

AI Bootcamp

Introduction to Machine Learning

Module 11 Day 1



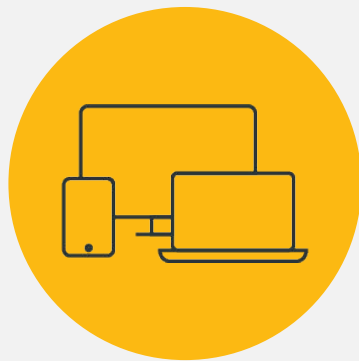
Class Objectives

By the end of class, you will be able to:

- 1 Recognize the differences between supervised and unsupervised machine learning.
- 2 Define clustering and how it is used in data science.
- 3 Apply the K-means algorithm to identify clusters in a given dataset.
- 4 Determine the optimal number of clusters for a dataset using the elbow method.

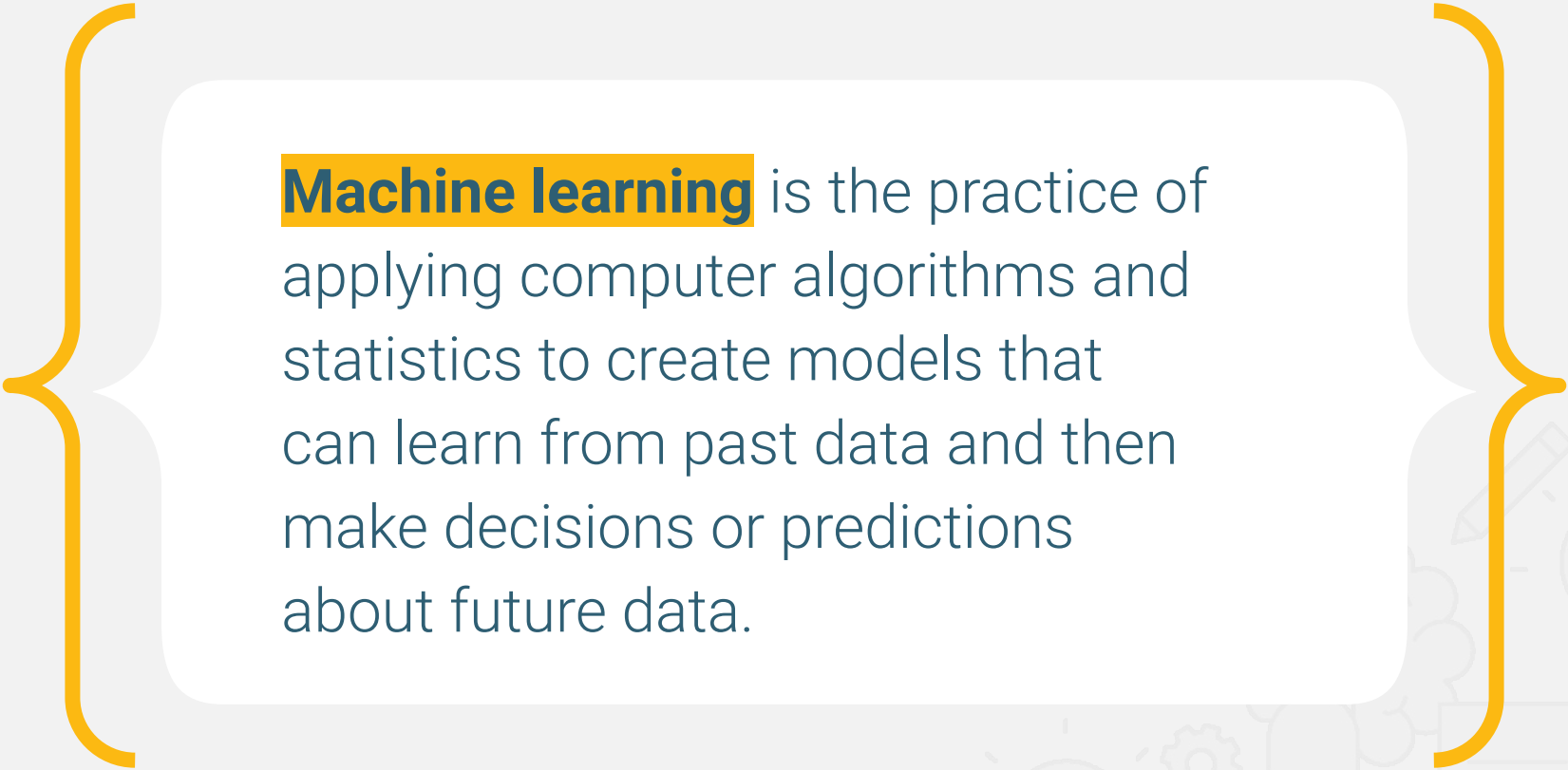


Welcome




Instructor **Demonstration**

Demystifying Machine Learning

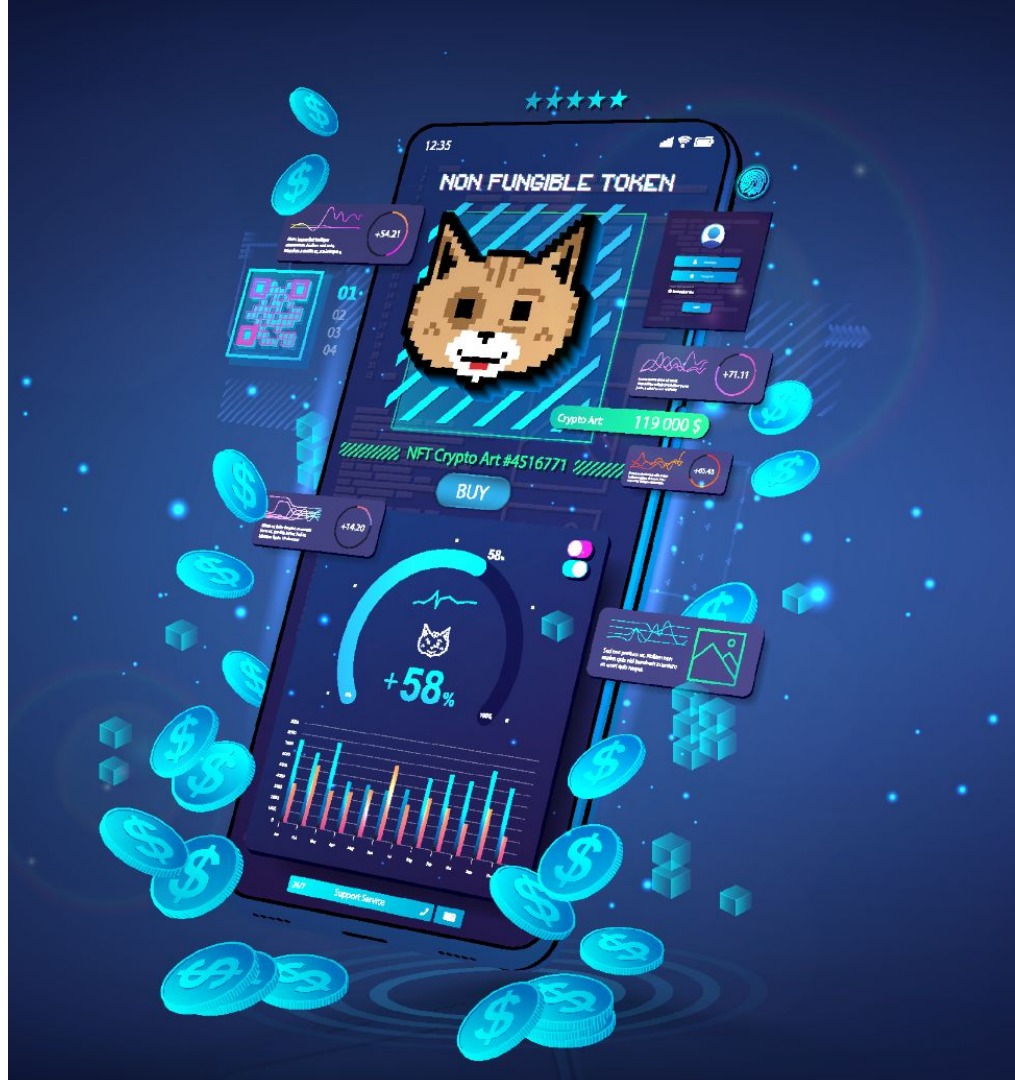


Machine learning is the practice of applying computer algorithms and statistics to create models that can learn from past data and then make decisions or predictions about future data.



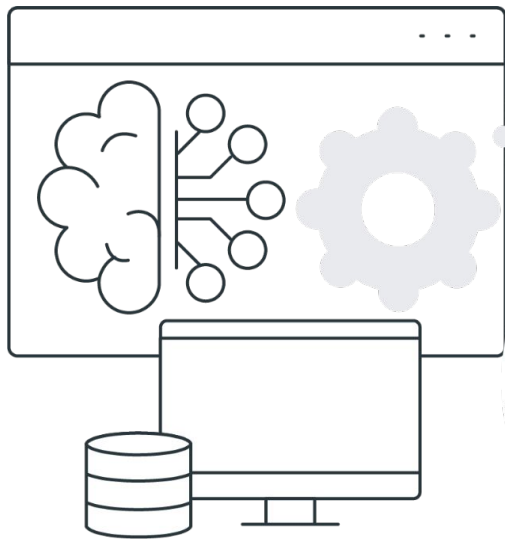
Machine learning is changing industries at an unprecedented pace.

Machine learning allows for decisions to be made more quickly and efficiently than ever before.

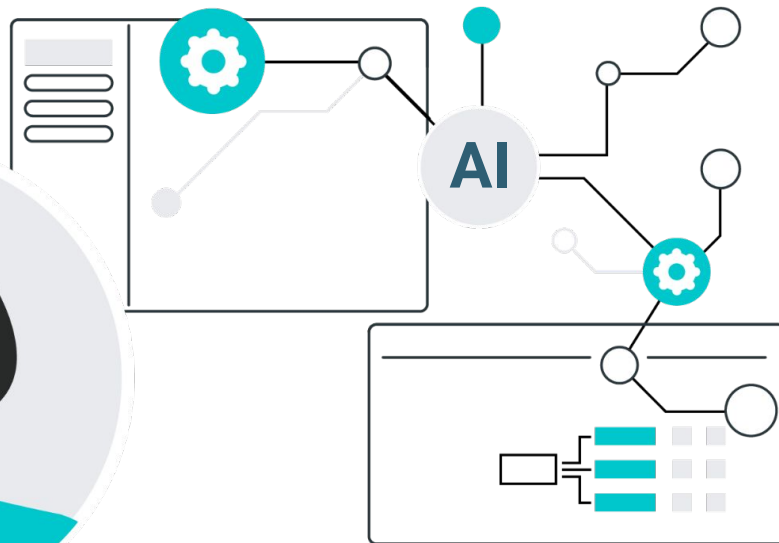
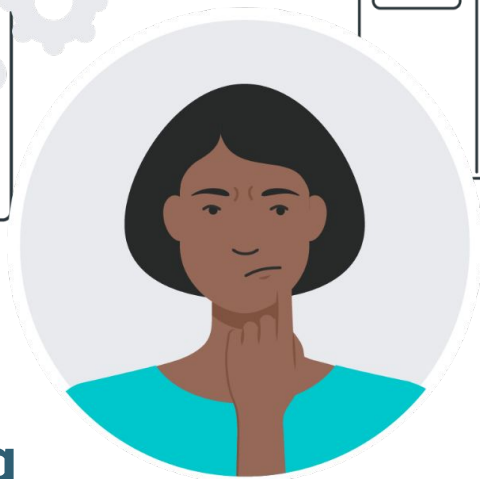


The Mysticism of Machine Learning

Despite the mainstream use of the term “machine learning,” most people still don’t know what machine learning *really* is.



Machine Learning



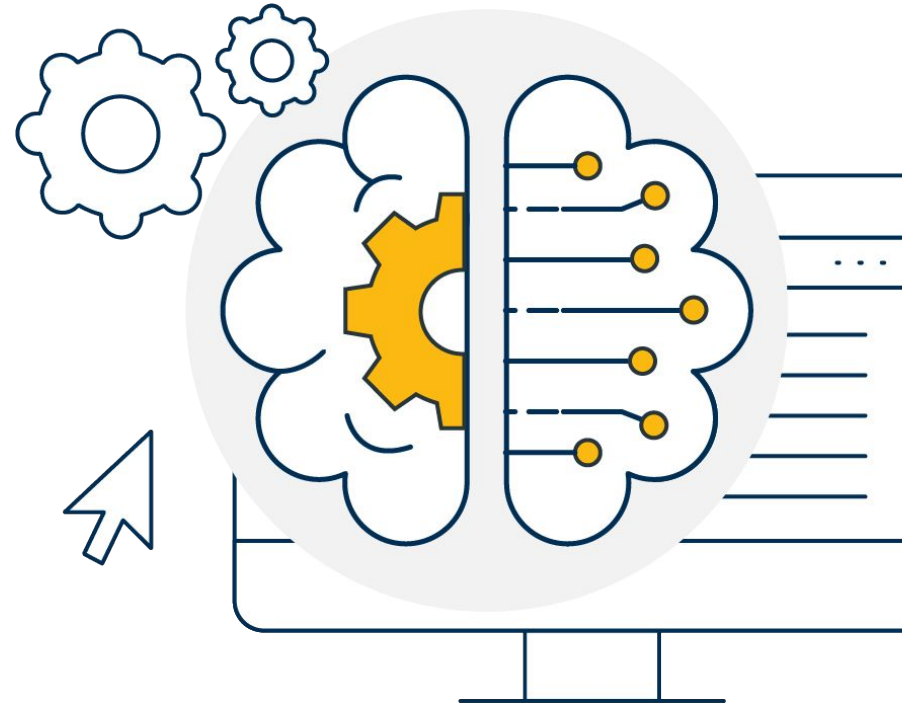
Artificial Intelligence

Machine Learning

Algorithms learn how to make decisions without needing anyone to program the logic directly.

They learn the patterns, behavior, and relationships on their own directly from the data.

They then use that knowledge to make decisions and predictions.





Here's an example of
how machine learning
can be useful.



Machine Learning

Imagine that you work as a fraud analyst in a bank, and you want to identify fraudulent transactions.

Option 1

Create a 5,000-line **if-else** decision structure that evaluates every price range and product category to determine if a transaction counts as fraudulent.

Option 2

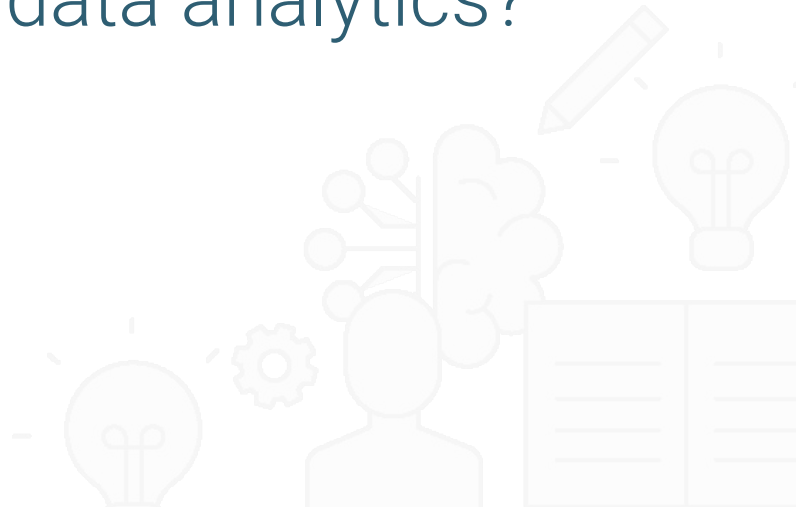
Use machine learning algorithms to review all of the transactions that an account owner has ever made. Then, group the transactions and predict whether the most recent transaction counts as fraudulent.



This is the kind of machine learning solution that you'll learn to build.



Why is machine learning
essential for data analytics?



Machine Learning in Data Analytics



Applications for machine learning vary widely, but all share the common goals of making more efficient decisions, predictions, and products.



Machine learning applications have streamlined operational processes across many industries.



Incorporating machine learning has helped businesses dramatically improve responsiveness to customer demands.



What are some **examples** of machine learning models that you've heard of?



Types of ML

Examples include:



Regression



Clustering



Neural networks



Deep learning

Types of ML

We can group all of these models into two main buckets:

01

Unsupervised Learning

The algorithm tries to make sense of an **unlabeled dataset** by extracting features and patterns on its own.

02

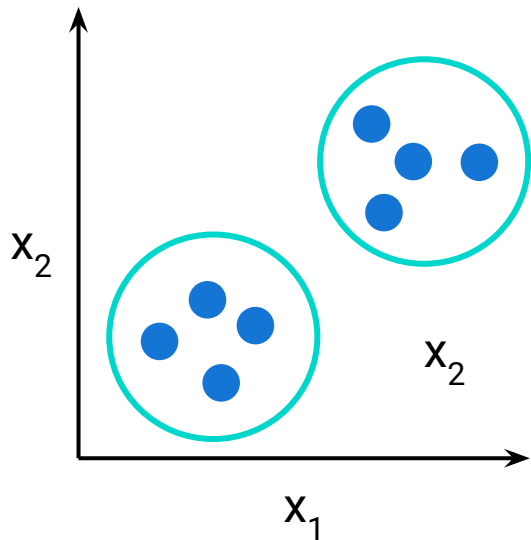
Supervised Learning

The algorithm learns on a **labeled dataset**, where each example in the dataset is tagged with the answer.

This provides an answer key that can be used to evaluate the accuracy of the model.

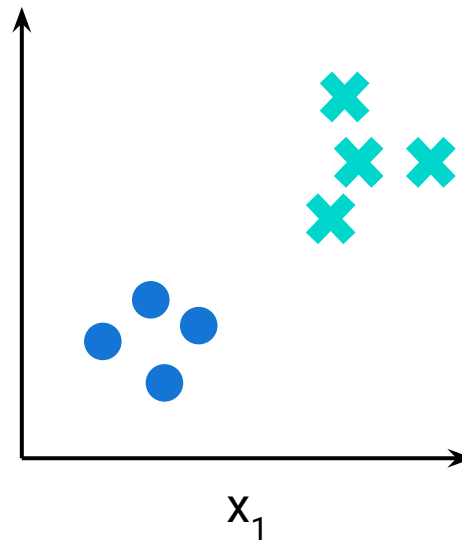
Unsupervised Learning vs. Supervised Learning

Unsupervised Learning



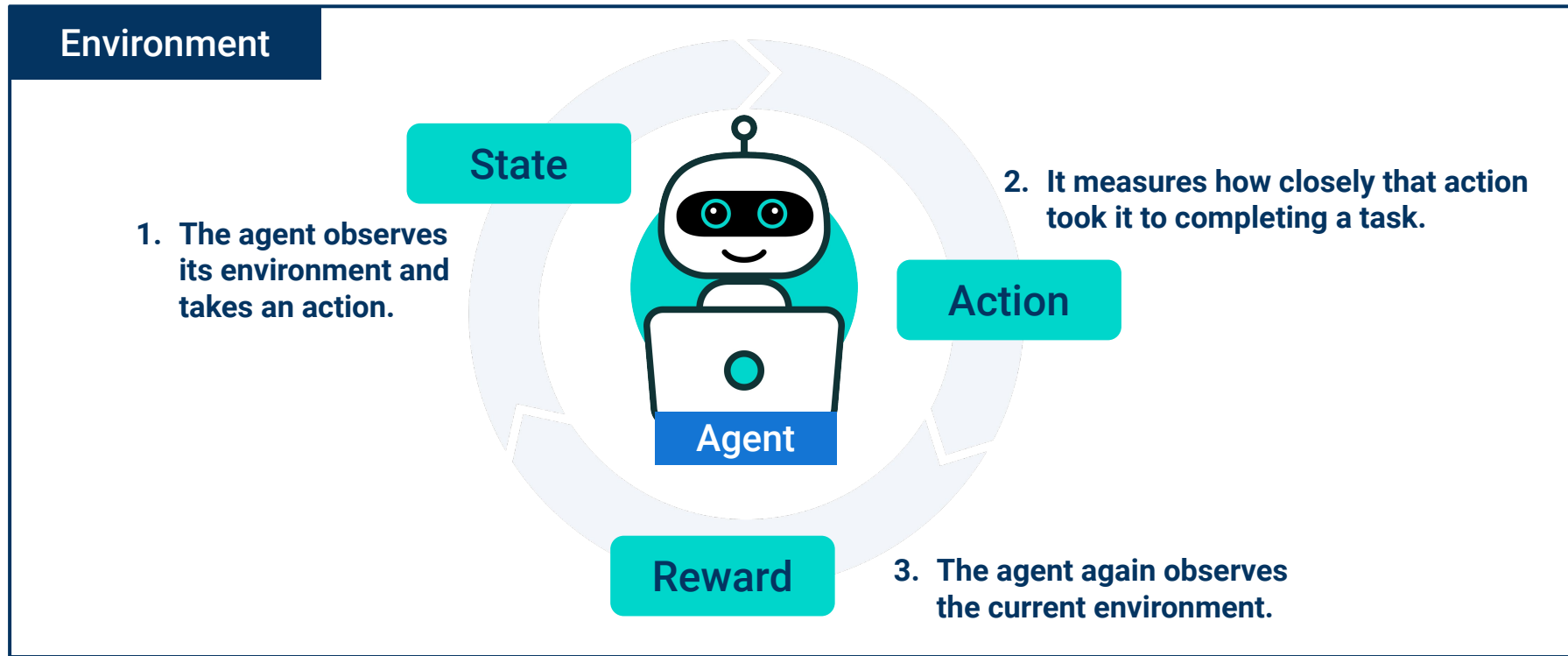
vs.

Supervised Learning

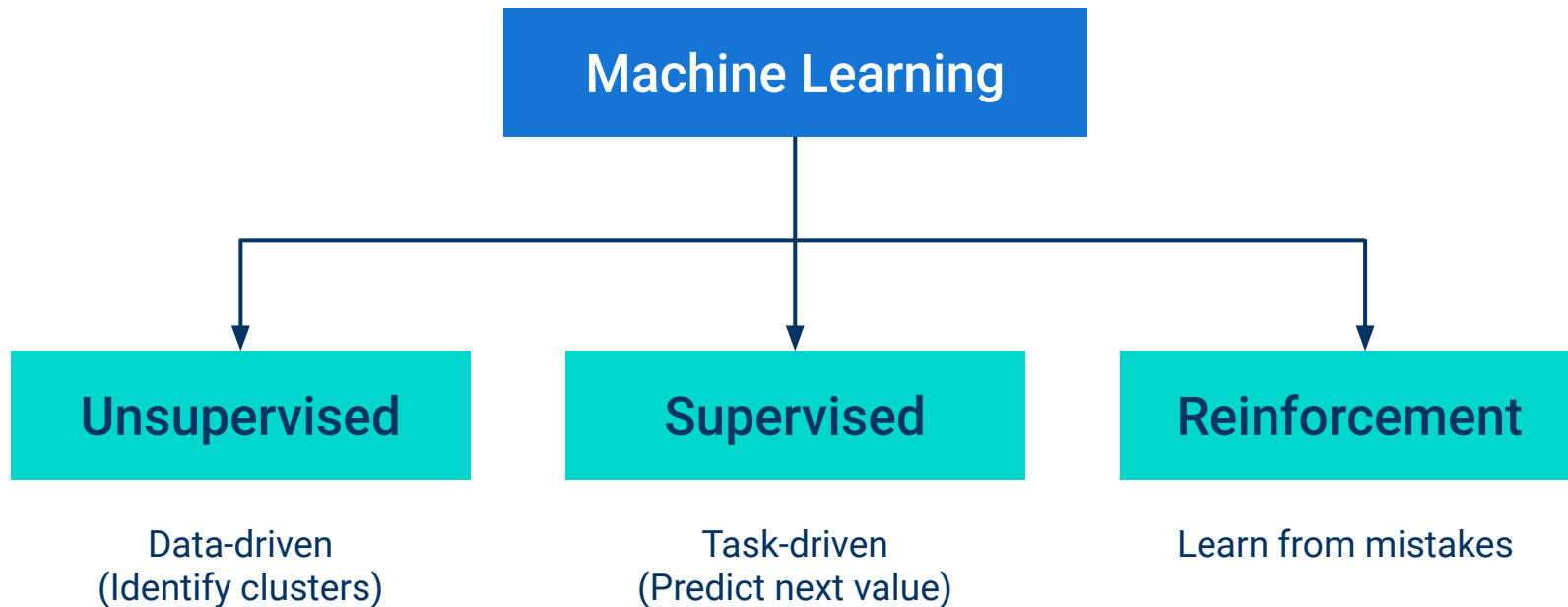


Reinforcement Learning

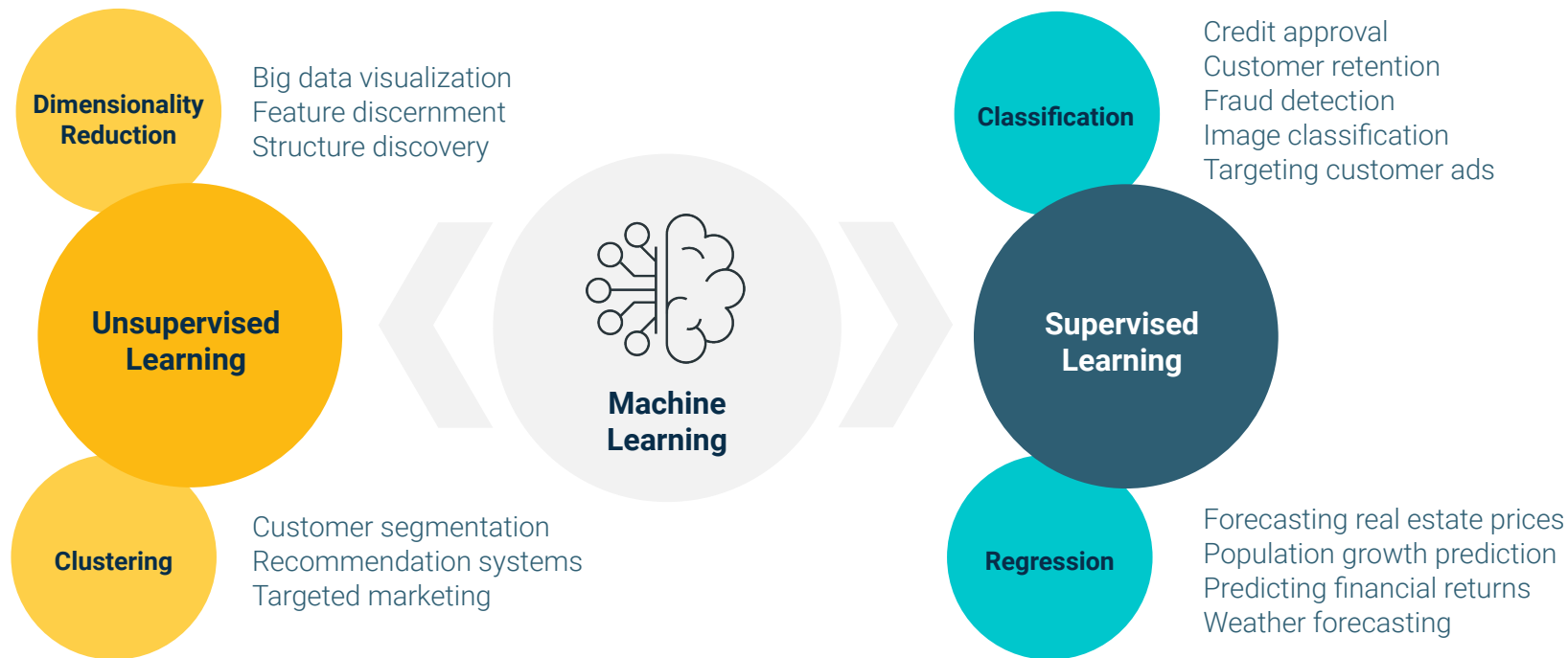
This third type of machine learning algorithm is used less frequently but still has important applications in data analytics.



Three Types of ML



Types of ML



Types of ML

Most Python libraries for machine learning use a common interface to build and use machine learning models.



Pandas

SciPy

TensorFlow

Matplotlib

Scikit Learn

NumPy



Questions?



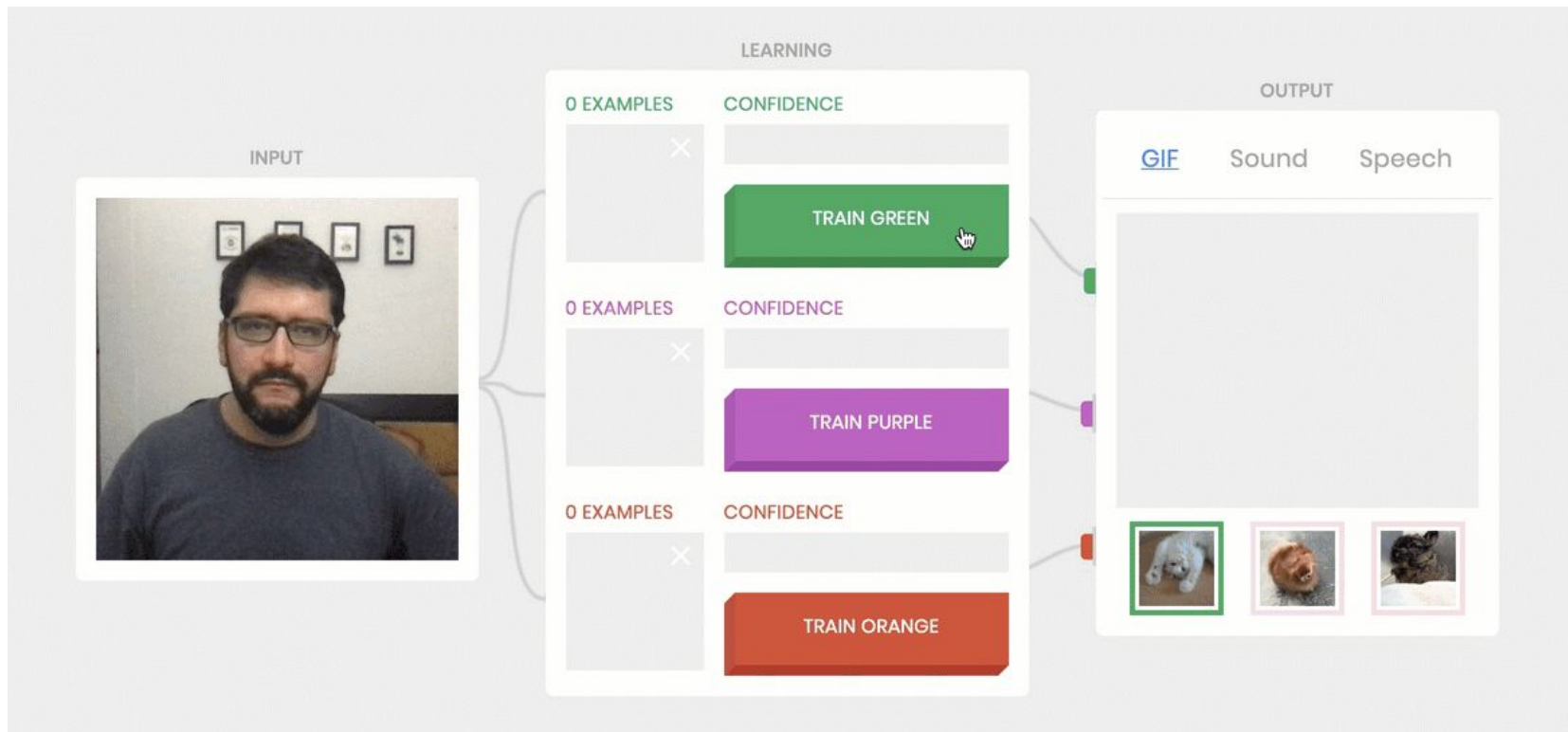


Instructor **Demonstration**

Machine Learning is Awesome

Teachable Machine in Action

The [Teachable Machine project from Google](#) shows the fundamental mechanism of a neural network by training a model that recognizes gestures from your webcam to predict one of three classes.






Instructor **Demonstration**

Introduction to Unsupervised Learning



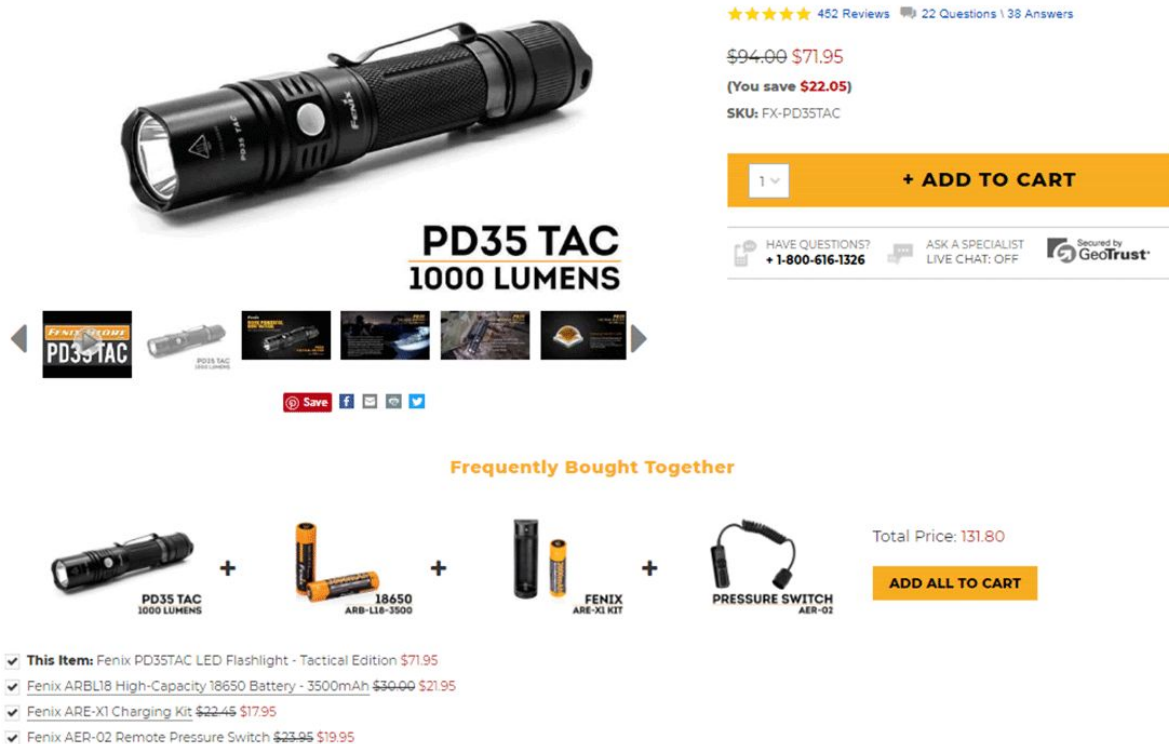
Unsupervised learning algorithms

use test data to construct models that categorize relationships among data points.



Introduction to Unsupervised Learning

For example, when you're reviewing a particular item for purchase on a website, unsupervised learning algorithms might be used to identify related items that are frequently bought together.



The screenshot displays a product page for the Fenix PD35 TAC flashlight. The main image shows the flashlight, with the text "PD35 TAC 1000 LUMENS" below it. To the right, the price is listed as \$94.00 (original) and \$71.95 (current), with a note "(You save \$22.05)". The SKU is FX-PD35TAC. Below the price is a quantity selector set to 1 and a yellow "+ ADD TO CART" button. Further down, there are links for "HAVE QUESTIONS? +1-800-616-1326", "ASK A SPECIALIST LIVE CHAT: OFF", and a "Secured by GeoTrust" logo. A horizontal carousel of smaller images shows the flashlight in various settings. Below this is a "Save" button and social media icons. The "Frequently Bought Together" section shows the flashlight plus three other items: two 18650 ARB-L18-3500 batteries, a Fenix ARE-X1 charging kit, and a pressure switch AER-02. The total price for the bundle is 131.80. A yellow "ADD ALL TO CART" button is present. At the bottom, a list of items with checkmarks shows the individual prices: Fenix PD35TAC LED Flashlight - Tactical Edition (\$71.95), Fenix ARBL18 High-Capacity 18650 Battery - 3500mAh (\$21.95), Fenix ARE-X1 Charging Kit (\$17.95), and Fenix AER-02 Remote Pressure Switch (\$19.95).

★★★★★ 452 Reviews 22 Questions 138 Answers

~~\$94.00~~ \$71.95
(You save **\$22.05**)
SKU: FX-PD35TAC

1 **+ ADD TO CART**

HAVE QUESTIONS?
+1-800-616-1326

ASK A SPECIALIST
LIVE CHAT: OFF

Secured by
GeoTrust

PD35 TAC 1000 LUMENS

Frequently Bought Together

PD35 TAC 1000 LUMENS + 18650 ARB-L18-3500 + FENIX ARE-X1 KIT + PRESSURE SWITCH AER-02

Total Price: 131.80

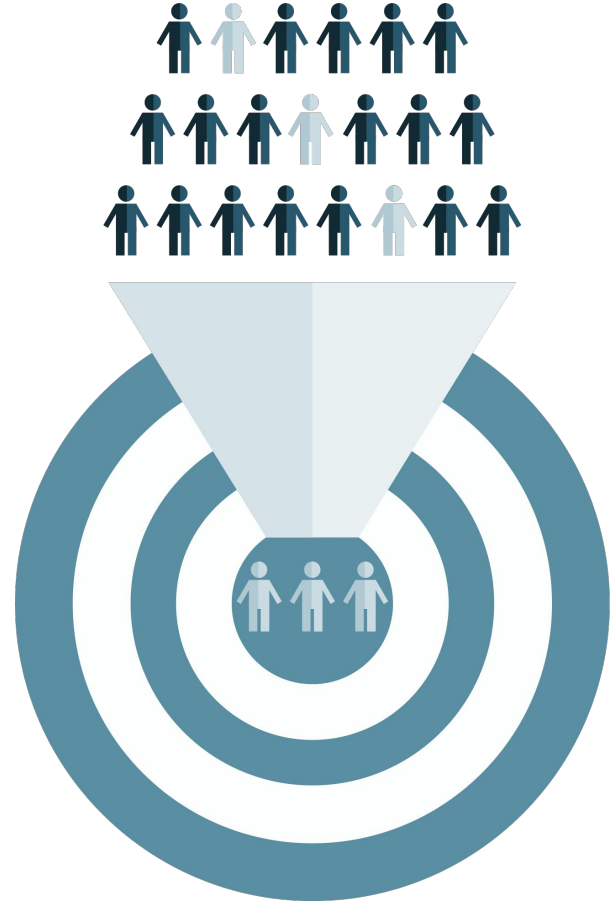
ADD ALL TO CART

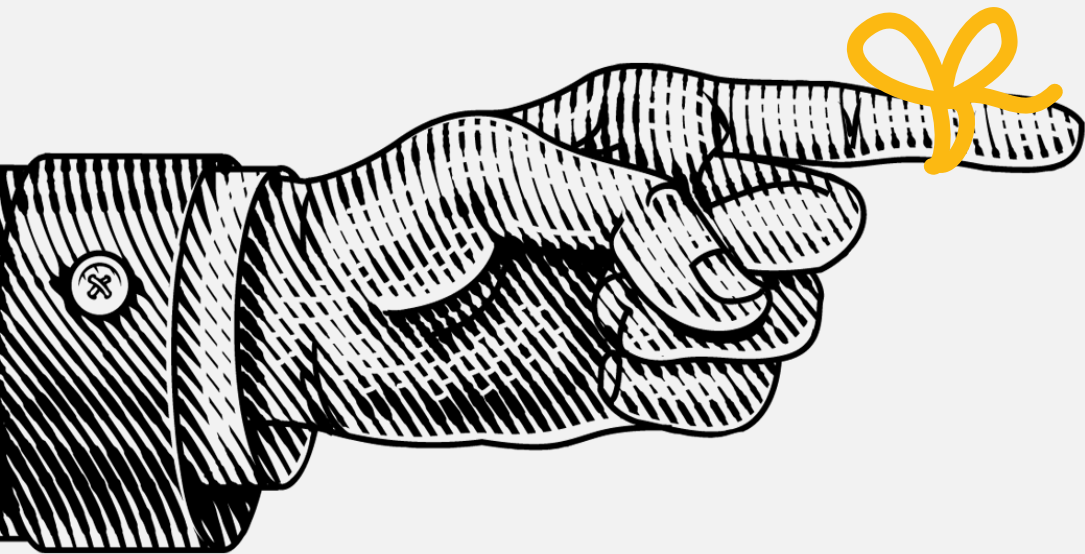
- ✓ **This Item:** Fenix PD35TAC LED Flashlight - Tactical Edition \$71.95
- ✓ Fenix ARBL18 High-Capacity 18650 Battery - 3500mAh ~~\$30.00~~ \$21.95
- ✓ Fenix ARE-X1 Charging Kit ~~\$22.45~~ \$17.95
- ✓ Fenix AER-02 Remote Pressure Switch ~~\$23.95~~ \$19.95



This power to recognize data patterns has broad applications in data analytics.

Unsupervised learning can be used to **identify clusters**, or related groups, of clients to target with product offerings or marketing campaigns.





Remember,

the two most frequently used methods of machine learning are **supervised learning** and **unsupervised learning**.



Supervised vs. Unsupervised Learning

Supervised Learning	Unsupervised Learning
Input data is labeled.	Input data is unlabeled.
Uses training datasets.	Uses input datasets.
Goal: Predict a class or value.	Goal: Determine patterns or group data, called data clusters.

Challenges of Unsupervised Learning

Unsupervised learning comes with challenges:

1 Because the data isn't labeled, we don't know if the output is correct.

2 The algorithm creates its own categories for the data, so an expert must determine if these categories are meaningful.

3 Even with challenges, unsupervised learning can be useful for a variety of applications, including the following customer segmentation tasks:

- Grouping customers by spending habits
- Finding fraudulent credit card charges
- Identifying unusual data points (outliers) within the dataset

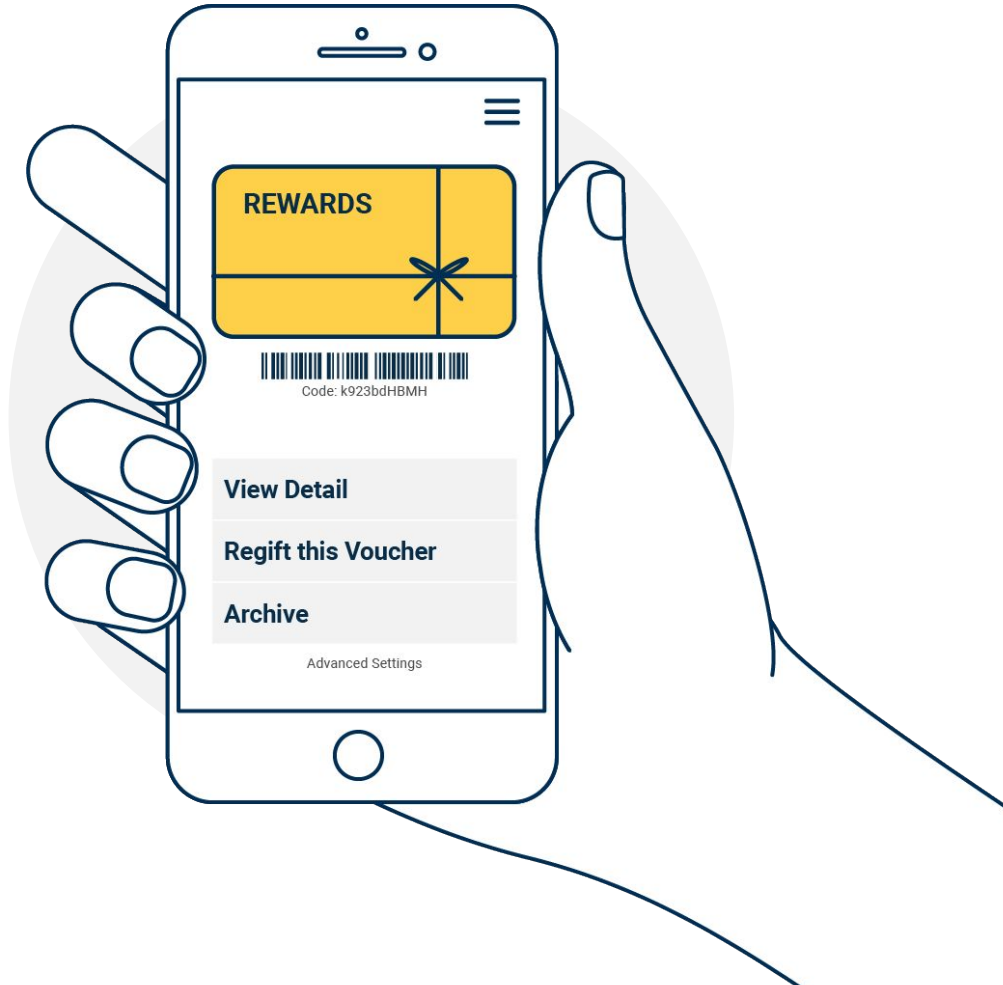


How might businesses
use **clustering**?



One possible answer:

Clustering can be used to group customers by spending habits and create customized offers via email or mobile apps.





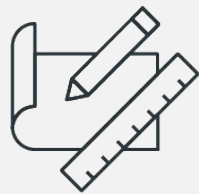
How might credit
card companies use
anomaly detection?





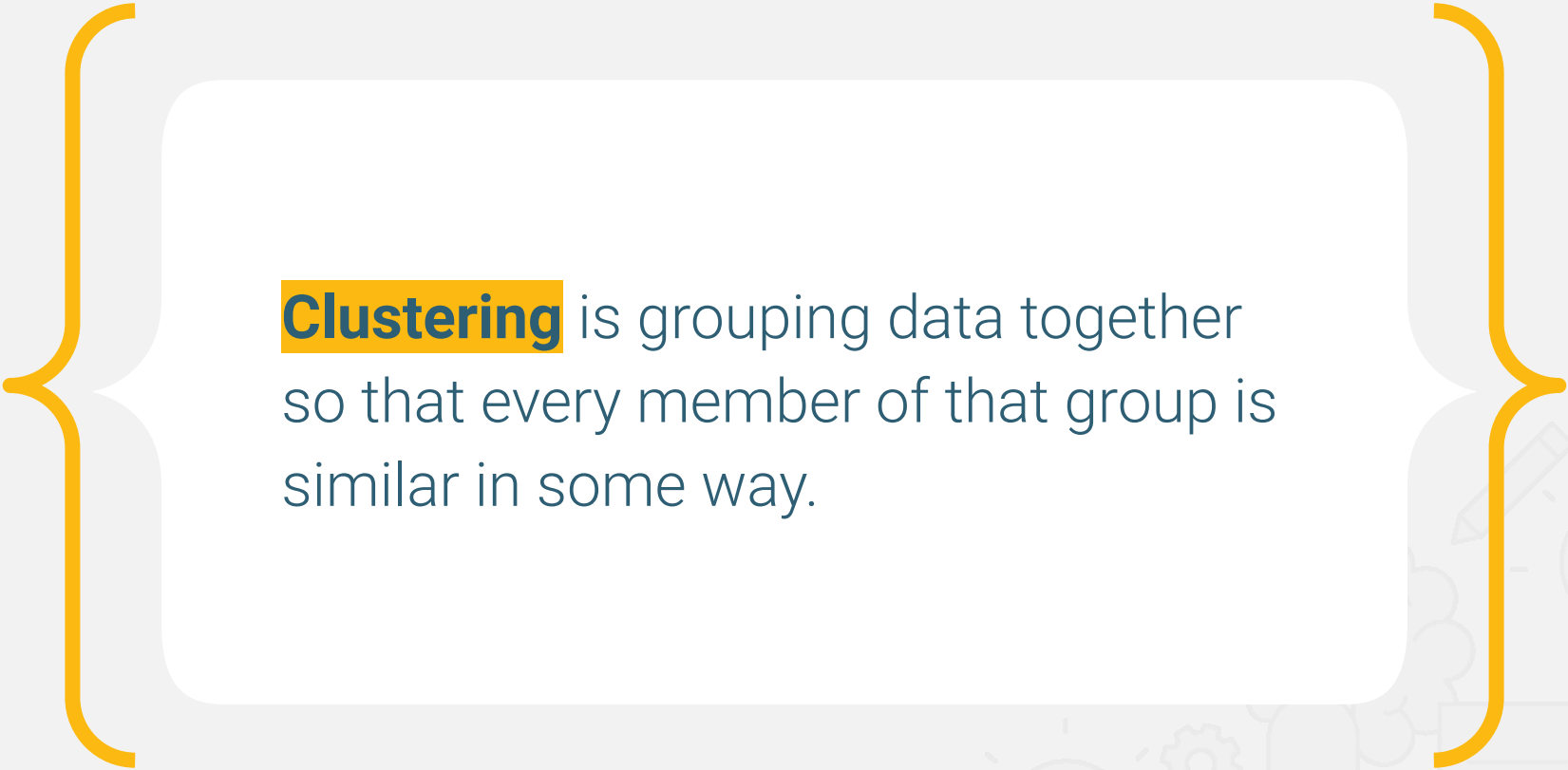
One possible answer:

Anomaly detection can be used to detect potential customers who might default on their loan by grouping transactions based on a variety of features.

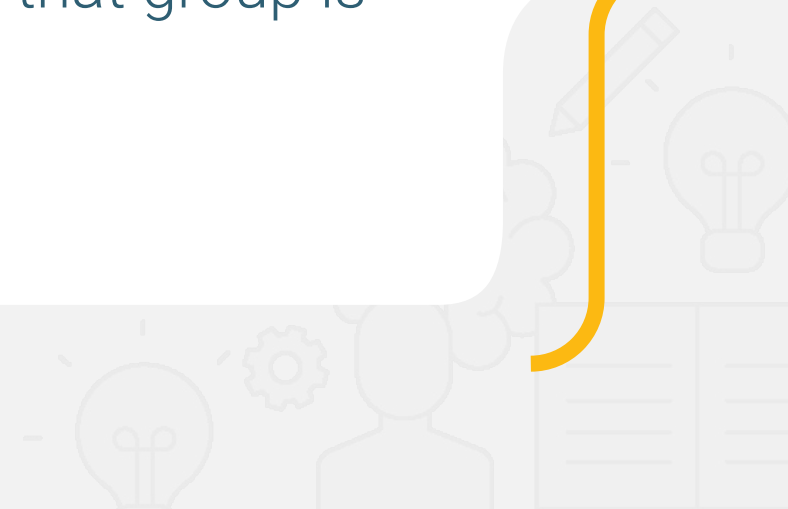


Clustering **Explained**



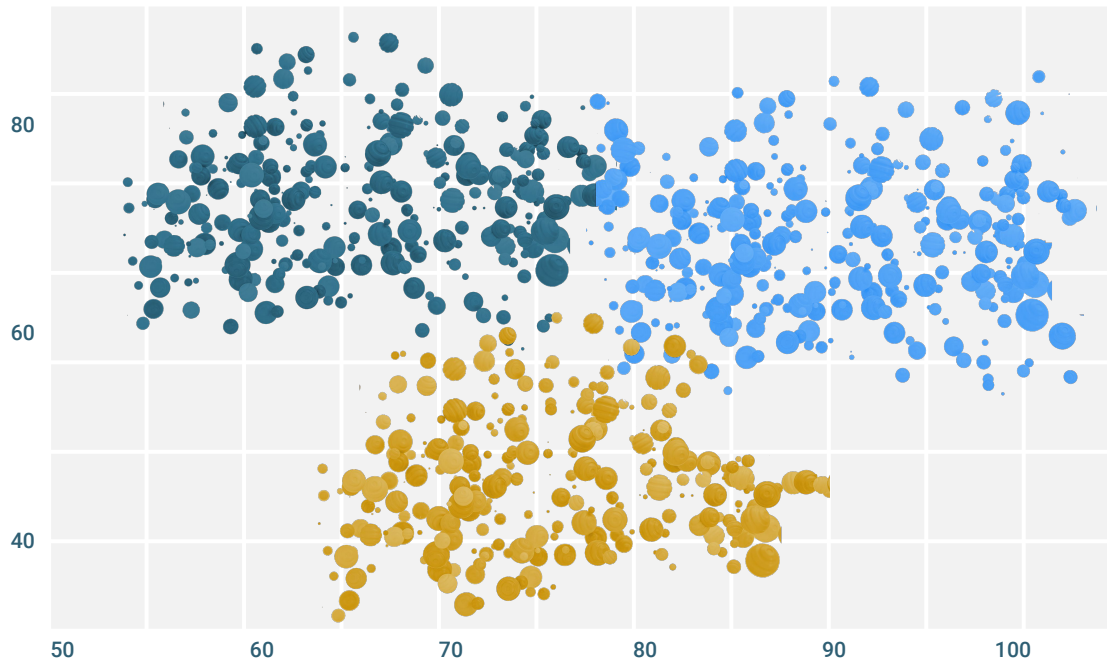


Clustering is grouping data together so that every member of that group is similar in some way.



Clustering Explained

Unsupervised learning models are often created using a clustering algorithm.





Instructor **Demonstration**

Clustering Explained

Clustering Explained

The process of clustering data points into groups is called **centering**.

01

In advanced analytics, centering helps to determine the number of classes or groups to create.

02

Centering improves the performance of logistic regression models by ensuring that all data points share the same starting mean value.

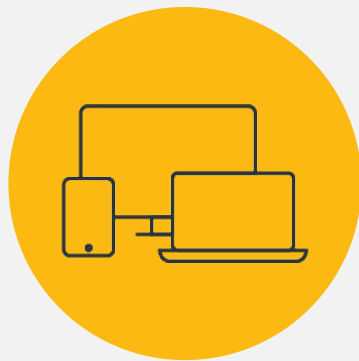
03

Data points with the same starting mean value are clustered together.



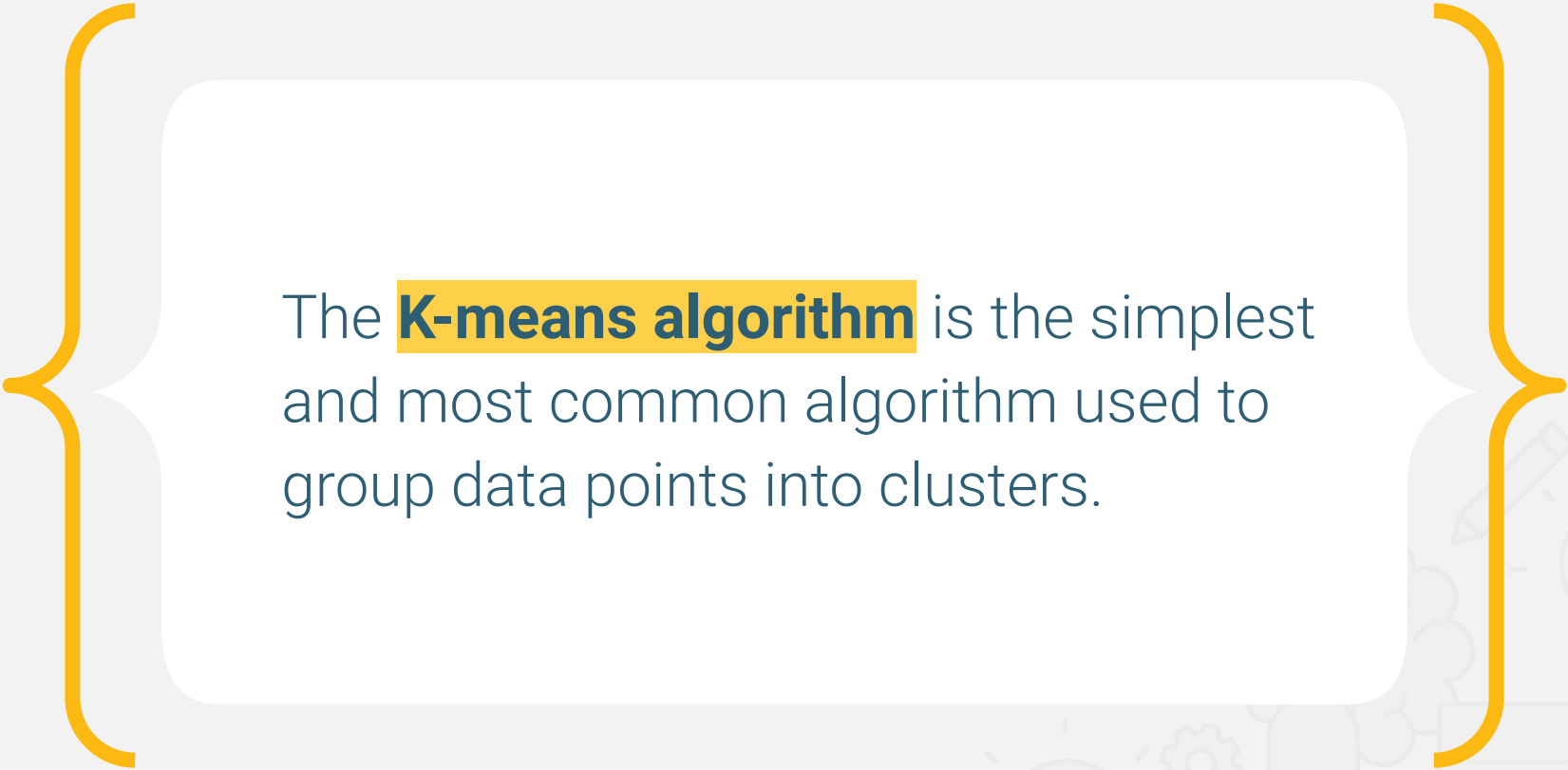
Questions?



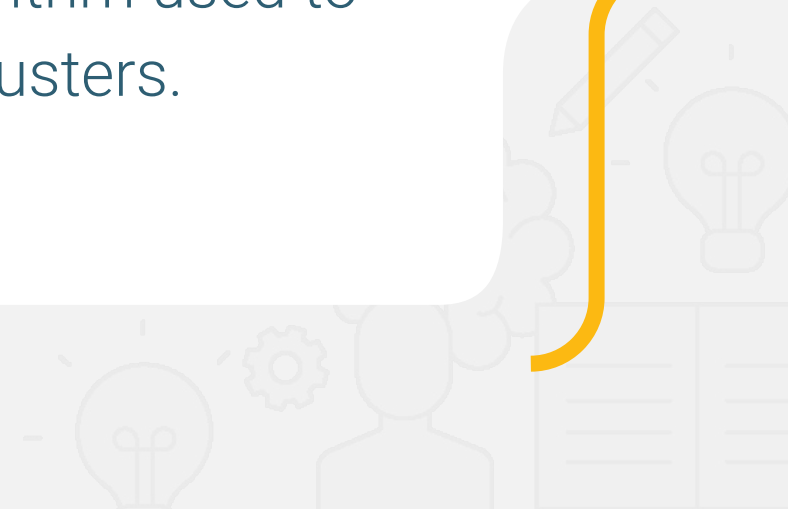


Instructor **Demonstration**

The K-Means Algorithm

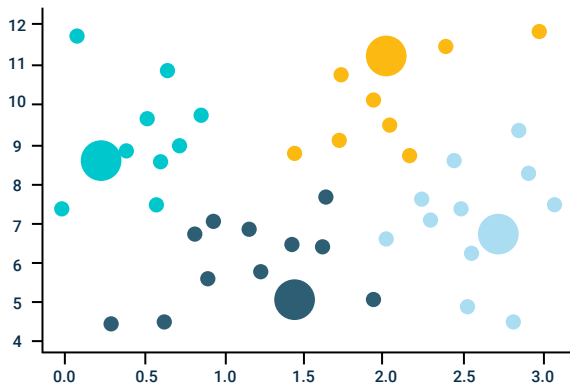


The **K-means algorithm** is the simplest and most common algorithm used to group data points into clusters.

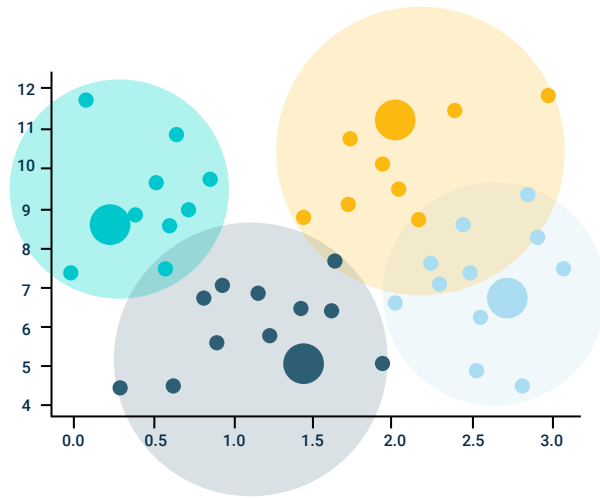


The K-Means Algorithm

K-means takes a predetermined number of clusters and then assigns each data point to one of those clusters.



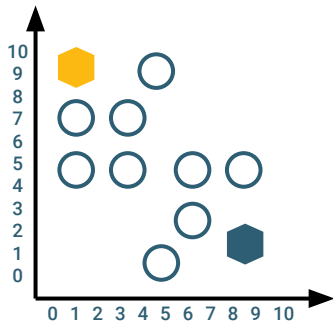
The algorithm assigns points to the closest cluster center.



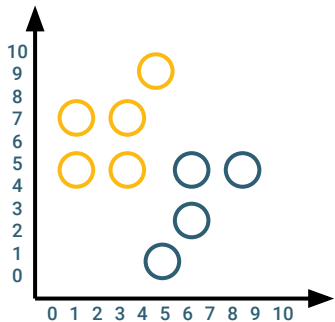
The algorithm readjusts the cluster's center by setting each center as the mean of all the data points contained within that cluster.

The K-Means Algorithm

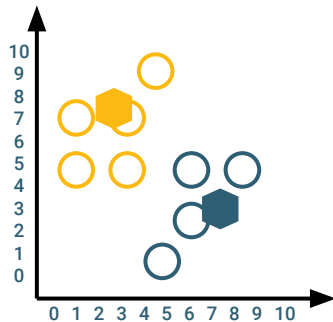
The K-means algorithm then repeats this process, again and again, each time getting a little bit better at separating the data points into distinct groups.



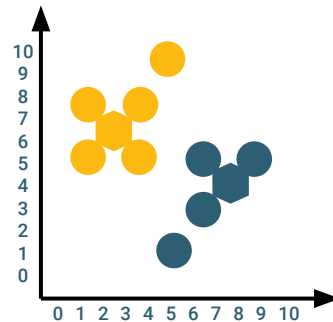
Randomly select k clusters



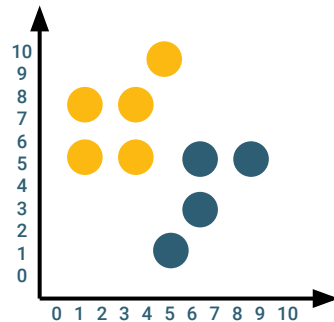
Each object assigned to similar centroid randomly



Cluster centers updated depending on new cluster mean



Reassign data points and update cluster centers



Reassign data points



Questions?





Activity:

Spending Beyond Our K-Means

In this activity, you will cluster the data into two different customer shopping segments and determine which segment reveals any relevant differences in customer shopping habits.

Suggested time:

20 minutes





Time's up!
Let's review



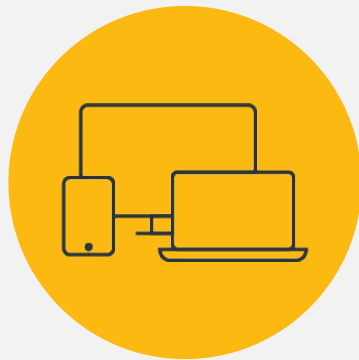
Questions?





Break

15 mins



Instructor **Demonstration**

Introduction to Clustering Optimization

Introduction to Clustering Optimization



The appropriate clustering algorithm and parameter settings depend on the individual dataset and intended use of the results.



Cluster analysis is not an automatic task.



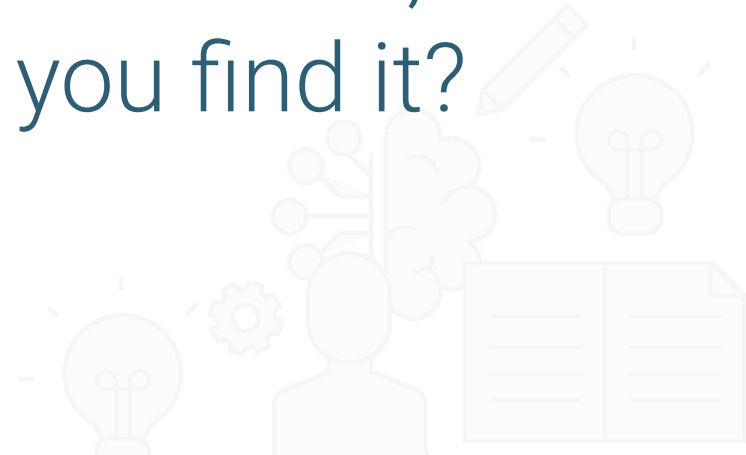
As a data professional, you will need to do some trial and error to find the optimal clusters.



This process includes modifying the data preprocessing and model parameters until the result achieves the desired properties.



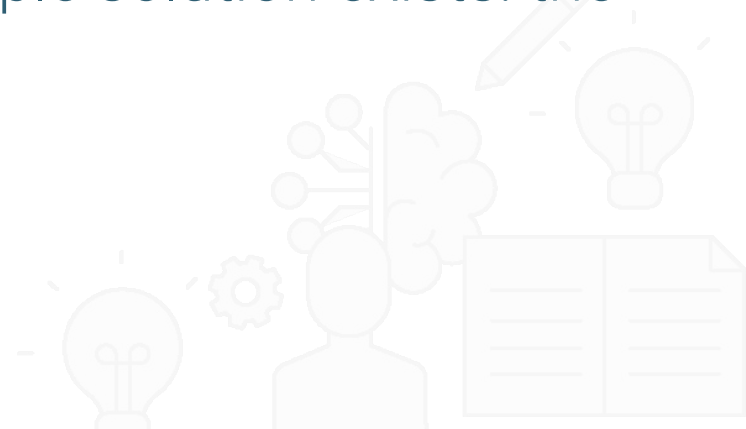
How do you know the optimal number of clusters, or **value of k** , and how do you find it?





One of the challenges of working with unlabeled data is the unknown number of existing segments, or clusters.

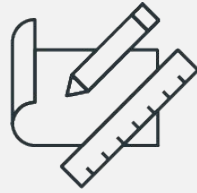
Fortunately, a simple solution exists: the **elbow method**.





Questions?



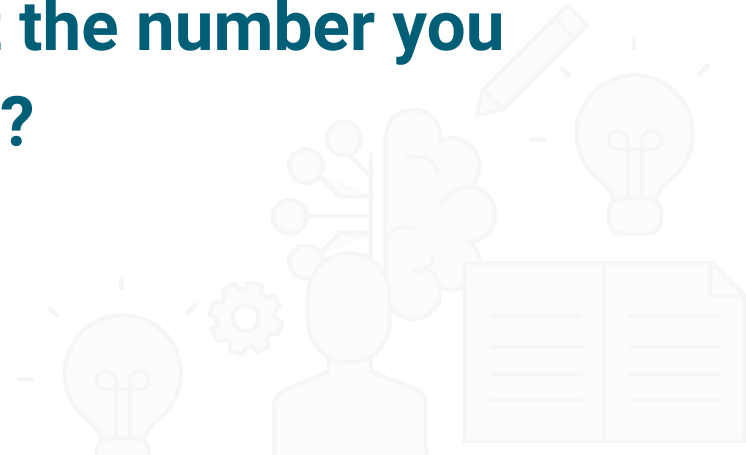


The **Elbow Method**



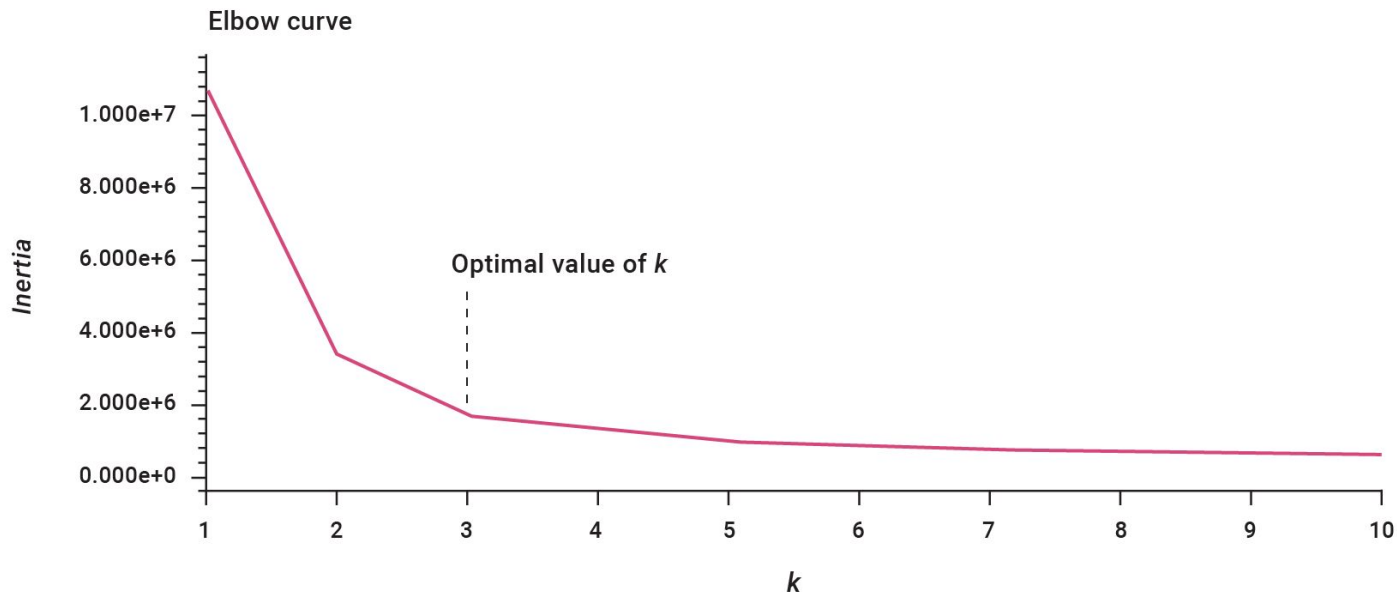


Since the K-means algorithm needs to have the number of clusters provided ahead of time, **how can you be sure that the number you chose is correct?**



The Elbow Method

One method to determine the optimal value of k , or the number of clusters in a dataset, is the **elbow method**.

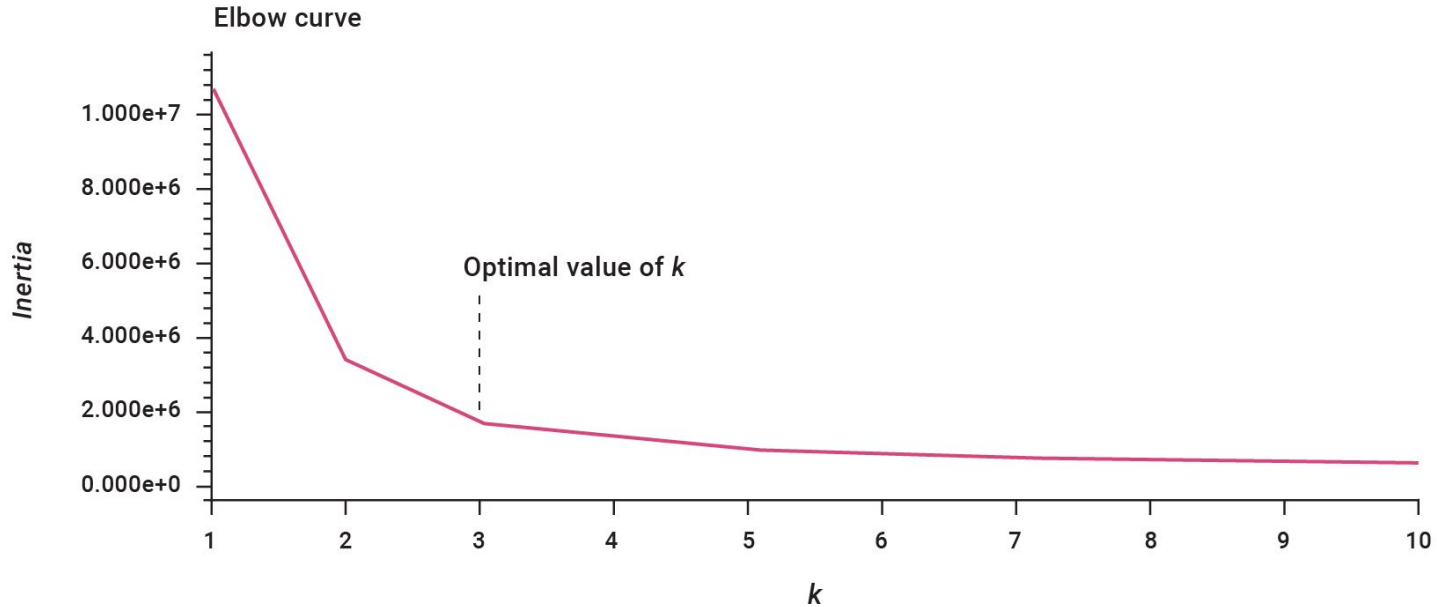


The elbow method runs the K-means algorithm for a range of possibilities for k , or the number of clusters.

The resulting elbow curve plots the number of clusters, x , versus an objective function called inertia.

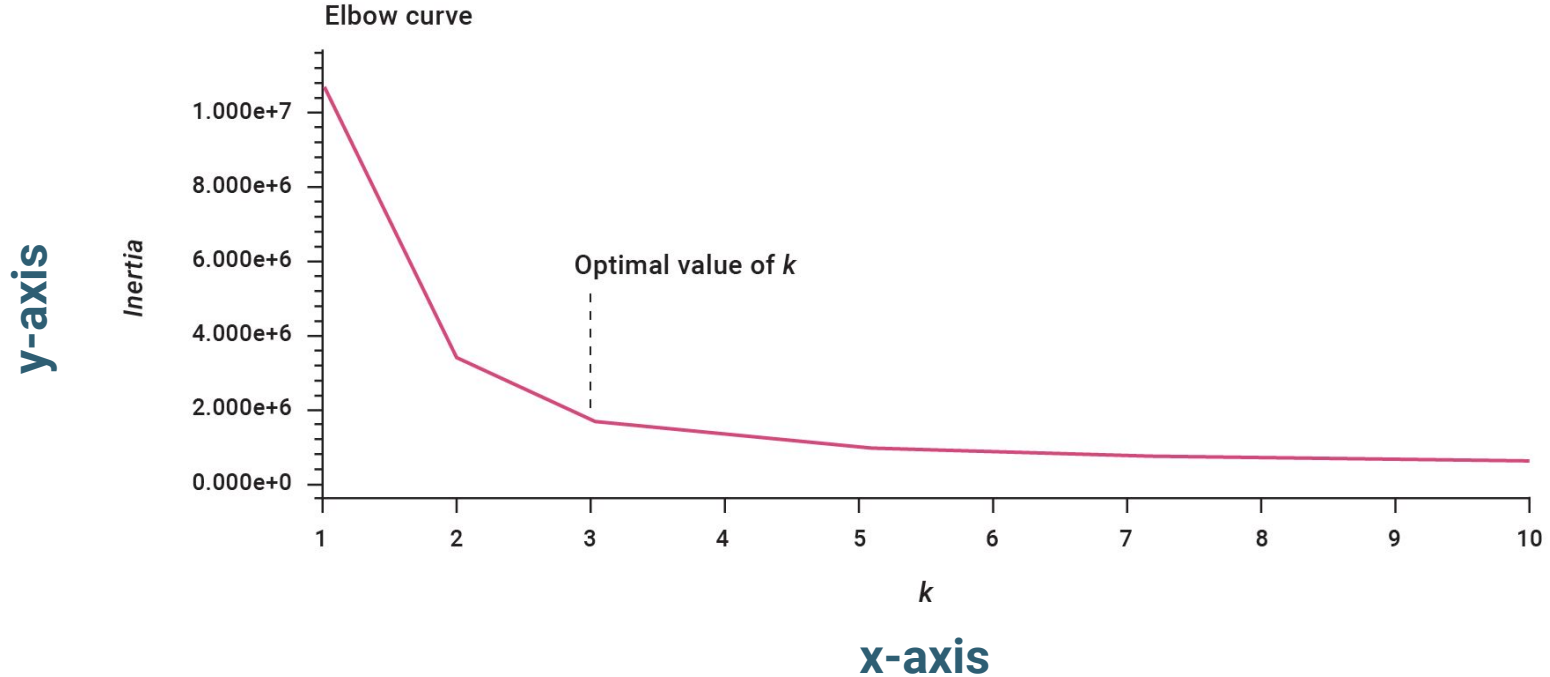
Elbow Curve

The **elbow curve** is commonly used to figure out the best value of k . It is essentially used to determine the number of clusters at which the data points become tightly clustered.



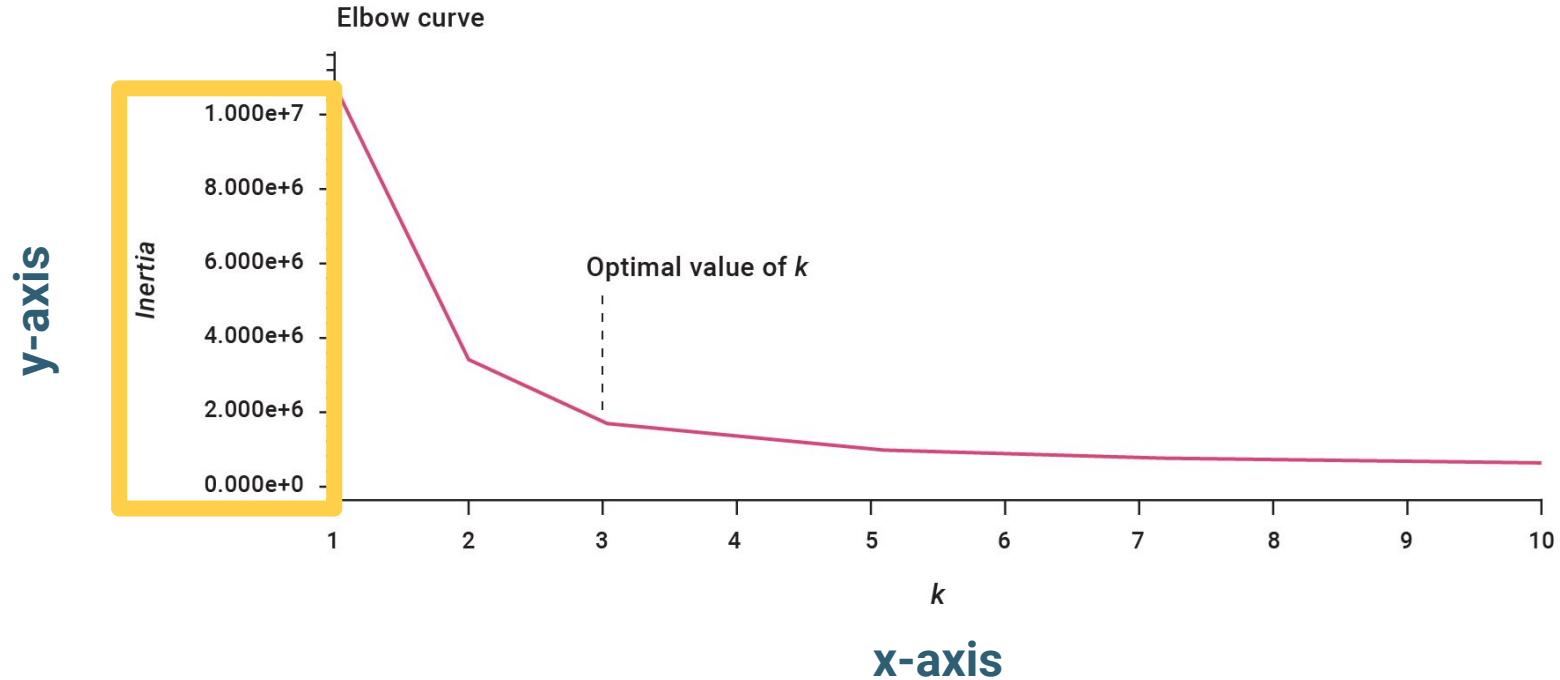
Elbow Curve

On the elbow curve, the x-axis is the value of clusters, while the y-axis is a metric used to assess the value of k .



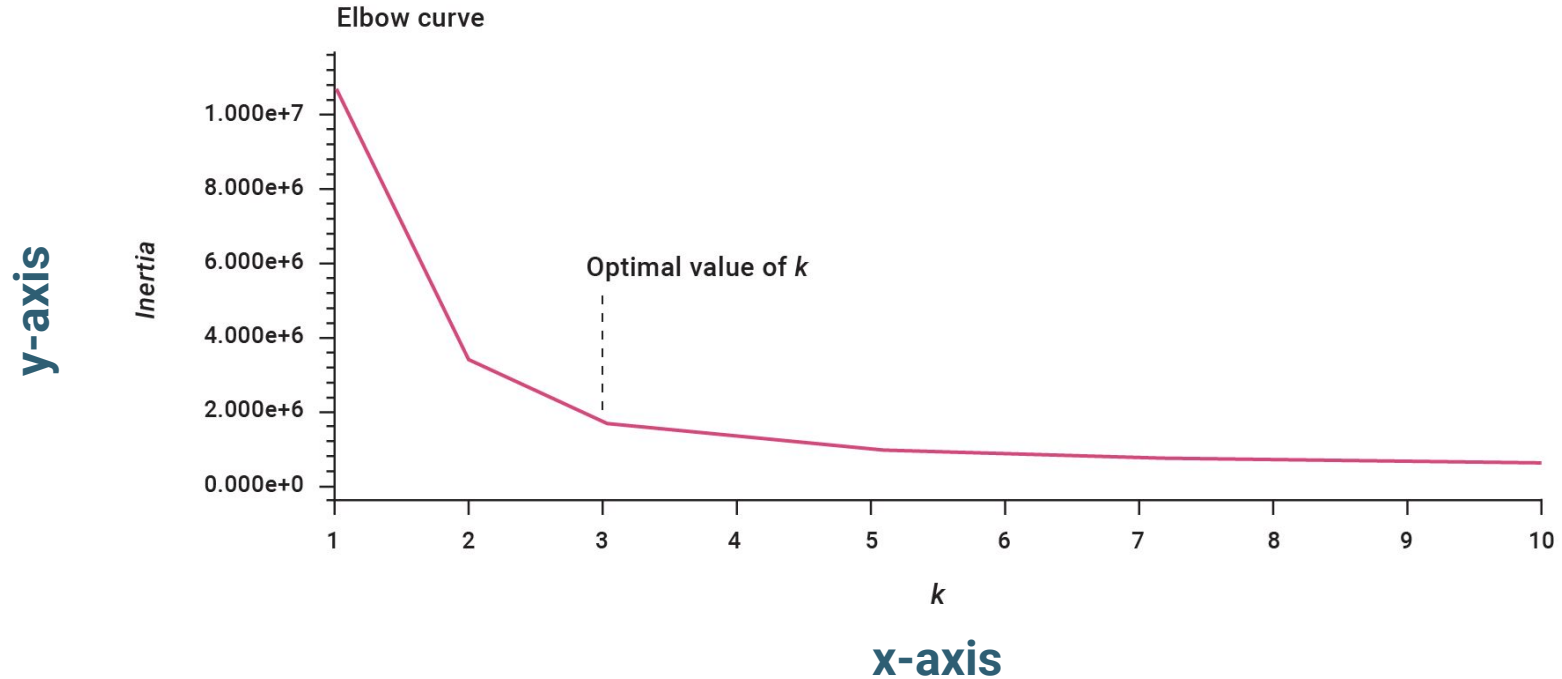
Elbow Curve

The **inertia** is commonly used as an objective function. It is the sum of the squared distances of samples to their closest cluster center.



Elbow Curve

A low inertia value means that the data points are tightly clustered around the cluster center.



Inertia

Inertia involves complicated math, but it is basically a measure of how concentrated the elements are in a dataset.

High Concentration

Datasets with a high concentration of elements (where elements are tightly grouped together) have a **low** inertia value.

This means that there is a **small standard deviation** for the elements in the cluster relative to the cluster mean value.

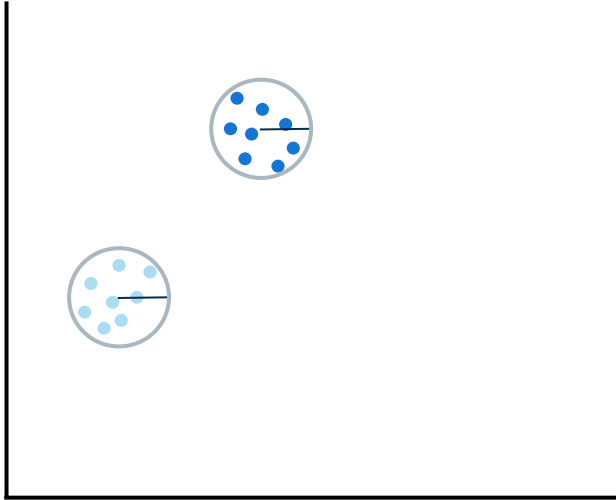
Low Concentration

Datasets with a low concentration of elements (where elements are spread out) have a **high** inertia value.

This means that there is a **high standard deviation** for the elements in the cluster relative to the cluster mean value.

Low Inertia

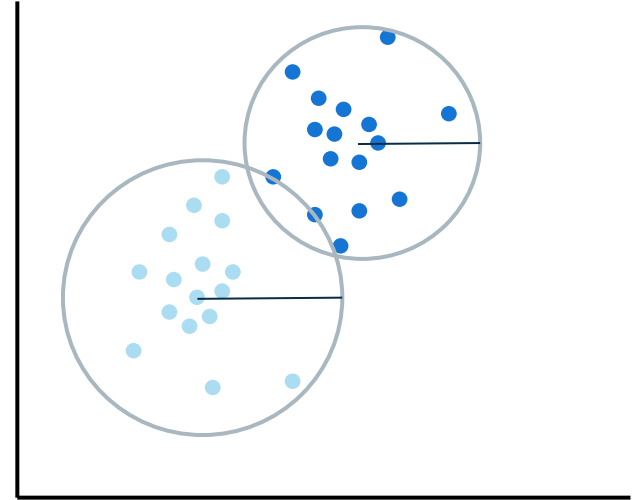
Radius of circle is small = small standard deviation from cluster mean



vs.

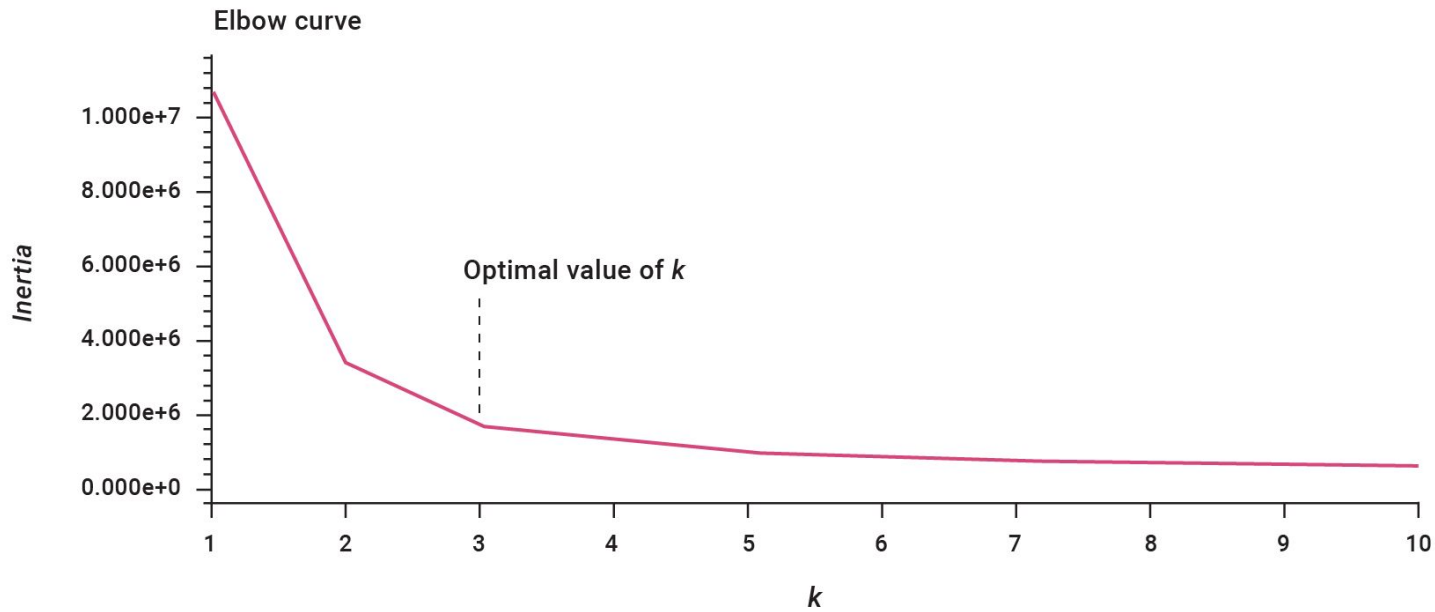
High Inertia

Radius of circle is large = large standard deviation from cluster mean



The Elbow Method

The goal is to find a value for k that corresponds to a measure of inertia that shows minimal change for each additional cluster (or value of k) that is added to the dataset. **The spot is indicated by the bend in the elbow.**





Instructor **Demonstration**

Introduction to Clustering Optimization



Questions?





Activity:

Finding the Best k

In this activity, you will use the elbow method to determine the optimal number of clusters that should be used to segment a dataset of stock pricing information.

Suggested time:

25 minutes



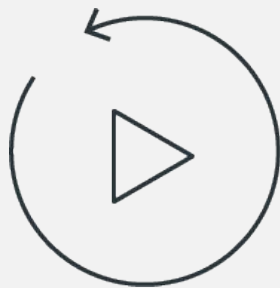


Time's up!
Let's review



Questions?





Let's **recap**



Review the Class Objectives

In this lesson, you learned how to:

- 1 Recognize the differences between supervised and unsupervised machine learning.
- 2 Define clustering and how it is used in data science.
- 3 Apply the K-means algorithm to identify clusters in a given dataset.
- 4 Determine the optimal number of clusters for a dataset using the elbow method.



Next

In the next lesson, you will learn how to preprocess the data that goes into these types of models, and you'll create models that can adapt and perform better on more complex types of data.



Questions?





The End