Conceptual Frameworks of Geospatial Analysis

Sourav Bhadra, Ph. D.

GIS4120/5120 Geospatial Analytics
8/31/2024
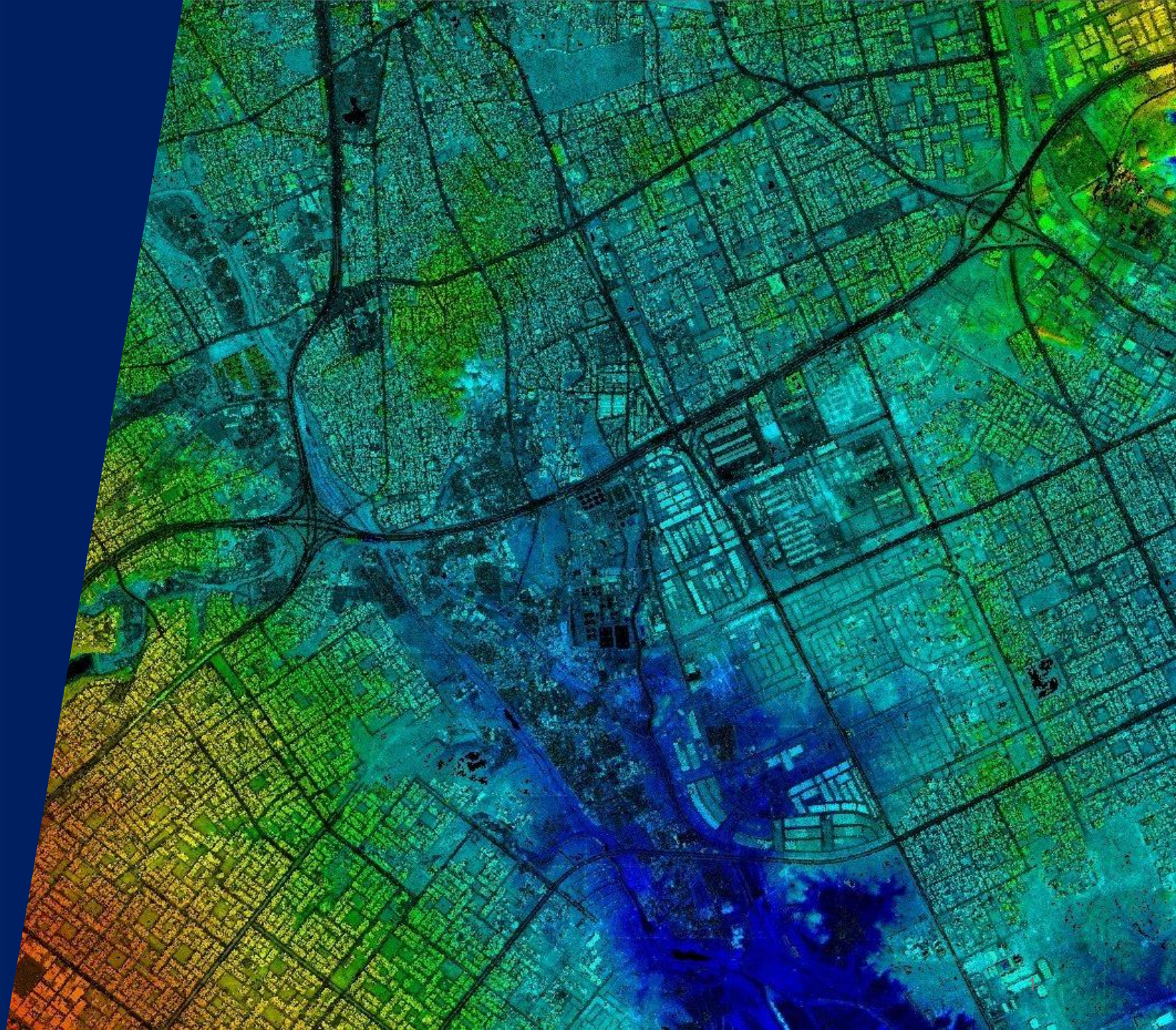
So, what is Geospatial Analytics?

Remote Sensing

GIS

Statistics
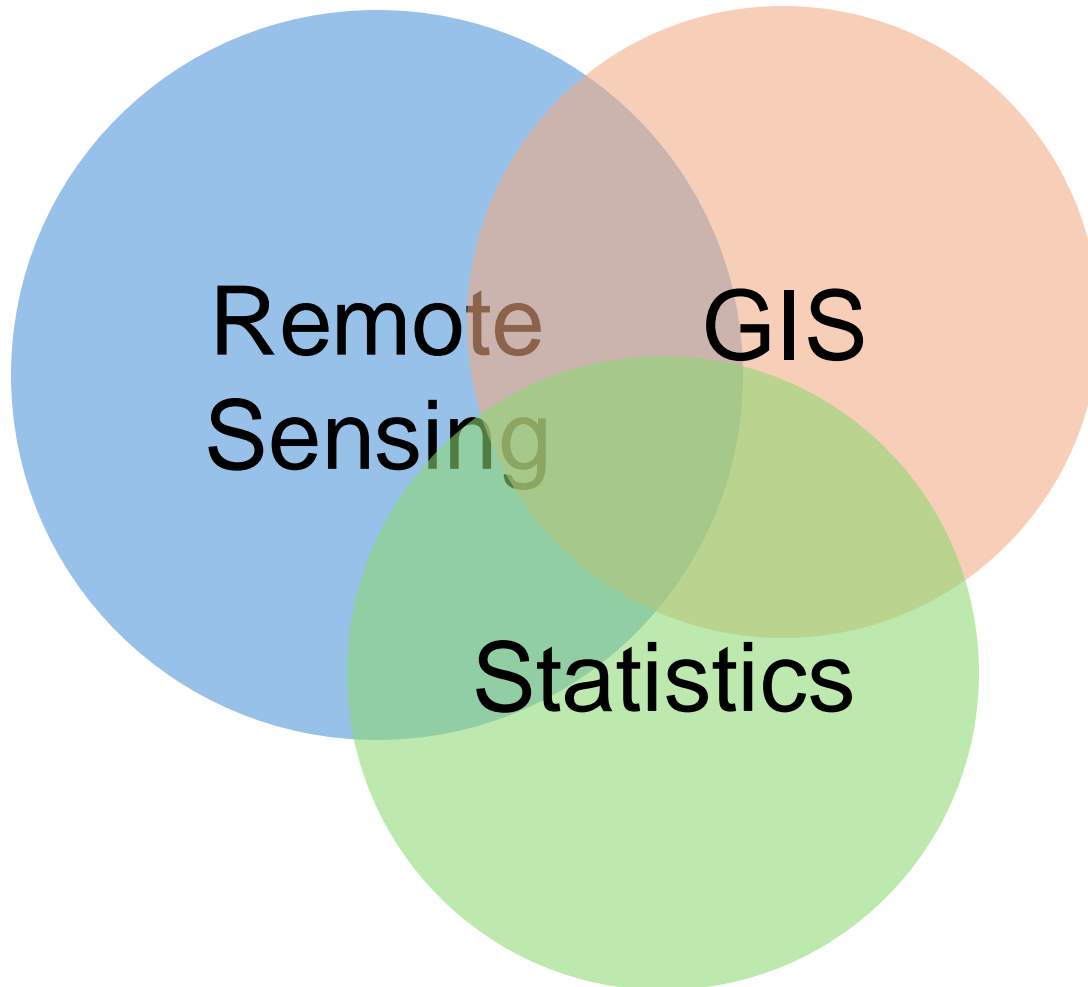
- Spatial Modeling
- Geocomputation
- Data Visualization
- Big Data
- Geographic Patterns
- Spatial Relationship
- Environmental Modeling
- Machine Learning

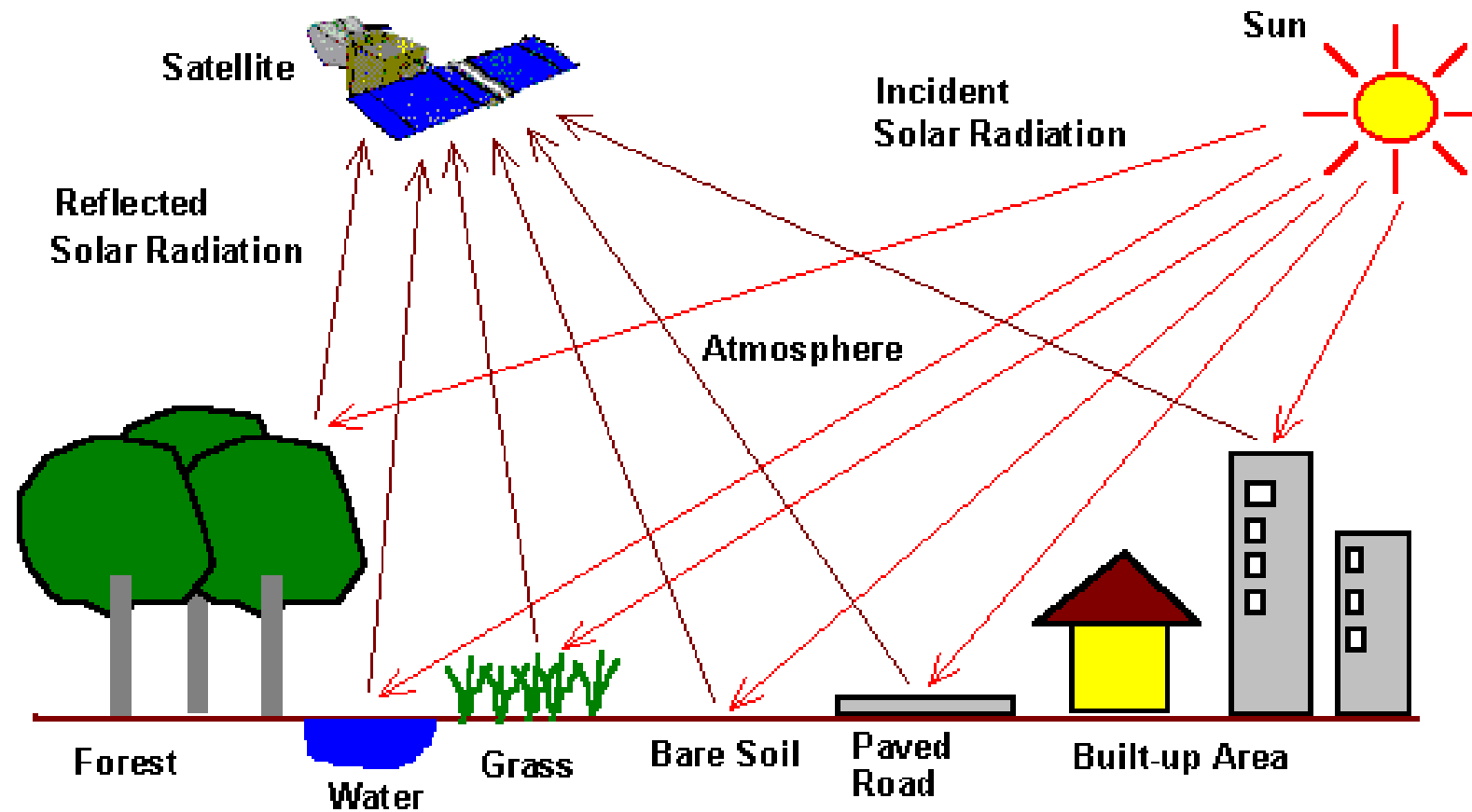# GIS is integral part of Geospatial Analytics

Data

Methods

Software
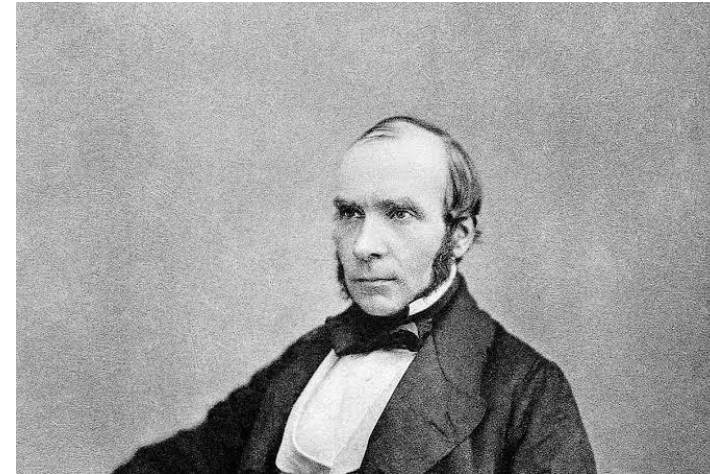
People

Hardware

Remote Sensing is another component

Spatial Statistics is the backbone of the analysis


Dr. John Snow (1813 – 1858)

John Snow's cholera map was the first spatial analysis to reveal the correlation between cholera outbreaks and contaminated water sources by mapping the locations of deaths relative to water pumps in London.

Source

What makes a "data" geospatial?

Abstraction of real-world relationship has been the backbone of mathematical representation
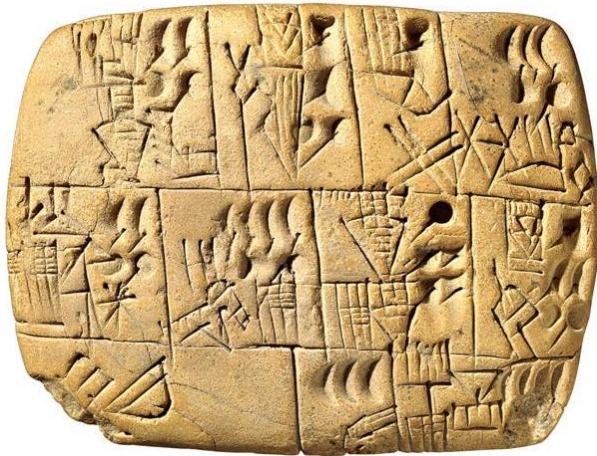
Source

|	
Count the number of Bisons	12, 14, 15
Differentiate between # Bison and weight of Meat	x = Bisons, y = Meat
Understand relationship between Bison and Meat	$y = mx + C$
More complex relationship at multiple dimensions	$ax^2 + bx + c = y$
Efficient solution to represent multiple equations	$\begin{bmatrix} 2 & -5 \\ 3 & 7 \end{bmatrix} = \begin{bmatrix} 3 \\ 7 \end{bmatrix}$

GIS recreates real world spatial data

Complex real data without any inherent meanings
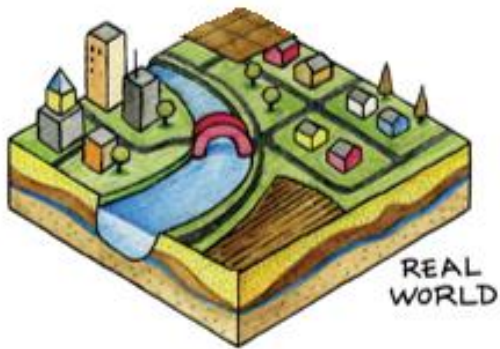
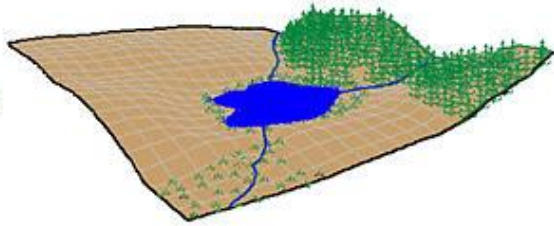Structured data representing different themes
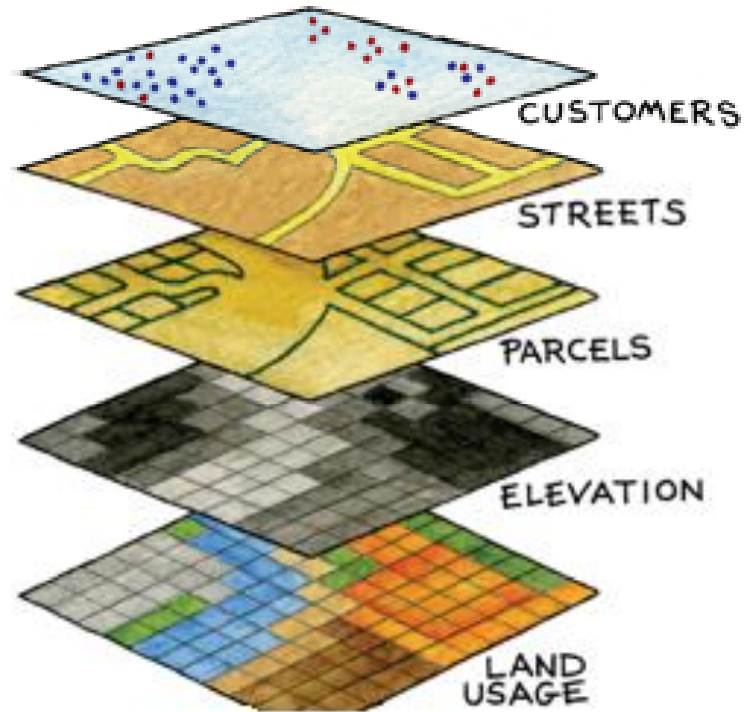
Specific information generated from analyzing the data

REAL WORLD

CUSTOMERS

STREETS

PARCELS

ELEVATION

LAND USAGE

Two types of geospatial data

The Real World

Vector Data

Raster Data

Images © University of Washington

Geospatial or coordinate data can be represented in two different data formats

Examples
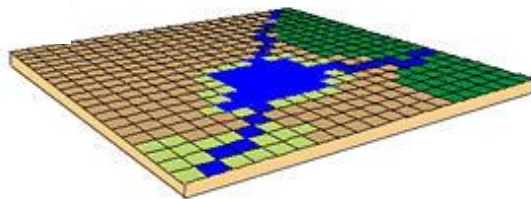
Vector:
e.g., points, lines, and polygons

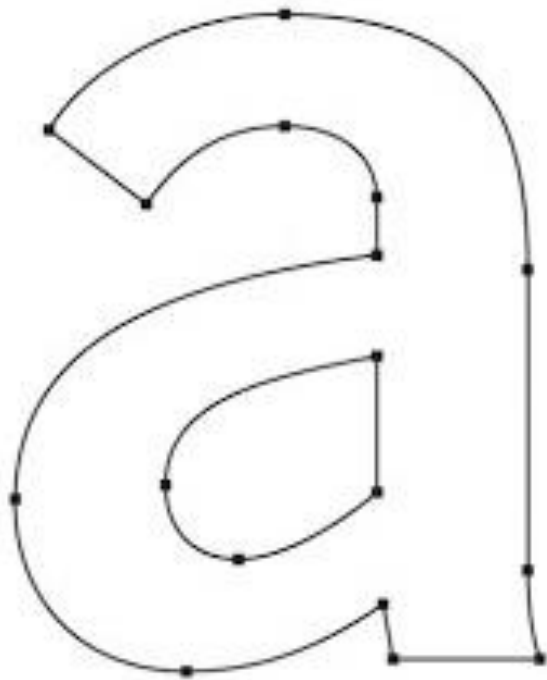School locations, road networks

Raster:
e.g., row and column matrix

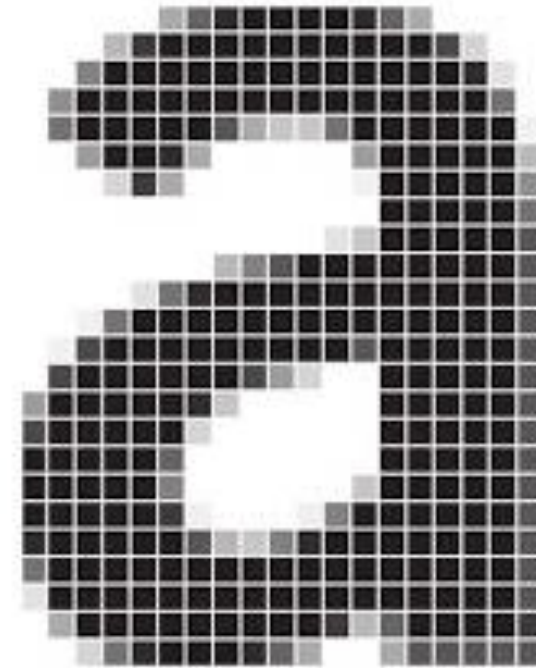Land use land cover, Digital Elevation Model

Vector vs Raster

Vector Model

Composed of coordinates

Raster Model

Composed of pixels

Image courtesy of Zina Yonten

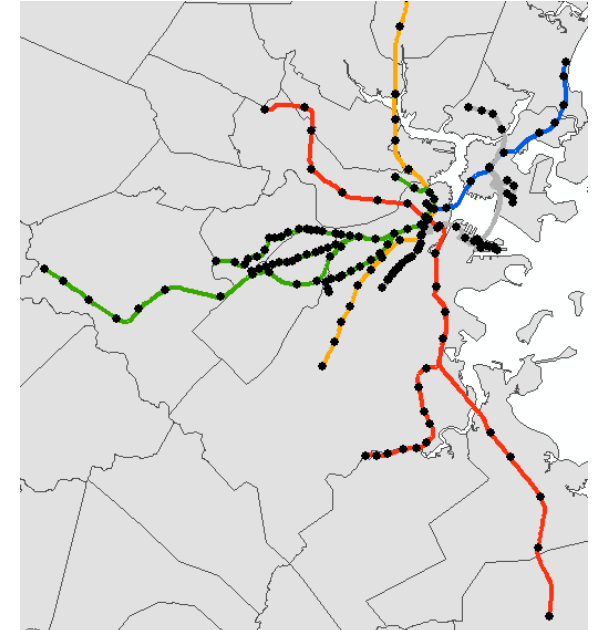# Points, lines and polygons as vector
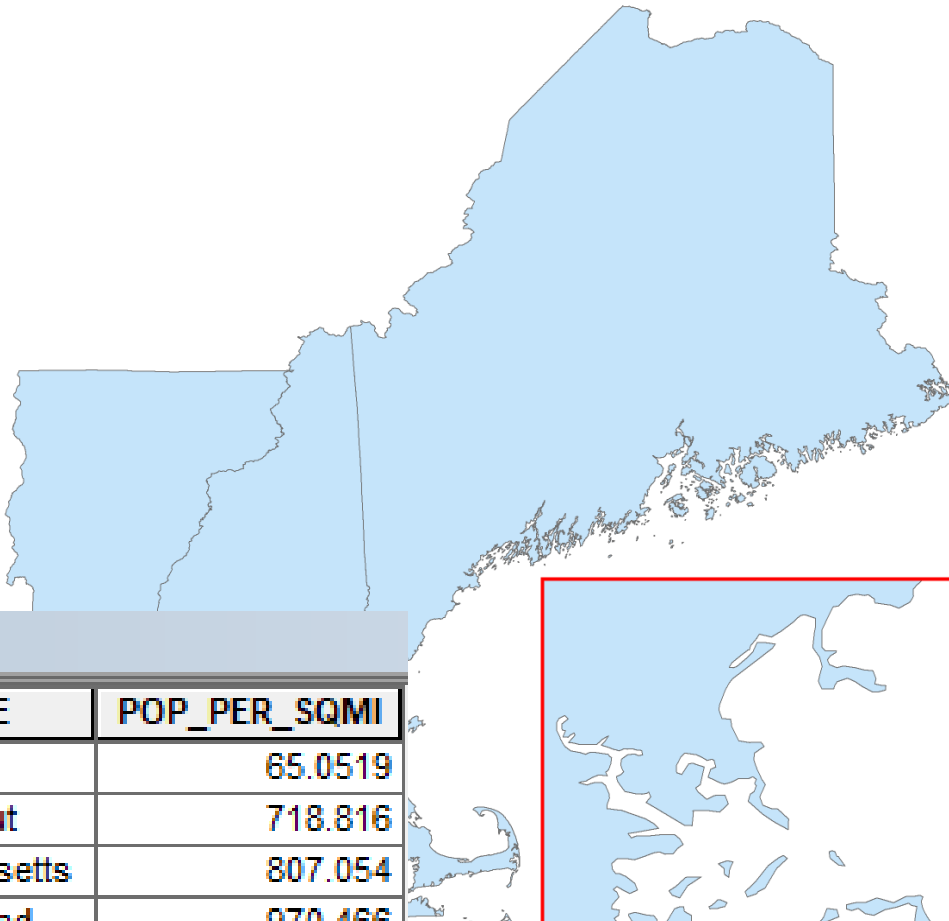
Points

Lines

Polygons

Combined

Vector data is represented in two fronts

Vectors have a **frontend geometry**
In this example the geometry   represents state polygons

Vectors have a **backend database**, normally called an 'attribute table'
- **rows** represent unique geometries (e.g. state polygons)
- **columns** represent a number of variables (theoretically infinite)
- Here each row (state) is symbolized by 'NAME' (categorical variable)

NewEngland

FID	Shape *	NAME	POP_PER_SQMI
0	Polygon	Vermont	65.0519
1	Polygon	Connecticut	718.816
2	Polygon	Massachusetts	807.054
3	Polygon	Rhode Island	970.466
4	Polygon	Maine	40.891
5	Polygon	New Hampshire	142.186

Vector data can be symbolized by different data

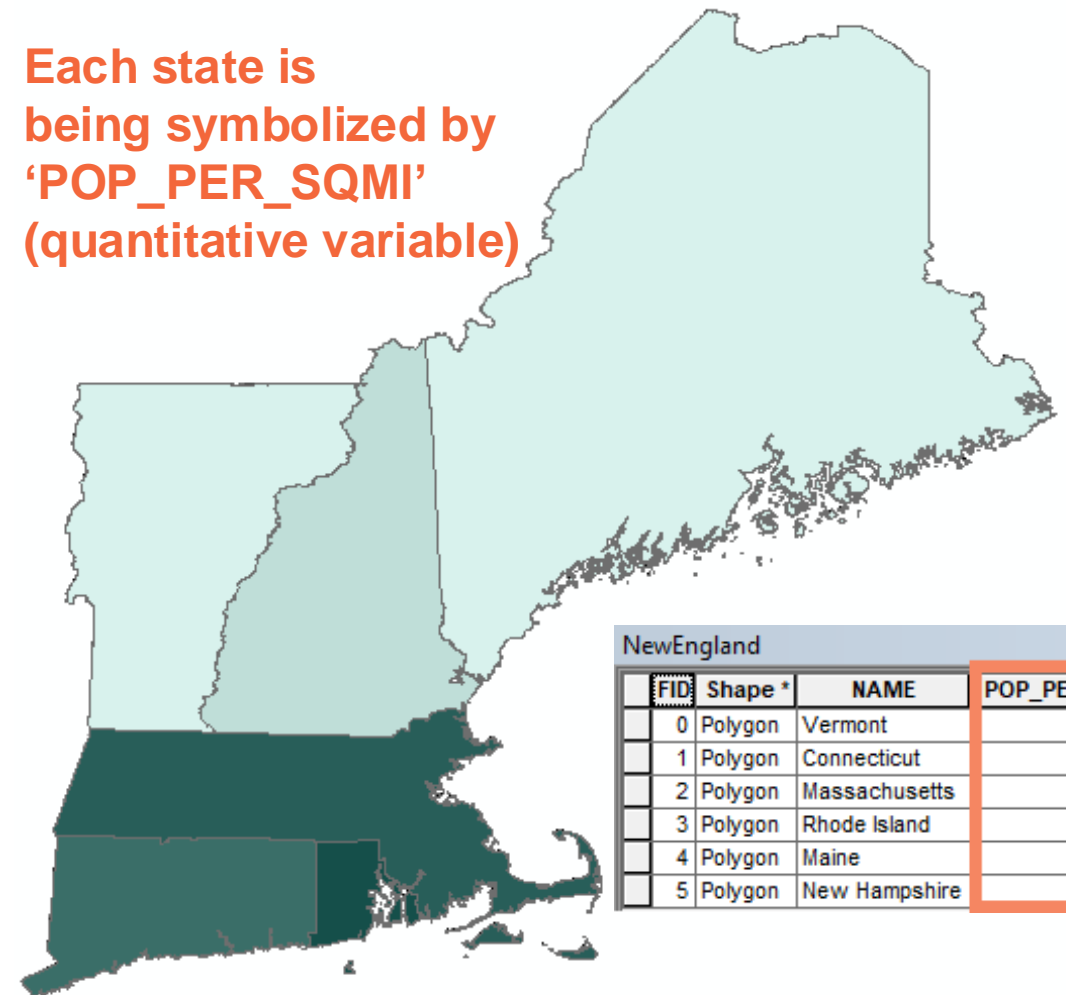**Each state is being symbolized by 'NAME' (qualitative variable)**
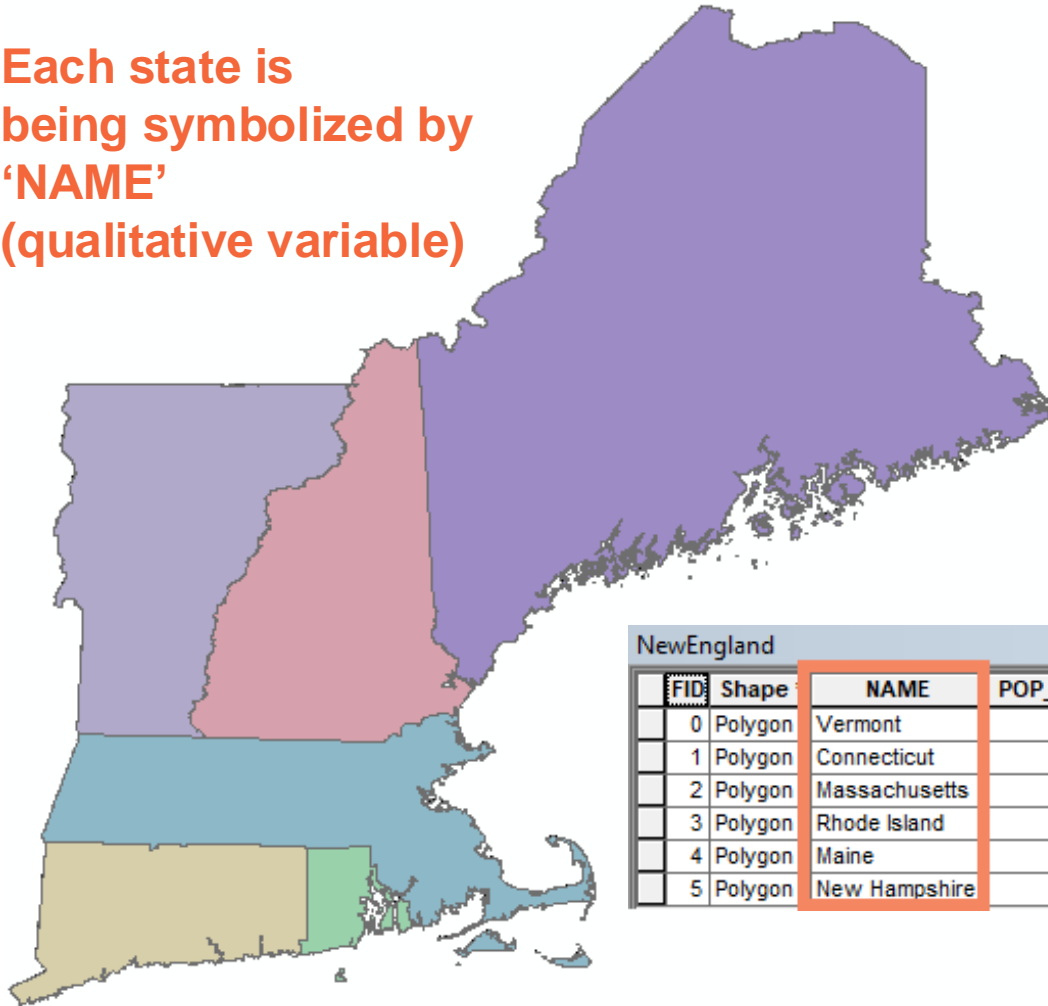
Each state is being symbolized by 'POP_PER_SQMI' (quantitative variable)

NewEngland

FID	Shape	NAME	POP_PER_SQMI
0	Polygon	Vermont	65.0519
1	Polygon	Connecticut	718.816
2	Polygon	Massachusetts	807.054
3	Polygon	Rhode Island	970.466
4	Polygon	Maine	40.891
5	Polygon	New Hampshire	142.186

NewEngland

FID	Shape *	NAME	POP_PER_SQMI
0	Polygon	Vermont	65.0519
1	Polygon	Connecticut	718.816
2	Polygon	Massachusetts	807.054
3	Polygon	Rhode Island	970.466
4	Polygon	Maine	40.891
5	Polygon	New Hampshire	142.186

Different levels of measurement for attribute data

Categorical

Numerical

Nominal

Ordinal

Interval

Ratio

- Gender, ethnicity, eye color, blood type
- Land use, land cover

- Income level
- Customer ratings
- Political orientation

No true zero.
- Temperature in Fahrenheit or Celsius
- Test scores

Has true zero.
- Age
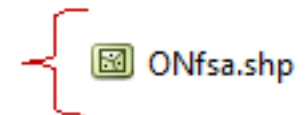- Height
- Weight
- Temperature in Kelvin

The most common vector data format is '.shp'

- "A" shapefile is a collection of several different files with different extensions

- Each file serves different purposes which in combine represent a vector data model

- The three required files are:
  - **SHP** is the feature geometry.
  - **SHX** is the shape index position.
  - **DBF** is the attribute data.

- **PRJ** is the projection system metadata.
  - **XML** is the associated metadata.
  - **SBN** is the spatial index for optimizing queries.
  - **SBX** optimizes loading times.

How a shapefile looks in Windows Explorer:

ONfsa.dbf
ONfsa.htm
ONfsa.prj
ONfsa.sbn
ONfsa.sbx
ONfsa.shp
ONfsa.shp.xml
ONfsa.shx
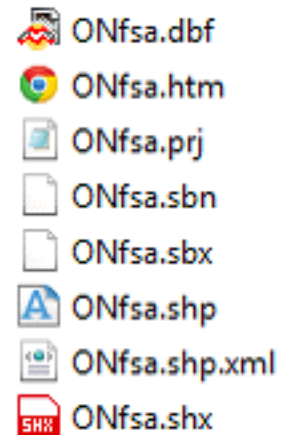
How a shapefile looks in ArcMap:

ONfsa.shp

There are other popular vector data formats

- .GEOJSON: GeoJSON is a format for encoding geographic data structures, such as points, lines, and polygons, using JavaScript Object Notation (JSON)

- .GML: GML allows for the use of geographic coordinates extension of XML.

- .KML: KML stands for Keyhole Markup Language. This GIS format is XML-based and is primarily used for Google Earth.

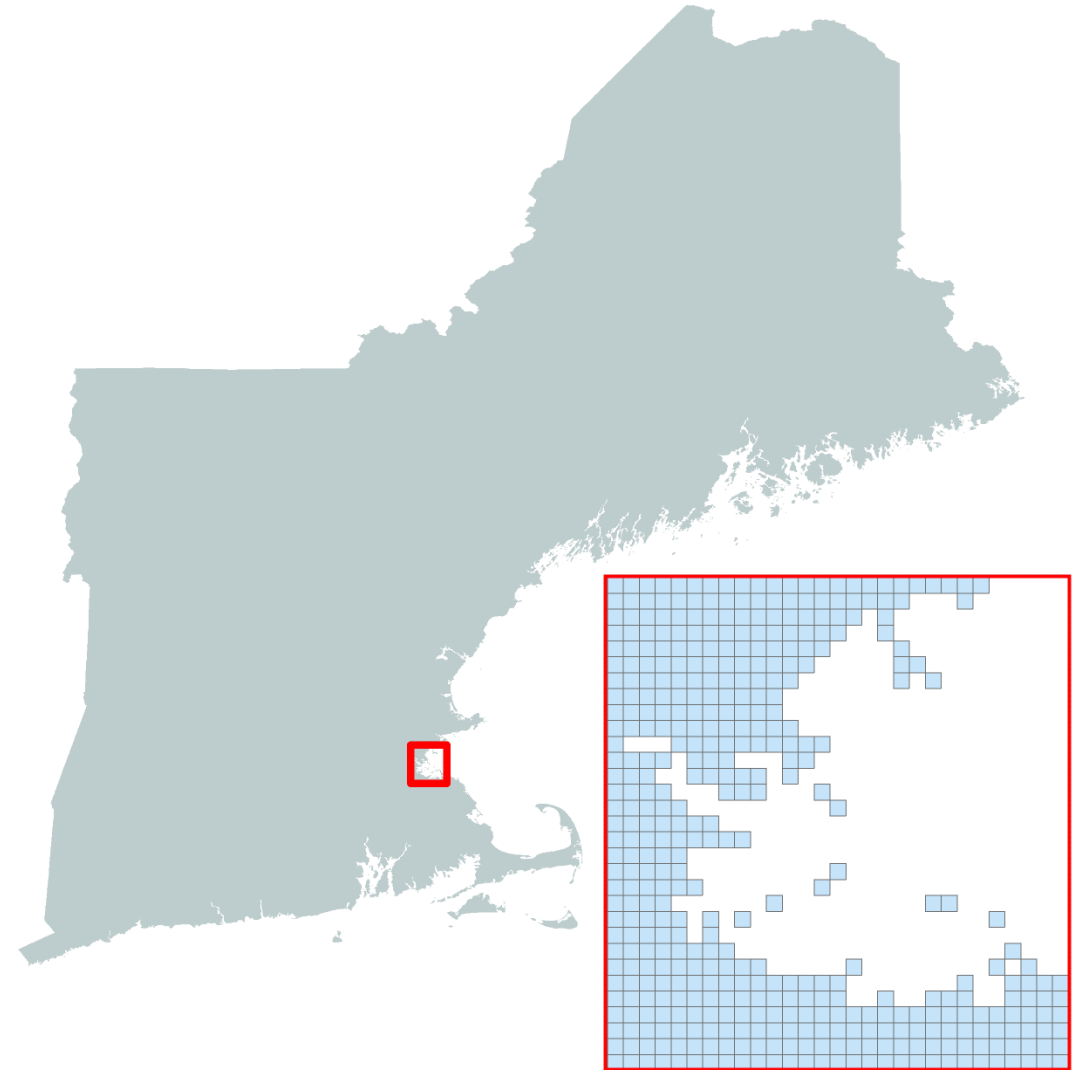- Learn more at: https://gisgeography.com/gis-formats/

Raster data model

Raster data includes aerial photographs, digital elevation models, and scanned maps.
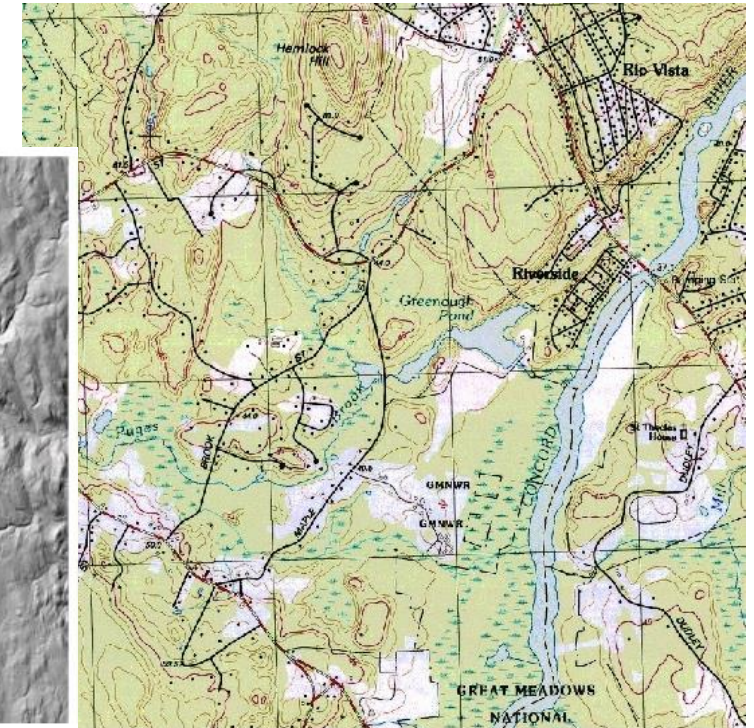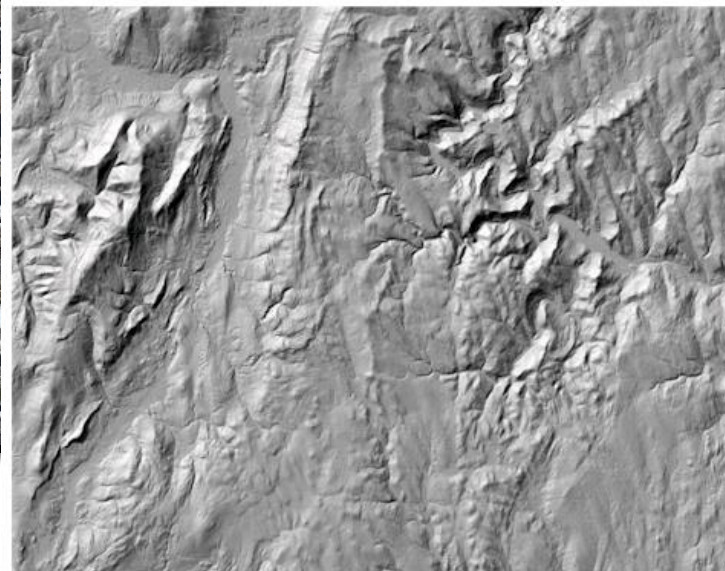
Raster data has a frontend cell matrix

- Where each cell has its own value

- A raster can only symbolize one variable at a time

- Multiple bands can be stacked together to represent multiple variables

Raster data has a frontend cell matrix

- Where each cell has its own value

- A raster can only symbolize one variable at a time

- Multiple bands can be stacked together to represent multiple variables

- **Here each cell/pixel is being symbolized by elevation value**
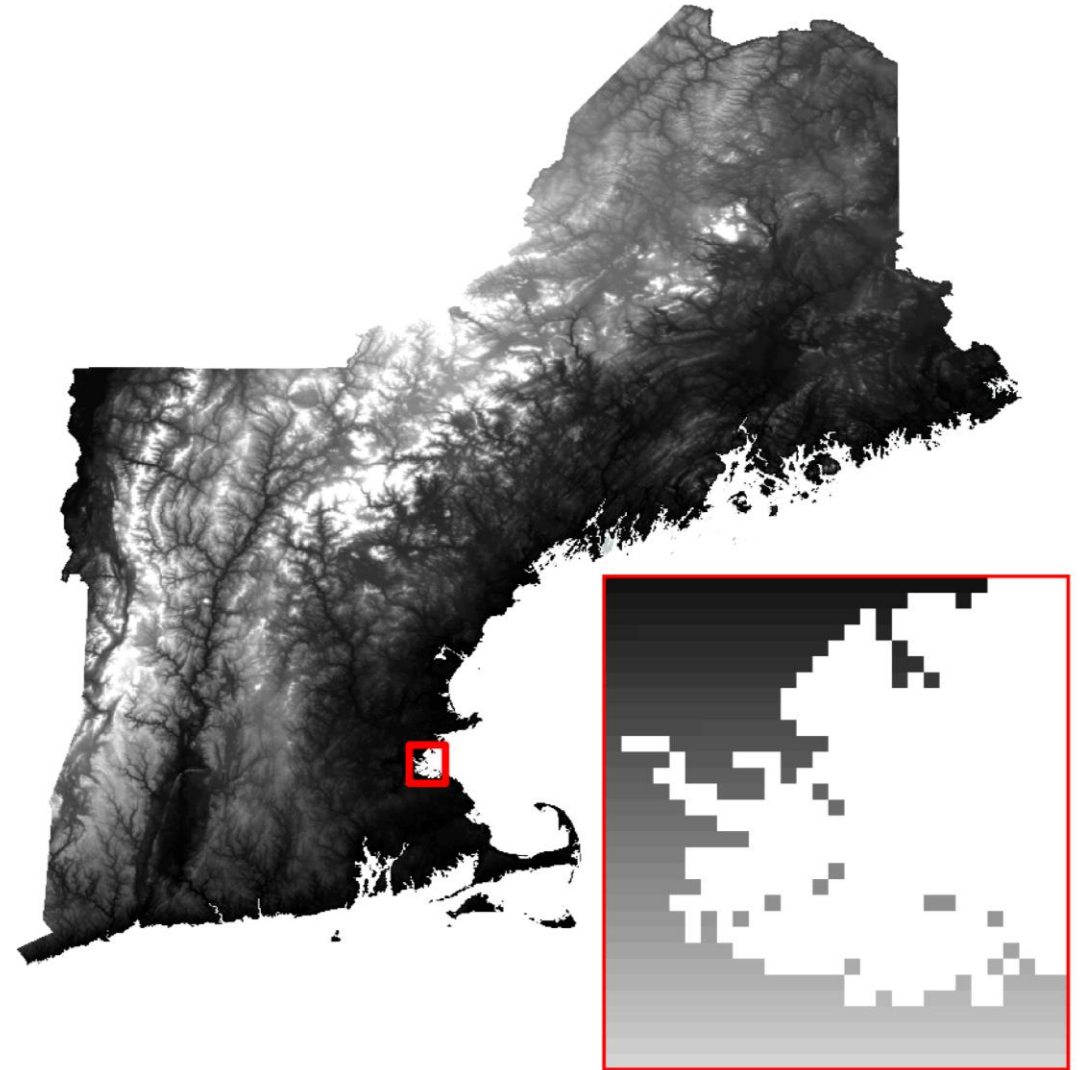
# Raster data can have a **backend database** too

- Raster data have a **backend database**, normally called an 'attribute table'

- **rows** represent unique values (1m, 2m, 3m, etc.)

- **columns** have specific variables
   1) unique 'ROW ID'
   2) unique 'VALUE'
   3) 'COUNT' of pixels with that 'VALUE'

Elevation

Rowid	VALUE	COUNT
0	1	1399
1	2	457
2	3	574
3	4	310
4	5	345
5	6	294
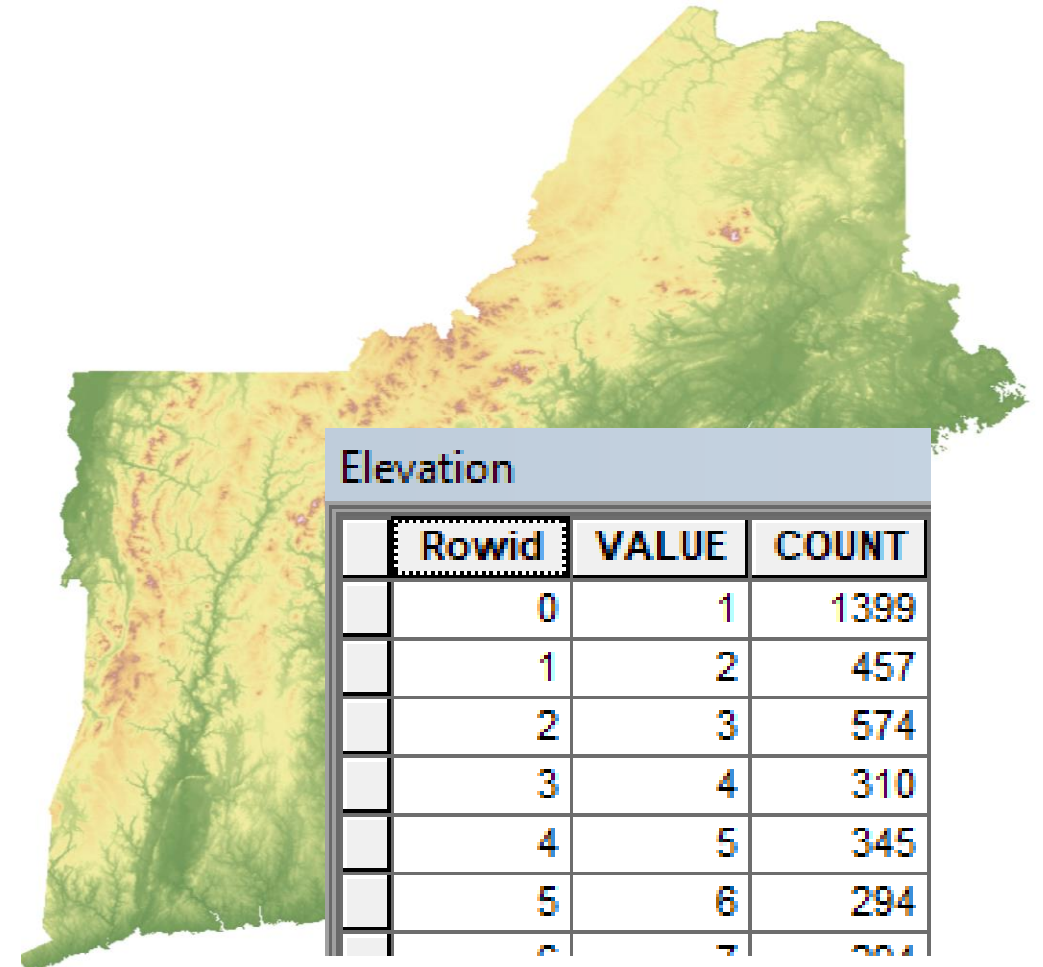
My go to choice as raster data format is GeoTIFF

- A .tif file stores metadata or attributes about the file as embedded tif tags.

- A GeoTIFF is a standard .tif image format with additional spatial (georeferencing) information embedded in the file as tags.

- Some example of tags:
  - Extent
  - Resolution
  - Coordinate Reference System (CRS) - we will introduce this concept in a later episode
  - Values that represent missing data (NoDataValue)
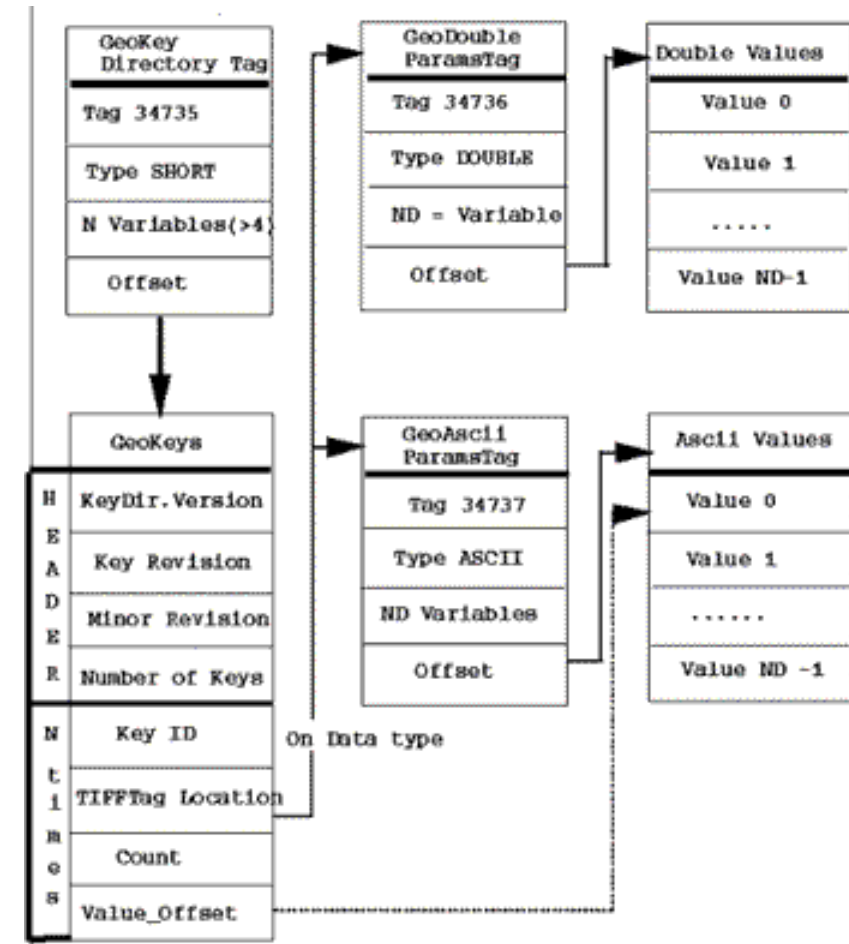
Source

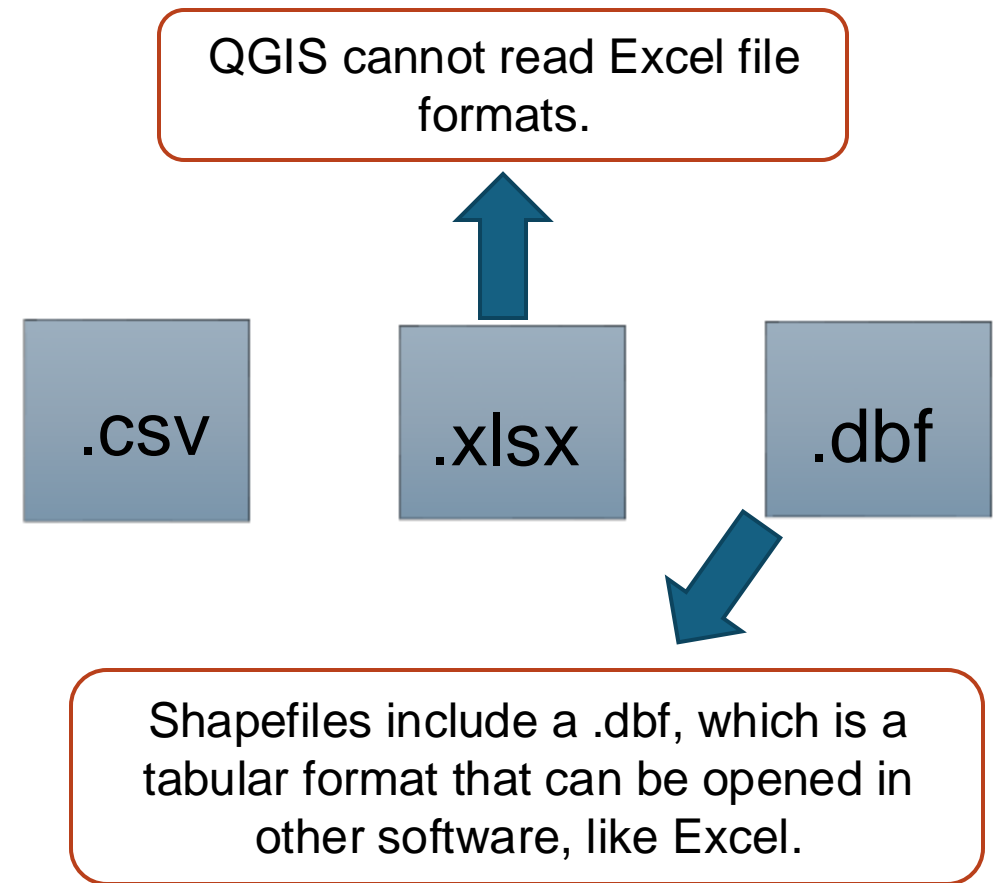# Tabular data is also a common data format needed in geospatial analytics

Tabular data can be transformed

into spatial data in two ways:

1. Joining
- **Use a shared unique identifier** (GEOID, name, etc.) to match up tabular data to the spatial data's attribute table.

2. Geocoding
- **Use lat/lon** coordinates in table to plot as points on map
- **Use addresses** to plot locations based on a street network

QGIS cannot read Excel file formats.

.csv   .xlsx   .dbf

Shapefiles include a .dbf, which is a tabular format that can be opened in other software, like Excel.

Geodatabase is a more complex and efficient data structure to represent multiple geospatial data

- ESRI/ArcGIS storage system
- A collection of geographic datasets of various types held in a common file system folder
- **Advantages:** larger files size limits, faster processing time when using analysis tools
- **Disadvantages:** can only be opened in ESRI software
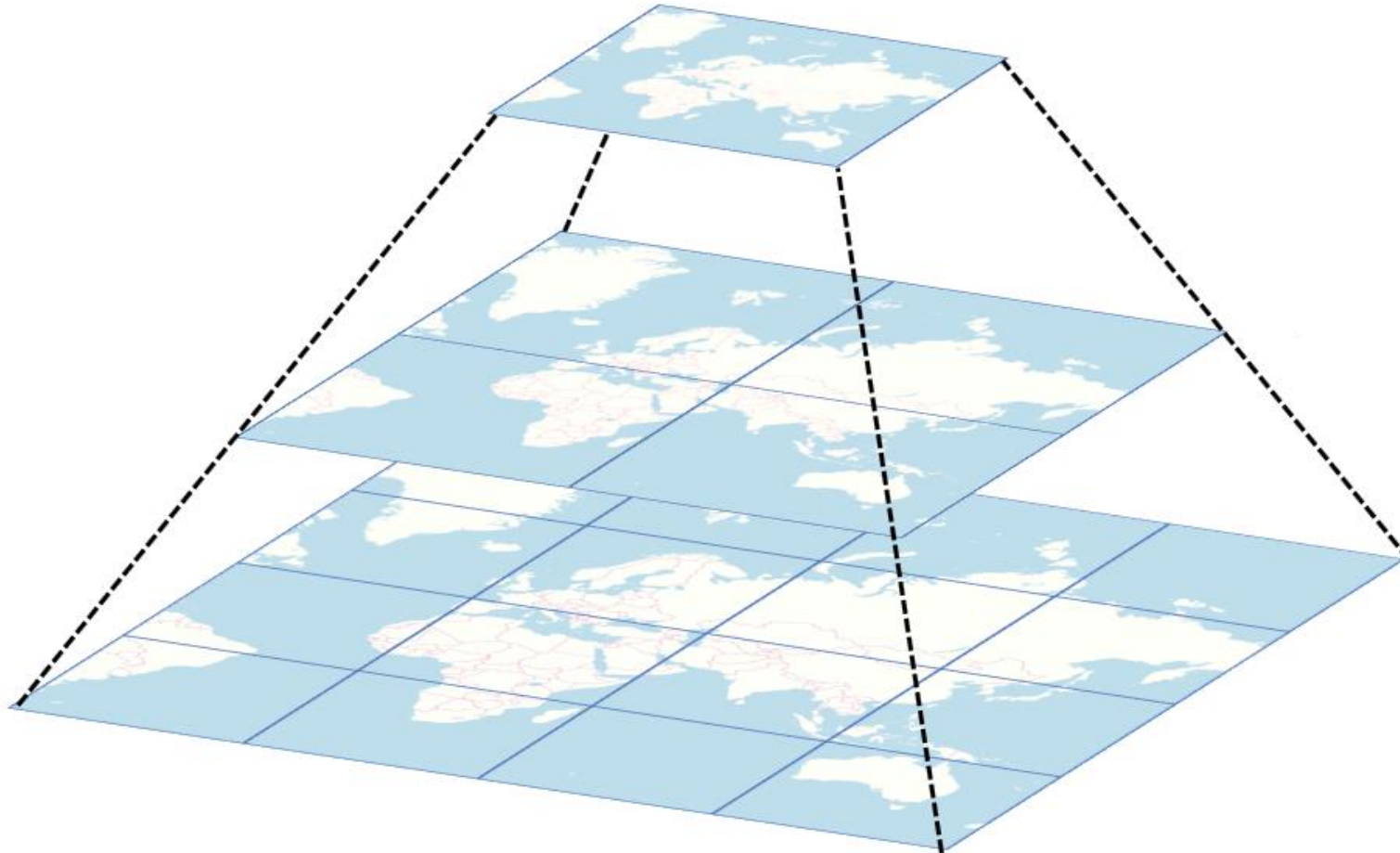- Learn more about using geodatabases in Pro.
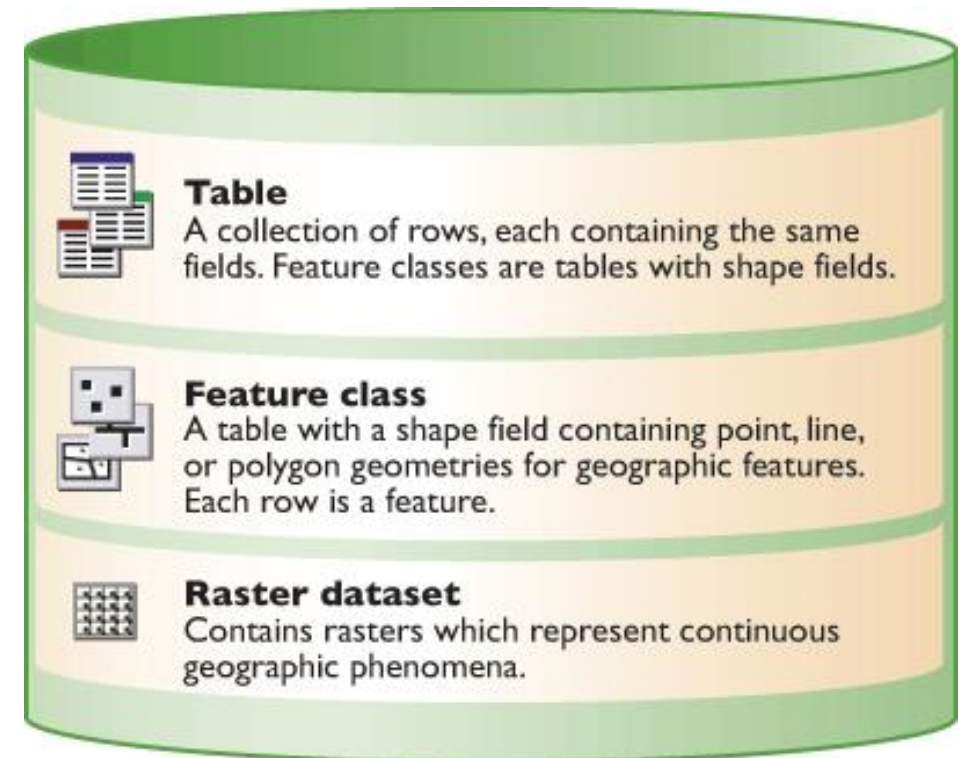
Source

Raster and vector tiles are now the face of online mapping

Vector tiles are replacing the use of raster tiles

- Vector tiles are a way of storing and serving geographic data in a compact and efficient format.

- Unlike traditional raster tiles, which are pre-rendered images, vector tiles contain vector data that can be rendered dynamically on the client side.

Why should we care about projections or CRS?

If a coordinate system is wrong or missing, data will not display in the correct location
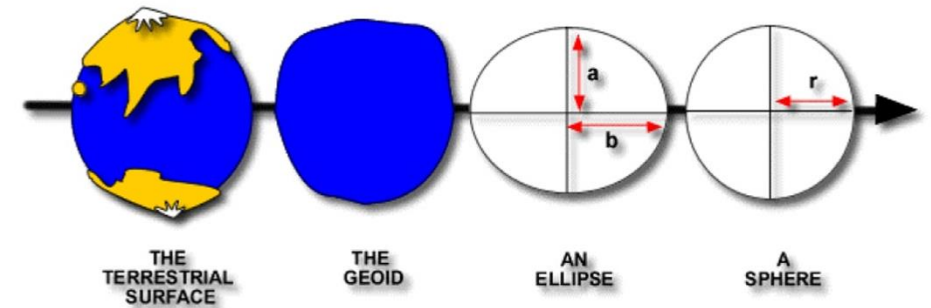

Source: Dan Mahr. https://ihatecoordinatesystems.com/

Earth's geoid is a complex shape to map

The geoid is an irregular-shaped model of the Earth's gravity field

Source
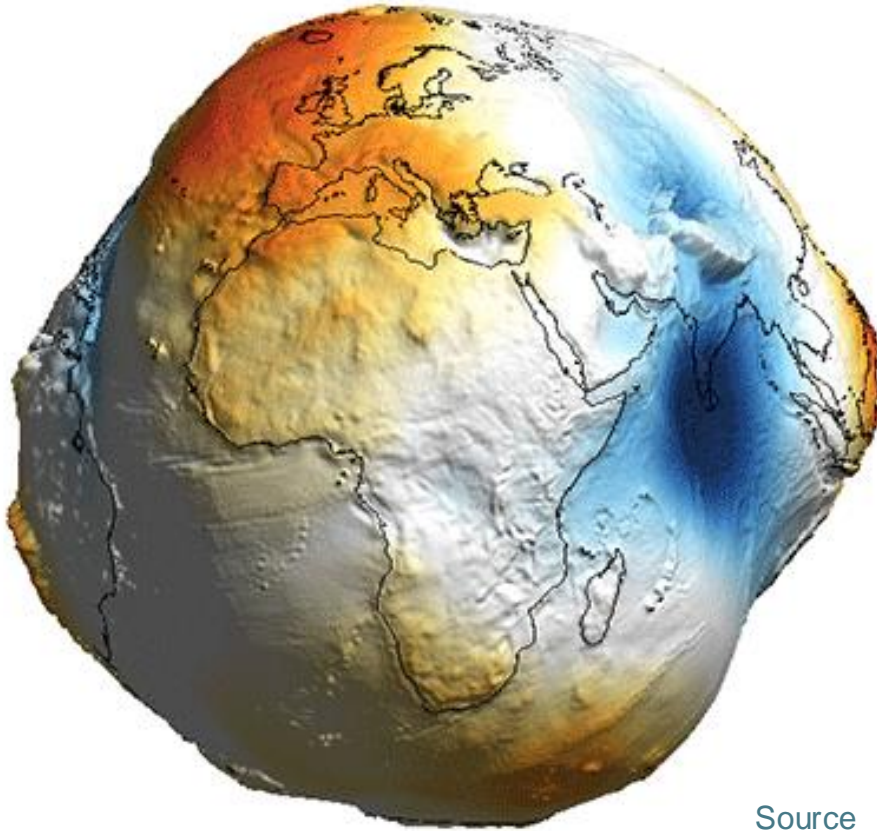
THE TERRESTRIAL SURFACE   THE GEOID   AN ELLIPSE   A SPHERE

- Sphere can be used as a first approximation to the geoid.
- Fits well locally
- Globally it is a poor approximation.
- Difference between equatorial and polar radius: 21km
- Rotational or biaxial ellipsoid is a good approximation both locally and globally.

Datum is an idealized mathematical representation of the Earth's 3D model
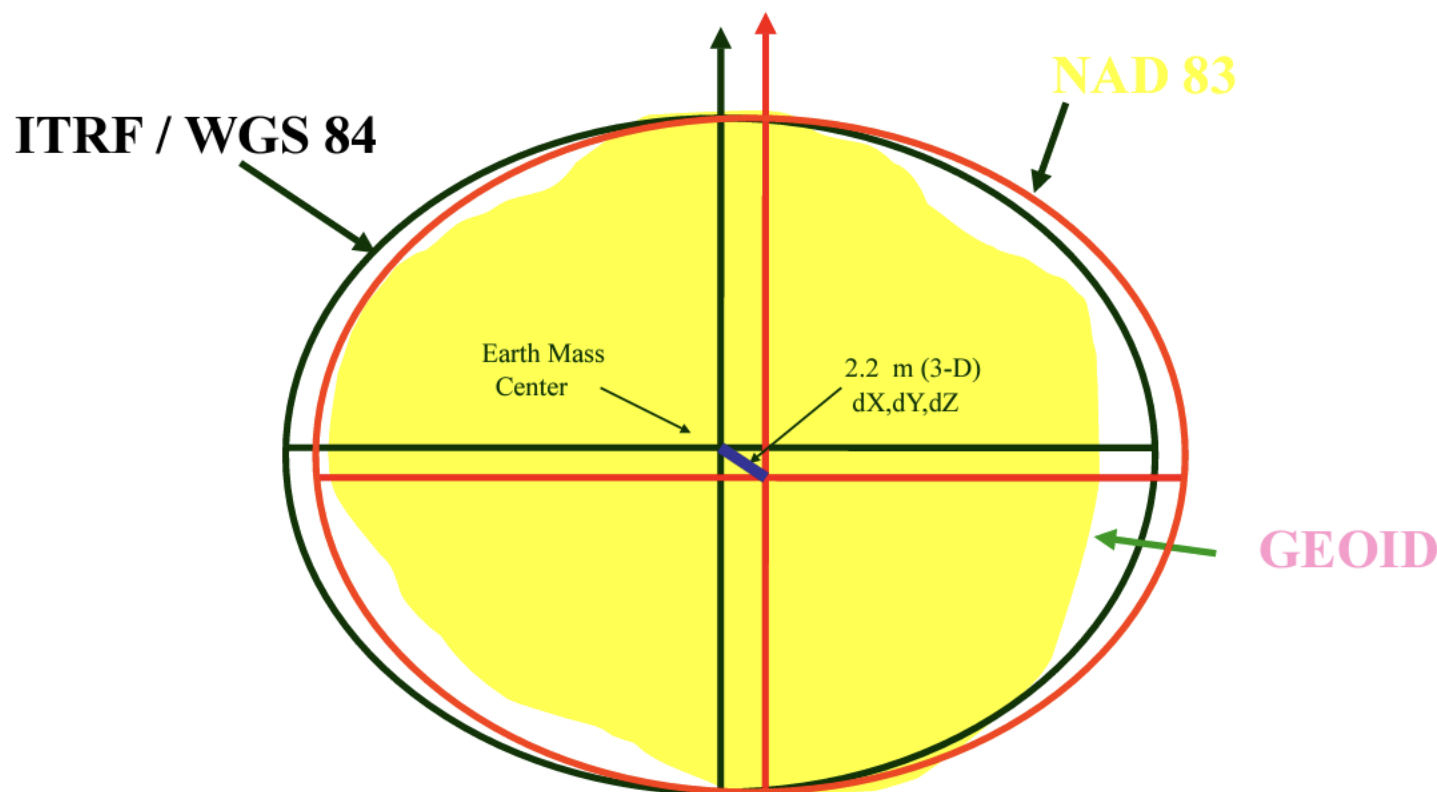
- Geographic Coordinate System (GCS) is the 3D representation of the Earth's coordinates.
- WGS84 (EPSG:4326), NAD83 (EPSG:4269)

Projection algorithm is applied to GCS to create a Projected Coordinate System (PCS)

- Imagine an orange as the Earth, and you want to be able to peel it in such a way as to lay the peel flat.

- Similarly, **projection is a method by which cartographers translate a 3D globe (spheroid or ellipsoid) to a 2D map surface.**
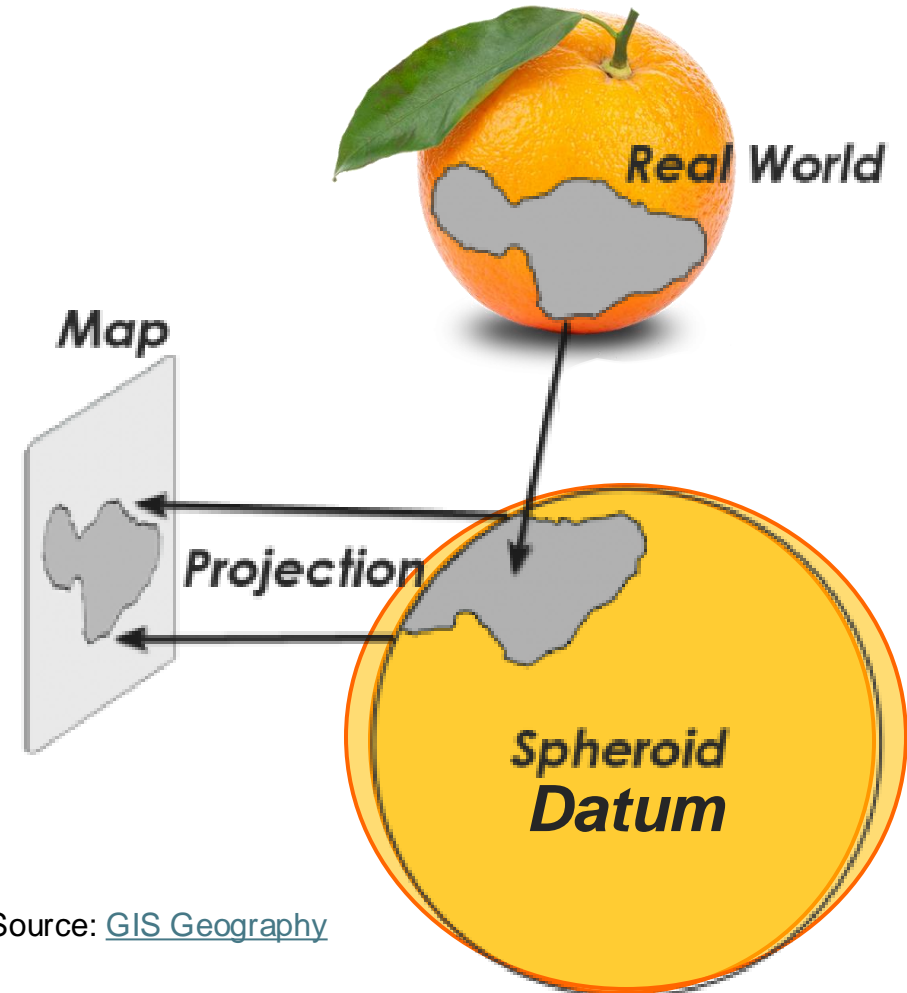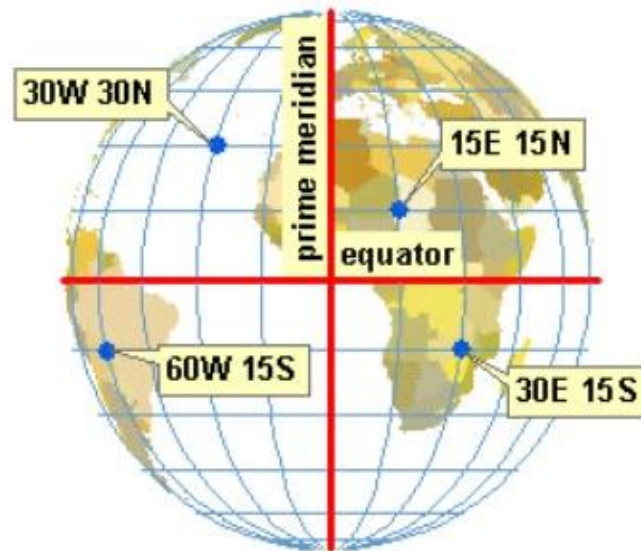

Source: GIS Geography

PCS is consisted of several items

- Geographic Coordinate System
- Projection Algorithm
- Linear Unit
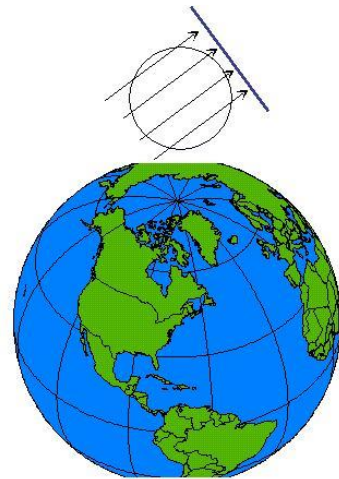- Parameters that center the system on a certain location

https://www.youtube.com/watch?v=kIID5FDi2JQ
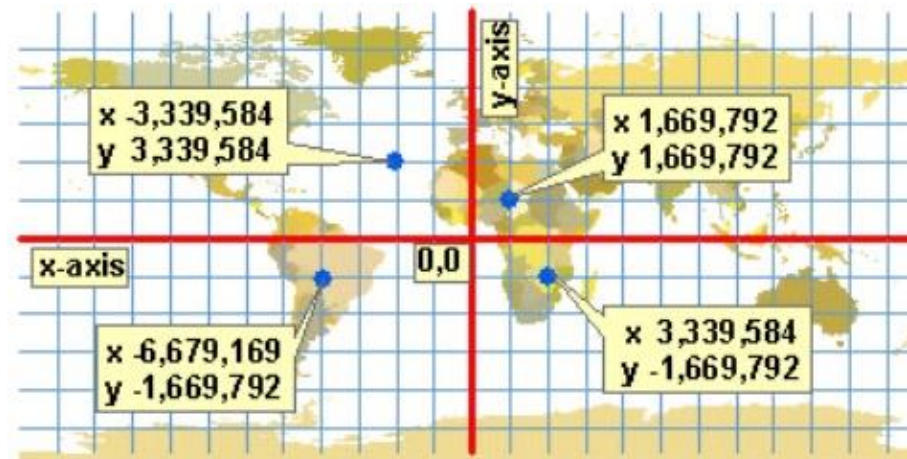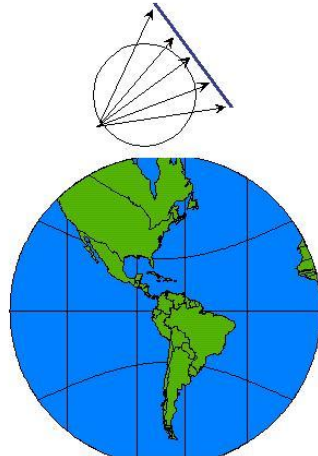

Source: Jochen Albrecht

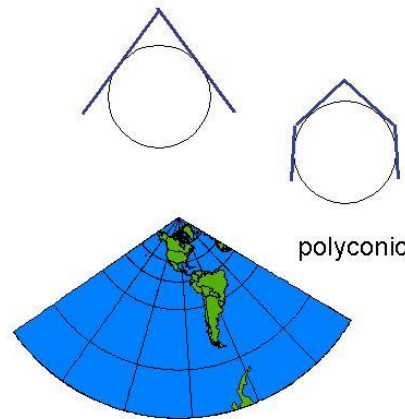# Lots of option for projection systems

- There are many different types of projections.
- Each have certain strengths and limitations in the following types of distortions: shape, area, distance, direction
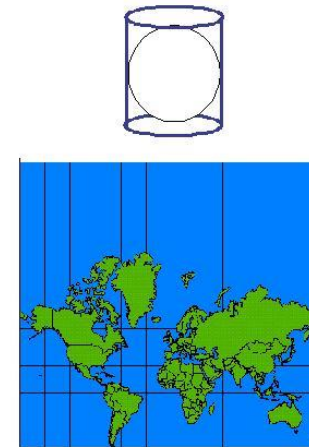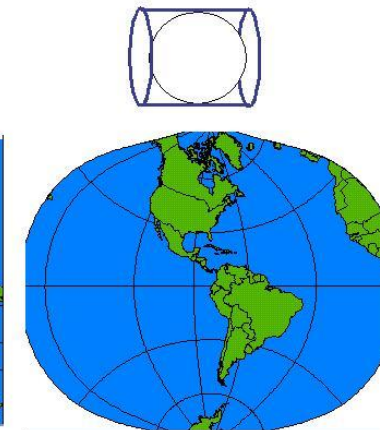

View from space
Orthographic/ Azimuthal

gnomonic
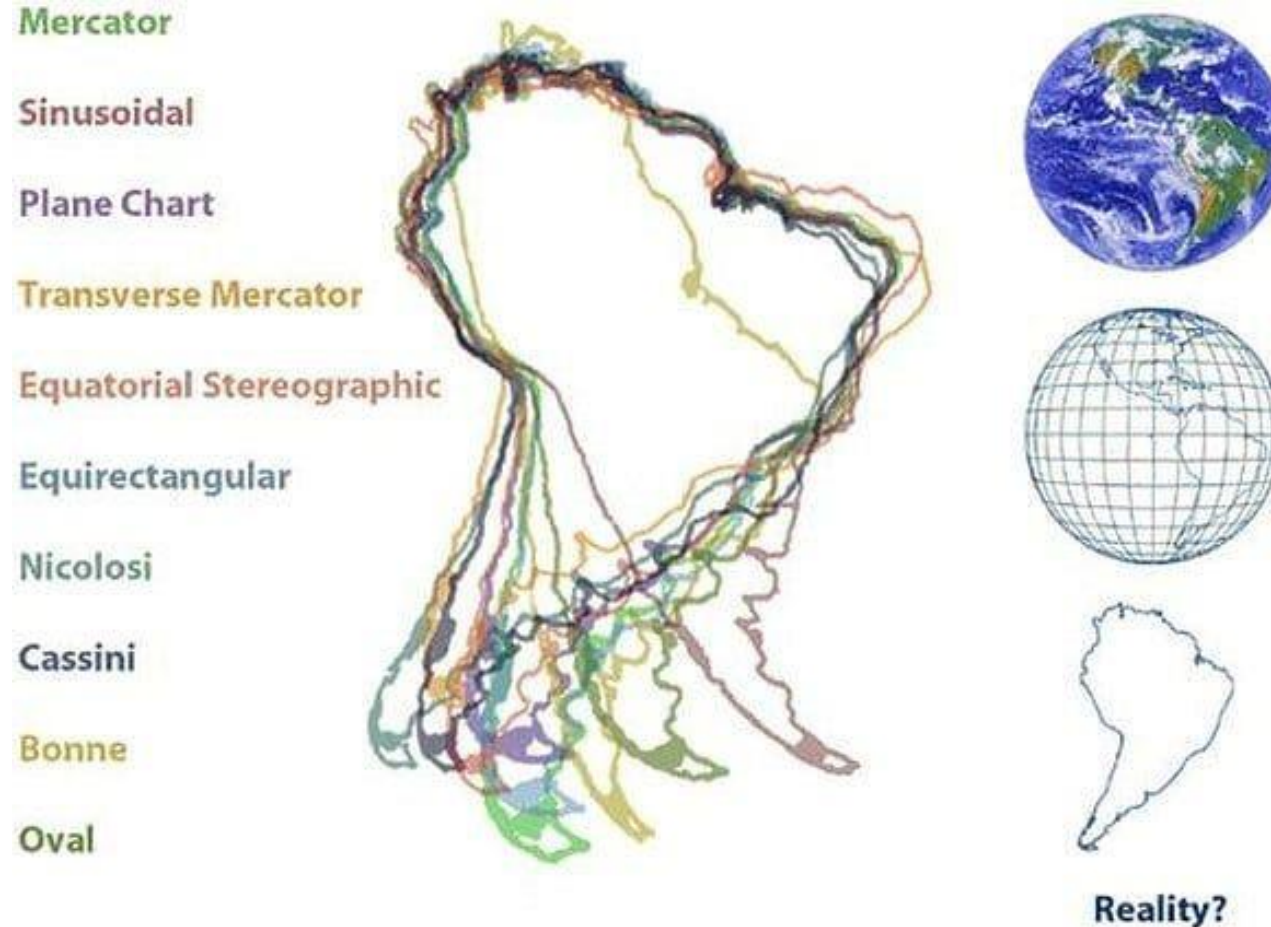
polyconic

conic
(Lambert conformal)

Mercator
cylindrical

Universal Transverse Mercator
(UTM, 1927, grid zone 17, 81W)

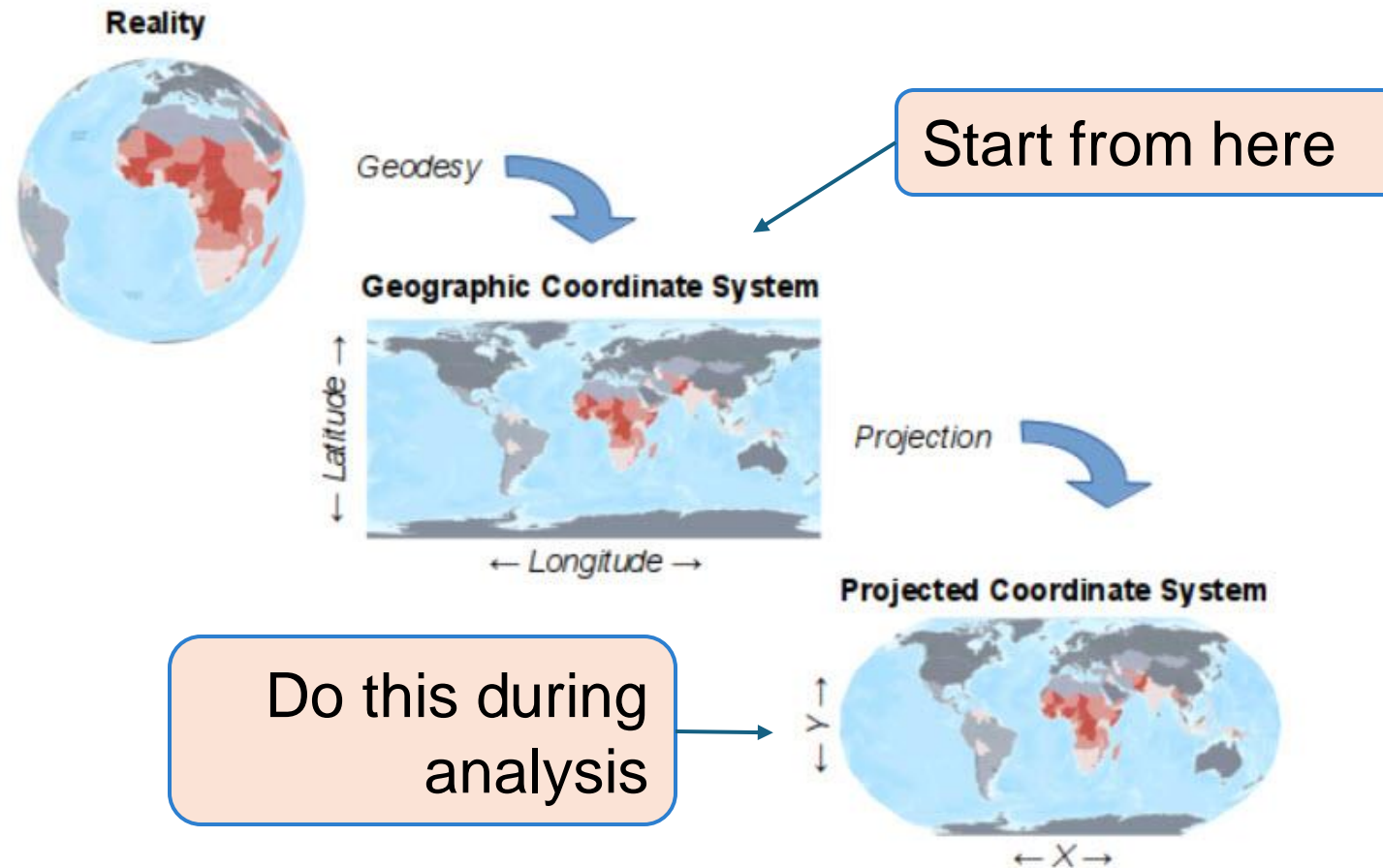# Analysis tools with shape, area or distance calculations require data to be in a projection
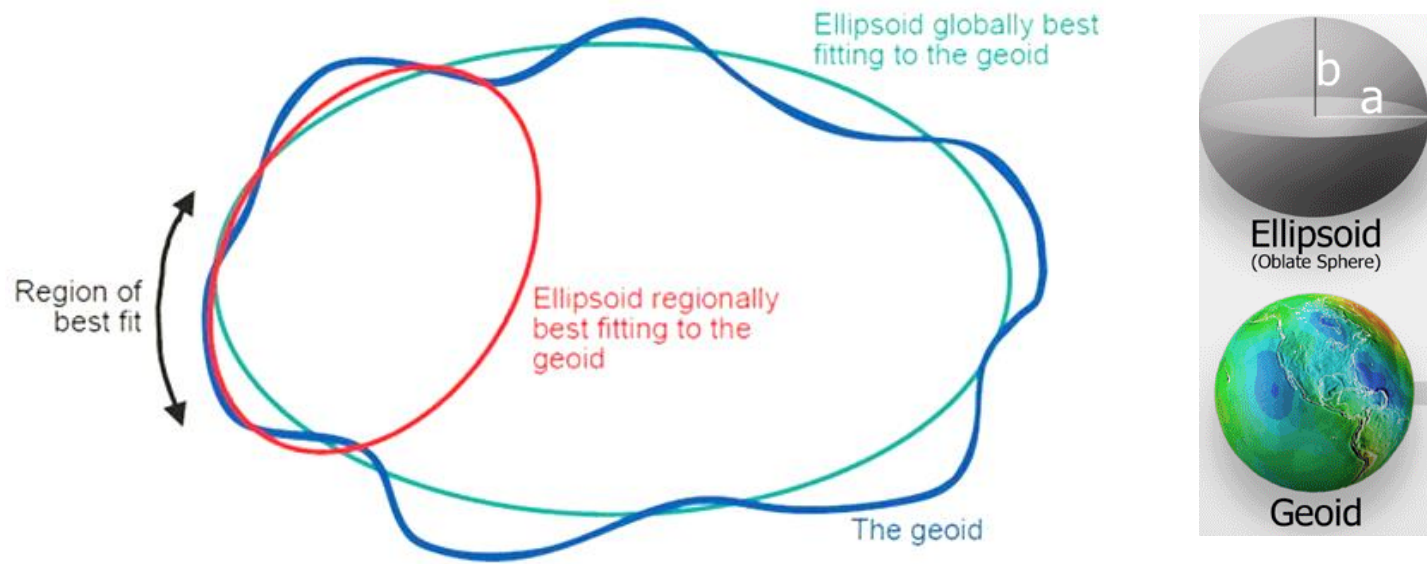

Mercator
Sinusoidal
Plane Chart
Transverse Mercator
Equatorial Stereographic
Equirectangular
Nicolosi
Cassini
Bonne
Oval

Reality?

So, what to do?

Start from here

Do this during analysis

Commonly encountered systems

- ## Geographic Coordinate System

  - NAD83 (North American Datum) – best fitting ellipsoid for North America

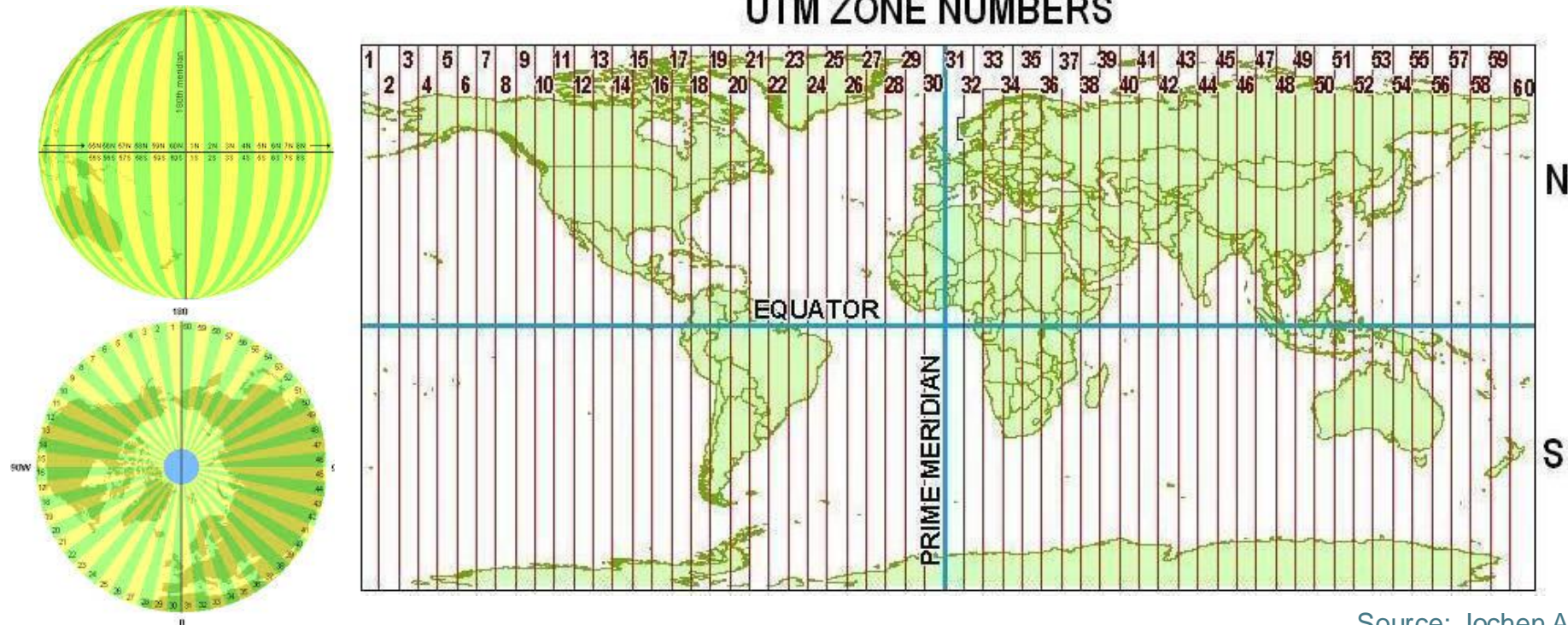 - WGS1984 (World Geodetic System) – best fitting ellipsoid for the globe/world

Courtesy of NOAA

Commonly encountered systems

- Projected Coordinate System

- UTM (Universal Transverse Mercator) – often best for large regions

Source: Jochen Albrecht

Always consider "scale" in your geospatial analysis

LOCAL

REGIONAL

GLOBAL

Source

# Always consider "scale" in your geospatial analysis

	NETWORK	FLOWS	SPATIAL CONSTRUCTS
LOCAL	• Transit systems • Street networks	• Commuting • Personal and social trips • Deliveries	• Activity space • District / Neighborhood • Terminal / Development zone • Town / City
REGIONAL	• Commuter rail • Regional air networks • National highway systems • National railway systems • Short sea shipping / feeders	• Intercity passenger flows • Distribution	• Metropolitan area • Market area • Hinterland / Corridor • Urban region
GLOBAL	• International air networks • Maritime shipping networks • Telecommunication networks	• Trade • Tourism and business trips • Migration	• Value chains • Landbridge • Trade area

Source

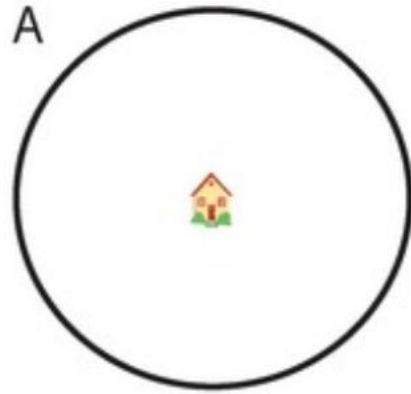# Neighborhood is a contextual concept

Neighborhood is defined as a circle centered on the house, extending equally in all directions.

Neighborhood is a contextual concept

Neighborhood is defined as a circle centered on the house, extending equally in all directions.

A

B
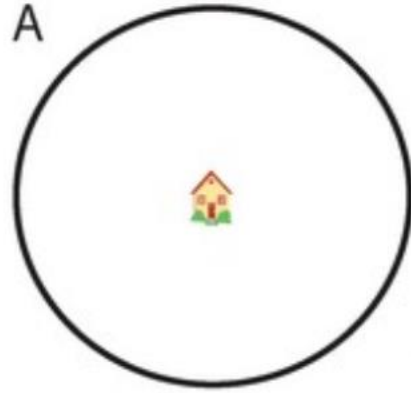
Neighborhood is equated with an existing zone, i.e., census tract or precinct, reflecting the common strategy of using existing aggregated data to characterize a household's surroundings.

Neighborhood is a contextual concept

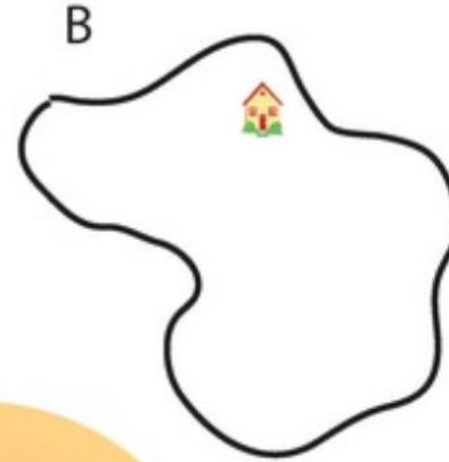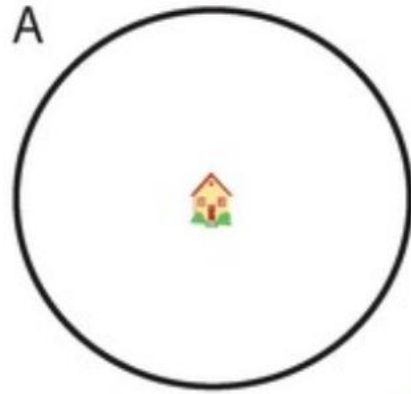Neighborhood is defined as a circle centered on the house, extending equally in all directions.

Neighborhood is equated with an existing zone, i.e., census tract or precinct, reflecting the common strategy of using existing aggregated data to characterize a household's surroundings.
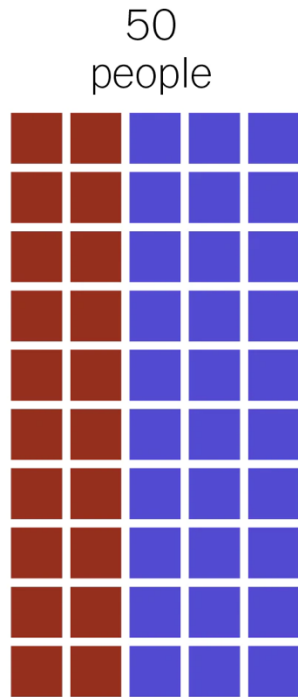
Weights are applied to surroundings based on distance, allowing neighborhood to be defined as a convolution with weight decreasing as a simple function of distance.
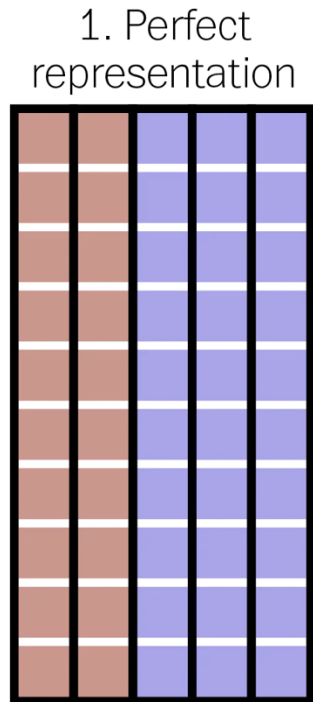
Gerrymandering – a controversial topic in elections

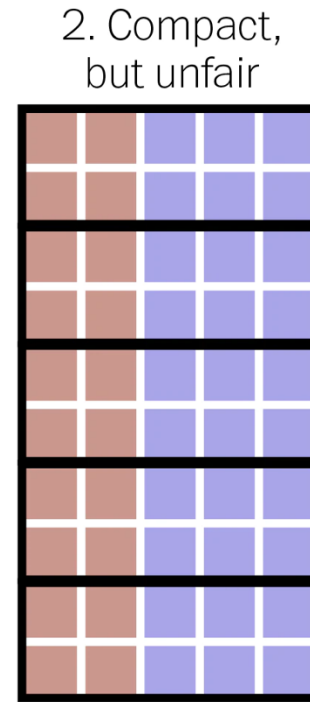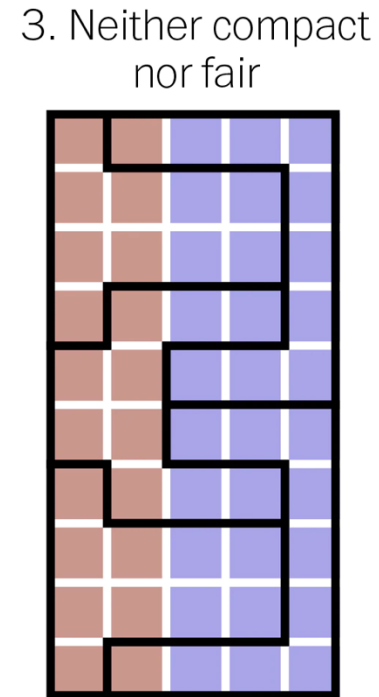Three different ways to divide 50 people into five districts

| 50 people	1. Perfect representation	2. Compact, but unfair	3. Neither compact nor fair
60% blue, 40% red	**3 blue districts, 2 red districts**	**5 blue districts, 0 red districts**	**2 blue districts, 3 red districts**
	BLUE WINS	**BLUE WINS**	**RED WINS**

"Disclaimer: The color is imaginary, and if it matches with any real-world scenario, I am not responsible for any resemblance or confusion." 🙂🙃

Source

Spatial sampling helps us making inference about real population without using the entire population

Sample

Collect and analyze data from sample

Make inferences about population

Population
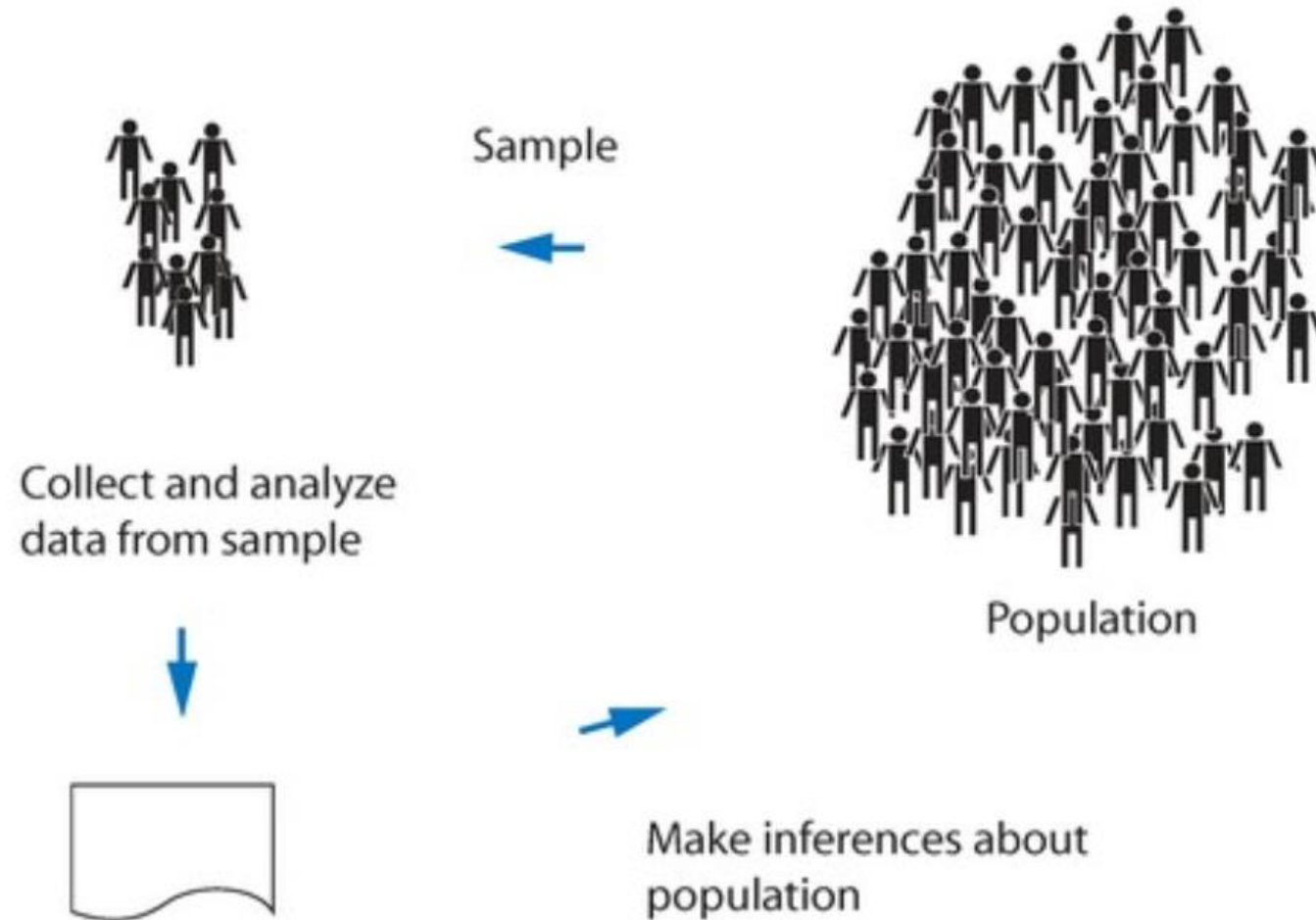
Econometrics is the statistical analysis of economic data and involves lots of spatial statistics

- Applies econometric techniques to spatial data, considering location-based dependencies.
- Accounts for the similarity between nearby spatial data points.
- Models how neighboring regions influence each other's economic outcomes.
- Adjusts for spatially correlated errors in data analysis.
- Used in analyzing geographic patterns in fields like urban planning and environmental economics.

- Location Quotient
- Spatial Regression
- Spatial Lag
- Shift-share analysis

Spatial heterogeneity is everywhere

Source

- The Earth's surface displays incredible spatial heterogeneity
- Analysis performed in one area cannot be simply transferred to another area without considering its heterogeneity
- Local relationship should be given priority in specific analysis

# Spatial dependence – the first rule of geography

- Tobler's rule: *"All things are related, but nearby things are more related than distant things."*

- Spatial dependence underlines the field of **Geostatistics**

- If we know the information of certain location, we can "interpolate/predict" similar information based on location

Mean Annual Precipitation [mm]
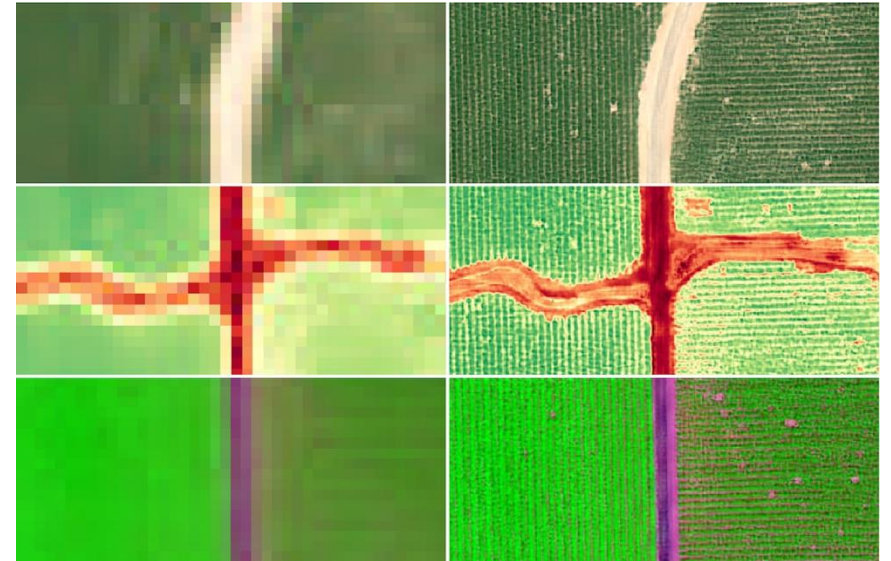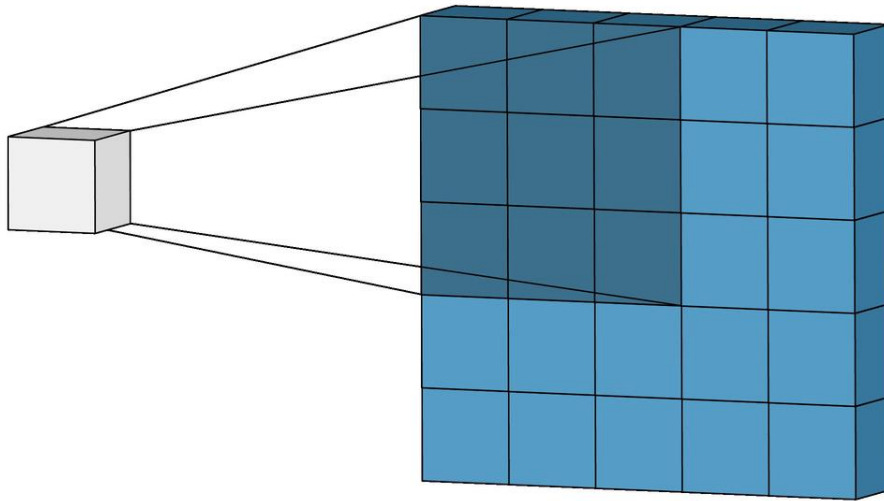
< 100 mm   > 3500 mm
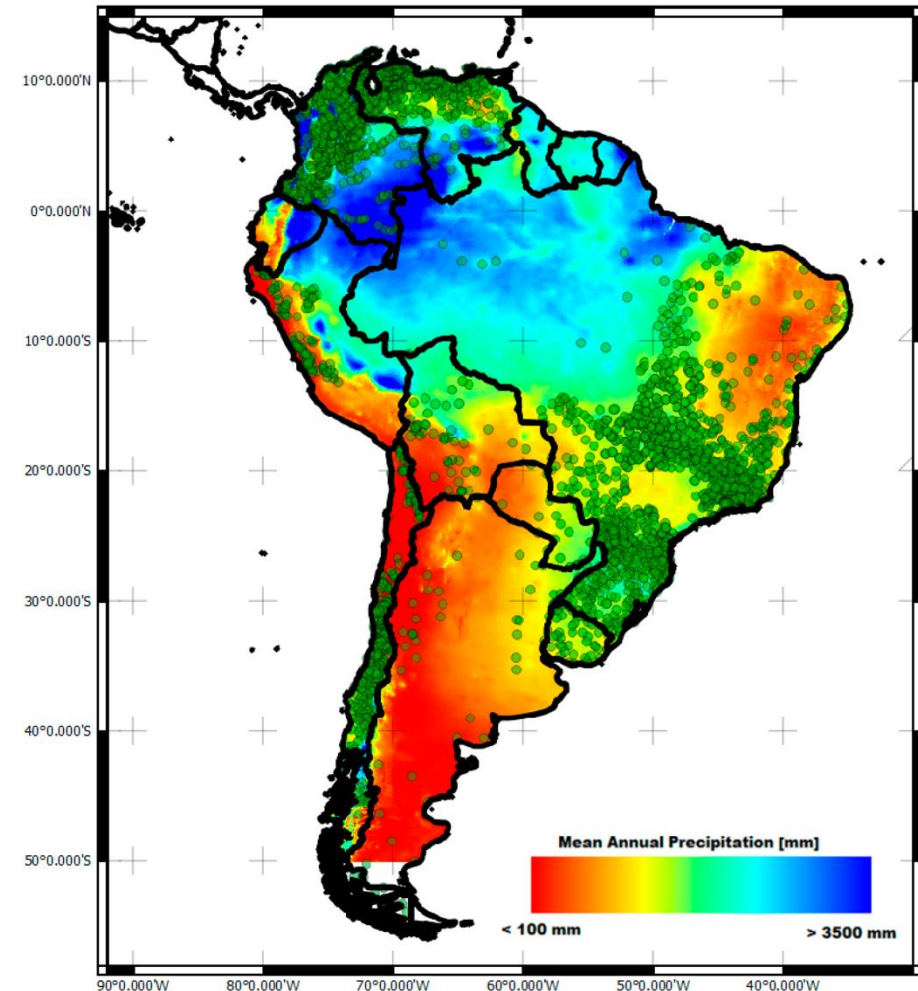
Source

Spatial interpolation uses geostatistics principles to interpolate unknown values at given location
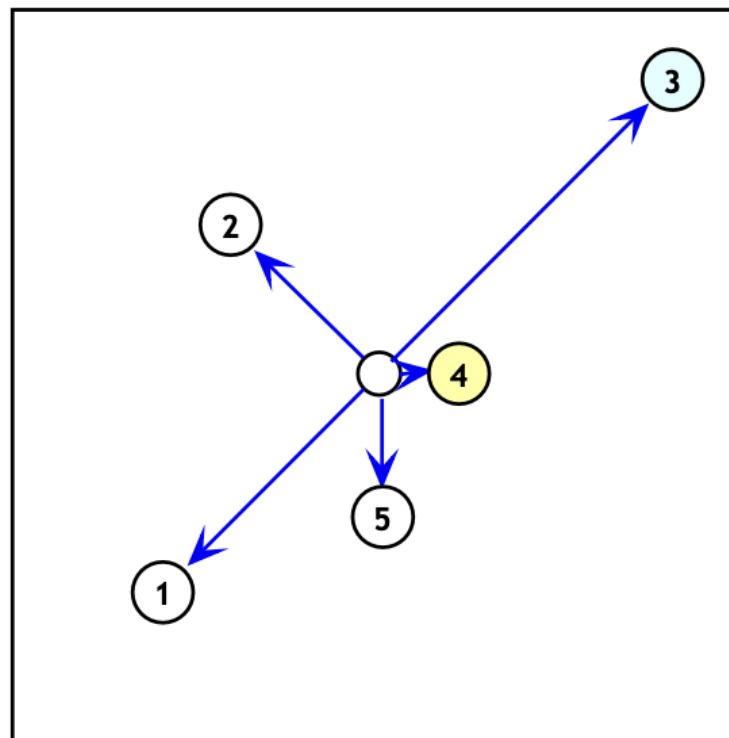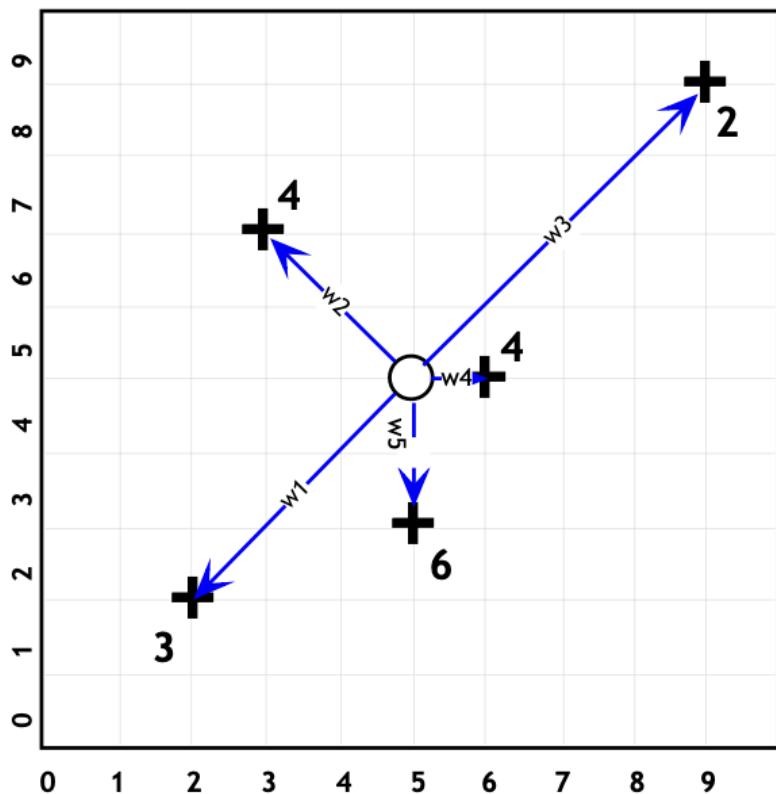
- We can only interpolate **continuous** variable, NOT discrete variables
- Deterministic interpolation and geostatistical interpolation

Smoothing often generalizes a complex spatial process and helps with the inference

Source

Summary

- The context of geospatial analytics
- Vector and raster data model
- Different formats of geospatial data
- Projection systems
- Spatial Statistics
  - Scale, neighborhood, spatial heterogeneity, dependence, interpolation, sampling, smoothing and sharpening
  - Spatial econometrics

Thank You