

# Exploratory Data Analysis

*Ivan Corneillet*

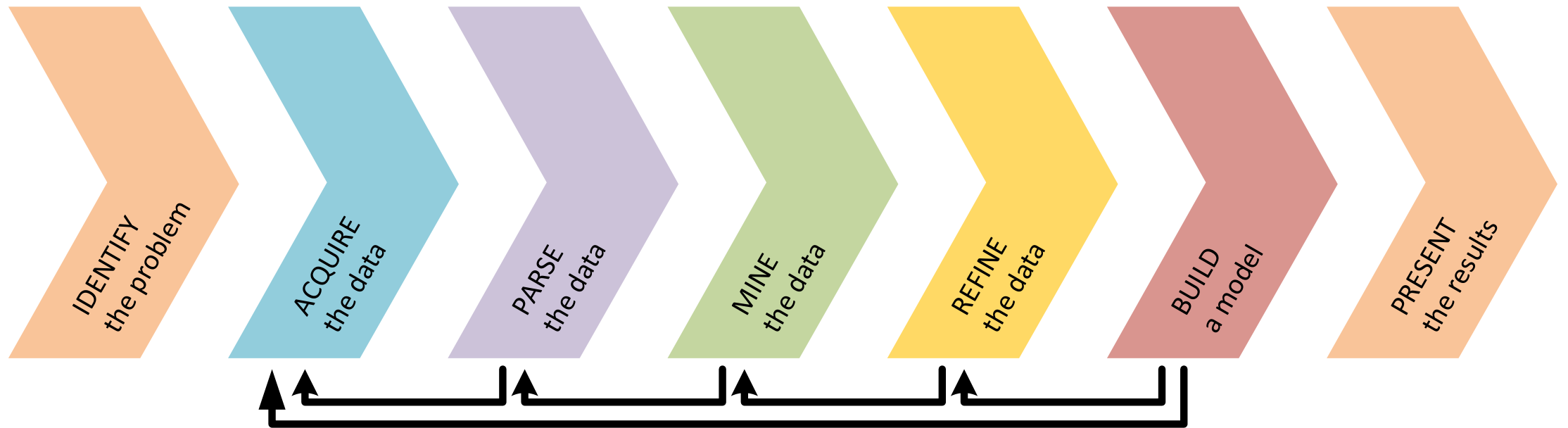
*Data Scientist*

# Learning Objectives

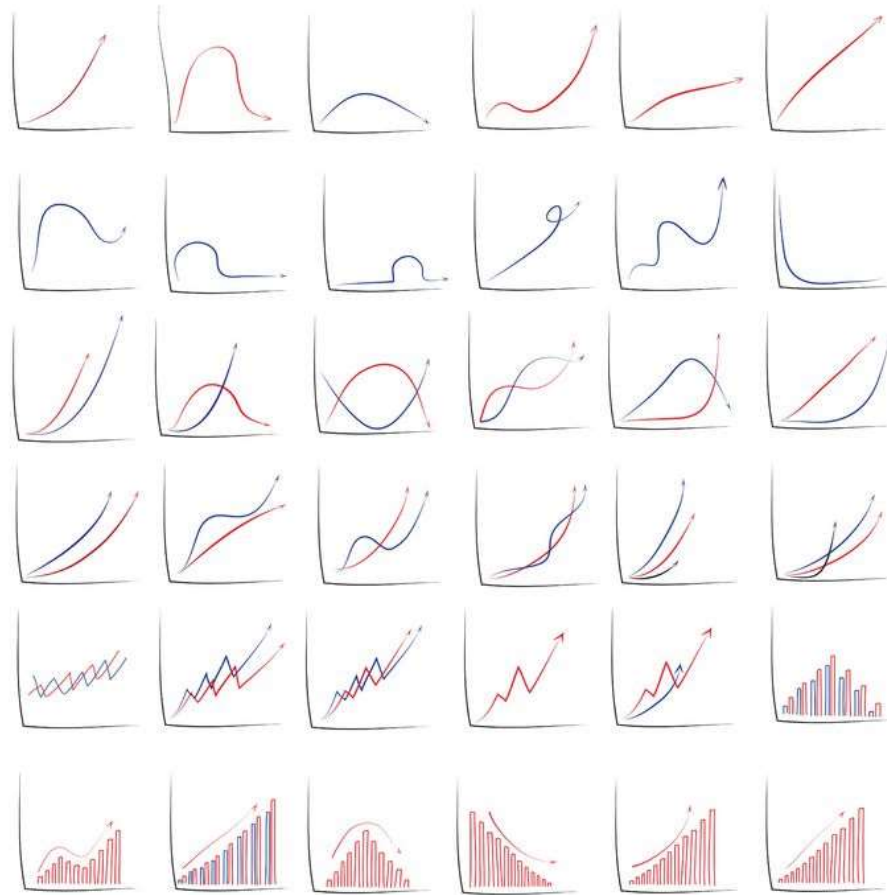
After this lesson, you should be able to:

- Identify variable types
- Use the *pandas* (and *NumPy*) libraries to analyze datasets using basic summary statistics: mean, median, mode, max, min, quartile, inter-quartile range, variance, standard deviation, and correlation
- Create data visualizations – including boxplots, histograms, and scatter plots – to discern characteristics and trends in a dataset

# The Data Science Workflow



# Today, our main goal is to gain enough descriptive statistics knowledge to perform exploratory data analysis



Napat Polchoke © 123RF.com

- Descriptive vs. inferential statistics; populations vs. samples
- Types of data and types of measurement scales
- Measures of central tendency and measures of dispersion
- Boxplots
- Outliers
- Histograms
- Correlation



DS

# Review

*The SMART Goals Framework for Data Science*

# The SMART Goals Framework for Data Science

([https://en.wikipedia.org/wiki/SMART\\_criteria](https://en.wikipedia.org/wiki/SMART_criteria))

<b>S</b> <sub>PECIFIC</sub>	The dataset and key variables are clearly defined
<b>M</b> <sub>EASURABLE</sub>	The type of analysis and major assumptions are articulated
<b>A</b> <sub>TTAINABLE</sub>	The question you are asking is feasible for your dataset and is not likely to be biased
<b>R</b> <sub>EPRODUCIBLE</sub>	Another person (or you in 6 months!) can read your state and understand exactly how your analysis is performed
<b>T</b> <sub>IME-BOUND</sub>	You clearly state the time period and population for which this analysis will pertain

Trends often change over time and vary by the population of source of your data. It is important to clearly define who/what you included in your analysis as well as the time period for the analysis

# Activity | A SMART Goal for your “Final” Project



## EXERCISE

### DIRECTIONS (15 minutes)

1. After the first few classes, you probably started brainstorming on an idea for your final project. If not, here's an opportunity!
2. If your idea is cool and interesting, that's great. But it is a SMART idea?
3. Assess your idea using the Data Science-tuned SMART Goal Framework
  - a. If you have just a couple of gaps, how can you close them?
  - b. On the other end, if you have too many and closing these gaps would be difficult, you might want to consider something else
4. After 5 minutes, share your idea and gaps in pairs and offer advice to each other, again using the SMART Framework (3 minutes each)

### DELIVERABLE

Answers to the above questions



DS

# Review

*The pandas Library*



# Activity | The *pandas* Library



## EXERCISE

### DIRECTIONS (10 minutes)

1. Using the dataset in the codealong (Part A), answer the following questions:

1. Subset the dataframe on the age and gender columns.
2. Subset the dataframe on the age column alone, first as a *DataFrame*, then as a *Series*
3. Subset the dataframe on the rows Bob and Carol
4. Subset the dataframe on the row Eve alone, first as a *DataFrame*, then as a *Series*
5. How old is Frank?
6. What is the men's mean age, the women's median age?

2. Annotate the next handout with the syntax on how to subset dataframes in *pandas*

3. When finished, share your answers with your table

### DELIVERABLE

Answers to the above questions

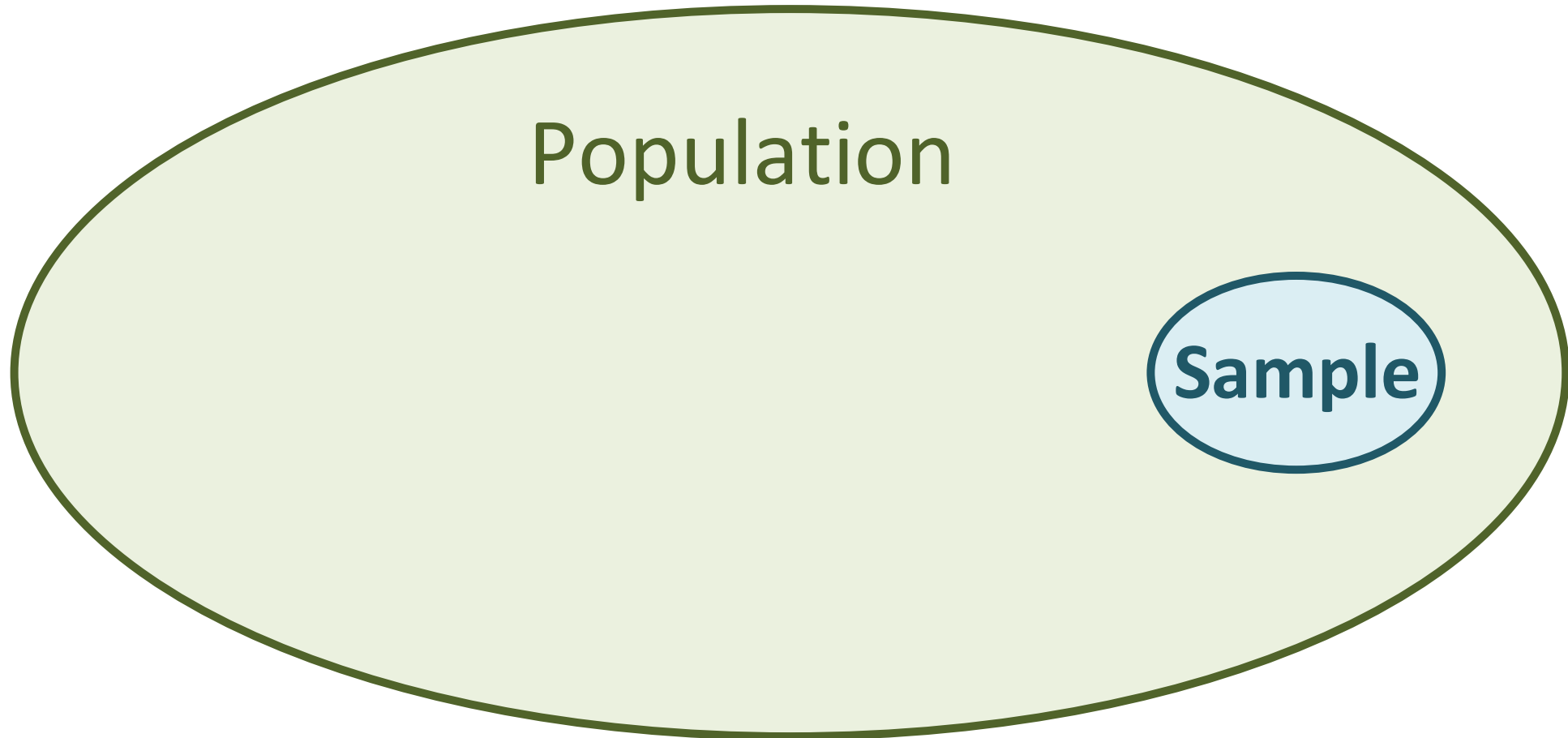
	DataFrame	Series
Column subsetting		
<b>by name</b>  (Columns names are stored in df.columns) (df.columns.get_loc('X1') returns X1's column index)	# New DataFrame with column named X1 df[ ['X1'] ]  # 2+ columns (in the order listed) df[ ['X1', 'X2', ...] ]	df['X1']  df.X1
<b>by location</b>	# New DataFrame with column at location i (numbering starts at 0) df[ [column_i] ]  # 2+ columns (in the order listed) df[ [column_i, column_j, ...] ]	
Row subsetting		
<b>by index label</b>	df.loc[ [index_label_i] ] df.loc[ [index_label_i, index_label_j, ...] ]  # Can use a range if the index is made of numbers (rows "a" to "b" included) df.loc[ index_label_a : index_label_b ]	df.loc[index_label_i]
<b>by location</b>	df.iloc[ [row_i] ] df.iloc[ [row_i, row_j, ...] ]  # (rows "a" to "b" excluded) df.iloc[row_a : row_b ] or df[row_a : row_b ]	df.iloc[location_i]
Cell/scalar lookup		
<b>by index label/column name</b>	df.at[index_label, 'X1']	
<b>by location</b>	df.iat[row_i, column_j]	



DS

# Populations and Samples

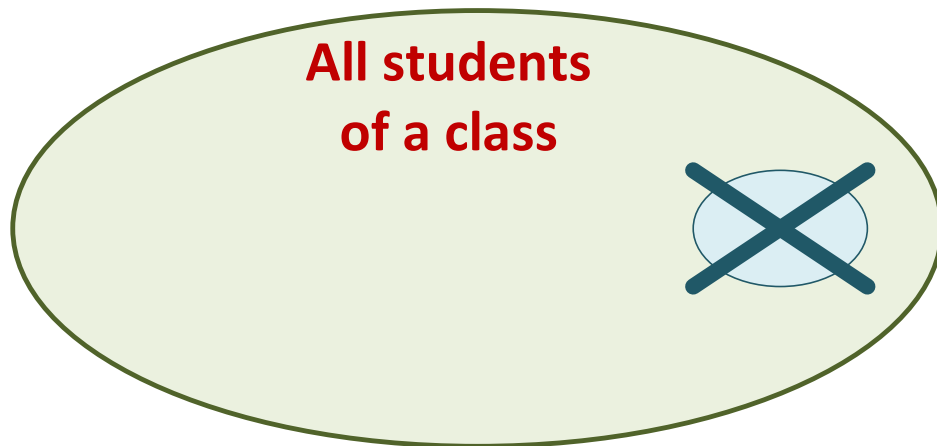
# Populations and Samples



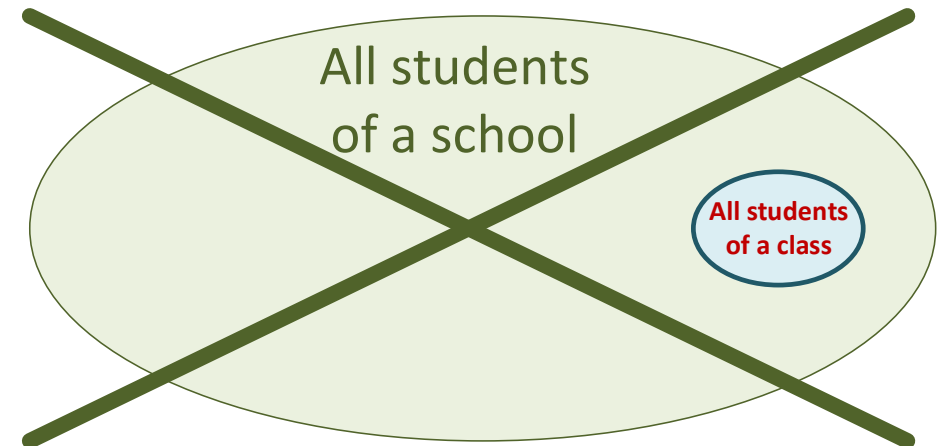
# A dataset may be considered either as a population or a sample, depending on the reason for its collection and analysis

- Students of a class are a population if the analysis describes the distribution of scores in that class
- But they are a sample if the analysis infers from their scores the scores of other students (e.g., all students from that school)

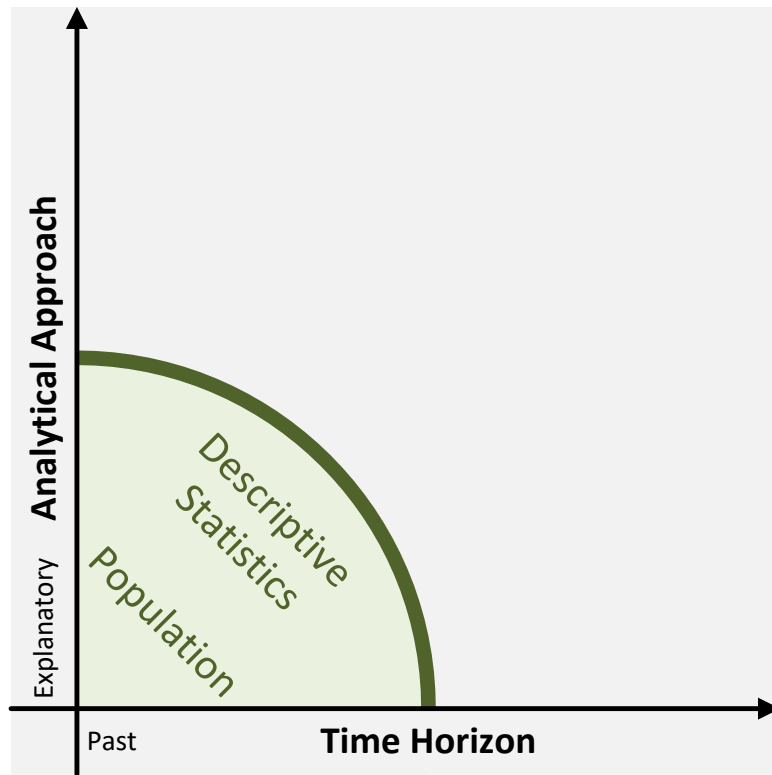
## Descriptive Statistics



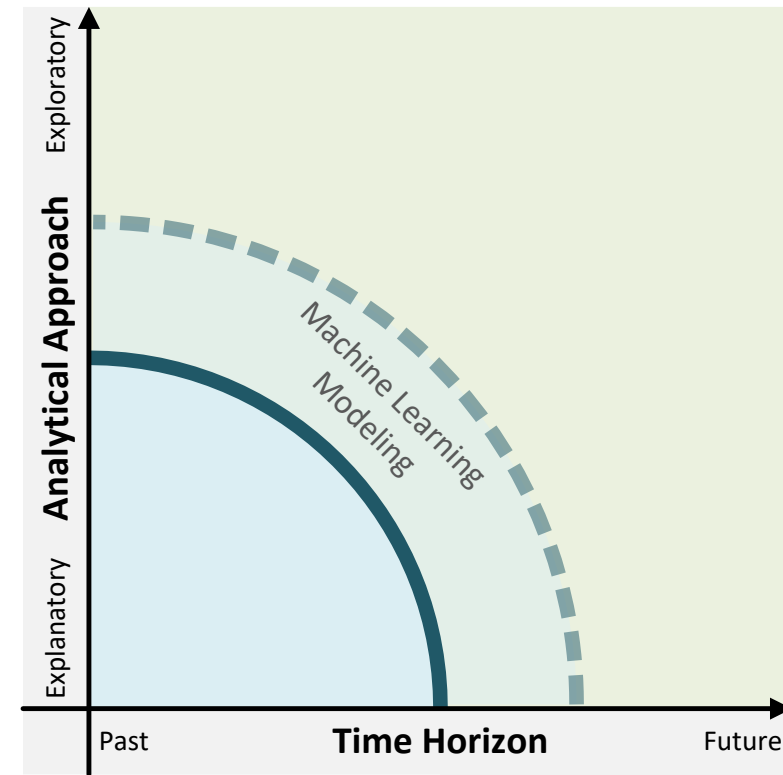
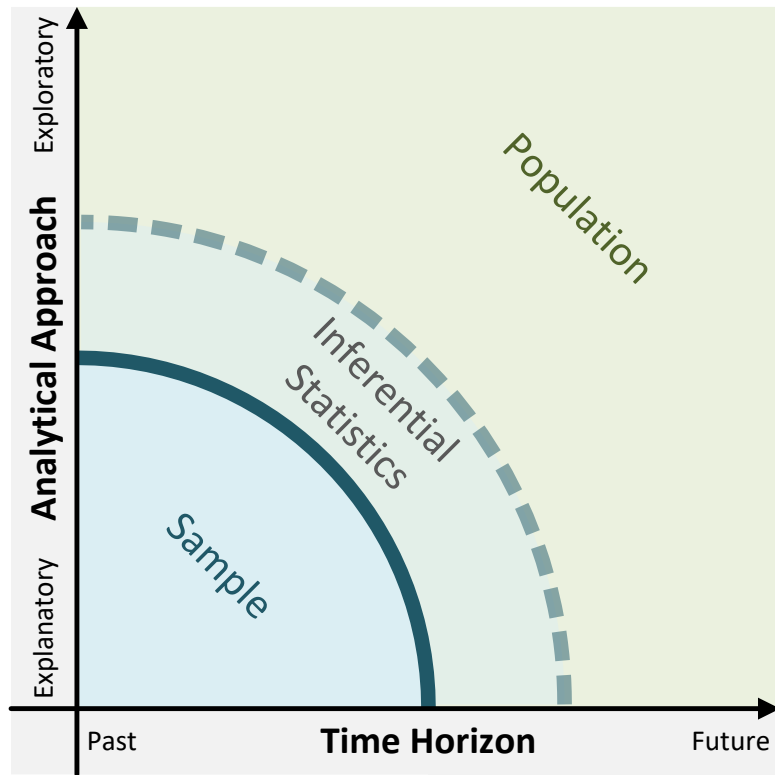
## Inferential Statistics



Exploratory data analysis is concerned with descriptive statistics (e.g., “what happened last quarter?” and “how many units were sold?”)



Machine learning modeling concerns itself with inferential statistics (e.g., “what if ...?”, “what will happen next?”, and “what if these trends continue?”)





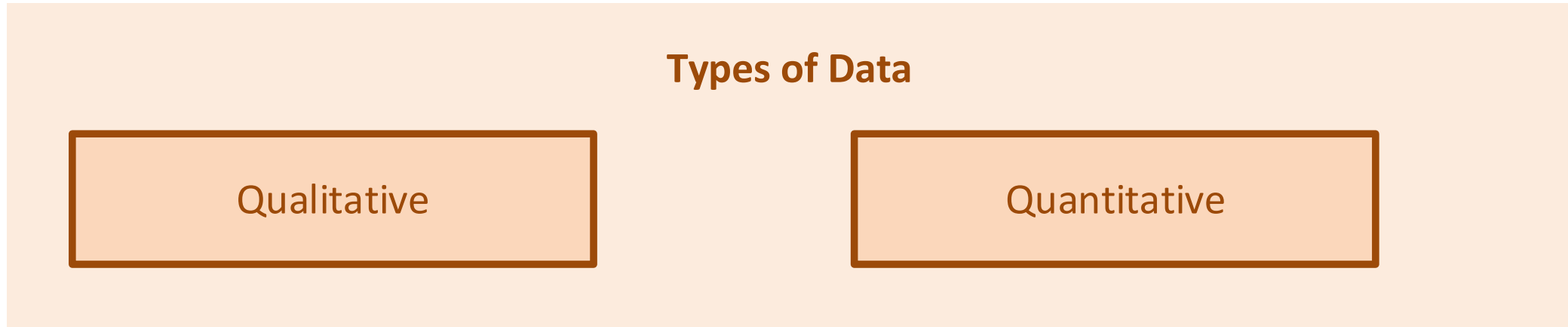
DS

# Types of Data

## Types of Measurement Scales



# Types of Data



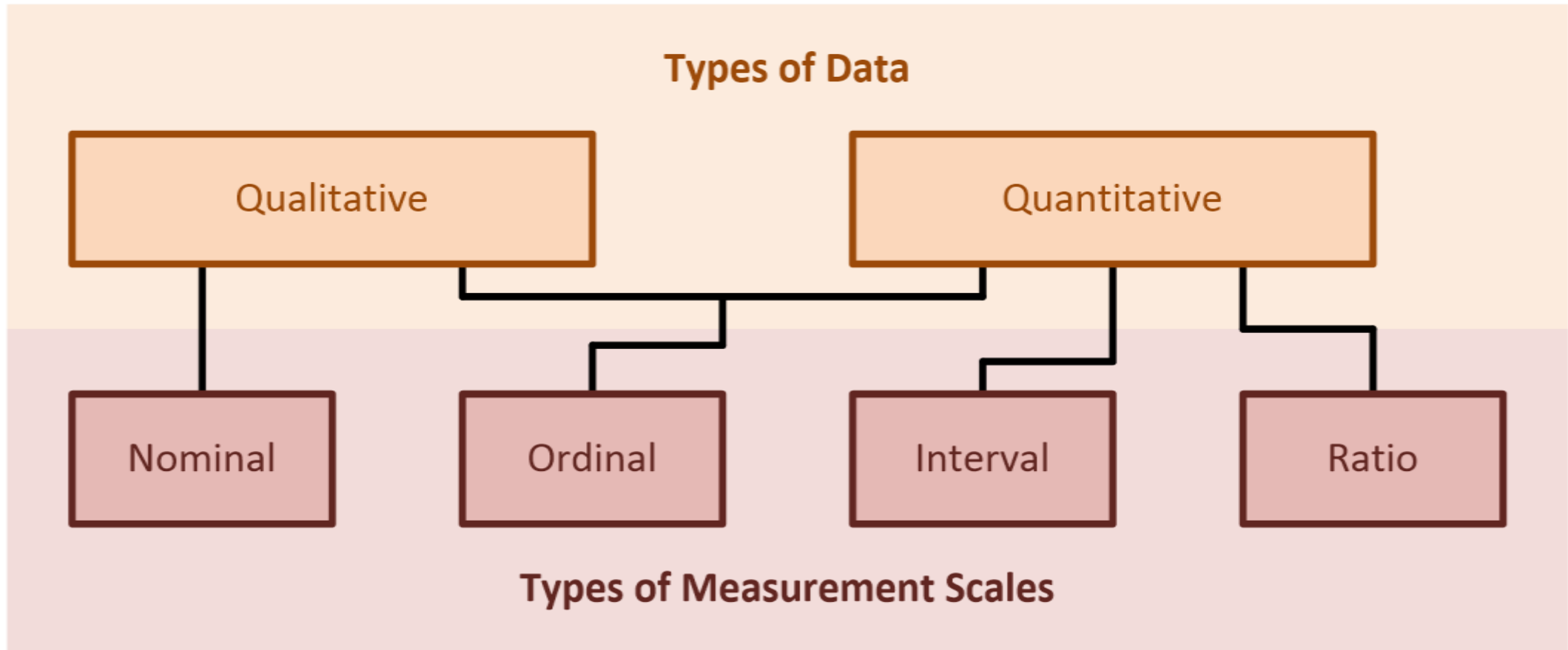
- Qualitative Data

- Uses descriptive terms to measure or classify something of interest, e.g., education level

- Quantitative Data

- Uses numerical values to describe something of interest, e.g., age

# Types of Measurement Scales



# Types of Measurement Scales (cont.)

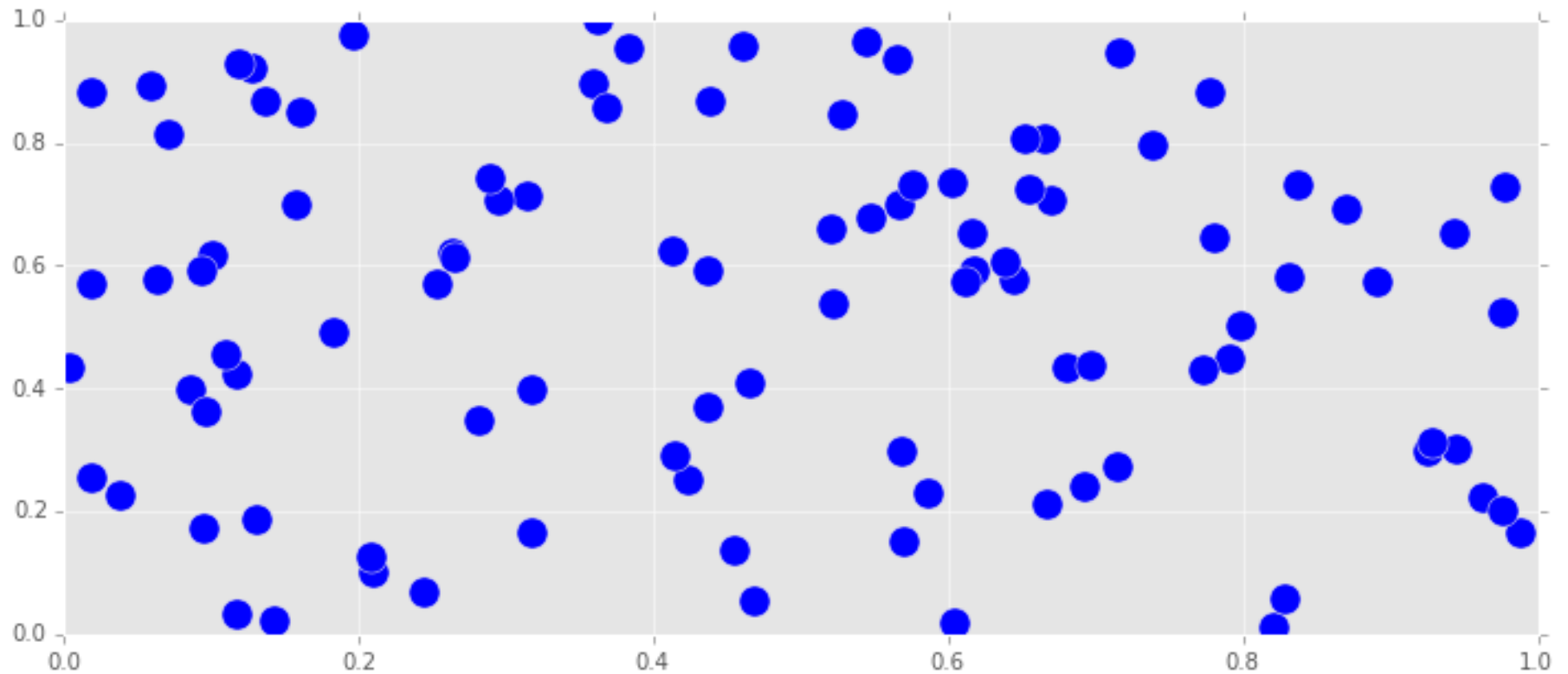
	Nominal	Ordinal	Interval	Ratio
<i>e.g.</i>	<i>Gender</i>	<i>Movie ratings</i>	<i>Temperature</i>	<i>Salary</i>
<b>Categorize?</b>	✓ (male, female)	✓	✓	✓
<b>Rank-order?</b>	✗	✓ (★ < 2★ < 3★ < 4★)	✓	✓
<b>Add and subtract?</b>	✗	✗ (4★ - 3★ ≠ ★)	✓ (75°C is 50°C warmer than 25°C)	✓
<b>Multiply and divide?</b>	✗	✗ (4★ not 4× better than 1★)	✗ (75°C not 3× as warm as 25°C) (0°C doesn't mean no temperature!)	✓ (Salary of \$200K is 2× that of \$100K) (\$0 means no salary ☹️)

DS

# Measures of Central Tendency Measures of Dispersion

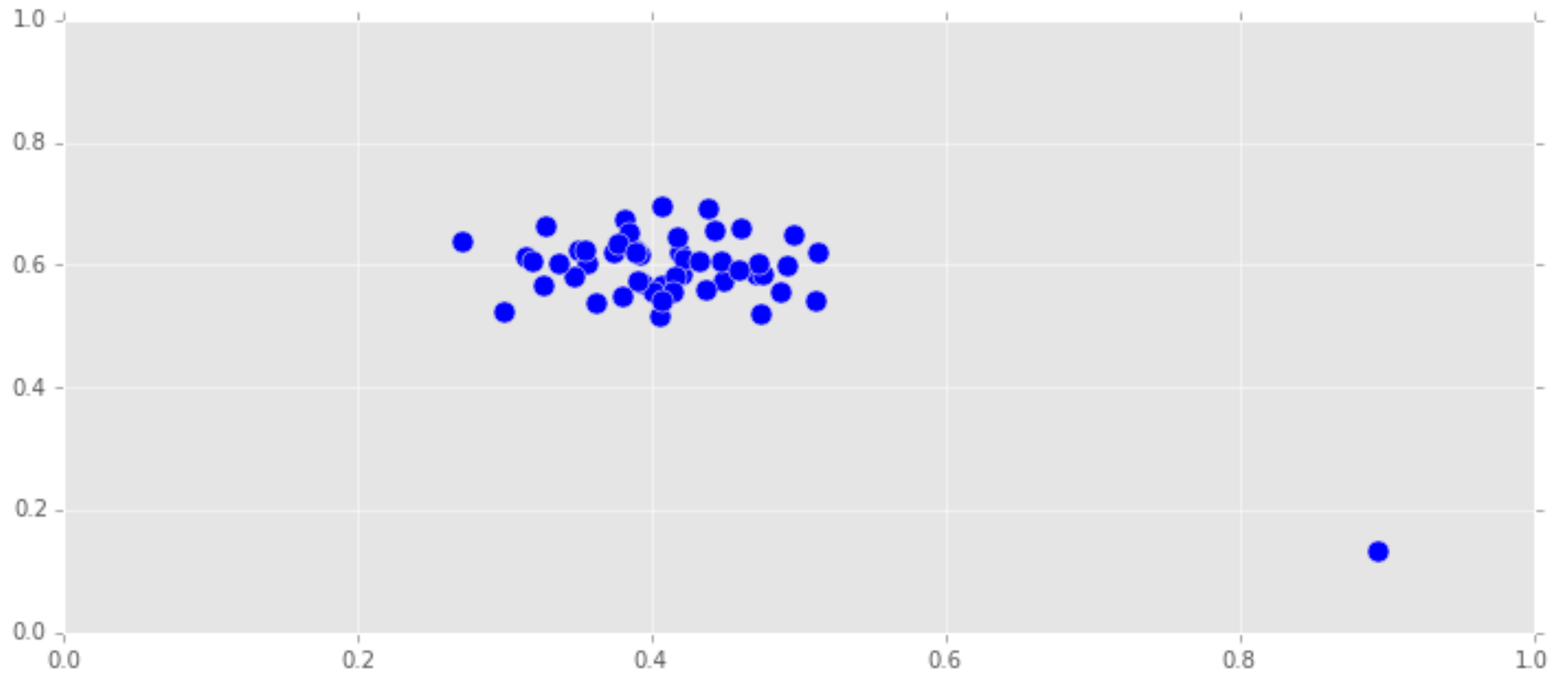
# Activity | How would you summarize this data?

EXERCISE



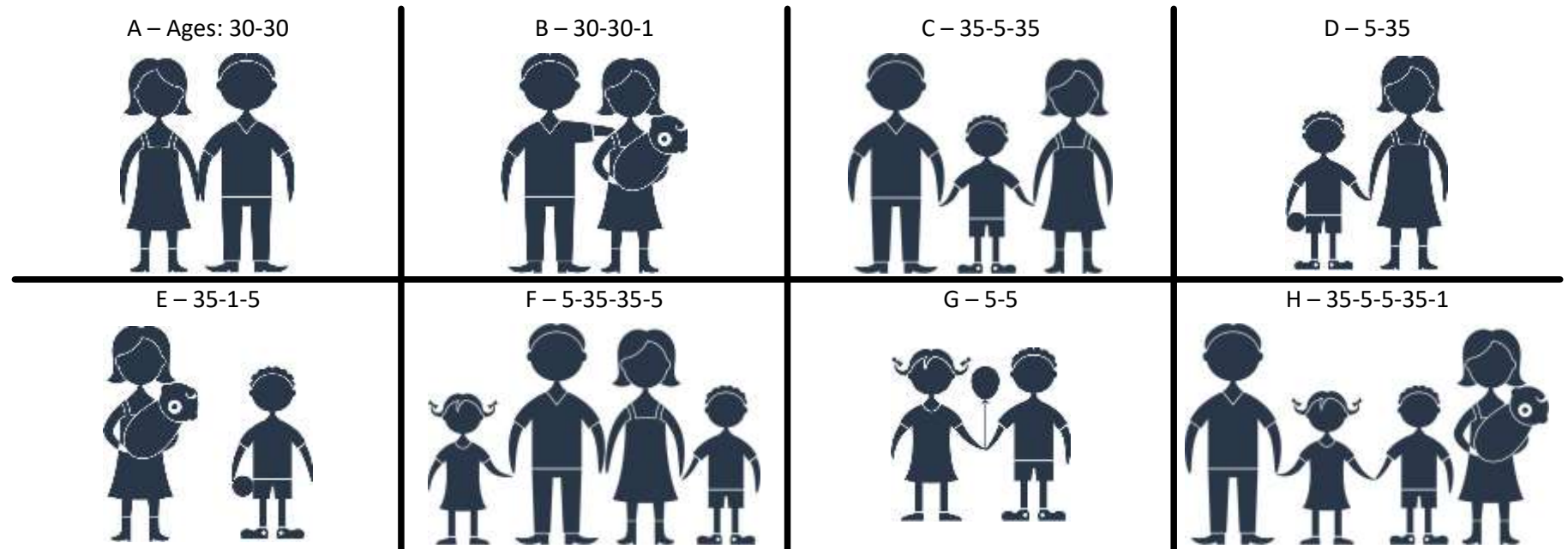
# Activity | How would you summarize this data? (cont.)

EXERCISE


















Activity | Measures of Central Tendency. What is the typical age for each of these 8 groups of people? (10 minutes)

EXERCISE



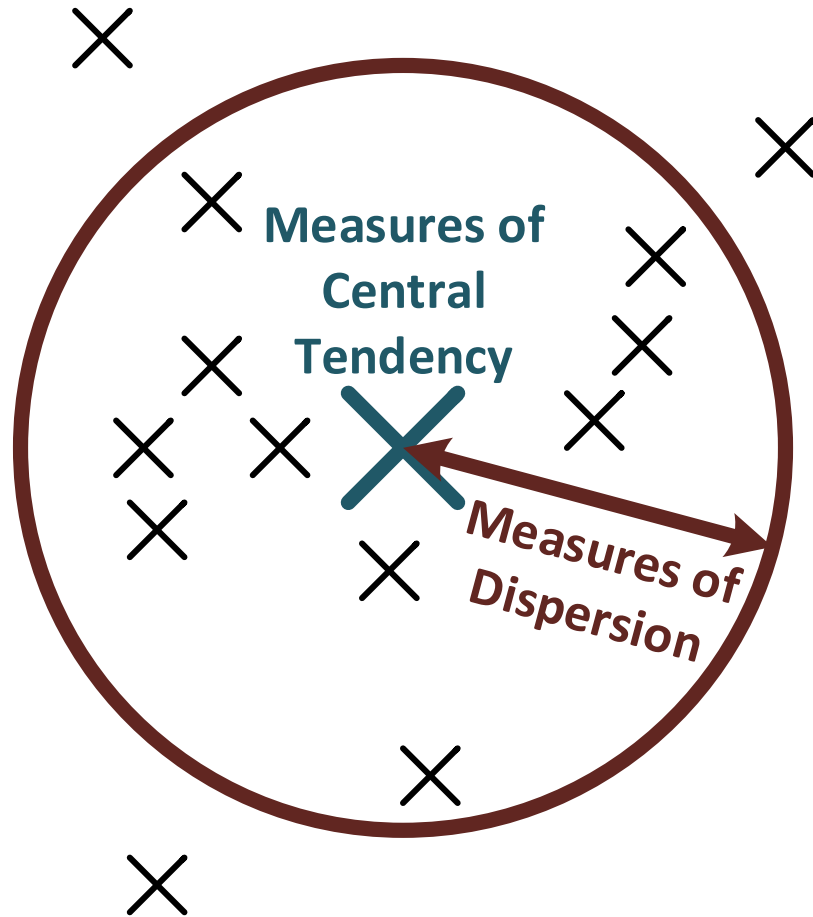
macrovector © 123RF.com

# Mean, Median, and Mode | Trade-offs

	Value is in the dataset	Value is easy to compute	Value is resistant to outliers	Corresponding measure of Dispersion	Used extensively by mathematical models
Mean	 (Unlikely)			 (Variance, standard deviation)	
Median	 (50% chance)	 (need to rank the values)		 (Interquartile Range)	
Mode	 (Always)	 (Need to count and rank the count)		 (Not really)	 (Mode might not be defined or you might have multiple values)



# Measures of Central Tendency and Measures of Dispersion



- Measures of Central Tendency
  - (Or measures of location)
  - Answer the question: “What’s the typical or common value for a variable?”
  - Mean, Median, Mode
- Measures of Dispersion
  - (Or measures of variability/spread)
  - Answer the question: “How far do values stray from the typical value?”
  - Variance, Standard Deviation, Range, Interquartile Range (IQR)

# (Arithmetic) Mean, Variance, and Standard Deviation

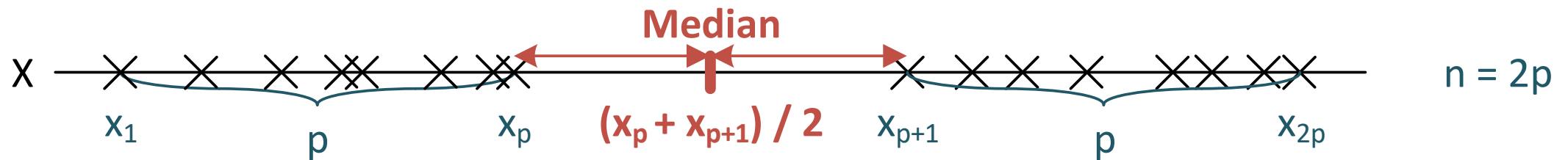
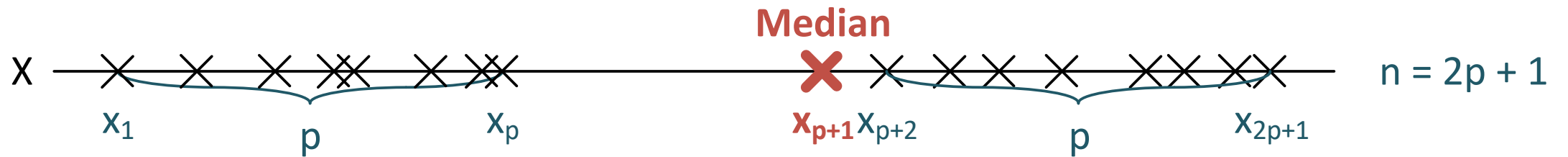
	Ordinal ✗	Nominal ✗	Interval ✓	Ratio ✓
	Population		Sample	
<b>(Arithmetic) Mean</b> <i>(a.k.a., the first moment)</i> (Mean has unit of $X:[X]$ )	$\mu = \frac{1}{N} \sum_{i=1}^N x_i = E[X^1]$ (mu)		$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ (x-bar)	
<b>Variance</b> <i>(a.k.a., the second moment)</i> $[X^2]$	$\sigma^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2$ $= E[(X - \mu)^2]$ (sigma-squared)		$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$	
<b>Standard Deviation</b> $[X]$	$\sigma = \sqrt{\sigma^2}$ (sigma)		$s = \sqrt{s^2}$	

(mean, variance, and standard deviations are based on the values of  $x_i$ )

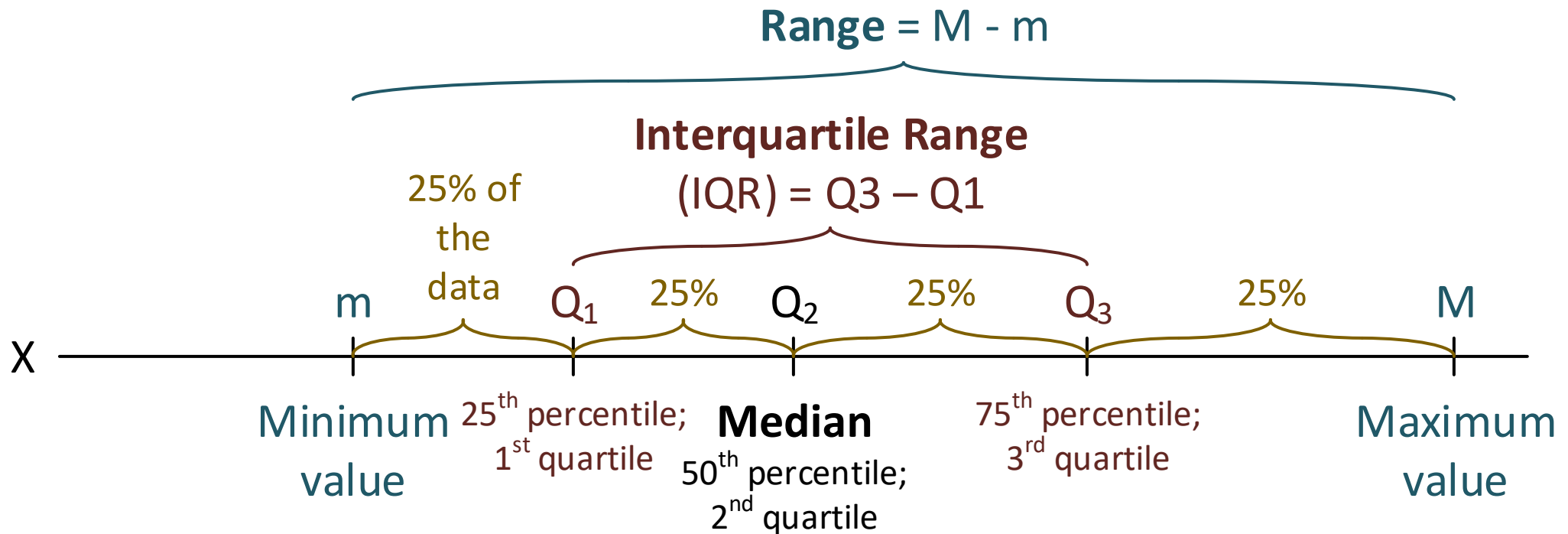
DS

# Median, Range, and Interquartile Range

# Median



# Median, Range, and Interquartile Range



# Median, Range, and Interquartile Range (cont.)

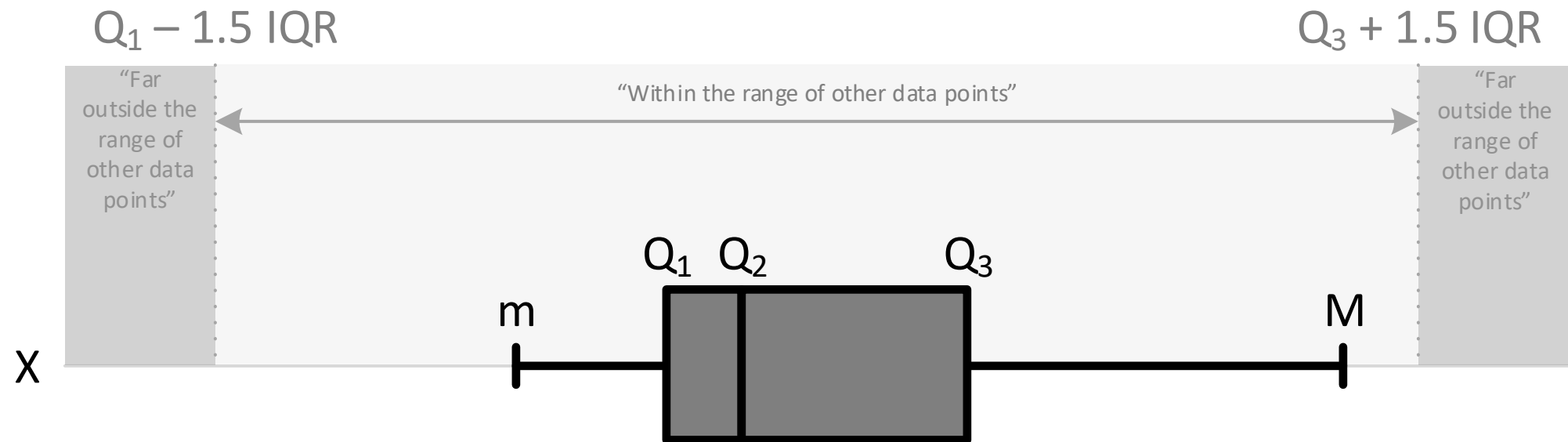
Nominal ✖		Ordinal ✖		Interval ✓		Ratio ✓	
Median		$median = \begin{cases} x_{p+1} & \text{if } n = 2p + 1 \\ \frac{x_p + x_{p+1}}{2} & \text{if } n = 2p \end{cases}$					
Range		$range = x_n - x_1$					
Percentile		$q_k = \begin{cases} x_{[p]} & \text{if } p = \frac{nk}{100} \text{ not integer} \\ \frac{x_p + x_{p+1}}{2} & \text{otherwise} \end{cases}$					
Quartile		$Q_1 = q_{25}; Q_3 = q_{75}$					
Interquartile Range		$IQR = Q_3 - Q_1$					

(median, range, and interquartile range are based on the ranks of  $x_i$ ;  $x_i$  ranked from smallest to largest)

DS

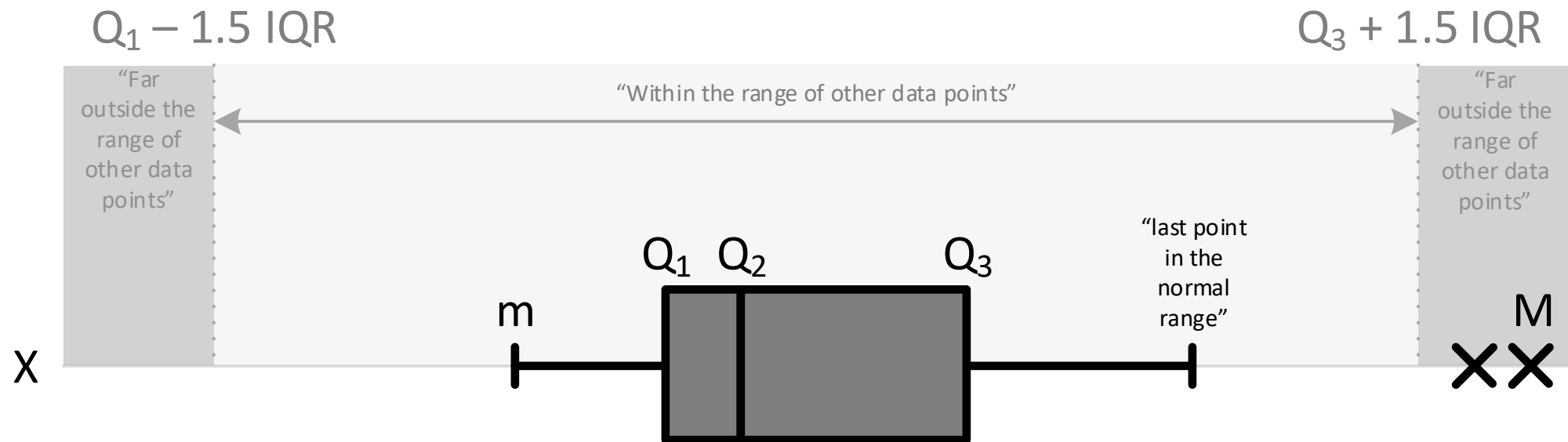
# Boxplots

# Boxplot #1 | Median, Range, Interquartile Range; no Outliers





# Boxplot #2 | Median, Range, Interquartile Range; with Outliers

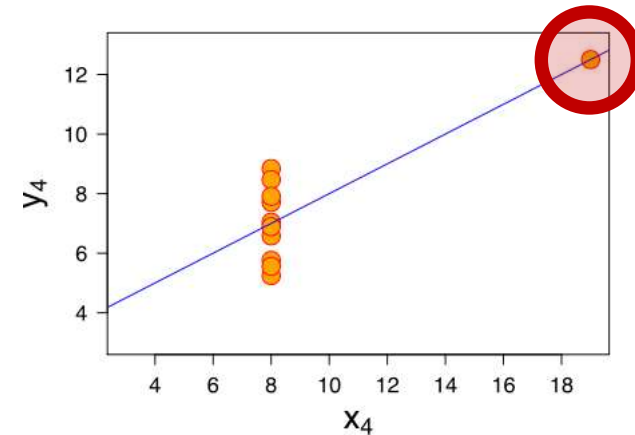
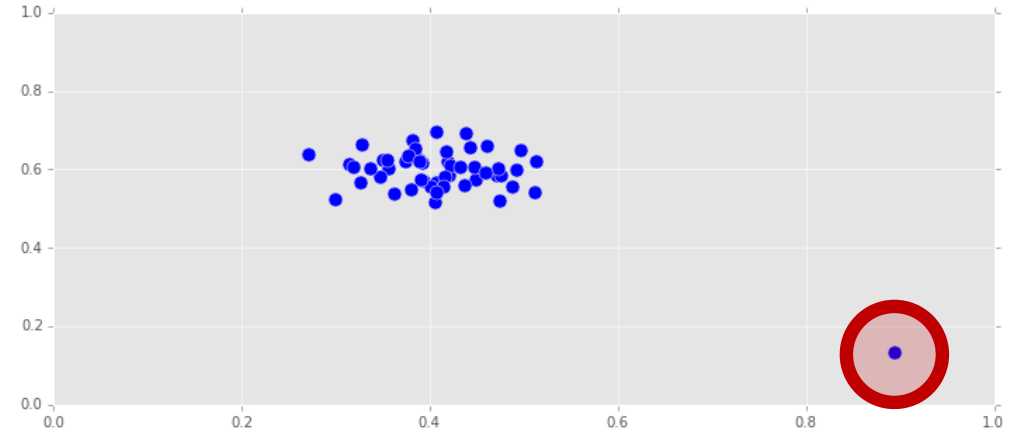


DS

# Outliers

# Think twice before discarding outliers; they might be the most important points of your dataset

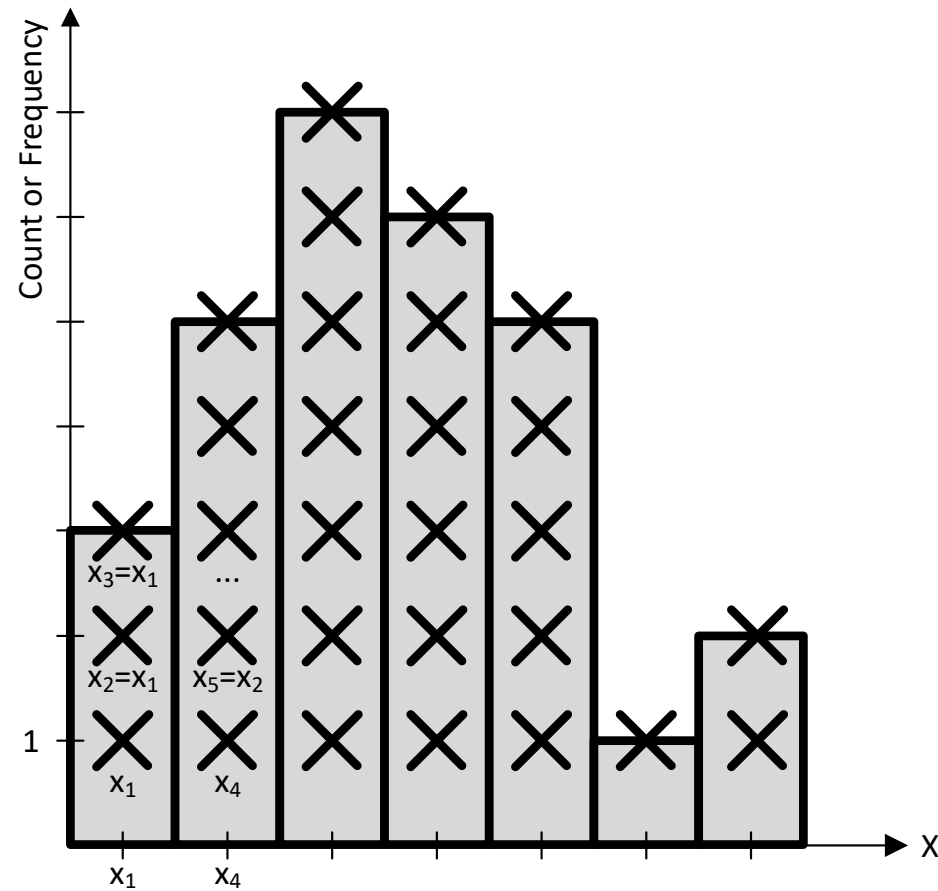
- Outliers are values that are “far” from the central tendency
- No formal definition among statisticians on how to define outliers (how do you define “far”?)
- However, general agreement that they be identified and dealt with appropriately (e.g., keep or discard)
  - They might be the most important points of your dataset



DS

# Histograms

Histograms.  $x_1 = x_2 = x_3 < x_4 = x_5 \dots$





# Mode

# Modes and Histograms

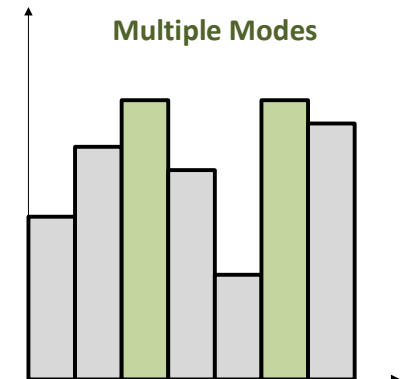
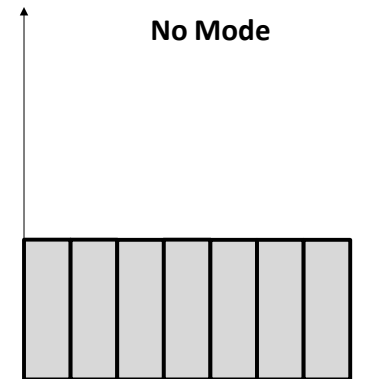
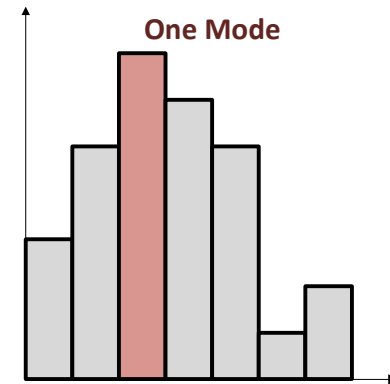
Nominal ✓

Ordinal ✓

Interval ✓

Ratio ✓

- The Mode is the value(s) that occur(s) most often





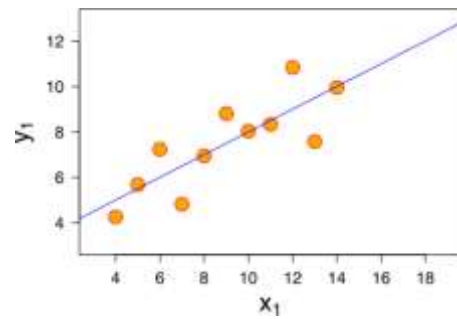
DS

# Plot the Data!

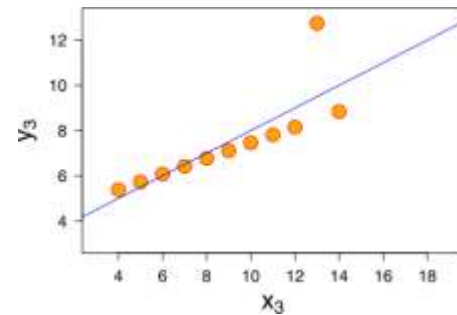


Don't rely on basic statistic properties and **plot the data!** 4 datasets (Anscombe's quartet) that have nearly identical simple statistical properties, yet are very different

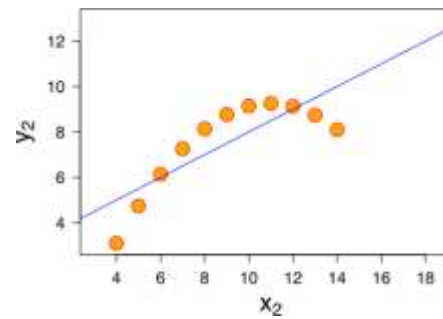
Scatter plot appears to be a simple linear relationship, corresponding to two variables correlated and following the assumption of normality.



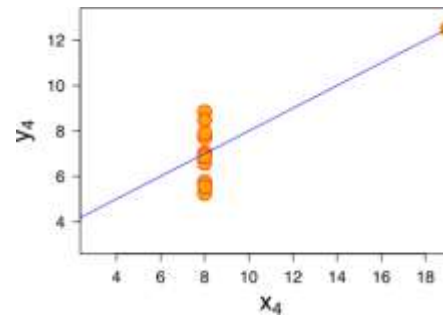
Distribution is linear, but with a different regression line, which is offset by the one outlier which exerts enough influence to alter the regression line.



Not distributed normally; while an obvious relationship between the two variables can be observed, it is not linear, and the linear correlation is not relevant.



Example when one outlier is enough to produce a high correlation coefficient, even though the relationship between the two variables is not linear.



Property	Value
Mean of $x_i$	9
Sample variance of $x_i$	11
Mean of $y_i$	7.50
Sample variance of $y_i$	4.122 or 4.127
Correlation between $x_i$ and $y_i$	0.816
Linear regression line in each case	$y_i = 3.00 + 0.500 x_i$

DS

# (Linear) Correlation

# Correlation

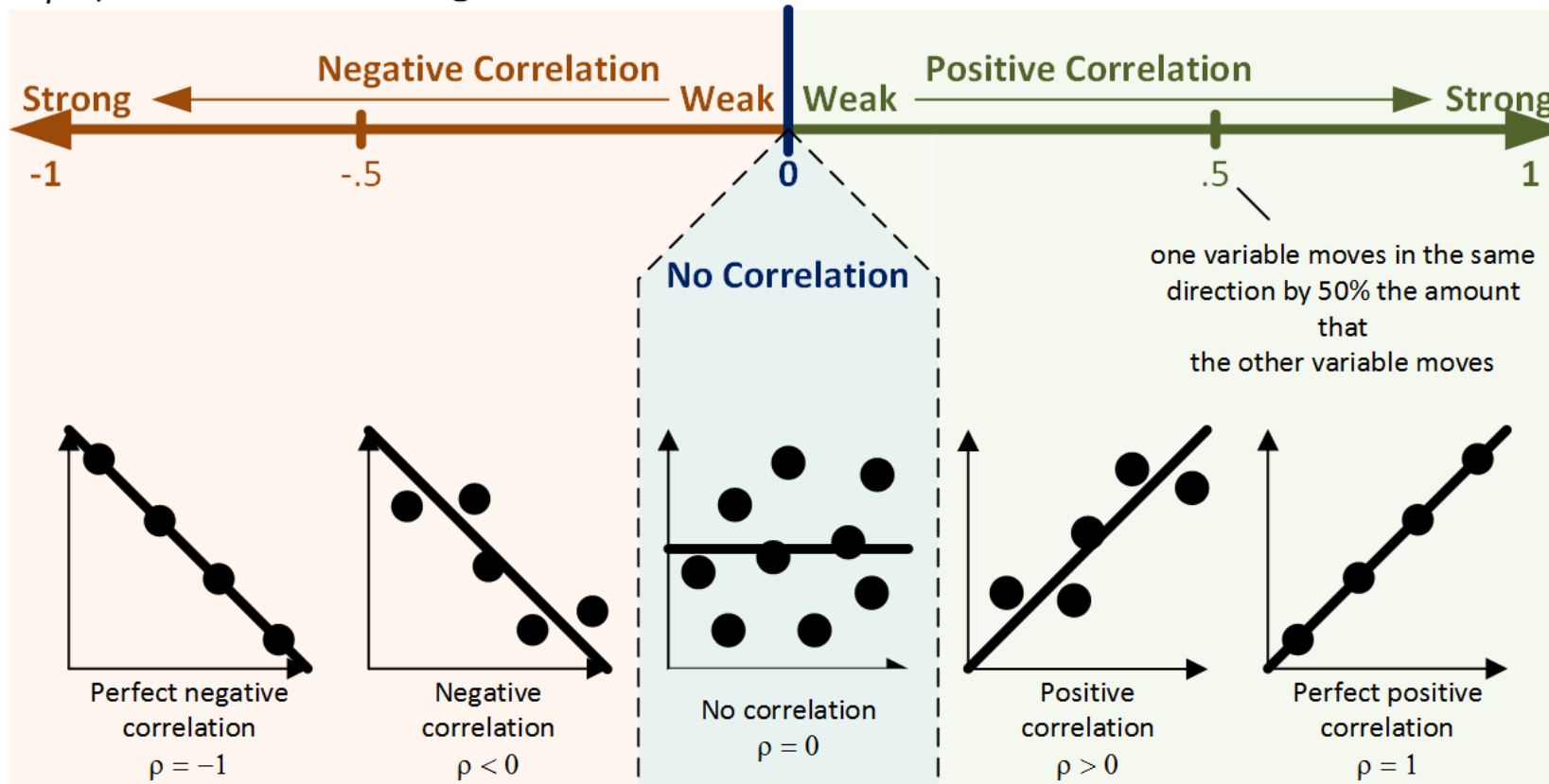
- A measure of strength and direction for a **linear association** between two random variables

$$\rho_{X,Y} = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y}$$

- $\rho = 0$  means that the two variables don't have a linear association
  - It doesn't imply that they are independent!

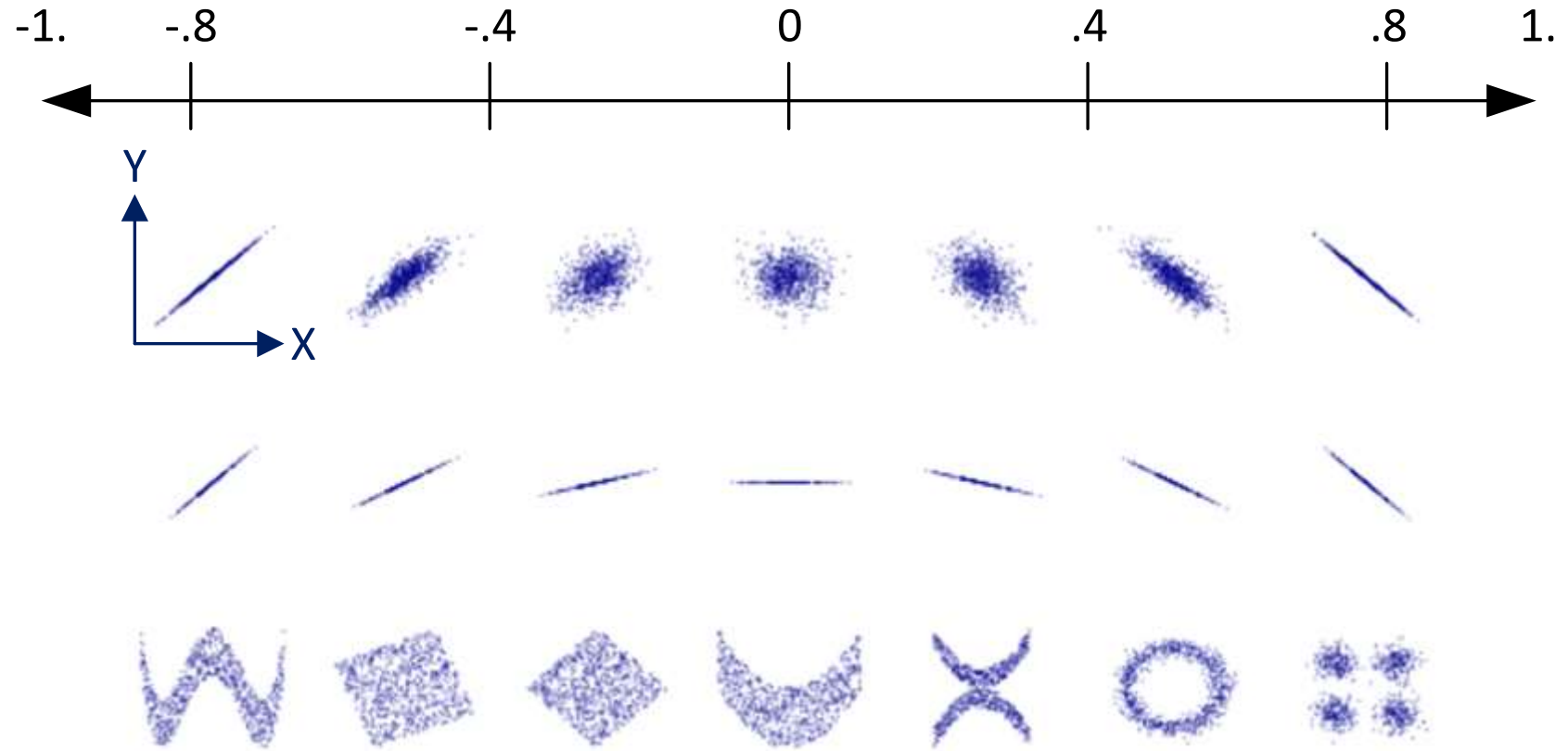
# Correlation (cont.)

$\rho$  quantifies the strength and direction of movements of two random variables



Activity | What are the correlations between  $X$  and  $Y$  for the following scatter plots (5 minutes)

EXERCISE



Slides © 2016 Ivan Corneillet Where Applicable  
Do Not Reproduce Without Permission