



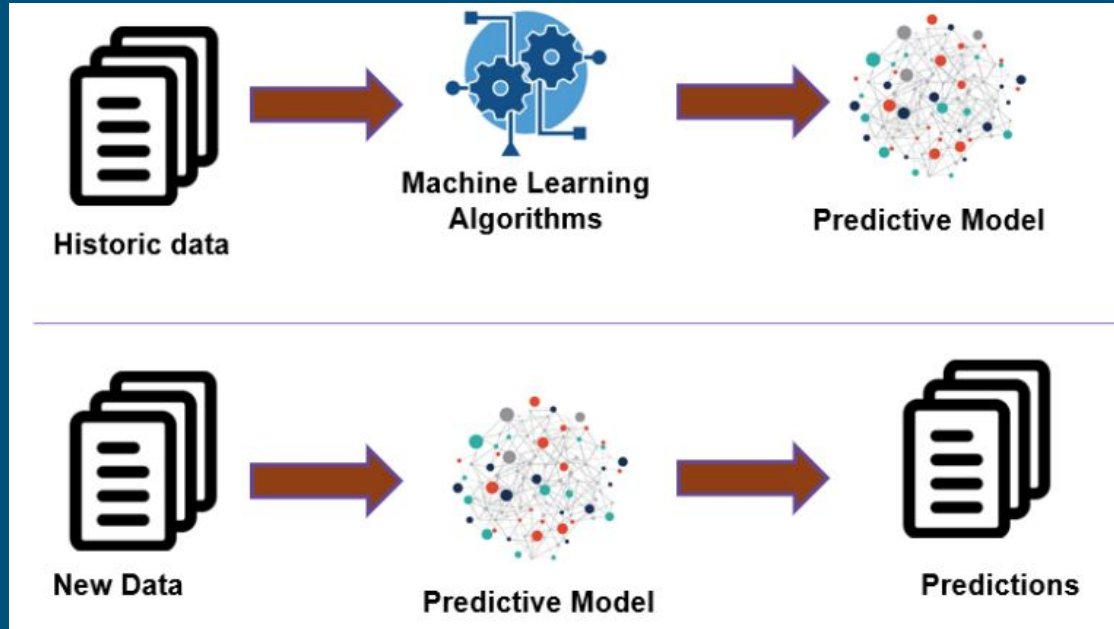
# Linear Regression

---

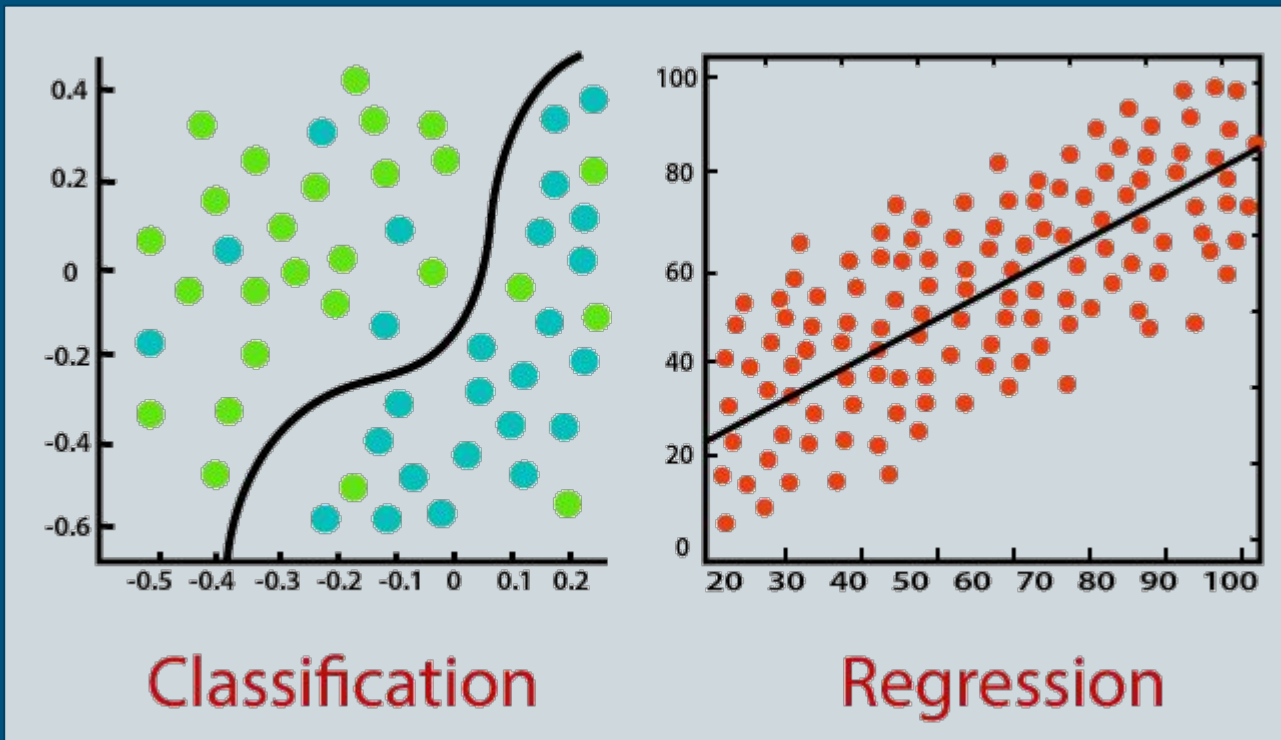
Dr. Dongchul Kim



# Prediction



# Regression vs Classification



# Regression vs Classification



## Regression

What is the temperature going to be tomorrow?

PREDICTION

84°



## Classification

Will it be Cold or Hot tomorrow?

PREDICTION

COLD

HOT



# Data (Train and Test)

---

$X$

	$x_1$	$x_2$	...	$x_{m-1}$	$x_m$
1					
2					
.					
.					
n					

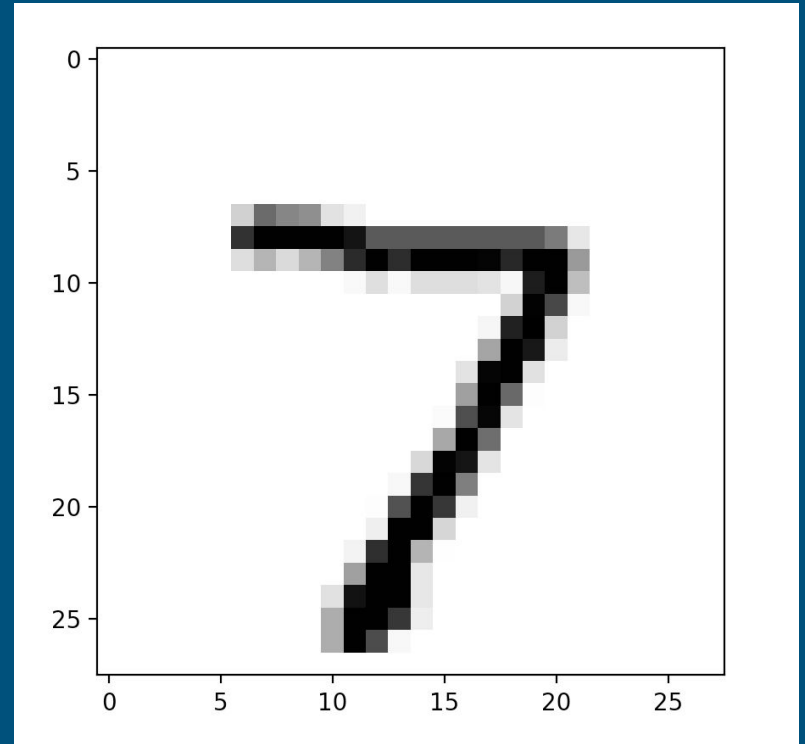
$y$

$y$



# MNIST data

- 60,000 train sample
- 10,000 test sample.
- 28 x 28 pixels (784 features)
- <http://yann.lecun.com/exdb/mnist/>



# THE MNIST DATABASE

## of handwritten digits

[Yann LeCun](#), Courant Institute, NYU

[Corinna Cortes](#), Google Labs, New York

[Christopher J.C. Burges](#), Microsoft Research, Redmond

The MNIST database of handwritten digits, available from this page, has a training set of 60,000 examples, and a test set of 10,000 examples. It is a subset of a larger set available from NIST. The digits have been size-normalized and centered in a fixed-size image.

It is a good database for people who want to try learning techniques and pattern recognition methods on real-world data while spending minimal efforts on preprocessing and formatting.

Four files are available on this site:

[train-images-idx3-ubyte.gz](#): training set images (9912422 bytes)


[train-labels-idx1-ubyte.gz](#): training set labels (28881 bytes)

[t10k-images-idx3-ubyte.gz](#): test set images (1648877 bytes)

[t10k-labels-idx1-ubyte.gz](#): test set labels (4542 bytes)



# Iris (Classification)



## Iris

Donated on 6/30/1988

A small classic dataset from Fisher, 1936. One of the earliest known datasets used for evaluating classification methods.

Dataset Characteristics	Subject Area	Associated Tasks
Tabular	Biology	Classification
Feature Type	# Instances	# Features
Real	150	4

### Dataset Information

**What do the instances in this dataset represent?**  
Each instance is a plant

**Additional Information**  
This is one of the earliest datasets used in the literature on classification methods and widely used in statistics and machine learning. The data set contains 3 classes of 50 instances each, where each class refers to a type of iris plant. One class is linearly separable from the other 2; the latter are not linearly separable from each other...


**SHOW MORE** ▾

**Has Missing Values?**  
No

### Introductory Paper

[The Iris data set: In search of the source of virginica](#)  
By A. Unwin, K. Kleinman. 2021  
Published in Significance, 2021

**DOWNLOAD**

 **IMPORT IN PYTHON**


**CITE**

**352 citations**  
**617968 views**

### Keywords

ecology

### Creators

 **R. A. Fisher**

### DOI

10.24432/C56C76

### License

This dataset is licensed under a **Creative Commons Attribution 4.0 International (CC BY 4.0)** license.


This allows for the sharing and adaptation of the datasets for any purpose, provided that the appropriate credit is given.

# Iris

---

```
5.1,3.5,1.4,0.2,Iris-setosa  
4.9,3.0,1.4,0.2,Iris-setosa  
4.7,3.2,1.3,0.2,Iris-setosa  
4.6,3.1,1.5,0.2,Iris-setosa  
5.0,3.6,1.4,0.2,Iris-setosa  
5.4,3.9,1.7,0.4,Iris-setosa  
4.6,3.4,1.4,0.3,Iris-setosa  
5.0,3.4,1.5,0.2,Iris-setosa  
4.4,2.9,1.4,0.2,Iris-setosa  
4.9,3.1,1.5,0.1,Iris-setosa  
5.4,3.7,1.5,0.2,Iris-setosa  
4.8,3.4,1.6,0.2,Iris-setosa  
4.8,3.0,1.4,0.1,Iris-setosa  
4.3,3.0,1.1,0.1,Iris-setosa
```

# Auto MPG (Regression)



## Auto MPG


Donated on 7/6/1993

Revised from CMU StatLib library, data concerns city-cycle fuel consumption



Dataset Characteristics	Subject Area	Associated Tasks
Multivariate	Other	Regression

Feature Type	# Instances	# Features
Real, Categorical, Integer	398	7

**DOWNLOAD**

 **IMPORT IN PYTHON**


**CITE**

 15 citations  
 96143 views

### Keywords

automobile

### Creators

 R. Quinlan

### DOI

10.24432/C5859H

### License

This dataset is licensed under a [Creative Commons Attribution 4.0 International](#) (CC BY 4.0) license.

This allows for the sharing and adaptation of the datasets for any purpose, provided that the appropriate credit is given.

### Dataset Information

**Additional Information**  
This dataset is a slightly modified version of the dataset provided in the StatLib library. In line with the use by Ross Quinlan (1993) in predicting the attribute "mpg", 8 of the original instances were removed because they had unknown values for the "mpg" attribute. The original dataset is available in the file "auto-mpg.data-original"....

**SHOW MORE** ▾

**Has Missing Values?**  
Yes

### Variables Table

Variable Name	Role	Type	Demographic	Description	Units	Missing Values
displacement	Feature	Continuous				no

# Auto MPG

1	18.0	8	307.0	130.0	3504.	12.0	70	1	"chevrolet chevelle malibu"
2	15.0	8	350.0	165.0	3693.	11.5	70	1	"buick skylark 320"
3	18.0	8	318.0	150.0	3436.	11.0	70	1	"plymouth satellite"
4	16.0	8	304.0	150.0	3433.	12.0	70	1	"amc rebel sst"
5	17.0	8	302.0	140.0	3449.	10.5	70	1	"ford torino"
6	15.0	8	429.0	198.0	4341.	10.0	70	1	"ford <u>galaxie</u> 500"
7	14.0	8	454.0	220.0	4354.	9.0	70	1	"chevrolet impala"
8	14.0	8	440.0	215.0	4312.	8.5	70	1	"plymouth fury iii"
9	14.0	8	455.0	225.0	4425.	10.0	70	1	"pontiac catalina"
10	15.0	8	390.0	190.0	3850.	8.5	70	1	"amc ambassador dpl"

# Linear Regression

---

Numerous **factors** contribute to the fluctuation of house prices, including the year of construction, location, and number of rooms. By leveraging relevant information associated with these factors, it becomes feasible to forecast house prices.

Let us denote the information that influences house prices as "**x**" and the corresponding house price as "**y**." In this context, "**x**" represents the **independent variable**, while "**y**," being contingent on the value of "**x**," is referred to as the **dependent variable**.

# Linear Regression

---

Linear regression entails the prediction of the dependent variable using the independent variable. In cases where a single independent variable, denoted as "x," fails to provide a comprehensive explanation alone, multiple independent variables such as "x1," "x2," and "x3" can be employed.

The relationship between the independent variable and the dependent variable can be expressed through a linear function:

$$y = ax + b$$

In this equation, "x" signifies the independent variable, while "y" represents the dependent variable. Consequently, the value of "y" varies depending on the value of "x." However, to achieve precise calculations, it is necessary to ascertain the values of "a" and "b."

# Linear Regression

---

- With the knowledge of "**a**" and "**b**," we can reliably determine the value of "y" given "x." The formula mentioned above, recognized as the linear formula, is the foundation of linear regression.
- Linear regression finds extensive practical applications, primarily falling into two broad categories:
  - a. **Prediction and Forecasting**
  - b. **Explanation of Variation in the Response Variable**

# Prediction and Forecasting

---

Linear regression enables the fitting of a predictive model to an observed dataset comprising values of both the response and explanatory variables.

Once such a model is developed, it can be employed to make **predictions** for the response variable when additional values of the explanatory variables are available, even without accompanying response values.

This application is particularly useful for prediction and forecasting.



# Explanation of Variation in the Response Variable

---

Linear regression analysis can be applied to elucidate the extent to which variation in the response variable can be attributed to fluctuations in the explanatory variables.

It quantifies the strength of the **relationship** between the response and explanatory variables. Furthermore, linear regression helps identify if certain explanatory variables lack a linear relationship with the response altogether or determine which subsets of explanatory variables contain redundant information about the response.

This application aims to provide insights into the factors driving variation in the response variable.

# Example: Predict your exam score

---

First, let's look at an example of a simple linear regression with only one independent variable.

$X$

	$x_1$
1	
2	
.	
.	
n	

$y$

$y$

# Example: Predict your exam score

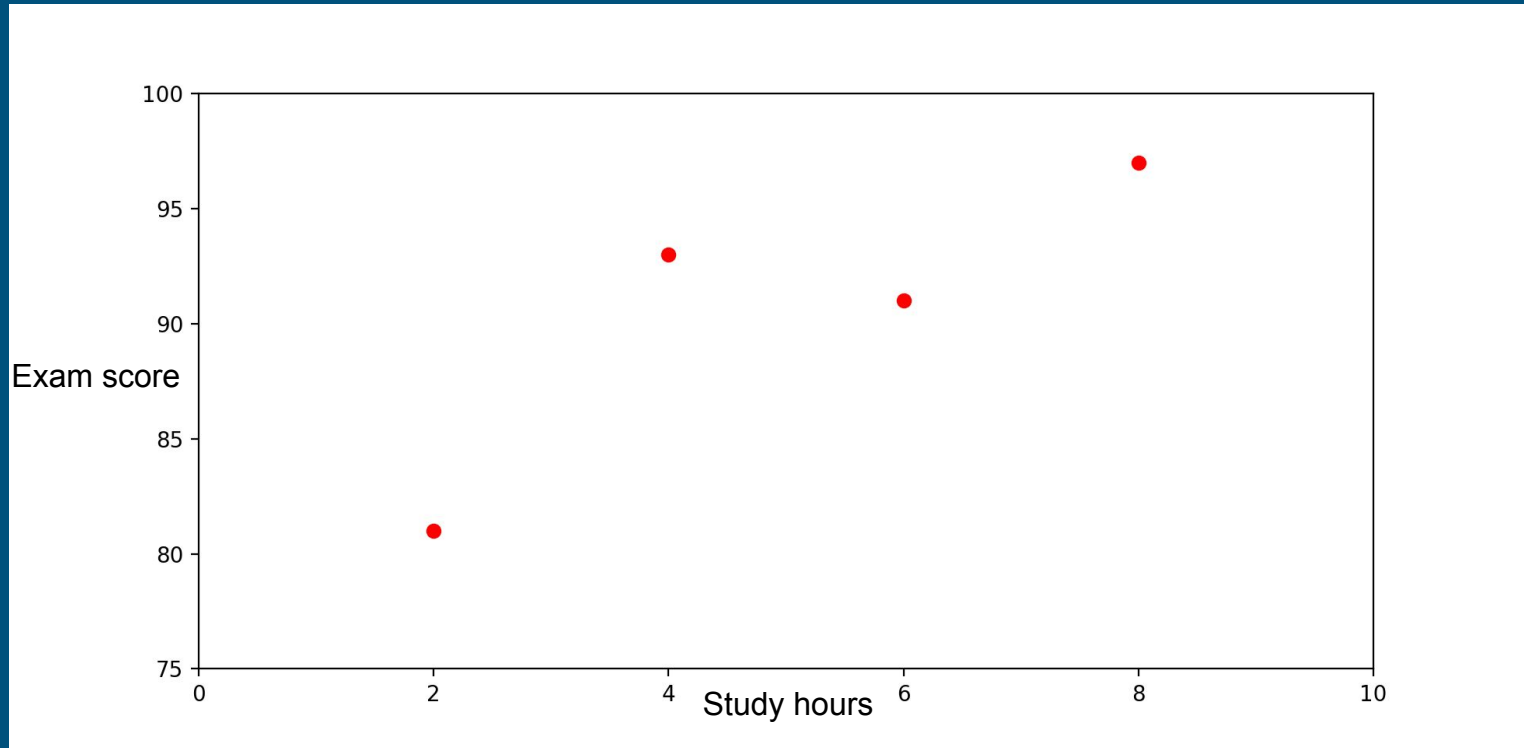
---

Of the many factors that determine your score, consider only the time you study.

Study hours	Score
2 hours	81
4 hours	93
6 hours	91
8 hours	97

# Example: Predict your exam score

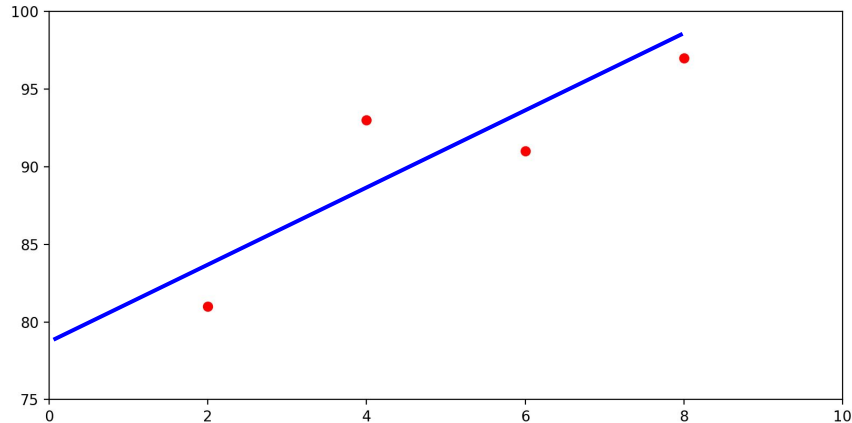
---



# Example: Predict your exam score

---

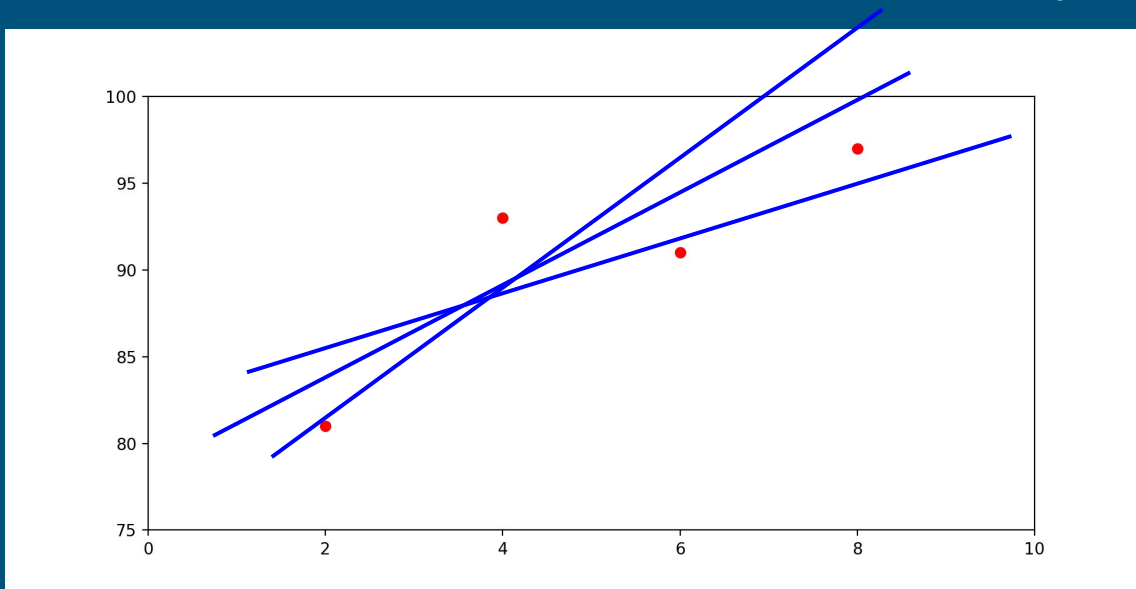
We need a model to represent this phenomena/event/data. Intuitively, we can observe that the data seems to be **linear** with the left side down and the right side up. Therefore, a linear function will be the best model to represent the data.



# Example: Predict your exam score

---

However, we are not sure which one has the best fit to the data yet.



# Example: Predict your exam score

---

Linear Function:

$$y = ax + b$$

**y** represents the score.

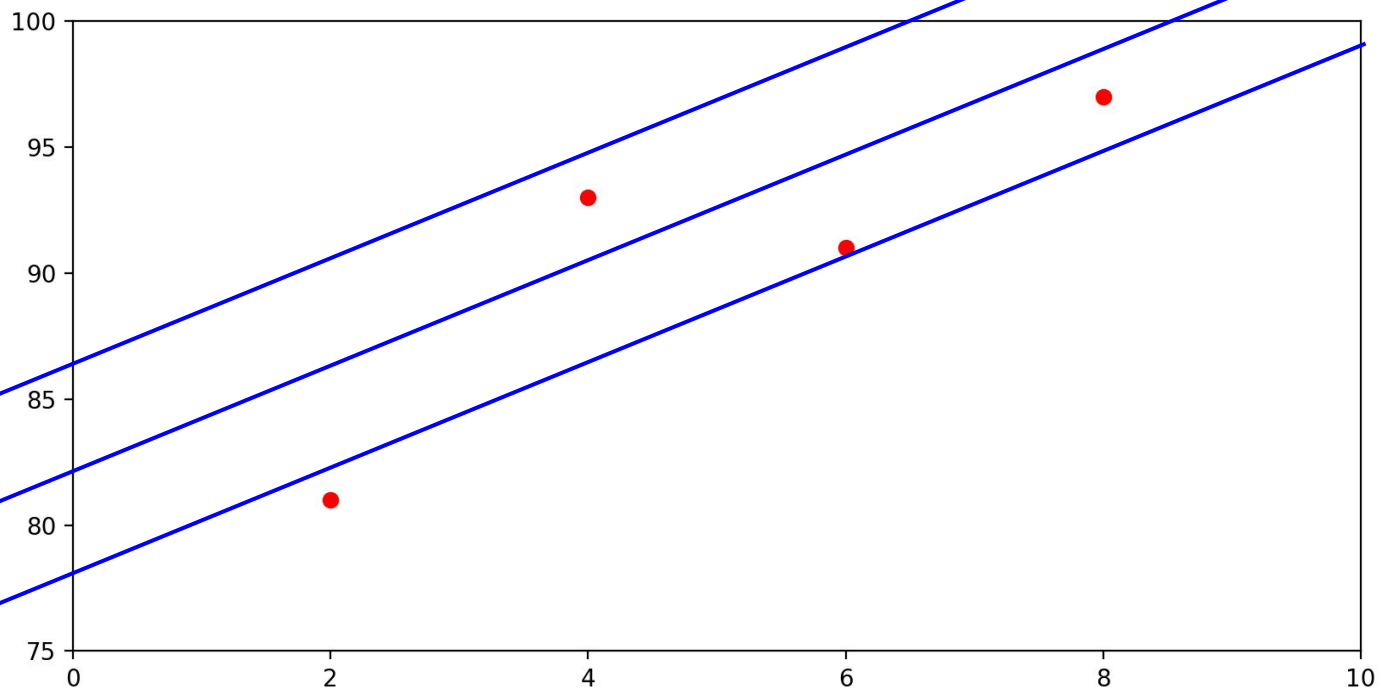
**x** represents the study time.

**a**: The slope of the linear function.

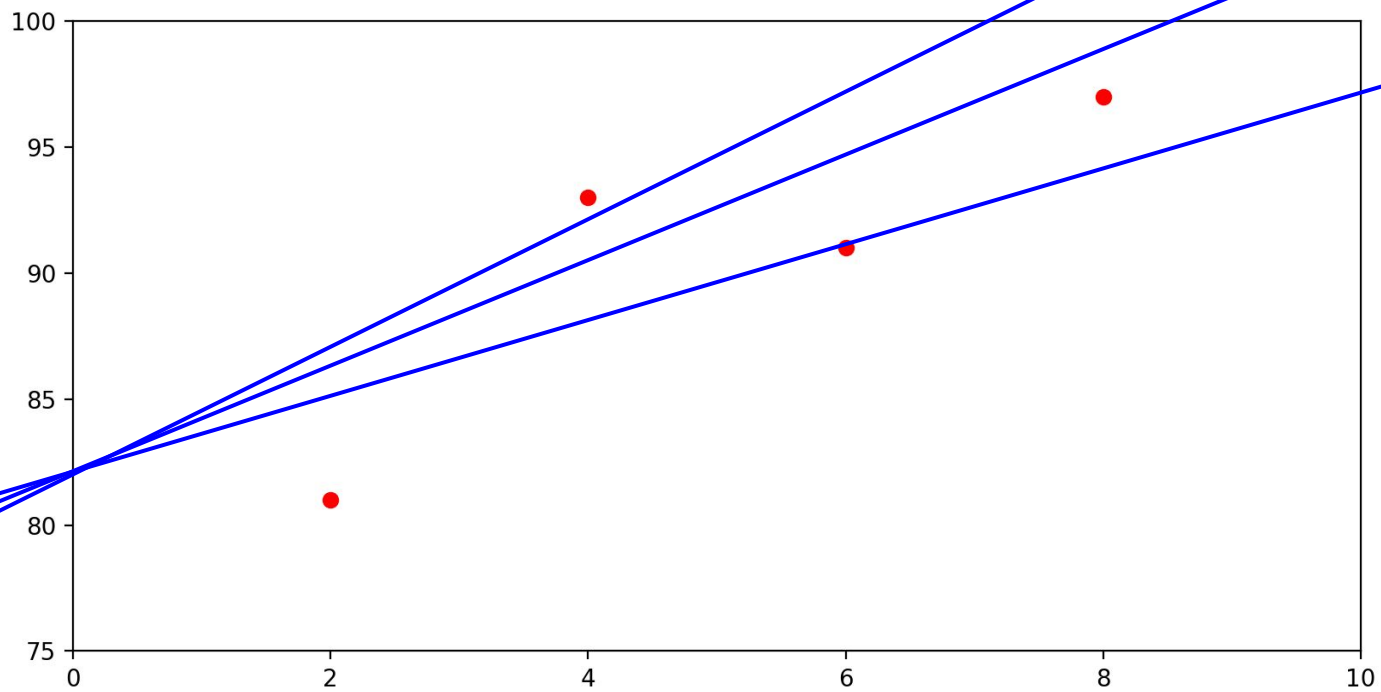
**b**: The y-intercept of the linear function.

The values of "a" and "b" play a crucial role in determining the quality of our model's fit to the data. They dictate the relationship between the study time ("x") and the resulting score ("y"). The slope ("a") indicates the rate of change in the score for each unit increase in study time. The y-intercept ("b") represents the score value when the study time is zero.

Essentially, by adjusting the values of "a" and "b" in the linear function, we can optimize our model's ability to capture the relationship between study time and scores, thereby enhancing its predictive capabilities.







# How to choose the best/optimal $a$ and $b$

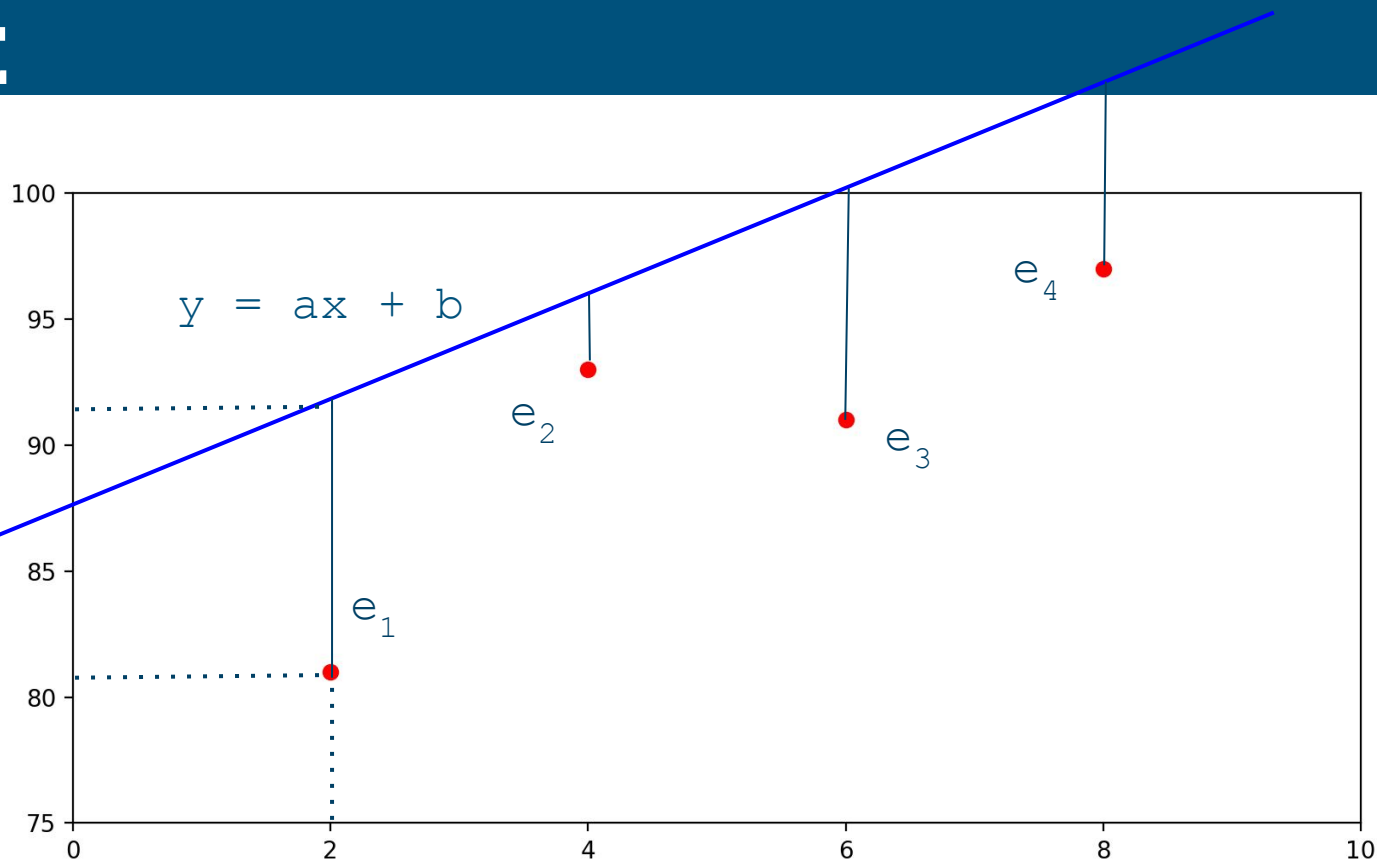
---

## Line Evaluation and Error Minimization:

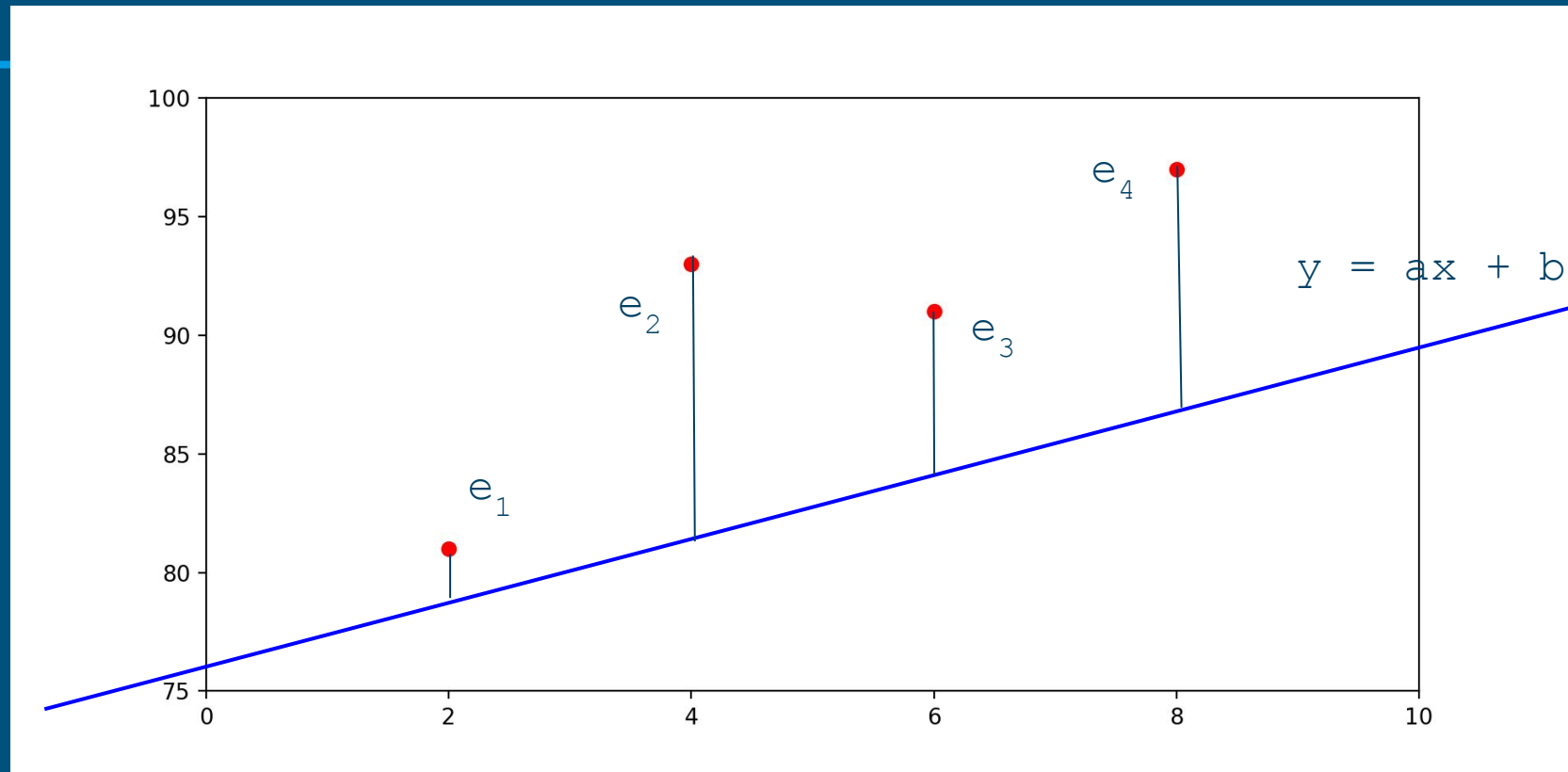
When drawing a line, it is essential to assess its accuracy. This evaluation involves determining the error associated with the line. To achieve this, we utilize the Mean Squared Error (MSE) method. By employing MSE, we iteratively refine the lines to minimize the error.

The algorithm's objective is to continuously search for lines that exhibit smaller errors (MSE), enabling the creation of increasingly precise models.

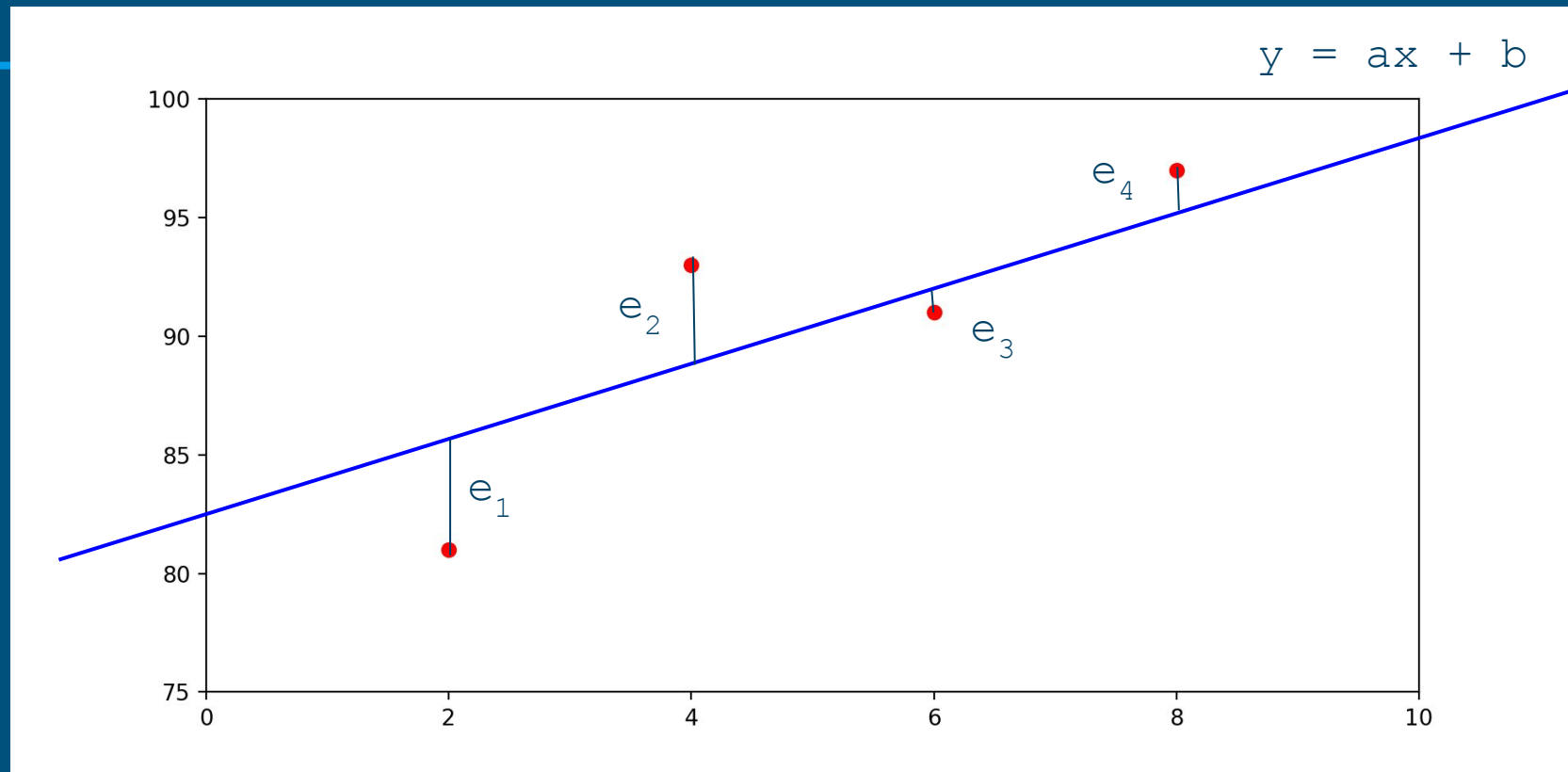
# MSE



# MSE



# MSE



# MSE

---

Linear model =  $y = H(x) = ax + b$

$$\text{Square Error} = \mathbf{e}^2 = (e_1)^2 + (e_2)^2 + (e_3)^2 + (e_4)^2$$

$$\mathbf{e}^2 = (H(x_1) - y_1)^2 + (H(x_2) - y_2)^2 + (H(x_3) - y_3)^2 + (H(x_4) - y_4)^2$$

$$\mathbf{e}^2 = (H(\mathbf{x}) - \mathbf{y})^2$$

$$\text{MSE} = \mathbf{e}^2 / 4$$

```
1 import numpy as np
2
3 a_b = np.array([3, 76])
4 data = np.array([[2, 81], [4, 93], [6, 91], [8, 97]])
5 x = data[:, 0]
6 y = data[:, 1]
7
8 mse = sum(((a_b[0] * x + a_b[1]) - y)**2)/4
9
10 print(mse)
```

# Lab 8

---

Estimate a MSE of the linear model (arbitrary  $a = 1.5$  and  $b = 5.0$ ) for the given example data below. Upload .py or .ipynb file (source code) and captured output image file.

x	y
2.2	6.14
1.3	4.72
4.2	11.17
5.8	14.23
3.4	9.55
8.7	22.49



# Next

---

*How to find best (optimal)  $a$  and  $b$ ?*

*We are going to talk about an algorithm to find the optimal  $a$  and  $b$  next time.*