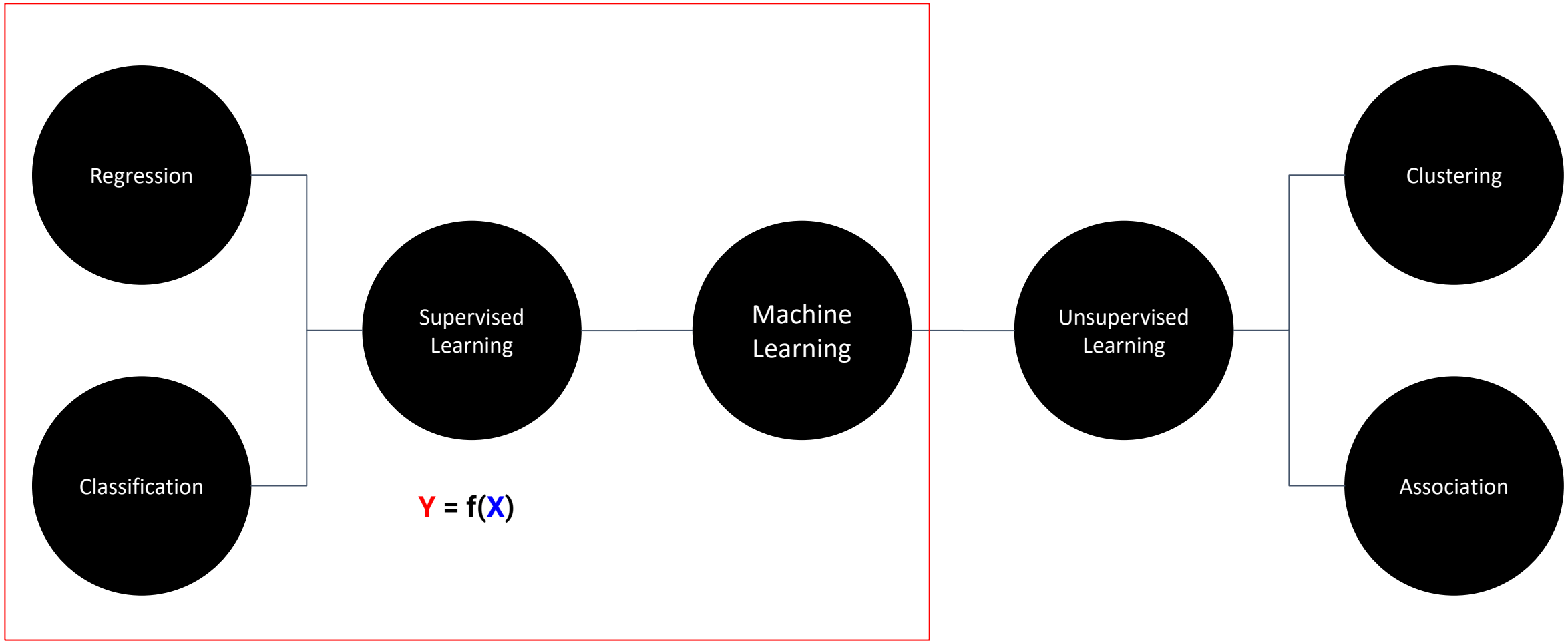
Data Prediction Model and Machine Learning

Online course #4

Classification: Random Forest


Regression
Classification
Supervised Learning
Y = f(X)
Machine Learning
Unsupervised Learning
Clustering
Association

Power of collective intelligence

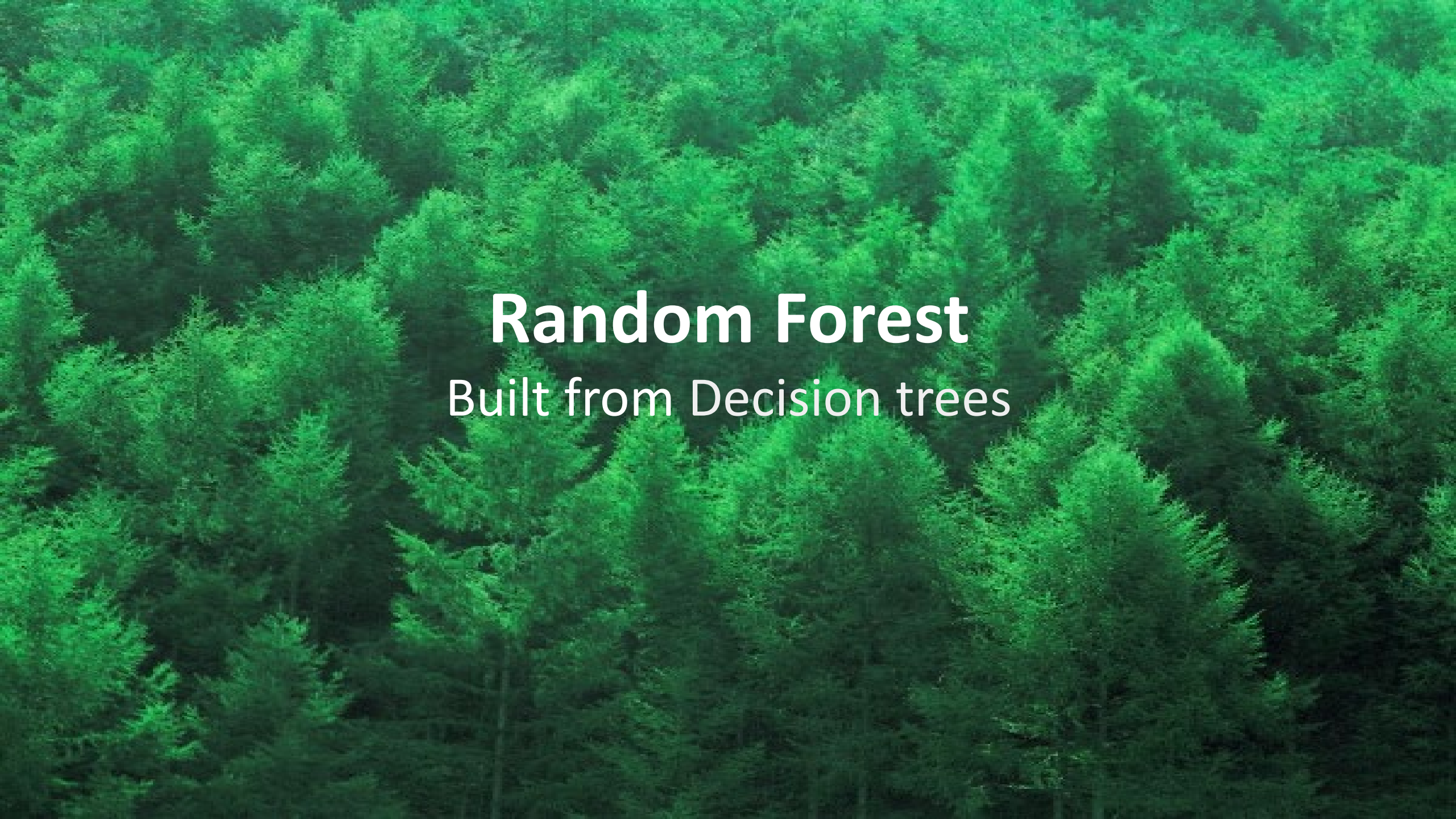
1. Each individual has a certain degree of knowledge of the problem
2. Judges independently
3. Participate seriously

→ **The more the better**

Collective intelligence >>>> One great individual

Why? How?


COLLECTIVE
©Fursys Inc. All rights reserved.



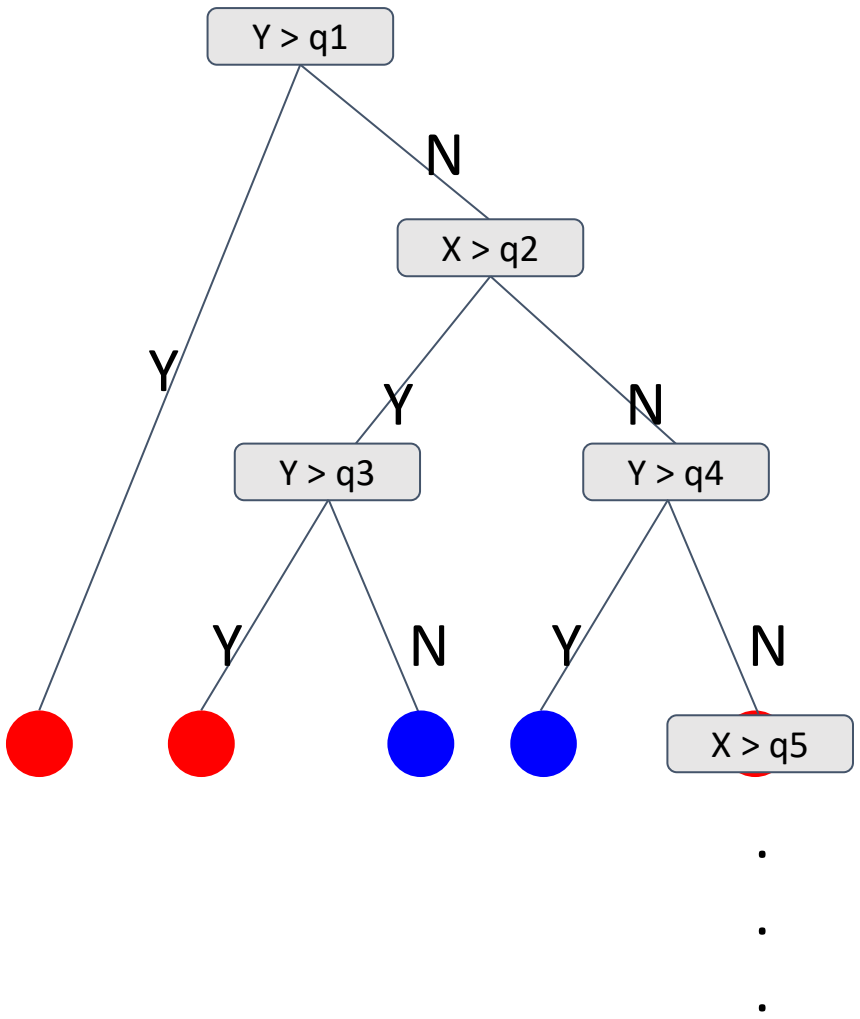
Random Forest

Built from Decision trees

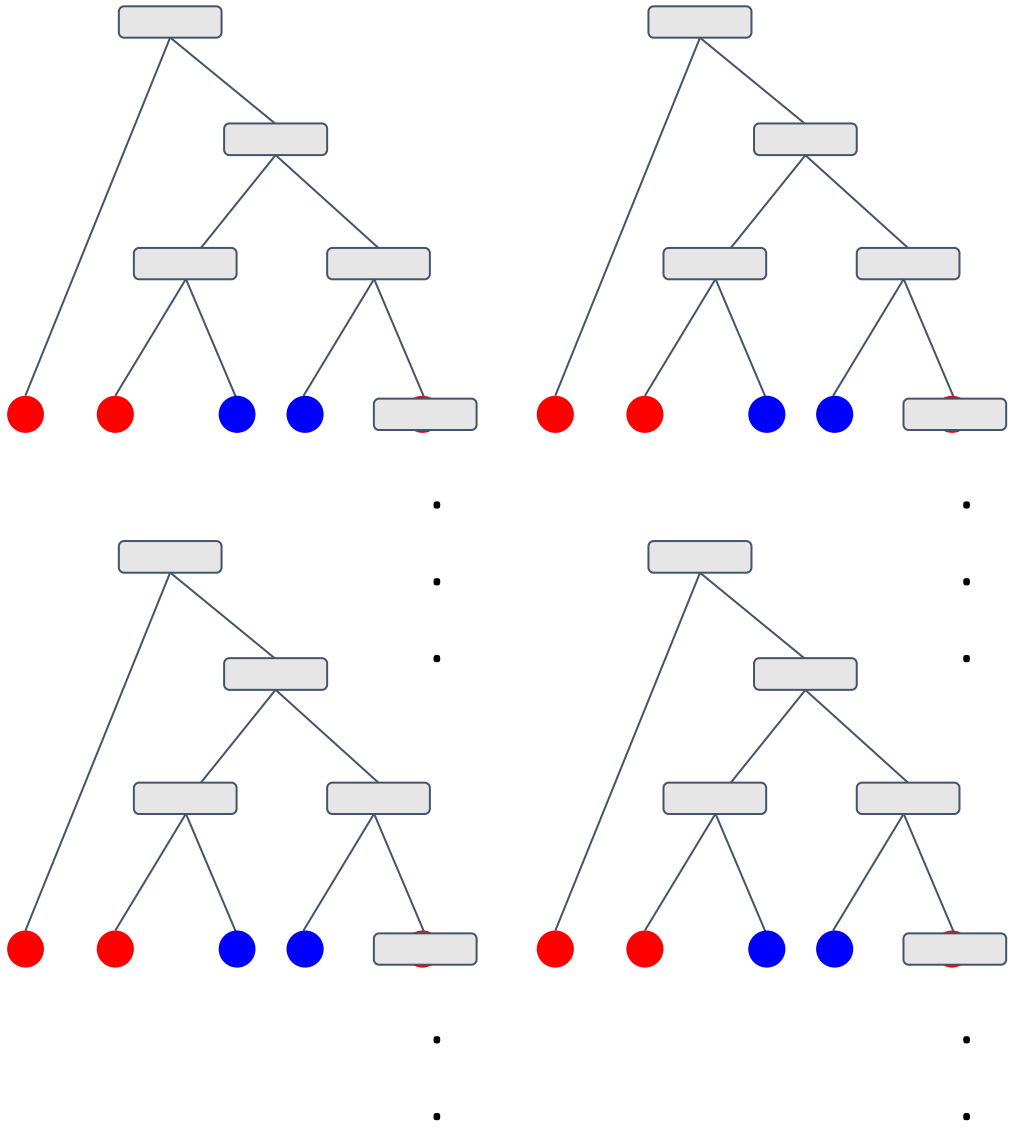
Let’s vote

(feat. The Lord of the ring)

Decision Tree



Random Forest



Step 1. Create a bootstrap data set

E.g.) Titanic dataset

Surv.	Sex	Age	Pclass	#.sib	#.par
1	F	14	1	2	2
1	F	28	2	3	0
1	M	10	1	0	2
0	M	45	3	3	1
0	M	23	2	3	0
0	M	14	1	0	0
0	F	70	2	1	0
0	M	60	2	1	0

Imagine that these 8 observations are the entire dataset to build a tree

Step 1. Create a bootstrap data set

E.g.) Titanic dataset

Surv.	Sex	Age	Pclass	#.sib	#.par
1	F	14	1	2	2
1	F	28	2	3	0
1	M	10	1	0	2
0	M	45	3	3	1
0	M	23	2	3	0
0	M	14	1	0	0
0	F	70	2	1	0
0	M	60	2	1	0

Bootstrapped Dataset

Surv.	Sex	Age	Pclass	#.sib	#.par

- To create a bootstrap data set that is the same size as the original. We just randomly select samples from the original data set.
- The important detail is that we're allowed to pick the same sample more than once

Step 1. Create a bootstrap data set

E.g.) Titanic dataset

Surv.	Sex	Age	Pclass	#.sib	#.par
1	F	14	1	2	2
1	F	28	2	3	0
1	M	10	1	0	2
0	M	45	3	3	1
0	M	23	2	3	0
0	M	14	1	0	0
0	F	70	2	1	0
0	M	60	2	1	0

first sample randomly selected.

Bootstrapped Dataset

Surv.	Sex	Age	Pclass	#.sib	#.par

Step 1. Create a bootstrap data set

E.g.) Titanic dataset

Surv.	Sex	Age	Pclass	#.sib	#.par
1	F	14	1	2	2
1	F	28	2	3	0
1	M	10	1	0	2
0	M	45	3	3	1
0	M	23	2	3	0
0	M	14	1	0	0
0	F	70	2	1	0
0	M	60	2	1	0

Bootstrapped Dataset

Surv.	Sex	Age	Pclass	#.sib	#.par
1	M	10	1	0	2

Step 1. Create a bootstrap data set

E.g.) Titanic dataset

Surv.	Sex	Age	Pclass	#.sib	#.par
1	F	14	1	2	2
1	F	28	2	3	0
1	M	10	1	0	2
0	M	45	3	3	1
0	M	23	2	3	0
0	M	14	1	0	0
0	F	70	2	1	0
0	M	60	2	1	0

Bootstrapped Dataset

Surv.	Sex	Age	Pclass	#.sib	#.par
1	M	10	1	0	2
0	M	23	2	3	0

Step 1. Create a bootstrap data set

E.g.) Titanic dataset

Surv.	Sex	Age	Pclass	#.sib	#.par
1	F	14	1	2	2
1	F	28	2	3	0
1	M	10	1	0	2
0	M	45	3	3	1
0	M	23	2	3	0
0	M	14	1	0	0
0	F	70	2	1	0
0	M	60	2	1	0

Bootstrapped Dataset

Surv.	Sex	Age	Pclass	#.sib	#.par
1	M	10	1	0	2
0	M	23	2	3	0
0	M	14	1	0	0
1	F	28	2	3	0
1	F	14	1	2	2
1	F	28	2	3	0
0	M	14	1	0	0
0	M	60	2	1	0

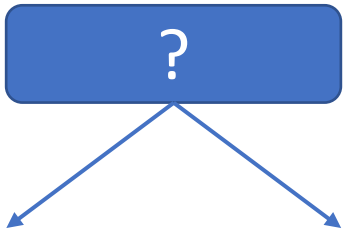
Step 2. Create a decision tree using the bootstrapped dataset, but only use a random subset of variables (or columns) at each step

- Consider two variables or columns at each step!

Bootstrapped Dataset

Surv.	Sex	Age	Pclass	#.sib	#.par
1	M	10	1	0	2
0	M	23	2	3	0
0	M	14	1	0	0
1	F	28	2	3	0
1	F	14	1	2	2
1	F	28	2	3	0
0	M	14	1	0	0
0	M	60	2	1	0

Step 2. Create a decision tree using the bootstrapped dataset, but only use a random subset of variables (or columns) at each step

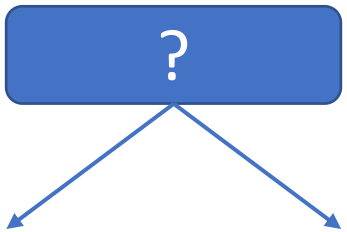


Instead of considering all 5 variables

Bootstrapped Dataset

Surv.	Sex	Age	Pclass	#.sib	#.par
1	M	10	1	0	2
0	M	23	2	3	0
0	M	14	1	0	0
1	F	28	2	3	0
1	F	14	1	2	2
1	F	28	2	3	0
0	M	14	1	0	0
0	M	60	2	1	0

Step 2. Create a decision tree using the bootstrapped dataset, but only use a random subset of variables (or columns) at each step

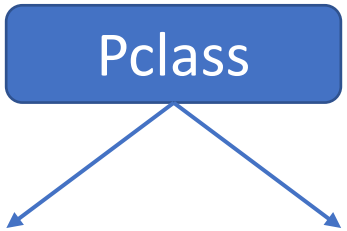


Randomly select two

Bootstrapped Dataset

Surv.	Sex	Age	Pclass	#.sib	#.par
1	M	10	1	0	2
0	M	23	2	3	0
0	M	14	1	0	0
1	F	28	2	3	0
1	F	14	1	2	2
1	F	28	2	3	0
0	M	14	1	0	0
0	M	60	2	1	0

Step 2. Create a decision tree using the bootstrapped dataset, but only use a random subset of variables (or columns) at each step

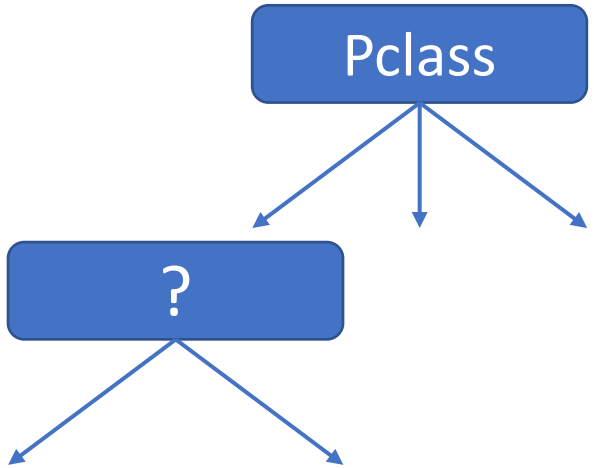


Randomly select two

Bootstrapped Dataset

Surv.	Sex	Age	Pclass	#.sib	#.par
1	M	10	1	0	2
0	M	23	2	3	0
0	M	14	1	0	0
1	F	28	2	3	0
1	F	14	1	2	2
1	F	28	2	3	0
0	M	14	1	0	0
0	M	60	2	1	0

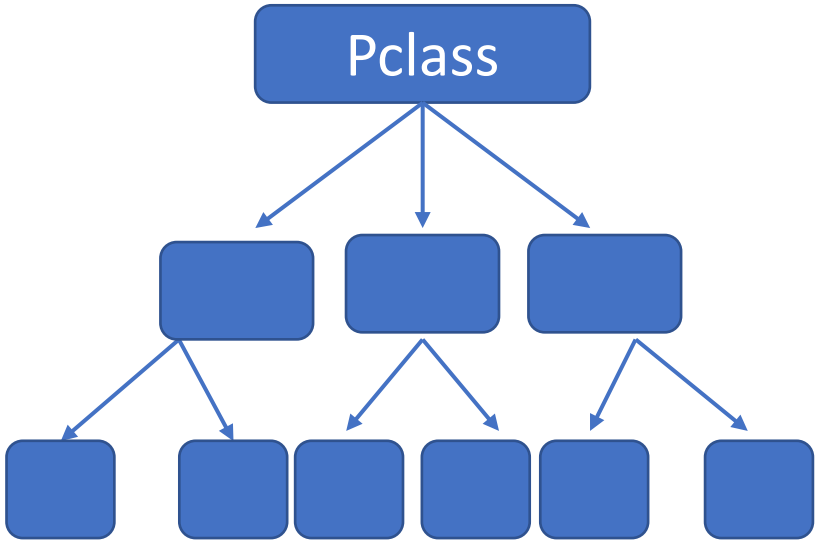
Step 2. Create a decision tree using the bootstrapped dataset, but only use a random subset of variables (or columns) at each step



Bootstrapped Dataset

Surv.	Sex	Age	Pclass	#.sib	#.par
1	M	10	1	0	2
0	M	23	2	3	0
0	M	14	1	0	0
1	F	28	2	3	0
1	F	14	1	2	2
1	F	28	2	3	0
0	M	14	1	0	0
0	M	60	2	1	0

Step 2. Create a decision tree using the bootstrapped dataset, but only use a random subset of variables (or columns) at each step

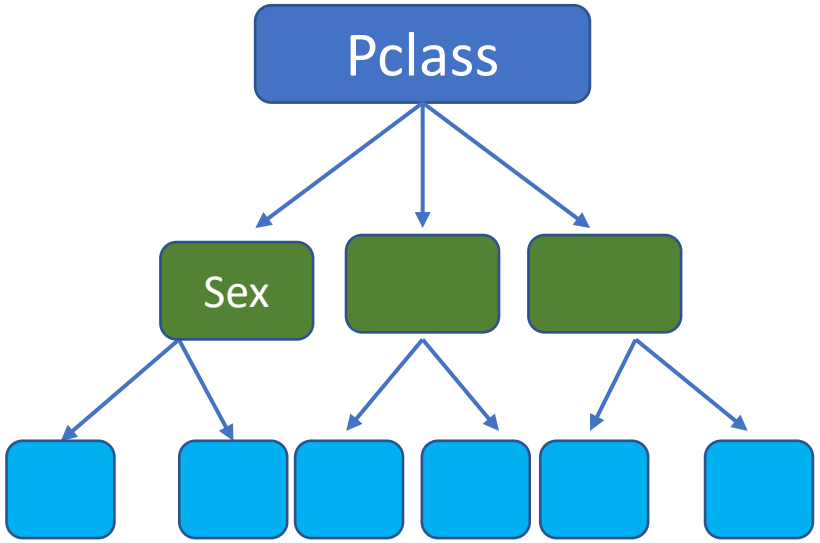


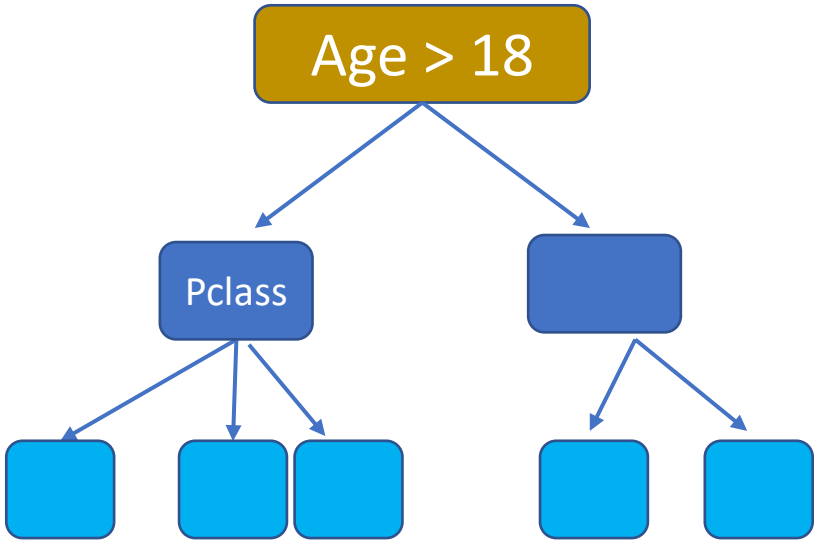
Bootstrapped Dataset

Surv.	Sex	Age	Pclass	#.sib	#.par
1	M	10	1	0	2
0	M	23	2	3	0
0	M	14	1	0	0
1	F	28	2	3	0
1	F	14	1	2	2
1	F	28	2	3	0
0	M	14	1	0	0
0	M	60	2	1	0

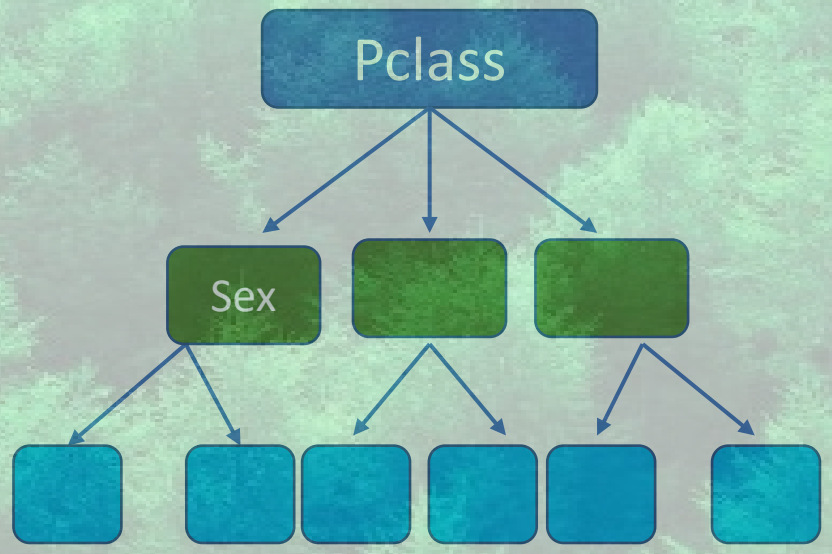
- Using a bootstrap data set
- Only considering a random subset of variables at each step

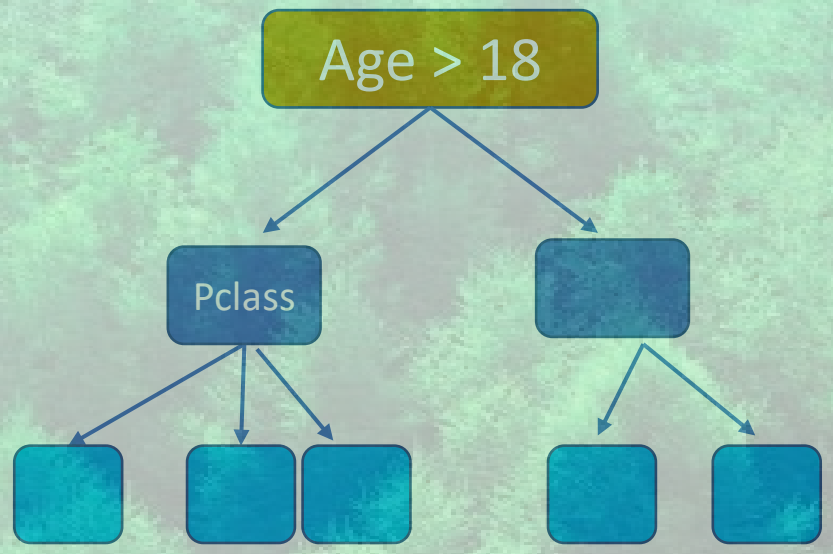
Step 3. Go back to the step 1 and repeat

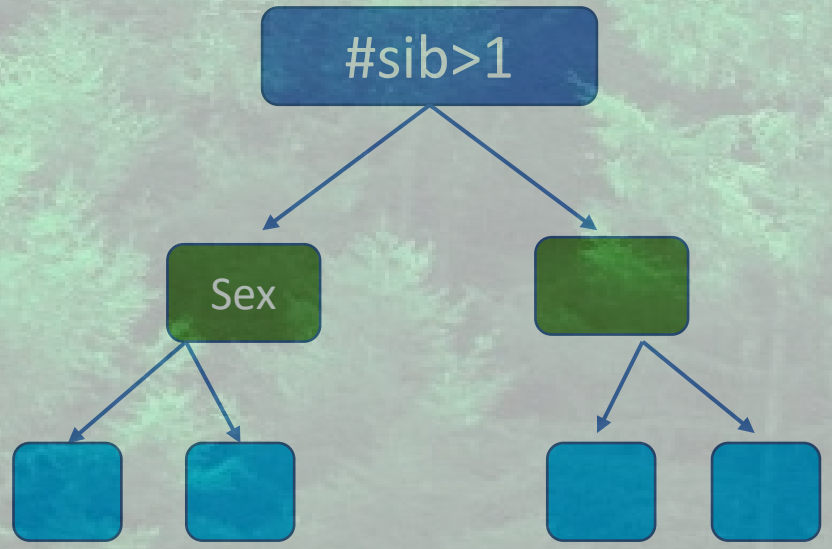


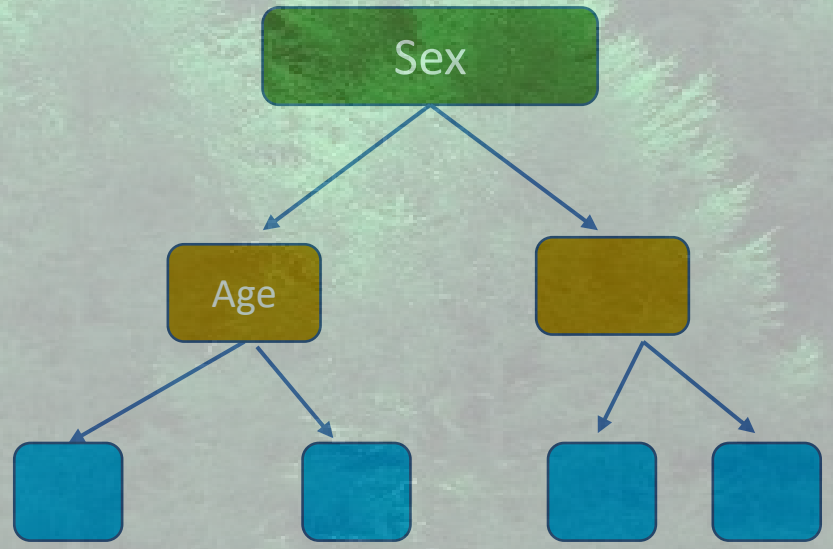


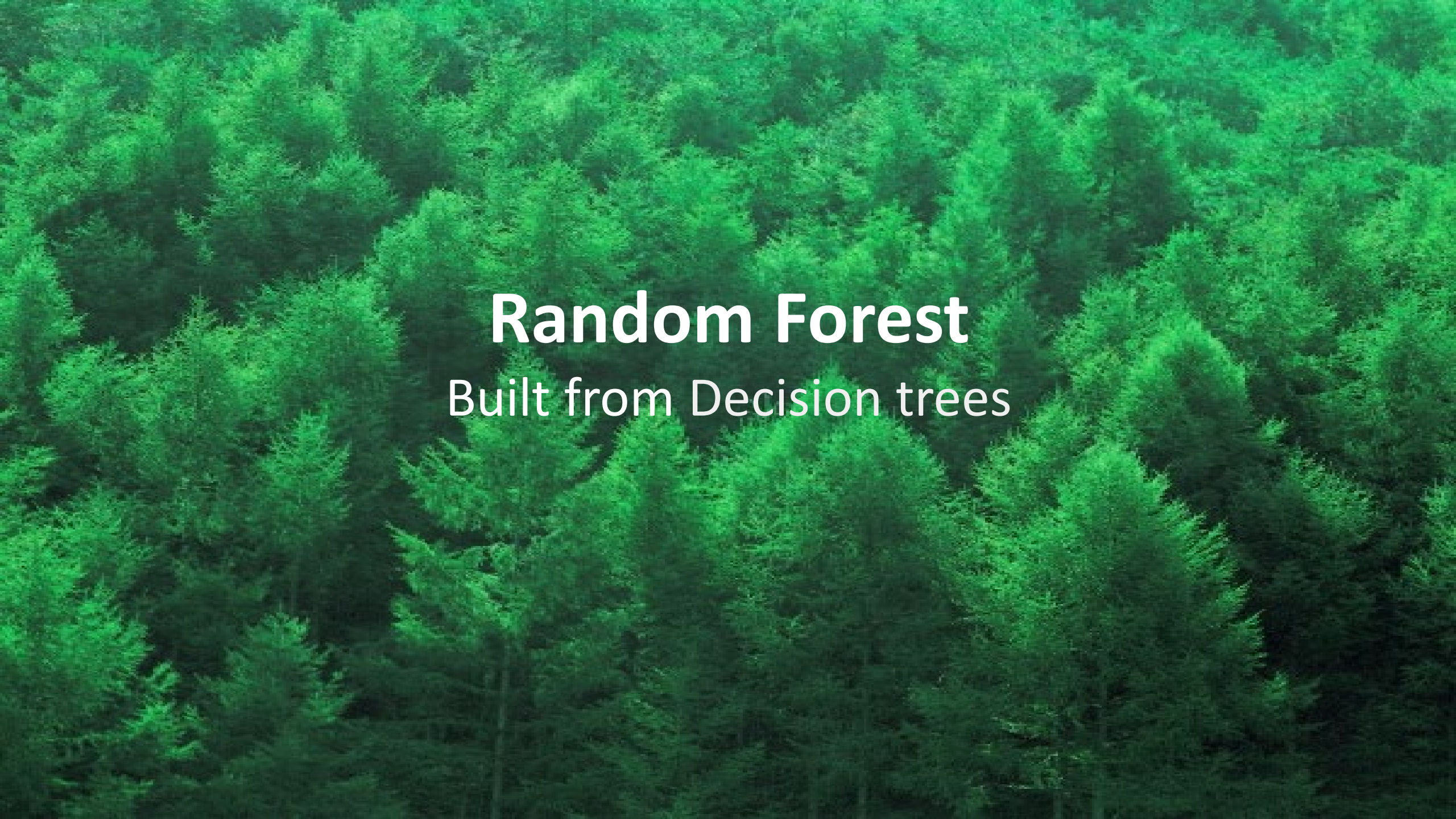
Step 3. Go back to the step 1 and repeat







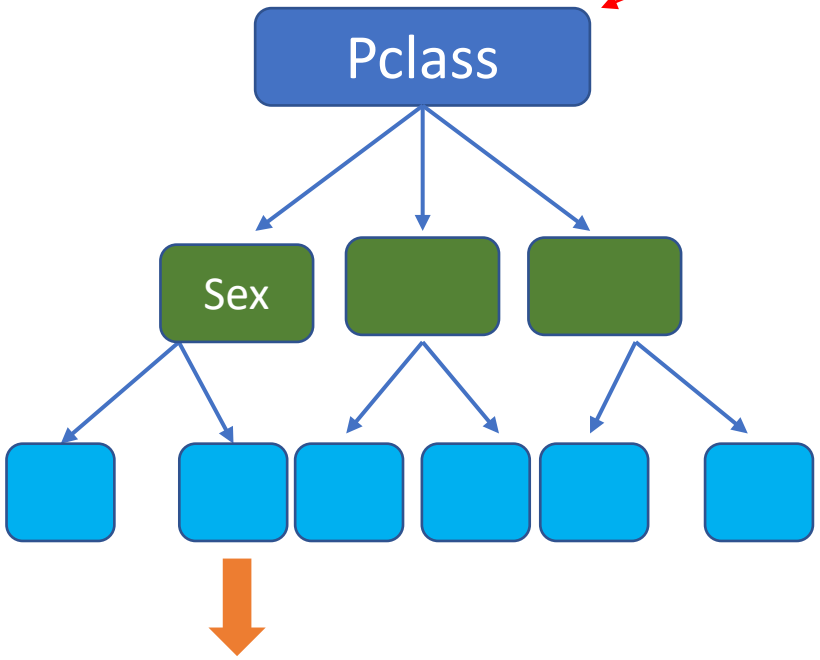




Random Forest

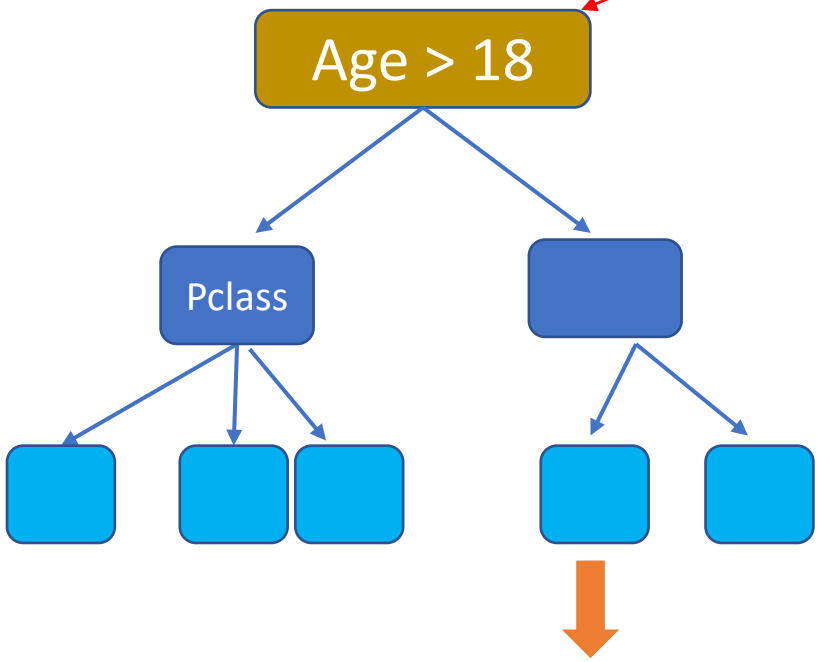
Built from Decision trees

Prediction



Surv.	Sex	Age	Pclass	#.sib	#.par
	F	14	1	2	2

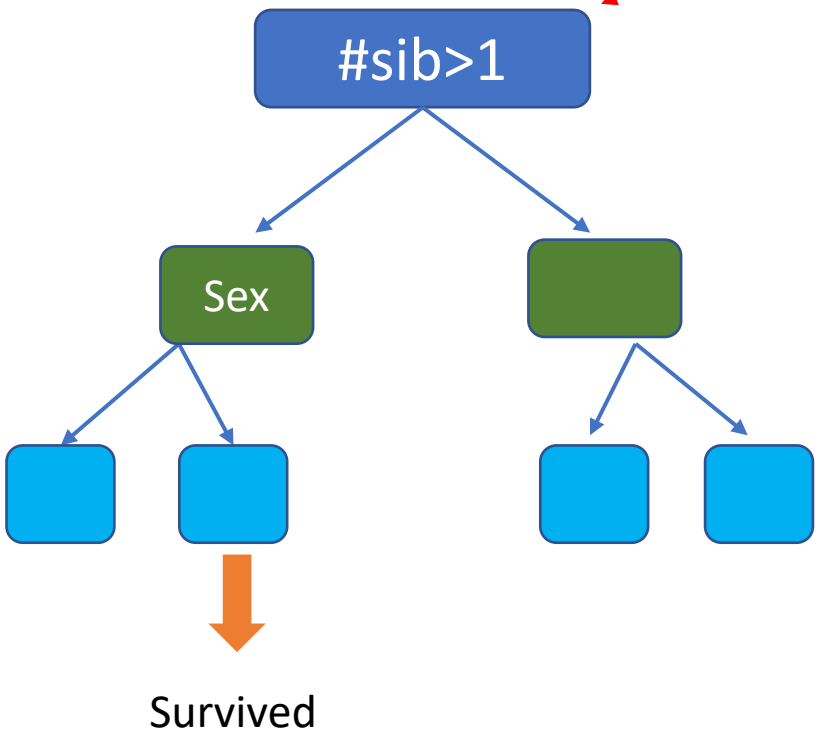
Prediction



Survived

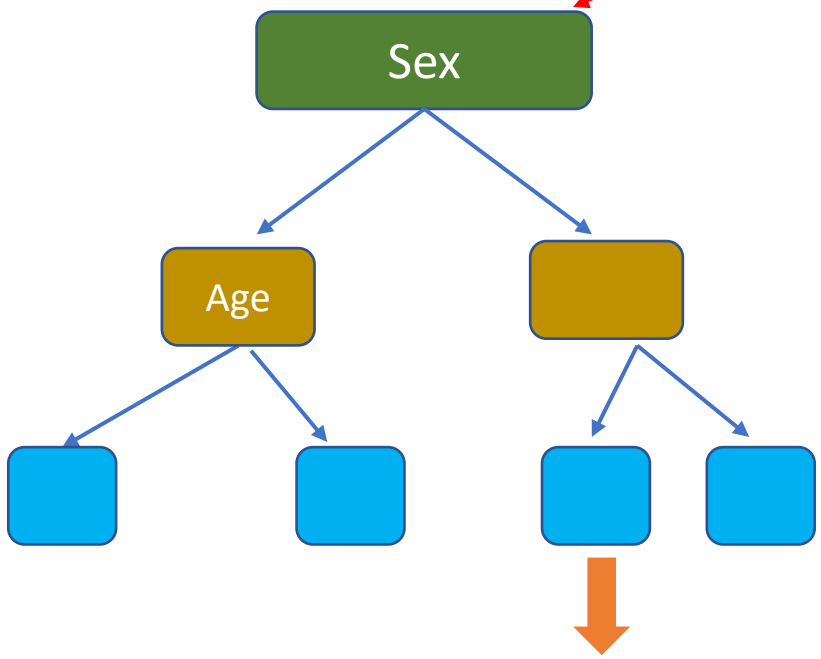
Surv.	Sex	Age	Pclass	#.sib	#.par
	F	14	1	2	2

Prediction



Surv.	Sex	Age	Pclass	#.sib	#.par
	F	14	1	2	2

Prediction



Surv.	Sex	Age	Pclass	#.sib	#.par
1	F	14	1	2	2

Not survived

By Majority rule

(Survived: 3 vs. Not survived: 1)

→ **Conclusion: Survived!**

Trees vs. Random Forest

Trees	Random Forest
Yield insight into decision rules	Has smaller prediction variance and therefore usually a better general performance
Rather fast	Easy to tune parameters
Easy to tune parameters	**(Cons)** Rather slow
(Cons) Prediction of trees tend to have a high variance	**(Cons)** Black box: rather difficult to get insights into decision rules