Types of Clustering

Item	Centroid-Based (k-Means)	Density-Based (DBSCAN)	Hierarchical (Agglomerative)	Distribution-Based (GMM)
Definition	Group data by distance from centers	Group high density regions and separate sparse regions	Build cluster hierarchy by repeating merge/split groups by similarity	Assumes data has multi probability distributions and tries to model them
Non-Tech Idea	Magnets and iron fillings	Group close people in parties, individuals are noise	Group cloths shelves items in tree-style: shirts/pants | shirts → short/long-sleeve	Mixed candy jar. Each flavor has a typical shape and size. Try to fit shapes to data.
Tech Idea	Init pre-defined no. of centroids, assign points by Euclidean distance, shift centroids to avg of assigned points, repeat till stable	Identify core points of neighboring points within a set distance. Grow clusters outward from these points. Sparse points are noise	Each point is a cluster. Then merge closest pairs step by step until all become one cluster. Cut at different levels to specify # clusters	GMM assumes cluster is generated by PDF (Gauss), estimates PDF parameters and assign points based on probability
Example	Segment customers by avg spending in online store	Traffic cars close together in jams are clusters; isolated ones are noise.	Group photo album: holiday vs family events → holiday city → holiday city activity	Classify students by exam scores. Identify overlapping PDF & assign probabilities
Operation	1. Specify k (Elbow/Silhouette) 2. Place k centers randomly 3. Assign points to centers 4. Move centers to assigned points average 5. Repeat until they stay still	1. Set radius R & min points 2. Count neighbors within R for each point 3. Enough neighbors → Core point 4. Add reachable points to expand. 5. Isolated points → noise 6. All points covered → stop	1. Every point → cluster 2. distance bet. clusters 3. merge 2 closest clusters 4. repeat for new clusters 5. stop when all are 1 tree 6. choose level to cut tree	1. Choose # PDFs 2. Init PDF mean & variance 3. Est. probability of each point belonging to each PDF 4. Max step: update parameters (m & v) based on probability. 5. Repeat 3 & 4 until parameters stabilize
Advantages	- Fast & good with large data - Easy to implement	- Any cluster shape (round) - Good with noise & outliers	- No initial # clusters - tree visualizes structure	- Soft assignments (prob.) - Cluster overlaps (k-Means)
Disadvantages	- Set k upfront - Non-round clusters	- Selecting R is tricky - Varying density clusters	- Computationally expensive - Can't reverse merge/split	- Specify # and type of PDF - Assumes data fits PDF
Comparison	- Centroid: round clusters & large data - Density: Complex shapes, noise, outliers			
		- Hierarchy: multi-level structure is important (nested) - Density: very large or noisy data		
	- Distribution: clusters overlap or elliptical - Centroid: distributions unknown or highly non-Gaussian			