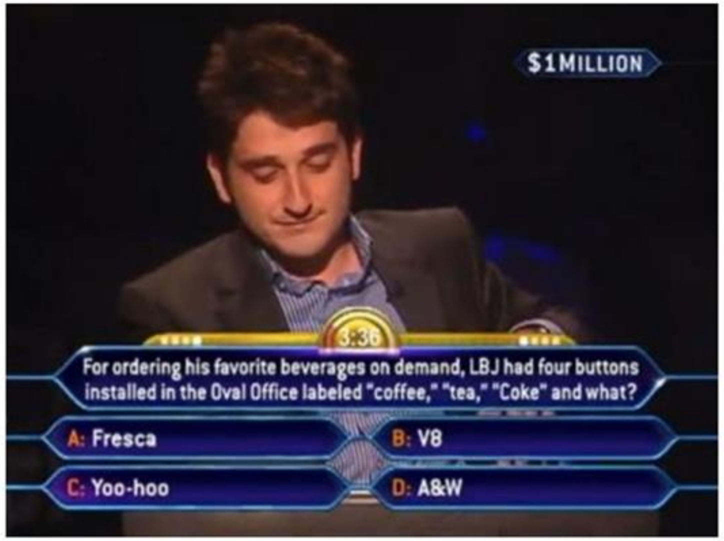
ENSEMBLE LEARNING

ENSEMBLE METHODS

In the world of Data Mining, Ensemble learning techniques attempt to make the performance of the predictive models better by improving their accuracy.

Ensemble Learning is a process using which multiple machine learning models (such as classifiers) are strategically constructed to solve a particular problem.

“Who wants to be a millionaire?”



For ordering his favorite beverages on demand, LBJ had four buttons installed in the Oval Office labeled "coffee," "tea," "Coke" and what?

A: Fresca

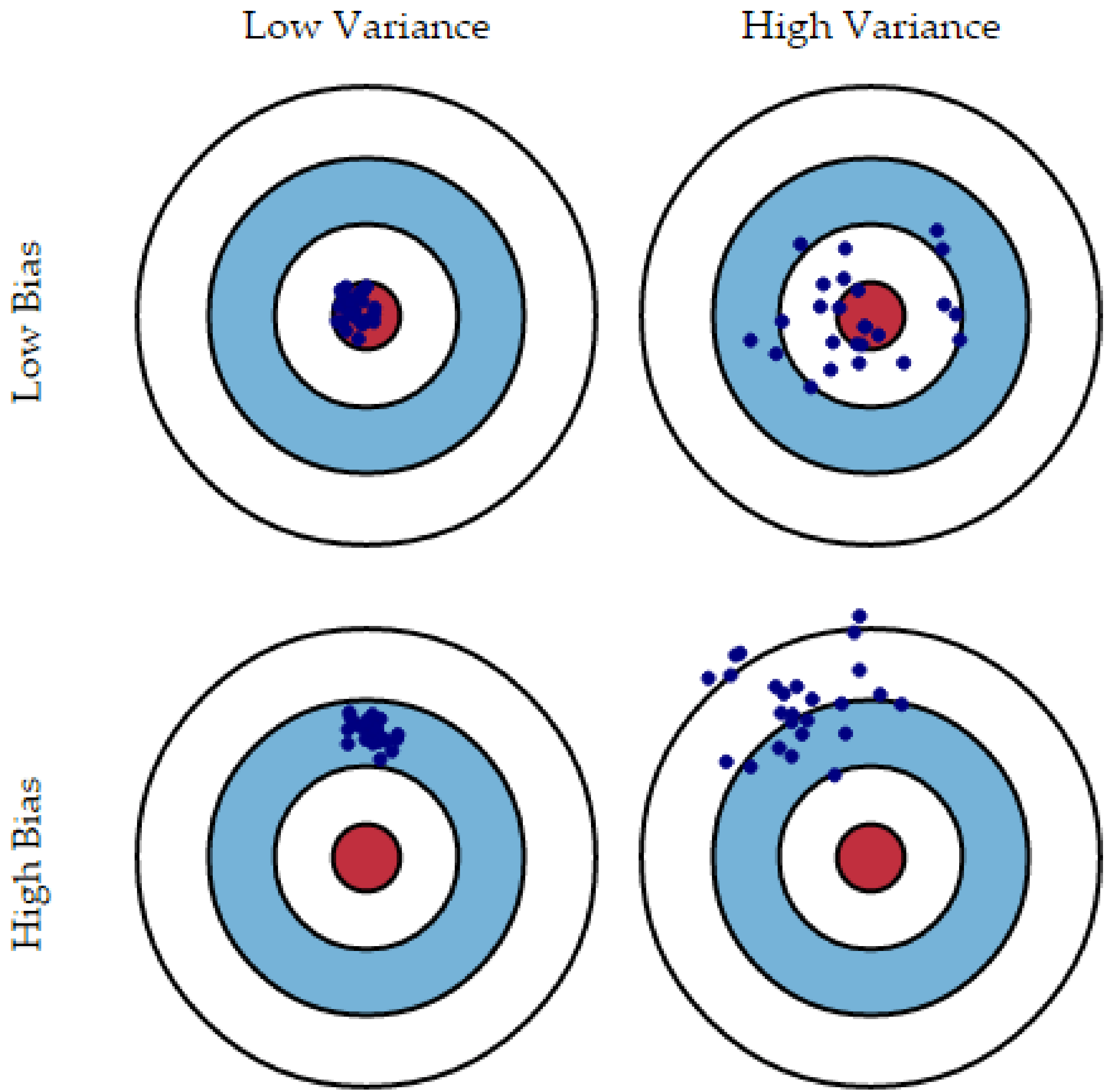
B: V8

C: Yoo-hoo

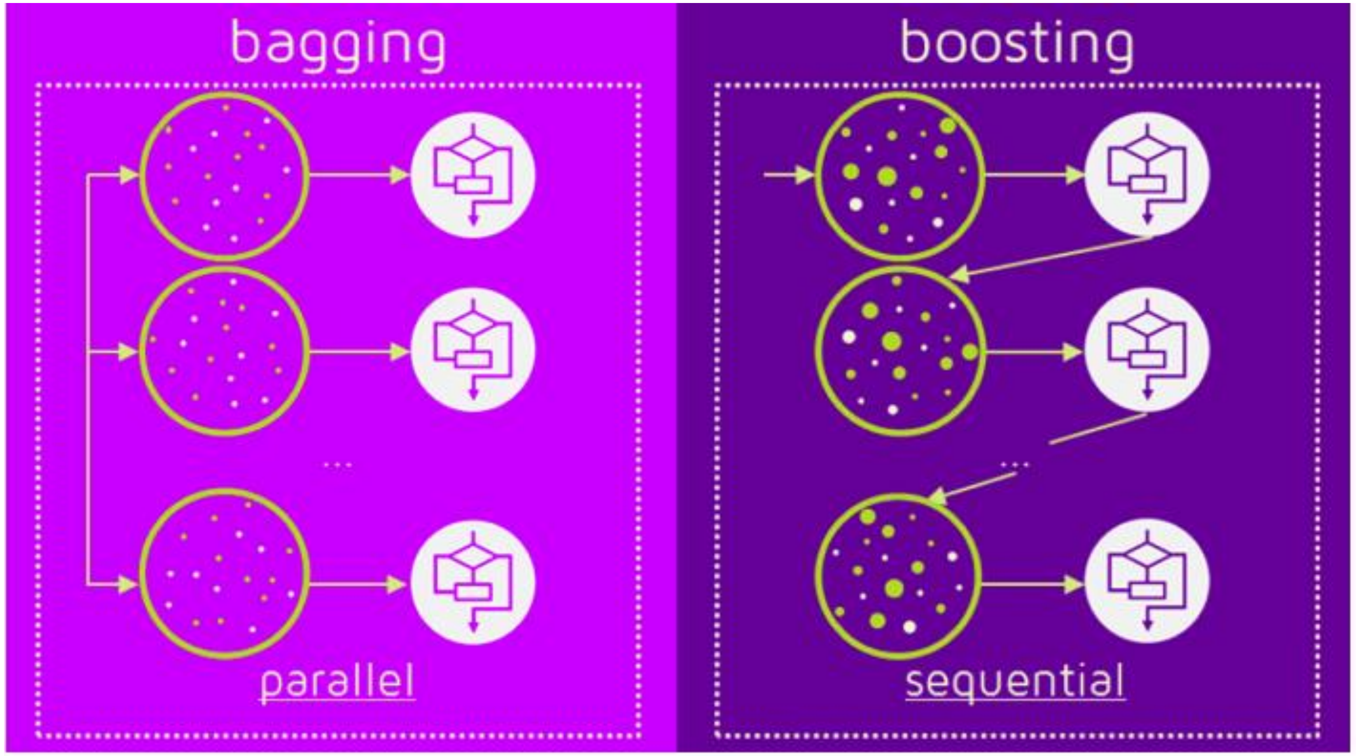
D: A&W

Various options for getting help:

MODEL ERROR AND REDUCING THIS ERROR WITH ENSEMBLES



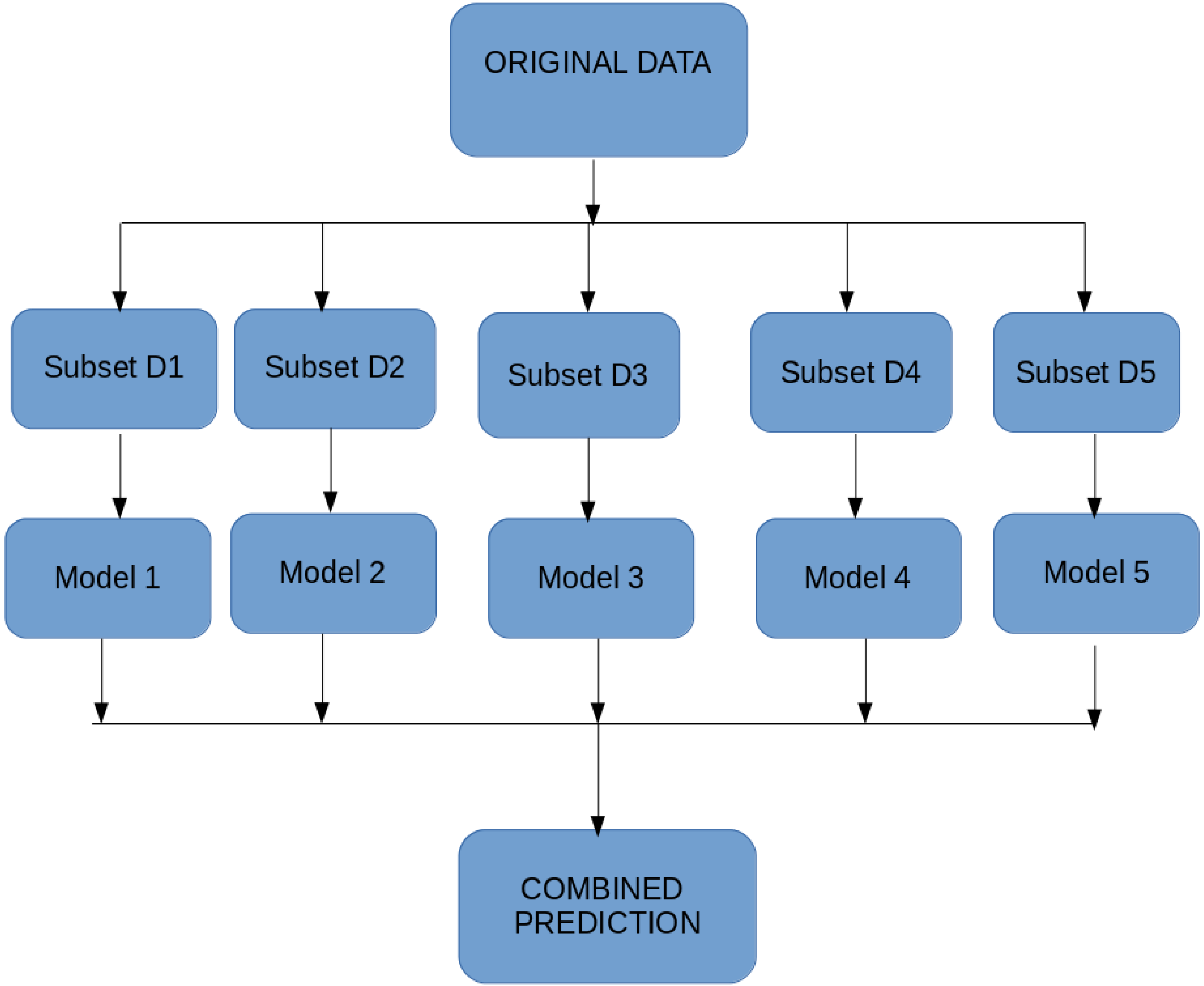
DIFFERENT TYPES OF ENSEMBLE LEARNING METHODS



BAGGING

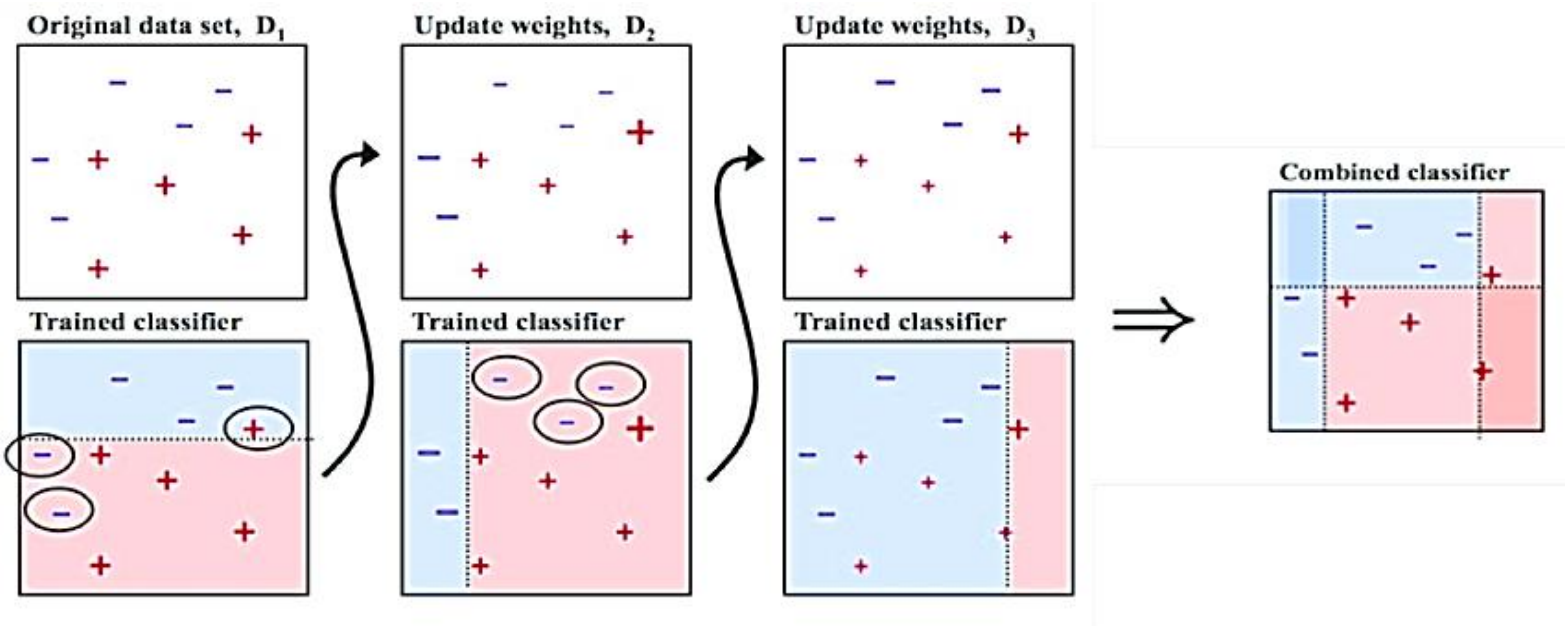
- The idea behind **bagging** is combining the results of multiple models (for instance, all decision trees) to get a generalized result.
- Here's a question: If you create all the models on the same set of data and combine it, will it be useful? There is a high chance that these models will give the same result since they are getting the same input.
- So how can we solve this problem? One of the techniques is bootstrapping.

- **Bootstrapping** is a sampling technique in which we create subsets of observations from the original dataset, with replacement. The size of the subsets is the same as the size of the original set.
- **Bagging (or Bootstrap Aggregating)** technique uses these subsets (bags) to get a fair idea of the distribution (complete set). The size of subsets created for bagging may be less than the original set.


ORIGINAL DATA
Subset D1
Subset D2
Subset D3
Subset D4
Subset D5
Model 1
Model 2
Model 3
Model 4
Model 5
COMBINED PREDICTION

BOOSTING

- Here’s another question for you: If a data point is incorrectly predicted by the first model, and then the next (probably all models), will combining the predictions provide better results? Such situations are taken care of by boosting.
- **Boosting** is a sequential process, where each subsequent model attempts to correct the errors of the previous model. The succeeding models are dependent on the previous model.


Original data set, D_1
Update weights, D_2
Update weights, D_3
Trained classifier
Trained classifier
Trained classifier
Combined classifier

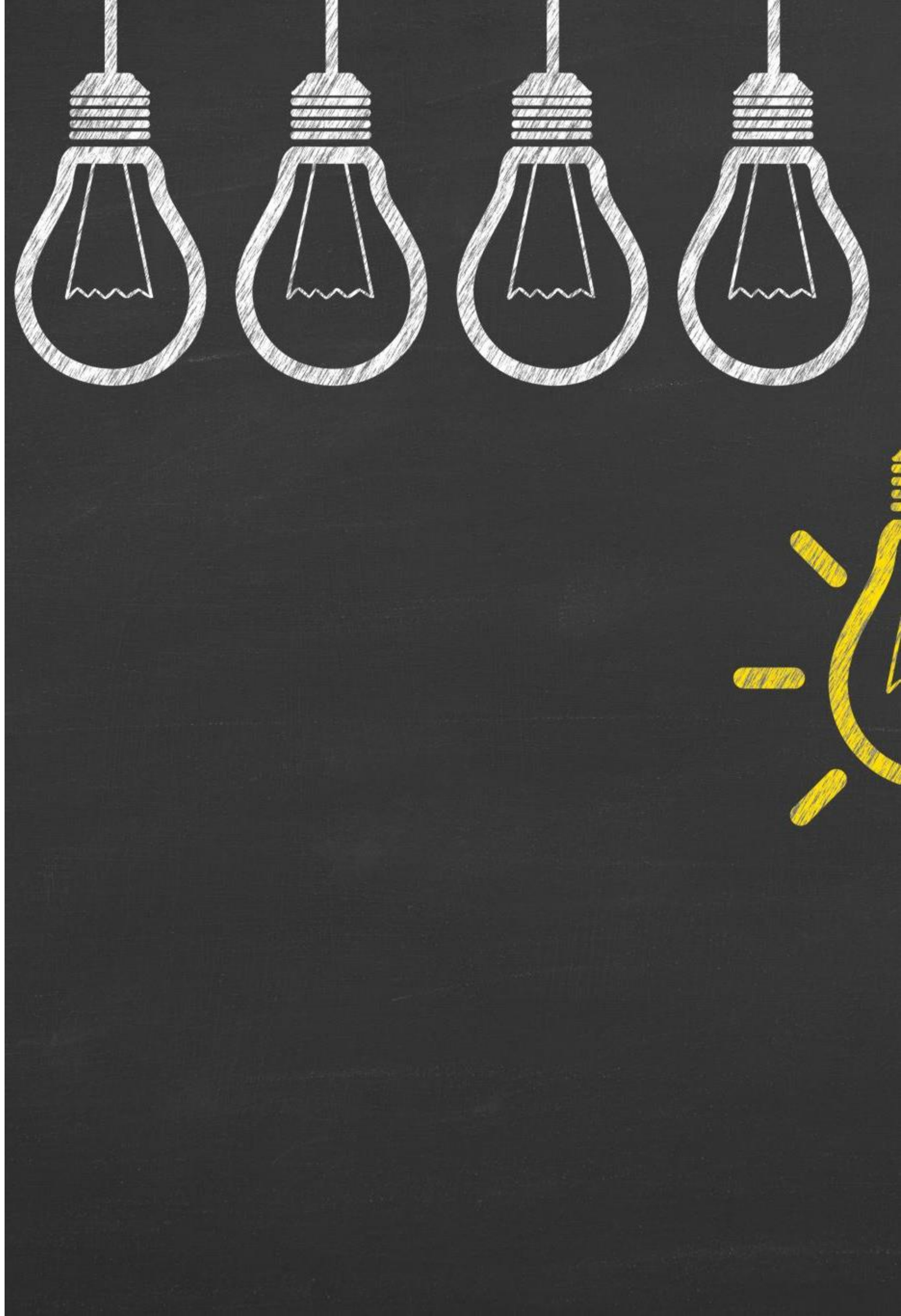
BOOSTING ALGORITHM

- **Boosting** algorithm combines a number of weak learners to form a strong learner. The individual models would not perform well on the entire dataset, but they work well for some part of the dataset. Thus, each model actually boosts the performance of the ensemble.

WHICH ONE IS BETTER?

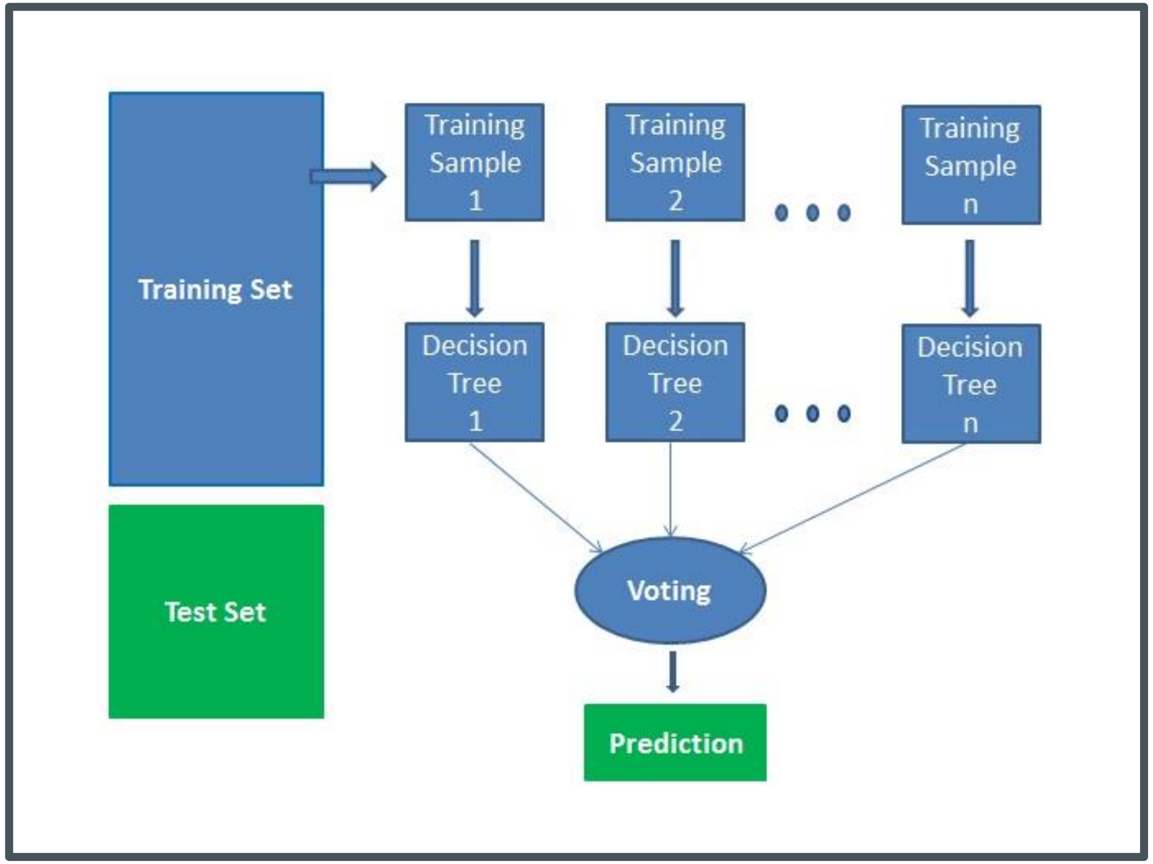
- **Bagging** to decrease the model's variance.
- **Boosting** to decreasing the model's bias.

- There's not an outright winner; it depends on the data, the simulation and the circumstances.



RANDOM FORESTS CLASSIFIERS

RANDOM FORESTS



- **Random forests** is a supervised learning algorithm. It can be used both for classification and regression. It is also the most flexible and easy to use algorithm.
- A forest is comprised of trees. It is said that the more trees it has, the more robust a forest is.
- Random forests creates decision trees on randomly selected data samples, gets prediction from each tree and selects the best solution by means of voting.
- It also provides a pretty good indicator of the feature importance.
- It is a type of Bagging Ensemble Learning.

PROS

- Random forests is considered as a highly accurate and robust method because of the number of decision trees participating in the process.
- It does not suffer from the overfitting problem. The main reason is that it takes the average of all the predictions, which cancels out the biases.
- You can get the relative feature importance, which helps in selecting the most contributing features for the classifier.

CONS

- Random forests is slow in generating predictions because it has multiple decision trees. Whenever it makes a prediction, all the trees in the forest have to make a prediction for the same given input and then perform voting on it. This whole process is time-consuming.
- The model is difficult to interpret compared to a decision tree, where you can easily make a decision by following the path in the tree.

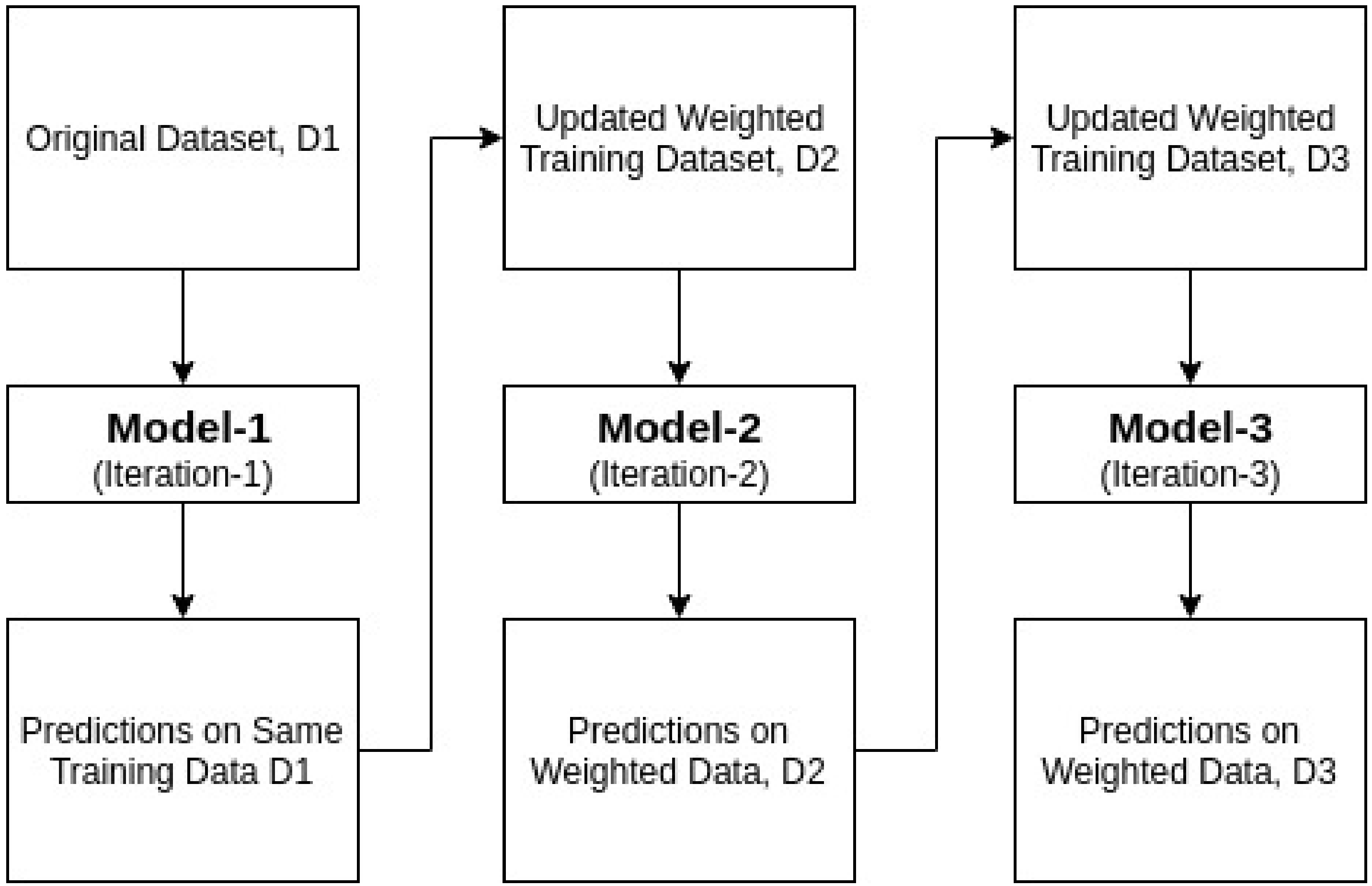
BOOSTING ALGORITHMS

- Boosting algorithms such as **AdaBoost, Gradient Boosting, and XGBoost** are widely used to win the data science competitions.

ADABOOST CLASSIFIER

- **Ada-boost** or Adaptive Boosting is one of ensemble boosting classifier proposed by Yoav Freund and Robert Schapire in 1996.
- It combines multiple classifiers to increase the accuracy of classifiers. Ada-Boost is an iterative ensemble method.
- The basic concept behind Ada-boost is to set the weights of classifiers and training the data sample in each iteration such that it ensures the accurate predictions of unusual observations.

ADABOOST ALGORITHM



Original Dataset, D1

Model-1
(Iteration-1)

Predictions on Same Training Data D1

Updated Weighted Training Dataset, D2

Model-2
(Iteration-2)

Predictions on Weighted Data, D2

Updated Weighted Training Dataset, D3

Model-3
(Iteration-3)

Predictions on Weighted Data, D3

ADABOOST AND RANDOM FOREST IN PYTHON. (DEMO)