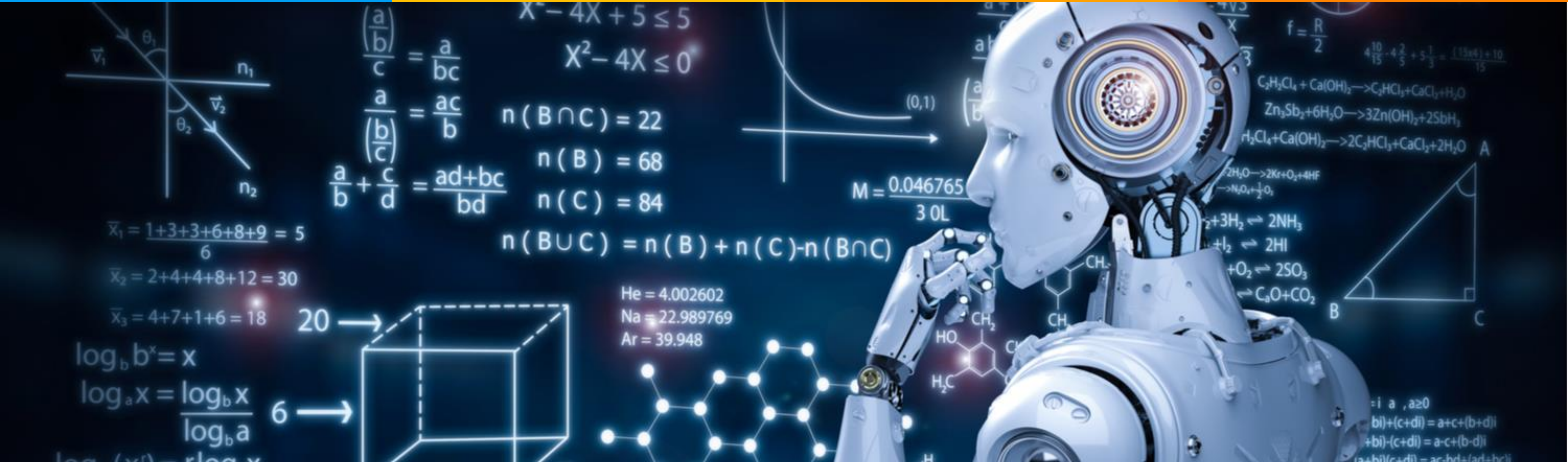
Duration: 60 min | Srikanth Shenoy | AML 5102 | M.E. AI & ML



Lecture 18: Ensemble Learning Basics & Random Forests

Recap

- Decision Tree
- Gini Impurity
- DT Pruning

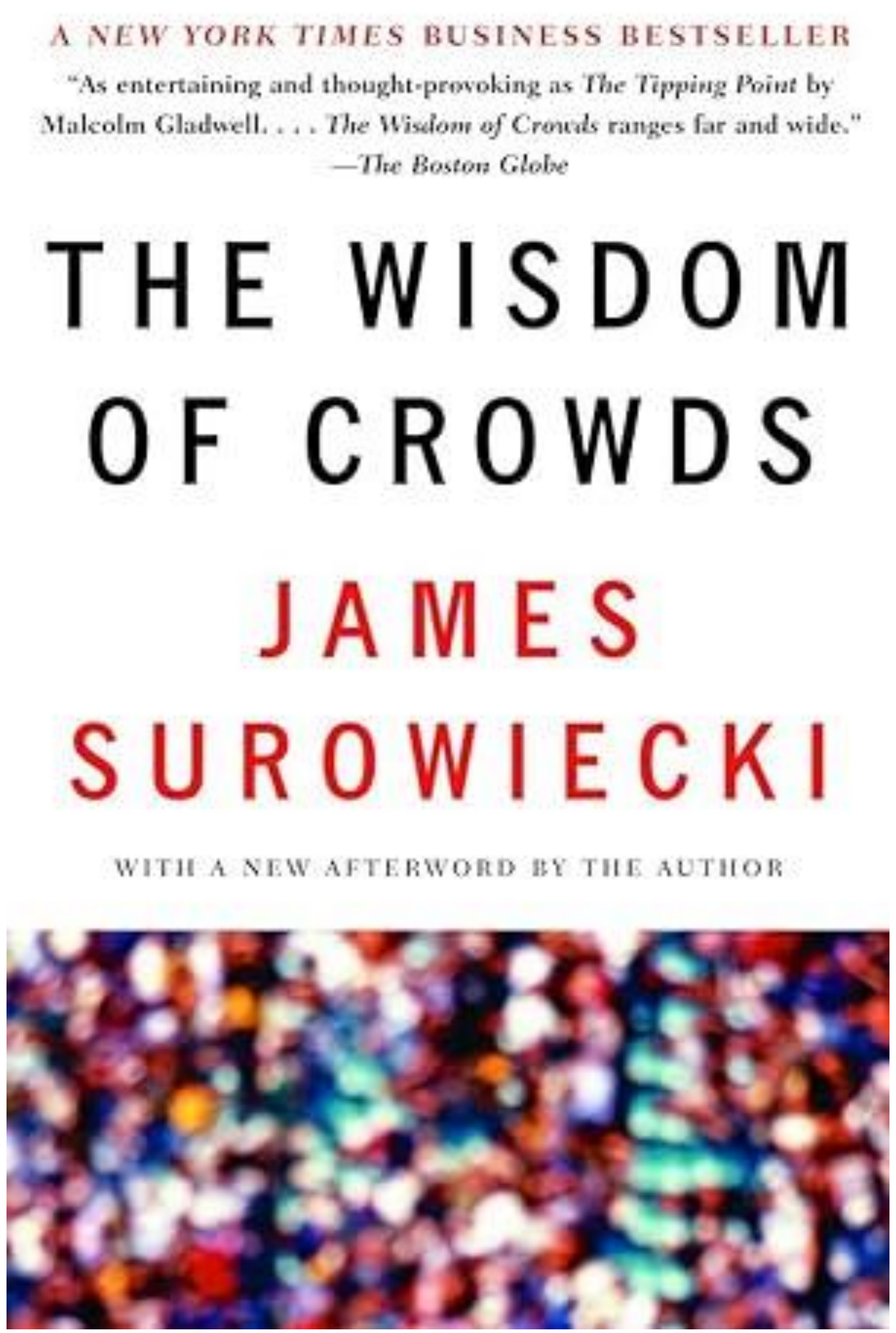
Wisdom of the Crowds

A game of guessing?

- How many candies in the jar?
- How to get the most correct answer?
- No answer is correct, some answers are useful
- Especially if averaged over a group

The wisdom of crowds

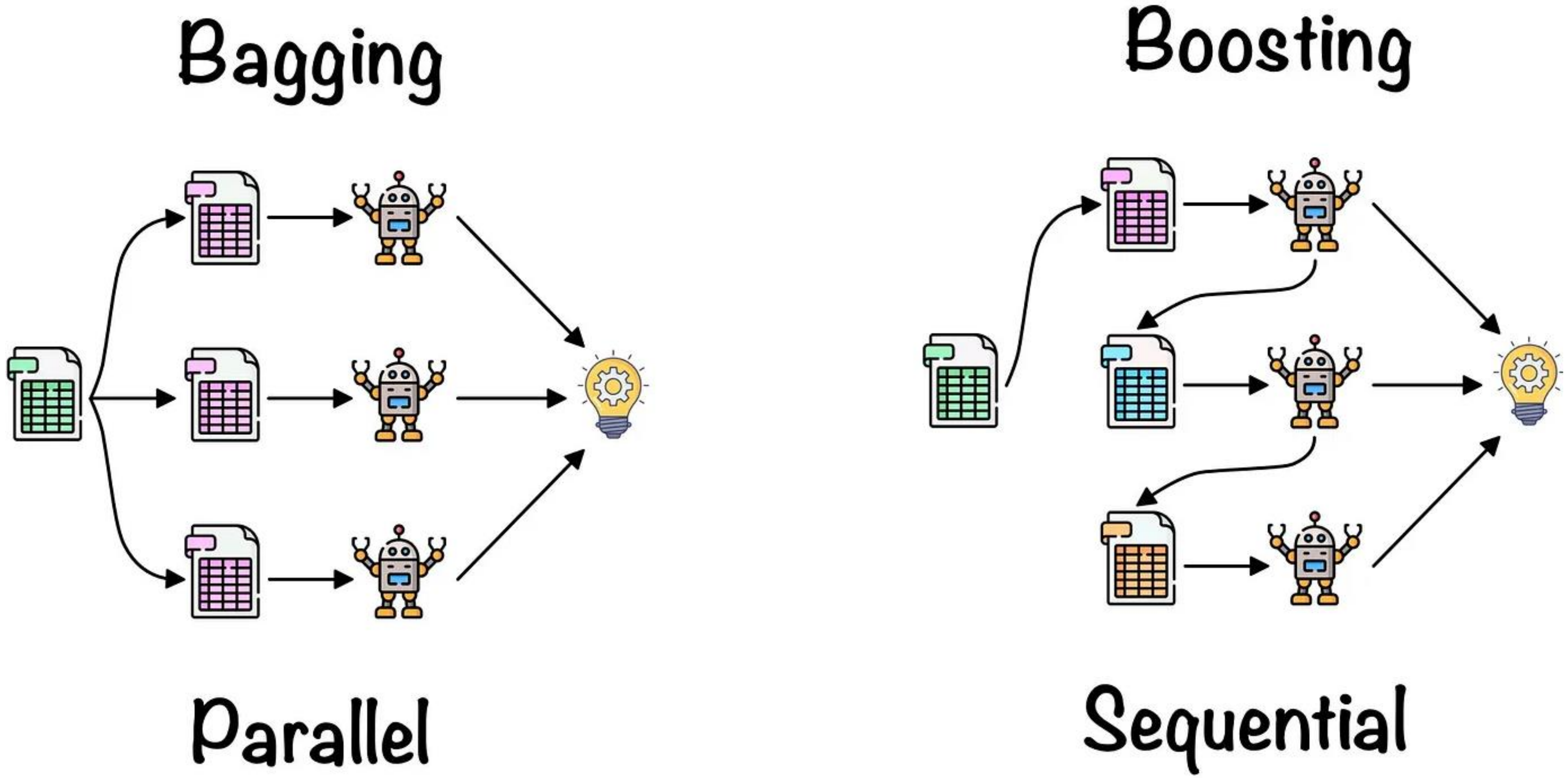
- “Under the right circumstances, groups are remarkably intelligent, and are often smarter than the smartest people in them”
- Using a Group for better decision
  - Diverse background
  - Independent decision by each individual
  - Good method for aggregation



Ensemble Learning

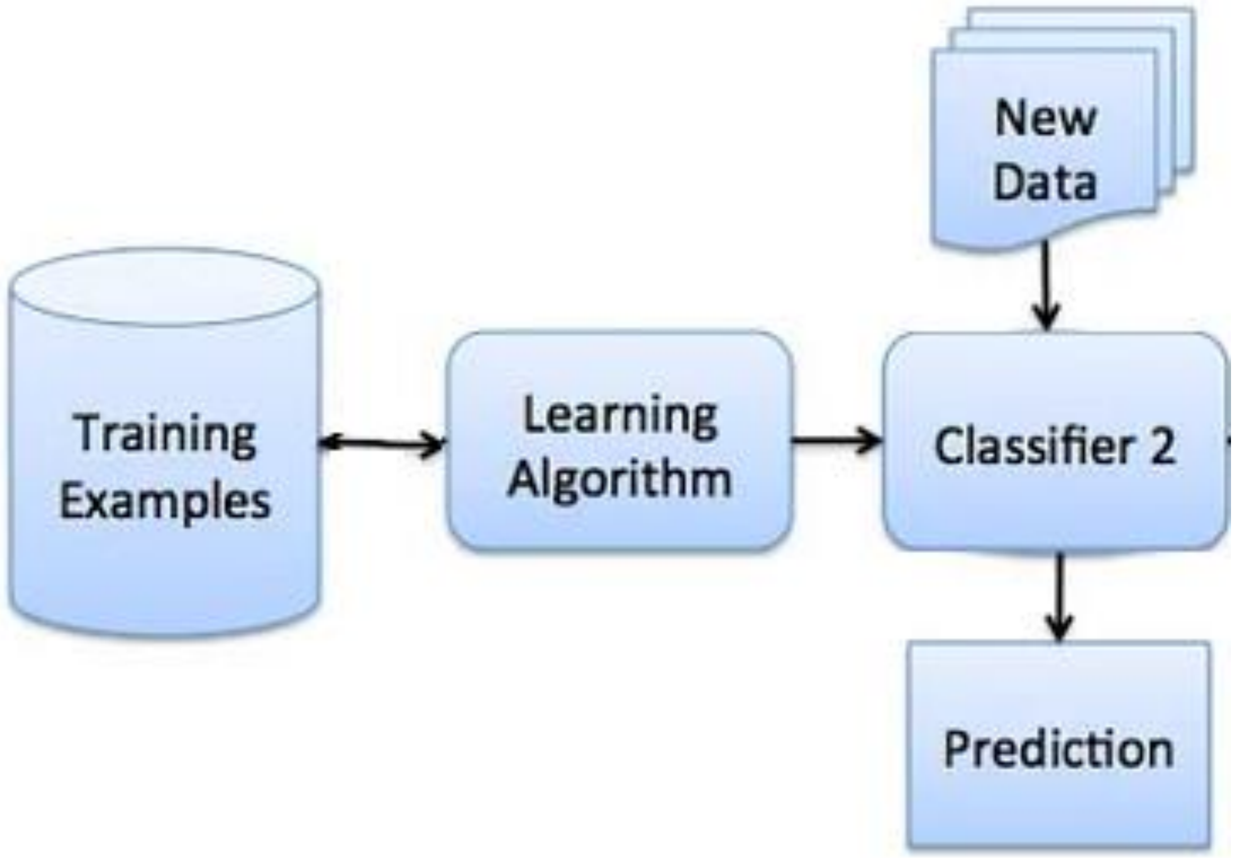
Ensemble Learning

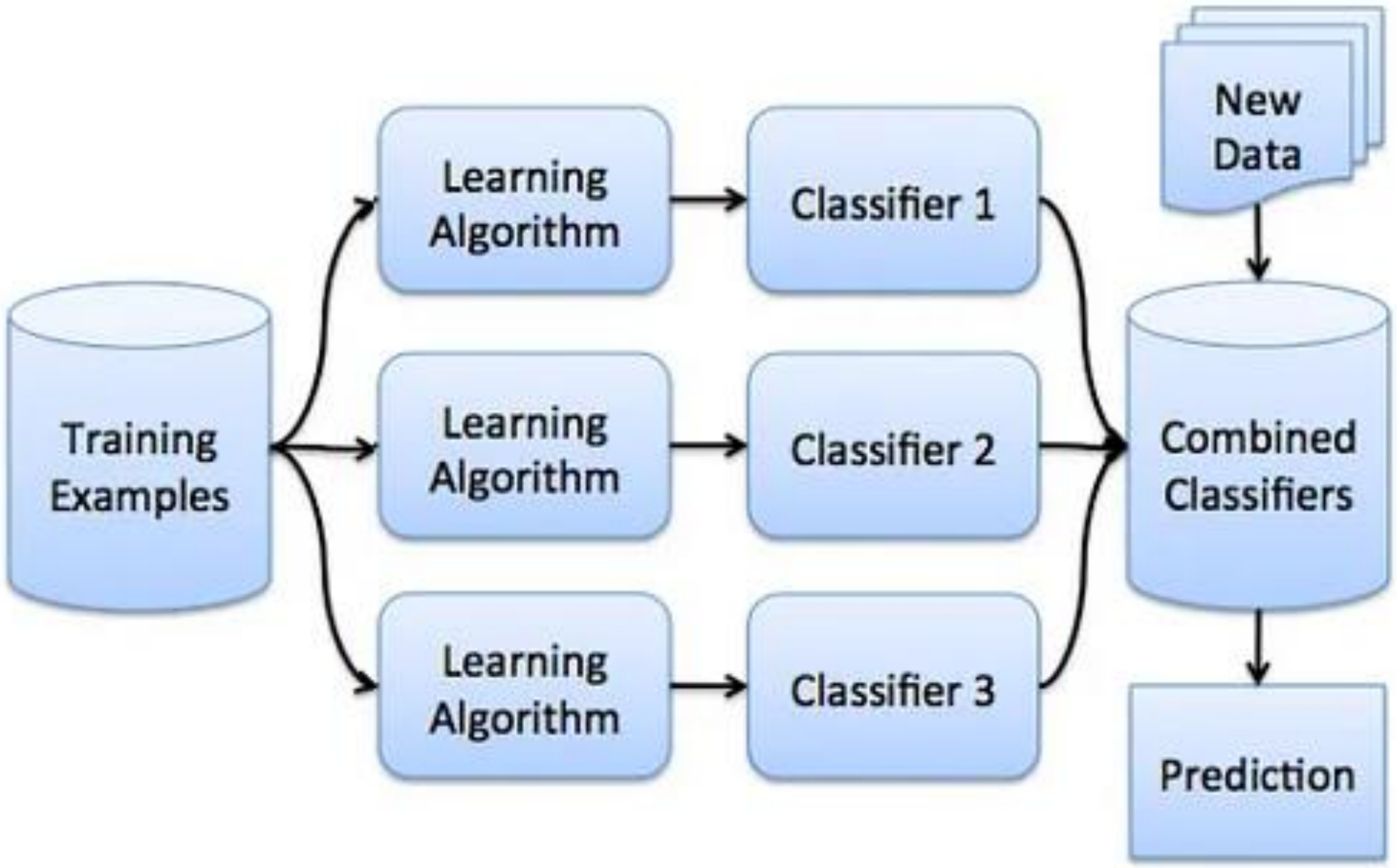
- Multiple ML model used together for prediction



Majority Voting/Averaging models

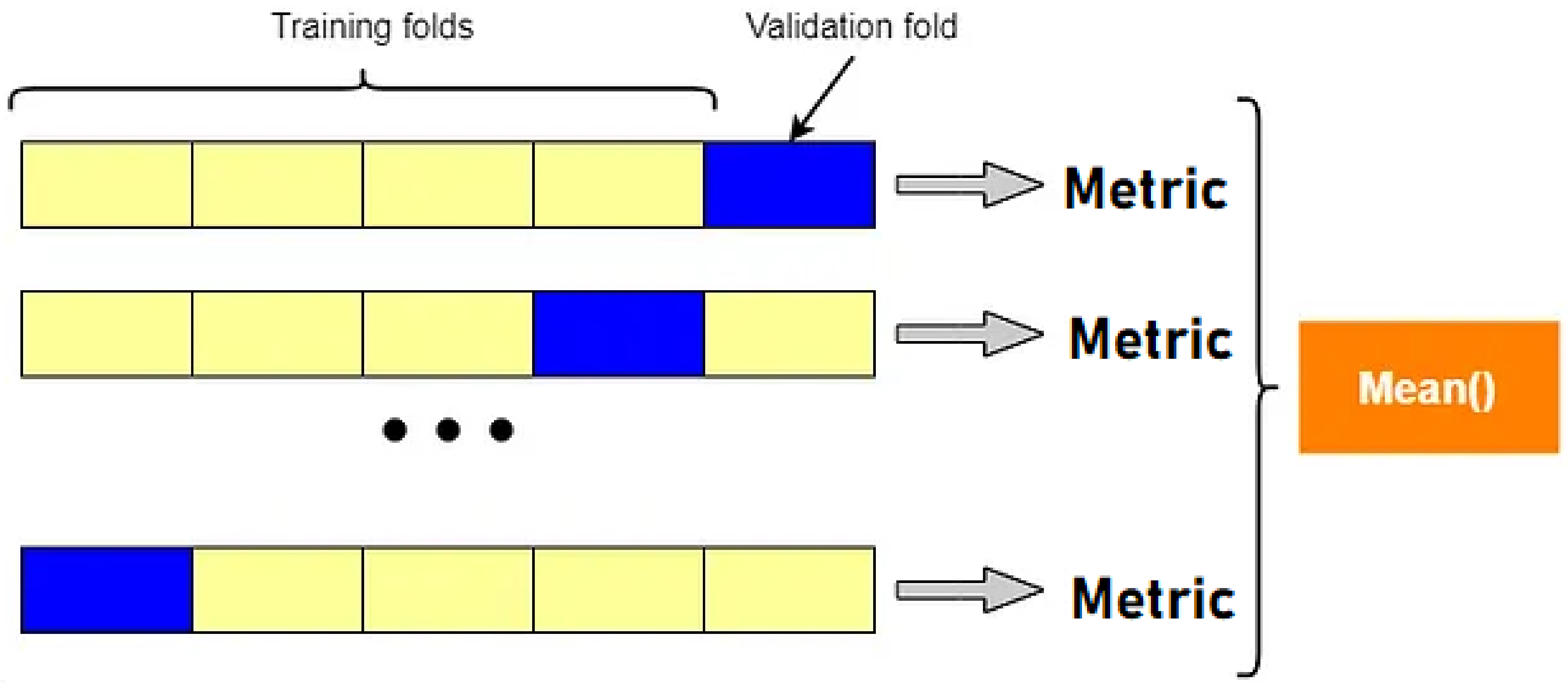
- High variance from single model
- How about using multiple models?
  - Will it reduce variance?
  - Not if data is correlated





A second look at K-Fold CV

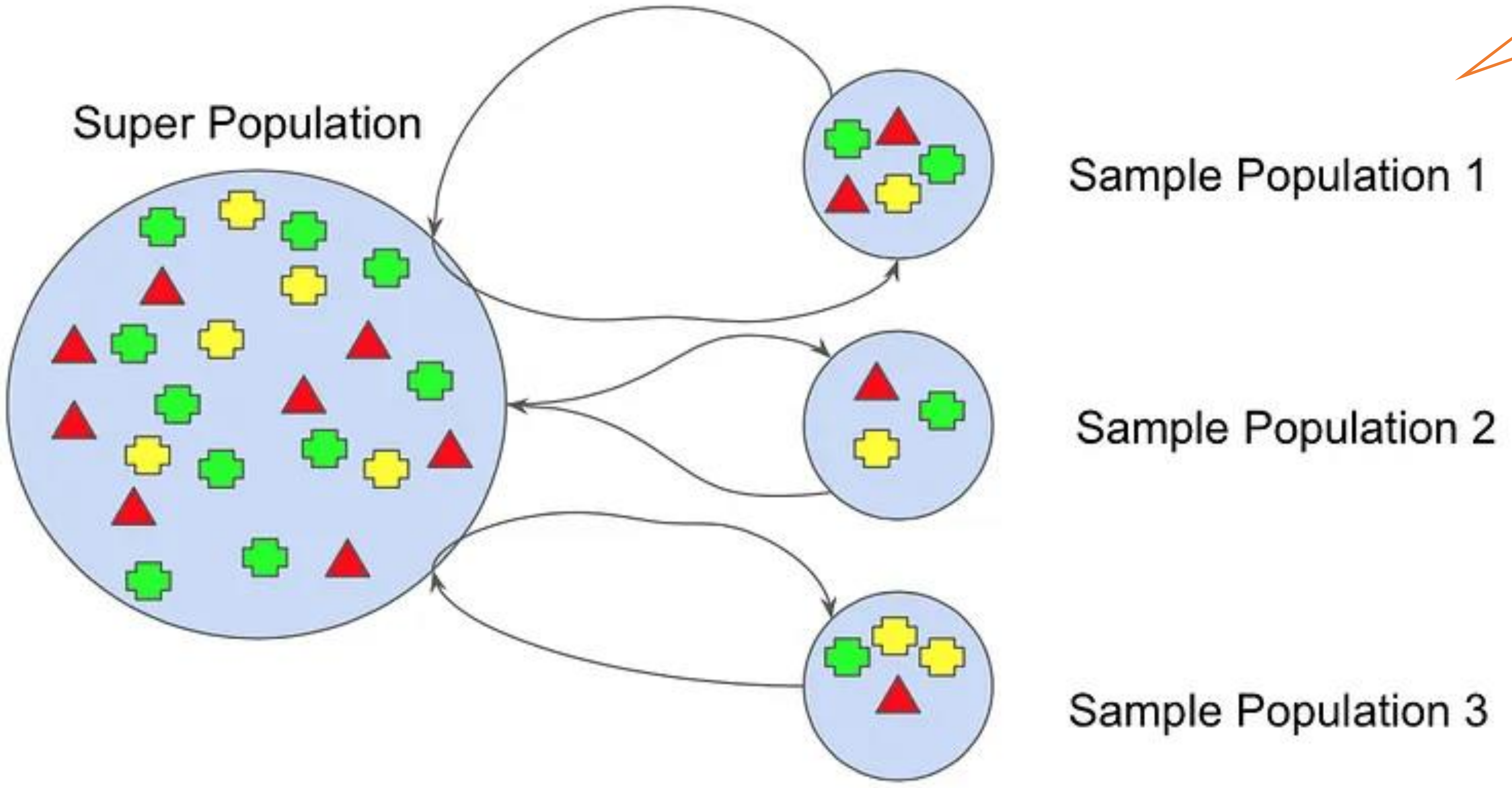
- Most data is repeated across folds
- IID from one record to next.
- Highly correlated from one fold to another



Solution: Bagging

- Bagging = Bootstrapping + Aggregation
- Bootstrapping: Sampling with Replacement

Data becomes uncorrelated



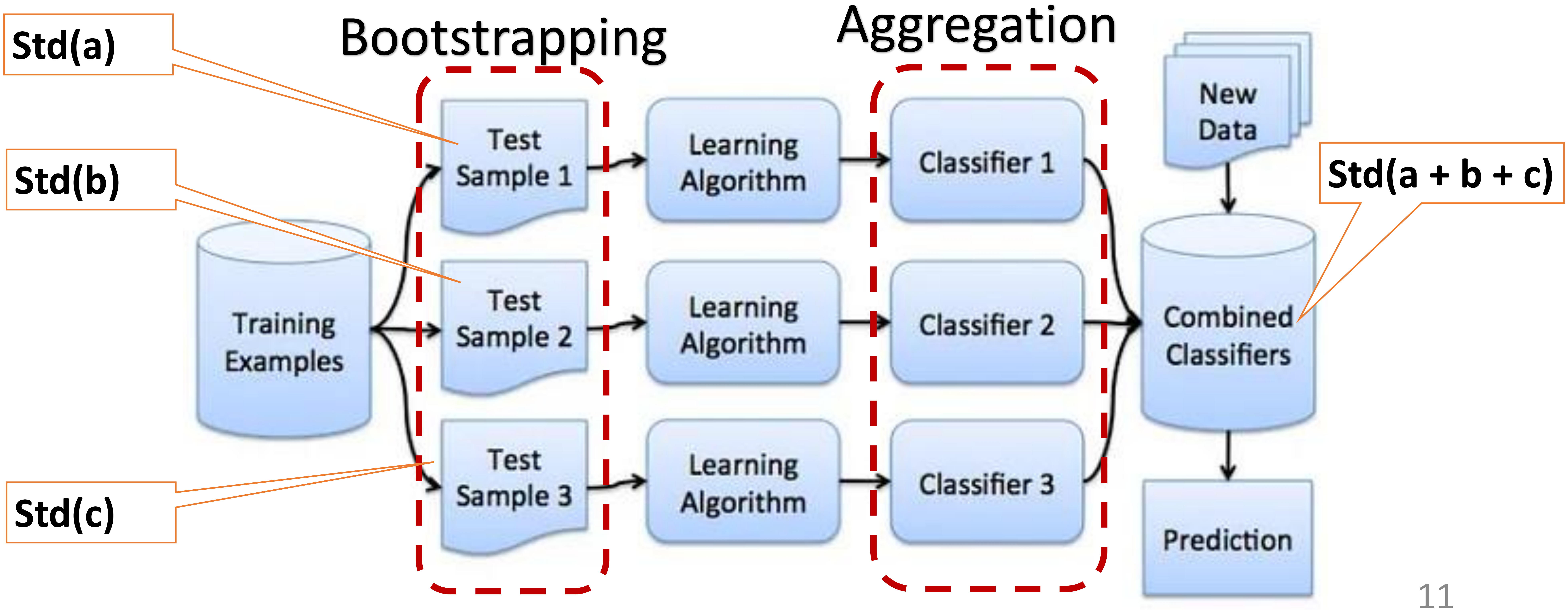
Std(a)

Std(b)

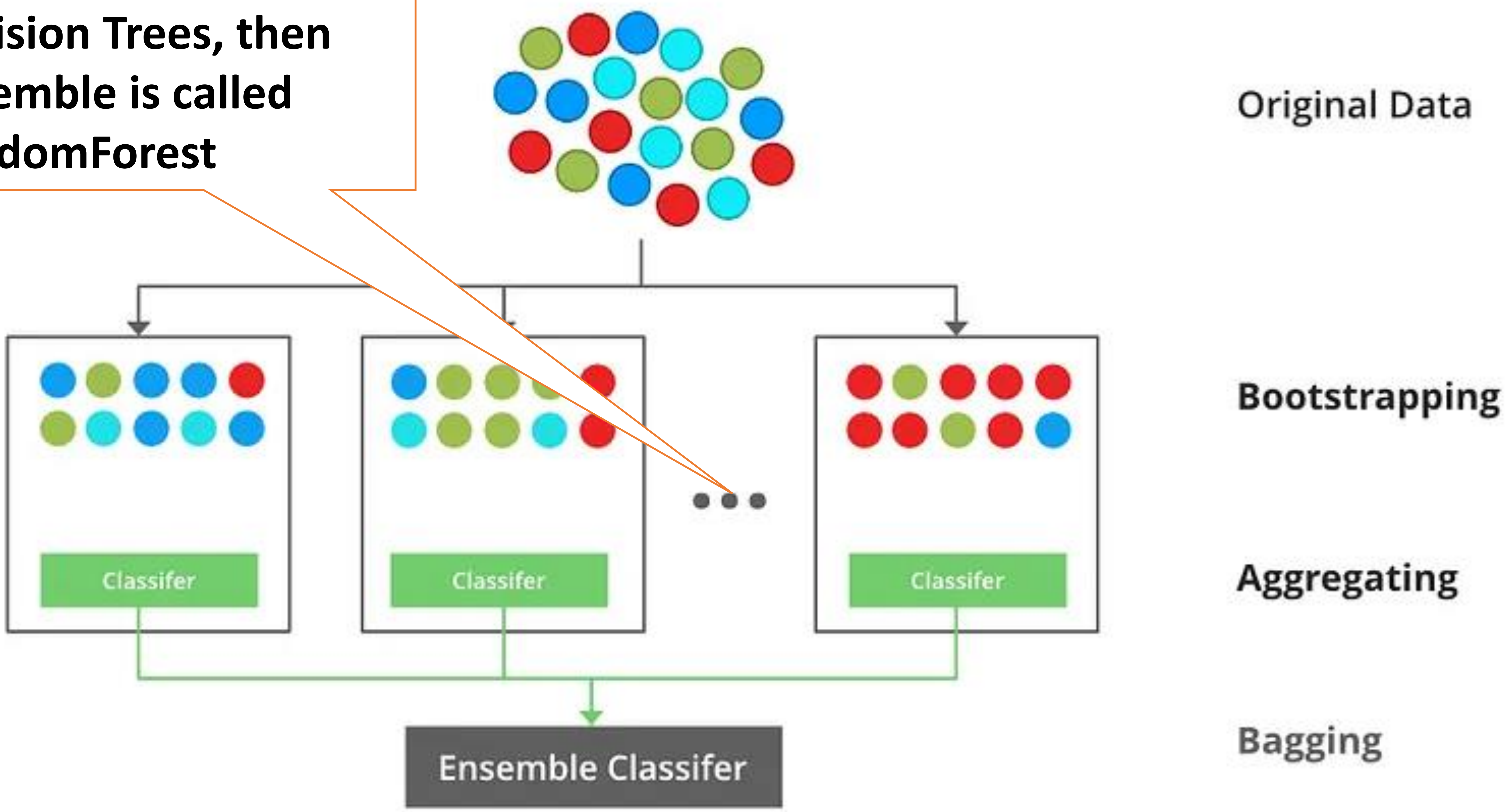
Std(c)

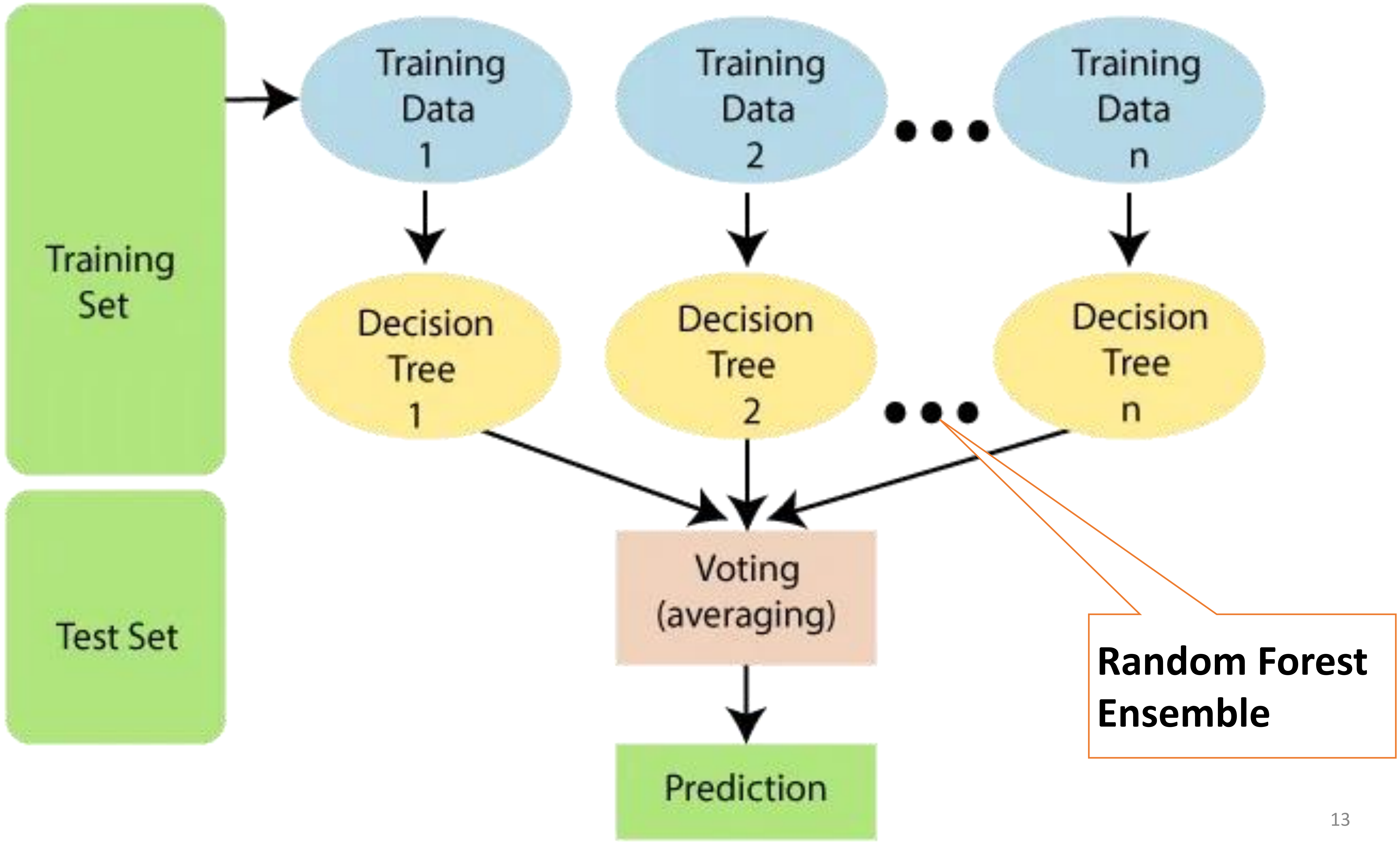
Solution: Bagging (Contd.)

- Aggregation = Combining classifiers



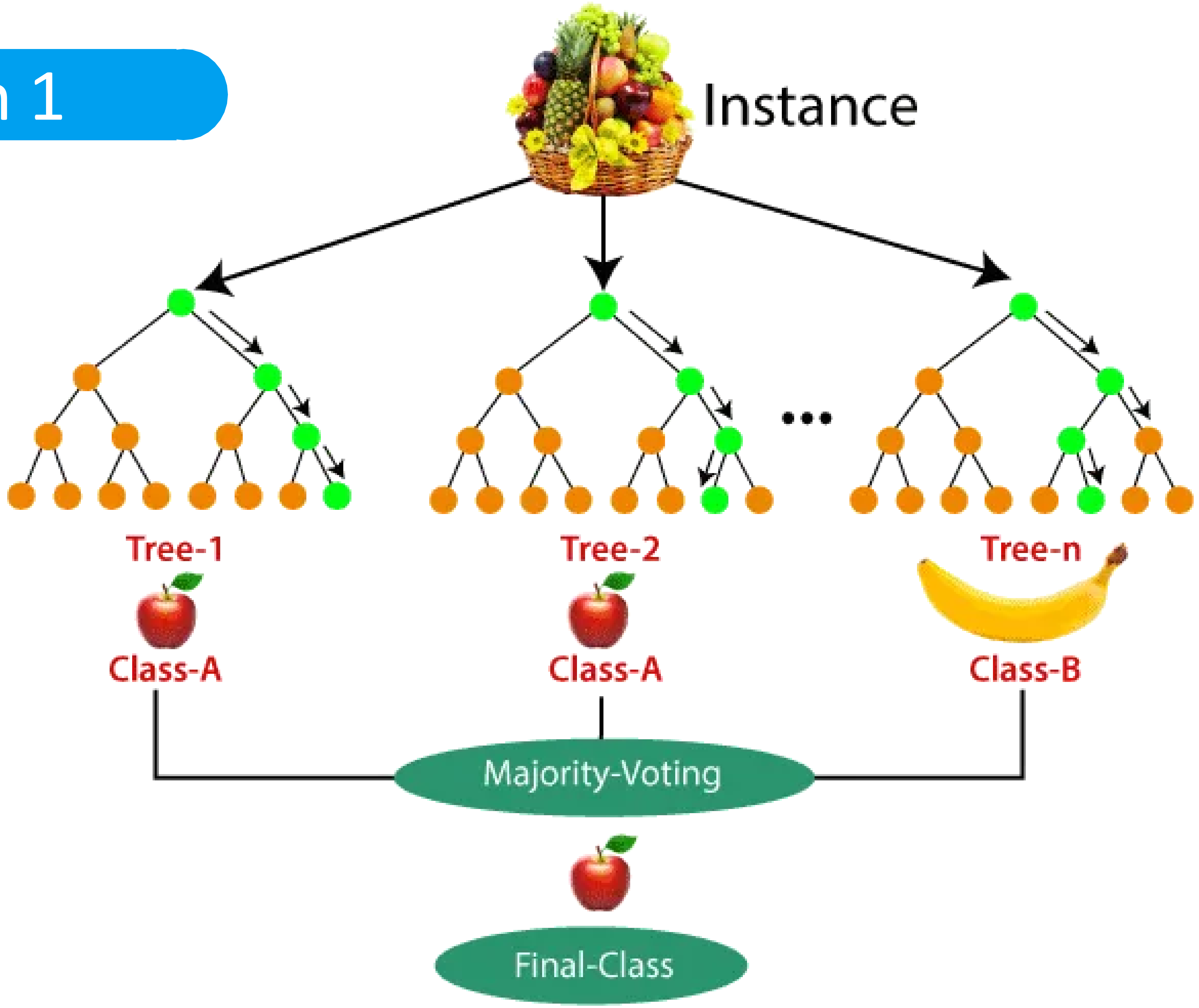
When all classifiers are Decision Trees, then ensemble is called RandomForest




Training Set
Test Set
Training Data 1
Training Data 2
Training Data n
Decision Tree 1
Decision Tree 2
Decision Tree n
Voting (averaging)
Prediction
Random Forest Ensemble

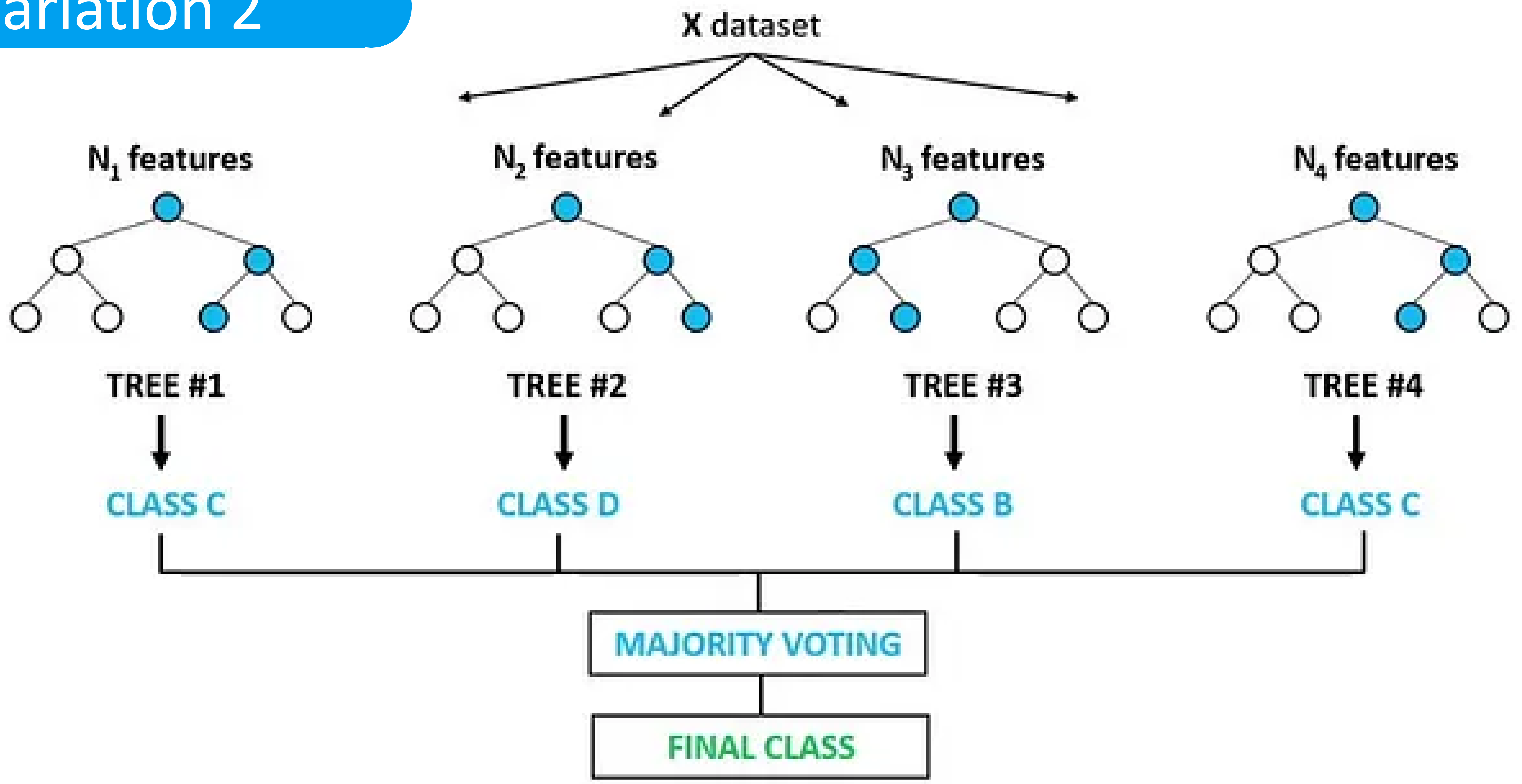
Random Forest

Variation 1



Instance
Tree-1
Class-A
Tree-2
Class-A
...
Tree-n
Class-B
Majority-Voting
Final-Class

Variation 2


X dataset
N1 features
N2 features
N3 features
N4 features
TREE #1
TREE #2
TREE #3
TREE #4
CLASS C
CLASS D
CLASS B
CLASS C
MAJORITY VOTING
FINAL CLASS

Random Forest hyperparameters

- Decision Tree hyper params
  - Criteria = gini/entropy
  - Max tree depth, Max leaf nodes
  - Min samples in split, Min samples in leaf
- Bagging hyperparams: Bootstrap Y/N
- Random Forest hyper params
  - Number of trees (n_estimators)
  - Max samples per tree
  - Max features per tree

Random Forest advantages

- Feature correlation does not matter
- Feature distribution does not matter
- No need to scale data
- Works well even when data is missing
- Overcomes the problem of overfitting
- Not very expensive
- Flexible and high accuracy

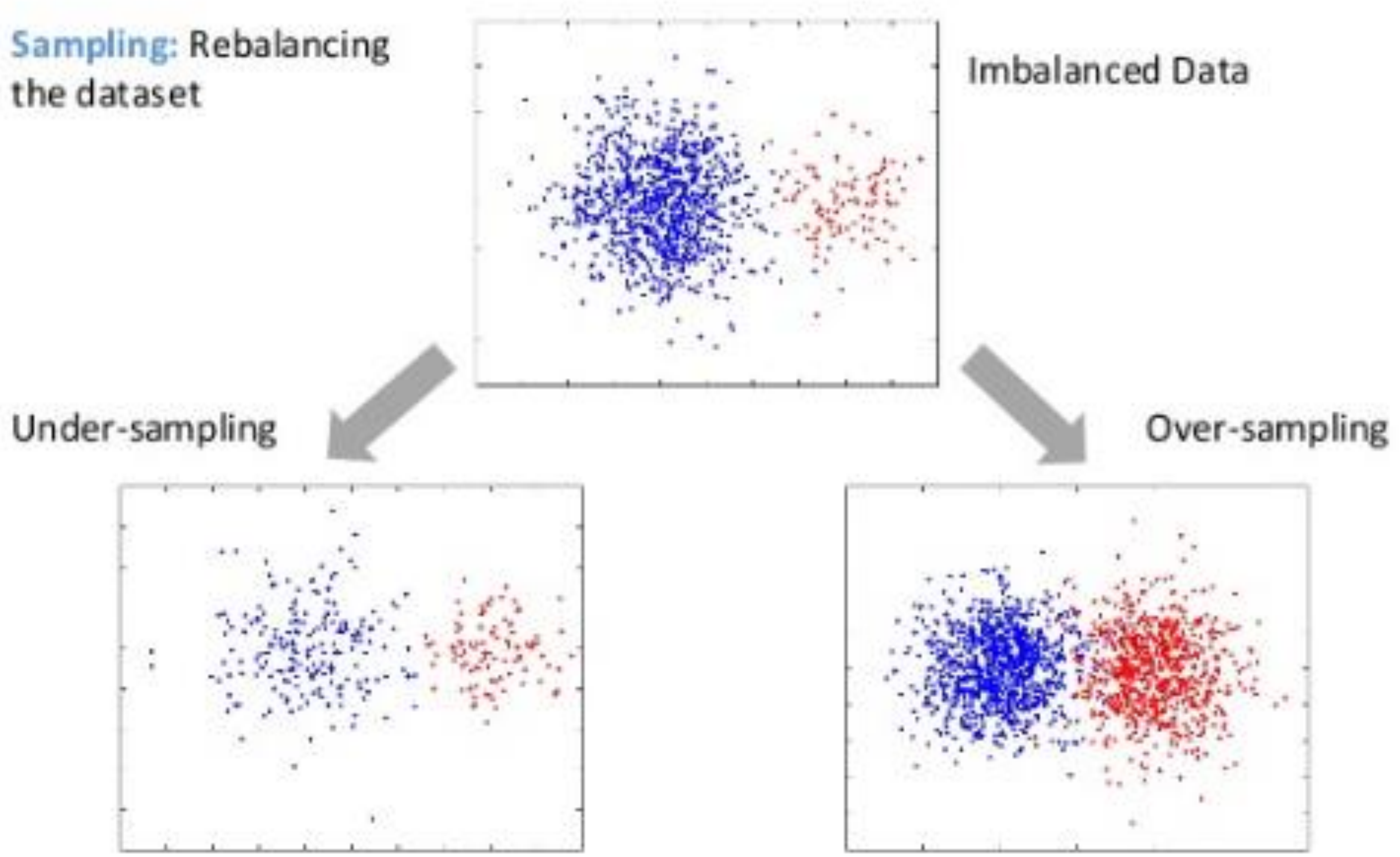
Imbalanced data

Accuracy metric and Imbalanced dataset

- A dataset has 98% majority class and 2% minority class
- A model was developed and it gave accuracy of 0.9499
- Is it a good model?

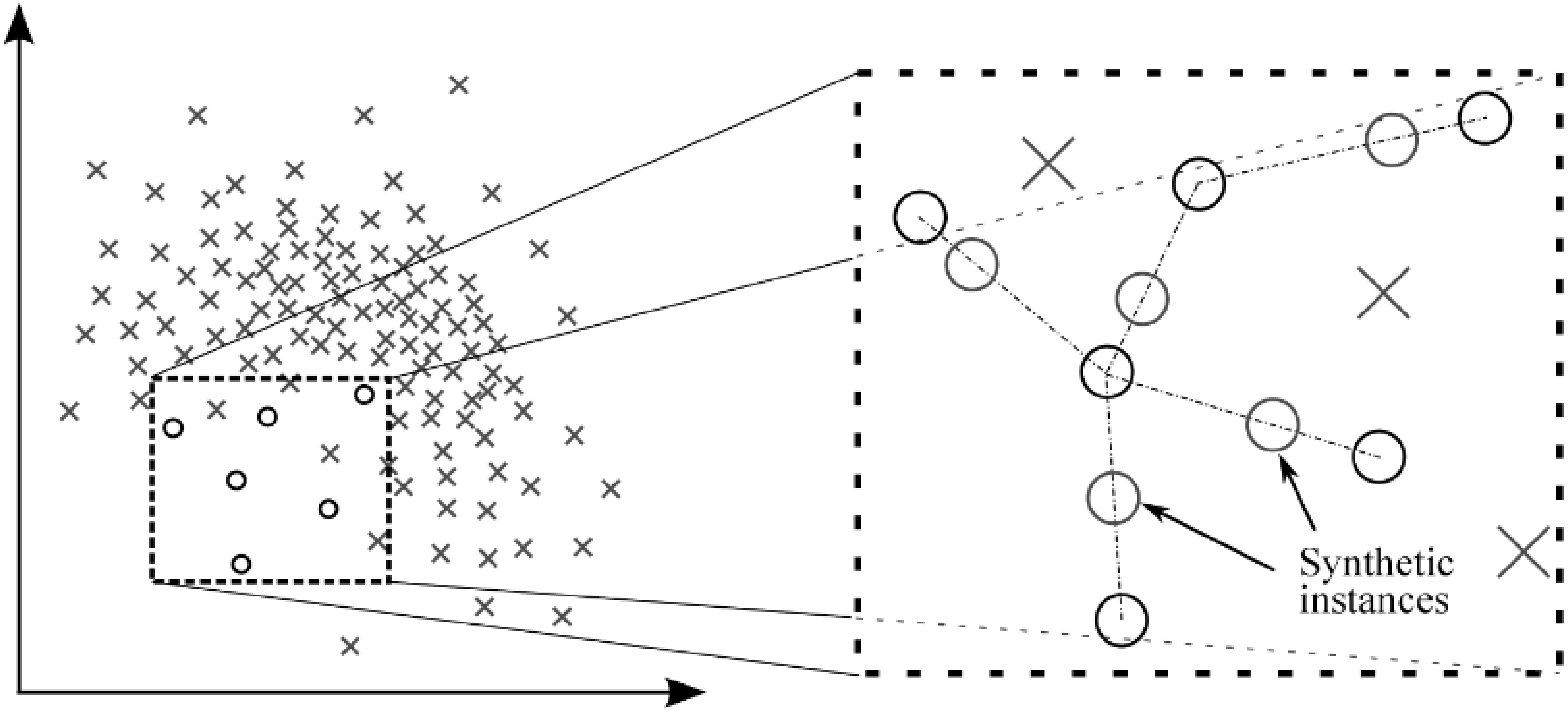
Oversampling minority class

- Synthetic Minority Oversampling Technique: SMOTE



SMOTE (Continued)

- Uses KNN on minority class

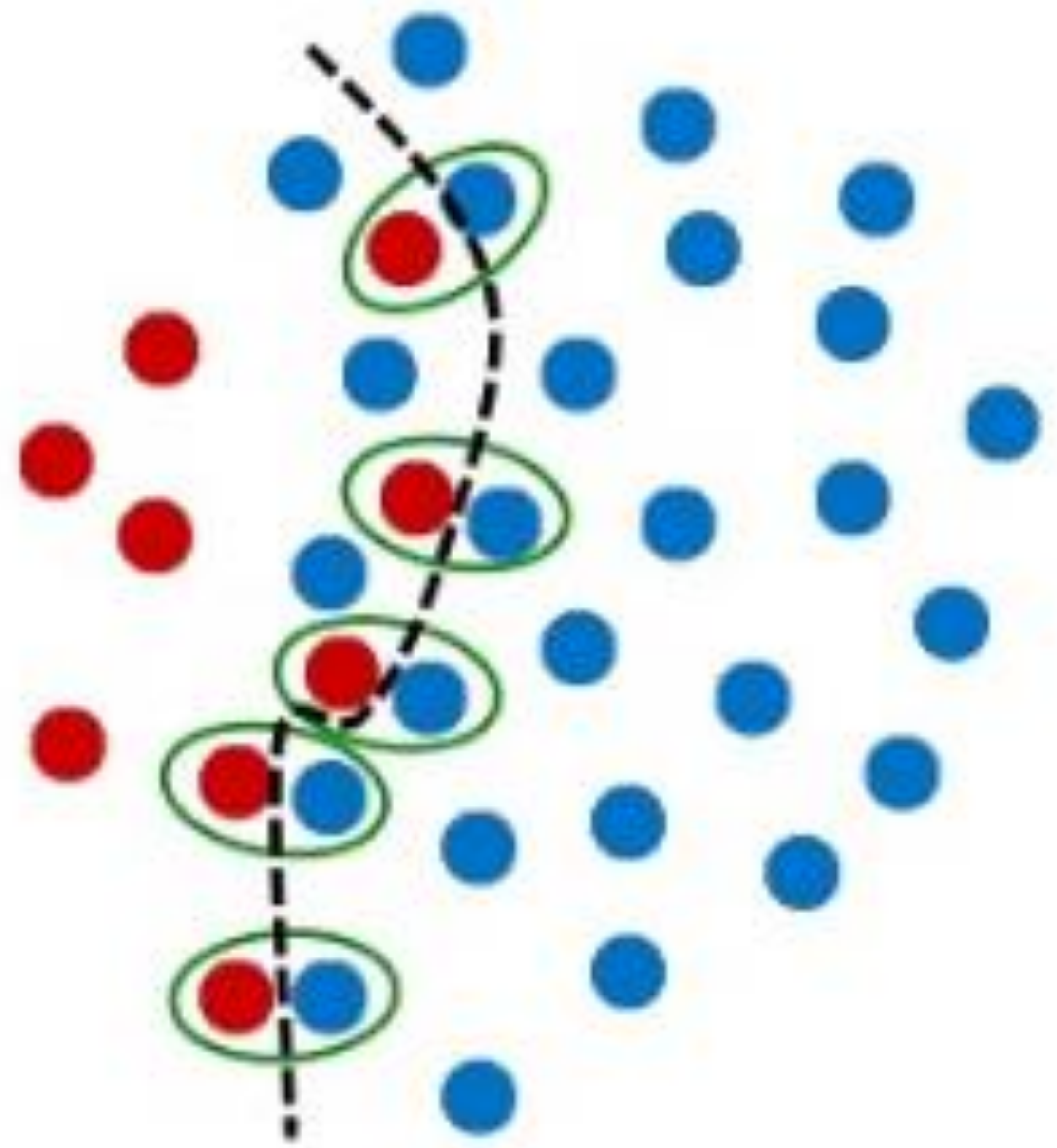


SMOTE (Continued)

- kNN SMOTE
- DBSMOTE: Uses DBSCAN clustering algorithm for SMOTE

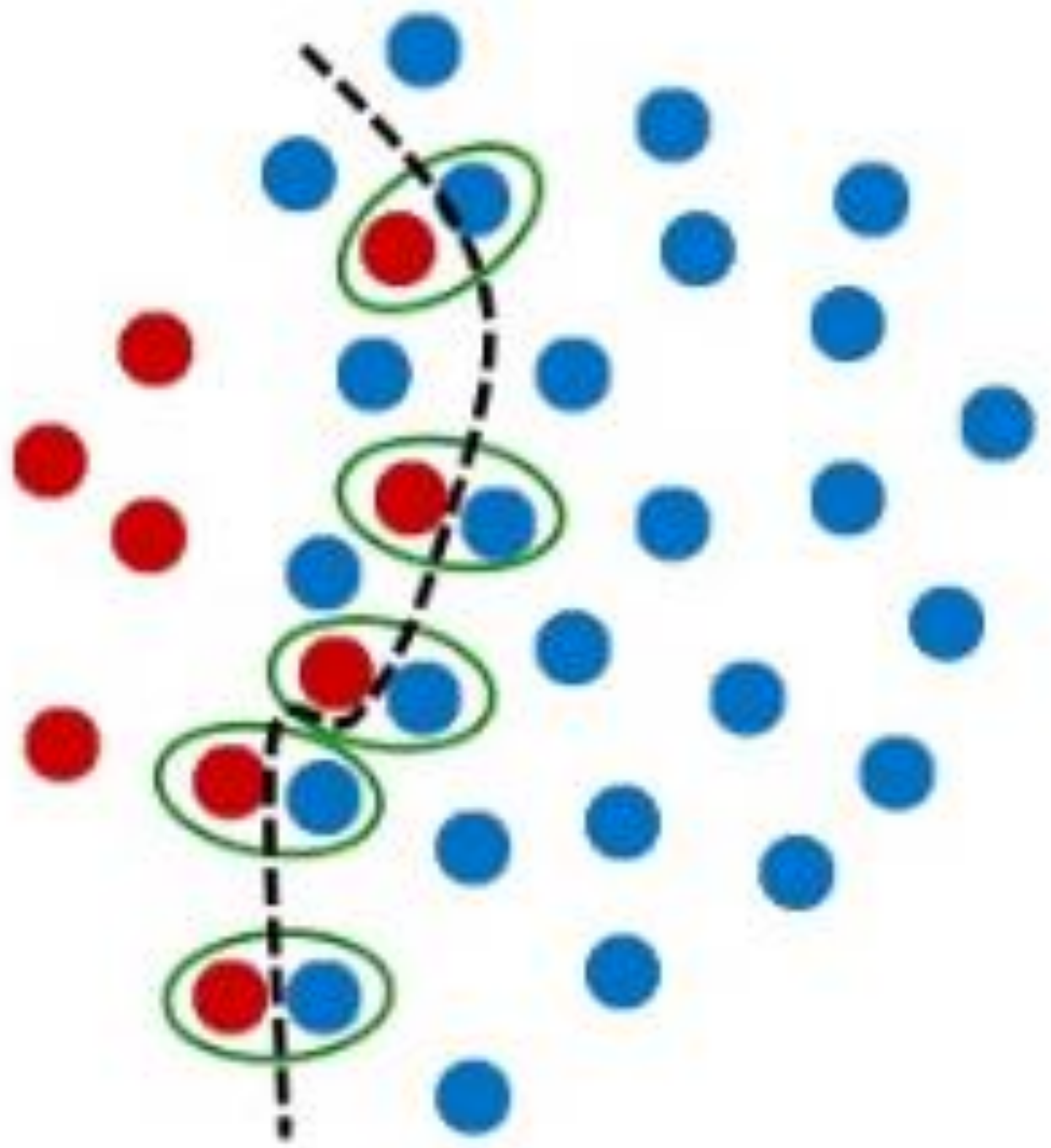
Undersampling majority class

- Random Undersampling
- Tomek Links



Combined oversampling and undersampling

- SMOTE Tomek



Handling Imbalanced dataset in Random Forest

- Class weight
  - Reciprocal of proportion of records per class
  - Balanced – default value
  - Balanced sub sample – each tree gets sub samples based on class weight

- Let us look at all of this in lab


QUESTIONS