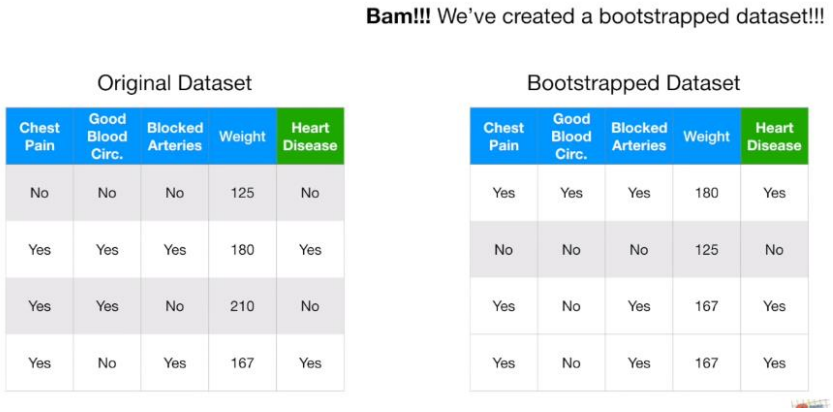
## Random Forest

在一般 Classification And Regression Tree 使用时，CART 能够对训练数据有很好的表现，但是 CART 不够 Flexible 当遇见新的数据。Random Forest 有着 CART 的简单性，也有着更好的表现在没看过的数据里。

### Bootstrapped Dataset

在 Random Forest 里面，使用了一个叫 Bootstrapped 的概念。Bootstrapped 就是从原来的 Dataset 里随机选出新的一组有着同样数目的 Dataset，这个新的 Dataset 是可以有重复的。



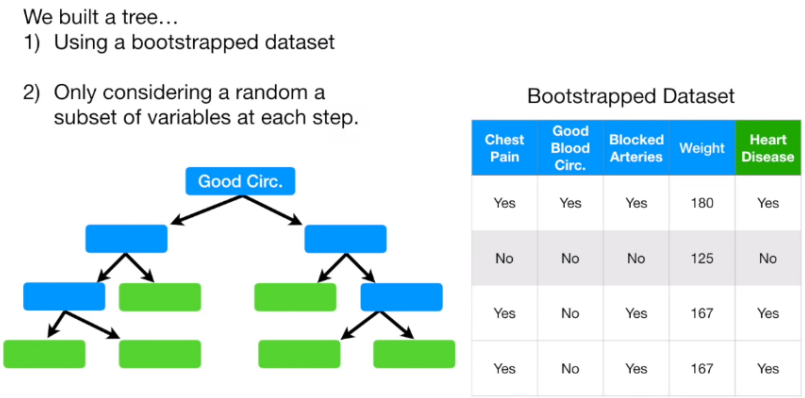

| Chest Pain | Good Blood Circ. | Blocked Arteries | Weight | Heart Disease |
|---|---|---|---|---|
| No | No | No | 125 | No |
| Yes | Yes | Yes | 180 | Yes |
| Yes | Yes | No | 210 | No |
| Yes | No | Yes | 167 | Yes |

| Chest Pain | Good Blood Circ. | Blocked Arteries | Weight | Heart Disease |
|---|---|---|---|---|
| Yes | Yes | Yes | 180 | Yes |
| No | No | No | 125 | No |
| Yes | No | Yes | 167 | Yes |
| Yes | No | Yes | 167 | Yes |

要使用 Random Forest，通常都会 Bootstrapped 很多比新的 Data 出来训练出新的树。

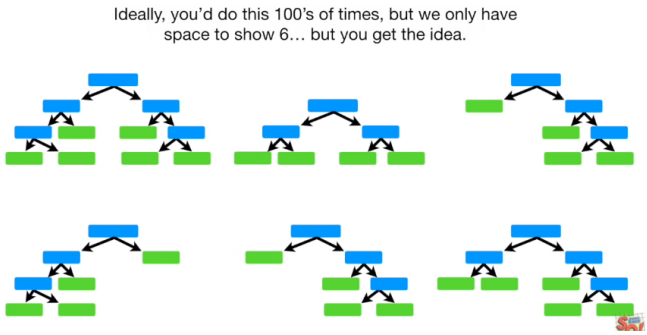
### Create a Tree

当已经 Bootstrapped 好新的一组 Dataset 之后，就可以开始训练 Decision Tree。在 Random Forest 的训练里，在对 Node 做出 Threshold 设定的时候，不会使用所有 Parameters 来做选择 (Gini Impurity)，而是设定好要选取几个 Parameters 之后，在 Bootstrapped Dataset 里面随机选取定好的 Parameters 数目。当选出 Root 之后，再 Random 选择实现定好的 Parameters 数目从没被选到的所有 Parameters 里面。 直到这个树完整建立好。



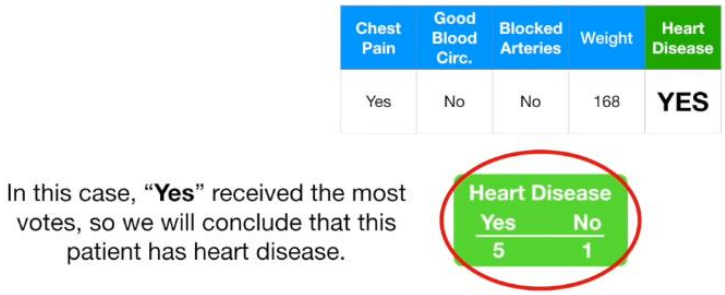

| Chest Pain | Good Blood Circ. | Blocked Arteries | Weight | Heart Disease |
|---|---|---|---|---|
| Yes | Yes | Yes | 180 | Yes |
| No | No | No | 125 | No |
| Yes | No | Yes | 167 | Yes |
| Yes | No | Yes | 167 | Yes |

当第一颗树建立好后，继续建立其他的树，使用同样的方法。




通常建立的树是比较多的，这也是 Random Forest 的特点。

**Random Forest 运作**




当需要对一笔 Data 进行预测的时候，将这笔 Data Feed 进每一刻 Generate 好的树，然后把全部 Prediction 总结，选出被预测最多的答案。

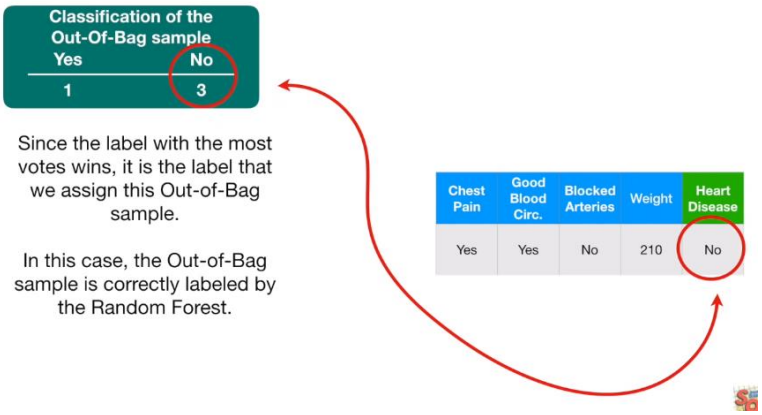
在这里使用 **B**ootstrapping 和 **Agg**regate 来做出选择，这个就是 Bagging。

**测试 Random Forest**

收集在 Bootstrapped Dataset 里没被选到的所有 Dataset (Out-of-Bag Dataset)，然后对其树进行测试


上图没有在 Bootstrap 里的 data 只有一个，但是当 Dataset 大的时候，会有比较多 Data 不在 Bootstrapped dataset 里。

会对每一个树使用 Out-of-Bag Dataset 来测试。




当所有 Out-of-Bag Dataset 都已经测试完后，就能计算有多少比 Dataset 是被预测错误，也就是 Out-of-Bag Error。

Parameters 数目选择

选择 Number of Parameters 通常是 2^N，一般选择几个不同的 Number of Parameters 之后，看看哪一个表现比较好，然后使用表现最好的 Number of Parameters。