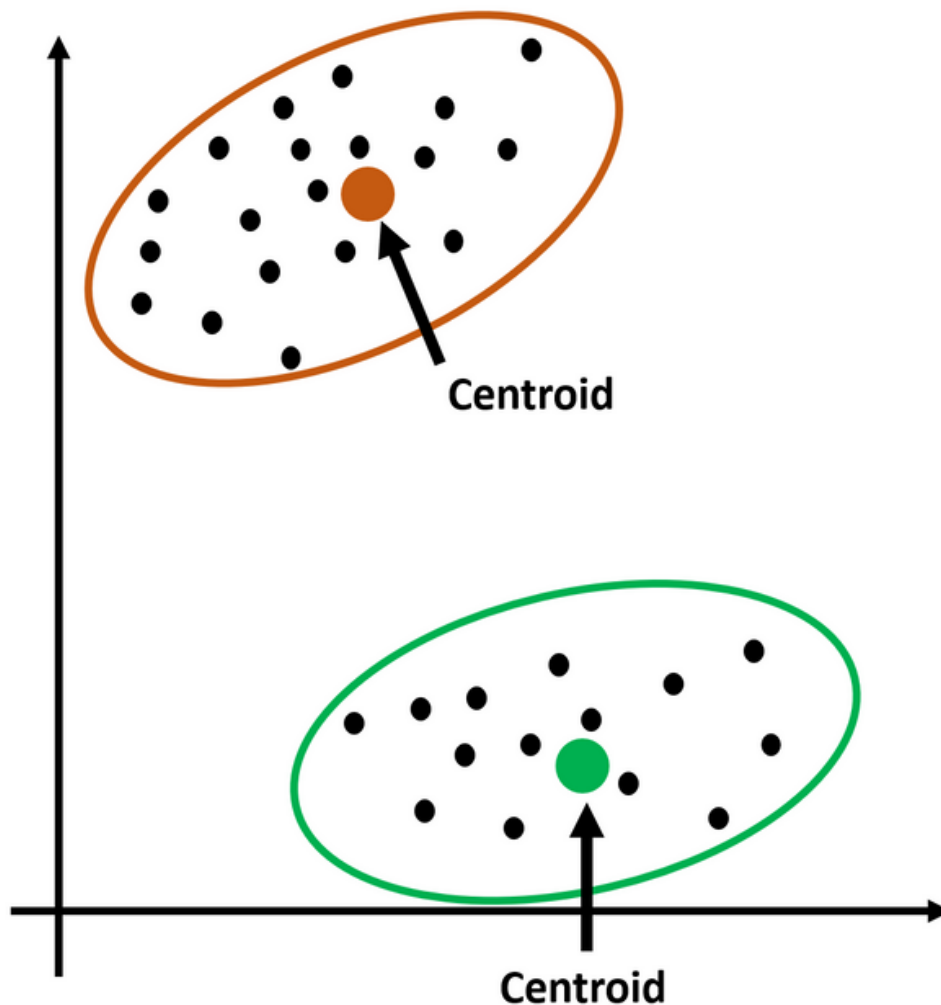


# K-MEANS CLUSTERING



**A BRIEF, INTUITIVE INTRODUCTION**

---

## Introducing Cluster Analysis

"**Cluster analysis** groups data objects based only on information found in the data that describes the objects and their relationships.

The goal is that the objects within a group be similar (or related) to one another and different from (or unrelated to) the objects in other groups.

The greater the similarity (or homogeneity) within a group and the greater the difference between groups, the better or more distinct the clustering"

Source: Introduction to Data Mining by Pang-Ning, Michael Steinbach, and Vipin Kumar, first edition May 2, 2005.

Because cluster analysis has no external information about groups (i.e., **labels**), it belongs to a form of machine learning known as **unsupervised learning**.

Because so much data is unlabeled, cluster analysis is a widely used tool to discover structure in data and produce new insights.

BTW - The words "groups" and "clusters" mean the same thing.

---

## Introducing K-Means

The **k-means algorithm** is a prototype-based, complete partitioning clustering technique.

Whoa! That was a mouthful. Let's break that down:

1. K-means uses cluster centers (**centroids**) based on the average data of cluster members.
2. The clusters produced by k-means are spherical.
3. Every data point (**observation**) is assigned to a cluster - even outliers.
4. Every observation will be assigned to a single cluster.

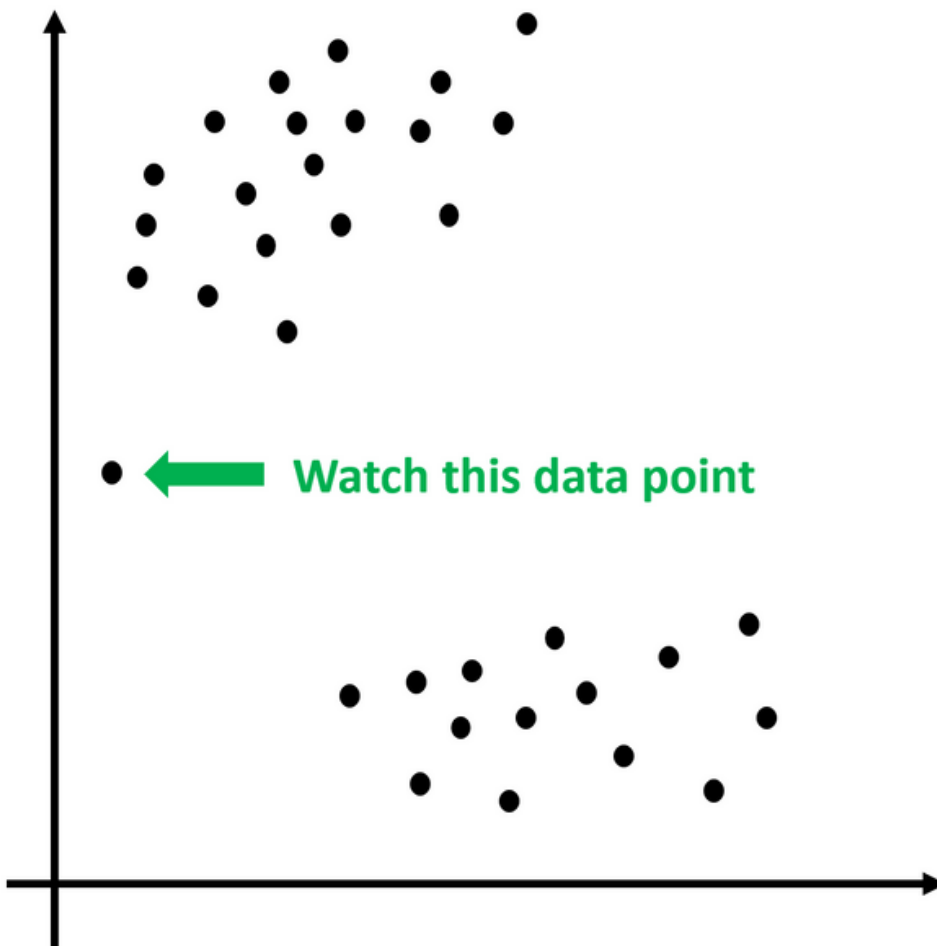
K-means is one of the most popular of all clustering techniques.

A big part of k-means' popularity is the algorithm's simplicity - it is easy to understand how k-means works intuitively.

## A Contrived Example

Here's the k-means algorithm:

1. Select  $k$  points as the initial cluster centroids.
2. Form  $k$  clusters by assigning each observation to its closest centroid.
3. Recompute each centroid as the average of all cluster members.
4. Stop if no centroid changes, otherwise, go to 2.

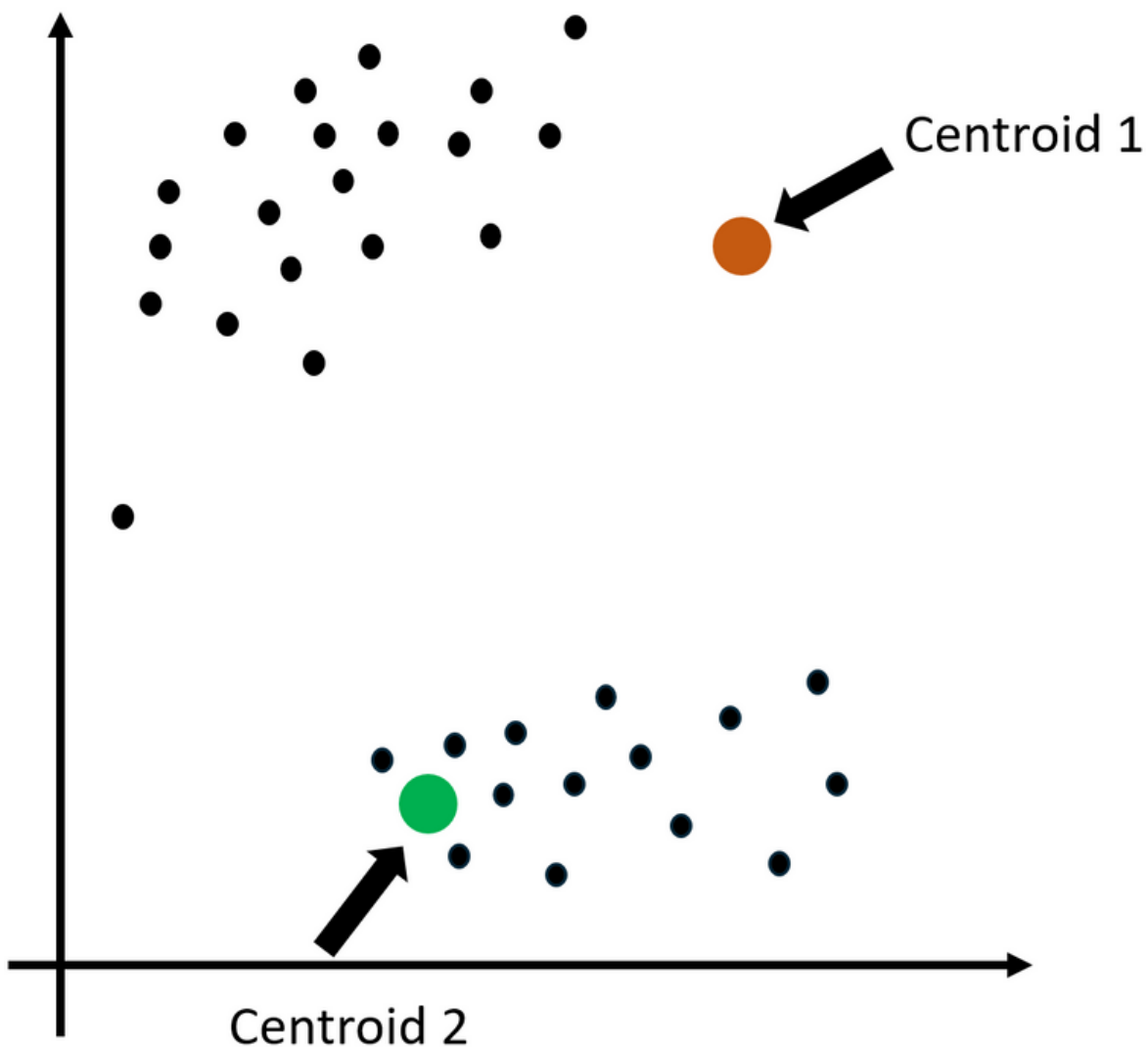


This is the dataset for the contrived example.

For this dataset, assume  $k = 2$ .

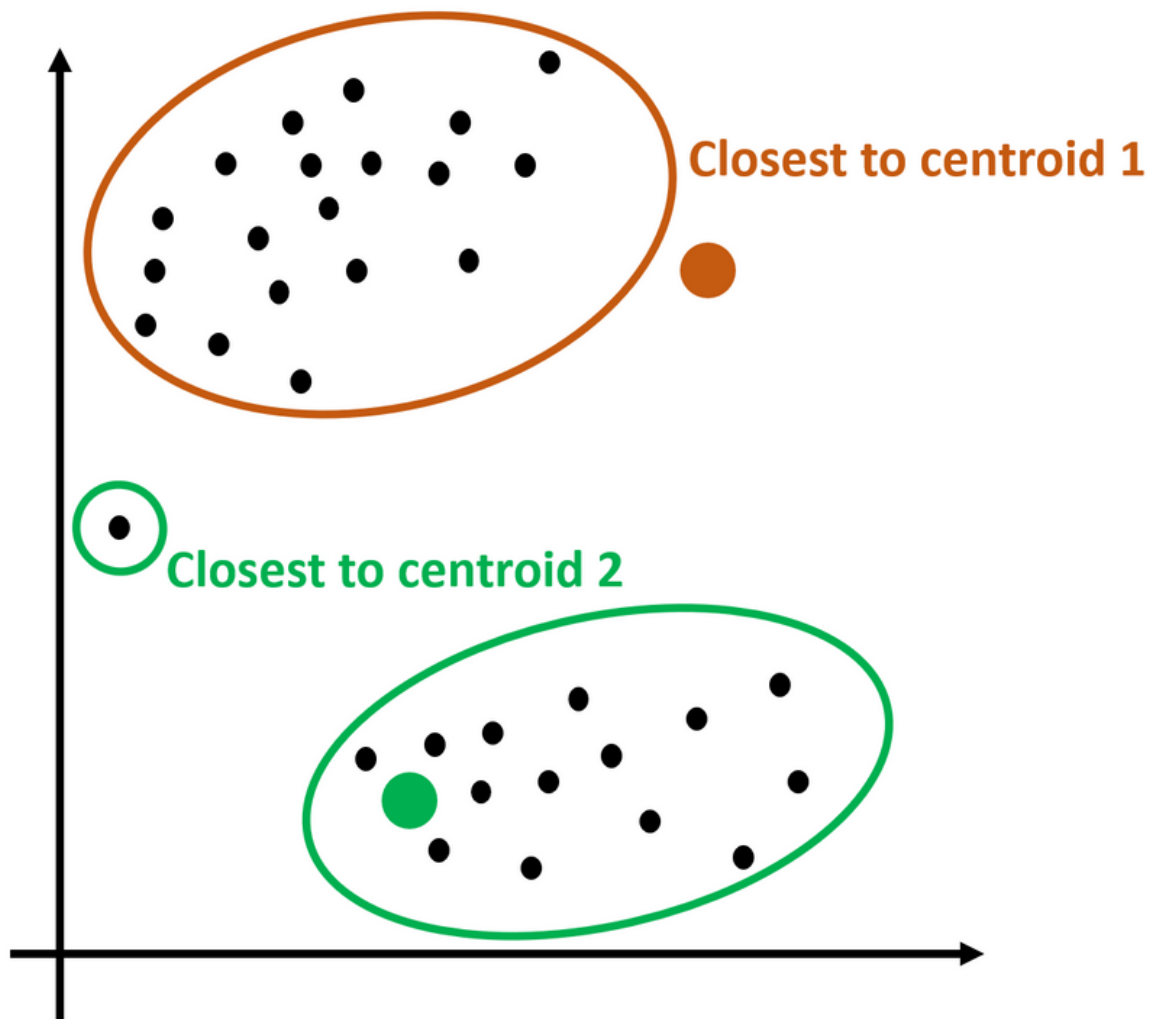
K-means needs someplace to start working with the data.

Once the number of clusters is chosen, k-means “throws out” some random cluster centers (i.e., **centroids**) as a starting point.

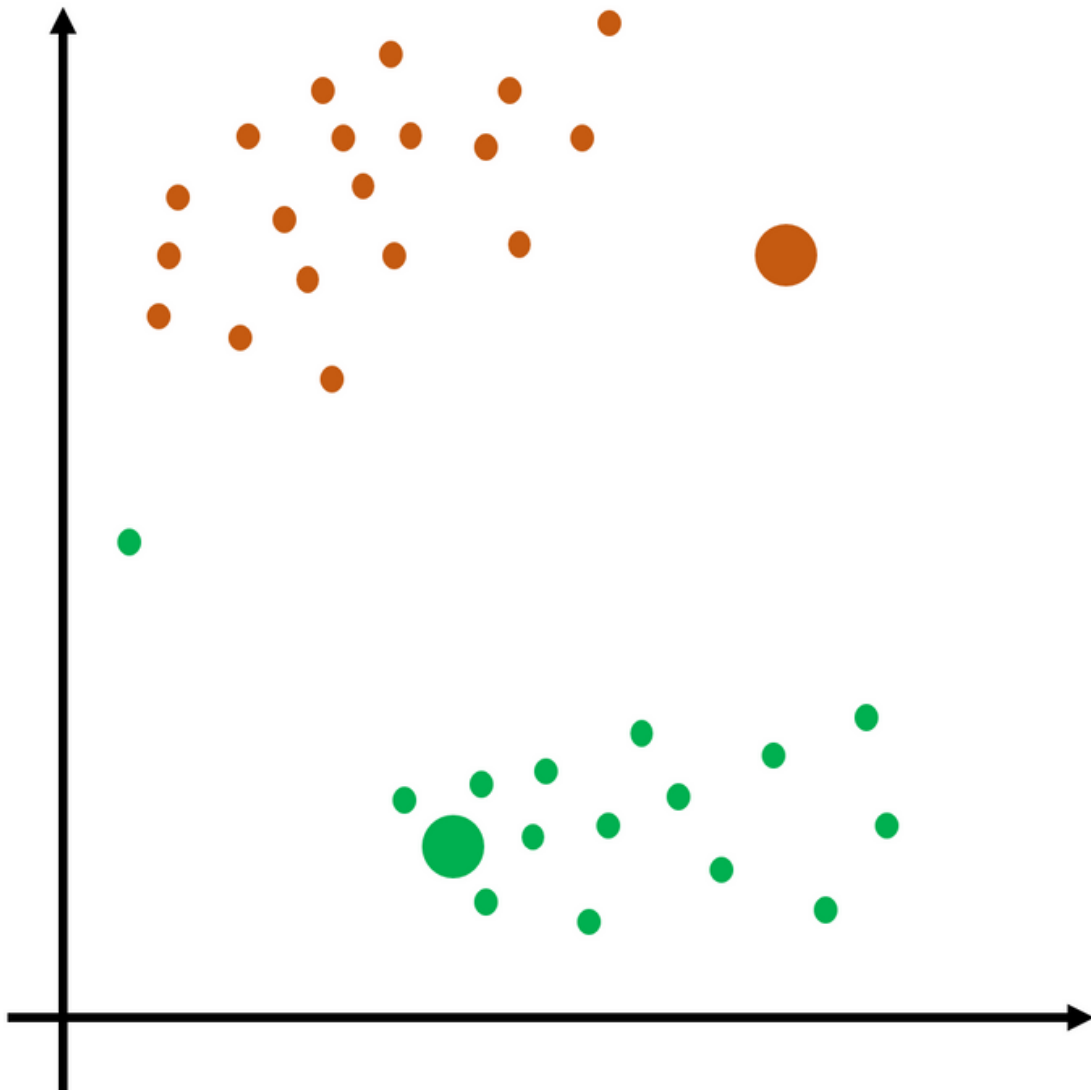


K-means then looks at each “point” (e.g., customer) to cluster in turn:

1. What’s the distance from the point to centroid 1?
2. What’s the distance from the point to centroid 2?

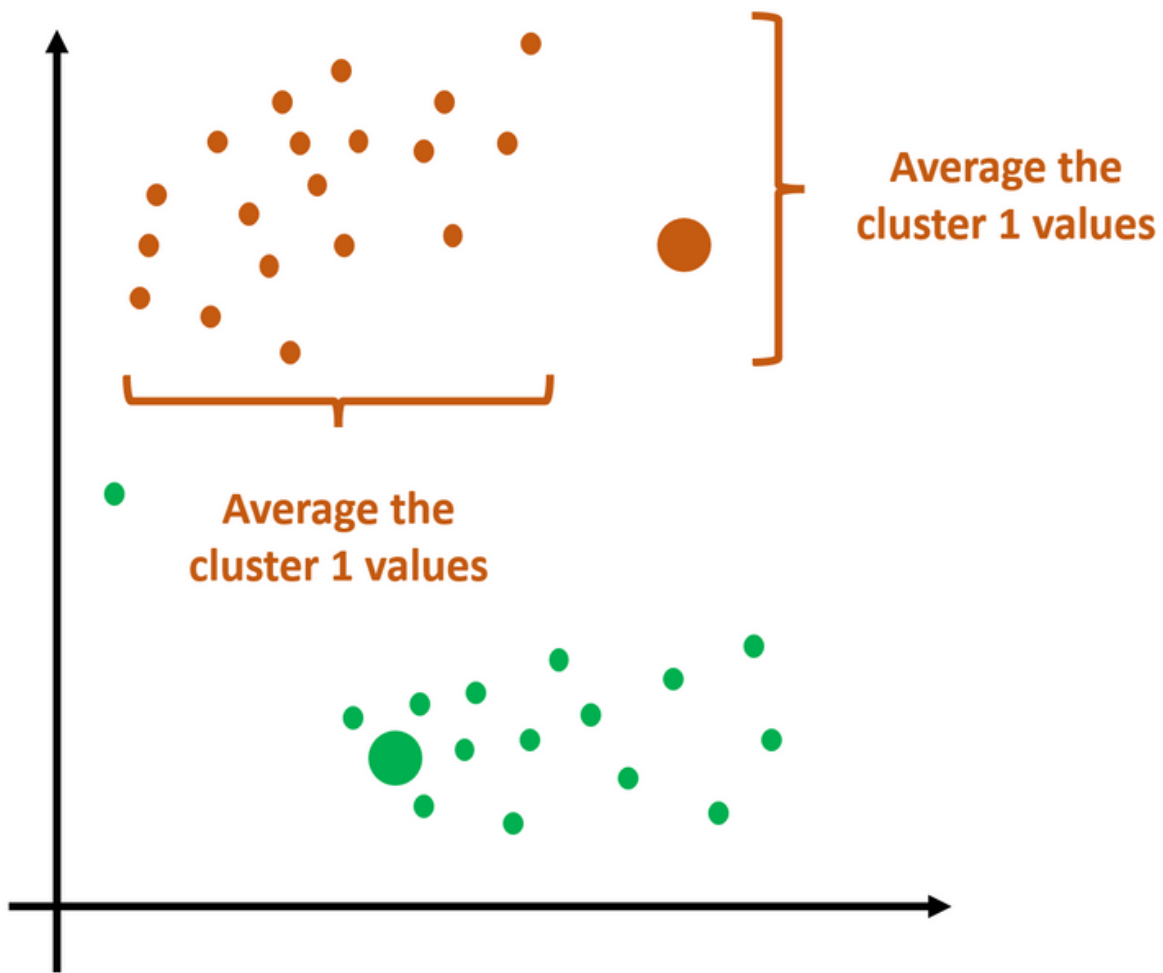


Based on proximity, k-means then assigns each data point to a cluster.



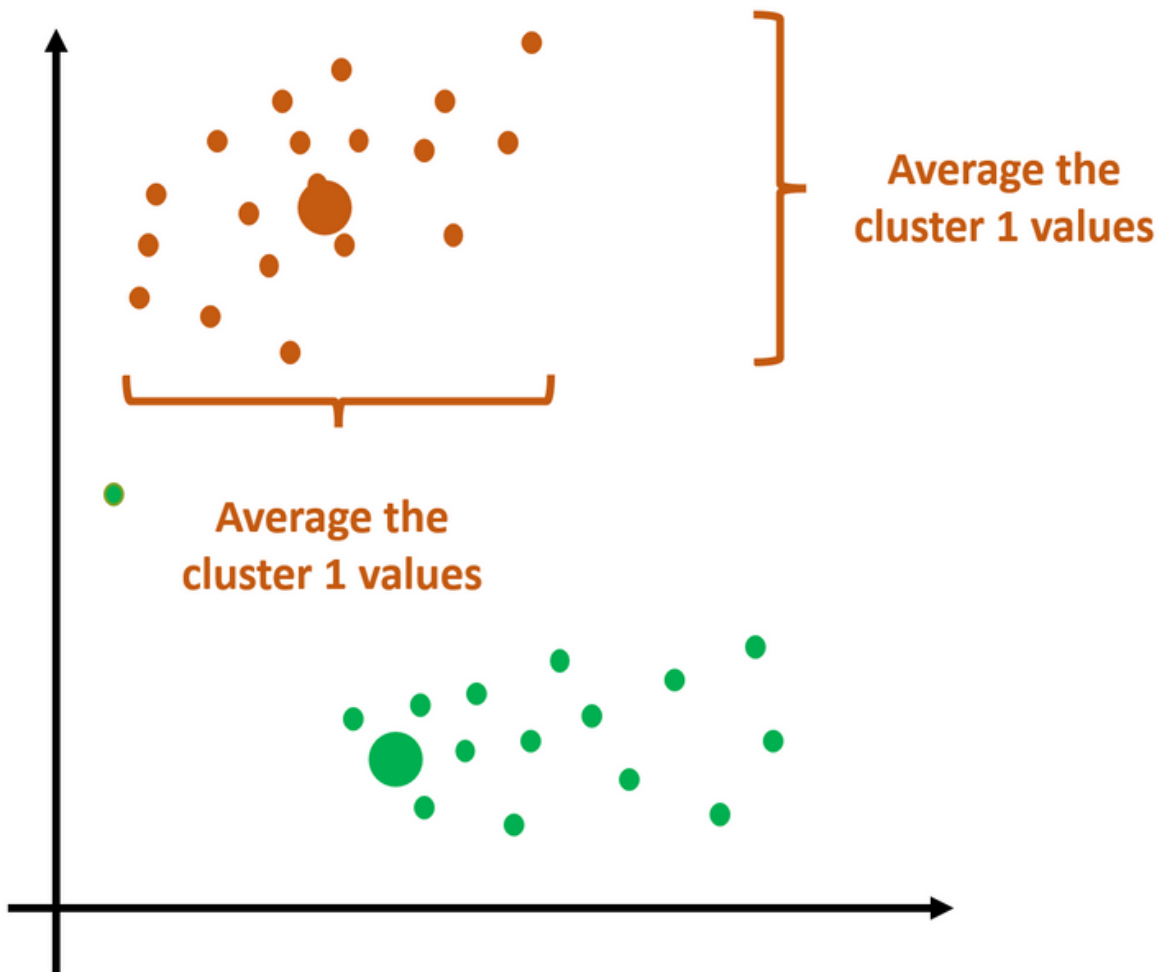
Moving the centroids is where the name “k-means” comes from.

Centroids are moved based on the average values of the data points within the cluster.

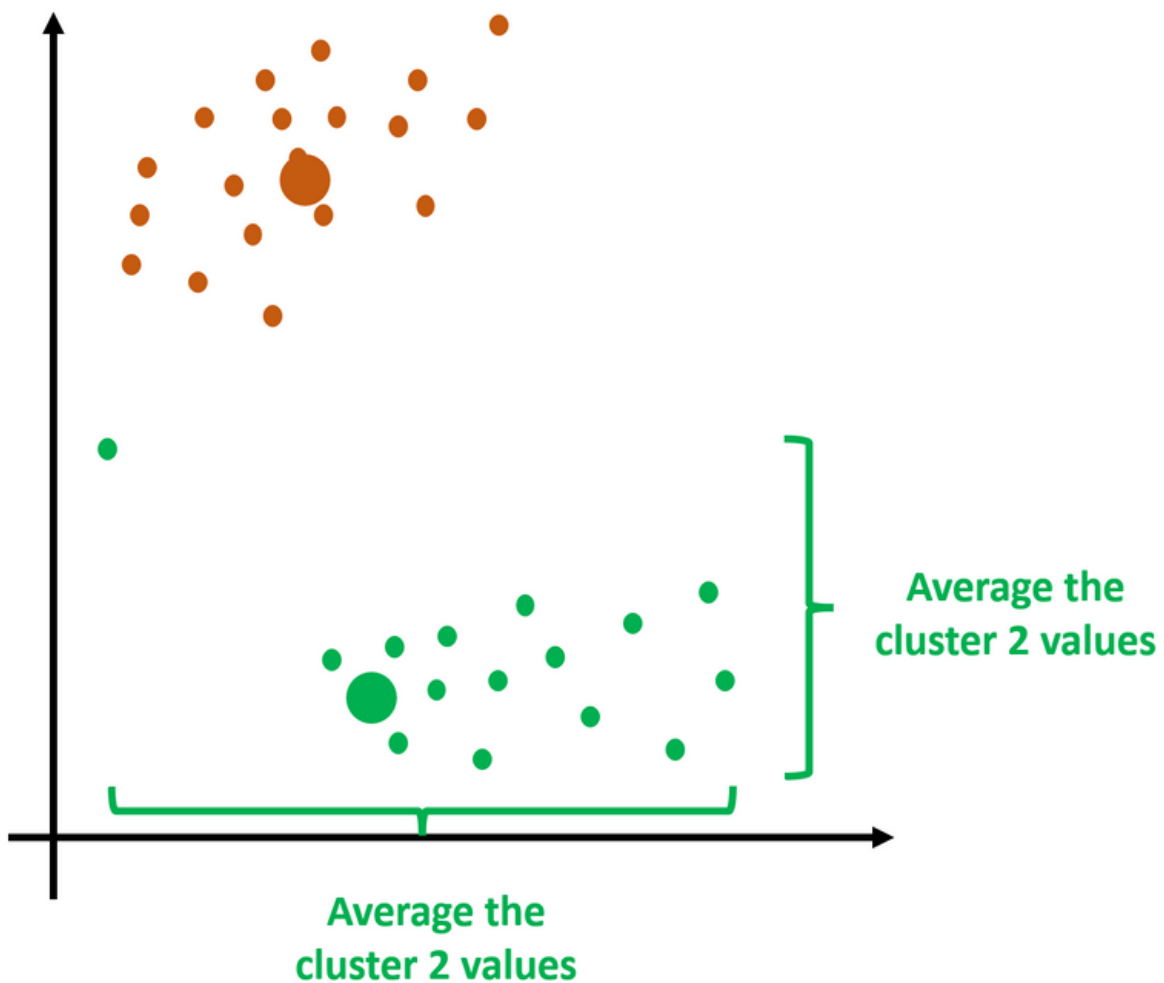




The algorithm moves the centroid (i.e., cluster center).

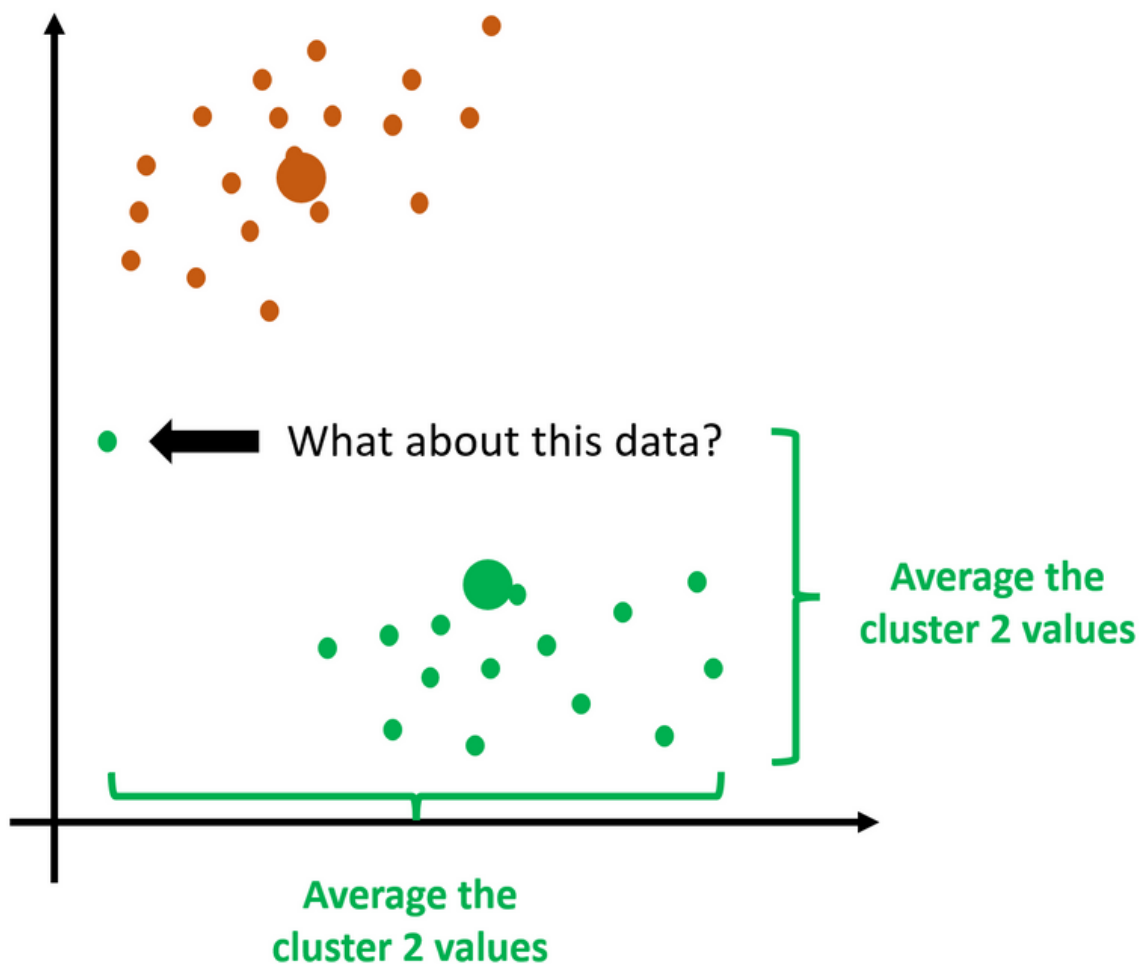


The moving process is repeated for each centroid.



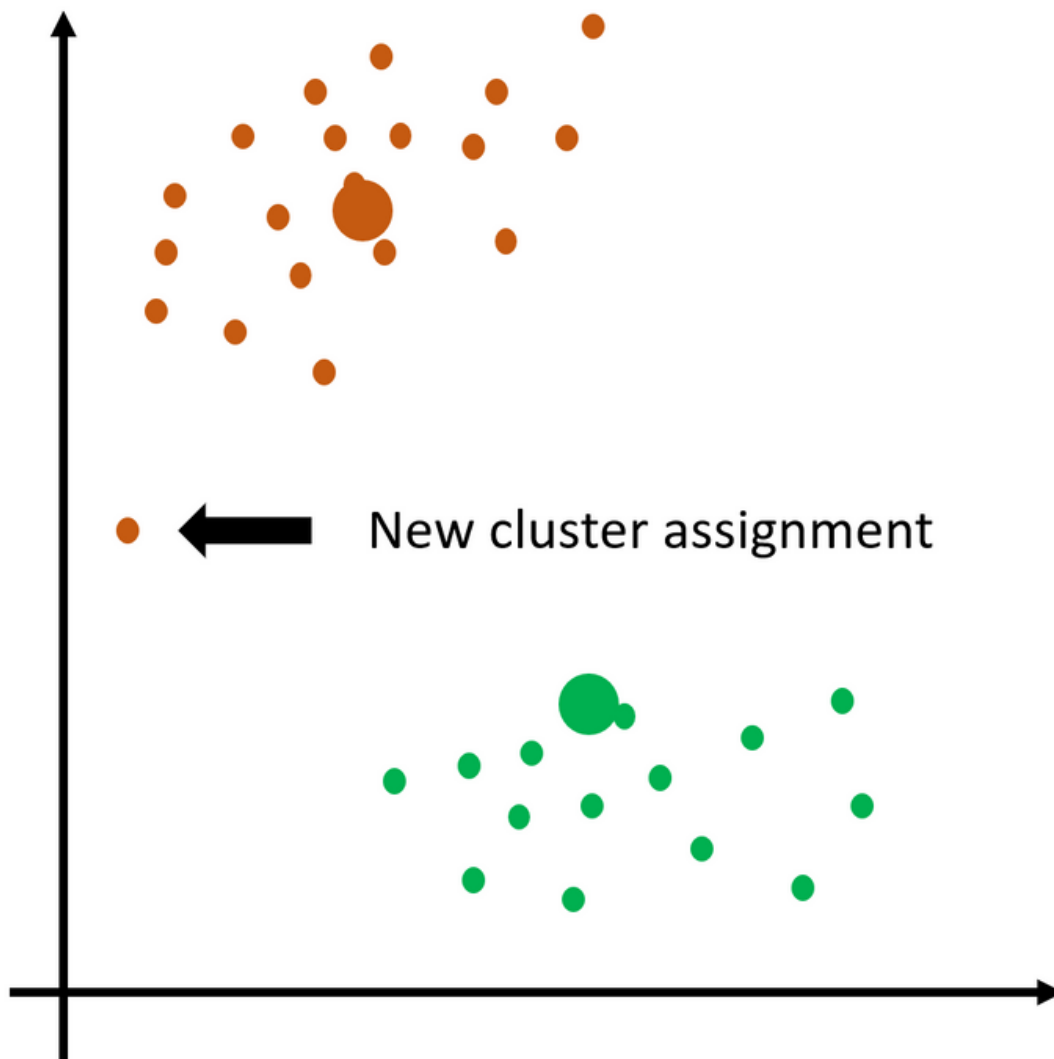
K-means is an iterative process (i.e., algorithm).

The process is repeated until a stopping condition is reached.

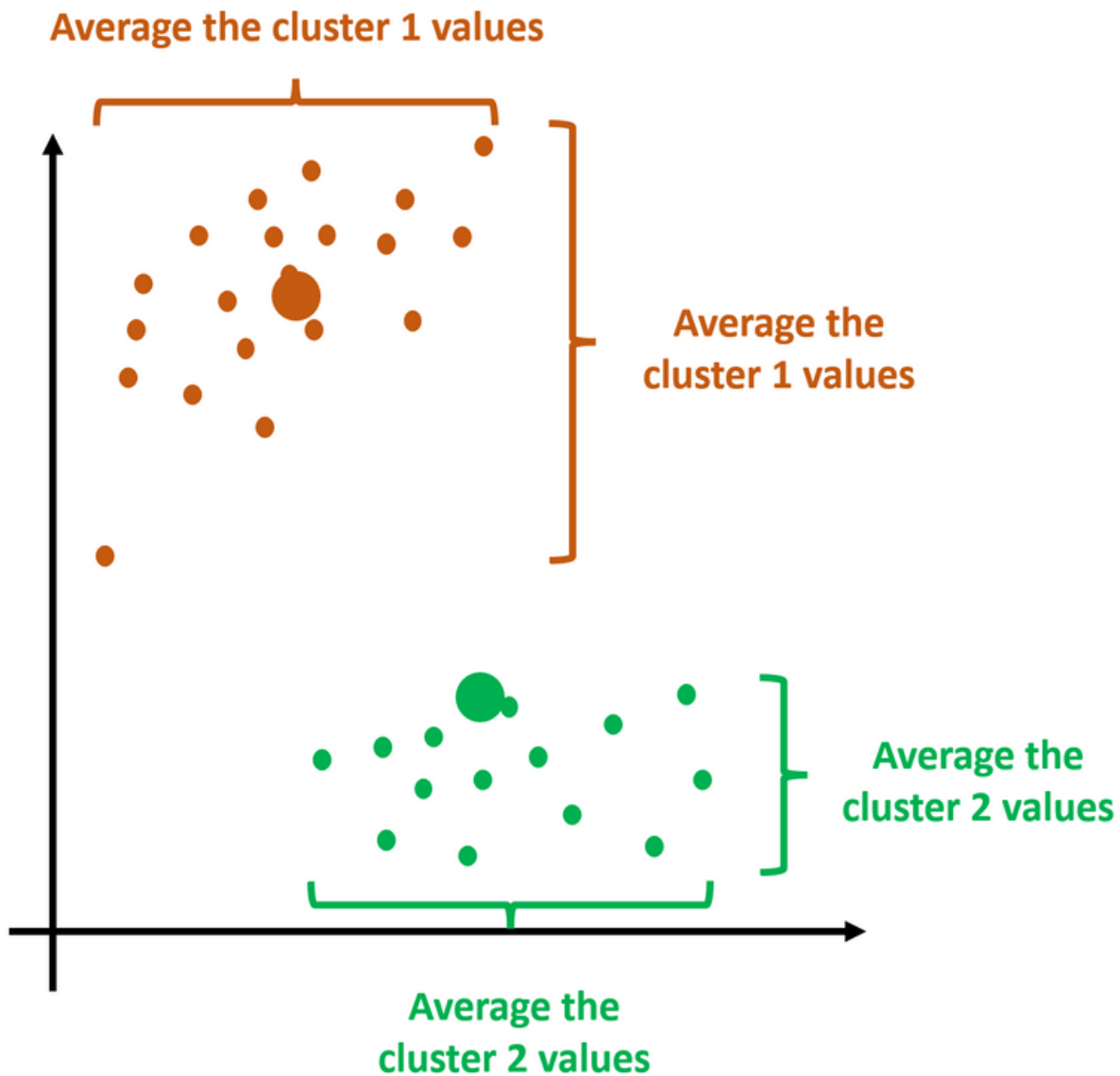


K-means looks at each “point” (i.e., document) to cluster in turn:

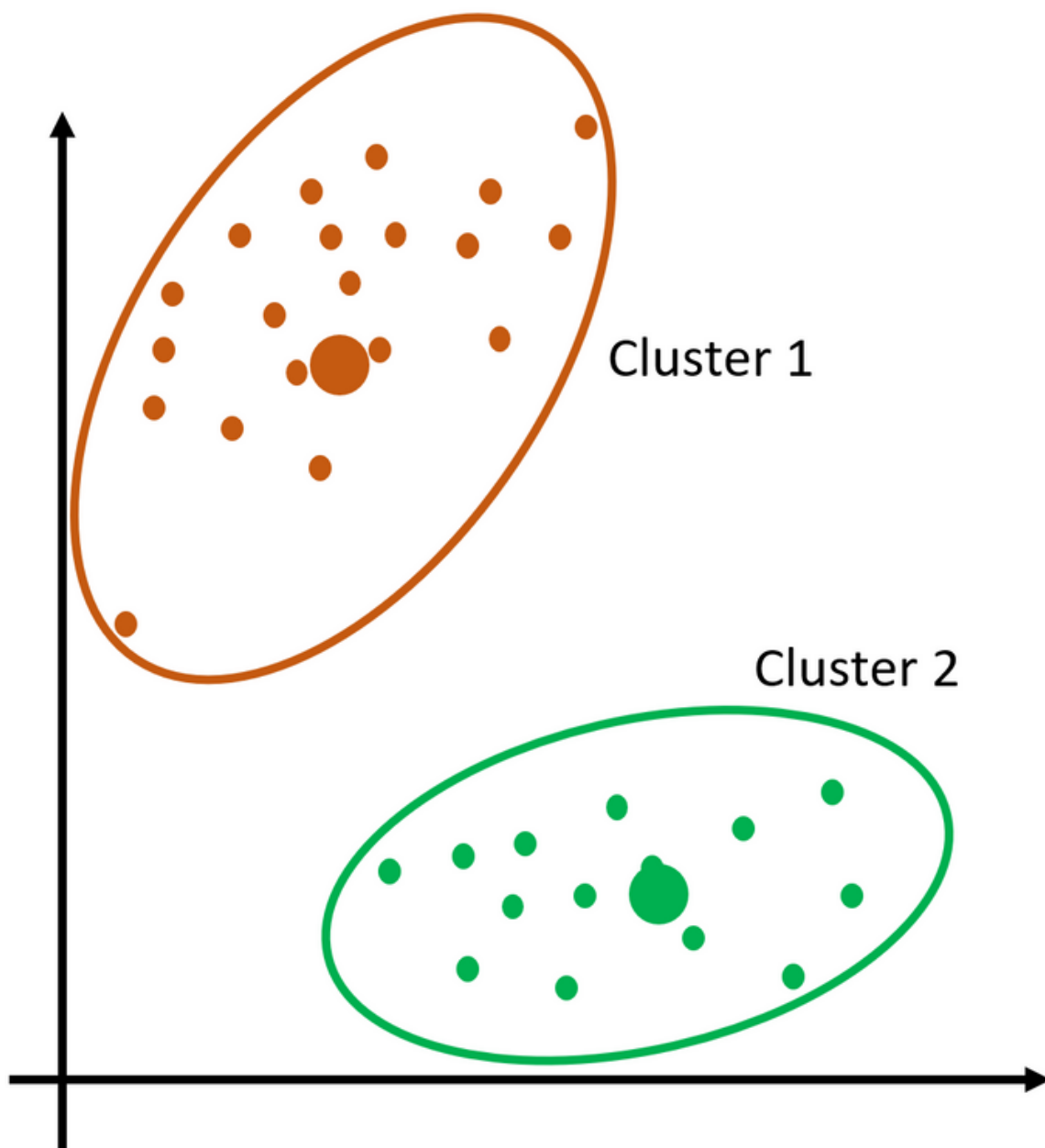
1. What’s the distance from the point to centroid 1?
2. What’s the distance from the point to centroid 2?



The centroids are again moved based on the new point-to-cluster assignments.



The k-means algorithm stops when no data points are assigned to a new cluster.



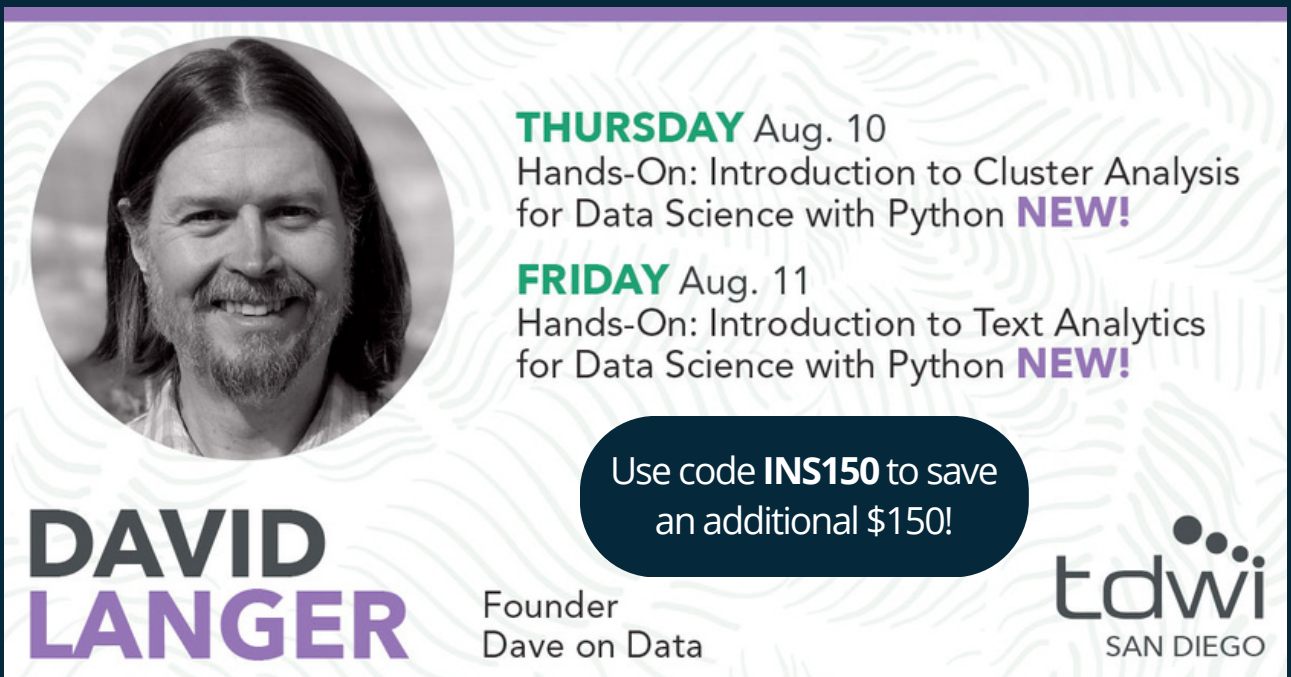
## Jumpstart Your Data Science Skills

The content in this document comes from the following live training course:

- Cluster Analysis for Data Science with Python

This will be one of 5 classroom experiences I will deliver at the TDWI San Diego conference next August. **These courses include 23 hands-on Python labs to jumpstart your data science skills.**

Be sure to check with your manager. TDWI is an approved training vendor for many organizations.



**THURSDAY** Aug. 10  
Hands-On: Introduction to Cluster Analysis  
for Data Science with Python **NEW!**

**FRIDAY** Aug. 11  
Hands-On: Introduction to Text Analytics  
for Data Science with Python **NEW!**

**DAVID**  
**LANGER**

Founder  
Dave on Data

Use code **INS150** to save  
an additional \$150!

**tdwi**  
SAN DIEGO

### Top-Rated Classroom Training



My classroom teaching is consistently top-rated by attendees. I combine an engaging style with many hands-on labs to build skills. Like Tyler, I can empower you with "go do" tools you can apply at work immediately.

No experience with Python? No worries!

Attendees of my live Python training courses will get **free access to a 4-hour Python online tutorial**. The Python Quick Start gives you the foundation you need to start learning data science.



---

### About the Author



My name is Dave Langer and I am the founder of Dave on Data.

I'm a hands-on analytics professional, having used my skills with Excel, SQL, and R/Python to craft insights, advise leaders, and shape company strategy.

I'm also a skilled educator, having trained 100s of working professionals in live in-person classroom settings and 1000s more via live virtual training and online courses.

In the past, I've held analytics leaderships roles at Schedulicity, Data Science Dojo, and Microsoft.

Drop me an email if you have any questions:  
[dave@daveondata.com](mailto:dave@daveondata.com)