Introduction to Machine Learning

Data Boot Camp

Lesson 19.1


WELCOME

Class Objectives

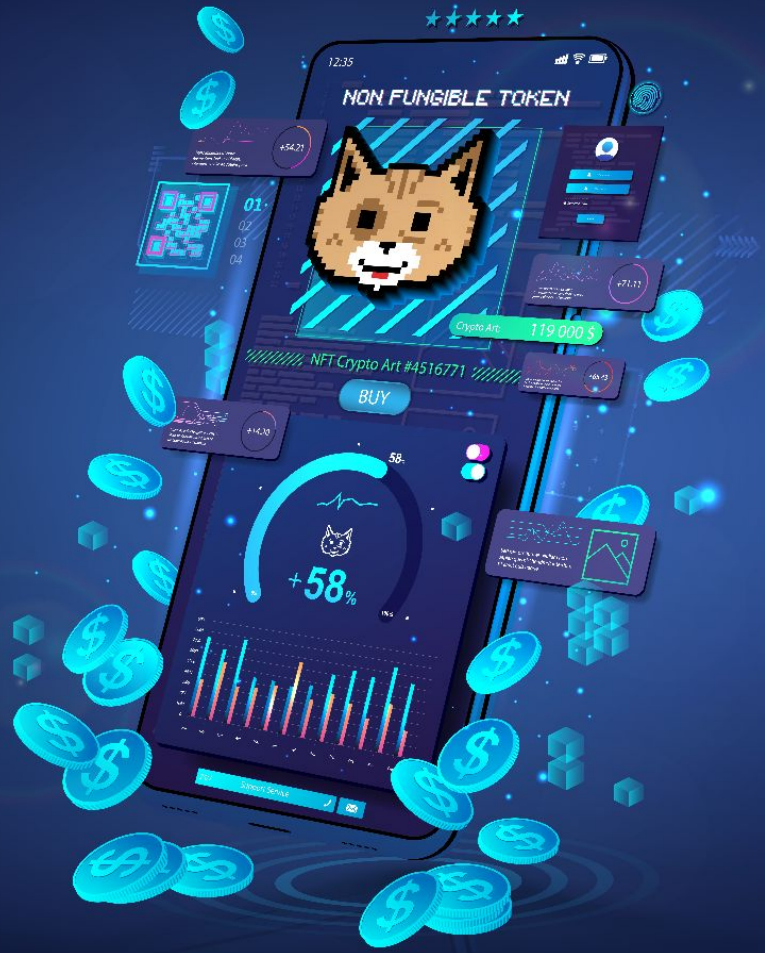
By the end of today's class, you'll be able to:

- Recognize the differences between supervised and unsupervised machine learning
- Define clustering and apply the K-means algorithm to identify clusters in a given dataset.
- Determine the optimal number of clusters for a dataset using the elbow method.
- Describe the purpose of data scaling and preprocessing and apply scaling to a dataset.
- Convert categorical data to indicator variables.

Introduction to Machine Learning

Today, machine learning is changing the world at an unprecedented pace.

These changes are being driven by automation which enables decisions to be made more quickly and efficiently than ever before.



The Mysticism of Machine Learning

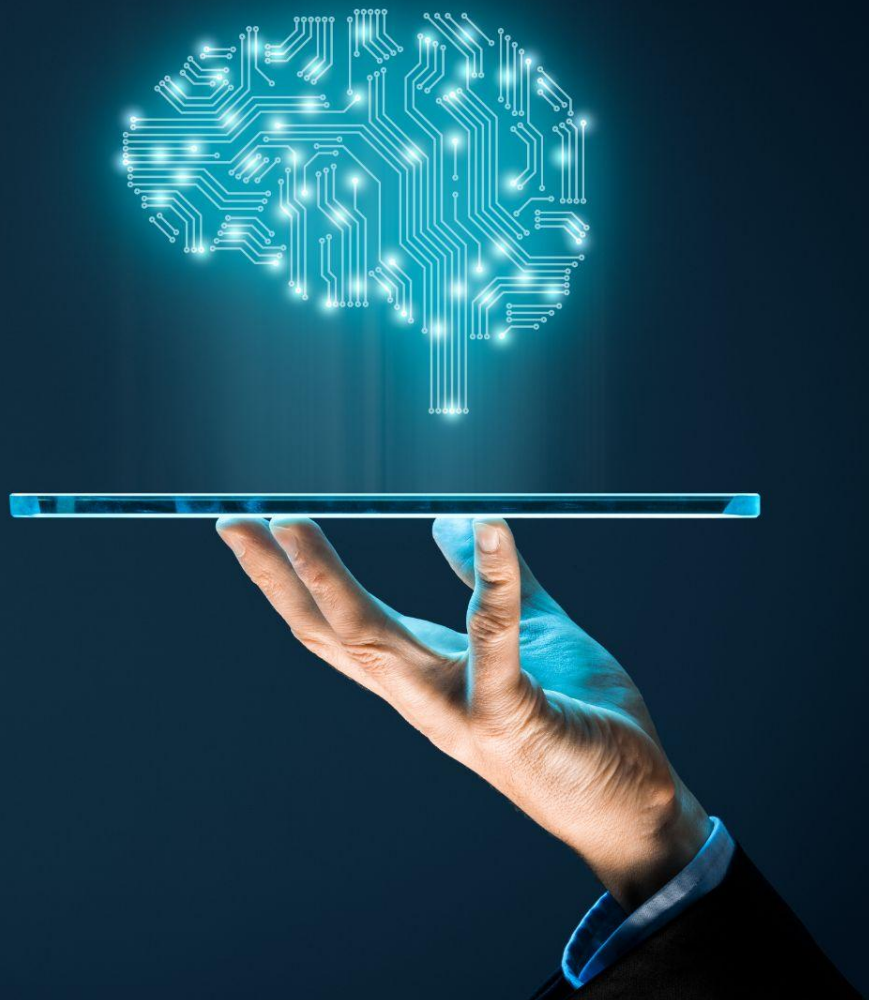
Despite the mainstream use of the term “machine learning,” most people still don't know what machine learning *really* is.

Machine learning 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.

Machine Learning

Algorithms learn how to make decisions without needing anyone to program all that logic.

They learn the patterns, behavior, and relationships of historical data on their own, and then they use that knowledge to make decisions and predictions in the present.



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, have the model group the transactions and predict whether the most recent transaction is fraudulent.

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

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

Types of Machine Learning

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.

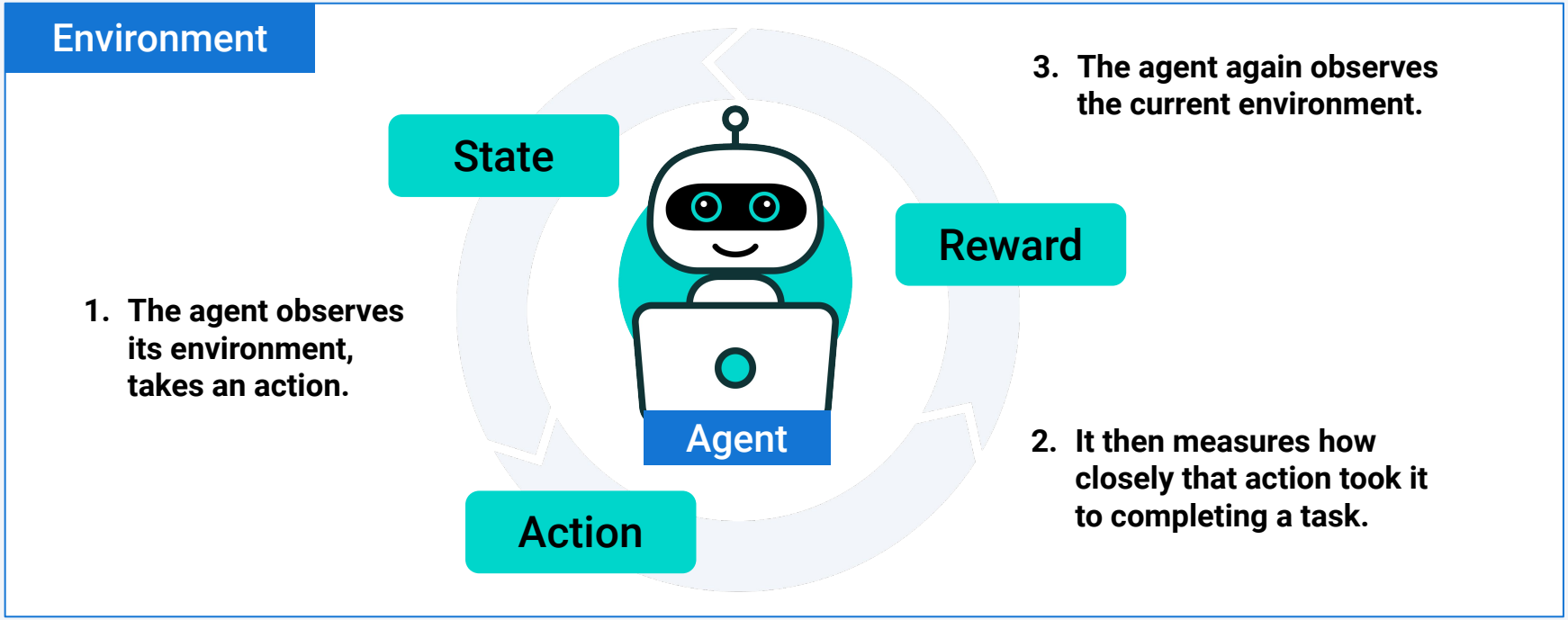
This provides an answer key that can be used to evaluate the accuracy of the training data.

02 Unsupervised learning

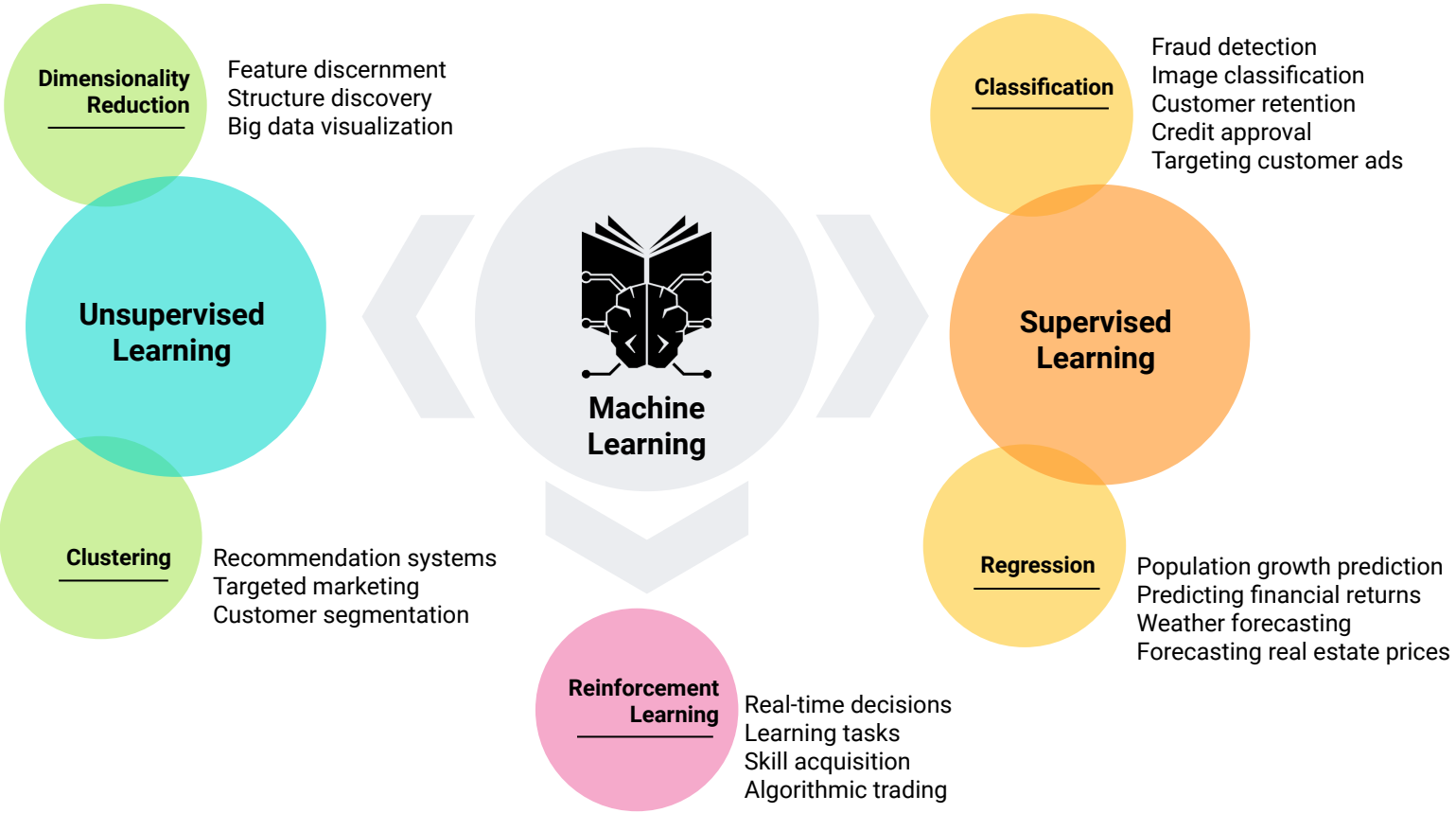
The algorithm tries to make sense of an **unlabeled dataset** by extracting features and patterns on its own.

Reinforcement Learning

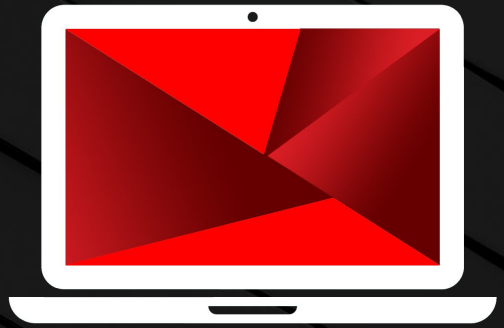
This third type of machine learning algorithm is used less frequently but still has important applications.



Types of Machine Learning



Questions?

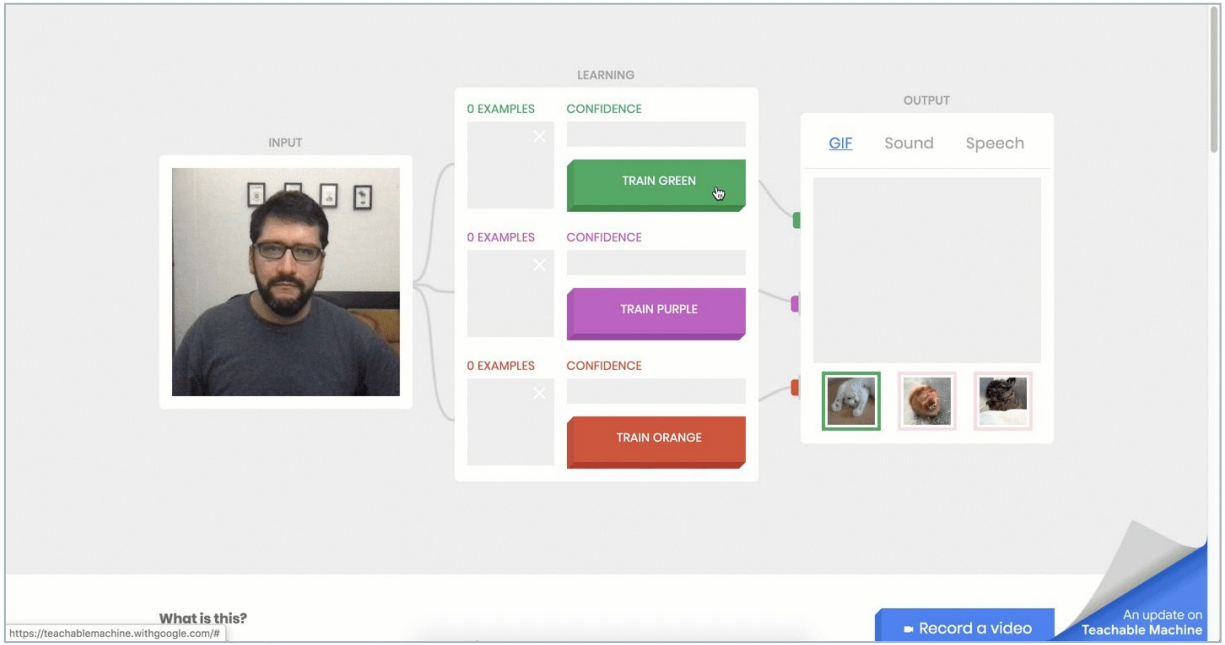


Instructor Demonstration

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.

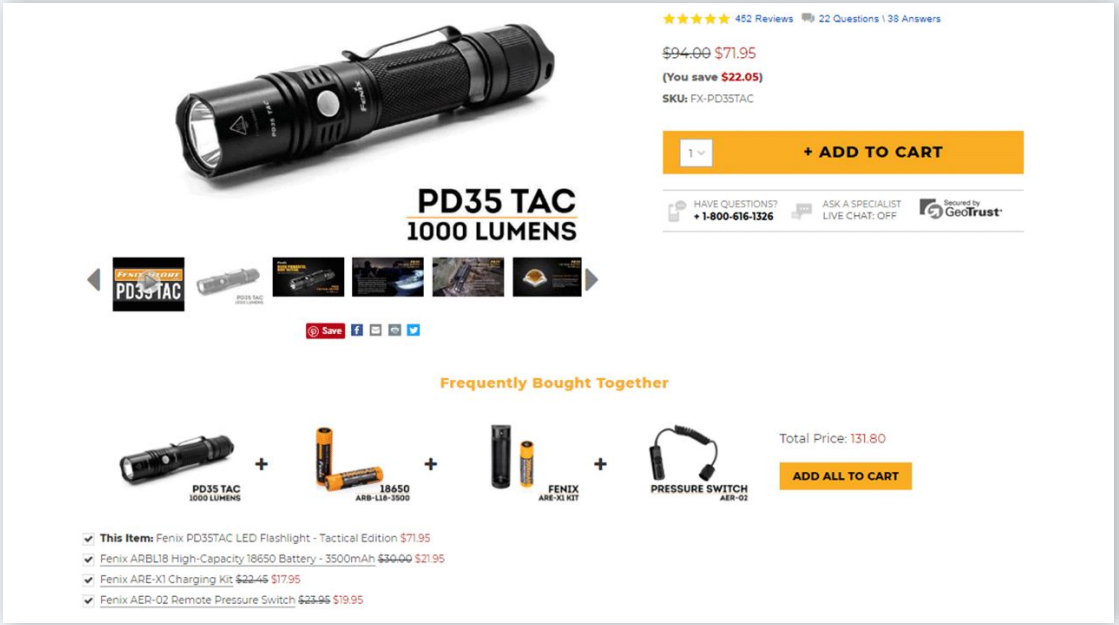


Introduction to Unsupervised Learning

Unsupervised learning algorithms find relationships among unlabeled 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.



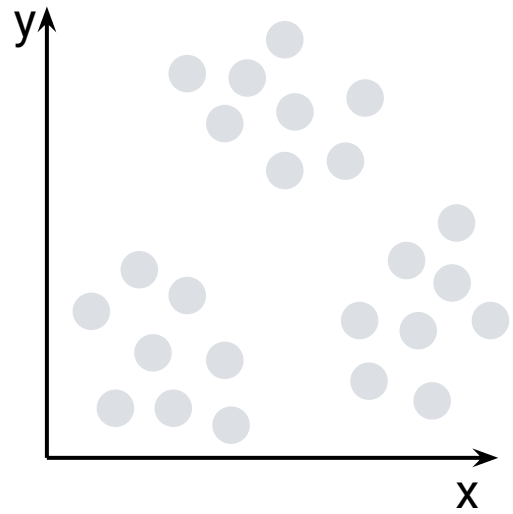
This power to recognize data patterns has broad applications in finance.

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

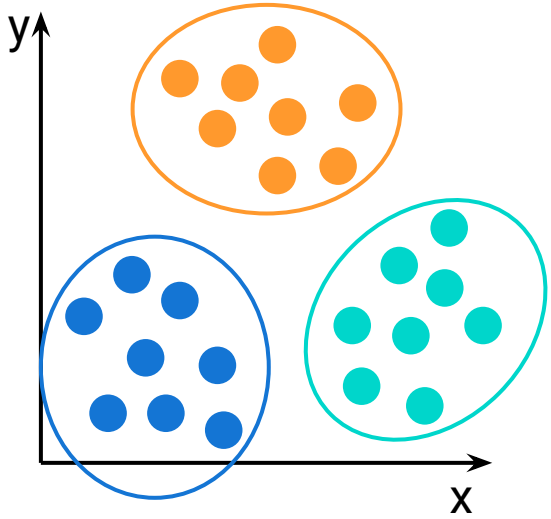
Introduction to Unsupervised Learning

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

Before K-means

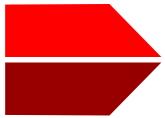


After K-means

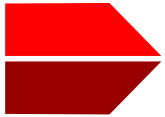


Challenges of Unsupervised Learning

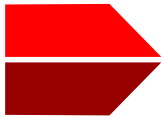
Unsupervised learning comes with challenges:



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:

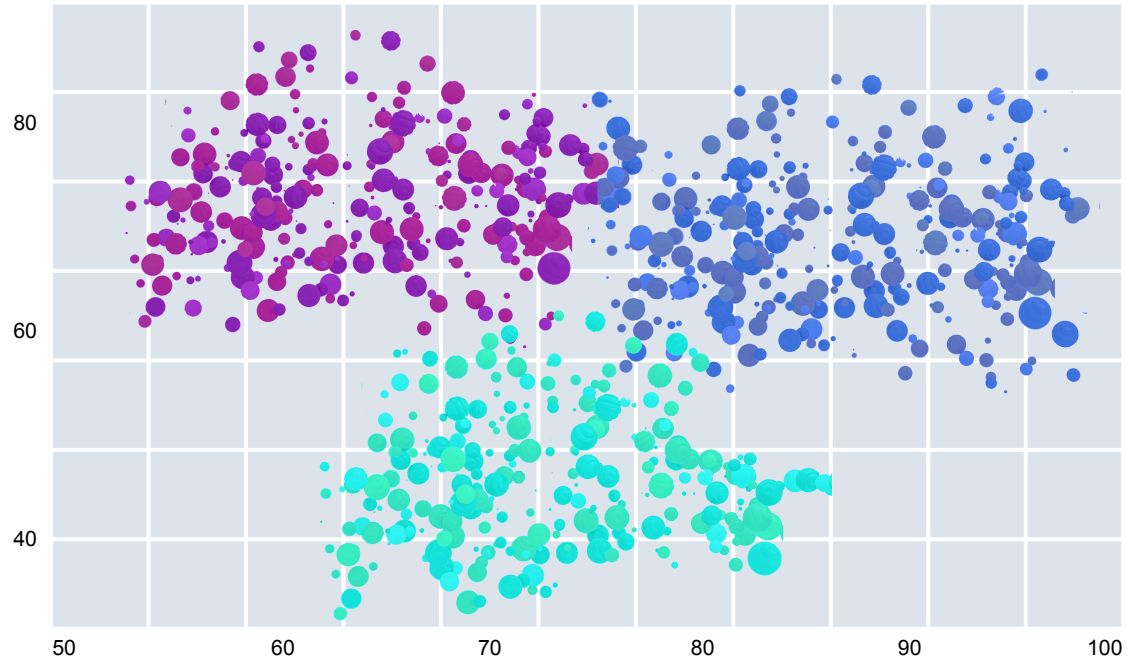
Anomaly detection can be used to detect potentially fraudulent credit card transactions by grouping transactions into “normal” or “abnormal.”

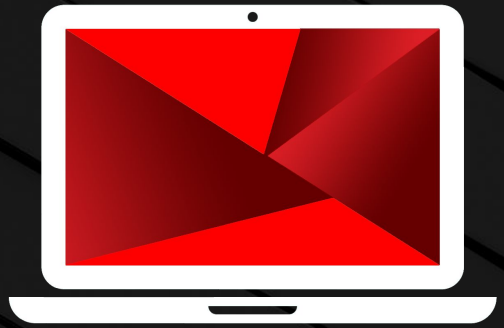
Clustering Explained

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

Clustering Explained

Unsupervised learning models are often created using a clustering algorithm.





Instructor Demonstration

Clustering Explained

Questions?

Time to Code

The K-means Algorithm

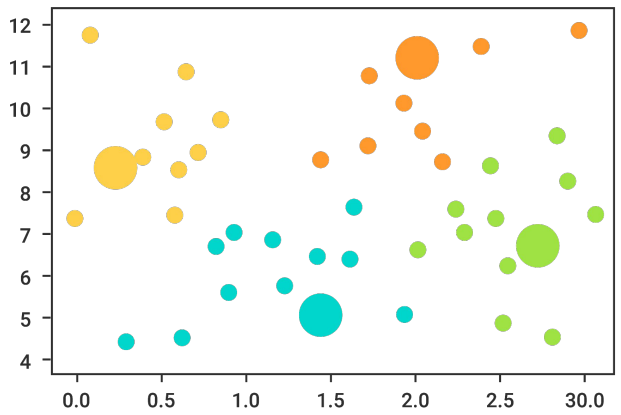
Suggested Time:

15 minutes

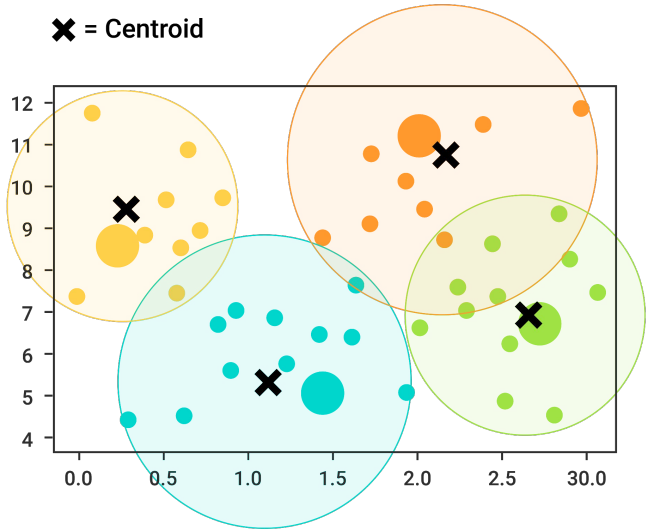
The **K-means algorithm** is the simplest and most common algorithm used to group data points into clusters.

The K-means Algorithm

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



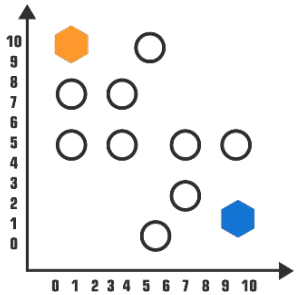
The algorithm assigns points to the closest cluster center.



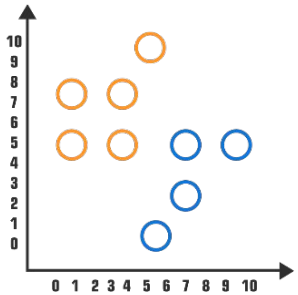
The algorithm readjusts the cluster's center by setting each center as the mean of all the data points contained within that cluster.

The K-means Algorithm

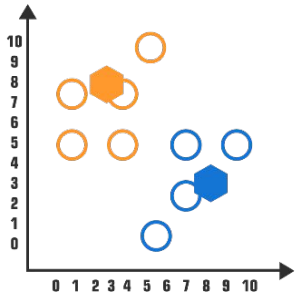
The K-means algorithm then repeats this process, again and again, each time getting a little bit better at separating the data points into distinct groups.



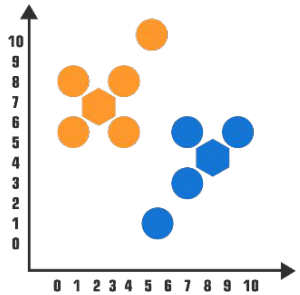
Randomly select starting points for each centroid



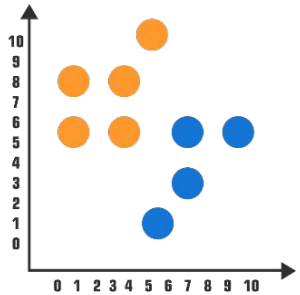
Each point assigned to the closest centroid



Cluster centers move to the mean position of all points in the cluster



Repeat



Stop process after a certain number of iterations or when centers stop moving

Questions?

Introduction to Clustering Optimization

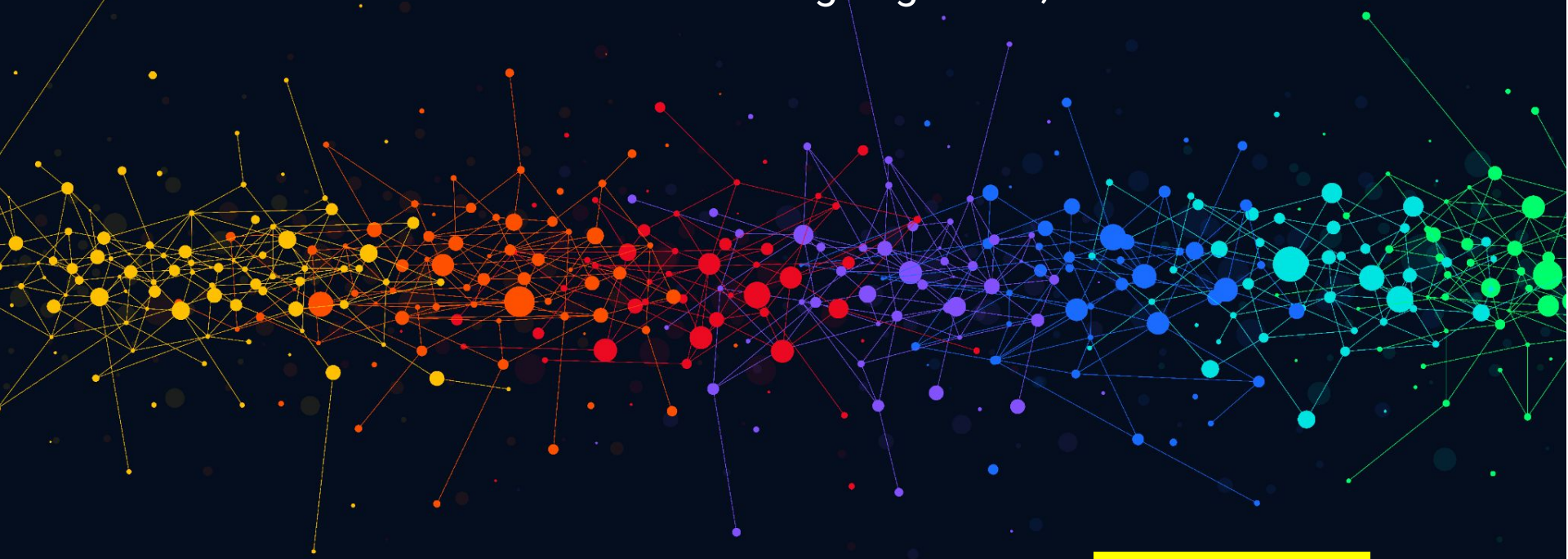
Introduction to Clustering Optimization

- 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 can involve modifying the data preprocessing steps and model parameters until the desired results are achieved.

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?

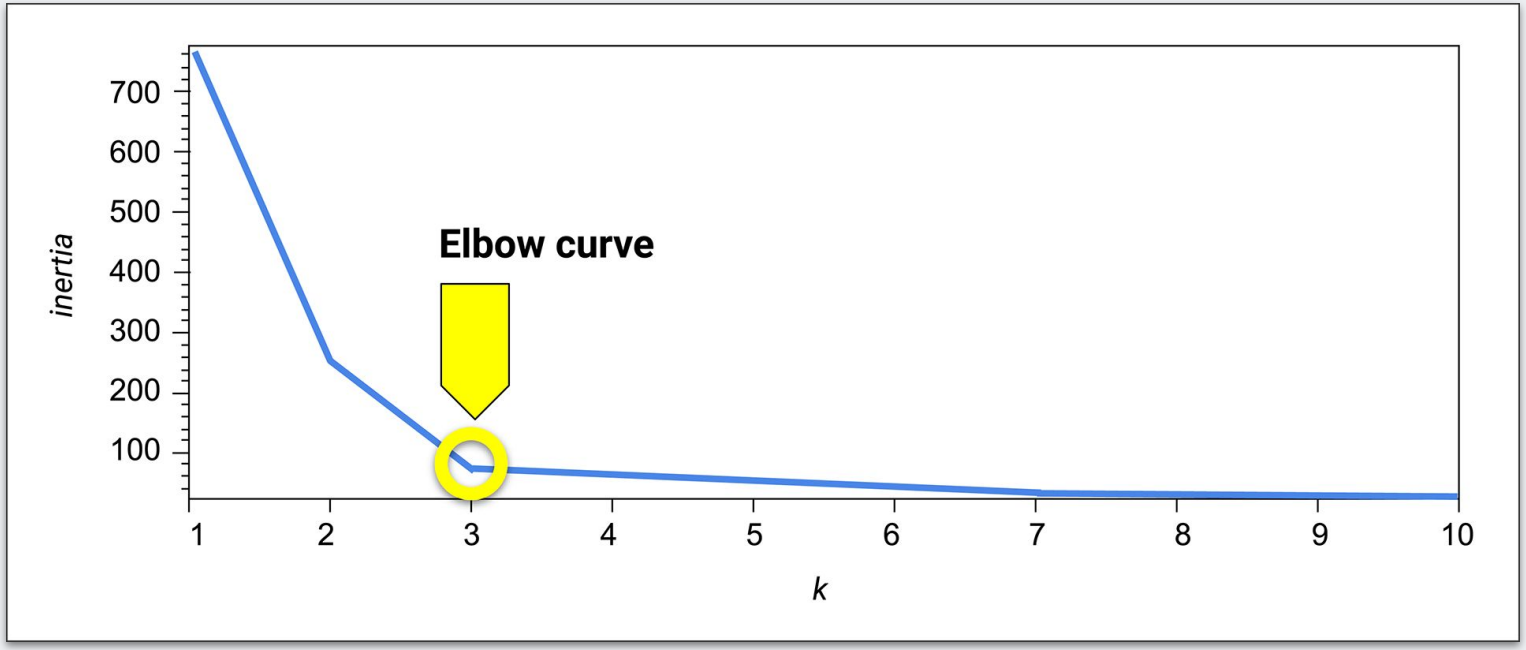
One of the challenges of working with unlabeled data is the unknown number of existing segments, or clusters.



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

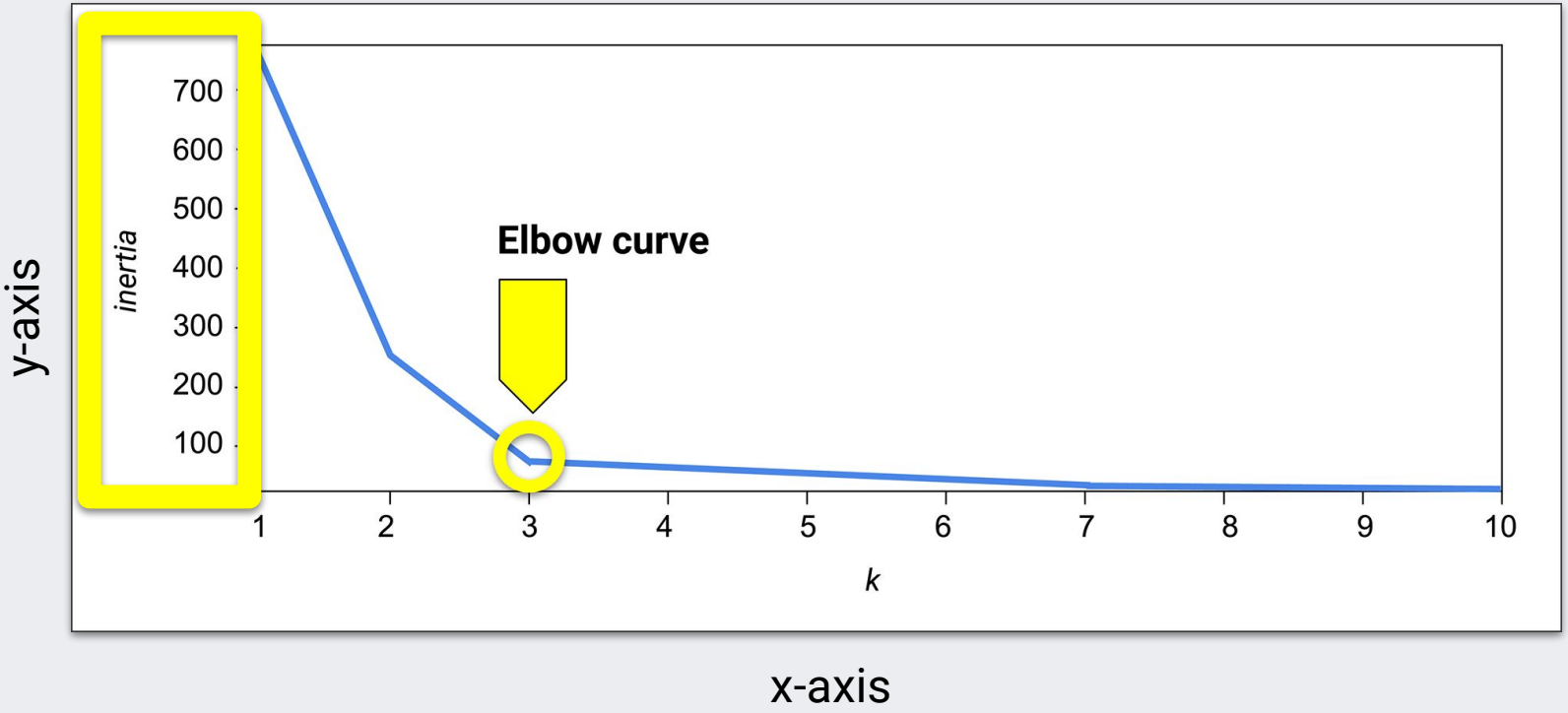
Elbow Curve

The **elbow curve** is commonly used to figure out the best value of k.
It is used to determine the number of clusters needed to form tightly grouped clusters.



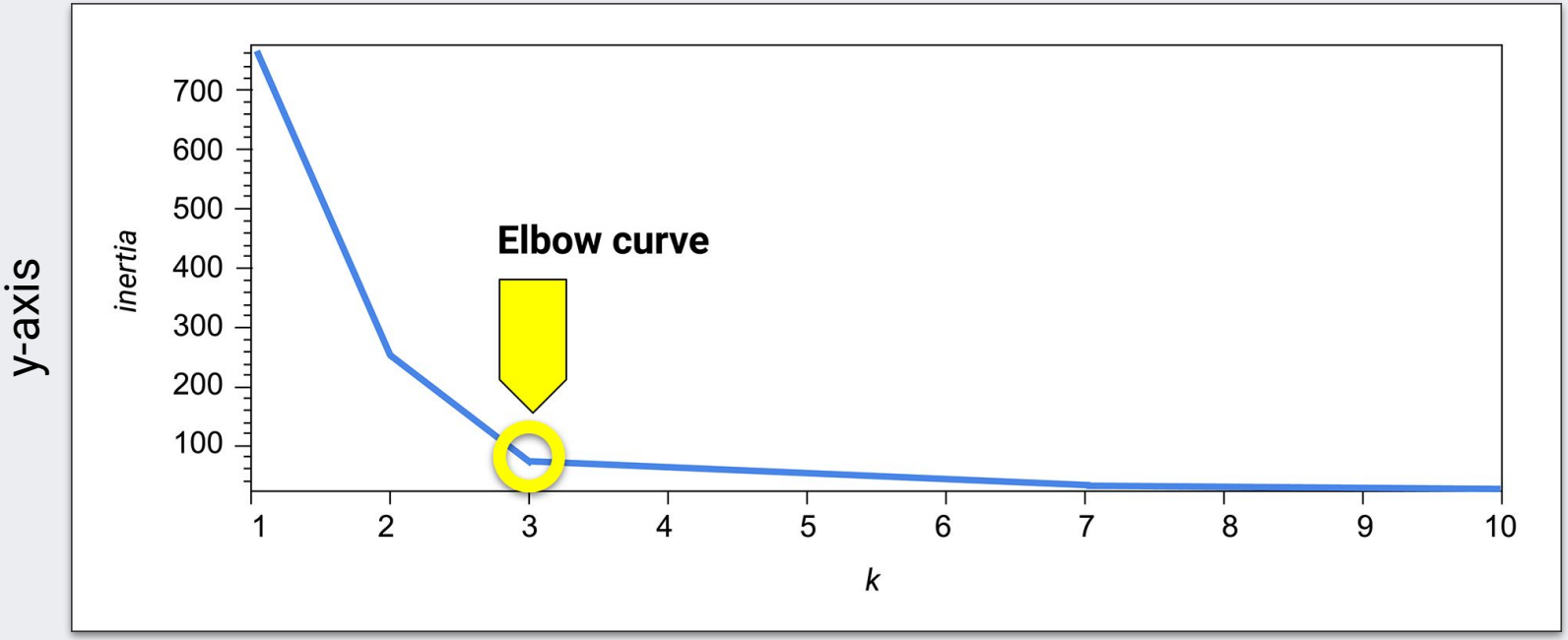
Elbow Curve

Inertia is commonly used as an objective function. It is a measure of how far, on average, each point is from the center of it's cluster.



Elbow Curve

A low inertia value means that the data points are tightly clustered, and therefore implies that the data points in each cluster are truly similar to one another.



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.

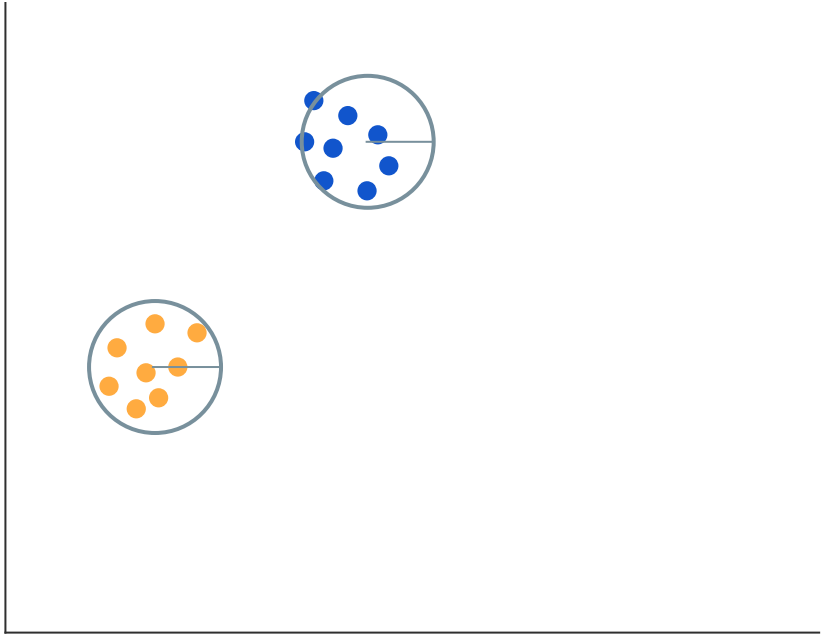
Low concentration

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

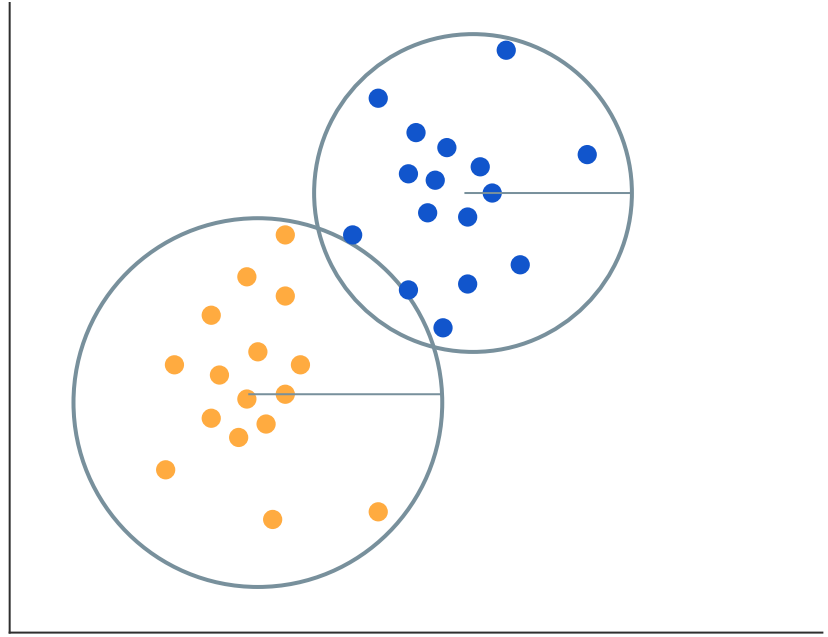
Radius of circle is small =
small standard deviation from cluster mean



vs.

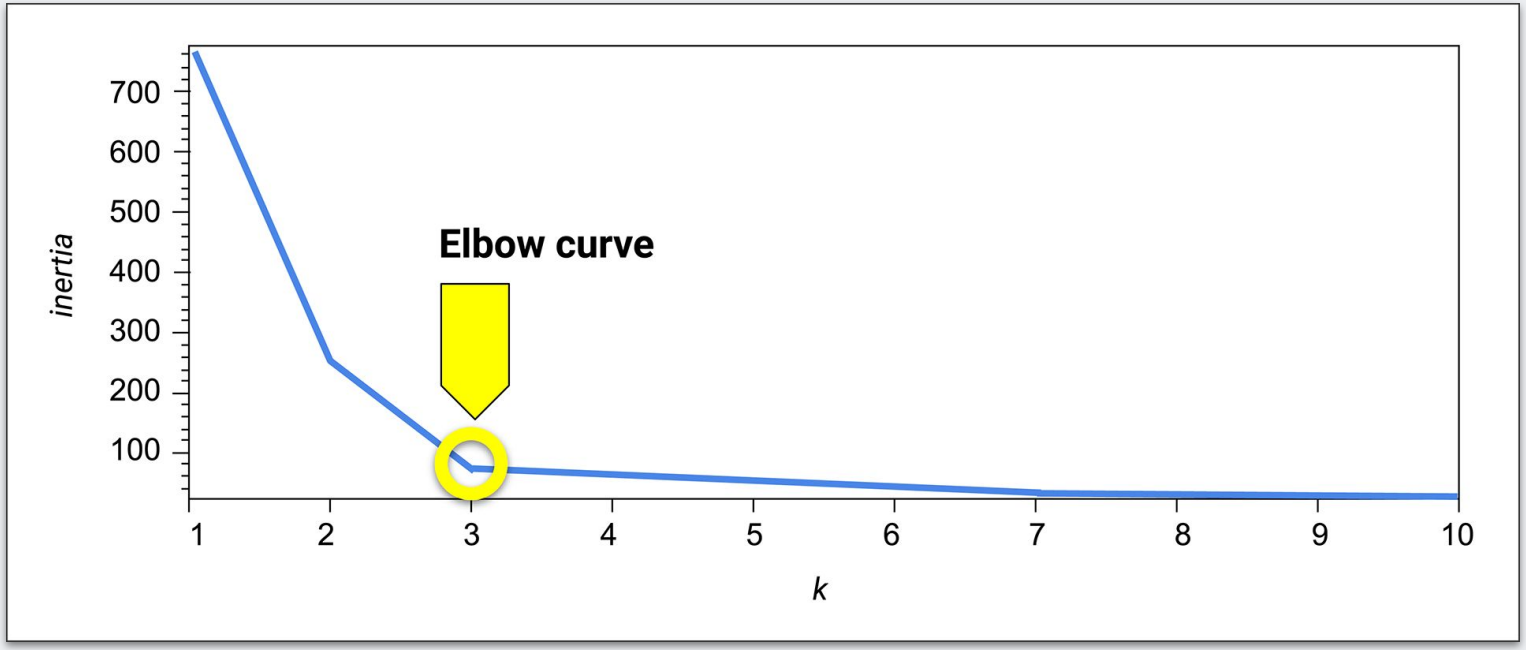
High Inertia

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?


Break

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.

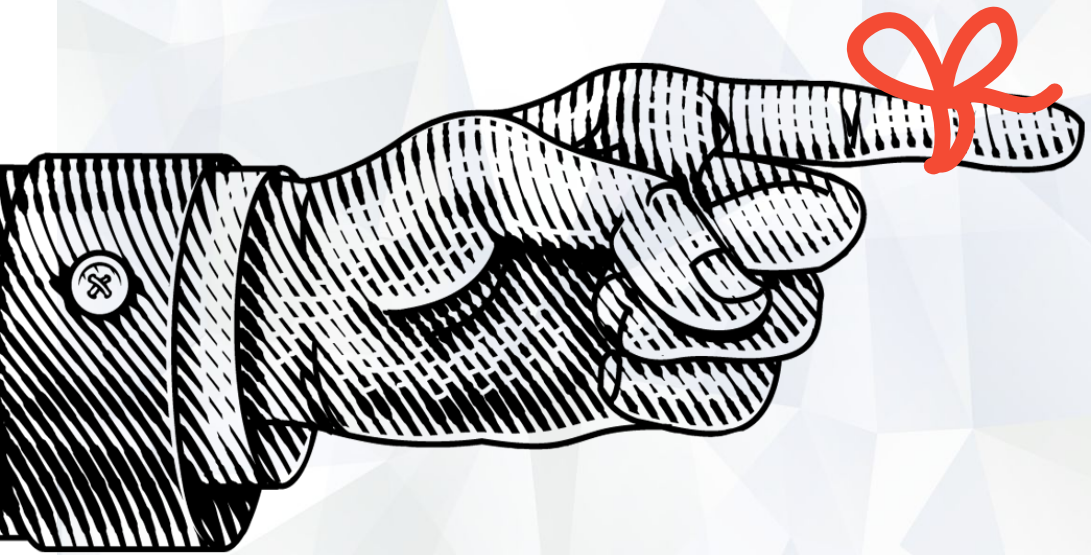
Suggested Time:

20 minutes

Time's Up! Let's Review.

Questions?

Preprocessing Data



Remember,

an important aspect of machine learning is data preparation, or data preprocessing.

Preprocessing Data

We can import a dataset into a Pandas DataFrames, but that doesn't mean all the data is ready for immediate analysis by a machine learning model.

Preprocessing Data

We should consider the following factors when feeding data to a machine learning model:

01

Most machine learning models cannot directly work with data that is in the form of strings or text.

We must encode, or convert, these elements into numeric categories.

02

Machine learning algorithms have trouble learning about data with wildly different scales.

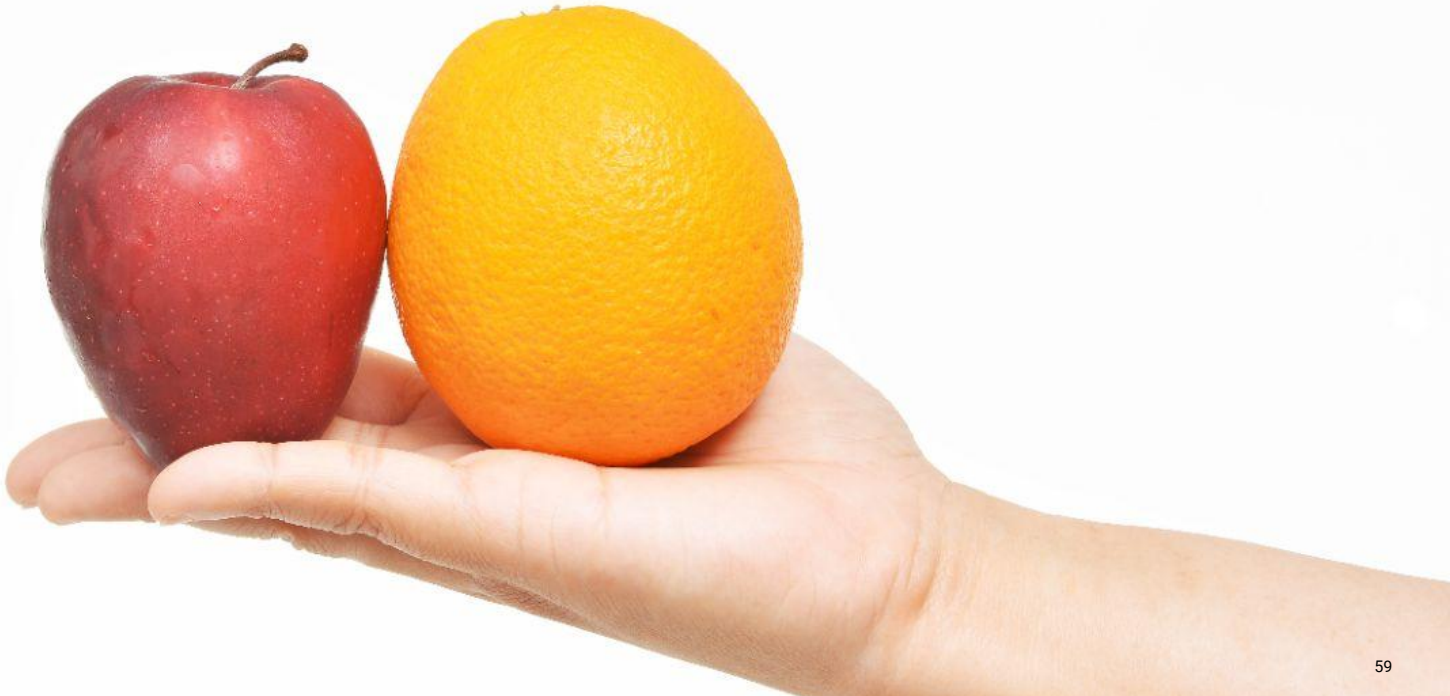
03

Missing values are difficult for machine learning models to navigate.

Scaling Data

When we **scale data**, we eliminate the measurement units and scale the numeric values from all columns to a similar scale.

We can then more easily compare data of **differing natures**.



Encoding is a preprocessing technique for creating a numeric version of a column that has a limited number of possible values.

Preprocessing Data

Instead of manually transforming our data, we can use functions from Pandas and the scikit-learn library to simplify our data preparation.

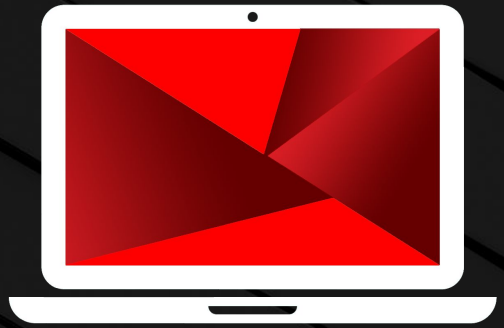
Preprocessing Data

Let’s review one of our credit card spending datasets to illustrate how to scale and encode data.

```
# Read in the CSV file and create the Pandas DataFrame
df_shopping = pd.read_csv(
    Path("../Resources/shopping_data.csv")
)

# Review the DataFrame
df_shopping.head()
```

	CustomerID	Card Type	Age	Annual Income	Spending Score
0	1	Credit	19	15000	39
1	2	Credit	21	15000	81
2	3	Debit	20	16000	6
3	4	Debit	23	16000	77
4	5	Debit	31	17000	40



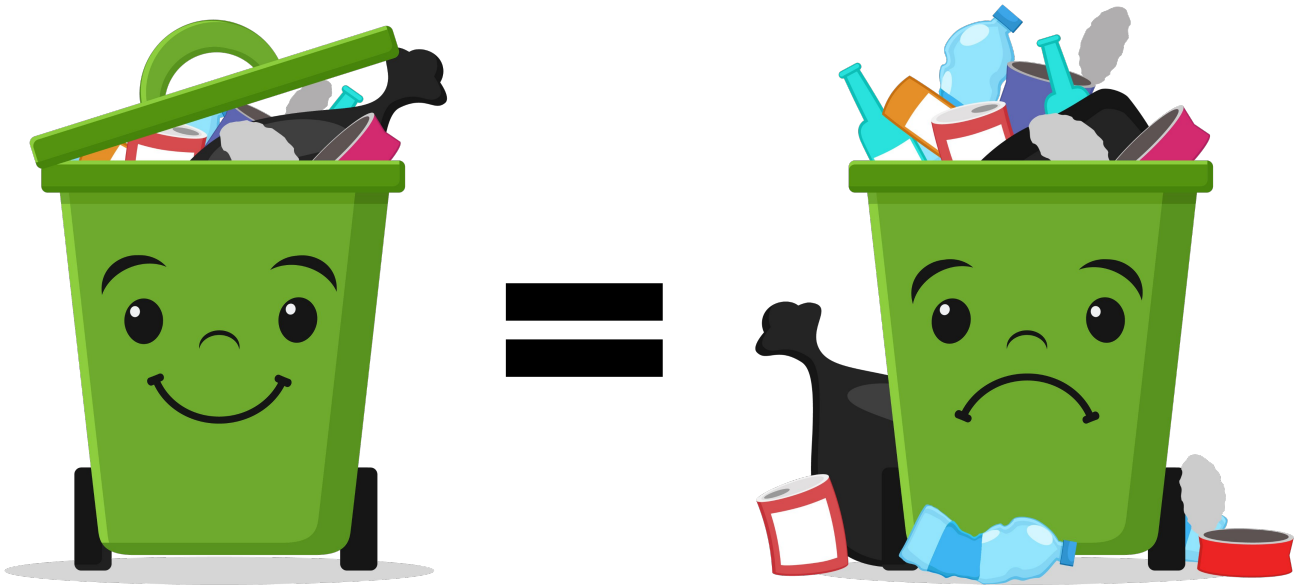
Instructor Demonstration

Scaling and Encoding

Questions?

Preprocessing Data

There is a saying, **“Garbage in, garbage out.”**
The data going into a machine learning model must be clean for the predictions coming out of it to be accurate.



Time to Code

Preprocessing Data

Suggested Time:

20 minutes

Questions?

Activity: Standardizing Stock Data

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

Suggested Time:

25 minutes

Time’s Up! Let’s Review.

Questions?

THE END