C4 - Ensemble learning
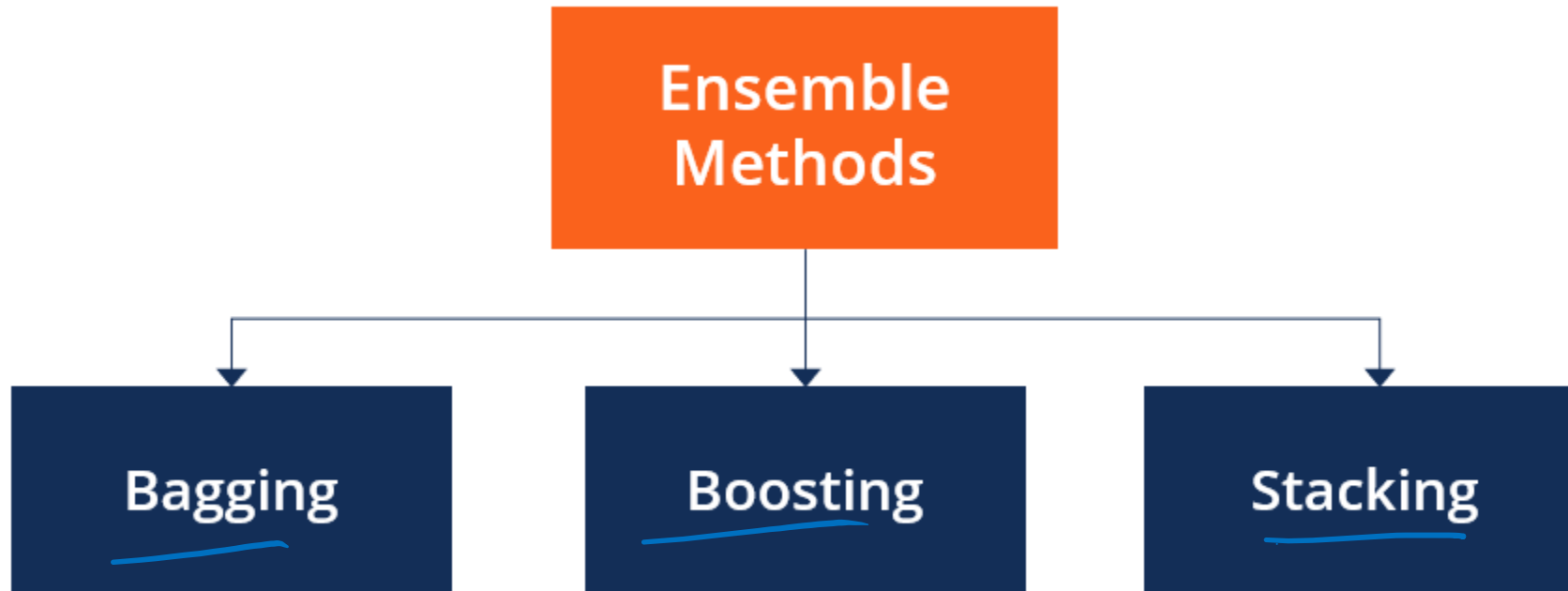
Ensemble Learning: visual generalization

Final ensemble prediction

Meta-model

Predictions

P₁ P₂ P₃ P₄

Decision Tree Logistic Regression KNN Other …

Training

Training set

weak learners

1

Voting Classifier

Predictions

1 0 1

Decision Tree Logistic Regression KNN

Province

New data point
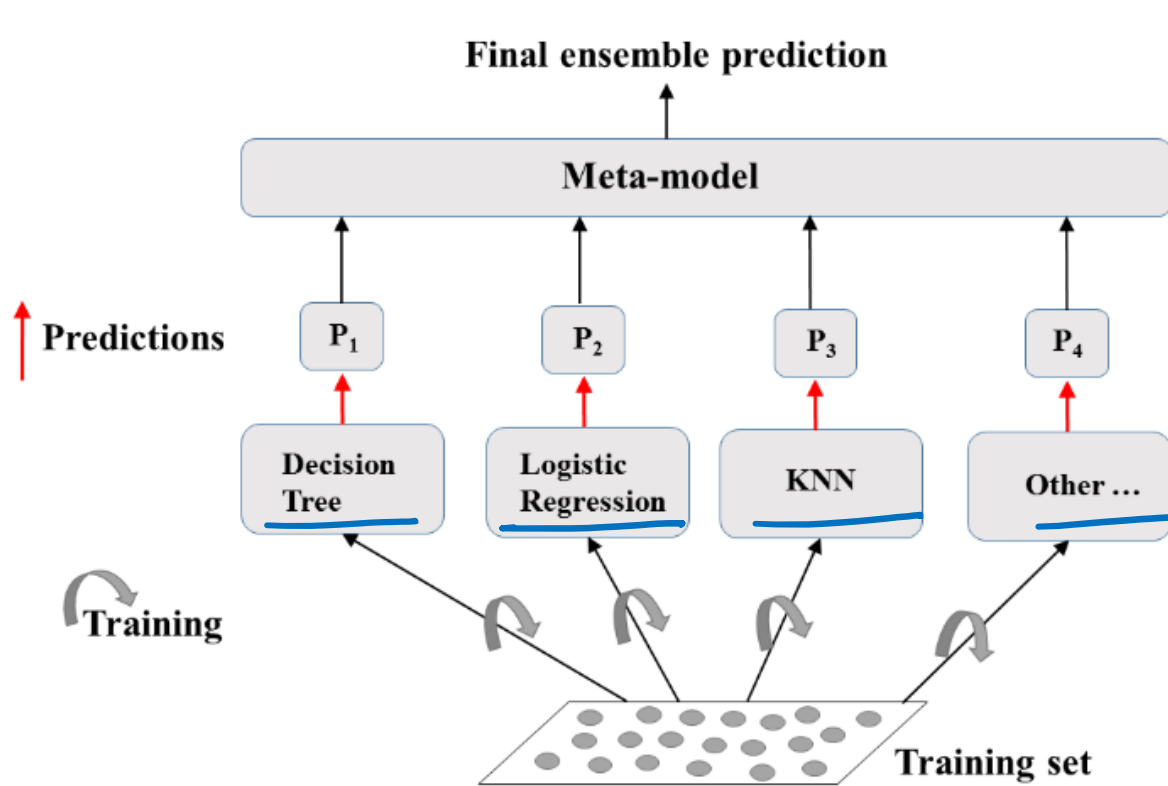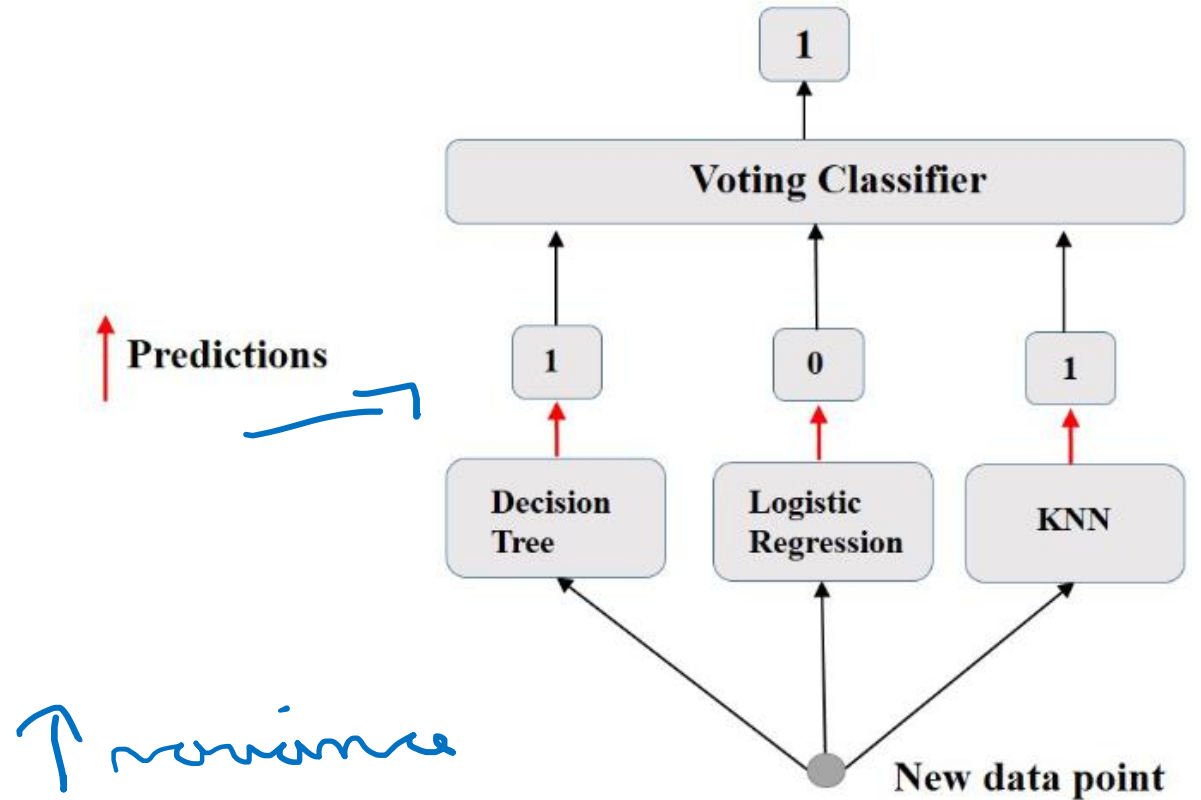
Bagging → bootstrap + aggregating

initial dataset

L bootstrap samples

weak learners fitted on
each bootstrap sample
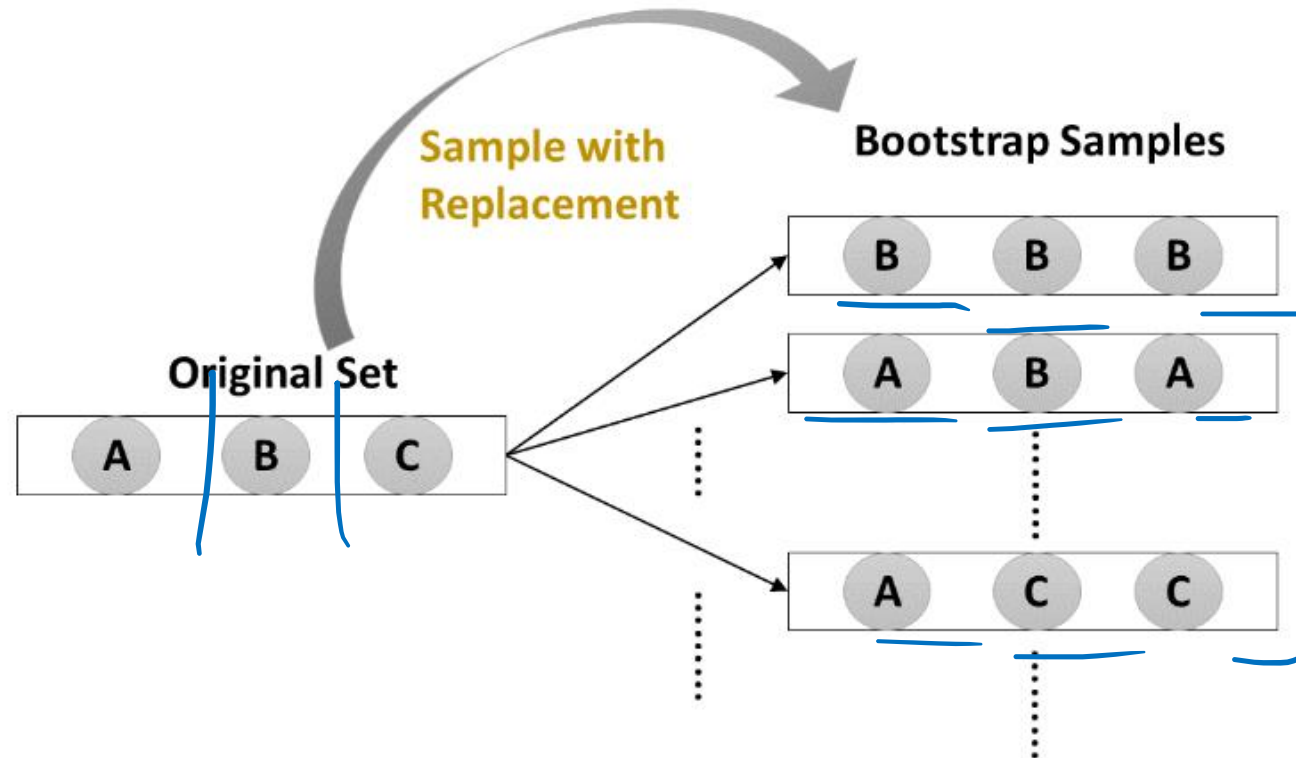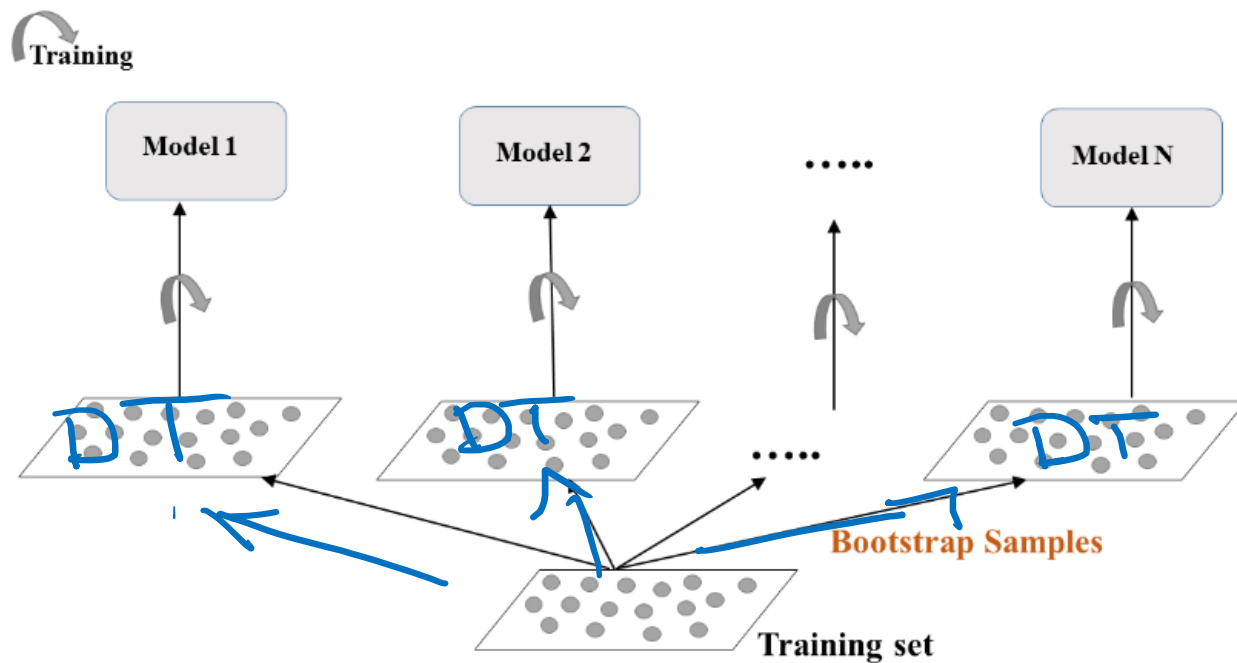
ensemble model (kind of average
of the weak learners)
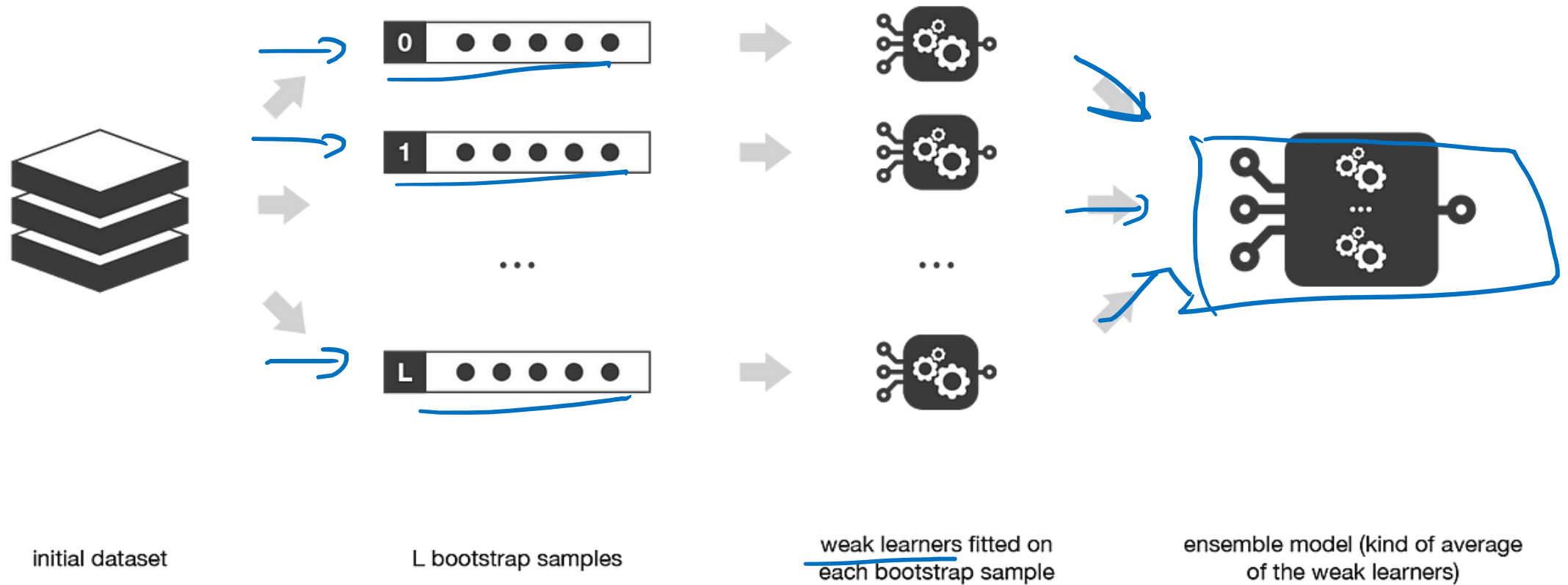
Bagging

One algorithm
≠ subsets of training set
Reduces variance of individual models in the ensemble

Sample with Replacement

Bootstrap Samples

Original Set

| A | B | C |

| B | B | B |

| A | B | A |

| A | C | C |

Bagging

Moda/Avg

Training

| Model 1 | Model 2 | ..... | Model N |

DT   DT   DT

.....

Bootstrap Samples

Training set

Final Prediction

Bagging

| Model 1 | ..... | Model N |

New instance
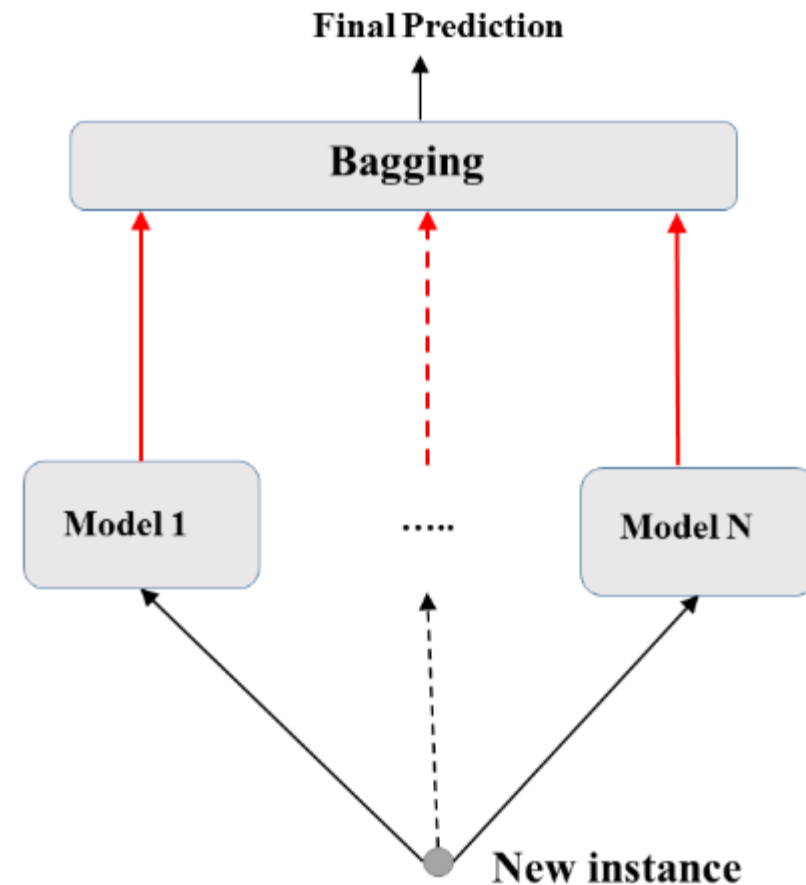
OOB (Out of Bagging Evaluation)

$$OOB \text{ score} = \frac{OOB_1 + \ldots + OOB_N}{N}$$

A B C.D

OOB$_1$ • • • • •   OOB$_N$

Model 1   • • • • •   Model N

Train   Evaluate

C D B B

A B A   OOB samples   • • • • •   • • • • •   A

Bootstrap Samples
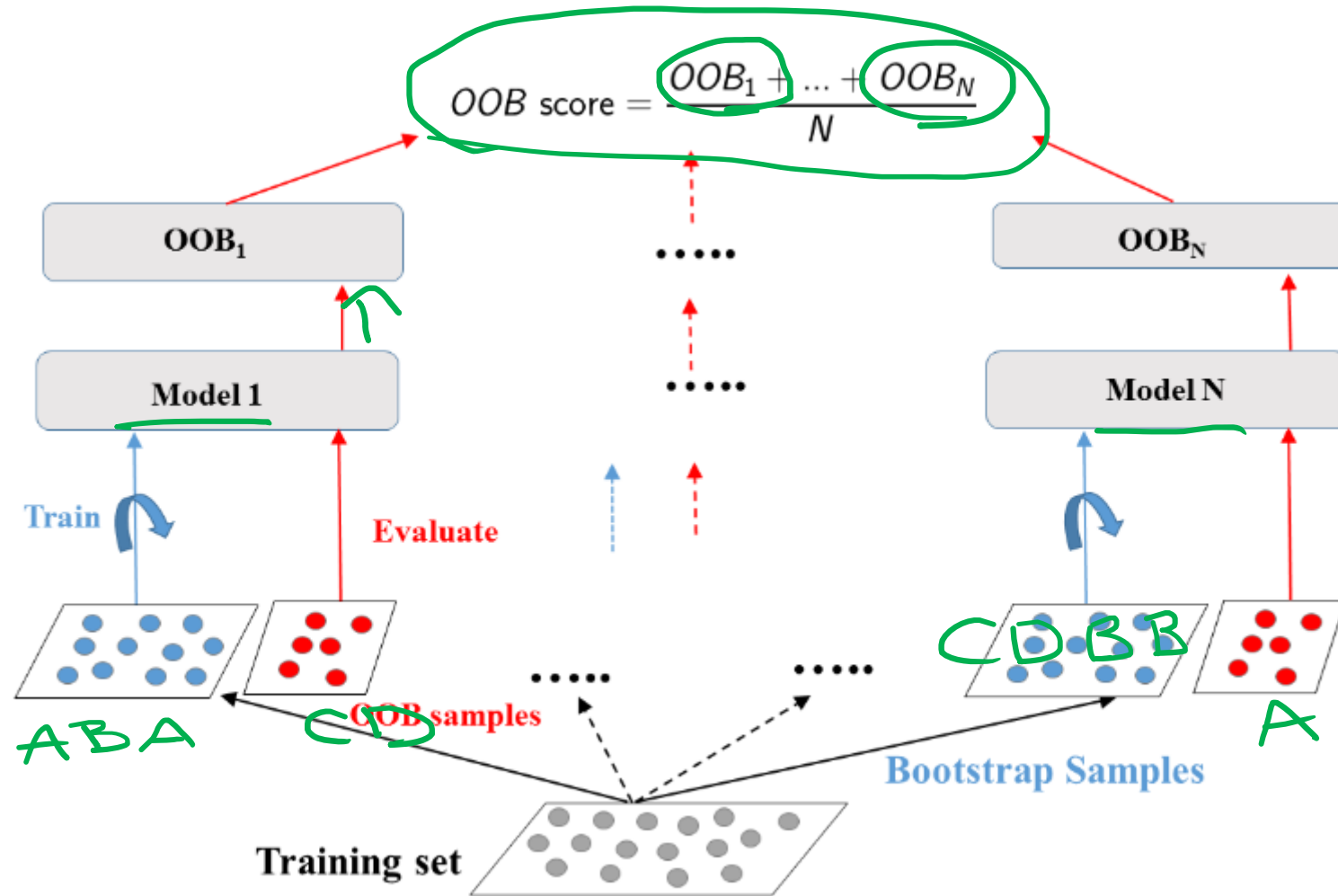
Training set

Random Forest

Sample d features at each split without replacement

Decision Tree 1

Decision Tree K

f_1 f_P f_2

f_1 f_P f_2

f_1 f_2 f_3

Training

Bootstrap Samples
(size= # of samples)
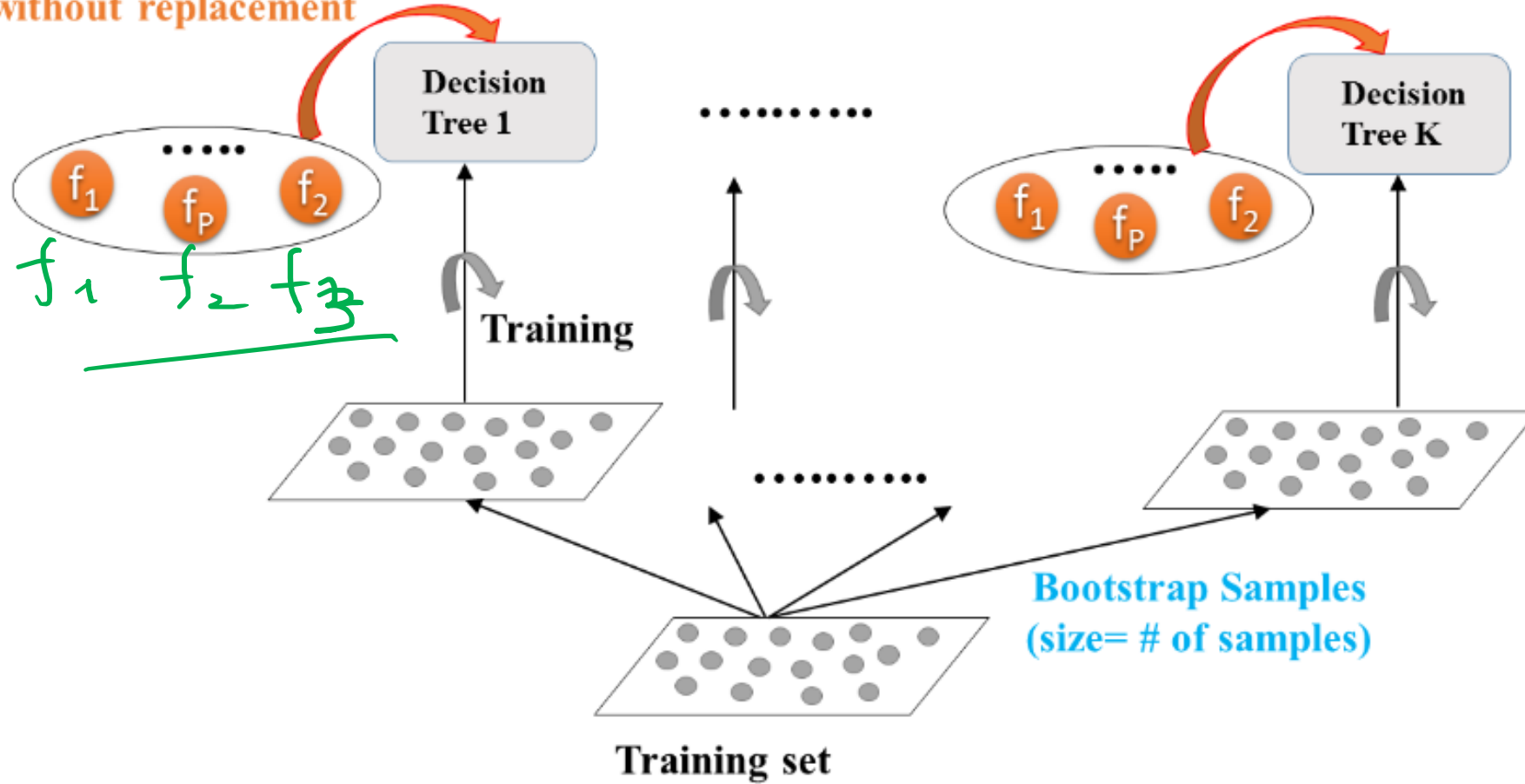
Training set

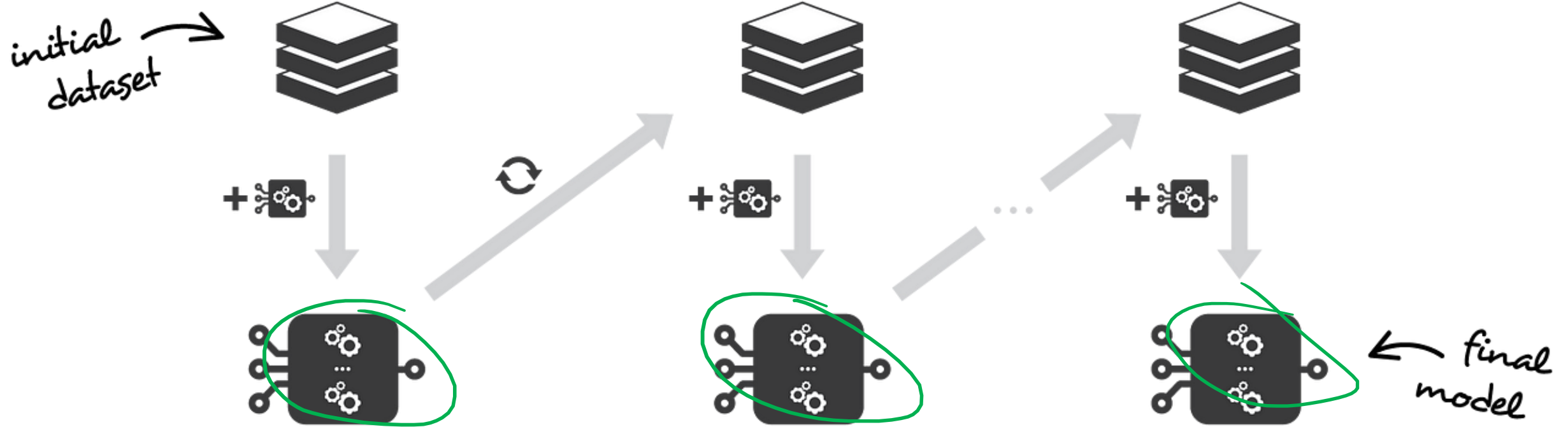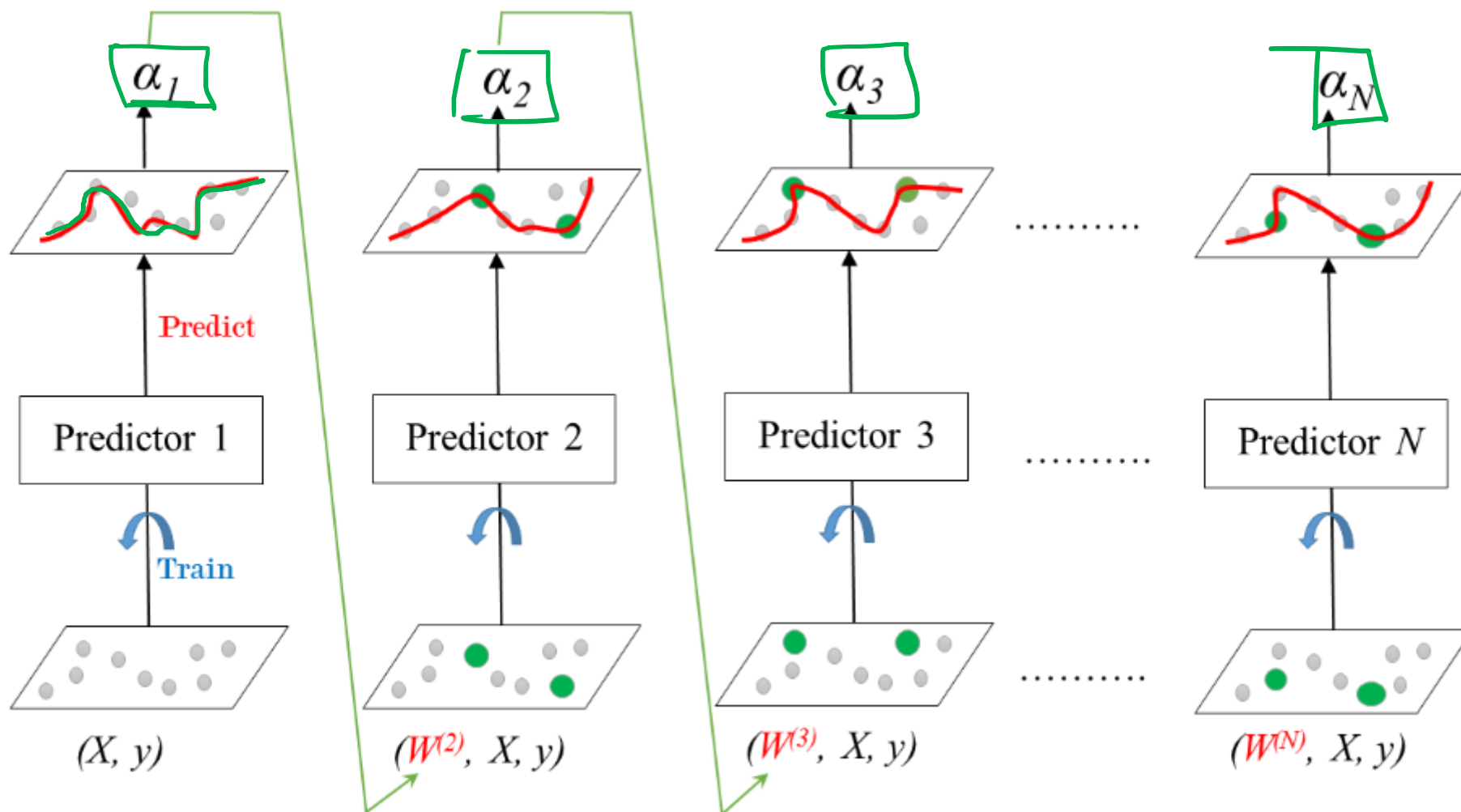# Random Forests
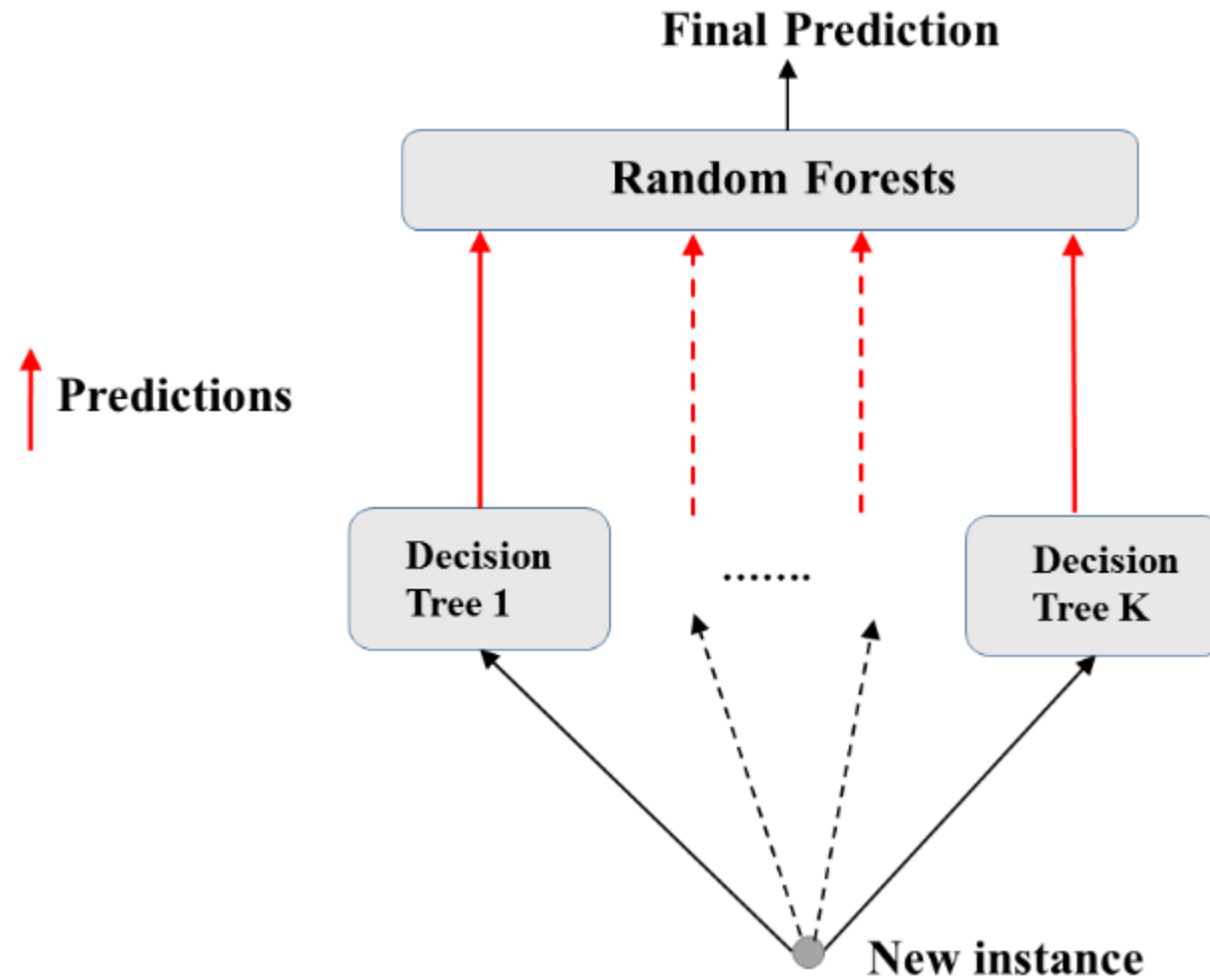
Boosting

Boosting

train a weak model and aggregate it to the ensemble model

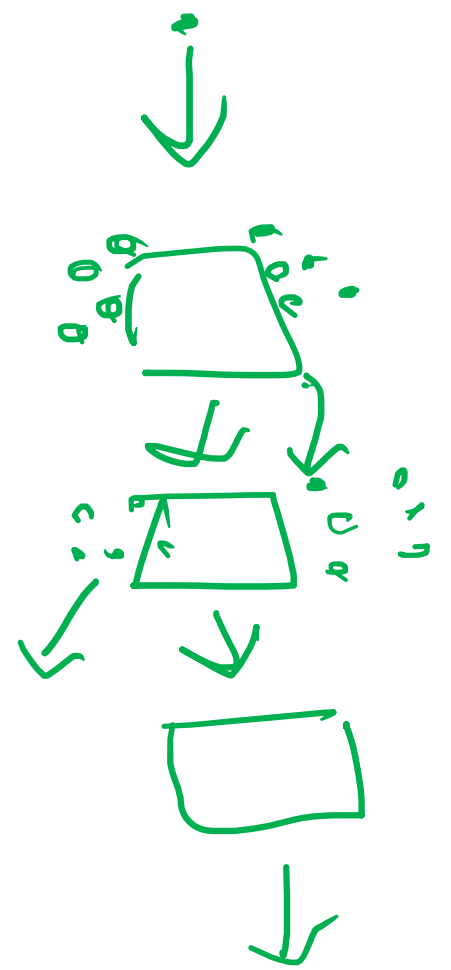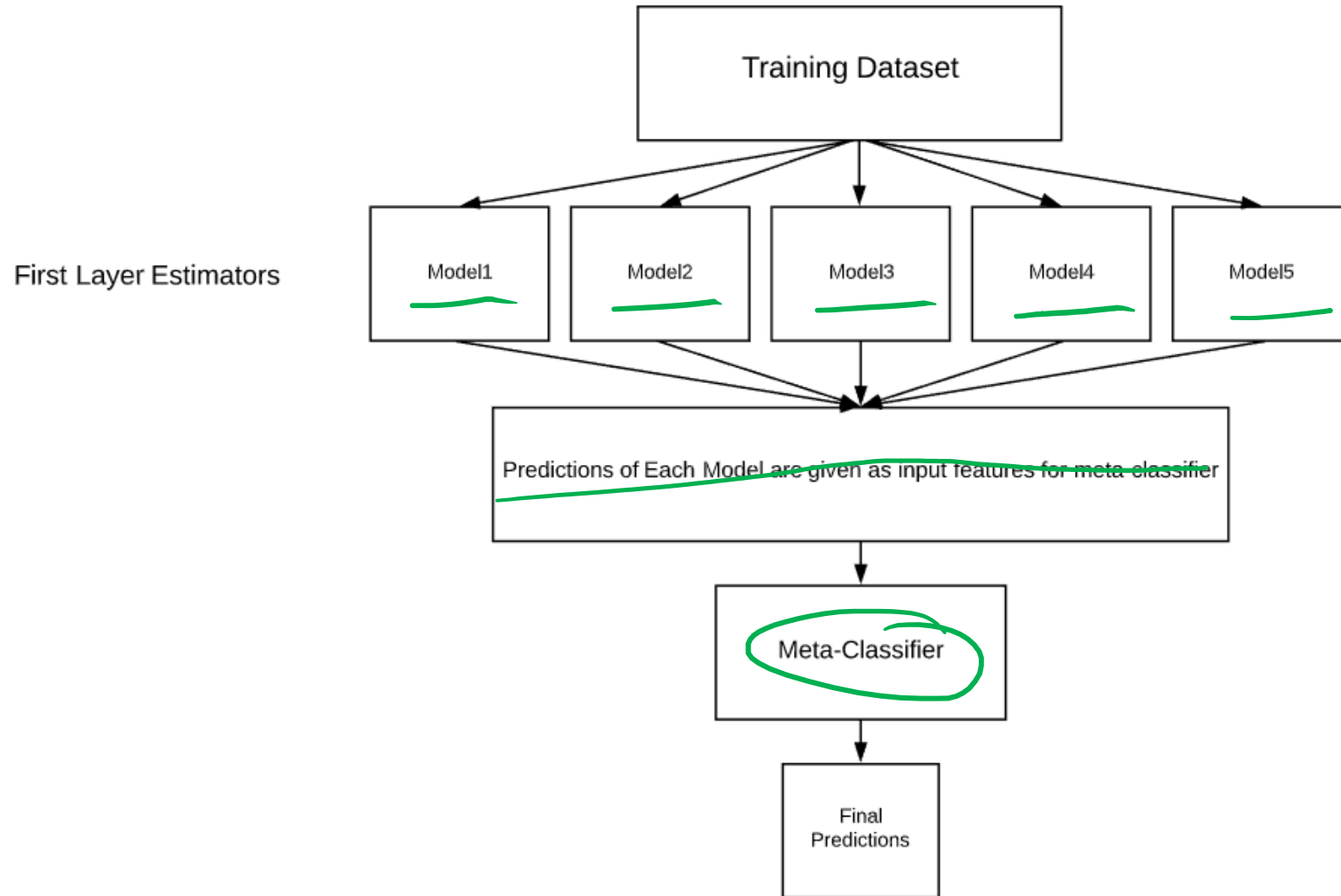update the training dataset (values or weights) based on the current ensemble model results

initial dataset

final model

AdaBoost

GBM (Gradient Boosting Machine)

$r_1 = y_1 - \hat{y}_1$ $r_2 = r_1 - \hat{r_1}$ $r_3 = r_2 - \hat{r_2}$ $r_N = r_{N-1} - \hat{r}_{N-1}$

Predict

Tree 1   Tree 2   Tree 3   ....   Tree N

Train

(X, y) $(X, \boldsymbol{\eta} \, r_1)$ $(X, \boldsymbol{\eta} \, r_2)$ $(X, \boldsymbol{\eta} \, r_{N-1})$

Stacking

Bibliography

https://app.datacamp.com/learn/courses/machine-learning-with-tree-based-models-in-python

https://towardsdatascience.com/ensemble-methods-bagging-boosting-and-stacking-c9214a10a205

https://www.geeksforgeeks.org/stacking-in-machine-learning/