



# **Introduction to Data Science**

## **(Lecture 2)**

**Dr. Mohammad Pourhomayoun**

Associate Professor

Computer Science Department

California State University, Los Angeles



# Review: What is Data Science?



# What is Data Science?

- **Data Science** is an interdisciplinary field of research that aims to design and develop automated or semi-automated techniques to extract knowledge (information) from large-scale data and use it for future purposes such as prediction, decision making, or recommendation.
- It can be an integration of statistics, machine learning, big data processing, predictive analytics, and computing.



# New Sources of Data

- Social Networks: Facebook, Twitter, ...
- World Wide Web
- Online Activities: Amazon, ebay, ...
- Smart Phone Activities
- Electrical Health Records (EHR)
- Body and wearable sensors
- ...



# Only Some of the Applications!



Stock Market Prediction



Real State Prediction



Online Shopping  
and Advertisements



Recommendation Systems



Self-Driving Cars



Healthcare

# Ingredients



# Ingredients

- **Data:** Rapid growth of massive datasets
  - E.g. WWW, Social Networks, Online Activities, Smart Phone, Wearables, Sensor networks, Science, ...

Data



facebook



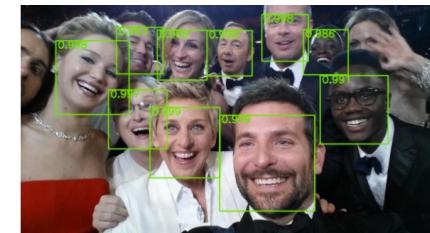
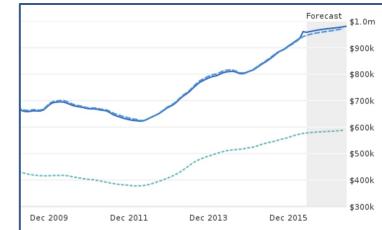
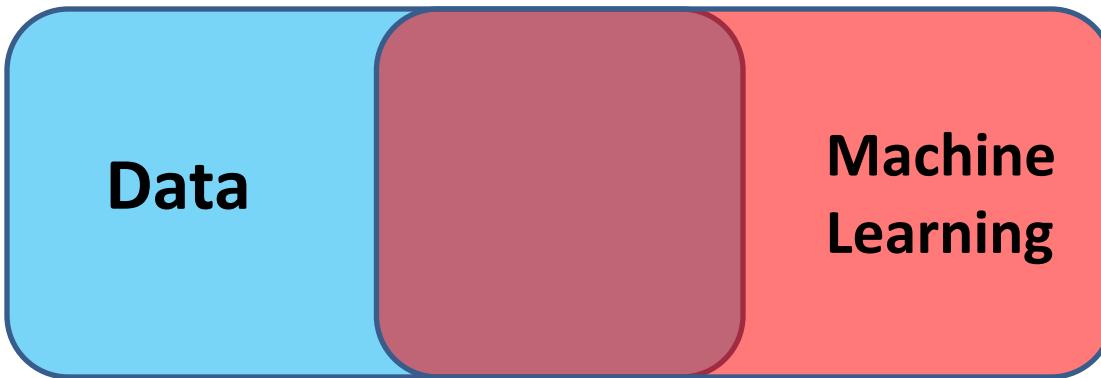
Google

amazon



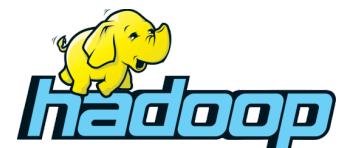
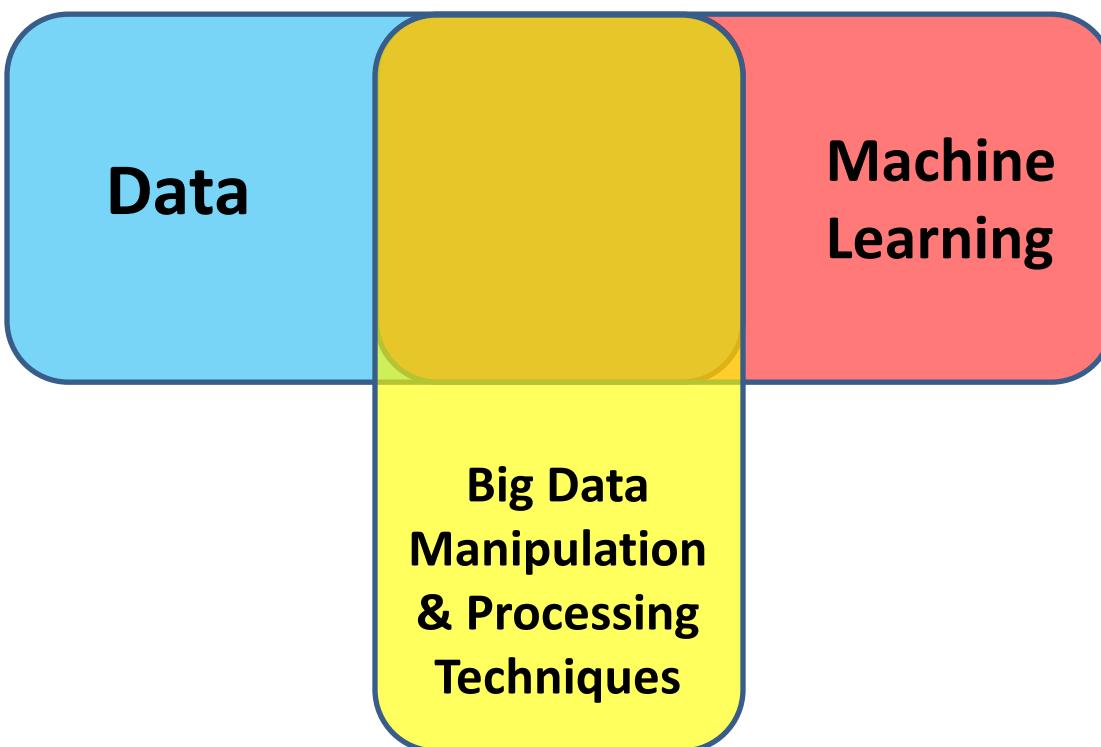
# Ingredients

- **Machine Learning:** It is applied Everywhere:
  - E.g., recommendation system, market prediction, speech recognition, Face detection, Fraud detection, Spam filtering, vehicle control, Medical diagnosis, ...



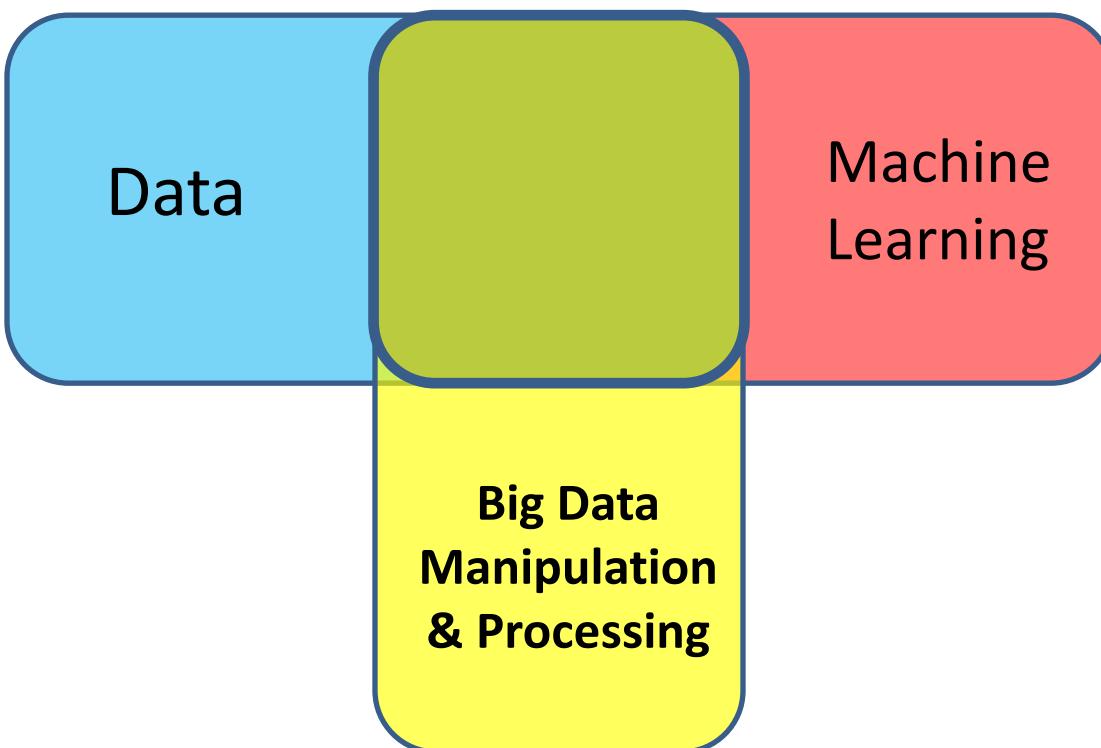
# Ingredients

- **Big Data Manipulation & Processing:**
  - Large-Scale Data Processing, Distributed Computing, Cloud Computing



# Ingredients

- Data, Machine Learning Algorithms, Big Data Manipulation Techniques



facebook



Google

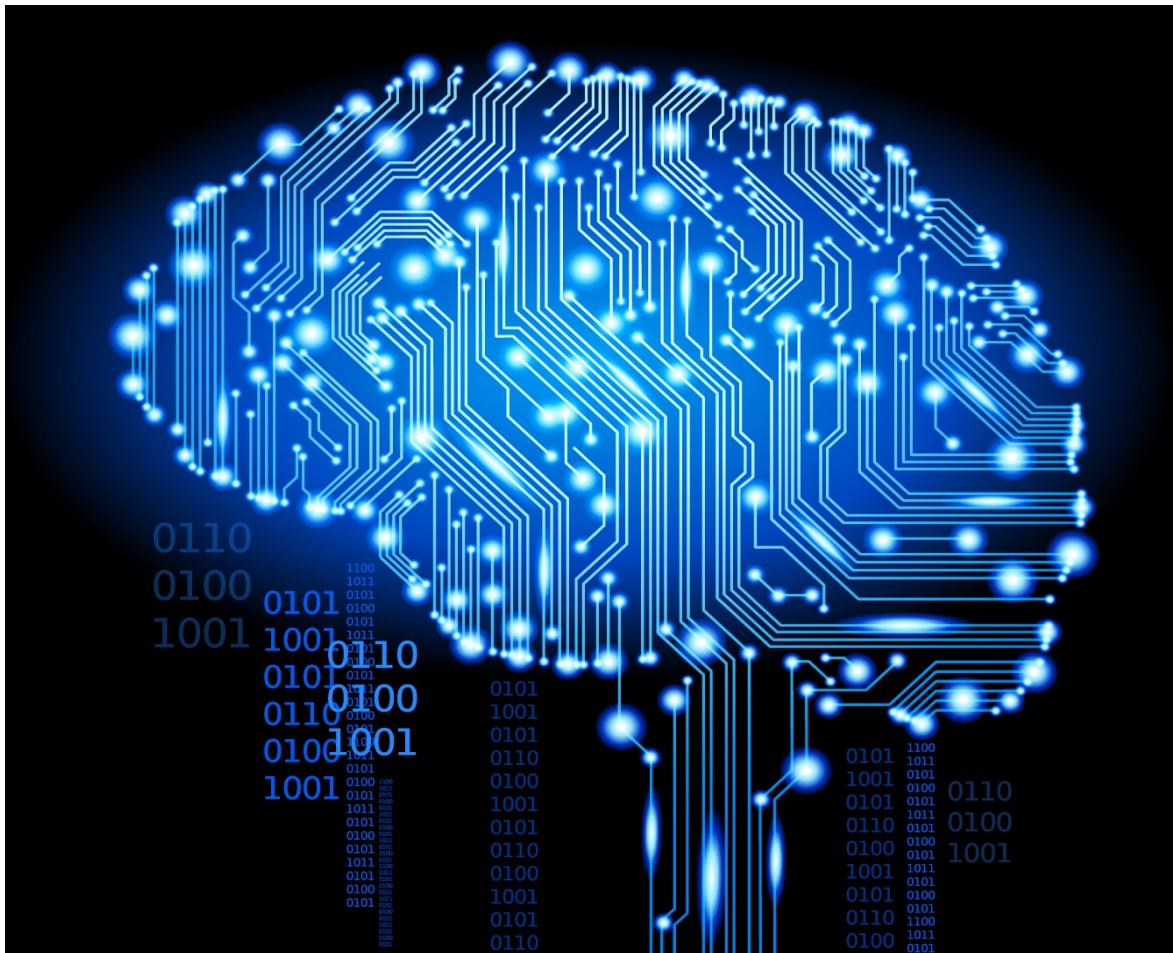
amazon





# Data Analytics and Machine Learning

# What is Machine Learning?



# What is Machine Learning?

- **A Definition:** Designing and constructing algorithms or methods that give computers the ability to learn from past data, without being explicitly programmed, and then make predictions on future data.
- **Another Definition:** A set of algorithms that can automatically detect and extract patterns in past data, and then use the extracted patterns to predict on future data, or to perform other kinds of decision making.



# What is Machine Learning?

- **A Definition:** Designing and constructing algorithms or methods that give computers the ability to learn from past data, without being explicitly programmed, and then make predictions on future data.
- **Another Definition:** A set of algorithms that can automatically detect and extract patterns in past data, and then use the extracted patterns to predict on future data, or to perform other kinds of decision making.

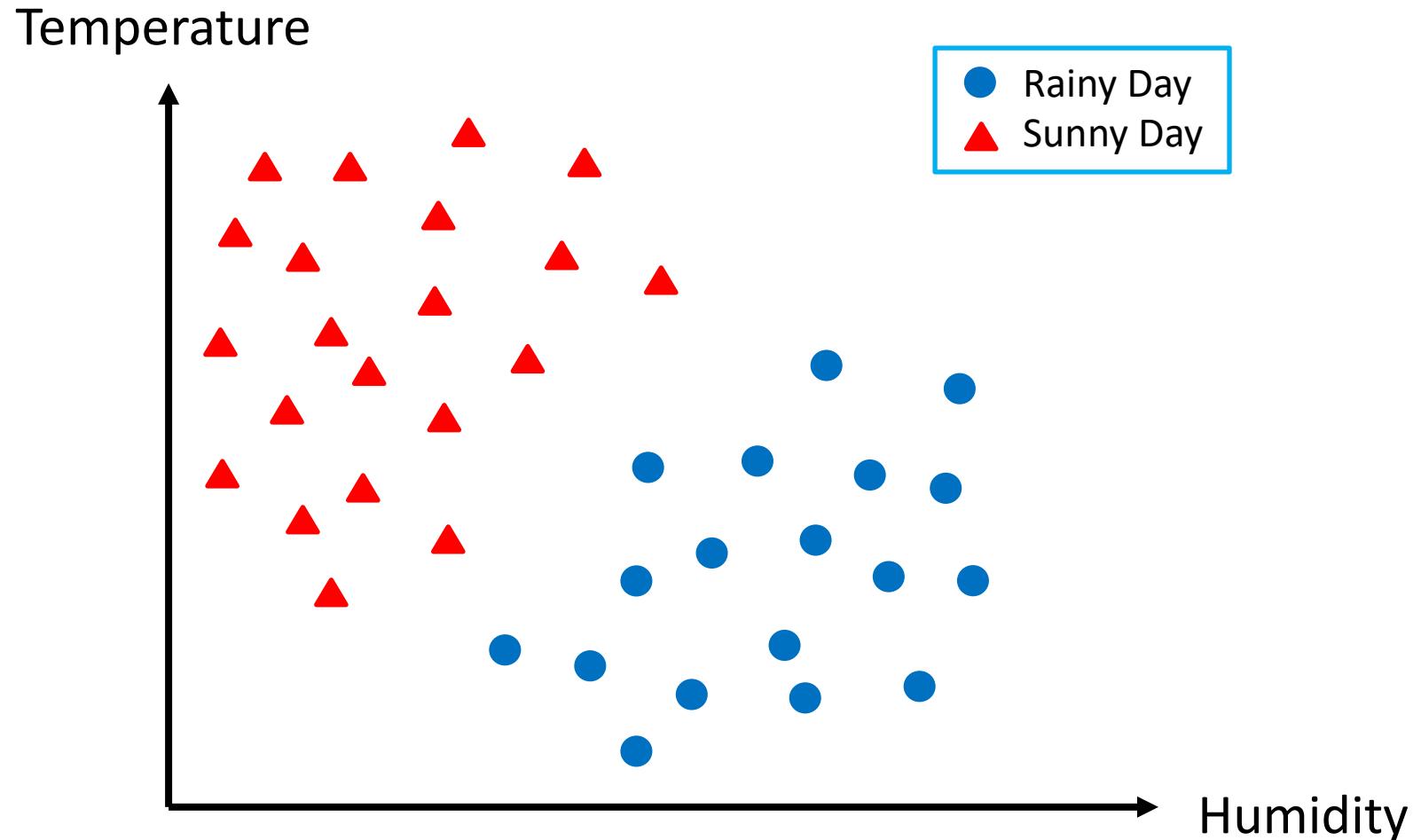


# Example: Weather Forecasting

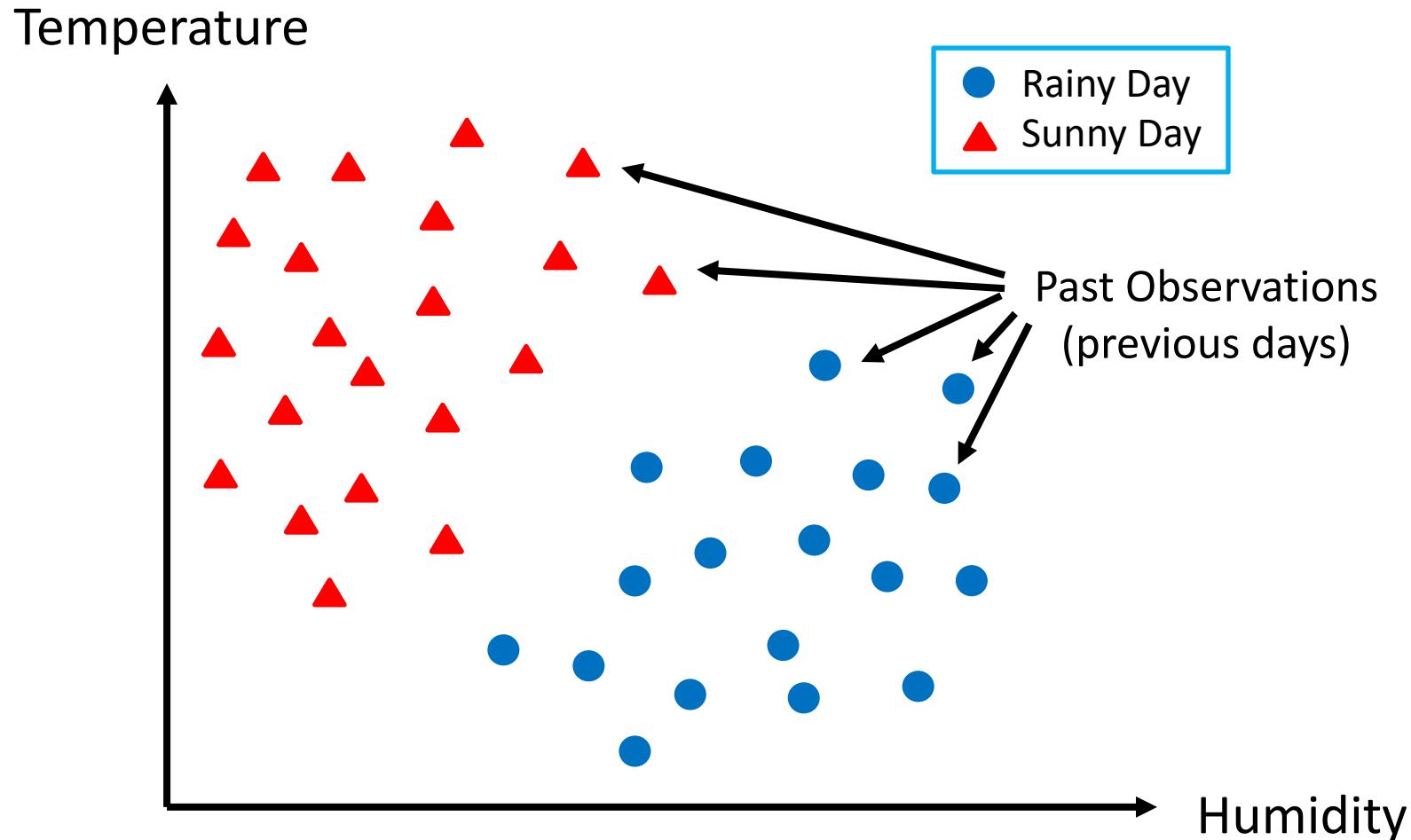
- Suppose that we have the **Temperature** and **Humidity** of the **past 30 days**.
- We also know whether those days were **Sunny** or **Rainy**.
  
- **Questions:** Now, If we know the Temperature and Humidity of tomorrow, can we predict tomorrow's outlook (predict whether tomorrow is rainy or sunny)?



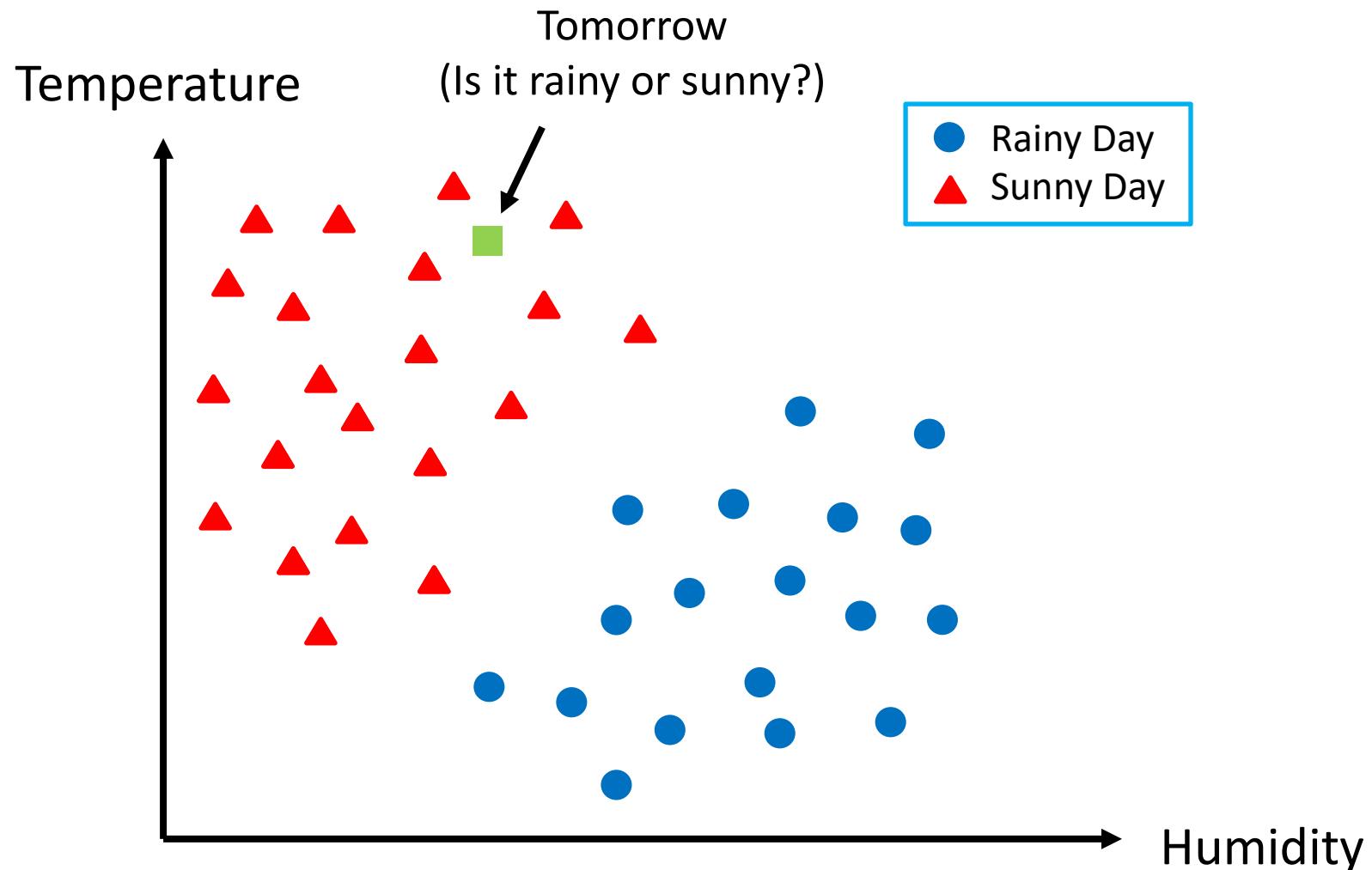
# Example: Weather Forecasting



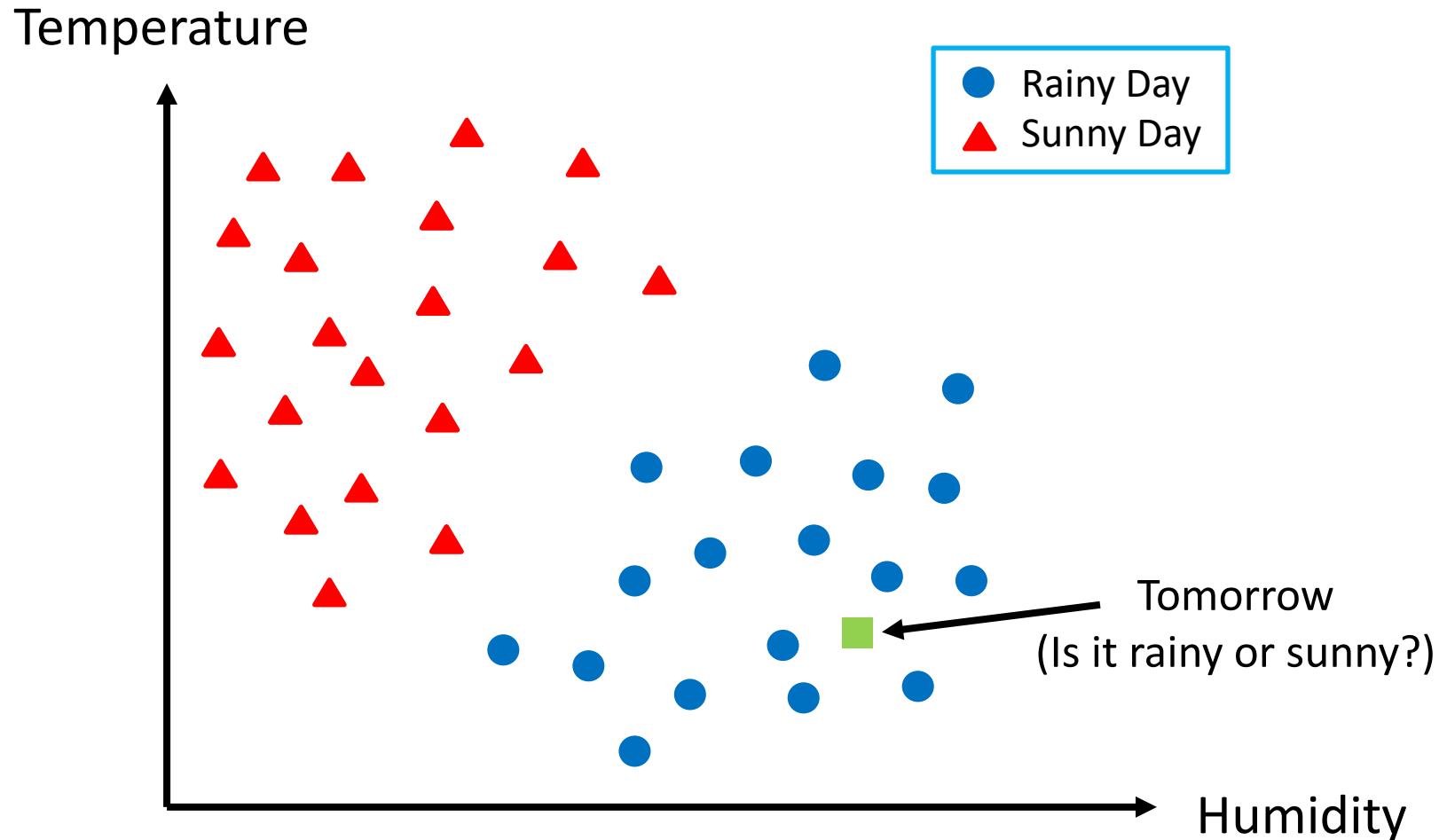
# Example: Weather Forecasting



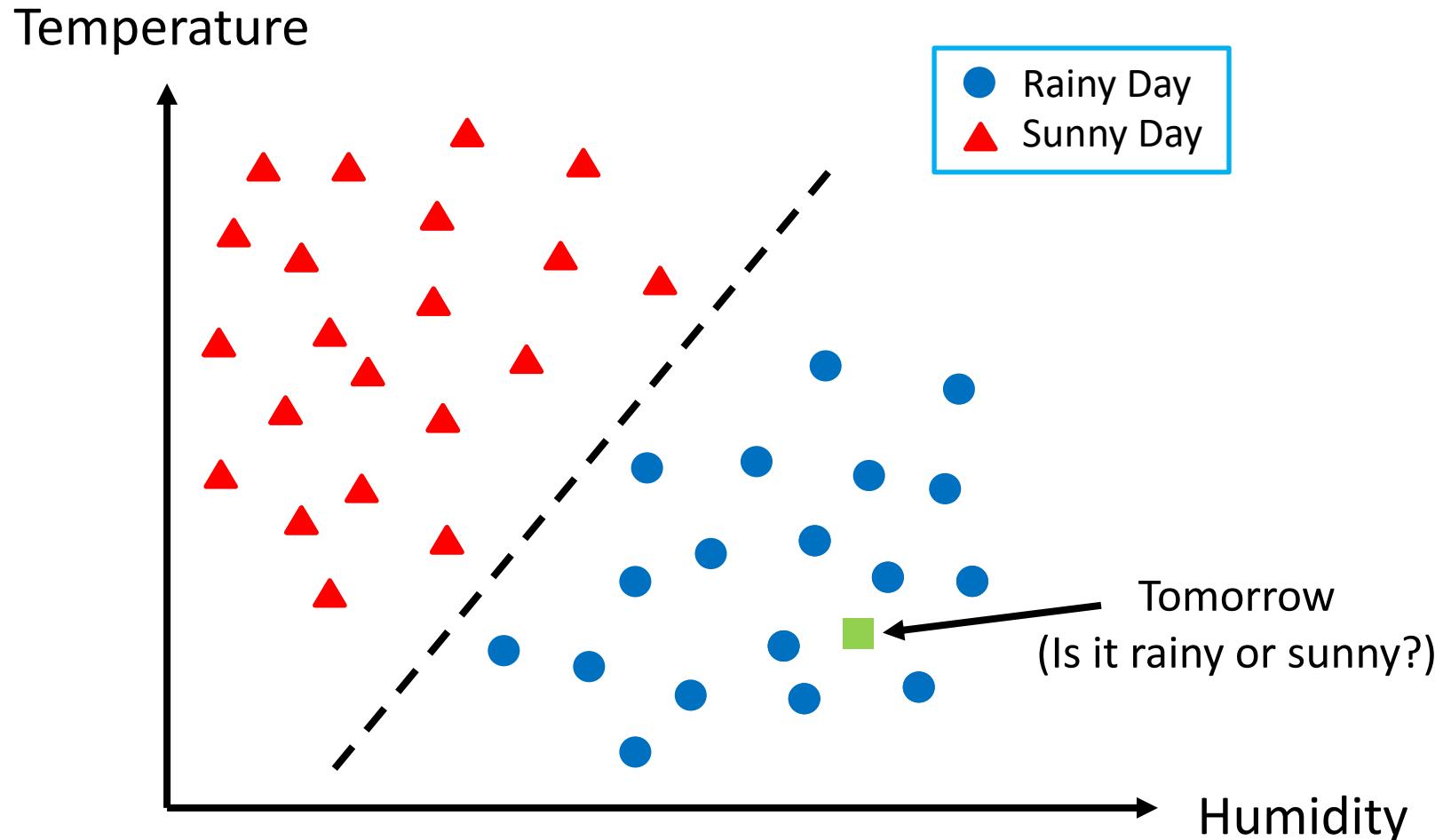
# Example: Weather Forecasting



# Example: Weather Forecasting

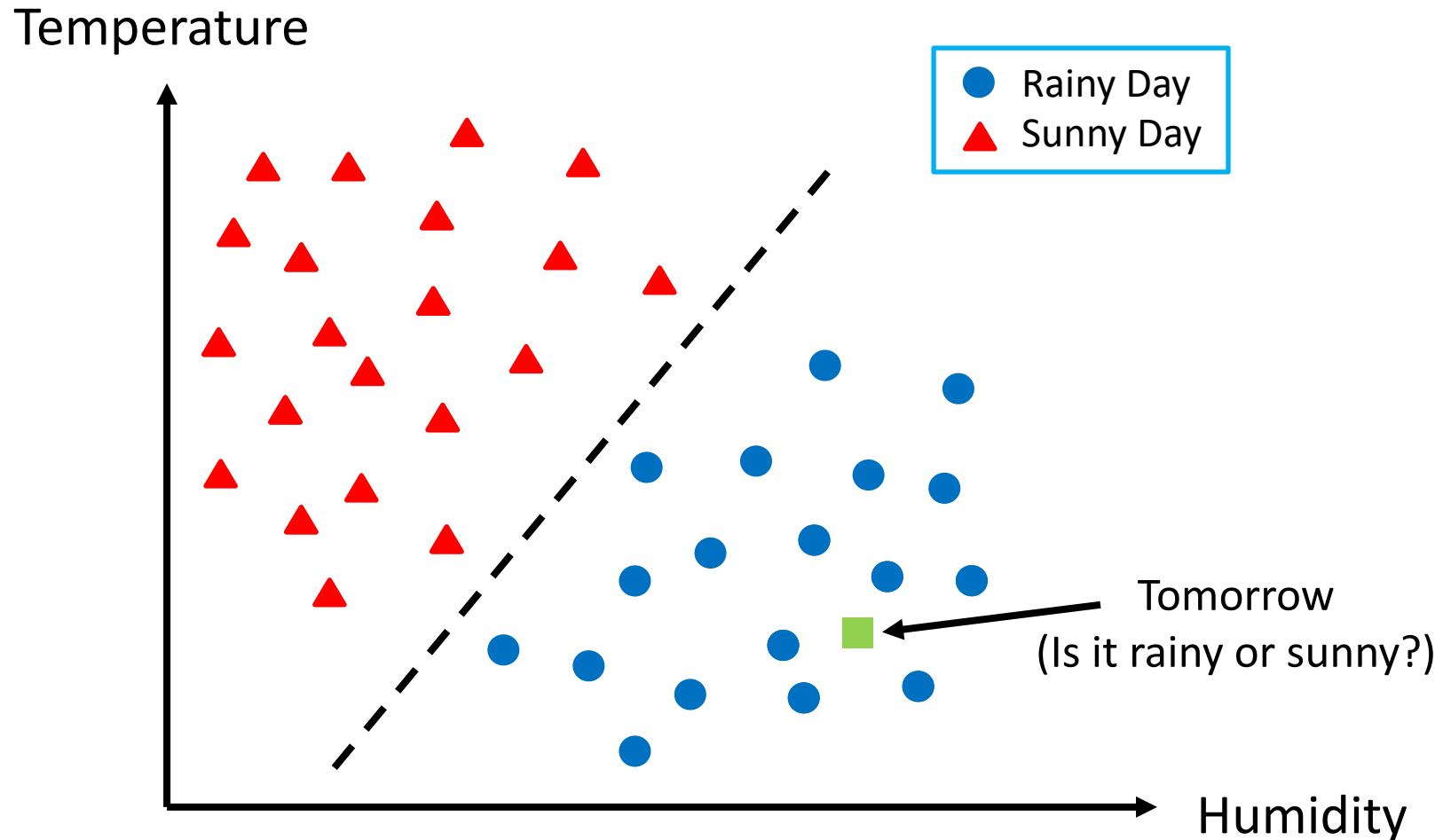


# Example: Weather Forecasting



# Example: Weather Forecasting

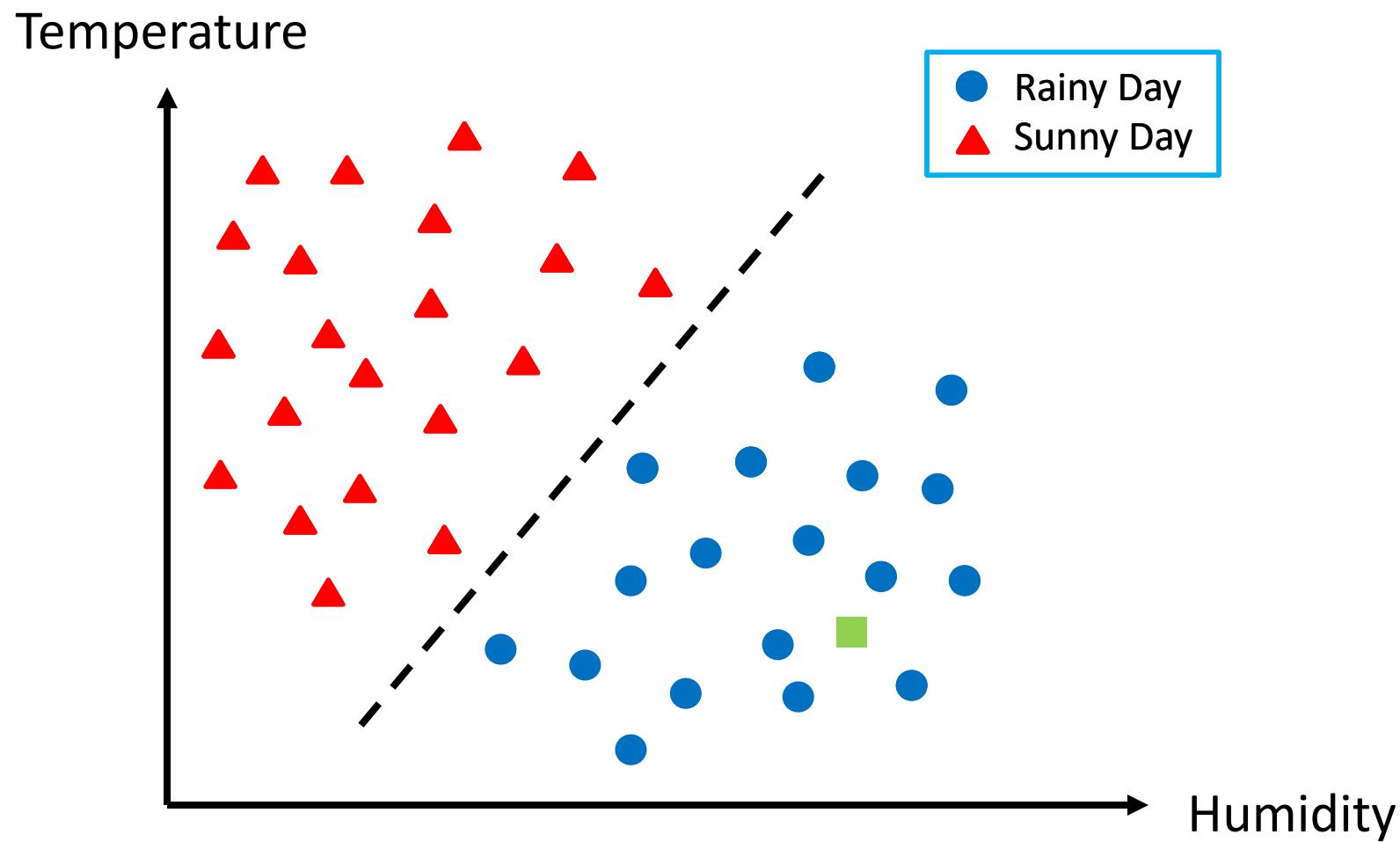
This is a super simple example, but did you get the idea?

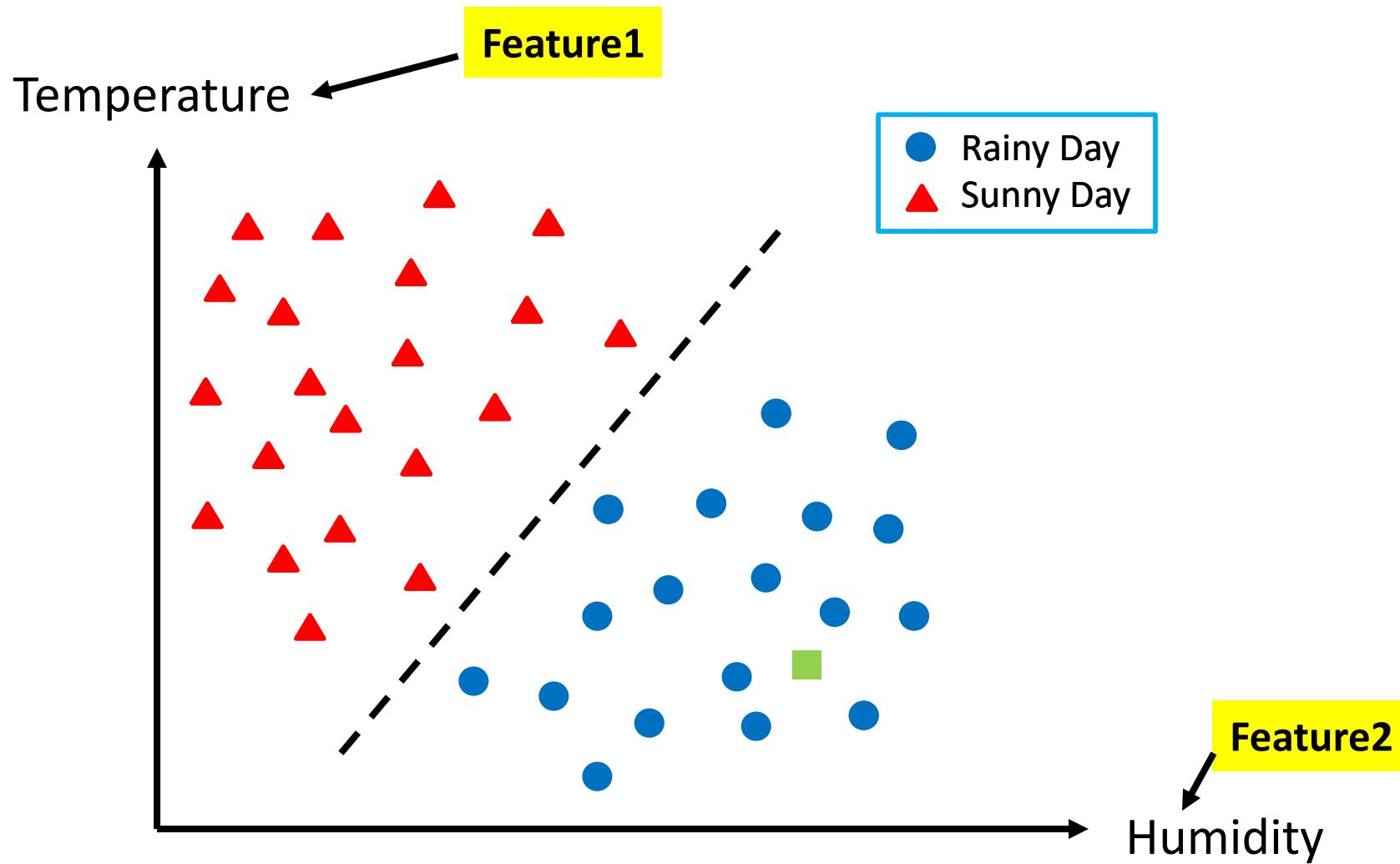


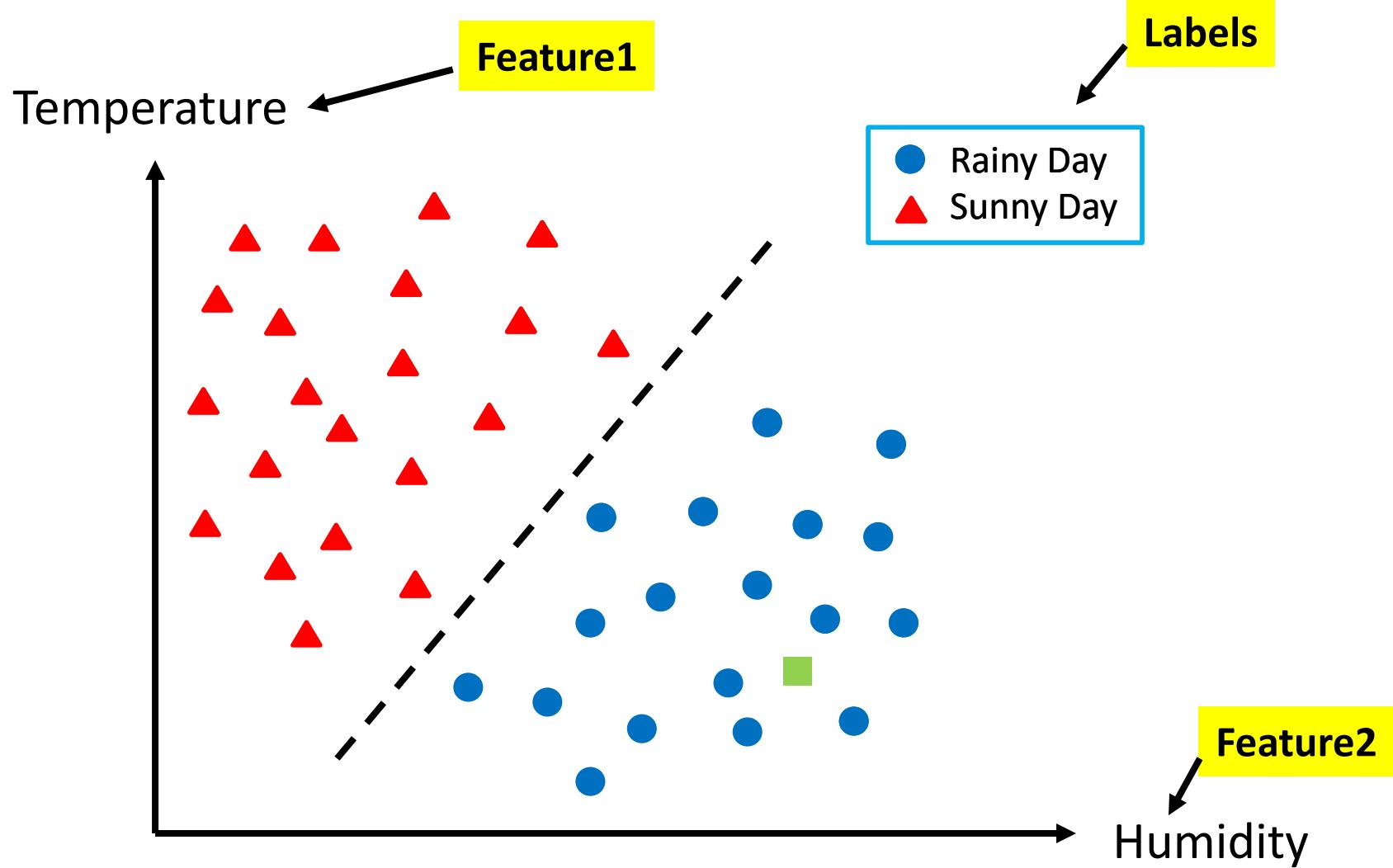
# Terminology

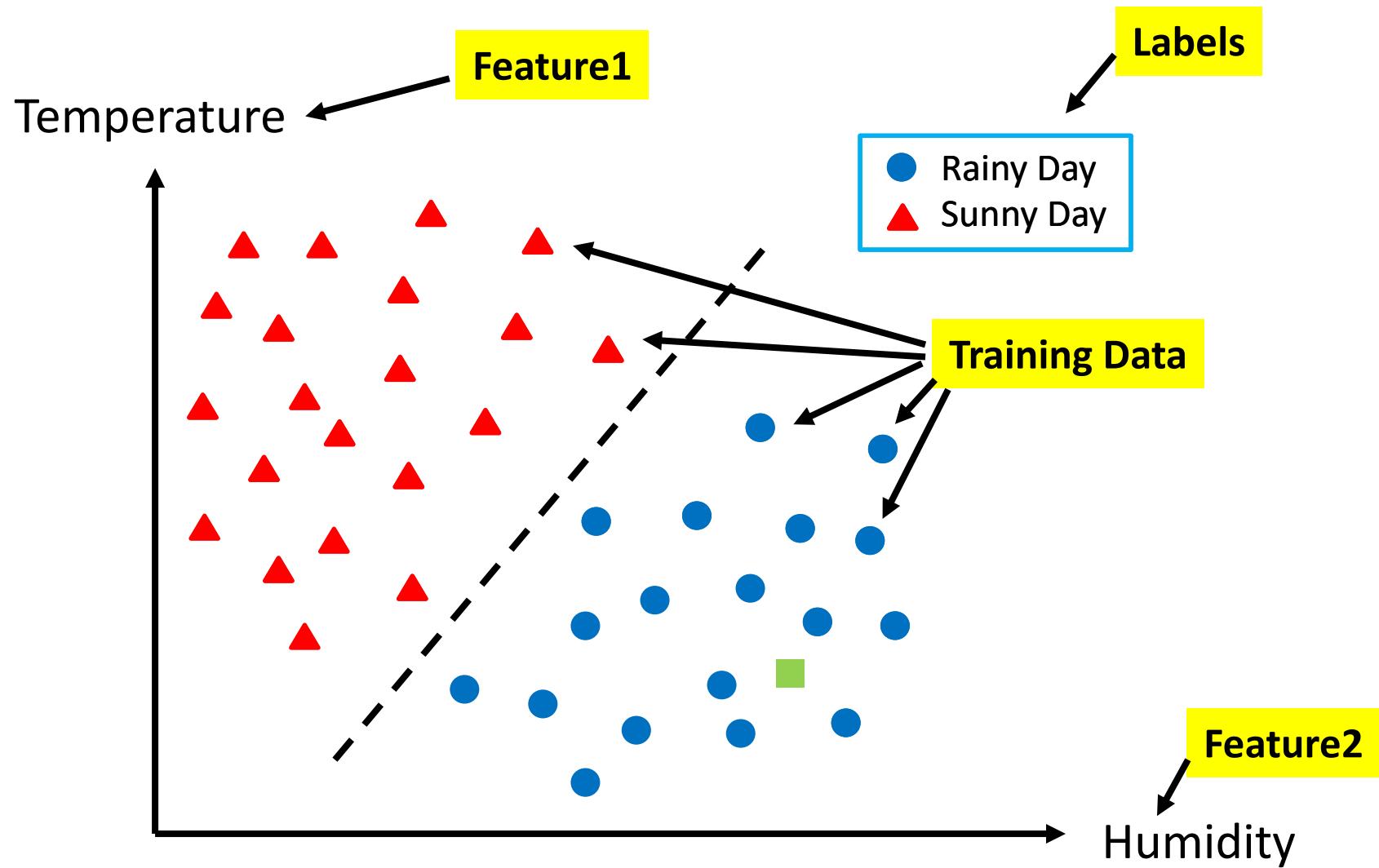
- **Observations:** Data Samples (Data Examples).
- **Features (inputs):** Attributes that represent an observation, e.g., temperature, humidity
- **Labels (outputs):** Values assigned to observations (also called class, target), e.g., rainy/sunny
- **Training Data:** Past observations given to the Machine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

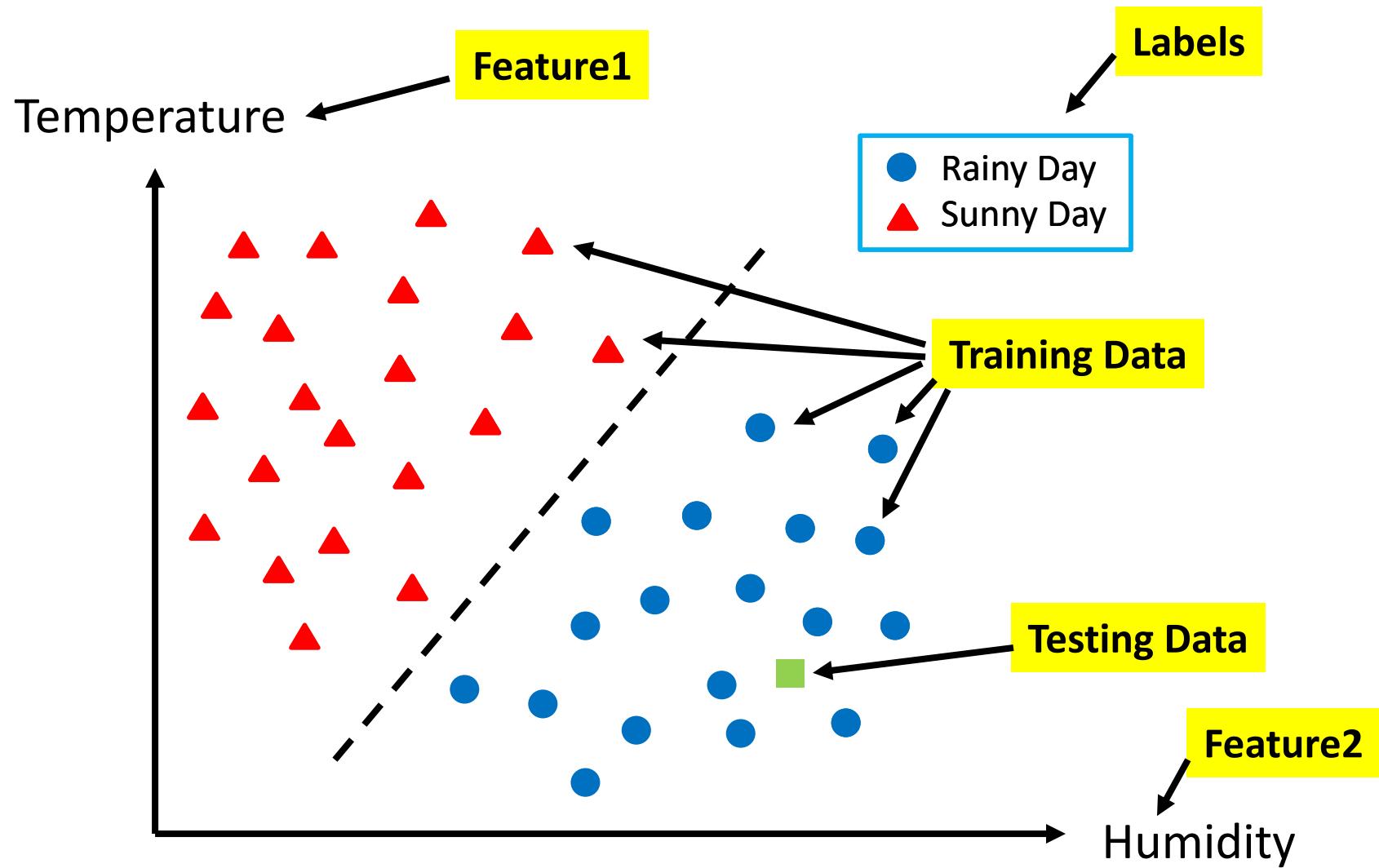


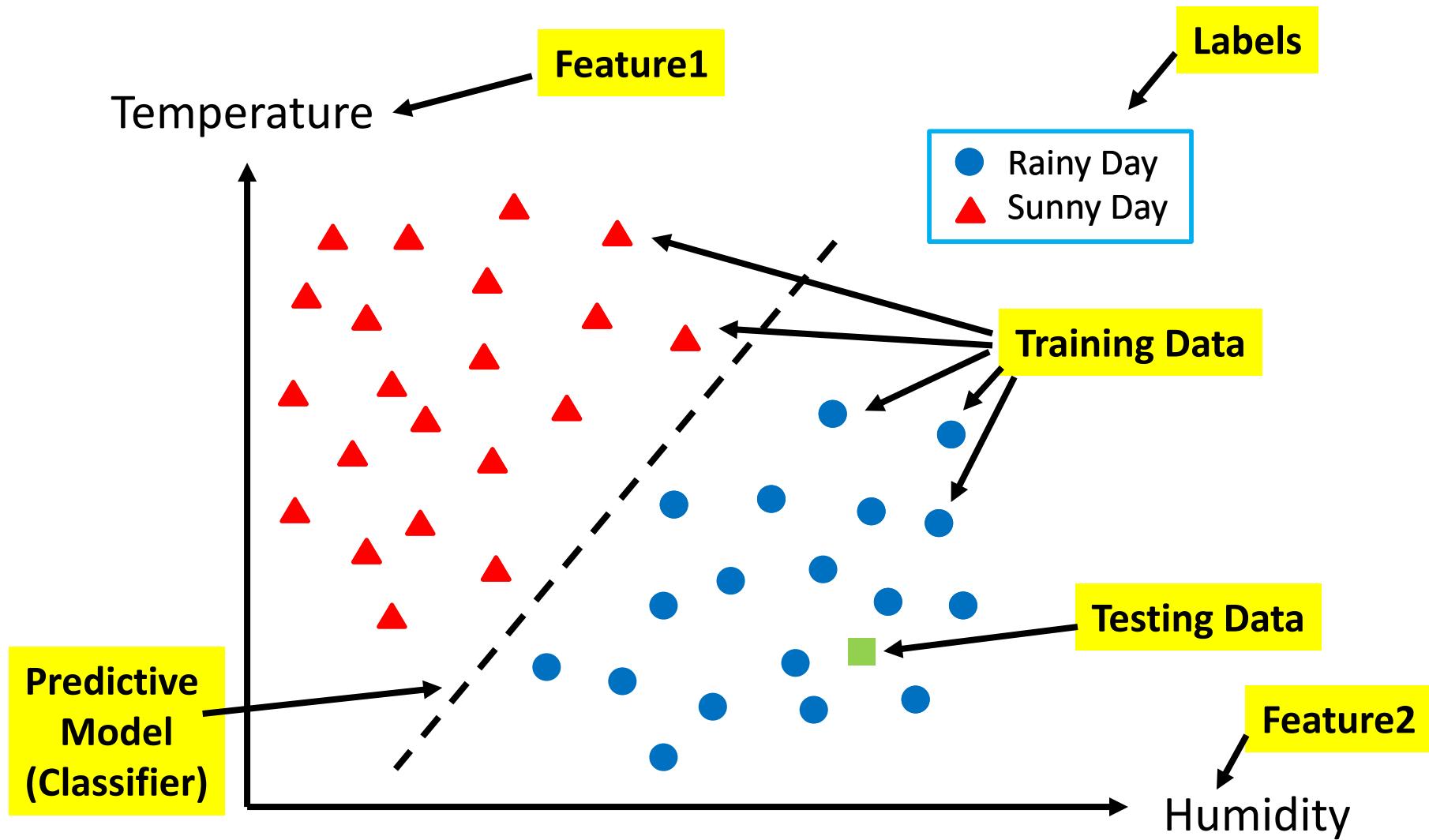










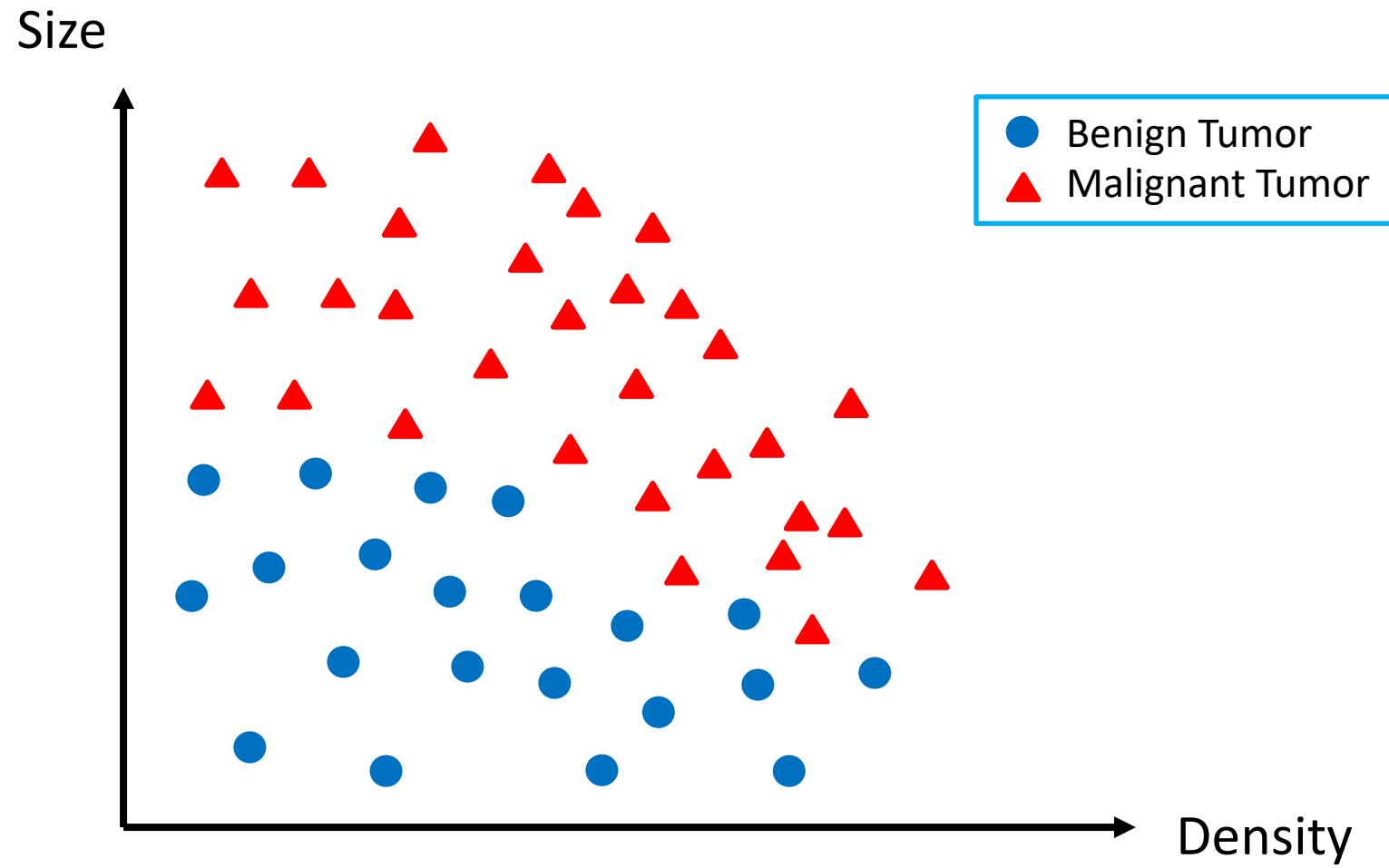


# Another Example: Predicting/Diagnosing Cancer

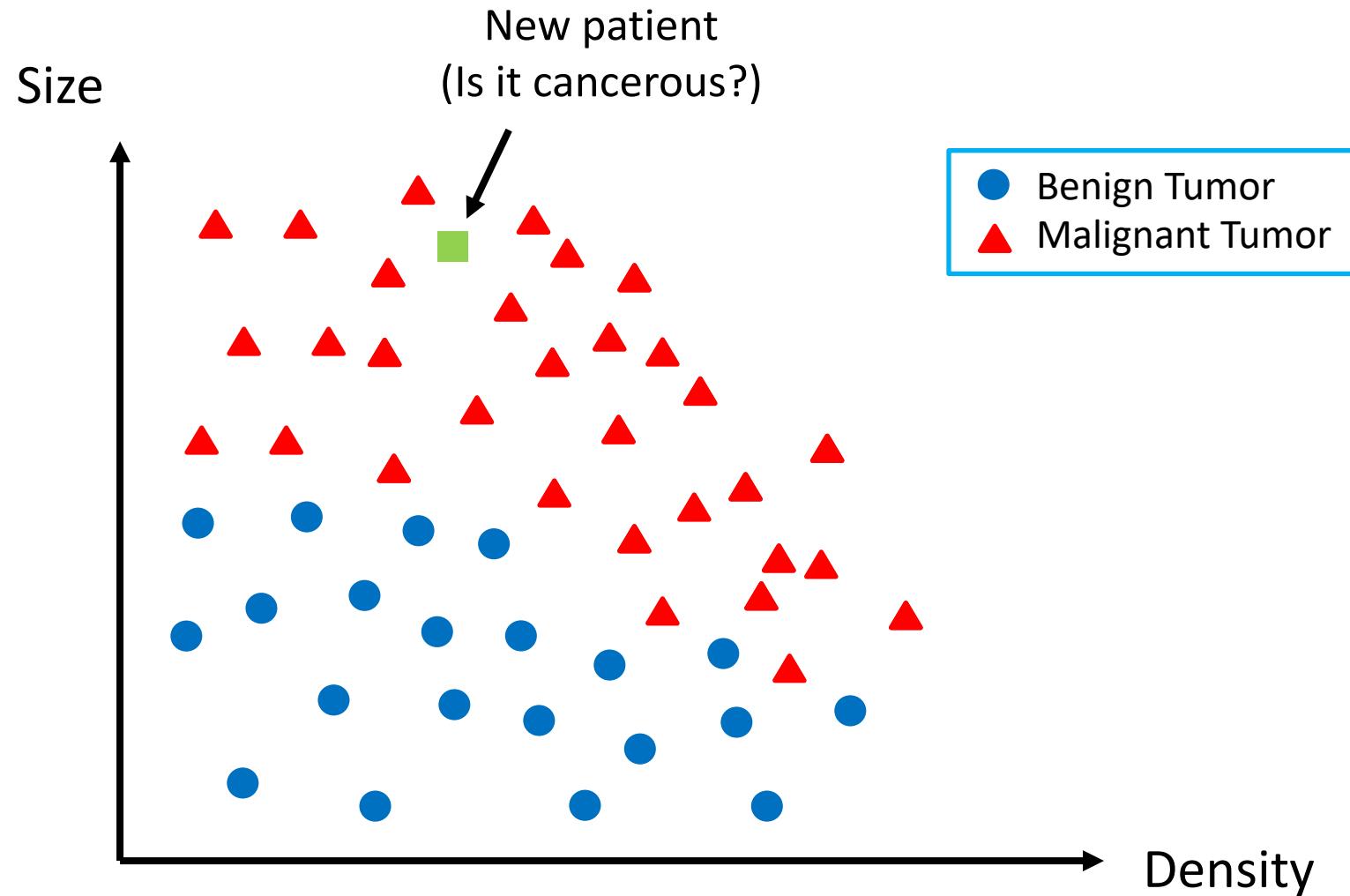
- Data: Suppose that we have the **Size** and **Density** of 100 tumors observed in 100 patients in the **past**.
- We also know whether those tumors were **Malignant** or **Benign**.
- **Questions:** Now, If we know the **Size** and **Softness** of a tumor in a **new** patient, can we **predict** if it is **Malignant** or **Benign**?



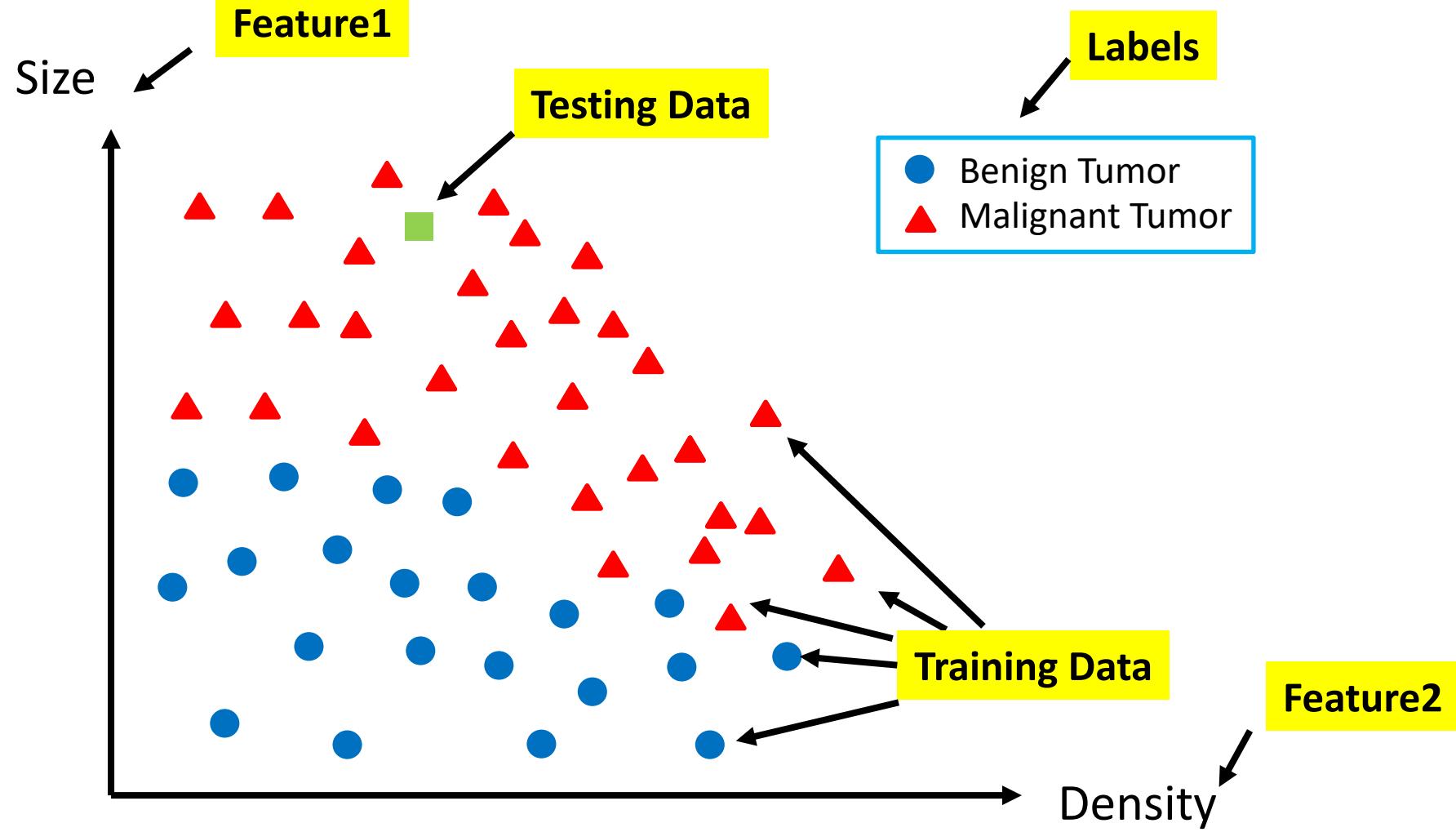
# Example: Predicting Cancer



# Example: Predicting Cancer



# Example: Predicting Cancer



# Example

- Suppose that we have medical data of **100 patients** diagnosed with a tumor in their body. We know the **age** and **gender** of patients. For each patient, we also know the **size** and **density** of tumors. After Biopsy examination, we know which one of these patients has **malignant** tumor and which one has **benign** tumors.
- Now, for **5 new patients** with tumor, having the **size** and **density** of tumors, and **age** and **gender** of patients, we would like to predict the type of tumor **without biopsy procedure**.
- Can you define **training data**, **testing data**, **features**, and **labels** for this problem?



# More Terminology

- **Training Stage (Modeling):** Building a predictive model based on the training dataset (past data).
  - 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)





# Machine Learning Settings

# Common Learning Settings

- Supervised learning
- Unsupervised learning
- Semi-supervised learning
- Reinforcement learning
- Transfer learning
- Active learning



- **Supervised learning:** Learning from labeled observations.
  - In training stage, the algorithm is presented with features and their known labels, and the goal is to train a model that maps future inputs to new labels.
  - E.g. Neural Networks (ANN), K-Nearest Neighbors classifier (KNN), Decision Tree classifier, Linear Regression, Logistic Regression, Polynomial Regression, Random Forest, ...
- **Unsupervised learning:** Learning from unlabeled observations.
  - The algorithm is presented **Only** with features! The goal is to Discover hidden patterns and structure from features alone. It is like a Data Exploration to find hidden patterns.
- **Semi-supervised learning:** Labels are provided only for a part of the training data. The remaining data is unlabeled.
- **Reinforcement learning:** Learning from an *agent* taking *actions* in an *environment* so as to maximize a long-term *reward*. E.g. Game Theory, Control Theory.
- **Transfer learning:** Learning from a dataset while solving a problem, and then applying the **extracted knowledge** to a different but related dataset/problem.
- **Active learning:** Similar to Semi-Supervised Learning, but the algorithm is able to interactively query the user or some other information source to obtain the labels as needed.



# Common Learning Settings

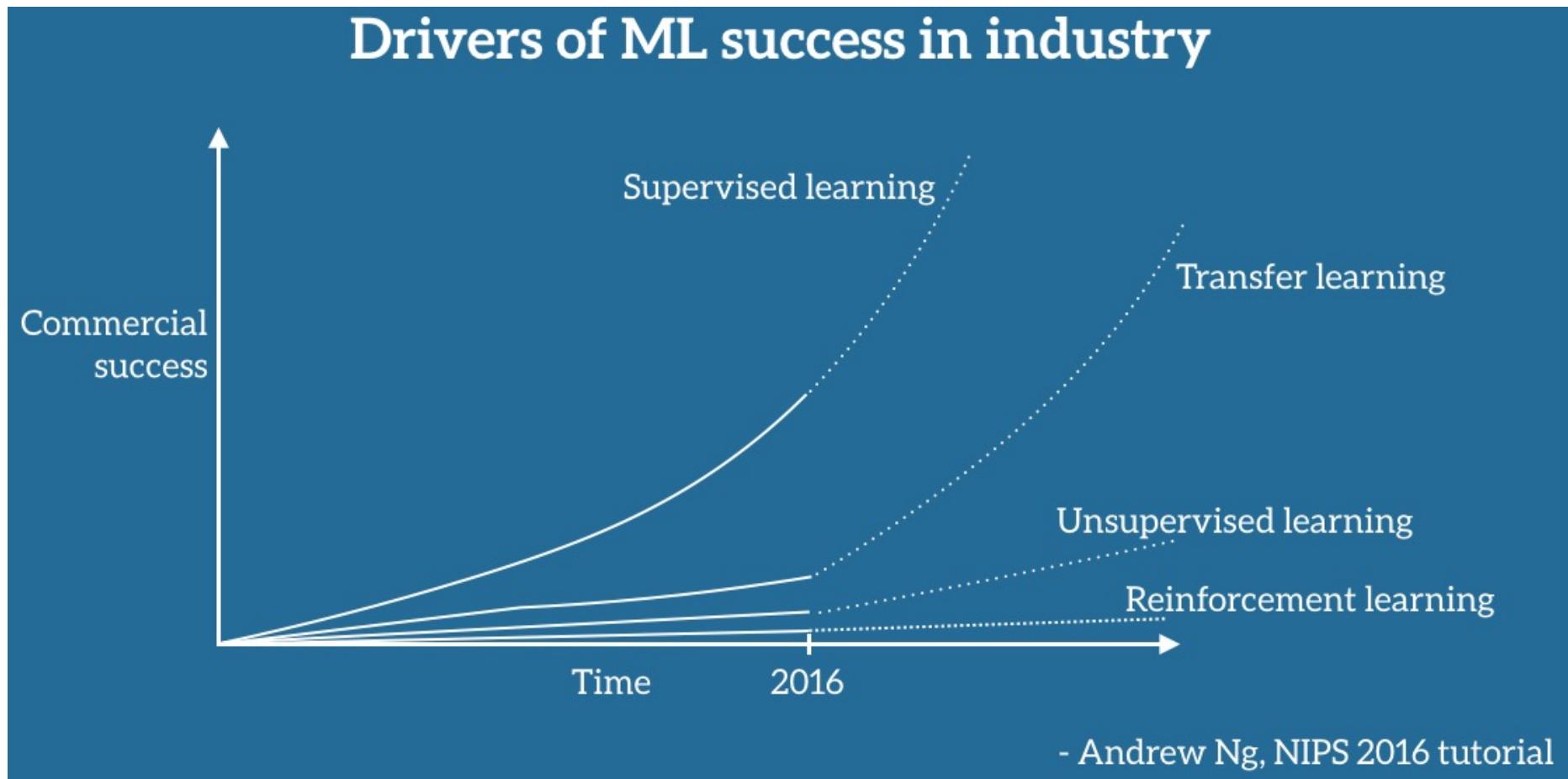
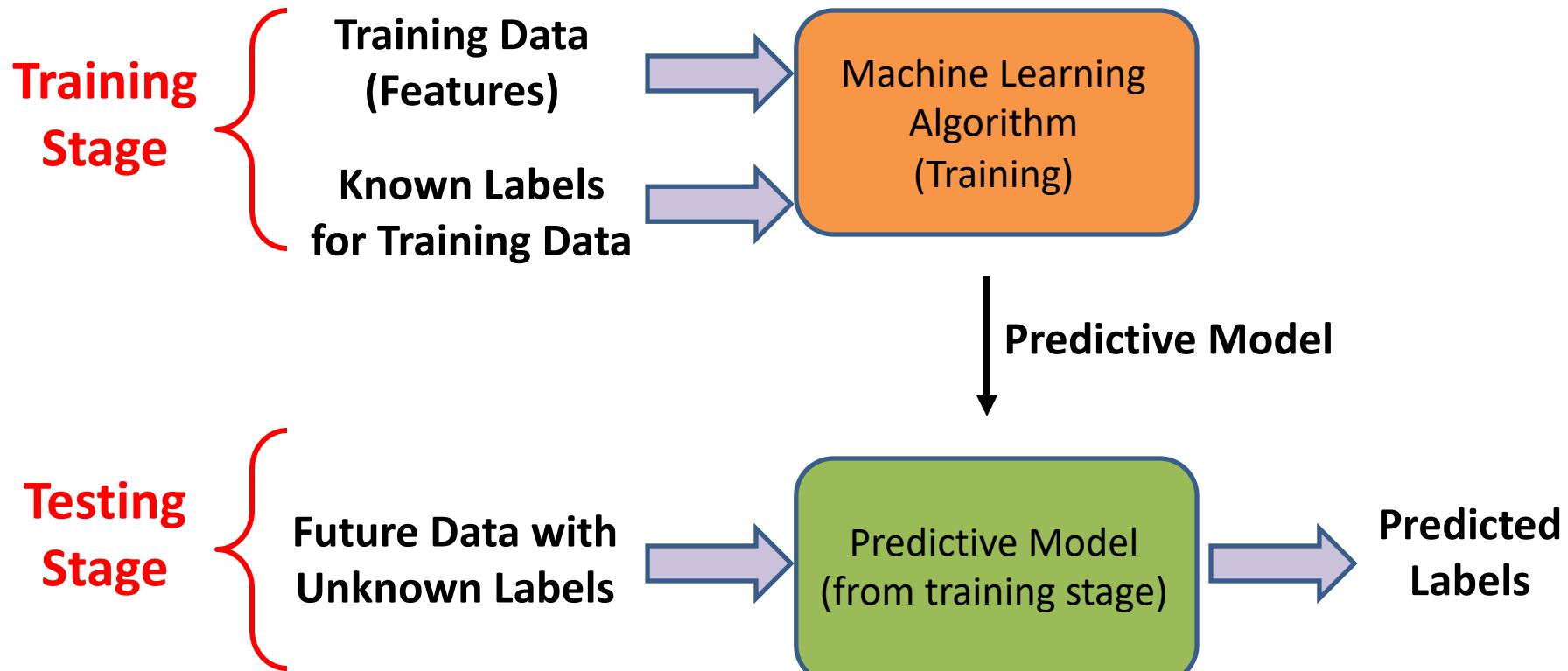


Figure Reference: Andrew Ng



# Supervised Learning: Learning from labeled Data

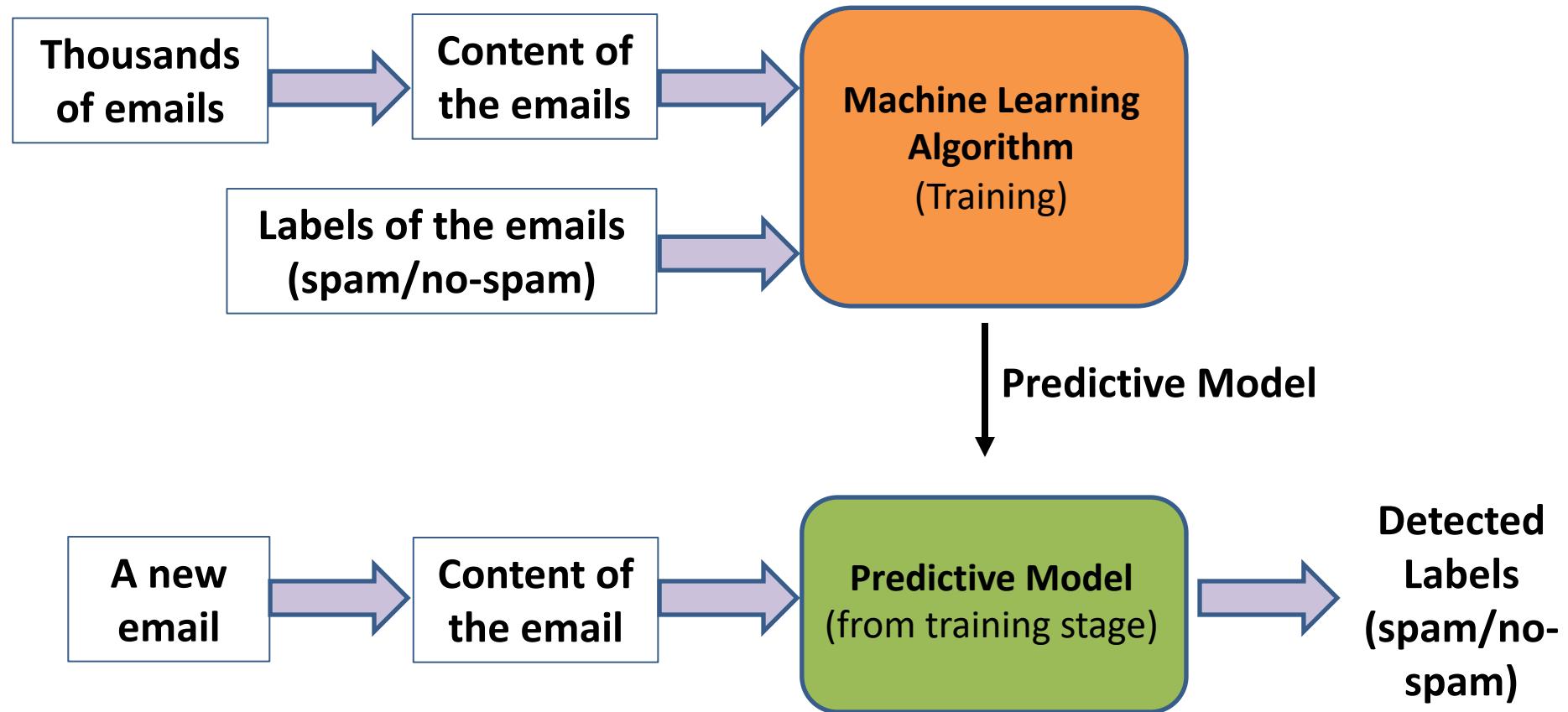


# Example for Supervised Learning: Spam Detection

- The algorithm will learn from previous data (previous emails) that what combination of words indicate suspicious emails.



# Example for Supervised Learning: Spam Detection

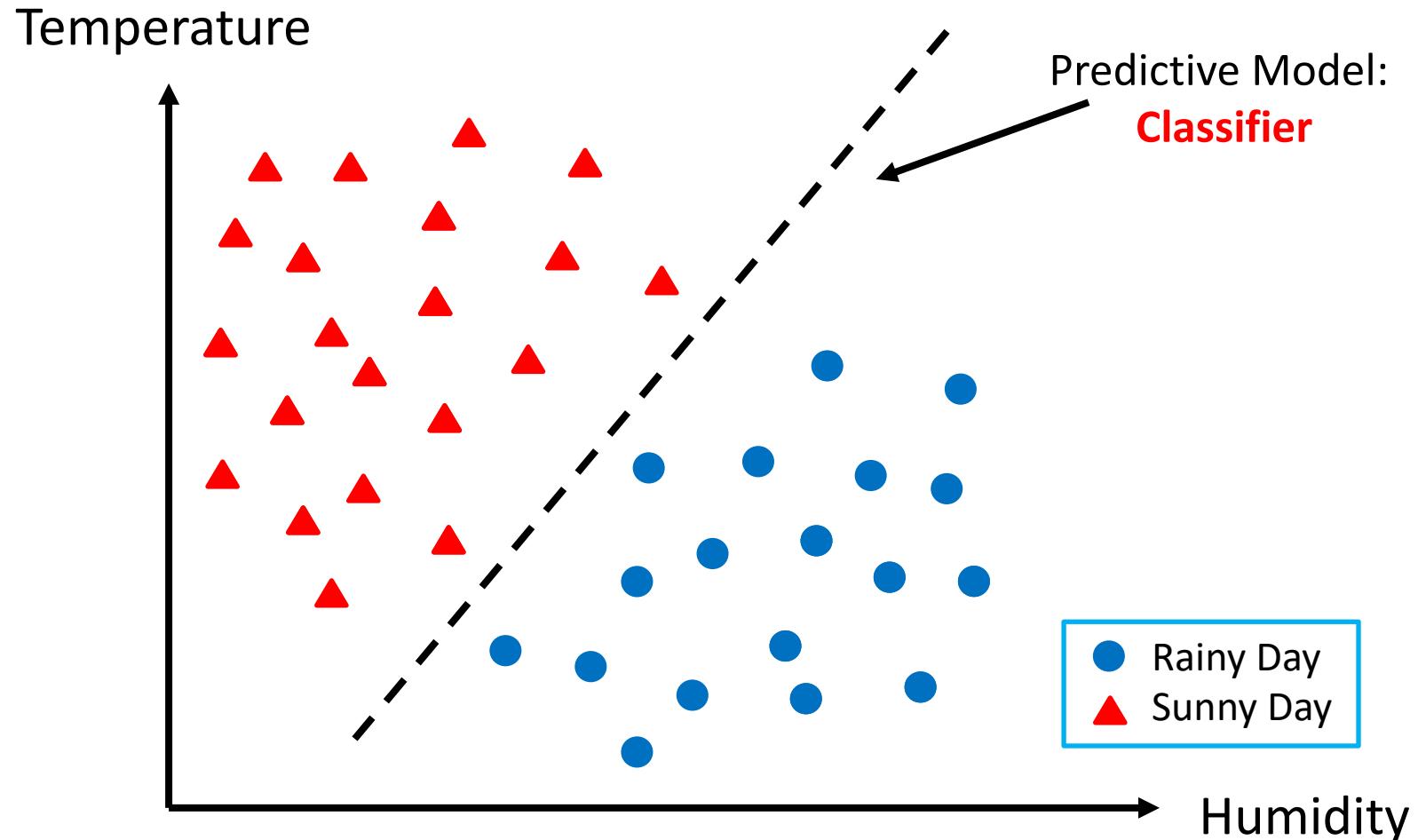


# Two Important Approaches of Supervised Learning

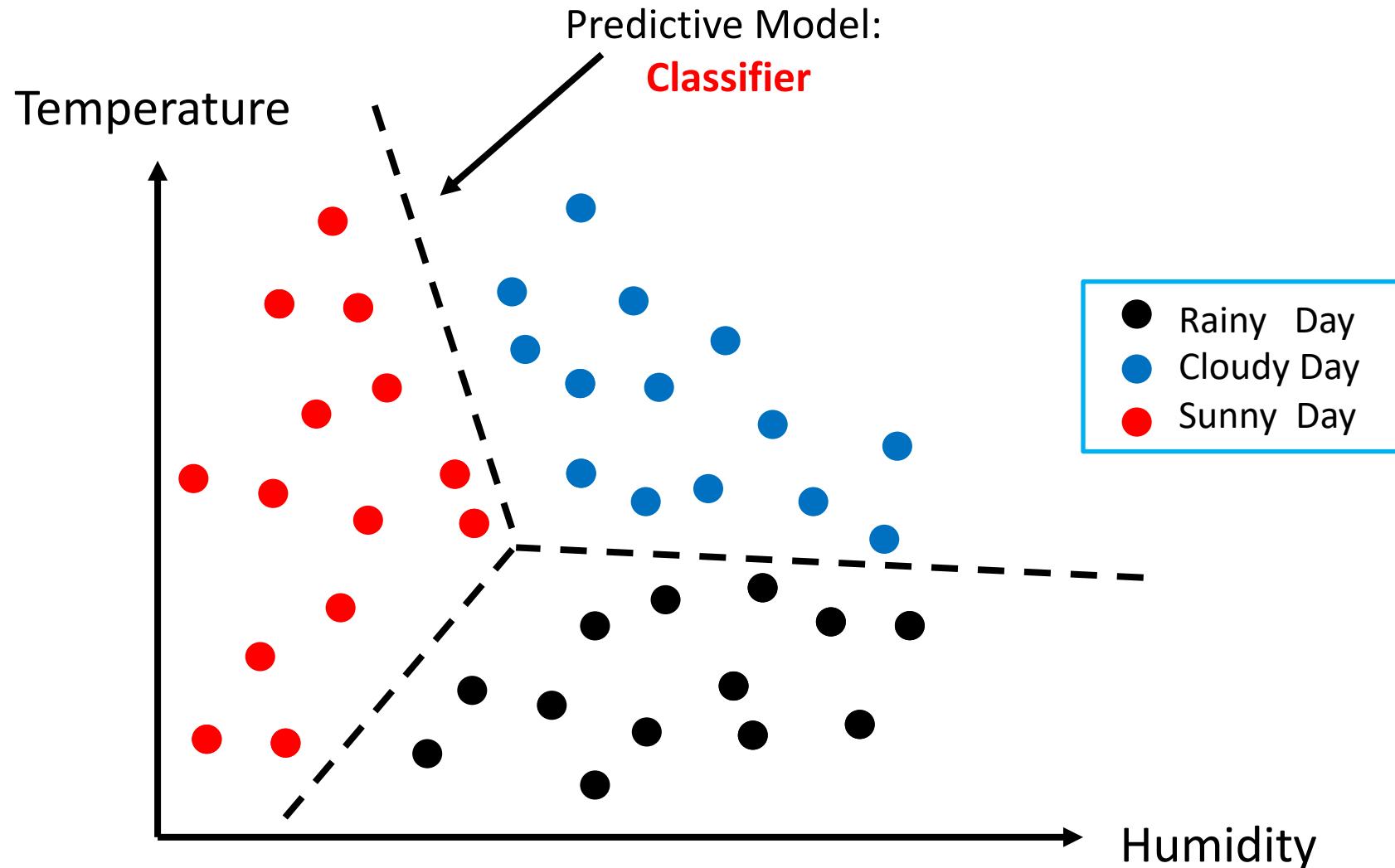
- **Classification:** Predict a discrete valued output for each observation.
  - Labels are **discrete (categorical)**
  - Labels can be binary (e.g., rainy/sunny, spam/non-spam,) or non-binary (e.g., rainy/sunny/cloudy, object recognition (100classes))
- **Regression:** Predict a continuous valued output for each observation.
  - Labels are **continuous (numeric)**, e.g., stock price, housing price
  - Can define ‘closeness’ when comparing prediction with true values



# Classification Example: Binary Label



# Classification Example: Multiple Label

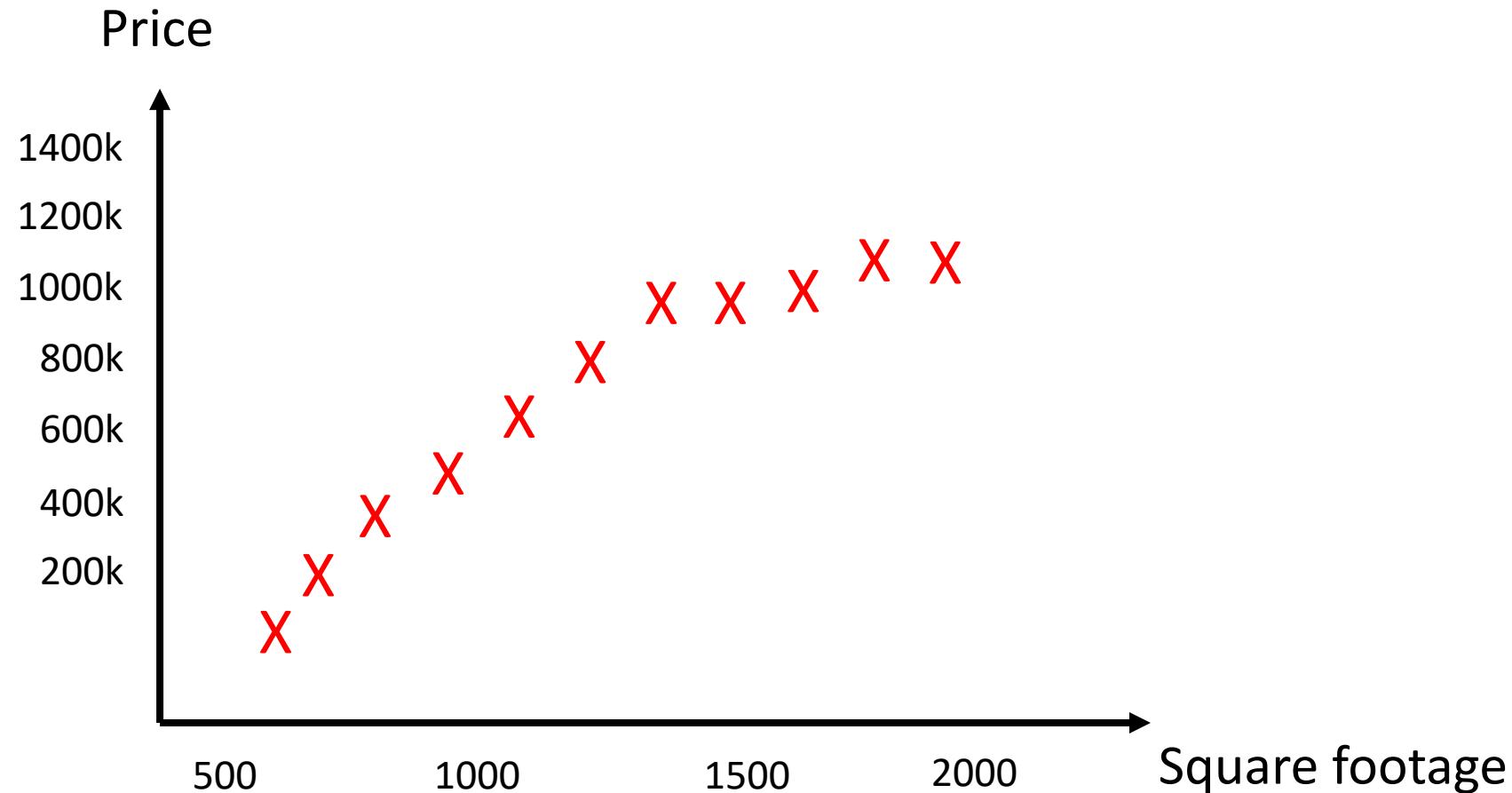


# Classification Example: Multiple Label

- **Recognizing flowers**
  - Types of Iris (Labels): setosa, versicolor, virginica
  - Features: sepal length, sepal width, petal length, petal width



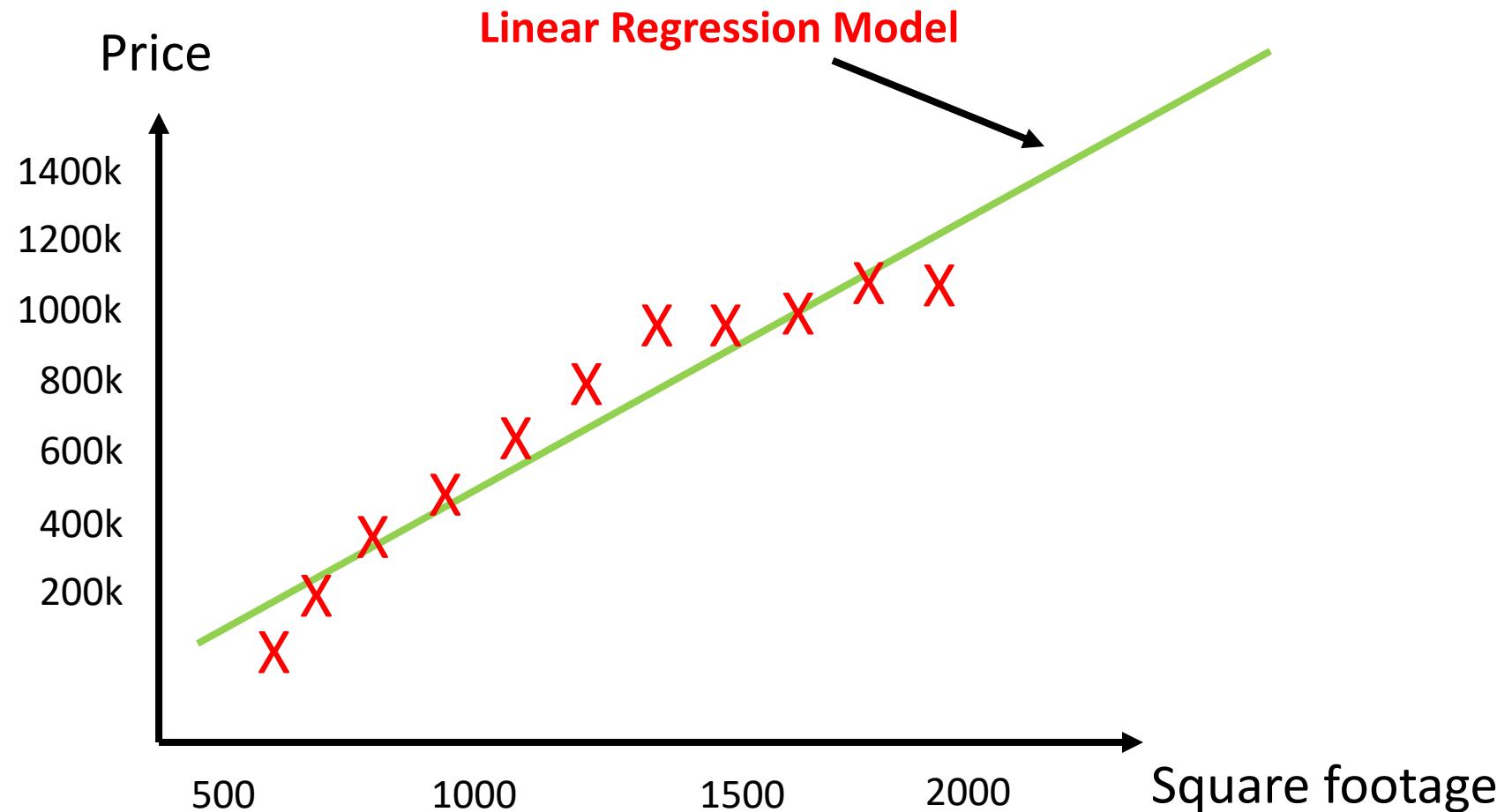
# Regression Example: Housing Price



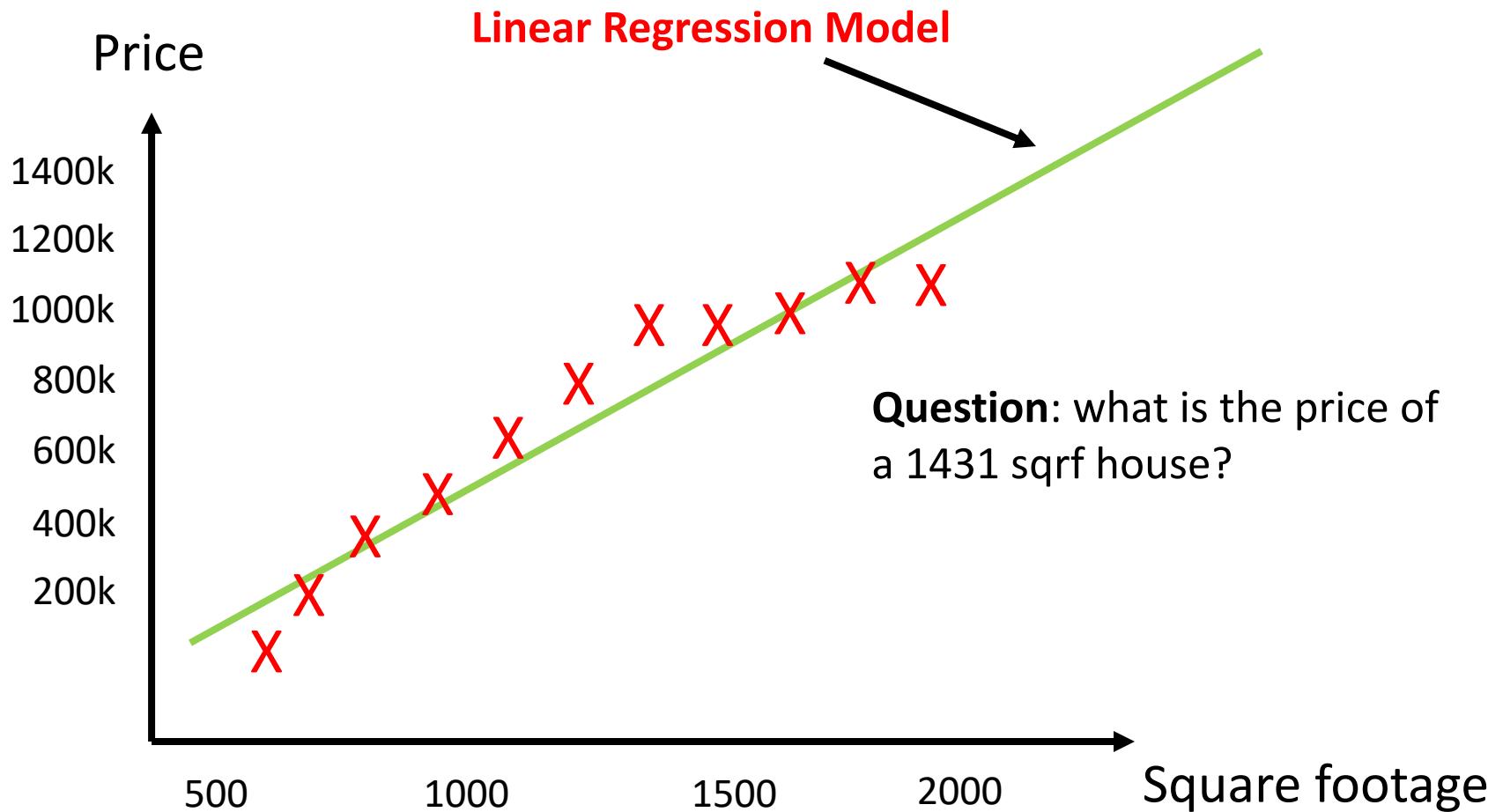
# Regression Example: Housing Price



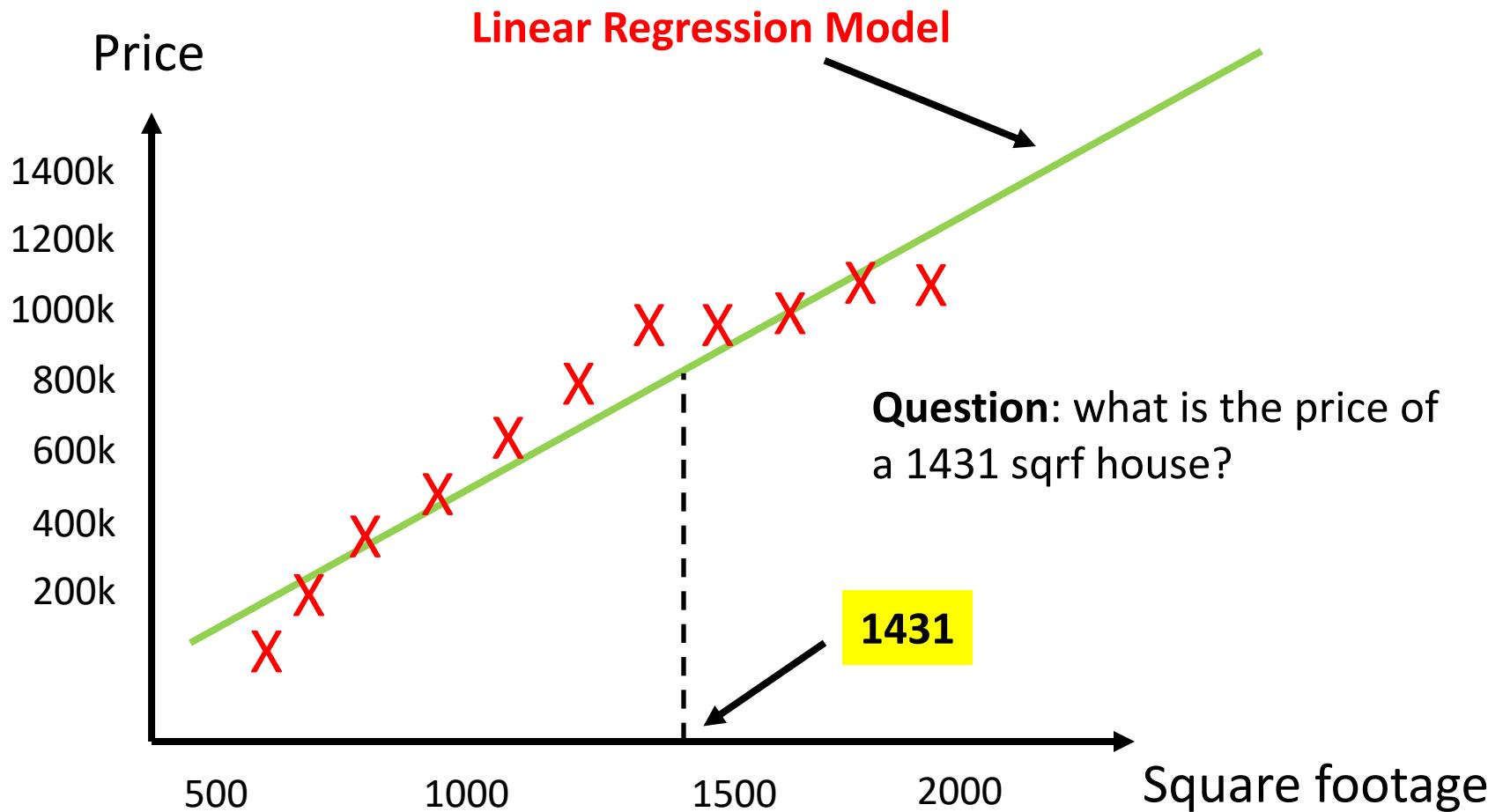
# Regression Example: Housing Price



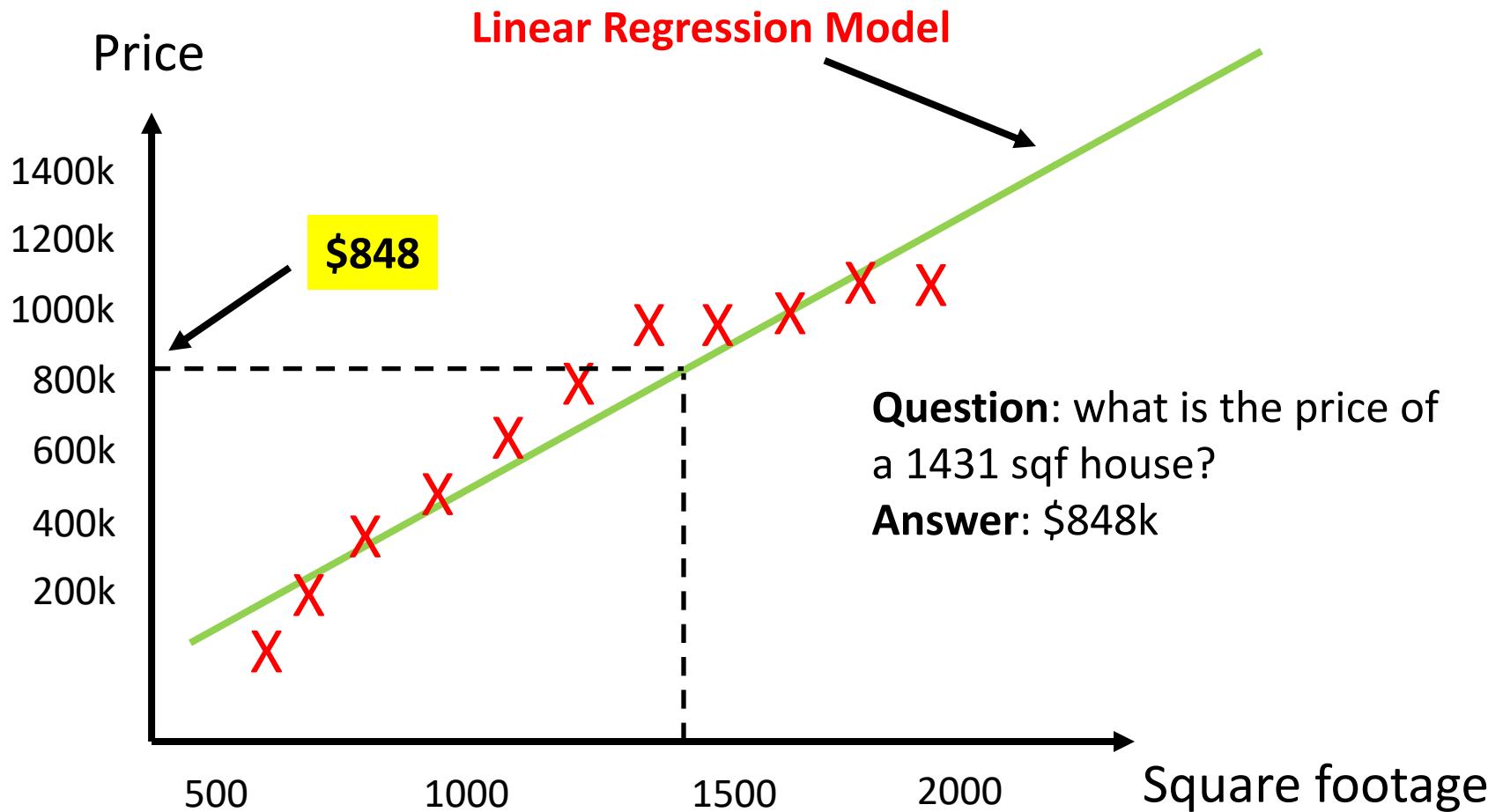
# Regression Example: Housing Price



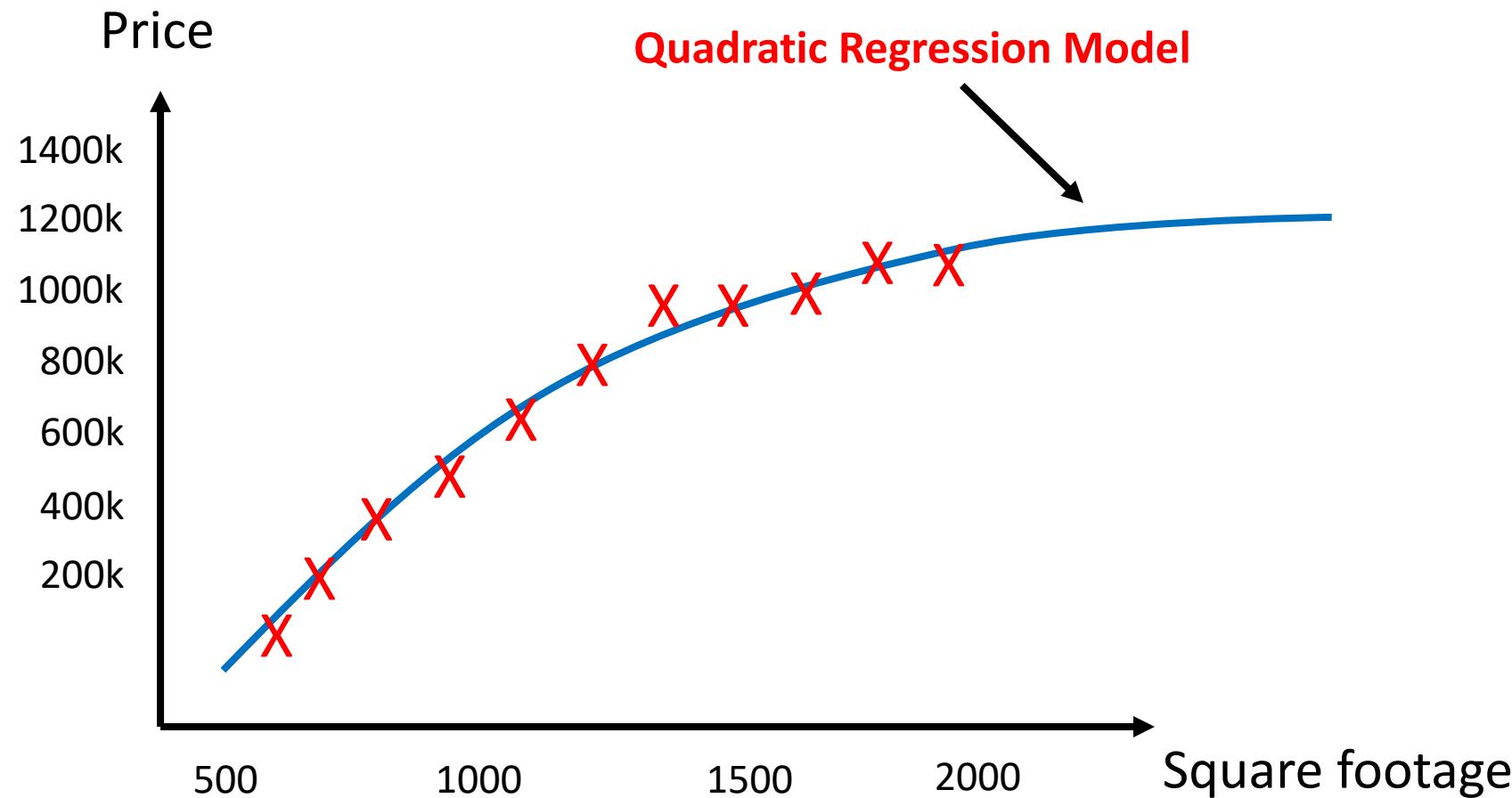
# Regression Example: Housing Price



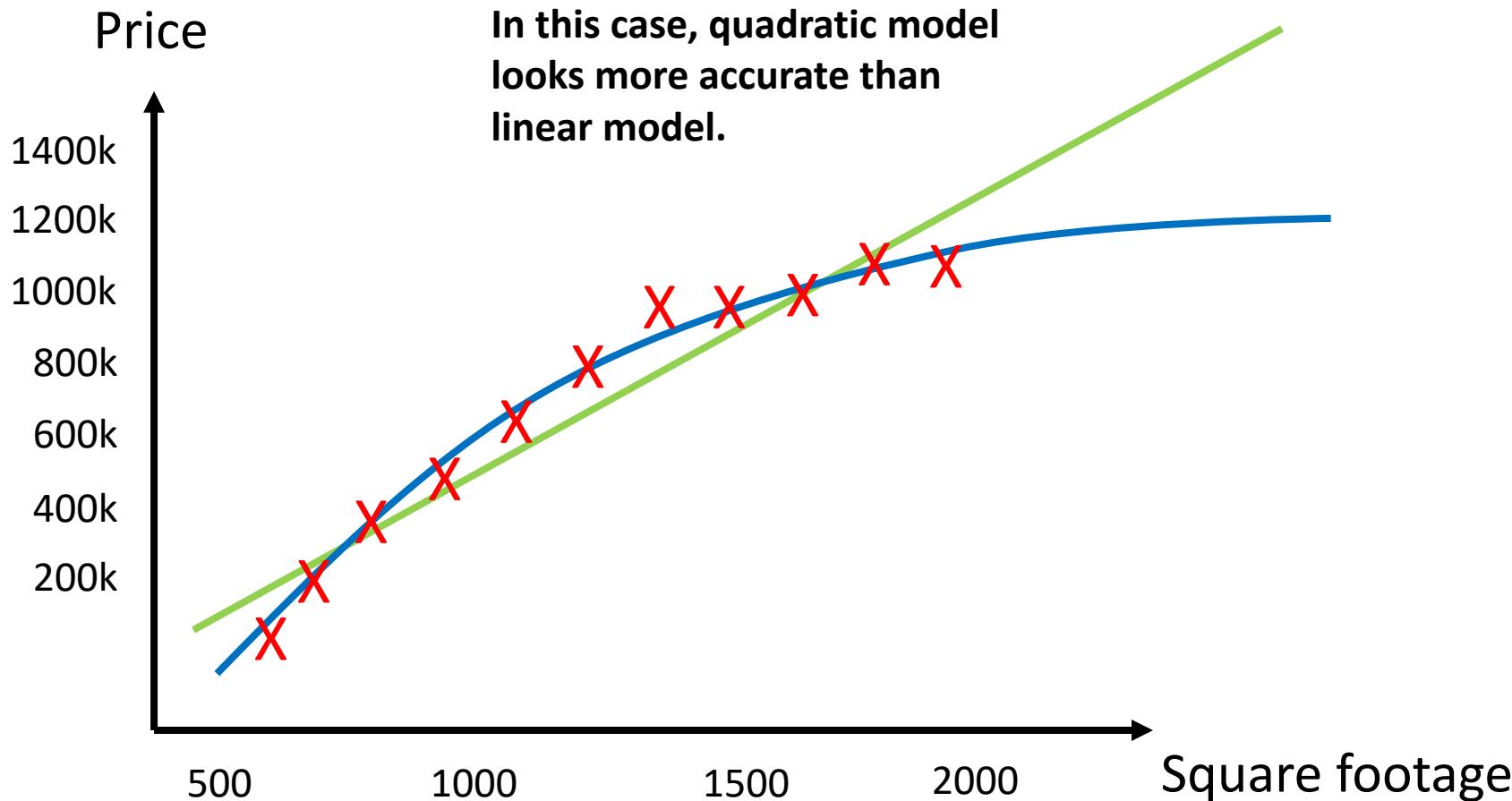
# Regression Example: Housing Price



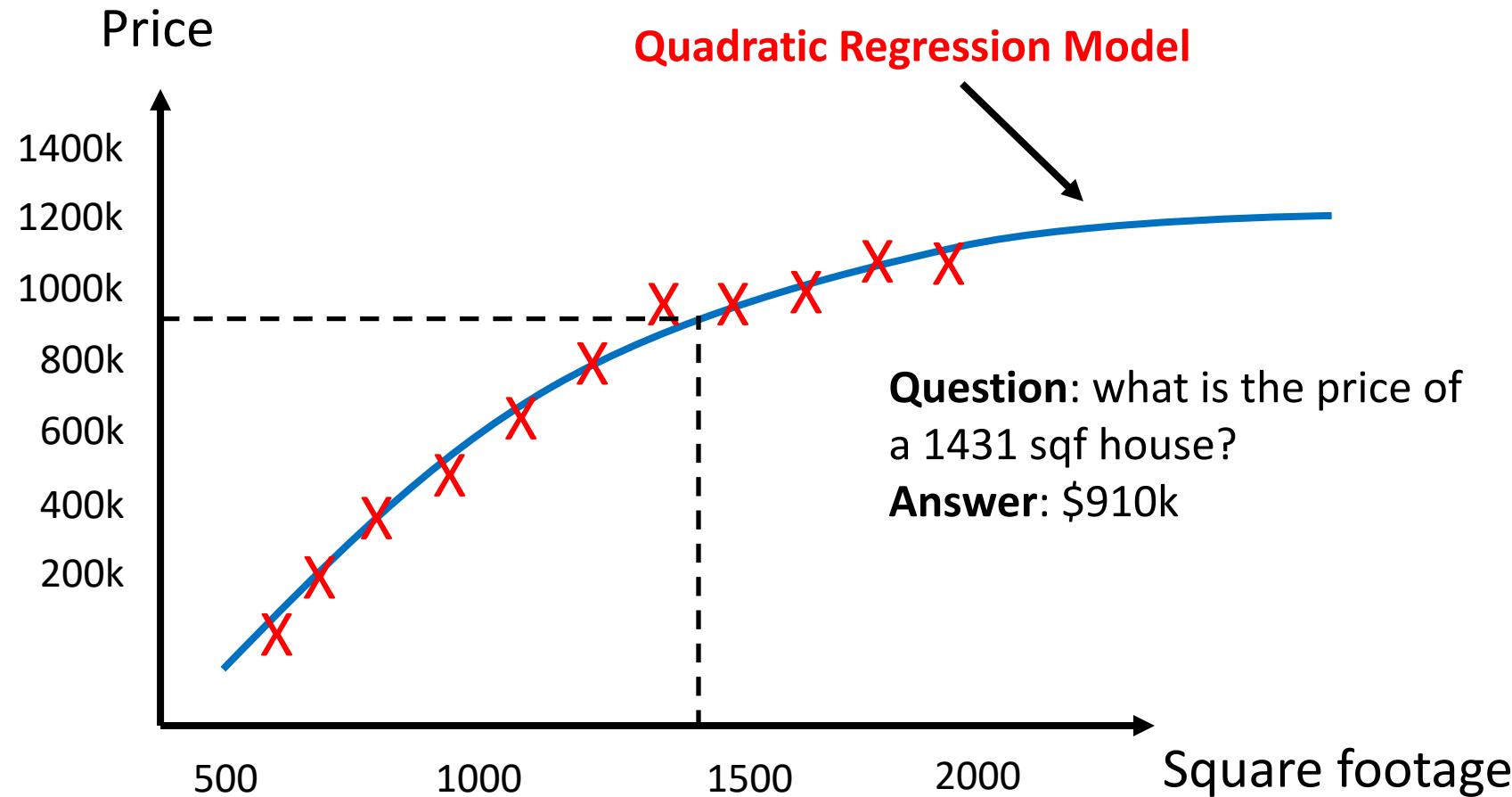
# Regression Example: Housing Price



# Regression Example: Housing Price



# Regression Example: Housing Price





*Thank You!*

**Questions?**