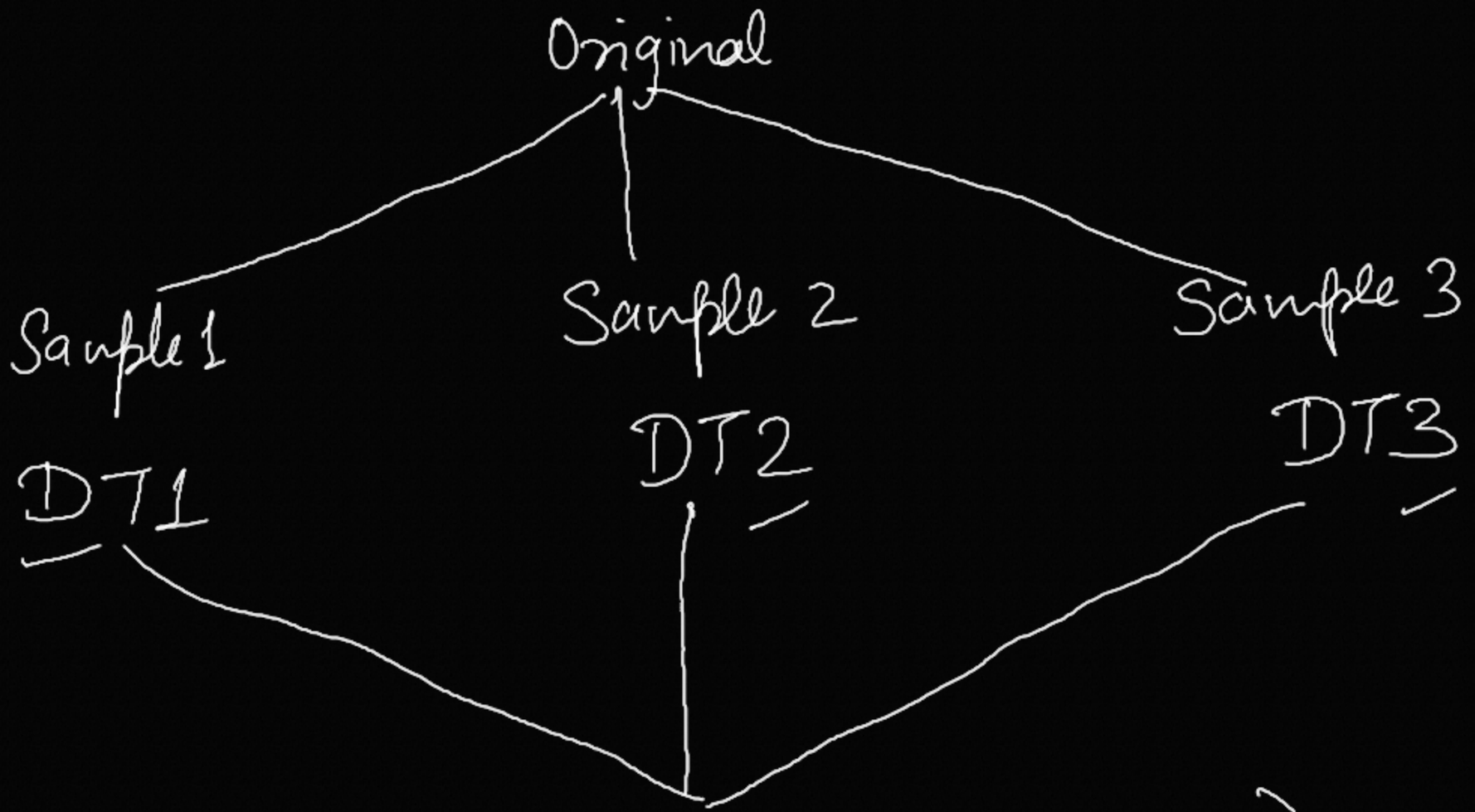
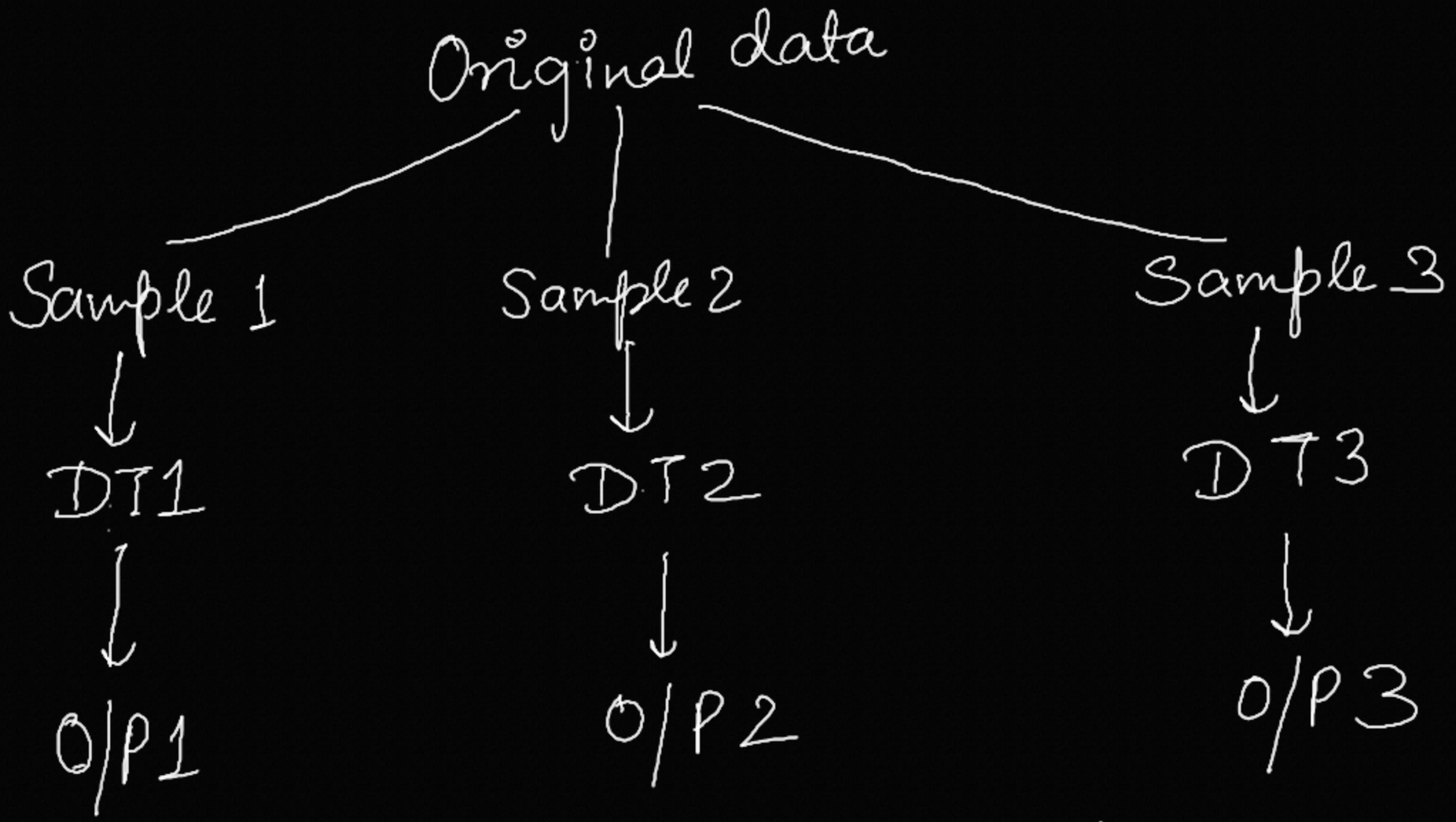
Random Forest

* Regression and Classification
* Introduced → Leo Breiman & Adele Cutler 2001.

* No assumptions
* OOB error estimation → Out of Bag (Internal validation data ($\frac{1}{3}$rd of data))


Original
Sample 1
Sample 2
Sample 3
DT1
DT2
DT3
RF (accuracy)

General



Reg → Mean , Classification → Voting

Mathematical

feature 1	feature 2	Target
1	2	0
2	3	1 0
3	4	1
4	5	1
5	6	1

① Create Samples / Subset

Subset 1

f1	f2	T
1	2	0
2	3	0
3	4	1
4	5	1

Subset 2

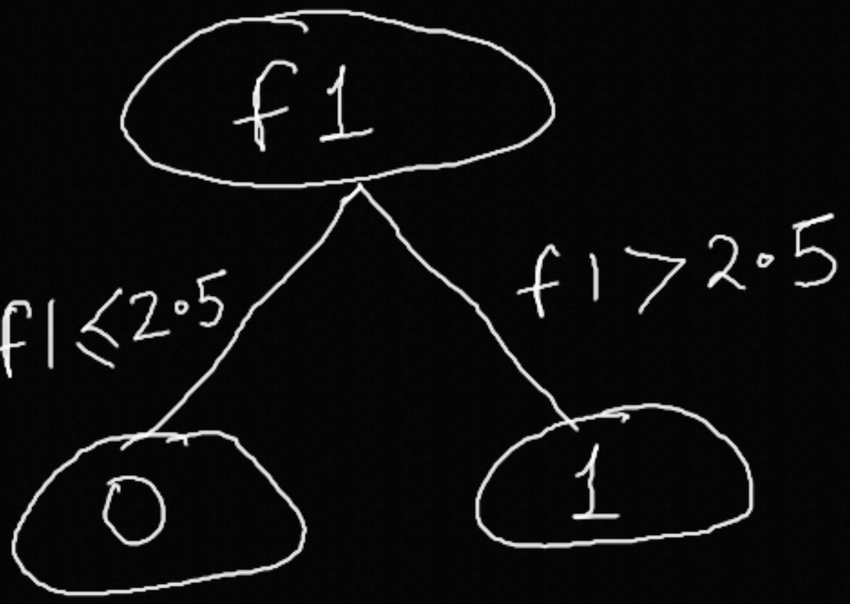
f1	f2	T
2	3	0
3	4	1
4	5	1
5	6	1

Subset 3

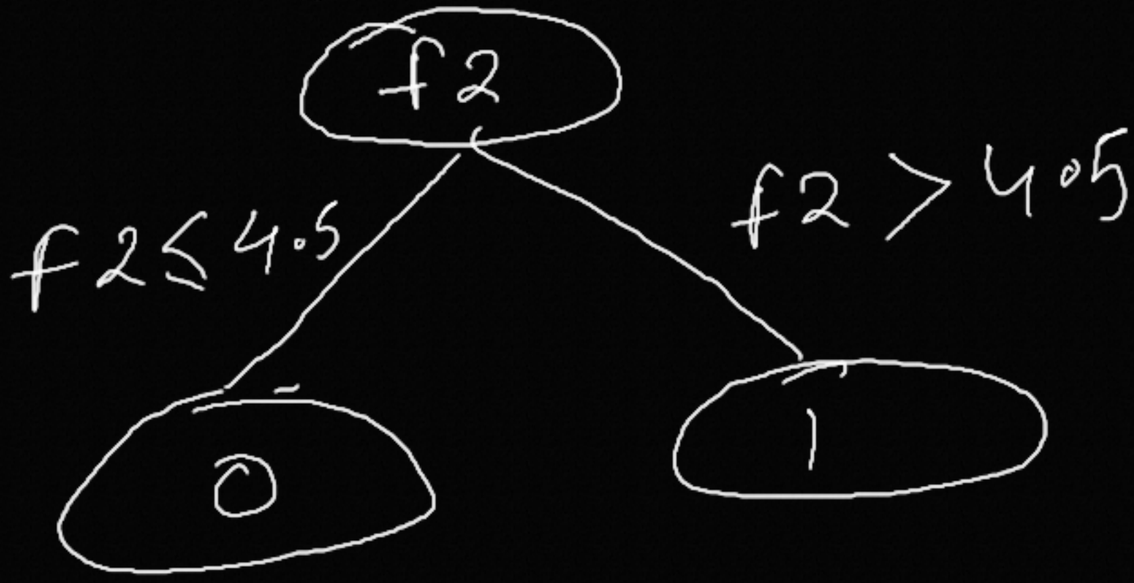
f1	f2	T
1	2	0
3	4	1
4	5	1
5	6	1

② Create DT for each sample

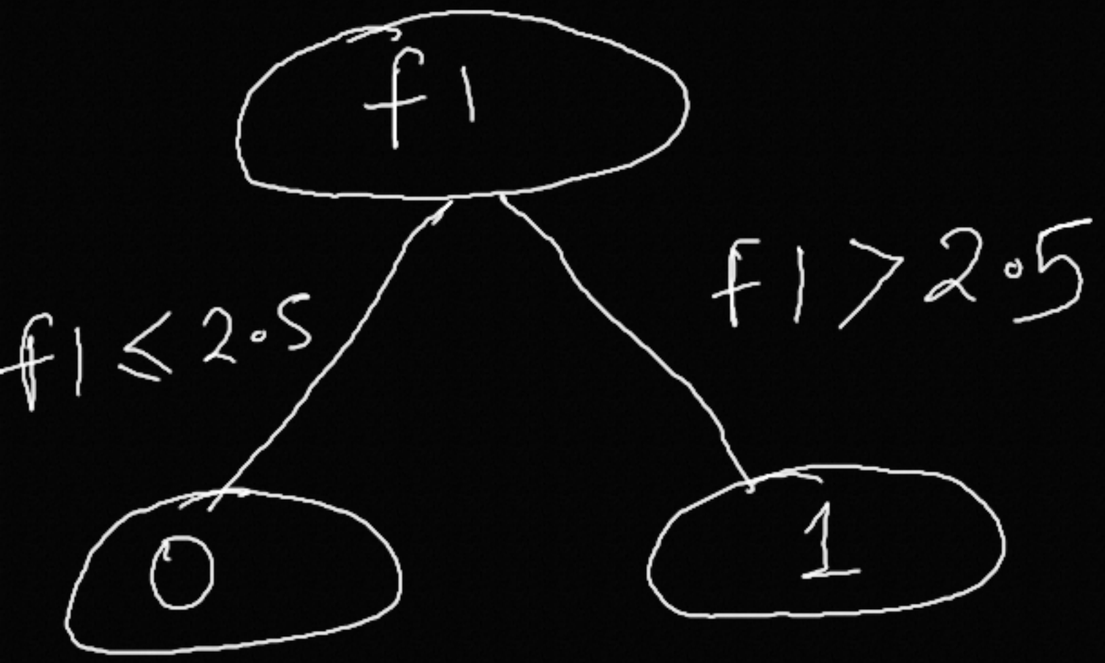
(a) DT1



(b) DT2



(c) DT3.



Model Training Completed

* Prediction

f1 = 3 , f2 = 4

Apply on each DT

DT 1 → 1

DT 2 → 0

DT 3 → 1

Voting → final = 1
o/p

When to Use

* Complex & Non-linear data
* Very large dataset
* They can perform good even we have missing values
* Prevent overfitting

When Not to Use

* High-D data with less rows / data points
* When majority features are categorical
* Highly computational intensive
* Large-scale problems

fact : Rf are also called "black-box model"