Model Evaluation Strategies

Why do we need to evaluate model?

- To estimate the predictive performance of our model on future (unseen) data.
- To fine-tune the model to increase its predictive performance
- To select the right machine learning algorithm that is best-suited for the problem at hand

Machine Learning(Model) Evaluation

- How do you evaluated machine learned model?

- Goal: Select the model that best performs on unseen(future) data i.e., best generalization capability.

Resubstitution Approach

- Use entire train data for learning as well as evaluation
- The accuracy returned by model on same training set used for learning is called as Resubstitutionaccuracy
- Does this approach makes sense???

Issues with Resubstitution Approach

- Model may not have enough data to fully learn the concept (but on training data we don't know this)
- For noisy data, the model may overfit the training data

Hence, Resub error is always too optimistic i.e., under-estimates true error of model

Resampling Approach

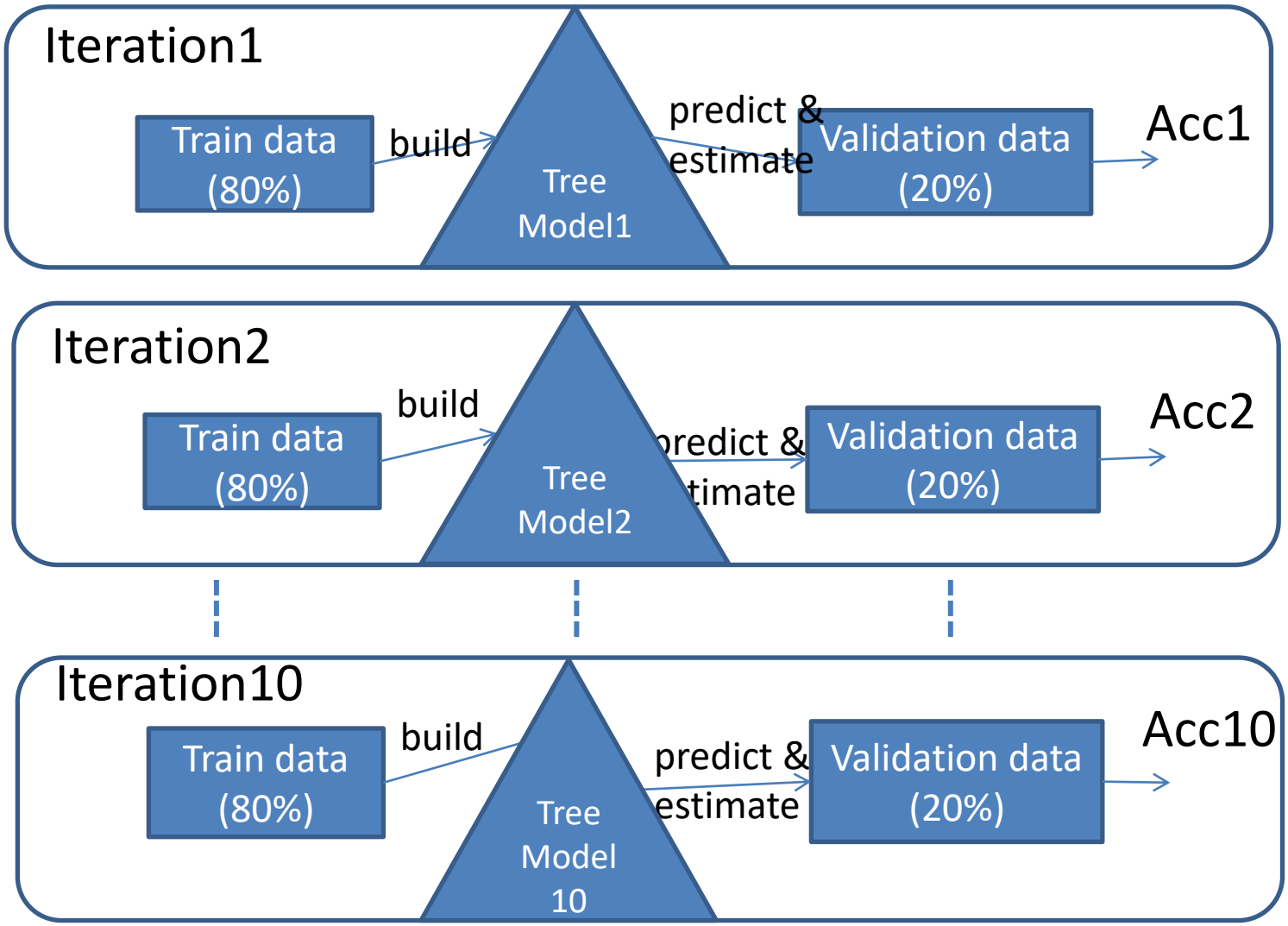
- Hold some observations from train-set, called as validation set
- Use a validation set to estimate how well the model perform on new unseen data(out of sample)
- Resampling methods try to "inject variation" in the system to approximate the model's performance on future samples.

Resampling Methods

- Repeated Holdout*
- Cross Validation*
- Bootstrapping

*Stratified resampling is preferred

Repeated holdout



Compute Avg Acc

Repeated Holdout Idea

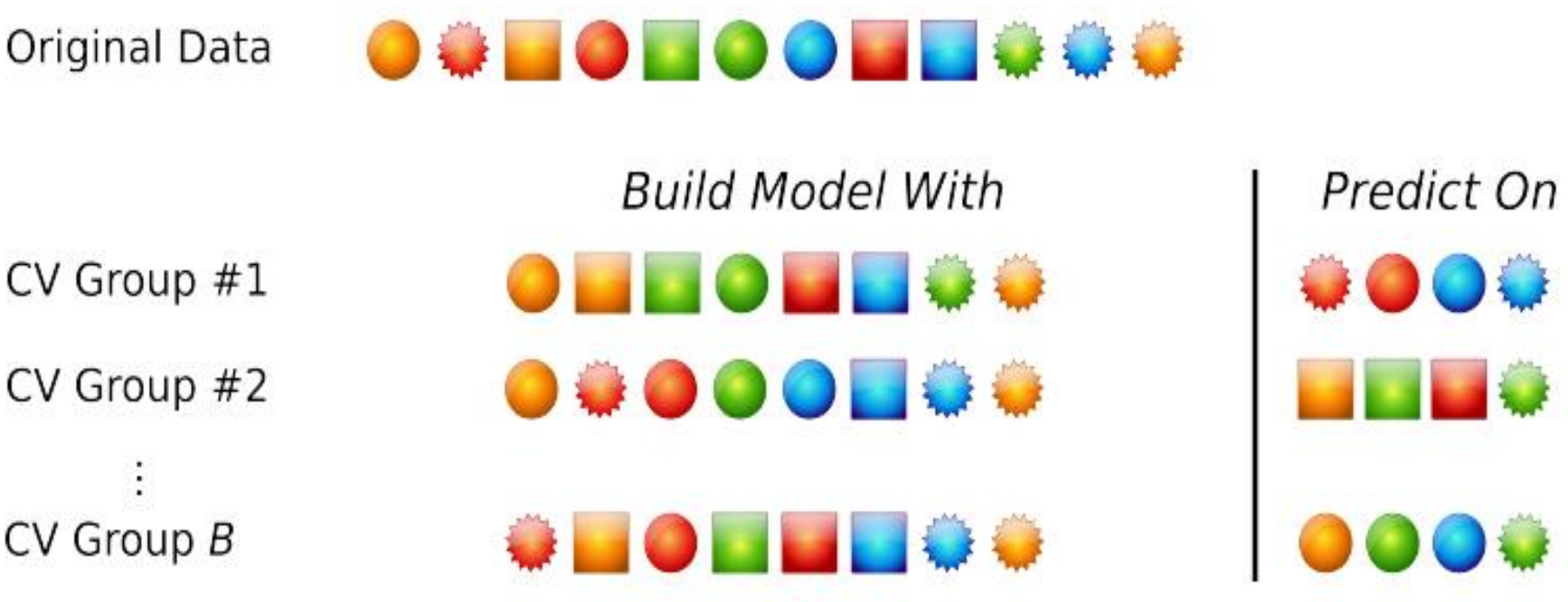
- n = total number of training data points
  k = number of repetitions

- for i = 1:k
  - Randomly hold 30% of data for validation
  - train on remaining 70% of data
  - Acc(i) = accuracy on 30% held data

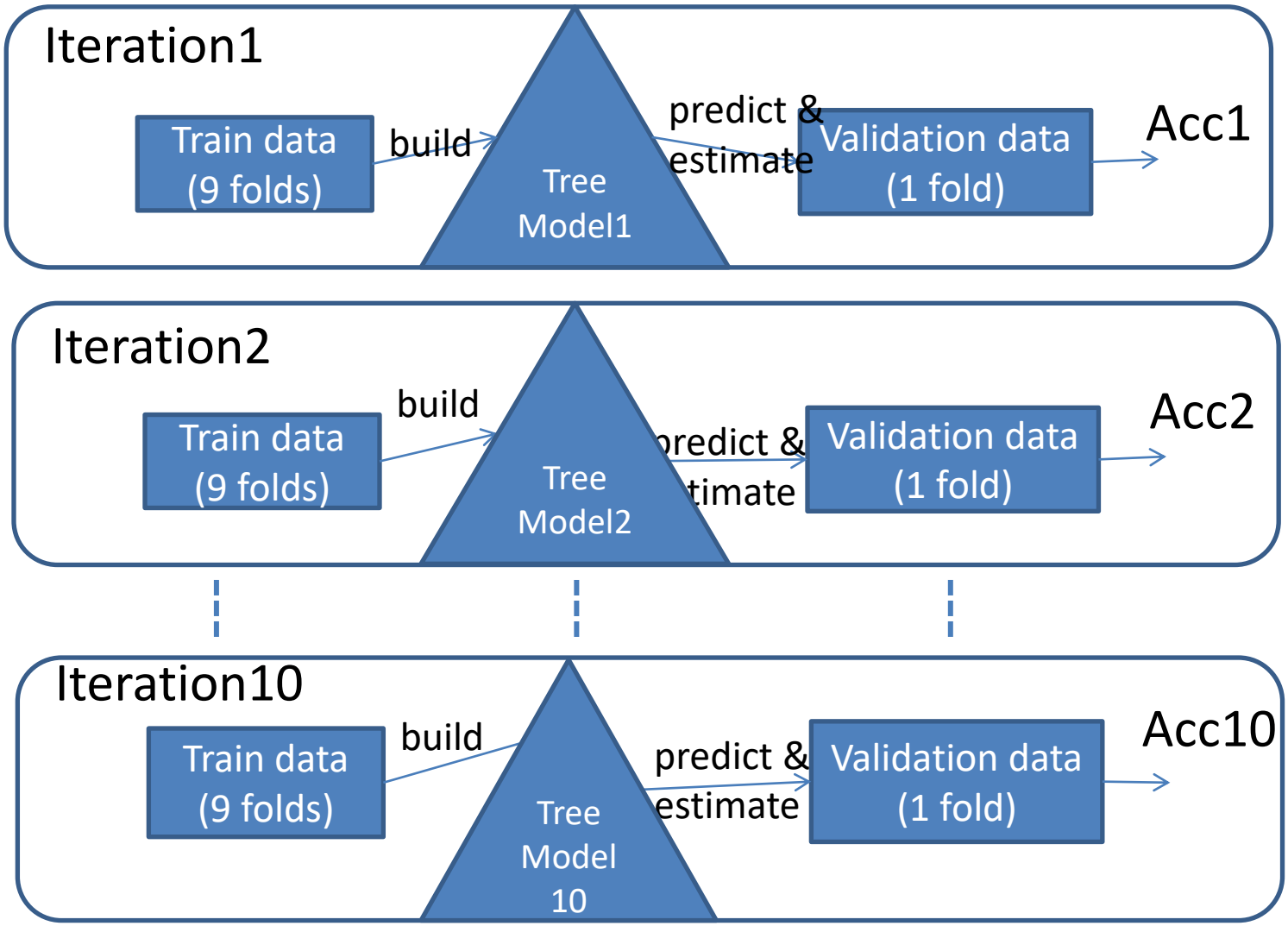
- Repeated holdout Accuracy = $1/k \ \Sigma_i$ Acc(i)

- Common value for k is 10

Repeated Holdout Illustrated



K-fold cross validation



Compute Avg Acc

K-Fold CV Idea

- n = total number of training data points
  k = number of folds

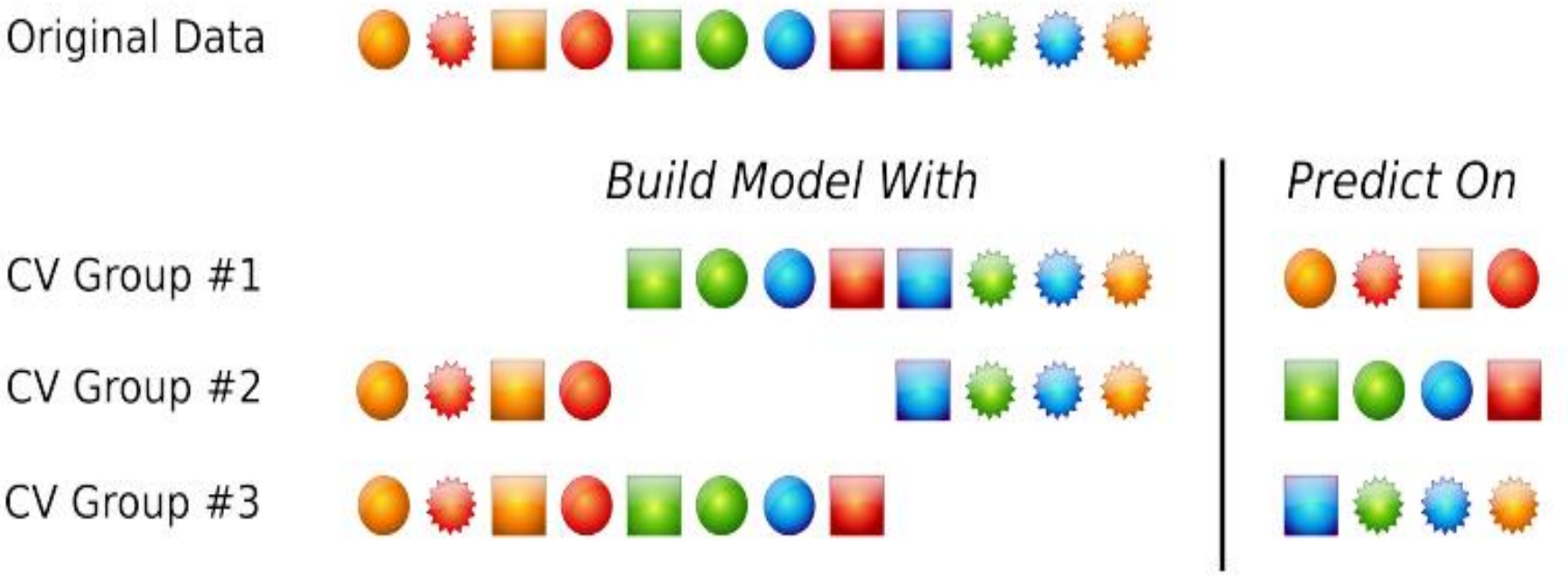
- Randomly partition our full data set into k disjoint subsets (each roughly of size n/k)

- for i = 1:k
  - train on all folds of data except i^{th}
  - Acc(i) = accuracy on i^{th} fold

- Cross-Validation Accuracy = $1/k \ \Sigma_i$ Acc(i)

- Common values for k are 5 and 10 and called as "leave-one-out" when k = n

K-Fold CV Illustrated



Bootstrapping Idea

Bootstrap sample: sample with replacement from a dataset

The probability that a given example is not selected for a bootstrap sample of size n:

$$(1 - 1/n)^n$$

This has a limit as n goes to infinity: 1/e = 0.368

Conclusion: each bootstrap sample is likely to leave out about a third of the examples.

Bootstrapping Illustrated