RANDOM FORESTS

- Bagging can greatly improve the performance of CART
  - Averaging over ensemble prediction, in case of regression trees
  - Majority vote over ensemble prediction, for classification trees

- However, the 'B' bootstrapped dataset are correlated!

  Therefore, the variance reduction due to averaging is diminished

Recall

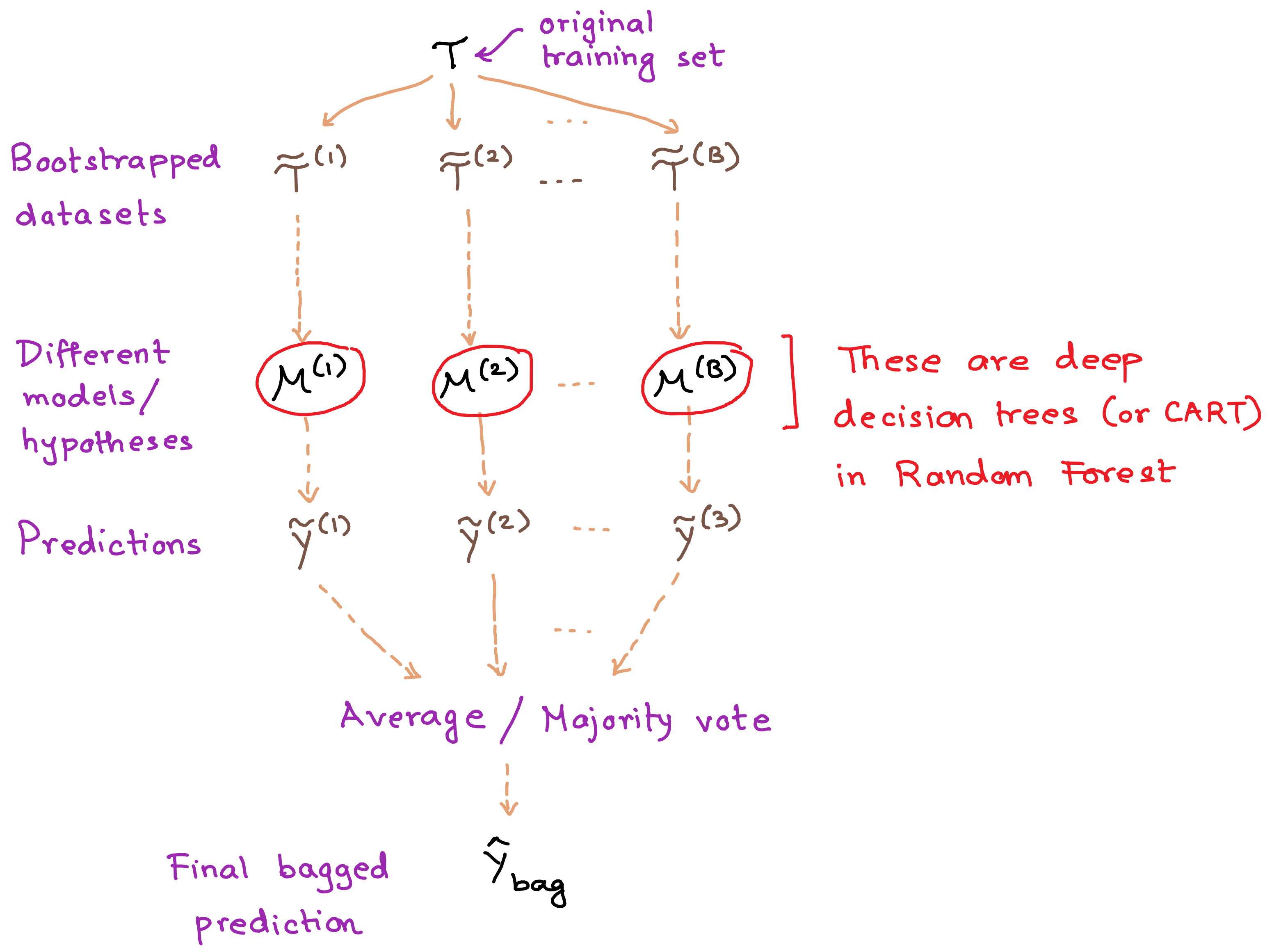
$$\mathrm{Var}\left[\frac{1}{B}\sum_{b=1}^{B} z_b\right] = \frac{1-\rho}{B}\sigma^2 + \rho\sigma^2$$

- No variance reduction when $\rho = 1$
- Highest variance reduction when $\rho = 0$

- Idea of Random Forest: De-correlate the 'B' trees by injecting additional randomness when constructing each tree

Random Forest = Bagging + Decision Trees (with random feature subset selection)

Bagging



Random Feature Subsets

- While growing a decision tree, one selects the best input feature x_j from all 'p' input variables $x_1, x_2, \ldots, x_p$ for splitting a node

- In random forest, we pick a random subset consisting of $q \leq p$ features and only consider these 'q' input features for possible splits

Example

A bootstrapped dataset

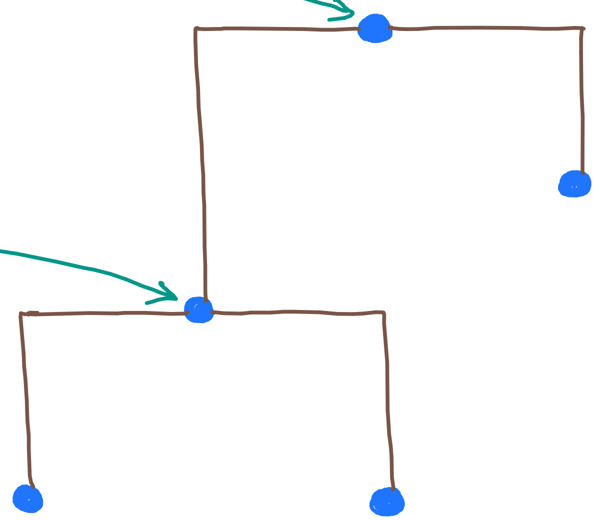
x_1	x_2	x_3	x_4
1.1	3.5	1.7	0.2
1.1	3.5	1.7	0.2
-0.9	3.3	0.5	0.1
1.2	3.2	-1	0.2
-0.9	3.3	0.5	0.1

$p = 4$ (# of inputs)

$q = 2$ random subsets

~~x_1~~	x_2 ✓	~~x_3~~	x_4 ✓
~~1.1~~	3.5	~~1.7~~	0.2
~~1.1~~	3.5	~~1.7~~	0.2
~~-0.9~~	3.3	~~0.5~~	0.1
~~1.2~~	3.2	~~-1~~	0.2
~~-0.9~~	3.3	~~0.5~~	0.1

~~x_1~~	~~x_2~~	x_3 ✓	x_4 ✓
~~1.1~~	~~3.5~~	1.7	0.2
~~1.1~~	~~3.5~~	1.7	0.2
~~-0.9~~	~~3.3~~	0.5	0.1
~~1.2~~	~~3.2~~	-1	0.2
~~-0.9~~	~~3.3~~	0.5	0.1



Random forest algorithm

Inputs : $\mathcal{T} = \{ \underline{x}^{(i)}, y^{(i)} \}_{i=1}^{N}$; $\underline{x} \in \mathbb{R}^p$

for $b = 1$ to B, do (can run in parallel)

(a) Draw a bootstrap dataset $\tilde{\mathcal{T}}^{(b)}$ of size N from $\mathcal{T}$

(b) Grow a regression (or classification) tree by repeating the steps below, until a minimum node size is reached :

Thumb rule
$q = \sqrt{p}$ (for CT)
$q \approx p/3$ (for RT)

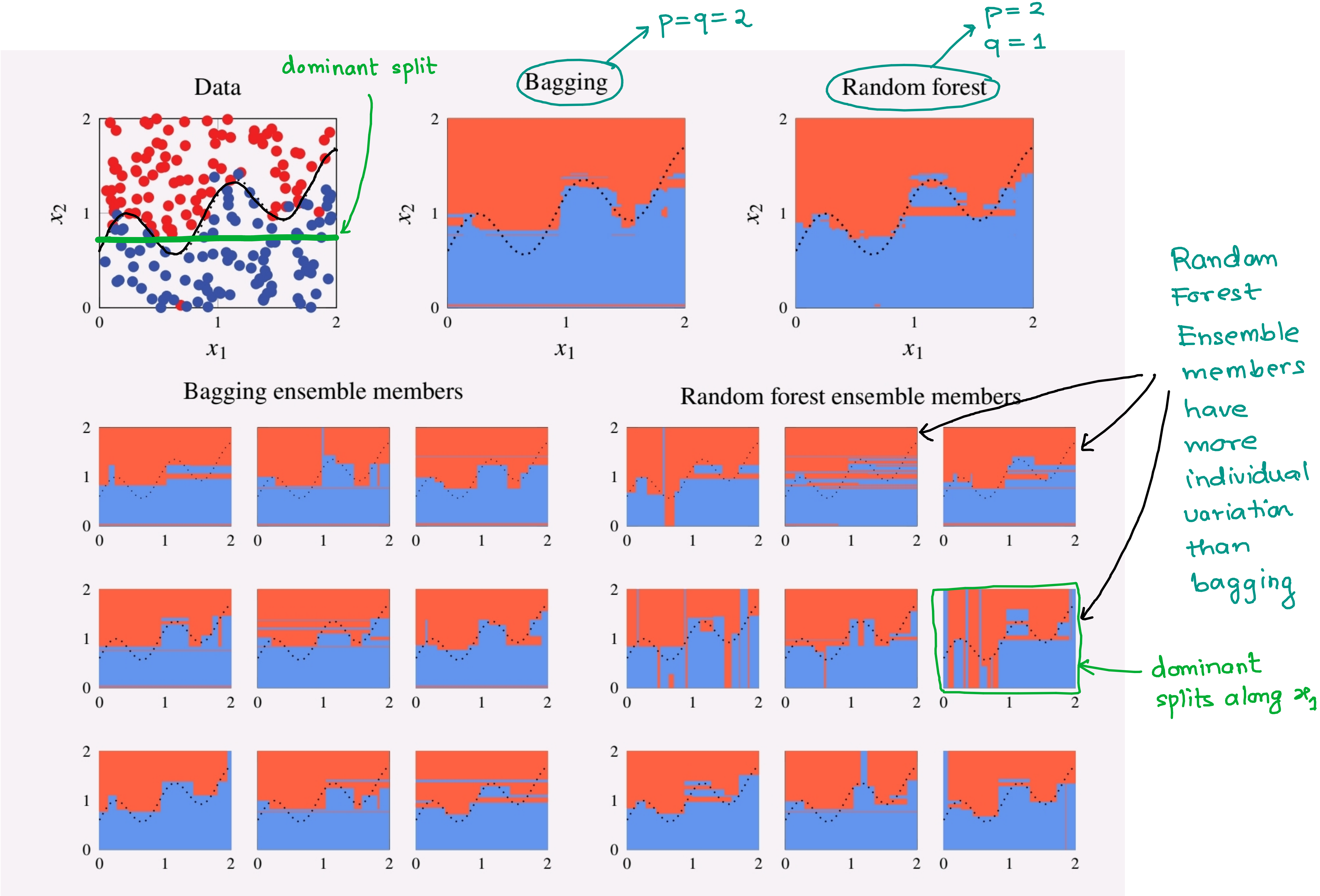
- Select a random subset consisting of $q \leq p$ inputs
- Find the best splitting variable x_j among the 'q' selected inputs
- Split the node into two children with $\{ x_j \leq s \}$ and $\{ x_j > s \}$

Final model is the average of the 'B' ensemble members

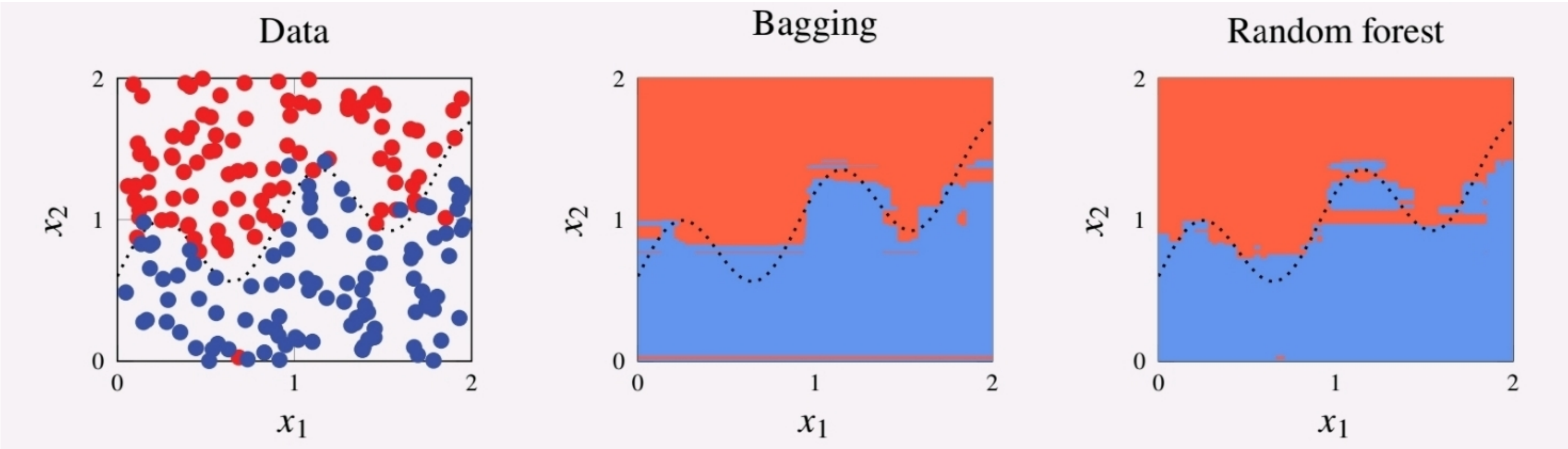
$$\boxed{\hat{y}_{rf} = \frac{1}{B} \sum_{b=1}^{B} \tilde{y}^{(b)}}$$

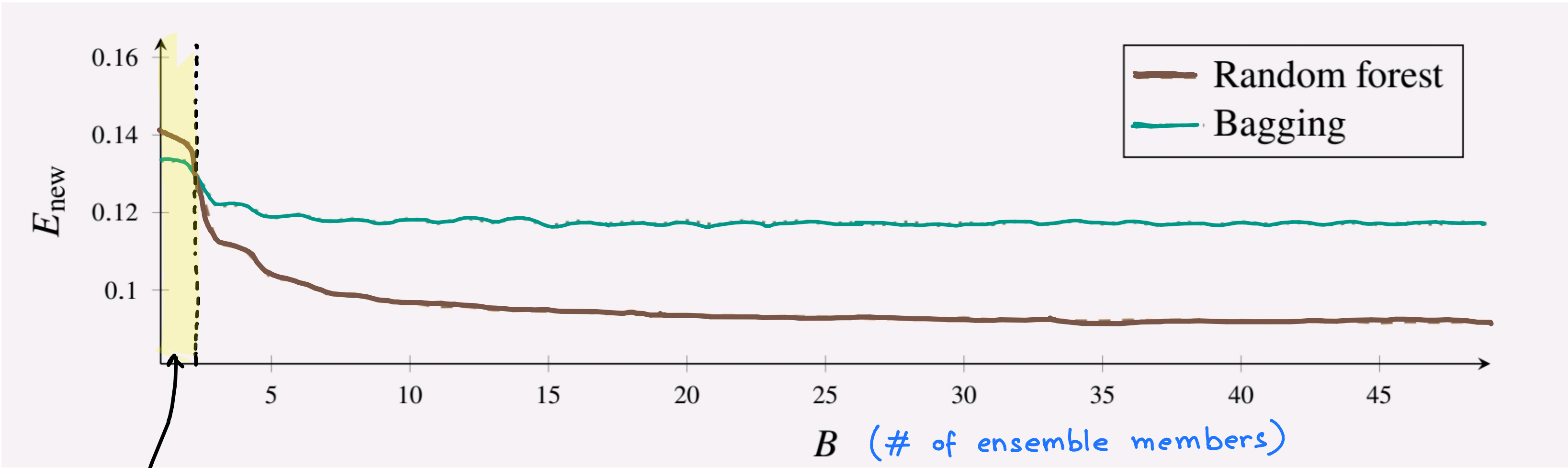
Example of binary classification

$B = 9$ ensemble members



Random Forest Ensemble members have more individual variation than bagging





For very small B, bagging performs better than random forests

However, as the number of ensemble member increases, test error decreases more for random forests

- For identically distributed random variables $\{z_b\}_{b=1}^{B}$

$$\mathrm{Var}\left[\frac{1}{B}\sum_{b=1}^{B} z_b\right] = \frac{1-\rho}{B}\sigma^2 + \rho\sigma^2$$

- The random input selection used in random forests:

  - increases the bias, but often very slowly ↓
  - adds to the variance (σ^2) of each tree ↓
  - reduces the correlation (ρ) between member trees ↑↑↑

- The reduction in correlation typically has a dominant effect

  ⇒ leads to an overall reduction in error

- Bagging is a general technique → can be used with any base model

  Random forests consider base models as classification or regression trees