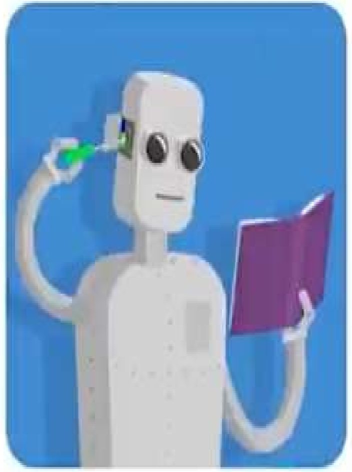
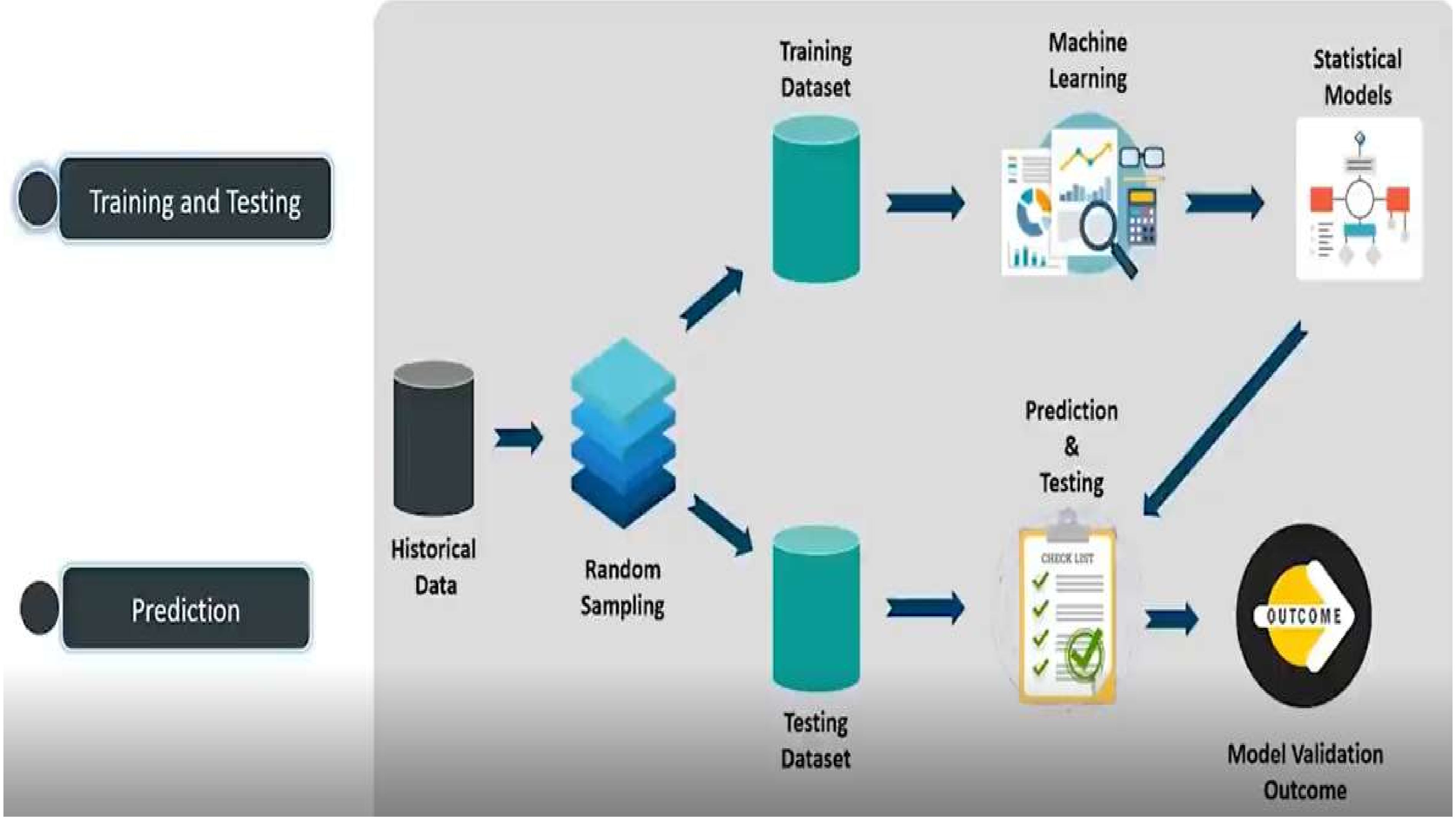
Supervised Learning

Supervised learning is where you have input variables (x) and an output variable (Y) and you use an algorithm to learn the mapping function from the input to the output



It is called Supervised Learning because the process of an algorithm learning from the training dataset can be thought as a teacher supervising the learning process

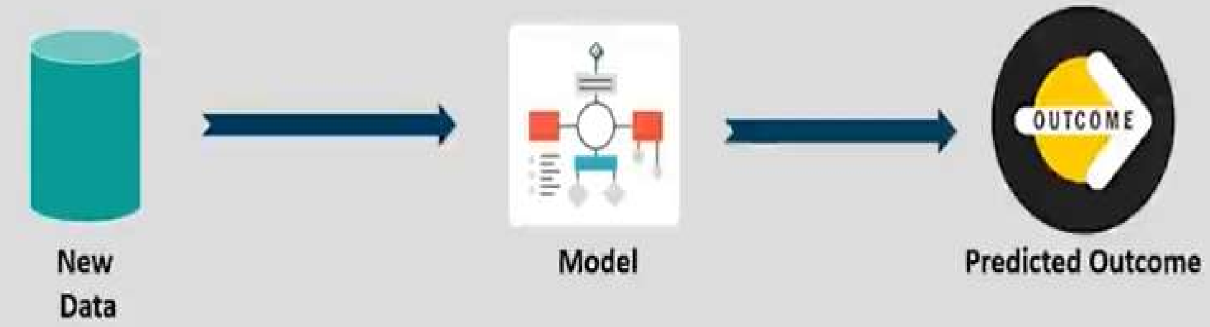
Supervised Learning



Supervised Learning

Training and Testing

Prediction



Supervised Learning Algorithms

- Linear Regression
- Logistic Regression
- Decision Tree
- Random Forest
- Naïve Bayes Classifier


What is
Regression?
“Regression analysis is a form of predictive modelling technique which investigates the relationship between a dependent and independent variable”

Uses of Regression

Three major uses for regression analysis are

- Determining the strength of predictors
- Forecasting an effect, and
- Trend forecasting

Linear vs Logistic Regression

Basis	Linear Regression	Logistic Regression
Core Concept	The data is modelled using a straight line	The probability of some obtained event is represented as a linear function of a combination of predictor variables.
Used with	Continuous Variable	Categorical Variable
Output/Prediction	Value of the variable	Probability of occurrence of event
Accuracy and Goodness of fit	measured by loss, R squared, Adjusted R squared etc.	Accuracy, Precision, Recall, F1 score, ROC curve, Confusion Matrix, etc

Linear Regression

Selection Criteria

- Classification and Regression Capabilities
- Data Quality
- Computational Complexity
- Comprehensible and Transparent

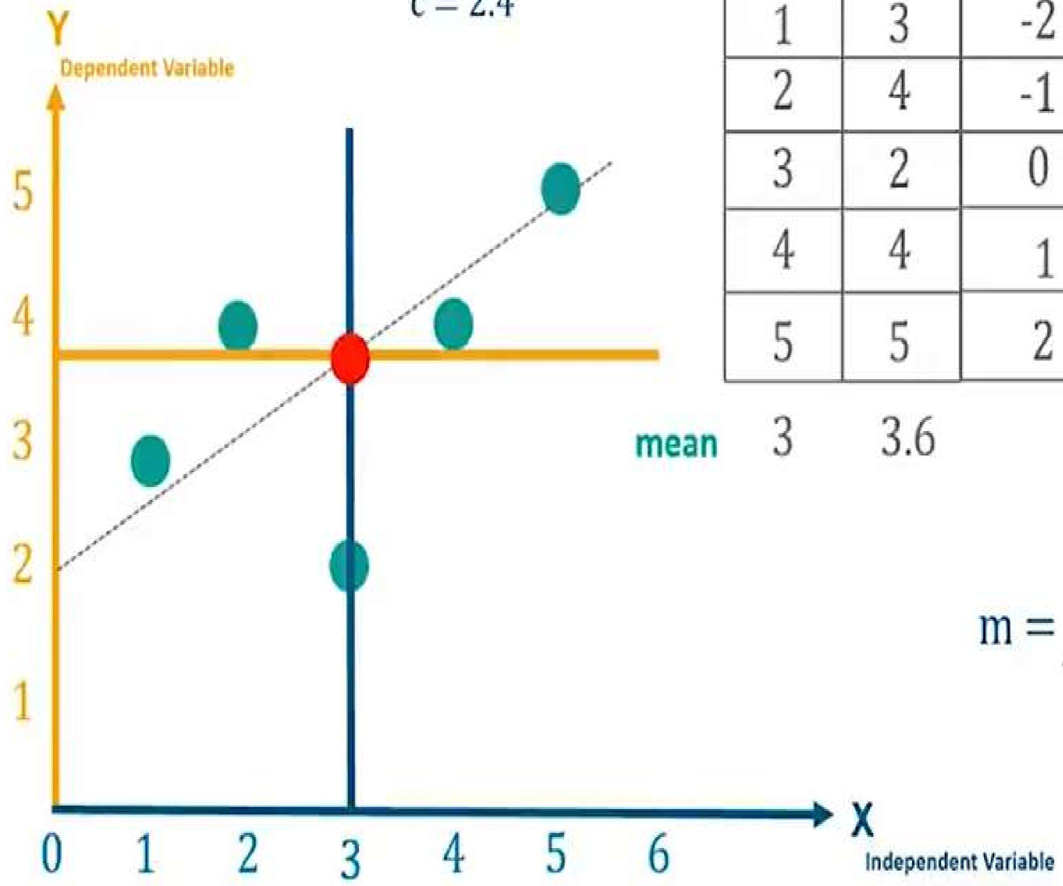
Where is Linear Regression used?

- Evaluating Trends and Sales Estimates
- Analyzing the Impact of Price Changes
- Assessment of risk in financial services and insurance domain

Understanding Linear Regression Algorithm

$y = mx + c$

$c = 2.4$



	x	y	$x - \dot{x}$	$y - \dot{y}$	$(x - \dot{x})^2$	$(x - \dot{x})(y - \dot{y})$
	1	3	-2	-0.6	4	1.2
	2	4	-1	0.4	1	-0.4
	3	2	0	-1.6	0	0
	4	4	1	0.4	1	0.4
	5	5	2	1.4	4	2.8
mean	3	3.6			$\Sigma = 10$	$\Sigma = 4$

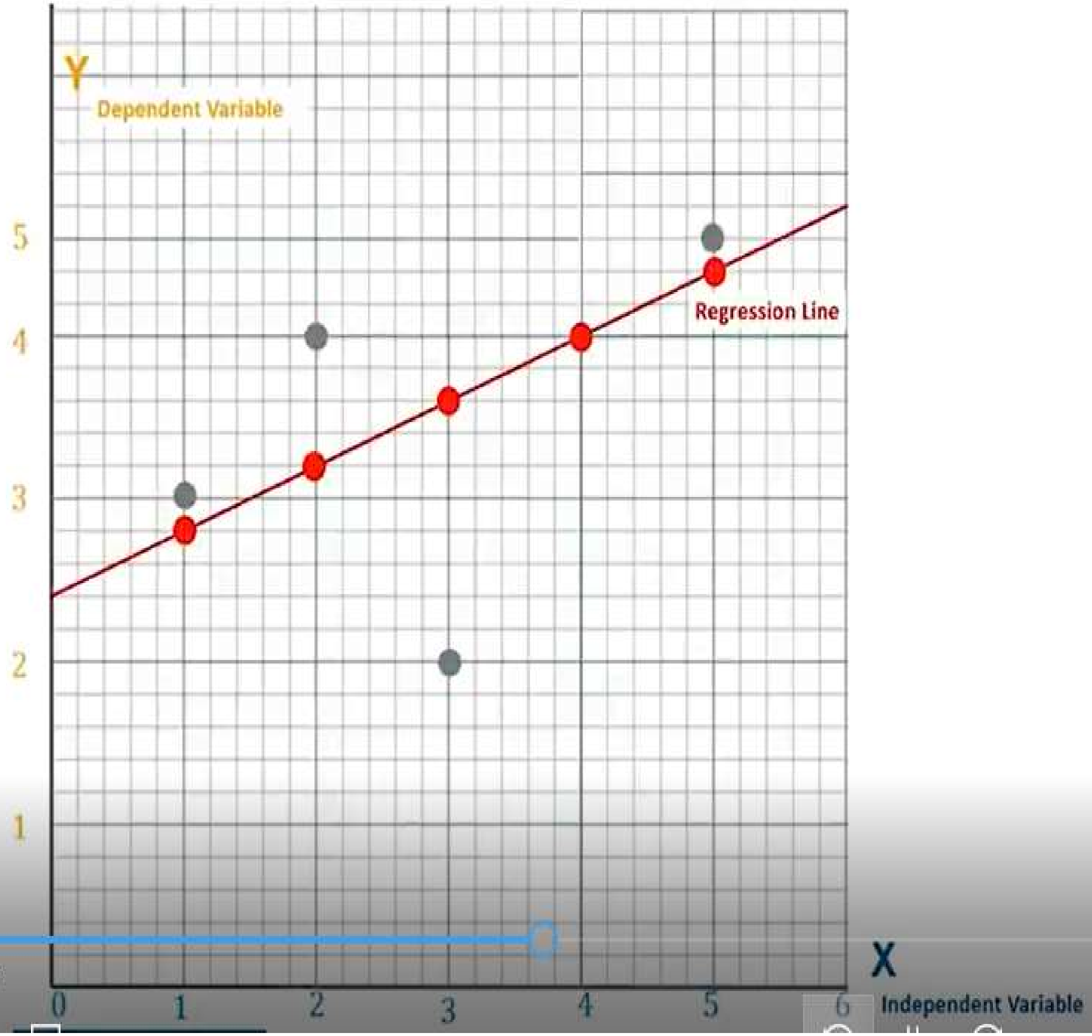
$$m = \sum \frac{(x - \dot{x})(y - \dot{y})}{(x - \dot{x})^2} = \frac{4}{10}$$

$m = 0.4$

$c = 2.4$

$y = 0.4x + 2.4$

Mean Square Error

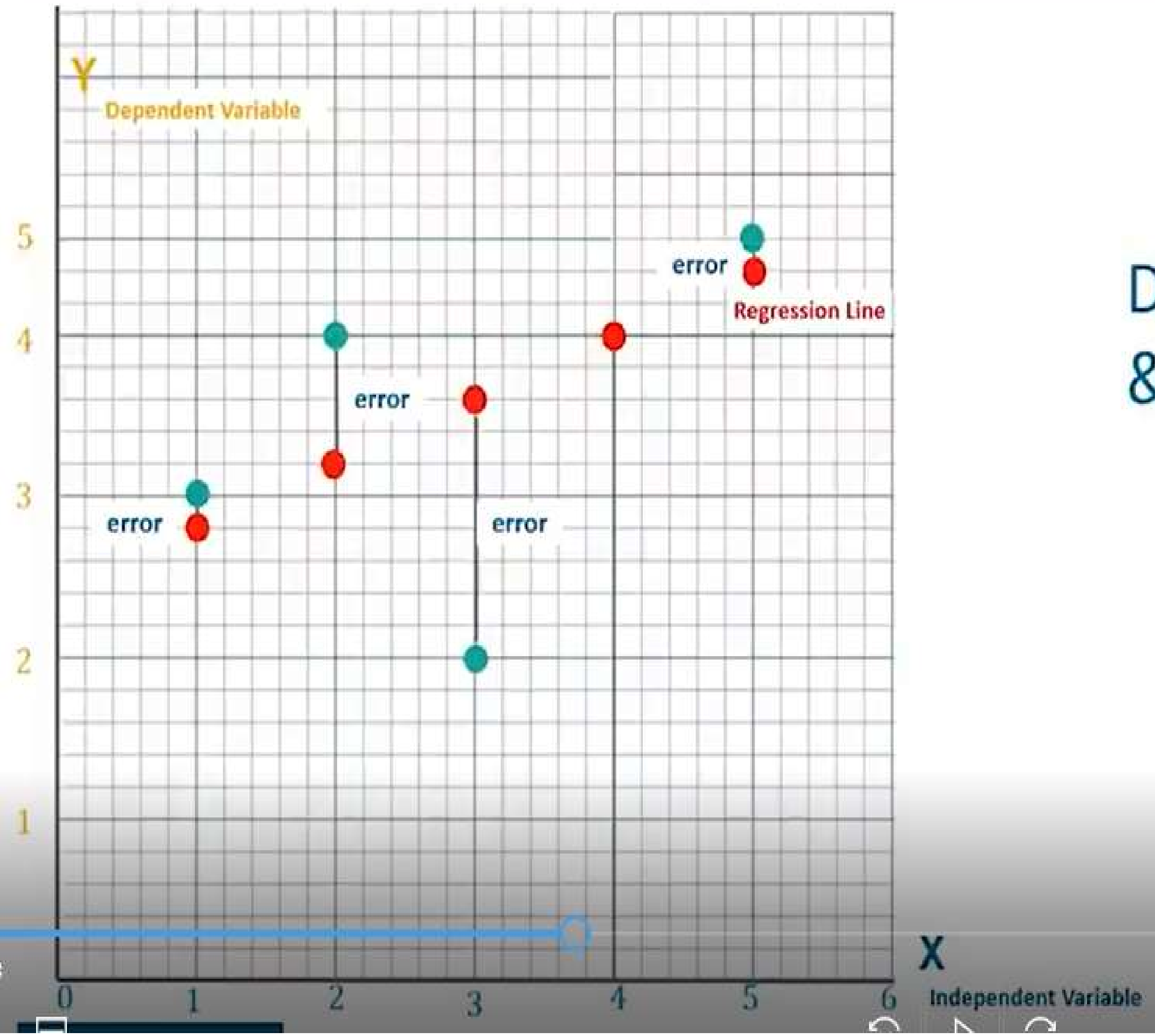


$m = 0.4$
$c = 2.4$
$y = 0.4x + 2.4$

For given $m = 0.4$ & $c = 2.4$, lets predict values for y for $x = \{1,2,3,4,5\}$

$y = 0.4 \times 1 + 2.4 = 2.8$
$y = 0.4 \times 2 + 2.4 = 3.2$
$y = 0.4 \times 3 + 2.4 = 3.6$
$y = 0.4 \times 4 + 2.4 = 4.0$
$y = 0.4 \times 5 + 2.4 = 4.4$

Mean Square Error

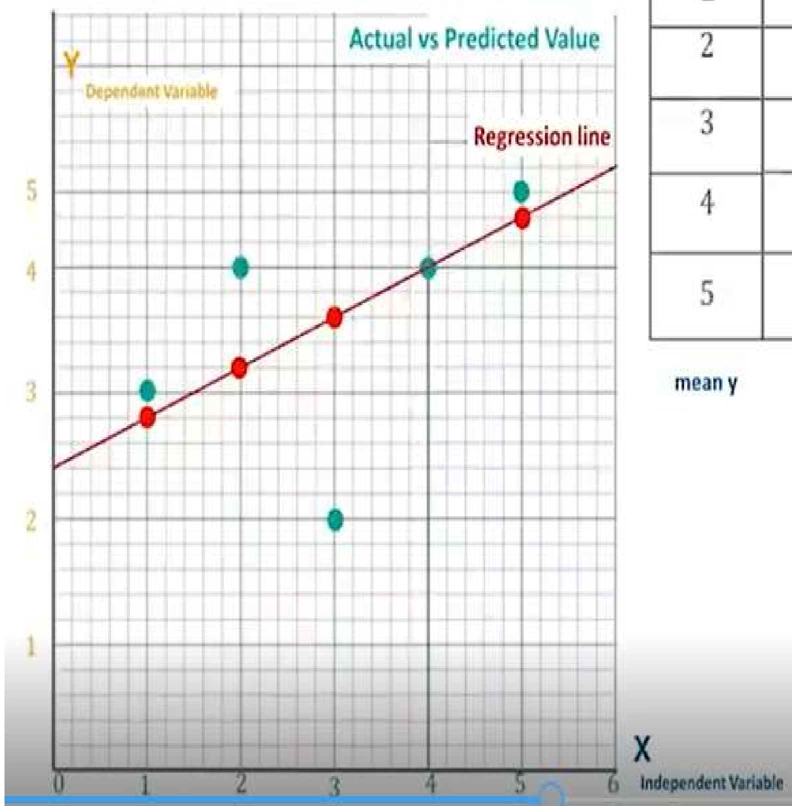


Distance between actual & predicted value

What is R-Square?

- **R-squared** value is a statistical measure of how close the data are to the fitted regression line
- It is also known as **coefficient of determination**, or the **coefficient of multiple determination**

Calculation of R^2



x	y	$y - y$	$(y - y)^2$	y_p	$(y_p - y)$	$(y_p - y)^2$
1	3	- 0.6	0.36	2.8	-0.8	0.64
2	4	0.4	0.16	3.2	-0.4	0.16
3	2	−1.6	2.56	3.6	0	0
4	4	0.4	0.16	4.0	0.4	0.16
5	5	1.4	1.96	4.4	0.8	0.64
mean y	3.6		$\sum$ 5.2			$\sum$ 1.6

$$R^2 = \frac{1.6}{5.2} = \frac{\sum \left((y_p - \bar{y}\right)^2}{\sum (y - \dot{y})^2}$$