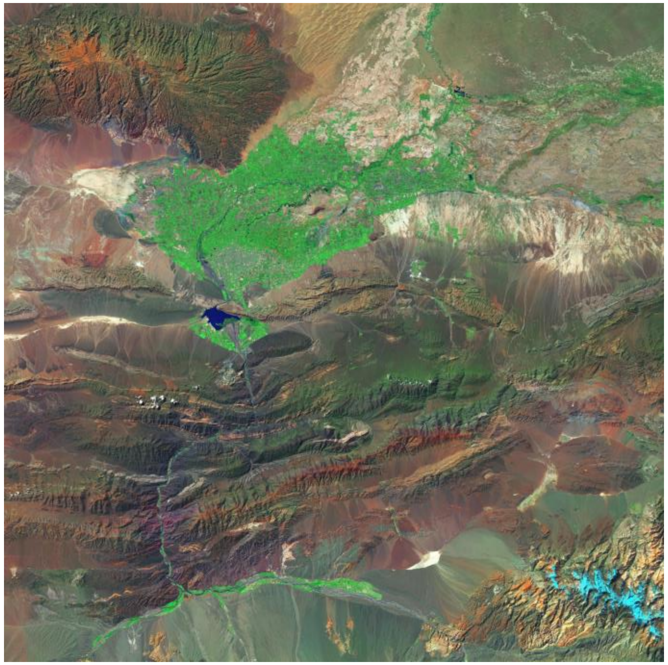
CSE 594: Spatial Data Science & Engineering

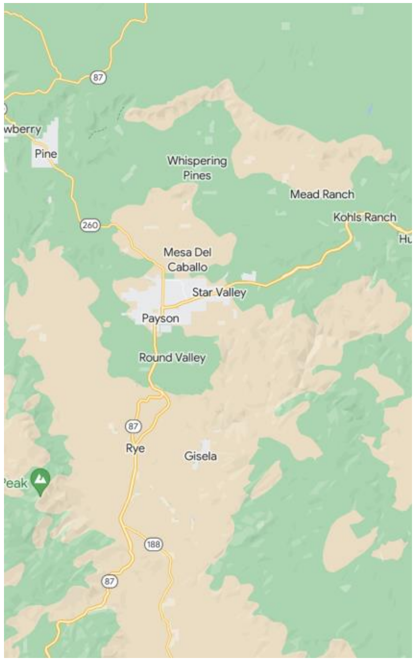
Lecture 2

Spatial Data Science Basic Concepts

Spatial Data

- Any data that reference geographical areas or locations either directly or indirectly
- Each data instance is related to one or more geographical locations
- May contain non-spatial attributes also.



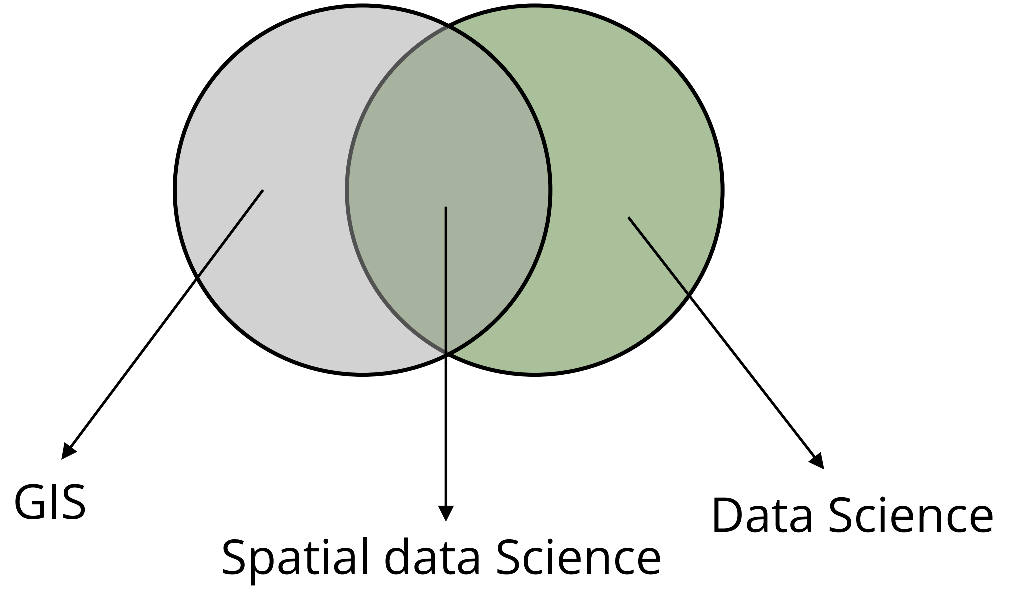


Spatial Data Examples

- Road networks and transportation data
- NASA satellite imagery
- Climate and weather data
- Rivers, farms, and ecological data
- Elevation data
- Census results
- Crime repots
- Accident reports
- Soil inventories
- Vegetation inventories
- Housing data

Spatial Data Science

- Subset of data science that focuses on unique characteristics of spatial data
- Getting insights from data using spatial analytical methods, spatial machine learning and deep learning algorithms



Applications of Spatial Data Science

- Analyzing outbreak of a new disease in some specific locations. Figuring out how the disease spreads and the reasons for spreading.

- Reducing pickup wait time by analyzing historical uber pickup location and time periods.

- Predicting the regions where house prices will increase the most in next five years.

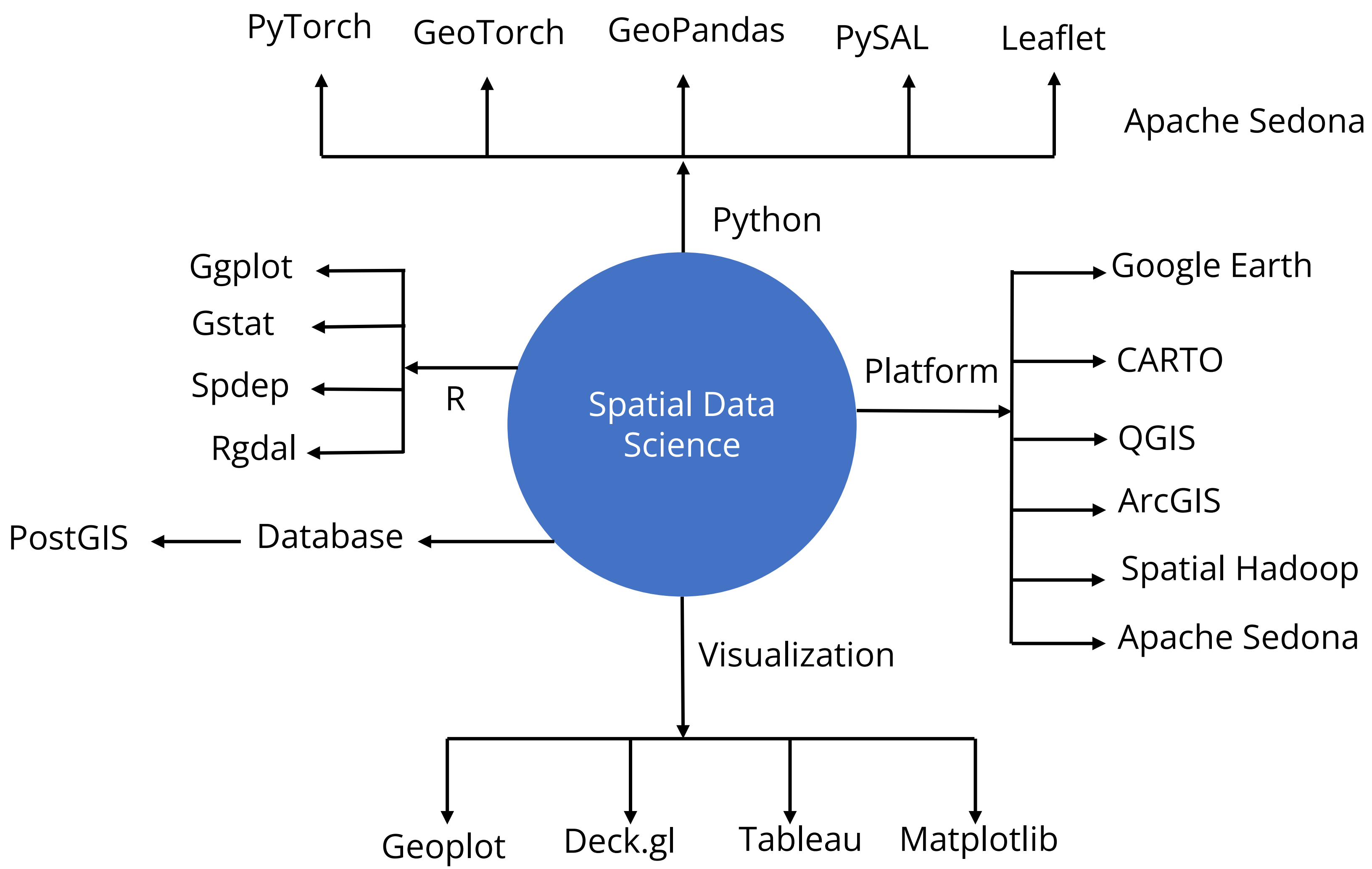
- Exploring the crime reasons at a location analyzing the crime rate, literacy rate, and other factors of the corresponding location along with neighbors.

- Flood risk analysis to detect areas susceptible to flooding

Applications of Spatial Data Science

- Detecting natural disaster, spread of wildfire, wind speed and direction
- Predicting traffic speed and traffic volume at various road intersections
- Classification of satellite images, such as roof top detection, urban and rural area detection, land cover classification

Tools for Spatial Data Science



Why Spatial Data is Special?

- Gigantic volume
- Everything we do in day-to-day life has some kind of spatial information
- Understand where and what is occurring in your neighborhood
- Combine information from many independent sources and derive new set of information
- Spatial attributes always have direct or indirect impact on neighbors
- Used for solving complex location-oriented problems, finding trends and patterns
- Location-oriented problems are complex to solve, regular solutions result in poor accuracy

Spatial Data Types

Geographic Data

- Data that can be mapped to a sphere (earth)
- Refers to latitude and longitude information of an object
- Example – GPS data

Geometric Data

- Data that can be mapped to a two-dimensional flat surface
- Example – building floor plans
- Google map uses geometric data to provide directions

Spatial Data Formats/Models

Vector Data

- Comprised of vertices and paths
- Usually stored in .shp files (known as shape files)
- Example – points, lines, and polygons

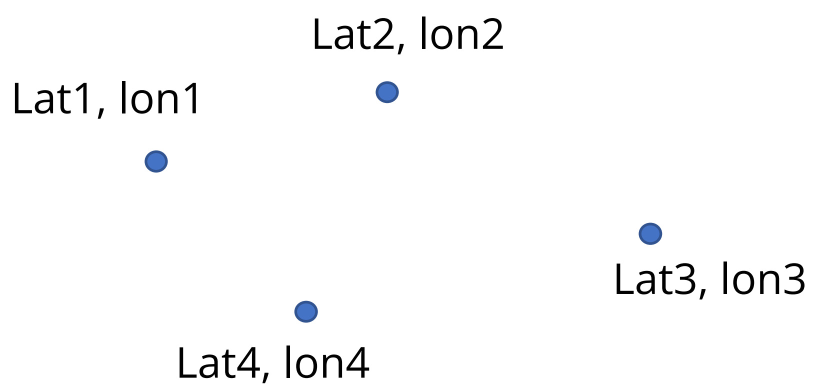
Raster Data

- Comprised of pixels or grid cells
- Usually regularly spaced and square/rectangle
- Each pixel has its own attribute values

Vector Data Model

Points Data

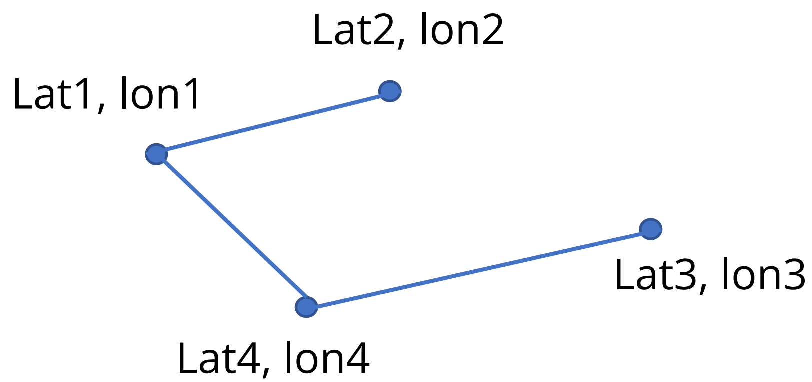
- Represented by Latitude and longitude (XY coordinates)
- Based on a specific coordinate reference system (CRS)
- Points of interests, such as restaurants, hospitals, grocery shops, etc.
- Location objects which are very small



Vector Data Model

Lines Data

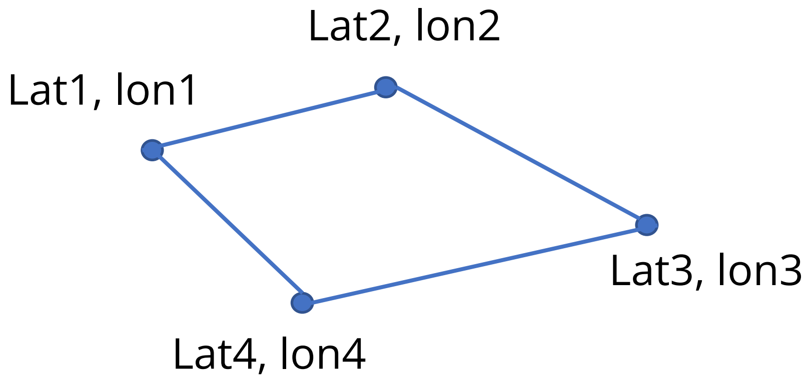
- Lines are formed by connecting points
- Points in a line follows an order, each point represents a vertex
- Roads, highways, rivers



Vector Data Model

Polygons Data

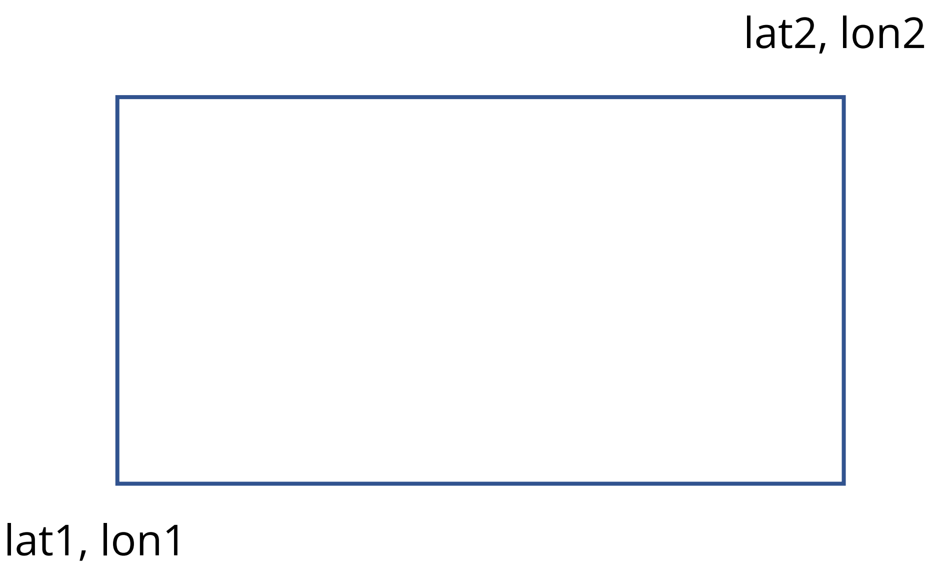
- Formed by connecting vertices and closing the path
- Similar to lines, connecting vertices follow an order
- First and last vertices (coordinates pairs) are same
- Represents boundaries, large areas



Other Elements of Vector Data Model

Envelop Data

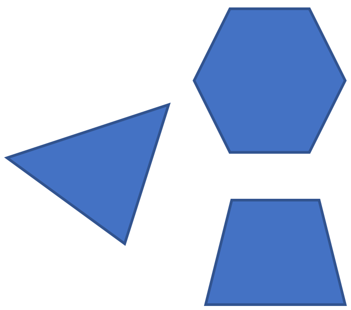
- An axis-aligned box described by the coordinates of the lower left corner and the coordinates of the upper right corner
- Sometimes they can be described by center, width, and height



Other Elements of Vector Data Model

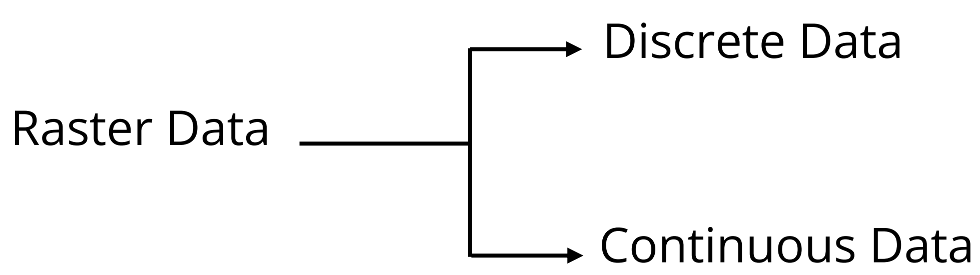
Multipolygon Data

- A collection of polygons
- Useful for gathering a group of polygons into one geometry
- An example use is to represent states or countries that include islands, or that are otherwise made up of non-overlapping shapes



Raster Data Model

- Represented as pixels similar to images
- Each pixel represents one or more attribute values
- Pixels in a satellite image might have red, green, blue band values, pixels in an elevation map represent heights, pixels might also represent land cover, rainfall, temperature, etc.



Raster Data Model

Discrete Raster Data

- Each pixel can have only one of few specific values
- Each pixel value represents a type
- Example – pixel might represent land cover classes, such as 1 for urban, 2 for rural, 3 for forest.

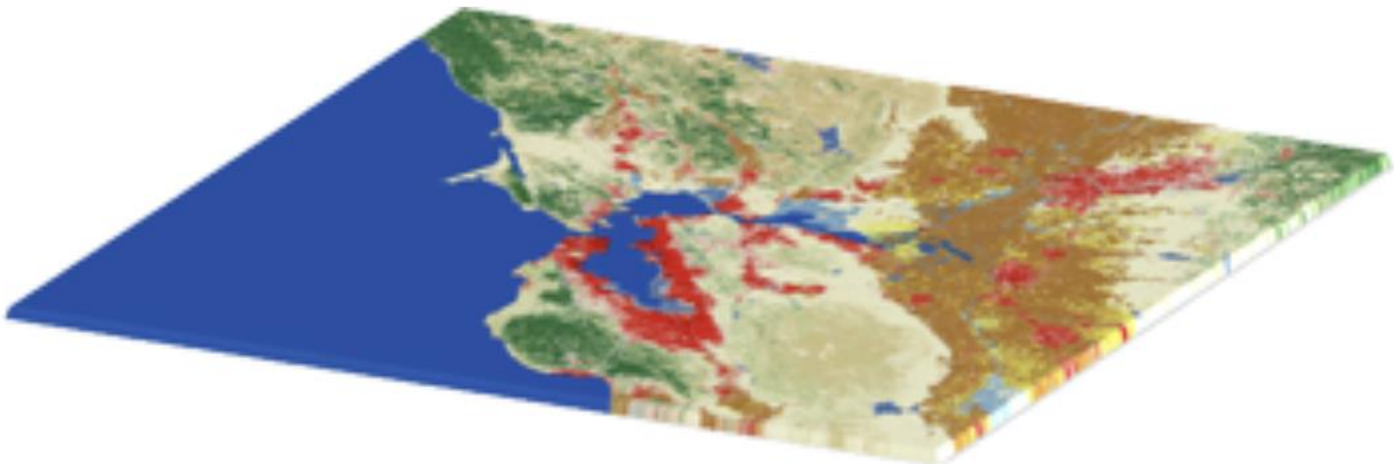


Image Sourcehttps://gisgeography.com/spatial-data-types-vector-raster/

Raster Data Model

Continuous Raster Data

- Each pixel has gradually changing data
- Example – temperature, rainfall, elevation, etc.
- A raster depicting oil spill represents the change in oil concentration from high to low

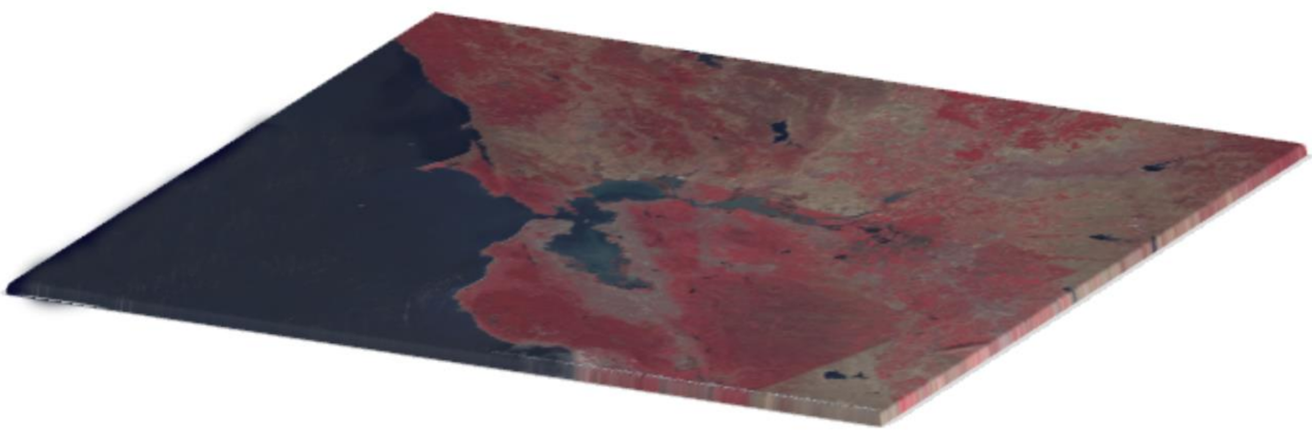


Image Sourcehttps://gisgeography.com/spatial-data-types-vector-raster/

Pros and Cons of Vector Data

Pros

- Good for graphical representation
- Provides higher geographic accuracy as it isn’t dependent on grid size

Cons

- Not good for storing continuous data
- Storing continuous data requires substantial generalization
- Proximity calculation is computation intensive

Pros and Cons of Raster Data

Pros

- Good for storing remote sensing data
- Because of fixed cell sizes, easy to map to geographic positions
- Map algebra operations are quick and easy

Cons

- Cannot be used to create network datasets
- Can become potentially very large in case of high-resolution grids

Storing Data as Vector Type VS Raster Type

- Coordinates vs pixels
- Scale objects to various sizes vs fixed size
- Restriction on file size?

Satellite Image vs Regular Image

Regular Image

- Images captured with traditional cameras
- Taken from much closer location to earth surface – clear and good resolution
- Usually does not have additional data tags except pixel values
- Usually contains two or three channels

Satellite Image

- Captured with electronic scanners incorporated in satellites
- Contains lots of noises
- Has geographical information of tagged within the image
- May contain more than three bands

Georeferencing

- The process of assigning coordinates to vectors or rasters so they can be oriented accurately on a model of the Earth's surface
- In the case of satellite image, the process incudes adding geographic information to the image so that GIS or mapping software can place the image in its appropriate real-world location

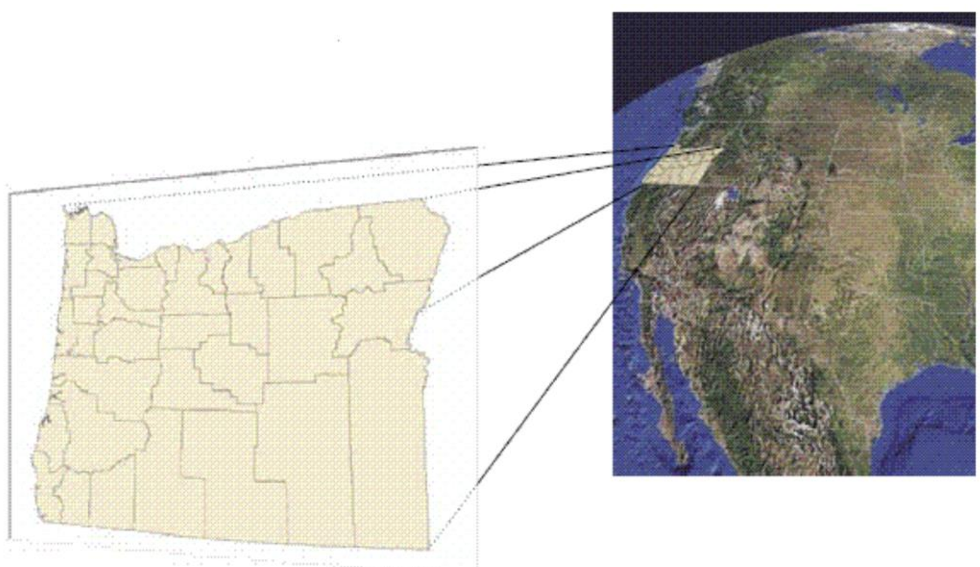


Image Source: http://gisresources.com/georeferencing-2/

Geocoding

- The process of converting address (such as street address) into geographic coordinates (such as latitude and longitude)
- The referenced coordinates can be used to place markers on a map
- Reverse geocoding is the process of converting coordinates into human-readable address

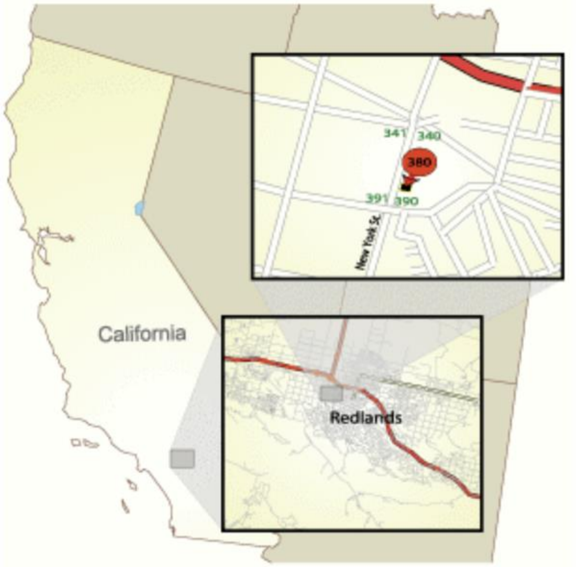


Image Source: https://desktop.arcgis.com/en/arcmap/latest/manage-data/geocoding/what-is-geocoding.htm

Coordinate Reference System (CRS)

- Map projections portray the surface of the earth on a flat piece of paper or computer screen
- Coordinate reference system defines how the projection relates to the real places on earth
- Three types of map projections are commonly used – cylindrical projection, conical projection, and planar projection
- None of the projections are fully accurate
- Distortions happen for angular conformity, distance, and area

Coordinate Reference System (CRS)

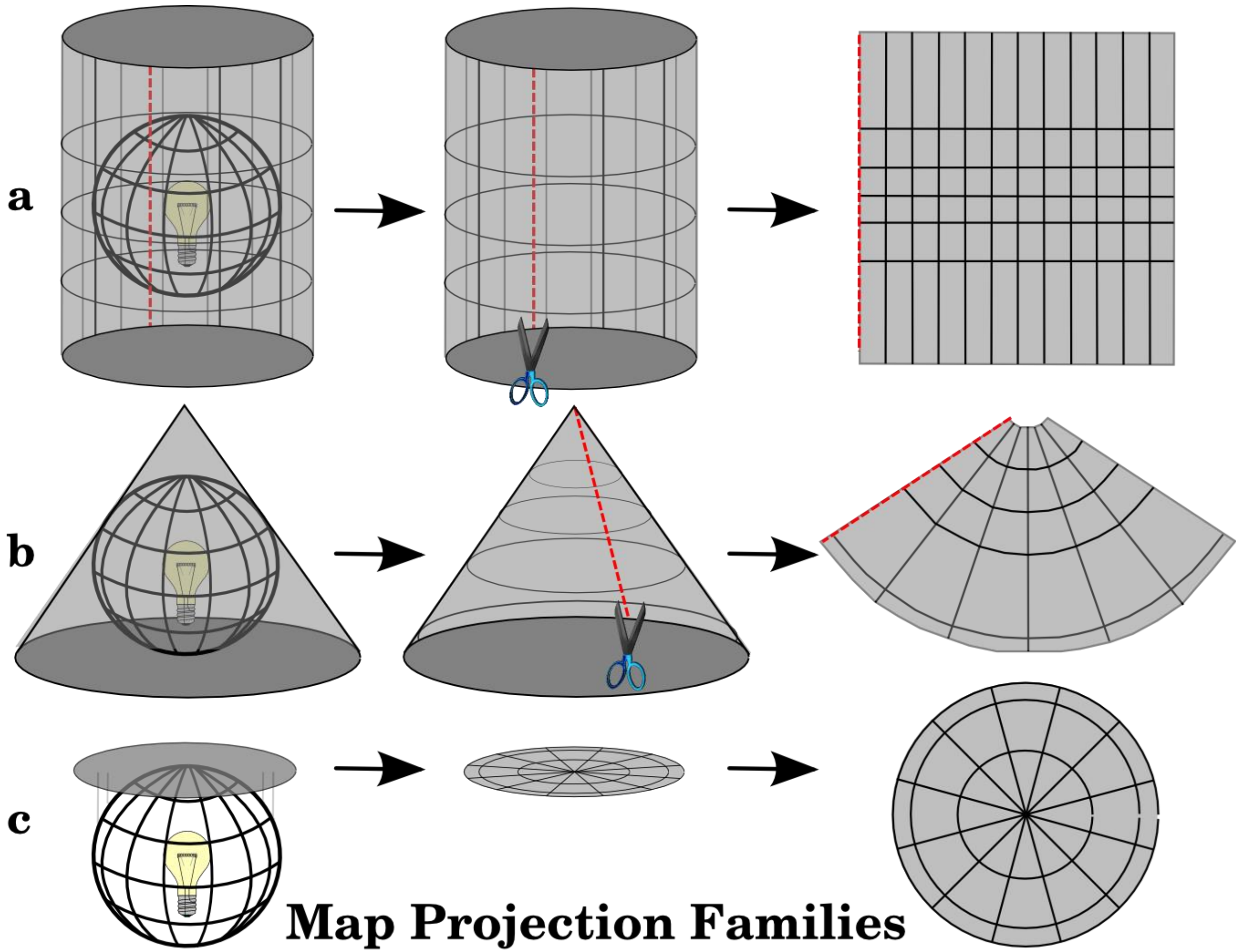


Image Source: https://docs.qgis.org/3.22/en/docs/gentle_gis_introduction/coordinate_reference_systems.html

Coordinate Reference System (CRS)

Comparing Map Projections

- Cylindrical map projections are accurate near the equator but distorts distances and sizes near the pole
- Shapes of small areas are preserved well by cylindrical projection
- Conical projection is good for a regional map instead of a complete world map
- Planar projections accommodate circular regions better than rectangular regions because area and shape distortion are circular around the point of contact
- Planar projections are used most often to map polar regions

Spatial File Formats

- Vector Data Files
  - Geodatabase
  - KML/KMZ
  - Open Street Map
  - Shape File
  - GeoJSON File
  - WKB/WKT
- Raster Data Files
  - JPG/JPEG
  - GIF
  - PNG
  - BMP
  - TIF/TIFF/GeoTIFF

Vector Data Files

Geodatabase

- A collection of files in a folder on disc that hold related geospatial data
- Mandatory file .gdb is kept alongside some other files in the same folder
- Can store, query, and manage both spatial and nonspatial data
- Contains system tables for managing geospatial functionality plus user data
- Default maximum size is 1 TB, can be increased to 256 TB
- Recommended native file format by ESRI for data storage in ArcGIS

Vector Data Files

KML/KMZ

- KML stands for Keyhole Markup Language
- Developed by Keyhole, then acquired by Google
- Originally used for viewing geographical data in Google Earth
- KMZ files are zipped files with a main KML file and associated support files

```
<?xml version="1.0" encoding="UTF-8"?>
<kml xmlns="http://earth.google.com/kml/2.2">
  <Placemark>
    <name>Stonehenge, England</name>
    <description>Stonehenge was built about 2500BC
    </description>
    <Point>
      <coordinates>-1.826752,51.179045
      </coordinates>
    </Point>
  </Placemark>
</kml>
```

Source: http://web.archive.org/web/20120111193310/http://code.google.com/apis/kml/documentation/whatiskml.html

Vector Data Files

Open Street Map

- File extensions .osm, .bz2, .pbf
- Contains XML formatted data in the form of nodes, connections, relations, and tags
    - ➢ Nodes are geographic positions stored as pairs of latitude and longitude
    - ➢ Connections/ways are represented as sorted list of nodes
    - ➢ Relations denote barriers, u turns, area with holes
    - ➢ Tag are metadata about the map objects, describes features such as buildings, roads

Vector Data Files

GeoJSON File

- An open standard geospatial data interchange format that represents simple geographic features and their nonspatial attributes


```
{ "type": "FeatureCollection",
  "features": [
    { "type": "Feature",
      "geometry": {"type": "Point", "coordinates": [102.0, 0.5]},
      "properties": {"prop0": "value0"}
      },
    { "type": "Feature",
      "geometry": {
        "type": "LineString",
        "coordinates": [
          [102.0, 0.0], [103.0, 1.0], [104.0, 0.0], [105.0, 1.0]
          ]
        },
      "properties": {
        "prop0": "value0",
        "prop1": 0.0
        }
      },
    { "type": "Feature",
       "geometry": {
         "type": "Polygon",
         "coordinates": [
           [ [100.0, 0.0], [101.0, 0.0], [101.0, 1.0],
             [100.0, 1.0], [100.0, 0.0] ]
           ]
       },
       "properties": {
         "prop0": "value0",
         "prop1": {"this": "that"}
         }
       }
     ]
   }
```


Source: https://geojson.org/geojson-spec.html

Vector Data Files

Shape File

- The most popular geospatial file format
- A set of three mandatory files along with some optional files under the same folder
- Mandatory files .shp contains feature geometry, .shx contains indexing info, .dbf contains attribute data
- Optional files .prj contains the projection metadata, .xml contains associated metadata

Vector Data Files

WKB/WKT

- WKT stands for Well-Known Text representation
- Made up of three components: geometry type, coordinate type and coordinate list
- Coordinate type indicates whether or not the geometry has Z coordinates and a referencing system
- WKT is a textual representation of spatial information as a text markup language
- WKB stands for Well-Known Binary representation, a binary equivalent of WKT, represented as a contiguous stream of bytes

Raster Data Files

TIF/TIFF/GeoTIFF

- Stands for Tagged Image File Format
- Sizes are large, as only lossless compression is used
- Contains much more additional information as metadata, for example – number of channels
- GeoTIFF images contain geospatial information as metadata, such as geographical location of the source, coordinate reference system, etc.

Geometry Object Representations

Geometry	Representation
Point	Point(lat lon)
LineString	LineString(lat1 lon1, lat2 lon2, lat3, lon3)
Polygon	Polygon((lat1 lon1, lat2 lon2, lat3 lon3, lat4 lon4, lat1 lon1))
MultiPolygon	Polygon(((lat1 lon1, lat2 lon2, lat3 lon3, lat4 lon4, lat1 lon1)), ((lat5 lon5, lat6 lon6, lat7 lon7, lat8 lon8, lat5 lon5)))

Loading and Representing Spatial Datasets

Loading Spatial Data in Apache Sedona

File Format	Loader Method
Shape File	ShapefileReader.readToGeometryRDD(sparkContext, path_to_dataset)
WKB File	WkbReader.readToGeometryRDD(sparkContext, path_to_dataset)
WKT File	WktReader.readToGeometryRDD(sparkContext, path_to_dataset)
GeoJSON File	GeoJsonReader.readToGeometryRDD(sparkContext, path_to_dataset)
GeoTIFF File	spark.read.format("geotiff").options(**options_dict).load(path_to_dataset)

Loading and Representing Spatial Datasets

Representation of Spatial Vector File

Nonspatial Attribute 1	Nonspatial Attribute 2	Geometry
---	--	POLYGON ((933100.92 192536.09, 933091.01 192572.17, 933088.58 192604.97, 933779.28 195908.73, 933841.76 195957.79, 933100.92 192536.09))
---	---	MULTIPOLYGON (((1033269.24 172126.00, 1033439.64 170883.95, 1033473.26 170808.21, 1033269.24 172126.00)), ((1033422.35 157944.65, 1033419.99 157936.99, 1033408.21 157938.17, 1033422.35 157944.65)))
---	---	POLYGON ((933100.92 192536.09, 933091.01 192572.17, 933088.58 192604.97, 933779.28 195908.73, 933841.76 195957.79, 933100.92 192536.09))
---	---	POLYGON ((933100.92 192536.09, 933091.01 192572.17, 933088.58 192604.97, 933779.28 195908.73, 933841.76 195957.79, 933100.92 192536.09))

Loading and Representing Spatial Datasets

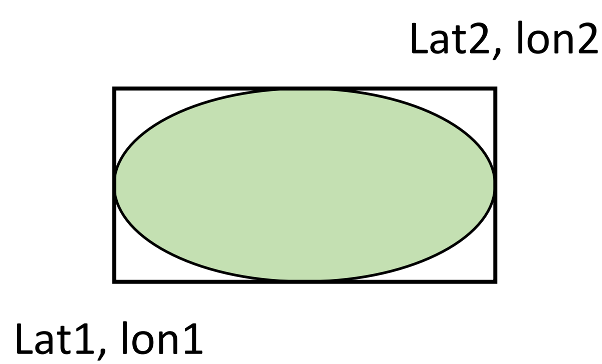
Representation of GeoTIFF Raster File

- root
  - Image: struct
    - origin: string
    - geometry: string
    - height: integer
    - width: integer
    - nBands: integer
    - data: array
      - element: diuble

Geospatial Geometry Operations

Minimum Bounding Rectangle (MBR)

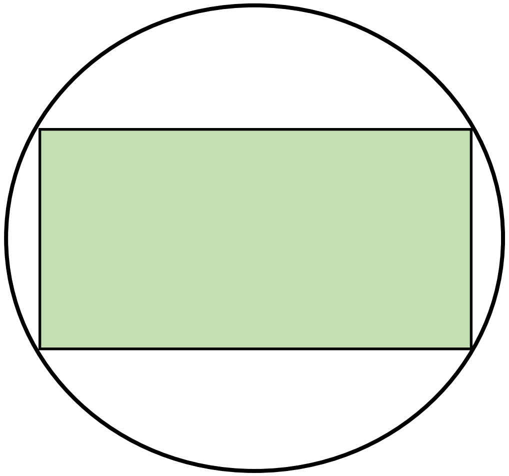
- MBR of a geometry is the smallest rectangle that covers the complete geometry
- The bounding geometry formed by the minimum and maximum (X,Y) coordinates
- Also known as Envelop
- Exception: MBR of a point or line does not form any rectangle



Geospatial Geometry Operations

Minimum Bounding Circle (MBC)

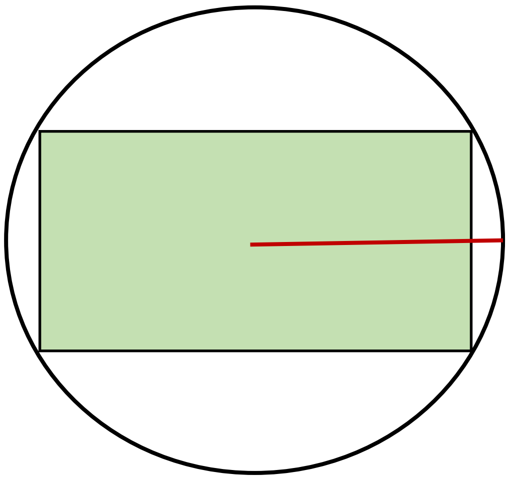
- The smallest circle that contains a complete geometry



Geospatial Geometry Operations

Minimum Bounding Radius

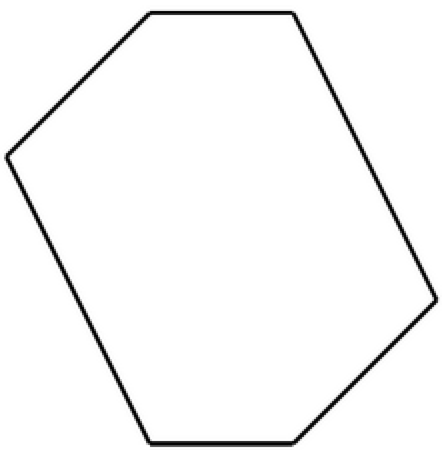
- The radius of the smallest circle that contains a complete geometry



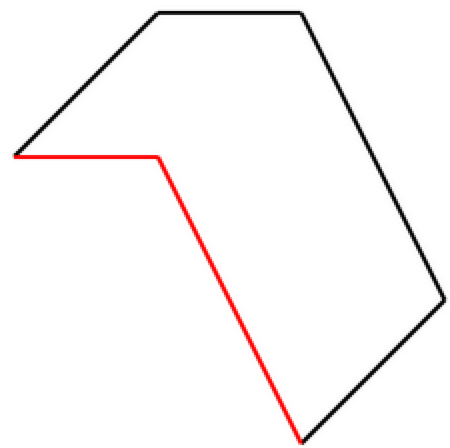
Geospatial Geometry Operations

Convex Hull of a Polygon

- Convex hull of a geometry is the smallest convex region enclosing the complete geometry
- Convex polygon means the polygon has no corner that is bent inwards



Convex Polygon



Non-convex Polygon

Other geospatial Geometry Operations

- ST_Distance(A, B) :- Returns Euclidean distance between geometries A and B
- ST_Within(A, B) :- Returns True if geometry A is fully contained by geometry B
- ST_Length (A) :- Returns the perimeter of geometry A
- ST_Area (A) :- Returns the area of geometry A
- ST_Centroid (A) :- Returns the center point of geometry A
- ST_Transform (A, crsSource, crsDest) :- Transforms the coordinate reference system of geometry A from crsSource to crsDest
- ST_Intersection(A, B) :- Returns the intersection geometry of A and B