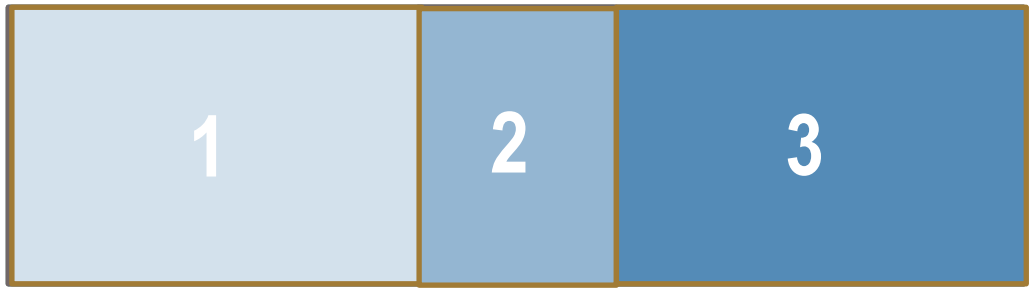
- **Case study:** Lines must not have dangles
  - Trim
  - Extent
  - Snap



IDEAMAP SUDAN

Topology Solutions for Polygon Errors

- **Case study: Polygons must not overlap**
  - Subtract
  - Merge
  - Create feature



IDEAMAP SUDAN

Practical overview:

- Topology is built on data base level
- Target layers should be determined along with their ranks
- Define Topology rules
- Validate data
- Fix errors

The most important thing when dealing with errors is the consistency

Questions?

Selftest