Data Science and Artificial Intelligence

Machine Learning

Bias and Variance

Lecture No. 1

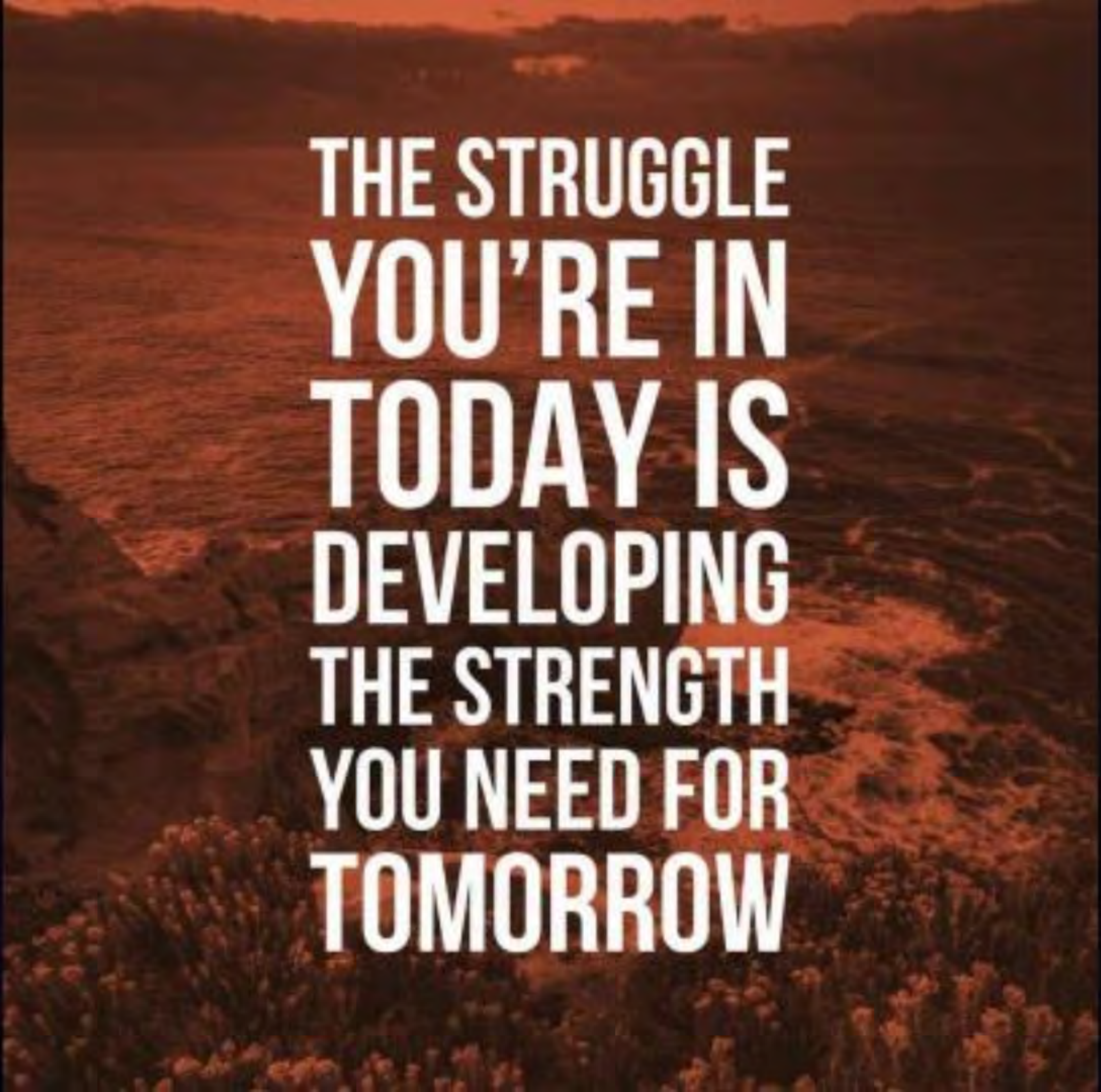
By- SIDDHARTH SABHARWAL SIR

Recap of Previous Lecture

- **Topic** — Time and space Complexity of DT
- **Topic** — Bias
- **Topic** — Variance
- **Topic**
- **Topic**

Topics to be Covered

- **Topic** Bias
- **Topic** Variance
- **Topic** Questions
- **Topic** Feature selection
- **Topic** Ensemble learning


THE STRUGGLE
YOU'RE IN
TODAY IS
DEVELOPING
THE STRENGTH
YOU NEED FOR
TOMORROW

$$\text{Bias} \Rightarrow \underbrace{\frac{1}{N}\sum_{i=1}^{N}|y_i - E(\hat{Y})|}$$ MAE → Training error.

REVISION

Bias ⇒ Training error

- $\text{Bias} = Y - E(\hat{Y})$ → It is avg of Predicted Value from all Predictors.

$\boxed{\text{data}}$ → 400 points Subsets ⇒ $\underbrace{\text{10 subsets}}$ ⇒ $\underbrace{\text{10 models}}$

REVISION

Variance $\Rightarrow E\left[\left(E(\hat{Y})-\hat{Y}\right)^2\right]$

(Proxy to test error)

REVISION

Overfit/underfit/Balanced fit

(done)

PW

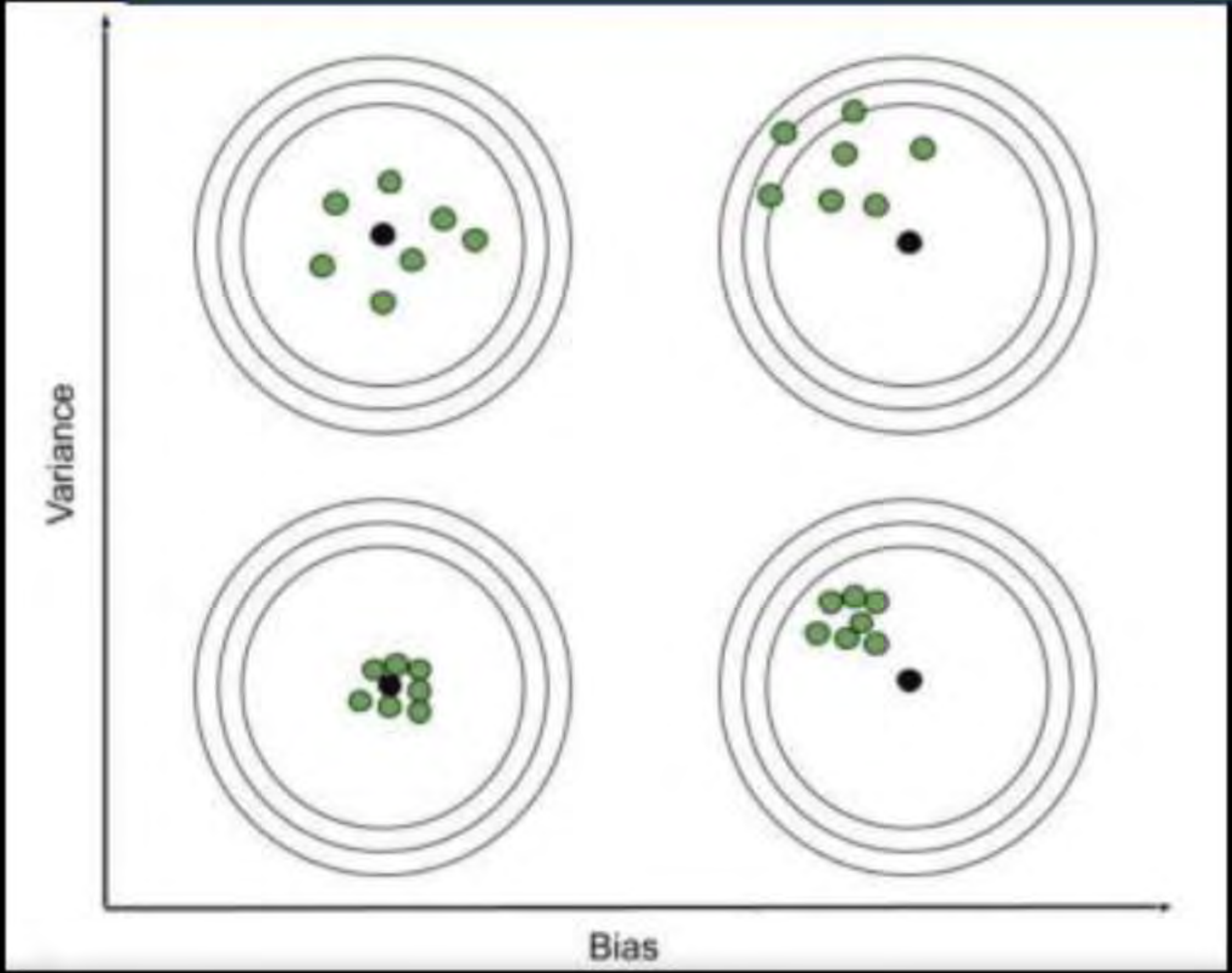
Different Combinations of Bias-Variance

There can be four combinations between bias and variance.

- **High Bias, Low Variance:** underfitting.
- **High Variance, Low Bias:** overfitting.
- **High-Bias, High-Variance:** A model is not able to capture the underlying patterns in the data (high bias) and is also too sensitive to changes in the training data (high variance). As a result, the model will produce inconsistent and inaccurate predictions on average.
- **Low Bias, Low Variance:** A model is able to capture the underlying patterns in the data (low bias) and is not too sensitive to changes in the training data (low variance). This is the ideal scenario for a machine learning model, as it is able to generalize well to new, unseen data and produce consistent and accurate predictions. But in practice, it's not possible.

Bias and Variance

Different Combinations of Bias-Variance



Looking at Bias and Variance in another way

$$\text{Bias} = Y - E(\hat{Y})$$

- Y: actual fxn, actual relation b/w Y and X
- $E(\hat{Y})$: avg of the model created

So we have some data
→ we created subsets and created model from each subset

$\rightarrow \hat{Y} \Rightarrow 2x+5$

$\rightarrow \hat{Y} \Rightarrow 2.3x+4.5$

$\rightarrow \hat{Y} \Rightarrow 1.9x+6.2$

$$E(\hat{Y}) \Rightarrow \frac{1}{3}(6.2x+15.7)$$

$$\boxed{\text{Bias} = Y - \frac{1}{3}(6.2x+15.7)}$$

Bias and Variance

Looking at Bias and Variance in another way

Variance $\Rightarrow E\left(\left(E(\hat{Y}) - \hat{Y}\right)^2\right)$

$\Rightarrow$ Expectation of $\left(E(\hat{Y}) - \hat{Y}\right)^2$ over the whole data.

So we have some data

$\hookrightarrow$ we created subsets and created model from each subset

$\rightarrow \hat{Y} \Rightarrow 2x+5$

$\rightarrow \hat{Y} \Rightarrow 2.3x+4.5$

$\rightarrow \hat{Y} \Rightarrow 1.9x+6.2$

$E(\hat{Y}) \Rightarrow \frac{1}{3}(6.2x+15.7)$

Mean Square error

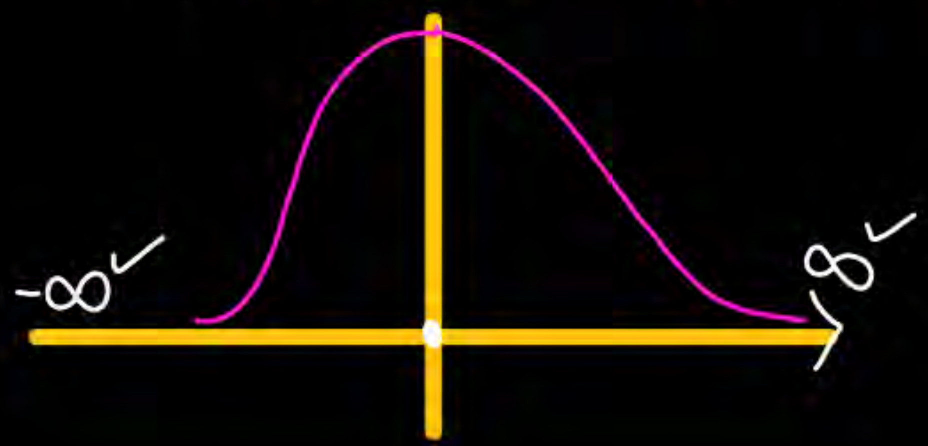
$\Rightarrow Y \Rightarrow$ actual function $\Rightarrow$ fix $\rightarrow$ Const

$\Rightarrow \hat{Y} \Rightarrow$ Predicted function by one subset $\Rightarrow$ Randomness.

$E(\hat{Y}) \Rightarrow$ average of all predicted fxn.

Noise
Gaussian PDF $\leftarrow$ $\varepsilon \Rightarrow$ noise in the data
Zero mean
σ^2 Variance

The actual values in data $\Rightarrow (Y+\varepsilon)$ — noise in data



$$\left\{ \widetilde{\epsilon_i^2} = \text{MSV of } \epsilon_i = \text{mean}^2 + \text{Variance} \atop = 0 + \sigma^2 \right\}$$

Mean square error $\Rightarrow$

$$\Rightarrow E\Big[\big(\underbrace{Y + \epsilon_i}_{\text{label of data}} - \hat{Y}\big)^2\Big] \Rightarrow \widetilde{(Y - \hat{Y} + \epsilon_i)^2}$$

$$\Rightarrow \widetilde{(Y-\hat{Y})^2 + \epsilon_i^2 + 2\epsilon_i(Y-\hat{Y})}$$

assume independent noise

$$\Rightarrow \widetilde{(Y-\hat{Y})^2} + \widetilde{\epsilon_i^2} + 2\tilde{\epsilon_i}\widetilde{(Y-\hat{Y})}$$

$$\Rightarrow E\big((Y-\hat{Y})^2\big) + \sigma^2 + \cancel{2 \times 0\,\widetilde{(Y-\hat{Y})}}$$

$$\Rightarrow \underbrace{E\big[(Y-\hat{Y})^2\big] + \sigma^2}$$

$$\text{So MSE} = E\big((\overset{\text{Const}}{Y} - \hat{Y})^2\big) + \sigma^2$$

$$\Rightarrow E\left[Y - \hat{Y} + E(\hat{Y}) - E(\hat{Y})\right]^2 + \sigma^2$$

$$\Rightarrow E\left[\big(Y - E(\hat{Y})\big) - \big(\hat{Y} - E(\hat{Y})\big)\right]^2 + \sigma^2$$

$$\Rightarrow E\left[\underbrace{\big(\overset{\text{Const}}{Y} - E(\hat{Y})\big)^2}_{\text{Const}}\right] + \underbrace{E\left[\hat{Y} - E(\hat{Y})\right]^2}_{\text{Variance}} - 2E\left[\big(Y - E(\hat{Y})\big)\big(\hat{Y} - E(\hat{Y})\big)\right] + \sigma^2$$

- $E(c) = C$
- Y = Const function
- $\hat{Y} \Rightarrow$ has Randomness
- $E(\hat{Y}) \Rightarrow$ Constant fxn.

$$(Y - E(\hat{Y}))^2 + E\left[(E(\hat{Y}) - \hat{Y})^2\right] - 2E\left[Y\hat{Y} - E(\hat{Y})\hat{Y} - YE(\hat{Y}) + E(\hat{Y})E(\hat{Y})\right] + \sigma^2$$

$$\text{Bias}^2 + \text{Variance} + \sigma^2 - 2\left[\cancel{YE(\hat{Y})} - \cancel{E(\hat{Y})E(\hat{Y})} - \cancel{YE(\hat{Y})} + \cancel{E(\hat{Y})E(\hat{Y})}\right]$$

$$\boxed{MSE = \underbrace{\text{Bias}^2 + \text{Variance}}_{\text{Controllable Part}} + \sigma^2}$$

Uncontrollable Part.

Bias and Variance

Mathematics for Bias and Variance

X	Y
1	5
3	6
2	10
8	16

$X \Rightarrow$ Random

$Y \Rightarrow$ Random

$\rightarrow$ Known

$\rightarrow$ Known

$\hat{Y} \Rightarrow$

$\hat{Y} \Rightarrow$ Predict

$\hat{Y} \Rightarrow$

$\hat{Y} \Rightarrow$

$E(\hat{Y}) \Rightarrow$

Bias and Variance

Mathematics for Bias and Variance

$$\Rightarrow \boxed{\left(MSE = \underbrace{Bias^2}_{\text{Train error}} + \underbrace{Var}_{\text{Test error}} + \sigma^2 \right)}$$

1000 points → 400 points for models training

$\hat{Y}$

$$(Y + \varepsilon_i - \hat{Y})^2$$

Bias and Variance

Region for the Least Value of Total Error

MSE

$$(Y+\varepsilon-\hat{Y})^2$$

$\Downarrow$

Tr error & Test error

Both effect

Conclusion

1. $\text{Bias} = Y - E(\hat{Y})$
2. $\text{Variance} = E\left(\left(\hat{Y} - E(\hat{Y})\right)^2\right)$
3. Variance ↓
4. Bias ↓
5. $MSE = \text{Bias}^2 + \text{Variance} + \sigma^2$

Bias and Variance

Practice

ABC

3) Which of the following statements are True? Check all that apply: ***1 point***

True

underfit

- [x] If a learning algorithm is suffering from high bias, only adding more training examples may **not** improve the test error significantly.
- [x] A model with more parameters is more prone to overfitting and typically has a higher variance.

  → Complex → Overfit

- [x] When debugging learning algorithms, it is useful to plot a learning curve to understand if there is a high bias or high variance problem.
- [ ] Increasing degree of the polynomial in curve fitting will increase the bias in the model

degree inc → overfit → Bias↓↓

Ridge

Practice

5) Suppose you have implemented a regularized linear regression model. You observe that on the held out testing set, the model makes unacceptably large errors with its predictions. However, you observe that the model performs well (has a low error) on the training set. Which of the following steps can be incorporated to lower the error on testing dataset. Select all that apply. ***1 point***

- ✓ Try using a smaller set of the features
- ~~Try decreasing the regularization parameter λ~~
- ✓ Get more training examples
- Use fewer training examples

• Overfitting Bias↓ Var↑↑

• λ inc ⇒ β's → 0

$(x^1, x^2, x^3) \rightarrow$ linear Reg $\longrightarrow \beta_0 + \beta_1 x^1 + \beta_2 x^2 + \beta_3 x^3$

$(\underbrace{x^1, x^2, x^3, (x^1)^2, x_1 x_2, x_3^2}_{\text{Polynomial dimension}}) \longrightarrow$ linear Reg $\Rightarrow \beta_0 + \beta_1 x^1 + \beta_2 x^2$

$+ \beta_3 x^3 + \beta_4 (x^1)^2 + \beta_5 x_1 x_2$

$+ \beta_6 (x_3)^2$

Practice

6) Suppose you have implemented a regularized linear regression model. You observe that on the held out testing set, the model makes unacceptably large errors with its predictions. Furthermore, you observe that the model performs **poorly** on the training set. Which of the following steps can be incorporated to lower the error on the testing dataset. Select all that apply ***1 point***

- [x] Try to obtain an additional set of features
- [ ] Try increasing the regularization parameter λ
- [x] Get more training examples
- [x] Try adding polynomial features

$\Rightarrow$ $\underbrace{\text{poore train \& test}}$

underfit

- λ is v. large
  β's → 0
- Reduce λ

Practice

7) Suppose you are training a regularized linear regression model. Check which of the following ***1 point*** statements are true? Select all that apply.

- The regularization parameter λ value is chosen so as to give the lowest training set error
- The regularization parameter λ value is chosen so as to give the lowest cross validation error
- The regularization parameter λ value is chosen so as to give the lowest test set error
- The performance of a learning algorithm on the training set will typically be better than its performance on the test set

Bias and Variance

Practice

Q1. Impact of high variance on the training set ?

- A. overfitting
- B. underfitting
- C. both underfitting & overfitting
- D. depends upon the dataset

Bias and Variance

Practice

Q2. How does the bias-variance decomposition of a ridge regression estimator compare with that of ordinary least squares regression?

- A. ridge has larger bias, larger variance
- B. ridge has smaller bias, larger variance
- C. ⇒ ridge has larger bias, smaller variance
- D. ridge has smaller bias, smaller variance

⇒ Ridge ⇒ large bias
low variance

OLS
overfit
Bias = 0
Variance = high

Bias and Variance

Practice

Q4. You trained a binary classifier model which gives very high accuracy on the training data, but much lower accuracy on validation data. Which is false.

- A. this is an instance of overfitting
- B. this is an instance of underfitting
- C. the training was not well regularized
- D. the training and testing examples are sampled

Bias and Variance

high variance ⇒ overfit

Practice

Q5. Suppose your model is demonstrating high variance across the different training sets. Which of the following is NOT valid way to try and reduce the variance?

- A. increase the amount of training data in each training set ✓
- B. improve the optimization algorithm being used for error minimization.
- C. decrease the model complexity OK
- D. reduce the noise in the training data

ex OLS → Ridge Reg

Practice

Q6. Which of the following are components of generalization Error?

- A. bias
- B. variance
- C. both of them
- D. none of them

data $\rightarrow$ Subsets $\rightarrow$ we create models from subset

$\Downarrow$

models $\Rightarrow$ hypothesis.

Practice

we created many models from data

Q7. Which one of the following is suitable? 1. When the hypothesis space is richer, overfitting is more likely. 2. when the feature space is larger , overfitting is more likely.

→ not possible

→ dimensions

Hypothesis space ⇒

- A. true, false
- B. false, true
- C. true, true
- D. false, false

large No of D

⇓

Large Complex

⇓

Overfitting

Practice

Q8. MLE estimates are often undesirable because

Skip

- A. they are biased
- B. they have high variance
- C. they are not consistent estimators
- D. none of the above

Practice

Q9. Suppose, you got a situation where you find that your linear regression model is under fitting the data. In such situation which of the following options would you consider?

- A. you will add more features
- B. you will remove some features
- C. all of the above
- D. none of the above

Underfit → Balanced fit

Practice

Q10. We have been given a dataset with n records in which we have input attribute as x and output attribute as y. Suppose we use a linear regression method to model this data. To test our linear regressor, we split the data in training set and test set randomly. Now we increase the training set size gradually. As the training set size increases, What do you expect will happen with the mean training error?

- ~~A.~~ increase
- ~~B.~~ decrease
- C. remain constant
- D. can't say

Bias and Variance

Practice

Q11. We have been given a dataset with n records in which we have input attribute as x and output attribute as y. Suppose we use a linear regression method to model this data. To test our linear regressor, we split the data in training set and test set randomly. What do you expect will happen with bias and variance as you increase the size of training data?

- A. bias increases and variance increases
- B. bias decreases and variance increases
- C. bias decreases and variance decreases ⇒ model becomes better
- D. bias increases and variance decreases

Bias and Variance

Practice

Q12. Regarding bias and variance, which of the following statements are true? (Here 'high' and 'low' are relative to the ideal model.

(i) Models which overfit are more likely to have high bias → No 0 Bias

(ii) Models which overfit are more likely to have low bias ✓

(iii) Models which overfit are more likely to have high variance ✓

(iv) Models which overfit are more likely to have low variance X

Bias and Variance

Practice

Q13. In terms of bias and variance. Which of the following is true when you fit degree 2 polynomial?

wrt linear regression.

- A. bias will be high, variance will be high
- B. bias will be low, variance will be high Overfit ⇒ model Complex ↑↑
- C. bias will be high, variance will be low
- D. bias will be low, variance will be low

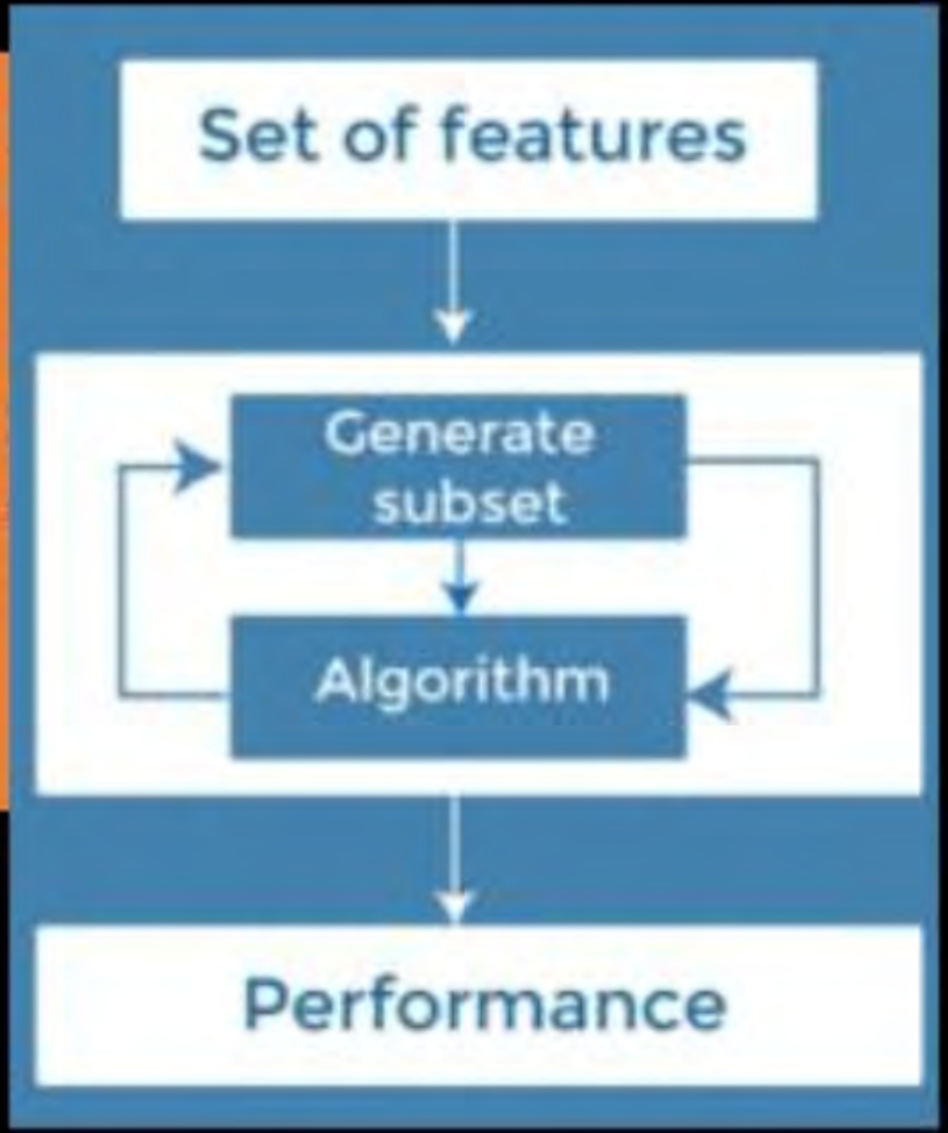
Feature Selection Methods

The Role of Feature Selection

1. To reduce the dimensionality of feature space.
2. To speed up a learning algorithm.
3. To improve the predictive accuracy of a classification algorithm.
4. To improve the comprehensibility of the learning results.

Wrapper Methods

- Here selection of features is done by considering it as a search problem, in which different combinations are made, evaluated, and compared with other combinations. It trains the algorithm by using the subset of features iteratively.
- These are computationally extensive



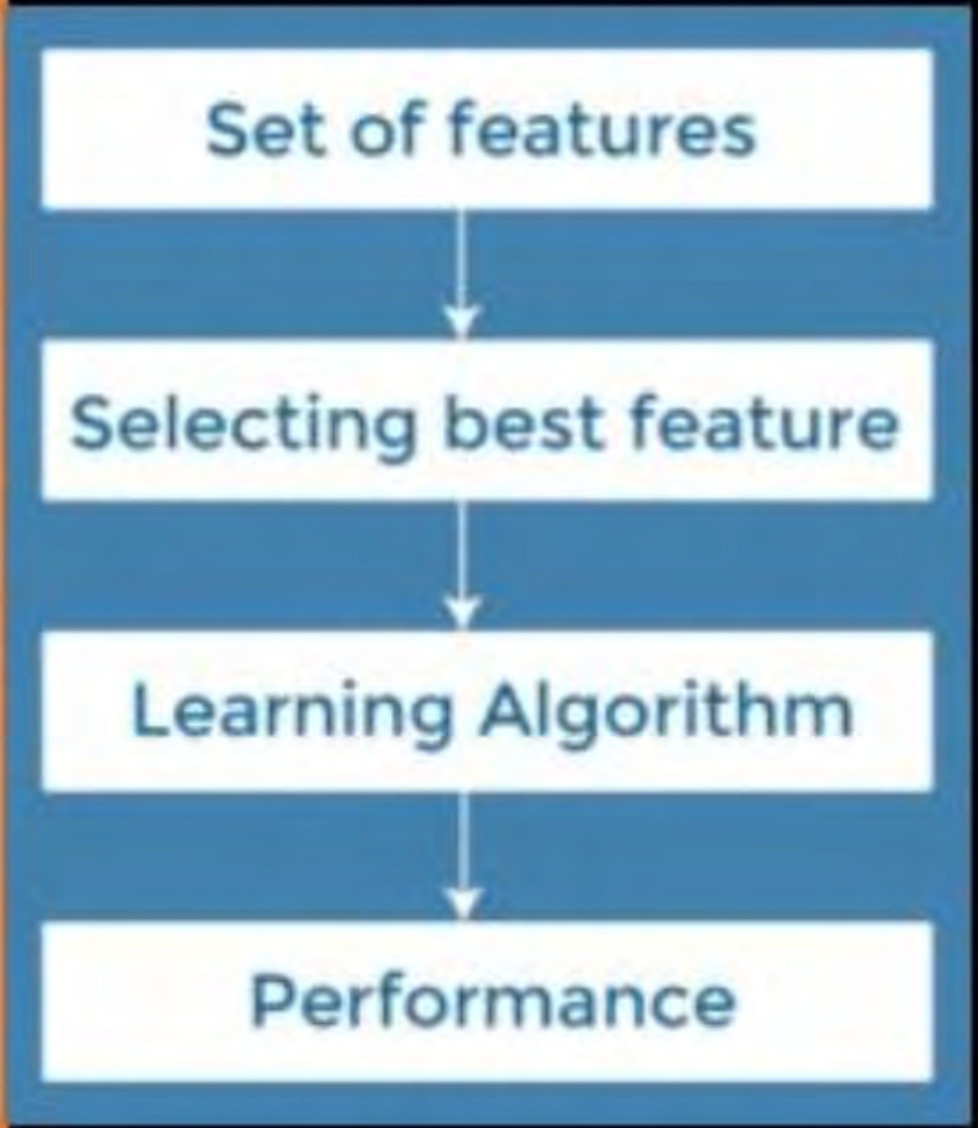
Wrapper Methods

Some techniques of wrapper methods are: (Forward- and Backward-Stepwise Selection)

- Forward selection - Forward selection is an iterative process, which begins with an empty set of features. After each iteration, it keeps adding on a feature and evaluates the performance to check whether it is improving the performance or not. The process continues until the addition of a new variable/feature does not improve the performance of the model.
- Backward elimination - Backward elimination is also an iterative approach, but it is the opposite of forward selection. This technique begins the process by considering all the features and removes the least significant feature. This elimination process continues until removing the features does not improve the performance of the model.
- Exhaustive Feature Selection- Exhaustive feature selection is one of the best feature selection methods, which evaluates each feature set as brute-force. It means this method tries & make each possible combination of features and return the best performing feature set.

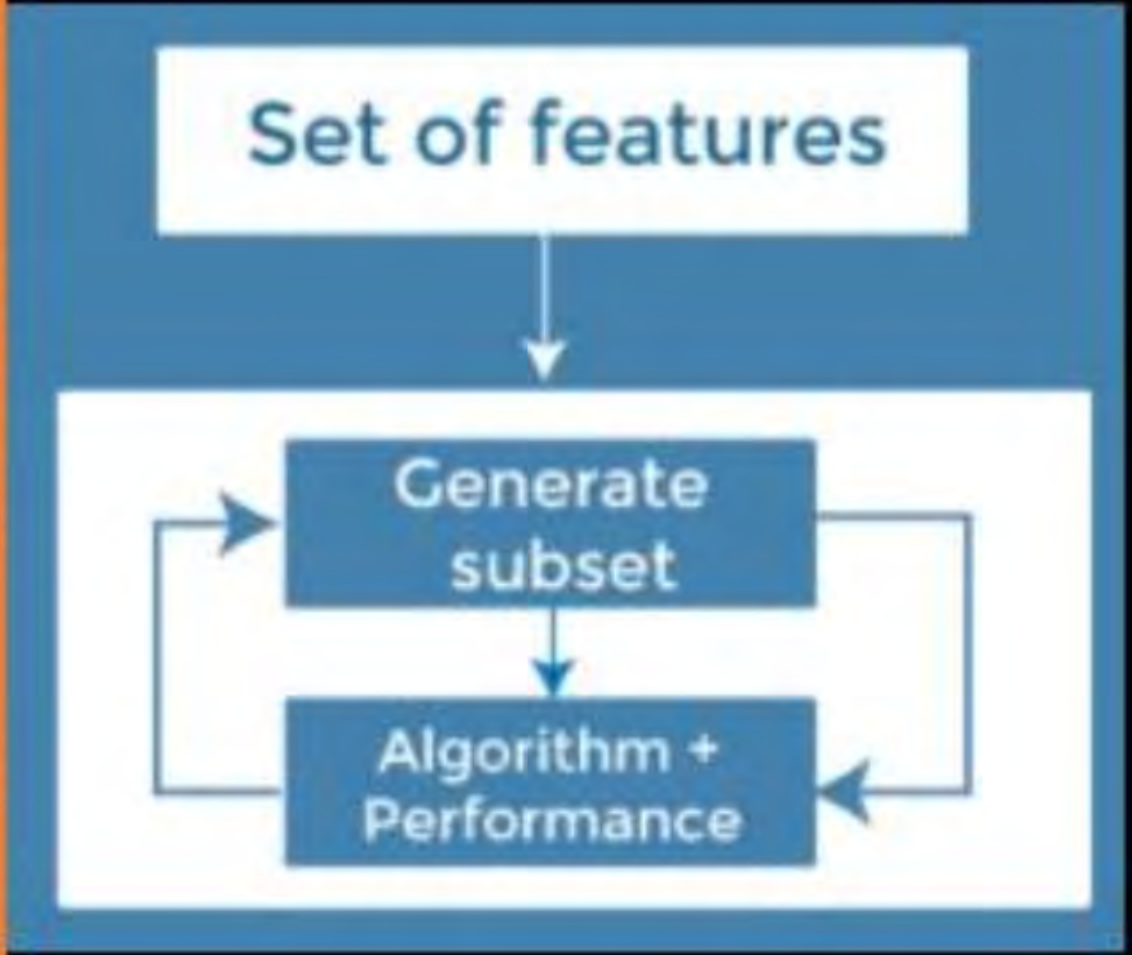
Filter Methods

- In Filter Method, features are selected on the basis of statistics measures. This method does not depend on the learning algorithm and chooses the features as a pre-processing step.
- Actually we find the features which are having maximum correlation with the output or label.
- The filter method filters out the irrelevant feature and redundant columns from the model by using different metrics through ranking.
- The advantage of using filter methods is that it needs low computational time and does not overfit the data.



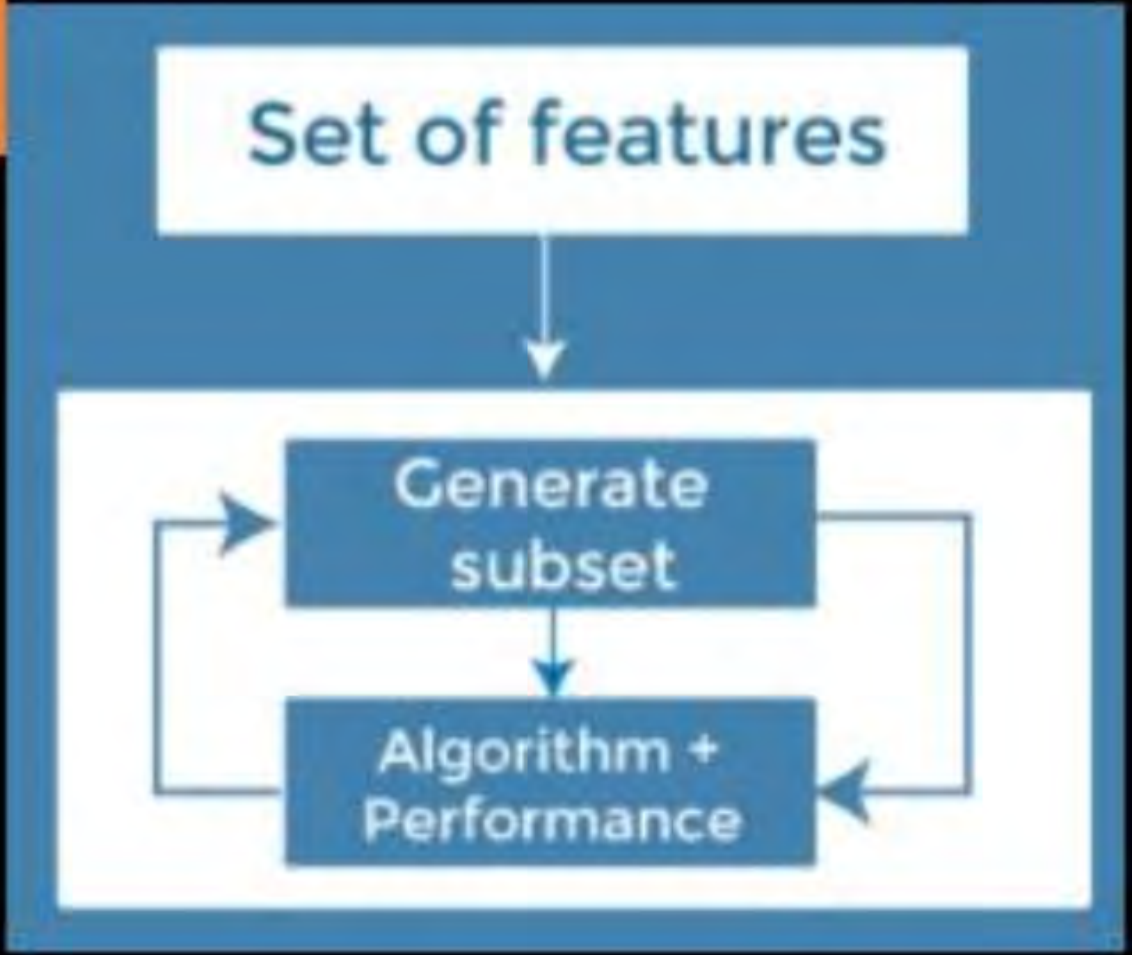
Embedded Methods

- The above methods are used when the dataset is small. But when the dataset is large then we use Embedded methods
- Embedded methods combined the advantages of both filter and wrapper methods by considering the interaction of features along with low computational cost. These are fast processing methods similar to the filter method but more accurate than the filter method.
- Regularisation and Tree based methods.
- These methods are also iterative, which evaluates each iteration, and optimally finds the most important features that contribute the most to training in a particular iteration.



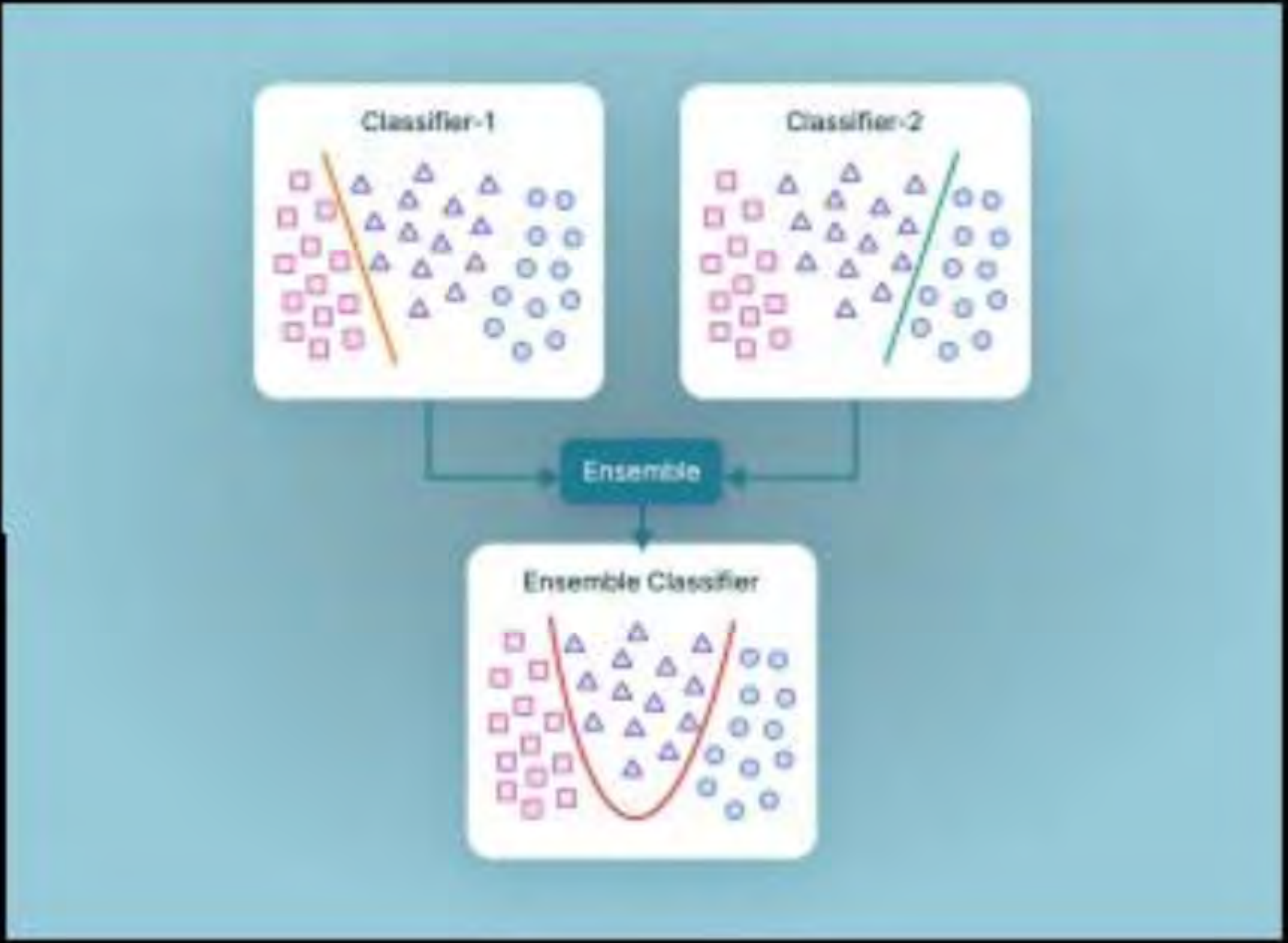
Embedded Methods

- Regularisation – Ridge regression
- Tree based methods –Random forest etc.



Ensemble learning

- ❖ Don't consult only one expert but consult many expert before taking the final decision.
- ❖ Ensemble learning helps improve machine learning results by combining several models.



- combine the outputs of diverse models to create a more precise prediction.

Few simple but powerful techniques, namely:

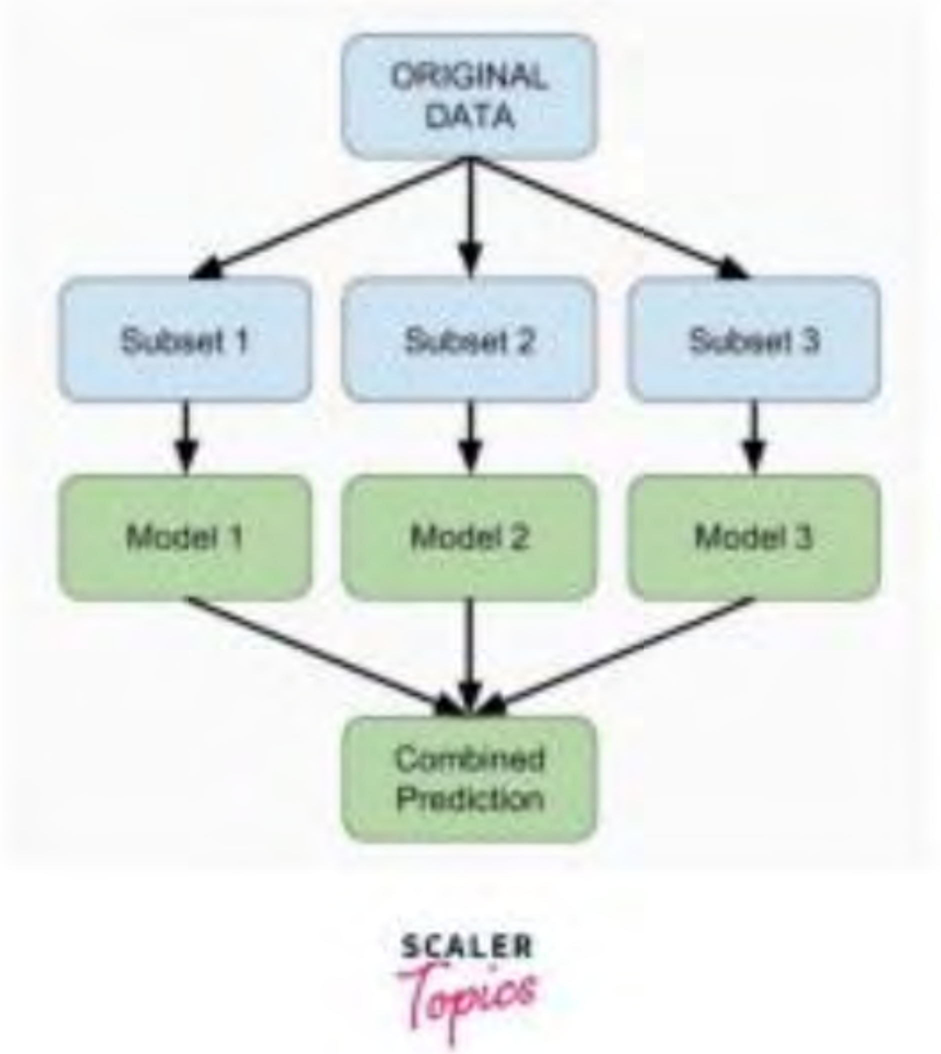
1. Max Voting
2. Averaging
3. Weighted Averaging

Ensemble learning

- This make the model more generalised and thus the test error and the train error so bias and variance decreases.
- Lets prove that the variance decreases...

Ensemble learning

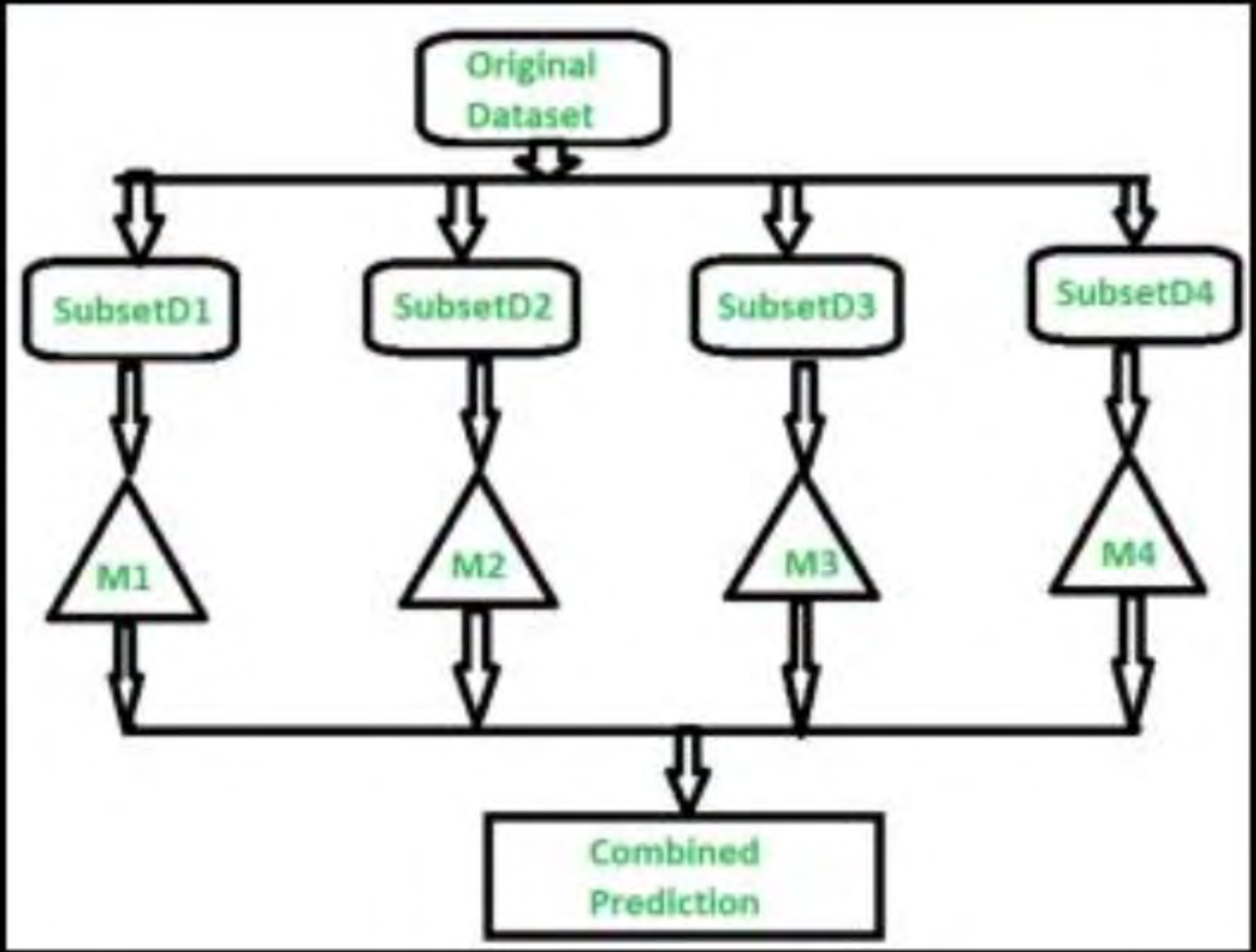
- All these models are called the base learners.
- These base learners can take different algorithms.
- And also we can give different training data to each of the model.
- These base learners are also called weak learners.



Ensemble learning (bagging/Bootstraping)

❖ **Types of Ensemble Classifier – Bagging:**

1. In Bootstrapping Multiple subsets are created from the original data set with equal tuples, selecting observations with replacement.
2. But in Bagging we can create subset of different sizes
3. A base model is created on each of these subsets. (these are called the weak model)
4. Each model is learned in parallel from each training set and independent of each other.
5. The final predictions are determined by combining the predictions from all the models.

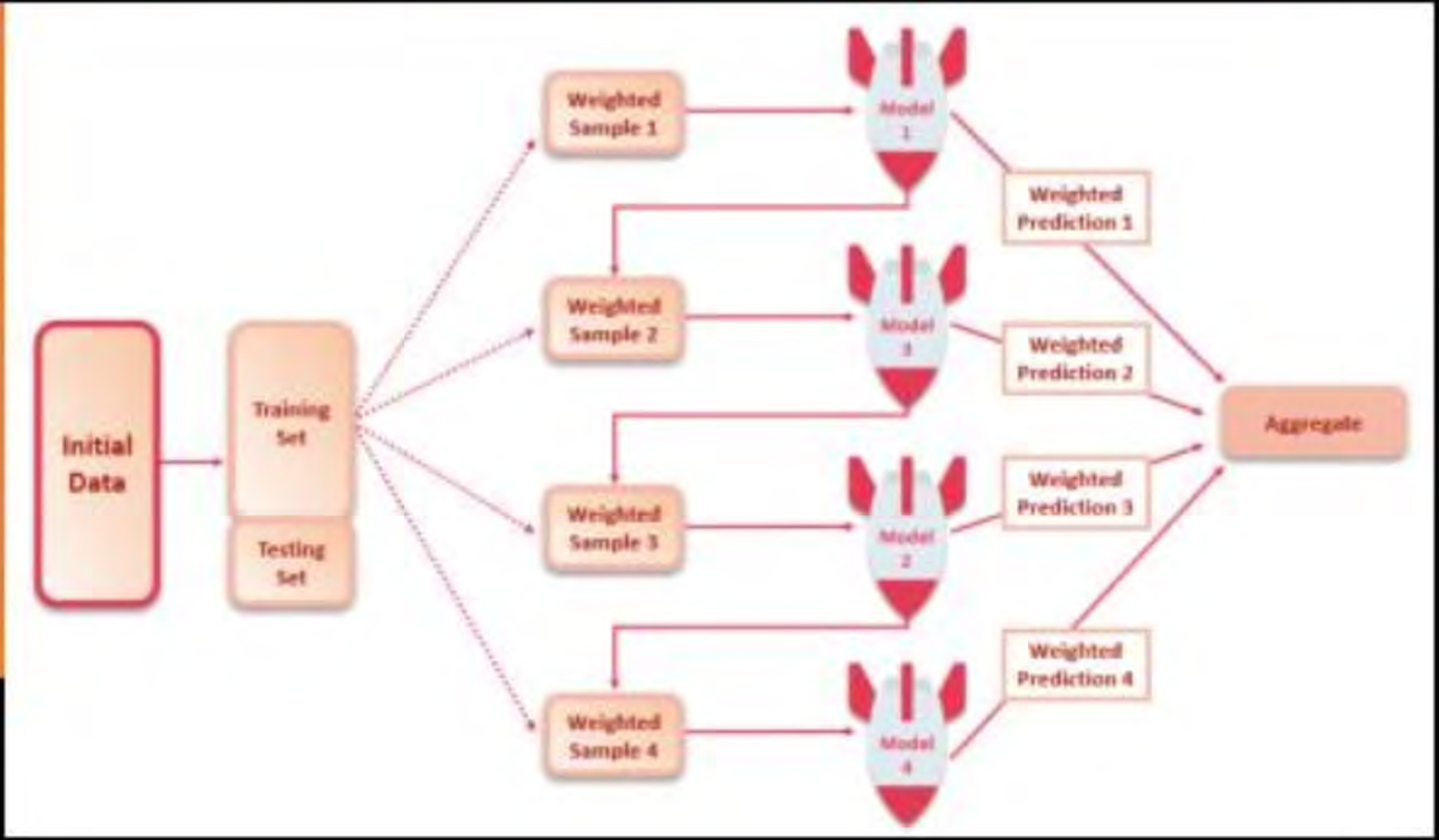


Ensemble learning

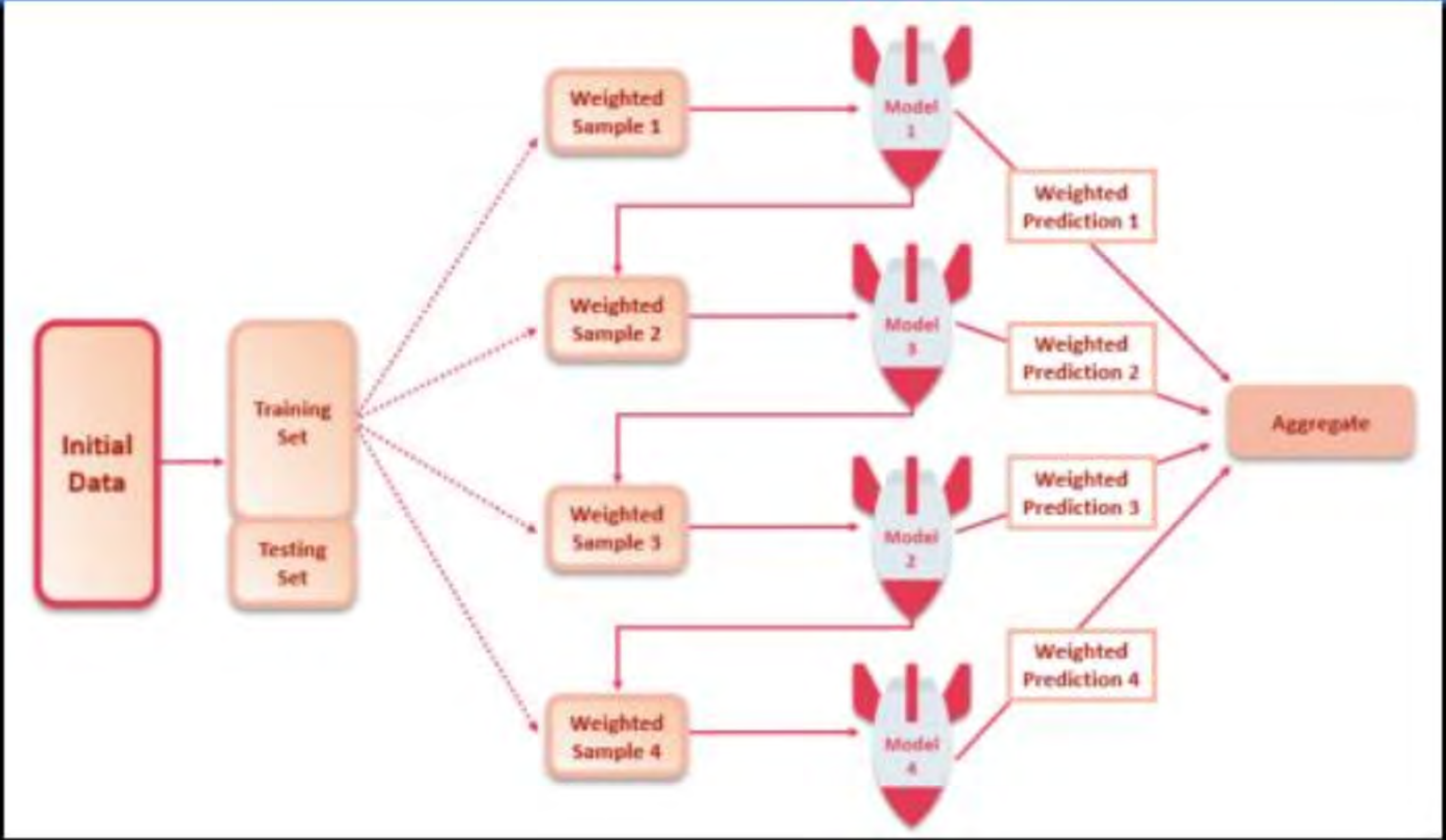
- Actually problem in DT is that it become too large on big dataset thus we use Ensemble learning here, So we break the training data into subsets and then train my model on these subsets.
- But we can have the problem that the models are not able to catch the pattern and lead to more bias...

Ensemble learning

- Types of Ensemble Classifier – Boosting:
- This is like Bagging.
- But this is not a parallel process rather a sequential process...
- Here we first learn a model and find the error on the data and then train next model where we have more error...



Ensemble learning - Boosting



Ensemble learning - Boosting

1. Samples generated from the training set are assigned the **same weight** to start with. These samples are used to train a homogeneous weak learner or base model.
2. The prediction error for a sample is calculated – **the greater the error, the weight of the sample increases**. Hence, the sample becomes more important for training the next base model.
3. The individual learner is weighted too – **does well on its predictions, gets a higher weight assigned to it**. So, a model that outputs good predictions will have a higher say in the final decision.
4. The weighted data is then passed on to the following base model, and steps 2) and 3) are repeated until the **data is fitted well enough to reduce the error below a certain threshold**.
5. When new data is fed into the boosting model, it is passed through all individual base models, and **each model makes its own weighted prediction**.
6. Weight of these models is used to generate the final prediction. The predictions are scaled and **aggregated to produce a final prediction**.

Bias and Variance

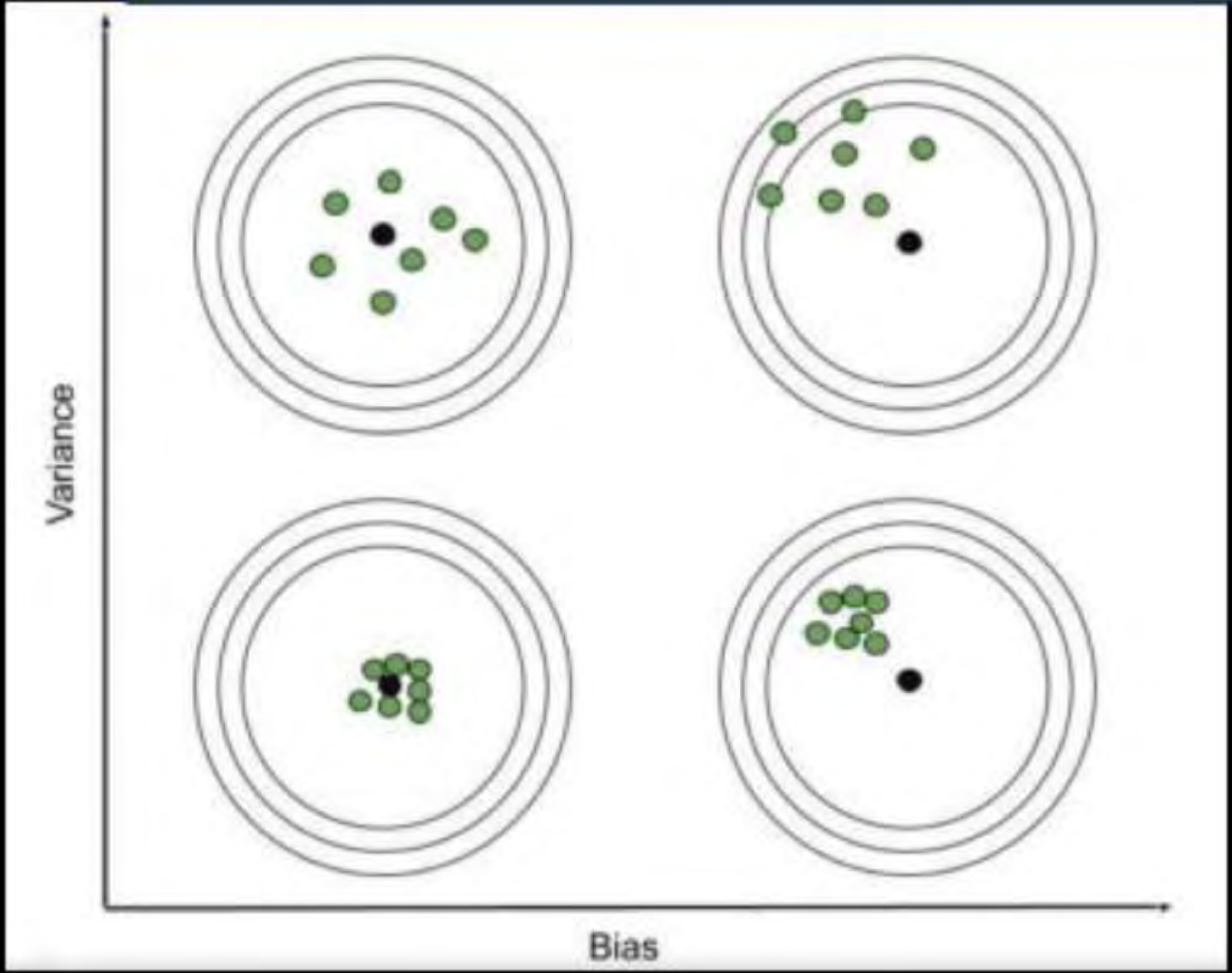
Different Combinations of Bias-Variance

There can be four combinations between bias and variance.

- **High Bias, Low Variance:** underfitting.
- **High Variance, Low Bias:** overfitting.
- **High-Bias, High-Variance:** A model is not able to capture the underlying patterns in the data (high bias) and is also too sensitive to changes in the training data (high variance). As a result, the model will produce inconsistent and inaccurate predictions on average.
- **Low Bias, Low Variance:** A model is able to capture the underlying patterns in the data (low bias) and is not too sensitive to changes in the training data (low variance). This is the ideal scenario for a machine learning model, as it is able to generalize well to new, unseen data and produce consistent and accurate predictions. But in practice, it's not possible.

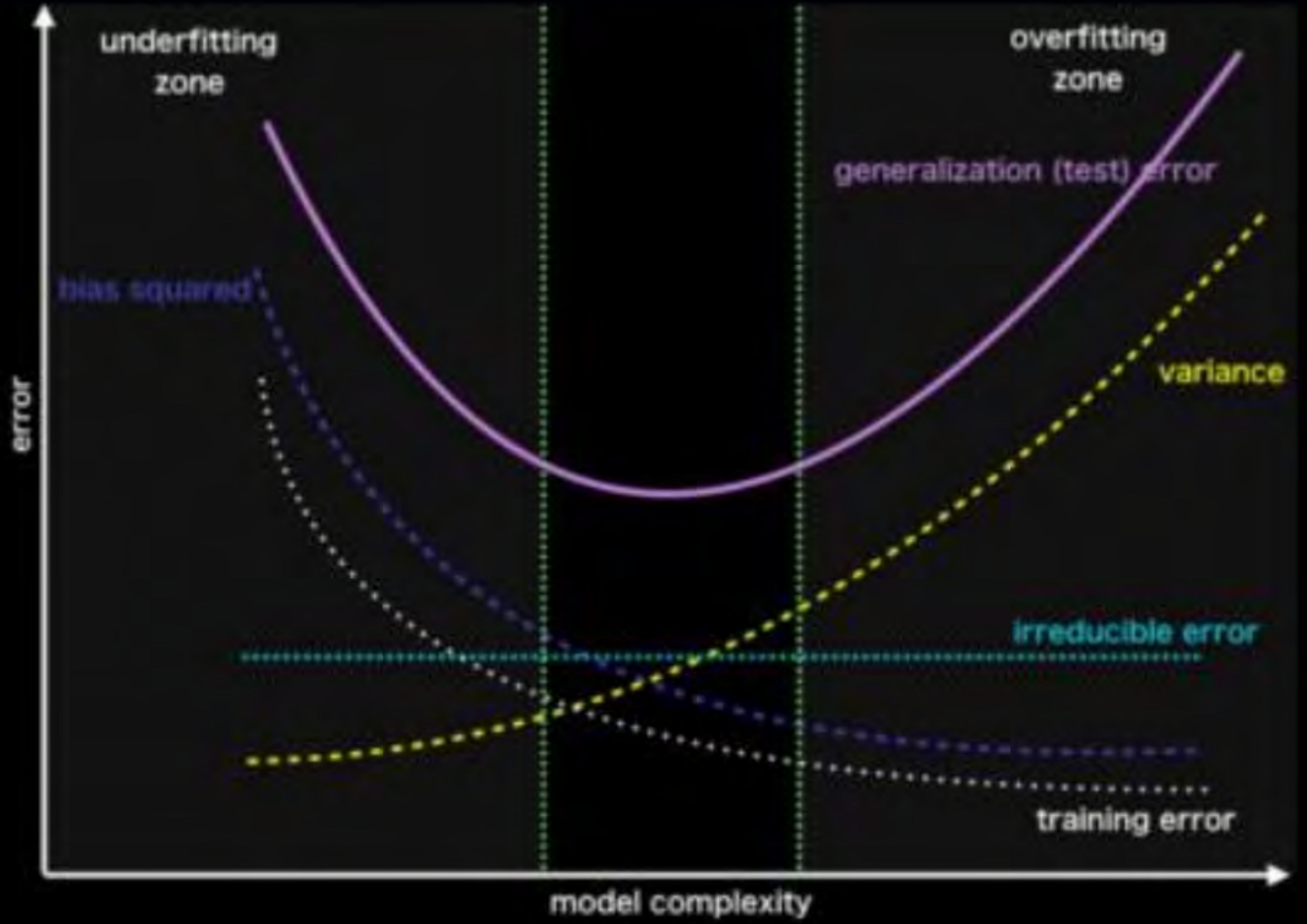
Bias and Variance

Different Combinations of Bias-Variance



Bias and Variance

Bias Variance Tradeoff



THANK - YOU