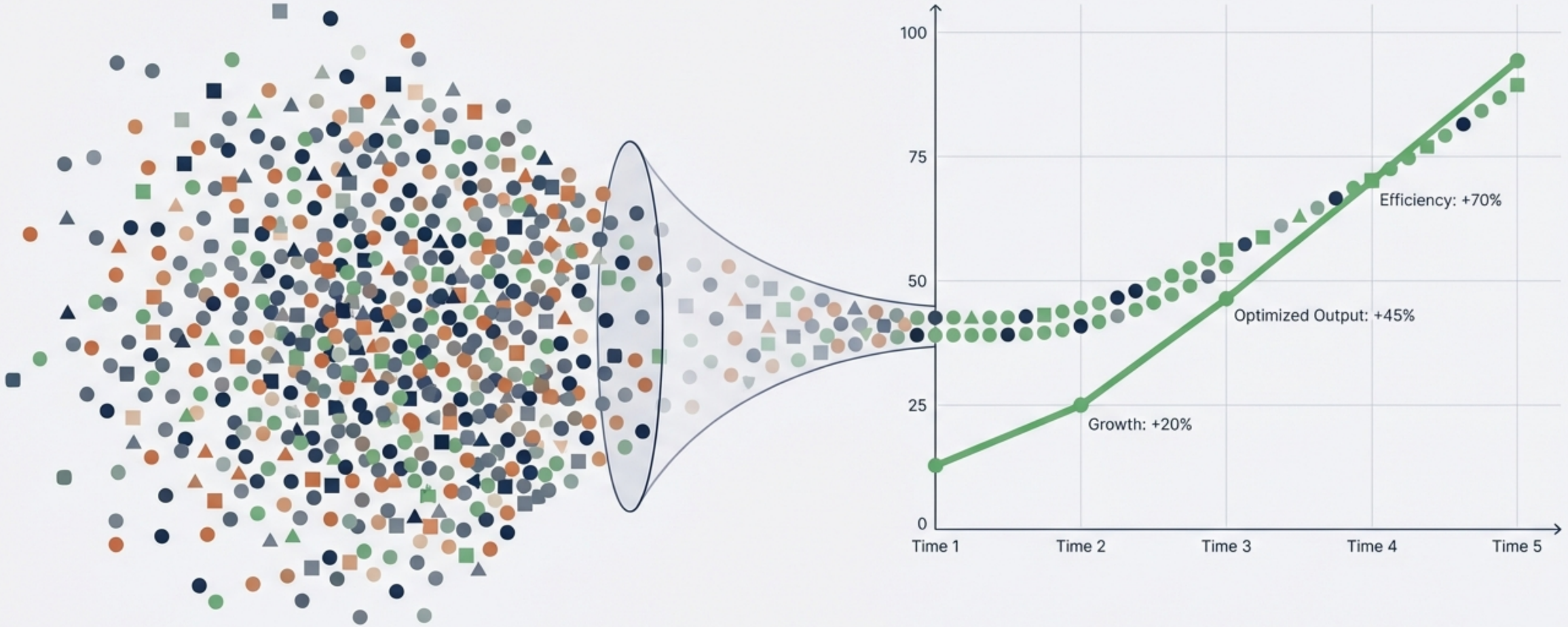
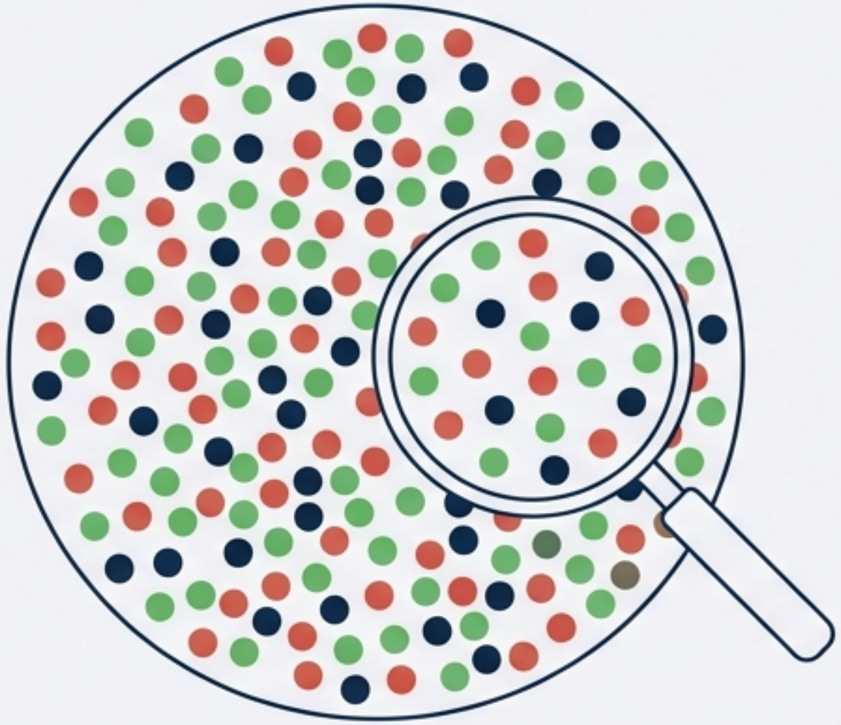


Descriptive Data Analysis: From Raw Numbers to Reliable Insights

A Comprehensive Guide to Sampling, Analysis, and Statistical Validation

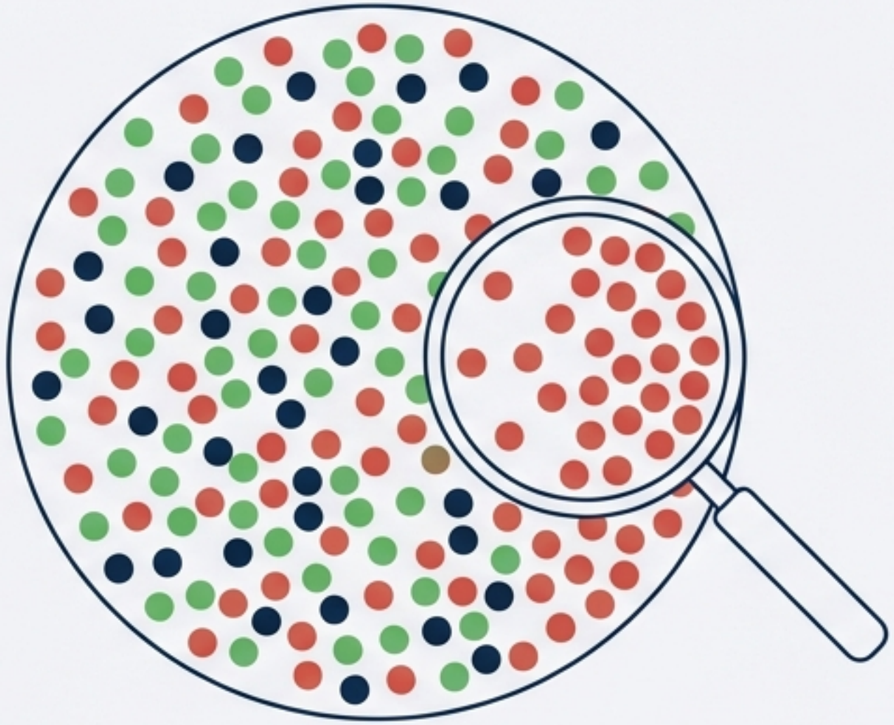
The Art of Sampling: Capturing the Whole by Measuring a Part

Unbiased Sample



Represents the population accurately.
Chosen at random. Large enough to be reliable.

Biased Sample



Does not represent the population.
Preferential screening of specific groups.

Key Insight: The trade-off between accuracy and the cost of data collection.

The Sampling Menu: Probability vs. Non-Probability

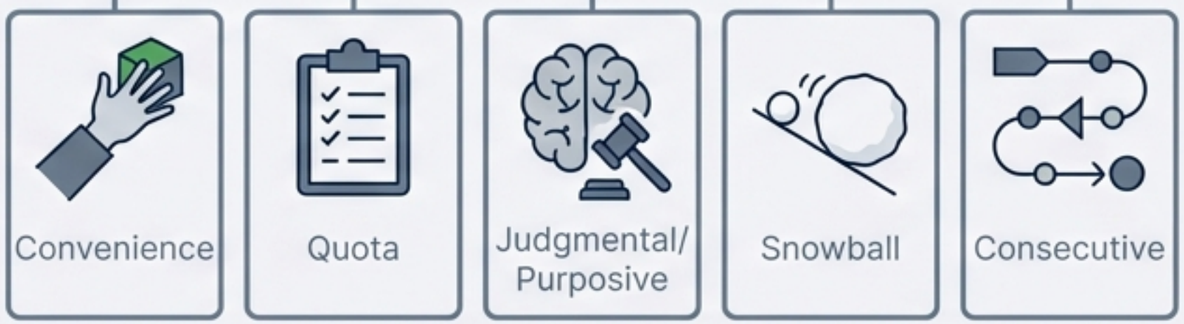
Sampling Methods

Probability Sampling (Random)

Objective: Representative & Unbiased

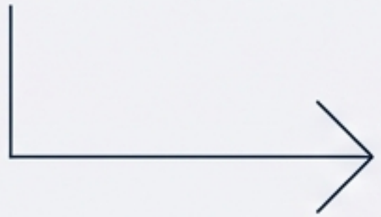
Non-Probability Sampling (Subjective)

Objective: Exploratory & Fast



Taming the Chaos: Organizing Raw Data

Raw Data
43, 65, 71, 76, 98, 82, 95, 83, 84, 96



The Stem and Leaf Plot

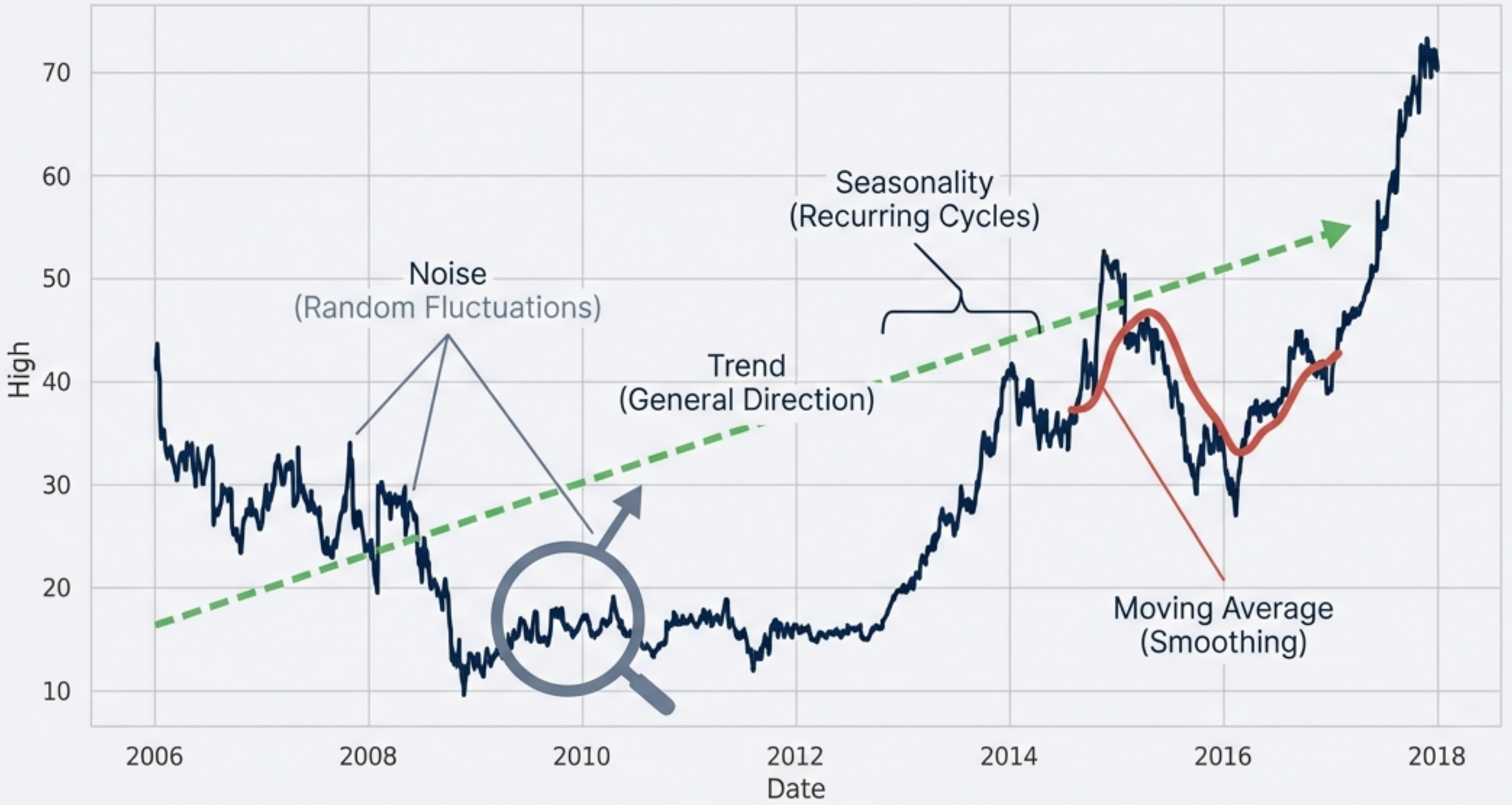
Stem (Tens)	Leaf (Ones)
4	3
6	5
7	1 6
8	2 3 4
9	5 6 8

Frequency Distribution Table (Grouped)

Class Interval	Frequency	Tally
40-59	1	|
60-79	3	|||
80-99	6	𝍸 |

Key Insight: Stem and Leaf plots preserve granular detail (e.g., exact values like 43), while Frequency Tables sacrifice detail for scalability.

Visualizing Time: Anatomy of a Trend



Time Series Components

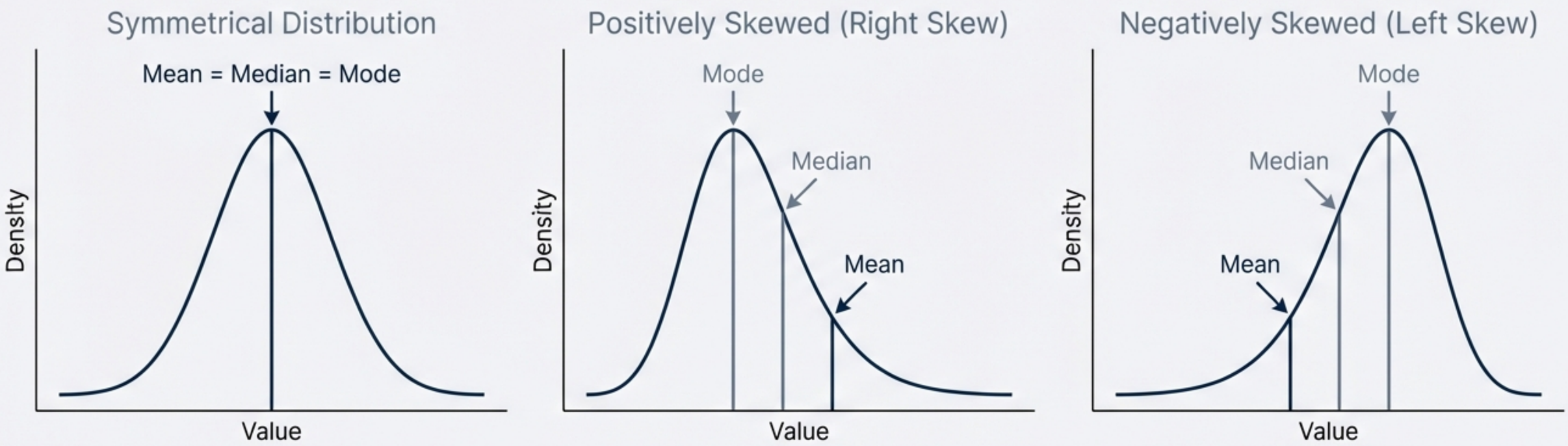
Trend: Long-term movement.

Seasonality: Regular periodic variation.

Noise: Irregular, random variation.

Moving Average: Technique to filter noise.

Finding the Center: Measures of Central Tendency

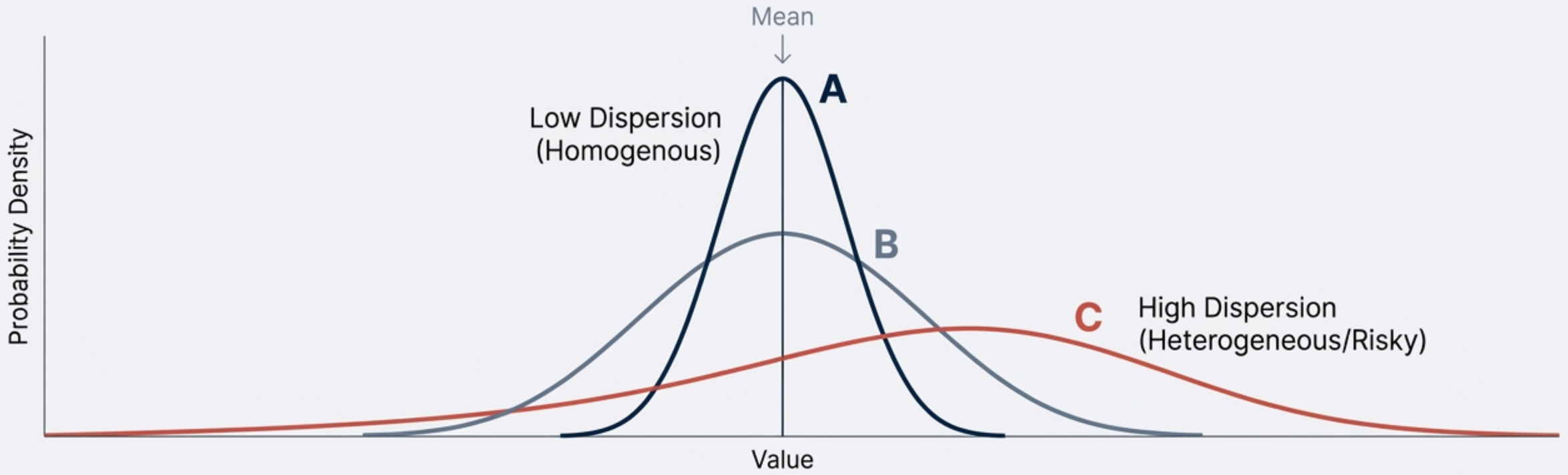


Key Definitions

- **Mean:** Arithmetic Average (Sensitive to outliers).
- **Median:** Middle Value (Resistant to outliers).
- **Mode:** Most Frequent Value (Best for peaks).

Empirical Relation: $2 \times \text{Mean} + \text{Mode} = 3 \times \text{Median}$

Measuring Variability: The Spread of the Data



Absolute Measures
Same units as data

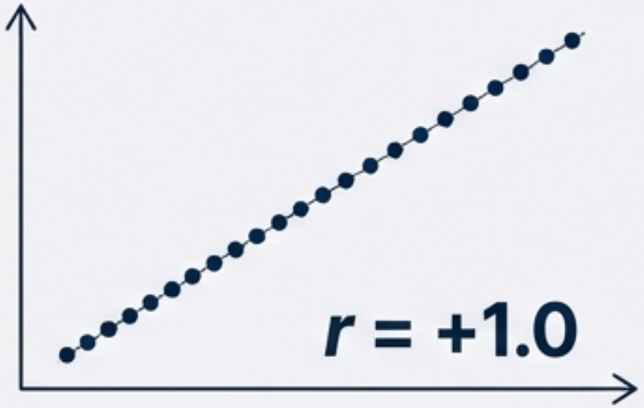
- **Range:** Max Value - Min Value.
- **Variance (σ^2):** Average squared deviation.
- **Standard Deviation ($\sigma$$):** Square root of variance (Most common).

Relative Measures
Unitless

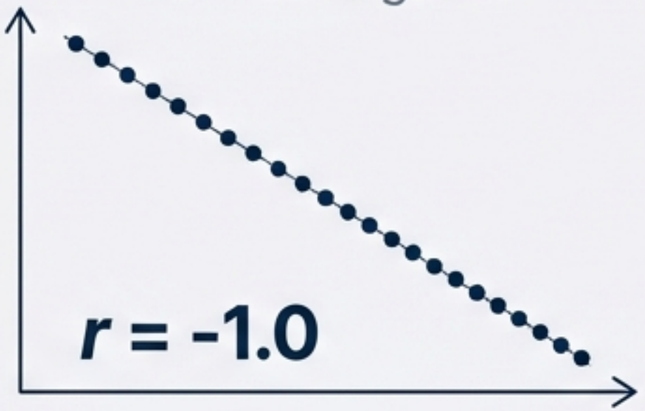
- **Coefficients of Dispersion:** Used to compare datasets with different units (e.g. Height vs Weight).

Connecting the Dots: Correlation Analysis

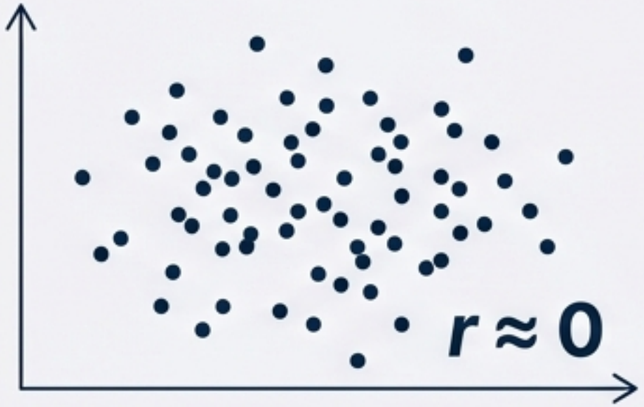
Perfect Positive



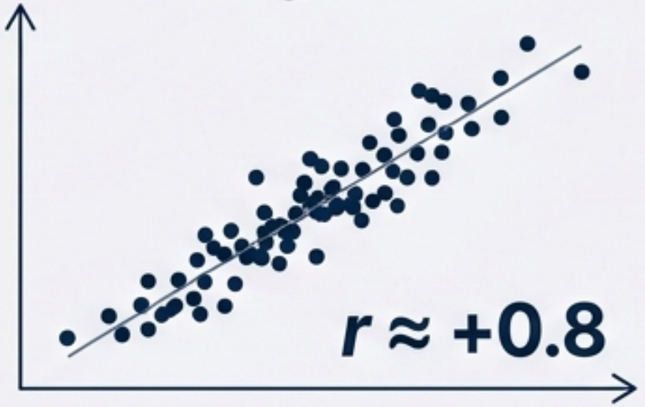
Perfect Negative



No Correlation

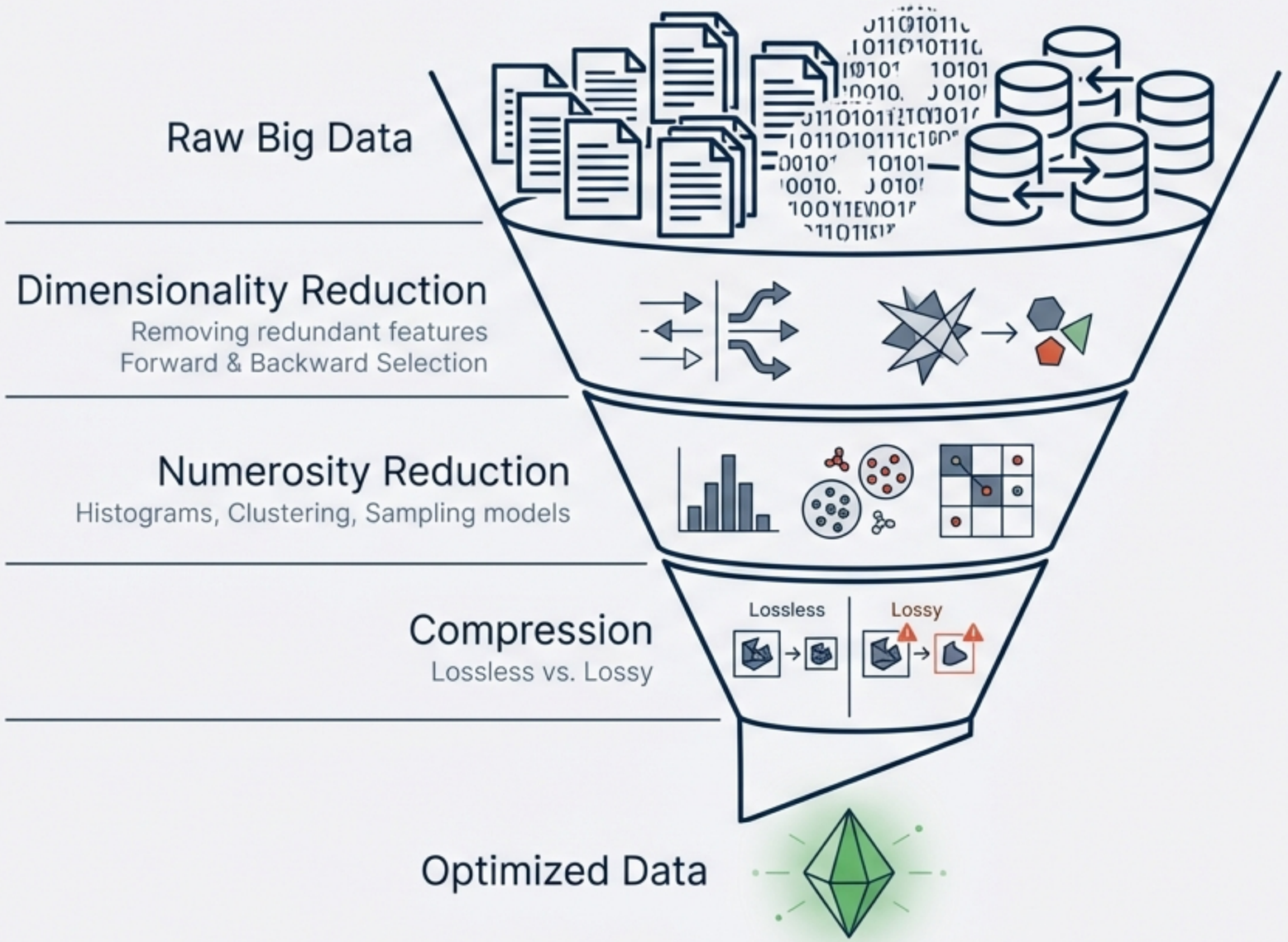


Strong Positive



- **Pearson Coefficient (r):** A measure of linear relationship strength.
- **Critical Warning:** Correlation measures co-variation, NOT causation.

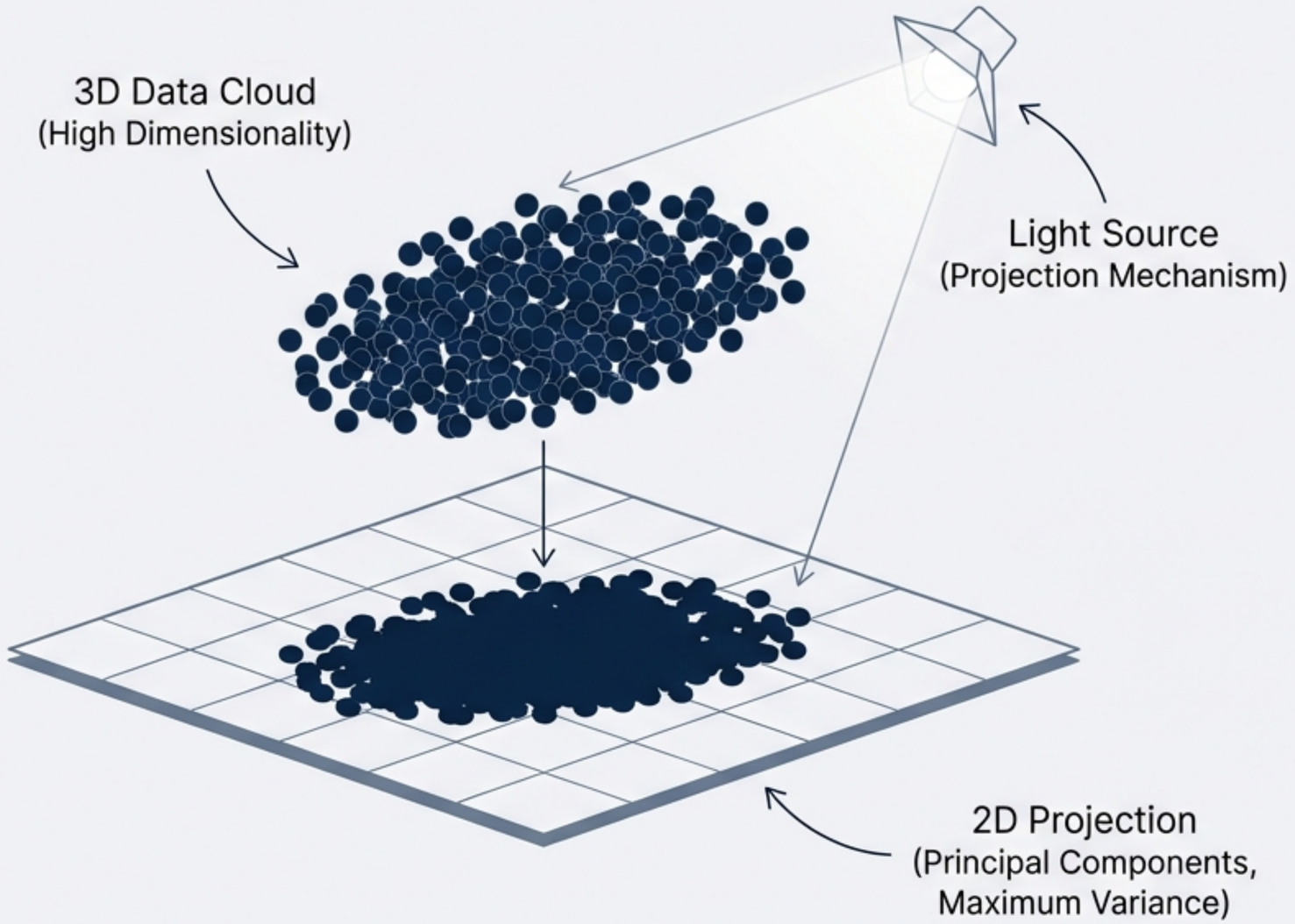
Shrinking the Haystack: Data Reduction Techniques



Goal:

Reduce dataset size and **complexity** while preserving the **integrity** of the information.

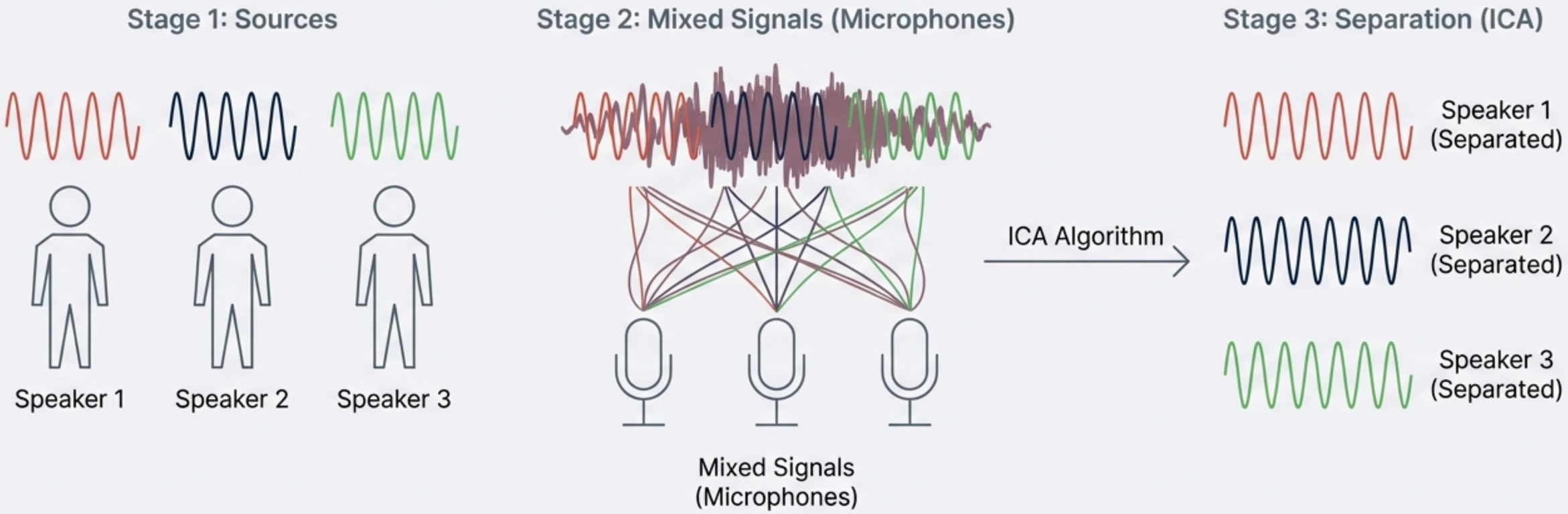
Principal Component Analysis (PCA): Simplifying Complexity



Key Concepts & Trade-offs

- **Goal**: Reduce dimensions while preserving **Maximum Variance.**
- **Mechanism**: Orthogonal transformation to create uncorrelated variables.
- **Trade-off**: Easier visualization (2D/3D) at the cost of interpretability.

The Cocktail Party Problem: Independent Component Analysis (ICA)

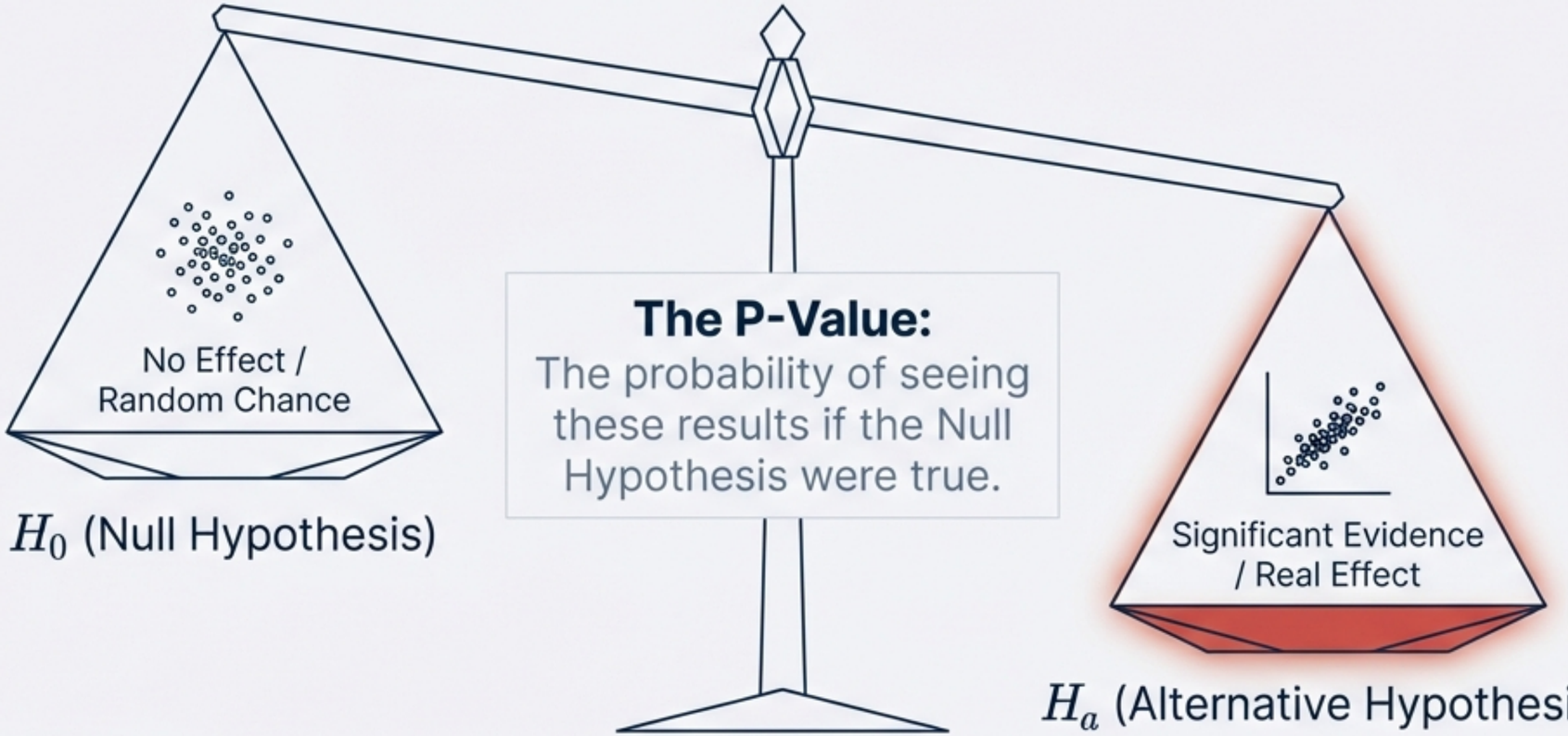


PCA vs. ICA

PCA looks for Variance (Compression). ICA looks for Independence (Separation).

Application: Audio processing, removing noise from brain signals.

The Truth Test: Hypothesis Testing Framework



Decision Rule:

If P-value **≤ 0.05** (Significance Level) ⟶ Reject Null Hypothesis (Result is **Statistically Significant**).

Choosing the Right Test & Managing Errors

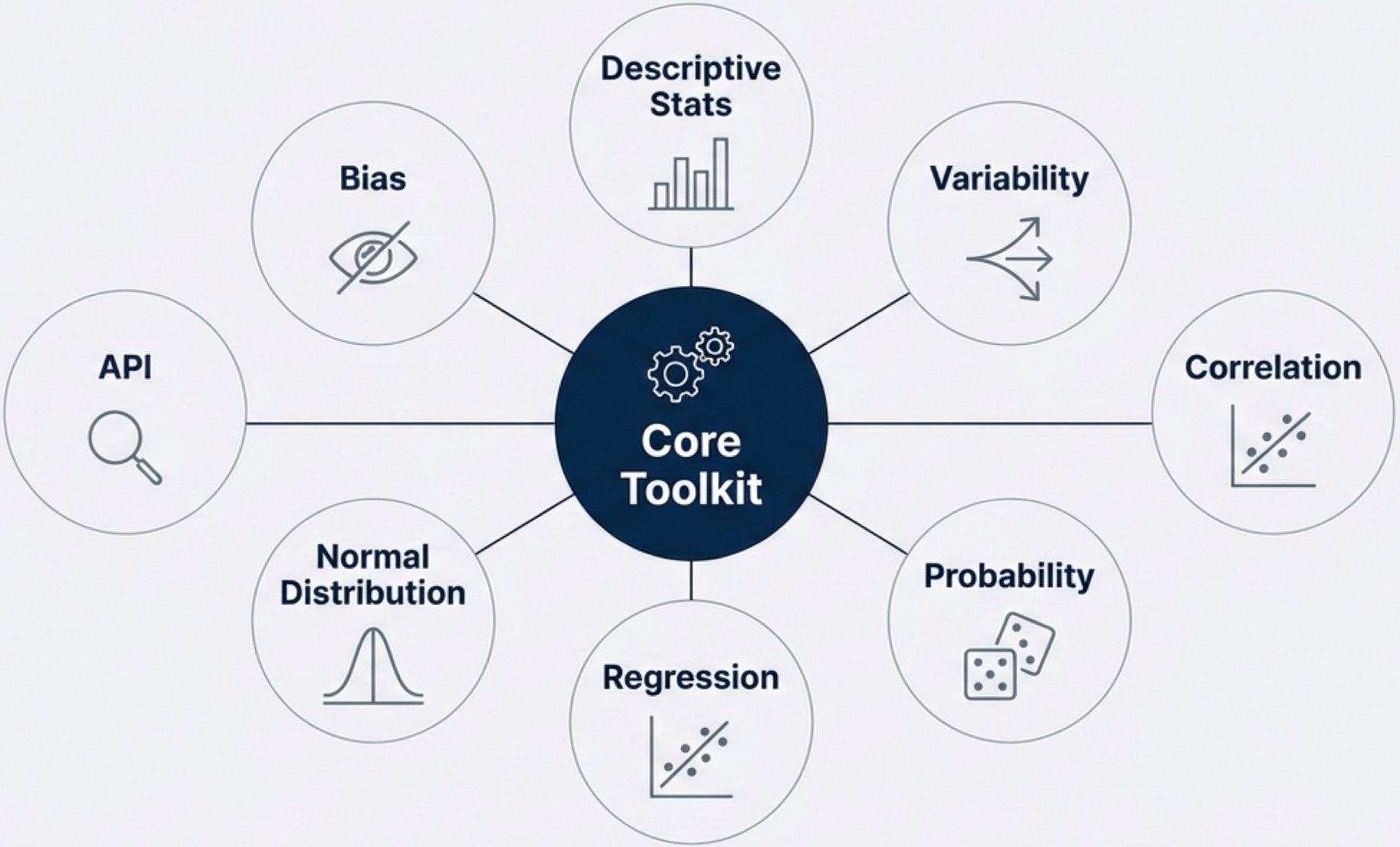
Parametric Tests Assumes Normal Distribution	T-tests (Means), ANOVA (Groups), Z-tests
Non-Parametric Tests No Distribution Assumption / Ordinal Data	Mann-Whitney U, Wilcoxon, Kruskal-Wallis

Warning: Managing Statistical Errors

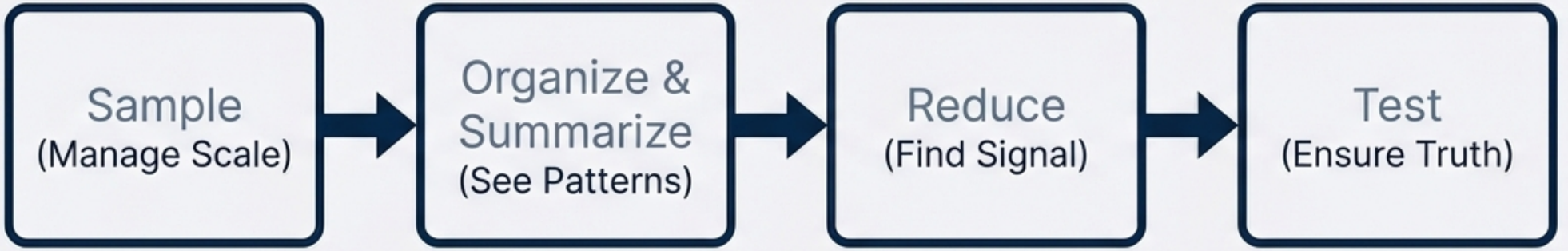
Type I Error (False Positive):
Rejecting a true Null Hypothesis.

Type II Error (False Negative):
Failing to reject a false Null Hypothesis.

The 7 Pillars of Data Science Statistics



The Cycle of Discovery



"The goal of central tendency is to offer a single value that represents data... but the goal of the analyst is to interpret that value to drive decisions."

Inter Formulas

Mean:

$$\bar{x} = \frac{\sum x}{N}$$

Variance:

$$\sigma^2 = \frac{\sum (x - \mu)^2}{N}$$

Correlation:

$$r = \frac{\mathrm{cov}(X, Y)}{\sigma_x \sigma_y}$$