Unsupervised Learning

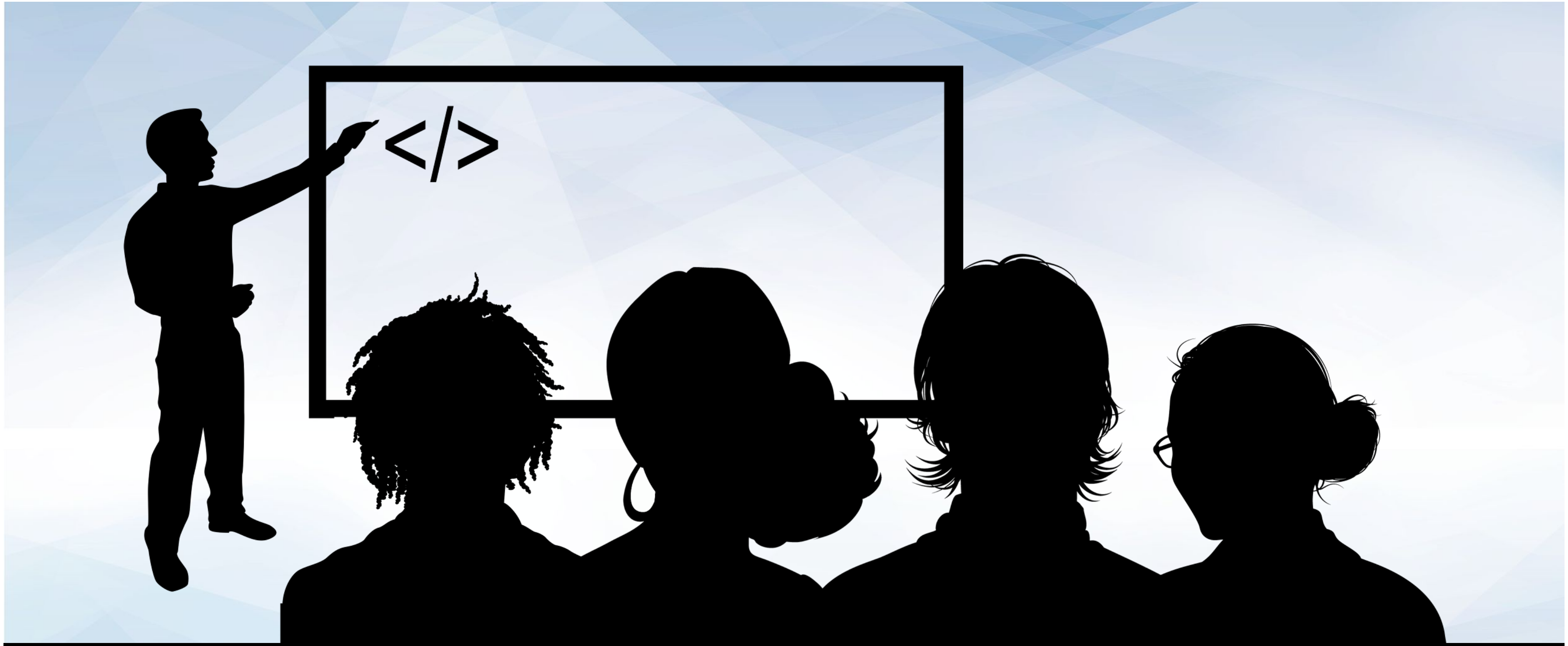
Data Boot Camp

Lesson 20.1

Class Objectives

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

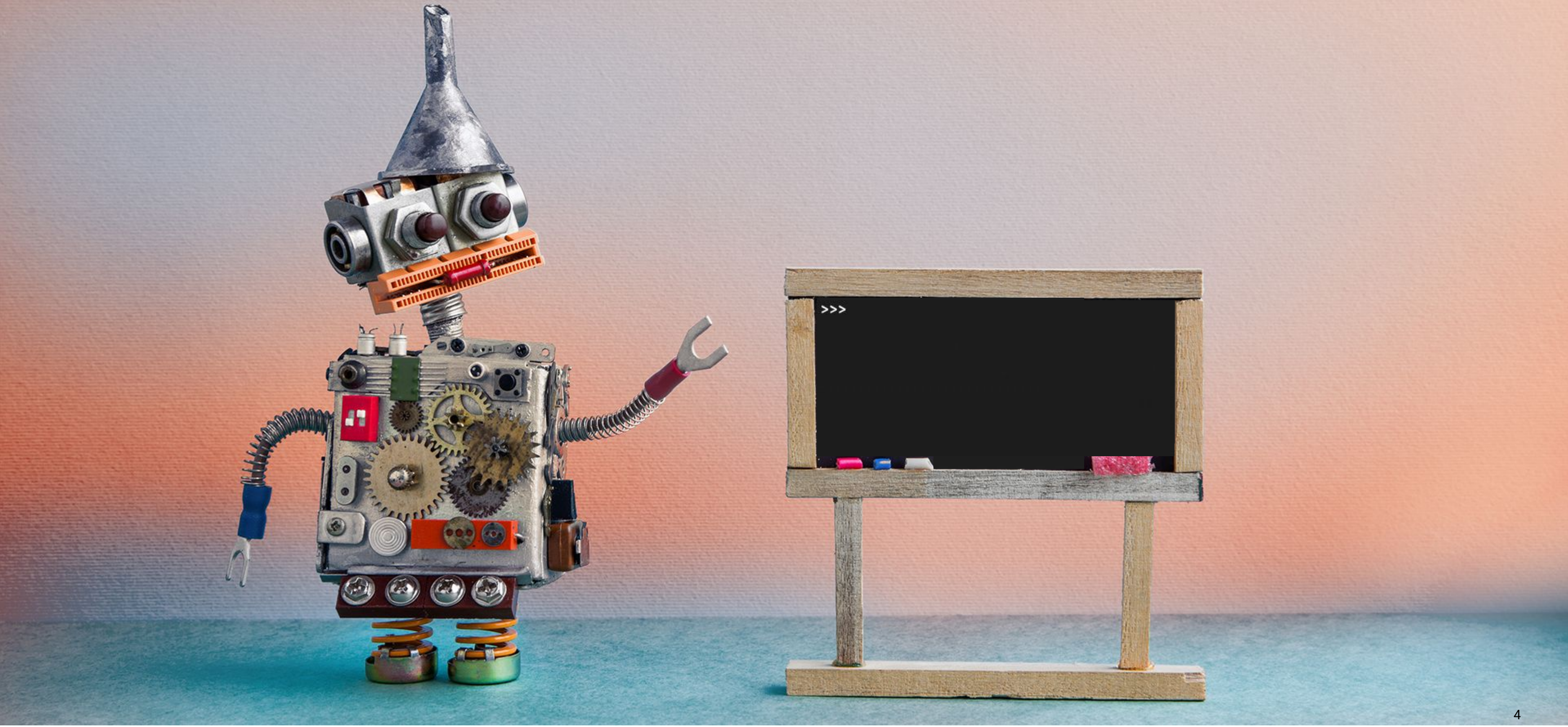
- Recognize the differences between supervised and unsupervised learning.
- Apply the k-means algorithm to identify clusters on a given dataset.
- Apply feature engineering techniques to a dataset to use with the k-means algorithm.
- Speed up machine-learning algorithms using principal component analysis.



Instructor Demonstration

Welcome Class

Instructor Do: Welcome Class



Activity: Supervised Learning with KNN

In this activity, you will use KNN model to predict weather a tumor is malignant or benign.

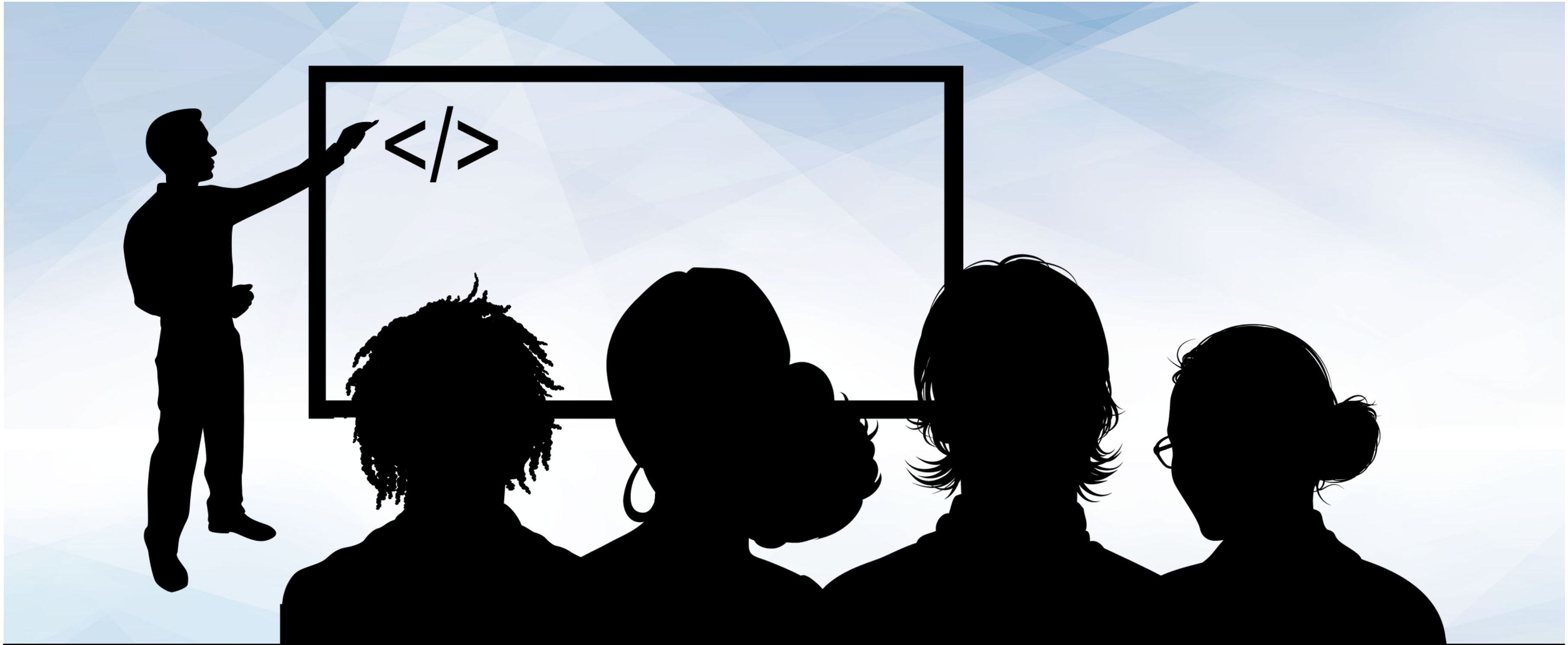
Suggested Time:
15 Minutes

Instructions:

Activity: Supervised Learning with KNN

- **In this activity you will use a KNN model to predict whether a tumor is malignant or benign.**
- Use `bread_cancer.csv` as your dataset.
  - The database has 30 columns. The last, `target`, states whether a tumor sample is benign or malignant.
- Split the dataset into data and target (x and y).
  - Further split the dataset into training and testing sets.
- Standardize the data with the `StandardScaler` module.
  - Create standardized sets for x training data and x test data.
- Instantiate a **KNN** model with k (`n_neighbors`) set to 9.
- Train the model and create predictions with the x test set.
- Use the `accuracy_score` module to assess the accuracy of the KNN model.

Let's Review

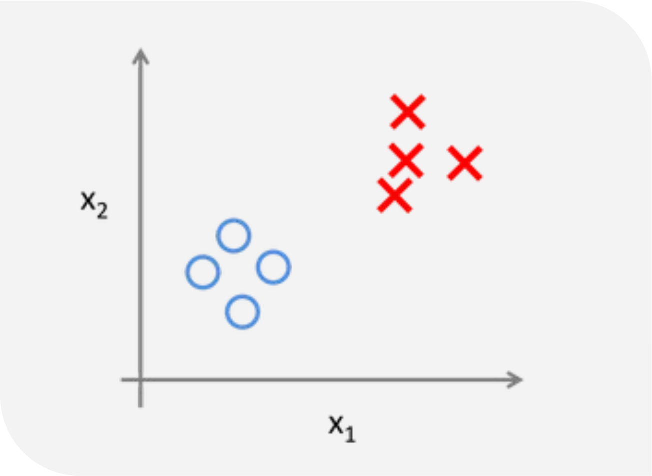


Instructor Demonstration

Introduction to Unsupervised Learning

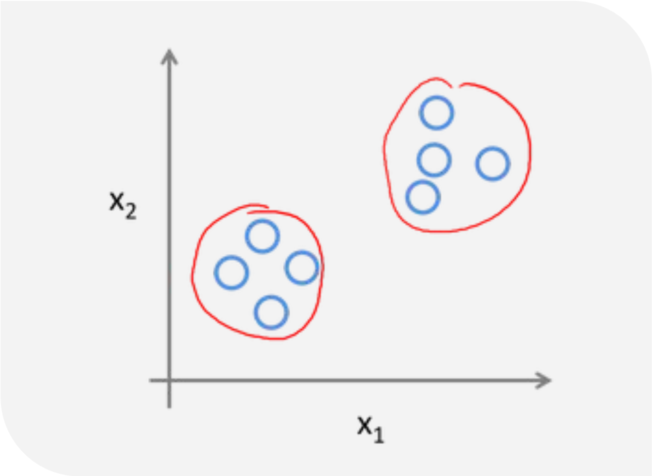
Instructor Do: Introduction to Unsupervised Learning

Supervised Learning



- Input data is labeled.
- Uses training datasets.
- **Goal:** Predict a class or value.

Unsupervised Learning



- Input data is unlabeled.
- Uses just input datasets.
- **Goal:** Determine patterns or grouping data.

Applications of Unsupervised Learning

Instructor Do: Introduction to Unsupervised Learning

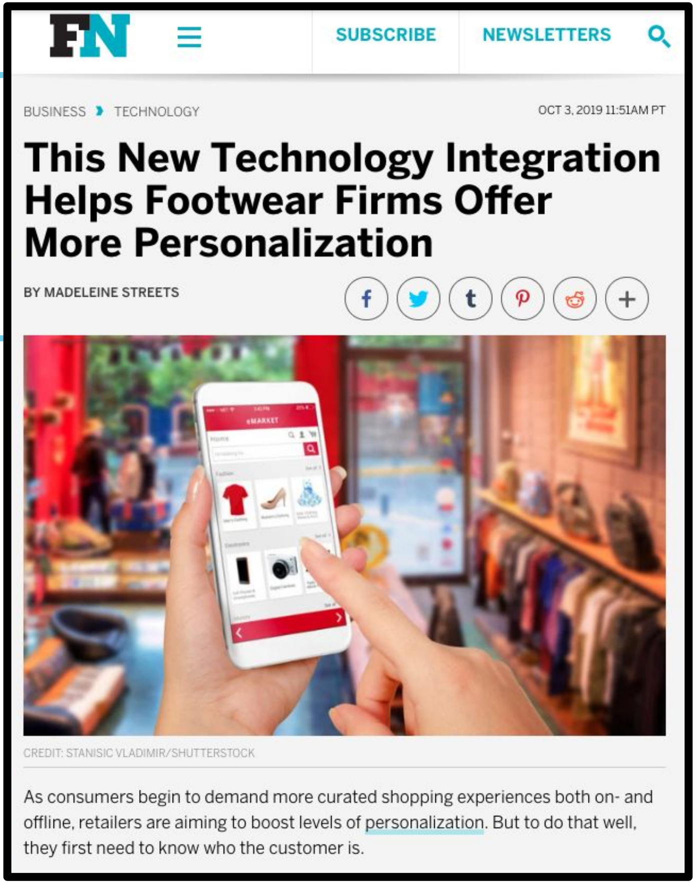
- ➔ Clustering

Allows automatic splitting of the dataset into groups according to similarity.

It can be used for customer segmentation and targeting.

Instructor Do: Introduction to Unsupervised Learning

We can group customers on a retail chain by shopping habits, so we can send customized offers by e-mail or using a mobile app to increase sales.



OCT 3, 2019 11:51AM PT

This New Technology Integration Helps Footwear Firms Offer More Personalization

BY MADELEINE STREETS

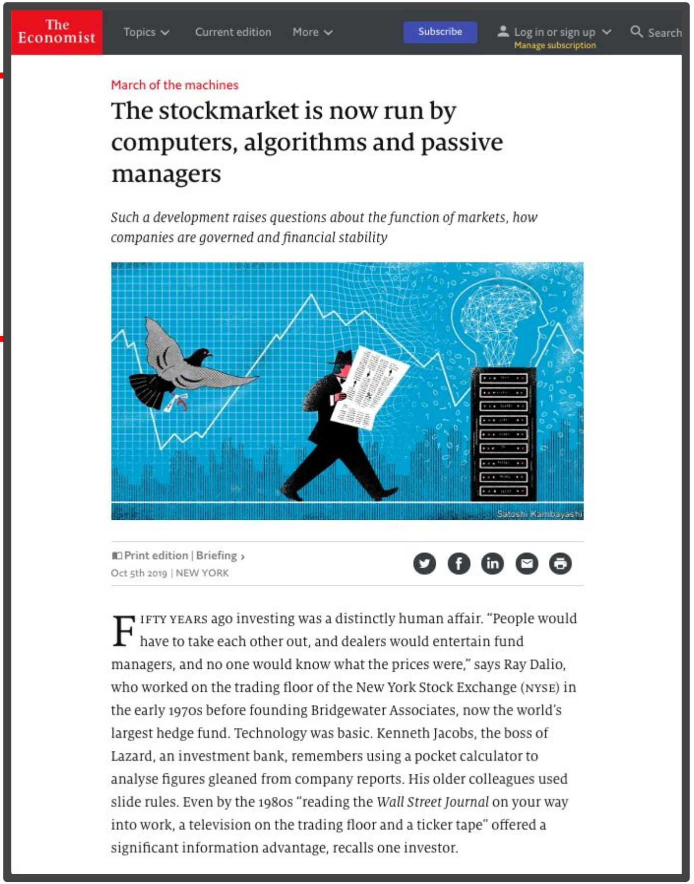
CREDIT: STANISIC VLADIMIR/SHUTTERSTOCK

As consumers begin to demand more curated shopping experiences both on- and offline, retailers are aiming to boost levels of personalization. But to do that well, they first need to know who the customer is.

(Footwear News)

Instructor Do: Introduction to Unsupervised Learning

We can use unsupervised learning to cluster stock data, so we can create investment portfolios according to the resulting groups.



March of the machines

The stockmarket is now run by computers, algorithms and passive managers

Such a development raises questions about the function of markets, how companies are governed and financial stability

Print edition | Briefing ›
Oct 5th 2019 | NEW YORK

FIFTY YEARS ago investing was a distinctly human affair. "People would have to take each other out, and dealers would entertain fund managers, and no one would know what the prices were," says Ray Dalio, who worked on the trading floor of the New York Stock Exchange (NYSE) in the early 1970s before founding Bridgewater Associates, now the world's largest hedge fund. Technology was basic. Kenneth Jacobs, the boss of Lazard, an investment bank, remembers using a pocket calculator to analyse figures gleaned from company reports. His older colleagues used slide rules. Even by the 1980s "reading the *Wall Street Journal* on your way into work, a television on the trading floor and a ticker tape" offered a significant information advantage, recalls one investor.

(Economist)

Applications of Unsupervised Learning

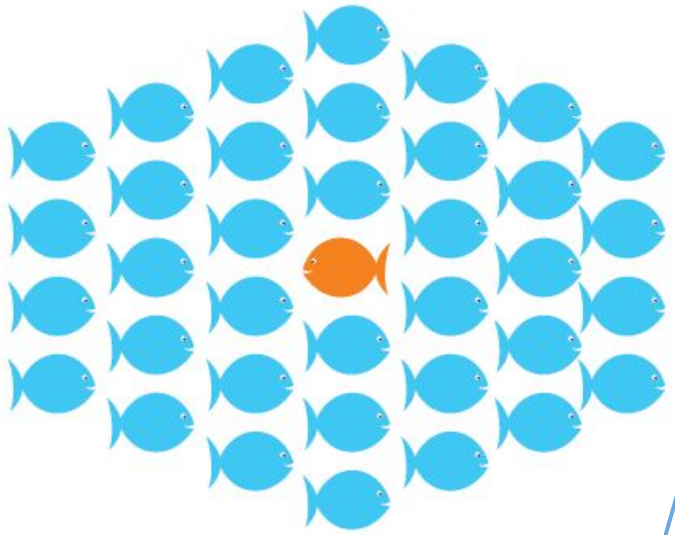
Instructor Do: Introduction to Unsupervised Learning

➔ Anomaly Detection

Automatically discovers unusual data points in a dataset.

It's useful in:

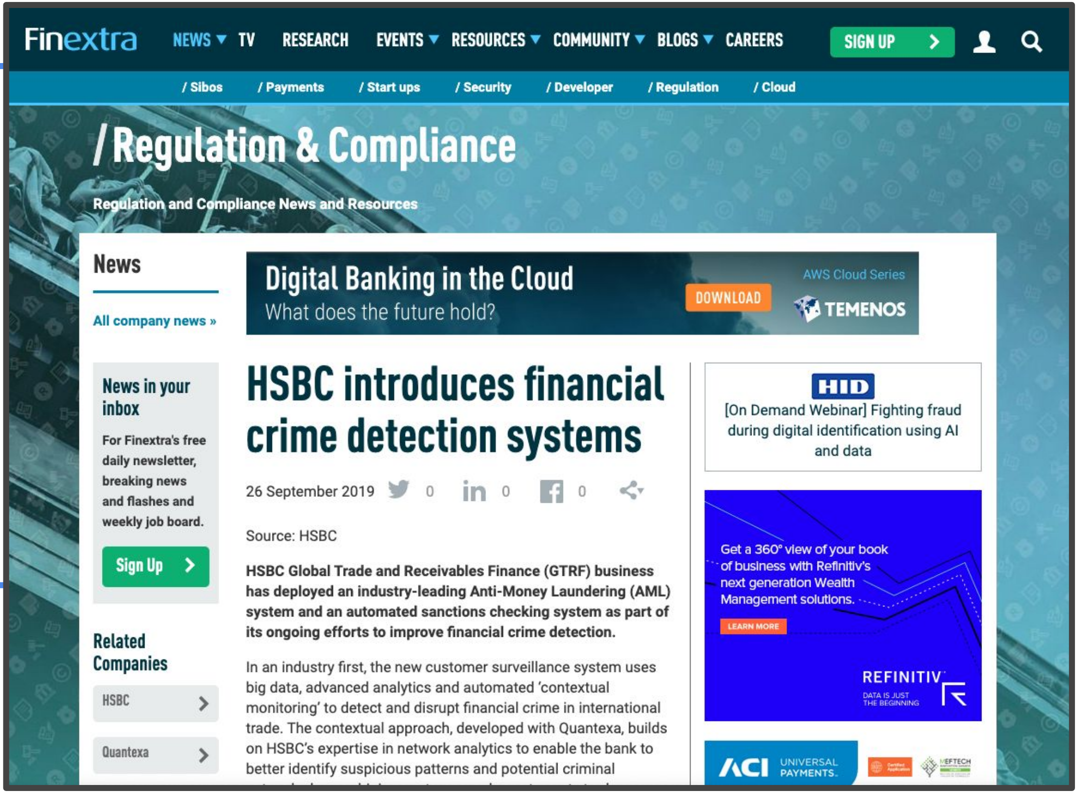
- Identifying fraudulent transactions
- Discovering faulty pieces of hardware
- Identifying an outlier caused by a human error during data entry



Instructor Do: Introduction to Unsupervised Learning

Having thousands of transactions per day on credit card operations, it's hard to identify anomalous or fraudulent transactions.

We can use unsupervised learning to find patterns among transactions data to identify anomalies and potential fraudulent transactions.



/ Regulation & Compliance

Regulation and Compliance News and Resources

News

All company news »

News in your inbox

For Finextra's free daily newsletter, breaking news and flashes and weekly job board.

Sign Up >

Related Companies

HSBC >

Quantexa >

HSBC introduces financial crime detection systems

26 September 2019 0 0 0

Source: HSBC

HSBC Global Trade and Receivables Finance (GTRF) business has deployed an industry-leading Anti-Money Laundering (AML) system and an automated sanctions checking system as part of its ongoing efforts to improve financial crime detection.

In an industry first, the new customer surveillance system uses big data, advanced analytics and automated 'contextual monitoring' to detect and disrupt financial crime in international trade. The contextual approach, developed with Quantexa, builds on HSBC's expertise in network analytics to enable the bank to better identify suspicious patterns and potential criminal

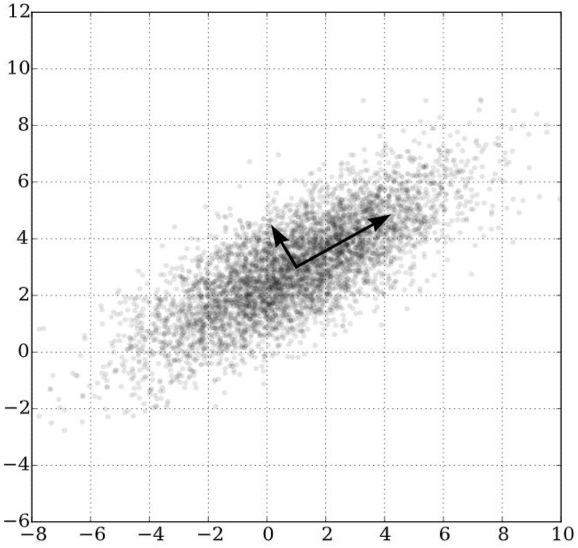
(FinExtra)

Applications of Unsupervised Learning

Instructor Do: Introduction to Unsupervised Learning

➔ Anomaly Detection

- Reduce the number of features while preserving much of the useful data.



Instructor Do: Introduction to Unsupervised Learning

Supervised Learning Approach

- Is this person satisfied?
- How much is this customer going to spend next month?

Upbeat Millennial
purchases Matcha tea drinks most often

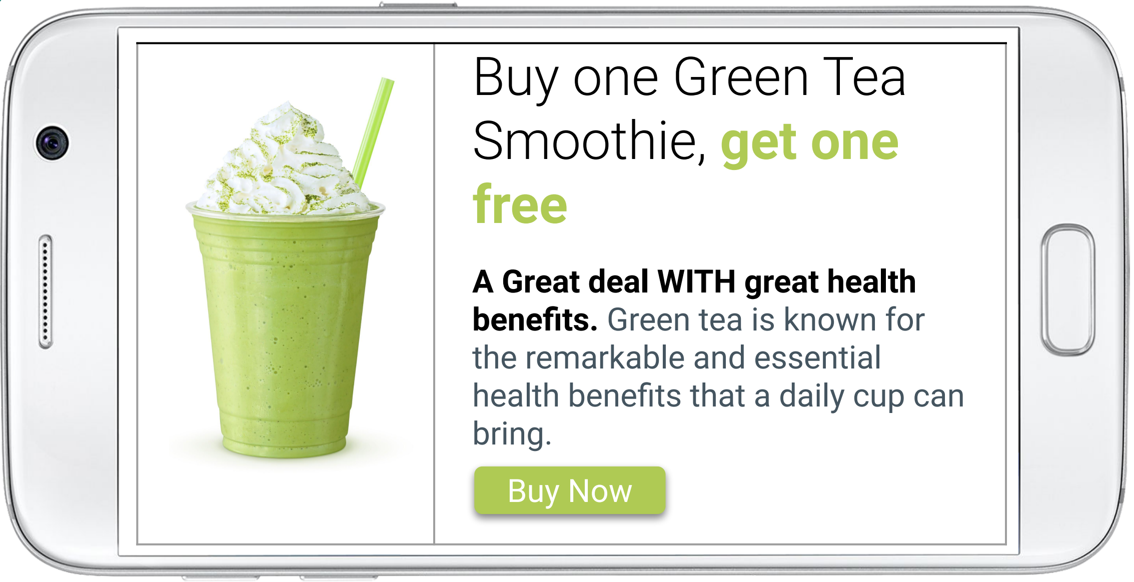
Behaviors
Average Purchase per month: $64.32

Motivations
The most important factor in purchase is knowing that their tea is sustainably sourced.

Instructor Do: Introduction to Unsupervised Learning

Unsupervised Learning Approach

- How can I create a customized offer to customers?



Buy one Green Tea Smoothie, **get one free**

A Great deal WITH great health benefits. Green tea is known for the remarkable and essential health benefits that a daily cup can bring.

Customer Segmentation

Instructor Do: Introduction to Unsupervised Learning

It is the division of potential customers in a given market into discrete groups.

That division is based on customers having similarities such as:

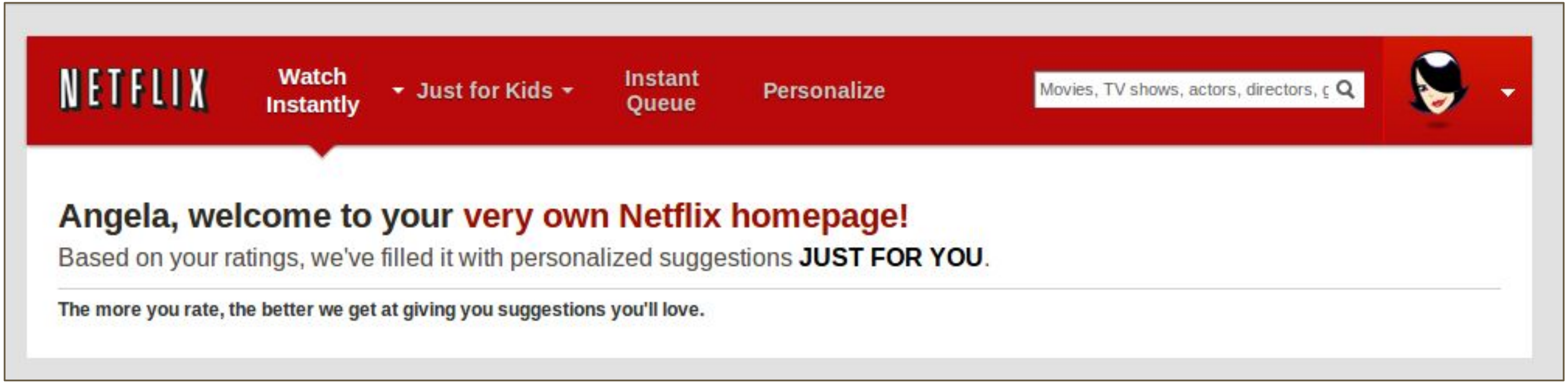
- Customer needs (e.g. a particular product can satisfy some of them)
- Responses to online marketing channels
- Buying habits (e.g. best day for buying, weekly spend)

Customer Segmentation in Action

Instructor Do: Introduction to Unsupervised Learning

Some facts about how customer segmentation is driving revenue in leading companies:

75% of Netflix viewer activity
is driven by recommendation



Angela, welcome to your very own Netflix homepage!

Based on your ratings, we've filled it with personalized suggestions **JUST FOR YOU**.

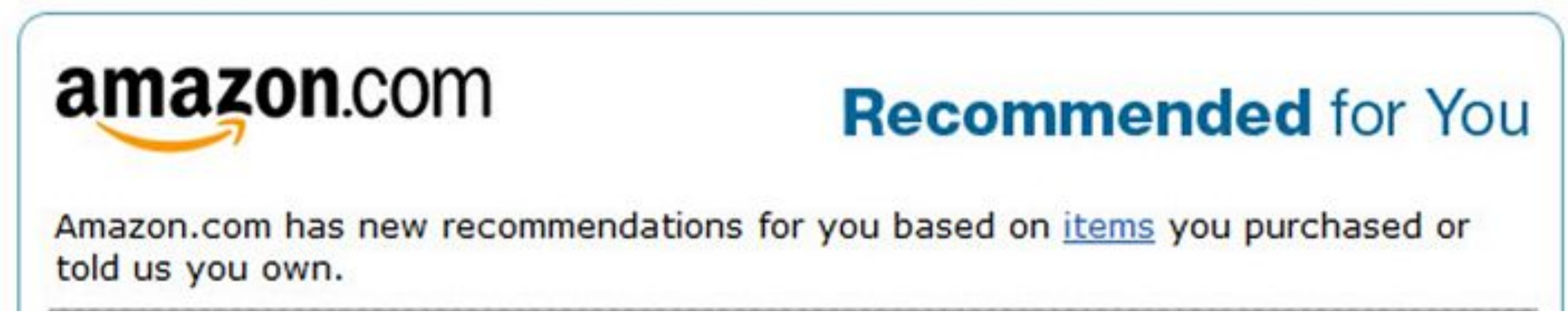
The more you rate, the better we get at giving you suggestions you'll love.

Customer Segmentation in Action

Instructor Do: Introduction to Unsupervised Learning

Some facts about how customer segmentation is driving revenue in leading companies:

35% of Amazon's sales are generated through their recommendation engine



amazon.com

Recommended for You

Amazon.com has new recommendations for you based on items you purchased or told us you own.

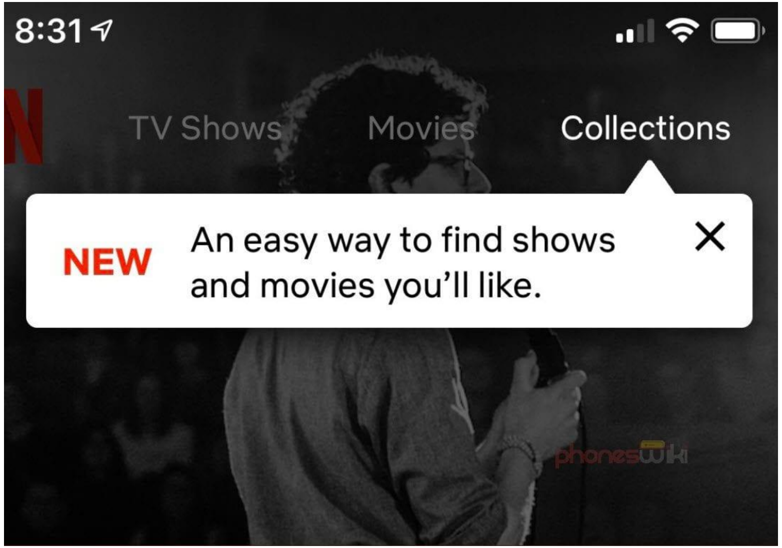
(Source)

Customer Segmentation in Action

Instructor Do: Introduction to Unsupervised Learning

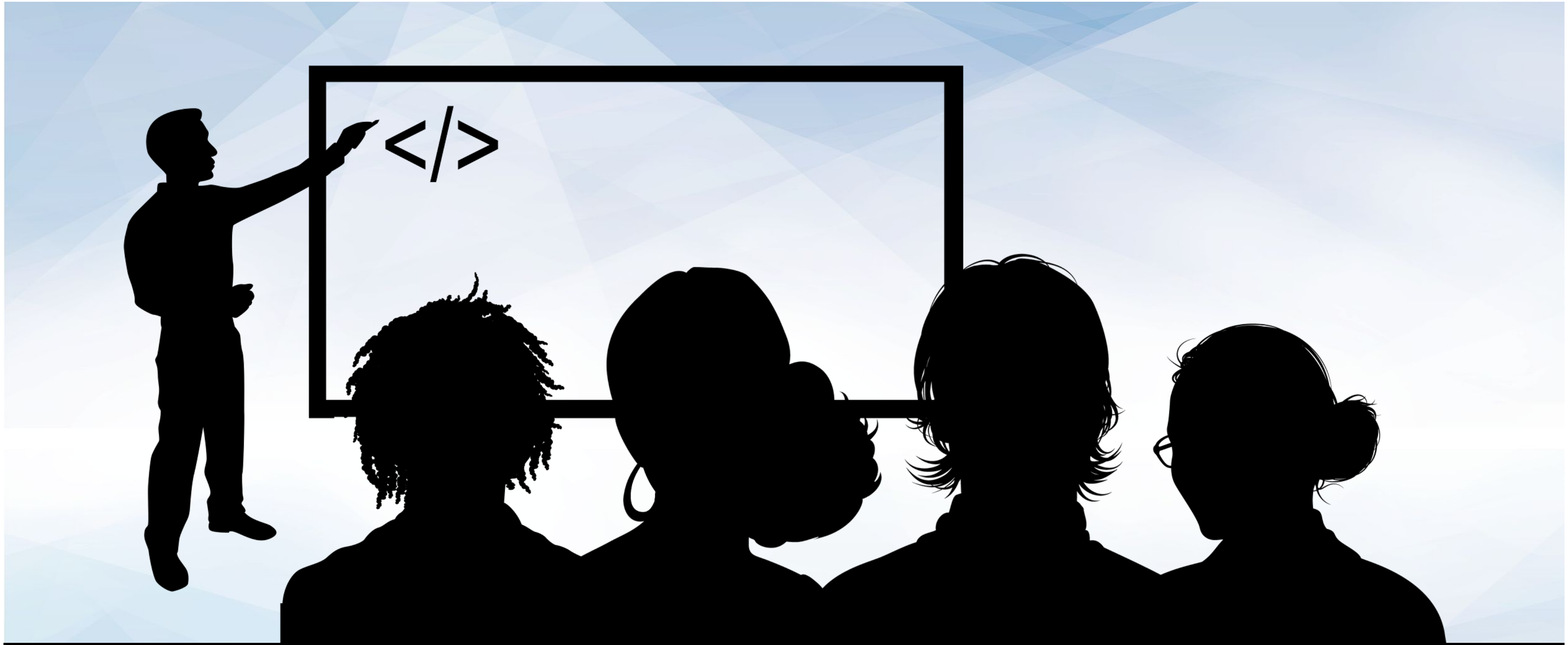
Some facts about how customer segmentation is driving revenue in leading companies:

Netflix's recommendation system saves the company an estimated **$1 Billion** per year through reduced churn



An easy way to find shows and movies you'll like.

(Source)

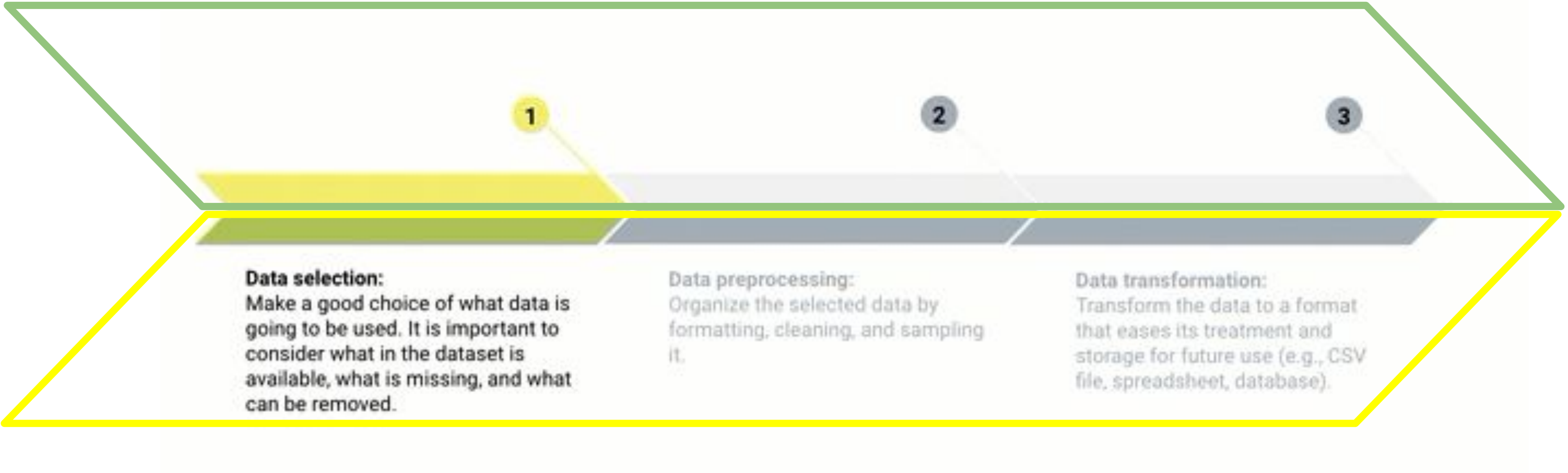


Instructor Demonstration

Data Preparation for Unsupervised Learning

Discipline and organization are key

Instructor Do: Data Preparation for Unsupervised Learning



Data selection:
Make a good choice of what data is going to be used. It is important to consider what in the dataset is available, what is missing, and what can be removed.

Data preprocessing:
Organize the selected data by formatting, cleaning, and sampling it.

Data transformation:
Transform the data to a format that eases its treatment and storage for future use (e.g., CSV file, spreadsheet, database).

Activity: Understanding Customers

In this activity, you will assist an e-commerce company to increase revenue by creating custom offers to its customers as part of their business growth strategy. You will be given access to a dataset containing sales data in order to perform some data preparation tasks to kickstart this project.

Suggested Time:
20 Minutes

Instructions:

Activity: Understanding Customers

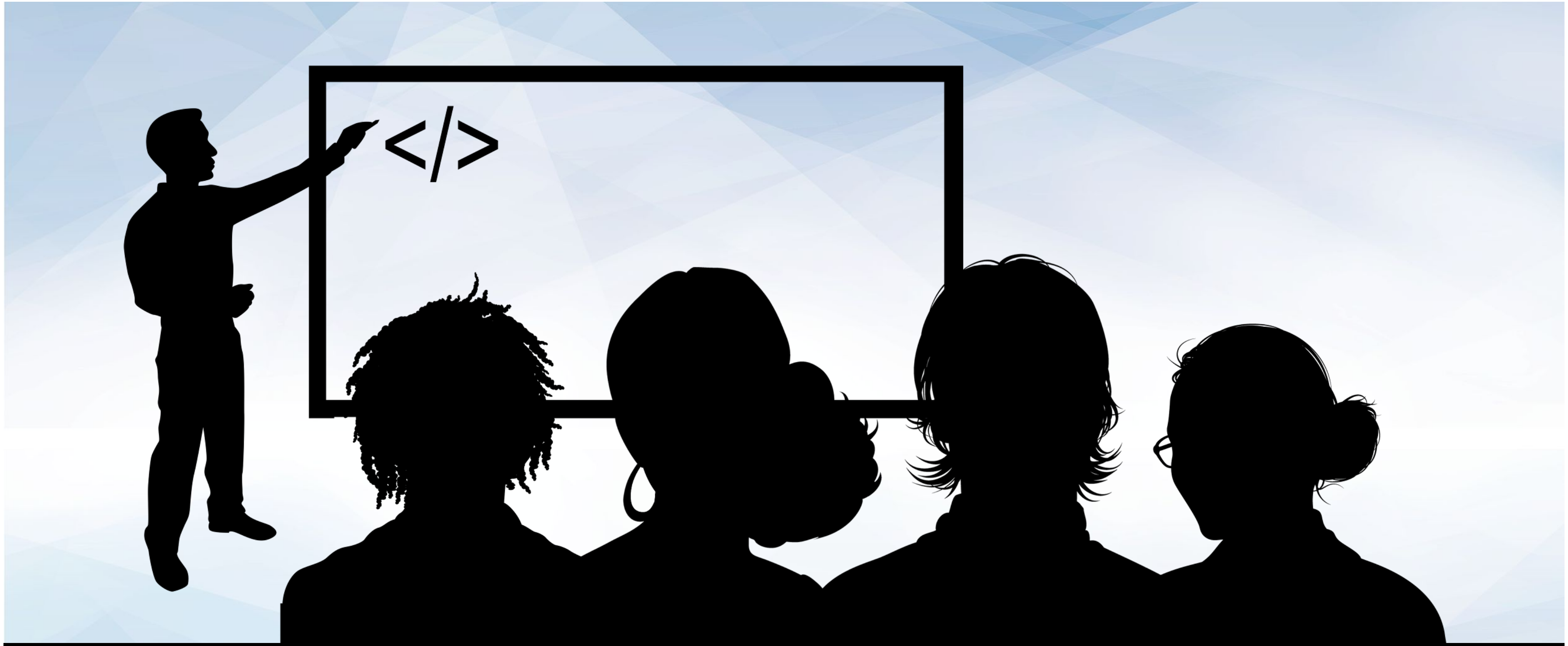
- You are given a dataset that contains historical data from purchases at an online store made by 200 customers. In this activity, you will put **your data-preprocessing skills** to work.
- Use the starter Jupyter Notebook and perform the following tasks:
  - Load the data into Pandas DataFrame and preview it.
  - List the DataFrame's data types to ensure that they're aligned to the type of data stored in each column. Are there any columns whose data type needs to be changed? If so, make the corresponding adjustments.
  - Another best practice is to drop any column that would be unnecessary. Are there any unnecessary columns that need to be dropped? If so, make the corresponding adjustments.
  - Remove all rows with `null` values, if any.
  - Remove duplicate entries, if any.
- To use unsupervised learning algorithms, all the features should be numeric and on similar scales. Perform the following data transformations:
  - The `Previous Shopper` column contains categorical data; anytime you have categorical variables, you should transform them to a numerical value. In this case, transforming `Yes` to `1` and `No` to `0` is a feasible solution.
  - Scale the following features with Scikit-learn's `StandardScaler`: `Age`, `Annual Income`, `Spending Score (1-100)`.
  - Once you are done with data preprocessing, save the cleaned DataFrame in a new `CSV` file.

Let's Review

Countdown timer

15:00

(with alarm)



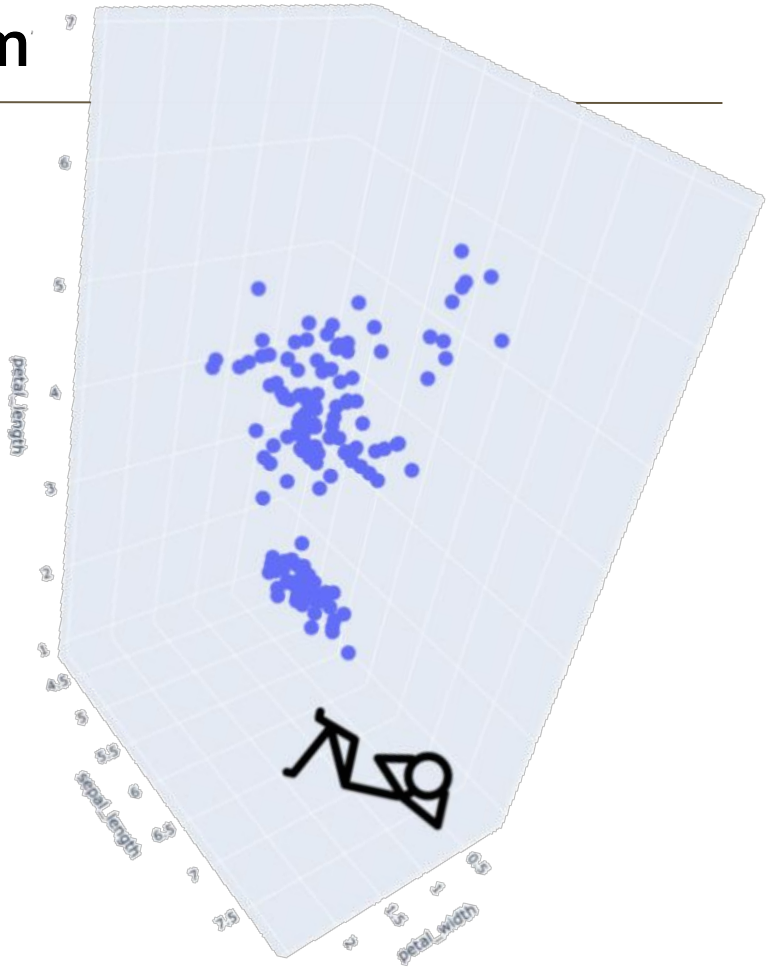
Instructor Demonstration

The K-Means Algorithm

What K-Means can do for me?

Instructor Do: The K-Means Algorithm

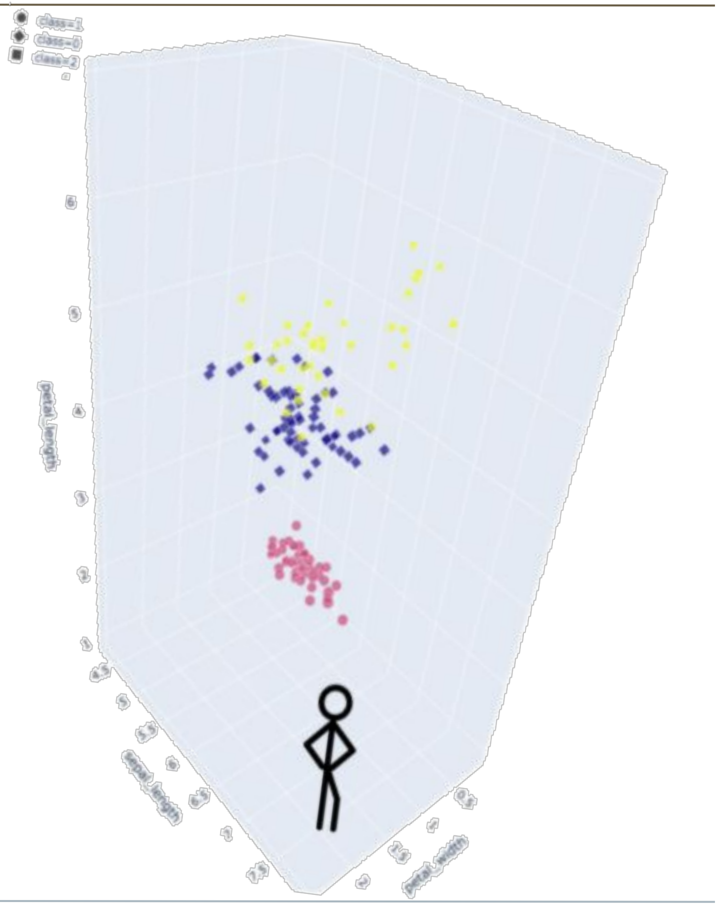
- Imagine that you are in a room full of small spheres (data points).
- Each sphere represents a flower (iris) and the axes represent features of flowers.



What K-Means can do for me?

Instructor Do: The K-Means Algorithm

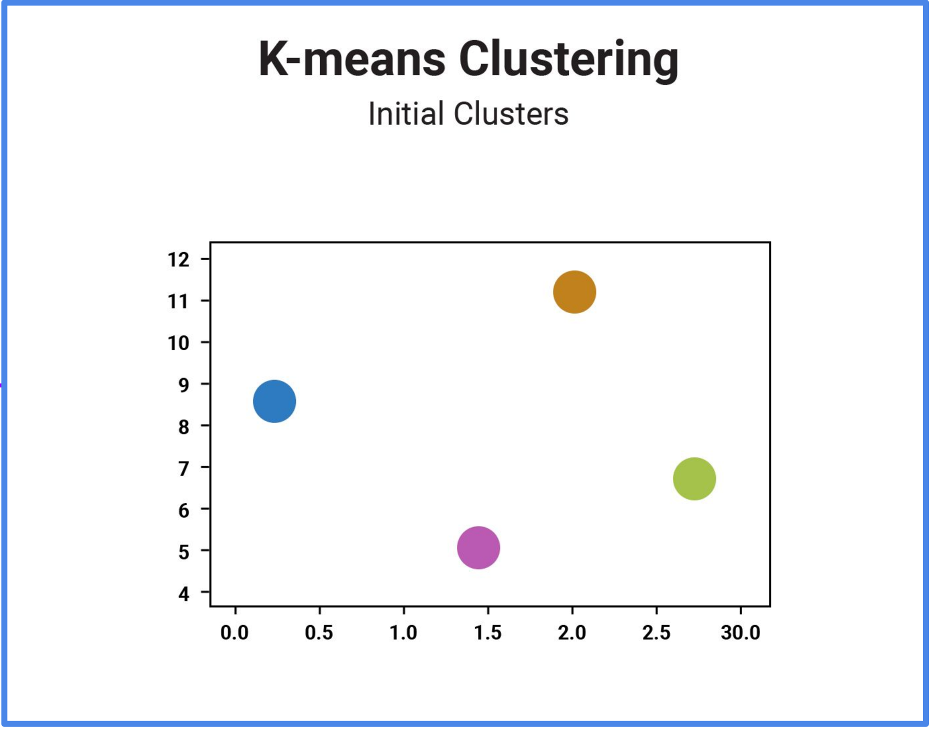
- K-means is an unsupervised learning algorithm used to identify clusters.



How K-Means Algorithm Works?

Instructor Do: The K-Means Algorithm

- K-Means algorithm groups the data into **k clusters**, where belonging to a cluster is based on some similarity or distance measure to a centroid.
- A **centroid** represents a data point that is the arithmetic mean position of all the points on a cluster.



How K-Means Algorithm Works?

Instructor Do: The K-Means Algorithm

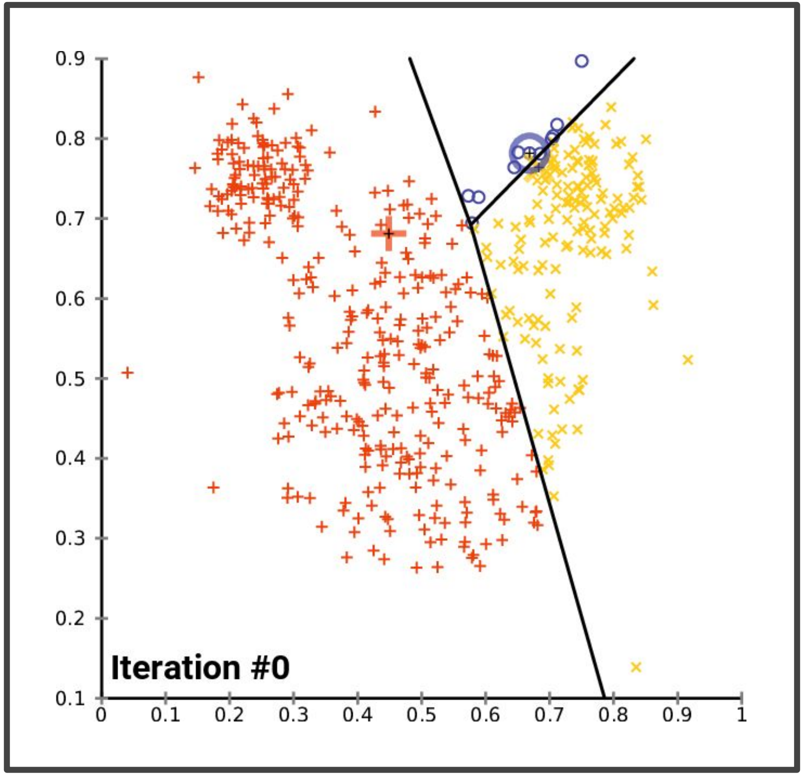
Algorithm at a glance

01 Randomly initialize the *k* starting centroids.

02 Each data point is assigned to its nearest centroid.

03 The centroids are recomputed as the mean of the data points assigned to the respective cluster.

04 Repeat steps 1 through 3 until the stopping criteria is triggered.

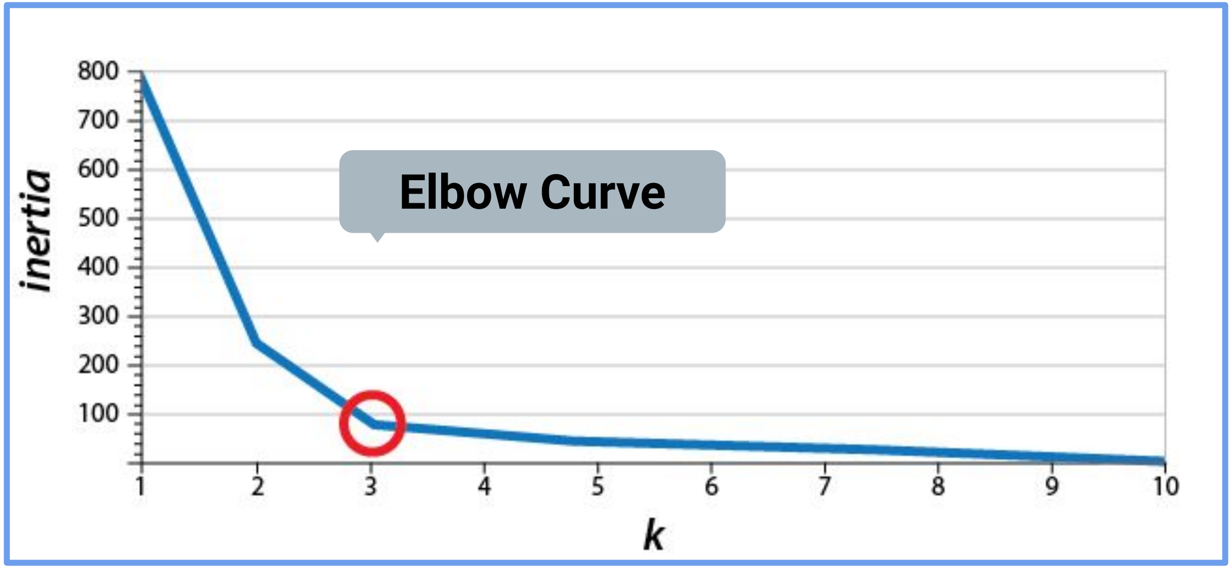


What is the Best Number for k?

Instructor Do: The K-Means Algorithm

This is done using an **elbow curve**, where the *x axis* is the *K*-value and the *y axis* is some *objective function*.

A common objective function is the *inertia*.



Questions?

Activity: Customers Segmentation

In this activity, you will give continuation to the project with the e-commerce company. Now that you have prepared the data, it’s time to start looking for patterns in the customer data. The CFO has asked you to group customer based on their spending habits. You decided to use k-means to perform this task!

Suggested Time:
20 Minutes

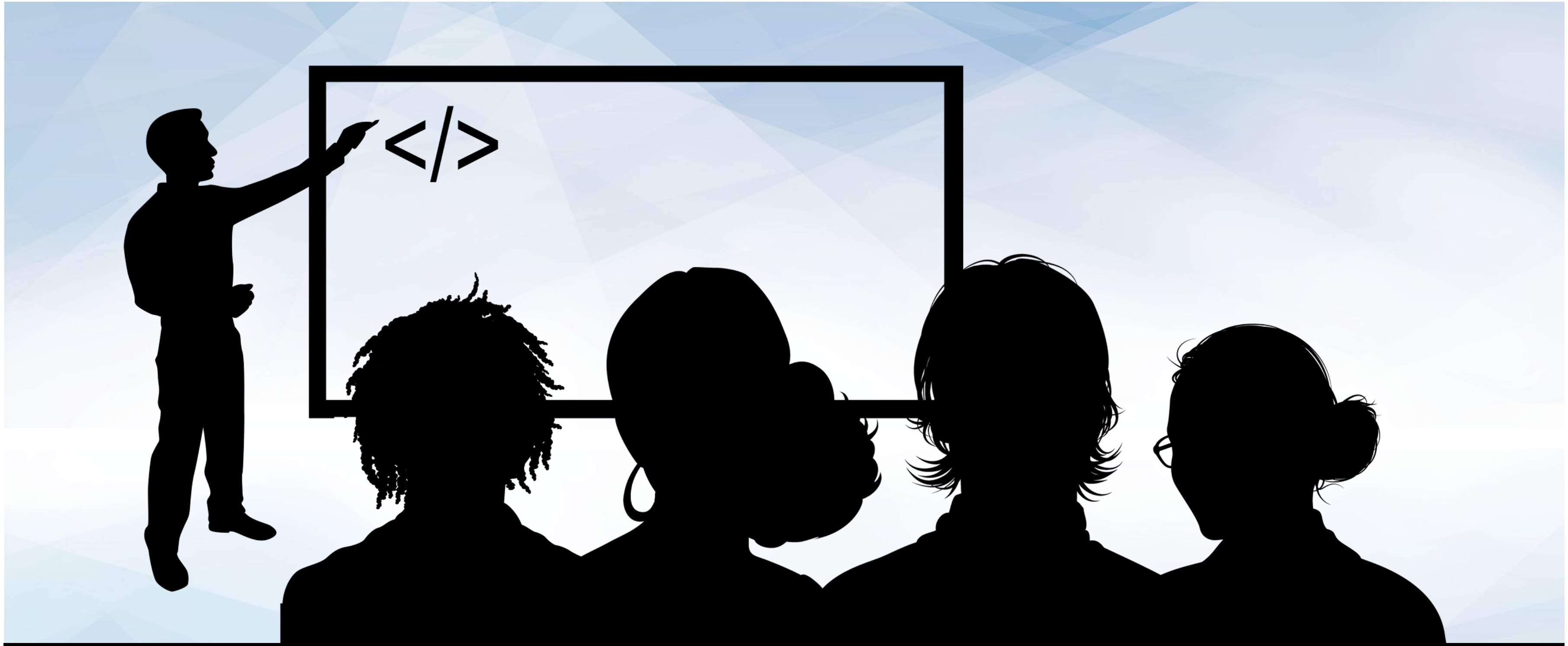
Instructions:

Activity: Customers Segmentation

- Accomplish the following tasks and use k-means to cluster the customer data.
  - Load the dataset (which you previously cleaned) into a DataFrame.
  - Find the best number(s) of clusters using the elbow curve.
  - Create a 2-D scatter plot to analyze the clusters using `x="Annual Income"` and `y="Spending Score (1-100)"`.

- **Bonus:**
  - Create a function called `get_clusters(k, data)` that finds the `k` clusters using k-means on `data`.
    - `data` represents a dataframe.
    - The function should use k-means to identify clusters in the dataset.
    - The function should add a new column containing the cluster value of each sample (row).
    - The function should return a copy of the new dataframe.
  - Create a function called `show_clusters(df)` that will create a scatter plot of a dataframe's `Annual Income` and `Spending Score (1-100)` columns, and color by the cluster.

Let’s Review



Instructor Demonstration

Speed Up Machine Learning with PCA

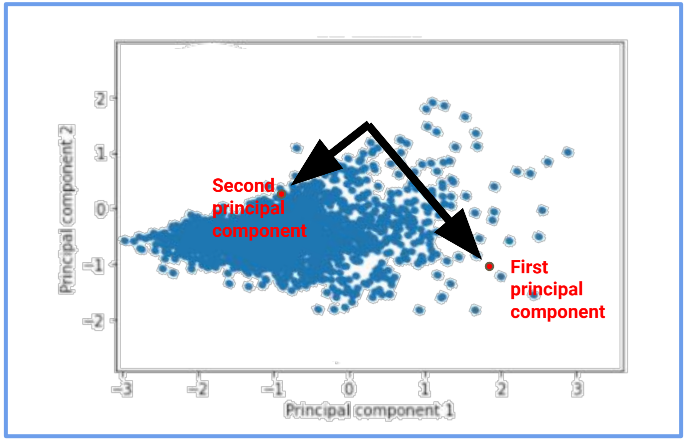
Instructor Do: Speed Up Machine Learning with PCA

Why use it?

- Simply put, PCA was designed to save both time and computing resources when dealing with enormous datasets.

How does it work?

- It does so by reducing the number of input features (or dimensions).
- The PCA algorithm transforms a large set of variables into a smaller one that contains most of the information in the original large set.



- This plot illustrates well what PCA does.
- PCA is mainly used for dimensionality reduction, not for visualization.
- We will cover **t-SNE** next class, which is mostly used to visualize high dimensional data.

Activity: PCA in Action

In this activity, you will use PCA to reduce the dimensions of the consumers shopping database from `4` to `2` features. After applying PCA, you will use the principal components data to fit a k-means model with `k=6` and make some conclusions.

Suggested Time:

20 Minutes

Instructions:

Activity: Customers Segmentation

- Load the dataset.
- Standardize the data of all the features.
- Apply PCA to reduce the dataset to 2 dimensions.
- Compute the explained variance.
- Is the explained variance sufficiently high at `n_components=3`? If not, try reducing to 3 dimensions instead.
- Train the k-means algorithm with the reduced data at `k=5`.

- **Bonus:**
    - Install Plotly for Python in your current virtual environment. Uncomment and run the code at the end of the notebook to visualize the dataset in 3 dimensions.

Let's Review