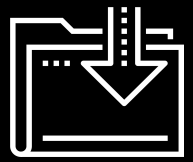
Introduction to Machine Learning

Data Boot Camp

Lesson 19.1



Class Objectives

By the end of this lesson, you will be able to:

Recognize the differences between supervised and unsupervised machine learning (ML).

Define clustering and how it is used in data analytics.

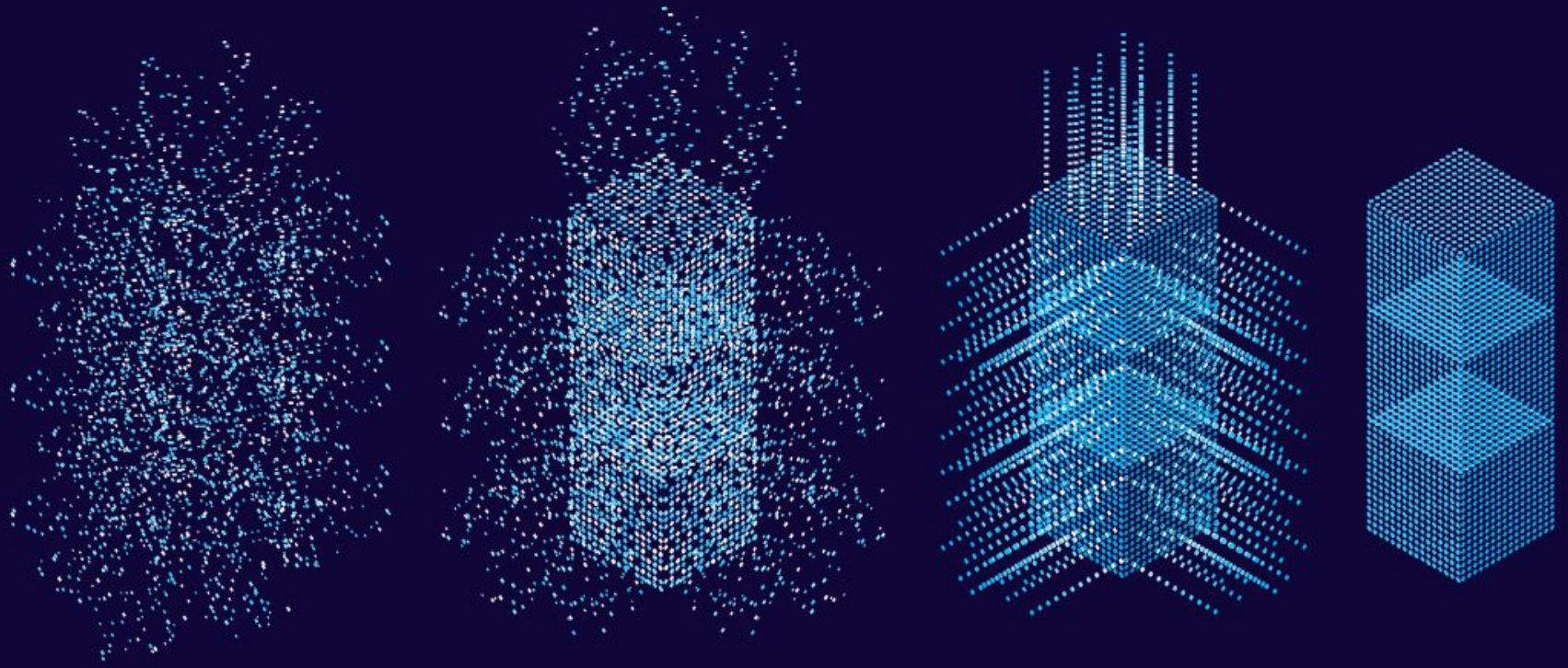
Apply the K-means algorithm to identify clusters in a given dataset.

Use the elbow method to determine the optimal number of clusters for a dataset.


WELCOME

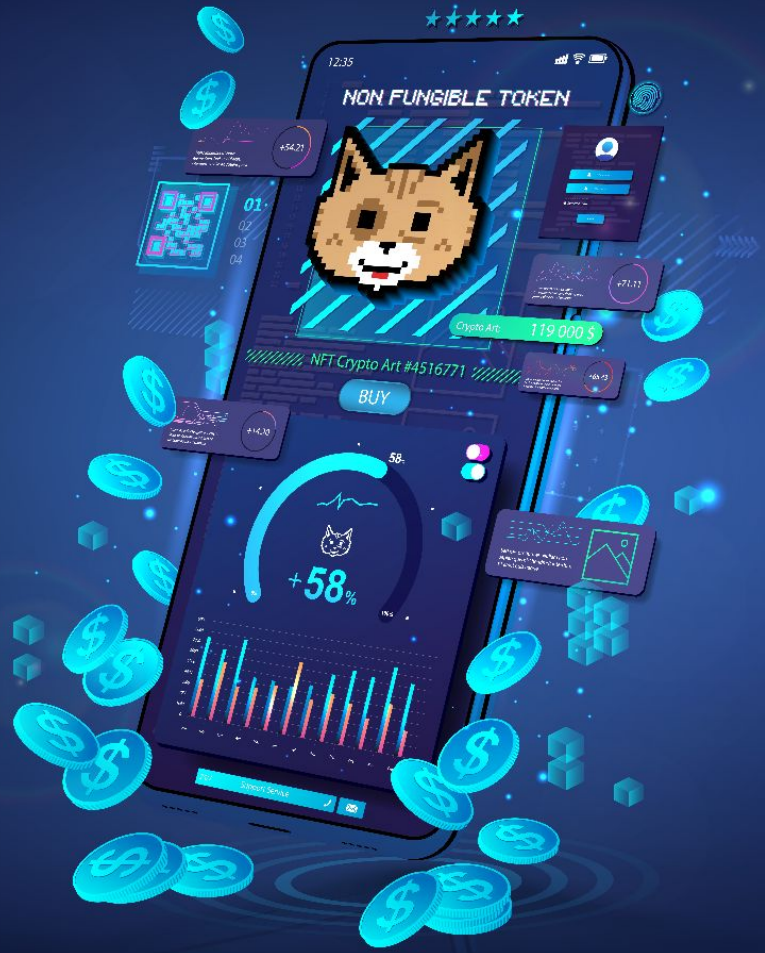
Demystifying Machine Learning

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



Today, machine learning is changing industries at an unprecedented pace.

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



This Week's Challenge Assignment

Create a machine learning model that groups cryptocurrencies to assemble investment portfolios that are based on the profitability of those cryptocurrencies.

The Mysticism of Machine Learning

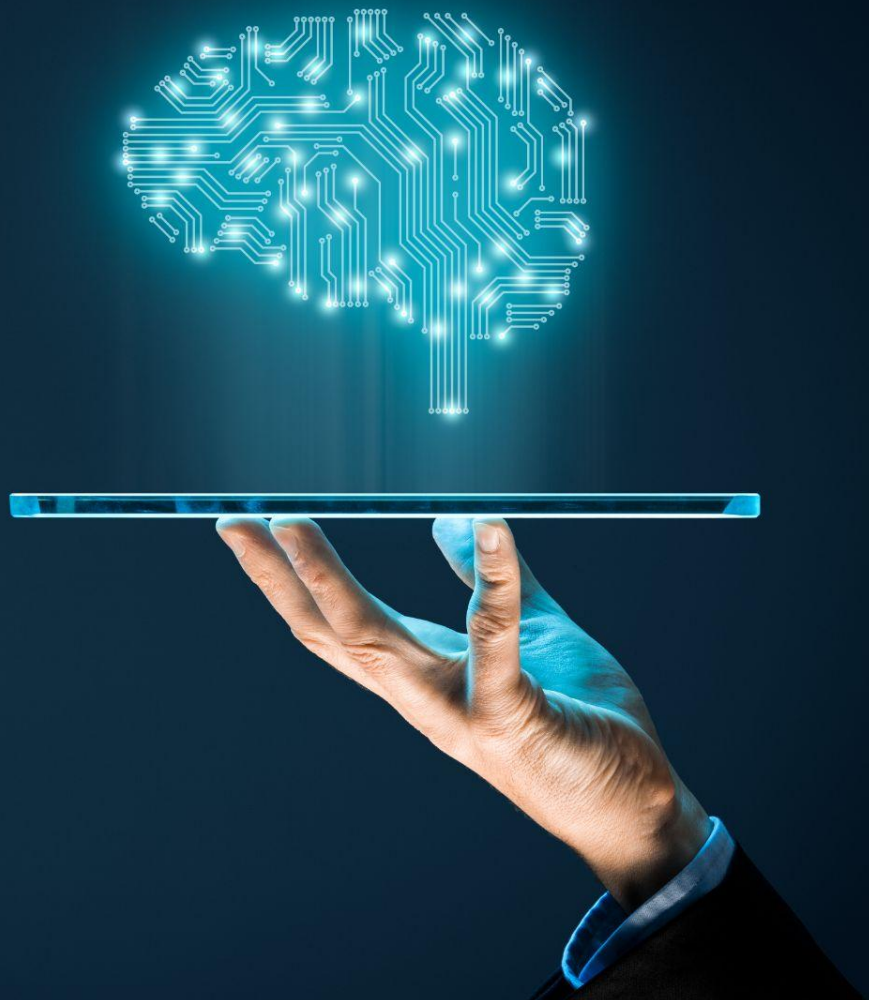
Despite the mainstream use of the term “machine learning,” most people still don't know what machine learning *really* is.

Machine learning 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

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, and then they 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, you can 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 Supervised learning

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

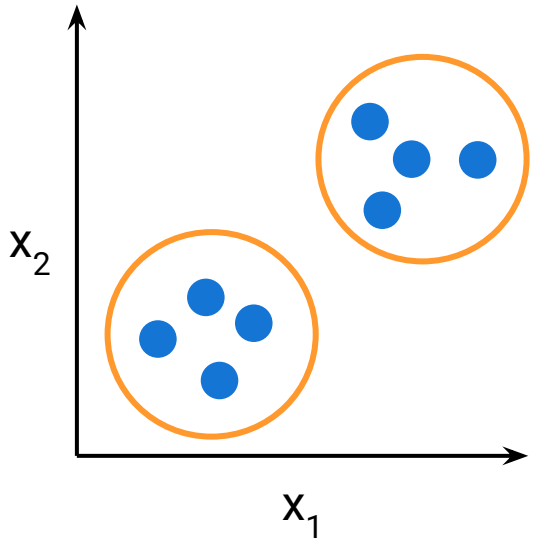
This provides an answer key that can be used to evaluate the accuracy of the training data.

02 Unsupervised learning

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

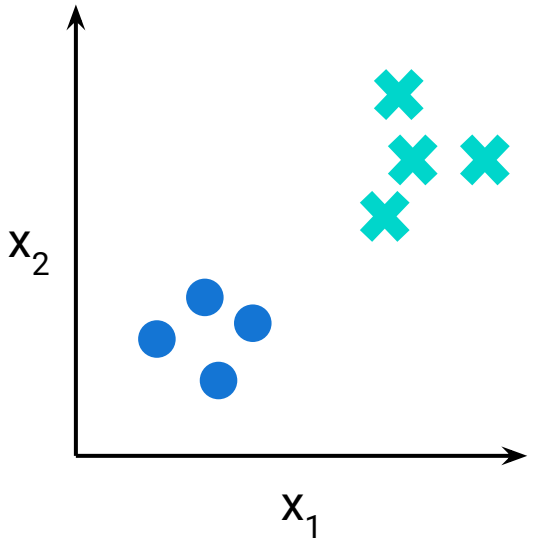
Supervised Learning vs. Unsupervised Learning

Supervised Learning



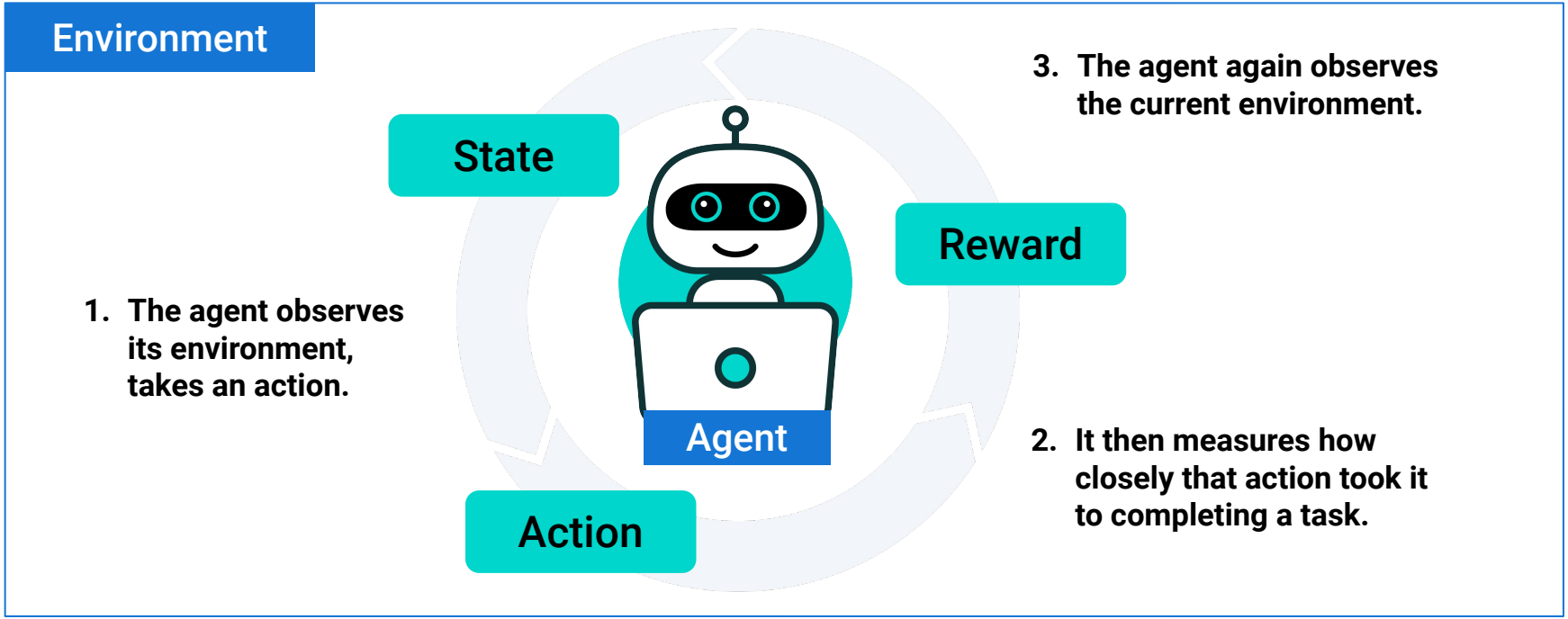
vs.

Unsupervised Learning

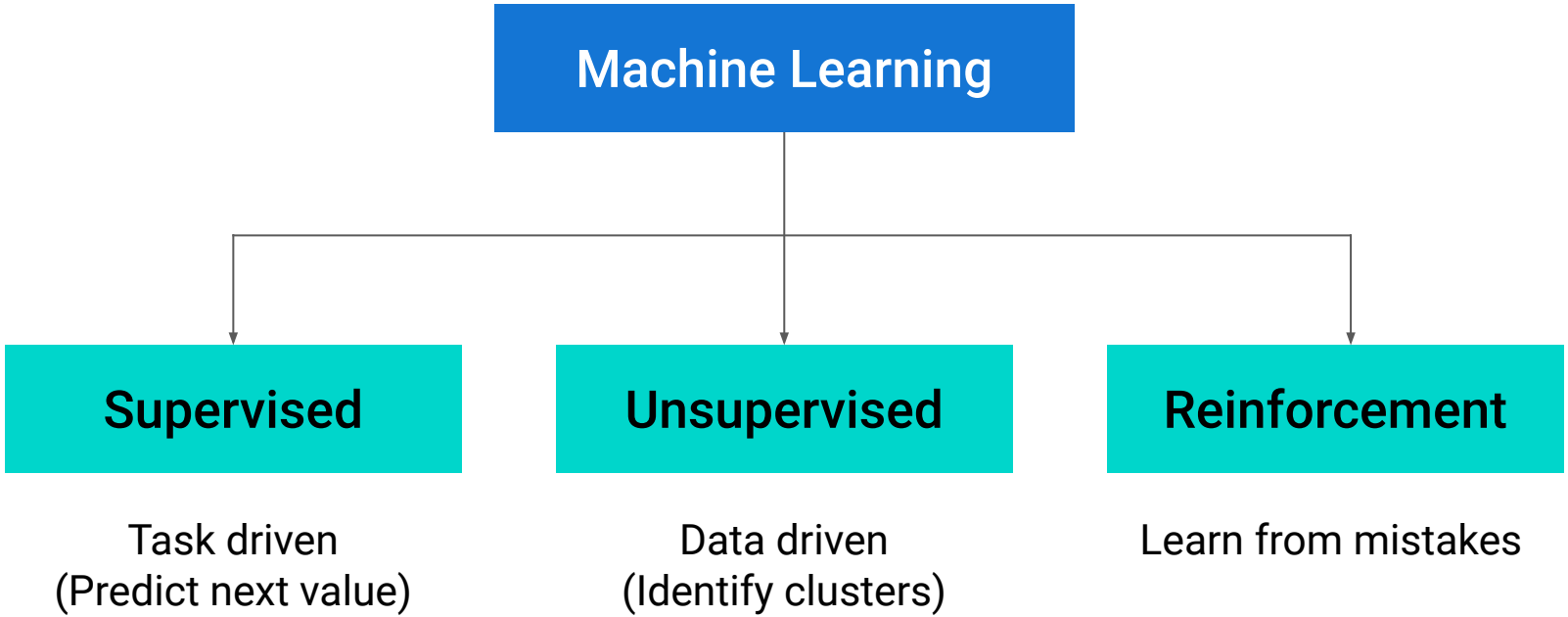


Reinforcement Learning

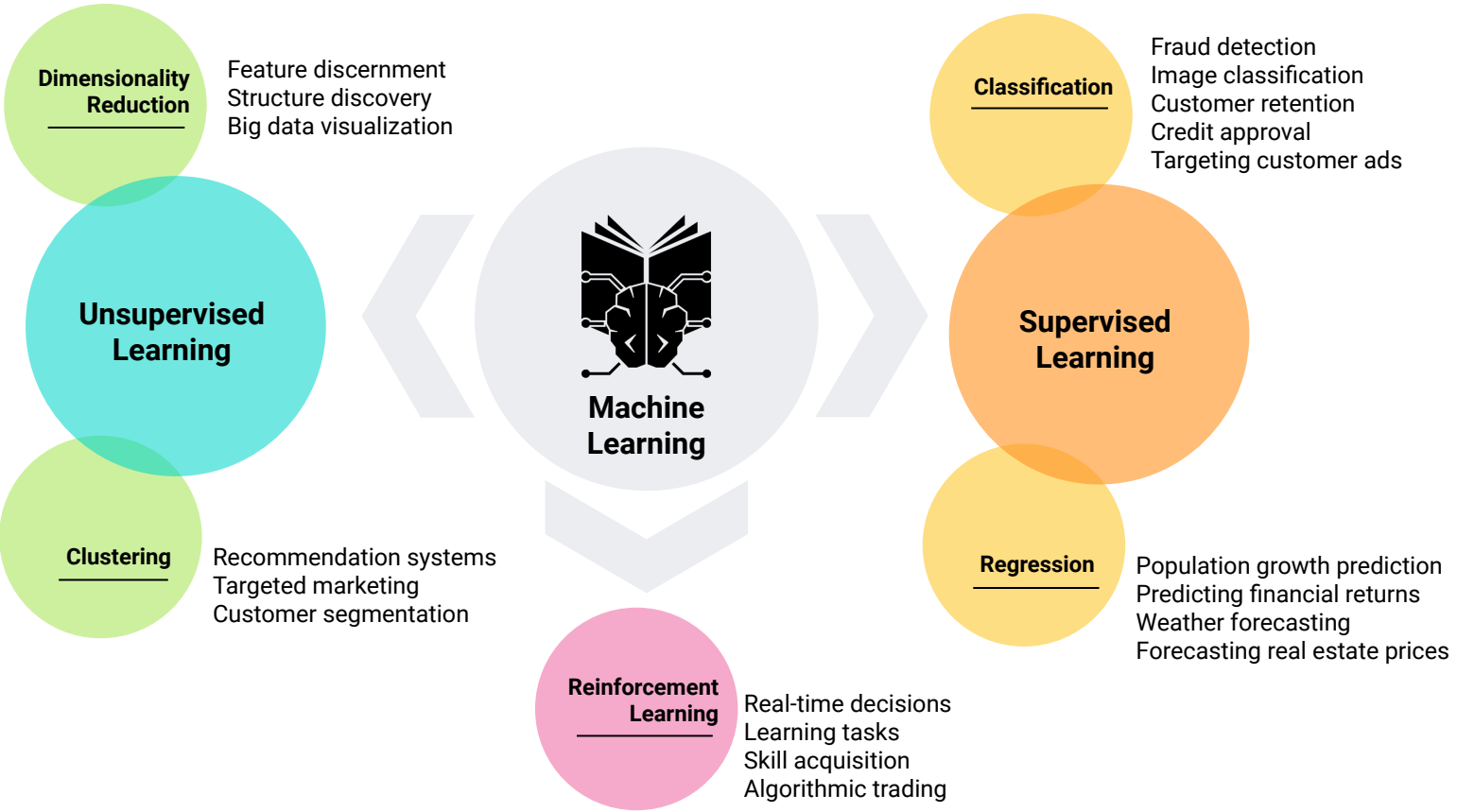
This third type of machine learning algorithm is used less frequently but still has important applications in data analytics.



Three Types of ML



Types of ML


Machine Learning
Unsupervised Learning
Dimensionality Reduction
Feature discernment
Structure discovery
Big data visualization
Clustering
Recommendation systems
Targeted marketing
Customer segmentation
Supervised Learning
Classification
Fraud detection
Image classification
Customer retention
Credit approval
Targeting customer ads
Regression
Population growth prediction
Predicting financial returns
Weather forecasting
Forecasting real estate prices
Reinforcement Learning
Real-time decisions
Learning tasks
Skill acquisition
Algorithmic trading

Types of ML

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

scikit learn

NumPy

pandas

SciPy

matplotlib

TensorFlow

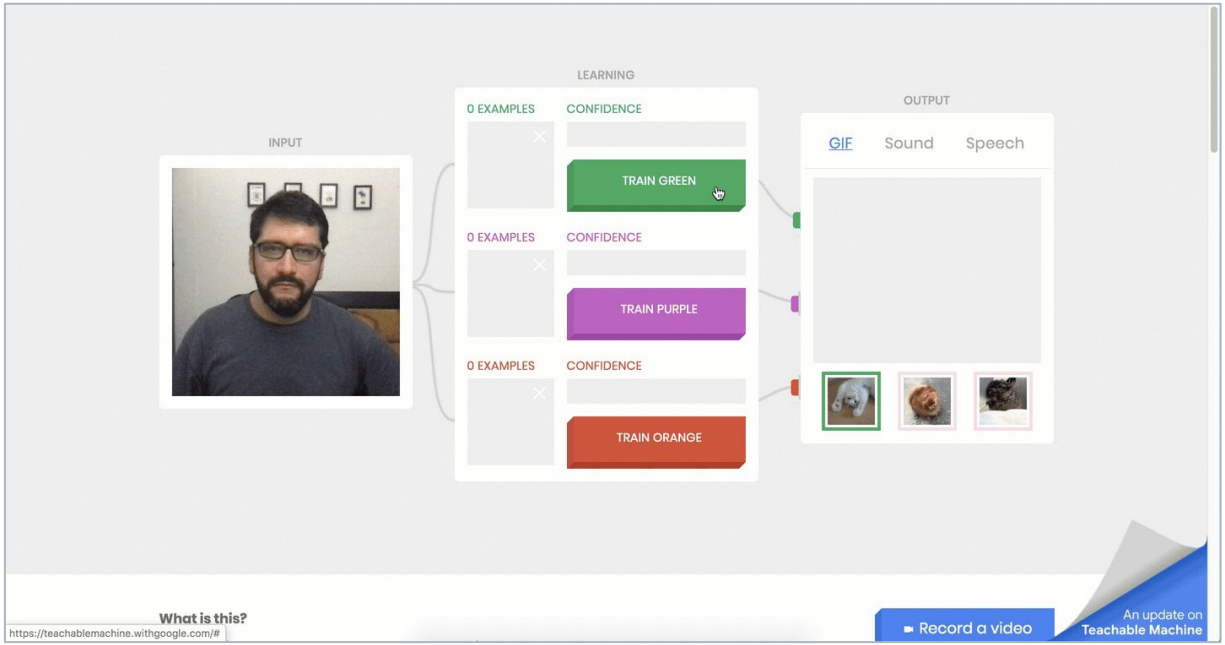
Questions?

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.

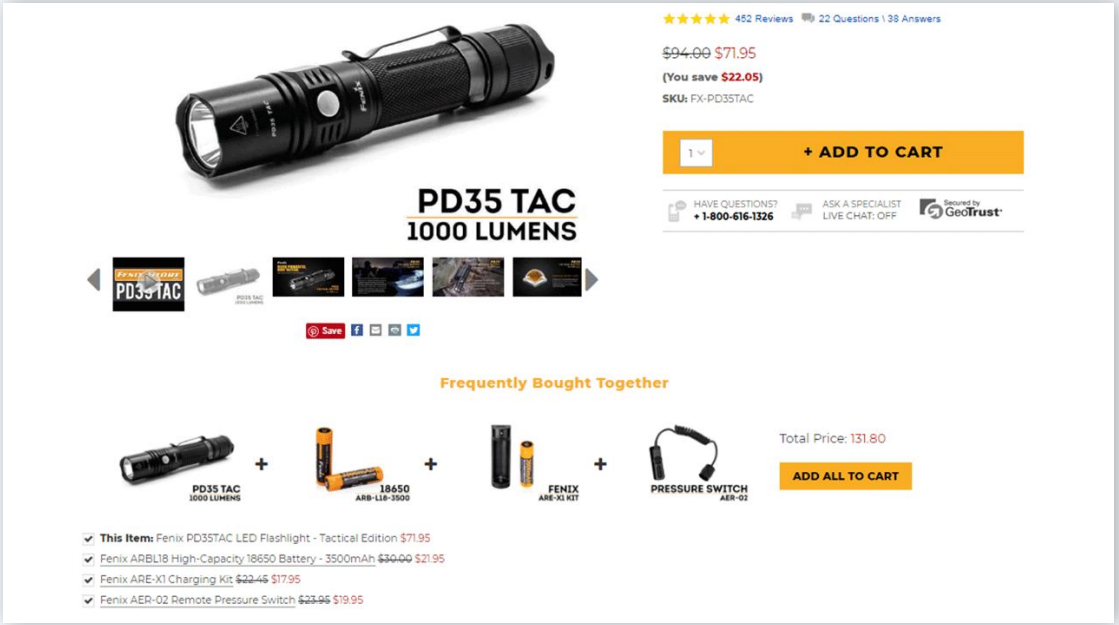


Introduction to Unsupervised Learning

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.



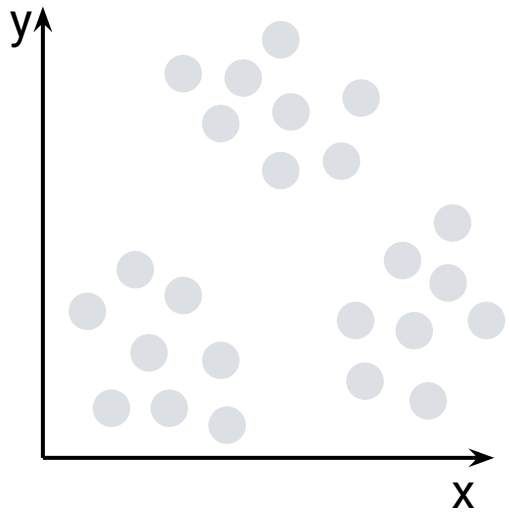
This power to recognize data patterns has broad applications in data analytics.

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

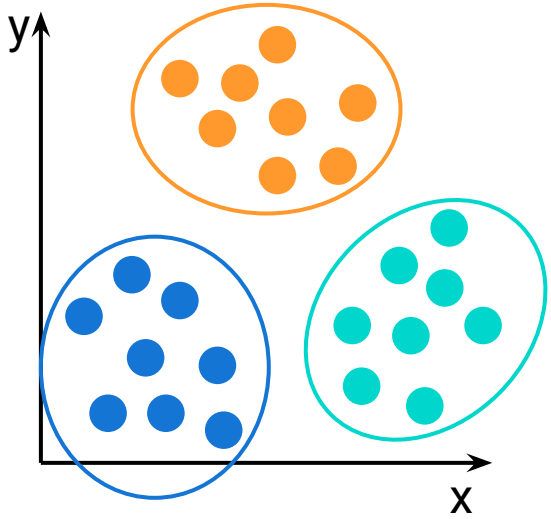
Introduction to Unsupervised Learning

The **K-means algorithm** is used for marketing use cases because of its ability to segment customers for more targeted ads.

Before K-means



After K-means



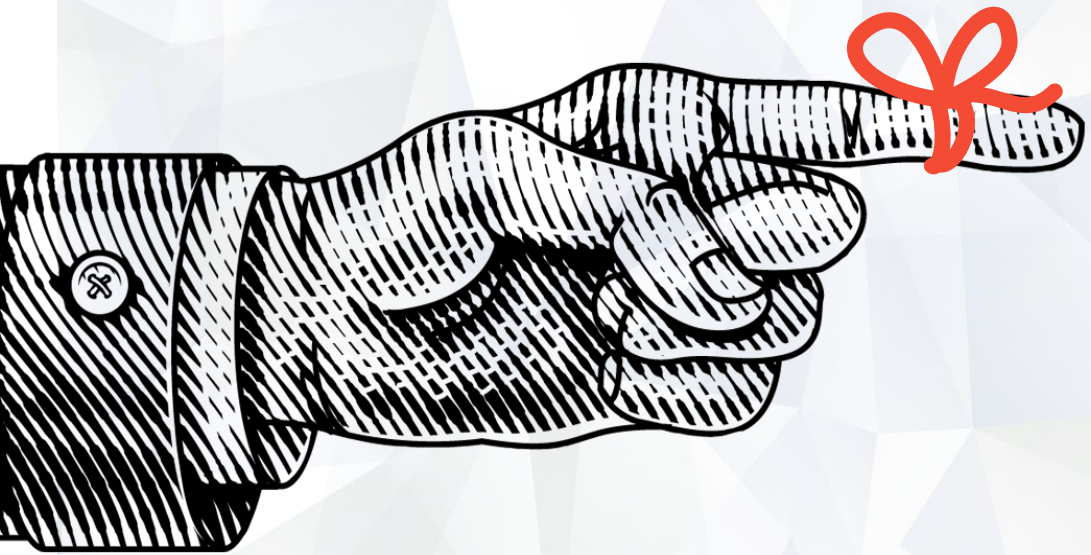
Introduction to Unsupervised Learning

In the lesson:

You'll apply unsupervised learning by using the K-means algorithm to define customer segments, or clusters.

In the Challenge assignment:

You'll apply the K-means algorithm to cluster data and prepare a cryptocurrency portfolio proposal to the company's board of directors.



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:

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

The algorithm is creating its own categories for the data, so an expert is needed to determine if these categories are meaningful.

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 clustering be used by businesses?

One possible answer:

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

How might anomaly detection be used by credit card companies?

One possible answer:

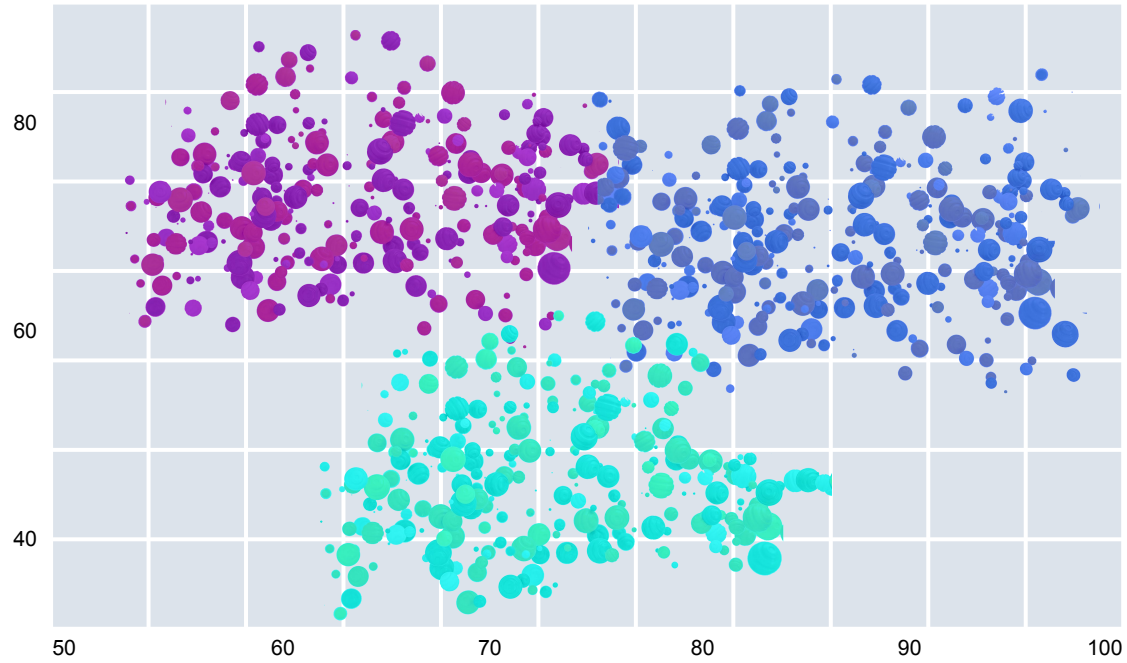
Anomaly detection can be used to detect potentially fraudulent credit card transactions by grouping transactions into “normal” or “abnormal.”

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**.

- In advanced analytics, centering helps to determine the number of classes or groups to create.
- Centering improves the performance of logistic regression models by ensuring that all data points share the same starting mean value.
- Data points with the same starting mean value are clustered together.

Questions?

Time to Code

The K-means Algorithm

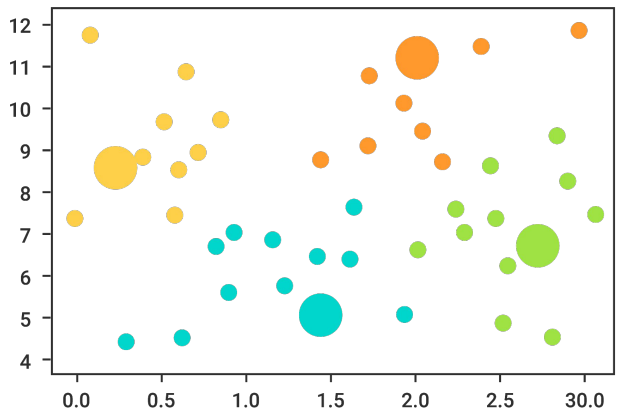
Suggested Time:

15 Minutes

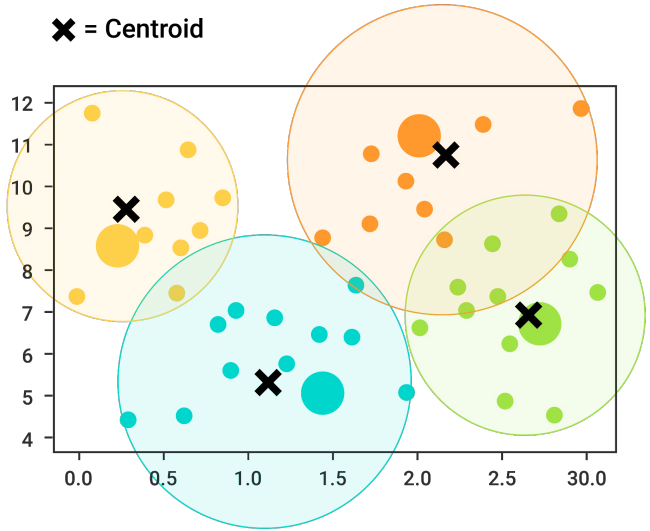
The **K-means algorithm** is the simplest and most common algorithm used to group data points into clusters.

The K-means Algorithm

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



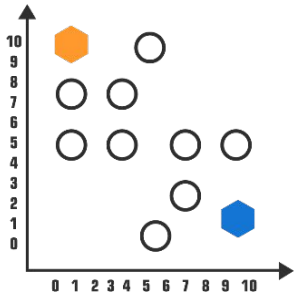
The algorithm assigns points to the closest cluster center.



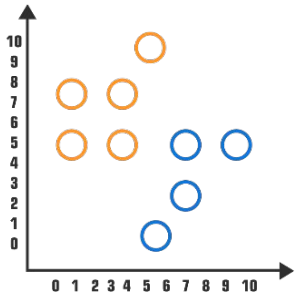
The algorithm readjusts the cluster's center by setting each center as the mean of all the data points contained within that cluster.

The K-means Algorithm

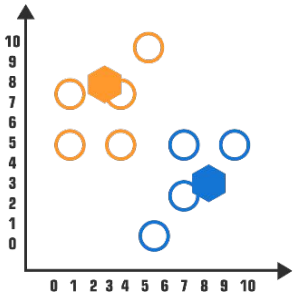
The K-means algorithm then repeats this process, again and again, each time getting a little bit better at separating the data points into distinct groups.



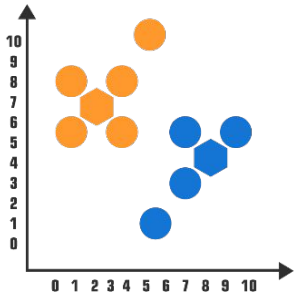
Randomly select K-clusters



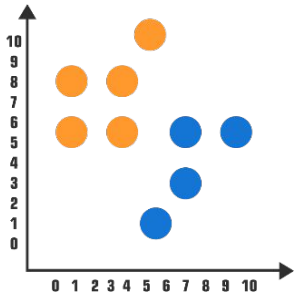
Each object assigned to similar centroid randomly



Cluster centers updating depending on renewed cluster mean



Reassign data points; update cluster centers



Reassign data points

Questions?

Activity: Segmenting Customers

In this activity, you will use the K-means algorithm to segment customer data for mobile versus in-person banking service ratings.

Suggested Time:

20 Minutes

Time's Up! Let's Review.

Questions?


Break

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, called the **elbow method**.

Questions?

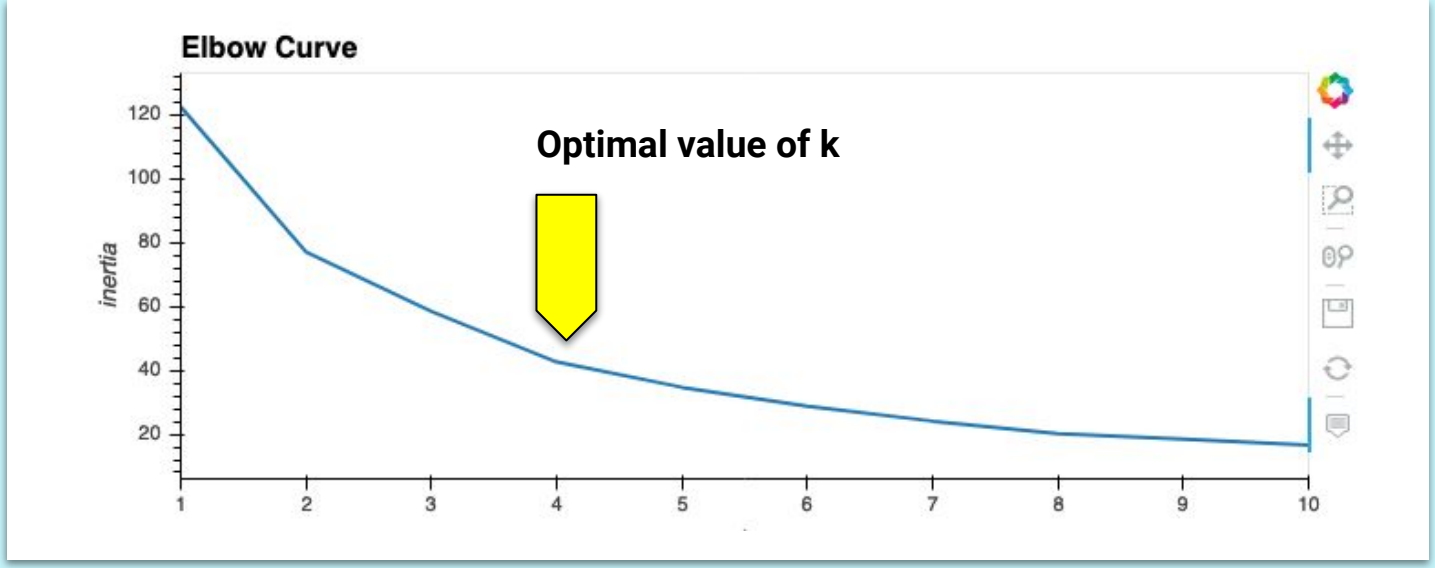
The Elbow Method

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

The Elbow Method

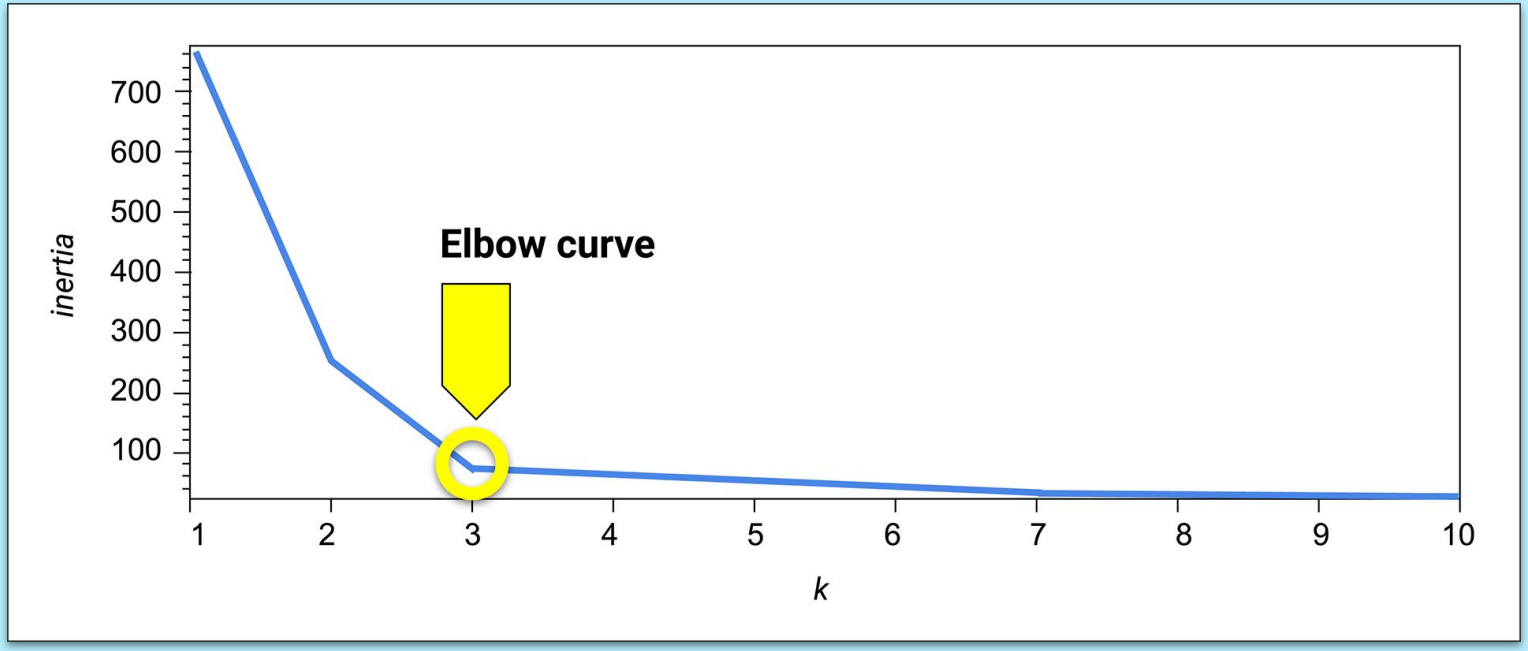
One method for determining the optimal value of k, or the number of clusters in a dataset, is the **elbow method**.

- 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

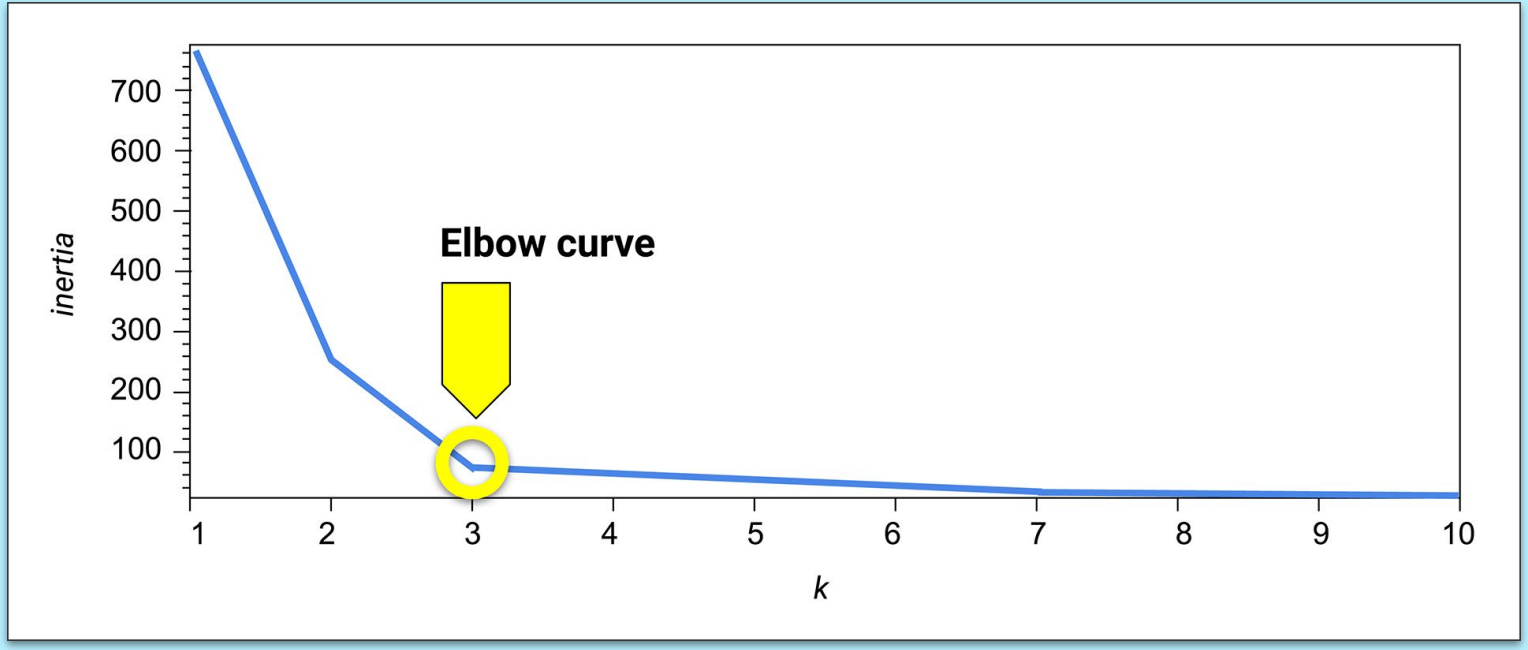
The **elbow curve** is commonly used to figure out the best value of k.
It is essentially used to determine the number of clusters at which the data points become tightly clustered.



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**.

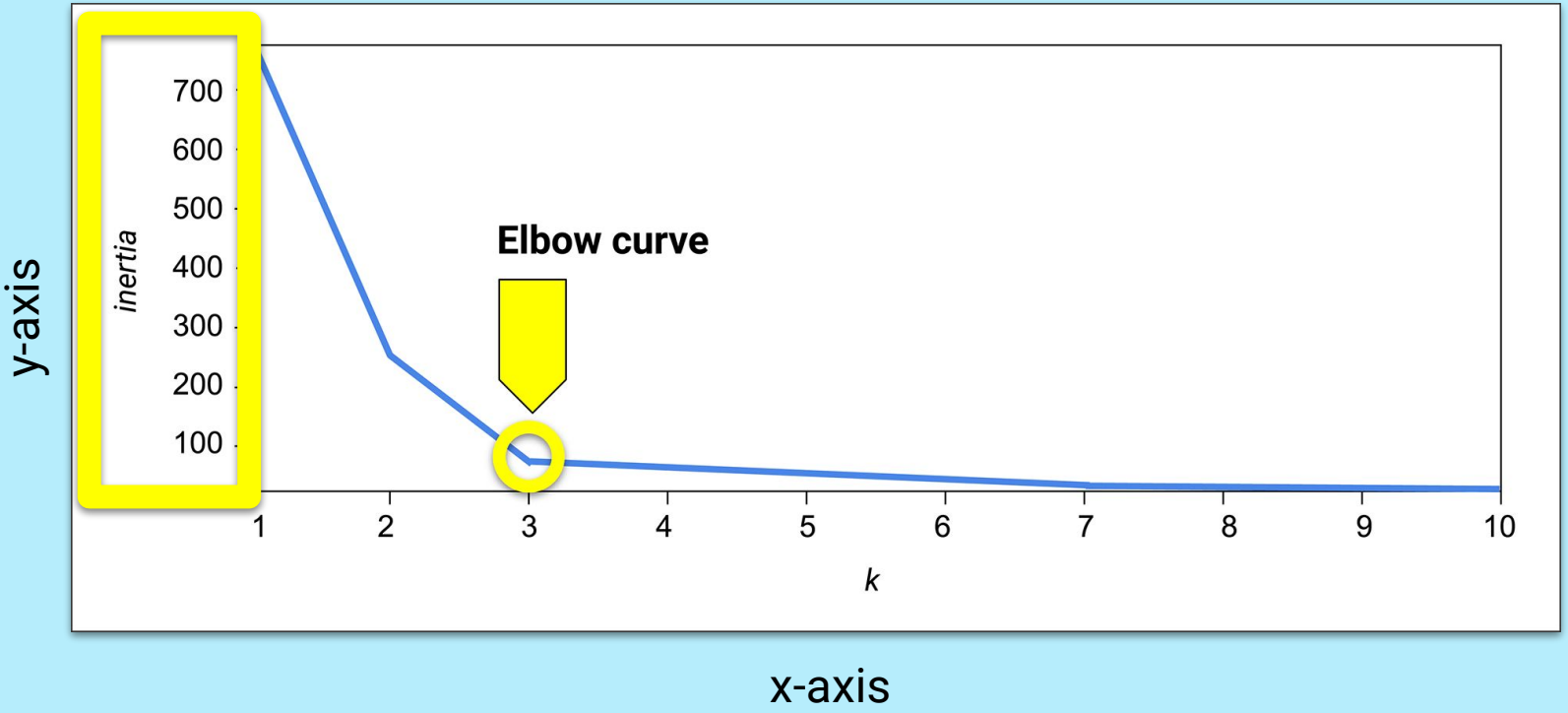
y-axis



x-axis

Elbow Curve

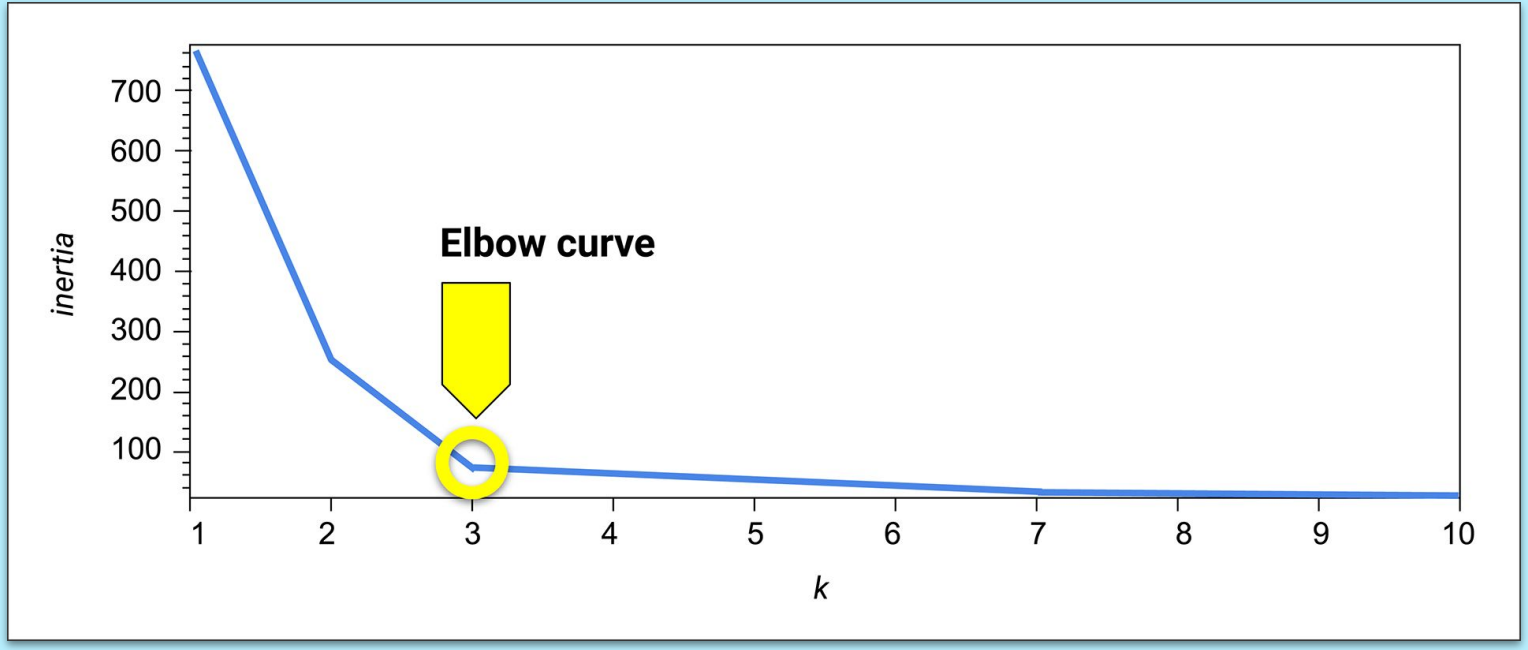
The **inertia** is commonly used as an objective function. It is the sum of the squared distances of samples to their closest cluster center.



Elbow Curve

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

y-axis



x-axis

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.

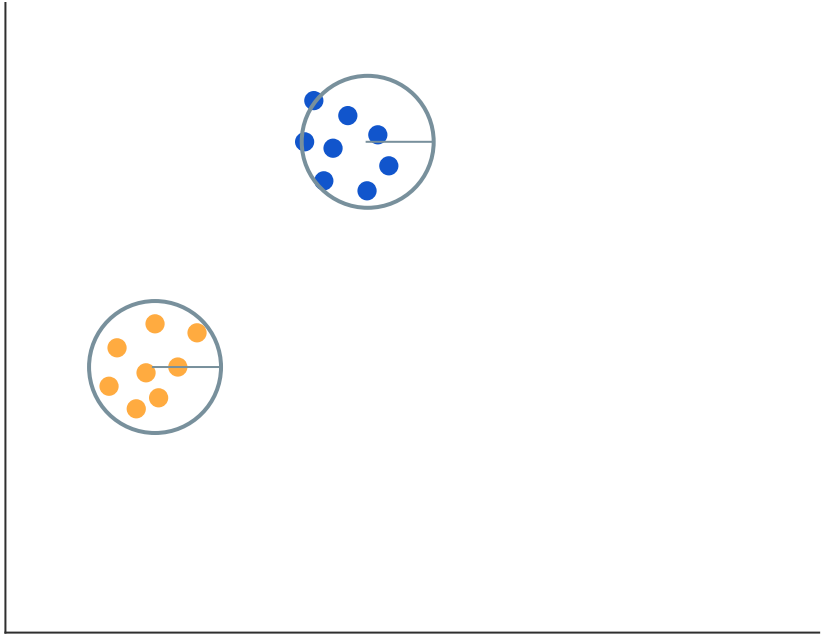
Low concentration

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

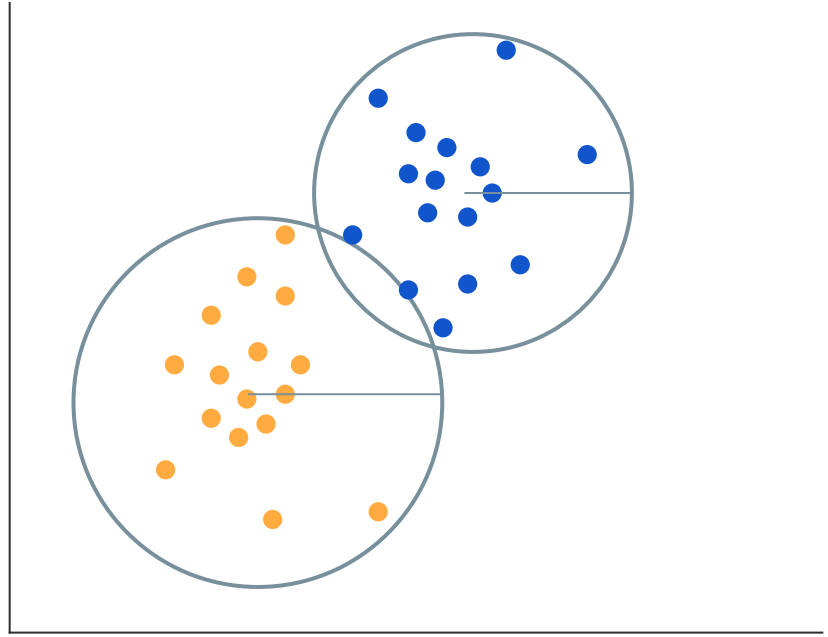
Radius of circle is small =
small standard deviation from cluster mean



vs.

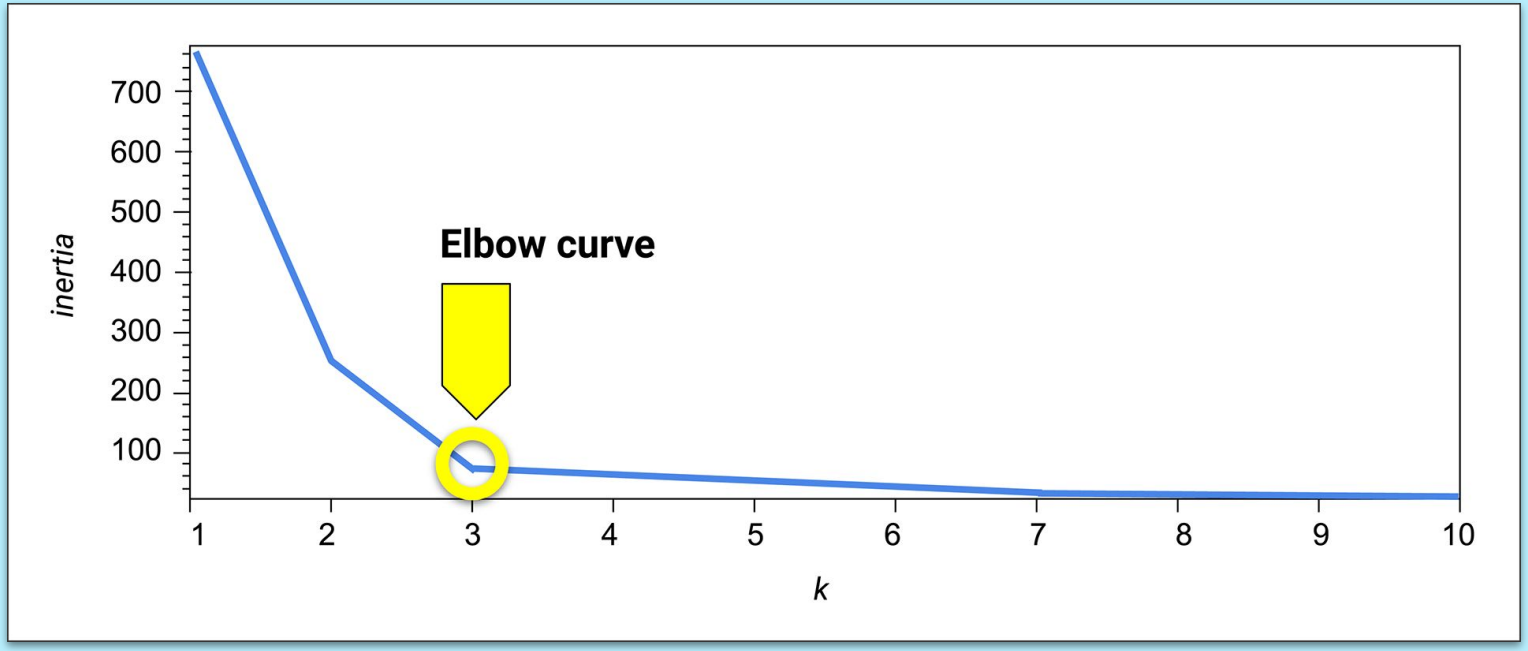
High Inertia

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**.



Time to Code

The Elbow Method

Suggested Time:

20 Minutes

Questions?

Activity: Finding 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?


Recap

Recap

Congratulations on acquiring your first machine learning skills! You can now:

- Explain the differences between supervised and unsupervised machine learning.
- Describe the purpose of clustering and how it's used in data analytics.
- Use the K-means algorithm to identify dataset clusters, and how to optimize this algorithm by using the elbow method.
- Optimize the K-means algorithm by using the elbow method.

Next Class

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