In [1]:
```
from IPython.display import Image
Image(filename='logo.PNG', height=340, width=900)
```

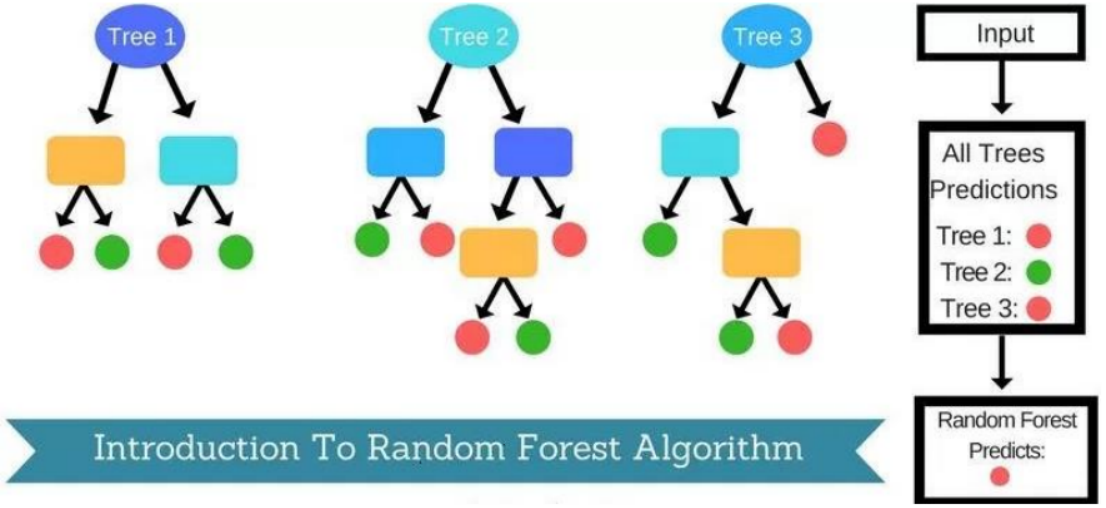
Out[1]:

RANDOM FOREST CLASSIFICATION - INTRODUCTION

Random forest algorithm is a supervised classification algorithm. As the name suggest, this algorithm creates the forest with a number of trees

In [2]:
```
Image(filename='RFIntro.PNG', height=340, width=900)
```

Out[2]:



REAL LIFE EXAMPLE:

Suppose Aanya decides to buy a house and settle in one of these cities – Mumbai, Delhi or Bangalore.

She wants to make a right and optimal choice considering all the perspectives since it is a very important decision for her.

So she decided to ask her best friend about the places she may like. Then her friend started asking about her opinions. It's just like her best friend will ask, You have visited X city. Did you like it?, etc.

Based on the answers which are given by Aanya, her best friend will start recommending the place she may like. Here her best friend forms the decision tree with the answer given by Aanya.

As Aanya's best friend may recommend HER the best place by virtue of being her friend. The model will be biased with the closeness of their friendship. So she decided to ask few more friends to recommend the best place she may like.

Now her friends asked some random questions and each one recommended one place to Aanya. Now Aanya considered the place which has highest votes from her friends as the final place to settle down.

In the above decision process, two main interesting algorithms decision tree algorithm and random forest algorithm are used.

-- DECISION TREE APPROACH To recommend the best place to Aanya, her best friend asked some questions. Based on the answers given by Aanya, she recommended a place. This is decision tree algorithm approach.

Here Aanya's best friend is the decision tree. The vote (recommended place) is the leaf of the decision tree (Target class). The target is finalized by a single person. In a technical way of saying, using an only single decision tree.

-- RANDOM FOREST APPROACH In this case when Aanya asked her friends to recommend the best place to settle among the three. Each friend asked her different questions and came up with their recommendation of a place.

Later Aanya considered all the recommendations and calculated the votes. Votes basically, to pick the popular place from the recommended places from all her friends. Here, each friend is the tree and the combination of all friends (trees) will form the forest. This forest is the random forest, as each friend asked random questions to recommend the best place to settle down among the three.

Random Forest Algorithm - Working

1. Assume number of cases in the training set is **N**. Then, sample of these N cases is taken at random but with replacement. This sample will be the training set for growing the tree
2. If there are **M** input variables, a number **m<M** is specified such that at each node, **m** variables are selected at random out of the M. The best split on these m is used to split the node. The value of m is held constant while we grow the forest. **three possible values for m: ½√M, √M, and 2√M**

3. Each tree is grown to the largest extent possible. Predict new data by aggregating the predictions of the ntree trees (i.e., majority votes for classification, average for regression).

Advantages of Random Forest

1. The same random forest algorithm or the random forest classifier can use for both classification and the regression task
2. Random forest classifier will handle the missing values
3. It doesn't overfit the model
4. Can model the random forest classifier for categorical values also

In [ ]: