Lecture 5
Introduction to Spatial Data

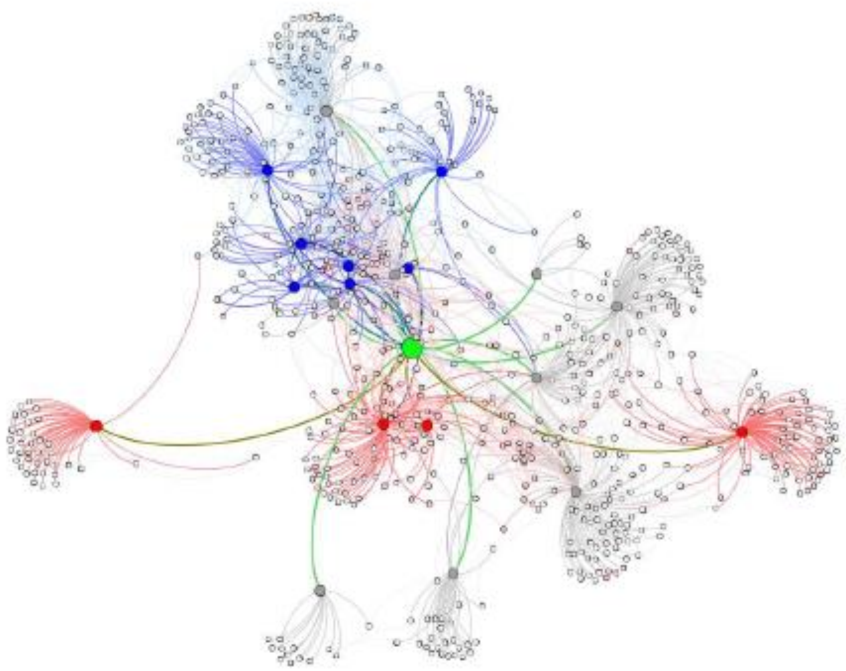
Pierre Biscaye
Université Clermont Auvergne

Data Science for Economics

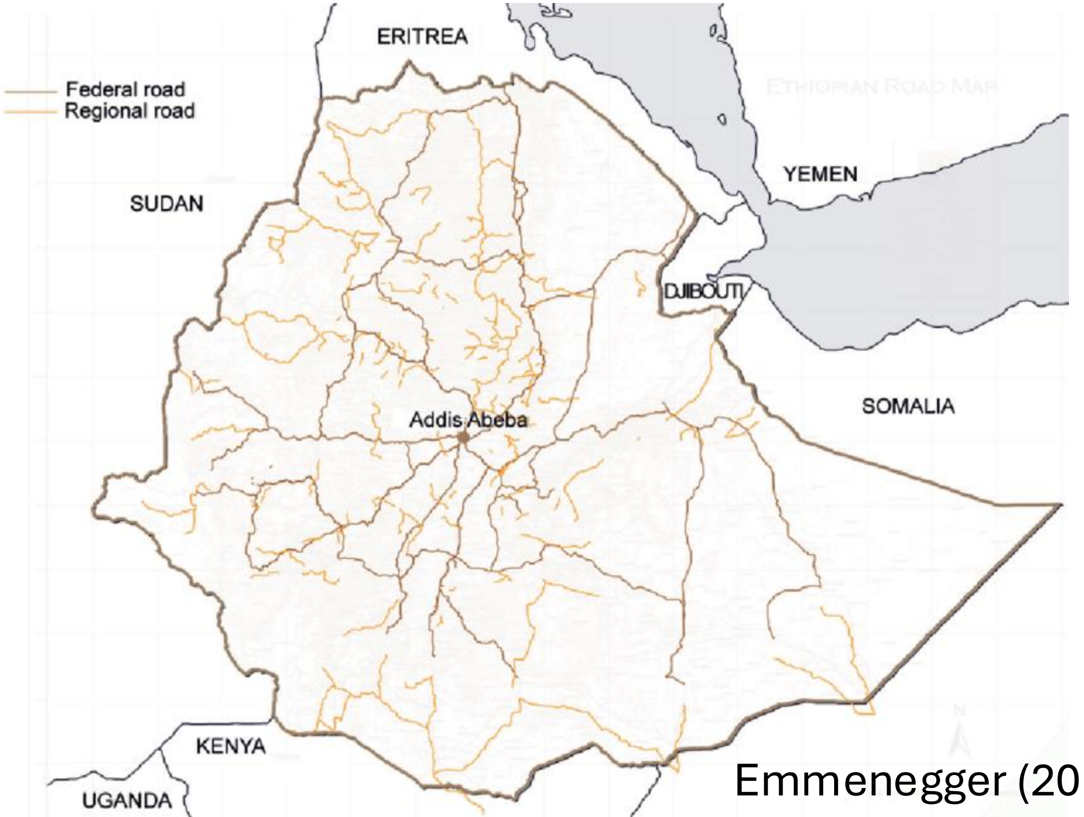
Note: Materials for this lecture are drawn from Sol Hsiang's Spatial Analysis course at UC Berkeley

What does it mean for data to have a spatial component?

- Information that can be mapped over space
- Not necessarily physical space! Ex: social network links



Blumenstock et al. (2019)



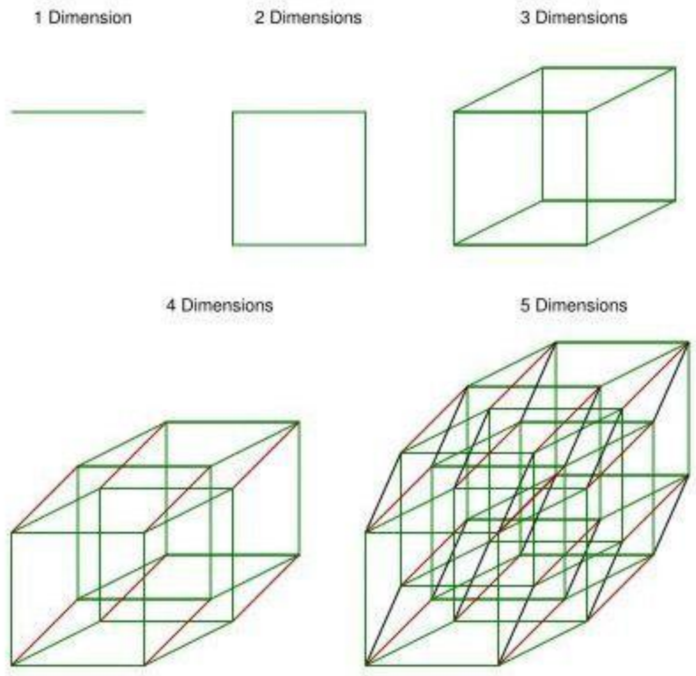
Emmenegger (2012)

Many types of data with spatial components

- Geographic data: land cover, land use, topography, elevation
- Climate and weather data, maps of disasters extent
- Remote sensing data: satellite imagery, nighttime lights
- GPS data: mobile GPS logs, locations of households, firms, cities, etc.
- Administrative boundaries
- Transportation networks
- Population density and settlement
- Geolocated mobile or internet use data
- Network graphs

Source

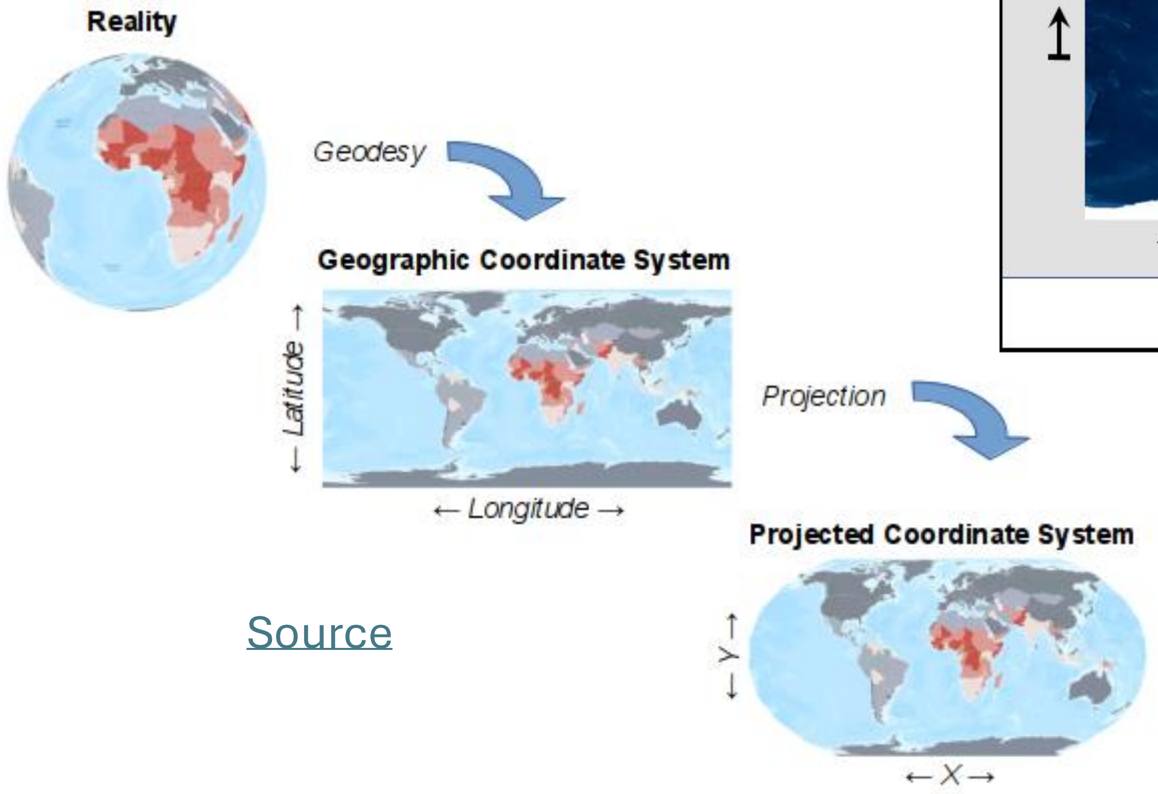
Important spatial concepts



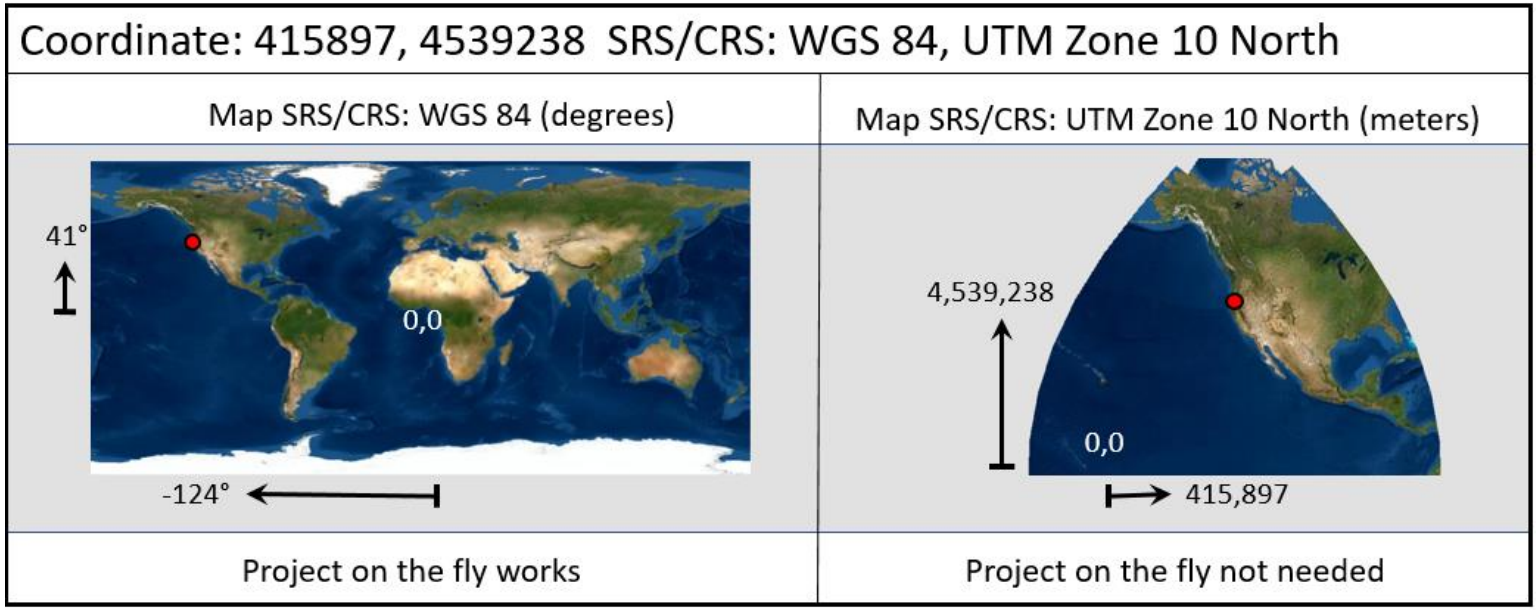
Source

- **Dimensions**: smallest quantity of elements needed to uniquely identify an object in space
- **Coordinate reference system**: set of indices to uniquely identify all locations within a space
- **Indices**: arguments of a coordinate system
  - Ex: latitude and longitude
- **Changing CRS** (coordinate reference systems): if two systems describe the same space, there must exist transformations to change from one to the other
  - Ex: latitude/longitude and mailing addresses both identify locations in 2D physical space
  - It is often necessary to change CRS to make different datasets consistent with one another; this may involve loss of information (ex: apartment numbers)

Knowing the CRS of a dataset is critical for correct interpretation!



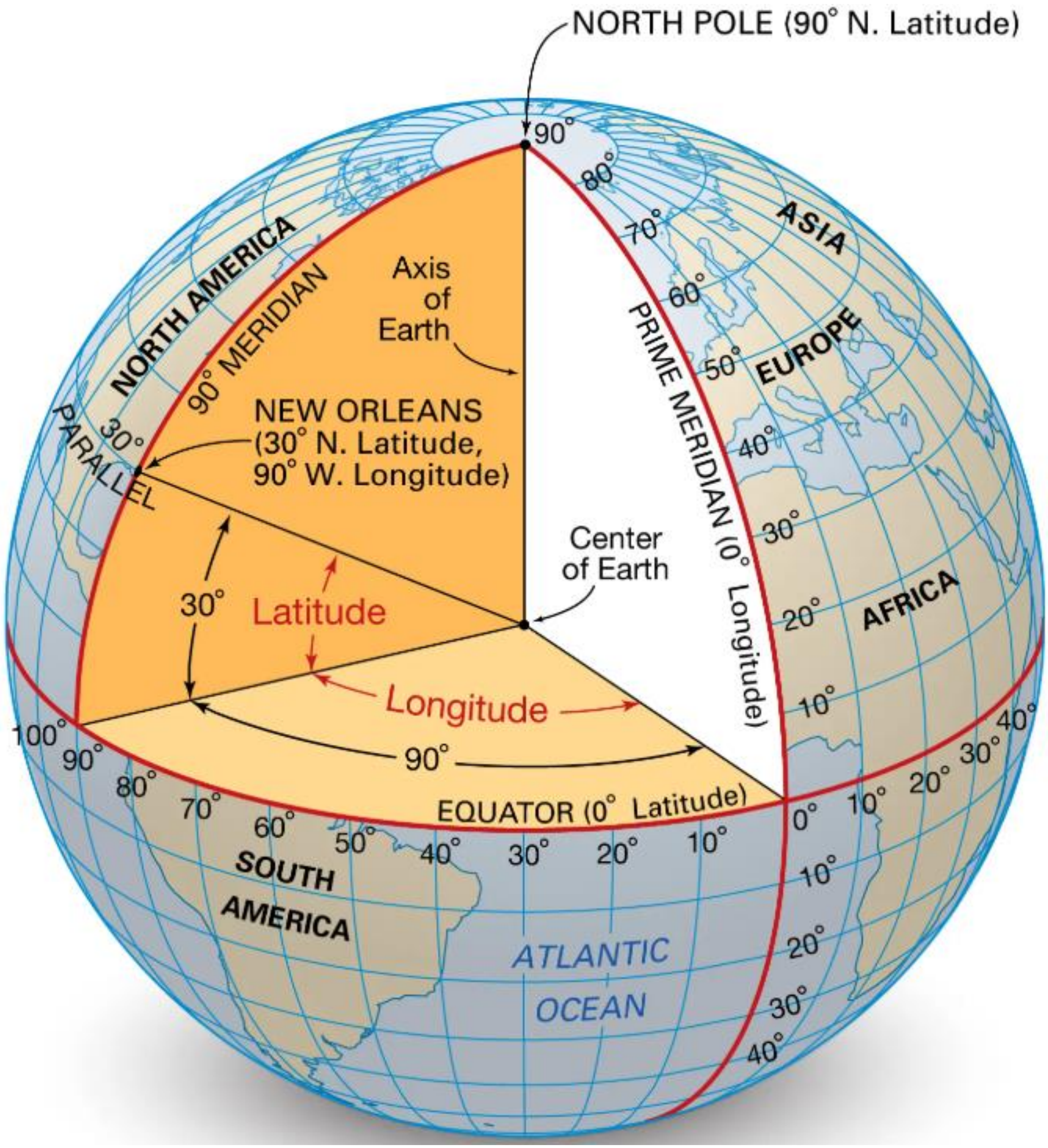
Source



Source

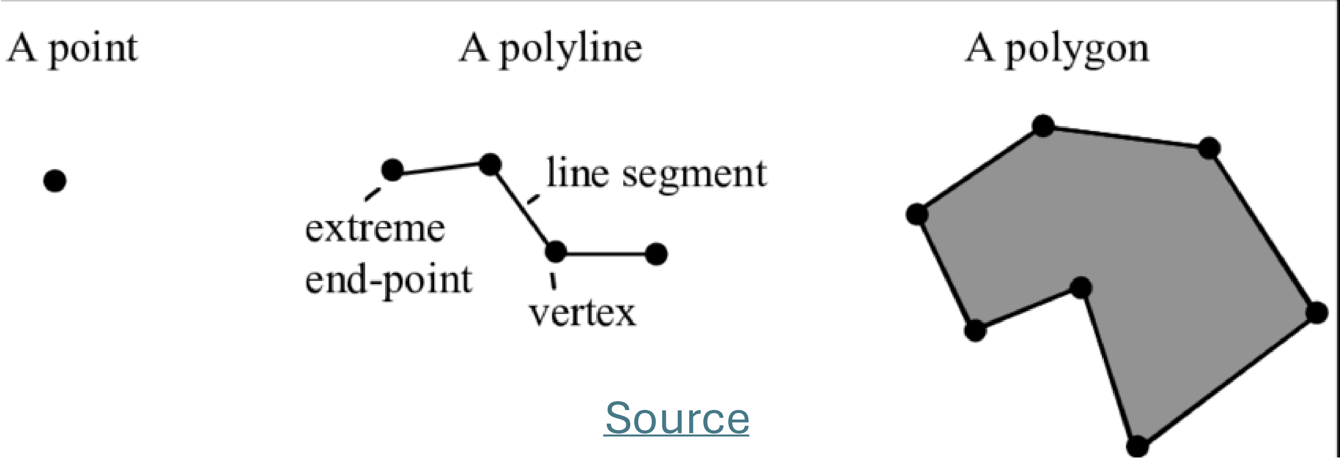
Positions on Earth

- Coordinates in geographic CRS given by latitude and longitude
- **Latitude:** angle relative to equator
  - Distance in km of 1º latitude is the same everywhere: 111.11 km
  - Half circumference of Earth divided by 180 degrees
- **Longitude:** angle relative to "prime meridian" at equator
  - Distance in km of 1º longitude depends on latitude: 111.11 km * cos(lat)



Source

Spatial shapes

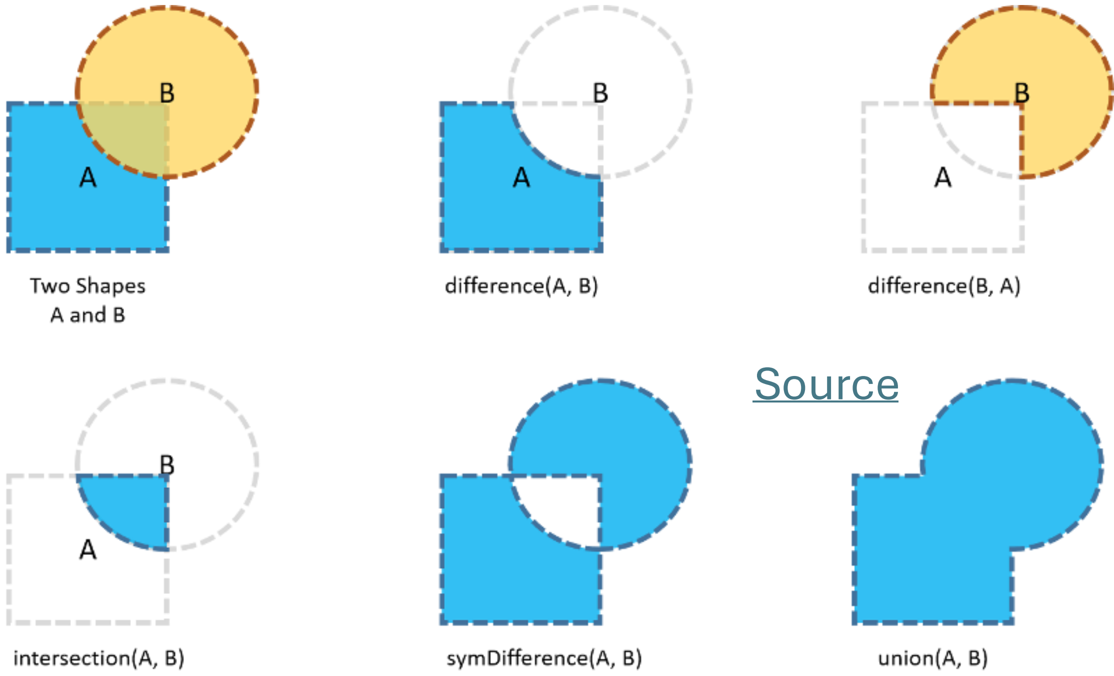


Source

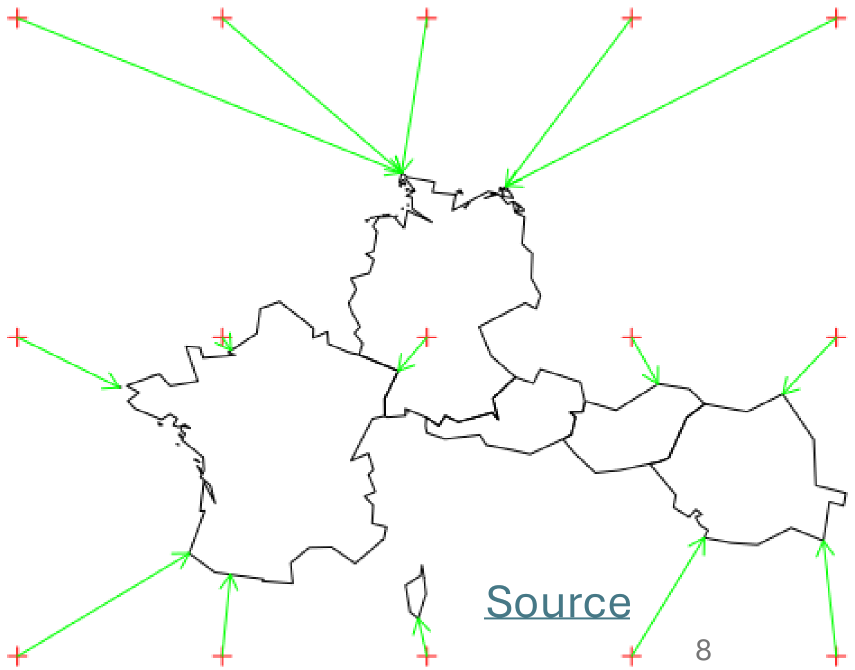
- **Point** (x,y)
- **Line** { (x1,y1), (x2,y2) }
- **Polyline**: sequence of line segments with sequential endpoints
- **Polygon**: polyline with identical first and last vertex
  - Convex: a line between any 2 points within polygon remains within polygon
  - Concave: not convex
  - Can take any shape
- **Multi-polygons**: set of polygons forming an object of interest
- **Buffers**: set of points within a given distance of polyline
- **Network**: set of points (vertices) and connections (edges)

Operations with shapes

- **Intersection**: area in common
- **Union**: combined area
- **Difference**: one area minus area of the other
- **Centroid**: average location of set of locations
- **Center of mass**: weighted average location
- **Distance**: to point, line, boundary, along a polyline, etc.
- **Connectivity** (in a network)

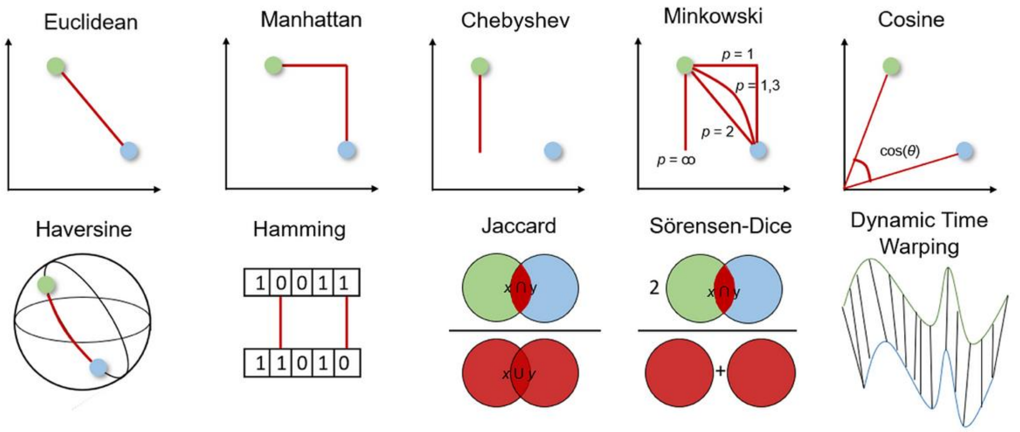


Source



Source

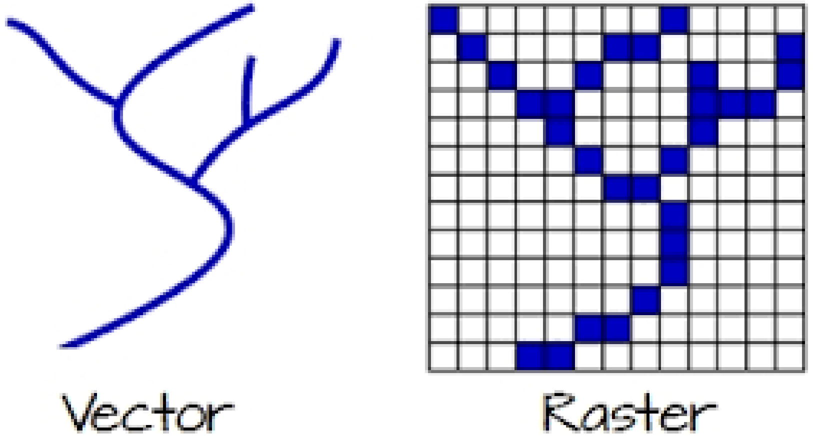
Measuring "distance"



Source

- D(**x**_1, **x**_2) is a measure of how different two locations **x**_1 and **x**_2 are from each other
  - Important concept for analyzing spatial relationships
  - Indices of **x** could be anything
- Euclidean distance: D(**x**_1, **x**_2) = sqrt(sum_i((**x**_1(i) - **x**_2(i))^2))
- With geospatial data, important to account for **nature of distance** of interest
  - Manhattan distance: distance when movement is constrained to a grid
  - Great-circle (geodesic) distance: shortest distance on surface of a sphere
- Other distance measures: Minkowski, Haversine, Mahalanobis, Cosine, Chebyshev, etc.

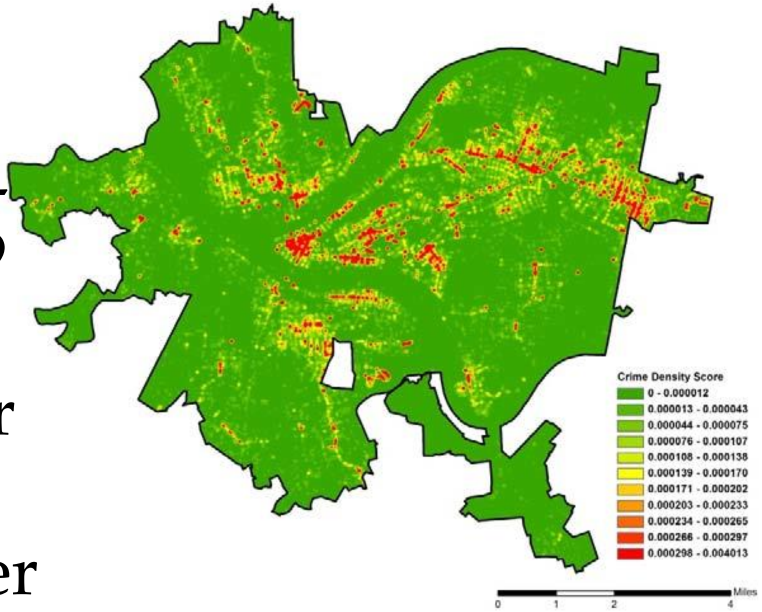
Fields and rasters



Source

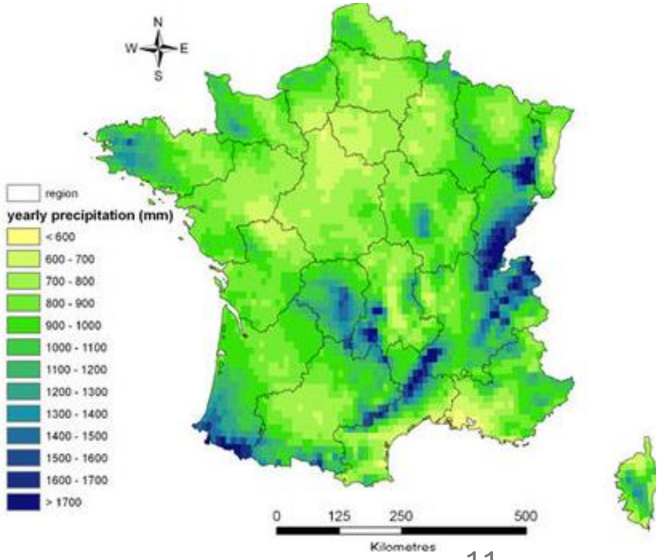
- **Field**: a function over space that takes on a value at every location
  - Example: temperature, population density
  - Vector fields: multiple values at every location
  - Dynamic fields: different values over time at every location
- **Raster**: approximation of a field using a grid, where points within grid cell all take on the same value
  - Identify positions by center of each grid cell
- **Resolution**: describes quality of approximation in raster
  - Photography: pixels/area
  - Elements depend on dimensions of the field
    - Geospatial resolution given in meters or degrees of the side of the grid cell
  - Data size increases rapidly with resolution

Spatial intensity and clustering



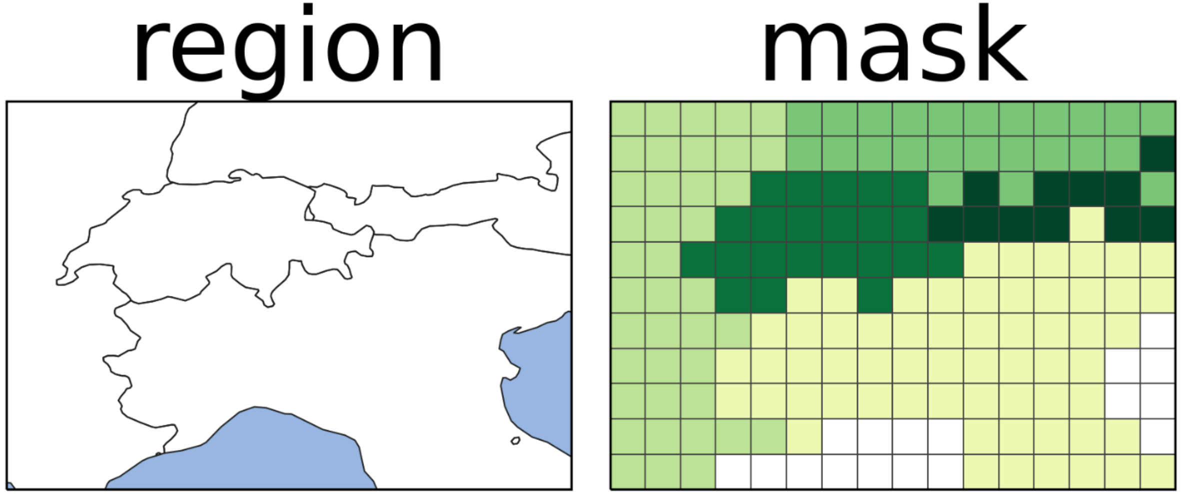
[Source]

- **Intensity**: Where do we observe higher values over space?
- **Clustering**: How do values relate to each other over space?
- For **points/events**: intensity and clustering of event locations
  - Example: count of events in a grid cell, kernel density (with weights as function of distance), distance to nearest neighbor, measures of centrality/dispersion, expected number of events at a given distance
- For **fields**: intensity and clustering of field values
  - Example: average value in a grid cell or other shape, average correlation between values at a given distance



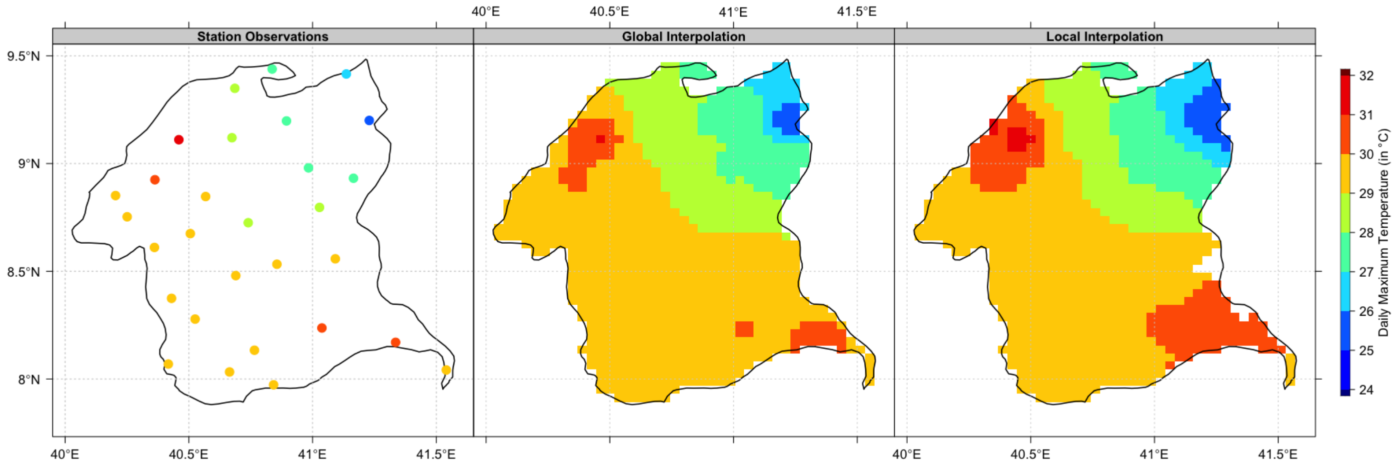
[Source]

Map algebra



Source

- Computing **functions of fields**
  - Each point assigned a value
  - Example: combine population and grid area → population density
- **Masking**: binary field indicating area of interest
- **Interpolation**: observe {Z_i} at subset of locations {**X**_i} where **X** is a vector of fields, and want to know Z everywhere
  - Polynomial regression: predict Z based on **X**
  - Nearest neighbor matching: replace with value of closest **X**
  - Inverse distance weighting: average value of other observations (within some distance) weighted by distance



Source