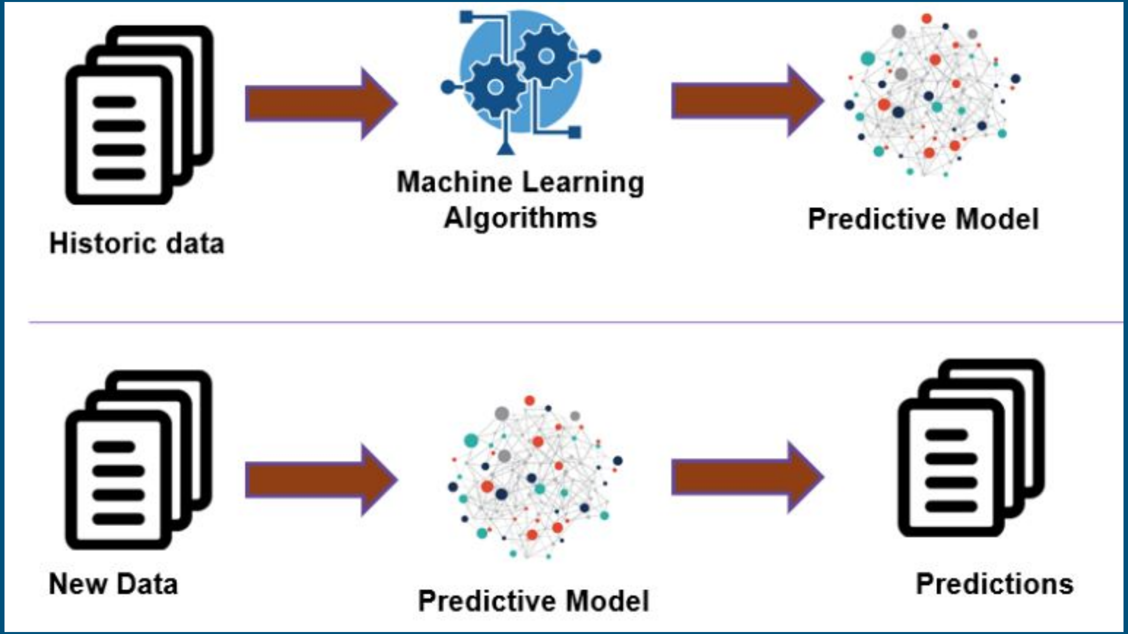
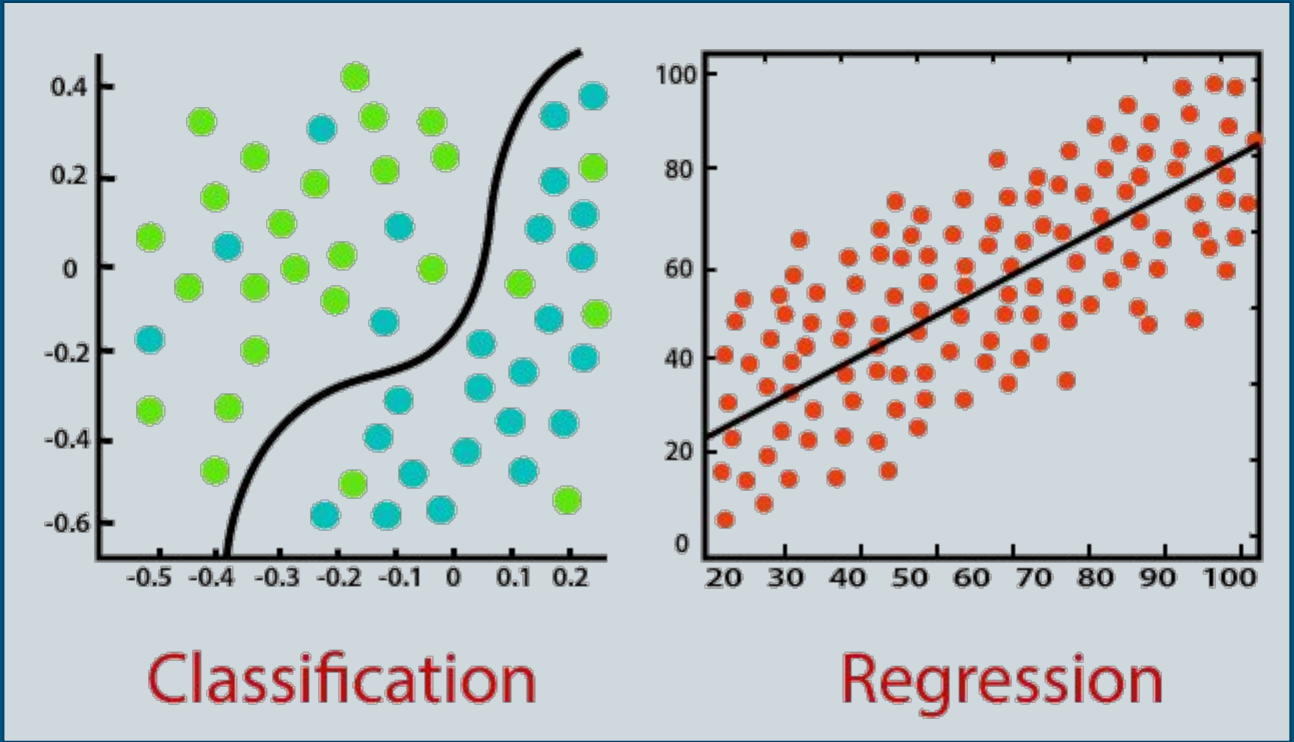
Linear Regression

Dr. Dongchul Kim

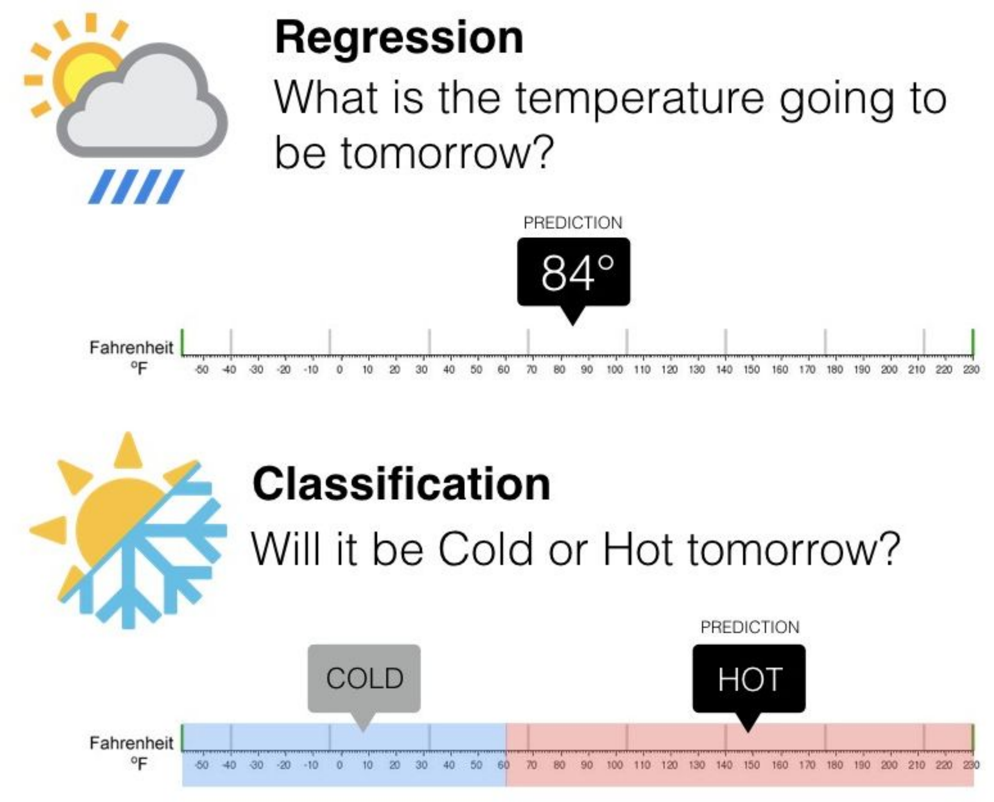
Prediction


Historic data
Machine Learning
Algorithms
Predictive Model
New Data
Predictive Model
Predictions

Regression vs Classification



Regression vs Classification



Data (Train and Test)

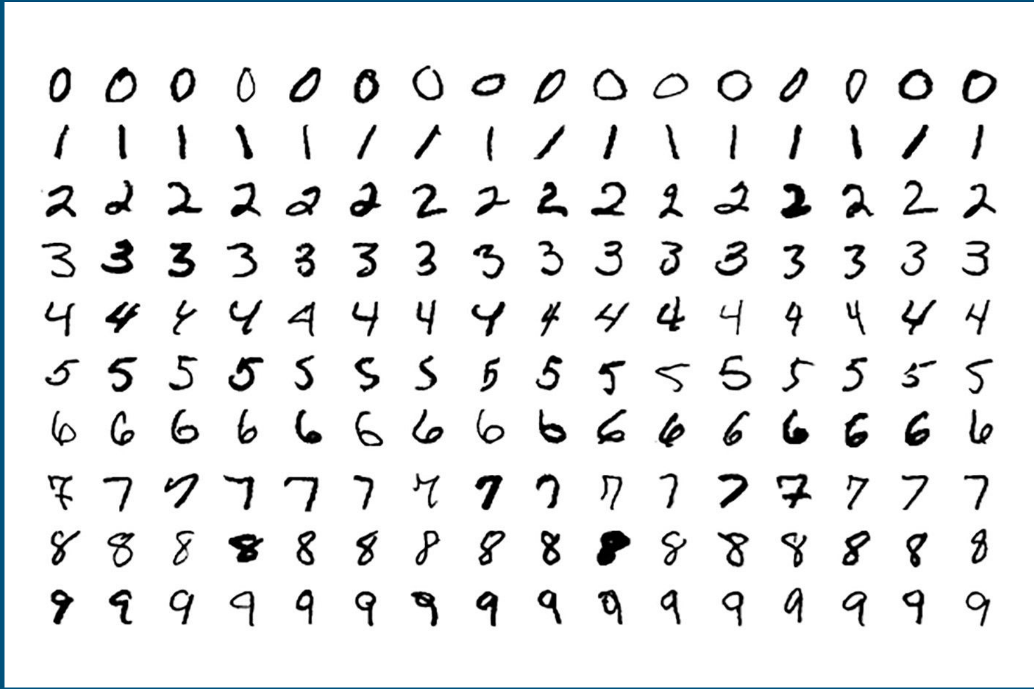
X

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

y

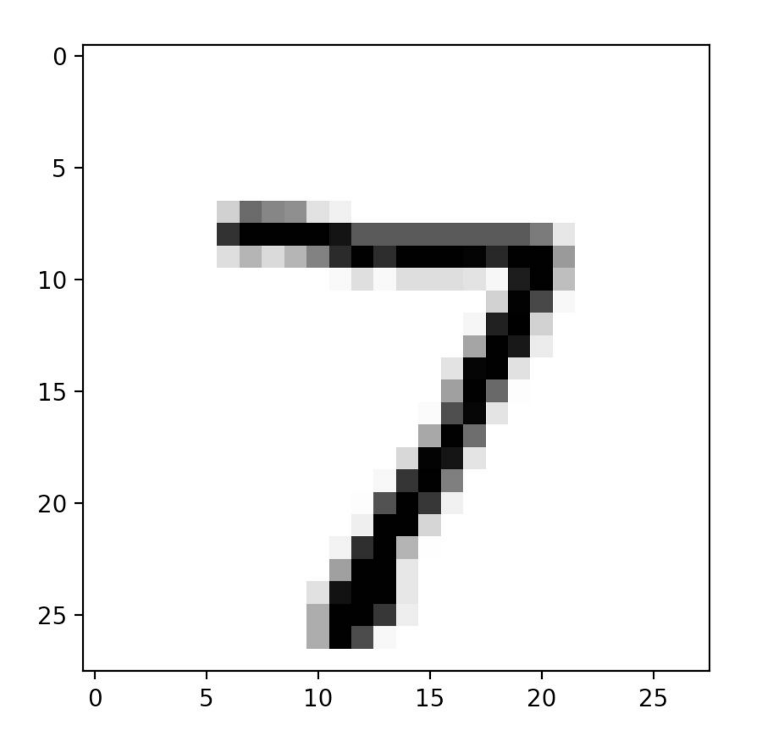
y

MNIST data (Classification)



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)

UCI

Machine Learning Repository

Center for Machine Learning and Intelligent Systems

Iris Data Set

Download: Data Folder, Data Set Description

Abstract: Famous database; from Fisher, 1936

Data Set Characteristics:	Multivariate	**Number of Instances:**	150	**Area:**	Life
Attribute Characteristics:	Real	**Number of Attributes:**	4	**Date Donated**	1988-07-01
Associated Tasks:	Classification	**Missing Values?**	No	**Number of Web Hits:**	3500413

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)

UCI

Machine Learning Repository

Center for Machine Learning and Intelligent Systems

Auto MPG Data Set

Download: Data Folder, Data Set Description

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

Data Set Characteristics:	Multivariate	**Number of Instances:**	398	**Area:**	N/A
Attribute Characteristics:	Categorical, Real	**Number of Attributes:**	8	**Date Donated**	1993-07-07
Associated Tasks:	Regression	**Missing Values?**	Yes	**Number of Web Hits:**	571036

Auto MPG

```
18.0   8   307.0      130.0      3504.      12.0   70   1    "chevrolet chevelle malibu"
15.0   8   350.0      165.0      3693.      11.5   70   1    "buick skylark 320"
18.0   8   318.0      150.0      3436.      11.0   70   1    "plymouth satellite"
16.0   8   304.0      150.0      3433.      12.0   70   1    "amc rebel sst"
17.0   8   302.0      140.0      3449.      10.5   70   1    "ford torino"
15.0   8   429.0      198.0      4341.      10.0   70   1    "ford galaxie 500"
14.0   8   454.0      220.0      4354.       9.0   70   1    "chevrolet impala"
14.0   8   440.0      215.0      4312.       8.5   70   1    "plymouth fury iii"
14.0   8   455.0      225.0      4425.      10.0   70   1    "pontiac catalina"
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

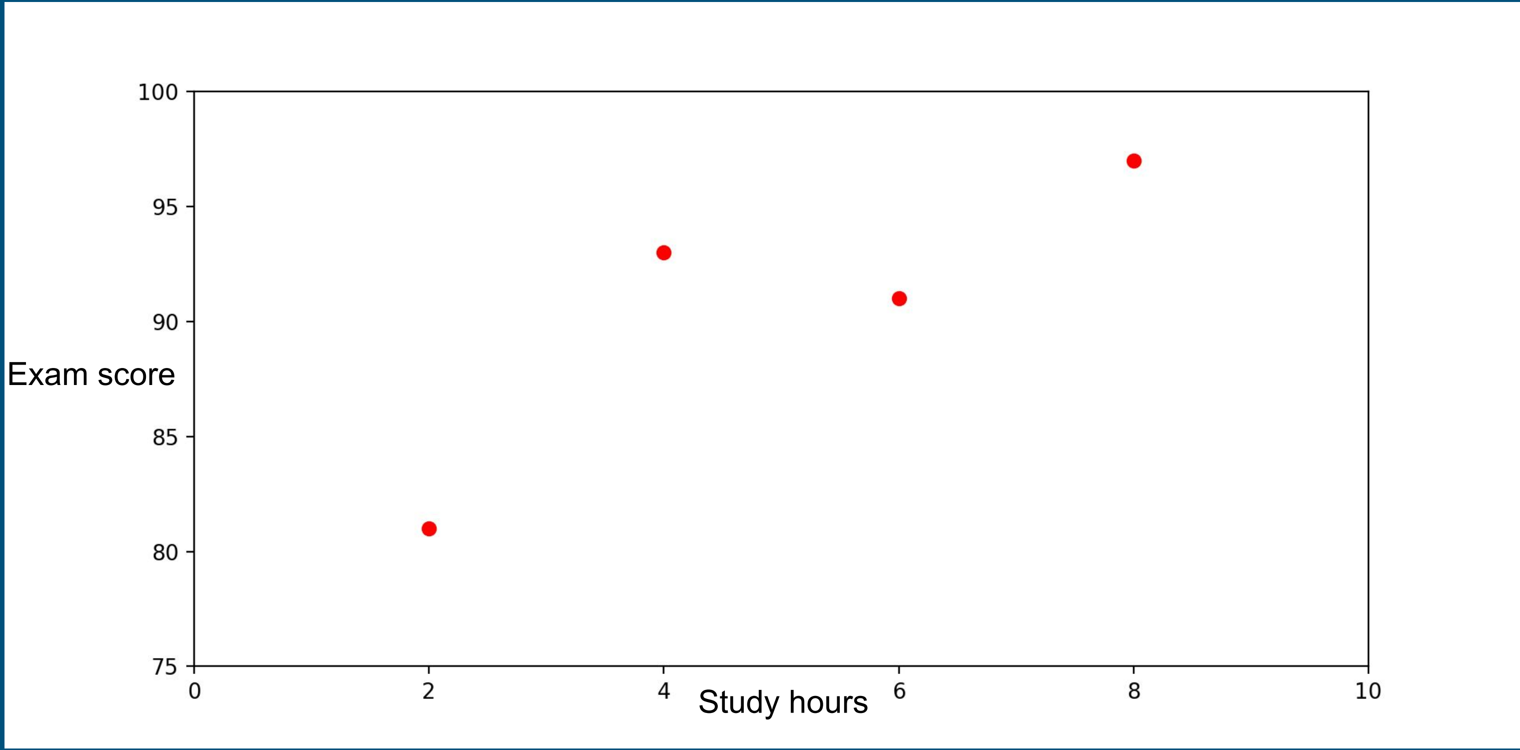
y

Example: Predict your exam score

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

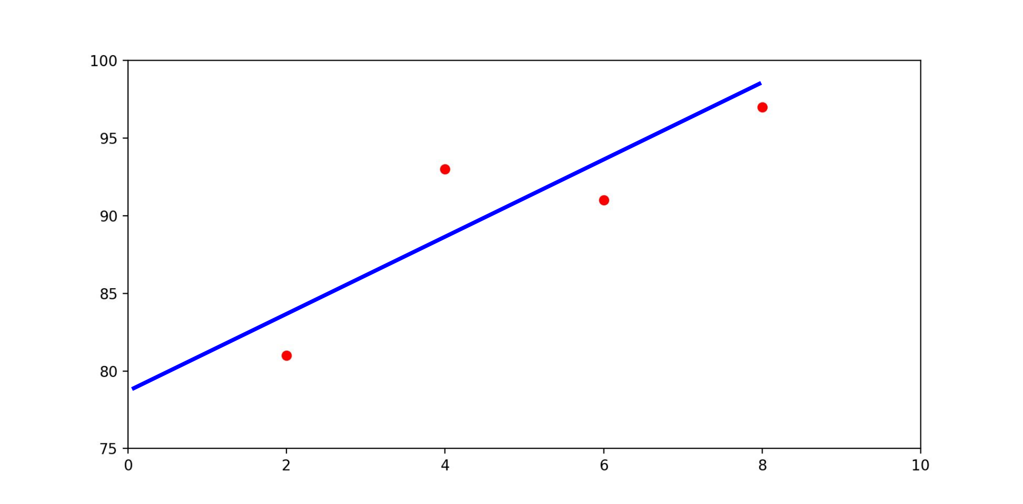
Study hours	Score
2 hours	81
4 hours	93
6 hours	91
8 hours	97

Example: Predict your exam score



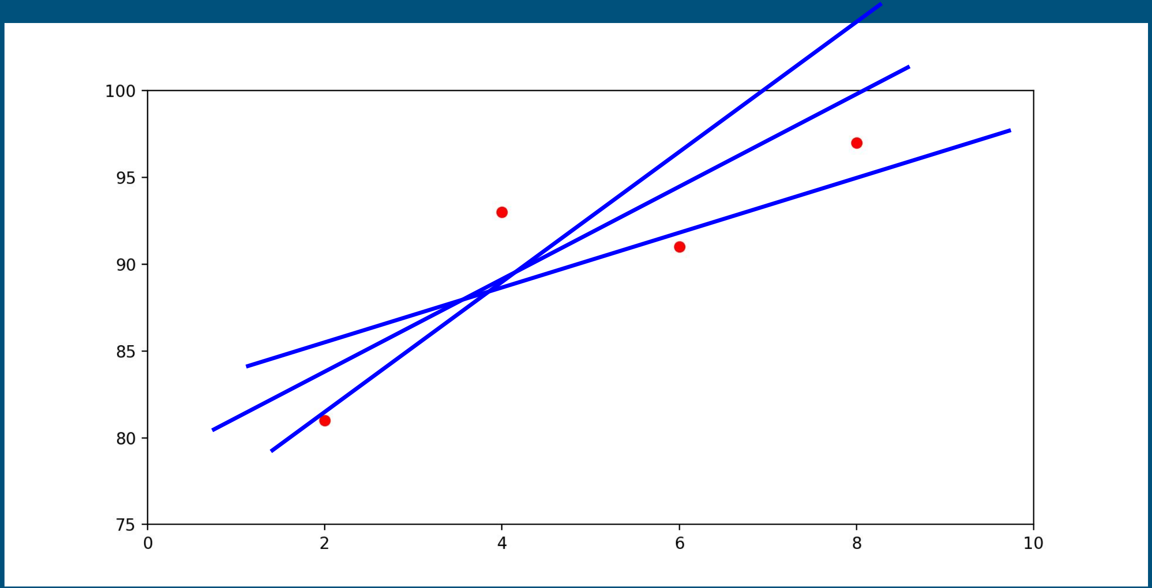
Example: Predict your exam score

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.

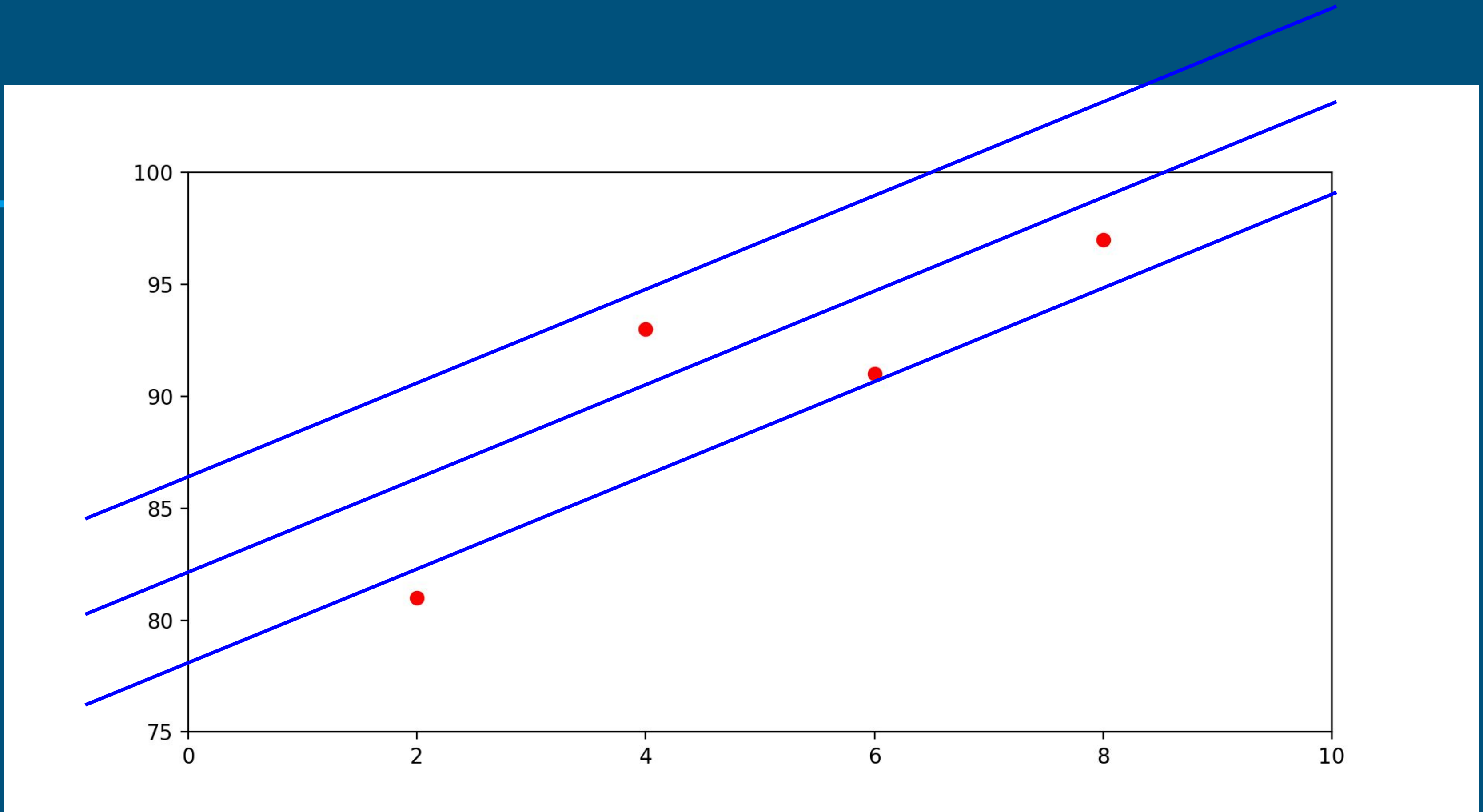


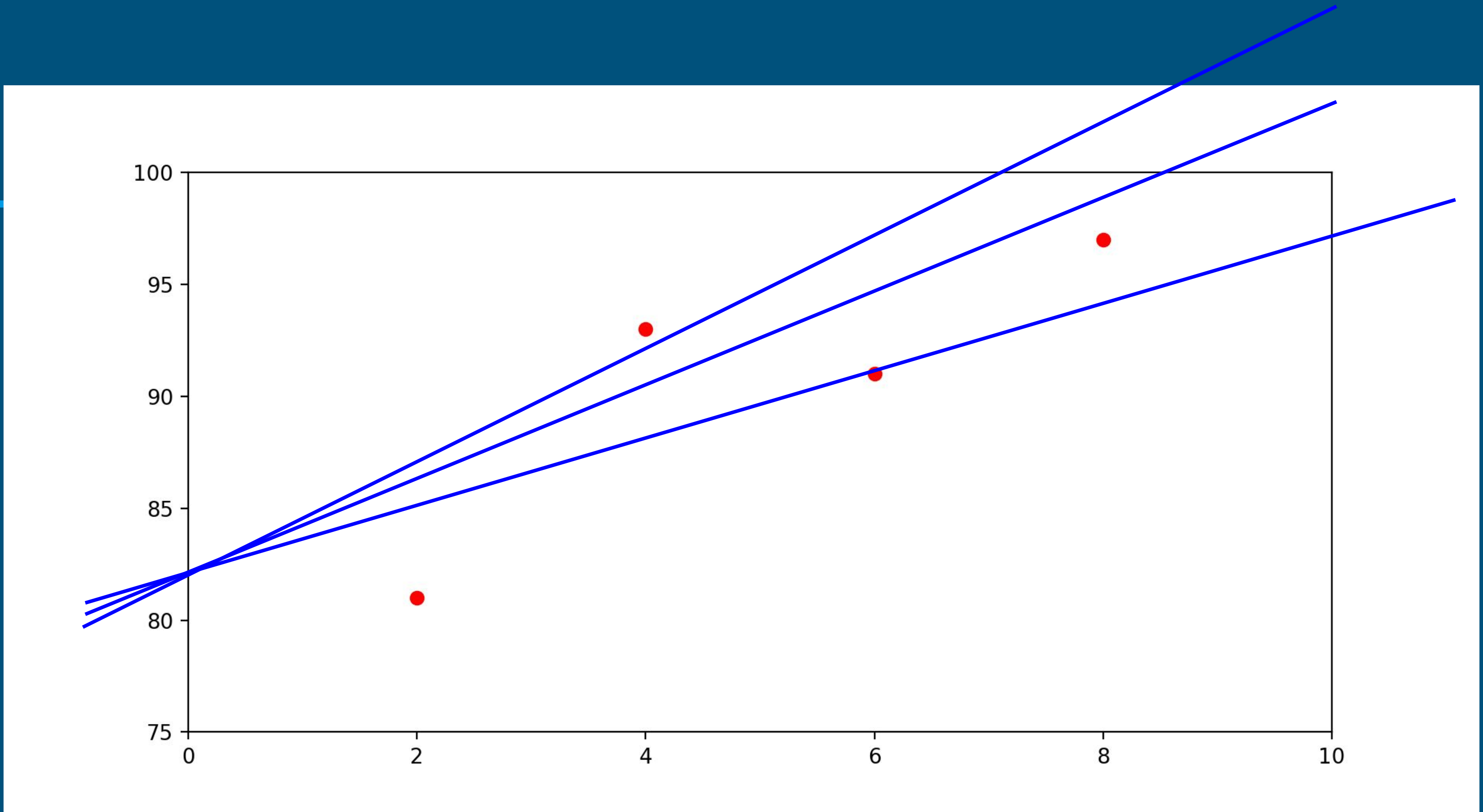
Example: Predict your exam score

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.


100
95
90
85
80
75
0
2
4
6
8
10


100
95
90
85
80
75
0
2
4
6
8
10

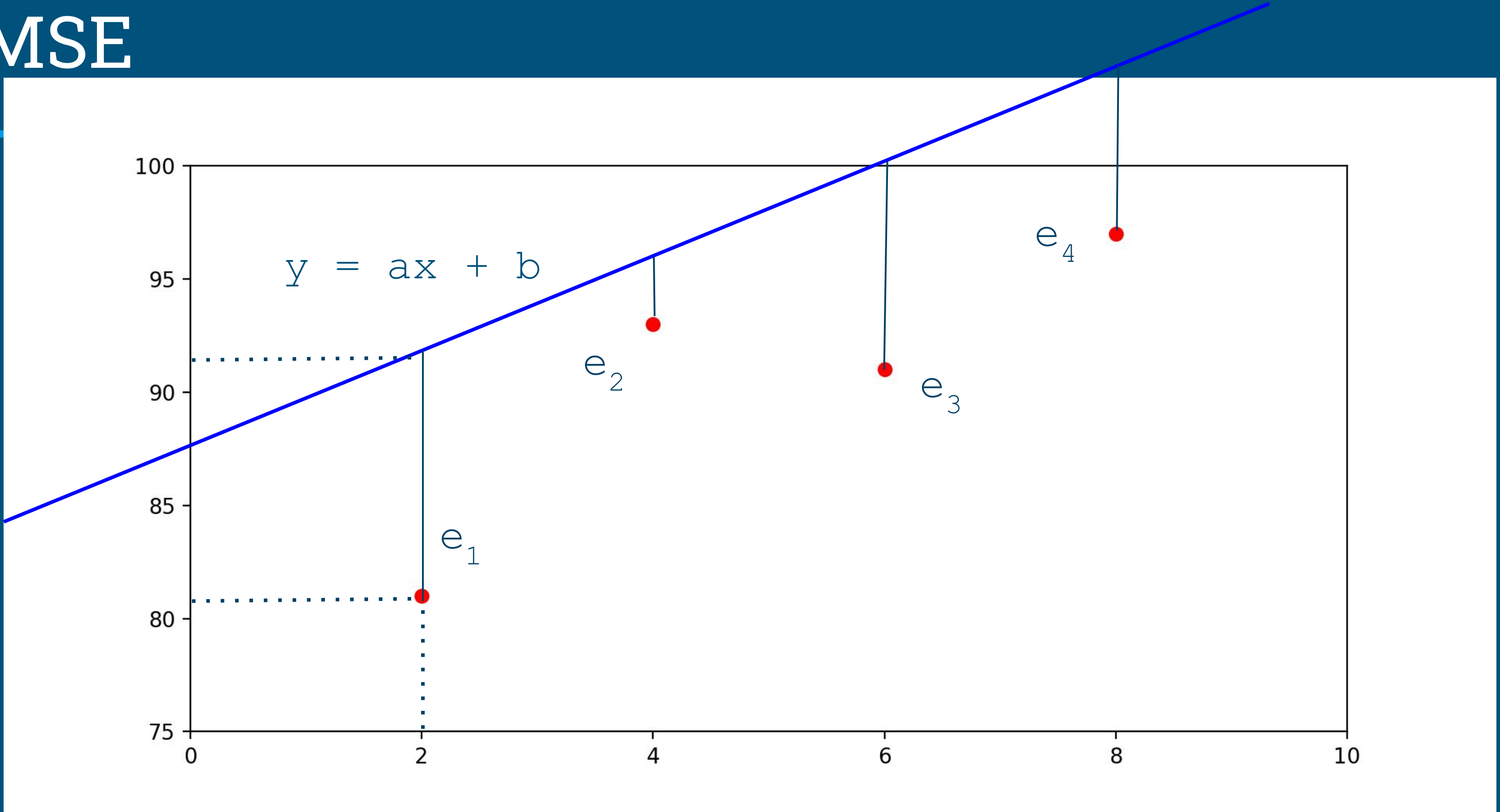
How to choose the best/optimal `a` and `b`

Line Evaluation and Error Minimization:

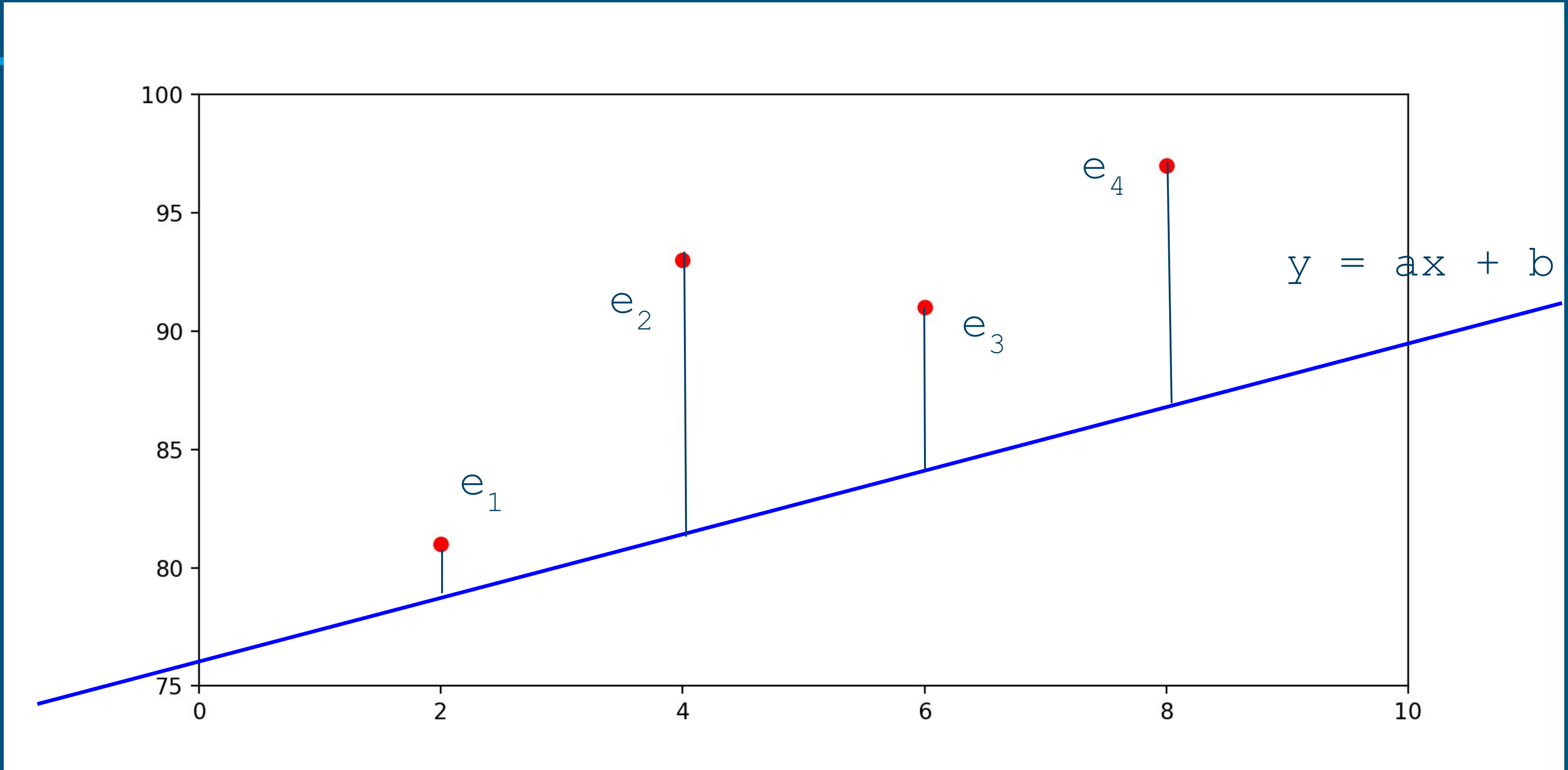
When drawing a line, it is essential to assess its accuracy. This evaluation involves determining the error associated with the line. To achieve this, we utilize the Mean Squared Error (MSE) method. By employing MSE, we iteratively refine the lines to minimize the error.

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

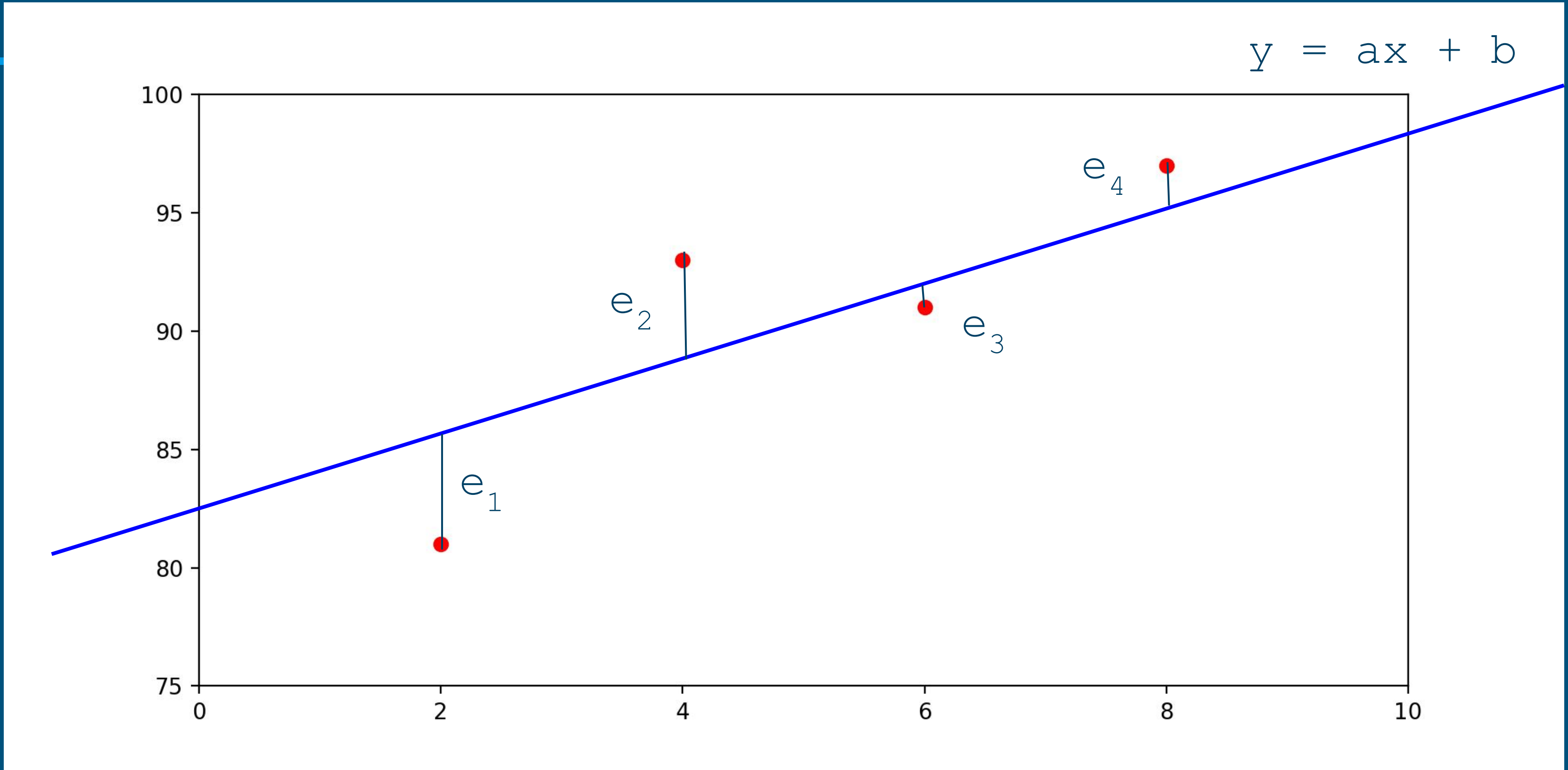
MSE



MSE



MSE



MSE

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

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

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

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

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

```python
import numpy as np

a_b = np.array([3, 76])
data = np.array([[2, 81], [4, 93], [6, 91], [8, 97]])
x = data[:, 0]
y = data[:, 1]

mse = sum(((a_b[0] * x + a_b[1]) - y)**2)/4

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

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.