K-MEANS
CLUSTERING

Centroid

Centroid

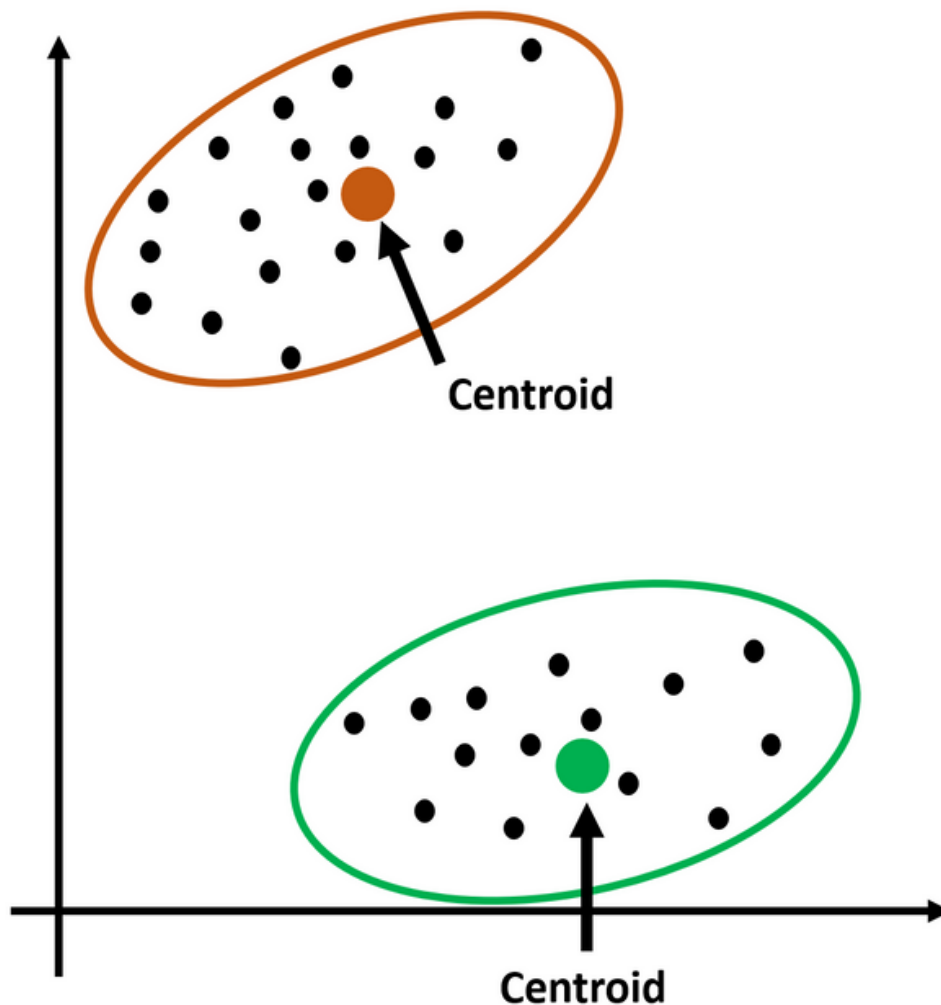## A BRIEF, INTUITIVE INTRODUCTION

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.
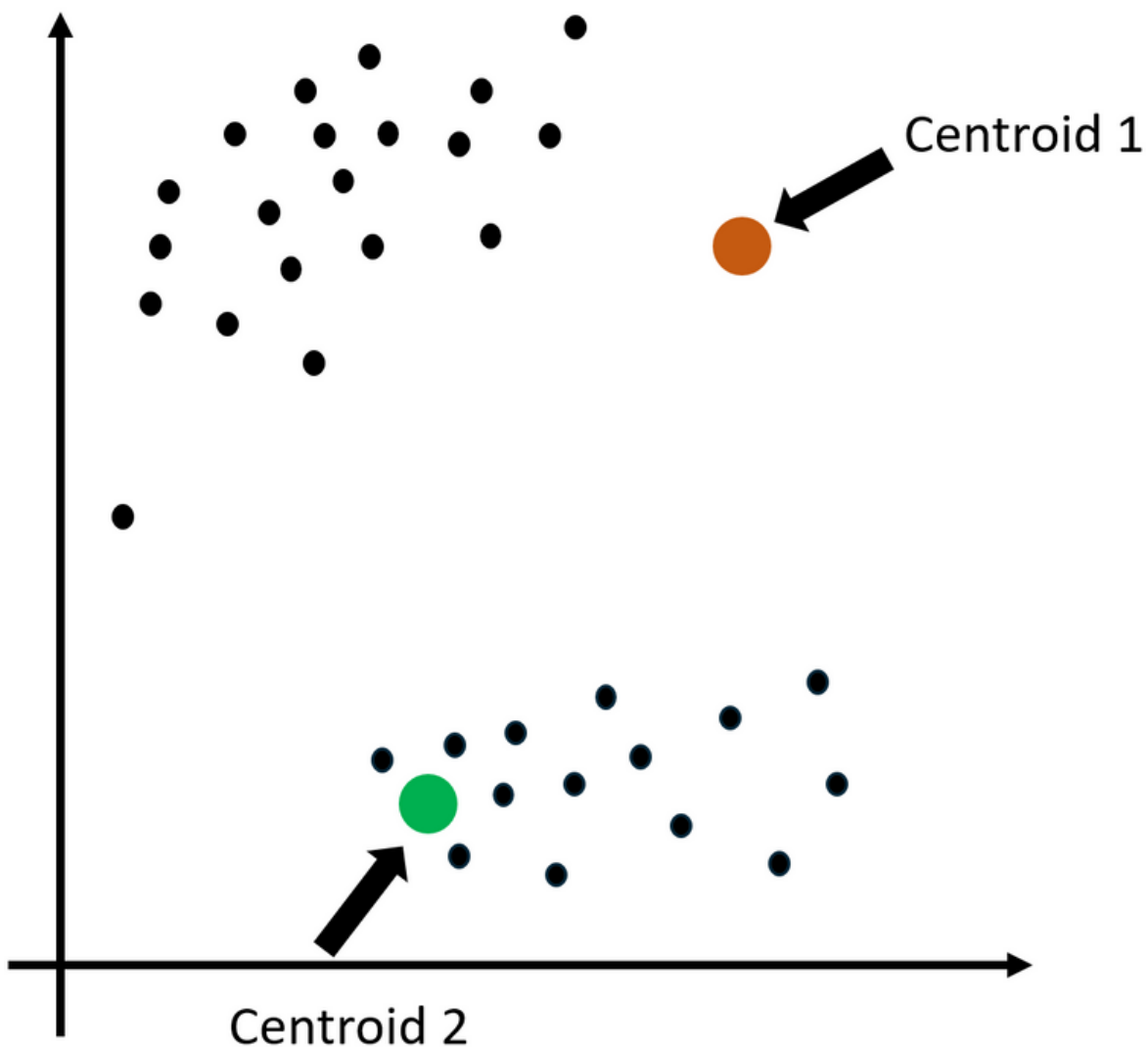4. Stop if no centroid changes, otherwise, go to 2.

Watch this data point
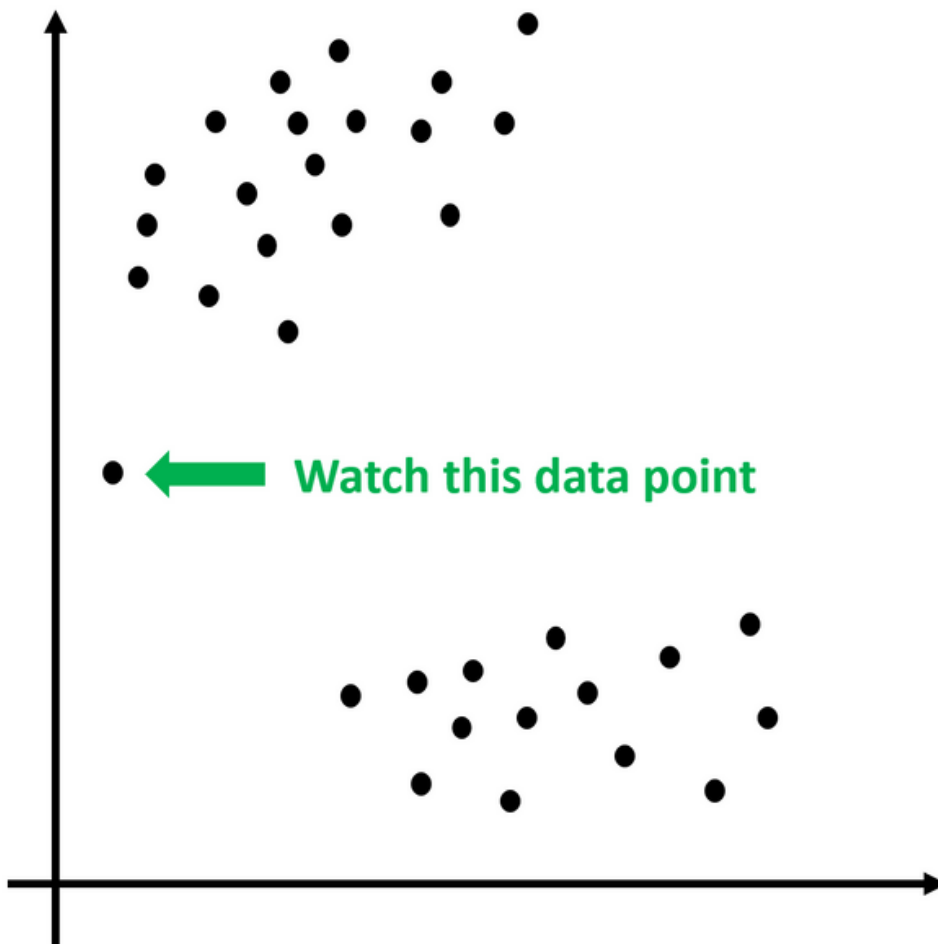
This is the dataset for the contrived example.

For this dataset, assume k = 2.

Dave ON DATA

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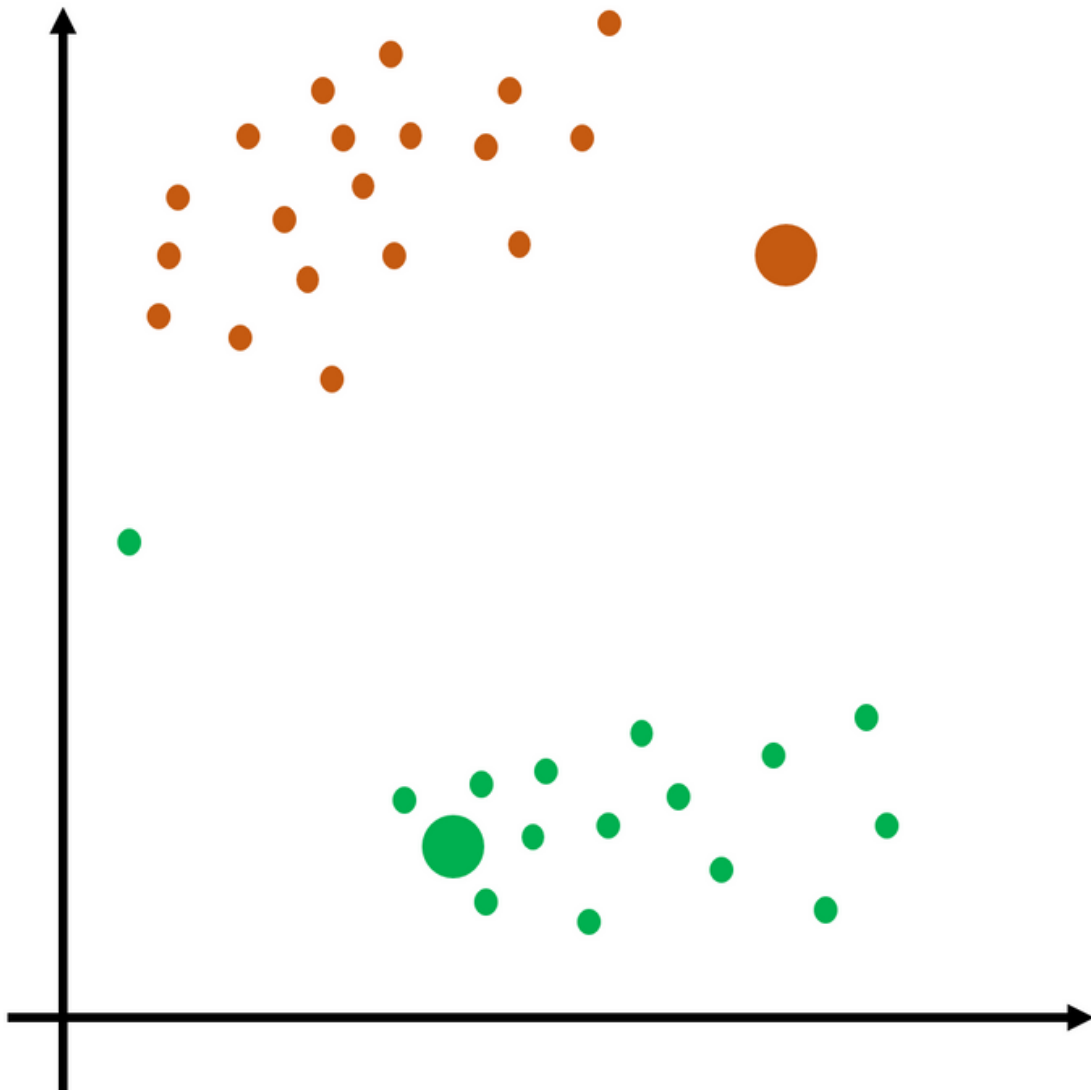
K-means then looks at each "point" (e.g., customer) to cluster in turn:

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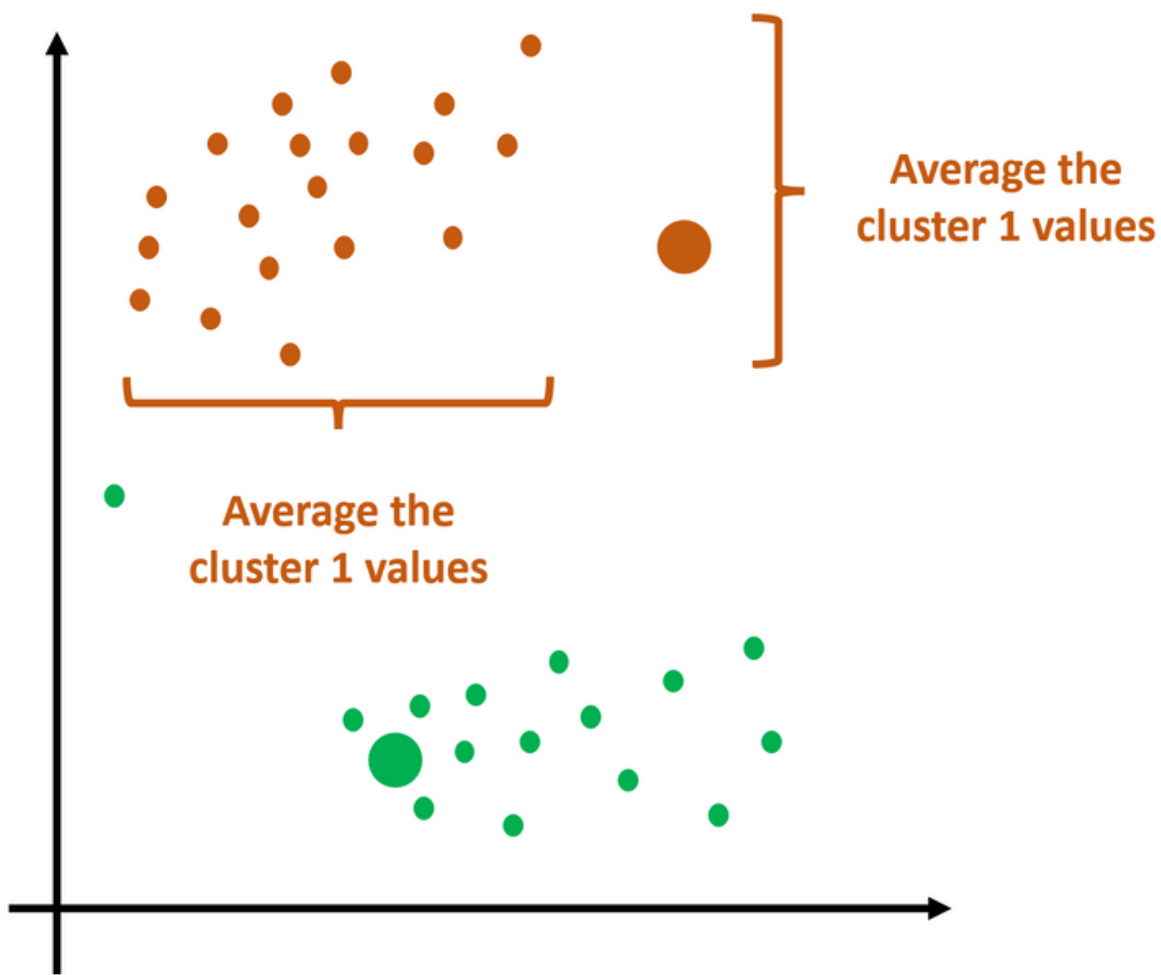2. What's the distance from the point to centroid 2?

Closest to centroid 1

Closest to centroid 2
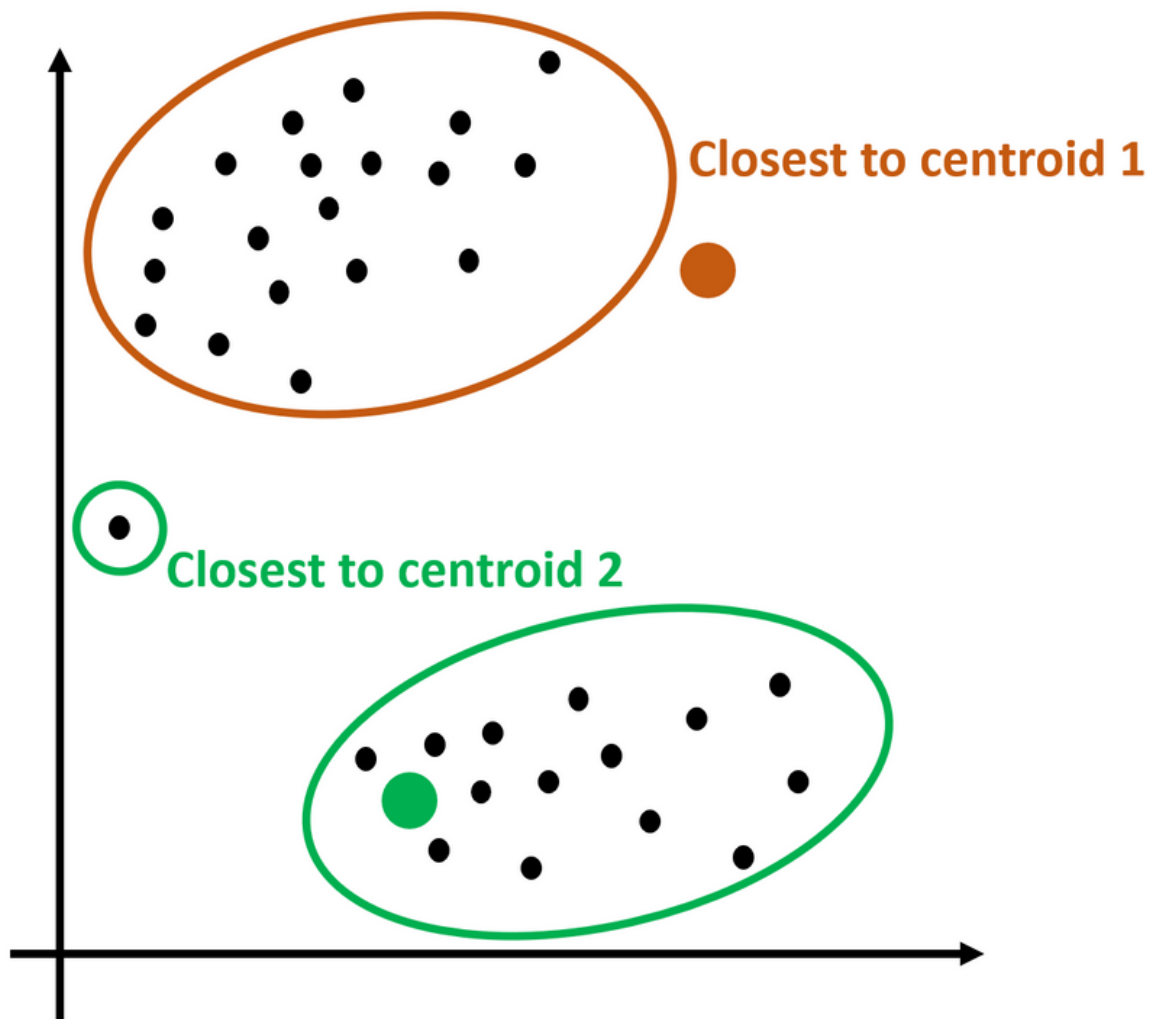
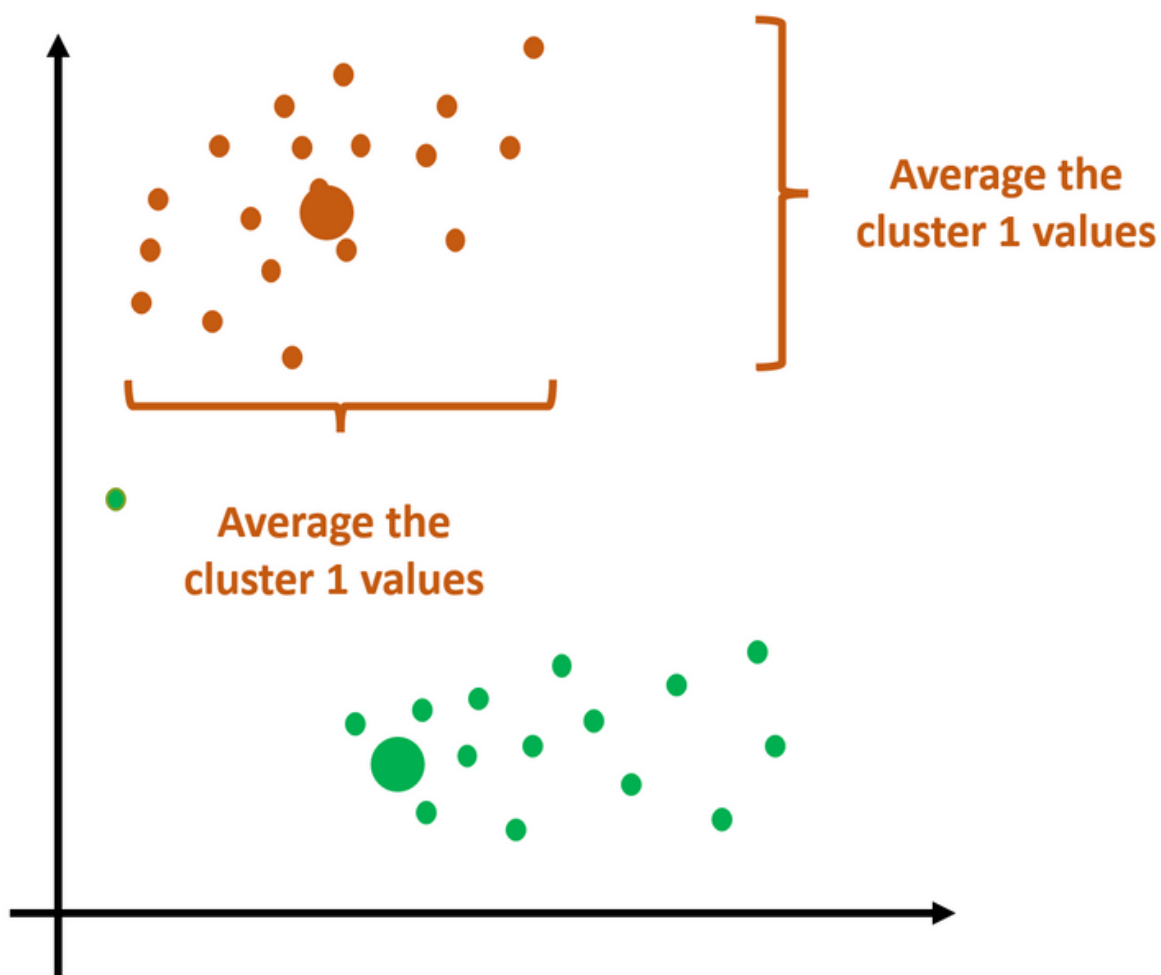Based on proximity, k-means then assigns each data point to a cluster.

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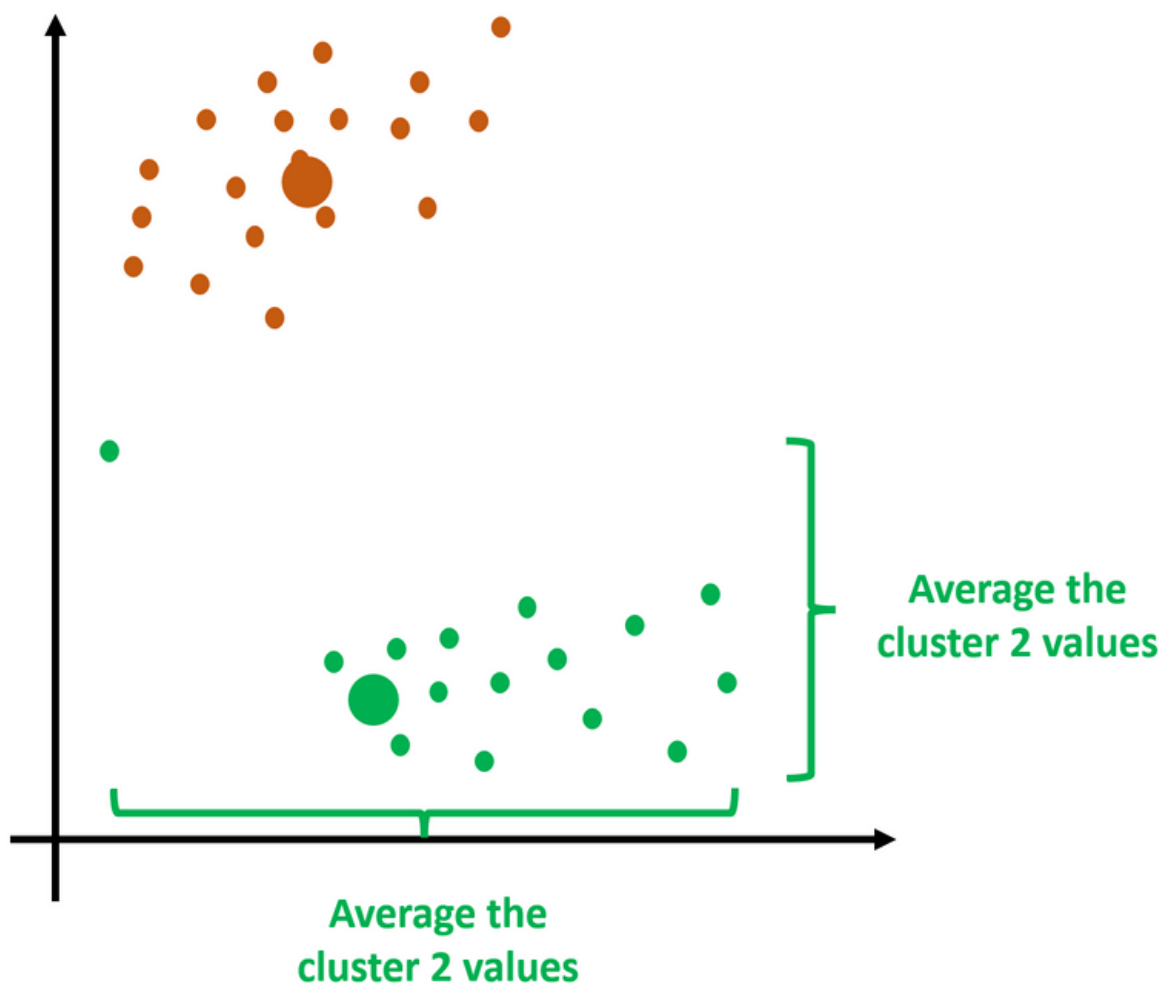
The algorithm moves the centroid (i.e., cluster center).

Average the cluster 1 values

Average the cluster 1 values

The moving process is repeated for each centroid.

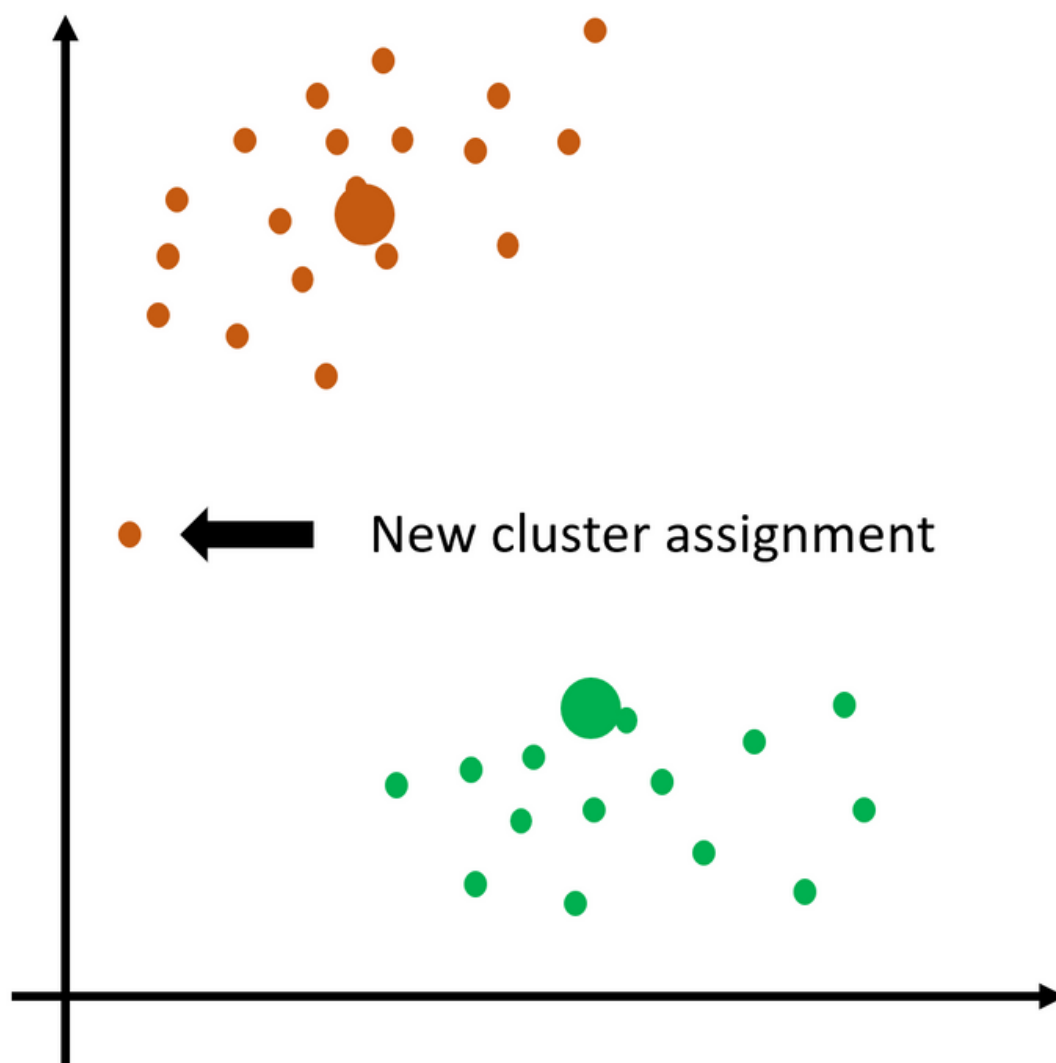Average the cluster 2 values

Average the cluster 2 values

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

The process is repeated until a stopping condition is reached.
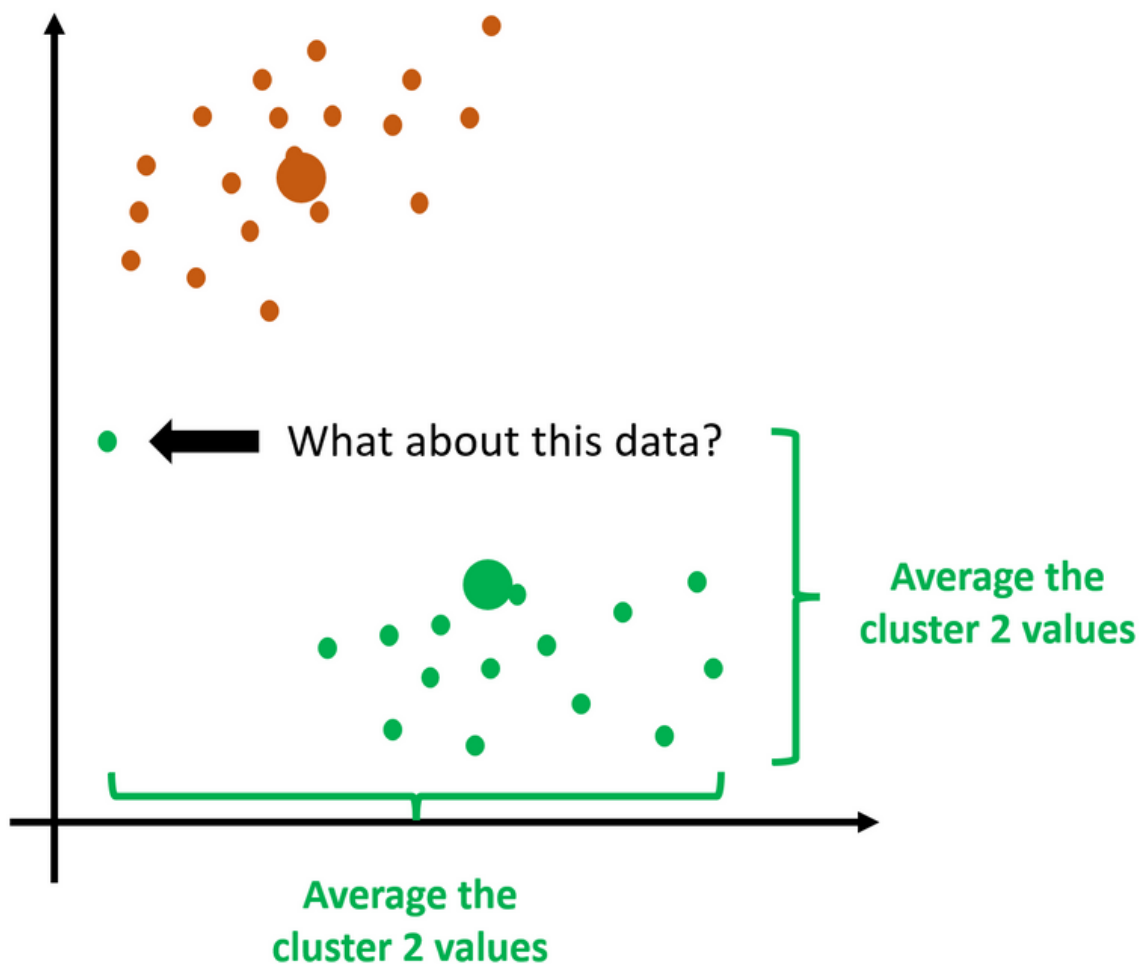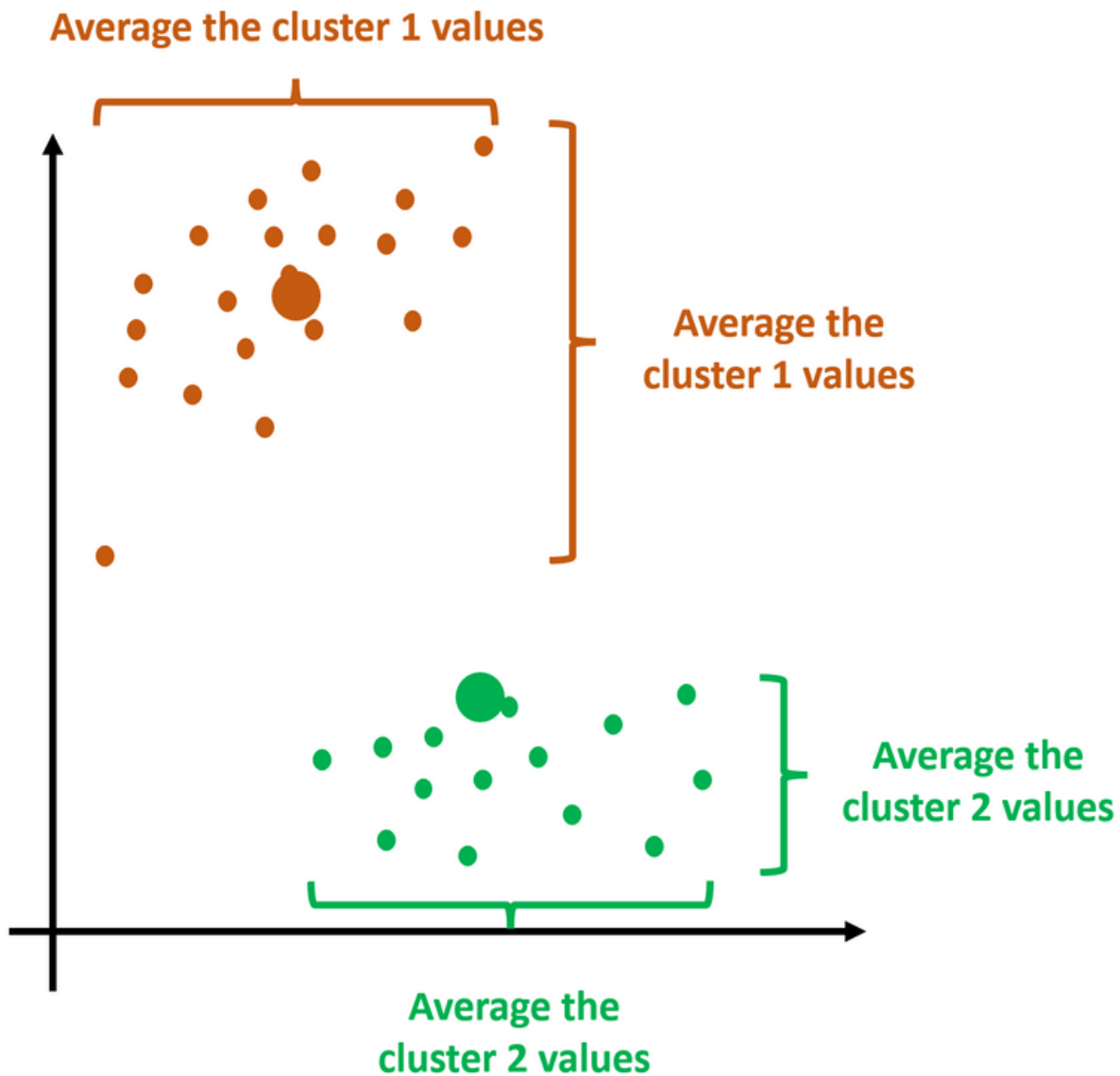
K-means looks at each "point" (i.e., document) to cluster in turn:
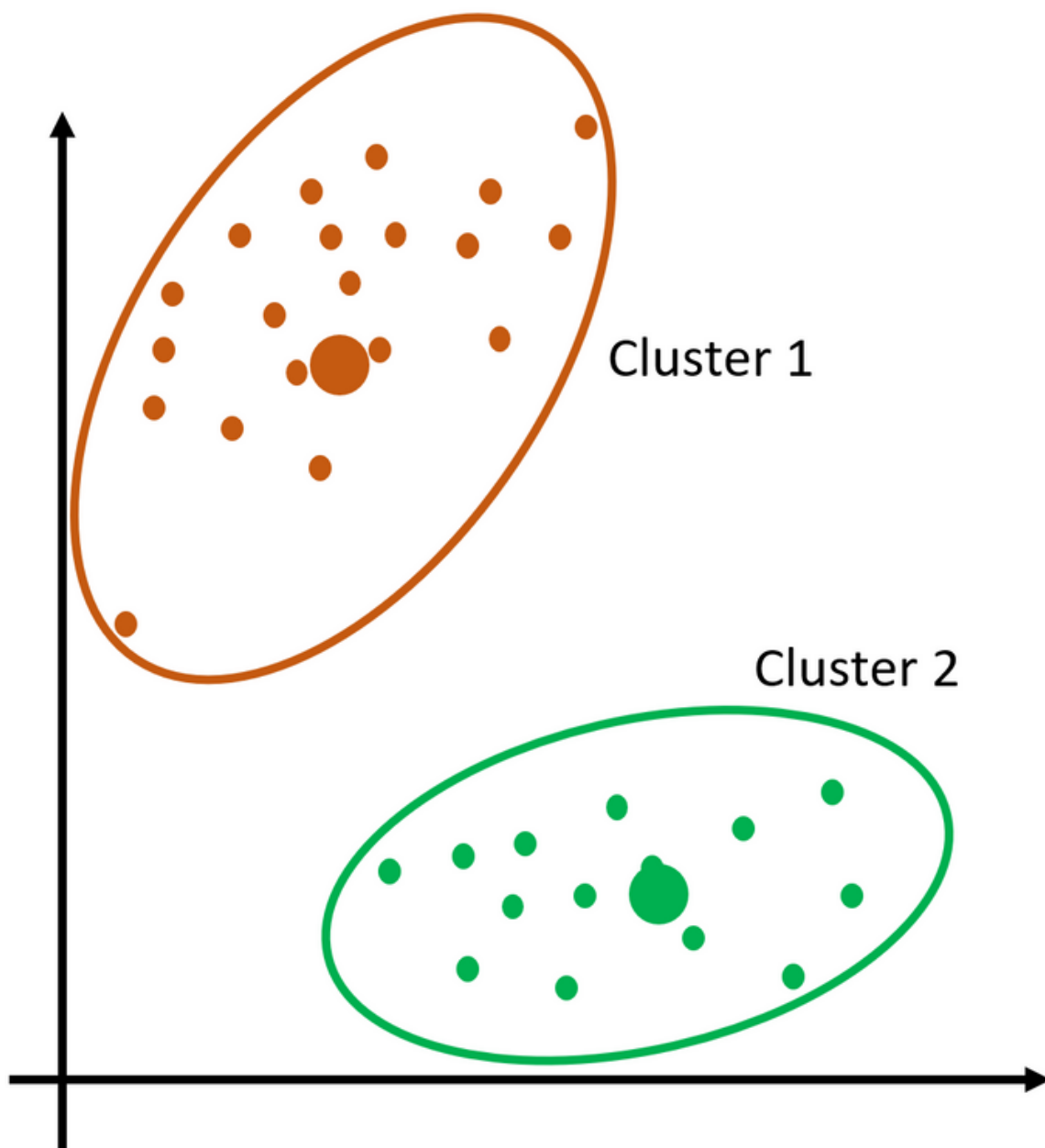
1. What's the distance from the point to centroid 1?
2. What's the distance from the point to centroid 2?

New cluster assignment

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.

Cluster 1

Cluster 2

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 is one of 3 classroom experiences I will deliver as part of the TDWI-certified Machine Learning Bootcamp this May. These courses include 14 hands-on Python labs to jumpstart your ML skills.

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

Andrew Koutalakis • 1st
Data Integration Manager, Global Customer Intelligence at IDEXX
2d • Edited •

What a week at the #TDWI Transform Las Vegas 2024 conference! Earning my Machine Learning Bootcamp Certificate over three days with David Langer was easily the most interesting part of my stay in Vegas. Dave's lectures and hands-on training are well organized, relevant and fun - imagine that!

I met so many great people from a variety of industries and occupations, companies large and small. The events and day to day conference operations were so well structured. Staff were friendly and accomdating to every need, and communication was handled very well. I won't wait so long next time to attend another TDWI Transform event!

https://bit.ly/TDWIChicago2024MLBootcamp

How it Works

The following three live training courses are being offered at the TDWI Transform Chicago conference on May 1-3:

- Machine Learning with Python Made Easy
- Data Wrangling for Machine Learning with Python
- Introduction to Cluster Analysis for Data Science with Python

Attending all three courses will earn you a TDWI Machine Learning Bootcamp certificate.

These hands-on courses are designed to quickly build machine learning (ML) skills.

Previous attendees have used these skills immediately after returning to work from the conference.

No experience with Python? No problem!

You will receive free access to a 4-hour Python Quick Start online tutorial.

https://bit.ly/TDWIChicago2024MLBootcamp

14 Hands-On Labs to Build Your Skills

May 1st – Machine Learning with Python Made Easy
- Lab 1 – Decision Trees
- Lab 2 – Random Forests
- Lab 3 – Feature Engineering
- Lab 4 – Model Testing
- Lab 5 – Model Improvement

May 2nd - Data Wrangling for Machine Learning with Python
- Lab 1 – Data Profiling
- Lab 2 – The Mighty pandas
- Lab 3 – Wrangling Strings
- Lab 4 – Joining Data

May 3rd – Cluster Analysis for Data Science with Python
- Lab 1 – K-Means Clustering
- Lab 2 – Optimizing K-Means
- Lab 3 – Optimizing DBSCAN
- Lab 4 – Dimensionality Reduction
- Lab 5 – Categorical Data

Dave
ON DATA

Register Now. Seats are Limited.

You can **save $1,000** off my Machine Learning Bootcamp!

Use promo code **FACULTY35** to get the savings.

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