Topics in Data Science

(Lecture 1)

Mohammad Pourhomayoun
Assistant Professor
Computer Science Department
California State University, Los Angeles

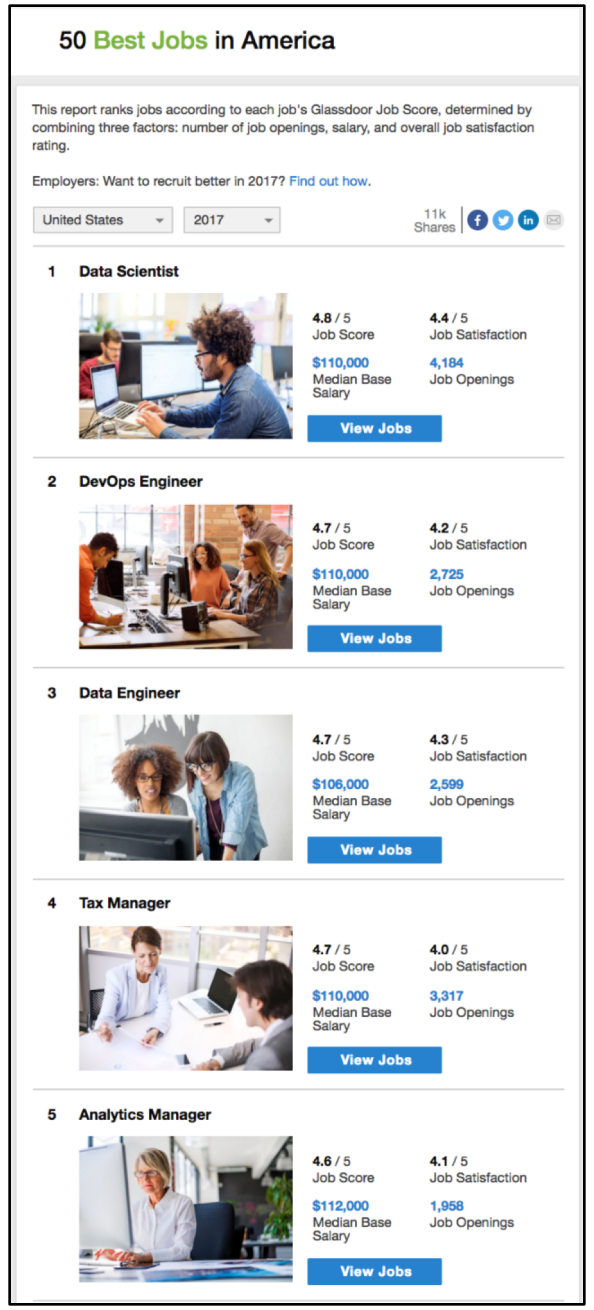
Review: What is Data Science?

Data Science

Data Science is an interdisciplinary field of research that aims to design and develop automated techniques to **extract knowledge from large-scale data** and use it for **future purposes** such as **prediction** or **decision making**.

Who is a Data Scientist?

- **Glassdoor:**
  - "**Data Scientist**" is rated **#1 in the list of Best Jobs in America in 2017 and 2018**.
  - In this list, the jobs are determined by combining three key factors:
    - **number of job openings**
    - **salary**
    - **career opportunities rating**



Why Is Data Science So Important Now?

- Why is Data Science an important topic these days?

  (why didn't anyone talk about it 10 years ago?)

- Because now we have:
  1. New Sources of Data that did not exist before.
  2. New Capabilities to acquire, store, and process data.
  3. Thanks to 1 & 2, we were able to develop new algorithms and methods to better extract knowledge from raw data!

CS5661: Course Overview

- **Course Overview**: In this course, we will cover more advanced algorithms, techniques, and tools for machine learning and data processing including Artificial Neural Networks, Deep Learning, Convolution Neural Networks, and other advanced machine learning methods such as Advanced Ensemble Learning and SVM, as well as techniques in Data Processing, Data Analytics, Dimensionality Reduction, and visualization (as time permits!).

- We will cover both theoretical and practical aspects of these methods.

Course Overview

- **Instructor**: Mohammad Pourhomayoun
- **Email:** mpourho@calstatela.edu
- **Class:** Friday, 11:30 – 2:00 PM, ASCB 132
- **Office Hours:** Wed: 1:30 PM - 2:30 PM, Fri: 2:30 PM - 3:30 PM
- **Office:** ET, A408

Evaluation

- Assignments: 40%
  - Group assignments
  - Theoretical Problems and Implementation and Programming
  - Assignments are due at the beginning of class on the due date. **Late submissions will not be accepted**.
  - Copying homework/project from other groups is considered as cheating!

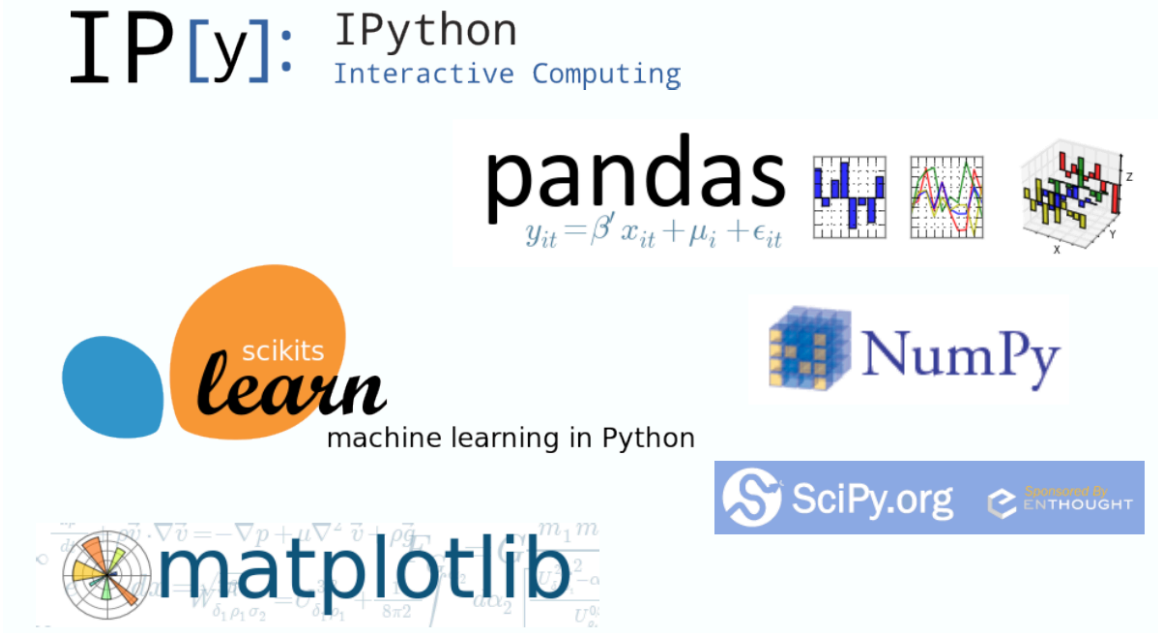
- Final Project: 20%
- Final Exam: 40%
- Participation: 5-10%

In this class…

- Feel free to participate in class discussions …, There is nothing to be ashamed of! **WE ARE ALL FRIENDS!** ☺
- Your question is important! Please feel free to ask!
- Don’t hesitate to interrupt when you have a question.
- Please let me know if you want me to repeat or clarify something.
- Your Feedback is precious to me!

Python Programming

- In this class, we use Python for all homeworks and projects. Python is very powerful and highly popular for Data Science purposes. We can name it the main programming language for Data Science!
- Python includes unique powerful libraries for data science. Also, most of Data Science frameworks support python.



Review of
Main Concepts and Definitions

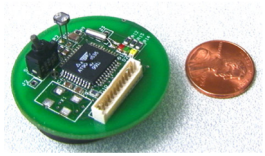
What is Data Science?

Data Science is an interdisciplinary field of research that aims to design and develop automated techniques to **extract knowledge from large-scale data** and use it for **future purposes** such as **prediction** or **decision making**.

- It can be an integration/extension of statistics, machine learning, predictive analytics, and computing.






Only Some of the Applications!

Stock Market Prediction

Zillow®

REDFIN™

Real State Prediction

amazon

facebook

Online Shopping and Advertisements

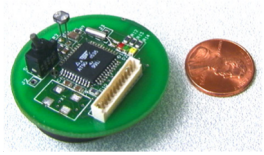
Recommendation Systems

Self-Driving Cars

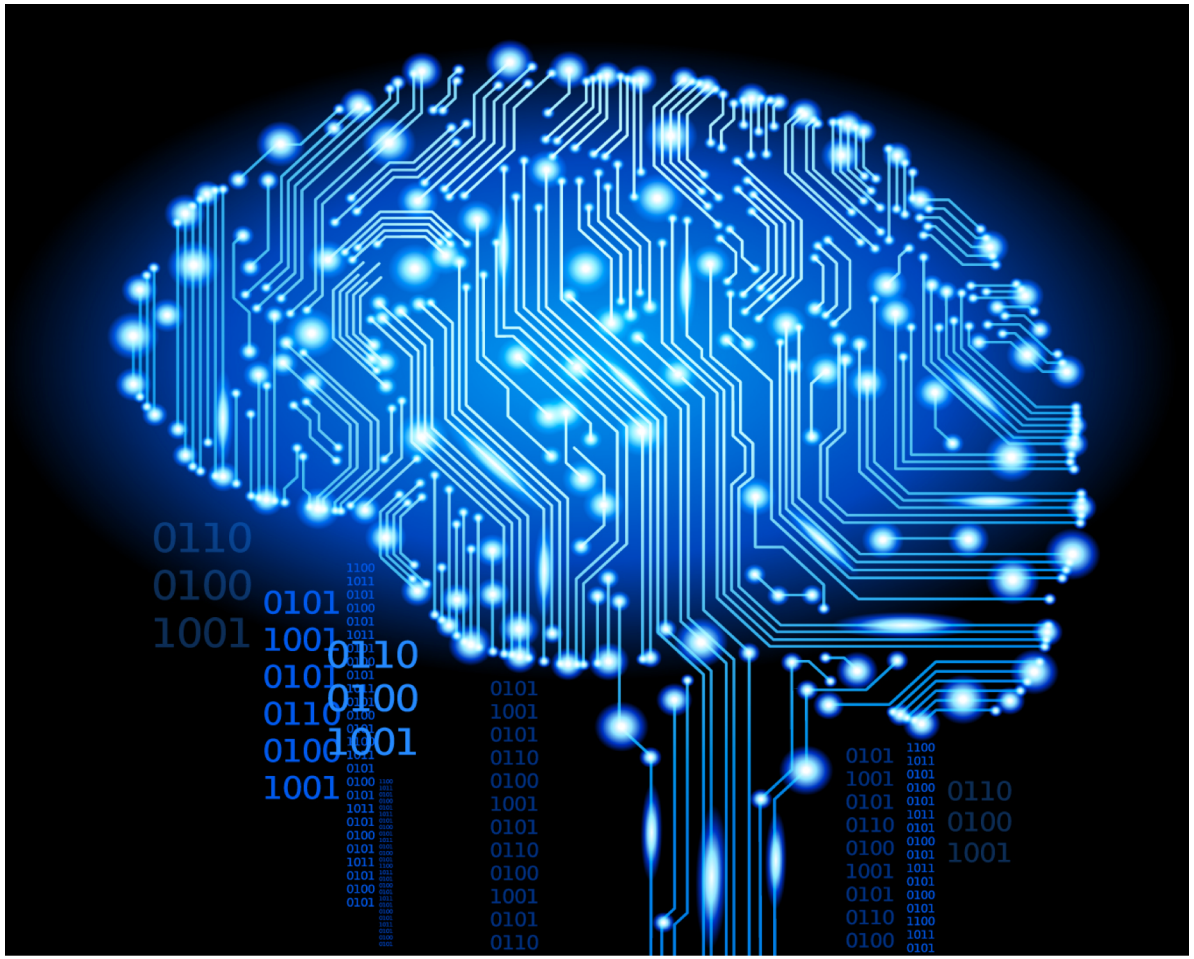
Healthcare

Ingredients

- **Data**
  - E.g. WWW, Social Networks, Online Activities, Smart Phone, Wearables, Sensor networks, Science, …
- **Machine Learning Algorithms**
  - E.g., recommendation system, market prediction, speech recognition, Face detection, Fraud detection, Spam filtering, vehicle control, Medical diagnosis, ...
- **Big Data Manipulation Techniques**
  - Large-Scale Data Processing, Distributed Computing, Cloud Computing

What is Machine Learning?



Review: What is Machine Learning?

- **A Definition:** Designing and constructing methods that learn from **existing data** and make predictions on **future data**.

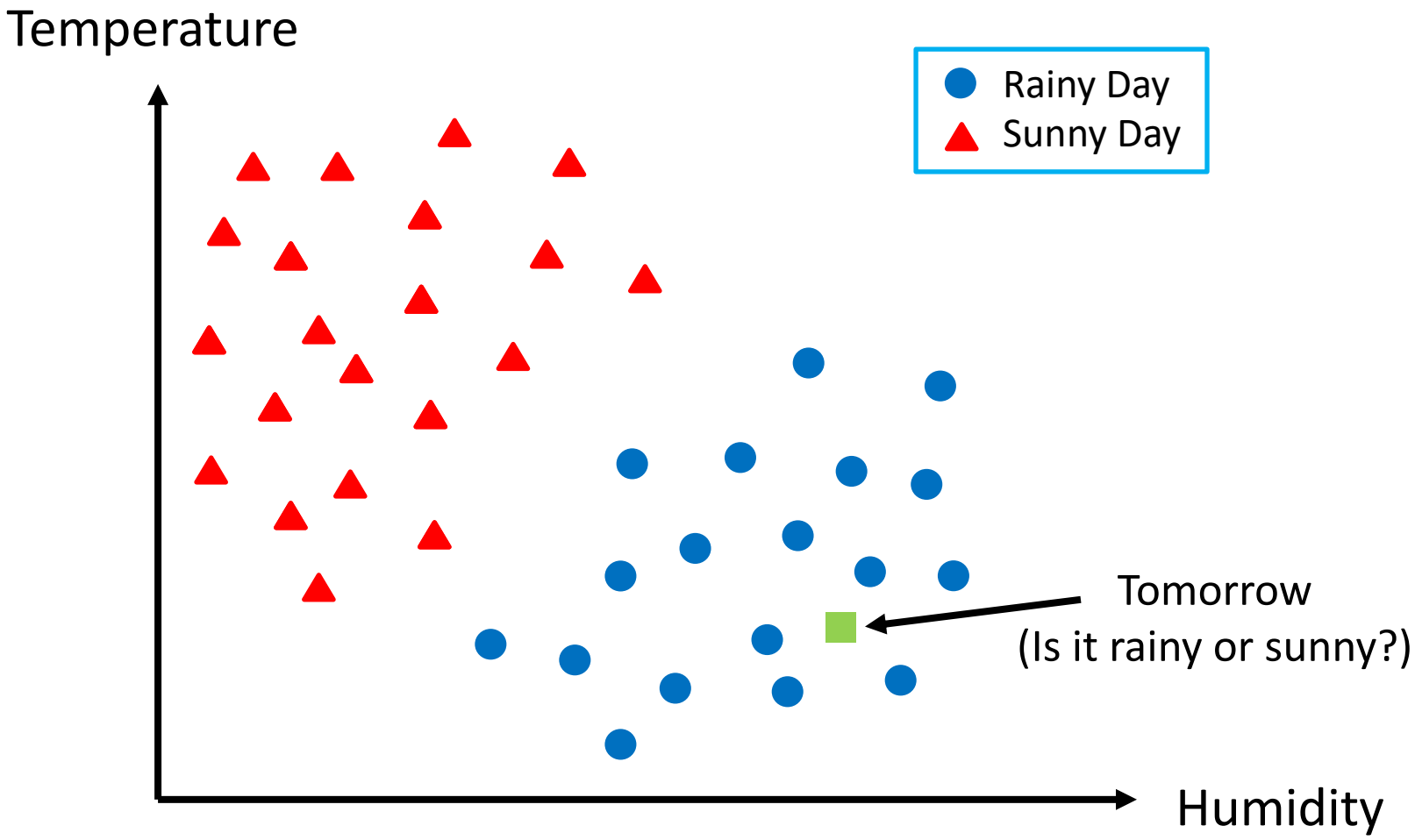
- **Another Definition:** A set of algorithms that can automatically detect and extract patterns in **existing data**, and then use the extracted patterns to predict on **future data**, or to perform other kinds of decision making.

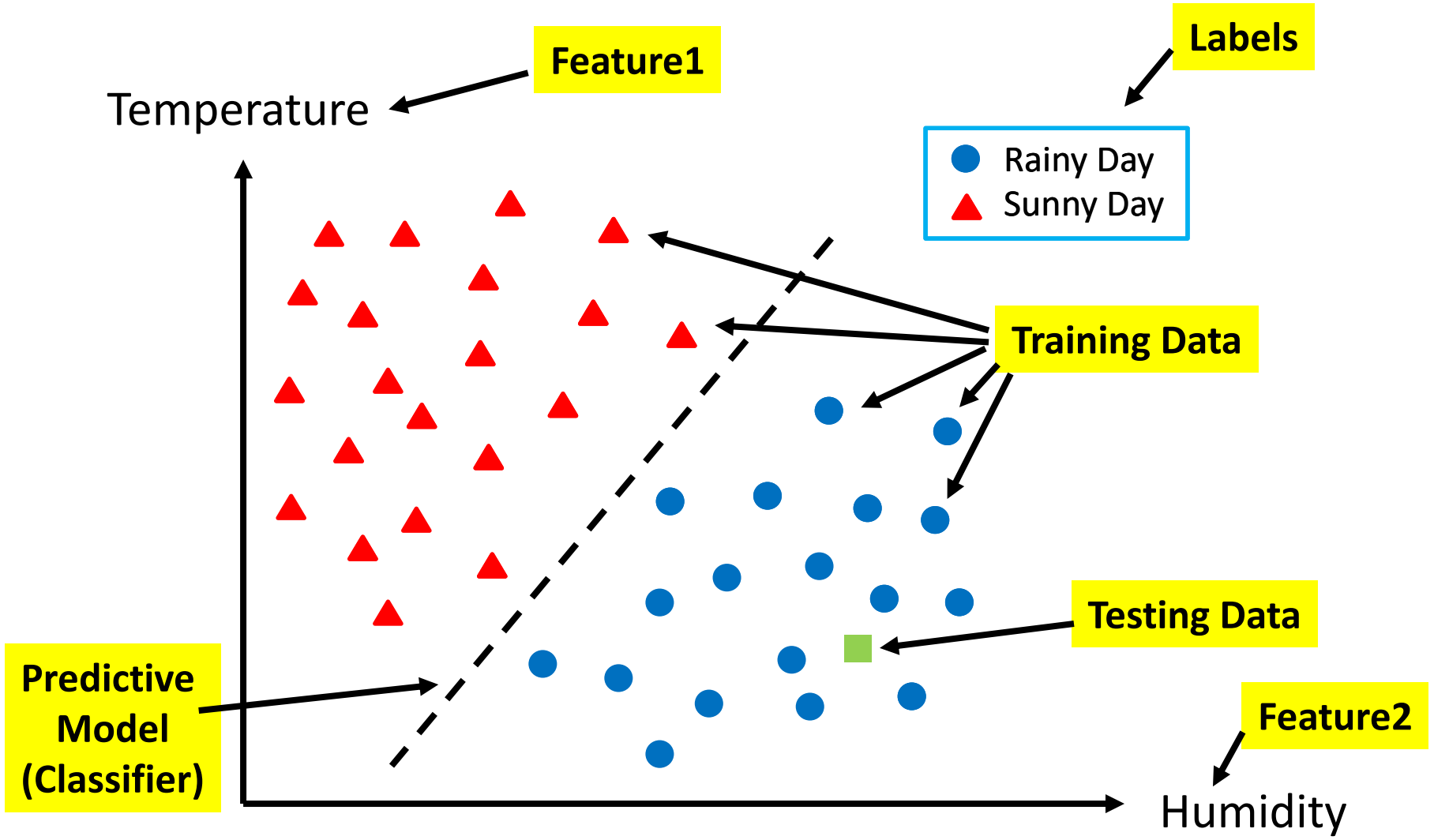
Machine Learning Algorithms

- **ML Algorithms Learned in CS4661:**
  - KNN, Decision Tree, Linear Regression, Logistic Regression, Kmeans, Stochastic gradient descent, Batch gradient descent, Random Forest, ...

- **More advanced ML Algorithms covered in CS5661/4662:**
  - Support Vector Machine (SVM), Artificial Neural Networks (ANN), Deep Learning methods, Convolutional Neural Networks, Boosting/Bagging methods, Principal Component Analysis (PCA), ...

Example: Weather Forecasting




Feature1
Temperature
Labels
Rainy Day
Sunny Day
Training Data
Testing Data
Predictive Model (Classifier)
Feature2
Humidity

Review: Terminology

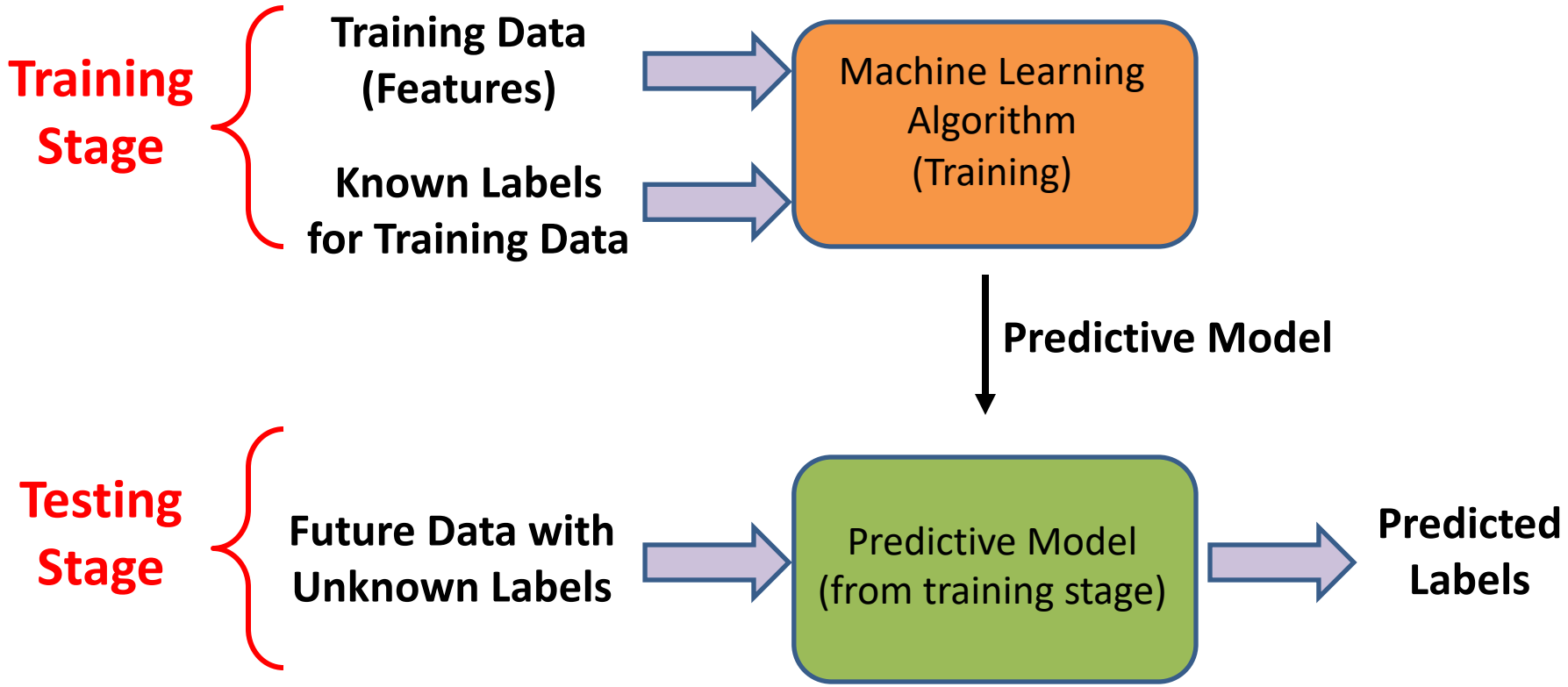
- **Observations**: Data Samples (Data Examples).
- **Features (inputs)**: Attributes that represent an observation, e.g., temperature, humidity
- **Labels (outputs):** Values assigned to observations (also called **class**, **target**), e.g., rainy/sunny
- **Training Data:** Past observations given to the ML algorithm for training. E.g. temperature and humidity of the past 30 days, along with the label for each day.
- **Testing Data:** Observations given to a "predictive model" for prediction.

Review: More Terminology

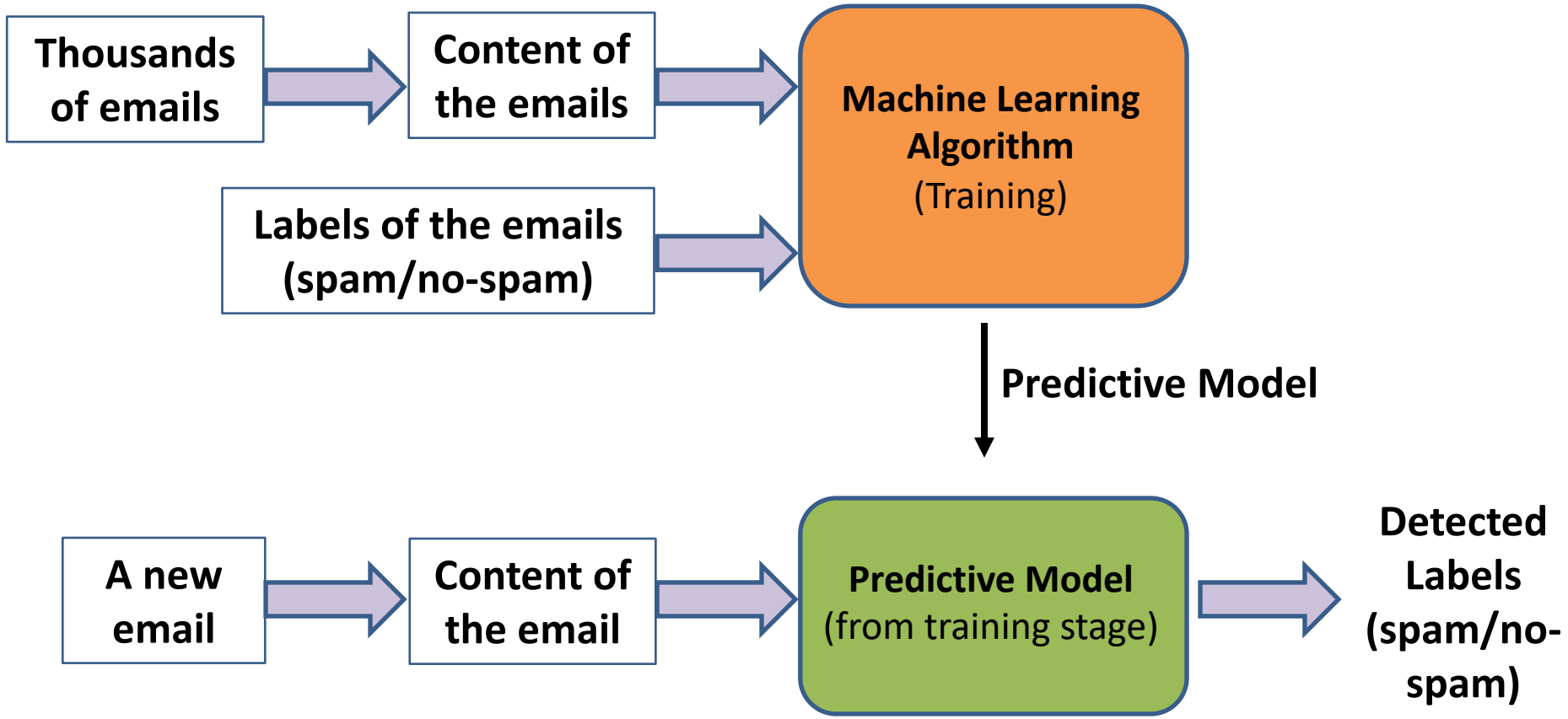
- **Training Stage (Modeling):** Building a predictive model based on the training dataset.
  - The model does not have to be perfect. As long as it is close, it is useful.
  - We should tolerate randomness and mistakes.

- **Testing Stage (Prediction):** Applying the trained model to forecast what is going to happen in future (on future testing data)

Supervised Learning: Learning from labeled Data



Example for Supervised Learning: Spam Detection

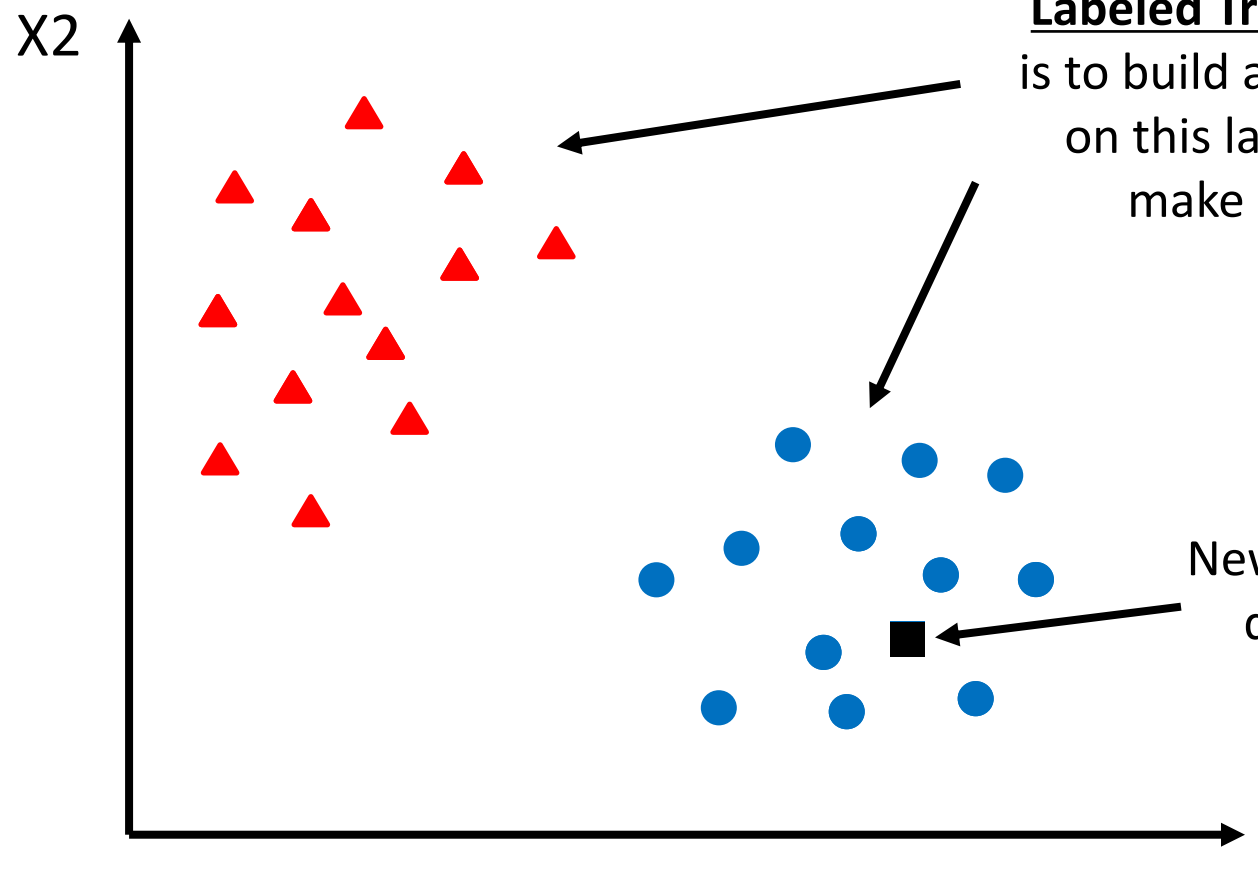


Review: More Terminology

- **Supervised learning:** Learning from labeled observations.
  - The algorithm is presented with **training inputs and their known labels**, and the goal is to train a model that maps future inputs to new labels.

- **Unsupervised learning:** Learning from unlabeled observations.
  - **Discover hidden patterns and latent structure** from features alone.
  - Data exploration

Supervised Learning

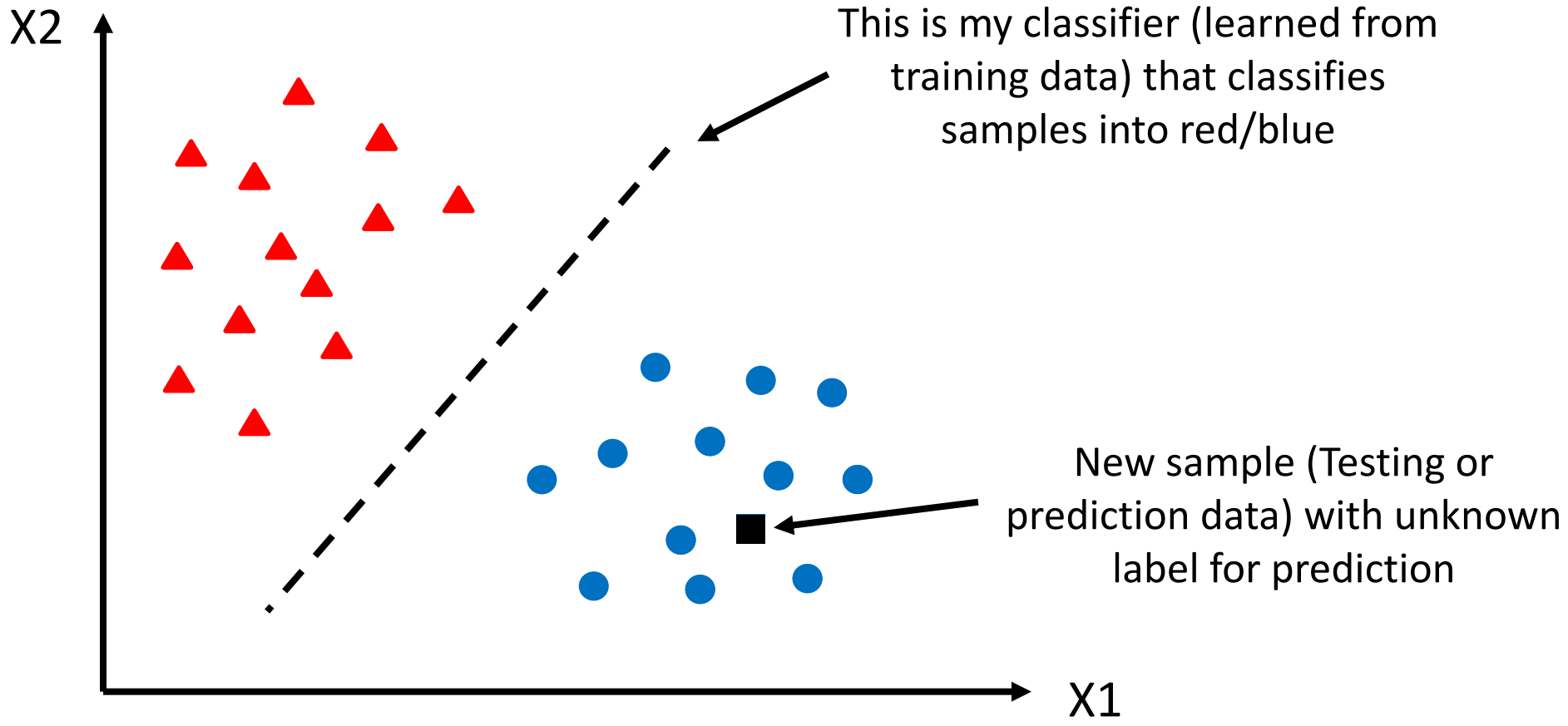


Labeled Training Dataset: the goal is to build a predictive model based on this labeled training data, to make prediction in future

New sample (Testing or prediction data) with unknown label for prediction

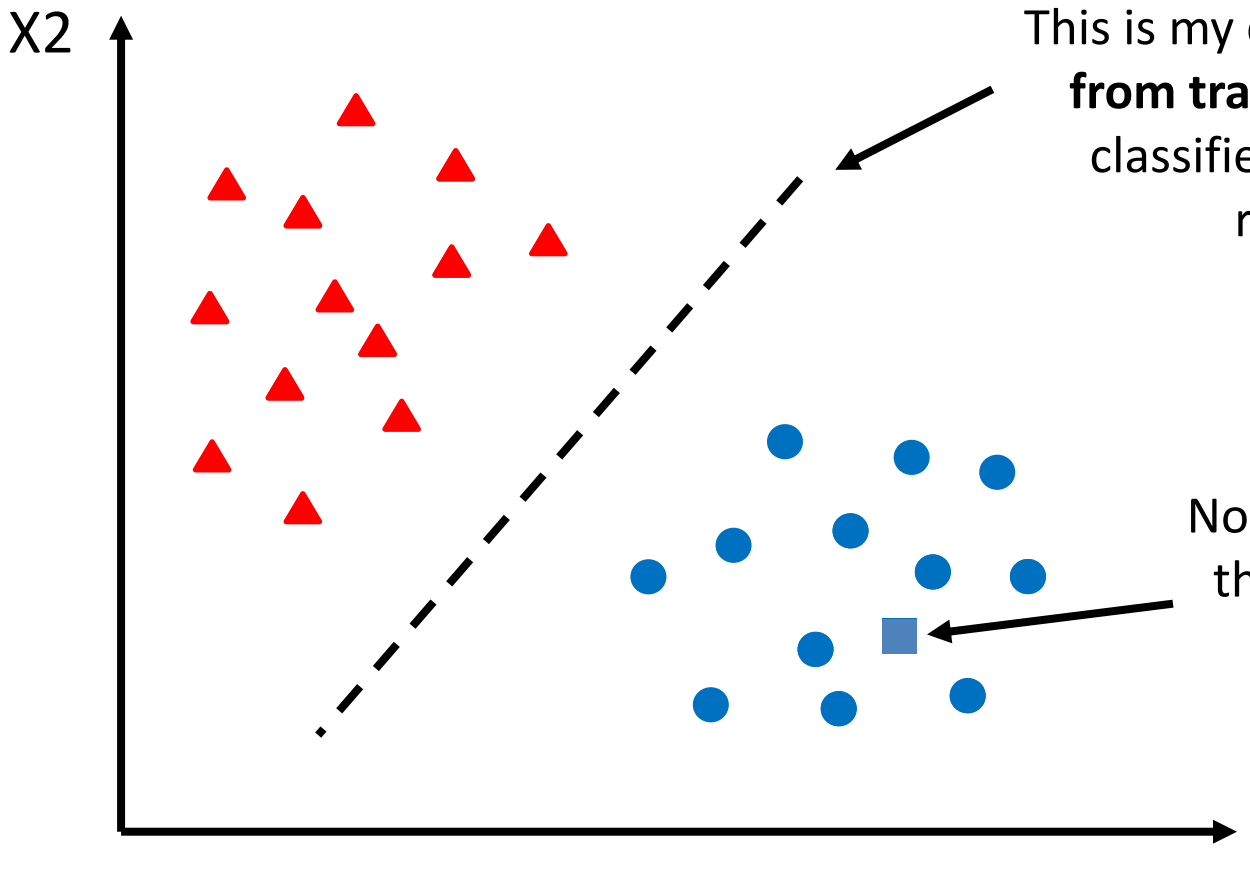
Training set: $\{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), (x^{(3)}, y^{(3)}), \ldots, (x^{(m)}, y^{(m)})\}$

Supervised Learning



Training set: $\{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), (x^{(3)}, y^{(3)}), \ldots, (x^{(m)}, y^{(m)})\}$

Supervised Learning

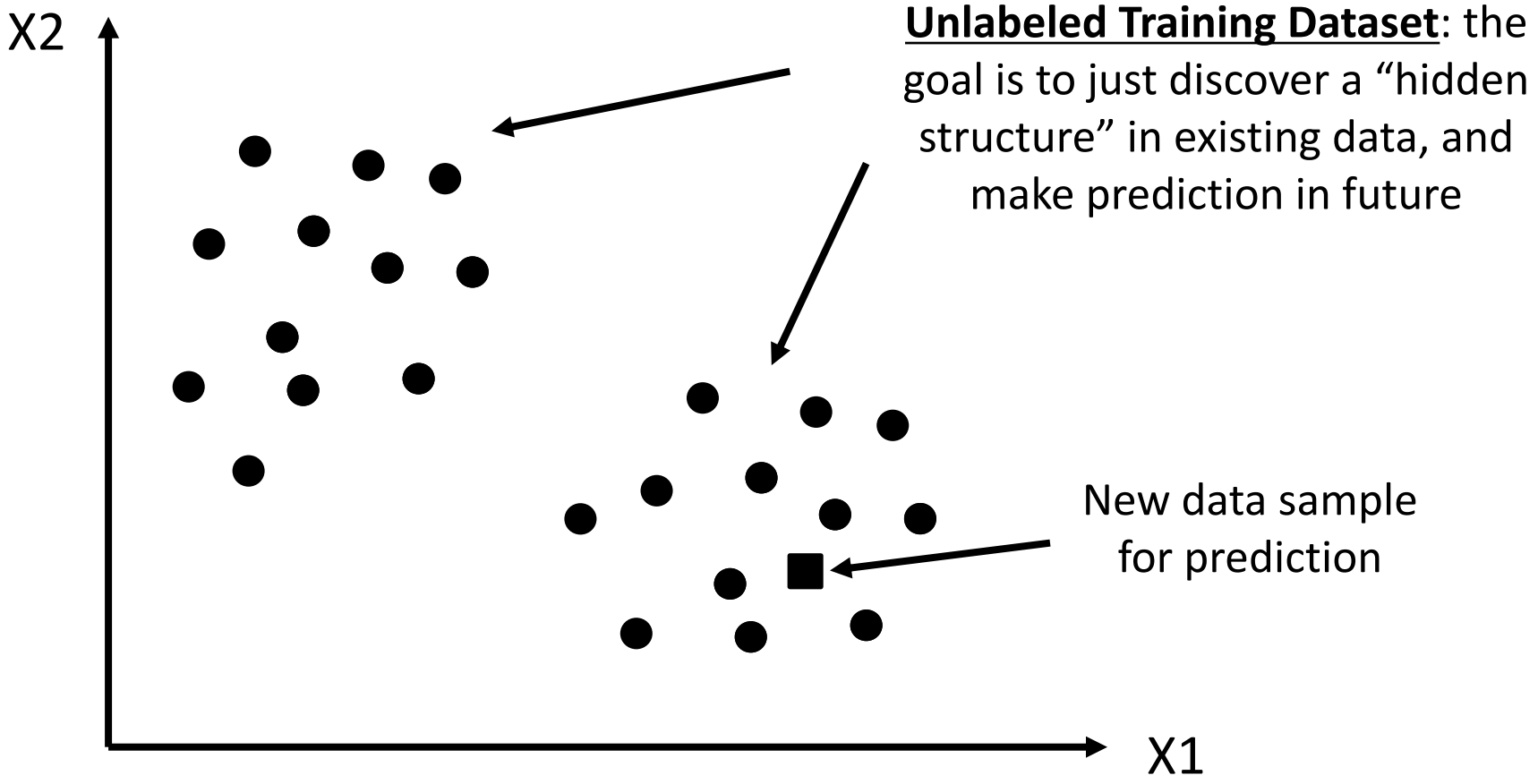


This is my classifier (**learned from training data**) that classifies samples into red/blue

Now, we can conclude (or predict) that the new sample (Testing or prediction data) is Blue!

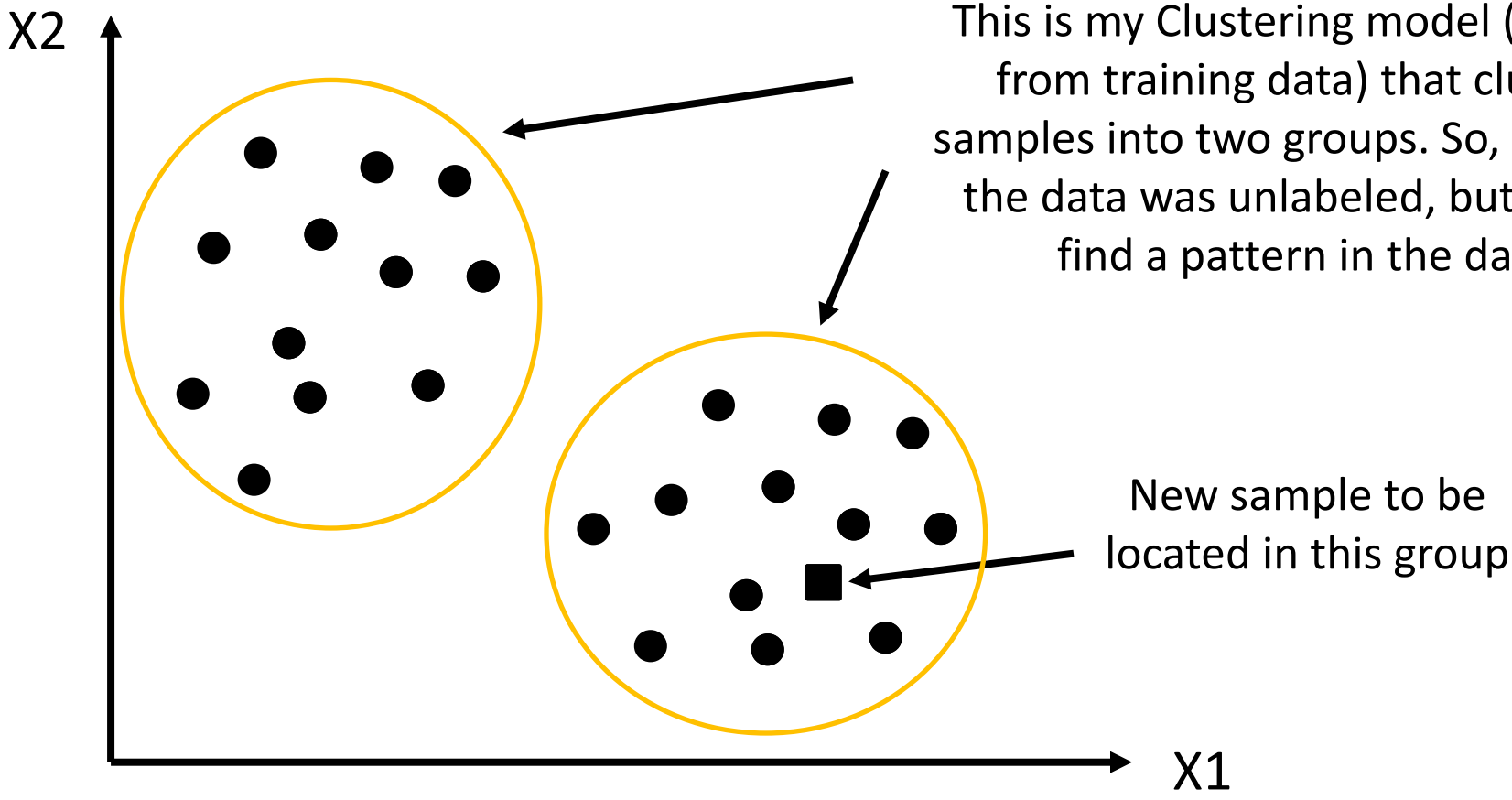
Training set: $\{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), (x^{(3)}, y^{(3)}), \ldots, (x^{(m)}, y^{(m)})\}$

Unsupervised Learning



Training set: $\{x^{(1)}, x^{(2)}, x^{(3)}, \ldots, x^{(m)}\}$

Unsupervised Learning



This is my Clustering model (learned from training data) that clusters samples into two groups. So, although the data was unlabeled, but I could find a pattern in the data.

New sample to be located in this group

Training set: $\{x^{(1)}, x^{(2)}, x^{(3)}, \ldots, x^{(m)}\}$

Thank You!

Questions?