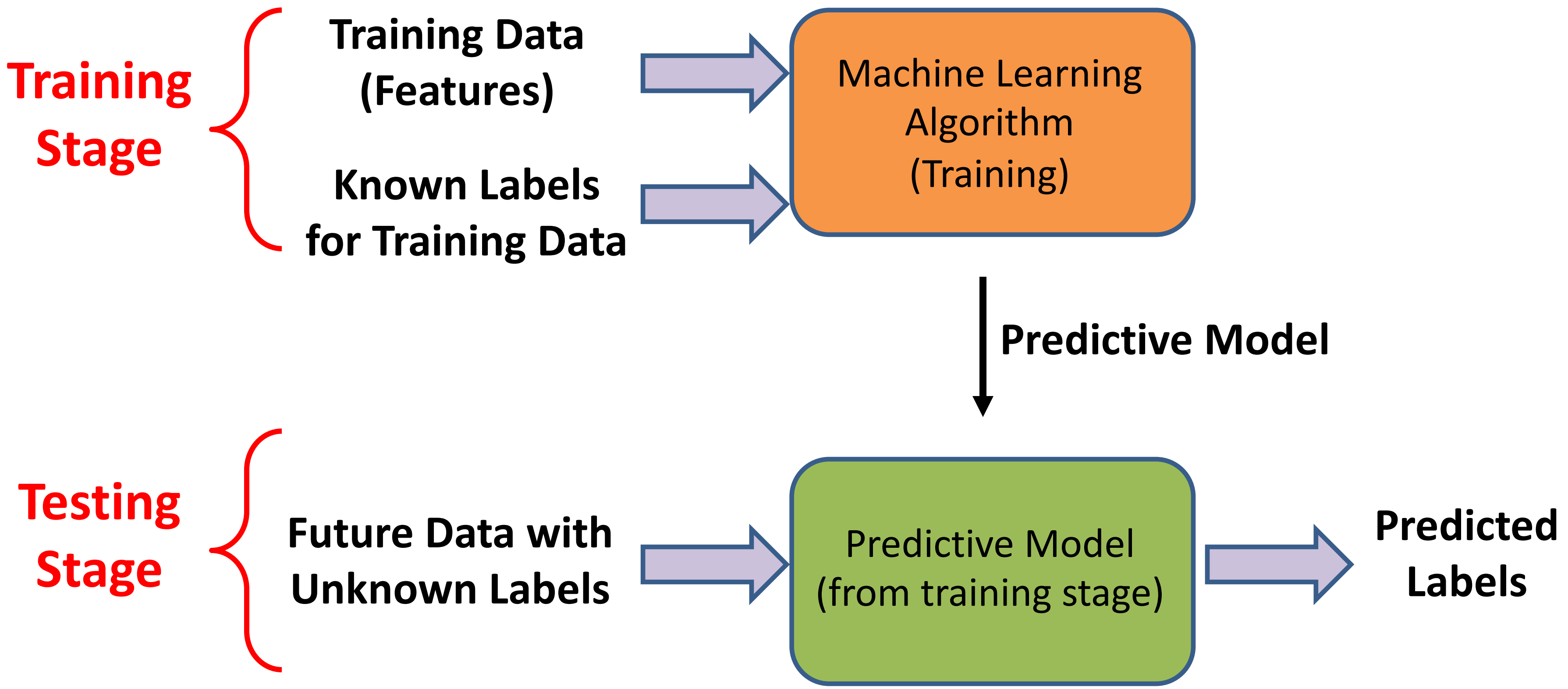
Introduction to Data Science

(Lecture 10)

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

Linear Regression

Review: Supervised Learning: Learning from labeled Data



Review: Two main approaches in supervised learning

- **Classification**: Predicts a discrete valued output.
  - Labels are discrete (categorical)
  - Labels can be binary (e.g., rainy/sunny, spam/non-spam,)
    or non-binary (e.g., rainy/sunny/cloudy, Setosa/Versicolor/Virginica)

- **Regression:** Predicts a continuous valued output.
  - Labels are continuous, e.g., stock price, housing price
  - Can define 'closeness' when comparing prediction with true values

Regression Examples

- Predicting a ***company's future stock price*** using its profit and other financial information.
- Predicting ***housing price*** based on the property information (size, # rooms, …).
- Predicting ***annual rainfall*** based on the weather information.
- Predicting a ***risk of granting a credit card*** based on financial and personal information.
- **…**

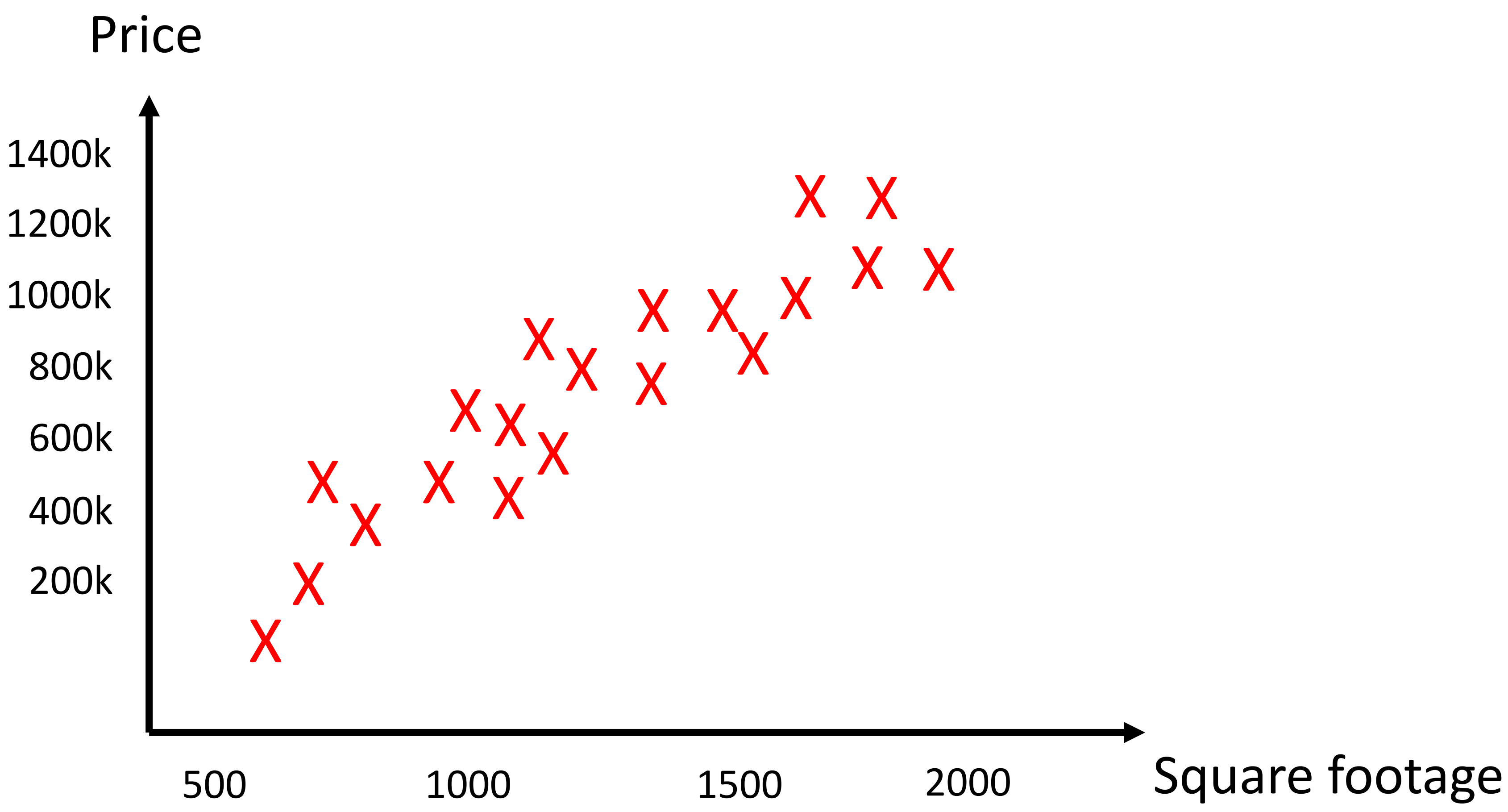
Special Case

- Let’s start with a simple, special case:

  Linear Regression with **single input** variable (**one feature**).

  It is also called “***Univariate Linear Regression.***”

Regression Example: Housing Price



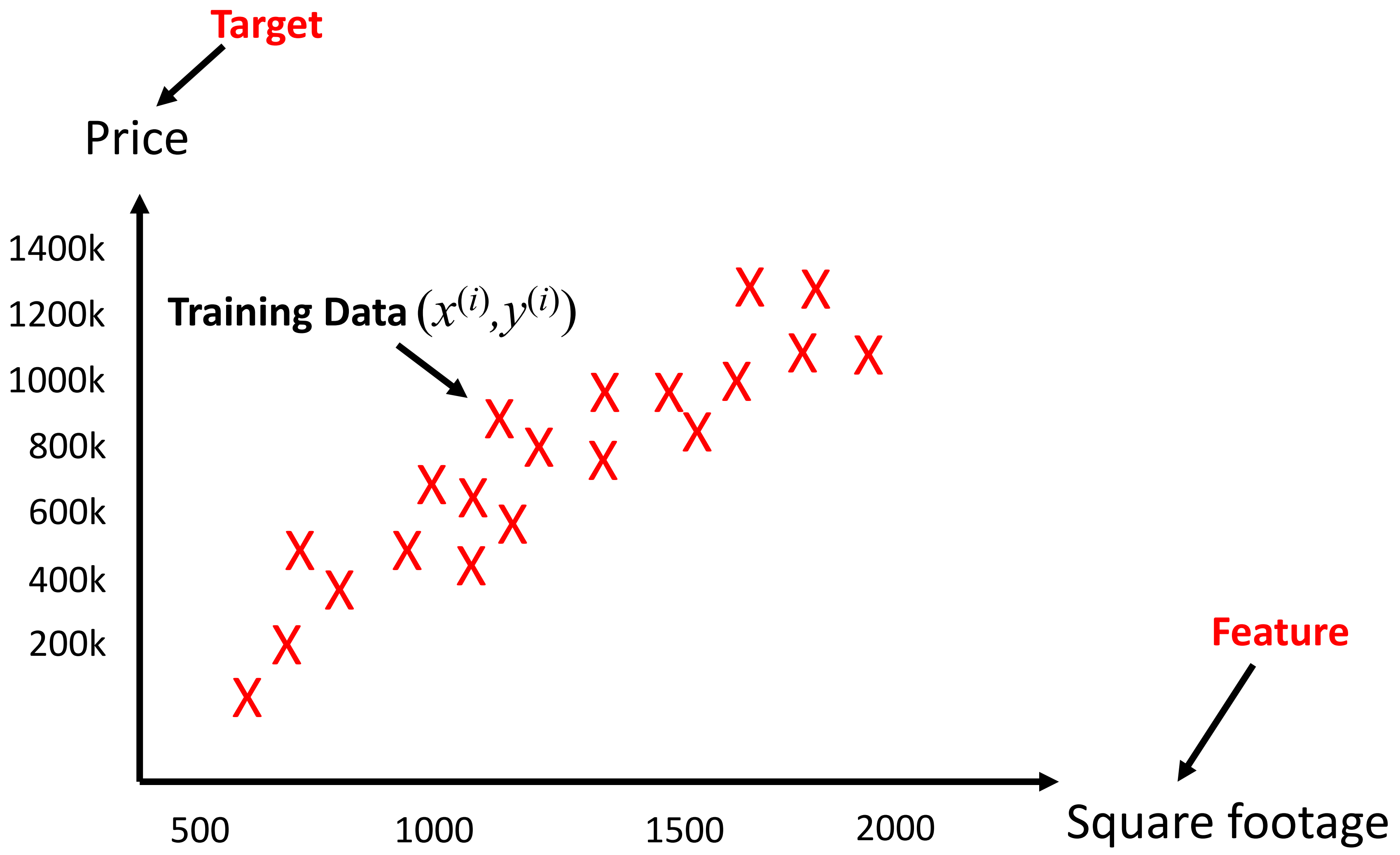
Training Dataset (past sales records)

Important Notation:

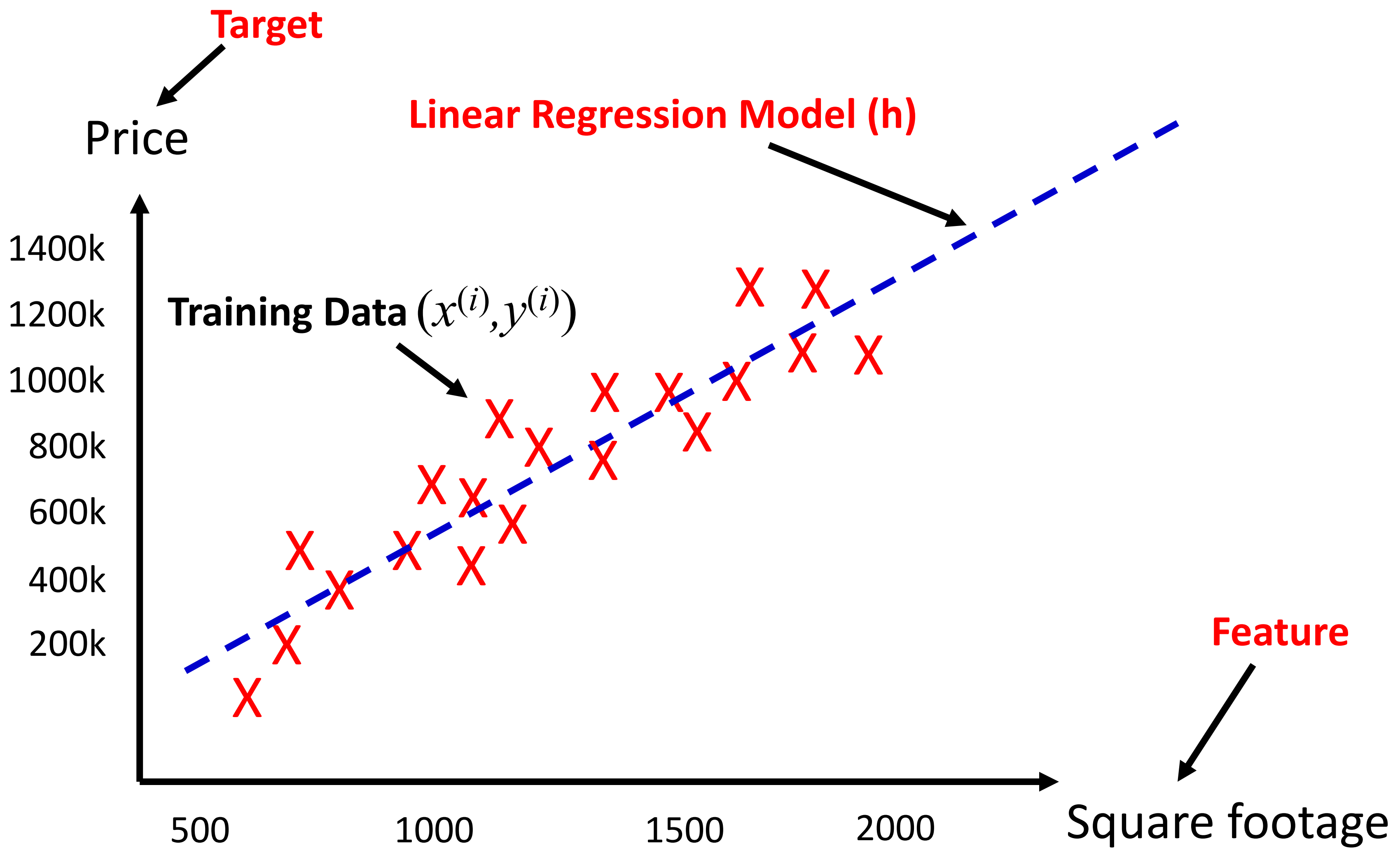
- $\boldsymbol{x}$'s = "input" variable (**features**)
- y's = "output" variable (**target**)
- $(\boldsymbol{x},y)$ = one training sample
- $(\boldsymbol{x}^{(i)},y^{(i)})$ = i^{th} training sample (**Note**: (i) is index not exponent!)
- m = Number of training examples

Size in feet^2 (***x***)	Price (*y*)
1000	410K
1200	600K
1230	620K
1340	645K
…	…

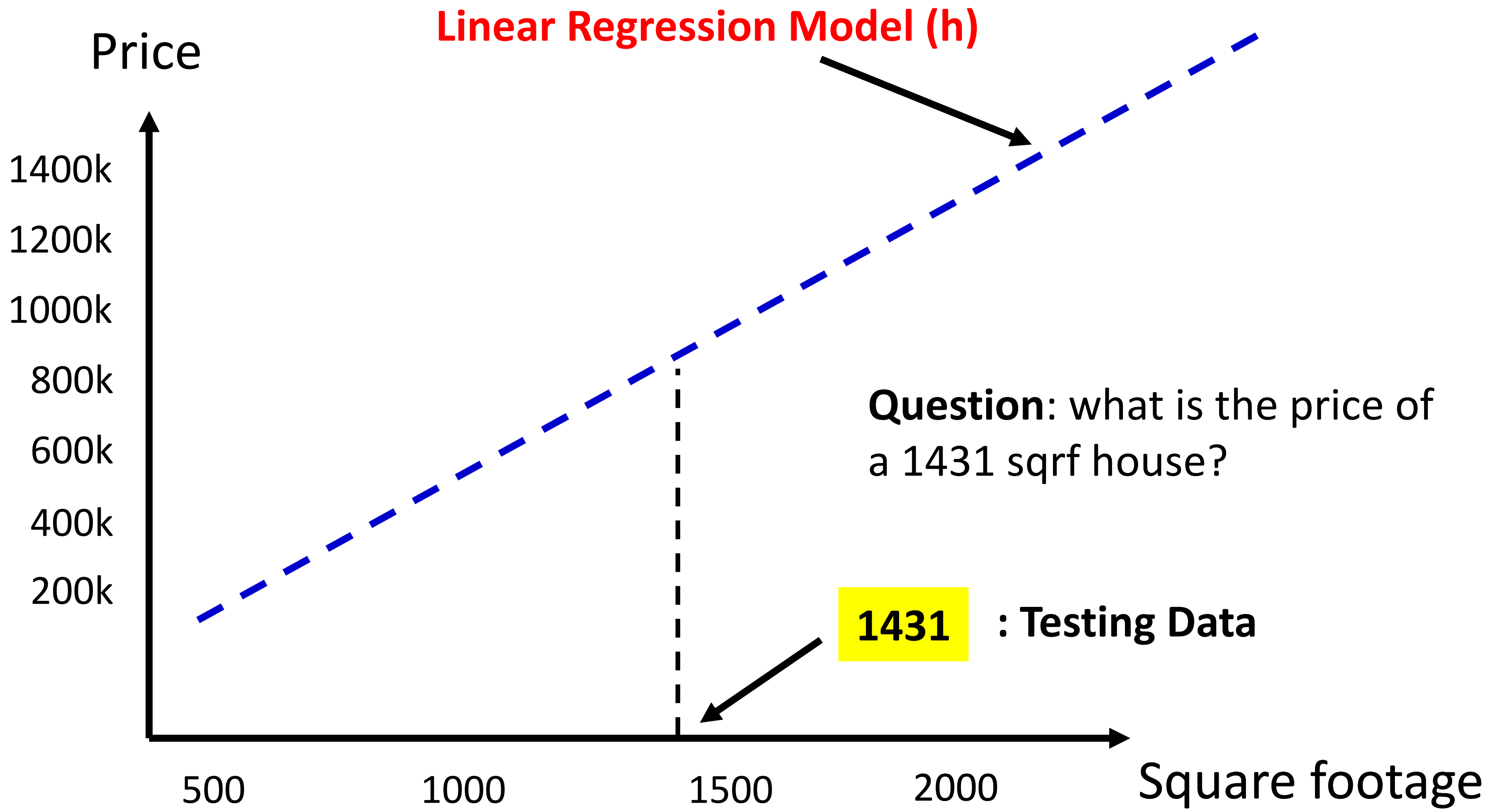
Regression Example: Housing Price



Regression Example: Housing Price

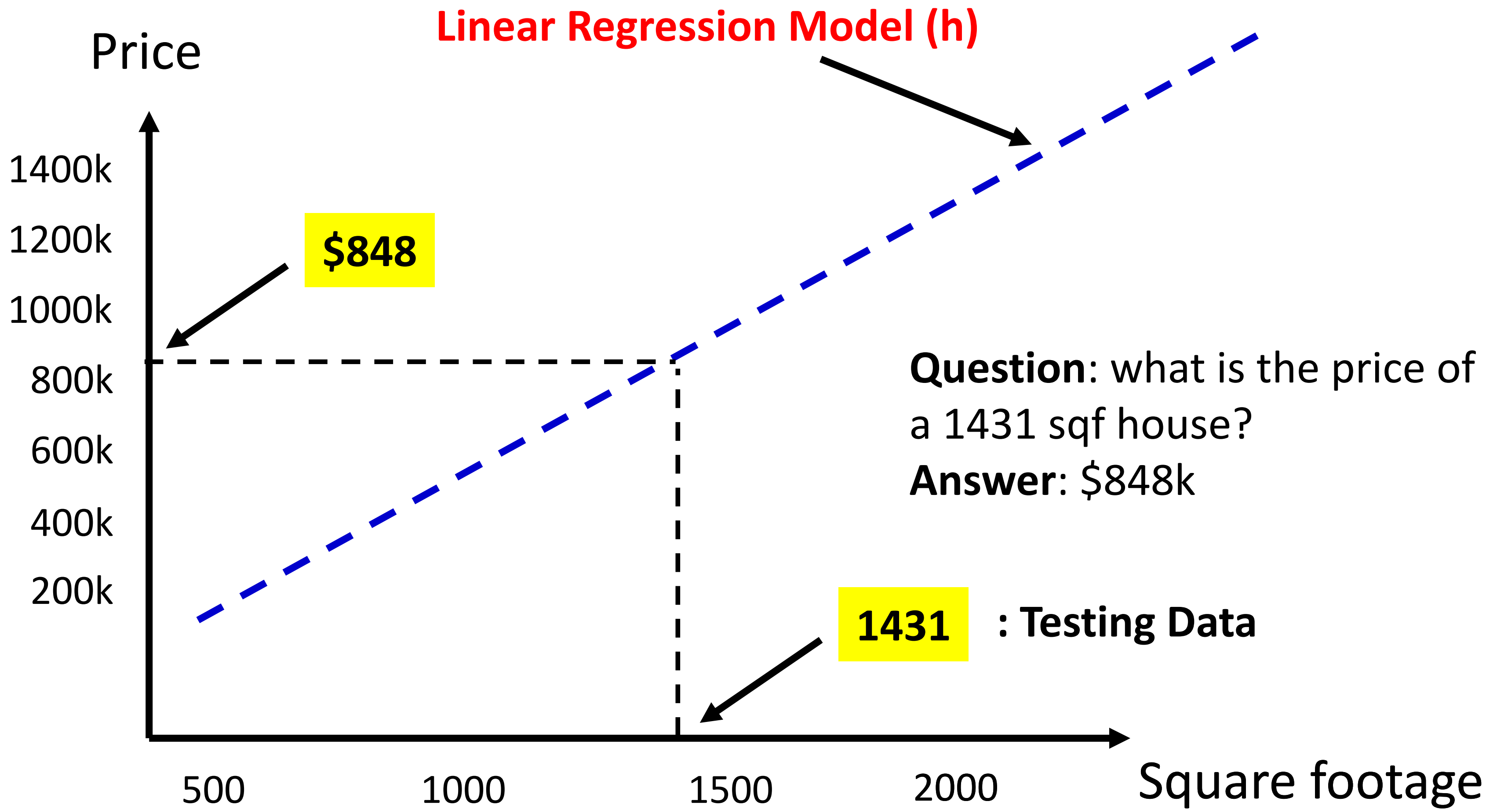


Regression Example: Housing Price

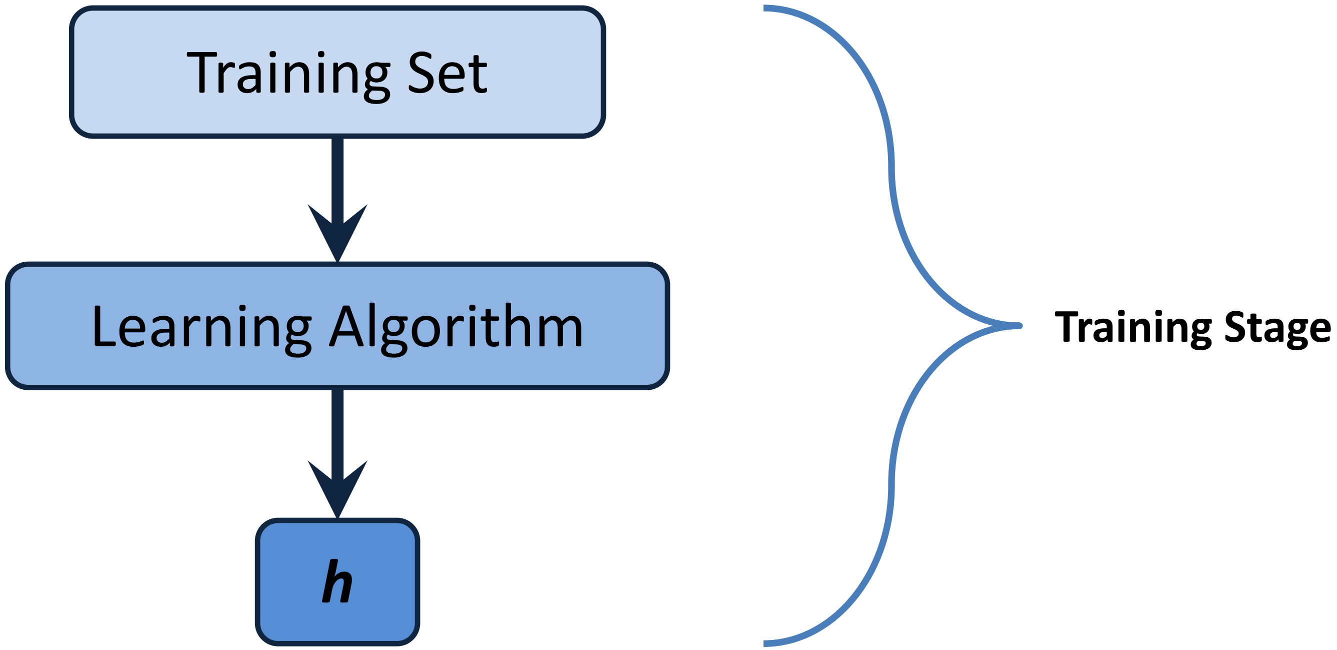


Question: what is the price of a 1431 sqrf house?

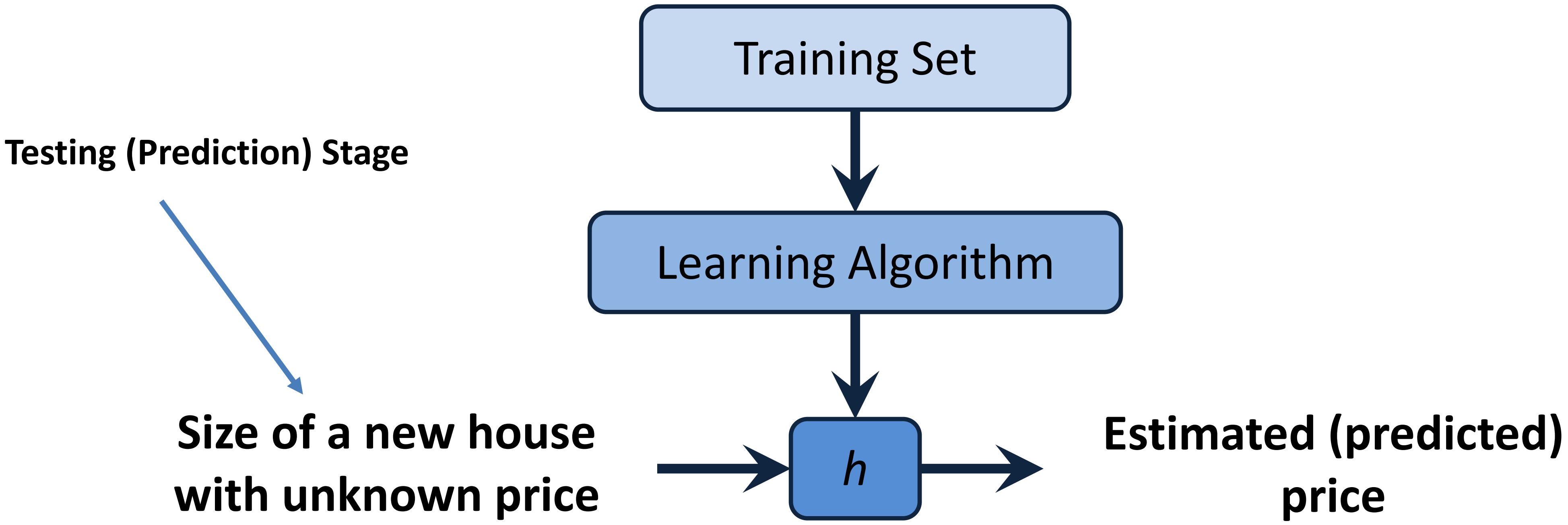
Regression Example: Housing Price



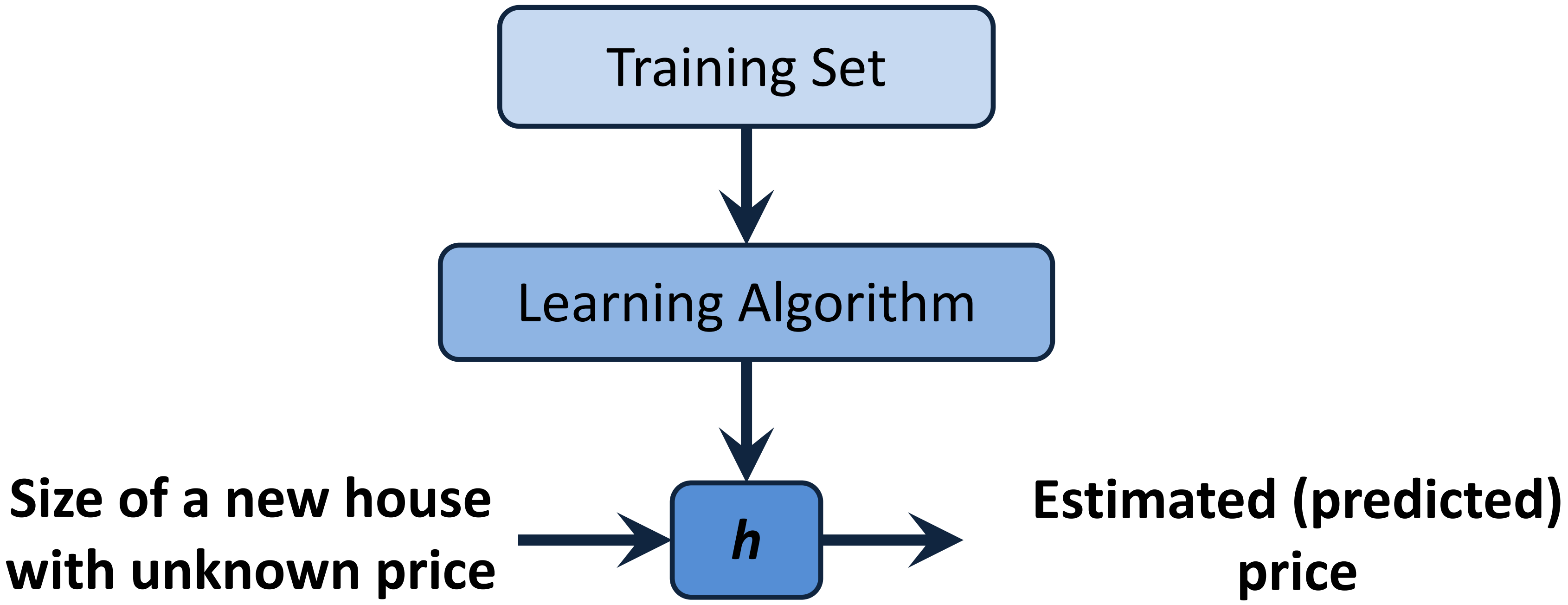
Question: what is the price of a 1431 sqf house?
Answer: $848k



- *"h"* is the <u>Regression Model</u> (Predictive Model) built by machine learning algorithm based on the training set. It will be later used to predict or estimate the "target" (e.g. price) on new observations.



- ***"h"*** is the Regression Model (Predictive Model) built by machine learning algorithm based on the training set. It will be later used to predict or estimate the "target" (e.g. price) on new observations.



- In the case of **Linear Regression** with **One Variable**, the Regression model (h) will be a line:

$$h_\theta(x) = \theta_0 + \theta_1 x$$

- In the case of Linear regression with one variable, the Regression model (h) will be a line:

$$h_\theta(x) = \theta_0 + \theta_1 x$$

- This line should be the **Best Fit** to the **training samples**!

How to find the Regression Model?

How to find the Regression Model*

Training Sets (x,y)
(past sales records):

Size in feet^2 (*x*)	Price (*y*)
1000	410K
1200	600K
1230	620K
1340	645K
…	…

Regression Model (Hypothesis): $h_\theta(x) = \theta_0 + \theta_1 x$

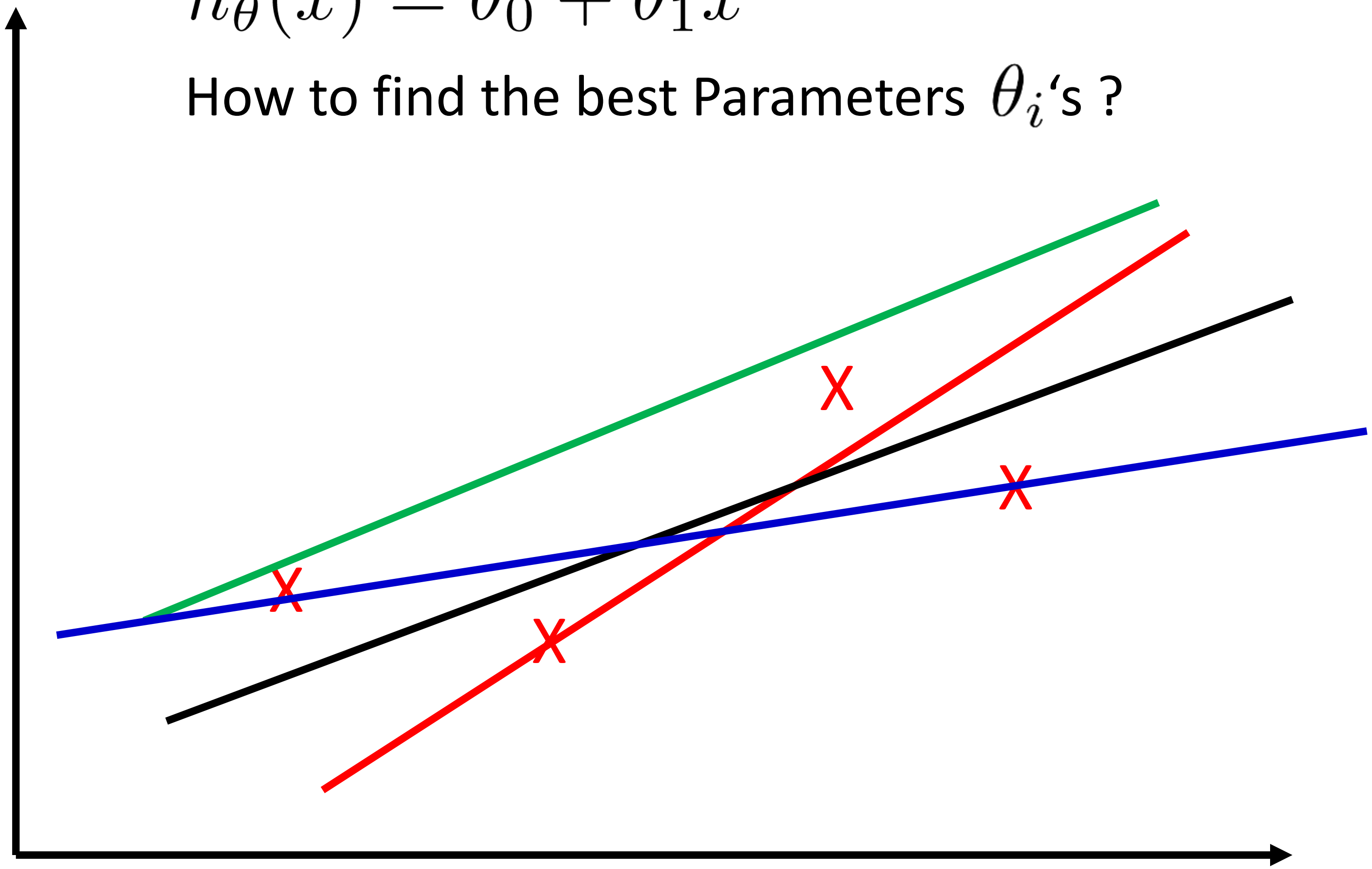
Question: How to find the best fit line? In other word, How to find the Parameters θ_0, θ_1 that correspond to the best line?

* Reference: Andrew Ng, Machine Learning, Stanford University.

How to Select the Parameters: A Simple Example

$$h_\theta(x) = \theta_0 + \theta_1 x$$

How to find the best Parameters θ_i‘s ?



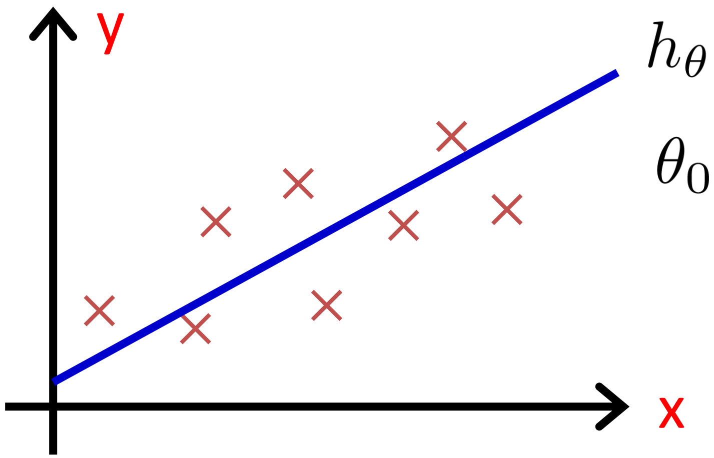
How to Select the Parameters

- How to find the best Parameters θ_i's ?

Answer: We should find the **Best Fit** to the training samples. To do that, we have to Choose θ_0, θ_1 so that $h_\theta(x)$ (the blue line) is as close as possible to all y 's (actual red points) for **ALL training examples** (x, y)

Training Set:

Size in feet^2 (*x*)	Price (*y*)
1000	410K
1200	600K
1230	620K
1340	645K
…	…

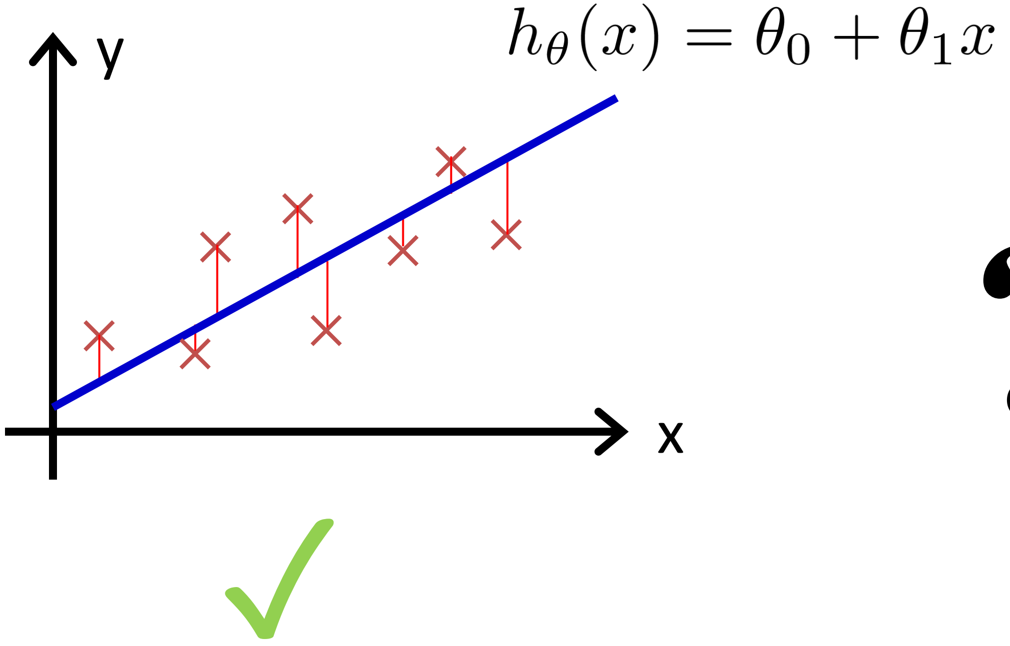


$$h_\theta(x) = \theta_0 + \theta_1 x$$

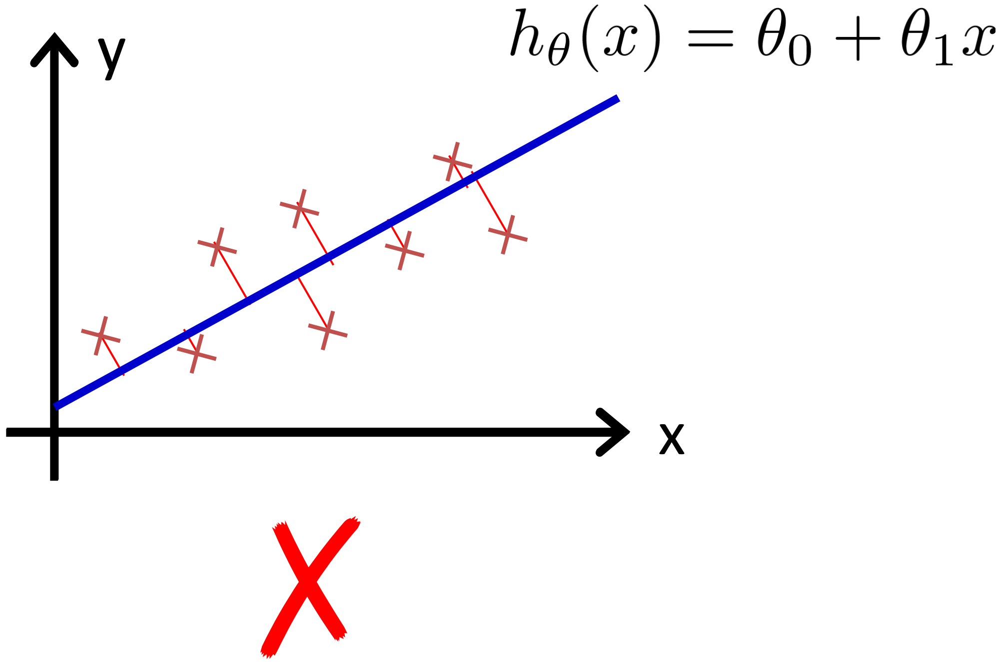
θ_0, θ_1 =?

Which one?

- **Question:** Which **distance** should be minimum to find the **best fit** line? The distance to the line Or the vertical distance?



?

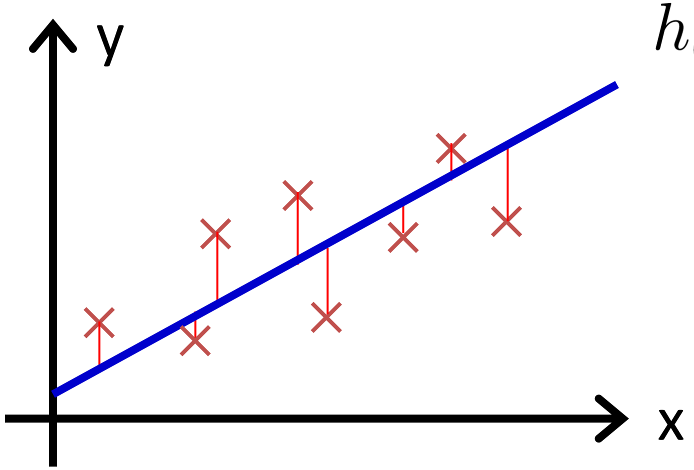


How to Select the Parameters

- **Answer**: Choose θ_0, θ_1 so that $h_\theta(x)$ is as close as possible to y for all training examples (x, y)
- In other word, we have to minimize the differences between "prediction $h(x^{(i)})$ " and "actual value y" in **training set** to find the best "fit" to our training data.

Training Set:

Size in feet^2 (*x*)	Price (*y*)
1000	410K
1200	600K
1230	620K
1340	645K
…	…



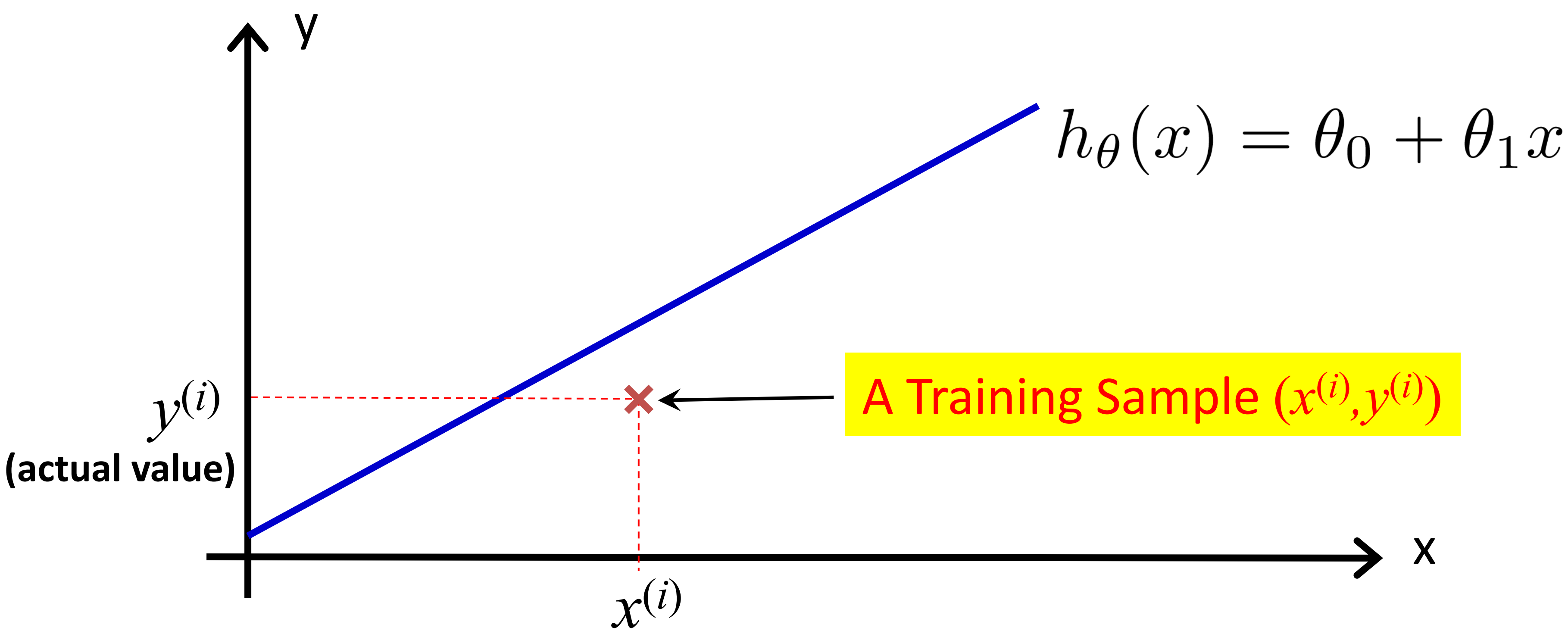
$$h_\theta(x) = \theta_0 + \theta_1 x$$

θ_0, θ_1 =?

How to Select the Parameters

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

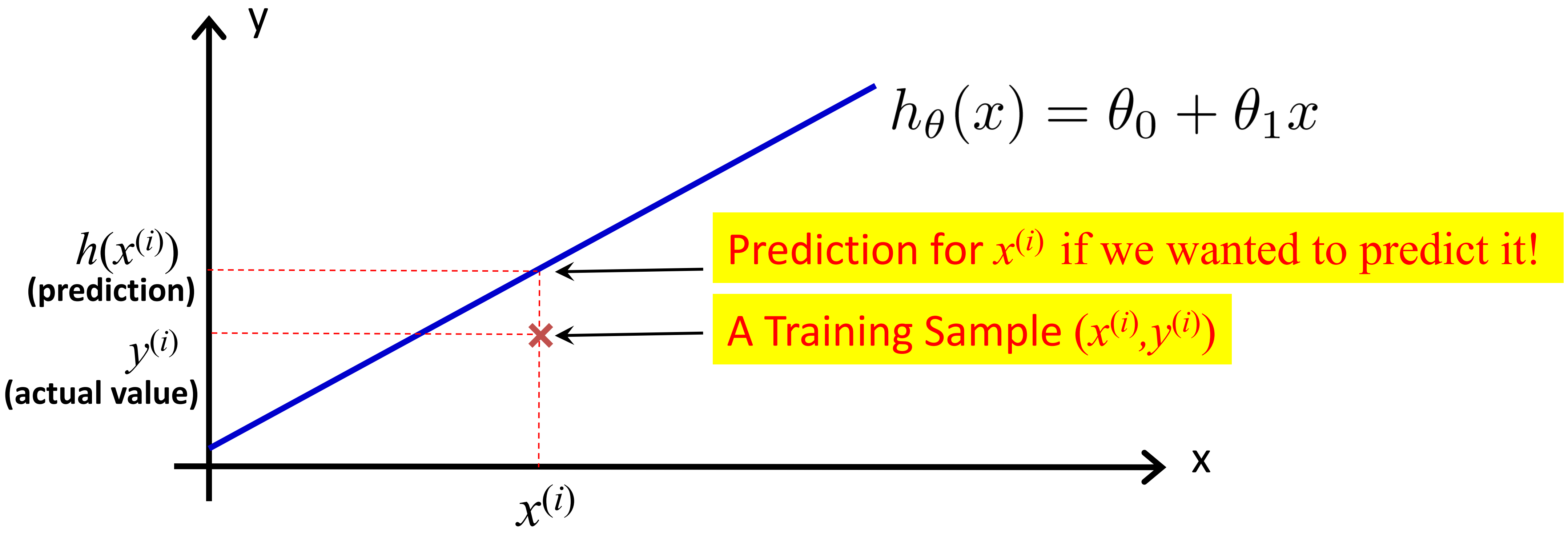
Note: (i) is index!



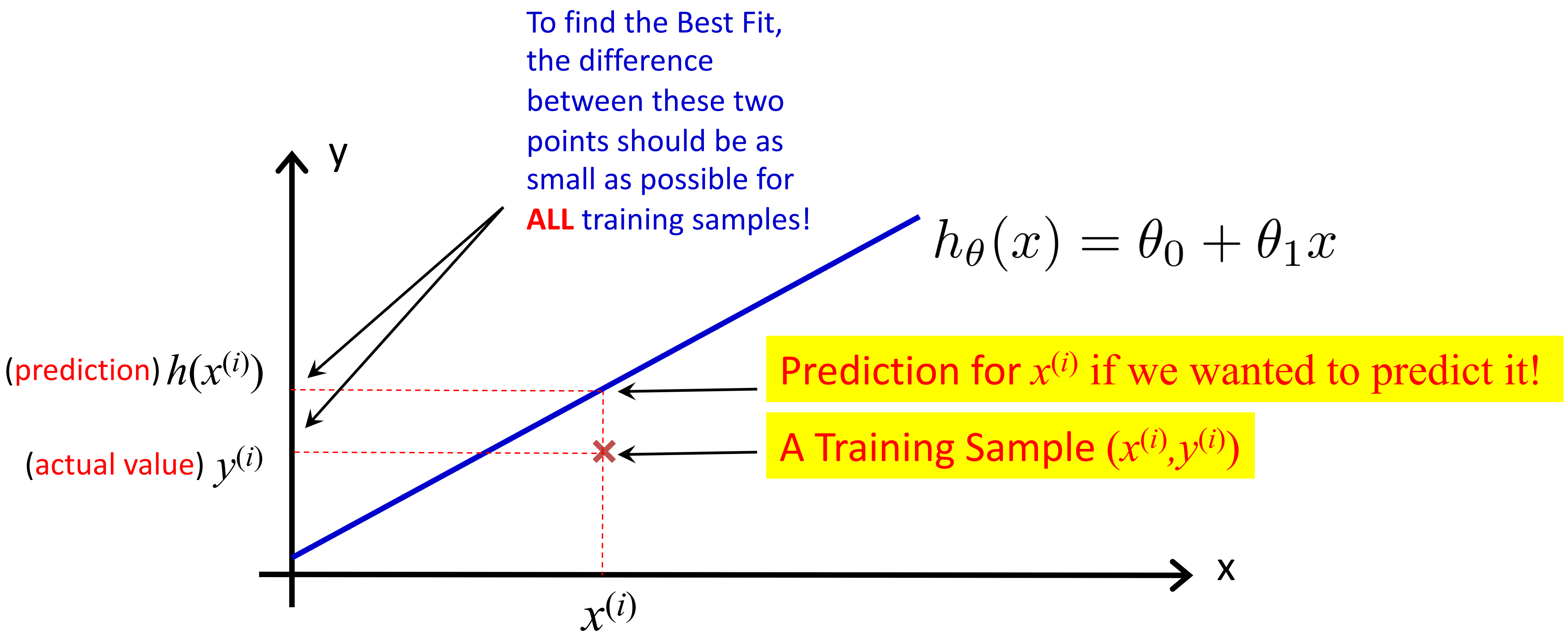
How to Select the Parameters

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

Note: (i) is index!



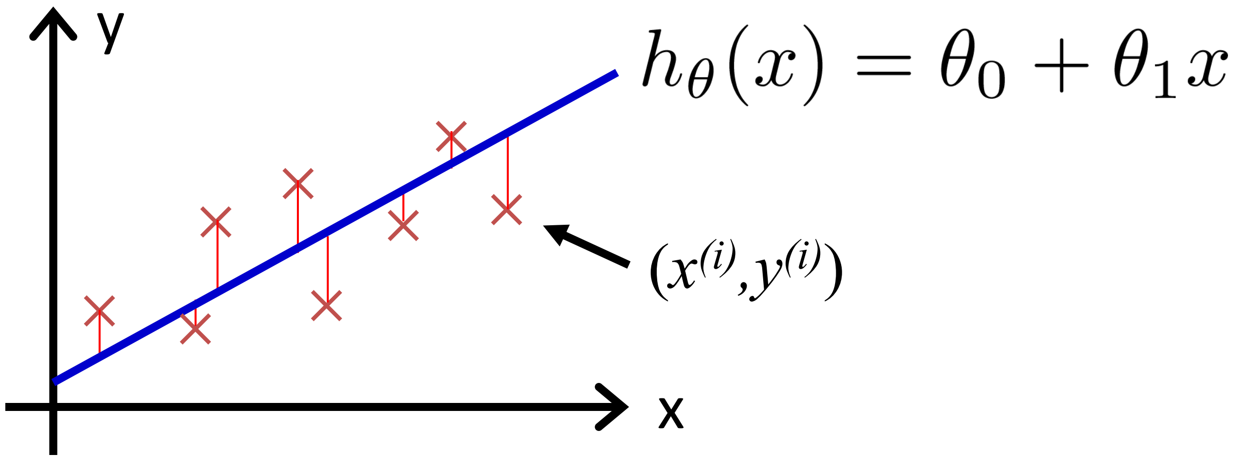
- Thus, we just need to minimize $(h(x^{(i)}) - y^{(i)})^2$ for **every** training sample *i.*
- *Question: Why squared?*



Let's define a **Cost Function** $\boldsymbol{J(\theta_{0,}\ \theta_1)}$ as the summation (or average) of **all** differences between "**training samples**" and the **"regression line"**.

Cost Function: $$J(\theta_0, \theta_1) = \frac{1}{2m}\sum_{i=1}^{m}\left(h_\theta(x^{(i)}) - y^{(i)}\right)^2$$

Thus, the Best Fit Line is the line with minimum Cost Function $\boldsymbol{J(\theta_{0,}\ \theta_1)}$.

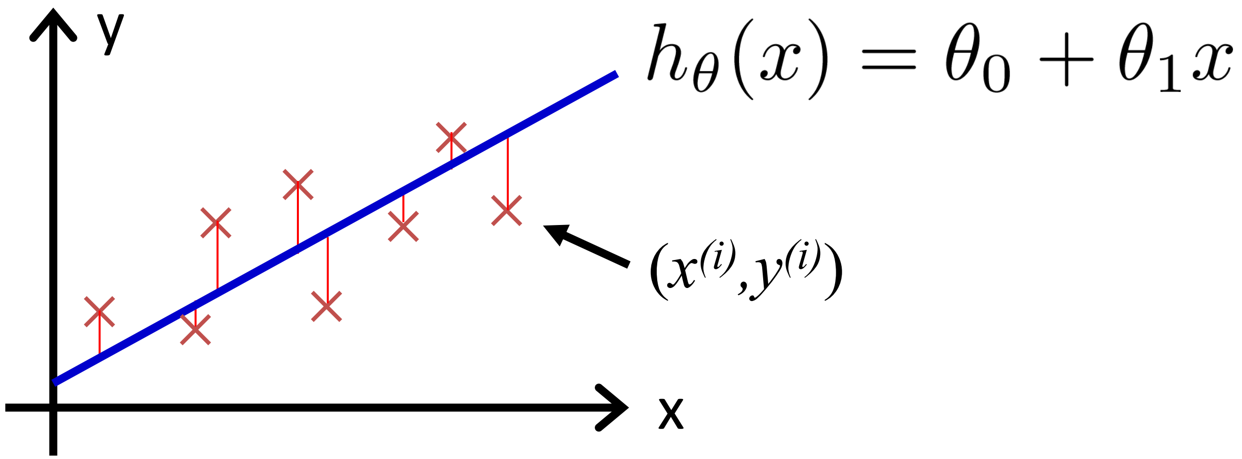


Training Set: $\{(x^{(1)},y^{(1)}), (x^{(2)},y^{(2)}), \ldots, (x^{(m)},y^{(m)})\}$

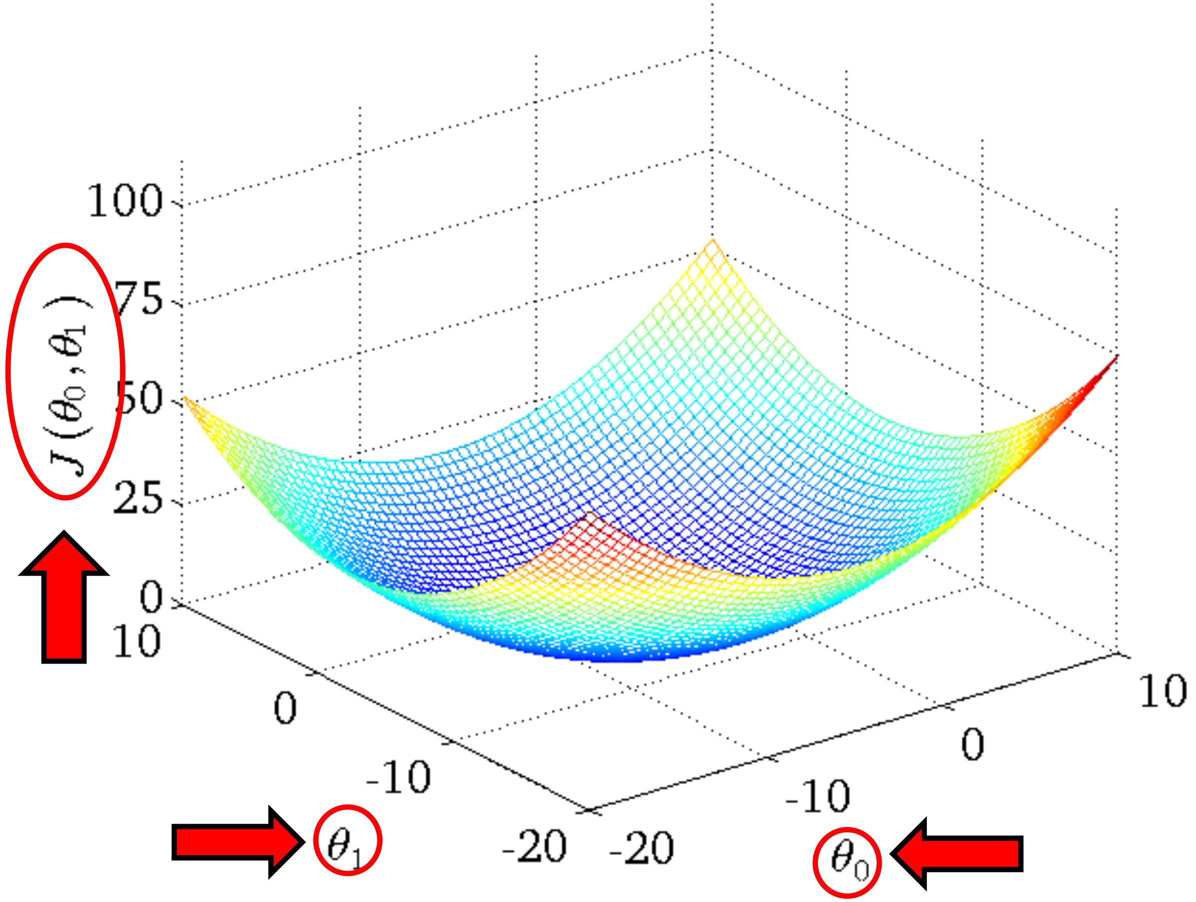
Cost Function: $$J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)}\right)^2$$

Goal: To find the best parameters $\boldsymbol{\theta_0}, \boldsymbol{\theta_1}$ that minimizes the **Cost Function** $\boldsymbol{J(\theta_0, \theta_1)}$:

$$\underset{\theta_0, \theta_1}{\text{minimize}}\ J(\theta_0, \theta_1)$$



- It turns out that the **Cost Function** $J(\theta_0, \theta_1)$ is a Convex (Bowl-Shaped) function. So, it has a global minimum! Example:

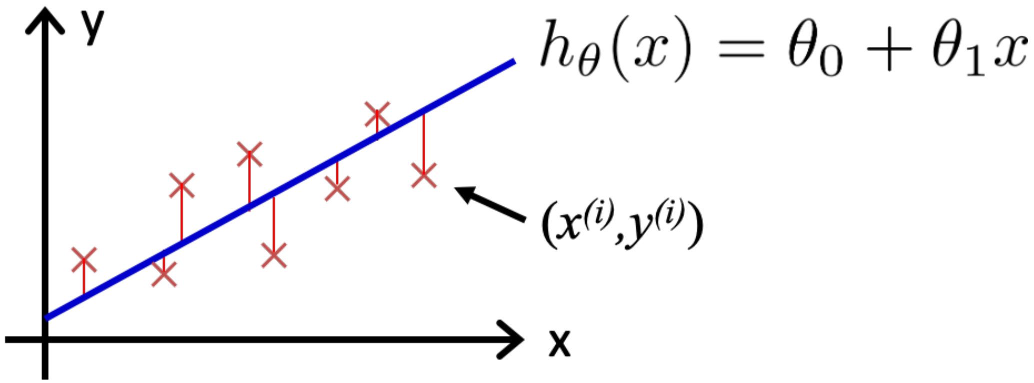


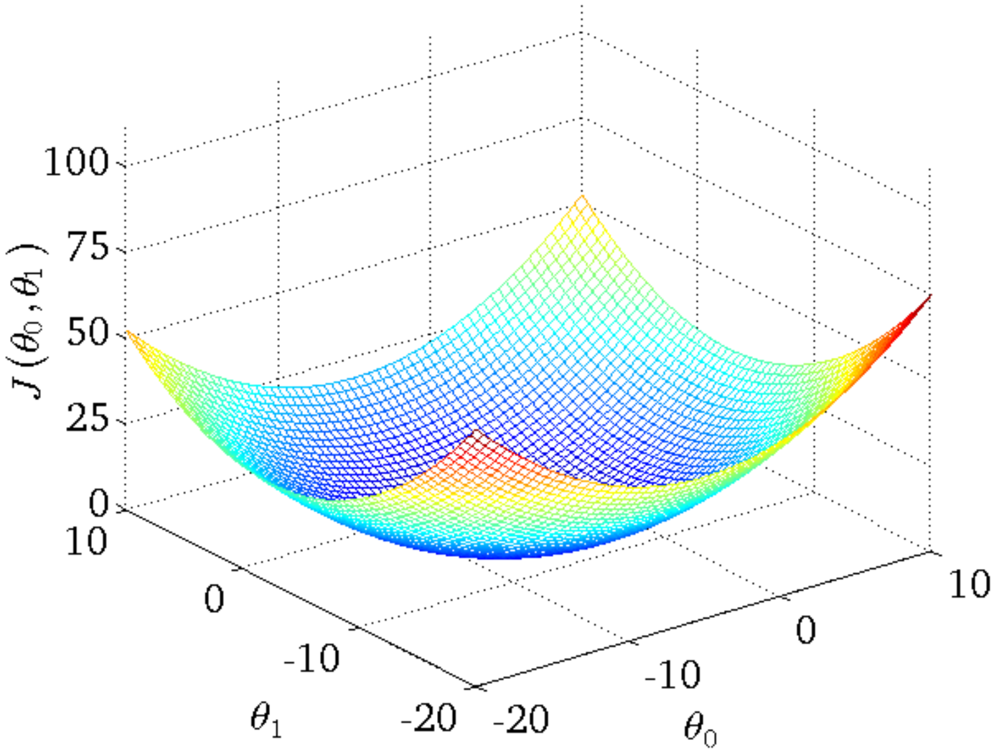
* Reference: Andrew Ng, Machine Learning, Stanford University.

Regression Model: $h_\theta(x) = \theta_0 + \theta_1 x$

Cost Function: $J(\theta_0, \theta_1) = \frac{1}{2m}\sum_{i=1}^{m}\left(h_\theta(x^{(i)}) - y^{(i)}\right)^2$

Goal: $\underset{\theta_0, \theta_1}{\text{minimize}}\ J(\theta_0, \theta_1)$





Thank You!

Questions?