Session 4
DBSCAN

Blue AI Team

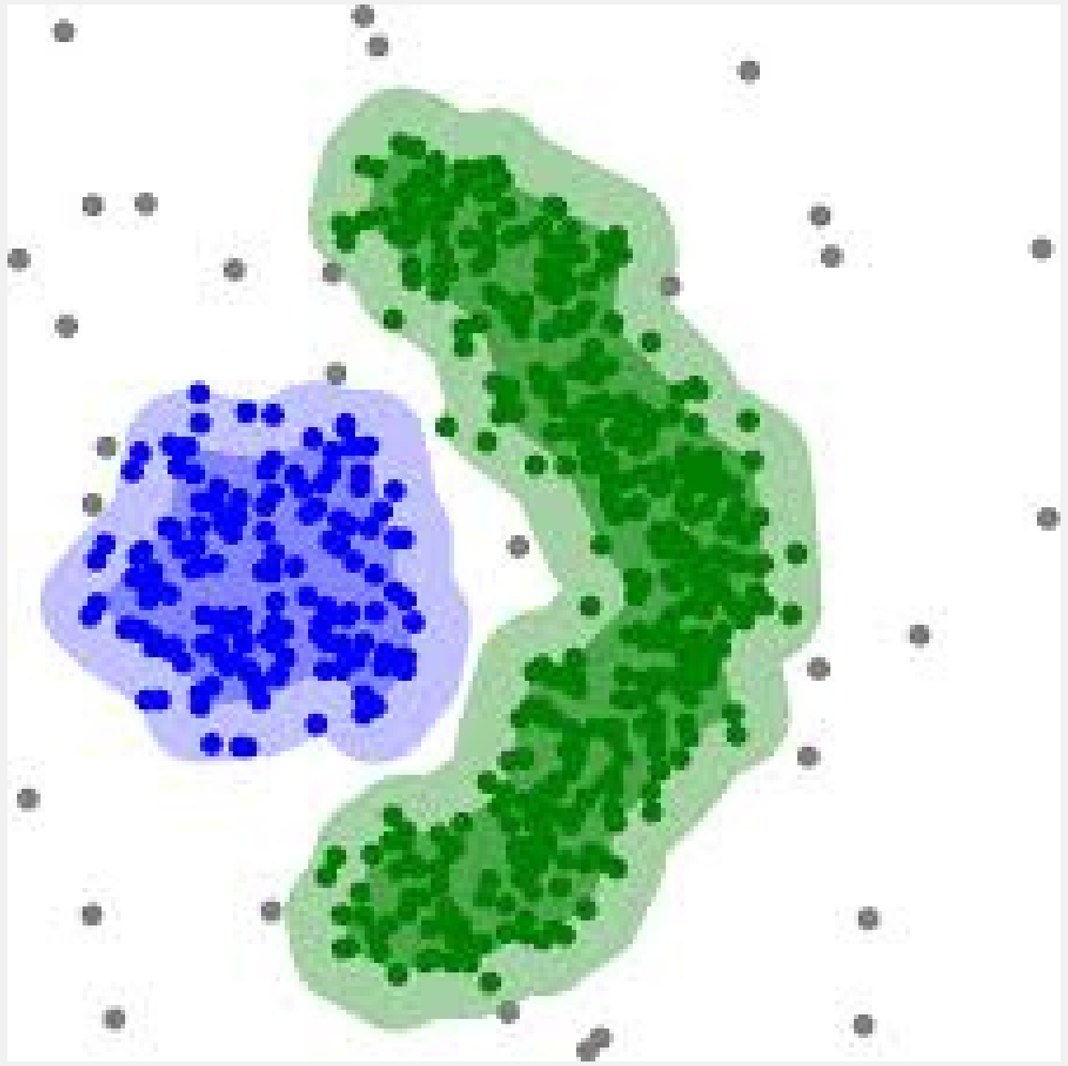
Collecting Samples

1. Simple Random
2. Systematic
3. Stratified
4. Cluster

- **Cluster Sampling** is a data collection technique.
- **Machine learning clustering** is a pattern recognition technique.

What is DBSCAN?

- A density-based clustering algorithm.
- Clusters data points that are closely packed together
- Marks outliers as noise.
- automatically finds clusters based on density.



How Does DBSCAN Work?

DBSCAN works by categorizing data points into three types:

1️⃣ Core Points

✅ A core point has at least MinPts neighbors within a specified radius ε (epsilon).

✅ It forms the center of a cluster and helps expand it.

How Does DBSCAN Work?

1️⃣ Border Points

✅ A border point is close to a core point but does not have enough neighbors to be a core itself.

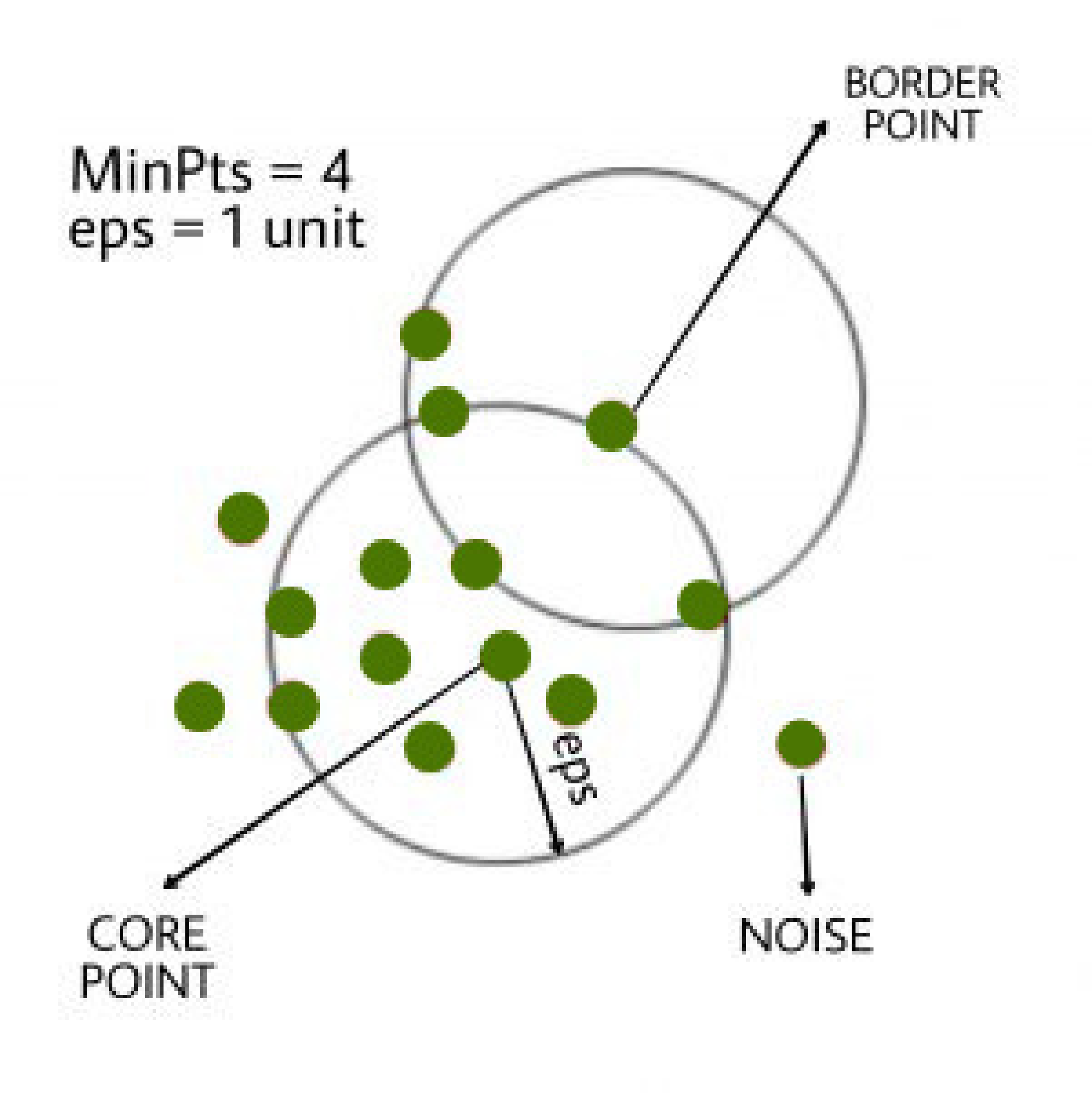
✅ It belongs to a cluster but does not expand it.

How Does DBSCAN Work?

1️⃣ Noise (Outlier) Points

- ✅ A noise point is isolated with too few neighbors within ε.
- ✅ It does not belong to any cluster.

How Does DBSCAN Work?



1. Pick a random unvisited point and check how many neighbors are within ε.
2. If MinPts are found, mark it as a core point and start forming a cluster.
3. Expand the cluster by adding all reachable border points.
4. If a point doesn’t meet the density threshold, it’s labeled noise.
5. Repeat until all points are visited and categorized.

Key Parameters in DBSCAN

1) **eps**: This defines the radius of the neighborhood around a data point.

2) **MinPts**: This is the minimum number of points required within the **eps** radius to form a dense region.

How to set your parameters?

Choosing Epsilon (eps)

If the distance between two points is less than or equal to **eps**, they are considered neighbors. Choosing the right **eps** is crucial:

If **eps** is too small, most points will be classified as noise.

Choosing Epsilon (eps)

If **eps** is too large, clusters may merge, and the algorithm may fail to distinguish between them.

Choosing Epsilon (eps)

Keep it balanced :)

Choosing MinPts in DBSCAN

MinPts (Minimum Points) determines how many neighboring points are needed to form a core point and start a cluster.

How to Set MinPts?

A general rule of thumb:

- MinPts ≥ D + 1, where D is the number of dimensions in the dataset.
- In most cases, a minimum value of MinPts = 3 is recommended.

Choosing MinPts in DBSCAN

Why is MinPts Important?

✅ **Too low MinPts**: Can form small, meaningless clusters and misclassify noise.

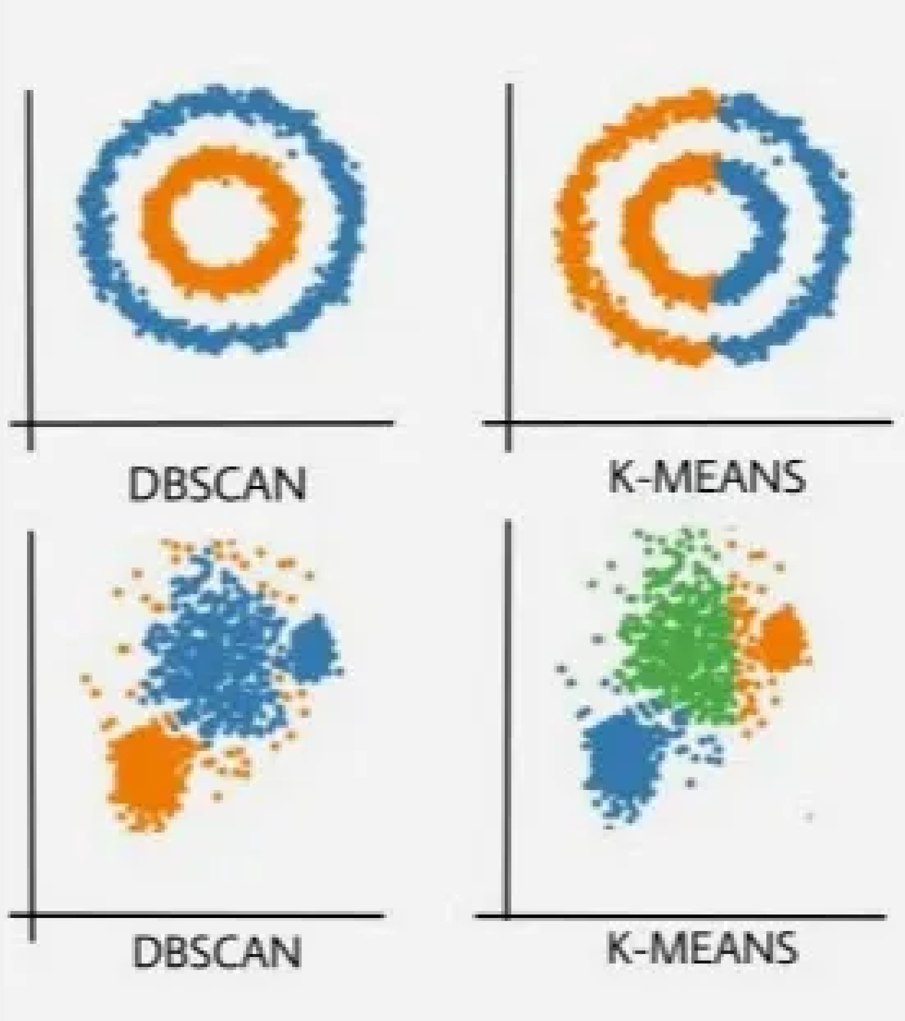
✅ **Too high MinPts**: Can miss smaller clusters and label important points as noise.

When to Use DBSCAN?

DBSCAN is the best choice when dealing with **complex datasets** that traditional clustering methods struggle with.

Use DBSCAN When:

✅ Clusters are irregularly shaped: Unlike K-Means, which assumes circular clusters, DBSCAN detects arbitrary shapes.



When to Use DBSCAN?

Use DBSCAN When:

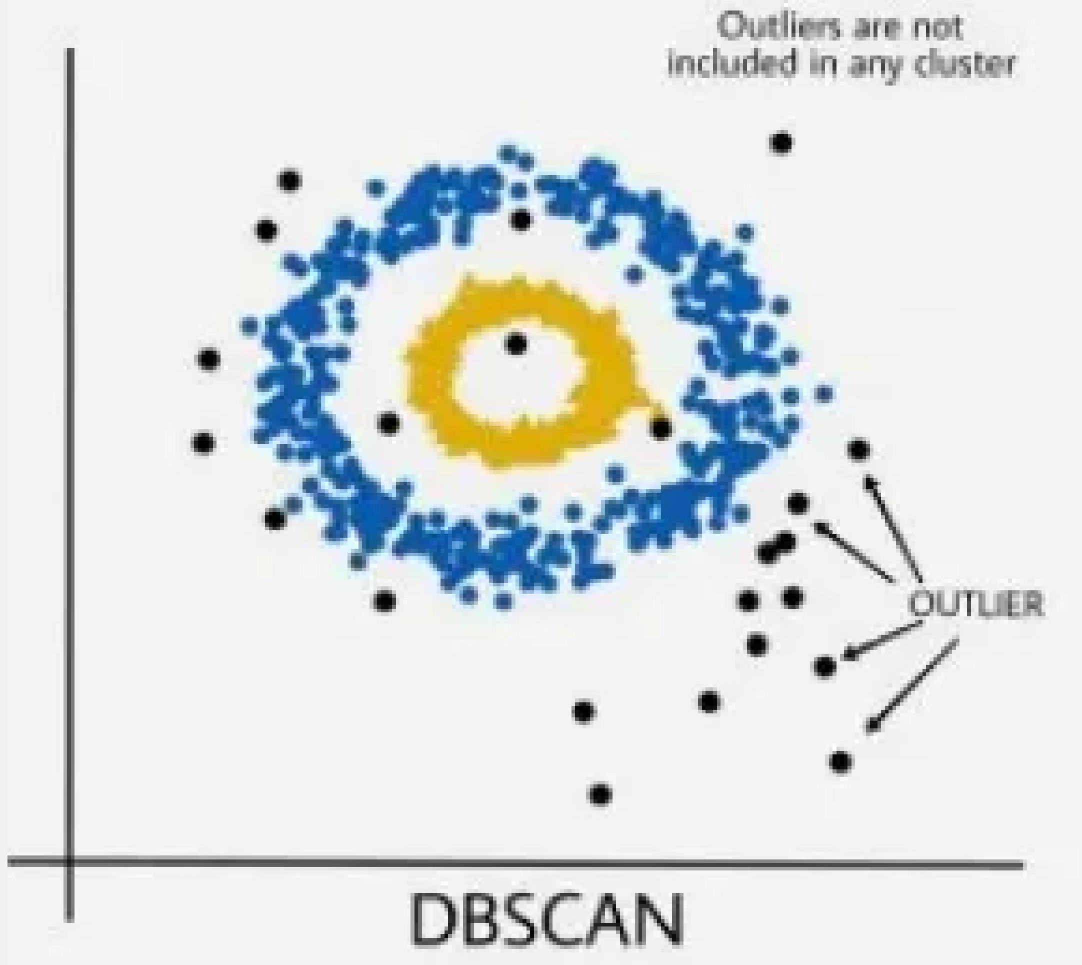
✅ **Clusters have different densities:** Unlike Hierarchical Clustering, DBSCAN can adapt to varying densities without merging everything into one big cluster.

✅ **You don’t know the number of clusters:** K-Means requires predefining K, but DBSCAN discovers clusters naturally.

When to Use DBSCAN?

Use DBSCAN When:

✅ **Clusters have different densities:** Unlike Hierarchical Clustering, DBSCAN can adapt to varying densities without merging everything into one big cluster.



When to Use DBSCAN?

❌ Avoid DBSCAN When:

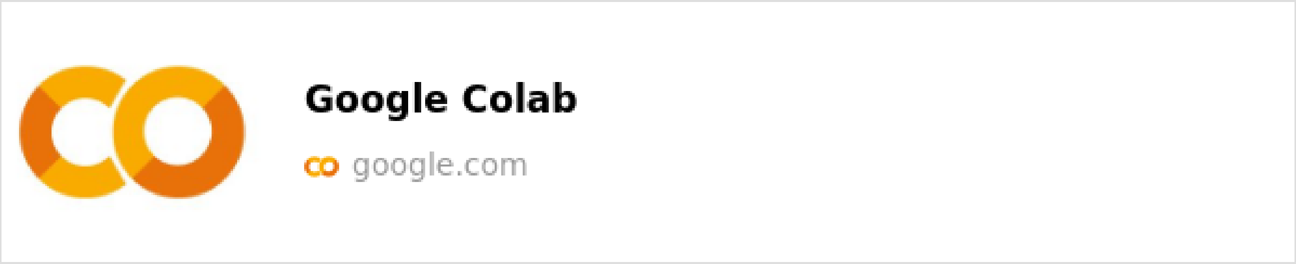
🚫 Clusters have similar densities and well-separated → K-Means is better.

🚫 **Dataset is too large:** DBSCAN has higher computational complexity than K-Means.

🚫 **You have high-dimensional data:** DBSCAN struggles because distances become meaningless.

- 🔹 Solution? Use **PCA** to reduce dimensions before applying DBSCAN.

Code for DBSCAN

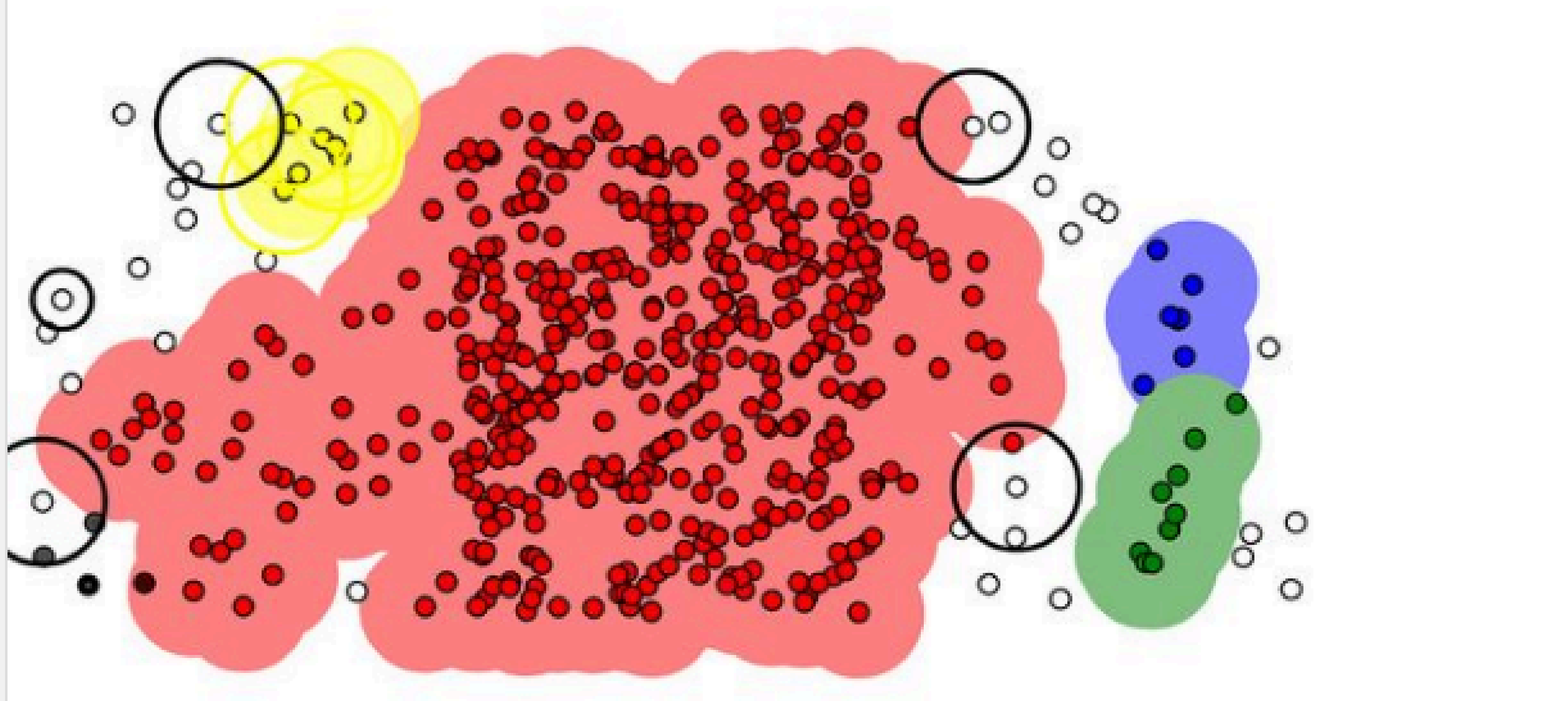


Google Colab

google.com

https://colab.research.google.com/drive/1wrEFL9mVXu_lRdD3zrD58LqBKzAb7HB7?usp=sharing

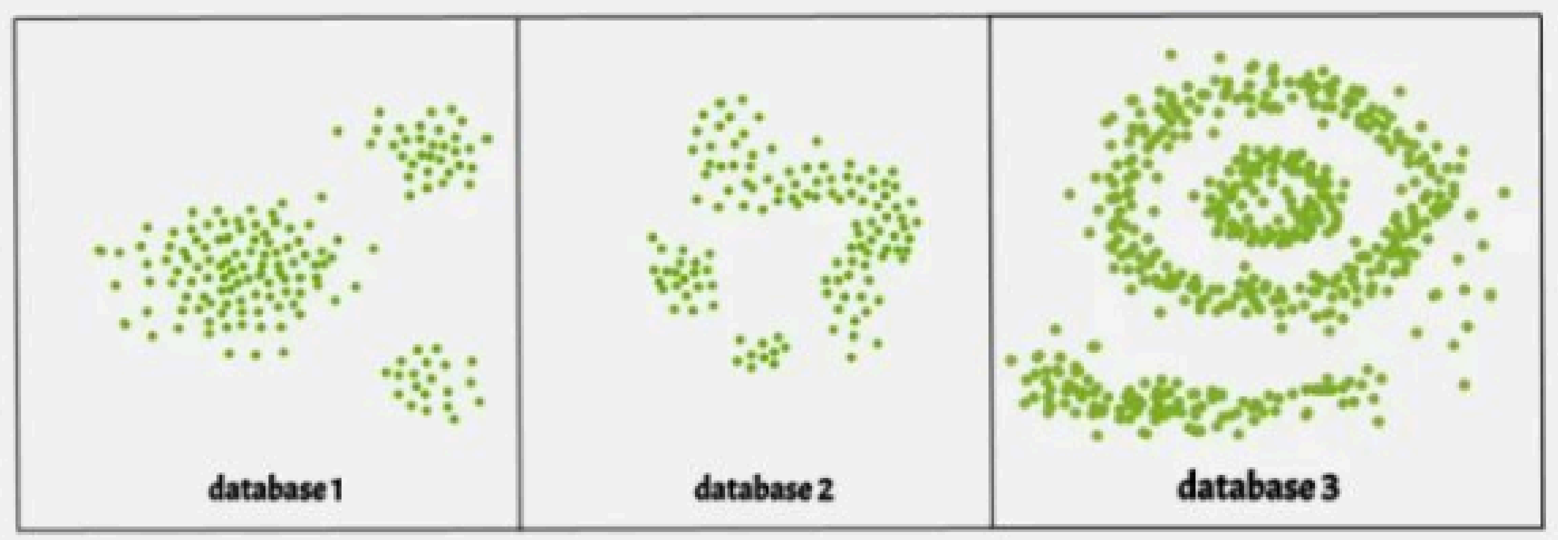
Resources



Clustering Like a Pro: A Beginner’s Guide to DBSCAN

Data clustering is a fundamental task in machine learning and data analysis. One powerful technique that has gained prominence is...

Medium / Dec 26, 2023



DBSCAN Clustering in ML | Density based clustering

DBSCAN is a density-based clustering algorithm that effectively identifies arbitrary-shaped clusters and handles noise, distinguishing it from K-Means and hierarchical clustering, which assume compact, spherical...

GeeksforGeeks / Jan 29

Thank you very much!

Blue AI Team

Youssef Moustafa

Nagham Wael