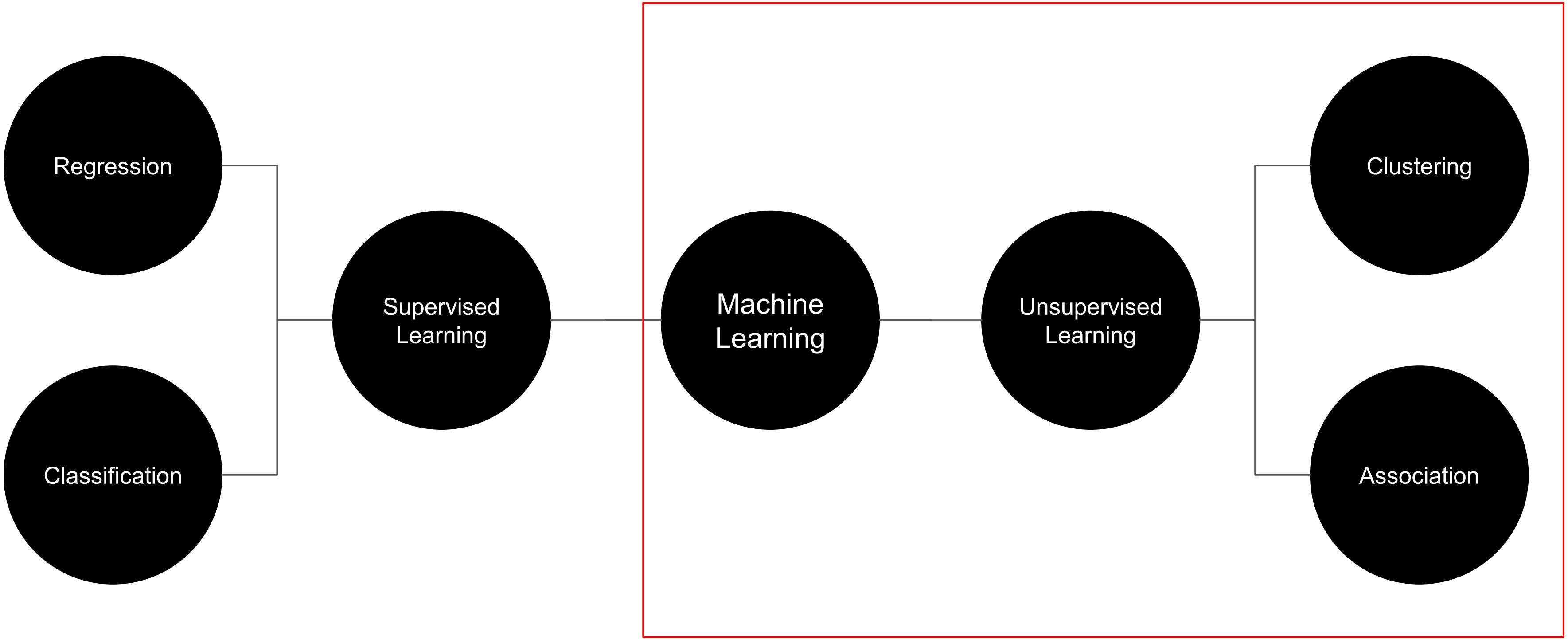
Data Prediction Model and Machine Learning

Online course #9

K-Means Clustering


Regression
Classification
Supervised Learning
Machine Learning
Unsupervised Learning
Clustering
Association

Unsupervised Learning

When unsupervised learning is useful..

- Unsupervised machine learning finds all kind of unknown patterns in data
- Unsupervised methods help you to find features which can be useful for categorization
- It is easier to get unlabelled data from a computer than labelled data, which needs manual intervention

Unsupervised Learning (Example)

Type 1: Clustering

Clustering is an important concept when it comes to unsupervised learning. It mainly deals with finding a structure or pattern in a collection of uncategorized data. Clustering algorithms will process your data and find natural clusters(groups) if they exist in the data. You can also modify how many clusters your algorithms should identify. It allows you to adjust the granularity of these groups.

sample

Cluster/group

Unsupervised Learning (Example)

Type 2: Association

Association rules allow you to establish associations amongst data objects inside large databases. This unsupervised technique is about discovering interesting relationships between variables in large databases. For example, people that buy a new home most likely to buy new furniture.

Other Examples:

- Groups of shopper based on their browsing and purchasing histories
- Movie group by the rating given by movies viewers

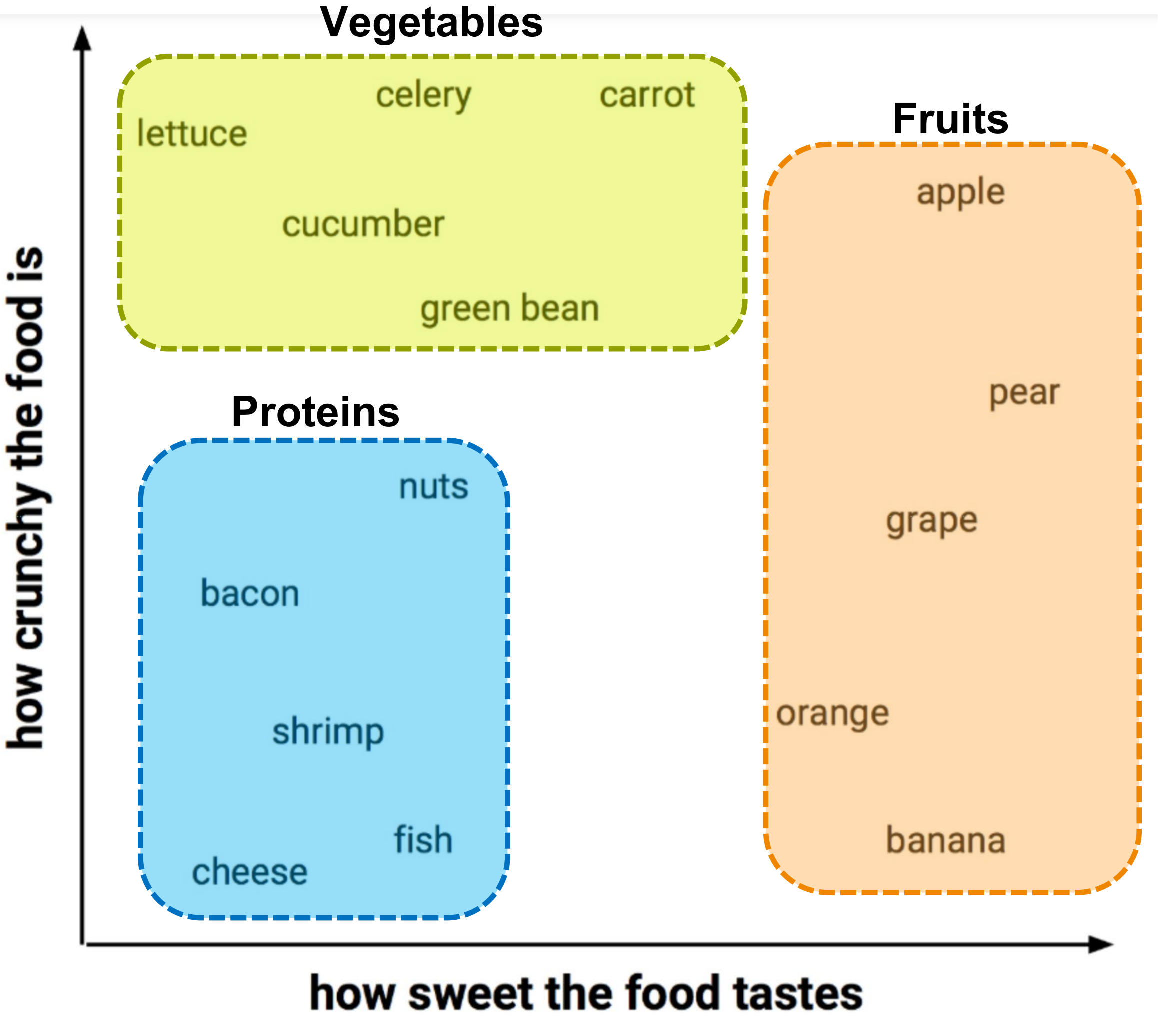
K-NN (Nearest Neighbours) ~ **K-means** Clustering

“Birds of a feather flock together”

類類相從

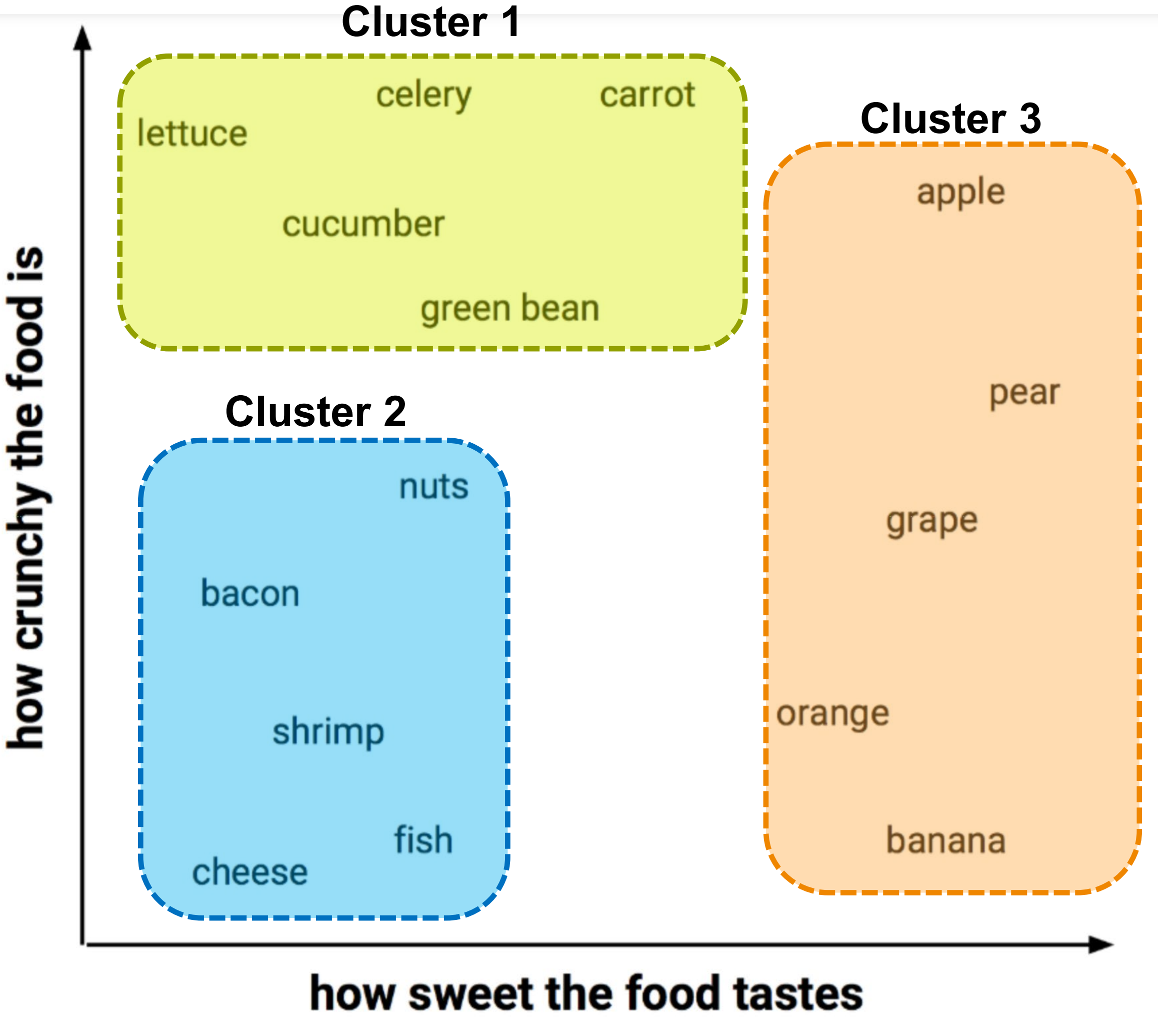
K-NN
(Nearest Neighbours)

- **Vege**: Crunchy but not sweat
- **Fruit**: Mostly sweet
- **Protein**: not so crunchy and not sweet as well



K-means clustering

- **Cluster 1**: Crunchy
- **Cluster 2**: not so crunchy and not sweet as well but not sweat
- **Cluster 3**: Mostly sweet



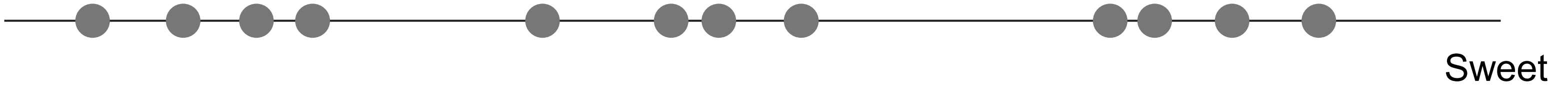
How it works? K-means clustering

How it works? K-means clustering

Step 1: Choose the number of clusters you want to identify in your data

K-means clustering

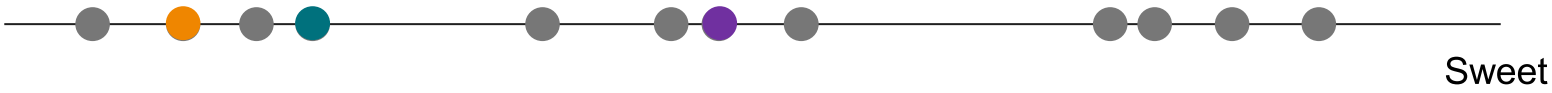
K=3



How it works? K-means clustering

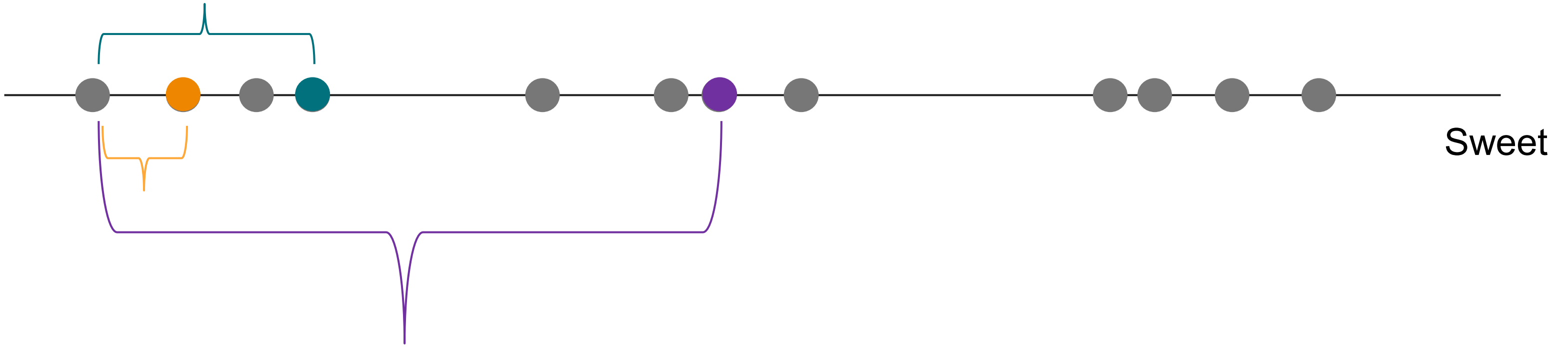
Step 2: Randomly select 3 distinct data points

They will be the initial 3 clusters' centroids



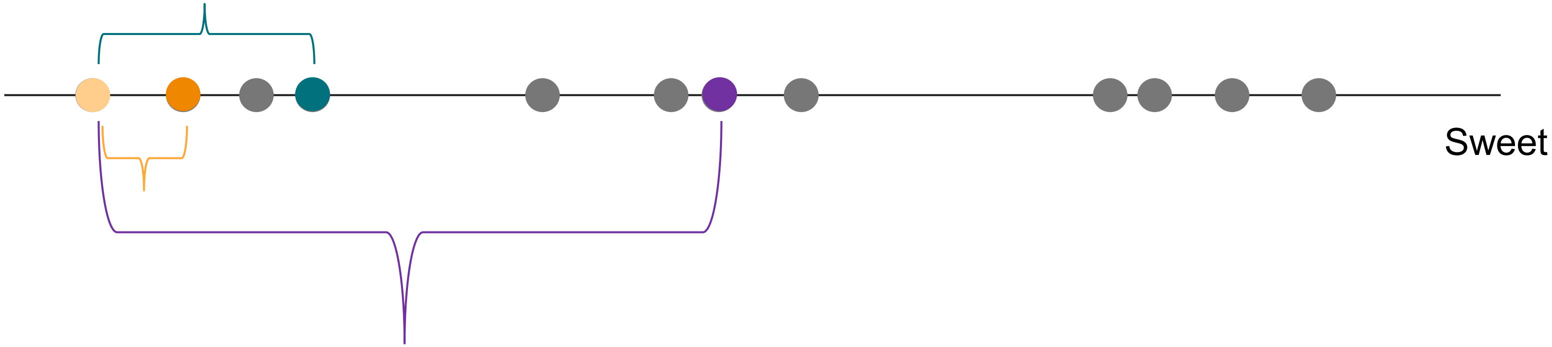
How it works? K-means clustering

Step 3: Measure the distance btw the 1st point and the three clusters' centroids



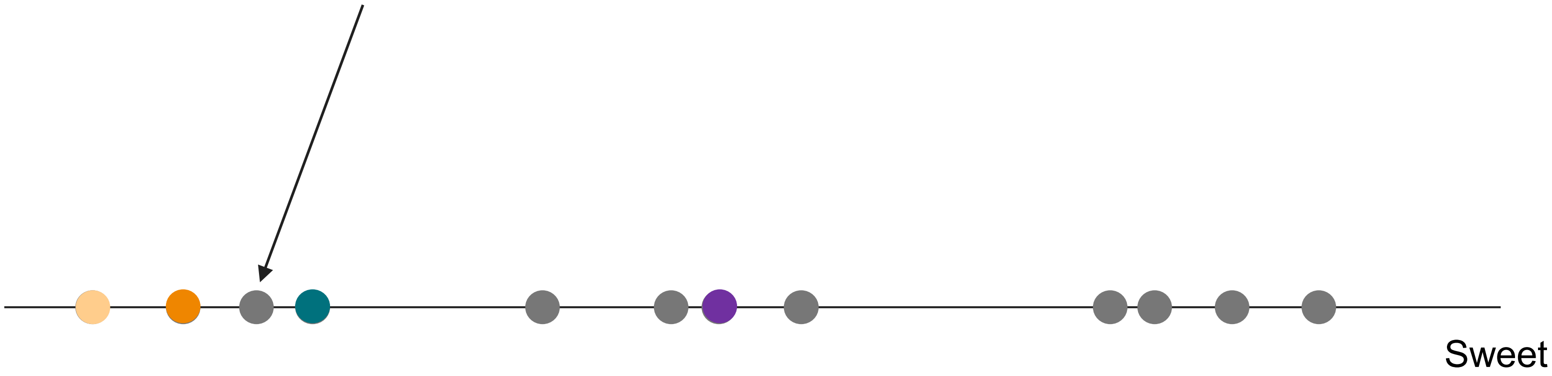
How it works? K-means clustering

Step 4: Assign the first data point to the nearest cluster



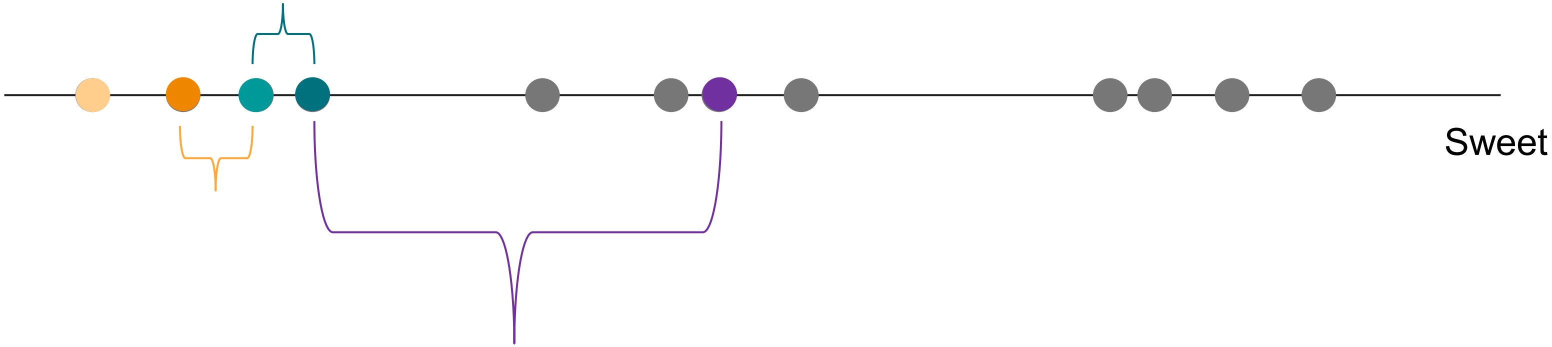
How it works? K-means clustering

Step 5: Do the same thing for the other data points left



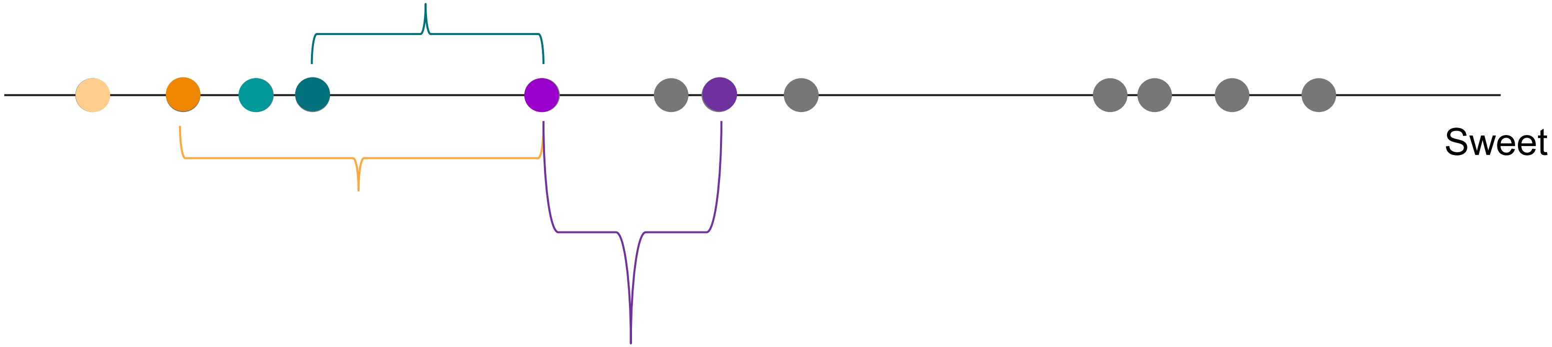
How it works? K-means clustering

Step 5: Do the same thing for the other data points left



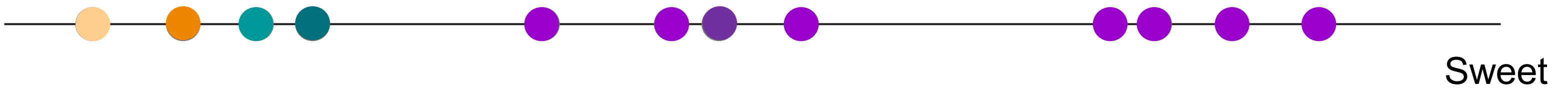
How it works? K-means clustering

Step 5: Do the same thing for the other data points left



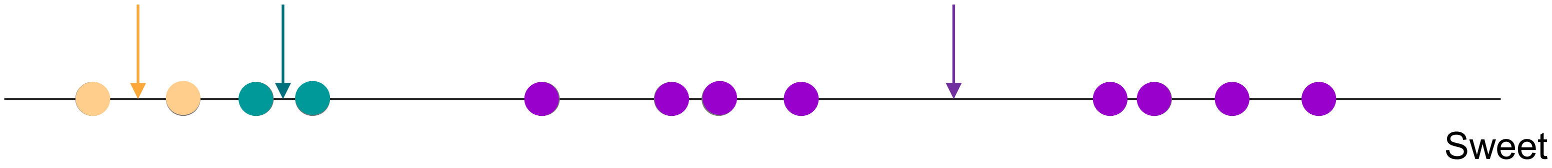
How it works? K-means clustering

Step 5: Do the same thing for the other data points left



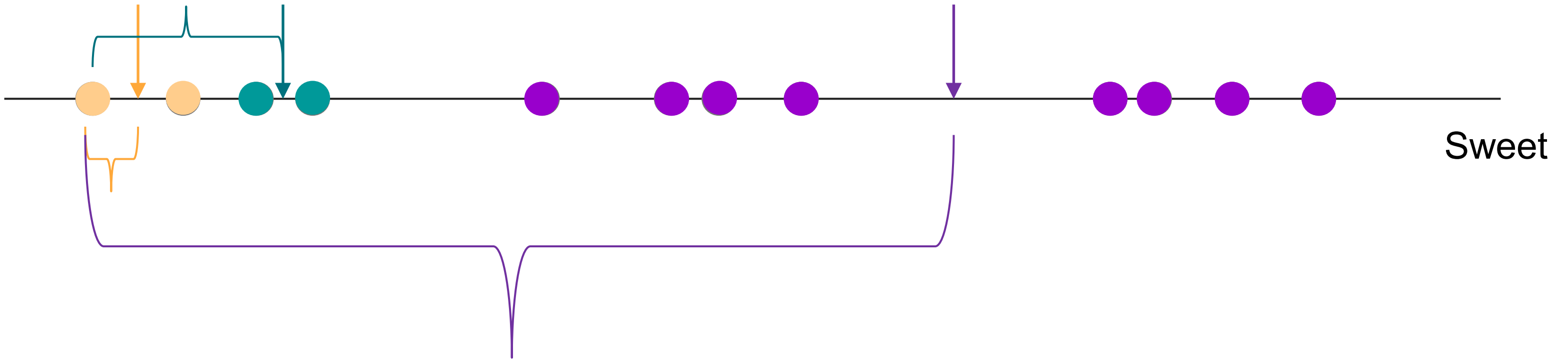
How it works? K-means clustering

Step 6: Calculate the mean of each cluster = Reassigning each cluster’s centroid



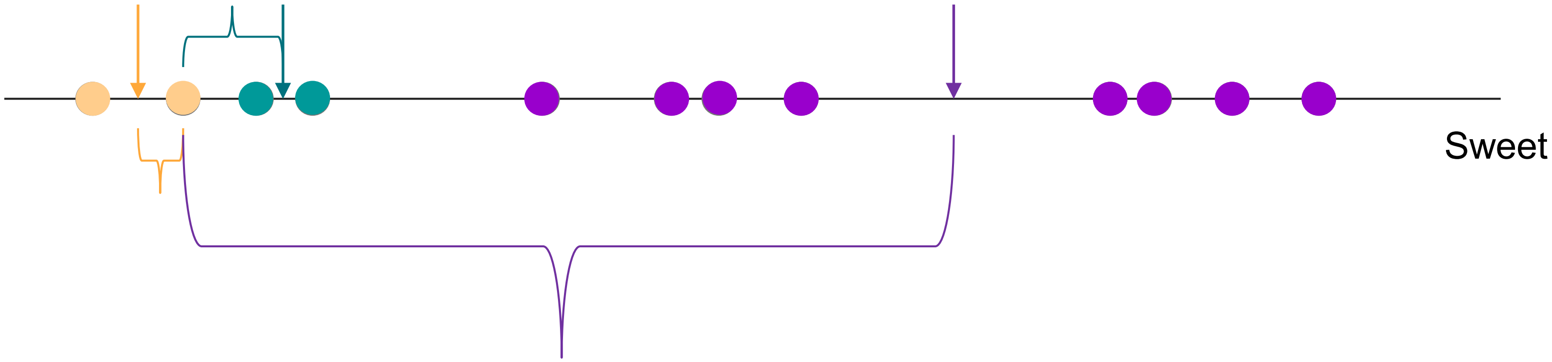
How it works? K-means clustering

Step 7: Repeat measuring the distance from each data point to clusters’ centroids



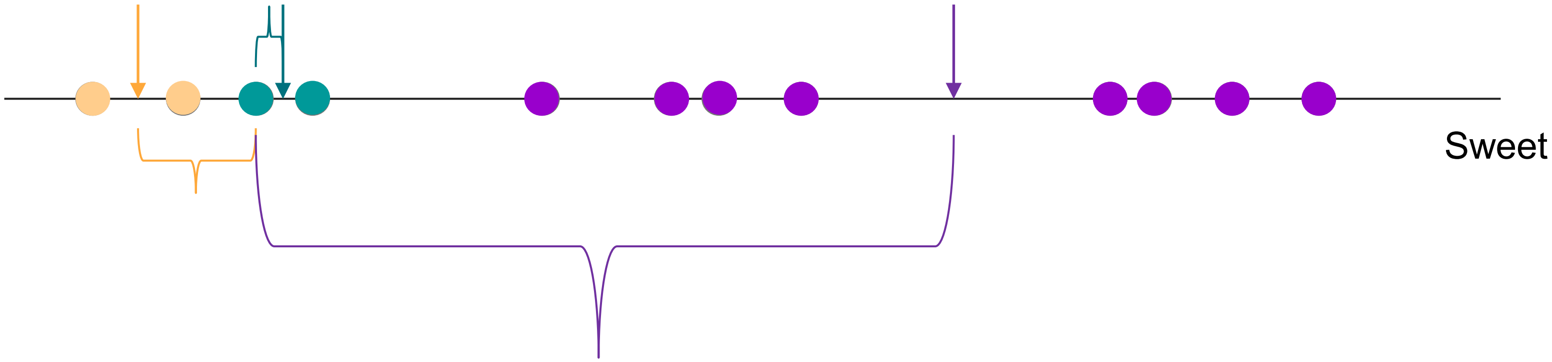
How it works? K-means clustering

Step 7: Repeat measuring the distance from each data point to clusters' centroids



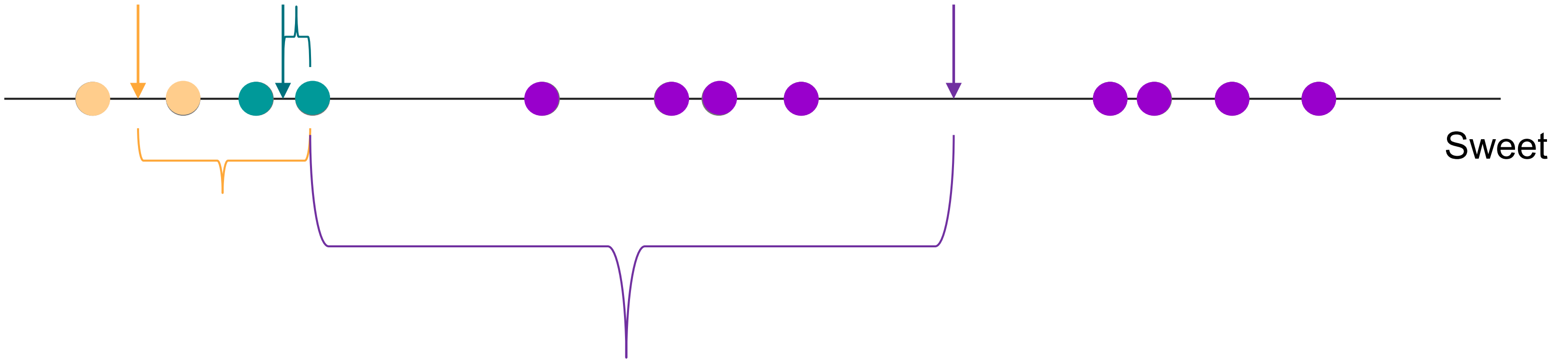
How it works? K-means clustering

Step 7: Repeat measuring the distance from each data point to clusters' centroids



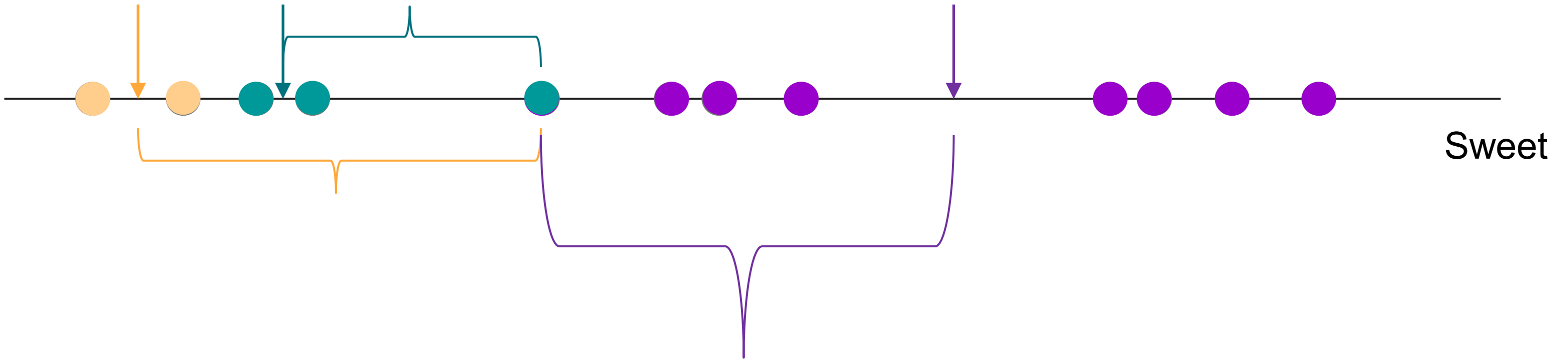
How it works? K-means clustering

Step 7: Repeat measuring the distance from each data point to clusters' centroids



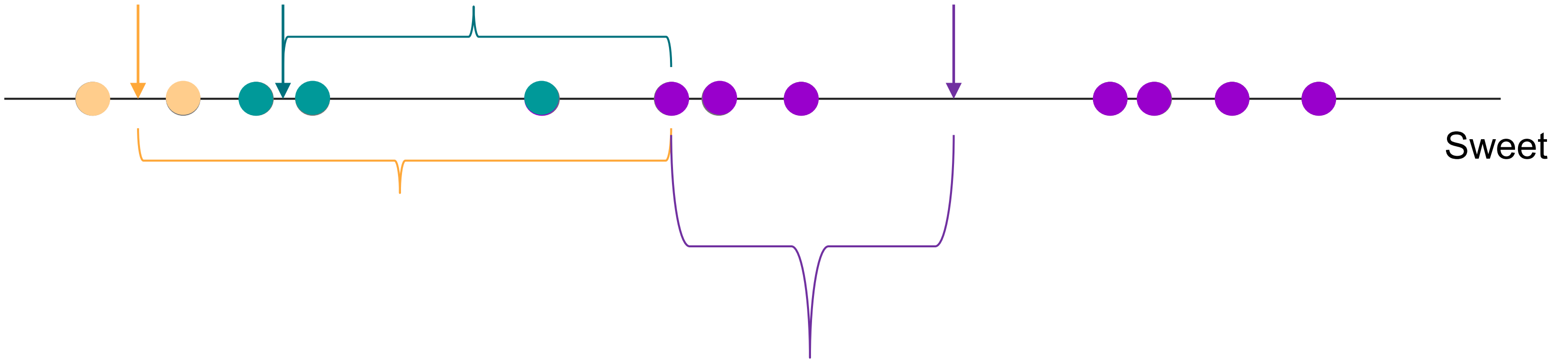
How it works? K-means clustering

Step 7: Repeat measuring the distance from each data point to clusters' centroids



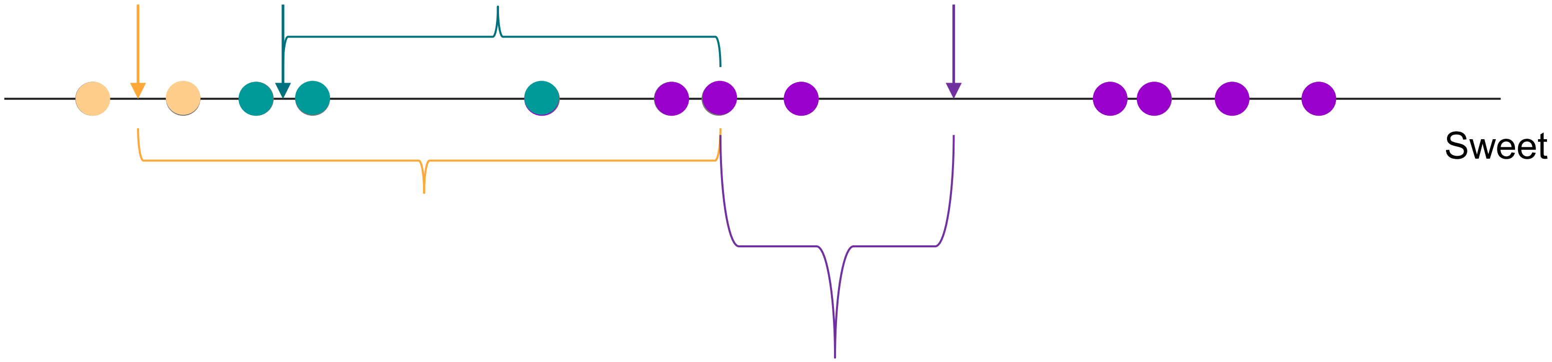
How it works? K-means clustering

Step 7: Repeat measuring the distance from each data point to clusters' centroids



How it works? K-means clustering

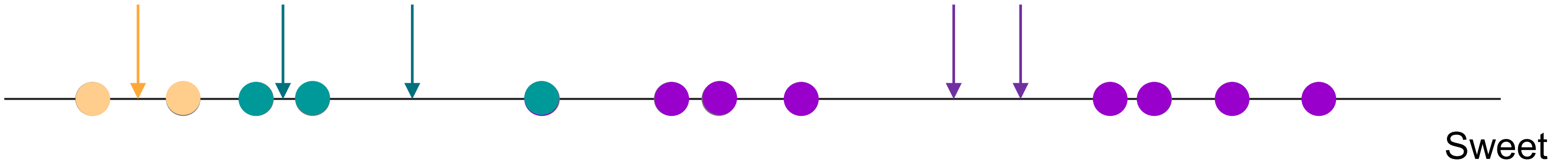
Step 7: Repeat measuring the distance from each data point to clusters' centroids



How it works? K-means clustering

Since the clustering changed we go back to **Step 6**: Reassigning each cluster's centroid

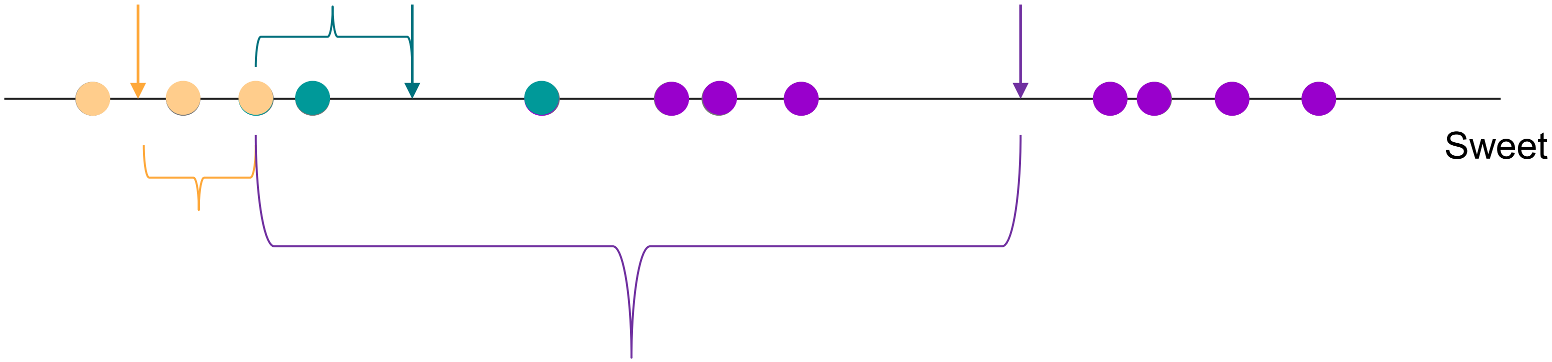
Step 6: Reassigning each cluster's centroid



How it works? K-means clustering

Since the clustering changed we go back to **Step 6**: Reassigning each cluster’s centroid

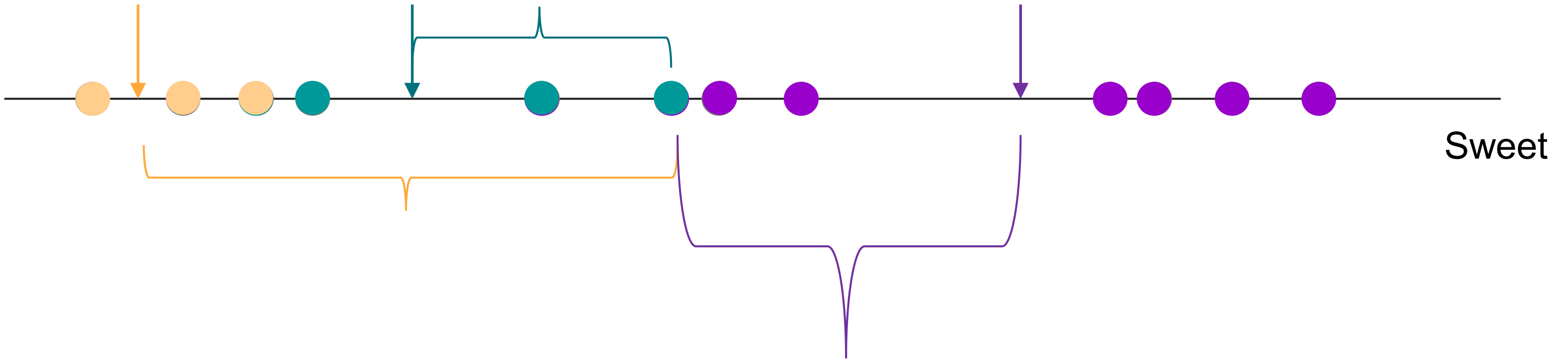
Step 7: Repeat measuring the distance from each data point to clusters’ centroids



How it works? K-means clustering

Since the clustering changed we go back to **Step 6**: Reassigning each cluster's centroid

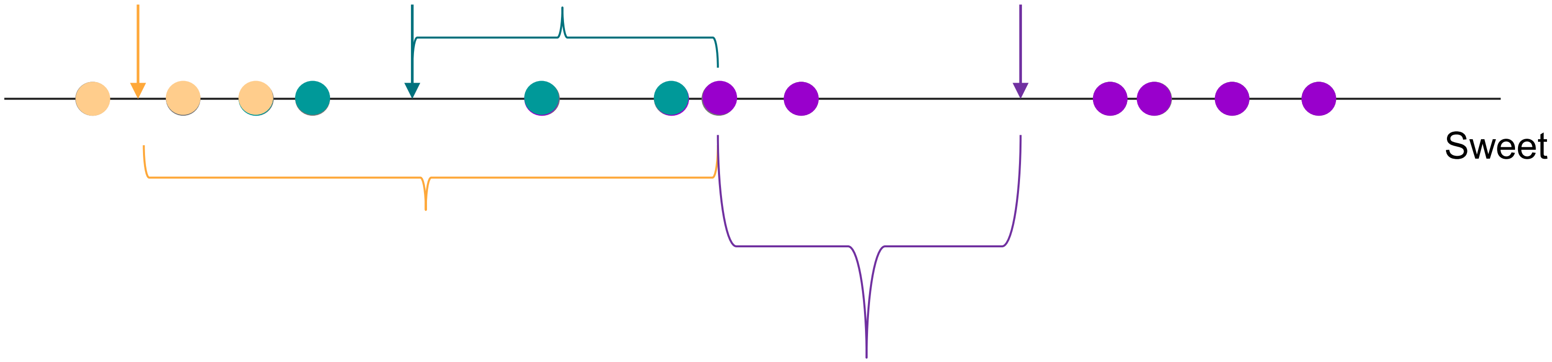
Step 7: Repeat measuring the distance from each data point to clusters' centroids



How it works? K-means clustering

Since the clustering changed we go back to **Step 6**: Reassigning each cluster's centroid

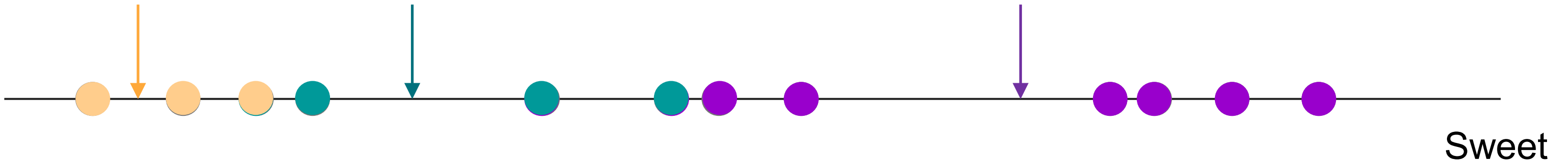
Step 7: Repeat measuring the distance from each data point to clusters' centroids



How it works? K-means clustering

Since the clustering changed we go back to **Step 6**: Reassigning each cluster's centroid

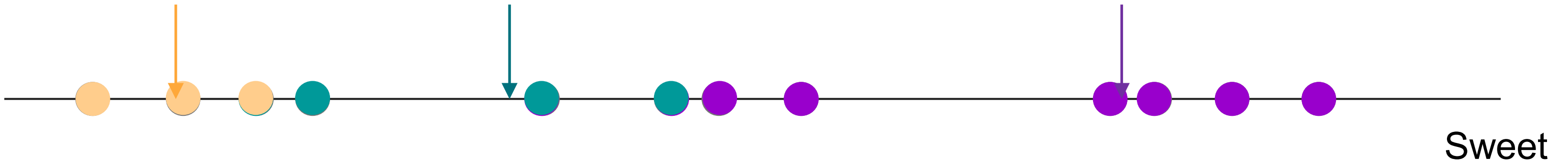
Step 6: Reassigning each cluster's centroid



How it works? K-means clustering

Since the clustering changed we go back to **Step 6**: Reassigning each cluster’s centroid

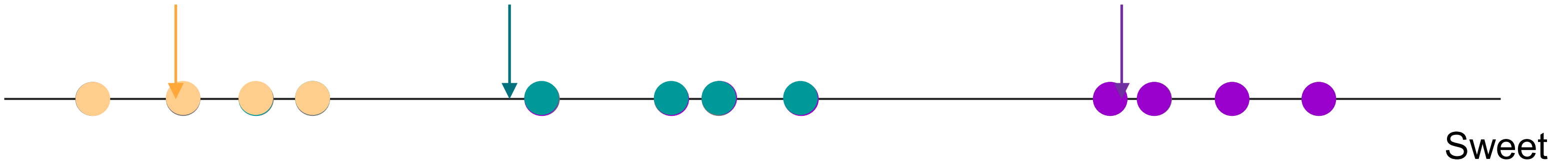
Step 6: Reassigning each cluster’s centroid



How it works? K-means clustering

Since the clustering changed we go back to **Step 6**: Reassigning each cluster's centroid

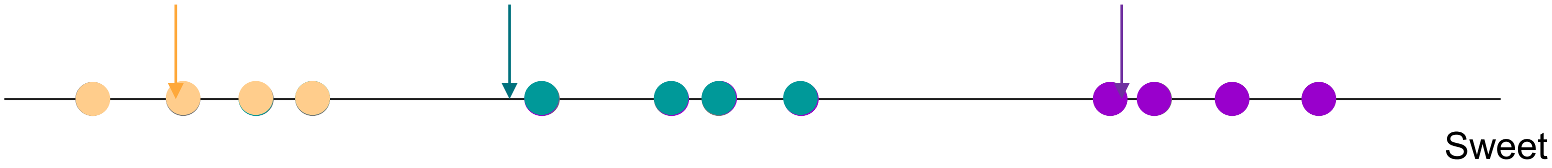
Step 7: Repeat measuring the distance from each data point to clusters' centroids



How it works? K-means clustering

Since the clustering changed we go back to **Step 6**: Reassigning each cluster's centroid

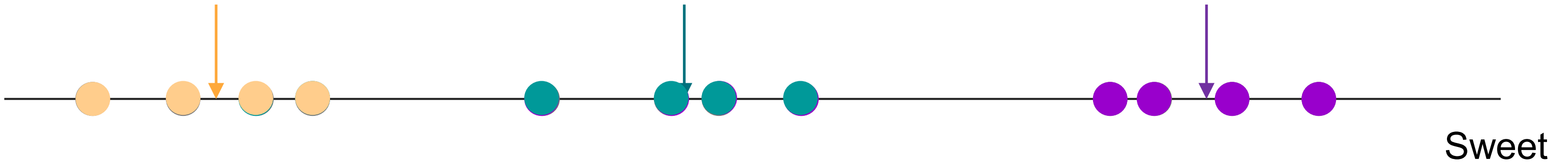
Step 6: Reassigning each cluster's centroid



How it works? K-means clustering

Since the clustering changed we go back to **Step 6**: Reassigning each cluster's centroid

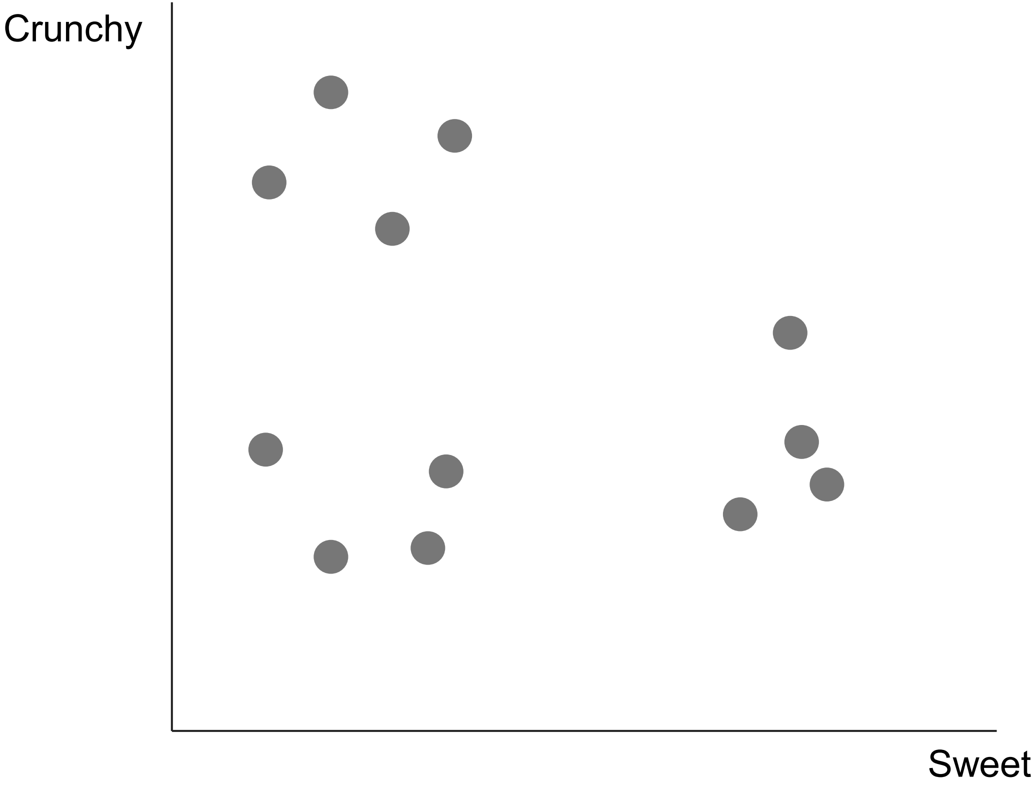
Step 7: Repeat measuring the distance from each data point to clusters' centroids



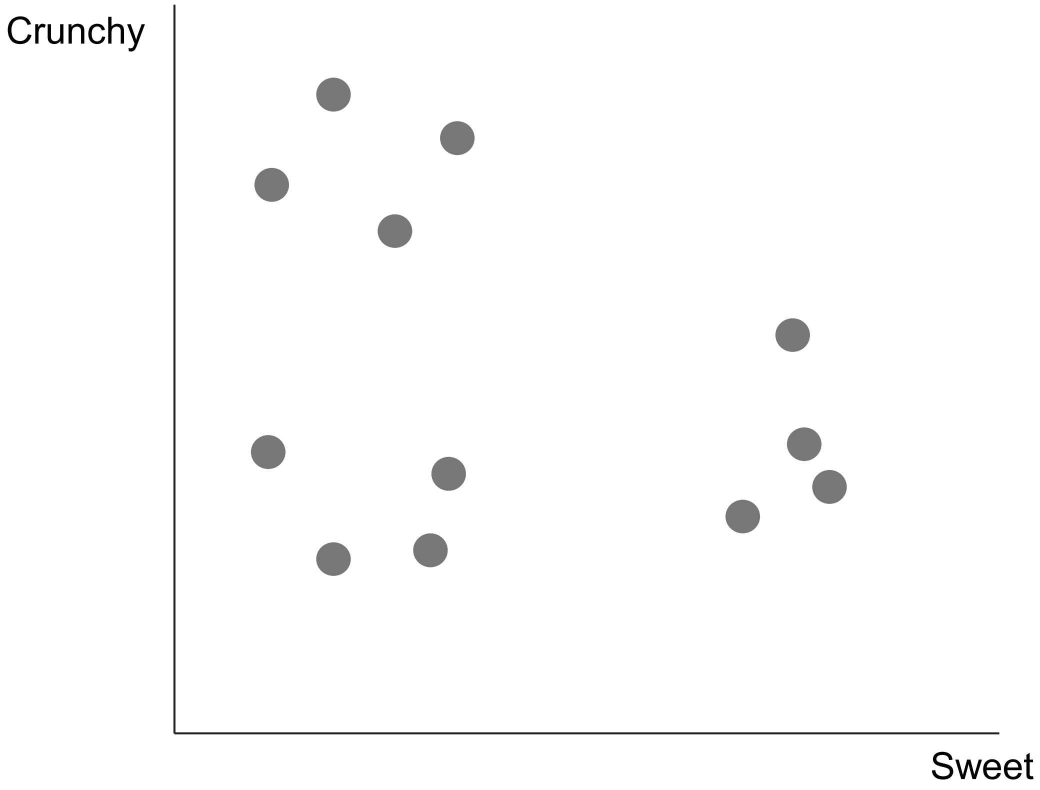
Since the clustering did not change, the algorithm stops

How it works? K-means clustering

How it works? K-means clustering

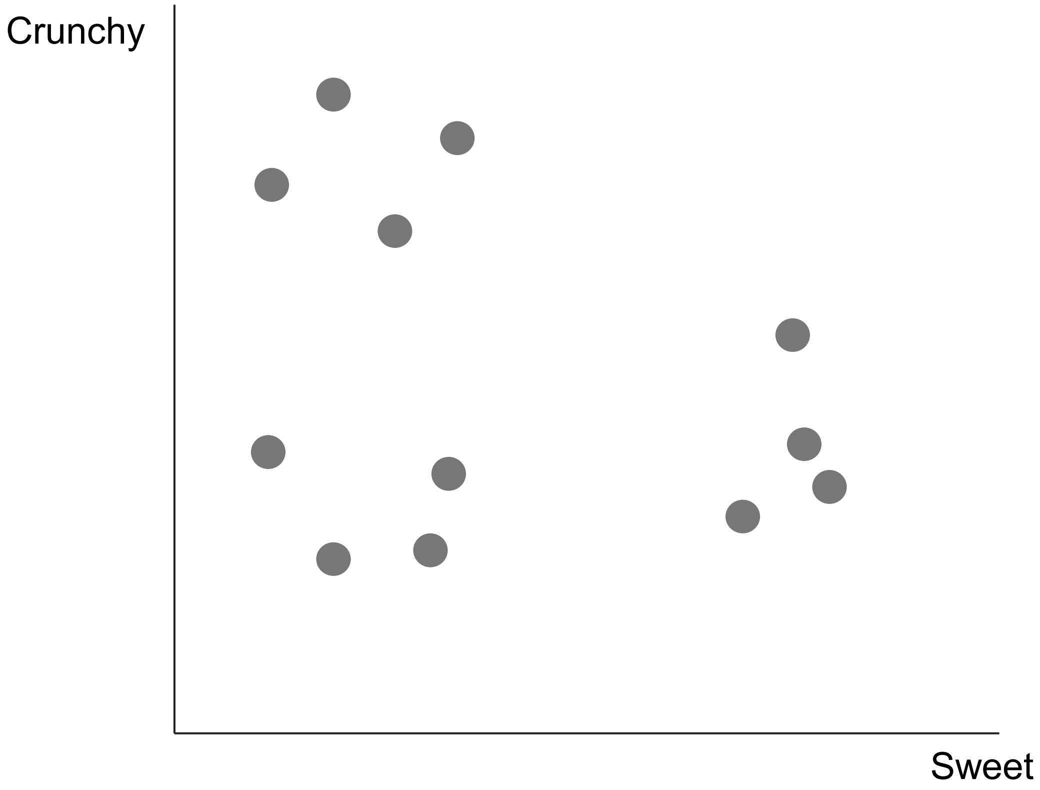


How it works? K-means clustering



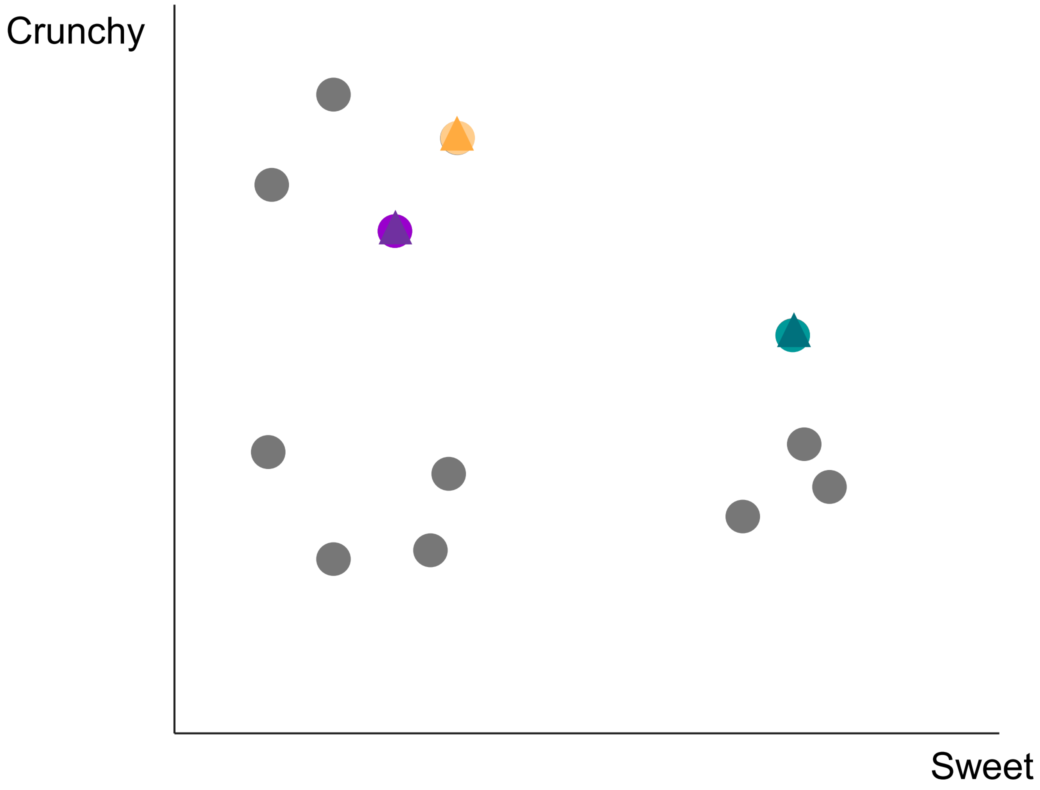
- **Step 1**: Choose the number of clusters you want to identify in your data
- **Step 2**: Randomly select ‘K’ distinct data points
- **Step 3**: Measure the distance btw each data point and the three clusters’ centroids
- **Step 4**: Assign each data point to the nearest cluster
- **Step 5**: Calculate the mean of each cluster (Centroid reassign)
- **Step 6**: Repeat measuring the distance from each data point to clusters’ centroids
- **Step 7**: Going back to step 5 if there is any change at Step 6 OR END clustering

How it works? K-means clustering



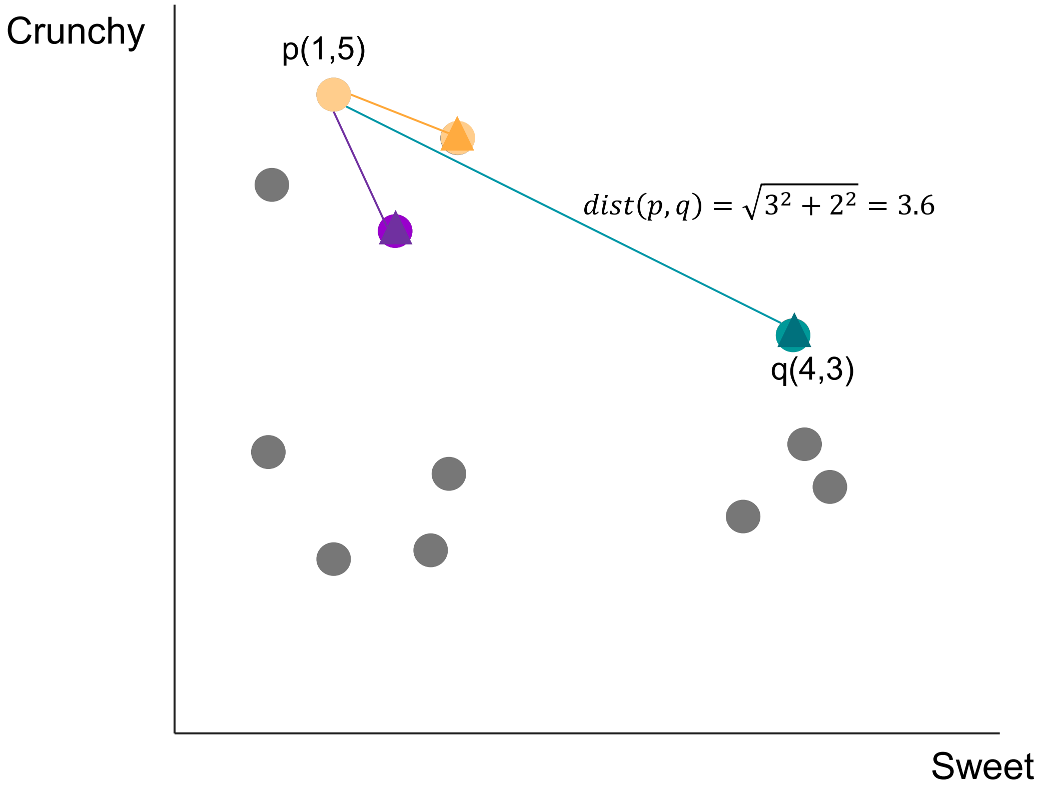
- **Step 1**: Choose the number of clusters you want to identify in your data
- **Step 2**: Randomly select ‘K’ distinct data points
- **Step 3**: Measure the distance btw each data point and the three clusters’ centroids
- **Step 4**: Assign each data point to the nearest cluster
- **Step 5**: Calculate the mean of each cluster (Centroid reassign)
- **Step 6**: Repeat measuring the distance from each data point to clusters’ centroids
- **Step 7**: Going back to step 5 if there is any change at Step 6 OR END clustering

How it works? K-means clustering



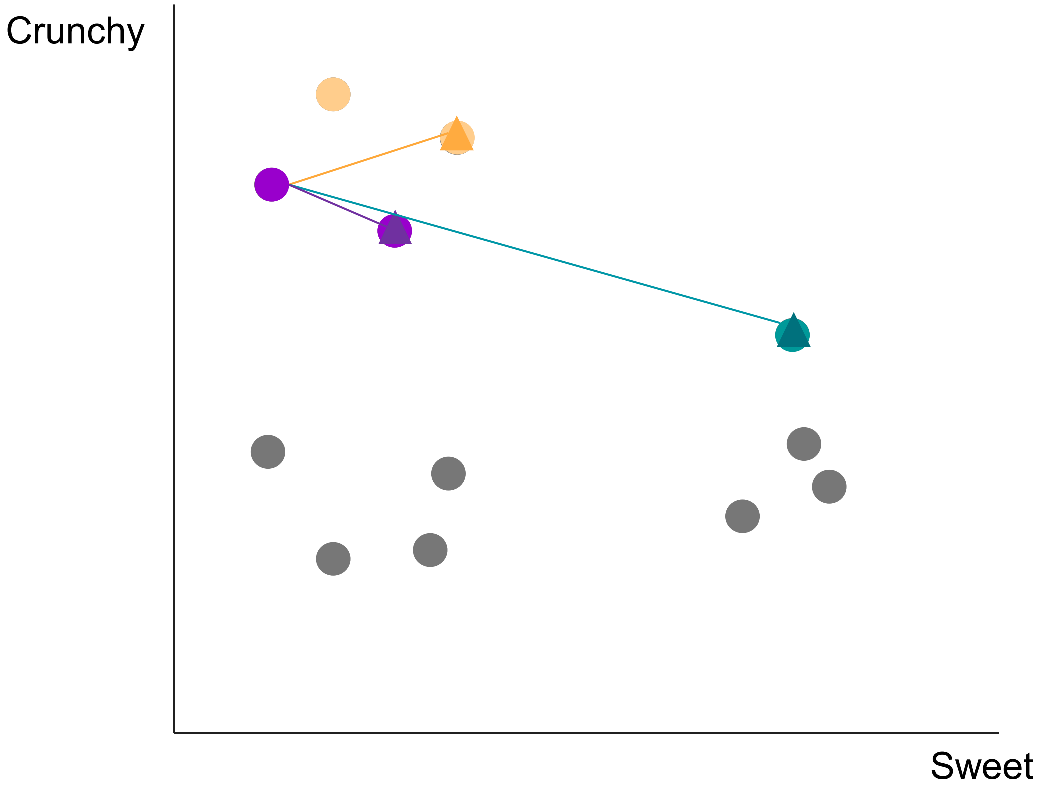
- **Step 1**: Choose the number of clusters you want to identify in your data
- **Step 2**: Randomly select 'K' distinct data points
- **Step 3**: Measure the distance btw each data point and the three clusters' centroids
- **Step 4**: Assign each data point to the nearest cluster
- **Step 5**: Calculate the mean of each cluster (Centroid reassign)
- **Step 6**: Repeat measuring the distance from each data point to clusters' centroids
- **Step 7**: Going back to step 5 if there is any change at Step 6 OR END clustering

How it works? K-means clustering



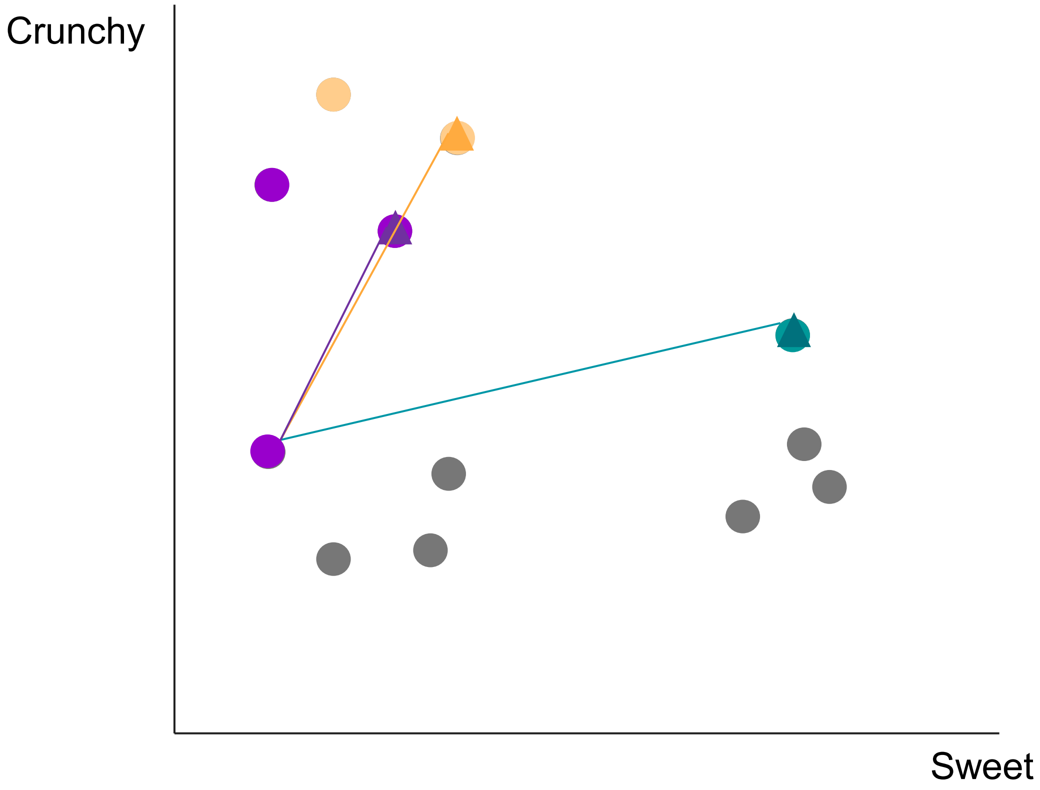
- **Step 1**: Choose the number of clusters you want to identify in your data
- **Step 2**: Randomly select 'K' distinct data points
- **Step 3**: Measure the distance btw each data point and the three clusters' centroids
- **Step 4**: Assign each data point to the nearest cluster
- **Step 5**: Calculate the mean of each cluster (Centroid reassign)
- **Step 6**: Repeat measuring the distance from each data point to clusters' centroids
- **Step 7**: Going back to step 5 if there is any change at Step 6 OR END clustering

How it works? K-means clustering



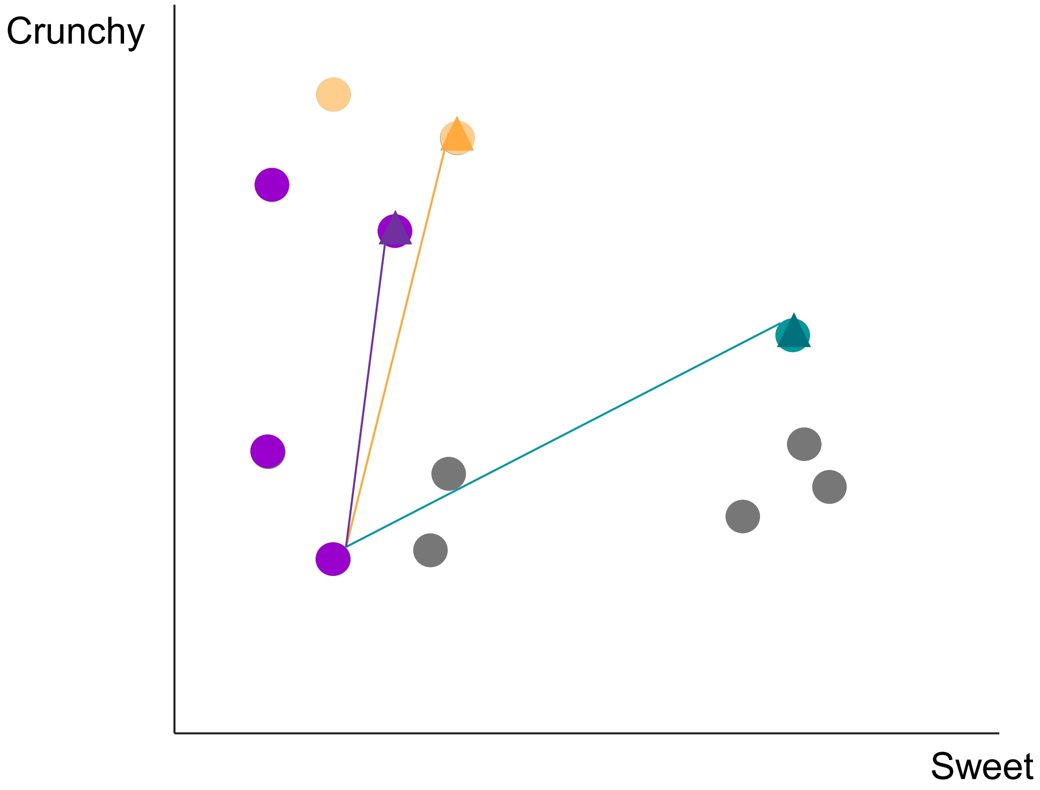
- **Step 1**: Choose the number of clusters you want to identify in your data
- **Step 2**: Randomly select 'K' distinct data points
- **Step 3**: Measure the distance btw each data point and the three clusters' centroids
- **Step 4**: Assign each data point to the nearest cluster
- **Step 5**: Calculate the mean of each cluster (Centroid reassign)
- **Step 6**: Repeat measuring the distance from each data point to clusters' centroids
- **Step 7**: Going back to step 5 if there is any change at Step 6 OR END clustering

How it works? K-means clustering



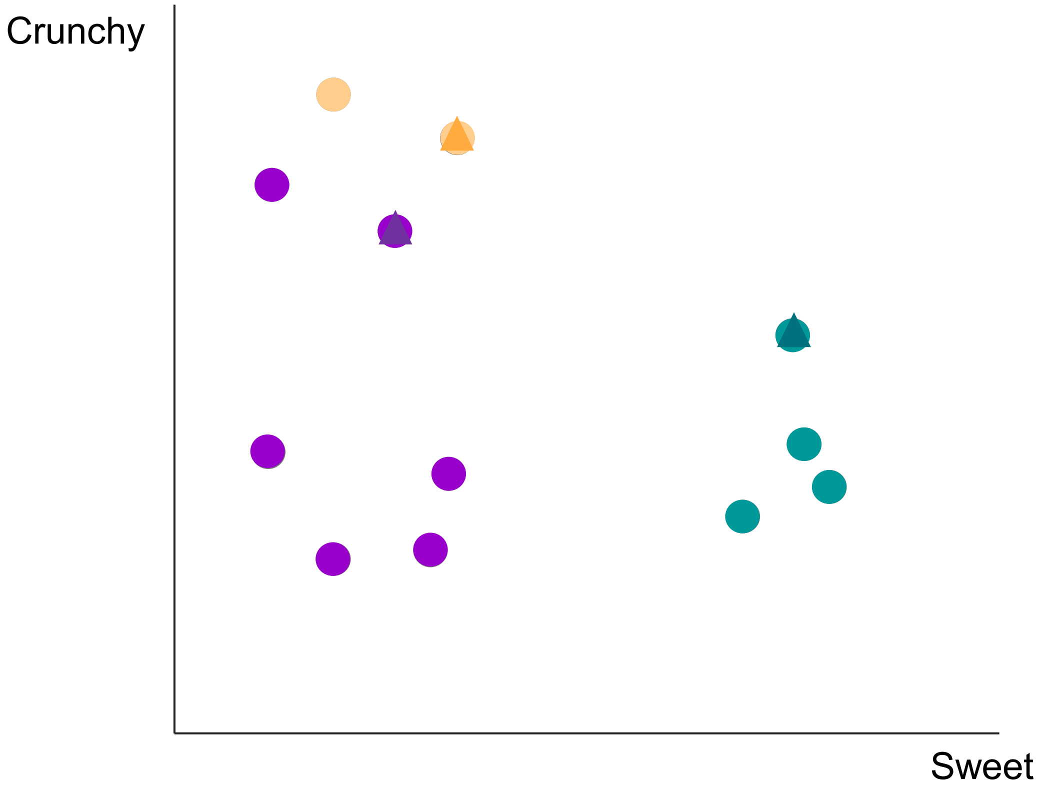
- **Step 1**: Choose the number of clusters you want to identify in your data
- **Step 2**: Randomly select ‘K’ distinct data points
- **Step 3**: Measure the distance btw each data point and the three clusters’ centroids
- **Step 4**: Assign each data point to the nearest cluster
- **Step 5**: Calculate the mean of each cluster (Centroid reassign)
- **Step 6**: Repeat measuring the distance from each data point to clusters’ centroids
- **Step 7**: Going back to step 5 if there is any change at Step 6 OR END clustering

How it works? K-means clustering



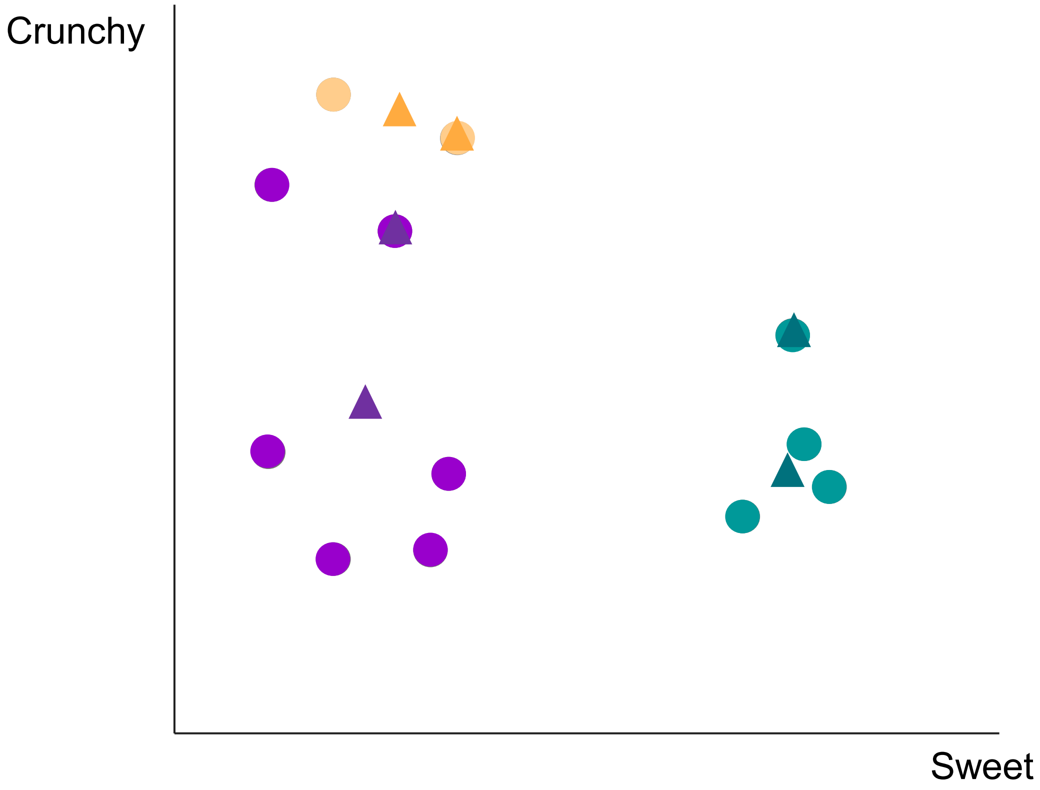
- **Step 1**: Choose the number of clusters you want to identify in your data
- **Step 2**: Randomly select ‘K’ distinct data points
- **Step 3**: Measure the distance btw each data point and the three clusters’ centroids
- **Step 4**: Assign each data point to the nearest cluster
- **Step 5**: Calculate the mean of each cluster (Centroid reassign)
- **Step 6**: Repeat measuring the distance from each data point to clusters’ centroids
- **Step 7**: Going back to step 5 if there is any change at Step 6 OR END clustering

How it works? K-means clustering



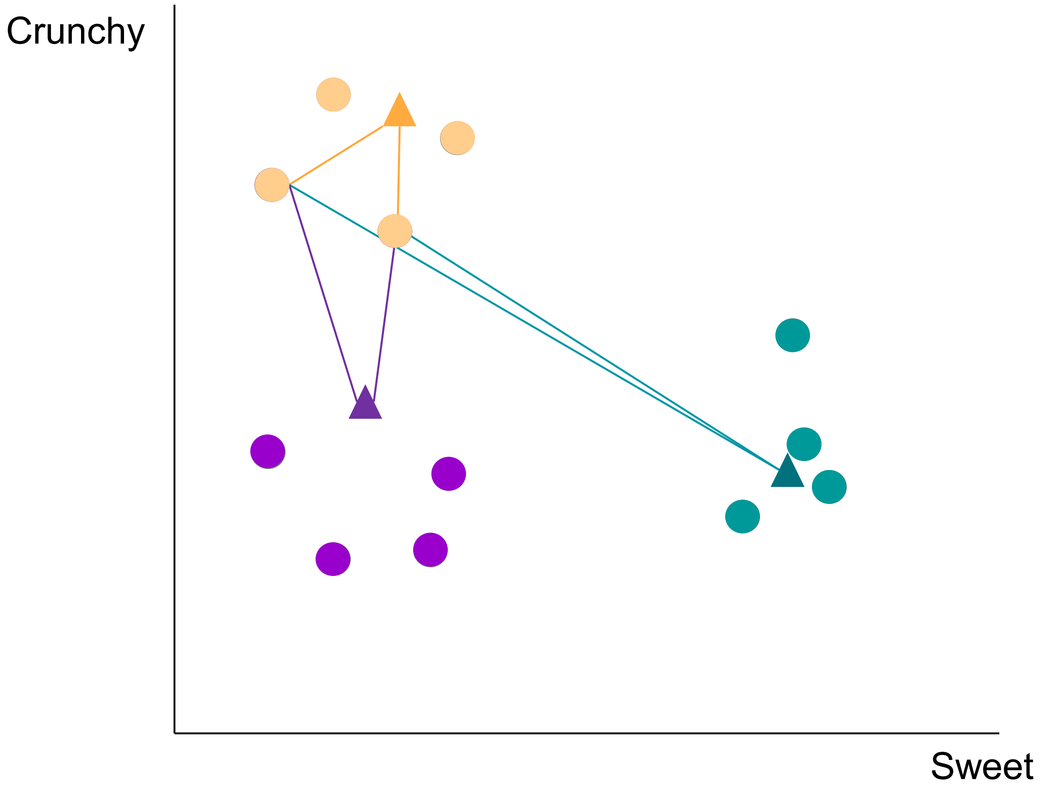
- **Step 1**: Choose the number of clusters you want to identify in your data
- **Step 2**: Randomly select ‘K’ distinct data points
- **Step 3**: Measure the distance btw each data point and the three clusters’ centroids
- **Step 4**: Assign each data point to the nearest cluster
- **Step 5**: Calculate the mean of each cluster (Centroid reassign)
- **Step 6**: Repeat measuring the distance from each data point to clusters’ centroids
- **Step 7**: Going back to step 5 if there is any change at Step 6 OR END clustering

How it works? K-means clustering



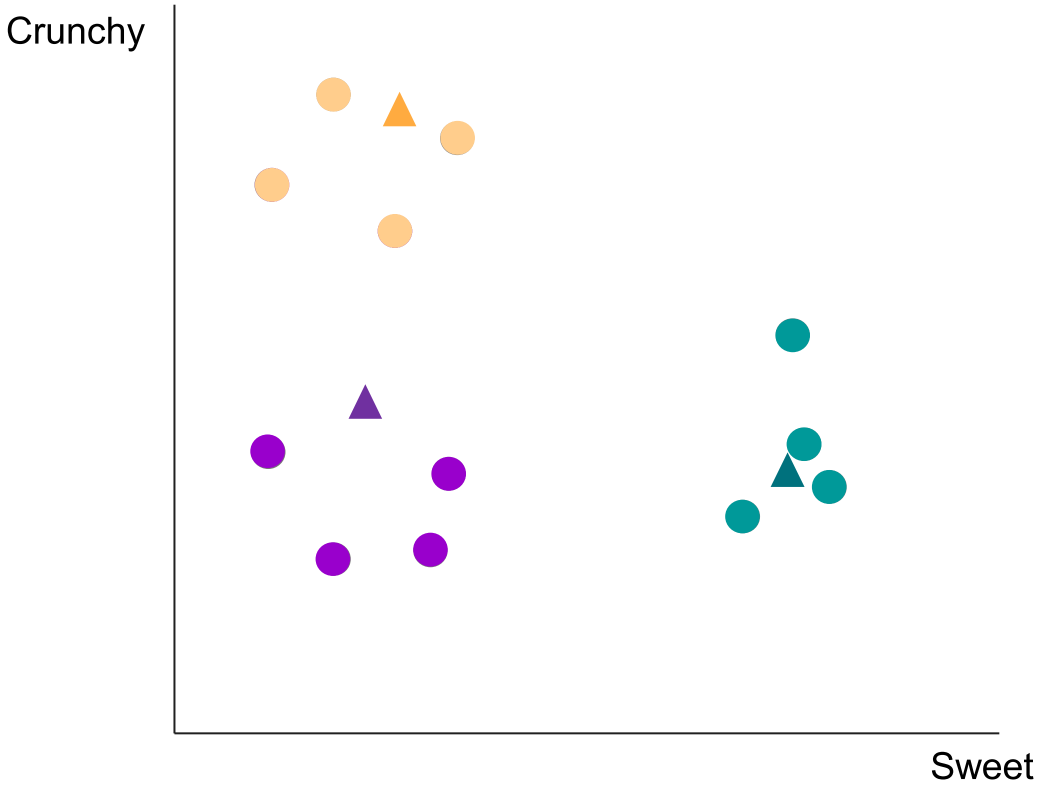
- **Step 1**: Choose the number of clusters you want to identify in your data
- **Step 2**: Randomly select ‘K’ distinct data points
- **Step 3**: Measure the distance btw each data point and the three clusters’ centroids
- **Step 4**: Assign each data point to the nearest cluster
- **Step 5**: Calculate the mean of each cluster (Centroid reassign)
- **Step 6**: Repeat measuring the distance from each data point to clusters’ centroids
- **Step 7**: Going back to step 5 if there is any change at Step 6 OR END clustering

How it works? K-means clustering



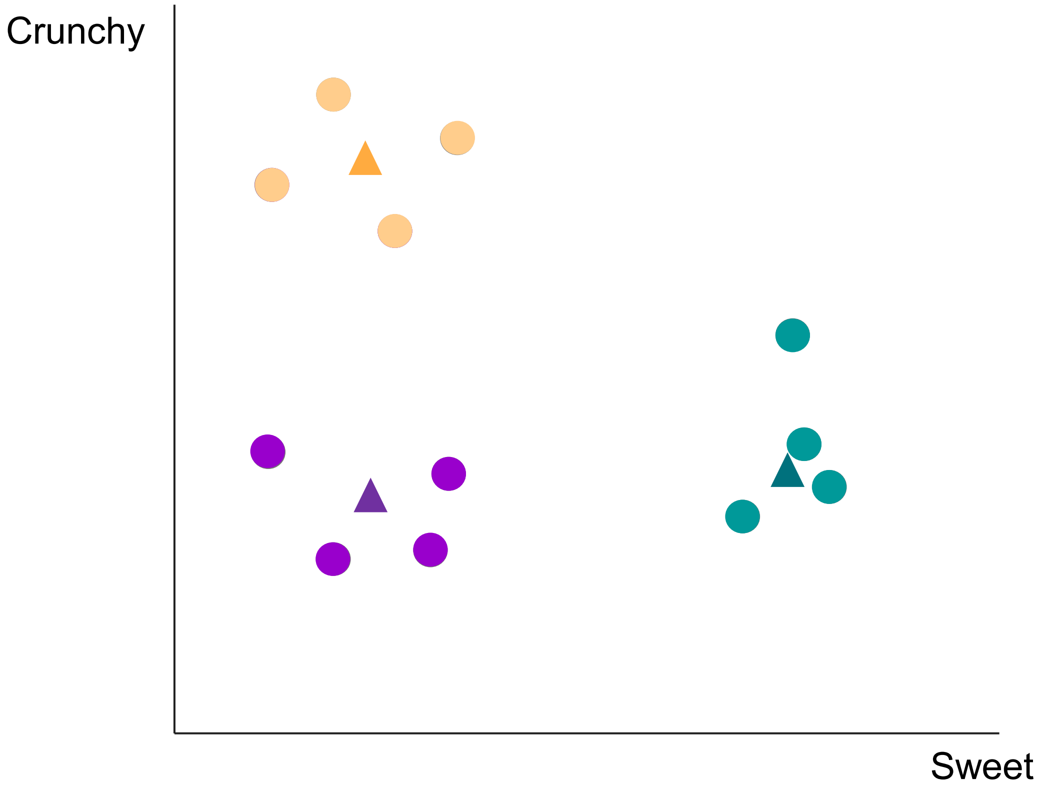
- **Step 1**: Choose the number of clusters you want to identify in your data
- **Step 2**: Randomly select ‘K’ distinct data points
- **Step 3**: Measure the distance btw each data point and the three clusters’ centroids
- **Step 4**: Assign each data point to the nearest cluster
- **Step 5**: Calculate the mean of each cluster (Centroid reassign)
- **Step 6**: Repeat measuring the distance from each data point to clusters’ centroids
- **Step 7**: Going back to step 5 if there is any change at Step 6 OR END clustering

How it works? K-means clustering



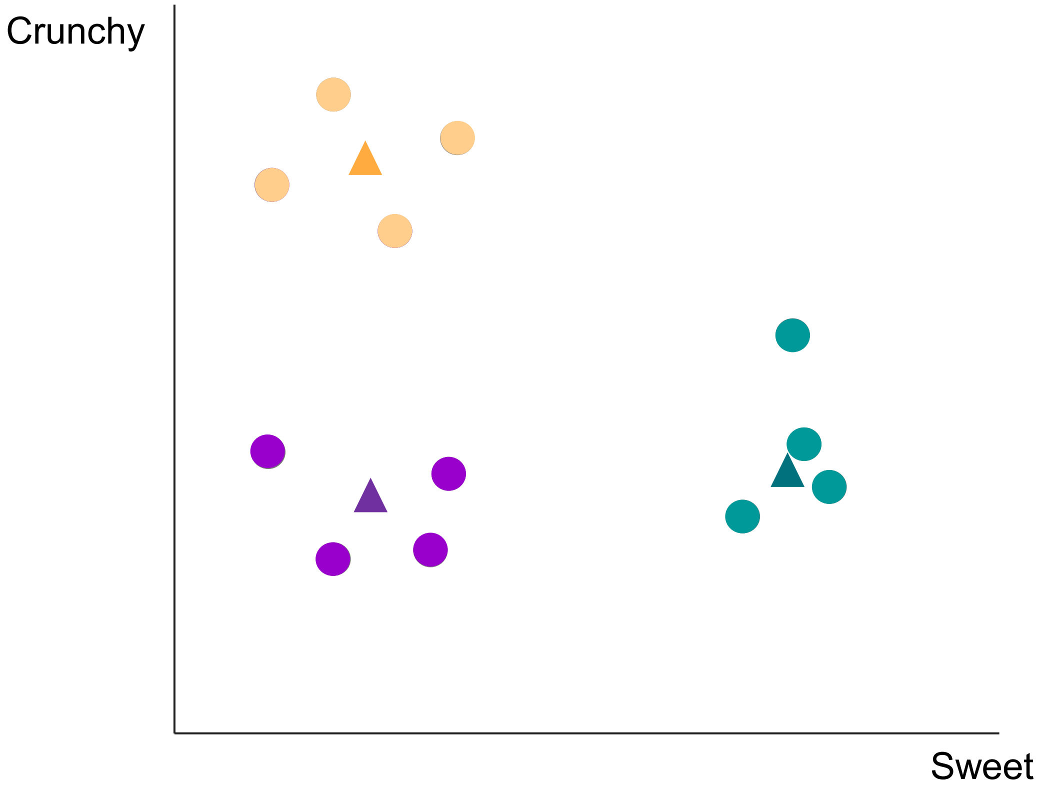
- **Step 1**: Choose the number of clusters you want to identify in your data
- **Step 2**: Randomly select ‘K’ distinct data points
- **Step 3**: Measure the distance btw each data point and the three clusters’ centroids
- **Step 4**: Assign each data point to the nearest cluster
- **Step 5**: Calculate the mean of each cluster (Centroid reassign)
- **Step 6**: Repeat measuring the distance from each data point to clusters’ centroids
- **Step 7**: Going back to step 5 if there is any change at Step 6 OR END clustering

How it works? K-means clustering



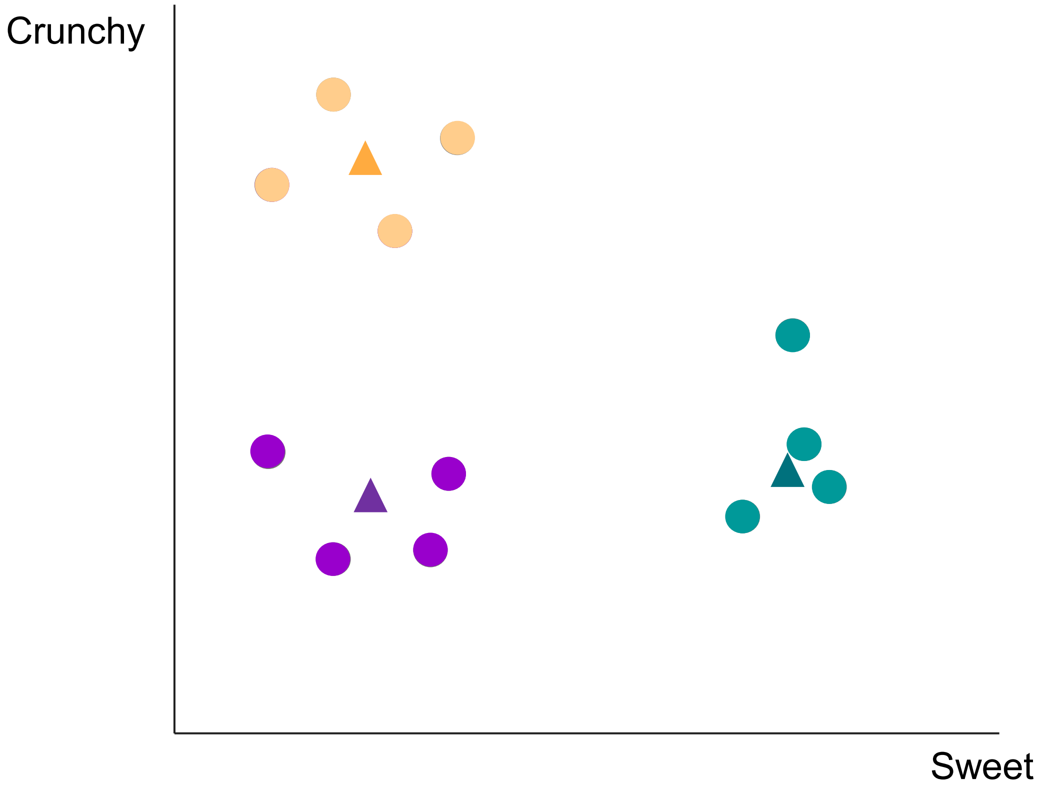
- **Step 1**: Choose the number of clusters you want to identify in your data
- **Step 2**: Randomly select 'K' distinct data points
- **Step 3**: Measure the distance btw each data point and the three clusters' centroids
- **Step 4**: Assign each data point to the nearest cluster
- **Step 5**: Calculate the mean of each cluster (Centroid reassign)
- **Step 6**: Repeat measuring the distance from each data point to clusters' centroids
- **Step 7**: Going back to step 5 if there is any change at Step 6 OR END clustering

How it works? K-means clustering



- **Step 1**: Choose the number of clusters you want to identify in your data
- **Step 2**: Randomly select ‘K’ distinct data points
- **Step 3**: Measure the distance btw each data point and the three clusters’ centroids
- **Step 4**: Assign each data point to the nearest cluster
- **Step 5**: Calculate the mean of each cluster (Centroid reassign)
- **Step 6**: Repeat measuring the distance from each data point to clusters’ centroids
- **Step 7**: Going back to step 5 if there is any change at Step 6 OR END clustering

How it works? K-means clustering



- **Step 1**: Choose the number of clusters you want to identify in your data
- **Step 2**: Randomly select ‘K’ distinct data points
- **Step 3**: Measure the distance btw each data point and the three clusters’ centroids
- **Step 4**: Assign each data point to the nearest cluster
- **Step 5**: Calculate the mean of each cluster (Centroid reassign)
- **Step 6**: Repeat measuring the distance from each data point to clusters’ centroids
- **Step 7**: Going back to step 5 if there is any change at Step 6 OR END clustering

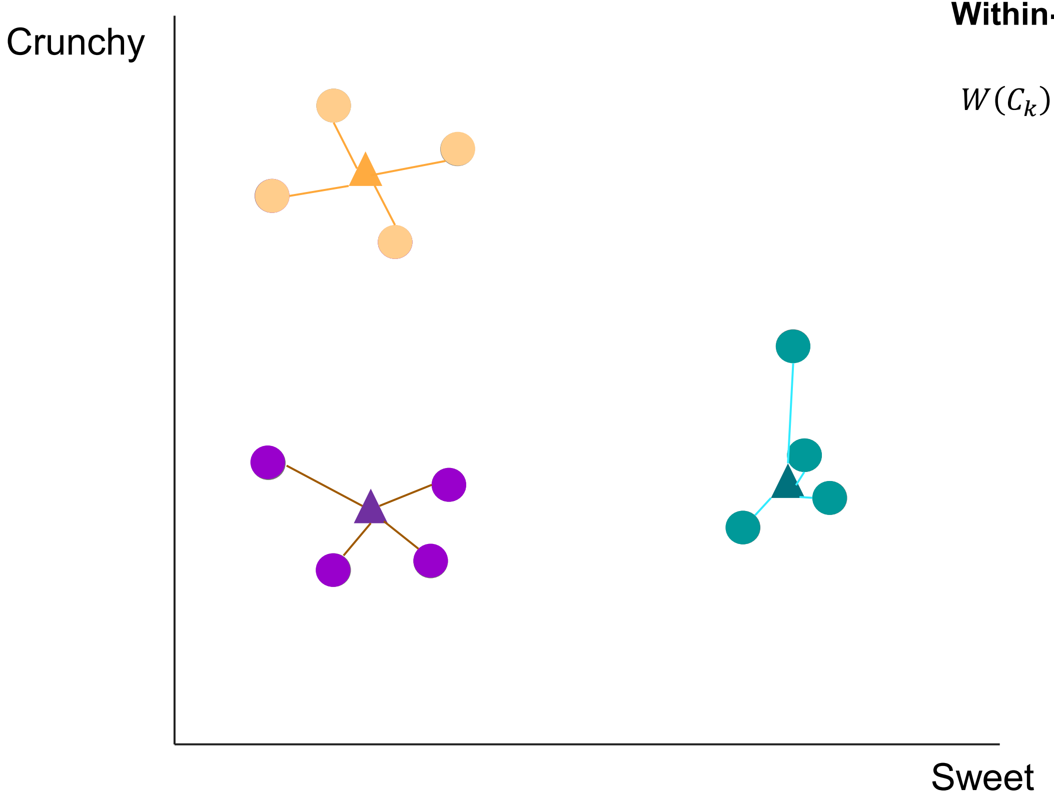
K number of clusters (We can adjust)

K-means clustering

Because we set each cluster’s centroid. In other word, we have K means of clusters

How to choose the number of clusters (k) ?

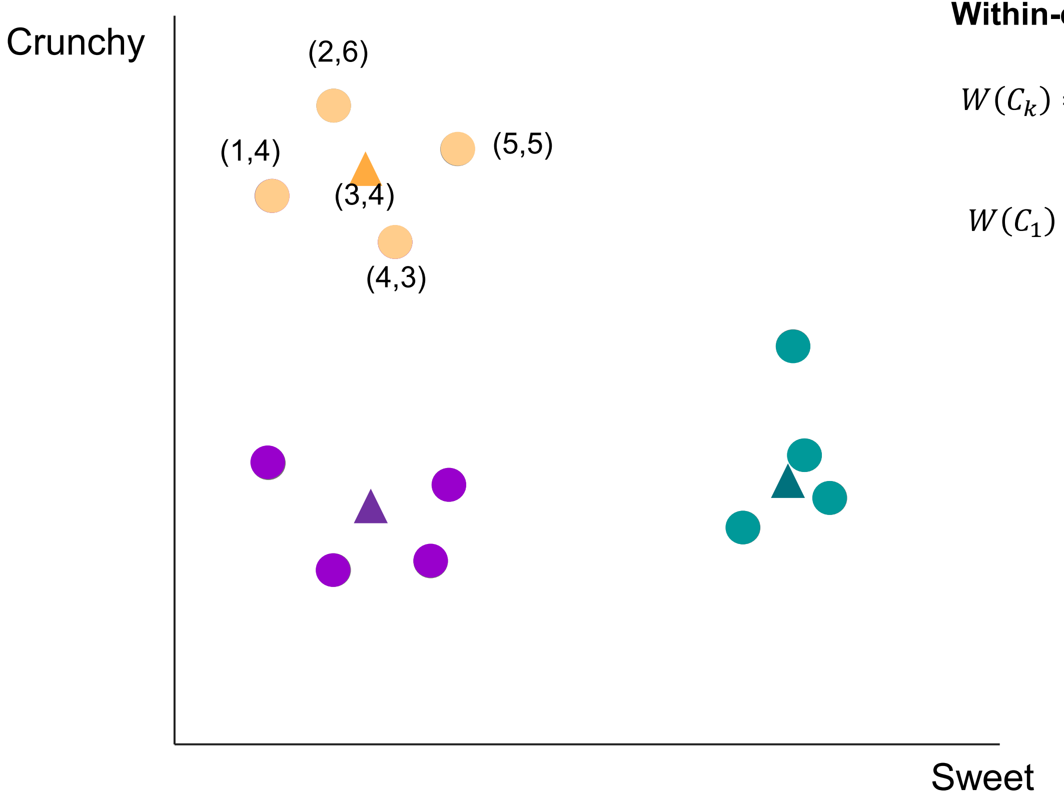
How to choose K?



Within-cluster variation (Intra-cluster variation)

$$W(C_k) = \sum_{x_i \in C_k} (x_i - \mu_k)^2$$

How to choose K?



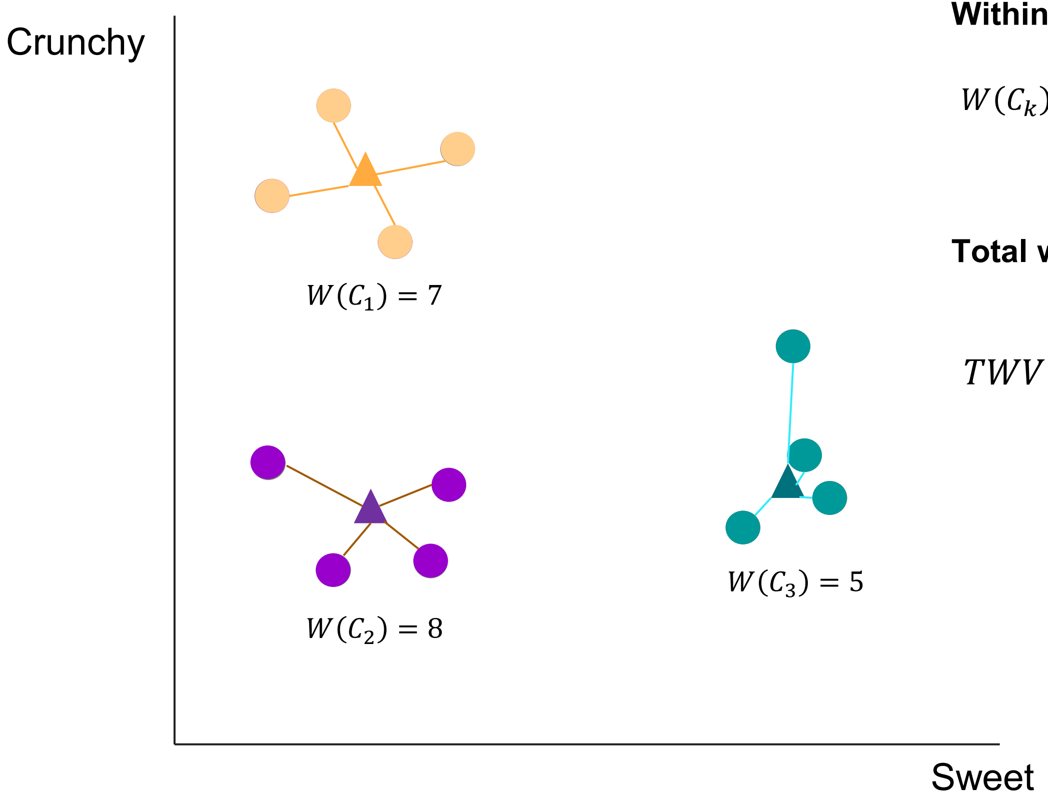
Within-cluster variation (Intra-cluster variation)

$$W(C_k) = \sum_{x_i \in C_k} (x_i - \mu_k)^2$$

$$W(C_1) = (x_1 - \mu_1)^2 + (x_2 - \mu_1)^2 + (x_3 - \mu_1)^2 + (x_4 - \mu_1)^2$$

$$W(C_1) = \left(\sqrt{(2-3)^2 + (6-4)^2}\right)^2 + \left(\sqrt{(1-3)^2 + (4-4)^2}\right)^2 + \left(\sqrt{(5-3)^2 + (5-4)^2}\right)^2 + \left(\sqrt{(4-3)^2 + (3-4)^2}\right)^2$$

How to choose K?



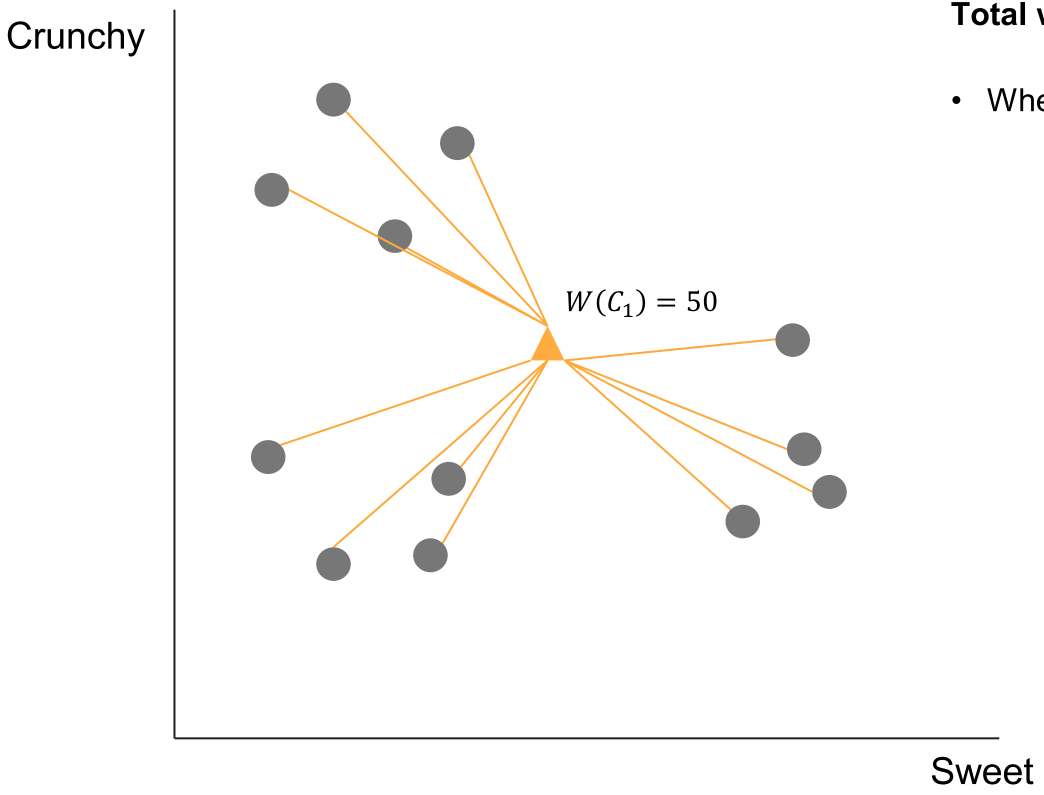
Within-cluster variation (Intra-cluster variation)

$$W(C_k) = \sum_{x_i \in C_k} (x_i - \mu_k)^2$$

Total within-cluster variation

$$TWV(C_3) = \sum_{k=1}^{3} W(C_k) = \sum_{k=1}^{3} \sum_{x_i \in C_k} (x_i - \mu_k)^2 = 20$$

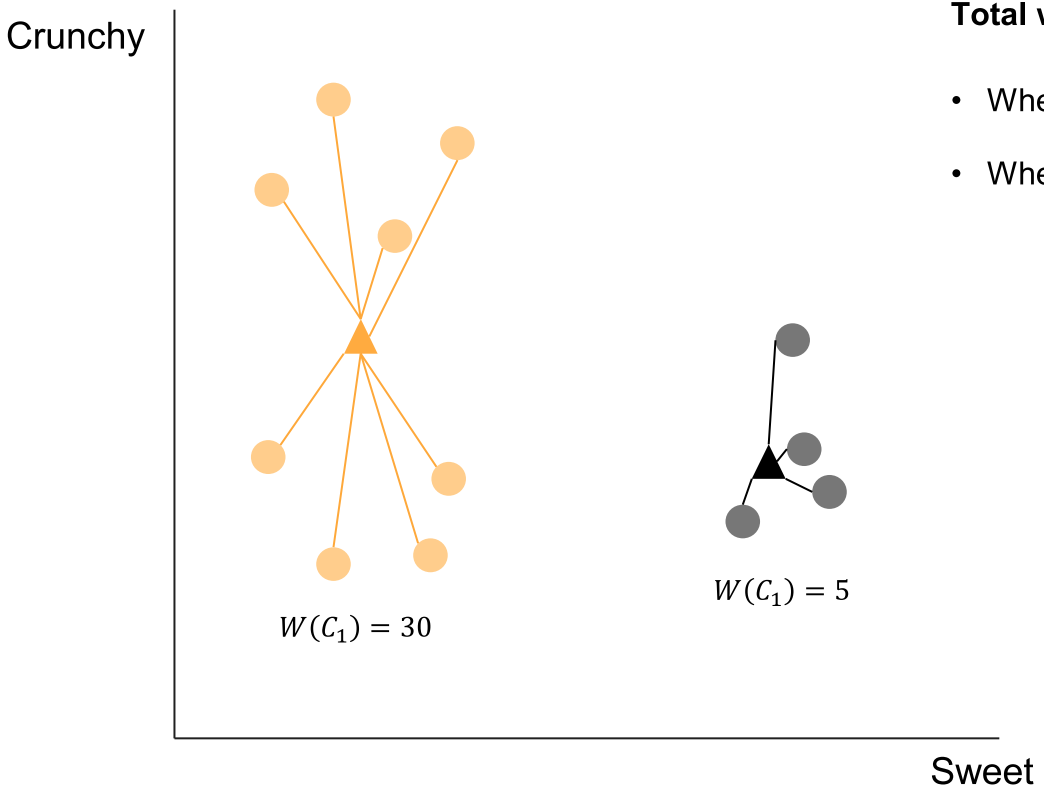
How to choose K?



Total within-cluster variation

- When K=1: TWC=50

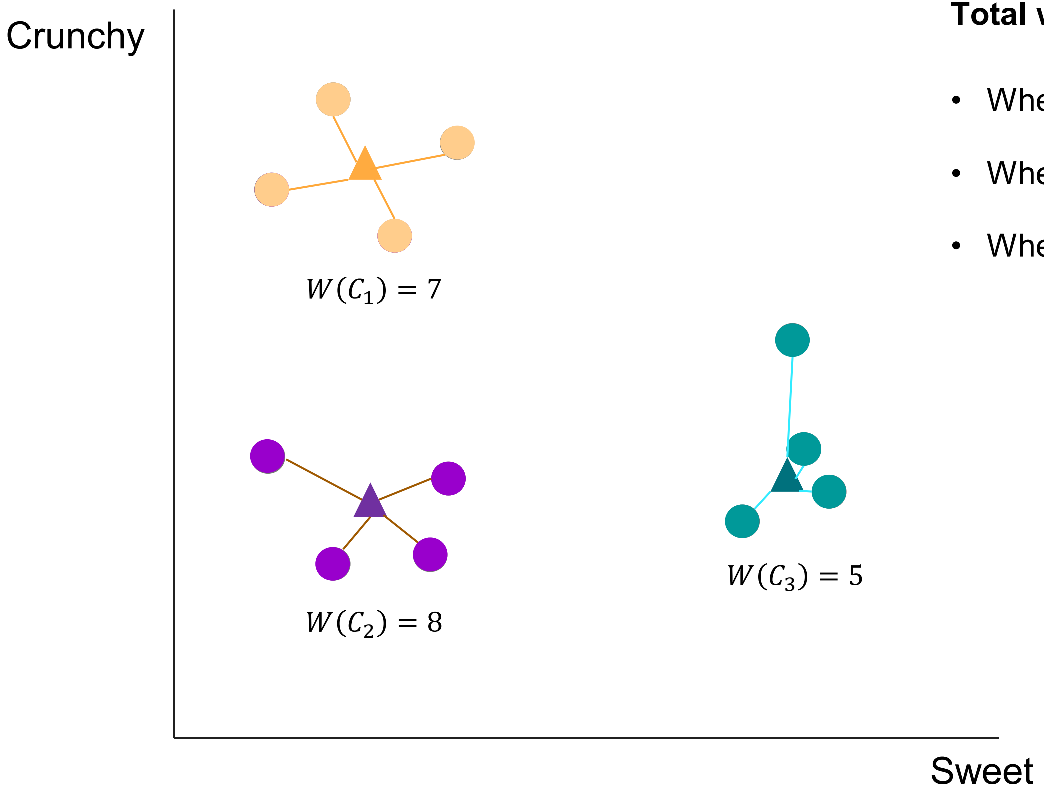
How to choose K?



Total within-cluster variation

- When K=1: TWC=50
- When K=2: TWC=35

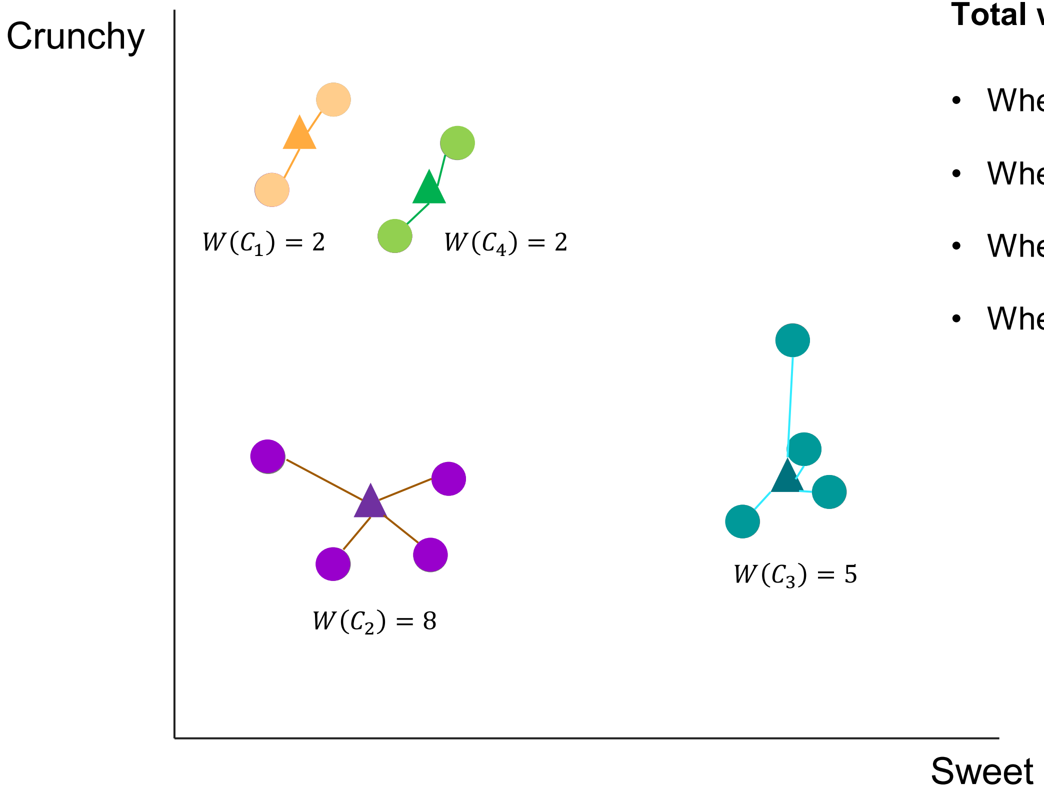
How to choose K?



Total within-cluster variation

- When K=1: TWC=50
- When K=2: TWC=35
- When K=3: TWC=20

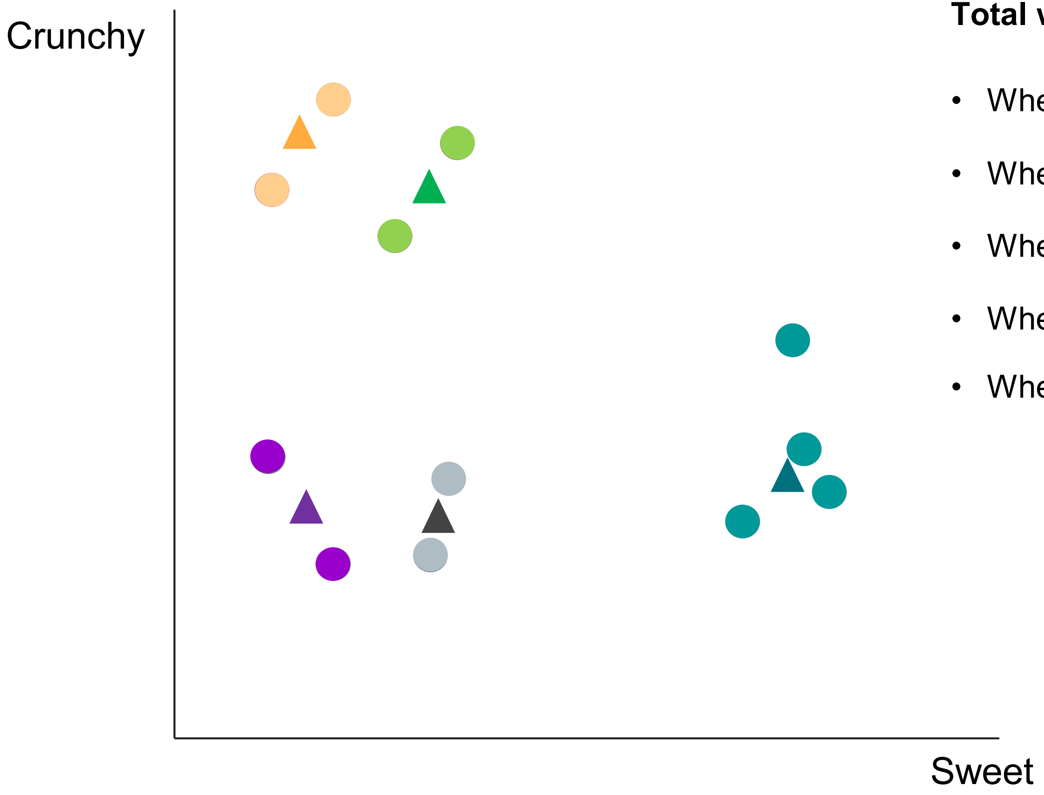
How to choose K?



Total within-cluster variation

- When K=1: TWC=50
- When K=2: TWC=35
- When K=3: TWC=20
- When K=4: TWC=17

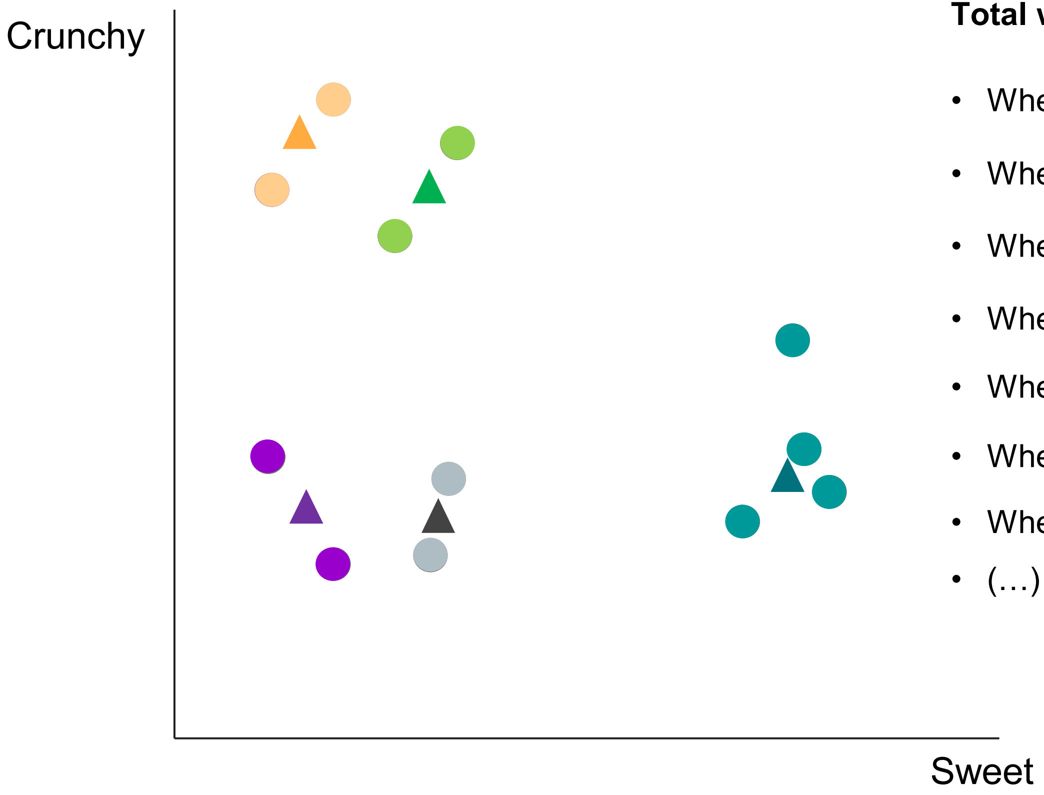
How to choose K?



Total within-cluster variation

- When K=1: TWC=50
- When K=2: TWC=35
- When K=3: TWC=20
- When K=4: TWC=17
- When K=5: TWC=15

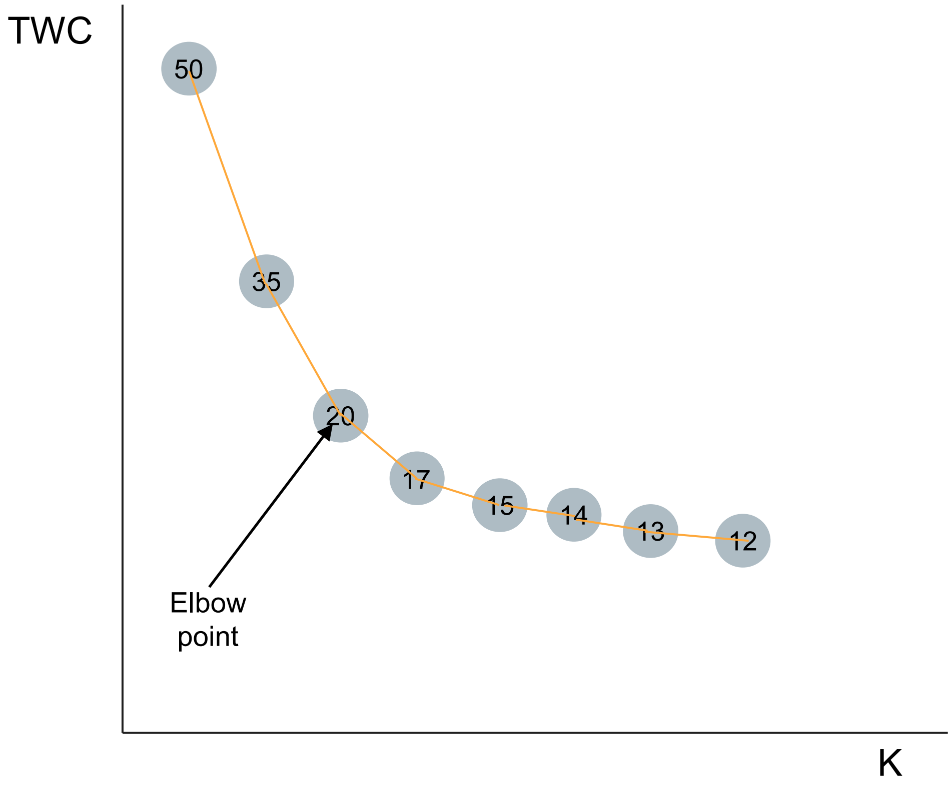
How to choose K?



Total within-cluster variation

- When K=1: TWC=50
- When K=2: TWC=35
- When K=3: TWC=20
- When K=4: TWC=17
- When K=5: TWC=15
- When K=6: TWC=14
- When K=7: TWC=13.5
- (…)

How to choose K?



Total within-cluster variation

- When K=1: TWC=50
- When K=2: TWC=35
- When K=3: TWC=20
- When K=4: TWC=17
- When K=5: TWC=15
- When K=6: TWC=14
- When K=7: TWC=13.5
- (…)

Elbow method is one of the 30 or more methods to find appropriate k in k-means clustering