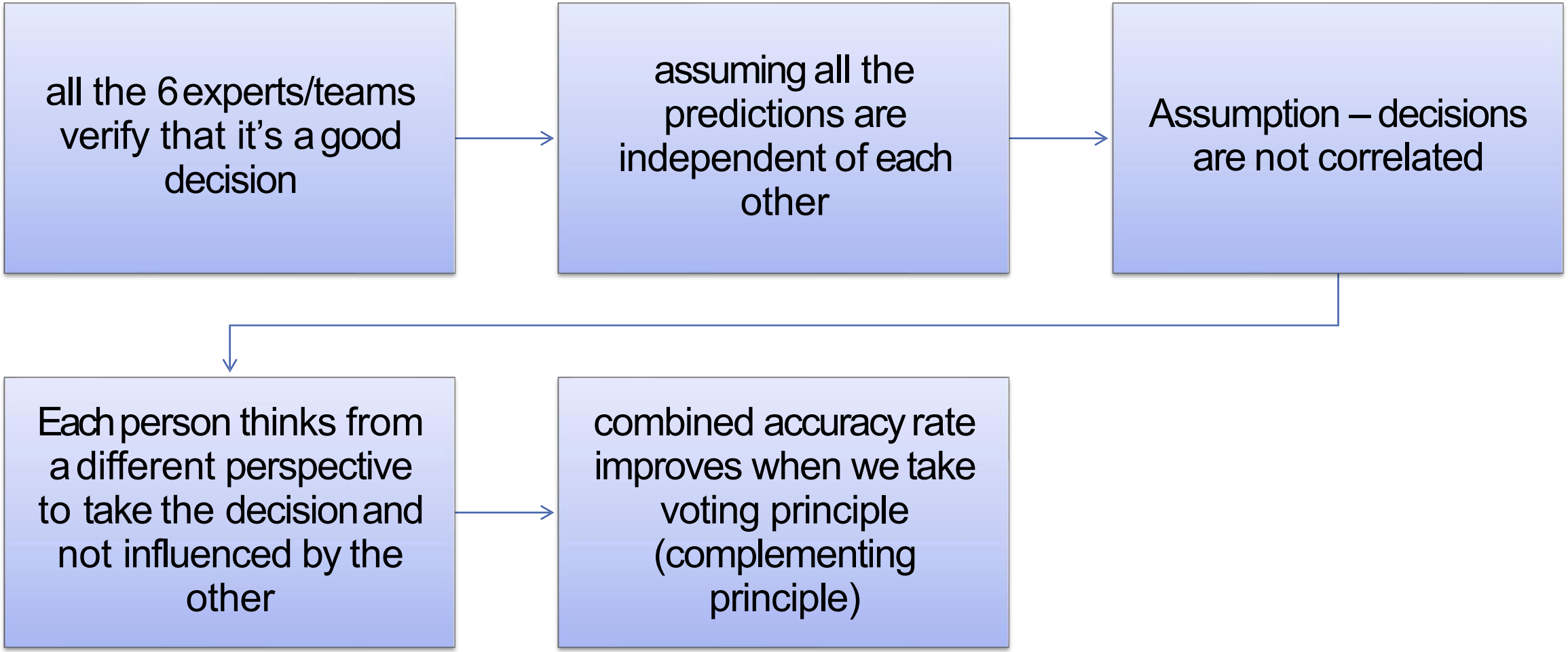
Random Forest

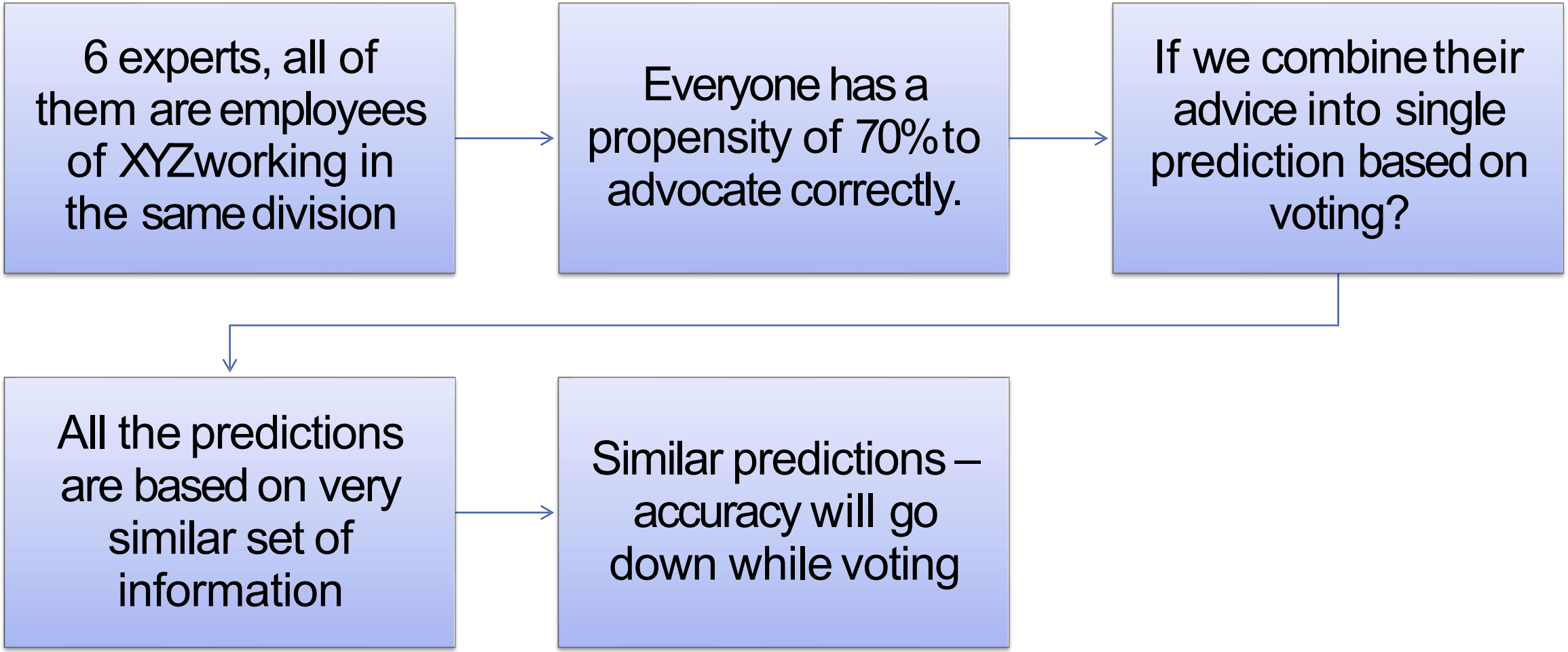
Basic steps - Classification algorithms

Should I invest in a company – ask the experts

Scenario1 - Combine all the info – informed decision



Scenario 2 – info from similar sources



Ensemble learning

- Machine learning technique that combines several base models in order to produce one optimal predictive model.
- Weak classifiers
- Different set of variables for each classifier
- Combine into single prediction

What is a boot strapped dataset

Sno	X1	X2	Y
1	432	29	Yes
2	529	34	Yes
3	125	67	No
4	144	29	No

Random sample rows **with replacement**

Sno	X1	X2	Y
4	144	29	No
2	529	34	Yes
3	125	67	No

Sno	X1	X2	Y
3	125	67	No
4	144	29	No
4	144	29	No

Sno	X1	X2	Y
3	125	67	No
2	529	34	Yes
3	125	67	No

Using a random set of variables every time

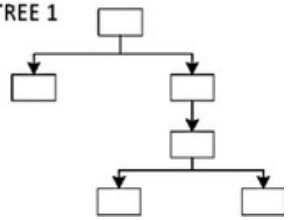
Sno	X1	X2	X3	X4	Y
1	432	29	313	6	Yes
2	529	34	379	2	Yes
3	125	67	317	4	No
4	144	29	103	8	No

Random sample rows **with replacement**

Random subset of X variables

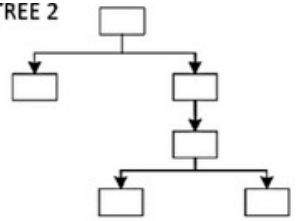
Sno	X1	X2	Y
4	144	29	No
2	529	34	Yes
3	125	67	No

TREE 1



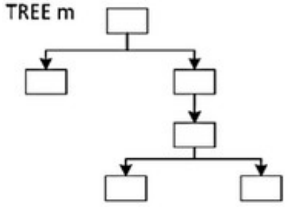
Sno	X3	X4	Y
3	317	4	No
4	103	8	No
4	103	8	No

TREE 2

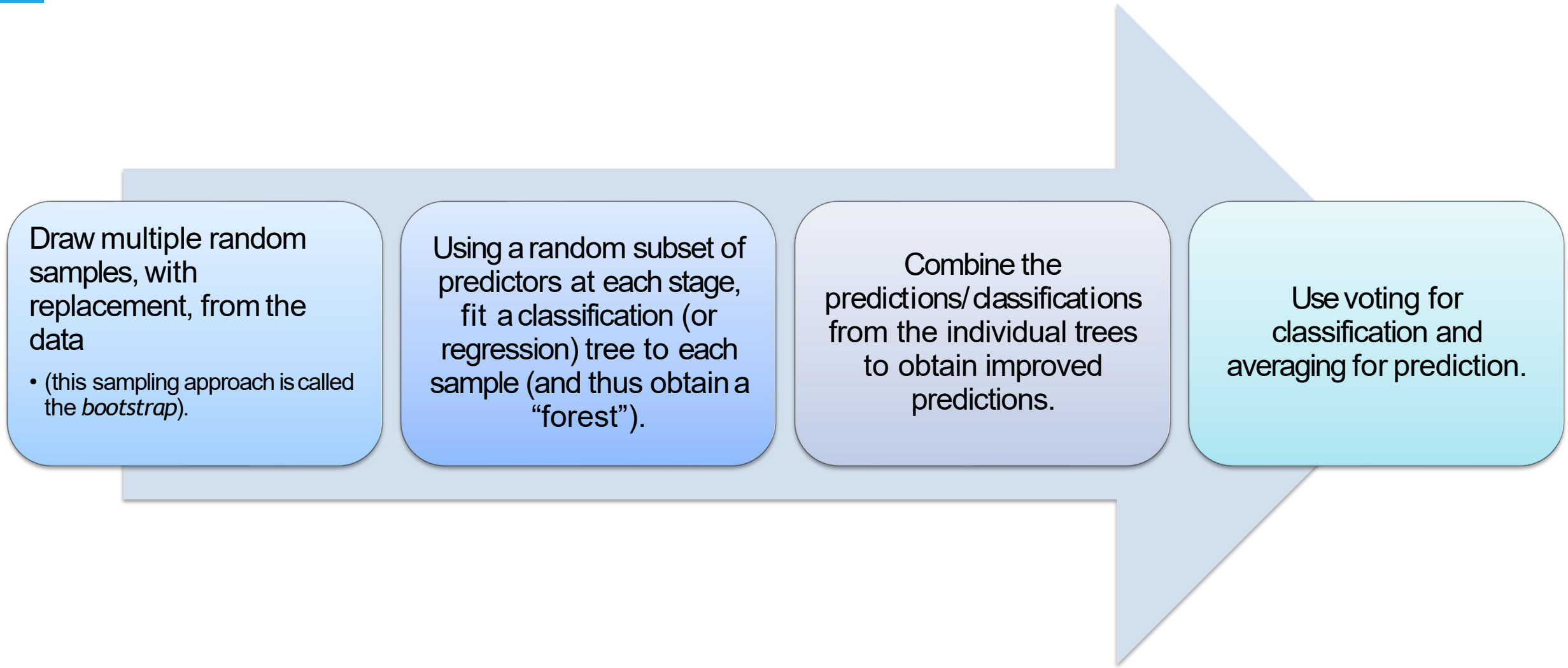


Sno	X1	X3	Y
3	125	317	No
2	529	379	Yes
3	125	317	No

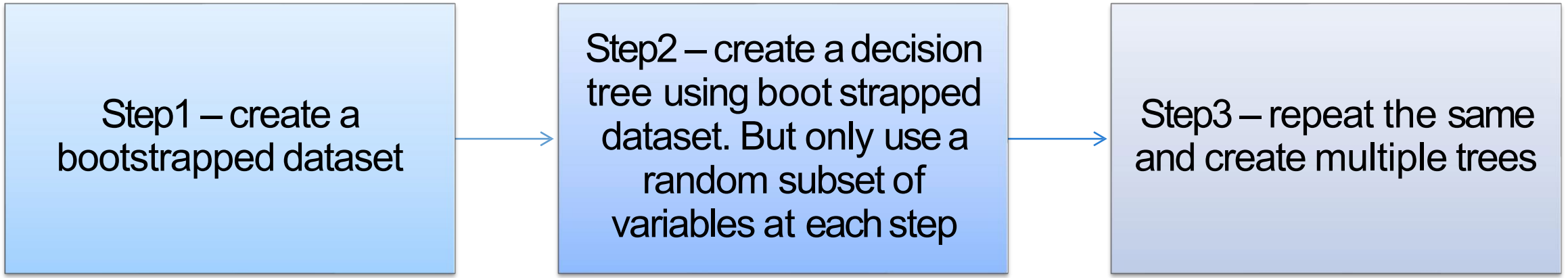
TREE m



Basic idea of random forest



Steps in random forest algorithm



Out of bag data points

Sno	X1	X2	X3	X4	Y
1	432	29	313	6	Yes
2	529	34	379	2	Yes
3	125	67	317	4	No
4	144	29	103	8	No

Sno	X1	X2	Y
4	144	29	No
2	529	34	Yes
3	125	67	No

Sno	X3	X4	Y
3	317	4	No
4	103	8	No
4	103	8	No

Sno	X1	X3	Y
3	125	317	No
2	529	379	Yes
3	125	317	No

- When we create a bootstrapped dataset, ~1/3 of the original data does not end up in the boot strapped dataset
- This is called out-of-bag dataset

How to calculate accuracy

- OOB samples used to measure how accurate our random forest is
- by the ratio of out of bag samples correctly classified by the random forest model
- Proportion of OOB samples incorrectly classified – out of bag error

How to decide on how many variables to use per step?

- Compare OOB error for using 2 variables per step, 3 variables and so on
- Choose the most accurate set of variables
- Typically we start by using square root of number of variables
- Then try a few settings above and below the value

Summary of Random forest

Consists of a large number of individual decision trees that operate as an ensemble

Each tree in the random forest spits out a class prediction

class with most votes becomes model's prediction

fundamental concept - wisdom of crowds

A large number of relatively uncorrelated models (trees) operating as a committee will outperform any of the individual models.

Overall flow of the RF classification process