Unit III Syllabus

- **Bias, Variance, Generalization, Under-fitting and Over-fitting**
- **Linear Regression**
- **Regression: Lasso Regression and Ridge Regression**
- **Gradient Descent Algorithm and SGD (Over and Above)**
- **Evaluation Metrics: MAE, RMSE and R2**

Errors in Machine Learning

Error:

- In Machine Learning, error is used to see how accurately our model can predict on data it uses to learn; as well as new, unseen data.
- Based on our error, we choose the machine learning model which performs best for a particular dataset.

In supervised machine learning an algorithm learns a model from training data.

Input Data

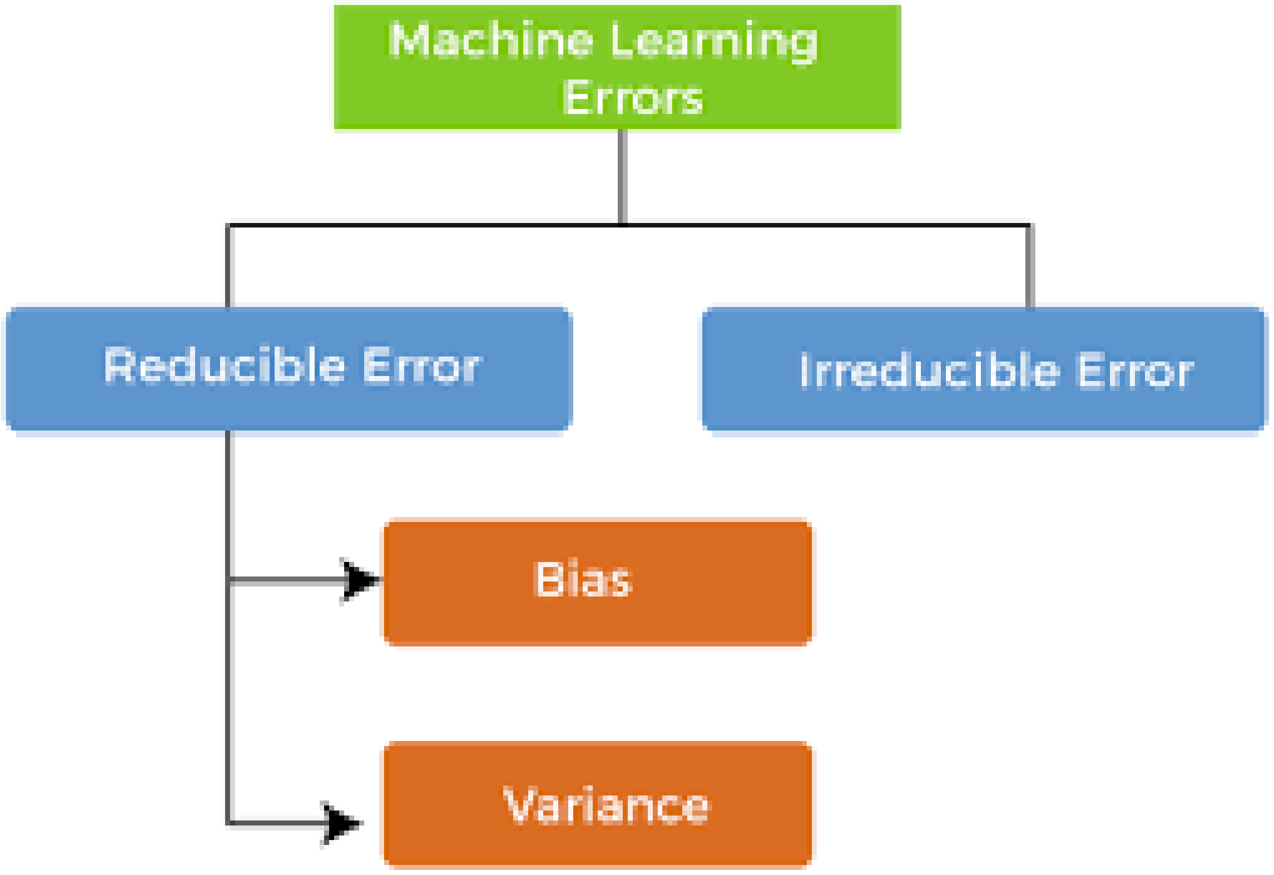
Output Variable or Target ← $$Y = f(X) + E$$ → **Prediction Error**

Mapping Function

Errors in Machine Learning

Error:

- In Machine Learning, error is used to see how accurately our model can predict on data it uses to learn; as well as new, unseen data.
- Based on our error, we choose the machine learning model which performs best for a particular dataset.



Bias

- **Bias:**

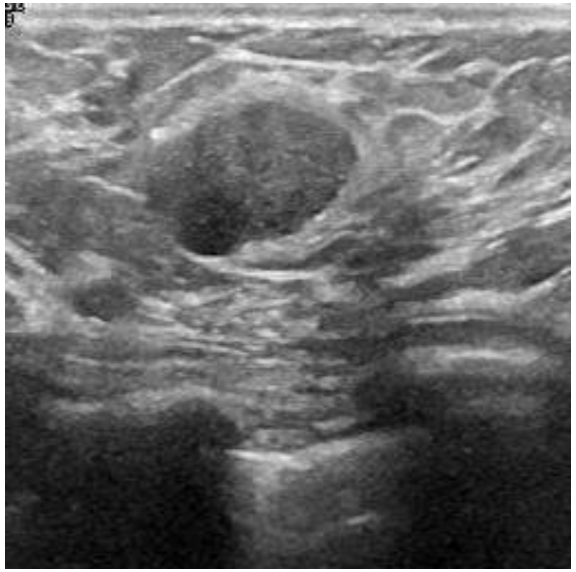
  It is the error/difference between average model prediction and the **actual values**/ground truth.

X1	X2	X3	X4	Predicted Values (Y)	Actual Values	Bias
23	1.2	2	3	3.1	4	0.9
23	1.4	3	4	3.6	4	0.4
45	1.1	1	8	4.4	5	0.6
56	1.0	4	9	3.8	4	0.2
12	1.7	5	2	4.5	6	1.5
34	1.9	4	1	5.9	7	1.1
Average				**4.22**	**5**	**0.78**

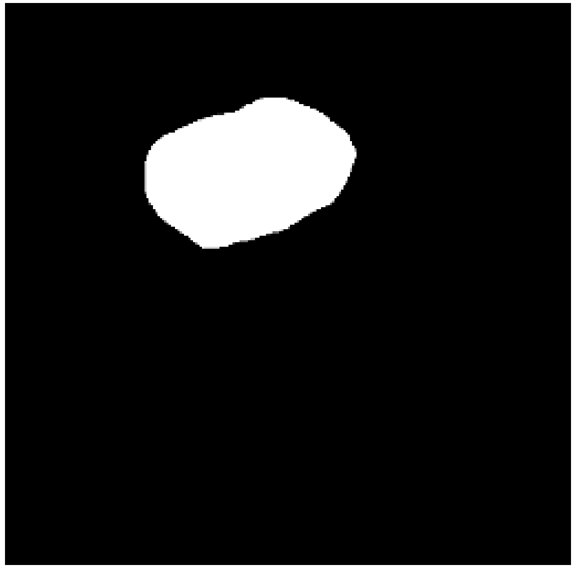
Bias

- **Bias:**

  It is the error/difference between average model prediction and the actual values/**ground truth**.



Input Image

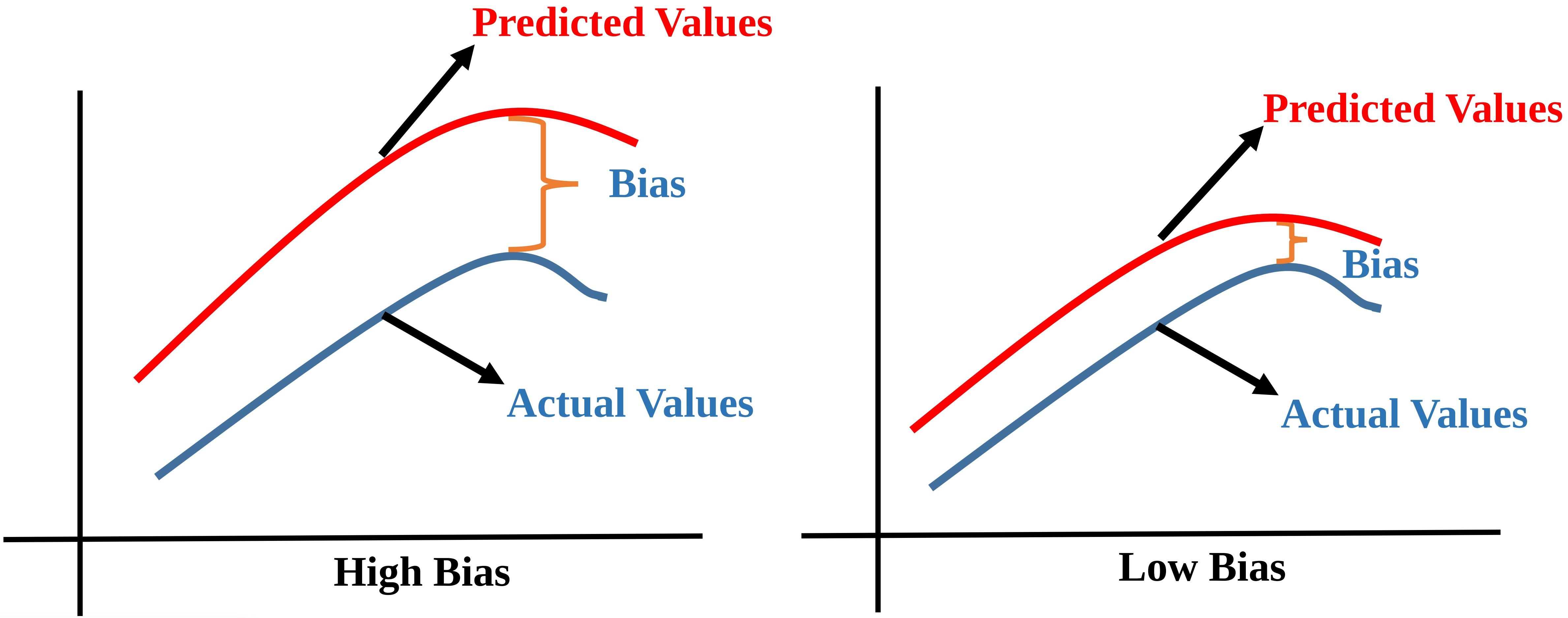


Ground Truth

Predicted Image

Bias

- A model with a **higher bias** would not match the data set closely.
- A **low bias** model will closely match the training data set.



Signs of a High Bias ML Model

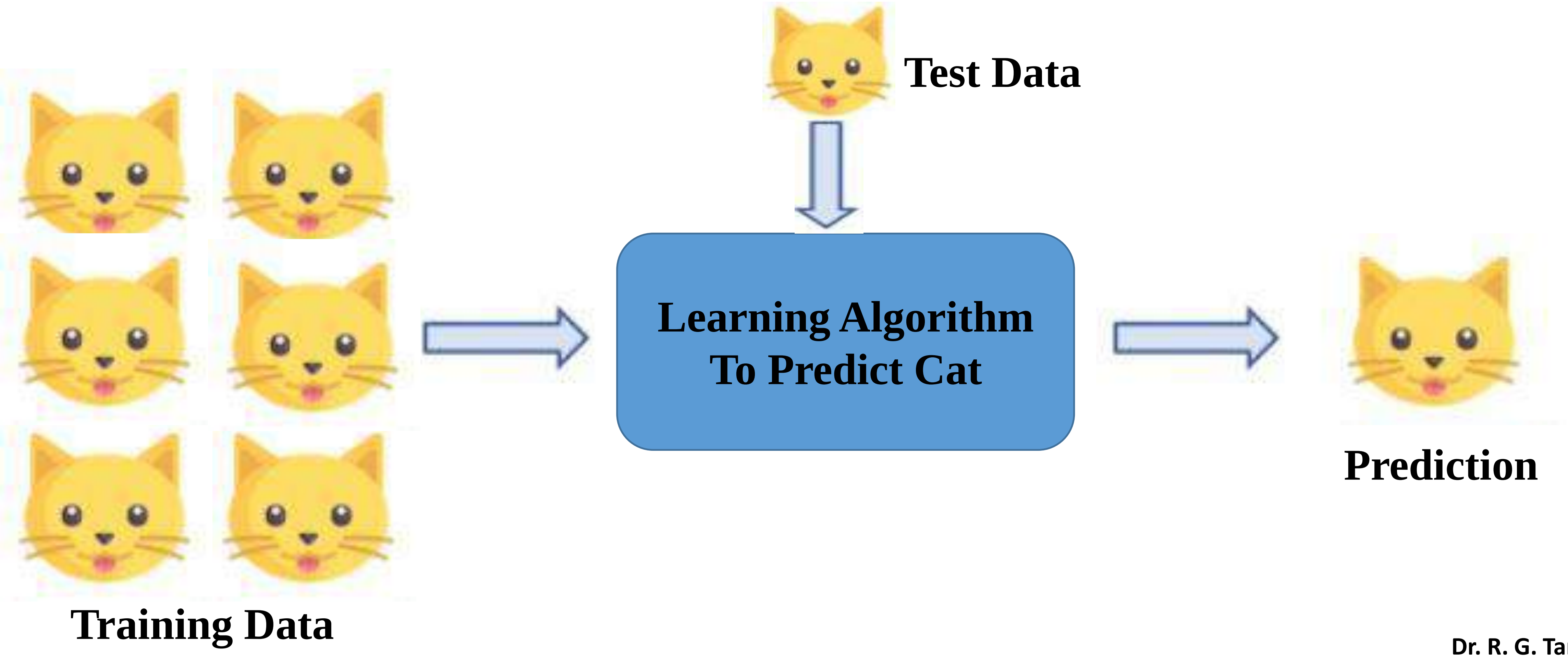
- Failure to capture data trends
- Underfitting
- Overly simplified
- High error rate

Variance

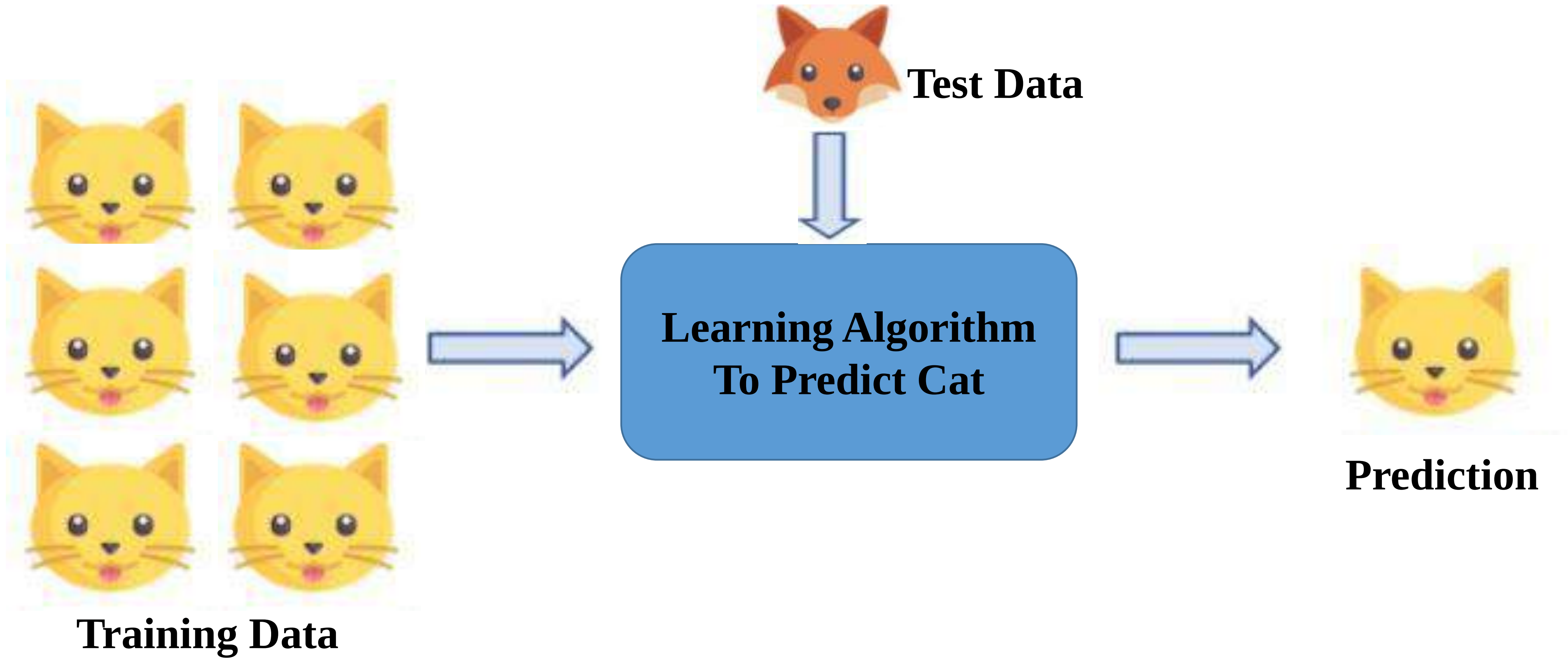
- **Variance:**

  Variance refers to the changes in the model when using different portions of the training data set.

- Models with **high bias** will have **low variance**.
- Models with **high variance** will have a **low bias**.

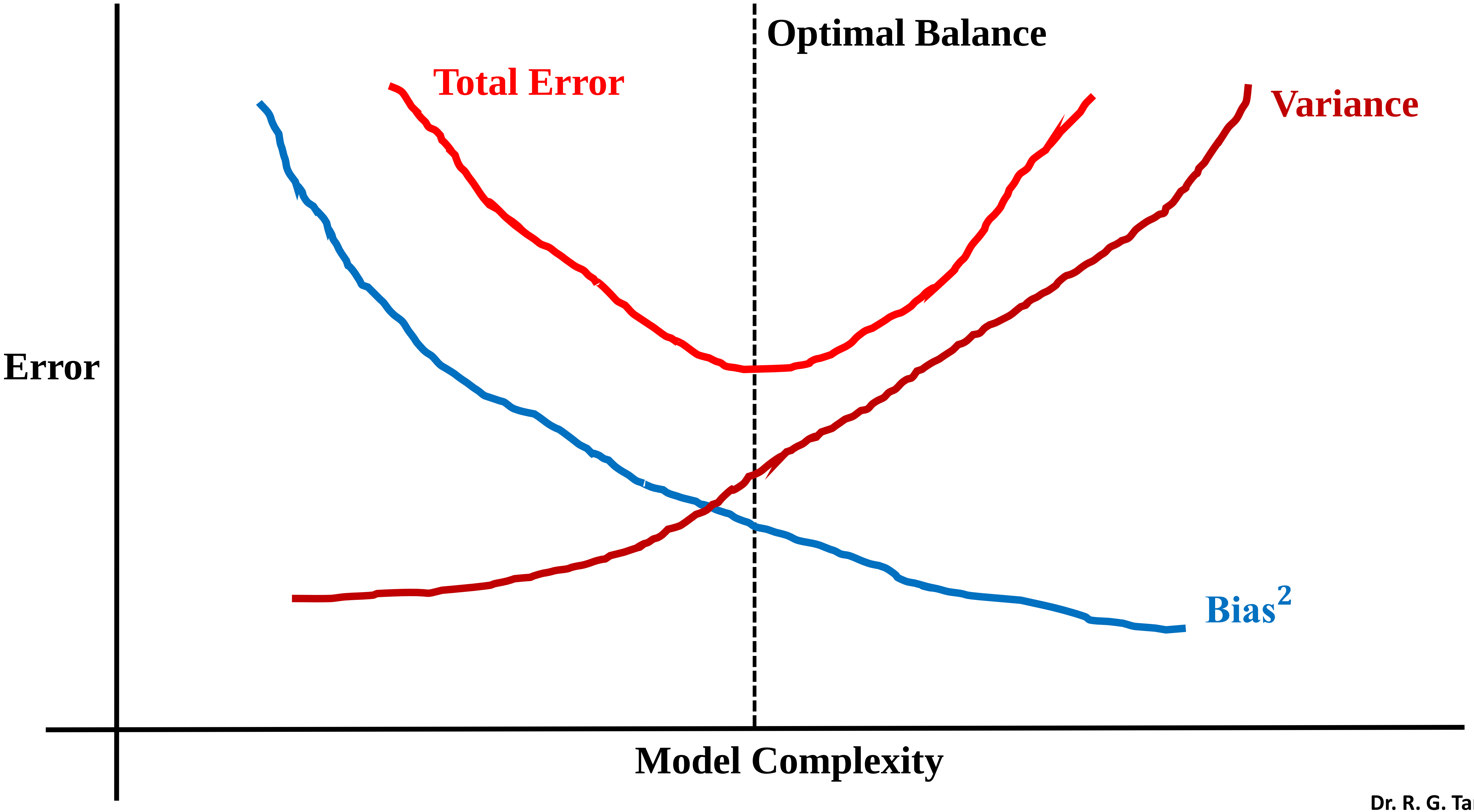


Variance



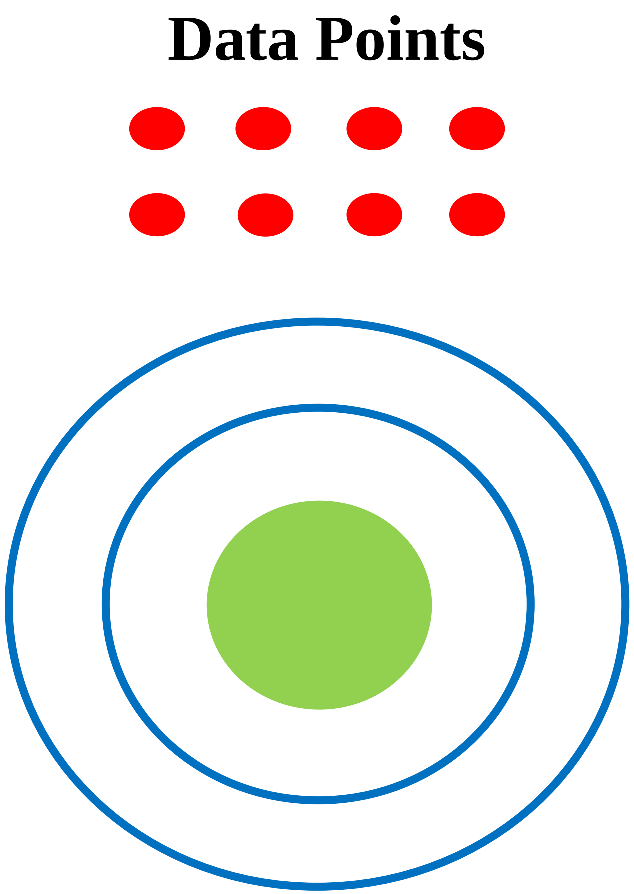
- Signs of a High Variance ML Model
- Noise in data set
- Overfitting
- Complexity
- Forcing data points together

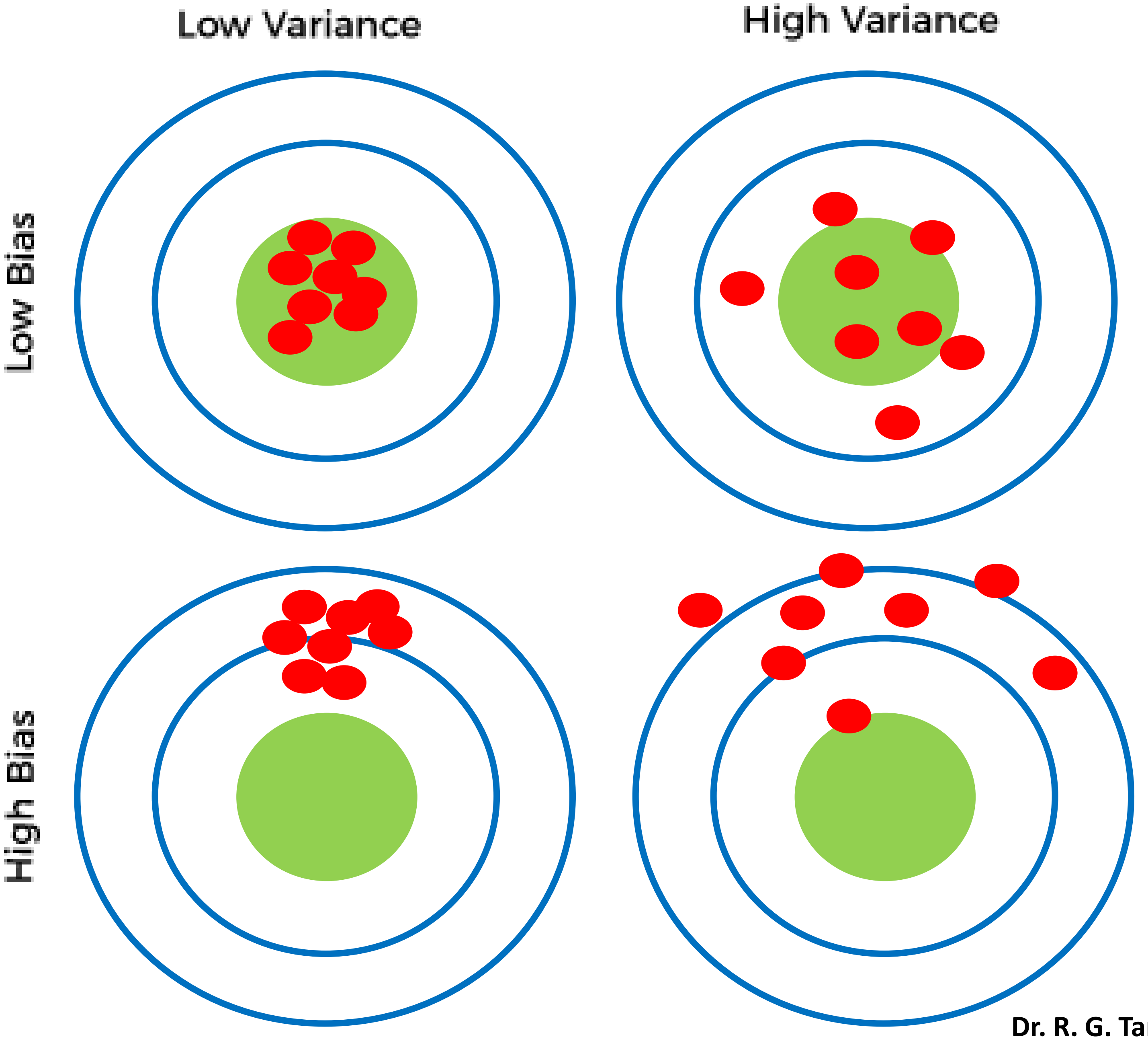
Bias and Variance Trade-off



Dr. R. G. Tambe

Bias and Variance Trade-off



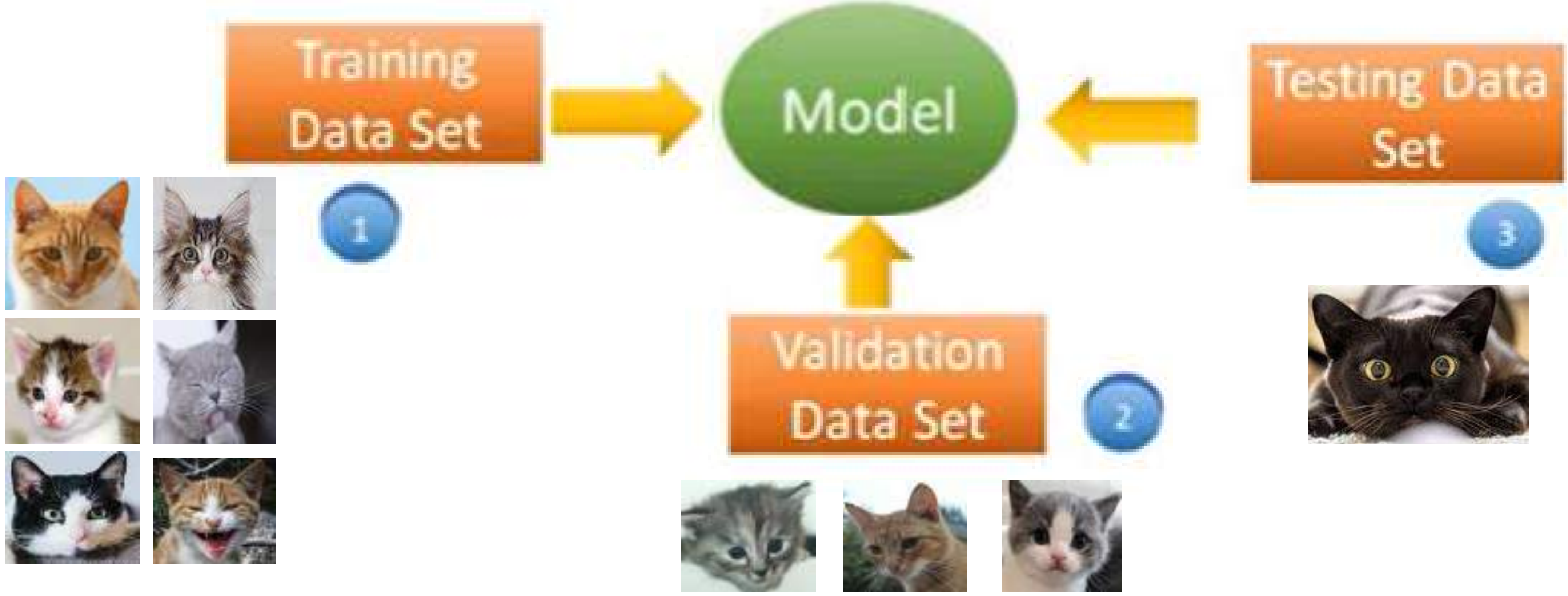


Dr. R. G. Tambe

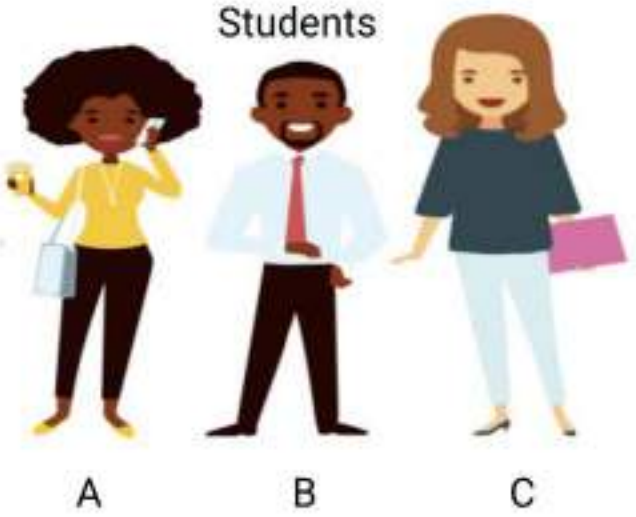
Generalization

A model trained on the training set, predicts the right output for new instances is called **Generalization**

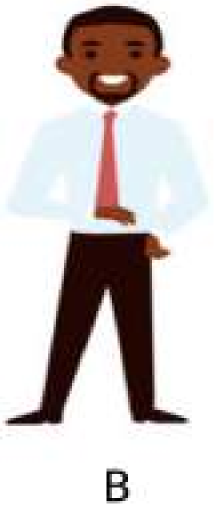
- The goal of a good machine learning model is to generalize well from the training data to any data from the problem domain.
- This allows us to make predictions in the future on data the model has never seen.



Overfitting and Underfitting



- Hobby = chating
- Not interested in class
- Doesn't pay much attention to professor

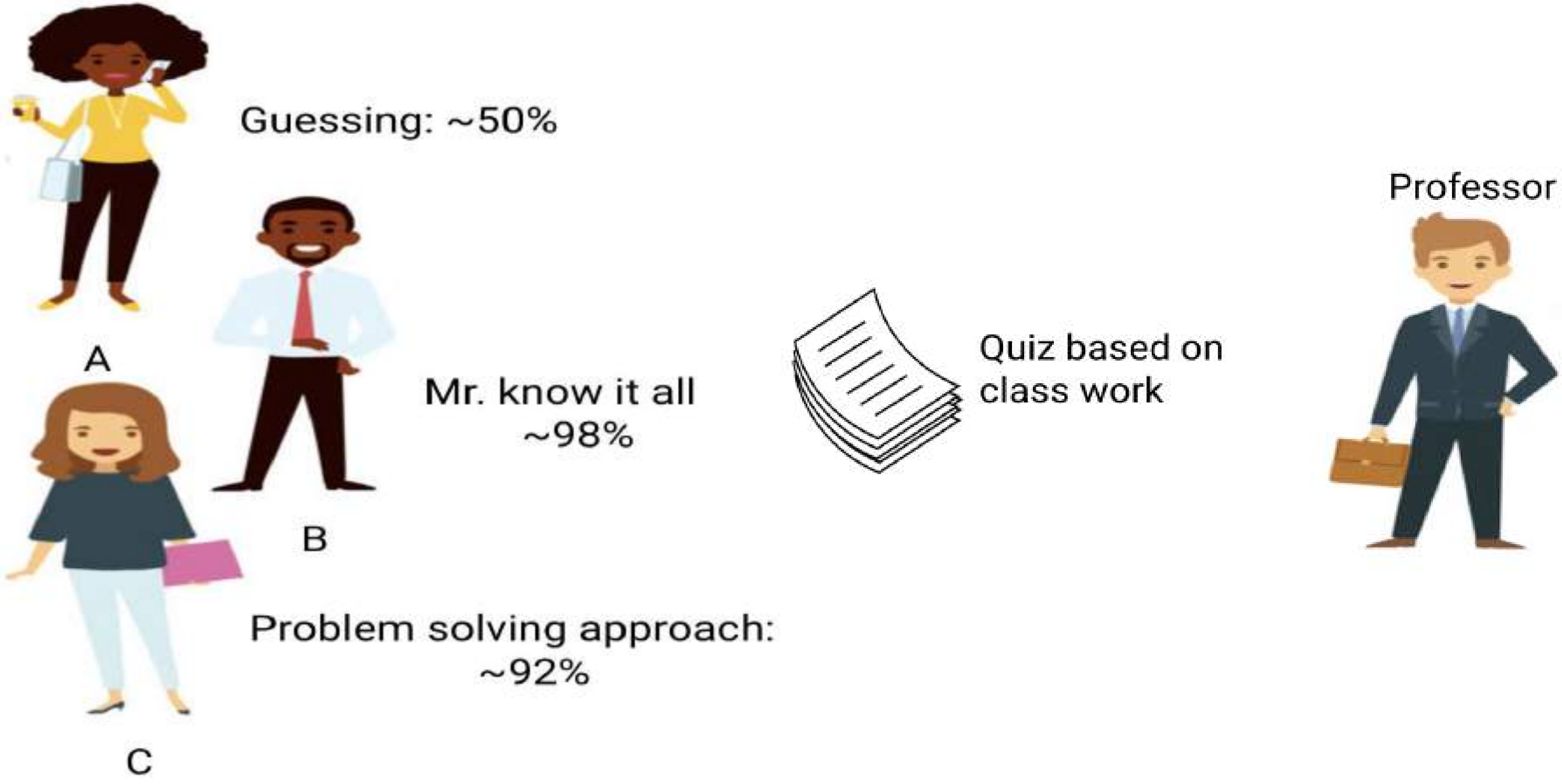


- Hobby = to be best in class.
- Mugs up everything professor says.
- Too much attention to the class work.

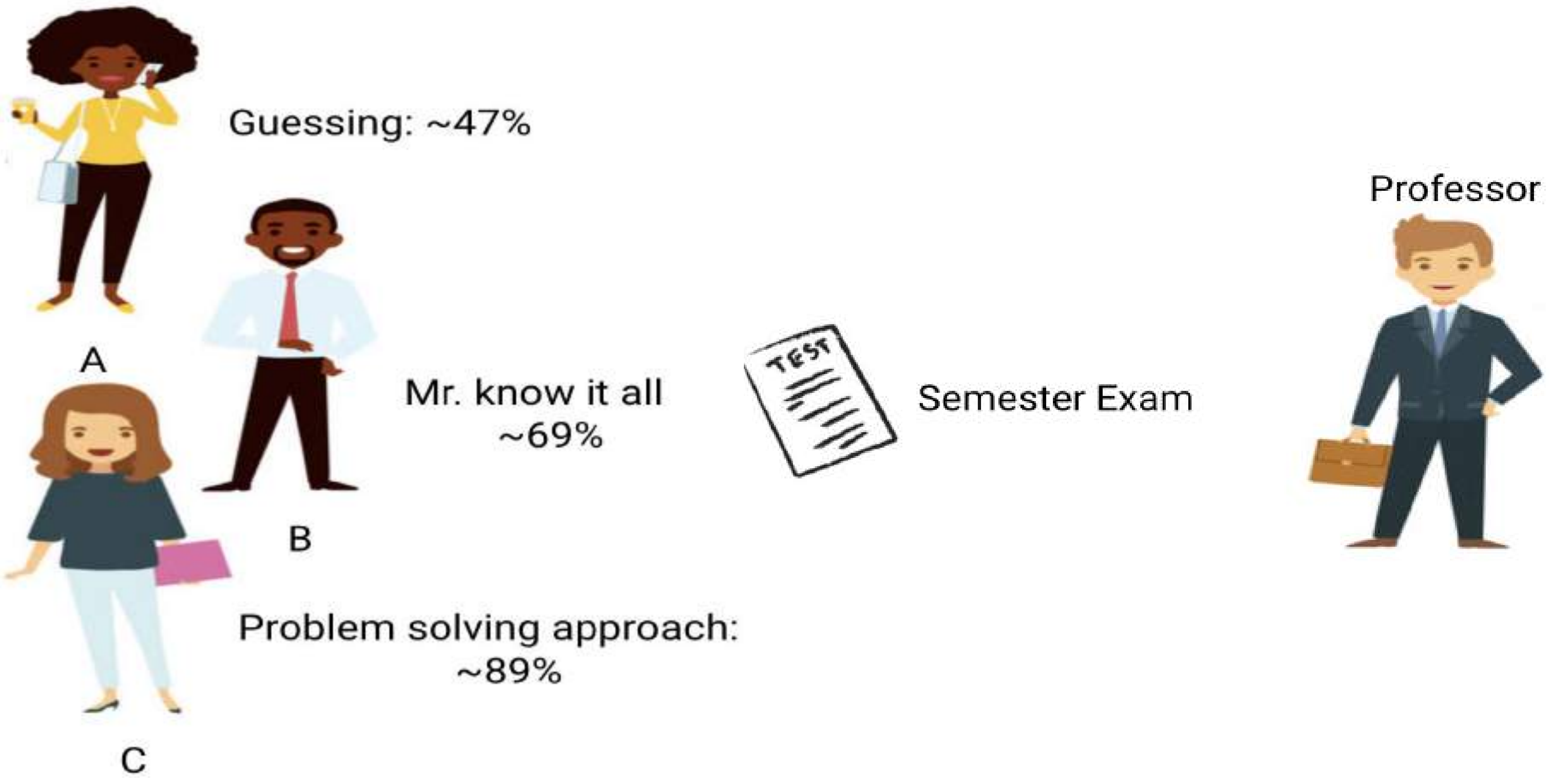
- Hobby = learning new things
- Eager to learn concepts.
- Pays attention to class and learns the idea behind solving a problem.

Dr. R. G. Tambe

Overfitting and Underfitting



Overfitting and Underfitting



Overfitting and Underfitting

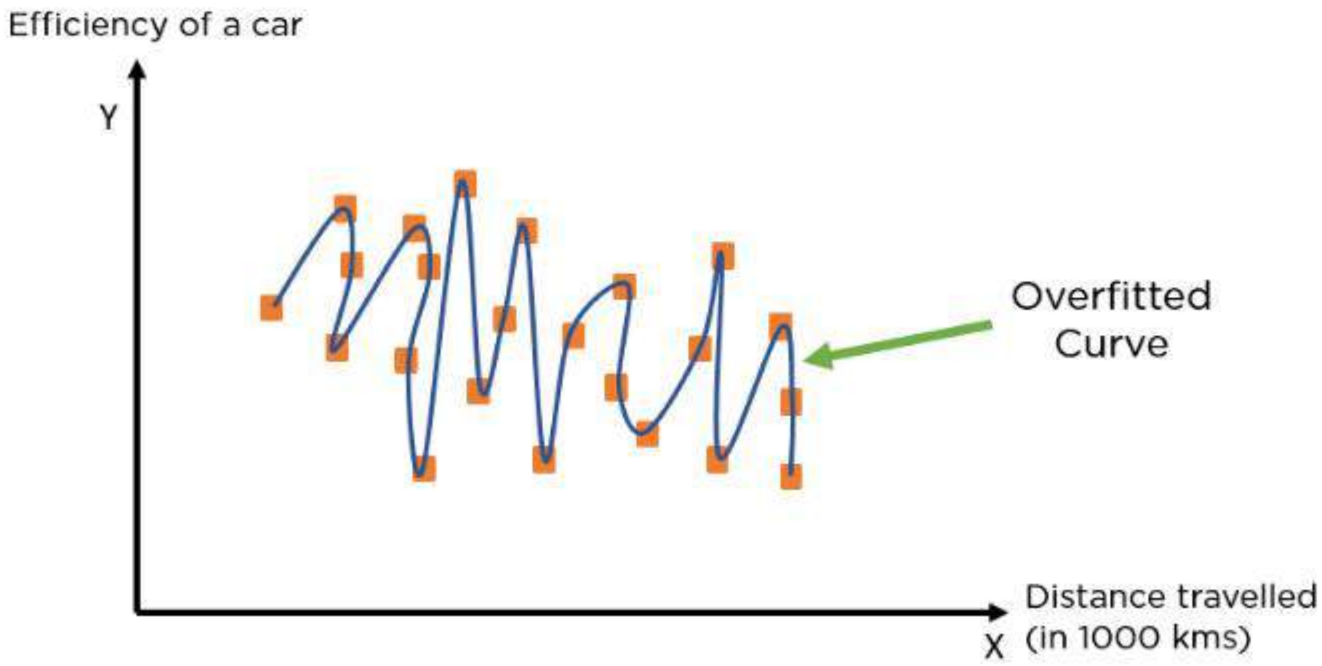
		
A	B	C
Not interested in learning	Memorizing the lessons	Conceptual Learning
Class test ~50% Test ~47%	Class test ~98% Test ~69%	Class test ~92% Test ~89%
Underfit	**Overfit**	**Best Fit**

Dr. R. G. Tambe

Overfitting and Underfitting

Overfitting

- When a model performs very well for training data but has poor performance with test data (new data), it is known as **overfitting**.
- In this case, the machine learning model learns the details and noise in the training data such that it negatively affects the performance of the model on test data.



Low Bias and High Variance

Overfitting and Underfitting

Overfitting

- When a model performs very well for training data but has poor performance with test data (new data), it is known as **overfitting**.
- In this case, the machine learning model learns the details and noise in the training data such that it negatively affects the performance of the model on test data.

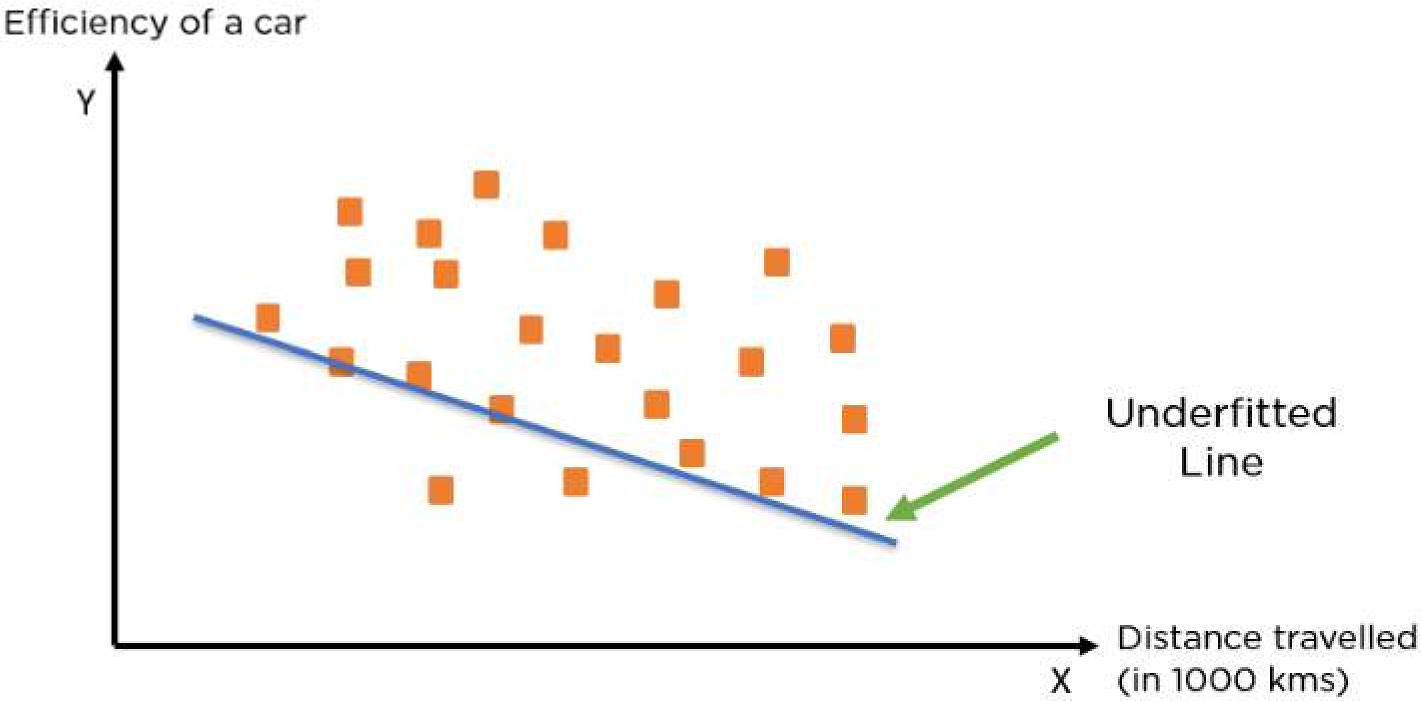
Reasons for Overfitting

- Data used for training is not cleaned and contains noise (garbage values) in it.
- The model has a high variance.
- The size of the training dataset used is not enough.
- The model is too complex.

Overfitting and Underfitting

Underfitting

- When a model has not learned the patterns in the training data well and is unable to generalize well on the new data, it is known as **underfitting**.
- An underfit model has poor performance on the training data and will result in unreliable predictions.



High Bias and Low Variance

Overfitting and Underfitting

Underfitting

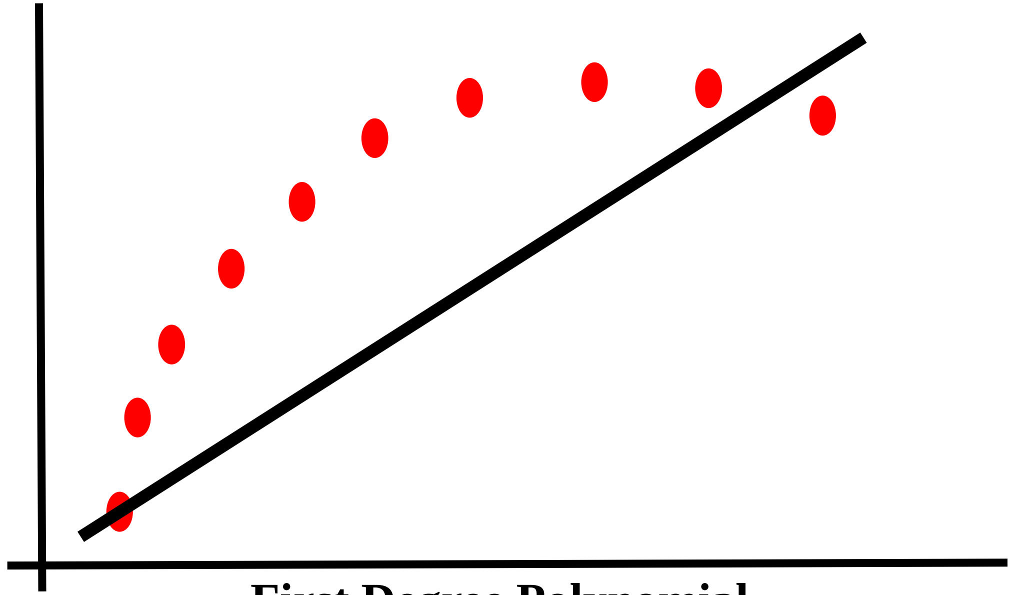
- When a model has not learned the patterns in the training data well and is unable to generalize well on the new data, it is known as **underfitting**.
- An underfit model has poor performance on the training data and will result in unreliable predictions.

Reasons for Underfitting

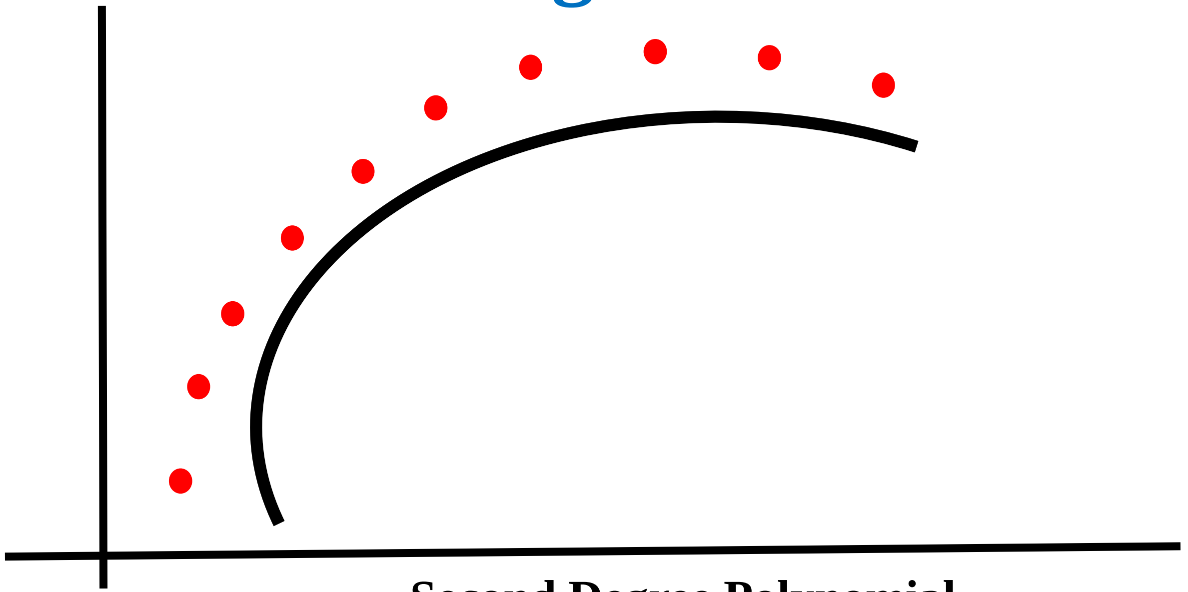
- Data used for training is not cleaned and contains noise (garbage values) in it.
- The model has a high bias.
- The size of the training dataset used is not enough.
- The model is too simple.

Overfitting and Underfitting

Training Data

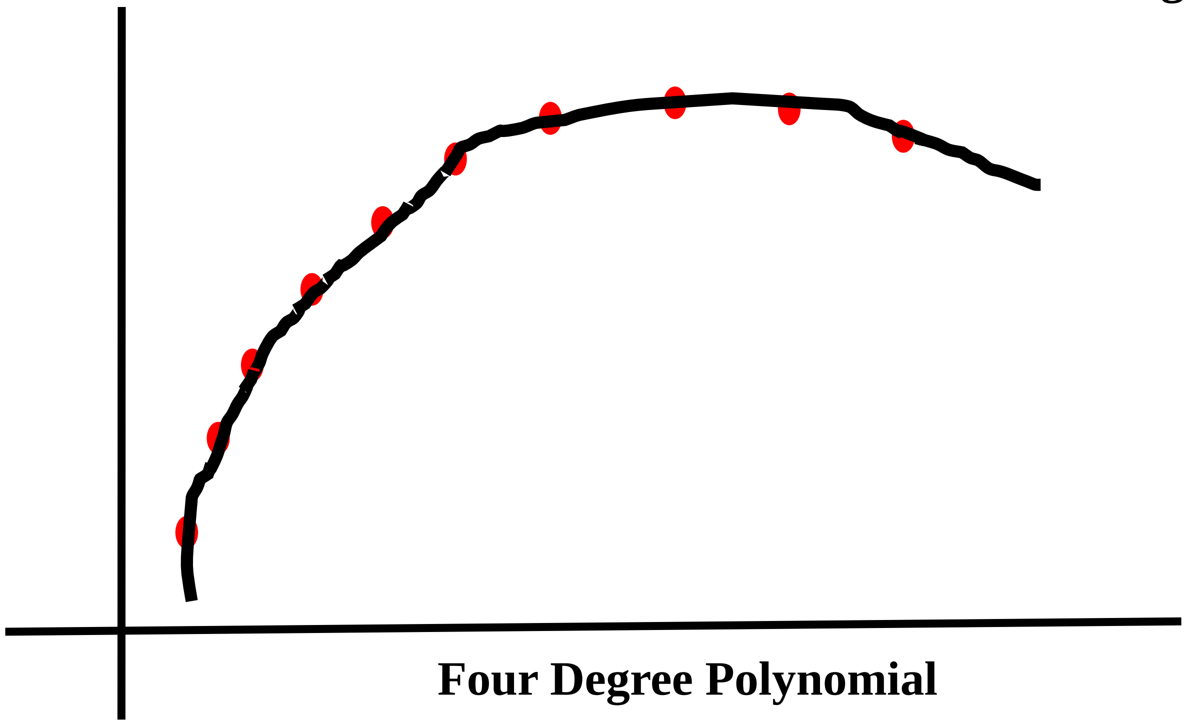


First Degree Polynomial



Second Degree Polynomial

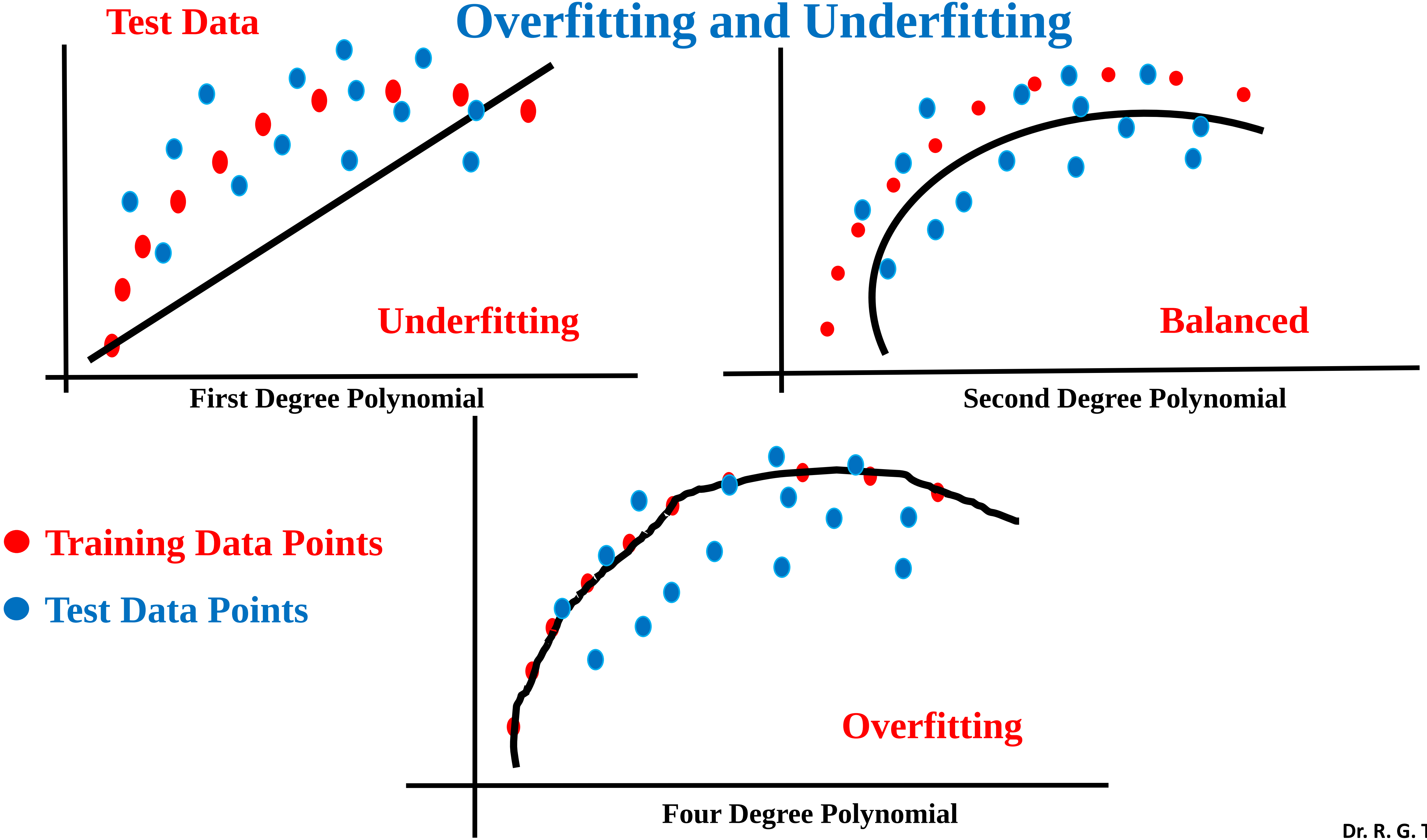
● Training Data Points



Four Degree Polynomial

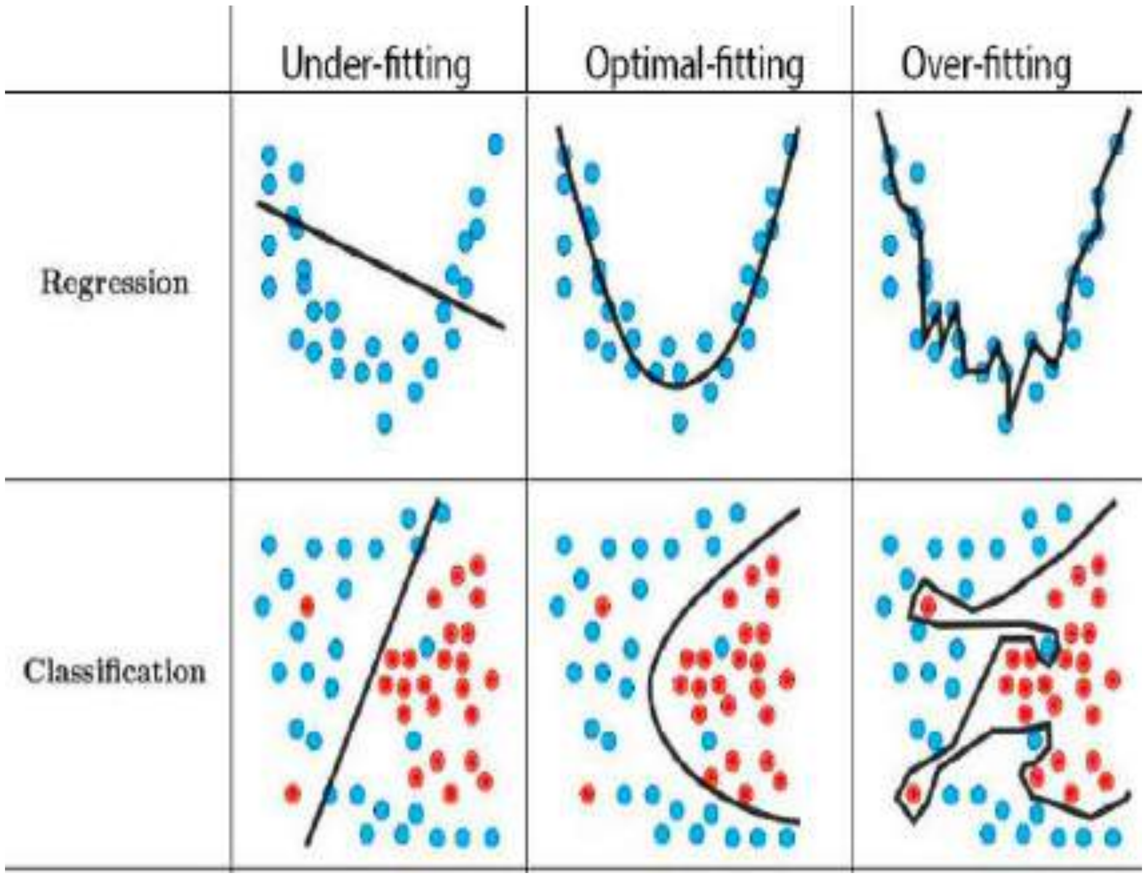
Dr. R. G. Tambe

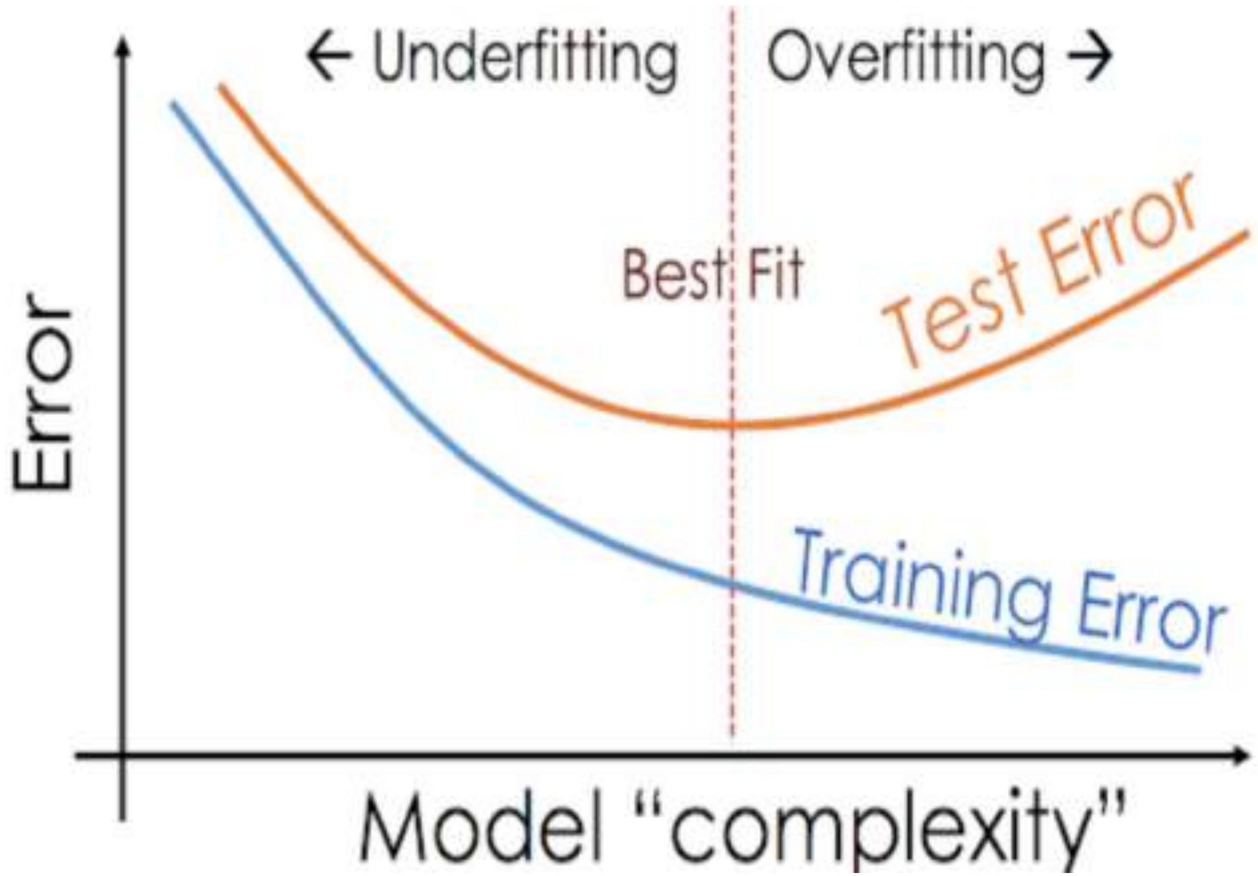
Overfitting and Underfitting



Dr. R. G. Tambe

Overfitting and Underfitting





Overfitting and Underfitting

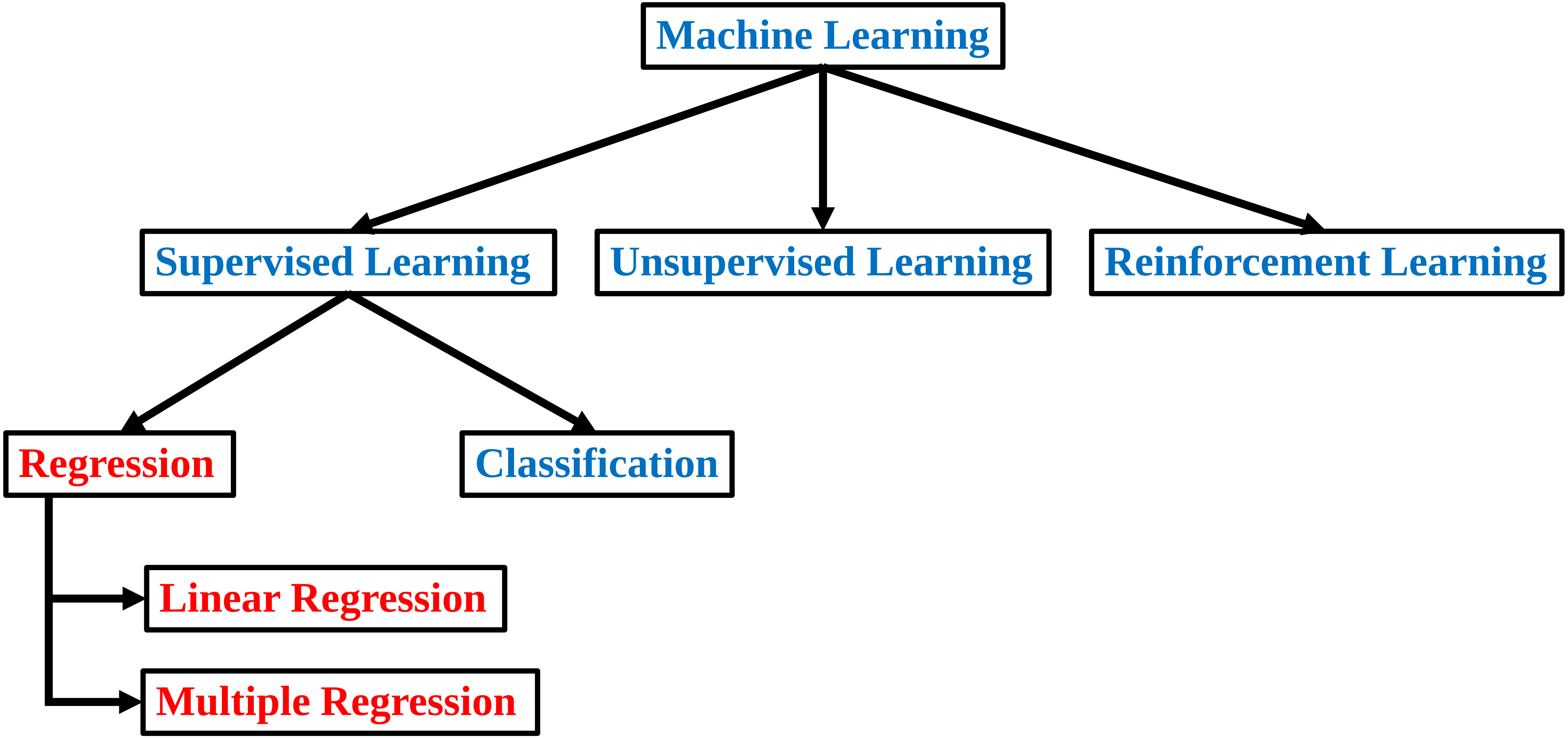
Ways to Tackle Overfitting

- Using K-fold cross-validation.
- Using Regularization techniques such as Lasso and Ridge.
- Training model with sufficient data.
- Adopting Ensembling Techniques.

Ways to Tackle Underfitting

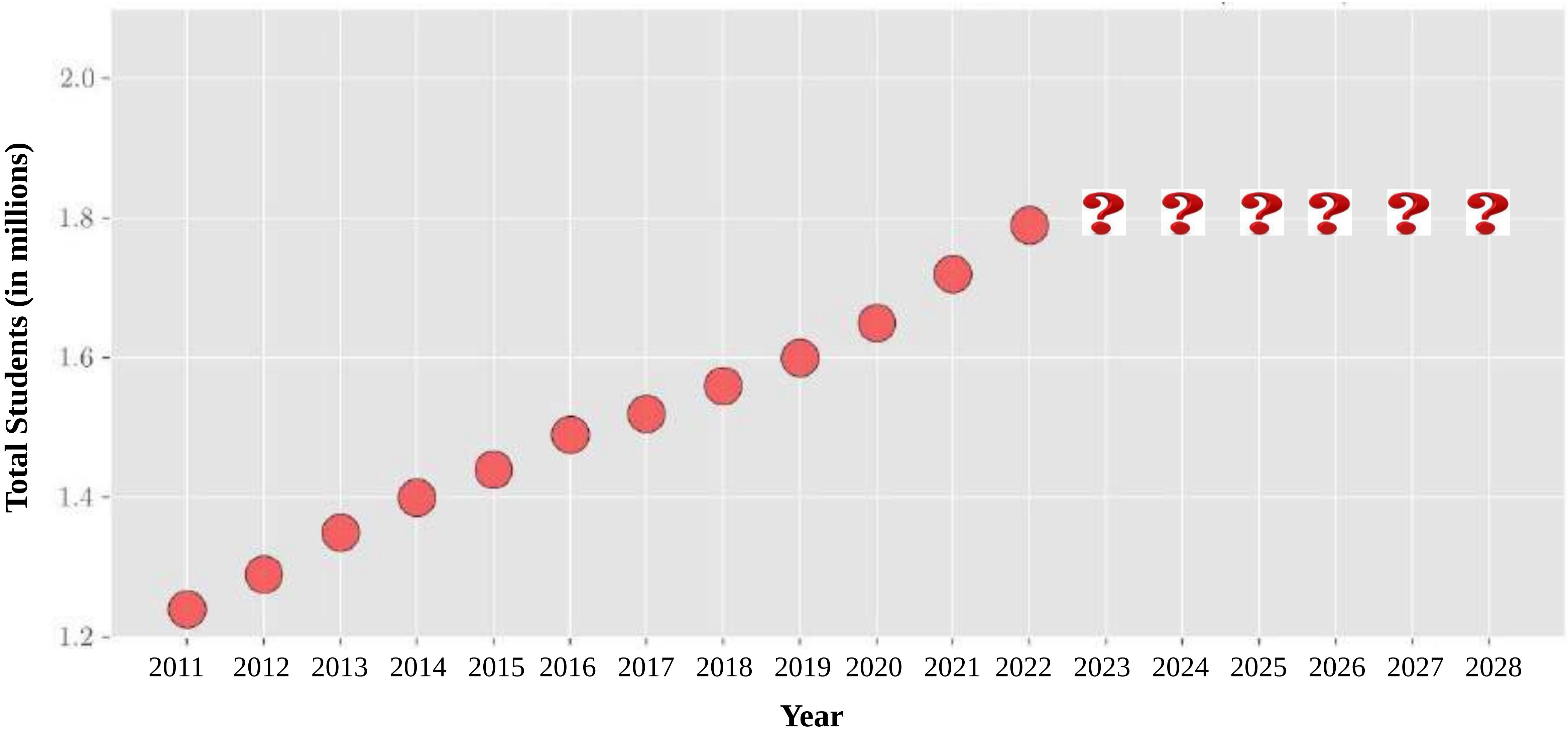
- Increase the number of features in the dataset.
- Increase model complexity.
- Reduce noise in the data.
- Increase the duration of training the data.

Types in Machine Learning



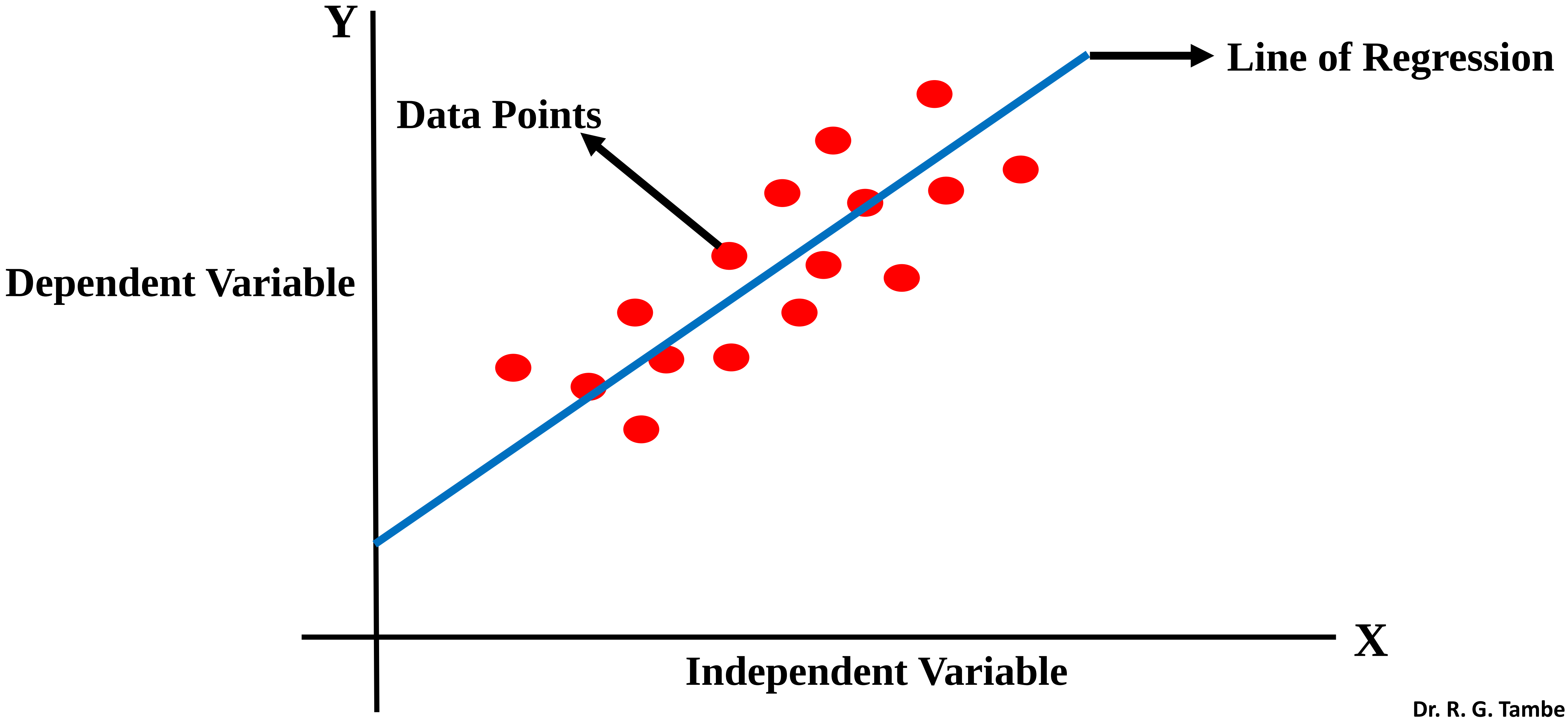
Regression

Number of College Graduates with Master Degree in India

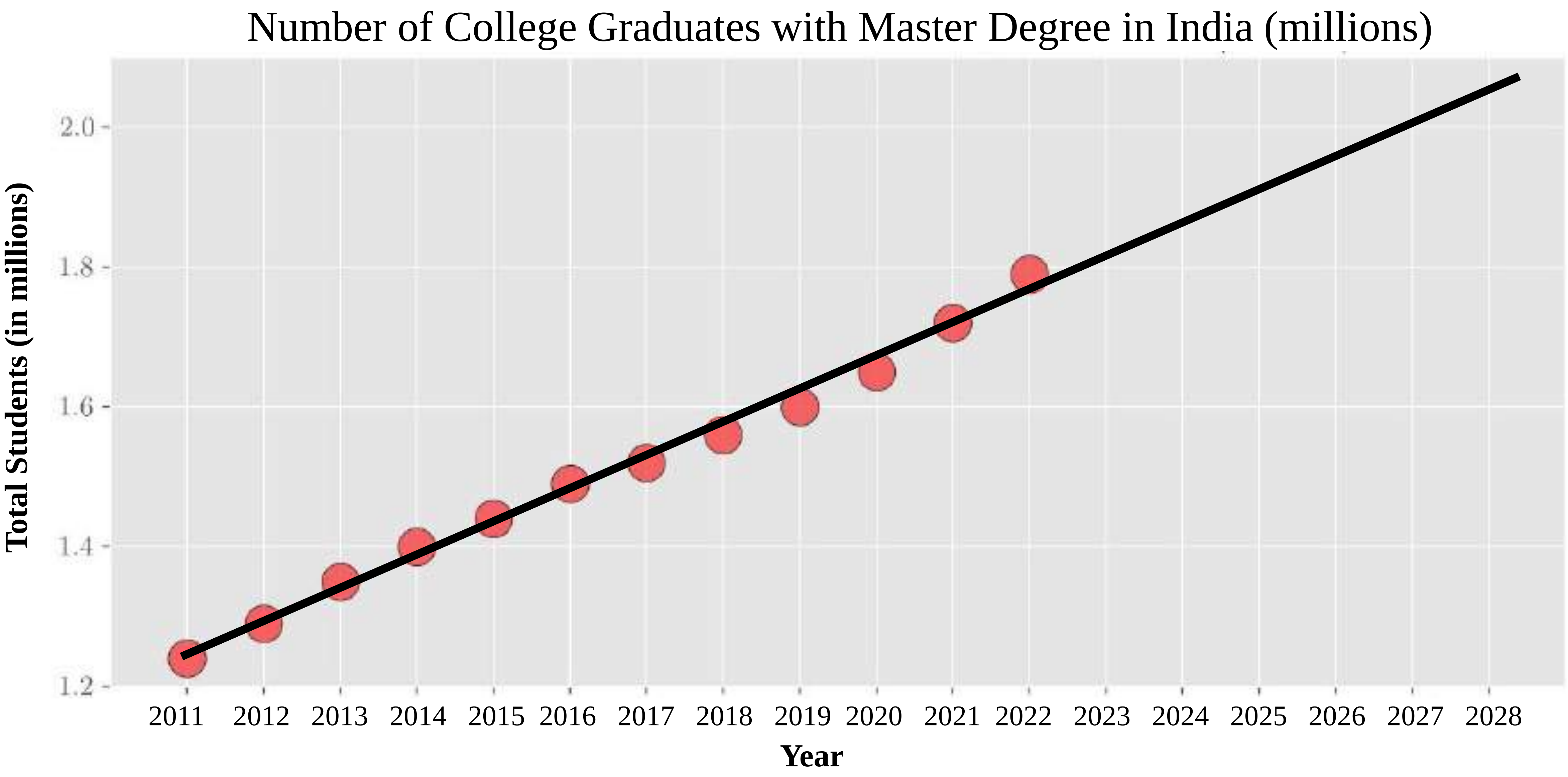


Regression

Regression Analysis is the process of estimating the relationship between a dependent variable and independent variables.

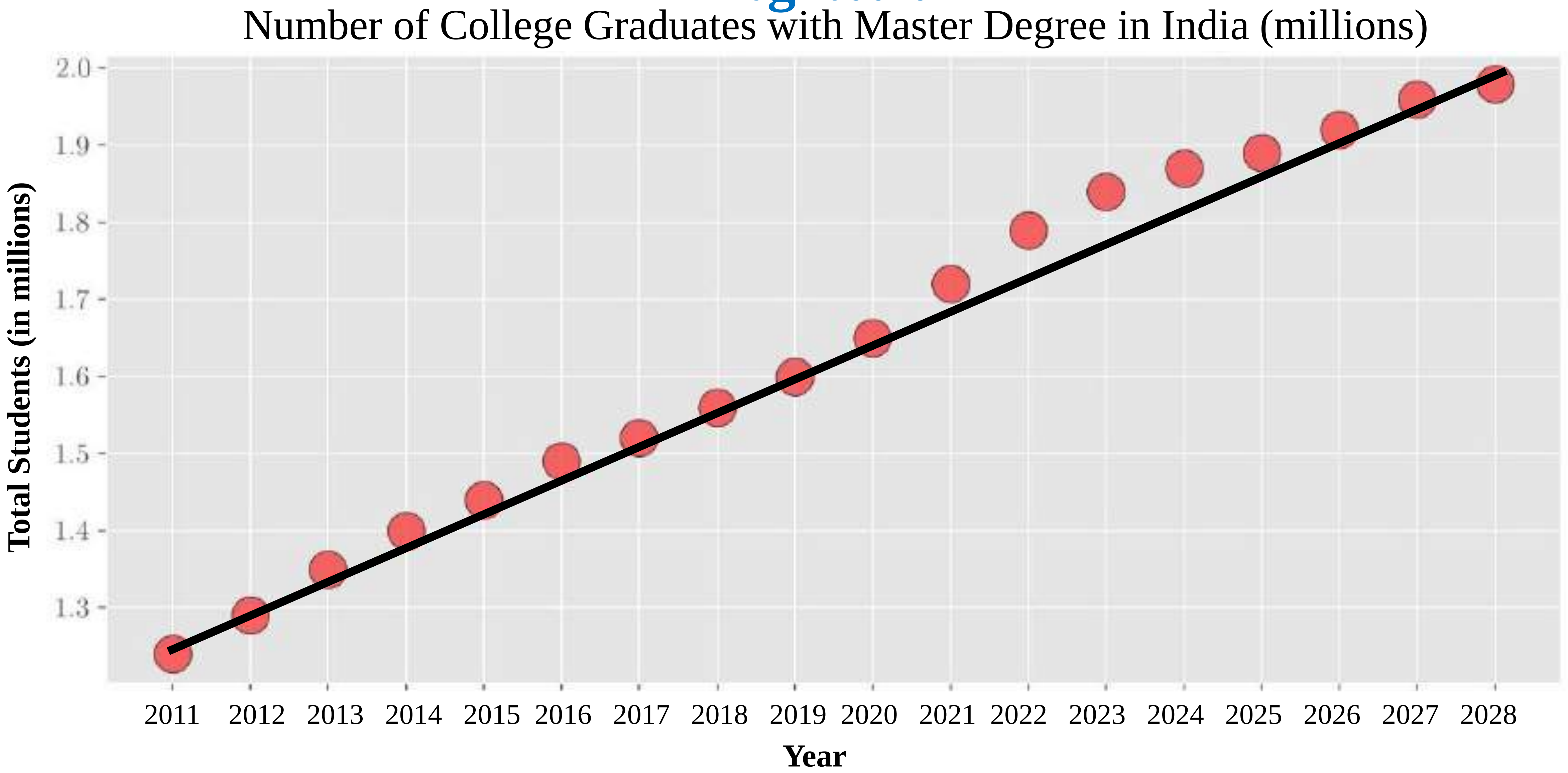


Regression



Dr. R. G. Tambe

Regression



Process of **fitting a function to a set of data points** is known as **regression analysis**.

Regression

Regression Analysis is the process of estimating the relationship between a dependent variable and independent variables.

Regression

Simple Linear Regression

The number of **independent variables** is **one** and there is a **linear relationship** between the **independent(x)** and **dependent(y)** variable.

Multiple Regression

The number of **independent variables** is **more then one** and there is a **linear relationship** between the **independent(x)** and **dependent(y)** variable.

Linear Regression

Regression Analysis is the process of estimating the relationship between a dependent variable and independent variables.

Simple Linear Regression:

The number of **independent variables** is **one** and there is a **linear relationship** between the **independent(x)** and **dependent(y)** variable.

$$y = \propto_0 + \propto_1 (x) + \varepsilon$$

y = dependent variable
x = independent variable
$\propto_0$ and $\propto_1$ = Regression Coefficients
ε = Residual Error

Linear Regression

Regression Analysis is the process of estimating the relationship between a dependent variable and independent variables.

Multiple Linear Regression:

The number of **independent variables** is **more then one** and there is a **linear relationship** between the **independent(x)** and **dependent(y)** variable.

$$y = \propto_0 + \propto_1 x_1 + \propto_2 x_2 + \propto_3 x_3 + \ldots + \propto_n x_n + \varepsilon$$

y = dependent variable

$x_1, x_2, \ldots, x_n$ = independent variable

$\propto_0, \propto_1, \propto_2, \ldots, \propto_n$ = Regression Coefficients

ε = Residual Error

Linear Regression

Given data points, predict value of Glucose level if Age of person is 55. Further calculate regression coefficient for the same.

Subjects/Samples	Age	Glucose Level
1	43	99
2	21	65
3	25	79
4	42	75
5	57	87
6	59	81
7	55	86.327

$$\hat{y}_i = \propto_0 + \propto_1 x_i$$

$$\propto_0 = \frac{SS_{xy}}{SS_{xx}}$$ y intercept

$$\propto_1 = \bar{y} - b\bar{x}$$ Slope of Line

Linear Regression

Given data points, predict value of Glucose level if Age of person is 55. Further calculate regression coefficient for the same.

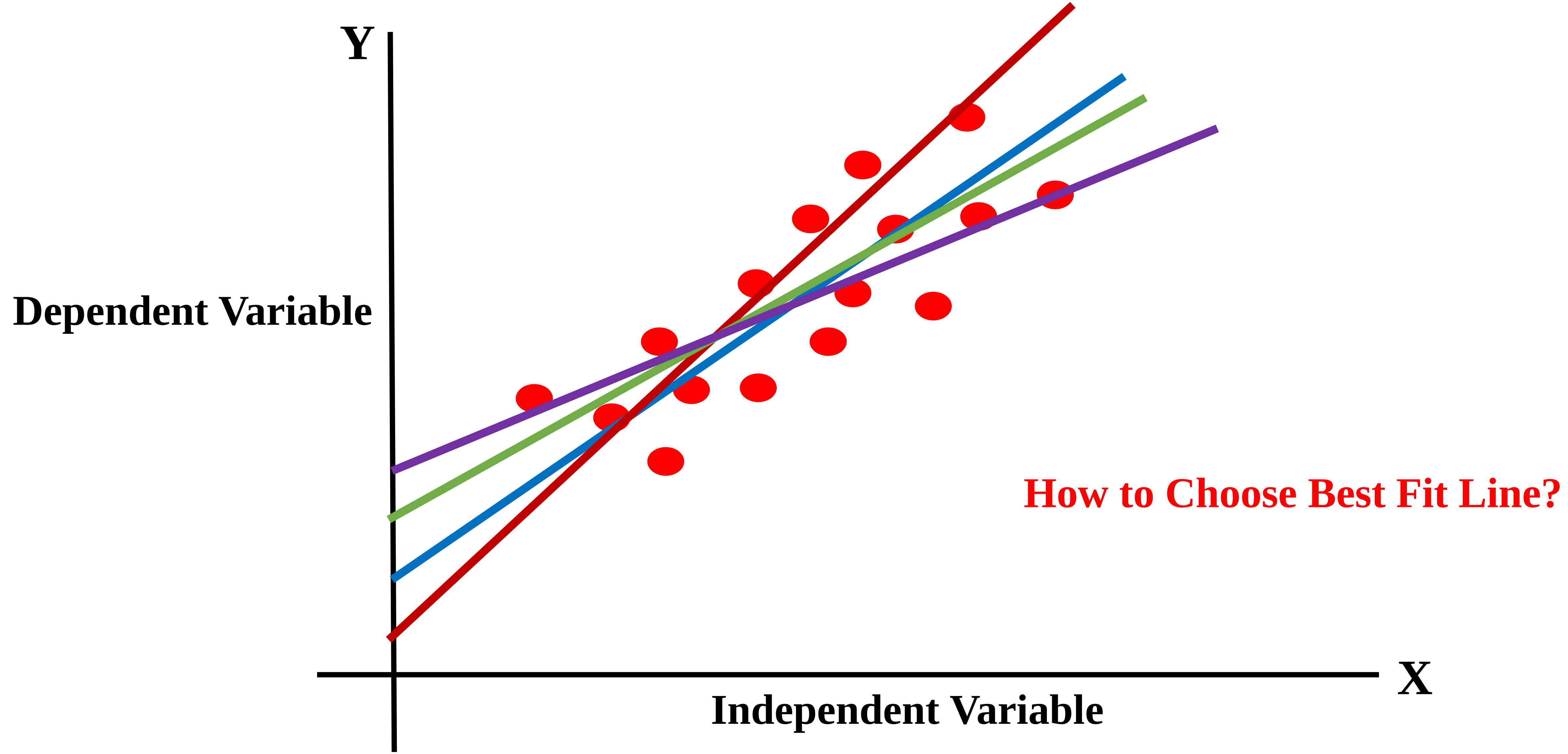
Subjects/Samples	Age	Glucose Level
1	43	99
2	21	65
3	25	79
4	42	75
5	57	87
6	59	81
7	55	86.327

$$\hat{y}_i = \propto_0 + \propto_1 x_i$$

$$\propto_0 = \frac{(\sum y)(\sum x^2) - (\sum x)(\sum xy)}{n(\sum x^2) - (\sum x)^2}$$

$$\propto_1 = \frac{n(\sum xy) - (\sum x)(\sum y)}{n(\sum x^2) - (\sum x)^2}$$

Linear Regression



Dr. R. G. Tambe

Linear Regression

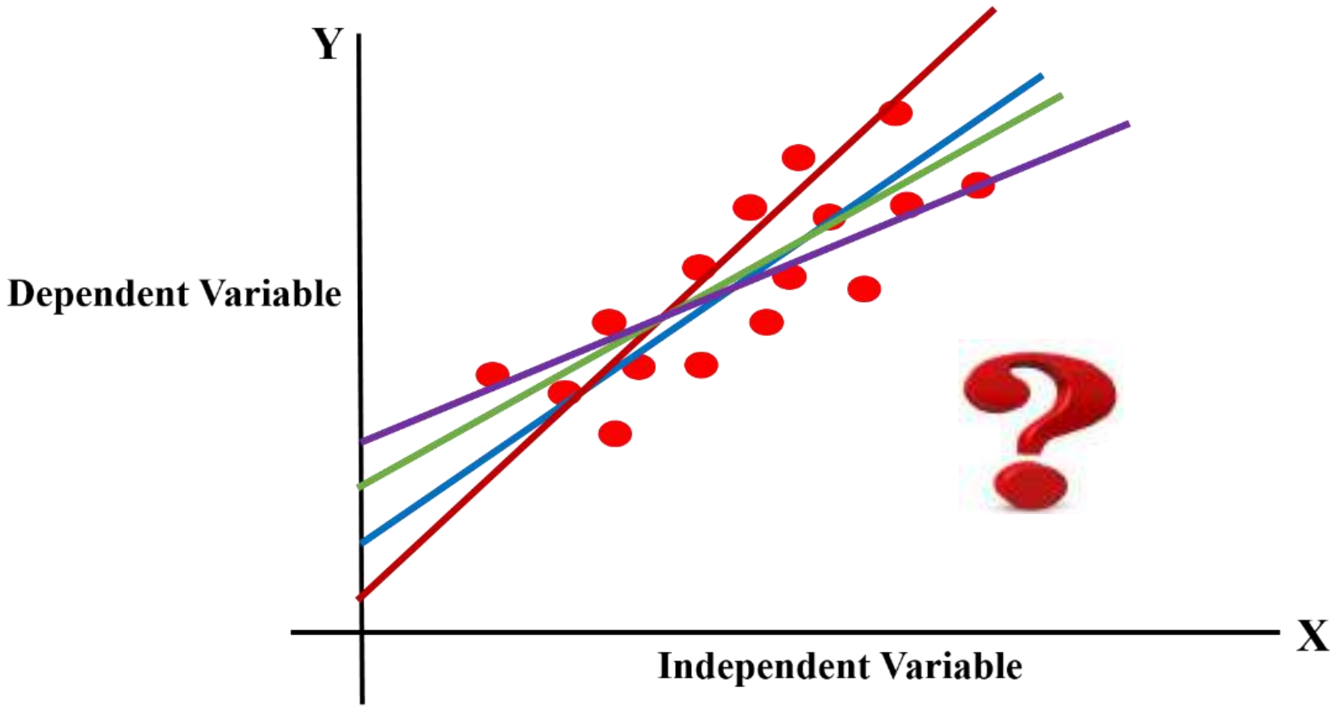
Cost Function:

We can use cost function to select Best Fit Line.

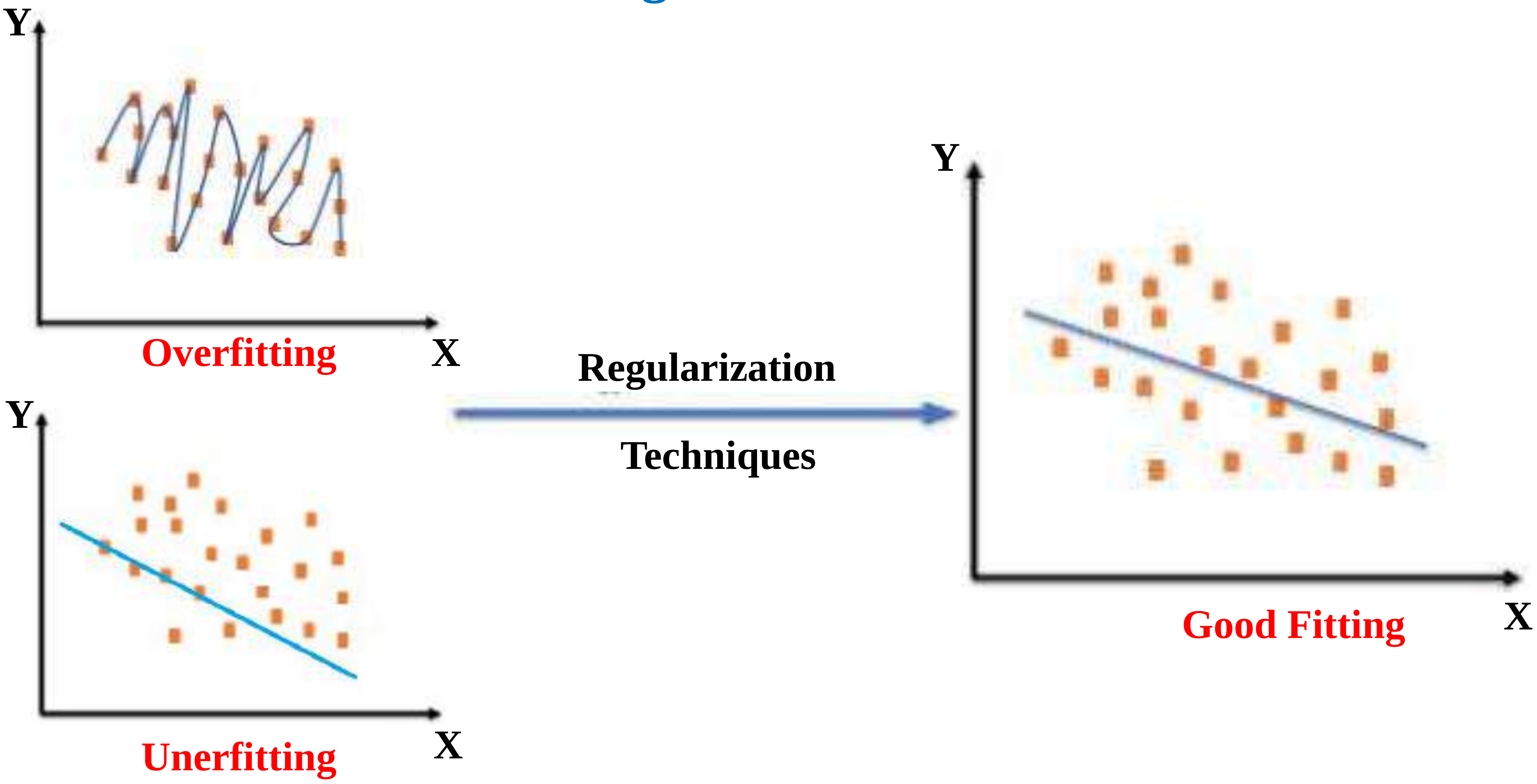
$$\text{Cost Function} = \frac{1}{2n}\sum_{i=1}^{n}(\hat{y} - y)^2$$

$$J(n) = \frac{1}{2n}\sum_{i=1}^{n}(\hat{y} - y)^2$$

$\hat{y}$ = Predicted Value
y = Actual Value
n = No. of Data Points



Regularization



Regularization

Regularization is implemented to avoid overfitting of the data, especially when there is a large variance between train and test set performances.

$$y = \propto_0 + \propto_1 x_1 + \propto_2 x_2 + \propto_3 x_3 + \ldots + \propto_n x_n + \varepsilon$$

With regularization, the number of features used in training is kept constant, yet the magnitude of the coefficients ($\propto$) as seen in above equation, is reduced.

The fitting procedure involves a loss function, known as **residual sum of squares** or **RSS**.

$$RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

y_i ----- Actual Value

$\hat{y}_i$ ----- Predicted Value and $\hat{y}_i = \propto_0 + \propto_i x_i$

Regularization

- Regularization
  - Ridge (L2) Regularization
  - Lasso (L1) Regularization

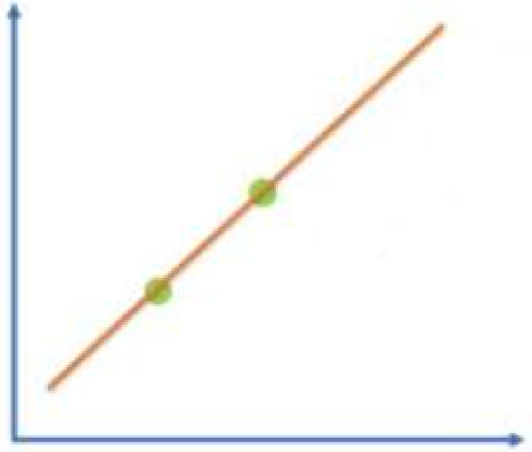
Ridge Regularization

Ridge Regression, it modifies the over-fitted or under fitted models by adding the penalty equivalent to the sum of the squares of the magnitude of coefficients.

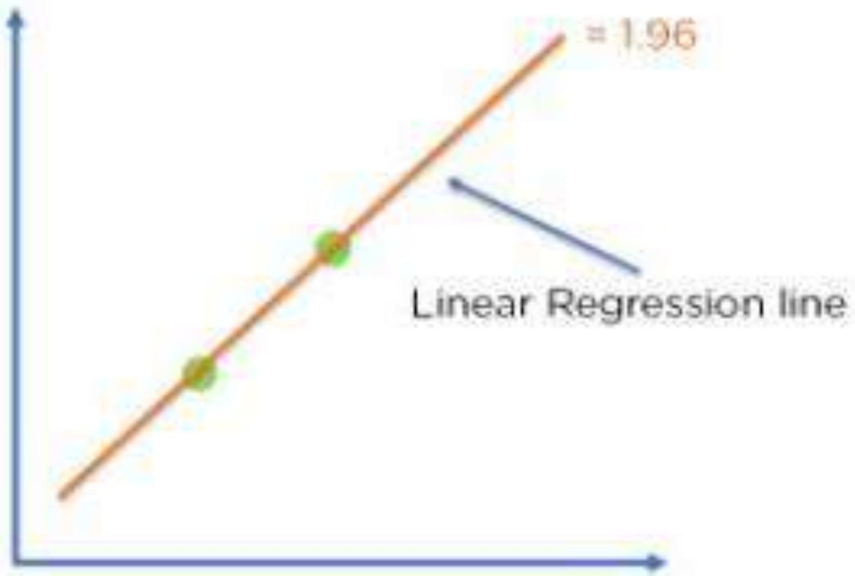
$$\textbf{Cost Function} = \text{RSS} + \boldsymbol{\lambda} \sum_{\mathbf{j=1}}^{\mathbf{p}} (\propto_{\mathrm{j}})^{\mathbf{2}}$$

$\text{RSS} = \sum_{i=1}^{n}(y_i - \propto_0 - \propto_i x_i)^2$, λ ------ Penalty for error and $\lambda > 0$, $\propto_j$ ------ Slope of line or curve

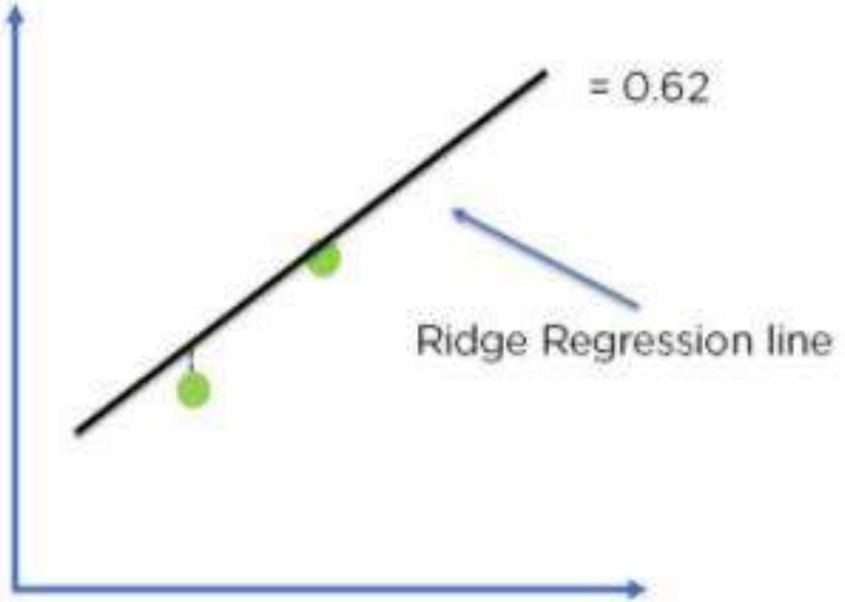
By changing the values of the penalty function, we are controlling the penalty term. The higher the penalty, it reduces the magnitude of coefficients. It shrinks the parameters.



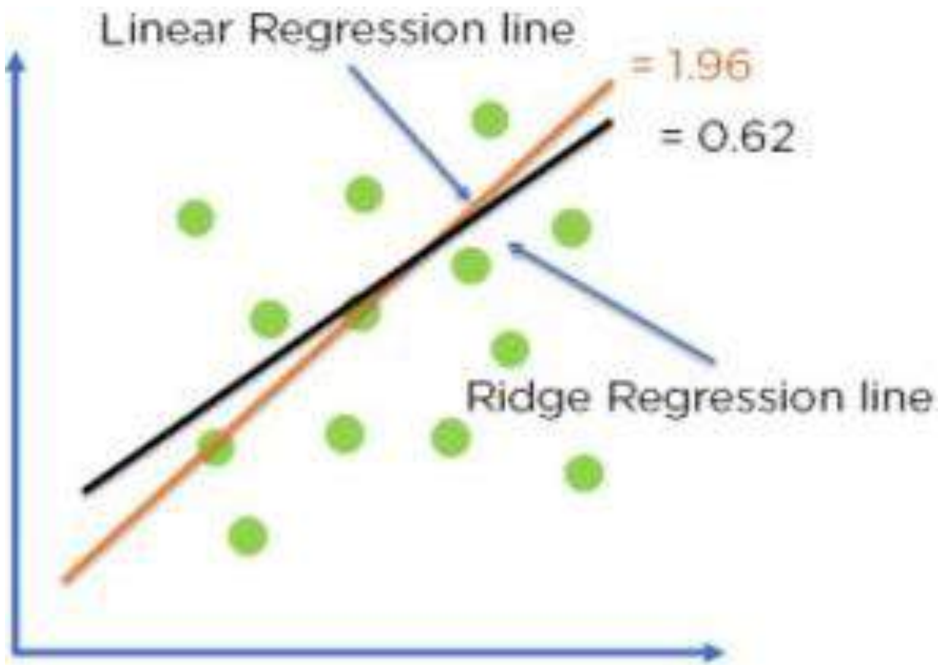
Ridge Regularization



Linear regression model



Ridge regression model



Optimization of model fit using Ridge Regression

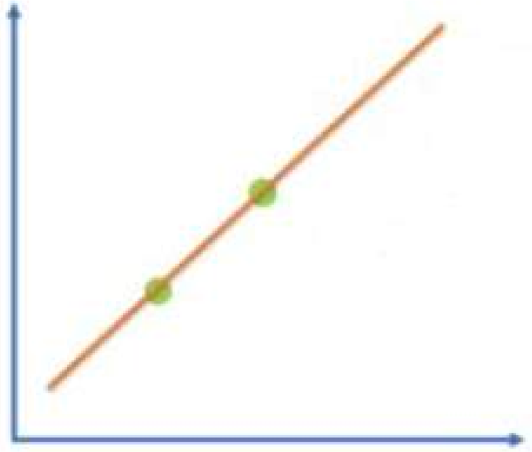
Lasso Regularization

Lasso Regression, it modifies the over-fitted or under fitted models by adding the penalty equivalent to the sum of the absolute values of coefficients.

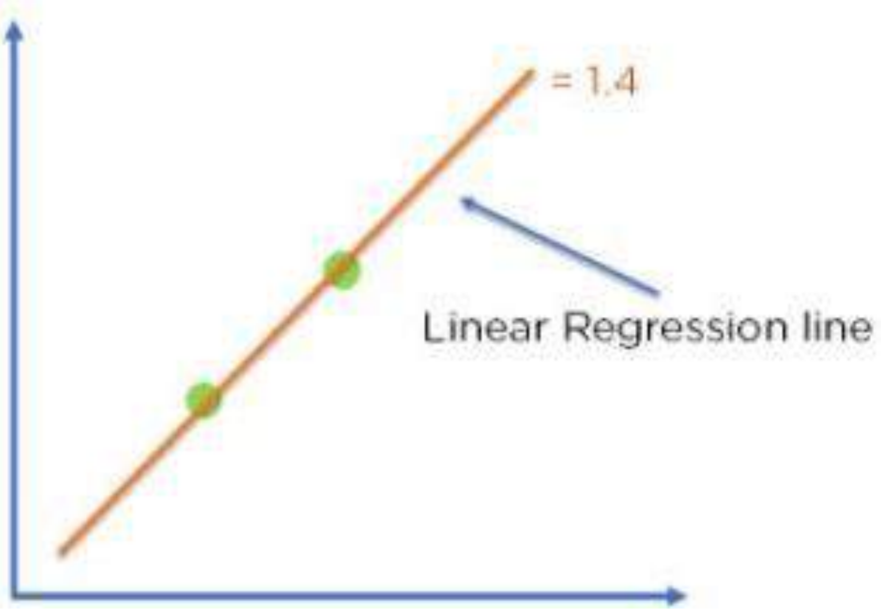
$$\textbf{Cost Function} = \mathrm{RSS} + \boldsymbol{\lambda} \sum_{\mathbf{j=1}}^{\mathbf{p}} (\propto_{\mathrm{j}})$$

$\mathrm{RSS} = \sum_{i=1}^{n}(y_i - \propto_0 - \propto_i x_i)^2$, λ ------ Penalty for error and $\lambda > 0$,$\propto_j$ ------ Slope of line or curve

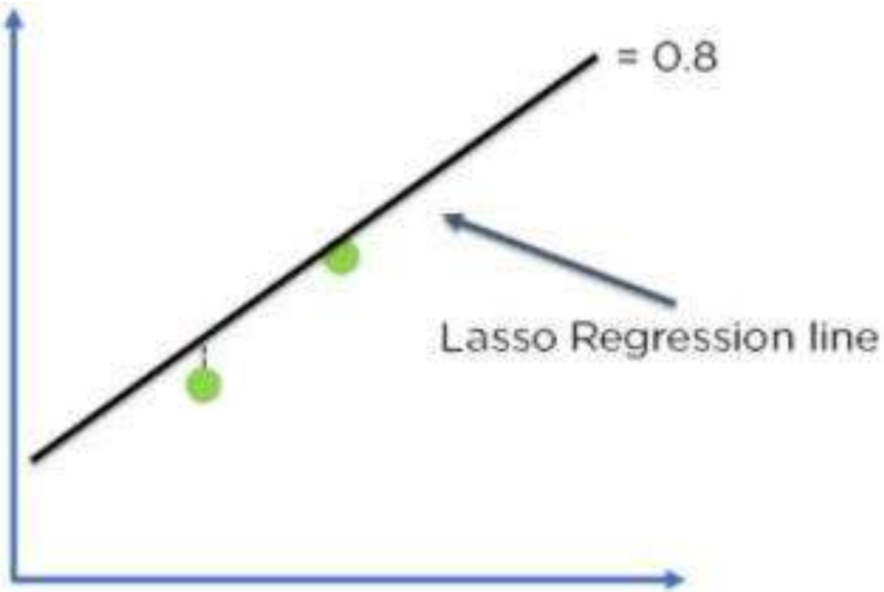
This means that the coefficient sum can also be 0, because of the presence of negative coefficients.



Lasso Regularization

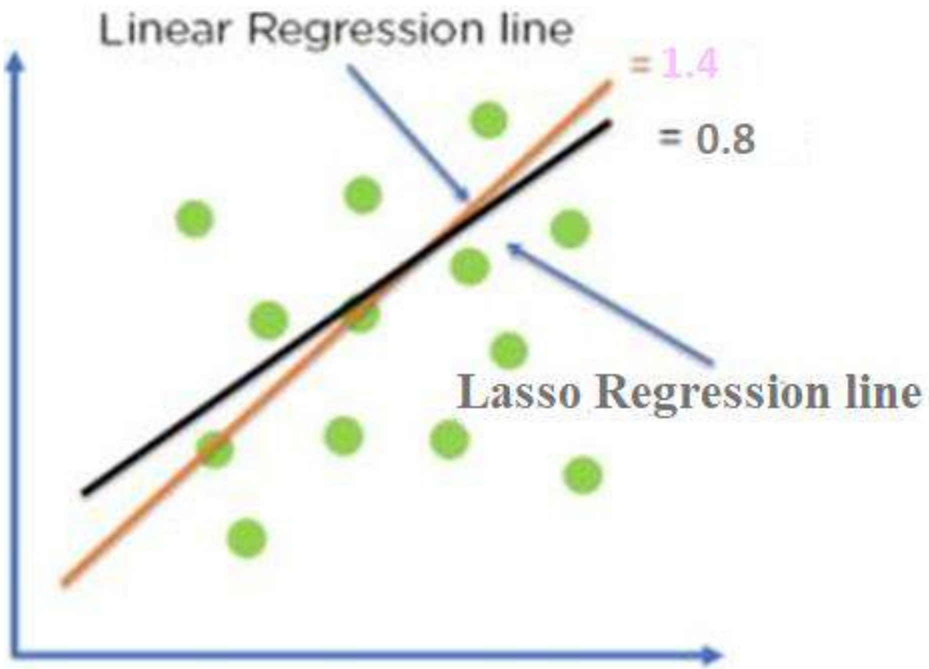


Linear regression model



Lasso regression model

Optimization of model fit using Lasso Regression

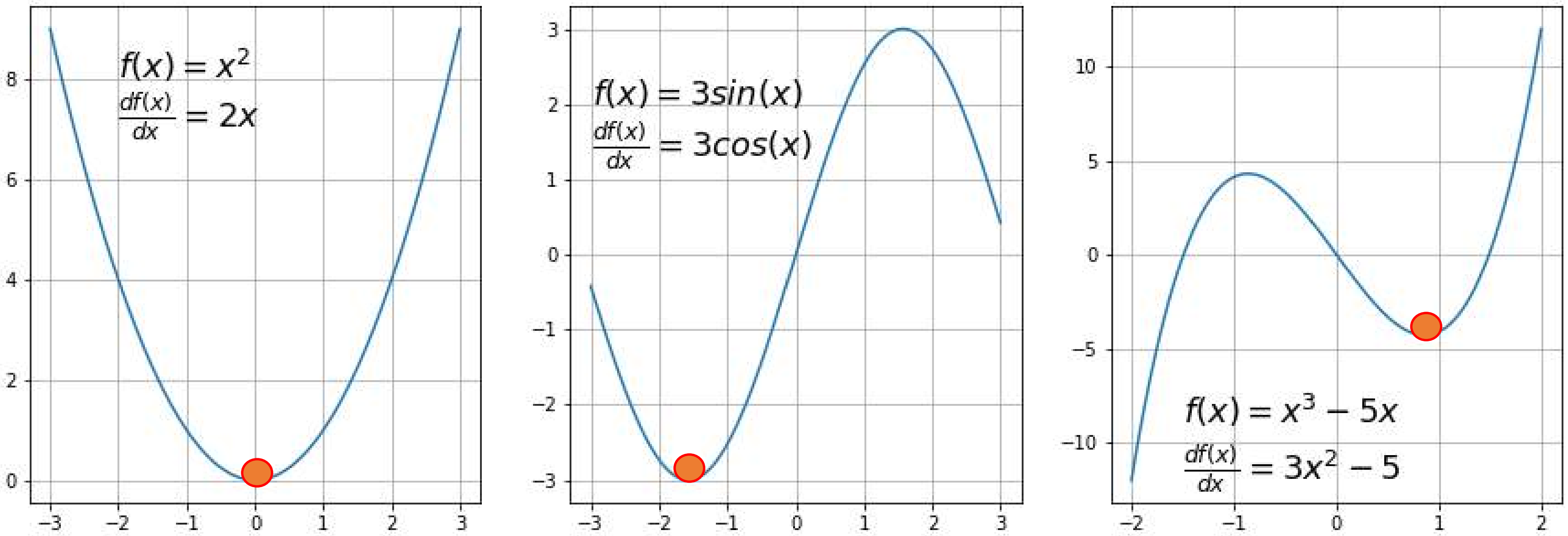


Gradient Descent Algorithm

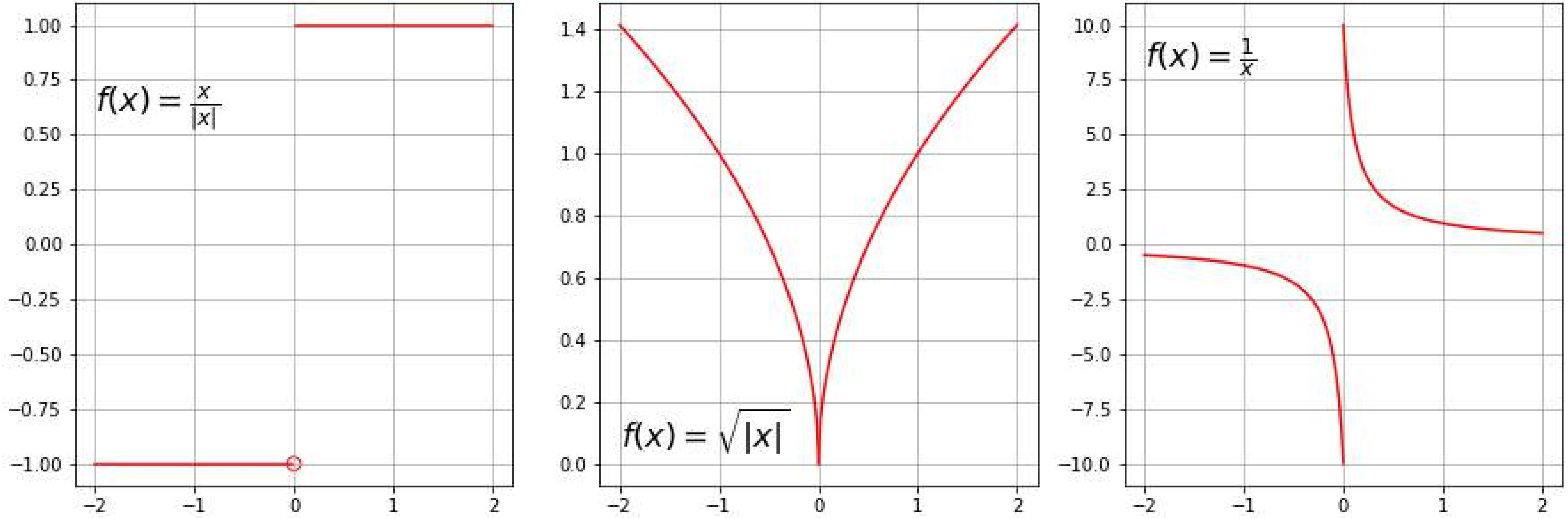
- **Gradient Descent** is an **optimization algorithm** for finding a **local minimum/minima** of a differentiable and convex function.
- Gradient descent is simply used in machine learning to find the **values of coefficients** that **minimize a cost function** as far as possible.
- Gradient descent algorithm does not work for all functions. There are two specific requirements.
  A function has to be:
  - **Differentiable**
  - **Convex**

Gradient Descent Algorithm

- **Differentiable Functions**

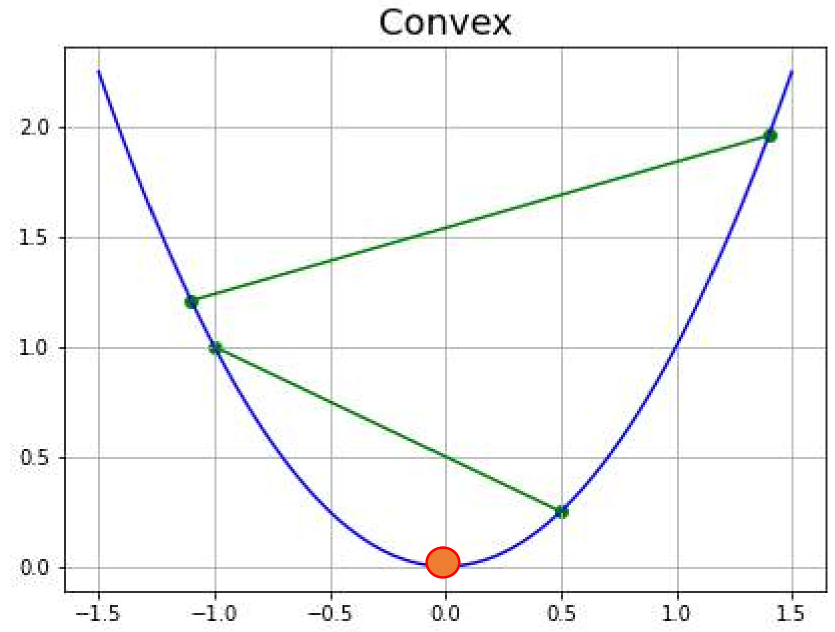


- **Non-Differentiable Functions**

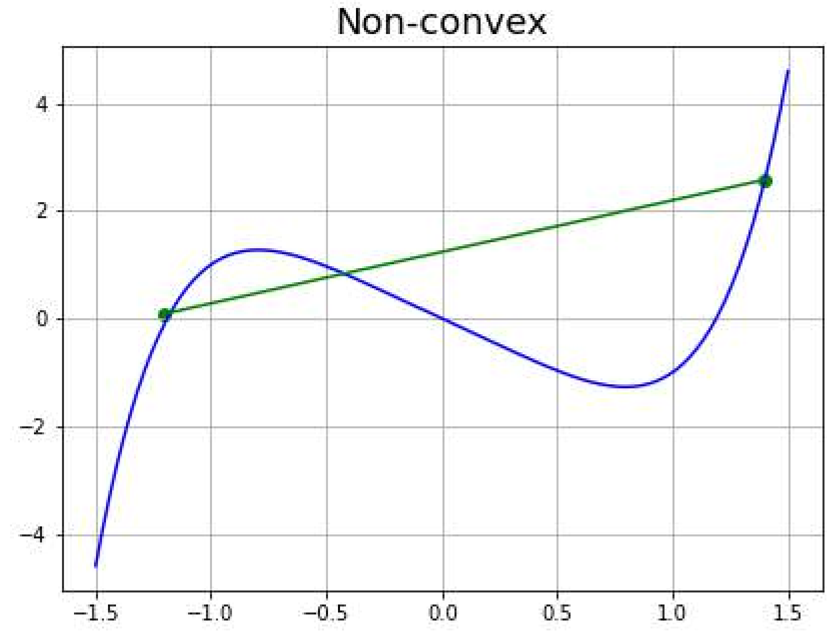


Gradient Descent Algorithm

- **Convex Function**



- **Non-Convex Function**



Gradient Descent Algorithm

Gradient

- A **gradient** simply measures the change in all weights with regard to the change in error.

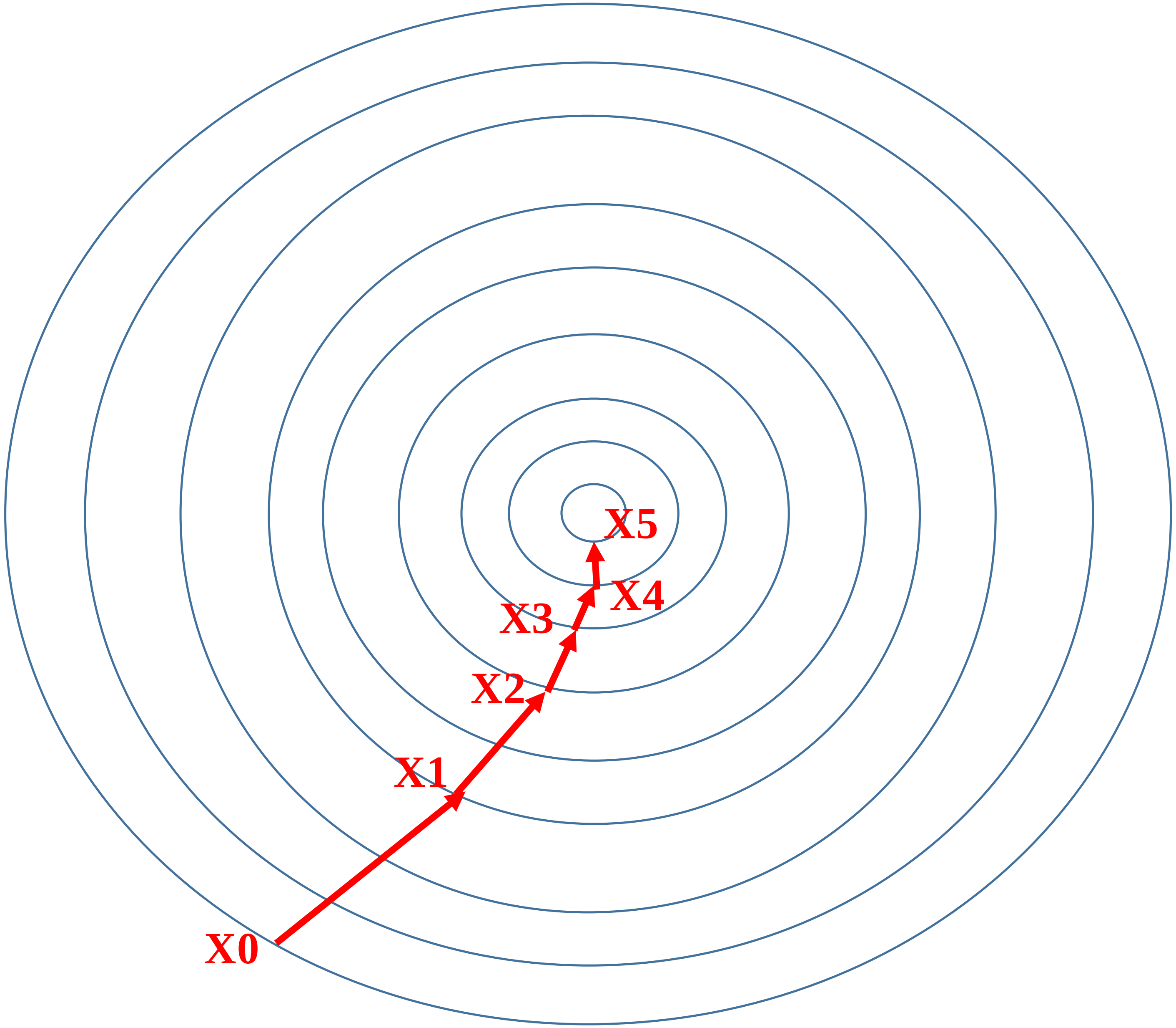
- You can also think of a gradient as the **slope of a function**. The higher the gradient, the steeper the slope and the faster a model can learn. But if the slope is zero, the model stops learning.

- In mathematical terms, a gradient is a **partial derivative** with respect to its inputs.

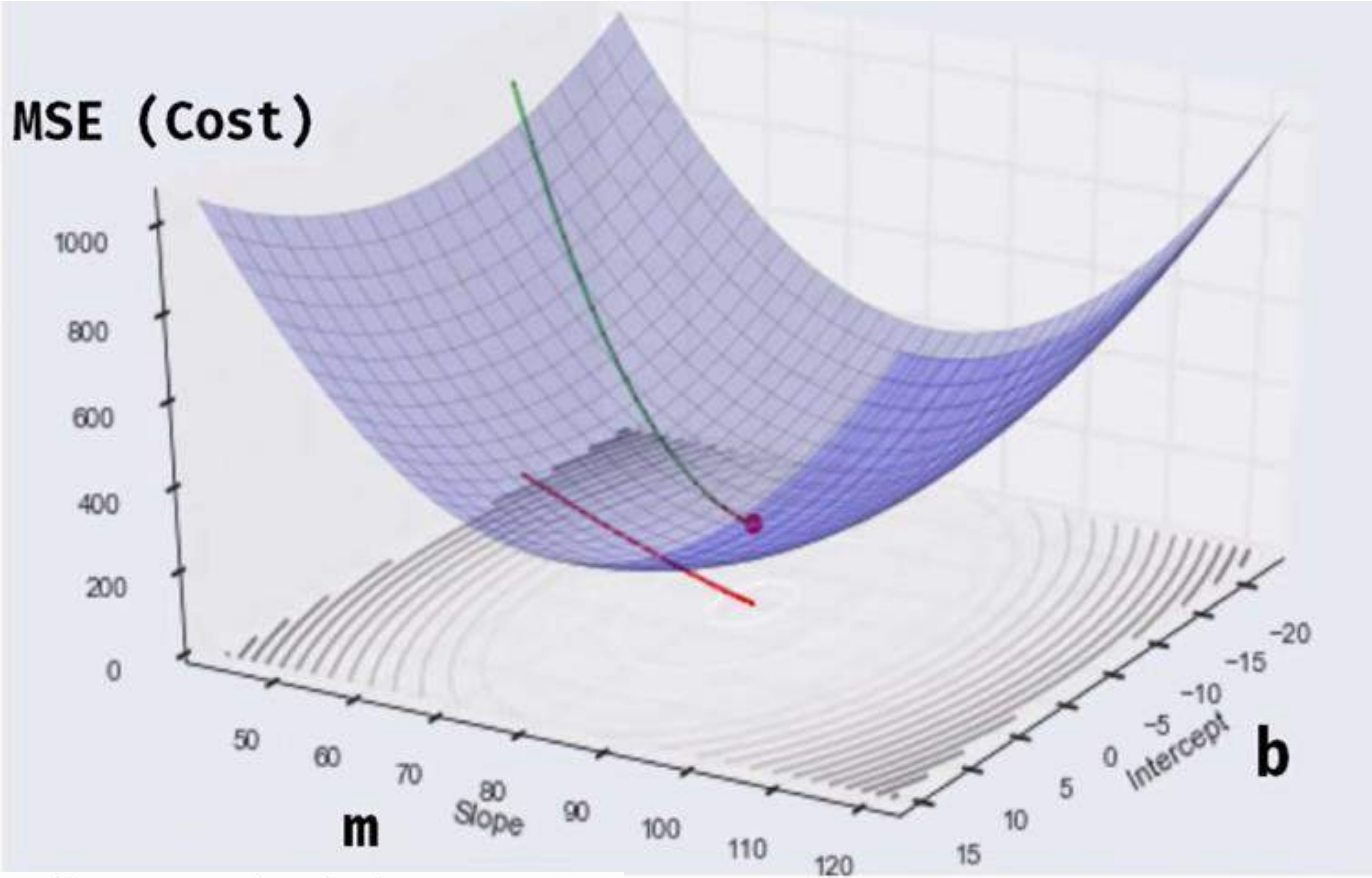
- A **gradient** measures how much the output of a function changes if you change the inputs a little bit.

Gradient Descent Algorithm

Gradient

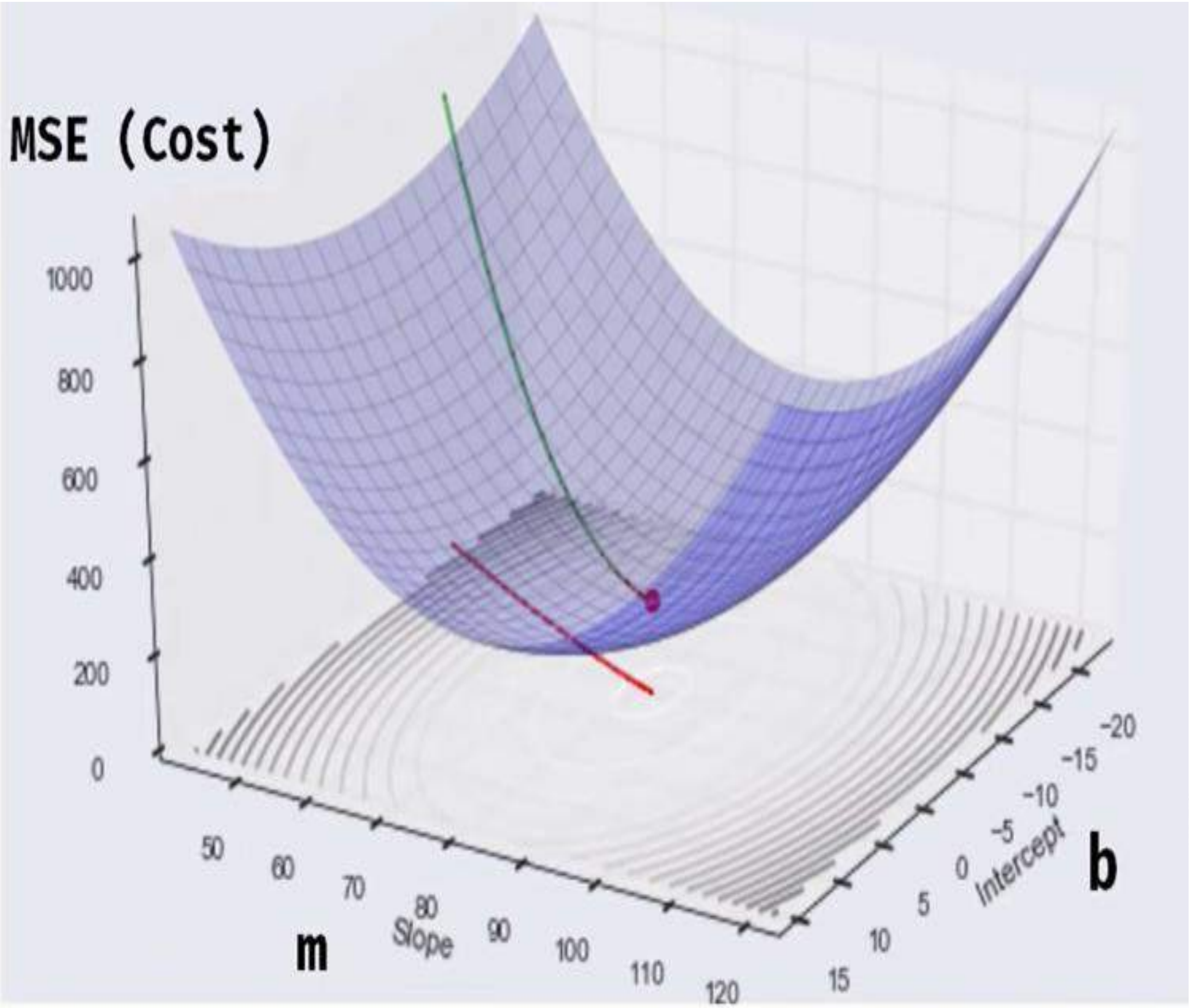


Gradient Descent Algorithm

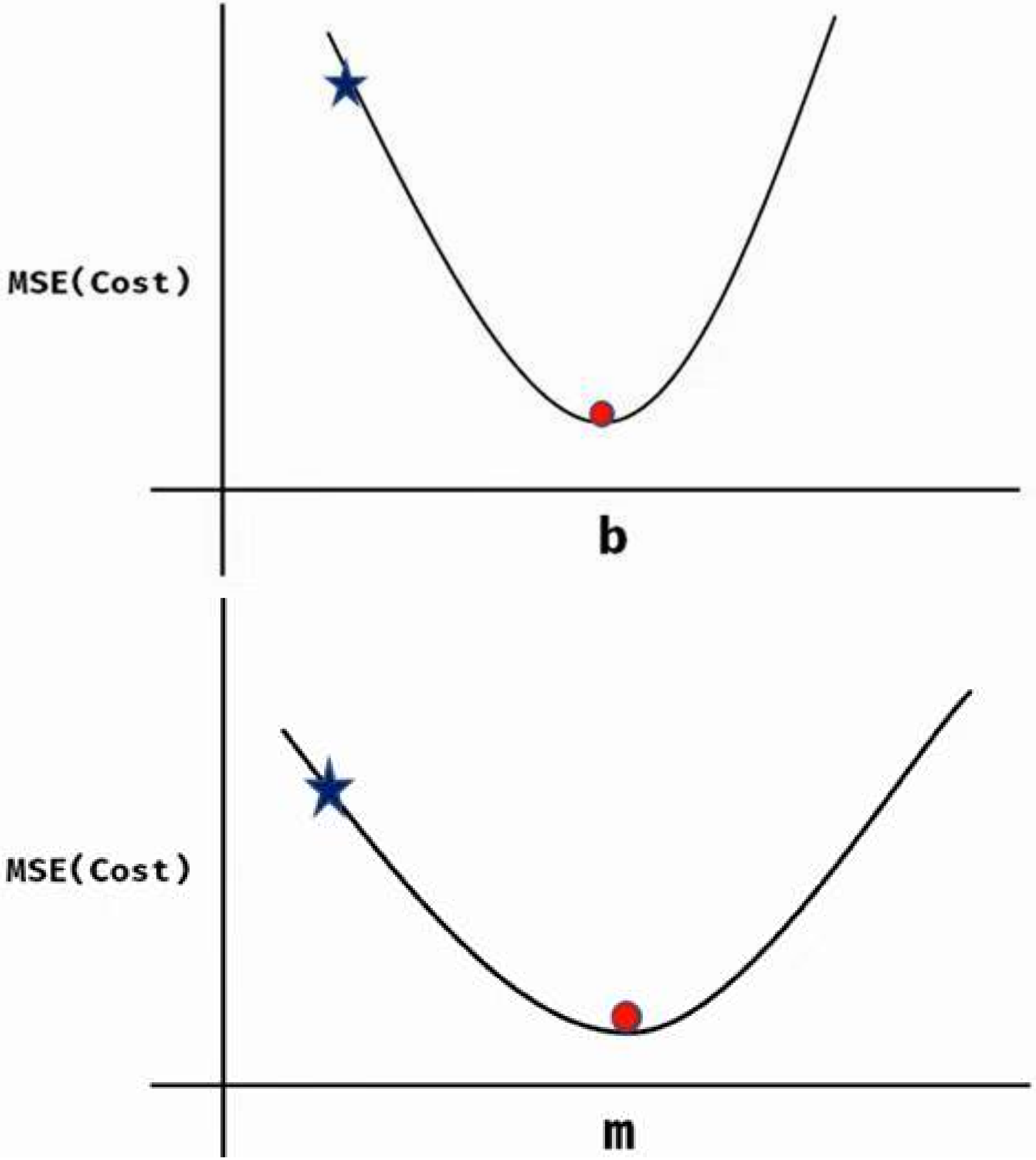


Reference: https://am207.github.io/2017/wiki/gradientdescent.html

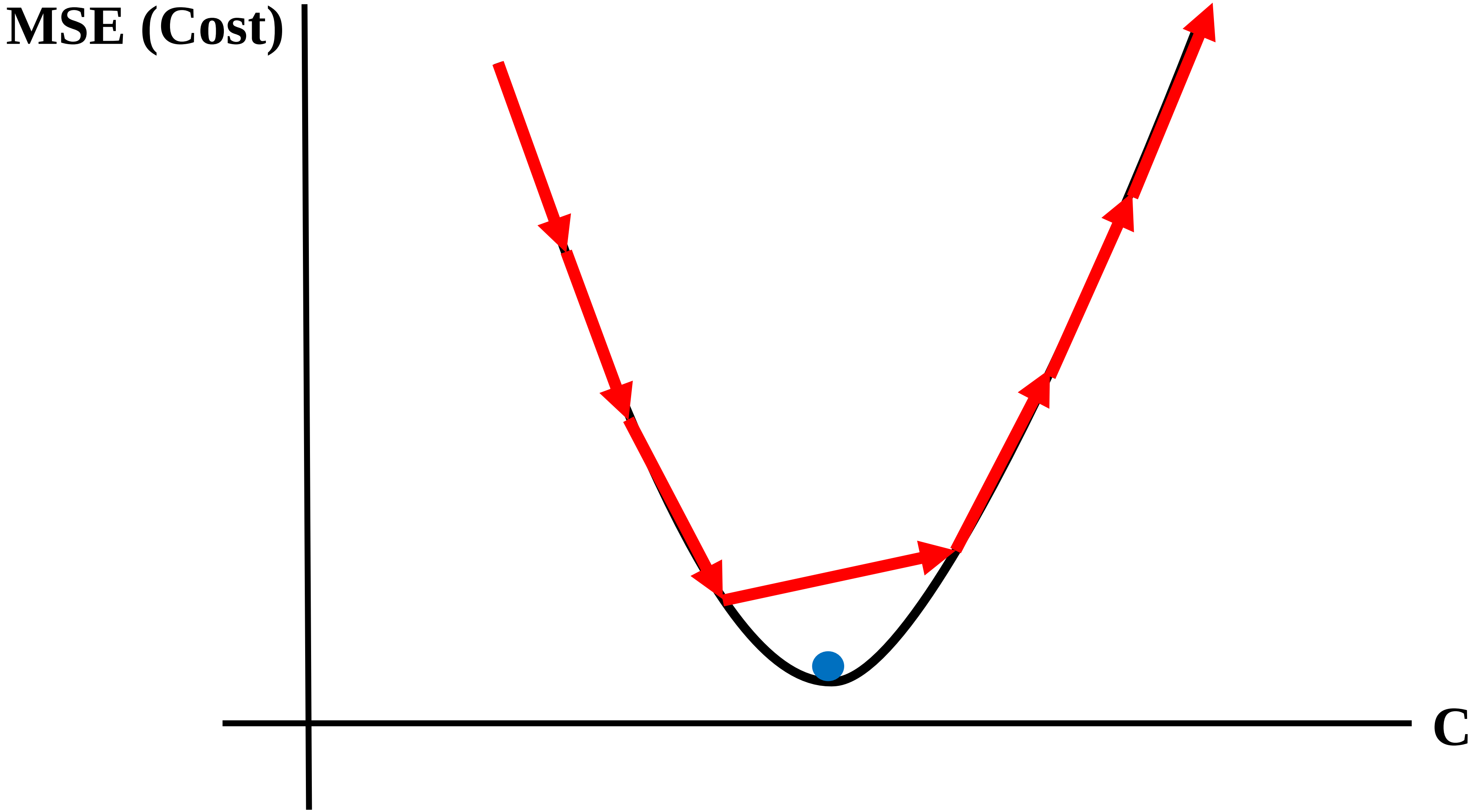
Gradient Descent Algorithm



Reference: https://am207.github.io/2017/wiki/gradientdescent.html

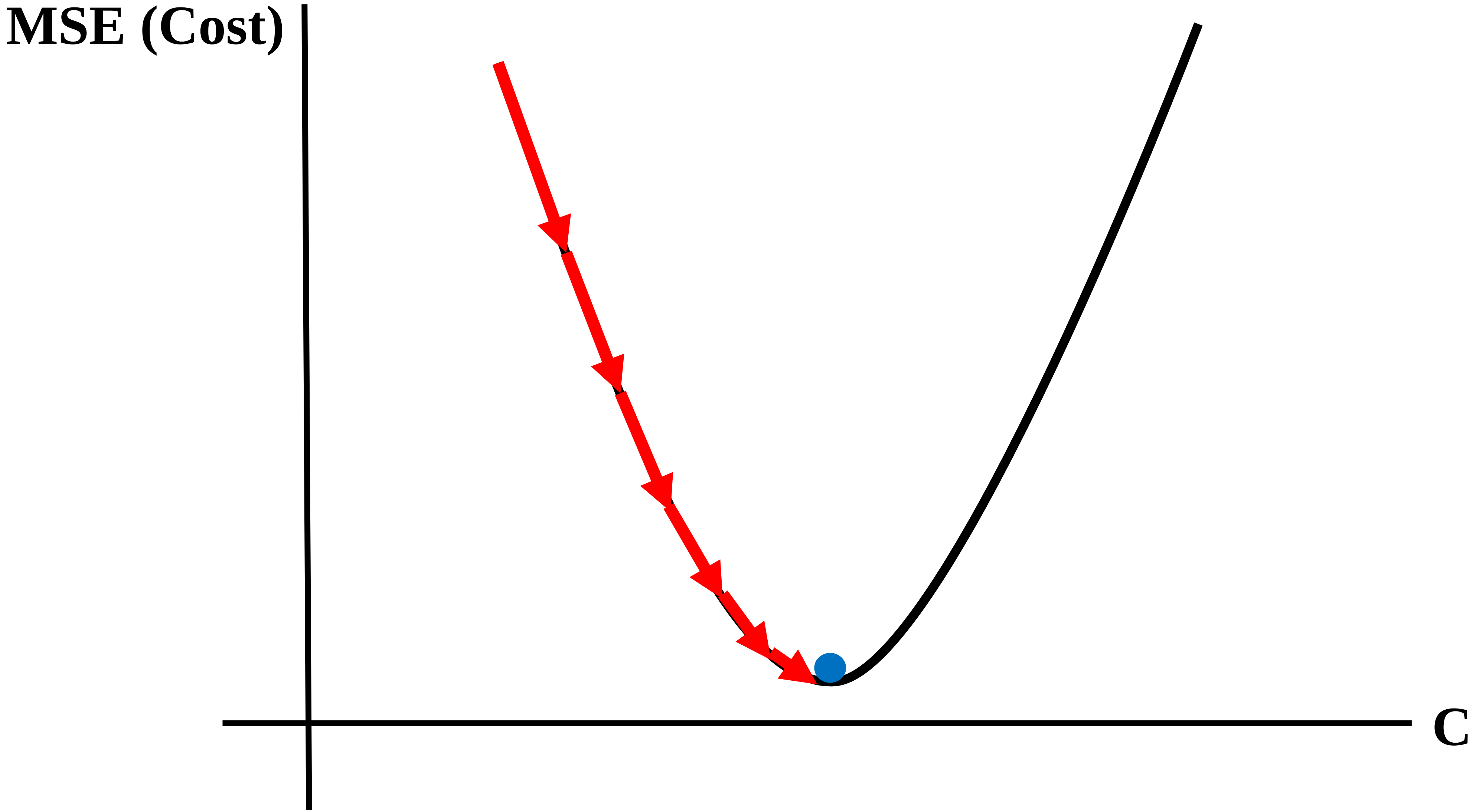


Gradient Descent Algorithm

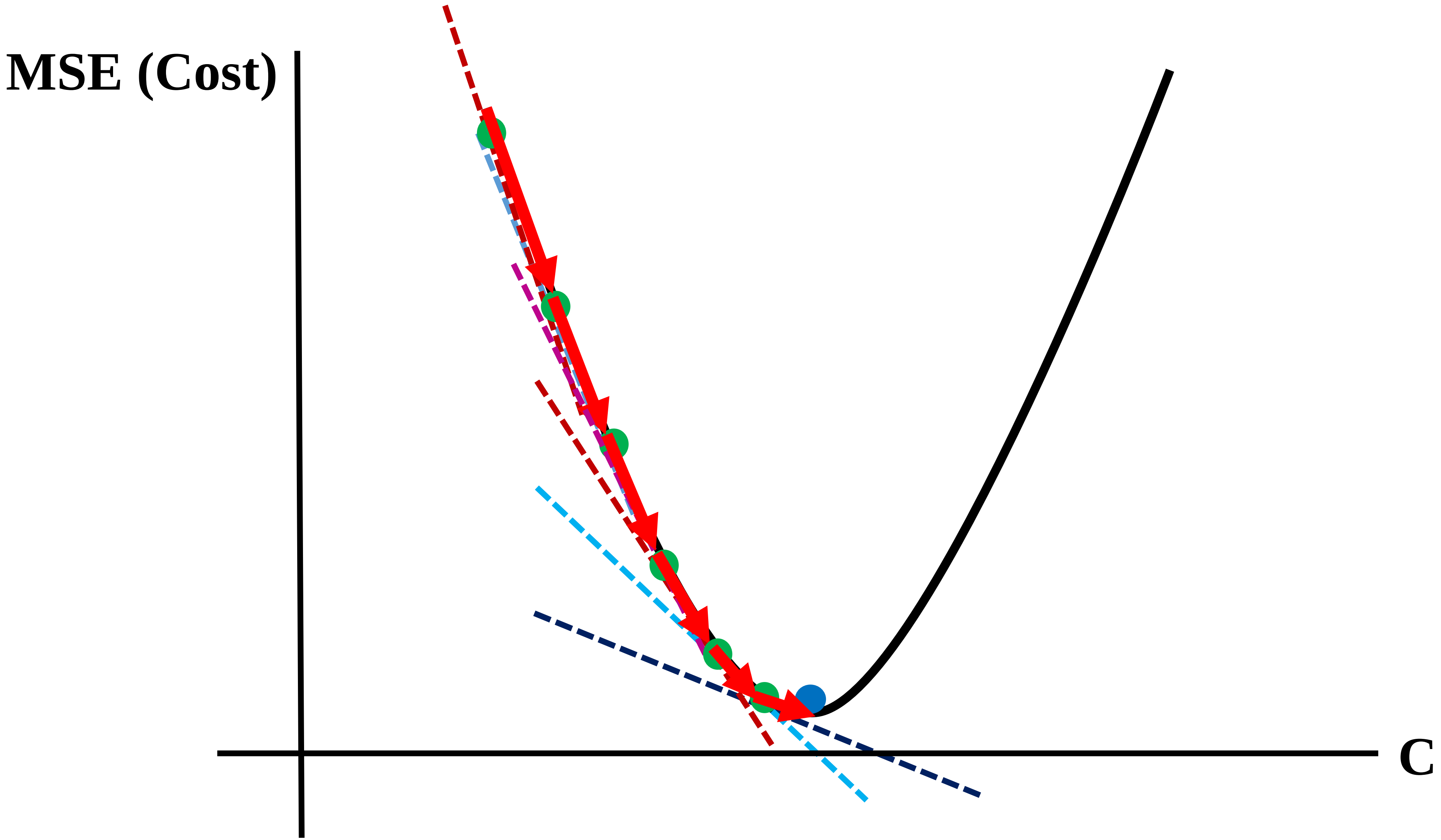


Dr. R. G. Tambe

Gradient Descent Algorithm



Gradient Descent Algorithm



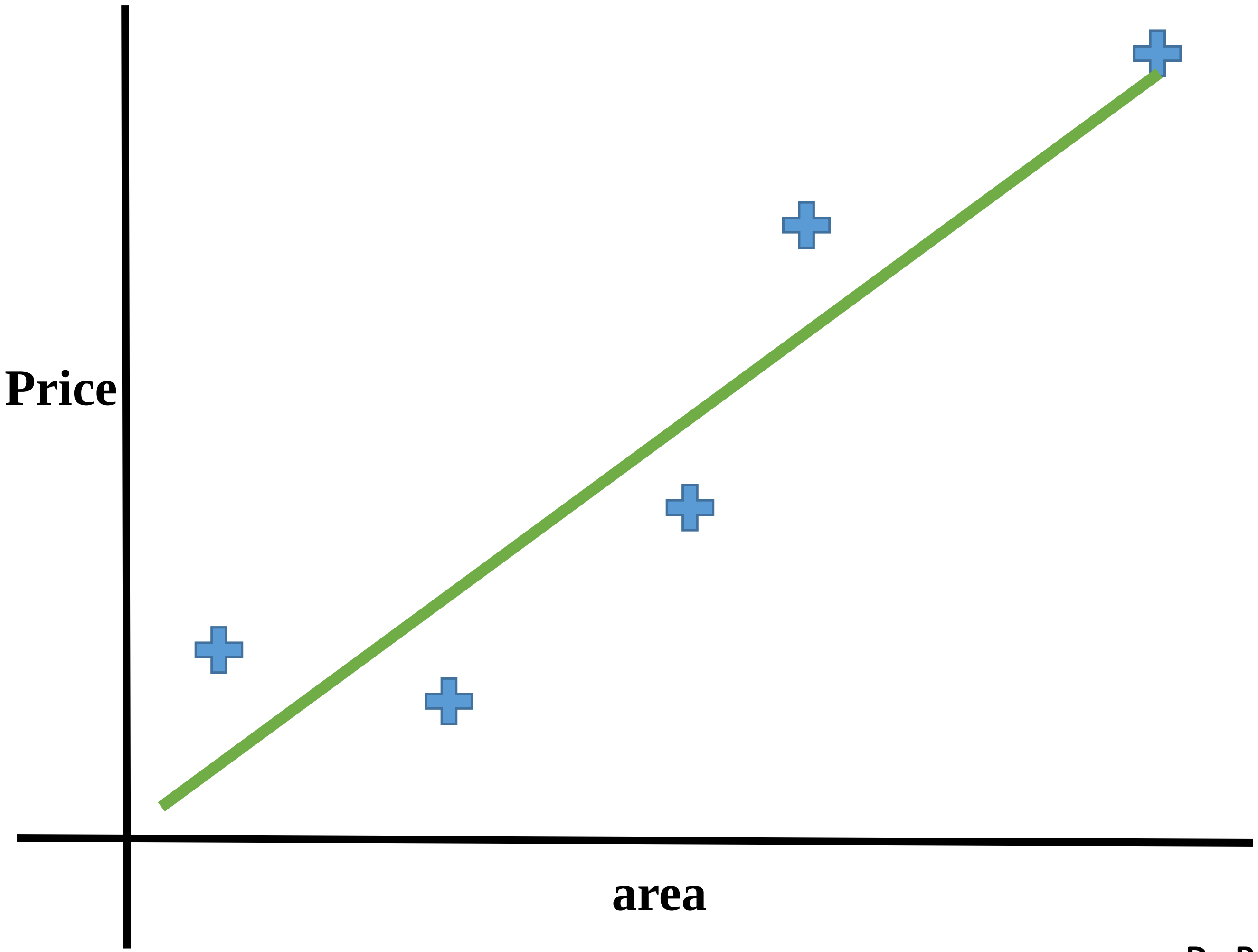
Dr. R. G. Tambe

Gradient Descent Algorithm

Gradient Descent Algorithm:

- An **Algorithm** to **Minimize** the **Function** by **Optimizing** its **Parameters**.

area	price
2600	550000
3000	565000
3200	610000
3600	680000
4000	725000



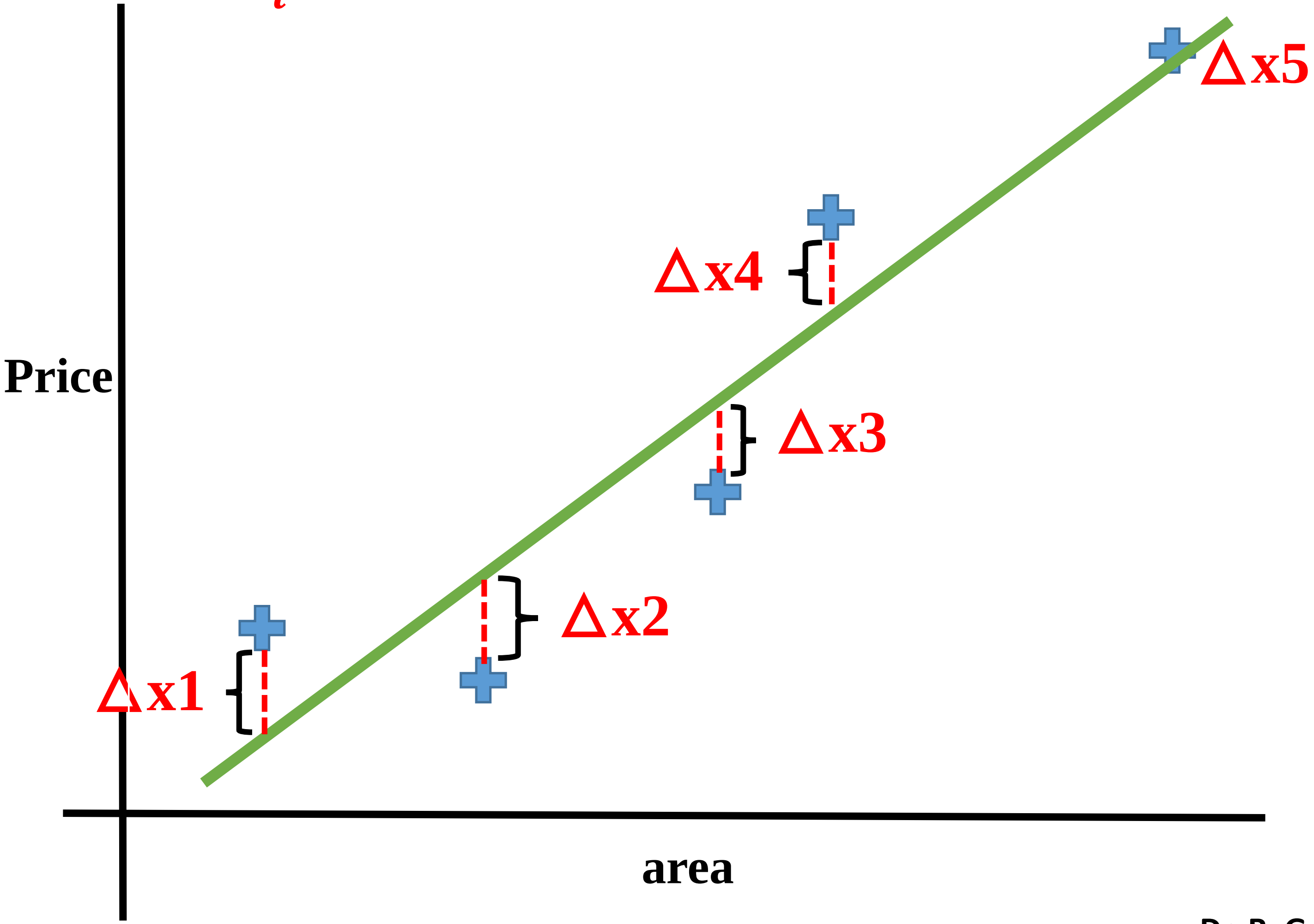
Gradient Descent Algorithm

Gradient Descent Algorithm:

- An **Algorithm** to **Minimize** the **Function** by **Optimizing** its **Parameters**.

area	price
2600	550000
3000	565000
3200	610000
3600	680000
4000	725000

ΔX_i = Actual Value – Predicted Value



Gradient Descent Algorithm

Gradient Descent Algorithm:

- An **Algorithm** to **Minimize** the **Function** by **Optimizing** its **Parameters**.

$$J(n) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

$$Mean\ Squared\ Error(MSE) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

$where,$

$y_i = Actual\ Value$

$\hat{y}_i = Predicted\ Value = mx_i + c$

Gradient Descent Algorithm

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \qquad MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - (mx_i + c\,))^2$$

Taking Partial Derivatives w.r.t. to slope (m).

$$\frac{d(MSE)}{dm} = \frac{2}{n}\sum_{i=1}^{n}(-x_i)(y_i - (mx_i + c\,))$$

$$0 = \sum_{i=1}^{n}(-x_i y_i) + \ mx_i^2 + cx_i))$$

$$\sum_{i=1}^{n} mx_i^2 + \sum_{i=1}^{n} cx_i = \sum_{i=1}^{n} x_i y_i$$

Gradient Descent Algorithm

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \qquad MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - (mx_i + c\,))^2$$

Taking Partial Derivatives w.r.t. to intercept (c).

$$\frac{d(MSE)}{dc} = \frac{2}{n}\sum_{i=1}^{n} -(y_i - (mx_i + c\,)$$

$$0 = \sum_{i=1}^{n} -y_i + \; mx_i + c)$$

$$\sum_{i=1}^{n} x_i + \sum_{i=1}^{n} c = \sum_{i=1}^{n} y_i$$

Dr. R. G. Tambe

Gradient Descent Algorithm

$$\sum_{i=1}^{n} mx_i^2 + \sum_{i=1}^{n} cx_i = \sum_{i=1}^{n} x_i y_i \qquad \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} c = \sum_{i=1}^{n} y_i$$

$$\frac{d(MSE)}{dm} \qquad \frac{d(MSE)}{dc}$$

To updated value of sploe (m) and intercept (c) are given by,

$$m_{new} = m_{old} - \lambda \left(\frac{d(MSE)}{dm_{old}} \right)$$

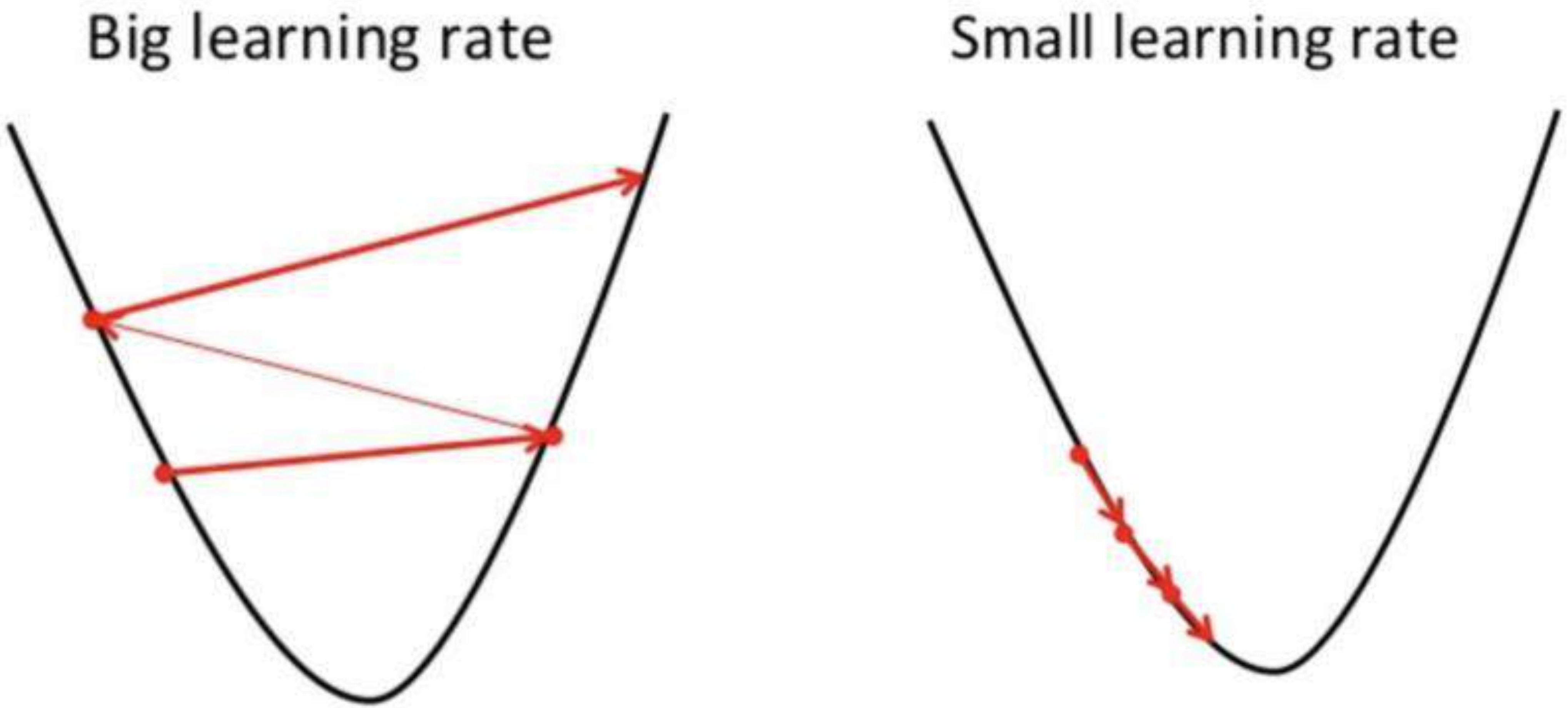
$$c_{new} = c_{old} - \lambda \left(\frac{d(MSE)}{dc_{old}} \right)$$

where,
$\lambda = Learning\ Rate$

Gradient Descent Algorithm

Learning Rate:

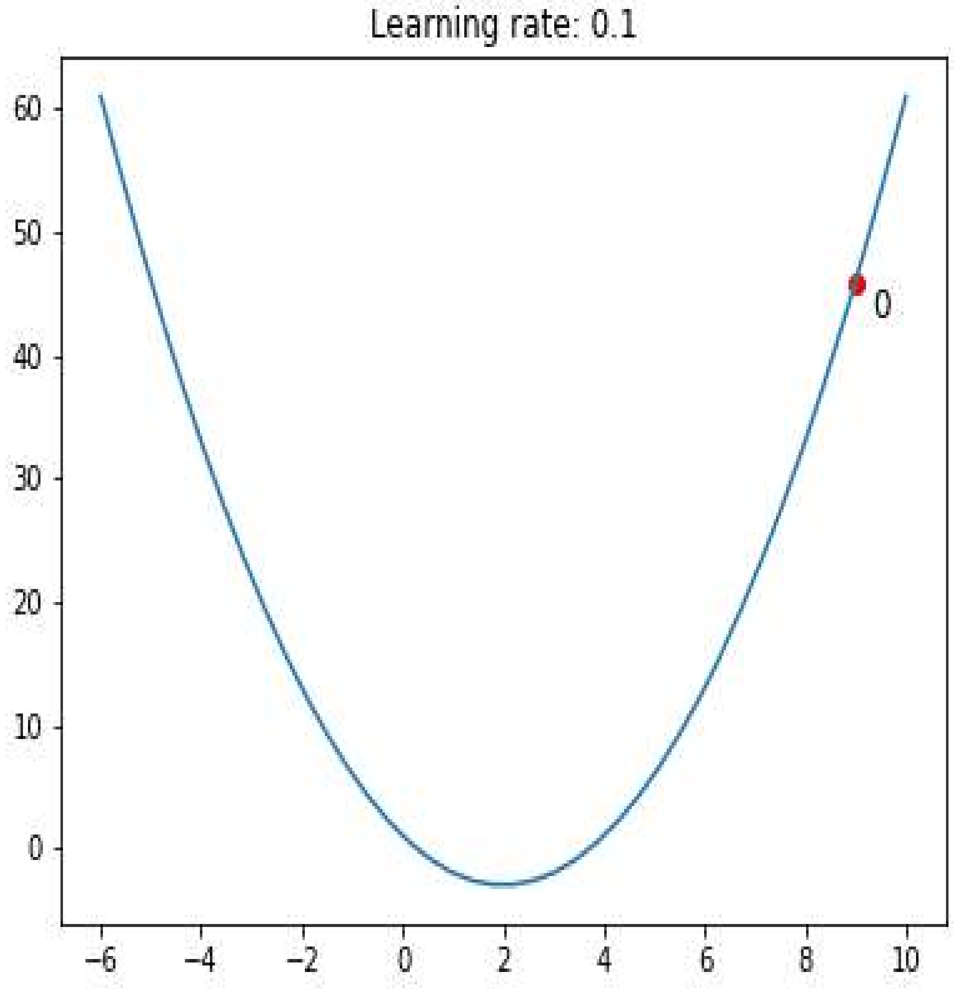
How big the steps the gradient descent takes into the direction of the local minimum are determined by the **learning rate**.

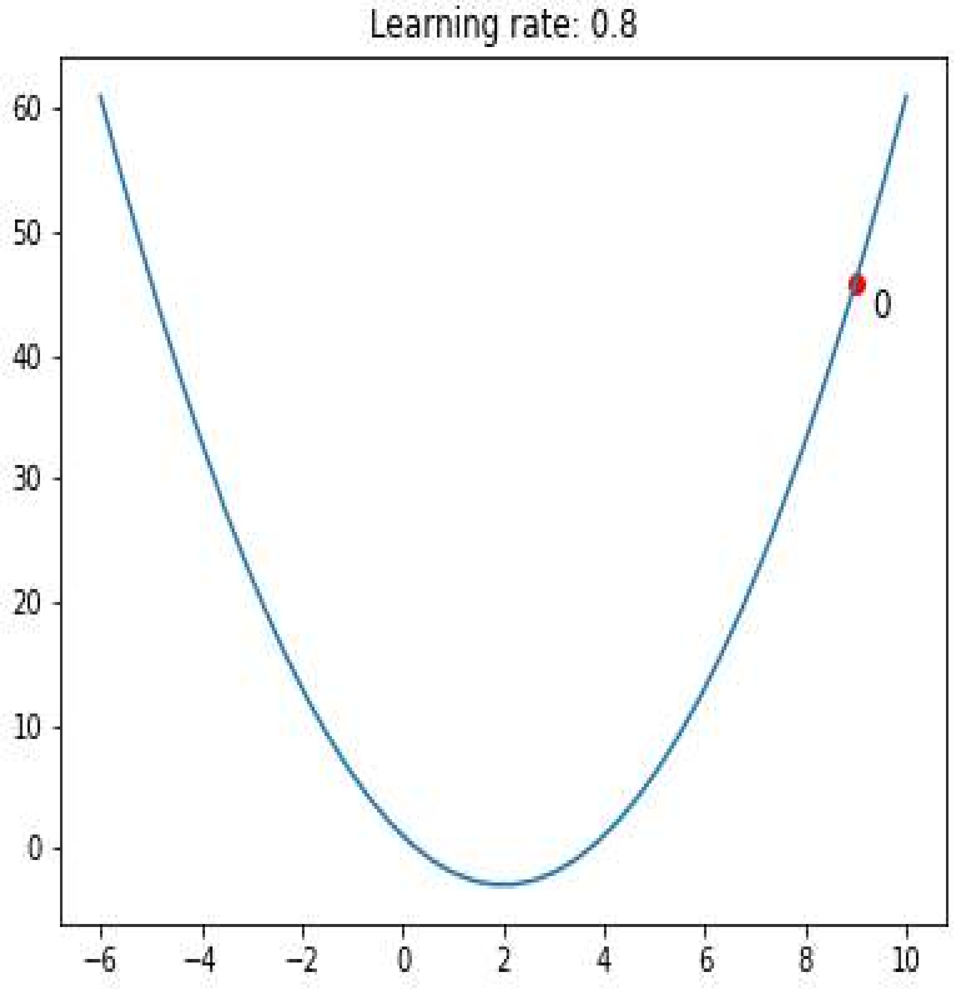


Gradient Descent Algorithm

Learning Rate:

How big the steps the gradient descent takes into the direction of the local minimum are determined by the **learning rate**.

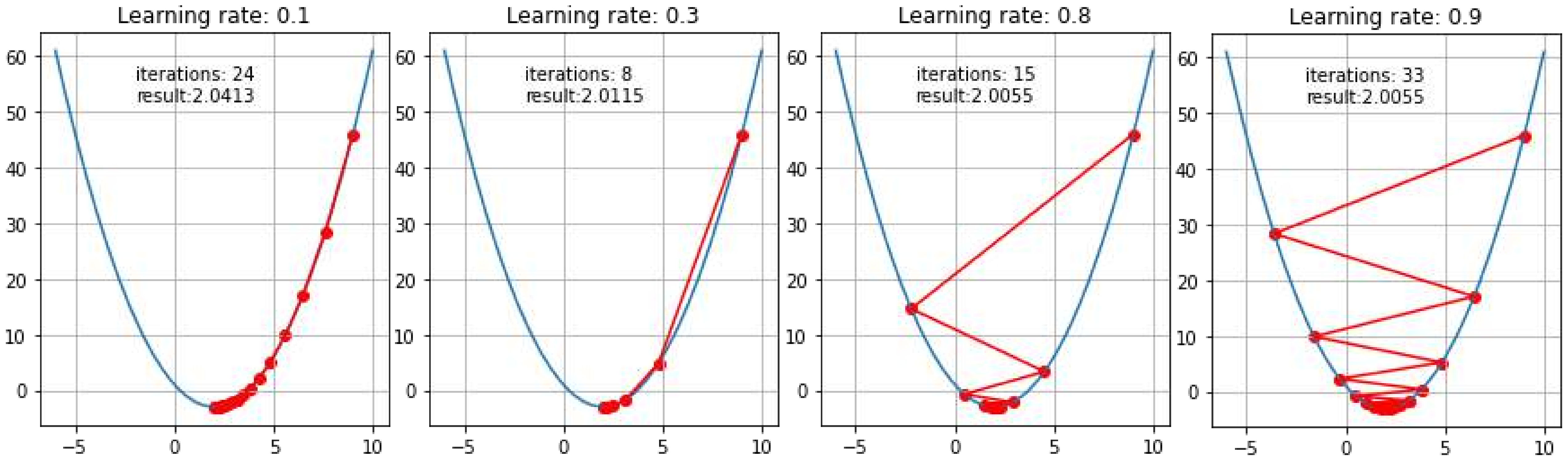




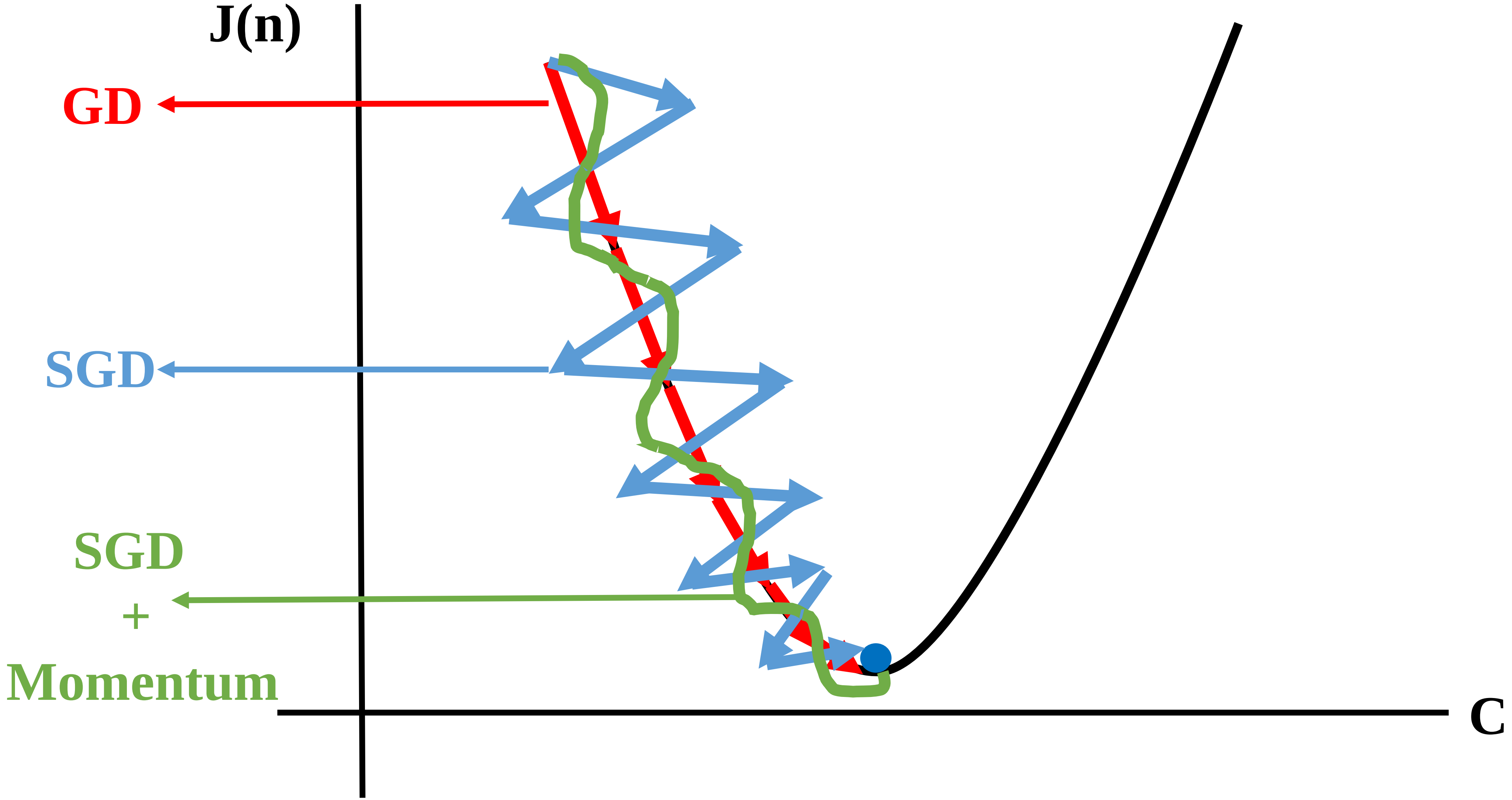
Gradient Descent Algorithm

Learning Rate:

How big the steps the gradient descent takes into the direction of the local minimum are determined by the **learning rate**.



Gradient Descent Algorithm



Dr. R. G. Tambe

Evaluation Metrics

$$Mean\ Squared\ Error(MSE) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

$$Mean\ Absolute\ Error(MAE) = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

$$R^2 = \frac{n(xy)-(\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

Assignment III

Q1. What is Bias and Variance? Explain trade-off between bias and variance.

Q2. What is underfitting and overfitting? Explain Reasons and ways to avoid underfitting and overfitting.

Q3. What is Regression Analysis? Explain Lasso Regression in detail.

Q4. Explain Gradient Descent Algorithm with its limitation.

Also Solve Problems on Next Slide as Q5.

Check your assignment on or before 24/09/2023

Q5 Given data points, predict value of Marks obtained by students if Correct_Ans of student is 11. Further calculate Regression Coefficient, Regression Line Equation and R^2 for the same.

Subjects/Samples	Correct_Ans	Marks
1	17	94
2	13	73
3	12	59
4	15	80
5	16	93
6	14	85
7	16	66
8	16	79
9	18	77
10	19	91
11	**11**	**?**

References

Test Books

1. Bishop, Christopher M., and Nasser M. Nasrabadi, “Pattern recognition and machine learning”,Vol. 4. No. 4. New York: springer, 2006.

2. Ethem Alpaydin, “ Introduction to Machine Learning”, PHI 2nd Edition-2013

Reference Books

1. Tom Mitchell, “Machine learning”, McGraw-Hill series in Computer Science.

2. Shalev-Shwartz, Shai, and Shai Ben-David, “Understanding machine learning: From theory to algorithms”, Cambridge university press, 2014.

3. Jiawei Han, Micheline Kamber, and Jian Pie, “Data Mining: Concepts and Techniques”, Elsevier Publishers 3 Edition.

4. Hastie, Trevor, et al., “The elements of statistical learning: data mining, inference, and prediction”, Vol. 2. New York: springer, 2009.

5. McKinney, “Python for Data Analysis “,O' Reilly media.

6. Trent hauk, “Scikit-learn”, Cookbook , Packt Publishing, ISBN: 9781787286382

7. Goodfellow I.,Bengio Y. and Courville, “ A Deep Learning”, MIT Press, 2016