Advanced Machine Learning and Deep Learning

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

CS4662: Course Overview

- **Course Overview**: In this course, we will cover more advanced algorithms, techniques, and tools for machine learning and data processing including Advanced Ensemble Learning, Support Vector Machine (SVM), Artificial Neural Networks and Deep Learning, Principal Component Analysis (PCA), as well as techniques in Data Processing, Data Analytics, Dimensionality Reduction, and visualization.

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

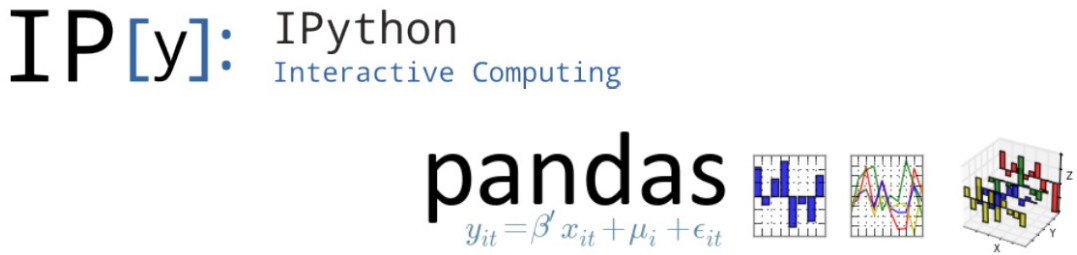
Evaluation

- Homework Assignments: 30%
  - Theoretical Problems, Implementation, Programming
  - **Late submissions will not be accepted**!
  - Copying homework/project from others is considered as cheating and is not tolerable!
- Quiz: 25%
- Final Project (group project): 20%
- Final Exam: 25%
- Participation Bonus: 5-10%

- In this course, we work with **Python 3.x** and its libraries.

Python Programming

- In this class, we will teach and use Python.
- We will work with new libraries, frameworks and tools for data science, data processing, machine learning, and deep learning.



Review of
Main Concepts and Definitions

Review: What is Data Science?

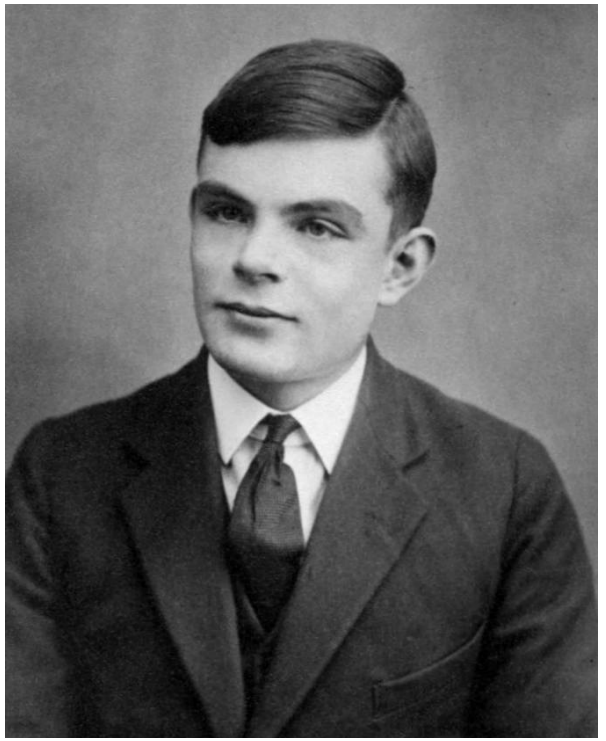
Data Science is an interdisciplinary field of research that aims to design and develop automated techniques to **extract knowledge from large-scale data** and use it for **future purposes** such as **prediction** or **decision making**.

What is Artificial Intelligence (AI)?

- **Artificial Intelligence (AI)** is a branch of computer science and Data Science that aims to build machines (computers) that **can mimic human intelligence**, such as "learning" , ”decision making” and "problem solving".

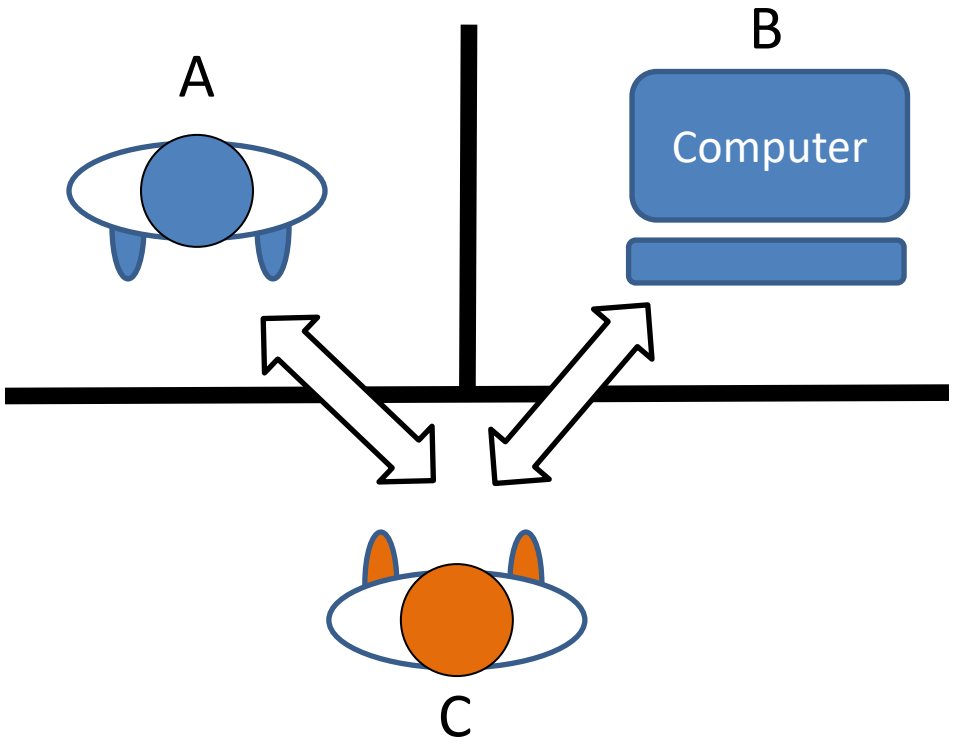
History

- Alan Turing (1912 –1954).
- Turing published a paper in 1950, titled "*Computing Machinery and Intelligence*", that raised this question: "***Can Machines think?***"
- Alternative question: ***Can a computer do well in "Imitation Game"?***
- A modified version of that is called "*Turing Test*".



Turing Test

- If the evaluator cannot distinguish the computer from the human (after communicating with written questions/answers), the computer is said to have **passed the Turing Test**.
- If a computer can pass the test, it could be considered "*intelligent*".



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.

Only Some of the Applications!

Online Shopping and Advertisements

Speech Recognition and Natural Language Processing

Marketing

Computer Vision

Recommendation Systems

Self-Driving Cars

Healthcare

Social Media

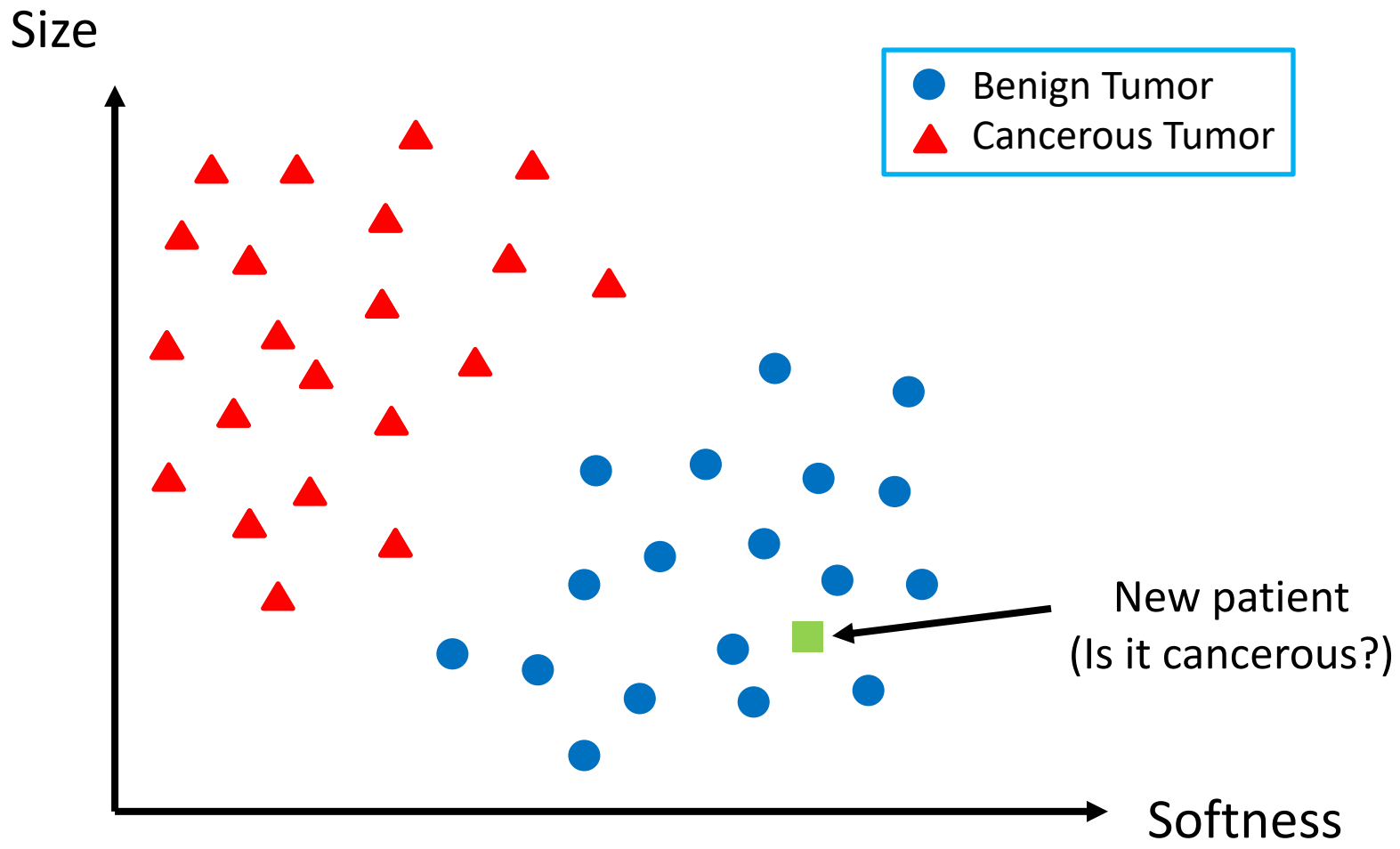
Machine Learning Algorithms

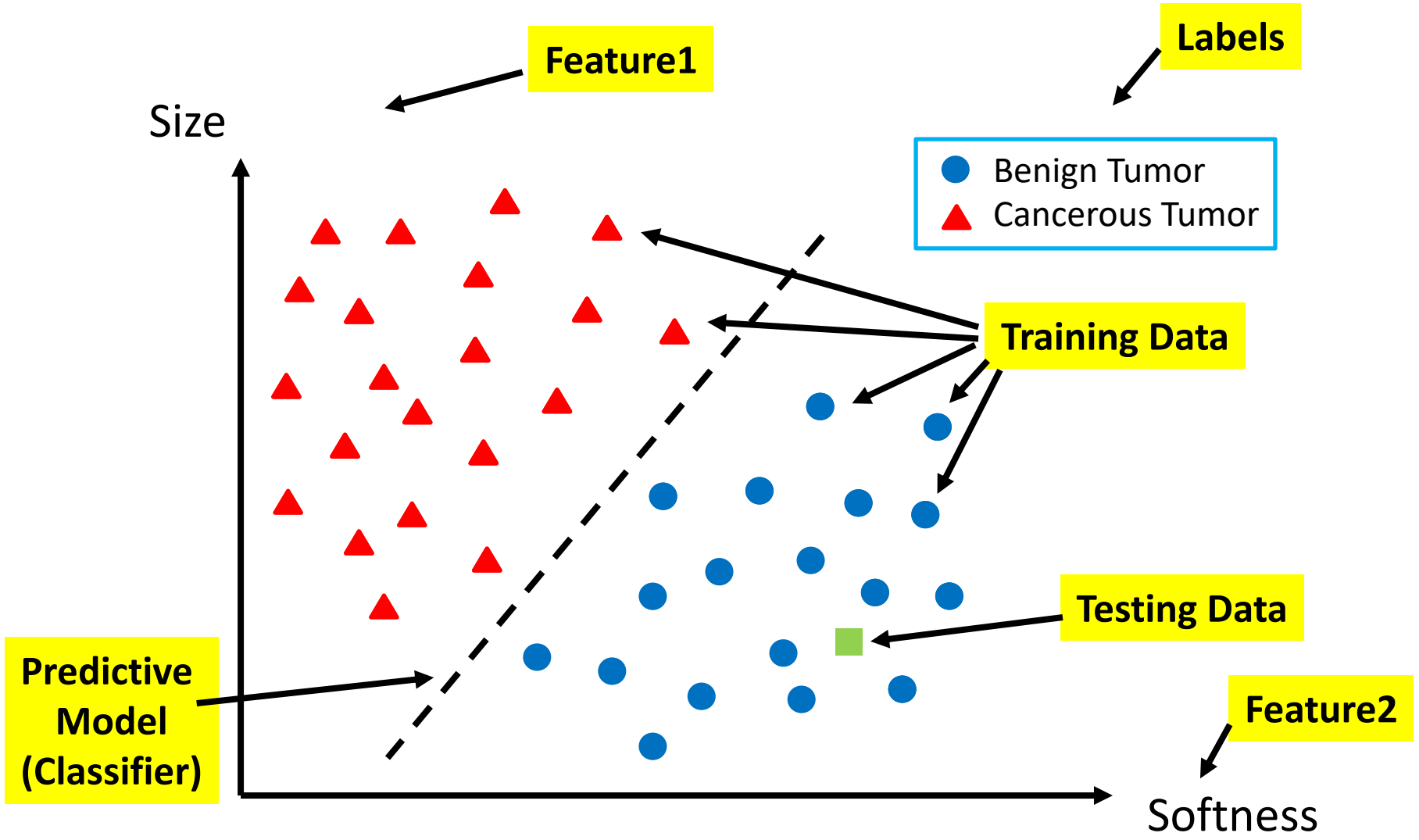
- **ML Algorithms Learned in CS4661:**
  - KNN, Decision Tree, Linear Regression, Logistic Regression, Kmeans, Stochastic gradient descent, Batch gradient descent, Random Forest, ...

- **More advanced ML Algorithms covered in CS4662:**
  - Artificial Neural Networks (ANN), Deep Learning methods, Convolutional Neural Networks, Support Vector Machine (SVM), Boosting/Bagging methods, Principal Component Analysis (PCA), ...

Example: Predicting Cancer

This is a super simple and too ideal example, but it illustrates the idea.




Feature1
Labels
Size
Benign Tumor
Cancerous Tumor
Training Data
Testing Data
Predictive Model (Classifier)
Feature2
Softness

Review: Terminology

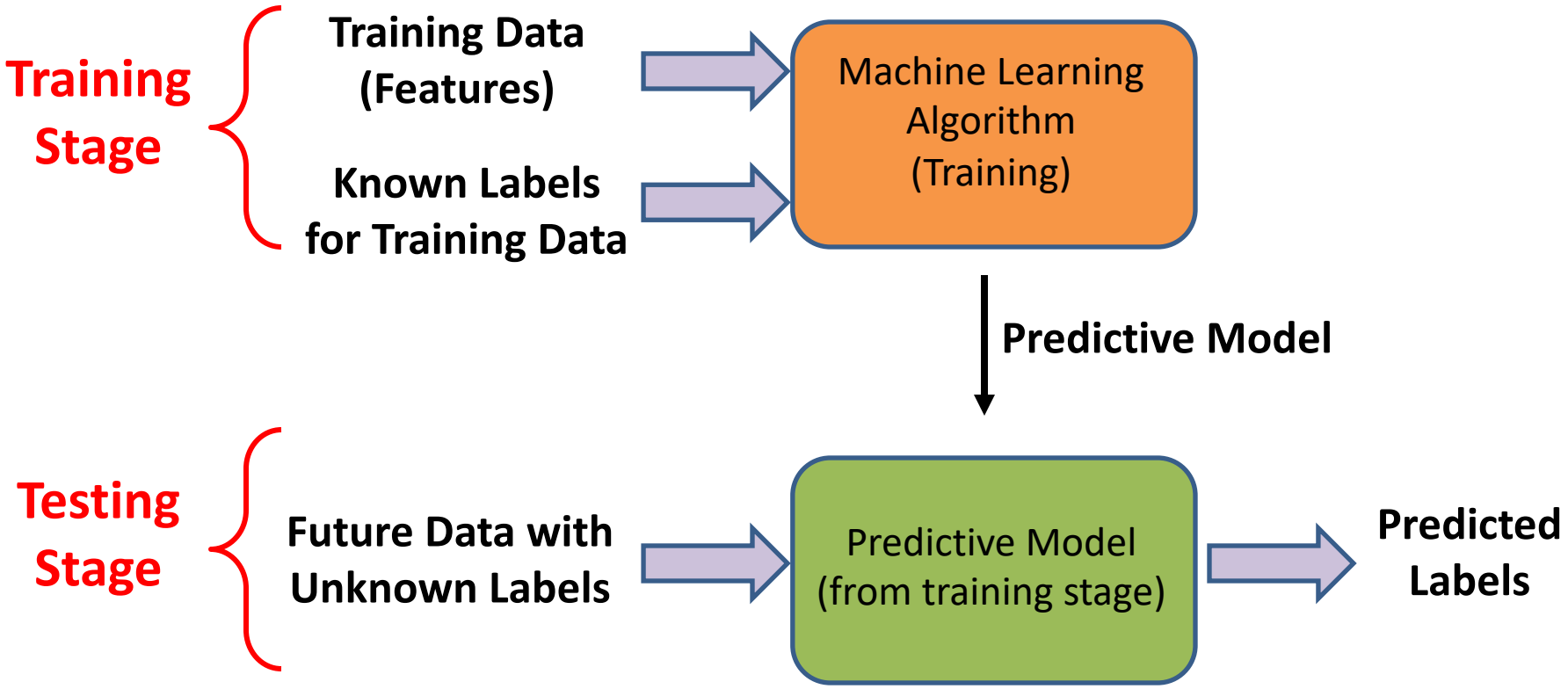
- **Observations**: Data Samples (Data Examples).
- **Features (inputs)**: Attributes that represent an observation, e.g., temperature, humidity
- **Labels (outputs):** Values assigned to observations (also called **class**, **target**), e.g., rainy/sunny
- **Training Data:** Past observations given 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

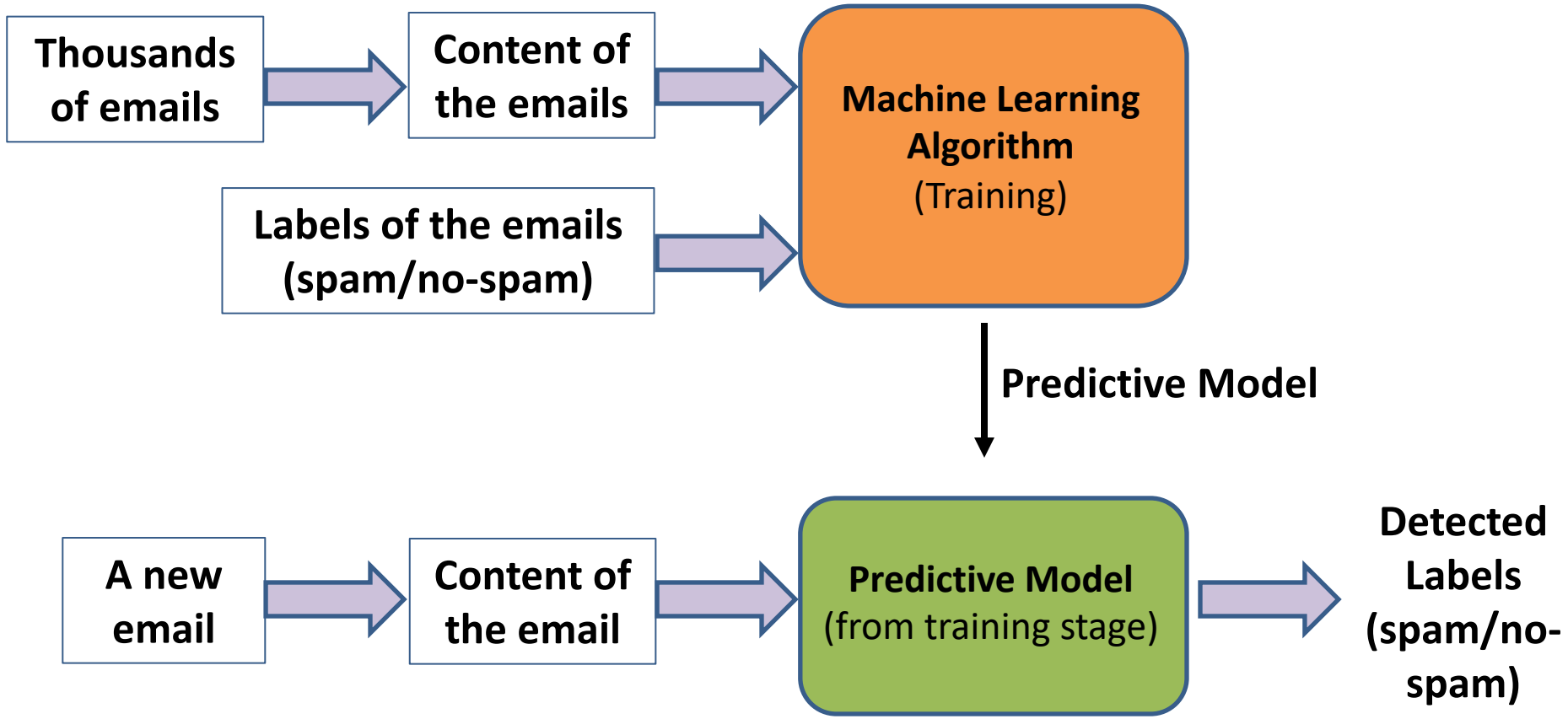
- **Training Stage (Modeling):** Building a predictive model based on the training dataset.
  - The model does not have to be perfect. As long as it is close, it is useful.
  - We should tolerate randomness and mistakes.

- **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

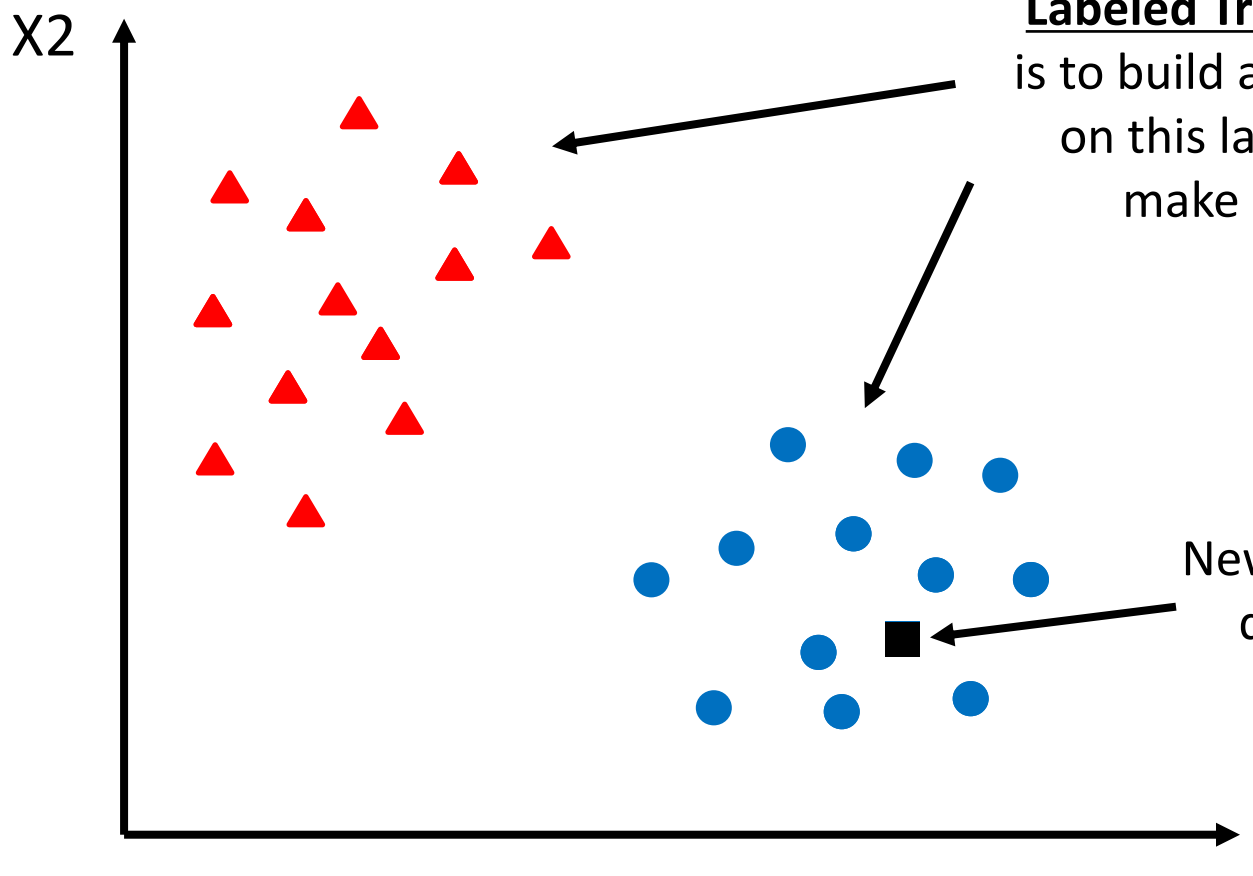


Review: More Terminology

- **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

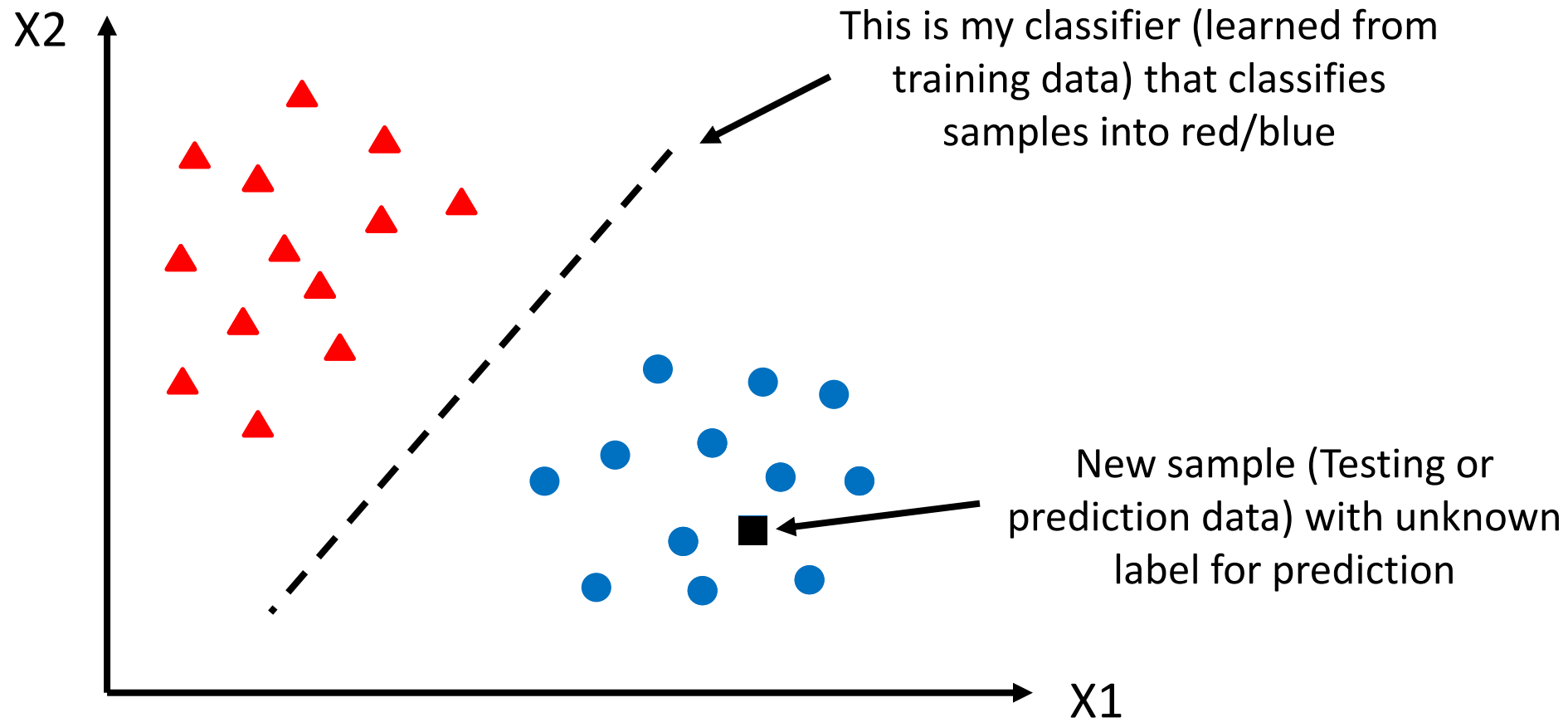


Labeled Training Dataset: the goal is to build a predictive model based on this labeled training data, to make prediction in future

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

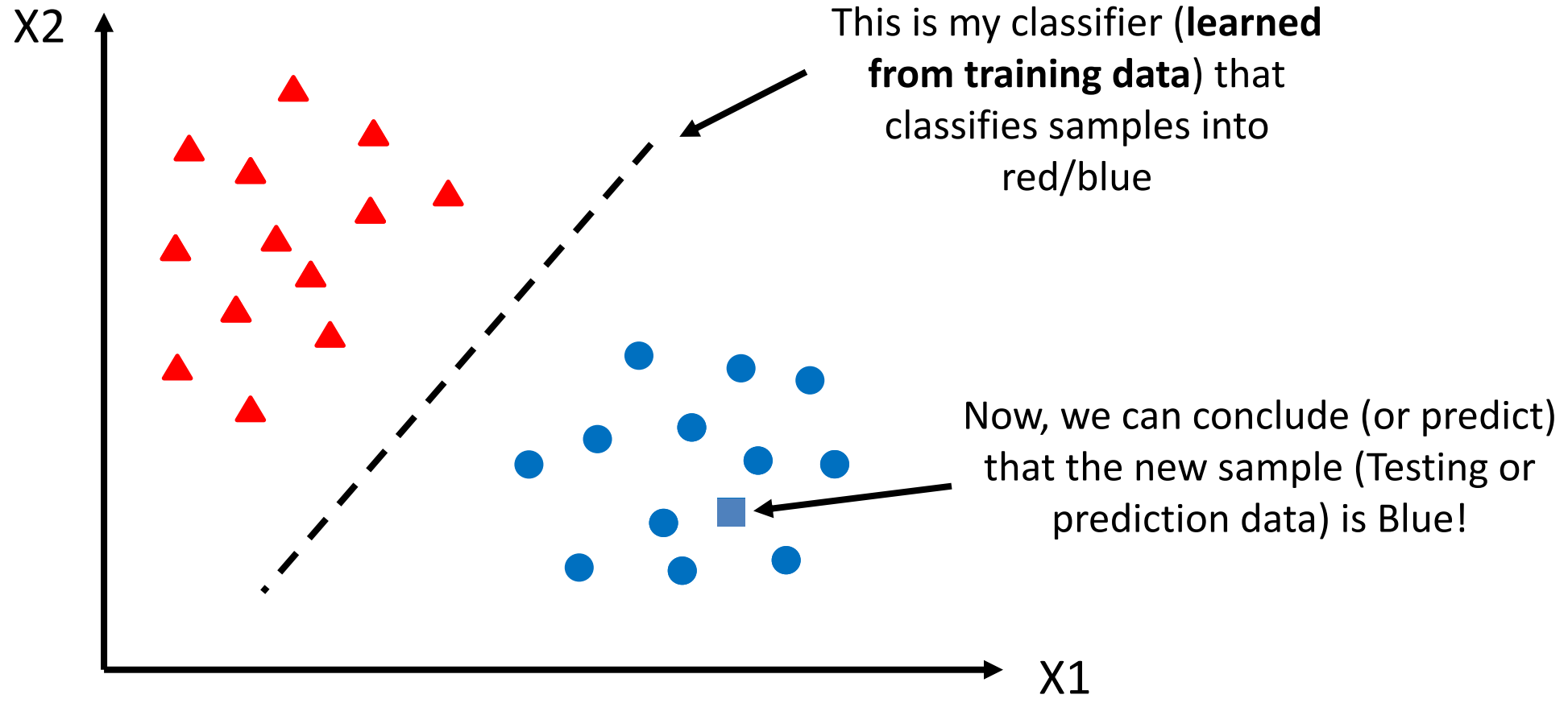
Training set: $\{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), (x^{(3)}, y^{(3)}), \ldots, (x^{(m)}, y^{(m)})\}$

Supervised Learning



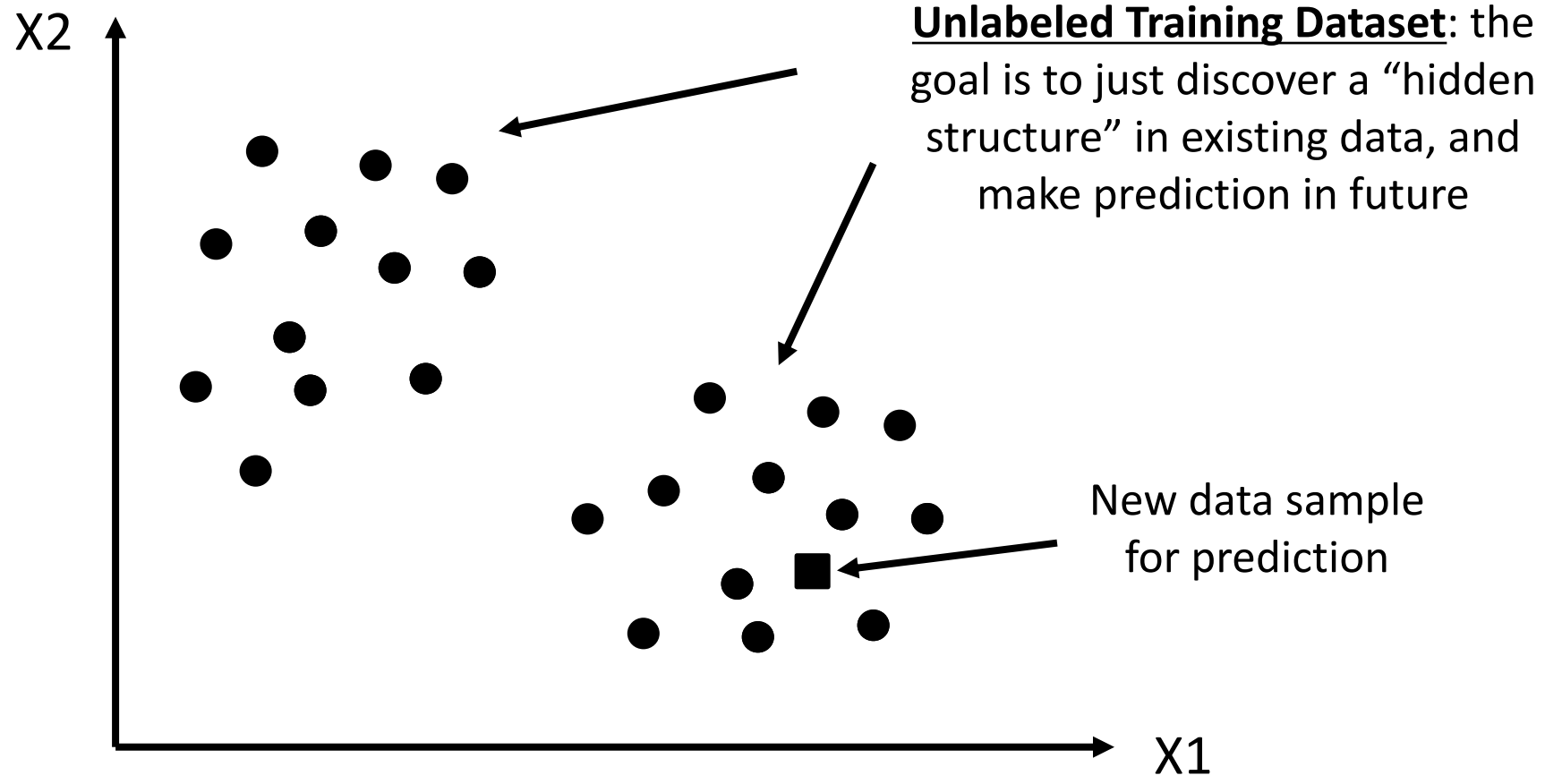
Training set: $\{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), (x^{(3)}, y^{(3)}), \ldots, (x^{(m)}, y^{(m)})\}$

Supervised Learning



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

Unsupervised Learning

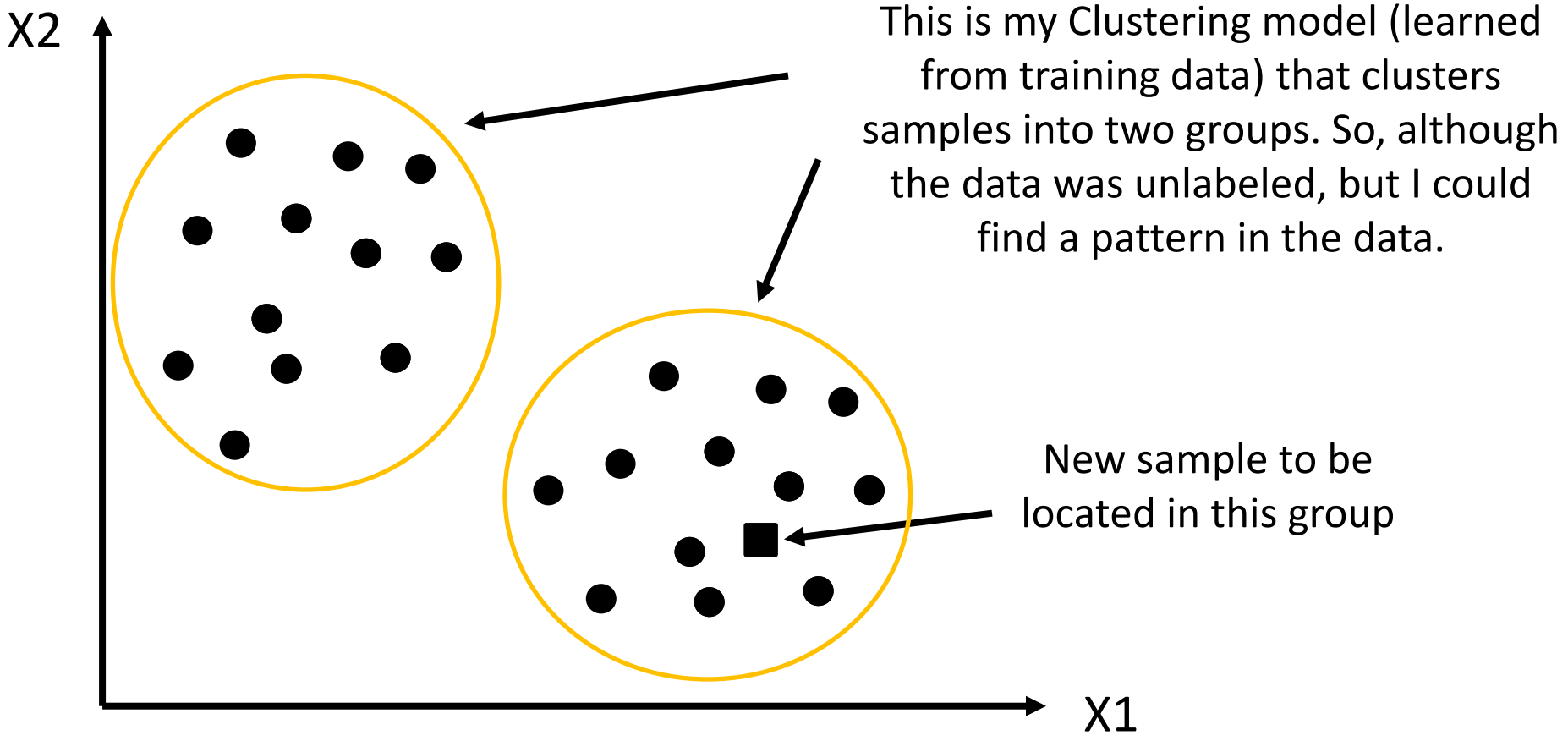


Unlabeled Training Dataset: the goal is to just discover a "hidden structure" in existing data, and make prediction in future

New data sample for prediction

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

Unsupervised Learning



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

Why/When to use Unsupervised Learning?

1. **When the Label is Unknown.**
   - e.g. image processing, fraud detection, anomaly detection, search for alien life!
2. **When the data is Unlabeled because we cannot afford labeling the data.**
   - e.g. big unlabeled data on web!
3. **Sometimes, we don't care about the label, we just want to categorize the data.**
   - e.g. clustering in marketing.
4. **Sometimes, applying an unsupervised algorithm prior to a supervised learning can improve the prediction results!**
   - we'll see an example in next slides.
1. **Sometimes, We want to manipulate the data w/o considering the label of that.**
   - e.g. dimensionality reduction.

Any hidden patterns or structure?

Figure: Ian H. Witten, etc., Data Mining: Practical Machine Learning Tools and Techniques, 2011.

Thank You!

Questions?