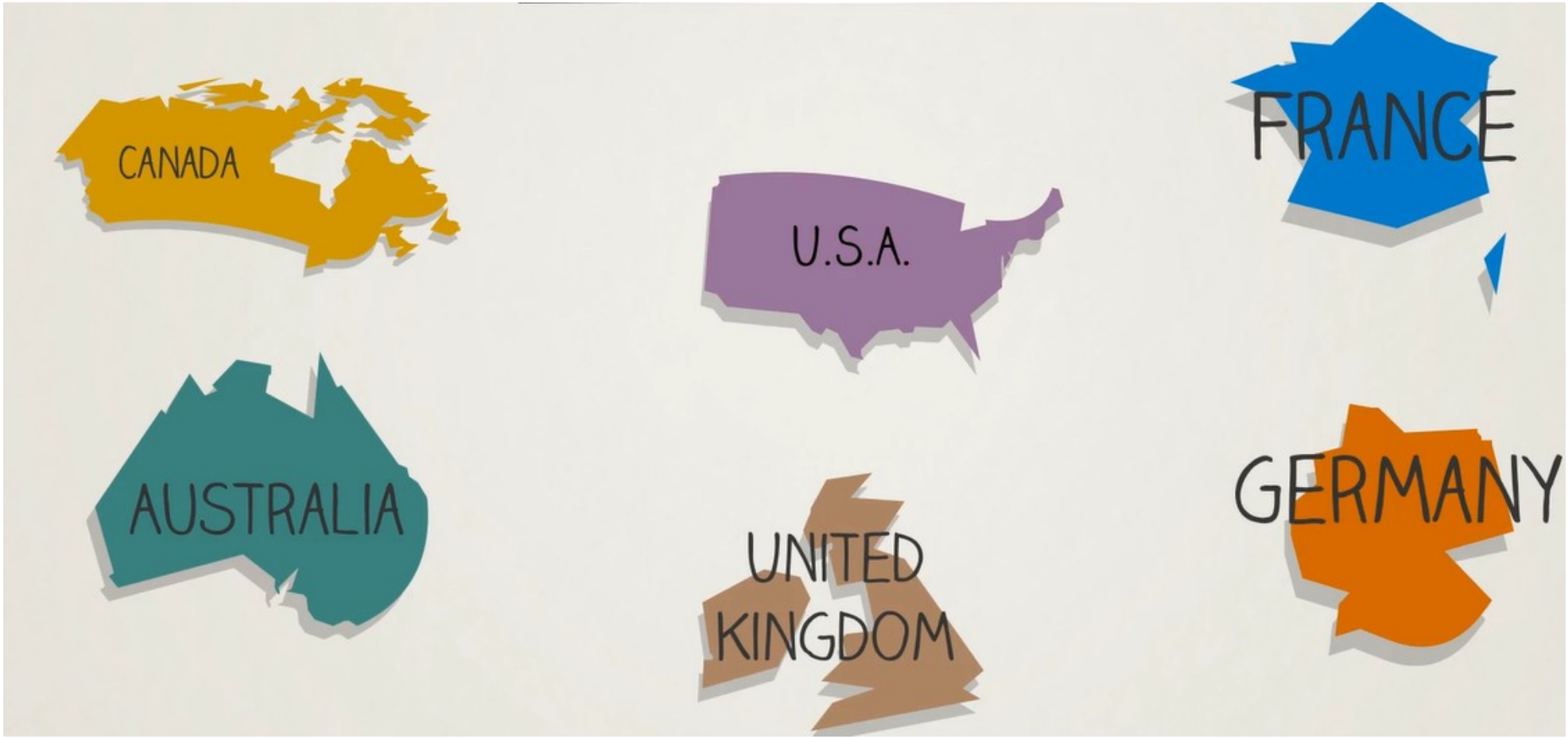
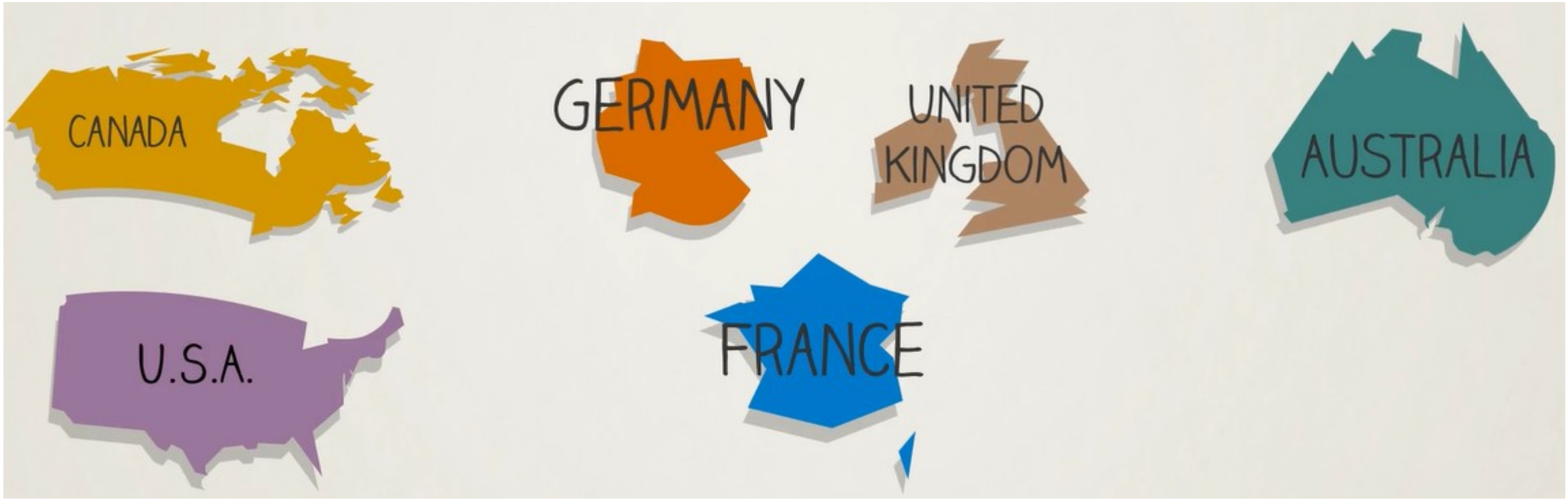
# Cluster Analysis

- This is a fundamentally different data science method.
- Technically speaking cluster analysis is a multivariate statistical technique that groups observations on the basis some of their features or variables that are described by.
- Intuitively speaking, observations in a dataset can be divided into different groups and sometimes this is very useful

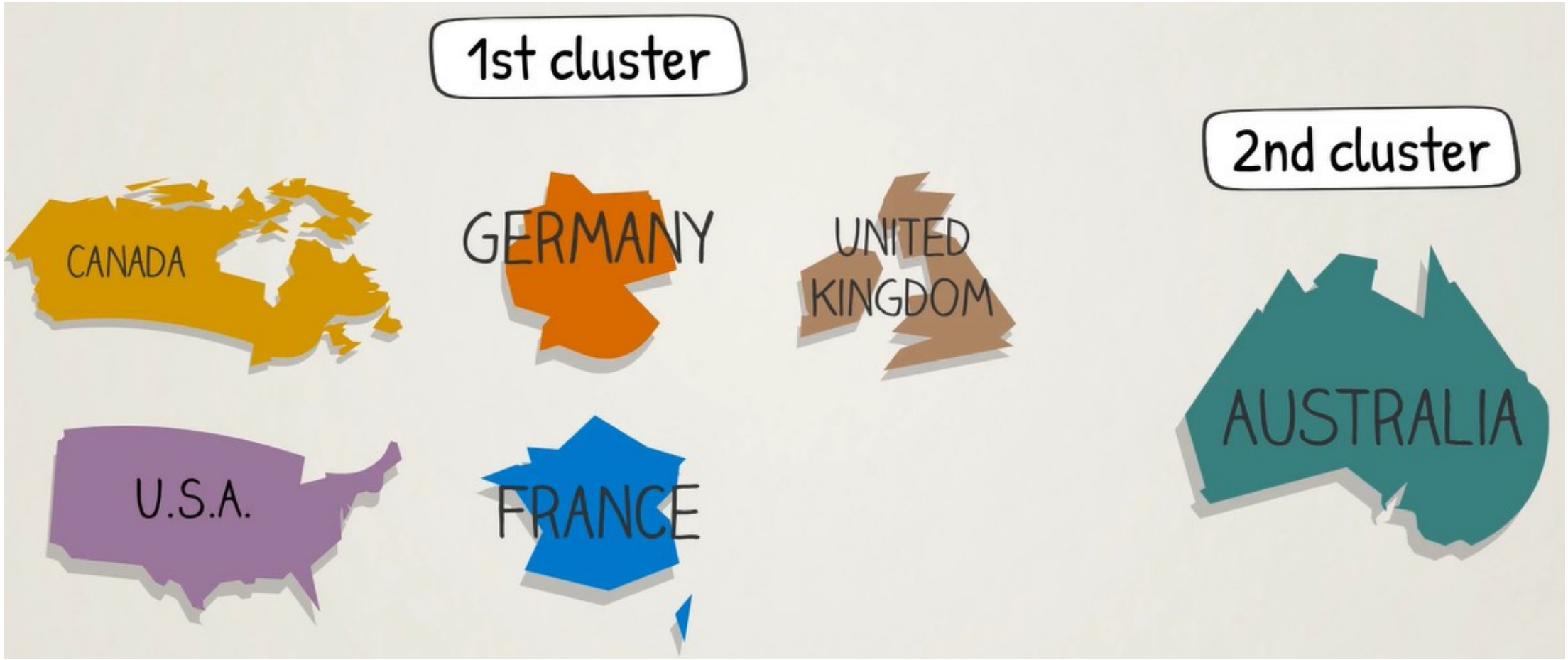
# Example



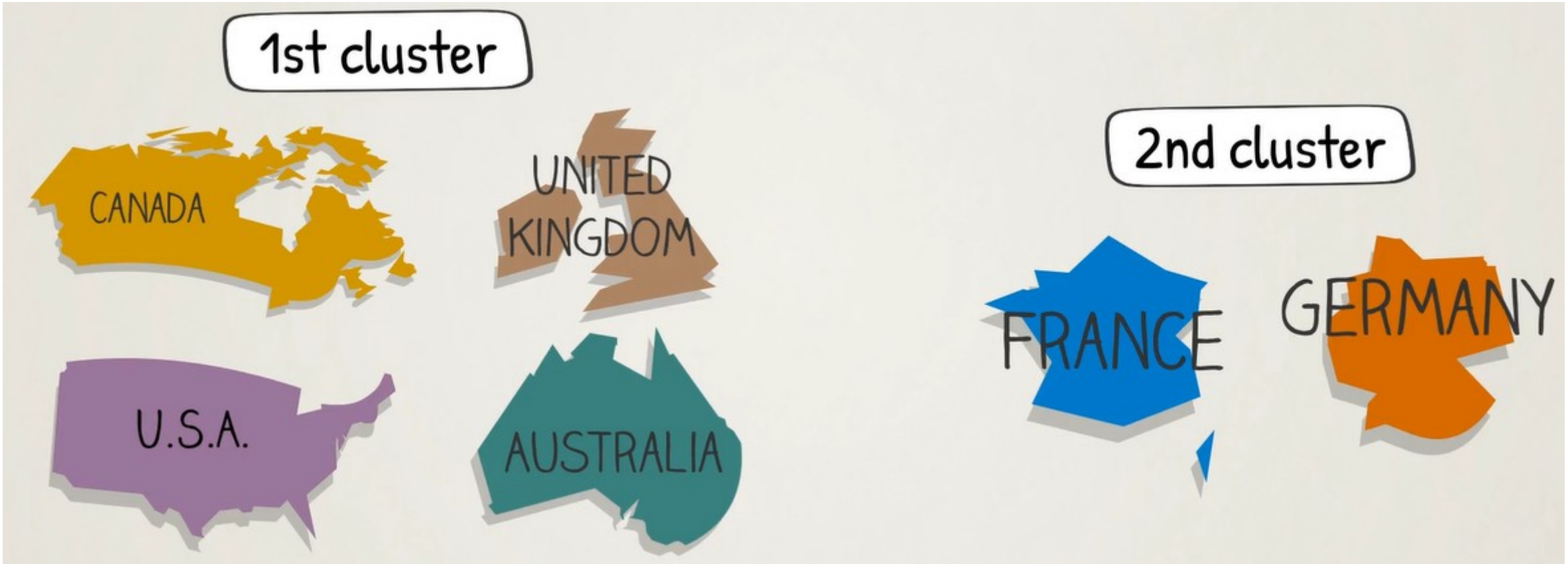
# Example



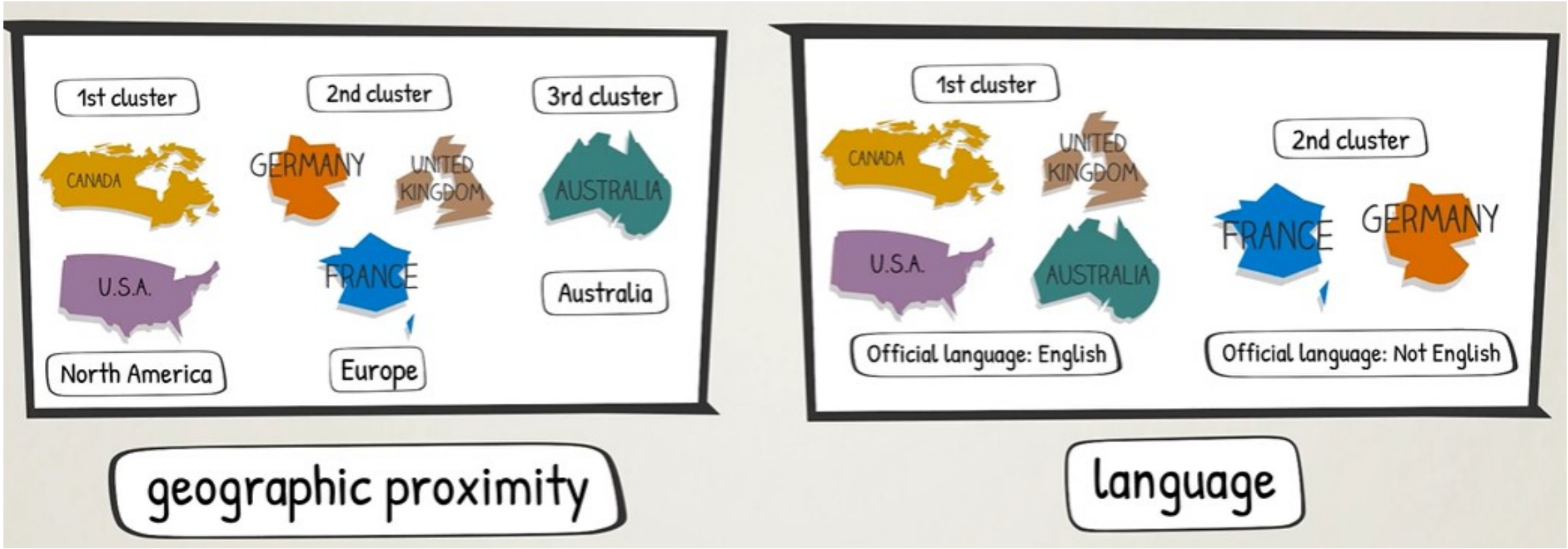
# Example


1st cluster
2nd cluster
CANADA
GERMANY
UNITED KINGDOM
AUSTRALIA
U.S.A.
FRANCE

# Example


1st cluster
CANADA
UNITED KINGDOM
U.S.A.
AUSTRALIA
2nd cluster
FRANCE
GERMANY

# Example


1st cluster
2nd cluster
3rd cluster
CANADA
GERMANY
UNITED KINGDOM
AUSTRALIA
U.S.A.
FRANCE
Australia
North America
Europe
geographic proximity
1st cluster
CANADA
UNITED KINGDOM
2nd cluster
FRANCE
GERMANY
U.S.A.
AUSTRALIA
Official language: English
Official language: Not English
language

# Cluster Analysis – Final Goal

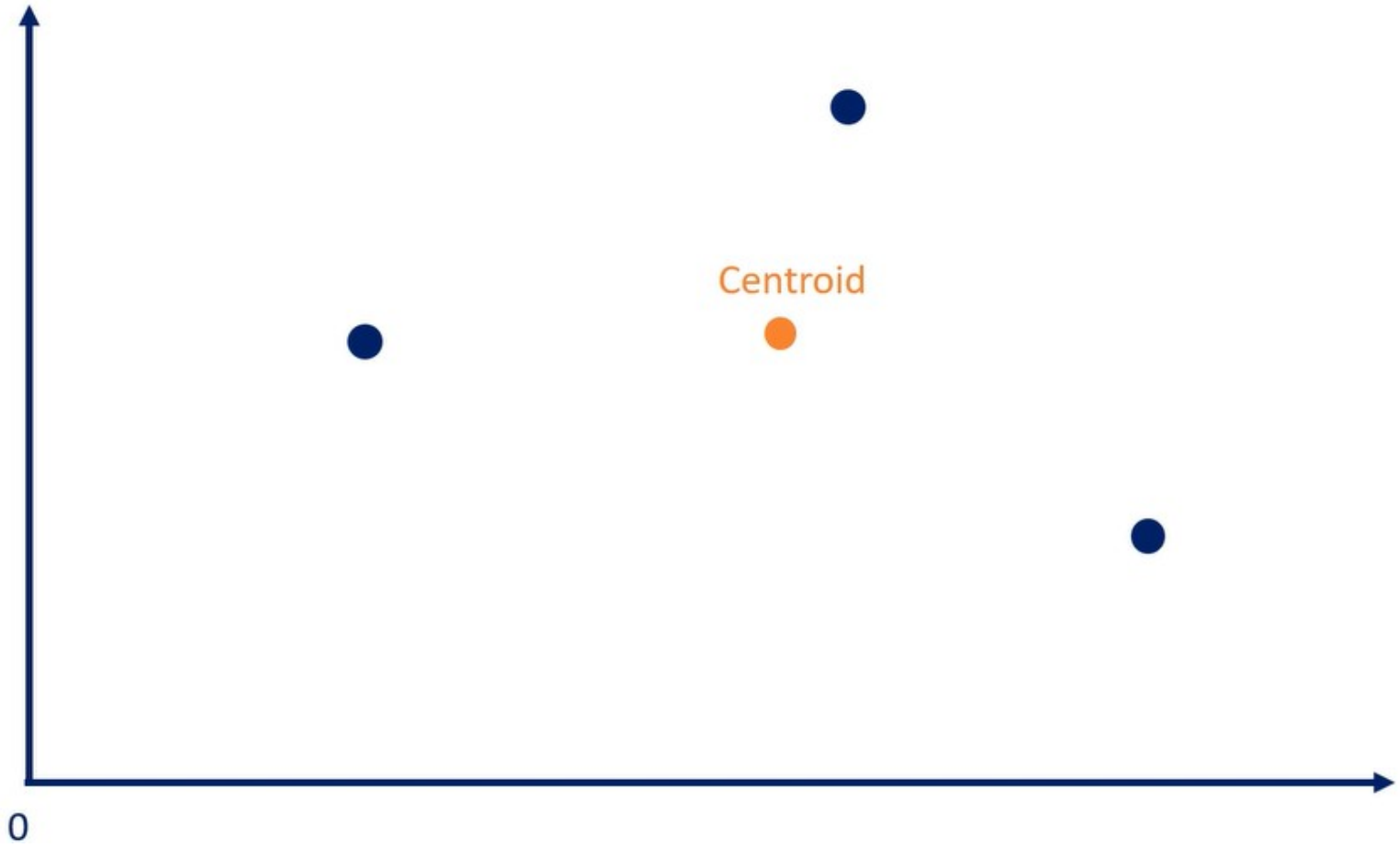
- The goal of clustering is to maximize the similarity of observations within a cluster and maximize the dissimilarity between clusters

# Cluster Analysis – agenda

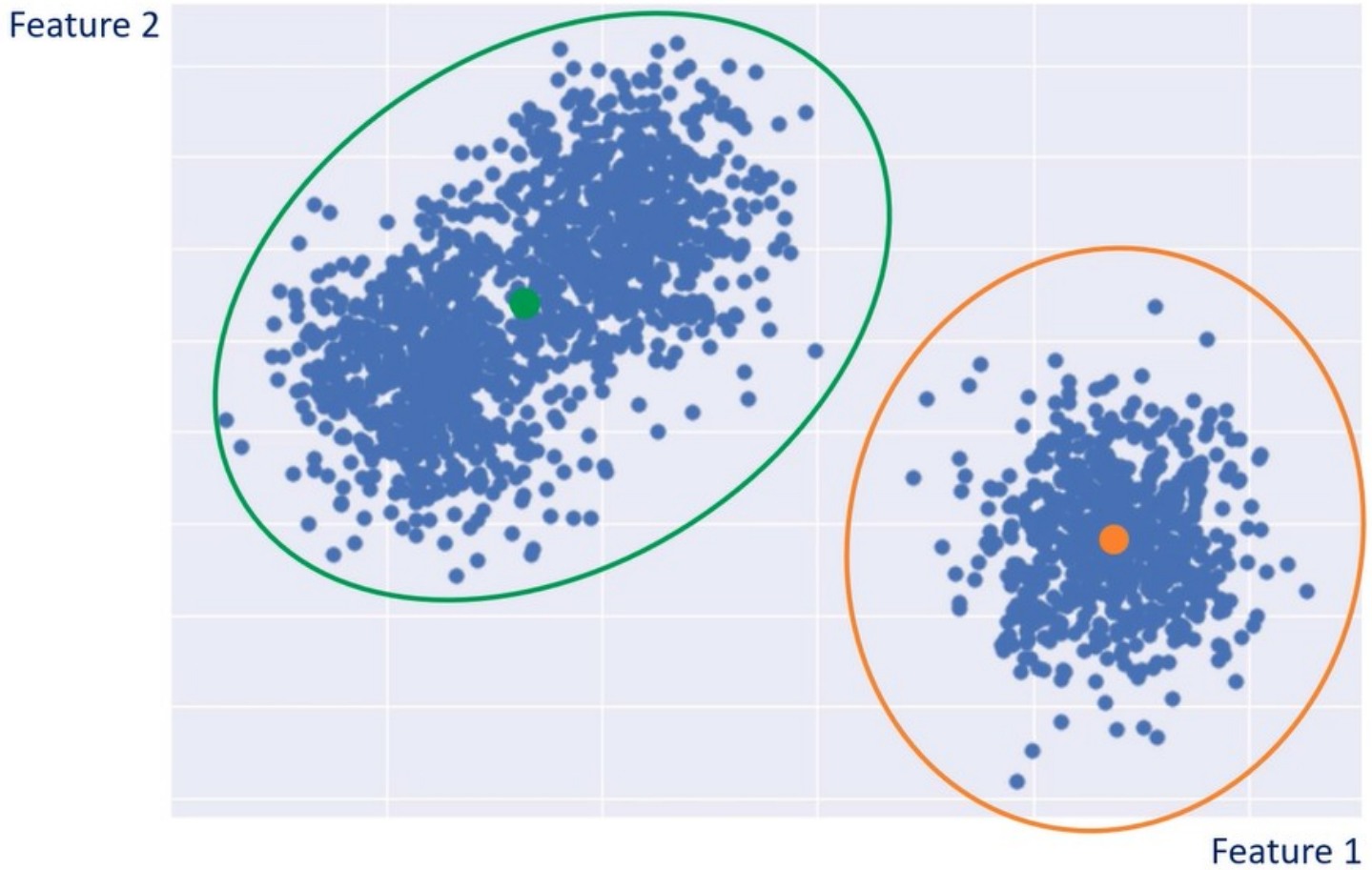
- Perform a clustering problem
- How to find the optimal number of clusters
- How to identify appropriate features
- How to interpret results

# What is a centroid?

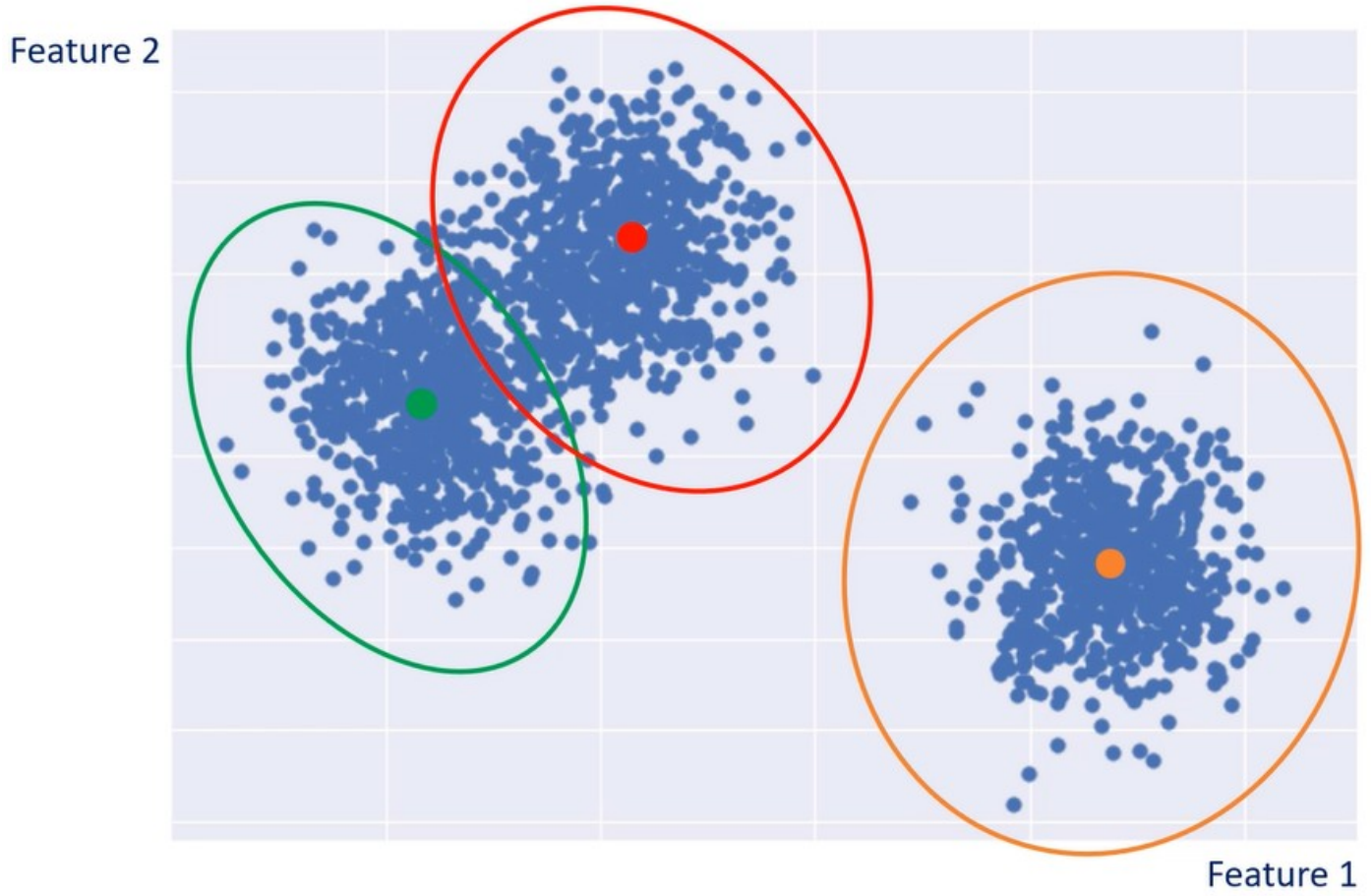
- A centroid is the mean position of a group of points (aka center of mass)



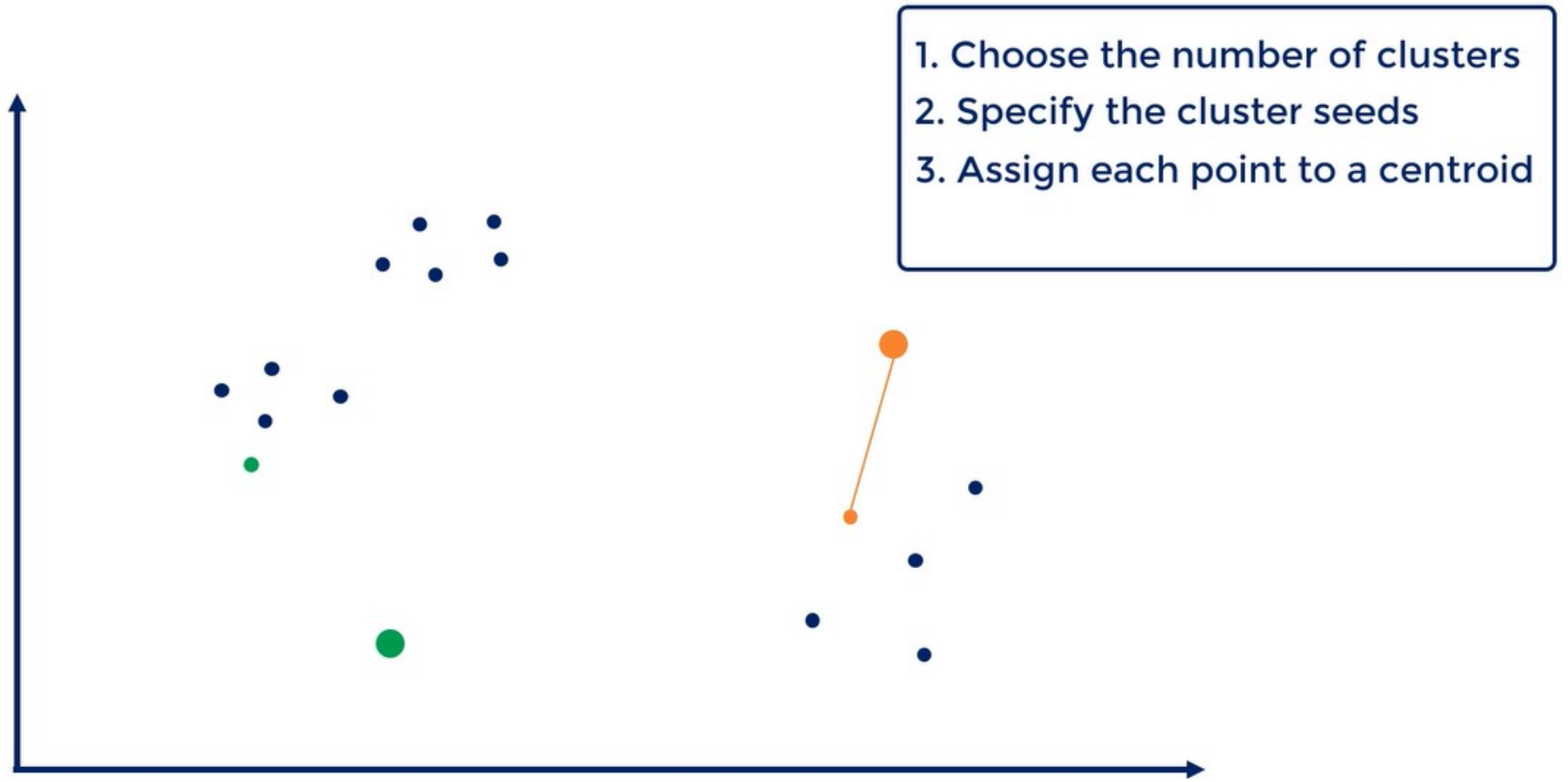
# Clusters



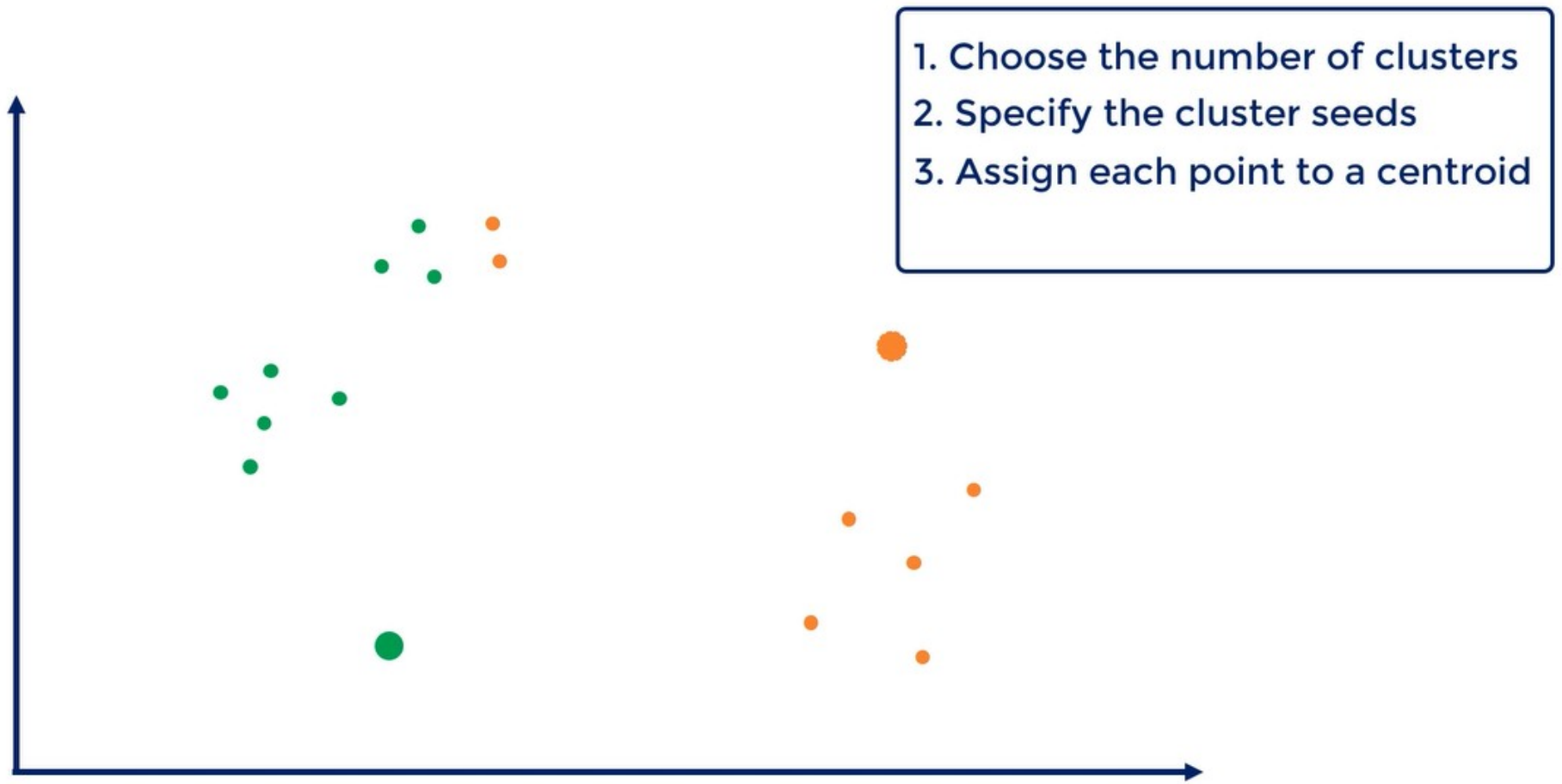
# Clusters



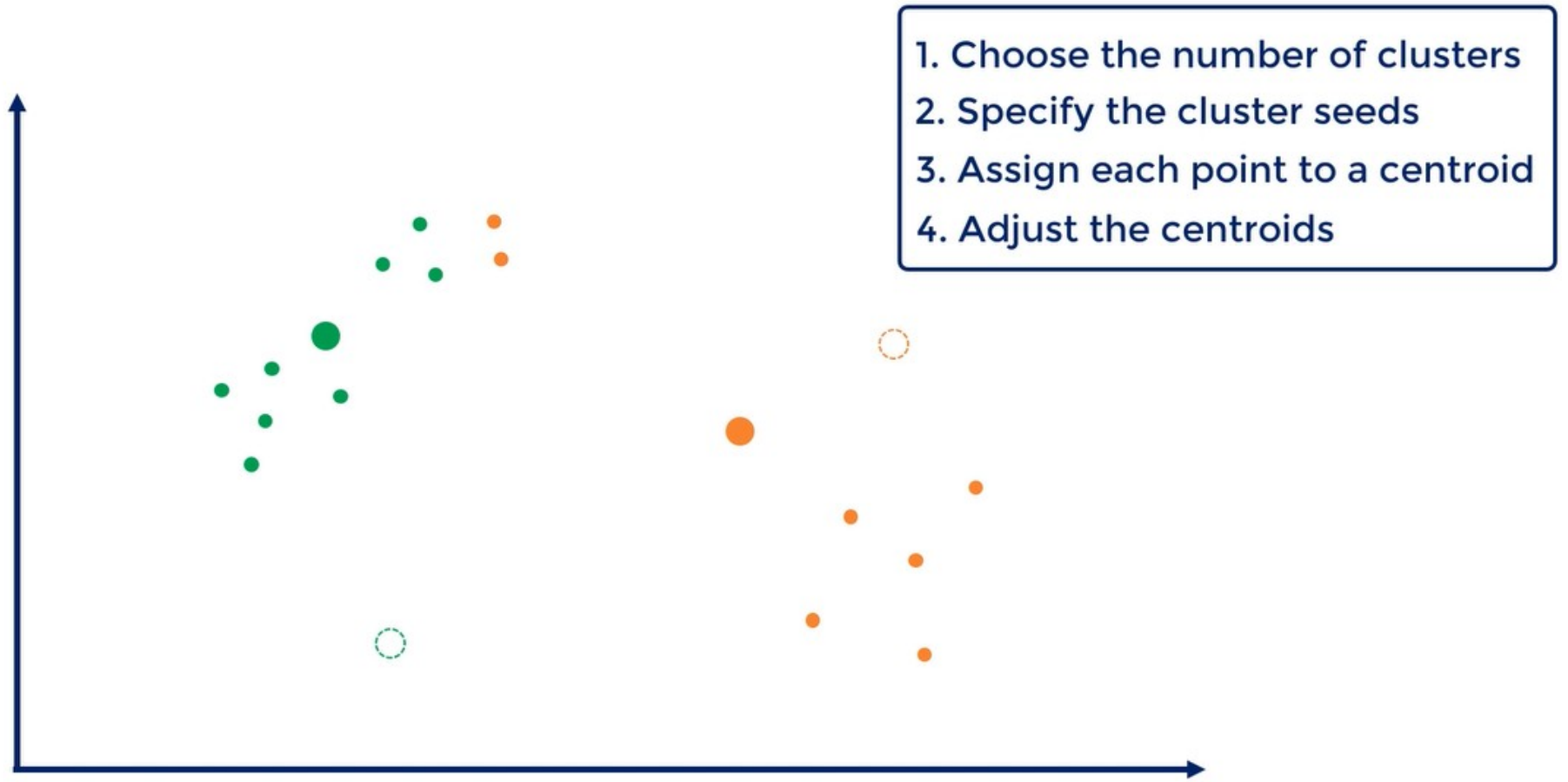
# K-Means Clustering



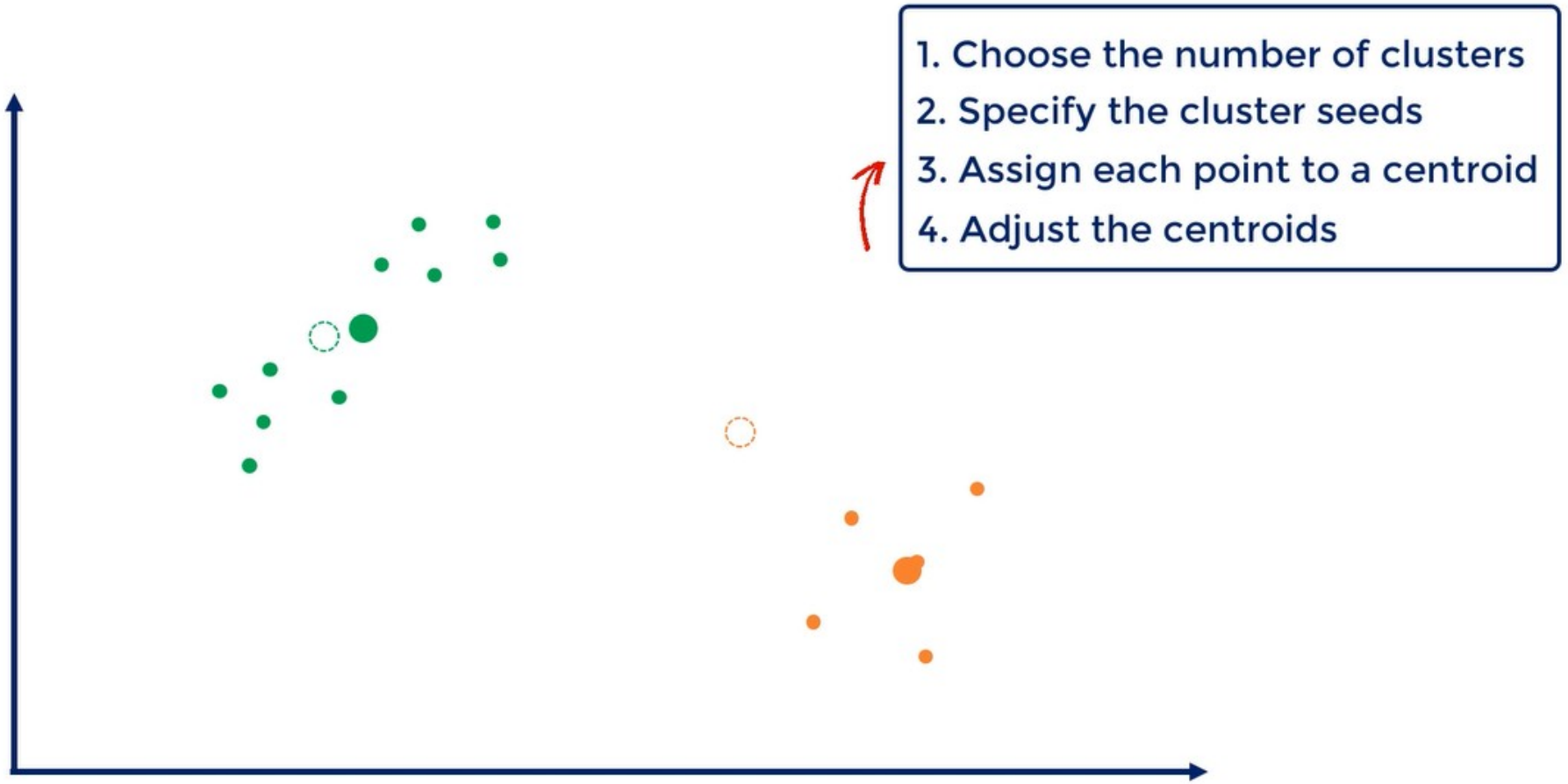
# K-Means Clustering



# K-Means Clustering



# K-Means Clustering



# K-Means Clustering