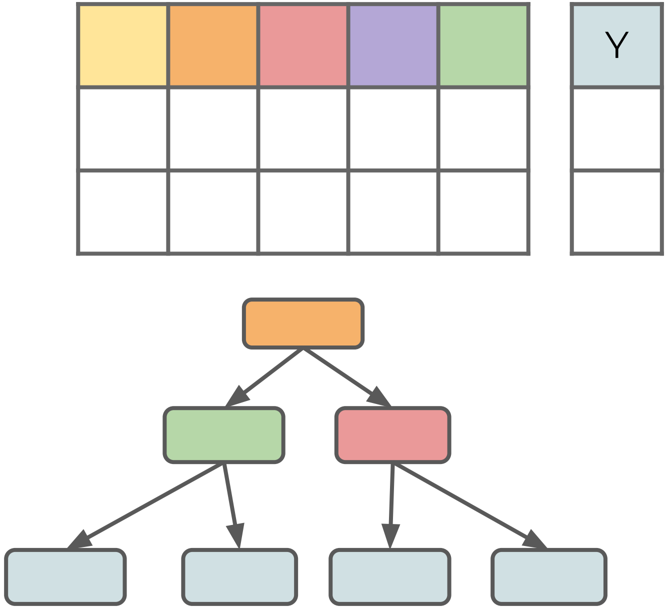
Random Forests

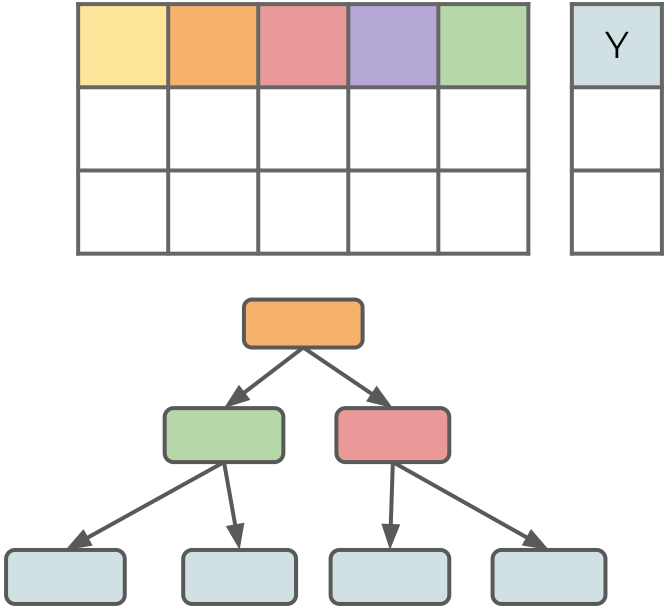
Random Forests

- Decision Tree restricted by gini impurity:



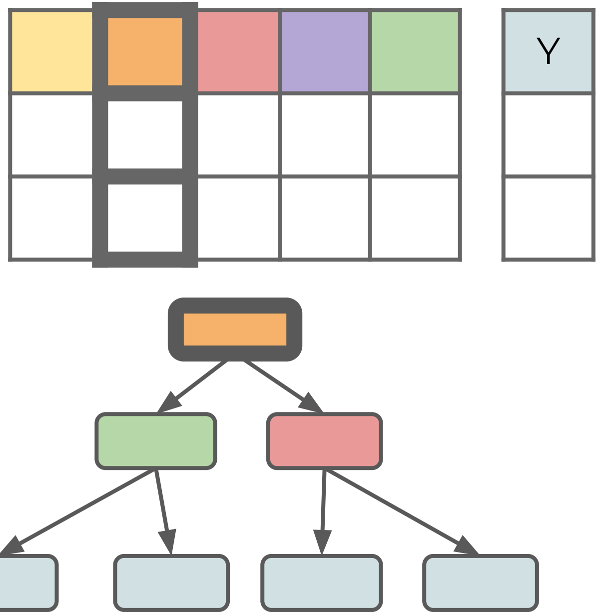
Random Forests

- No guarantee of using all features!



Random Forests

- Root feature has huge influence over tree.



Random Forests

- We could try adjusting rules, such as:
    - Splitting Criterion (Information Gain)
    - Minimum Gini Impurity Decrease
    - Setting Depth Limits
    - Limits on number of leaf nodes

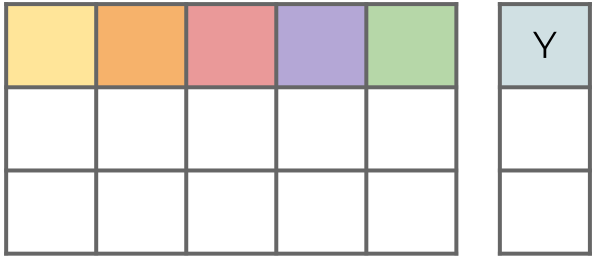
Random Forests

Main ideas

Random Forests

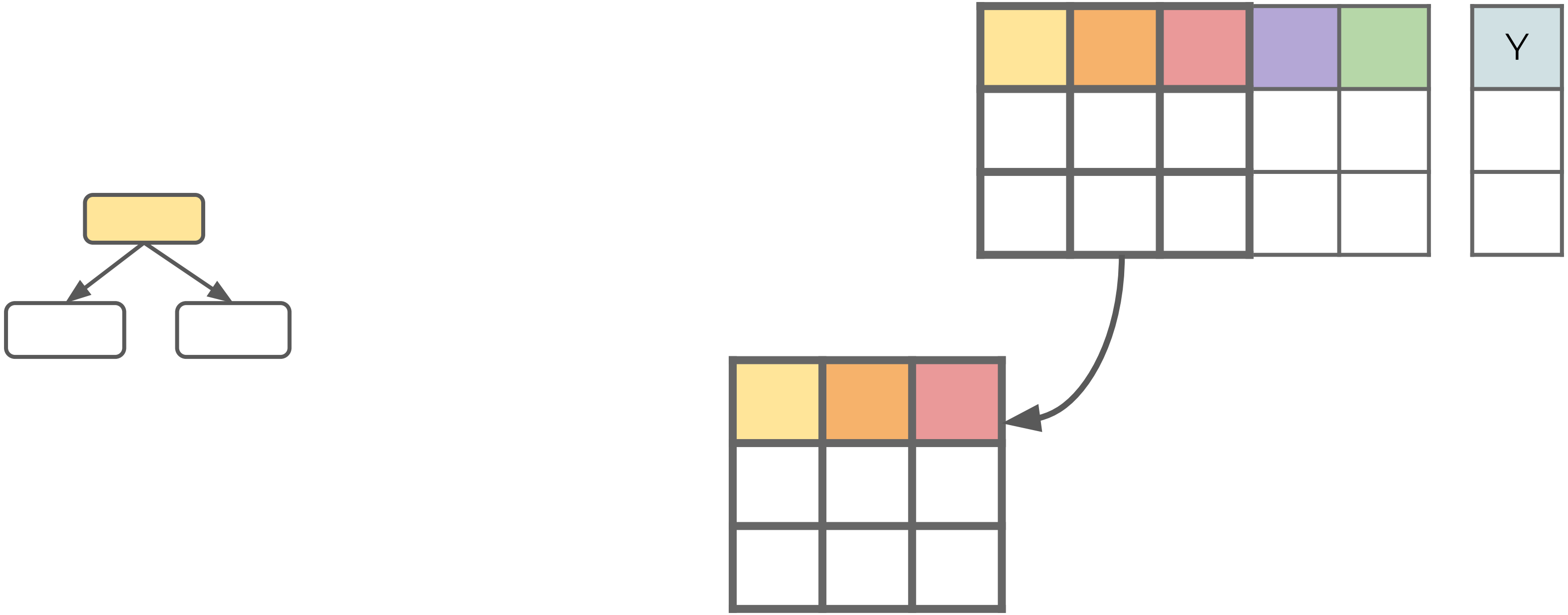
- Known as **ensemble** learners, since they rely on an ensemble of models (multiple decision trees).

Random Forests

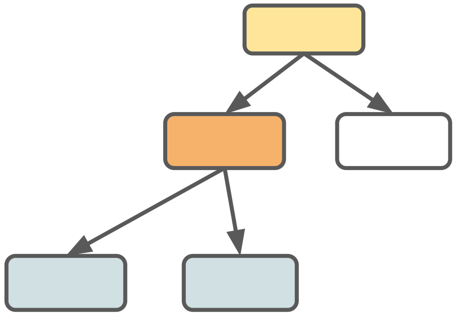


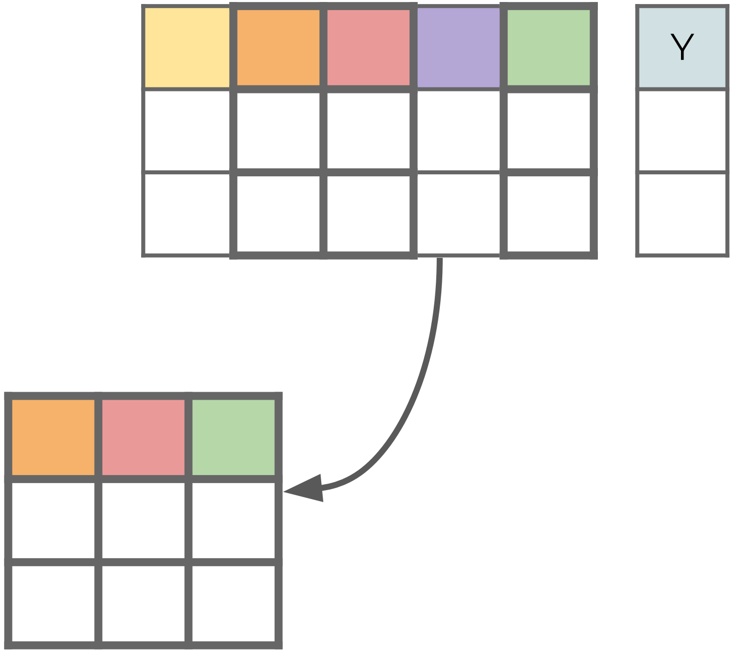
Create subsets of randomly picked features at each potential split

Random Forests

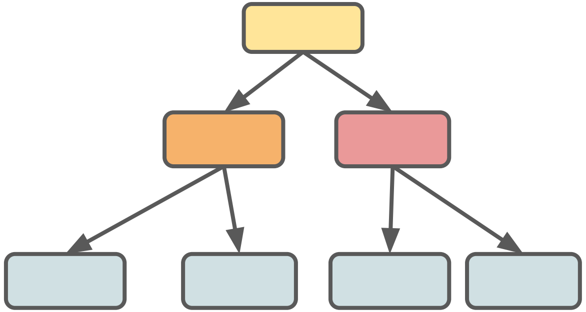


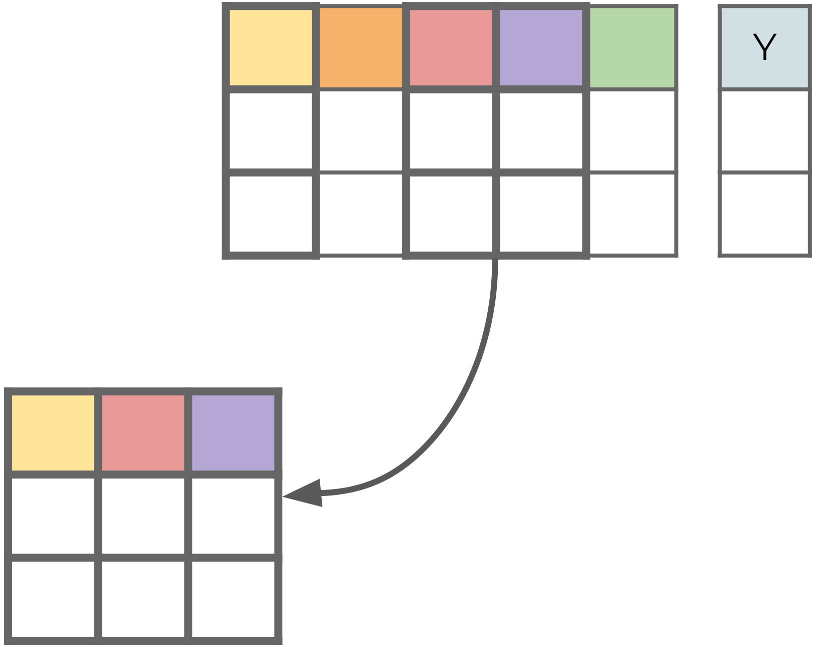
Random Forests



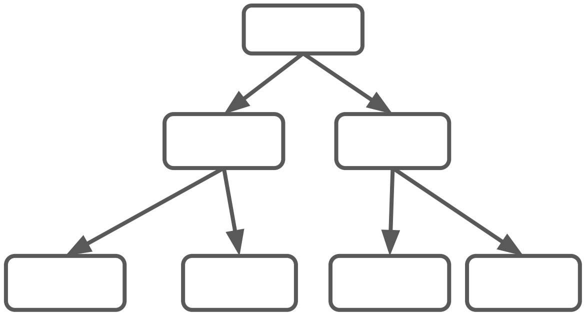


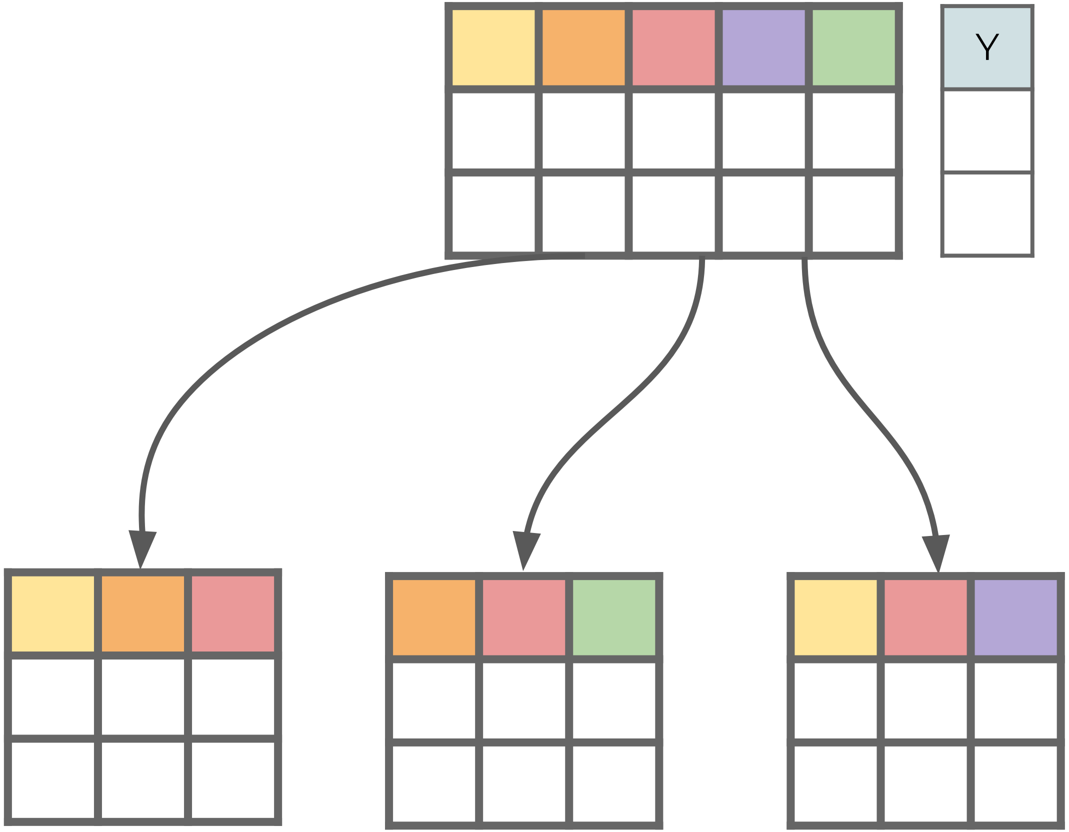
Random Forests





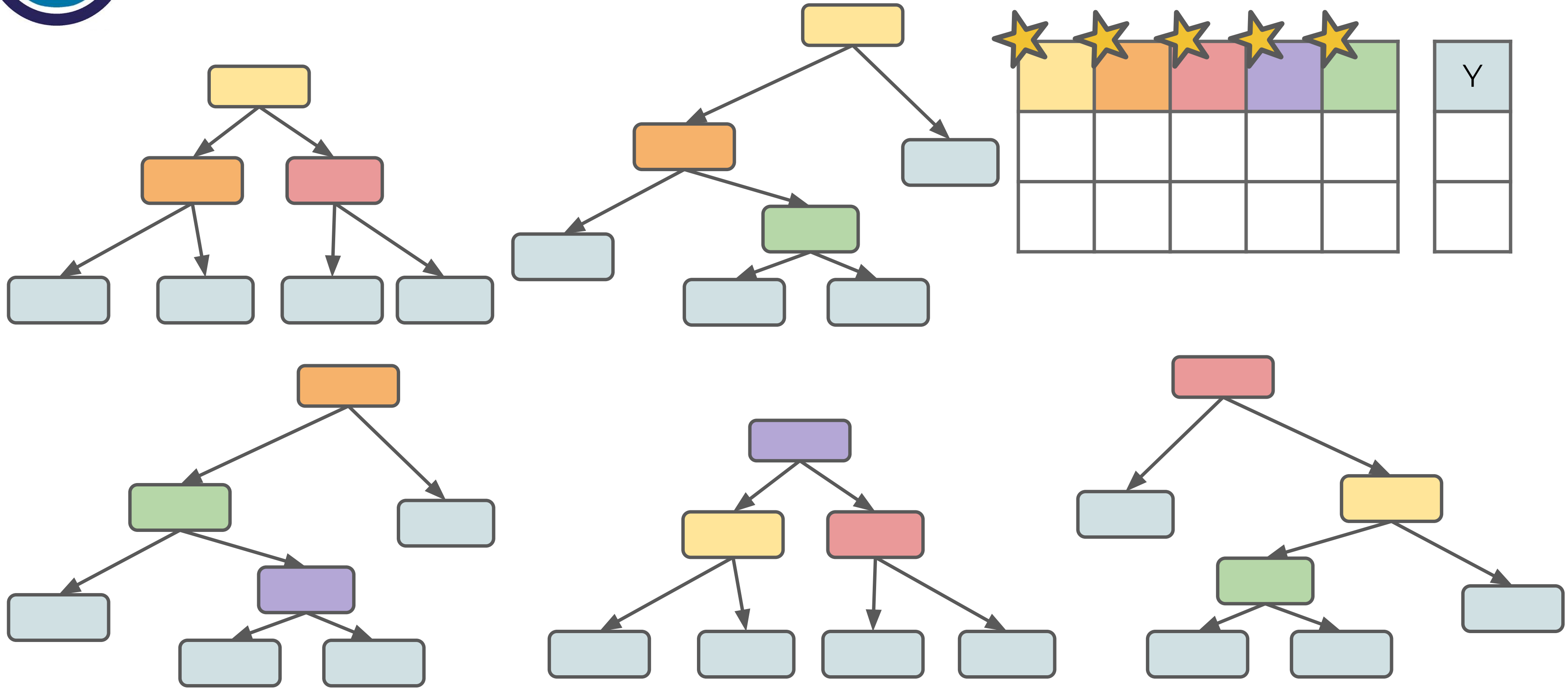
Random Forests



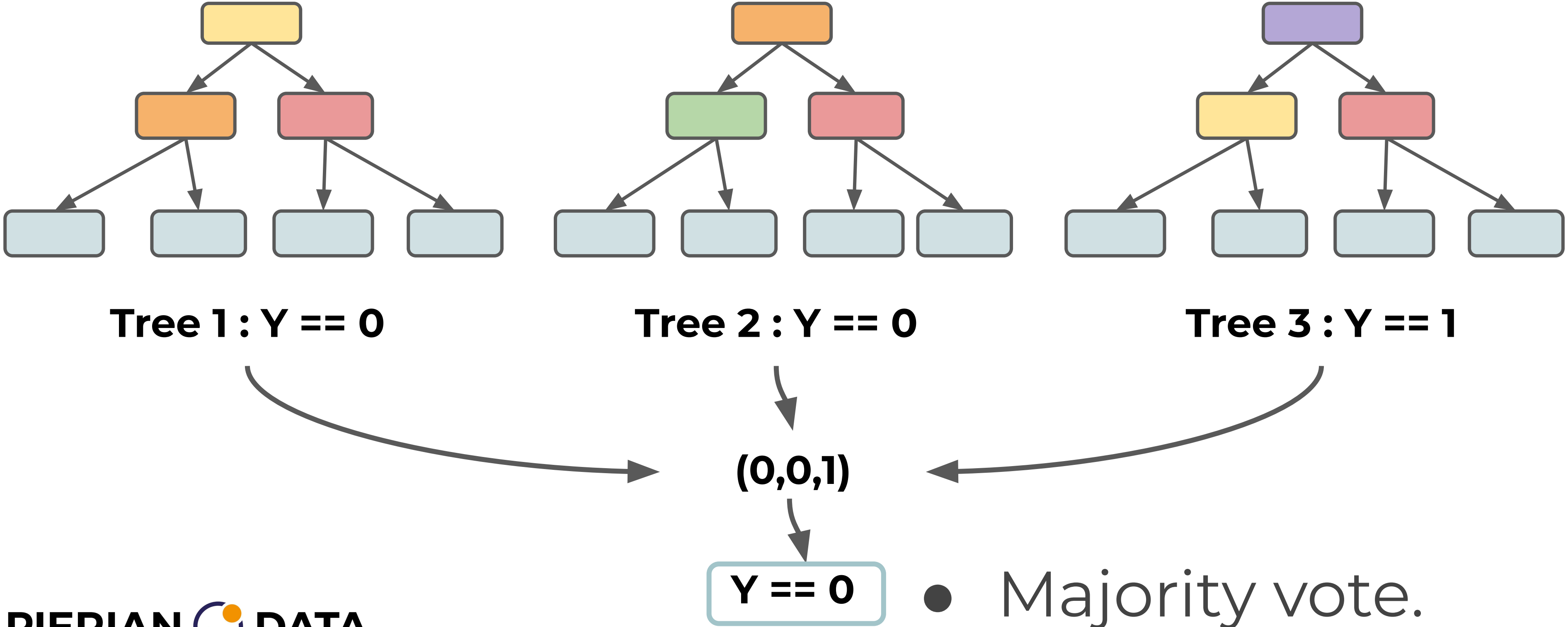


Random Forests

- All features used!

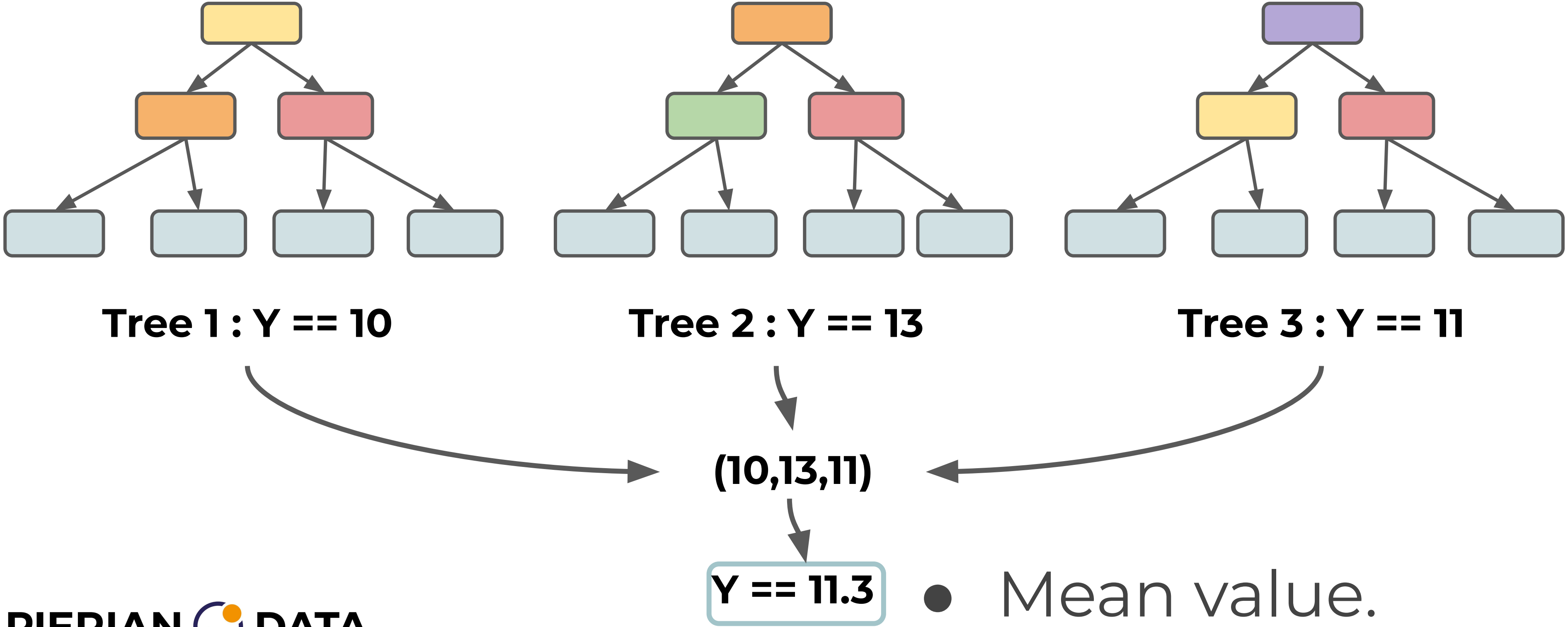


Random Forests



- Majority vote.

Random Forests



- Mean value.

Random Forests

Theory and Intuition: Hyperparameters

Random Forests

class sklearn.tree.**DecisionTreeClassifier**(**, criterion='gini', splitter='best', max_depth=None, min_samples_split=2, min_samples_leaf=1, min_weight_fraction_leaf=0.0, max_features=None, random_state=None, max_leaf_nodes=None, min_impurity_decrease=0.0, min_impurity_split=None, class_weight=None, ccp_alpha=0.0*) ¶ [source]

class sklearn.ensemble.**RandomForestClassifier**(*n_estimators=100, *, criterion='gini', max_depth=None, min_samples_split=2, min_samples_leaf=1, min_weight_fraction_leaf=0.0, max_features='auto', max_leaf_nodes=None, min_impurity_decrease=0.0, min_impurity_split=None, bootstrap=True, oob_score=False, n_jobs=None, random_state=None, verbose=0, warm_start=False, class_weight=None, ccp_alpha=0.0, max_samples=None*) ¶ [source]

Random Forests

```
class sklearn.tree.DecisionTreeClassifier(*, criterion='gini', splitter='best', max_depth=None, min_samples_split=2, min_samples_leaf=1, min_weight_fraction_leaf=0.0, max_features=None, random_state=None, max_leaf_nodes=None, min_impurity_decrease=0.0, min_impurity_split=None, class_weight=None, ccp_alpha=0.0) ¶ [source]
```

```
class sklearn.ensemble.RandomForestClassifier(n_estimators=100, *, criterion='gini', max_depth=None, min_samples_split=2, min_samples_leaf=1, min_weight_fraction_leaf=0.0, max_features='auto', max_leaf_nodes=None, min_impurity_decrease=0.0, min_impurity_split=None, bootstrap=True, oob_score=False, n_jobs=None, random_state=None, verbose=0, warm_start=False, class_weight=None, ccp_alpha=0.0, max_samples=None) ¶ [source]
```

Random Forests

class `sklearn.ensemble.RandomForestClassifier`(*n_estimators=100*, *, *criterion='gini', max_depth=None, min_samples_split=2, min_samples_leaf=1, min_weight_fraction_leaf=0.0, max_features='auto', max_leaf_nodes=None, min_impurity_decrease=0.0, min_impurity_split=None, bootstrap=True, oob_score=False, n_jobs=None, random_state=None, verbose=0, warm_start=False, class_weight=None, ccp_alpha=0.0, max_samples=None*) ¶ [source]

Random Forests

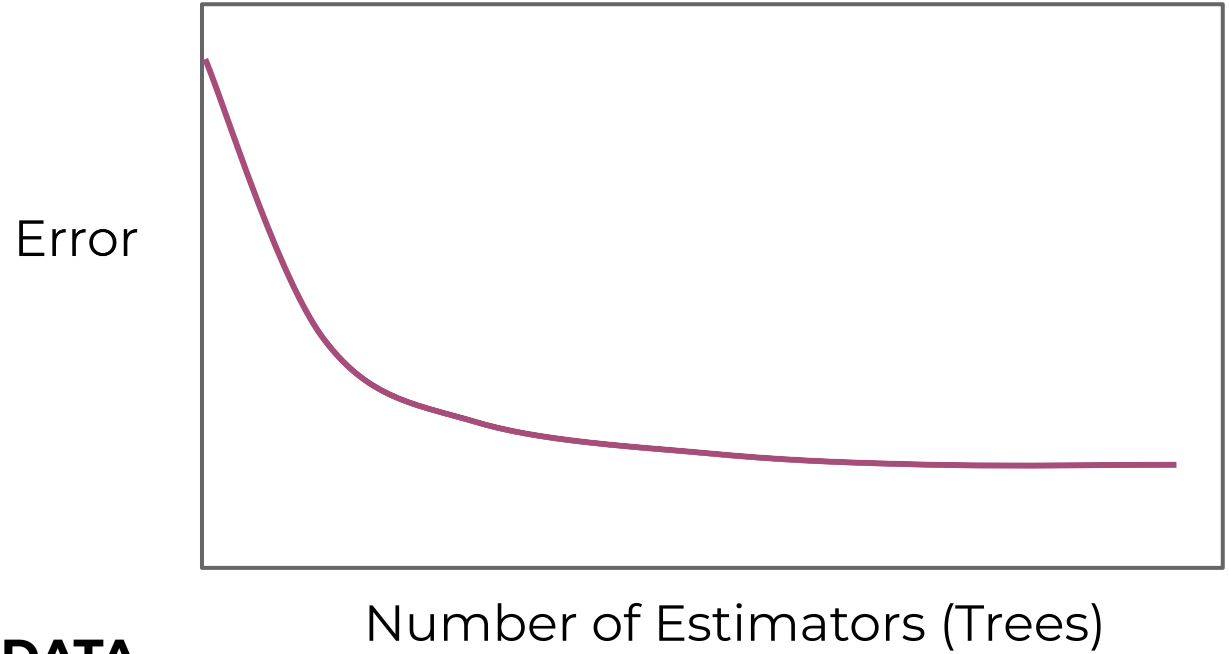
- Random Forest Hyperparameters:
  - Number of Estimators
  - Number of Features per estimator
  - Bootstrap Samples

Random Forests

- How to choose number of trees?
  - Reasonable Default Value: 100
  - Publications suggest 64-128 trees.
  - Cross Validate a grid search of trees.
  - Plot Error versus number of trees.
    - Should notice diminishing error reduction after some N trees (similar to KNN).

Random Forests

- Error vs. Trees



Random Forests

- After a certain number of trees, two things that can occur:
  - Different random selections don’t reveal any more information.
    - Trees become highly correlated.
  - Different random selections are simply duplicating trees that have already been created.

Random Forests

- Number of Features in Subset?
    - Current suggested convention is **sqrt(N)** in the subset given **N** features.
    - **N/3** may be more suitable for regression tasks, typically larger than **sqrt(N)**.
  - should be treated as a tuning parameter, with **sqrt(N)** as a good starting point.

Random Forests

- Hyperparameter Review:
    - Number of Estimators:
        - Start with 100 as default, feel free to grid search for lower and higher values.
    - Number of Features for Selection:
        - Start with sqrt(N), grid search for other possible values (N/3).

Random Forests

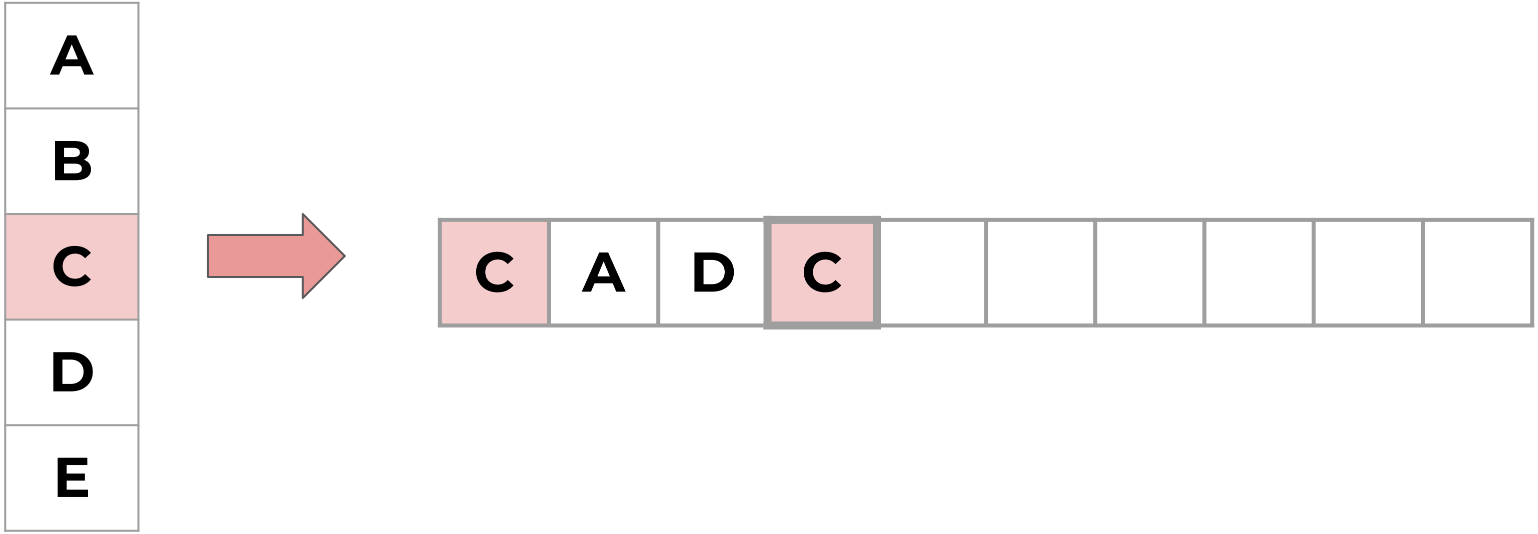
Theory and Intuition: Hyperparameters
Bootstrap Samples and OOB Error

Random Forests

- Bootstrap Samples
    - *Allow for bootstrap sampling of each training subset of features?*
- First, let’s understand “bootstrapping” in general terms…

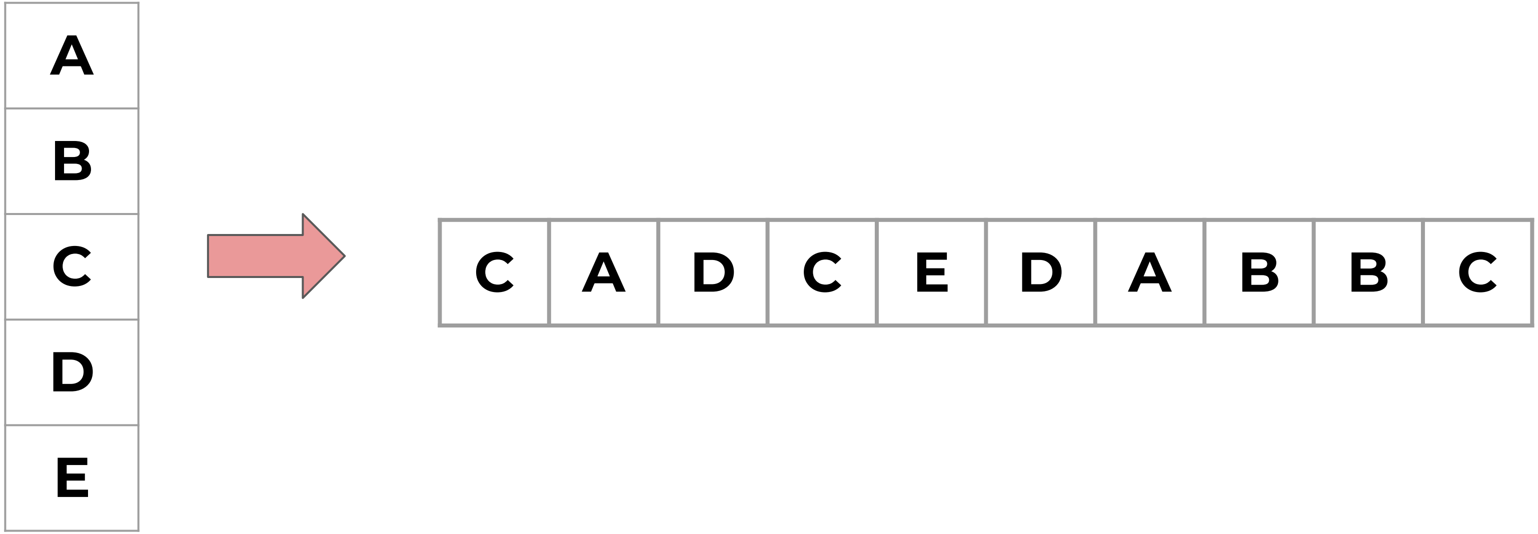
Random Forests

- Bootstrapping = sampling with replacement.



Random Forests

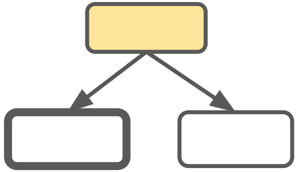
- Bootstrapping = sampling with replacement.

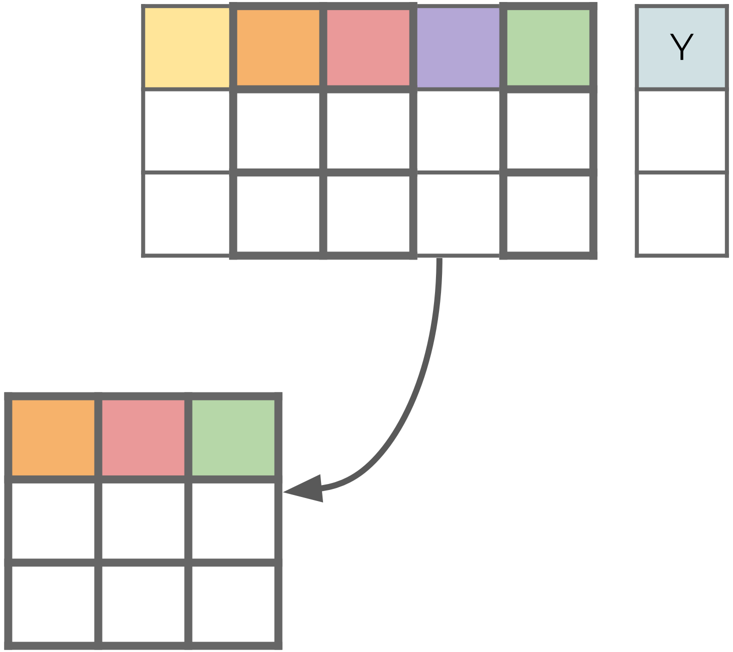


Random Forests

- Bootstrapping in Random Forest
    - Recall for each split we are randomly selecting a **subset of features**.
    - This random subset of features helps create more diverse trees that are not correlated to each other.

Random Forests

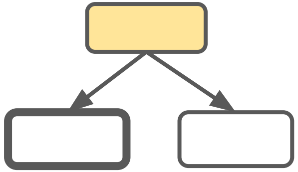




Random Forests

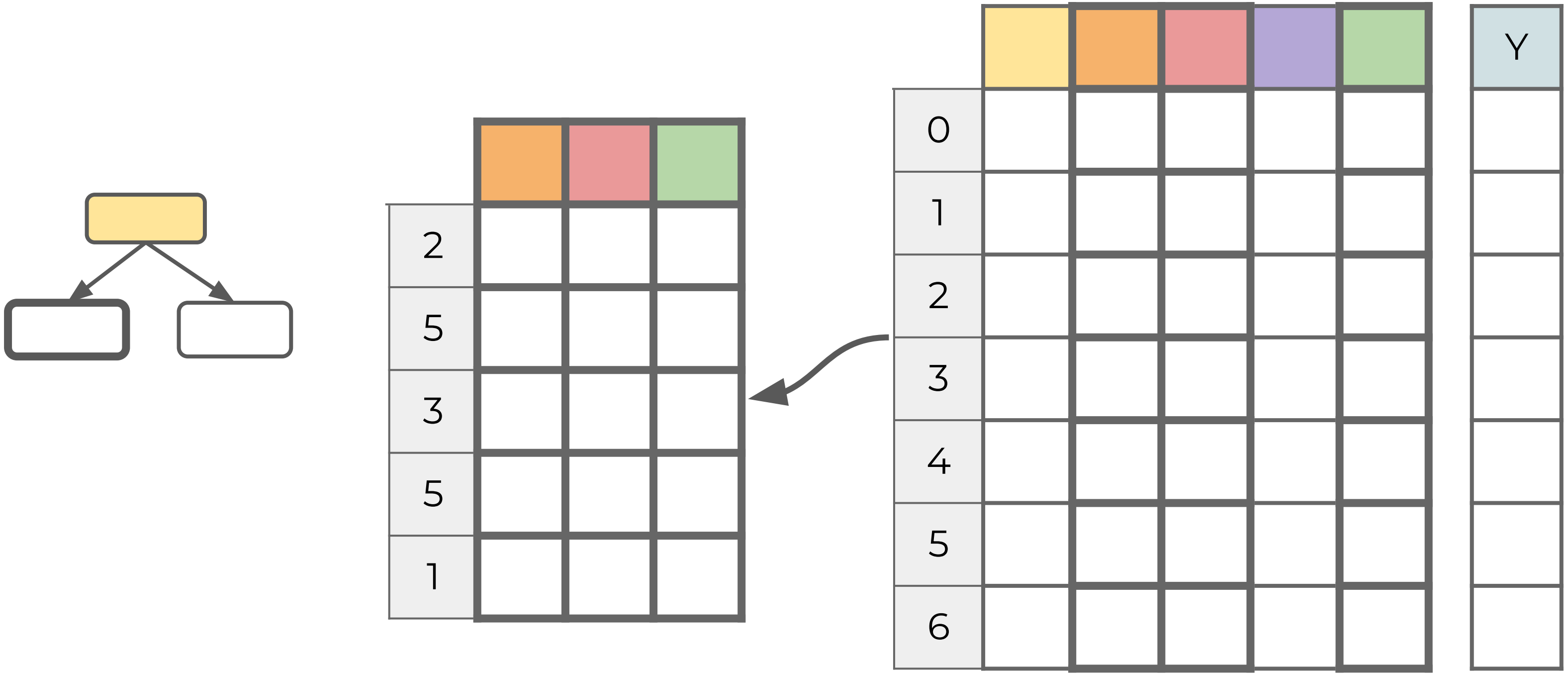
- Bootstrapping in Random Forest
  - To further differentiate trees, we could **bootstrap a selection of rows** for each split.
  - This results in <u>**two randomized training components**</u>:
    - Subset of Features Used
    - Bootstrapped rows of data

Random Forests



						Y
0						
1						
2						
3						
4						
5						
6						

Random Forests



Random Forests

- Bootstrapping can be set to False during training (it is True by default).
- Bootstrapping is yet another hyperparameter meant to reduce correlation between trees, since trees are then trained on different subsets of feature columns and data rows!

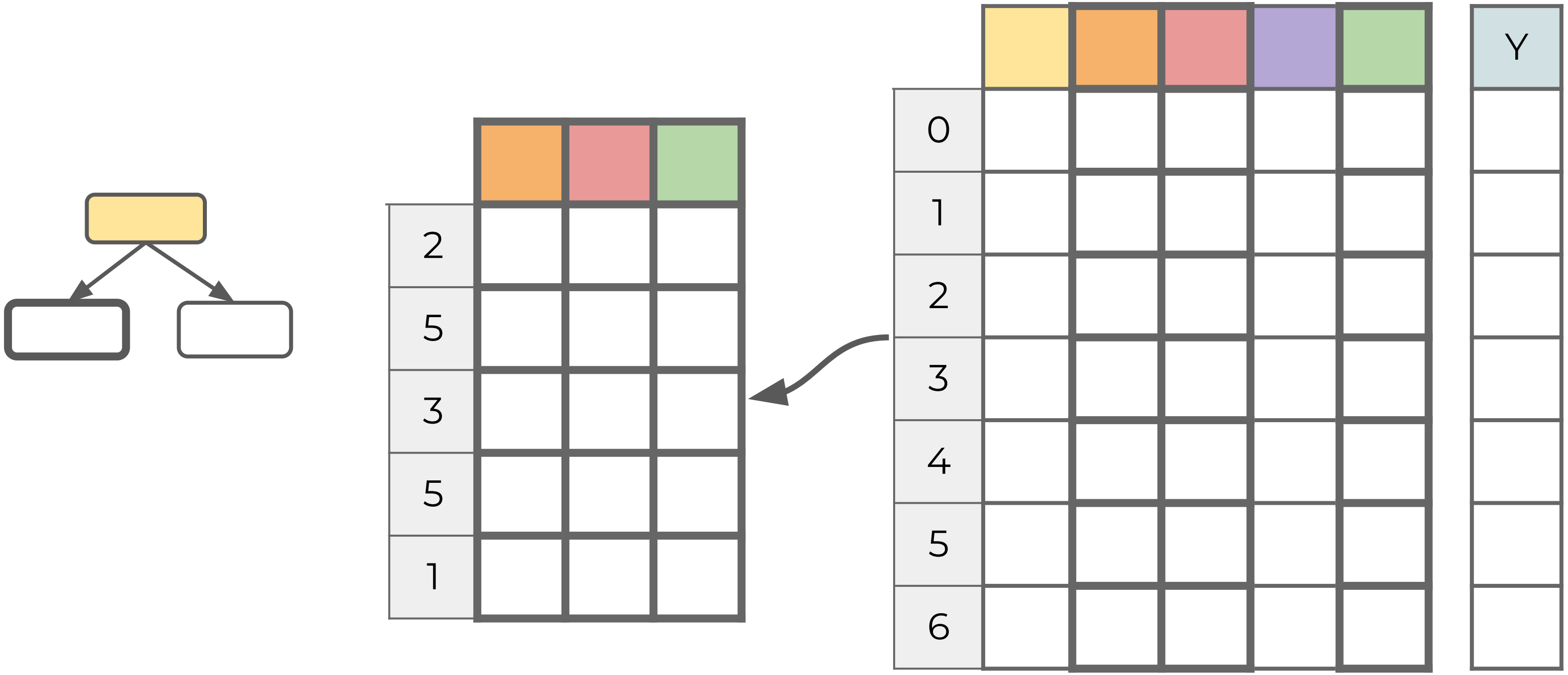
Random Forests

- What is Bagging?
    - Recall to actually use a Random Forest, we use **b**ootstrapped data and then calculate a prediction based on the **ag**gregated prediction of the trees:
        - Classification: Most Voted Y Class
        - Regression: Average Predicted Ys

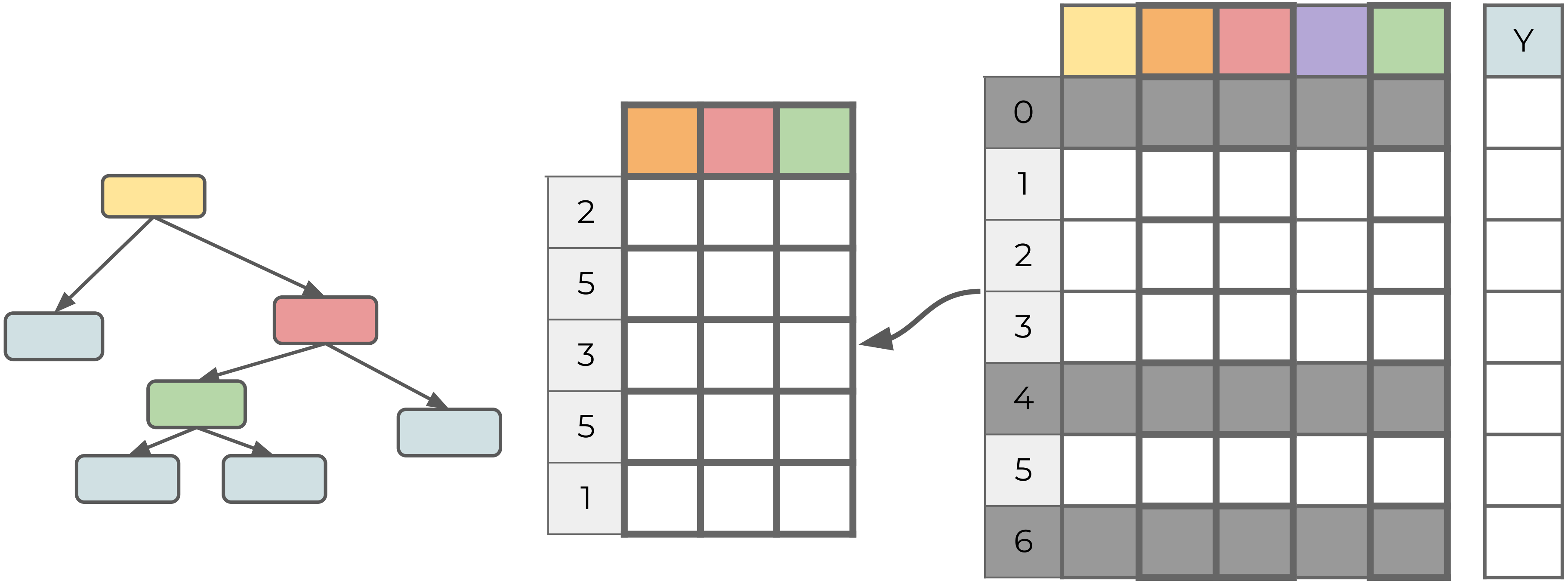
Random Forests

- What is Bagging?
  - If we performed bootstrapping when building out trees, this means that for certain trees, certain rows of data were not used for training.

Random Forests



Random Forests



Random Forests