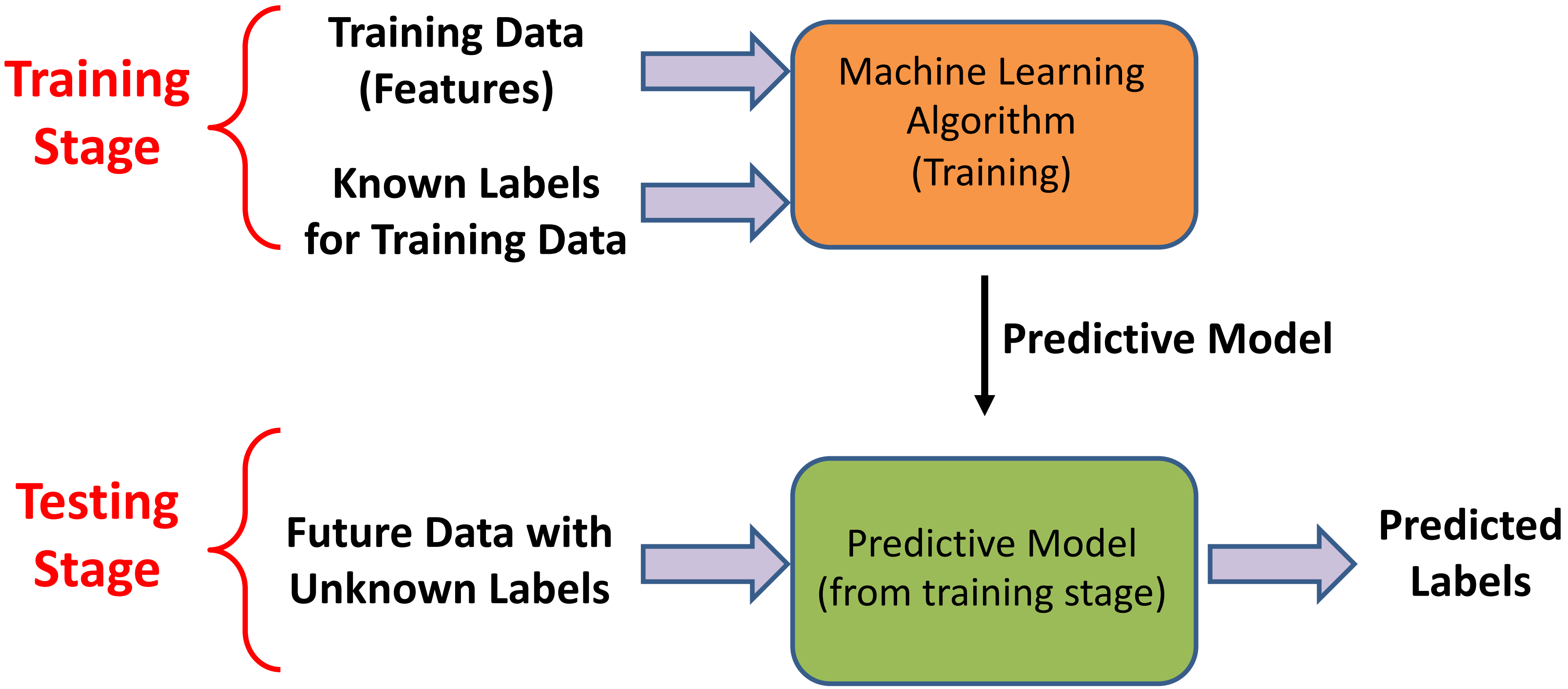
Advanced Machine Learning and Deep Learning

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

Supervised Learning: Learning from labeled Data

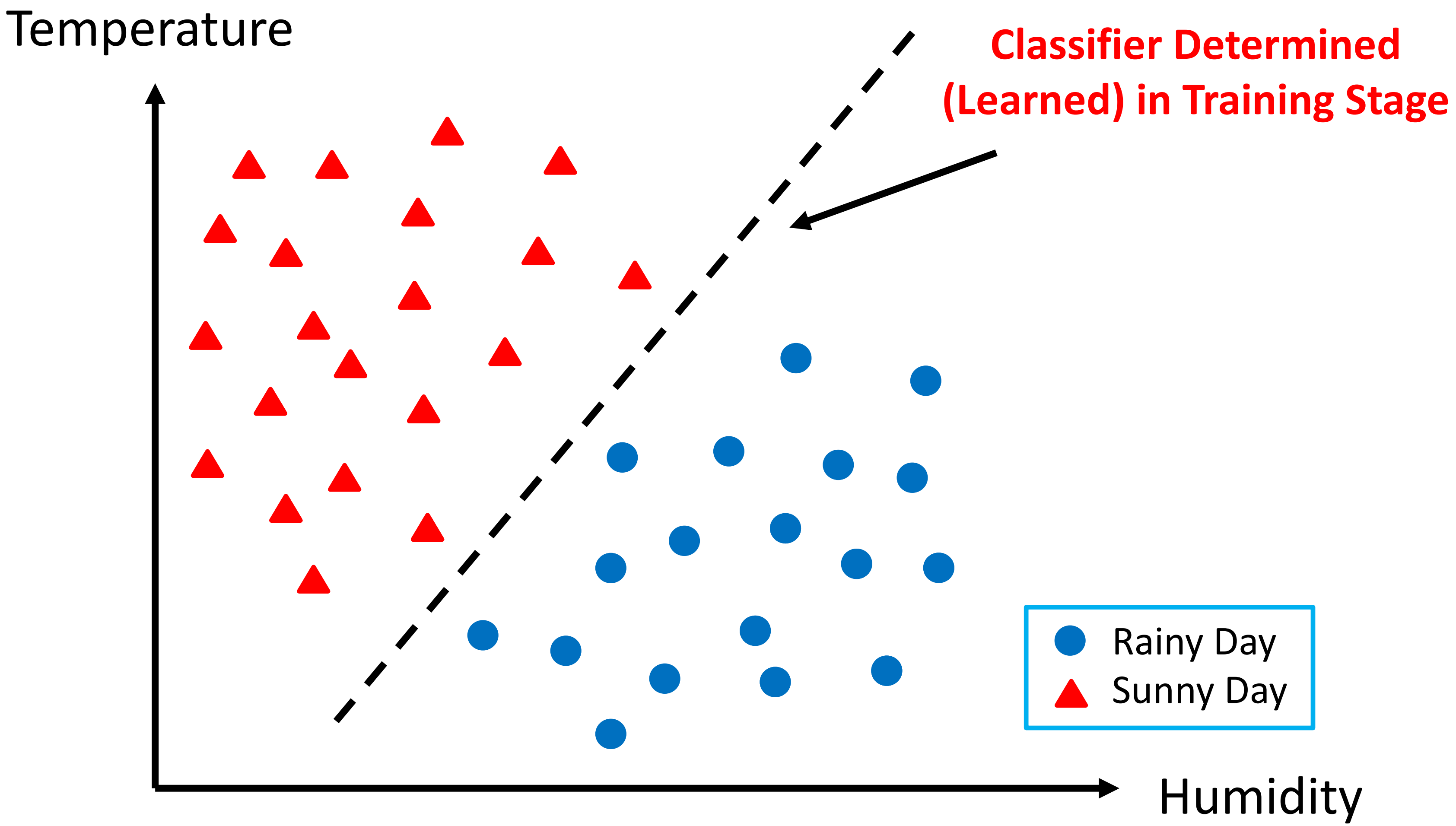


Two Basic Approaches of Supervised Learning

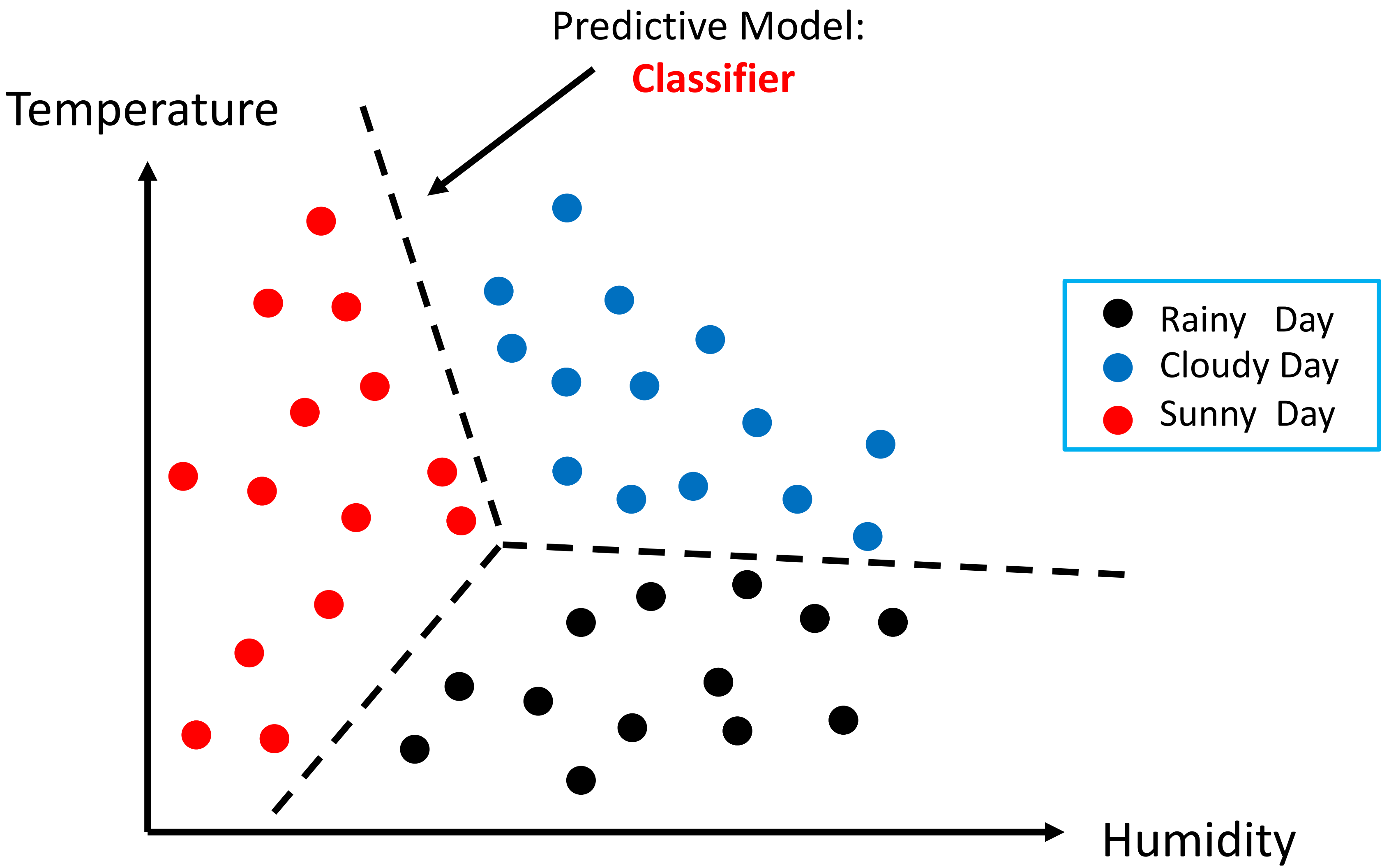
- **Classification**: Predict a **discrete** valued output for each observation.
  - Labels are discrete (categorical)
  - Labels can be binary (e.g., rainy/sunny, spam/non-spam,) or non-binary (e.g., rainy/sunny/cloudy)

- **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

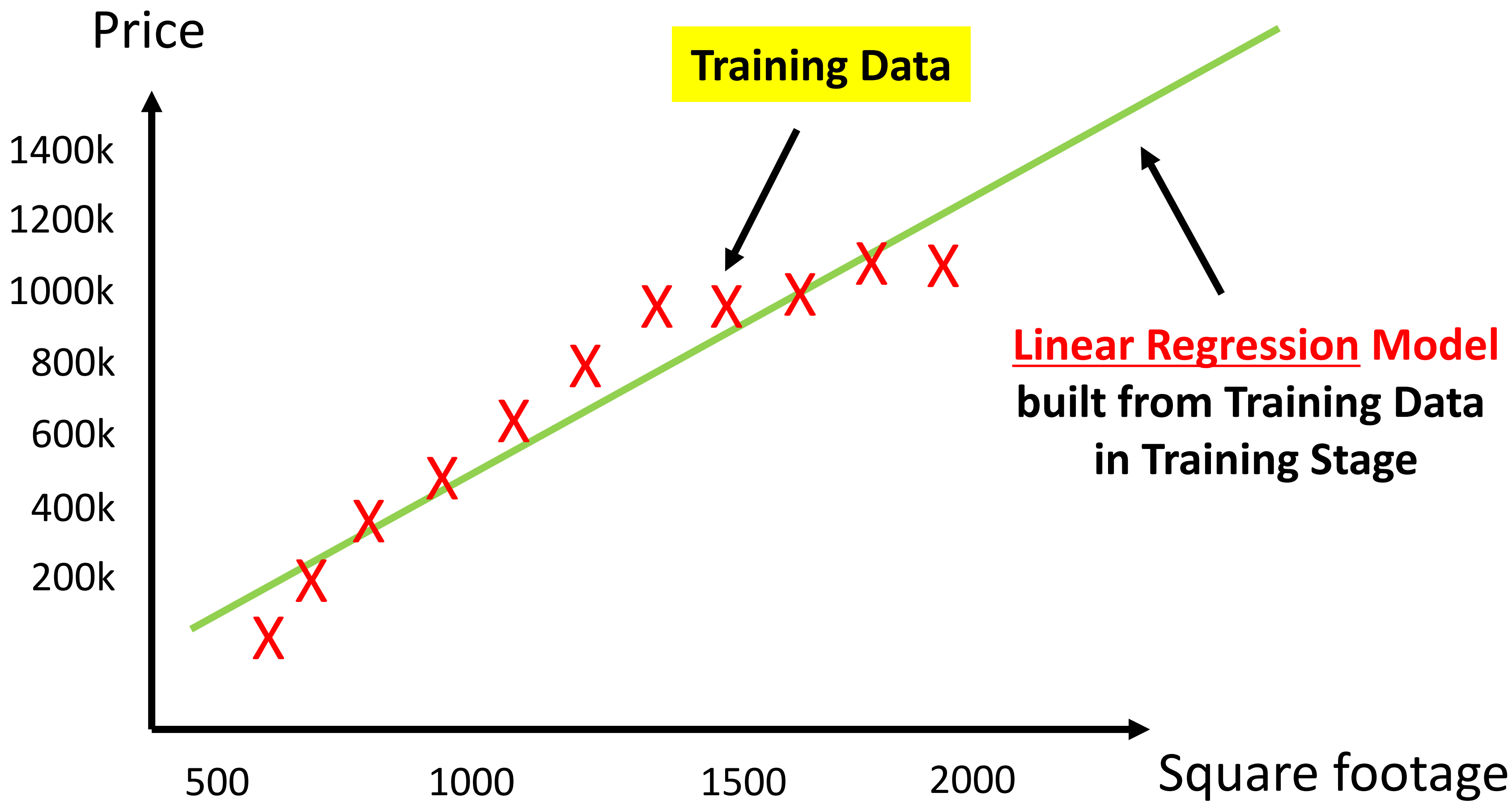


Classification Example: Multiple Label

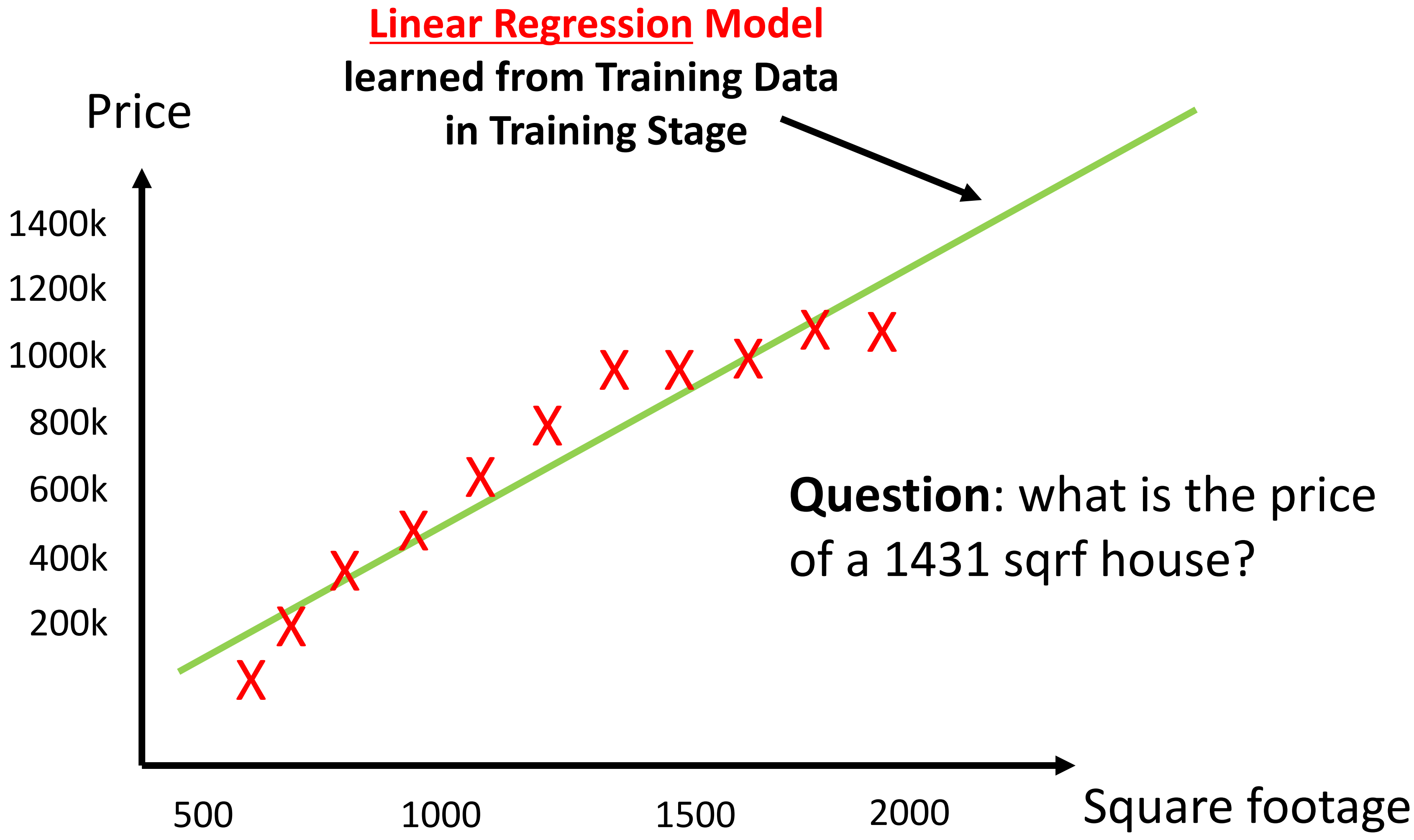


Regression Example: Housing Price

Regression Example: Housing Price

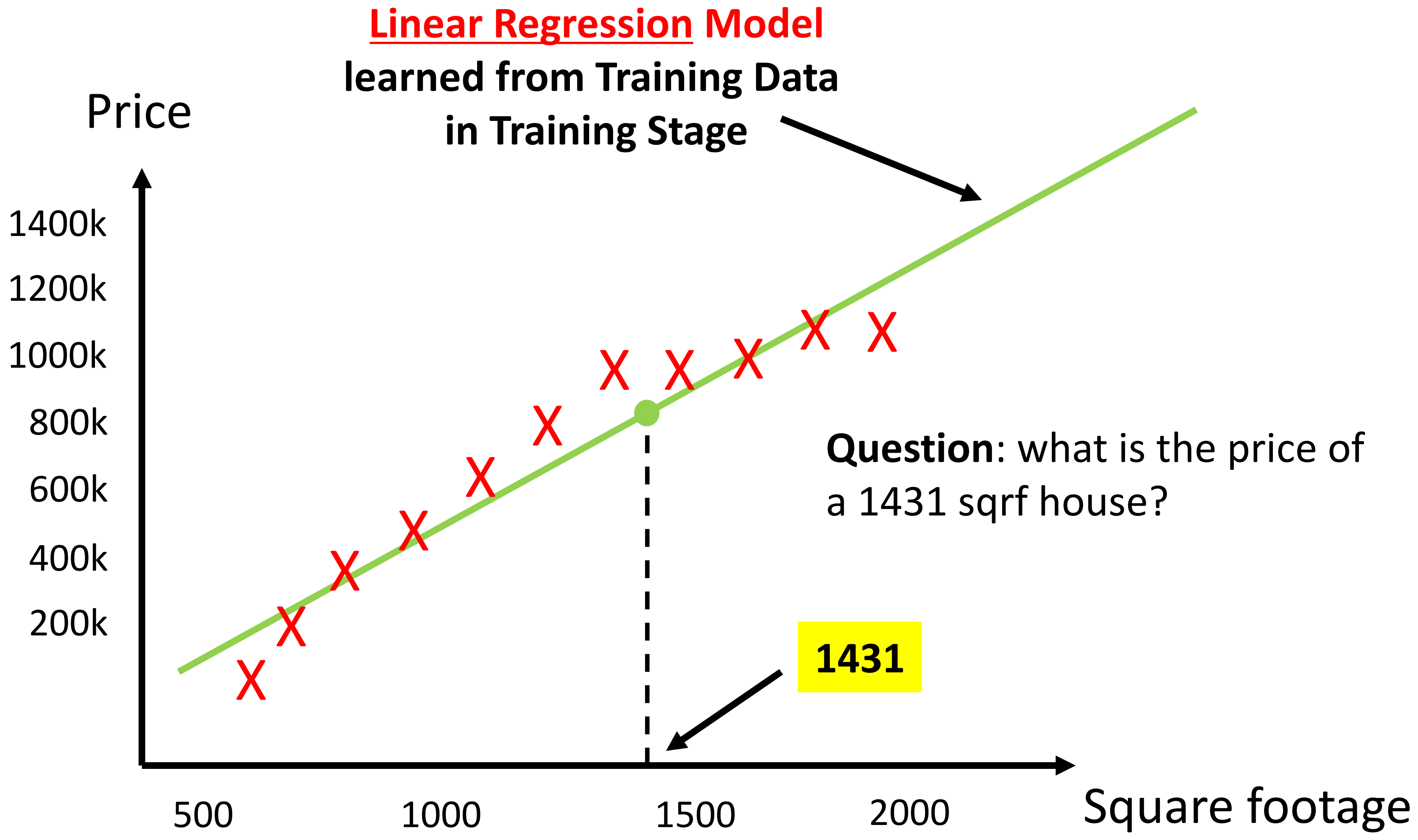


Regression Example: Housing Price



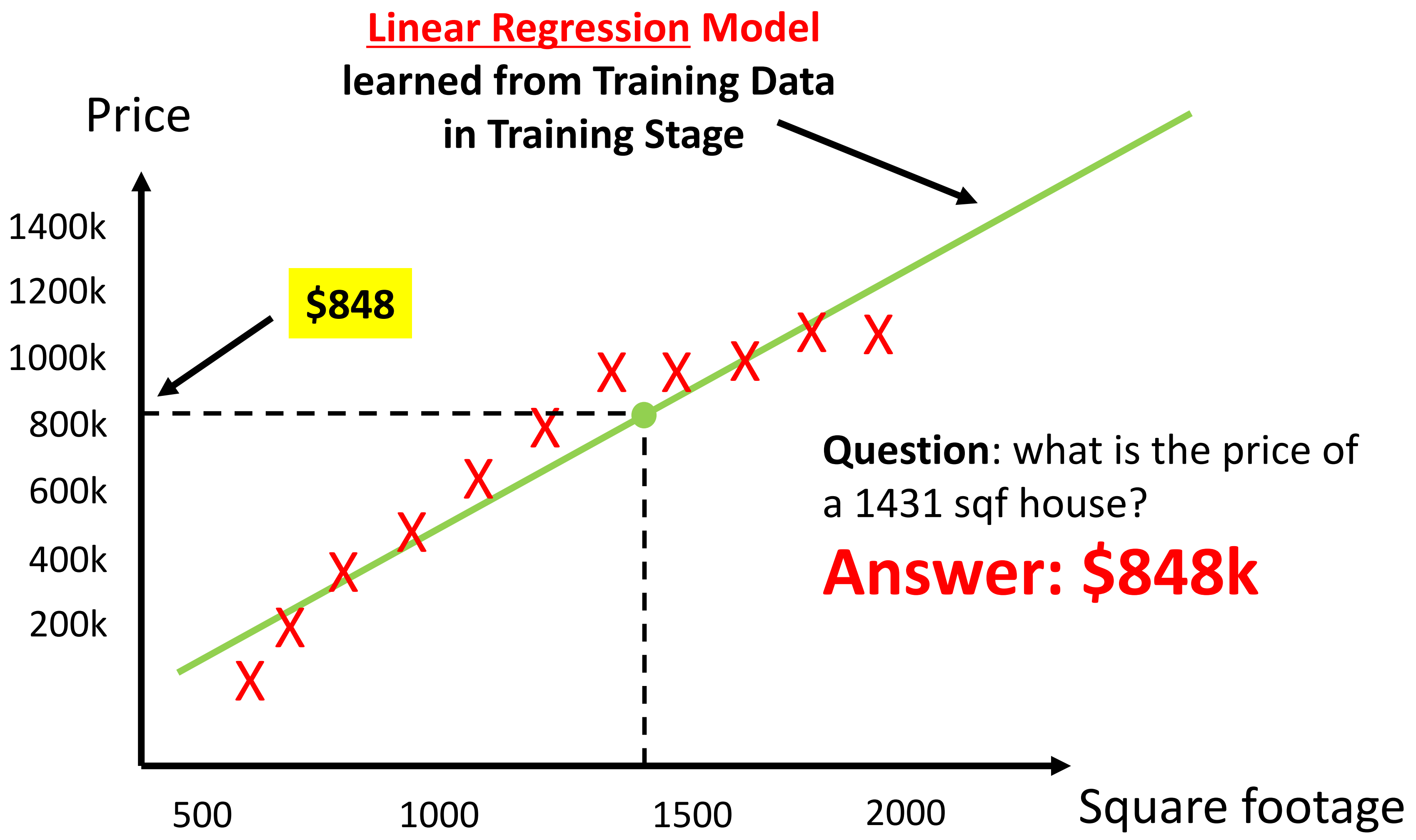
Question: what is the price of a 1431 sqrf house?

Regression Example: Housing Price



Question: what is the price of a 1431 sqrf house?

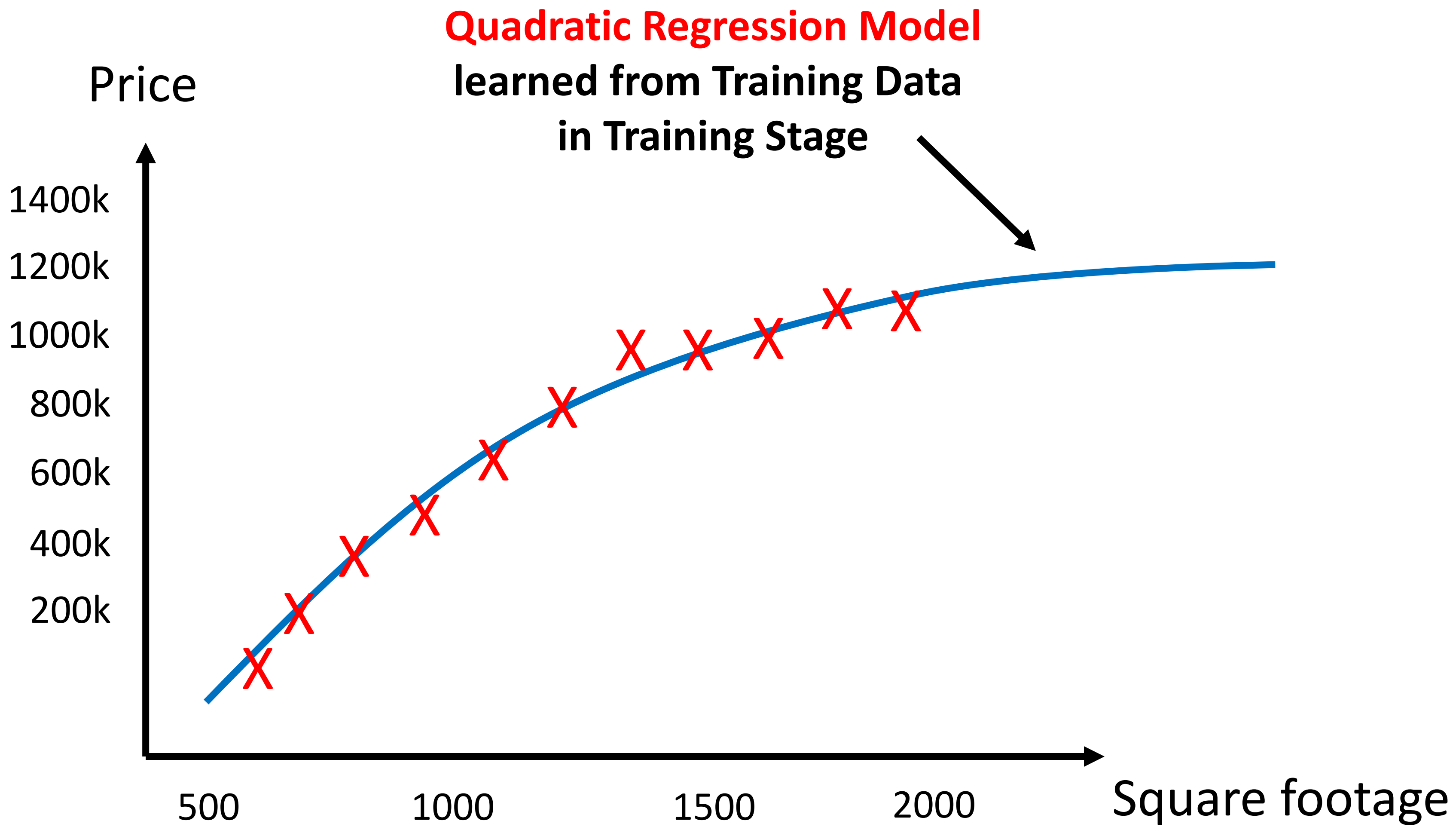
Regression Example: Housing Price



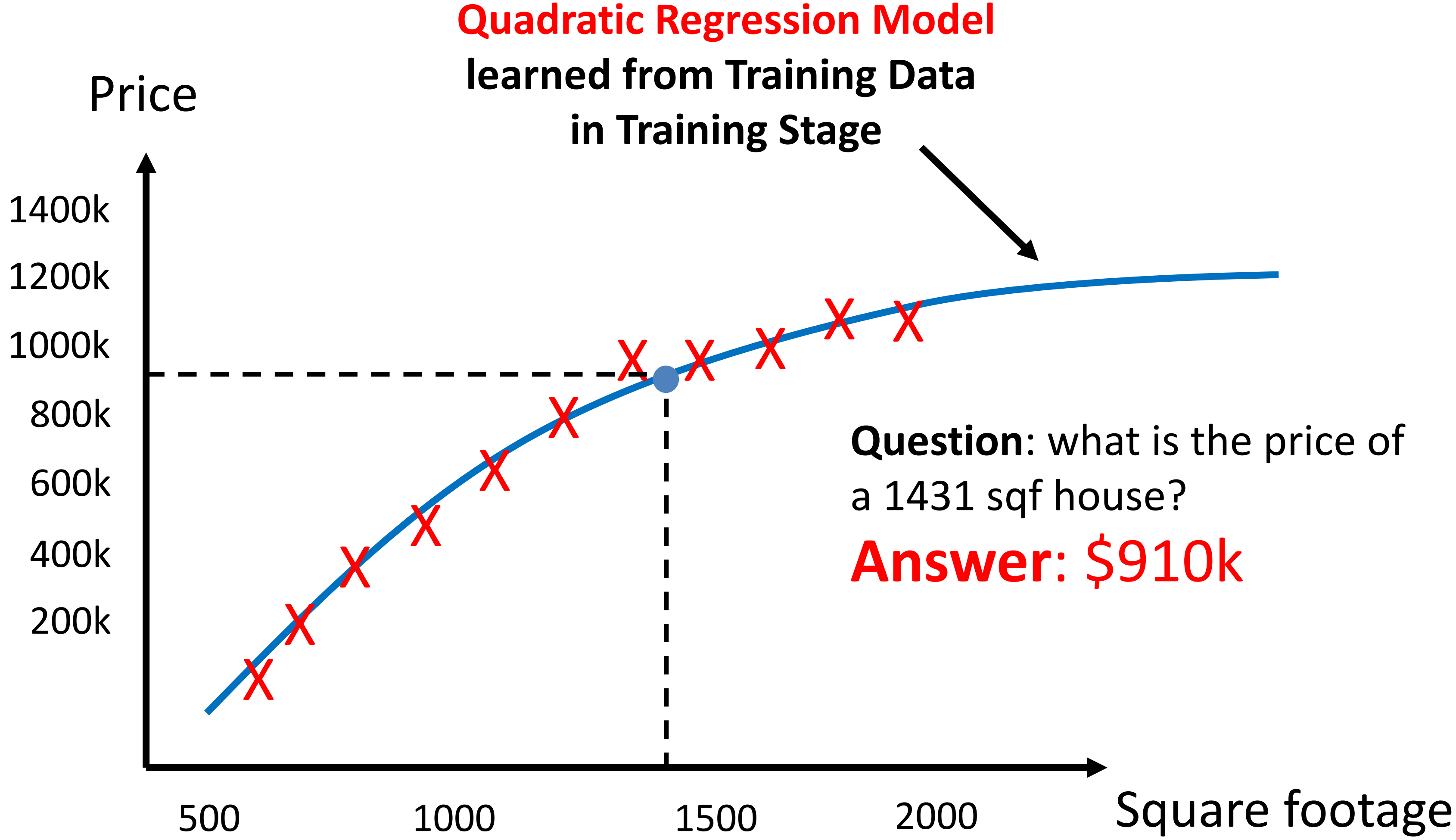
Question: what is the price of a 1431 sqf house?

Answer: $848k

Regression Example: Housing Price



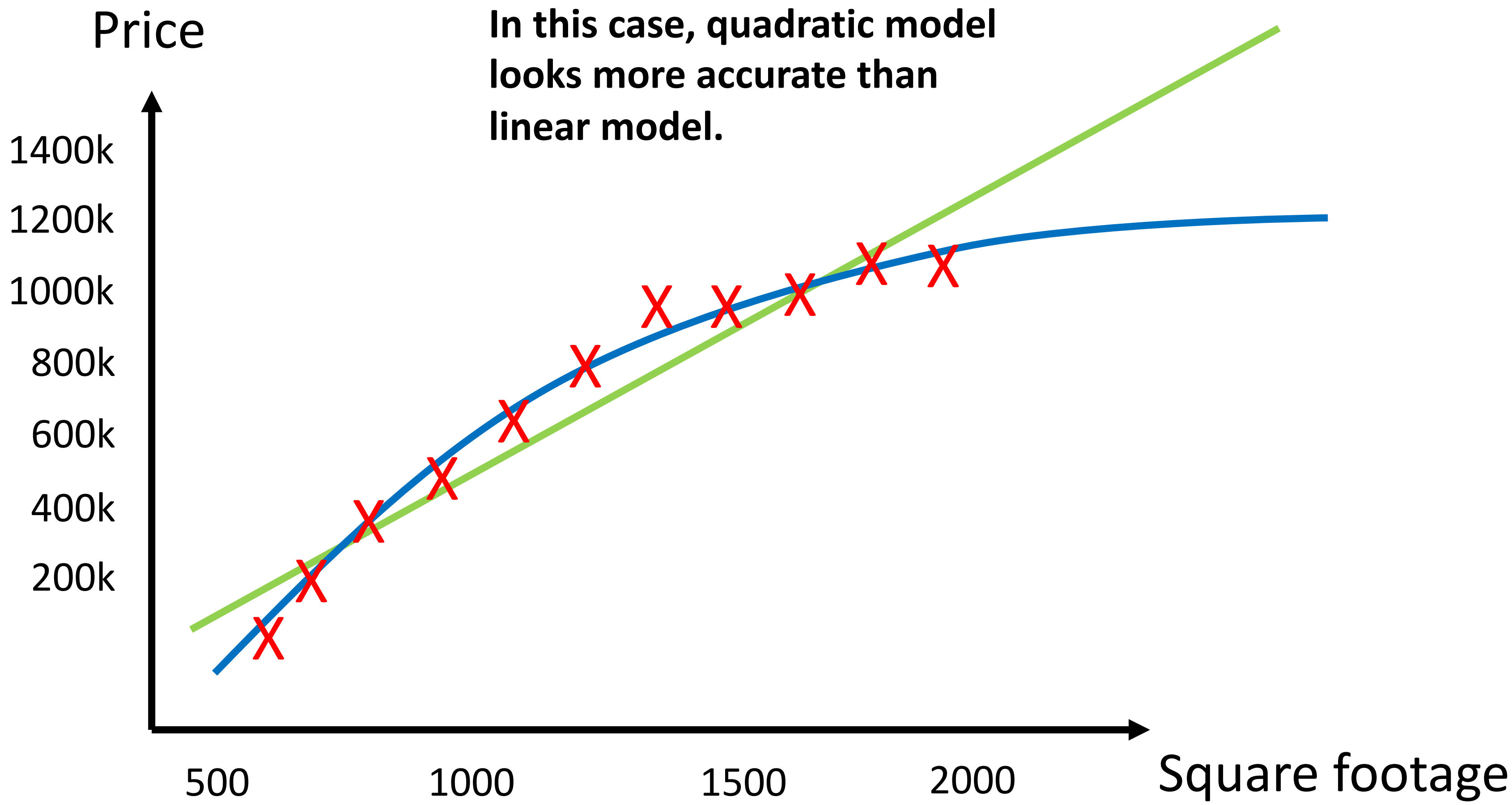
Regression Example: Housing Price



Question: what is the price of a 1431 sqf house?

Answer: $910k

Regression Example: Housing Price



Feature Table

- *Training dataset*: $\{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_N, y_N)\}$: N data samples used for training.

	sepal length	sepal width	petal length	petal width	Label	
$\boldsymbol{x}_1$	5.3	3.7	1.5	0.2	setosa	y_1
$\boldsymbol{x}_2$	5	3	2	0.2	setosa	y_2
$\boldsymbol{x}_3$	7.0	3.2	4.7	1.4	versicolor	y_3
	6.4	3.2	4.5	1.5	versicolor	
	6.3	2.7	4.9	1.8	virginica	
	7.9	3.8	6.4	2	virginica	

- *Training dataset*: $\{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_N, y_N)\}$ with known label.
- Now, we have a new sample with unknown label: $(\boldsymbol{x}\ , y=?)$

	sepal length	sepal width	petal length	petal width	Label	
$\boldsymbol{x}_1$	5.3	3.7	1.5	0.2	setosa	y_1
$\boldsymbol{x}_2$	5	3	2	0.2	setosa	y_2
	7.0	3.2	4.7	1.4	versicolor	
:	6.4	3.2	4.5	1.5	versicolor	:
	6.3	2.7	4.9	1.8	virginica	
$\boldsymbol{x}_N$	7.9	3.8	6.4	2	virginica	y_N
$\boldsymbol{x}$	7	3.9	5.9	1.3	???	$y=?$

The Problem of Overfitting

The Problem of Overfitting

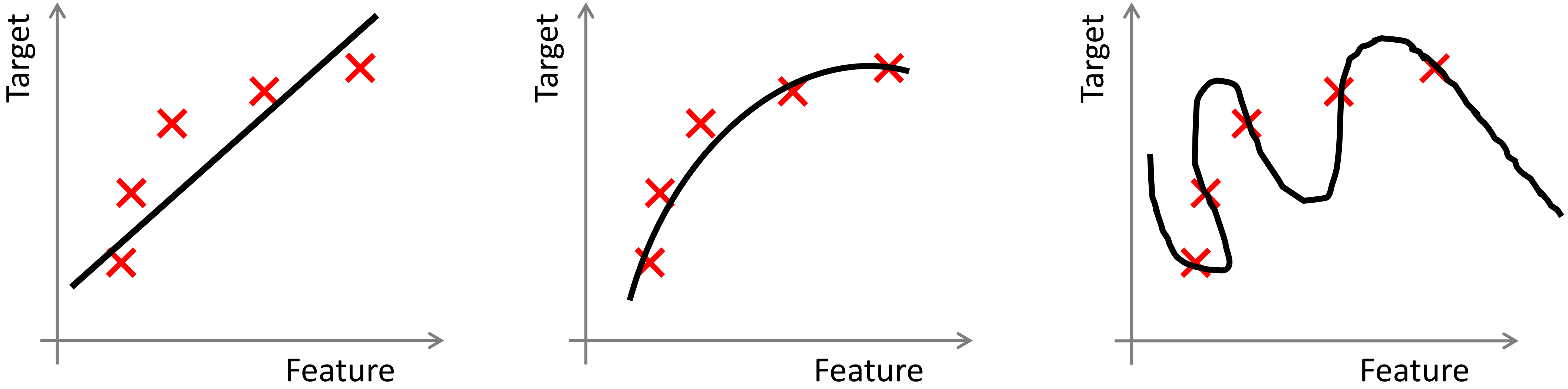
- **Overfitting** happens when the predictive model (classification model or regression model) **fits too much** with the **training samples** so that it starts capturing, learning, and representing the **noise and randomness or outlier samples** of the training dataset.

- Overfitting provides excellent accuracy for training data, but poor results for future data samples (testing set)!

The Problem of Overfitting

- **Overfitting** occurs when a model is excessively complex. The two main reasons that makes a model too complex are:
  1. having **too many features**.
  2. having a **complex model with very high order**.

Example of Overfitting for Regression



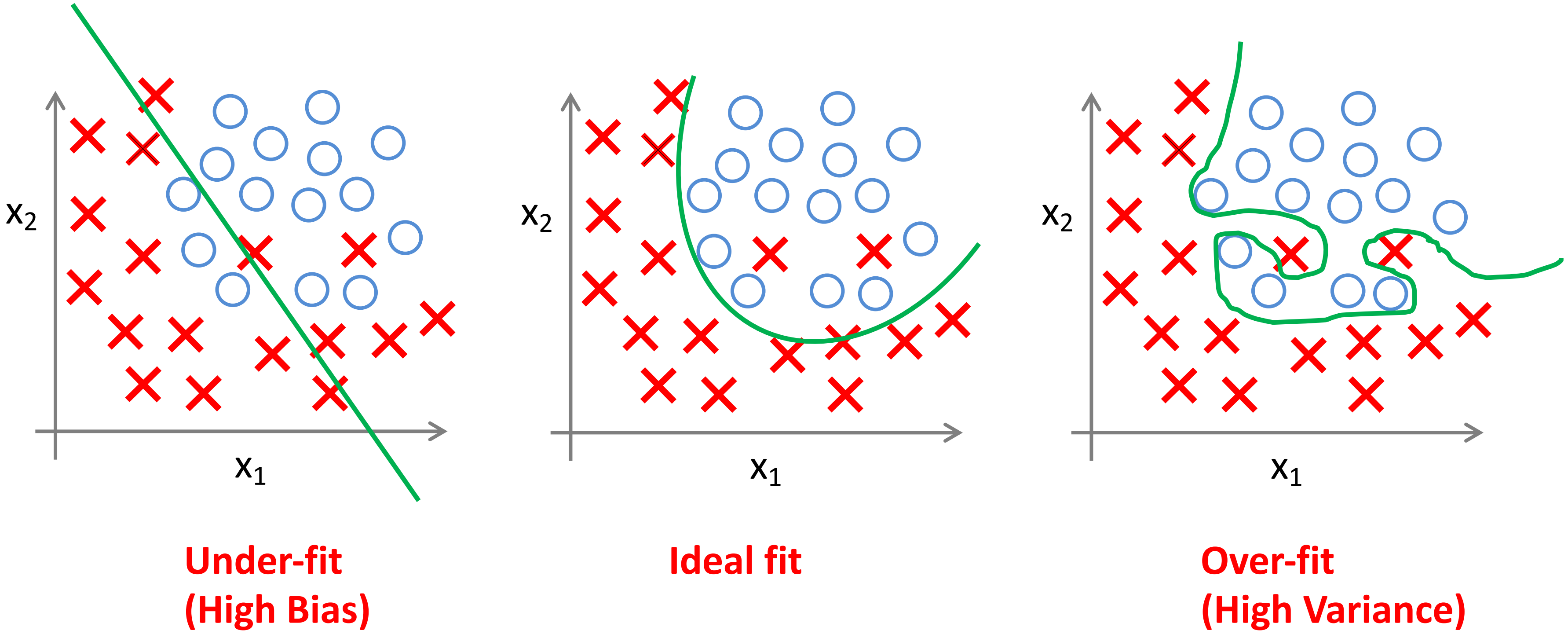
Under-fit
(High Bias)

Ideal fit

Over-fit
(High Variance)

*Reference: Andrew Ng, Stanford University.

Example of Overfitting for Classification



*Reference: Andrew Ng, Stanford University.

Addressing the Overfitting Problem: Approach 1: Dimensionality Reduction

- **Approach 1**: **Dimensionality Reduction:**
  - Reduce the number of features x (e.g. rather using 20 features for prediction, use only the best 3 features)

$$\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \ldots + \theta_{20} x_{20} \quad \rightarrow \quad \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3$$

We can:

a) Manually select which features to keep.

b) Detecting the **best features** using automated **Feature Selection** and/or **Dimensionality Reduction** algorithms (will be covered later).

Feature Selection

- **Feature selection** is an important step in machine learning. The classic feature selection algorithms usually focus on specific metrics to **quantify the relevance and/or redundancy of each feature** with the goal of finding **the smallest subset of features that provides the maximum amount of useful information** for prediction.
- Thus, the **main goal of feature selection algorithms is to eliminate redundant or irrelevant features** in a given feature set.
- Applying an effective feature selection algorithm not only decreases the complexity of the system by reducing the dimensionality, but also increases the performance of the classifier by avoiding overfitting and also removing irrelevant and confusing features.

Addressing the Overfitting Problem: Approach 2: Regularization

- **Approach 2**: **Regularization**:
  - Keep all features, but reduce the magnitude/values of parameters of the model (θ_j) to simplify the model.

$$\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 + \theta_4 x_1^2 + \theta_5 x_2^2 + \theta_6 x_2 x_3 + \theta_6 x_2 x_3^2 + \ldots$$

$$\rightarrow \quad \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 + \theta_4 x_1^2 + \theta_6 x_2 x_3$$

Thank You!

Questions?