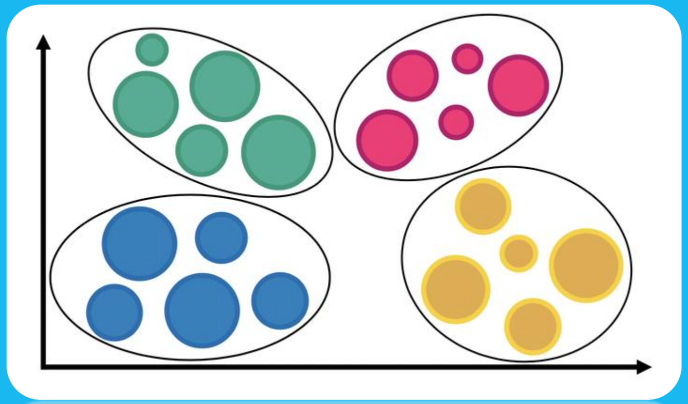
دوره‌ی آموزشی «علم داده»

Data Science Course

جلسه بیست و یکم - (بخش دوم)

ورود به دنیای خوشه‌بندی

Clustering vs. Classification

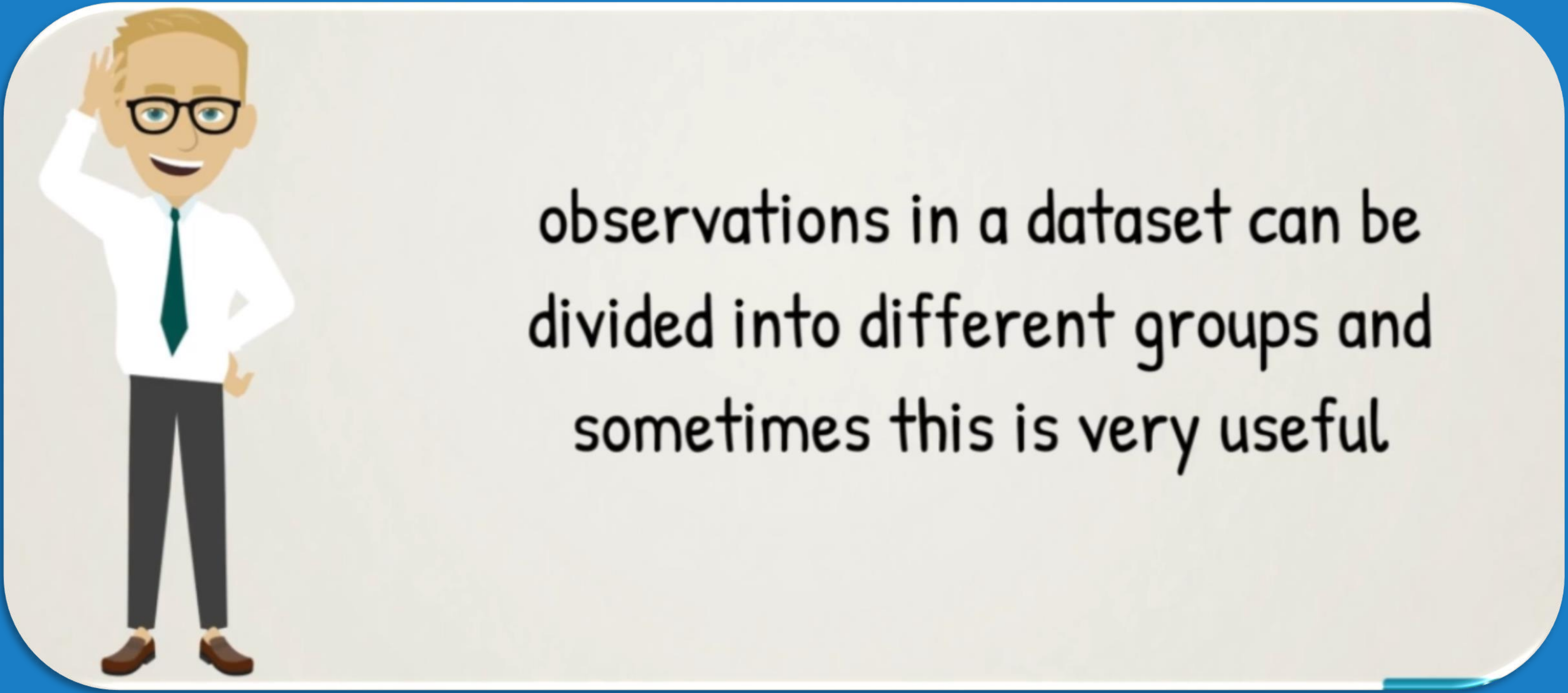


مدرس: محمد فزونی

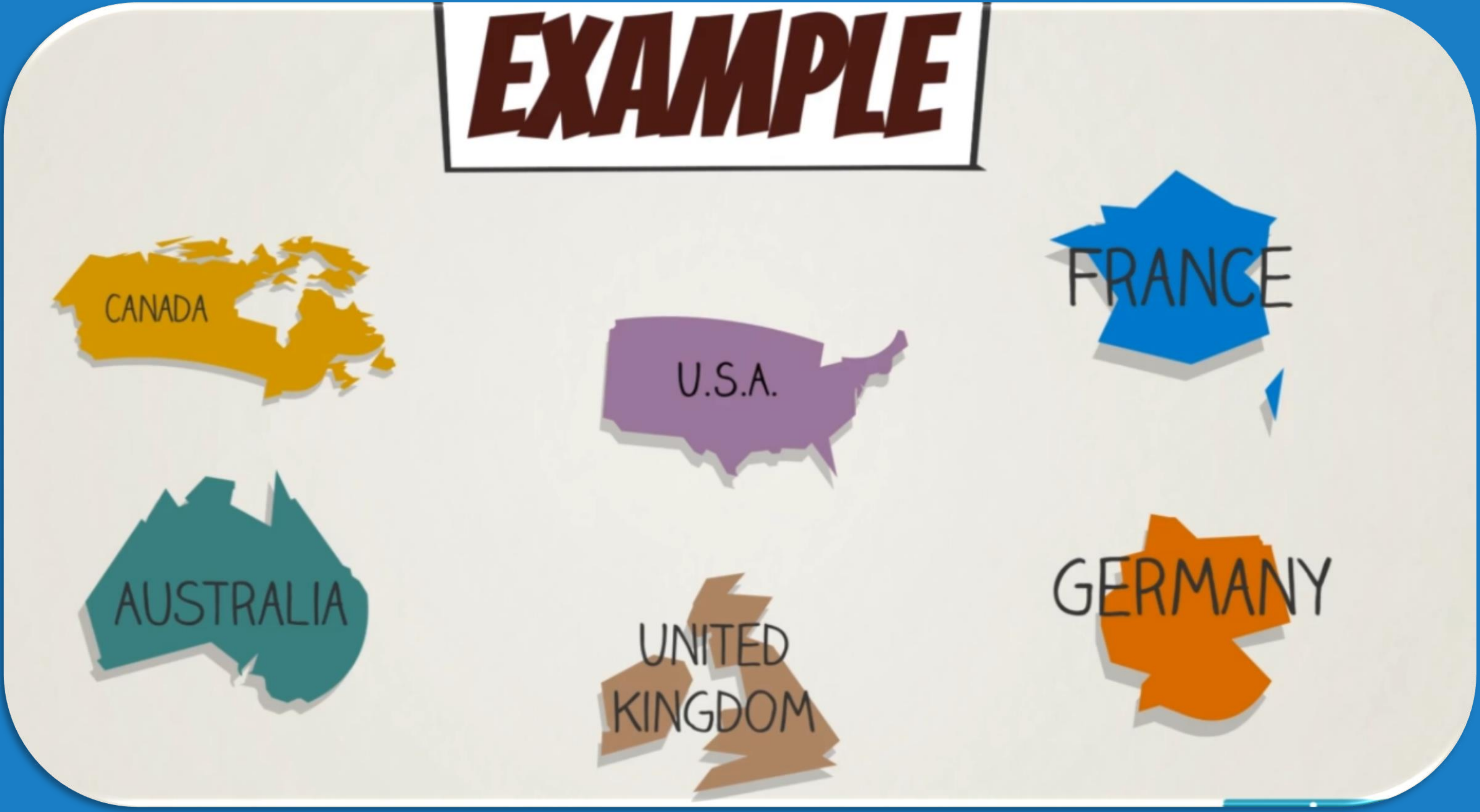
عضو هیئت علمی دانشگاه گنبدکاووس

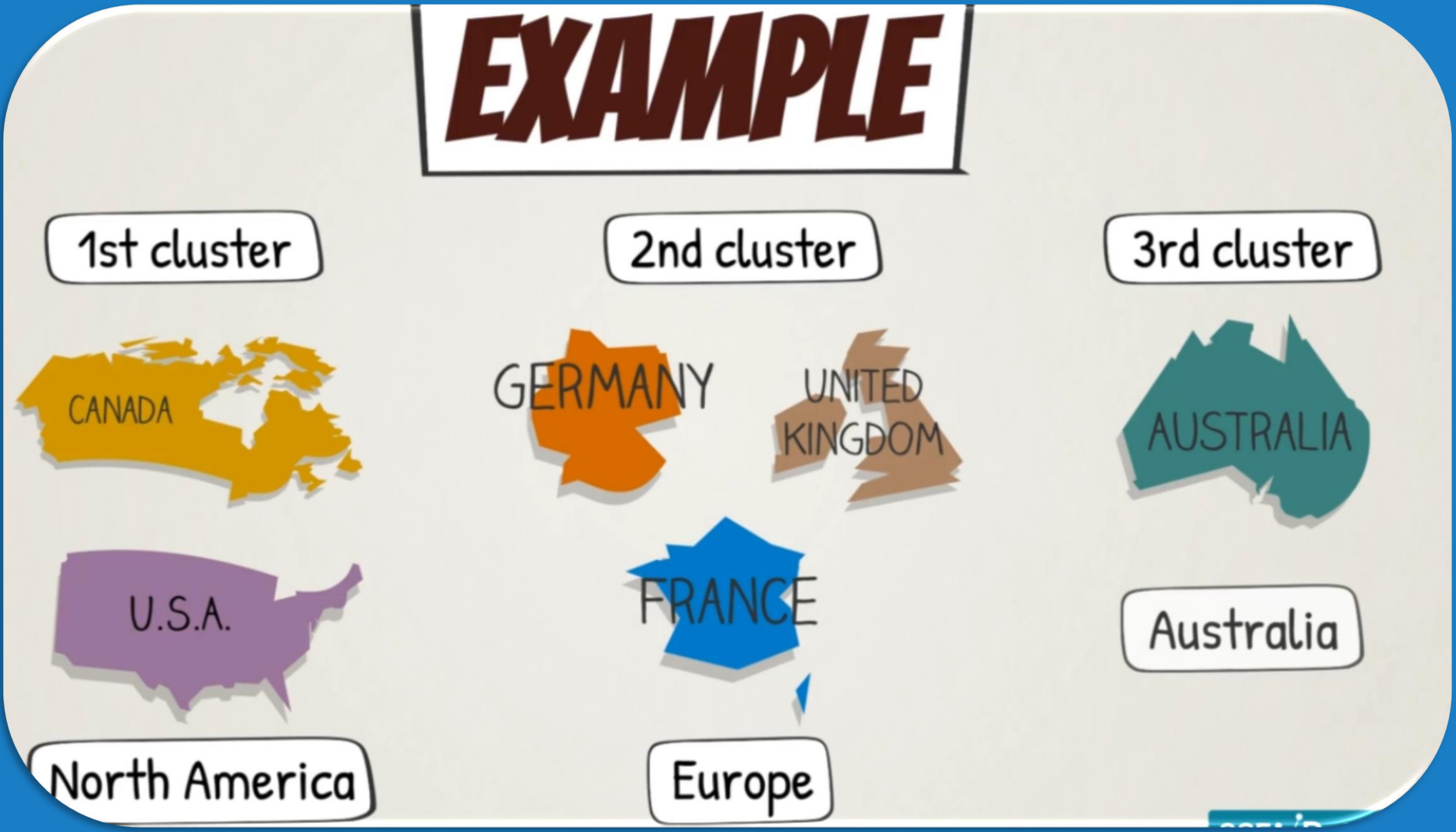
CLUSTER ANALYSIS

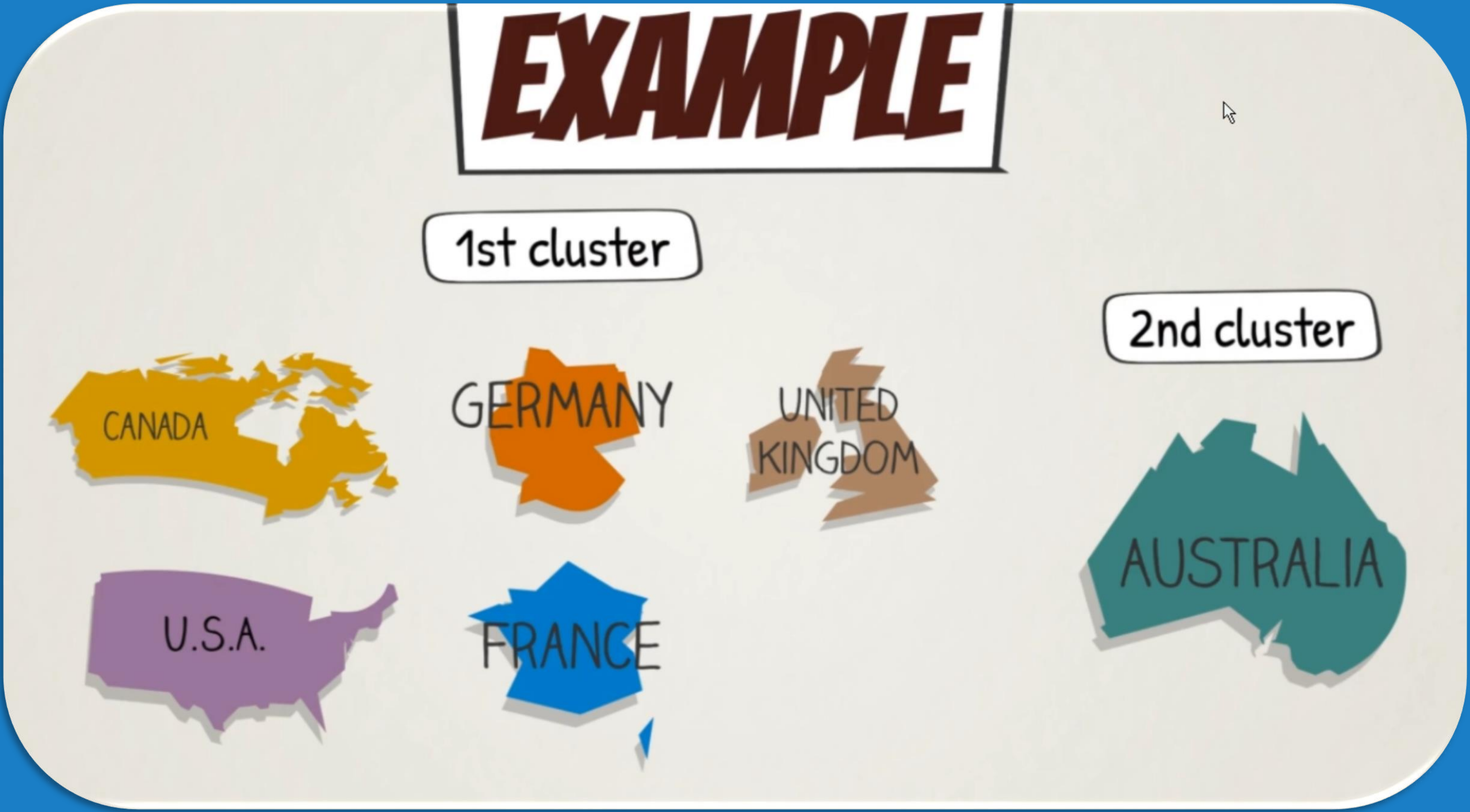
Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group are more similar to each other than to those in other groups.

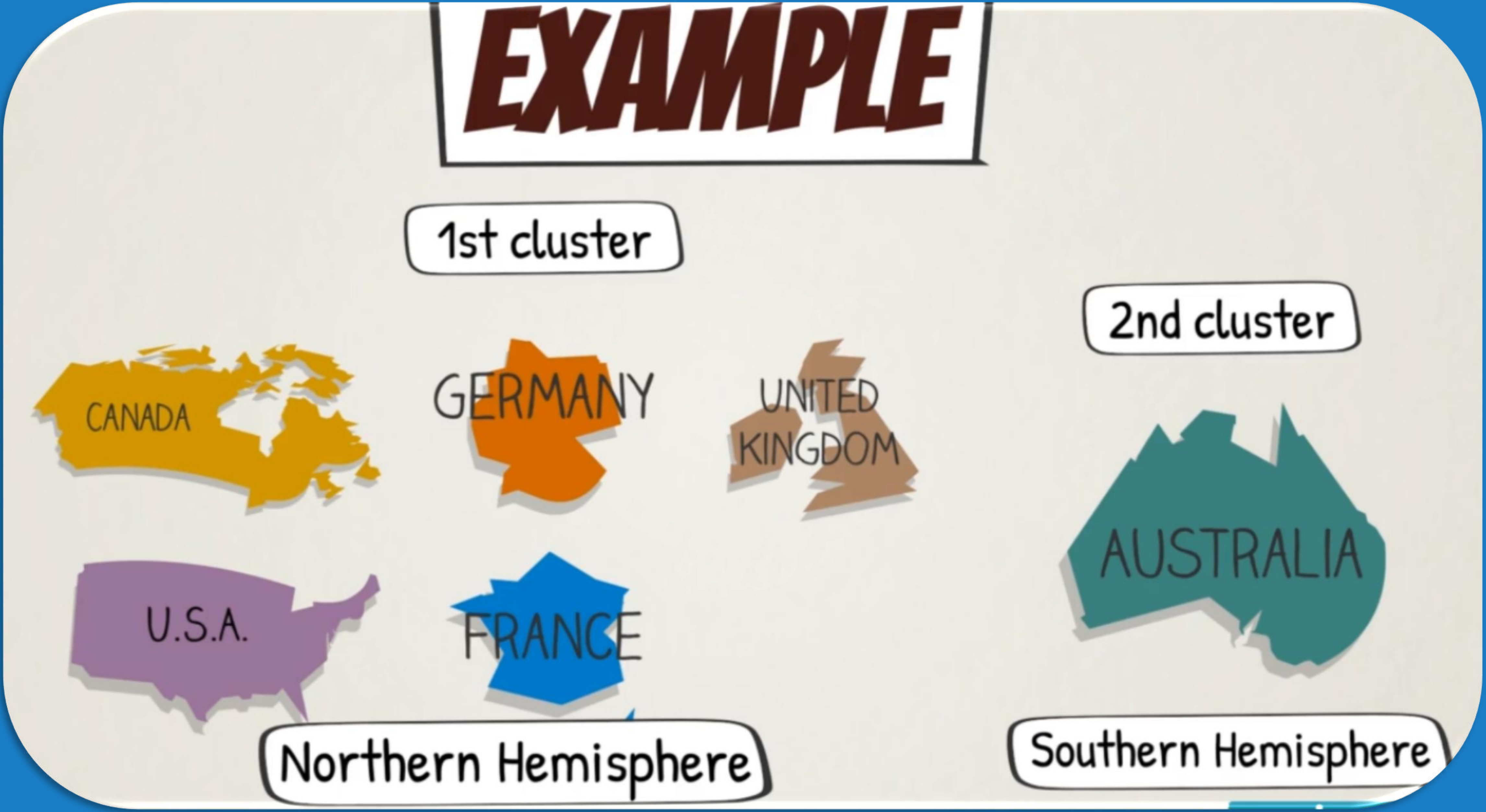


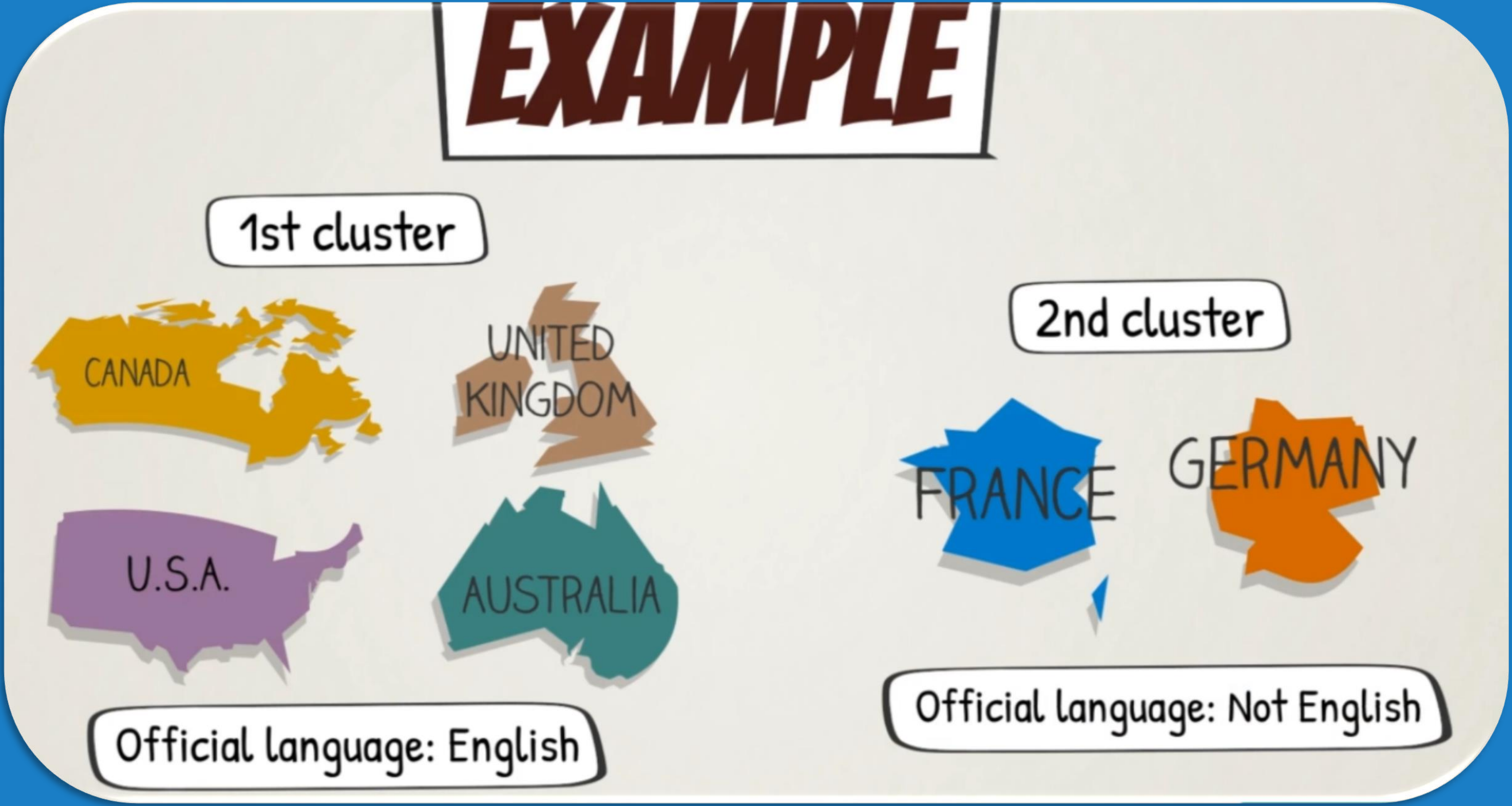
observations in a dataset can be divided into different groups and sometimes this is very useful

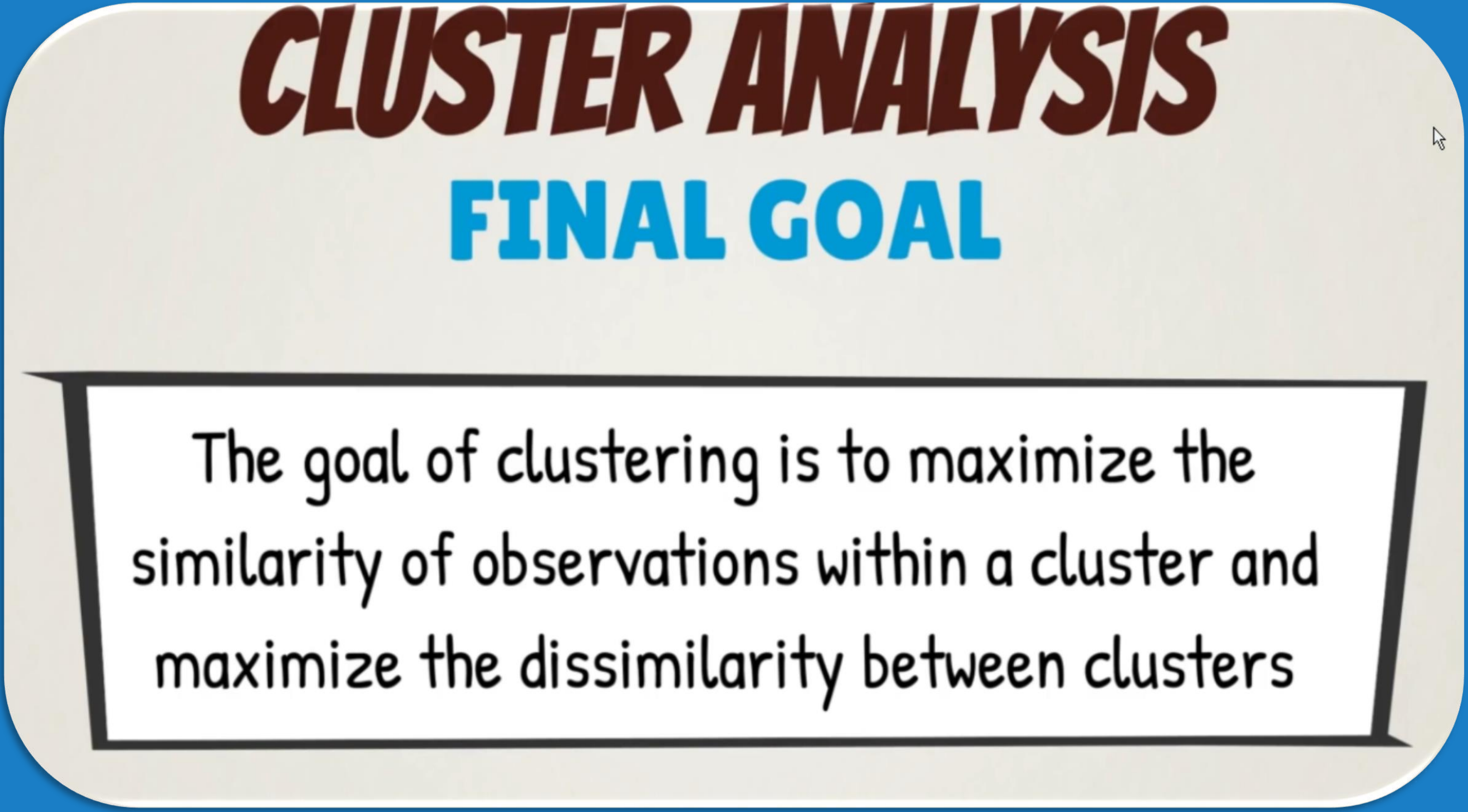

EXAMPLE
CANADA
U.S.A.
FRANCE
AUSTRALIA
UNITED KINGDOM
GERMANY


EXAMPLE
1st cluster
2nd cluster
3rd cluster
CANADA
GERMANY
UNITED KINGDOM
AUSTRALIA
U.S.A.
FRANCE
Australia
North America
Europe


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


EXAMPLE
1st cluster
2nd cluster
CANADA
GERMANY
UNITED KINGDOM
AUSTRALIA
U.S.A.
FRANCE
Northern Hemisphere
Southern Hemisphere


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

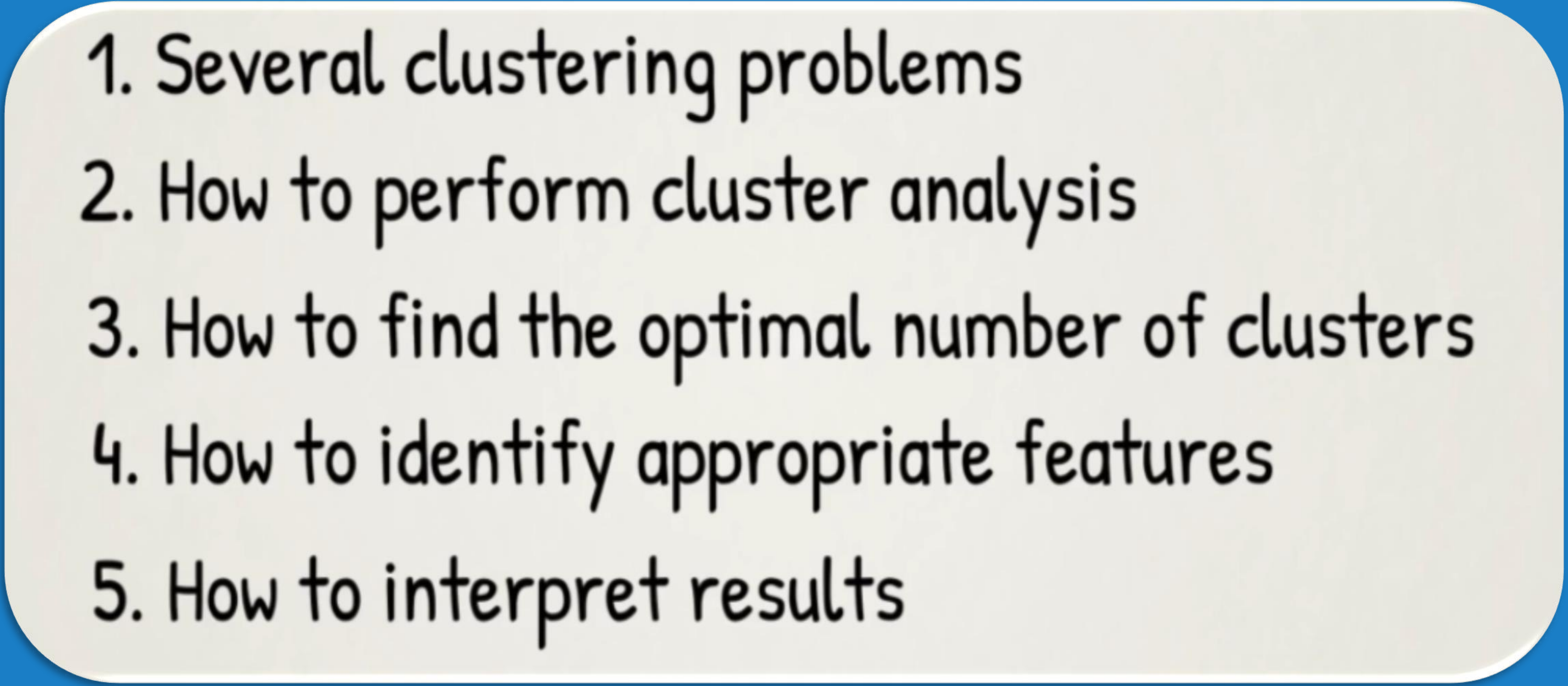


CLUSTER ANALYSIS

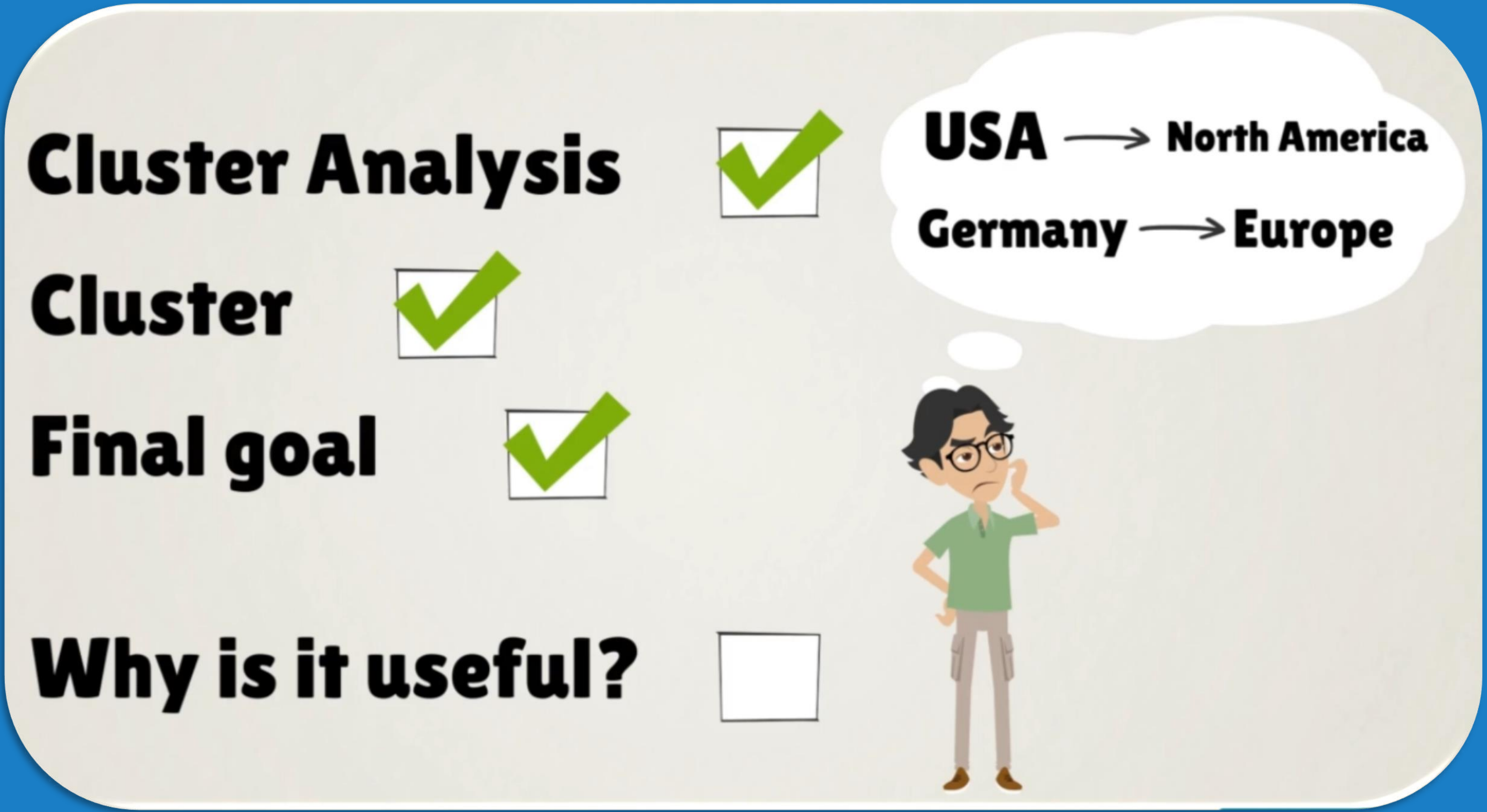
FINAL GOAL

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

What are we going to do in the sequel?



1. Several clustering problems
2. How to perform cluster analysis
3. How to find the optimal number of clusters
4. How to identify appropriate features
5. How to interpret results



Cluster Analysis ☑

Cluster ☑

Final goal ☑

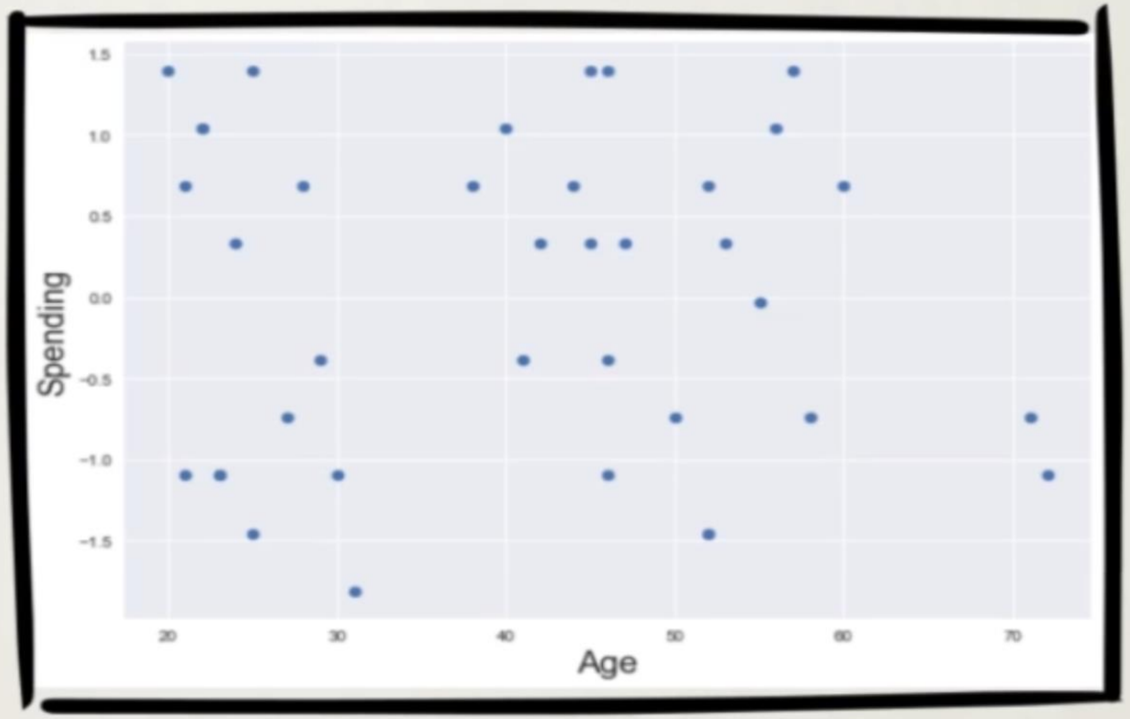
Why is it useful? □

USA → North America

Germany → Europe

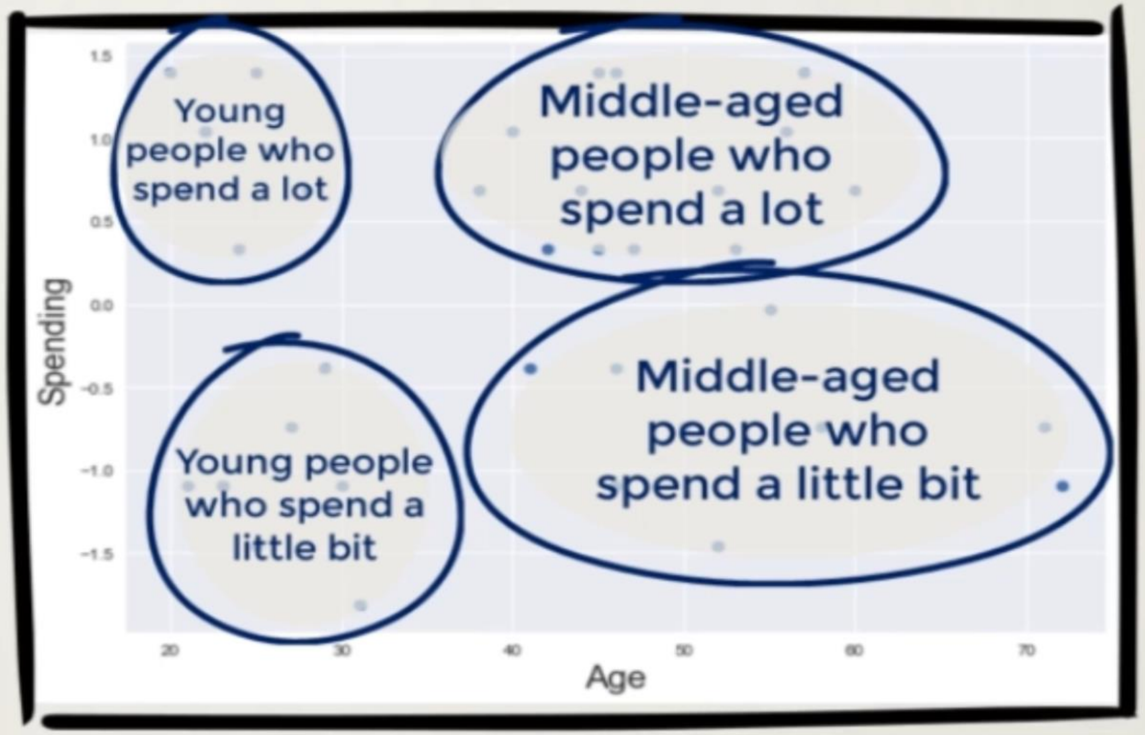
Market segmentation

Scatter plot



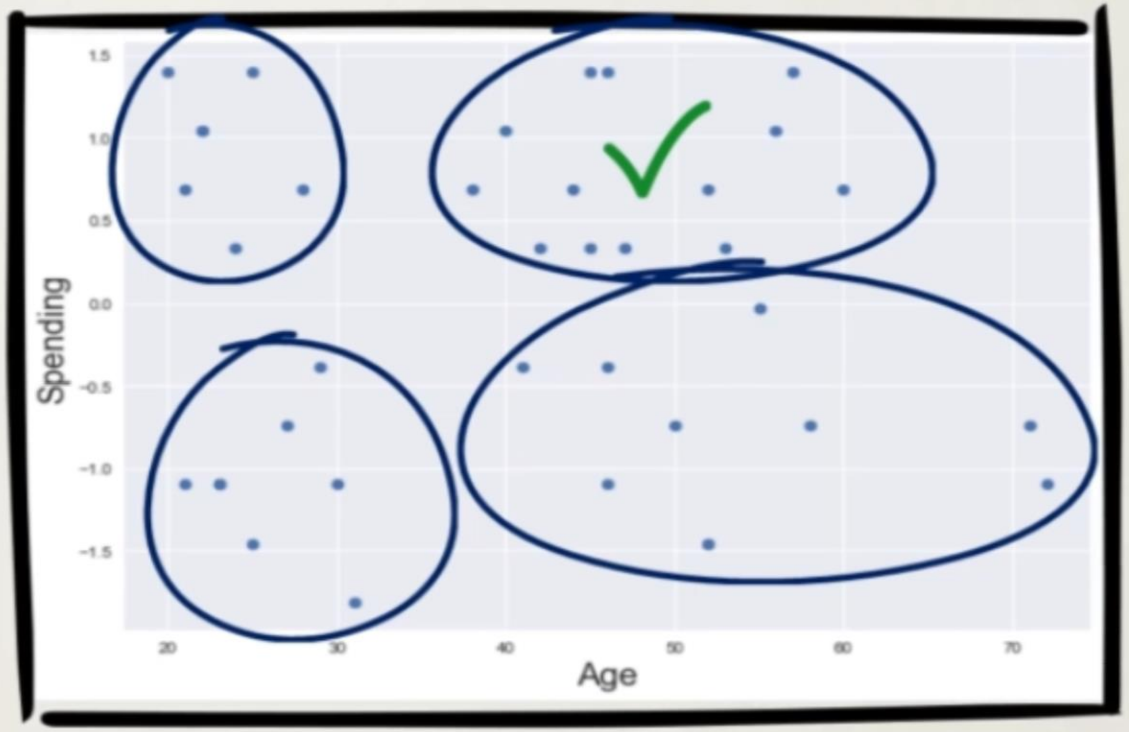
Market segmentation

Scatter plot

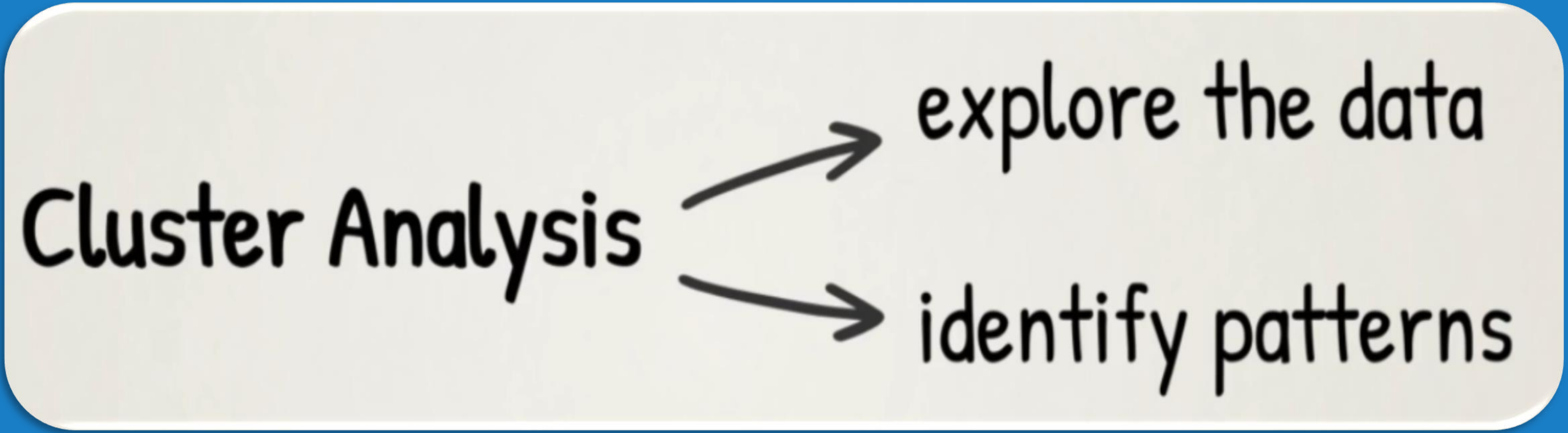

Young people who spend a lot
Middle-aged people who spend a lot
Young people who spend a little bit
Middle-aged people who spend a little bit
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
Spending
20
30
40
50
60
70
Age

Market segmentation

Scatter plot



So, briefly speaking …

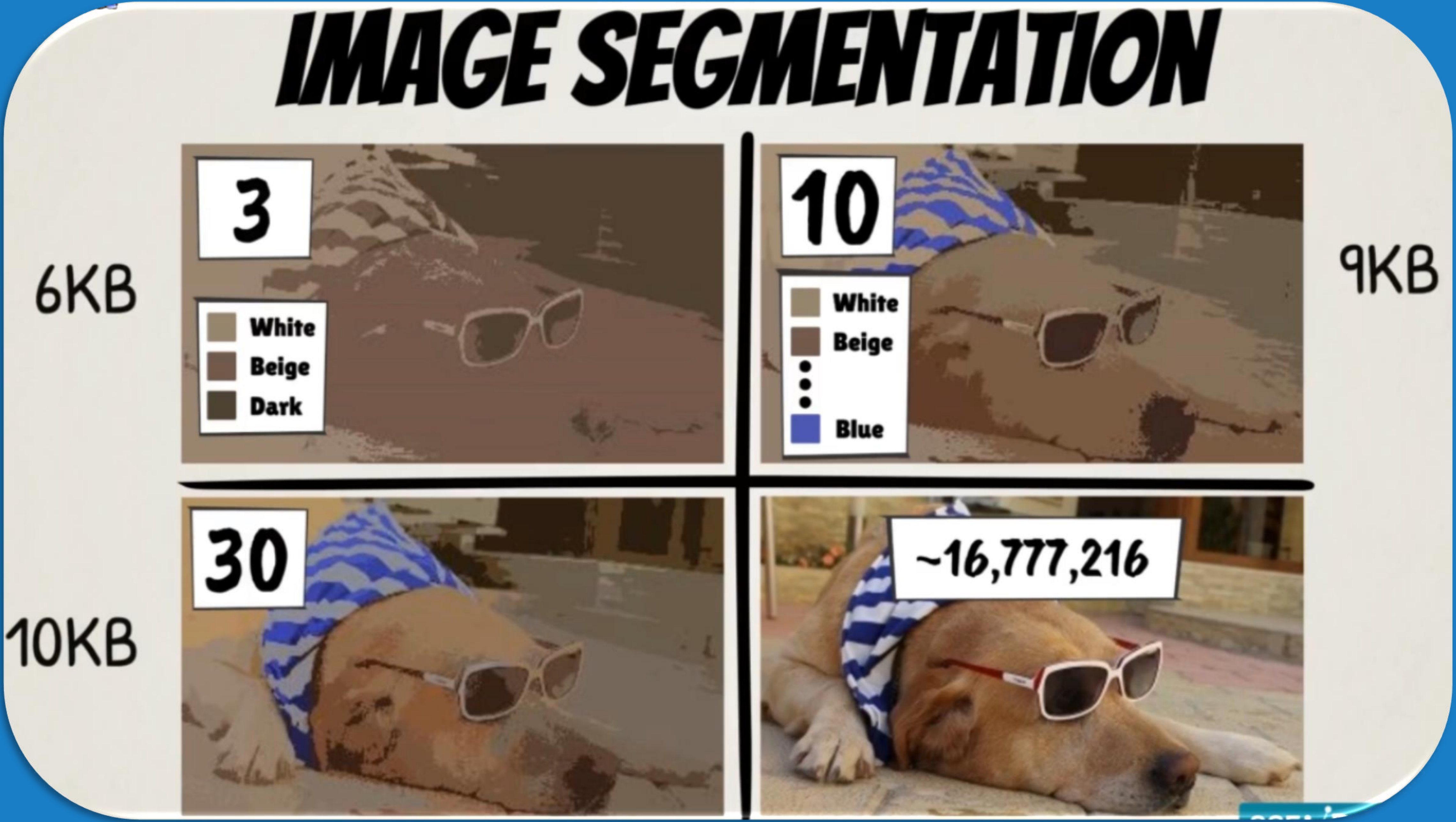


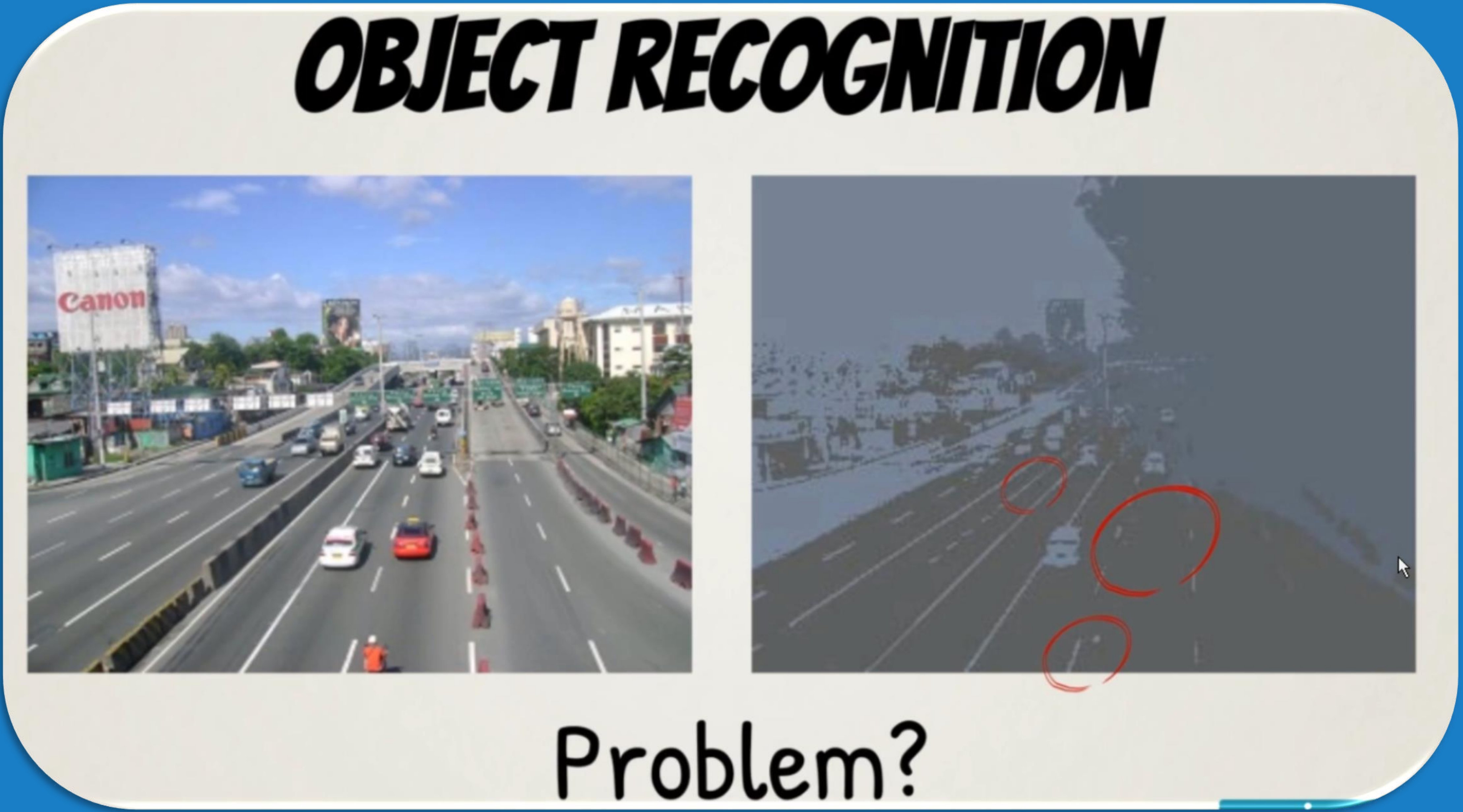
Cluster Analysis

explore the data

identify patterns

IMAGE SEGMENTATION

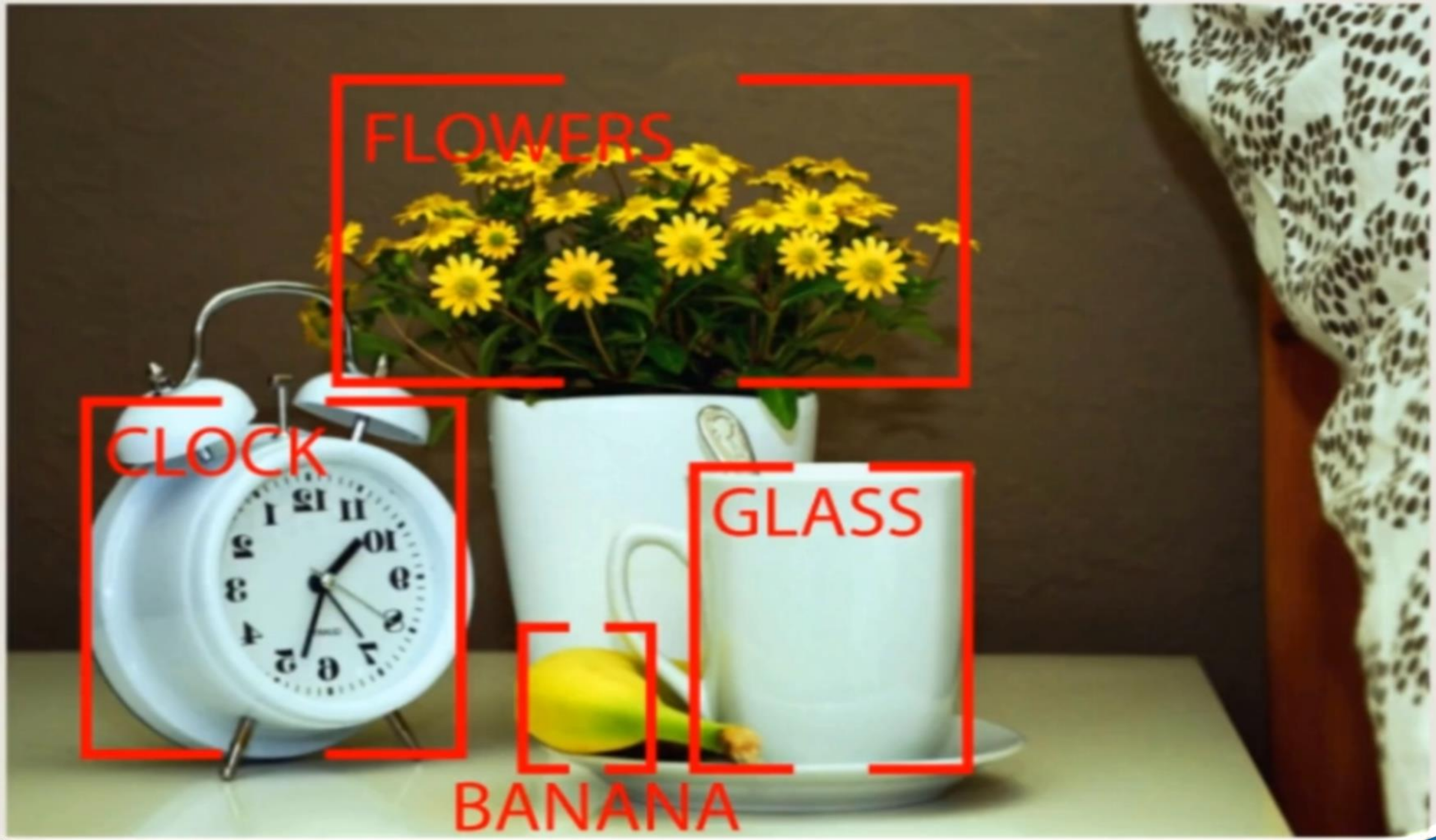




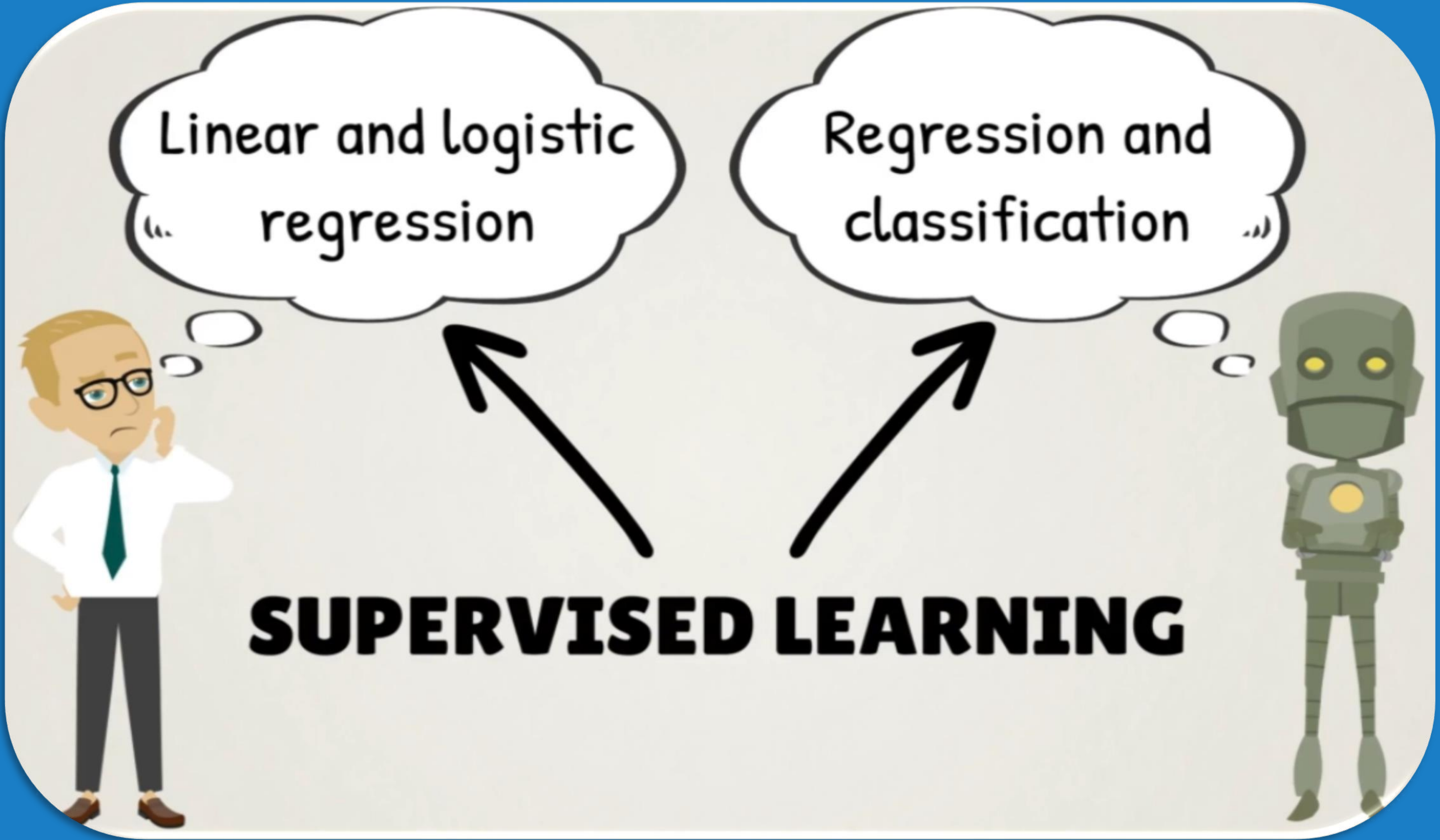
OBJECT RECOGNITION

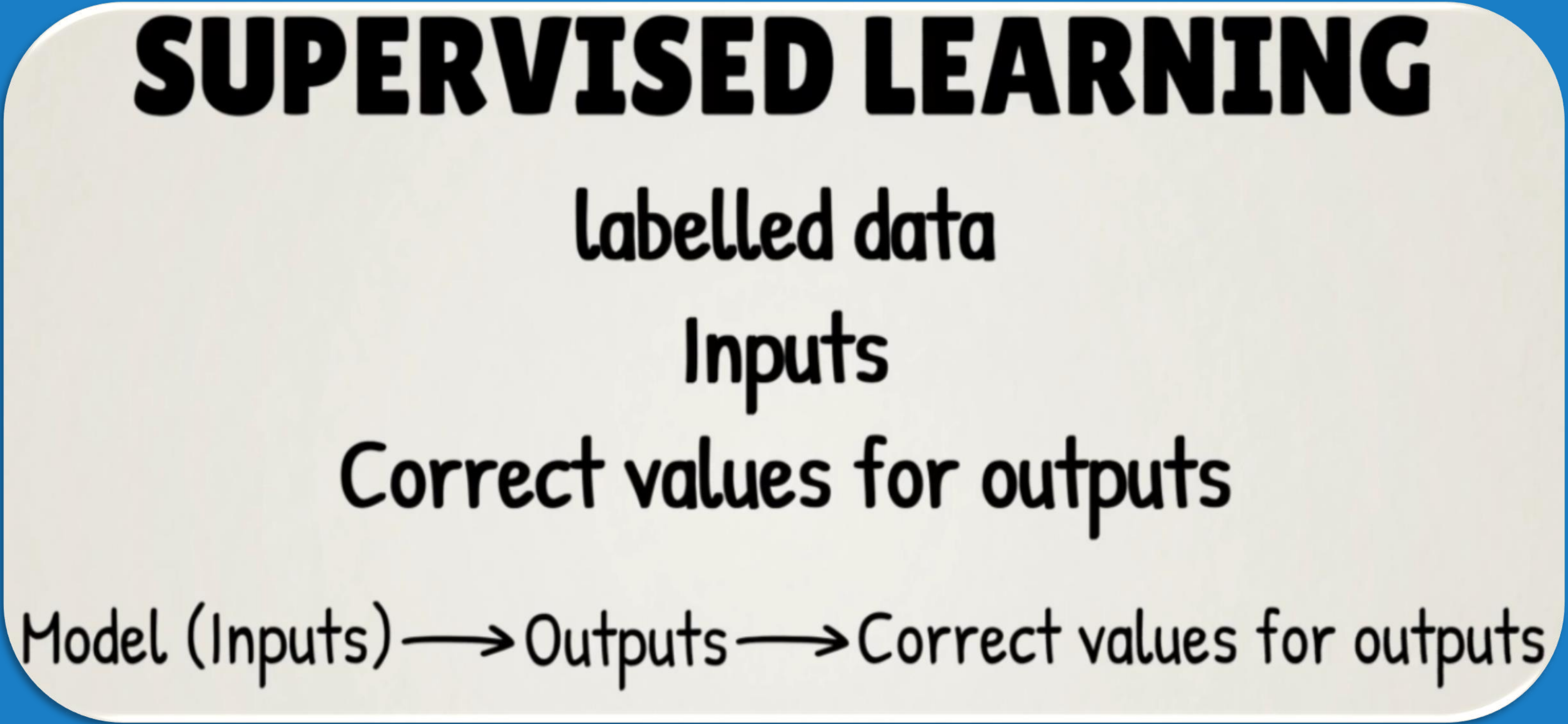
Problem?

OBJECT RECOGNITION



Linear and logistic regression
Regression and classification


Linear and logistic regression
Regression and classification
SUPERVISED LEARNING



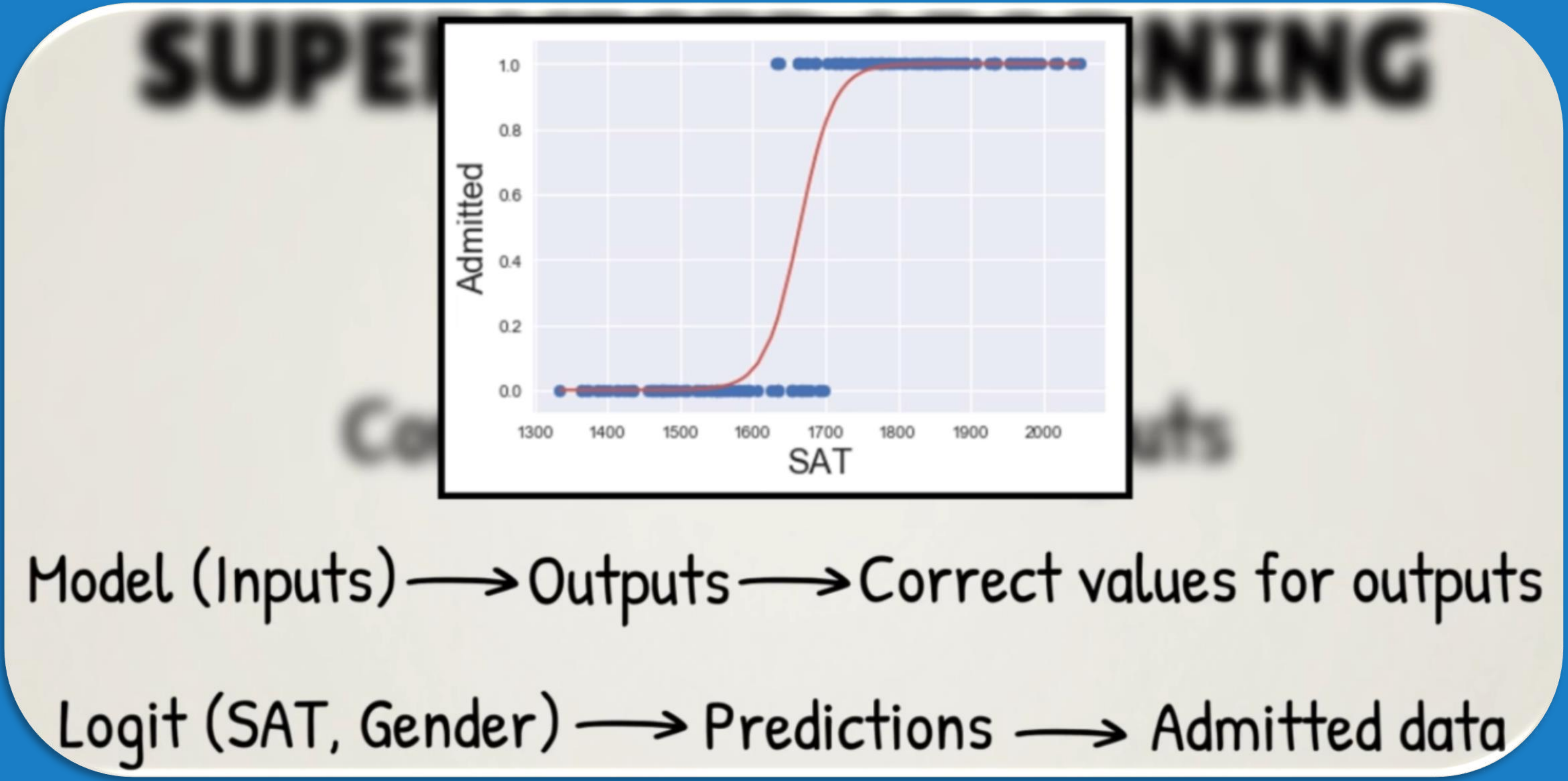
SUPERVISED LEARNING

labelled data

Inputs

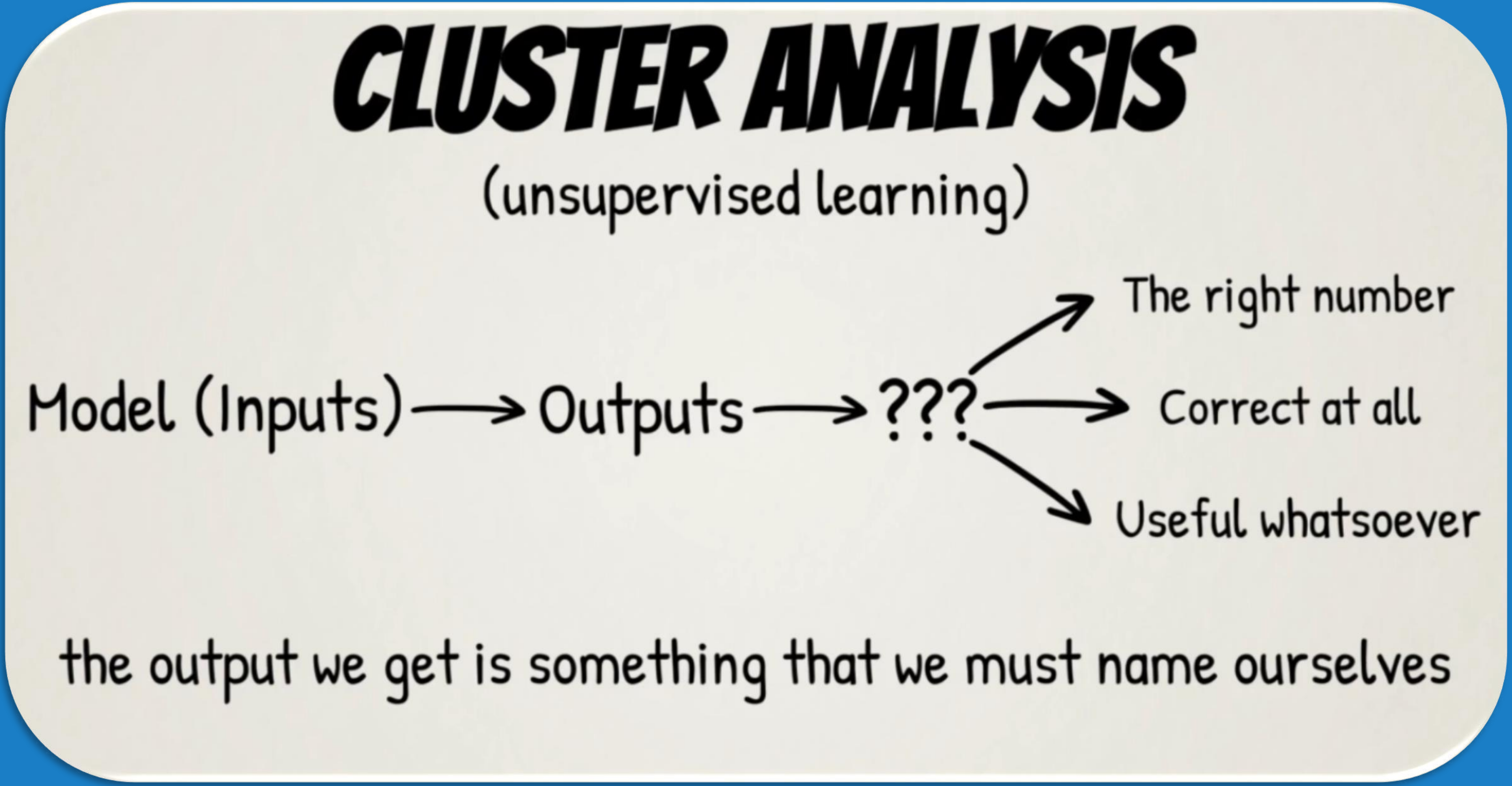
Correct values for outputs

Model (Inputs) → Outputs → Correct values for outputs



Model (Inputs) → Outputs → Correct values for outputs

Logit (SAT, Gender) → Predictions → Admitted data



CLUSTER ANALYSIS

(unsupervised learning)

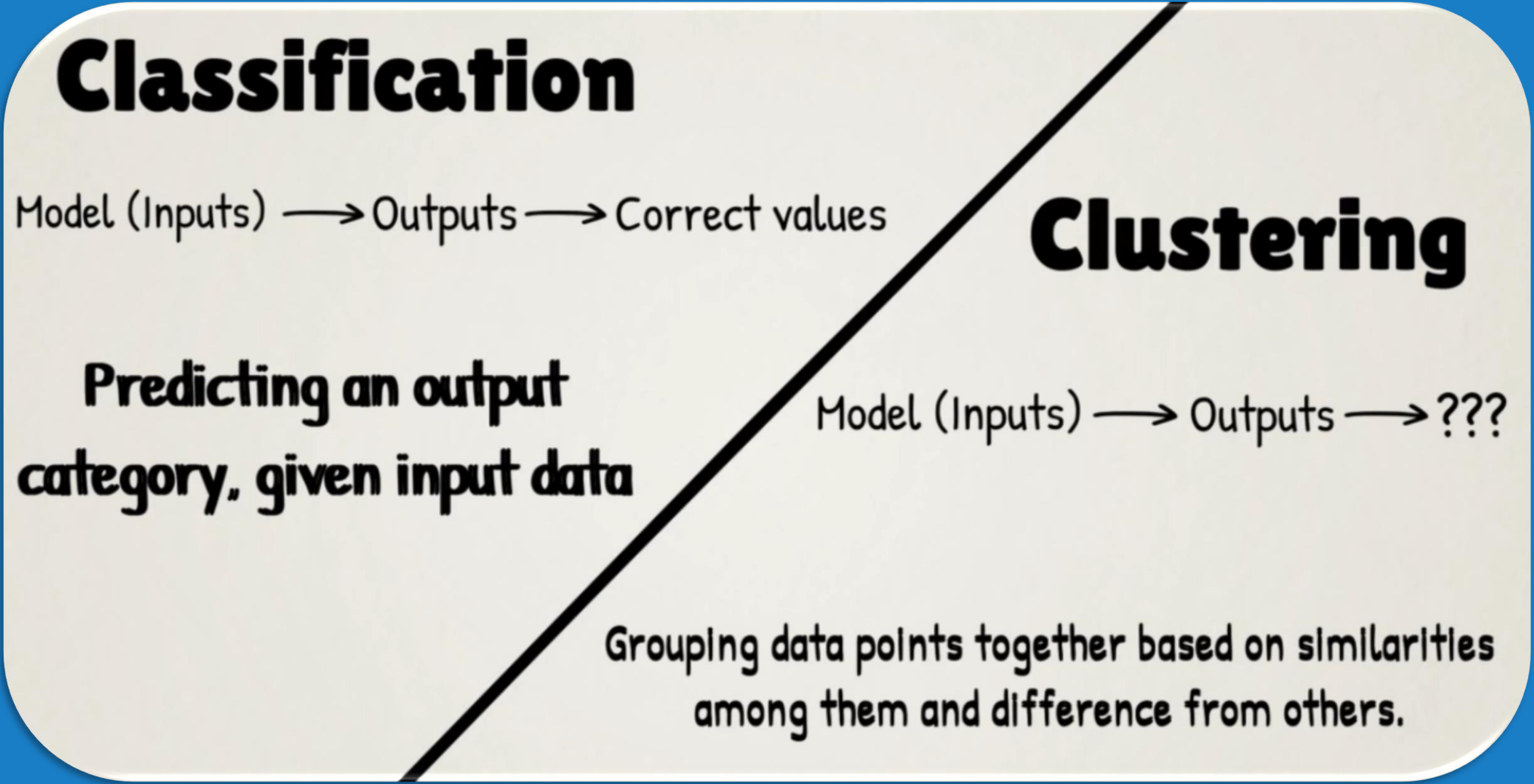
Model (Inputs) → Outputs → ???

??? → The right number

??? → Correct at all

??? → Useful whatsoever

the output we get is something that we must name ourselves



Classification

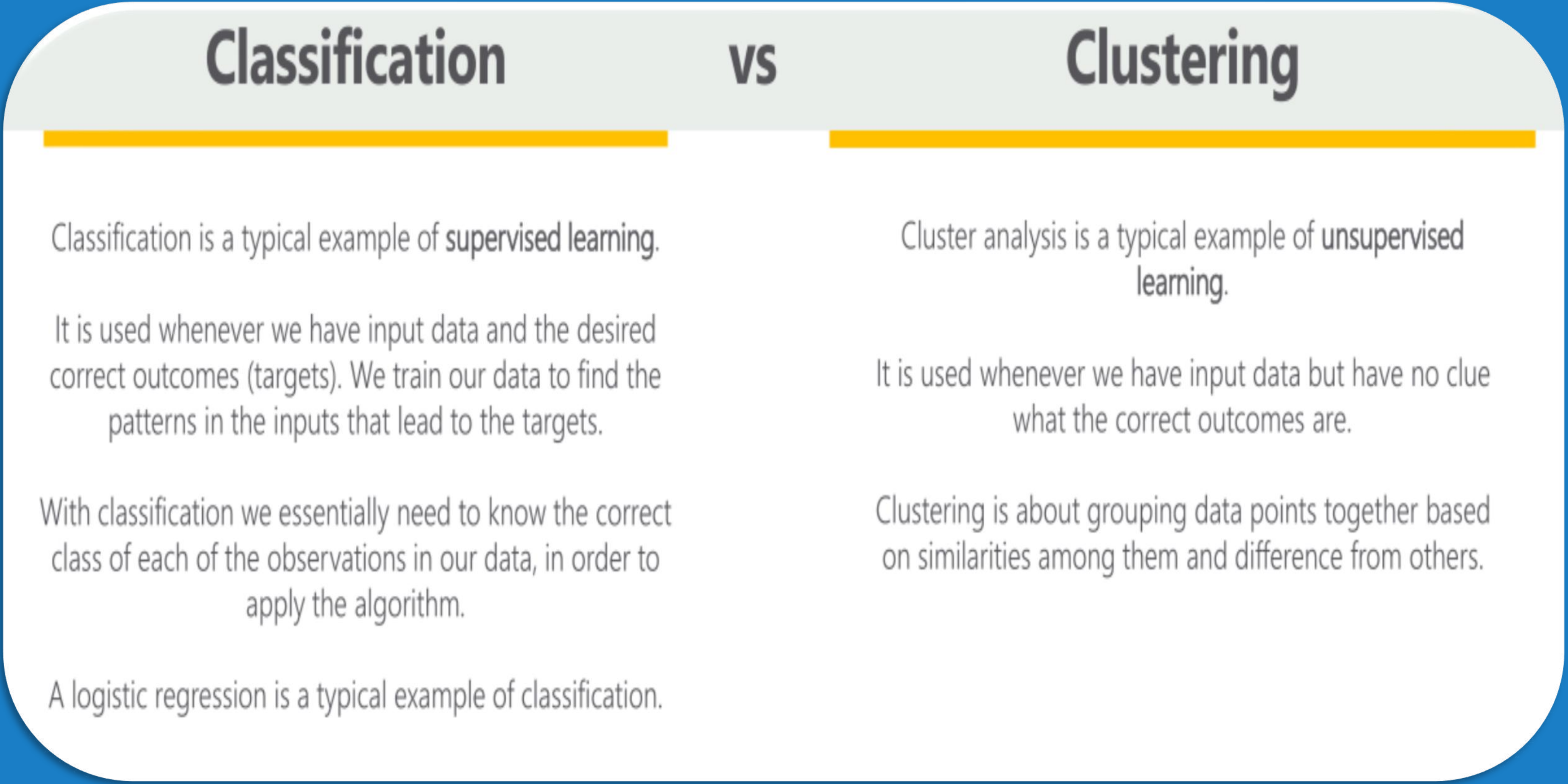
Model (Inputs) → Outputs → Correct values

Predicting an output category, given input data

Clustering

Model (Inputs) → Outputs → ???

Grouping data points together based on similarities among them and difference from others.



Classification vs Clustering

Classification is a typical example of **supervised learning**.

It is used whenever we have input data and the desired correct outcomes (targets). We train our data to find the patterns in the inputs that lead to the targets.

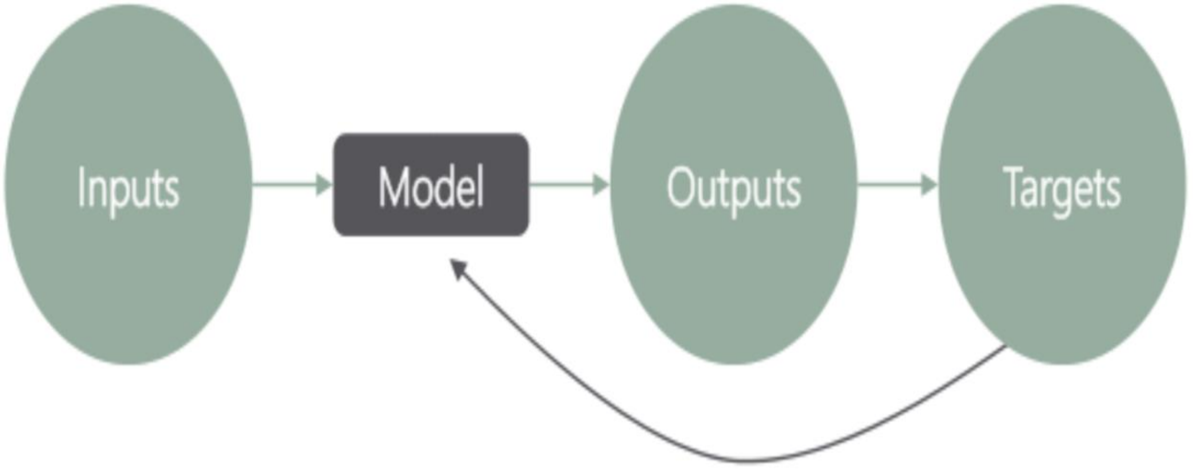
With classification we essentially need to know the correct class of each of the observations in our data, in order to apply the algorithm.

A logistic regression is a typical example of classification.

Cluster analysis is a typical example of **unsupervised learning**.

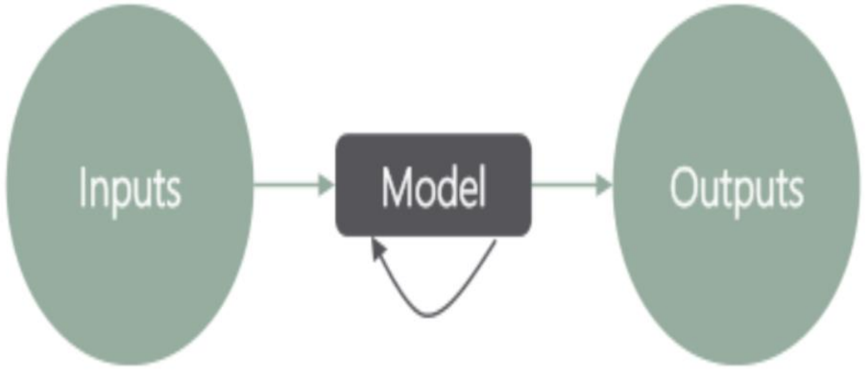
It is used whenever we have input data but have no clue what the correct outcomes are.

Clustering is about grouping data points together based on similarities among them and difference from others.



We use the targets (correct values) to adjust the model to get better outputs.

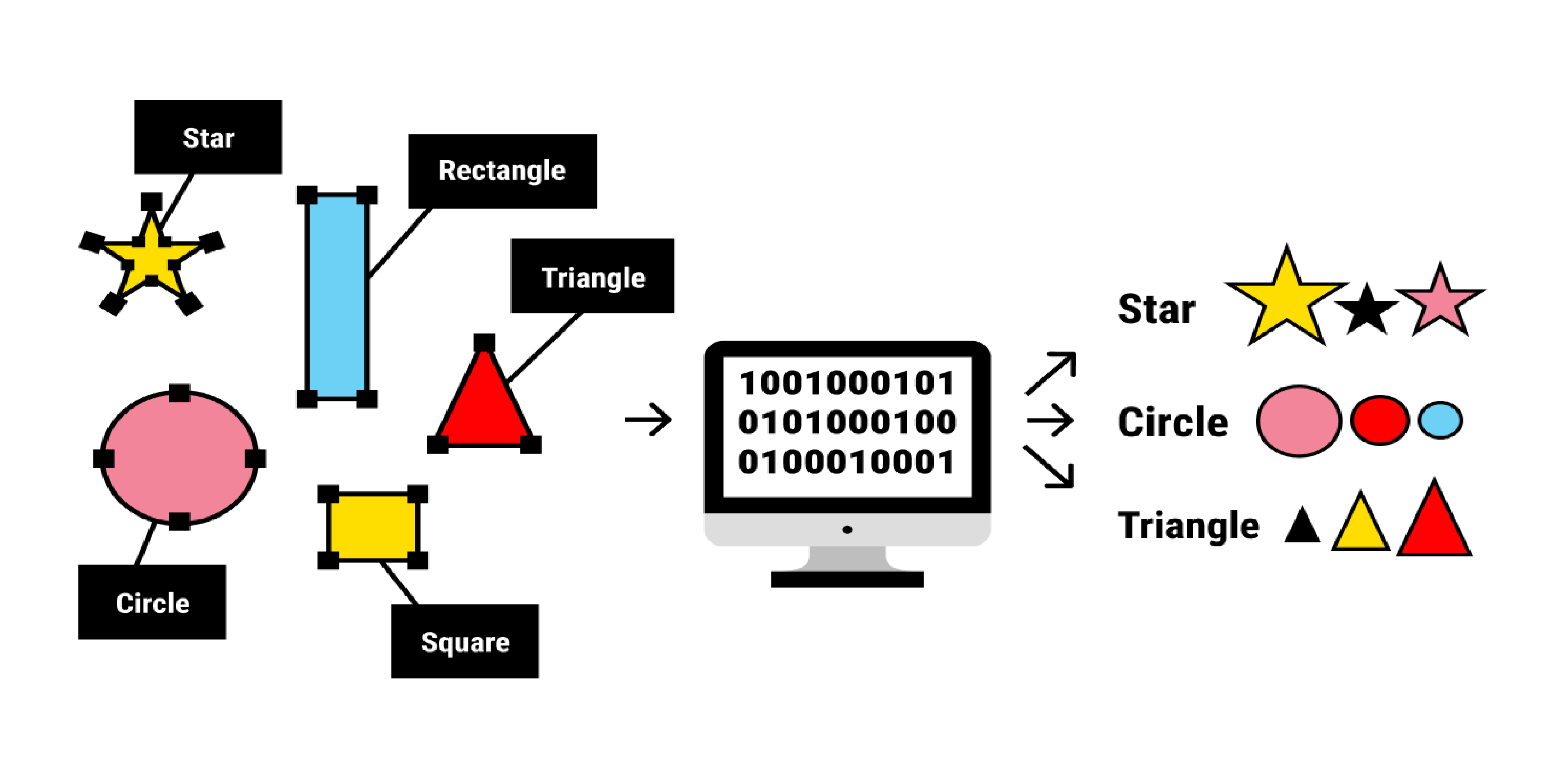
Supervised Learning



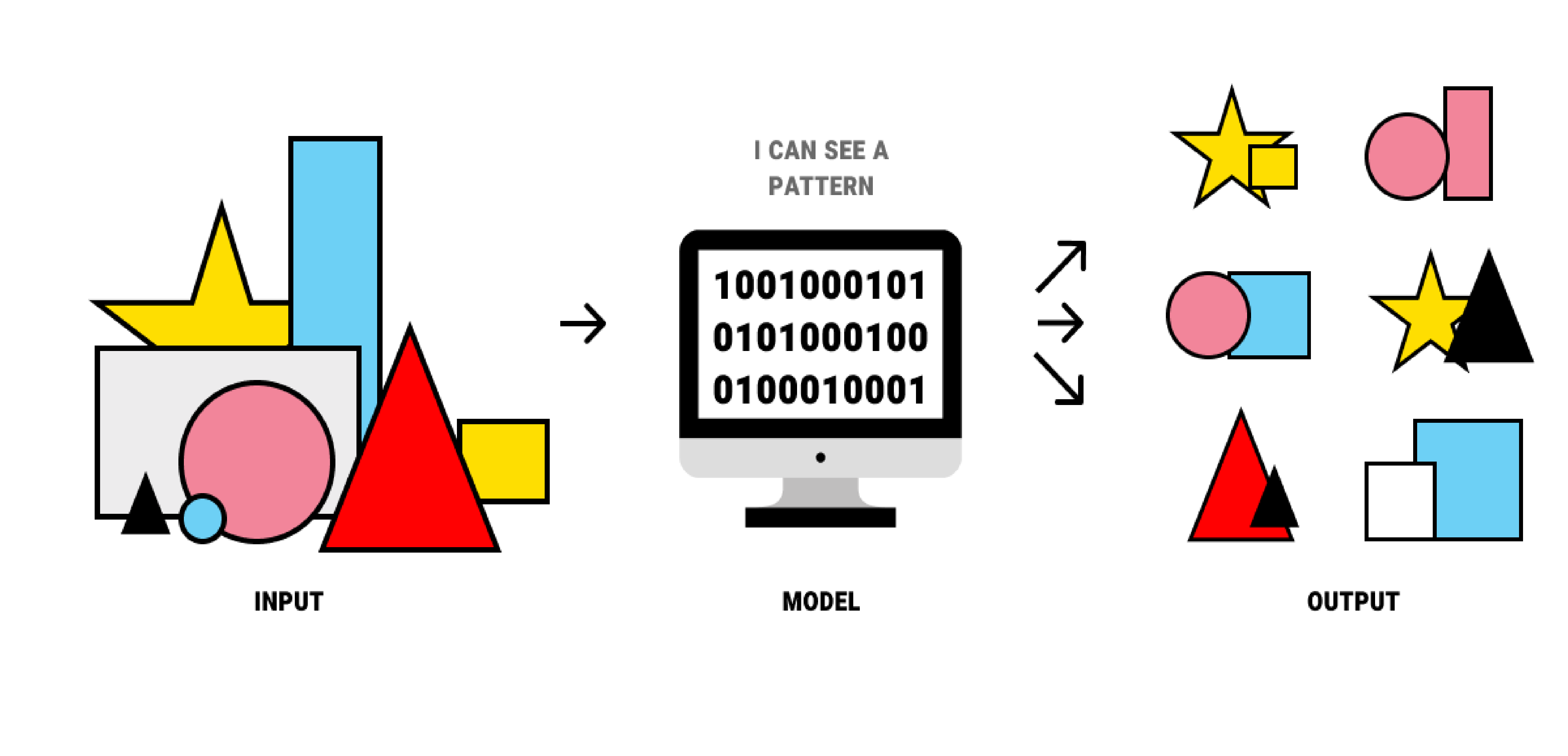
There is no feedback loop, therefore, the model simply finds the outputs it deems best.

Unsupervised Learning

Classification



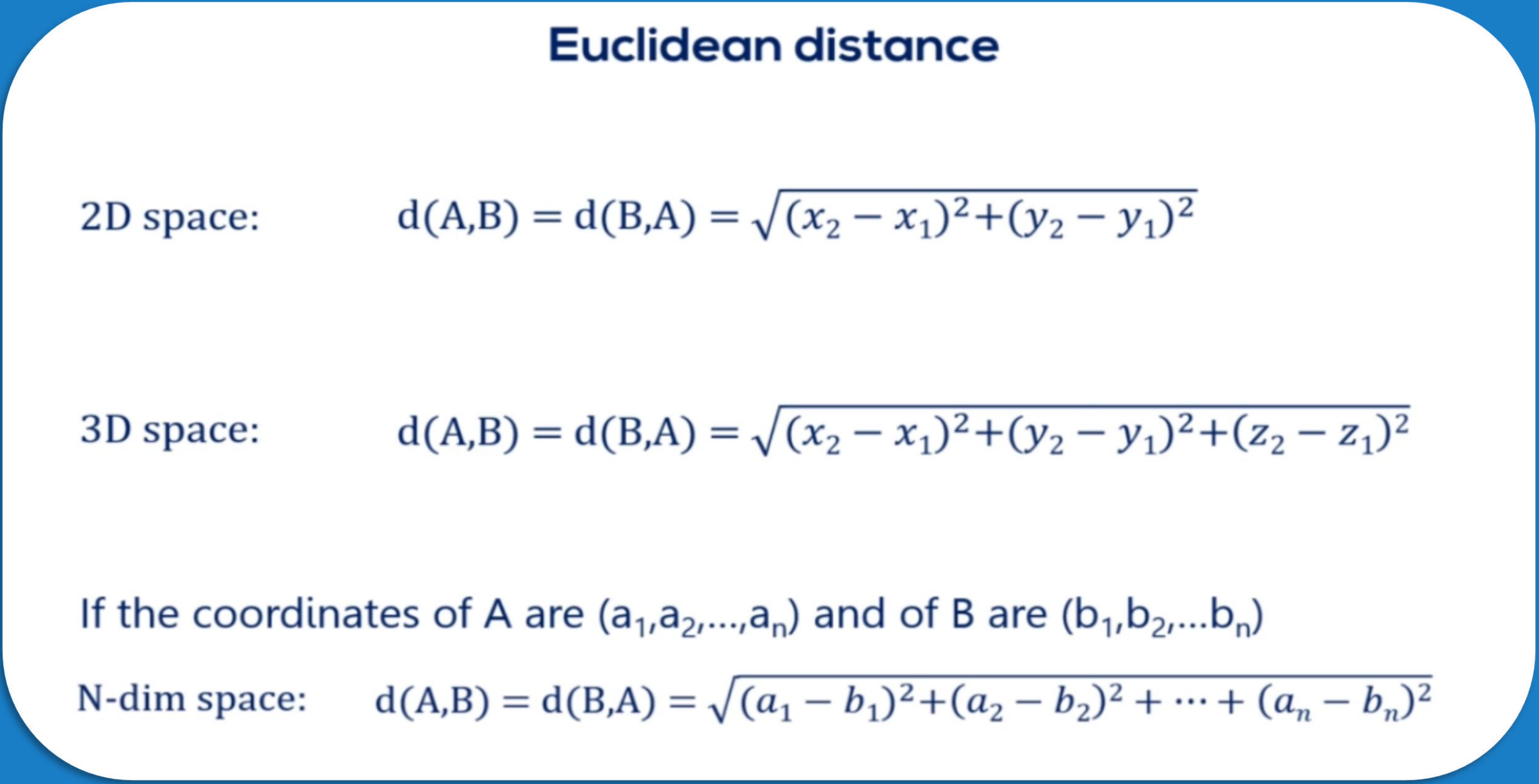
Clustering


I CAN SEE A PATTERN
1001000101
0101000100
0100010001
INPUT
MODEL
OUTPUT

MATH PREREQUISITES

distance between two data points

centroid



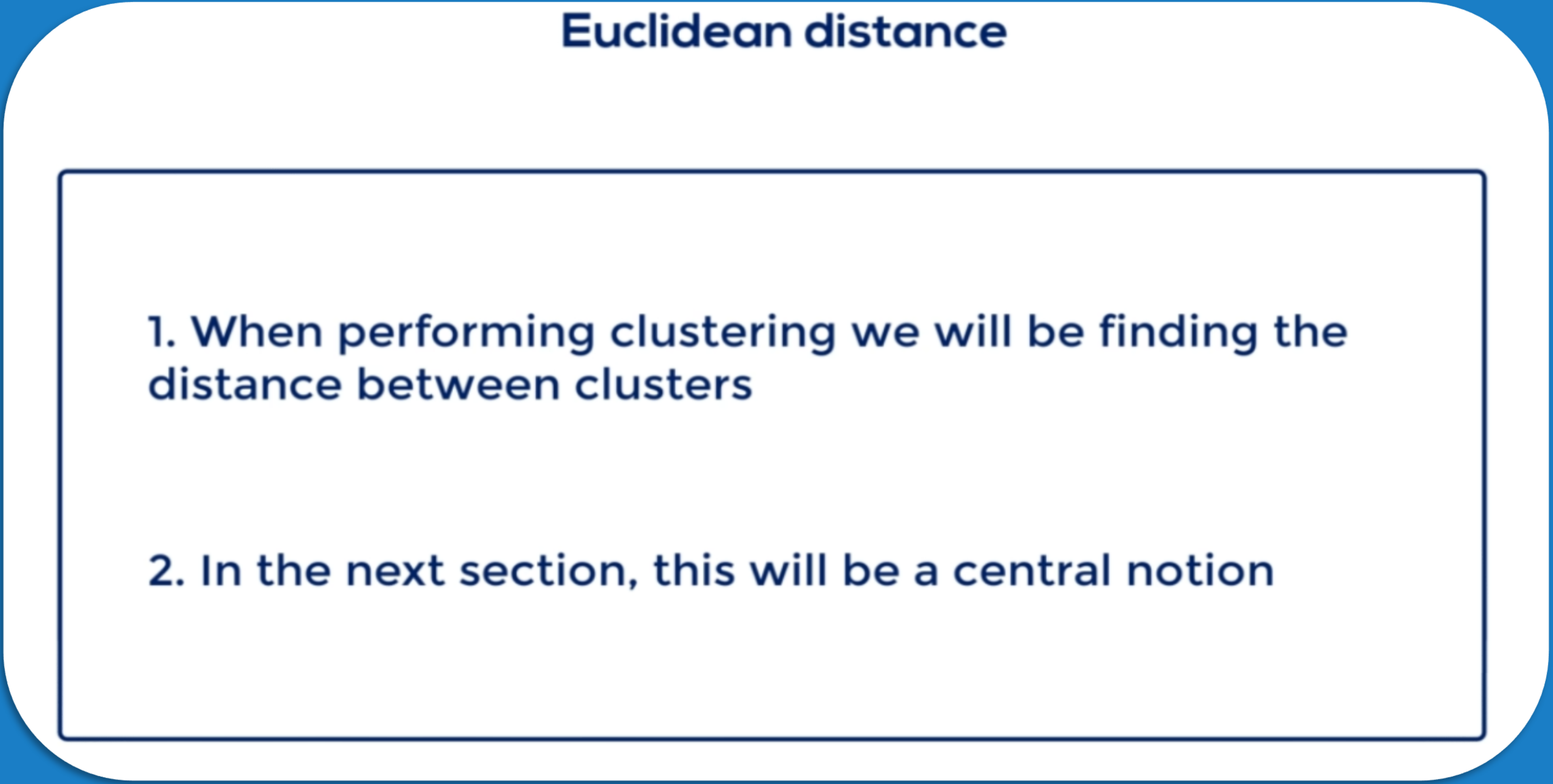
Euclidean distance

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,…,a_n) and of B are (b_1,b_2,…b_n)

N-dim space: $$d(A,B) = d(B,A) = \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2 + \cdots + (a_n - b_n)^2}$$

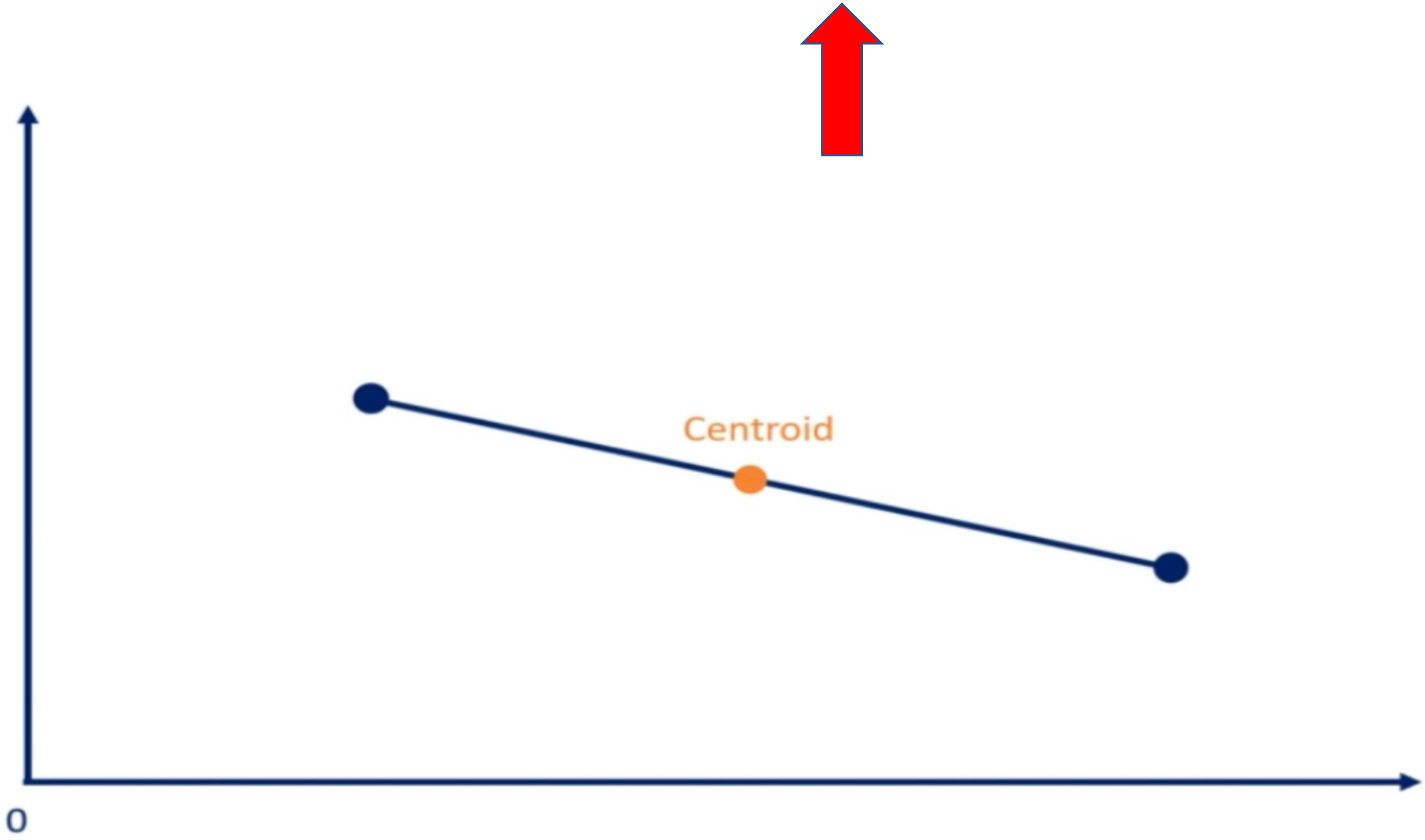


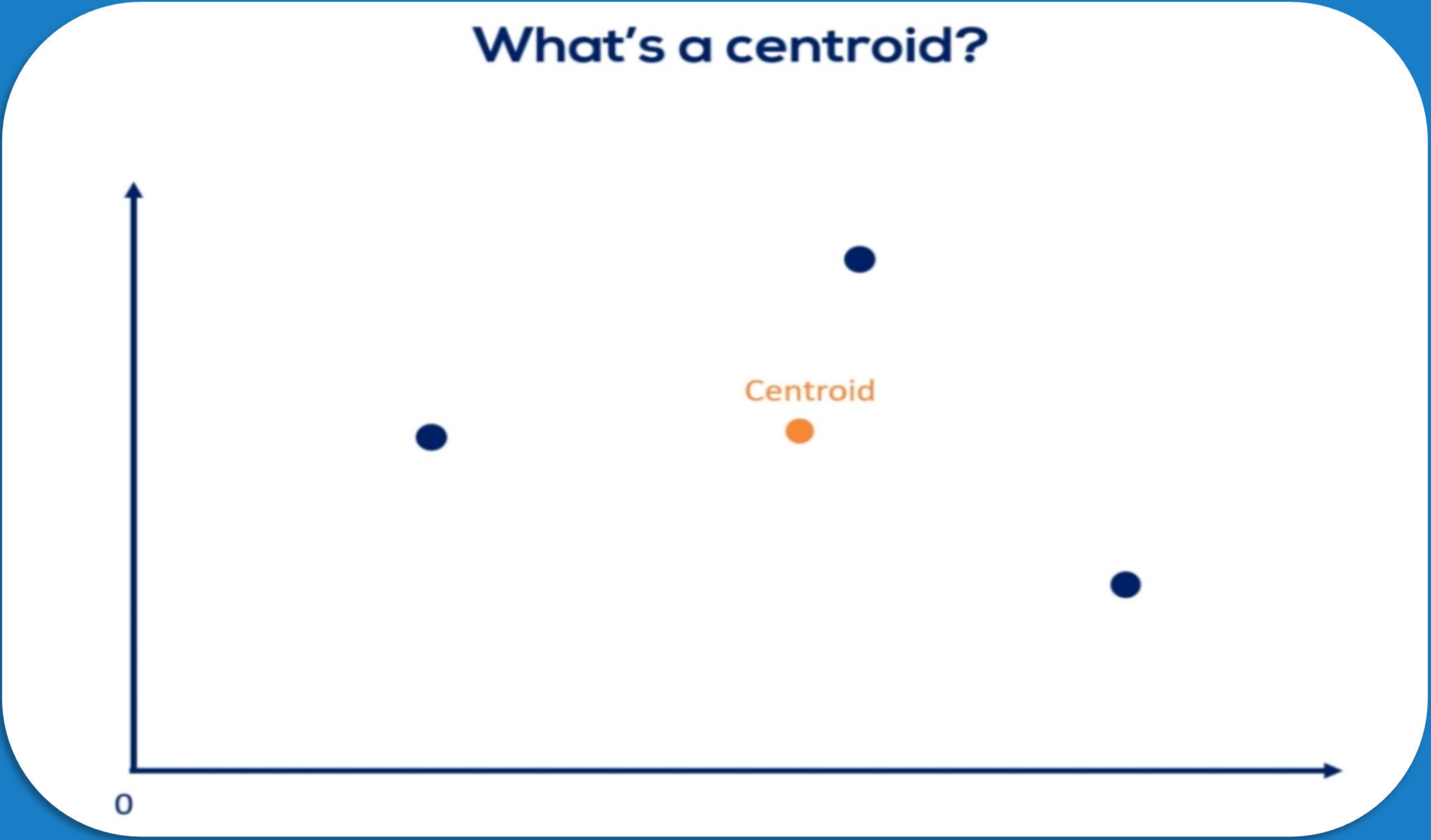
Euclidean distance

1. When performing clustering we will be finding the distance between clusters

2. In the next section, this will be a central notion

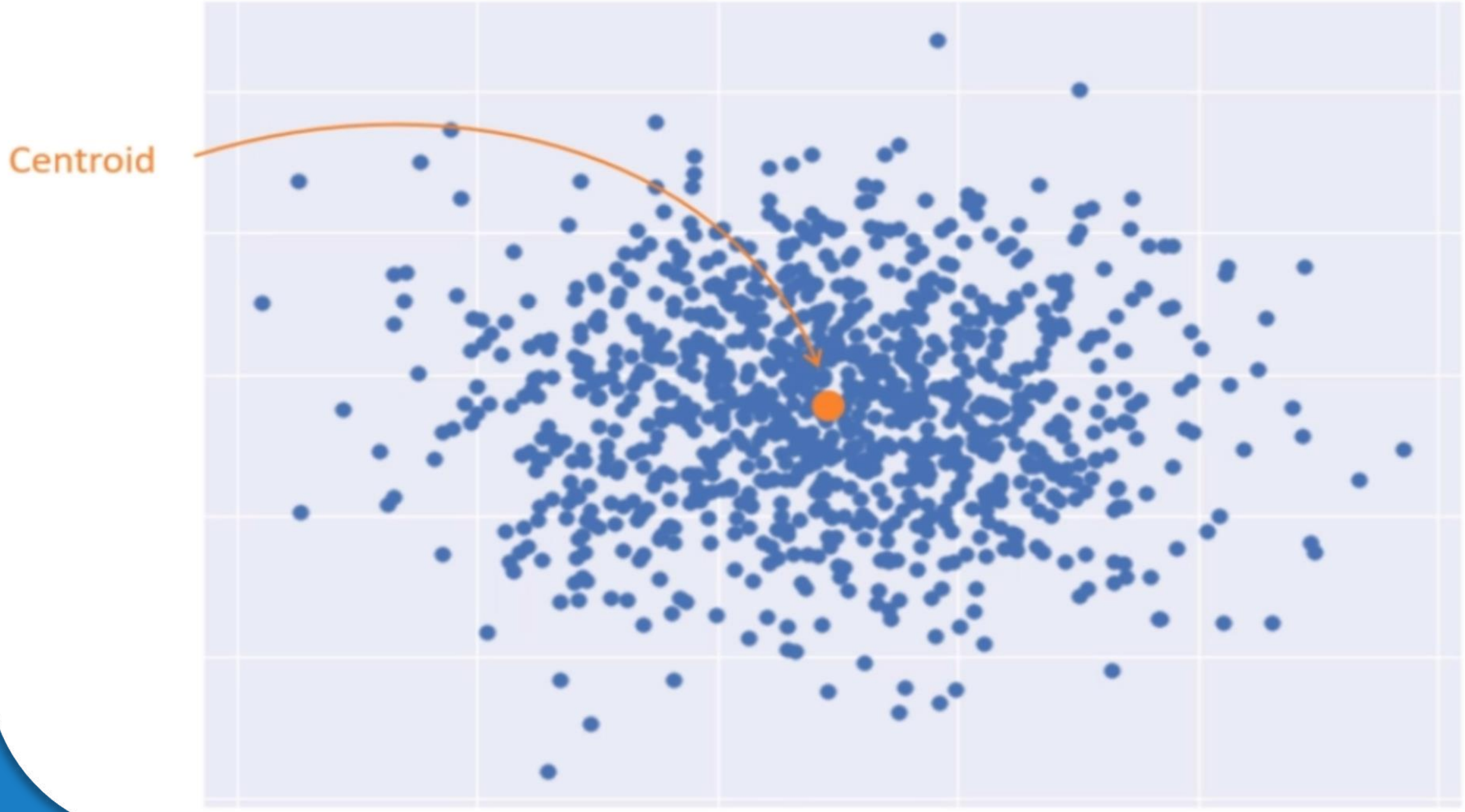
What's a centroid?

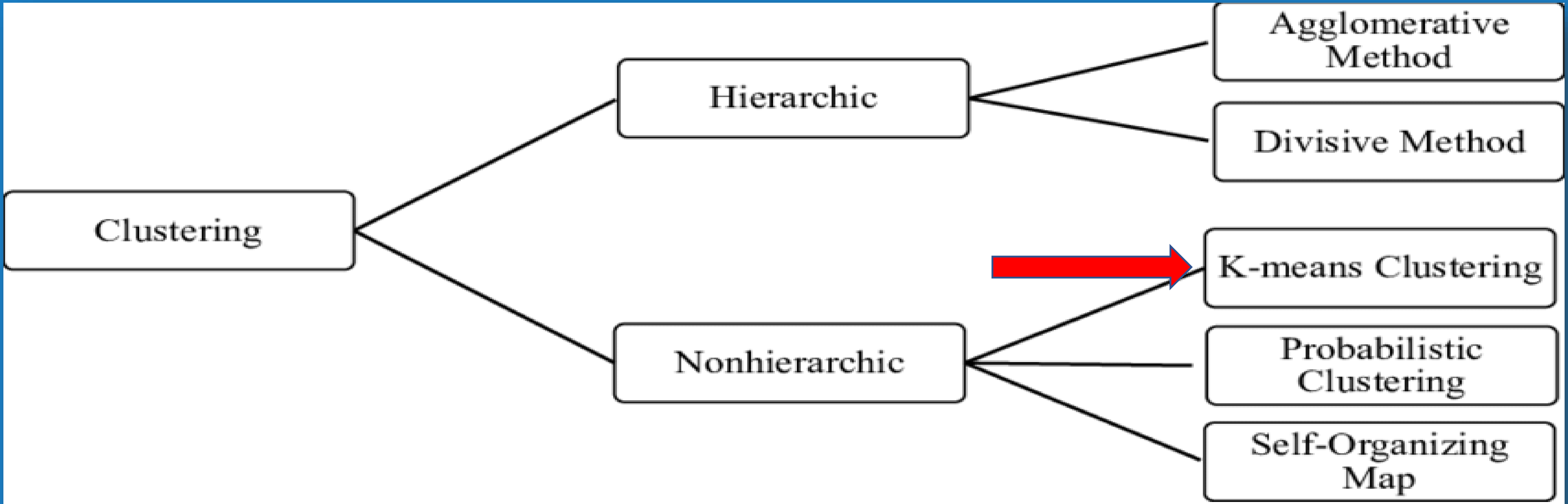




What's a centroid?

What's a centroid?





What is the difference between **hierarchical clustering** and **non-hierarchical clustering**?

- In non-hierarchical clustering, such as the k-means algorithm, the relationship between clusters is undetermined.

- Hierarchical clustering repeatedly links pairs of clusters until every data object is included in the hierarchy.