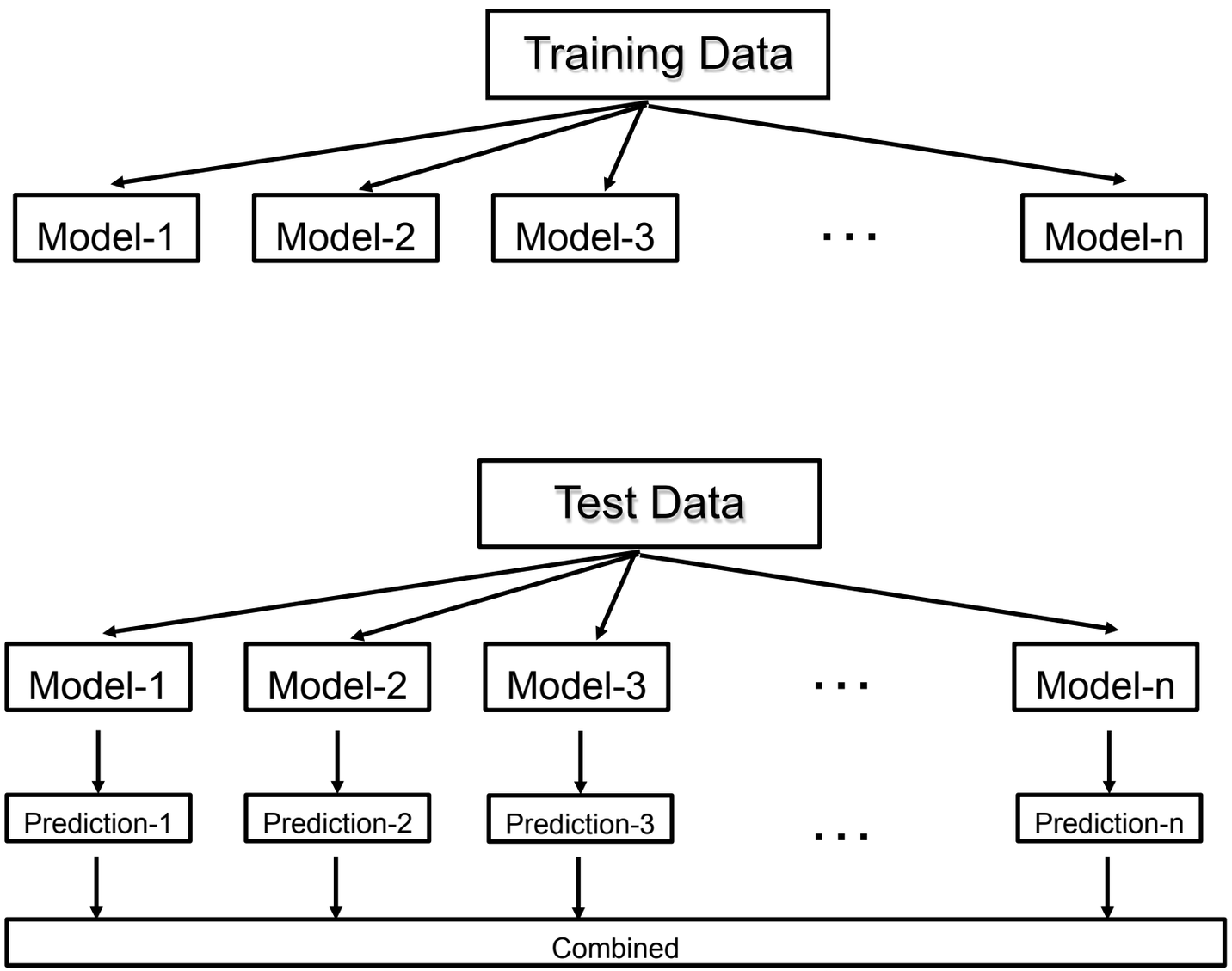
Ensemble Methods

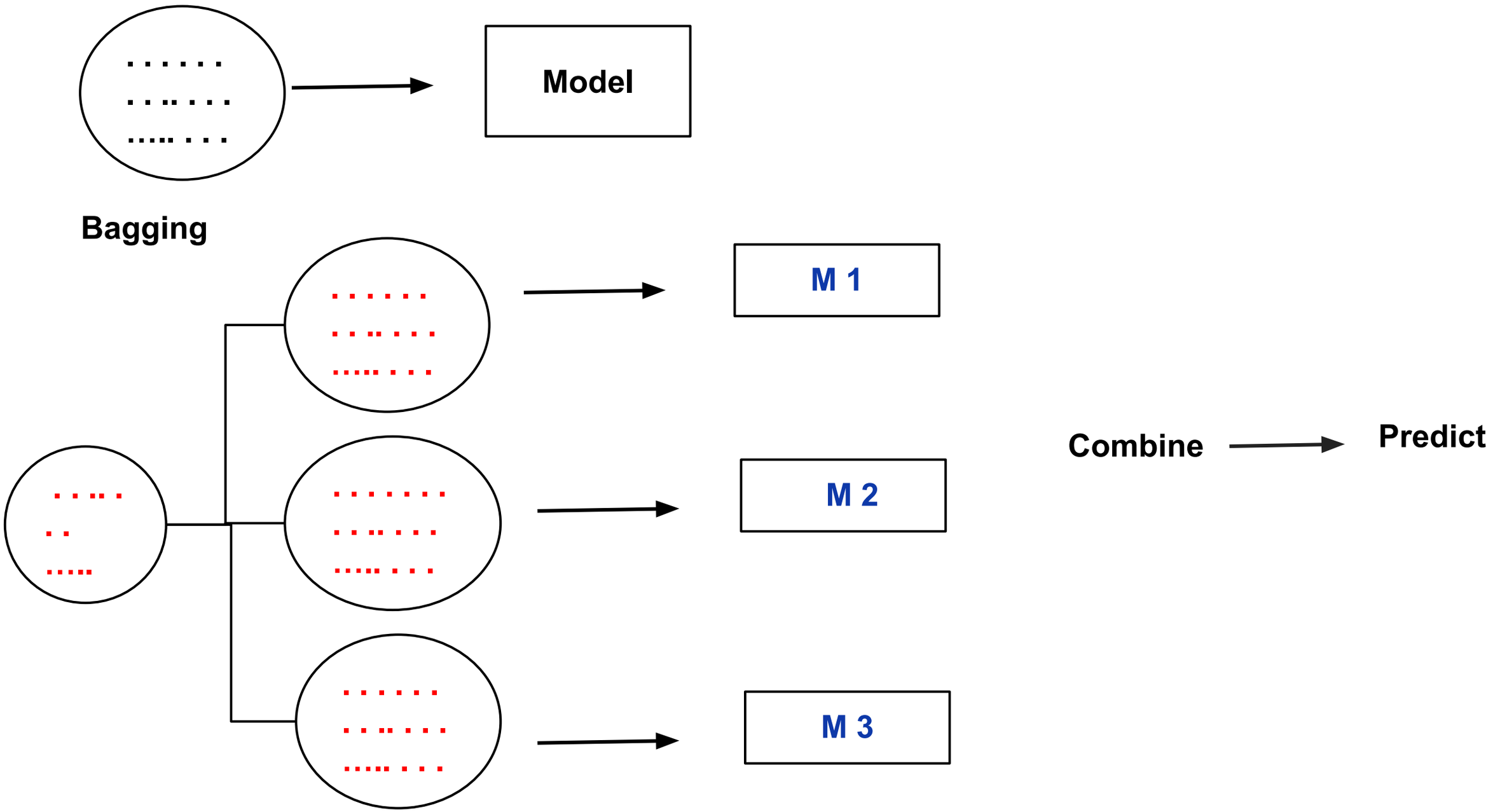
- Ensembles are machine learning methods for combining predictions from multiple separate models.
- The central motivation is rooted under the belief that a committee of experts working together can perform better than a single expert.



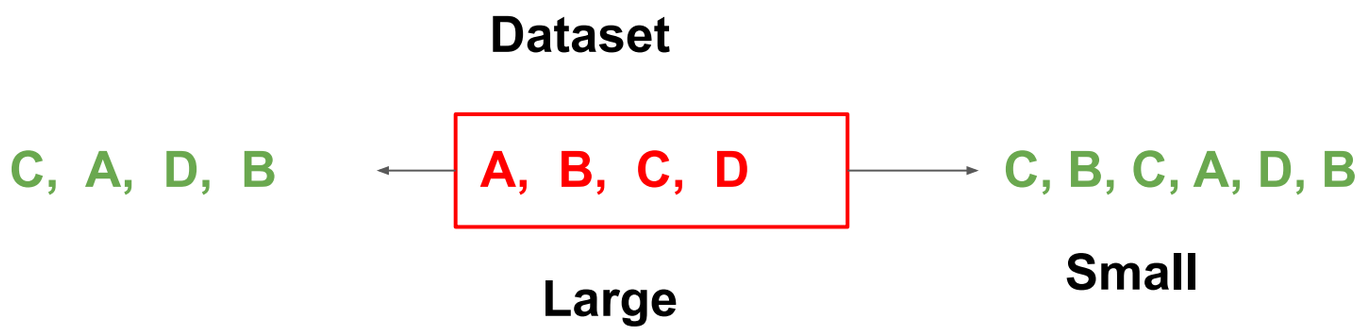
Ensemble Methods

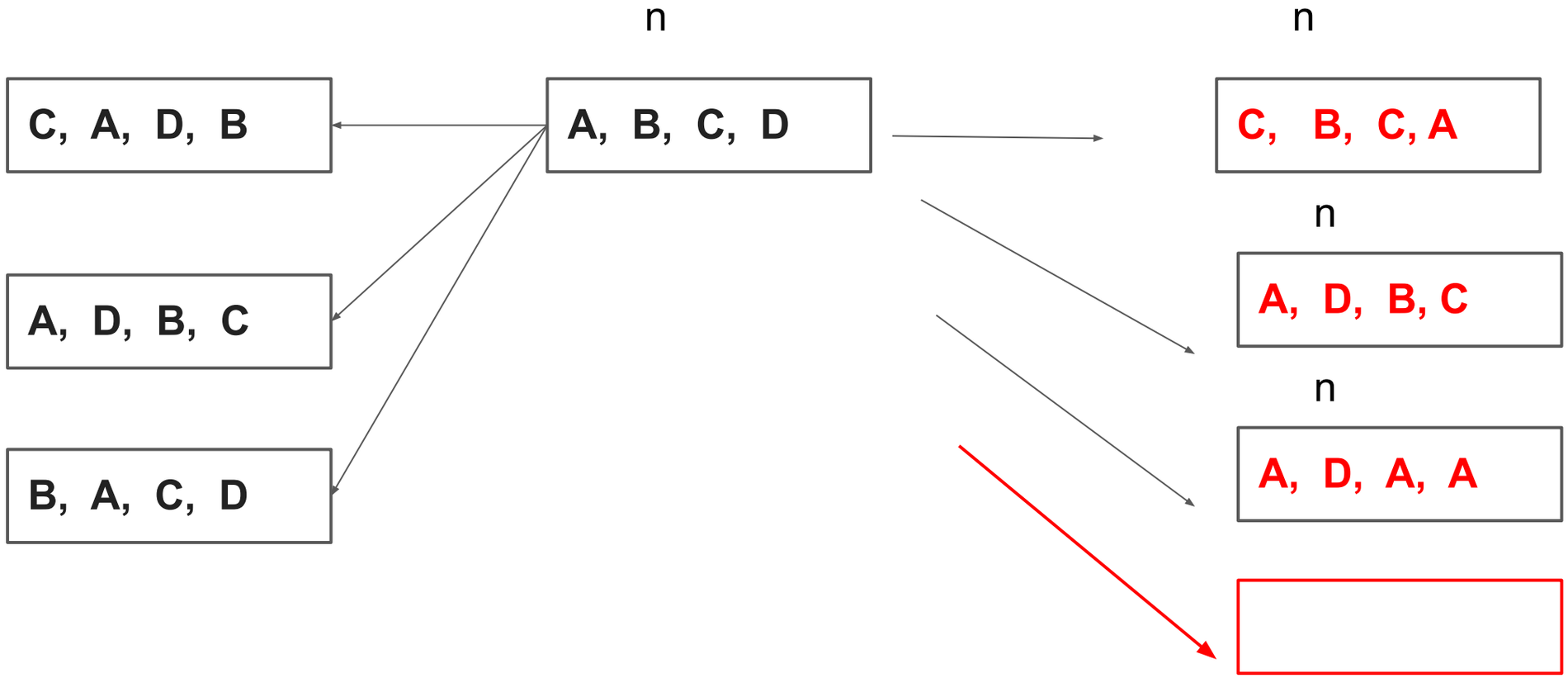
Truth	90% M1	90% M2	90% M3	90%	90% M10	
Y	✓	✓	✓	✓	X	✓
Y	X	X	X	X	X	X	✓
N	✓	✓	✓	X	X	X	✓
..	✓	X	✓	X	X	✓	✓
...	✓	✓	✓	✓	.....	.....	✓
...	✓	✓	✓	✓	✓	.....	...
Y	✓	✓	✓	✓		✓	...
N	✓	✓	✓	✓	✓	✓	...

Bagging

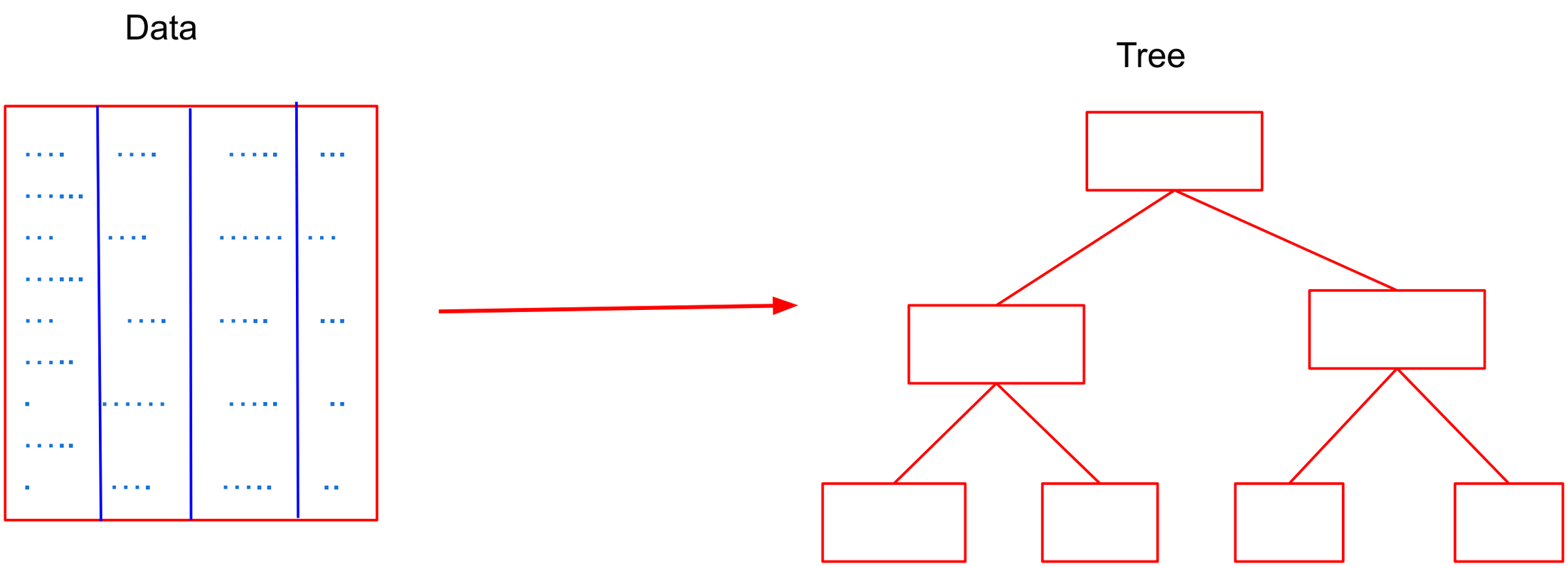


Why Sampling with Replacement?



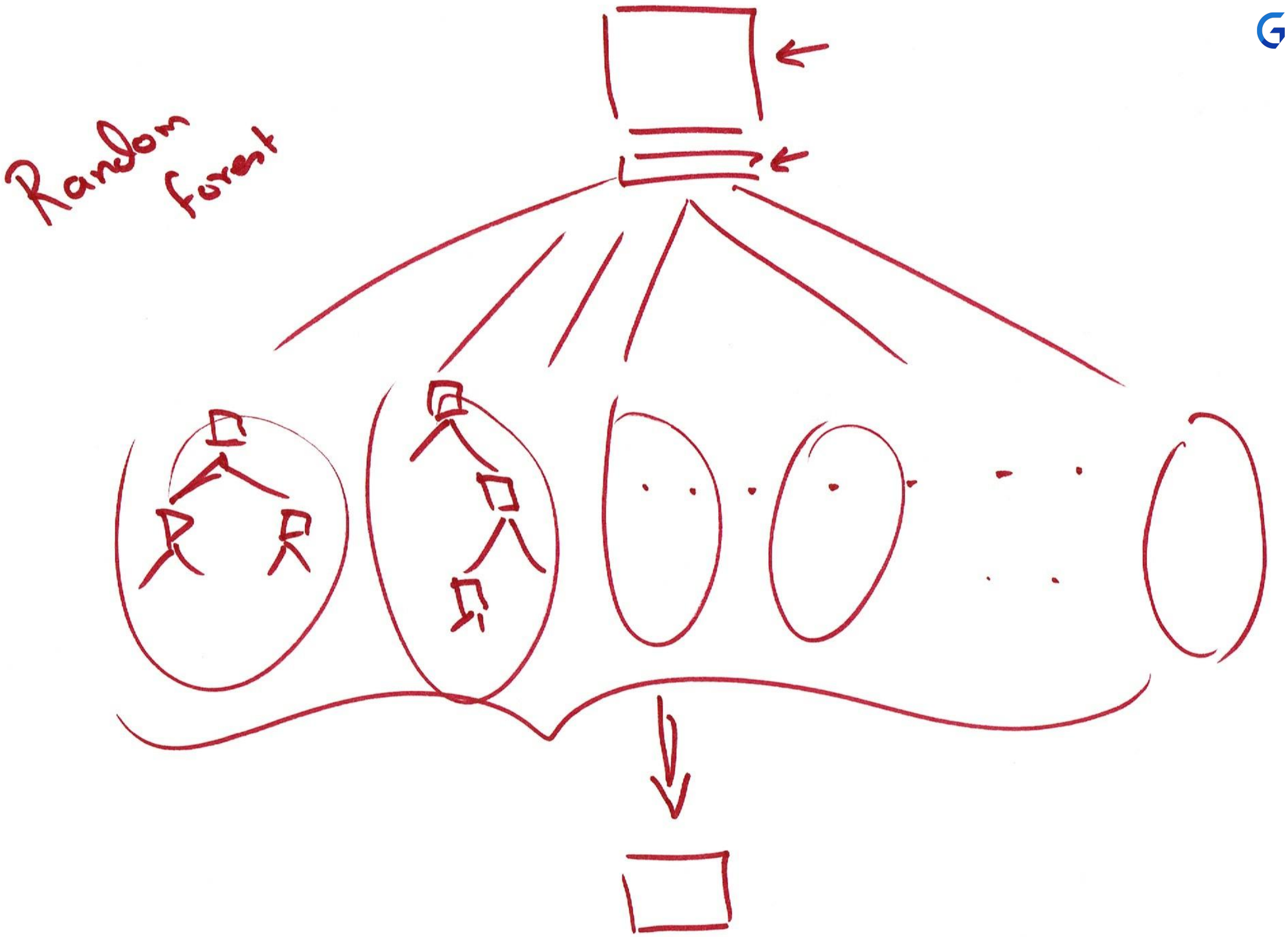


Random Forest



Tree to a Forest

- Decision trees are very sensitive to even small changes in the data - usually called unstable.
- Can we get a whole bunch of decision trees to work together to yield a better and more robust prediction?
- Then for prediction we could use the mean for regression trees and mode for classification trees
- While individual trees are tend to over-fit training data, averaging corrects this.



The General Ideas

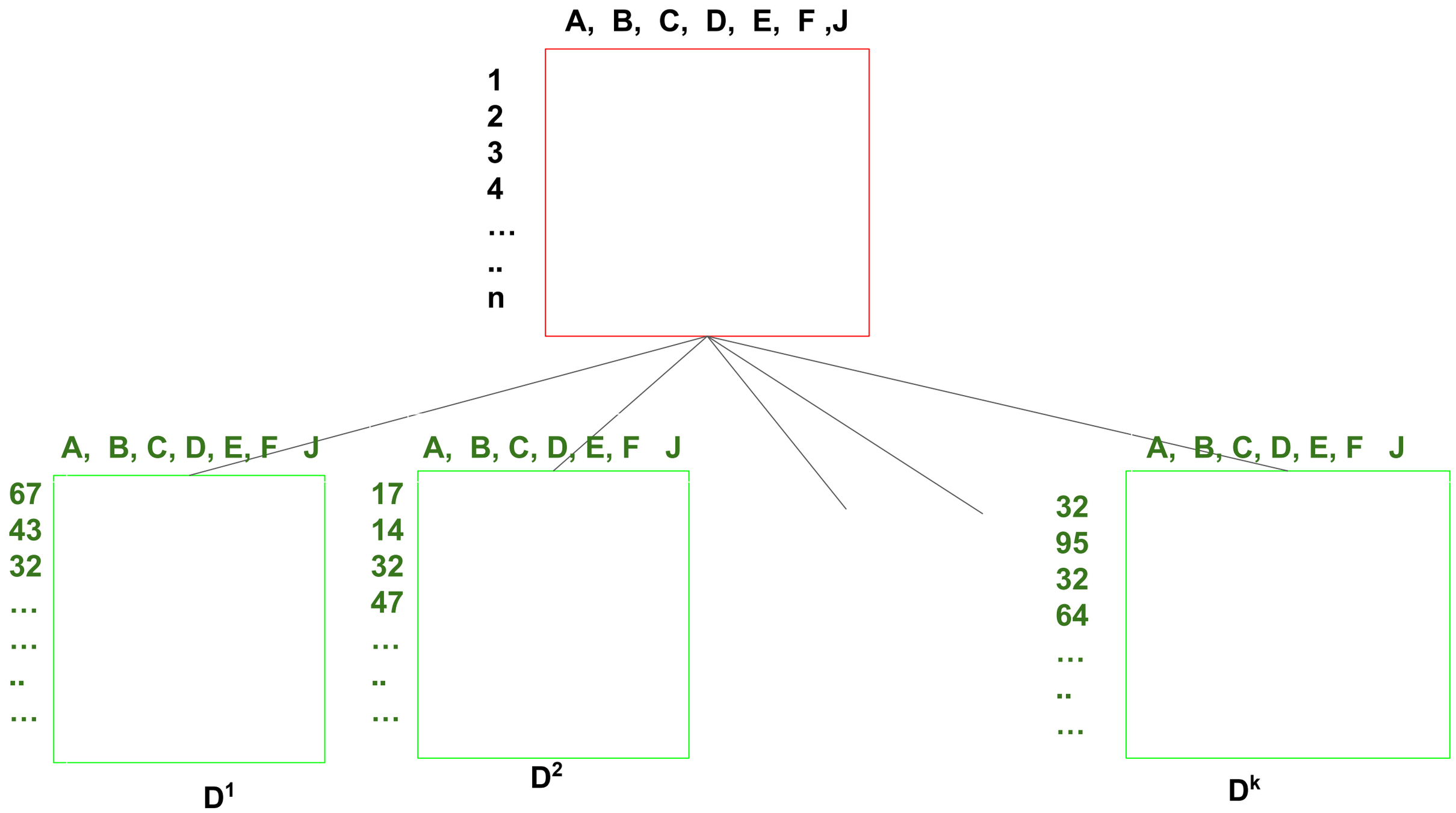
- The general procedure of using multiple models (trees, in this case) to obtain better predictive performance is called ensemble learning.

- Bootstrap aggregating. also called bagging:

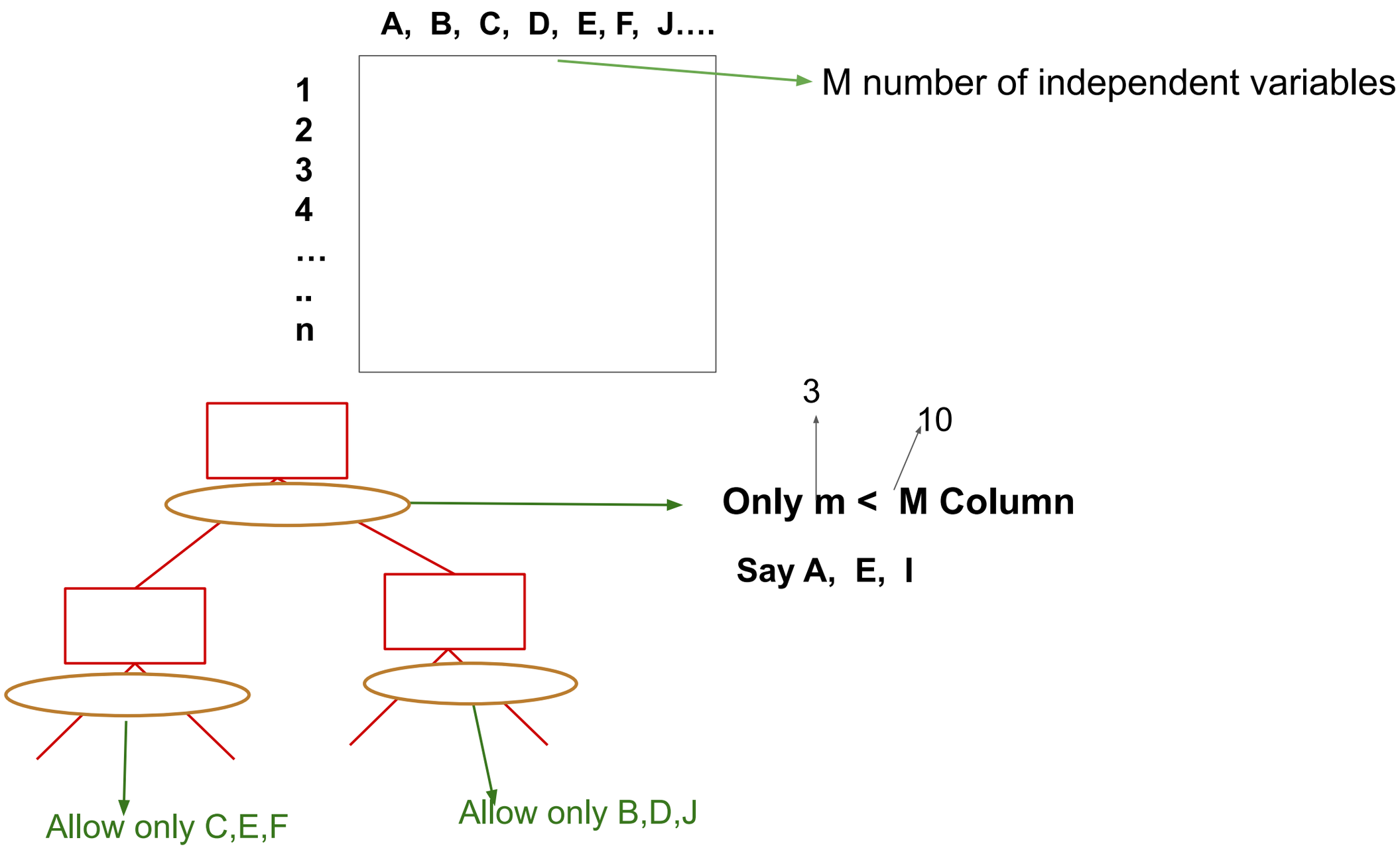
    - Generate new training subsets of the original, each of the same size (usually the size of the data) by sampling with replacement.

    - By sampling with replacement, some observations may be repeated in each subset.

Random Forest



Random Forest



Random forests

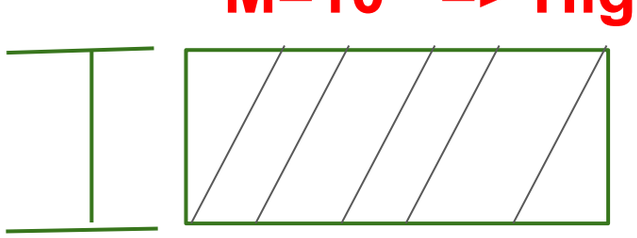
- Random Sampling with replacement
- For each subset build a decision tree. However, only use **m** randomly pick independent variables for each node's branching possibilities.
- Do not prune
- While predicting:
  - Use each tree to make individual predictions
  - Combine predictions using voting:
    - Means for regression
    - Modes for classification

Random Forest

Say M = 10 => A B C D ◯ ◯ ◯ J

1

M=10 => High Tree Correlation



Good

2

M=2 => Your trees are weak