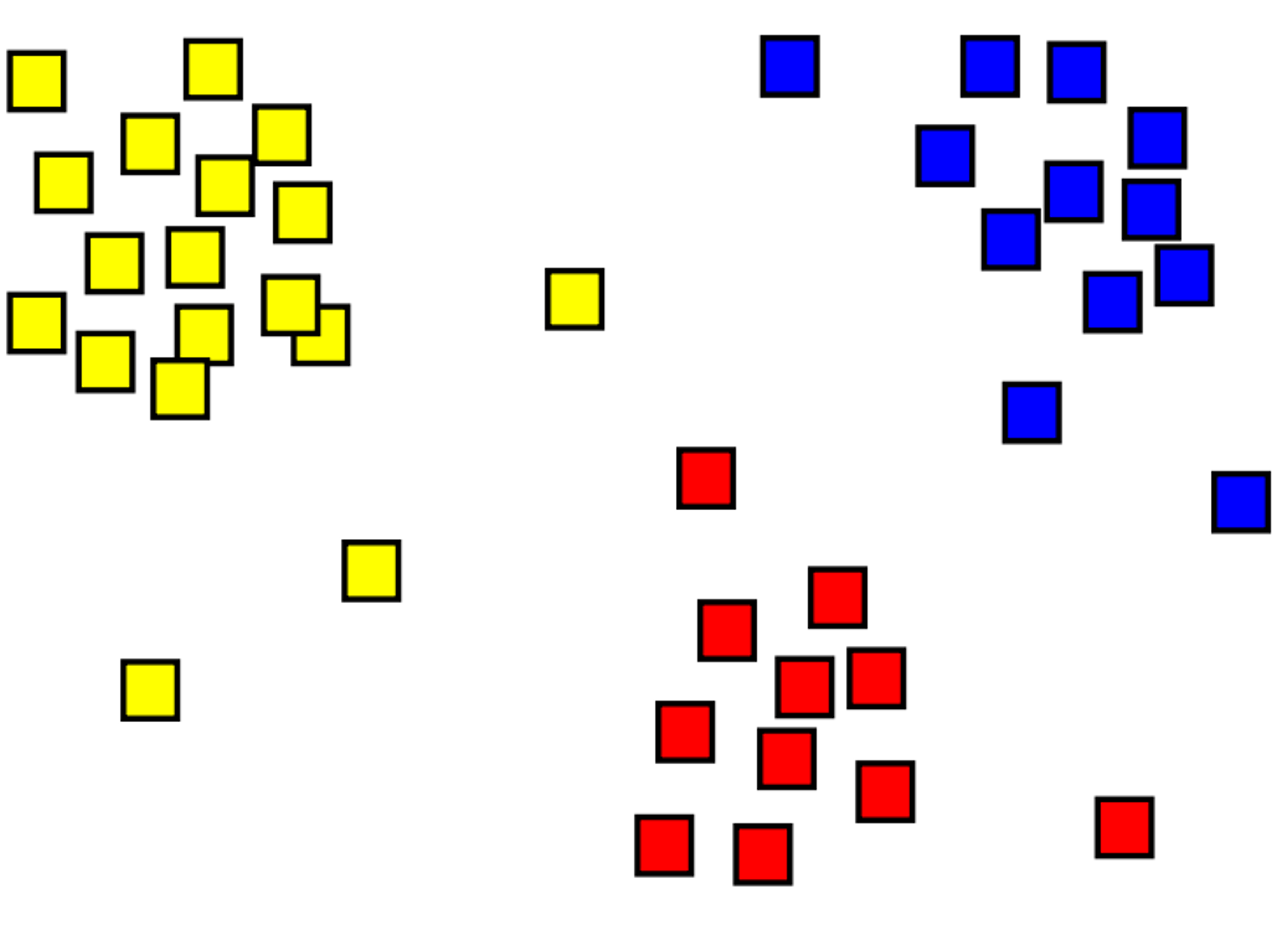
**Introduction**

Completed100 XP

* 2 minutes

Clustering is the process of grouping objects with similar objects. For example, in the image below we have a collection of 2D coordinates that have been clustered into three categories - top left (yellow), bottom (red), and top right (blue).



A major difference between clustering and classification models is that clustering is an ‘unsupervised’ method, where ‘training’ is done without labels. Instead, models identify examples that have a similar collection of features. In the preceding image, examples that are in a similar location are grouped together.

Clustering is common and useful for exploring new data where patterns between data points, such as high-level categories, aren't yet known. It's used in many fields that need to automatically label complex data, including analysis of social networks, brain connectivity, spam filtering, and so on.

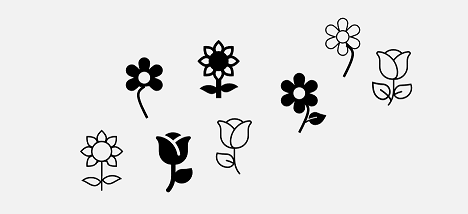
**What is clustering?**

Completed100 XP

* 5 minutes

*Clustering* is a form of *unsupervised* machine learning in which observations are grouped into clusters based on similarities in their data values, or *features*. This kind of machine learning is considered unsupervised because it doesn't make use of previously known *label* values to train a model. In a clustering model, the label is the cluster to which the observation is assigned, based purely on its features.

For example, suppose a botanist observes a sample of flowers and records the number of petals and leaves on each flower.



It may be useful to group these flowers into clusters based on similarities between their features.

There are many ways this grouping could be done. For example, if most flowers have the same number of leaves, they could be grouped into flowers with many vs few petals. Alternatively, if both petal and leaf counts vary considerably there may be a pattern to discover, such as flowers with many leaves also having many petals. The goal of the clustering algorithm is to find the optimal way to split the dataset into groups. What ‘optimal’ means depends on both the algorithm used and the dataset that is provided.

Although this flower example may be simple for a human to achieve with only a few samples, as the dataset grows to thousands of samples or to more than two features, clustering algorithms become useful to quickly dissect a dataset into groups.

# Evaluate different types of clustering

Completed100 XP

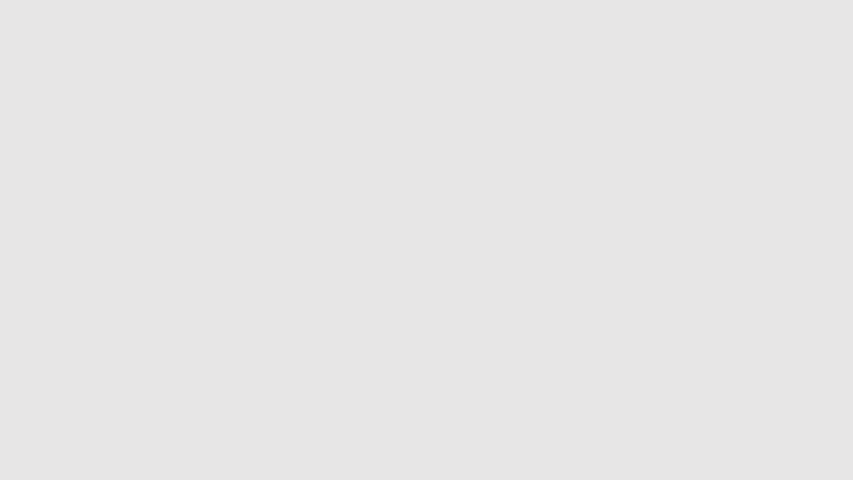
* 5 minutes

## Training a clustering model

There are multiple algorithms you can use for clustering. One of the most commonly used algorithms is K-Means clustering that, in its simplest form, consists of the following steps:

1. The feature values are vectorized to define n-dimensional coordinates (where n is the number of features). In the flower example, we have two features: number of petals and number of leaves. So, the feature vector has two coordinates that we can use to conceptually plot the data points in two-dimensional space.
2. You decide how many clusters you want to use to group the flowers, and call this value ***k***. For example, to create three clusters, you would use a k value of 3. Then k points are plotted at random coordinates. These points become the center points for each cluster, so they're referred to as centroids.
3. Each data point (in this case flower) is assigned to its nearest centroid.
4. Each centroid is moved to the center of the data points assigned to it based on the mean distance between the points.
5. After the centroid is moved, the data points may now be closer to a different centroid, so the data points are reassigned to clusters based on the new closest centroid.
6. The centroid movement and cluster reallocation steps are repeated until the clusters become stable or a predetermined maximum number of iterations is reached.

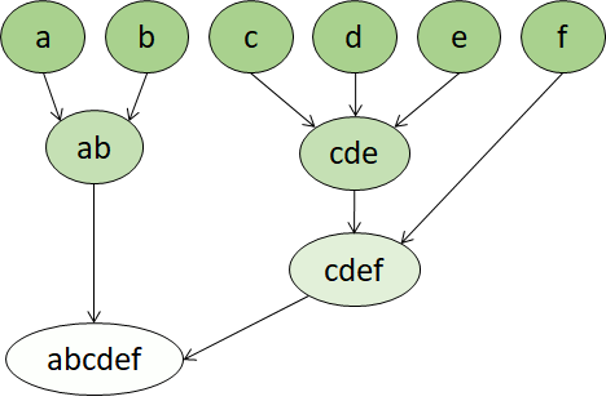
The following animation shows this process:



## Hierarchical Clustering

Hierarchical clustering is another type of clustering algorithm in which clusters themselves belong to a larger group, which belong to even larger groups, and so on. The result is that data points can be clusters in differing degrees of precision: with a large number of very small and precise groups, or a small number of larger groups.

For example, if we apply clustering to the meanings of words, we may get a group containing adjectives specific to emotions ('angry,' 'happy,' and so on). This group belongs to a group containing all human-related adjectives (‘happy’, ‘handsome,’ ‘young’), which belongs to an even higher group containing all adjectives (‘happy,’ ‘green,’ ‘handsome,’ ‘hard,’ etc.).



Hierarchical clustering is useful for not only breaking data into groups, but understanding the relationships between these groups. A major advantage of hierarchical clustering is that it doesn't require the number of clusters to be defined in advance. And, it can sometimes provide more interpretable results than nonhierarchical approaches. The major drawback is that these approaches can take longer to compute than simpler approaches and sometimes aren't suitable for large datasets.