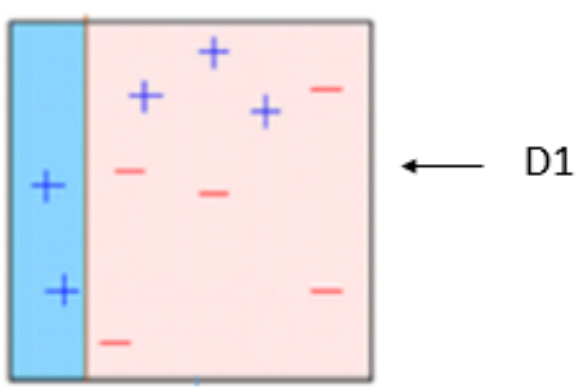
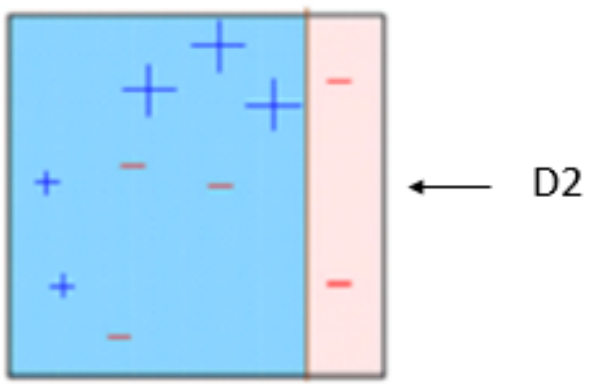
지난 회 '군중은 똑똑하다' 편에서 Random Forest 모델에 대하여 알아보았습니다. Random Forest는 수많은 의사결정 트리들을 모아서 의사결정 '숲'을 만든 후 투표에 의한 결정을 내린는 앙상블 방식이 요점이었습니다. 오늘은 Random Forest 외에도 데이터 사이언티스트들이 자주 쓰는 또 다른 모델, 숲의 군중보다 조금 더 똑똑한 군중, AdaBoost (Adaptive Boosting)에 대해서 설명해 보도록 하겠습니다.

AdaBoost 역시 Random Forest처럼 의사결정 트리 기반의 (tree-based) 모델입니다. 하지만 Random Forest처럼 각각의 트리들이 독립적으로 존재하지는 않습니다. 예를 들어, 아래의 상황에서 '+'와 '-'들을 구분해 내는 것이 목적이라고 가정합시다.



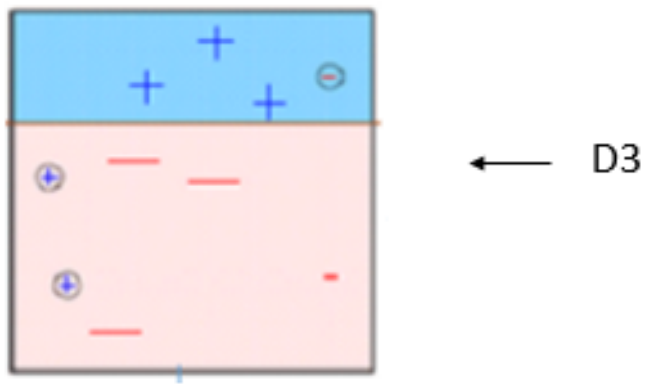
출처: Analytics Vidhya

AdaBoost는 첫 번째 의사결정 트리를 생성합니다. 좌측에서부터 약 1/5 지점을 기준으로 왼쪽은 '+', 오른쪽은 '-'라는 아주 단순한 결정을 내립니다. 한번 그은 선으로는 최선의 결과이긴 하지만, 이 결정은 두 기호를 완벽하게 분리하지 못합니다. 세 개의 '+'들이 '-'라고 잘못 지정되었네요. 저희는 틀린 것을 또 틀리고 싶지는 않습니다. 다음번에는 저 세 개의 '+'들을 정확하게 맞추기 위해서 '가중치'를 크게 조정합니다. '가중치'가 상향되었으니, 틀리면 다른 이미 맞춘 기호들을 틀리는것보다 피해가 큽니다. 재조정된 '가중치'로 AdaBoost는 두번째 의사결정 트리를 생성합니다.



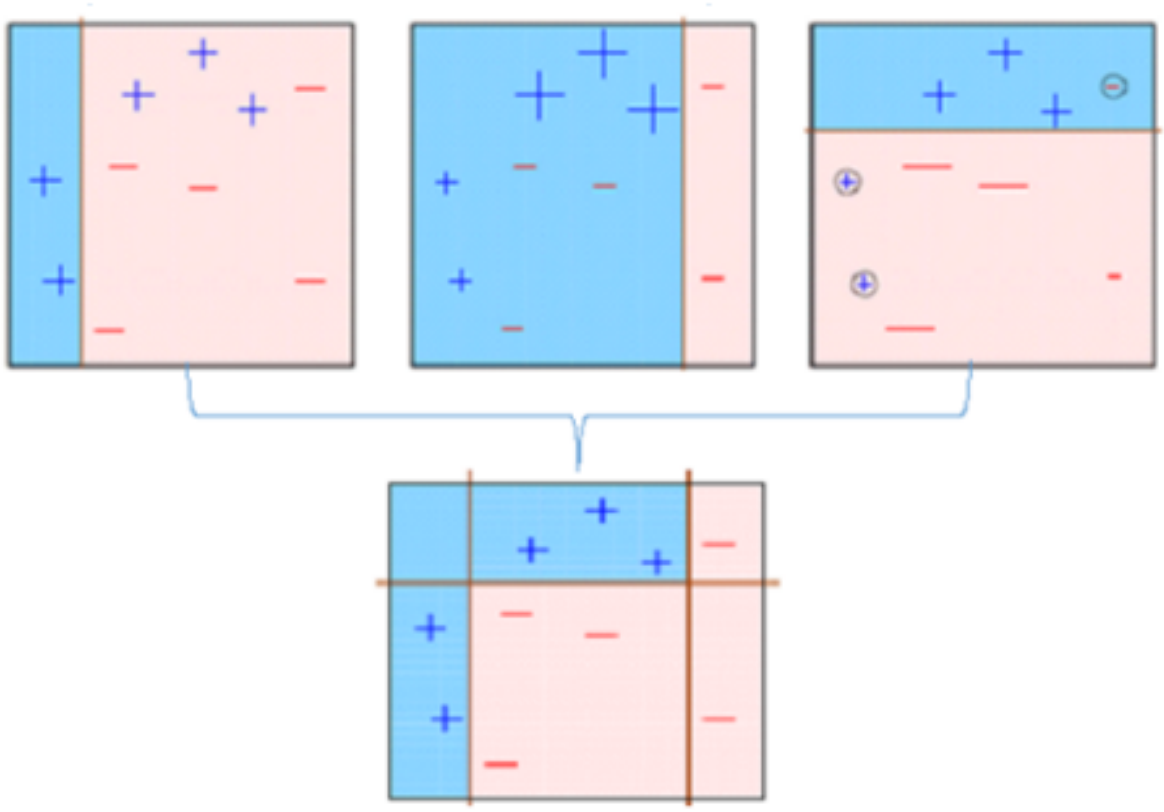
저번에 놓친 +들이 더 커진 것을 볼 수 있습니다.

이번에는 우측에서 약 1/5 지점을 기준으로 왼쪽은 '+,' 오른쪽은 '-'라는 결정을 내립니다. 역시 한 번의 선으로 나눌 수 있는 최선의 결과이고, 세 개의 '+'들을 모두 맞추기는 했지만, 이번에는 세 개의 '-'가 +라고 잘못 지정되었습니다.
AdaBoost는 같은 과정을 반복합니다.



이번에는 상단으로부터 약 1/3 지점의 가로선을 기준으로 기호들이 나뉘는군요!
기호들의 '배점'을 주의 깊게 살펴봅시다. 첫번째와 두번째 의사 결정 트리에서 성공적으로 맞춘 세 개의 기호들의 (원으로 표시되어있는) 배점은 줄어든 반면에, 두번째 트리가 틀린 기호들의 배점은 또다시 상향 조정되었습니다.

이처럼 stump 트리들이 생성되었으니, 이제는 Random Forest와 같이 생성한 stump 트리들을 합칠 차례입니다. 세가지 트리들을 합친다면 ...



위와 같은 다소 복잡한 경계들이 형성되고, 훨씬 더 정확한 예측을 할 수 있게 됩니다.

Random Forest와 AdaBoost의 근본적인 차이는 각각의 트리를 생성하는 방식에 있습니다. Random Forest는 각각의 트리들이 상당히 우수합니다. 퀴즈 프로그램에 나간다고 생각한다면, 국어, 수학, 역사, 음악, 미술, 체육 등 각기 다른 분야에 우수한 전교 5등과 함께 출전하는 것과 같은 원리입니다. 이들은 퀴즈 프로에 나가기 전까지 같은 학교에 있을 필요도, 서로 알 필요도 없습니다. 내신은 같이 출전한 전교 1등 팀에게 밀릴지도 모릅니다. 하지만, 내신에 중요한 '국영수' 위주의 학습을 위주로 한 1등 팀은 서로 아는 것도 비슷하여 철학 문제가 출제된다면 다 같이 틀려, 퀴즈 대회에서만큼은 전교 5등 팀 Random Forest가 전교 1등팀을 꺾을 수 있습니다.

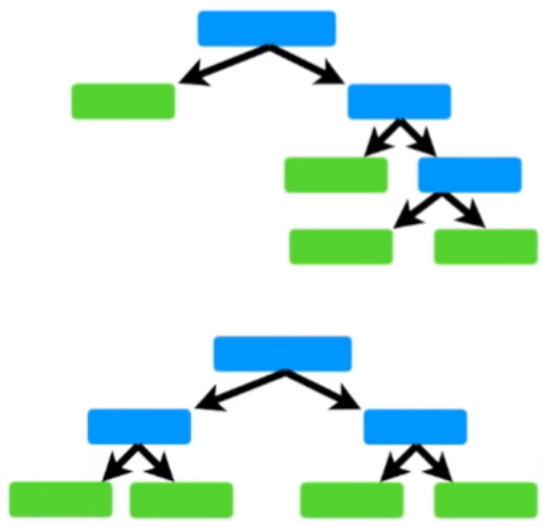
하지만 이러한 강적도 전교 50등이 모인 AdaBoost팀의 적수는 되지 못합니다. AdaBoost팀은 Random Forest팀과는 다르게 같은 동네에서 자란 죽마고우들 입니다. 개개인들은 공부를 잘하지 못하지만, 서로의 강점과 약점들은 빠짐없이 알고 있습니다. 친구들이 자주 틀리는 문제들을 위주로 공부해서, 내신을 몰라도 퀴즈 대회에서 만큼은 최강의 팀이 되죠.

그럼 Random Forest 같은것은 필요 없고, AdaBoost만 쓰면 되냐고요? 그것은 아닙니다. AdaBoost, 확실히 정확도는 높지만 팀을 구성하는데 오래 걸립니다. AdaBoost는 퀴즈 대회에 출전하고 싶은 학생 한명을 고른 후, 친구들을 한명씩 불러와야 하는 반면에, Random Forest는 대회에 참가하고 싶은 전교 5등을 구하는 공고문 하나만 올리면 끝나기 때문입니다. 이 것을 전문적인 용어로 "병행 연산 (parallel computing)"이라고 합니다. Random Forest의 트리들은 여러 대의 컴퓨터로 만들 수 있어서 연산적인 측면의 어드밴티지가 있죠. 그래서 시간적인 제약이 많은 현실 세계에서는 Random Forest가 많이 사랑받고 있는 이유입니다.

병렬 연산 가능.

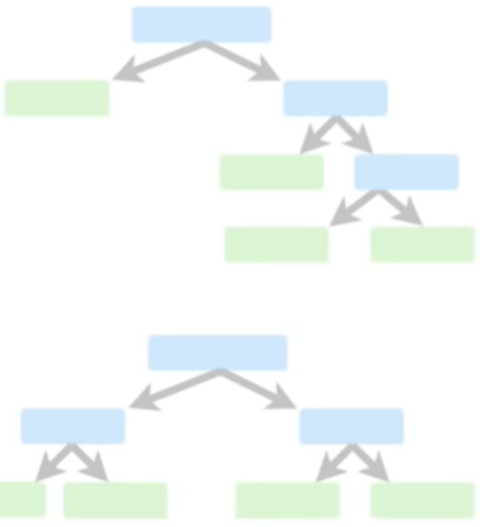
하지만 AdaBoost의 이러한 협력 방식, 어떤 의미에서는 진정한 협력이 아닐까요? 정확도만이 배울 점이아닌 것 같습니다.

In a **Random Forest**, each time you make a tree, you make a full sized tree.

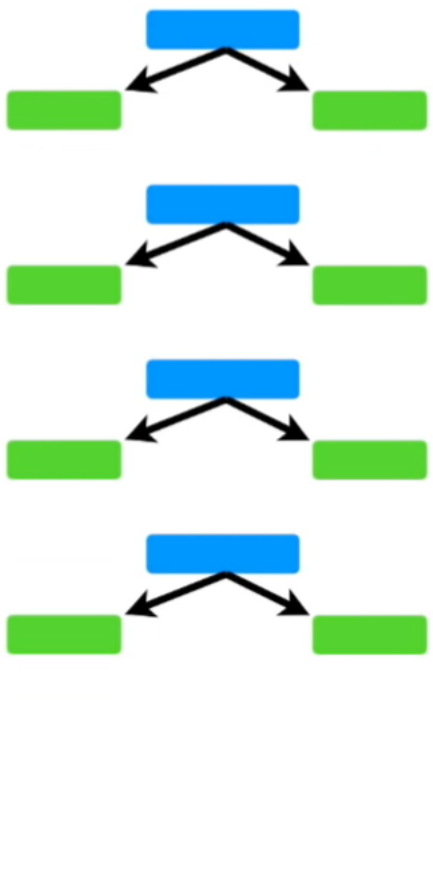


Some trees might be bigger than others, but there is no predetermined maximum depth.

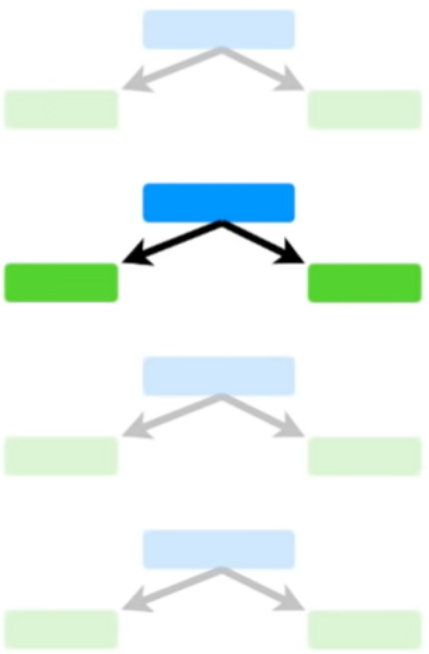
'가지치기'를 하지 않음



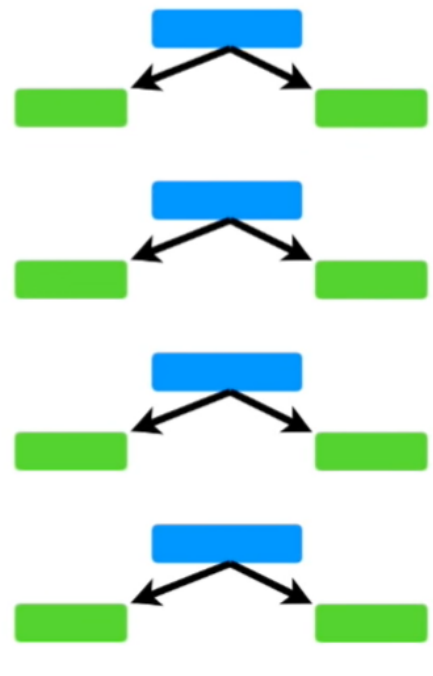
In contrast, in a **Forest of Trees** made with **AdaBoost**, the trees are usually just a **node** and two **leaves**.



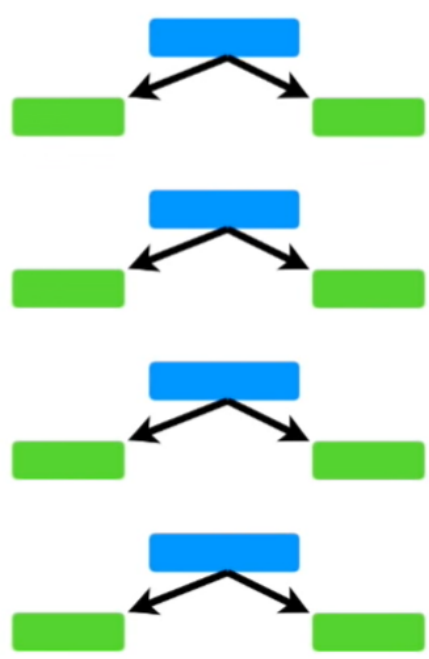
A tree with just one node and two leaves is called a ***stump***.



…so this is really a
Forest of Stumps
rather than trees.



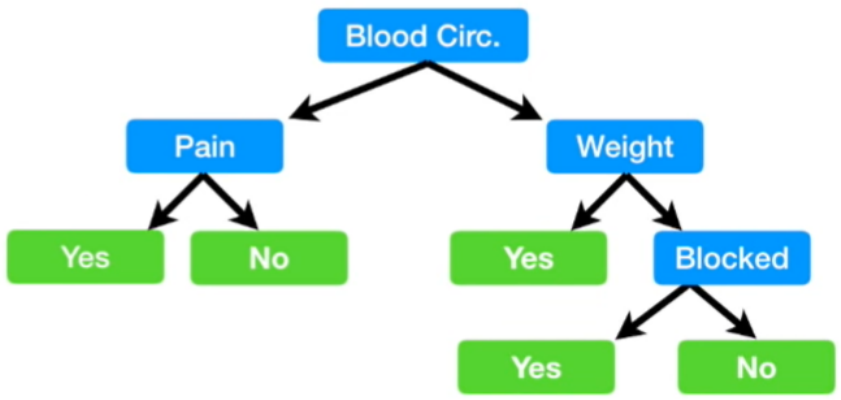
Stumps are not great at making accurate classifications.



For example, if we were using this data to determine if someone had heart disease or not…

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

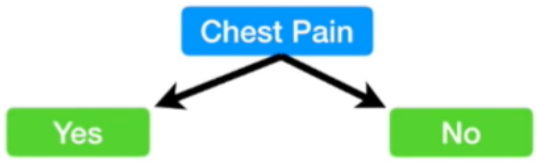
…then a full sized **Decision Tree** would take advantage of all **4** variables that we measured (**Chest Pain**, **Blood Circulation**, **Blocked Arteries** and **Weight**) to make a decision…



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
No	No	No	125	No
Yes	Yes	Yes	180	Yes
Yes	Yes	No	210	No
Yes	No	Yes	167	Yes

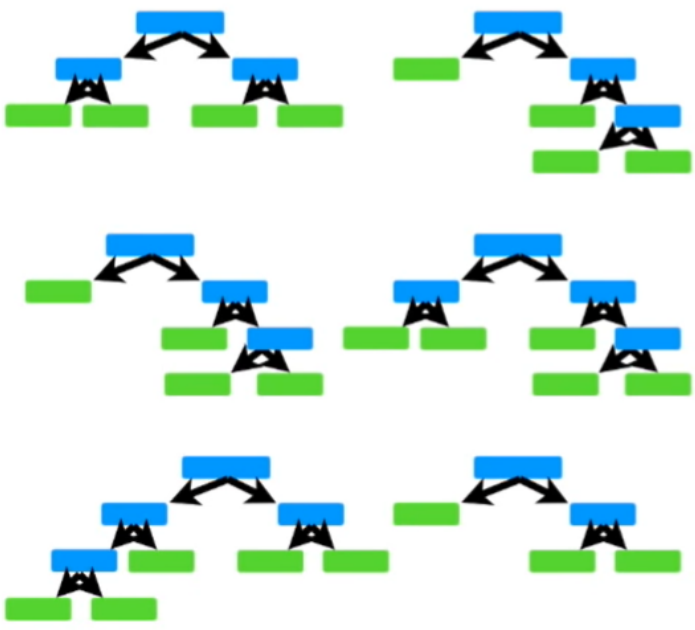
...but a **Stump** can only use one variable to make a decision.

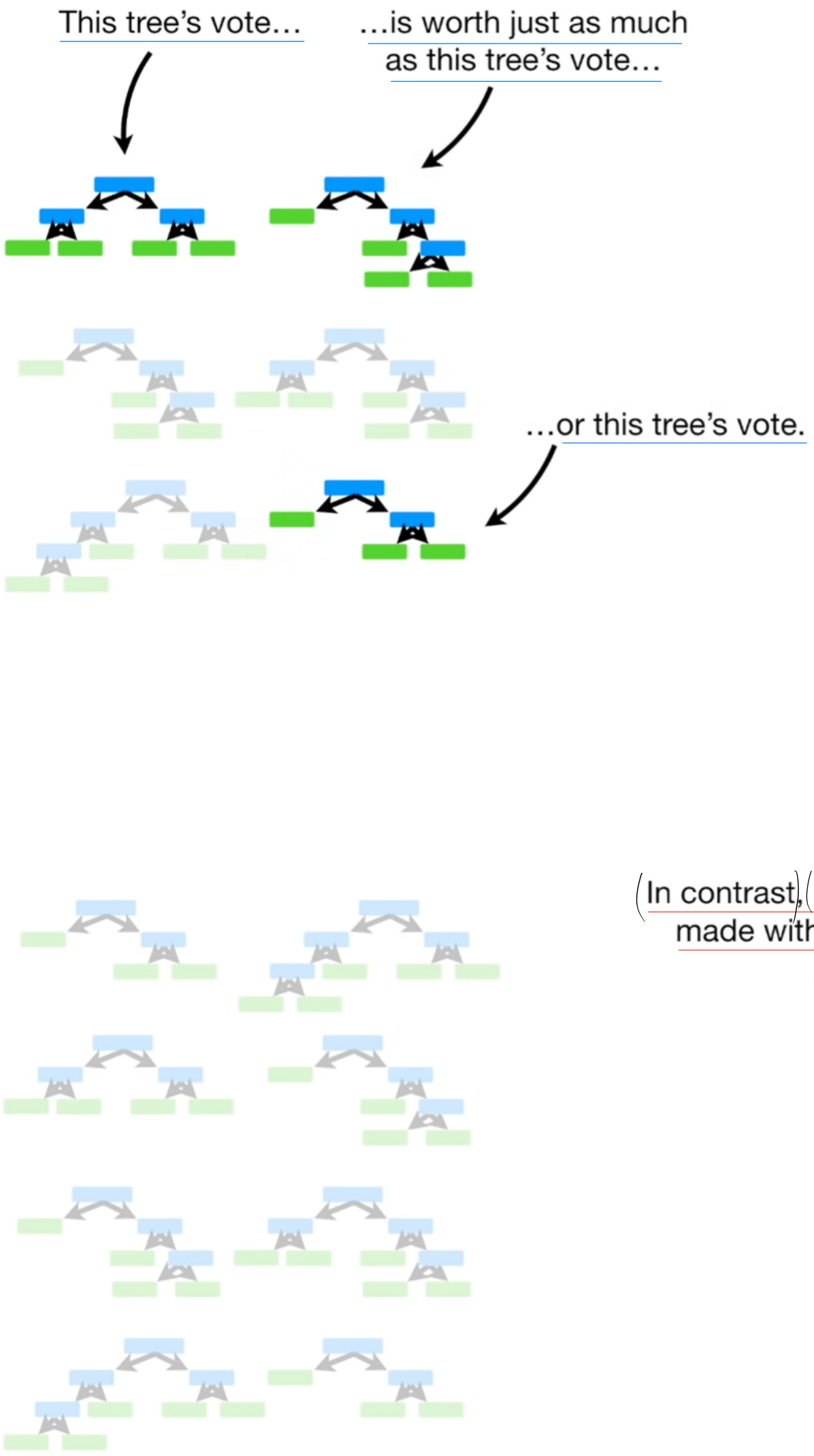


Thus, **Stumps** are technically "weak learners".

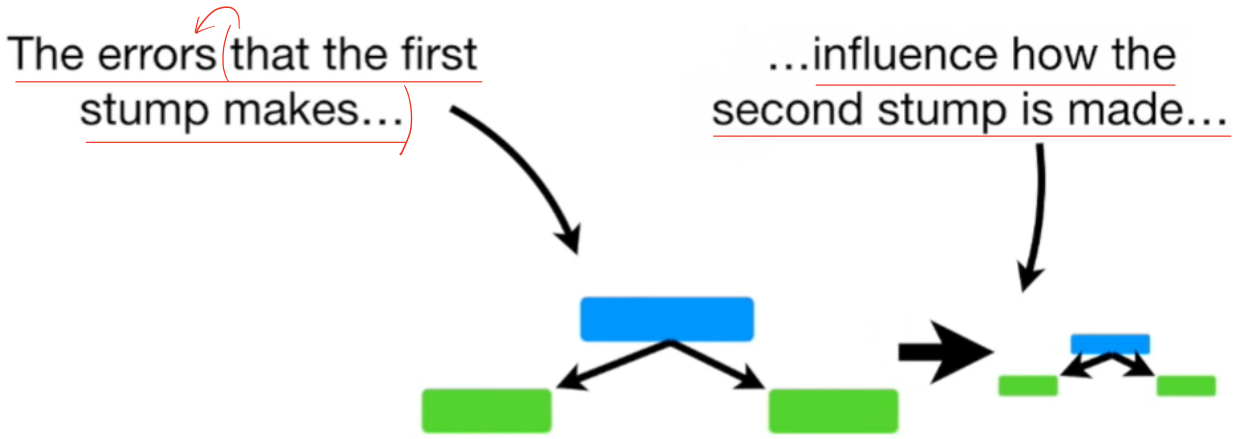
However, that's the way **AdaBoost** likes it, and it's one of the reasons why they are so commonly combined.

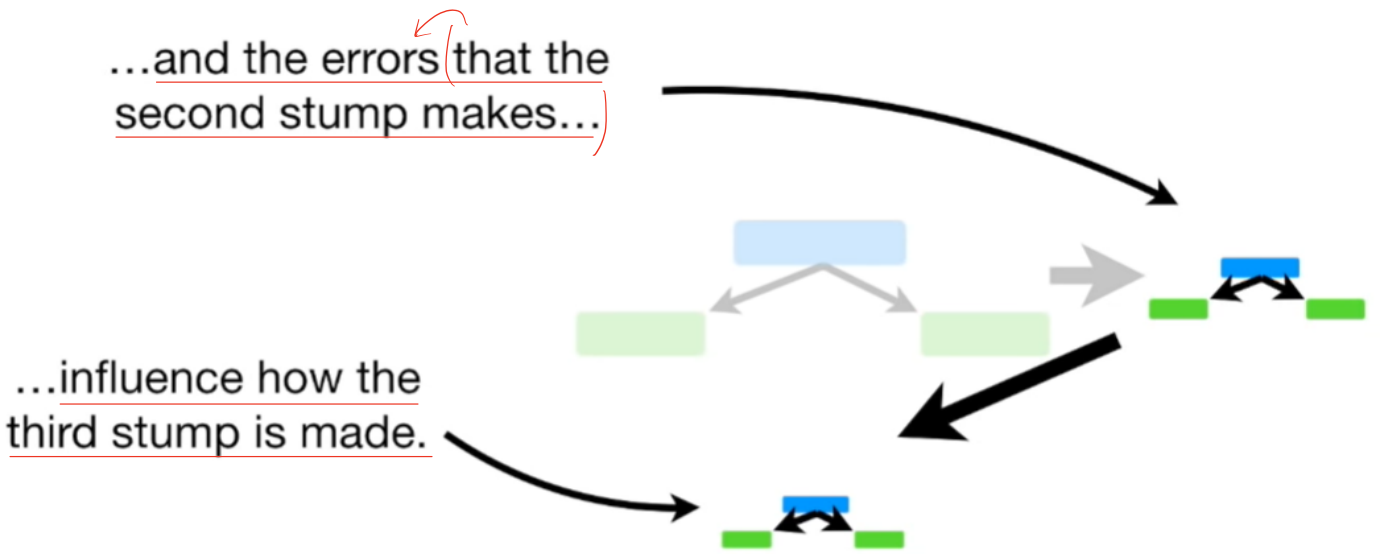
In a **Random Forest**, each tree has an equal vote on the final classification.





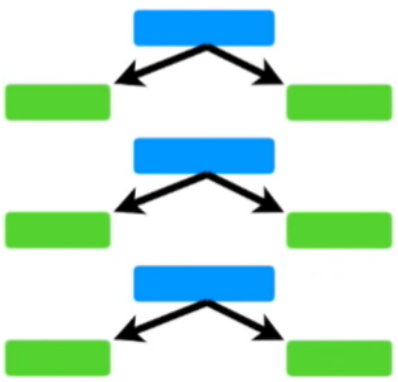
(In contrast, in a **Forest of Stumps** made with **AdaBoost**, order is important.)


The errors that the first stump makes…
…influence how the second stump is made…

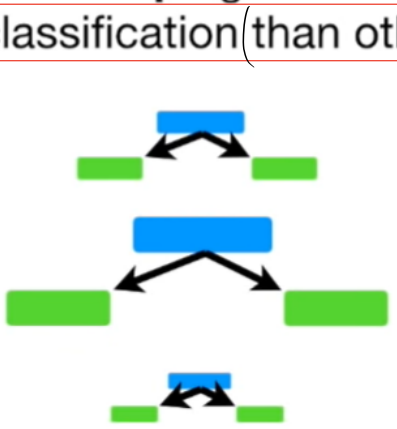

…and the errors that the second stump makes…
…influence how the third stump is made.

To review, the three ideas behind **AdaBoost** are…

1) **AdaBoost** combines a lot of "weak learners" to make classifications. The weak learners are almost aways **stumps**.

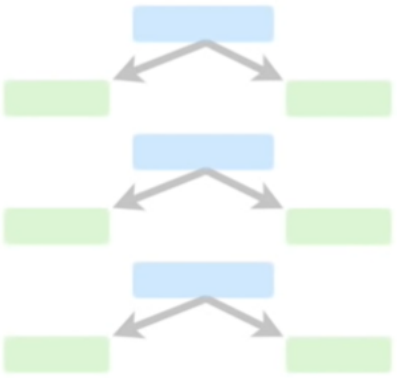


2) Some **stumps** get more say in the classification than others.



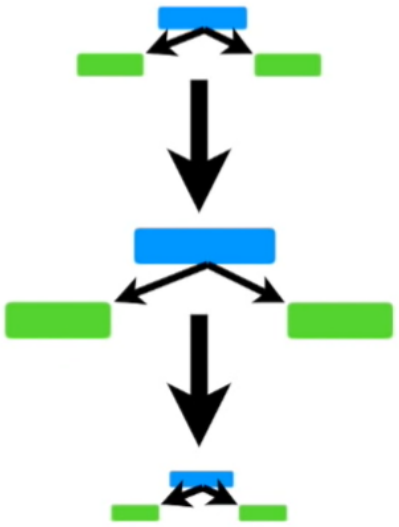
To review, the three ideas behind **AdaBoost** are…

1) **AdaBoost** combines a lot of "weak learners" to make classifications. The weak learners are almost aways **stumps**.



2) Some **stumps** get more say in the classification than others.

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



Now let's dive into the nitty gritty detail of how to create a **Forest of Stumps** using **AdaBoost**.

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

First, we'll start with some data.

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

We create a **Forest of Stumps** with **AdaBoost** to predict if a patient has heart disease.

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

We will make these predictions based on a patient's **Chest Pain** and **Blocked Artery** status and their **Weight**.

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

The first thing we do is give each sample a weight that indicates how important it is to be correctly classified.

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

NOTE: The **Sample Weight** is different from the **Patient Weight**, and I'll do the best I can to be clear about which of the two I'm talking about.

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.

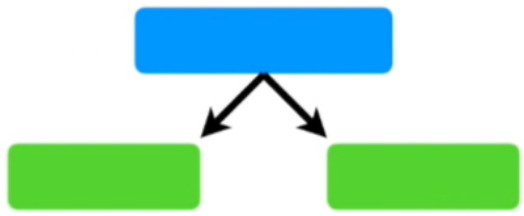
Chest Pain	Blocked Arteries	Patient Weight	Heart Disease	Sample Weight
Yes	Yes	205	Yes	1/8
No	Yes	180	Yes	1/8
Yes	No	210	Yes	1/8
Yes	Yes	167	Yes	1/8
No	Yes	156	No	1/8
No	Yes	125	No	1/8
Yes	No	168	No	1/8
Yes	Yes	172	No	1/8

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!

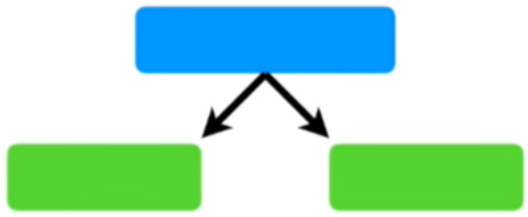
Chest Pain	Blocked Arteries	Patient Weight	Heart Disease	Sample Weight
Yes	Yes	205	Yes	1/8
No	Yes	180	Yes	1/8
Yes	No	210	Yes	1/8
Yes	Yes	167	Yes	1/8
No	Yes	156	No	1/8
No	Yes	125	No	1/8
Yes	No	168	No	1/8
Yes	Yes	172	No	1/8

Now we need to make the first **stump** in the forest.



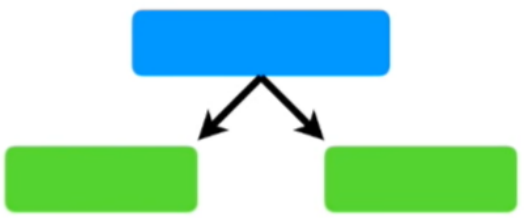
Chest Pain	Blocked Arteries	Patient Weight	Heart Disease	Sample Weight
Yes	Yes	205	Yes	1/8
No	Yes	180	Yes	1/8
Yes	No	210	Yes	1/8
Yes	Yes	167	Yes	1/8
No	Yes	156	No	1/8
No	Yes	125	No	1/8
Yes	No	168	No	1/8
Yes	Yes	172	No	1/8

This is done finding the variable, **Chest Pain**, **Blocked Arteries** or **Patient Weight**, that does the best job classifying the samples.



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

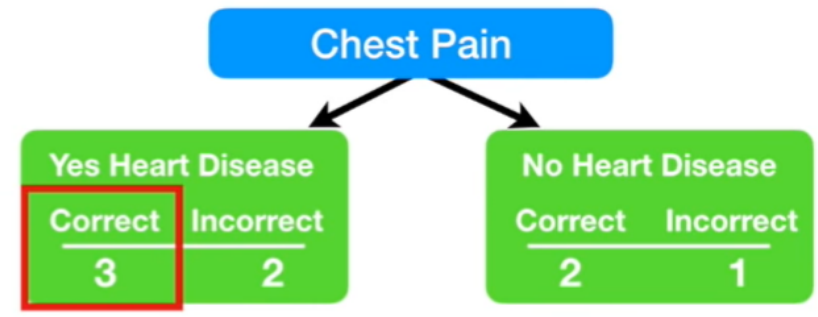
NOTE: Because all of the weights are the same, we can ignore them right now.



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

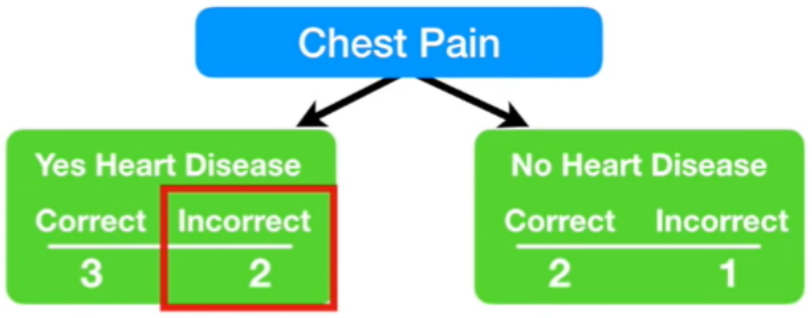
We start by seeing how well **Chest Pain** classifies the samples.

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



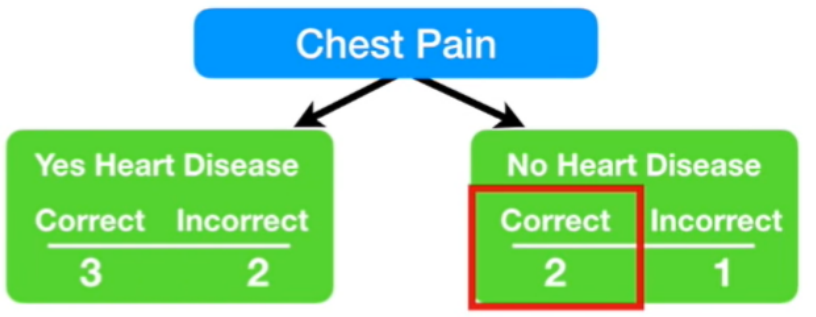
Of the 5 samples with **Chest Pain**, 3 were *correctly* classified as having **Heart Disease**…

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



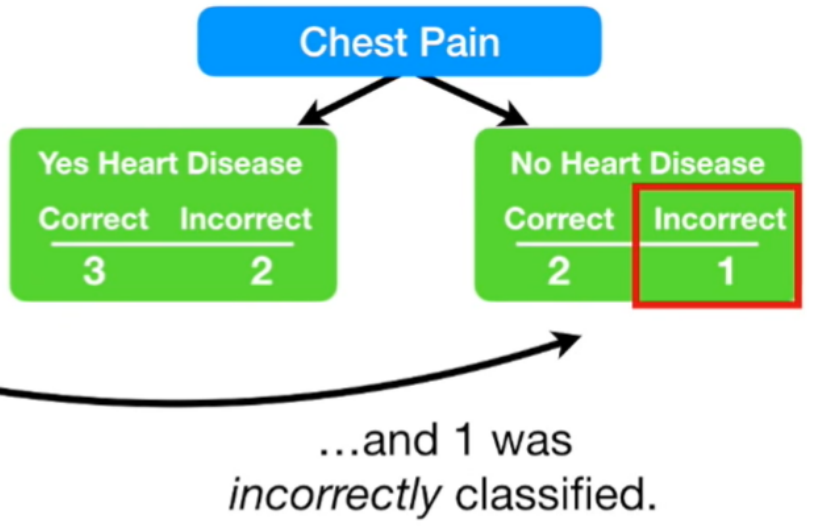
…and 2 were *incorrectly* classified.

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



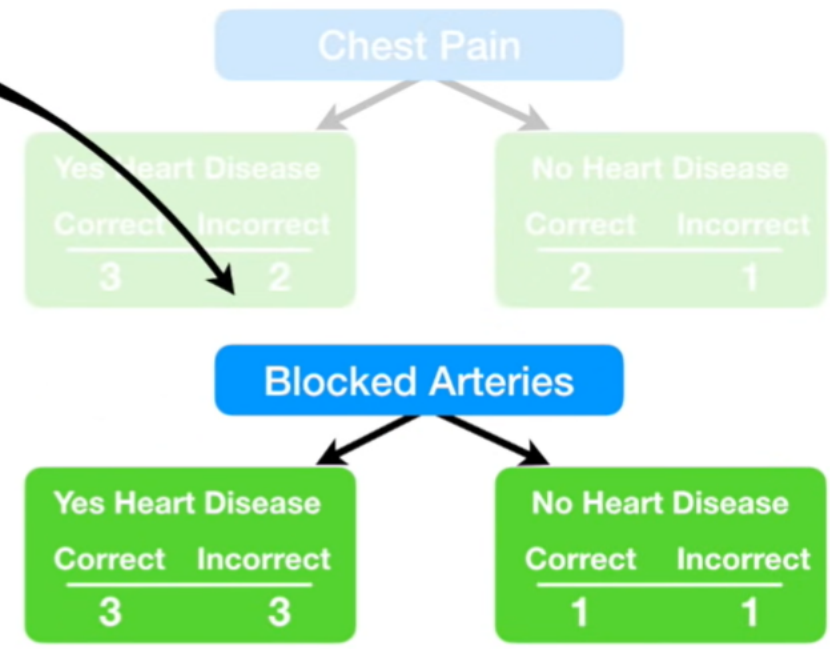
Of the 3 samples *without* **Chest Pain**, 2 were *correctly* classified as *not* having **Heart Disease**…

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



...and 1 was *incorrectly* classified.

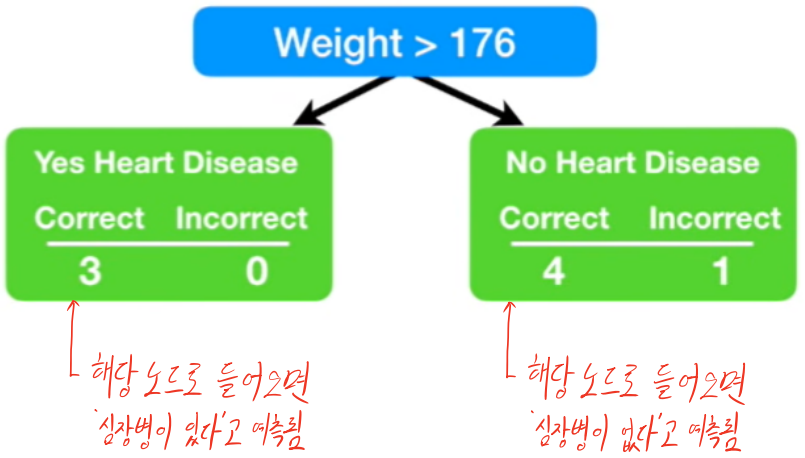
Chest Pain	Blocked Arteries	Patient Weight	Heart Disease	Sample Weight
Yes	Yes	205	Yes	1/8
No	Yes	180	Yes	1/8
Yes	No	210	Yes	1/8
Yes	Yes	167	Yes	1/8
No	Yes	156	No	1/8
No	Yes	125	No	1/8
Yes	No	168	No	1/8
Yes	Yes	172	No	1/8



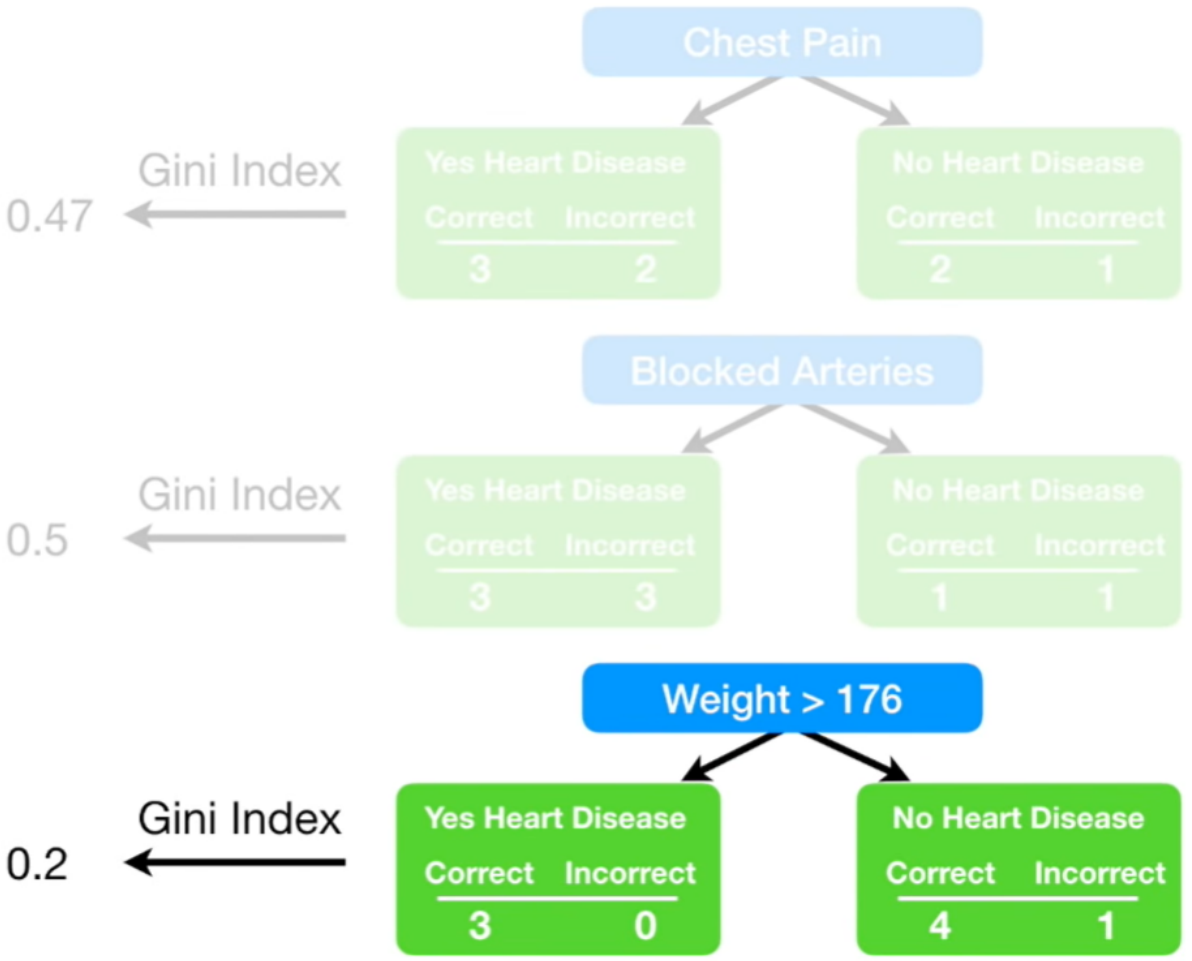
Now we do the same thing for **Blocked Arteries**...

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

…and for **Patient Weight.**



Now we calculate the **Gini Index** for the three stumps.



Chest Pain

Yes Heart Disease: Correct 3, Incorrect 2

No Heart Disease: Correct 2, Incorrect 1

Gini Index 0.47

Blocked Arteries

Yes Heart Disease: Correct 3, Incorrect 3

No Heart Disease: Correct 1, Incorrect 1

Gini Index 0.5

The **Gini Index** for **Patient Weight** is the lowest…

Weight > 176

Yes Heart Disease: Correct 3, Incorrect 0

No Heart Disease: Correct 4, Incorrect 1

Gini Index 0.2

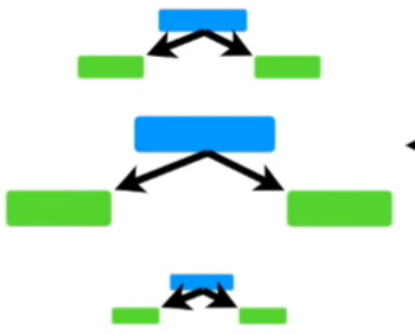
…so this will be the first stump in the forest.

Weight > 176

Yes Heart Disease: Correct 3, Incorrect 0

No Heart Disease: Correct 4, Incorrect 1

Now we need to determine how much say this stump will have in the final classification.



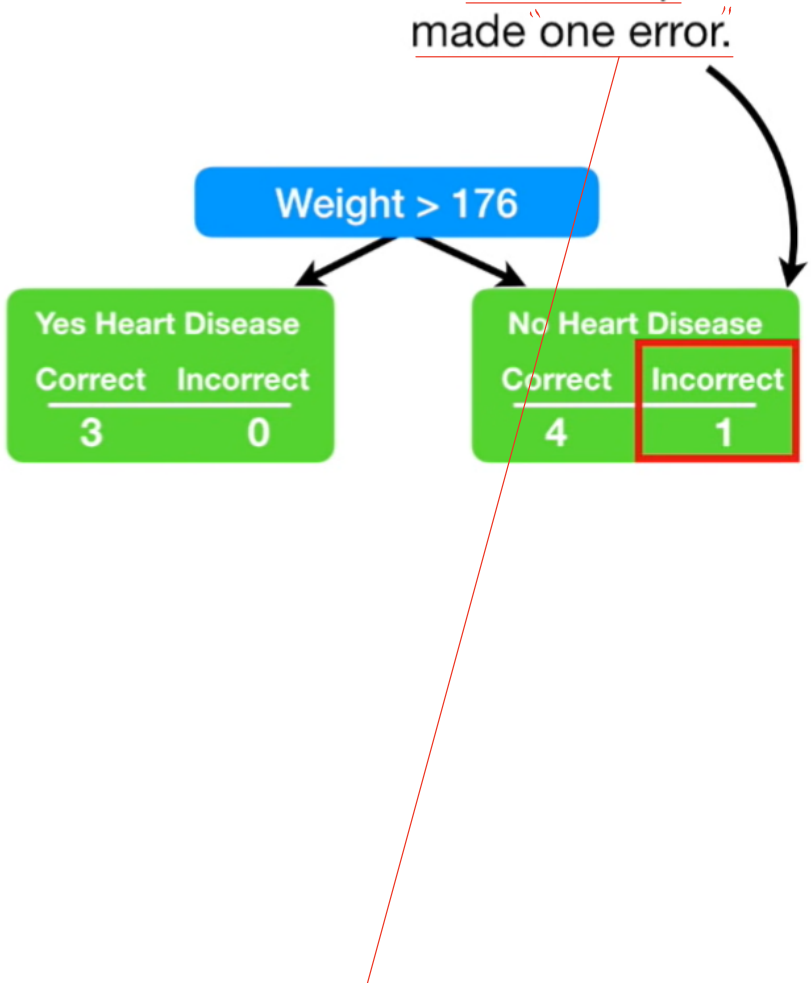
Remember, some stumps get more say in the final classification than others.

Weight > 176
Yes Heart Disease
Correct
Incorrect
3
0
No Heart Disease
Correct
Incorrect
4
1

We determine how much say a stump has in the final classification based on how well it classified the samples.

전차가 적은 stump가 더 많은 "말의 영향력"을 가진다.

This stump made "one error."



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

This patient, who weighs less than 176, *has* heart disease, but the stump says they do not.

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

The **Total Error** for a stump is the sum of the weights associated with the *incorrectly* classified samples.

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

Thus, in this case, the **Total Error** is **1/8**.

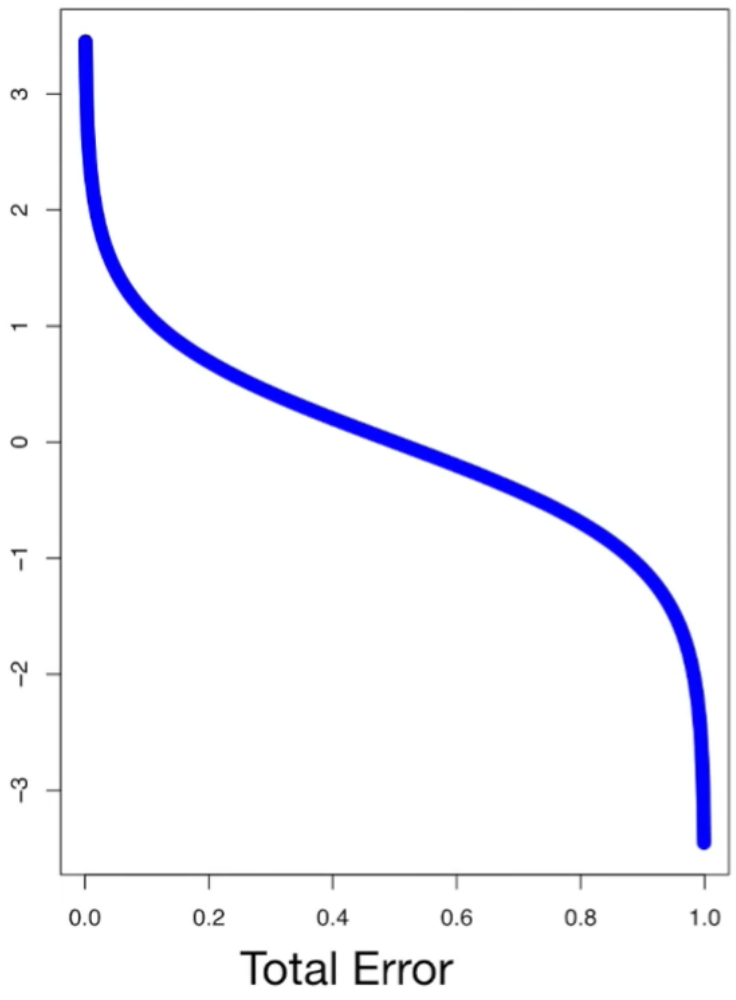
Chest Pain	Blocked Arteries	Patient Weight	Heart Disease	Sample Weight
Yes	Yes	205	Yes	1/8
No	Yes	180	Yes	1/8
Yes	No	210	Yes	1/8
Yes	Yes	167	Yes	1/8
No	Yes	156	No	1/8
No	Yes	125	No	1/8
Yes	No	168	No	1/8
Yes	Yes	172	No	1/8

NOTE: Because all of the **Sample Weights** add up to **1**, **Total Error** will always be between **0**, for a perfect stump, and **1**, for a horrible stump.

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

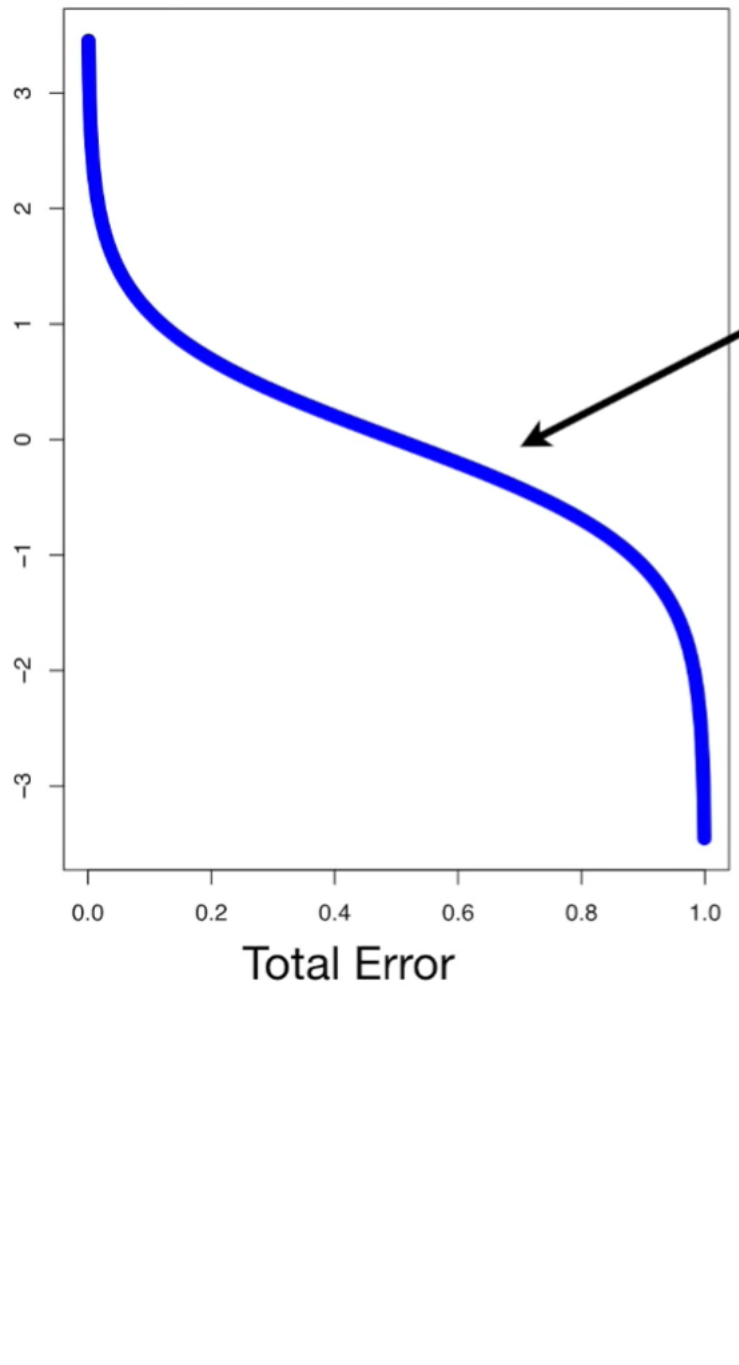
We use the **Total Error** to determine **Amount of Say** this stump has in the final classification with the following formula:

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



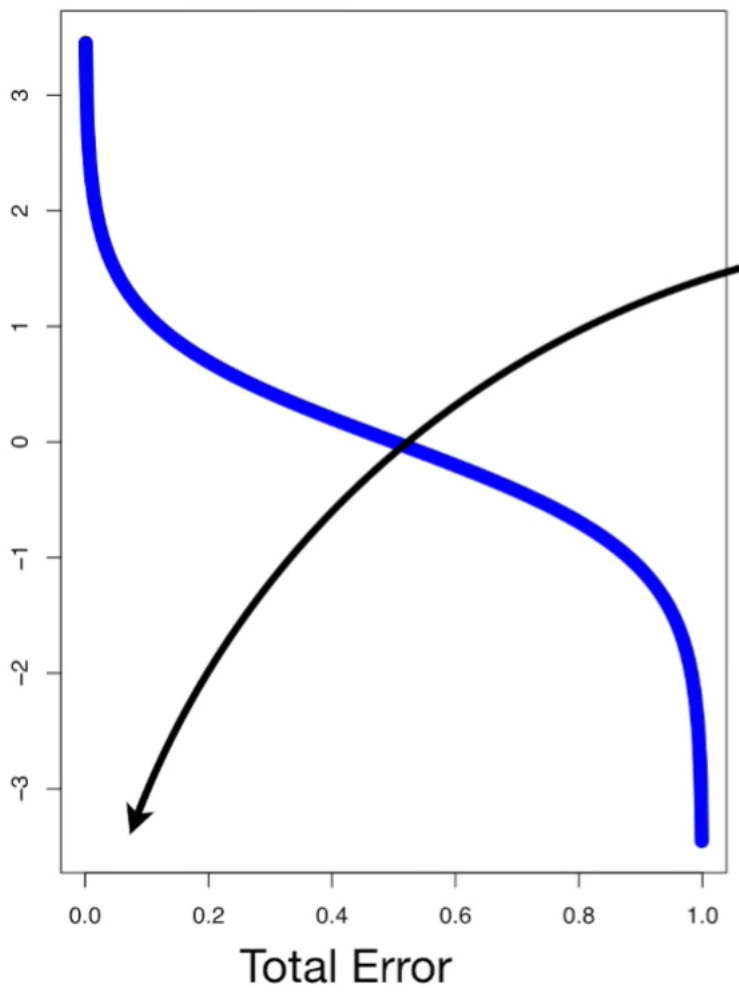
We can draw a graph of the **Amount of Say** by plugging in a bunch of numbers between **0** and **1** for **Total Error**.

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



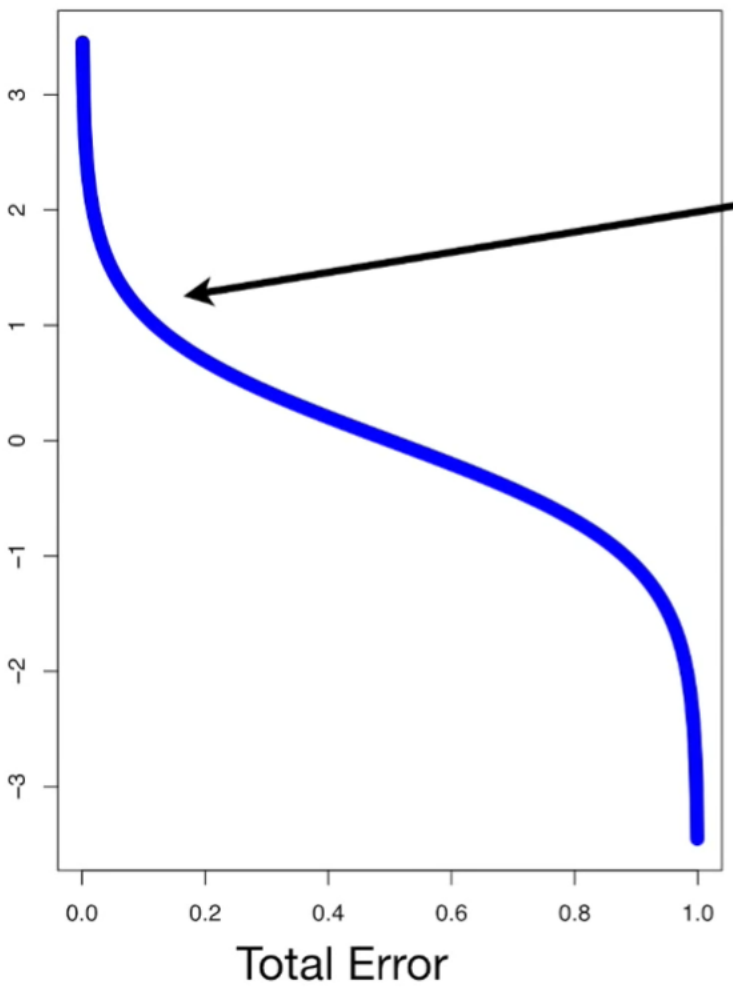
The **Blue Line** tells us the **Amount of Say** for **Total Error** values between **0** and **1**.

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



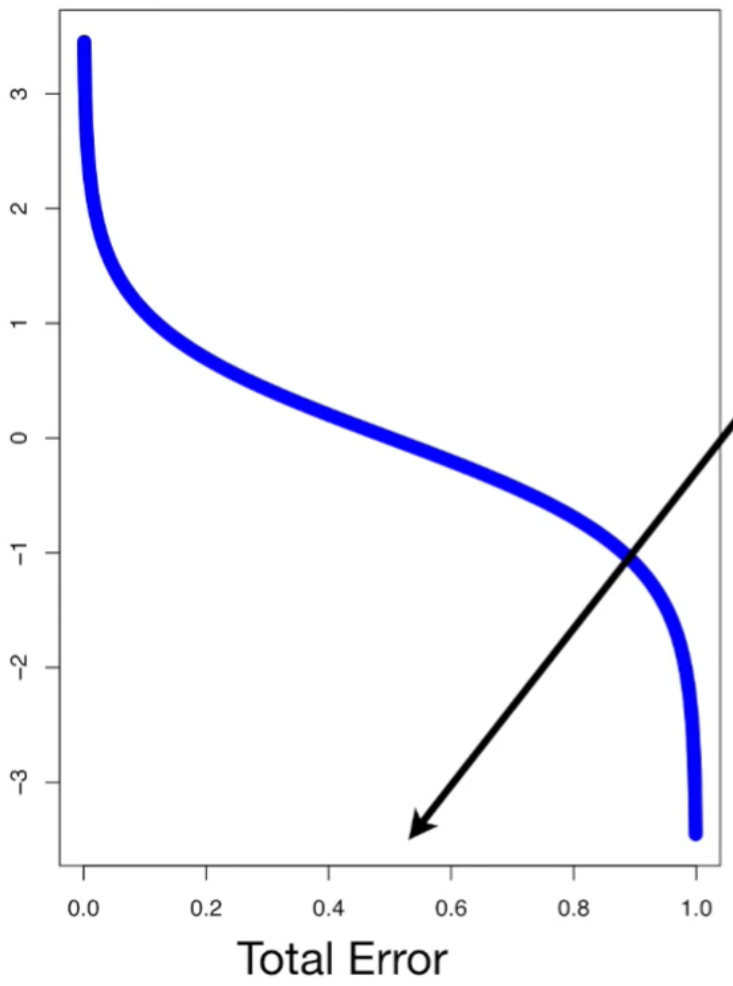
When a stump does a good job, and the **Total Error** is small…

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



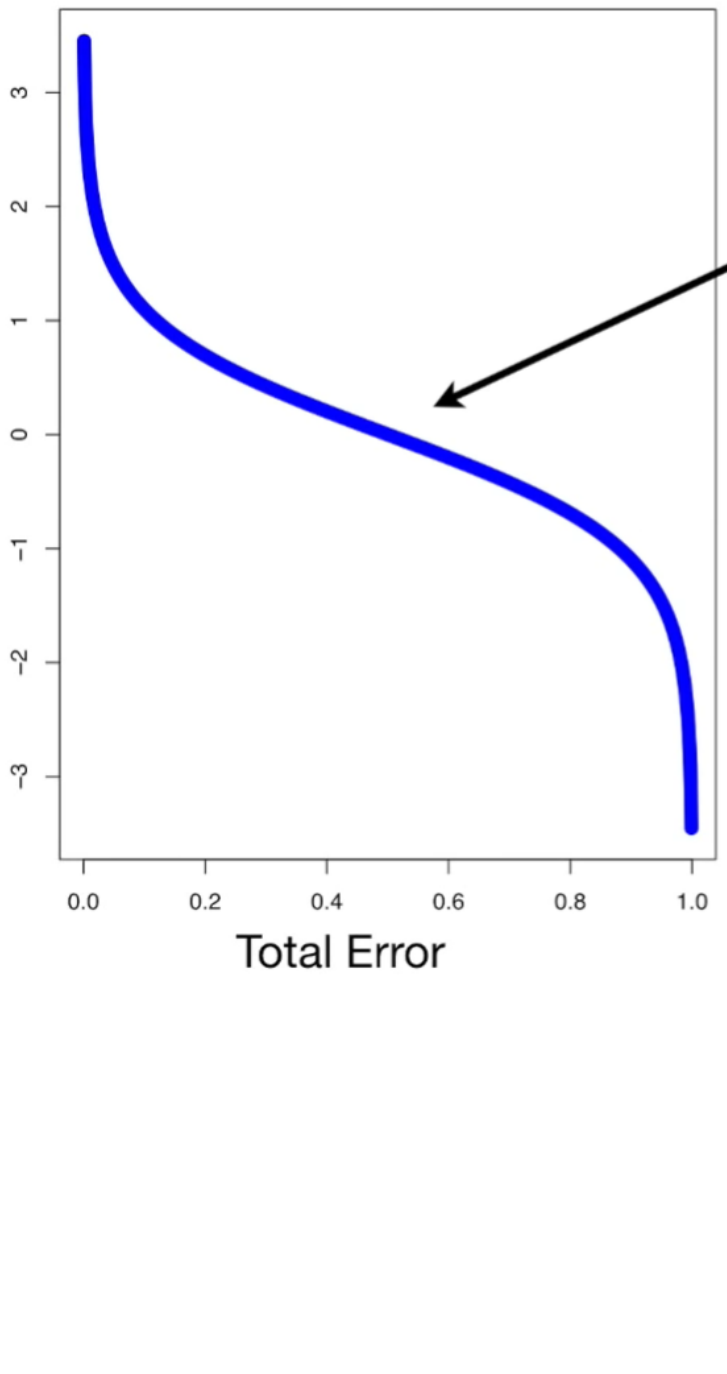
…then the **Amount of Say** is a relatively large, *positive* value.

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



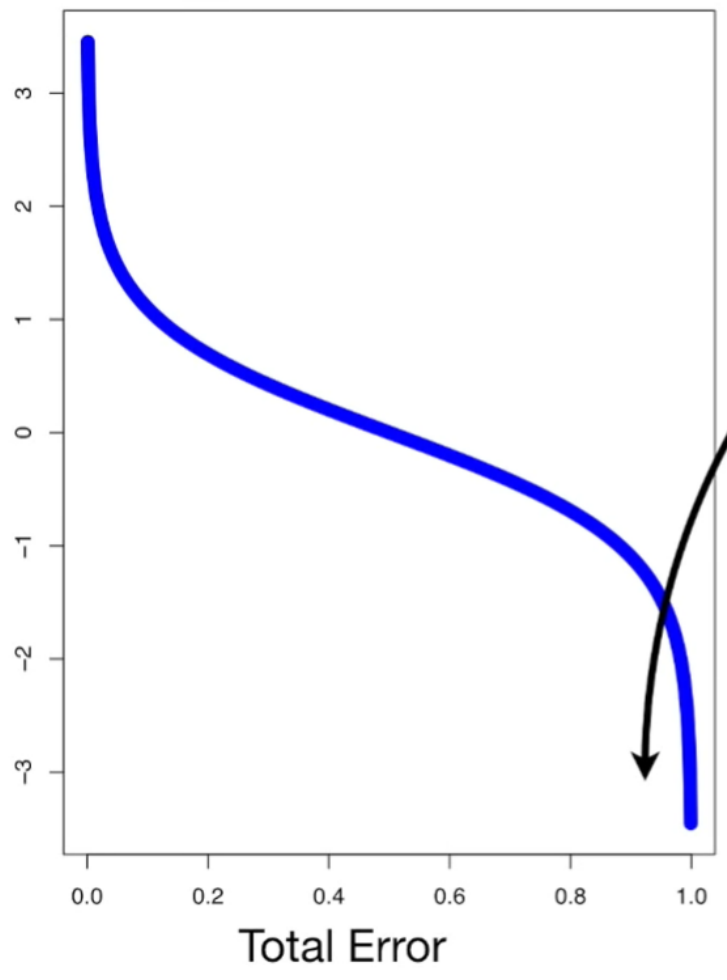
When a stump is no better at classification than flipping a coin (i.e. half of the samples are correctly classified and half are incorrectly classified) and **Total Error** = **0.5**…

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



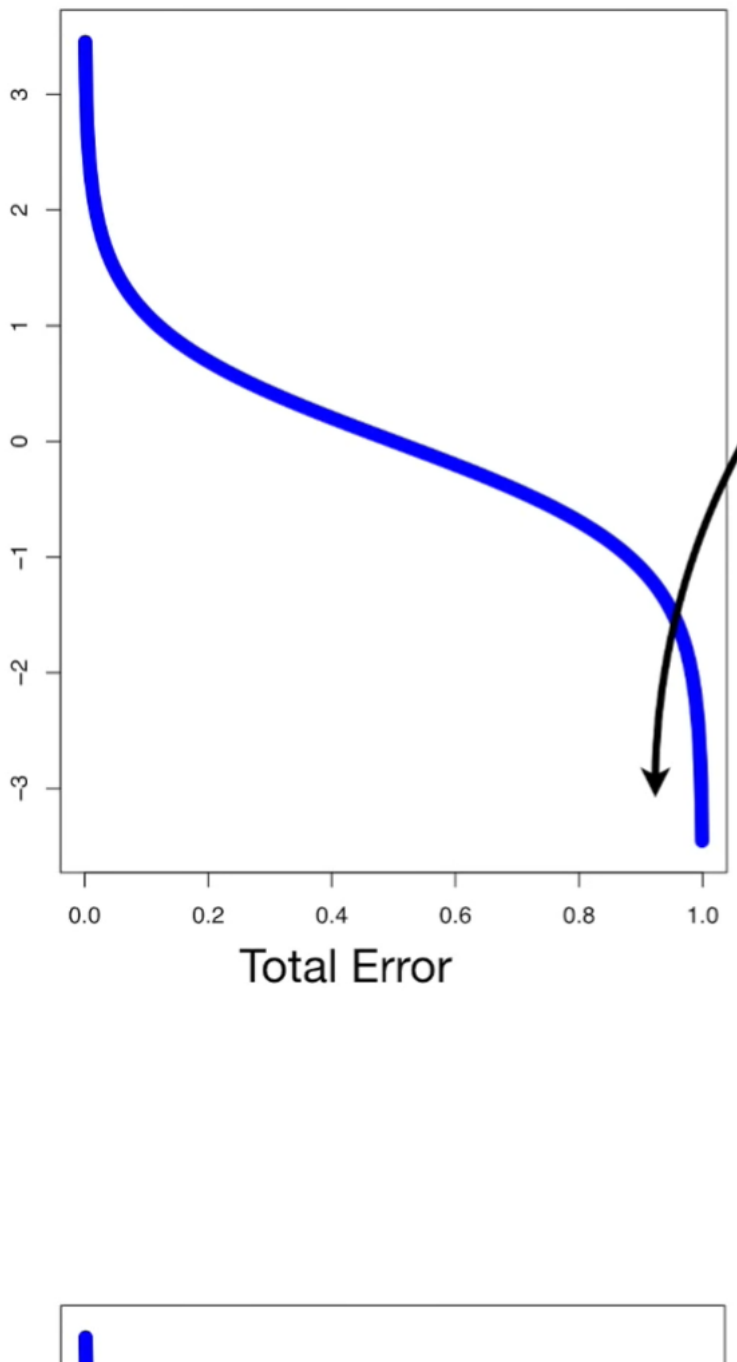
…then the **Amount of Say** will be **0**.

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



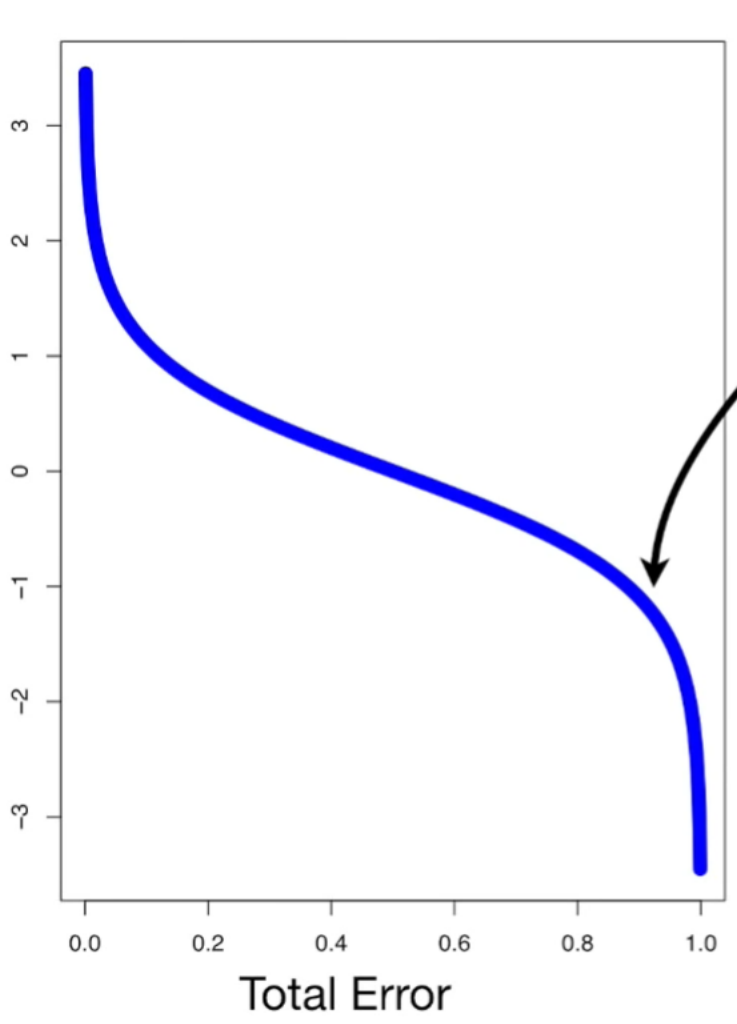
And when a stump does a terrible job and the **Total Erro**r is close to **1**…

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



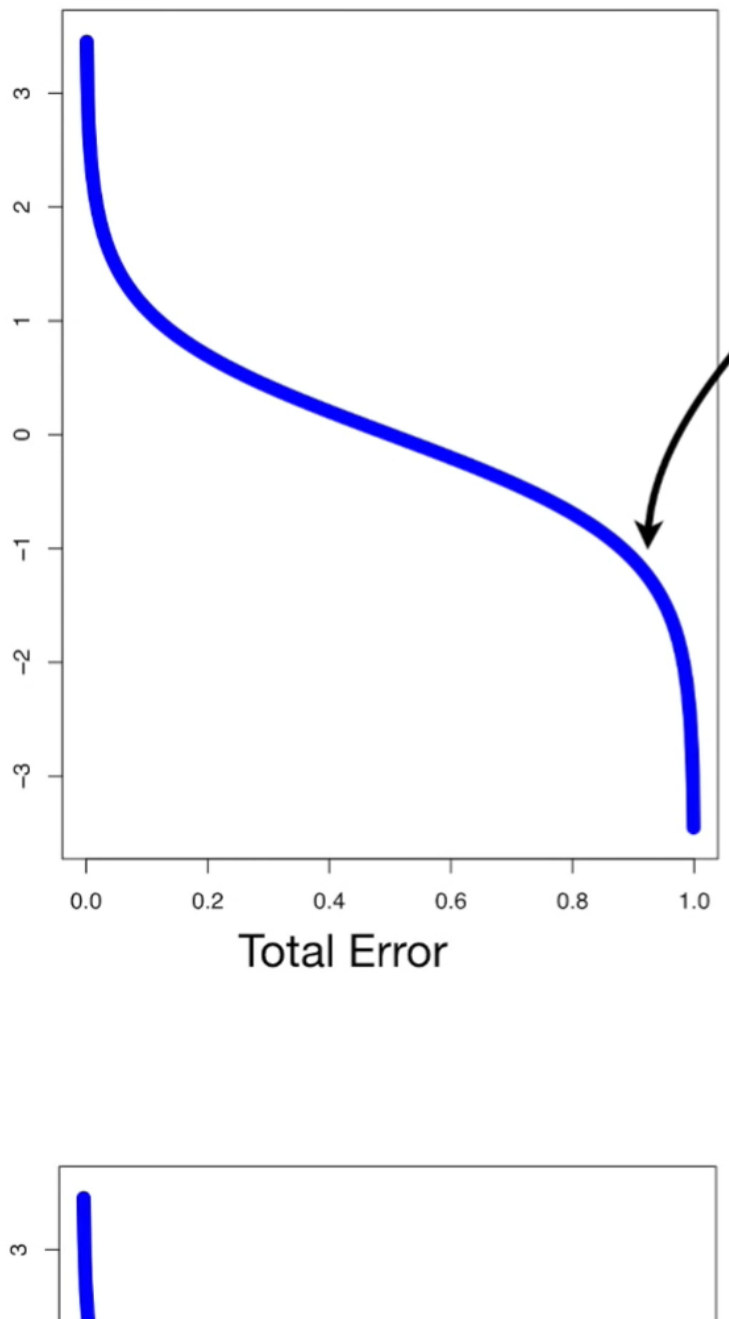
…in other words, if the stump consistently gives you the opposite classification…

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



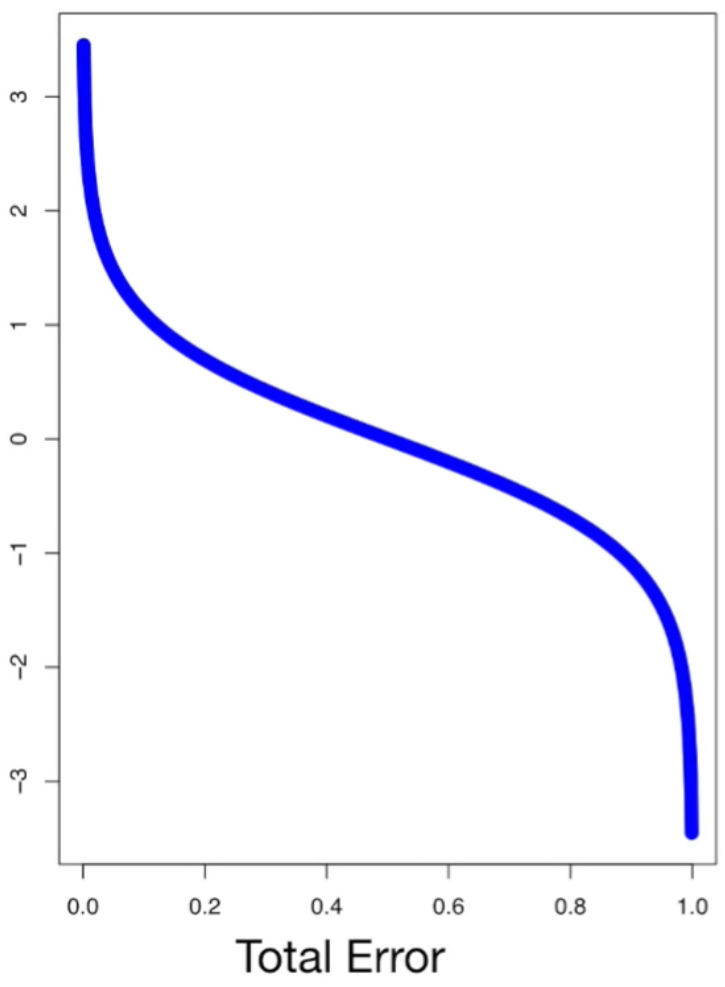
…then the **Amount of Say** will be a large *negative* value.

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



So if a stump votes for "Heart Disease", the negative **Amount of Say** will turn that vote into "*Not* Heart Disease".

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

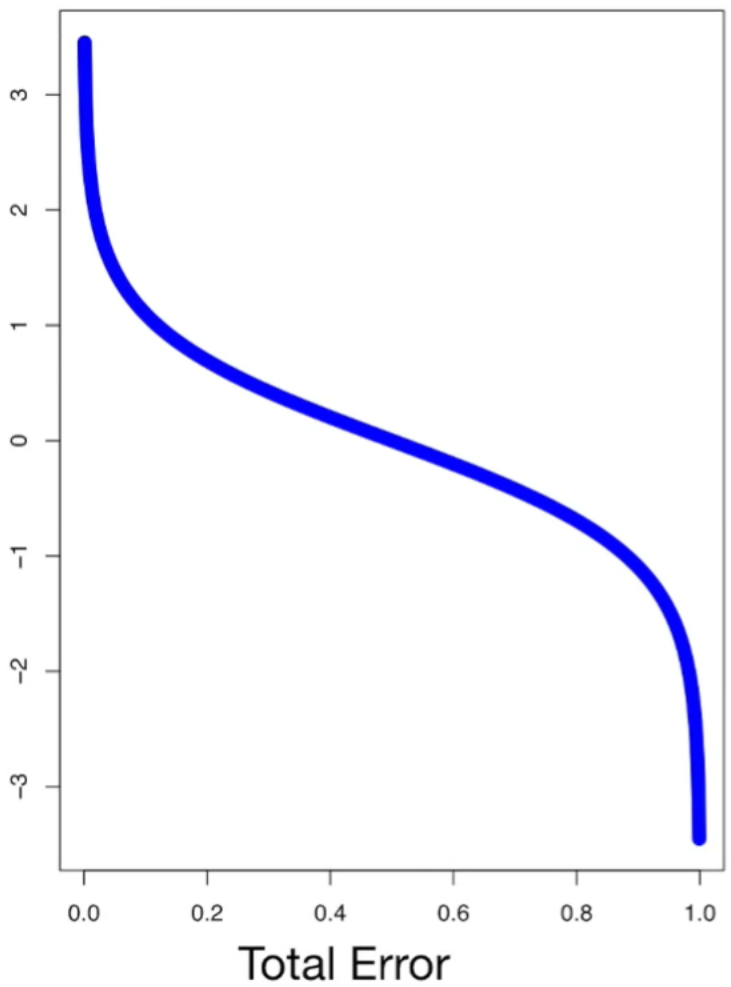


NOTE: If **Total Error** is **1** or **0**, then this equation will freak out.

$\log 0$: error

$\frac{x}{0}$: error

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



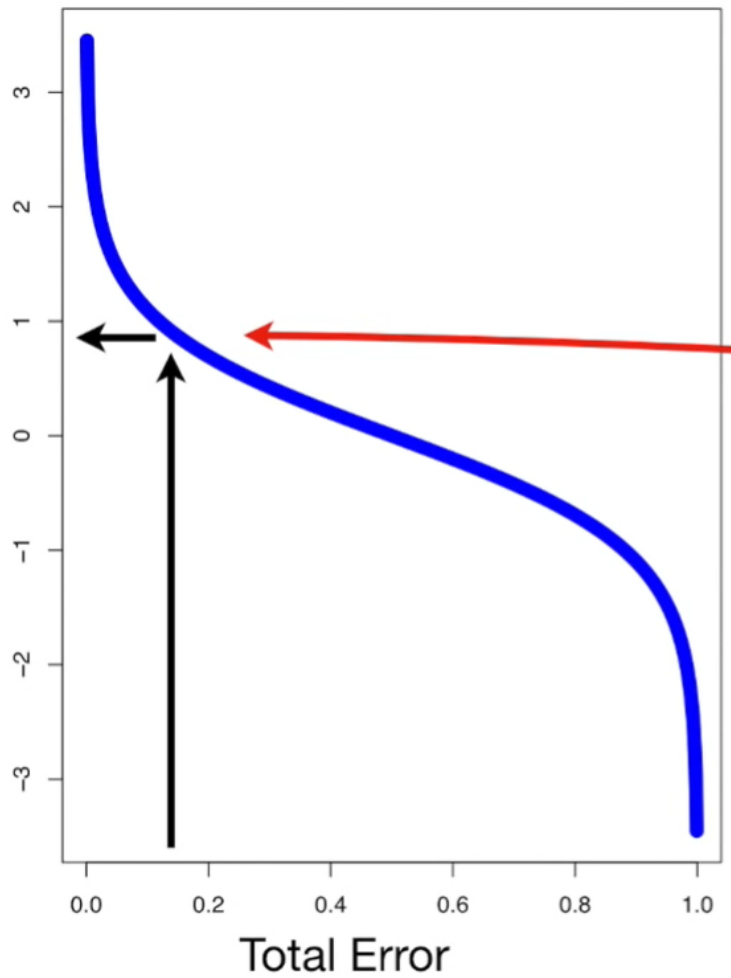
In practice a small error term is added to prevent this from happening.

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

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
Yes	No	168	No	1/8
Yes	Yes	172	No	1/8

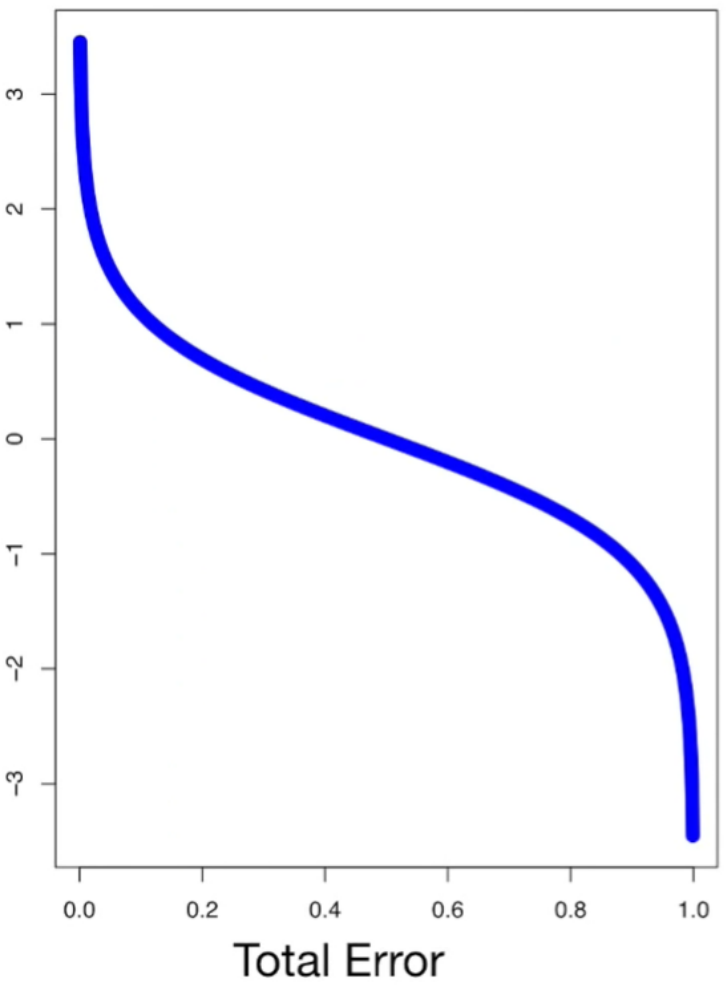
With **Patient Weight > 176**, the **Total Error** is **1/8**, so we just plug and chug…

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



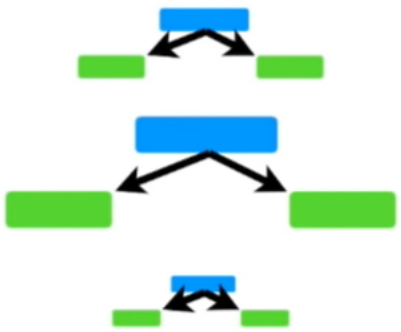
…and the **Amount of Say** that this stump has on the final classification is **0.97**.

$$\text{Amount of Say} = \frac{1}{2}\log(7) = 0.97$$



Now that we've worked out how much say this stump gets when classifying a sample….

Now we know how the **Sample Weights** for the *incorrectly* classified samples are used to determine the **Amount of Say** each stump gets.



Now we need to learn how to modify the weights so that the next stump will take the errors that the current stump made into account.

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

Let's go back to the first stump that we made.

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

When we created this stump, all of the **Sample Weights** were the same…

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

…and that meant we did not emphasize the importance of correctly classifying any particular sample…

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

…but since this stump *incorrectly* classified this sample…

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

…we will emphasize the need for the next stump to correctly classify it by increasing its **Sample 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

…and decreasing all of the other **Sample Weights**.

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

Let's start by increasing the **Sample Weight** for the *incorrectly* classified sample.

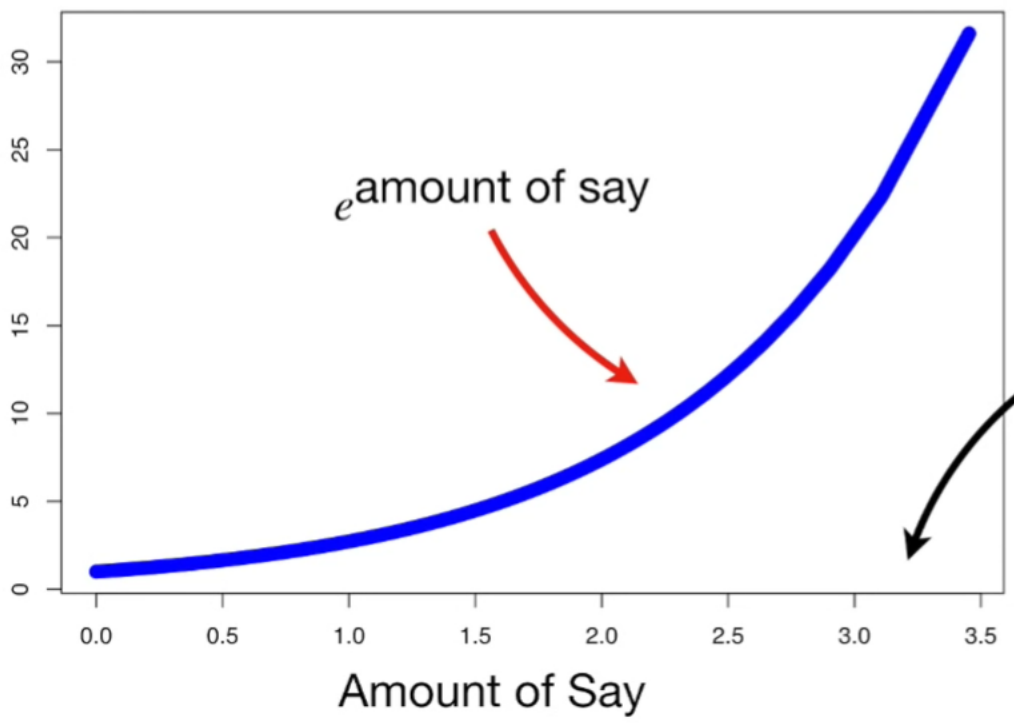
Chest Pain	Blocked Arteries	Patient Weight	Heart Disease	Sample Weight
Yes	Yes	205	Yes	1/8
No	Yes	180	Yes	1/8
Yes	No	210	Yes	1/8
Yes	**Yes**	**167**	**Yes**	**1/8**
No	Yes	156	No	1/8
No	Yes	125	No	1/8
Yes	No	168	No	1/8
Yes	Yes	172	No	1/8

$$\text{New Sample Weight} = \text{sample weight} \times e^{\text{amount of say}}$$

This is the formula we will use to *increase* the **Sample Weight** for the sample that was *incorrectly* classified.

$$\text{New Sample Weight} = \text{sample weight} \times e^{\text{amount of say}}$$

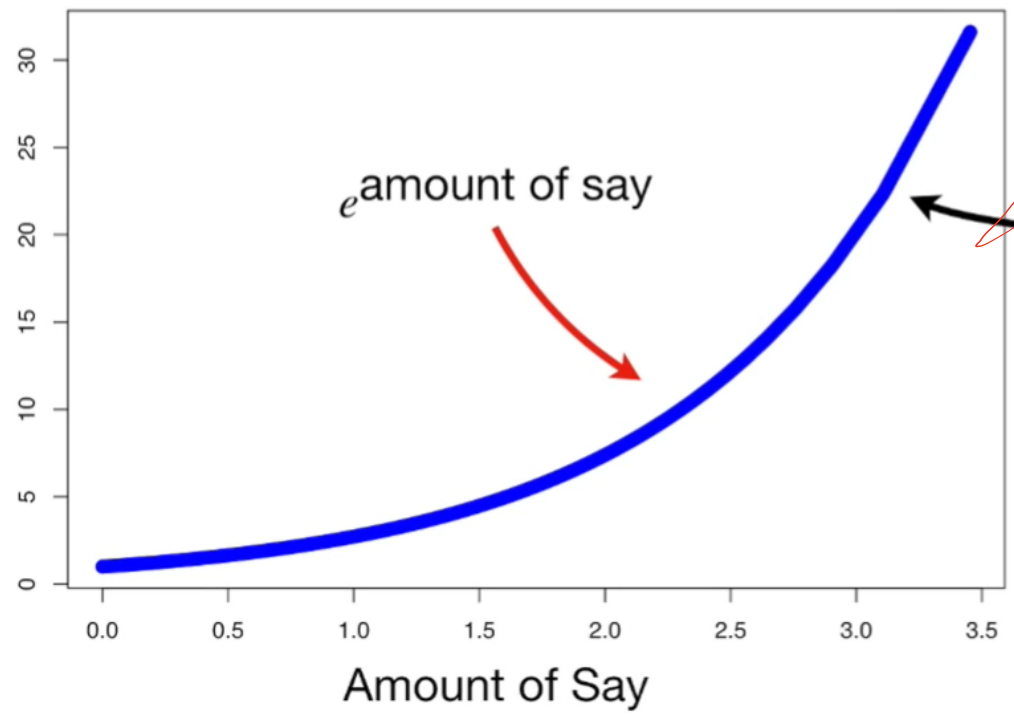
$$= \frac{1}{8} e^{\text{amount of say}}$$



When the **Amount of Say** is relatively large, (i.e. the last stump did a good job classifying samples)…

$$\text{New Sample Weight} = \text{sample weight} \times e^{\text{amount of say}}$$

$$= \frac{1}{8} e^{\text{amount of say}}$$

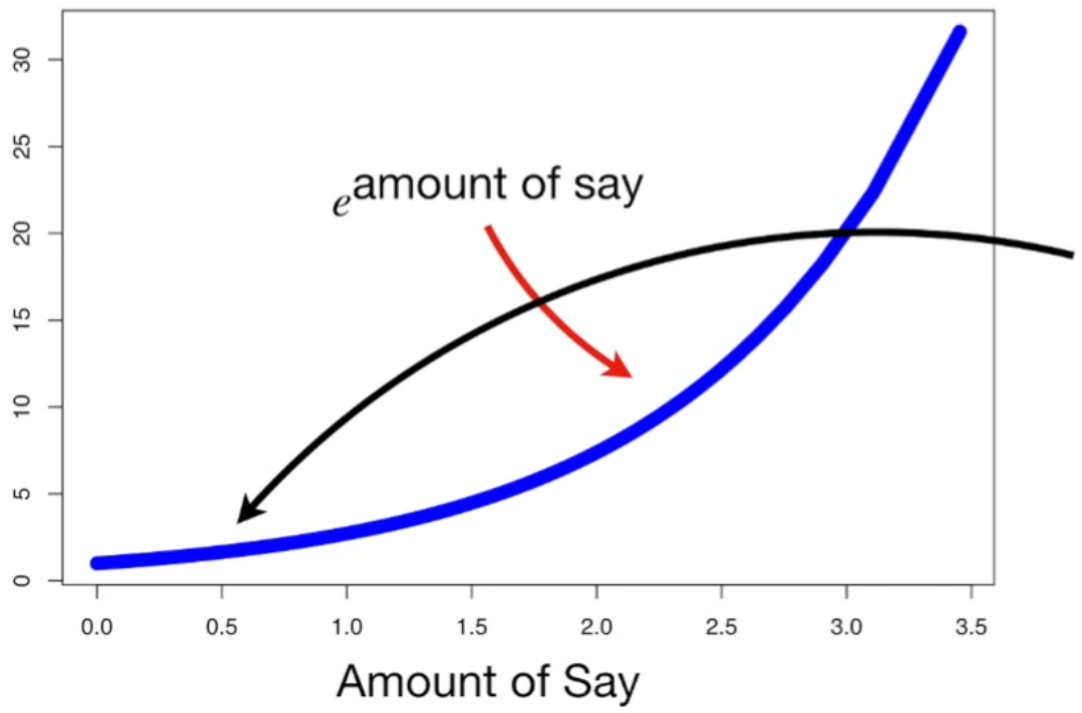


…then we will scale the previous **Sample Weight** with a large number.

This means that the **New Sample Weight** will be much larger than the old one.

$$\text{New Sample Weight} = \text{sample weight} \times e^{\text{amount of say}}$$

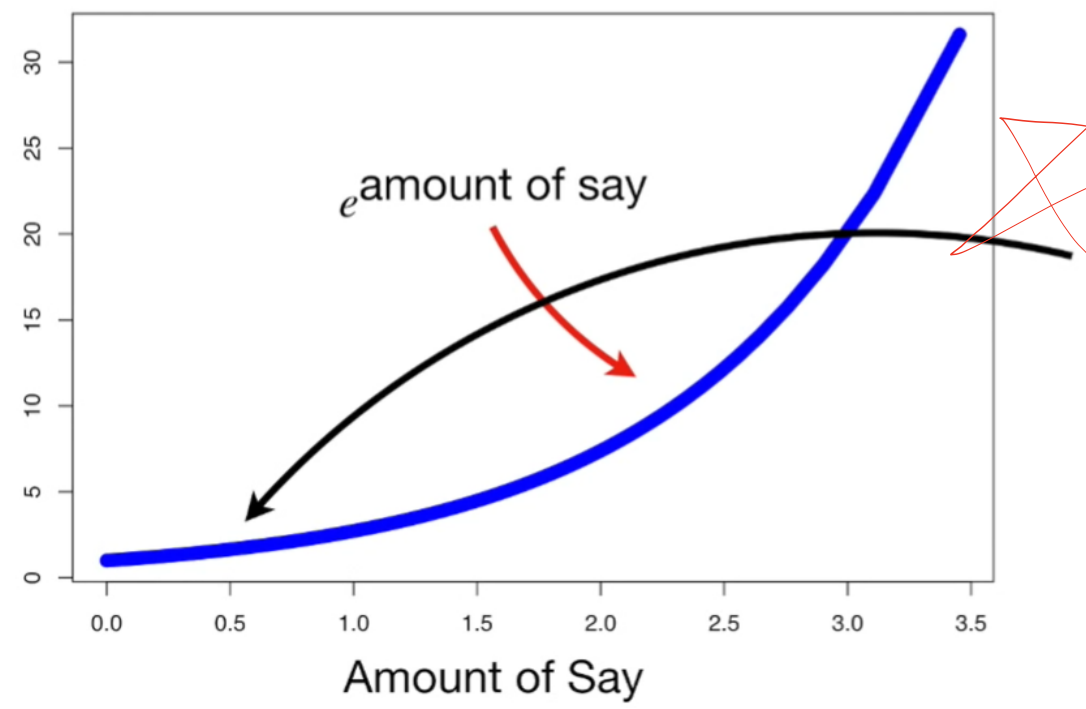
$$= \frac{1}{8} e^{\text{amount of say}}$$



And when the **Amount of Say** is relatively low (i.e. the last stump did not do a very good job classifying samples)…

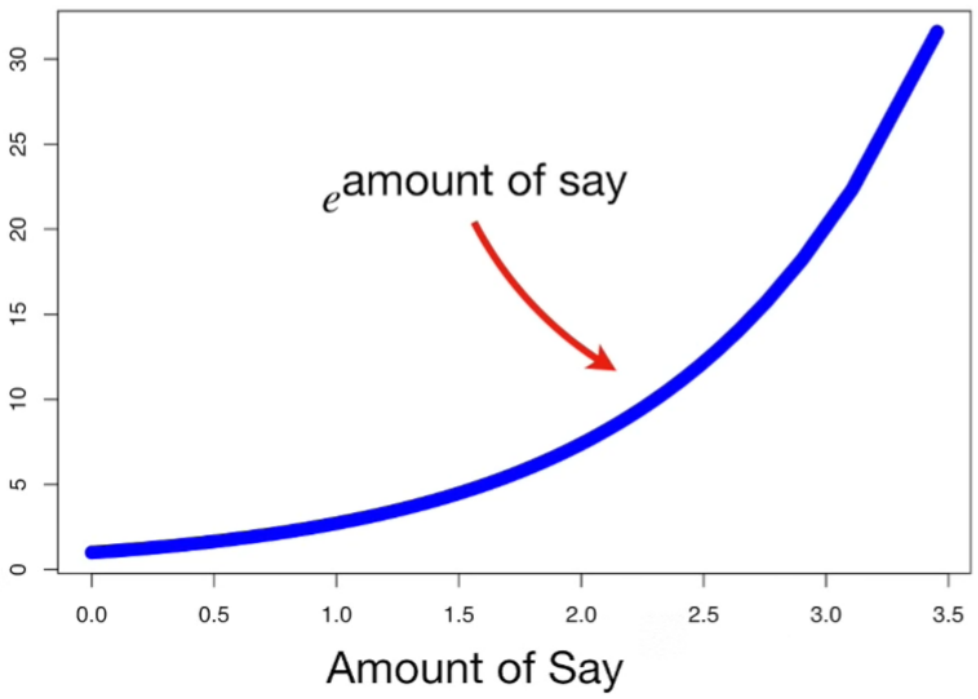
$$\text{New Sample Weight} = \text{sample weight} \times e^{\text{amount of say}}$$

$$= \frac{1}{8} e^{\text{amount of say}}$$



…then the previous **Sample Weight** is scaled by a relatively small number.

This means that the **New Sample Weight** will only be a little larger than the old one.



$$\text{New Sample Weight} = \text{sample weight} \times e^{\text{amount of say}}$$

$$= \frac{1}{8} e^{\text{amount of say}}$$

$$= \frac{1}{8} e^{0.97} = \frac{1}{8} \times 2.64 = 0.33$$

That means the new **Sample Weight** is **0.33**, which is *more* than the old one (**1/8** = **0.125**).

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

Now we need to *decrease* the **Sample Weights** for all of the *correctly* classified samples.

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

$$\text{New Sample Weight} = \text{sample weight} \times e^{\text{-amount of say}}$$

This is the formula we will use to *decrease* the **Sample Weights**.

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

$$\text{New Sample Weight} = \text{sample weight} \times e^{\text{-amount of say}}$$

The big difference is the *negative* sign in front of **Amount of Say**

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

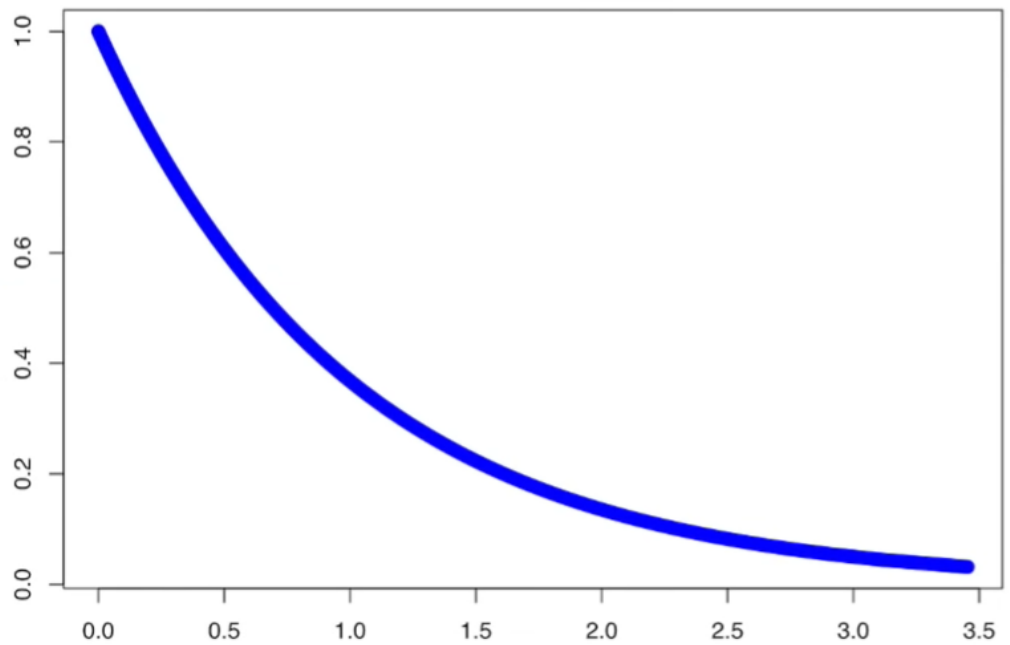
$$\text{New Sample Weight} = \text{sample weight} \times e^{\text{-amount of say}}$$

$$= \frac{1}{8} e^{\text{-amount of say}}$$

Just like before, we plug in the **Sample Weight**…

$$\text{New Sample Weight} = \text{sample weight} \times e^{\text{-amount of say}}$$

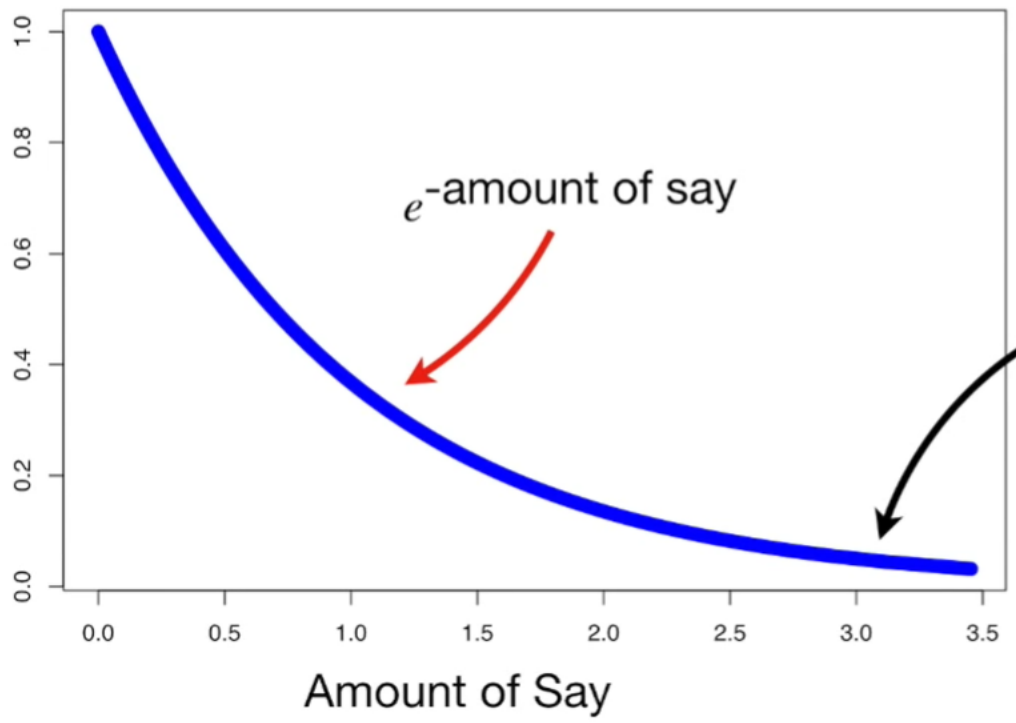
$$= \frac{1}{8} e^{\text{-amount of say}}$$



…and just like before, we can get a better understanding of how this will scale the **Sample Weight** by plotting a graph using different values for **Amount of Say**.

New Sample Weight = sample weight × e-amount of say

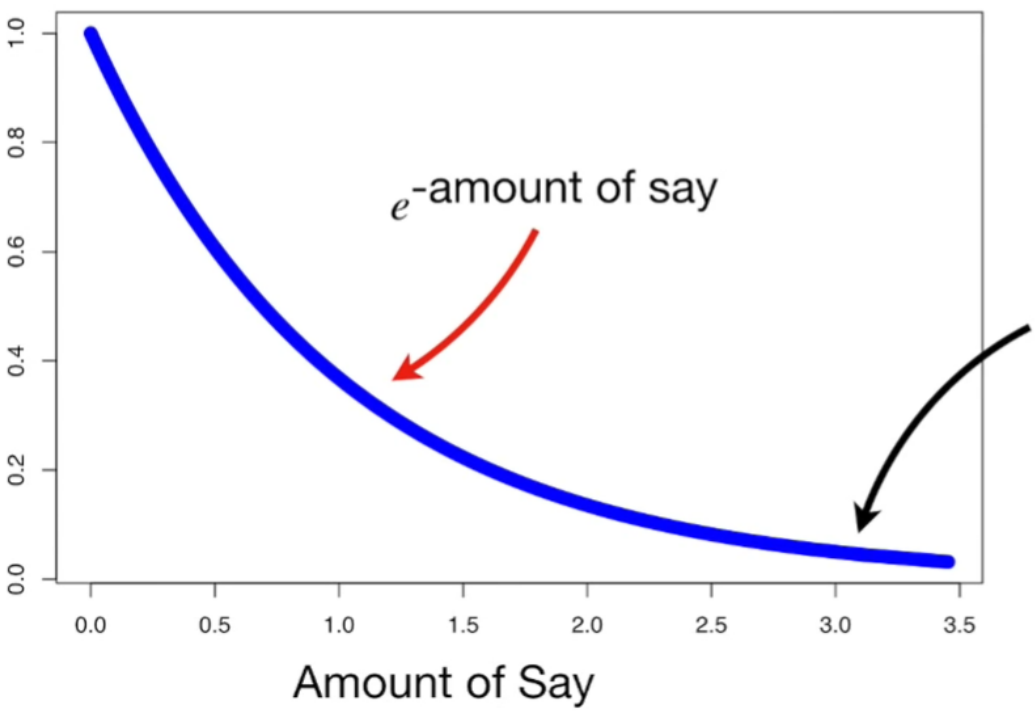
= $\frac{1}{8}$ e-amount of say



When the **Amount of Say** is relatively large…

New Sample Weight = sample weight × e-amount of say

= $\frac{1}{8}$ e-amount of say

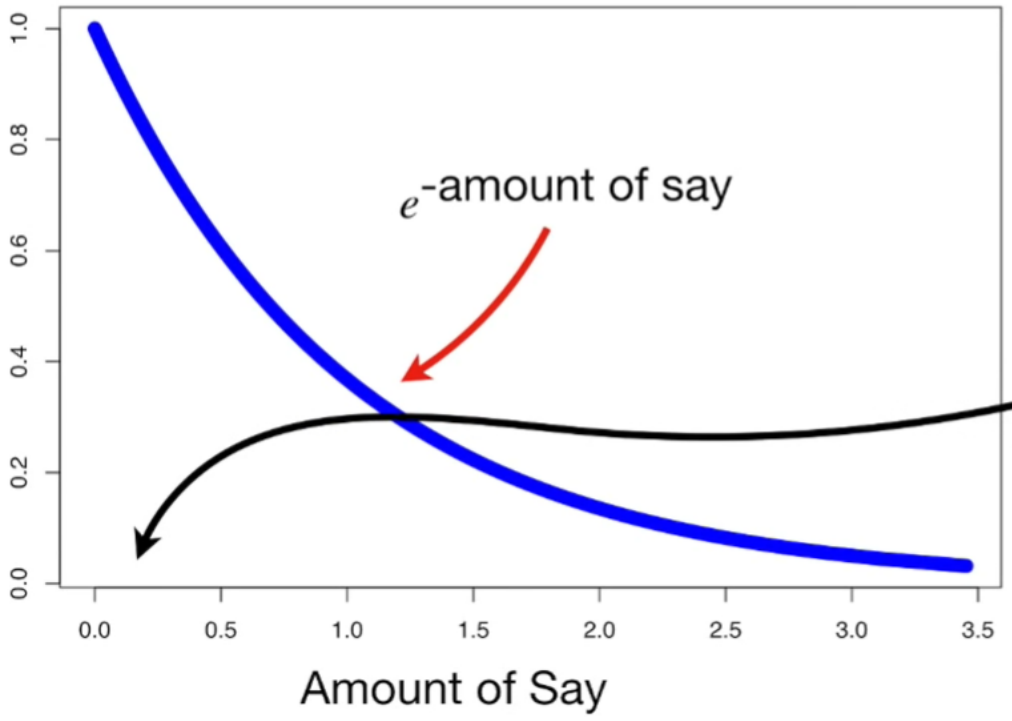


…then we scale the **Sample Weight** by a value very close to 0.

This will make the **New Sample Weight** very small.

$$\text{New Sample Weight} = \text{sample weight} \times e^{\text{-amount of say}}$$

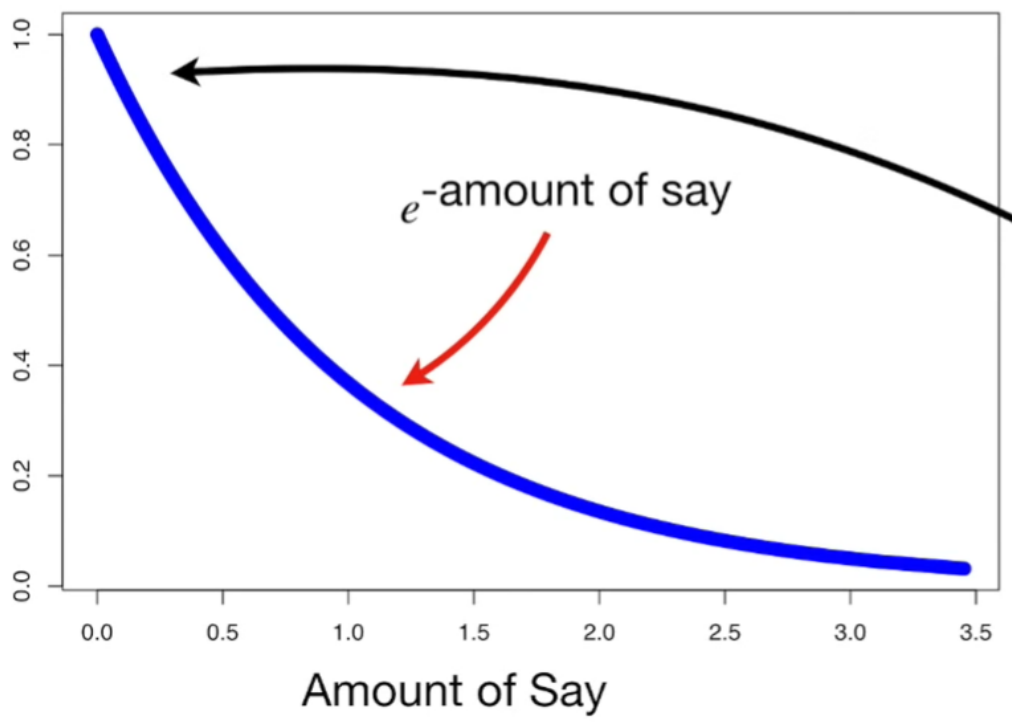
$$= \frac{1}{8} e^{\text{-amount of say}}$$



If the **Amount of Say** for the last stump is relatively small…

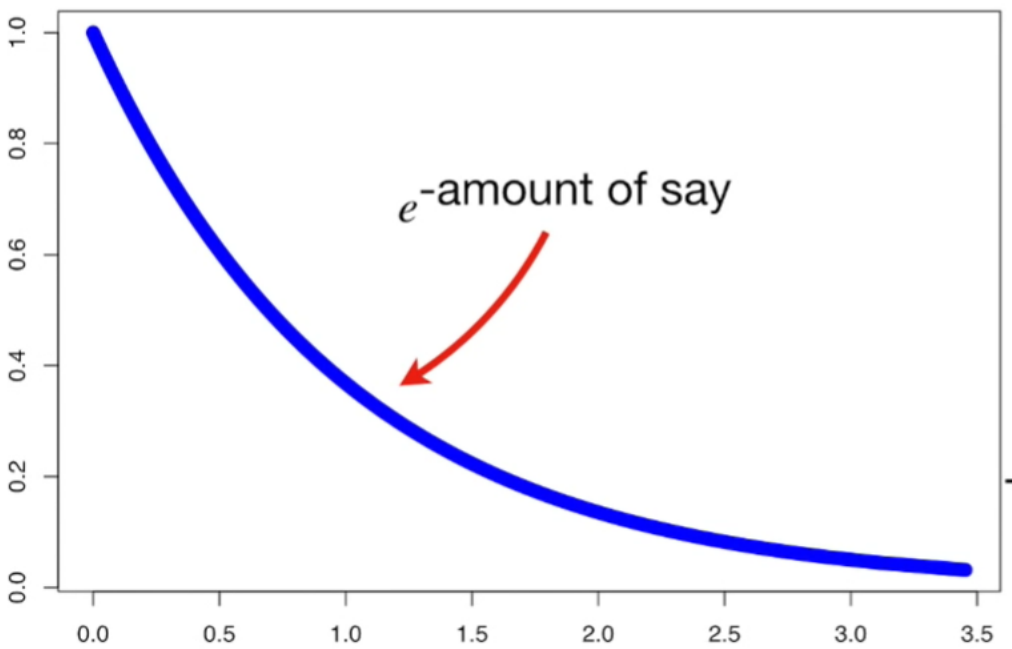
$$\text{New Sample Weight} = \text{sample weight} \times e^{\text{-amount of say}}$$

$$= \frac{1}{8} e^{\text{-amount of say}}$$



…then we will scale the **Sample Weight** by a value close to **1**.

This means that the **New Sample Weight** will be just a little smaller than the old one.



$$\text{New Sample Weight} = \text{sample weight} \times e^{\text{-amount of say}}$$

$$= \frac{1}{8} e^{\text{-amount of say}}$$

$$= \frac{1}{8} e^{-0.97} = \frac{1}{8} \times 0.38 = 0.05$$

The new **Sample Weight** is **0.05**, which is *less* than the old one (**1/8** = **0.125**).

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

Now we need to normalize the **New Sample Weights** so that they will add up to 1.

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

Right now, if you add up the **New Sample Weights**, you get **0.68**.

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

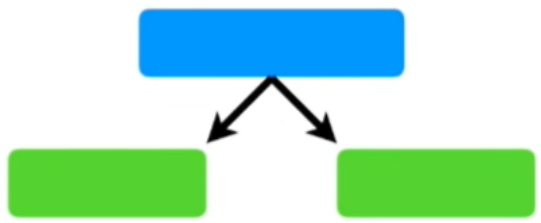
So we divide each **New Sample Weight** by **0.68** to get the normalized values.

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.

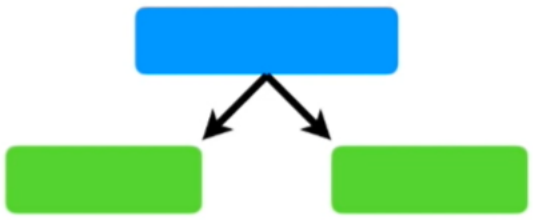
Chest Pain	Blocked Arteries	Patient Weight	Heart Disease	Sample Weight
Yes	Yes	205	Yes	0.07
No	Yes	180	Yes	0.07
Yes	No	210	Yes	0.07
Yes	Yes	167	Yes	0.49
No	Yes	156	No	0.07
No	Yes	125	No	0.07
Yes	No	168	No	0.07
Yes	Yes	172	No	0.07

Now we can use the modified **Sample Weights** to make the second **stump** in the forest.



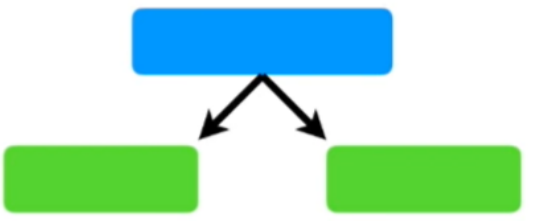
Chest Pain	Blocked Arteries	Patient Weight	Heart Disease	Sample Weight
Yes	Yes	205	Yes	0.07
No	Yes	180	Yes	0.07
Yes	No	210	Yes	0.07
Yes	Yes	167	Yes	0.49
No	Yes	156	No	0.07
No	Yes	125	No	0.07
Yes	No	168	No	0.07
Yes	Yes	172	No	0.07

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**.



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

Alternatively, instead of using a **Weighted Gini Index**, we can make a new collection of samples that contains duplicate copies of the samples with the largest **Sample Weights**.

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

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

Then we pick a random number between 0 and 1…

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	Heart Disease

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	Heart Disease

…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	Heart Disease

…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

…and if the number is between **0.21** and **0.70** (**0.21** + **0.49** = **0.70**), 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

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
No	Yes	156	No

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

…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

Then I pick another random number and get **0.42**…

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

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

Then I pick **0.83**…

…and 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

…and 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

Then I pick **0.51**…

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

…and 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	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

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
Yes	Yes	167	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	No	0.07
Yes	No	168	No	0.07
Yes	Yes	172	No	0.07

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

…and use the new collection of 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

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.

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

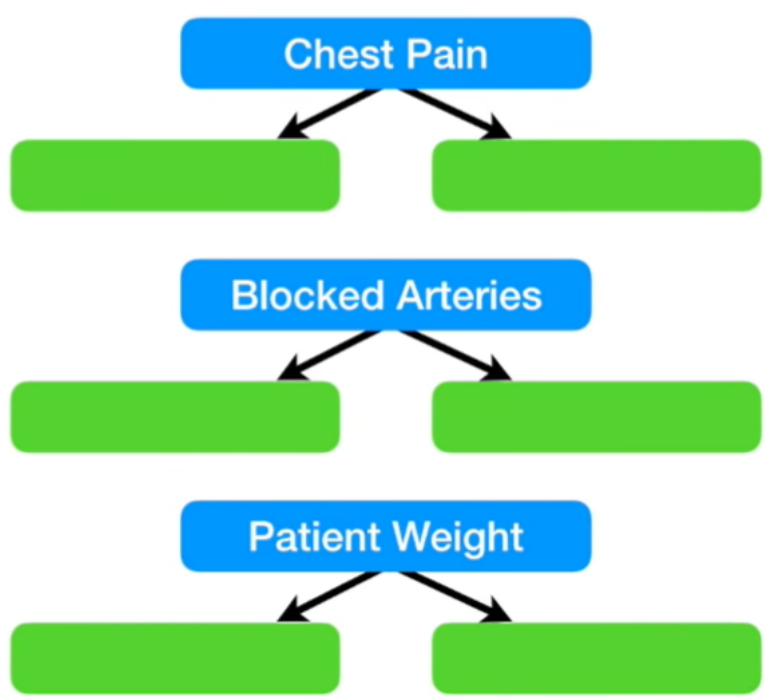
However, that doesn't mean the next stump will not emphasize the need to correctly classify these samples.

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	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.

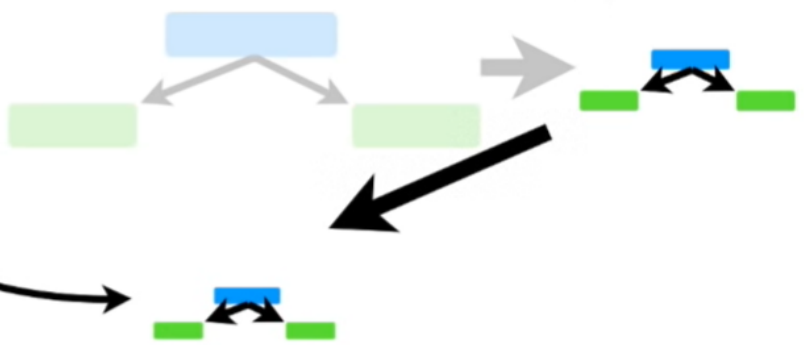


So that is how the errors that the first tree makes…

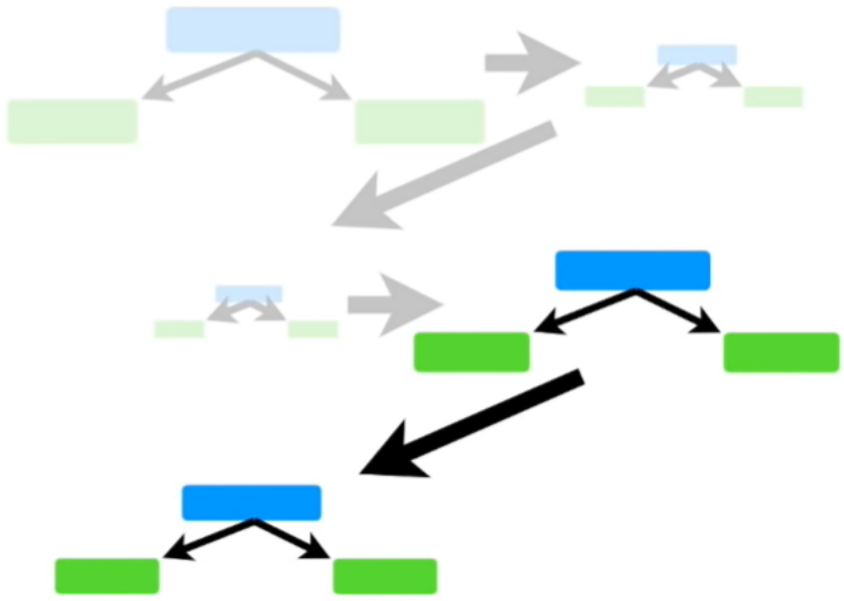
…influence how the second tree is made…

…and the errors that the second tree makes…

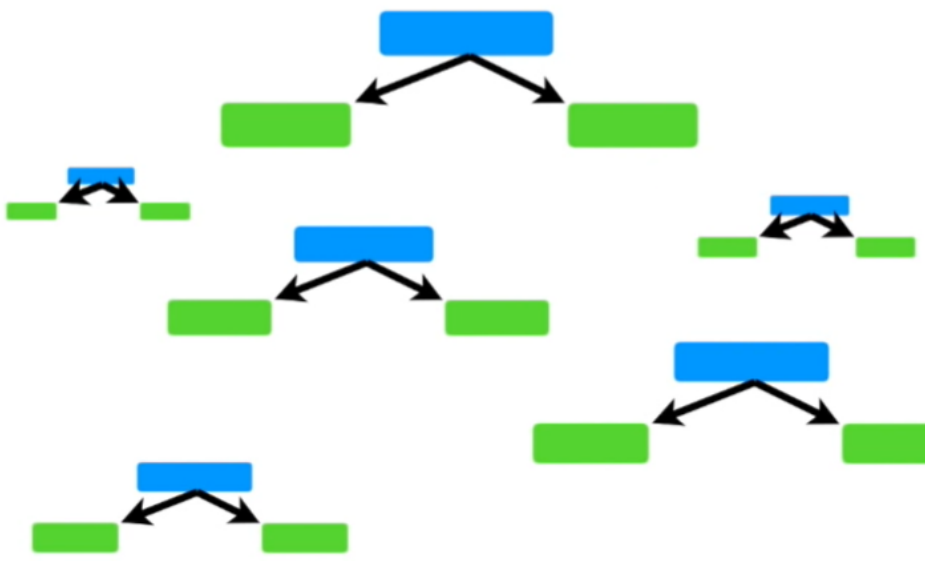
…influence how the third tree is made.



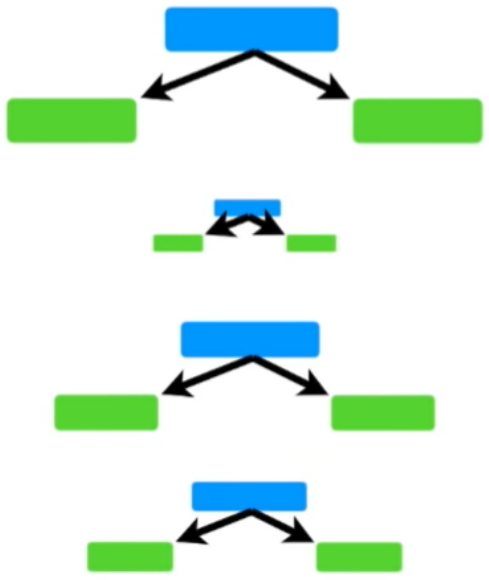
etc. etc. etc.



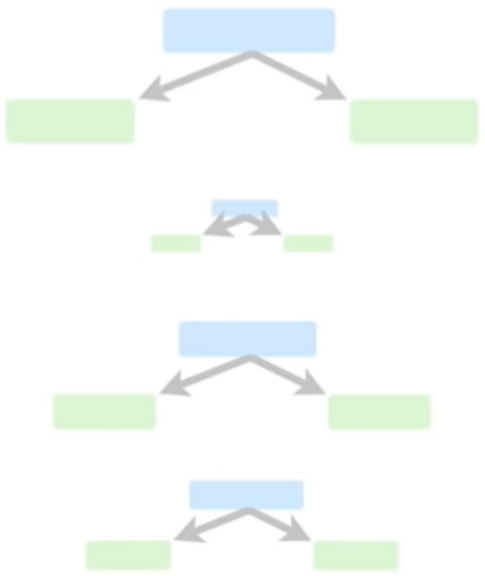
Now we need to talk about how a forest of stumps created by **AdaBoost** makes classifications…



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

Amount of Say

0.97

0.32

0.78

0.63

Does Not Have Heart Disease

…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

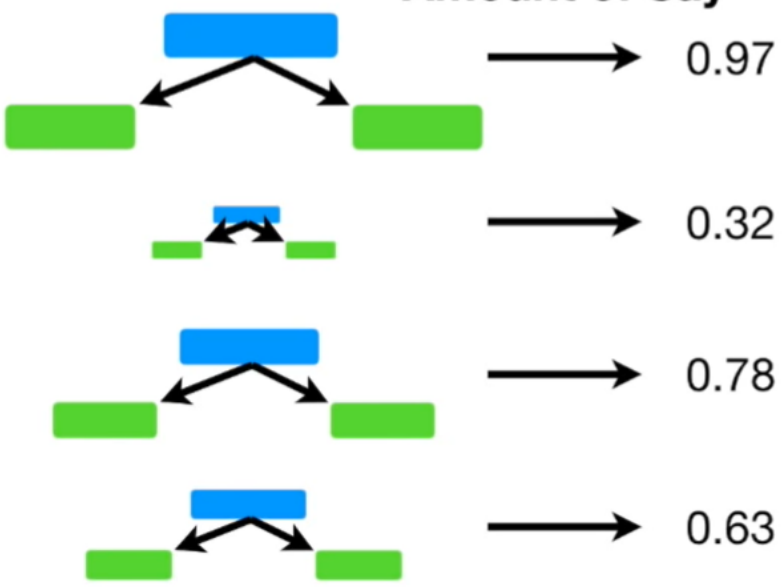
0.41

0.82

Now we add up the **Amounts of Say** for this group of stumps…

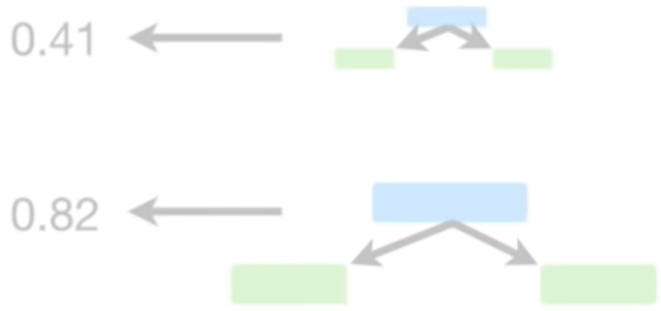
Has Heart Disease **Total = 2.7**

Amount of Say



Does Not Have Heart Disease

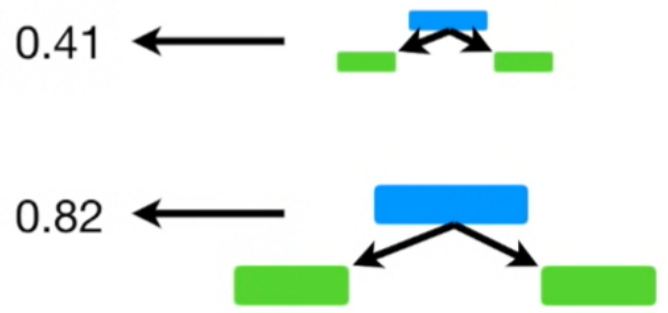
Amount of Say



…and for this group of stumps…

Total = 1.23 Does Not Have Heart Disease

Amount of Say



Ultimately, the patient is classified as **Has Heart Disease** because this is the larger sum.

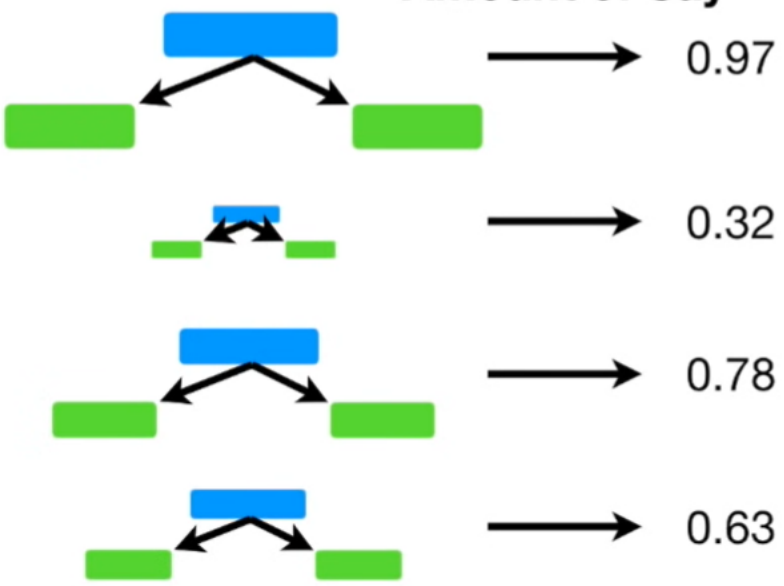
Has Heart Disease

Total = 2.7

Total = 1.23

Does Not Have Heart Disease

Amount of Say



Amount of Say