

Clustering Method Selection Guide

K-Means

Pros:

Fast Scalable Simple

Cons:

Fixed K Spherical Sensitive

*Well-separated,
spherical clusters*

DBSCAN

Pros:

No K needed Any shape Noise handling

Cons:

Parameters Density Memory

*Arbitrary shapes,
noise present*

Hierarchical

Pros:

Dendrogram No K upfront Interpretable

Cons:

Slow Memory No undo

*Need hierarchy,
small datasets*

GMM

Pros:

Soft clustering Flexible Probabilistic

Cons:

Complex Slow Assumptions

*Overlapping,
elliptical clusters*

Mean Shift

Pros:

No K Robust Modes

Cons:

Very slow Bandwidth Memory

*Mode seeking,
computer vision*

Key Question: Do you know the number of clusters?

YES → K-Means, GMM

NO → DBSCAN, Hierarchical, Mean Shift