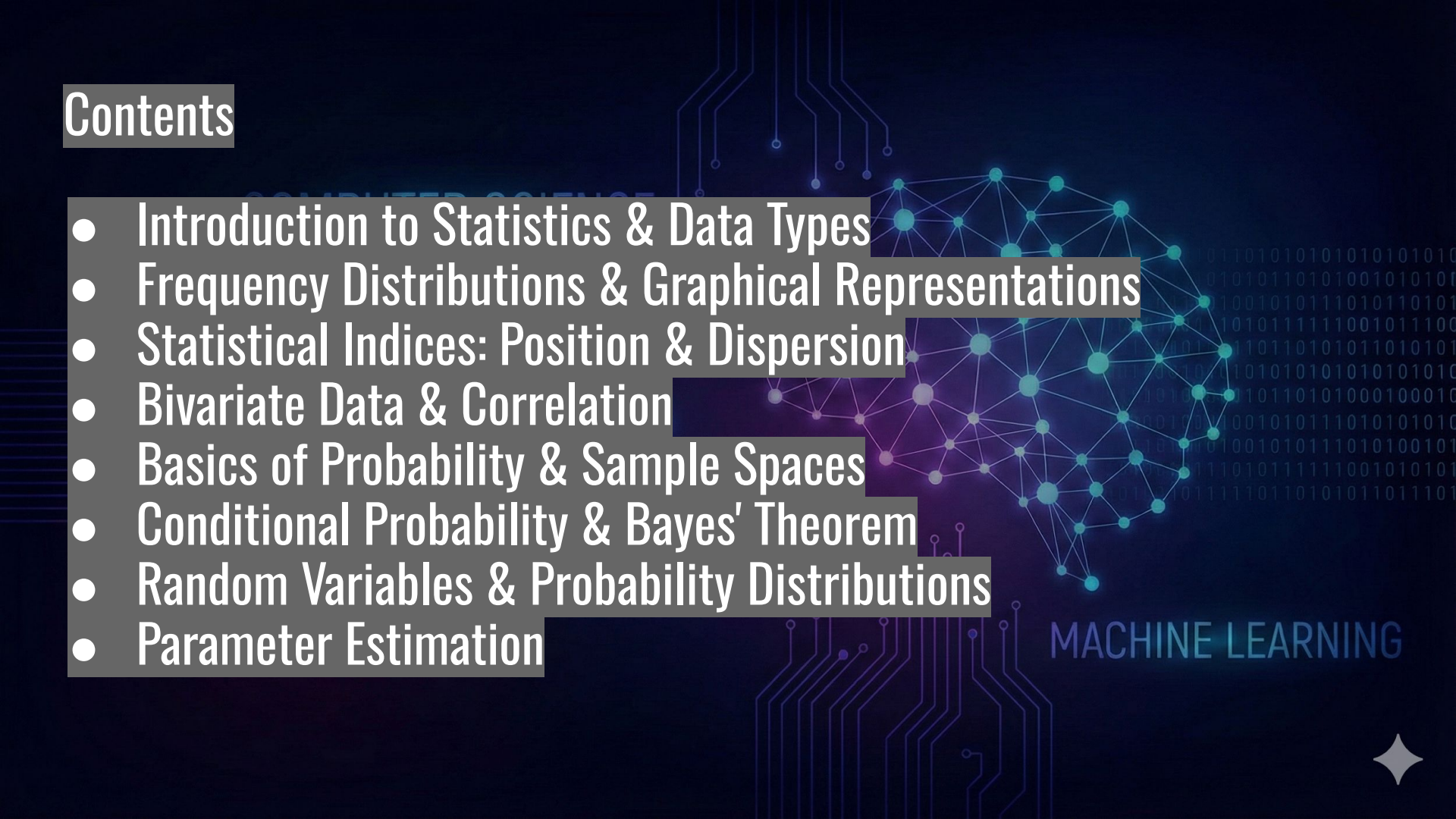
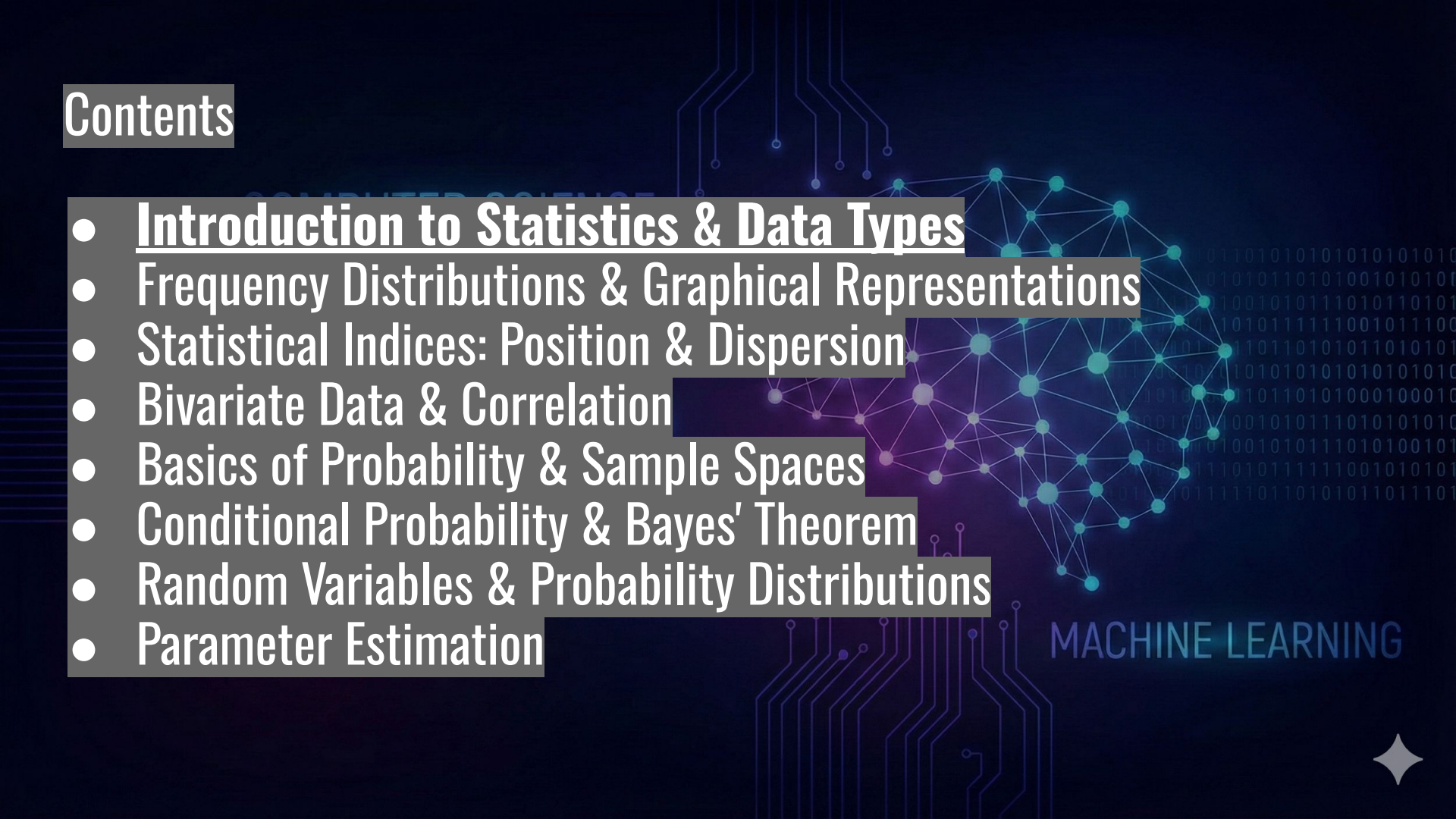
ANALYSIS AND MACHINE LEARNING TECHNIQUES WITH BASICS OF COMPUTER SCIENCE AND STATISTICAL LEARNING - ANALYSIS AND MACHINE LEARNING TECHNIQUES WITH BASICS OF COMPUTER SCIENCE AND STATISTICAL LEARNING

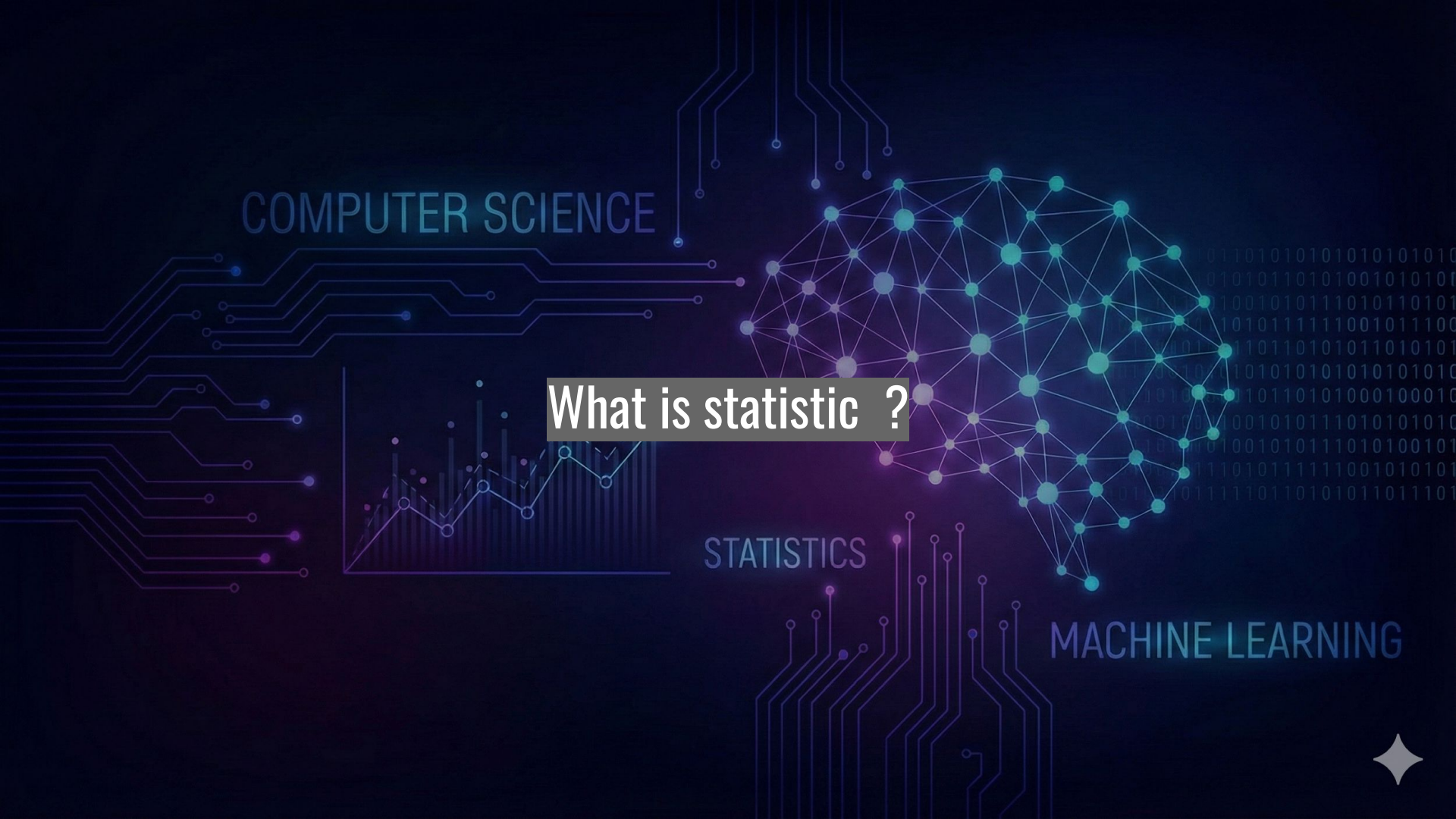
Prof. Loriano Storchi
loriano@storchi.org
https://www.storchi.org/

Contents



Contents




COMPUTER SCIENCE
What is statistic ?
STATISTICS
MACHINE LEARNING

Introduction to Statistics

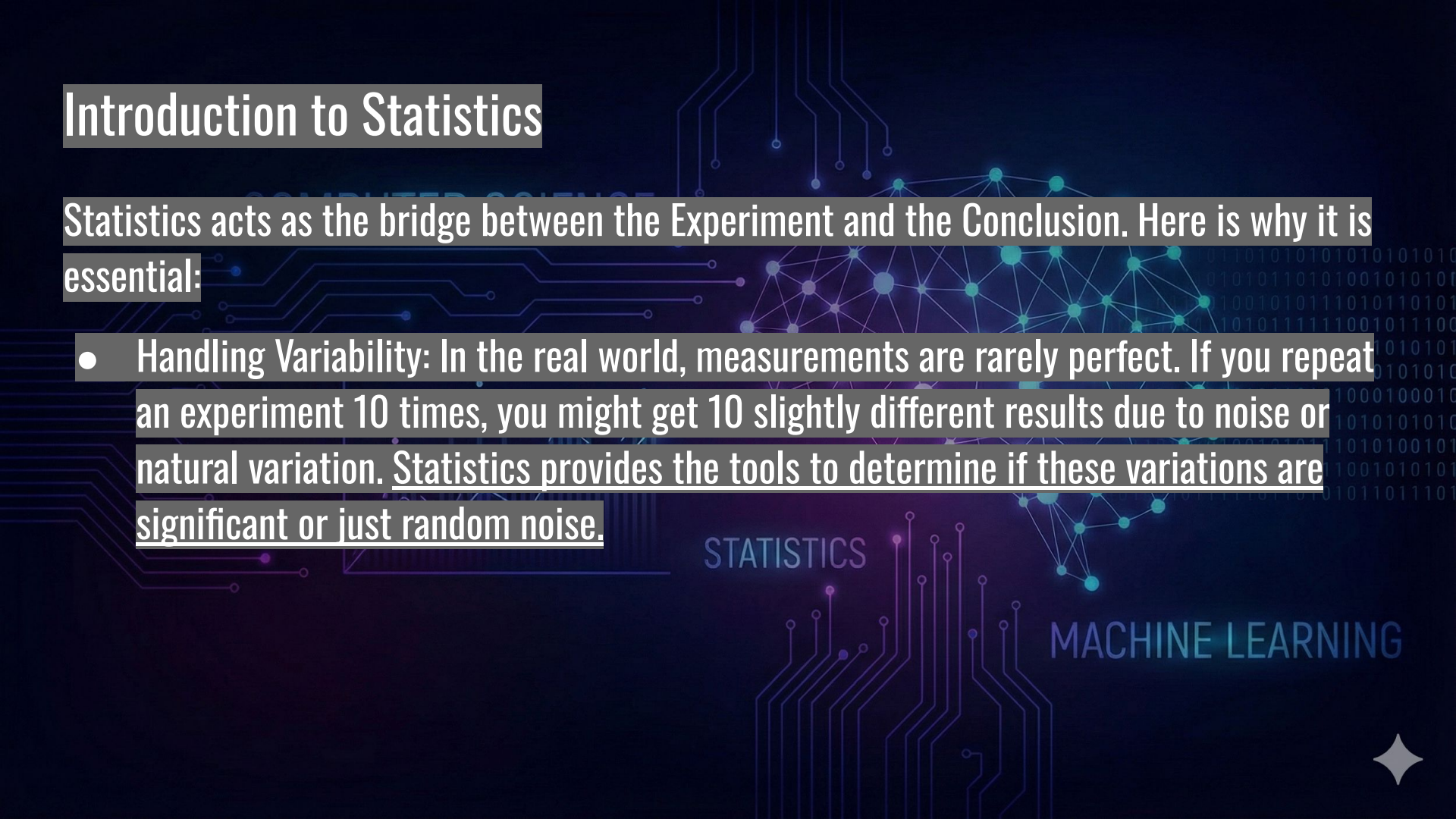
Statistics is the mathematical science involved in the application of quantitative principles to the collection, analysis, and presentation of numerical data.

- Descriptive Statistics: The art of summarizing data. This involves organizing raw data into meaningful graphs, tables, and metrics (like the mean, median, or standard deviation) to understand "what happened" in an experiment.
- Inferential Statistics: The art of making conclusions. This involves using data from a small sample to make generalizations (inferences) about a larger population. This is where you calculate the probability that your experimental results are valid and not just due to luck

Introduction to Statistics

Statistics acts as the bridge between the Experiment and the Conclusion. Here is why it is essential:

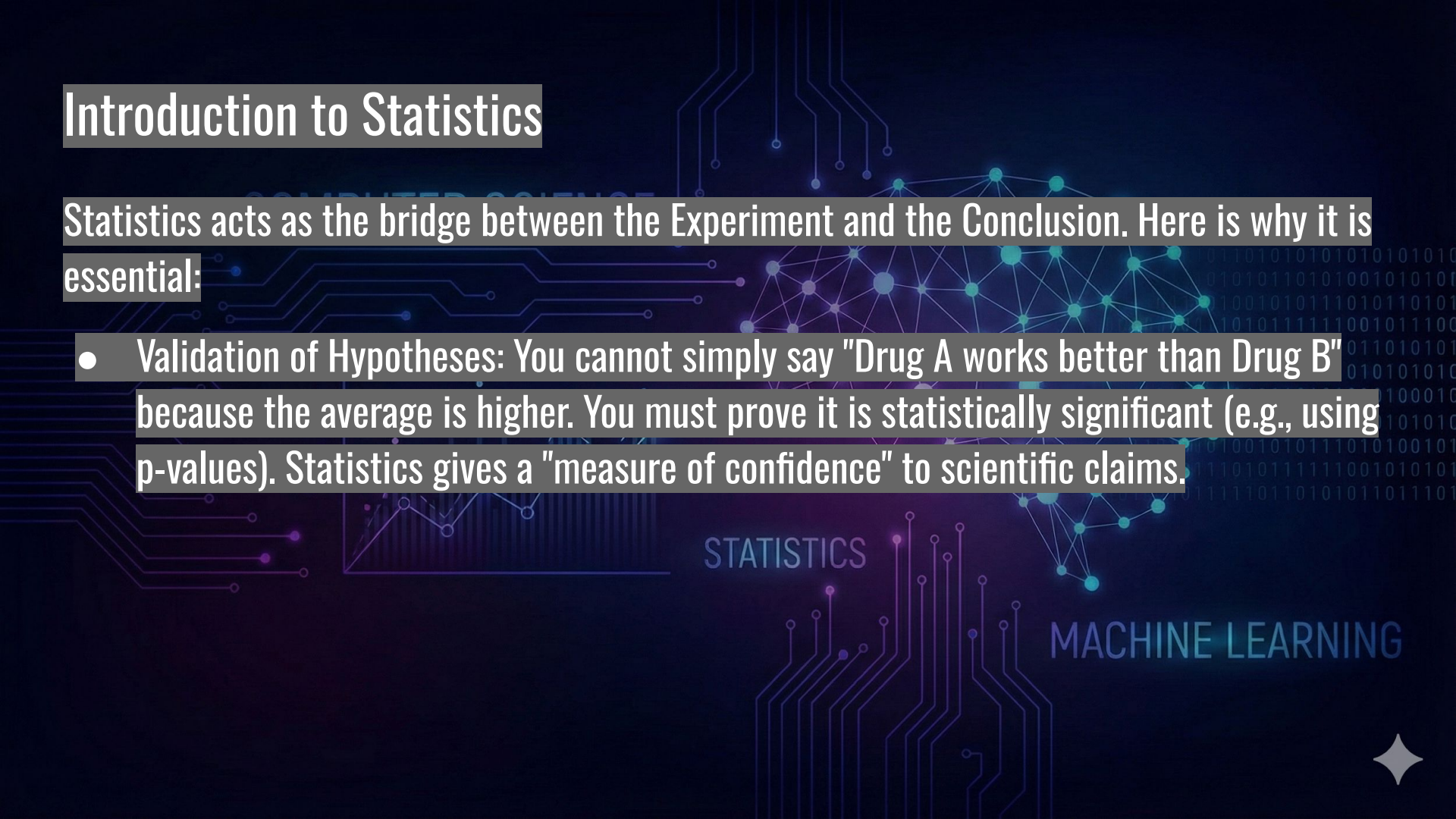
- Handling Variability: In the real world, measurements are rarely perfect. If you repeat an experiment 10 times, you might get 10 slightly different results due to noise or natural variation. Statistics provides the tools to determine if these variations are significant or just random noise.



Introduction to Statistics

Statistics acts as the bridge between the Experiment and the Conclusion. Here is why it is essential:

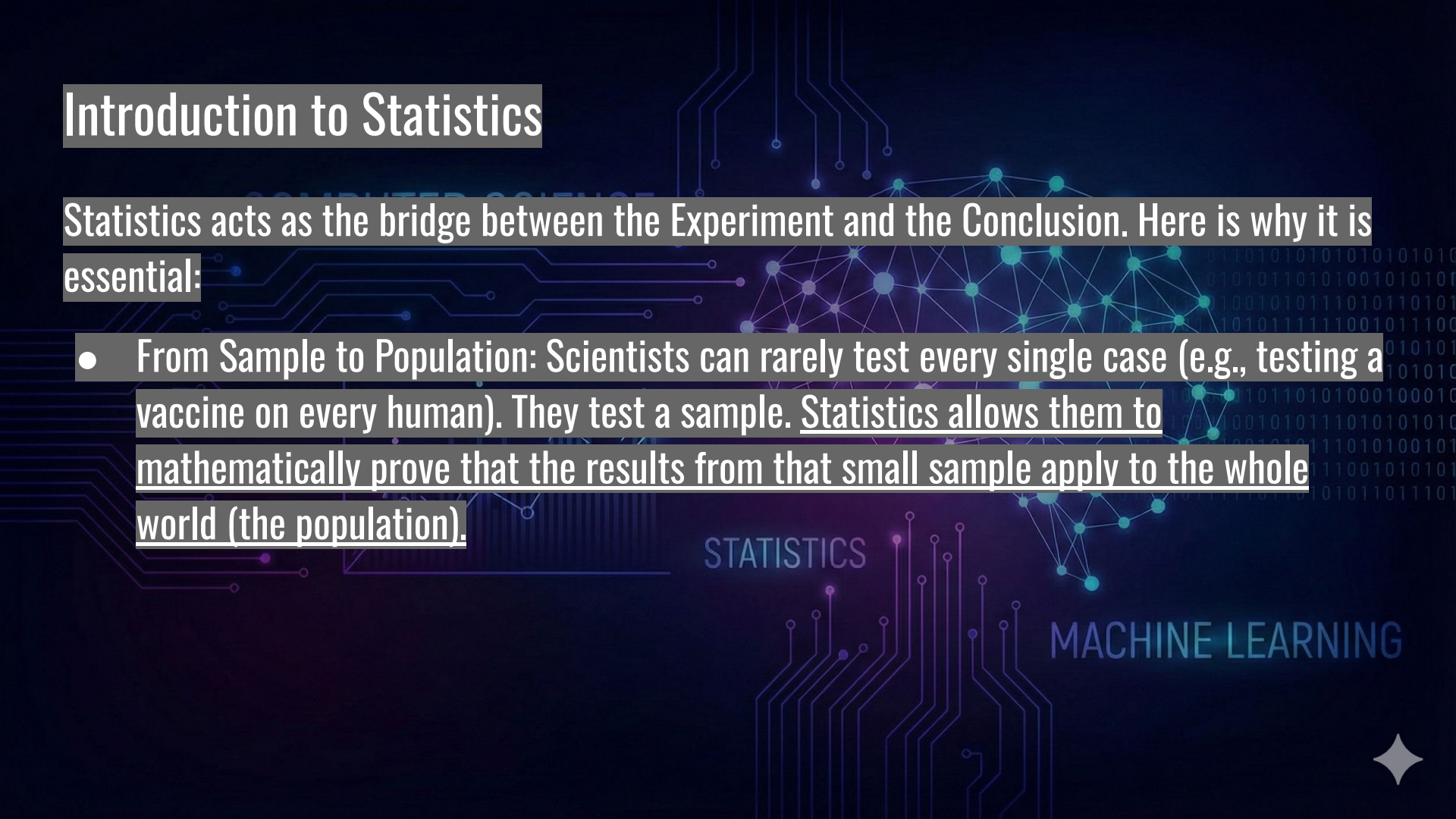
- Validation of Hypotheses: You cannot simply say "Drug A works better than Drug B" because the average is higher. You must prove it is statistically significant (e.g., using p-values). Statistics gives a "measure of confidence" to scientific claims.



Introduction to Statistics

Statistics acts as the bridge between the Experiment and the Conclusion. Here is why it is essential:

- From Sample to Population: Scientists can rarely test every single case (e.g., testing a vaccine on every human). They test a sample. Statistics allows them to mathematically prove that the results from that small sample apply to the whole world (the population).



1. Example of Descriptive Statistics

SCENARIO

A scientist is testing a new fertilizer on 100 tomato plants.

THE TASK

After 4 weeks, she measures the height of every single plant.

THE DESCRIPTIVE STAT

Average height: 45 cm. Standard deviation: 5 cm. (Most plants are between 40 and 50 cm).

WHY

Simply describing the data at hand. No broader claims yet—just summarizing what happened in her specific garden.

2. Inferential Statistics: Pharmaceutical Case Study

SCENARIO

A pharmaceutical firm develops a new drug to lower blood pressure in the global population.

THE CLINICAL TRIAL

Scientists test the drug on a sample of 250 diverse volunteers and measure the mean reduction in pressure.

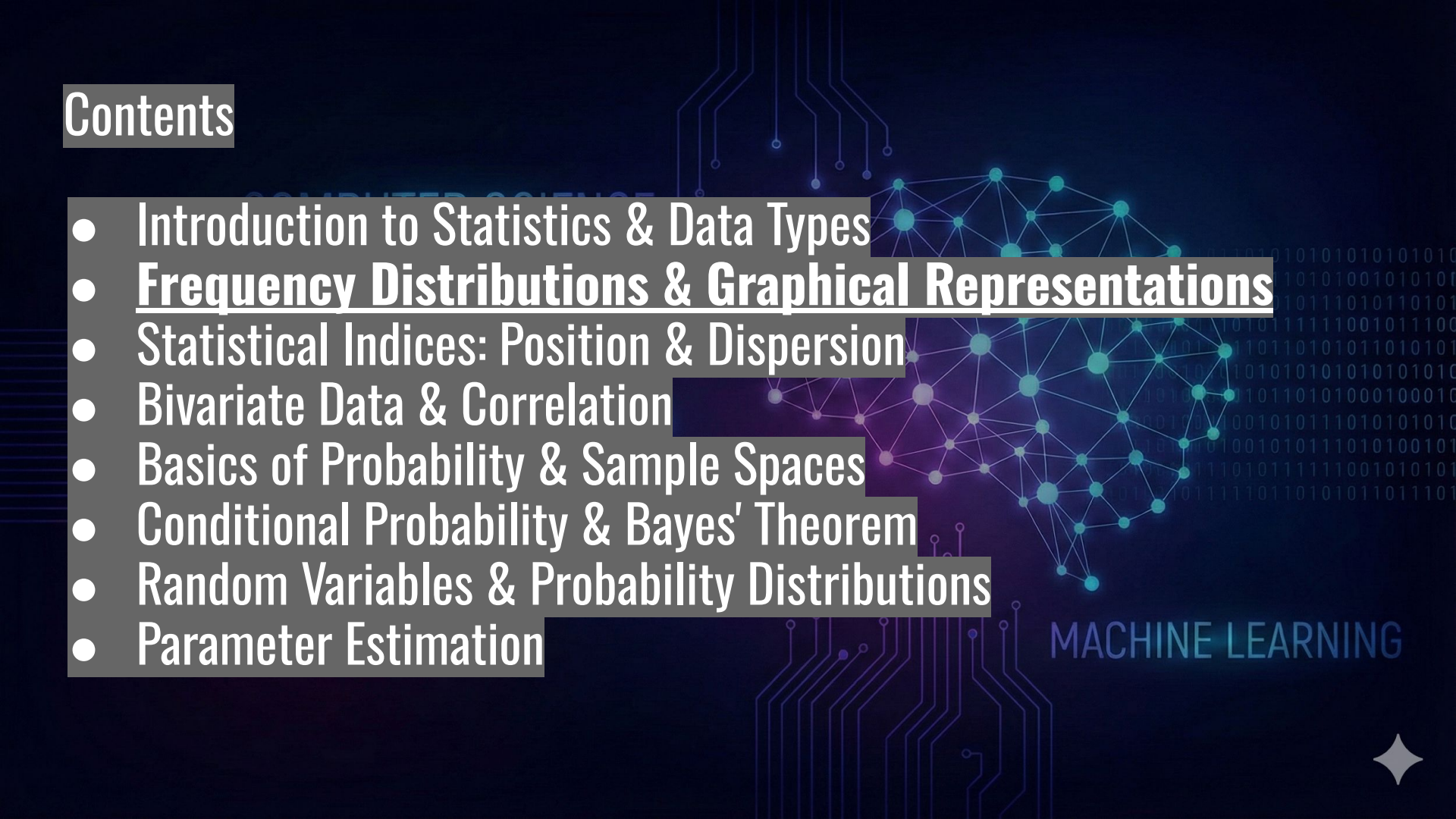
THE INFERENTIAL CLAIM

Based on the trial results, researchers conclude with 95% confidence that the drug will be effective for millions of patients worldwide.

WHY

Statistics allow scientists to mathematically prove that results from a small sample (the trial) apply to the whole world (the population).

Contents



Transitioning to Practical Data Handling

Variable Classification

Defining the nature of the data we collect:

- Numerical (Discrete/Continuous)
- Categorical (Qualitative)

Understanding measurement scales is the first step in analysis.

Frequency Tables

Organizing raw data into meaningful summaries:

- Absolute Frequencies
- Relative & Percentage

Converting chaos into structured information for decision making.

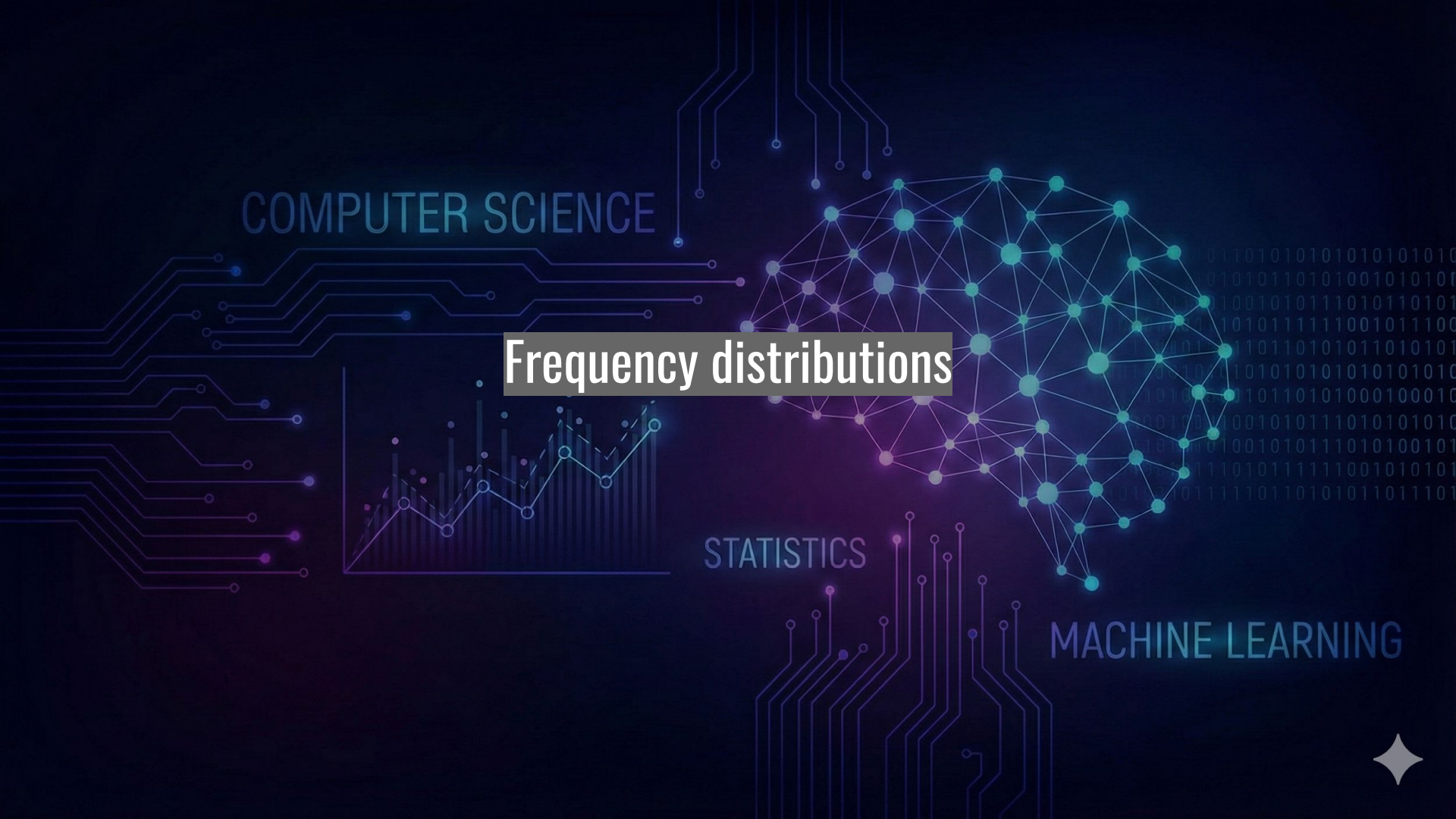
Graphical Representation

Visualizing patterns and distributions:

- Histograms for trends

Moving from broad theory to visual evidence.

Next Step: Mastering the tools to describe "what happened" in our specific dataset.


COMPUTER SCIENCE
Frequency distributions
STATISTICS
MACHINE LEARNING

Introduction and Variable Classification

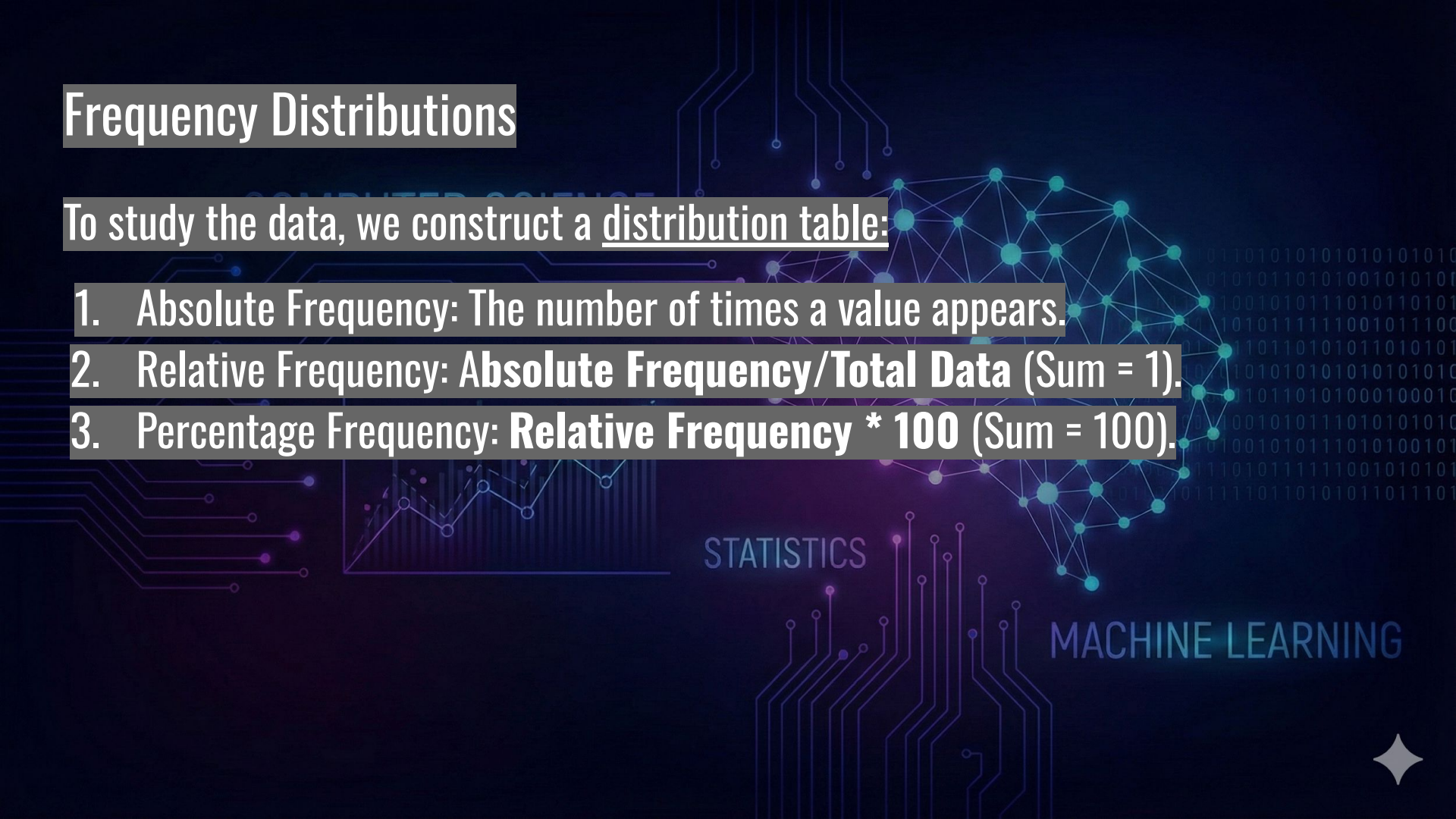
Classification of Variables:

- **Numerical**: The values assumed are numbers.
  - **Discrete: Finite or countable set of values (e.g., Number of births in a family).**
  - **Continuous: Set of continuous values, such as R or an interval (e.g., Height in cm).**
- **Categorical**: The values are not numbers (e.g., Eye color).

Frequency Distributions

To study the data, we construct a distribution table:

1. Absolute Frequency: The number of times a value appears.
2. Relative Frequency: A**bsolute Frequency/Total Data** (Sum = 1).
3. Percentage Frequency: **Relative Frequency * 100** (Sum = 100).



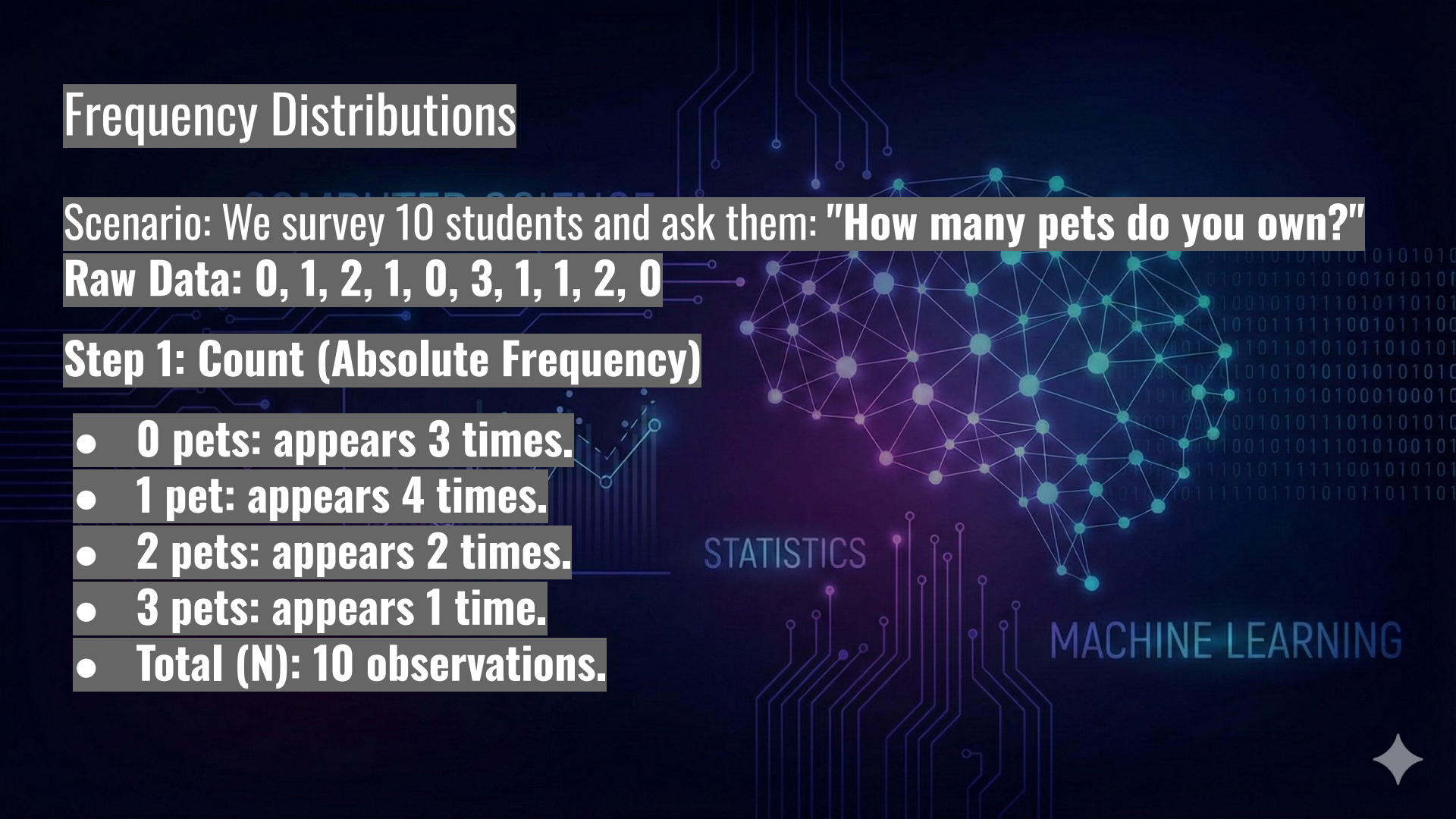
Frequency Distributions

Scenario: We survey 10 students and ask them: **"How many pets do you own?"**

Raw Data: 0, 1, 2, 1, 0, 3, 1, 1, 2, 0

Step 1: Count (Absolute Frequency)

- **0 pets: appears 3 times.**
- **1 pet: appears 4 times.**
- **2 pets: appears 2 times.**
- **3 pets: appears 1 time.**
- **Total (N): 10 observations.**



Frequency Distributions

Step 2: Calculate Relative & Percentage

Number of Pets (xi)	Absolute Freq (ni)	Relative Freq (fi=ni/N)	Percentage Freq (pi=fi×100)
0	3	3/10=0.3	30%
1	4	4/10=0.4	40%
2	2	2/10=0.2	20%
3	1	1/10=0.1	10%
TOTAL	10	1	100%

Interpretation

Absolute Frequency

Tells us the exact count of occurrences.

4 students have 1 pet.

Relative Frequency

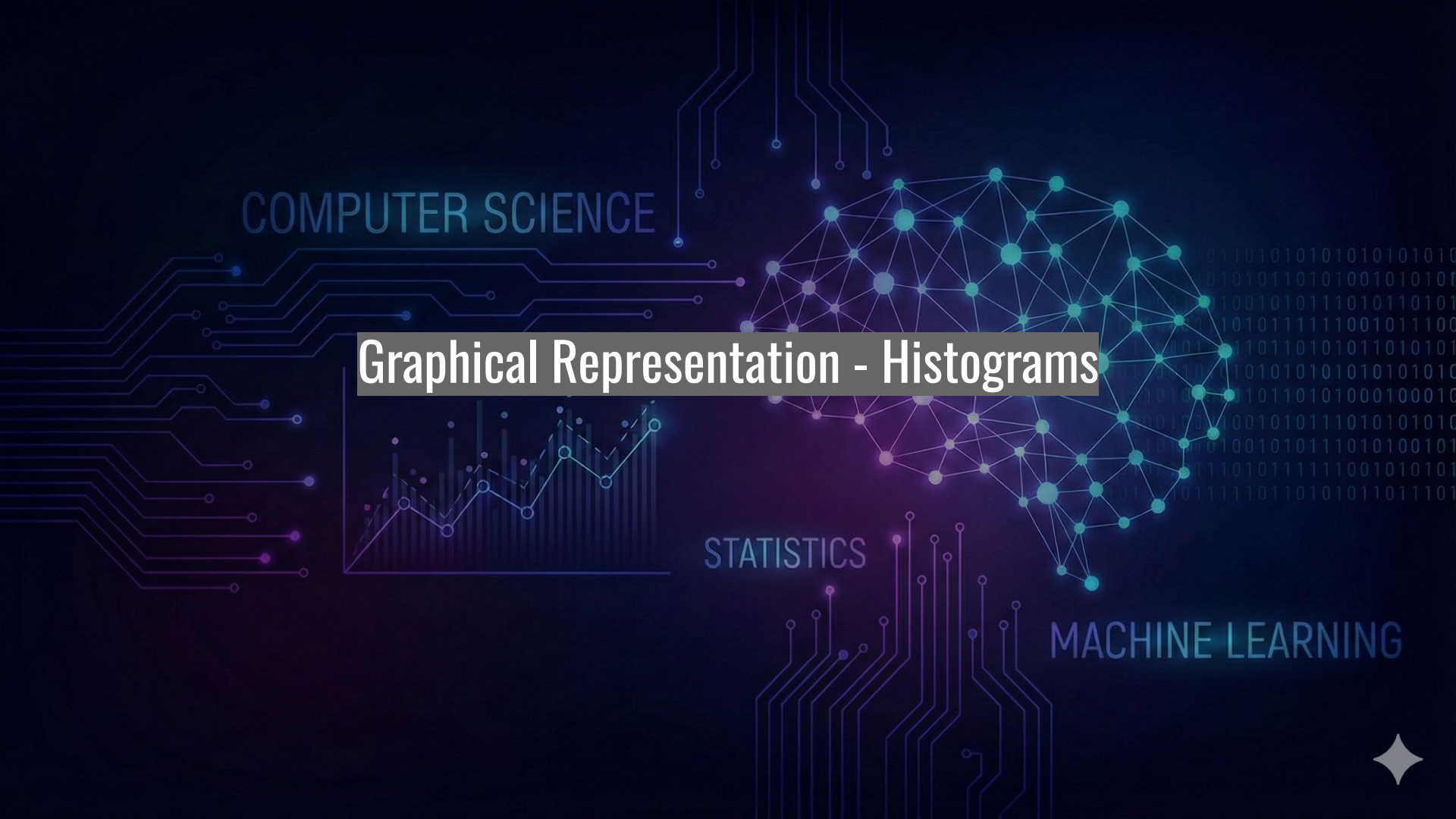
Crucial for calculating probabilities.

Probability of picking a student with 1 pet is 0.4.

Percentage

The best way to communicate findings clearly.

"40% of the class has 1 pet."


COMPUTER SCIENCE
Graphical Representation - Histograms
STATISTICS
MACHINE LEARNING

Graphical Representation - Histograms

Used for numerical variables.

Structure

Consists of adjacent rectangles where the base corresponds to class width.

Area

Proportional to the frequency (relative or percentage) of the class.

Height Calculation

Calculated by dividing the frequency by the class width.

Important: If classes have unit width, the height coincides with the frequency.

Height = Frequency / Class Width

Histogram Construction Example

Data: Total of 200 observations.

Class	Frequency	Freq. %	Width	Height (Freq%/Width)
110<D≤130	20	10%	20	10/20=0.5
130<D≤150	40	20%	20	20/20=1.0
150<D≤170	60	30%	20	30/20=1.5
170<D≤210	80	40%	40	40/40=1.0

The total area of the rectangles will equal 100 (if using percentages) or 1 (if using relative frequencies).

Histogram Construction Example

Sum = 200

Data: Total of 200 observations.

Class	Frequency	Freq. %	Width	Height (Freq%/Width)
110<D≤130	20	10%	20	10/20=0.5
130<D≤150	40	20%	20	20/20=1.0
150<D≤170	60	30%	20	30/20=1.5
170<D≤210	80	40%	40	40/40=1.0

The total area of the rectangles will equal 100 (if using percentages) or 1 (if using relative frequencies).

Histogram Construction Example

Data: Total of 200 observations.

Sum = 200
20/200 = 0.1
0.1 * 100 = 10

Class	Frequency	Freq. %	Width	Height (Freq%/Width)
110<D≤130	20	10%	20	10/20=0.5
130<D≤150	40	20%	20	20/20=1.0
150<D≤170	60	30%	20	30/20=1.5
170<D≤210	80	40%	40	40/40=1.0

The total area of the rectangles will equal 100 (if using percentages) or 1 (if using relative frequencies).

Histogram Construction Example

130-110 = 20

Data: Total of 200 observations.

Class	Frequency	Freq. %	Width	Height (Freq%/Width)
110<D≤130	20	10%	20	10/20=0.5
130<D≤150	40	20%	20	20/20=1.0
150<D≤170	60	30%	20	30/20=1.5
170<D≤210	80	40%	40	40/40=1.0

The total area of the rectangles will equal 100 (if using percentages) or 1 (if using relative frequencies).

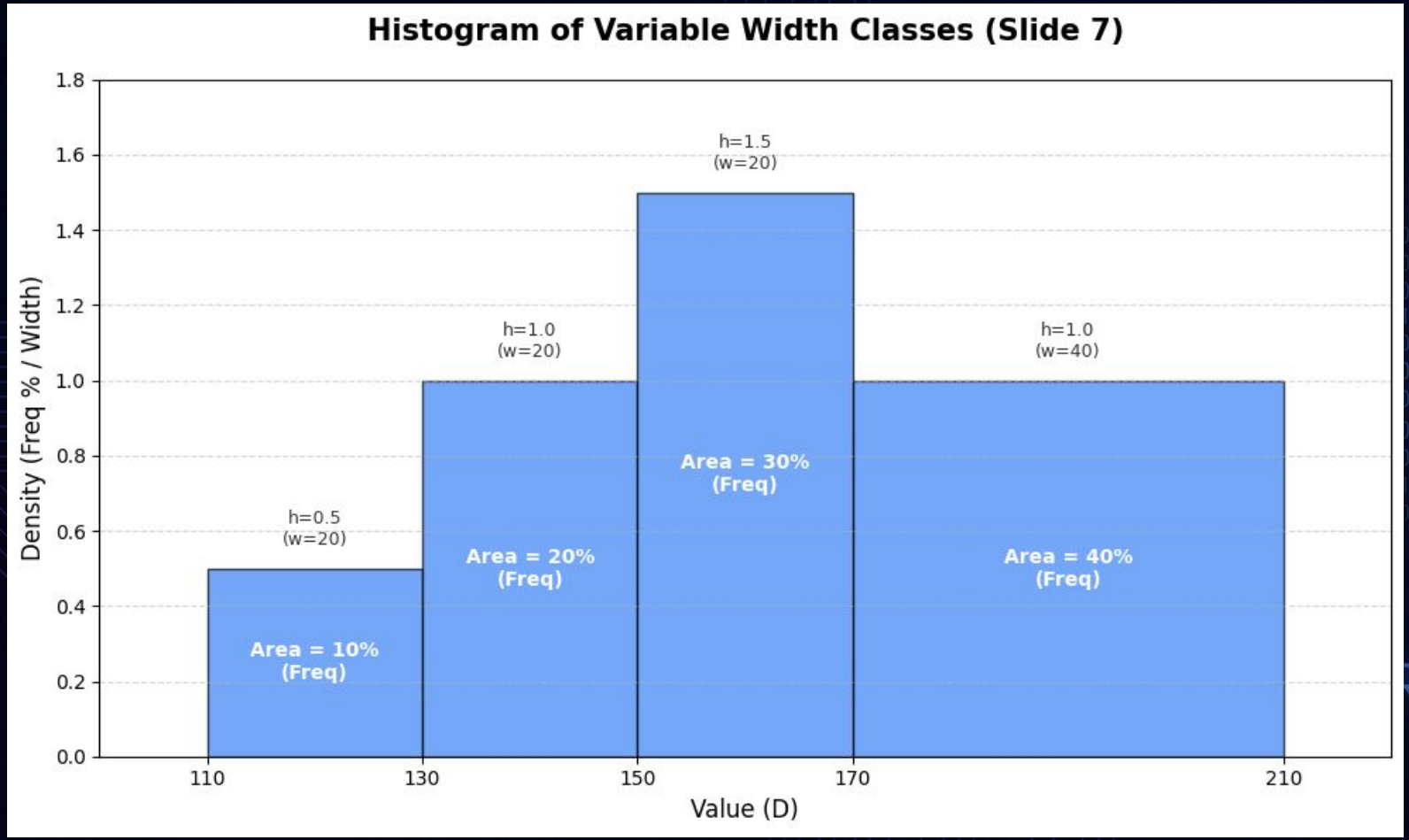
Histogram Construction Example

Data: Total of 200 observations.

10/20

Class	Frequency	Freq. %	Width	Height (Freq%/Width)
110<D≤130	20	10%	20	10/20=0.5
130<D≤150	40	20%	20	20/20=1.0
150<D≤170	60	30%	20	30/20=1.5
170<D≤210	80	40%	40	40/40=1.0

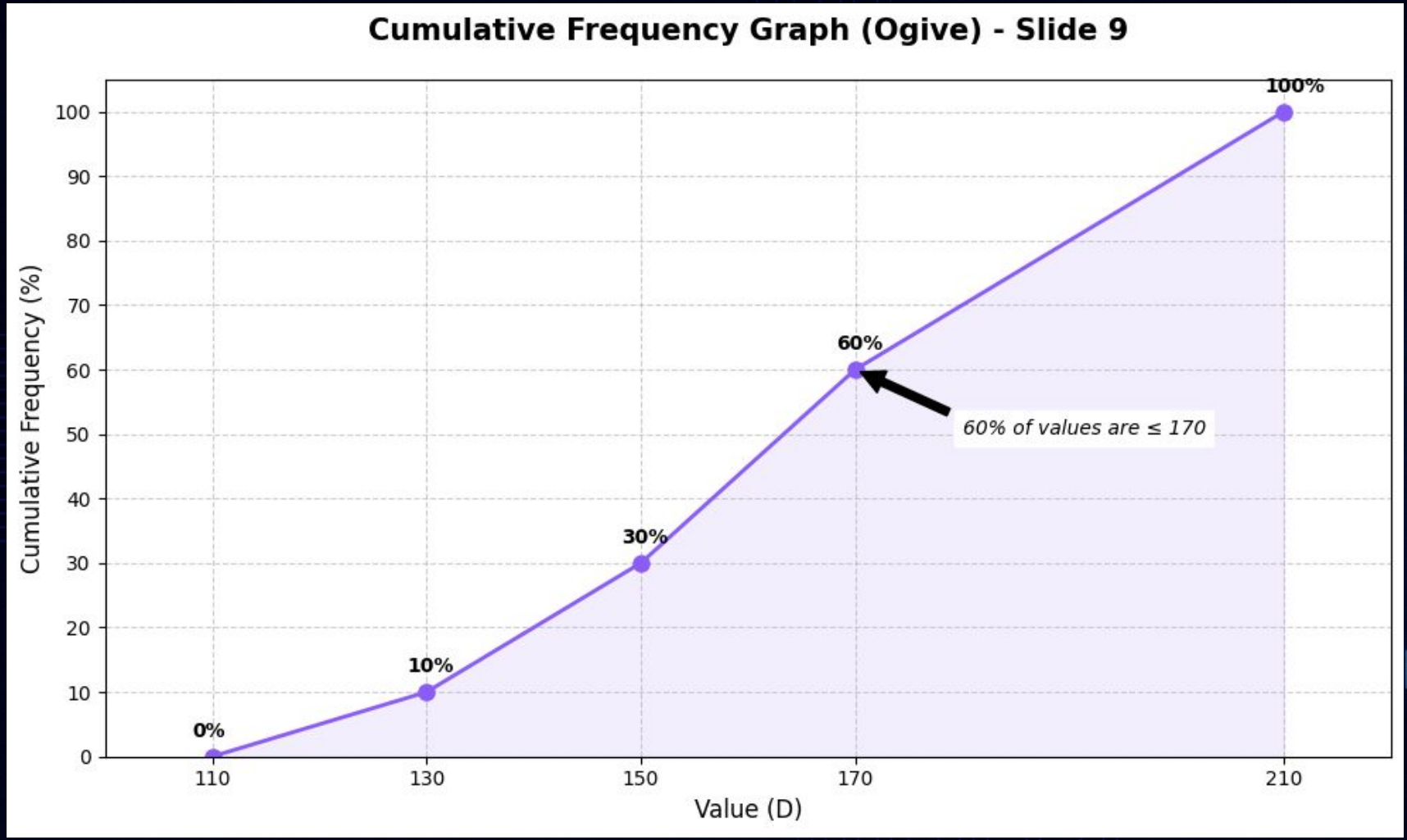
The total area of the rectangles will equal 100 (if using percentages) or 1 (if using relative frequencies).


Histogram of Variable Width Classes (Slide 7)
Density (Freq % / Width)
Value (D)
1.8
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
110
130
150
170
210
h=0.5
(w=20)
h=1.0
(w=20)
h=1.5
(w=20)
h=1.0
(w=40)
Area = 10%
(Freq)
Area = 20%
(Freq)
Area = 30%
(Freq)
Area = 40%
(Freq)

Cumulative Frequencies

Valid only for numerical variables.

- Definition: The sum of all frequencies less than or equal to the upper boundary of a fixed class.
- Indicates how many individuals have a value ≤ to a certain threshold (e.g., 60% of the population has a value ≤ 170).


Cumulative Frequency Graph (Ogive) - Slide 9
Cumulative Frequency (%)
Value (D)
0%
10%
30%
60%
100%
60% of values are ≤ 170
0
10
20
30
40
50
60
70
80
90
100
110
130
150
170
210

Unit Width Histograms (w = const)

In most practical applications, class widths are kept constant (e.g. w = 1 or unit width).

Key Property:

When the class width (w) is constant and equal to unity:

Height = Frequency

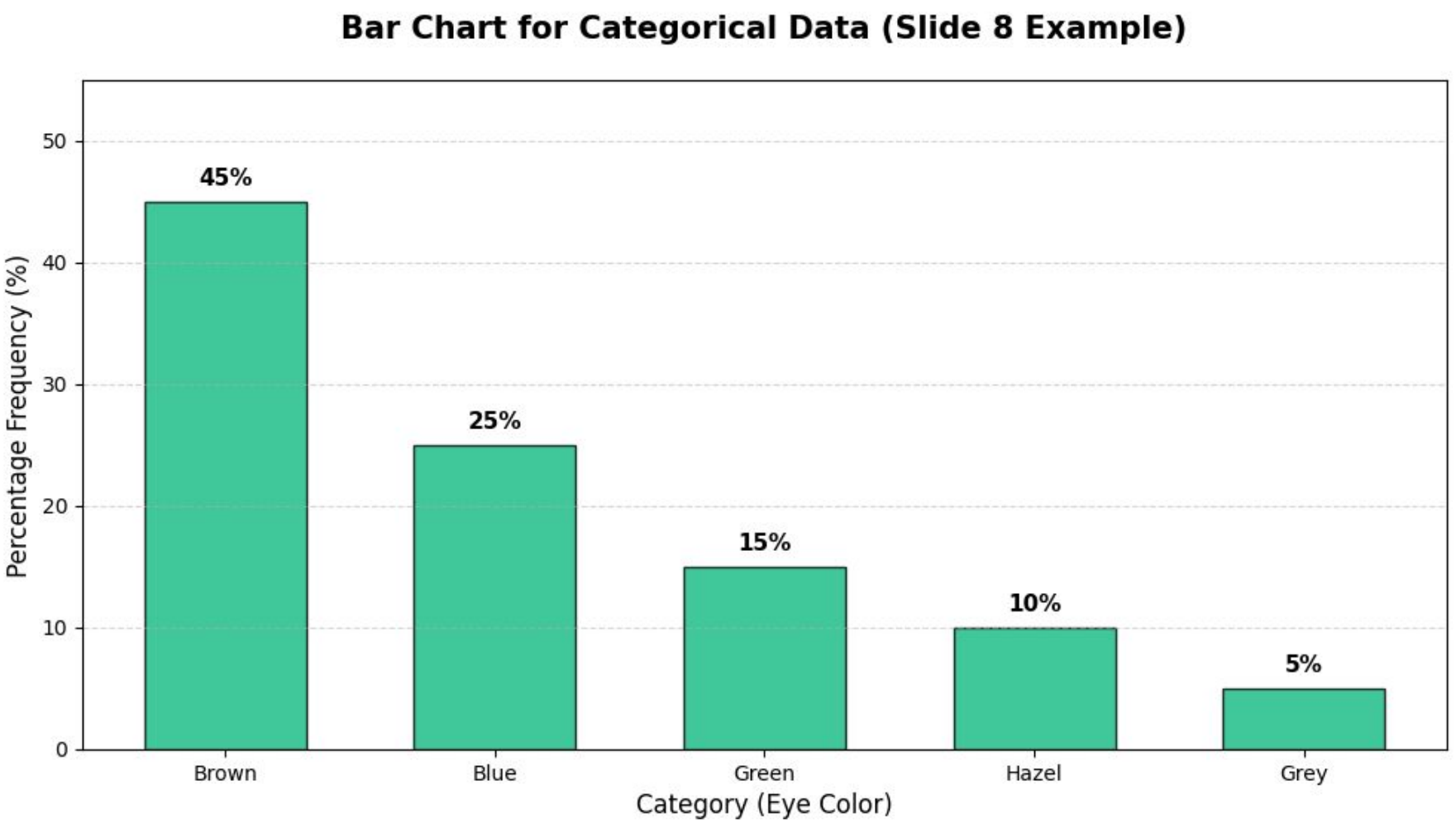
The height of the bar directly represents the density because the divisor (width) is 1.

Standard representation simplifies visual interpretation of density.

Bar Charts

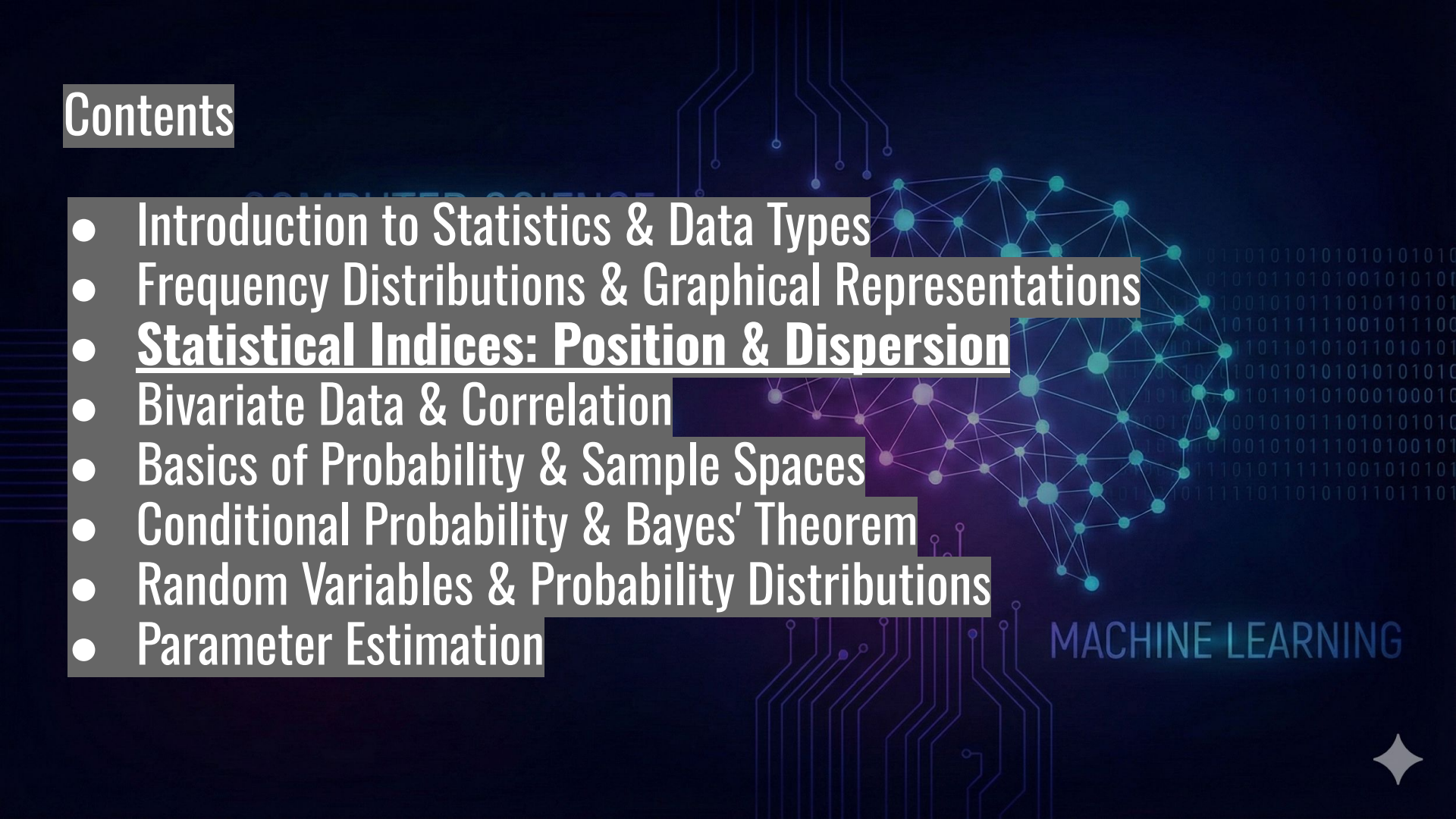
Used for categorical variables (e.g., Eye color).

- Non-adjacent rectangles.
- The height directly represents the relative or percentage frequency.
- Categories are reported on the x-axis in an arbitrary order.



Note: Bars are non-adjacent (separated) because the variable is categorical, not continuous.

Contents



SECTION 03

STATISTICAL INDICES

POSITION

The "Center of Gravity"

Mean • Median • Mode

DISPERSION

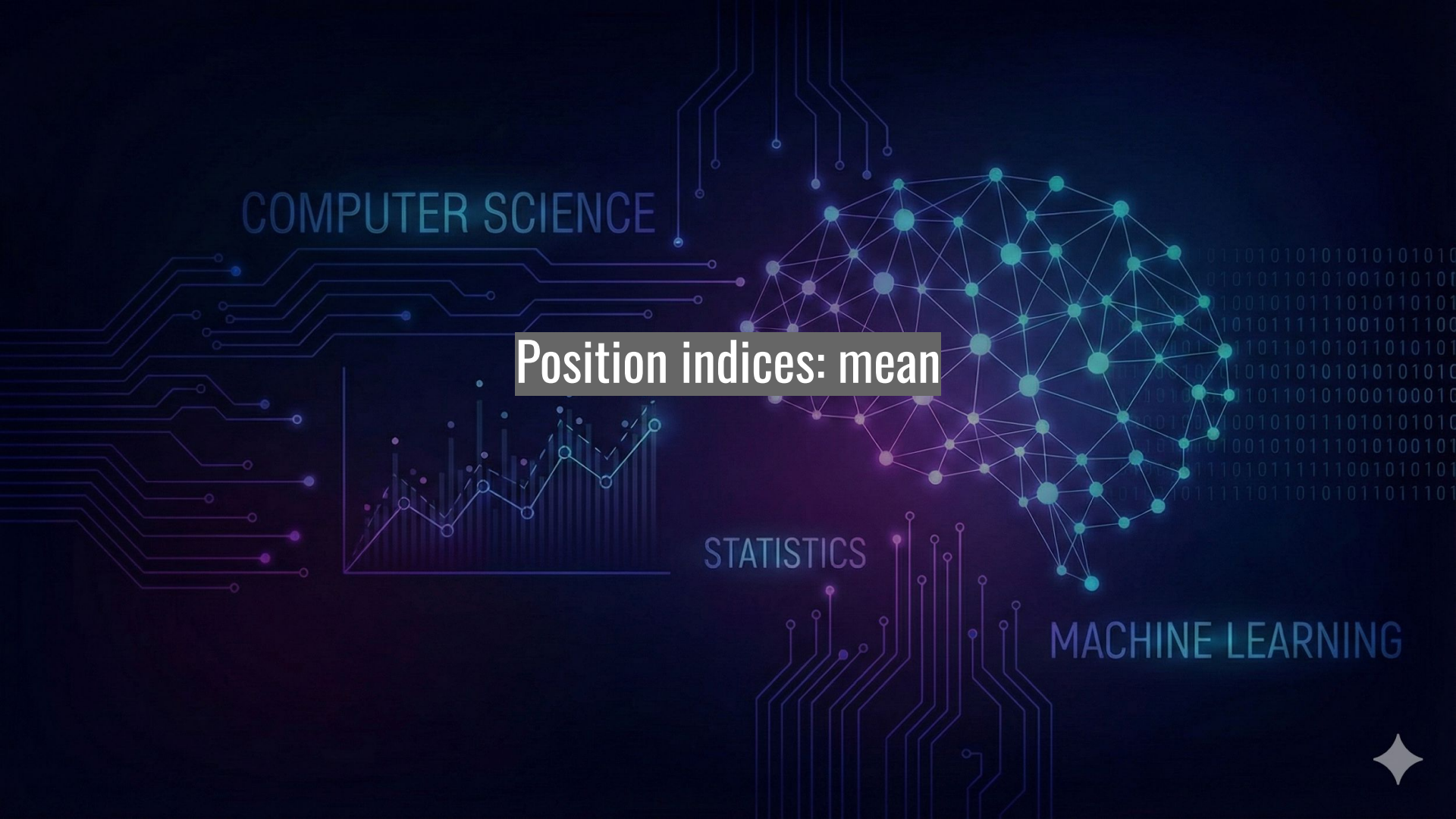
The "Spread" of Data

Variance • Std. Deviation

Introduction to Statistical Indices

To describe a dataset efficiently, we use two types of indices:

- **Position Indices: Help us understand "where" the distribution is located (e.g., Mean, Median).**
- **Dispersion Indices: Measure how "spread out" the data is around the central value (e.g., Variance, Standard Deviation).**


COMPUTER SCIENCE
Position indices: mean
STATISTICS
MACHINE LEARNING

Position Indices - The Arithmetic Mean

The Mean is the most common position index.

Where:N is the total population, **x_i are the values, f_i are frequencies, and p_i are relative frequencies.**

Note: The mean acts as the "center of gravity" of the data. It may be a value that does not actually exist in the dataset (e.g., the average family has 2.4 children).

$$\overline{X} = \frac{1}{N} \sum_{i=1}^{n} x_i f_i = \sum_{i=1}^{n} x_i p_i$$

Position Indices - The Arithmetic Mean

When data is grouped into classes (e.g., $110 < D \leq 130$), we cannot use exact individual values.

Methodology:

- Identify the Central Value (Midpoint) of each class.
- Assume all observations in that class coincide with the midpoint.
- Calculate the weighted average using these midpoints.

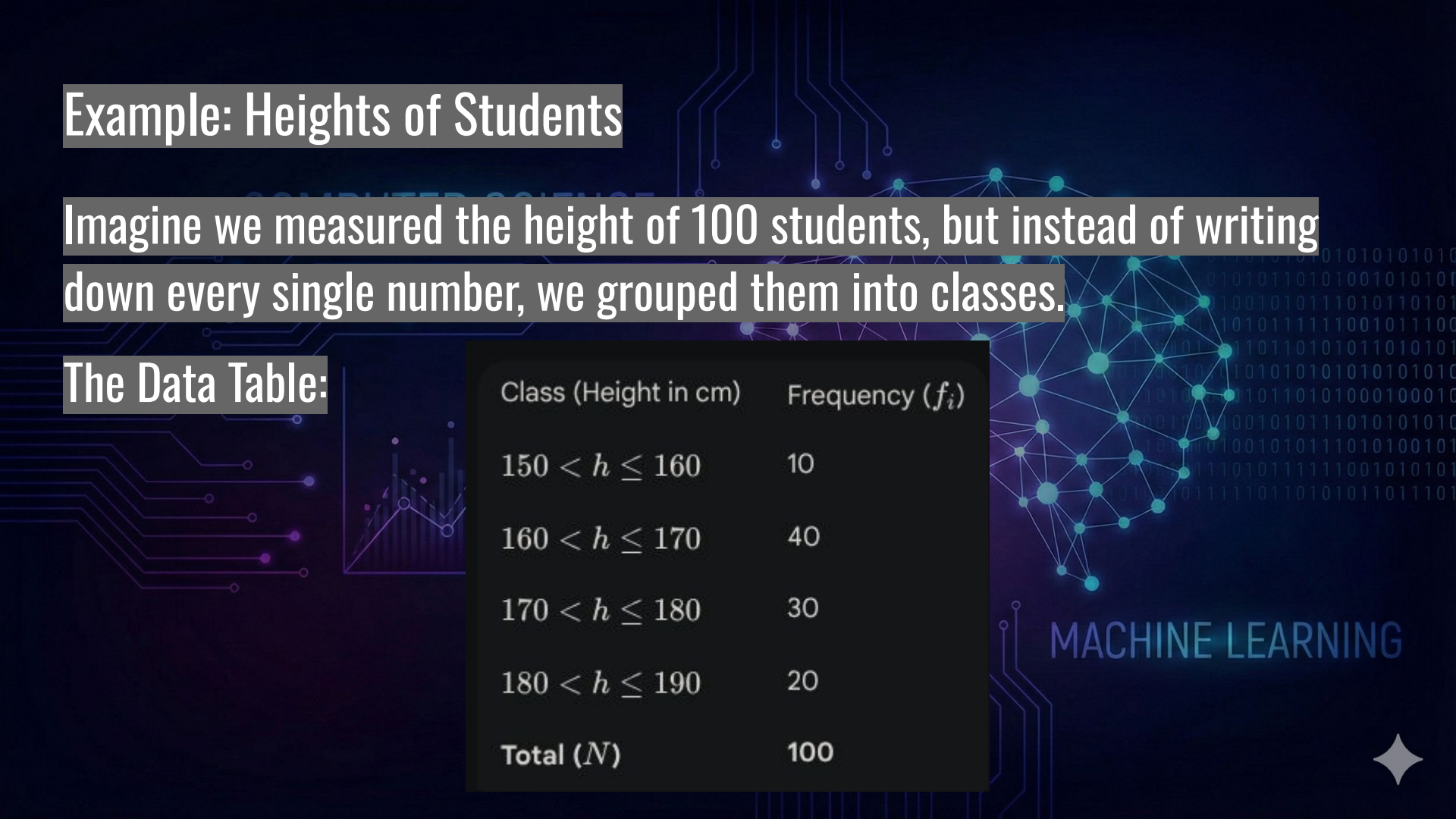
the Mean for Grouped Data is an approximation of the true mean computed with non-grouped data, so it is not exactly equal, but it is usually very close.

Example: Heights of Students

Imagine we measured the height of 100 students, but instead of writing down every single number, we grouped them into classes.

The Data Table:

Class (Height in cm)	Frequency (f_i)
$150 < h \leq 160$	10
$160 < h \leq 170$	40
$170 < h \leq 180$	30
$180 < h \leq 190$	20
Total (N)	**100**



Example: Heights of Students

Find the Central Value (Midpoint) for each class x_i: The midpoint is the average of the class boundaries: (Lower + Upper)/2.

- Class 1 (150-160): $\frac{150+160}{2} = 155$
- Class 2 (160-170): $\frac{160+170}{2} = 165$
- Class 3 (170-180): $\frac{170+180}{2} = 175$
- Class 4 (180-190): $\frac{180+190}{2} = 185$

Example: Heights of Students

Multiply Midpoint by Frequency ($x_i * f_i$): We assume every student in the "150-160" group is exactly 155 cm tall

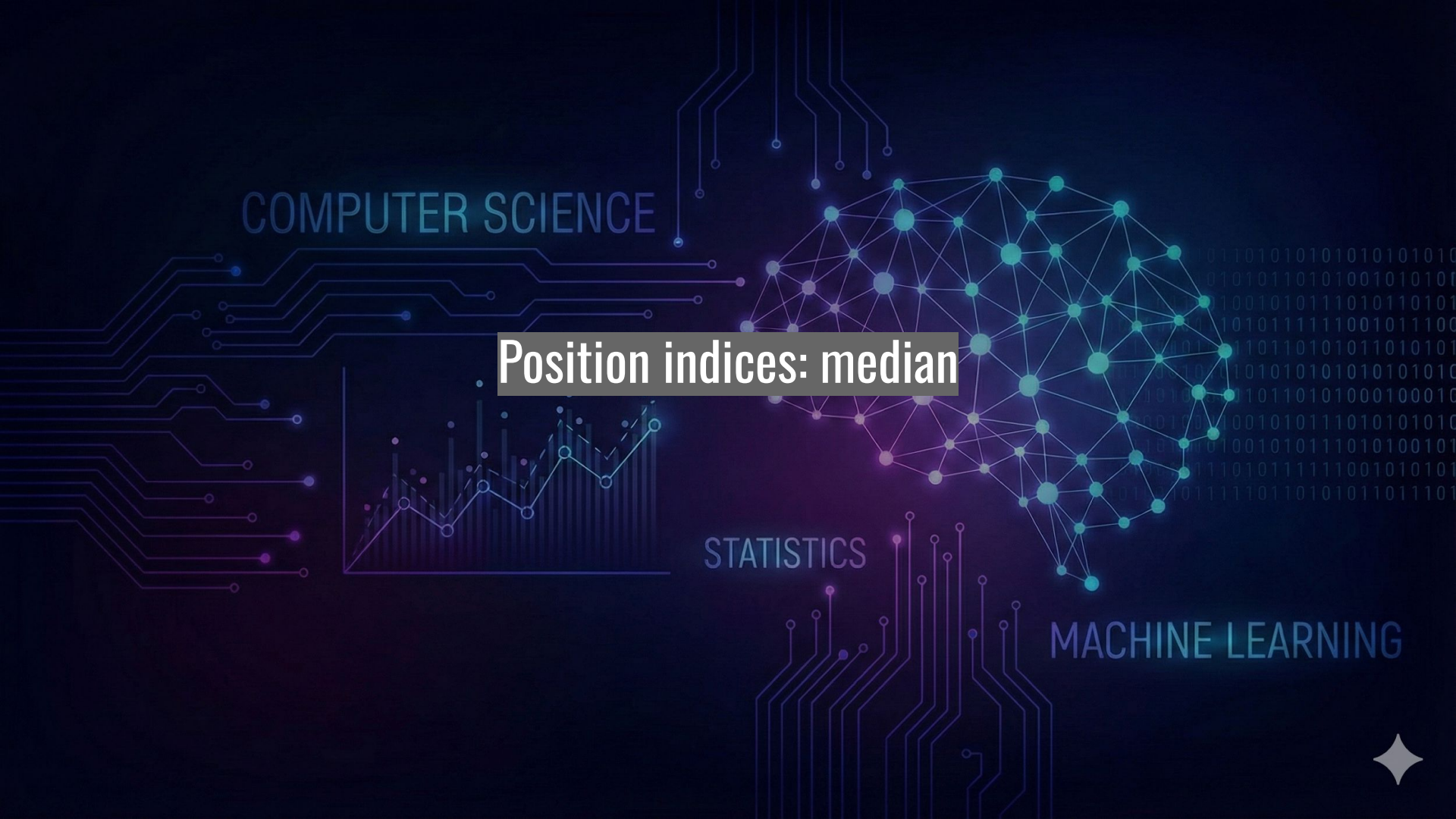
- $155 \times 10 = 1,550$
- $165 \times 40 = 6,600$
- $175 \times 30 = 5,250$
- $185 \times 20 = 3,700$

Example: Heights of Students

Sum the Results and Divide by Total Observations (N)

$$1,550 + 6,600 + 5,250 + 3,700 = \mathbf{17,100}$$

$$\text{Mean } (\overline{X}) = \frac{17,100}{100} = \mathbf{171} \text{ cm}$$

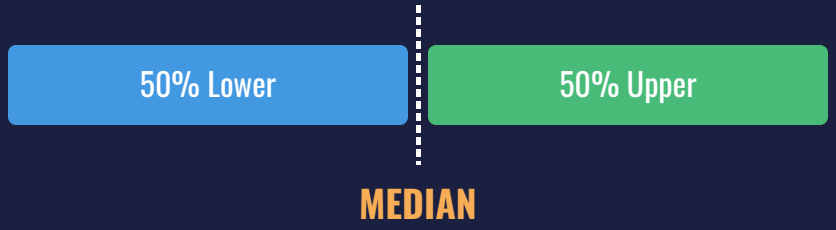

COMPUTER SCIENCE
Position indices: median
STATISTICS
MACHINE LEARNING

Position Indices: The Median

Definition

The Median is the value that splits the ordered population into two equal parts.

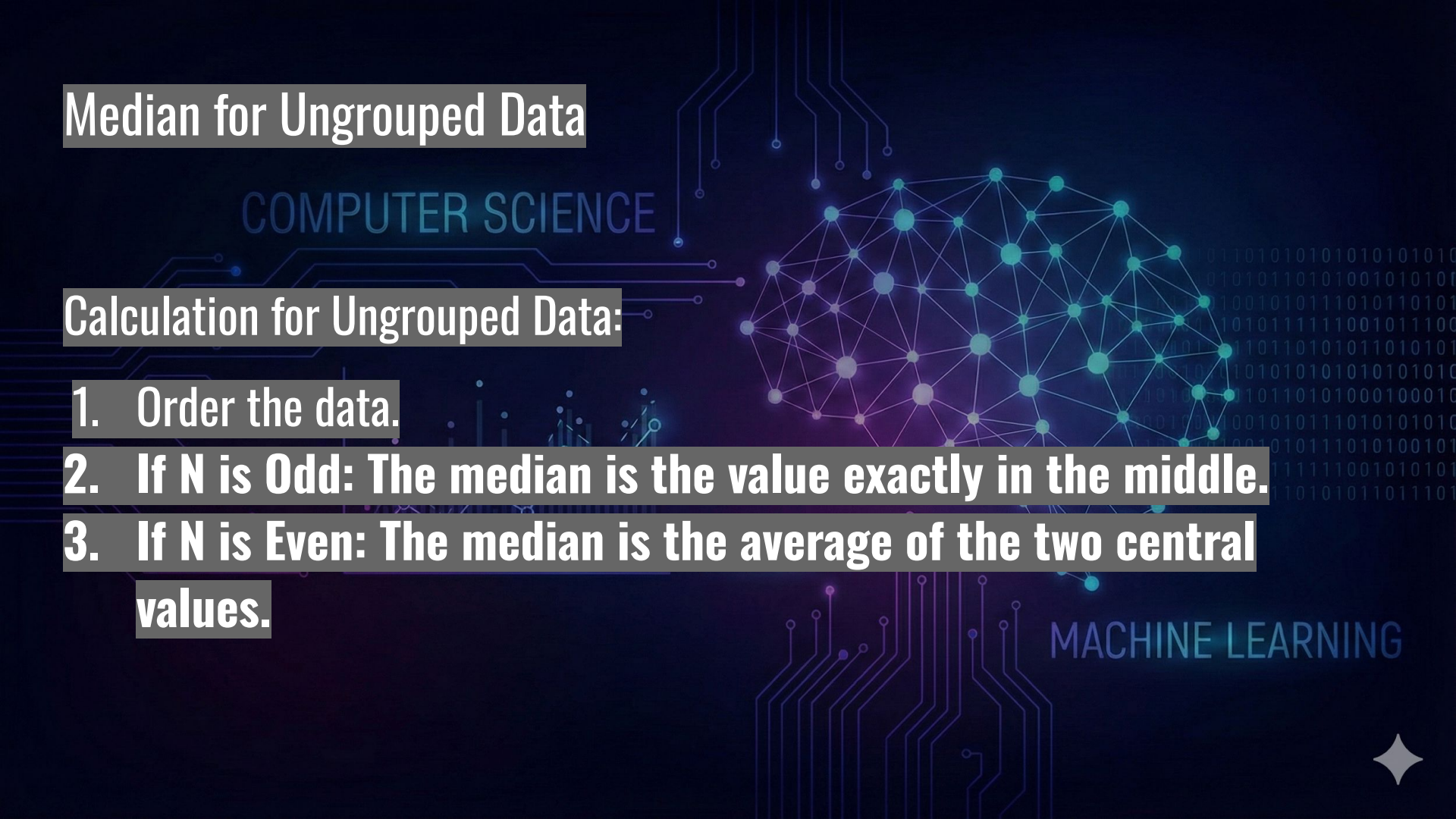
Ordered Population



Key Characteristics

- **Order Dependent:** Data must be sorted from lowest to highest before calculation.
- **Robust Statistic:** Unlike the mean, the median is not heavily influenced by extreme outliers.
- **Geometric Center:** It represents the "middle" observation in a dataset.

Median for Ungrouped Data



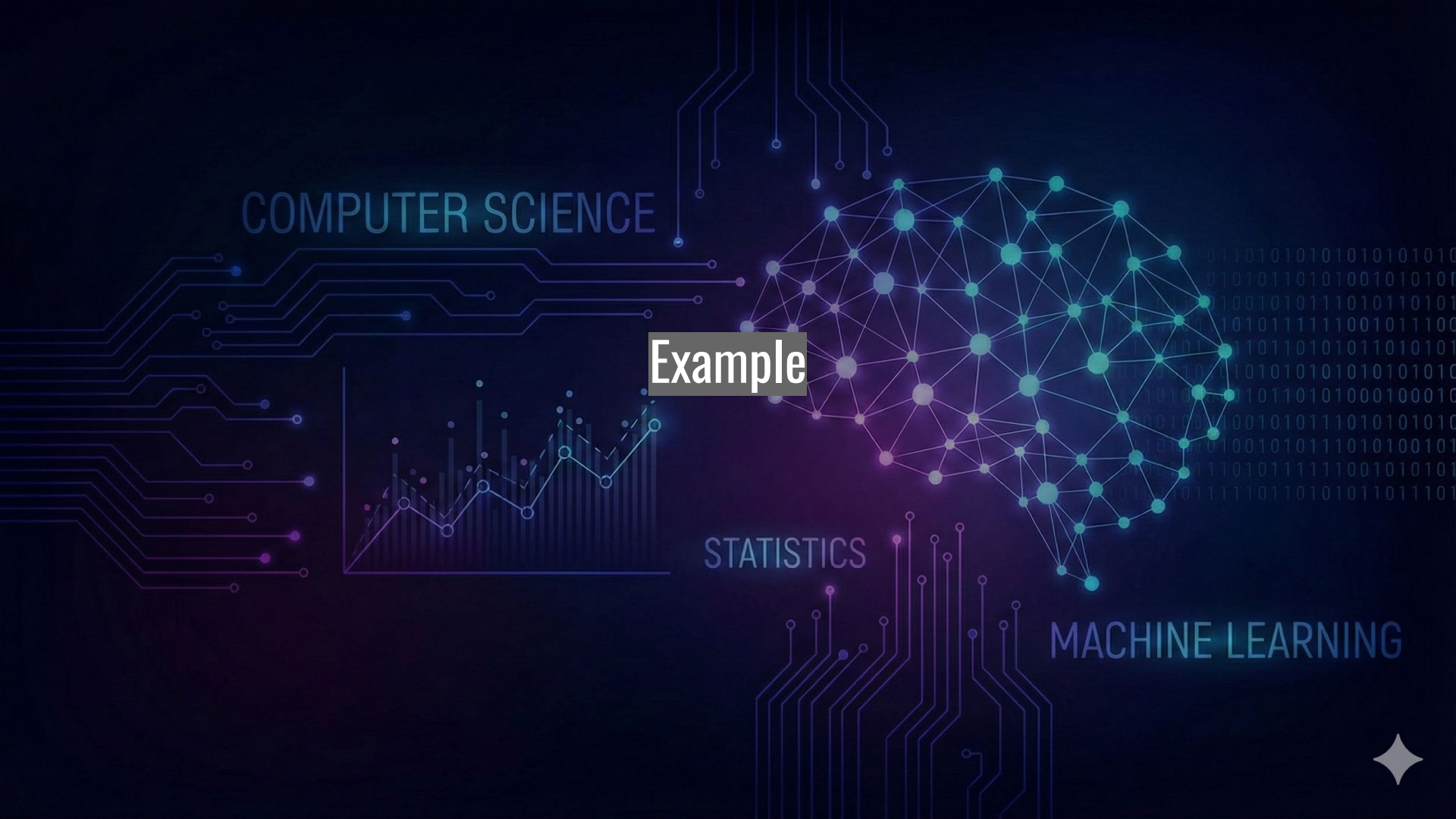
Calculation for Ungrouped Data:

1. Order the data.
2. **If N is Odd: The median is the value exactly in the middle.**
3. **If N is Even: The median is the average of the two central values.**

Median for Grouped Data (Geometric Method)

For grouped data (histograms), the Median is the value on the x-axis that divides the area of the histogram exactly in half.

1. **Using Histogram: Find x such that Area(Left) = Area(Right).**
2. **Using Ogive (Cumulative Freq): Find x such that the Cumulative Frequency is 50% (or 0.5).**


COMPUTER SCIENCE
Example
STATISTICS
MACHINE LEARNING

Position Indices - The Median

Imagine a small class of 5 students took a test. Here are their scores: 85, 92, 76, 88, 95

Case 1: Odd Number of Observations (N = 5)

1. Order the Data: First, we must arrange the scores from lowest to highest.
    a. 76, 85, 88, 92, 95
2. Find the Middle: Since there are 5 numbers, the middle one is the 3rd one.
    a. 76, 85, [88], 92, 95

The Median is 88.(Note: It splits the class perfectly: 2 students did worse, 2 students did better.)

Position Indices - The Median

Even Number of Observations (N = 6)

Now, imagine a 6th student joins and scores a 79. The new dataset is: 85, 92, 76, 88, 95, 79

1. Order the Data:
    a. 76, 79, 85, 88, 92, 95
2. Find the Middle: Since there are 6 numbers, there is no single middle number.
3. The "middle" falls between the 3rd and 4th numbers.
    a. 76, 79, [85, 88], 92, 95

Average the Middle Two: {85 + 88}/2 = 86.5

Histogram Analysis: Median vs. Mean

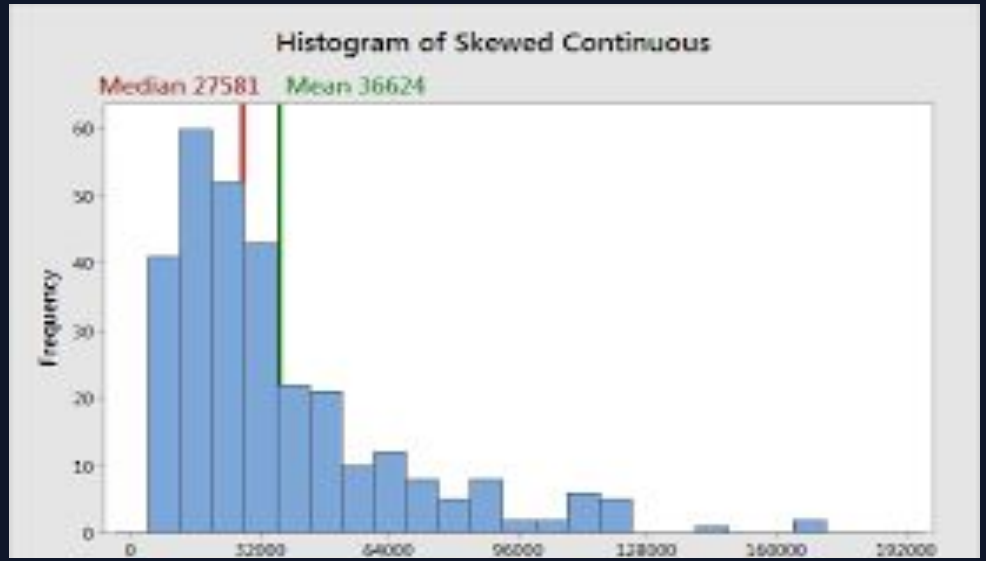
Visualizing Central Tendency in Skewed Distributions

The Median (Area Half-Point)

The value that splits the total histogram area into two equal halves. It minimizes the sum of absolute deviations.

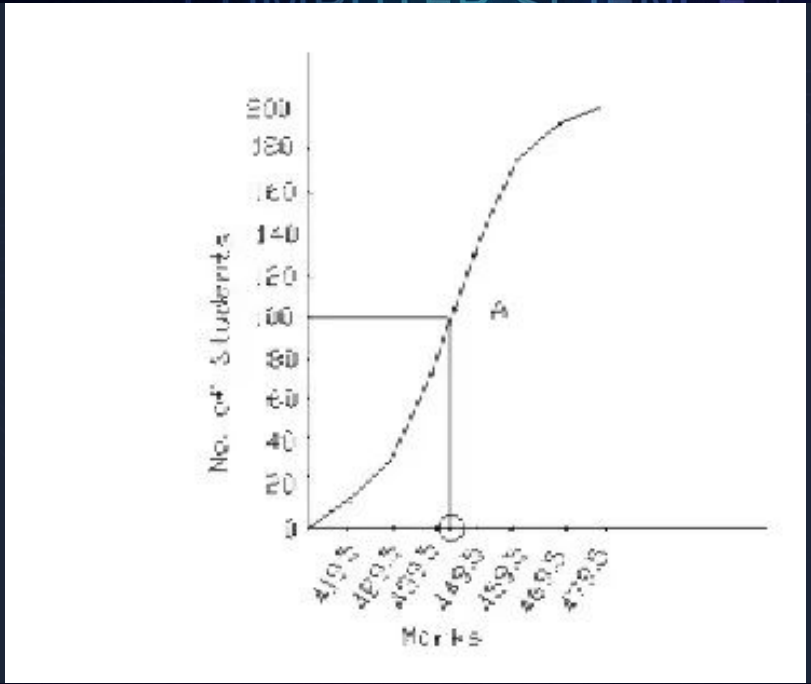
The Mean (Balance Point)

The mathematical center of gravity where the distribution would balance. Highly sensitive to outliers and extreme values.



*In right-skewed data, the **Mean** is pulled toward the extreme values in the tail, while the **Median** remains more resistant to these outliers.*

Median Calculation via Ogive (Cumulative Graph)

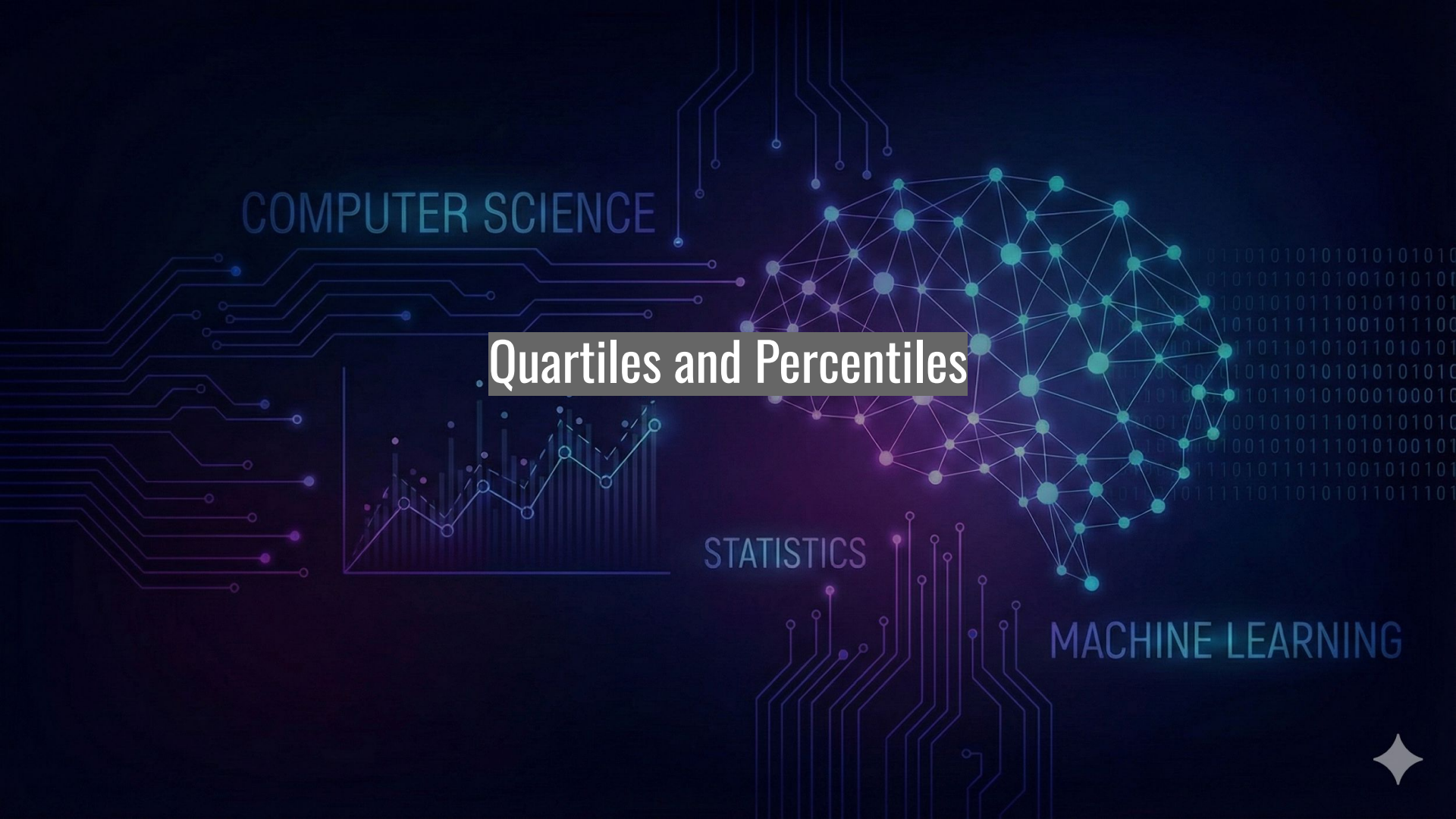


The Geometric Method

For grouped data, the Median is the value on the x-axis where the cumulative frequency reaches exactly 50% (or 0.5) of the total population.

Step-by-Step Estimation

- 1. Identify N/2: Locate the 50% point on the Y-axis (Number of Students).
- 2. Horizontal Projection: Draw a line from this point to the curve (Point A).
- 3. Vertical Projection: Drop a line from Point A down to the X-axis (Marks).
- 4. Read the Value: The intersection on the X-axis is the Estimated Median.


COMPUTER SCIENCE
Quartiles and Percentiles
STATISTICS
MACHINE LEARNING

Quartiles and Percentiles

These extend the concept of the Median.

- Quartiles (q): Divide the population into 4 equal parts.
  - (1st Quartile): Leaves 25% of data to the left
  - .(2nd Quartile): Leaves 50% to the left (= Median).
  - (3rd Quartile): Leaves 75% to the left.

Percentiles: Divide the population into 100 equal parts (e.g., 90th percentile).

The Fractiles: Slicing the Data (N = 40)

1st Quartile (Q_1)
Leaves 25% of data to the left.
Target CF = 10

2nd Quartile (Q_2)
Leaves 50% to the left (The Median).
Target CF = 20

3rd Quartile (Q_3)
Leaves 75% of data to the left.
Target CF = 30

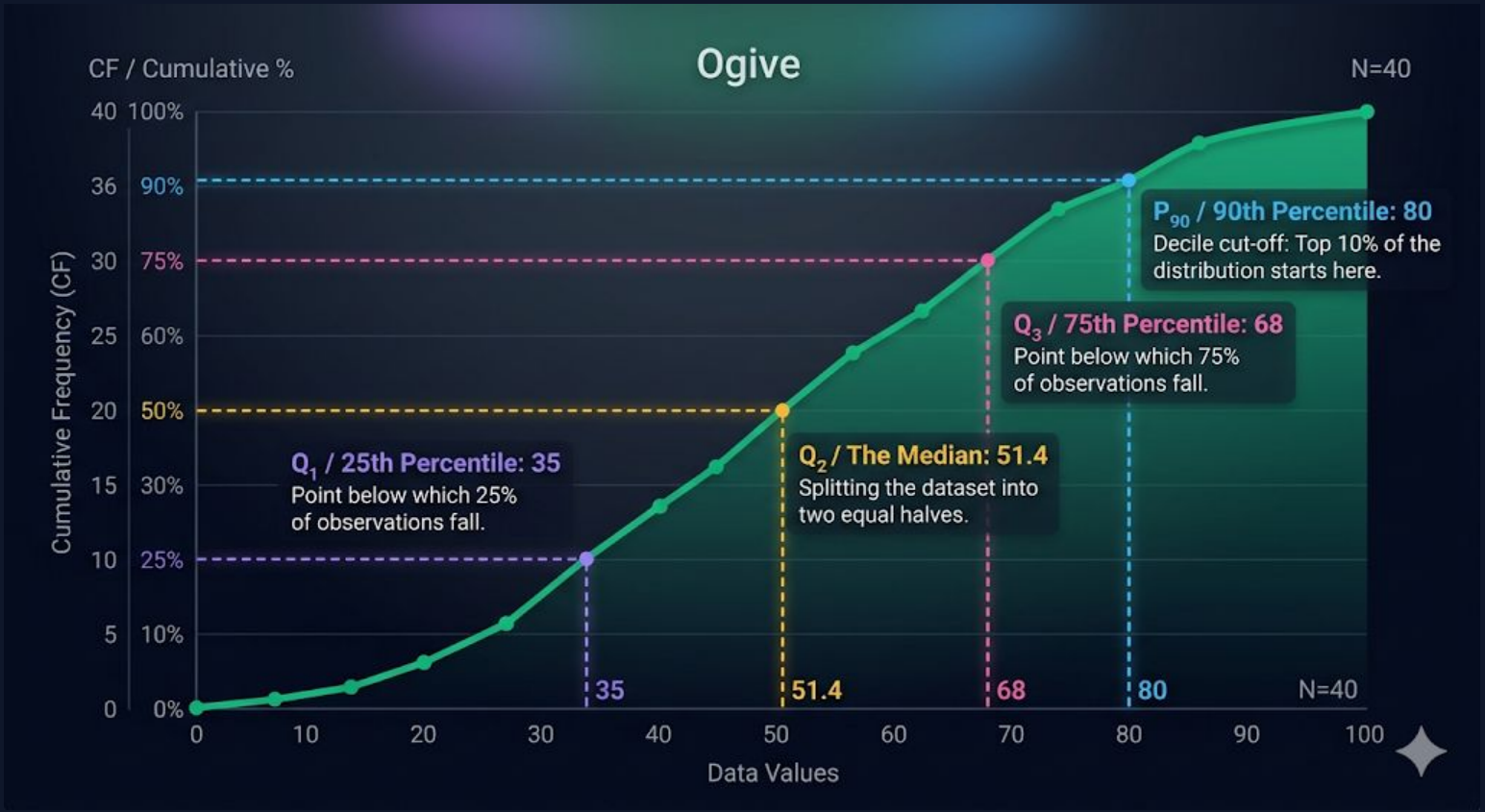
90th Percentile (P_{90})
Leaves 90% of data to the left.
Target CF = 36

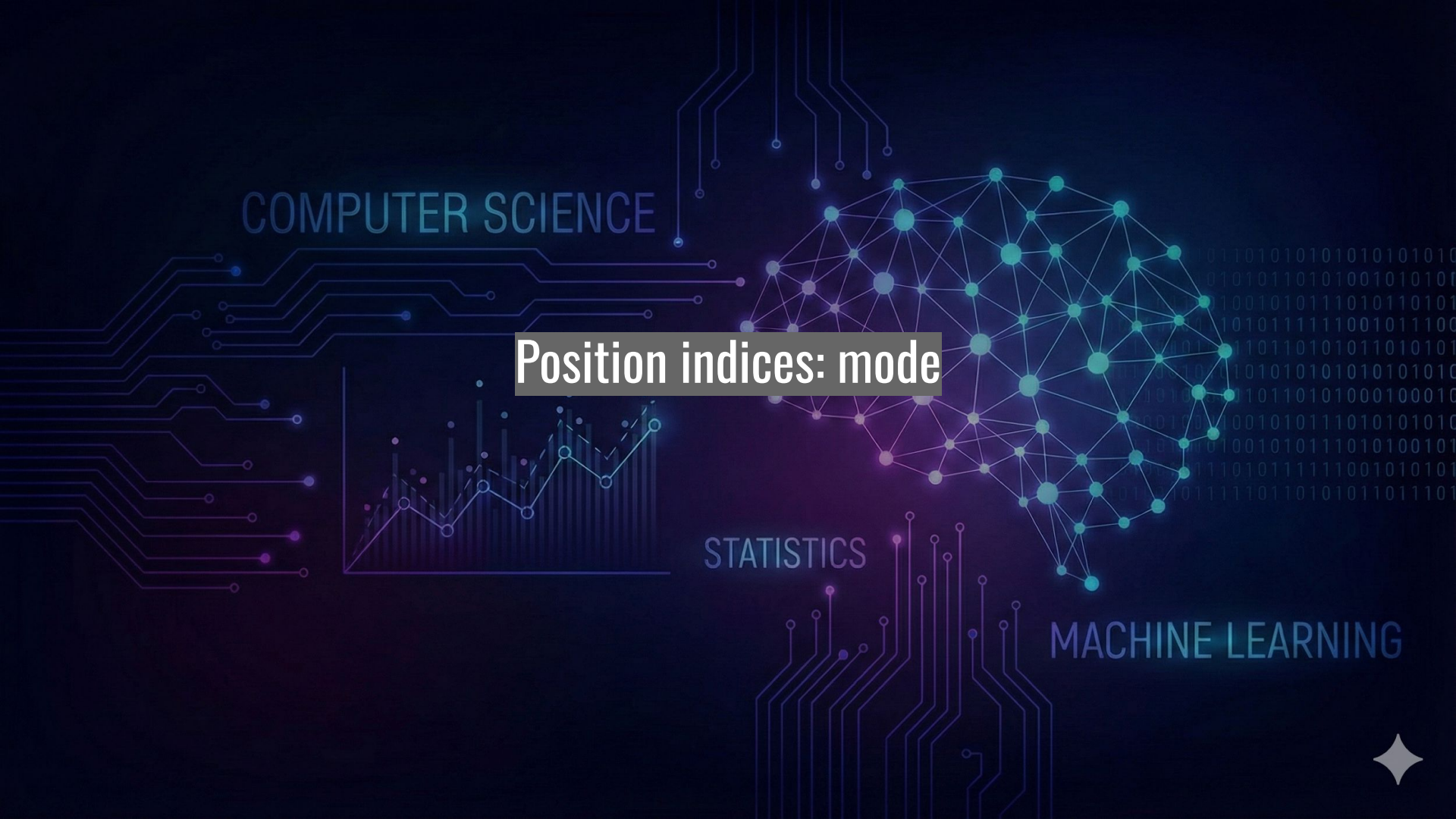
Class (Values)	Frequency (f)	Cumulative Freq (CF)	Contains...
0 - 20	4	4	
20 - 40	8	12	**Q_1 (Crosses 10)**
40 - 60	14	26	**Q_2 / Median (Crosses 20)**
60 - 80	10	36	**Q_3 (Crosses 30) & P_{90} (Exactly 36)**
80 - 100	4	40	

Histogram: Dividing the Area into 4 Equal Quarters

Total Area = 40. Q_1, Q_2, and Q_3 slice the histogram into 4 blocks, each with an exact area of 10.

The Ogive: Locating Specific Percentiles




COMPUTER SCIENCE
Position indices: mode
STATISTICS
MACHINE LEARNING

Position Indices: The Mode

Definition The Mode is simply the value that appears most frequently in a dataset.

- Unlike the Mean, the Mode can be used for categorical data (e.g., "The most popular eye color is Brown")
- .A dataset can have:
  - One Mode (Unimodal)
  - Two Modes (Bimodal)
  - Multimodal
  - No Mode (if all values are unique)

On a histogram, the Mode corresponds to the highest peak.

Position Indices: The Mode

.A dataset can have:

- One Mode (Unimodal)
- Two Modes (Bimodal)
- No Mode (if all values are unique)

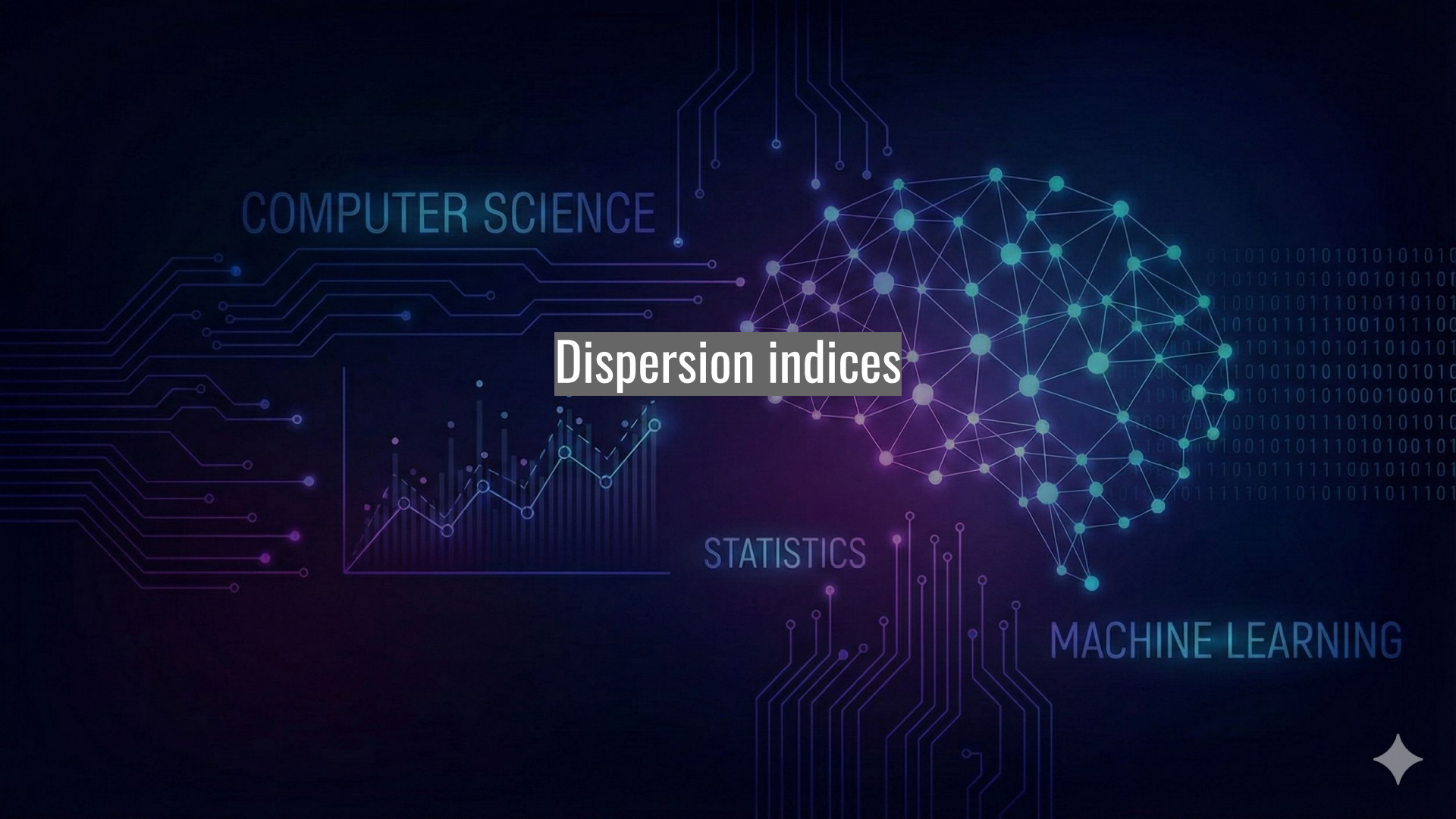
Position Indices: The Mode

Multimodal: Three or more numbers tie for the highest frequency.

- **Example: {1, 1, 2, 3, 3, 4, 5, 5} — The modes are 1, 3, and 5.**

No Mode: Every value in the dataset appears the exact same number of times (usually just once), meaning no number stands out.

- **Example: {2, 4, 6, 8, 10} — There is no mode.**


COMPUTER SCIENCE
Dispersion indices
STATISTICS
MACHINE LEARNING

Dispersion Indices

Knowing the center (Mean) is not enough; we need to know how concentrated the data is.

- Problem: The sum of deviations from the mean is always zero:
- Solution: We measure dispersion using squared deviations.

$$\sum(x_i - \overline{X}) = 0$$

Variance (s^2) and Standard Deviation (s)

Variance (s_X^2): The average of the squared deviations from the mean

Standard Deviation (s_X): The square root of the variance

Note: Variance is measured in [unit]^2, Standard Deviation in [unit].
They are always ≥ 0

$$s_X^2 = \frac{1}{N}\sum_{i=1}^{n}(x_i - \overline{X})^2 f_i$$

Calculation Shortcut for Variance

Calculating deviations for every point can be tedious. A mathematically equivalent and faster formula is

:The Variance is the Mean of the Squares minus the Square of the Mean.

1. Calculate the average of the squared values .
2. Calculate the square of the average .
3. Subtract the latter from the former.

$$s_X^2 = \overline{X^2} - (\overline{X})^2$$

The Great Divide: N vs. N-1

Population (Complete Data)

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

Used when you have data for every member of the group.

Sample (Subset)

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

Used when you have a random subset to estimate the whole.

The Bias Problem

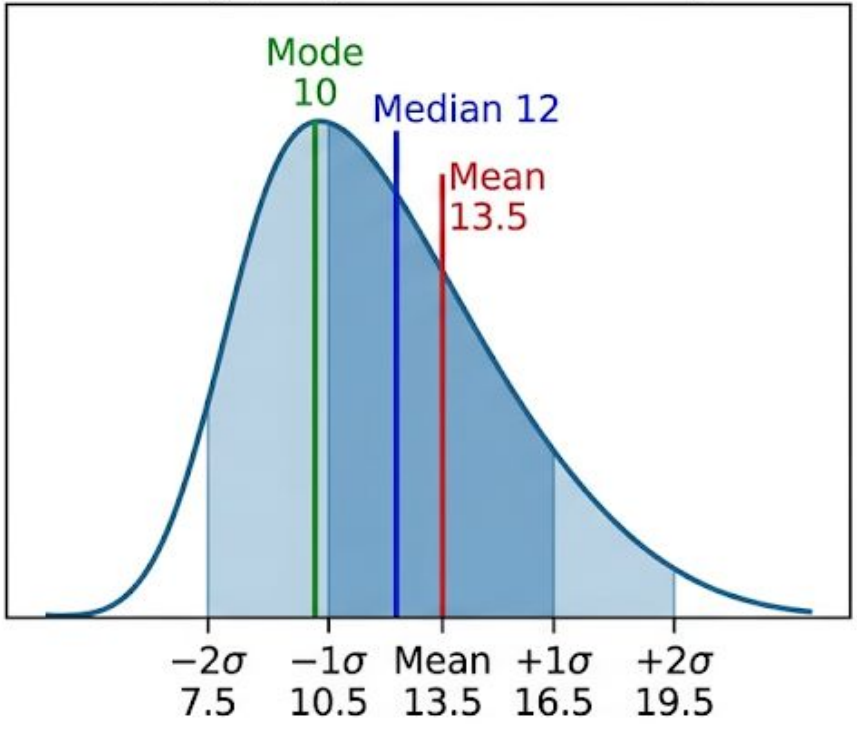
A sample is almost always less diverse than the full population. If we divide by N, we consistently **underestimate** the true variability.

Bessel's Correction

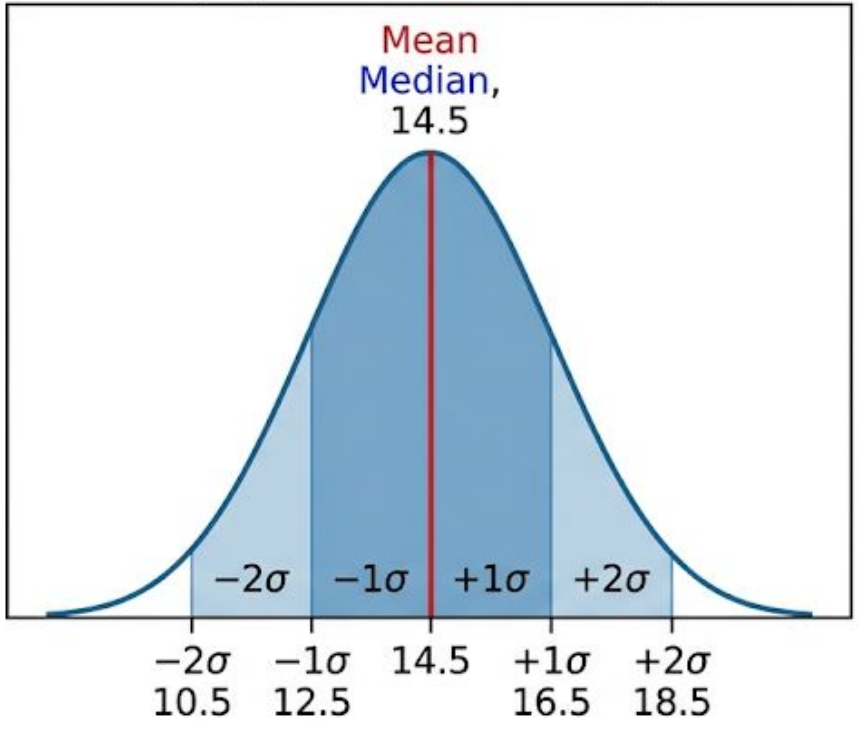
Dividing by **N-1** makes the result slightly larger. This mathematically corrects the bias, providing a more accurate estimate of the population's true spread.

Comparing Measures of Central Tendency and Dispersion

Skewed Distribution
(e.g., Right-Skewed Data)



Normal Distribution
(e.g., Bell Curve Data)



Note: Shaded regions represent ranges of Mean ± σ and Mean ± 2σ

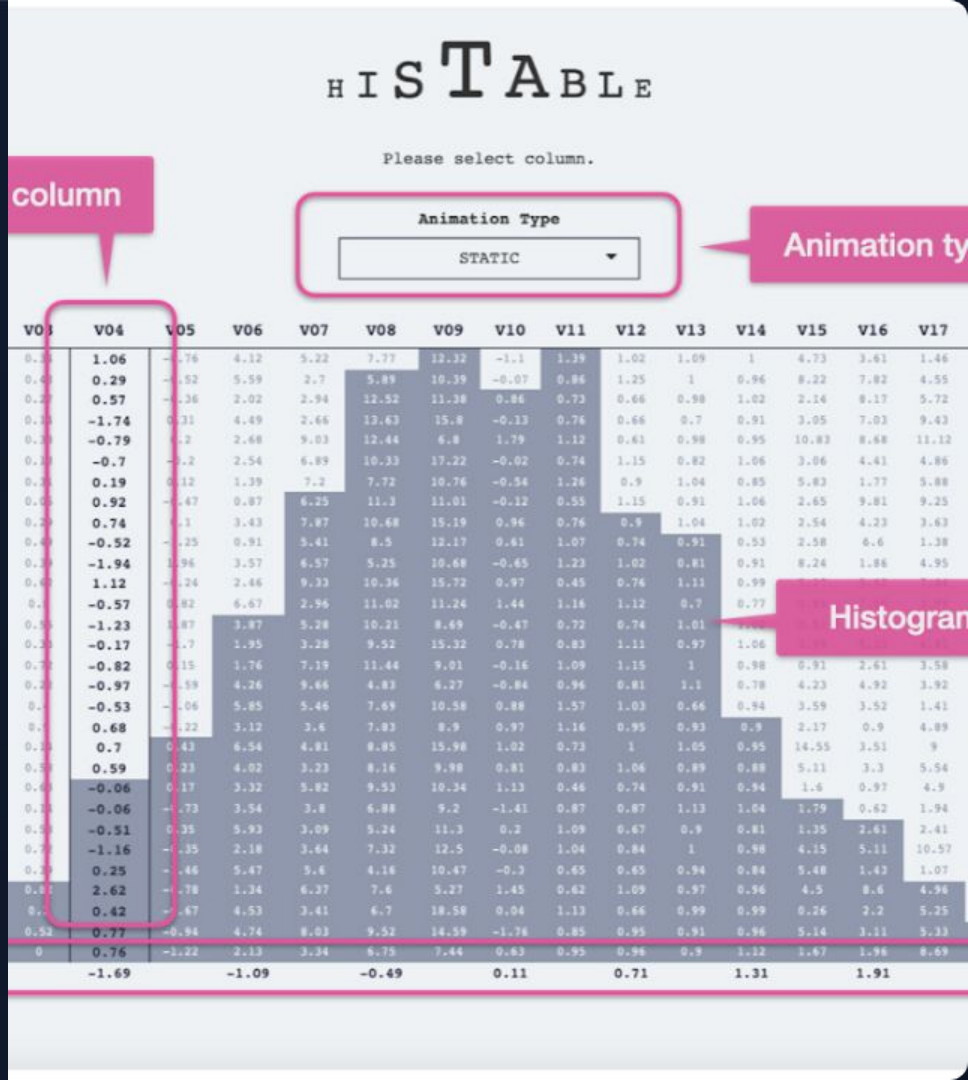
Mastering Histogram Binning

A comprehensive guide to finding the perfect bin size to reveal the true shape of your data distributions.

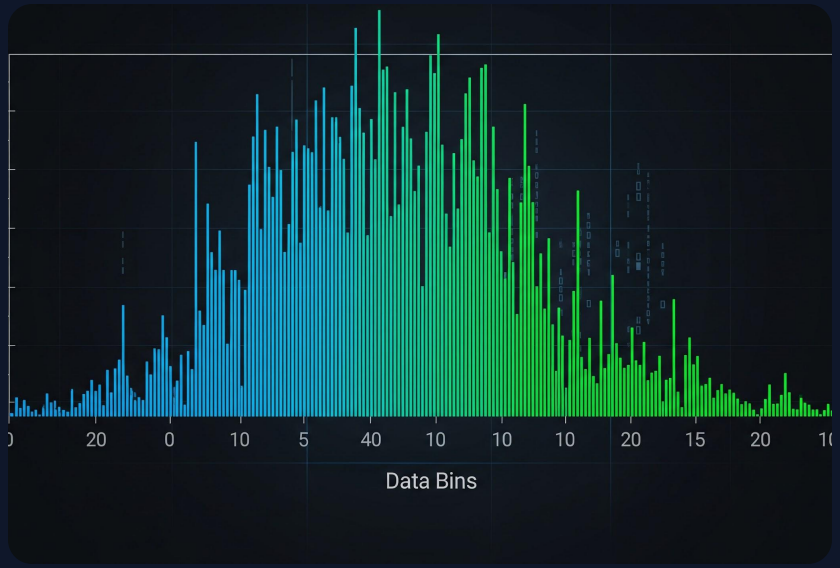
The Binning Problem

Determining the correct number of classes (bins) is crucial for accurate data visualization:

- **Overfitting (Too many bins):** The histogram will have too many "holes" and look like a comb. Statistical noise obscures the real trend.
- **Underfitting (Too few bins):** You will lose critical distribution details, and the data will appear as a single, indistinct block.
- **The Goal:** Find a statistical balance that accurately represents the underlying probability density function.



Visualizing Binning Errors



Overfitting: Too Many Bins

Visual Sign: The "Comb" effect. High variance and "holes" in the data.

Impact: Statistical noise obscures the true density function.

Underfitting: Too Few Bins

Visual Sign: A single, blocky mass. High bias.

Impact: Critical details of the distribution's shape are completely lost.

1. Square-Root Choice

The Baseline Method

This is the simplest and fastest method, frequently used in basic statistics courses and standard spreadsheet software.

$$k = \lceil \sqrt{n} \rceil$$

- **Calculates:** Number of bins (*k*) based on sample size (*n*).
- **When to use:** Excellent for small to medium-sized datasets.
- **Example:** For *n* = 100 measurements, you use *k* = 10 bins.

2. Sturges' Formula

The Statistical Default

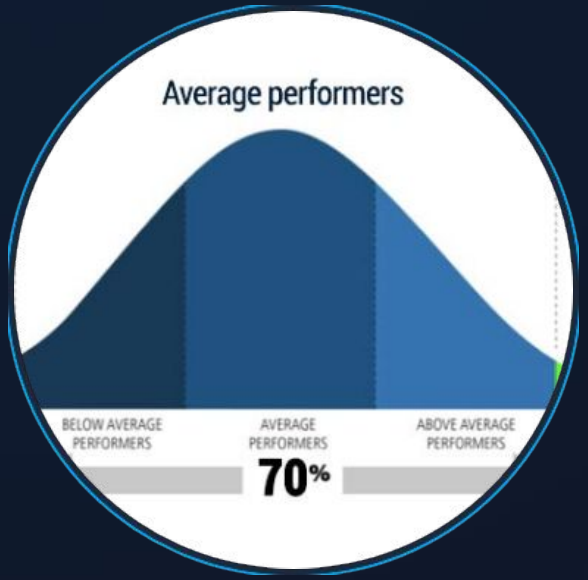
Sturges' formula is the default method in many statistical software packages, including R and basic Python plotting libraries.

$$k = \lceil 1 + \log_2(n) \rceil$$

Assumptions & Limitations

- **Assumption:** It implicitly assumes that the data follows an approximately normal (Gaussian) distribution.
- **When to use:** Works very well for datasets where **n < 200**.
- **When to avoid:** Poor choice when data has highly asymmetrical distributions (strongly skewed) or very large sample sizes.

3. Scott's Normal Reference Rule



Targeting Bin Width (W)

Unlike previous methods, Scott's Rule calculates the optimal bin **width** directly, relying on the data's dispersion (Standard Deviation, σ).

$$W = \frac{3.49 \cdot \sigma}{\sqrt[3]{n}}$$

When to use: Ideal for distributions close to normal but with larger datasets. Once *W* is found, calculate *k* by dividing the total range (Max - Min) by *W*.

4. Freedman-Diaconis Rule

Robust Against Outliers

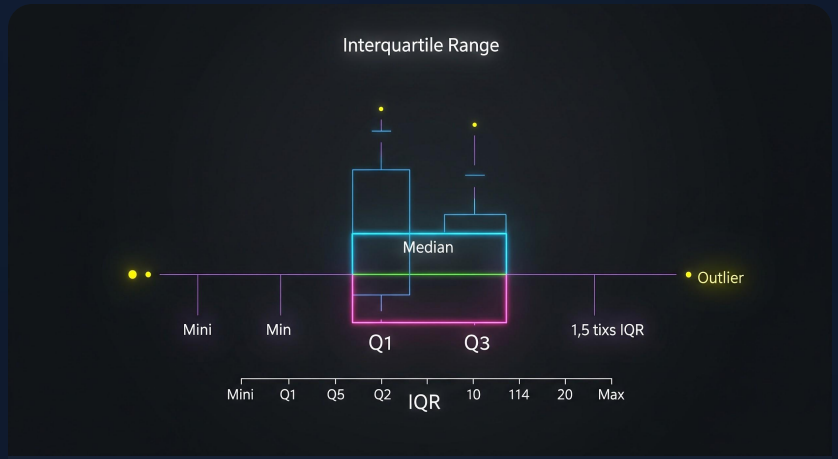
Considered one of the most robust methods available. Instead of standard deviation, it uses the Interquartile Range (IQR = Q3 - Q1), making it resistant to severe outliers.

$$W = 2 \cdot \frac{IQR}{\sqrt[3]{n}}$$

Application Scenario

- **Calculates:** Optimal bin width (*W*).
- **When to use:** Strongly recommended for complex real-world data, heavily skewed distributions, and datasets with "heavy tails".
- **Why it works:** By isolating the middle 50% of data (IQR), anomalous extremes do not distort the underlying bin scaling.

The Interquartile Range (IQR)



Mathematical Rigor

The difference **Q3 - Q1** defines the width of the central 50% of the distribution. It is the heart of the Freedman-Diaconis rule for defining histogram bins.

Robustness to Outliers

- **Beyond Standard Deviation:** Unlike variance, the IQR is not "skewed" by extreme individual values or anomalies in the tails.
- **Central Focus:** Describes dispersion by completely ignoring the tails, making it an extremely robust measure.
- **Application:** Ideal for real-world datasets with skewed distributions or "heavy tails".

Practical Application Tips

Mathematical rigor must be balanced with human readability. A chart is useless if it cannot be quickly interpreted.

1. Calculate

Determine the theoretical range or width (W) using one of the formulas (preferably Scott or Freedman-Diaconis for robust datasets).

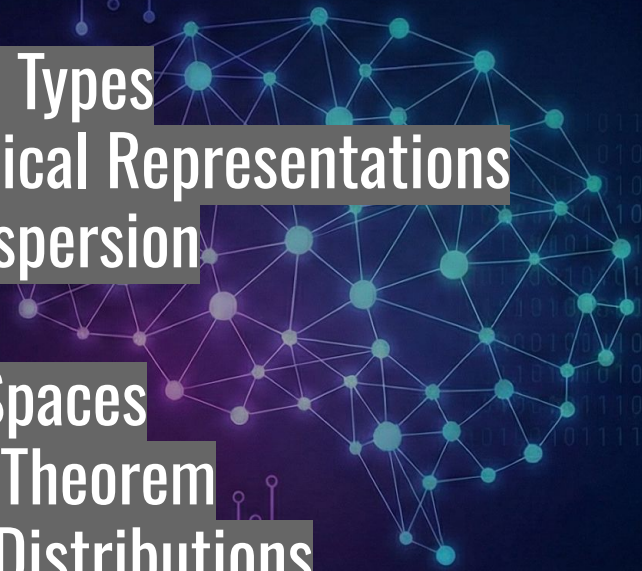
2. Round Safely

Always round the width to a "readable" number that scales well mathematically (e.g., multiples of 2, 5, 10, 20, or 50).

3. Optimize UX

If Scott's formula suggests W = 17.3, force classes of 20. An X-axis reading 0, 20, 40 is instantly clear; 0, 17.3, 34.6 is unreadable.

Contents

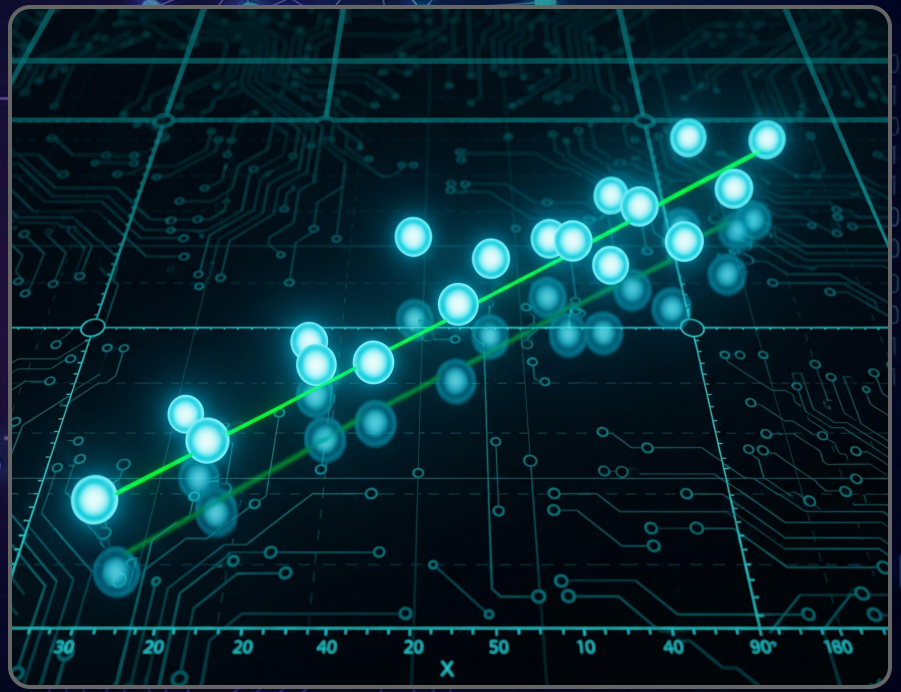


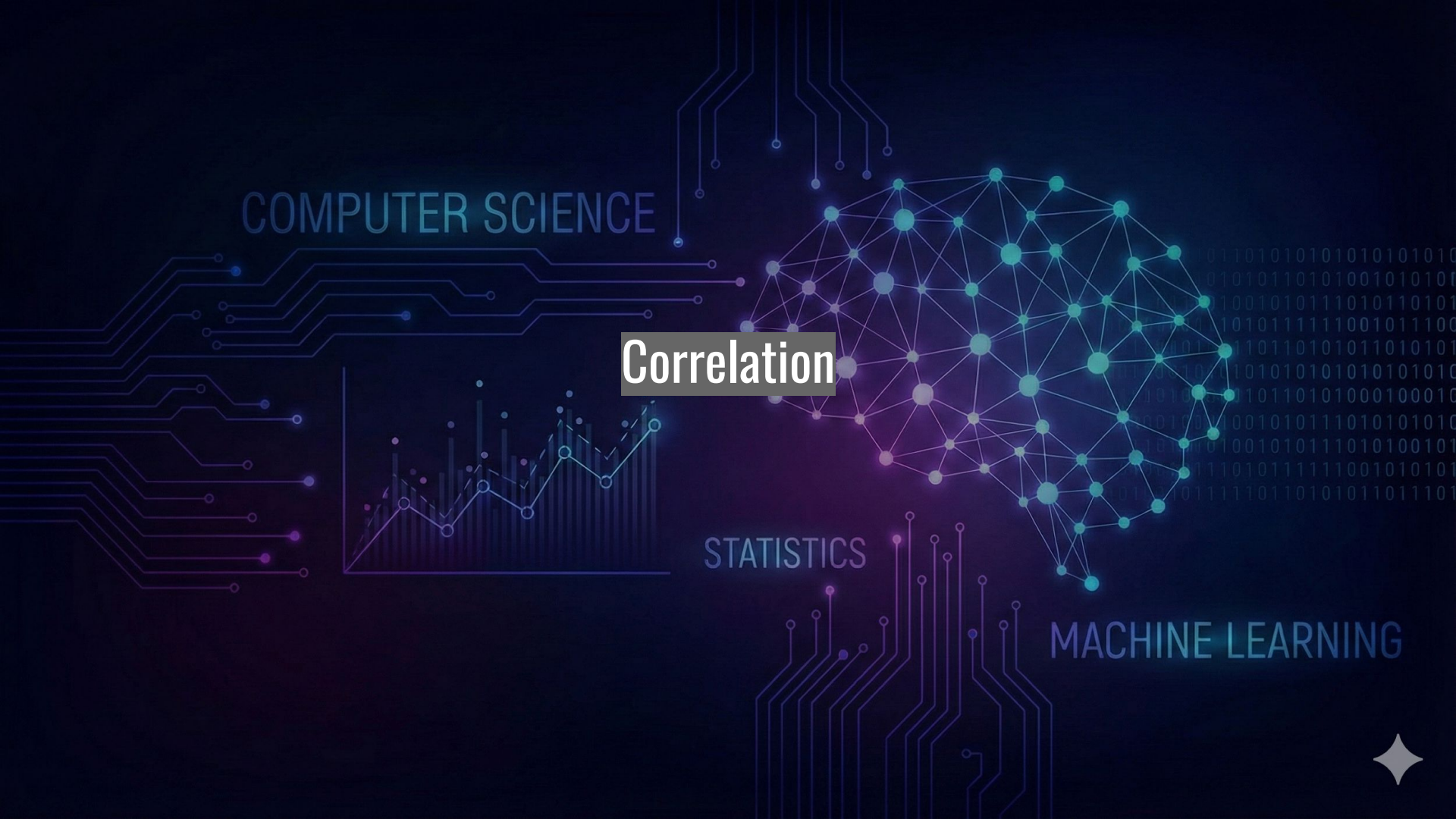
From Univariate to Bivariate Analysis

Shifting the Lens:

Move beyond analyzing single variables in isolation to explore how two distinct characteristics vary together.

- **The Scatterplot: Visualizing relationship patterns (linear, quadratic, or null).**
- **Covariance: Quantifying the directional relationship between two variables.**
- **Bivariate Pairs: Represented as (x_i, y_i) for each individual in the sample.**



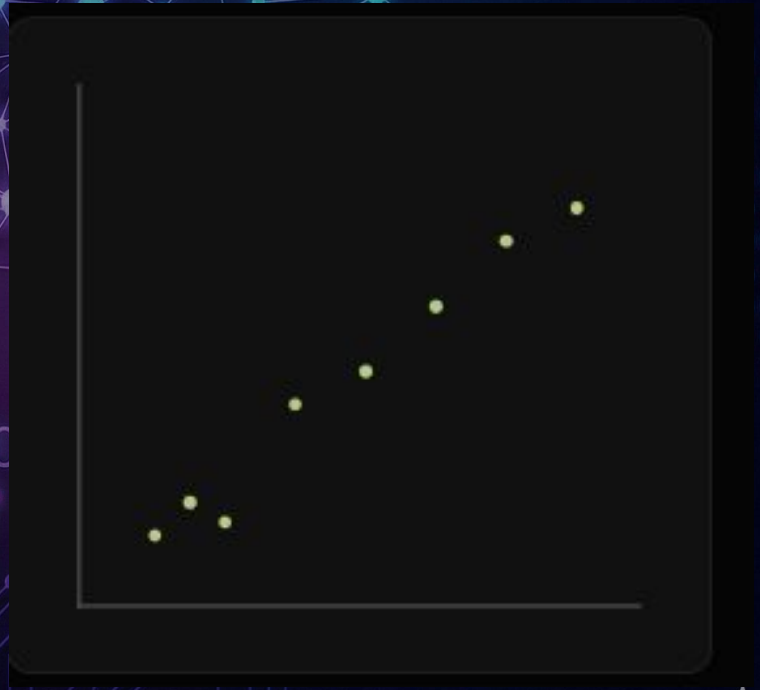

COMPUTER SCIENCE
Correlation
STATISTICS
MACHINE LEARNING

Bivariate Data & Scatterplots

The Concept. Often, we measure two characteristics for each individual (e.g., Height and Weight). We denote this as pairs: $(x_1, y,), (x,, y,), ..., (x,, y,)$

The Scatterplot: A graphical representation where each pair is a point on the Cartesian plane. It helps visual identification of patterns:

- **Linear trends**
- **Quadratic trends**
- **No correlation**



Covariance (s,)

Covariance measures how two variables vary together.

Interpretation

s, > 0: Direct correlation (Large X with Large Y).

s, < 0: Inverse correlation (Large X with Small Y).

s, = 0: Uncorrelated.

$$s_{XY} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{X})(y_i - \overline{Y})$$

$$s_{XY} = \overline{XY} - \overline{X} \cdot \overline{Y}$$

Correlation Coefficient

Why do we need it? Since covariance depends on units, we need a normalized index to measure the strength of the linear relationship.

Properties

- Always between -1 and 1:.
- 1: Perfect linear alignment.
- 0: No linear relationship.

$$\rho_{XY} = \frac{s_{XY}}{s_X s_Y}$$

Example

We want to see if there is a relationship between Study Hours (X) and Exam Scores (Y) for 5 students.

Data Points:

- Student A: 2 hours → Score 50
- Student B: 4 hours → Score 60
- Student C: 6 hours → Score 70
- Student D: 8 hours → Score 80
- Student E: 10 hours → Score 90

Example

First, we need the "center of gravity" for both variables, i.e., mean values

- Mean of Hours ($\overline{X}$):

$$\frac{2+4+6+8+10}{5} = \frac{30}{5} = 6$$

- Mean of Scores ($\overline{Y}$):

$$\frac{50+60+70+80+90}{5} = \frac{350}{5} = 70$$

Example

Now we see how far each student is from the average, i.e., the deviationa

Student	Hours (x_i)	$x_i - \overline{X}$	Score (y_i)	$y_i - \overline{Y}$	Product $(x - \overline{X})(y - \overline{Y})$
A	2	$2 - 6 = -4$	50	$50 - 70 = -20$	$(-4)(-20) = \mathbf{80}$
B	4	$4 - 6 = -2$	60	$60 - 70 = -10$	$(-2)(-10) = \mathbf{20}$
C	6	$6 - 6 = 0$	70	$70 - 70 = 0$	$(0)(0) = \mathbf{0}$
D	8	$8 - 6 = 2$	80	$80 - 70 = 10$	$(2)(10) = \mathbf{20}$
E	10	$10 - 6 = 4$	90	$90 - 70 = 20$	$(4)(20) = \mathbf{80}$
Sum					**200**

Example

Covariance is the average of these products.

$$s_{XY} = \frac{\text{Sum of Products}}{N} = \frac{200}{5} = 40$$

Interpretation: The covariance is positive (40 > 0), which means there is a direct relationship. As study hours go up, scores go up.

However, "40" is hard to interpret on its own because it depends on the units (Hours x Score)

Example

To normalize the covariance into a correlation coefficient, **we need the standard deviation of both variables.**

Variance of X (s_X^2):

$$\frac{(-4)^2 + (-2)^2 + 0^2 + 2^2 + 4^2}{5} = \frac{16 + 4 + 0 + 4 + 16}{5} = \frac{40}{5} = 8$$

Standard Deviation of X (s_X):

$$\sqrt{8} \approx \mathbf{2.828}$$

Example

To normalize the covariance into a correlation coefficient, **we need the standard deviation of both variables.**

Variance of Y (s_Y^2):

$$\frac{(-20)^2 + (-10)^2 + 0^2 + 10^2 + 20^2}{5} = \frac{400 + 100 + 0 + 100 + 400}{5} = \frac{1000}{5}$$

Standard Deviation of Y (s_Y):

$$\sqrt{200} \approx \mathbf{14.142}$$

Example

Now we divide the Covariance by the product of the Standard Deviations.

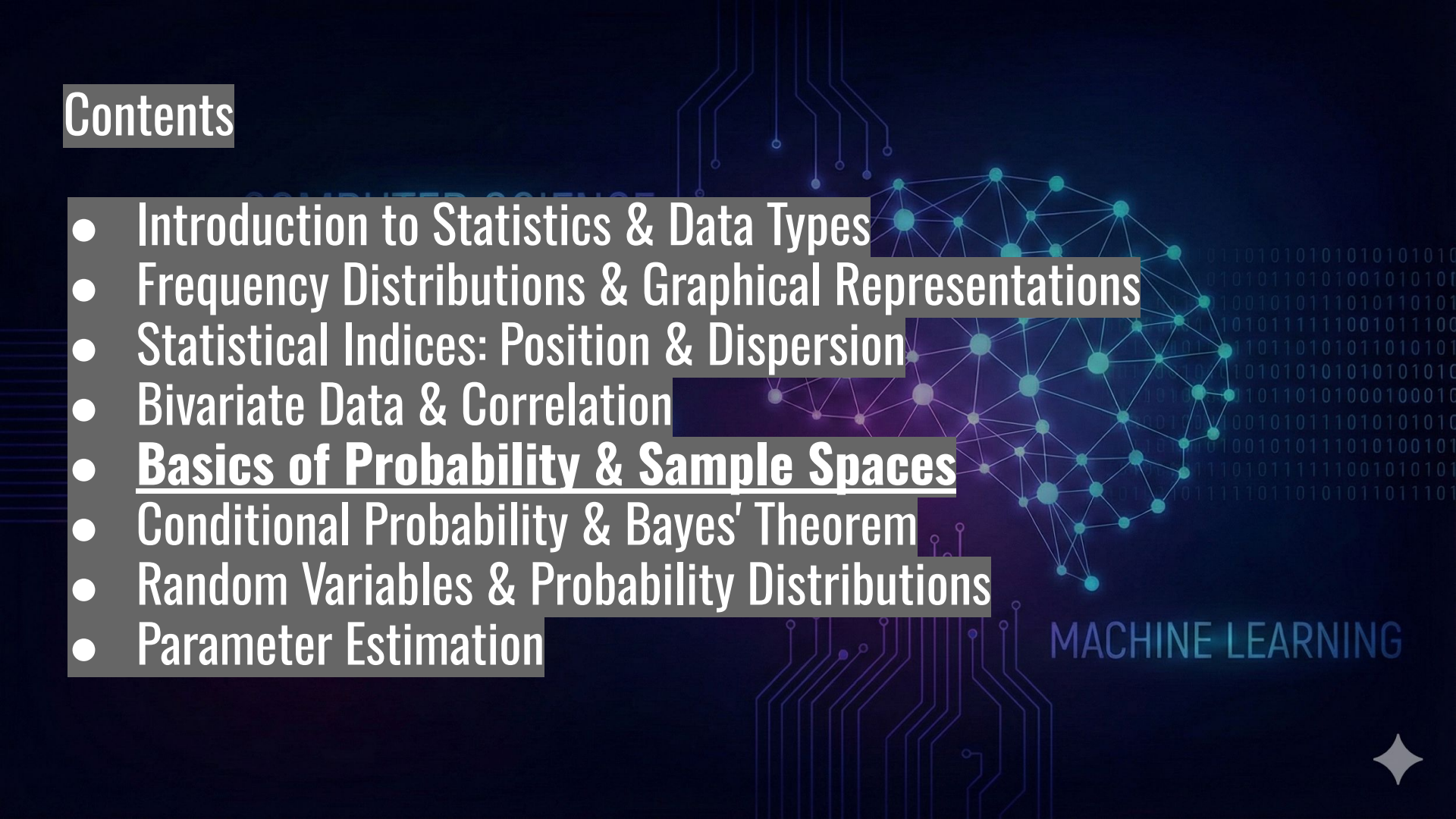
Final Result: The Correlation Coefficient is +1. This indicates a perfect positive linear correlation. Every extra hour of study adds exactly the same amount of points to the score.

$$r = \frac{s_{XY}}{s_X \cdot s_Y}$$

$$r = \frac{40}{2.828 \times 14.142}$$

$$r = \frac{40}{39.99} \approx \mathbf{1.0}$$

Contents



Section 05: Basics of Probability

The Conceptual Pivot: From Describing Data to Predicting Uncertainty

The Sample Space (Ω)

The set of all possible outcomes of a random experiment. It represents the **Certain Event**.

Example: A Die Roll

$\Omega = \{1, 2, 3, 4, 5, 6\}$

Events (A)

A subset of the sample space. An event occurs if the outcome belongs to this subset.

Example: Odd Result

$A = \{1, 3, 5\} \subseteq \Omega$

Kolmogorov Axioms

1. **Non-negativity:** $P(A) \geq 0$
2. **Normalization:** $P(\Omega) = 1$
3. **Additivity:** For disjoint events, $P(A \cup B) = P(A) + P(B)$

TRANSITION: Moving from *what has happened* (Descriptive) to the *mathematical model of what might happen* (Probability).

The Probability Space

Sample Space (Ω): The set of all possible outcomes of a random experiment. It represents the "Certain Event".

Example (Die Roll): Ω = {1, 2, 3, 4, 5, 6}

Event (A): A subset of Ω. An event occurs if the result of the experiment belongs to this subset.

Example (Odd Number): A = {1, 3, 5}

Operations on Events

We combine simple events to form complex ones using Set Theory logic.

Union ($A \cup B$)

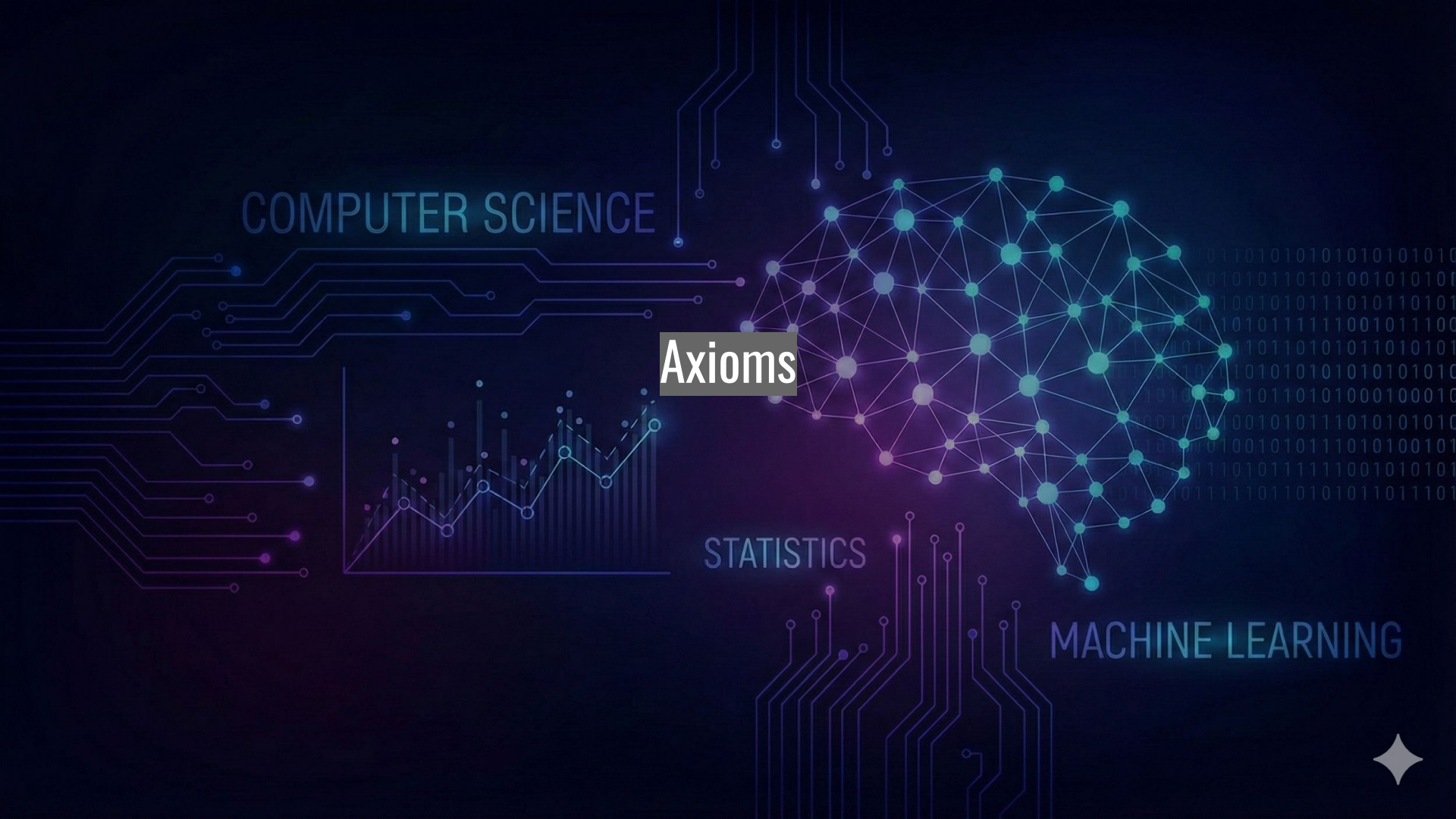
The event that happens if A OR B (or both) occur.

Intersection ($A \cap B$)

The event that happens if A AND B occur simultaneously.

Complement (A^c)

The event that happens if A does NOT occur.


COMPUTER SCIENCE
Axioms
STATISTICS
MACHINE LEARNING

The Axioms

Definition: Probability is a function satisfying specific rules.

1. Normalization
2. Additivity (for disjoint events):

Event space

$$\mathbb{P} : \mathcal{F} \rightarrow [0, 1]$$

$$\mathbb{P}(\Omega) = 1$$

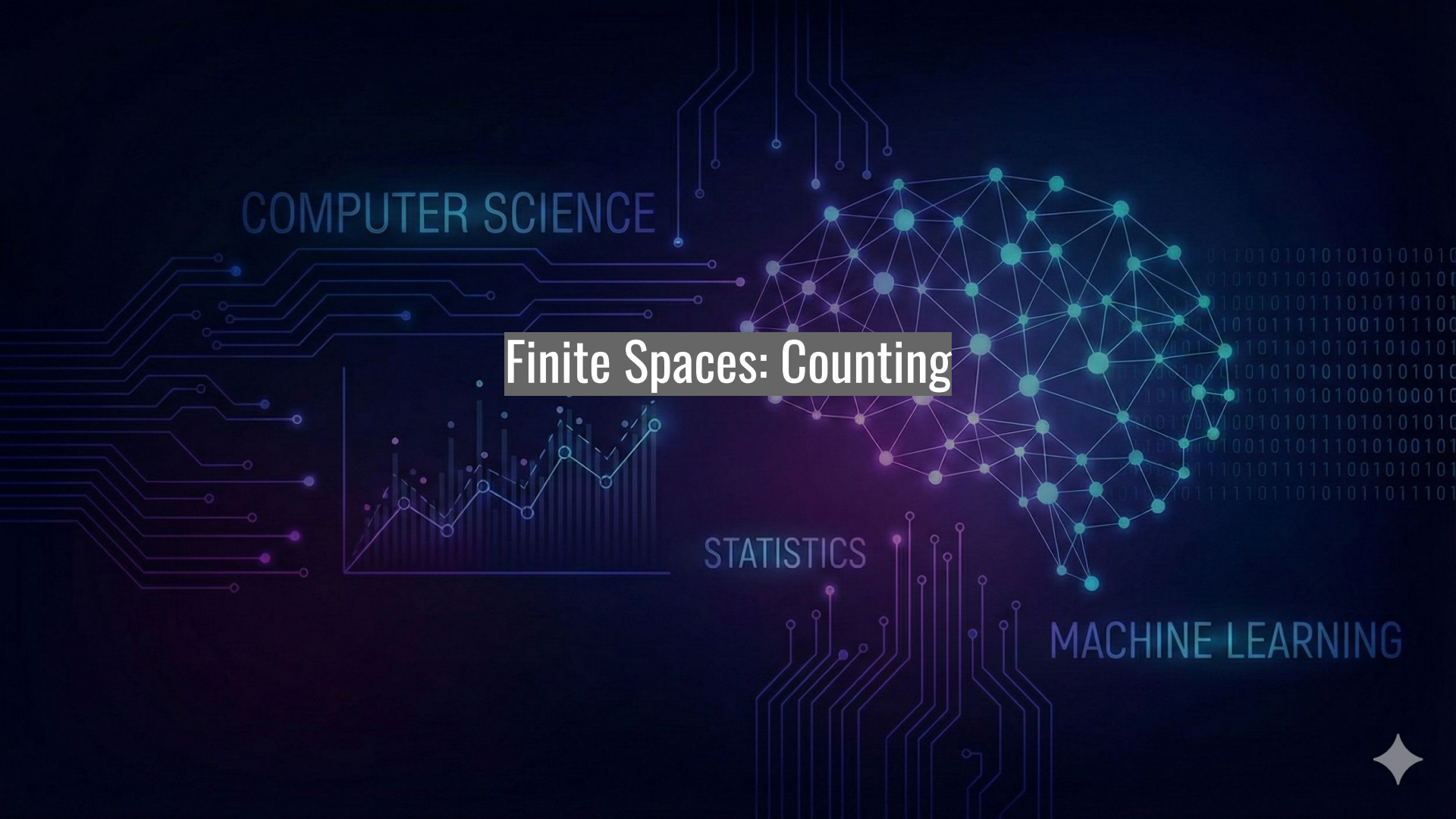
$$If A \cap B = \emptyset, then : \mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$$

The Axioms

Key Consequences From these axioms, we derive useful formulas:

$$\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$$

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$$


COMPUTER SCIENCE
Finite Spaces: Counting
STATISTICS
MACHINE LEARNING

Finite Spaces: Counting

Uniform Spaces (Equiprobable) If every outcome has the same likelihood (e.g., fair coin, fair dice), calculation is simple counting.

$$\mathbb{P}(A) = \frac{\text{Number of favorable cases}}{\text{Total number of cases}} = \frac{r}{n}$$

Finite Spaces: Counting

Non-Uniform Spaces If outcomes have different weight, we sum the individual probabilities of the elementary events in A. (picking colored marbles from a bag with unequal counts, e.g., more red than blue),

$$\mathbb{P}(A) = p_1 + p_2 + \dots$$

Finite Spaces: Counting

THE SAMPLE SPACE (Ω)

Red: 7, Blue: 3, Total: 10

Uniform vs Non-Uniform

In a **Uniform** space (like rolling a fair die), every outcome has the exact same weight ($\frac{1}{6}$).

In a **Non-Uniform** space, the elementary outcomes have *different* weights. If we randomly pick one marble from the bag, getting "Red" is much more likely than getting "Blue".

The Probability Rule

To find the probability of an Event A, we cannot just count the outcomes. We must **sum the individual weights (probabilities)** of every elementary event inside A.

$$P(A) = \sum_{\omega \in A} P(\omega)$$

The Scenario: The "Rigged" Race

Non-Uniform Spaces Imagine a race with 4 runners. They are not all equally fast.

- Runner A: The professional champion.
- Runner B: A decent amateur.
- Runner C: A complete beginner.
- Runner D: Has an injury.

The Scenario: The "Rigged" Race

The Sample Space (Ω)

- The possible winners are still finite: Ω = {A, B, C, D}

The Probability Assignment (Non-Uniform) Based on their skills, we assign probabilities (p_i) to each runner winning. These weights must sum to 1.

- P(A) = 0.50 (50% chance)
- P(B) = 0.30 (30% chance)
- P(C) = 0.15 (15% chance)
- P(D) = 0.05 (5% chance)

The Scenario: The "Rigged" Race

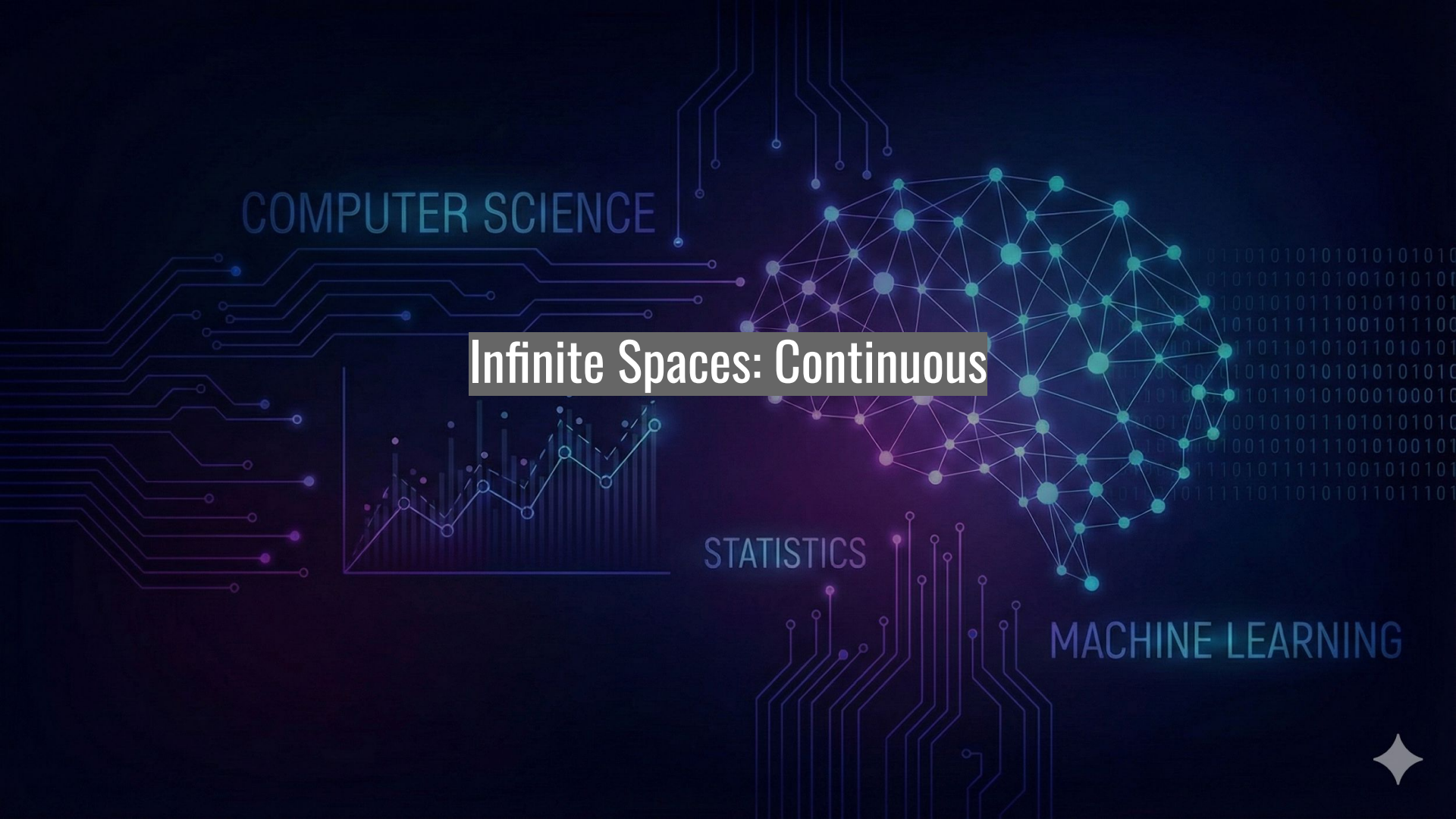
Calculating an Event:

Let's define Event E: "An amateur wins" (meaning B or C wins).

Since the outcomes are disjoint (A and B can't both win), we just sum the individual probabilities.

$$P(E) = P(B) + P(C)$$

$$P(E) = 0.30 + 0.15 = 0.45$$


COMPUTER SCIENCE
Infinite Spaces: Continuous
STATISTICS
MACHINE LEARNING

Continuous Probability: The Density Function

THE PROBLEM

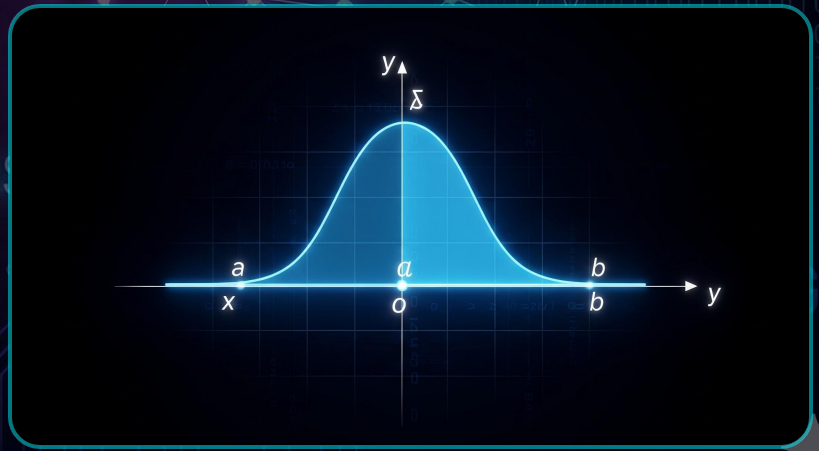
In continuous spaces (Time, Distance), the probability of hitting an exact single point is mathematically ZERO.

THE SOLUTION

We define probability as the Area under a curve f(x) over a specific interval [a, b].

The Density Function (f) represents the relative likelihood.

$$\mathbb{P}(a < X < b) = \int_a^b f(x)dx$$



The Two Conditions of a Valid PDF

NON-NEGATIVITY

A density function can never be negative. Likelihood must be zero or greater at every point.

$$f(x) \geq 0$$

For all real numbers x

NORMALIZATION

The total area under the entire curve must equal exactly 1 (100% total probability).

$$\int_{-\infty}^{\infty} f(x)\,dx = 1$$

Total integral over space

Continuous Probability: Infinite Spaces

THE SCENARIO

Imagine a bus arrives at a stop perfectly randomly anytime between 8:00 and 8:10.

TOTAL INTERVAL
10 Minutes

VARIABLE (X)
Arrival Time

X is the number of minutes past 8:00.

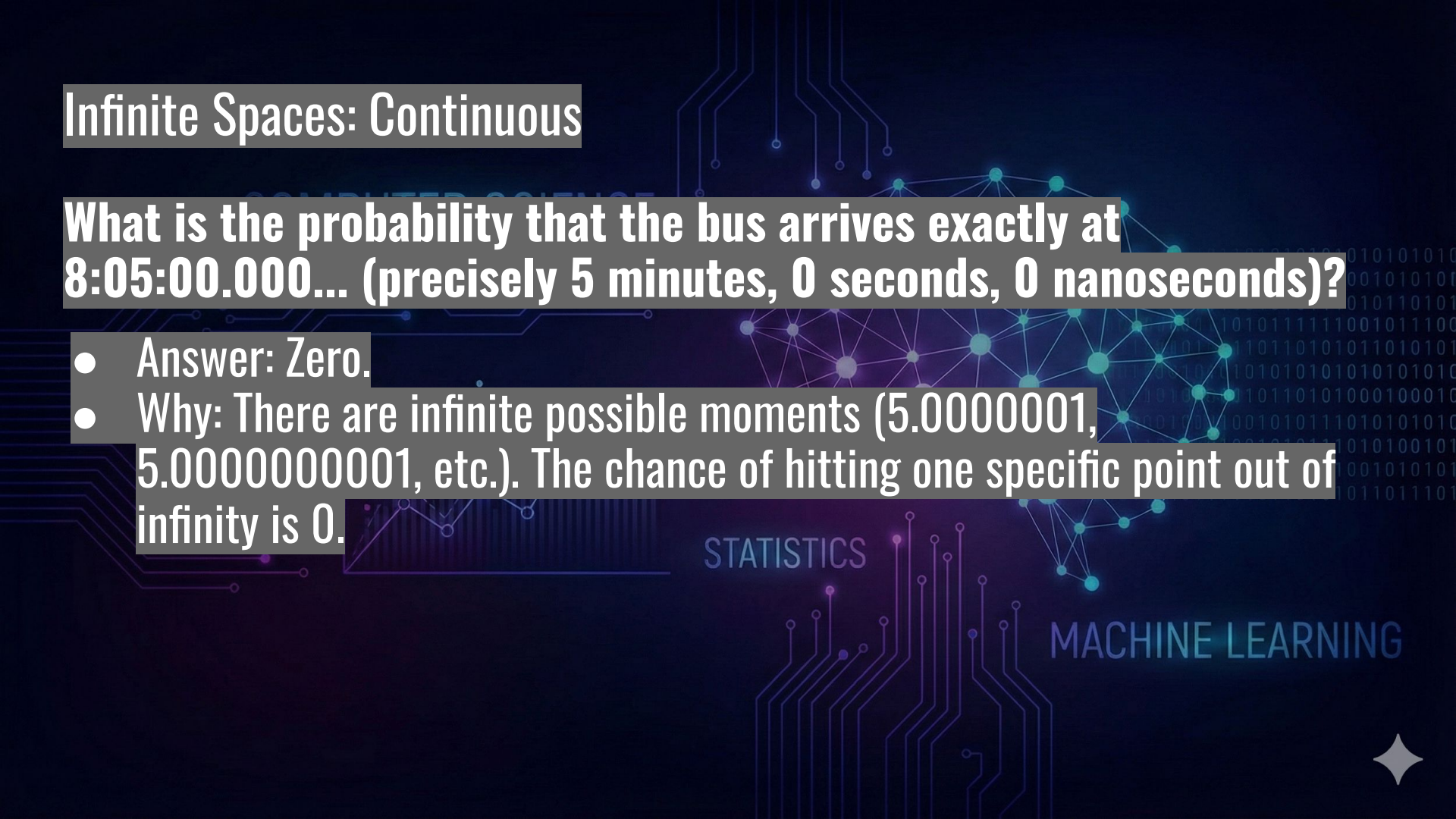
CONTINUOUS SAMPLE SPACE

8:00 — 8:10

Infinite Spaces: Continuous

What is the probability that the bus arrives exactly at 8:05:00.000... (precisely 5 minutes, 0 seconds, 0 nanoseconds)?

- Answer: Zero.
- Why: There are infinite possible moments (5.0000001, 5.0000000001, etc.). The chance of hitting one specific point out of infinity is 0.



Infinite Spaces: Continuous

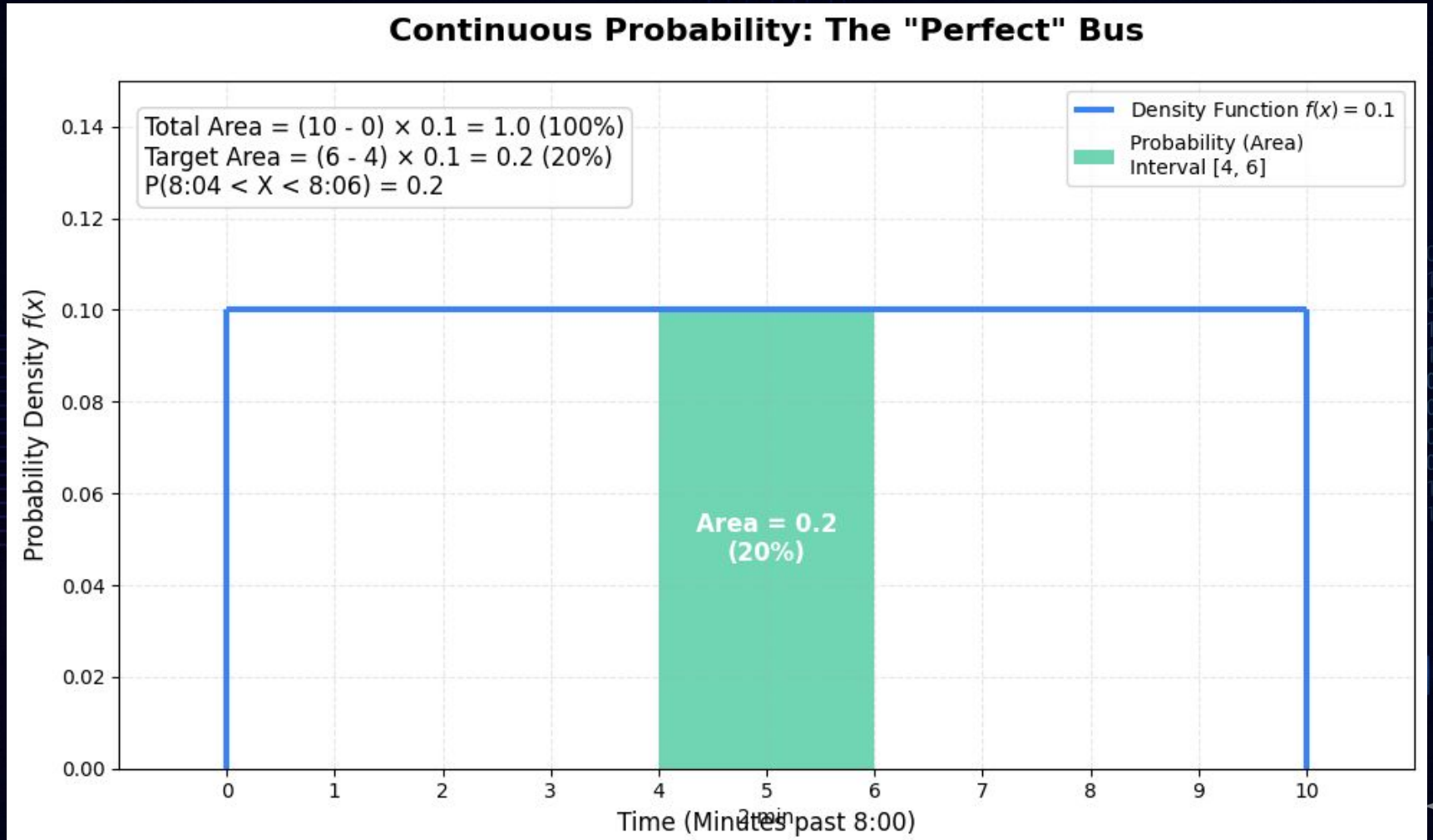
Instead, we ask: "What is the probability the bus arrives between 8:04 and 8:06?

- Calculation: The interval (4 to 6) is 2 minutes long.
- Total Time: 10 minutes.
- Probability: 2 / 10 = 0.2 (or 20%).

Infinite Spaces: Continuous

The Density Function (f(x)): Since the bus is equally likely to arrive at any time, the Density Function is a flat horizontal line (a rectangle).

- Height: 1/10 = 0.1 (This ensures total Area = 10 x 0.1 = 1).
- Area Calculation: To find the probability between minute 4 and minute 6:
  - Base = 6 - 4 = 2
  - Height = 0.1
  - Area = 2 x 0.1 = 0.2 (20%)


Continuous Probability: The "Perfect" Bus
Total Area = (10 - 0) × 0.1 = 1.0 (100%)
Target Area = (6 - 4) × 0.1 = 0.2 (20%)
P(8:04 < X < 8:06) = 0.2
Density Function f(x) = 0.1
Probability (Area)
Interval [4, 6]
Area = 0.2
(20%)
Probability Density f(x)
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
0
1
2
3
4
5
6
7
8
9
10
2 min
Time (Minutes past 8:00)

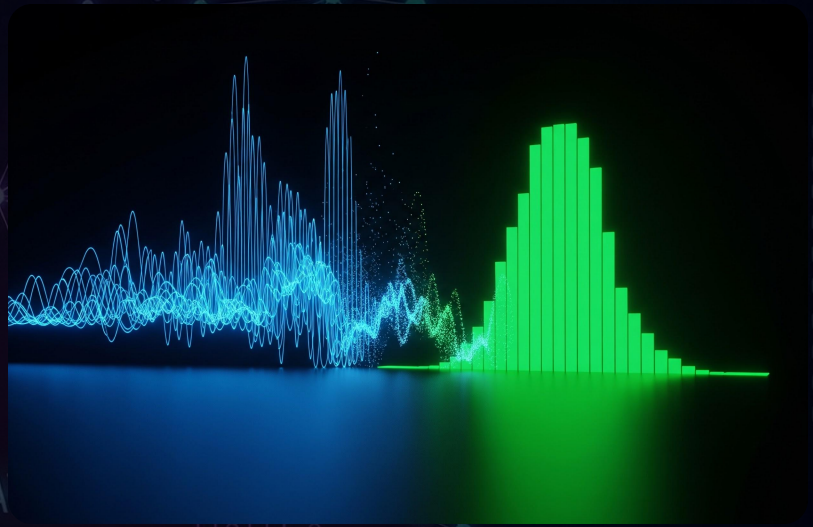
QM: THE BORN INTERPRETATION

HISTORICAL CONTEXT

This is the central concept connecting probability and quantum mechanics. Stated by Max Born in 1926, it provides the physical meaning for the wave function.

THE MEANING

The complex-valued wave function, $\Psi(x, t)$, is squared to derive the Probability Density Function (PDF) for finding a particle at a specific point.



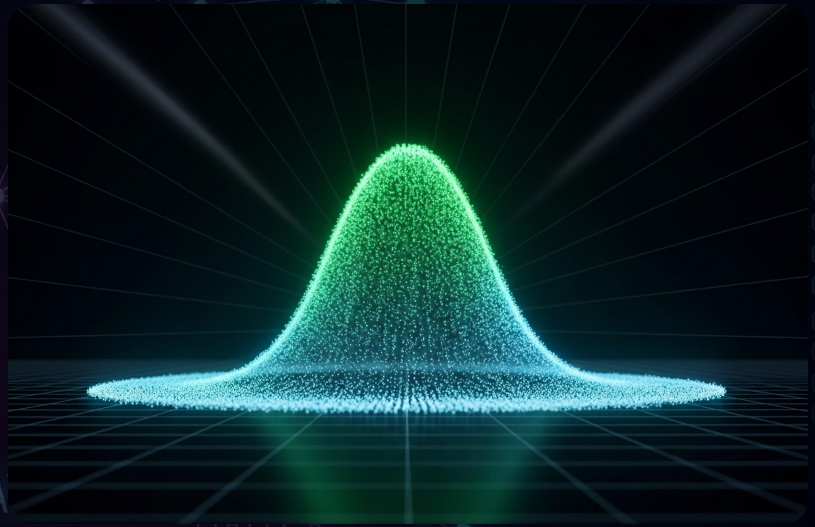
$$PDF(x, t) = |\Psi(x, t)|^2$$

THE KEY EQUATION

THE QUANTUM WAVE FUNCTION (Ψ)

FINDING THE PARTICLE

The distribution on the right is the probability density cloud for a particle in a quantum state.



QM PROBABILITY

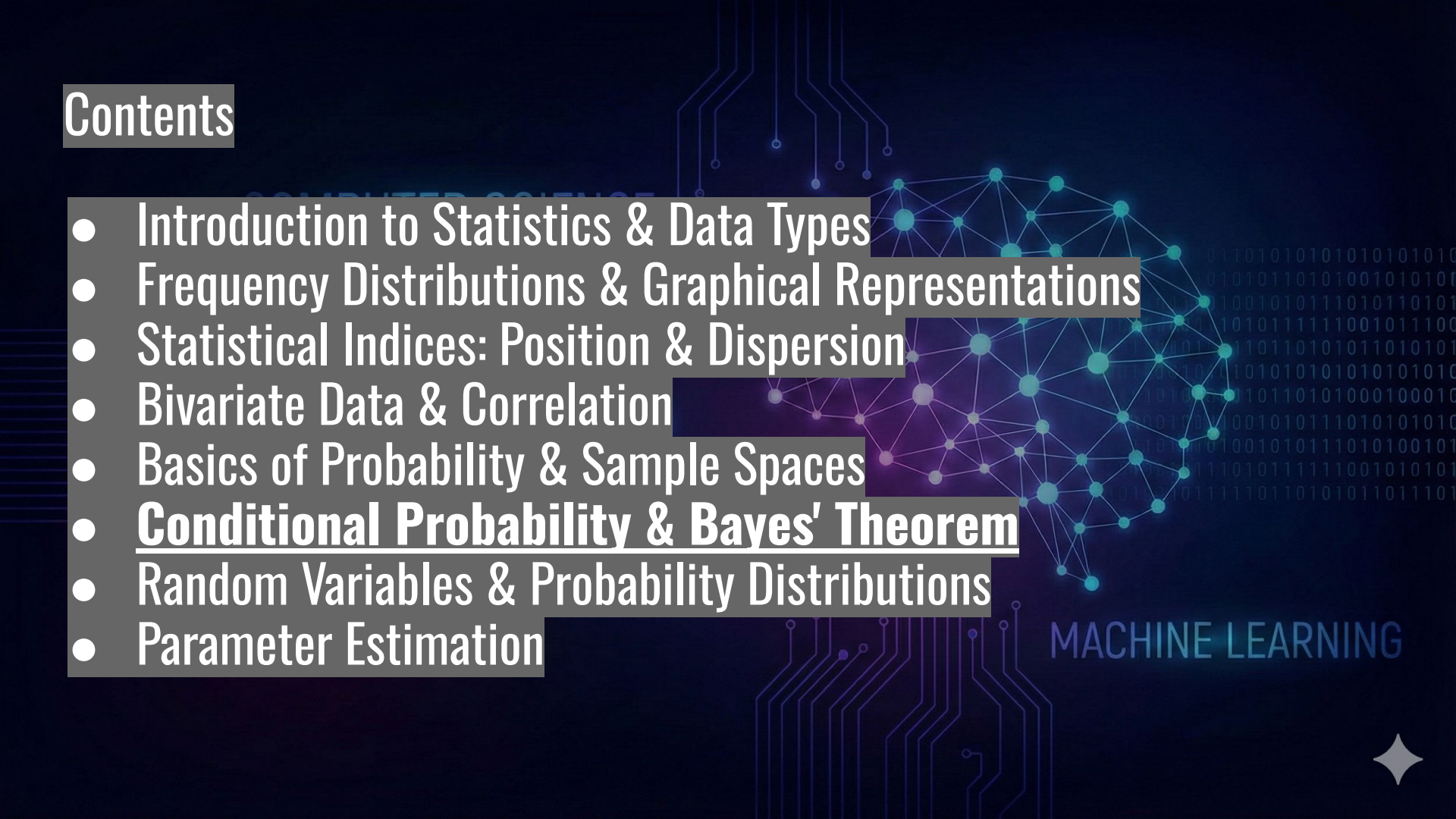
The probability of finding the particle between position a and b at time t is calculated by integrating the squared magnitude:

$$P(a, b) = \int_a^b |\Psi(x, t)|^2 \, dx$$

$$\int_{-\infty}^{+\infty} |\Psi(x, t)|^2 \, dx = 1$$

NORMALIZATION UNIVERSAL RULE

Contents



Advanced Probability: The Logic of Inference

THE PARADIGM SHIFT

Moving beyond static sample spaces into dynamic inference. We no longer just count outcomes; we model how new information reshapes our entire mathematical reality.

UPDATING BELIEFS

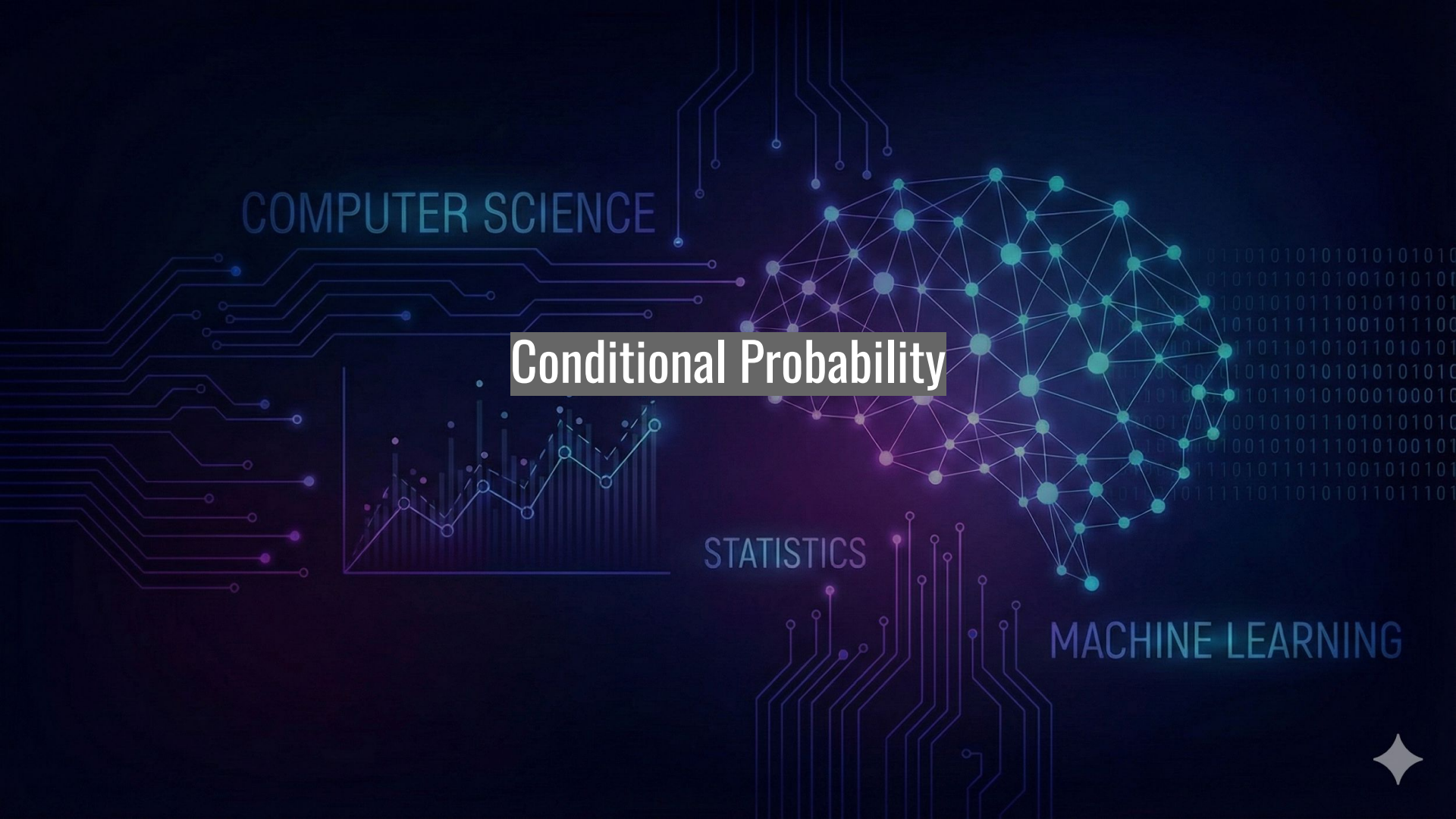
Probability is not fixed. As evidence (E) arrives, the prior likelihood of a hypothesis (H) is mathematically updated to a posterior probability.

THE CULMINATION: BAYES' THEOREM

Prior Knowledge
Initial Beliefs

$$P(H|E) = \frac{[P(E|H) * P(H)]}{P(E)}$$

Posterior Logic
Updated Reality


COMPUTER SCIENCE
Conditional Probability
STATISTICS
MACHINE LEARNING

Conditional Probability: The Impact of Information

CORE CONCEPT

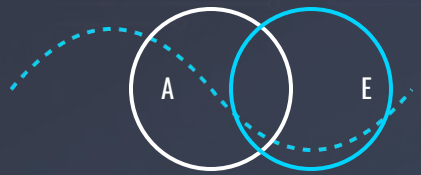
When an event E occurs, the sample space Ω shrinks to E. We calculate the likelihood of A within this new constrained reality.

Information updates our belief: "Given E, what is the chance of A?"

MATHEMATICAL DEFINITION

$$\mathbb{P}(A|E) = \frac{\mathbb{P}(A \cap E)}{\mathbb{P}(E)}$$

Mandatory Condition: **P(E) > 0**



The sample space Ω is replaced by E.

By knowing E happened, outcomes outside of E are now impossible. The probability of A is now normalized by the size of E.

Example: Rolling a Die

Scenario:

- You roll a fair 6-sided die.Sample Space Ω: {1, 2, 3, 4, 5, 6} (6 possible outcomes)

Let's define two events:

- Event A: Rolling a 4. (A = {4})
- Event E: Rolling an Even number. (E = {2, 4, 6})

Example: Rolling a Die

What is the probability of rolling a 4 (Event A), given that we know the result is Even (Event E)? We write this as P(A|E)

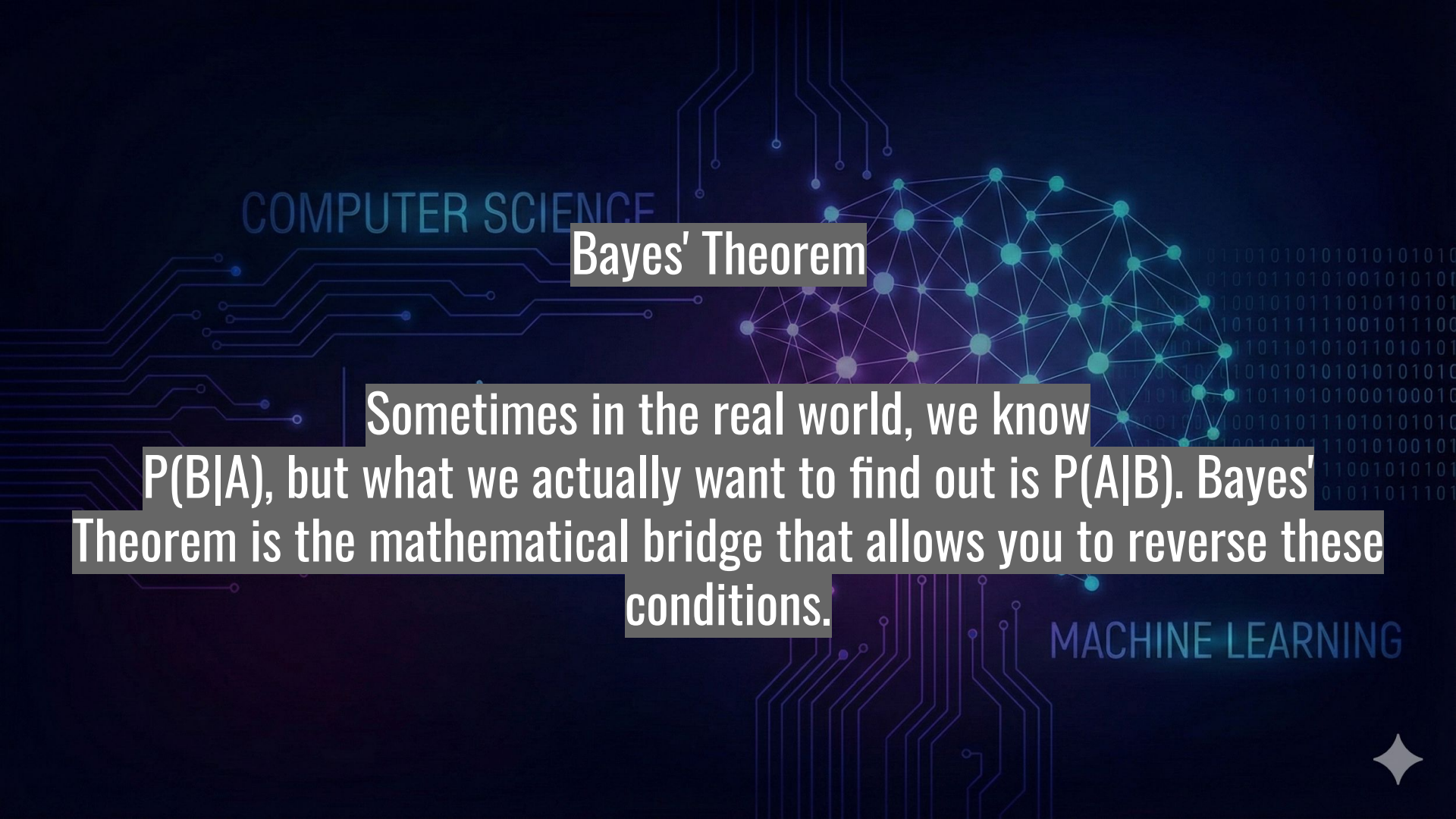
1. Find $P(A \cap E)$: What is the probability of rolling a number that is both "4" AND "Even"?
   a. The only number is 4. So, 1 outcome out of the original 6.
   b. $P(A \cap E) = \frac{1}{6}$
2. Find P(E):
   a. What is the probability of rolling an "Even" number? Outcomes are {2, 4, 6}. So, 3 outcomes out of the original 6.
   b. $P(E) = 3/6 = 1/2$.

Example: Rolling a Die

Apply Formula:

$$\mathbb{P}(A|E) = \frac{\mathbb{P}(A \cap E)}{\mathbb{P}(E)} = \frac{1/6}{3/6} = \frac{1}{3}$$

Result: 1/3. The formula matches our intuition perfectly!



Bayes' Theorem

Sometimes in the real world, we know P(B|A), but what we actually want to find out is P(A|B). Bayes' Theorem is the mathematical bridge that allows you to reverse these conditions.

The Equation

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

$$P(A \cap B) = P(A|B)P(B)$$

Because the labels A and B are arbitrary, we can also write it from the perspective of event A happening first

$$P(A \cap B) = P(B|A)P(A)$$

Updating our Beliefs

The theorem calculates the probability of Hypothesis **A** occurring, given that Evidence **B** has been observed.

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

It provides a rigorous way to reverse conditional probabilities. It allows us to calculate P(A|B) if we already know P(B|A).

Anatomy of the Formula

Let's break down the variables on the right side of the equation that allow us to calculate the final Posterior probability.

P(A) - The Prior

The initial, baseline probability of the hypothesis **A** being true, *before* any new evidence is collected or observed.

P(B|A) - The Likelihood

The probability of observing this specific evidence **B**, assuming that our hypothesis **A** is actually true.

P(B) - The Marginal

The total probability of observing the evidence **B** under *all* possible conditions, regardless of whether hypothesis A is true or false.

This acts as a normalizing constant. It ensures that the final calculated probability remains a valid percentage between 0 and 100%.

Real-World Impact

Medical Testing

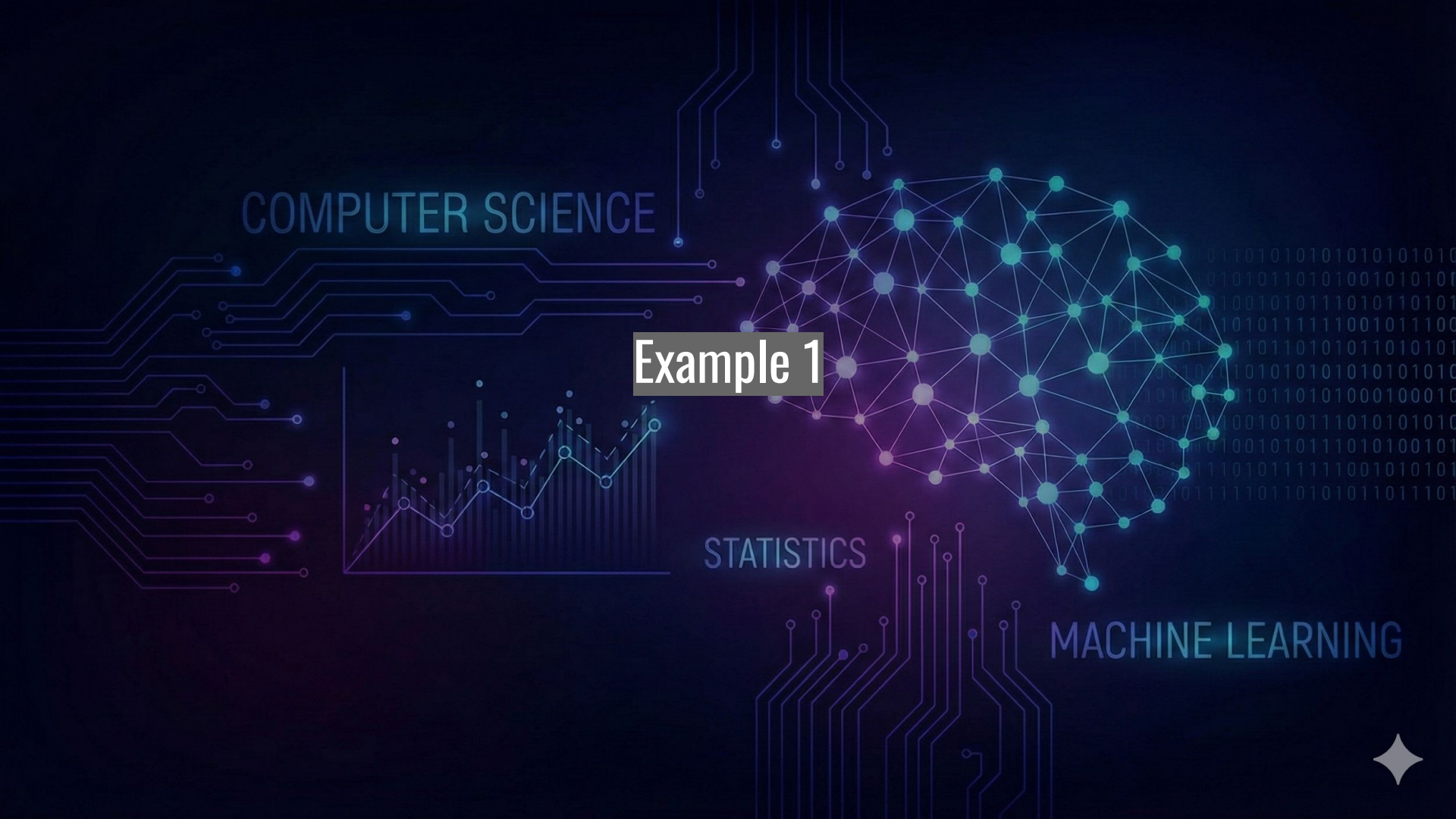
If a rare disease test comes back positive, Bayes' Theorem is used to calculate the actual probability of the patient being sick, factoring in the false-positive rate of the test.

Spam Filters

Machine learning models (Naive Bayes Classifiers) use this theorem to calculate the probability that an email is spam, given the specific combination of words it contains.

Risk Assessment

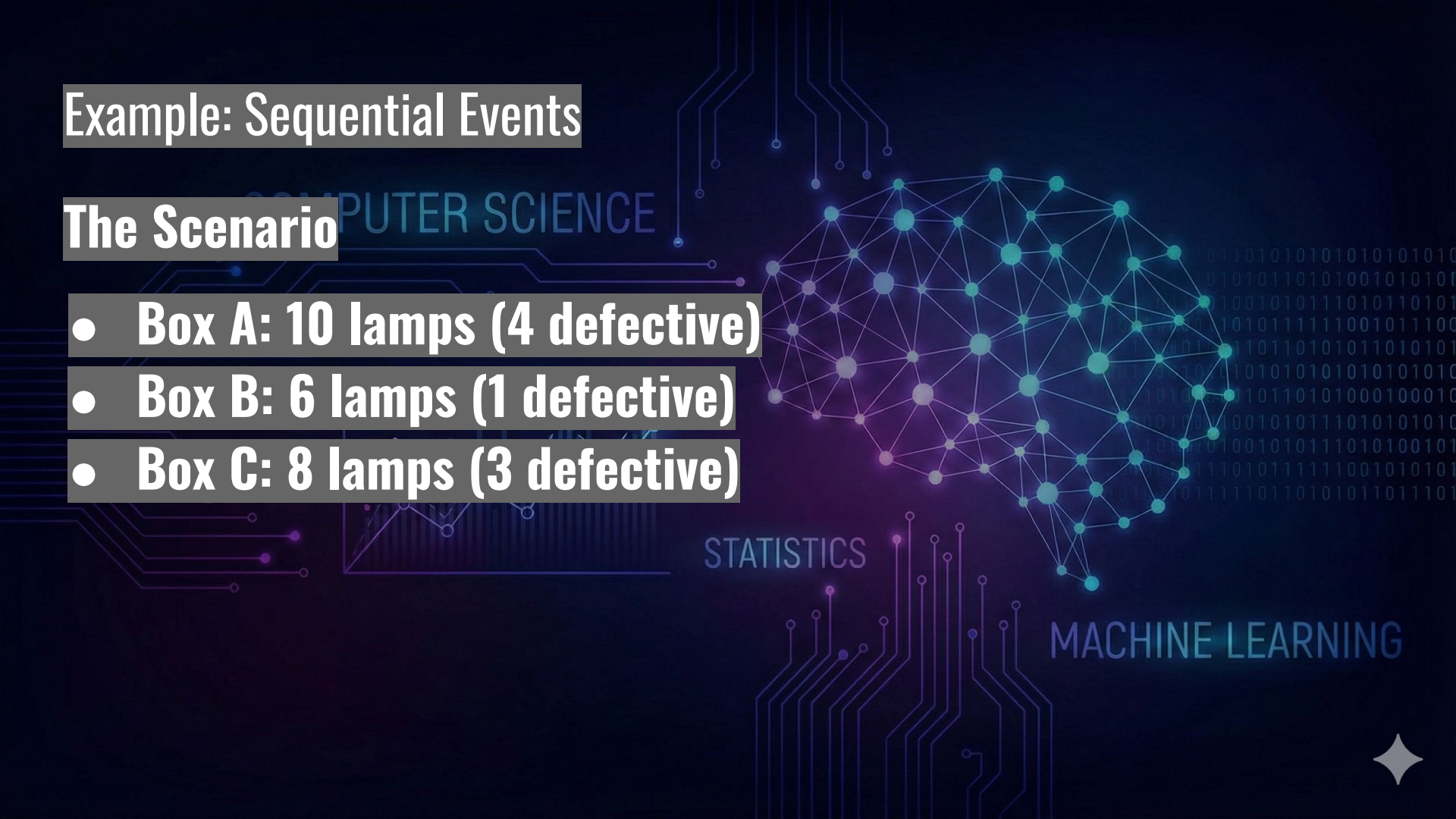
Insurance companies and quantitative finance firms use Bayesian inference to continuously update risk models and asset prices as new market data arrives in real-time.


COMPUTER SCIENCE
Example 1
STATISTICS
MACHINE LEARNING

Example: Sequential Events

The Scenario

- **Box A: 10 lamps (4 defective)**
- **Box B: 6 lamps (1 defective)**
- **Box C: 8 lamps (3 defective)**



Example: Sequenti

The Scenario

- Box A: 10 lam
- Box B: 6 lam
- Box C: 8 lam

Bayes' Theorem

Reversing the Condition It allows us to calculate the probability of a Cause given an Effect.

- "Given that the lamp is defective (D), what is the probability it came from Box A?"

$$\mathbb{P}(A|D) = \frac{\mathbb{P}(D|A)\mathbb{P}(A)}{\mathbb{P}(D)}$$

Calculate the Numerator (The Product Rule)

The top part of the fraction: P(D|A) P(A)

1. P(A): The probability of picking Box A at random is 1/3
2. P(D|A): The probability of picking a defective lamp, given you are looking in Box A, is 4/10 (since there are 10 lamps and 4 are defective).

$$P(D|A)P(A) = \frac{4}{10} \cdot \frac{1}{3} = \frac{4}{30}$$

Calculate the Denominator (Total Probability)

$$\mathbb{P}(D) = \mathbb{P}(D|A)\mathbb{P}(A) + \mathbb{P}(D|B)\mathbb{P}(B) + \mathbb{P}(D|C)\mathbb{P}(C)$$

$$= \frac{4}{10} \cdot \frac{1}{3} + \frac{1}{6} \cdot \frac{1}{3} + \frac{3}{8} \cdot \frac{1}{3} = \frac{113}{360}$$

Total probability of picking a defective lamp from any of the boxes.

Using the chain rule

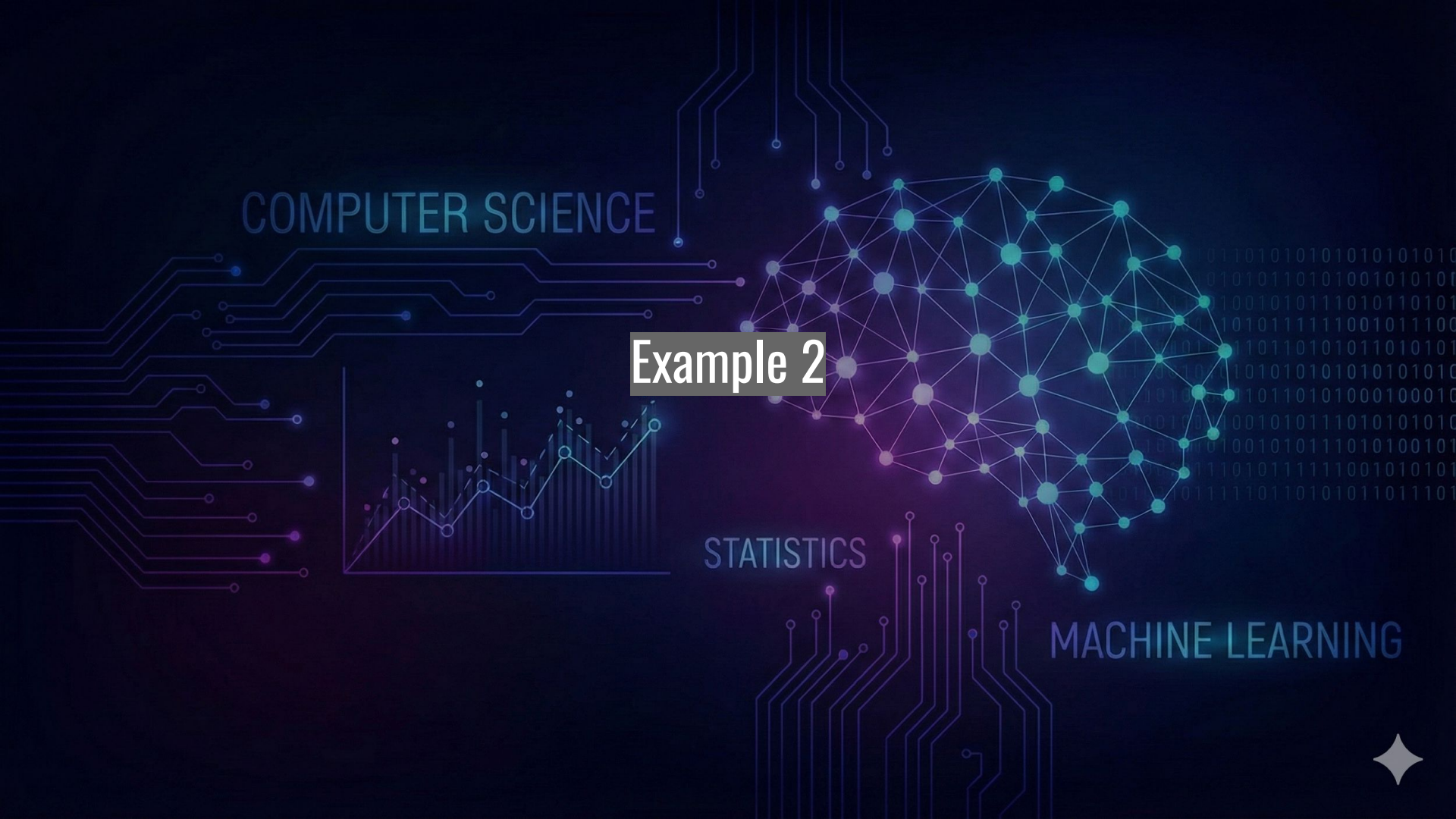
Put it all together in Bayes' Theorem

We found P(D|A) P(A) = 4/30 but 4/30 = 48/360 and P(D) = 113/360

$$P(A|D) = \frac{48/360}{113/360}$$

The estimated probability that the defective lamp came from Box A is 48/113, which is approximately 42.5%.

Even though you had an equal ⅓ = (33.3%) chance of picking Box A at the very start, knowing that the lamp you drew was defective increases the likelihood that you drew from Box A. This makes perfect logical sense because Box A has the highest concentration of defective lamps compared to the other boxes

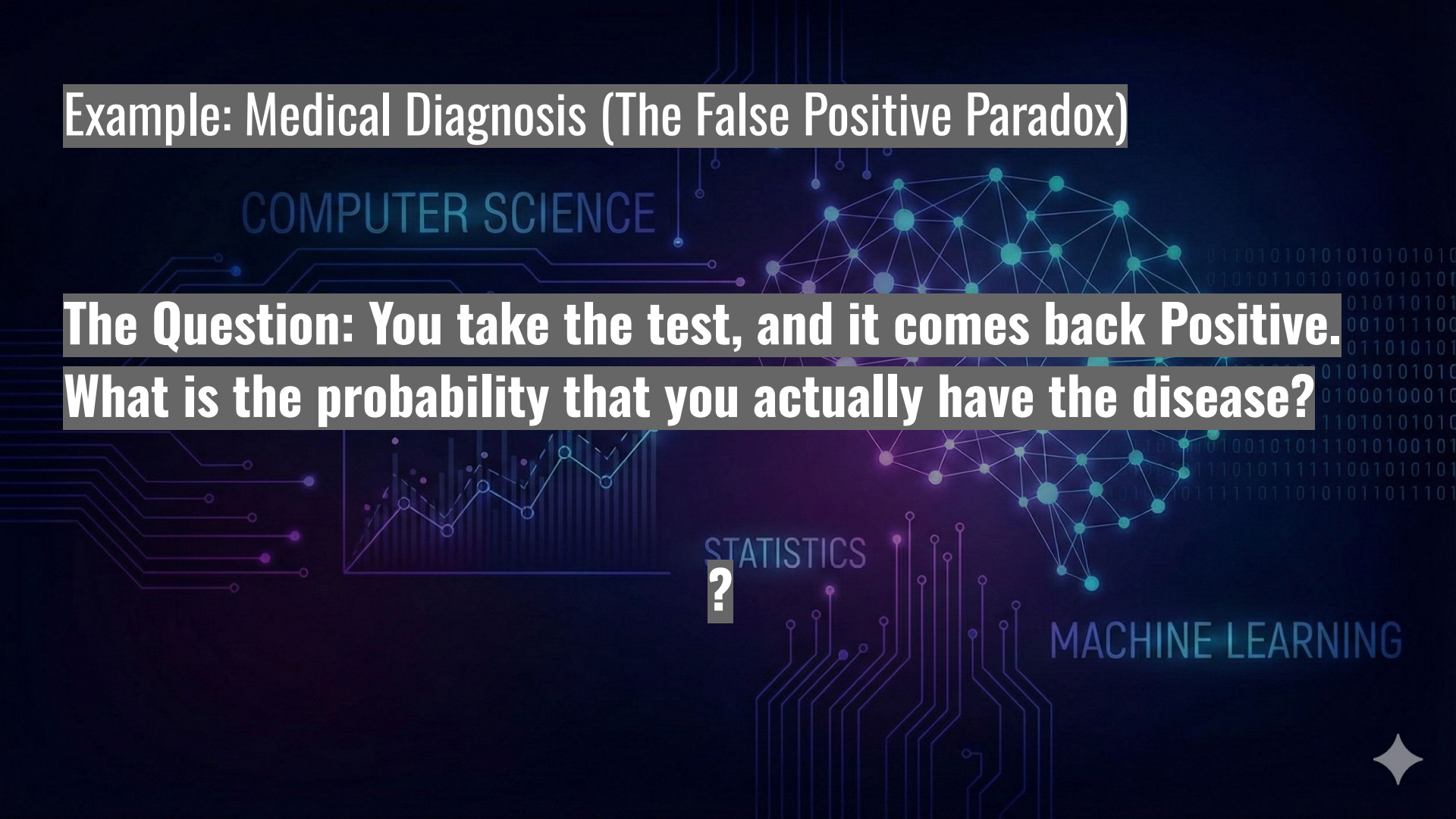

COMPUTER SCIENCE
Example 2
STATISTICS
MACHINE LEARNING

Example: Medical Diagnosis (The False Positive Paradox)

The Scenario: Imagine a rare disease affects 1% of the population. There is a test for this disease that is 99% accurate.

- **If you have the disease, the test is positive 99% of the time.**
- **If you don't have the disease, the test is negative 99% of the time (1% False Positive).**

Example: Medical Diagnosis (The False Positive Paradox)



The Question: You take the test, and it comes back Positive. What is the probability that you actually have the disease?

Example: Medical Diagnosis (The False Positive Paradox)

The Question: You take the test, and it comes back Positive. What is the probability that you actually have the disease?

Intuitive Guess: Most people guess 99%.

Example: Medical Diagnosis (The False Positive Paradox)

The Calculation (Using Bayes' Theorem):

1. Let D be "Has Disease"
2. Let + be "Positive Test".

We want to find P(D|+).

Example: Medical Diagnosis (The False Positive Paradox)

- **Prior Probability: P(D) = 0.01 (1% of people have it).**
- **Probability of Positive if Diseased: P(+|D) = 0.99.**
- **Probability of Positive if Healthy: P(+|H) = 0.01 (False Positive).**
- **Total Probability of Positive P(+):**
  - **$P(+) = P(+|D)P(D) + P(+|H)P(H)$**
  - **$P(+) = (0.99 \times 0.01) + (0.01 \times 0.99) = 0.0099 + 0.0099 = 0.0198$**

Example: Medical Diagnosis (The False Positive Paradox)

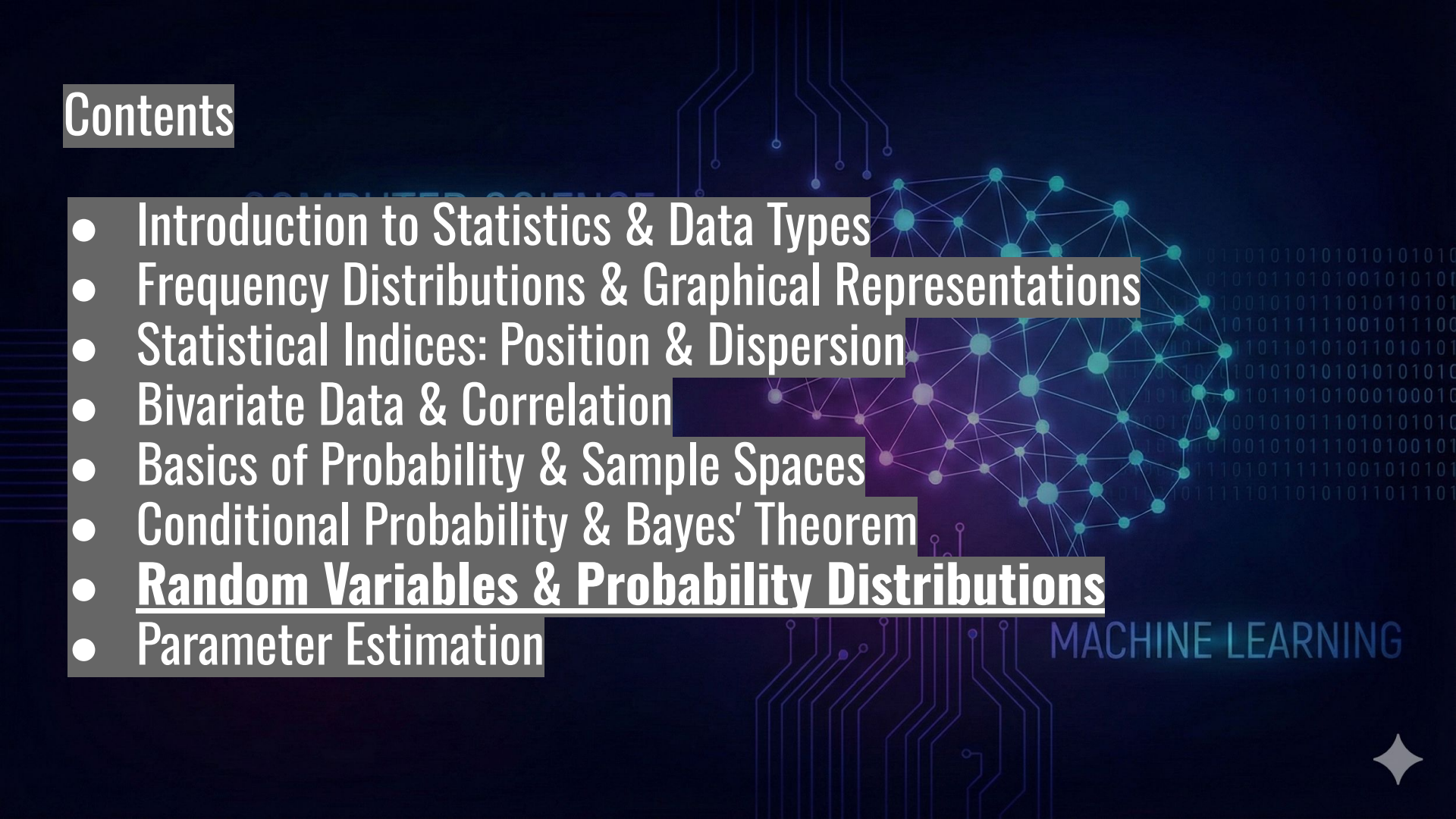
Bayes' Formula:

$$P(D|+) = \frac{P(+|D)P(D)}{P(+)} = \frac{0.0099}{0.0198} = 0.5$$

The Result:The probability is only 50%. Even with a positive test, it's a coin flip whether you are actually sick or just a false positive

Why? Because healthy people (99% of the population) vastly outnumber the sick people. Even a tiny 1% error rate on the massive healthy group produces as many false positives as there are true positives

Contents



Random Variables & Probability Distributions

Mathematical Mapping

A Random Variable (X) is a function that maps every outcome in the sample space (Ω) to a real number (ℝ).

Binomial Distribution

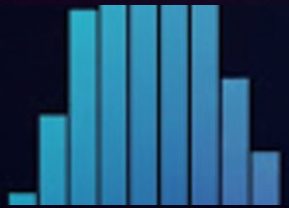
Models the total number of successes in 'n' independent Bernoulli trials with probability 'p'.

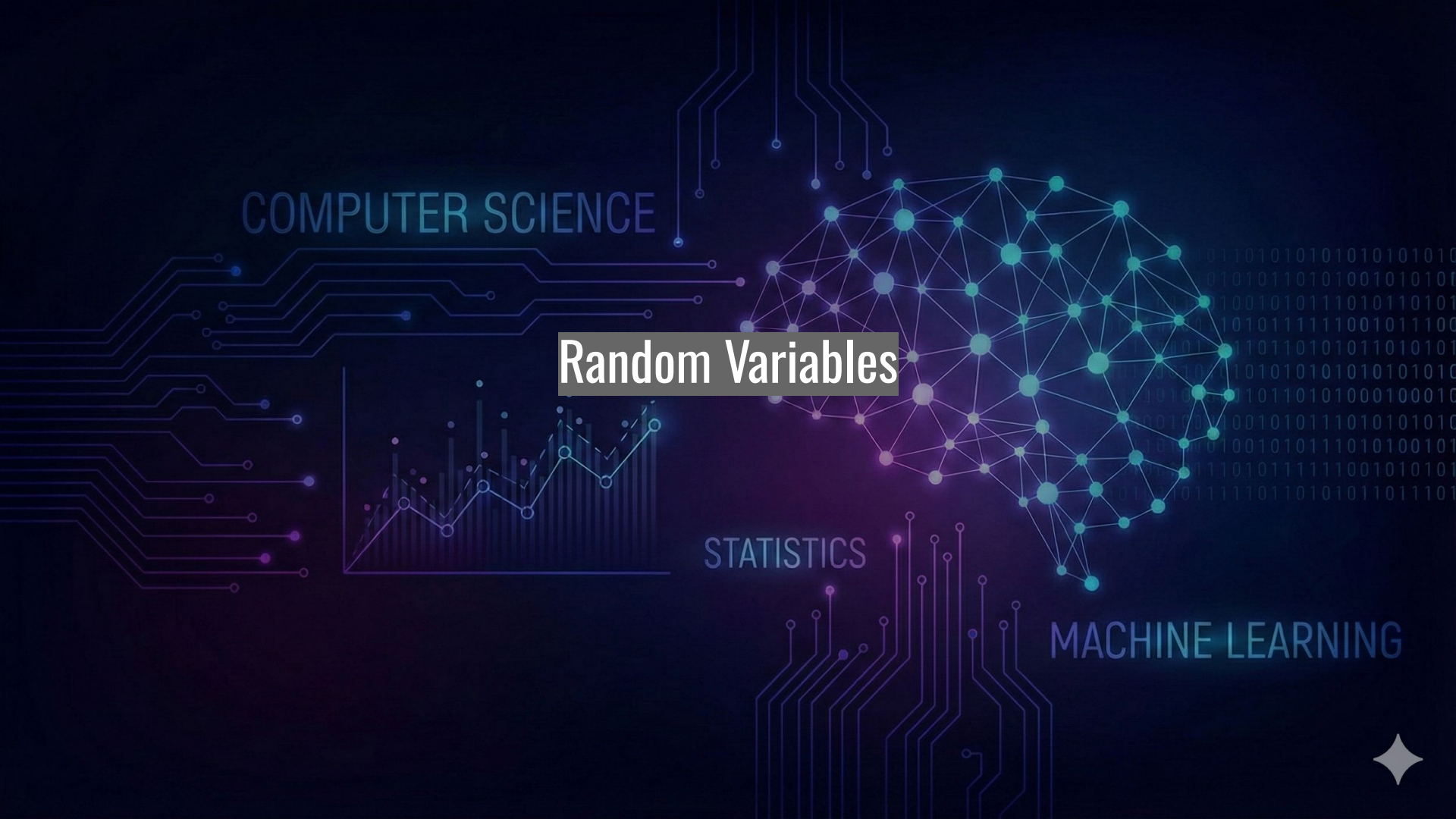
Poisson Distribution

Describes the number of independent events occurring in a fixed interval of time or space.

Normal Distribution

The classic "bell-shaped" curve defined by its mean (μ) and variance (σ^2).




COMPUTER SCIENCE
Random Variables
STATISTICS
MACHINE LEARNING

General Definition

What is a Random Variable? (X)

A function that associates a real number with every possible outcome of a random experiment (Ω).

$$X : \Omega \to E \subset \mathbb{R}$$

1 Discrete

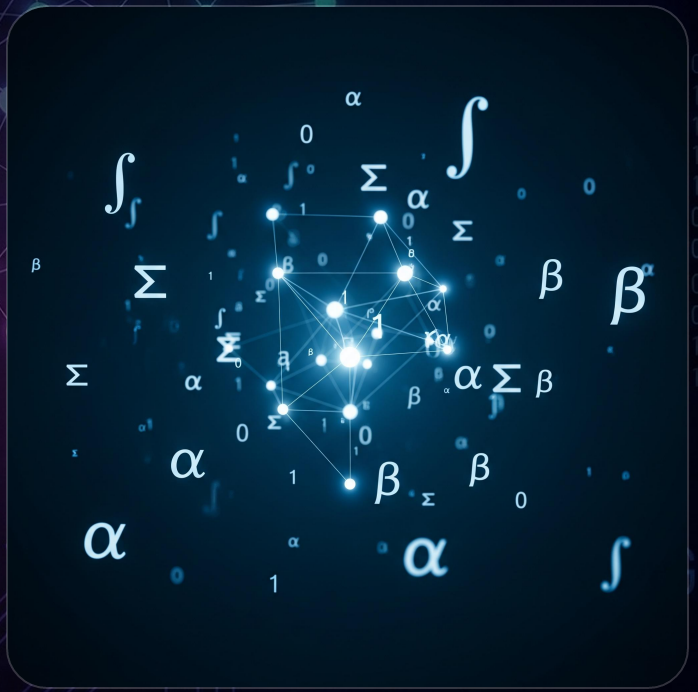
The set of values E is finite or countable.

e.g., Die rolls, Counting items.

Continuous

The set E is a continuous interval.

e.g., Time, Length, Temperature.

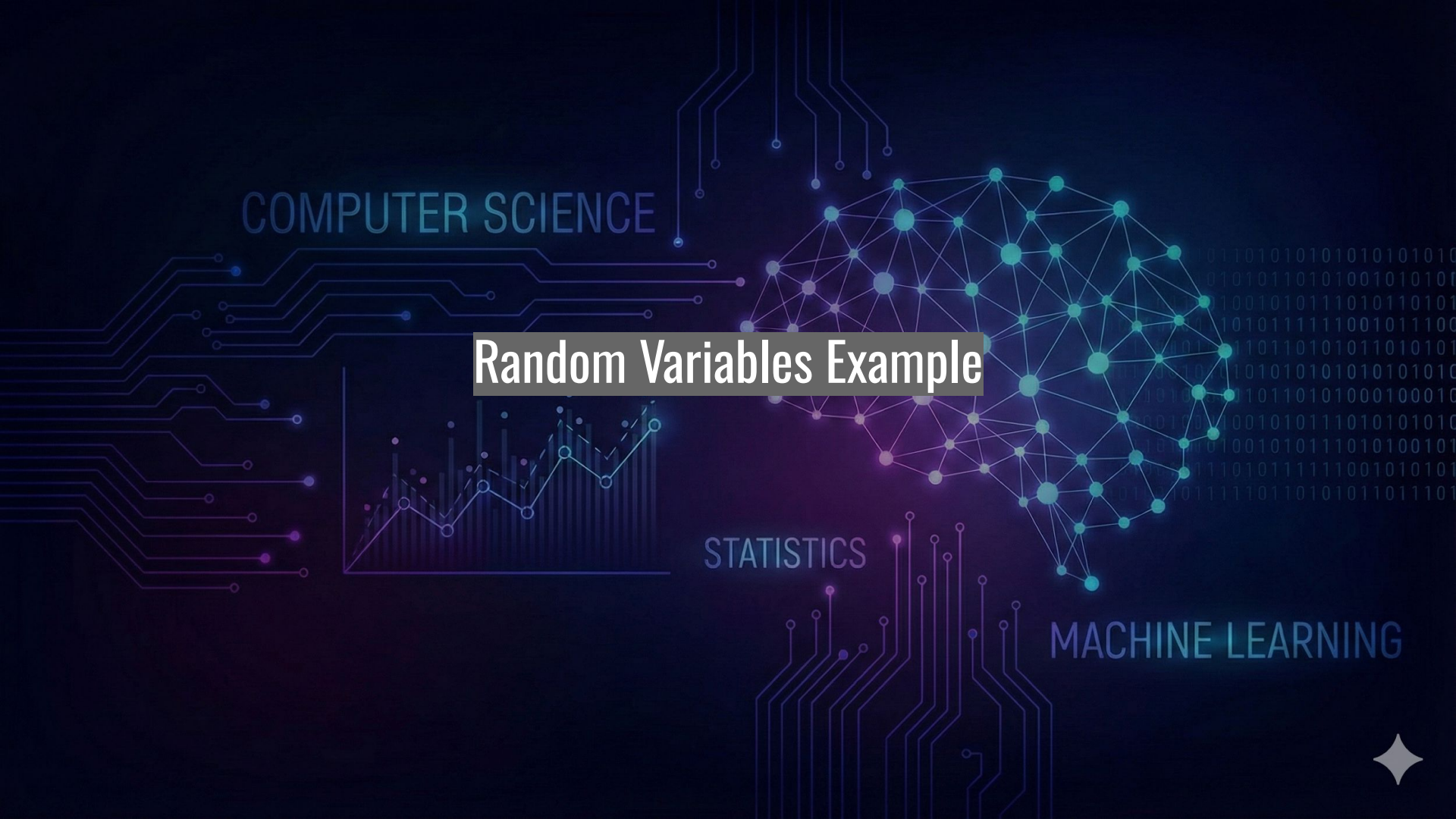


What is a Random Variable

The Random Variable translates real-world outcomes into numbers.

A random variable is actually a mapping rule or a translator

When you conduct an experiment (like tossing coins, rolling dice, or picking a lightbulb), the raw outcomes are often not numbers—they are physical events (e.g., "Heads, Tails," or "Defective, Working"). Mathematics cannot easily calculate with words. A random variable (X) is the rule you define to assign a specific numerical value to every possible physical outcome.



Random Variables Example

General Definition: Examples

Rolling a Single Die

Let X = 0 if even, X = 1 if odd. This maps the physical outcome to a numerical value.

Sample Space (Ω)

{1, 2, 3, 4, 5, 6}

Mapped Values (E)

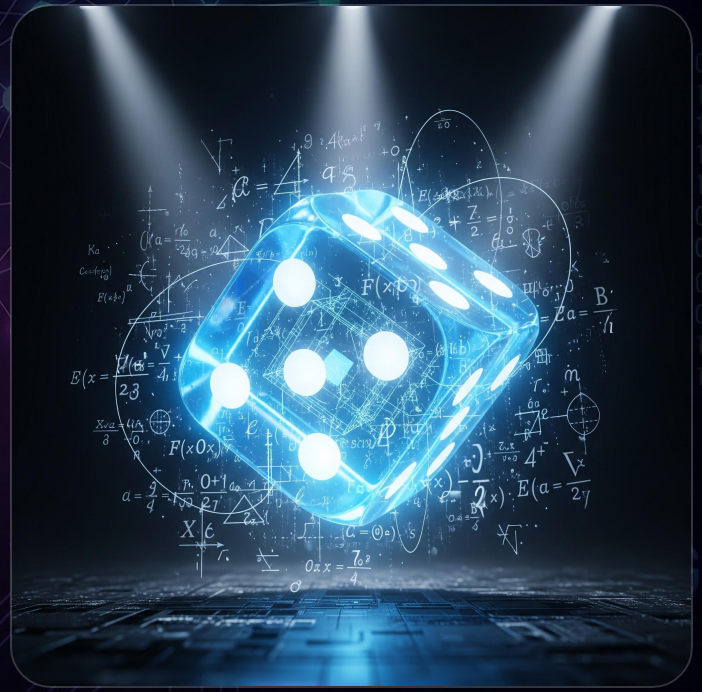
{0, 1}

Odd Outcomes (X=1)

X(1) = X(3) = X(5) = 1

Even Outcomes (X=0)

X(2) = X(4) = X(6) = 0

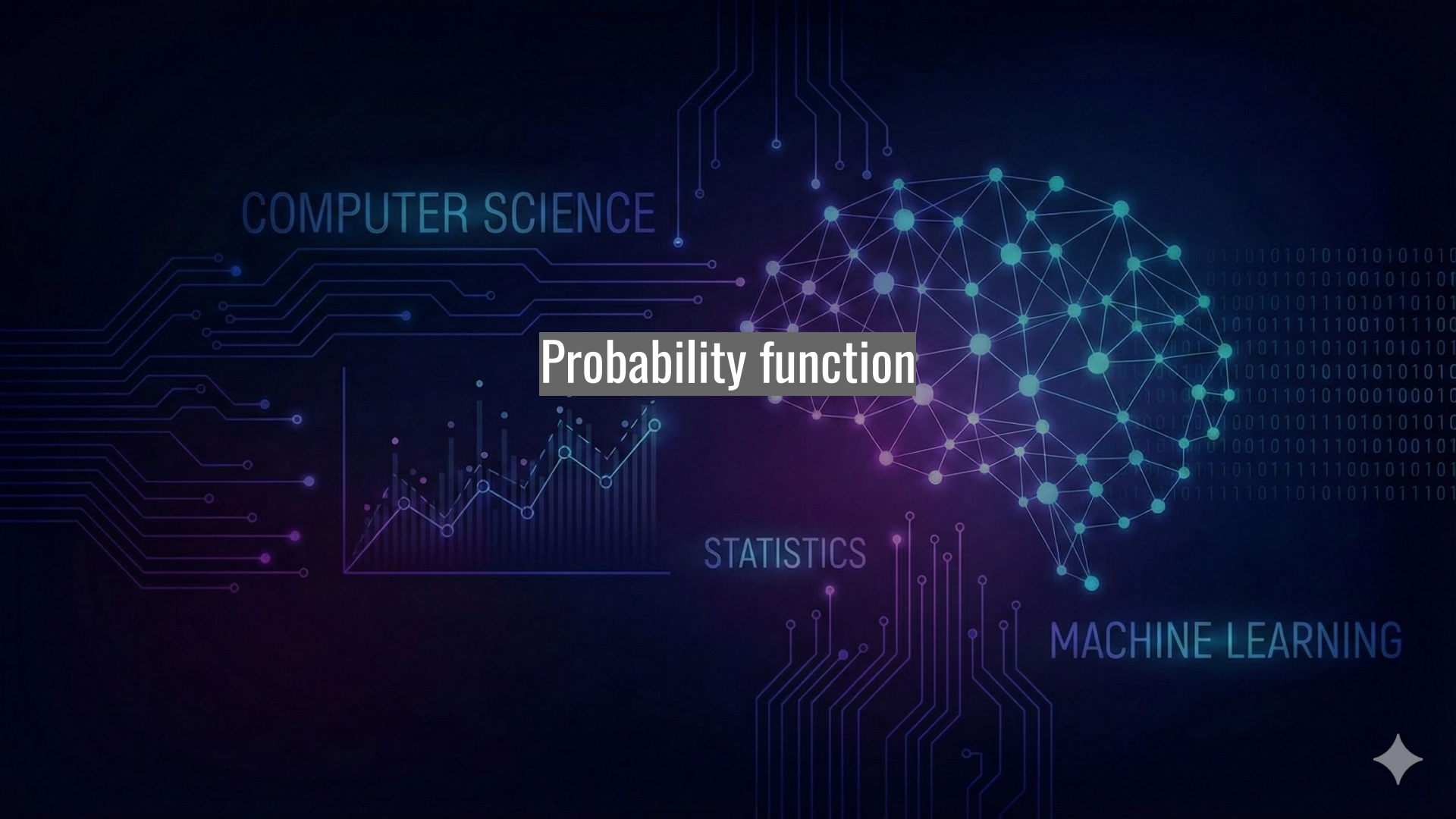


General Definition: Examples

Two dice are rolled. Let Y be the sum of the two numbers drawn. The sample space that describes all the possibilities in the two draws is

- $\Omega = \{(1, 1), (1, 2), \ldots, (1, 6), \ldots, (6, 1), (6, 2), \ldots, (6, 6)\}$
- the set of values assumed by X is $E = \{2, 3, \ldots, 12\}$

$Y((1, 1)) = 2, Y((2, 1)) = 3, Y((3, 2)) = 5.$


COMPUTER SCIENCE
Probability function
STATISTICS
MACHINE LEARNING

General Definition: Probability

Since a random variable is a function of random experiments, it assumes its values with a certain probability.

PROBABILITY QUESTIONS

$$P(X = 0) = ?$$

What is the probability that X is zero?

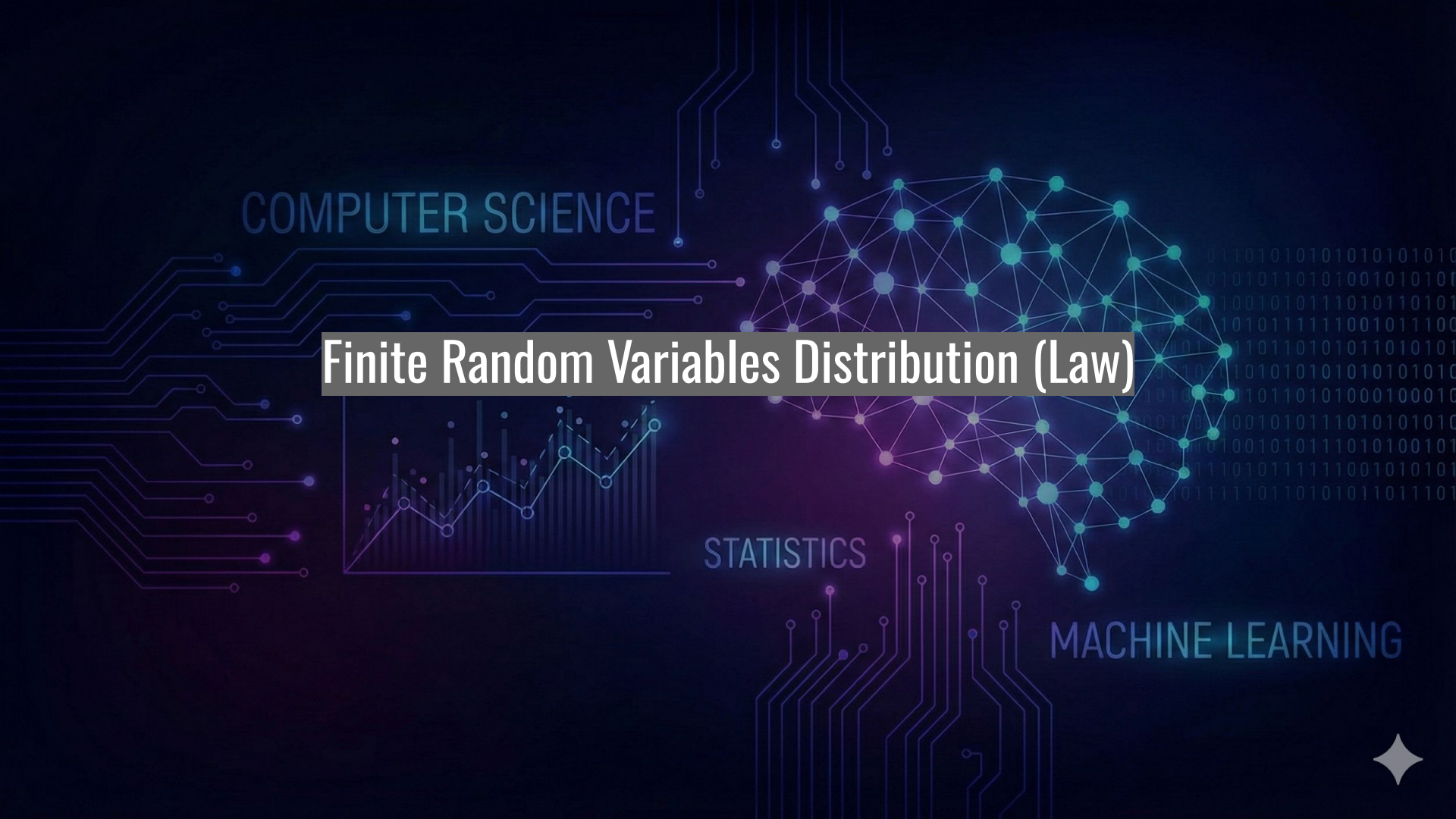
$$P(Y = 5) = ?$$

What is the probability that Y is five?

Probability function

The Probability Function translates random variables into probabilities.

Once the random variable has done its job of assigning numbers to all the outcomes, the Probability Function steps in. Its job is to take those numbers and calculate the exact likelihood (a value between 0 and 1) that each number will occur.Depending on whether your data is discrete or continuous, this is called a Probability Mass Function (PMF) or a Probability Density Function (PDF).


COMPUTER SCIENCE
Finite Random Variables Distribution (Law)
STATISTICS
MACHINE LEARNING

Distribution (Law)

The Distribution or "Law" of a finite random variable defines the complete mapping of outcomes to their likelihoods.

Mathematical Definition

$$P(X = x_i) = p_i$$

Probability of variable X assuming a specific value xi

$$\sum p_i = 1$$

The sum of all associated probabilities must equal unity

Finite Random Variables

Generalizing :

- A random variable X that takes values in $E = \{x_1, x_2, \ldots, x_n\}$
- Each with a certain probability: $P(X = x_i) = p_i$, $i = 1, 2, \ldots, n$.
- With $0 \leq p_i \leq 1$
- $\Sigma p_i = p_1 + p_2 + \ldots + p_n = P(\Omega) = 1$

Mean and Variance

Mean (μ_x)

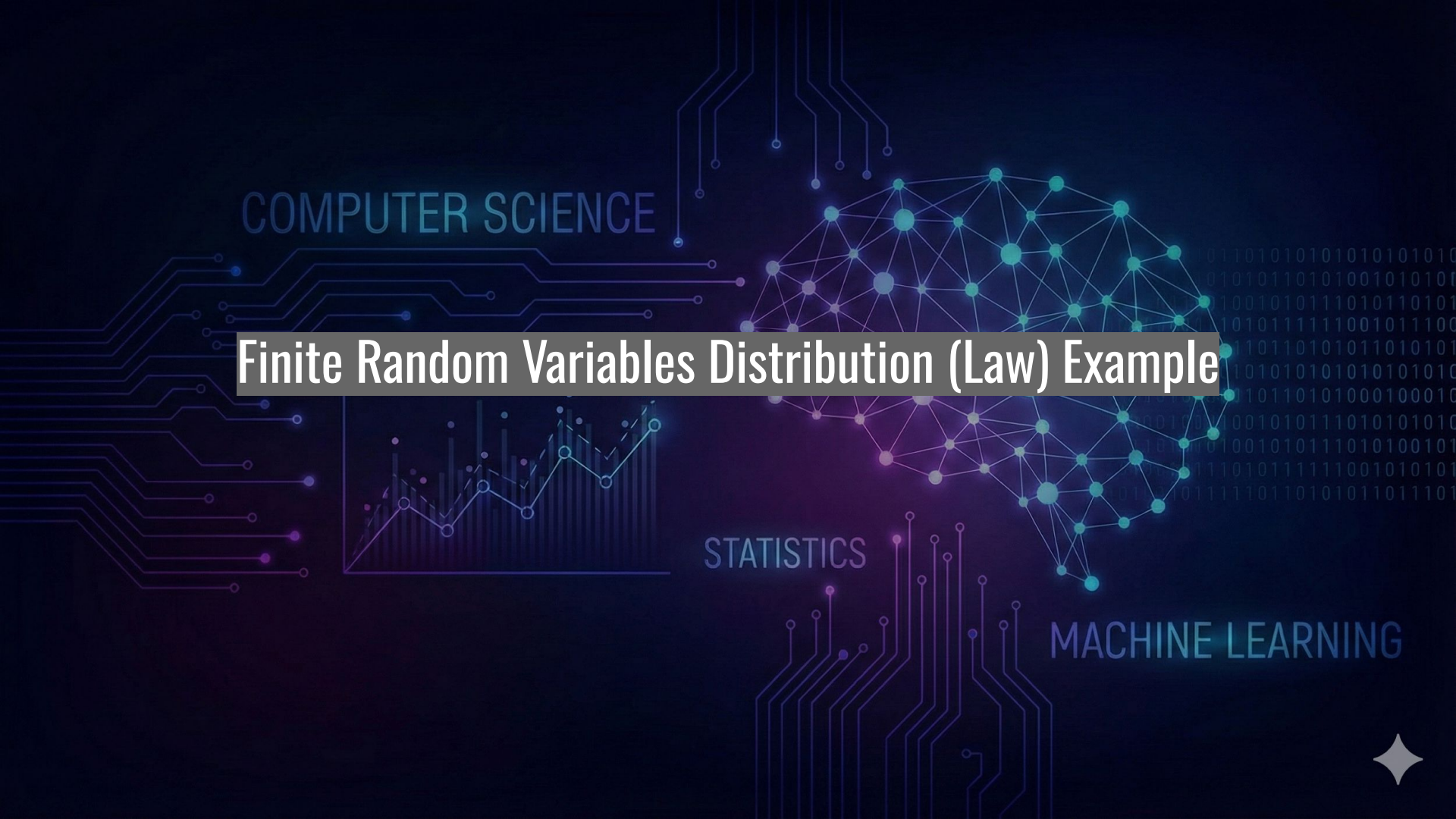
The **Expected Value** (Expectation). It serves as a position index representing the "center of gravity" of a distribution.

$$\mathbb{E}[X] = \sum x_i p_i$$

Variance (Var[X])

A measure of **dispersion** (spread) of the random variable around its mean value.

$$Var[X] = \sum (x_i - \mu_X)^2 p_i = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$$


COMPUTER SCIENCE
Finite Random Variables Distribution (Law) Example
STATISTICS
MACHINE LEARNING

Finite Random Variables: Examples

Considering the first example X, as already mentioned, can take on values in E = {0, 1}.

- **What does it mean that X = 0?**
  - **It means that the throw gave an even result.**

    Formalizing, we have that the event {X = 0} can be written as follows,
    - **{X = 0} = {2, 4, 6}**
    - **Probability P(X = 0) = P({2, 4, 6}) = 3/6 = 1/2**

Finite Random Variables: Examples

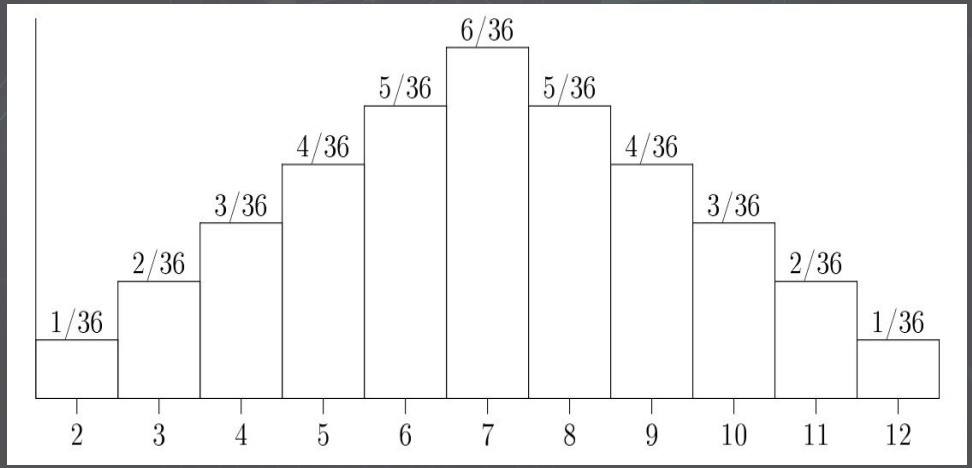
$$_Y = \overline{XY} - \overline{X} \cdot$$

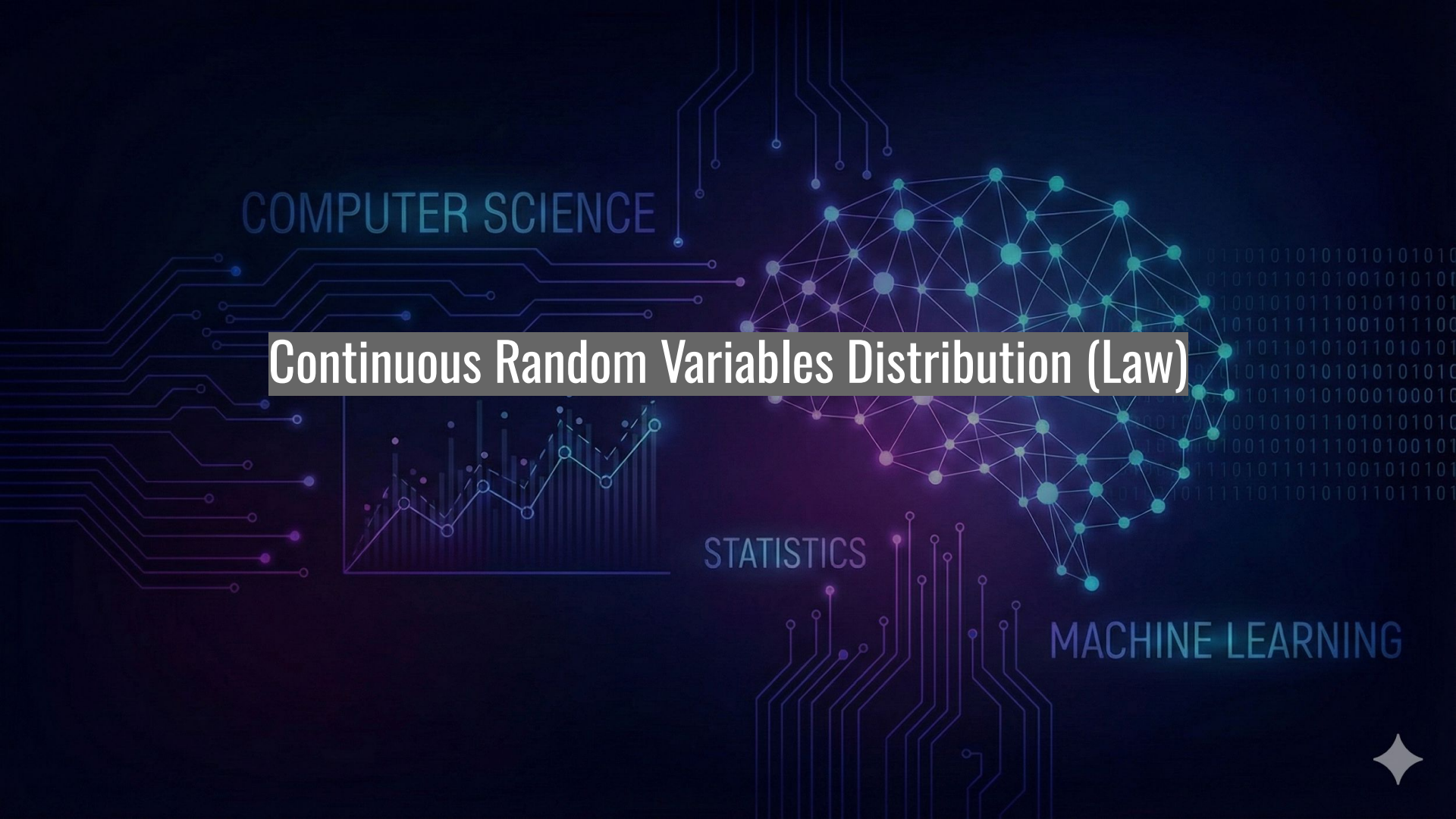
The set of values assumed by X and the probabilities with which they are assumed is called the **law or distribution of X**.

Two Dice Experiment

The random variable **Y** represents the sum of the results of two dice.

$\{Y = 2\} = \{(1, 1)\}$
$P(Y=2) = P(\{(1, 1)\}) = 1/36$




COMPUTER SCIENCE
Continuous Random Variables Distribution (Law)
STATISTICS
MACHINE LEARNING

Continuous Random Variables

Probability Density Function (f(x))

Probabilities are defined as the **area under the curve**.

Note: For continuous variables, P(X=x) = 0 for any single point.

$$P(a < X \leq b) = \int_a^b f(x)dx$$

Fundamental Properties

- Non-negativity: f(x) ≥ 0 for all x ∈ ℝ
- Normalization: The total area under the density curve must equal 1.

$$\int_{\mathbb{R}} f(x)\,dx = 1$$

The Mean: Expected Value

Conceptual Foundation

Transitioning from discrete to continuous domains requires a shift in mathematical operations:

- In **discrete math**, we use summation:
  - $\sum x\ p(x)$.
- In **continuous math**, we use integration.
- **Center of Gravity:** The Mean represents the physical balance point of the probability density function.

Mathematical Definition

The Expected Value E[X] (or μ_X) is calculated by integrating the value x weighted by the probability density f(x) over the entire range.

$$\mu_X = \int_{-\infty}^{+\infty} x \cdot f(x)\, dx$$

Variance & Standard Deviation

Conceptual Definition

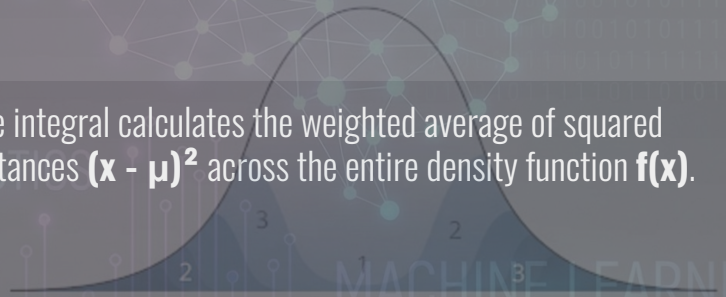
Variance (Var[X]) measures the **spread** of a distribution.

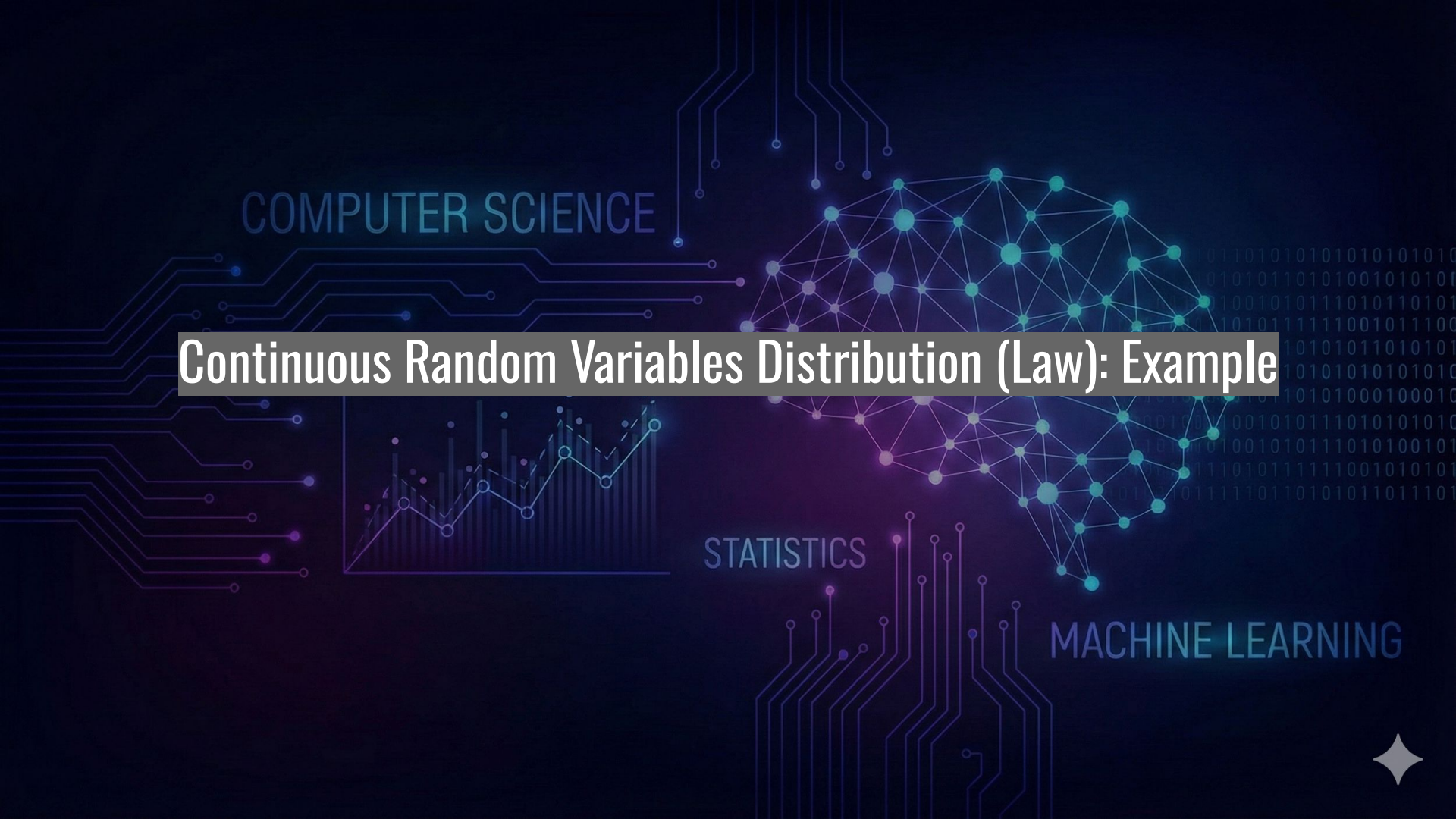
It represents the expected value of the squared deviation from the mean, quantifying how much the values differ from the average.

Mathematical Formula

$$\sigma_X^2 = \int_{-\infty}^{+\infty} (x - \mu)^2 f(x)\, dx$$

The integral calculates the weighted average of squared distances **(x - μ)²** across the entire density function **f(x)**.





Continuous Random Variables Distribution (Law): Example

Continuous Random Variables: Example

Probability Density Function f(x)

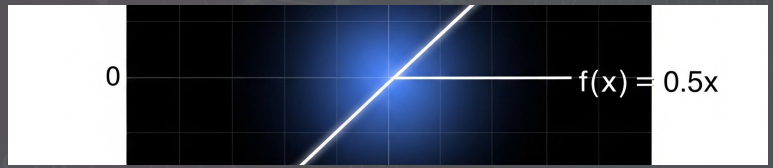
Let X be the continuous random variable with density function f given by:

$$f(x) = \begin{cases} \frac{1}{2}x & 0 \le x \le 2 \\ 0 \end{cases}$$

Normalization Property Verification

The total area under the density curve must equal 1 for the function to be a valid PDF.

We verify this by integrating over the range:

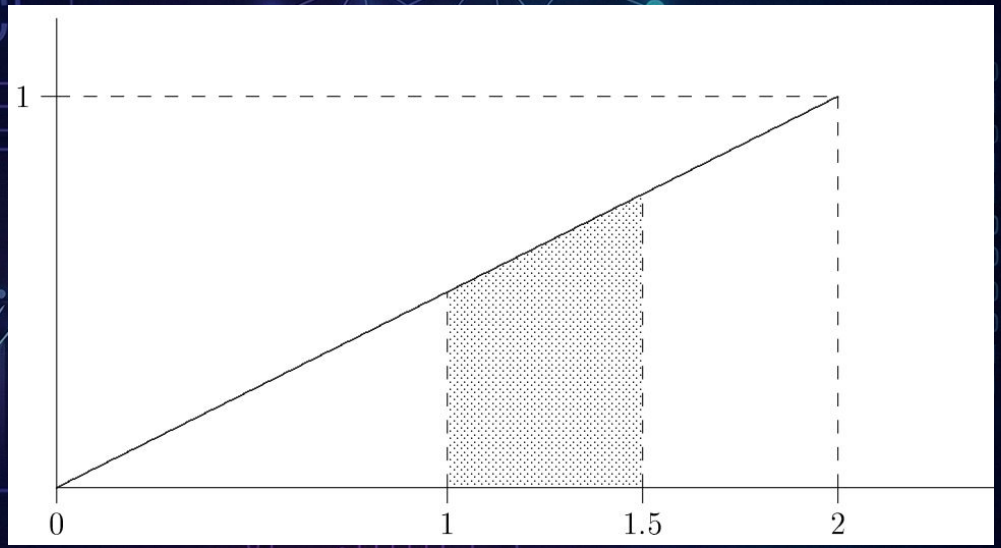


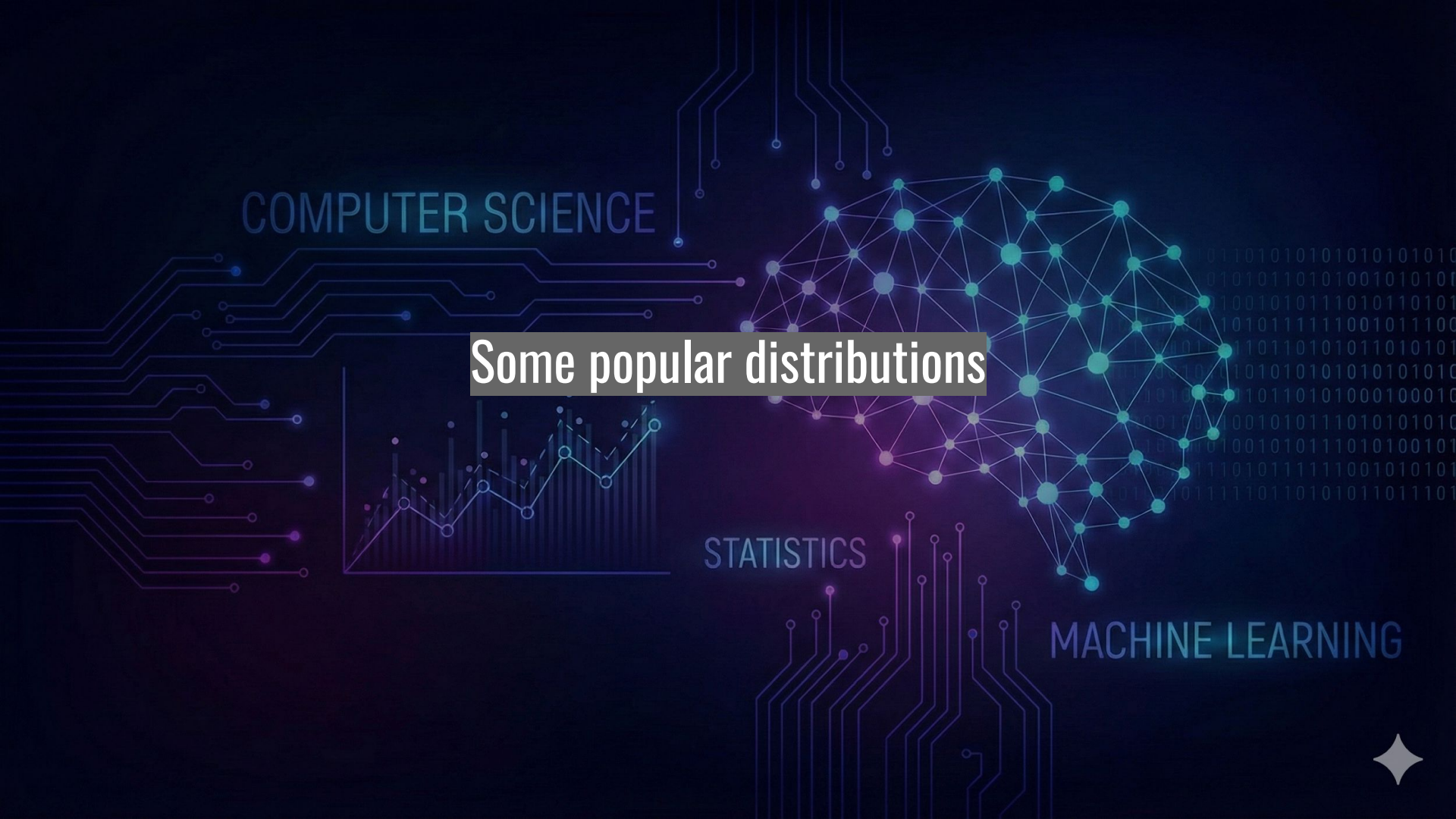
$$\int_{\mathbb{R}} f(x)\,dx = \int_0^2 \frac{1}{2}x\,dx = \frac{1}{4}x^2\Big|_0^2 = 1.$$

Continuous Random Variables: Example

Let's calculate P(1 < X < 1.5).

$$\mathbb{P}(1 \le X \le 1.5) = \int_1^{1.5} \frac{1}{2} x \, dx = \frac{1}{4} x^2 \Big|_1^{3/2} = \frac{5}{16}.$$

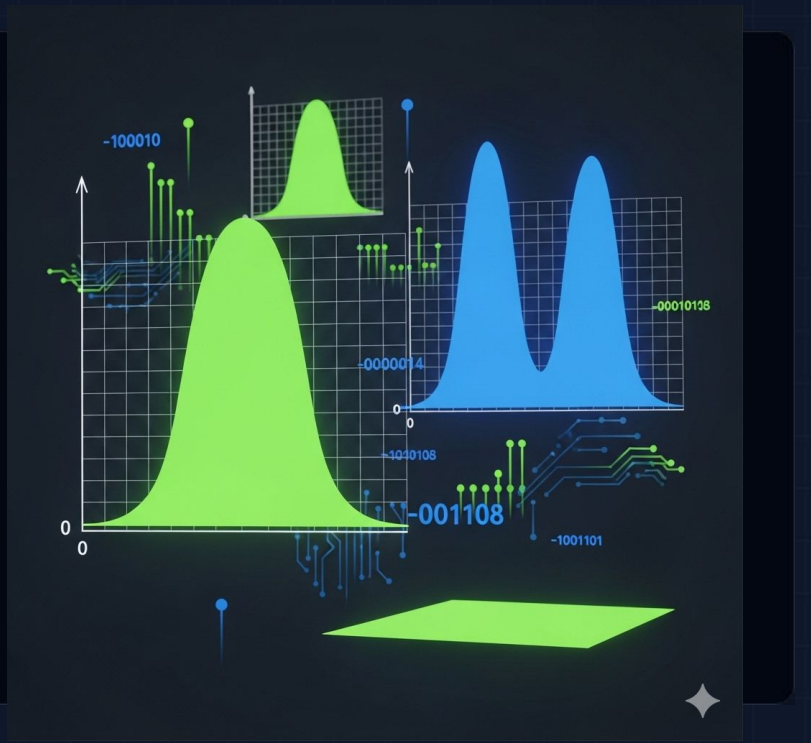



COMPUTER SCIENCE
Some popular distributions
STATISTICS
MACHINE LEARNING

The Core Concept

A statistical distribution is not an arbitrary curve; it is fundamentally defined by its **parameters**.

- **The Control Dials:** Think of parameters as the settings on a machine. When you adjust the parameters, the entire distribution curve shifts, stretches, or
- **Mathematical Dependence:** The mean (expected value) and standard deviation (variance) are not independent traits. They are directly calculated *from* the parameters.



Discrete Distributions

How parameters dictate the mean (μ) and variance (σ^2) in discrete spaces.

Binomial Distribution

Parameters: n (trials) and p (probability of success).

$$\mu = n \cdot p$$

$$\sigma^2 = n \cdot p \cdot (1 - p)$$

Poisson Distribution

Parameter: λ(lambda), the average rate of occurrence.

$$\mu = \lambda$$

$$\sigma^2 = \lambda$$

*In Poisson, the mean and variance are identical.

Continuous Distributions

How boundaries and direct parameters shape the continuous curves.

Uniform Distribution

Parameters: The boundary points a and b.

$$\mu = \frac{a + b}{2}$$

$$\sigma^2 = \frac{(b - a)^2}{12}$$

Normal Distribution

This is a highly special, convenient case where the parameters *are* the mean and variance.

$$N(\mu, \sigma^2)$$

*The curve's center is mapped directly to μ and its width is mapped directly to σ^2.

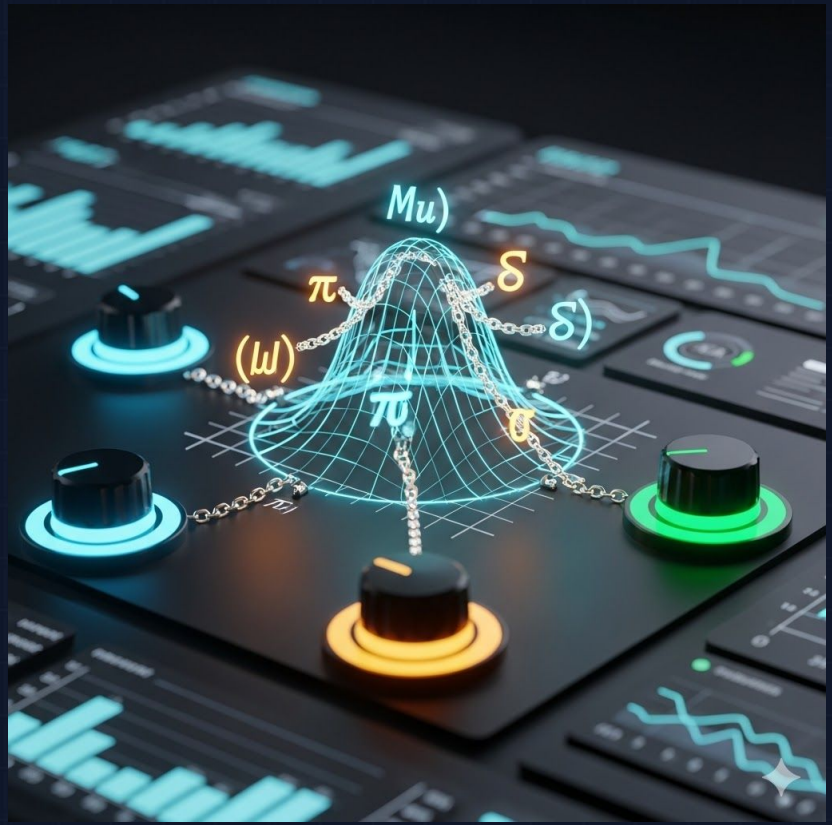
Adjusting the Dials

Dynamic Mathematical Models

When an analyst changes a parameter, they are mathematically reconstructing the entire probability space.

For example, in a Binomial distribution **Bi(n, p)**, if you increase the probability of success **p**, the peak of the curve physically slides to the right.

Simultaneously, the variance expands or contracts based exactly on the mathematical output of **np(1-p)**.

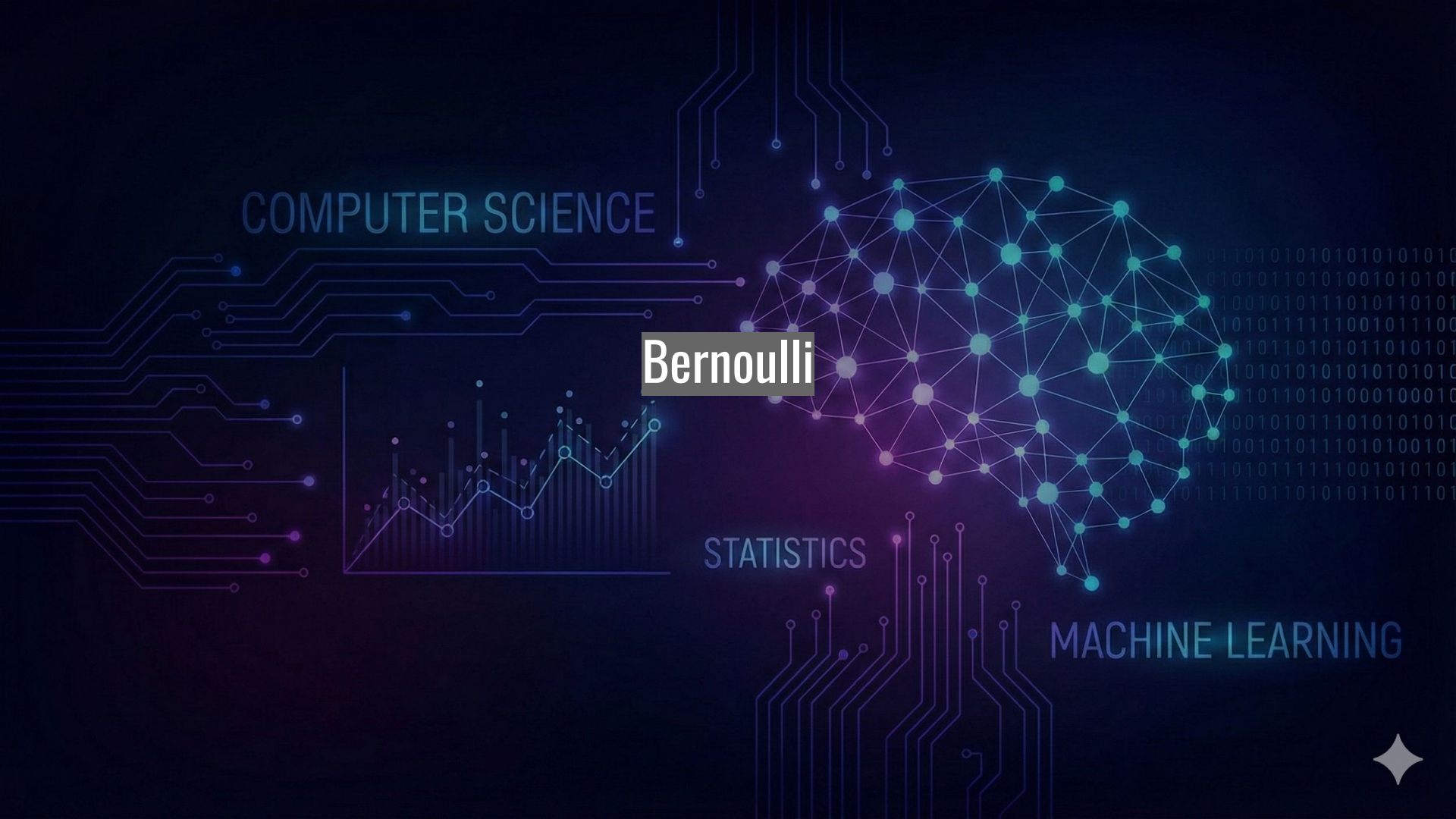


The Big Takeaway

100% Predictability

The core concept is absolute mathematical dependence.

You do not need to "measure" or "guess" the mean or variance of a known probability distribution. If you know the parameters that define the distribution, the expected value and standard deviation are automatically, deterministically known.


COMPUTER SCIENCE
Bernoulli
STATISTICS
MACHINE LEARNING

Bernoulli Distribution: **BE(P)**

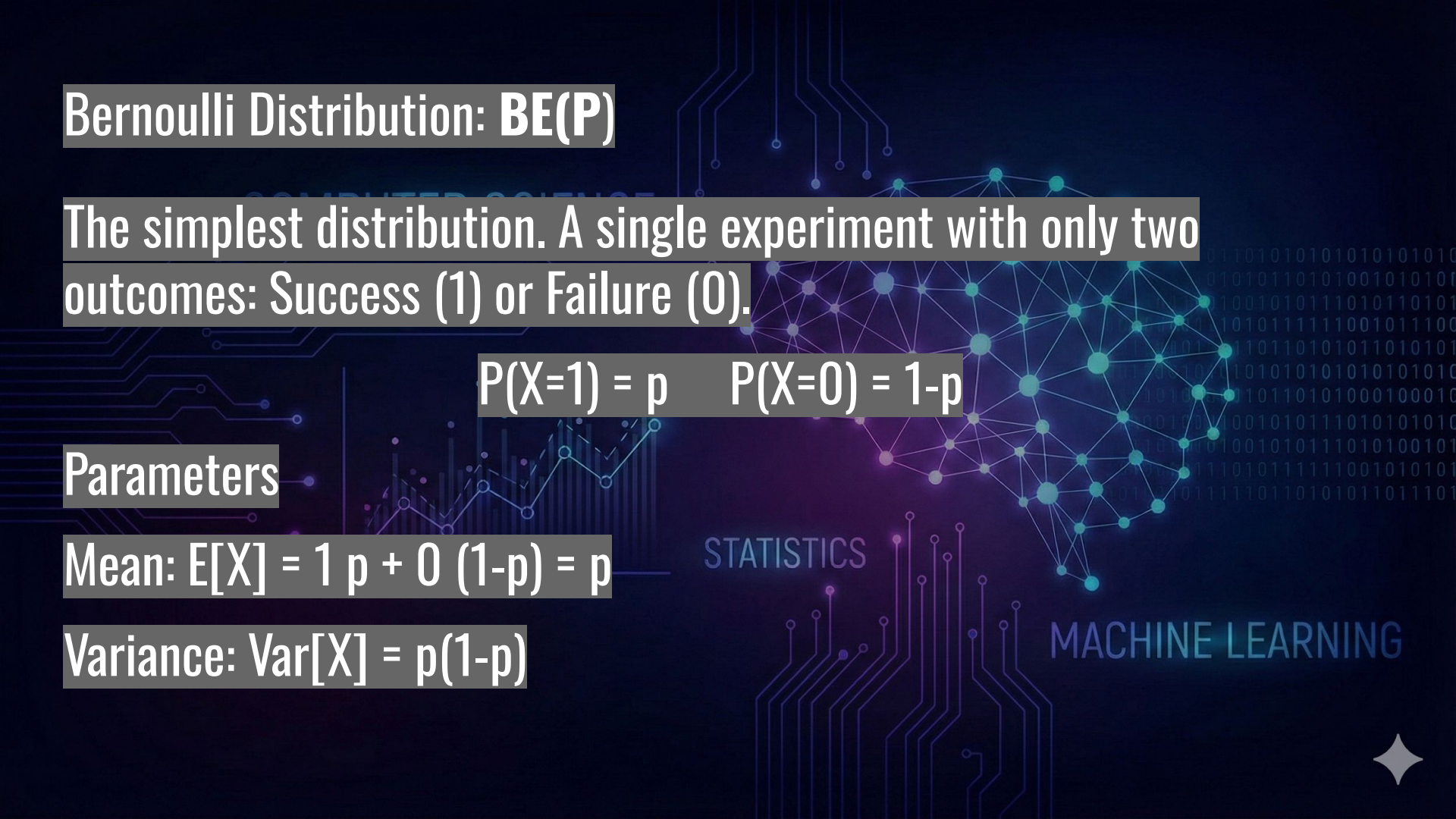
The simplest distribution. A single experiment with only two outcomes: Success (1) or Failure (0).

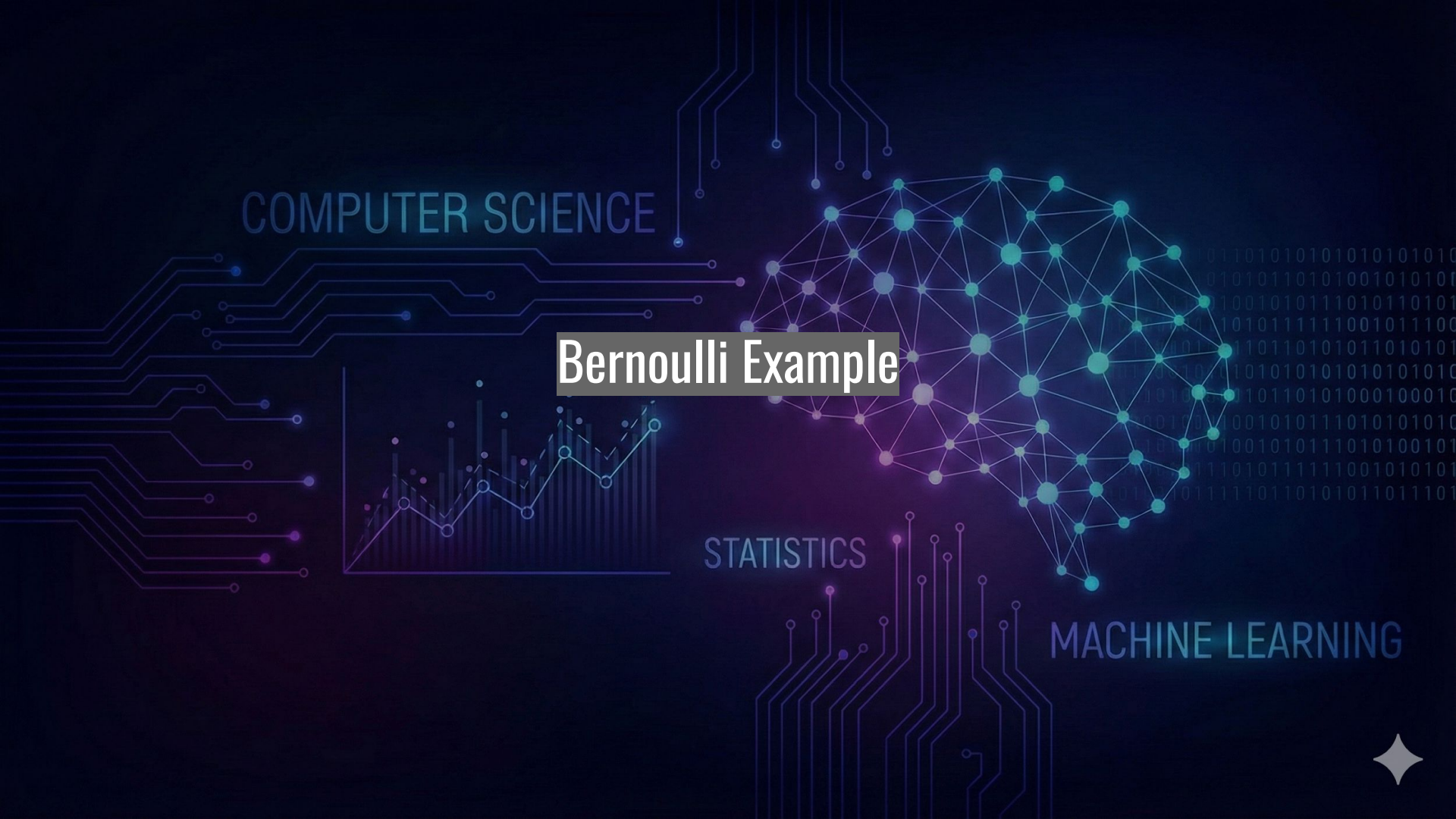
$$P(X=1) = p \qquad P(X=0) = 1-p$$

Parameters

Mean: $E[X] = 1\,p + 0\,(1-p) = p$

Variance: $Var[X] = p(1-p)$




COMPUTER SCIENCE
Bernoulli Example
STATISTICS
MACHINE LEARNING

Bernoulli Distribution: example

Experiment: Tossing a fair coin once. Outcomes: Heads (H) or Tails (T).

Random Variable (X):

- We define Heads as "Success" (1).
- We define Tails as "Failure" (0).

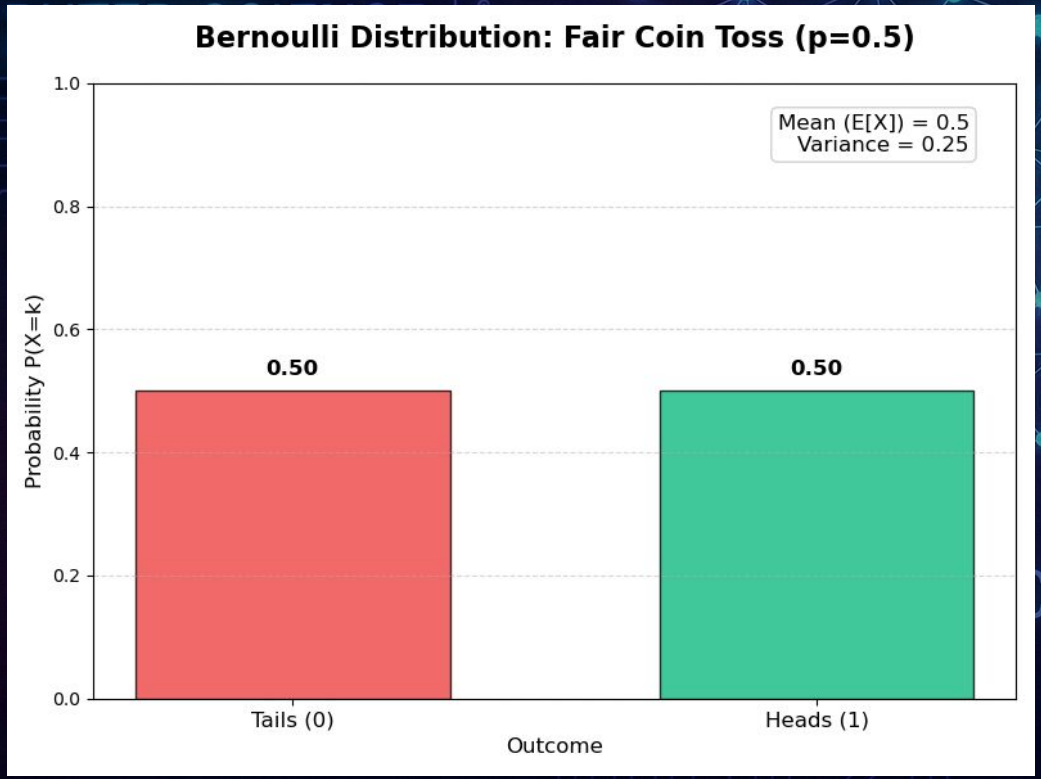
Parameters:

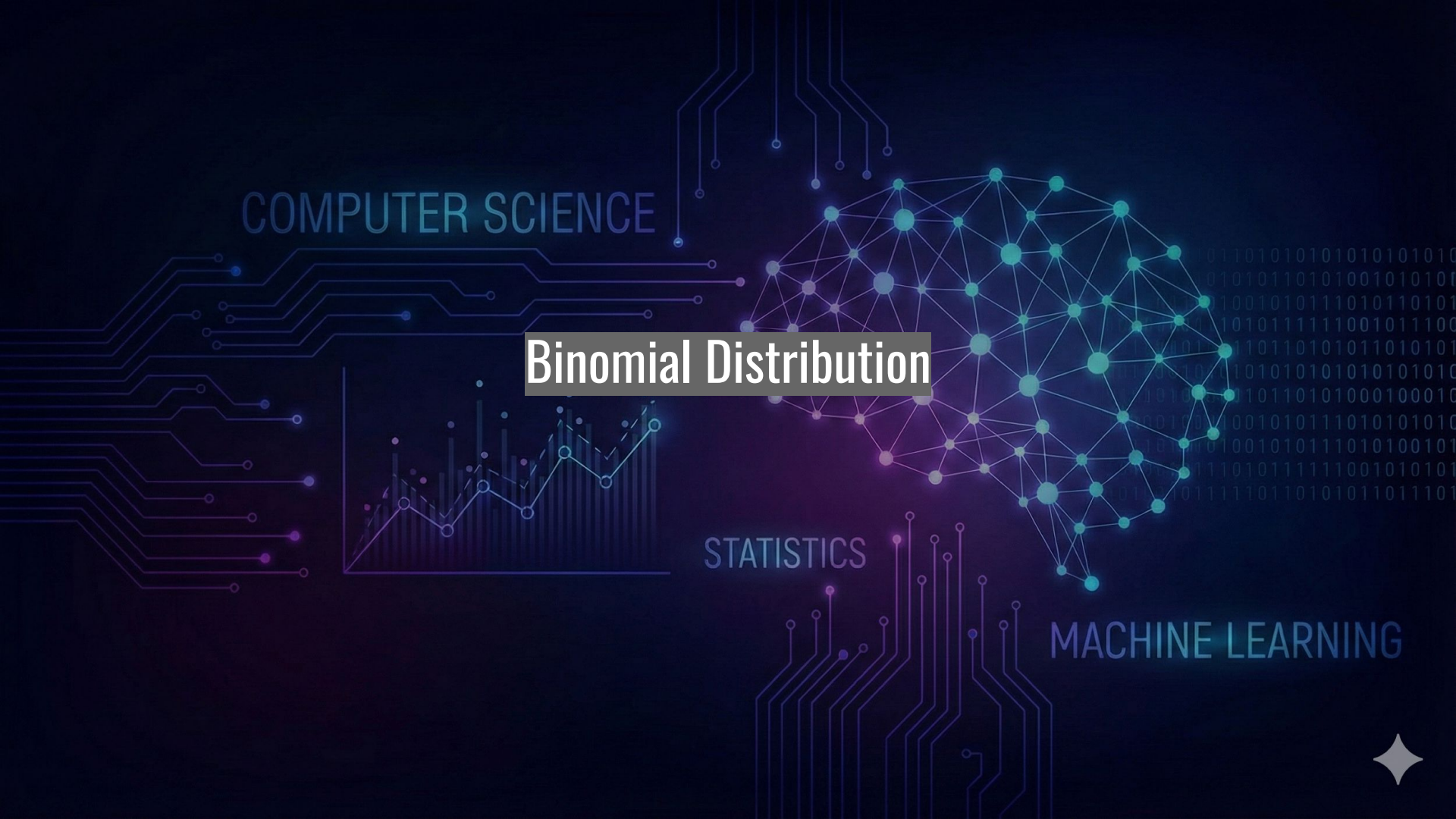
- p (Probability of Success): Since the coin is fair, p = 0.5.
- q (Probability of Failure): 1 - p = 0.5.

Statistics:

- Mean (μ) : p = 0.5 (If you tossed it infinitely, the average value of 0s and 1s would be 0.5).
- Variance (σ^2): p × (1-p) = 0.5 × 0.5 = 0.25.

Bernoulli Distribution: example




COMPUTER SCIENCE
Binomial Distribution
STATISTICS
MACHINE LEARNING

Binomial Distribution **Bi(n, p)**

Performing **n** independent Bernoulli trials and counting the total number of successes **k**, **p** is the probability of success (on each single trial).. **X** can take values from **0** (no successes) to **n** (all successes), that is, **E** = **{0, 1, 2, . . . , n}**.

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

Mean: E[X] = n × p

Variance: Var[X] = n × p × (1-p)

Binomial Distribution **Bi(n, p)**

In the context of the binomial distribution $\binom{n}{k}$ answers one specific question: If I have n total trials, how many different ways can I arrange exactly k successes?

n! = The factorial of your total number of trials.

k! = The factorial of your number of successes.

(n - k)! = The factorial of your number of failures.

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

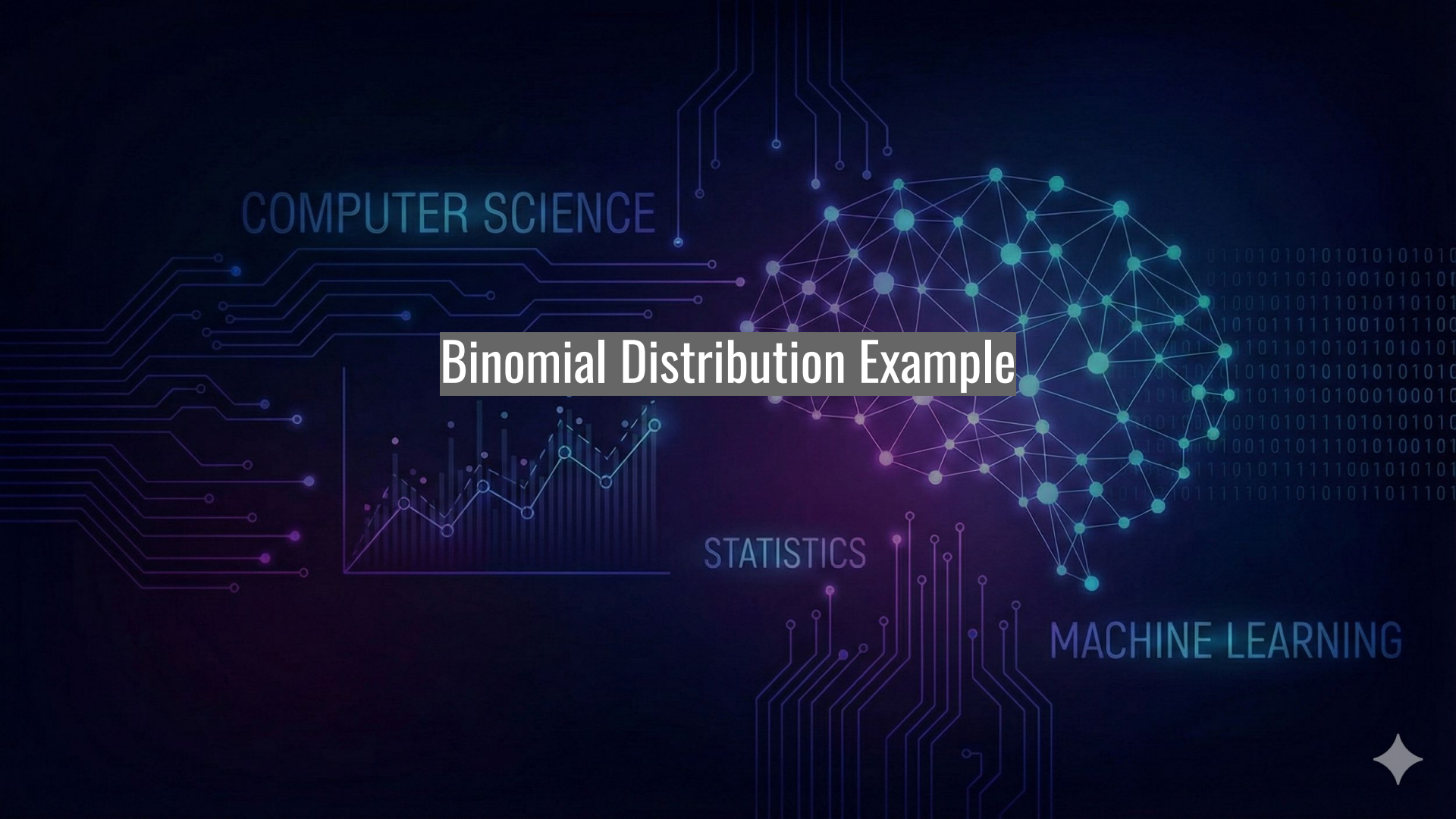
Binomial Distribution **Bi(n, p)**

Let's look at a small example to make it easy to visualize. Imagine you flip a coin 3 times (n=3) and you want exactly 2 heads (k=2). How many ways can that happen?

1. H H T (Heads on the first two, Tails on the last)
2. H T H (Heads on the first and last, Tails in the middle)
3. T H H (Tails on the first, Heads on the last two)

$$\binom{3}{2} = \frac{3!}{2!(3-2)!} = \frac{3!}{2! \times 1!}$$

$$\frac{3 \times 2 \times 1}{(2 \times 1) \times 1}$$


COMPUTER SCIENCE
Binomial Distribution Example
STATISTICS
MACHINE LEARNING

Binomial Distribution Example

We perform an experiment where we toss a fair coin n = 6 times. We are interested in the number of Heads (Successes) obtained.

Distribution: Binomial B(n, p)

Parameters:

- n = 6 (Number of trials)
- p = 0.5 (Probability of Heads per toss)

Random Variable (X): The total number of Heads. Possible values are {0, 1, 2, 3, 4, 5, 6}.

Binomial Distribution Example

We perform an experiment where we toss a fair coin **n = 6** times. We are interested in the total number of **Heads (Successes)** obtained.

Distribution

B(n, p)

Trials (n)

6

Probability (p)

0.5 (Fair Coin)

Random Variable (X)

The total number of Heads across all trials.

Possible Values

0 1 2 **3** 4 5 6

Where 3 is the most probable outcome (Expected Value) is the mean .

Binomial Distribution Example

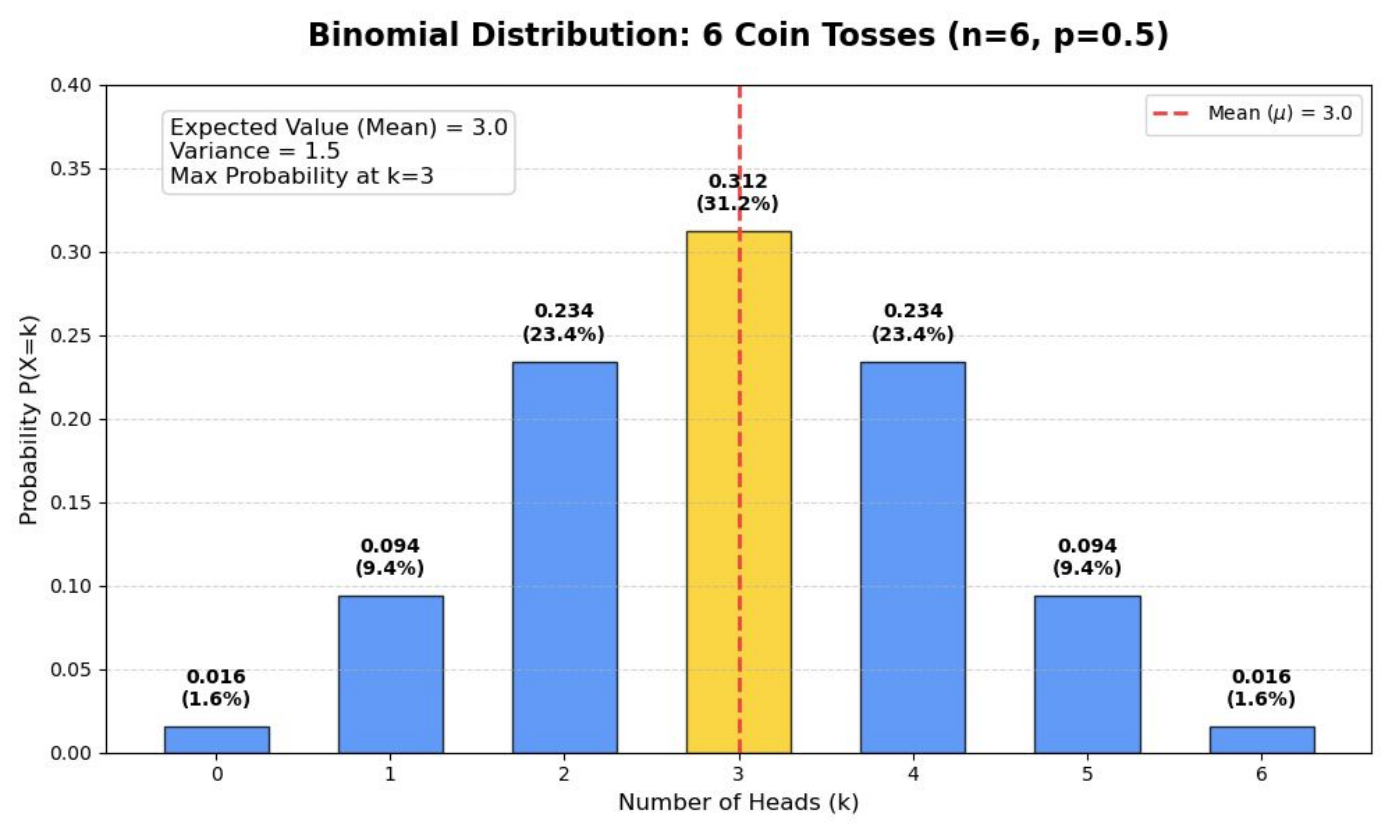
The Formula:

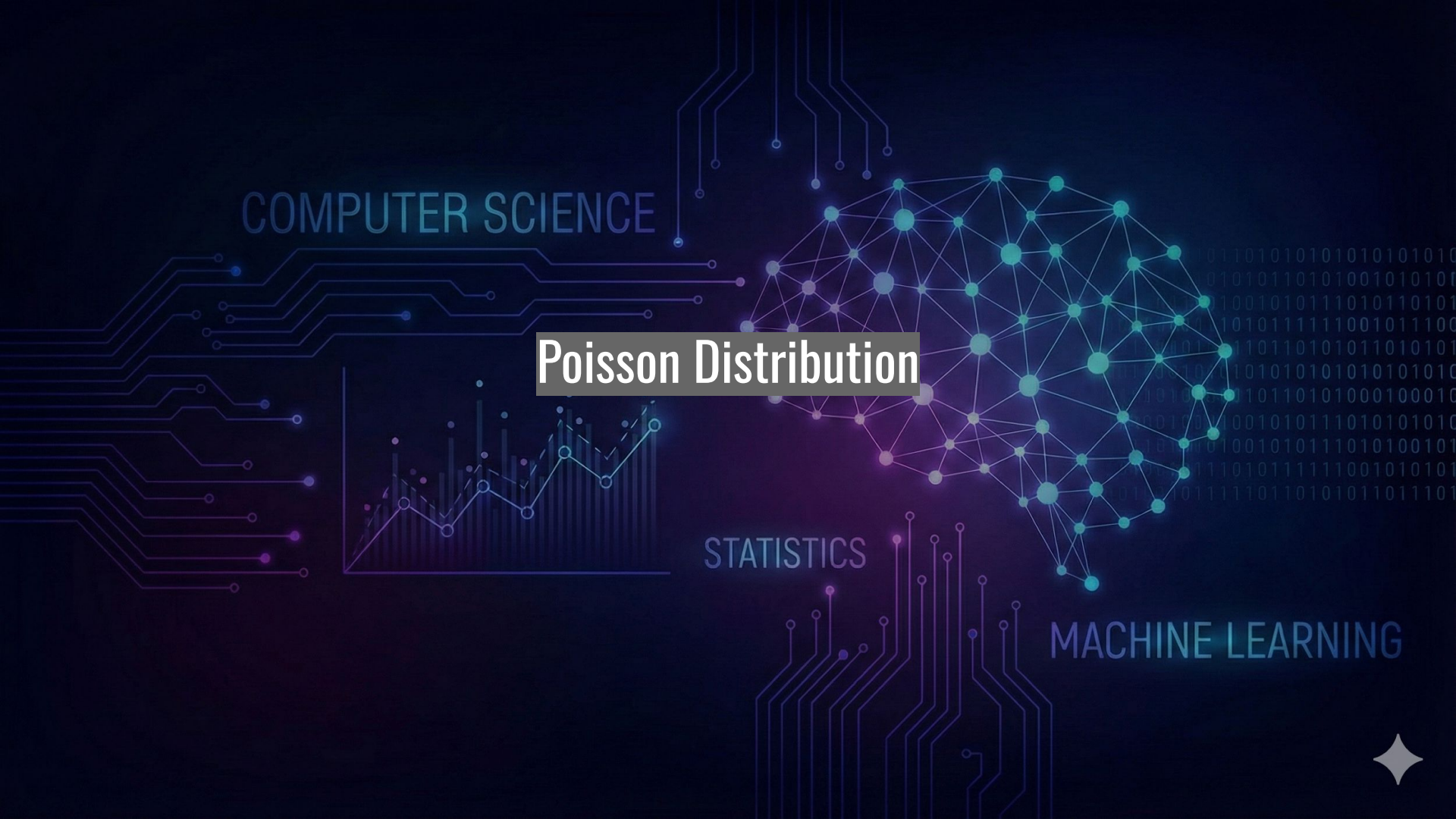
$$P(X = k) = \binom{6}{k} (0.5)^k (0.5)^{6-k}$$

Expected Value (Mean): μ = n × p = 6 × 0.5 = 3 (We expect 3 Heads on average).

Variance Var[X] = n × p × (1 - p) = 6 × 0.5 × (1 -0.5) = 1.5

Binomial Distribution Example




COMPUTER SCIENCE
Poisson Distribution
STATISTICS
MACHINE LEARNING

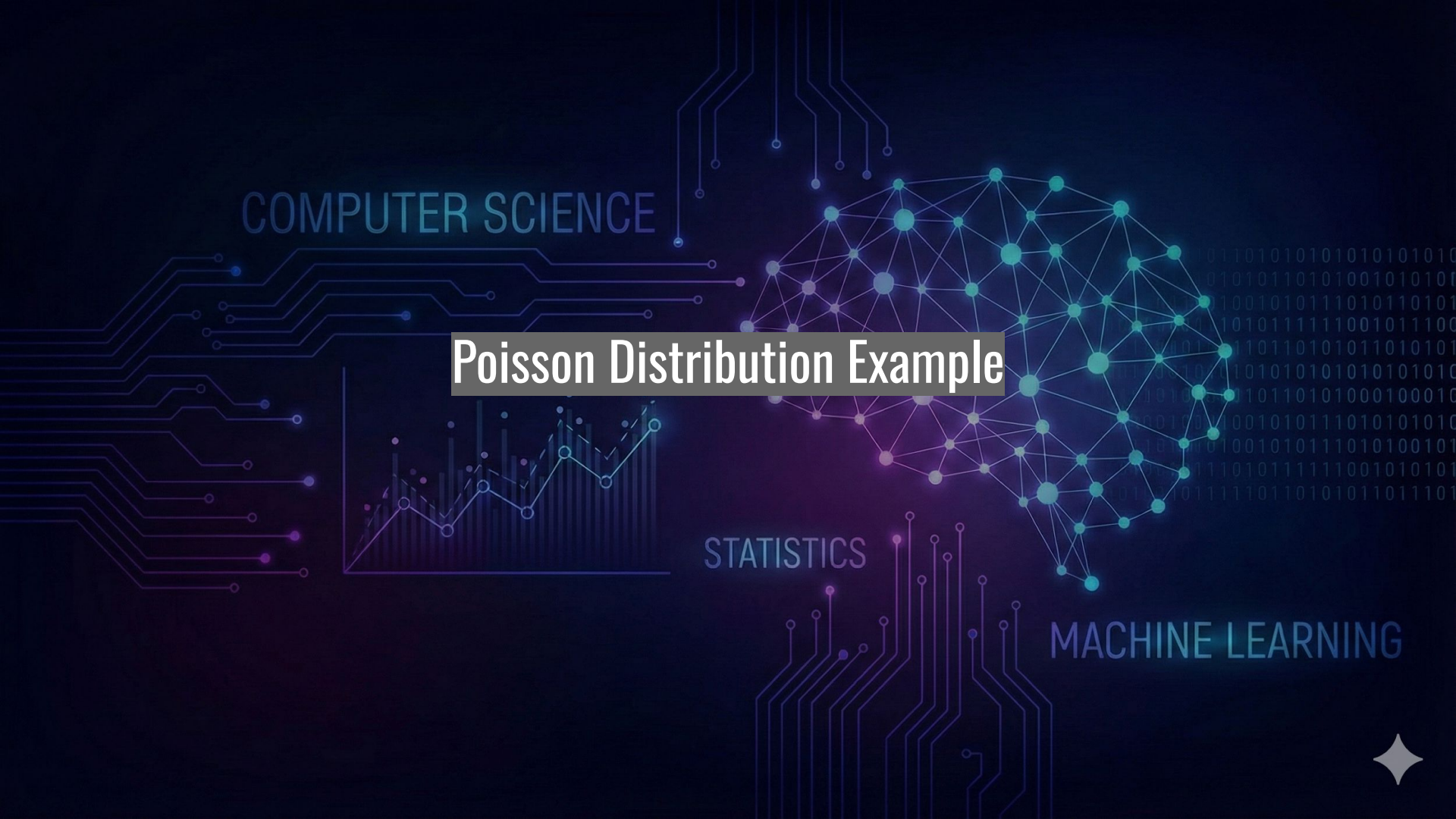
Poisson Distribution

Modeling the number of events occurring in a fixed interval of time or space (e.g., calls to a call center, cars passing), parameter **λ > 0** is the average number of event accurring per time interval and **k** number of accurences

$$P(X = k) = e^{-\lambda} \frac{\lambda^k}{k!}$$

Mean (E[X]): The expected number of events is exactly λ.

Variance (Var(X)): The spread (dispersion) of the data is also λ


COMPUTER SCIENCE
Poisson Distribution Example
STATISTICS
MACHINE LEARNING

Poisson Distribution Example

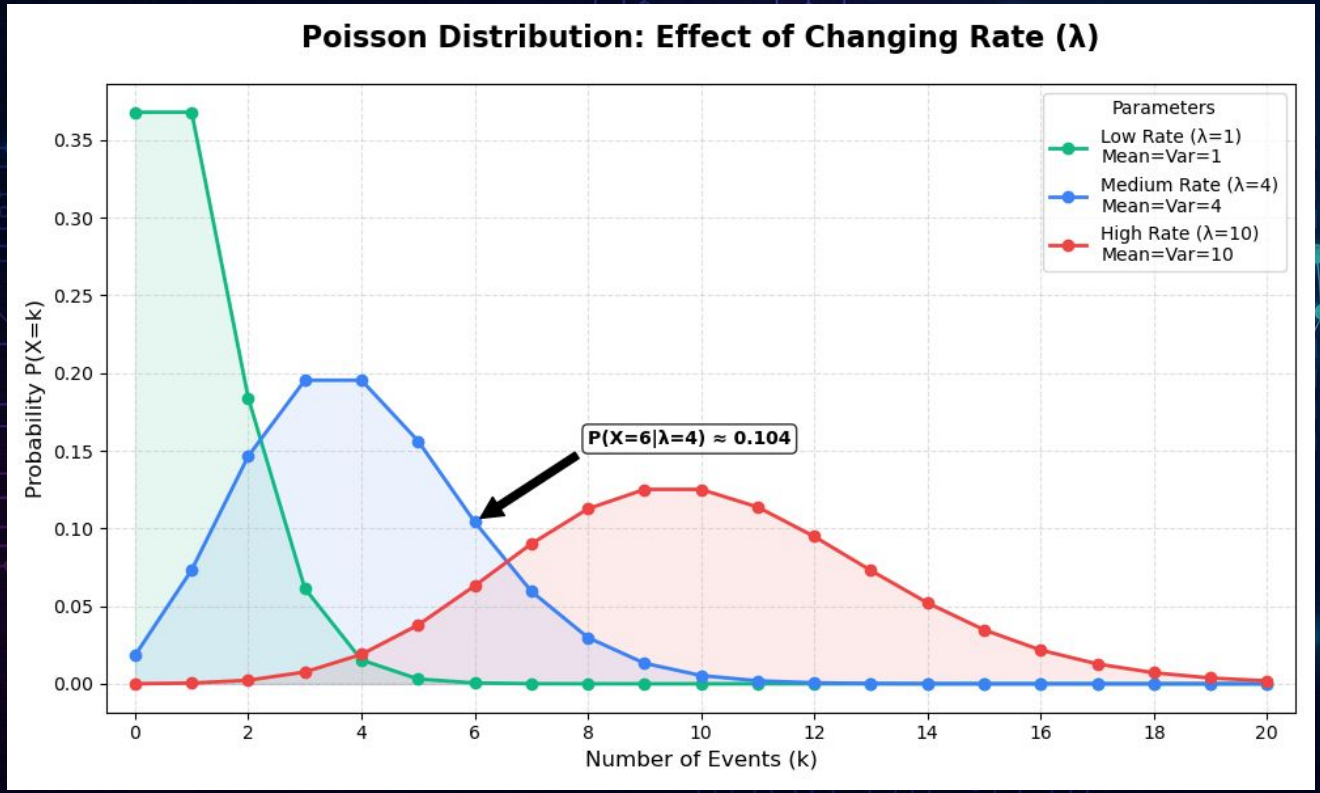
Imagine you manage a small help desk. On average, you receive 4 calls per minute.

- Variable (X): The number of calls received in a specific minute.
- Parameter (λ): λ = 4 calls/minute (average rate).
- We want to know:

What is the probability of receiving exactly 6 calls in the next minute (k=6)? (10.4 % of probability)

$$P(X = 6) = \frac{e^{-4} 4^6}{6!} \approx 0.104$$

Poisson Distribution Example



STATISTICAL MECHANICS

Gas Theory & The Poisson Distribution

Why ideal gases perfectly manifest the mathematics of independent, random events in space and time.

Clarifying the Confusion

When discussing Kinetic Gas Theory, we are actually dealing with **two different mathematical distributions** depending on what we are trying to measure.

Maxwell-Boltzmann

What it measures: The speeds or kinetic energies of the gas molecules.

This is a *continuous* distribution. It defines the famous bell-like curve showing that a few particles are very slow, a few are extremely fast, but most travel at an average speed.

Poisson Distribution

What it measures: The distinct number of particles in a specific tiny volume, or the number of collisions over time.

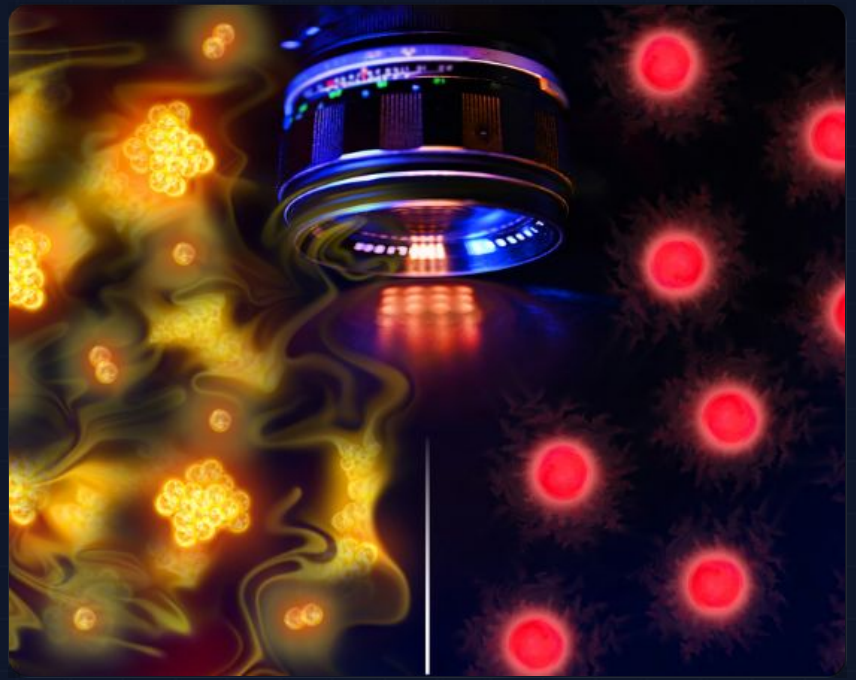
This is a *discrete* distribution. It predicts exactly how many individual events or items we can expect to count.

1. Particles in a Microscopic Volume

Imagine a large container filled with an ideal gas. If you zoom in and draw a microscopic, imaginary box inside that container...

How many gas molecules are inside that tiny box at any given exact millisecond?

This scenario perfectly maps to the Poisson Distribution because it satisfies all the strict mathematical requirements of a Poisson Process in 3D space.



Why the Math Matches the Physics

The physical properties of an ideal gas are identical to the assumptions required for a Poisson model.

Independence

In an ideal gas, molecules do not attract or repel each other (except during direct collisions). The position of Molecule A has absolutely no effect on the position of Molecule B.

Constant Rate (λ)

Because the gas is evenly spread out macroscopically, there is a known, constant average density. We know there is an average rate (λ) of particles per volume.

The "Rare Event"

The chance of one specific molecule out of trillions landing in your tiny box is incredibly small. A Binomial distribution with massive n and tiny p mathematically becomes Poisson!

2. The Number of Collisions

Tracking Events Over Time

Now, instead of looking at space, let's track one single gas molecule over time.

How many times will it crash into another molecule in the next 1 second?

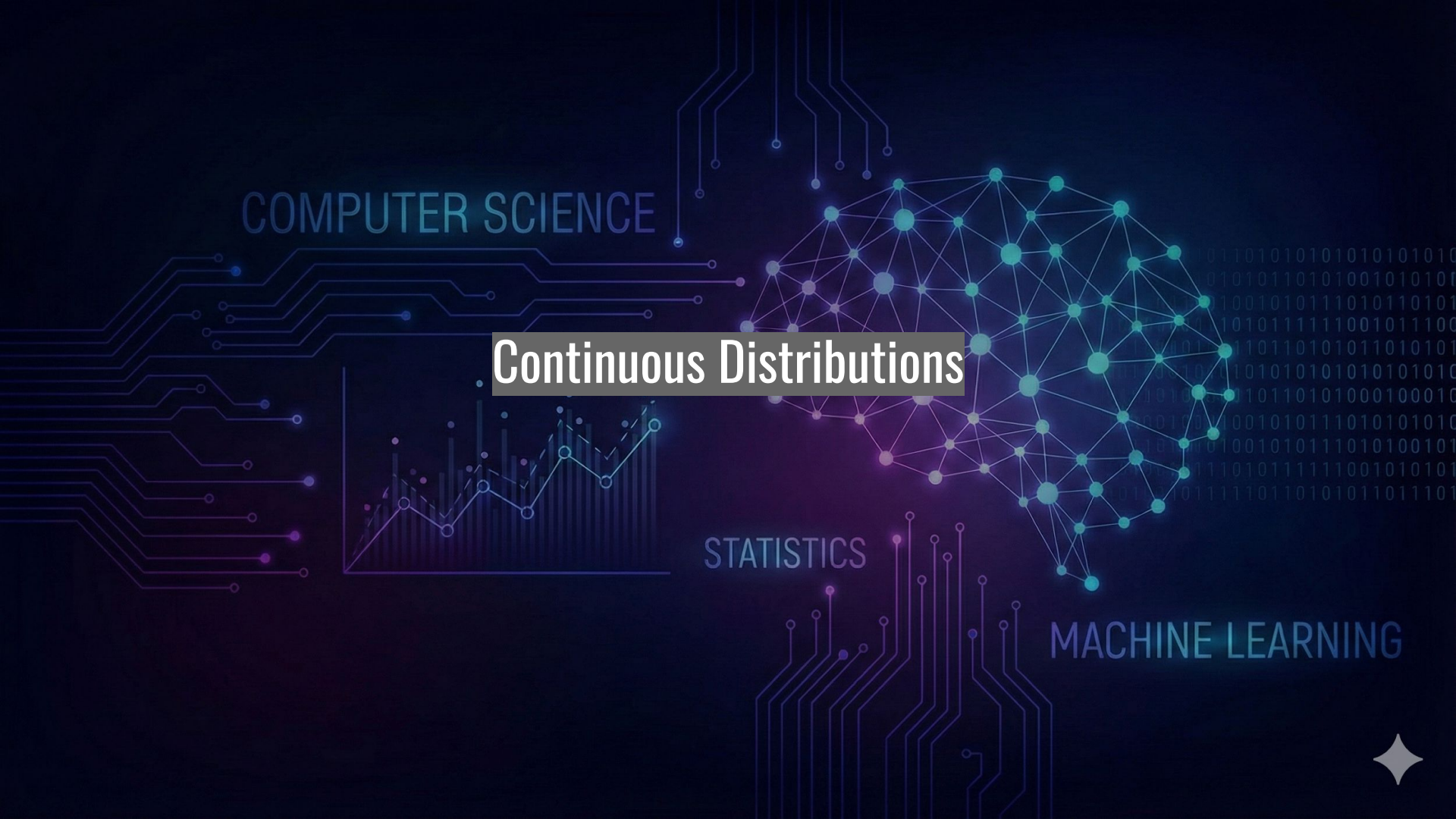
- Collisions happen at a constant average rate based on pressure and temperature.
- Collisions are "Memoryless". Knowing a collision happened 1 microsecond ago doesn't change the probability of one happening in the next microsecond.

The Physical Manifestation

The Ultimate Random Process

You see the Poisson distribution in gas theory because an ideal gas is the ultimate physical manifestation of **independent, random events happening at a constant average rate**.

Whether you are counting particles inside a microscopic volume, or counting the number of collisions over a span of time, the underlying mathematics are identical.


COMPUTER SCIENCE
Continuous Distributions
STATISTICS
MACHINE LEARNING

Uniform distribution (continued)

It describes a variable where every interval of the same length within the range [a, b] has the same probability.

Since probability is constant, the density function is a flat horizontal line (a rectangle).

$$f(x) = \frac{1}{b-a} \quad \text{for } a \leq x \leq b$$

Uniform distribution (continued)

Mean (Expected Value): Since the distribution is symmetric and rectangular, the mean is simply the midpoint of the interval.

$$\mathbb{E}[X] = \int_a^b x \cdot \left(\frac{1}{b-a}\right) dx = \frac{1}{b-a}\int_a^b x\,dx =$$

$$\frac{1}{b-a}\left[\frac{x^2}{2}\right]_a^b = \frac{1}{b-a}\cdot\frac{b^2-a^2}{2} =$$

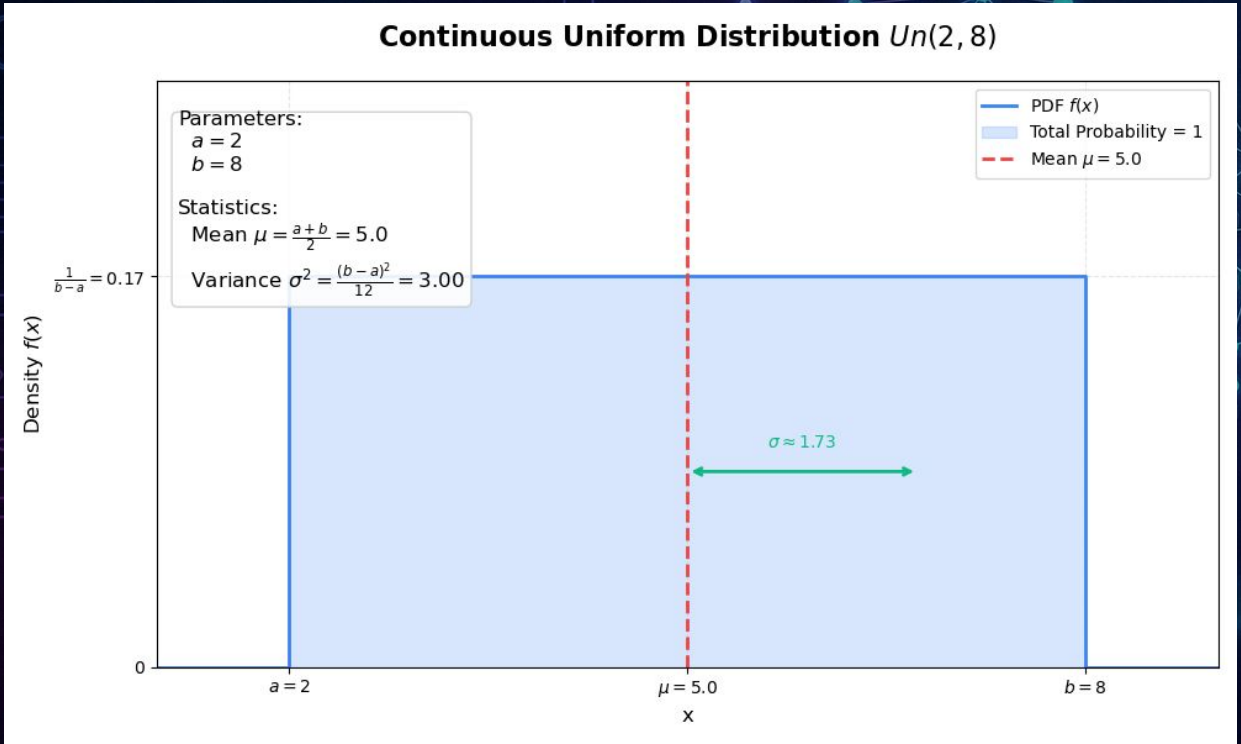
$$\frac{1}{b-a}\cdot\frac{(b-a)(b+a)}{2} = \frac{a+b}{2}$$

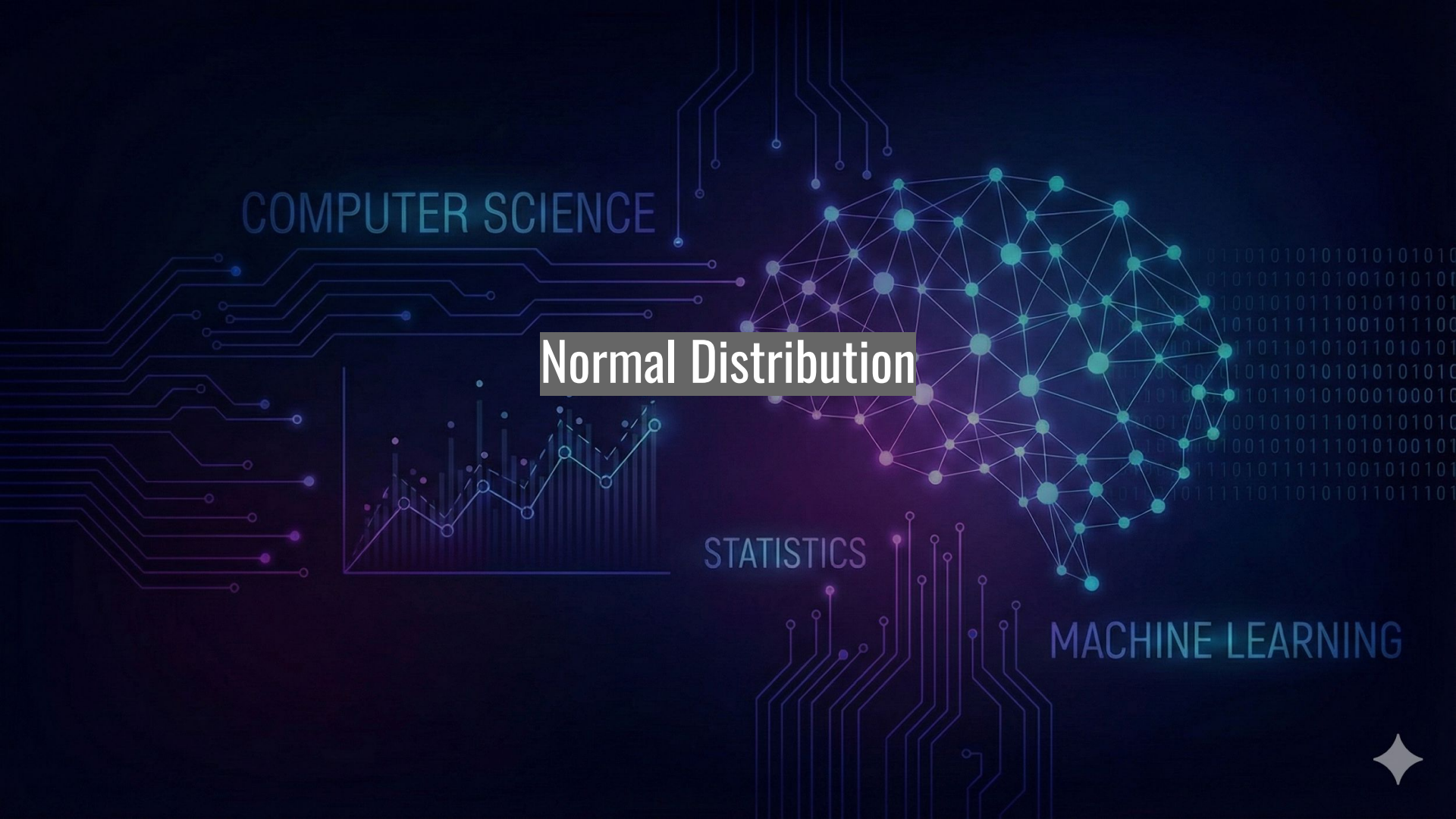
Uniform distribution (continued)

Variance: The spread depends on the length of the interval. Note the specific constant (12) in the denominator.

$$Var[X] = \frac{(b-a)^2}{12}$$

Uniform distribution (continued)




COMPUTER SCIENCE
Normal Distribution
STATISTICS
MACHINE LEARNING

Introduction to the Normal Distribution

The Most Important Distribution in Statistics

- The Normal Distribution (or Gaussian Distribution) is a continuous distribution defined for all real numbers.
- Real-World Examples: Many natural phenomena follow this model, such as rainfall amounts or various biological measurements.
- Approximation Tool: It is widely used to approximate other probability distributions.
- Key Characteristic: It is defined by its density function, which creates a classic "bell-shaped" curve.

Mathematical Definition

Parameters and Probability Density Function (PDF)

- A random variable X follows a Normal Distribution with parameters μ and σ^2

$$X \sim N(\mu, \sigma^2)$$

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Mathematical Definition

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Mean (E[X]): Represented by μ, it determines the center of the distribution.

Variance (Var[X]): Represented by σ^2, it determines the spread of the distribution.

Visualizing μ and σ

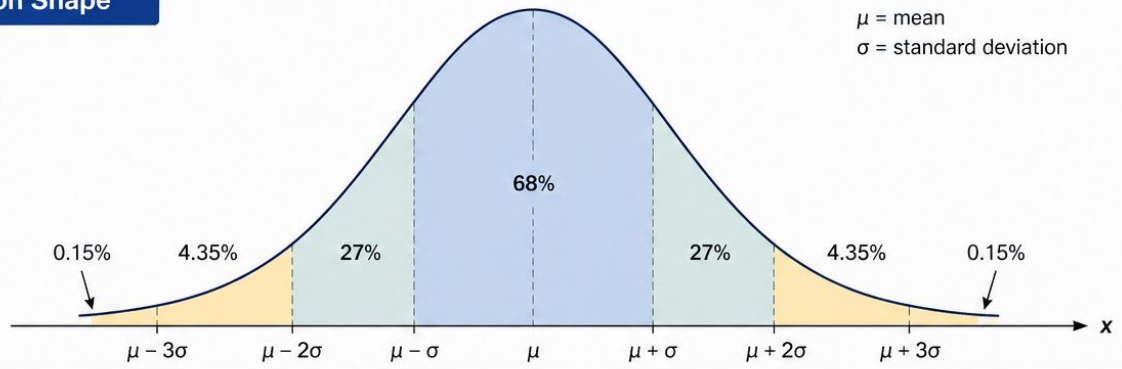
How Parameters Change the Curve

- Changing μ:
  - Changing the mean shifts (translates) the bell curve along the x-axis without changing its shape.
- Changing σ:
  - Small σ: The curve is tall and concentrated near the mean.
  - Large σ: The curve is short and more dispersed (spread out).
- Symmetry: All normal curves are symmetric around the line x = μ.

Normal Distribution: Shape and How It Changes

1. The Normal Distribution Shape

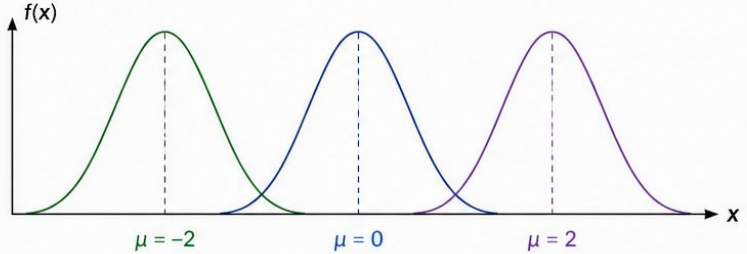
- The normal distribution is symmetric and bell-shaped.
- Mean (μ) is at the center.
- The total area under the curve is **1** (100%).
- About **68%, 95%, 99.7%** of data lie within **1σ, 2σ, 3σ** from the mean.



2. How It Changes

A. Changing the Mean (μ)

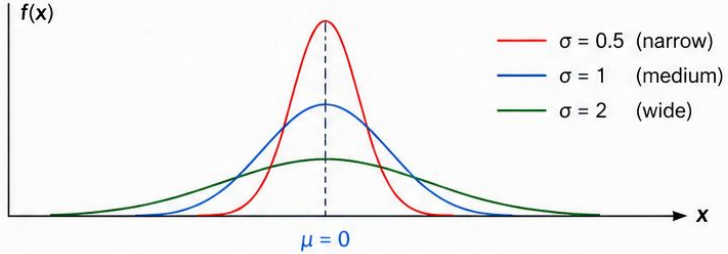
Changing μ shifts the curve left or right.
The shape and spread stay the same.



Same σ (spread), different μ (location)

B. Changing the Standard Deviation (σ)

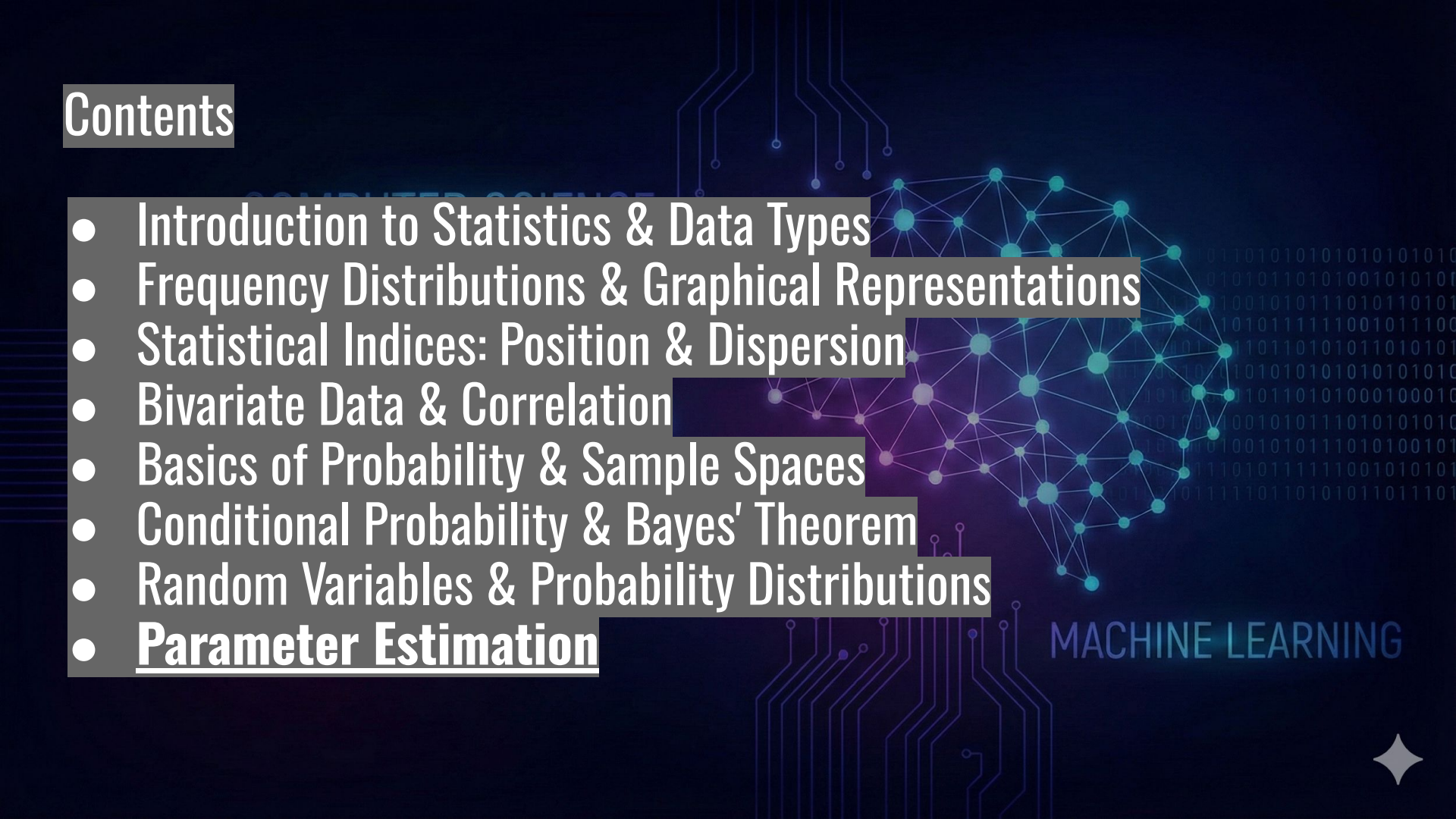
Changing σ changes the spread.
The mean (center) stays the same.

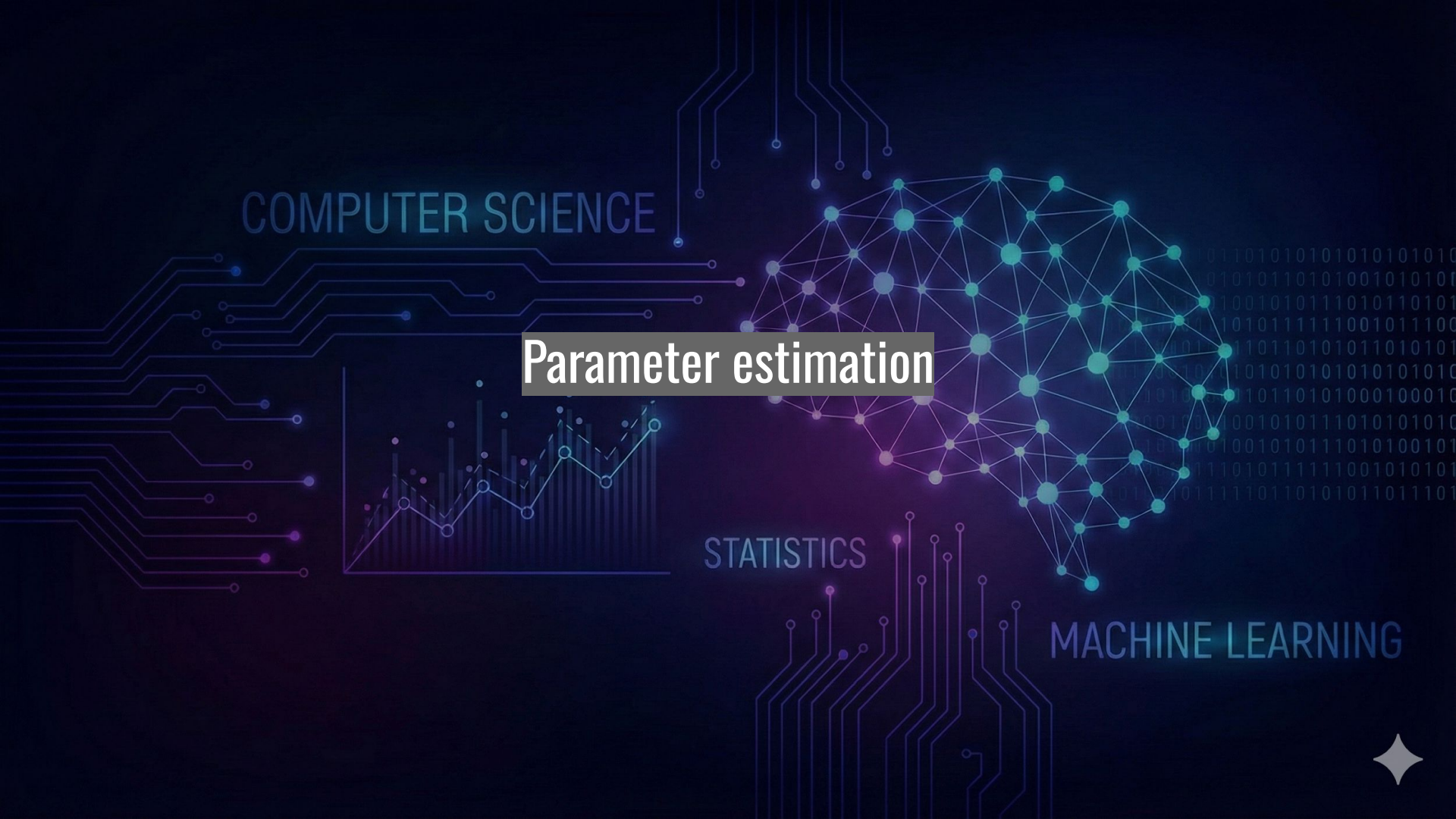


Same μ (center), different σ (spread)

Key Takeaway: μ controls the location (left or right). σ controls the spread (narrow or wide). Both determine the normal distribution.

Contents




COMPUTER SCIENCE
Parameter estimation
STATISTICS
MACHINE LEARNING

The Two Phases of Inference

Phase 2: The Deterministic World

After measuring the sample:

The variables collapse into certain, precise numerical values (x_1, x_2, ..., x_n) (e.g., 10.1 mm,9.9 mm). The uncertainty is gone; the numbers are fixed.

Lottery Analogy: Unscratched tickets (potential variance) vs. Scratched tickets (fixed money/trash).

Core Definitions

1. Statistic

- A random variable that is a function exclusively of the random sample.
- It contains no unknown parameters and can be calculated directly from data.

$$T = f(X_1, X_2, \ldots, X_n)$$

Core Definitions

2. Estimator

- A specific statistic chosen with the explicit goal of guessing the value of an unknown parameter ϑ (like μ or σ^2).
- Evaluated on its theoretical long-term properties, not a single lucky guess.

Core Definitions

3. Estimate

- The single numerical value obtained by actually calculating the estimator's formula on the observed data.
- It is a fixed, deterministic number.

$$\hat{\vartheta} = f(x_1, x_2, \ldots, x_n)$$

Evaluating an Estimator

Unbiased (Corretto)

An estimator is unbiased if, "on average", it hits the true parameter value. It has no systematic error.

$$\mathbb{E}[T] = \vartheta$$

Example: $\overline{X}_n$

The sample mean is always an unbiased estimator of the population mean μ.

Evaluating an Estimator

Consistent (Consistente)

An estimator is consistent if the estimate improves and converges to the true parameter as the sample size n approaches infinity

The Variance Problem

The Biased Approach

If we simply divide by n, our estimator T_1 systematically underestimates the true variance σ^2.

$$T_1 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \overline{X}_n)^2$$

$$\mathbb{E}[T_1] = \frac{n-1}{n} \sigma^2 \neq \sigma^2$$

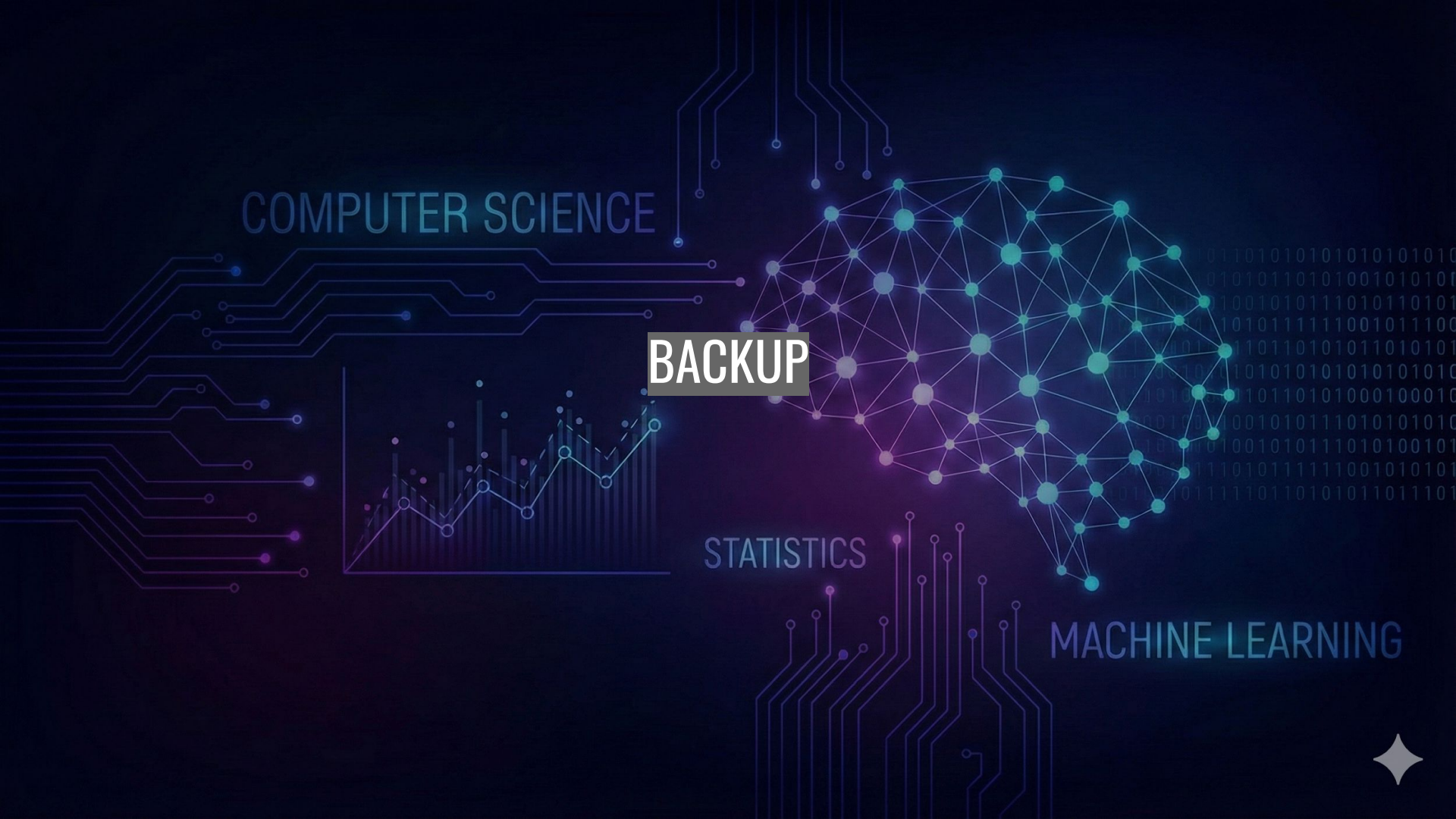
The Variance Problem

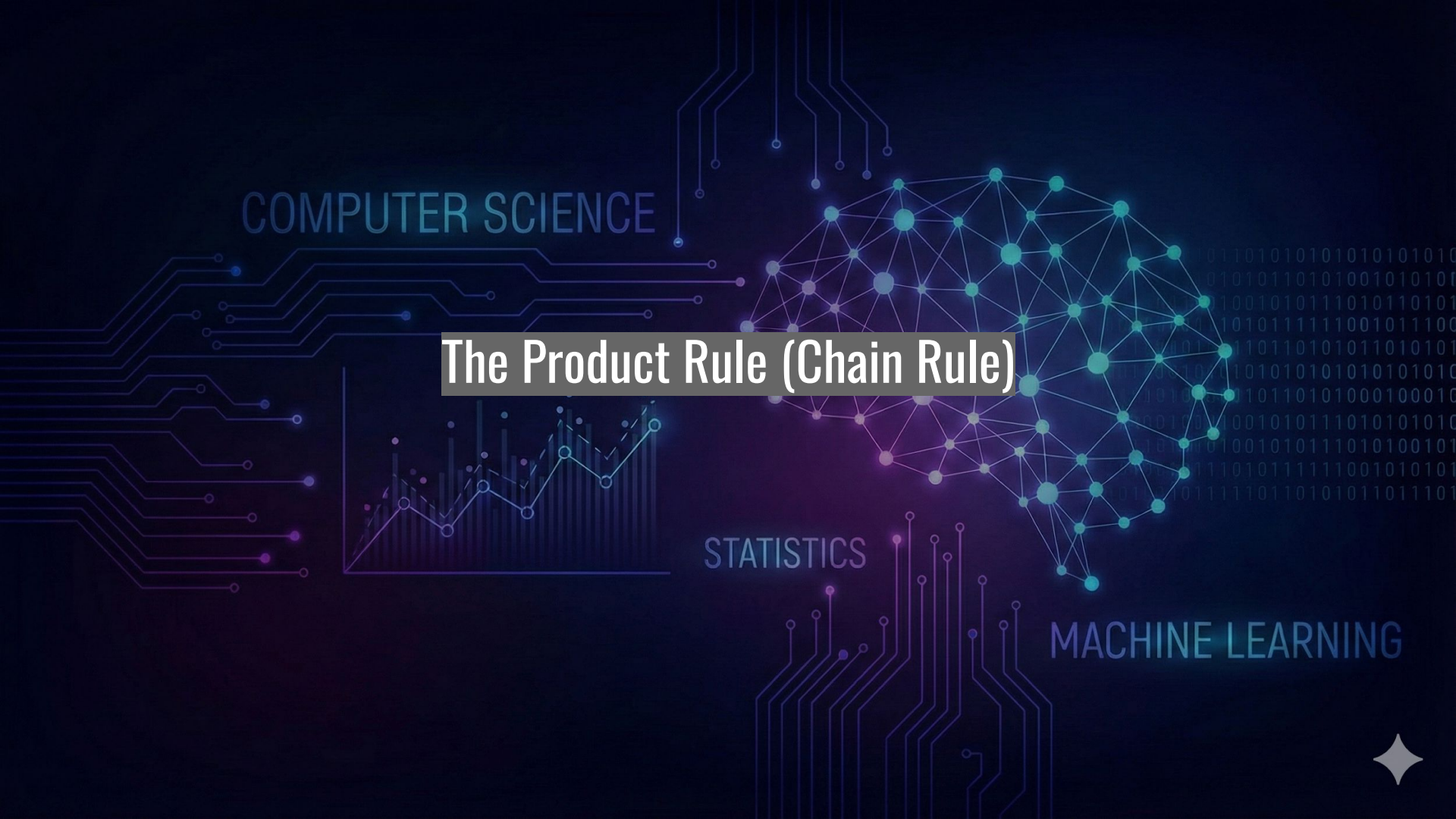
The Unbiased Correction

To fix this systematic error, we divide by n-1. This gives us the corrected sample variance S_n^2.

$$S_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X}_n)^2$$

$$\mathbb{E}[S_n^2] = \sigma^2$$


COMPUTER SCIENCE
BACKUP
STATISTICS
MACHINE LEARNING



The Product Rule (Chain Rule)

The Product Rule (Chain Rule)

Calculating Intersections: From the definition of conditional probability, we can derive a rule to find the probability that both events occur.

$$\mathbb{P}(A \cap E) = \mathbb{P}(A|E) \cdot \mathbb{P}(E)$$

Application: This is essential for sequential experiments (e.g., drawing two cards one after another without replacement).

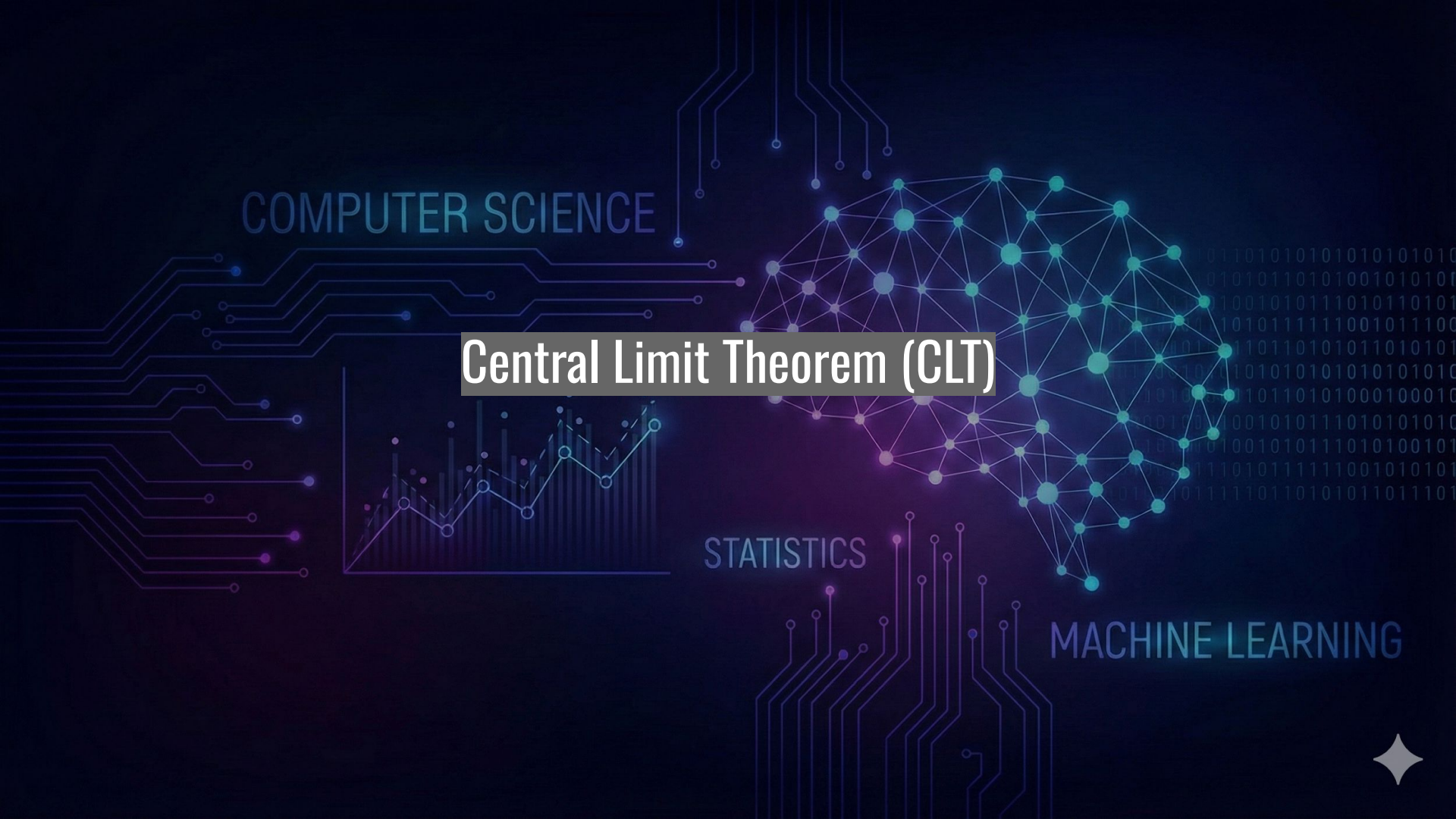
Independence

Definition: Two events are independent if knowing one occurred does not change the probability of the other

$$\mathbb{P}(A|B) = \mathbb{P}(A)$$

$$\text{or } \mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$$

Independent ≠ Disjoint (Mutually Exclusive). Disjoint events are highly dependent (if one happens, the other cannot!).


COMPUTER SCIENCE
Central Limit Theorem (CLT)
STATISTICS
MACHINE LEARNING

Central Limit Theorem (CLT)

The Power of Sums:

- **The sum ($S_n = X_1 + X_2 + ... + X_n$) of a large number of independent and identically distributed (therefore with the same mean μ and the same variance σ^2) random variables (X_i) tends toward a Normal Distribution, regardless of the original distribution.**

Parameters for S_n:

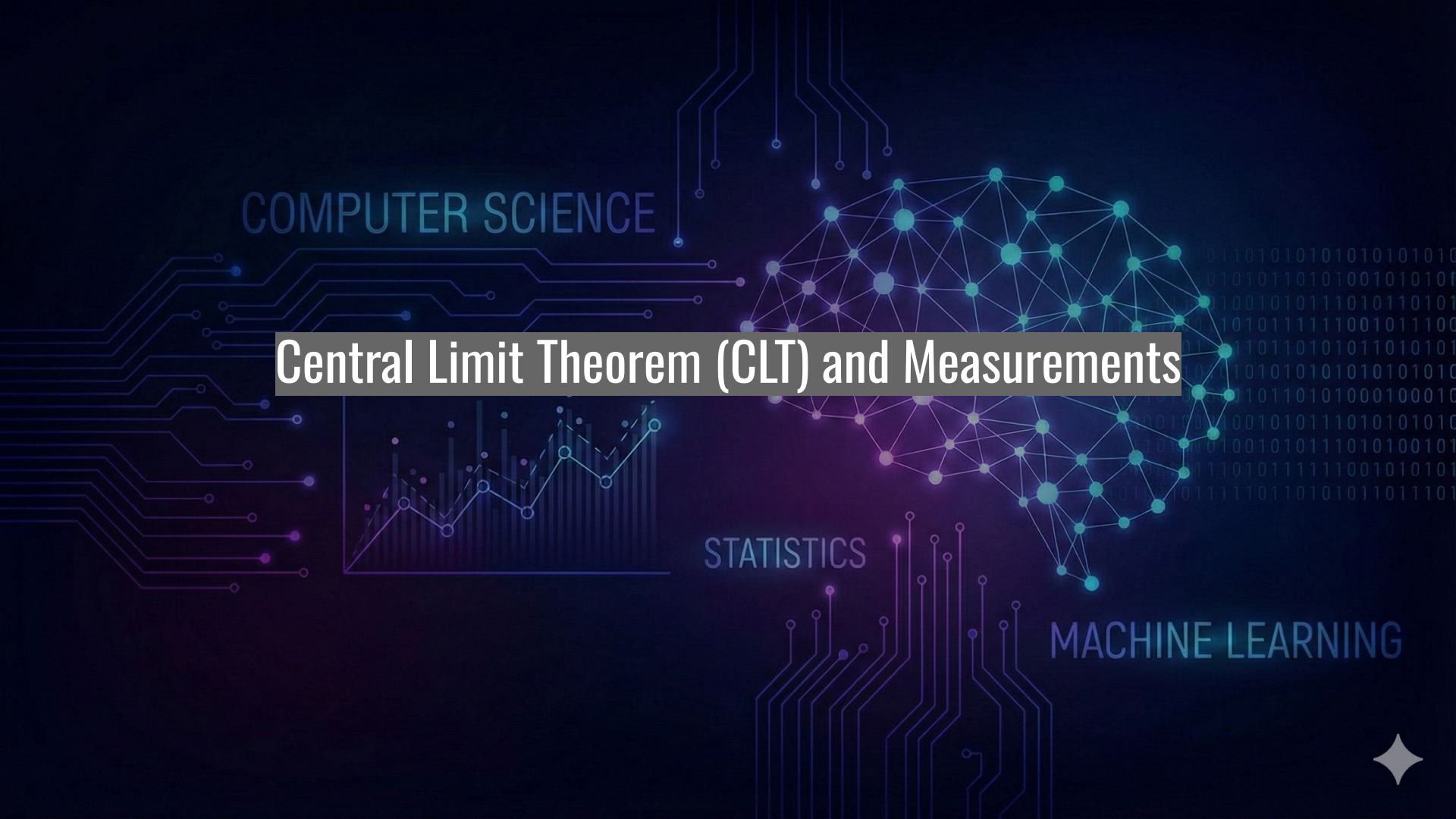
- $E[S_n] = n\mu$
- $Var[S_n] = n\sigma^2$

Central Limit Theorem (CLT)

One of the most frequent uses of the CLT is in "Sampling Theory," which helps us understand how a sample average relates to the true population average.

Sample Mean: The average of n independent observations will behave like a Gaussian variable as n grows large.

Reduced Spread: The variance of the sample mean is σ^2_n, meaning that as you collect more data, your estimate of the average becomes much more precise and "concentrated".


COMPUTER SCIENCE
Central Limit Theorem (CLT) and Measurements
STATISTICS
MACHINE LEARNING

Dealing with Experimental Error

Imagine you are a scientist or engineer trying to measure a physical quantity (like the temperature of a liquid, the length of a component, or the voltage of a battery).

- The True Value (μ): This is the actual value you want to find (e.g., the true temperature).
- The Instrument's Precision (σ): No instrument is perfect. Every time you take a measurement, there is some random error. The "noise" or spread of your instrument is represented by the standard deviation σ

The Single Measurement (X)

If you take just one measurement (X_1), your result is "noisy." It comes from a distribution where the values are scattered around the true mean μ with a variance of σ^2.

- Result: You might get a value that is far from the true target simply due to bad luck (random error).
- When you take a single measurement, you are essentially "picking one number" out of a hat that contains all possible outcomes defined by the probability distribution.

The Single Measurement (X)

Single Measurement (X): You draw from a wide curve $N(\mu,\sigma^2)$. The "target area" is broad, so missing the bullseye is easy

$$\text{Measurement}(X_i) = \text{True Value}(\mu) + \text{Random Error}$$

Many measurements

The Central Limit Theorem (CLT) explicitly states that the sum of a large number of **independent** random variables (measurements) tends to have a Normal distribution, regardless of the distribution of the single measurements.

- $S_n = X_1 + X_2 + ... + X_n$ is nomaly distributed $N(n\mu, n\sigma^2)$
- μ and σ^2 are the same as we are measuring the same thing

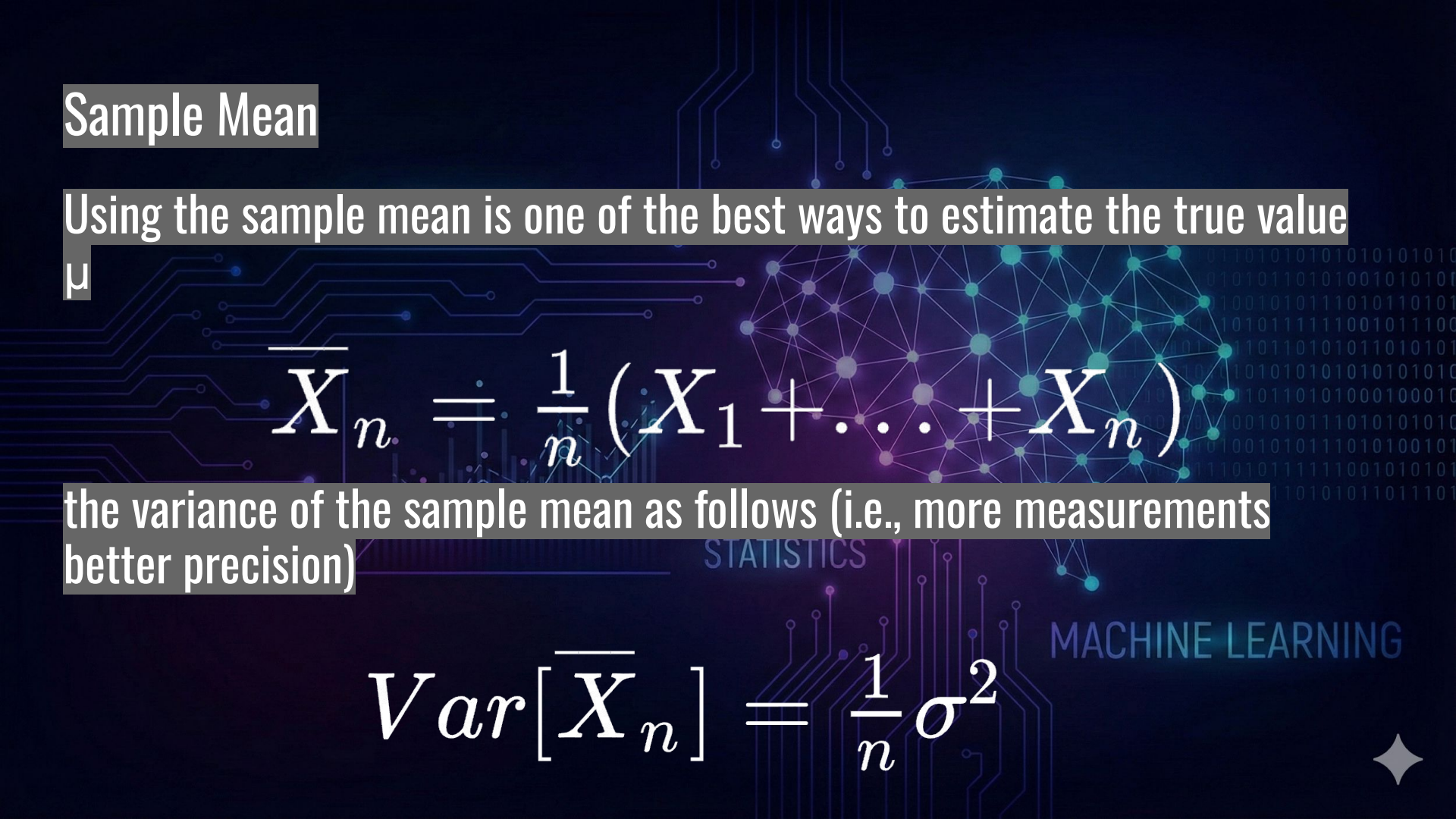
Sample Mean

Using the sample mean is one of the best ways to estimate the true value μ

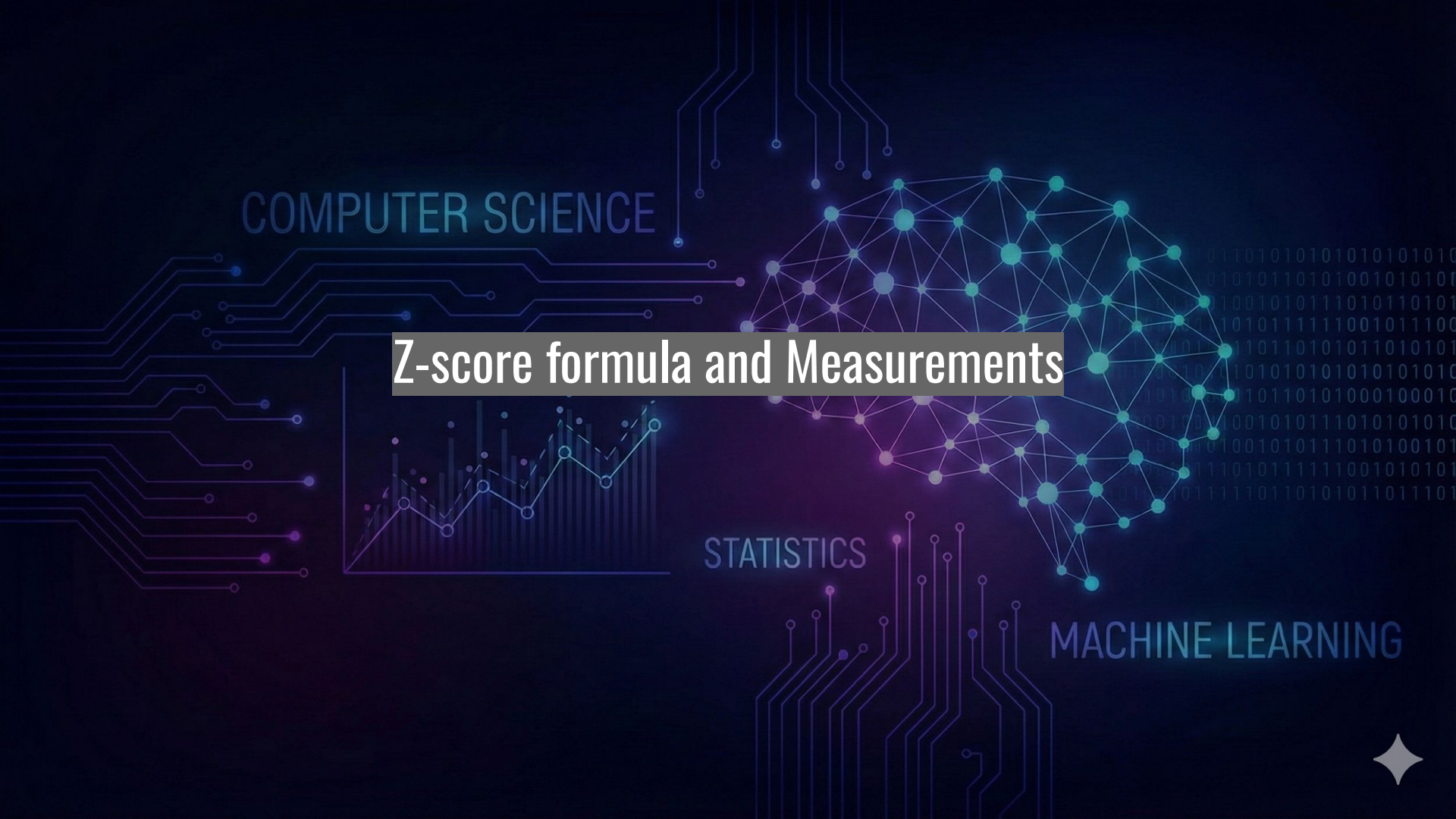
$$\overline{X}_n = \frac{1}{n}(X_1 + \cdots + X_n)$$

the variance of the sample mean as follows (i.e., more measurements better precision)

$$Var[\overline{X}_n] = \frac{1}{n}\sigma^2$$




EXPERIMENT:
Measuring Ball Rolling Time
2.15s
2.30s
2.10s
2.15s
2.10s
Data
DATA COLLECTION:
Individual Measurements (Ramp Time)
Non-Normal Population
Skewed Distribution of Individual Data Points (x)
THE PROCESS:
Sampling & Calculating Means
Sample 1 (n=5)
Sample 2 (n=5)
Sample 3 (n=5)
Sample 4 (n=5)
$\bar{x}_1$
$\bar{x}_2$
$\bar{x}_3$
$\bar{x}_3$
μ
CENTRAL LIMIT THEOREM IN ACTION:
Distribution of Sample Means
Normal (Gaussian) Distribution of Sample Means ($\bar{x}$)
μ
As Sample Size (n) Increases, the Distribution Approaches Normal
Frequency
μ
Sample Mean Value ($\bar{x}$)


COMPUTER SCIENCE
Z-score formula and Measurements
STATISTICS
MACHINE LEARNING

Central Limit Theorem (CLT)

One of the most frequent uses of the CLT is in "Sampling Theory," which helps us understand how a sample average relates to the true population average.

Standardization: This allows researchers to use the Z-score formula to determine if their sample results are statistically significant.

$$Z = \frac{\overline{X}_n - \mu}{\sigma / \sqrt{n}}$$

Z score example

Imagine you bought a batch of electronic components (resistors) that the manufacturer claims have a true mean resistance (μ) of 100 Ω

You suspect the batch might be defective (off-spec). You decide to perform an experiment to check this claim.

Z score example

- The Claim (Target Mean μ): 100 Ω
- The Instrument Error (σ): From the multimeter's manual, you know the standard deviation of measurements is 4 Ω.
- The Experiment: You measure 64 resistors (n=64).
- The Experimental Result: The average of your 64 measurements is 101.2 Ω.

$$\overline{X}_n = 101.2\Omega$$

Z score example

First, we determine how "tight" the distribution of the average should be if the manufacturer is telling the truth. Because we averaged 64 measurements, the variance drops significantly .

$$\text{Standard Error} = \frac{\sigma}{\sqrt{n}} = \frac{4}{\sqrt{64}} = \frac{4}{8} = 0.5\Omega$$

Meaning: If the true mean is really 100, the average of 64 measurements should typically fluctuate by only about 0.5 Ω.

Z score example

Now we calculate the Z-score to see how many "standard errors" away your result is from the claim. We use the standardization formula provided in the text :

$$Z = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} = \frac{101.2 - 100}{0.5} = 2.4$$

Variance & Standard Deviation

Computational Shortcut: Just like with discrete variables, it is easier to calculate using the "Second Moment".

$$Var[X] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$$

$$\sigma_X = \sqrt{Var[X]}$$