# GEOG 531 Quantitative Methods In Geographic Research

Statistical Data Types and Nature of Spatial Data

# Types of Statistics

- Based on domains:
  - **Classical statistics** (basic and fundamental)
  - Biostatistics
  - Econometrics
  - ……
  - **Spatial statistics (deal with spatial/geographic data)**
- Based on functions:
  - **Descriptive statistics** (summary)
  - **Inferential statistics** (forecasting, decision making)

# Descriptive Statistics

- Concerned with summarization and presentation of data
  - Numerical
    - mean/average, median, range
  - Tabular
    - tables
  - Graphical
    - histograms, scatter plot
  - Cartographical
    - Thematic maps

# Descriptive Statistics Example

Home sales of Milwaukee, Wisconsin, 2014 (subset)

| Address | Sqft | Sale $ |
|---|---|---|
| 129 W BOLIVAR AV | 1,246 | 130,000 |
| 1336 W VAN NORMAN | 1,143 | 87,000 |
| 137 W UNCAS AV | 1,521 | 150,000 |
| 1409 W WANDA AV | 1,059 | 168,000 |
| 1416 W HOWARD AV | 1,958 | 122,500 |
| 1420 W FOSTER AV | 1,208 | 136,000 |
| 1421 W KLEIN AV | 1,135 | 157,000 |
| 1430 W RAMSEY AV | 1,053 | 115,000 |
| 1444 W FOSTER AV | 1,070 | 134,900 |
| 1510 W CLAYTON CREST | 1,282 | 159,880 |
| 1521 W HOLMES AV | 936 | 123,000 |
| 1526 W VOGEL AV | 942 | 108,500 |
| 154 W WILBUR AV | 1,346 | 180,000 |
| 1564 W DENIS AV | 1,427 | 156,500 |
| 1566 W GRANGE AV | 1,326 | 100,000 |
| 1573 W BOLIVAR AV | 1,208 | 116,000 |
| 158 W TRIPOLI AV | 1,341 | 200,000 |
| 1647 W BOTTSFORD AV | 1,364 | 156,000 |
| 165 W VAN NORMAN A | 1,096 | 111,500 |
| 1700 W HOWARD AV | 2,988 | 170,000 |
| 1801 W BOLIVAR AV | 2,354 | 179,960 |
| 181 W SAVELAND AV | 1,373 | 148,000 |
| 1815 W GRANGE AV | 1,949 | 203,500 |
| 1816 W WHITAKER AV | 1, | |
| 1819 W HOLMES CT | 2, | |

Tabular

Can we use fewer numbers to describe the sale price?

| Mean | 147489.6 |
|---|---|
| Median | 148000 |
| Standard Deviation | 36391.91949 |
| Minimum | 87000 |
| Maximum | 245000 |

Numerical

Using a diagram?



Histogram of Home Sales Price
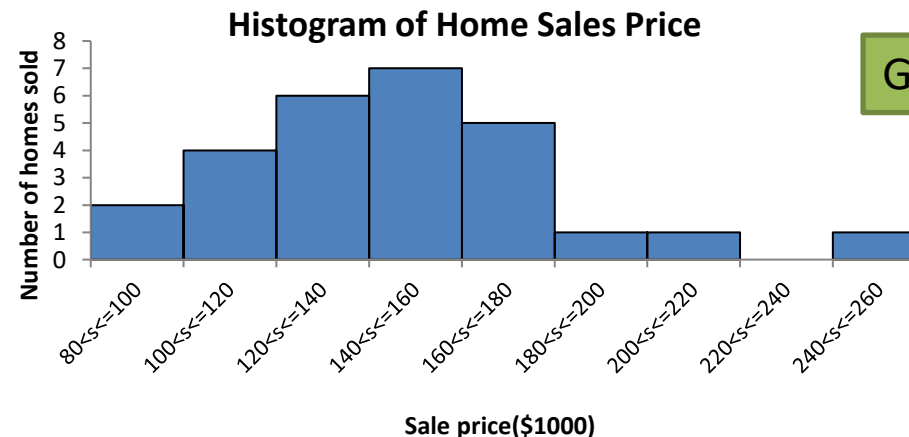
Graphical

# Descriptive Statistics Example: Per Capita State Income Taxes



https://www.caliper.com/featured-maps/maptitude-state-income-taxes-per-capita.html
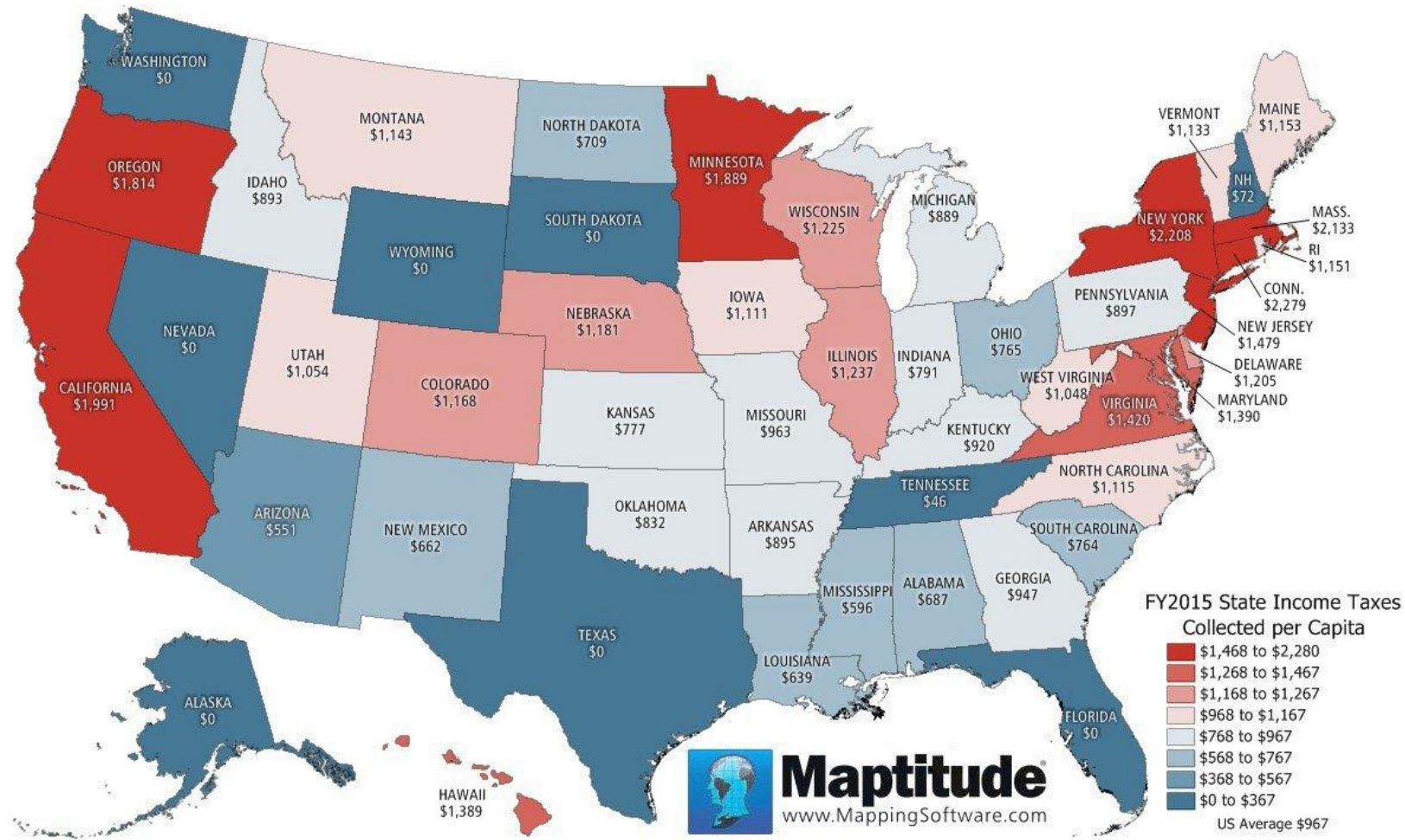
# Descriptive Statistics Example: Thematic map of unemployment rates



JUNE 2020 UNEMPLOYMENT RATES — SEASONALLY ADJUSTED

HIGHEST RATE: MASSACHUSETTS 17.4%
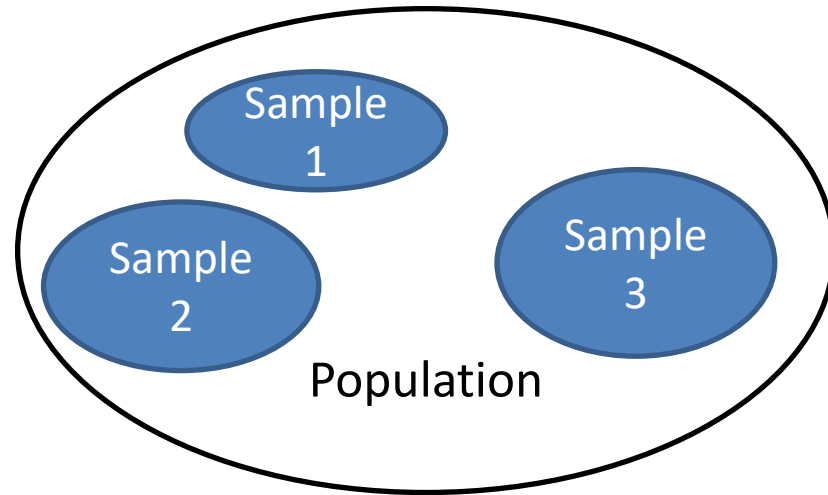NEW YORK 15.7%
NEW JERSEY 16.6%
LOWEST RATE: KENTUCKY 4.3%
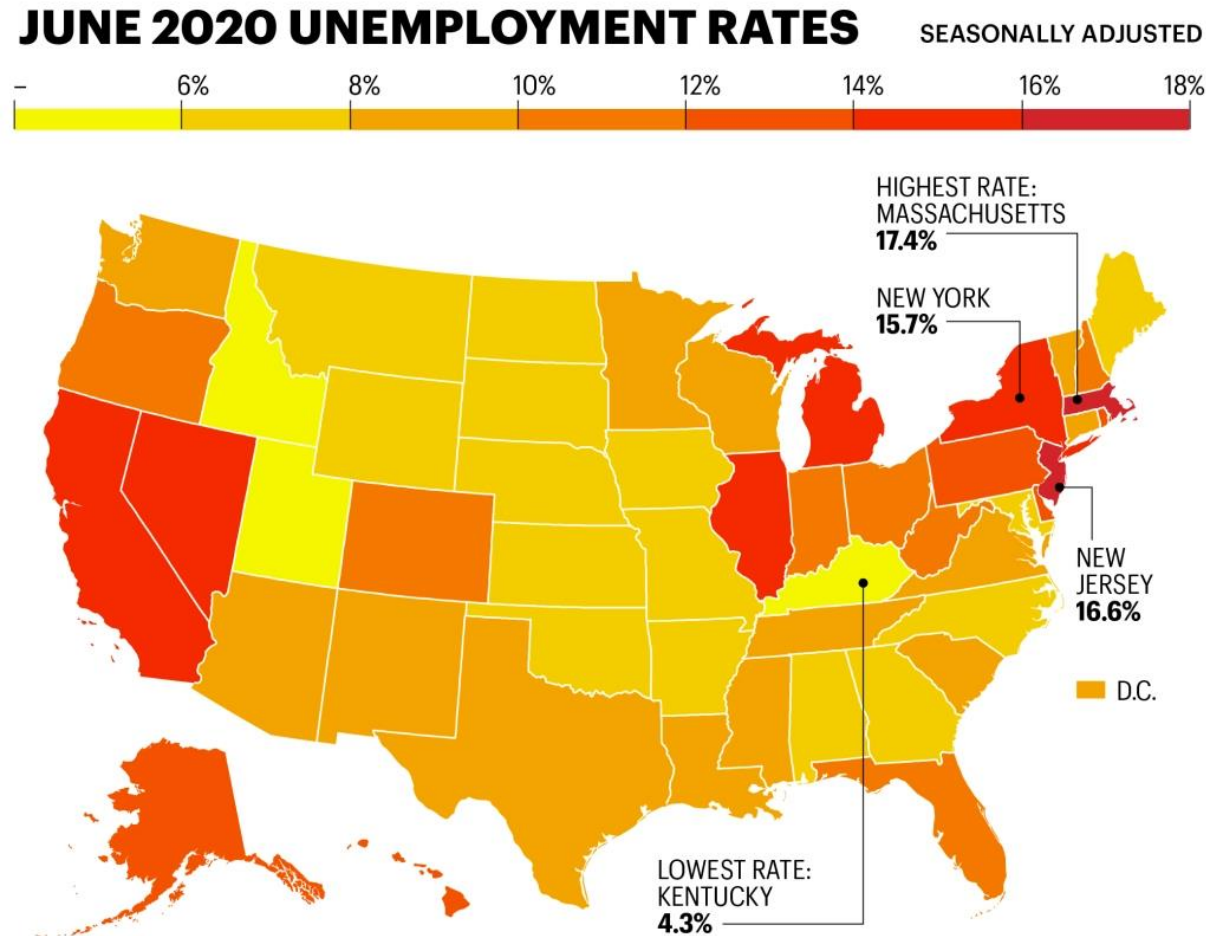
SOURCE: U.S. BUREAU OF LABOR STATISTICS

FORTUNE

# Population and Sample

- Population: all data
  - Many a times impossible to get (e.g., bodyweight of human population)

- Sample: a subset of population
  - Usually what we have
  - Cost effective, timely/fast



| Population | Sample |
|---|---|
| All trees in a forest | 50 trees from the forest |
| All students at USC | Students majored in Geography |
| All home sales in SC | 100 home sales in Columbia |

# Inferential Statistics

- Make statements about population based on a sample
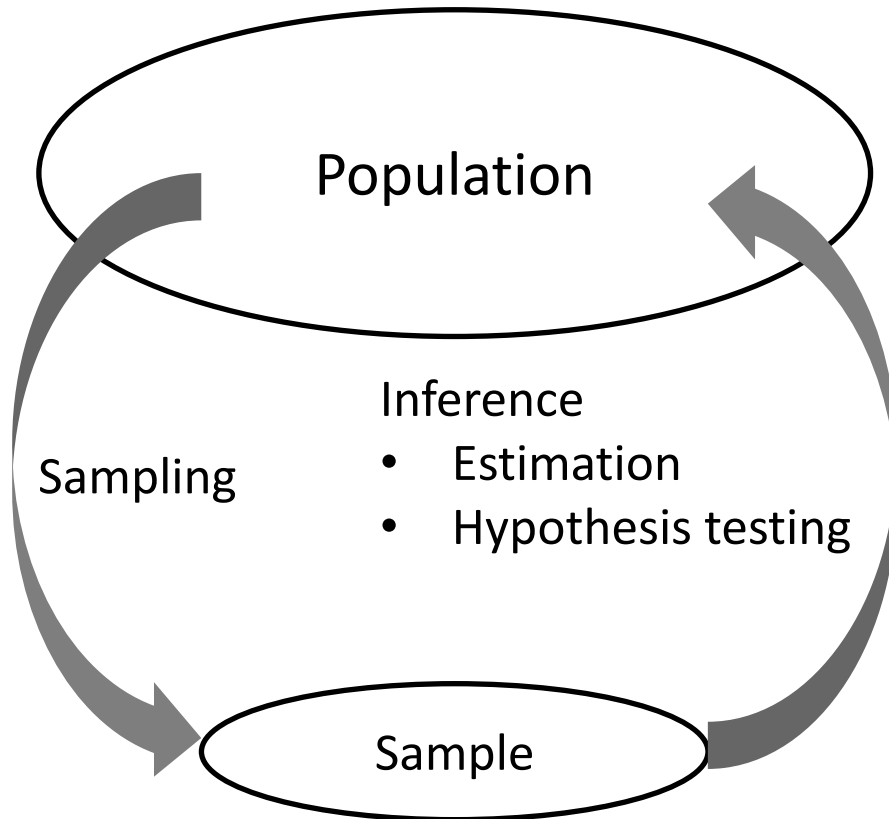


**Estimation**: estimate population properties based on a sample

**Hypothesis Testing**: determine if an observed pattern in the sample is likely to be true in the population

# Descriptive vs Inferential Statistics

- Descriptive Statistics: exploratory methods
  - Explore patterns and trends in the data
  - Suggest hypotheses
- Inferential Statistics: confirmatory methods
  - Confirm hypotheses
  - Either accept or reject a hypothesis
- Often used together

# Typical Steps of using Quantitative Methods

Study the spatial patterns of cancer cases in an area

- 1. Collect data
- 2. Display data
- 3. Discover patterns
- 4. Form hypothesis
  - Cancer cases are related to the distance from local power plant
- 5. Test hypothesis
  - Modeling
  - Statistical inference
- 6. Draw conclusions
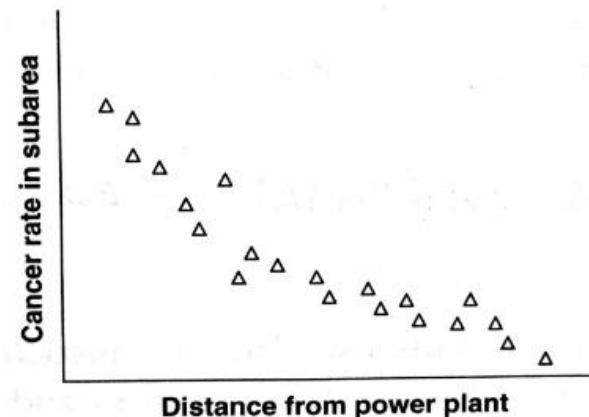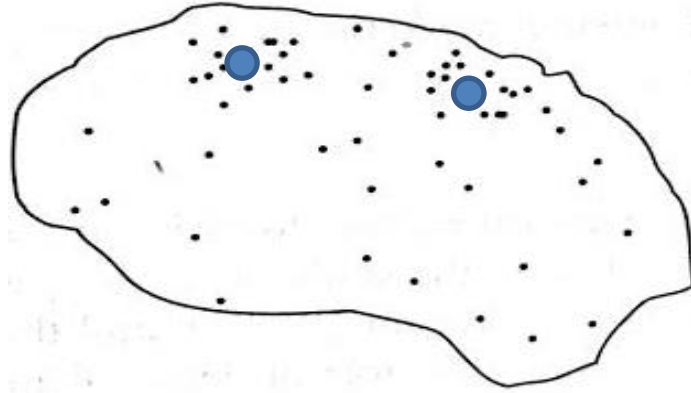
**The whole process is the core of the Scientific Method**

# Application Example

## Temporal and spatial changes in social vulnerability to natural hazards

Susan L. Cutter* and Christina Finch

Hazards and Vulnerability Research Institute, Department of Geography, University of South Carolina, Columbia, SC 29208

During the past four decades (1960–2000), the United States experienced major transformations in population size, development patterns, economic conditions, and social characteristics. These social, economic, and built-environment changes altered the American hazardscape in profound ways, with more people living in high-hazard areas than ever before. To improve emergency management, it is important to recognize the variability in the vulnerable populations exposed to hazards and to develop place-based emergency plans accordingly. The concept of social vulnerability identifies sensitive populations that may be less likely to respond to, cope with, and recover from a natural disaster. Social vulnerability is complex and dynamic, changing over space and through time. This paper presents empirical evidence on the spatial and temporal patterns in social vulnerability in the United States from 1960 to the present. Using counties as our study unit, we found that those components that consistently increased social vulnerability for all time periods were density (urban), race/ethnicity, and socioeconomic status. The spatial patterning of social vulnerability, although initially concentrated in certain geographic regions, has become more dispersed over time. The national trend shows a steady reduction in social vulnerability, but
(elderly and children), migration, and housing tenure (renter or owner). For example, the literature has cited many reasons why the elderly are more vulnerable in the event of a disaster: physical limitations that influence their inability or unwillingness to comply with mandatory evacuation orders; postdisaster psychological stress that impairs recovery and increases the need for additional social services; declining cognitive abilities to process hazard information necessitating specially targeted risk communication or warning messages; and fewer economic resources to repair damaged homes, especially by elderly residents on fixed incomes (15–18). Thus, the greater the proportion of elderly in a community, the more vulnerable it is and the longer it will take for the community to fully recover from the disaster's aftermath.

There have been some notable attempts to measure vulnerability. There are many national-level hazards and disasters indicator studies that incorporate social characteristics such as population numbers and distributions as a method for defining population exposures to a variety of hazard agents (19–25). Other studies incorporating vulnerability metrics focused on human-environmental systems at different subnational spatial scales: within India (26), U.S. watersheds (27), U.S. Great Plains

Cutter, S. L., & Finch, C. (2008). Temporal and spatial changes in social vulnerability to natural hazards. *Proceedings of the national academy of sciences*, *105*(7), 2301-2306.

# Quantitative methods used in this paper

- Descriptive statistics

- Global spatial statistics

- Local indicator of spatial autocorrelation
(LISA or the Local Moran's $I$)

- Correlation($r$)

- Simple linear regression

- Hypothesis test (F-test)

- Principal components analysis (PCA)

Cutter, S. L., & Finch, C. (2008). Temporal and spatial changes in social vulnerability to natural hazards. *Proceedings of the national academy of sciences*, *105*(7), 2301-2306.

# Data and Dataset

- Data: Measurable information
  -- e.g., population of a city

- Dataset: A collection of data consisting of observations, variables and data values

Home sales of Milwaukee, Wisconsin, 2014 (subset)

| Address | Sqft | Sale $ |
|---|---|---|
| 129 W BOLIVAR AV | 1,246 | 130,000 |
| 1336 W VAN NORMAN . | 1,143 | 87,000 |
| 137 W UNCAS AV | 1,521 | 150,000 |
| 1409 W WANDA AV | 1,059 | 168,000 |
| 1416 W HOWARD AV | 1,958 | 122,500 |
| 1420 W FOSTER AV | 1,208 | 136,000 |

- **Observations**: the elements under the study
- **Variables**: the properties of observation that measured
- **Data values**: the measurement of the properties of the observations

based on Steven Farber, Geographical Analysis , University of Utah

# Statistical Data Types

- Nominal, Ordinal, Interval, and Ratio
  - Indicate different levels of measurement/information
- Discrete and Continuous
  - Discrete data: can only take certain values(finite number of values), e.g., zip codes, number of sunny days in a year, COVID cases
  - Continuous data: can take any value(within a range, infinite number of values) e.g., height, temperature
- Quantitative and Qualitative
  - Quantitative: deals with numbers
  - Qualitative: deals with descriptions, e.g., color of eyes

# Nominal Data

- Also known as <u>Categorical</u> Data
- Observations are placed into a set of <u>unordered</u> categories
  - mutually exclusive
  - e.g., gender: {male, female},

    geographic location: {North, South, West, East}
  - Qualitative or quantitative?  →Qualitative
- Binary Data: when only two categories:, e.g., yes/no, success/failure

# Ordinal Data

- Observations are <u>ordered or ranked</u>
  - ordered by groups: weak order (e.g., low, average, high)
  - ordered by individual ranks: strong order (e.g., 1st, 2nd,3rd, … last)
  - Comparable: e.g. A is larger than B. But cannot tell how by how much.
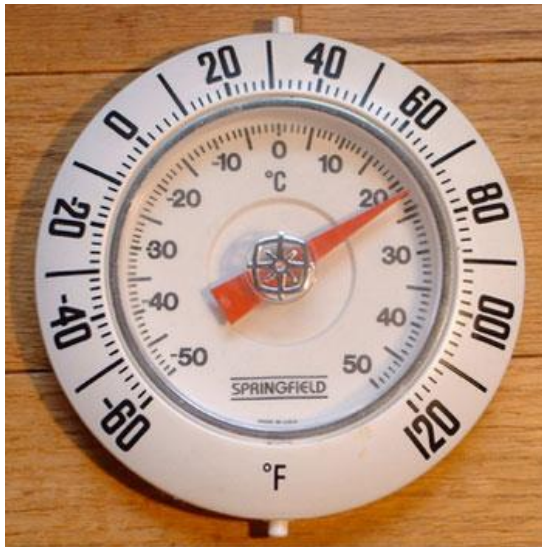  - Qualitative or quantitative?   →Qualitative

Country ranked by Area:

| Rank | Country | Area(sq km) |
|------|---------|-------------|
| 1 | Russia | 17,075,200 |
| 2 | Canada | 9,984,670 |
| 3 | United States | 9,826,630 |
| 4 | China | 9,596,960 |
| 5 | Brazil | 8,511,965 |

# Interval Data

- Not only indicates order, but also the difference between observations
- No natural zero point
- Can add/subtract, **cannot** multiply/divide



Example: temperature in Celsius or Fahrenheit

0 °C doesn't mean "no temperature"

30 °C is 15 °C warmer than 15 °C , but is not twice as warm as 15 °C  (15 °C  vs -15 °C )

Quantitative or Qualitative?    →Quantitative

# Ratio Data

- Interval data with a natural zero point
- +,-, x, ÷ is meaningful
- Most numerical data is ratio data
  - population in an area
  - height, weight …
  - temperature in Kelvin: absolute zero(no energy)

Quantitative or Qualitative?   →Quantitative

# Data Types Summary

Low      Level of information      High

| Properties | Nominal | Ordinal | Interval | Ratio |
|---|---|---|---|---|
| Distinctness (= ≠) | √ | √ | √ | √ |
| Order (< >) | | √ | √ | √ |
| Addition (+ -) | | | √ | √ |
| Multiplication (x ÷) | | | | √ |

qualitative discrete

quantitative continuous/discrete

# Data Type Exercise

**For the following measurements, identify their data types:**
– nominal, ordinal, interval or ratio?
– discrete or continuous?
– qualitative or quantitative?

• **Zip code**

• **Bronze, Silver, and Gold medals as awarded at the Olympics**

• **Temperature in Fahrenheit**

• **Number of patients in a hospital**

• **Angles as measured in degrees between 0° and 360°**

# Data Type Exercise

**For the following measurements, identify their data types:**
– nominal, ordinal, interval or ratio?
– discrete or continuous?
– qualitative or quantitative?

- **Zip code**

–nominal, discrete, qualitative

- **Bronze, Silver, and Gold medals as awarded at the Olympics**

–ordinal, discrete, qualitative

- **Temperature in Fahrenheit**

– interval, continuous, quantitative

- **Number of patients in a hospital**

–ratio, discrete, quantitative

- **Angles as measured in degrees between 0◦ and 360◦**

–ratio/interval , continuous, quantitative

# What is spatial data?

- ## Data with location
  - Absolute location: latitude, longitude on the Earth surface
  - Relative location:  e.g., distance to a highway

- ## Spatial data types
  - Point data: e.g.,  cities in a world map; event: flu cases, geo-tagged tweets
  - Areal/polygon data: e.g., US states, lakes
  - Line data : rivers, roads
  - Continuous/surface data: raster data, e.g., DEM(Digital Elevation Model, terrain)

# Spatial Data Examples


24 hours of Geo-tagged Tweets in New York
Point data


Areal(polygon) data


Line data
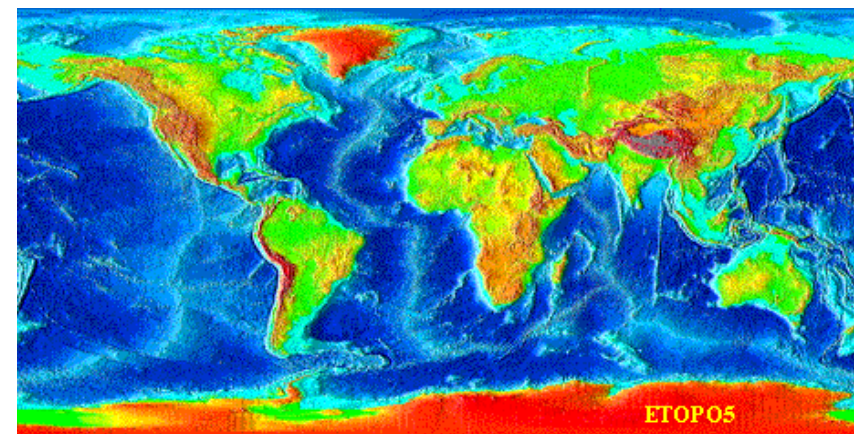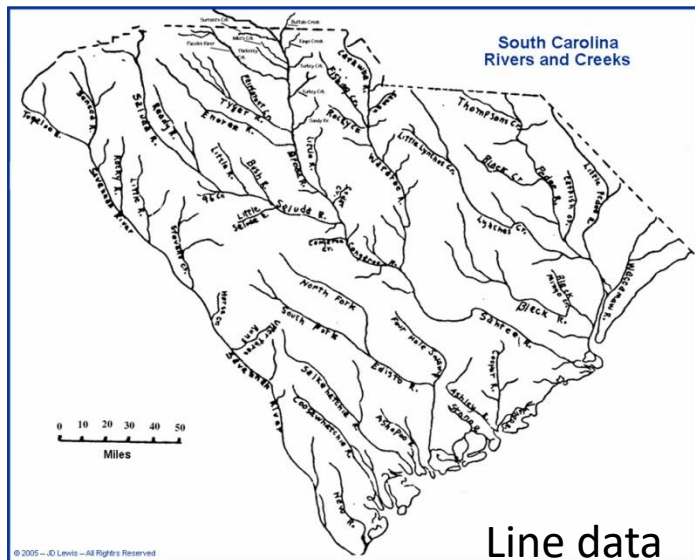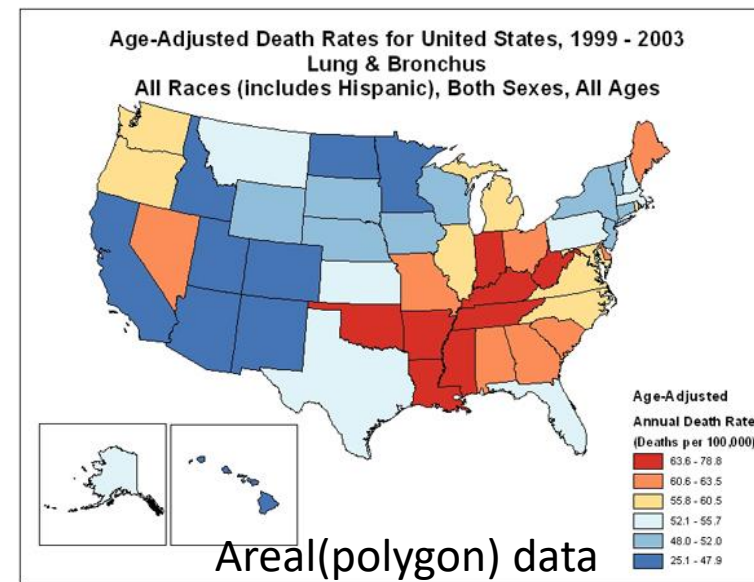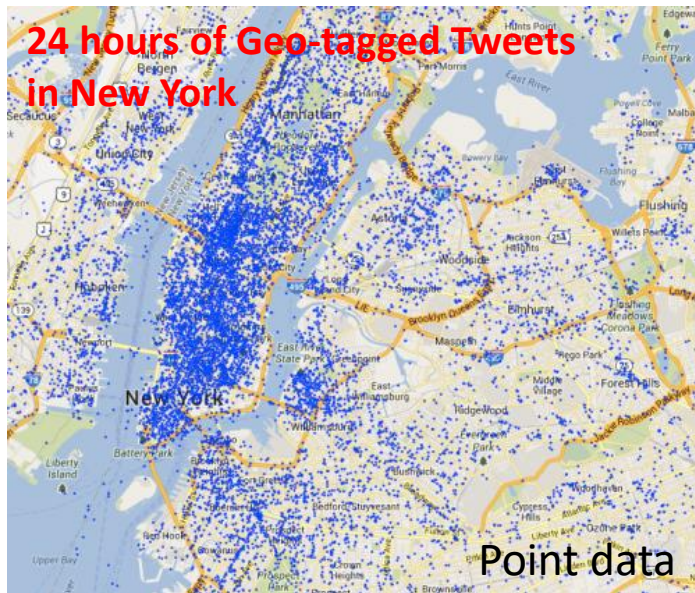

Continuous/Surface data

# Spatial Scales

- Point data vs. Areal data
- e.g., Cities



http://www.mapsofworld.com/usa/usa-capital-and-major-cities-map.html

Google Maps

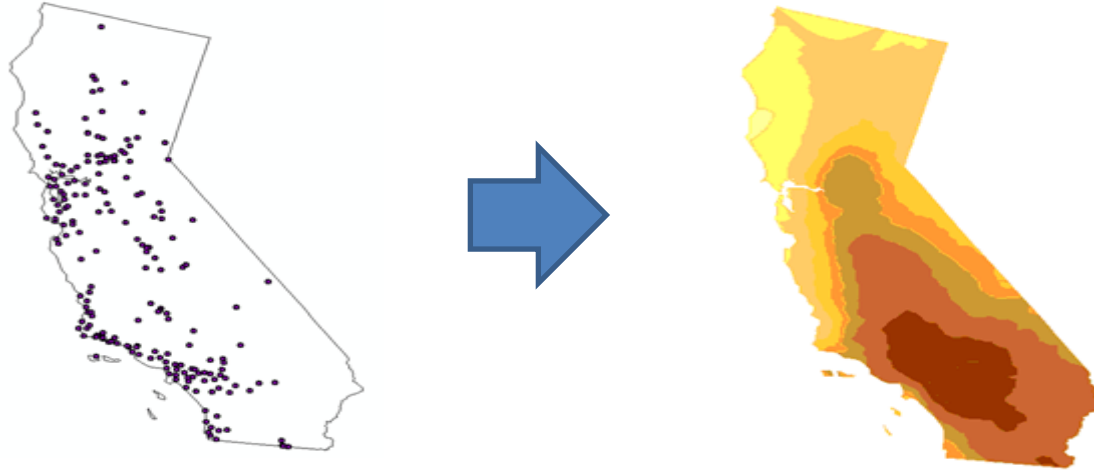Columbia as a point

Columbia as polygons

http://resources.arcgis.com/

# Data Conversion

- From point to area: using interpolation



- From area to point: using polygons' centroids



Generating points from state polygons

http://resources.arcgis.com/

http://giscommons.org/earth-and-map-preprocessing/