CLUSTERING IN MACHINE LEARNING

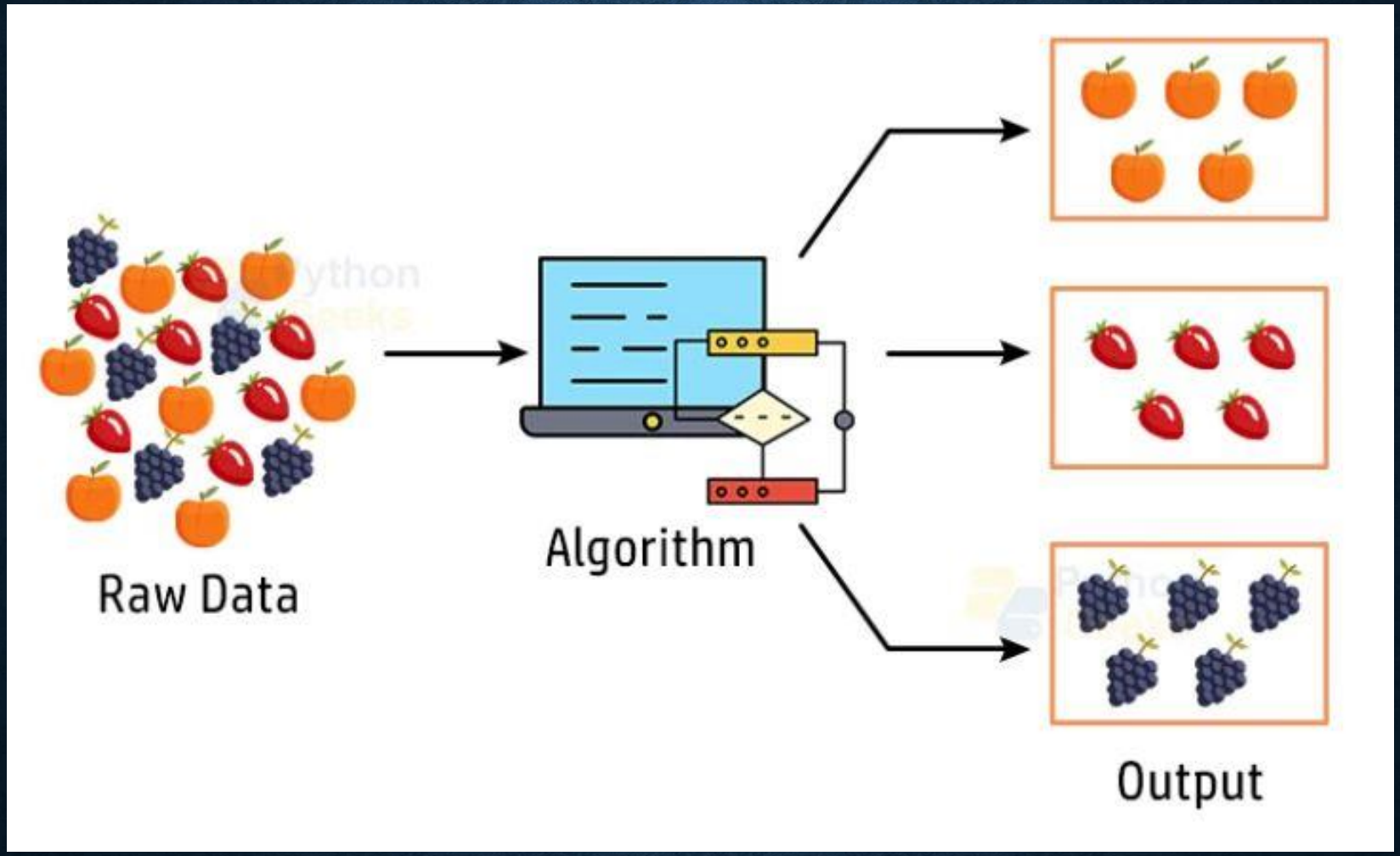
Definition, Types, and Examples

WHAT IS CLUSTERING?

- Clustering is an unsupervised machine learning technique used to group similar data points together into clusters based on similarity.

Goal: Points in the same cluster are more similar to each other than to points in other clusters.

Applications: Customer segmentation, anomaly detection, image compression, bioinformatics.



CLUSTERING EXAMPLE

PROTOTYPE-BASED CLUSTERING

- - Uses a prototype (center/mean) to represent each cluster.
- - Data points assigned to nearest cluster prototype.

- Example: K-Means Clustering
- - Chooses k cluster centers (means).
- - Iteratively assigns points and updates centers.
- - Works well for spherical, well-separated clusters.

HIERARCHICAL CLUSTERING

- - Builds a hierarchy (tree) of clusters.
- - Does not need number of clusters beforehand.

- Two main approaches:

1. Agglomerative (Bottom-Up): Start with single points, merge clusters.

2. Divisive (Top-Down): Start with one cluster, split into smaller clusters.

- - Results shown in a dendrogram.

DENSITY-BASED CLUSTERING

- - Groups data based on areas of high density separated by low-density regions.
- - Can detect clusters of arbitrary shapes and identify noise/outliers.

- Example: DBSCAN
- - Expands clusters from dense regions.
- - Parameters: eps (neighborhood radius), minPts (minimum points).
- - Good for non-linear shapes (e.g., spirals).

SUMMARY TABLE

Clustering Type	Example Method	Key Idea	Pros	Cons
Prototype-Based	K-Means	Assign points to nearest cluster mean	Simple, efficient	Needs k, struggles with irregular shapes
Hierarchical	Agglomerative / Divisive	Builds a tree of clusters	No need for k, dendrogram insight	Computationally expensive
Density-Based	DBSCAN	Groups dense regions, marks noise	Handles arbitrary shapes, finds outliers	Struggles with varying density