What We Have Learned

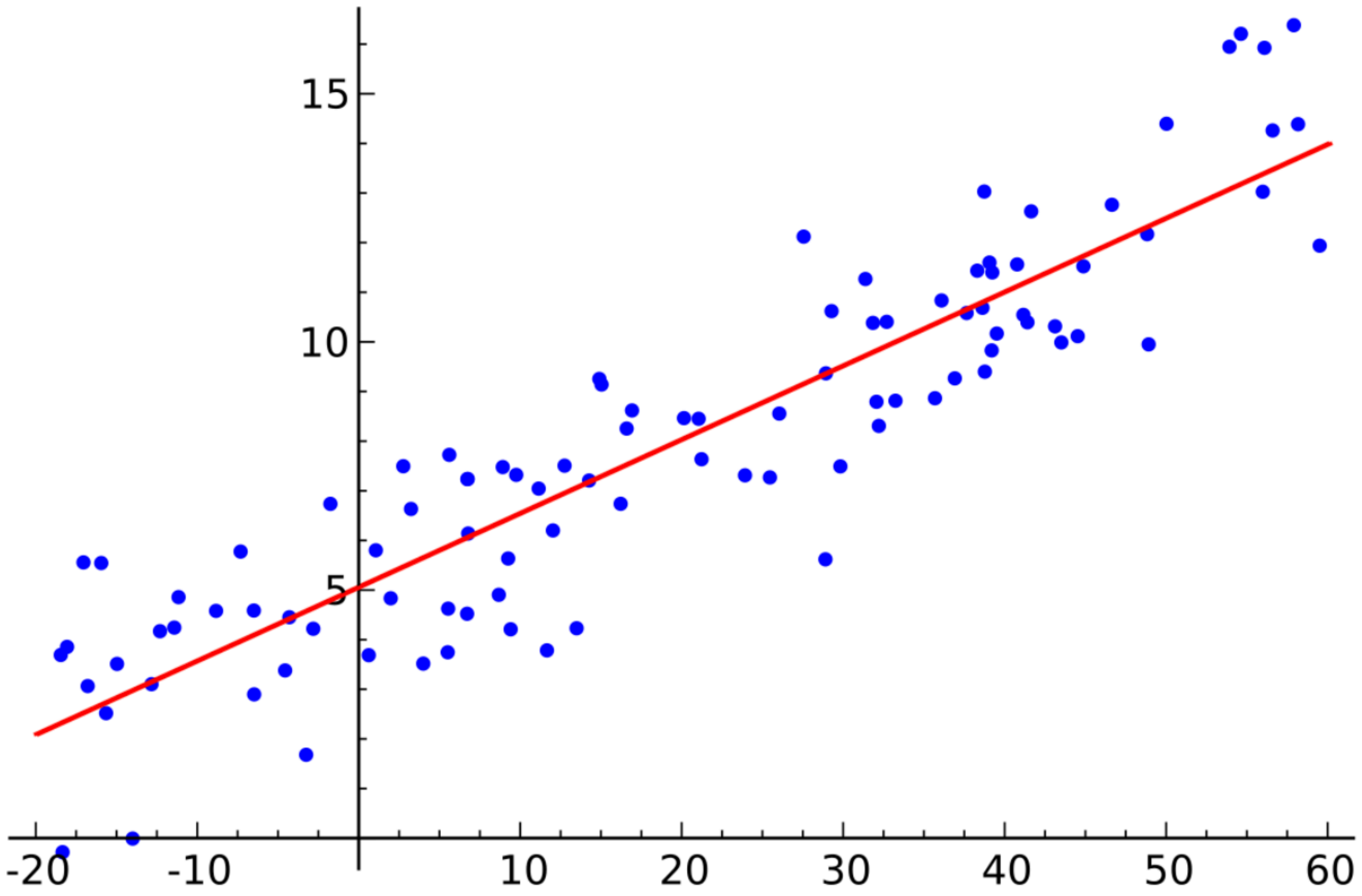
- Predicting a label:
  - K-nearest Neighbors (KNN)
  - Decision Trees
  - Support Vector Machine (SVM)
  - Naive Bayes
  - Logistic Regression
- How about predicting a continuous, numerical value?

Regression

- Regression is a method of modeling a target value base on independent predictors:
  - The target value is also called: dependent variable, in convention, Y.
  - The predictors is also called: independent variable, in convention, X.
- A method is mostly used for forecasting and finding out cause and effect relationship between variables.
- Regression techniques differ based on:
  - Number of predictors, or independent variables.
  - Type of relationship between X and Y.

Linear Regression

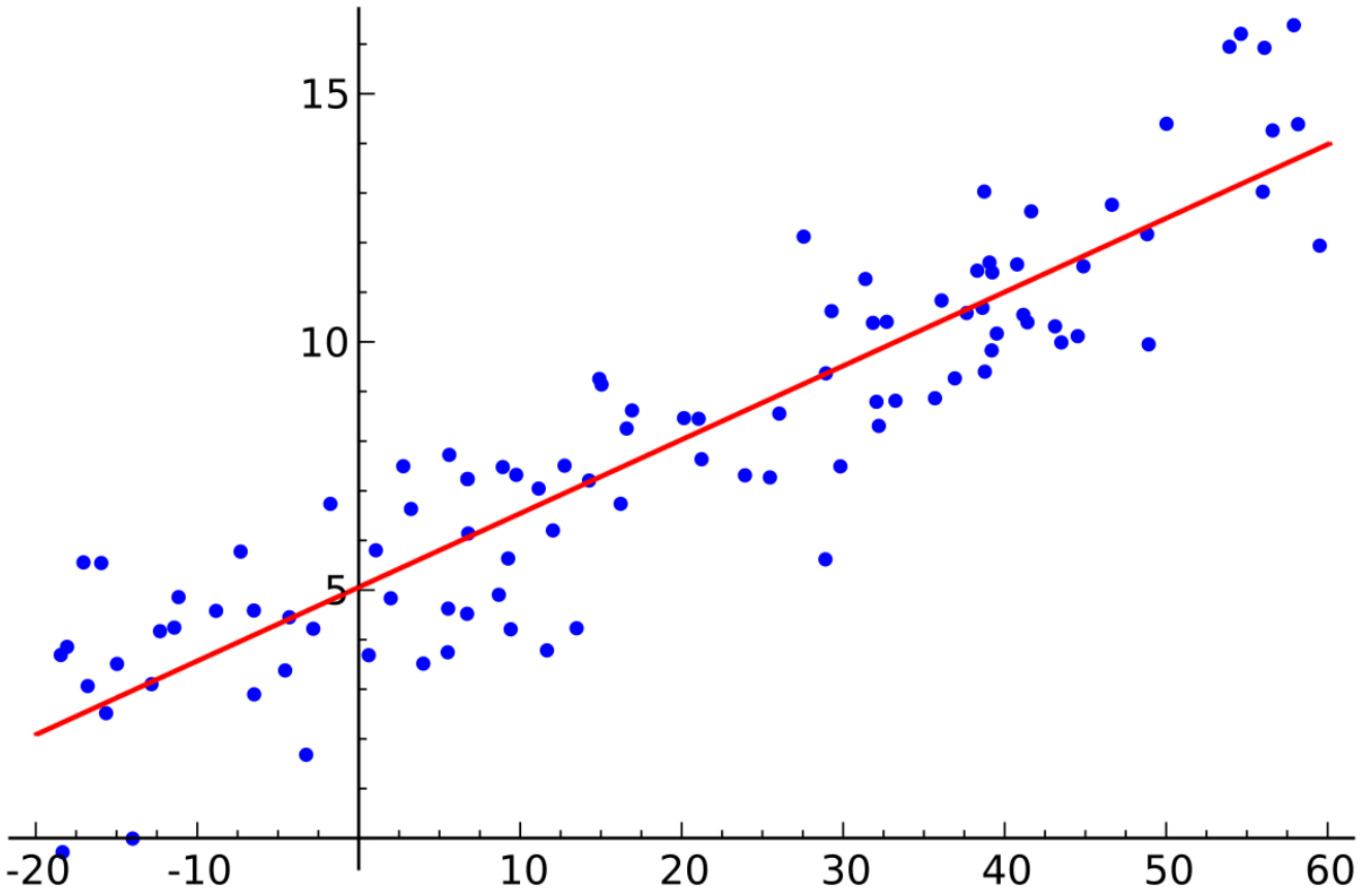
- Assumptions:
  - The true relationship is linear
  - Errors are normally distributed
  - Homoscedasticity of errors (or, equal variance around the line).
  - Independence of the observations
- Don't ignore these assumptions. They are real.



Linear Regression

Simple Linear Regression

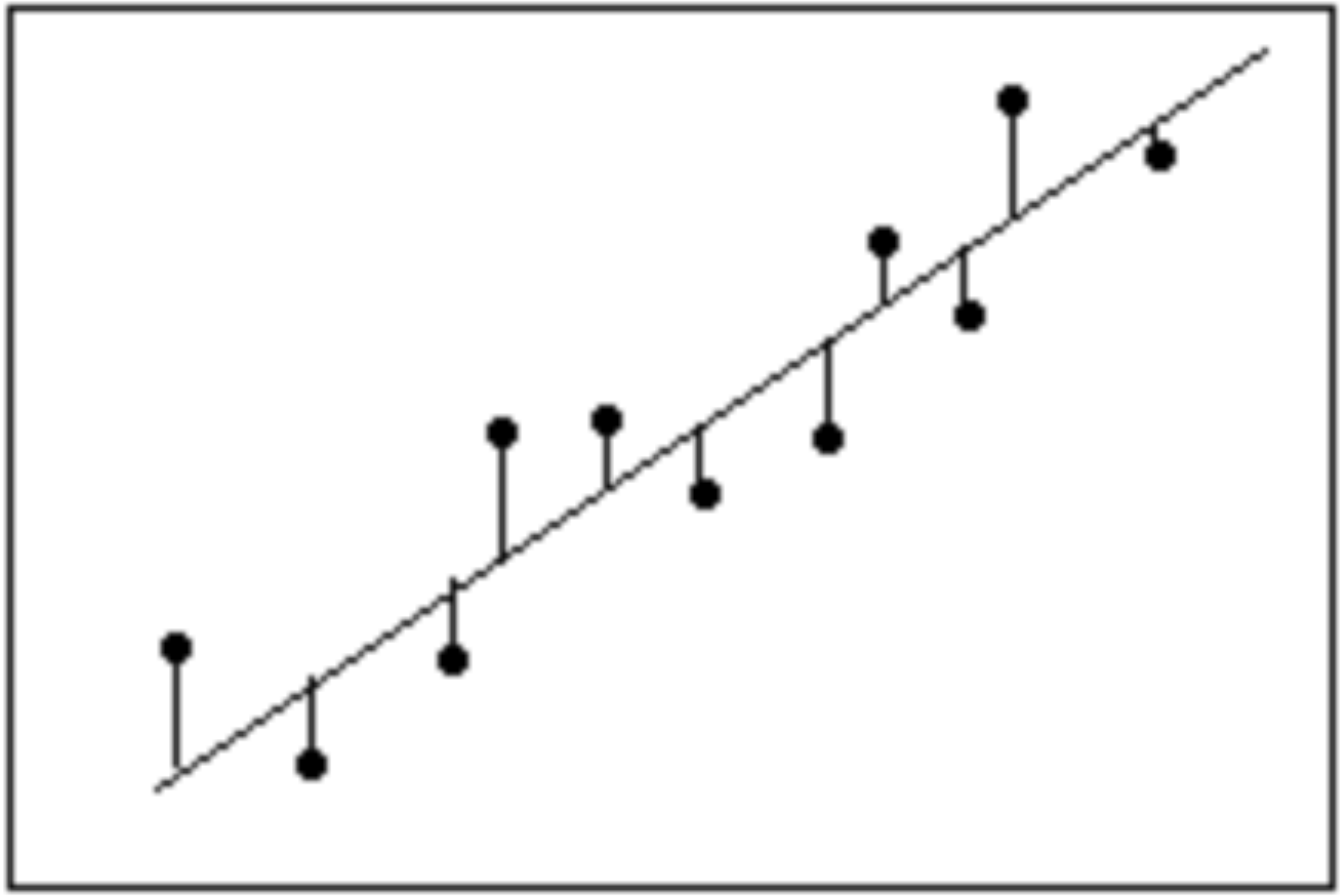
- Simple Linear regression is a type of regression where:
  - Number of independent variables is 1
  - The relationship between X and Y is linear.
- Also called: best fit straight line.
- Linear equation such as:
  - $y = a_0 + a_1 x$



Linear Regression

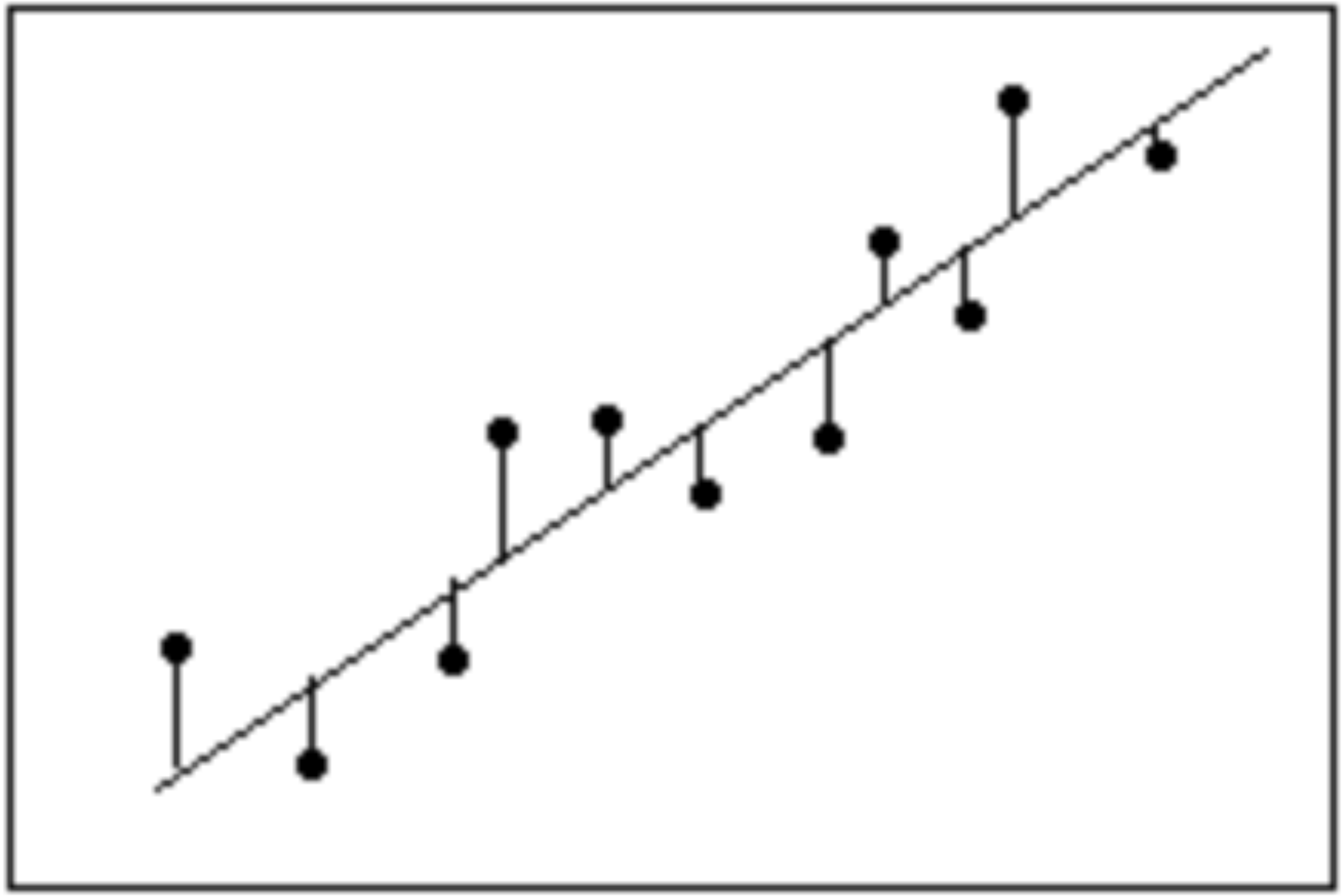
Error Measurement

- What is the error?
  - The difference between prediction and actual value.
  - $e_i = \hat{y}_i - y_i$, where $\hat{y}$ is the prediction



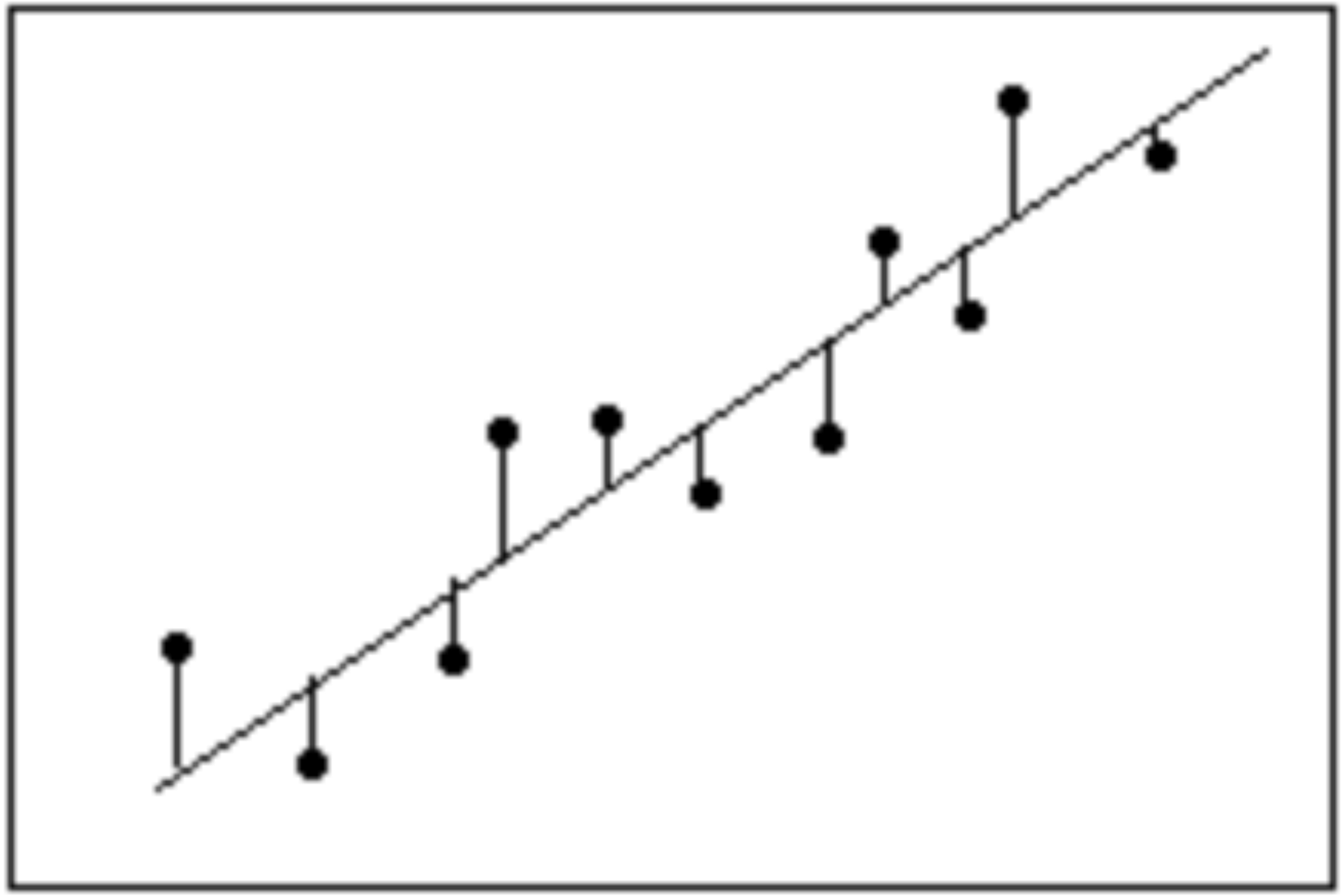
Cost Measurement

- What is the cost?
  - The sum of squared errors, or SSE
    - $\sum_{1}^{n} e_i^2 = \sum_{1}^{n} (\hat{y}_i - y_i)^2$
  - The mean of squared errors, or MSE
    - $\frac{1}{n}\sum_{1}^{n} e_i^2 = \frac{1}{n}\sum_{1}^{n} (\hat{y}_i - y_i)^2$



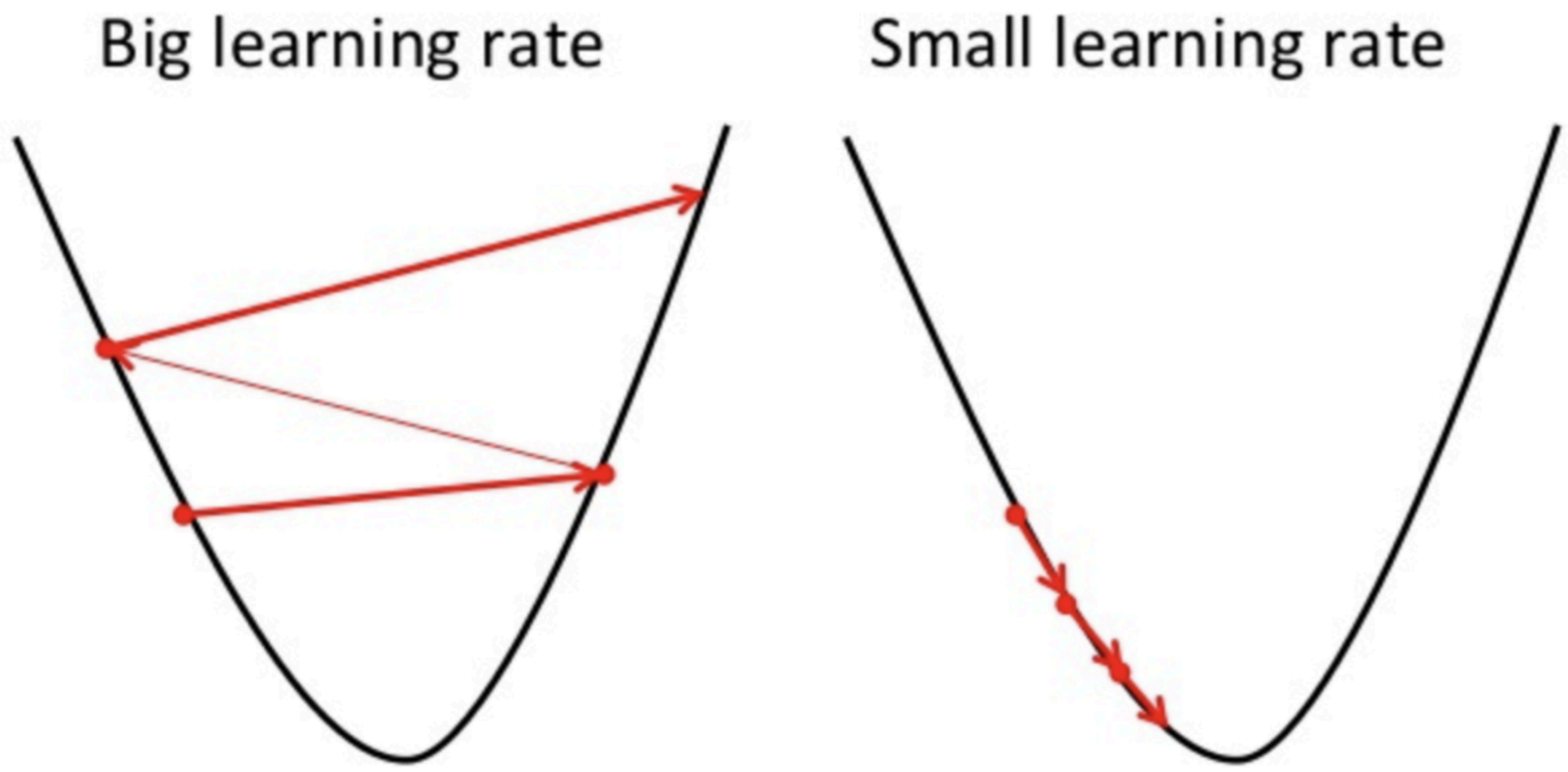
Optimization

- Optimization goal?
  - Minimize the cost. Which cost?
  - Minimize SE, LSSE: Least Sum of Squared Error, also called Ordinary Least Squared (OLS)
    - $min(\sum_{1}^{n}(\hat{y}_i - y_i)^2)$
  - Minimize MSE, MMSE: Minimal Mean of Squared Error
    - $min(\frac{1}{n}\sum_{1}^{n}(\hat{y}_i - y_i)^2)$
  - Why MSE is better for evaluation than SSE?



How to Reduce Cost?

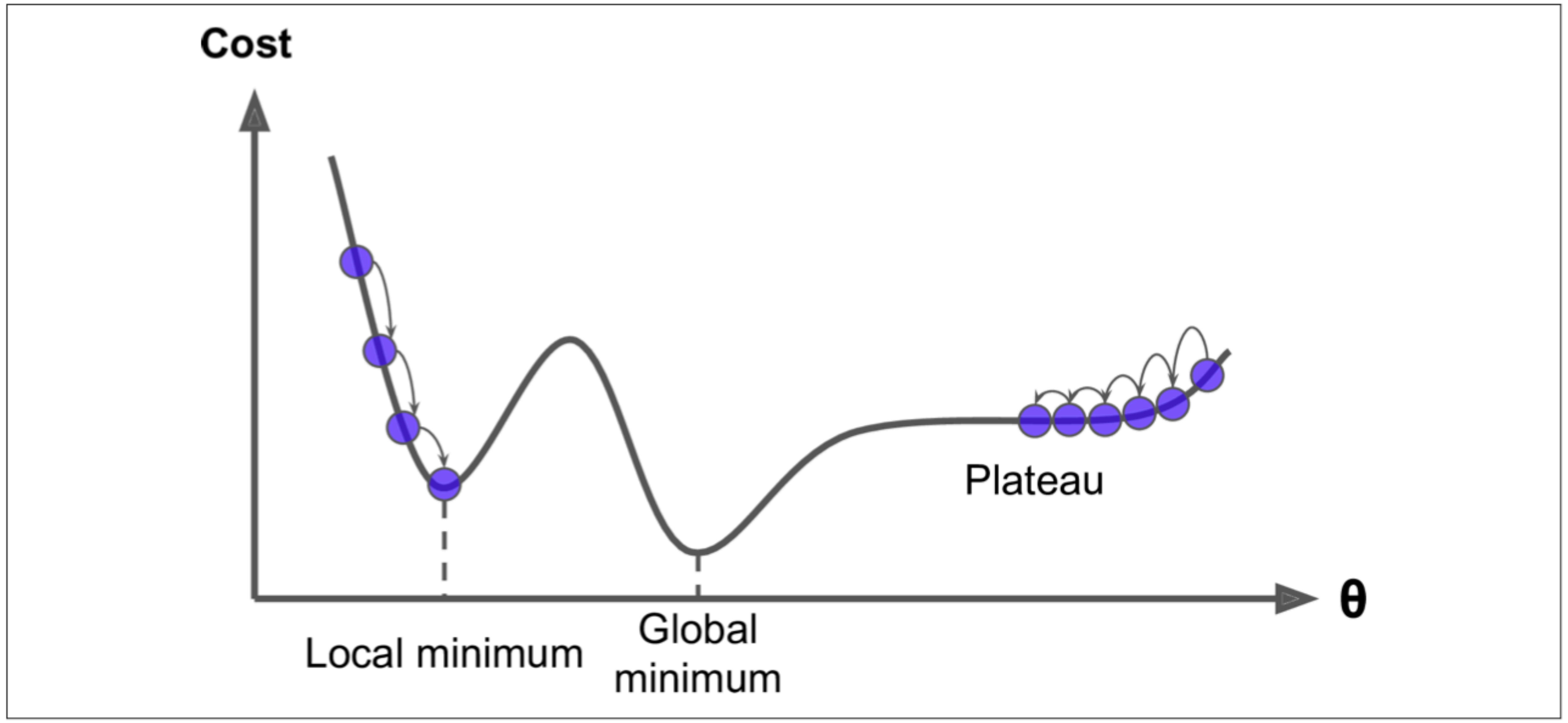
- Gradient Descent
  - Updating the coefficients of the line to reduce the cost
  - Use the knowledge of partial derivatives to know the direction to change
  - Decide how much to change
    - Learning rate: Big or Small?



Gradient Descent

Gradient Descent

- Problems:
  - Local VS Global
  - Plateau
- Scale of attribute:
  - All features must have similar scales
  - Normalization
  - StandardScaler from Scikit-Learn
- SGD:
  - Training data must be independent and identically distributed



Examples

- Diabetes dataset from Scikit-learn.
  - A toy dataset with 442 patients
  - 10 numerical independent variables to be used
  - 1 dependent numerical variable to be predicted
- Page: Diabetes Dataset

7.1.3. Diabetes dataset

Ten baseline variables, age, sex, body mass index, average blood pressure, and six blood serum measurements were obtained for each of n = 442 diabetes patients, as well as the response of interest, a quantitative measure of disease progression one year after baseline.

Data Set Characteristics:

Number of Instances:	442
Number of Attributes:	First 10 columns are numeric predictive values
Target:	Column 11 is a quantitative measure of disease progression one year after baseline
Attribute Information:	• age age in years • sex • bmi body mass index • bp average blood pressure • s1 tc, total serum cholesterol • s2 ldl, low-density lipoproteins • s3 hdl, high-density lipoproteins • s4 tch, total cholesterol / HDL • s5 ltg, possibly log of serum triglycerides level • s6 glu, blood sugar level