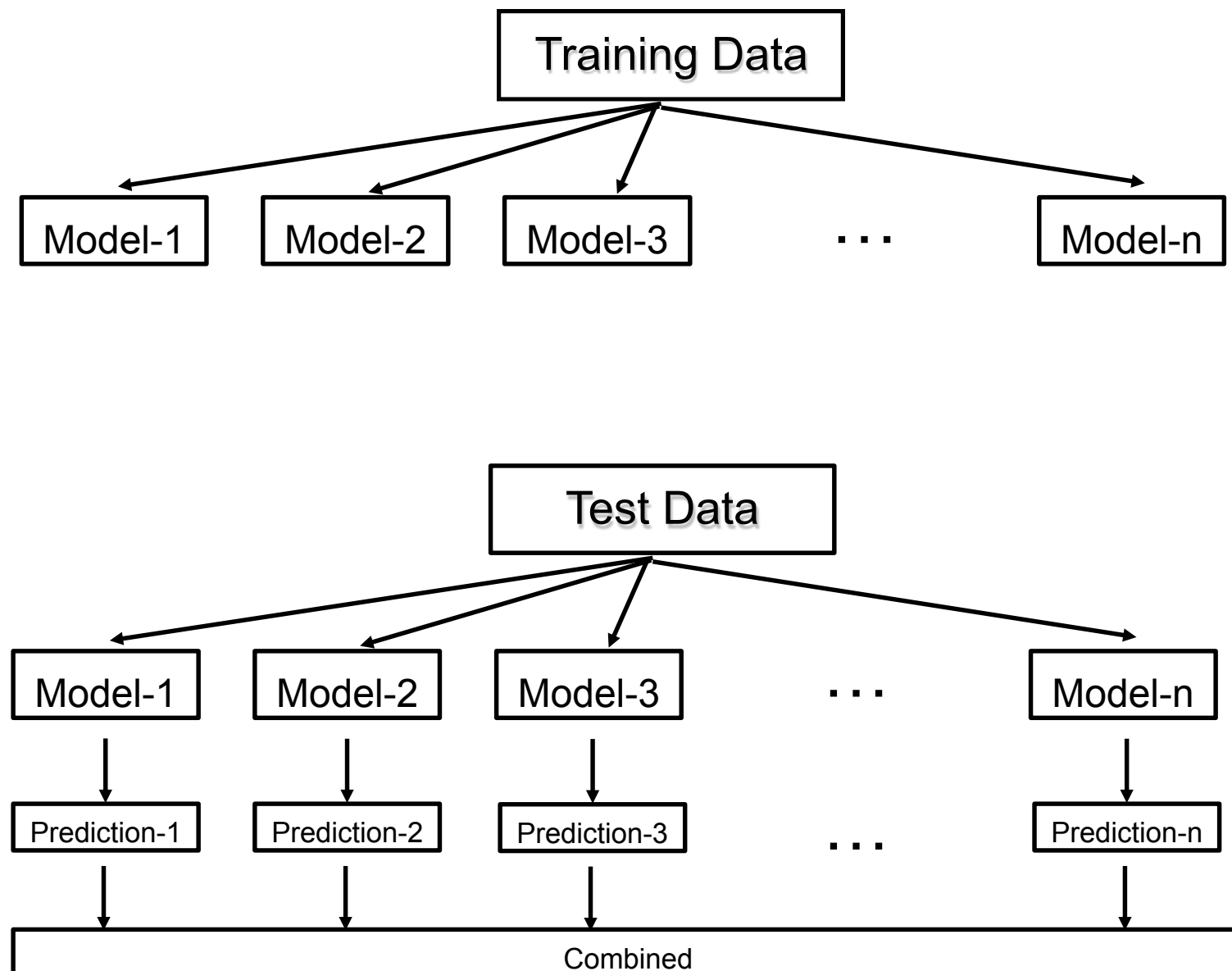# Ensemble Methods

- 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.
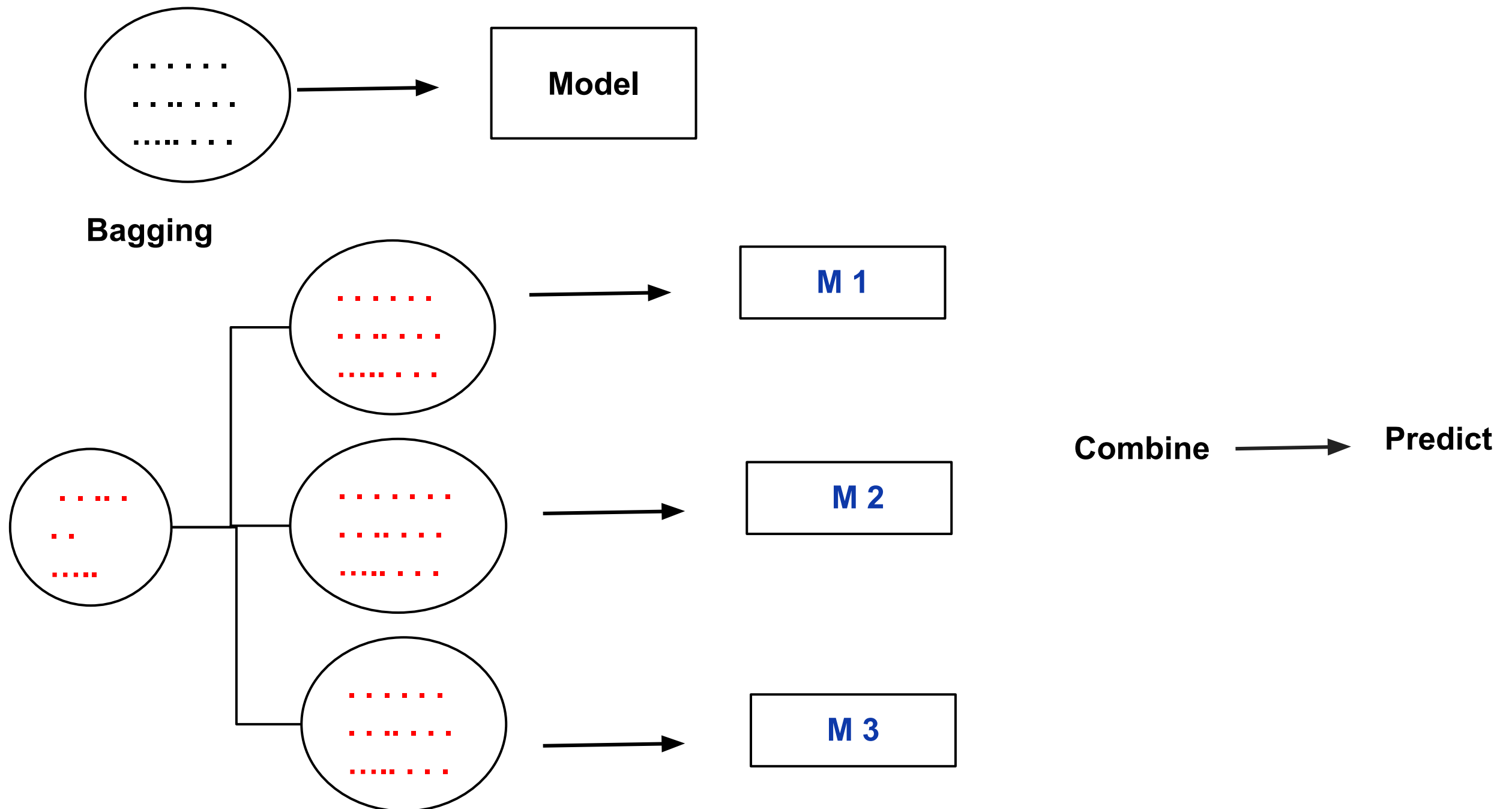
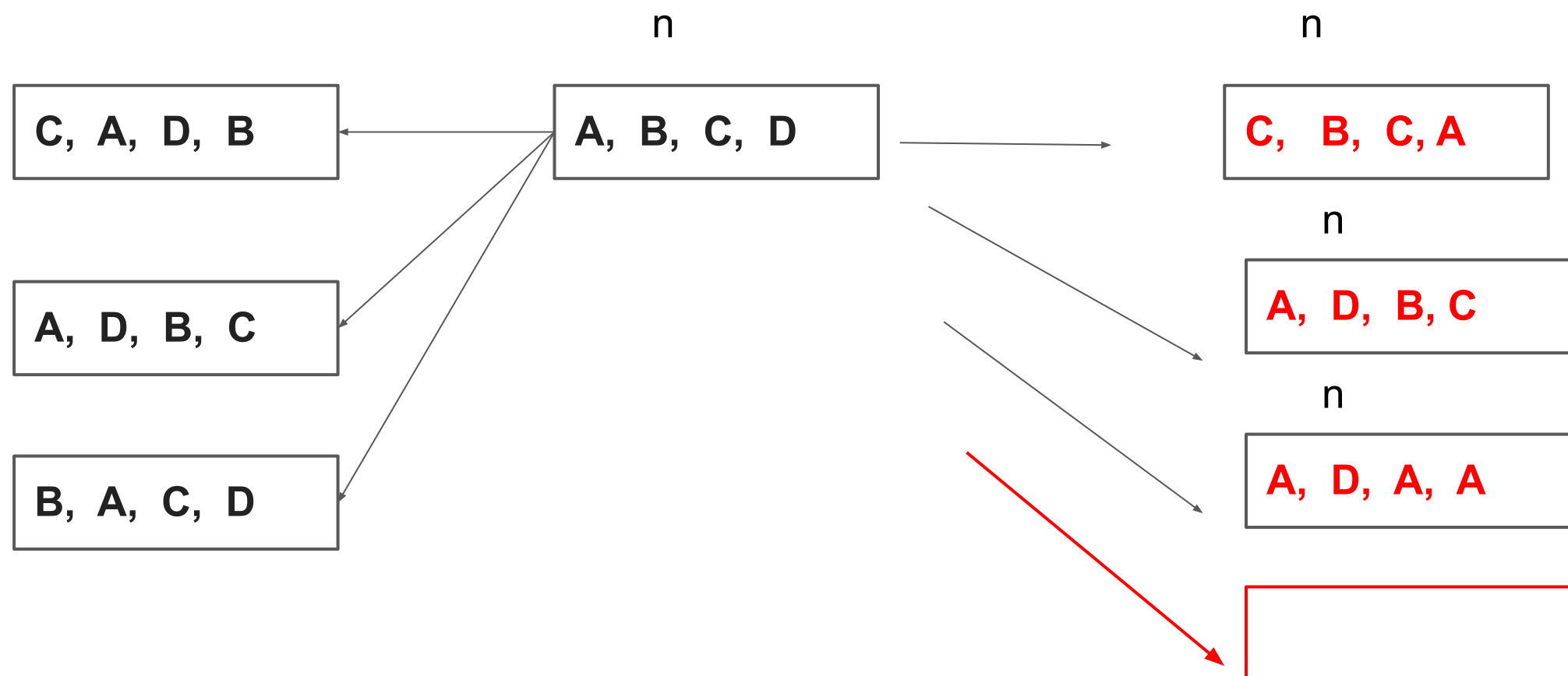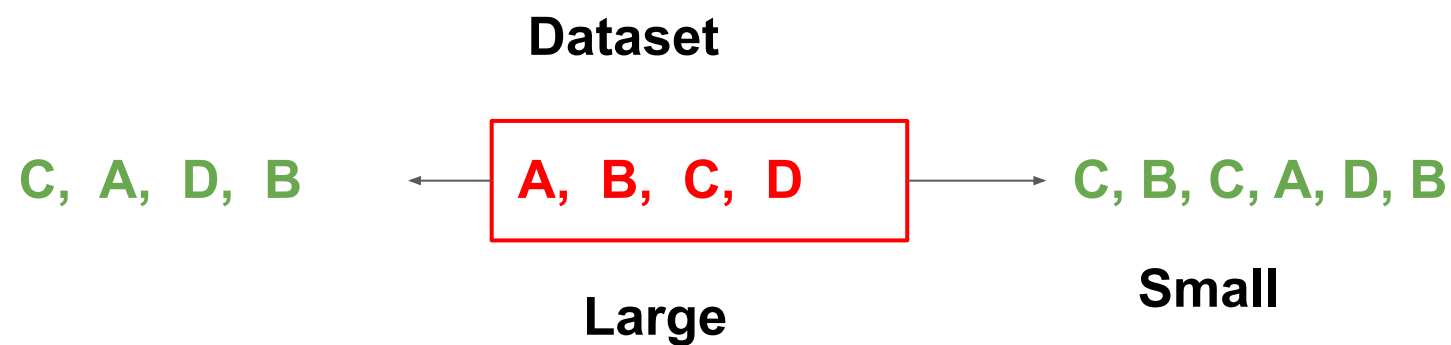# Ensemble Methods

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

Bagging

Combine → Predict

Why Sampling with Replacement?

Dataset

C, A, D, B ← A, B, C, D → C, B, C, A, D, B

Large   **Small**

n   n

| C, A, D, B | ← A, B, C, D → | C, B, C, A |

A, D, B, C

n

A, D, B, C

n

B, A, C, D

A, D, A, A

Random Forest
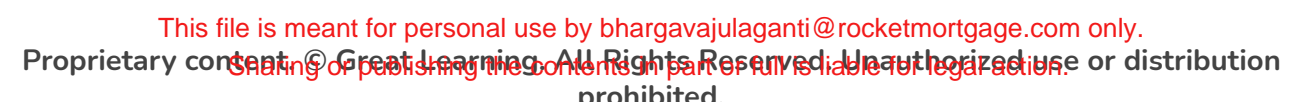

Data

Tree

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.
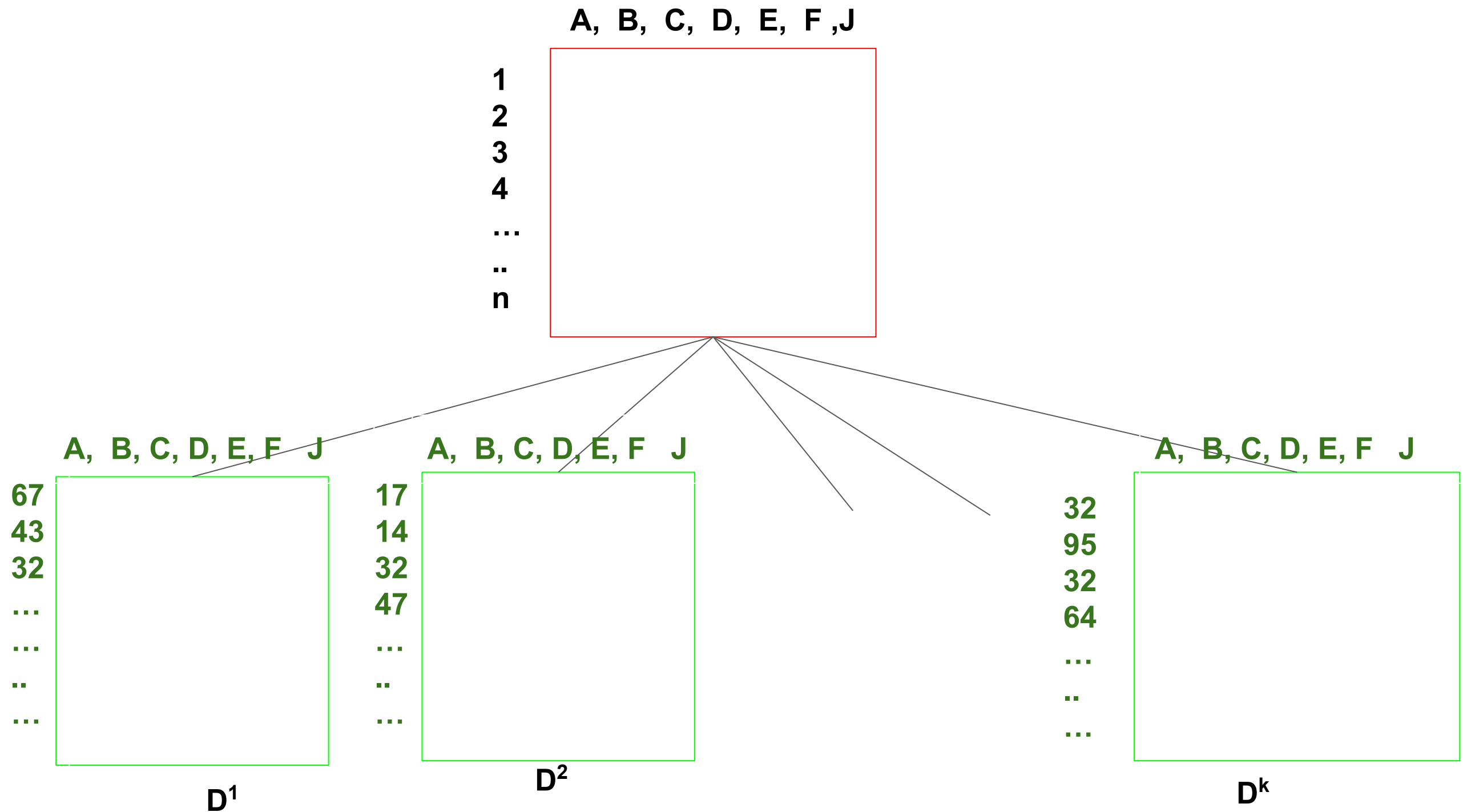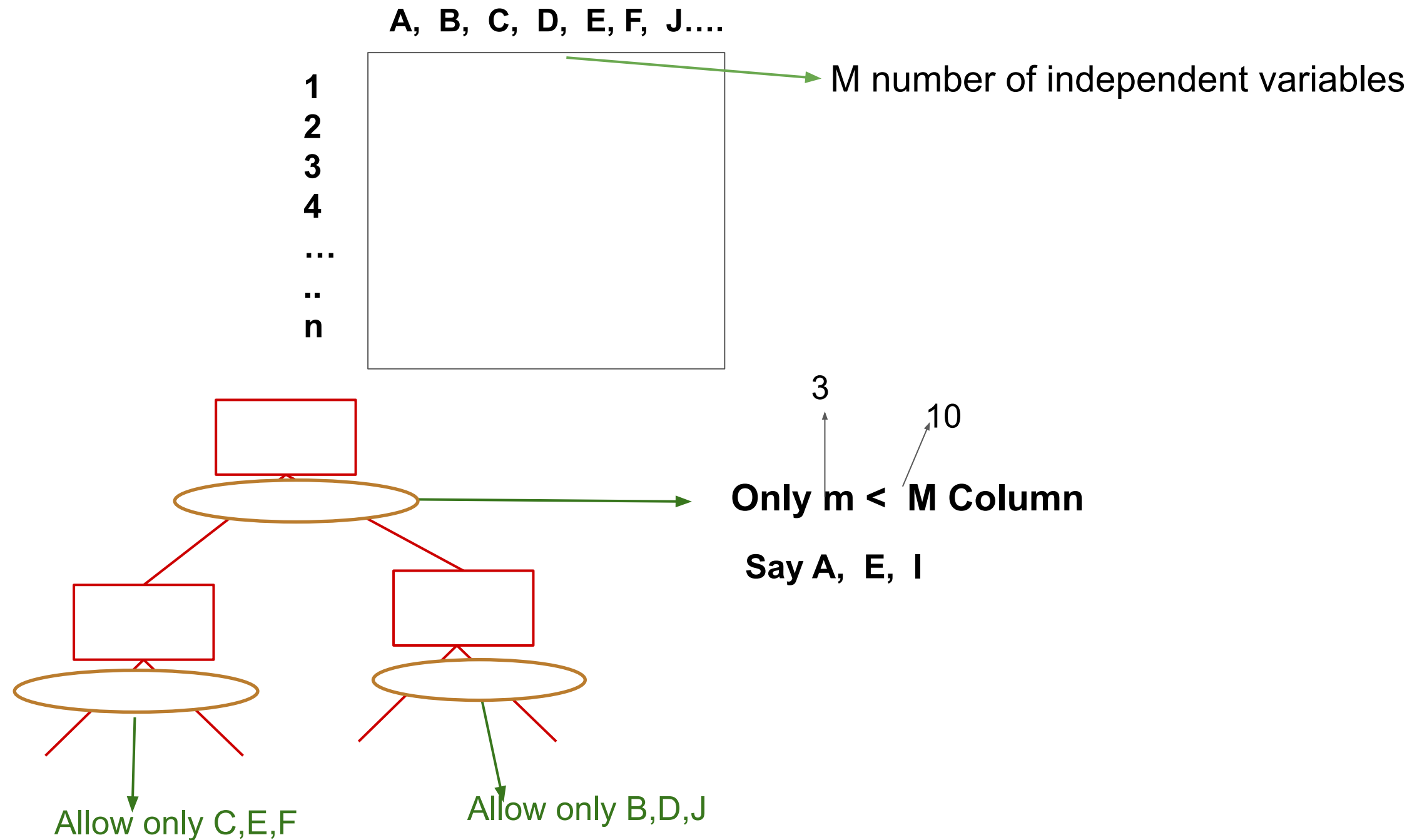
Random forest
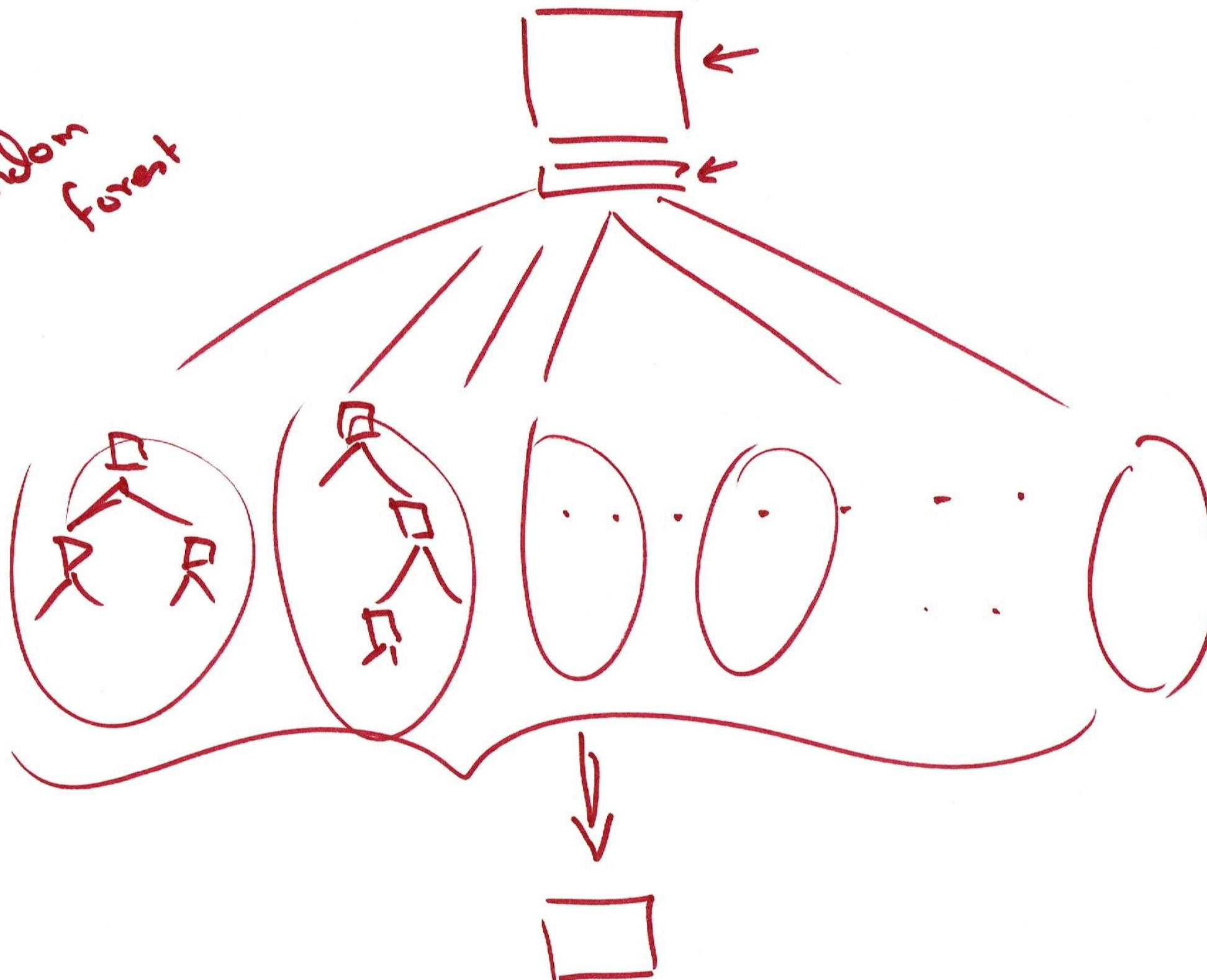
The General Ideas

- 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

A, B, C, D, E, F ,J

1
2
3
4
...
..
n

A, B, C, D, E,F J

67
43
32
...
...
..
...

D^1

A, B, C, D, E, F   J

17
14
32
47
...
..
...

D^2

A, B, C, D, E, F   J

32
95
32
64
...
..
...

D^k

Random Forest

A, B, C, D, E, F, J….

1
2
3
4
…
..
n

→ M number of independent variables

3

10

Only m < M Column

Say A, E, I

Allow only C,E,F

Allow only B,D,J

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.

  $< m$

- 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 ◯ ◯ ◯ Ⓙ

①

M=10   => High Tree Correlation

Good

②

M=2   => Your trees are weak