1. Explain Clustering algorithms and their use cases
2. Clustering is an **unsupervised learning** technique: you *don’t* have labels, and you want to group data points into clusters based on similarity.
3. The goal is that points in the same cluster are more similar to each other (according to some metric) than to points in other clusters.
4. Uses include exploring structure in data, reducing complexity, finding patterns, anomaly detection, etc.

**When to Use Which (Key Decision Factors)**

When choosing a clustering algorithm, consider:

* **Shape of clusters**: Are they spherical, convex, similar size? Or irregular/non-convex shapes?
* **Number of clusters known or unknown**: If you know how many you want, K-Means or GMM; if unknown, density-based or hierarchical may be better.
* **Size of dataset**: For very large datasets, algorithms with linear or near-linear time / memory are better.
* **Noise / outliers**: Some algorithms handle noise well (DBSCAN etc.), others not so much.
* **Dimensionality**: High-dimensional data can cause problems for distance/similarity metrics, degrade performance.
* **Mixed type data**: If features are categorical + numerical, you might need algorithms or pre-processing that support that (e.g. K-prototypes, mixed similarity metrics).
* **Interpretability**: Want to explain clusters to domain experts? Hierarchical trees, centroids, exemplars might help.

**Real-World Use Cases**

Here are several concrete use cases where clustering is very useful:

| **Use Case** | **Description / Examples** |
| --- | --- |
| **Customer Segmentation in Marketing** | Group customers based on purchasing behavior, demographics, engagement. Then tailor marketing / product recommendations accordingly. [GeeksforGeeks+2NIIT+2](https://www.geeksforgeeks.org/machine-learning/clustering-in-machine-learning/?utm_source=chatgpt.com) |
| **Anomaly / Outlier Detection** | In fraud detection, network security, sensor data etc., you can detect points that don’t belong to any cluster or that are far from cluster centers. Density-based methods are particularly useful. [NIIT+2Google for Developers+2](https://www.niit.com/india/Clustering-Techniques-for-Data-Science?utm_source=chatgpt.com) |
| **Image Segmentation / Computer Vision** | Segment an image into regions (e.g. different objects or backgrounds) based on similarity in color, texture etc. Mean-shift, spectral clustering used. [NIIT+1](https://www.niit.com/india/Clustering-Techniques-for-Data-Science?utm_source=chatgpt.com) |
| **Document / Text Clustering / Topic Discovery** | Group similar documents/articles into topics for organising large corpora, recommendation systems etc. [Analytics Vidhya+2Google for Developers+2](https://www.analyticsvidhya.com/blog/2023/11/types-of-clustering-algorithms-in-machine-learning/?utm_source=chatgpt.com) |
| **Social Network / Community Detection** | Clusters in graphs of social interaction. Identify communities or clusters of users who interact more among themselves. [develearn.in+1](https://www.develearn.in/blogs/clustering-techniques?utm_source=chatgpt.com) |
| **Bioinformatics, Genomics** | Clustering gene expression data, grouping similar proteins or microorganisms. Hierarchical clustering often used in phylogenetics. [develearn.in+1](https://www.develearn.in/blogs/clustering-techniques?utm_source=chatgpt.com) |
| **Geospatial Data / Spatial Analysis** | Clustering geolocations (e.g. hotspots, crime areas, resource allocation, etc.). DBSCAN or density methods useful. [GeeksforGeeks+1](https://www.geeksforgeeks.org/data-science/choosing-the-right-clustering-algorithm-for-your-dataset/?utm_source=chatgpt.com) |
| **Reducing / Summarizing Data** | Represent many data points by clusters / centroids for compression, summarization, visualization. Helps reduce complexity. [GeeksforGeeks+1](https://www.geeksforgeeks.org/machine-learning/clustering-in-machine-learning/?utm_source=chatgpt.com) |

Top of Form

Bottom of Form