DLI Accelerated Data Science Teaching Kit

Lecture 14.9 - Random Forest

Ensemble Learning

Learning methods that learn a single hypothesis, chosen form a hypothesis space that is used to make predictions.

Ensemble learning is to select a collection (ensemble) of hypotheses and combine their predictions in order to reducing the error rate.

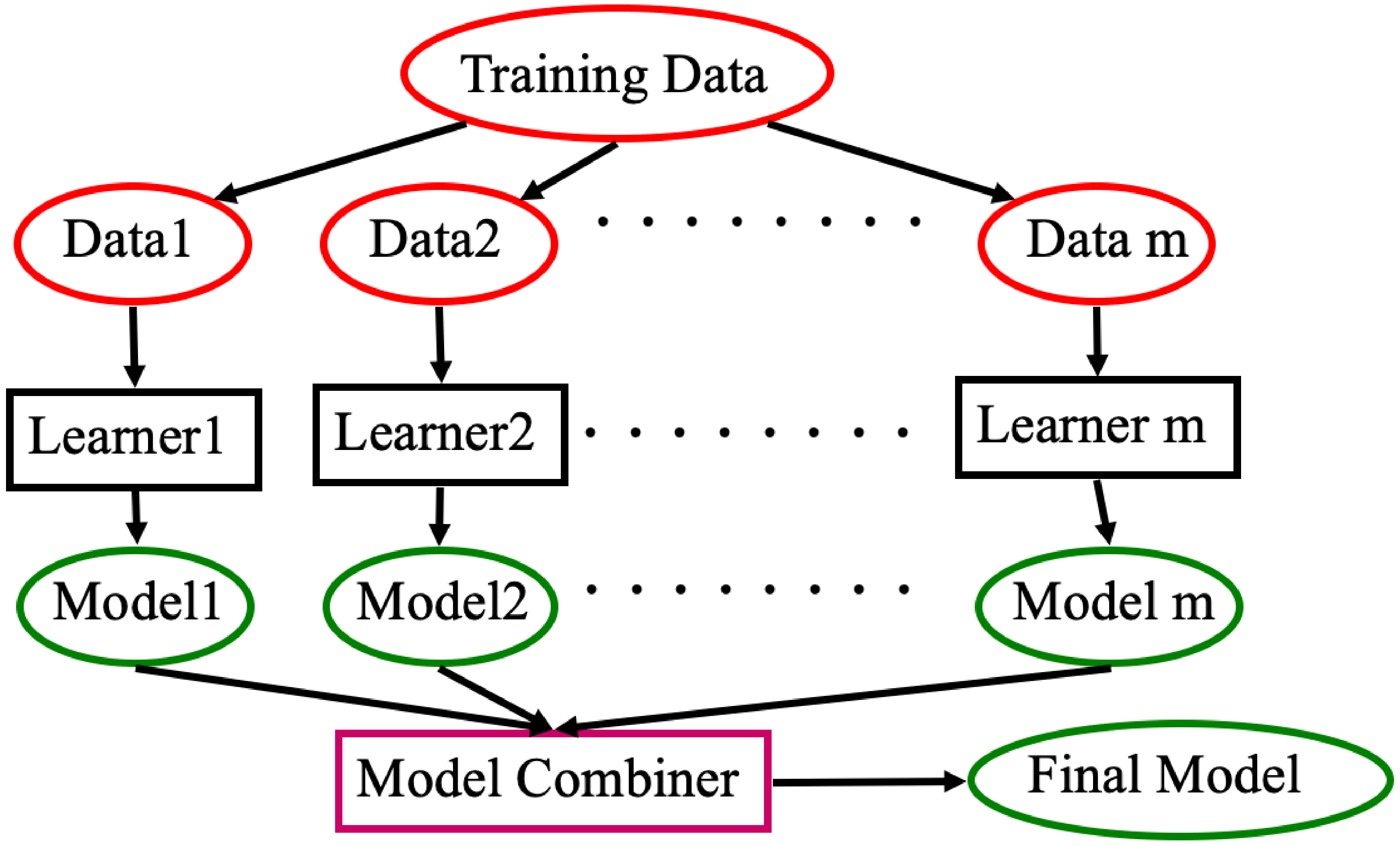
Example

- Generating 100 different decision trees from the same or different training set and have them vote on the best classification for a testing example.

Ensemble Learning

Learn multiple alternative definitions of a concept using different training data or different learning algorithms.

Combine decisions of multiple definitions, e.g., using majority voting.



Benefit

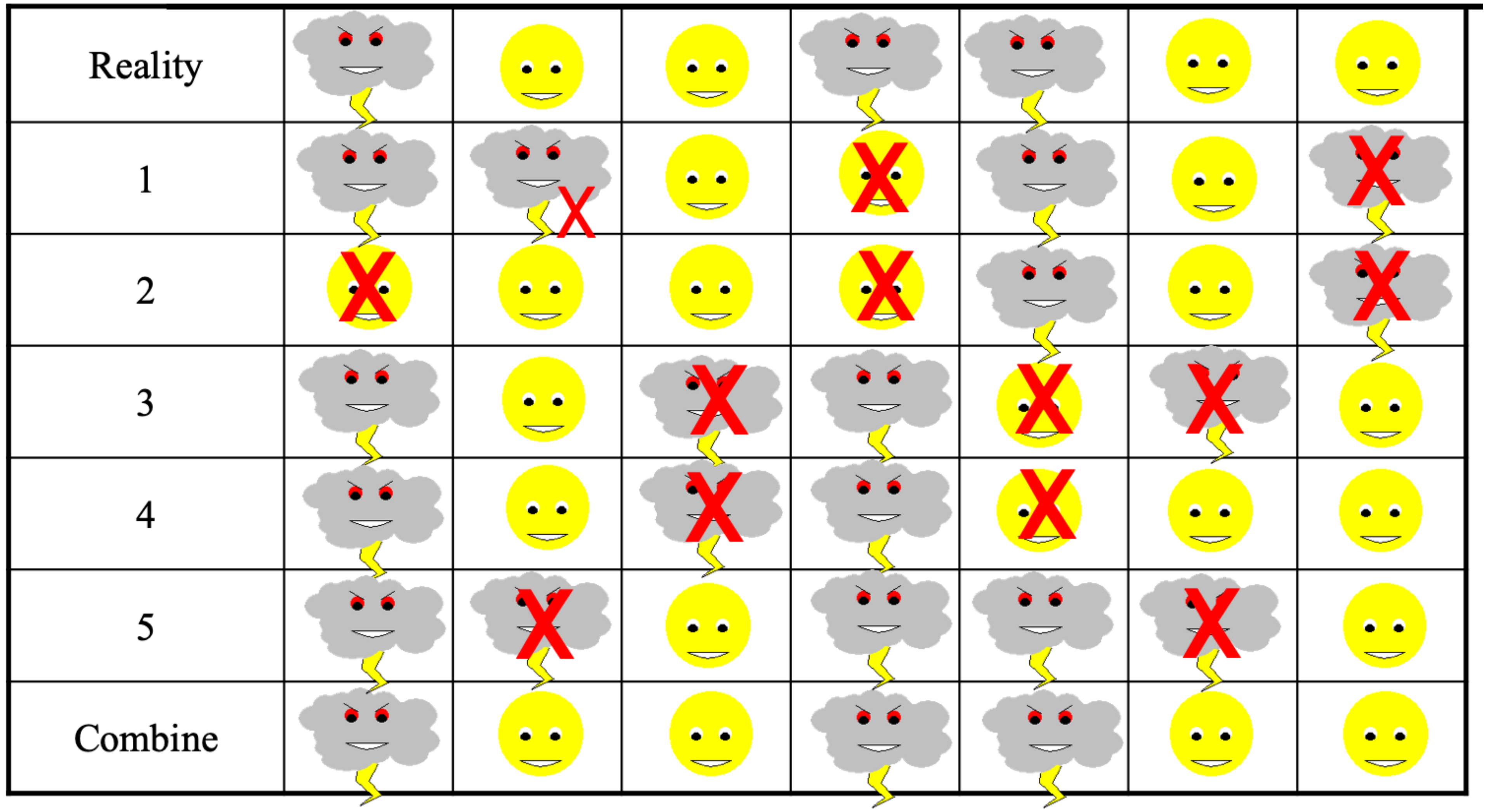
"No Free Lunch" Theorem

- No single algorithm wins all the time!

When combing multiple independent and diverse decisions each of which is at least more accurate than random guessing, correct decisions are reinforced.

Example: human ensembles are demonstrably better.

- How many jellybeans in the jar? Individual estimates vs. group average.
- Who Wants to be a Millionaire? Audience vote.



Weather Forecasting

Bagging Learning

Bagging

- Train M learners on M bootstrap samples
- Combine outputs by voting (e.g., majority vote)

Decreasing error by decreasing the variance in the results due to unstable learners

- Algorithms (like decision trees and neural networks) whose output can change dramatically when the training data is slightly changed.

Given a standard training set D of size n

For i = 1 .. M

- Draw a sample of size $n^* < n$ from D uniformly and with replacement
- Learn classifier C_i

Final classifier is a vote of C_1 .. C_M

Increases classifier stability/reduces variance

Boosting

Instead of sampling (as in bagging), re-weigh examples!

Samples are given weights.

At each iteration, a new hypothesis is learned (weak learner) and the samples are reweighted to focus the system on examples that the most recently learned classifier got wrong.

Final classification based on weighted vote of weak classifiers

C =0; /* counter*/
M = m; /* number of hypotheses to generate*/

1 Set same weight for all the examples (typically each example has weight = 1);

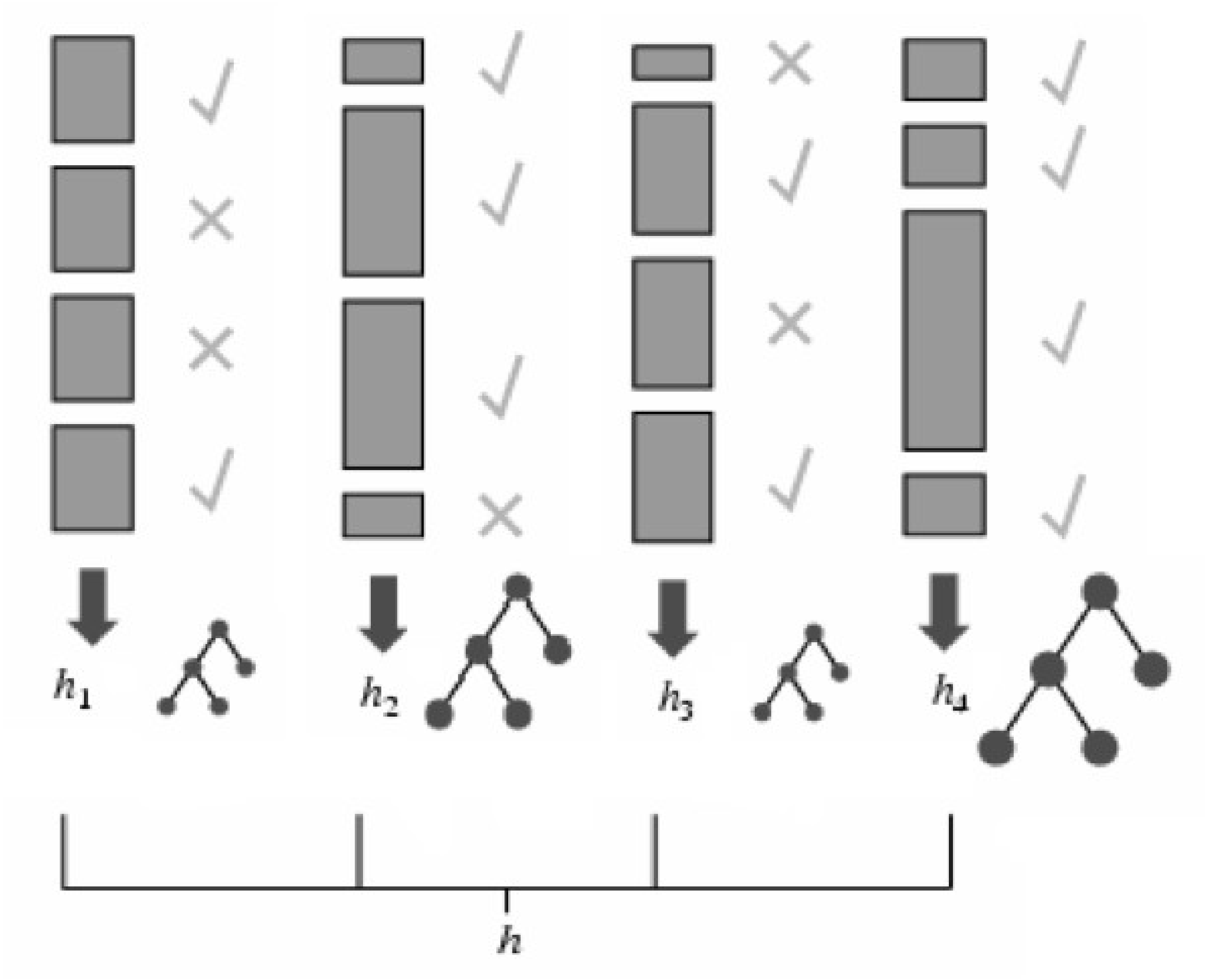
2 While (C < M)
  2.1 Increase counter C by 1.
  2.2 Generate hypothesis h_C .
  2.3 Increase the weight of the misclassified examples in hypothesis h_C
3 Weighted majority combination of all M hypotheses (weights according to how well it performed on the training set).

Adaptive Boosting

Each rectangle corresponds to a sample, with weight proportional to its height.

Crosses correspond to misclassified examples.

Size of decision tree indicates the weight of that hypothesis in the final ensemble.



Random Forest

An ensemble learning method for classification, regression and other tasks

Building multiple decision trees at training time

Outputting the class that is the mode of the classes or mean/average prediction of the individual trees

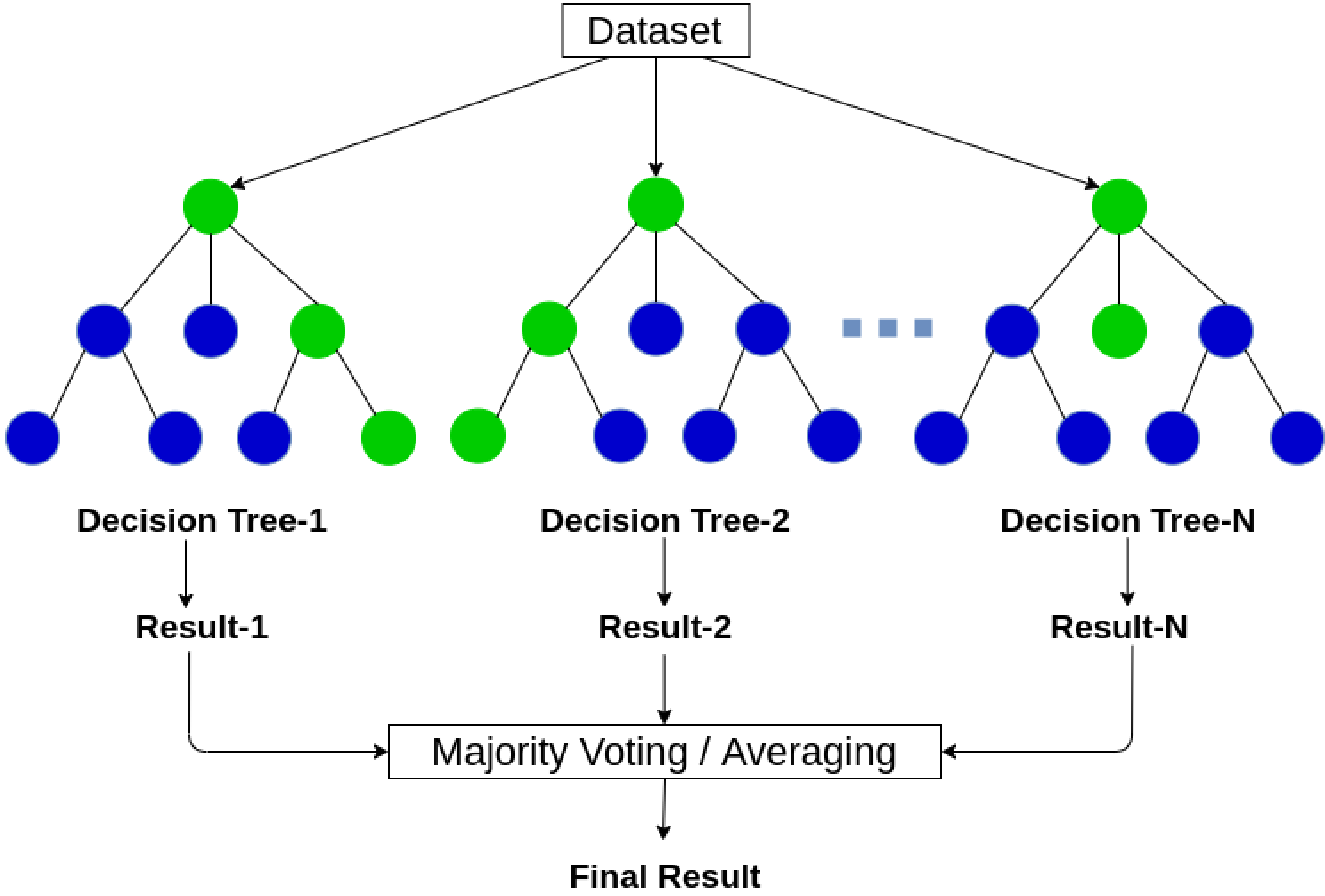
It generates an internal unbiased estimate of the generalization error as the forest building progresses.

Bagging of Decision Trees

Training random forests applies bagging learning to build tree learners.

Given a training set $X = x_1, ..., x_n$ with labels $Y = y_1, ..., y_n$, bagging repeatedly (N times) selects a random sample with replacement of the training set and fits trees on these samples.

After training, predictions for unseen samples x' can be made by averaging the predictions or by making the majority vote in the case of classification trees.



Source: https://www.analyticsvidhya.com/blog/2020/05/de -tree-vs-random-forest-algorithm/

Discussion on Random Forest

Strength

It is more robust.

It is faster to train (no reweighting, each split is on a small subset of data and feature).

It can handle missing/partial data.

It is easier to extend to online version.

Weakness

The feature selection process is not explicit.

Feature fusion is also less obvious.

Weak performance on small size of training data.

DLI Accelerated Data Science Teaching Kit

Thank You