Data Science

- Start of the Journey

What is the Relationship ?

Y = ?????????

X	Y
2	8
6	20
4	14
3	11
7	23
4	14
2	8
5	17

Relationship

$Y = 2 + 3(X)$

X	Y
2	8
6	20
4	14
3	11
7	23
4	14
2	8
5	17

Find the Y in ?

$Y = 2 + 3(X)$

X	Y
2	8
6	20
4	14
3	11
7	23
4	14
2	8
5	17
10	?
1	?

Value for Y with given X

$$Y = 2 + 3(X)$$

X	Y
2	8
6	20
4	14
3	11
7	23
4	14
2	8
5	17
10	32
1	5

Terminology

$Y = 2 + 3(X)$

Y = Model

2 = Intercept

3 = Slope

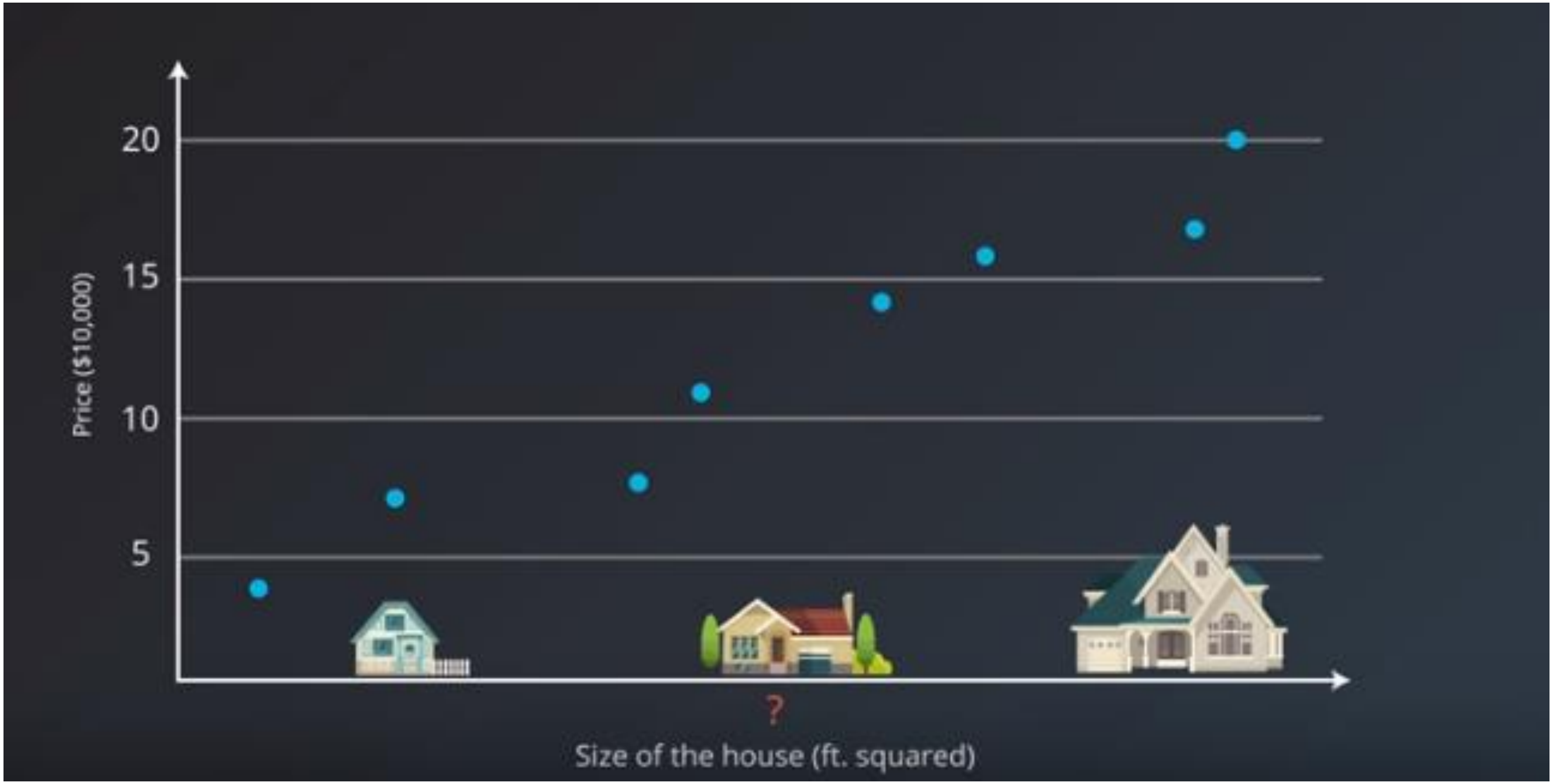
X = input

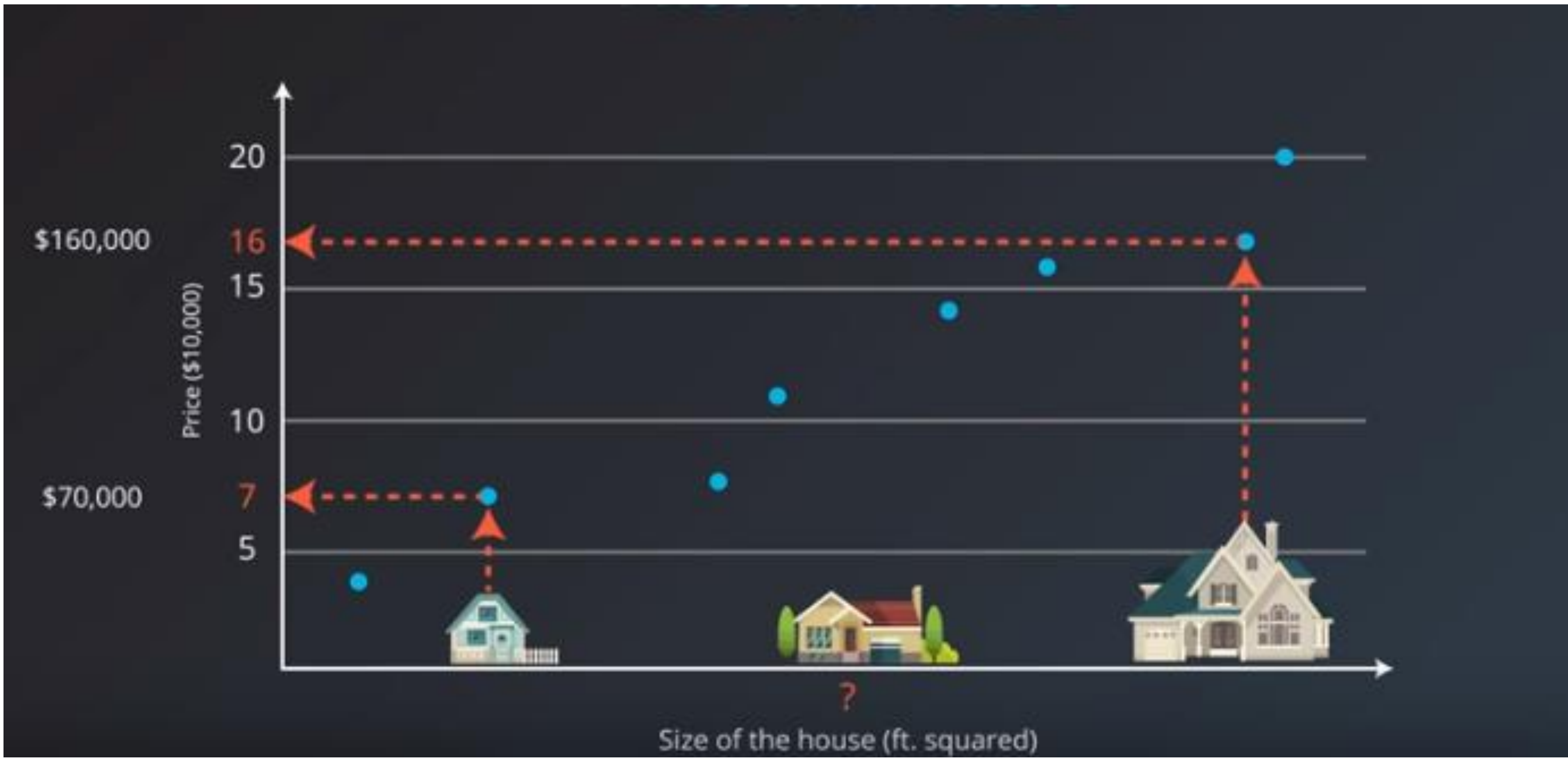
X	Y
2	8
6	20
4	14
3	11
7	23
4	14
2	8
5	17
10	32
1	5

Predict the price of House ?

Price of a House


20
16
15
Price ($10,000)
12
10
7
5
?
Size of the house (ft. squared)


20
15
10
5
Price ($10,000)
?
Size of the house (ft. squared)



$160,000
$70,000
Price ($10,000)
20
16
15
10
7
5
?
Size of the house (ft. squared)

It’s all about

- Finding the “best-fit” line is the **goal** of simple linear regression.

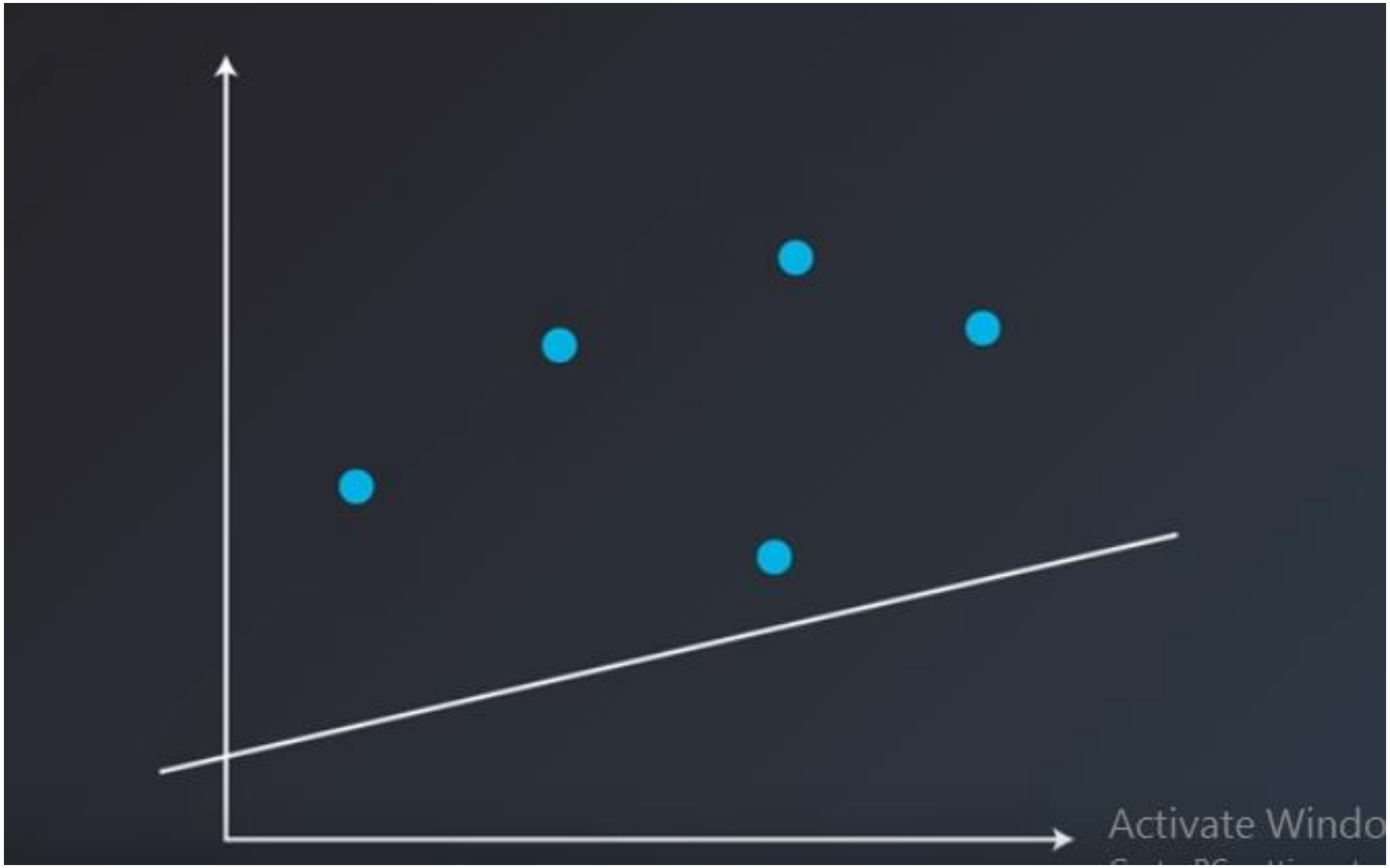
Linear Regression

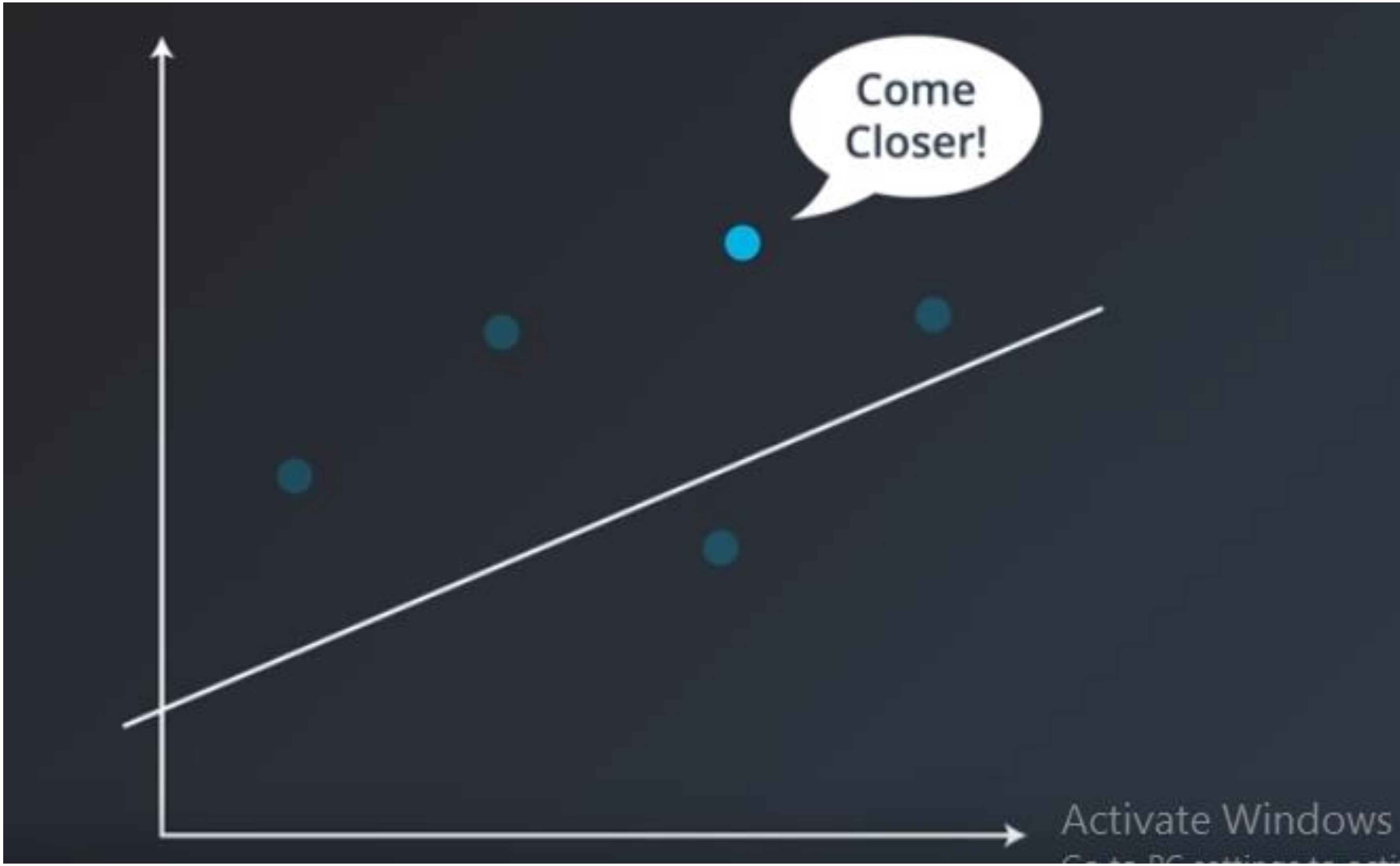
- Welcome to the world of data science

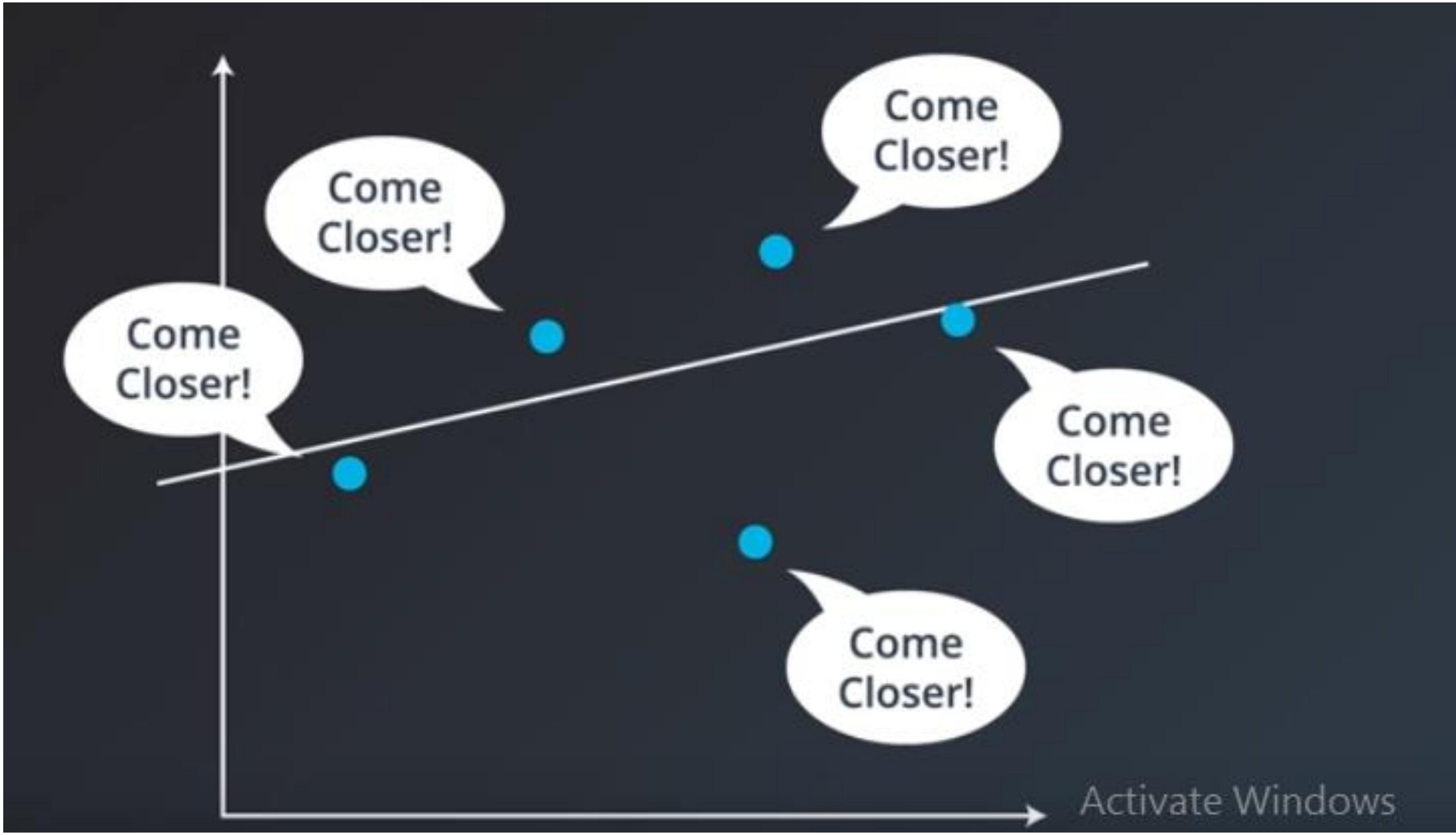
What is Simple Linear Regression?

- Simple Linear Regression is a method used to fit the **best straight line** between a set of data points.
- After a graph is properly scaled, the data points must "look" like they would fit a straight line, not a parabola, or any other shape.
- The line is used as a model in order to predict a variable y from another variable x.
- A regression line must involve 2 variables, the dependent and the independent variable.
- Finding the "best-fit" line is the **goal** of simple linear regression.

Fitting A Line



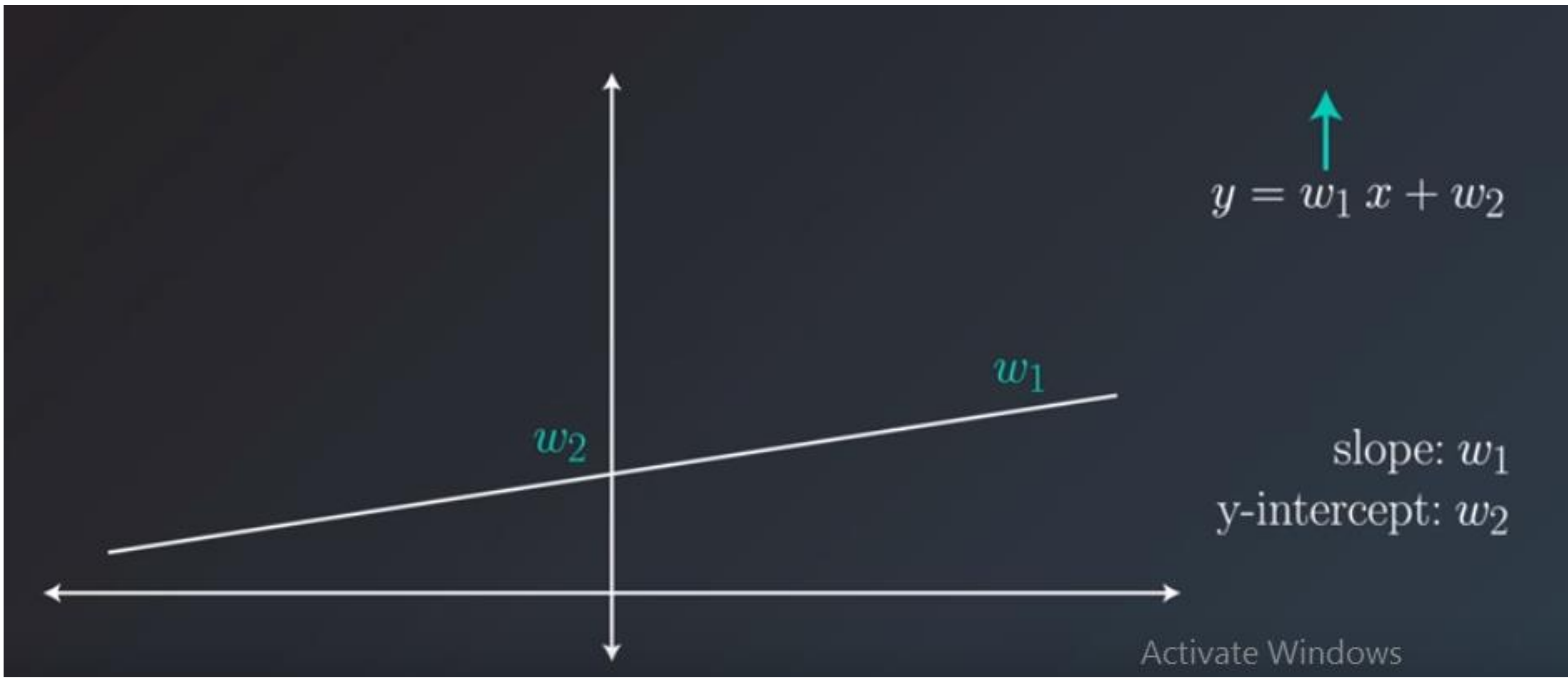

Come Closer!
Activate Windows

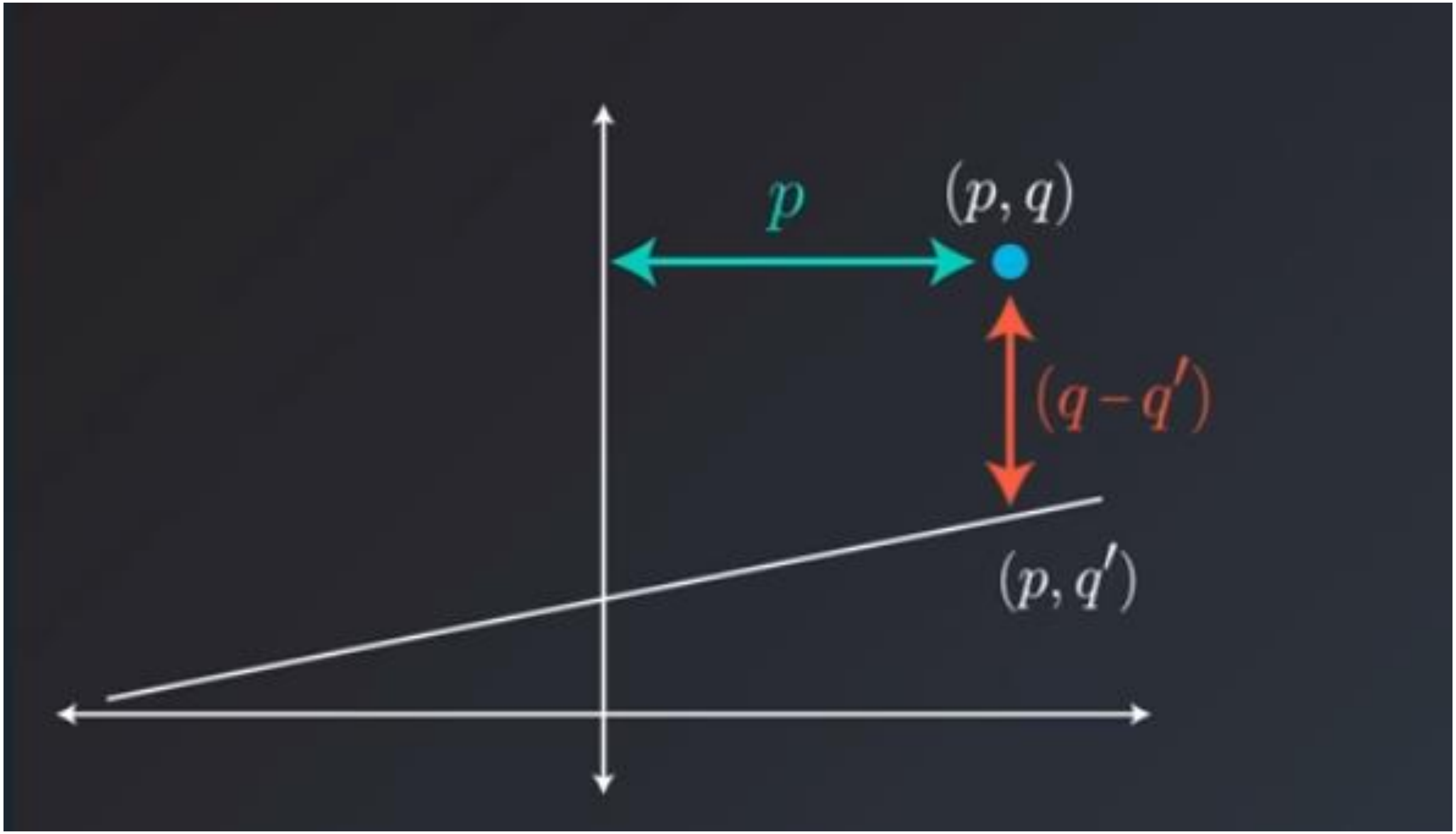

Come Closer!
Come Closer!
Come Closer!
Come Closer!
Come Closer!
Activate Windows

Equation of a Straight Line

$$y = w_1 x + w_2$$

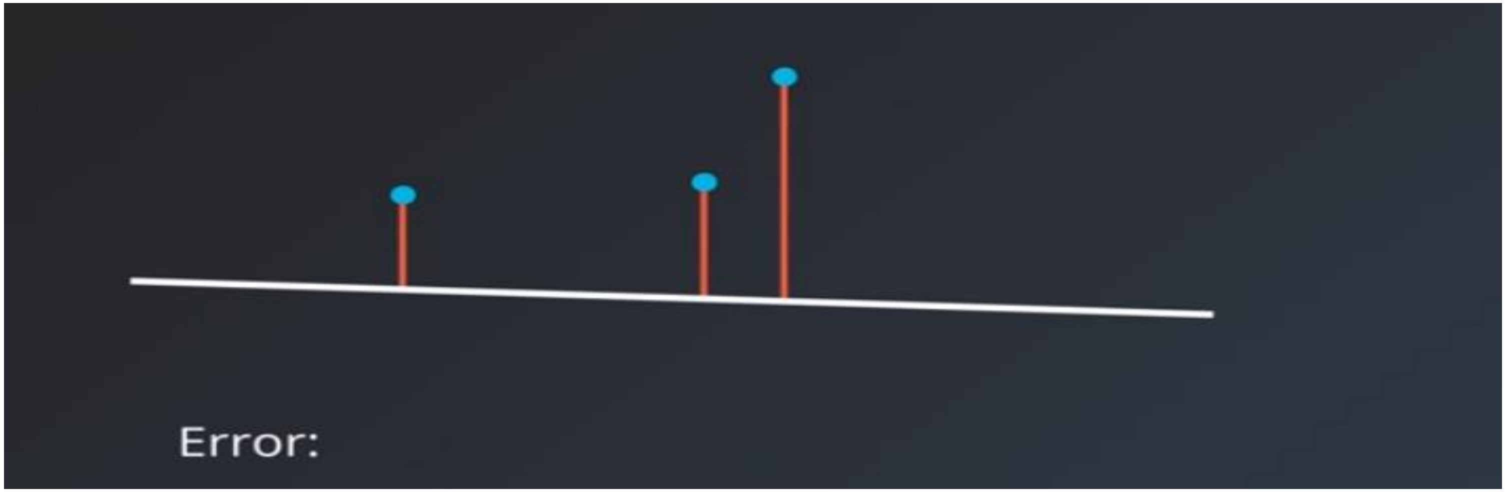
Moving A Line

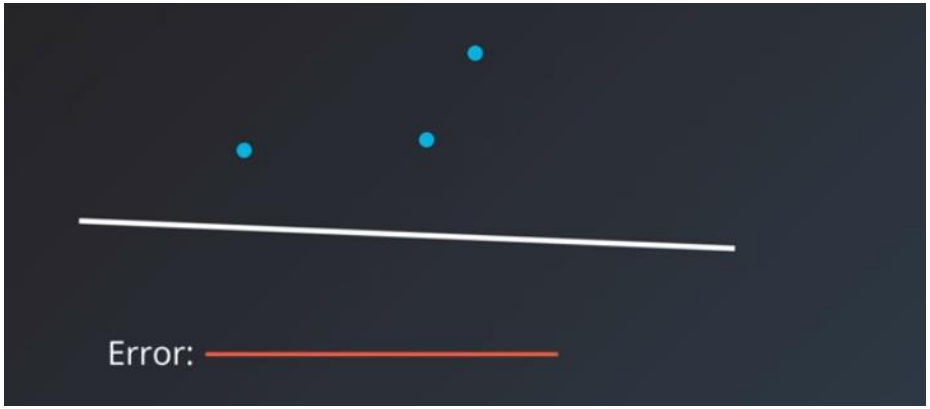


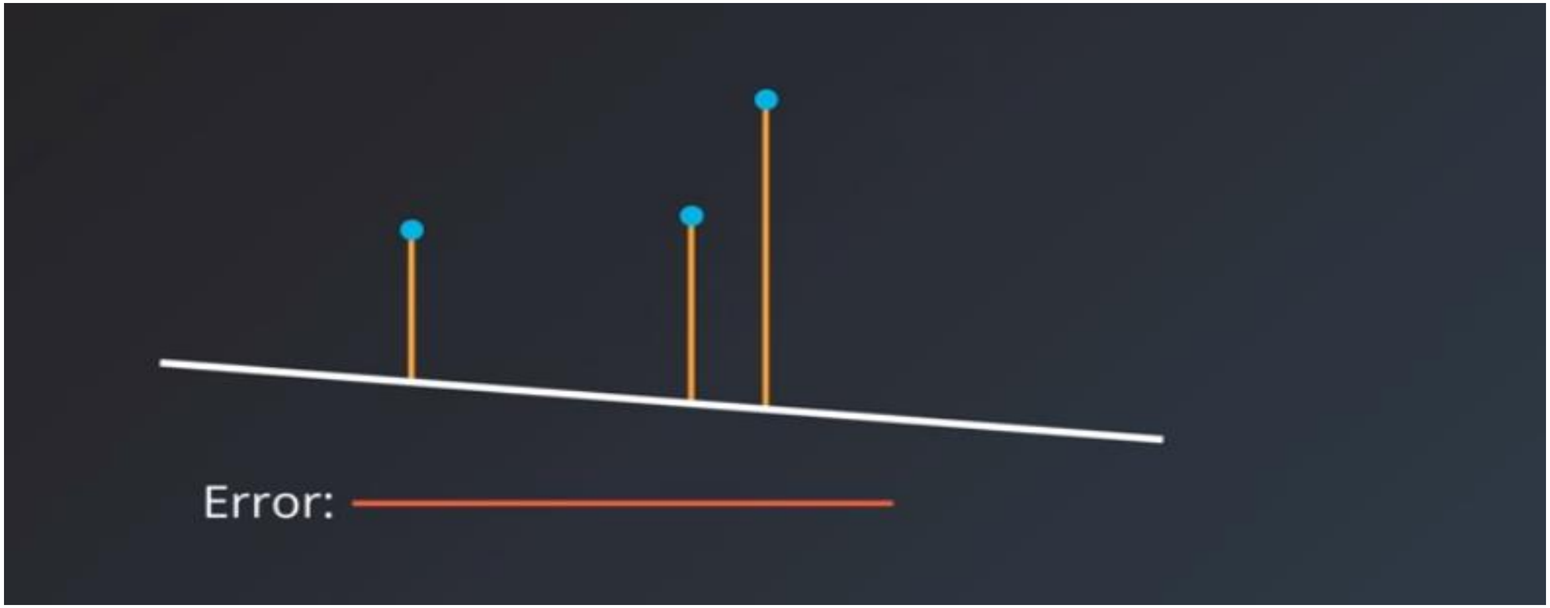


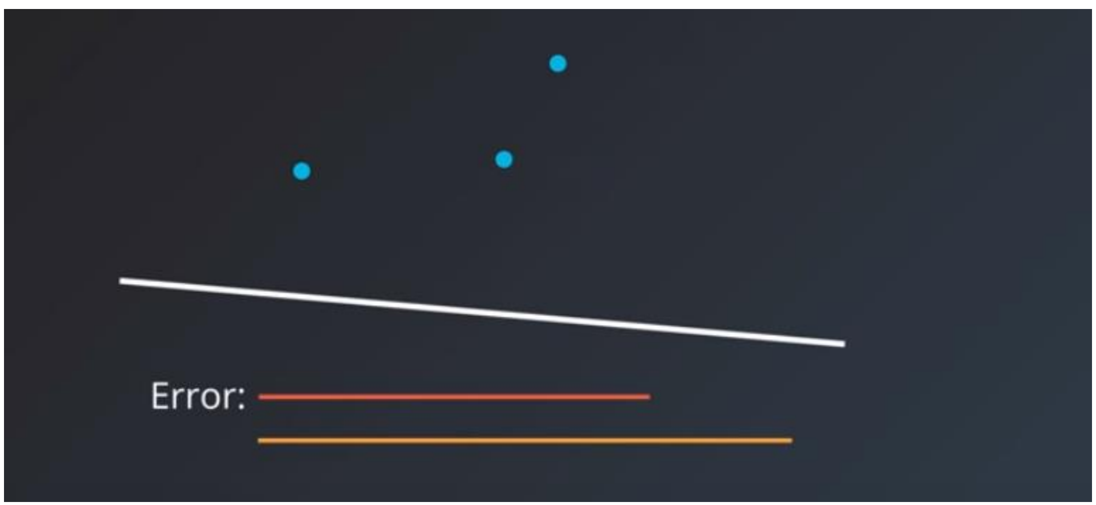
$$y = w_1 x + w_2$$

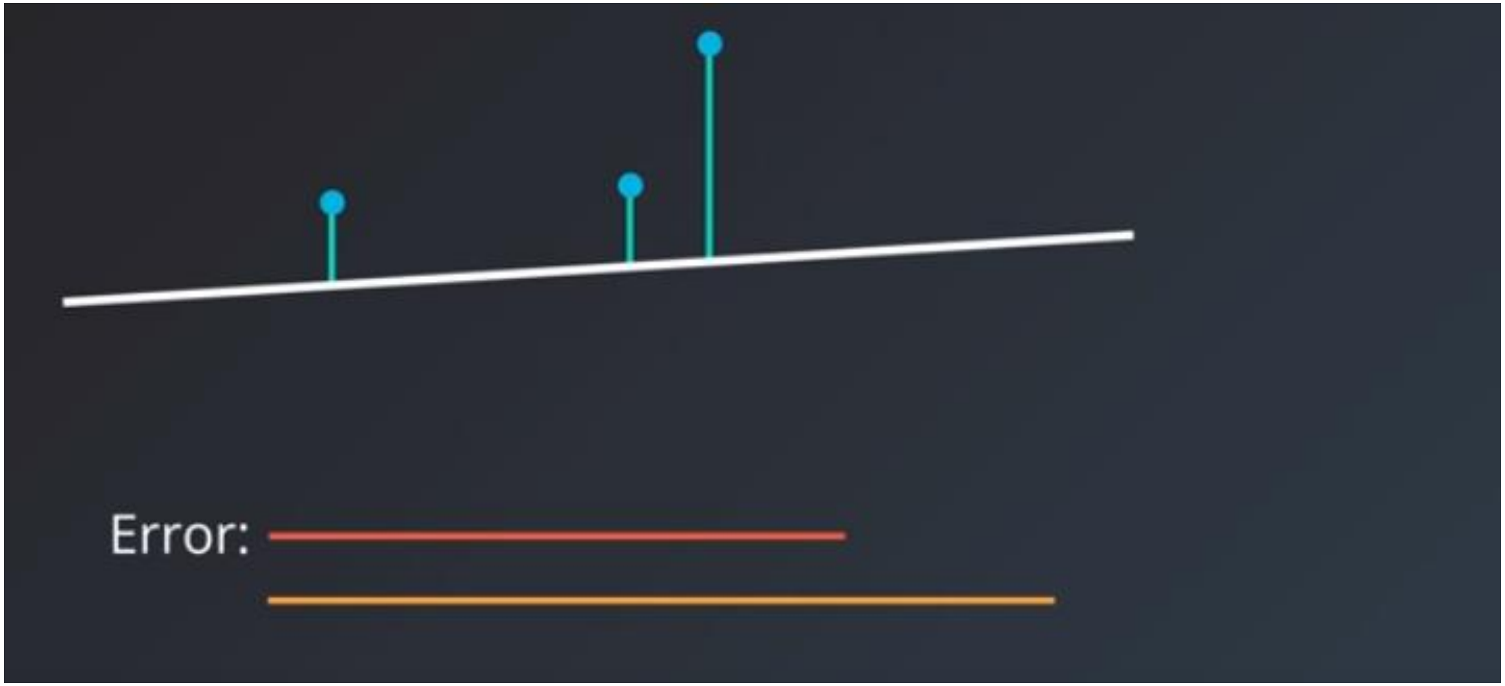
Line vs Error

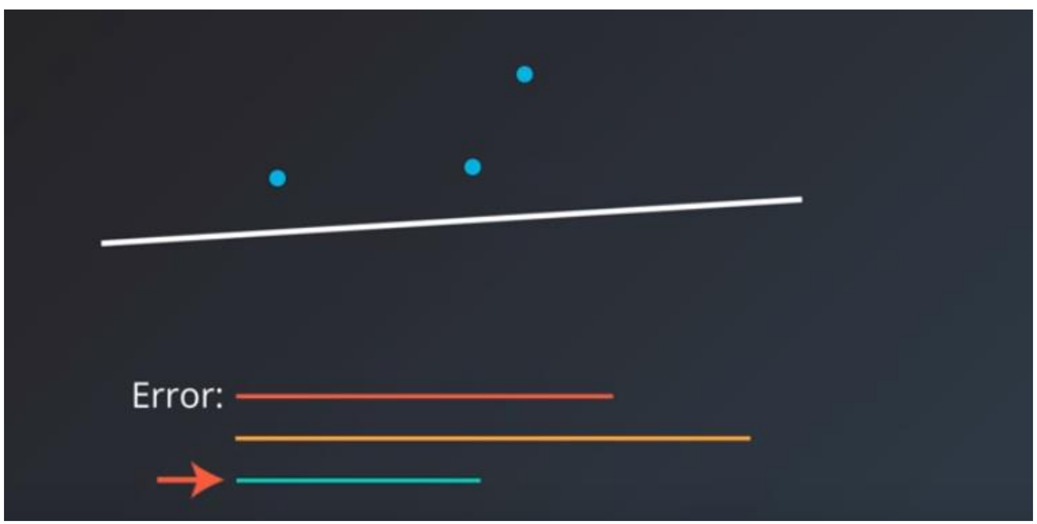





Error:


Error:


Error:

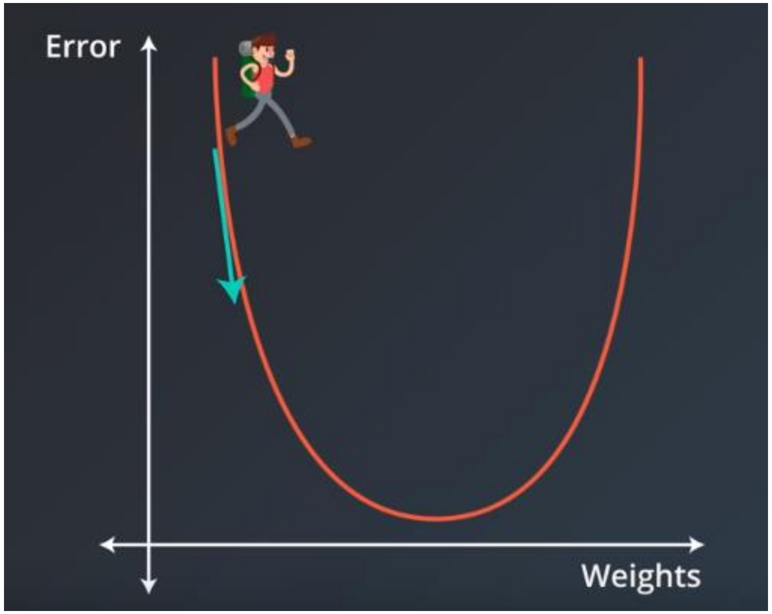

Error:

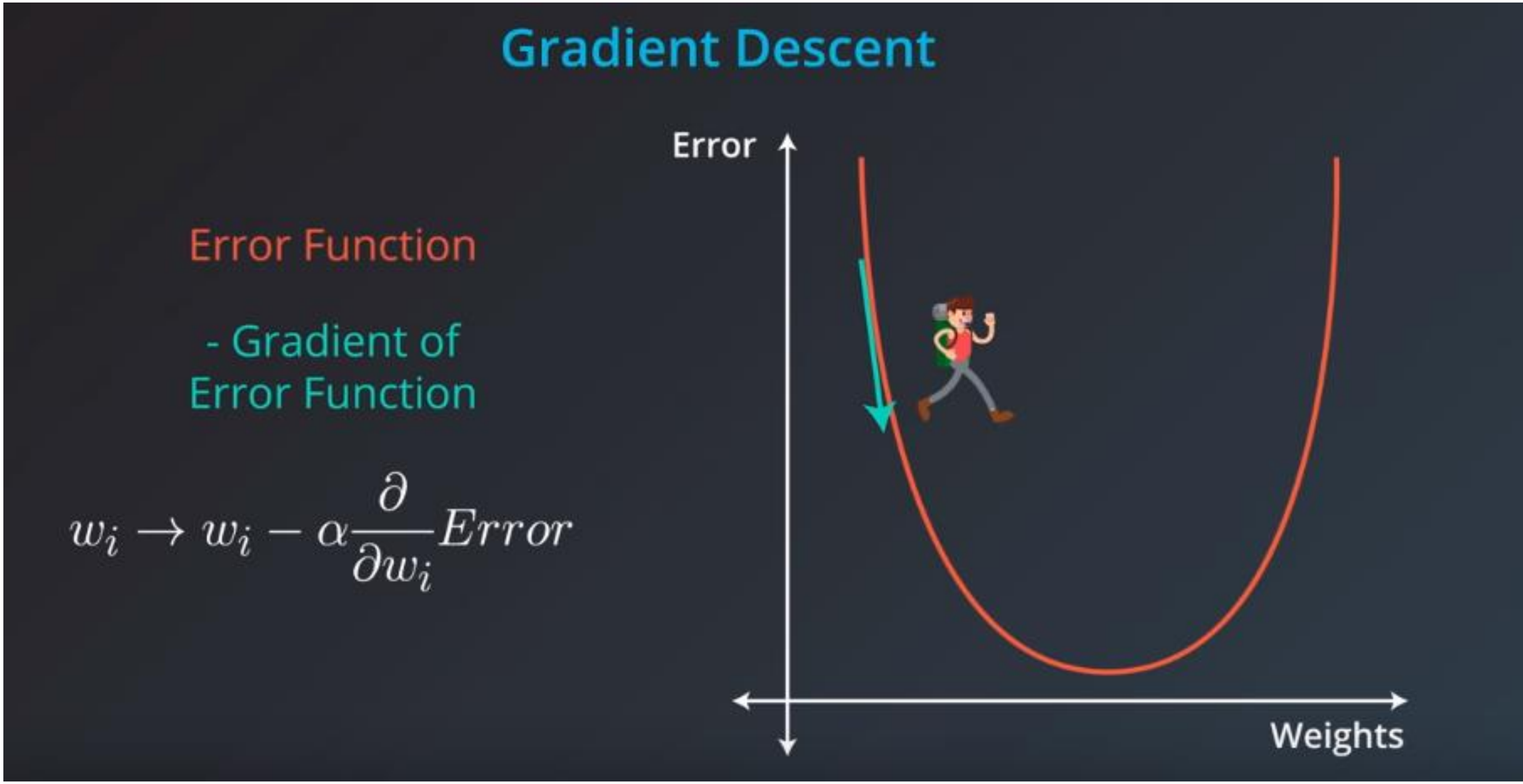
How Should the Line move ?

Gradient Descent

Error Function

\- Gradient of
Error Function




Gradient Descent
Error Function
- Gradient of
Error Function
$$w_i \to w_i - \alpha \frac{\partial}{\partial w_i} Error$$
Error
Weights

$$\frac{\partial}{\partial w_1} Error = -(y - \hat{y})\, x$$

$$\frac{\partial}{\partial w_2} Error = -(y - \hat{y})$$

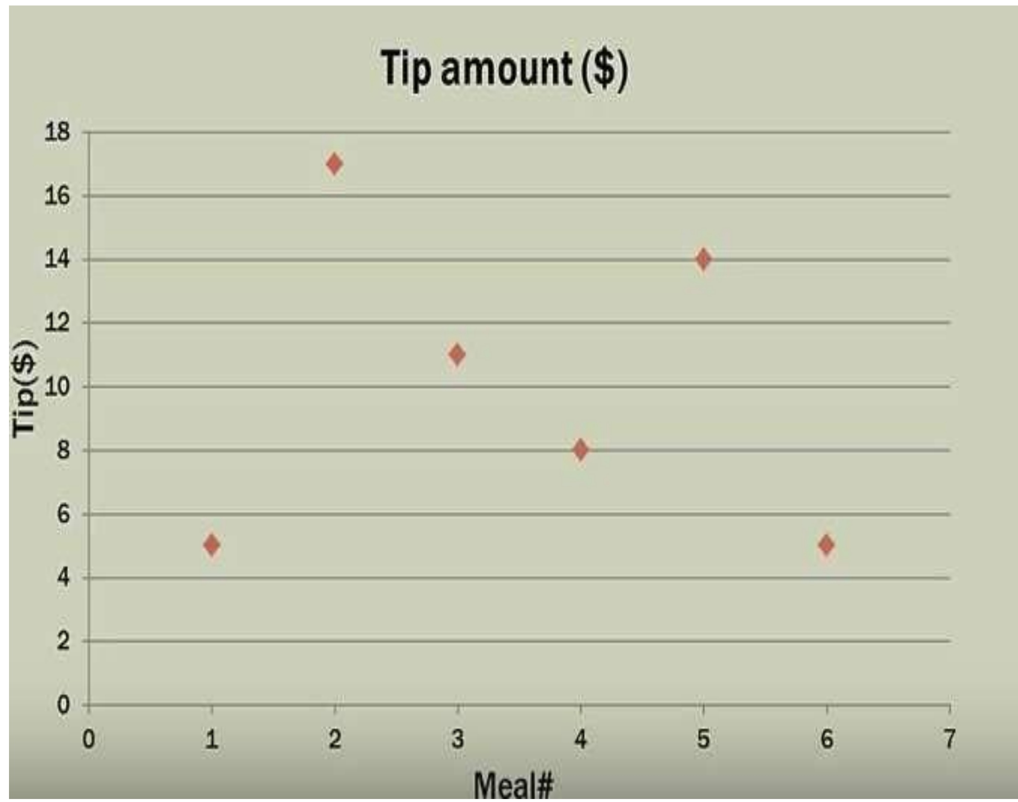
One Variable

-No Independent variable

- **Problem: A waiter wants to predict his next tip, but he forgot to record the bill amounts for previous tips.**
- **Here is a graph of his tips. The tips is the only variable. Let's call it the y variable.**
- **Meal# is not a variable. It is simply used to identify a tip.**

y variable

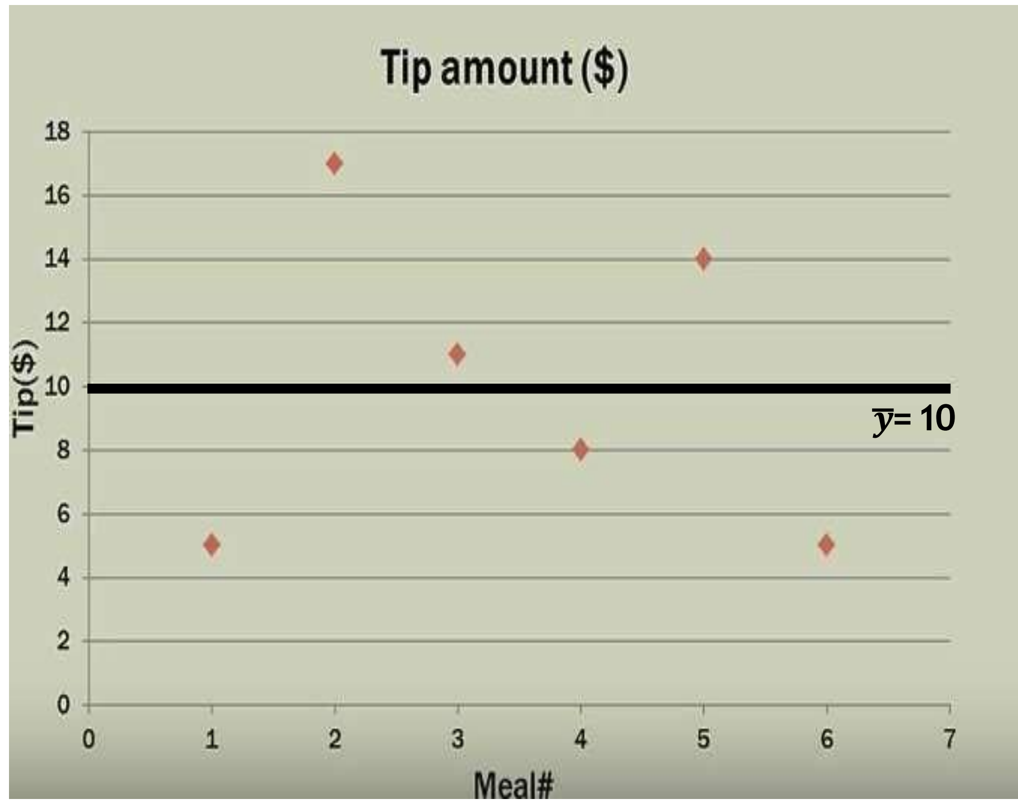
Meal#	Tip amount ($)
1	5.00
2	17.00
3	11.00
4	8.00
5	14.00
6	5.00



Can we come up with a model for this problem with only 1 variable?

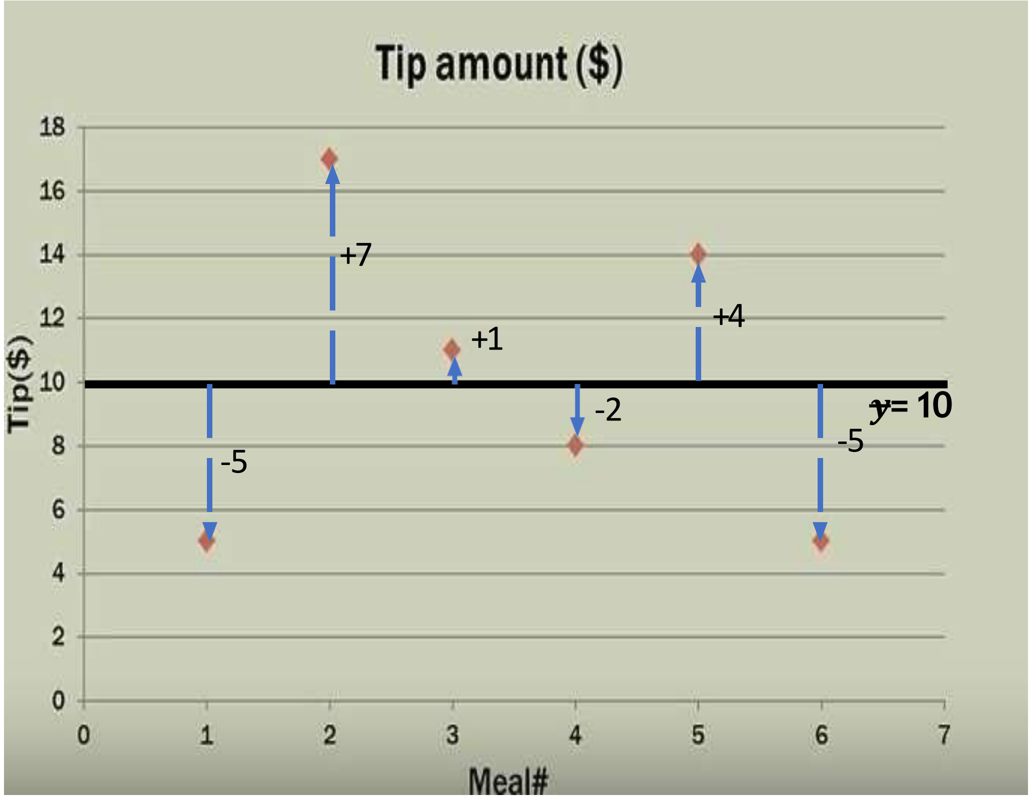
- The only option for our model is to use the mean of the Tips($)
- Tips are on the y access. We would call the mean $\bar{y}$ (y bar).
- The mean for the tip amounts is 10.
- The model for our problem is simply $\bar{y} = 10$.
- $\bar{y} = 10$ is our *best fit line* (represented by bold blackline).

Meal#	Tip amount ($)
1	5.00
2	17.00
3	11.00
4	8.00
5	14.00
6	5.00



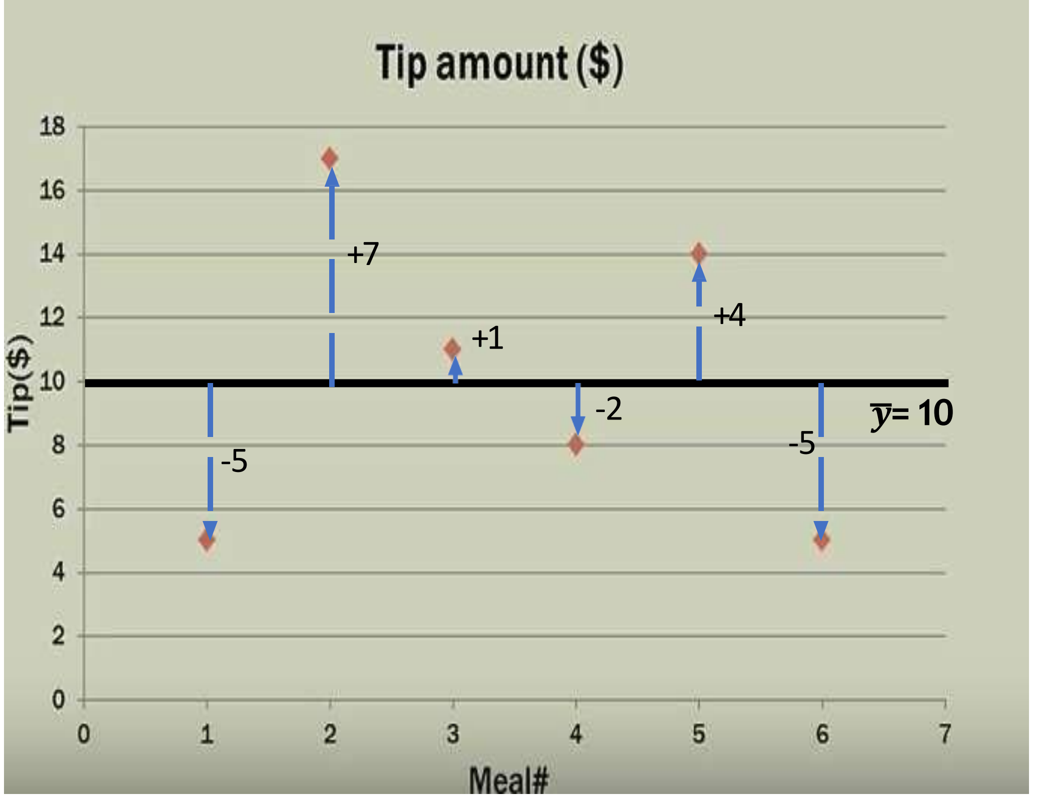
- Now, let's talk about goodness of fit. This will tell us how good our data points fit the line.
- We need to calculate the residuals (errors) for each point.

Meal#	Tip amount ($)
1	5.00
2	17.00
3	11.00
4	8.00
5	14.00
6	5.00



- The best fit line is the one that minimizes the sum of the squares of the residuals (errors).
- The error is the difference between the actual data point and the point on the line.
- SSE (Sum Of Squared Errors) = $(-5)^2 + 7^2 + 1^2 + (-2)^2 + 4^2 + (-5)^2 = 120$

Meal#	Tip amount ($)
1	5.00
2	17.00
3	11.00
4	8.00
5	14.00
6	5.00



- SST (Sum Of Squared Total) = SSR (Sum Of Squared Regression) + SSE is the Sum Of Squares Equation.
- Since there is no regression line (as we only have 1 variable), we can not make the SSE any smaller than 120, because SSR = 0.

Two Variables

- One Independent / Dependent variable

- **Repeating the Problem: As a waiter, how do we predict the tips we will receive for service rendered?**

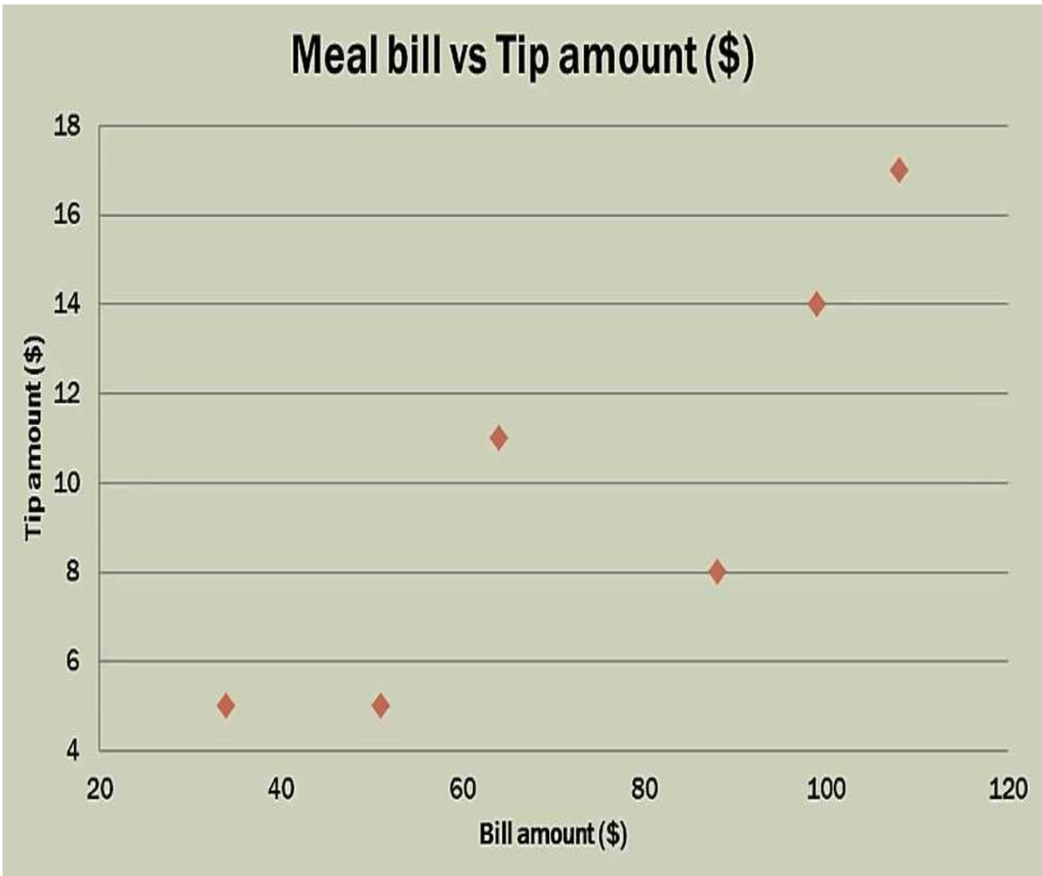
- **Let's say, we didn't forget to record the bill amount.**

Independent Variable (x)

Dependent Variable (y)

Total bill ($)	Tip amount ($)
34.00	5.00
108.00	17.00
64.00	11.00
88.00	8.00
99.00	14.00
51.00	5.00

If we scale the graph according to the data points available, we can then plot the points.

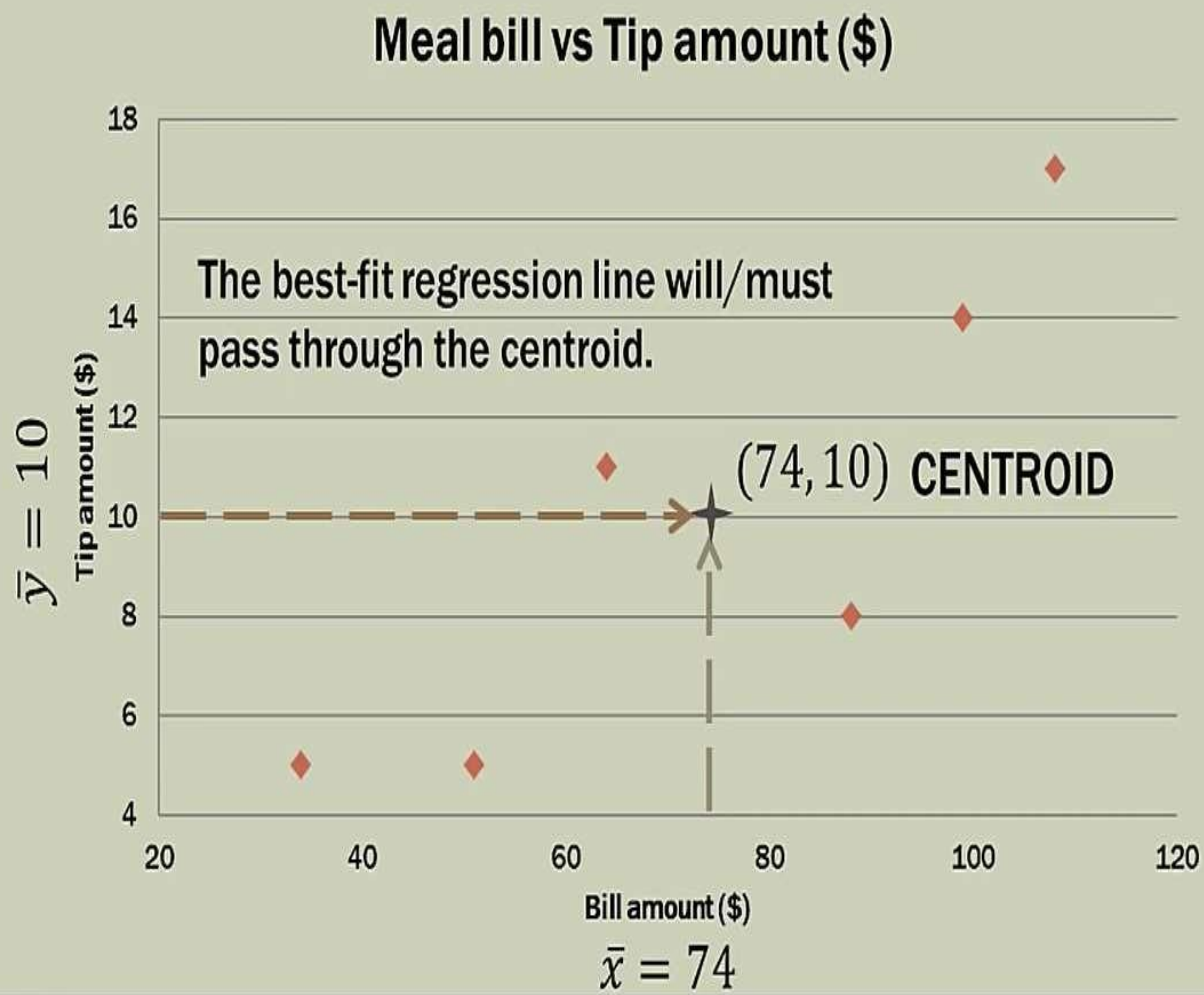


Bill ($)	Tip ($)
34.00	5.00
108.00	17.00
64.00	11.00
88.00	8.00
99.00	14.00
51.00	5.00

Does the data seem to fall along a line?

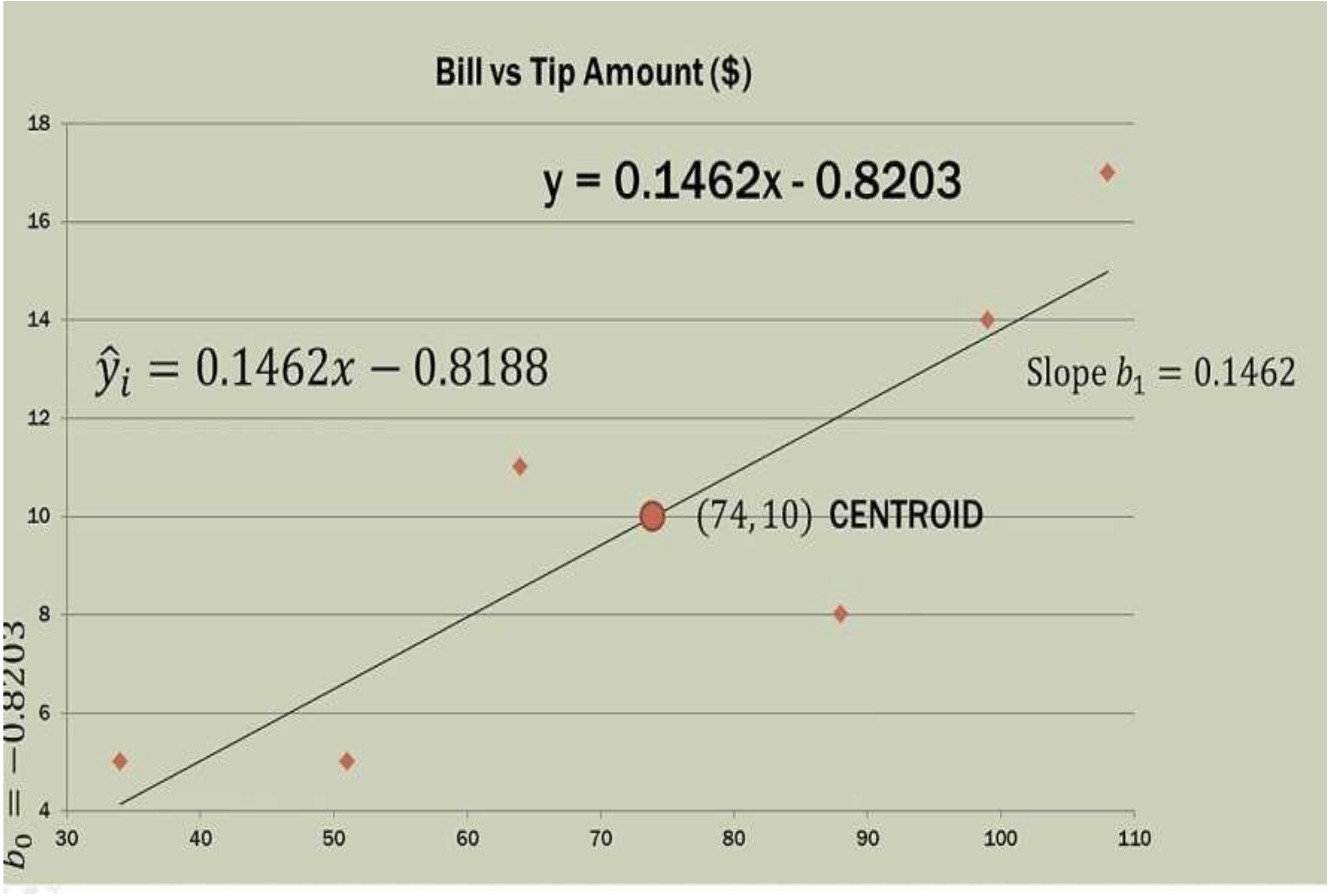
In this case, YES! Proceed.

If not...if it's a BLOB with no linear pattern, then stop.



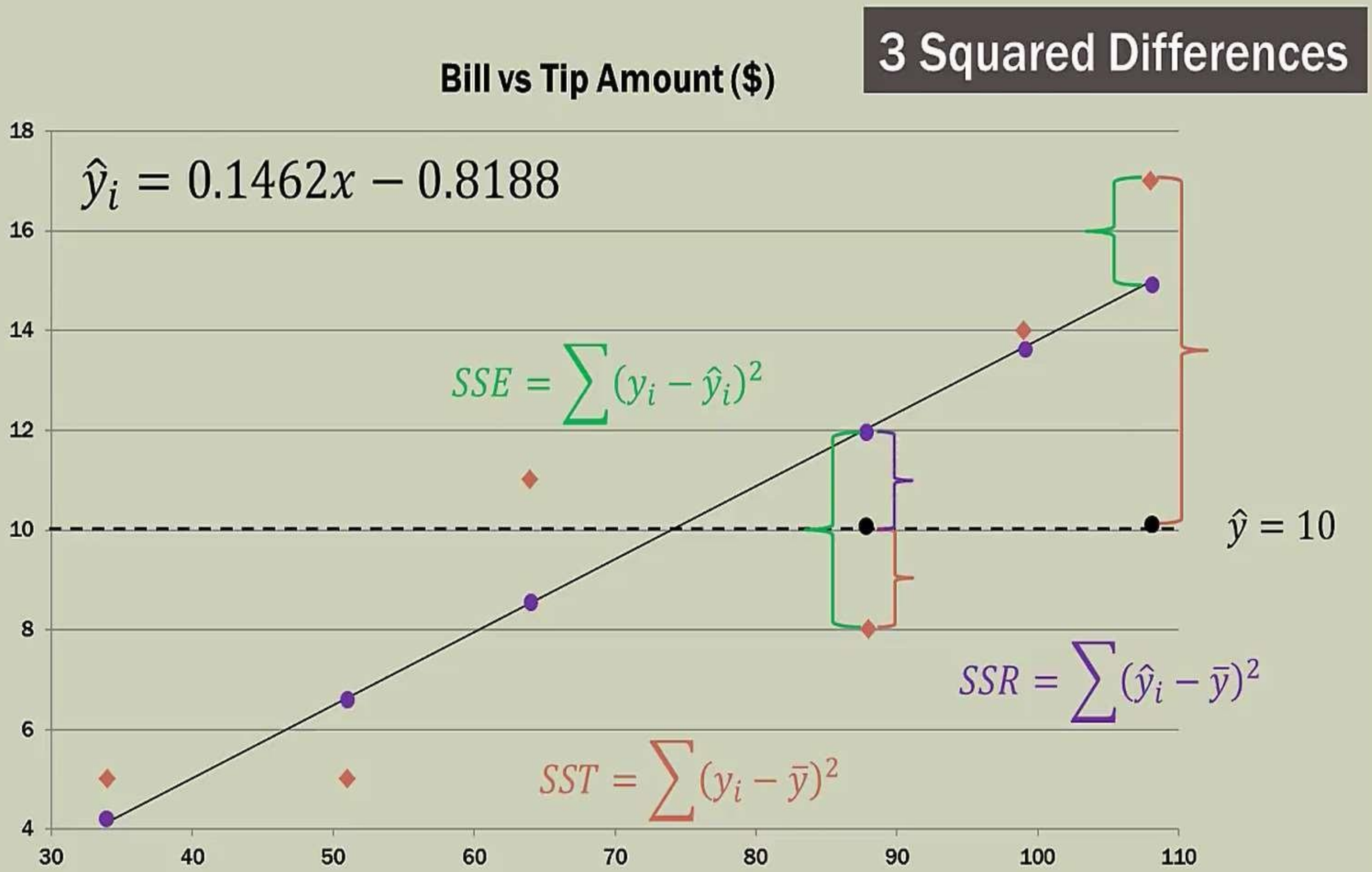
Bill ($)	Tip ($)
34.00	5.00
108.00	17.00
64.00	11.00
88.00	8.00
99.00	14.00
51.00	5.00
$\bar{x} = 74$	$\bar{y} = 10$

- (74,10) is the Centroid.
- We can calculate the linear regression in excel
- For comparison, Excel has calculated the regression equation very close to our manual calculation



Error Metrics

SST = SSR + SSE


3 Squared Differences
Bill vs Tip Amount ($)
$\hat{y}_i = 0.1462x - 0.8188$
$SSE = \sum (y_i - \hat{y}_i)^2$
$SSR = \sum (\hat{y}_i - \bar{y})^2$
$SST = \sum (y_i - \bar{y})^2$
$\bar{y} = 10$
18
16
14
12
10
8
6
4
30
40
50
60
70
80
90
100
110

Coefficient of Determination, R^2

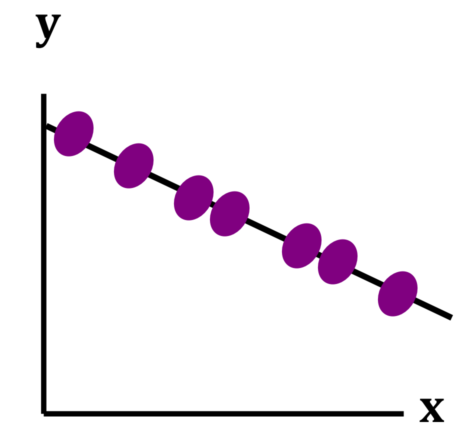
- The coefficient of determination is the portion of the total variation in the dependent variable that is explained by variation in the independent variable

- The coefficient of determination is also called R-squared and is denoted as R^2

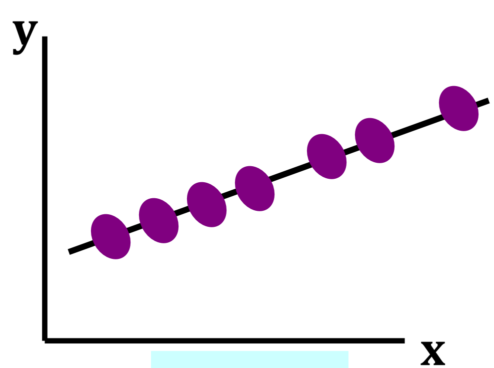
$$R^2 = \frac{SSR}{SST}$$

where $0 \le R^2 \le 1$

Examples of Approximate R^2 Values



$R^2 = 1$



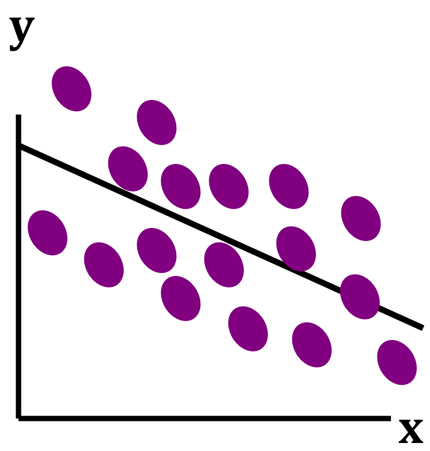
$R^2 = +1$

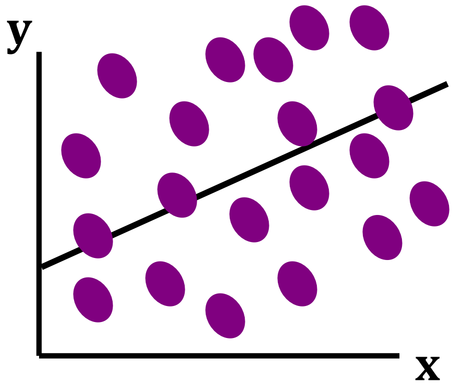
$R^2 = 1$

Perfect linear relationship between x and y:

100% of the variation in y is explained by variation in x

Examples of Approximate R^2 Values





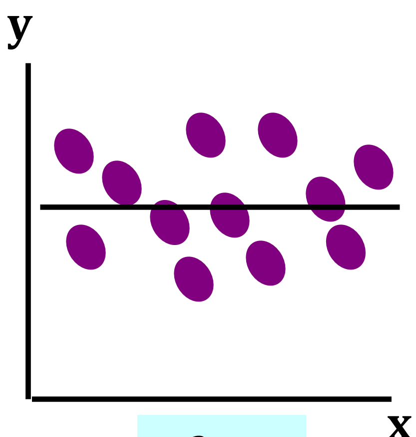
$0 < R^2 < 1$

Weaker linear relationship between x and y:

Some but not all of the variation in y is explained by variation in x

Examples of Approximate R^2 Values

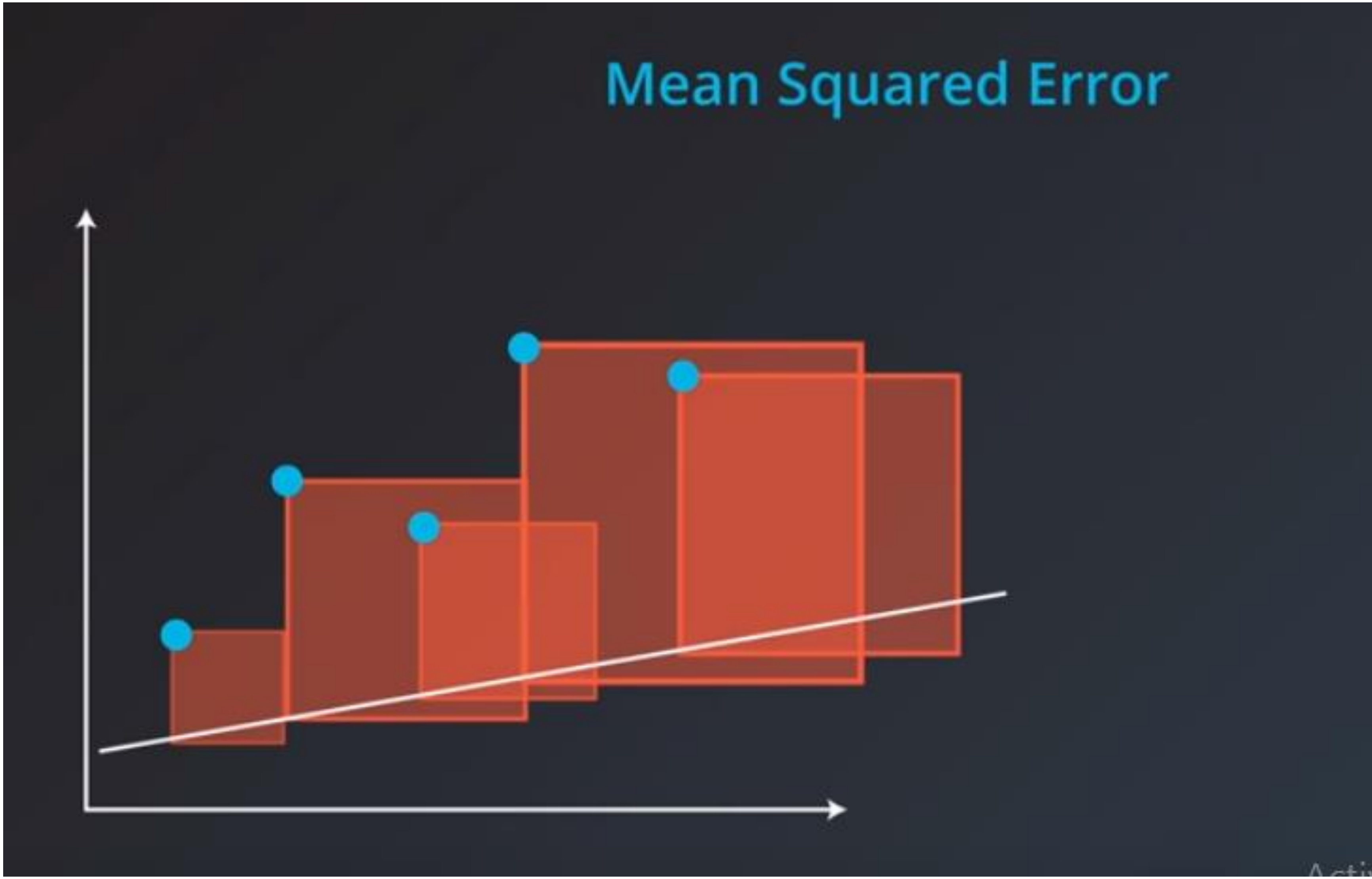
$R^2 = 0$

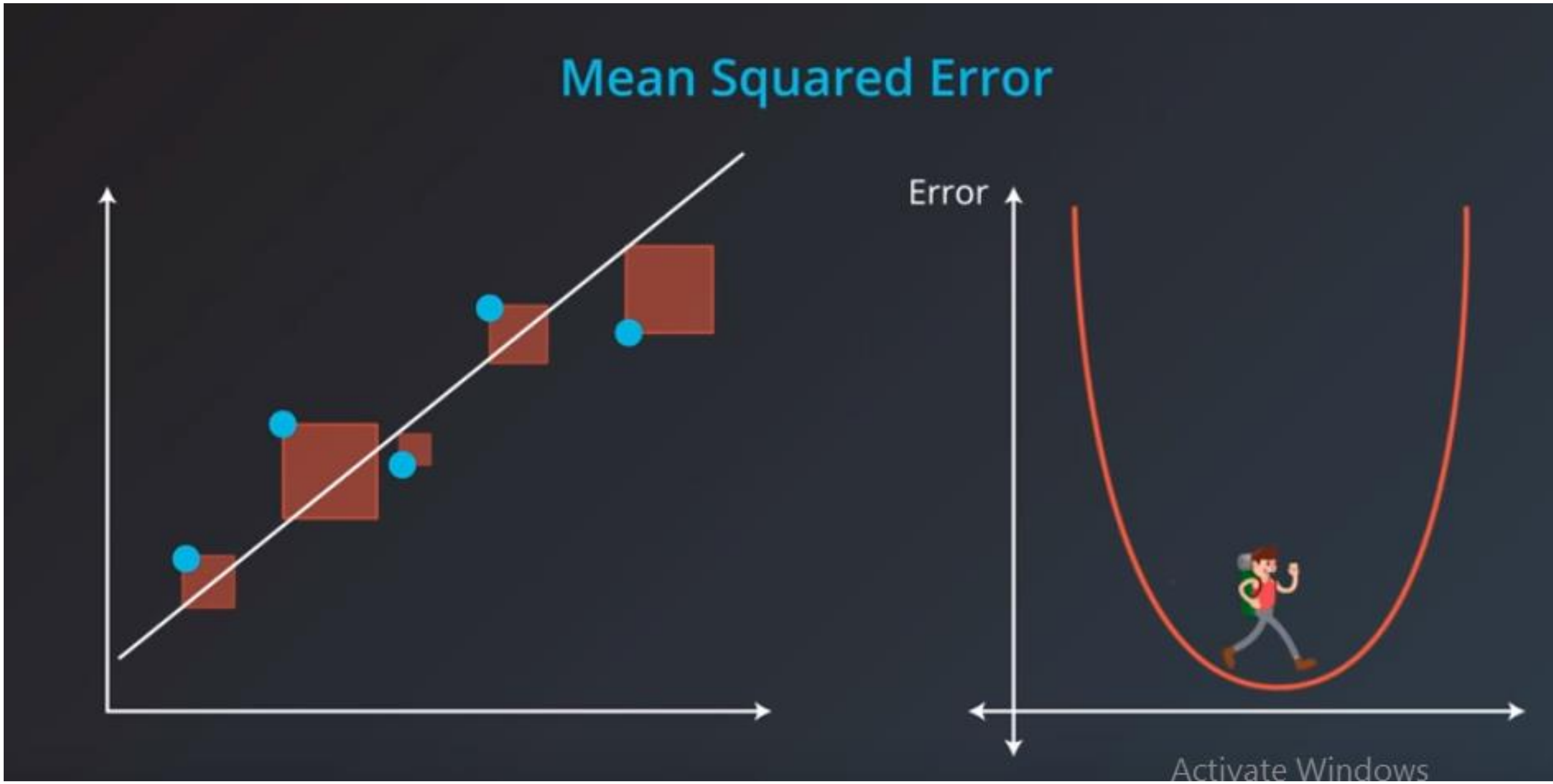


$R^2 = 0$

No linear relationship between x and y:

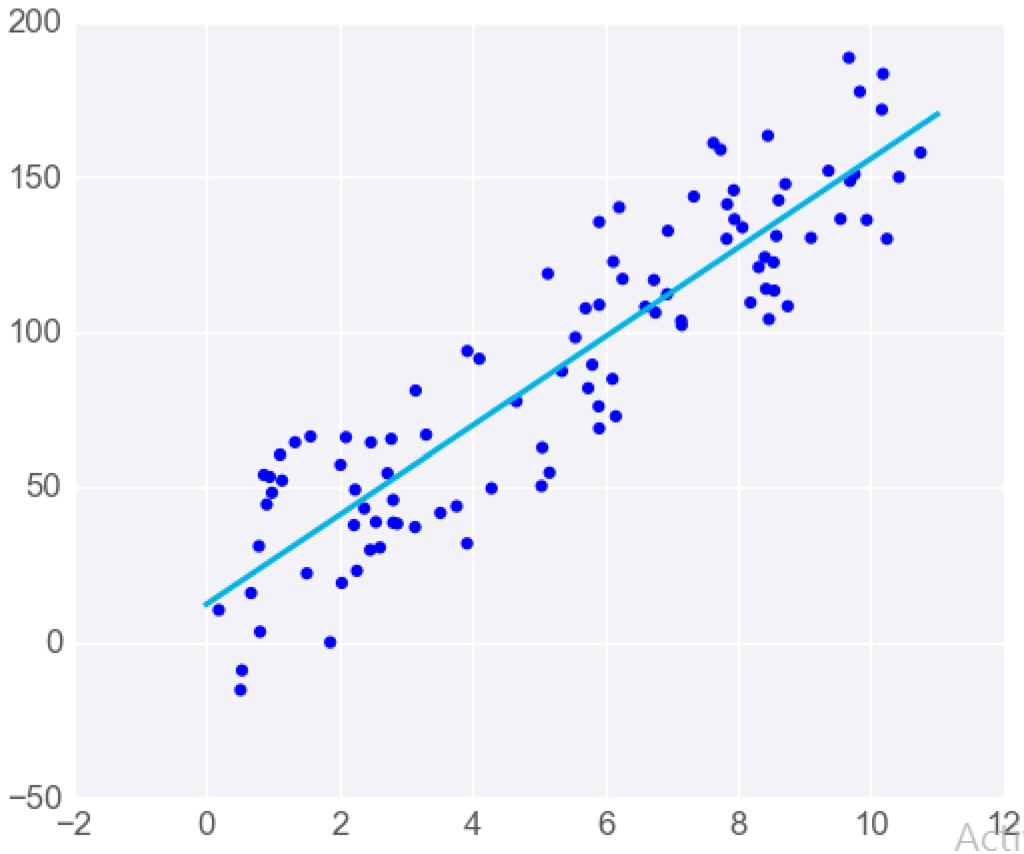
The value of Y does not depend on x. (None of the variation in y is explained by variation in x)


Mean Squared Error


Mean Squared Error
Error
Activate Windows

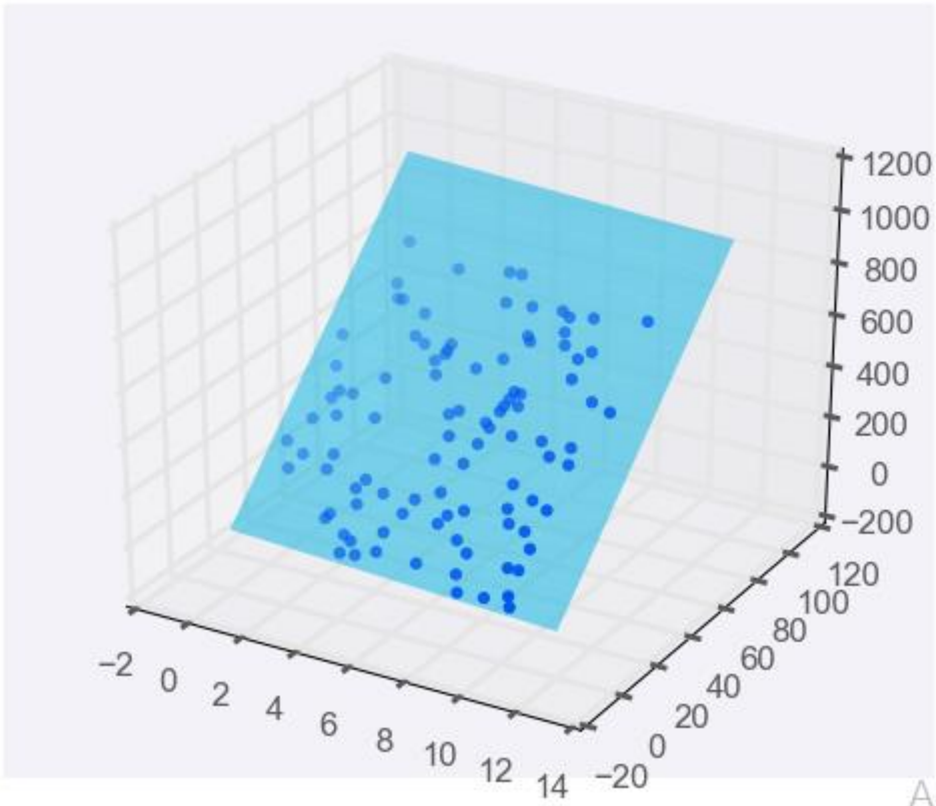
Visualization in N dimensions

Linear Regression – 1 Variable



Linear regression with one predictor variable

Linear Regression – 2 Variable

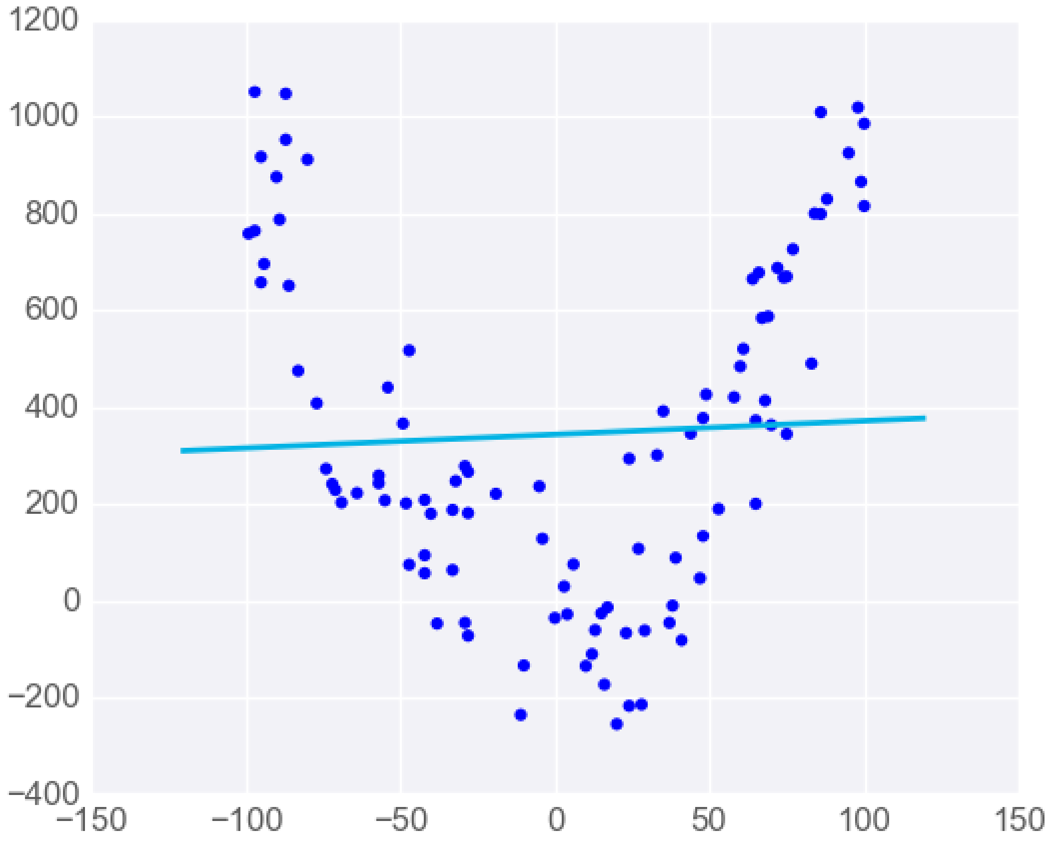


Linear regression with two predictor variables

When to use Linear Regression ?

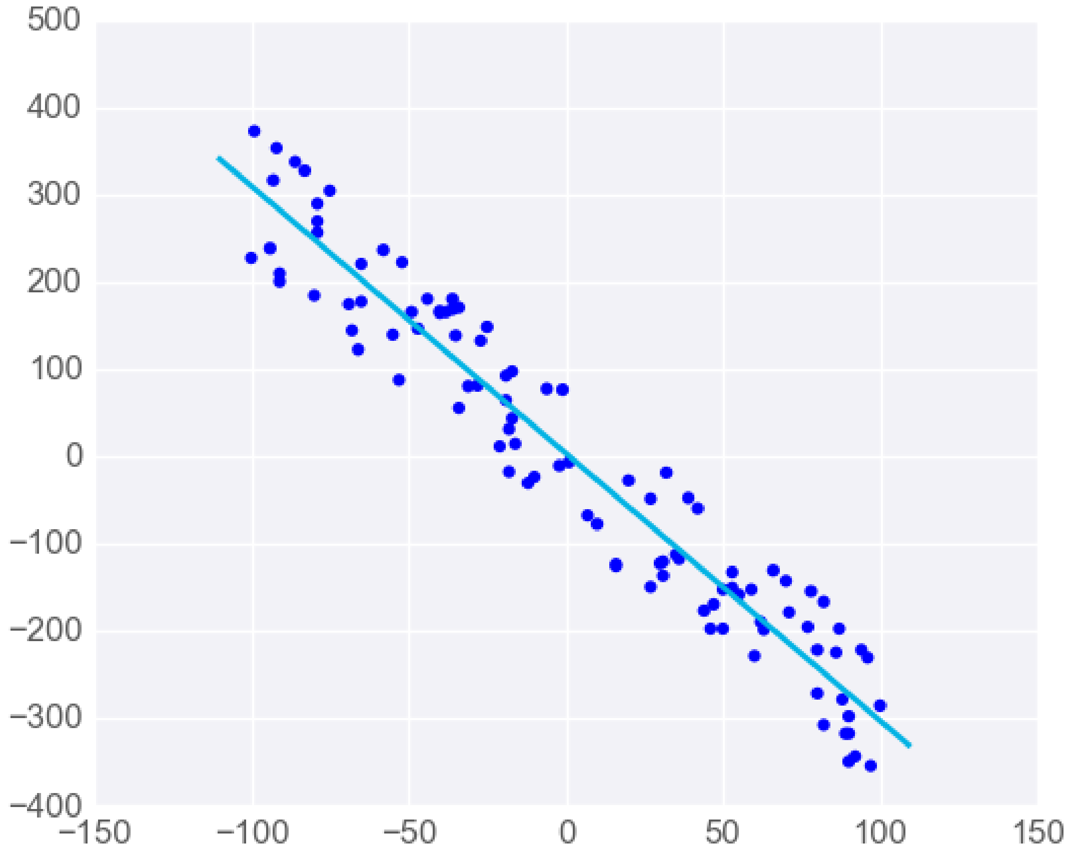
Linear Regression Warnings

Linear Regression Works Best When the Data is Linear



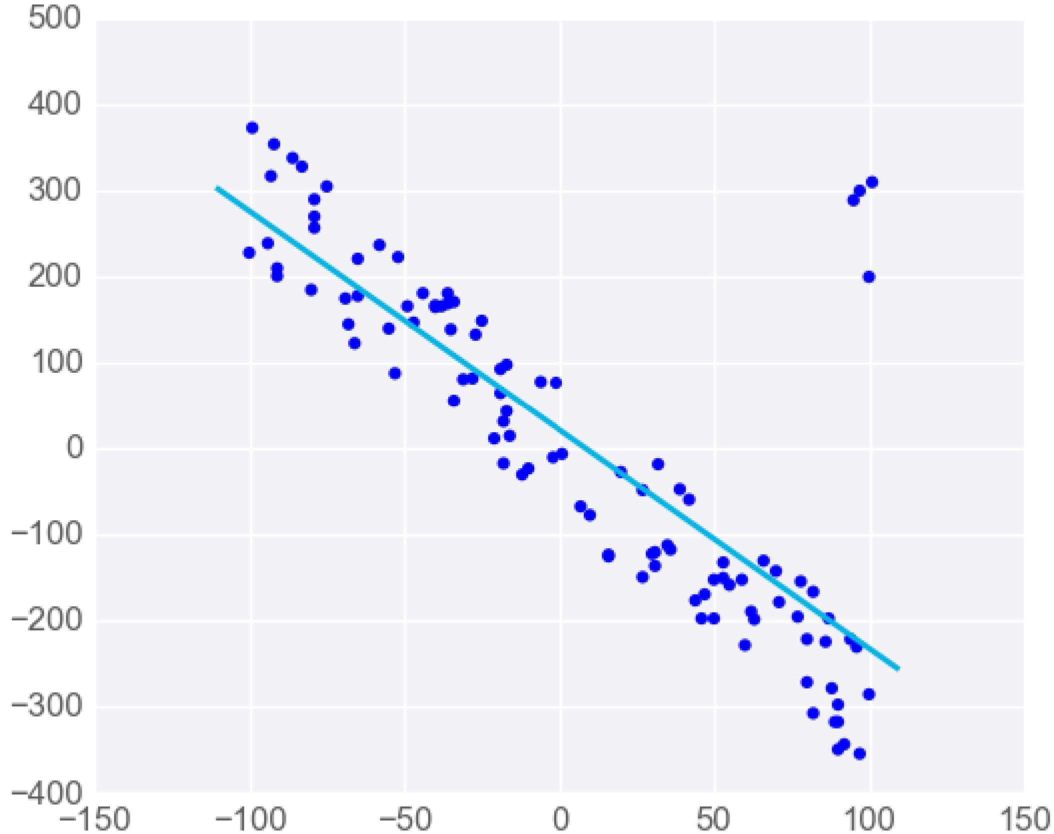
Linear Regression Warnings

Linear Regression is Sensitive to Outliers



Linear Regression Warnings

Linear Regression is Sensitive to Outliers

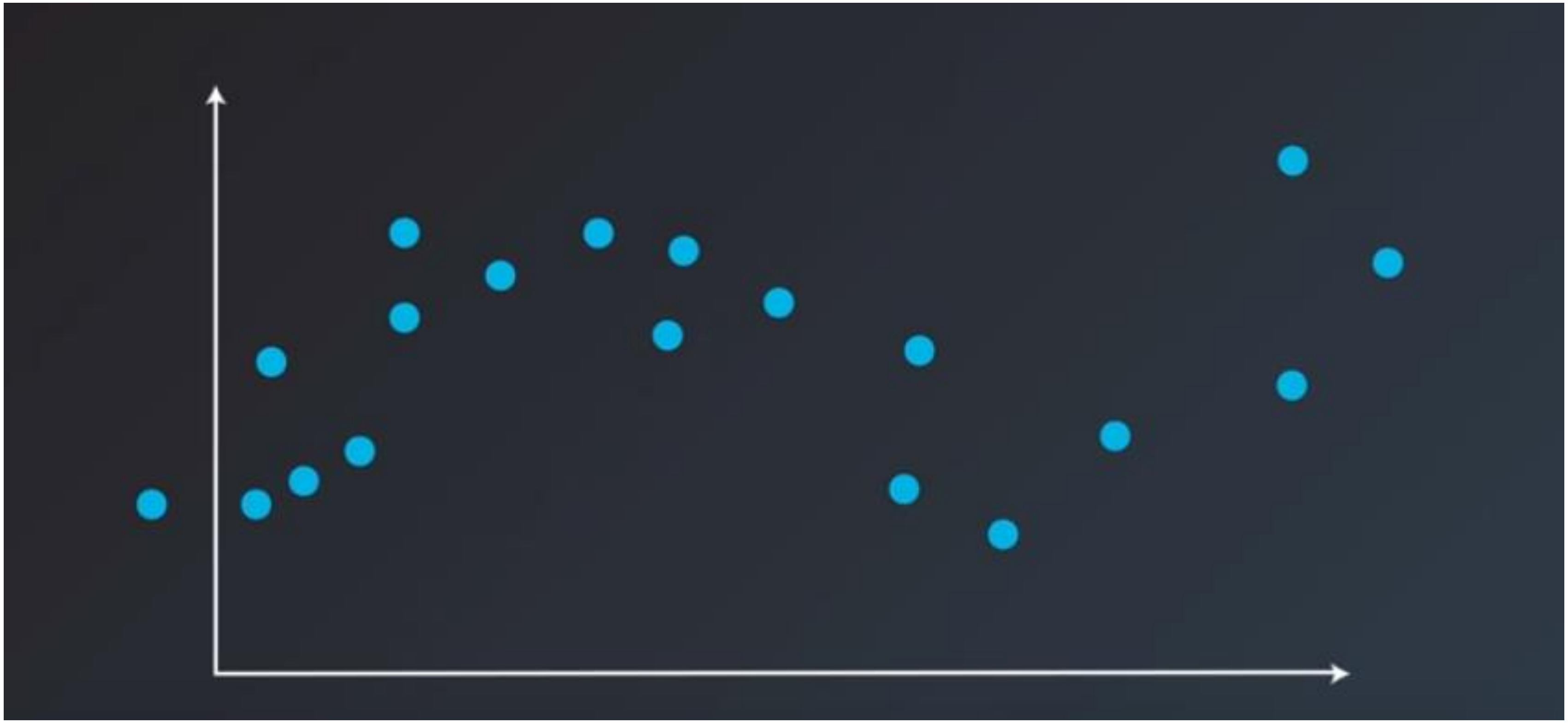


Linear Regression

- Extended in case of Non Linearity

Polynomial Regression

Polynomial Regression



Polynomial Regression