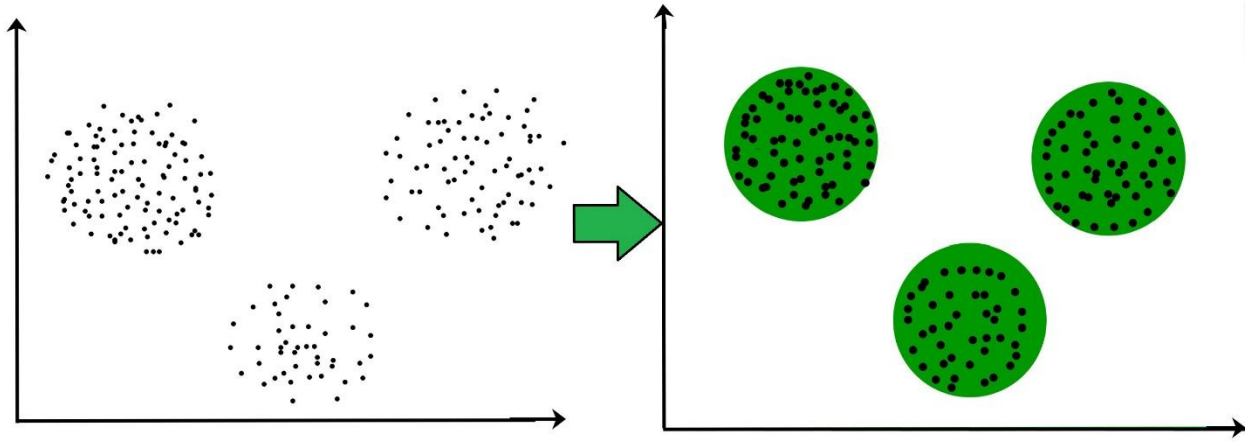
**What is Clustering?**

The task of **grouping data points based on their similarity with each other is called Clustering or Cluster Analysis**. This method is defined under the branch of unsupervised learning, which aims at gaining insights from unlabelled data points.

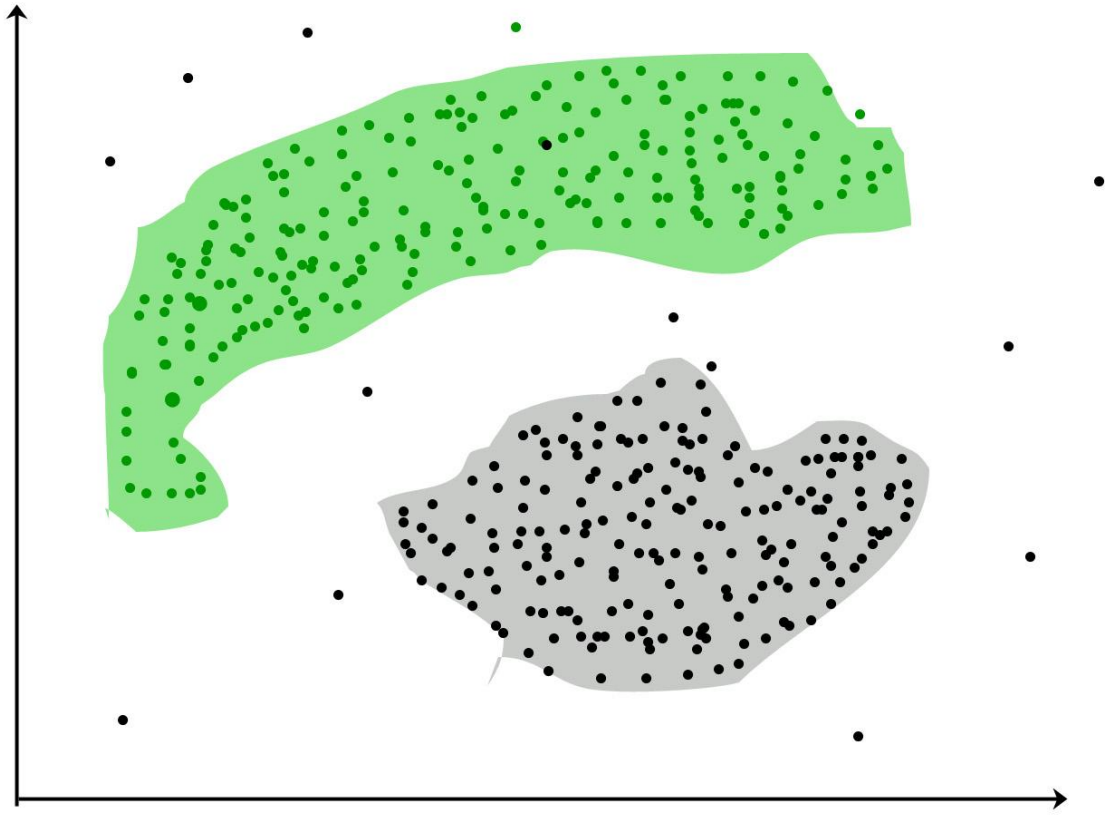
Think of it as you have a dataset of customers shopping habits. **Clustering can help you group customers with similar purchasing behaviors, which can then be used for targeted marketing, product recommendations, or customer segmentation**

For Example, In the graph given below, we can clearly see that there are 3 circular clusters forming on the basis of distance.



Now it is not necessary that the clusters formed **must be circular in shape**. The shape of clusters can be arbitrary. There are many algorithms that work well with detecting arbitrary shaped clusters.

For example, In the below given graph we can see that the clusters formed are not circular in shape.



**Types of Clustering**

Broadly speaking, there are 2 types of clustering that can be performed to group similar data points:

- **Hard Clustering:** In this type of clustering, each data point belongs to a cluster completely or not. For example, Let's say there are 4 data point and we have to cluster them into 2 clusters. So each data point will either belong to cluster 1 or cluster 2.

| Data Points | Clusters |
|---|---|
| A | C1 |
| B | C2 |
| C | C2 |

| Data Points | Clusters |
|---|---|
| D | C1 |

- **Soft Clustering:** In this type of clustering, instead of assigning each data point into a separate cluster, a probability or likelihood of that point being that cluster is evaluated. For example, Let's say there are 4 data point and we have to cluster them into 2 clusters. So we will be evaluating a probability of a data point belonging to both clusters. This probability is calculated for all data points.

| Data Points | Probability of C1 | Probability of C2 |
|---|---|---|
| A | 0.91 | 0.09 |
| B | 0.3 | 0.7 |
| C | 0.17 | 0.83 |
| D | 1 | 0 |

**Uses of Clustering**

Now before we begin with types of clustering algorithms, we will go through the use cases of Clustering algorithms. Clustering algorithms are majorly used for:

- **Market Segmentation:** Businesses use clustering to group their customers and use targeted advertisements to attract more audience.
- **Market Basket Analysis:** Shop owners analyze their sales and figure out which items are majorly bought together by the customers. For example, In USA, according to a study diapers and beers were usually bought together by fathers.
- **Social Network Analysis:** Social media sites use your data to understand your browsing behavior and provide you with targeted friend recommendations or content recommendations.

- **Medical Imaging:** Doctors use Clustering to find out diseased areas in diagnostic images like X-rays.
- **Anomaly Detection:** To find outliers in a stream of real-time dataset or forecasting fraudulent transactions we can use clustering to identify them.
- **Simplify working with large datasets:** Each cluster is given a cluster ID after clustering is complete. Now, you may reduce a feature set's whole feature set into its cluster ID. Clustering is effective when it can represent a complicated case with a straightforward cluster ID. Using the same principle, clustering data can make complex datasets simpler.

There are many more use cases for clustering but there are some of the major and common use cases of clustering. Moving forward we will be discussing Clustering Algorithms that will help you perform the above tasks.

**Types of Clustering Methods**

At the surface level, **clustering helps in the analysis of unstructured data. Graphing, the shortest distance, and the density of the data points are a few of the elements that influence cluster formation**. Clustering is the process of determining how related the objects are **based on a metric called the similarity measure**.

Similarity metrics **are easier to locate in smaller sets of features and harder as the number of features increases**. Depending on the type of clustering algorithm being utilized, several techniques are employed to group the data from the datasets. In this part, the clustering techniques are described. Various types of clustering algorithms are:

1. Centroid-based Clustering (Partitioning methods)
2. Density-based Clustering (Model-based methods)
3. Connectivity-based Clustering (Hierarchical clustering)

We will be going through each of these types in brief.

**1. Centroid-based Clustering (Partitioning methods)**

Centroid-based clustering organizes data points around central vectors (centroids) that represent clusters. Each data point belongs to the cluster with the nearest centroid. Generally, the similarity measure chosen for these algorithms are Euclidian distance, Manhattan Distance or Minkowski Distance.

The datasets are separated into a **predetermined number of clusters, and each cluster is referenced by a vector of values. When compared to the vector value, the input data variable shows no difference and joins the cluster.**

The major drawback for centroid-based algorithms is the requirement that we establish the number of clusters, "k," either intuitively or scientifically (using the Elbow Method) before any clustering machine learning system starts allocating the data points. Despite this limitation, it remains the most popular type of clustering due to its simplicity and efficiency. Popular algorithms of Centroid-based clustering are:

- K-means and
- K-medoids clustering

are some examples of this type clustering.

**2. Density-based Clustering (Model-based methods)**

Density-based clustering identifies clusters as areas of high density separated by regions of low density in the data space. Unlike centroid-based methods, density-based clustering **automatically determines the number of clusters and is less susceptible to initialization positions**. **Key Characteristics:**

- Can find arbitrarily shaped clusters
- Handles noise and outliers well
- Excels with clusters of different sizes and shapes
- Ideal for datasets with irregularly shaped or overlapping clusters
- Effectively manages both dense and sparse data regions
- Focus on local density allows detection of various cluster morphologies

*The most popular density-based clustering algorithm is DBSCAN and OPTICS (Ordering Points To Identify Clustering Structure).*

**3. Connectivity-based Clustering (Hierarchical clustering)**

Connectivity-based clustering builds a **hierarchy of clusters using a measure of connectivity based on distance** when organizing a collection of items based on their similarities. This method builds a **dendrogram**, a tree-like structure that visually represents the relationships between objects.

*At the base of the tree, each object starts as its own individual cluster. The algorithm then evaluates how similar the objects are to one another and begins merging the closest pairs of clusters into larger groups. This process continues iteratively, with clusters being combined step by step, until all objects are united into a single cluster at the top of the tree.*

There are 2 approaches for Hierarchical clustering:

- **Divisive Clustering:** It follows a top-down approach, here we consider all data points to be part one big cluster and then this cluster is divided into smaller groups.
- **Agglomerative Clustering:** It follows a bottom-up approach, here we consider all data points to be part of individual clusters and then these clusters are clubbed together to make one big cluster with all data points.

*Till now, we have understood **traditional "hard" clustering methods**, where each data point is assigned to exactly one cluster. These methods, like K-Means and hierarchical clustering, are powerful and widely used, but they have limitations when dealing with ambiguous or overlapping data. After learning all about hard clustering methods we can addresses these limitations with **soft clustering that** allows data points to belong to **multiple clusters simultaneously**, with varying degrees of membership. This approach is particularly useful when the boundaries between clusters are not clear-cut or when data points exhibit characteristics of more than one group.*

Two of the most popular soft clustering techniques are:

**4. Distribution-based Clustering**

Distribution-based clustering is a technique that assumes **data points are generated from a mixture of probability distributions (e.g., Gaussian, Poisson, etc.)**. The goal is to identify clusters by estimating the parameters of these distributions. In distribution-based clustering:

- Each cluster is represented by a probability distribution.
- Data points are assigned to clusters based on how likely they are to belong to each distribution.
- Unlike distance-based methods (e.g., K-Means), this approach can capture clusters of varying shapes, sizes, and densities.

Many real-world datasets, such as sensor data, financial data, or biological measurements, naturally follow statistical distributions. The most popular distribution-based clustering algorithm is Gaussian Mixture Model.

**5. Fuzzy Clustering**

Fuzzy clustering allows data points to belong to multiple clusters with varying degrees of membership.

- Each data point is assigned a membership value between 0 and 1 for every cluster.