Understanding and Implementing Clustering Models

Janani Ravi
CO-FOUNDER, LOONYCORN

www.loonycorn.com

Overview

Clustering as a form of unsupervised machine learning

Different families of clustering algorithms

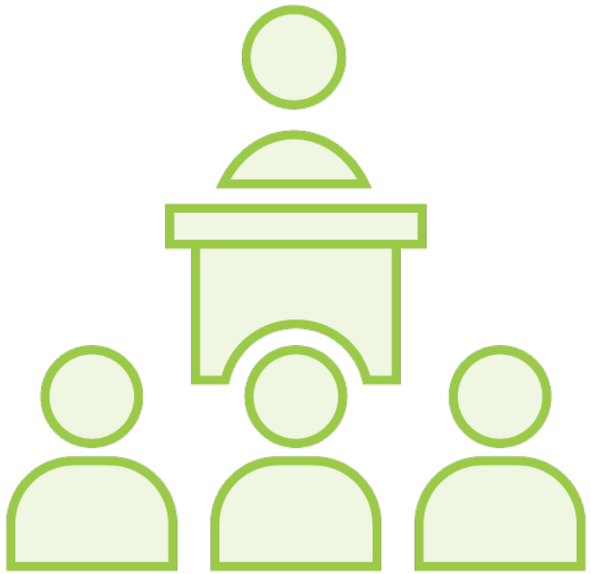
Choosing the right clustering algorithm

K-means clustering

Hierarchical clustering

Clustering Algorithms

Types of ML Algorithms



Supervised

Labels associated with the training data is used to correct the algorithm

Unsupervised

The model has to be set up right to learn structure in the data

Types of ML Algorithms

Supervised

Labels associated with the training data is used to correct the algorithm

Unsupervised

The model has to be set up right to learn structure in the data

Patterns in Data

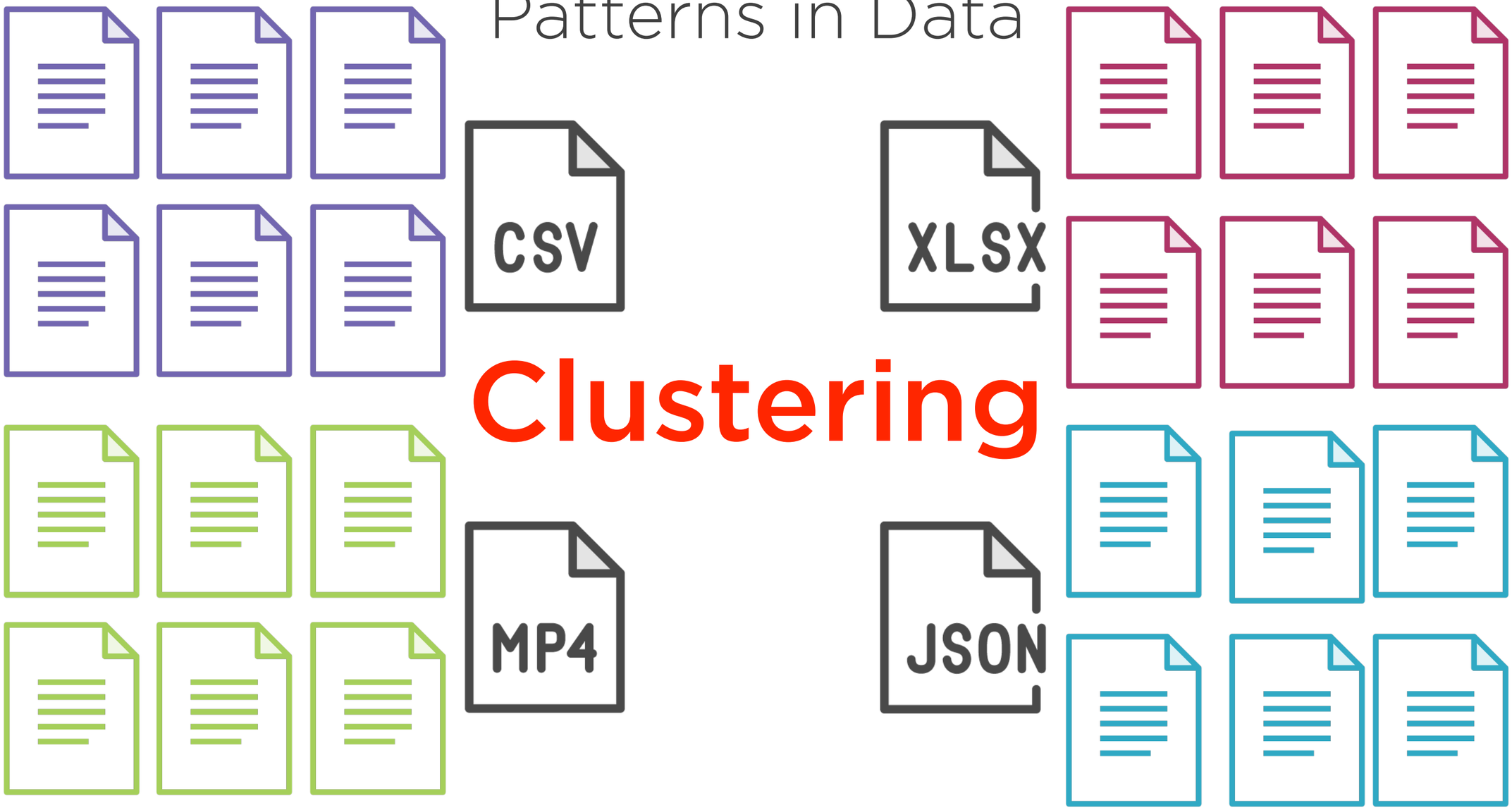
Patterns in Data

How do you make sense of this?

Patterns in Data

Group them based on some common attributes

Patterns in Data



Clustering

Clustering

A set of points, each representing a Facebook user

Clustering

Same group = similar

Different group = different

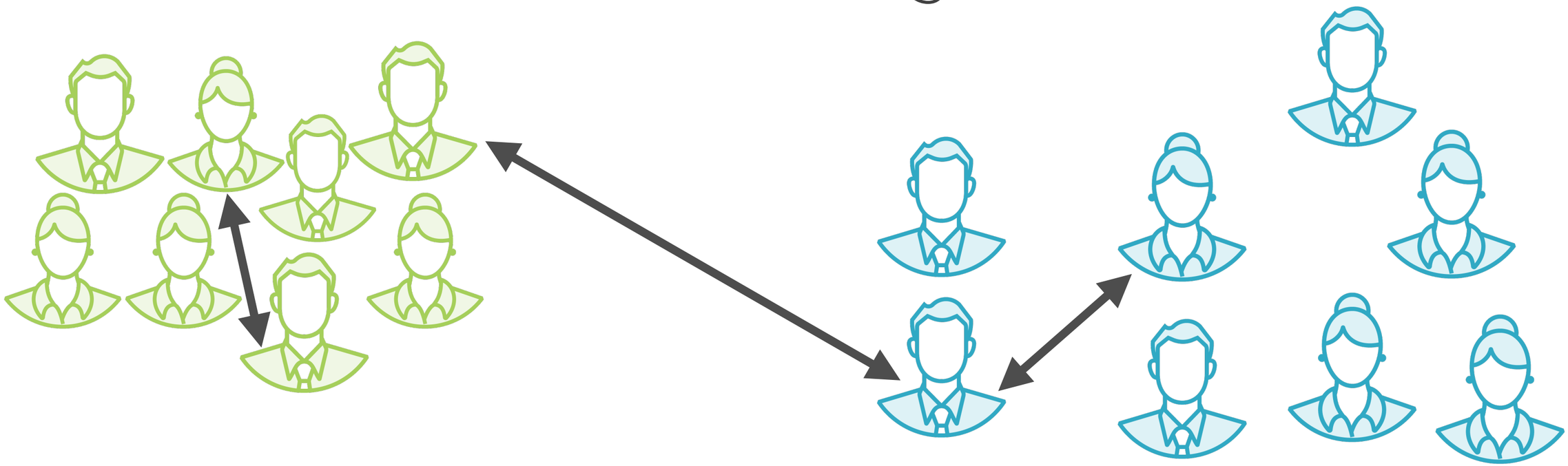
Users in a Cluster

May like the same kind of music

May have gone to the same high school

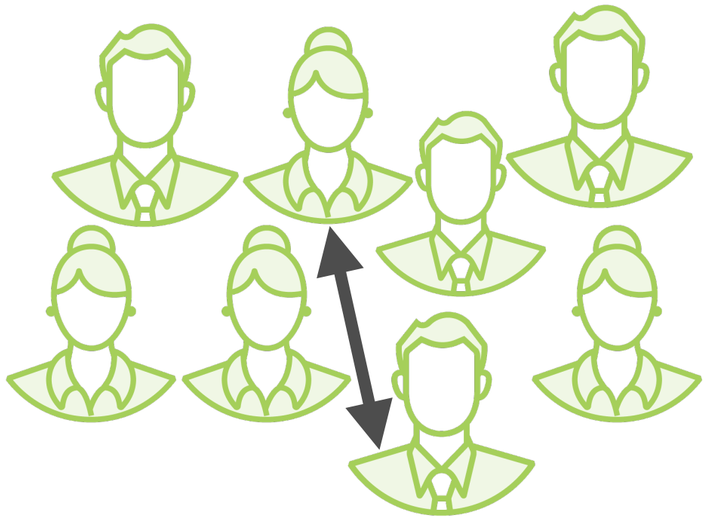
May enjoy the same kinds of movies

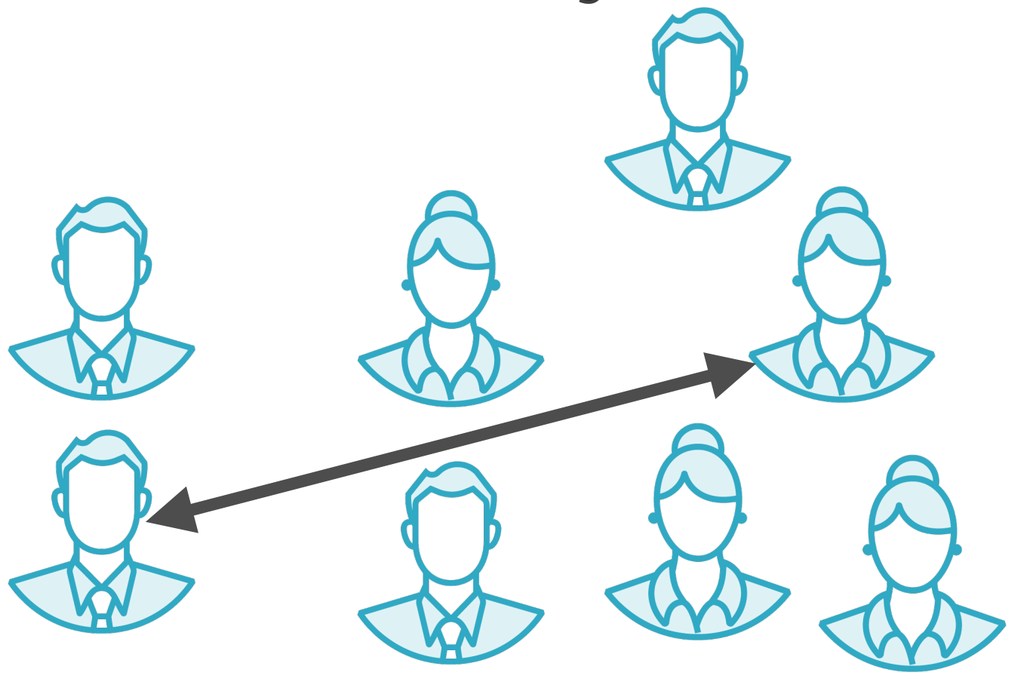
Clustering



The distance between users indicates how similar they are

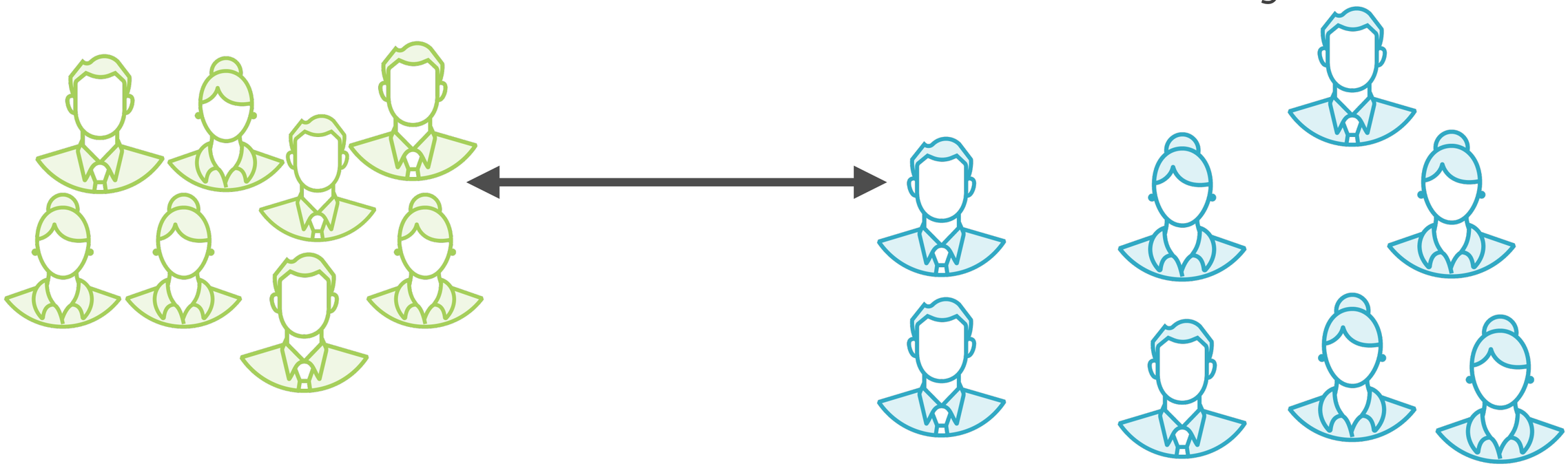
Maximize Intra-cluster Similarity





Distances between users in the same cluster should be small

Minimize Inter-cluster Similarity



Between users in different clusters distances should be large

Unsupervised ML Algorithms

Clustering

Identify patterns in data items e.g. K-means clustering

Autoencoding

Identify latent factors that drive data e.g. PCA

Unsupervised ML Algorithms

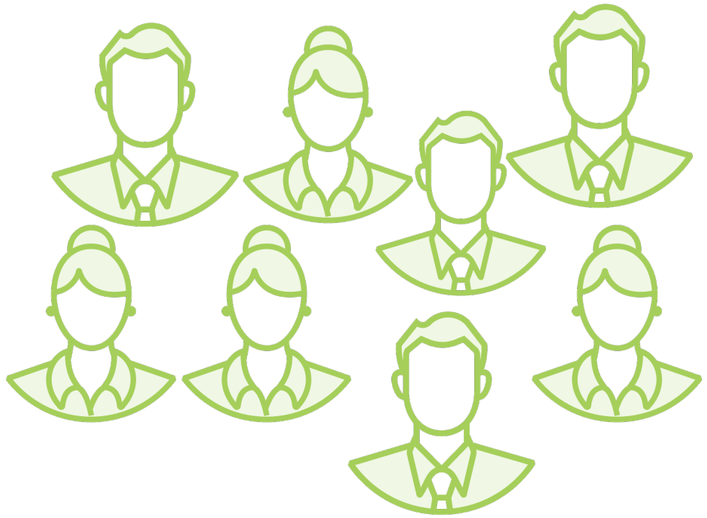
Clustering

Identify patterns in data items e.g. K-means clustering

Autoencoding

Identify latent factors that drive data e.g. PCA

Clustering

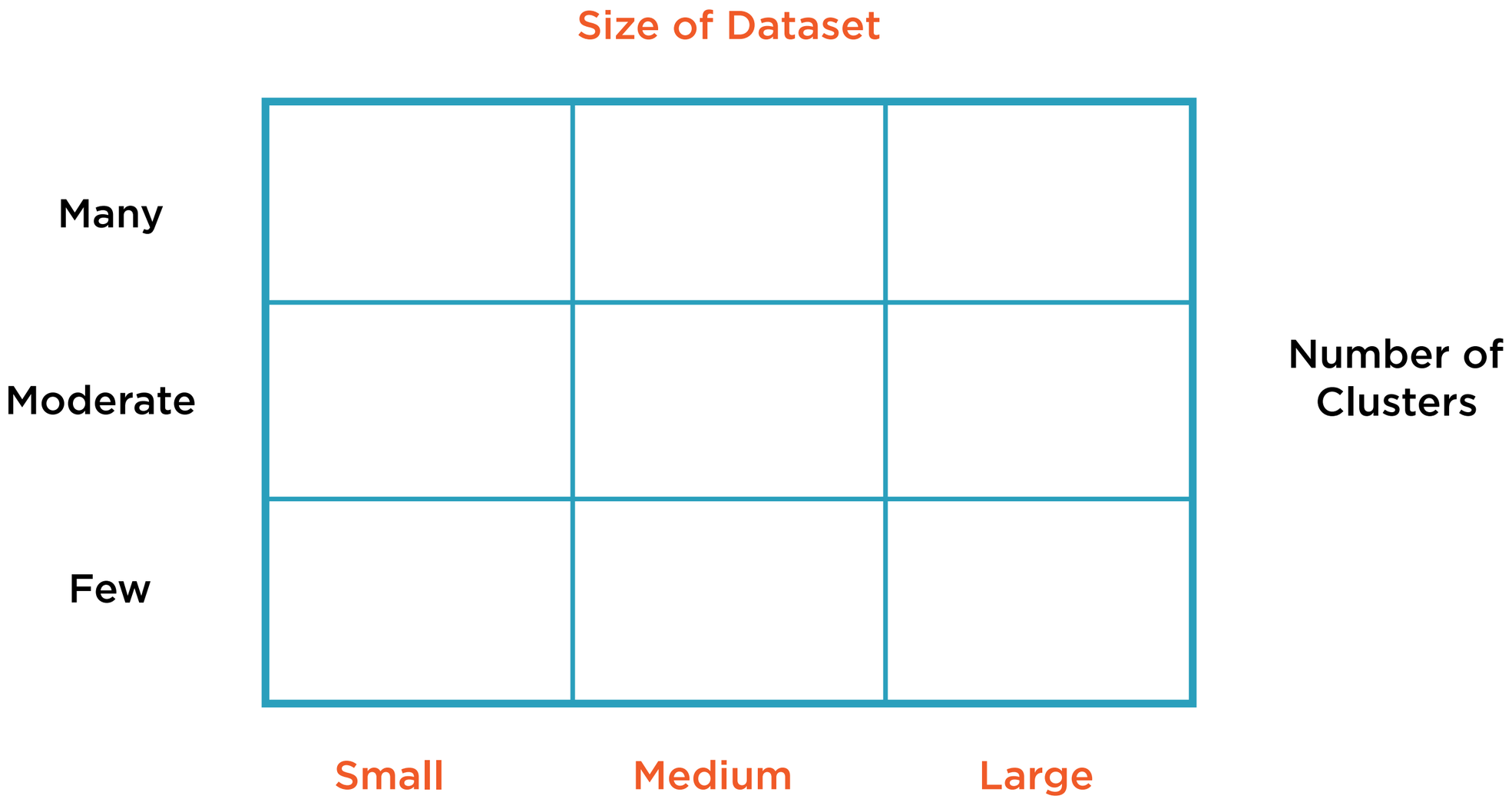


Same group = similar

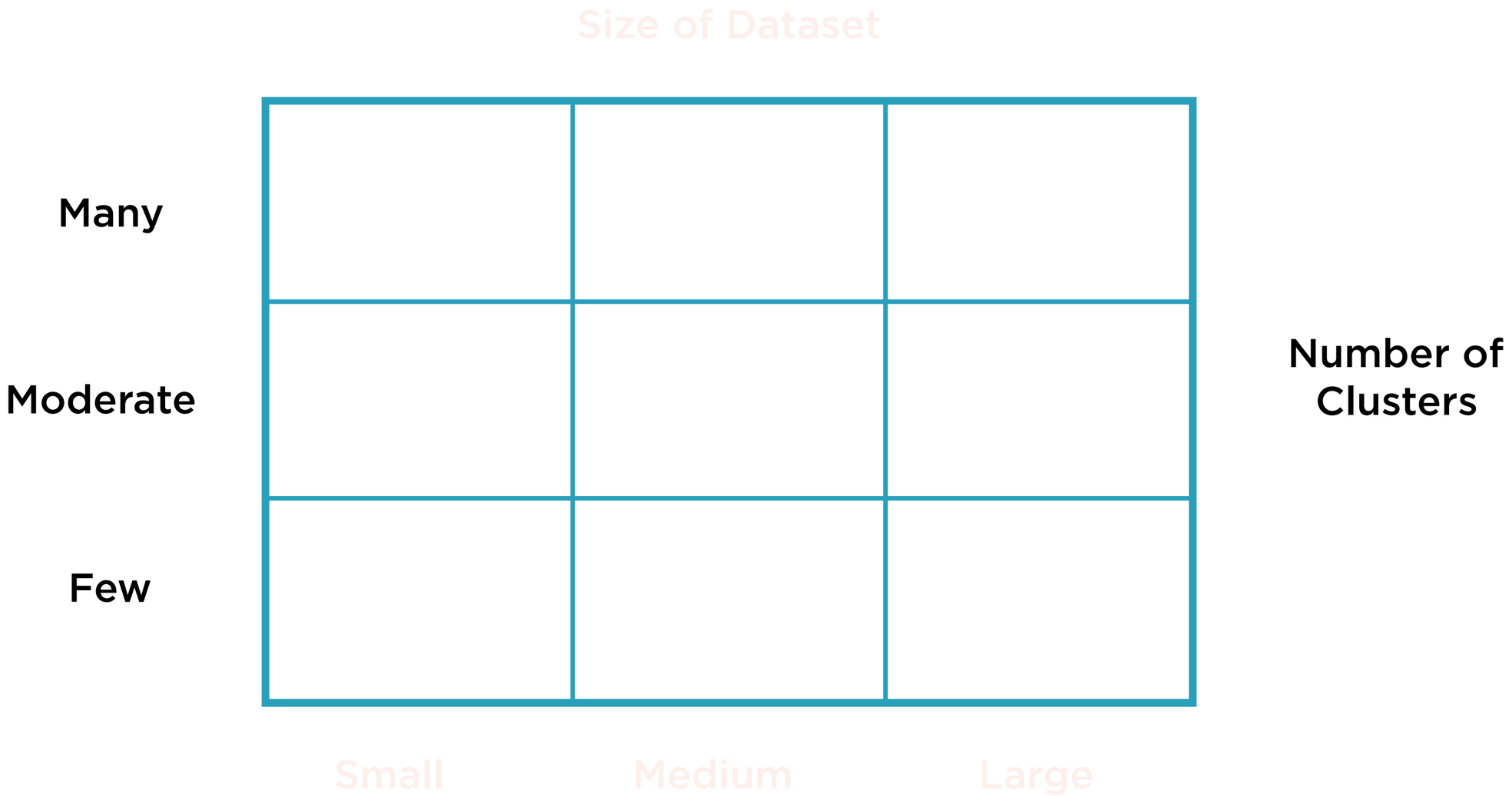
Different group = different

Choosing Clustering Algorithms

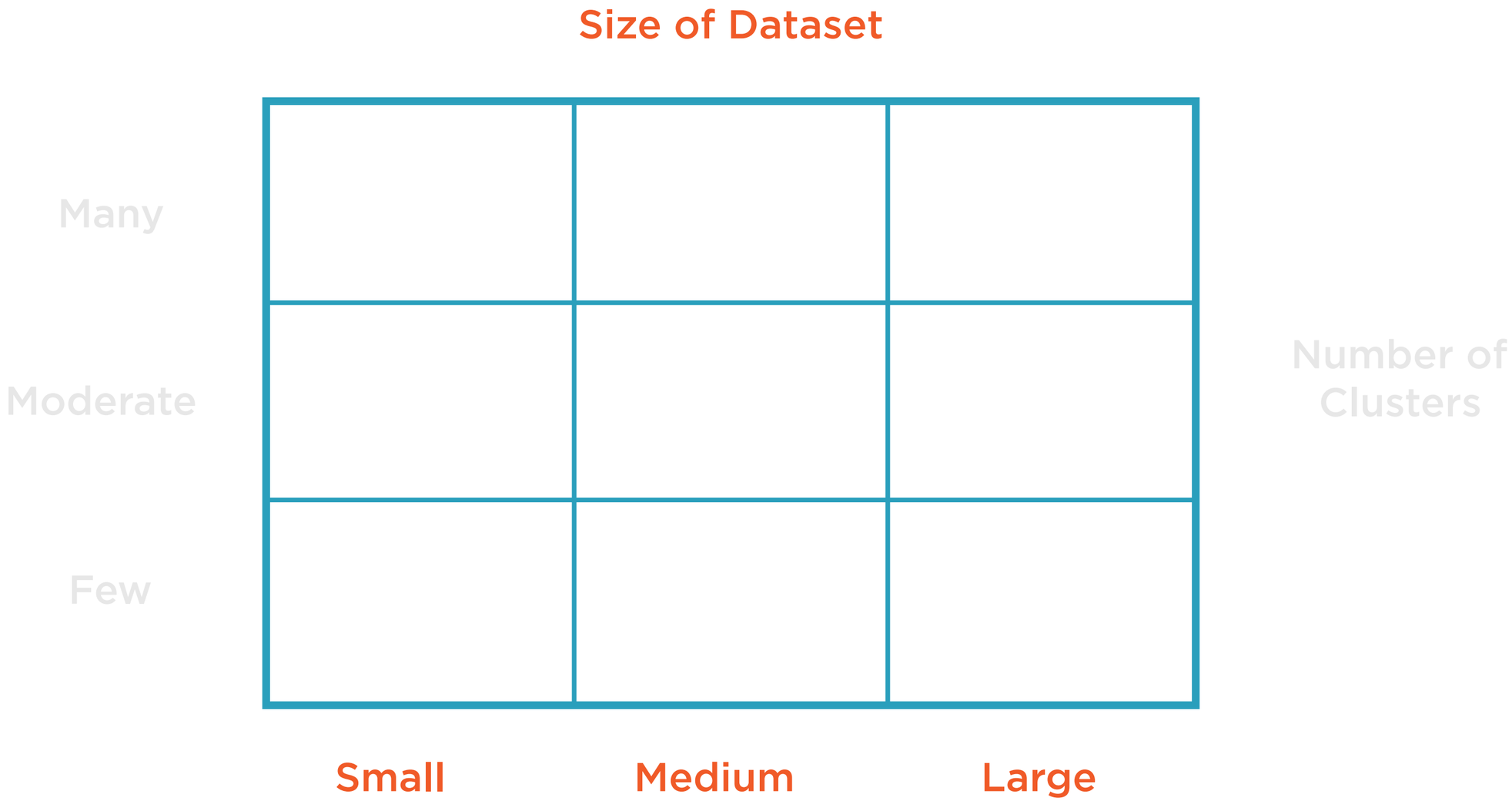
Choosing Clustering Algorithms



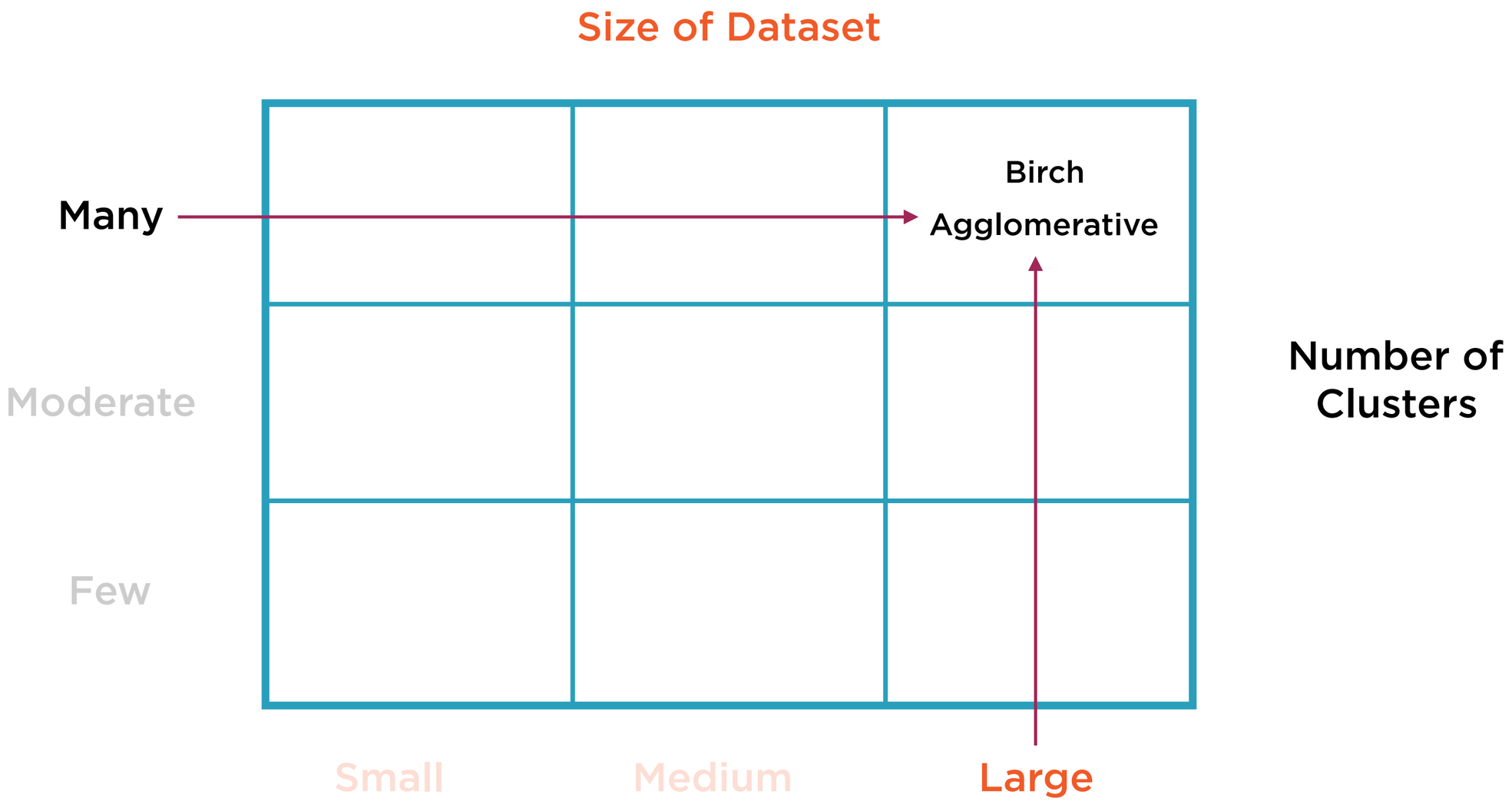
Choosing Clustering Algorithms



Choosing Clustering Algorithms



Choosing Clustering Algorithms



BIRCH, Agglomerative Clustering

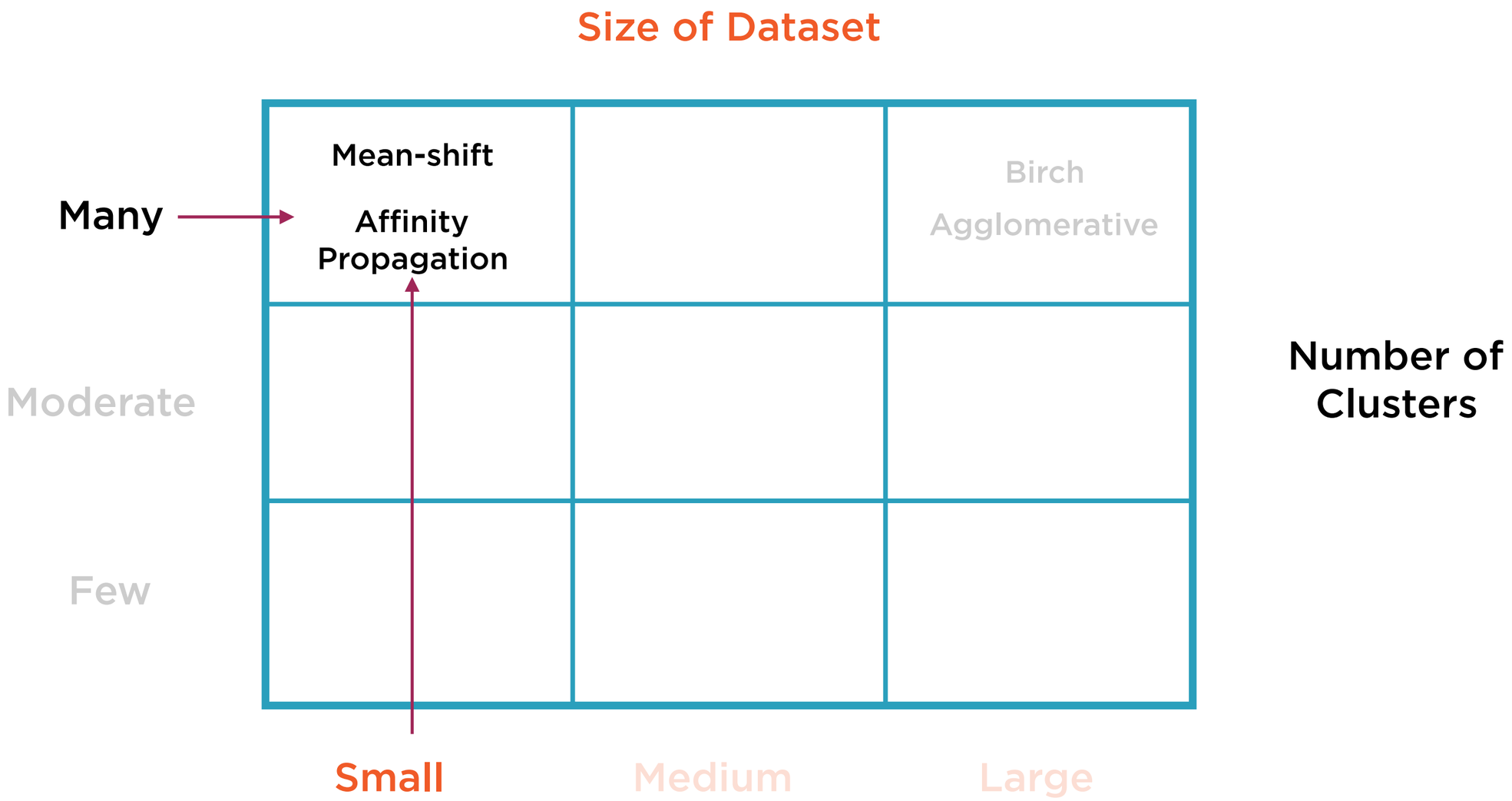
Hierarchical clustering algorithms

Also known as connectivity-based clustering

Build a tree representation of the data

Which may then be merged together into different numbers of clusters

Choosing Clustering Algorithms



Mean-shift, Affinity Propagation

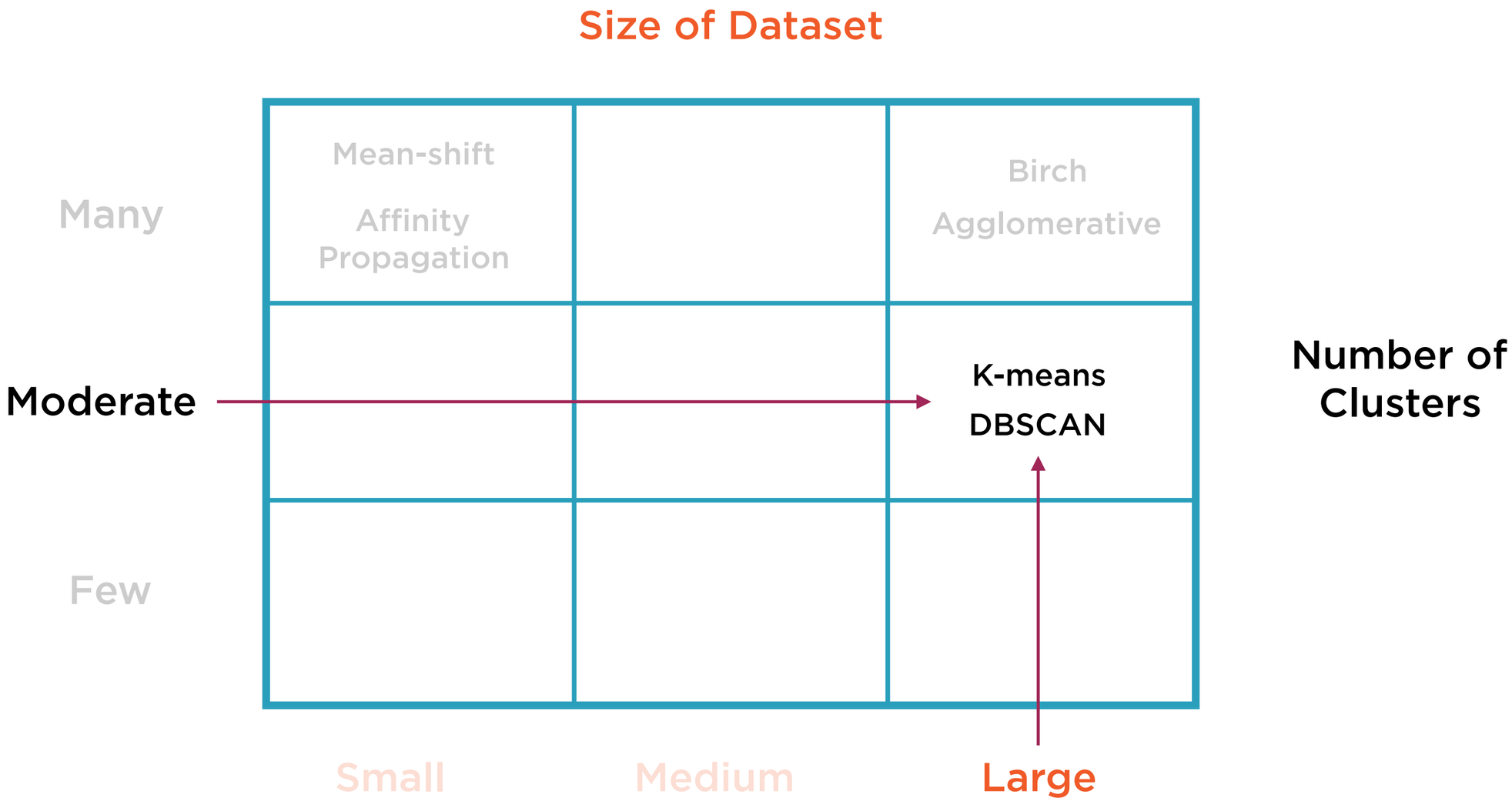
Small datasets, large number of clusters

Both work well with uneven cluster sizes and manifold shapes

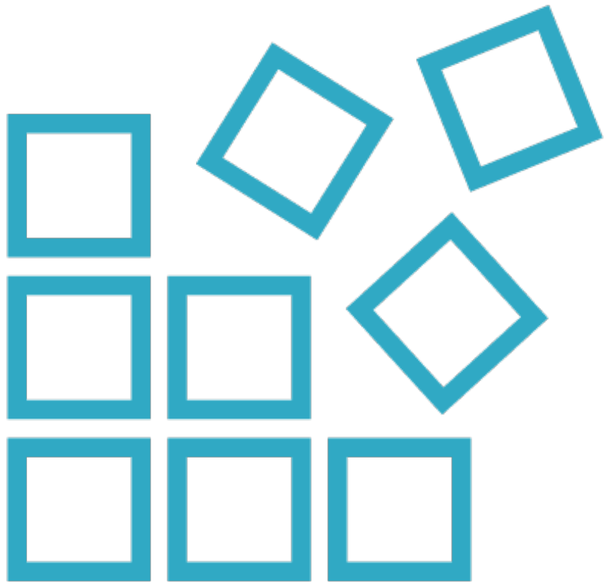
Mean-shift uses pairwise distances between points

Affinity Propagation does not need number of clusters to be specified

Choosing Clustering Algorithms



K-means, DBSCAN



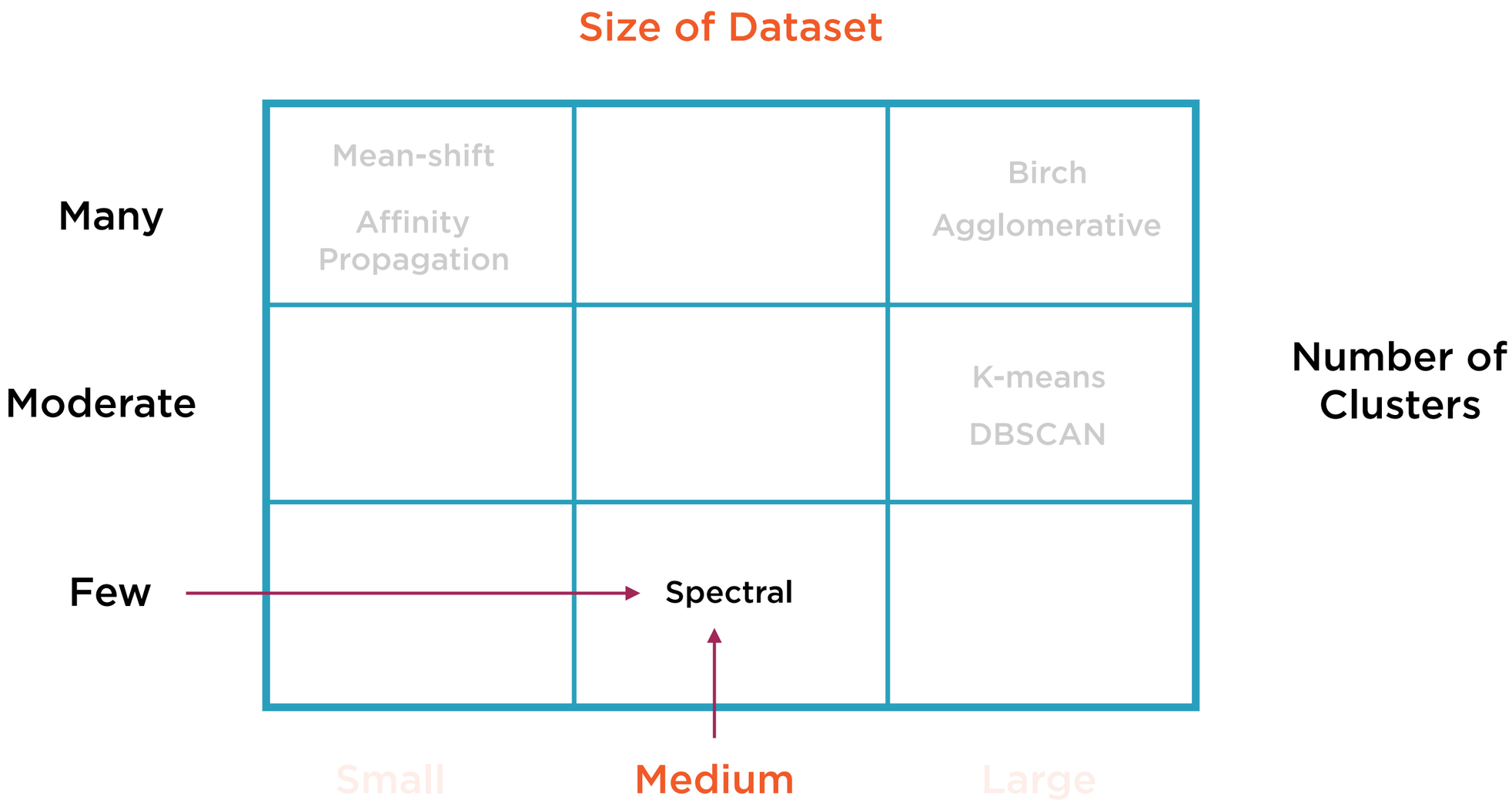
Large datasets, moderate number of clusters

K-means for even cluster sizes and flat surfaces

Mini-batch K-means tweaks algorithm to be much faster, almost as good

DBSCAN for uneven cluster sizes and manifolds

Choosing Clustering Algorithms



Spectral Clustering

Small datasets, small number of clusters

Simple to implement

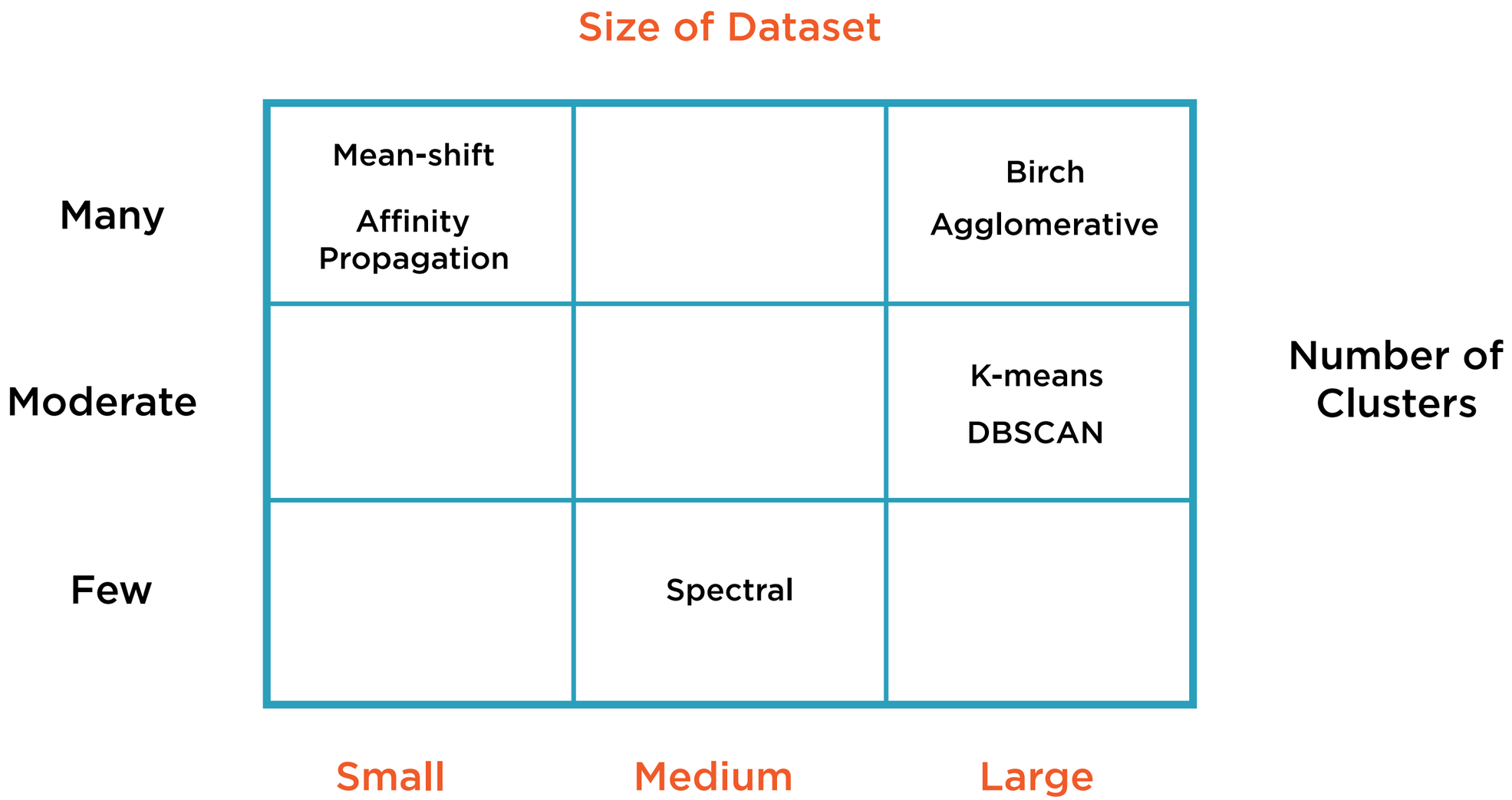
Intuitive results for data exploration

Even cluster sizes

Fine for manifolds

Relies on distances between points

Choosing Clustering Algorithms



Other Ways to Organize Clustering Algorithms

Hierarchical (Connectivity-based)

Centroid-based

Distribution-based

Density-based

Hierarchical Clustering

Group entities based on their connectivity; objects close to each other are more likely to be in the same cluster.

Centroid-based Clustering

Represent each cluster by a centroid (central vector) which may not be an actual entity at all. Cluster entities based on distance from these centroids.

Distribution-based Clustering

Entities that are likely from the same distribution are more likely to be in the same cluster. Work very well for artificially generated data which tend to be drawn by sampling random points from distributions.

Density-based Clustering

Define clusters based on regions of high density (concentration) of entities. Objects in sparse areas are often treated as noise.

Hard vs. Soft Clustering

Hard Clustering

Each point belongs in exactly one cluster

Virtually all famous clustering algorithms are hard clustering

Soft Clustering

Each point has a probability of being in each cluster

FCM (Fuzzy C-Means) is a relatively famous soft clustering algorithm

K-means Clustering

Clustering Objective

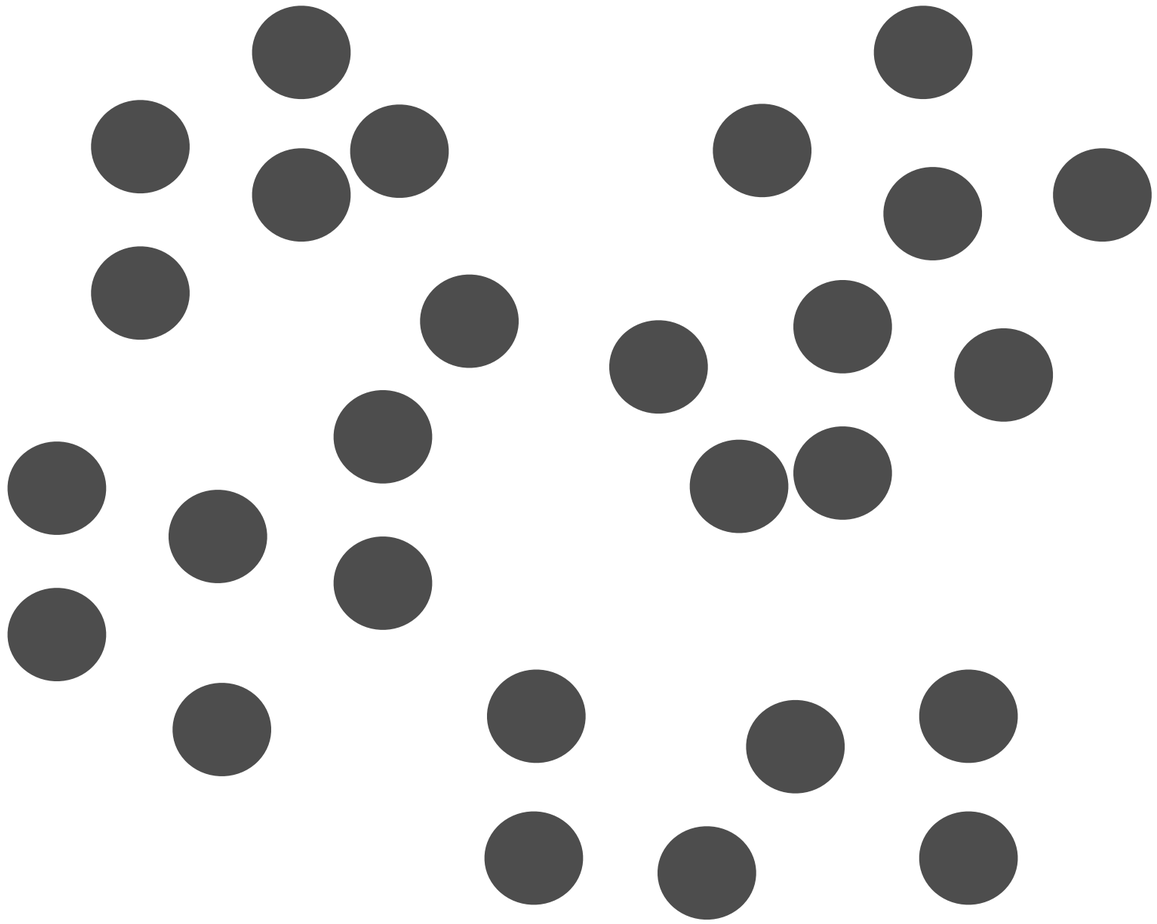
Centroid-based clustering algorithm

Maximize intra-cluster similarity

Minimize inter-cluster similarity

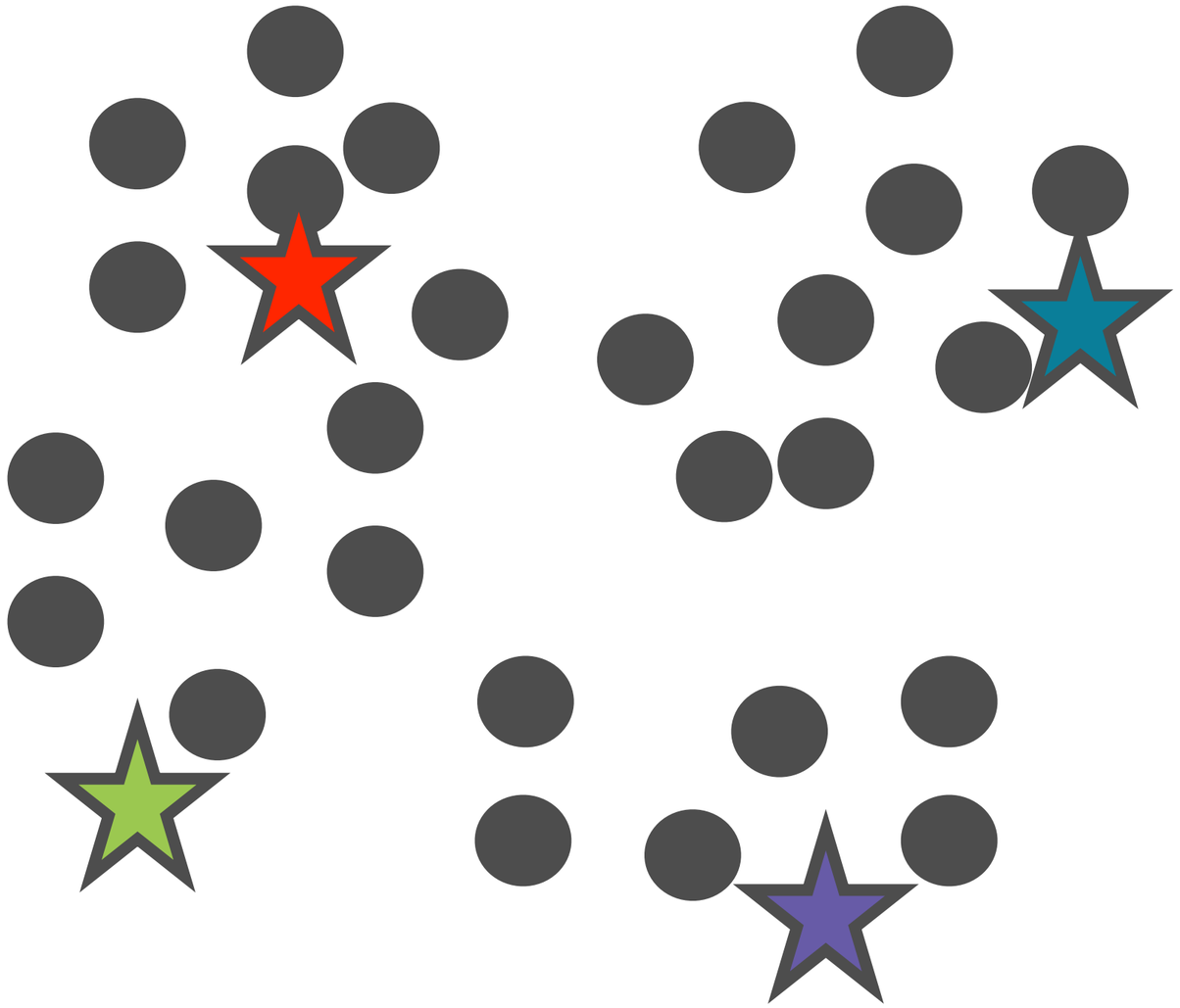
K-means Clustering

Initialize K centroids i.e means



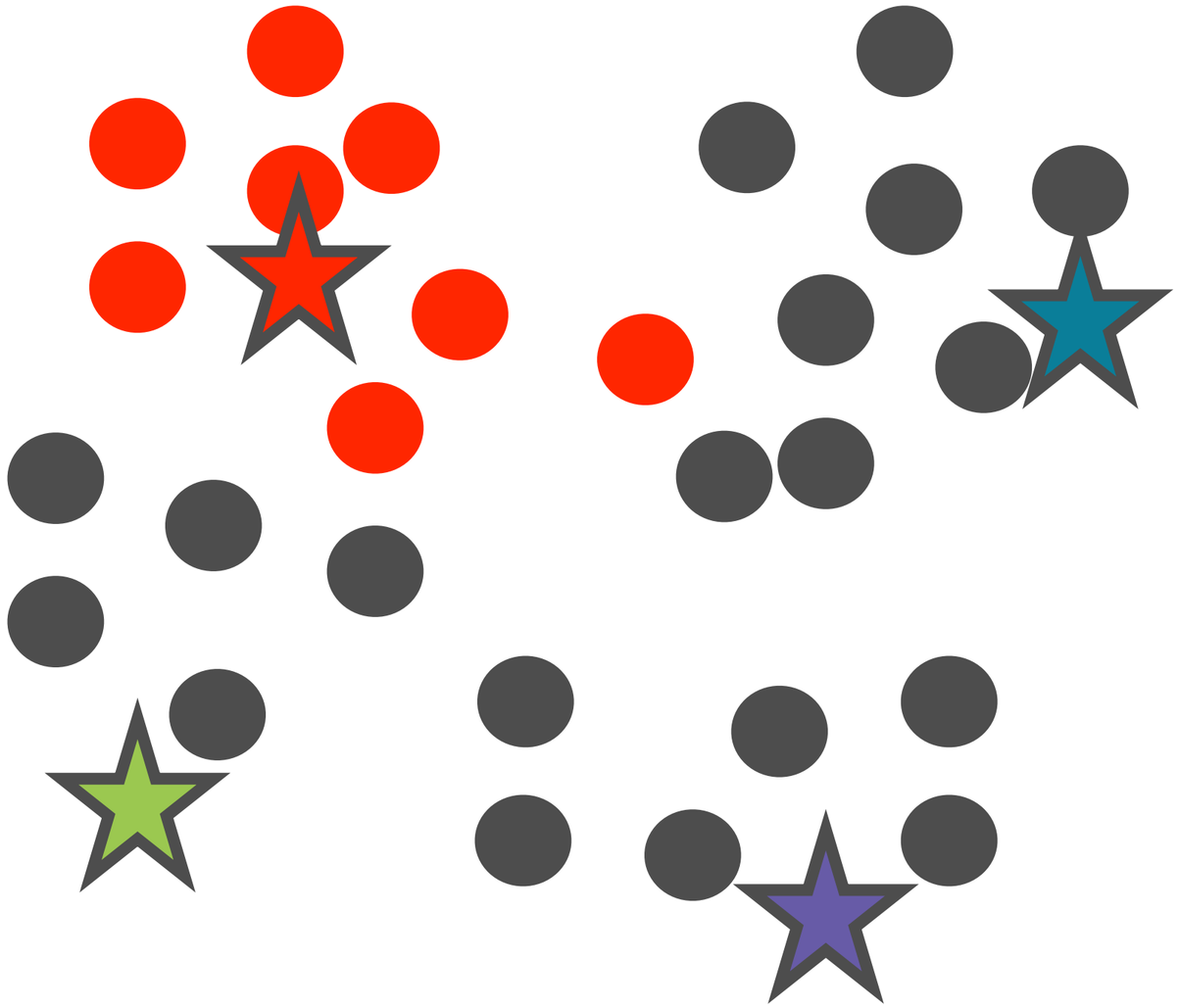
K-means Clustering

Assign each point to a cluster



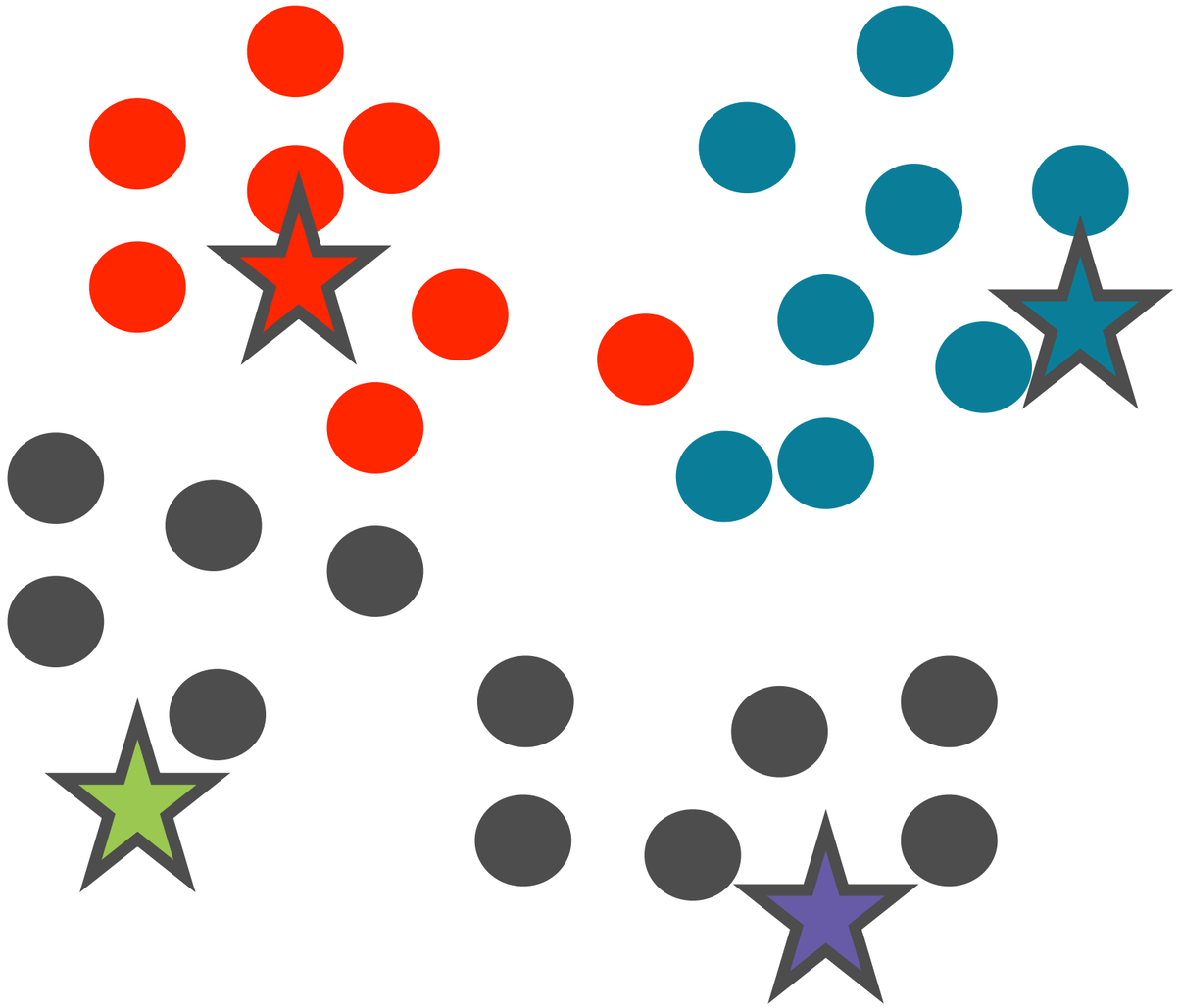
K-means Clustering

Assign each point to a cluster



K-means Clustering

Assign each point to a cluster



K-means Clustering

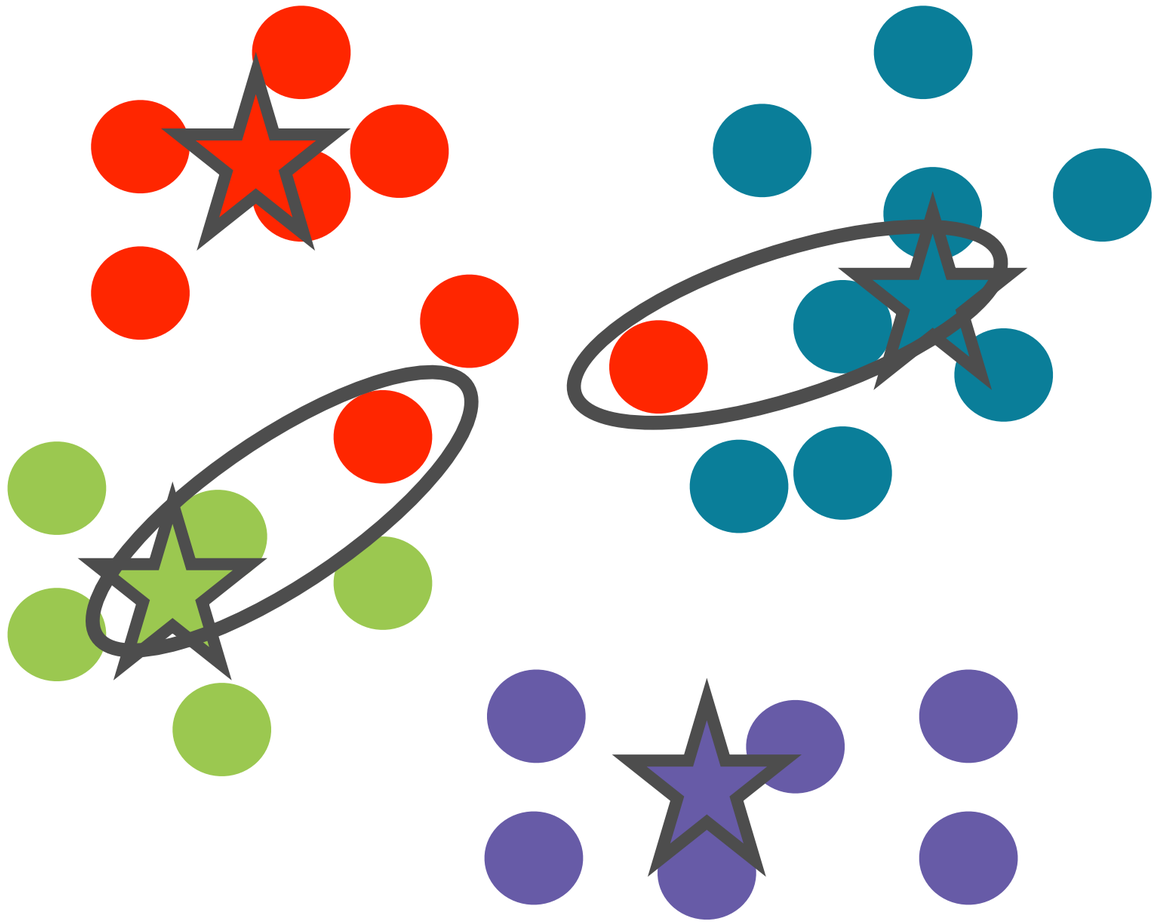
Assign each point to a cluster

K-means Clustering

Recalculate the mean for each cluster

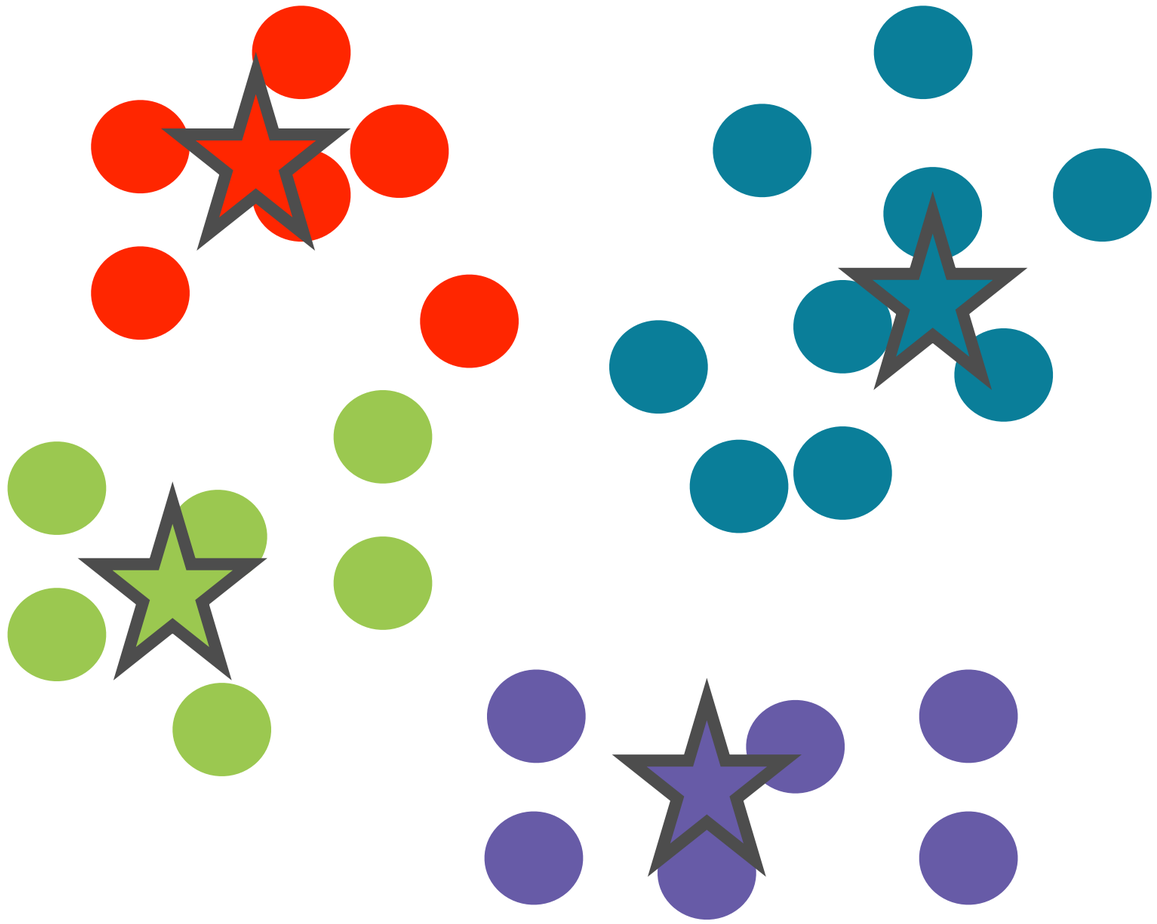
K-means Clustering

Re-assign the points to clusters

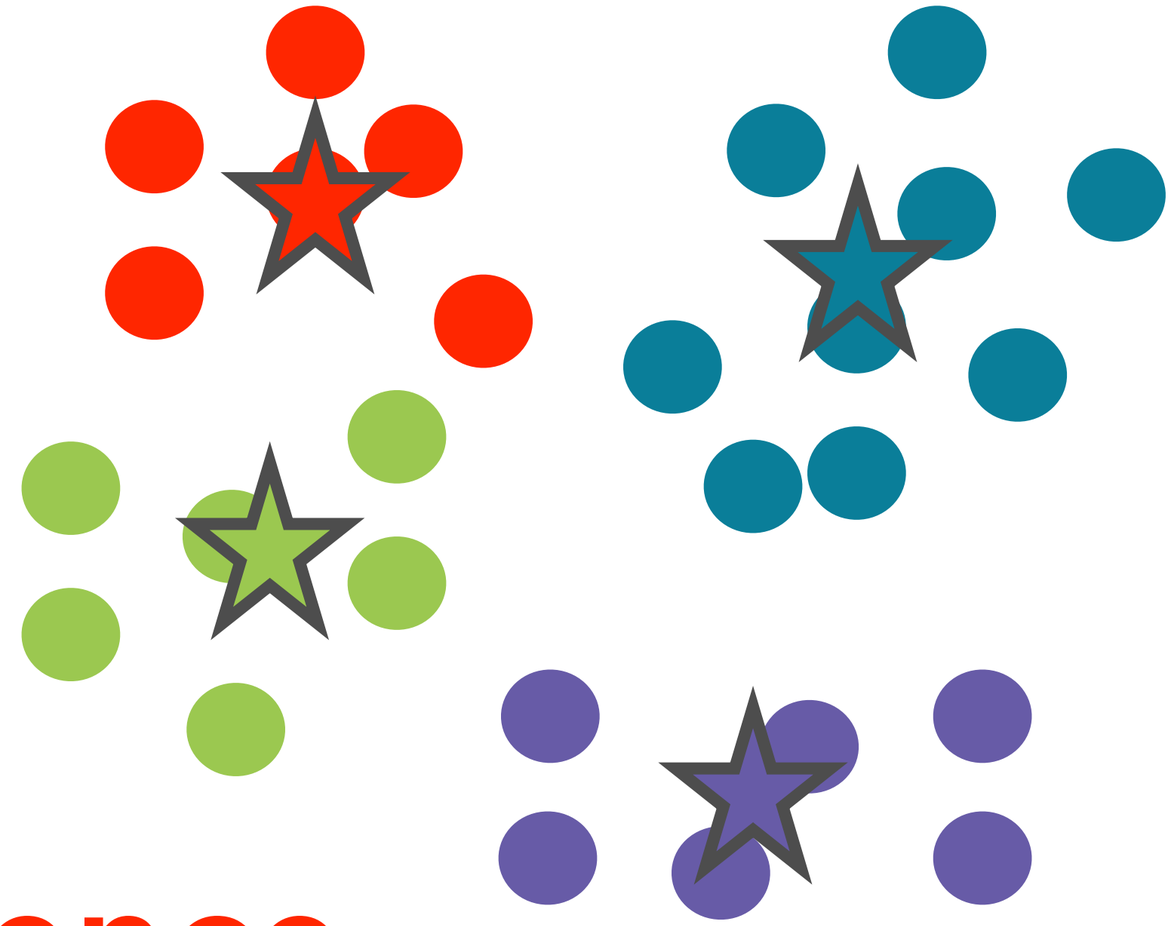


K-means Clustering

Iterate until points are in their final clusters

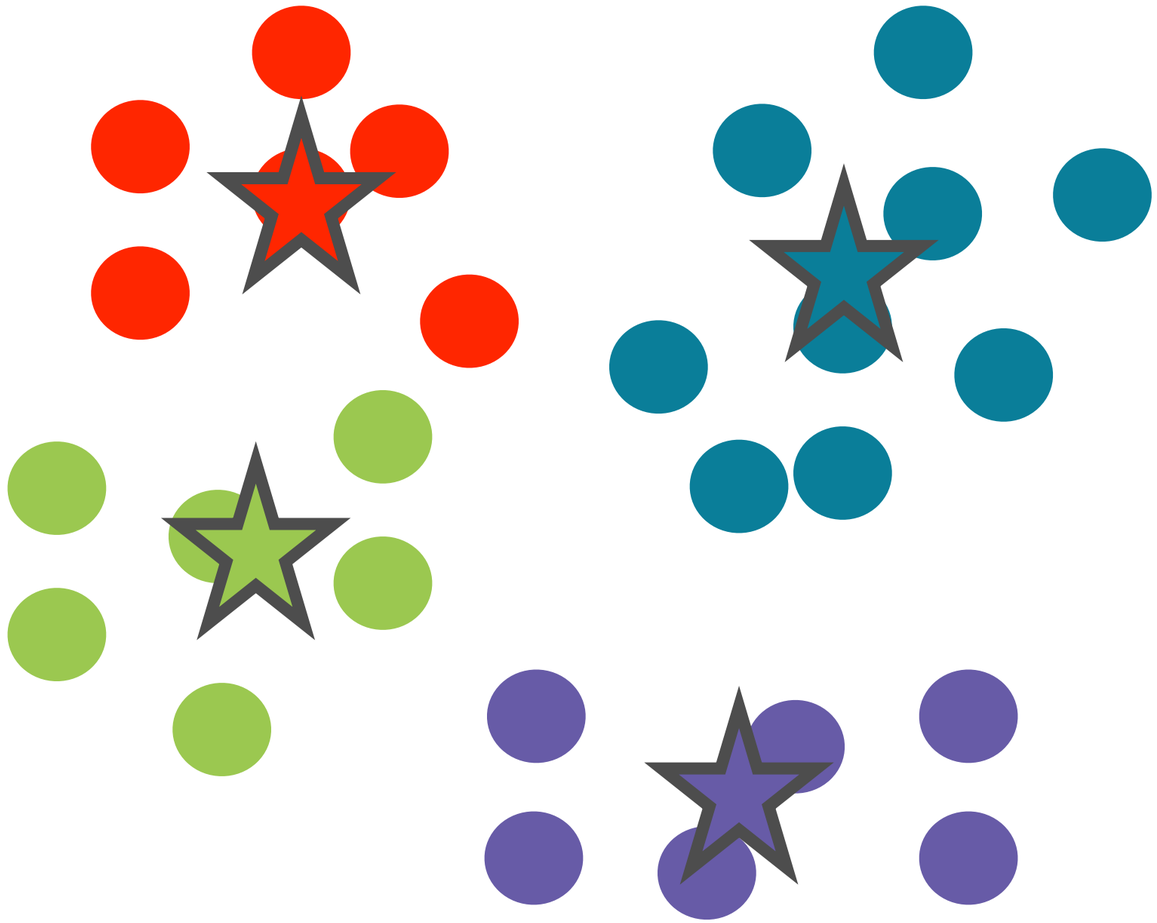


K-means Clustering

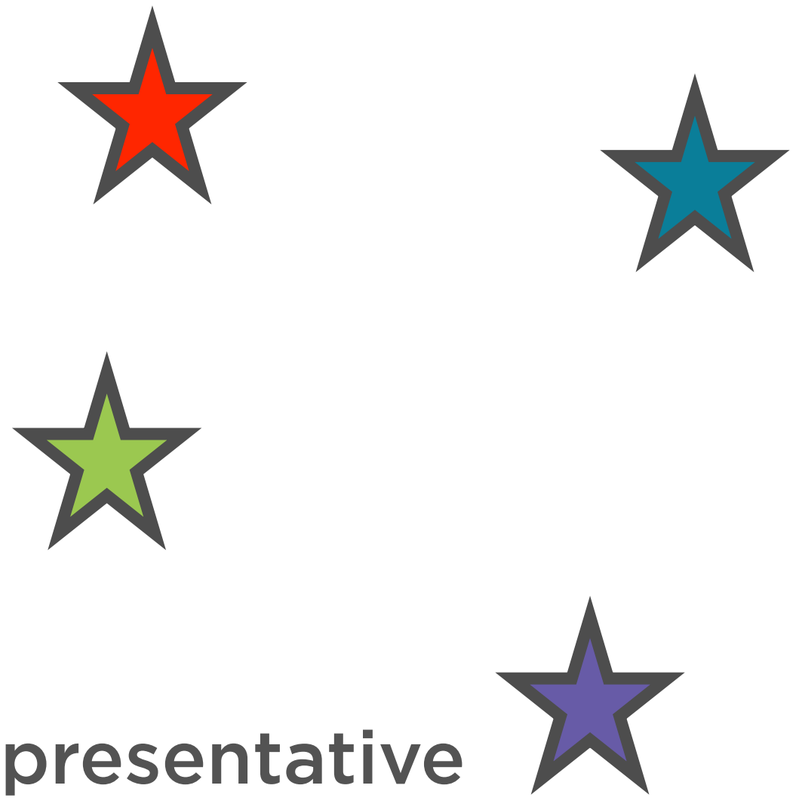


Convergence

K-means Clustering

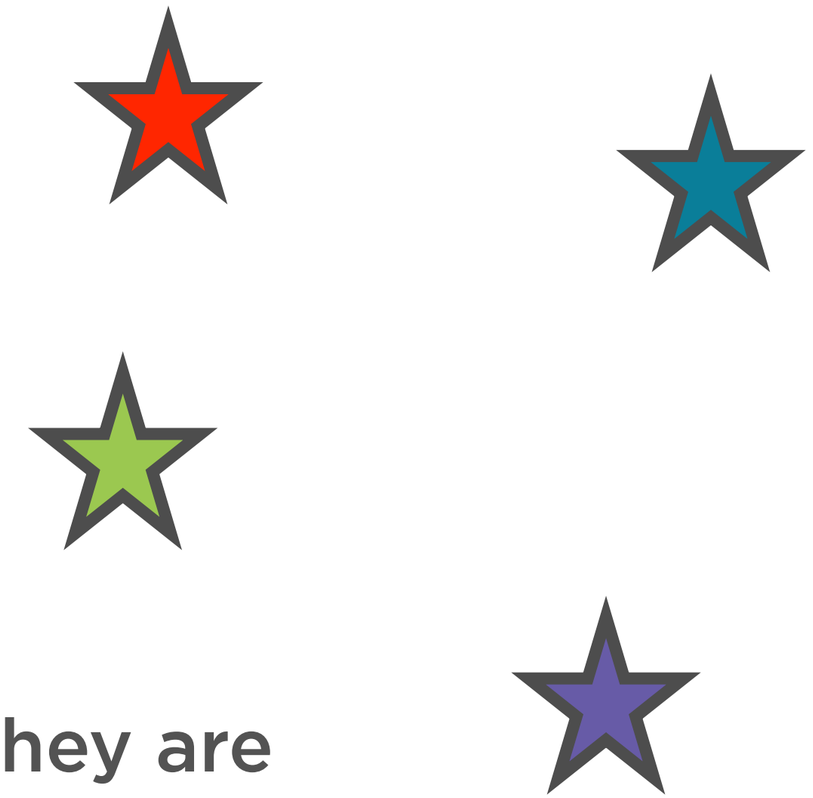


K-means Clustering



Each cluster has a representative point called a **reference vector**

K-means Clustering



Because of how they are calculated, these reference vectors are often called **centroids**

The **K-means Clustering** algorithm is a famous Machine Learning algorithm to achieve this

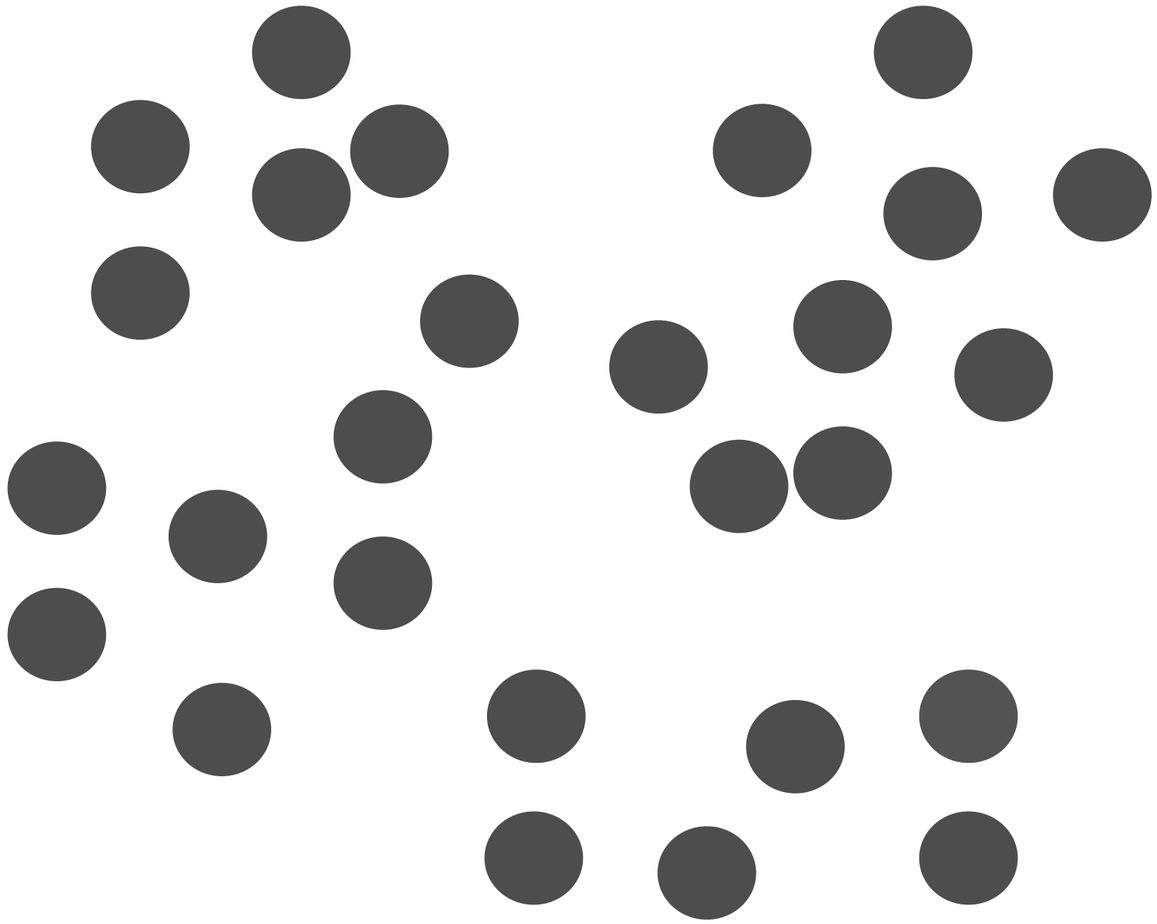
Hierarchical Clustering

Hierarchical Clustering

Group entities based on their connectivity; objects close to each other are more likely to be in the same cluster.

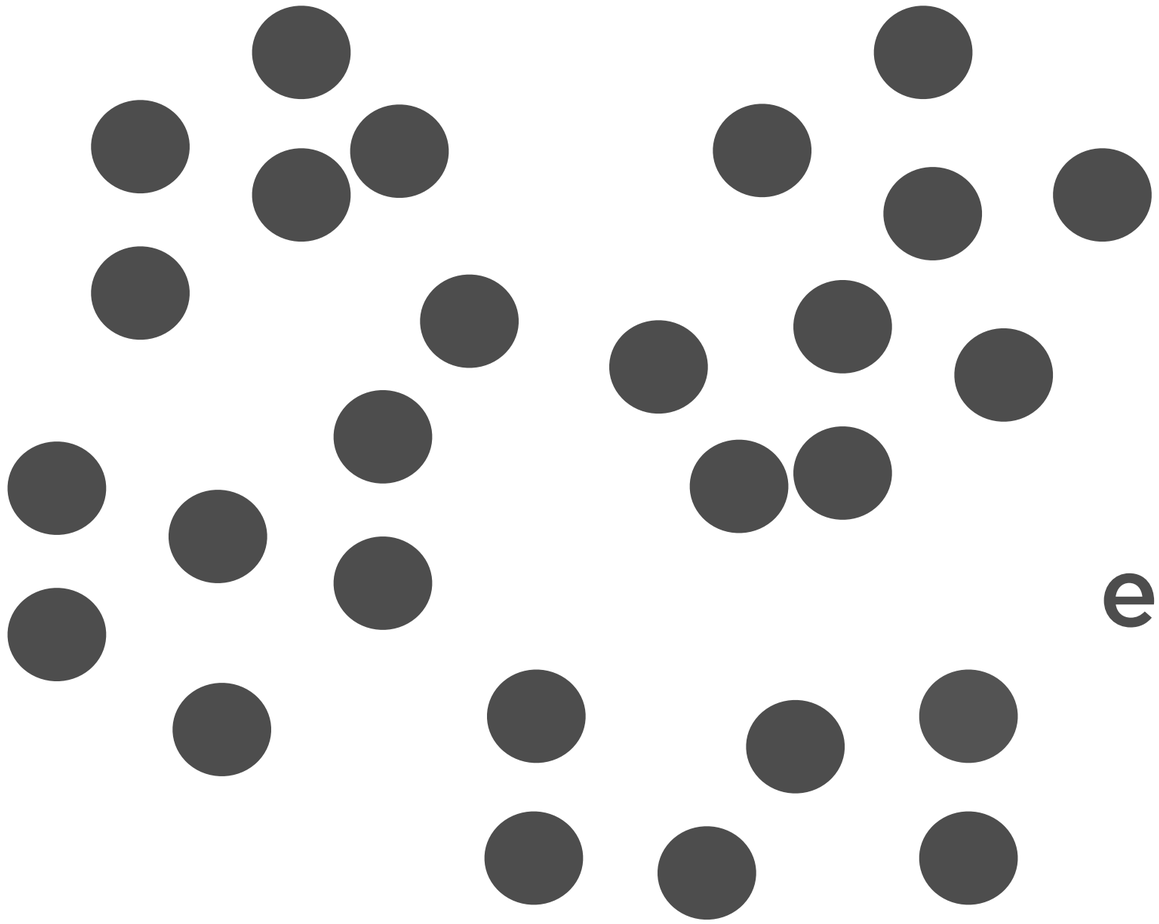
Hierarchical Clustering

Given t data points



Hierarchical Clustering

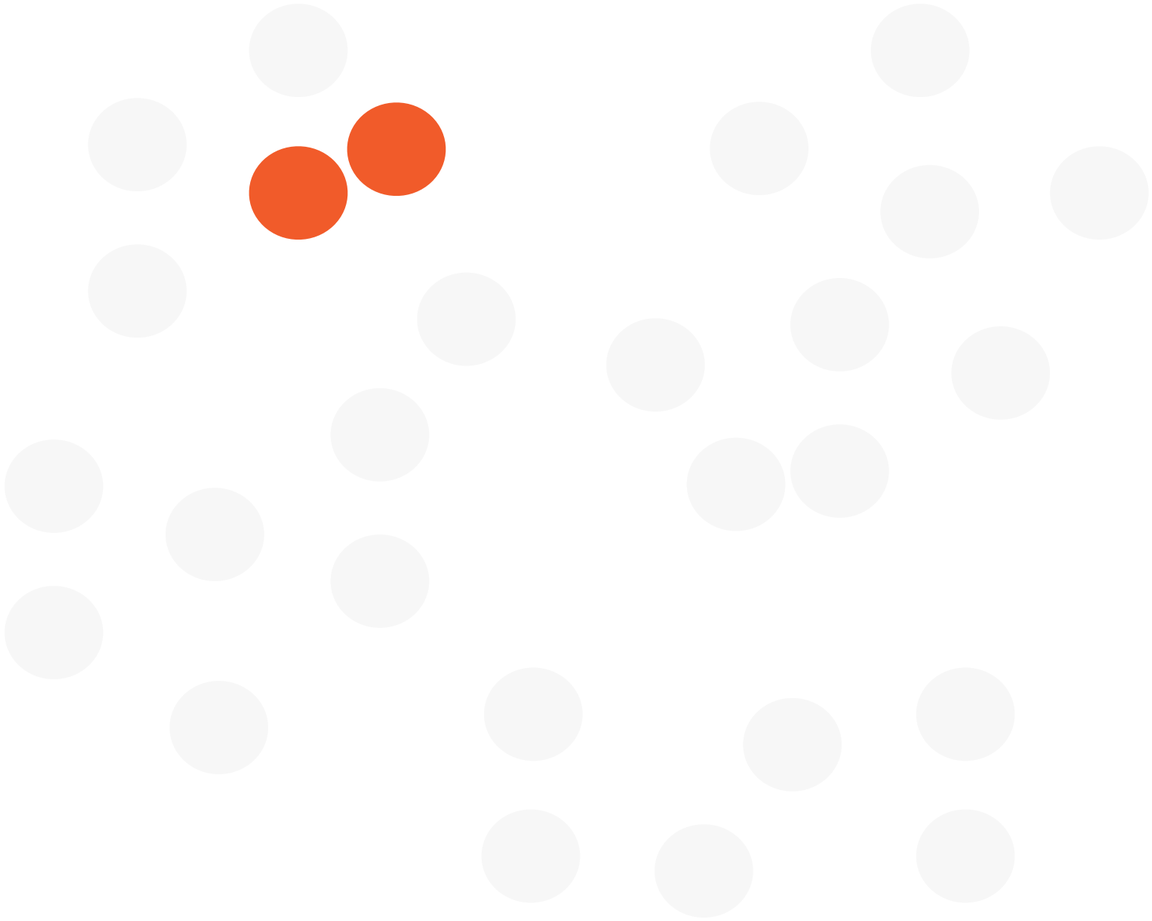
Start with t clusters, each with 1 point



t clusters, each of 1 point

Hierarchical Clustering

Merge the two clusters that are closest to each other



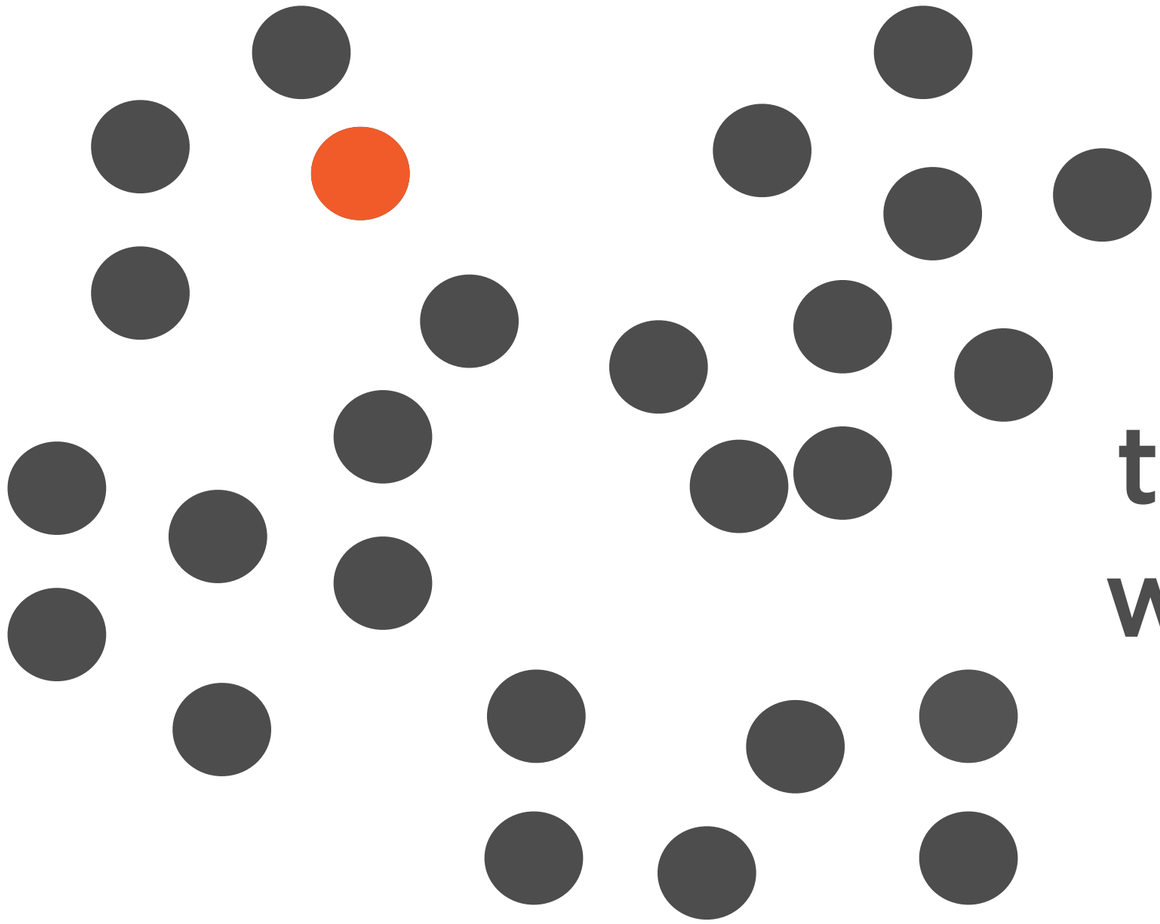
Hierarchical Clustering

Merge the two clusters that are closest to each other

t-1 clusters, 1 with 2 points

Hierarchical Clustering

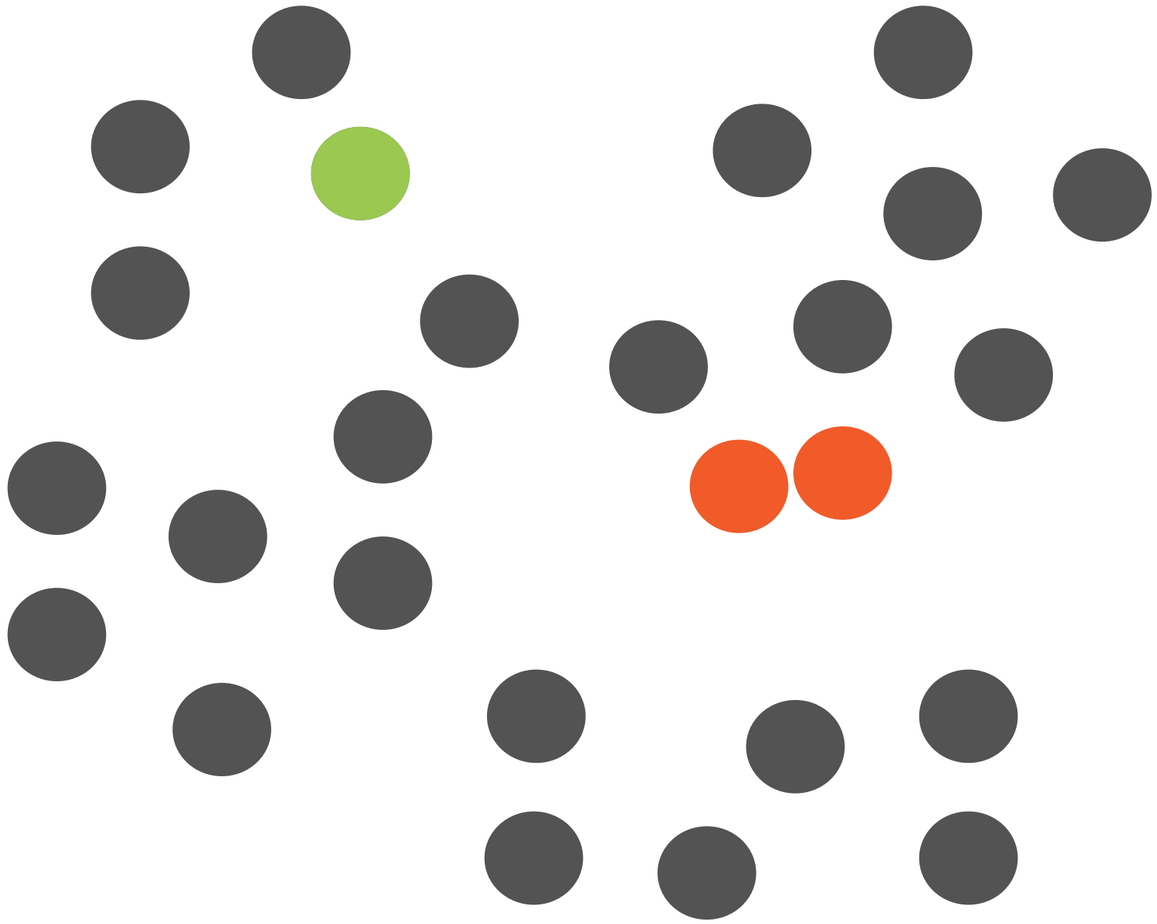
Rinse-and-repeat



t-1 clusters, 1 with 2 points

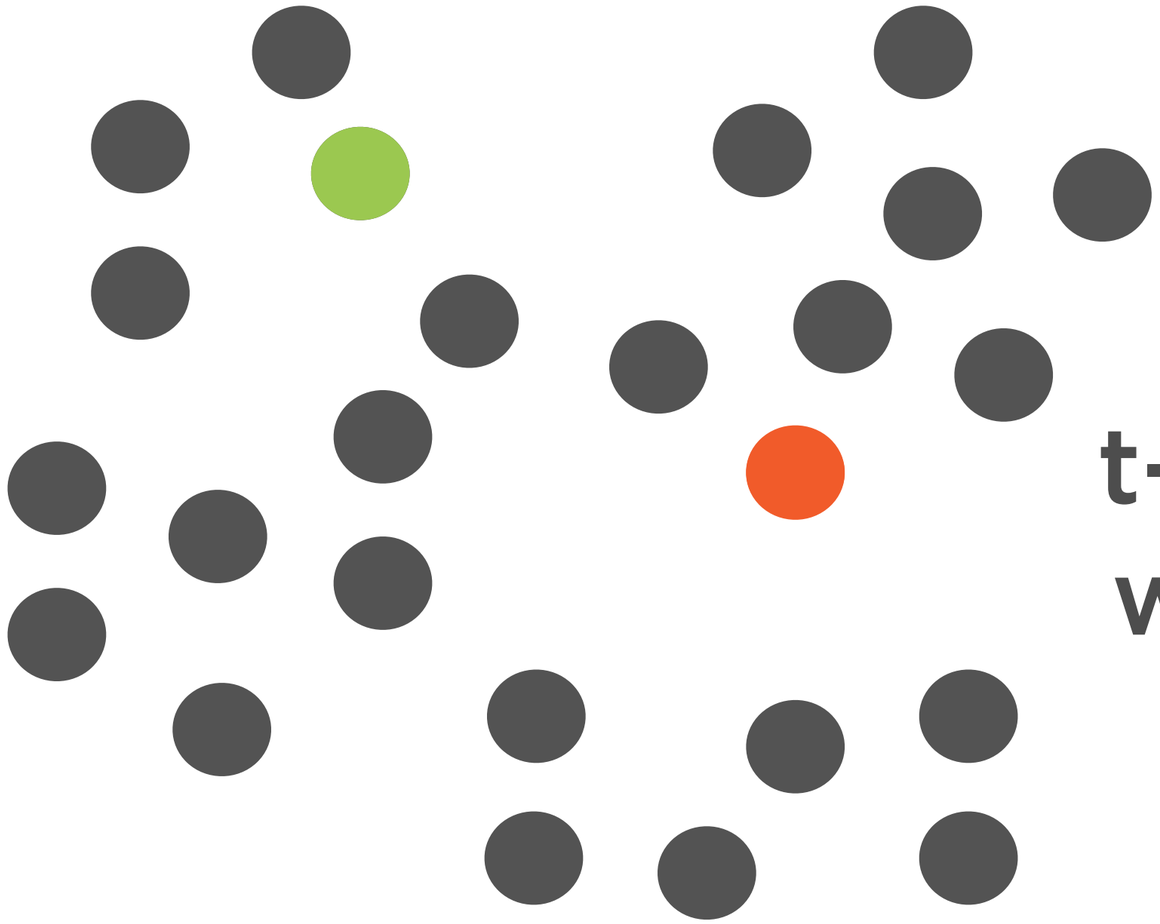
Hierarchical Clustering

Rinse-and-repeat



Hierarchical Clustering

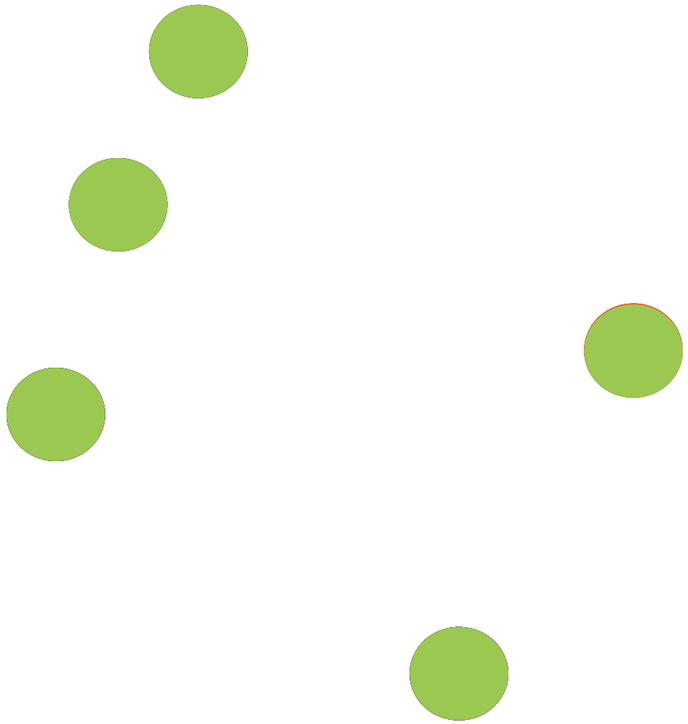
Rinse-and-repeat



t-2 clusters, 2 with 2 points

Hierarchical Clustering

Rinse-and-repeat



6 clusters, each with multiple points

Hierarchical Clustering

The number of clusters keeps reducing

2 clusters, each with multiple points

Hierarchical Clustering

The number of clusters keeps reducing

1 cluster, with all t points

Hierarchical Clustering

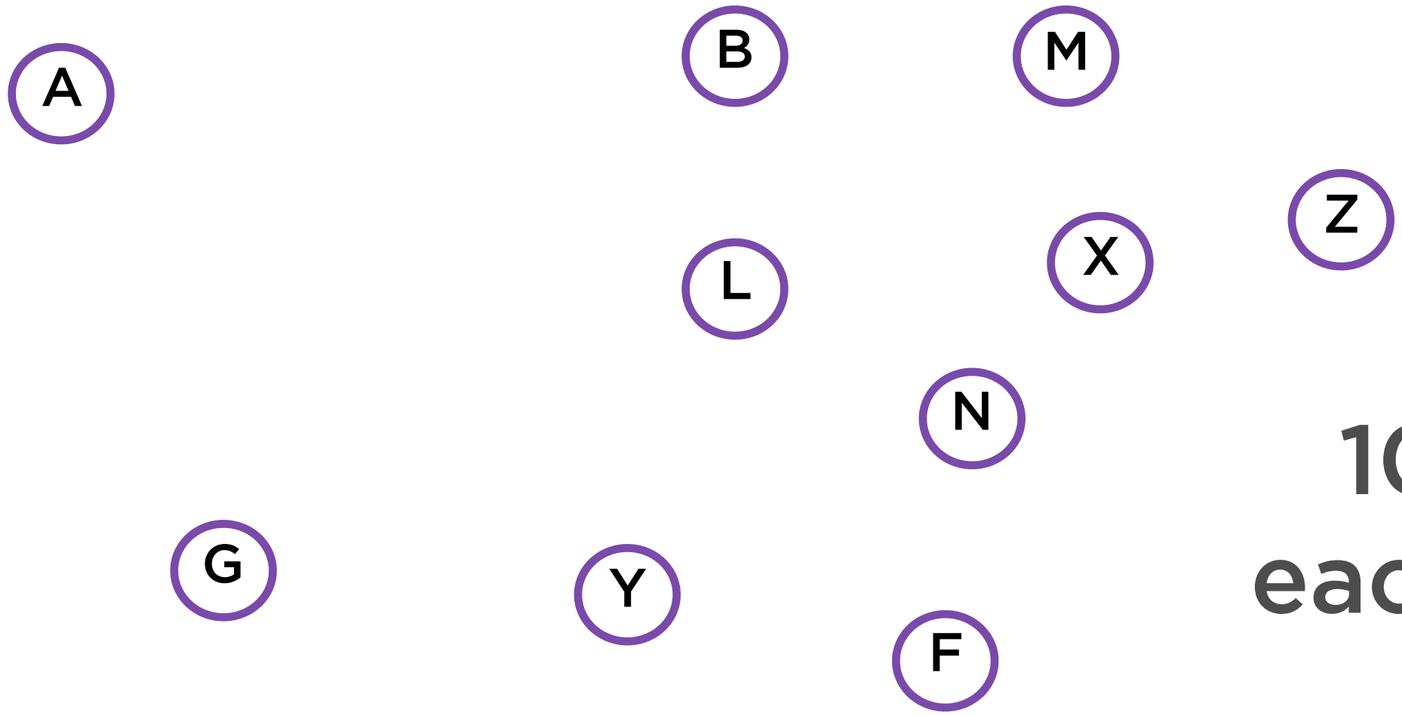
Until just 1 cluster remains

1 cluster, with all t points

Dendrogram

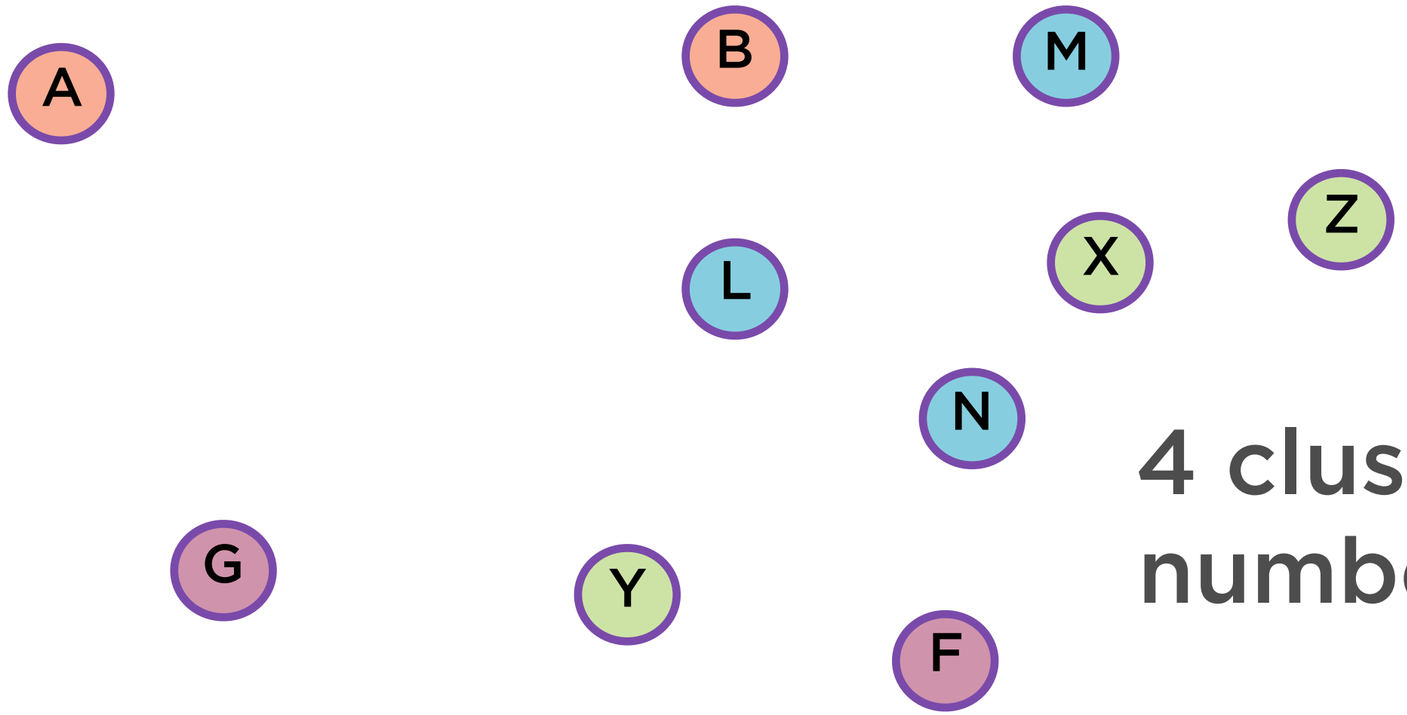
A tree diagram used to illustrate the arrangement of the clusters produced by hierarchical clustering

Dendrogram



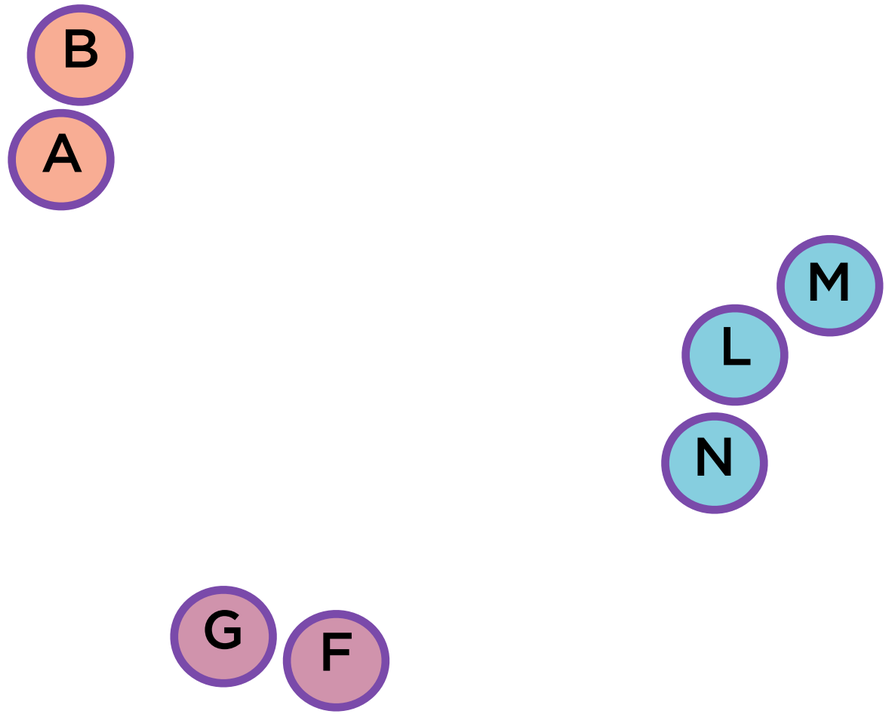
10 clusters,
each of 1 point

Dendrogram



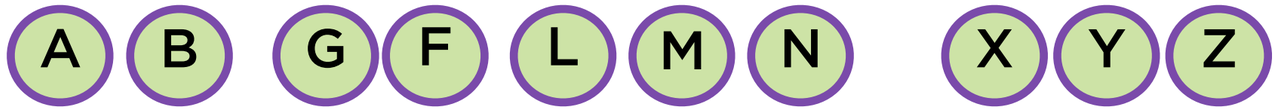
4 clusters, varying numbers of points

Dendrogram



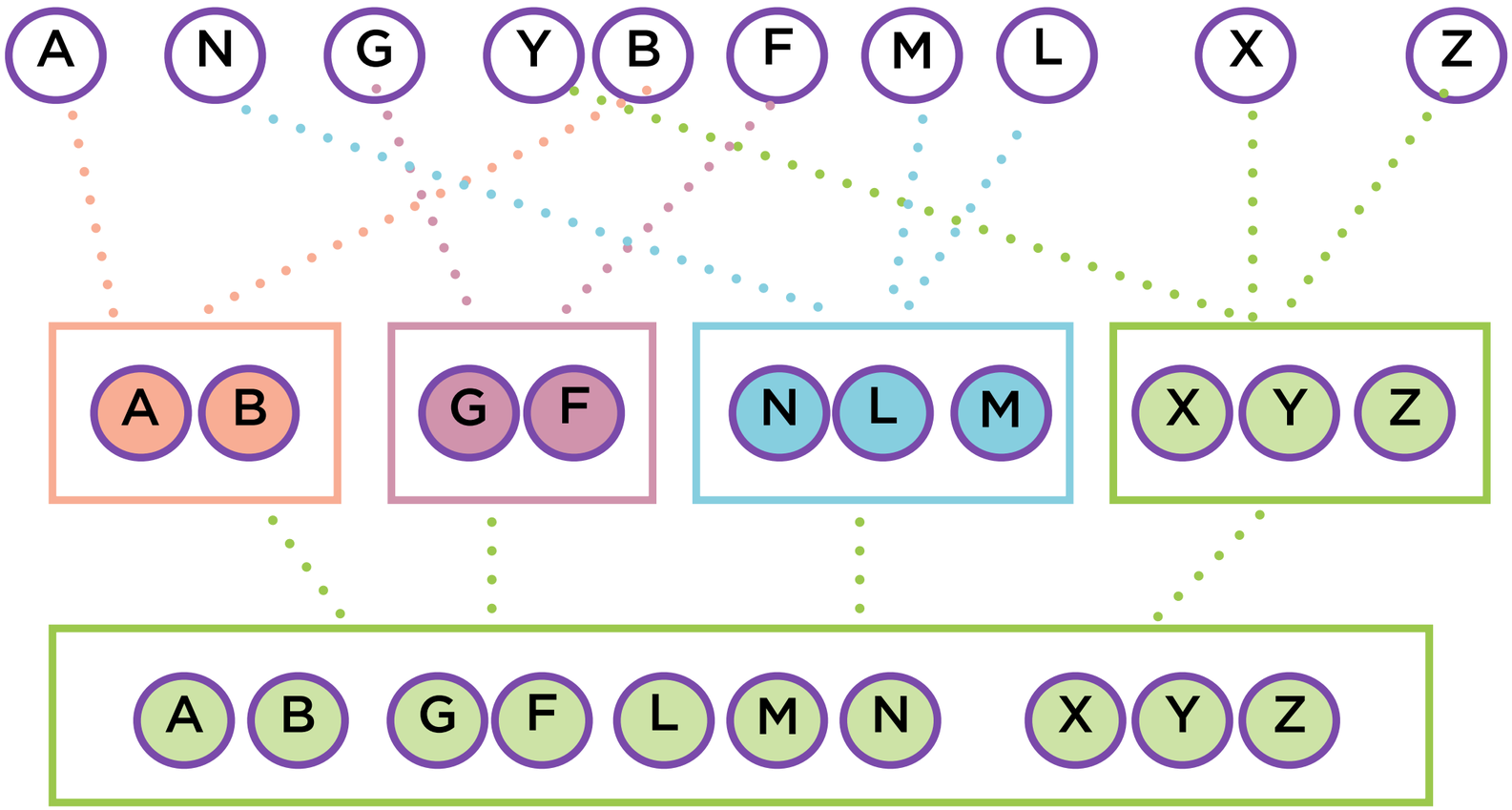
4 clusters, varying numbers of points

Dendrogram



1 clusters, all 10 points

Dendrogram

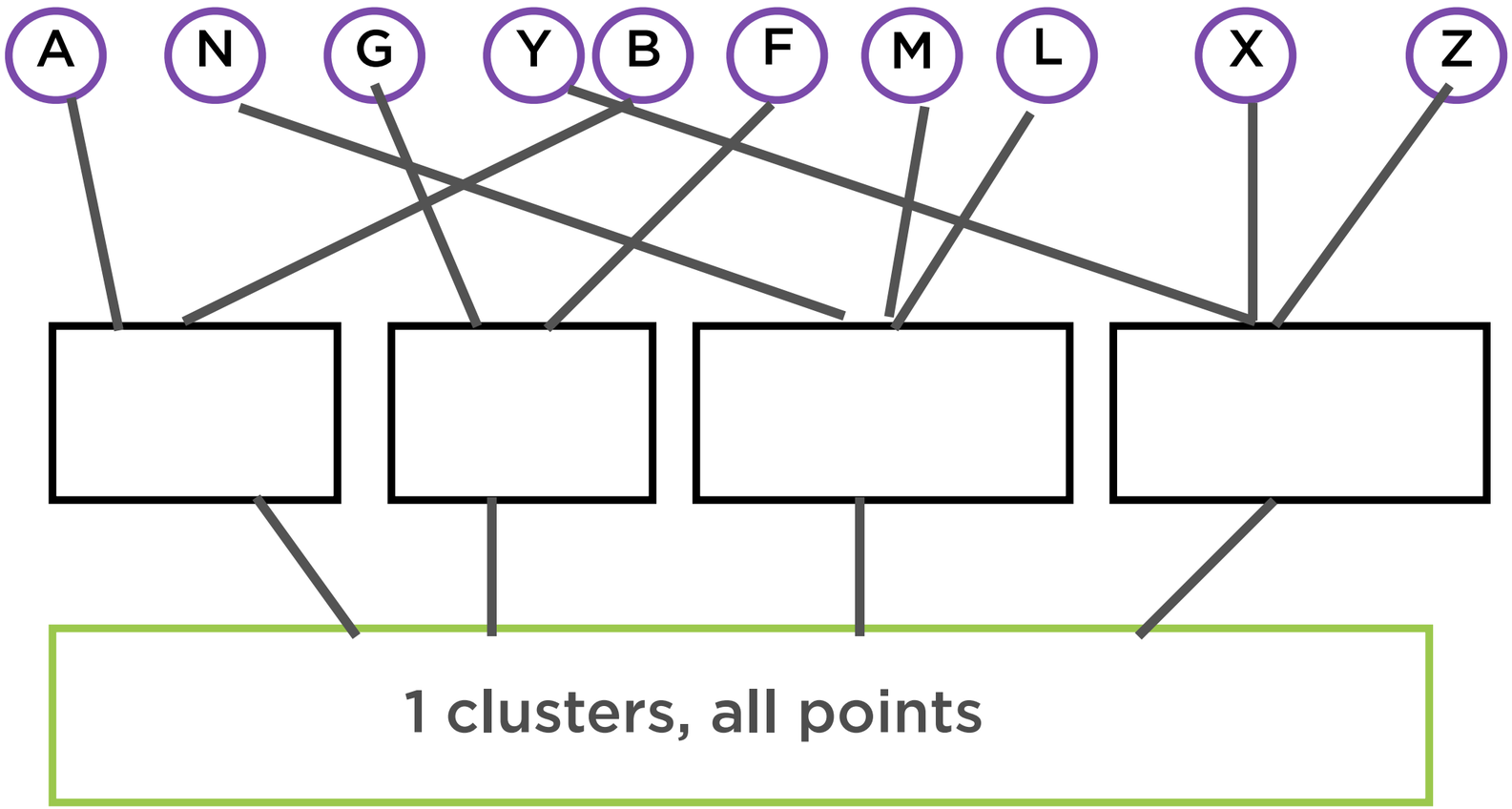


10 clusters, each of 1 point

4 clusters, varying numbers of points

1 clusters, all 10 points

Dendrogram



10 clusters, each of 1 point

Now, easy to vary number of clusters

Hierarchical Clustering

Agglomerative - start with many 1-point clusters, end with 1 big cluster

Divisive - start with 1 big cluster, end with many 1-point clusters

Demo

Performing k-means clustering on unlabeled data

Demo

Performing k-means clustering on labeled data

Demo

Performing agglomerative clustering on image data

Summary

Clustering as a form of unsupervised machine learning

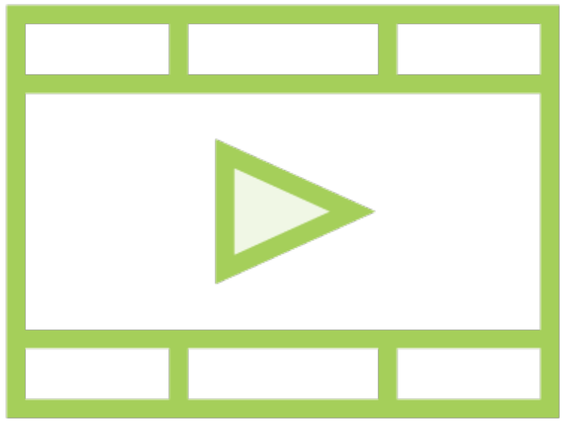
Different families of clustering algorithms

Choosing the right clustering algorithm

K-means clustering

Hierarchical clustering

Related Courses



Building Regression Models with scikit-learn

Building Classification Models with scikit-learn

Building Clustering Models with scikit-learn