Introduction to the Data Analytics

Statistical Toolbox

Summary

- **Concepts (25 mins)**
  - Where Most Data Analysis Goes Wrong
- **Tools (25 mins)**
  - Purpose
  - Examples
- **Questions (10 mins)**

Foundational Statistics

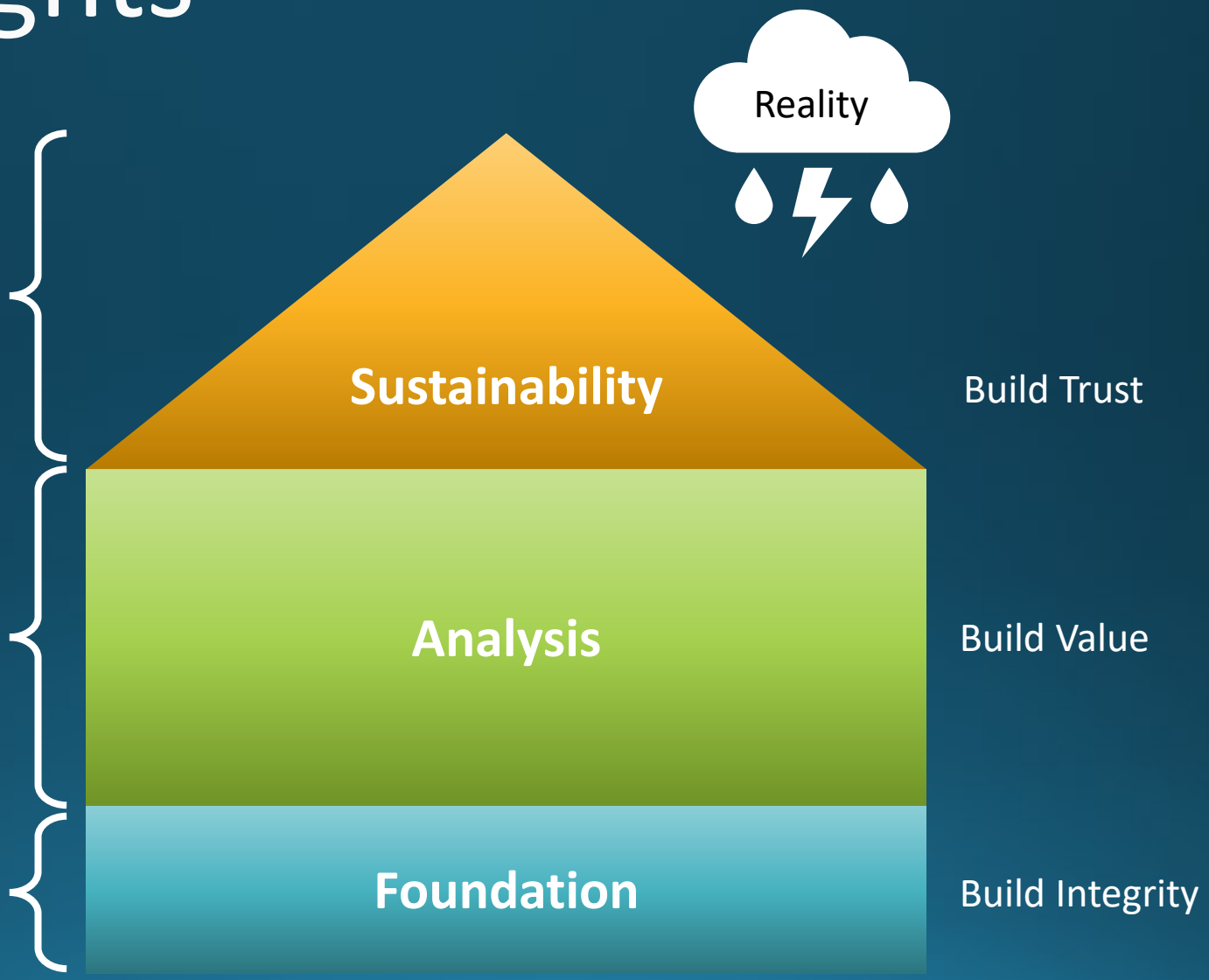
Concepts

Building Insights

Defensible Conclusions
Data Misinterpretation
Data Overinterpretation
Required Maintenance

Analysis is Responsive to Goals
Analysis adds to Confidence
Analysis adds to Knowledge
Foresight beats Hindsight

Consider Prior Knowledge
Have Clear Goals
Make Reasonable Assumptions



Build Trust

Build Value

Build Integrity

Statistics as a Toolbox

- **What is Statistics?**
  - Practice or science of collecting and analyzing numerical data.
- **Why is it Important?**
  - Used to make inferences and understand uncertainty.

Numbers

- **Continuous**
  - Unlimited Possible Values/Outcomes
    - What is the temperature outside? 74.4°F, 86.32539948°C
    - How tall are you? 71.4 in, 182.883 cm
- **Discrete**
  - Limited Possible Values/Outcomes
    - How many children do you have? 0, 1, 2, 3
    - What size shoe do you wear? 8, 8.5, 9, 9.5

Data vs. Process

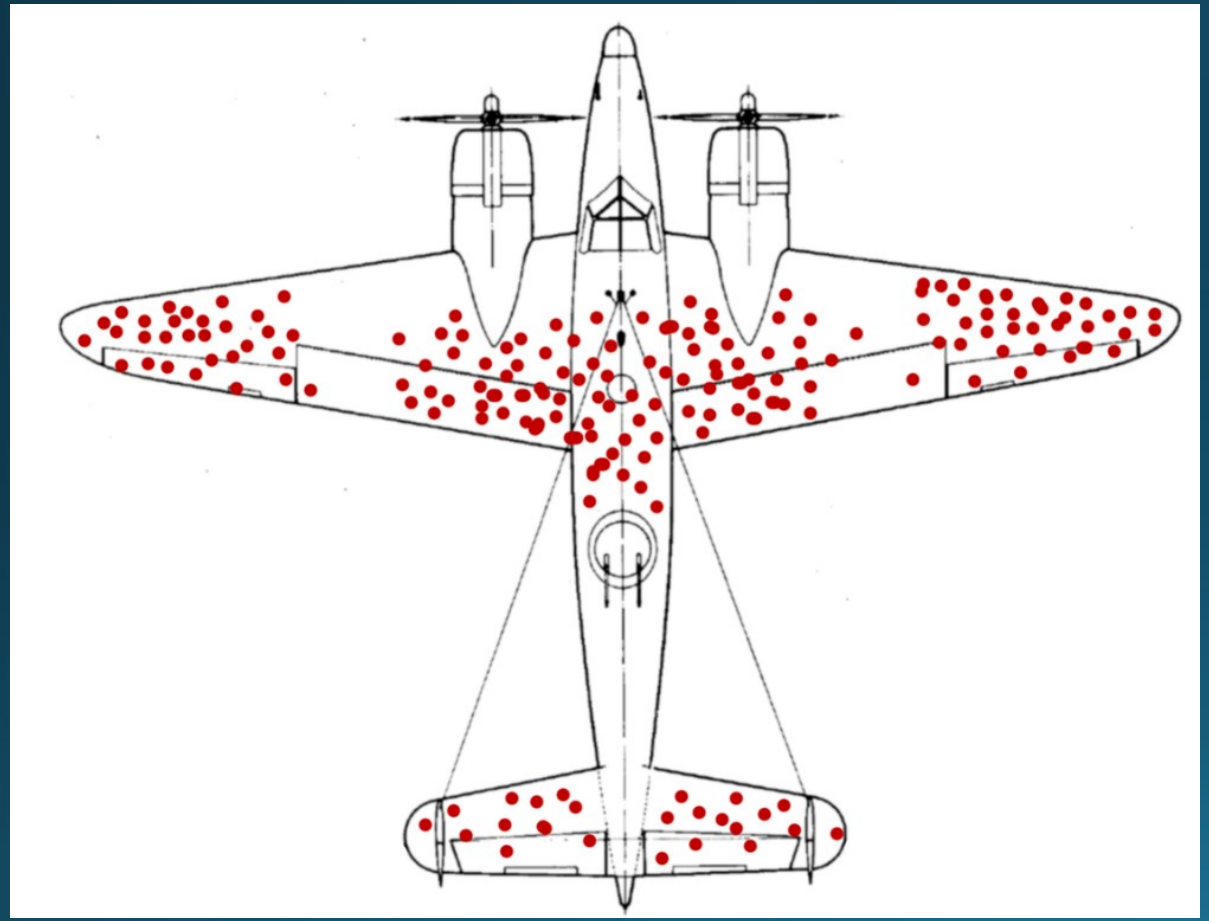
- What are **Data**?
  - Collection of observations (not facts).
  - Data are subject to imperfect observers and instrumentation.
  - Outcomes of one or more processes.
- What is a **Process**?
  - Actions, phenomena, and/or mechanisms that lead to creation, collection, or generation of data.

- Data is inherently tied to processes that generated it.
- Data reflects (or is result of) underlying processes.

Sample vs. Population

- What is a **Sample**?
  - I have SOME observations.
  - In most cases, we work with samples.
- What is a **Population**?
  - I have ALL the observations.
  - In rare cases, we work with populations.
- Often depends on how you frame a question.
  - Of all reported county incidents in 2020, how many were heart attacks?
  - How many heart attacks happened in the county this year?

Summarize With An Example



Consider:

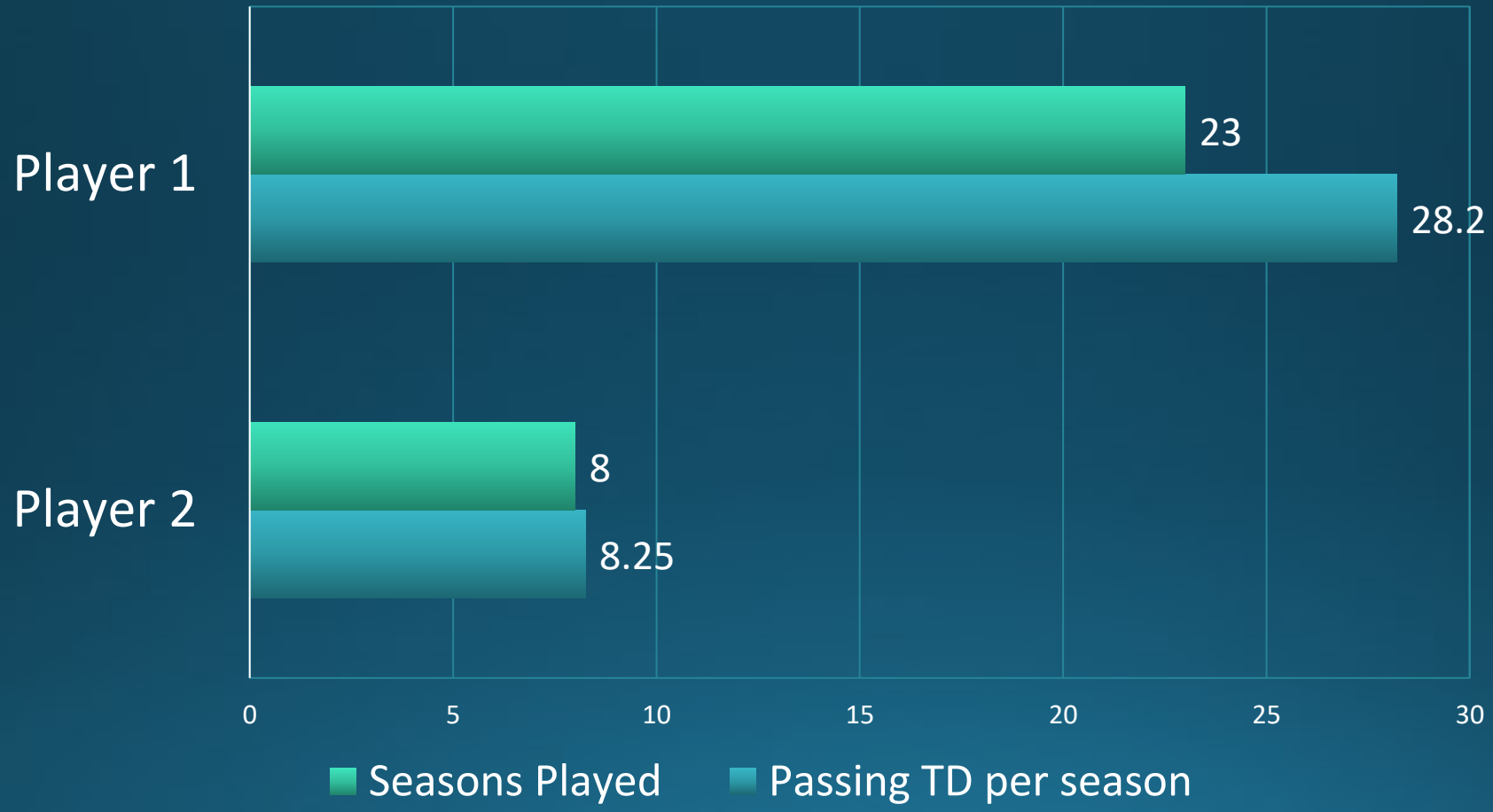
- Sample vs. Population
- Data vs. Process

Process Stationarity

- ✓Required assumption for most statistical tools
- ✓Property of the process generating the data
- ✓Depends on how data are grouped.
- ✓Depends on how data are normalized (rescaled).
- ✓Exists on a spectrum (Strong to Weak to Non-stationary).

- Not a characteristic of the data.
- There is no test for process stationarity.

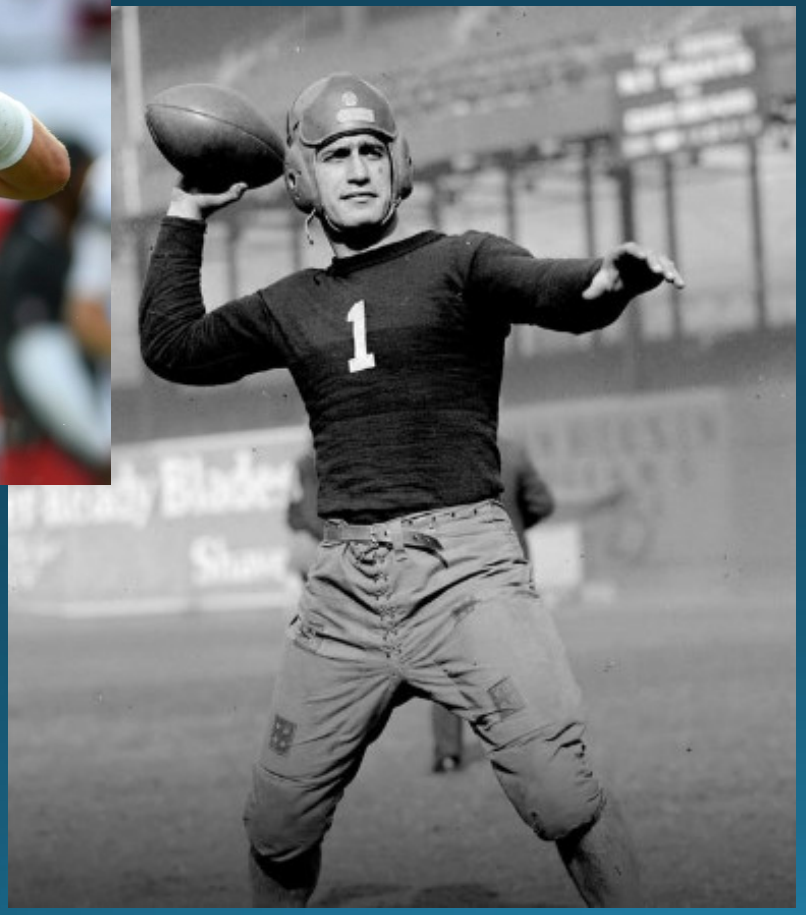
Who is the Better Quarterback?



Consider the Process

Tom Brady (2000-2022)

Benny Friedman (1927-1934)



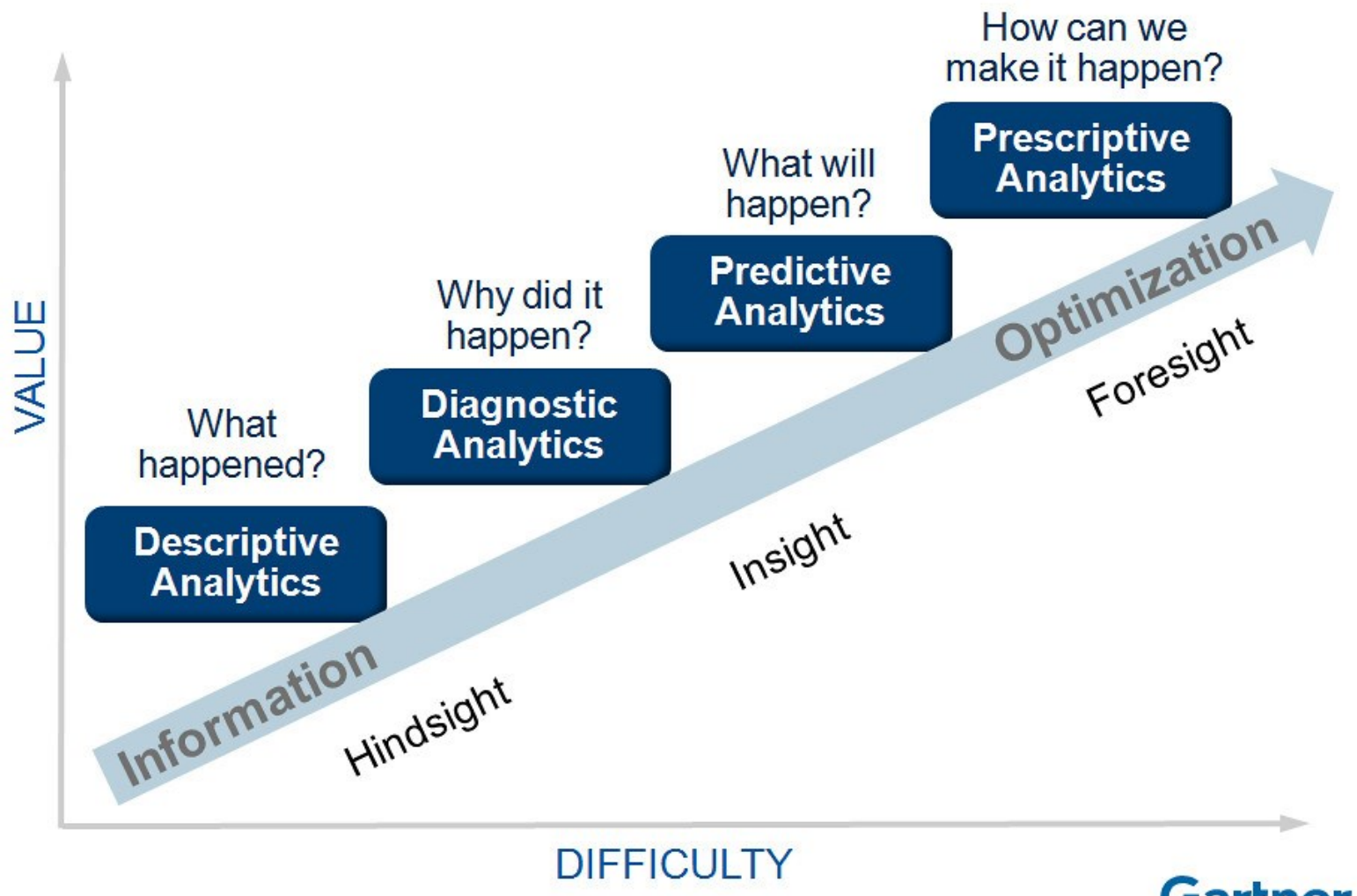
- Process Factors
  - Length of Season
  - Rules / Penalties
  - Pads / Helmet
  - Offensive Line

Do the collective processes seem stationary over time?

Foundational Statistics

Tools

Analytic Value Escalator



Gartner

Examples

Descriptive	Diagnostic	Predictive	Prescriptive
Distributions	Correlation Analysis	Linear Regression	Optimization Algorithms
Histograms	Regression Analysis	Logistic Regression	Decision Analysis
Central Tendency Metrics	Chi-Square Test	Time Series Analysis	Simulation
Dispersion Metrics	Variance Analysis	Decision Trees	Markov Chains
Box Plots	Normality Tests	Random Forest	Game Theory
Z Score	Residual Analysis	Neural Networks	Genetic Algorithms
Standard Error	Homogeneity Tests	Geostatistics (e.g., Kriging)	Reinforcement Learning
		Bayes Theorem	Bayes Theorem

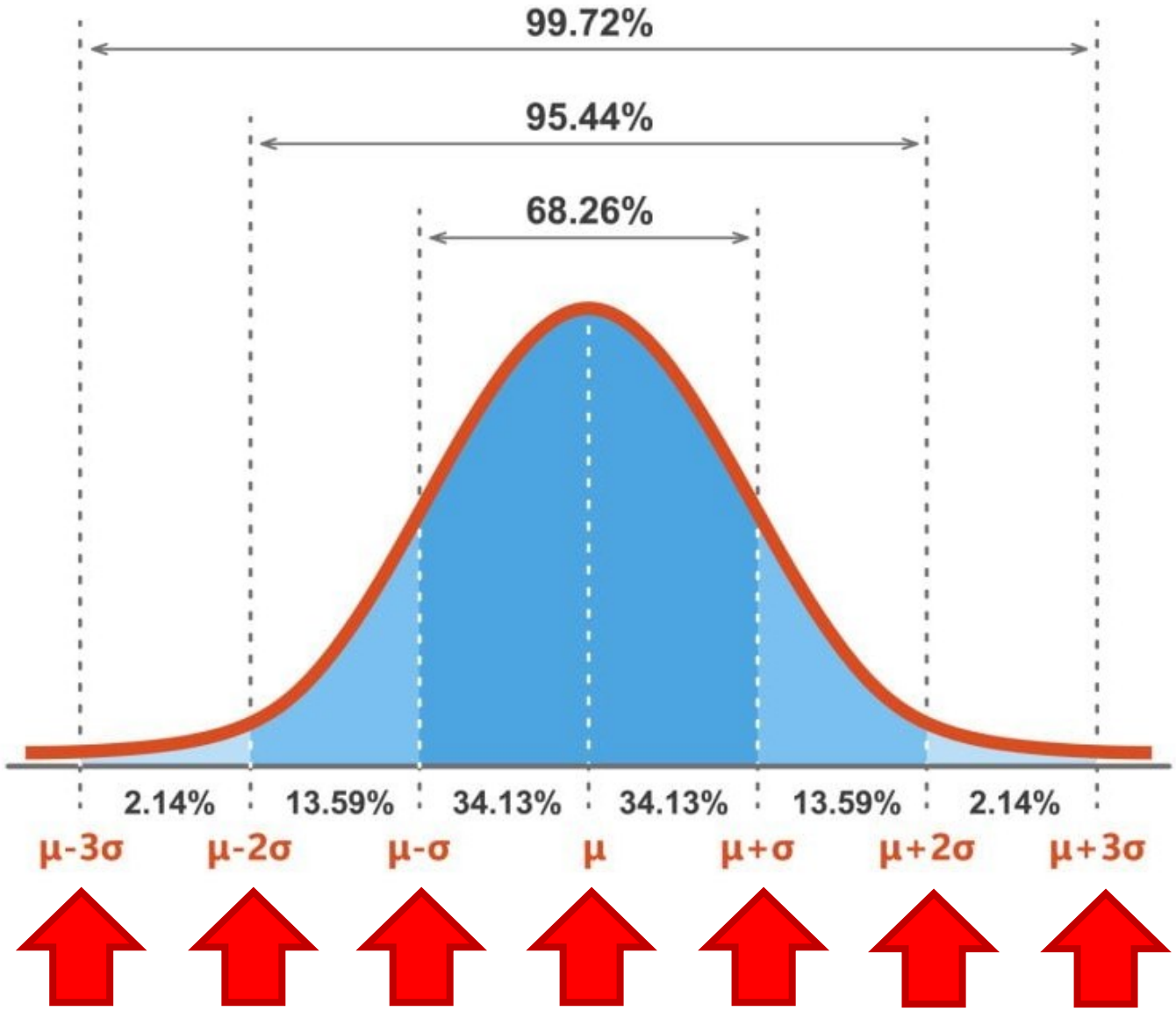
Distributions

Goal:

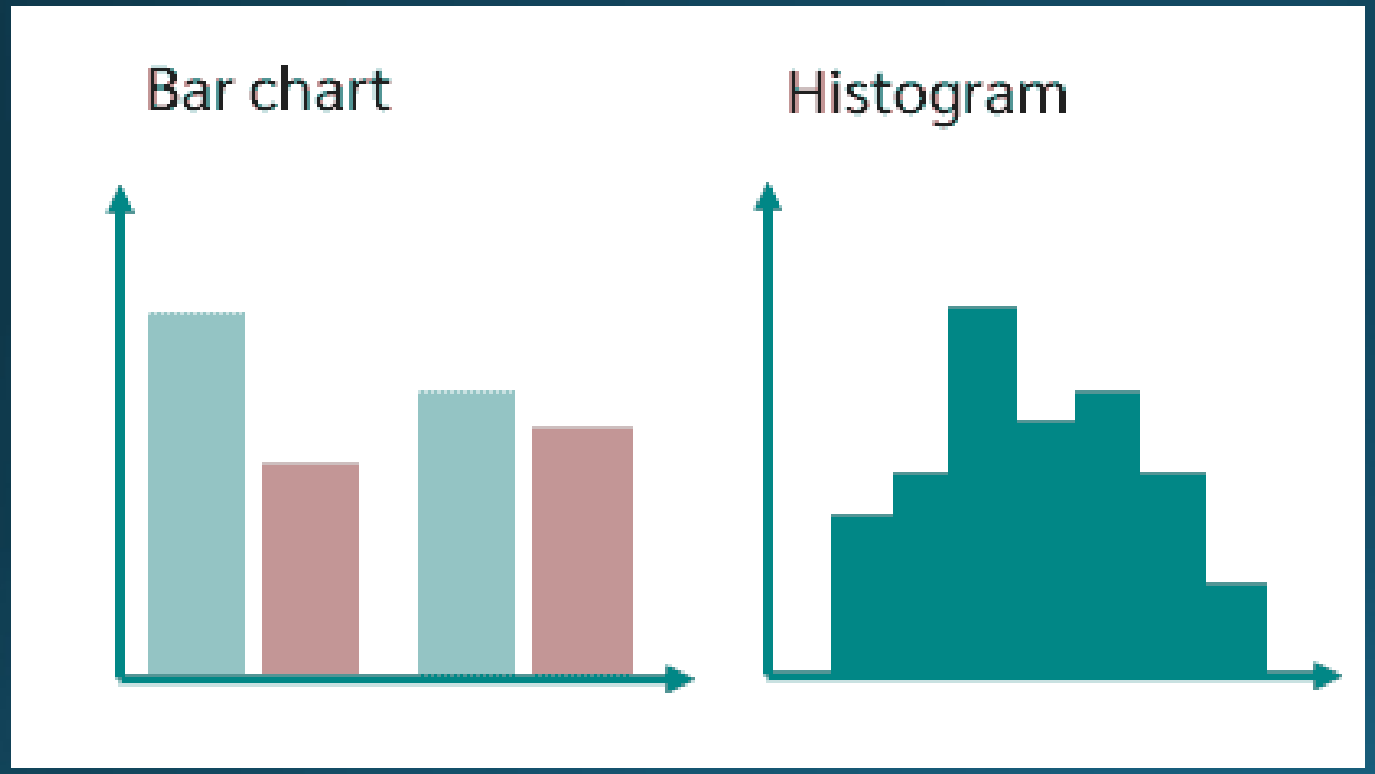
- Describe the **frequency** of values for one **continuous variable** within a data set.

μ = Mean

σ = Standard Deviation



Bar Charts vs. Histograms



Discrete Distributions

Proportional **Bar** Height

Continuous Distributions

Proportional **Bin** Area

Microsoft | Power BI

Cannot create Histograms.

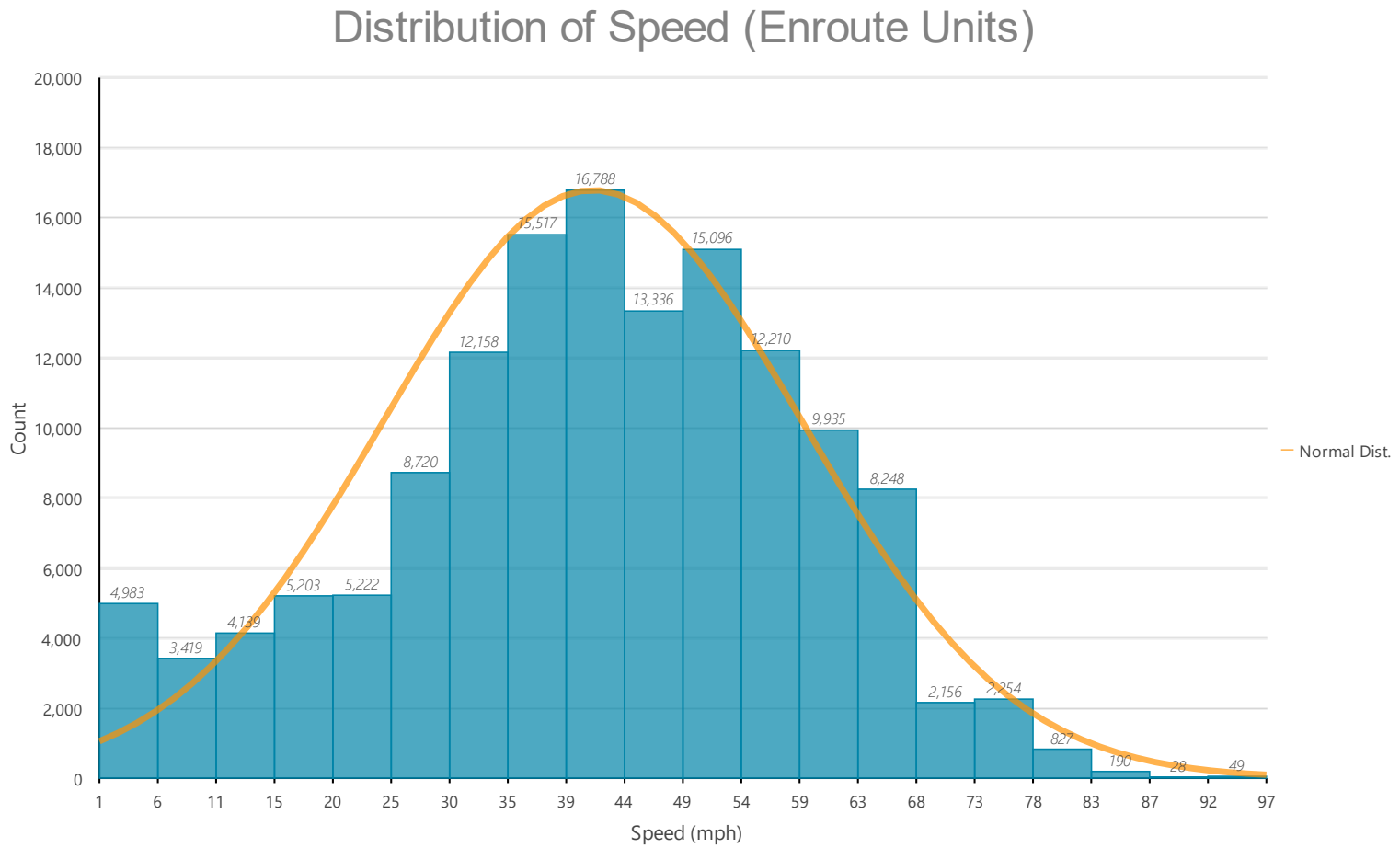
python™

Can do just about anything.

Example

Describe frequency of vehicle **speed** while engines are enroute to incidents.

2-weeks of data (April 2024).



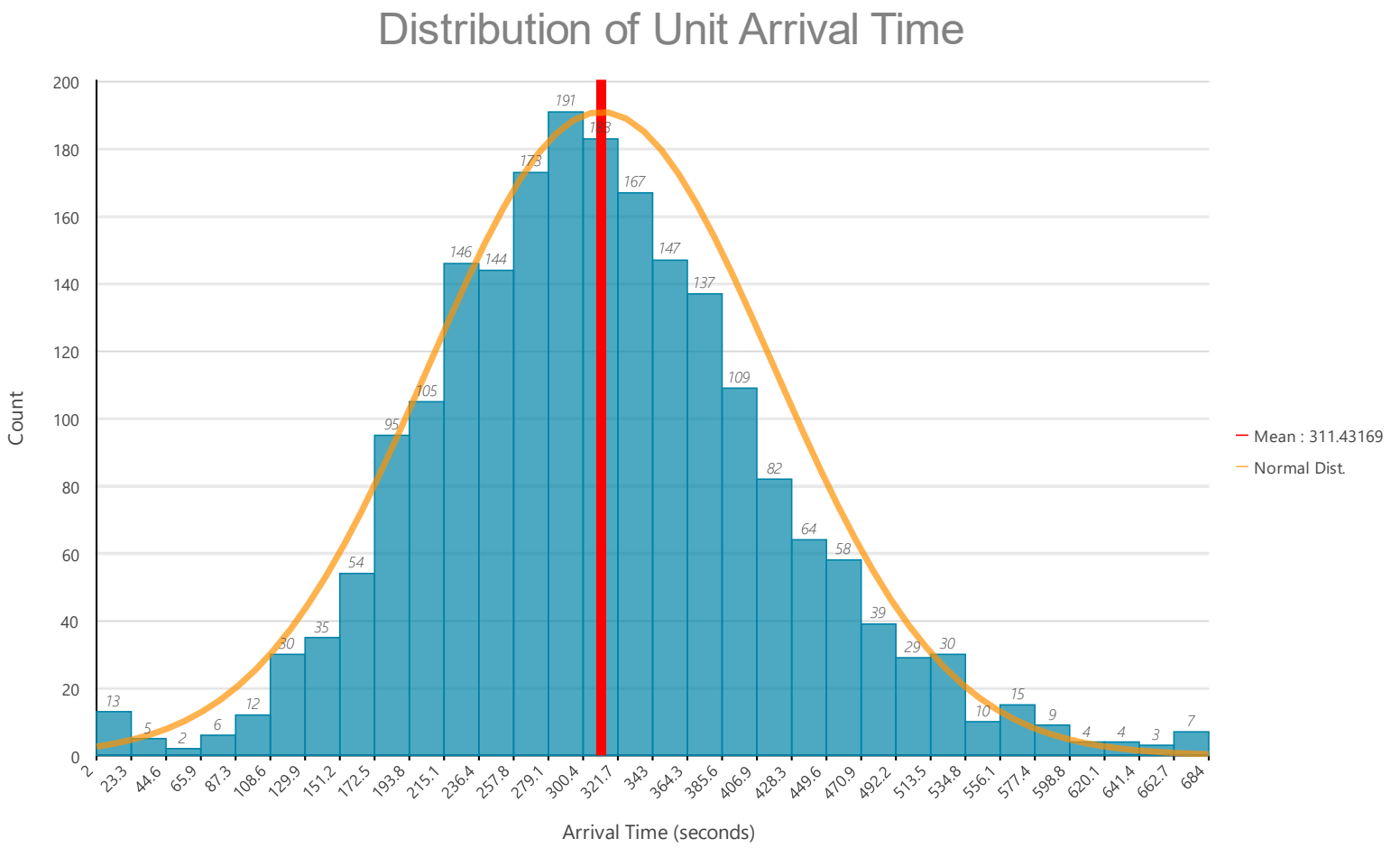
Central Tendency

- Mean, Median, Mode
- **Continuous Normal Distribution**: Mean
- **Continuous Skewed Distribution**: Median
- **Discrete Distribution**: Mode

Example

Describe frequency of unit **arrival time** to incidents.

Red bar = Mean

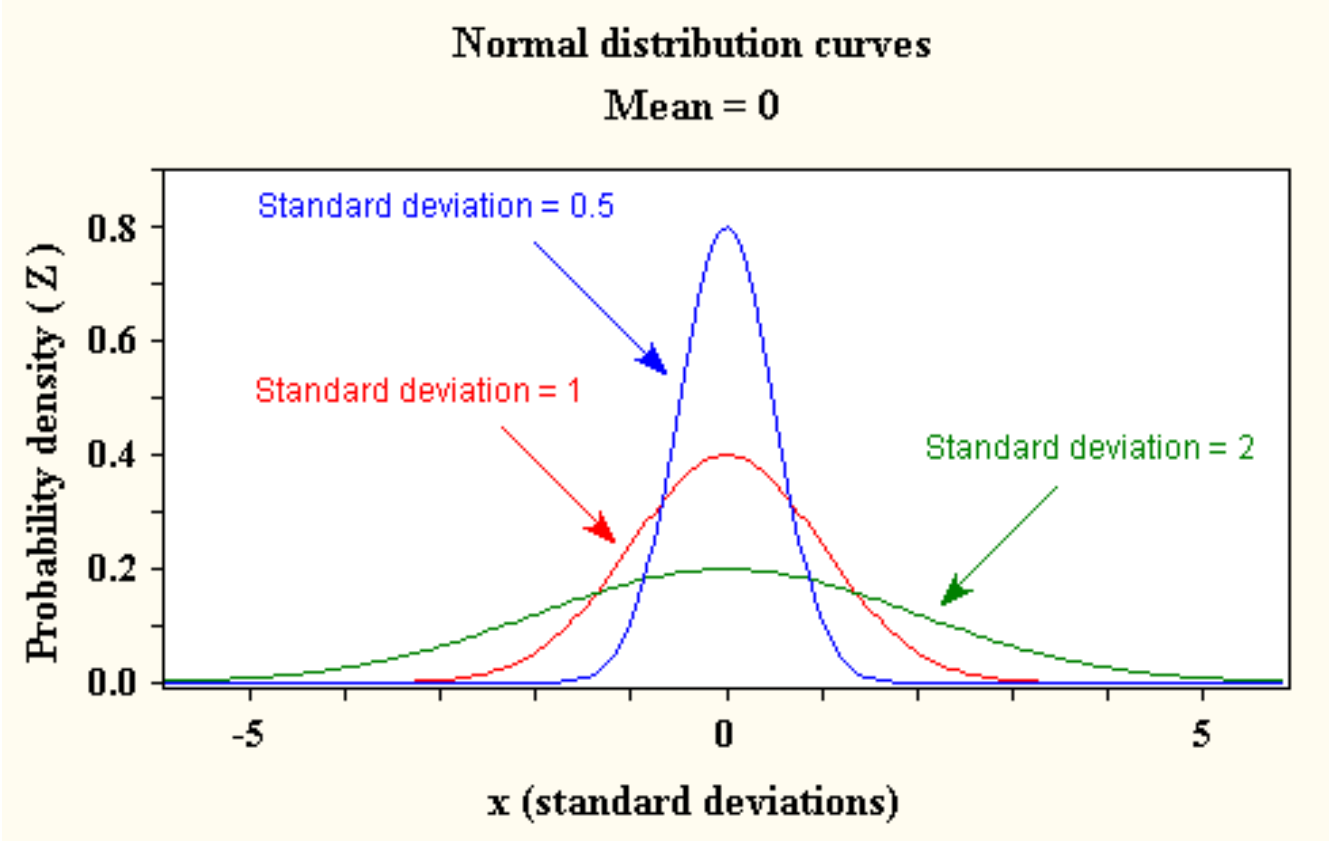


Distributions: 2 Distinct Types

- **Representation (Y-axis)**
  - Frequency: Count per bin
  - Relative Frequency: Count per bin ÷ Total Count
  - Probability: Continuous Function. No bins
- **Process Models**
  - Normal (Gaussian)
  - Chi-square
  - Binomial
  - Poisson
  - Uniform
  - Log-normal

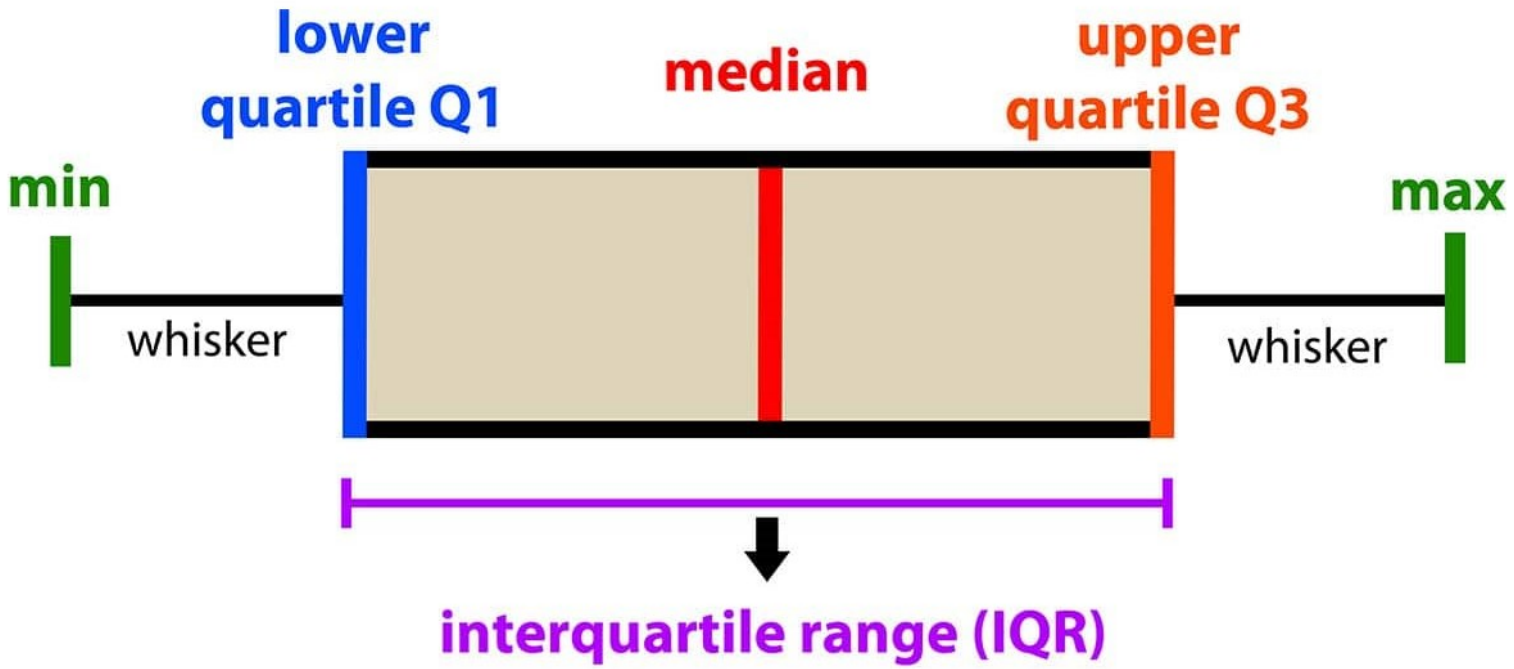
Measures of Dispersion

- Range
  - Minimum Value
  - Maximum Value
- Standard Deviation
- Variance

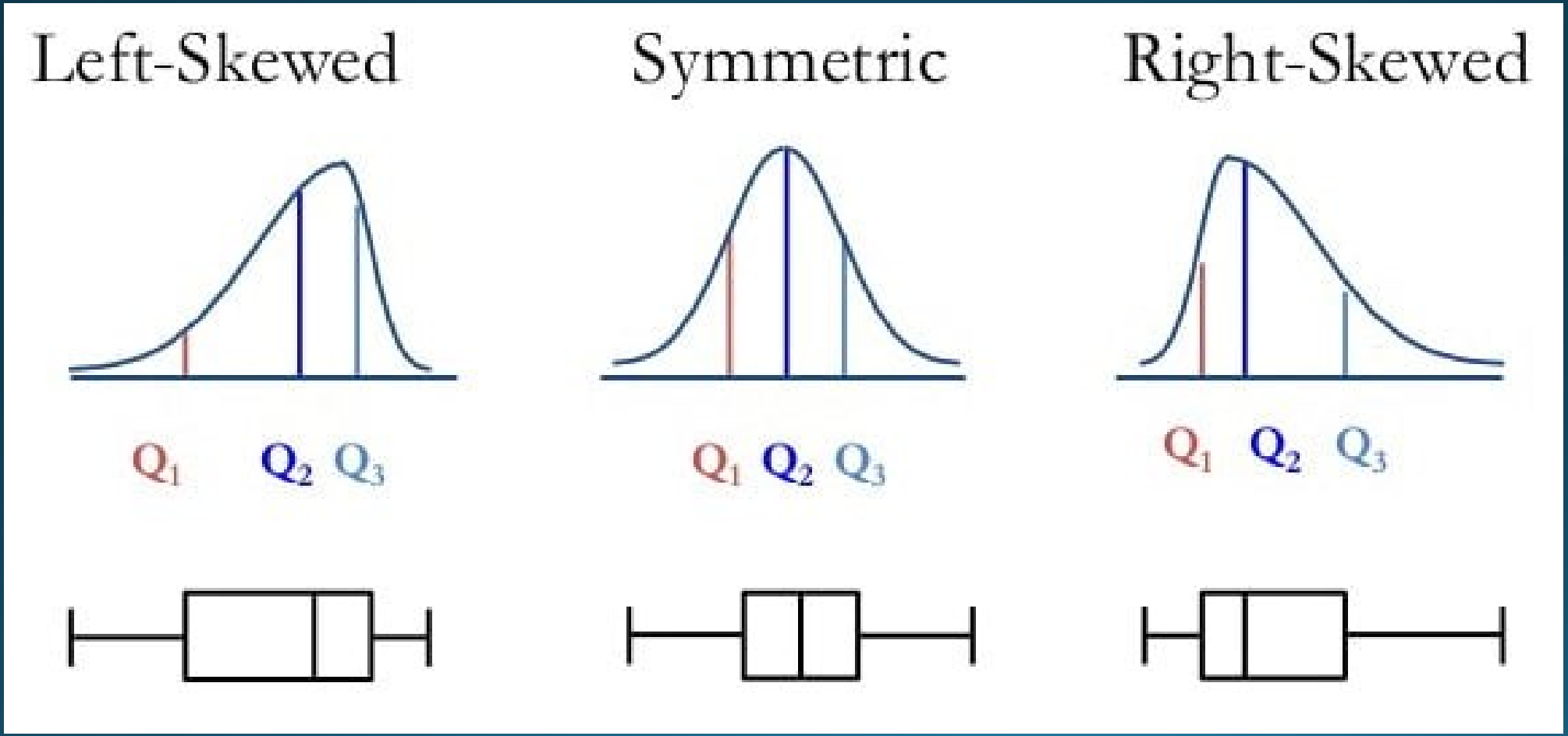


Box Plots

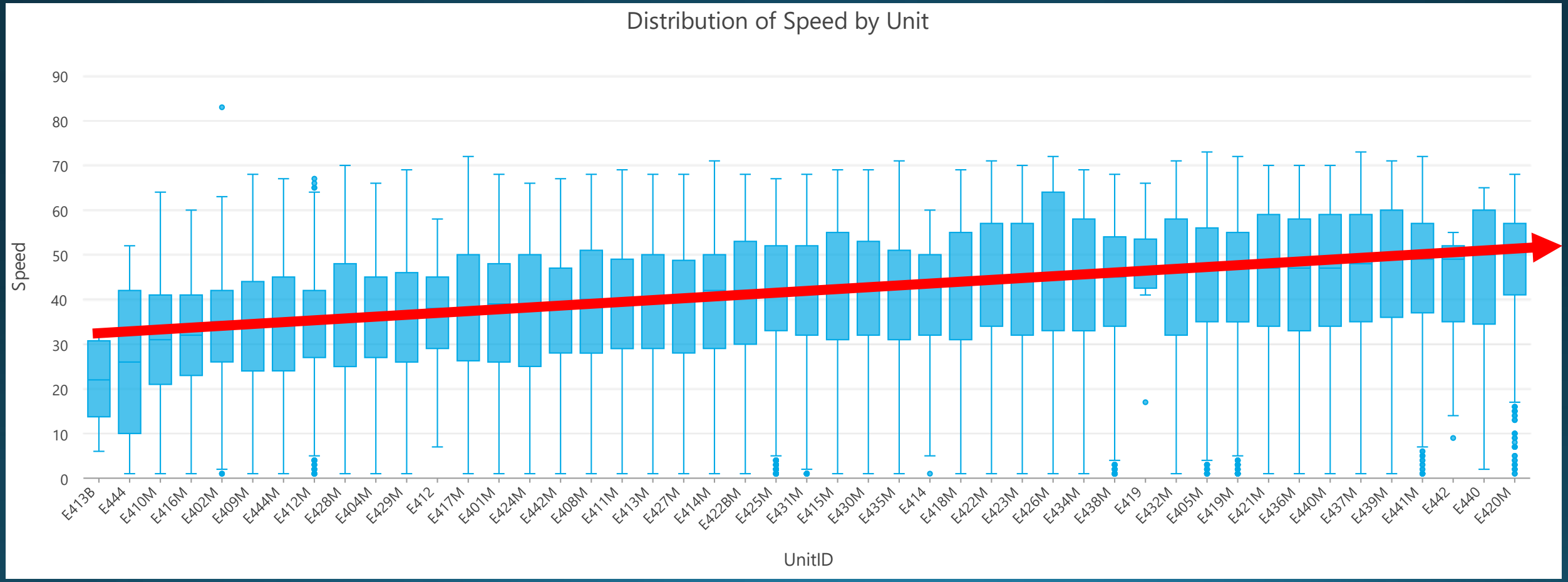
- Useful for quick description about a distribution.
- Useful for comparing different distributions.



Distribution Skew



Consider the Process



$$\mathrm{SE} = \frac{\sigma}{\sqrt{n}}$$

SE = standard error of the sample

σ = sample standard deviation

n = number of samples

Standard Error

- Measure of variability of the sample mean as an estimate of the population mean.
- Quantifies how much sample mean is likely to vary from the true population mean if multiple samples were taken from same population.
- Measure of Data's Uncertainty

Z Score

- Measurement that describes each value (X) related to the sample mean.
- Units of Standard Deviation (σ).
- Purpose:
  - Standardizing data for comparison.
  - Great to evaluate outliers

$$z = \frac{X - \mu}{\sigma}$$