Mastering Machine Learning

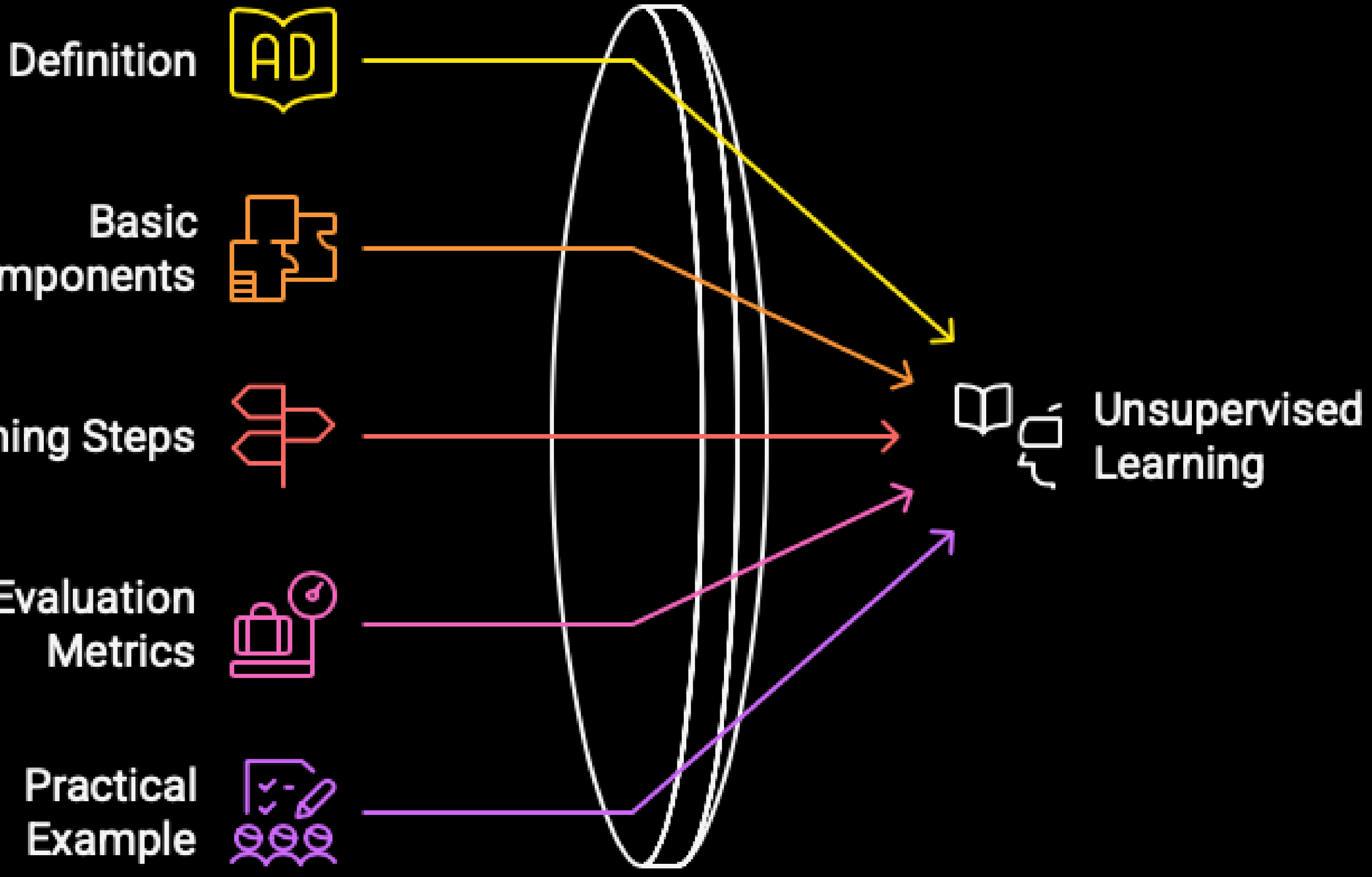
Model Training

A Complete Guide to Building Unsupervised Models

Edgar Rios Linares

Roadmap

Building Blocks of Unsupervised Learning



Definition

The model works with data that does not have predefined labels or outcomes.

Discover patterns in the data.

Definition

The model works with data that does not have labels or predefined results.

Does not require labeled data (lower cost).

Useful when the underlying structure of the data is not known.

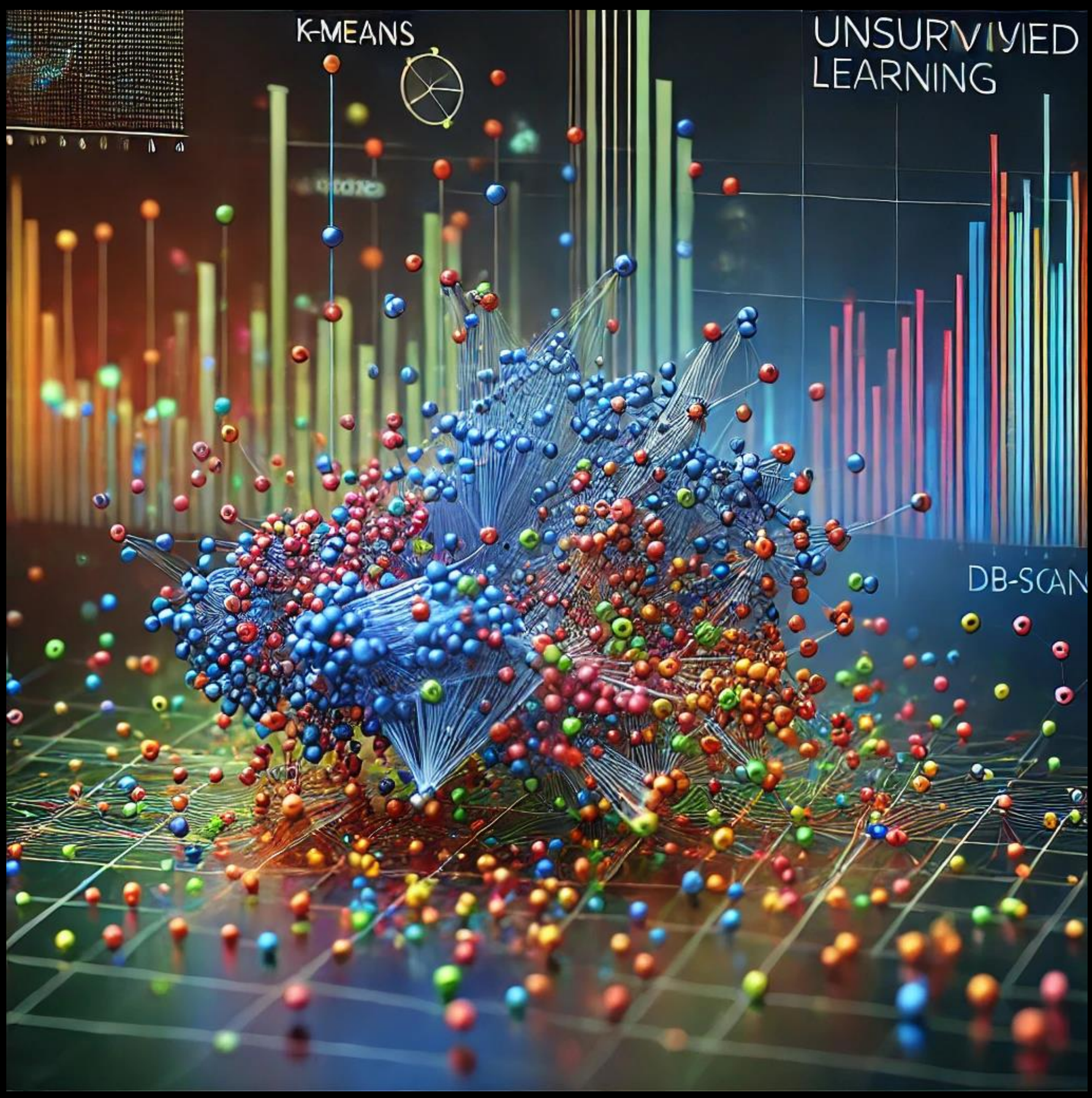
Ideal for exploring large volumes of data.

Components

Unlabeled dataset

Features

Objective function



Components

Unlabeled Dataset :

The data is "raw" and has no explicit target variable.

Features :

Variables used to train the model.

Objective Function :

Defines how the model should organize the data.

Training Steps

Preprocessing

Cleaning and normalizing data

Algorithm Selection

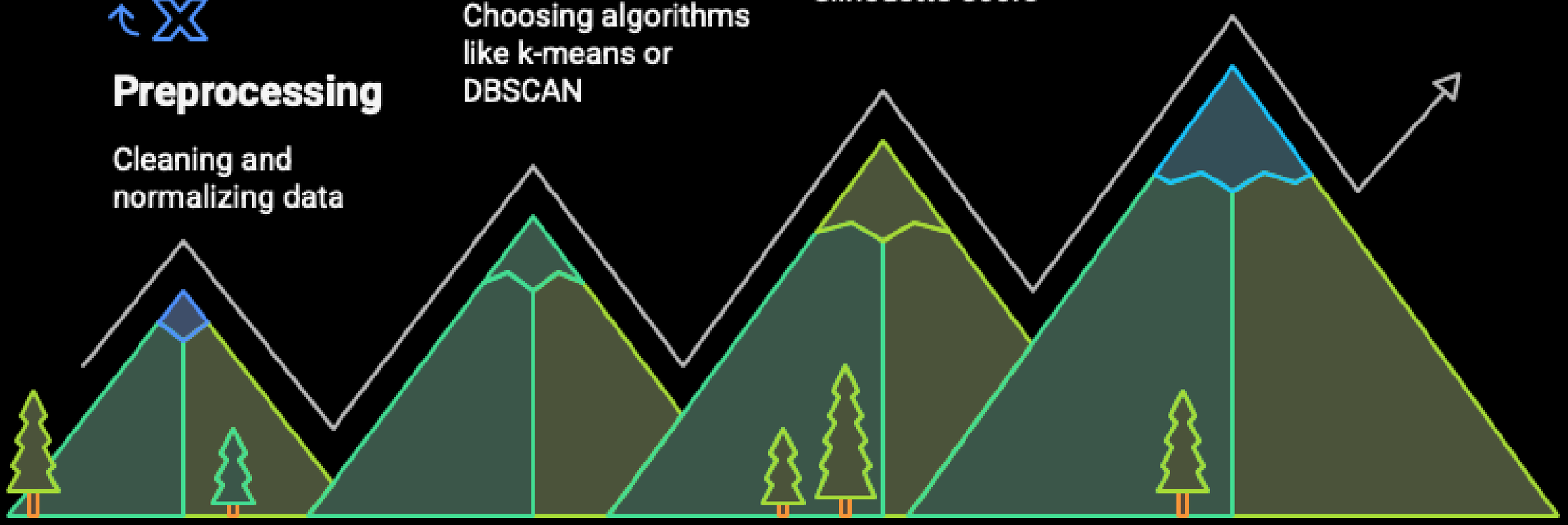
Choosing algorithms like k-means or DBSCAN

Parameter Optimization

Applying the Elbow Method and Silhouette Score

Performance Evaluation

Evaluating with Silhouette and Dunn's Coefficient



Training Steps

1. Preprocessing

Data Cleaning and Normalization

2. Model Selection

k-means, DBSCAN, PCA, Autoencoders

3. Parameter Optimization

Elbow Method, Silhouette Score

4. Performance Evaluation

Silhouette, intra-cluster and inter-cluster, Dunn's coefficient

Training Steps

1. Preprocessing

Outliers

Critical for identifying data anomalies.

Duplicates

Essential for ensuring data integrity.

Normalization with Logarithm

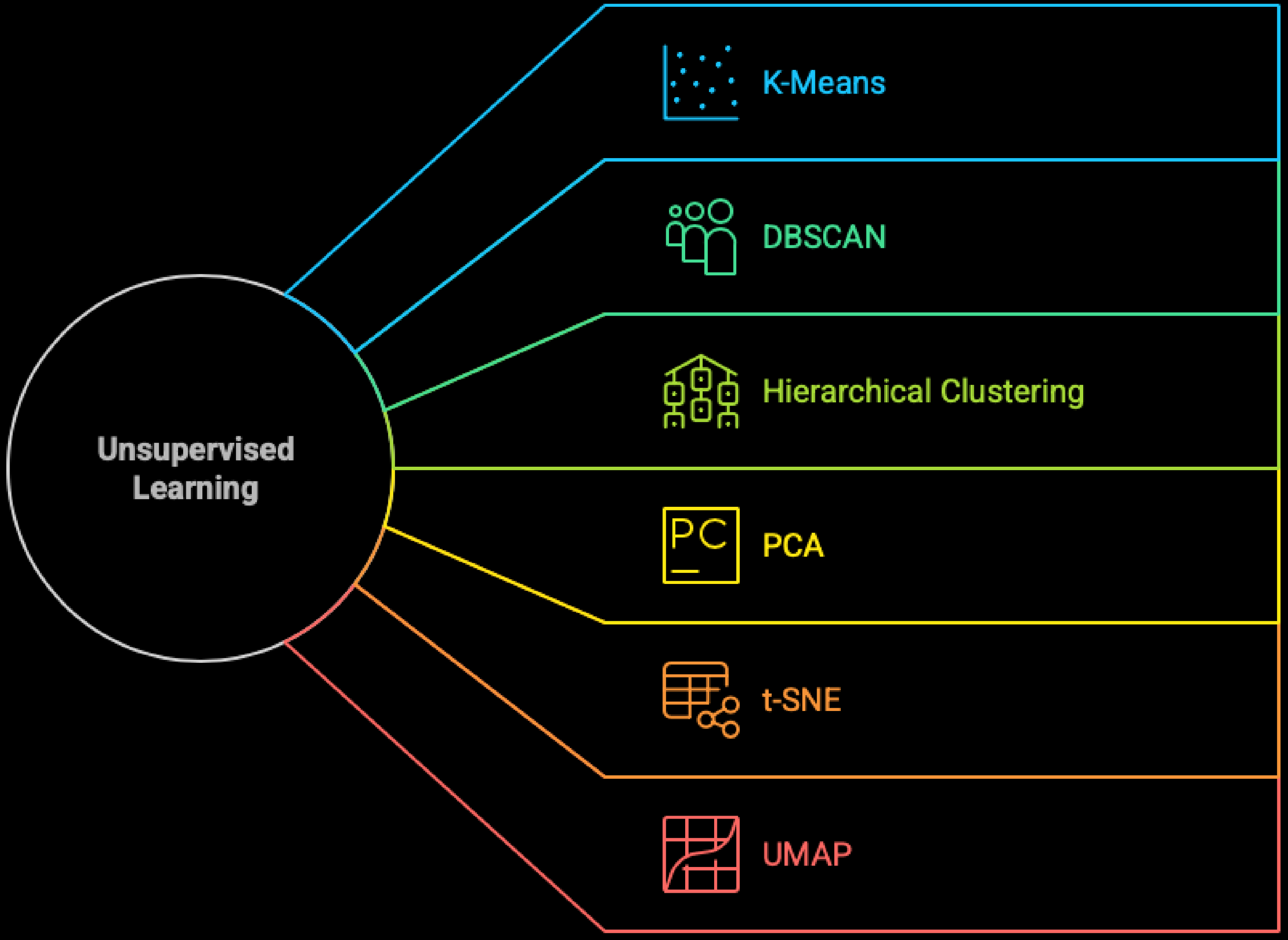
Advanced technique for transforming data distributions.

Min-Max Normalization

Simple method for scaling data to a specific range.

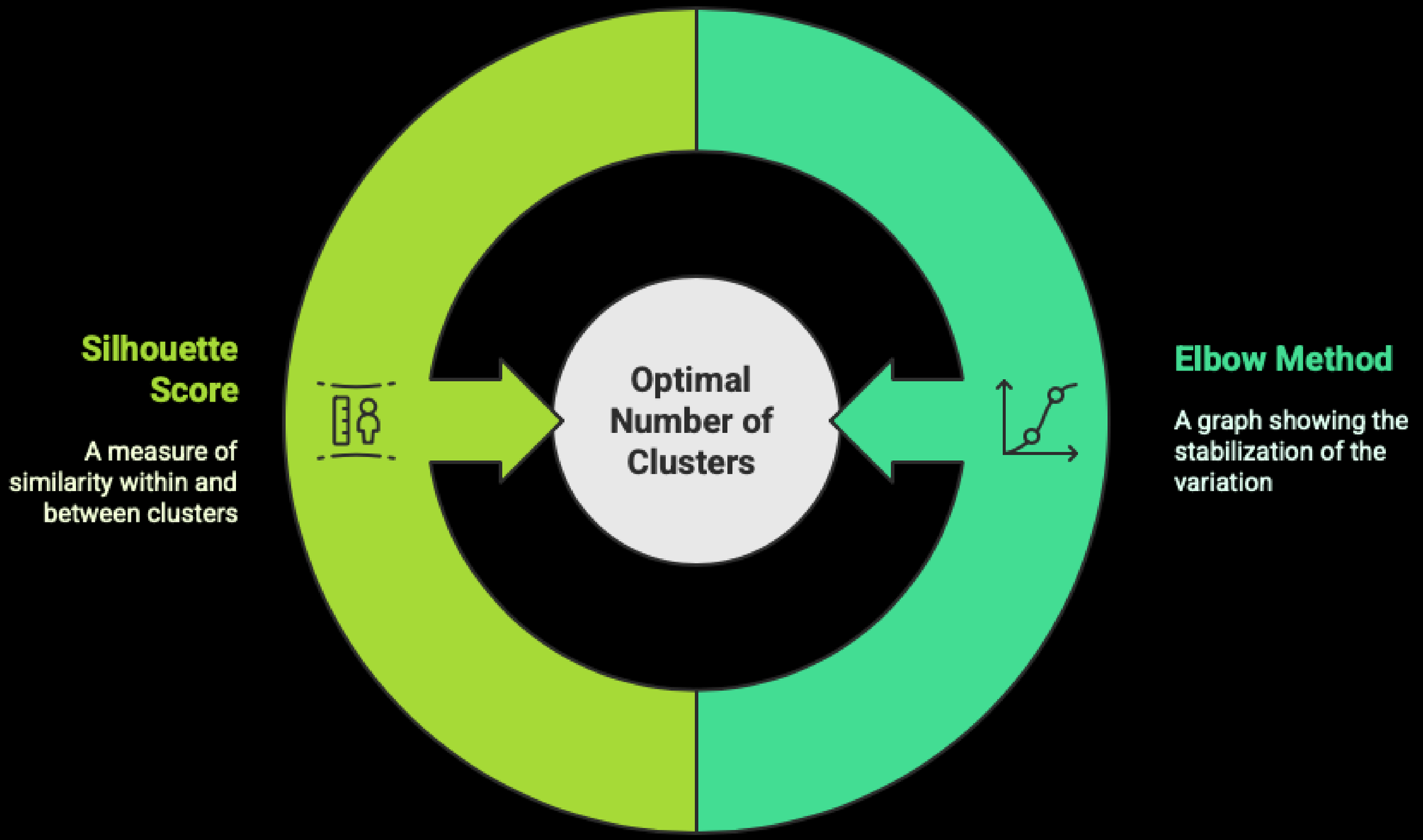
Training Steps

2. Model Selection



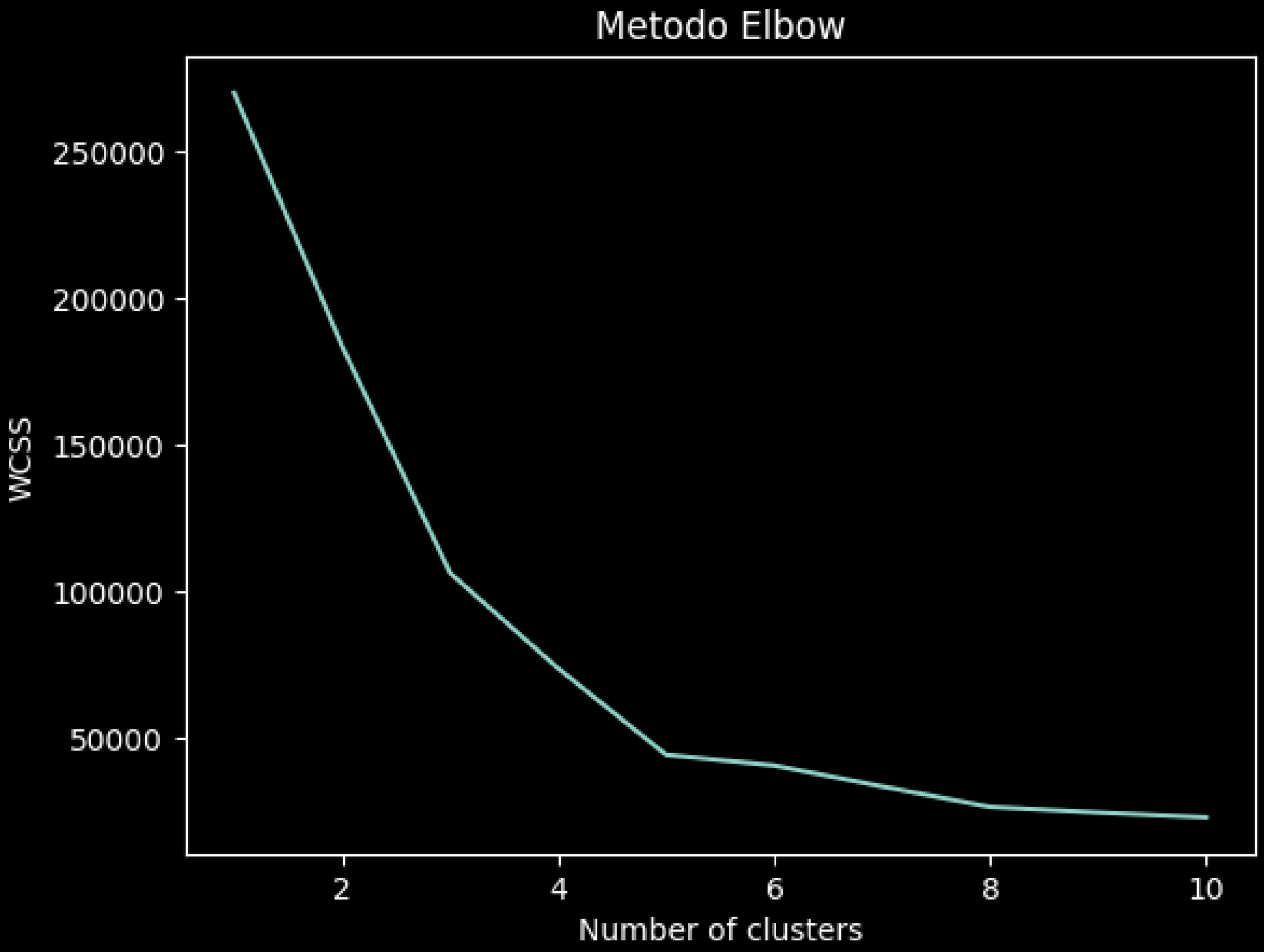
Training Steps

3. Parameter Optimization



Training Steps

3. Parameter Optimization

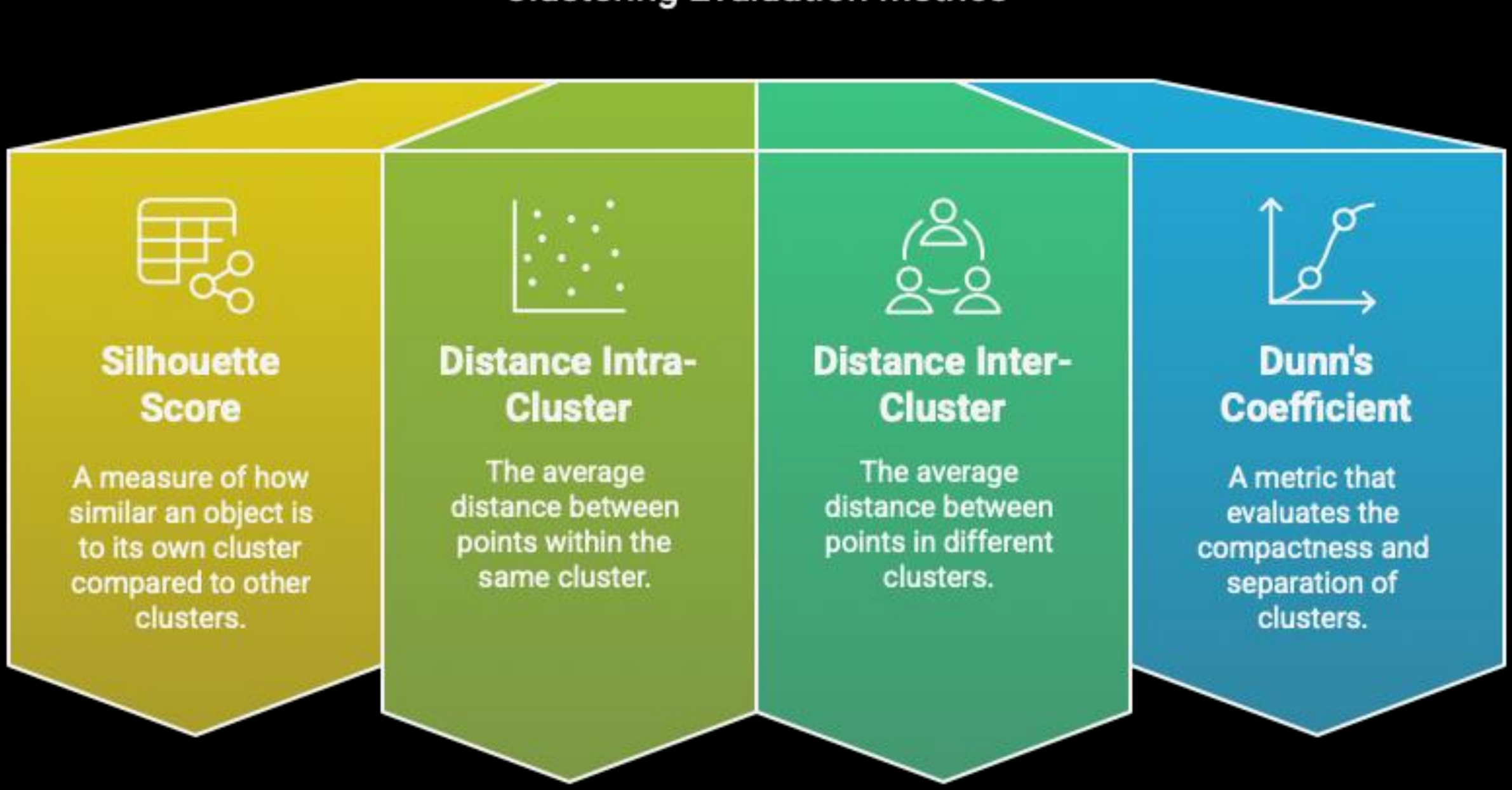


Training Steps

4. Performance Evaluation

Metric	Objective	Interpretation
Silhouette Score	How well the clusters are separated	≈1 is good, ≈0 is confusing, <0 is bad
Distance Intra-Cluster	Cohesion within a cluster	Lower values are better compaction
Distance Inter-Cluster	Separation between clusters	Higher values are better separation
Dunn's coefficient	Relationship between separation and cohesion	Higher values are better clustering

Evaluation Metrics



Practical Example

Segment customers in a bank

1.Preprocessing

Read dataset and Normalization

2. Model Selection

Select variables, use k-means model

3.Parameter Optimization

Review Elbow Method

4. Performance Evaluation

Silhouette, intra-cluster and inter-cluster, Dunn's coefficient

Practical Example

STEP 1

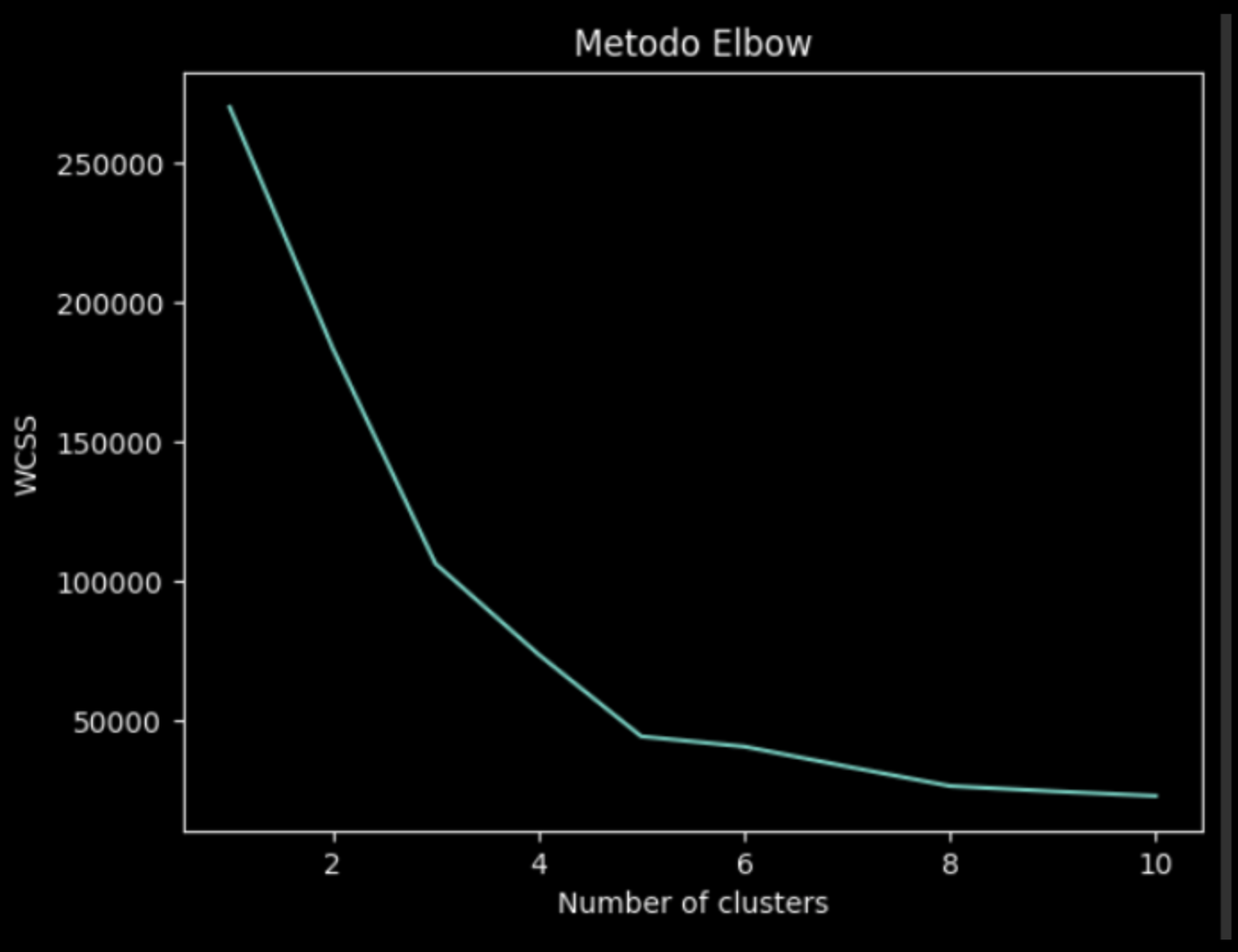
	CustomerID	Genre	Age	Annual Income (k$)	Spending Score (1-100)
0	1	Male	19	15	39
1	2	Male	21	15	81
2	3	Female	20	16	6
3	4	Female	23	16	77

STEP 2

```
[20] kmeans
```

```
KMeans(n_clusters=3, random_state=42)
```

STEP 3



STEP 4

```python
from sklearn.metrics import silhouette_score

silhouette_avg = silhouette_score(X, y_kmeans)
print(f"Silhouette Score: {silhouette_avg:.3f}")
```

```
Silhouette Score: 0.468
```

Educator in AI

Artificial Intelligence

Data Engineering

Machine Learning

Data Science

📍Linkedin —> https://www.linkedin.com/in/erlinares/

👋 Follow us on 𝕏: https://x.com/erlinares

💻 GitHub: https://github.com/erlinares/365_AI_Journey/

💬 Discord: https://discord.gg/5fFM2zh8

Edgar Rios Linares