Random Forest

Ravinder Singh

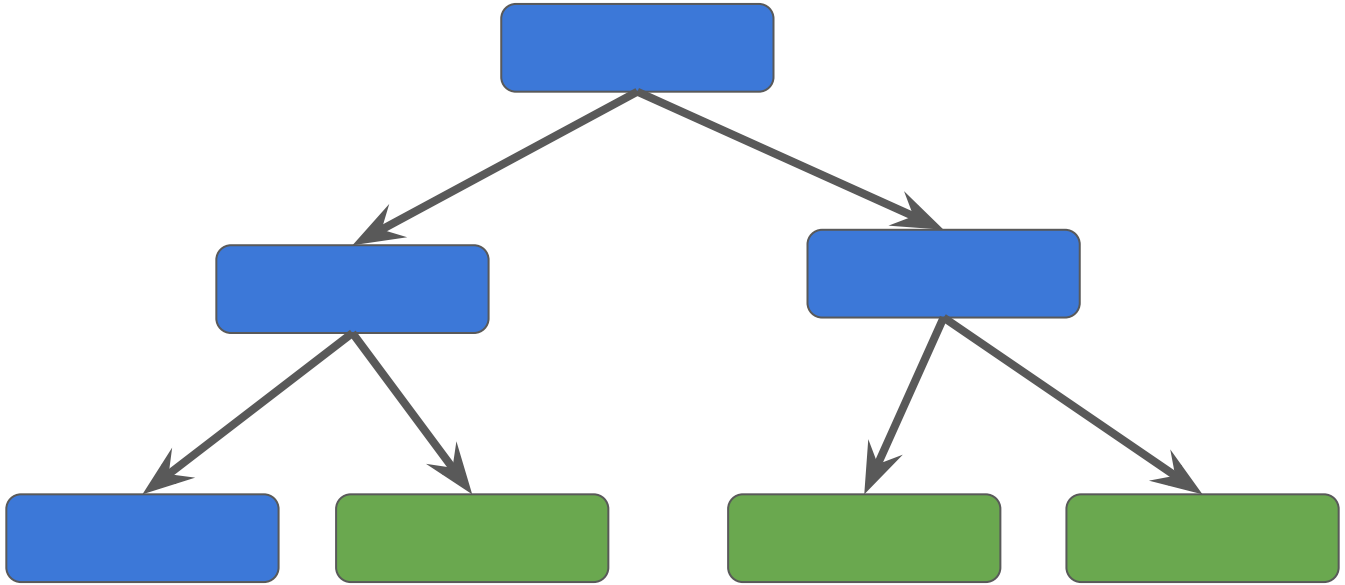
Random Forest

Random Forest is a type **Ensemble** Machine Learning algorithm called Bootstrap Aggregation or bagging.

Ensemble Machine Learning: general approach to machine learning that seeks better predictive performance by combining the predictions from multiple models.

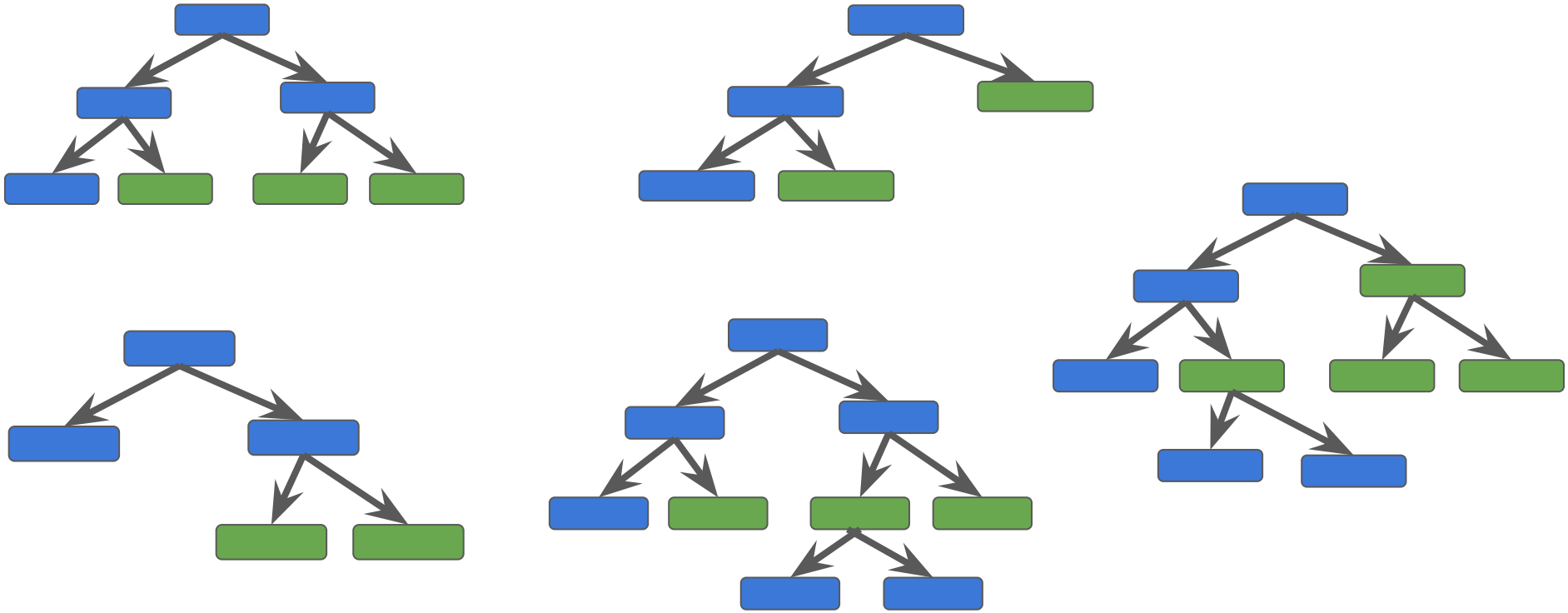
Why Random Forest?

Decision Trees are prone to overfitting!



Why Random Forest?

Combine simplicity of decision trees with flexibility to improve performance

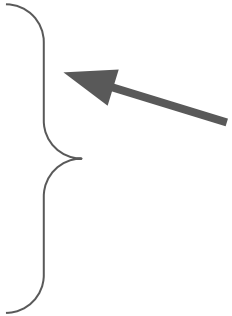


How to create a Random Forest

Step 1: ‘Bootstrapping’ a dataset

How to create a Random Forest

class	age	alone	sex	survived
1	adult	no	male	yes
2	child	no	female	yes
2	adult	yes	male	no
3	child	yes	male	no



For demo, we will use this dataset of 4 observations to build a decision tree

How to create a Random Forest

Original Dataset

class	age	alone	sex	survived
1	adult	no	male	yes
2	child	no	female	yes
2	adult	yes	male	no
3	child	yes	male	no

class age sex survived

Create a new dataset by choosing 4 observation at random from original dataset

We can pick same observation more than once to build this bootstrapped dataset

How to create a Random Forest

Original Dataset

class	age	alone	sex	survived
1	adult	no	male	yes
2	child	no	female	yes
2	adult	yes	male	no
3	child	yes	male	no

2nd sample

1st sample

3rd sample

4th sample

Bootstrapped Dataset

class	age	alone	sex	survived
2	child	no	female	yes
1	adult	no	male	yes
3	child	yes	male	no
3	child	yes	male	no

How to create a Random Forest

Step 2: Create a decision using the bootstrapped dataset and using random set of variables

Bootstrapped Dataset

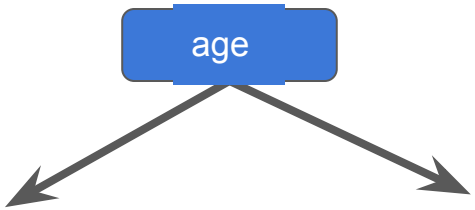
class	age	alone	sex	survived
2	child	no	female	yes
1	adult	no	male	yes
3	child	yes	male	no
3	child	yes	male	no

How to create a Random Forest

Randomly select two variables as candidates for root node

Bootstrapped Dataset

class	age	alone	sex	survived
2	child	no	female	yes
1	adult	no	male	yes
3	child	yes	male	no
3	child	yes	male	no

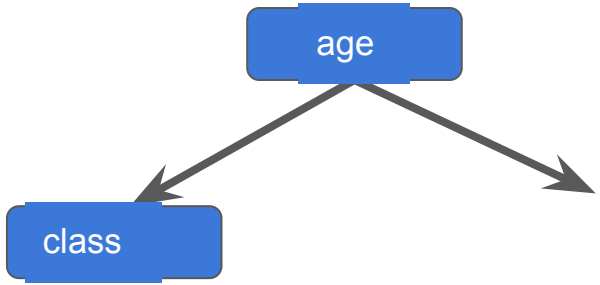


How to create a Random Forest

Again randomly select two random variables as candidate for the next node

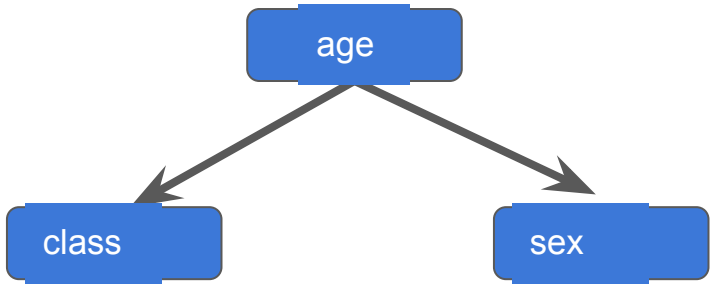
Bootstrapped Dataset

class	age	alone	sex	survived
2	child	no	female	yes
1	adult	no	male	yes
3	child	yes	male	no
3	child	yes	male	no



How to create a Random Forest

Again randomly select candidate for the next node

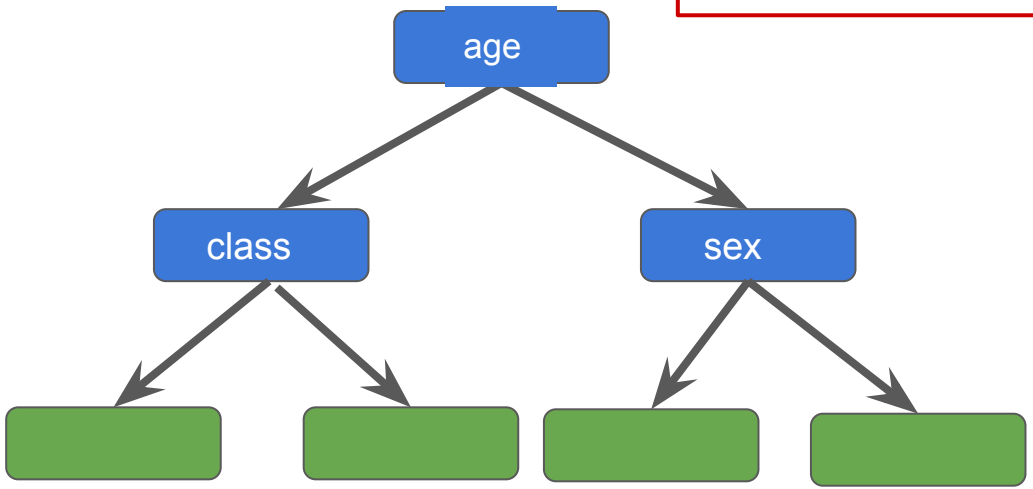


Bootstrapped Dataset

class	age	alone	sex	survived
2	child	no	female	yes
1	adult	no	male	yes
3	child	yes	male	no
3	child	yes	male	no

How to create a Random Forest

Continue building the tree as usual but choosing random set of variables to choose from at each node



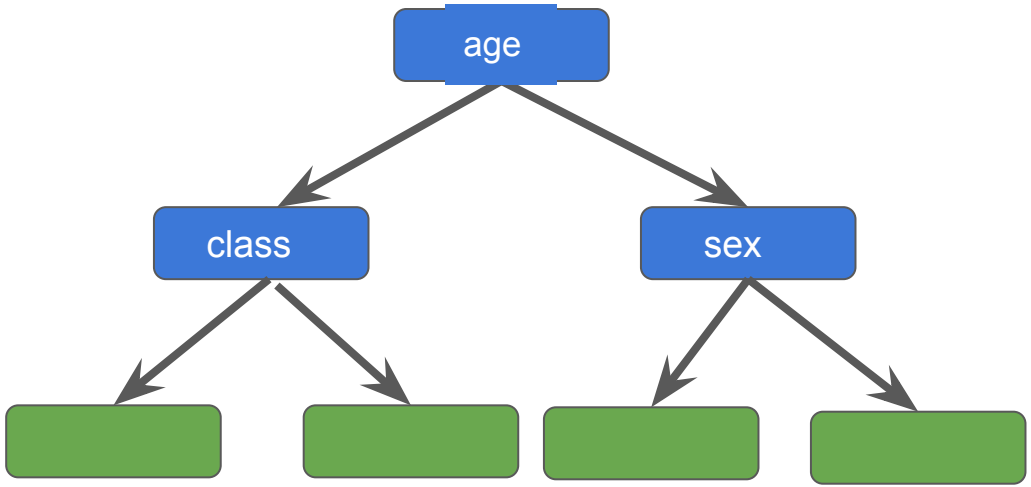
Bootstrapped Dataset

class	age	alone	sex	survived
2	child	no	female	yes
1	adult	no	male	yes
3	child	yes	male	no
3	child	yes	male	no

How to create a Random Forest

Summarize: we built a decision tree:

1. Using bootstrapped data
2. Choosing a random subset of variables at each node

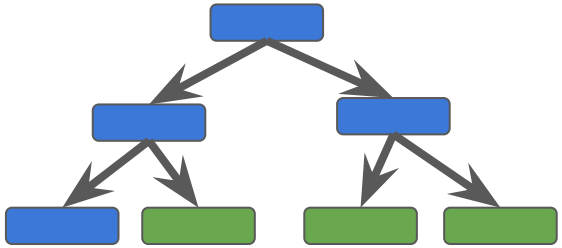


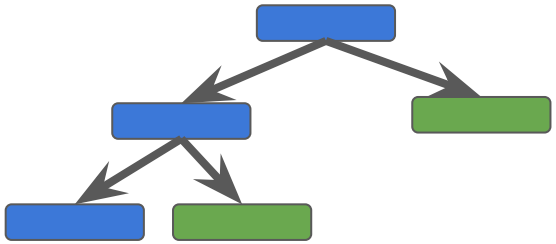
Bootstrapped Dataset

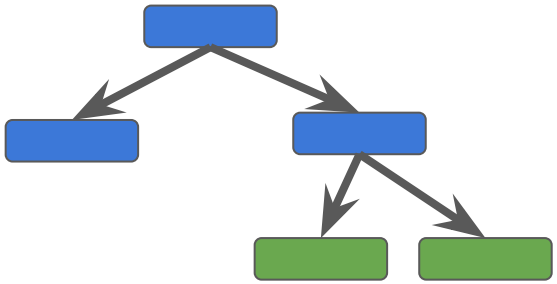
class	age	alone	sex	survived
2	child	no	female	yes
1	adult	no	male	yes
3	child	yes	male	no
3	child	yes	male	no

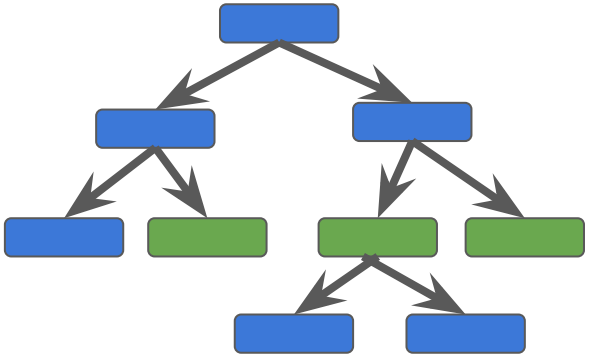
How to create a Random Forest

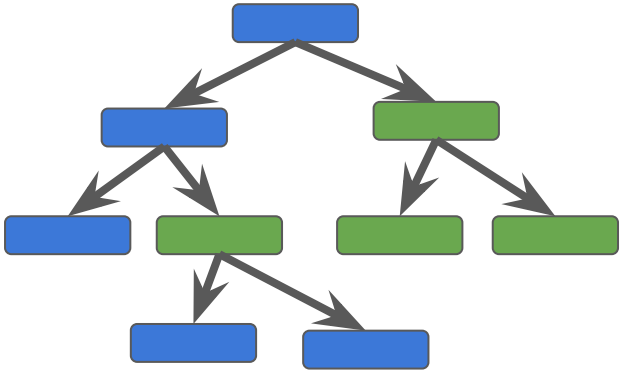
Now go back to step 1 and repeat: Make a new bootstrapped dataset and choose a random set of variables at each node to create new tree.





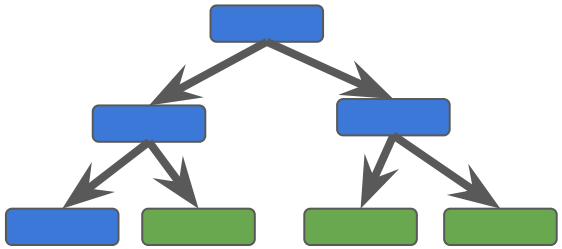


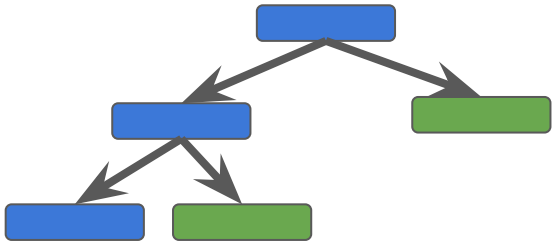


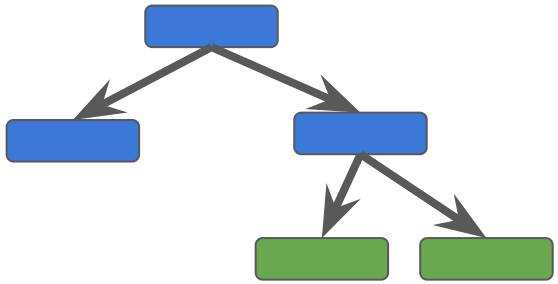


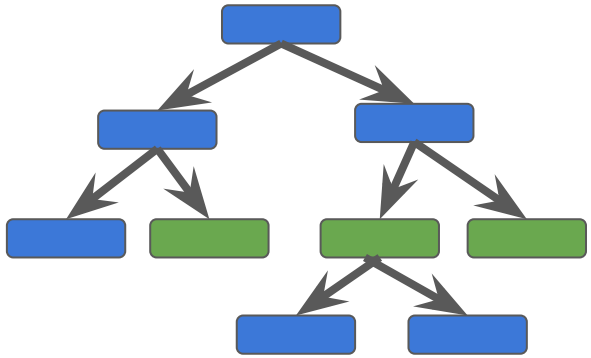
How to create a Random Forest

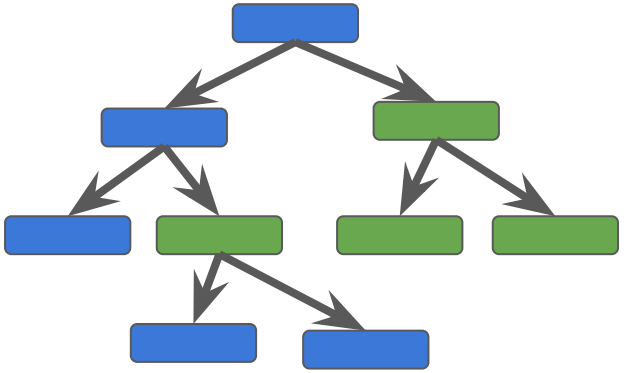
You do this 100's of times...but I have space only for 5 on this slide









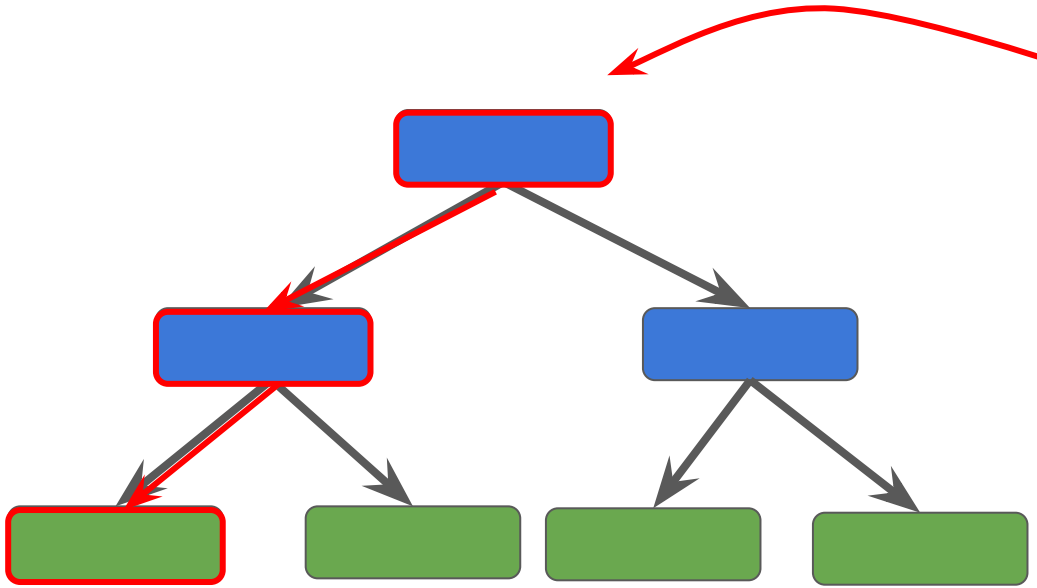


How Random Forest work?

What does the **first** decision tree predict for an observation?

Make Prediction for this observation

class	age	alone	sex	survived
2	child	no	female	??



Survived	Not Survived
1	0

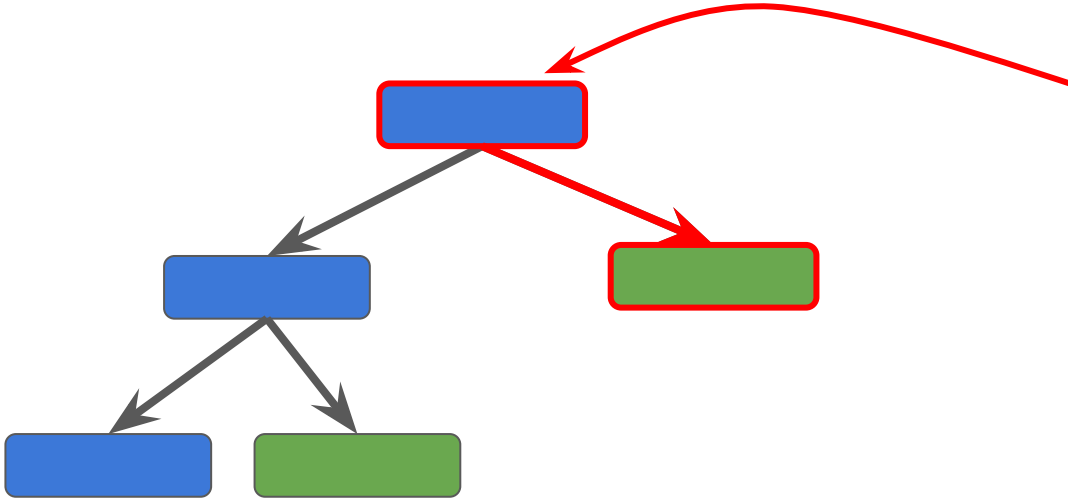
The first tree says, the passenger ‘survived’

How Random Forest work?

What does the **second** decision tree predict for an observation?

Make Prediction for this observation

class	age	alone	sex	survived
2	child	no	female	??



Survived	Not Survived
2	0

The second tree also says, the passenger ‘survived’

How Random Forest work?

After running the observation through all the trees, we see which option/label received the most votes

Make Prediction for this observation

class	age	alone	sex	survived
2	child	no	female	??

Bootstrapping the data plus **aggregating** the results to make a decision is called **bagging**

Survived	Not Survived
4	1

Choose ‘survived’ as the final prediction