Cluster analysis

Cluster analysis

Cluster analysis is a multivariate statistical technique that groups observations on the basis some of their features or variables that they are described by.

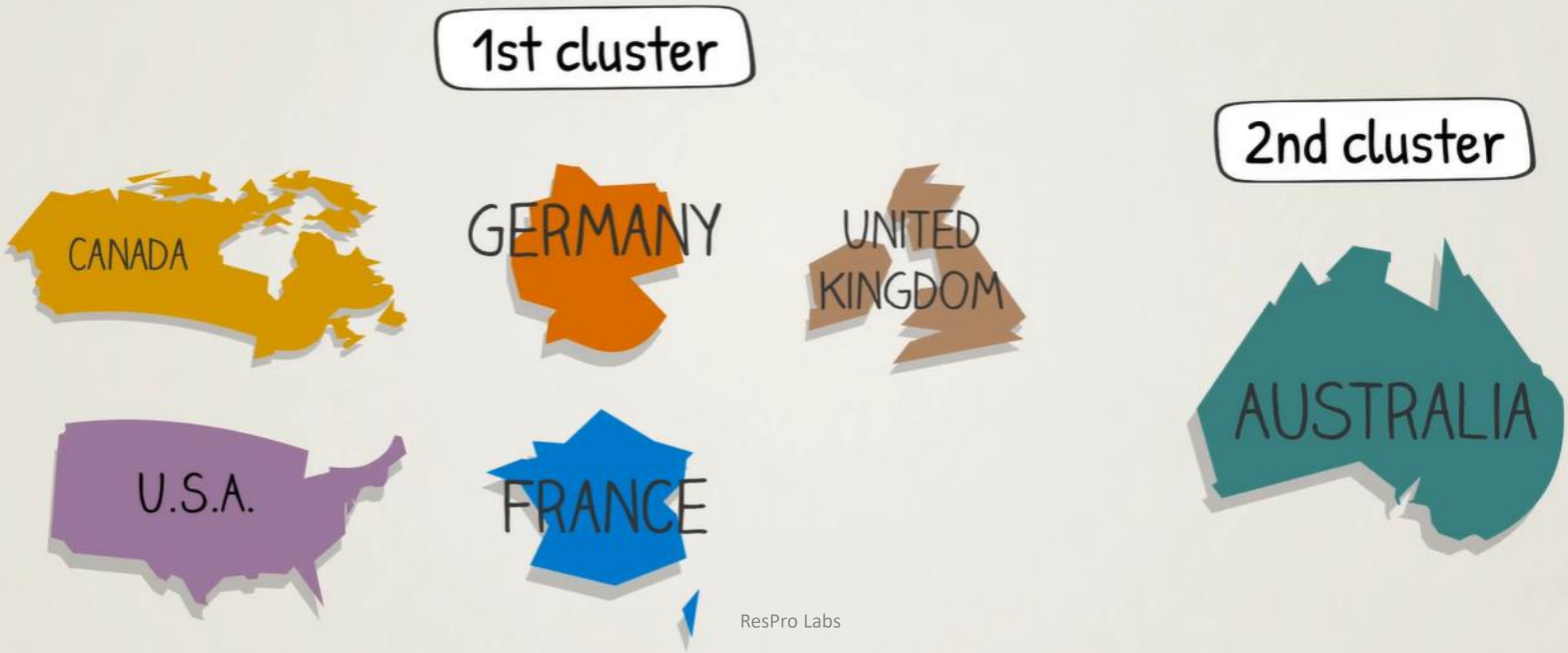
The goal of clustering is to maximize the similarity of observations within a cluster and maximize the dissimilarity between clusters.

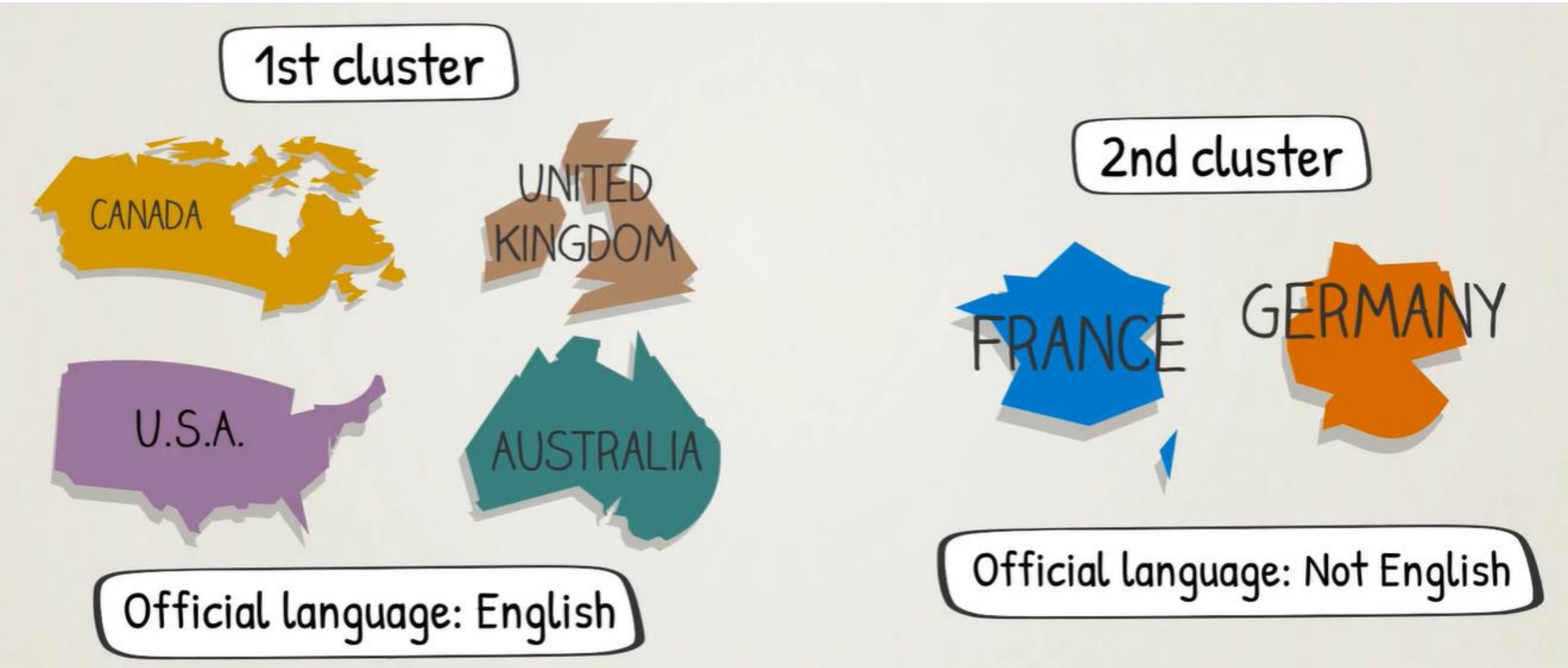
EXAMPLE

EXAMPLE

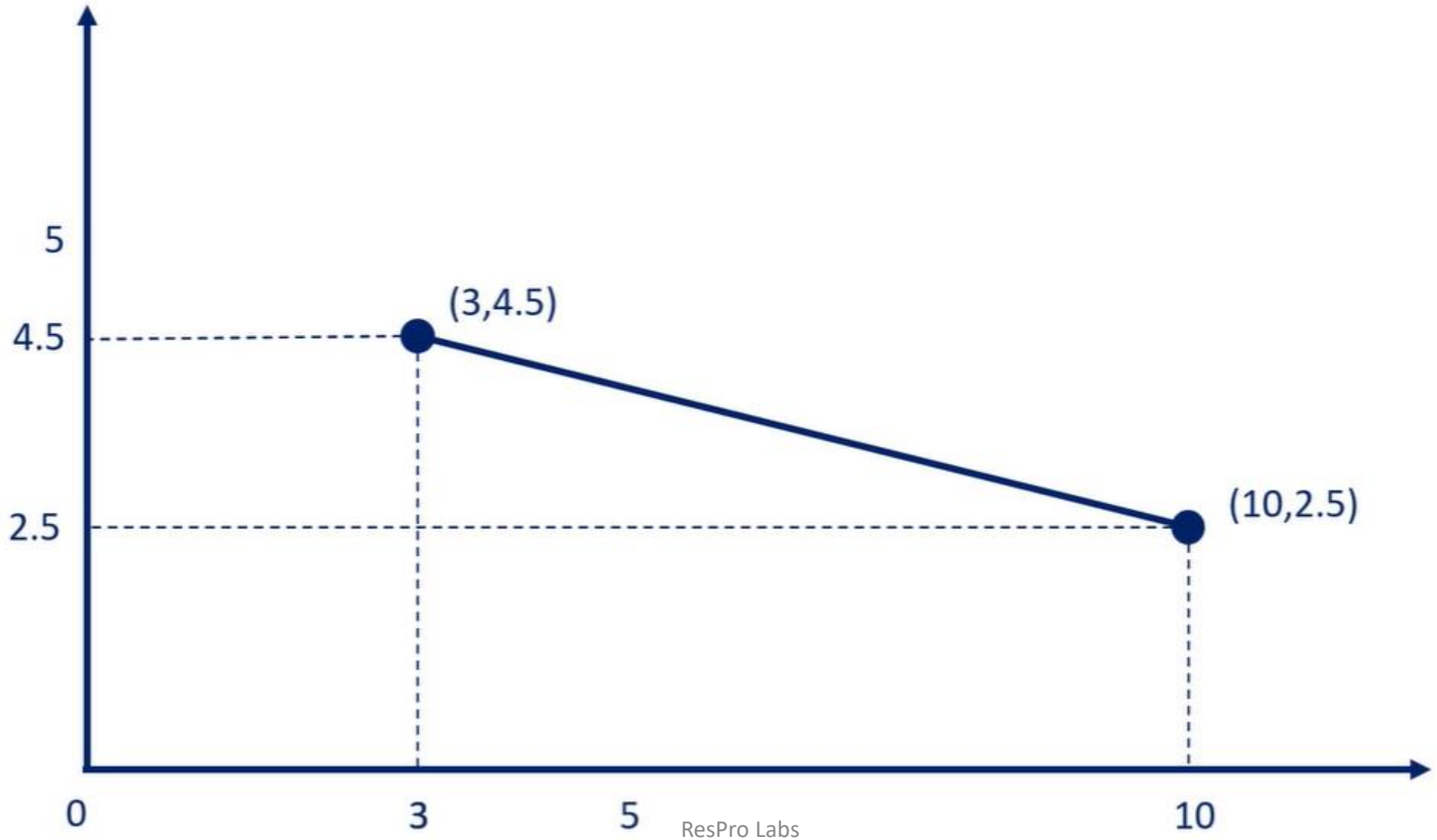

1st cluster
2nd cluster
3rd cluster
CANADA
GERMANY
UNITED KINGDOM
AUSTRALIA
U.S.A.
FRANCE

EXAMPLE

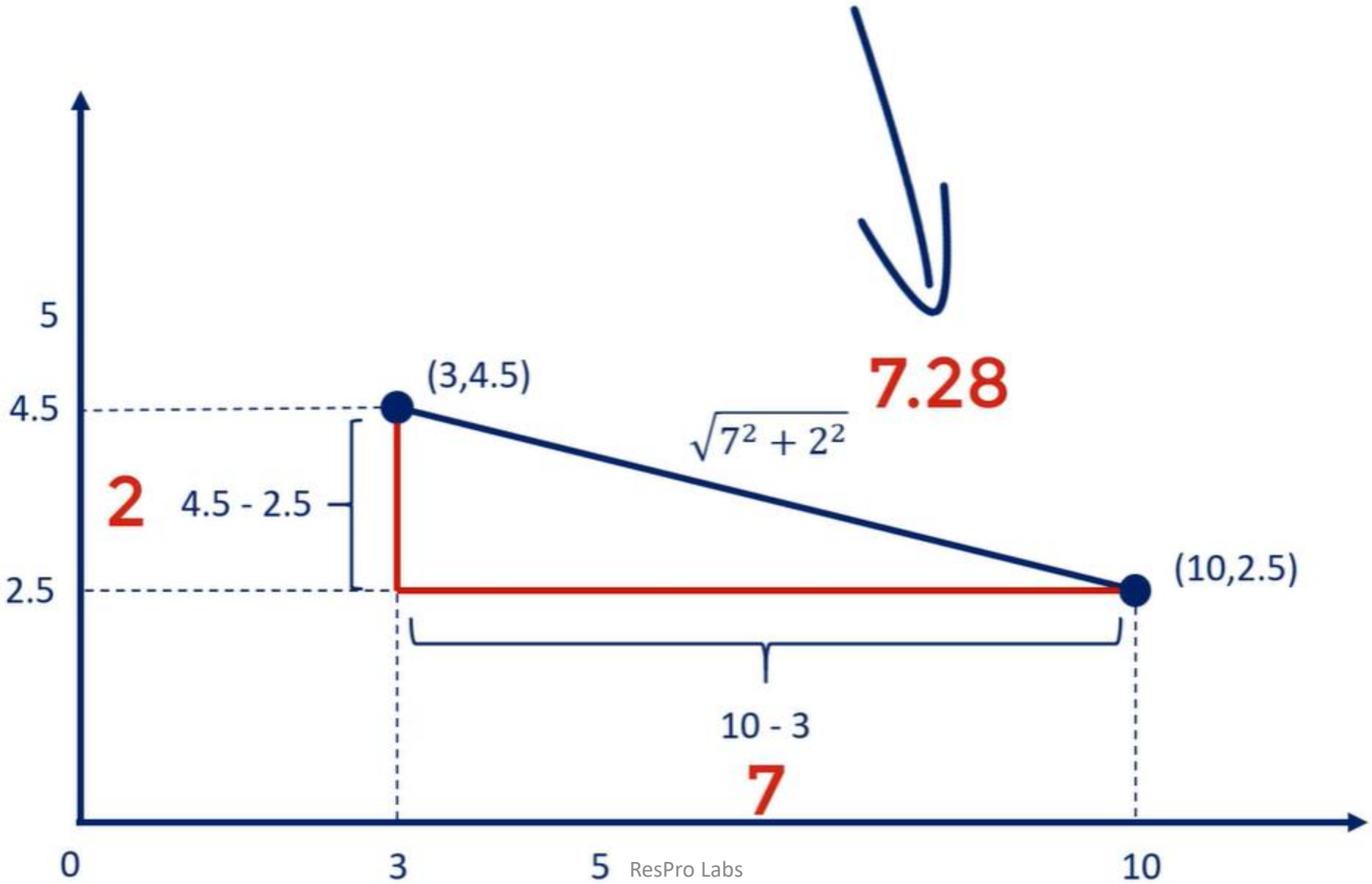

1st cluster
2nd cluster
CANADA
GERMANY
UNITED KINGDOM
AUSTRALIA
U.S.A.
FRANCE


1st cluster
CANADA
UNITED KINGDOM
U.S.A.
AUSTRALIA
Official language: English
2nd cluster
FRANCE
GERMANY
Official language: Not English

Euclidean distance



Euclidean distance



Euclidean distance

The most intuitive way to measure the distance between them is by drawing a straight line from one to the other. That's also known as Euclidean distance.

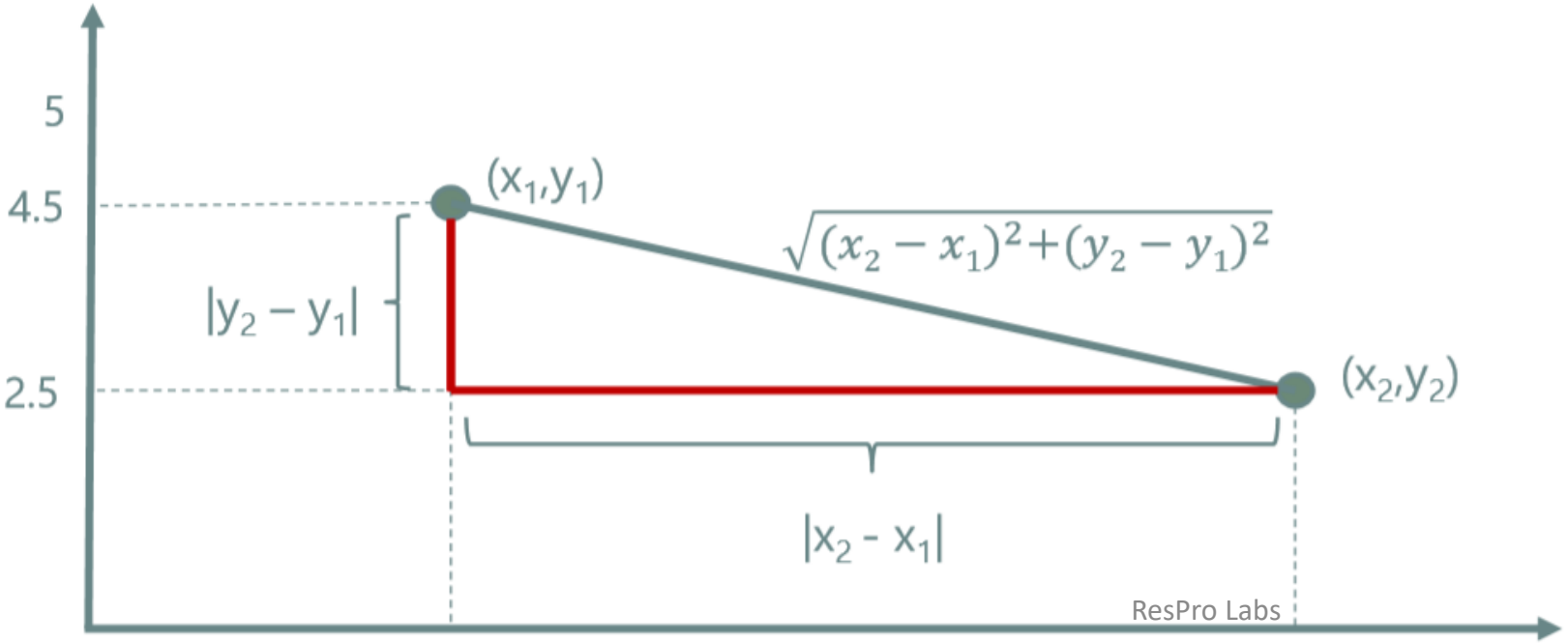
2D space: $d(A,B) = d(B,A) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$

3D space: $d(A,B) = d(B,A) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2}$

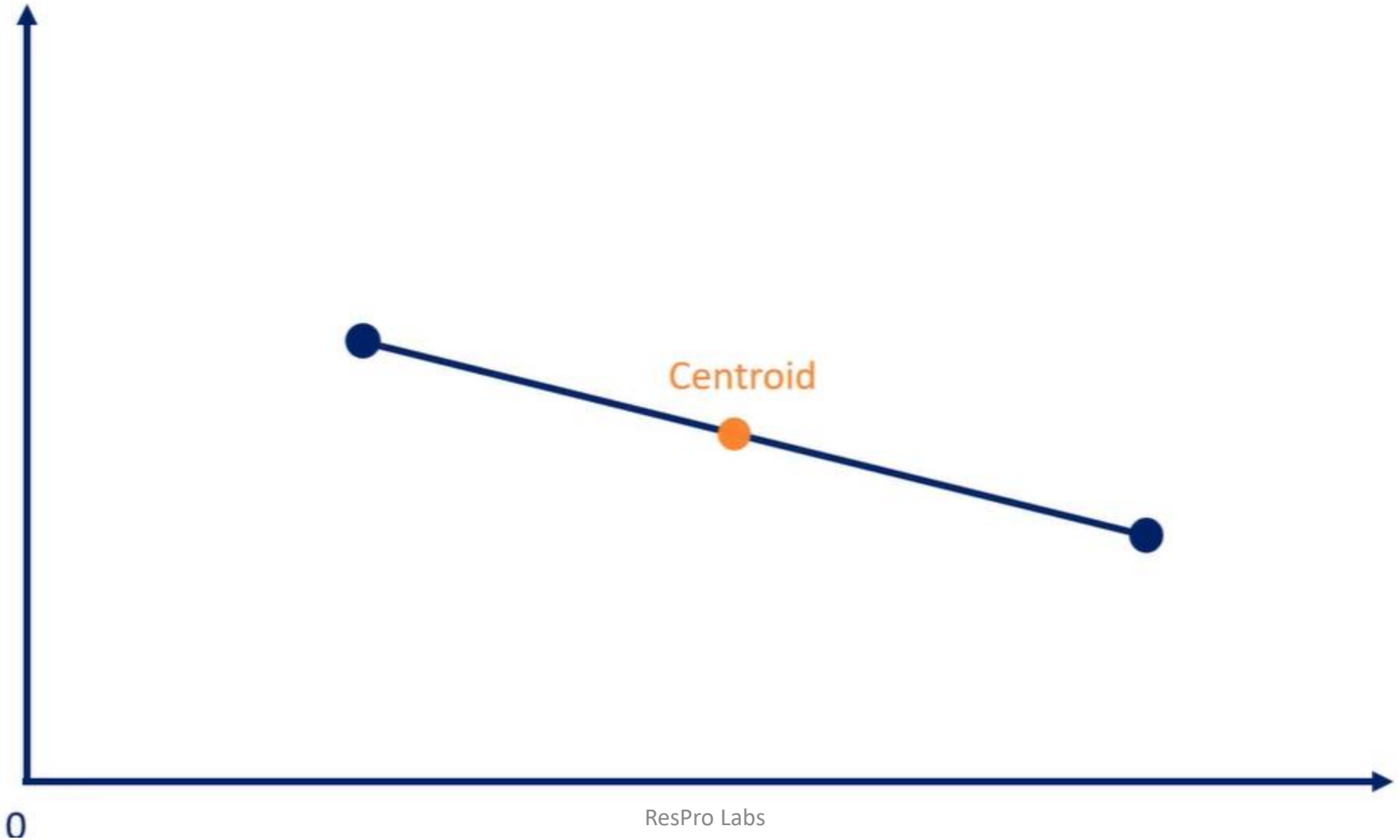
If the coordinates of A are $(a_1,a_2,\ldots,a_n)$ and of B are $(b_1,b_2,\ldots b_n)$

N-dim space:

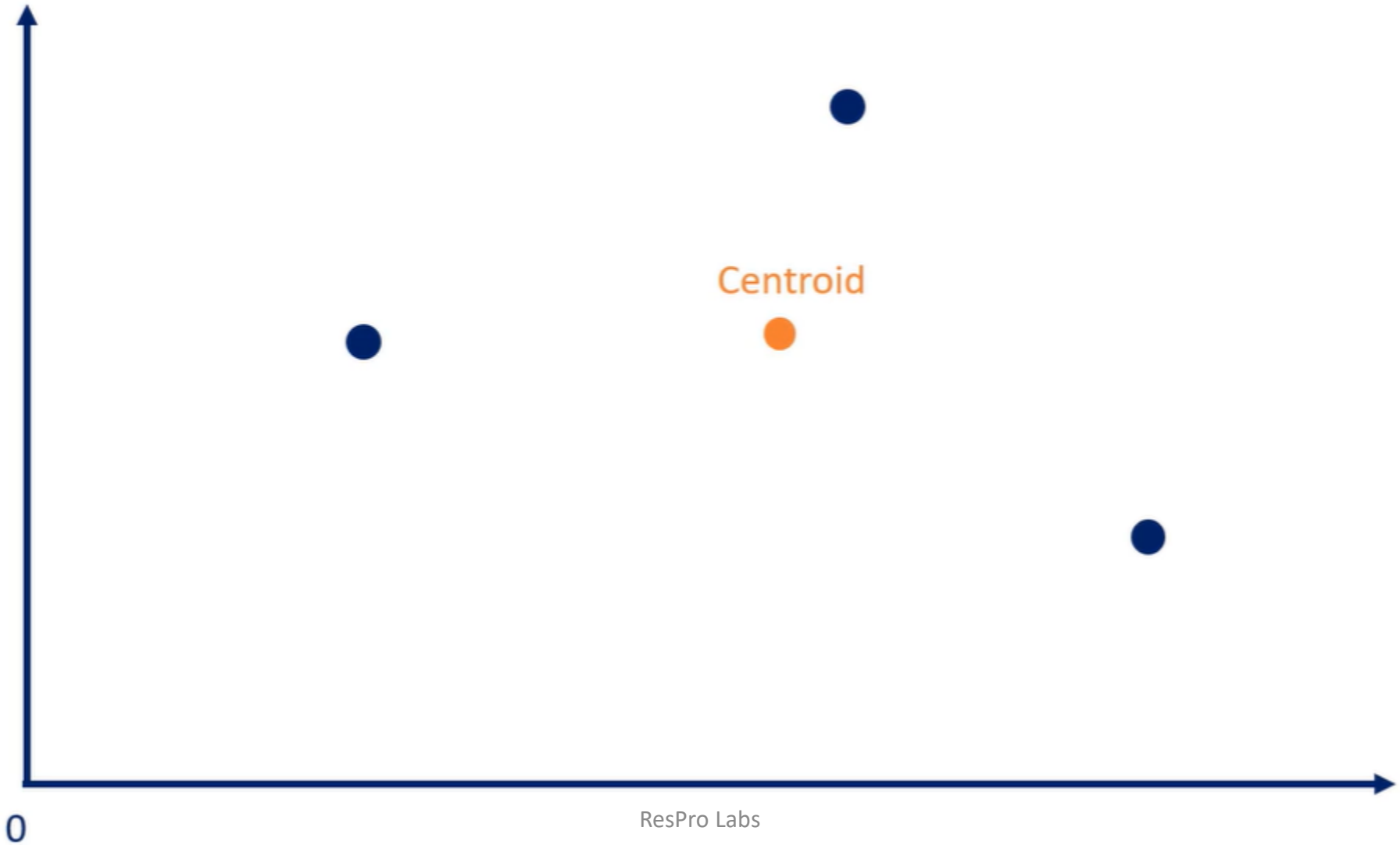
$$d(A,B) = d(B,A) = \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2 + \cdots + (a_n - bn)^2}$$



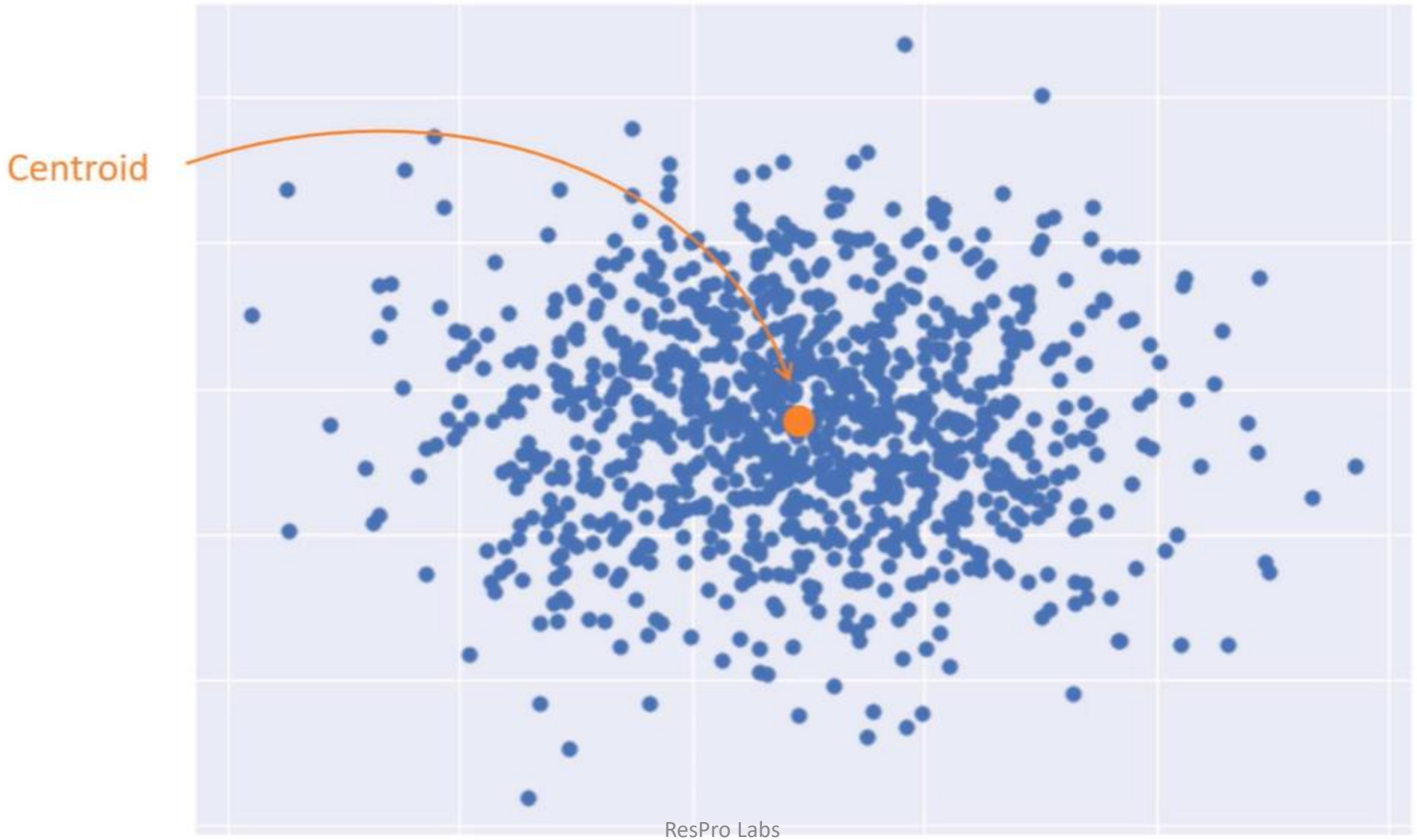
What's a centroid?



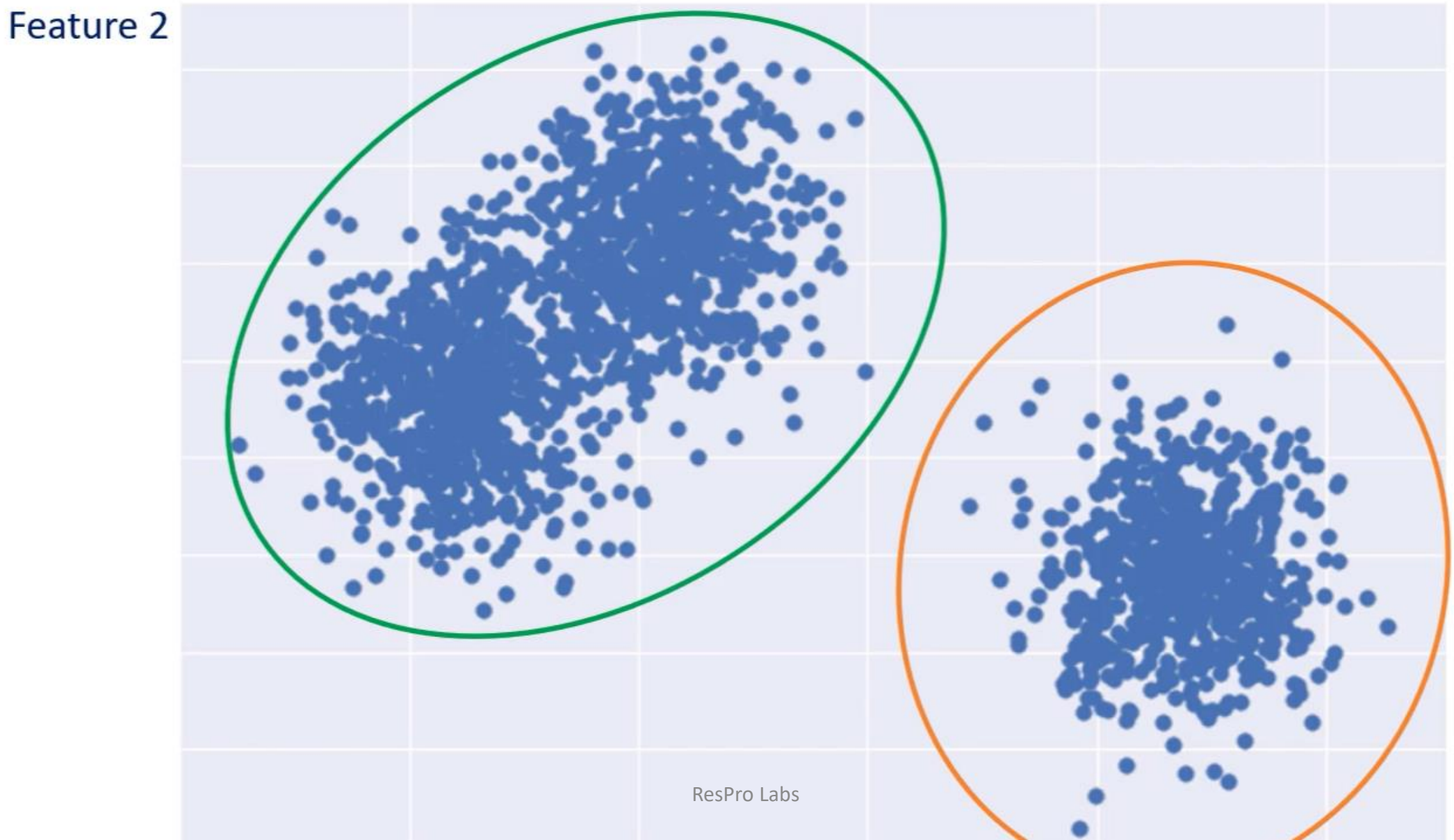
What's a centroid?



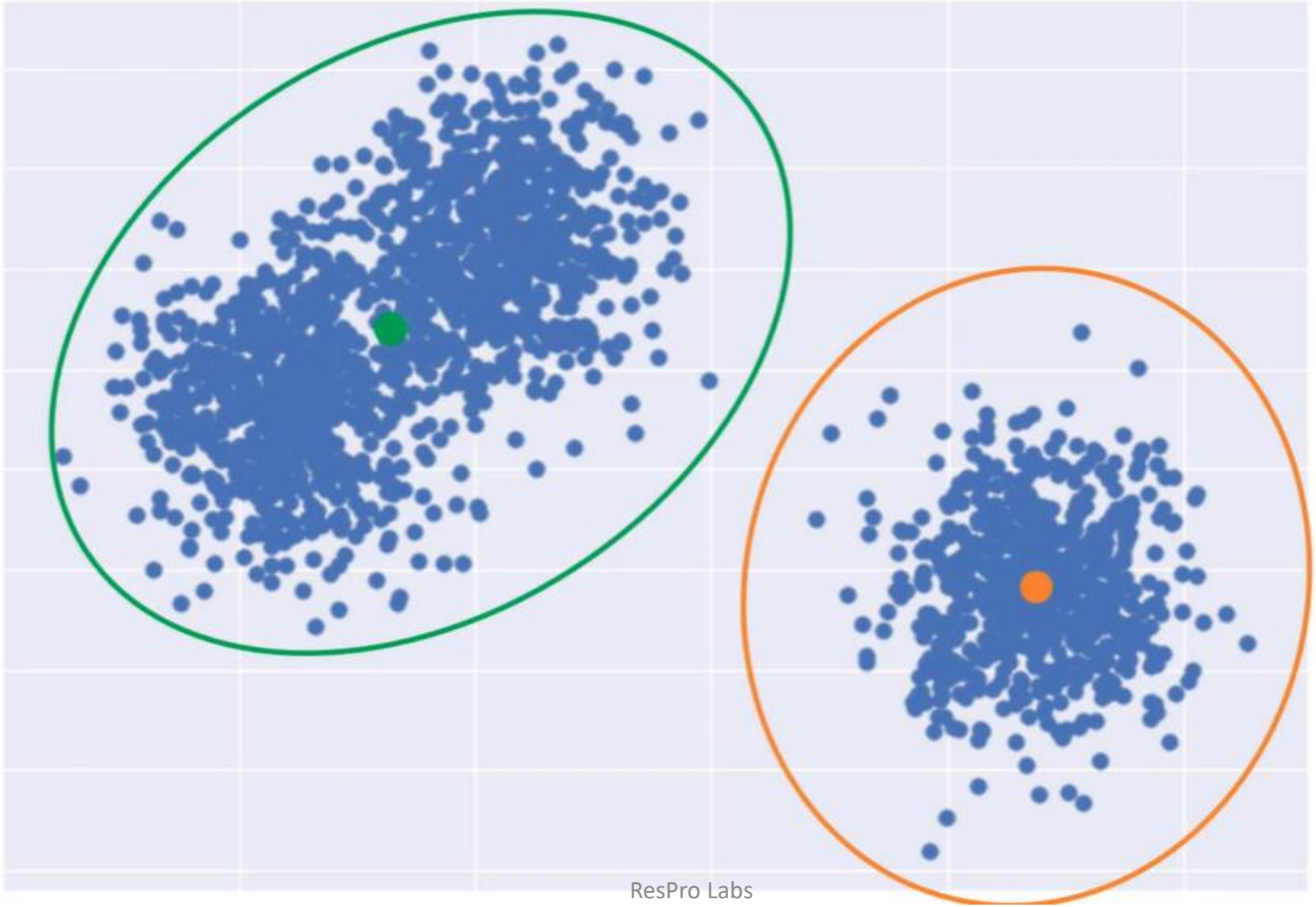
What's a centroid?



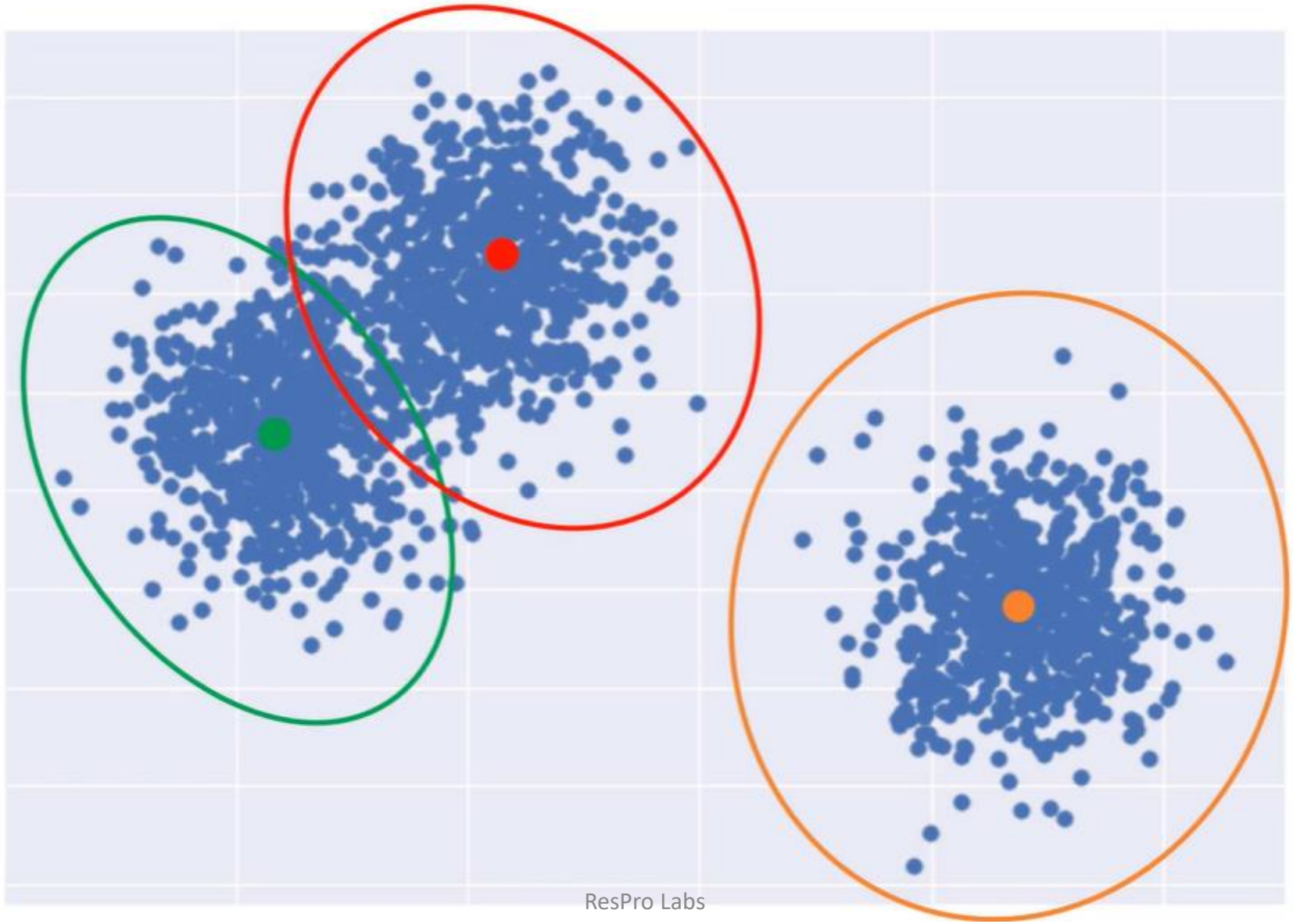
Clusters



Clusters



Clusters



K-means clustering

1. Choose number of clusters

Number – K, chosen by the person performing the clustering

2. Specify the number of seeds

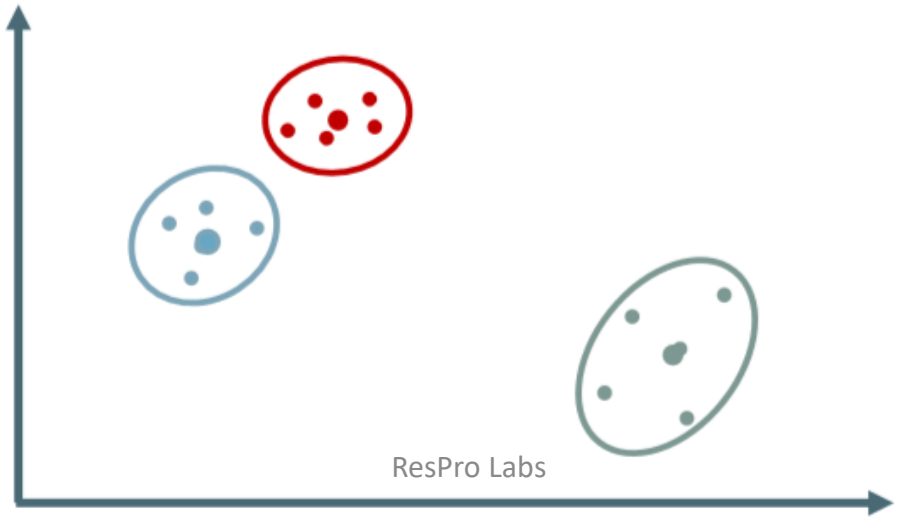
A seed is a starting centroid (can be chosen at random, with an algorithm or according to some prior knowledge)

3. Assign each point to a centroid

Based on proximity (measured by Euclidian distance)

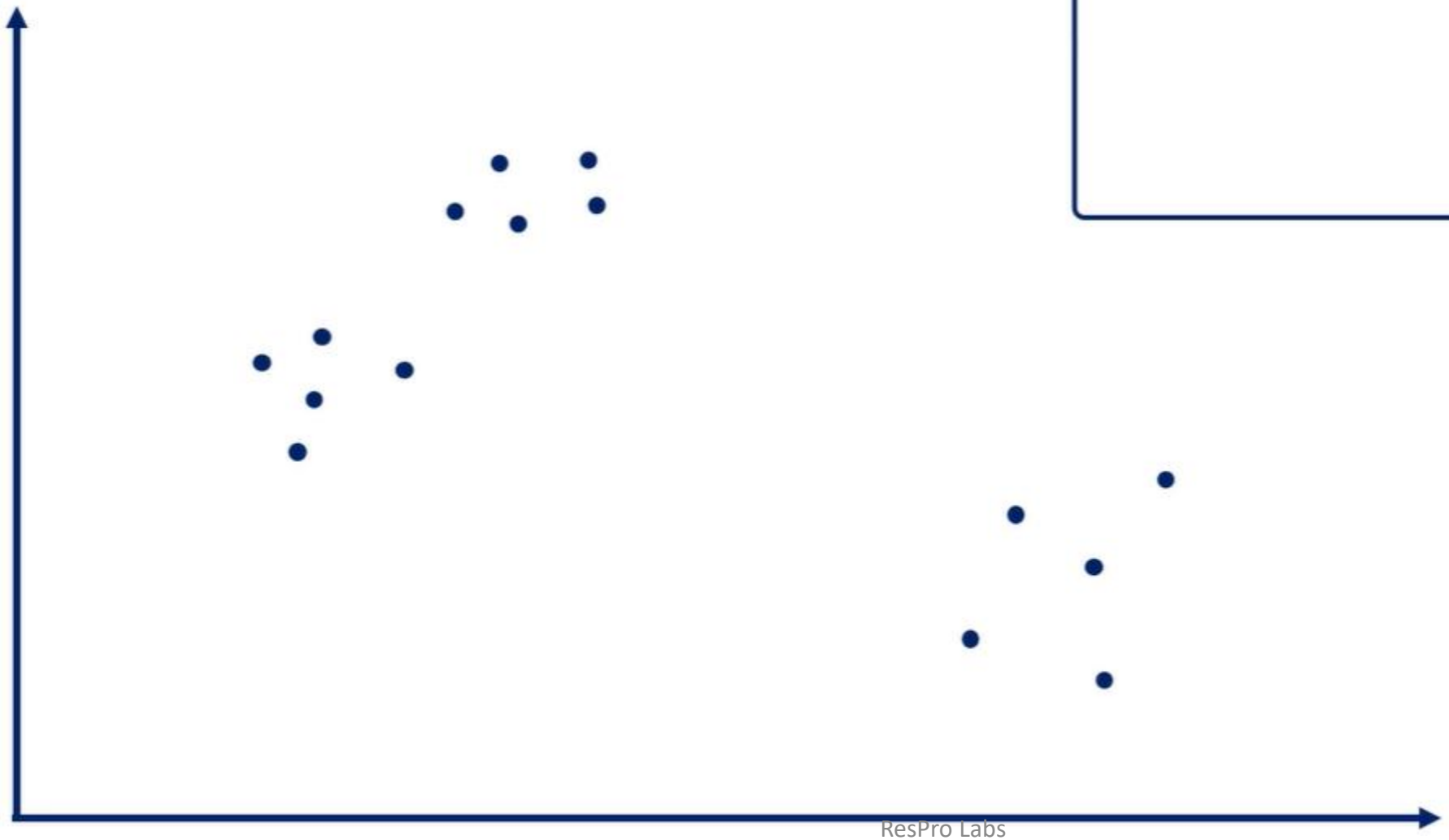
4. Adjust the centroids

Repeat 2. and 3. until there is you can no longer find a better clustering solution



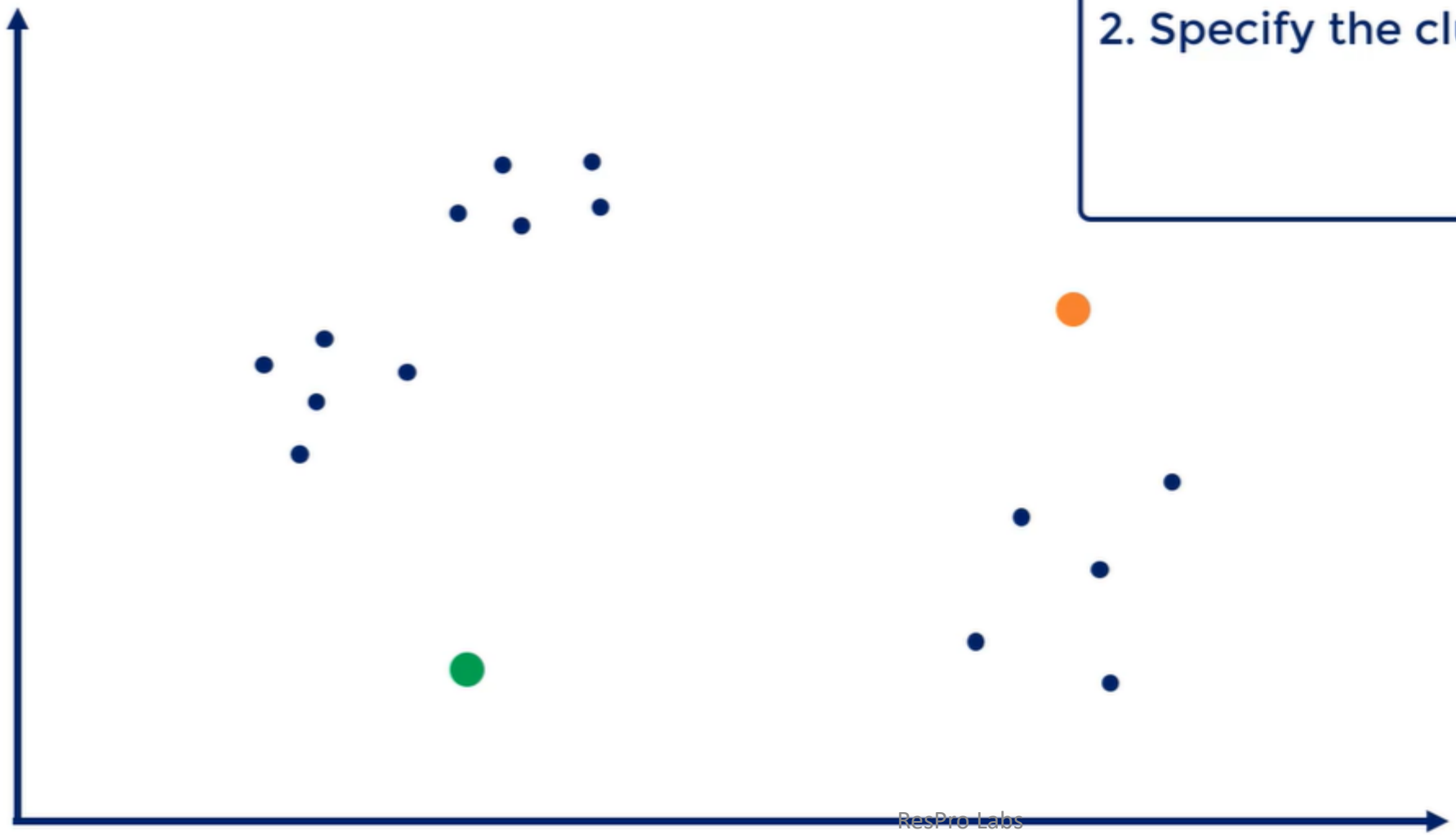
K-means clustering

1. Choose the number of clusters



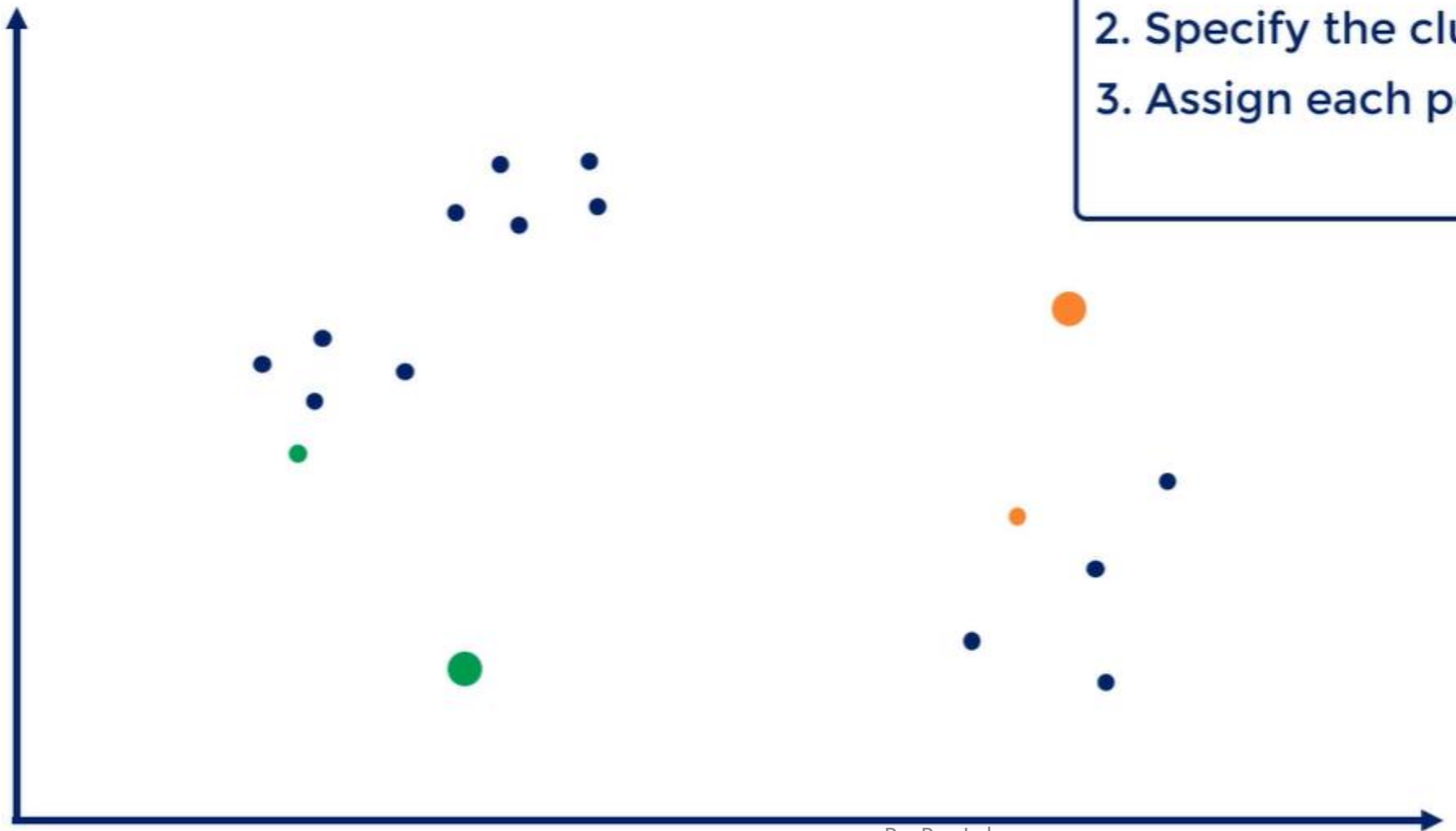
K-means clustering

1. Choose the number of clusters
2. Specify the cluster seeds



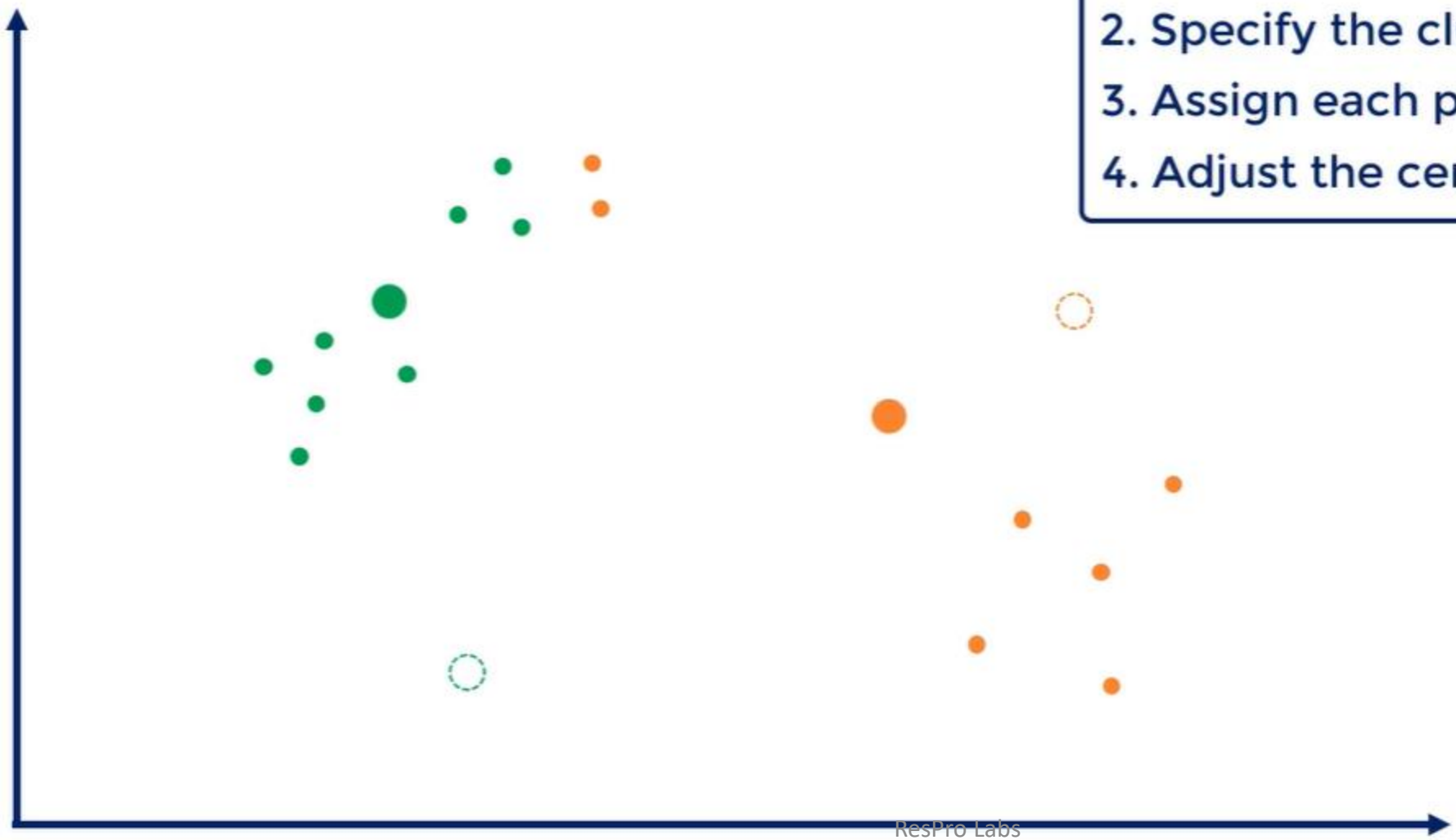
K-means clustering

1. Choose the number of clusters
2. Specify the cluster seeds
3. Assign each point to a centroid

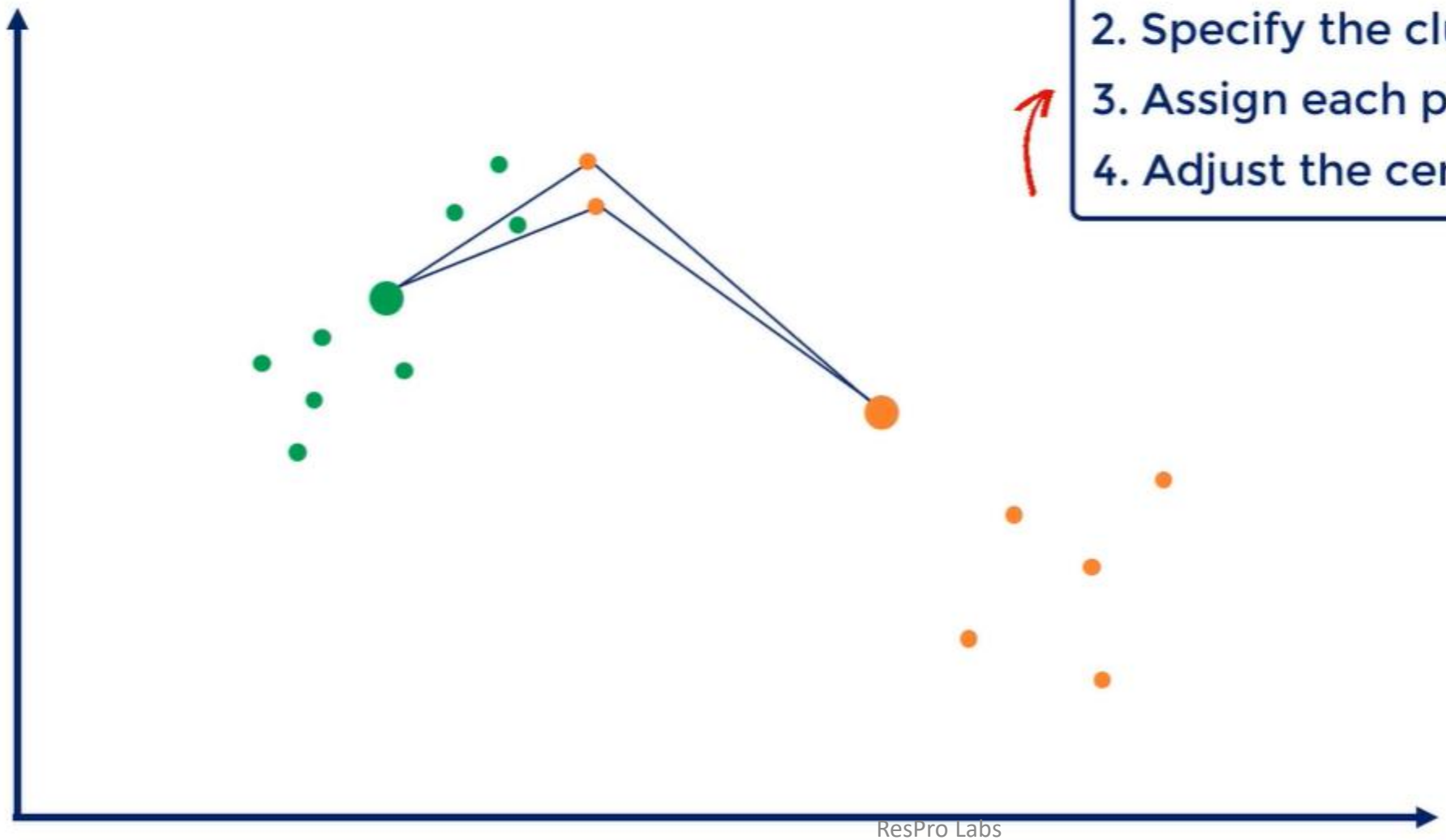


K-means clustering

1. Choose the number of clusters
2. Specify the cluster seeds
3. Assign each point to a centroid
4. Adjust the centroids



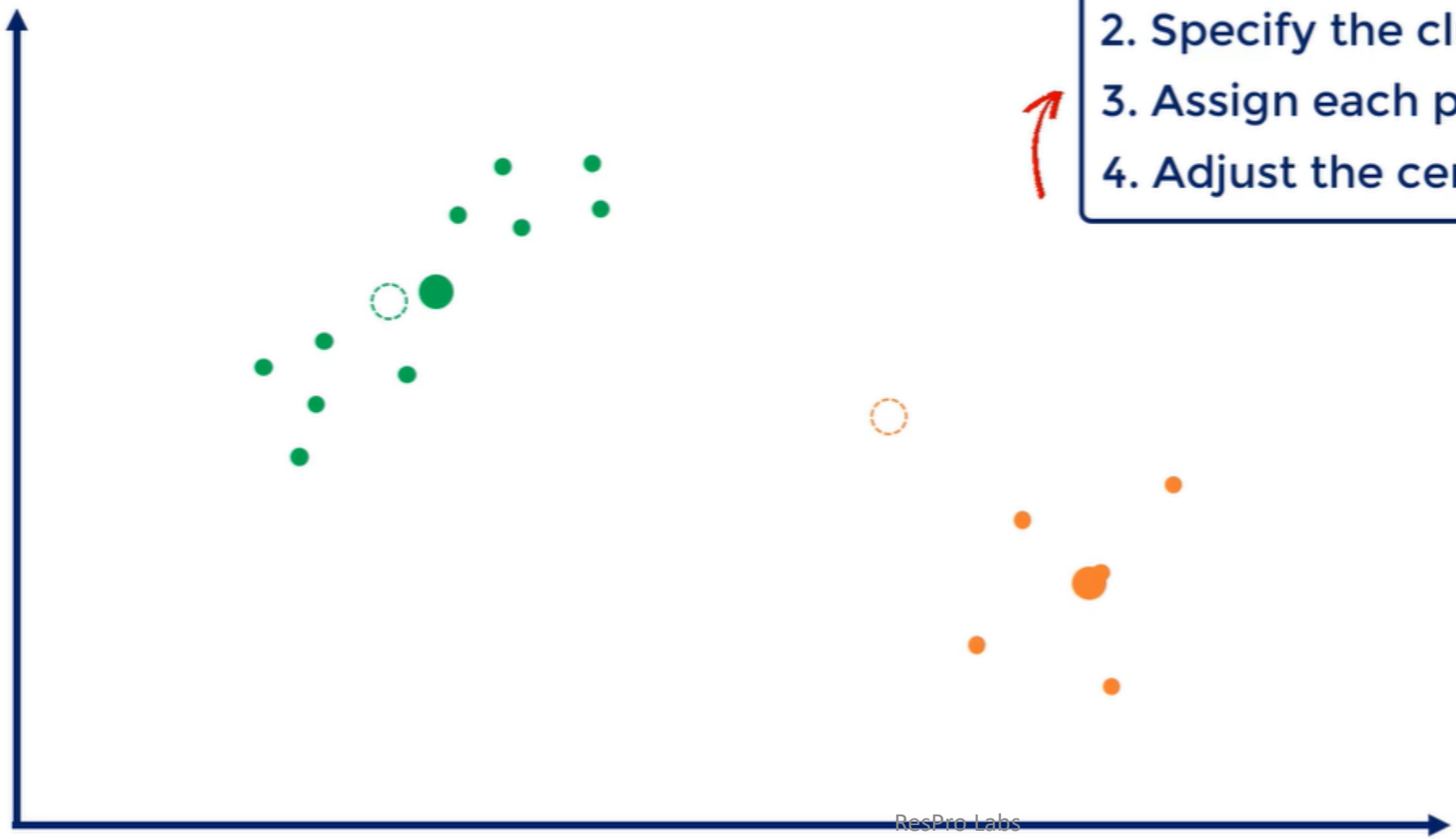
K-means clustering



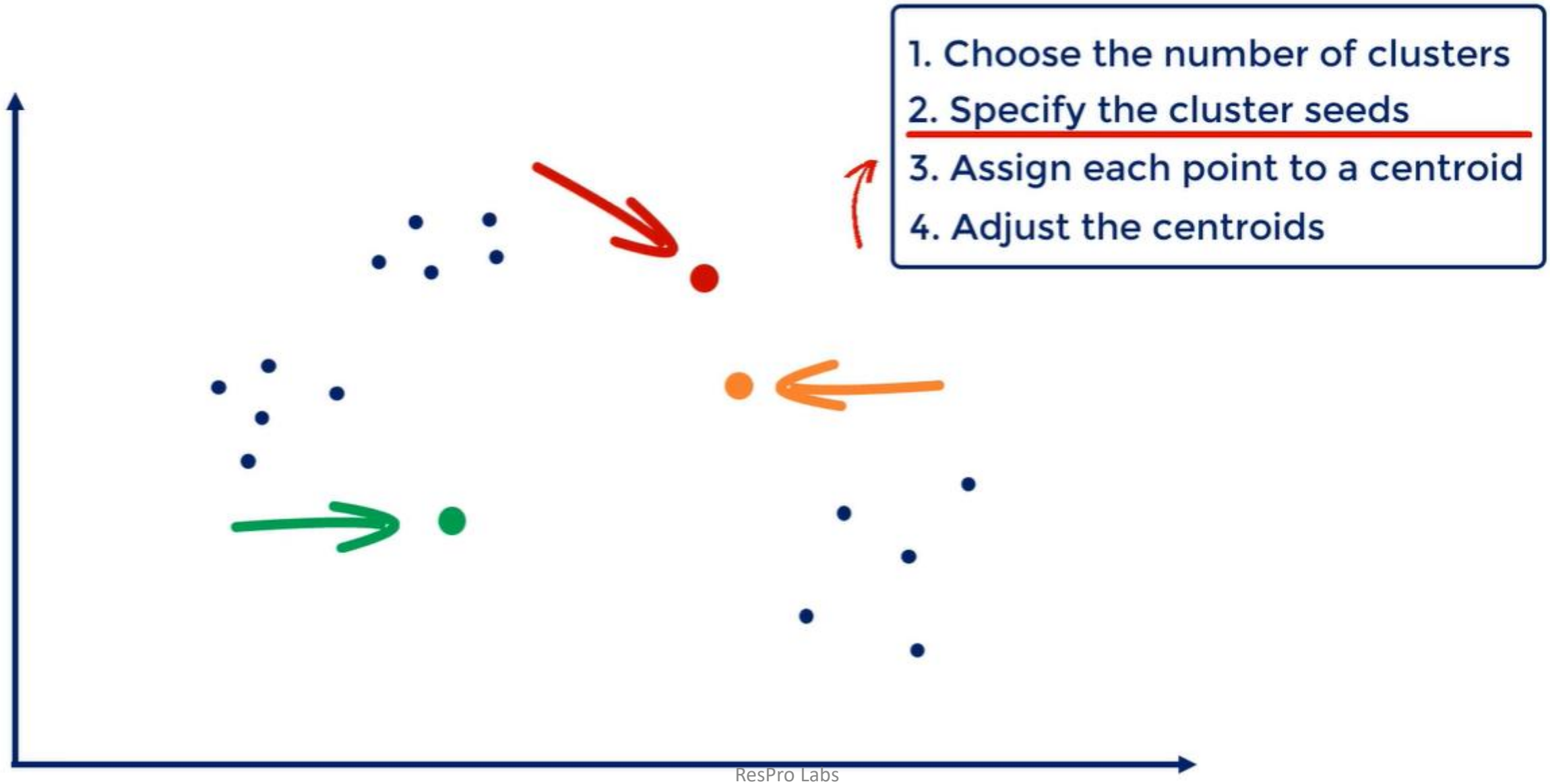
1. Choose the number of clusters
2. Specify the cluster seeds
3. Assign each point to a centroid
4. Adjust the centroids

K-means clustering

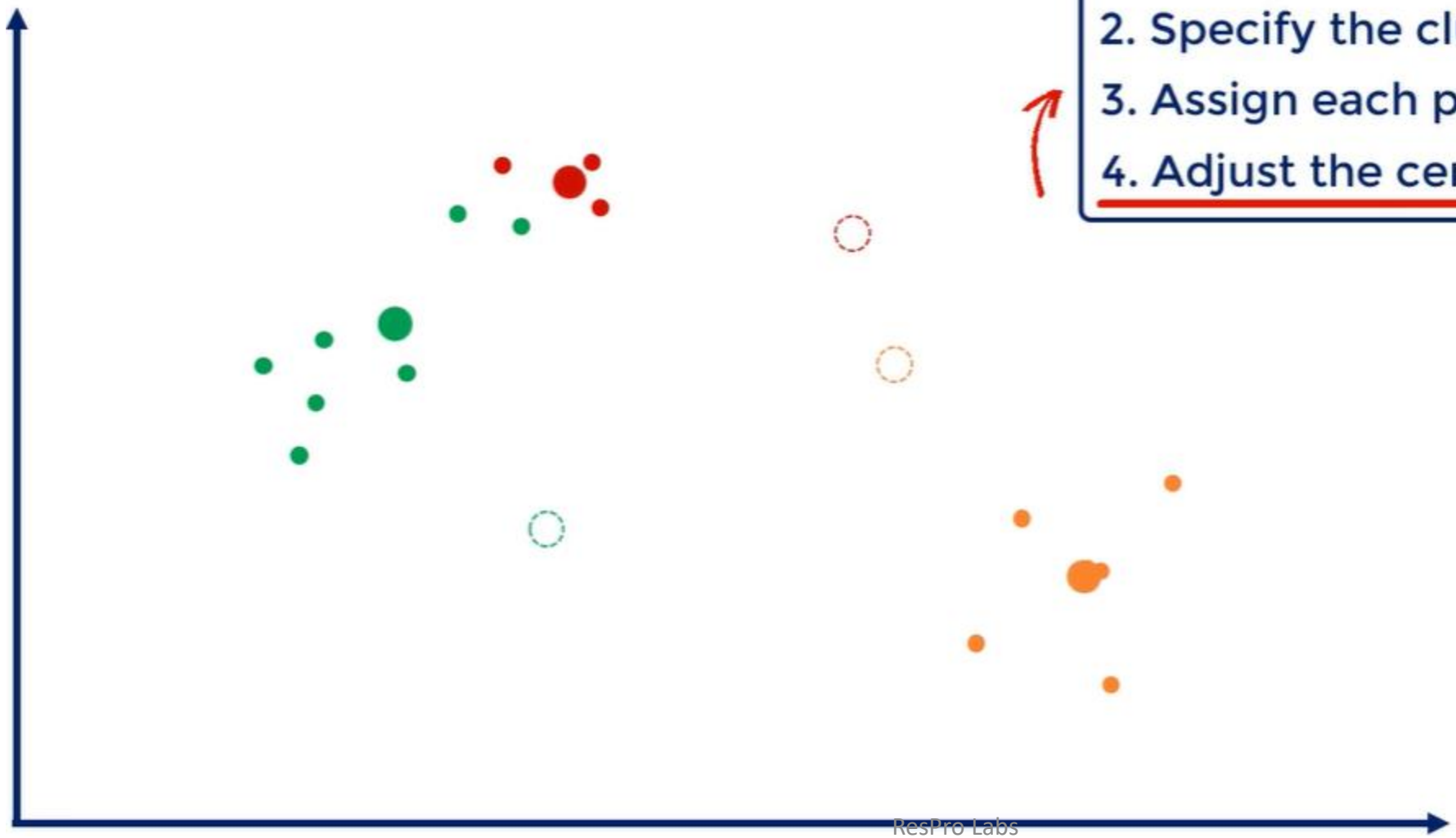
1. Choose the number of clusters
2. Specify the cluster seeds
3. Assign each point to a centroid
4. Adjust the centroids



K-means clustering

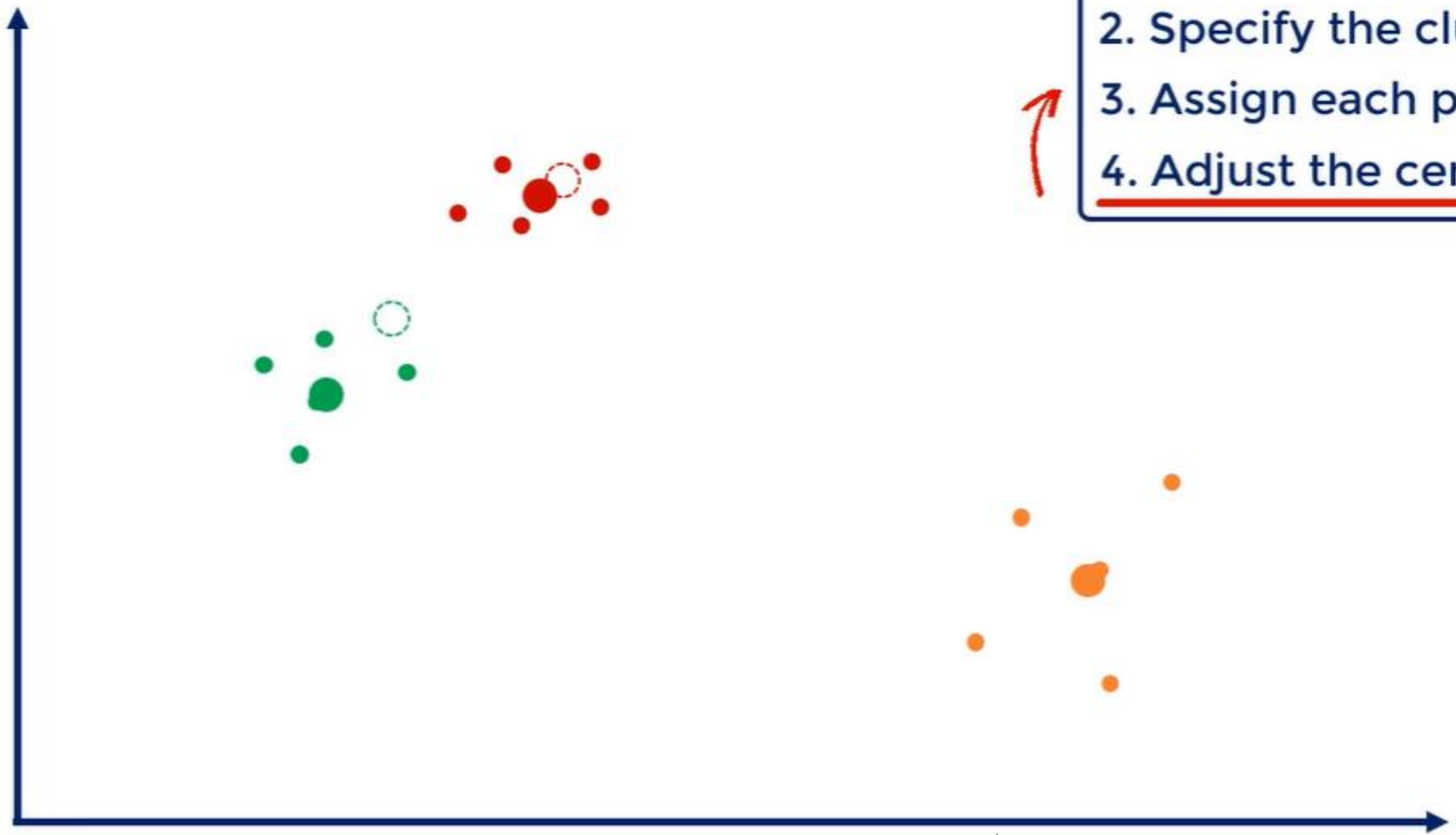


K-means clustering



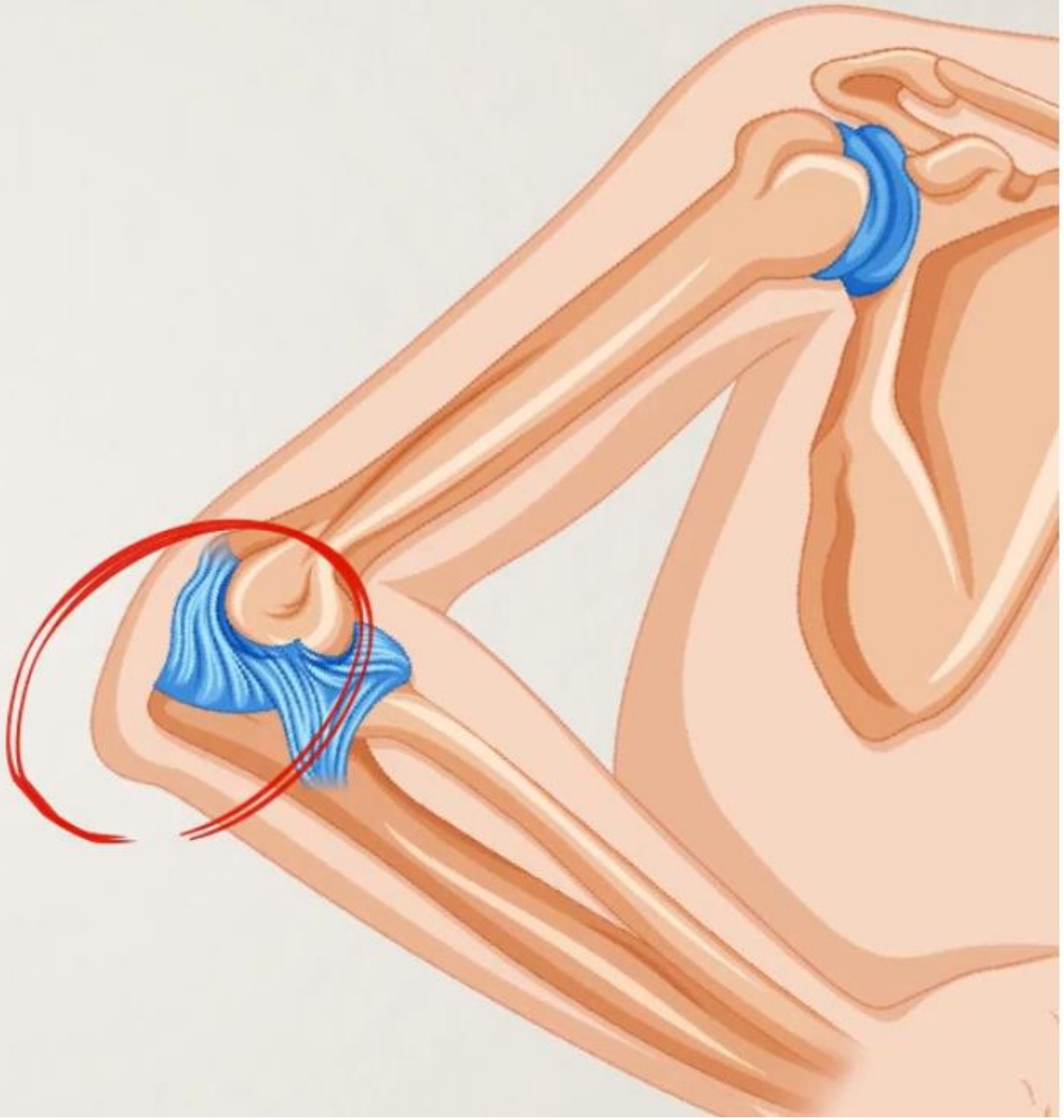
1. Choose the number of clusters
2. Specify the cluster seeds
3. Assign each point to a centroid
4. Adjust the centroids

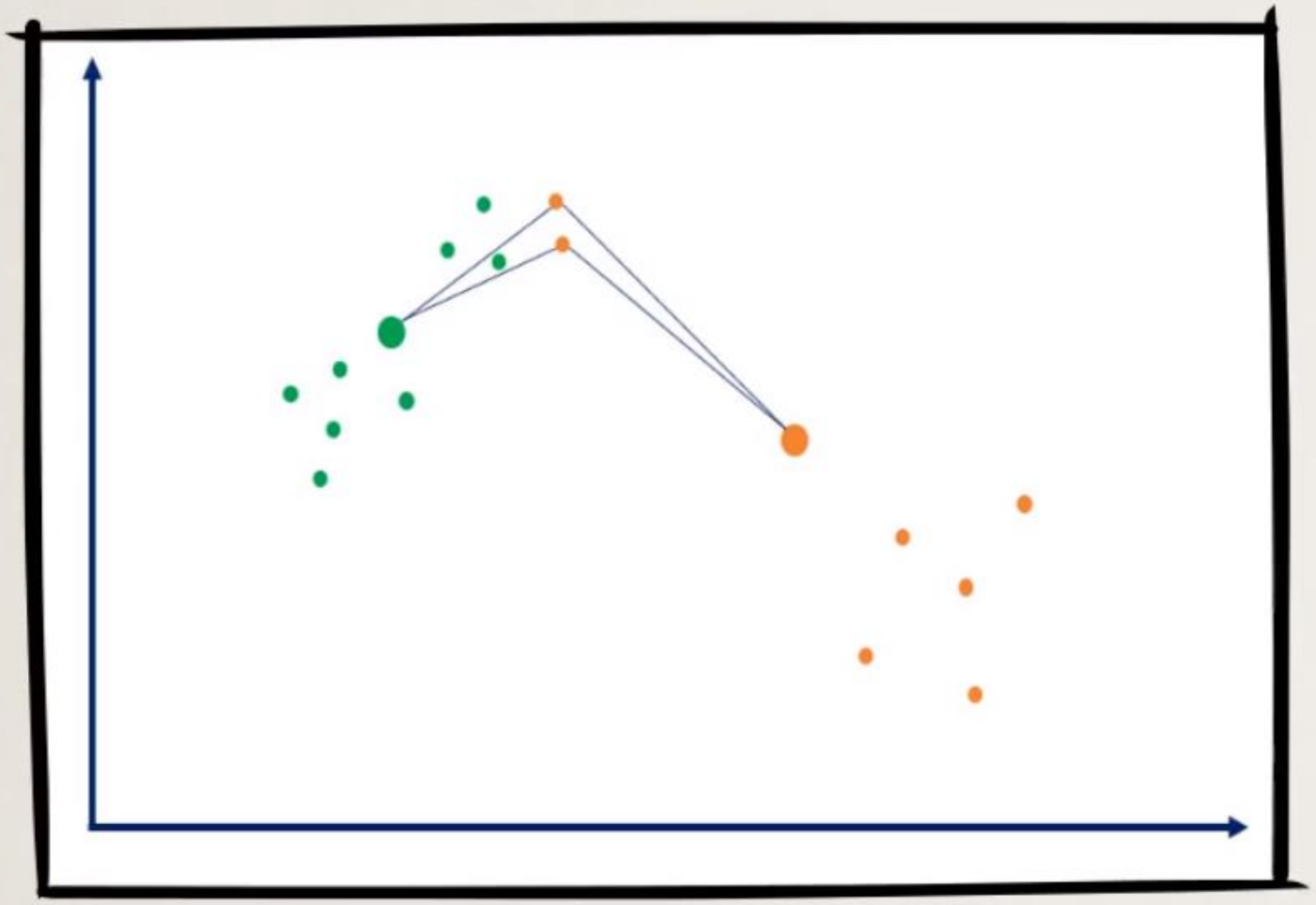
K-means clustering



1. Choose the number of clusters
2. Specify the cluster seeds
3. Assign each point to a centroid
4. Adjust the centroids

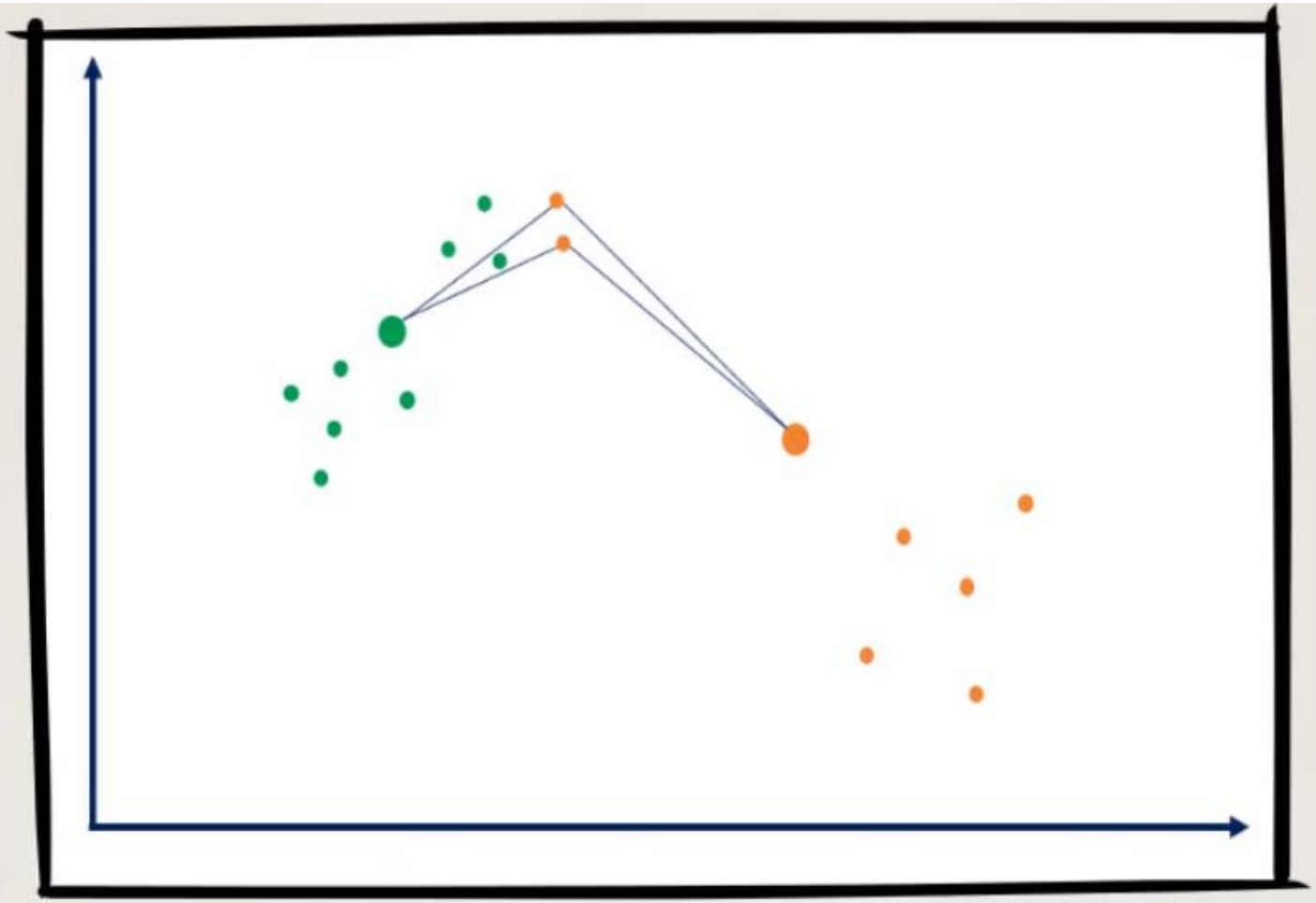
THE ELBOW METHOD





1) minimizing the distance between points in a cluster

2) maximizing the distance between clusters



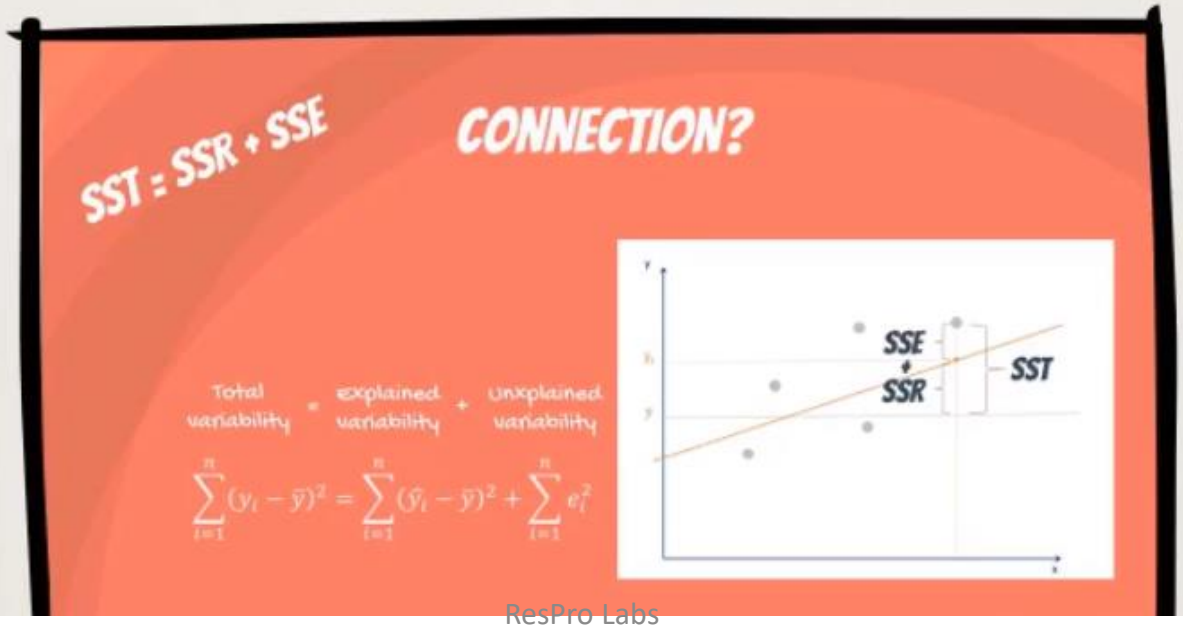
1) minimizing the distance between points in a cluster

=

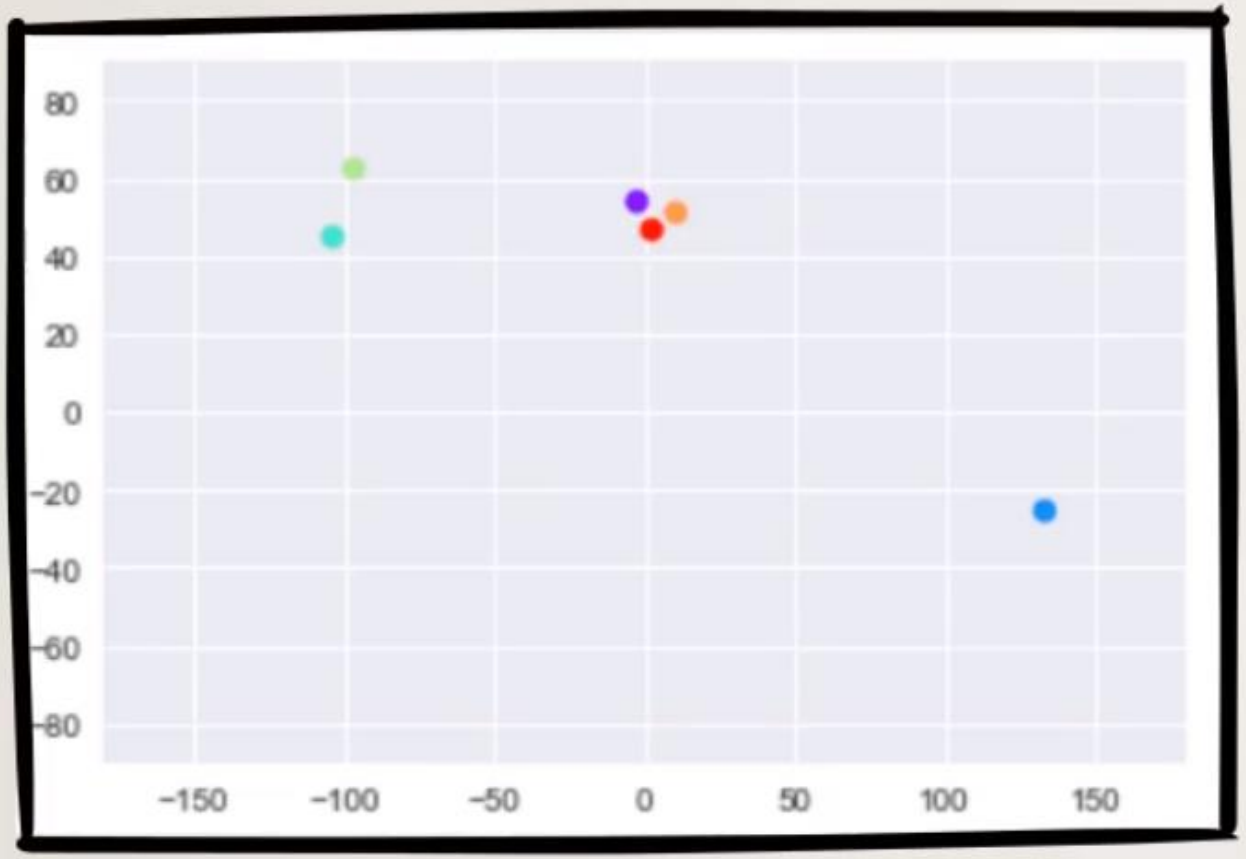
2) maximizing the distance between clusters

WCSS

similar to SST, SSR and SSE, WCSS is a measure developed within the ANOVA framework



WCSS

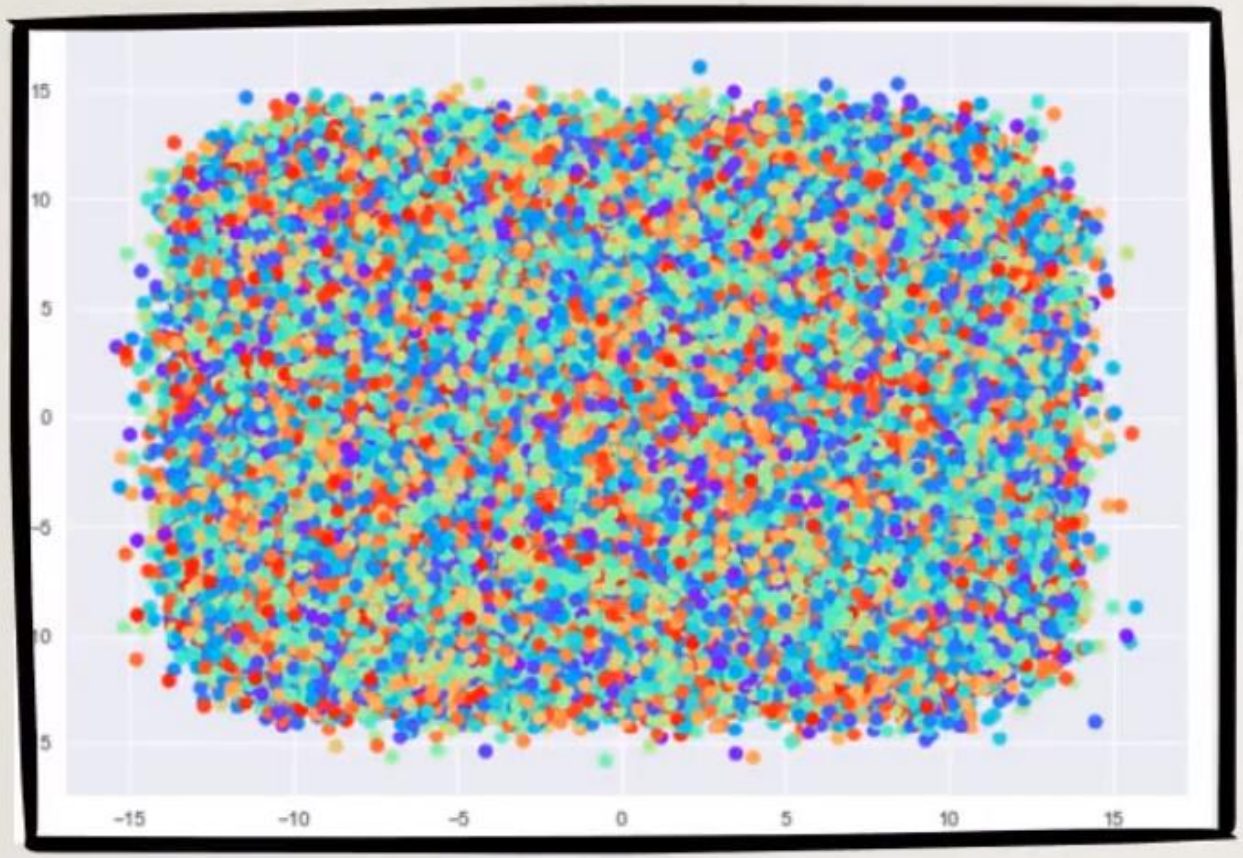


observations: 6

clusters: 6

WCSS = 0

WCSS

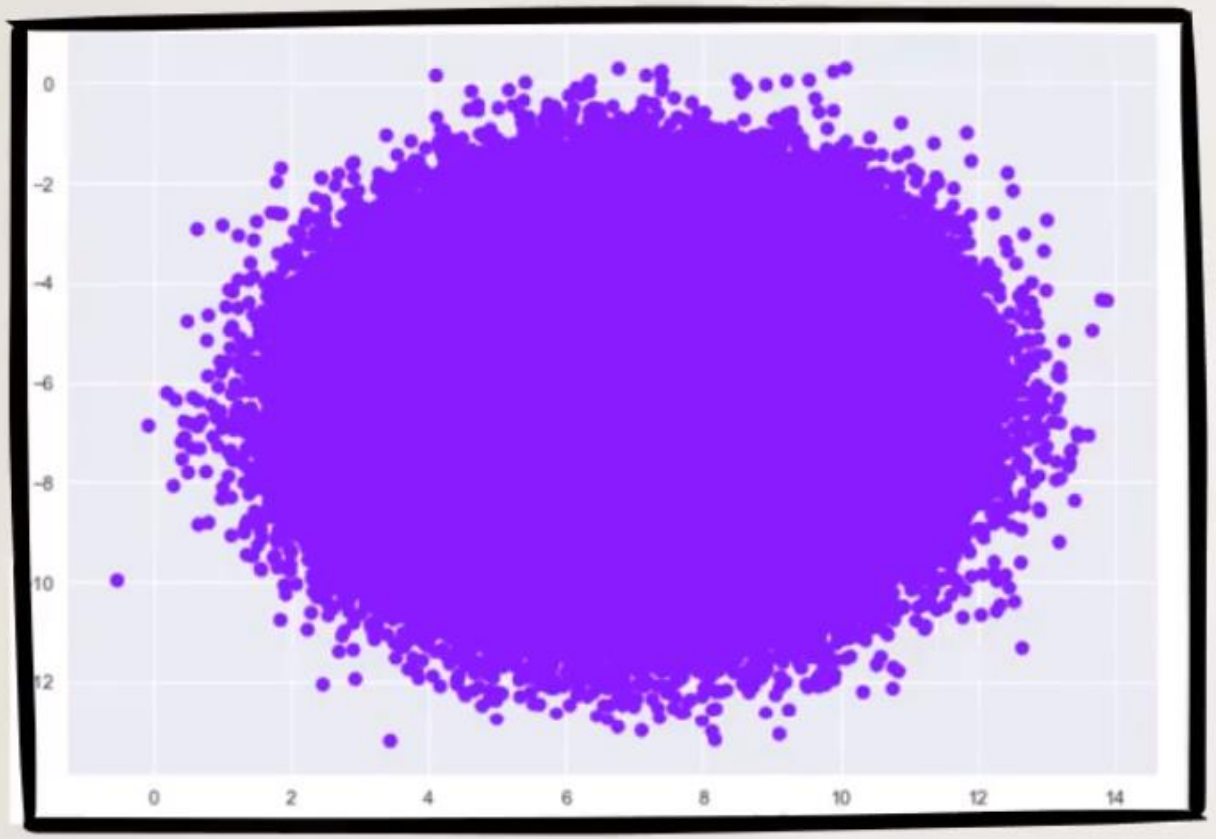


observations: **1,000,000**

clusters: **1,000,000**

WCSS = **0** = min

WCSS



observations: 1,000,000

clusters: 1

WCSS = max

WCSS

observations: 1,000,000

clusters: 1,000,000

WCSS = 0 = min

observations: 1,000,000

clusters: 1

WCSS = max

MIDDLE GROUND?

WCSS

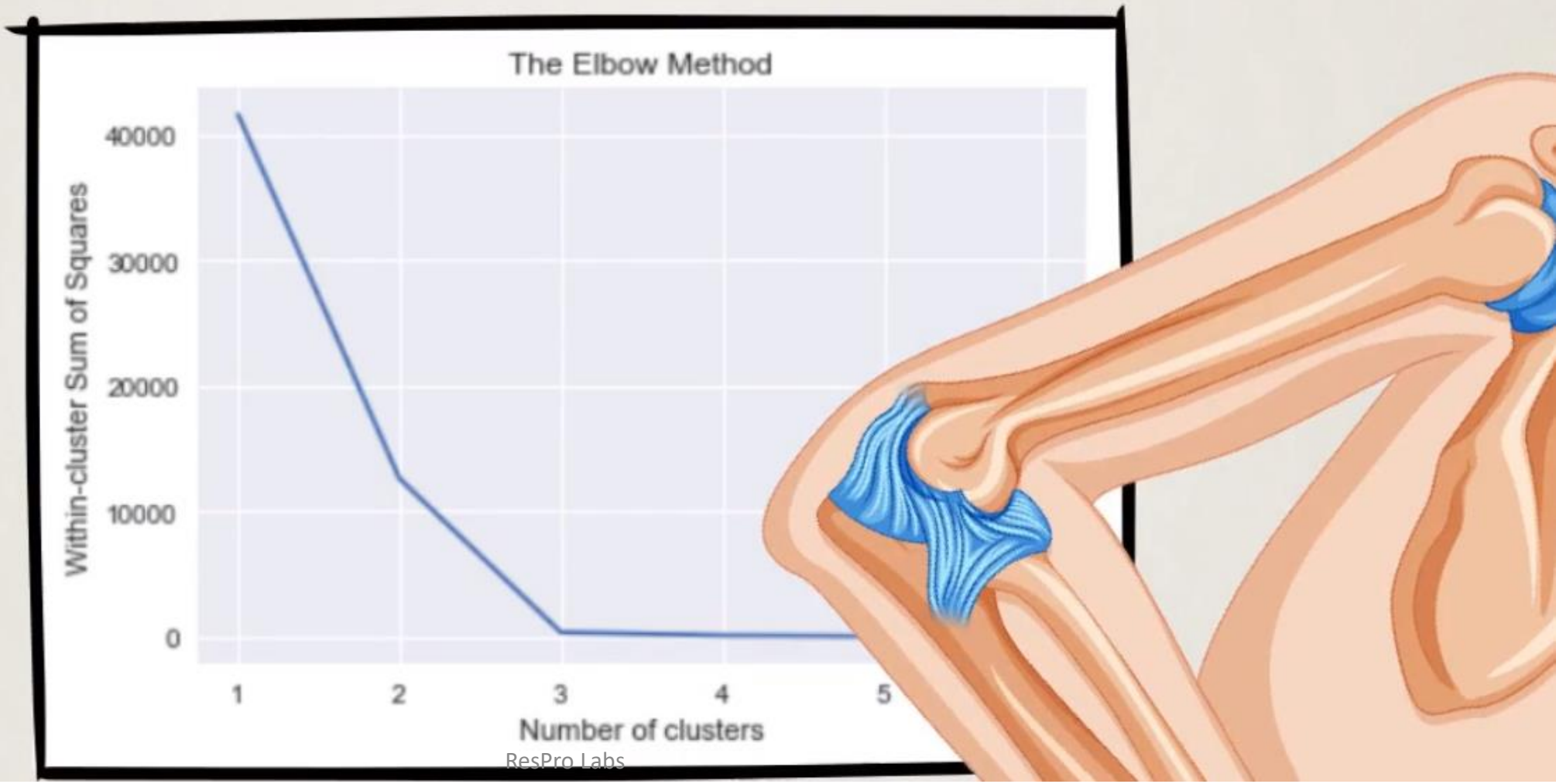
MIDDLE GROUND?

observations: **N**

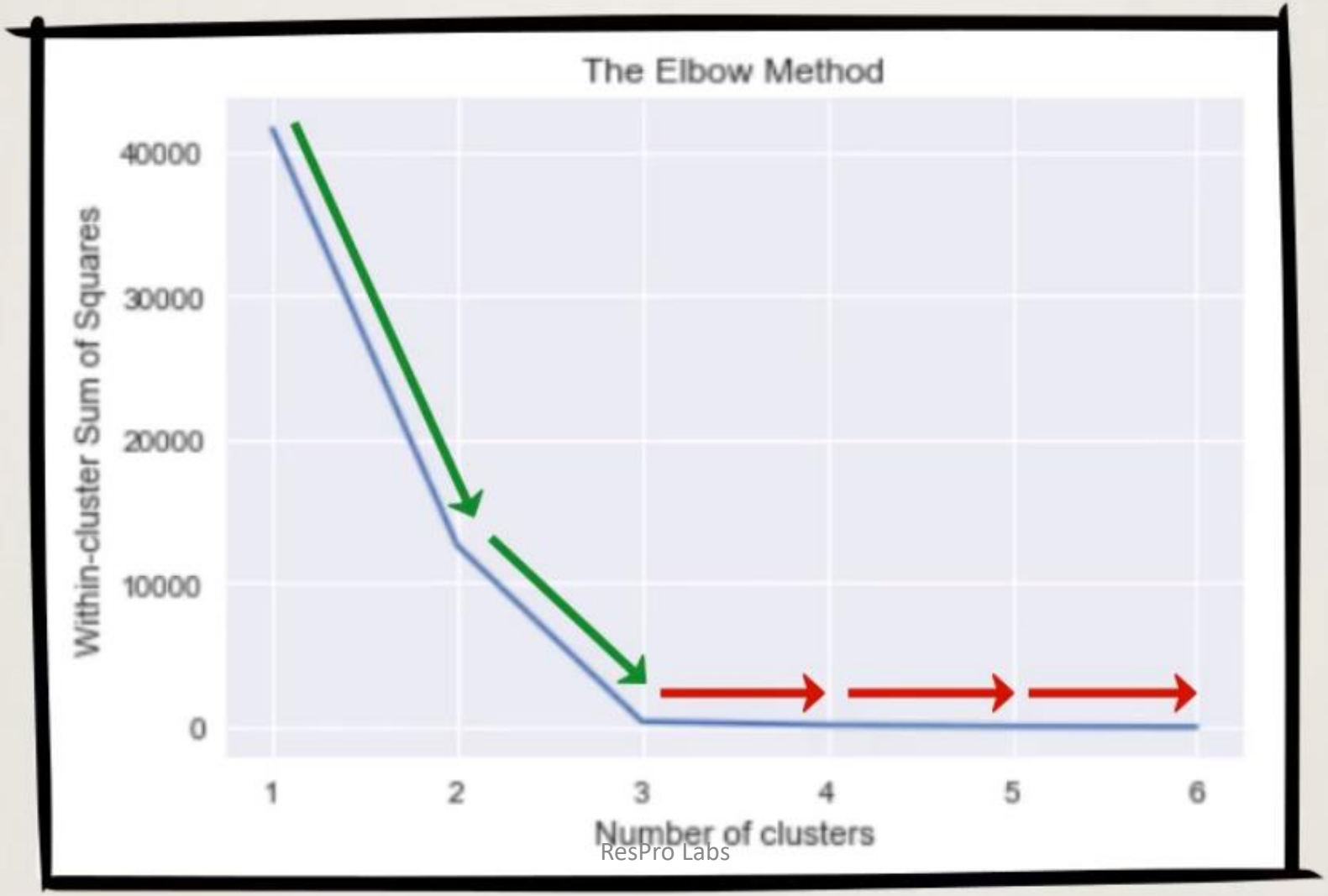
clusters: **SMALL**

WCSS = LOW

THE ELBOW METHOD



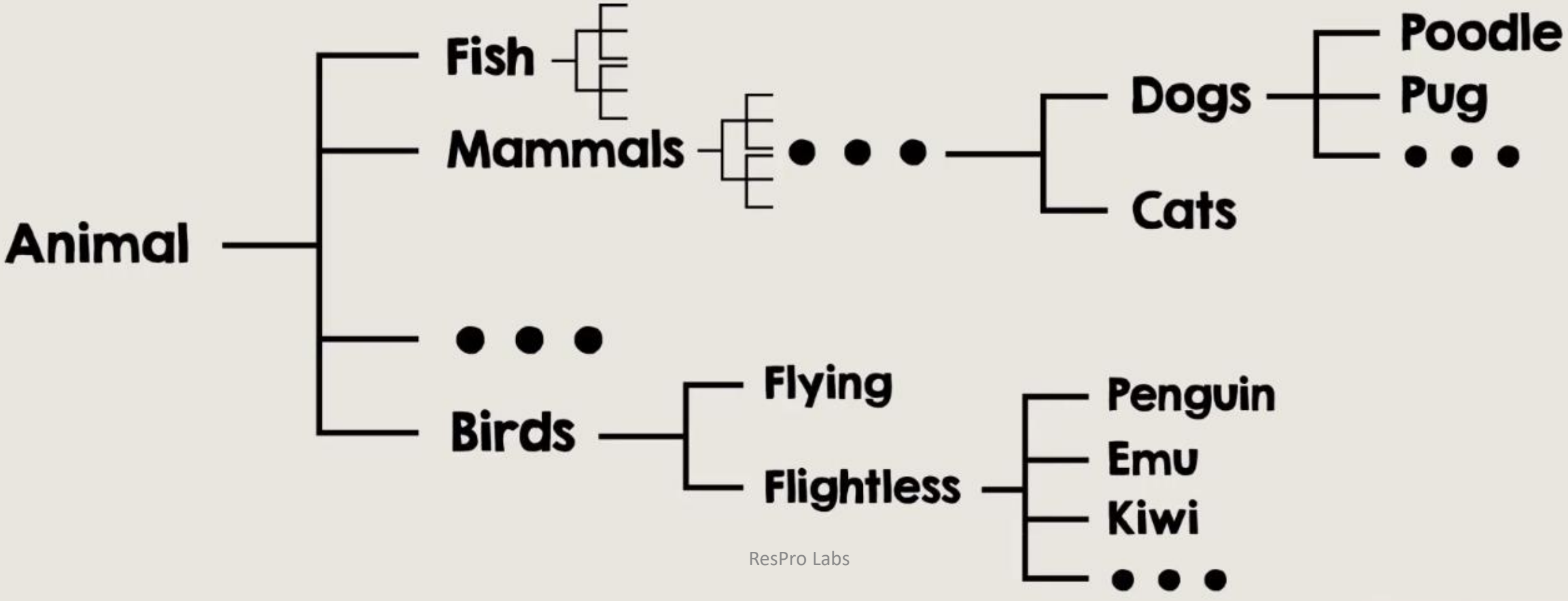
THE ELBOW METHOD



Clustering of Clustering

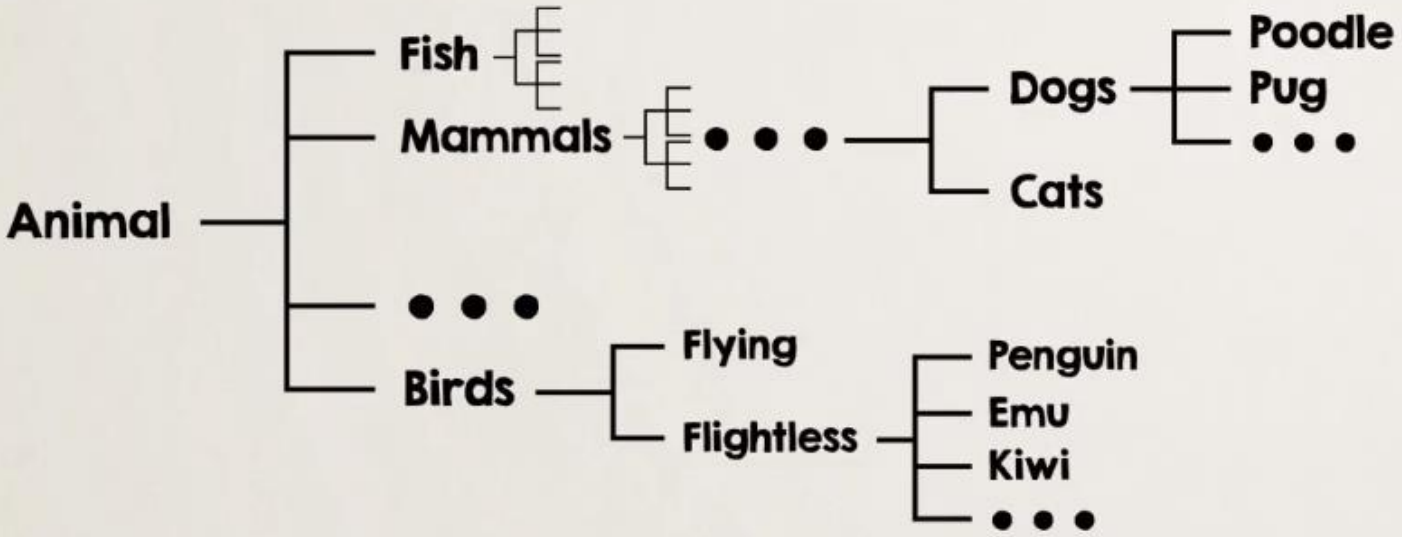
HIERARCHICAL

TAXONOMY OF THE ANIMAL KINGDOM

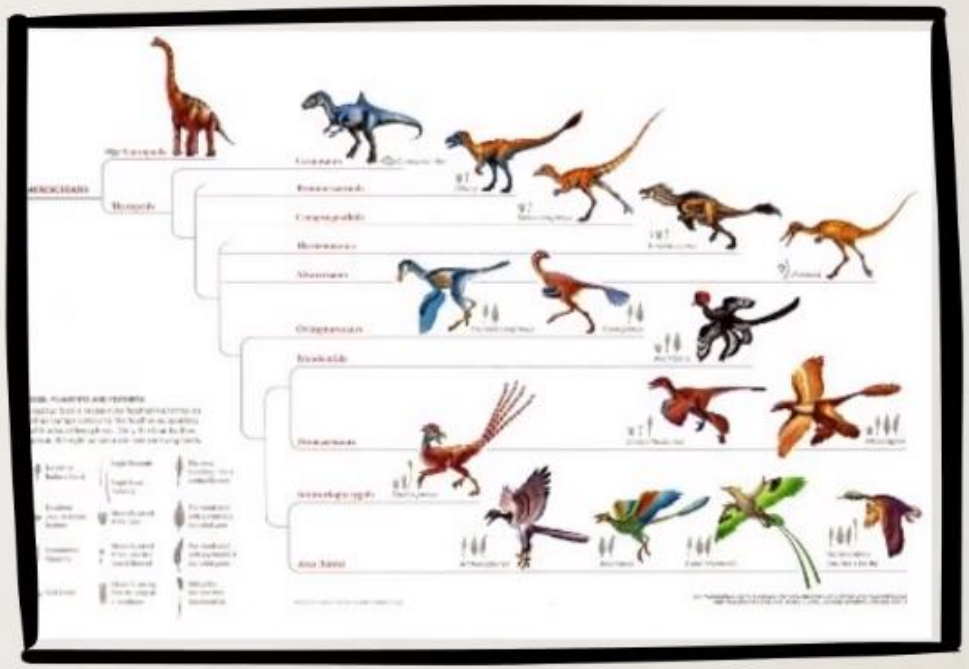


TYPES OF HIERARCHICAL CLUSTERING

AGGLOMERATIVE
(BOTTOM-UP)



DIVISIVE
(TOP-DOWN)



Types of clustering

With flat methods there is no hierarchy, but rather the number of clusters are chosen prior to clustering.

Flat methods have been developed because hierarchical clustering is much slower and computationally expensive.

Nowadays, flat methods are preferred because of the volume of data we typically try to cluster.

Historically, lustering was developed first. An example hierarchical cof clustering with hierarchy is taxonomy of the animal kingdom.

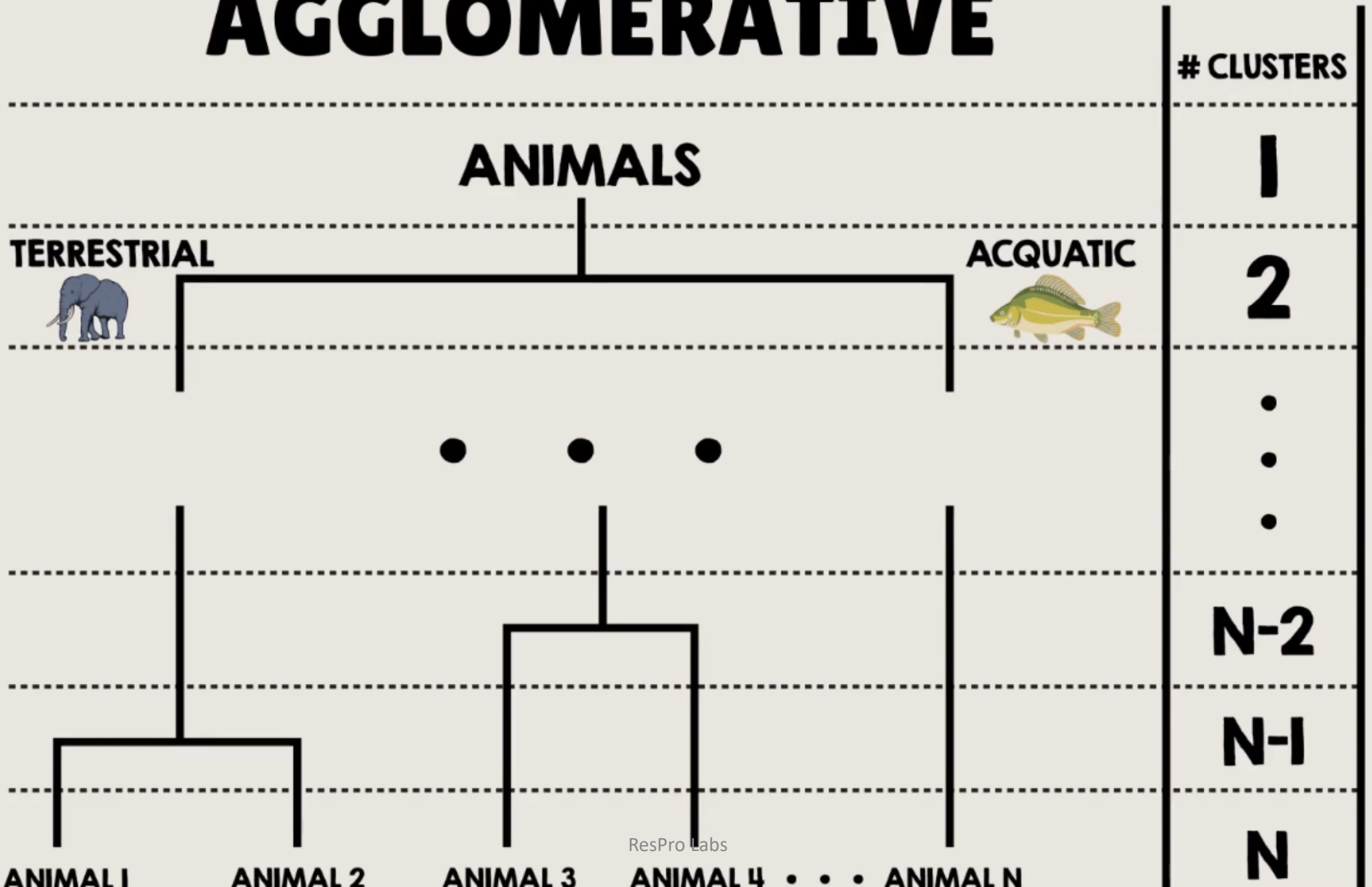
It is superior to flat clustering in the fact that it explores (contains) all solutions.

With divisive clustering we start from a situation where all observations are in the same cluster, e.g. from the dinosaurs. Then we split this big cluster into 2 smaller ones. Then we continue with 3, 4, 5, and so on, until each observation is its separate cluster.

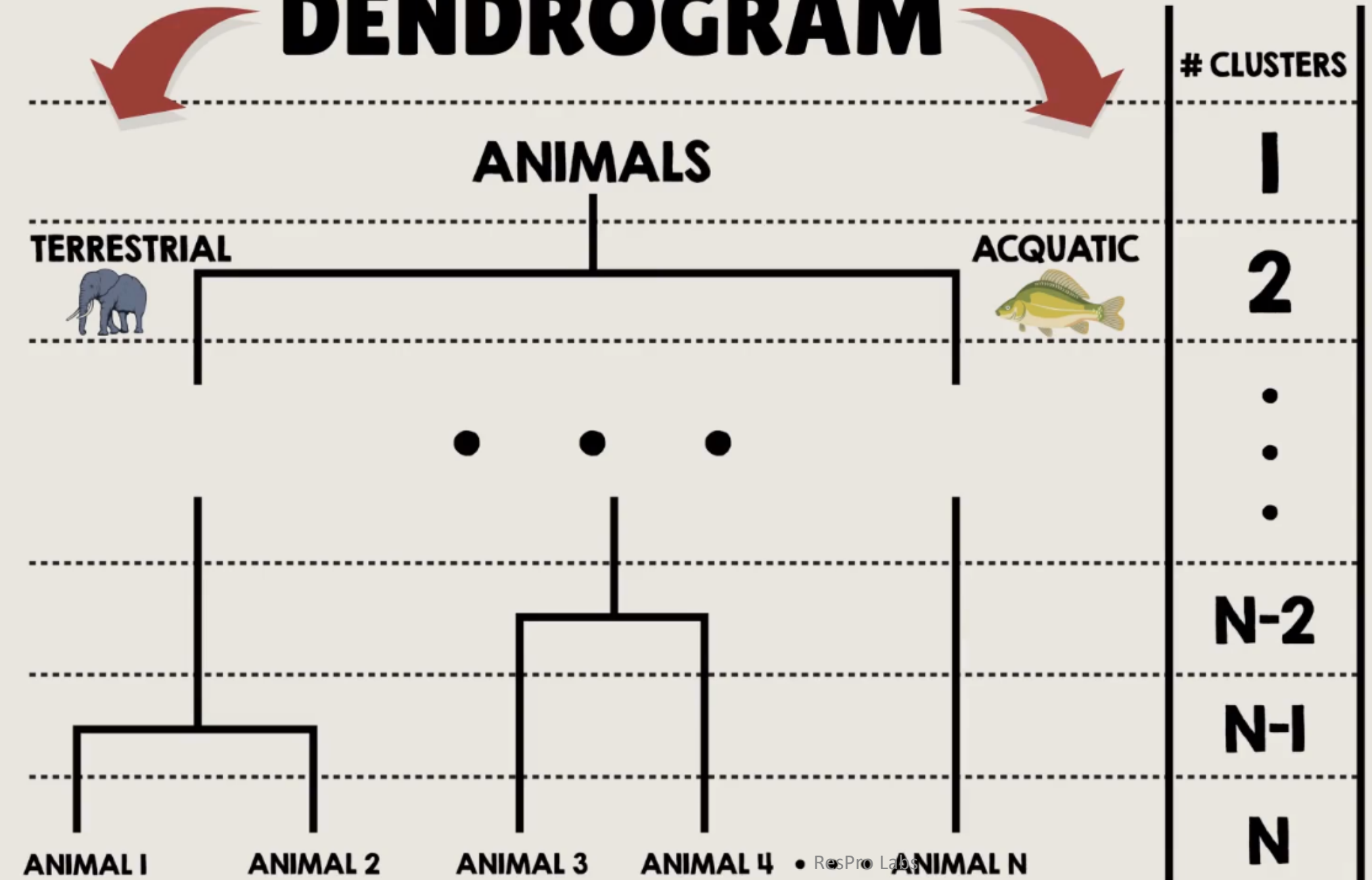
When it comes to agglomerative clustering, the approach is bottom up. We start from different dog and cat breeds, cluster them into dogs and cats respectively, and then we continue pairing up species, until we reach the animal cluster.

To find the combination of observations into a

AGGLOMERATIVE

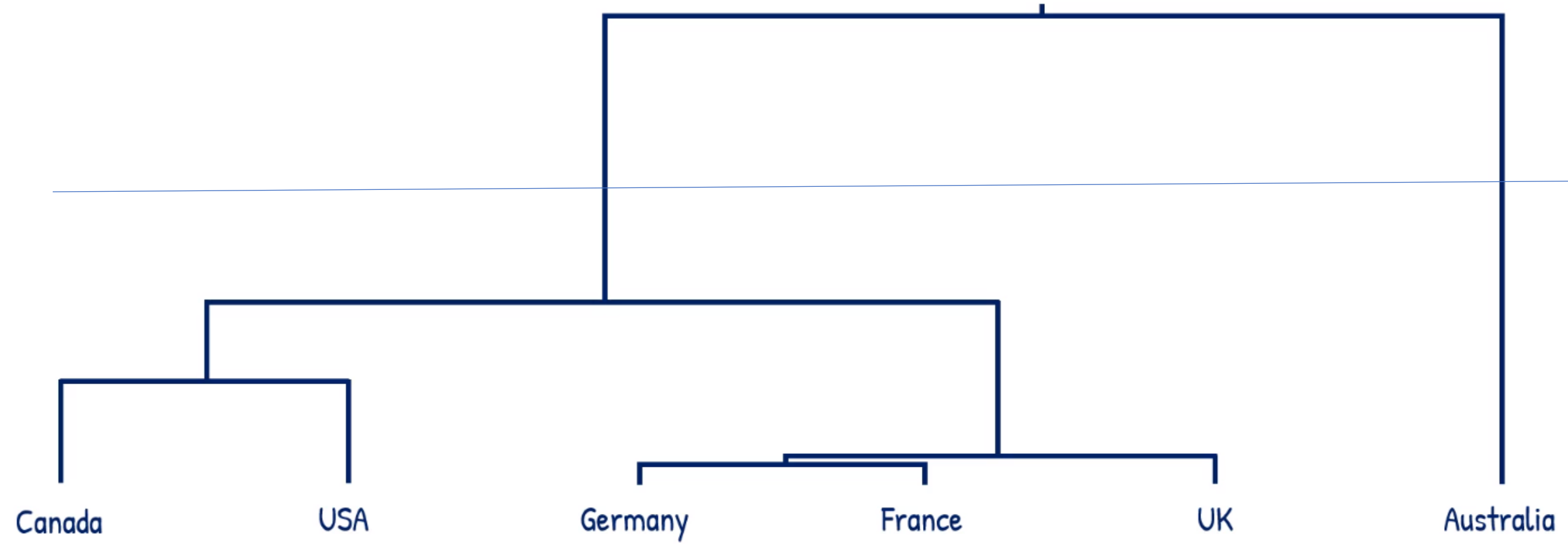


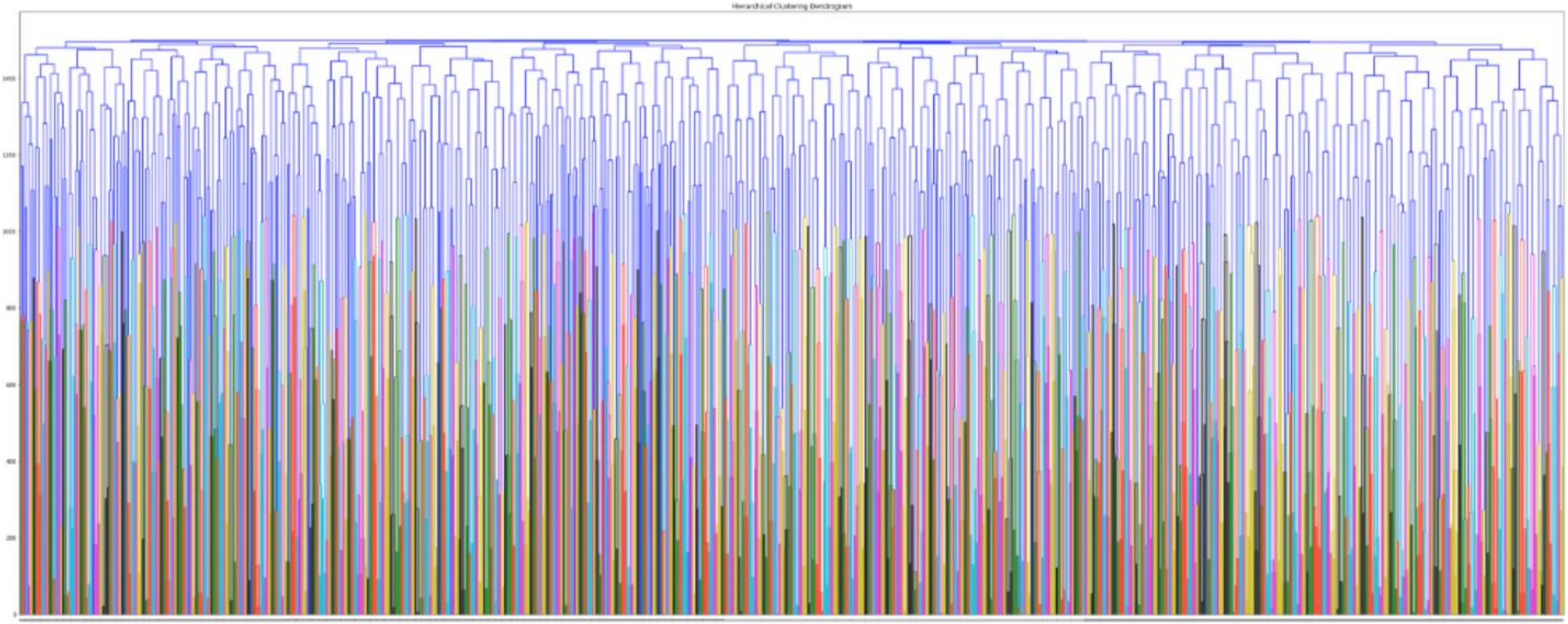
DENDROGRAM



ALL SOLUTIONS

DENDROGRAM




Hierarchical Clustering Dendrogram

K-means clustering - pros and cons

PROS

- Simple to implement
  (so many people can use it)
- Computationally efficient
  (it takes considerably less time than any hierarchical clustering model)
- Widely used
  (popular, therefore, in demand)
- Always yields a result
  (also a con as it may be deceiving)

CONS

- We need to pick K
  (often, we don't know how many clusters we need)
- Sensitive to initialization
  (but we can use methods such as kmeans++ to determine the seeds)
- Sensitive to outliers
  (by far the biggest downside of k-means)
- Produces spherical solutions