Data Science and Artificial Intelligence

Machine Learning

Bias and Variance

Lecture No. 3

By- SIDDHARTH SABHARWAL SIR

Recap of Previous Lecture

Topic

Topic

Topic

Topic

Topic

Featune selection Technique

Ensemble method → Bagging
→ Boosting

Topics to be Covered

Topic Random Forest

Topic MLE | MAP

Topic

Topic

Topic

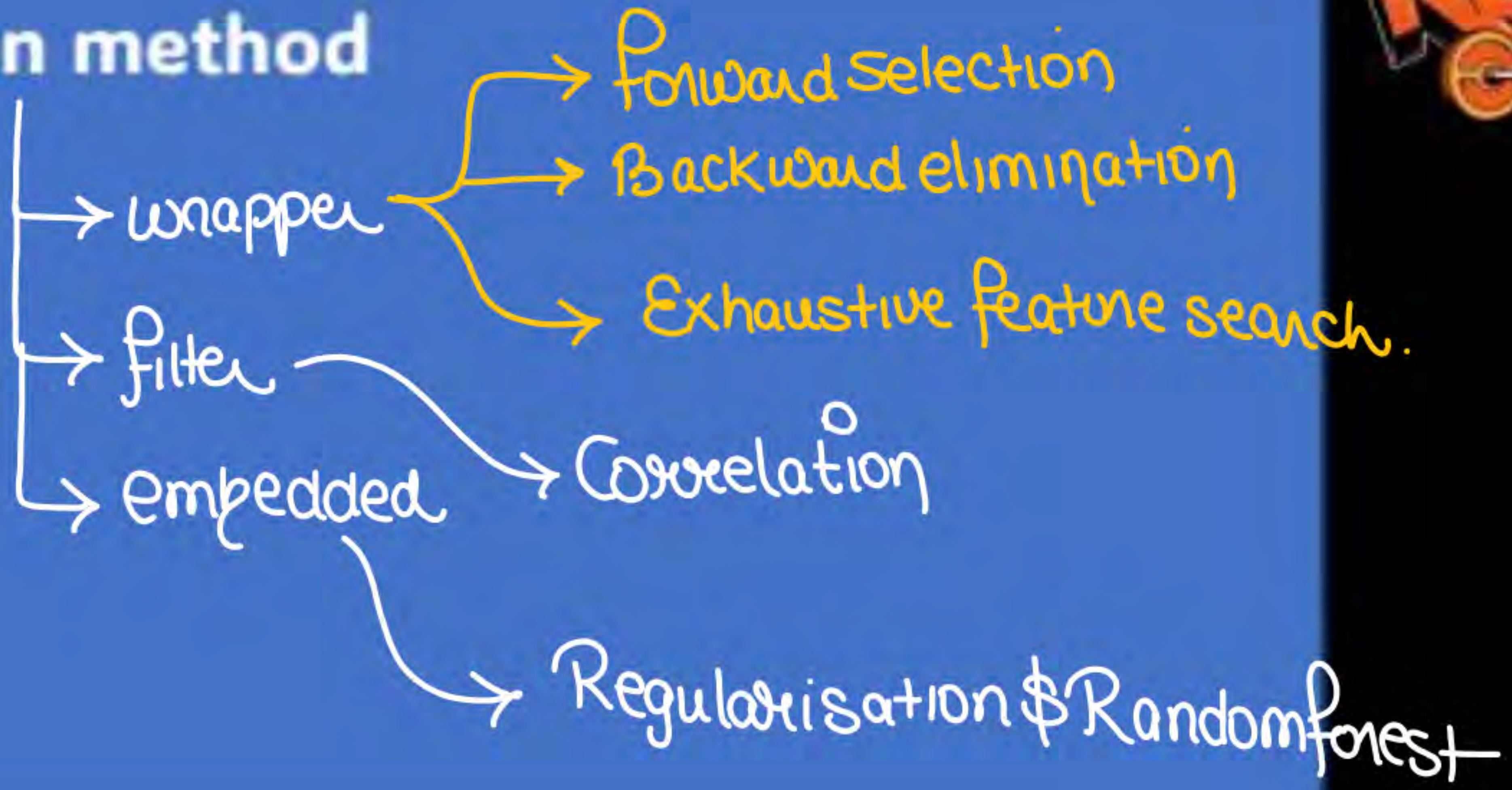
"Strength and growth come only through continuous effort and struggle."

- NAPOLEON HILL -

MARATHON HANDBOOK

@MARATHON.HANDBOOK

Feature selection method



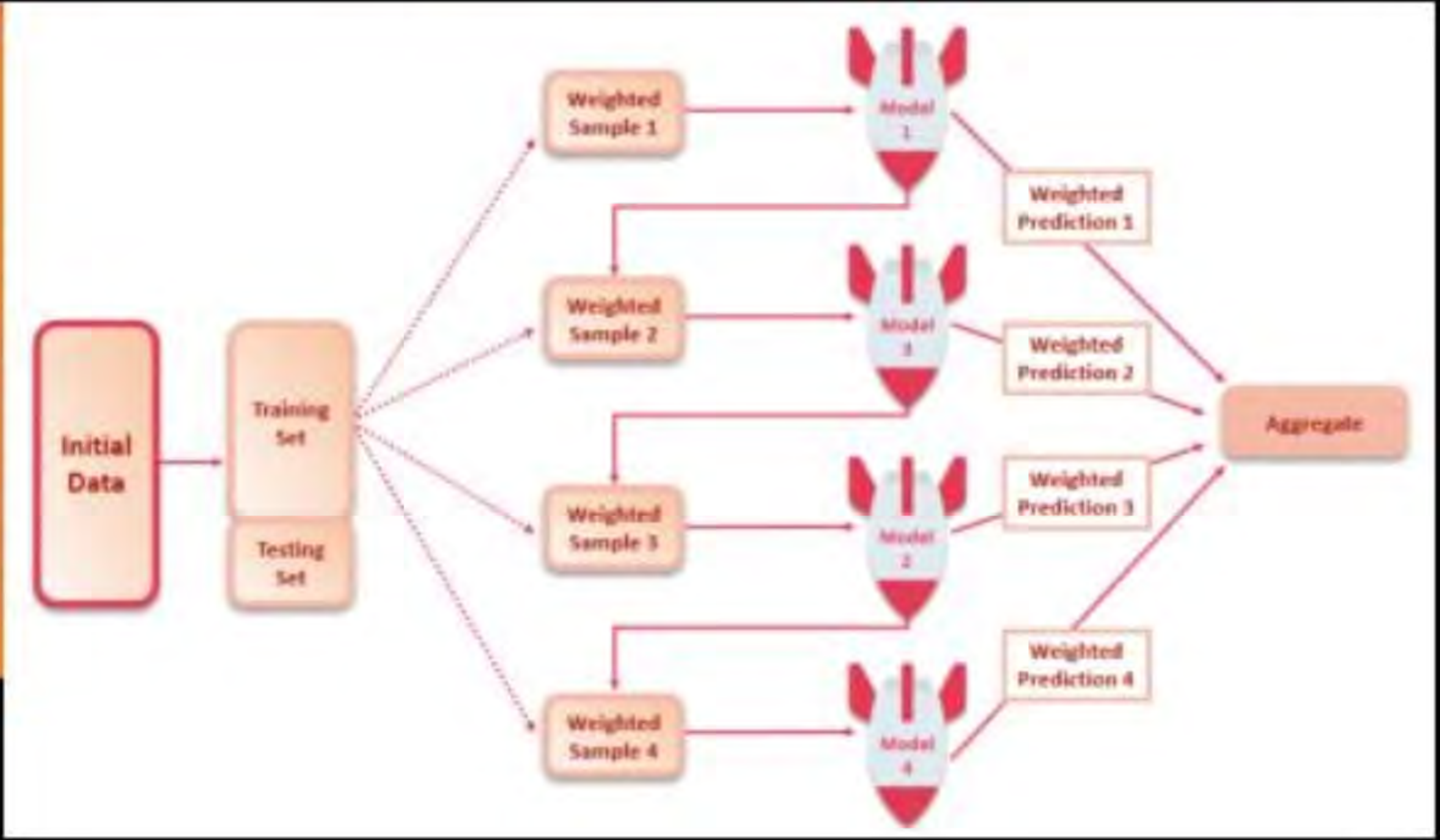
- wrapper
  - Forward selection
  - Backward elimination
  - Exhaustive feature search.
- filter
  - Correlation
- embedded
  - Regularisation & Random forest

Bias and Variance

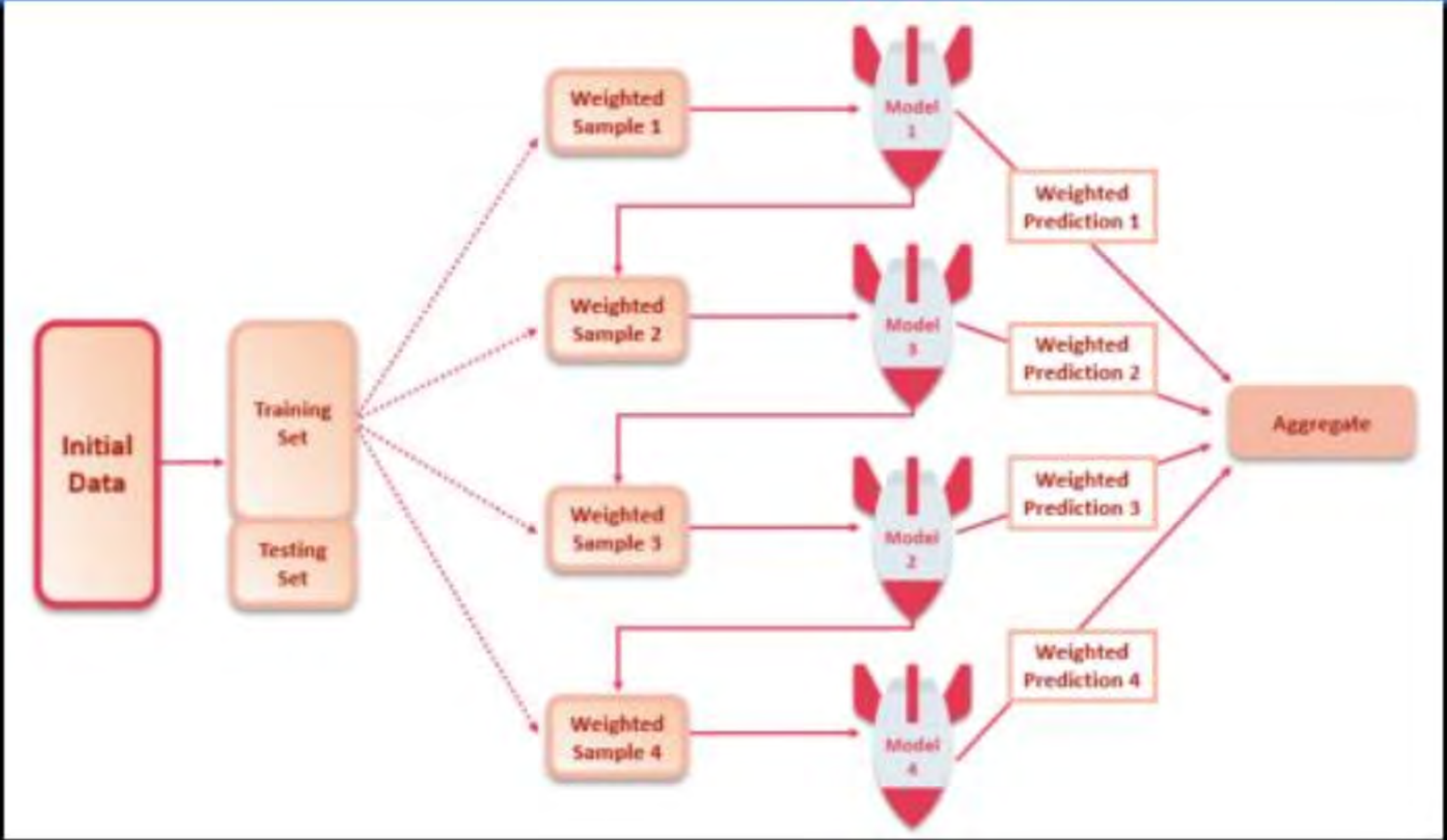
(Parallel Run) ← Bagging Boosting → Sequential

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**.

• ⇒ So $\underbrace{\left(\hat{y} = \hat{y}_1 w_1 + \hat{y}_2 w_2 + \hat{y}_3 w_3 + - - -\right)}_{\text{Both class/Reg} \Rightarrow}$

Algo 1 $w_1 = \cdot 1$ → ✓ Class 1

2 $w_2 = \cdot 4$ ✓ → Class 0

3 $w_3 = \cdot 2$ ✓ → Class 0

4 $w_4 = \cdot 7$ ✓ → Class 0

5 $w_5 = \cdot 3$ → Class 1 ✓

6 $w_6 = \cdot 4$ → Class 1 ✓

$$\hat{y} = \left(\hat{y}_1 w_1 + w_2 \hat{y}_2 + - - -\right)$$

$$\hat{y} = \text{Class1}(\cdot 8) + \text{Class0}\underbrace{(1\cdot 3)}$$

⇒ assign class with larger coeff

- ⇒ So $\underbrace{\left(\hat{y} = \hat{y}_1 w_1 + \hat{y}_2 w_2 + \hat{y}_3 w_3 + \cdots\right)}_{\text{Both class/Reg}}$ ⇒

Algo 1 $w_1 = .1$ → ✓ Class 1 ✓

2 $w_2 = .4$ ✓ → Class 0

3 $w_3 = .2$ ✓ → Class 1 ✓

4 $w_4 = .7$ ✓ → Class0

5 $w_5 = .3$ → Class 1 ✓

6 $w_6 = .4$ → Class 1 ✓

$$\hat{y} = \left(\hat{y}_1 w_1 + w_2 \hat{y}_2 + \cdots\right)$$

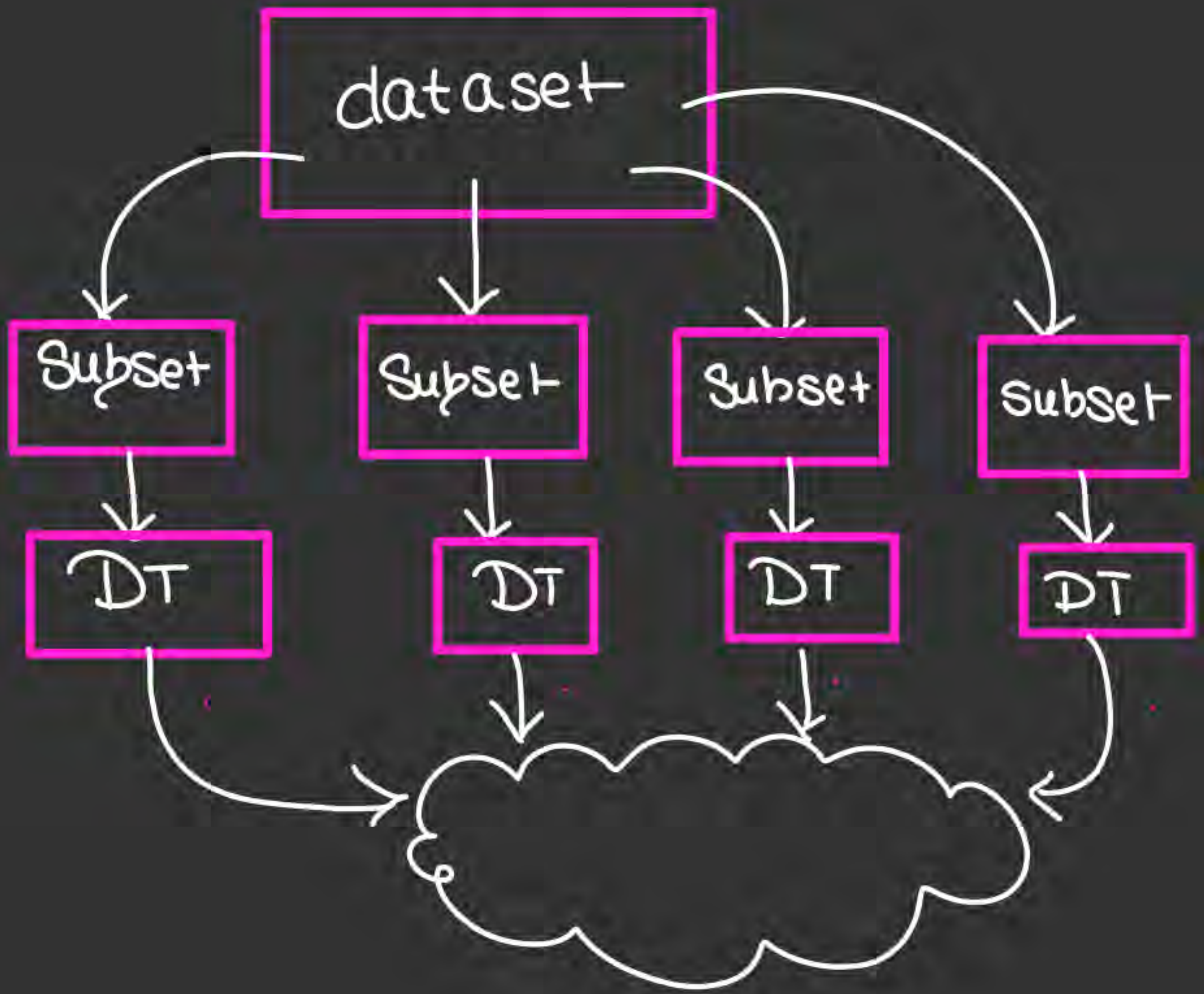
$$\hat{y} = 1.0 \text{ Class 1} + \underset{\downarrow}{(1.1)} \text{ Class0}$$

Class0 assign.

Ref only

⇒ Random forest ⇒ • if we have large dataset then, DT become computationally extensive and DT has property of overfitting

• So if we use ensemble learning technique in DT then we get Random forest.


dataset
Subset
Subset
Subset
Subset
DT
DT
DT
DT

Random Forest Algorithm

Power of Feature Selection .

Computationally extensive

Steps Involved in Random Forest Algorithm

- **Step 1:** In the Random forest model, a subset of data points and a subset of features is selected for constructing each decision tree. Simply put, n random records and m features are taken from the data set having k number of records.
- **Step 2:** Individual decision trees are constructed for each sample.
- **Step 3:** Each decision tree will generate an output.
- **Step 4:** Final output is considered based on *Majority Voting or Averaging* for Classification and regression, respectively.

Random Forest ⇒ Hyperparameters ⇒

1. No of subsets
2. No of dimension in subsets
3. Max depth of DT
4. Gini Impurity/IG/entropy.

Random Forest Algorithm

Important Hyperparameters in Random Forest

- Hyperparameters are used in random forests to either enhance the performance and predictive power of models or to make the model faster.
- **Number of decision trees to be constructed**
- **Maximum number of features a tree can use**
- **Splitting thresholds**

Random forest →

1) Reduce Risk of overfit

2) feature selection

3) low bias low Variance

4)

Random Forest Algorithm

no individual tree has whole data

Key Benefits

- □ Reduced risk of overfitting: when there's a robust number of decision trees in a random forest, the classifier won't overfit the model since the averaging of uncorrelated trees lowers the overall variance and prediction error.

✓

- □ Provides flexibility: Since random forest can handle both regression and classification tasks with a high degree of accuracy.
- □ Easy to determine feature importance: Random forest makes it easy to evaluate variable importance, or contribution, to the model. There are a few ways to evaluate feature importance. Gini importance and decrease in impurity are usually used to measure how much the model's accuracy decreases when a given variable is excluded.

Problems ⇒

1) Computationally extensive

2) more memory

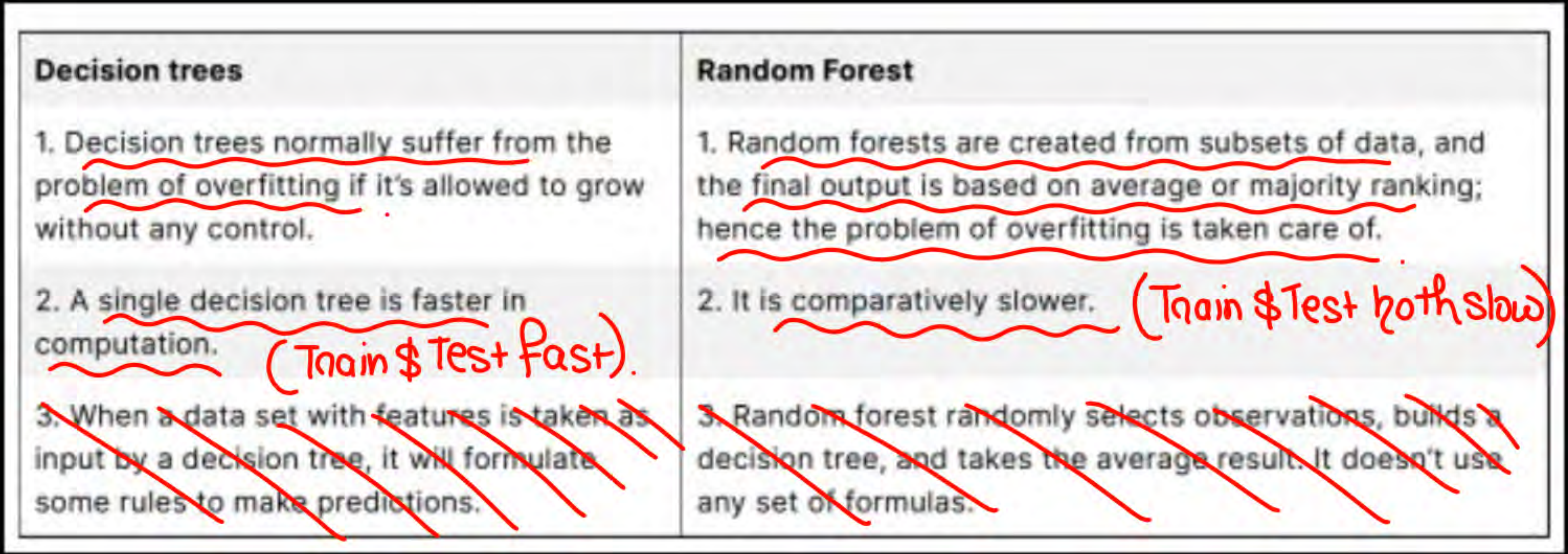
3) the subsets shd be good enough such that DT donot underfit

Random Forest Algorithm

Key Challenges

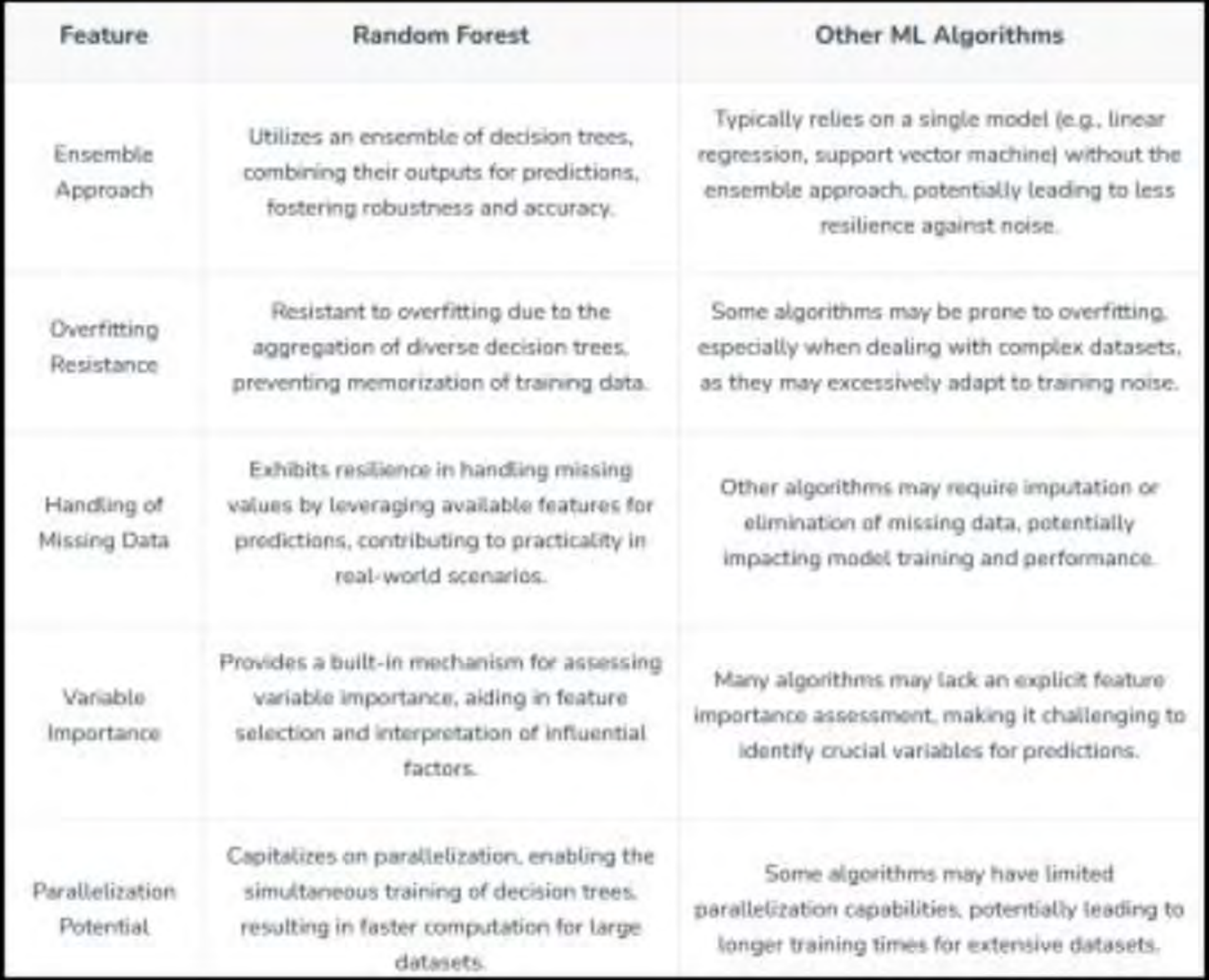
- ❑ Time-consuming process: Since random forest algorithms can handle large data sets, they can be provide more accurate predictions, but can be slow to process data as they are computing data for each individual decision tree.
- ❑ Requires more resources: Since random forests process larger data sets, they'll require more resources to store that data.
- ❑ More complex: The prediction of a single decision tree is easier to interpret when compared to a forest of them.

Random Forest Algorithm



Decision trees	Random Forest
1. Decision trees normally suffer from the problem of overfitting if it's allowed to grow without any control.	1. Random forests are created from subsets of data, and the final output is based on average or majority ranking; hence the problem of overfitting is taken care of.
2. A single decision tree is faster in computation. (Train & Test fast).	2. It is comparatively slower. (Train & Test both slow)
~~3. When a data set with features is taken as input by a decision tree, it will formulate some rules to make predictions.~~	~~3. Random forest randomly selects observations, builds a decision tree, and takes the average result. It doesn't use any set of formulas.~~

Random Forest Algorithm



Feature	Random Forest	Other ML Algorithms
Ensemble Approach	Utilizes an ensemble of decision trees, combining their outputs for predictions, fostering robustness and accuracy.	Typically relies on a single model (e.g., linear regression, support vector machine) without the ensemble approach, potentially leading to less resilience against noise.
Overfitting Resistance	Resistant to overfitting due to the aggregation of diverse decision trees, preventing memorization of training data.	Some algorithms may be prone to overfitting, especially when dealing with complex datasets, as they may excessively adapt to training noise.
Handling of Missing Data	Exhibits resilience in handling missing values by leveraging available features for predictions, contributing to practicality in real-world scenarios.	Other algorithms may require imputation or elimination of missing data, potentially impacting model training and performance.
Variable Importance	Provides a built-in mechanism for assessing variable importance, aiding in feature selection and interpretation of influential factors.	Many algorithms may lack an explicit feature importance assessment, making it challenging to identify crucial variables for predictions.
Parallelization Potential	Capitalizes on parallelization, enabling the simultaneous training of decision trees, resulting in faster computation for large datasets.	Some algorithms may have limited parallelization capabilities, potentially leading to longer training times for extensive datasets.

Maximum likelihood Estimation

What is MLE (lets see an example)

Maximum Likelihood Estimation (MLE) is a statistical method used to estimate the parameters of a probability distribution that best describe a given dataset. The fundamental idea behind MLE is to find the values of the parameters that maximize the likelihood of the observed data, assuming that the data are generated by the specified distribution.

Maximum likelihood Estimation

What is MLE (lets see an example)

- So whenever analysis is done we take a Small Sample from data

- Now our task is to use this Sample data to predict the Parameters of the Probability distribution of whole data

election data
100 Crore

Maximum likelihood Estimation

What is MLE (lets see an example)

- $\underbrace{\text{data distribution}}_{\text{data is a RV and data distribution} \Rightarrow \text{PDF of data}} \longrightarrow$ data का PDF

Gaussian

- $f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}$ $\longrightarrow (\sigma^2, \mu)$

- Exponential dist
  $f_X(x) = \lambda e^{-\lambda x} \longrightarrow (\lambda)$

Maximum likelihood Estimation

What is MLE (lets see an example)

ex ⇒ Let data has gaussian distribution

- If we have a sample of data ⇒ Containining points of value [Single D data] $(x_1, x_2, x_3, x_4, x_5 -----)$

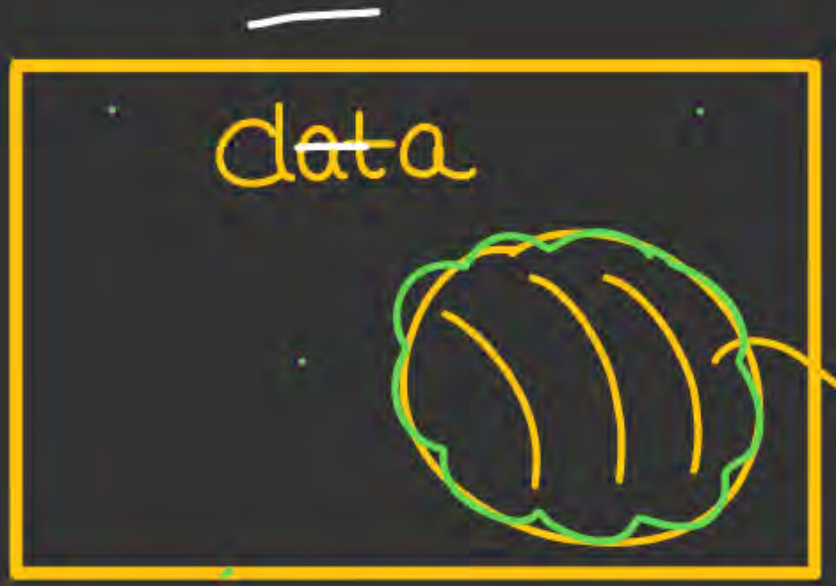
- These points in sample of data are independent from each other.

$$\underbrace{f_X(\underset{\uparrow}{x}) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}}_{}$$

Probability of getting these samples ⇒

$$P\left[x_1, x_2, x_3, x_4 - - - - - x_N / \mu, \sigma\right] \Rightarrow P(x_1)\,P(x_2)\,P(x_3)\,P(x_4) - - - P(x_N)$$

$$\left\{ \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x_1-\mu)^2/2\sigma^2} \cdot \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x_2-\mu)^2/2\sigma^2} - - - - - \right\}$$



Sample

$$\underbrace{P(x_1, x_2 --- x_N / \mu,\sigma) \checkmark = P(x_1/\mu,\sigma)\, P(x_2/\mu\sigma) ---}_{\text{This is the probability to get this sampleset}}$$

* So we always assume that this Sample represents whole data

* thus we want that the probability of getting this sample is maximium.

So* we assume that if we sample the data
the Probability of getting same sample values
is maximum

⇒ the parameters of the distribution of data shape
such that it maximises the probability of Sample
it maximises the likelihood of Sample.

So we have to maximise $\underbrace{P_{x_1} P_{x_2} P_{x_3} - - - P_{x_N}}$ $\Leftarrow$

Probab of Sample

$P(x_1, x_2 - - x_N / \mu, \sigma)$

Likelihood of Sample

$$\Rightarrow \left\{ \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right)^N e^{-(x_1-\mu)^2/2\sigma^2} \, e^{-(x_2-\mu)^2/2\sigma^2} - - - \, e^{-(x_N-\mu)^2/2\sigma^2} \right\}$$

data
Sample

assumption

* Sample represent whole data

* the Parameters of PDF of data Shape such that it max Probab of Sample.

So μ and σ^2 of the distribution of data shape

Such that the Probab of getting Sample is maximixed

Likelihood of Sample ⇒ $\mathcal{L} = \left\{ \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right)^N e^{-\sum_{i=1}^{N} (x_i-\mu)^2 / 2\sigma^2} \right\} \Bigg|_{max}$

$$(\log \mathcal{L})_{max} \Rightarrow \left\{ \log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^N + \log e^{-\sum_{i=1}^{N}(x_i-\mu)^2/2\sigma^2} \right\}$$

$$\Rightarrow \frac{\partial}{\partial u}\left\{ N\log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) - \sum_{i=1}^{N}(x_i-u)^2/2\sigma^2 \right\} = 0$$

$$-\cancel{2}\sum_{i=1}^{N}(x_i-\mu)/2\sigma^2 = 0$$

$$\sum_{i=1}^{N} x_i = N\mu$$

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i \checkmark$$

$$\frac{\partial}{\partial\sigma}\checkmark \left\{ N\log\left(\frac{1}{\sqrt{2\pi}\sigma}\right) - \sum_{i=1}^{N}\frac{(x_i-\mu)^2}{2\sigma^2} \right\} = 0$$

$$\frac{\partial}{\partial\sigma}\checkmark \left\{ -N\log\left(\sqrt{2\pi}\,\sigma\right) - \sum_{i=1}^{N}\frac{(x_i-\mu)^2}{2\sigma^2} \right\} = 0$$

$$\boxed{d\log x = \frac{1}{x}}$$

$$\left(\frac{-N\,\cancel{\sqrt{2\pi}}}{\cancel{\sqrt{2\pi}}\,\sigma}\right) - \frac{1}{\cancel{2}}\frac{\cancel{(-2)}}{\sigma^3}\sum_{i=1}^{N}(x_i-\mu)^2 = 0$$

;

$$\sum_{i=1}^{N}\frac{(x_i-\mu)^2}{\sigma^2} = N$$

$$\boxed{\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2}$$

Maximum likelihood Estimation

What is MLE (Logistic Regression)

Using Sample we need to find PDF of data

maximizing Probab / likelihood of Sample.

to predict parameter of the PDF of data (MLE)

Maximum likelihood Estimation

Probability Density Estimation & Maximum Likelihood Estimation

So what is Probability Density Estimation

- Probability Density: Assume a random variable x that has a probability distribution p(x). The relationship between the outcomes of a random variable and its probability is referred to as the probability density.

- The problem is that we don't always know the full probability distribution for a random variable. This is because we only use a small subset of observations to derive the outcome. This problem is referred to as Probability Density Estimation as we use only a random sample of observations to find the general density of the whole sample space.

Maximum likelihood Estimation

Probability Density Estimation & Maximum Likelihood Estimation

So what is Probability Density Estimation

- **Density Estimation:** It is the process of finding out the density of the whole population by examining a random sample of data from that population.

Maximum likelihood Estimation

Probability Density Estimation & Maximum Likelihood Estimation

Definition

- Maximum Likelihood Estimation

- our primary job is to analyse the data that we have been presented with.
- First thing would be to identify the distribution from which we have obtained our data.
- Next, we need to use our data to find the parameters of our distribution.
- Normal distributions, as we know, have mean (μ) & variance ($\sigma2$)
- Binomial distributions have the n and p.
- Exponential distributions have the inverse mean (λ).

Maximum likelihood Estimation

What is Maximum Likelihood Estimation (MLE)

Definition

- The goal of the maximum likelihood principle is to fit an optimal statistical distribution to some data.
- This makes the data easier to work with, makes it more general, allows us to see if new data follows the same distribution as the previous data, and lastly, it allows us to classify unlabelled data points.
- MLE stands for Maximum Likelihood Estimation. It is a method for estimating the parameters of a statistical model, given a dataset. The goal of MLE is to find the parameter values that maximize the likelihood of the data, given the model. This is done by choosing the values of the parameters that make the observed data most probable.

Maximum likelihood Estimation

What is Maximum Likelihood Estimation (MLE)

- we want to do now is obtain the parameter set θ that maximises the joint density function of the data vector; the so-called Likelihood function L(θ).
- This likelihood function can also be expressed as P(X|θ), which can be read as the conditional probability of X given the parameter set θ.

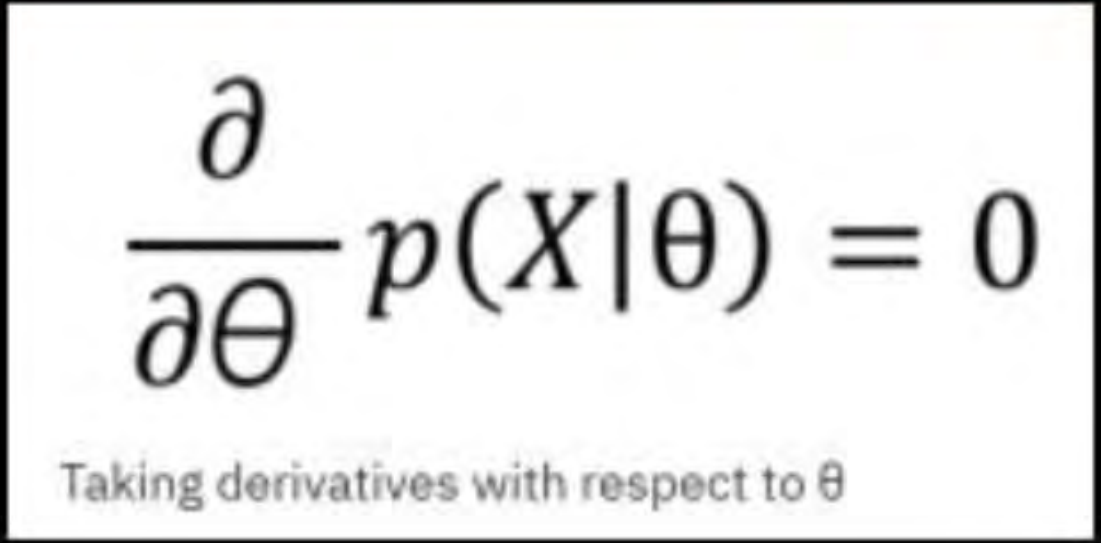
$$L(\theta) = p(X \mid \theta) = p(X(1), X(2), \ldots. X(n) \mid \theta)$$

X is the data matrix, and X(1) up to X(n) are each of the data points, and θ is the given parameter set for the distribution.

Maximum likelihood Estimation

What is Maximum Likelihood Estimation (MLE)

- To obtain this optimal parameter set, we take derivatives with respect to θ in the likelihood function and search for the maximum: this maximum represents the values of the parameters that make observing the available data as likely as possible.



$$\frac{\partial}{\partial \theta} p(X|\theta) = 0$$

Taking derivatives with respect to θ

Maximum likelihood Estimation

What is Maximum Likelihood Estimation (MLE)

- if the data points of X are independent of each other, the likelihood function can be expressed as the product of the individual probabilities of each data point given the parameter set:

$$L(\theta) = p(X \mid \theta) = \prod p(X(j) \mid \theta)$$

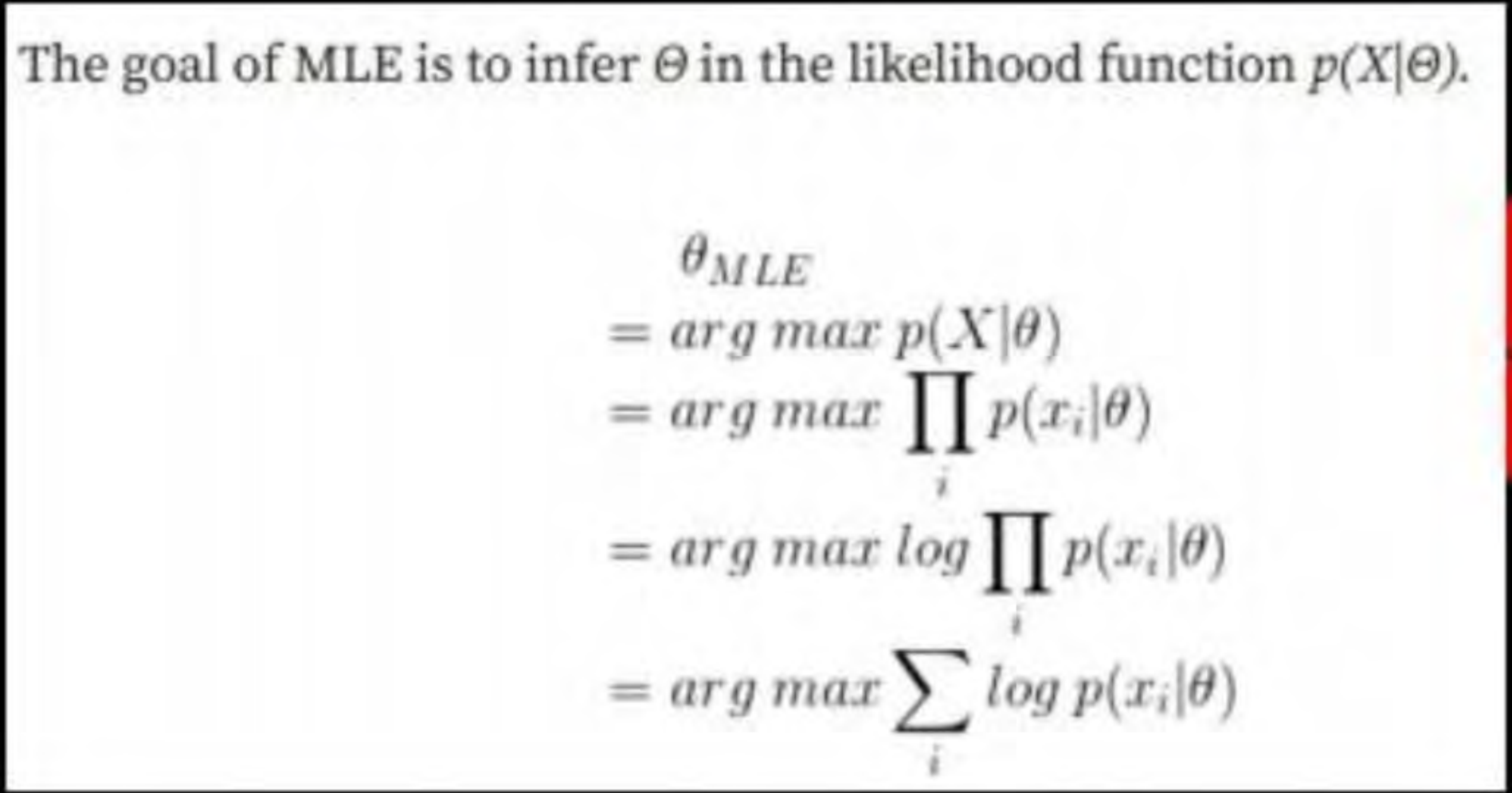
Taking the derivatives with respect to this equation for each parameter (mean, variance, etc...) keeping the others constant, gives us the **relationship between the value of the data points, the number of data points, and each parameter.**

What is Maximum Likelihood Estimation (MLE)

From the likelihood function we take log likelihood function

Maximum likelihood Estimation

What is Maximum Likelihood Estimation (MLE)



The goal of MLE is to infer Θ in the likelihood function $p(X|\Theta)$.

$$
\begin{aligned}
&\theta_{MLE}\\
&= arg\,max\,p(X|\theta)\\
&= arg\,max\,\prod_i p(x_i|\theta)\\
&= arg\,max\,log\prod_i p(x_i|\theta)\\
&= arg\,max\sum_i log\,p(x_i|\theta)
\end{aligned}
$$

Lets see some examples of MLE

Maximum Aposterori Probability Rule

What is Maximum Aposteriori Probability Rule(MAP)

Here we maximize ...

- MAP stands for Maximum A Posteriori probability. It is a method for estimating the parameters of a statistical model, given a dataset and some prior knowledge about the model. The goal of MAP is to find the parameter values that maximize the posterior probability of the data, given the model and the prior knowledge. This is done by choosing the values of the parameters that make the observed data most probable, given the prior knowledge.

Maximum Aposterori Probability Rule

What is Maximum Aposteriori Probability Rule(MAP)

Here we maximize ...

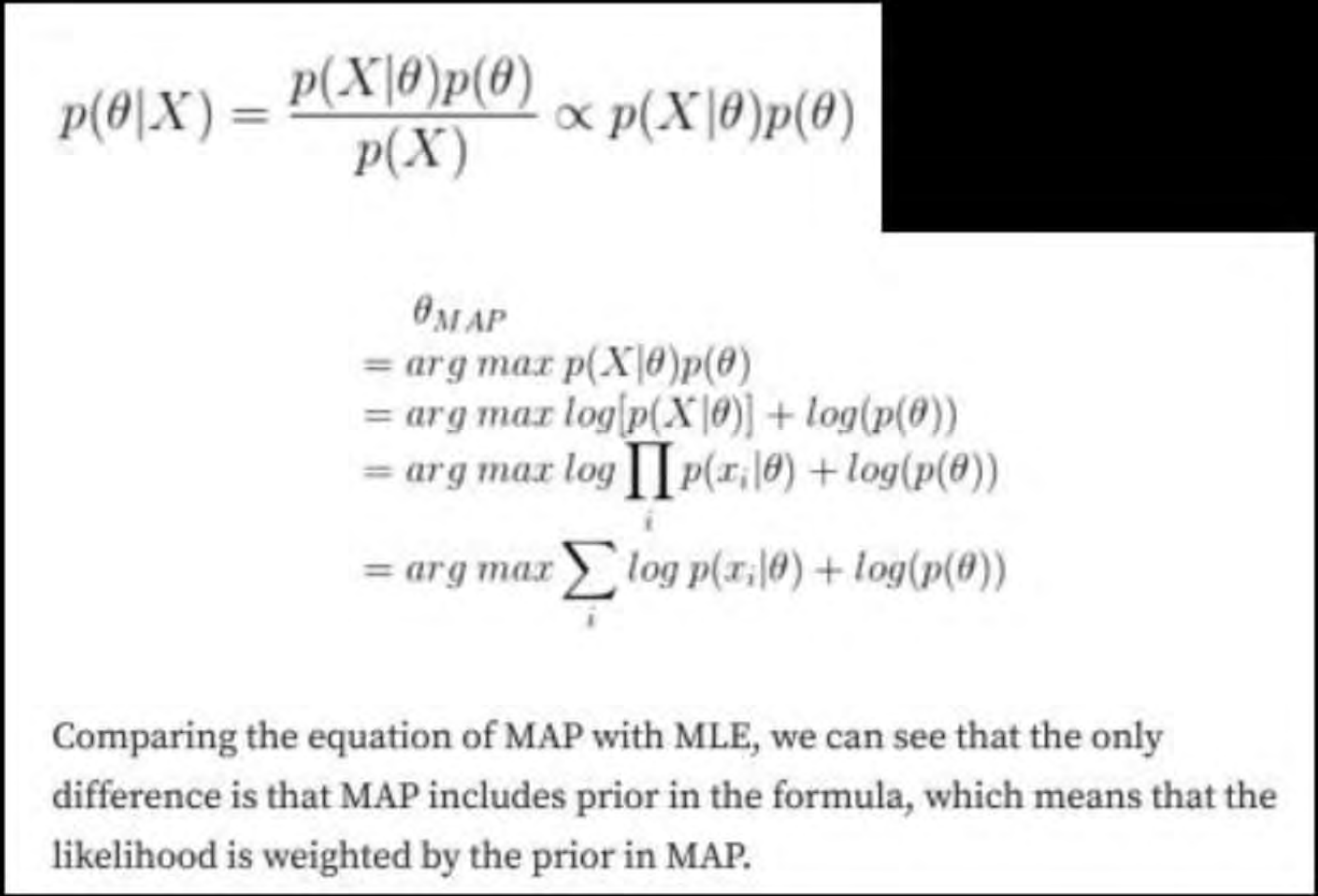
- The goal of MAP is to find the parameter values that maximize the posterior probability of the data, given the model and the prior knowledge.
- MAP is similar to MLE (Maximum Likelihood Estimation), but it incorporates prior knowledge about the model into the estimation process. This can be useful in cases where the data is limited or noisy, or where there is a need to incorporate domain-specific knowledge into the model.

Maximum Aposterori Probability Rule

What is Maximum Aposteriori Probability Rule(MAP)

Here we maximize ...

- MAP is similar to MLE (Maximum Likelihood Estimation), but it incorporates prior knowledge about the model into the estimation process. This can be useful in cases where the data is limited or noisy, or where there is a need to incorporate domain-specific knowledge into the model.



$$p(\theta|X) = \frac{p(X|\theta)p(\theta)}{p(X)} \propto p(X|\theta)p(\theta)$$

$$\begin{aligned}&\theta_{MAP}\\ &= arg\ max\ p(X|\theta)p(\theta)\\ &= arg\ max\ log[p(X|\theta)] + log(p(\theta))\\ &= arg\ max\ log\prod_i p(x_i|\theta) + log(p(\theta))\\ &= arg\ max\sum_i log\ p(x_i|\theta) + log(p(\theta))\end{aligned}$$

Comparing the equation of MAP with MLE, we can see that the only difference is that MAP includes prior in the formula, which means that the likelihood is weighted by the prior in MAP.

THANK - YOU