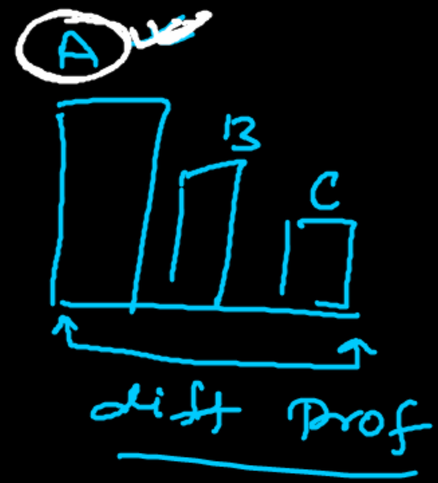
Ensemble Model

Consolidate gathering

Group of "musicians".



In ML, multiple models are brought together to build a powerful model.

$\{M_1 \quad M_2 \quad M_3 \quad \text{———} \quad M_K\}$ — baseline model

→ The multiple models may individual perform poorly but when you combined them, they become more powerful,

→ This combination to improve performance of several baseline models is called "ensemble learning".

Types of ensemble learning

- Voting
  - DT = 60%
  - Log = 51%
  - SVM = 55%
  - → 70%
  - Min - 51%
- Bagging
  - Random Forest
- Boosting
  - Ada boost
  - Gradient boost
  - XGBoost
- Stacking
  - DT
  - SVM
  - Log
  - KNN

→ High performing model

→ very powerful & elegant

→ Kaggle / Hackathon competition — most winners

→ Job / onboarding (interns / e-com) — very useful in real

Black box

Key aspect :-

↓

Generic

↓

expert

{ M_1 m_2 m_3 ——— m_k }

Real world case

↳ more diff. these models are, the better you can combine them.

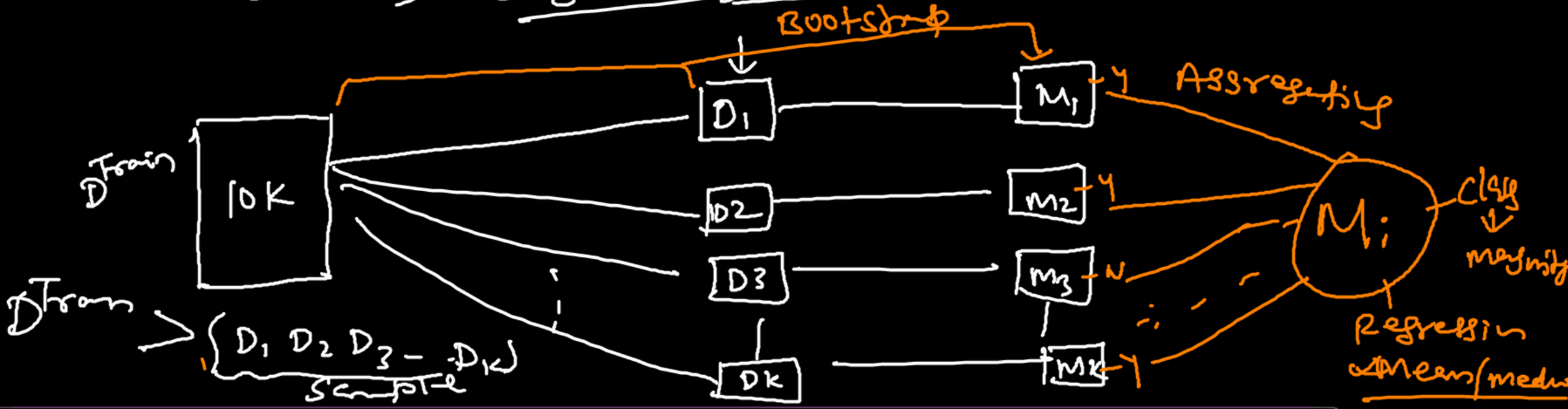
Bagging ⟷ Bootstrap Aggregating

↓

Short form

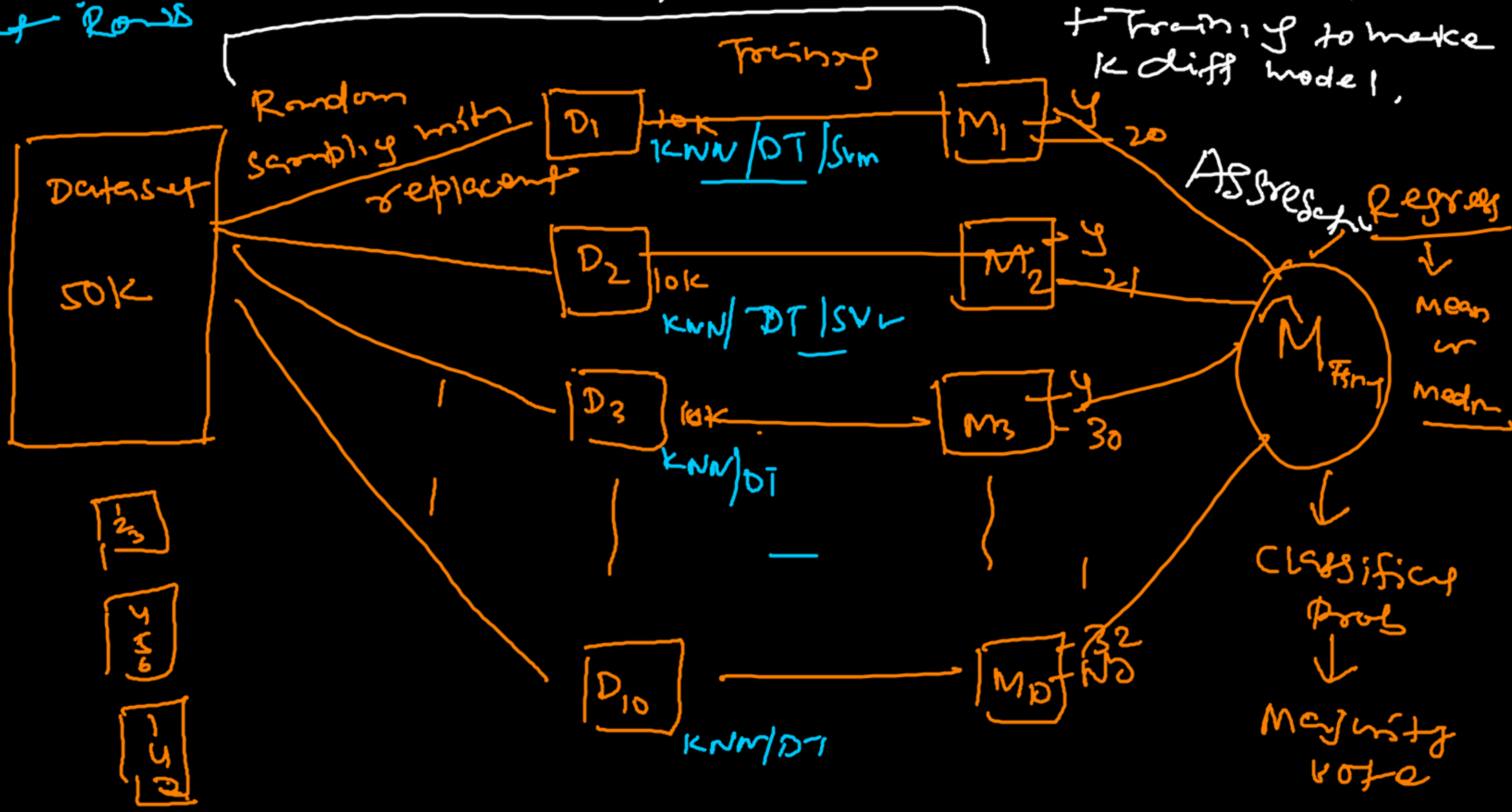
Bagging - Steps — Bootstrap Aggregation

↳ English Sample



Random Sampling with replacement Rows

Bootstrap = random sampling with replacement + Training to make k diff model,



Bagging model — handle high variance Problem

↳ overfitting Problem

* Bias - variance Trade Off - Interview Question

Bias $\propto \frac{1}{\text{Variance}}$

LB LV → Objective

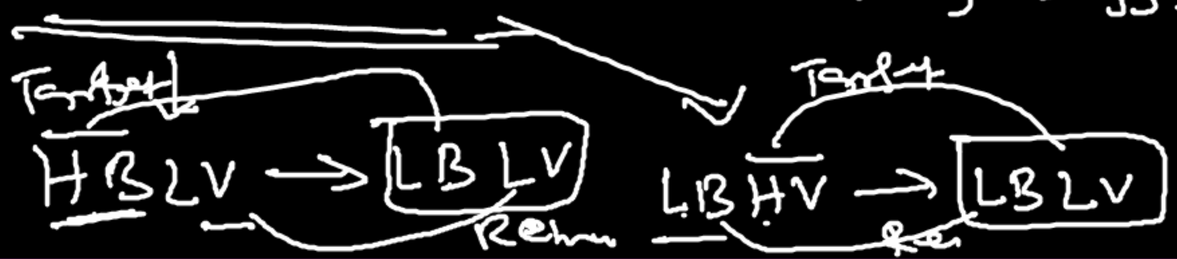
DT

LB HV

Linear

HB LV

Random Forest is a part of Bagging



Random Forest & their construction

↓

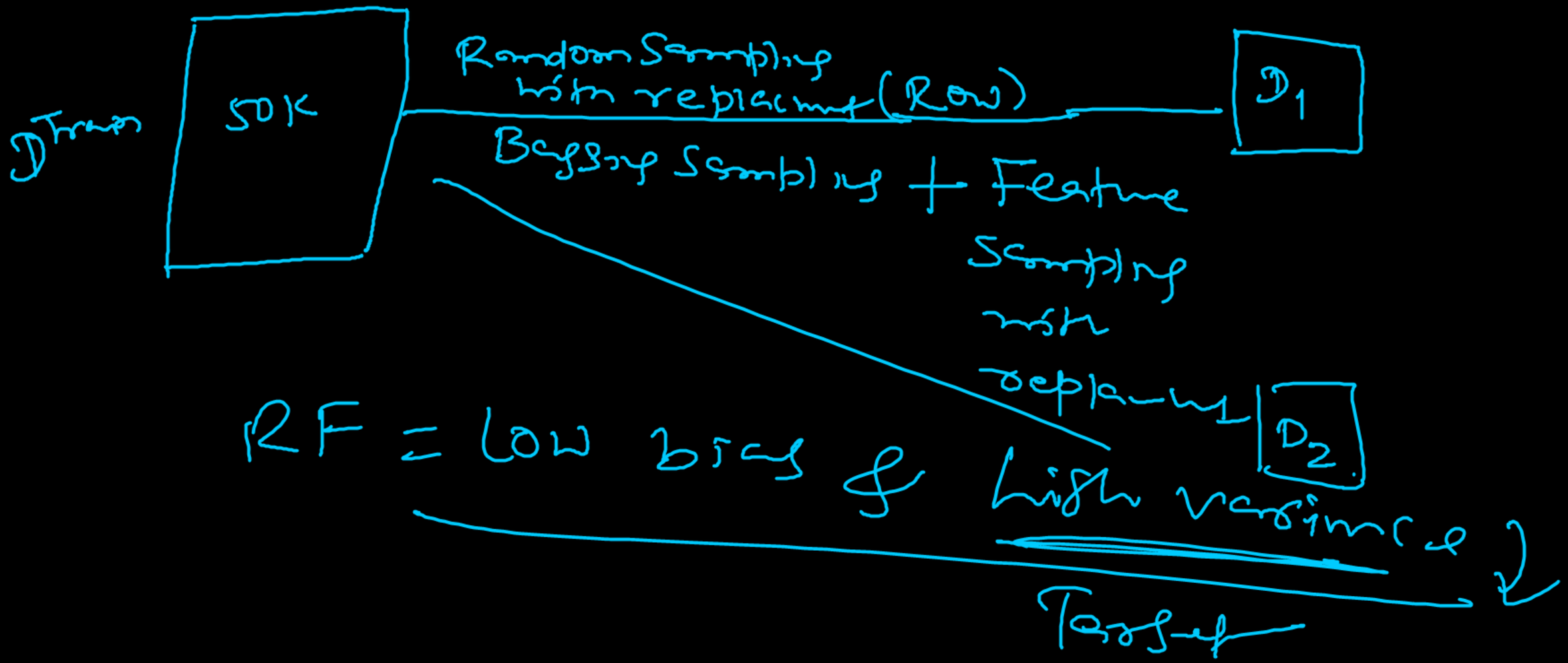
"Wisdom of the crowd" → different background

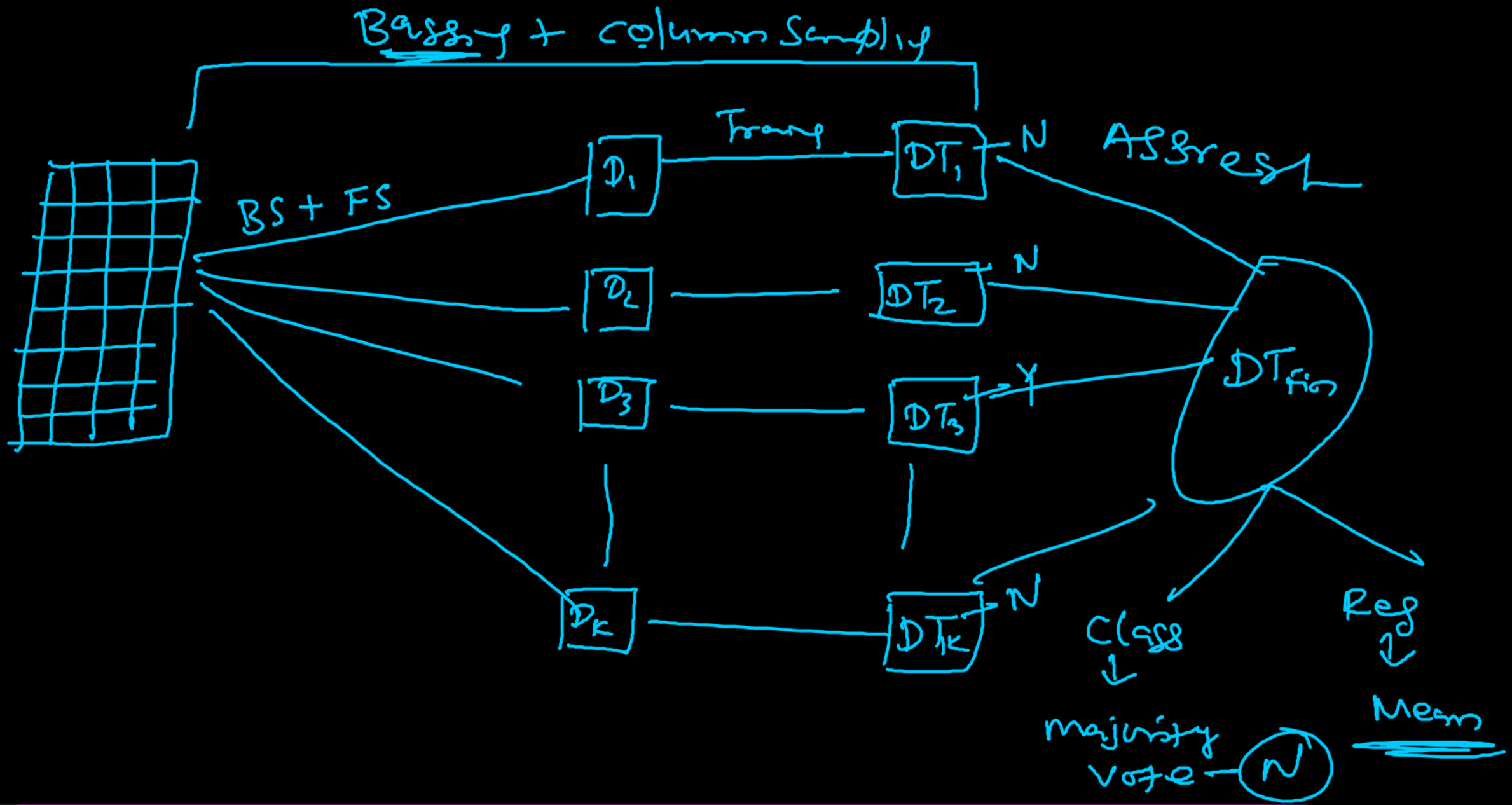
Random Forest - is a part of Bagging

↳ only creates multiple DT

RF :- DT + Bagging (Bootstrap + Aggregating) + Column Sampling

Base





Bagging — Random Forest → OOB — (Out of Bag)

RF :— DT → base learner

+ row Sampling with replacement

+ column Sampling with replacement

+ Aggregation — Majority → Classifier

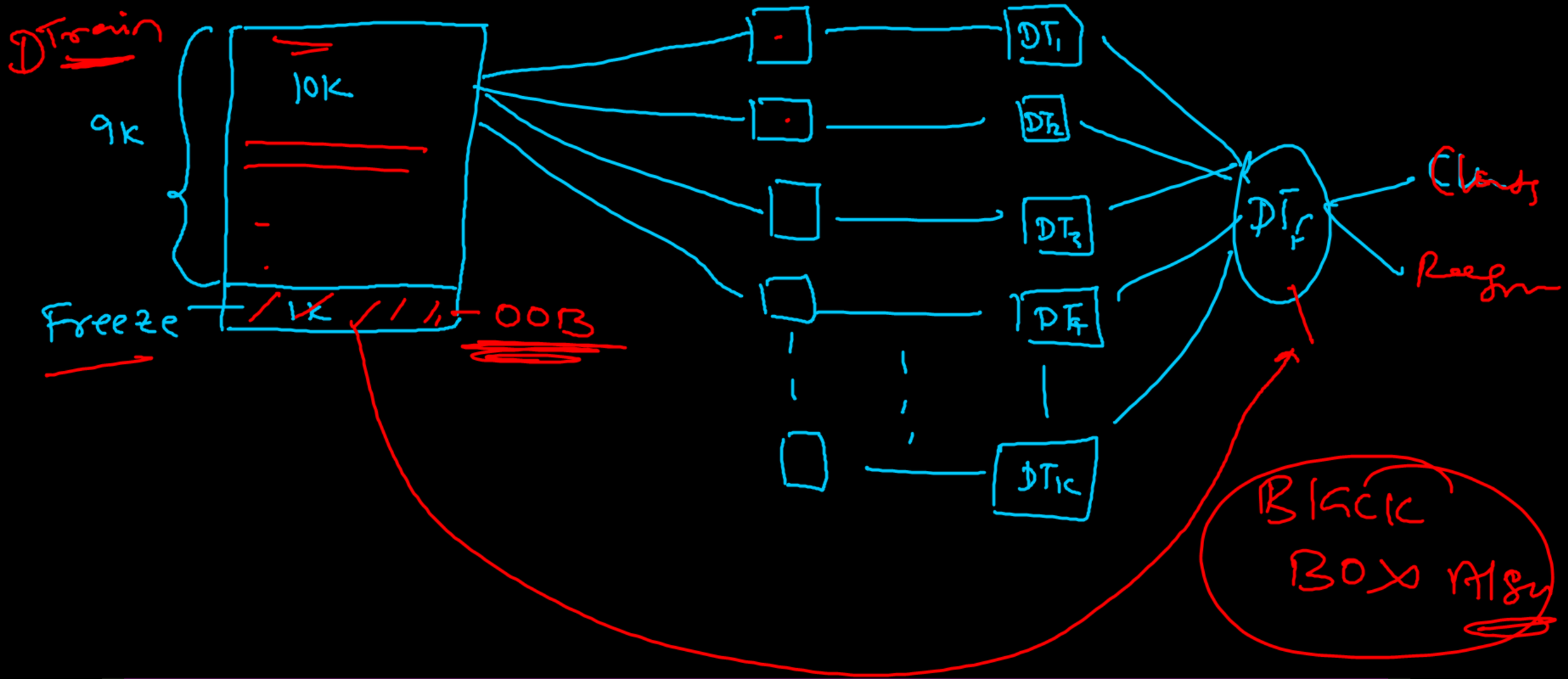
↓

Mean / Median

Regressor

Train & Test

OOB - Hyperparameters



Boosting — high Bias Problem (underfitting Problem)

- Ada Boost
- Gradient Boost → Objective
- XGBoost

SOTA model - State of the art model in ML

HB LV

Remain same

LB LV → objective

Bagging → high variance & low bias

- Random Forest

H V L B

Testset

remain same

LV LB

Ada DT = 90%.
RF = 88%.