What are we working with? Data Abstractions

Week 4 Lecture A

IAT 814

Lyn Bartram

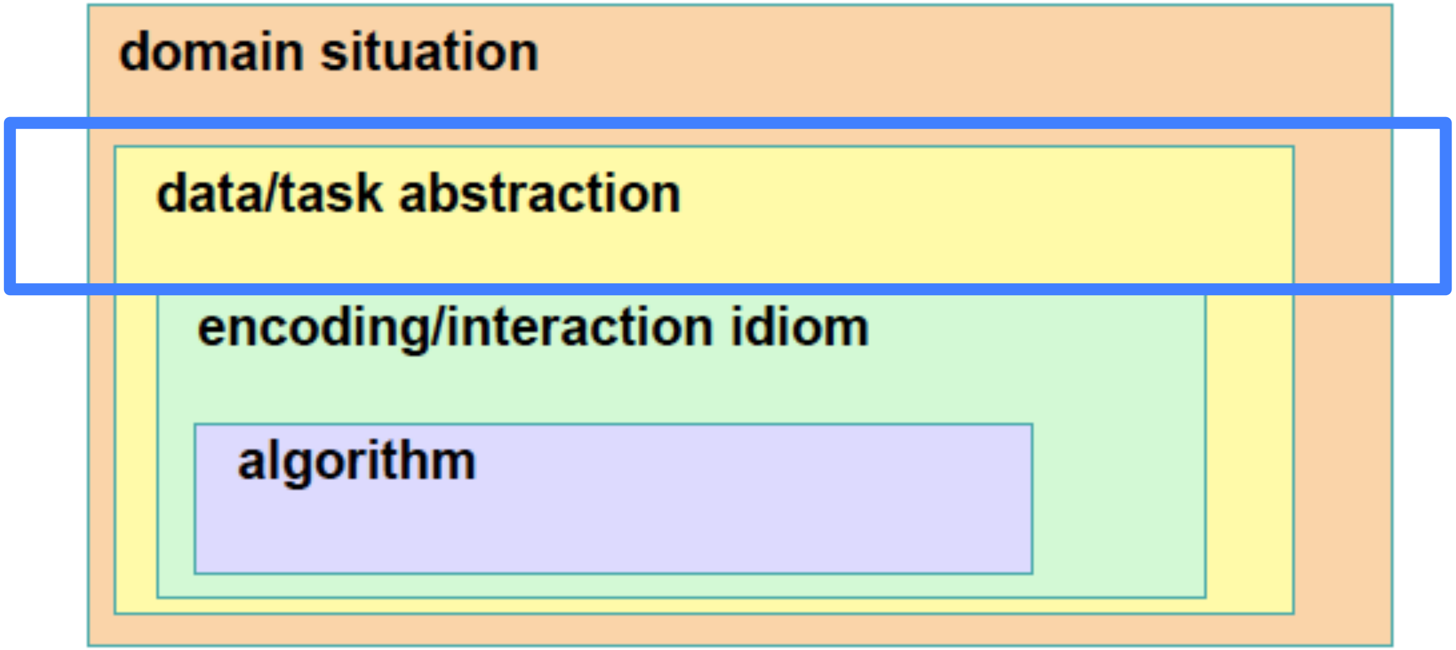
SFU

SCHOOL OF INTERACTIVE ARTS + TECHNOLOGY

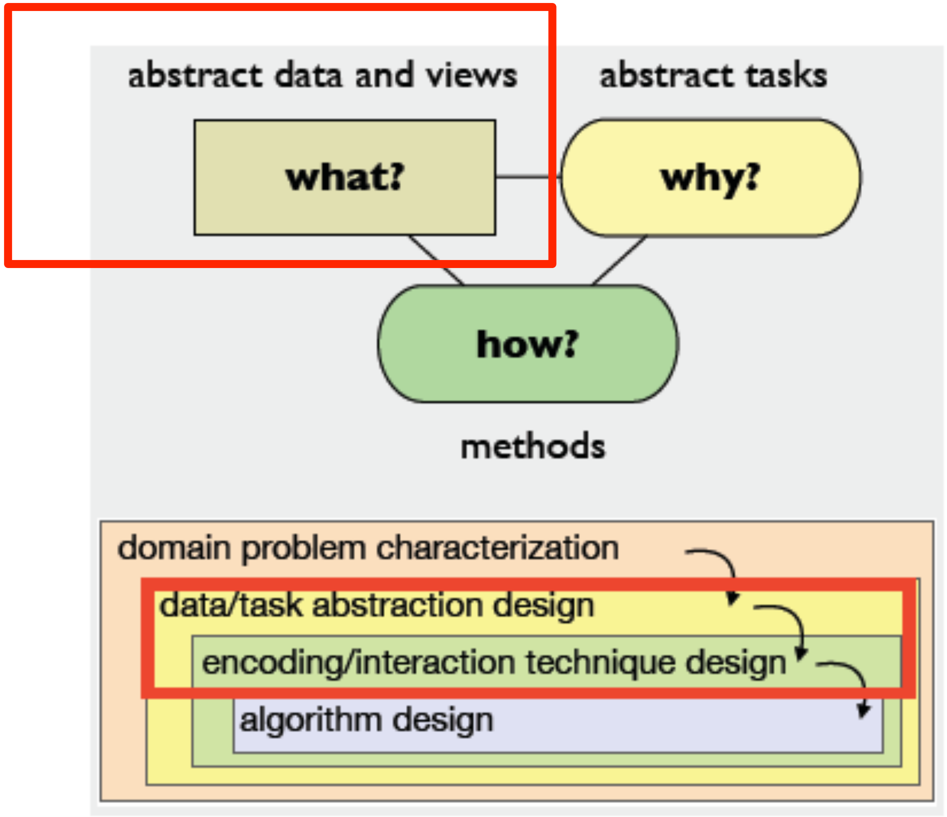
Munzner's What-Why-How

- What are we working with?
  - DATA abstractions, statistical methods
- Why are we doing it?
  - Task abstractions
- HOW are we doing it?

4 stages of visualization design



A Framework for Analysis (Munzner)



But what data ? Meaning?

- 14, 2.6, 30, 30, 15, 100001
- Basil, 7, 5, East-West

- Semantics:
  - what the data mean in the world
- Data type, model
  - Structural and mathematical representation
  - What you can derive from it (do with it)

Data forms

- Data comes in many different forms
- Typically, not in the way you want it
- How is it stored (in the raw)?

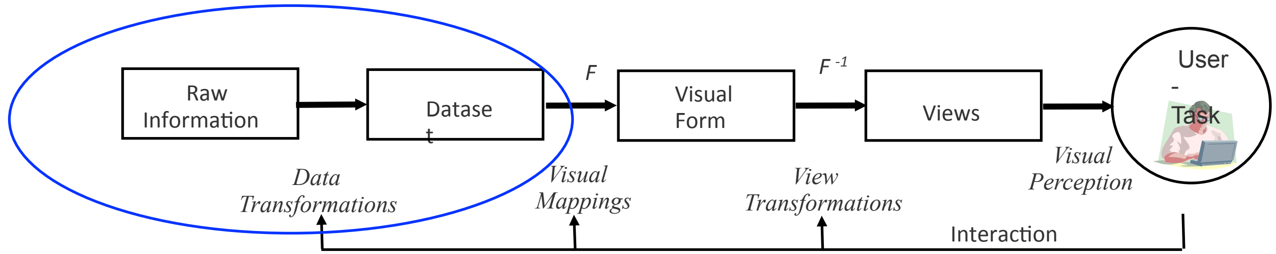
- Data set
- Data type

- Cars
  - make
  - model
  - year
  - miles per gallon
  - cost
  - number of cylinders
  - weights
  - ...

But WHAT numbers?????

- Data Models
- Types
- Metadata
- Aggregates
- Descriptive Statistics
- Distribution
- Clusters

Data models



- We take raw data and transform it into a form that is more workable
- Main idea: build a *model* of a *dataset*
  - Individual items are called *cases* or *records*
  - Items have *attributes* : an attribute is a *value* of a *variable or factor*
  - In vis terms, a *dimension*
- A model is an abstraction of the world, comprising abstractions

Vis Data types

- items
- Attributes
- Links
- Grids
- Positions

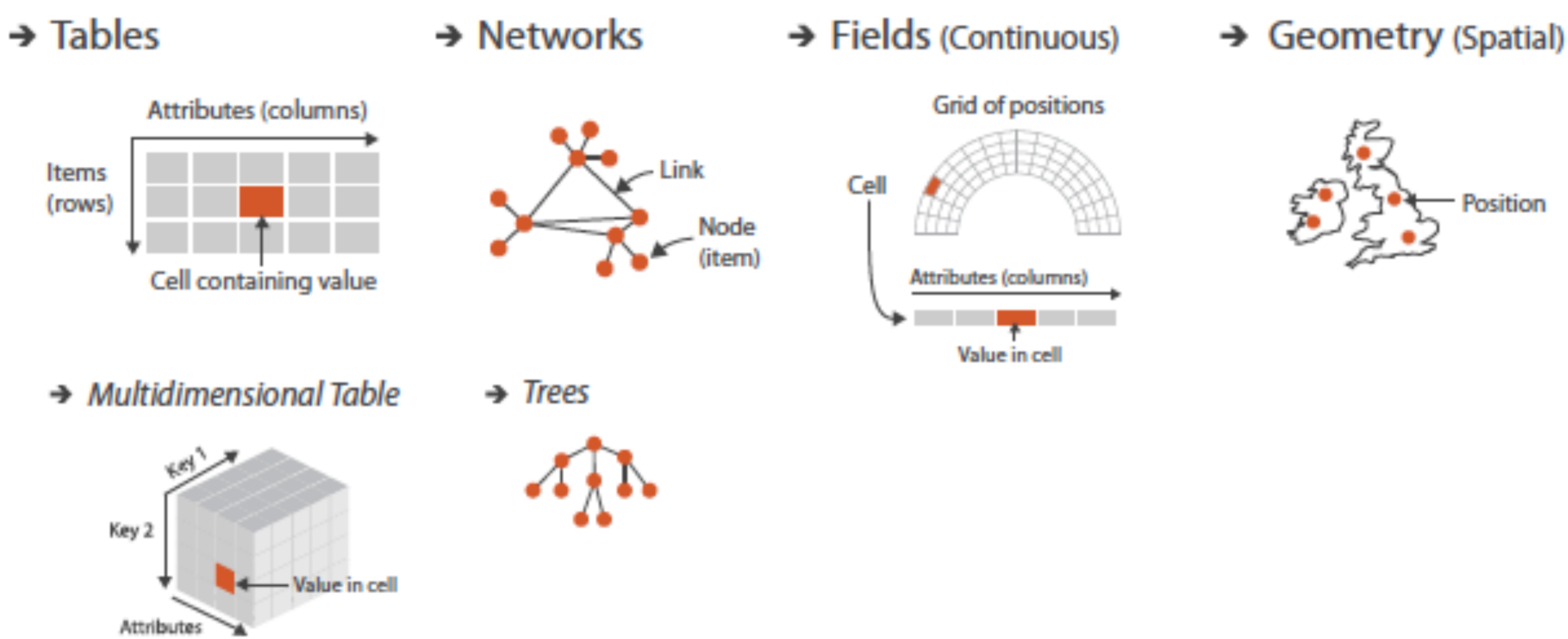
- “Type” here is with respect to what the data refers to
- These are essential components of vis data models

Data types

- Attribute
- Item
- Link
- Position
- grid

- Property that can be measured
- Individual entity (case)
- Relationship between items
- Spatial / location
- How to sample continuous data

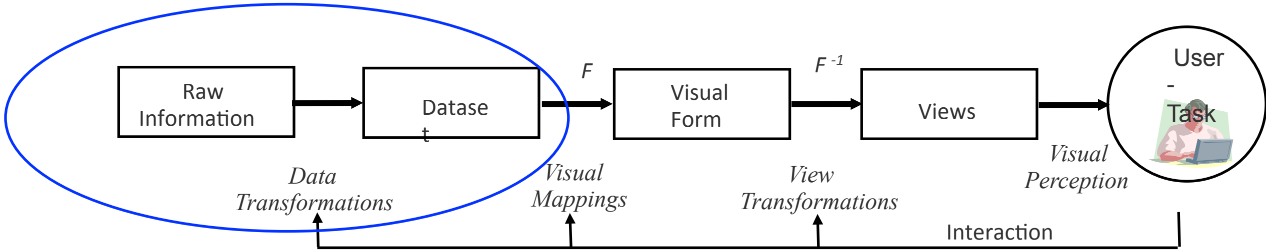
Dataset types [Munzner 2014]



Other multiple combinations

- Group items together
- Set == unordered group
- List == Ordered set
- Cluster == group based on similarity of some attribute
- Path == ordered set of segments in a network or tree

Tables



Dimensions

	Attr1	Attr2	Attr3
Item1	Cell11	Cell21	Cell31
Item2	Cell12	Cell22	Value32
Item3	Cell13	Cell23	Cell33

Example: Student Data

Item

Name	Mary	Tom	Louise
Student Num	65432101	98765651	89846251
Age	20	22	19
Entered SFU	Sep 2006	Jan 2004	Sep 2005
GPA	4.0	2.3	3.04

Example: Student Data

Name	Mary	Tom	Louise
Student Num	65432101	98765651	89846251
Age	20	22	19
Entered SFU	Sep 2006	Jan 2004	Sep 2005
GPA	4.0	2.3	3.04

Attribute/Dimension

Example: What kinds of data?

Name	Mary	Tom	Louise
Student Num	65432101	98765651	89846251
Age	20	22	19
Entered SFU	Sep 2006	Jan 2004	Sep 2005
GPA	4.0	2.3	3.04

How many attributes/dimensions?

- Data sets of dimensions 1, 2, 3 are common

- Number of attributes per class
- 1 - **Univariate** data
- 2 - Bivariate data
- 3 - Trivariate data
- >3 - Hypervariate data
  - These are the fun and interesting ones! But hard!

Attribute Types : categorical vs ordered

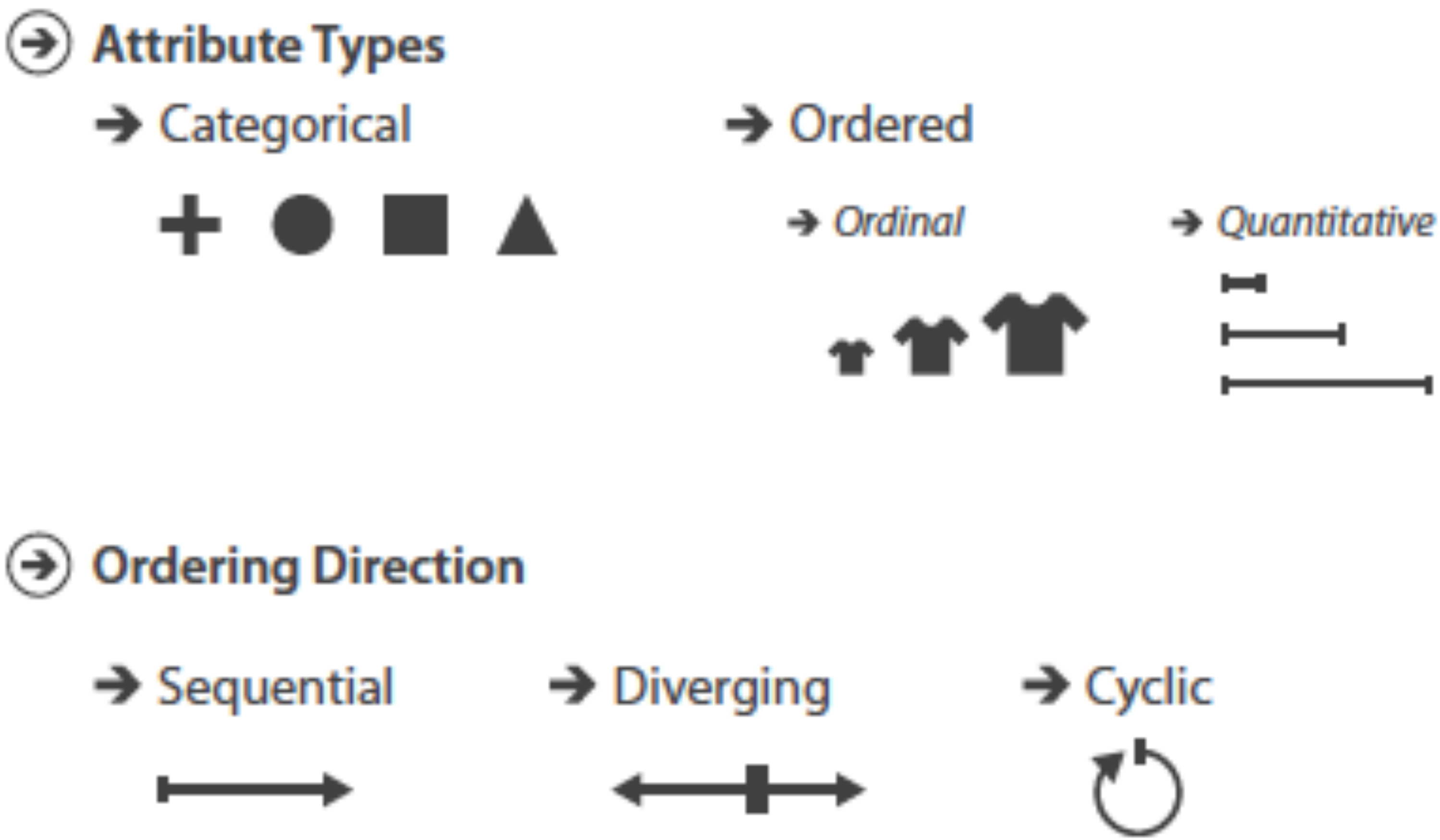
- **Nominal**: categorical,(equal or not equal to other values)
  - Example: gender, Student Number
  - No concept of relative relation other than inclusion in the set

- **Ordinal** : sequential (obeys < > relation, ordered set)
  - Example: Size of car, speed settings on road
  - Example: mild, medium, hot, suicide
  - Distance is not uniform

Ordinal Data Types : Quantitative

- **Interval** : Relative measurements, no fixed zero point.
  - Rank order among variables is explicit with an equal distance between points in the data set
  - days in a week
  - Can judge distance but not perform arithmetic
- **Ratio**: zero is fixed
  - Can say “twice as much as”
  - Example: account balance

Attributes


Attribute Types
Categorical
Ordered
Ordinal
Quantitative
Ordering Direction
Sequential
Diverging
Cyclic

Attribute characteristics:

- Continuous
  - Data can take any value within the range
  - Number grade (92.75%)
- Discrete: data can take only certain values
  - Example: number of students in a class (no half students)
  - Letter grade (A+)
- We can convert between these using **clustering** or **thresholds**

Dimensions/attributes: recap

- **Data Dimensions** are classified as:
  - **Quantitative** i.e. numerical
    - **Continuous** (e.g. pH of a sample, patient cholesterol levels)
    - **Discrete** (e.g. number of bacteria colonies in a culture)
  - **Categorical**
    - **Nominal** (e.g. gender, blood group)
    - **Ordinal** (ranked e.g. mild, moderate or severe illness). Often ordinal variables are re-coded to be **quantitative**.

Variables: the values of the dimension

- Variables are classified as:
  - **Dependent.** Variable of primary interest (e.g. blood pressure in an antihypertensive drug trial). What we want to know about.
  - **Independent/Predictor**
    - called a **Factor** when controlled by experimenter.
    - **A dimension** in visualization
  - **Random :** cannot be controlled or predicted
- These are experimental terms: how do they apply to analysis?

Data vs Conceptual Models

- data model: mathematical abstraction
  - set with operations
  - e.g. integers or floats with +,*
- conceptual model: mental construction
  - includes semantics, support data
  - e.g. navigating through city using landmarks
- conceptual model motivates derived data
- [Hanrahan, graphics.stanford.edu/courses/ cs448b-04-winter/lectures/encoding/ walk005.html]
- [Rethinking Visualization: A High-Level Taxonomy. Melanie Tory and Torsten M oller, Proc. InfoVis 2004, pp. 151-158.

Derived attributes can depend on task

data model

- 17, 25, -4, 28.6
- (floats)

conceptual model

- temperature

transform to type

- making toast
  - burned vs. not burned (N)
- classifying showers
  - hot, warm, cold (O)
- finding anamolies in local weather patterns
  - Continuous (Q)

Issues to consider

- Spatial/temporal frequencies on data
- Missing values
  - Interpolate
  - Show as missing explicitly?
  - Ignore?
- Special values
  - Of particular interest to visualize
  - Thresholds, ratio scales (consider sea level relative values)

But wait…. There's more !

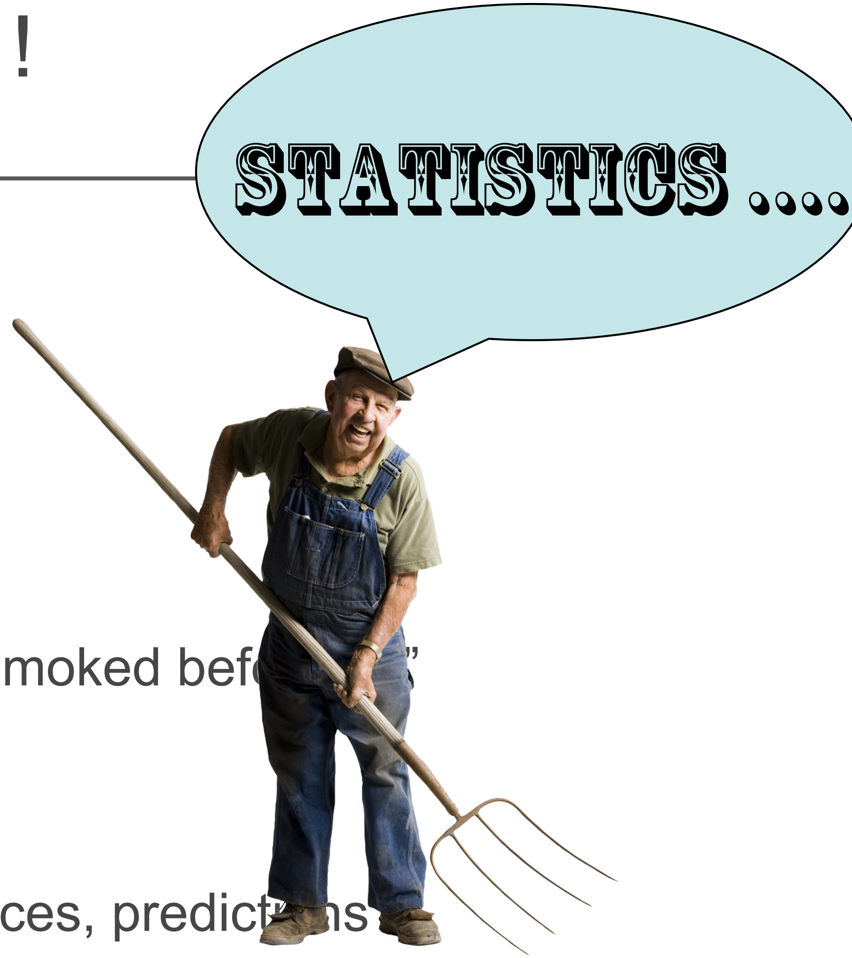
- Raw data
- Metadata
  - Data about the data
- Frequency data
  - "more than half the respondents smoked before 16"
- Derived data
  - Summaries, observations, inferences, predictions
  - "the odds of you getting ill from this pizza were 5 to 1"

Mary	Tom	Louise
65432101	98765651	89846251
20	22	19
Sep 2006	Jan 2004	Sep 2005
4.0	2.3	3.04

Metadata

- Descriptive information about the data
- Might be something as simple as the type of a variable, or could be more complex
  - For times when the table itself just isn't enough
  - Example: if variable1 is "l", then variable3 can only be 3, 7 or 16
- Missing values, uncertainty or importance are all examples of metadata

But wait…. There’s more !



- Raw data
- Metadata
  - Data about the data
- **Frequency data**
  - “more than half the respondents smoked bef ”
  - clustering
- **Derived data**
  - Summaries, observations, inferences, predict ns
  - “the odds of you getting ill from this pizza were 5 to 1”

Primary types of data analysis

- Qualitative
- **Descriptive**. Used to describe the distribution of a single variable or the relationship between two nominal variables (mean, frequencies, cross-tabulation)
- **Inferential** (Used to establish relationships among variables; assumes random sampling and a normal distribution)
- **Nonparametric** (Used to establish causation for small samples or data sets that are not normally distributed)

Descriptive Statistics: Univariate

- Range
- Min/Max
- Average
- Median
- Mode

Distribution Statistics

- Variance
- Error
- Standard Deviation
- Histograms and Normal Distributions

Frequency statistics

Most basic type of descriptive statistic

- An (Empirical) Frequency Distribution for a continuous variable presents the counts of observations grouped within pre-specified classes or groups
- A Relative Frequency Distribution presents the corresponding proportions of observations within the classes
- Visualizations: barcharts, histograms

Example use of frequency

- 40% of respondents are male.
- The mean level of income was $35,000
- 40% of all female voters cast their vote for Arnold compared to 52% of the male voters.

Range, Min, Max

- The Range
  - Difference between minimum and maximum values in a data set
  - Larger range usually (but not always) indicates a large spread or deviation in the values of the data set.

(73, 66, 69, 67, 49, 60, 81, 71, 78, 62, 53, 87, 74, 65, 74, 50, 85, 45, 63, 100)

Average = measure of centrality

- Measures of location indicate where on the number line the data are to be found. Common measures of location are:

- (i) the Arithmetic Mean,
- (ii) the Median, and
- (iii) the Mode

The "Average" ???

- The Average (Mean)
  - Sum of all values divided by the number of values in the data set.
  - One measure of central **location** in the data set.

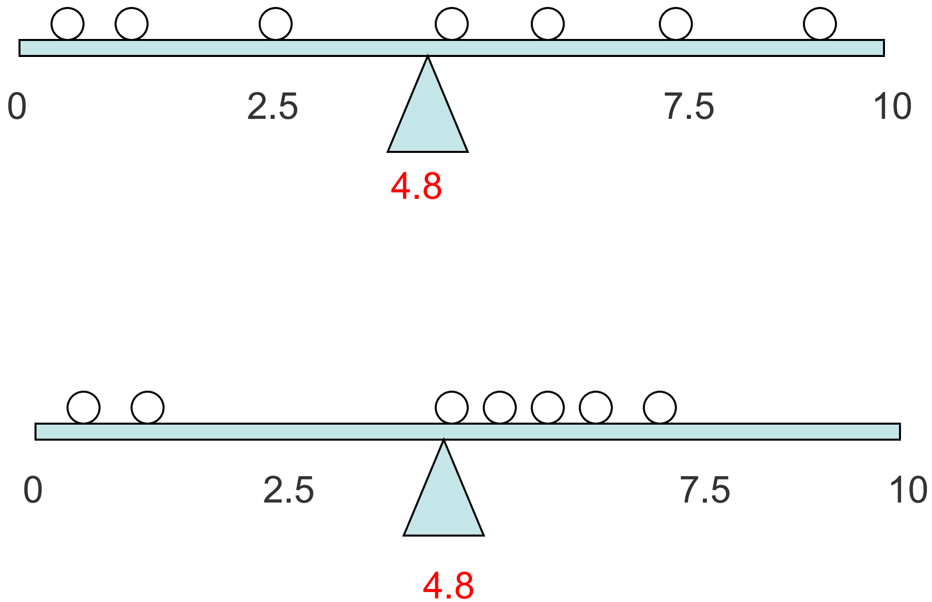
Average =

Average=(73+66+69+67+49+60+81+71+78+62+53+87+74+65+74+50+85+45+63+100)/20 = 68.6

The tyranny of the mean

- When might you not want to use the mean?

The mean is vulnerable to problems



The data may or may not be symmetrical around its average value

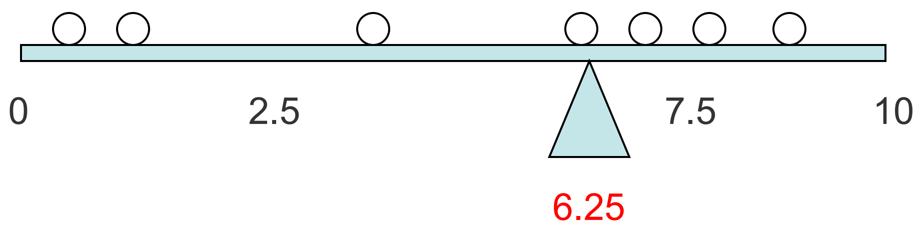
- The Median
  - The middle value in a sorted data set. Half the values are greater and half are less than the median.
  - Another measure of central location in the data set.

(45, 49, 50, 53, 60, 62, 63, 65, 66, 67, 69, 71, 73, 74, 74, 78, 81, 85, 87, 100)

Median: 68

(1, 2, 4, 7, 8, 9, 9)

- The Median
  - May or may not be close to the mean.
  - Combination of mean and median are used to define the skewness of a distribution.

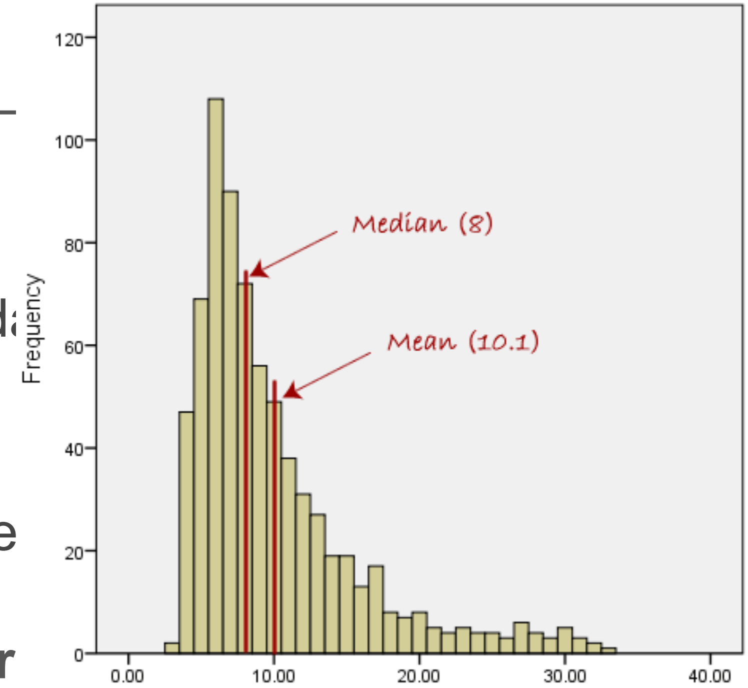


The Mode

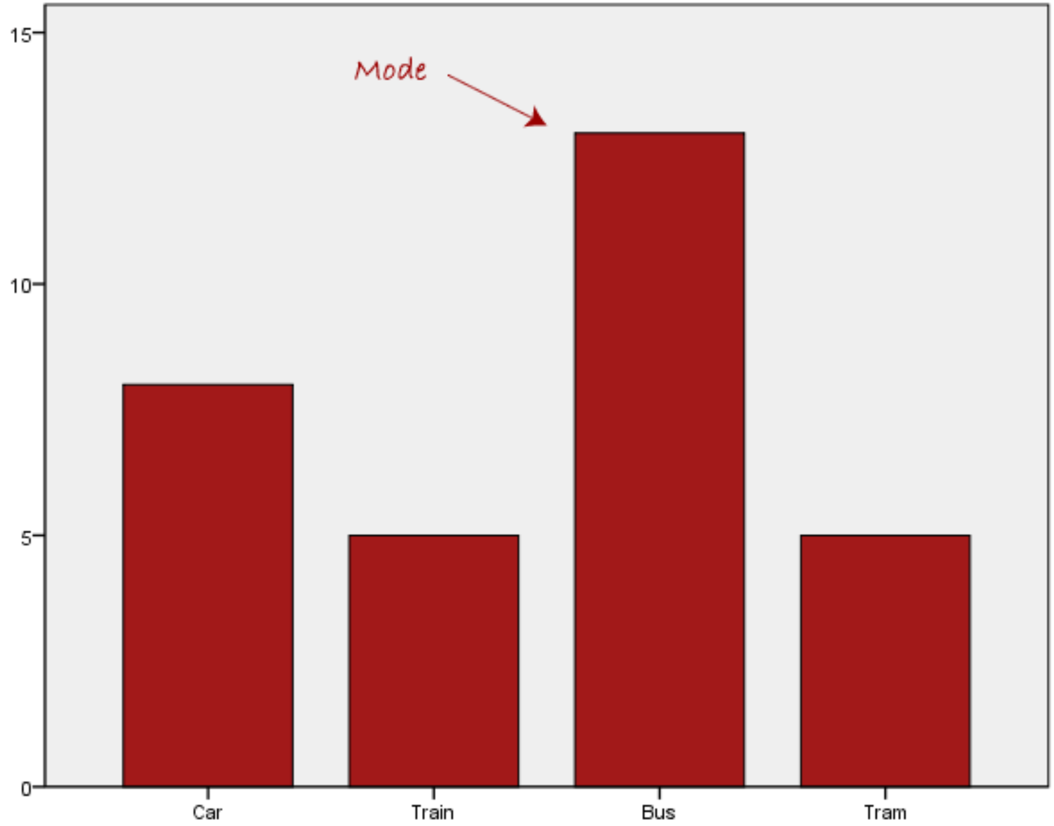
- The Mode
  - The most frequent occurring value.
  - Another measure of central location in the data set.
  - (45, 49, 50, 53, 60, 62, 63, 65, 66, 67, 69, 71, 73, 74, 74, 78, 81, 85, 87, 100)
  - Mode: 74
  - Generally not all that meaningful unless a larger percentage of the values are the same number

When do we use what?

- Dependent on how the data are **distributed**
  - Note if mean=median=mode then the da
    are said to be symmetrical
- Rule of thumb:
  - use mean if data are normally distribute
    and variance is within constraints
  - Use median to reduce effects of **outlier**



Mode for categorical frequency



Summary

Type of Variable	Best measure of central tendency
Nominal	Mode
Ordinal	Median
Interval/Ratio (not skewed)	Mean
Interval/Ratio (skewed)	Median

http://statistics.laerd.com/statistical-guides/measures-central-tendency-mean-mode-median.php

Data distribution

- Measures of dispersion characterise how spread out the distribution is, i.e., how variable the data are.
- Commonly used measures of dispersion include:
  1. Range
  2. Variance & Standard deviation
  3. Coefficient of Variation (or relative standard deviation)
  4. Inter-quartile range

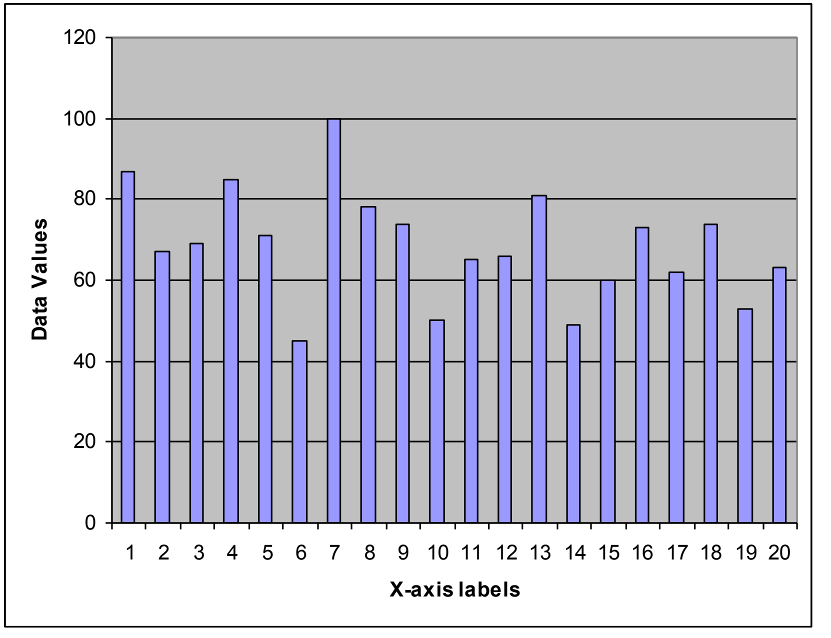
Measures of variance

- Variance
  - One measure of dispersion (deviation from the mean) of a data set. The larger the variance, the greater is the average deviation of each datum from the average value
- Standard Deviation
  - the average deviation from the mean of a data set.

- Variance and SD are critical in analysing your data distribution and determining how “meaningful” is the chosen average

Histograms and distribution

- We can't really tell much about this data set
- Even Min and Max are hard to see

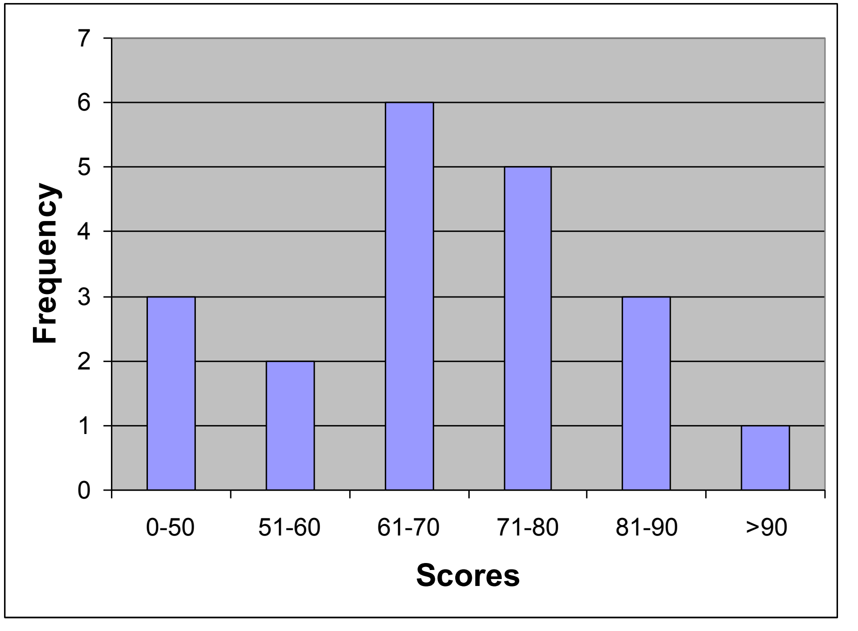


The data can be presented such that more statistical info can be estimated from the chart (average, standard deviation).

Plotting the distribution

- Determine a frequency table (bins)
- A histogram is a column chart of the frequencies

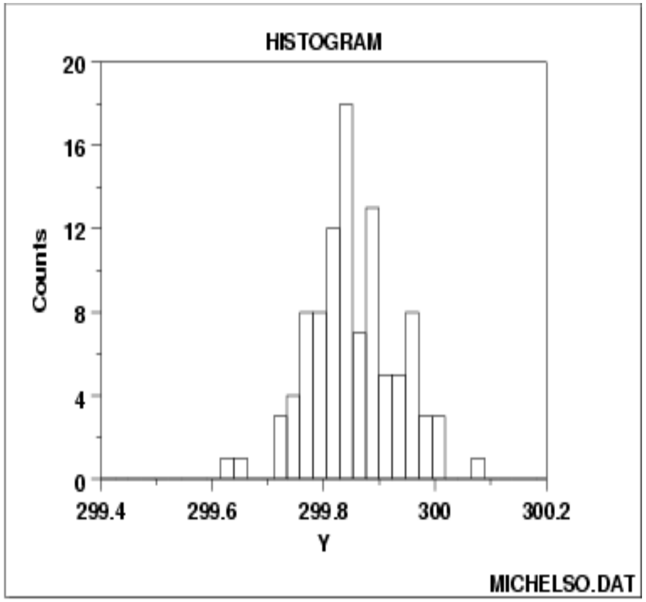
Category Labels	Frequency
0-50	3
51-60	2
61-70	6
71-80	5
81-90	3
>90	1



Histogram

- Most common form: split data range into equal-sized bins Then for each bin, count the number of points from the data set that fall into the bin.
  - Vertical axis: Frequency (i.e., counts for each bin)
  - Horizontal axis: Response variable

- The histogram graphically shows the following:
1. center (i.e., the location) of the data;
2. spread (i.e., the scale) of the data;
3. skewness of the data;
4. presence of outliers; and
5. presence of multiple modes in the data.



Issues with Histograms

- For small data sets, histograms can be misleading. Small changes in the data or to the bucket boundaries can result in very different histograms.
- Interactive bin-width example (online applet)
  - http://www.stat.sc.edu/~west/javahtml/Histogram.html

- For large data sets, histograms can be quite effective at illustrating general properties of the distribution.

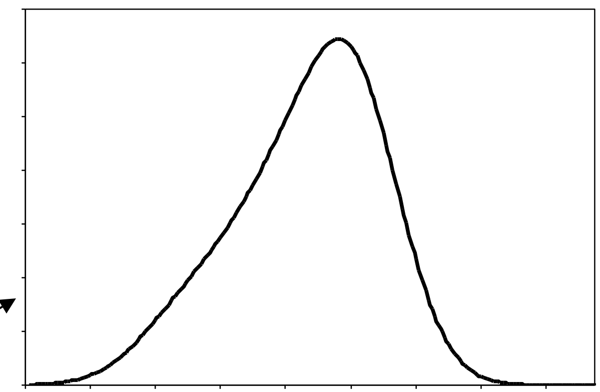
- Histograms effectively only work with 1 variable at a time
  - Difficult to extend to 2 dimensions, not possible for >2
  - So histograms tell us nothing about the relationships among variables

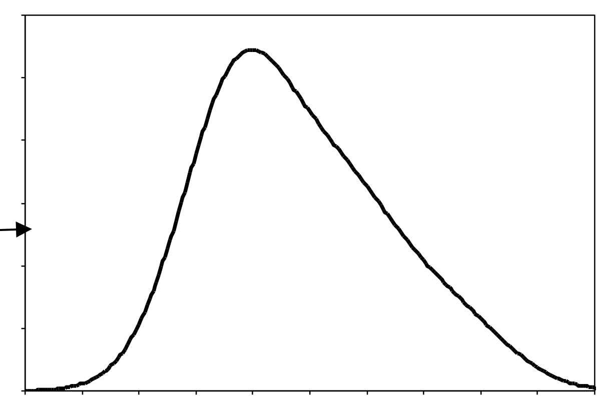
Normal and Skewed Distributions

- When data are skewed, the mean and SD can be misleading
- Skewness

  sk= 3(mean-median)/SD

  If sk>|1| then distribution is non-symetrical
- Negatively skewed
  - Mean<Median
  - Sk is negative
- Positively Skewed
  - Mean>Median
  - Sk is positive





Inter-quartile range

- The Median divides a distribution into two halves.

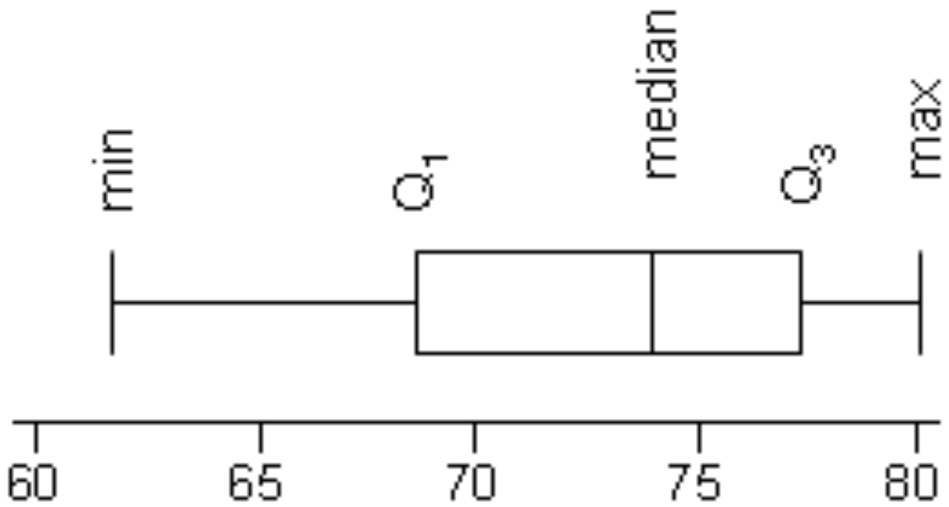
- The **first** and **third** quartiles (denoted $\mathbf{Q_1}$ and $\mathbf{Q_3}$) are defined as follows:
  - 25% of the data lie below Q_1 (and 75% is above Q_1),
  - 25% of the data lie above Q_3 (and 75% is below Q_3)

- The **inter-quartile range (IQR)** is the difference between the first and third quartiles, i.e.
  IQR = $\mathbf{Q_3 - Q_1}$

Box-plots

- A box-plot is a visual description of the distribution based on
  - Minimum
  - Q1
  - Median
  - Q3
  - Maximum
- Useful for comparing large sets of data

Example 1: Box-plot



Outliers

- An **outlier** is an datumwhich does not appear to belong with the other data
- Outliers can arise because of a measurement or recording error or because of equipment failure during an experiment, etc.
- An outlier might be indicative of a sub-population, e.g. an abnormally low or high value in a medical test could indicate presence of an illness in the patient.

Outlier Boxplot

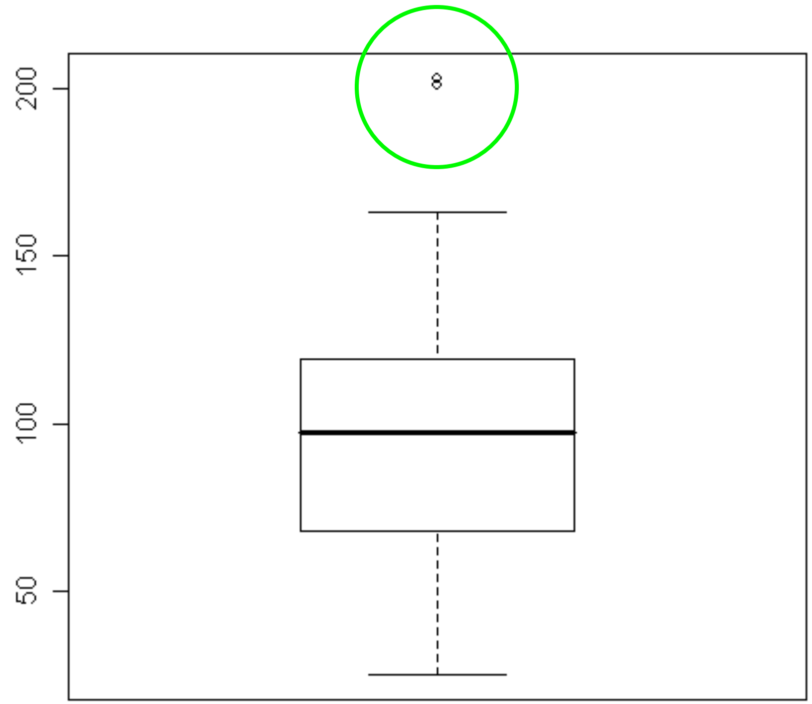
- Re-define the upper and lower limits of the boxplots (the whisker lines) as:

  Lower limit = Q_1-1.5×IQR, and

  Upper limit = Q_3+1.5×IQR

- Note that the lines may not go as far as these limits

- If a data point is < lower limit or > upper limit, the data point is considered to be an outlier.



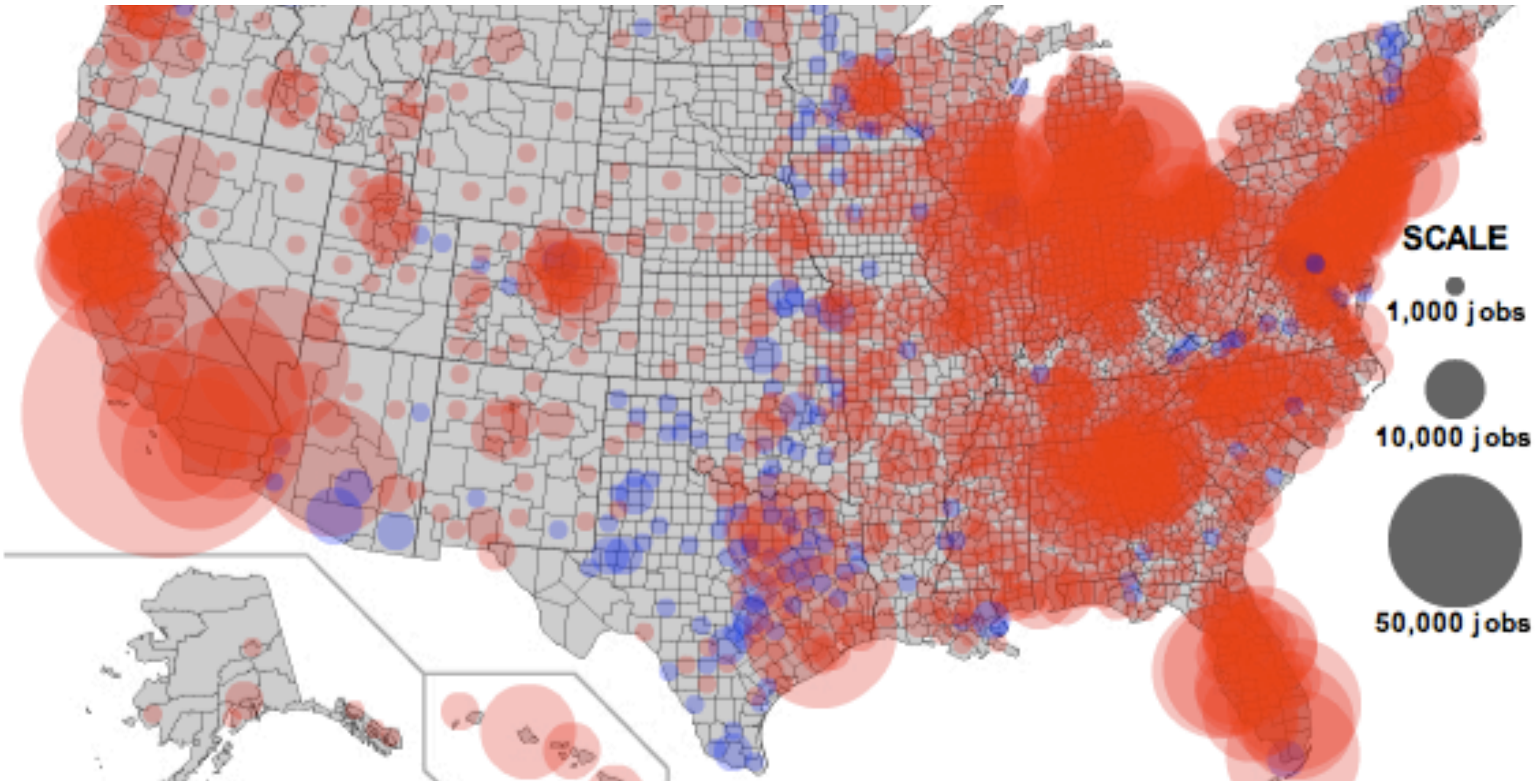
Recap: Distribution is important for Aggregation

- Visualization helps us see relations – or the trends of them - as visual patterns

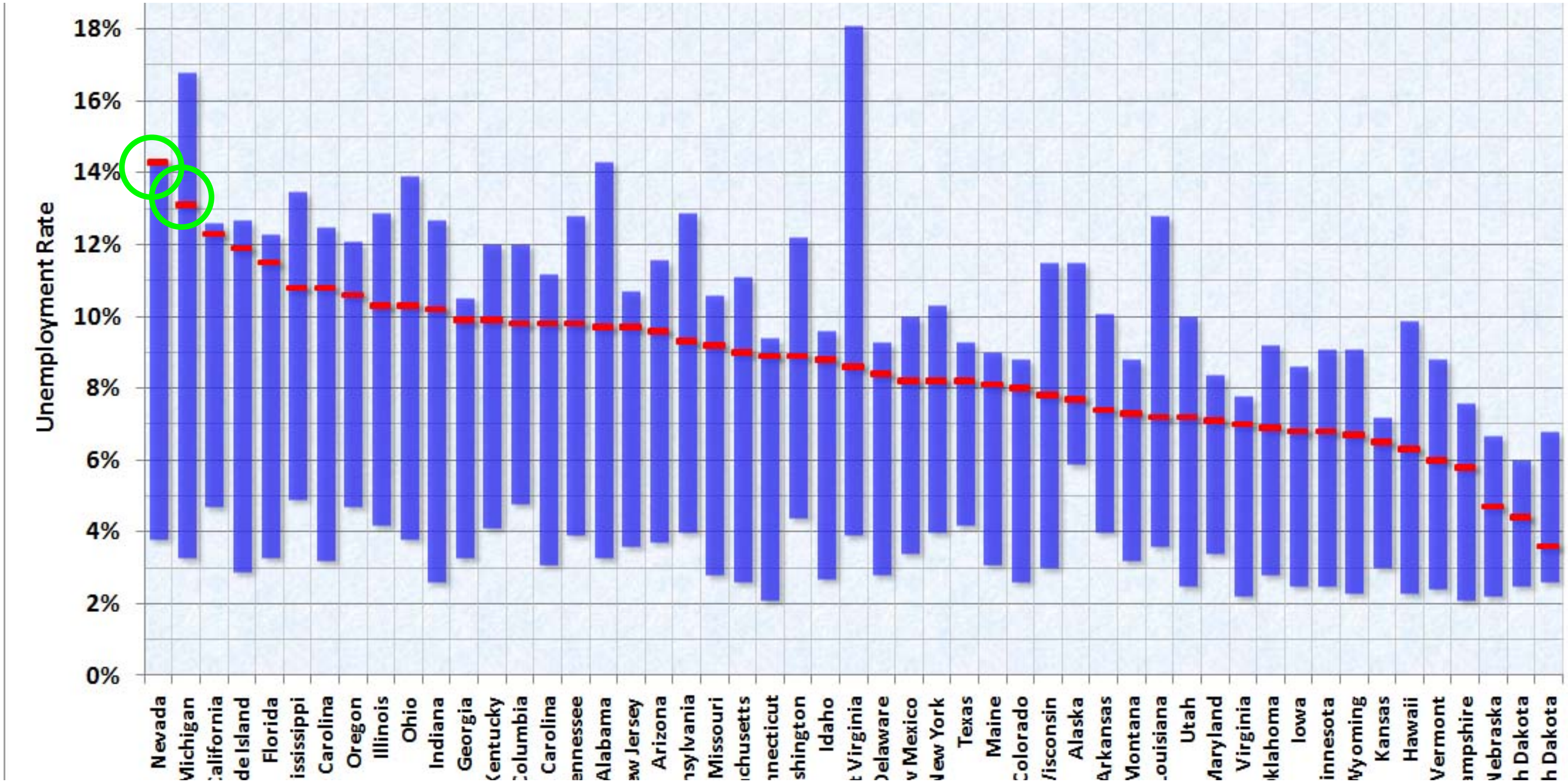
- a lot of what we visualize are the descriptive statistics
  - Example: mean income vs median income
  - Need to ensure that the univariate units of visualization are legit

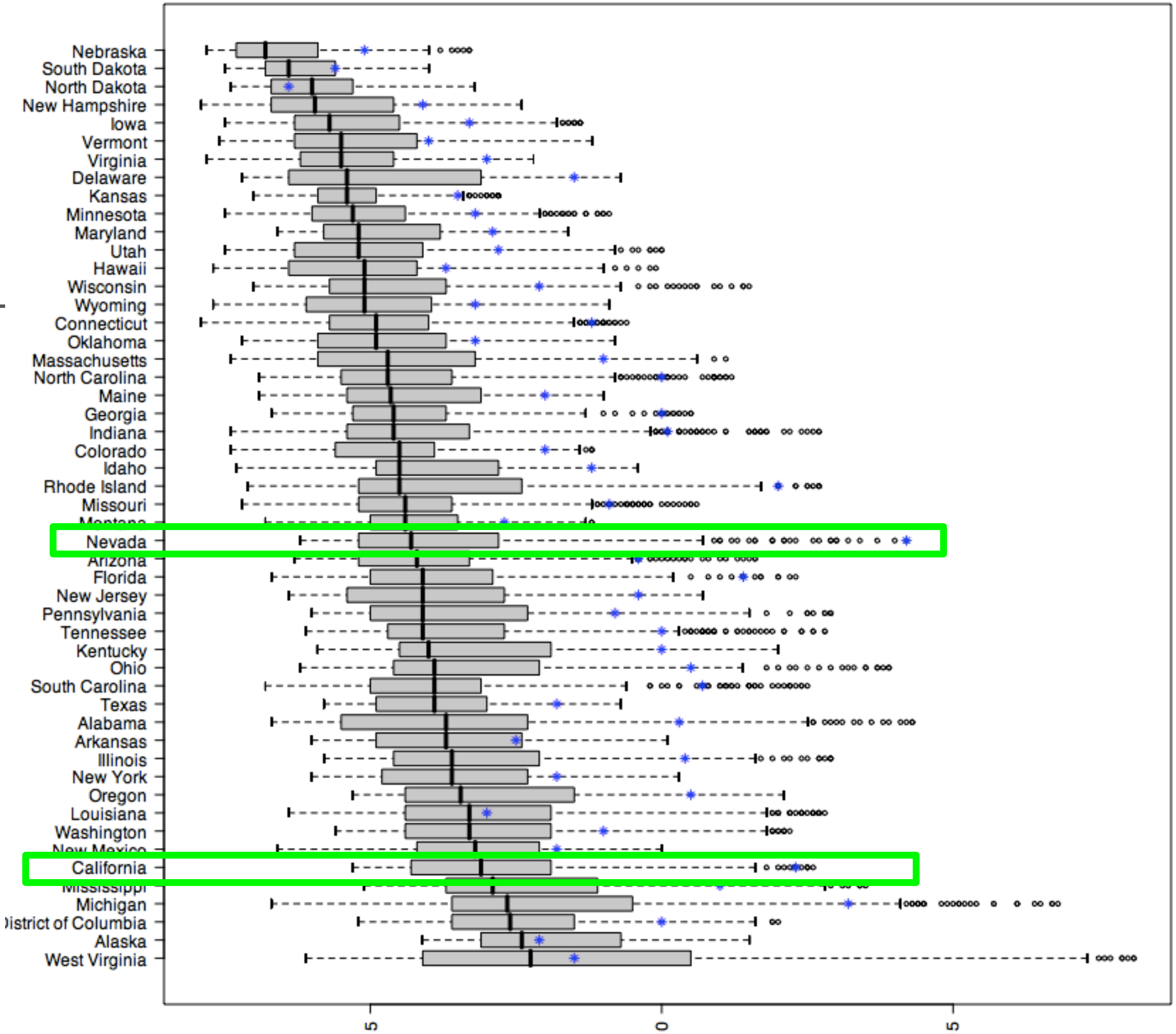
- Rule: check your core units /variables. If hey are descriptive, look at the distribution

Example: job losses in US over time



Example: job losses in US over time





Nebraska
South Dakota
North Dakota
New Hampshire
Iowa
Vermont
Virginia
Delaware
Kansas
Minnesota
Maryland
Utah
Hawaii
Wisconsin
Wyoming
Connecticut
Oklahoma
Massachusetts
North Carolina
Maine
Georgia
Indiana
Colorado
Idaho
Rhode Island
Missouri
Nevada
Florida
New Jersey
Pennsylvania
Tennessee
Kentucky
Ohio
South Carolina
Texas
Alabama
Arkansas
Illinois
New York
Oregon
Louisiana
Washington
California
Michigan
District of Columbia
Alaska
West Virginia
5
10
15