UNIT 5
K-MEANS CLUSTERING

Content Curated by Pollux M. Rey

FOR THIS UNIT...

01
What is K-Means Clustering?

02
How does it work?

03
How to choose the optimal number of clusters?

Content Curated by Pollux M. Rey

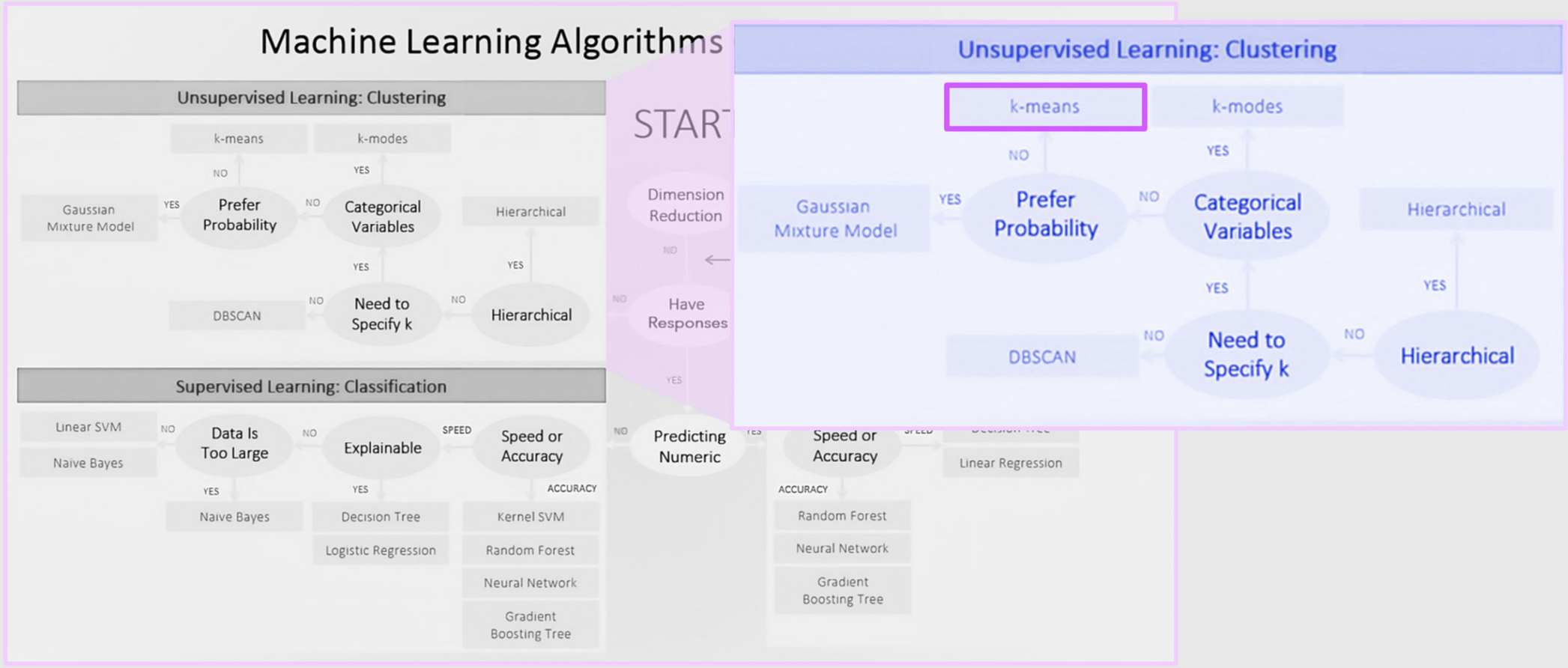
FOR THIS UNIT...

04

Applications

Content Curated by Pollux M. Rey

K-Means Clustering



Content Curated by Pollux M. Rey

https://blogs.sas.com/content/subconsciousmusings/2020/12/09/machine-learning-algorithm-use

K-Means Clustering

An unsupervised learning algorithm used for data clustering, which groups unlabeled data points into groups or clusters.

Content Curated by Pollux M. Rey

https://www.ibm.com/topics/k-means-clustering

K-Means Clustering

Petal Width	Petal Length	Iris Plant
5.8	2.2	Iris-virginica
1.5	0.2	Iris-setosa
4.9	1.5	Iris-versicolor
1.4	0.3	Iris-setosa
4.3	1.3	Iris-versicolor
5.6	1.8	Iris-virginica
5.4	2.3	Iris-virginica
4.2	1.2	Iris-versicolor
4.1	1.3	Iris-versicolor
5.8	2.2	Iris-virginica

Petal Width	Petal Length	Iris Plant
4.0	1.3	Iris-versicolor
1.4	0.2	Iris-setosa
1.4	0.2	Iris-setosa
1.3	0.2	Iris-setosa
1.4	0.2	Iris-setosa
5.0	1.9	Iris-virginica
1.5	0.2	Iris-setosa
1.4	0.2	Iris-setosa
4.5	1.5	Iris-versicolor
6.0	2.5	Iris-virginica

Petal Width	Petal Length	Iris Plant
5.9	2.1	Iris-virginica
5.2	2.0	Iris-virginica
1.4	0.2	Iris-setosa
1.6	0.2	Iris-setosa
3.0	1.1	Iris-versicolor
4.2	1.3	Iris-versicolor
5.2	2.3	Iris-virginica
5.1	1.8	Iris-virginica
5.1	1.9	Iris-virginica
4.7	1.4	Iris-versicolor

Content Curated by Pollux M. Rey

https://archive.ics.uci.edu/dataset/53/iris

K-Means Clustering

Petal Width	Petal Length	Iris Plant
5.8	2.2	Iris-virginica
1.5	0.2	Iris-setosa
4.9	1.5	Iris-versicolor
1.4	0.3	Iris-setosa
4.3	1.3	Iris-versicolor
5.6	1.8	Iris-virginica
5.4	2.3	Iris-virginica
4.2	1.2	Iris-versicolor
4.1	1.3	Iris-versicolor
5.8	2.2	Iris-virginica

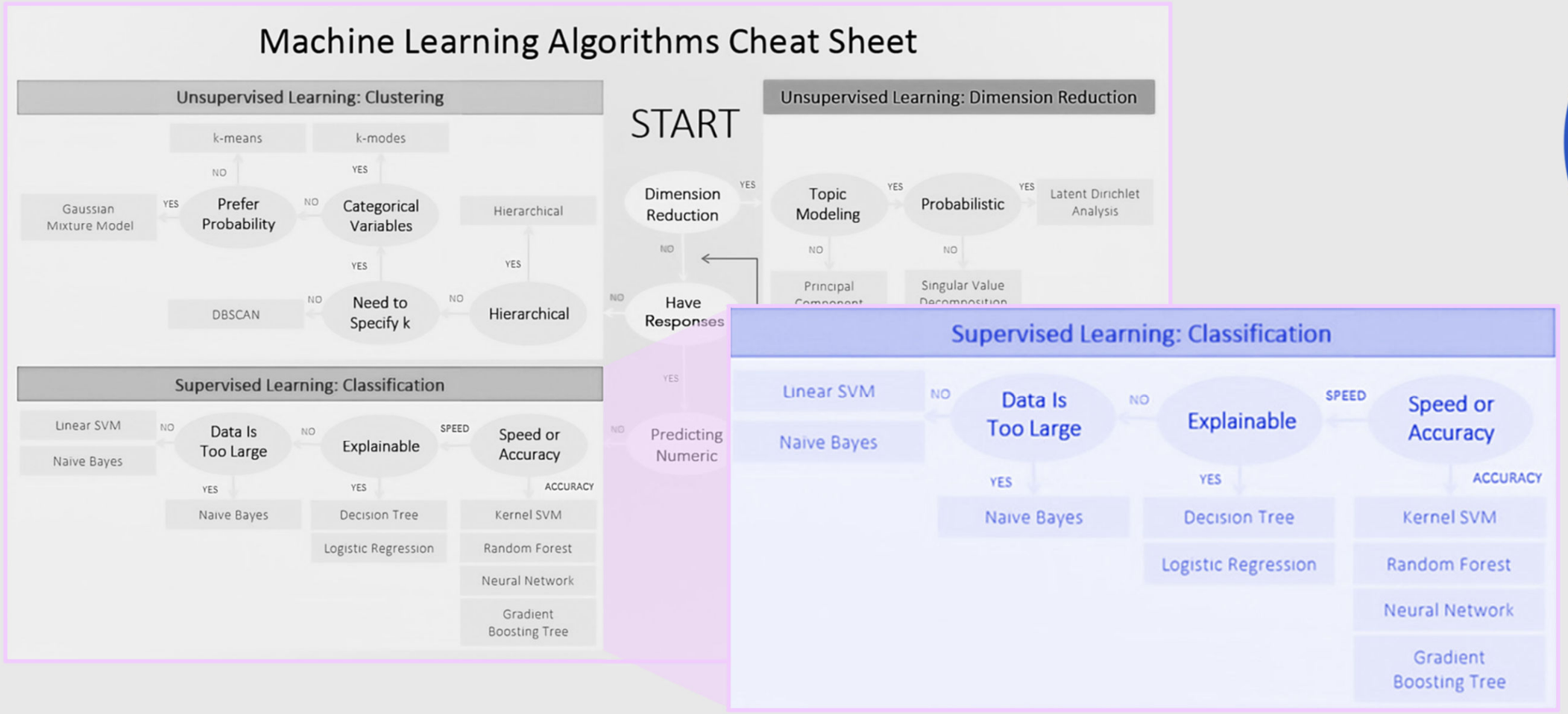
Petal Width	Petal Length	Iris Plant
4.0	1.3	Iris-versicolor
1.4	0.2	Iris-setosa
1.4	0.2	Iris-setosa
1.3	0.2	Iris-setosa
1.4	0.2	Iris-setosa
5.0	1.9	Iris-virginica
1.5	0.2	Iris-setosa
1.4	0.2	Iris-setosa
4.5	1.5	Iris-versicolor
6.0	2.5	Iris-virginica

Petal Width	Petal Length	Iris Plant
5.9	2.1	Iris-virginica
5.2	2.0	Iris-virginica
1.4	0.2	Iris-setosa
1.6	0.2	Iris-setosa
3.0	1.1	Iris-versicolor
4.2	1.3	Iris-versicolor
5.2	2.3	Iris-virginica
5.1	1.8	Iris-virginica
5.1	1.9	Iris-virginica
4.7	1.4	Iris-versicolor

Content Curated by Pollux M. Rey

https://archive.ics.uci.edu/dataset/53/iris

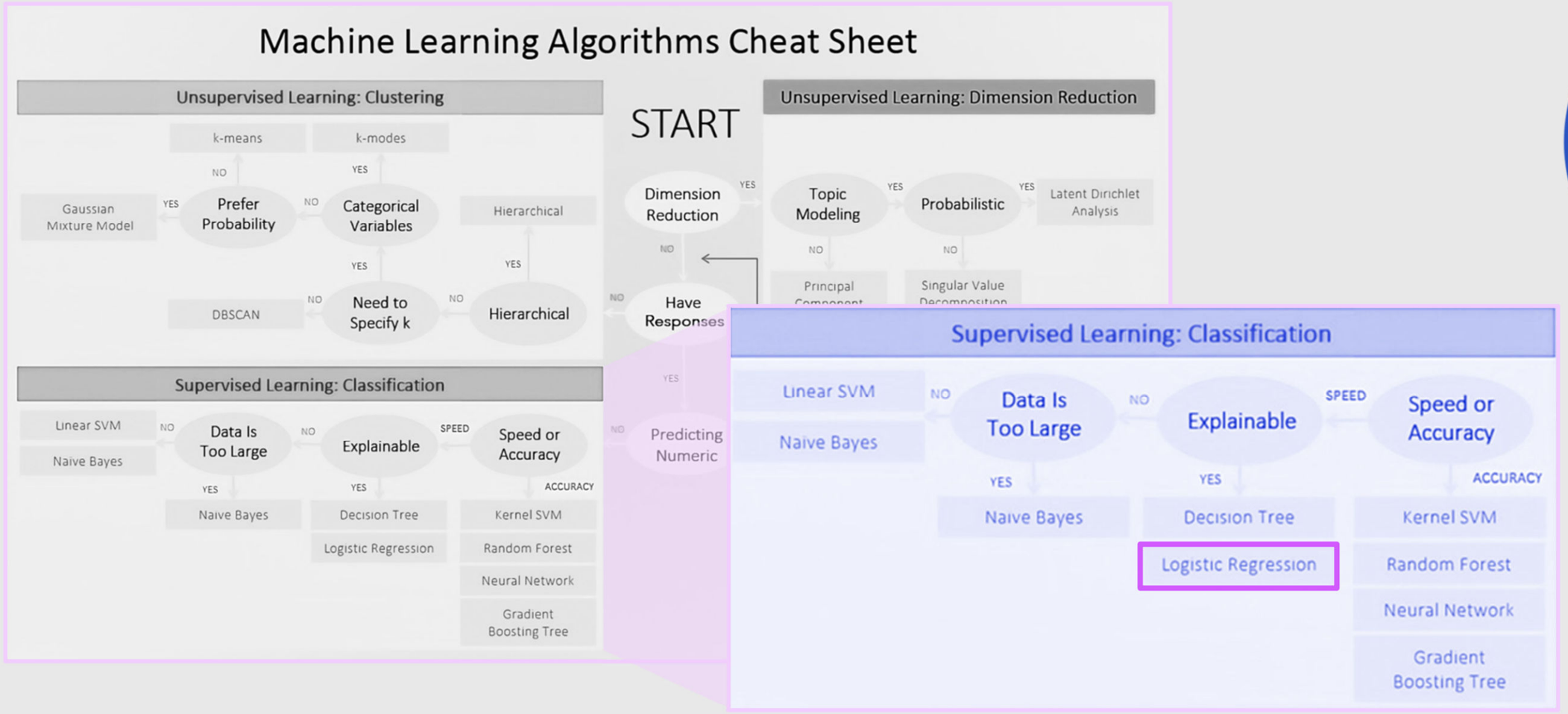
Classification Algorithms



Content Curated by Pollux M. Rey

https://blogs.sas.com/content/subconsciousmusings/2020/12/09/machine-learning-algorithm-use

Classification Algorithms



Content Curated by Pollux M. Rey

https://blogs.sas.com/content/subconsciousmusings/2020/12/09/machine-learning-algorithm-use

K-Means Clustering

Petal Width	Petal Length	Iris Plant
5.8	2.2	Iris-virginica
1.5	0.2	Iris-setosa
4.9	1.5	Iris-versicolor
1.4	0.3	Iris-setosa
4.3	1.3	Iris-versicolor
5.6	1.8	Iris-virginica
5.4	2.3	Iris-virginica
4.2	1.2	Iris-versicolor
4.1	1.3	Iris-versicolor
5.8	2.2	Iris-virginica

Petal Width	Petal Length	Iris Plant
4.0	1.3	Iris-versicolor
1.4	0.2	Iris-setosa
1.4	0.2	Iris-setosa
1.3	0.2	Iris-setosa
1.4	0.2	Iris-setosa
5.0	1.9	Iris-virginica
1.5	0.2	Iris-setosa
1.4	0.2	Iris-setosa
4.5	1.5	Iris-versicolor
6.0	2.5	Iris-virginica

Petal Width	Petal Length	Iris Plant
5.9	2.1	Iris-virginica
5.2	2.0	Iris-virginica
1.4	0.2	Iris-setosa
1.6	0.2	Iris-setosa
3.0	1.1	Iris-versicolor
4.2	1.3	Iris-versicolor
5.2	2.3	Iris-virginica
5.1	1.8	Iris-virginica
5.1	1.9	Iris-virginica
4.7	1.4	Iris-versicolor

Content Curated by Pollux M. Rey

https://archive.ics.uci.edu/dataset/53/iris

K-Means Clustering

Petal Width	Petal Length
5.8	2.2
1.5	0.2
4.9	1.5
1.4	0.3
4.3	1.3
5.6	1.8
5.4	2.3
4.2	1.2
4.1	1.3
5.8	2.2

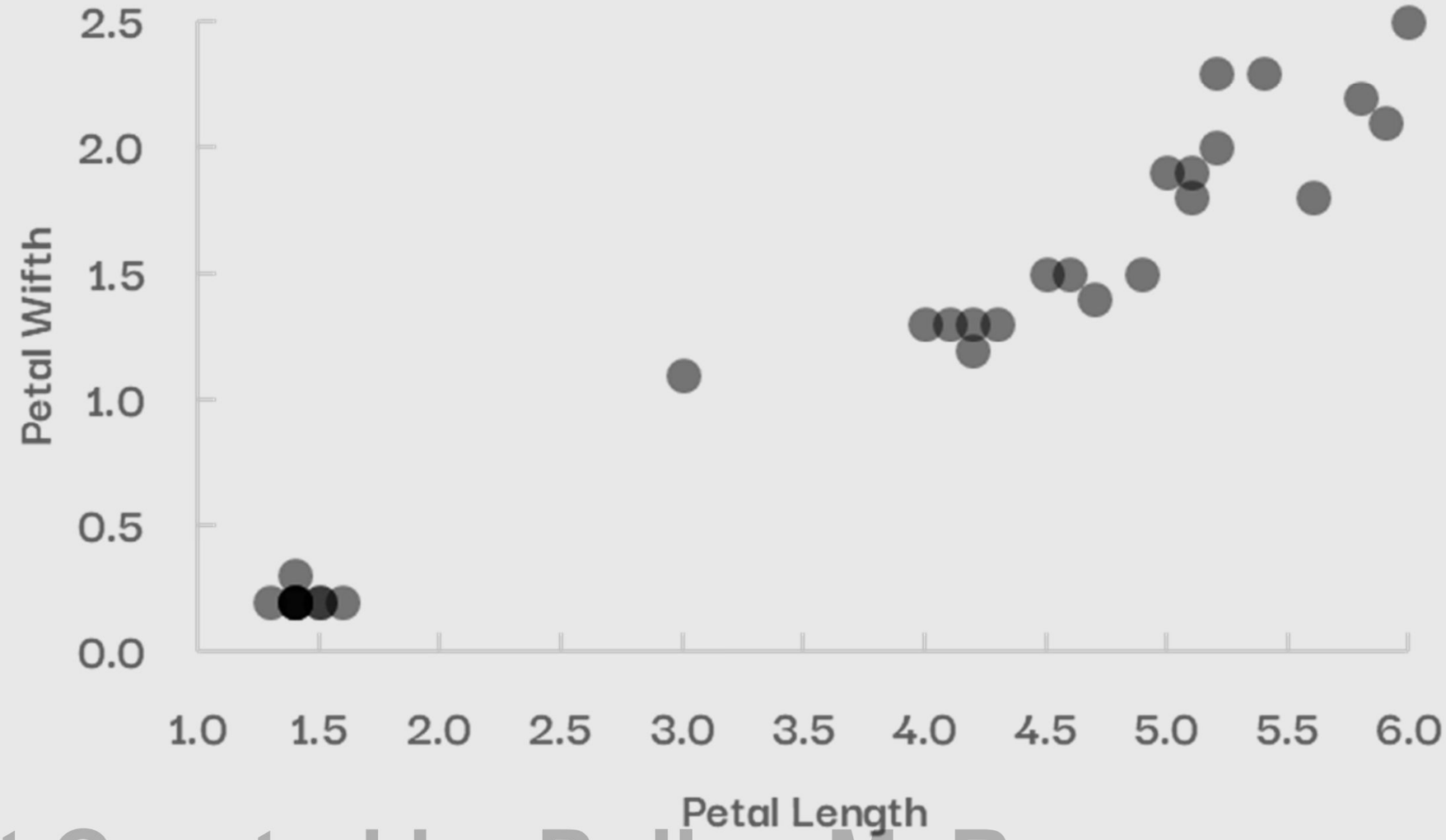
Petal Width	Petal Length
4.0	1.3
1.4	0.2
1.4	0.2
1.3	0.2
1.4	0.2
5.0	1.9
1.5	0.2
1.4	0.2
4.5	1.5
6.0	2.5

Petal Width	Petal Length
5.9	2.1
5.2	2.0
1.4	0.2
1.6	0.2
3.0	1.1
4.2	1.3
5.2	2.3
5.1	1.8
5.1	1.9
4.7	1.4

Content Curated by Pollux M. Rey

https://archive.ics.uci.edu/dataset/53/iris

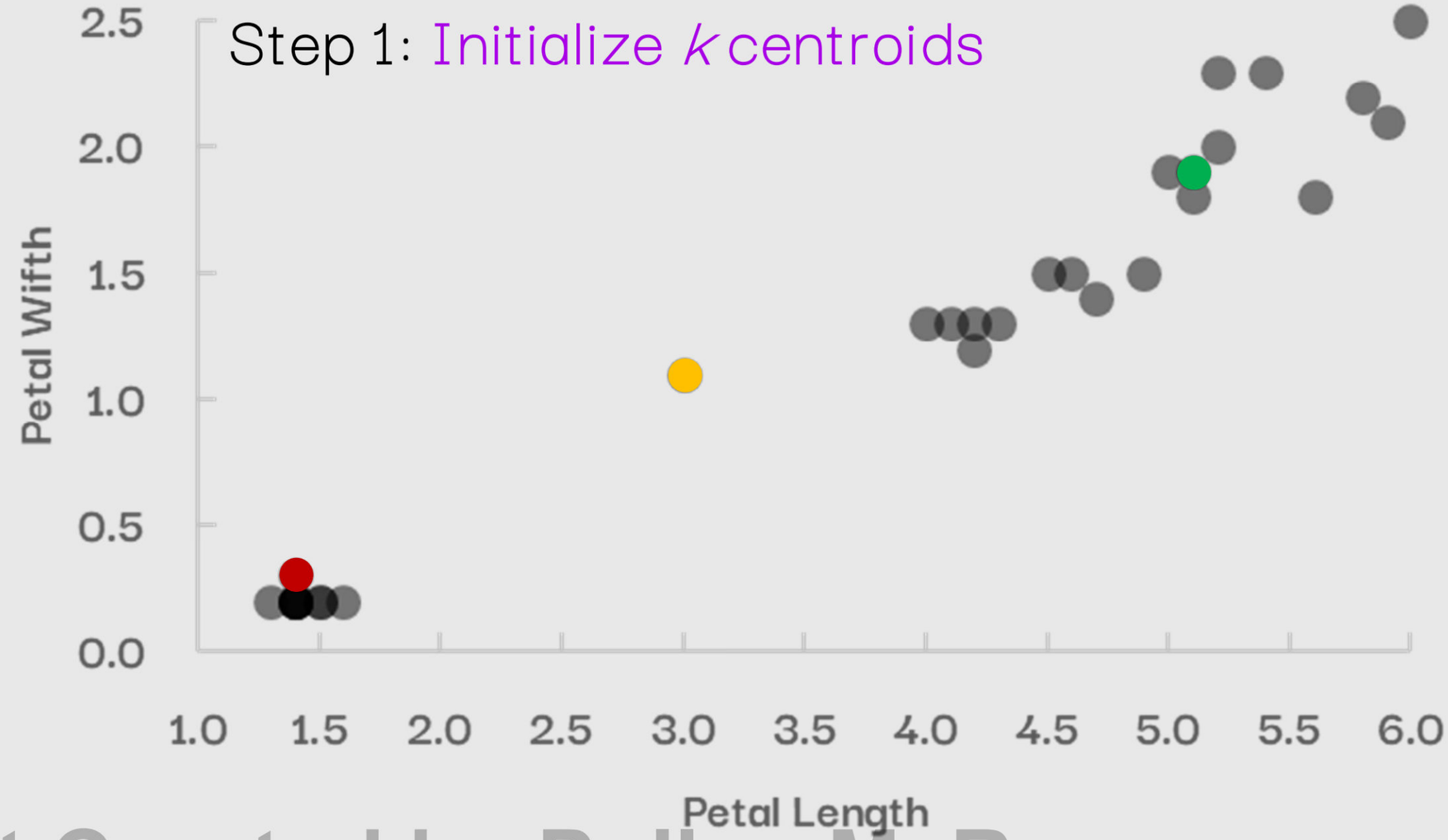
K-Means Clustering



Content Curated by Pollux M. Rey

https://www.ibm.com/topics/k-means-clustering

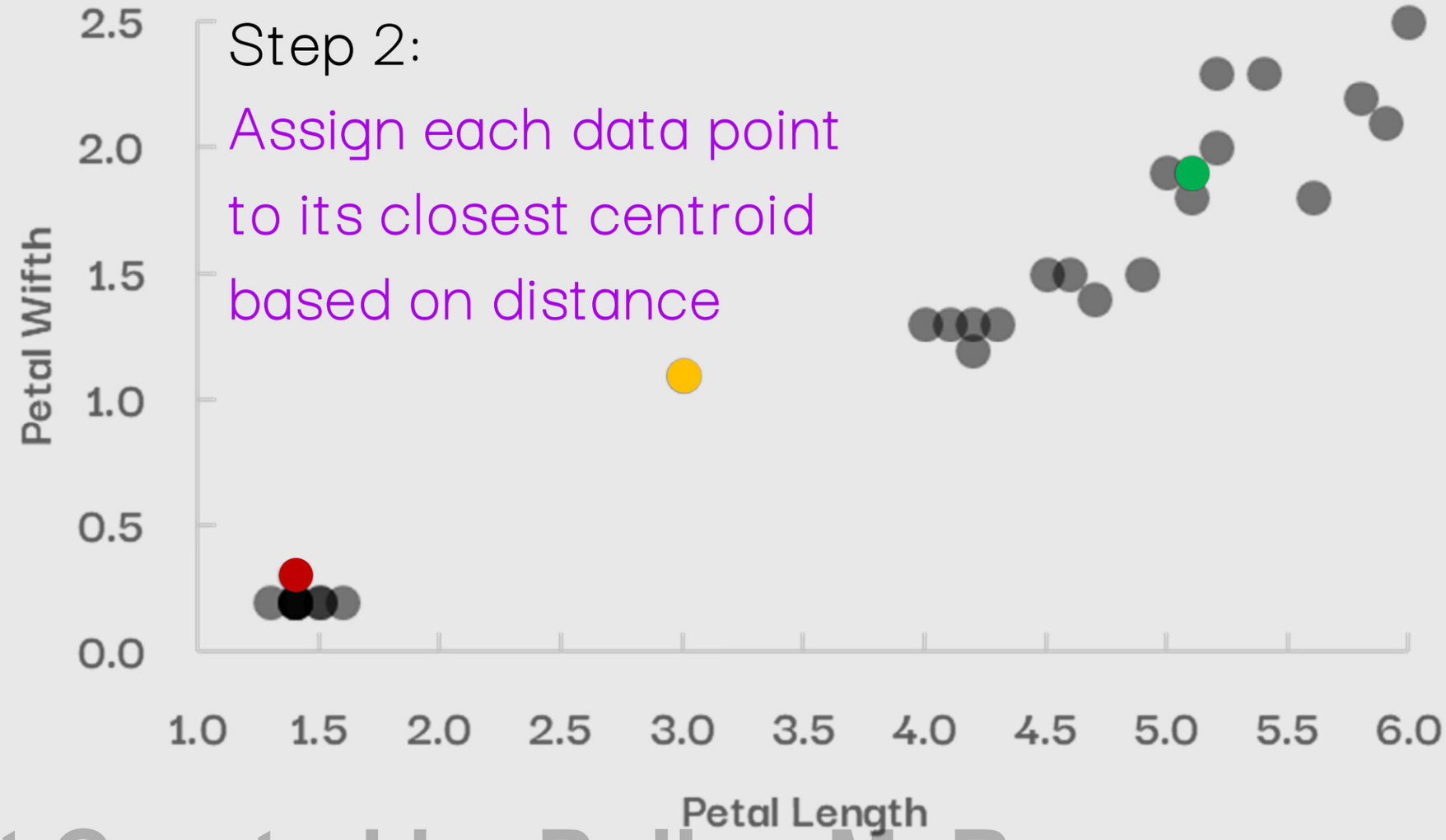
K-Means Clustering



Content Curated by Pollux M. Rey

https://www.ibm.com/topics/k-means-clustering

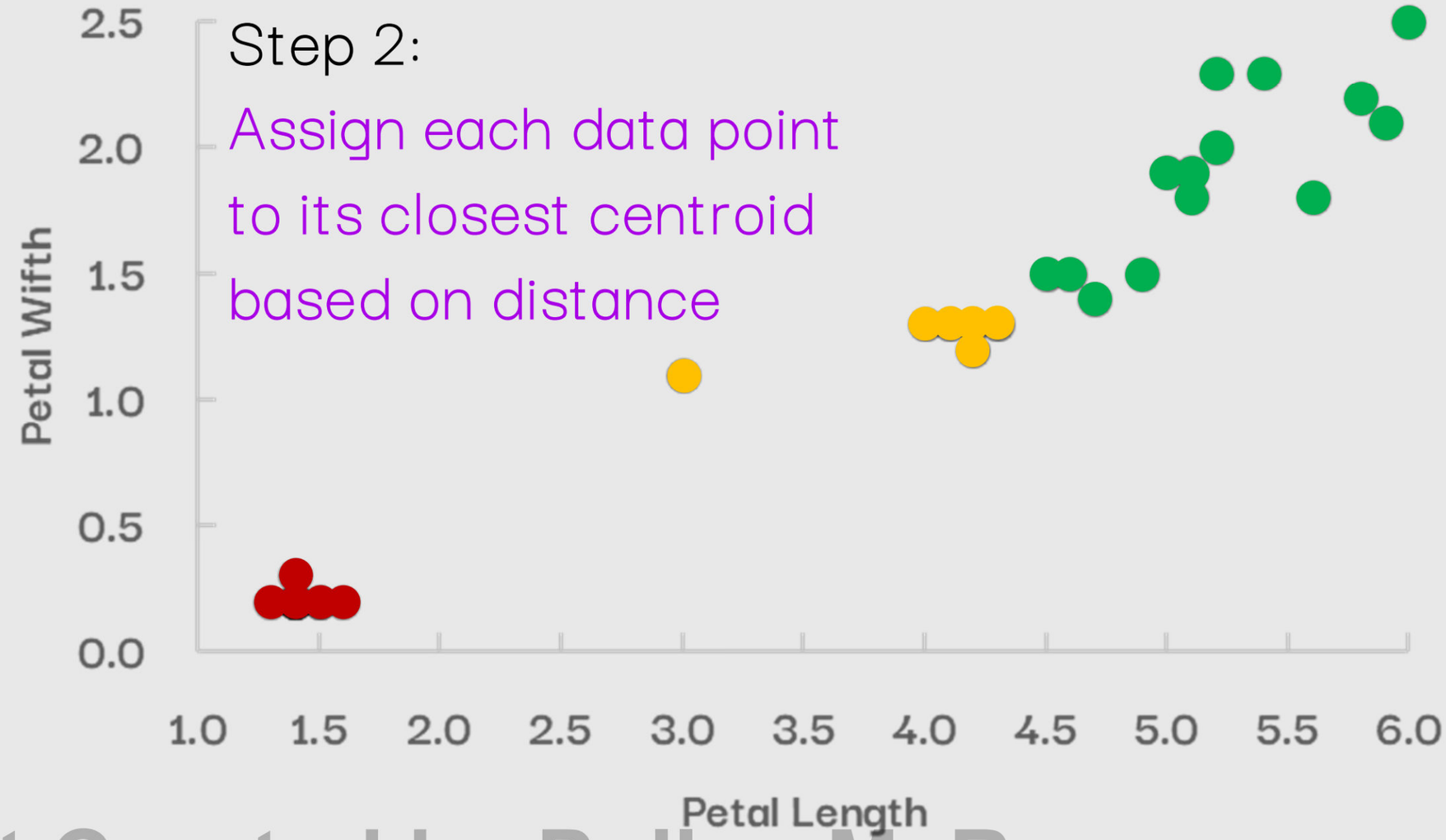
K-Means Clustering



Content Curated by Pollux M. Rey

https://www.ibm.com/topics/k-means-clustering

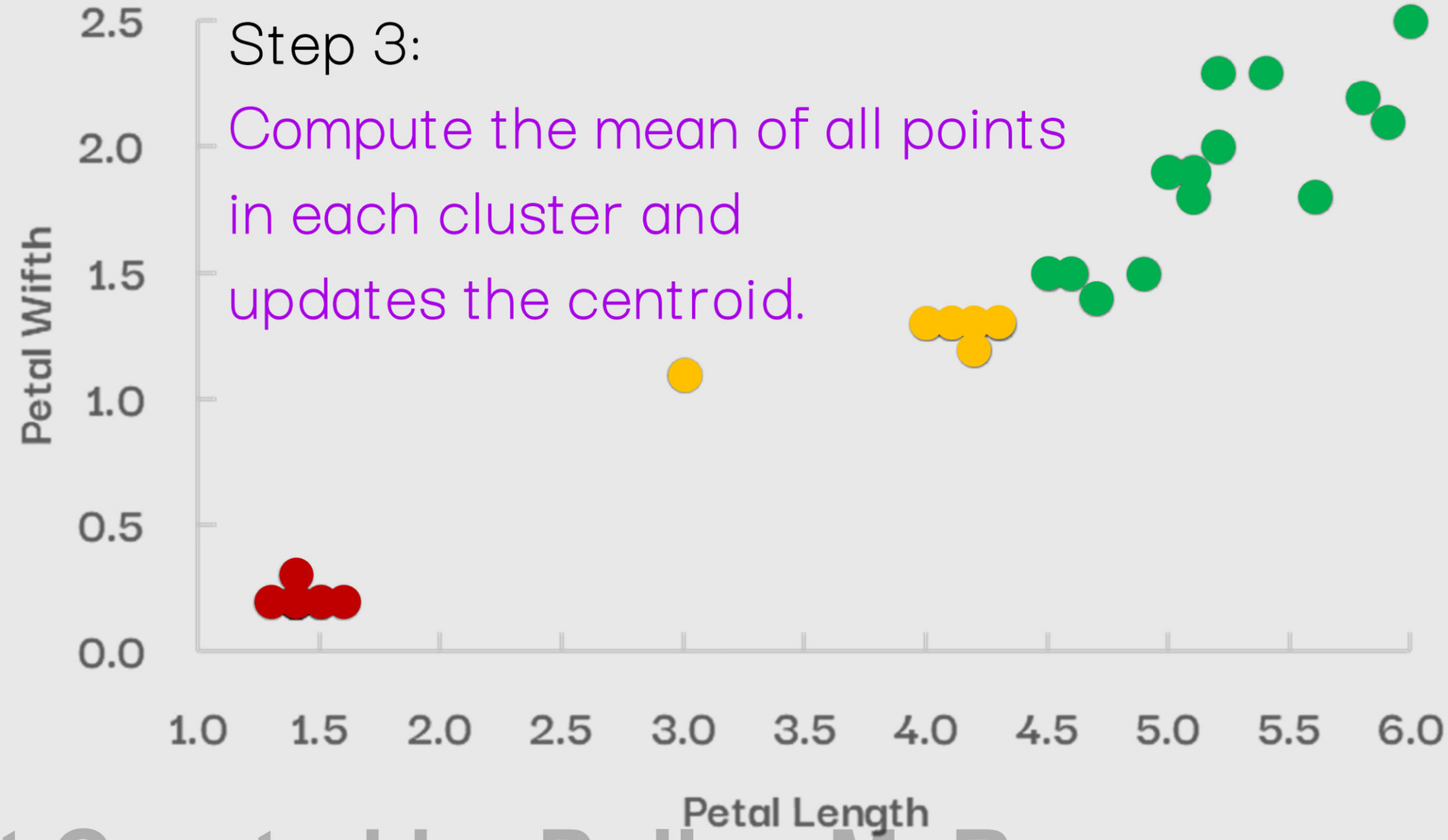
K-Means Clustering



Content Curated by Pollux M. Rey

https://www.ibm.com/topics/k-means-clustering

K-Means Clustering



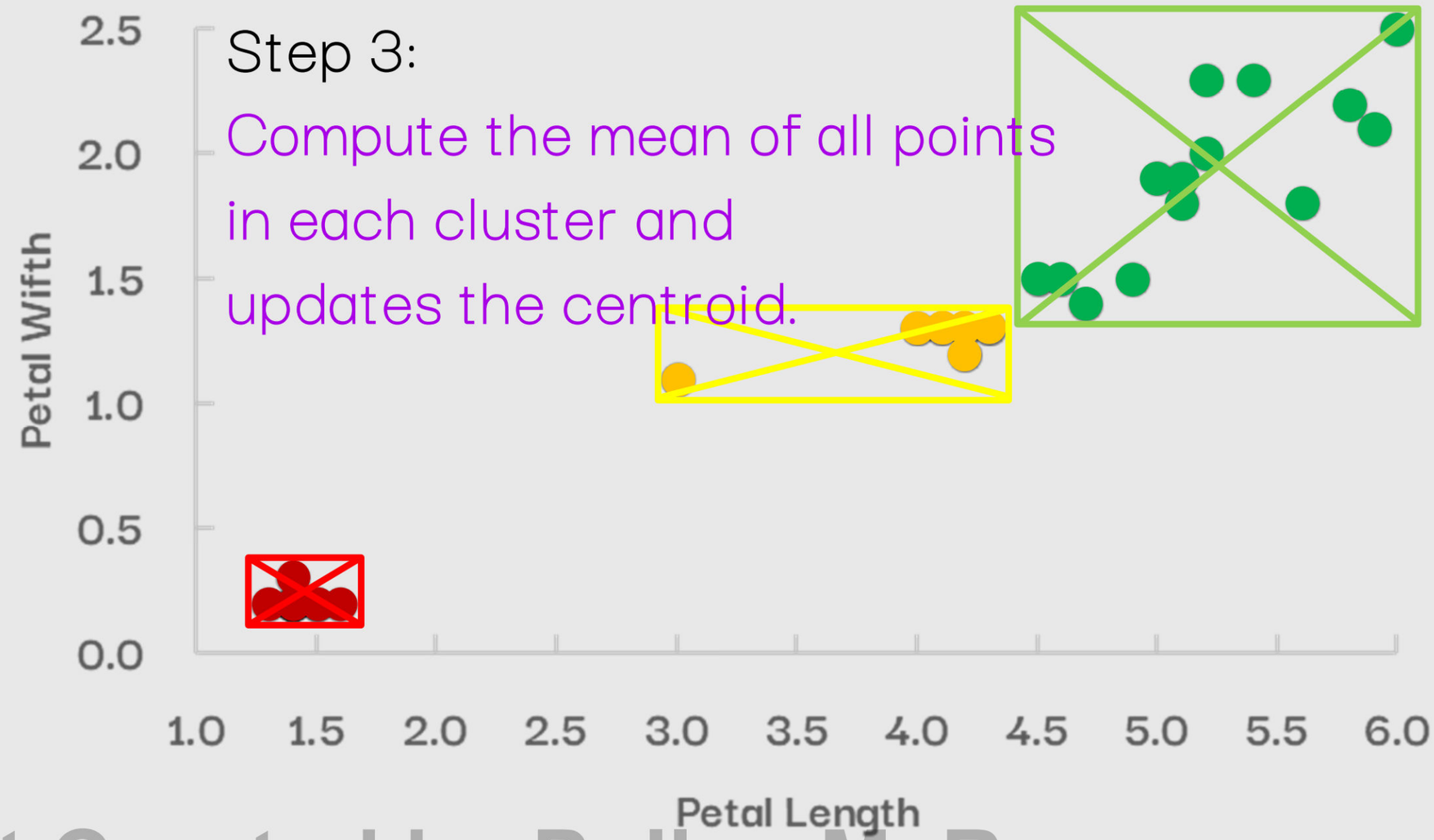
Content Curated by Pollux M. Rey

https://www.ibm.com/topics/k-means-clustering

K-Means Clustering

Step 3:

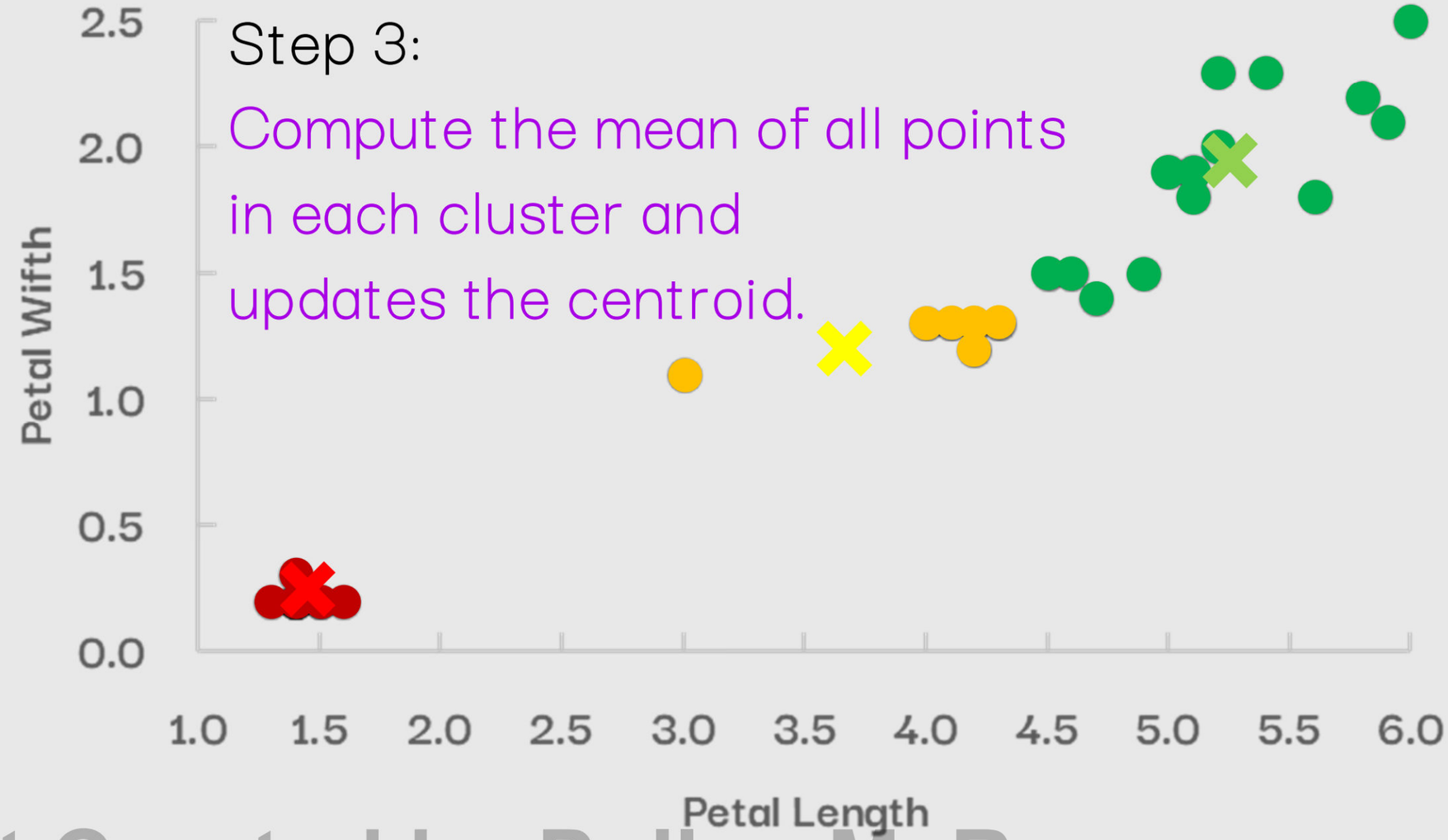
Compute the mean of all points in each cluster and updates the centroid.



Content Curated by Pollux M. Rey

https://www.ibm.com/topics/k-means-clustering

K-Means Clustering

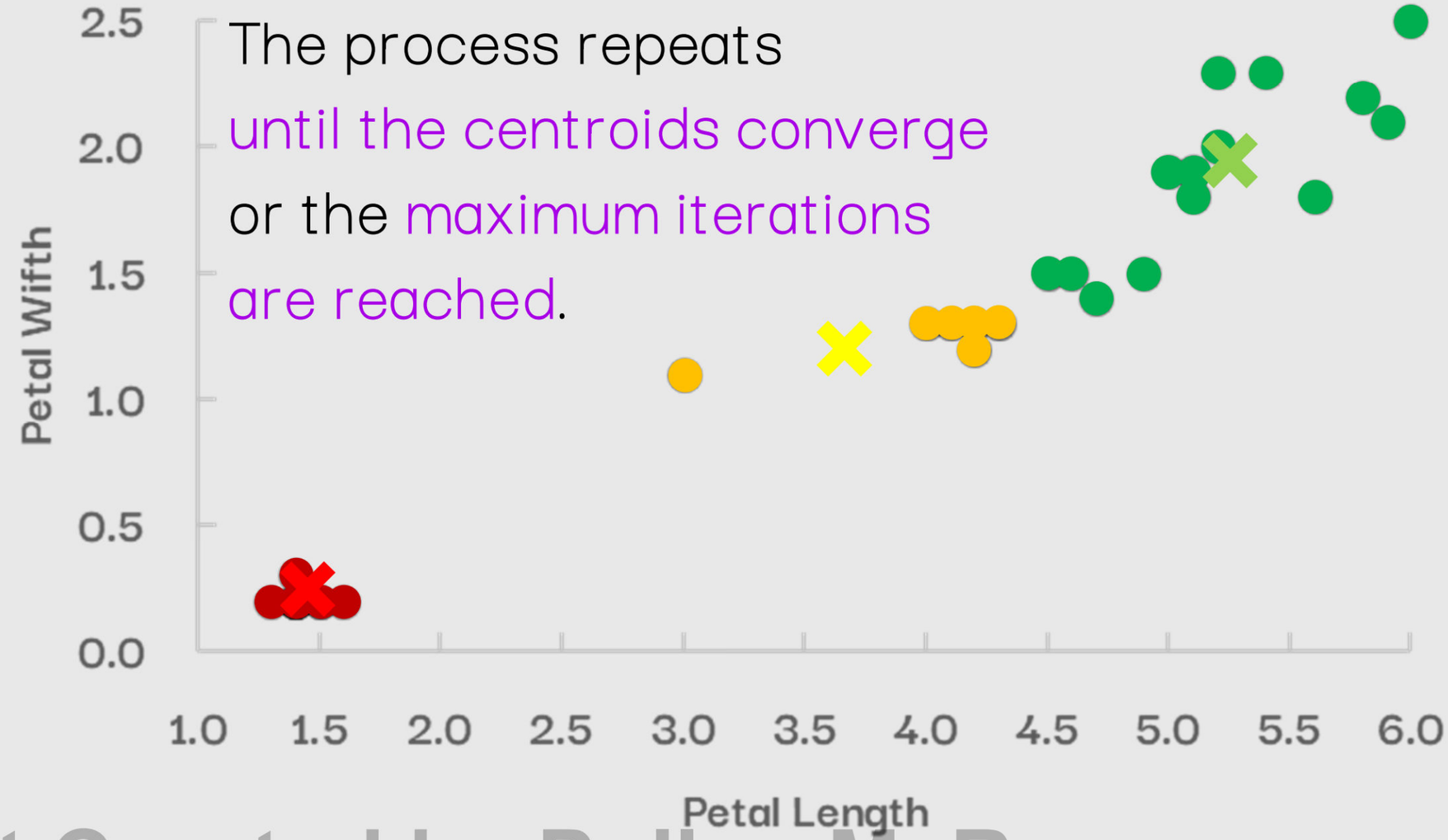


Content Curated by Pollux M. Rey

https://www.ibm.com/topics/k-means-clustering

K-Means Clustering

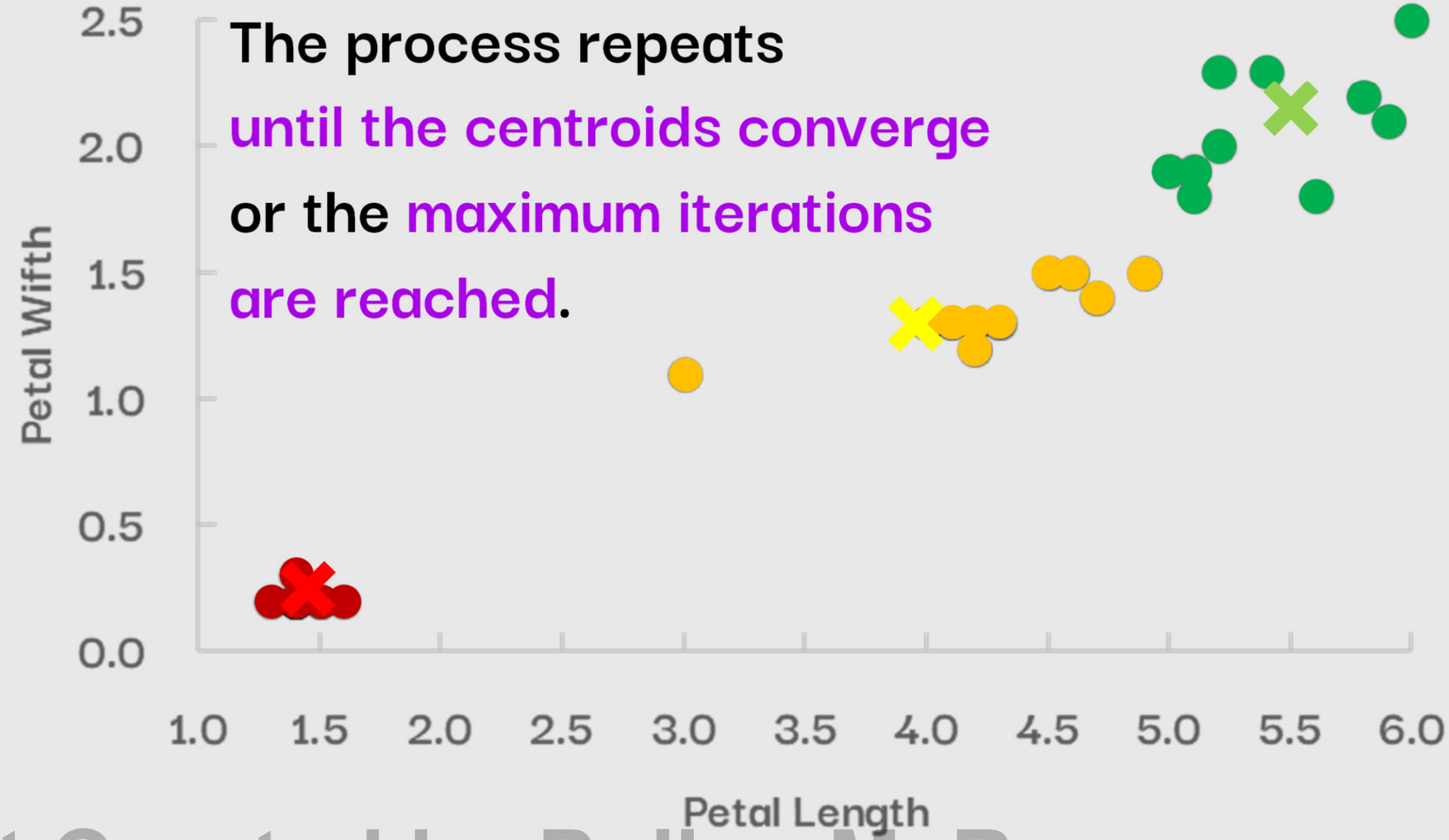
The process repeats
until the centroids converge
or the maximum iterations
are reached.



Content Curated by Pollux M. Rey

https://www.ibm.com/topics/k-means-clustering

K-Means Clustering



Content Curated by Pollux M. Rey

https://www.ibm.com/topics/k-means-clustering

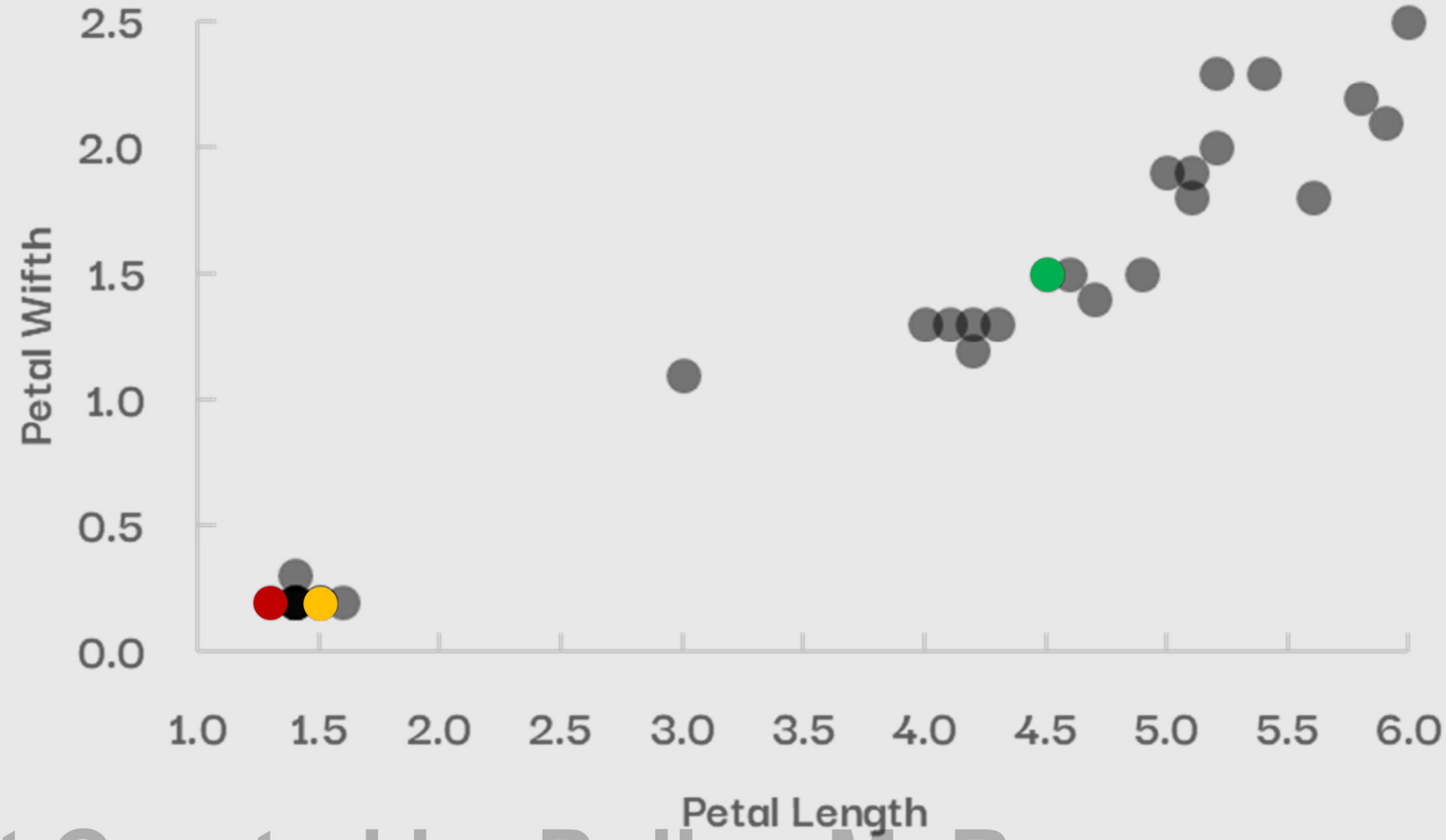
K-Means Clustering

K-means clustering is simple yet sensitive to initial conditions and outliers.

Content Curated by Pollux M. Rey

https://www.ibm.com/topics/k-means-clustering

K-Means Clustering



Content Curated by Pollux M. Rey

https://www.ibm.com/topics/k-means-clustering

K-Means Clustering

Content Curated by Pollux M. Rey

https://www.ibm.com/topics/k-means-clustering

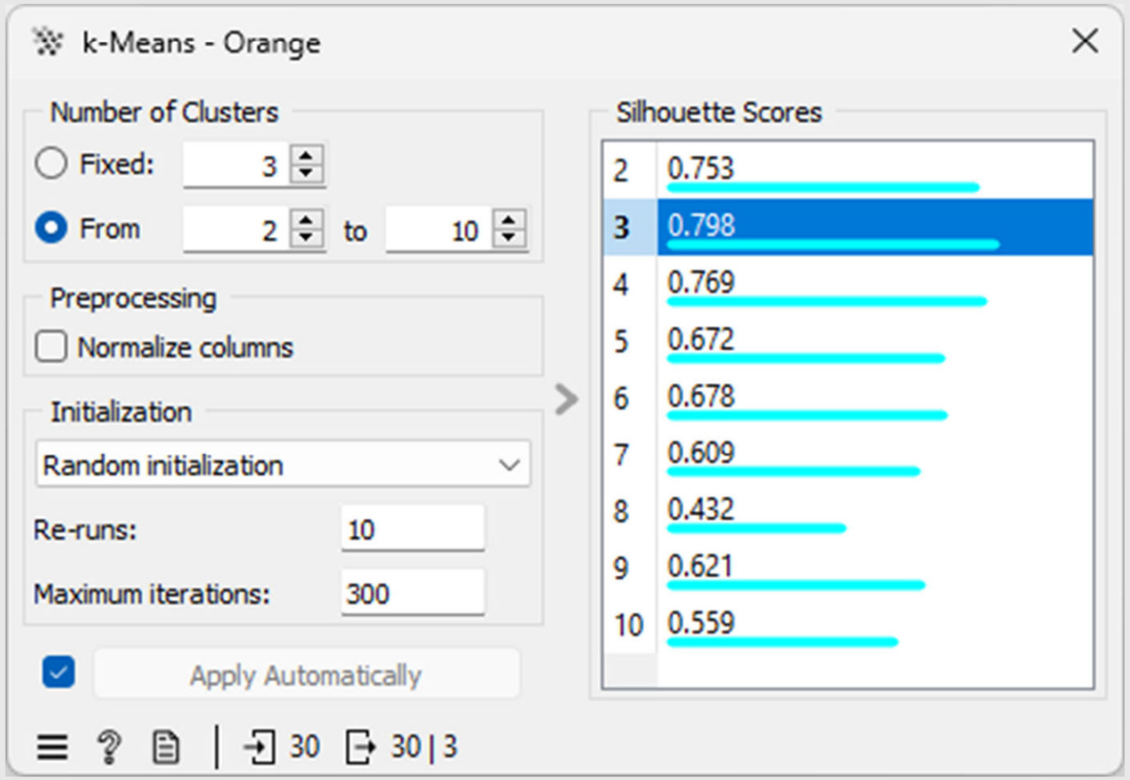
K-Means Clustering

It is important to optimize the centroid initialization and the number of clusters *k*, to achieve the most meaningful clusters.

Content Curated by Pollux M. Rey

https://www.ibm.com/topics/k-means-clustering

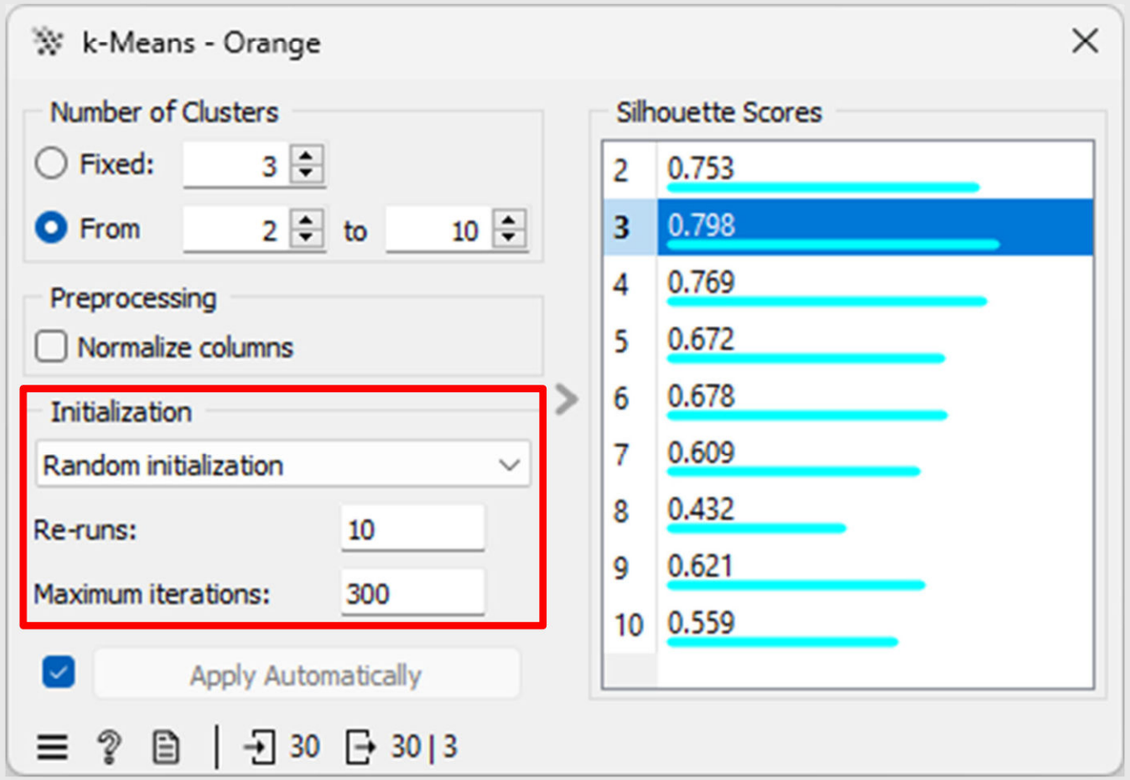
K-Means Clustering



Content Curated by Pollux M. Rey

https://www.ibm.com/topics/k-means-clustering

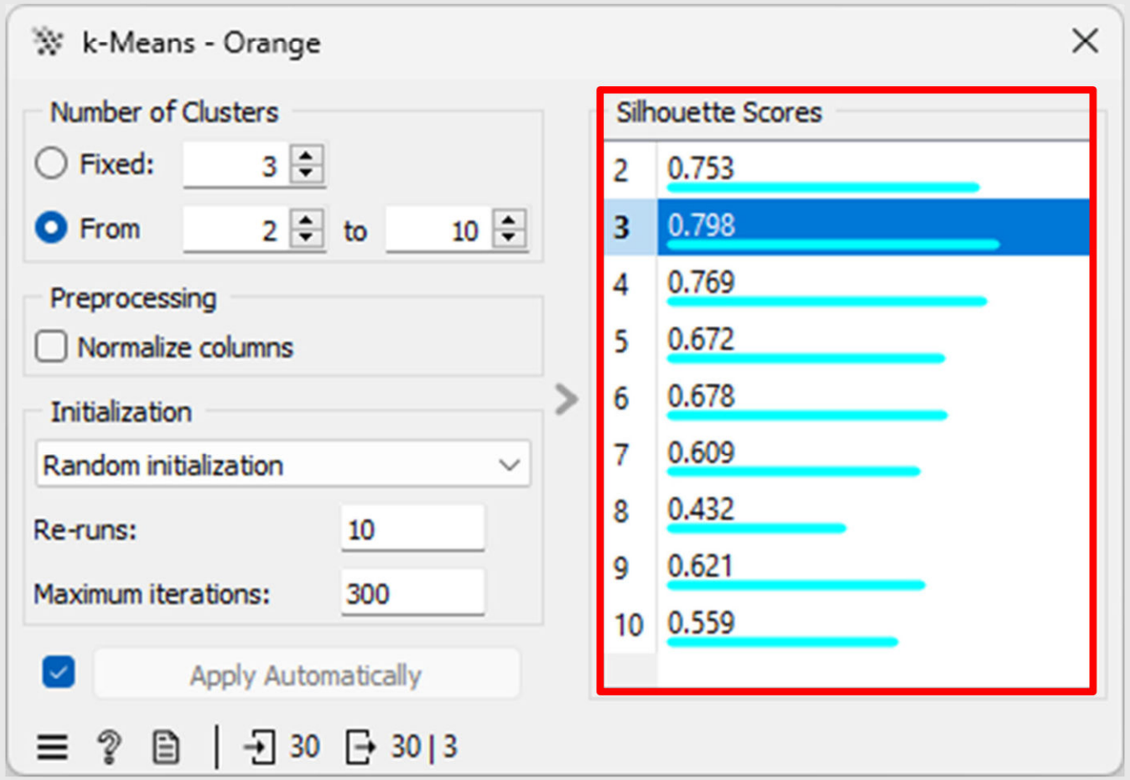
K-Means Clustering



Content Curated by Pollux M. Rey

https://www.ibm.com/topics/k-means-clustering

K-Means Clustering



Content Curated by Pollux M. Rey

https://www.ibm.com/topics/k-means-clustering

K-Means Clustering

StatQuest!

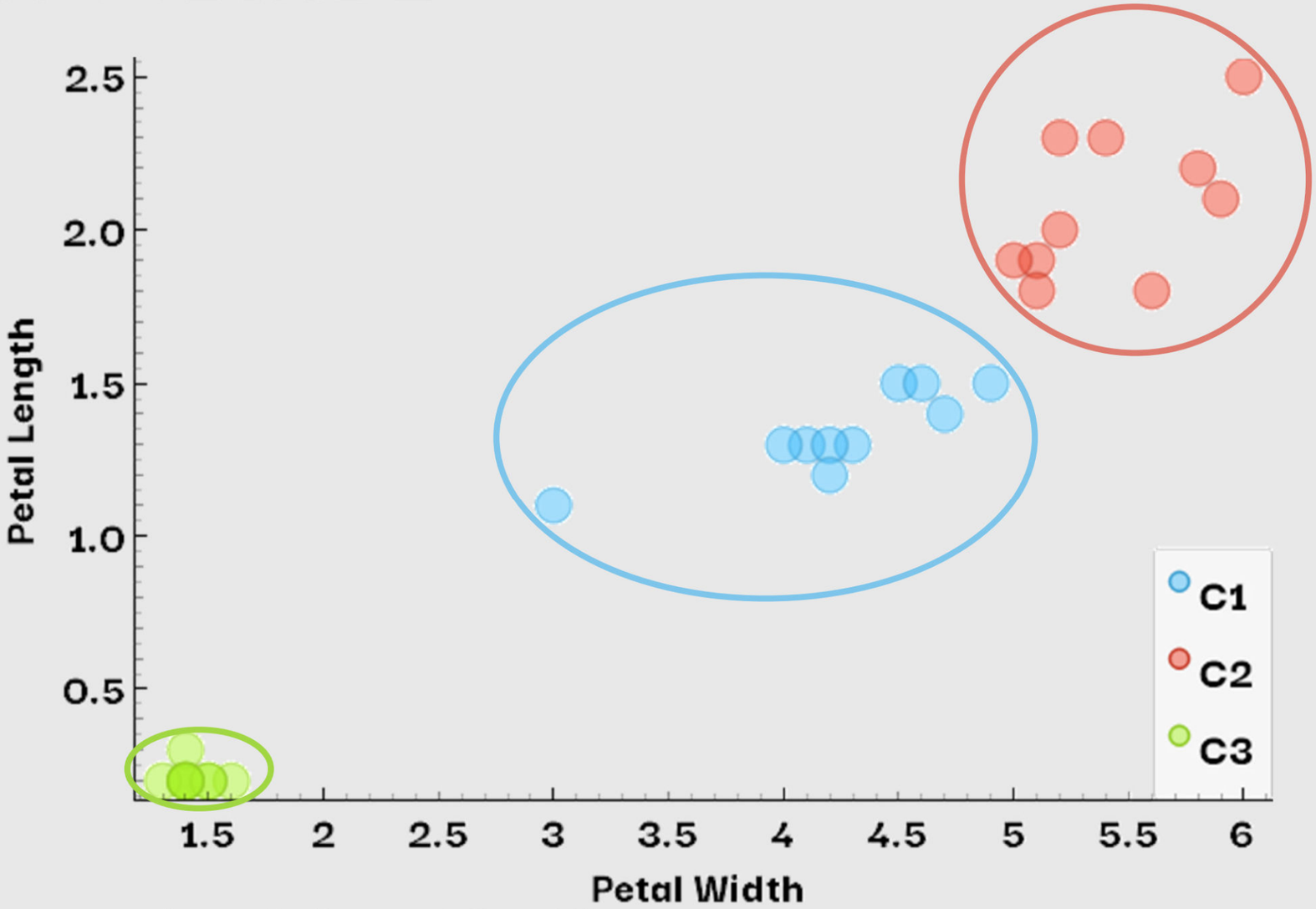
Content Curated by Pollux M. Rey

https://www.youtube.com/watch?v=4b5d3muPQmA

HOW TO CHOOSE THE OPTIMAL NUMBER OF CLUSTERS?

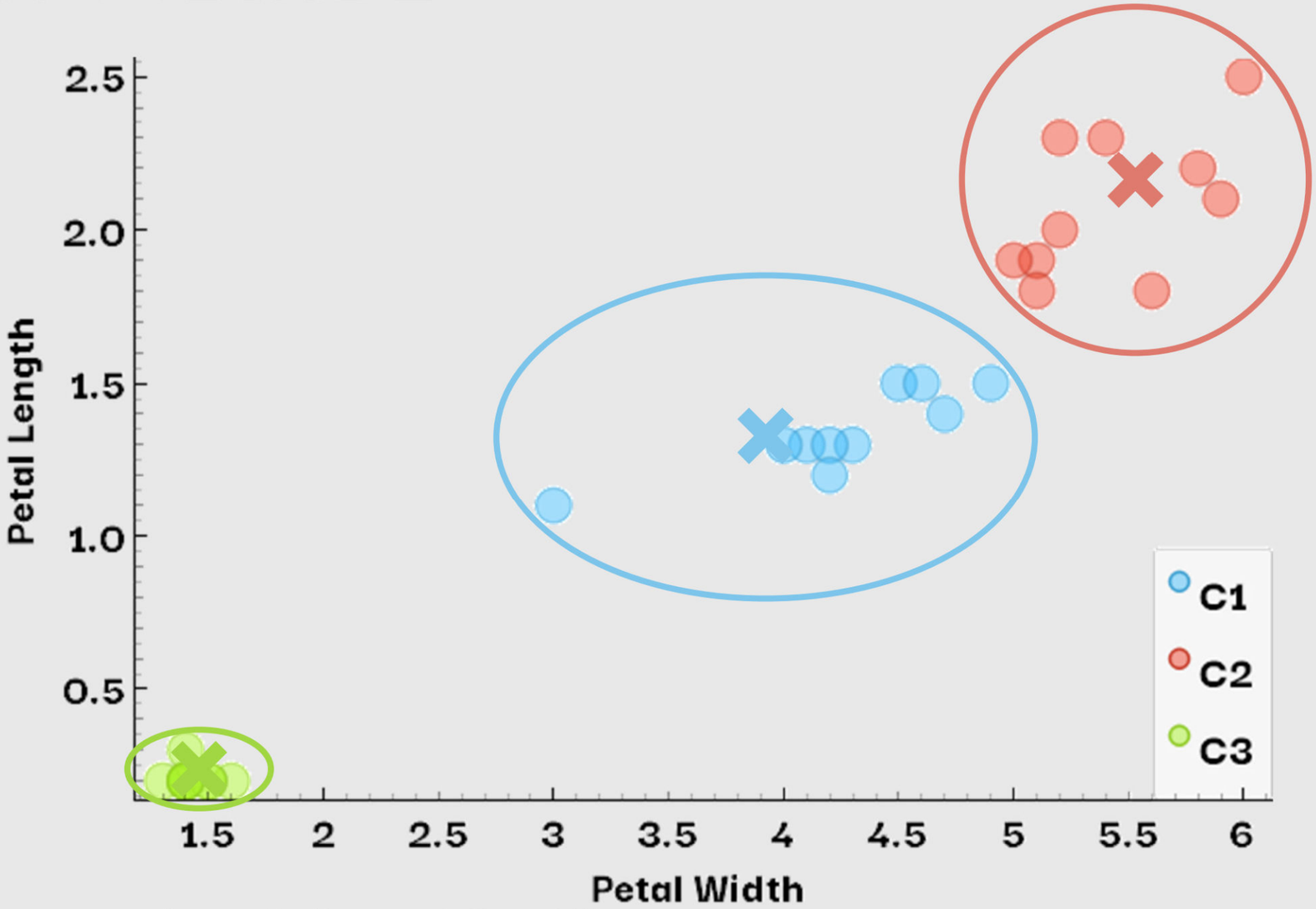
Content Curated by Pollux M. Rey

Elbow Method



Content Curated by Pollux M. Rey

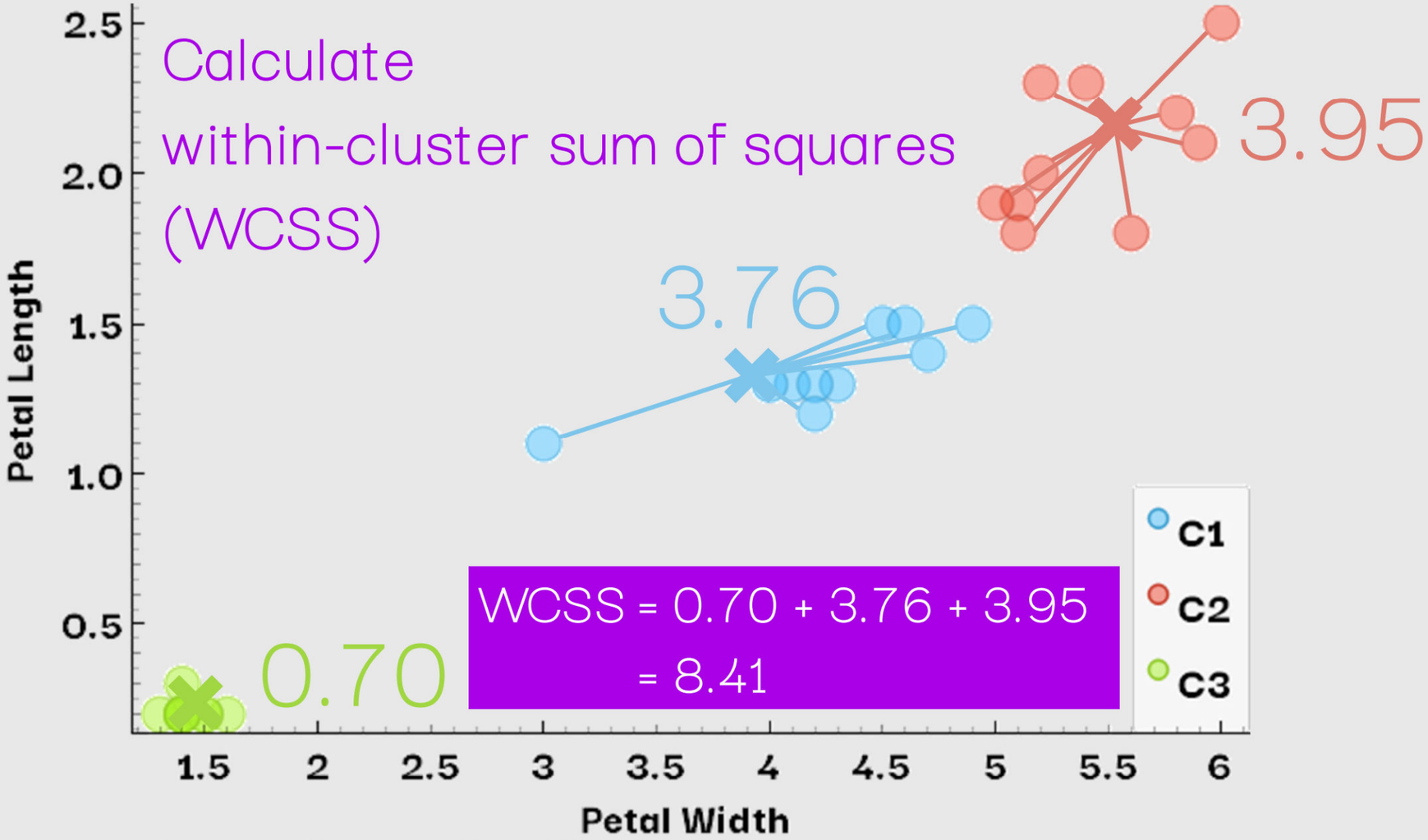
Elbow Method



Content Curated by Pollux M. Rey

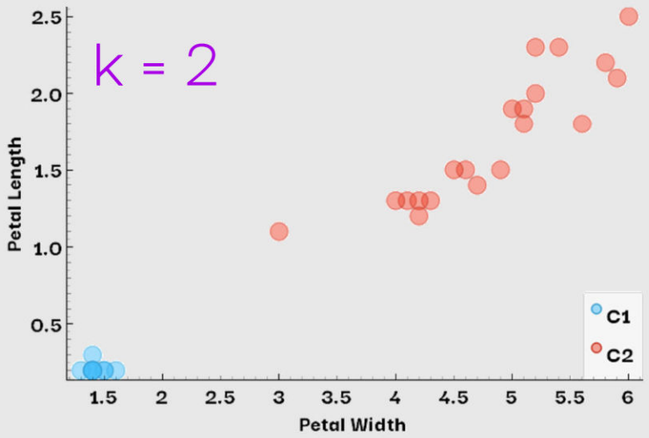
Elbow Method

Calculate
within-cluster sum of squares
(WCSS)

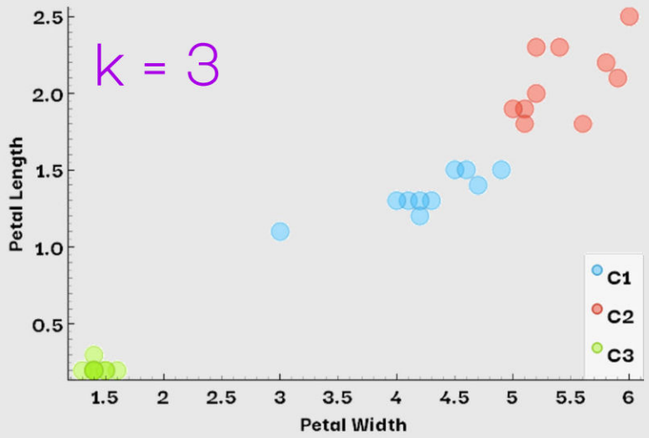


Content Curated by Pollux M. Rey

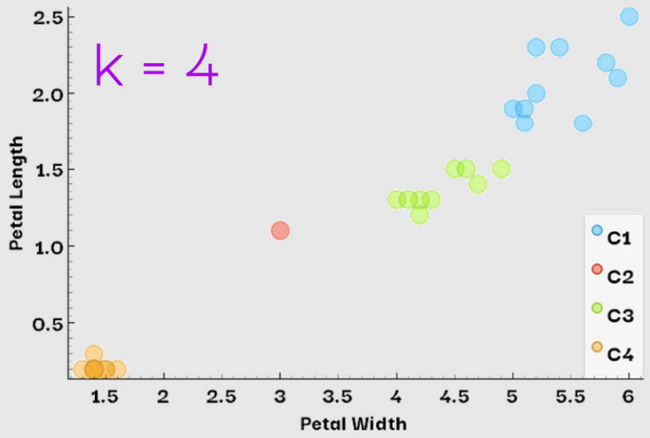
Elbow Method



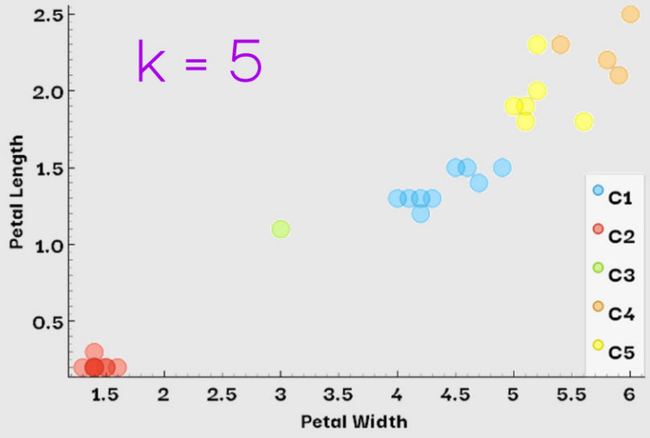
WCSS: 15.77



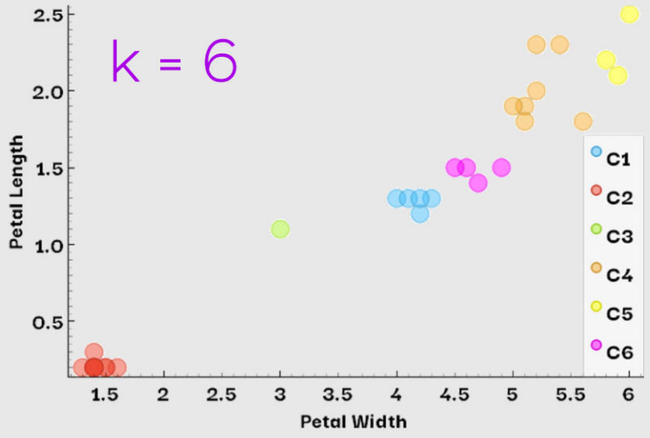
WCSS: 8.41



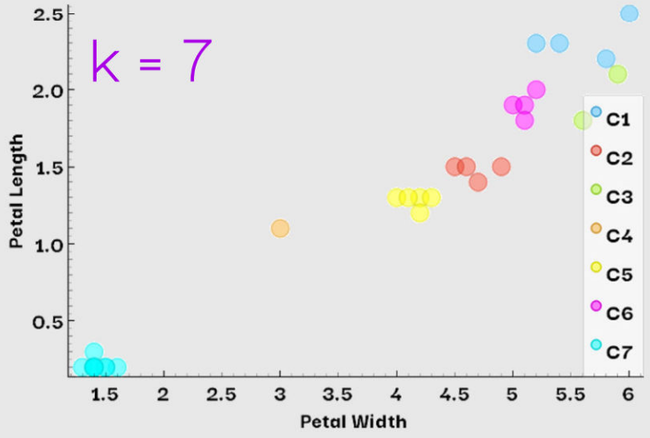
WCSS: 7.17



WCSS: 5.48



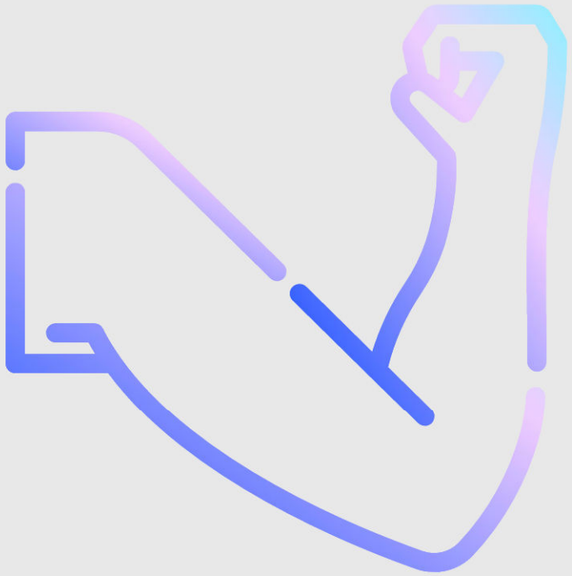
WCSS: 3.99

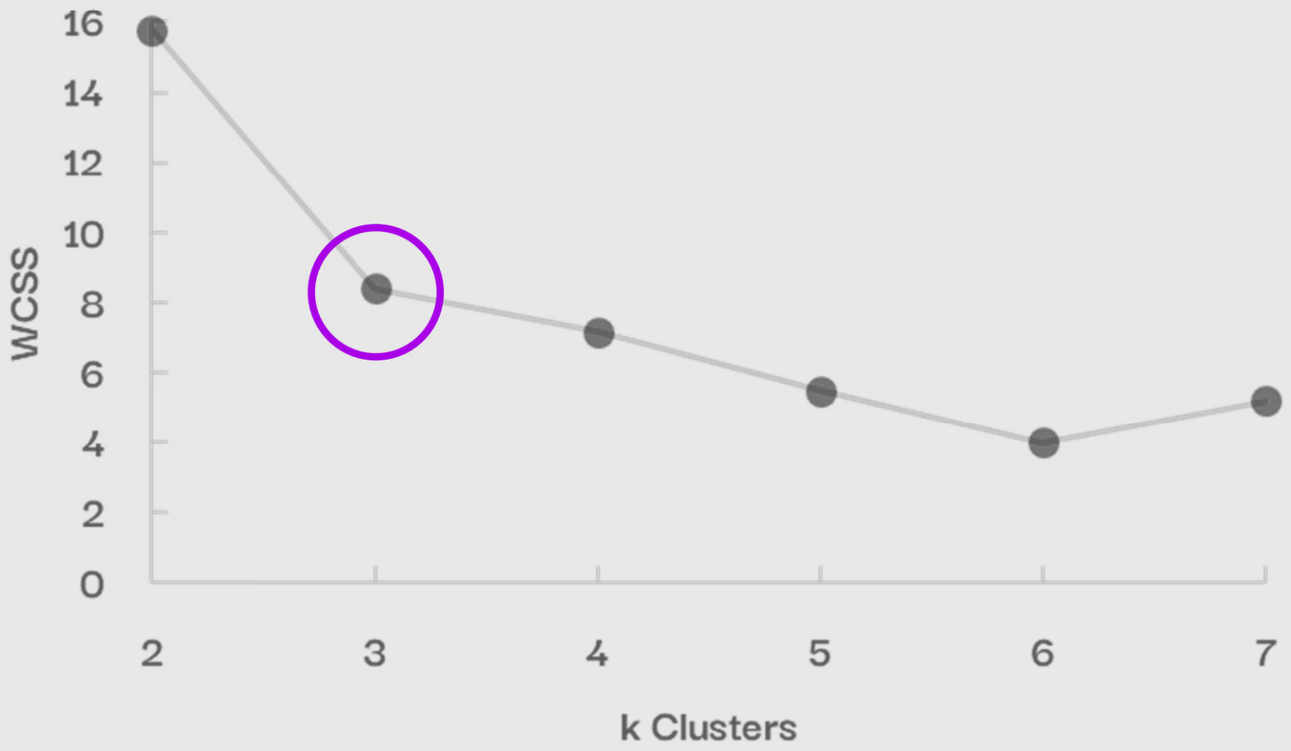


WCSS: 5.19

Content Curated by Pollux M. Rey

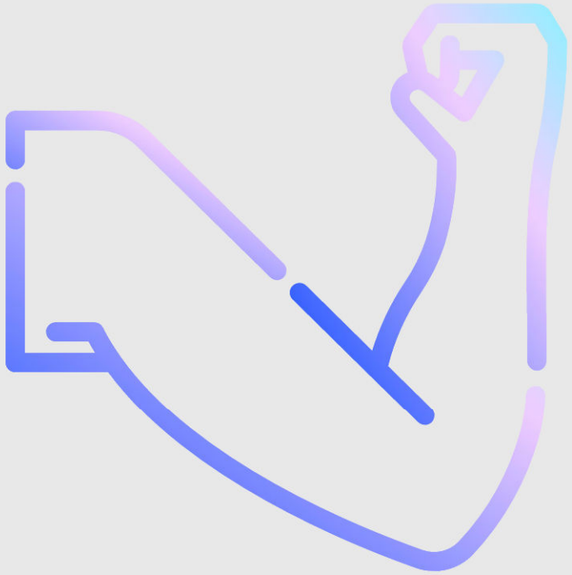
Elbow Method





Content Curated by Pollux M. Rey

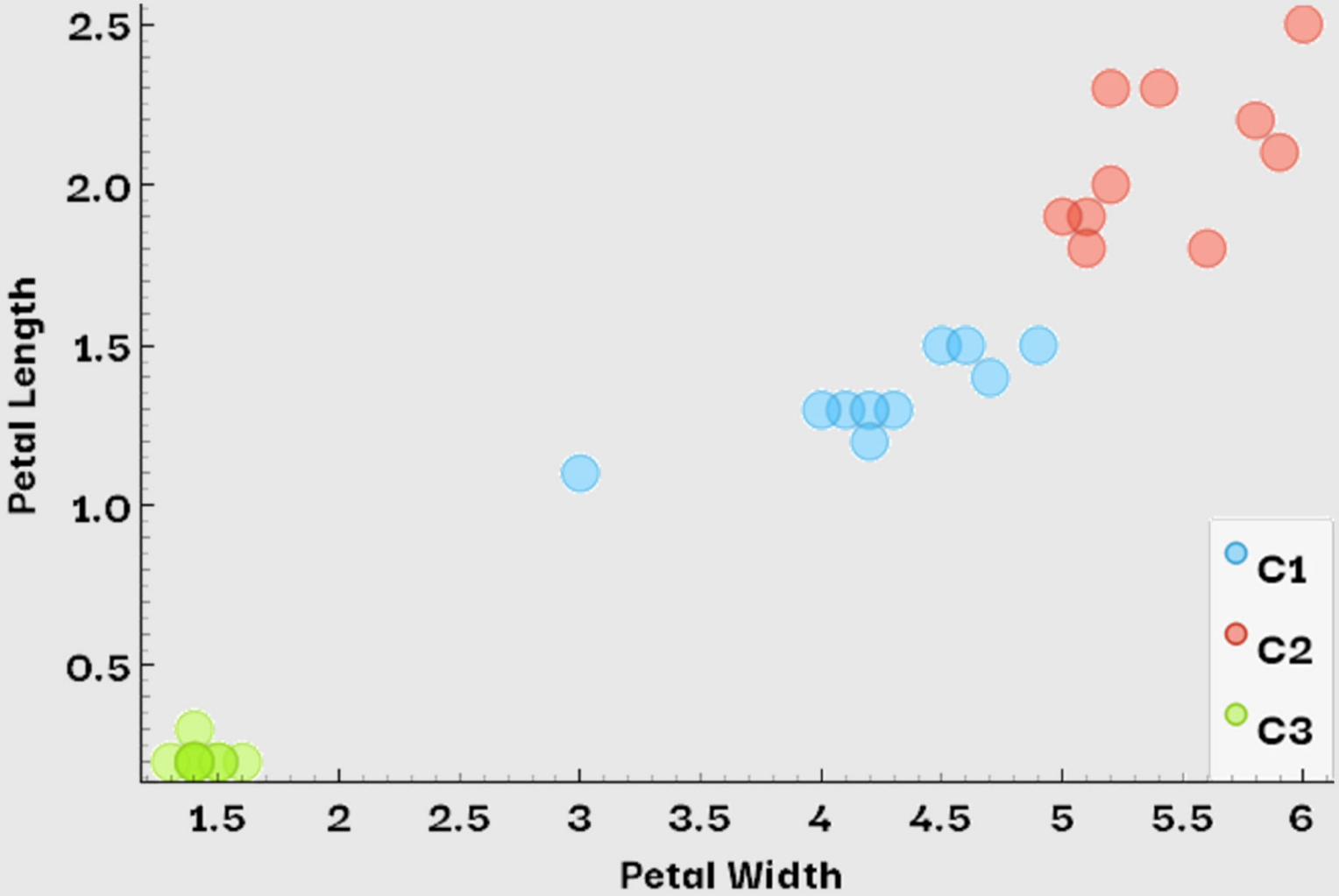
Elbow Method



The optimal number of clusters is 3.

Content Curated by Pollux M. Rey

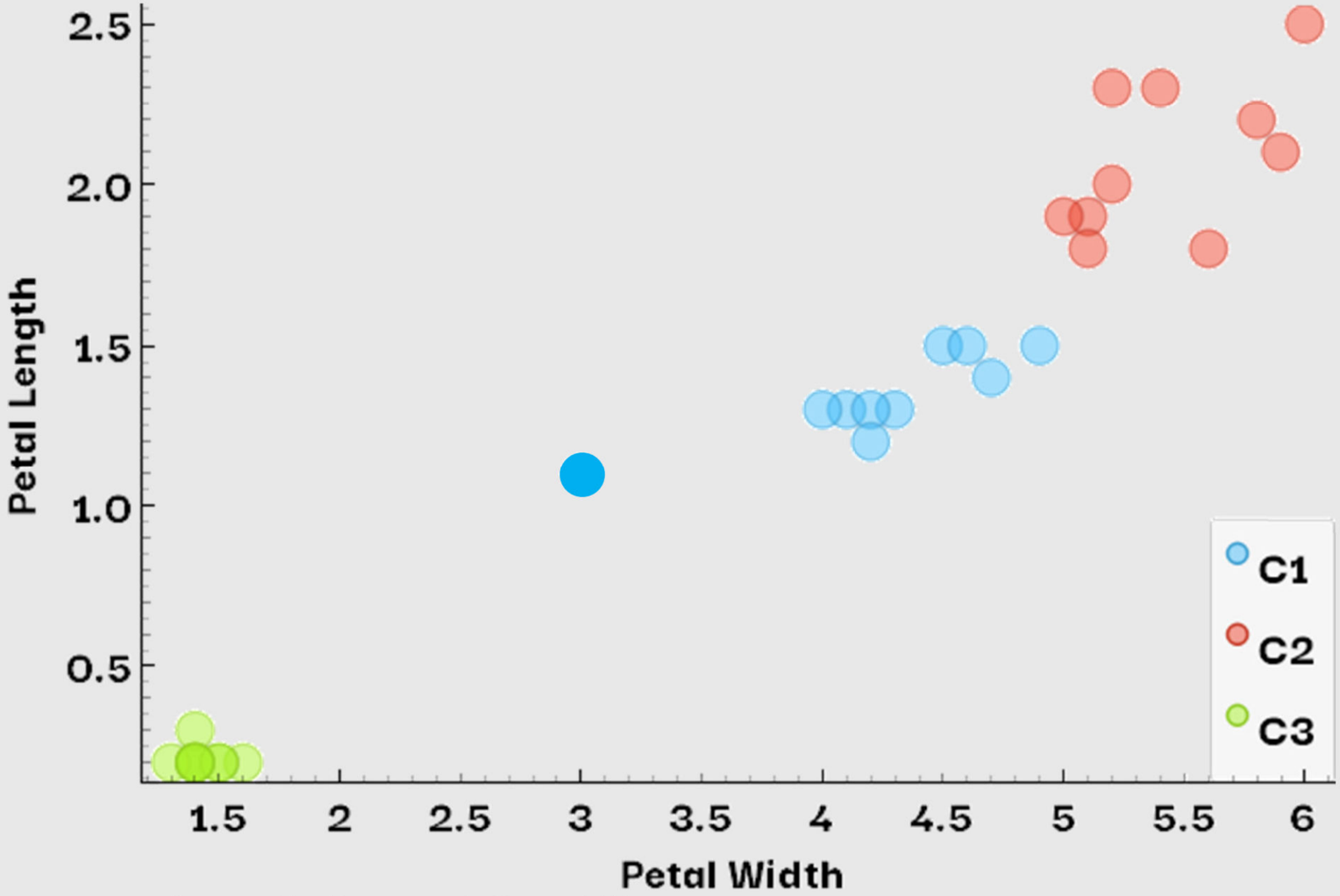
Silhouette Score



Content Curated by Pollux M. Rey

https://campus.datacamp.com/courses/cluster-analysis-in-r/k-means-clustering

Silhouette Score



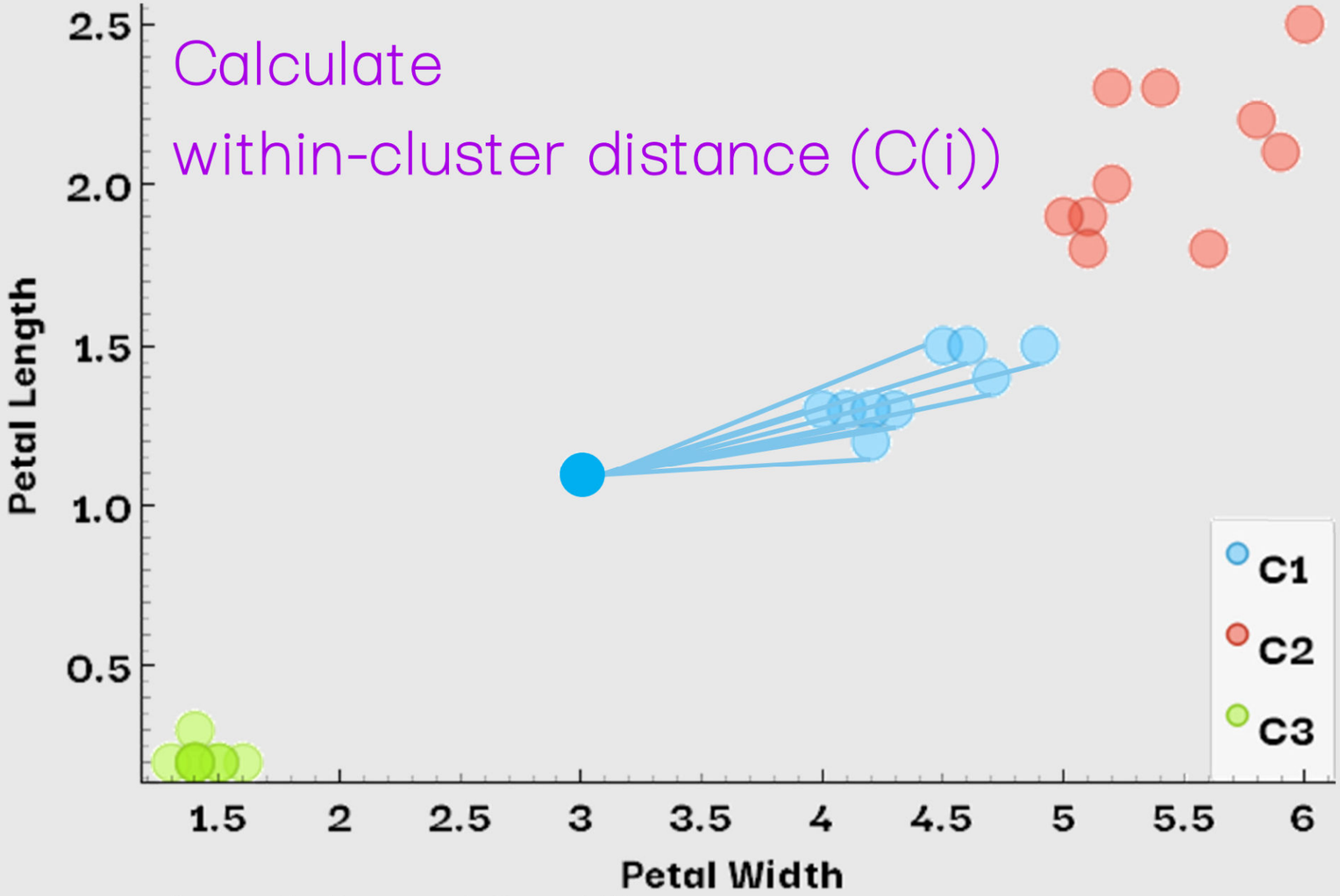
Content Curated by Pollux M. Rey

https://campus.datacamp.com/courses/cluster-analysis-in-r/k-means-clustering

Silhouette Score

Calculate

within-cluster distance (C(i))

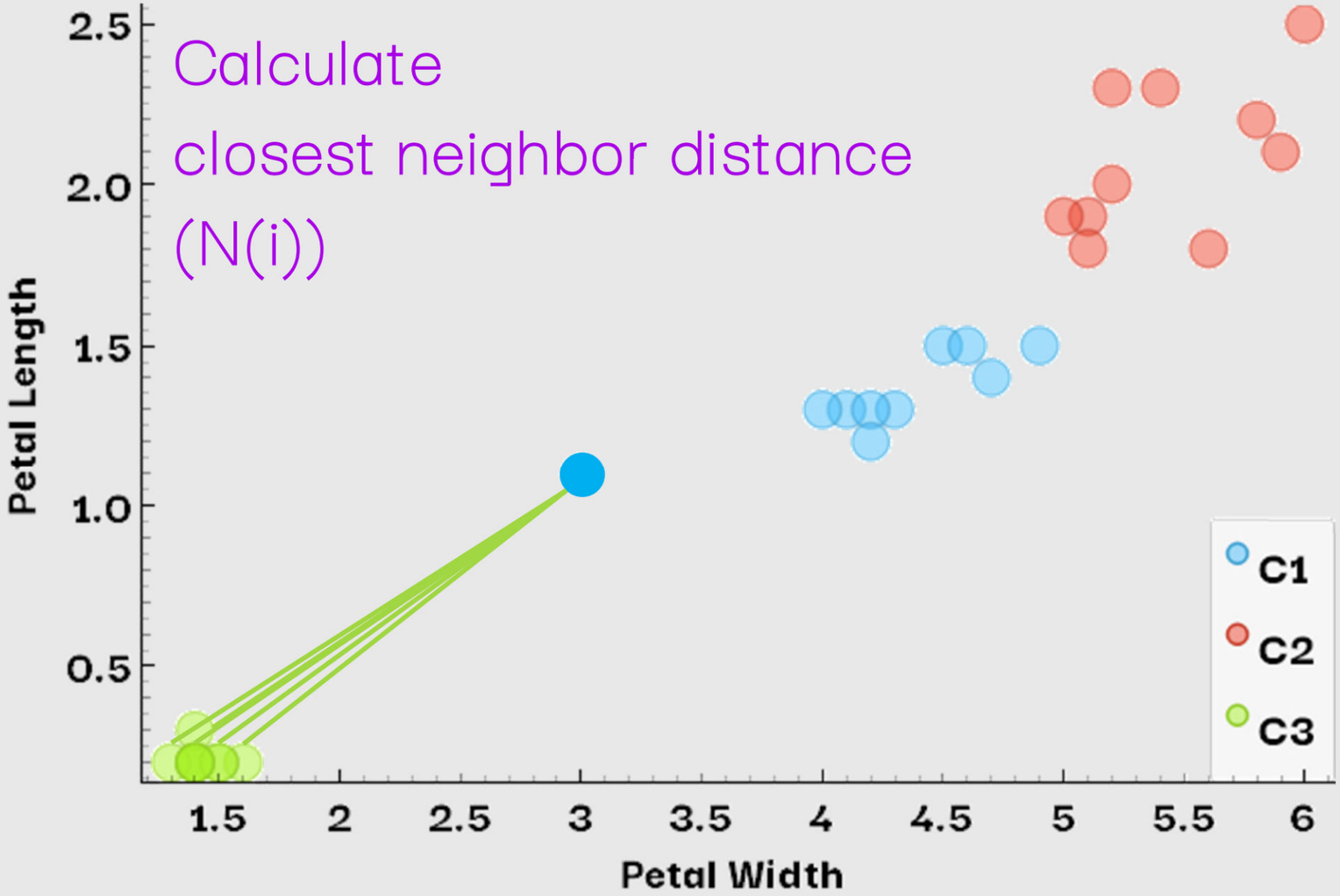


Content Curated by Pollux M. Rey

https://campus.datacamp.com/courses/cluster-analysis-in-r/k-means-clustering

Silhouette Score

Calculate
closest neighbor distance
(N(i))

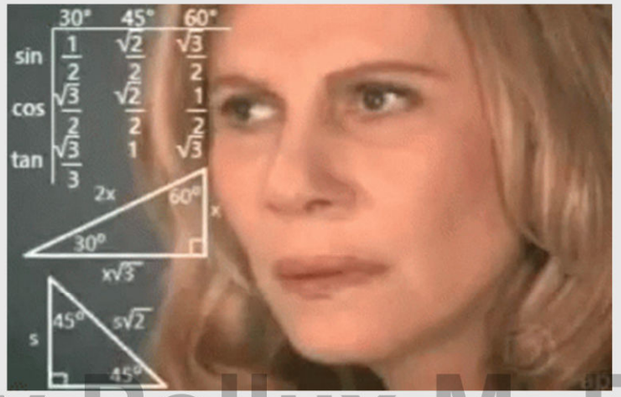


Content Curated by Pollux M. Rey

https://campus.datacamp.com/courses/cluster-analysis-in-r/k-means-clustering

Silhouette Score

$$s(i) = \begin{cases} 1 - \dfrac{C(i)}{N(i)} & \text{if } C(i) < N(i) \\ 0 & \text{if } C(i) = N(i) \\ \dfrac{N(i)}{C(i)} - 1 & \text{if } C(i) > N(i) \end{cases}$$



Content Curated by Pollux M. Rey

https://campus.datacamp.com/courses/cluster-analysis-in-r/k-means-clustering

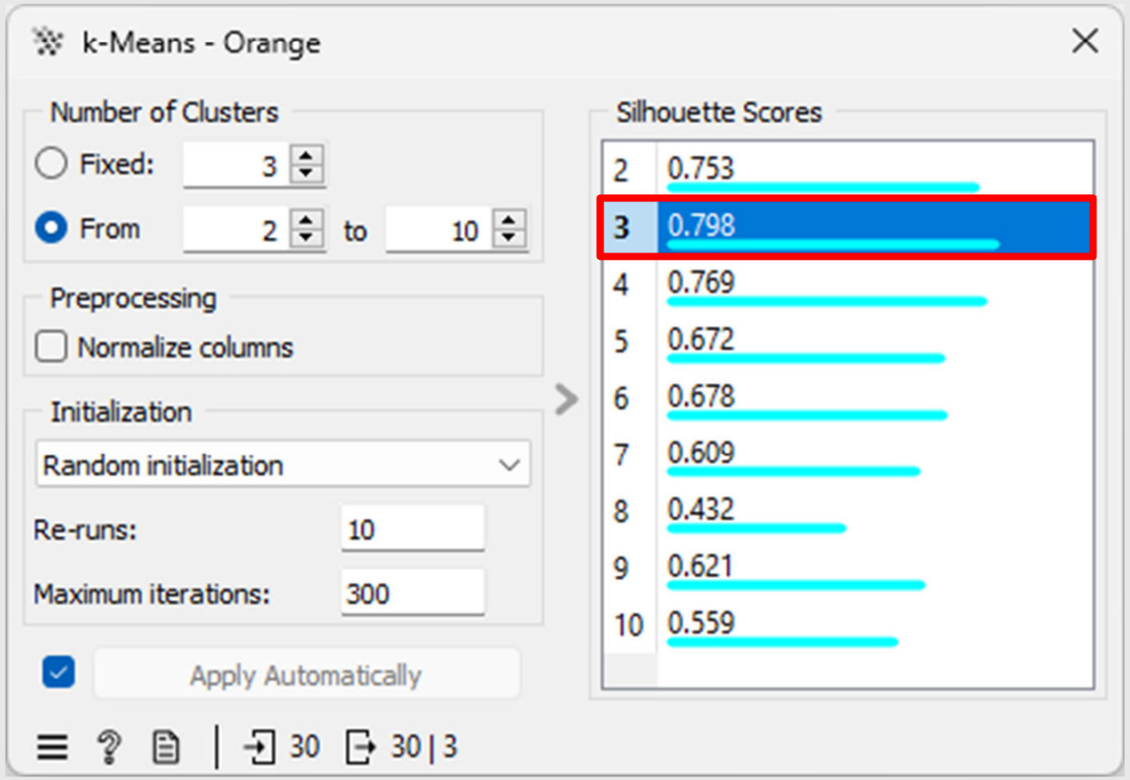
Silhouette Score

Silhouette Score	Interpretation
1	Observations are well matched to the assigned cluster.
0	Observations are on the border between two clusters.
-1	Observations may be assigned to the wrong cluster.

Content Curated by Pollux M. Rey

https://campus.datacamp.com/courses/cluster-analysis-in-r/k-means-clustering

Silhouette Score



Content Curated by Pollux M. Rey

Silhouette Score

Silhouette analysis: observation level performance

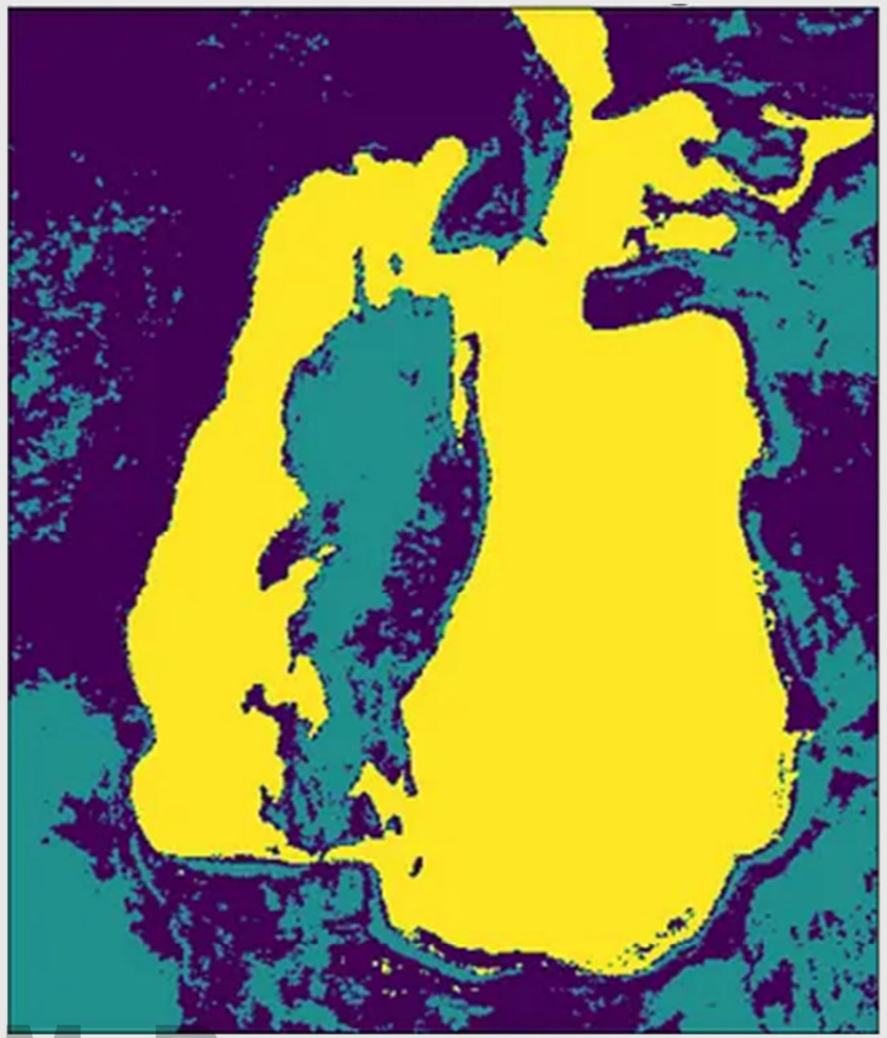
CLUSTER ANALYSIS IN R

datacamp

Content Curated by Pollux M. Rey

https://campus.datacamp.com/courses/cluster-analysis-in-r/k-means-clustering

K-Means Clustering **Applications**



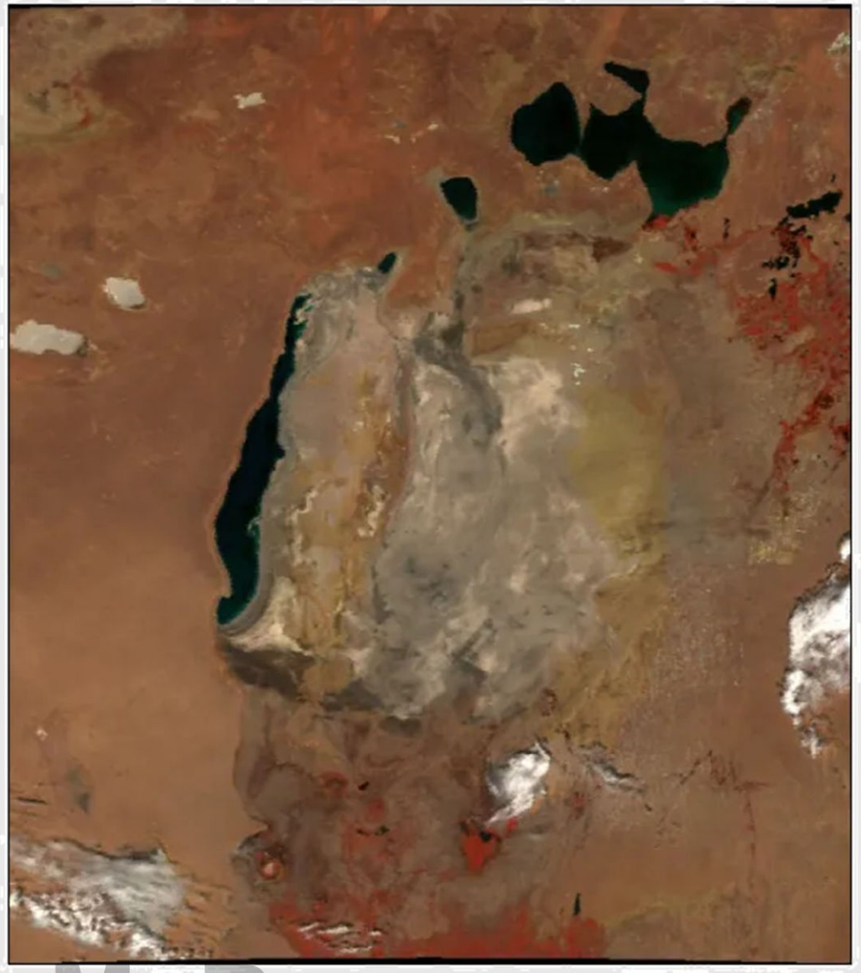
Content Curated by Pollux M. Rey

https://towardsdatascience.com/semantic-segmentation-of-remote-sensing-imagery-using-k-means-e4c165d9218e

K-Means Clustering Applications

2000-08-01

2023-08-01

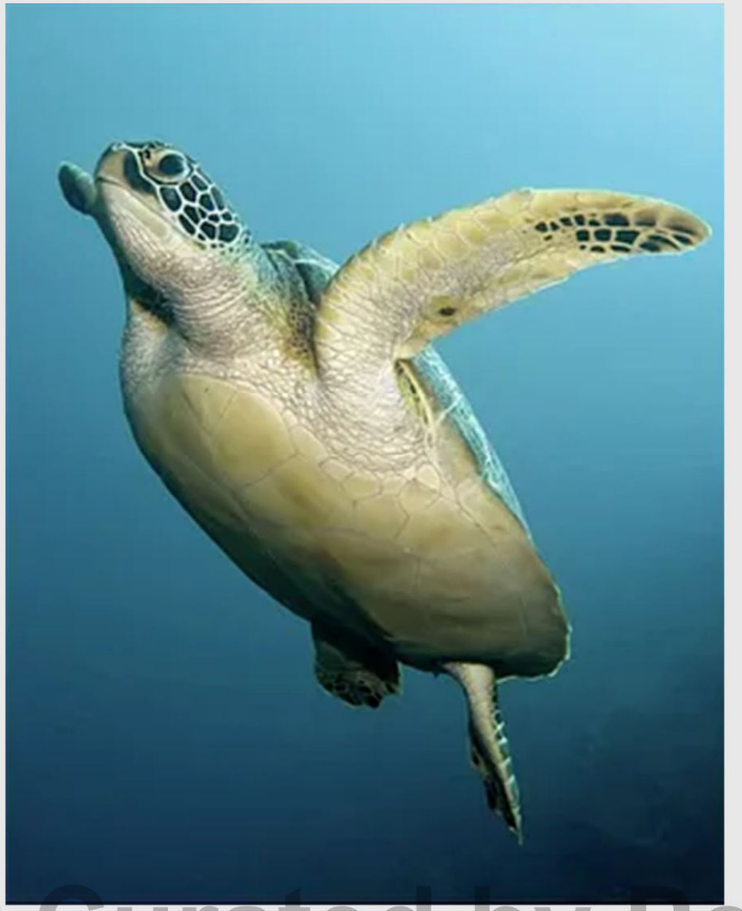


Content Curated by Pollux M. Rey

https://towardsdatascience.com/semantic-segmentation-of-remote-sensing-imagery-using-k-means-e4c165d9218e

K-Means Clustering **Applications**

Initial Image



Compressed Image

Content Curated by Pollux M. Rey

https://towardsdatascience.com/clear-and-visual-explanation-of-the-k-means-algorithm-applied-to-image-compression-b7fdc547e410

K-Means Clustering Applications

Content Curated by Pollux M. Rey

https://www.youtube.com/watch?v=yR7k19YBqiw

THANK YOU!

Content Curated by Pollux M. Rey