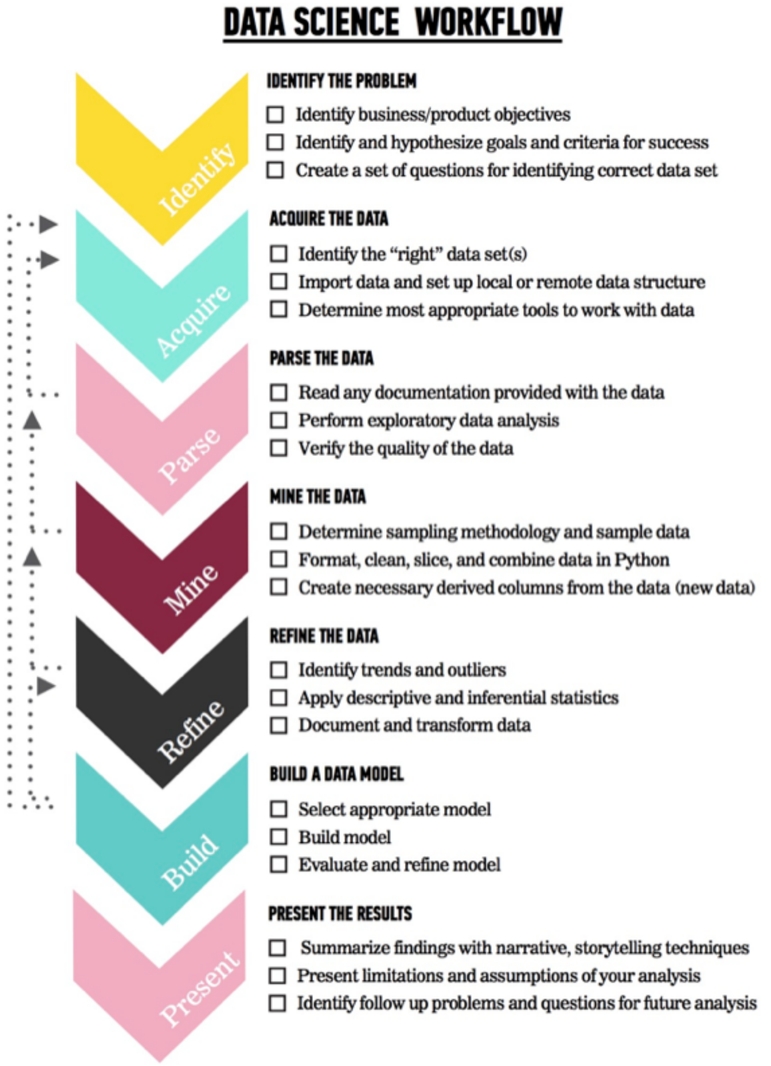
DSI

METHODS AND STATS

LET'S REVIEW THE DATA SCIENCE WORKFLOW

The steps:

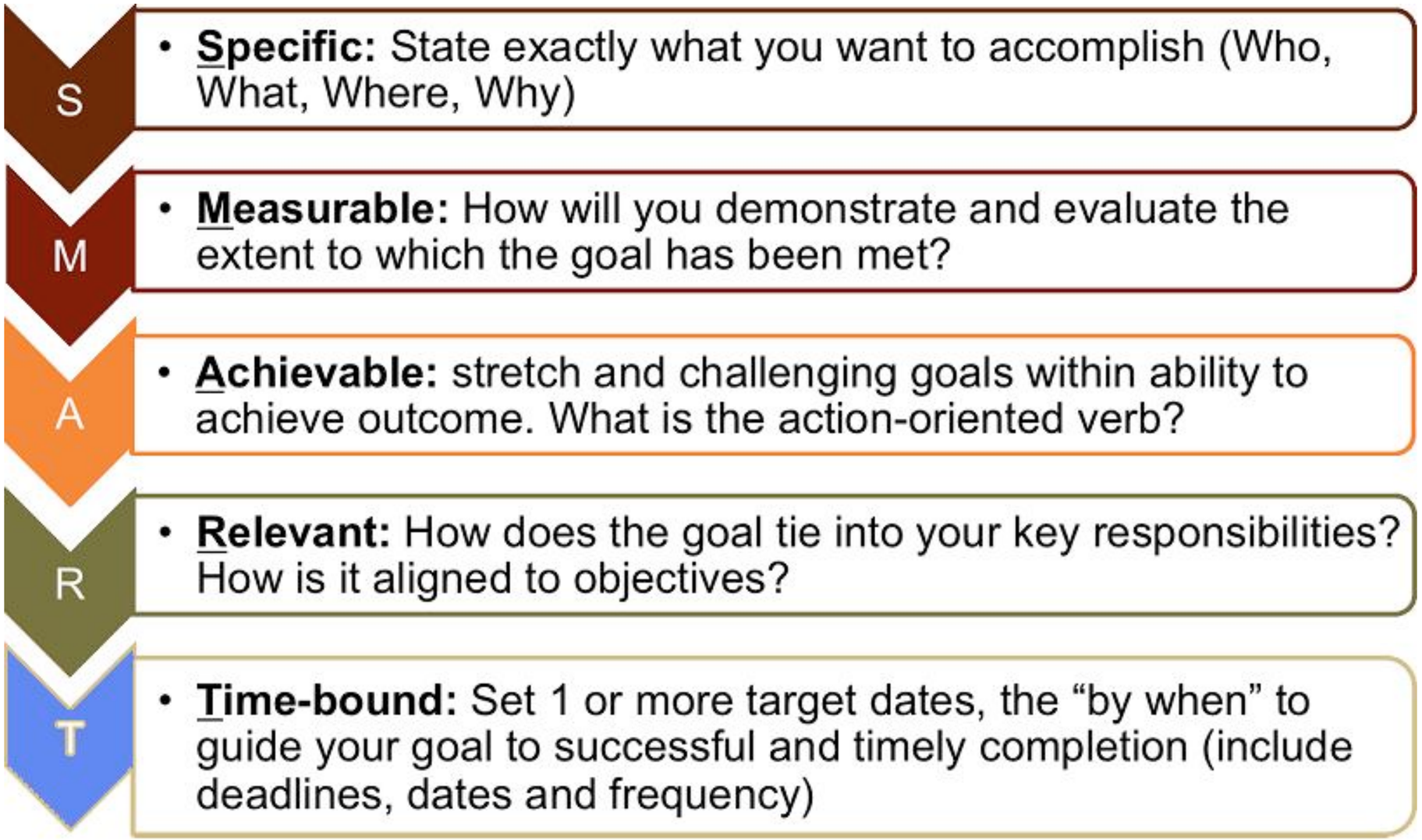
1. Identify the problem
2. Acquire the data
3. Parse the data
4. Mine the data
5. Refine the data
6. Build a data model
7. Present the results



WHY DO WE NEED A GOOD QUESTION?

- “A problem well stated is half solved.” -Charles Kettering
- Sets yourself up for success as you begin analysis
- Establishes the basis for reproducibility
- Enables collaboration through clear goals

WHAT IS A GOOD QUESTION?



NUMPY AND PANDAS INTRO

- What are Numpy and Pandas?
- Numpy uses arrays (lists) to do basic math and slice and index data.
- Pandas is built on Numpy.
- Pandas uses a data structure called a Dataframe.
  - Dataframes are similar to Excel tables: rows and columns.

NUMPY AND PANDAS INTRO

	A	B	C	D
2014-01-01	0.731803	2.318341	-0.126191	-0.903675
2014-01-02	0.161877	-0.892566	0.967681	-1.514520
2014-01-03	0.776626	1.797420	0.916972	0.634322
2014-01-04	2.020242	-0.763612	1.239145	-0.919727
2014-01-05	0.772058	0.417369	-0.957359	-0.916665
2014-01-06	-1.670217	-3.249906	2.017370	1.674340

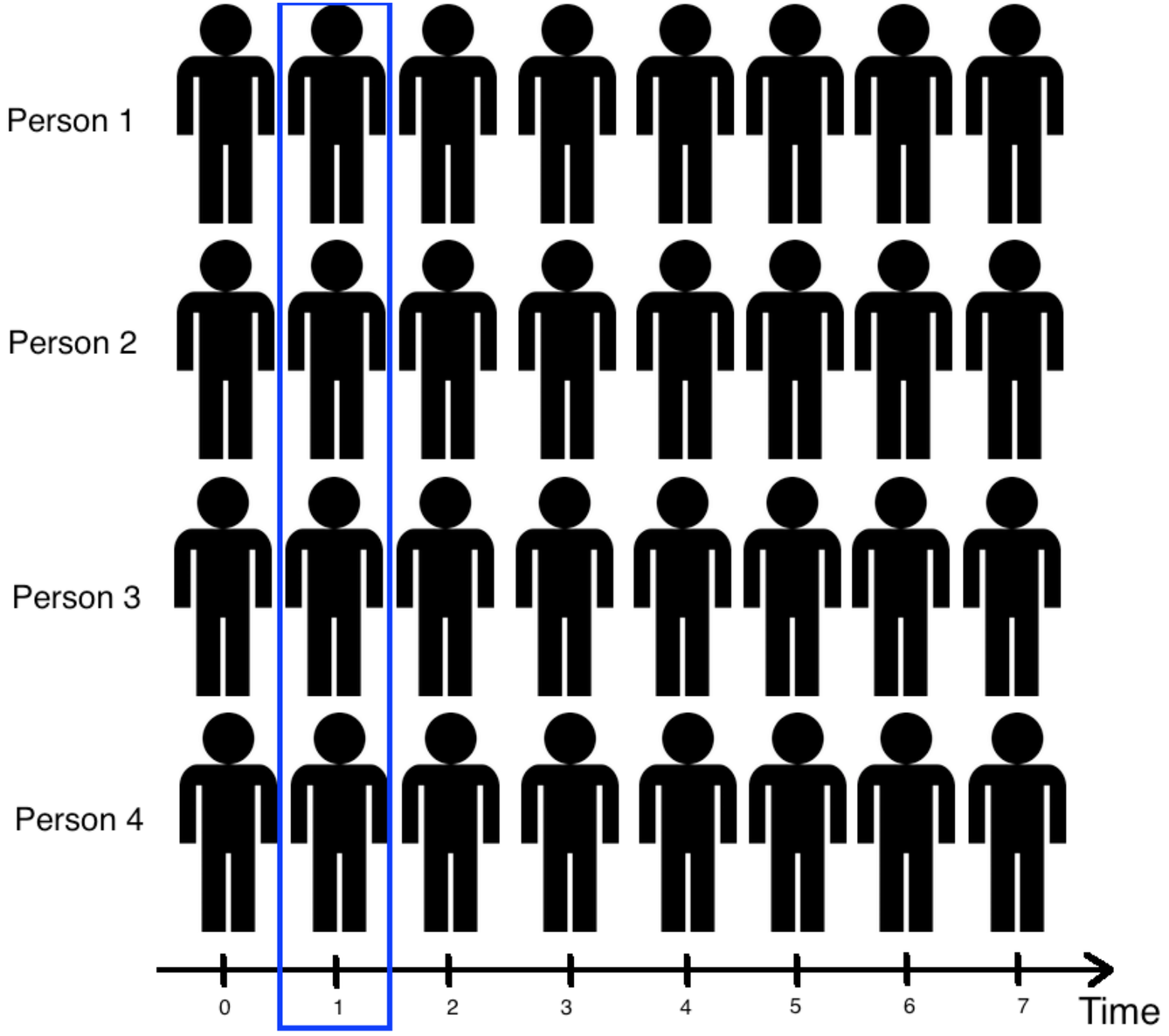
6 rows × 4 columns

CROSS-SECTIONAL DATA

- Strengths
  - Often comprehensive population based
    - Generalizability
  - Reduce cost compared to other types of data collection methods

- Weaknesses
  - Separation of cause and effect may be difficult (or impossible)
  - Cases with long duration or outliers can be over-represented

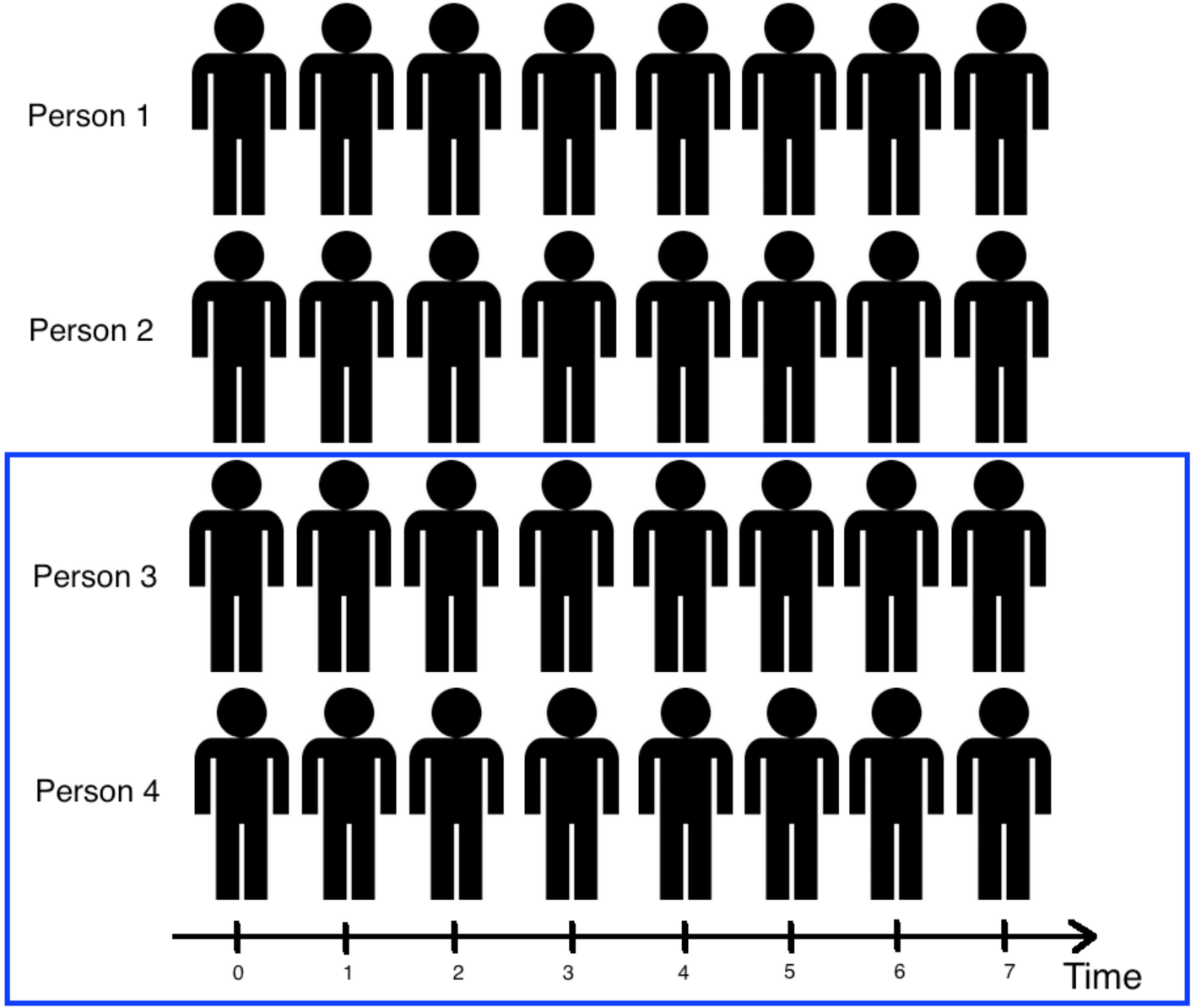
CROSS-SECTIONAL DATA



TIME SERIES/LONGITUDINAL DATA

- The information is collected over a period of time
- Strengths
  - Unambiguous temporal sequence - exposure precedes outcome
  - Multiple outcomes can be measured
- Weaknesses
  - Expense
  - Takes a long time to collect data
  - Vulnerable to missing data

TIME SERIES/LONGITUDINAL DATA



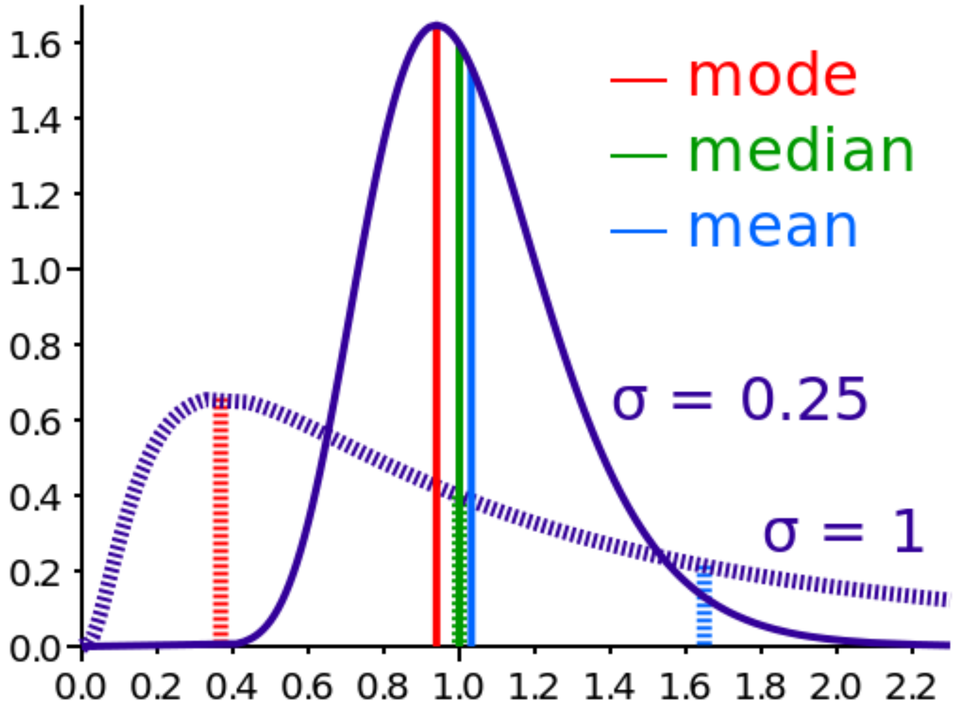
Statistics

- Mean
- Median
- Mode
- Max
- Min
- Quartile
- Interquartile Range
- Variance
- Standard Deviation
- Correlation

MEAN

- The mean of a set of values is the sum of the values divided by the number of values. It is also called the average.

$$\overline{X} = \frac{\sum X}{N}$$



MEAN EXAMPLE

- Find the mean of 19, 13, 15, 25, and 18.

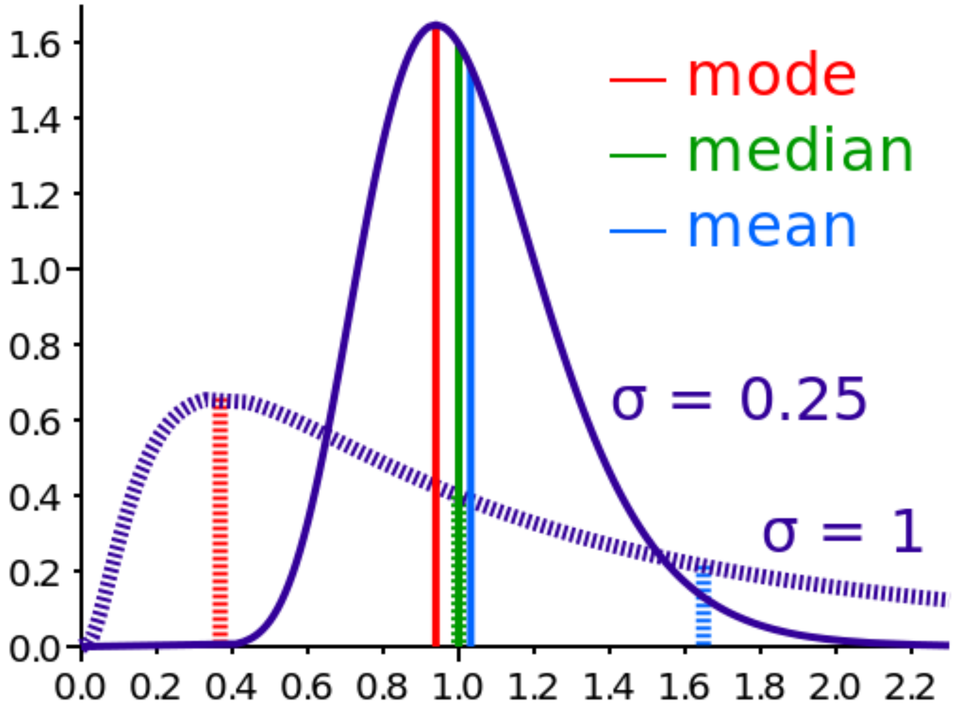
MEAN EXAMPLE

- Find the mean of 19, 13, 15, 25, and 18.

$$\frac{19 + 13 + 15 + 25 + 18}{5} = \frac{90}{5} = 18$$

MEDIAN

- The median refers to the midpoint in a series of numbers.
- To find the median
  - Arrange the numbers in order smallest to largest.
  - If there is an odd number of values, the middle value is the median.
  - If there is an even number of values, the average of the middle two values is the median.



MEDIAN EXAMPLE

- Find the median of 19, 29, 36, 15, and 20.

MEDIAN EXAMPLE

- Find the median of 19, 29, 36, 15, and 20.

Ordered Values:

15, 19, 20, 29, 36

20 is the median

MEDIAN EXAMPLE

- Find the median of 67, 28, 92, 37, 81, 75.

MEDIAN EXAMPLE

- Find the median of 67, 28, 92, 37, 81, 75.

Ordered Values:

28, 37, 67, 75, 81, 92

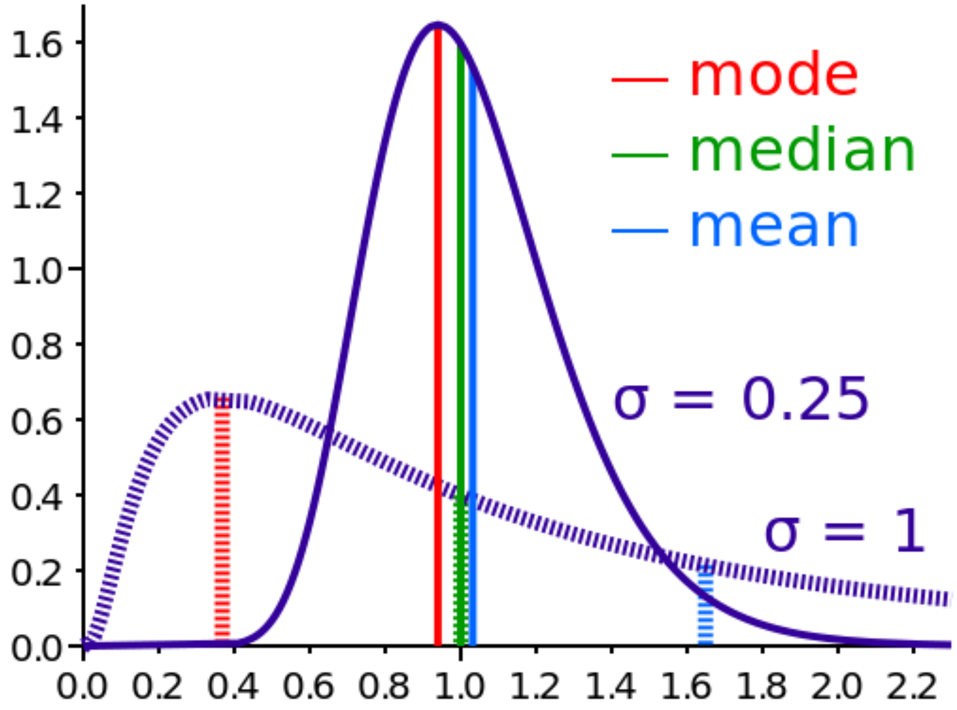
67 and 75 are the middle values.

$$\frac{67 + 75}{2} = \frac{142}{2} = 71$$

71 is the median.

MODE

- The mode of a set of values is the value that occurs most often.
- A set of values may have more than one mode or no mode.



MODE EXAMPLE

- Find the mode of 15, 21, 26, 25, 21, 23, 28, and 21.

MODE EXAMPLE

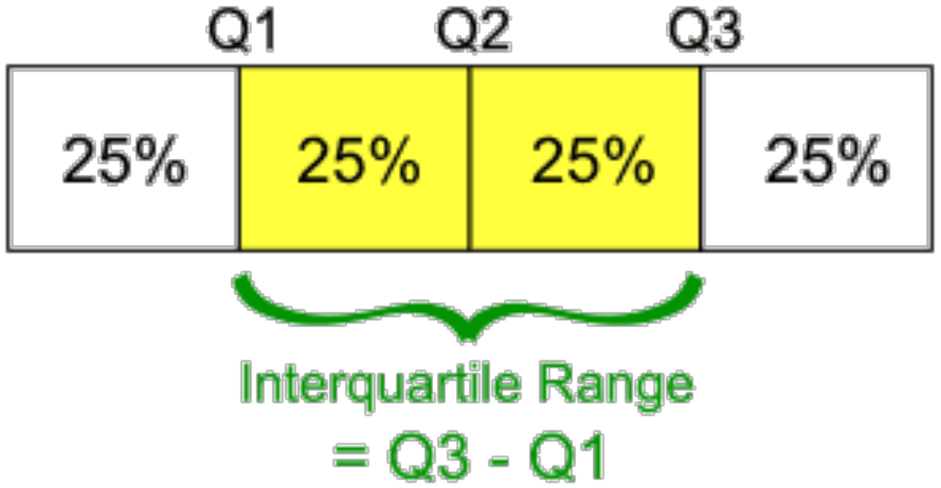
- Find the mode of 12, 15, 18, 26, 15, 9, 12, and 27.

MODE EXAMPLE

- Find the mode of 4, 8, 15, 21, and 23.

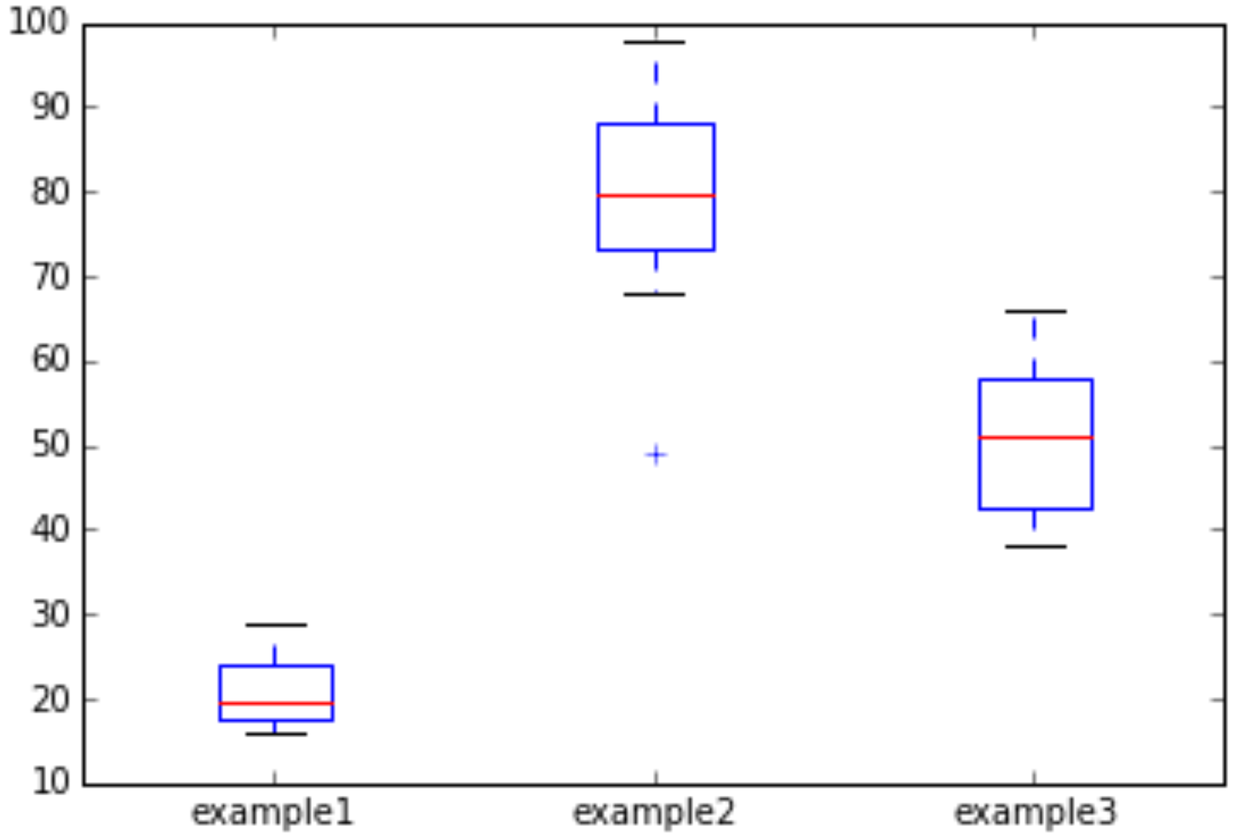
QUARTILES AND INTERQUARTILE RANGE

- Quartiles divide a rank-ordered data set into four equal parts.
- The values that divide each part are called first, second, and third quartiles, denoted Q1, Q2, and Q3, respectively.
- The interquartile range (IQR) is Q3 - Q1, a measure of variability (assuming relative normality).



CODEALONG PART 2: BOX PLOT

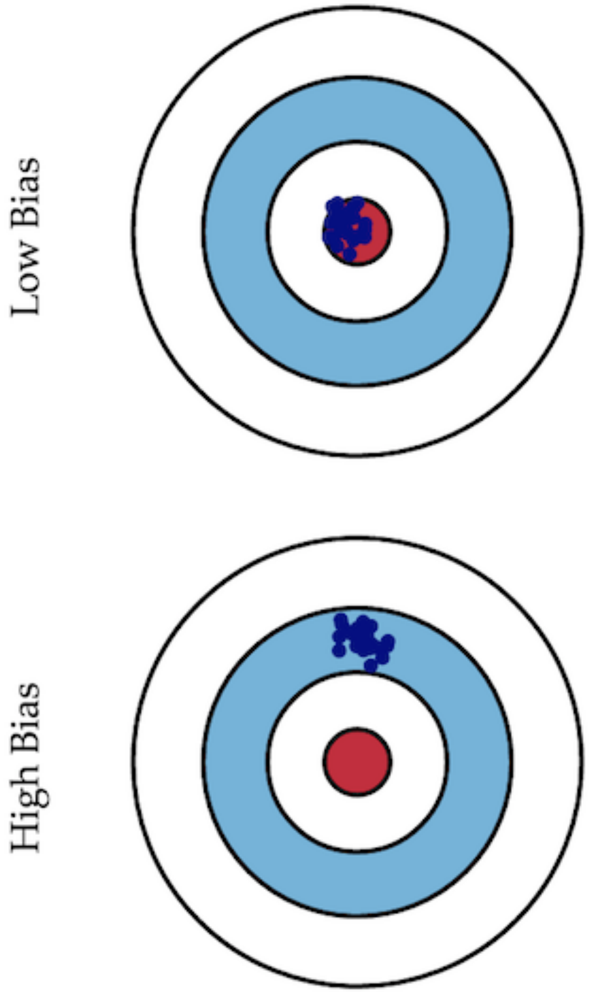
- Box plots give a nice visual of min, max, median, and the quartile and interquartile range.



BIAS VS. VARIANCE

- Error due to **bias** is calculated at the difference between the *expected prediction* of our model and the *correct value* we are trying to predict.

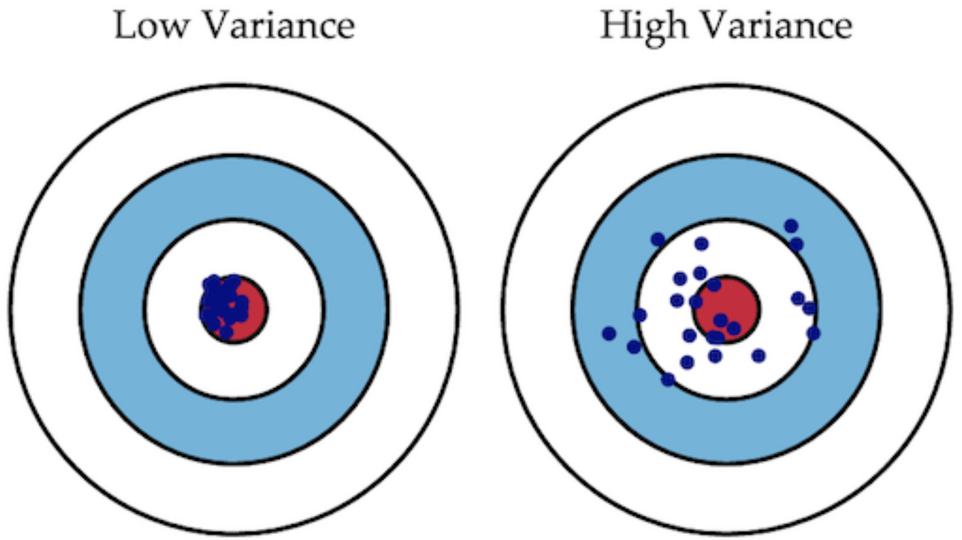
- Imagine creating multiple models on various datasets. **Bias** measures *how far off in general* models' predictions are from the correct value.



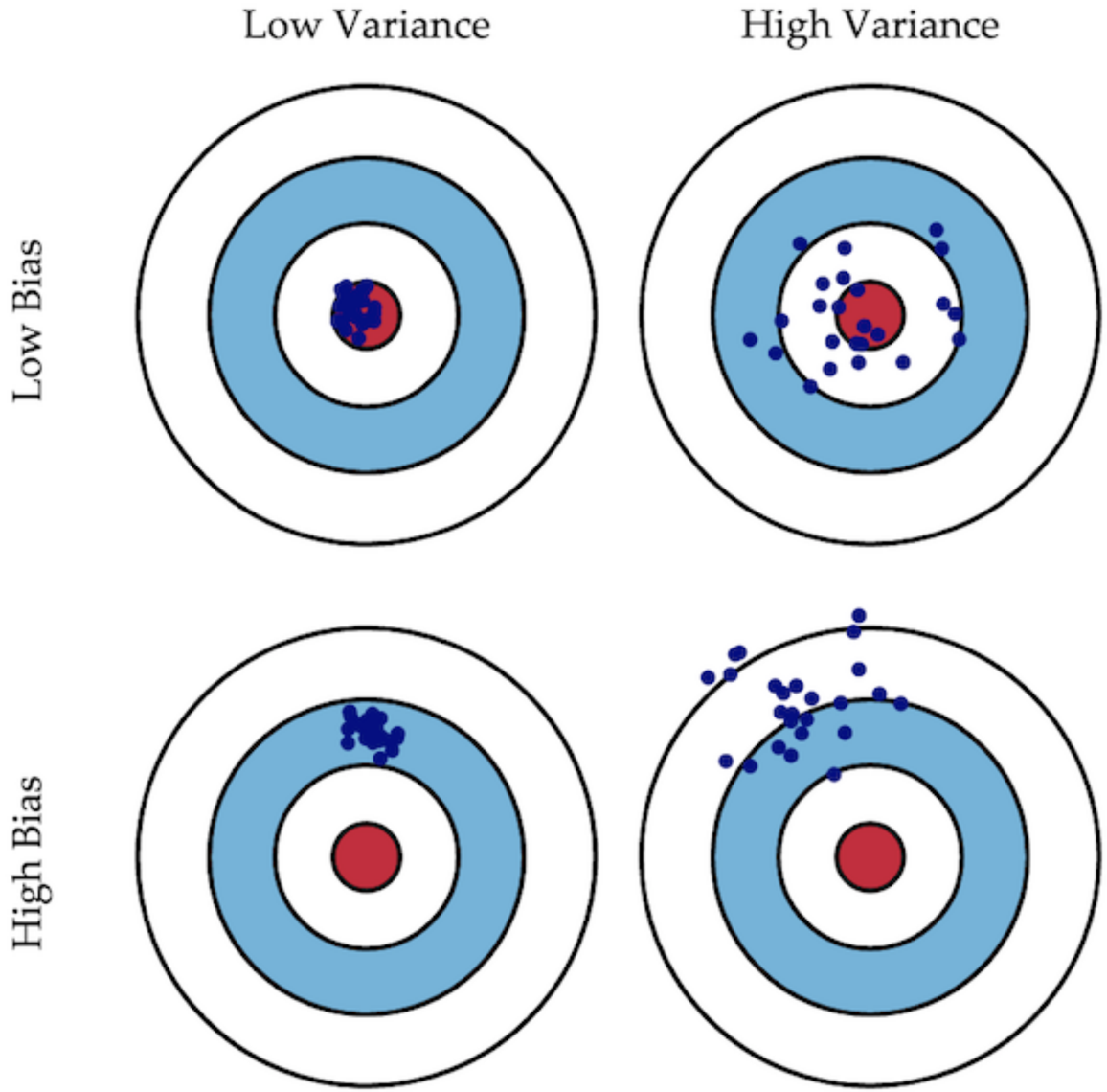
BIAS VS. VARIANCE

- Error due to **variance** is taken as the variability of a model prediction for a given point.

- Imagine creating multiple models on various datasets. The **variance** is *how much the predictions for a given point vary* between different realizations of the model.



BIAS VS. VARIANCE



STANDARD DEVIATION

- Standard deviation (SD, σ for population, s for sample) is a measure that is used to quantify the amount of variation or dispersion of a set of data values.

- Standard deviation is the square root of variance.

$$s = \sqrt{\frac{\sum(x - \bar{x})^2}{n - 1}}$$

STANDARD ERROR

- The standard error of the mean (SEM) quantifies the precision of the mean.
- It is a measure of how far your sample mean is likely to be from the true population mean.
- It generally increases with the size of an estimate, meaning a large standard error may not indicate the estimate of the mean is unreliable.
- It’s often better to compare the error in relation to the size of the estimate.

STANDARD ERROR

$$SE_{\bar{x}} = \frac{s}{\sqrt{n}}$$

CODEALONG PART 3: STANDARD DEVIATION & VARIANCE

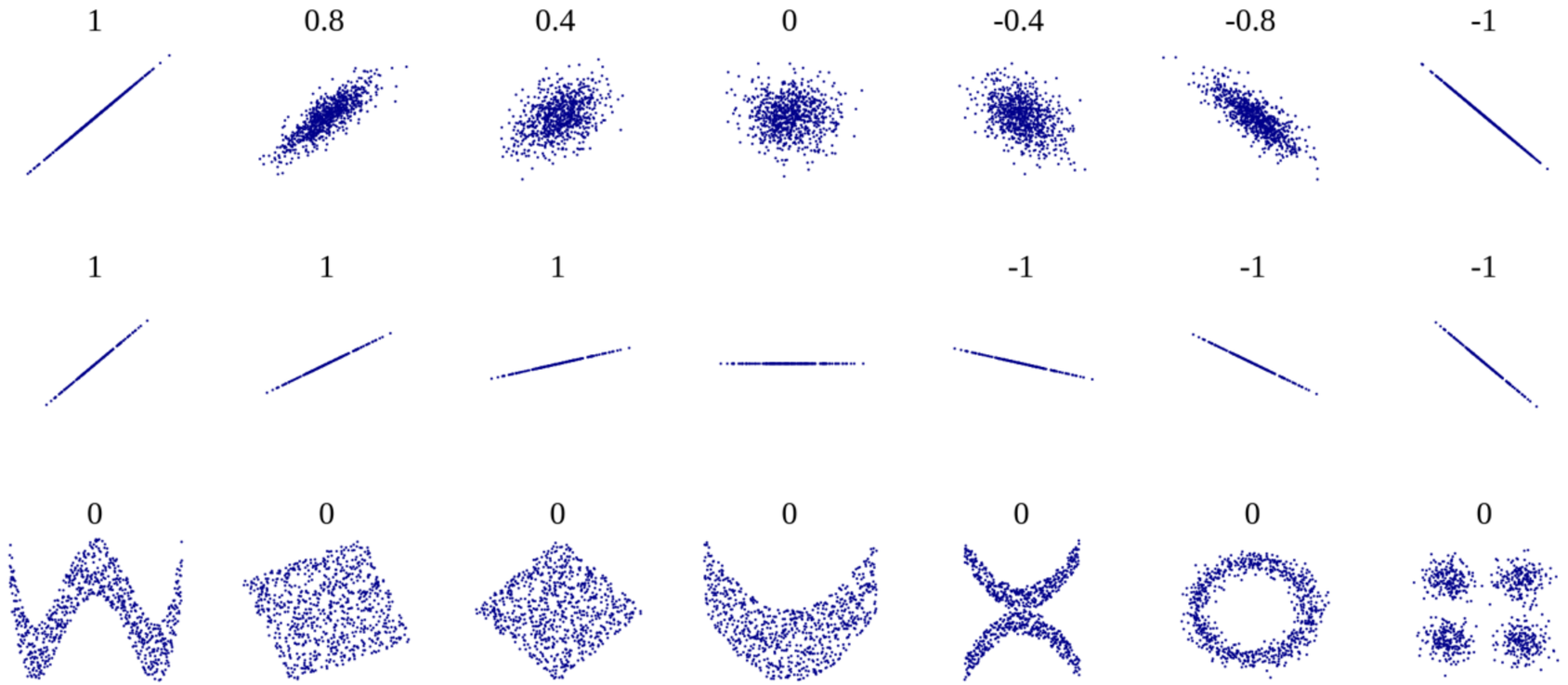
- You can calculate variance and standard deviation easily in Pandas.

```
Methods include:
.std() - Compute Standard Deviation
.var() - Compute variance
.describe() - short cut that prints out count, mean, std, min,
quartiles, max
```

CORRELATION

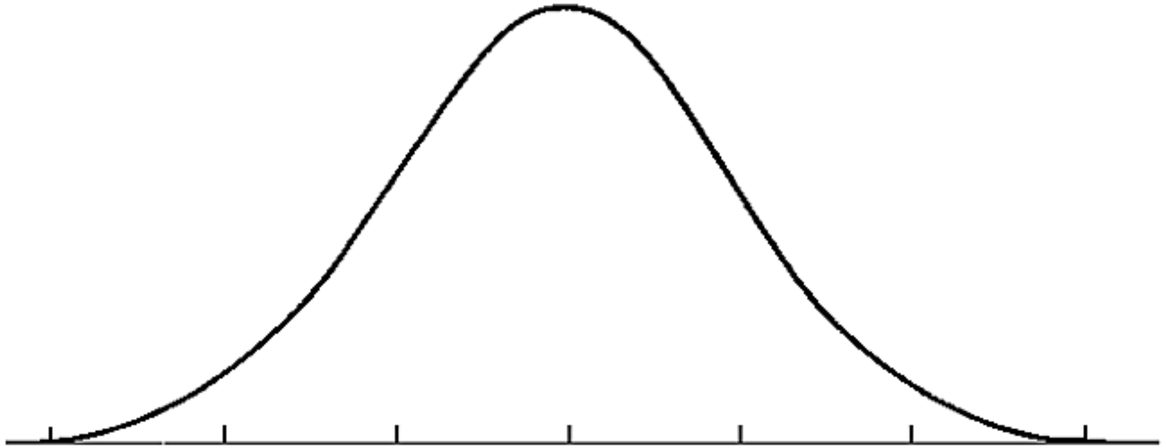
- The correlation measures the extent of interdependence of variable quantities.

- Example correlation values



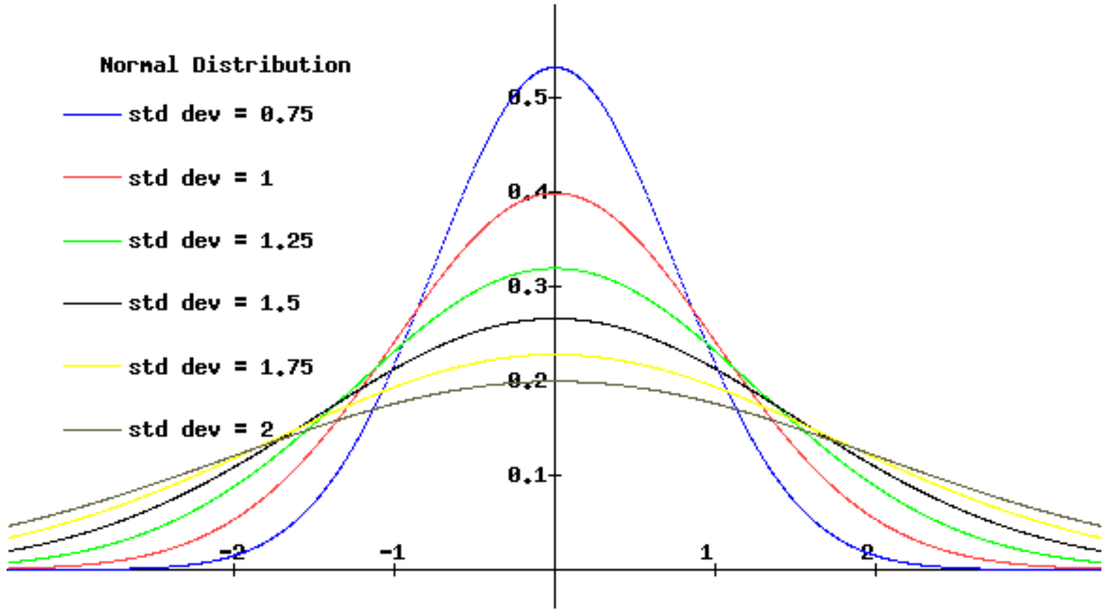
THE NORMAL DISTRIBUTION

- A normal distribution is often a key assumption to many models.
- The normal distribution depends upon the *mean* and the *standard deviation*.
- The *mean* determines the center of the distribution. The *standard deviation* determines the height and width of the distribution.



THE NORMAL DISTRIBUTION

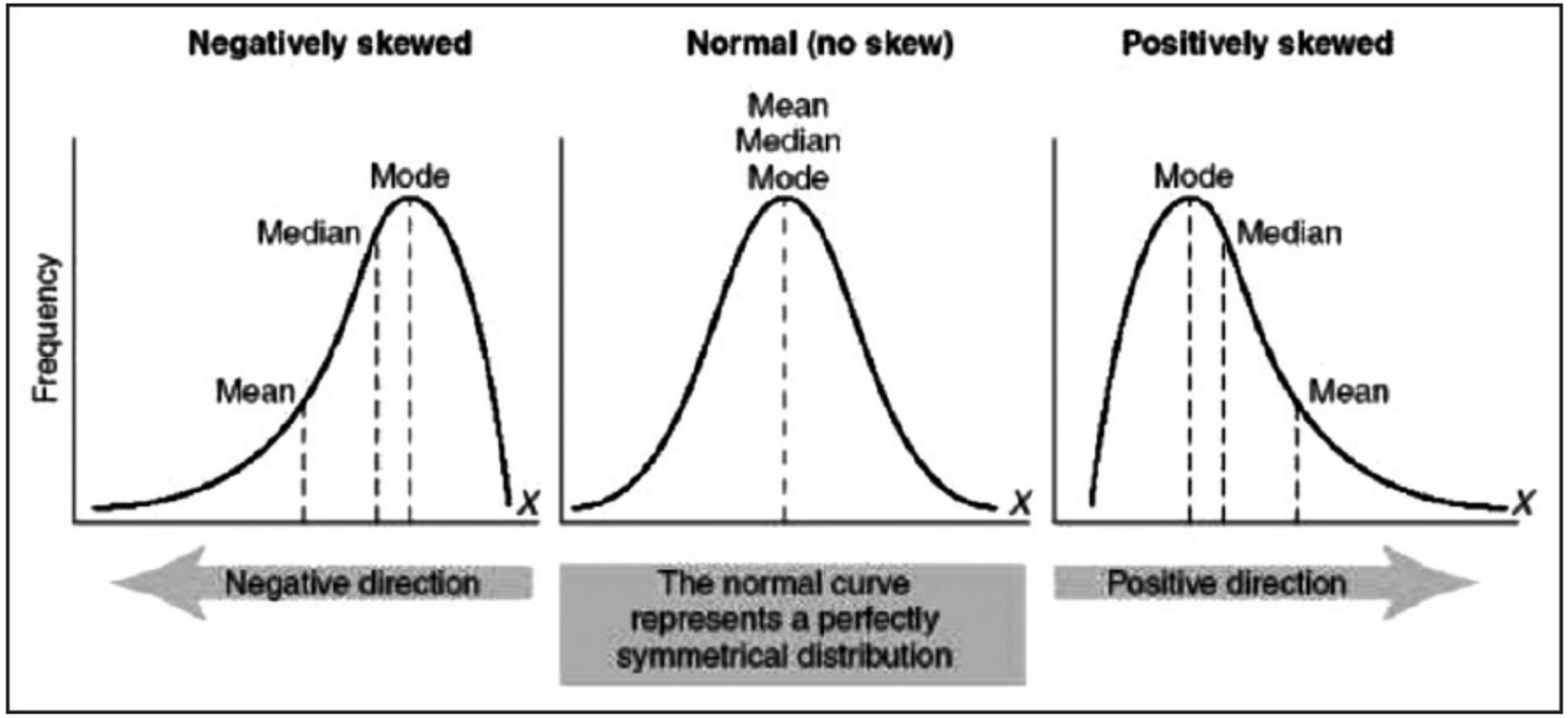
- Normal distributions are symmetric, bell-shaped curves.
- When the standard deviation is large, the curve is short and wide.
- When the standard deviation is small, the curve it tall and narrow.



SKEWNESS

- Skewness is a measure of the asymmetry of the distribution of a random variable about its mean.

- Skewness can be positive or negative, or even undefined.



KURTOSIS

- Kurtosis is a measure of whether the data are peaked or flat relative to a normal distribution.

- Datasets with high kurtosis tend to have a distinct peak near the mean, decline rather rapidly, and have heavy tails.