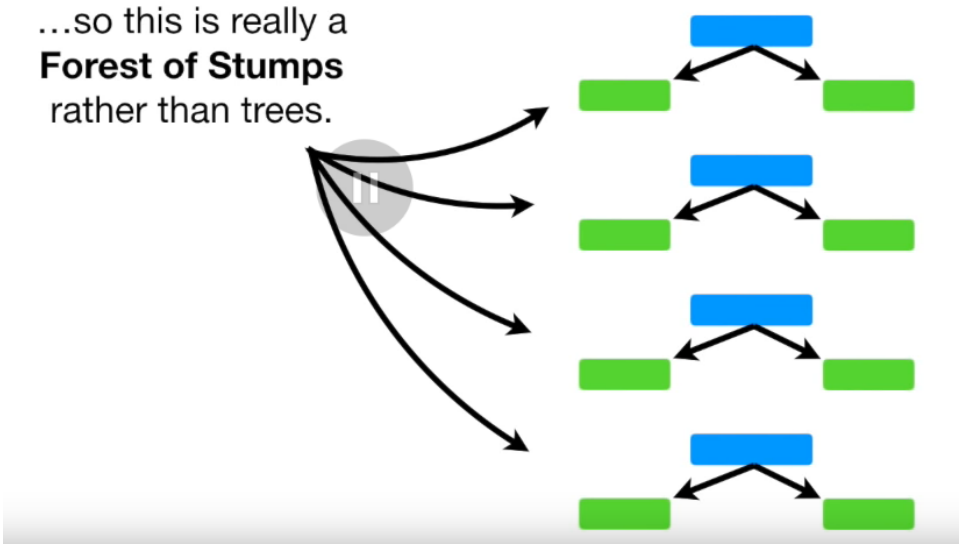
Random Forest和AdaBoost的区别：

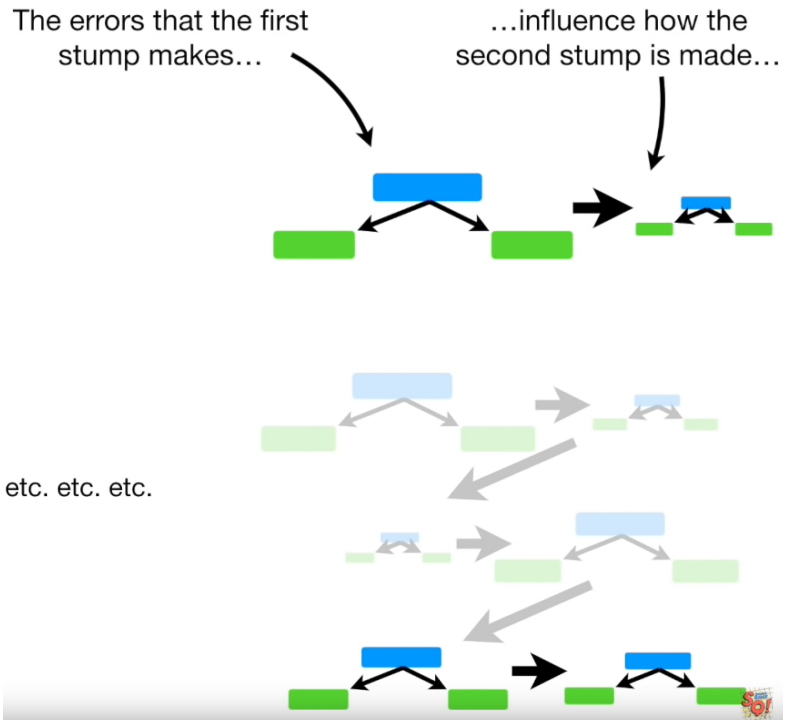
1. Random Forest中，每个树的深度是不固定的。在AdaBoost中，每个树均为只有两个叶节点的二叉树（Stump，树桩）。

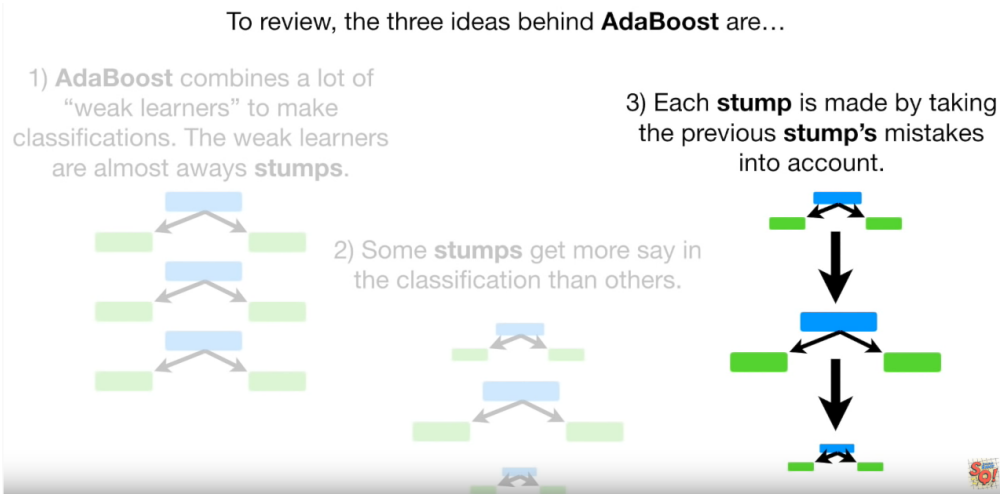


2.单个的stump对分类效果不好。因为每次只考虑一个特征。而决策树可以考虑多个特征。因此，stump是弱学习器（weak learner）。

在random forest中，每个决策树拥有相同的“话语权”（即权重）。而在AdaBoost中，某些stump权重会比其他大/小。

3. Random Forest中每个决策树是相互独立的，决策树构建的先后顺序没有影响。在AdaBoost中，先建立的stump的误差会对后一个stump产生影响。





AdaBoost原理

1.给数据集中的每个样本赋予weight

Chest Pain	Blocked Arteries	Patient Weight	Heart Disease	Sample Weight
Yes	Yes	205	Yes	1/8
No	Yes	180	Yes	1/8
Yes	No	210	Yes	1/8
Yes	Yes	167	Yes	1/8
No	Yes	156	No	1/8
No	Yes	125	No	1/8
Yes	No	168	No	1/8
Yes	Yes	172	No	1/8

At the start, all samples get the same weight…

$$\frac{1}{\text{total number of samples}} = \frac{1}{8}$$

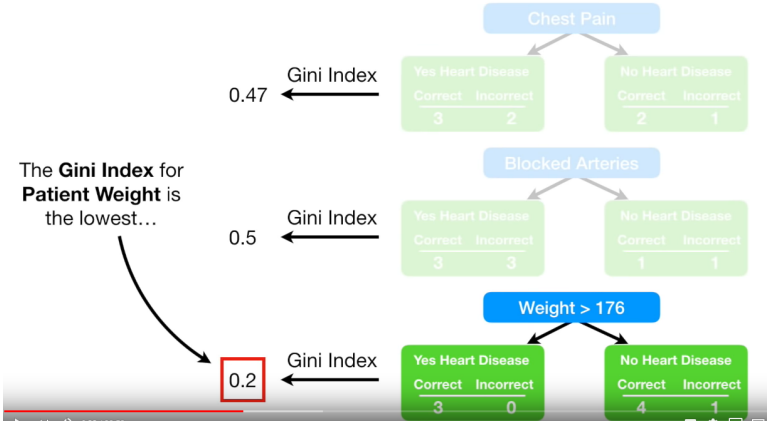
…and that makes the samples all equally important.

However, after we make the first stump, these weights will change in order to guide how the next stump is created.

In other words, we'll talk more about the **Sample Weights** later!

2.构建第一个stump

*因为每个sample的初始weight都是一样的，可以先忽略sample weight的影响。



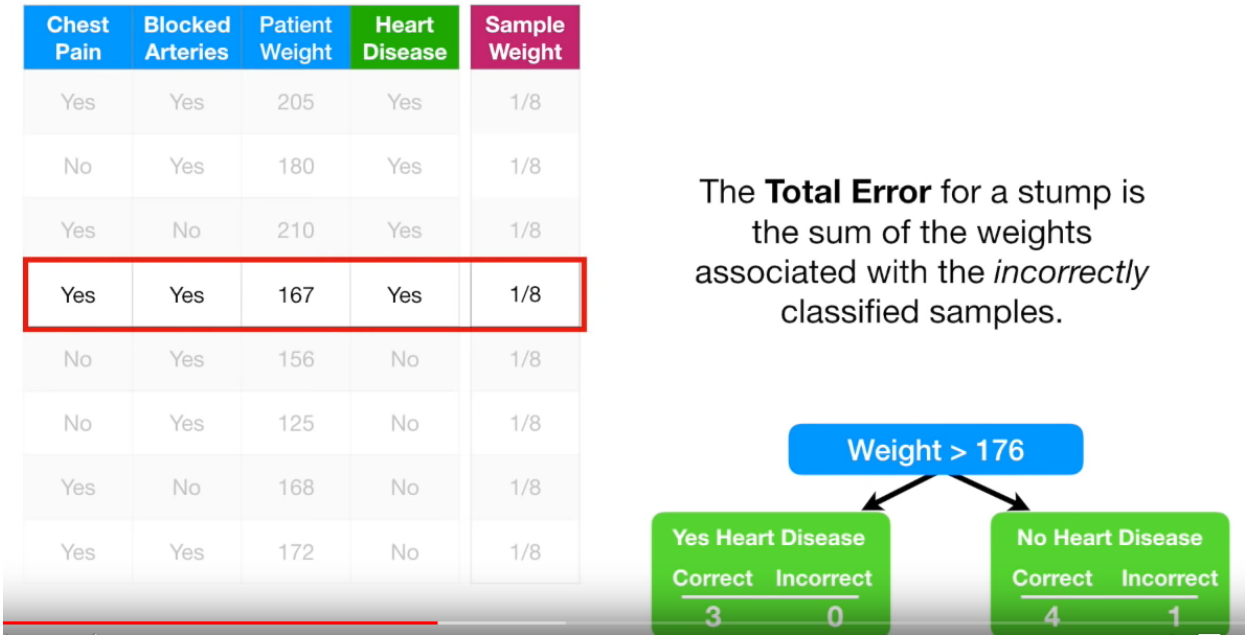
（1）计算每个feature的gini值

（2）选择Gini值最小的feature作为第一个stump的分裂依据

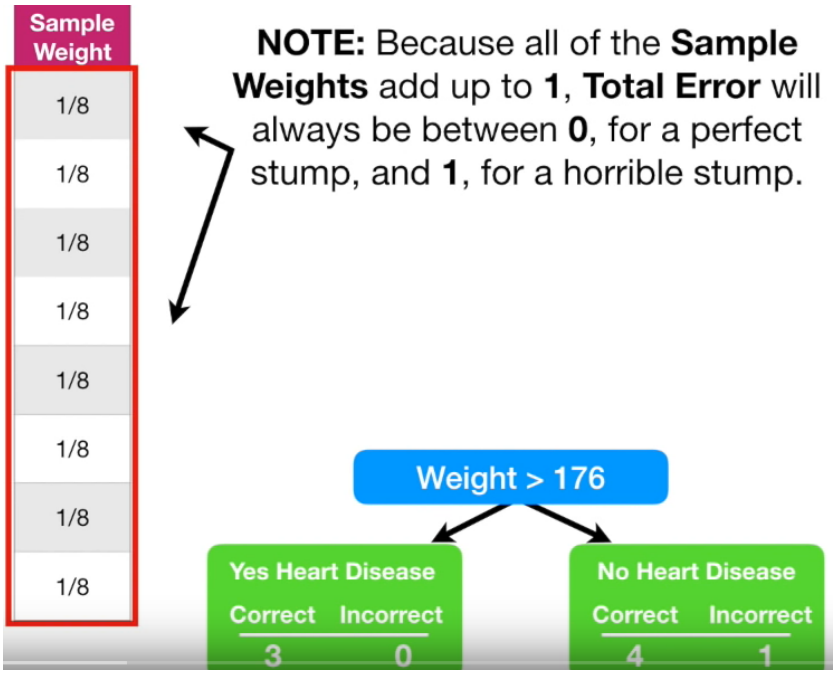
Weight>176的Gini值最小，选择此feature作为第一个stump（Weight>176是根据决策树章节中的相应内容决定的，即升序排序后，求相邻样本的平均值，计算每个平均值的Gini值）

3. 计算Stump在最终分类中的权重

3.1先计算total error



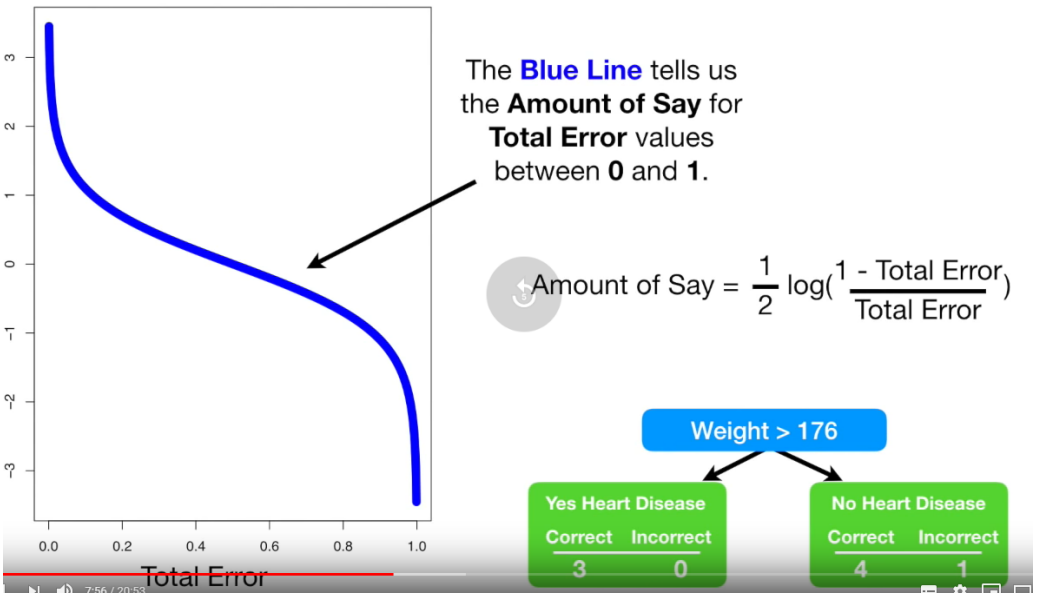
红框中的样本是被误分类的，一个stump的total error=sum(误分类的样本的sample weight)。因此，该stump的total error=1/8



3.2计算一个stump的Amount of Say

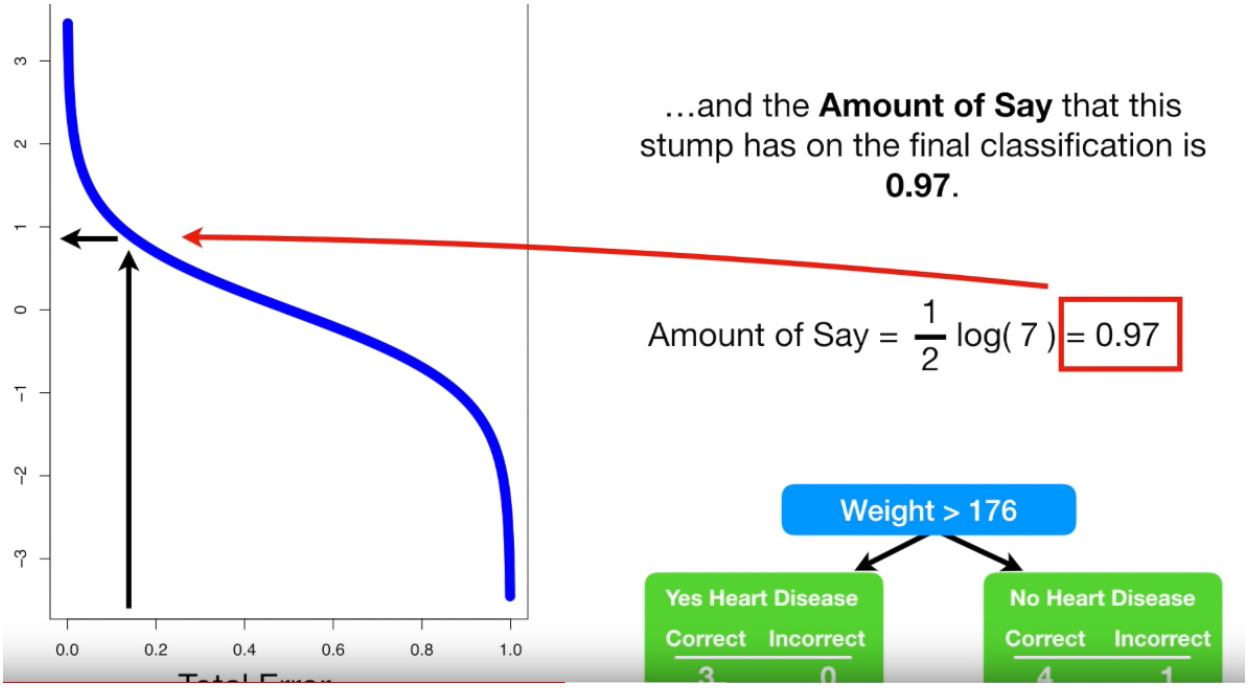
We use the **Total Error** to determine **Amount of Say** this stump has in the final classification with the following formula:

$$\text{Amount of Say} = \frac{1}{2}\log\left(\frac{1 - \text{Total Error}}{\text{Total Error}}\right)$$



由函数图像可知，当total error很接近1时，即stump的分类效果很差时，amount of say为很大的负数。反之为很大的正数。

*在实际使用中，amount of say的公式要加上一个很小的error项，以防total error=1或=0.



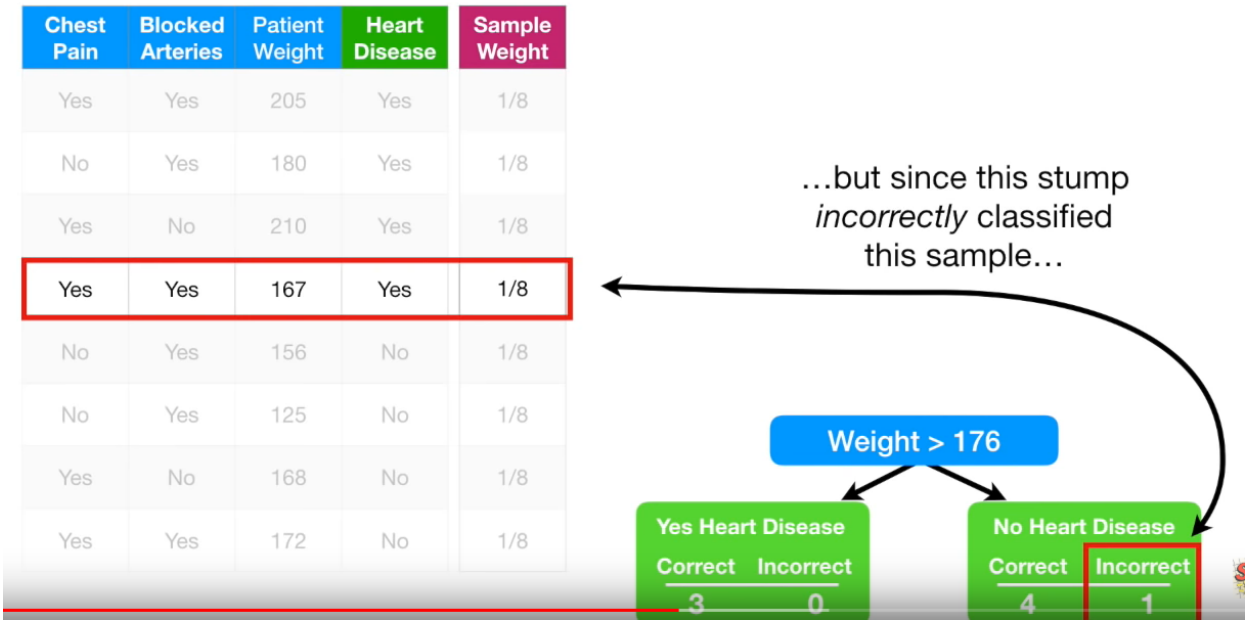
因此，该stump在最终分类时的权重为0.97。

Amount of Say越大，证明该stump分类效果越好，反之越差。

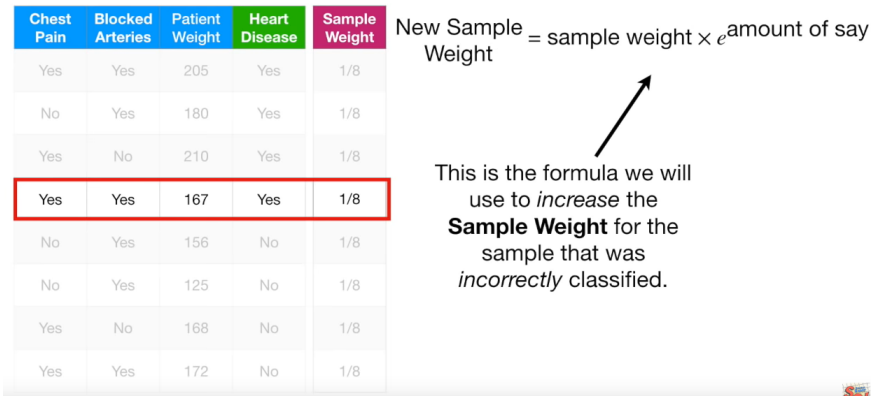
4.更新sample weight

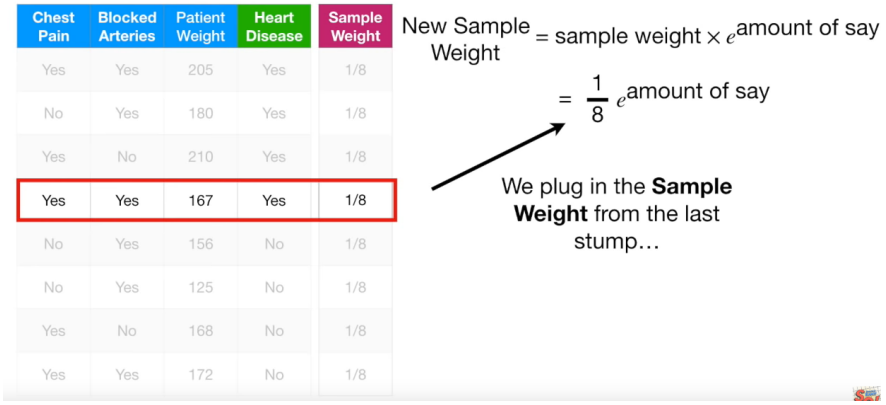
建立第一个stump时，所有sample的weight是相同的，没有侧重于要将某个sample分类正确。然而，第一个stump将某一个sample误分类了，因此后面一个stump要侧重于将这

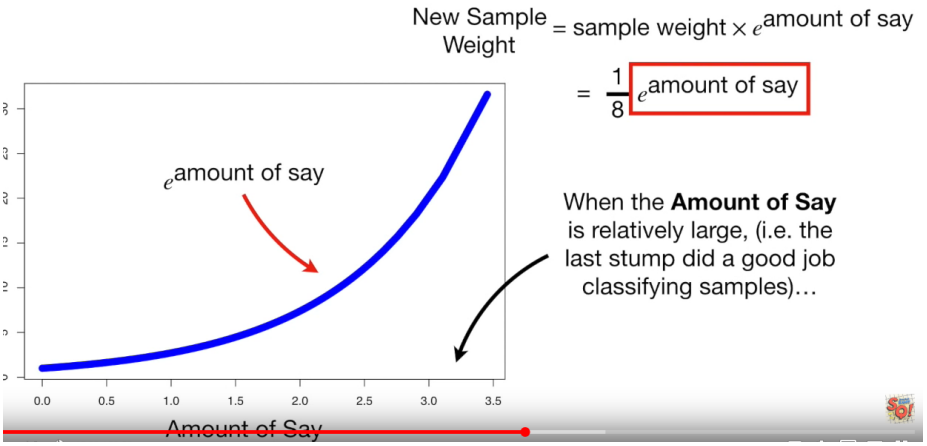
个误分类的sample正确分类，为此就要提高该sample的weight，降低其他已正确分类的sample的weight。

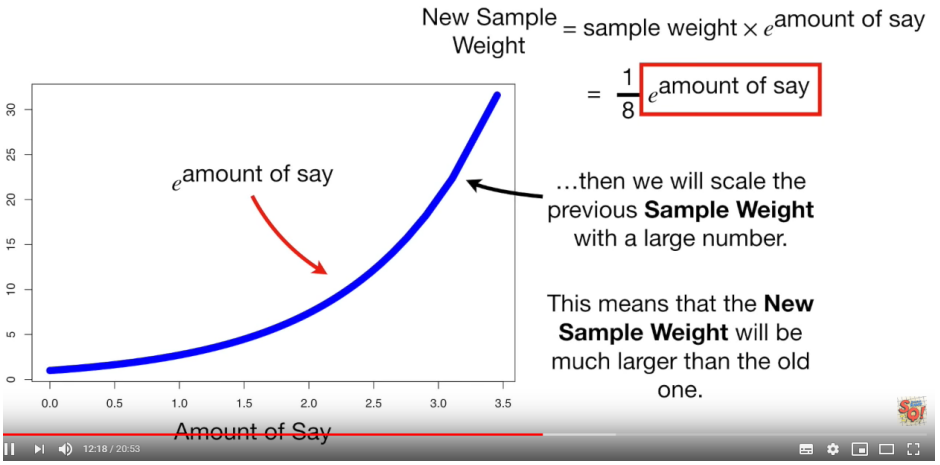


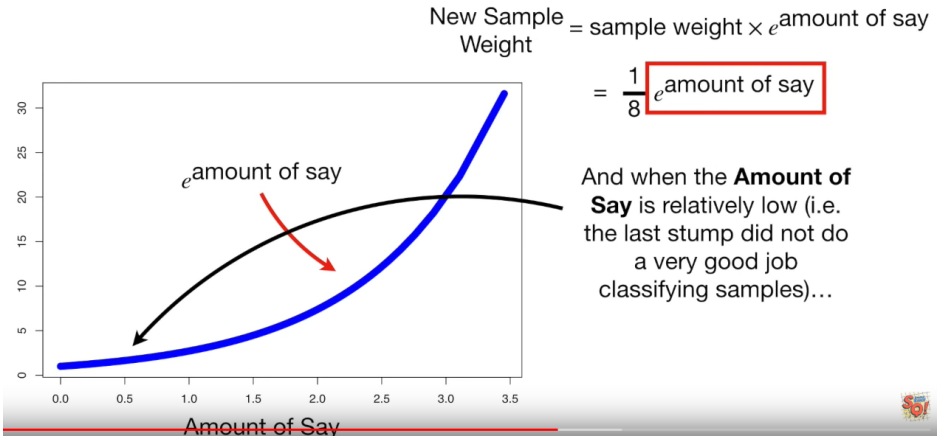
提高**误分类**sample的weight的公式：

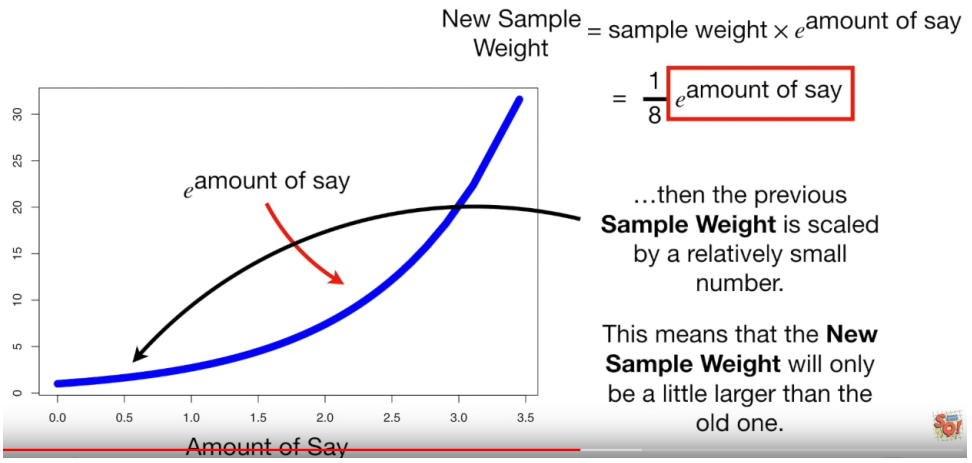


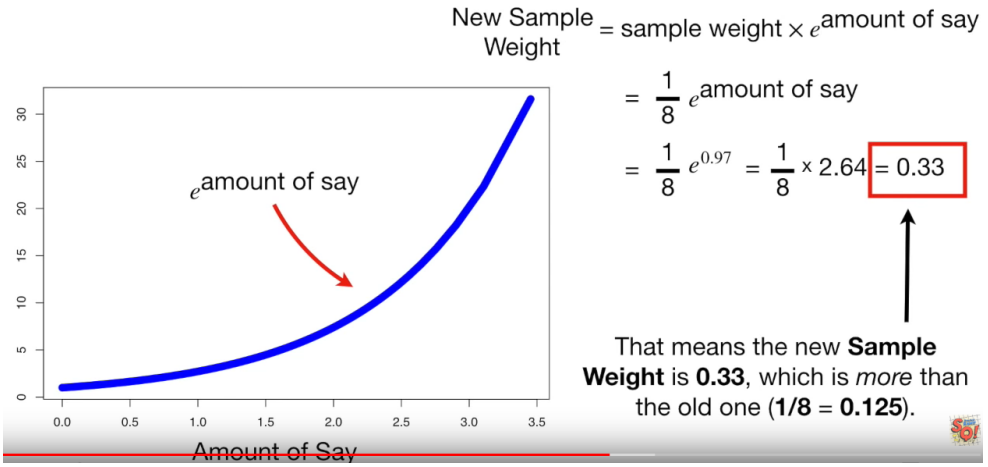



New Sample Weight = sample weight × $e^{\text{amount of say}}$
= $\frac{1}{8}e^{\text{amount of say}}$
$e^{\text{amount of say}}$
When the **Amount of Say** is relatively large, (i.e. the last stump did a good job classifying samples)…
0.0 0.5 1.0 1.5 2.0 2.5 3.0 3.5
Amount of Say

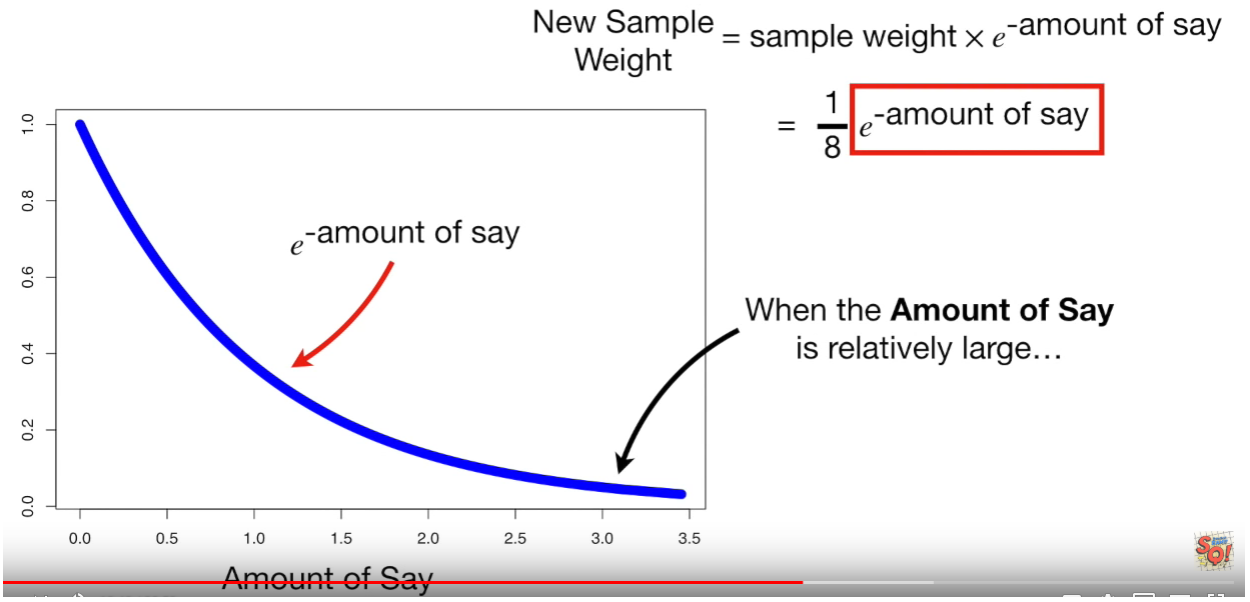

New Sample Weight = sample weight × $e^{\text{amount of say}}$
= $\frac{1}{8}e^{\text{amount of say}}$
$e^{\text{amount of say}}$
…then we will scale the previous **Sample Weight** with a large number.
This means that the **New Sample Weight** will be much larger than the old one.
0 5 10 15 20 25 30
0.0 0.5 1.0 1.5 2.0 2.5 3.0 3.5
Amount of Say
12:18 / 20:53

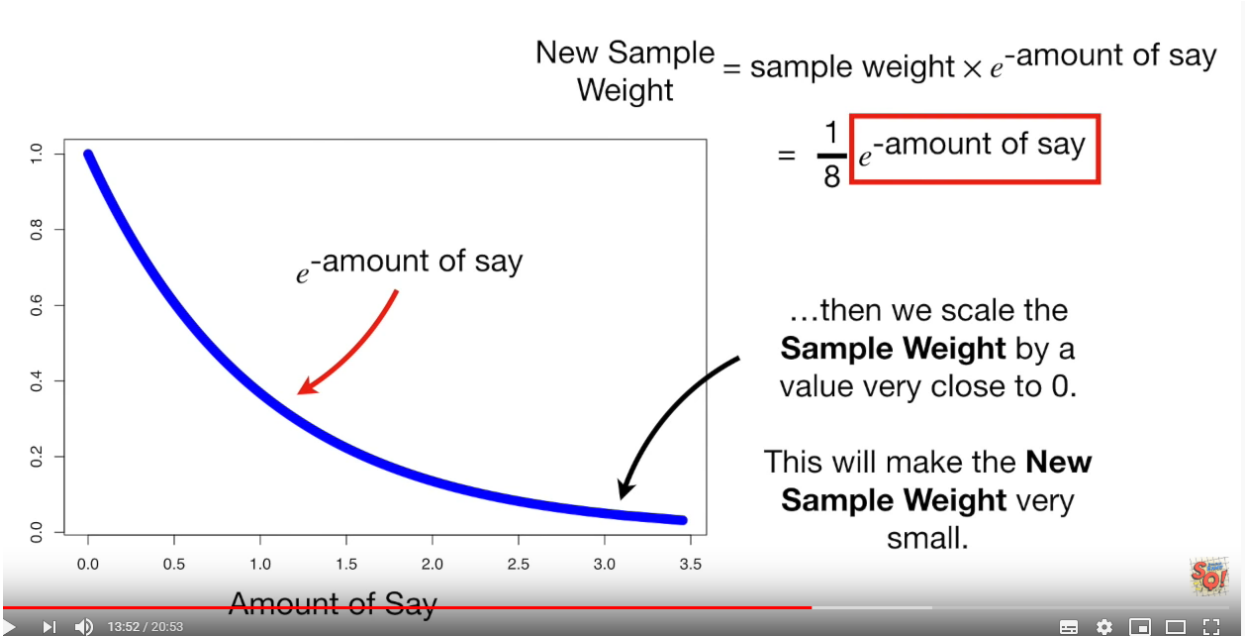

New Sample Weight = sample weight × $e^{\text{amount of say}}$
= $\frac{1}{8}e^{\text{amount of say}}$
$e^{\text{amount of say}}$
And when the **Amount of Say** is relatively low (i.e. the last stump did not do a very good job classifying samples)…
0 5 10 15 20 25 30
0.0 0.5 1.0 1.5 2.0 2.5 3.0 3.5
Amount of Say

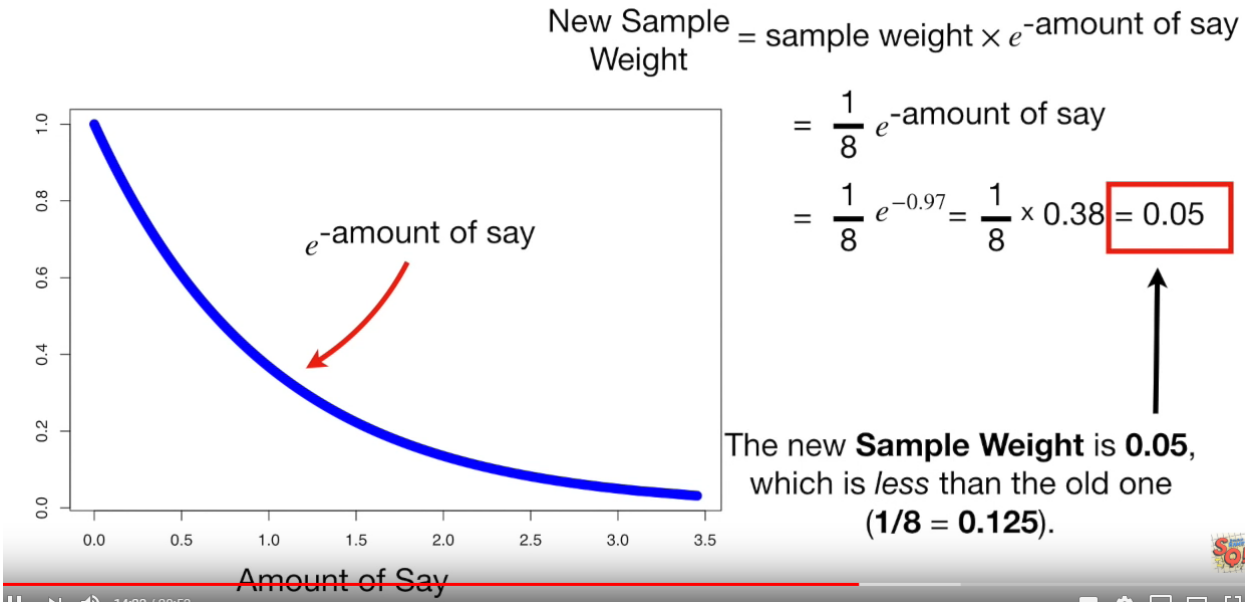




降低**正确分类**sample的weight:







更新完所有sample的weight:

Chest Pain	Blocked Arteries	Patient Weight	Heart Disease	Sample Weight	New Weight
Yes	Yes	205	Yes	1/8	0.05
No	Yes	180	Yes	1/8	0.05
Yes	No	210	Yes	1/8	0.05
Yes	Yes	167	Yes	1/8	0.33
No	Yes	156	No	1/8	0.05
No	Yes	125	No	1/8	0.05
Yes	No	168	No	1/8	0.05
Yes	Yes	172	No	1/8	0.05

All of the other samples get **0.05**.

5.归一化新的sample weight

因为所有sample weight的和要等于1，因此需要标准化

Chest Pain	Blocked Arteries	Patient Weight	Heart Disease	Sample Weight	New Weight	Norm. Weight
Yes	Yes	205	Yes	1/8	0.05	0.07
No	Yes	180	Yes	1/8	0.05	0.07
Yes	No	210	Yes	1/8	0.05	0.07
Yes	Yes	167	Yes	1/8	0.33	0.49
No	Yes	156	No	1/8	0.05	0.07
No	Yes	125	No	1/8	0.05	0.07
Yes	No	168	No	1/8	0.05	0.07
Yes	Yes	172	No	1/8	0.05	0.07

So we divide each **New Sample Weight** by **0.68** to get the normalized values.

weight(i)=weight(i)/所有weight的和

Chest Pain	Blocked Arteries	Patient Weight	Heart Disease	Sample Weight	New Weight	Norm. Weight
Yes	Yes	205	Yes	1/8	0.05	0.07
No	Yes	180	Yes	1/8	0.05	0.07
Yes	No	210	Yes	1/8	0.05	0.07
Yes	Yes	167	Yes	1/8	0.33	0.49
No	Yes	156	No	1/8	0.05	0.07
No	Yes	125	No	1/8	0.05	0.07
Yes	No	168	No	1/8	0.05	0.07
Yes	Yes	172	No	1/8	0.05	0.07

Now, when we add up the **New Sample Weights**, we get **1** (plus or minus a little rounding error).

Chest Pain	Blocked Arteries	Patient Weight	Heart Disease	Sample Weight
Yes	Yes	205	Yes	0.07
No	Yes	180	Yes	0.07
Yes	No	210	Yes	0.07
Yes	Yes	167	Yes	0.49
No	Yes	156	No	0.07
No	Yes	125	No	0.07
Yes	No	168	No	0.07
Yes	Yes	172	No	0.07

Now we just transfer the **Normalized Sample Weights** to the **Sample Weights** column, since those are what we will use for the next stump.

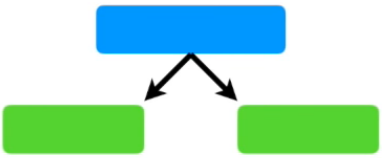
6.使用新的sample weight构造下一个stump（有两种方法）

第一种方法是：

理论上，可以通过sample weights来计算wighted gini index来确定构造下一个stump使用的特征（有待学习）。

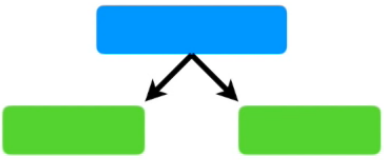
Chest Pain	Blocked Arteries	Patient Weight	Heart Disease	Sample Weight
Yes	Yes	205	Yes	0.07
No	Yes	180	Yes	0.07
Yes	No	210	Yes	0.07
Yes	Yes	167	Yes	0.49
No	Yes	156	No	0.07
No	Yes	125	No	0.07
Yes	No	168	No	0.07
Yes	Yes	172	No	0.07

In theory, we could use the **Sample Weights** to calculate **Weighted Gini Indexes** to determine which variable should split the next stump.



Chest Pain	Blocked Arteries	Patient Weight	Heart Disease	Sample Weight
Yes	Yes	205	Yes	0.07
No	Yes	180	Yes	0.07
Yes	No	210	Yes	0.07
Yes	Yes	167	Yes	0.49
No	Yes	156	No	0.07
No	Yes	125	No	0.07
Yes	No	168	No	0.07
Yes	Yes	172	No	0.07

The **Weighted Gini Index** would put more emphasis on correctly classifying this sample (the one that was misclassified by the last stump), since this sample has the largest **Sample Weight**.



第二种方法是，建立一个新的数据集，其中包含多个sample weight最大的样本：

（1）建立一个新的、空的、跟原数据集大小相同的collection

Chest Pain	Blocked Arteries	Patient Weight	Heart Disease	Sample Weight
Yes	Yes	205	Yes	0.07
No	Yes	180	Yes	0.07
Yes	No	210	Yes	0.07
Yes	Yes	167	Yes	0.49
No	Yes	156	No	0.07
No	Yes	125	No	0.07
Yes	No	168	No	0.07
Yes	Yes	172	No	0.07

So we start by making a new, but empty, dataset that is the same size as the original…

Chest Pain	Blocked Arteries	Patient Weight	Heart Disease

（2）在0~1中随机选择一个数，看这个数落在sample weight的哪个区间中

Chest Pain	Blocked Arteries	Patient Weight	Heart Disease	Sample Weight
Yes	Yes	205	Yes	0.07
No	Yes	180	Yes	0.07
Yes	No	210	Yes	0.07
Yes	Yes	167	Yes	0.49
No	Yes	156	No	0.07
No	Yes	125	No	0.07
Yes	No	168	No	0.07
Yes	Yes	172	No	0.07

…and we see where that number falls when we use the **Sample Weights** like a distribution.

Chest Pain	Blocked Arteries	Patient Weight	Heart Disease

Chest Pain	Blocked Arteries	Patient Weight	Heart Disease	Sample Weight
Yes	Yes	205	Yes	0.07
No	Yes	180	Yes	0.07
Yes	No	210	Yes	0.07
Yes	Yes	167	Yes	0.49
No	Yes	156	No	0.07
No	Yes	125	No	0.07
Yes	No	168	No	0.07
Yes	Yes	172	No	0.07

Chest Pain	Blocked Arteries	Patient Weight	H Di

If the number is between **0** and **0.07**, then we would put this sample into the new collection of samples…

Chest Pain	Blocked Arteries	Patient Weight	Heart Disease	Sample Weight
Yes	Yes	205	Yes	0.07
No	Yes	180	Yes	0.07
Yes	No	210	Yes	0.07
Yes	Yes	167	Yes	0.49
No	Yes	156	No	0.07
No	Yes	125	No	0.07
Yes	No	168	No	0.07
Yes	Yes	172	No	0.07

Chest Pain	Blocked Arteries	Patient Weight

…and if the number is between **0.07** and **0.14** (**0.07** + **0.07** = **0.14**), then we would put this sample into the new collection of samples…

Chest Pain	Blocked Arteries	Patient Weight	Heart Disease	Sample Weight
Yes	Yes	205	Yes	0.07
No	Yes	180	Yes	0.07
Yes	No	210	Yes	0.07
Yes	Yes	167	Yes	0.49
No	Yes	156	No	0.07
No	Yes	125	No	0.07
Yes	No	168	No	0.07
Yes	Yes	172	No	0.07

Chest Pain	Blocked Arteries	Patient Weight	D

…and if the number is between **0.14** and **0.21** (**0.14** + **0.07** = **0.21**), then we would put this sample into the new collection of samples…

Chest Pain	Blocked Arteries	Patient Weight	Heart Disease	Sample Weight
Yes	Yes	205	Yes	0.07
No	Yes	180	Yes	0.07
Yes	No	210	Yes	0.07
Yes	Yes	167	Yes	0.49
No	Yes	156	No	0.07
No	Yes	125	No	0.07
Yes	No	168	No	0.07
Yes	Yes	172	No	0.07

Chest Pain	Blocked Arteries	Patient Weight	Heart Disease

For example, imagine the first number I picked was **0.72**…

…then I would put this sample into my new collection of samples…

Chest Pain	Blocked Arteries	Patient Weight	Heart Disease	Sample Weight
Yes	Yes	205	Yes	0.07
No	Yes	180	Yes	0.07
Yes	No	210	Yes	0.07
Yes	Yes	167	Yes	0.49
No	Yes	156	No	0.07
No	Yes	125	No	0.07
Yes	No	168	No	0.07
Yes	Yes	172	No	0.07

Chest Pain	Blocked Arteries	Patient Weight	Heart Disease
No	Yes	156	No
Yes	Yes	167	Yes
No	Yes	125	No
Yes	Yes	167	Yes

NOTE: This is the second time that we have added this particular sample to the new collection of samples.

Chest Pain	Blocked Arteries	Patient Weight	Heart Disease	Sample Weight
Yes	Yes	205	Yes	0.07
No	Yes	180	Yes	0.07
Yes				
Yes				
No				
No	Yes	125	No	0.07
Yes	No	168	No	0.07
Yes	Yes	172	No	0.07

We then continue to pick random numbers and add samples to the new collection until we the new collection is the same size as the original.

Chest Pain	Blocked Arteries	Patient Weight	Heart Disease
No	Yes	156	No
Yes	Yes	167	Yes
No	Yes	125	No
Yes	Yes	167	Yes
Yes	Yes	167	Yes
Yes	Yes	172	No
Yes	Yes	205	Yes

Chest Pain	Blocked Arteries	Patient Weight	Heart Disease	Sample Weight
Yes	Yes	205	Yes	0.07
No	Yes	180	Yes	0.07
Yes	No	210	Yes	0.07
Yes	Yes	167	Yes	0.49
No	Yes	156	No	0.07
No	Yes	125		
Yes	No	168		
Yes	Yes	172		

Ultimately, this sample was added to the new collection of samples **4** times, reflecting its larger **Sample Weight**.

Chest Pain	Blocked Arteries	Patient Weight	Heart Disease
No	Yes	156	No
Yes	Yes	167	Yes
No	Yes	125	No
Yes	Yes	167	Yes
Yes	Yes	167	Yes
Yes	Yes	172	No
Yes	Yes	205	Yes
Yes	Yes	167	Yes

Now we get rid of the original samples…

Chest Pain	Blocked Arteries	Patient Weight	Heart Disease
No	Yes	156	No
Yes	Yes	167	Yes
No	Yes	125	No
Yes	Yes	167	Yes
Yes	Yes	167	Yes
Yes	Yes	172	No
Yes	Yes	205	Yes
Yes	Yes	167	Yes

7.使用新的dataset构造下一个stump

Chest Pain	Blocked Arteries	Patient Weight	Heart Disease	Sample Weight
No	Yes	156	No	1/8
Yes	Yes	167	Yes	1/8
No	Yes	125	No	1/8
Yes	Yes	167	Yes	1/8
Yes	Yes	167	Yes	1/8
Yes	Yes	172	No	1/8
Yes	Yes	205	Yes	1/8
Yes	Yes	167	Yes	1/8

Lastly, we give all the samples equal **Sample Weights**, just like before.

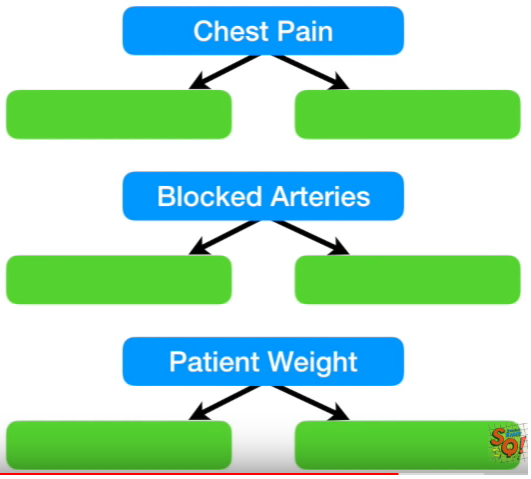
将sample weight重置为1/样本总数。但是这并不表示新的stump不会侧重于将此前误分类的sample。因为新的dataset中，误分类的样本重复出现多次。

Chest Pain	Blocked Arteries	Patient Weight	Heart Disease	Sample Weight
No	Yes	156	No	1/8
Yes	Yes	167	Yes	1/8
No	Yes	125	No	1/8
Yes	Yes	167	Yes	1/8
Yes	Yes	167	Yes	1/8
Yes	Yes	172	No	1/8
Yes	Yes	205	Yes	1/8
Yes	Yes	167	Yes	1/8

Because these samples are all the same, they will be treated as a block, creating a large penalty for being misclassified.

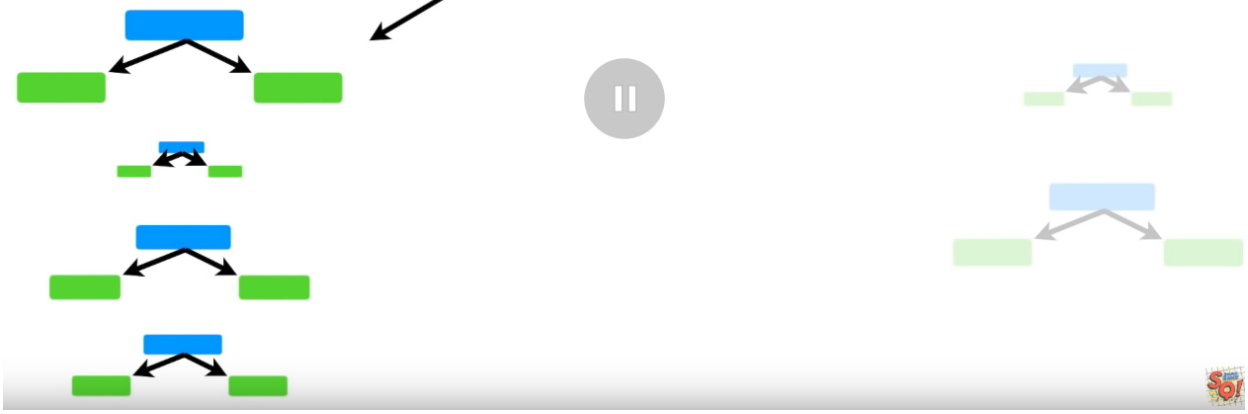
Chest Pain	Blocked Arteries	Patient Weight	Heart Disease	Sample Weight
No	Yes	156	No	1/8
Yes	Yes	167	Yes	1/8
No	Yes	125	No	1/8
Yes	Yes	167	Yes	1/8
Yes	Yes	167	Yes	1/8
Yes	Yes	172	No	1/8
Yes	Yes	205	Yes	1/8
Yes	Yes	167	Yes	1/8

Now we go back to the beginning and try to find the stump that does the best job classifying the new collection of samples.



如何使用AdaBoost做分类?

Imagine that these stumps classified a patient as **Has Heart Disease**…



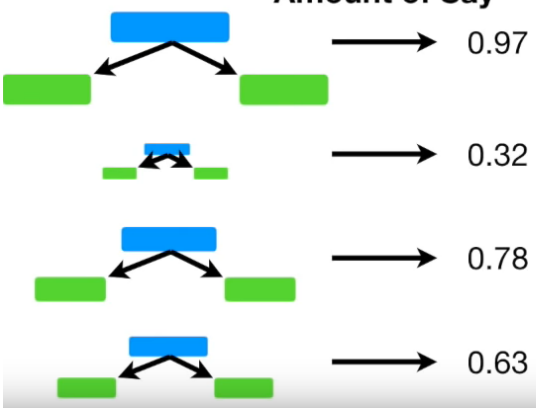
…and these stumps classified the patient as **Does Not Have Heart Disease**.

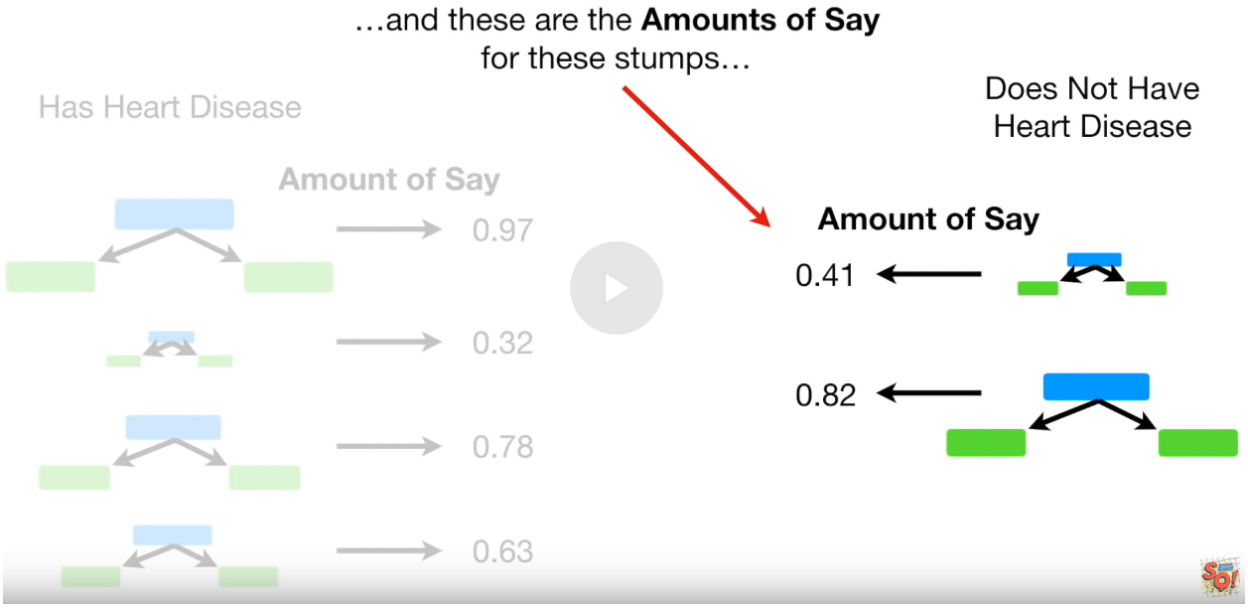
These are the **Amounts of Say** for these stumps…

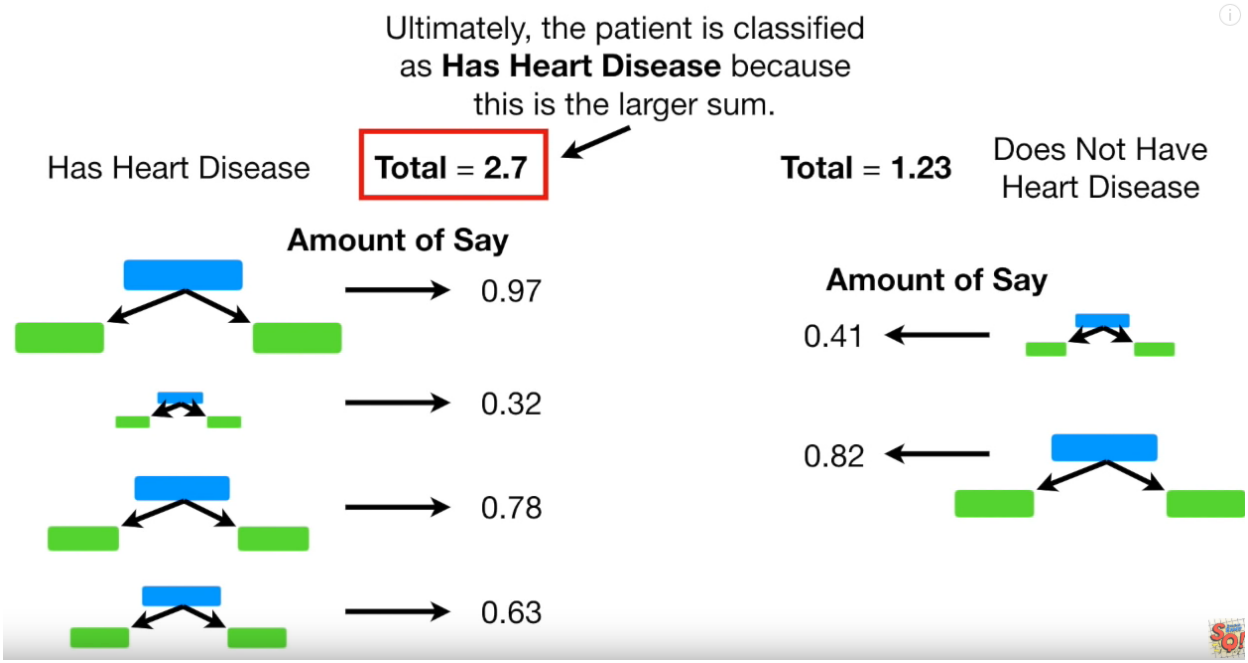
Has Heart Disease

Does Not Have Heart Disease

Amount of Say




…and these are the **Amounts of Say** for these stumps…
Has Heart Disease
Amount of Say
0.97
0.32
0.78
0.63
Does Not Have Heart Disease
Amount of Say
0.41
0.82


Ultimately, the patient is classified as **Has Heart Disease** because this is the larger sum.
Has Heart Disease
Total = **2.7**
Total = **1.23**
Does Not Have Heart Disease
Amount of Say
0.97
0.32
0.78
0.63
Amount of Say
0.41
0.82

If we have a **Weighted Gini Function**, then we use it with the **Sample Weights**, otherwise we use the **Sample Weights** to make a new dataset that reflects those weights.

3) Each **stump** is made by taking the previous **stump's** mistakes into account.