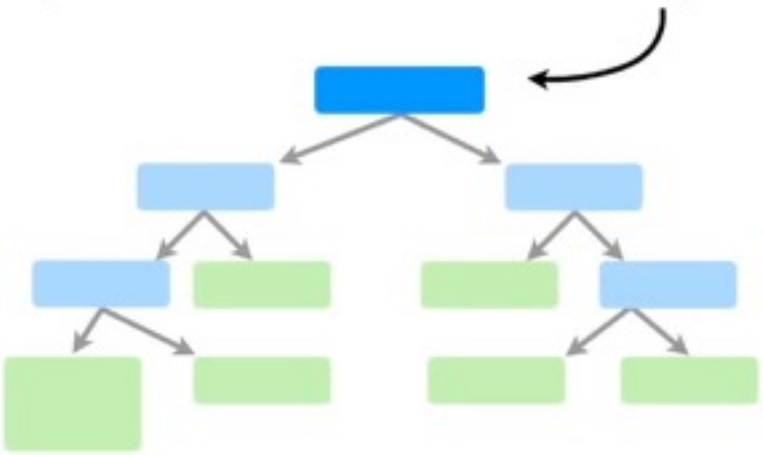
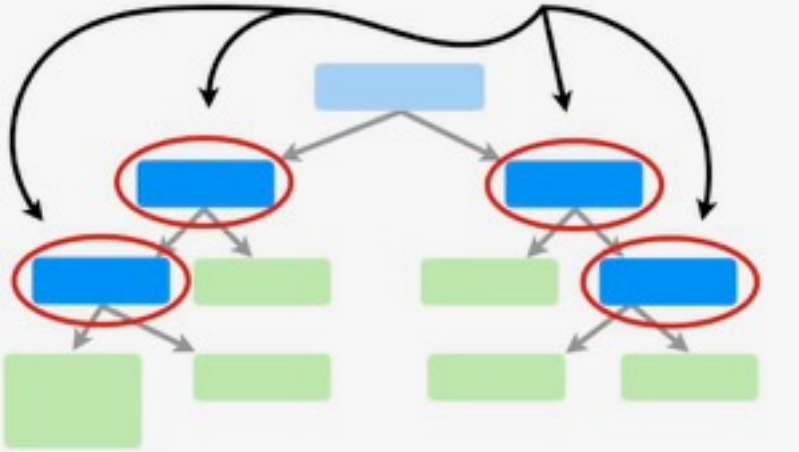
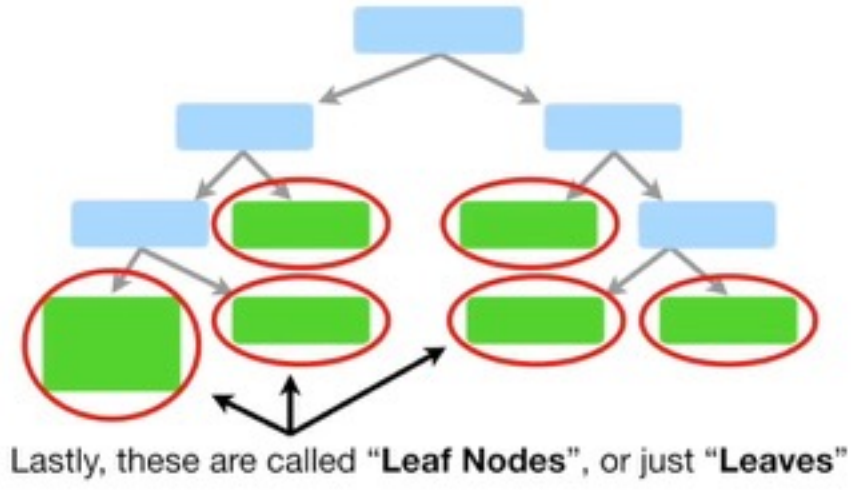
The very top of the tree is called the “**Root Node**” or just “**The Root**”



These are called “**Internal Nodes**”, or just “**Nodes**”.



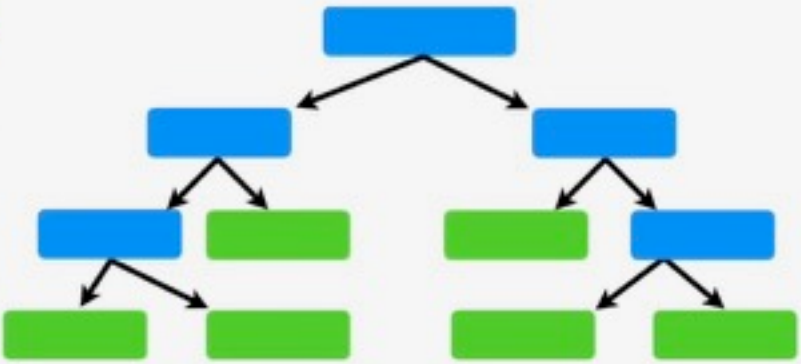


Lastly, these are called “**Leaf Nodes**”, or just “**Leaves**”

Now we are ready to talk about how to go from a raw table of data…

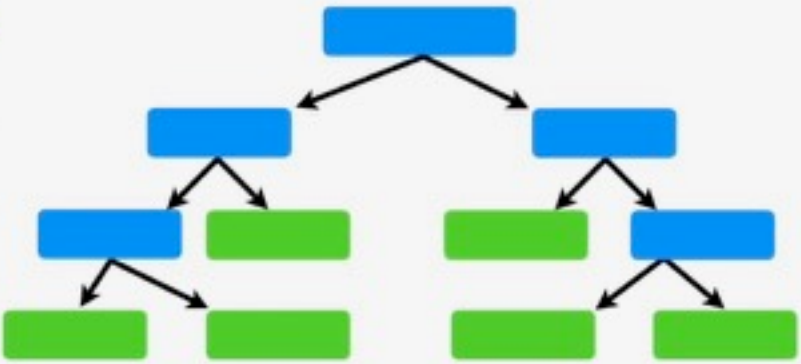
Chest Pain	Good Blood Circulation	Blocked Arteries	Heart Disease
No	No	No	No
Yes	Yes	Yes	Yes
Yes	Yes	No	No
Yes	No	???	Yes
etc…	etc…	etc…	etc…

…to a decision tree!!!



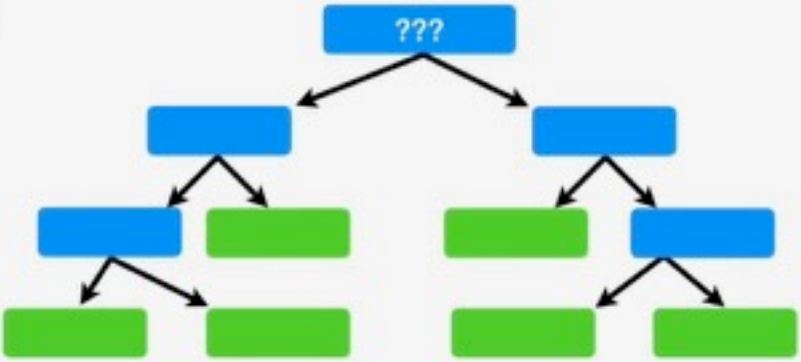
In this example, we want to create a tree that uses **chest pain**, **good blood circulation** and **blocked artery status** to predict…

Chest Pain	Good Blood Circulation	Blocked Arteries	Heart Disease
No	No	No	No
Yes	Yes	Yes	Yes
Yes	Yes	No	No
Yes	No	???	Yes
etc…	etc…	etc…	etc…



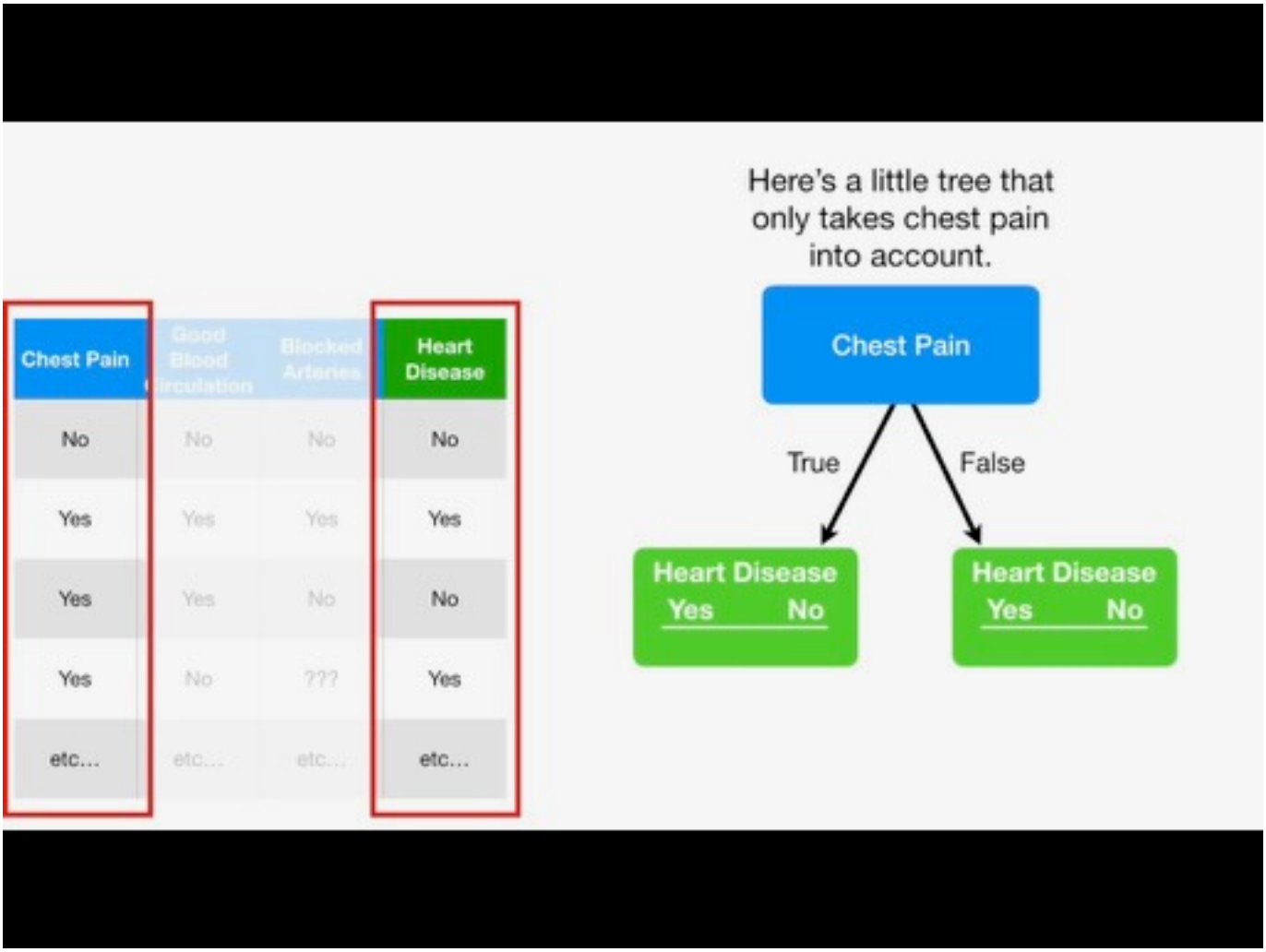
Chest Pain	Good Blood Circulation	Blocked Arteries	Heart Disease
No	No	No	No
Yes	Yes	Yes	Yes
Yes	Yes	No	No
Yes	No	???	Yes
etc…	etc…	etc…	etc…

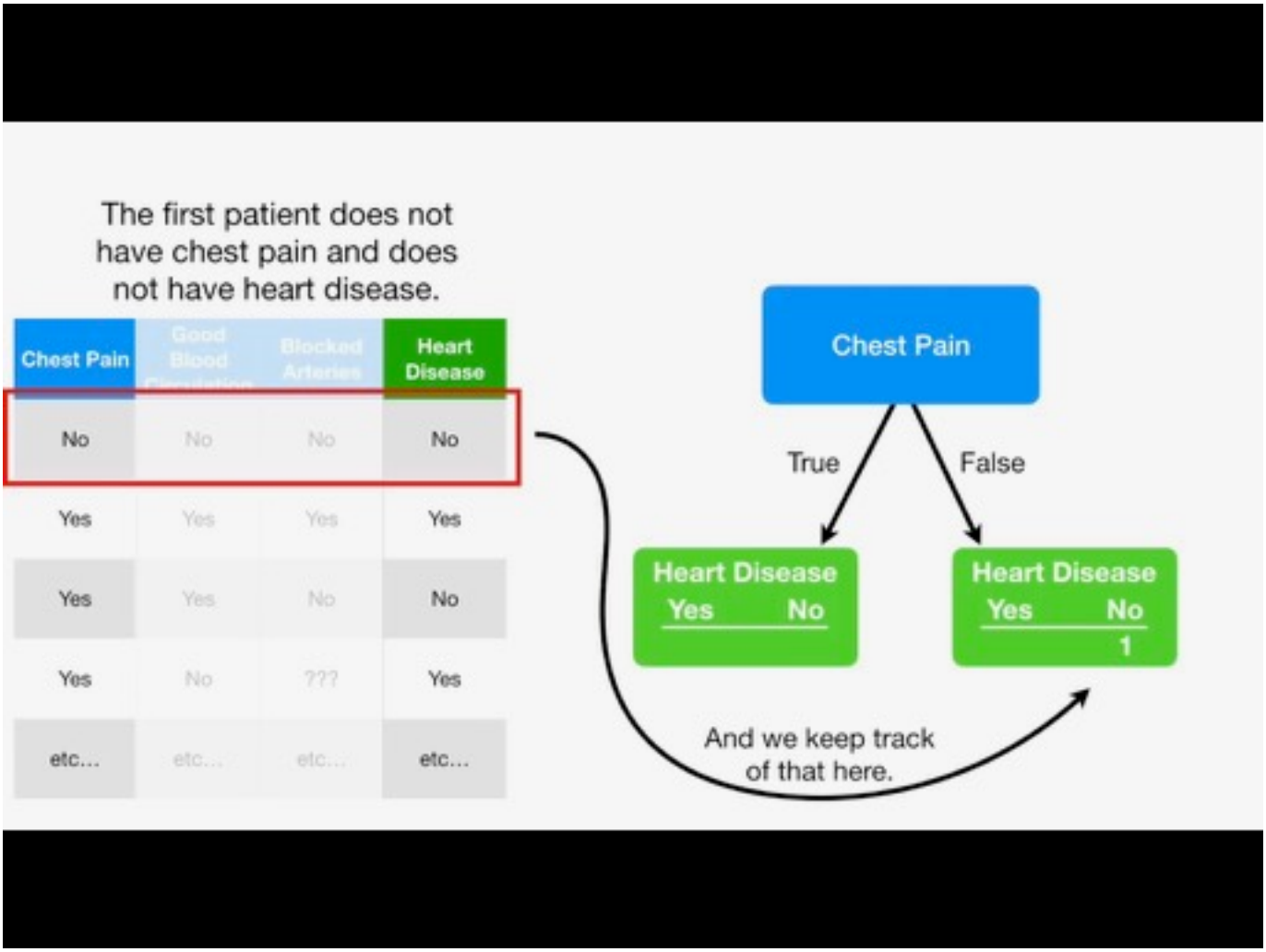
The first thing we want to know is whether **Chest Pain**, **Good Blood Circulation** or **Blocked Arteries** should be at the very top of our tree.

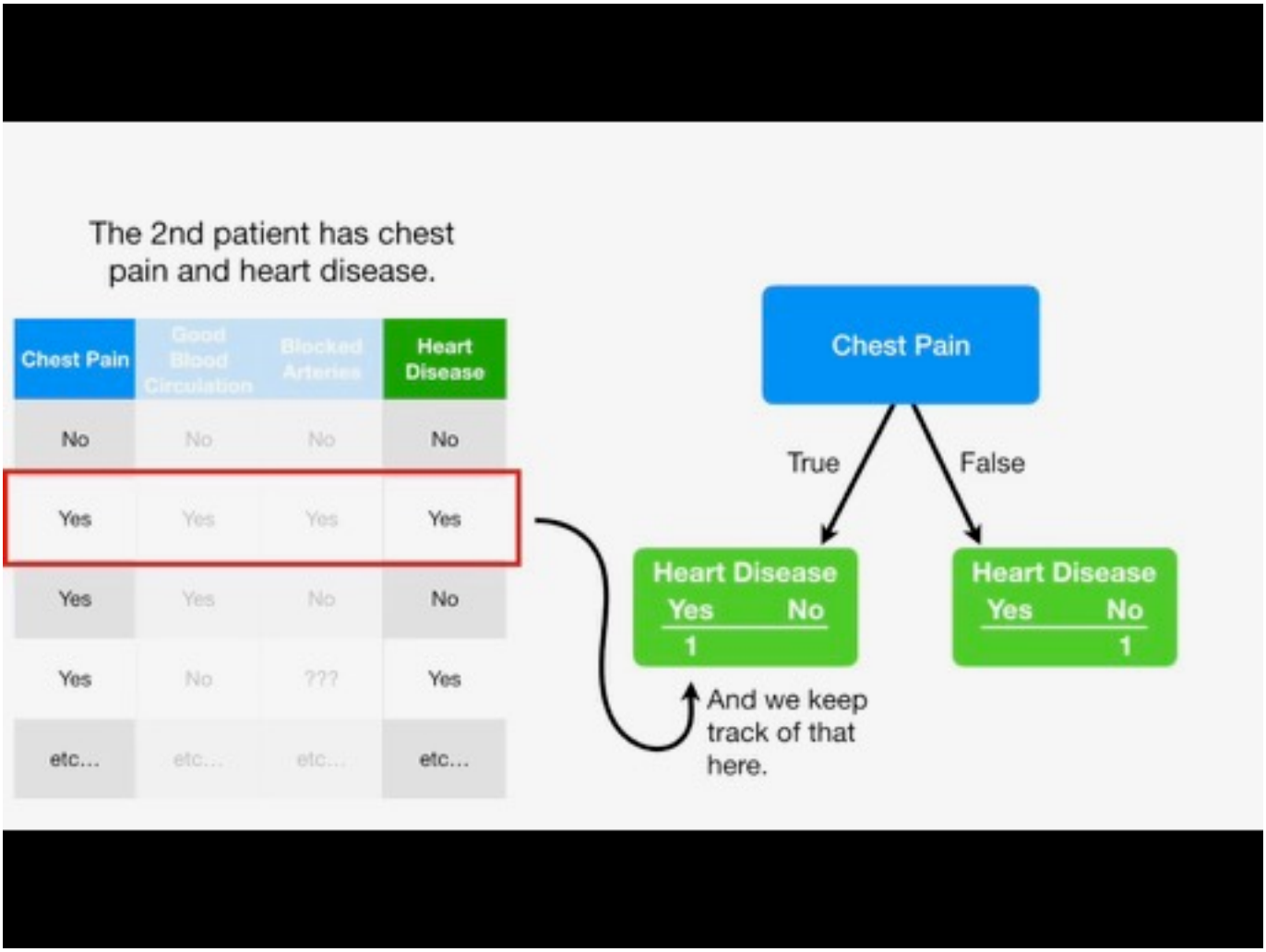


We start by looking at how well **Chest Pain** alone predicts heart disease…

Chest Pain	Good Blood Circulation	Blocked Arteries	Heart Disease
No	No	No	No
Yes	Yes	Yes	Yes
Yes	Yes	No	No
Yes	No	???	Yes
etc…	etc…	etc…	etc…

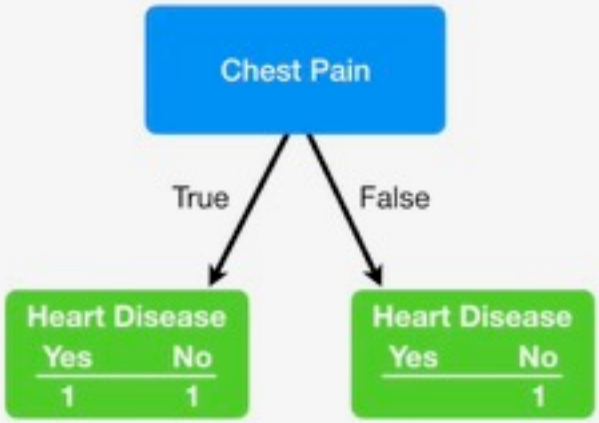

Chest Pain
Good Blood Circulation
Blocked Arteries
Heart Disease
No
No
No
No
Yes
Yes
Yes
Yes
Yes
Yes
No
No
Yes
No
???
Yes
etc...
etc...
etc...
etc...
Here's a little tree that only takes chest pain into account.
Chest Pain
True
False
Heart Disease
Yes
No
Heart Disease
Yes
No


The first patient does not have chest pain and does not have heart disease.
Chest Pain
Good Blood Circulation
Blocked Arteries
Heart Disease
No
No
No
No
Yes
Yes
Yes
Yes
Yes
Yes
No
No
Yes
No
???
Yes
etc...
etc...
etc...
etc...
Chest Pain
True
False
Heart Disease
Yes
No
Heart Disease
Yes
No
1
And we keep track of that here.

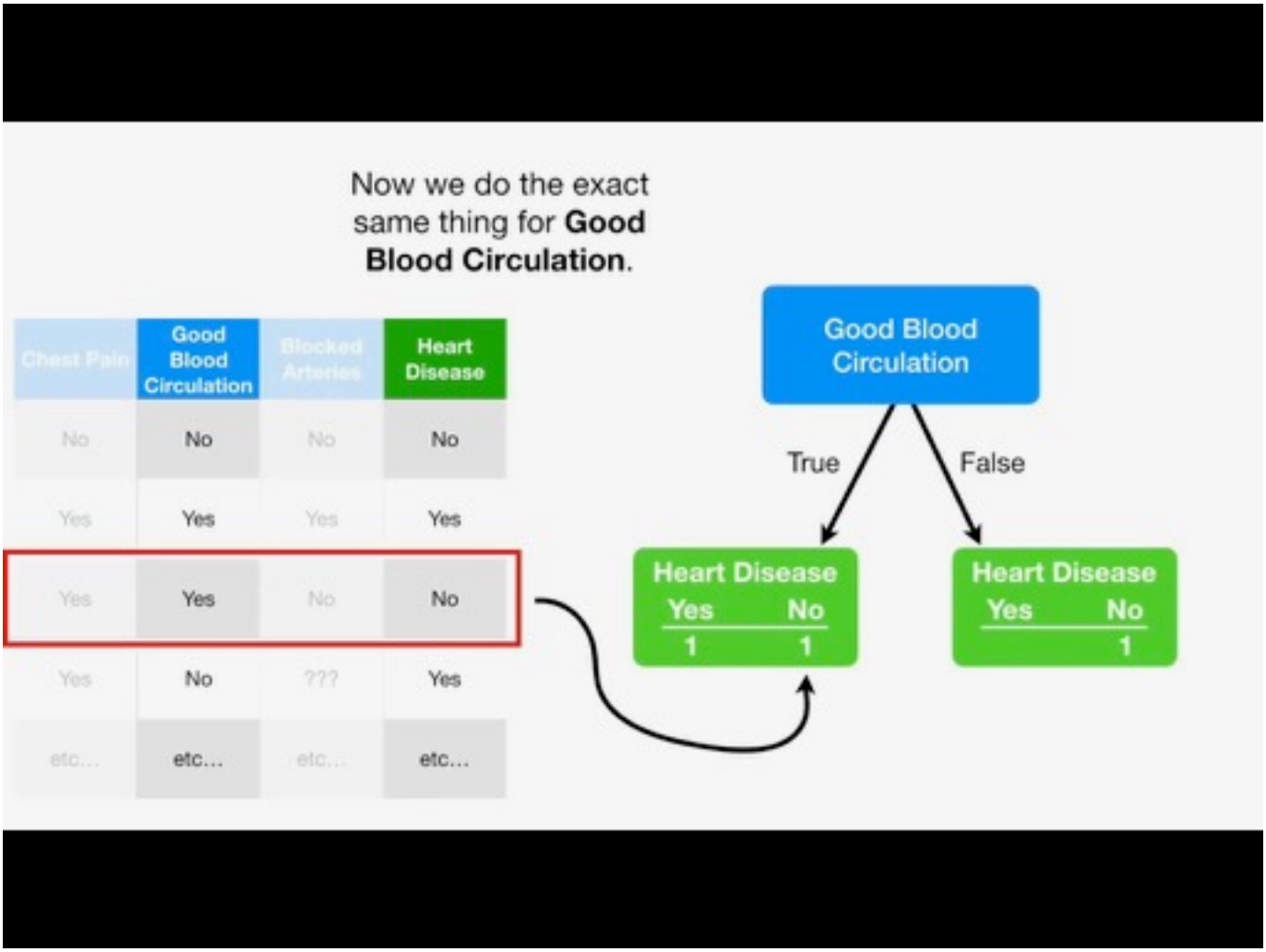

The 2nd patient has chest pain and heart disease.
Chest Pain
Good Blood Circulation
Blocked Arteries
Heart Disease
No No No No
Yes Yes Yes Yes
Yes Yes No No
Yes No ??? Yes
etc... etc... etc... etc...
Chest Pain
True
False
Heart Disease
Yes No
1
Heart Disease
Yes No
1
And we keep track of that here.

The 4th patient has chest pain and heart disease.

Chest Pain	Good Blood Circulation	Blocked Arteries	Heart Disease
No	No	No	No
Yes	Yes	Yes	Yes
Yes	Yes	No	No
Yes	No	???	Yes
etc...	etc...	etc...	etc...



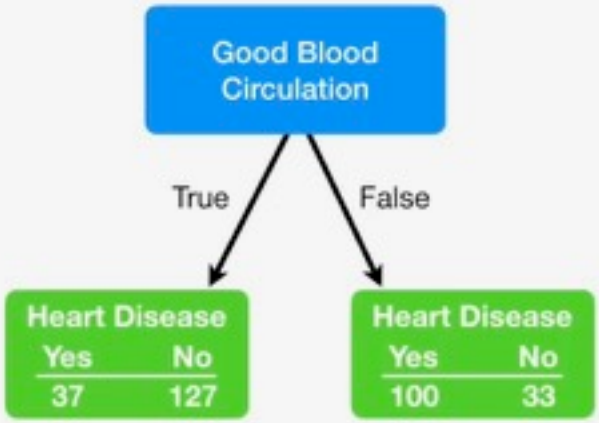

Chest Pain
Good Blood Circulation
Blocked Arteries
Heart Disease
No
No
No
No
Yes
Yes
Yes
Yes
Yes
Yes
No
No
Yes
No
???
Yes
etc...
etc...
etc...
etc...
Chest Pain
True
False
Heart Disease
Yes
No
105
39
Heart Disease
Yes
No
34
125
Ultimately, we look at chest pain and heart disease for all 303 patients in this study.


Now we do the exact same thing for **Good Blood Circulation**.
Chest Pain
Good Blood Circulation
Blocked Arteries
Heart Disease
No
No
No
No
Yes
Yes
Yes
Yes
Yes
Yes
No
No
Yes
No
???
Yes
etc…
etc…
etc…
etc…
Good Blood Circulation
True
False
Heart Disease
Yes
No
1
1
Heart Disease
Yes
No
1


Chest Pain
Good Blood Circulation
Blocked Arteries
Heart Disease
No
No
No
No
Yes
Yes
Yes
Yes
Yes
Yes
No
No
Yes
No
???
Yes
etc...
etc...
etc...
etc...
Chest Pain
True
False
Heart Disease
Yes
No
105
39
Heart Disease
Yes
No
34
125
Ultimately, we look at chest pain and heart disease for all 303 patients in this study.

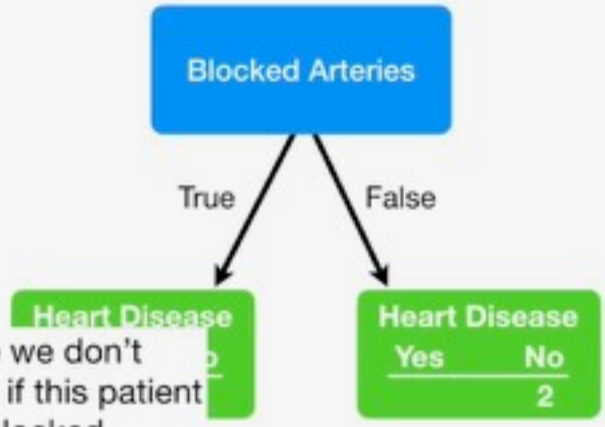
Now we do the exact same thing for **Good Blood Circulation**.

Chest Pain	Good Blood Circulation	Blocked Arteries	Heart Disease
No	No	No	No
Yes	Yes	Yes	Yes
Yes	Yes	No	No
Yes	No	???	Yes
etc...	etc...	etc...	etc...

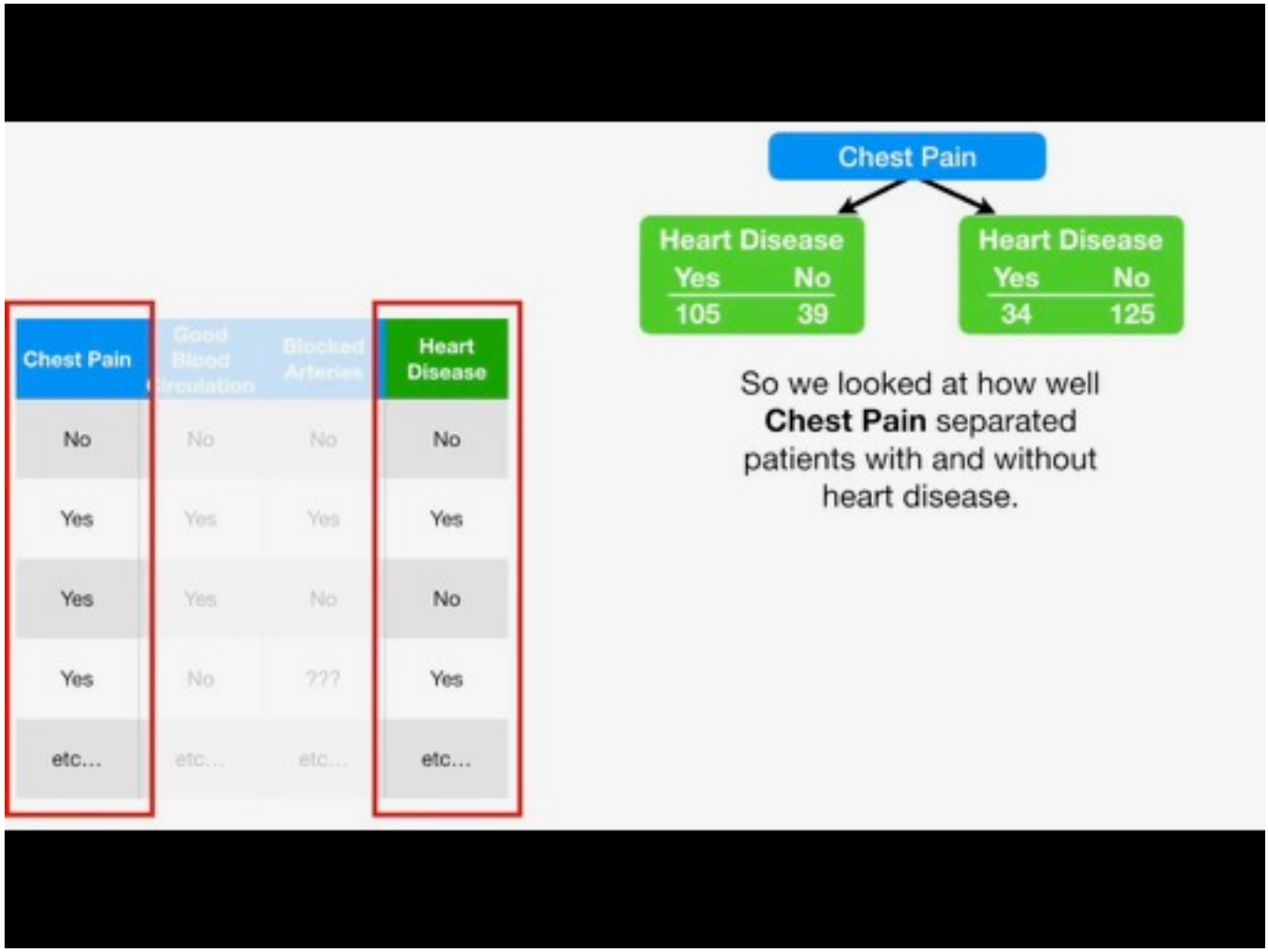


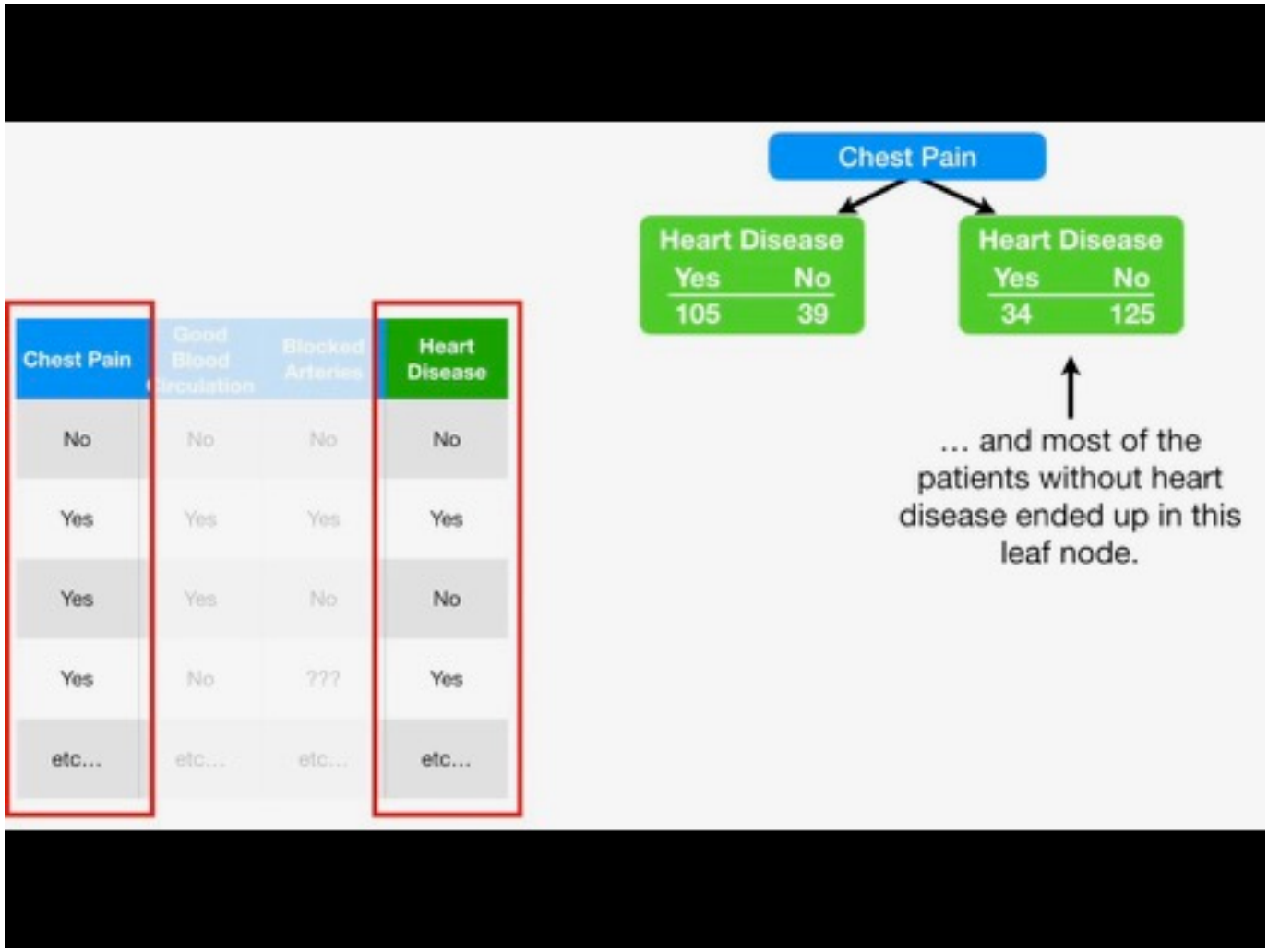
Chest Pain	Good Blood Circulation	Blocked Arteries	Heart Disease
No	No	No	No
Yes	Yes	Yes	Yes
Yes	Yes	No	No
Yes	No	???	Yes
etc...	etc...	etc...	etc...

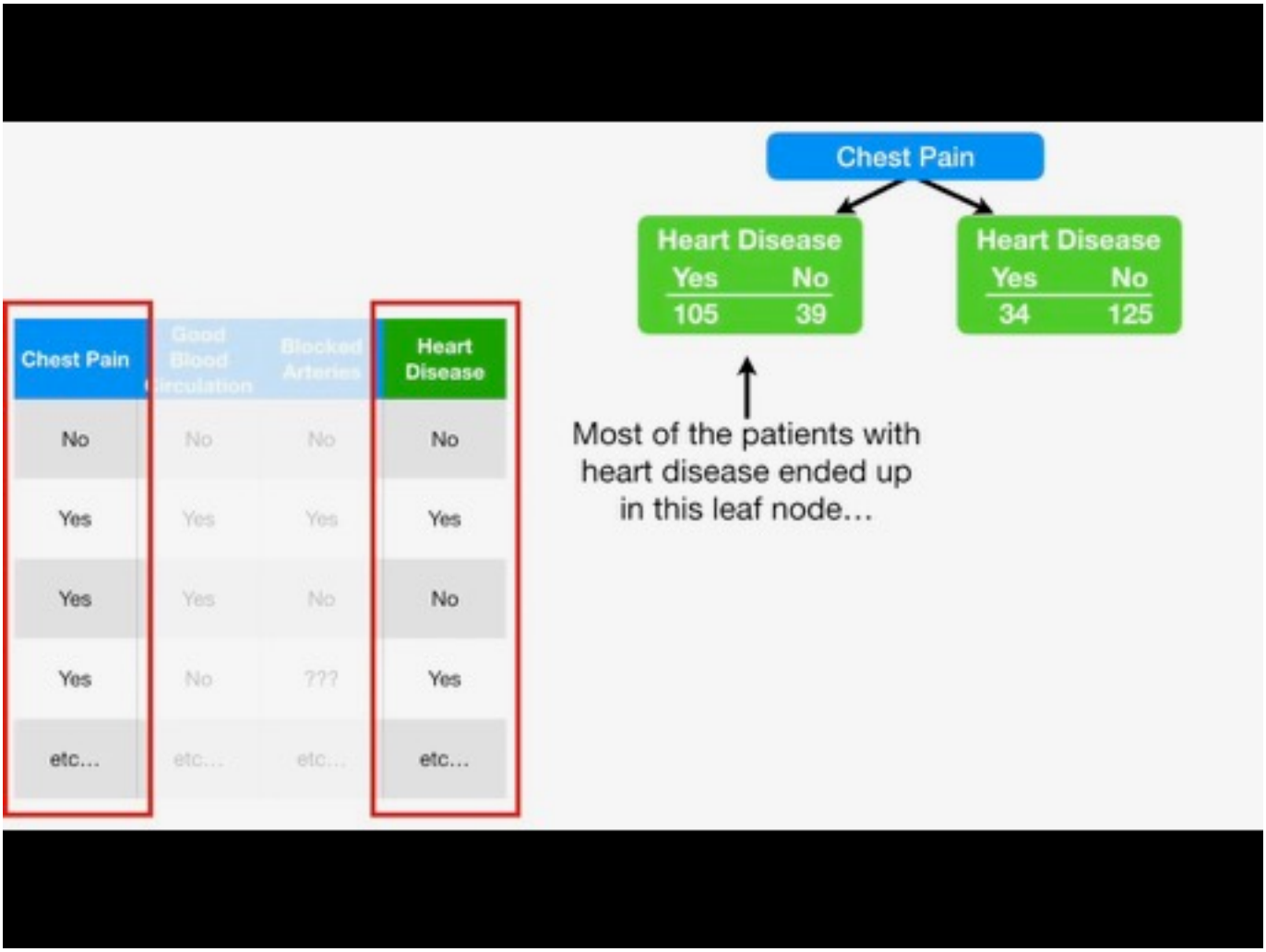
Since we don't know if this patient had blocked arteries or not, we'll skip it.

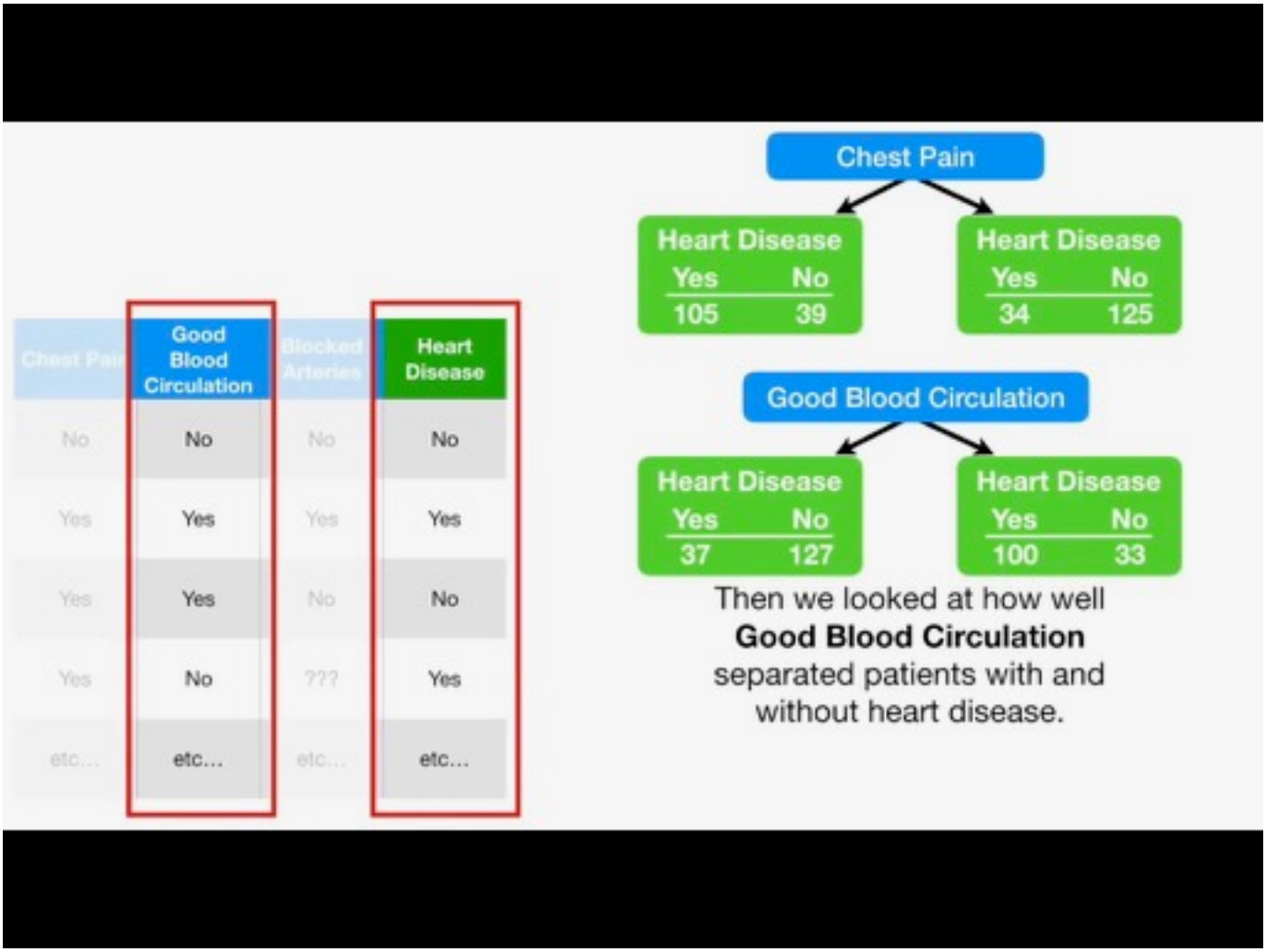


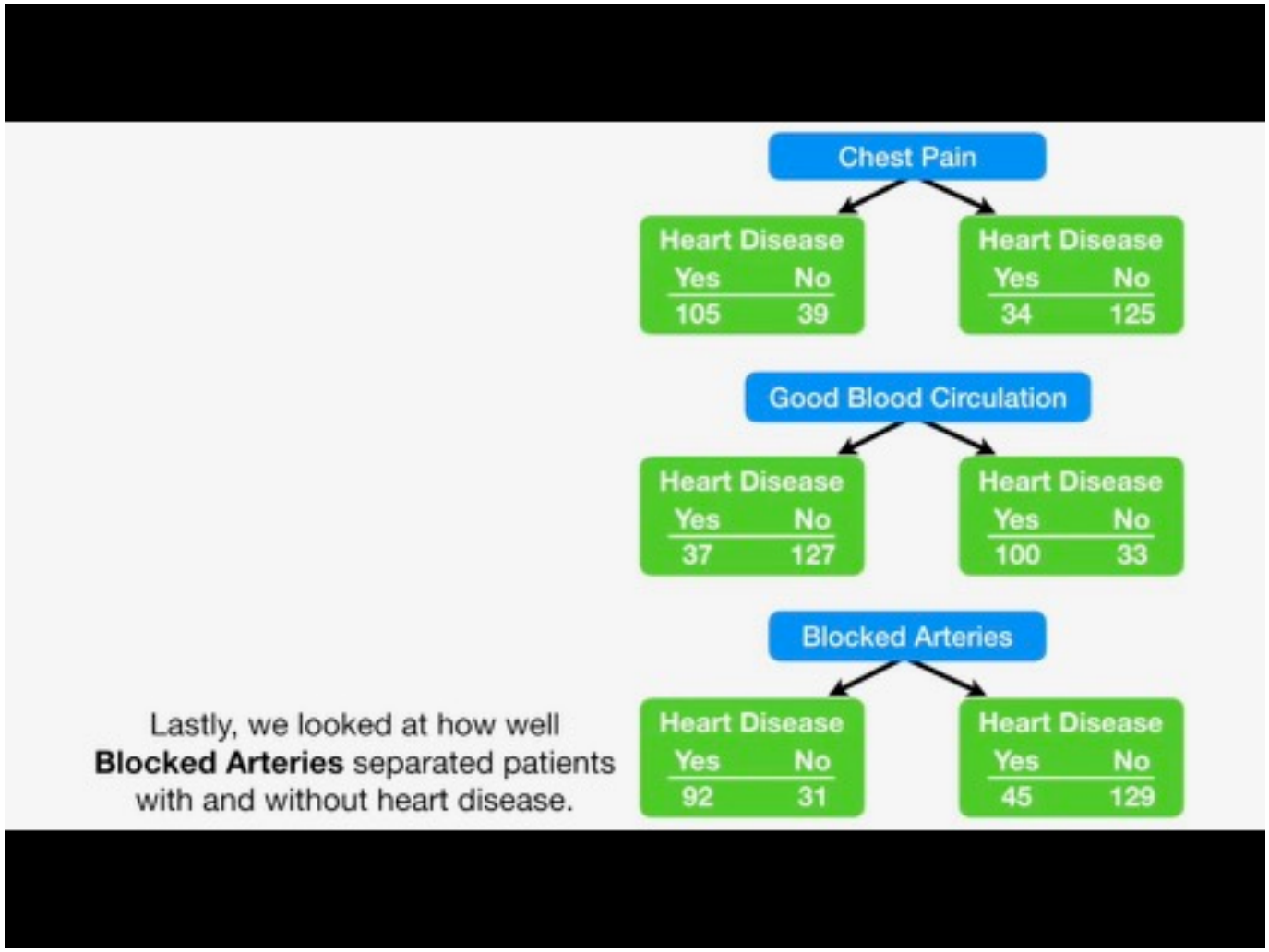

Chest Pain
Good Blood Circulation
Blocked Arteries
Heart Disease
No
No
No
No
Yes
Yes
Yes
Yes
Yes
Yes
No
No
Yes
No
???
Yes
etc...
etc...
etc...
etc...
Blocked Arteries
True
False
Heart Disease
Yes
No
92
31
Heart Disease
Yes
No
45
129

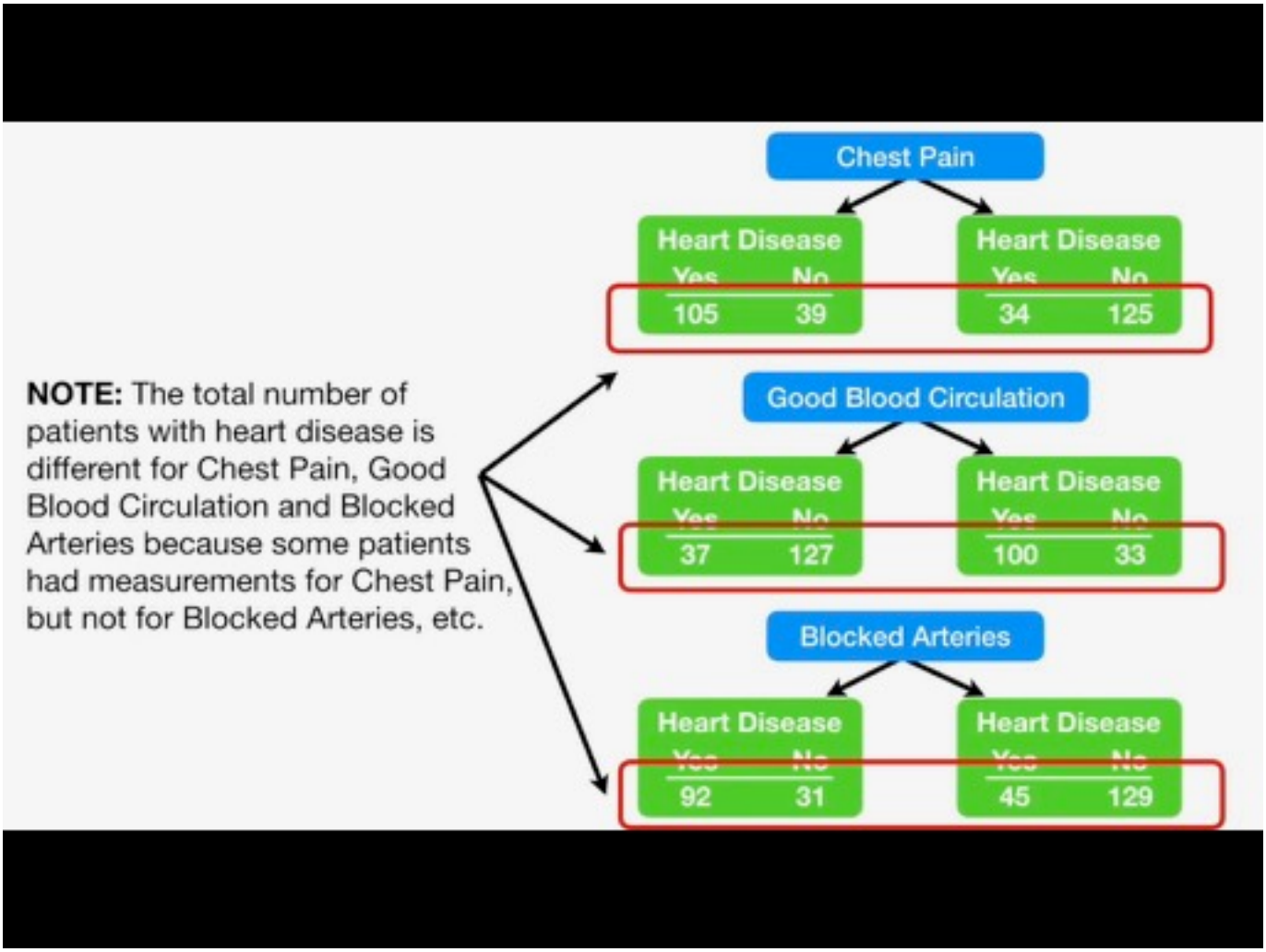

Chest Pain
Heart Disease
Yes No
105 39
Heart Disease
Yes No
34 125
So we looked at how well **Chest Pain** separated patients with and without heart disease.
Chest Pain
Good Blood Circulation
Blocked Arteries
Heart Disease
No No No No
Yes Yes Yes Yes
Yes Yes No No
Yes No ??? Yes
etc... etc... etc... etc...

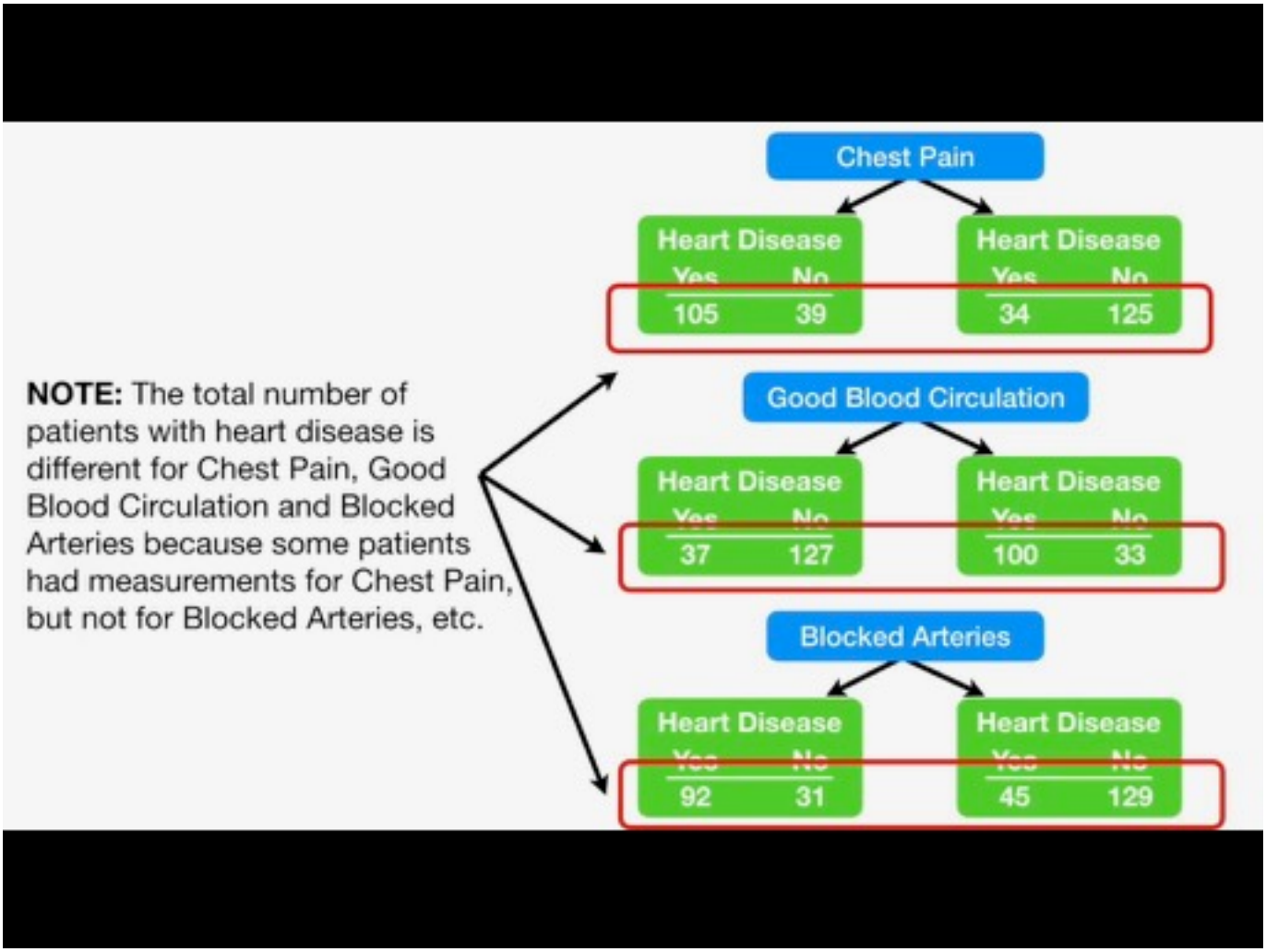

Chest Pain
Heart Disease
Yes No
105 39
Heart Disease
Yes No
34 125
… and most of the patients without heart disease ended up in this leaf node.
Chest Pain
Good Blood Circulation
Blocked Arteries
Heart Disease
No No No No
Yes Yes Yes Yes
Yes Yes No No
Yes No ??? Yes
etc… etc… etc… etc…


Chest Pain
Heart Disease
Yes No
105 39
Heart Disease
Yes No
34 125
Most of the patients with heart disease ended up in this leaf node…
Chest Pain
Good Blood Circulation
Blocked Arteries
Heart Disease
No No No No
Yes Yes Yes Yes
Yes Yes No No
Yes No ??? Yes
etc… etc… etc… etc…

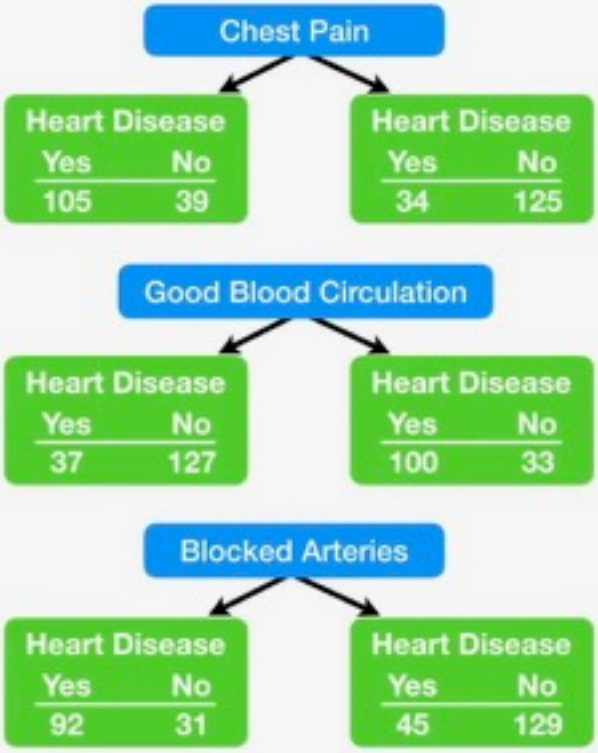

Chest Pain
Heart Disease
Yes No
105 39
Heart Disease
Yes No
34 125
Good Blood Circulation
Heart Disease
Yes No
37 127
Heart Disease
Yes No
100 33
Then we looked at how well
Good Blood Circulation
separated patients with and
without heart disease.
Chest Pain
Good Blood Circulation
Blocked Arteries
Heart Disease
No No No No
Yes Yes Yes Yes
Yes Yes No No
Yes No ??? Yes
etc... etc... etc... etc...


Chest Pain
Heart Disease
Yes No
105 39
Heart Disease
Yes No
34 125
Good Blood Circulation
Heart Disease
Yes No
37 127
Heart Disease
Yes No
100 33
Blocked Arteries
Heart Disease
Yes No
92 31
Heart Disease
Yes No
45 129
Lastly, we looked at how well
Blocked Arteries separated patients
with and without heart disease.


Chest Pain
Heart Disease
Yes No
105 39
Heart Disease
Yes No
34 125
NOTE: The total number of patients with heart disease is different for Chest Pain, Good Blood Circulation and Blocked Arteries because some patients had measurements for Chest Pain, but not for Blocked Arteries, etc.
Good Blood Circulation
Heart Disease
Yes No
37 127
Heart Disease
Yes No
100 33
Blocked Arteries
Heart Disease
Yes No
92 31
Heart Disease
Yes No
45 129

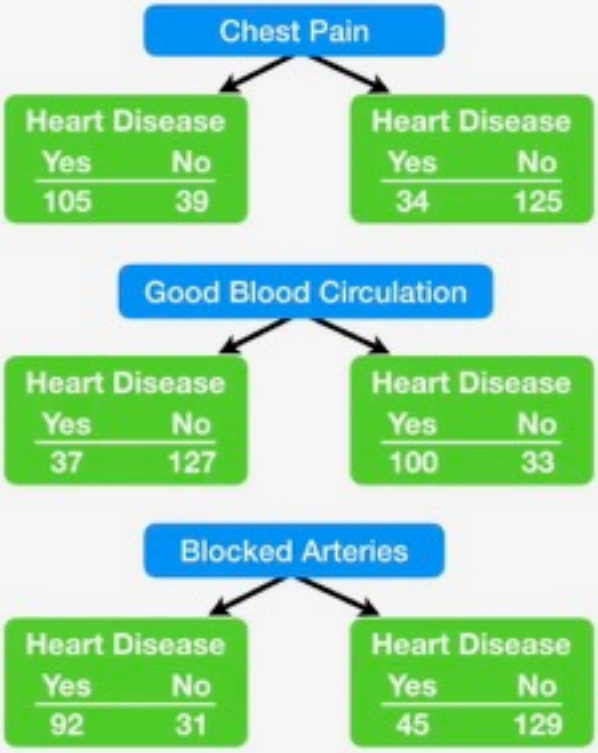

Chest Pain
Heart Disease
Yes No
105 39
Heart Disease
Yes No
34 125
NOTE: The total number of patients with heart disease is different for Chest Pain, Good Blood Circulation and Blocked Arteries because some patients had measurements for Chest Pain, but not for Blocked Arteries, etc.
Good Blood Circulation
Heart Disease
Yes No
37 127
Heart Disease
Yes No
100 33
Blocked Arteries
Heart Disease
Yes No
92 31
Heart Disease
Yes No
45 129

Because none of the leaf nodes are 100% “YES Heart Disease” or 100% “NO Heart Disease”, they are all considered “**impure**”.



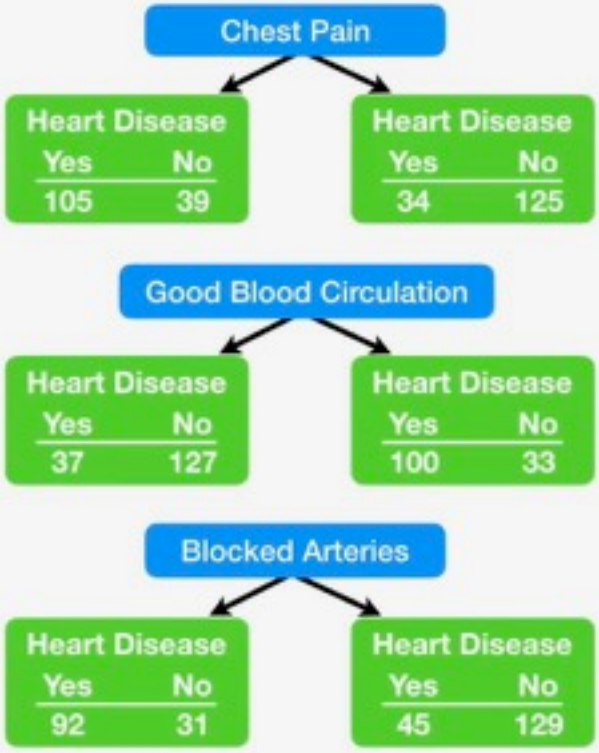
Because none of the leaf nodes are 100% “YES Heart Disease” or 100% “NO Heart Disease”, they are all considered “**impure**”.

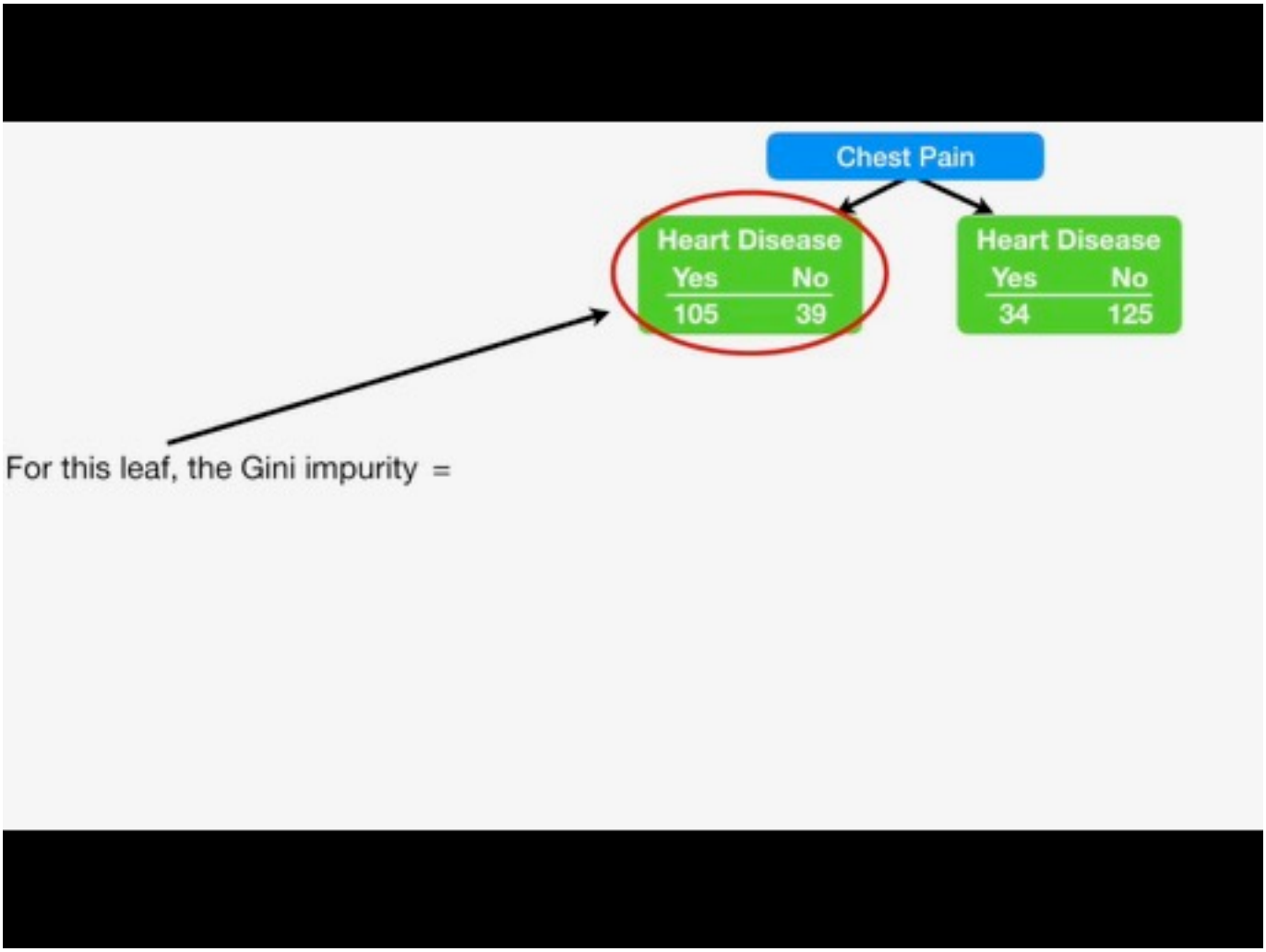
To determine which separation is best, we need a way to measure and compare “**impurity**”.

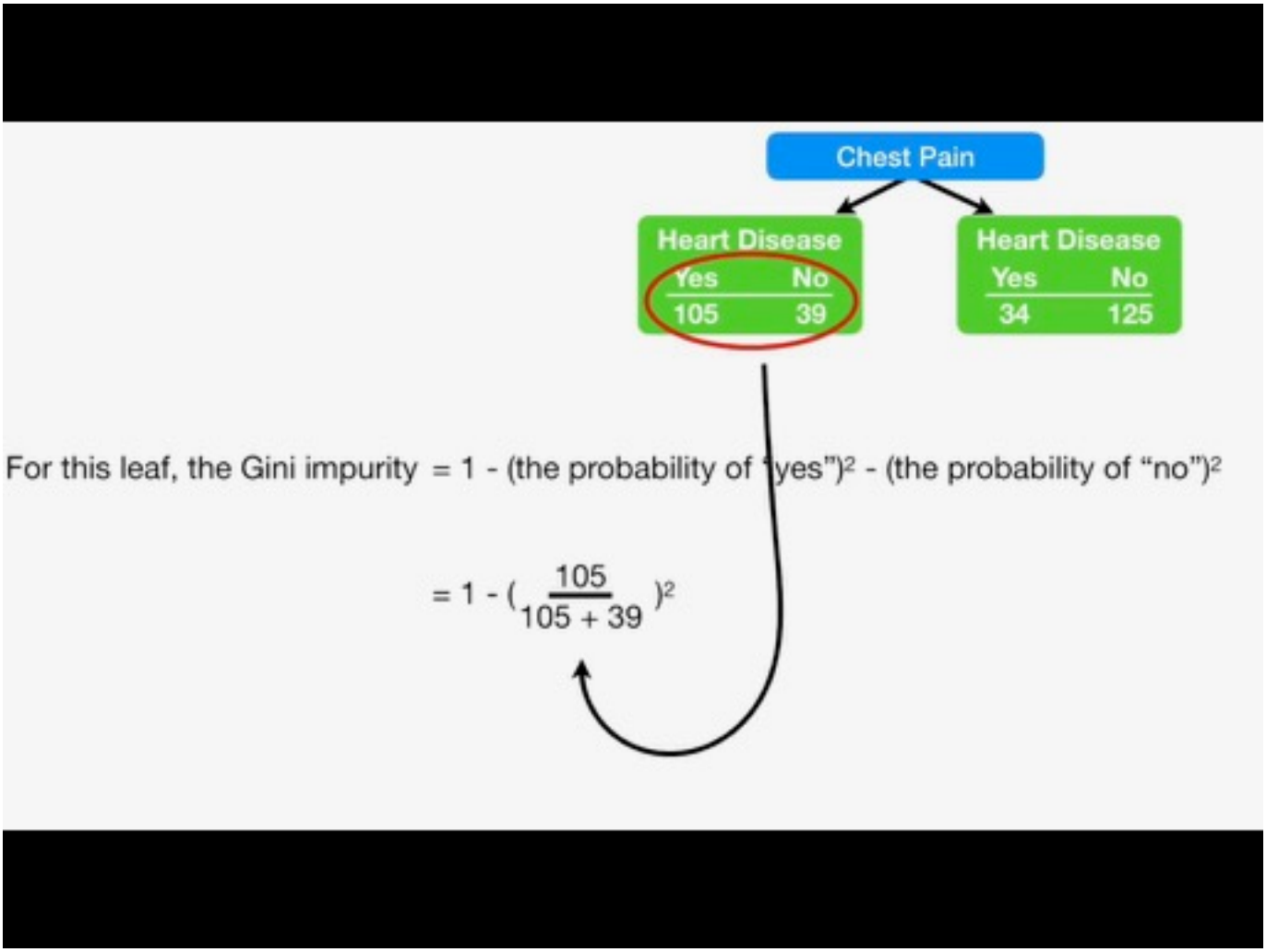


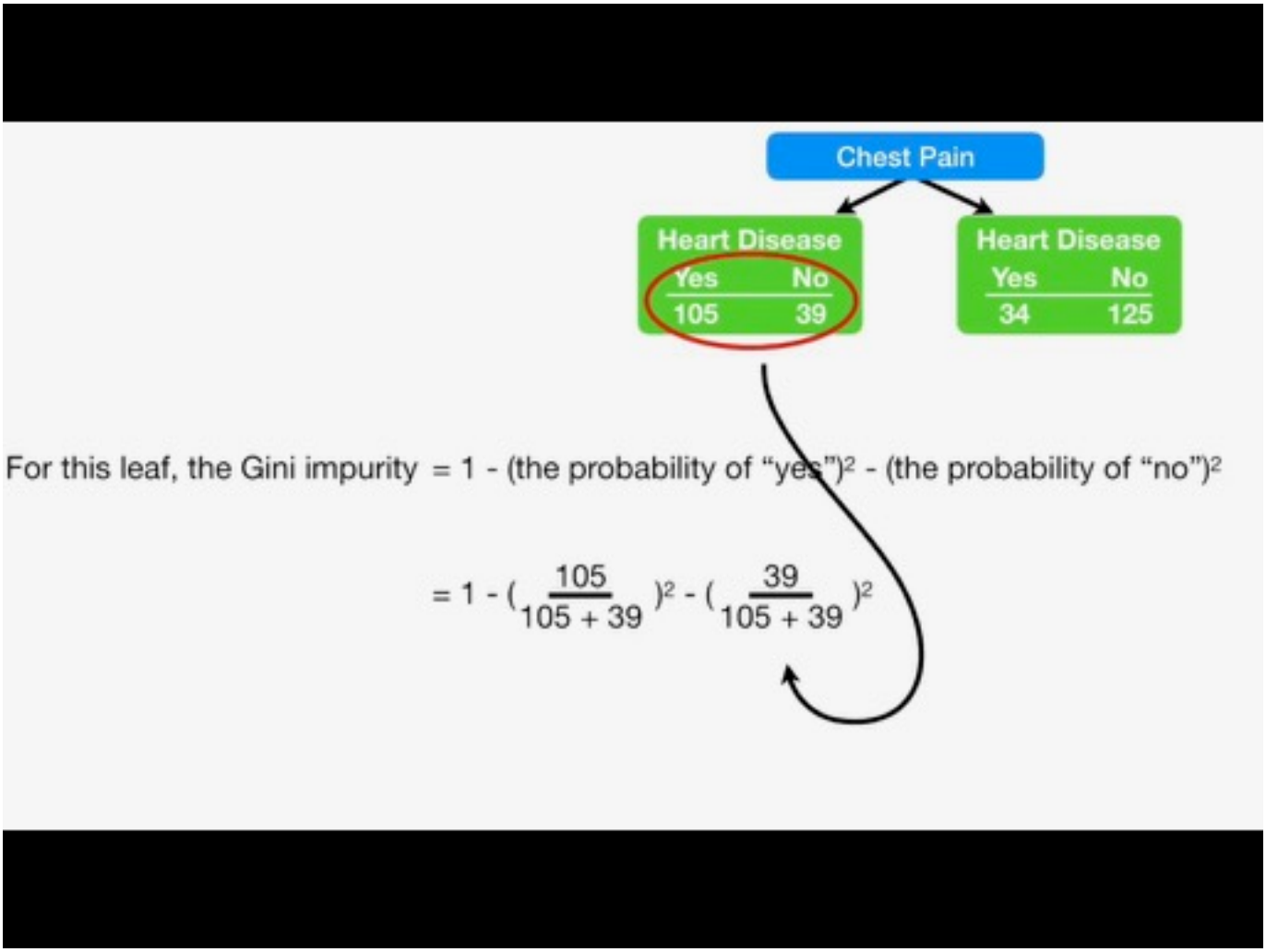
There are a bunch of ways to measure impurity, but I'm just going to focus on a very popular one called "**Gini**".

To be honest, I don't know why it's called Gini. I looked around on the internet and couldn't find anything. However, if you know, please put it in the comments below. I would love to know!!!




Chest Pain
Heart Disease
Yes No
105 39
Heart Disease
Yes No
34 125
For this leaf, the Gini impurity =

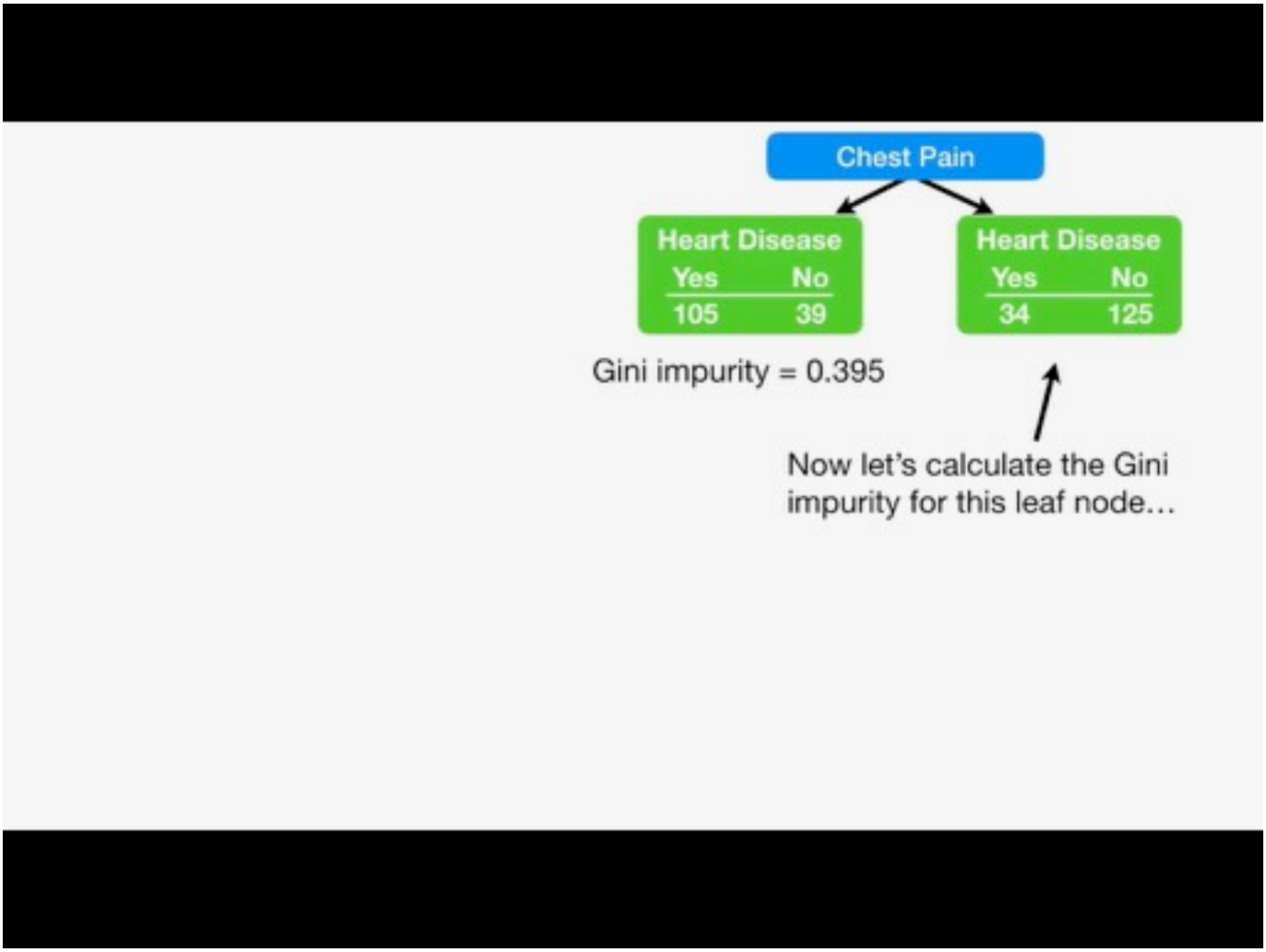

Chest Pain
Heart Disease
Yes
No
105
39
Heart Disease
Yes
No
34
125
For this leaf, the Gini impurity = 1 - (the probability of "yes")² - (the probability of "no")²
= 1 - (105 / 105 + 39)²

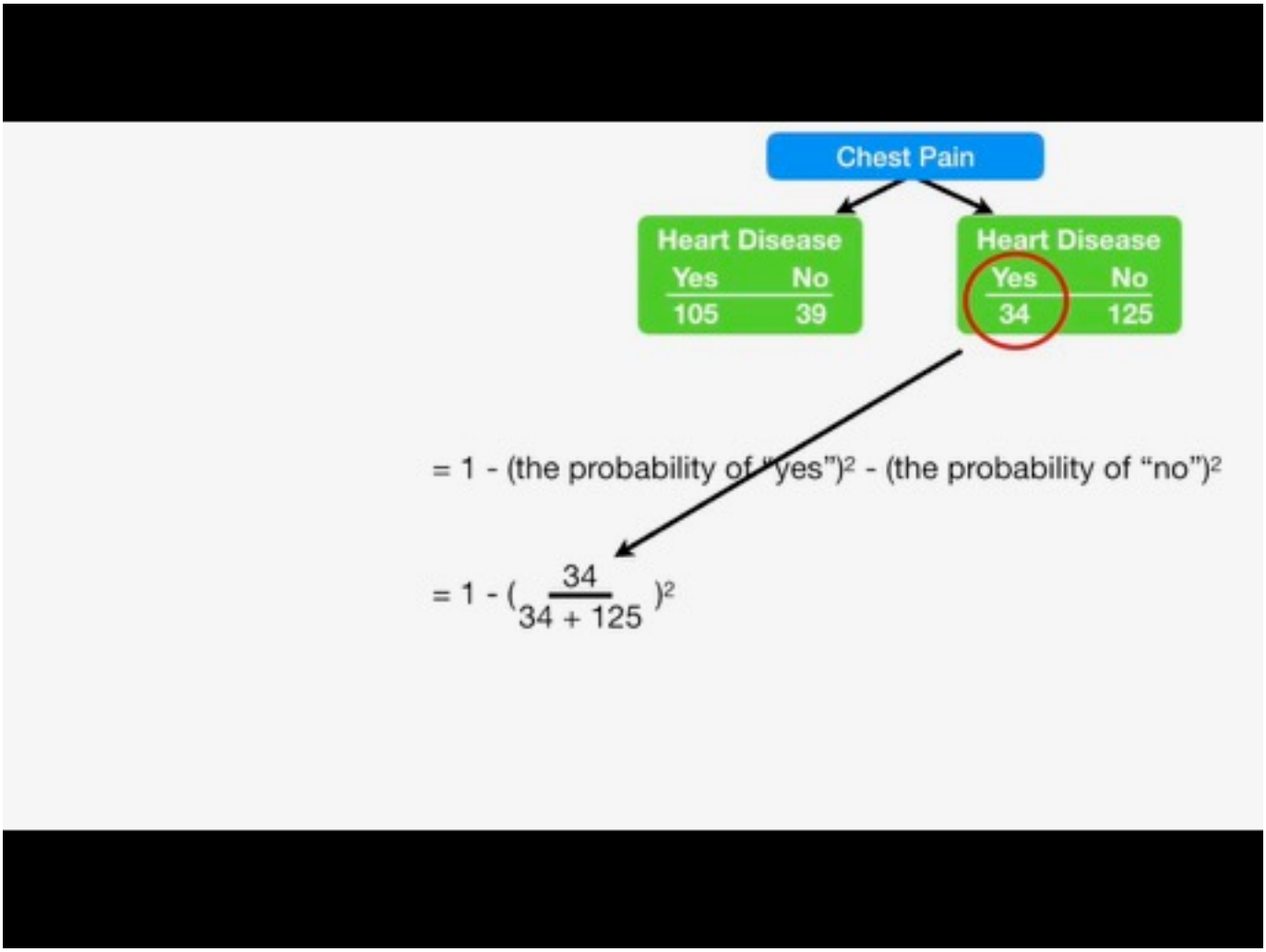

Chest Pain
Heart Disease
Yes
No
105
39
Heart Disease
Yes
No
34
125
For this leaf, the Gini impurity = 1 - (the probability of "yes")² - (the probability of "no")²
= 1 - (105 / 105 + 39)² - (39 / 105 + 39)²

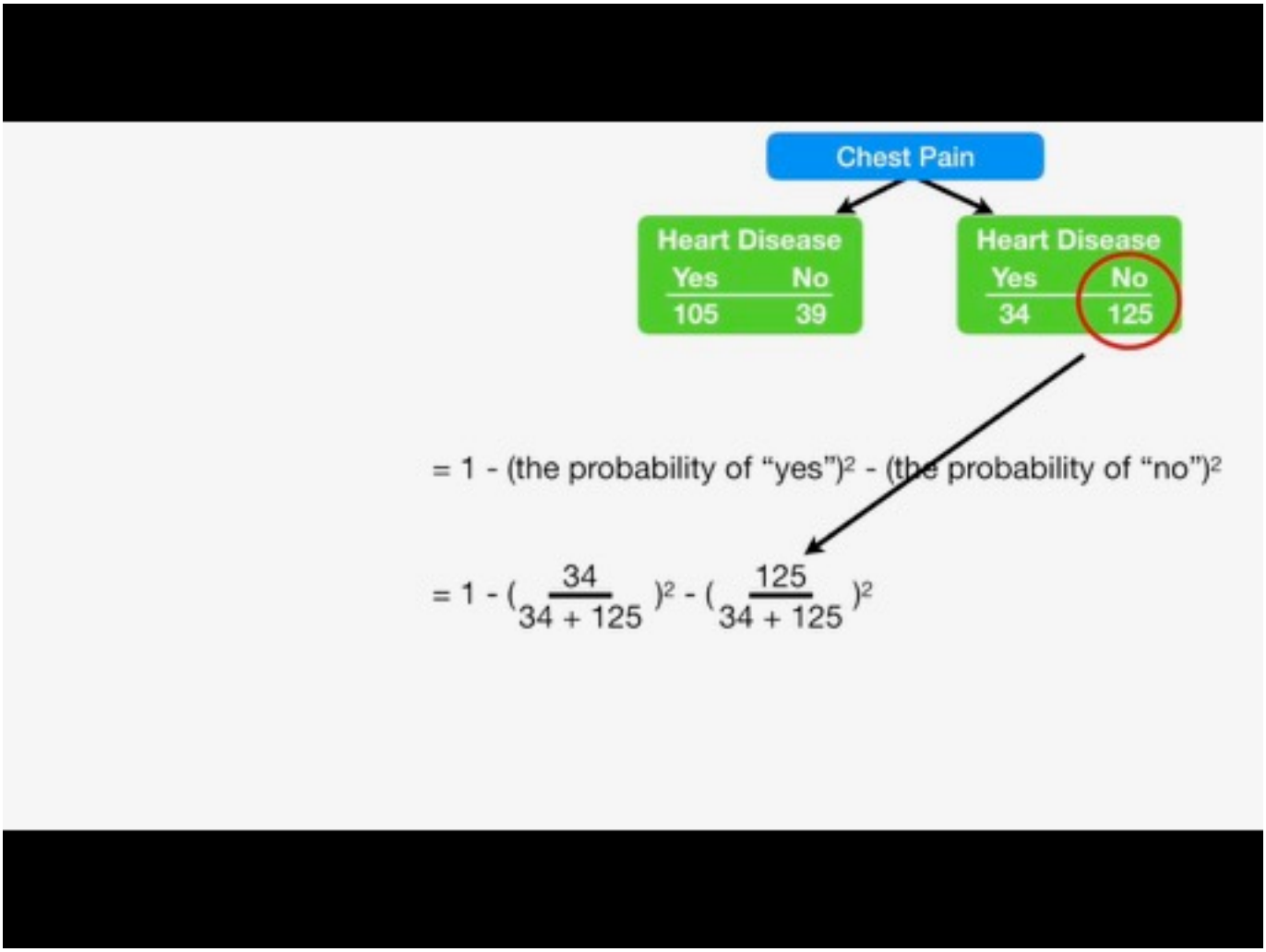
For this leaf, the Gini impurity = 1 - (the probability of "yes")^2 - (the probability of "no")^2

$$= 1 - \left(\frac{105}{105 + 39}\right)^2 - \left(\frac{39}{105 + 39}\right)^2$$

$$= 0.395$$


Chest Pain
Heart Disease
Yes
No
105
39
Heart Disease
Yes
No
34
125
Gini impurity = 0.395
Now let's calculate the Gini impurity for this leaf node...

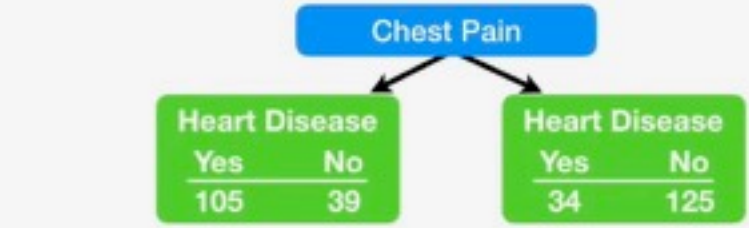

Chest Pain
Heart Disease
Yes
No
105
39
Heart Disease
Yes
No
34
125
= 1 - (the probability of "yes")² - (the probability of "no")²
= 1 - (34 / 34 + 125)²


Chest Pain
Heart Disease
Yes No
105 39
Heart Disease
Yes No
34 125
= 1 - (the probability of "yes")² - (the probability of "no")²
= 1 - (34 / 34 + 125)² - (125 / 34 + 125)²

$$= 1 - (\text{the probability of “yes”})^2 - (\text{the probability of “no”})^2$$

$$= 1 - \left(\frac{34}{34 + 125}\right)^2 - \left(\frac{125}{34 + 125}\right)^2$$

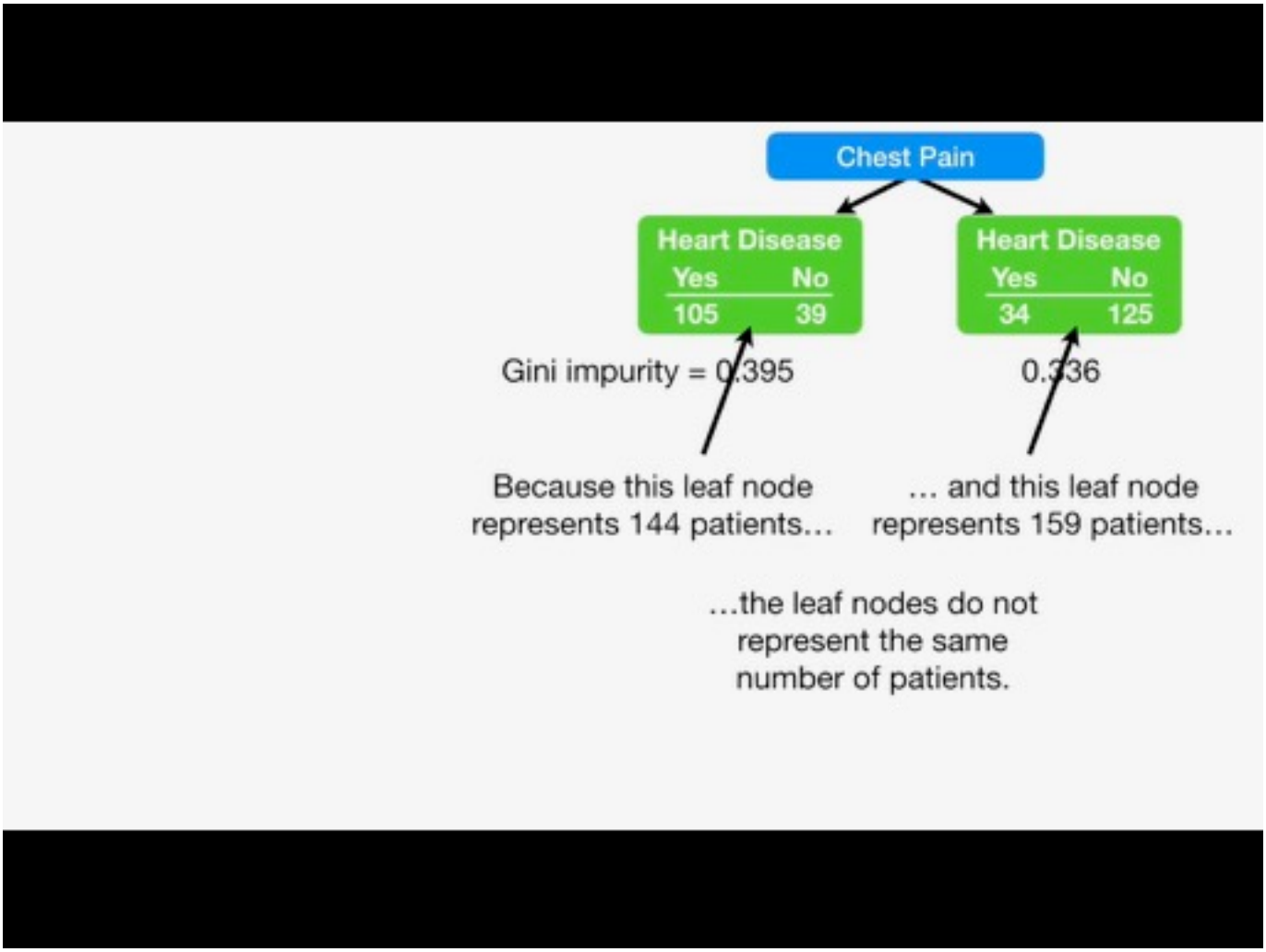
$$= 0.336$$

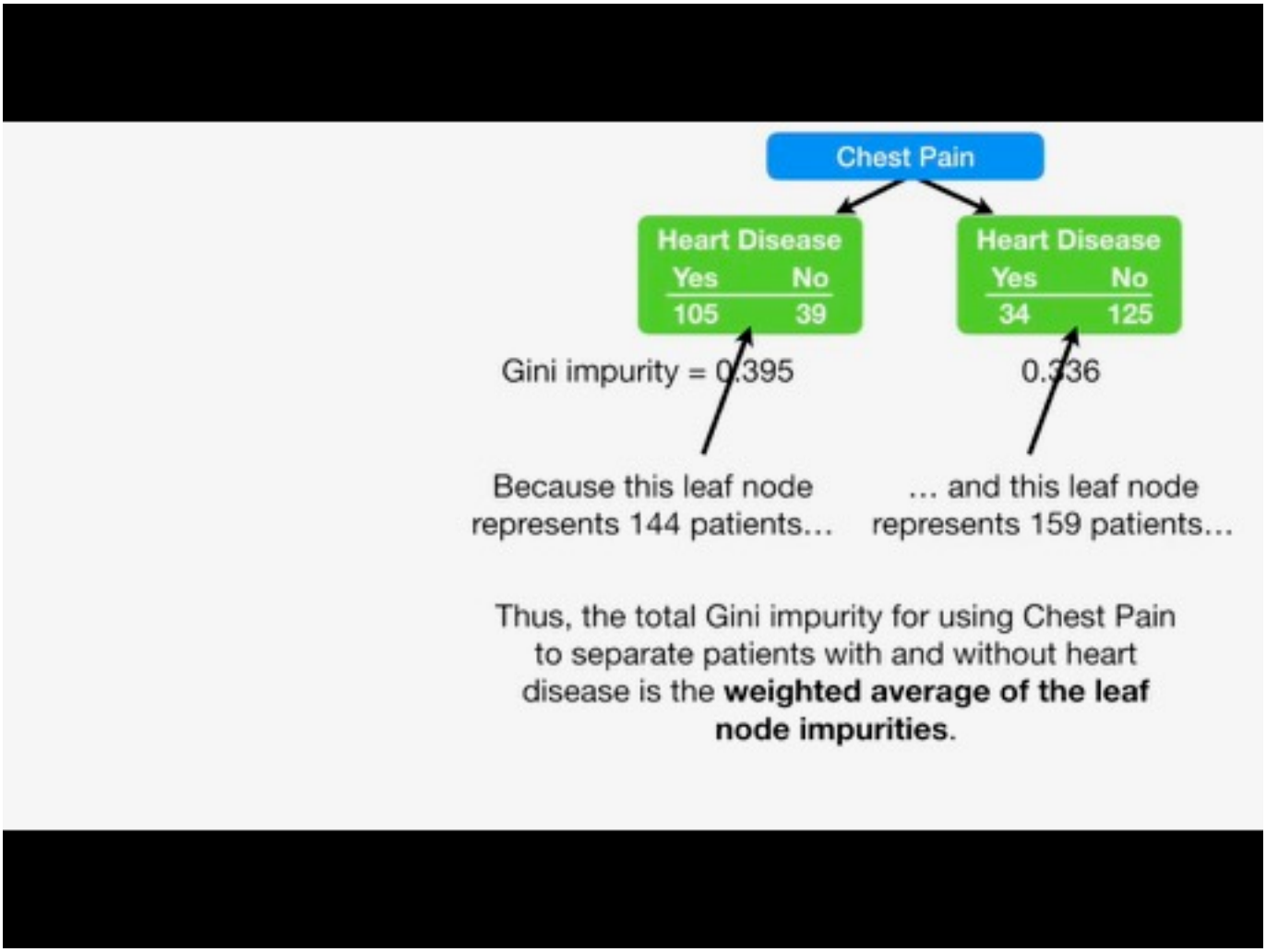


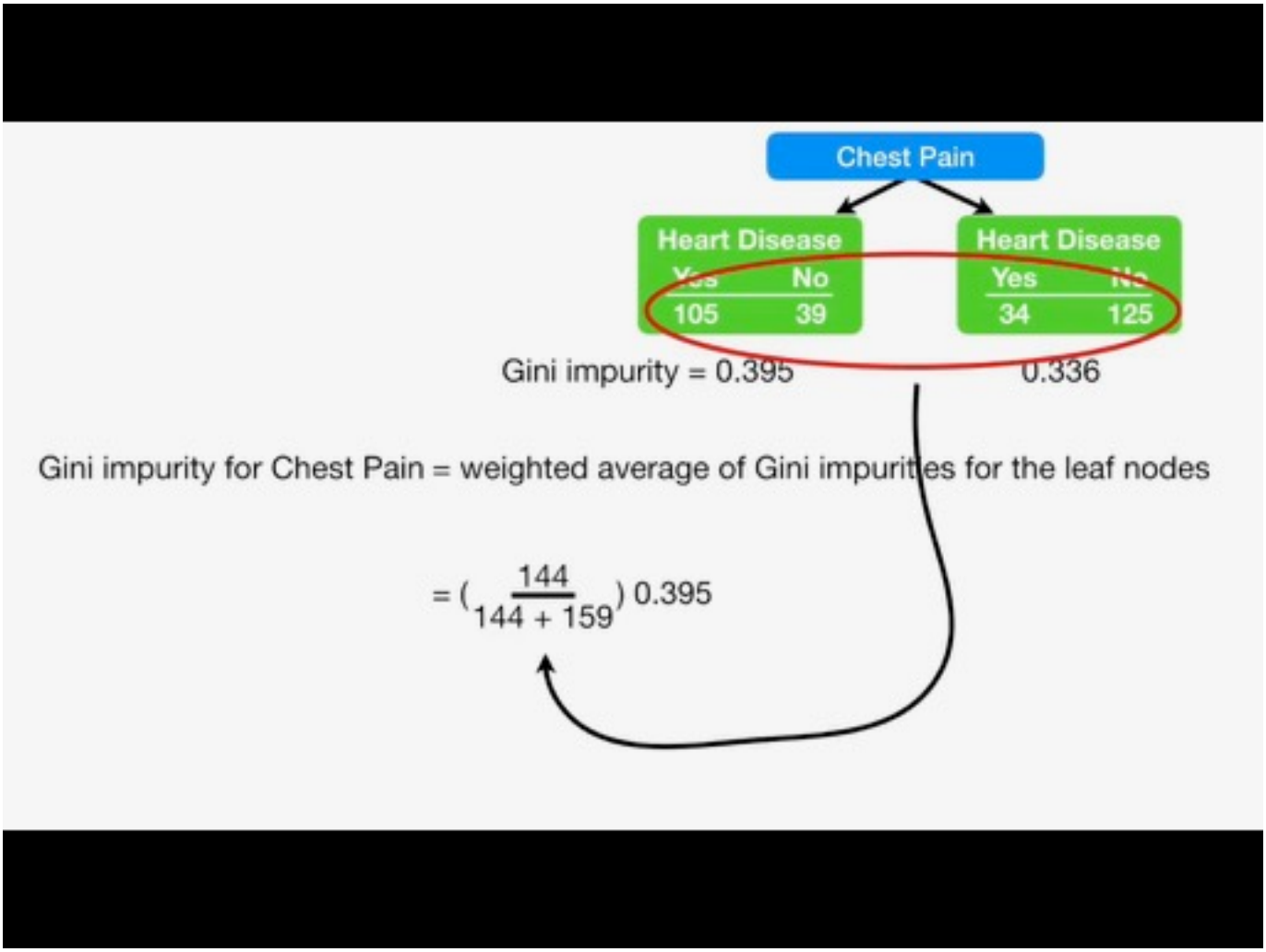
Gini impurity = 0.395

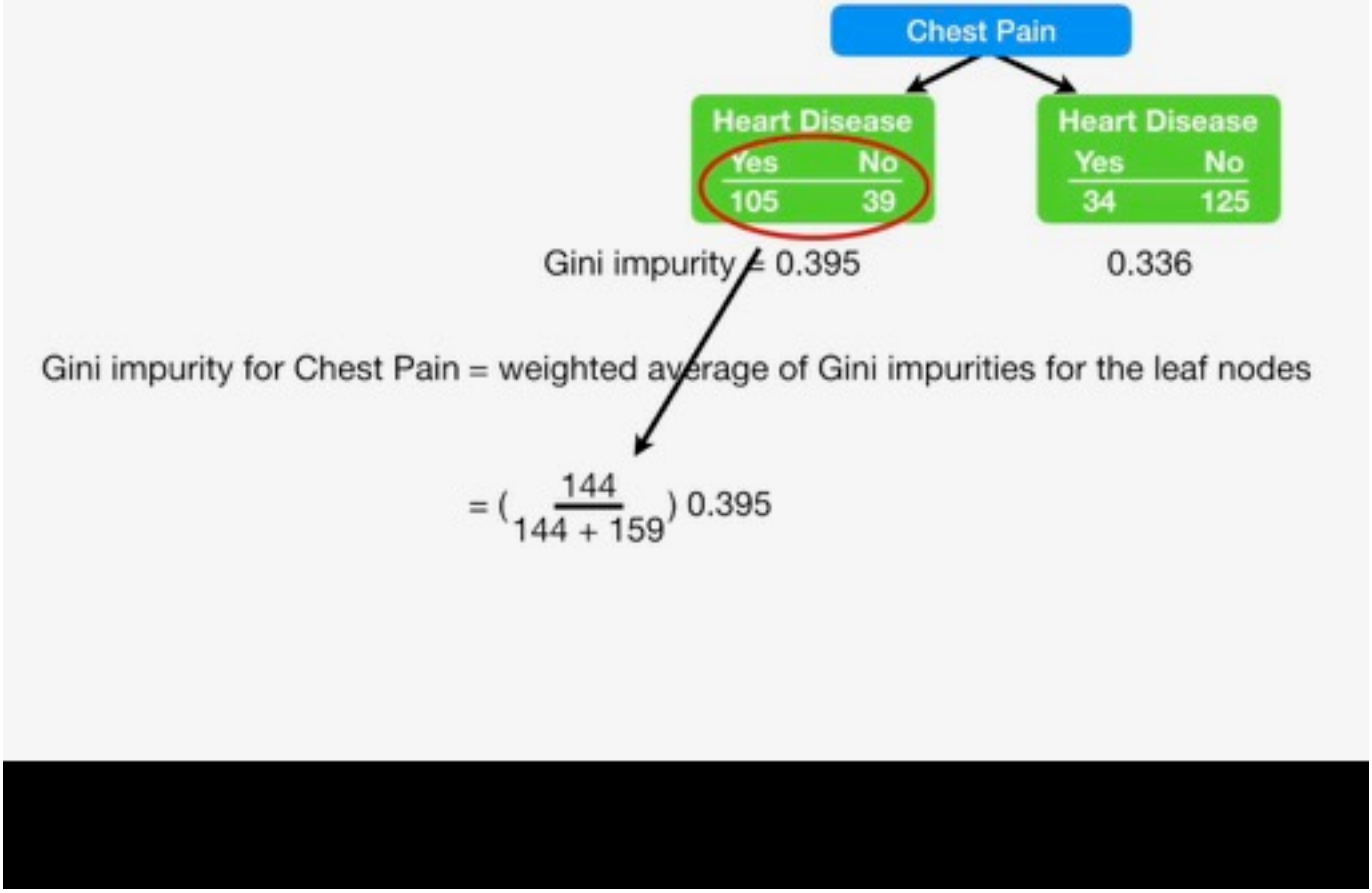
0.336

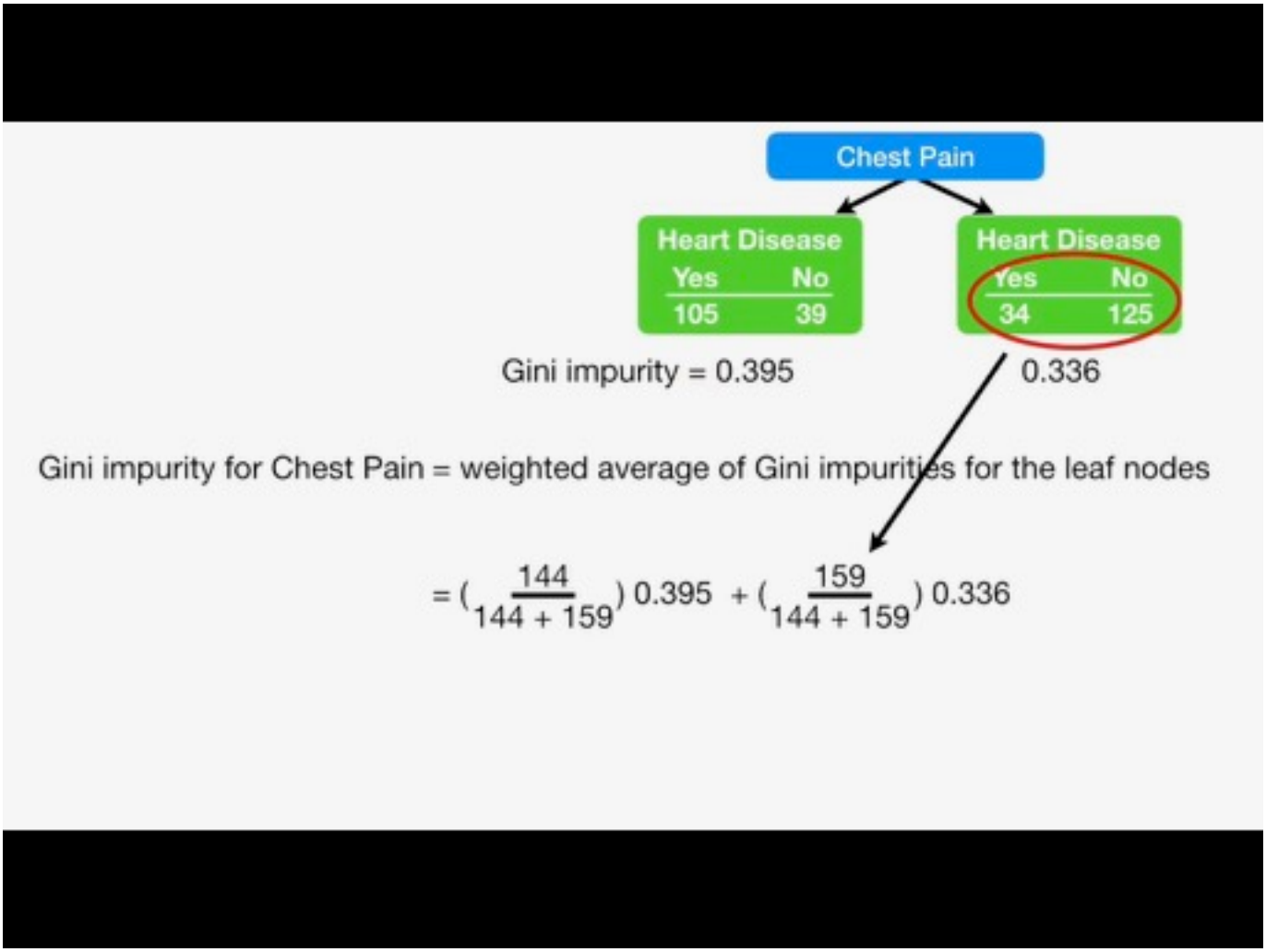
Now that we have measured the Gini impurity for both leaf nodes, we can calculate the total Gini impurity for using Chest Pain to separate patients with and without heart disease.

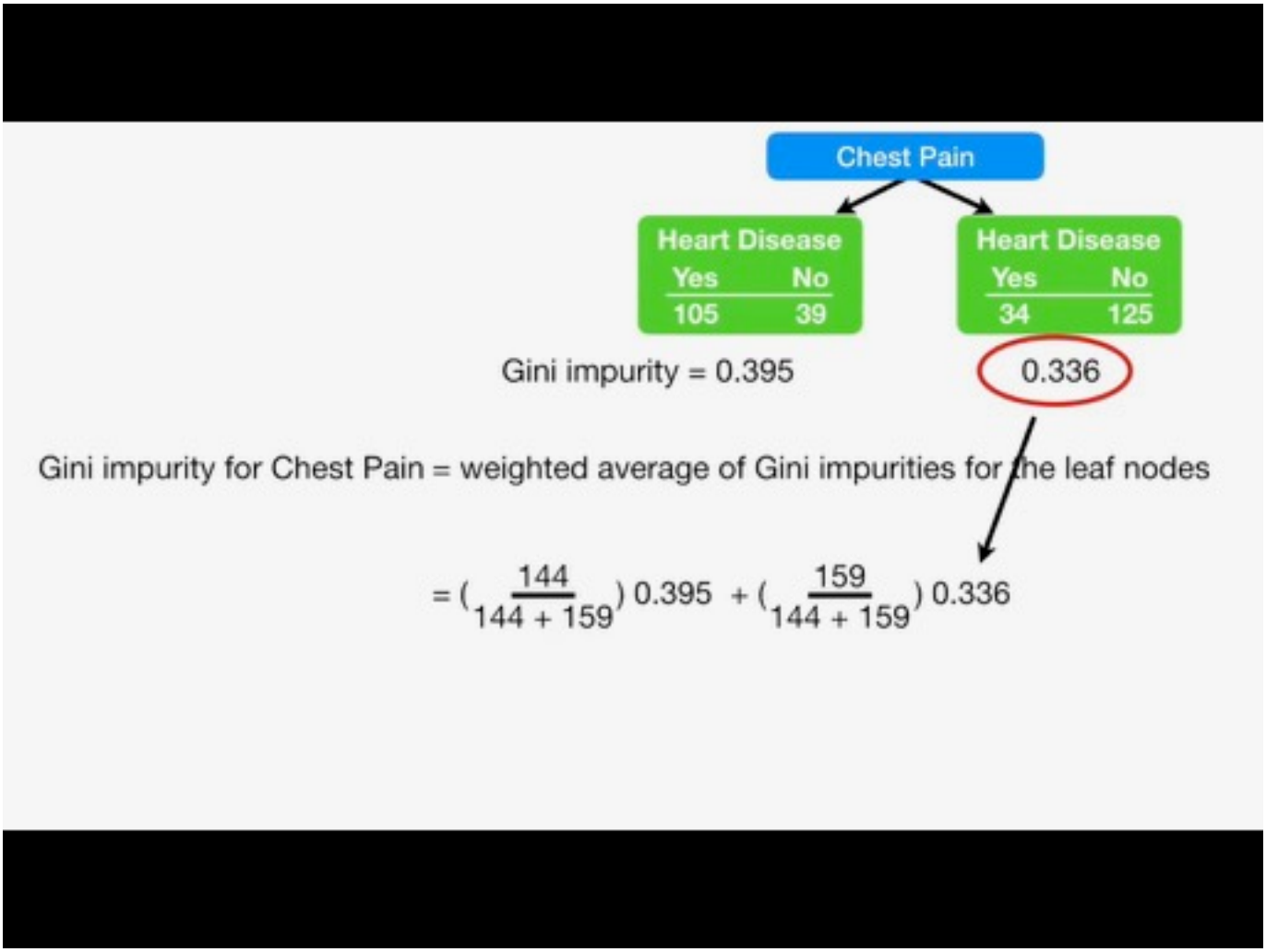

Chest Pain
Heart Disease
Yes No
105 39
Heart Disease
Yes No
34 125
Gini impurity = 0.395
0.336
Because this leaf node represents 144 patients...
... and this leaf node represents 159 patients...
...the leaf nodes do not represent the same number of patients.


Chest Pain
Heart Disease
Yes
No
105
39
Heart Disease
Yes
No
34
125
Gini impurity = 0.395
0.336
Because this leaf node represents 144 patients…
… and this leaf node represents 159 patients…
Thus, the total Gini impurity for using Chest Pain to separate patients with and without heart disease is the **weighted average of the leaf node impurities**.


Chest Pain
Heart Disease
No
105
39
Heart Disease
Yes
34
125
Gini impurity = 0.395
0.336
Gini impurity for Chest Pain = weighted average of Gini impurities for the leaf nodes
= (144 / 144 + 159) 0.395


Chest Pain
Heart Disease
Yes
No
105
39
Heart Disease
Yes
No
34
125
Gini impurity = 0.395
0.336
Gini impurity for Chest Pain = weighted average of Gini impurities for the leaf nodes
= (144 / 144 + 159) 0.395


Chest Pain
Heart Disease
Yes No
105 39
Heart Disease
Yes No
34 125
Gini impurity = 0.395
0.336
Gini impurity for Chest Pain = weighted average of Gini impurities for the leaf nodes
$= \left(\frac{144}{144 + 159}\right) 0.395 + \left(\frac{159}{144 + 159}\right) 0.336$


Chest Pain
Heart Disease
Yes No
105 39
Heart Disease
Yes No
34 125
Gini impurity = 0.395
0.336
Gini impurity for Chest Pain = weighted average of Gini impurities for the leaf nodes
$= (\frac{144}{144 + 159}) 0.395 + (\frac{159}{144 + 159}) 0.336$

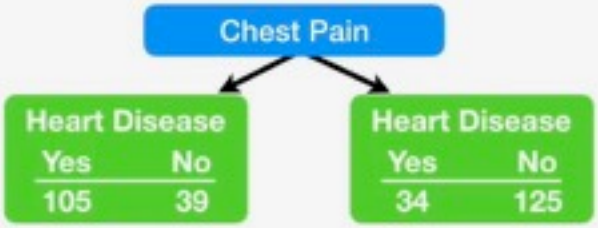
Gini impurity = 0.395 0.336

Gini impurity for Chest Pain = weighted average of Gini impurities for the leaf nodes

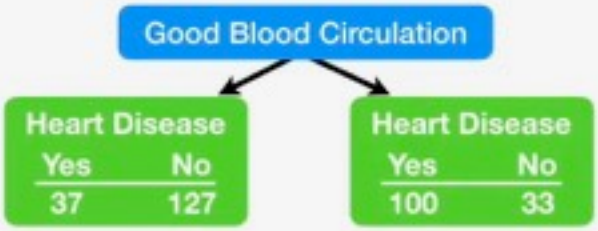
$$= \left(\frac{144}{144 + 159}\right) 0.395 + \left(\frac{159}{144 + 159}\right) 0.336$$

$$= 0.364$$

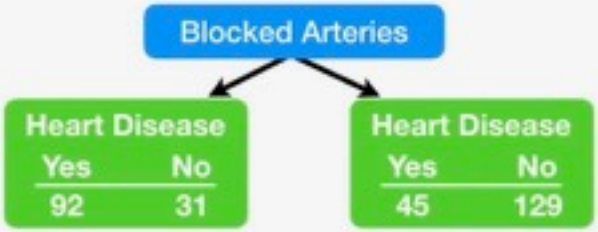
Gini impurity for Chest Pain = 0.364



Gini impurity for Good Blood Circulation = 0.360



Gini impurity for Blocked Arteries = 0.381

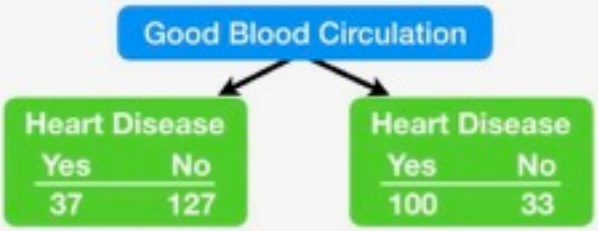


Gini impurity for Chest Pain = 0.364

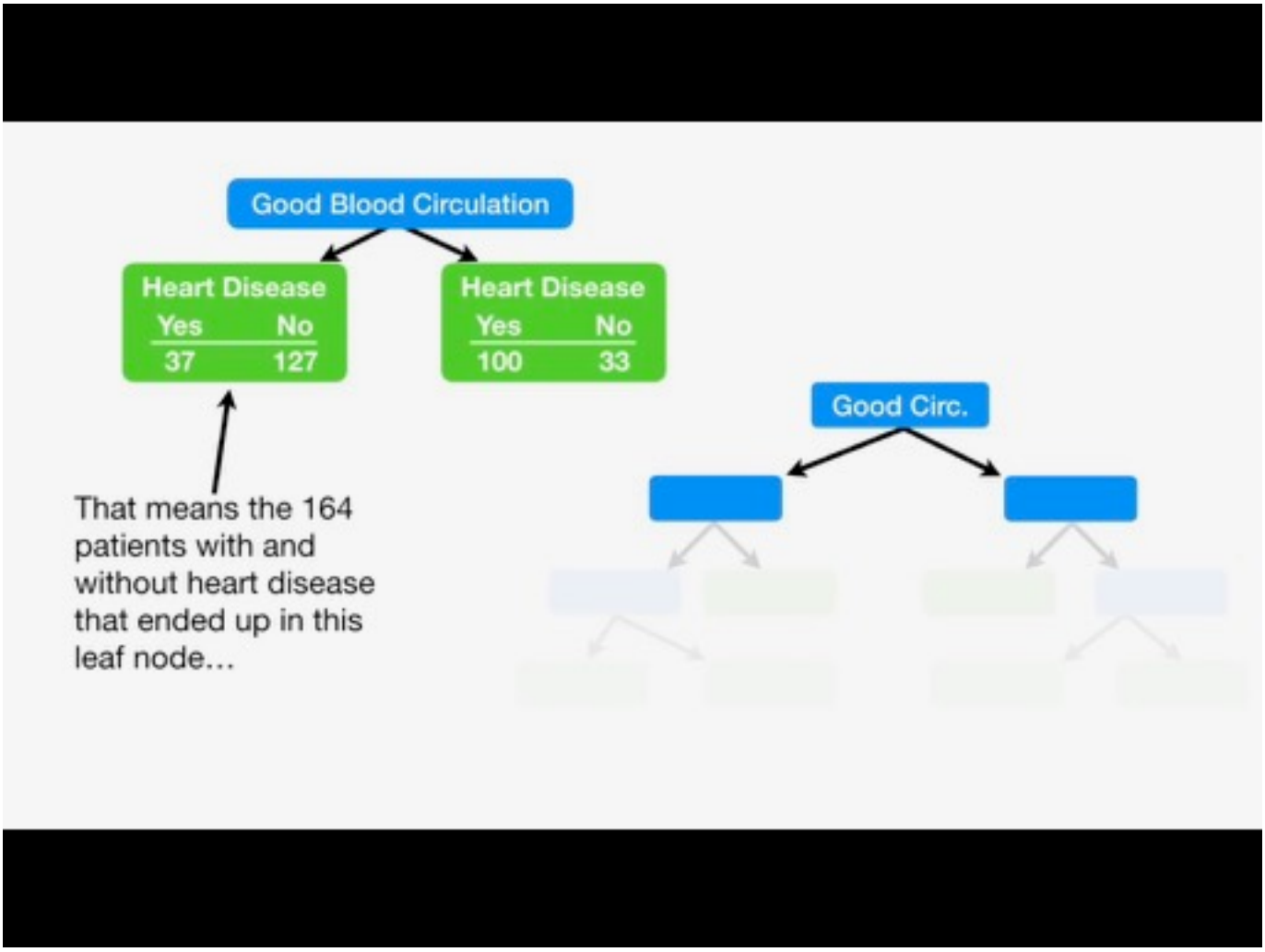
Gini impurity for Good Blood Circulation = 0.360

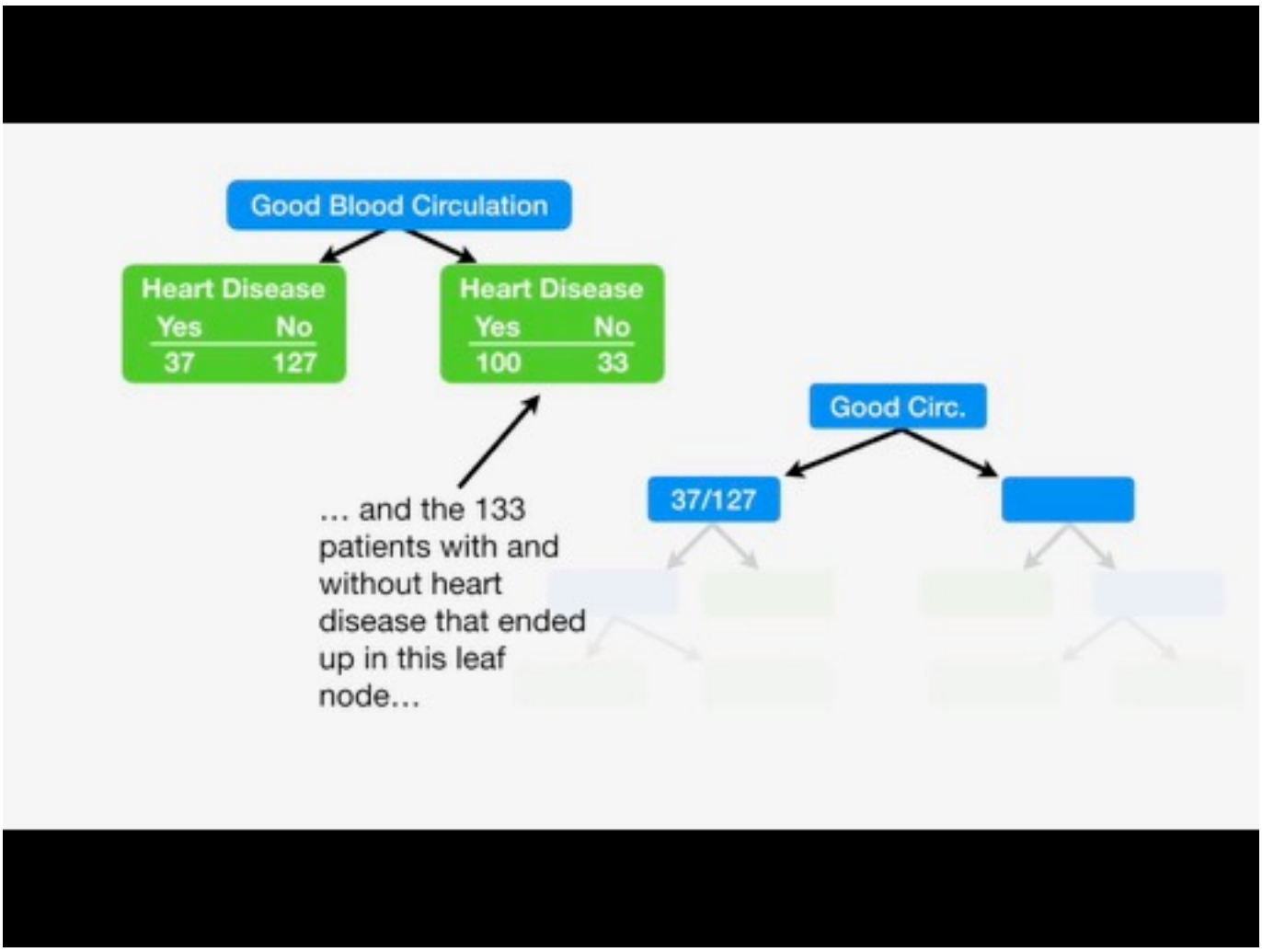
Good Blood Circulation has the lowest impurity (it separates patients with and without heart disease the best)...

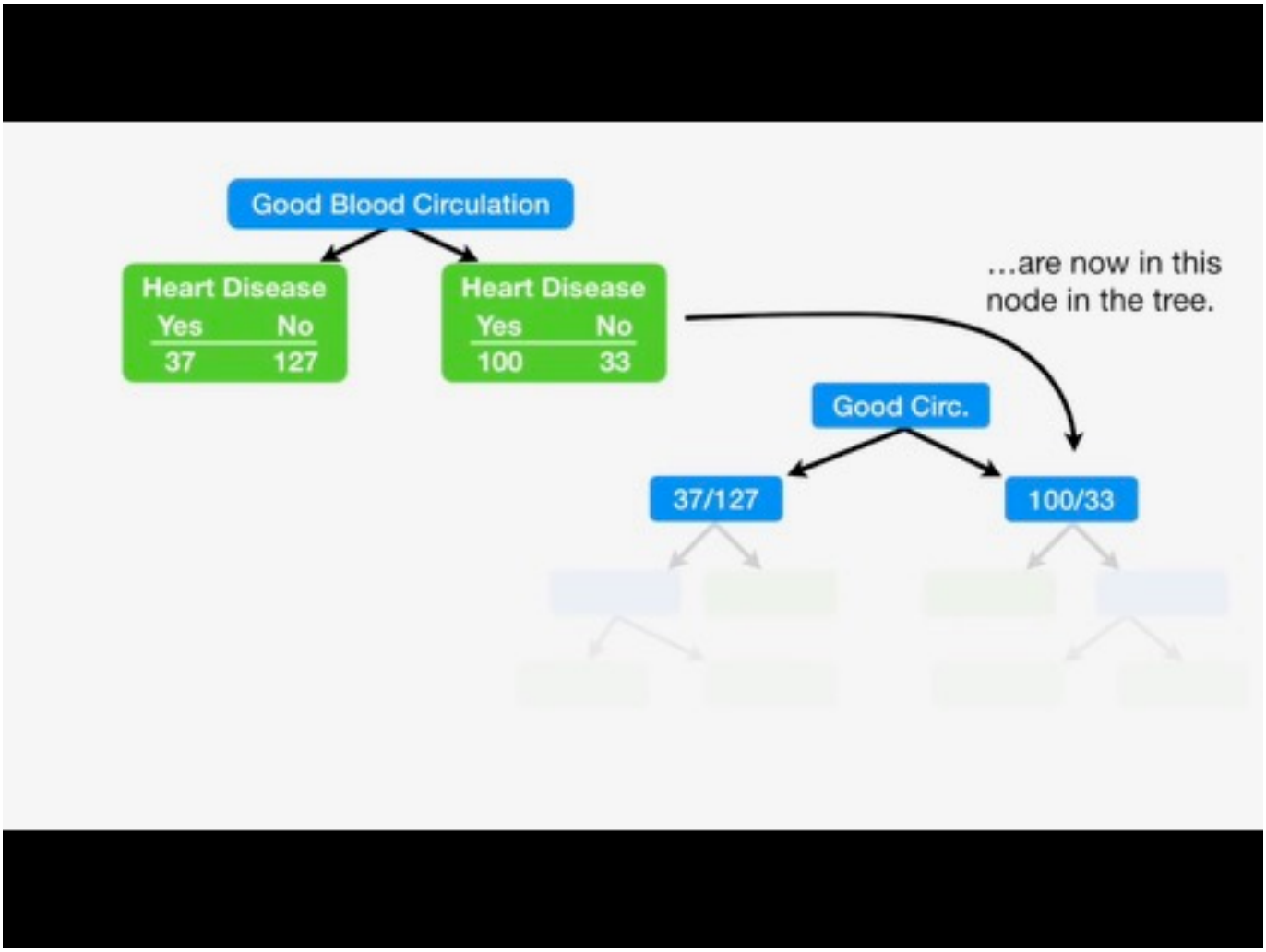
Gini impurity for Blocked Arteries = 0.381



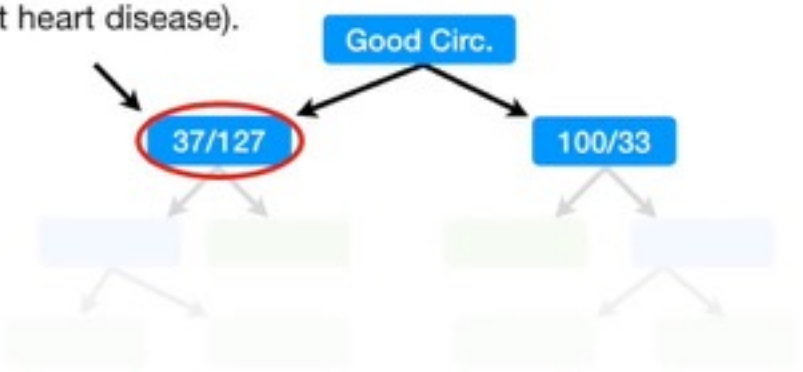
When we divided all of the patients using **Good Blood Circulation**, we ended up with "impure" leaf nodes.

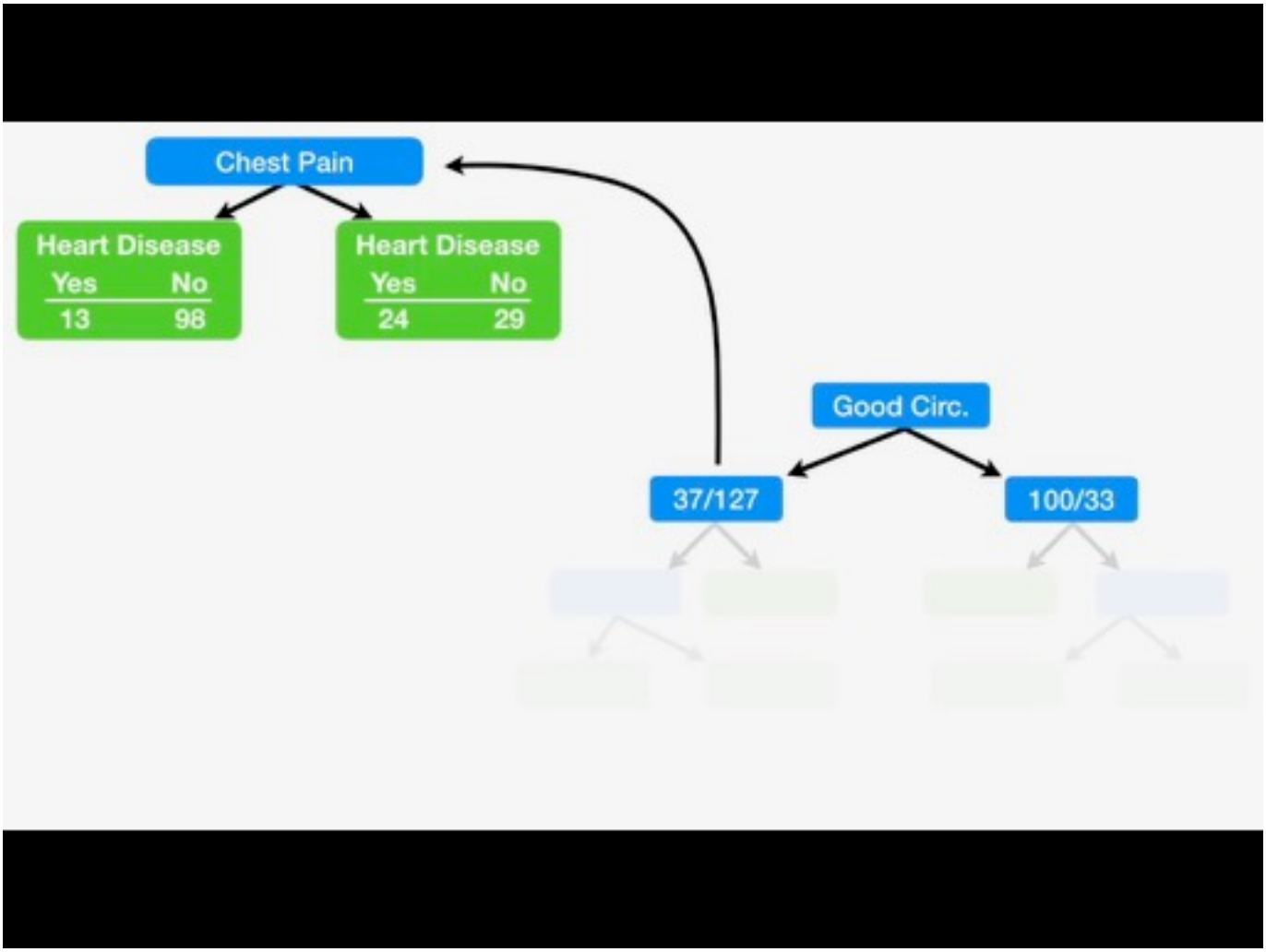

Good Blood Circulation
Heart Disease
Yes No
37 127
Heart Disease
Yes No
100 33
That means the 164 patients with and without heart disease that ended up in this leaf node…
Good Circ.

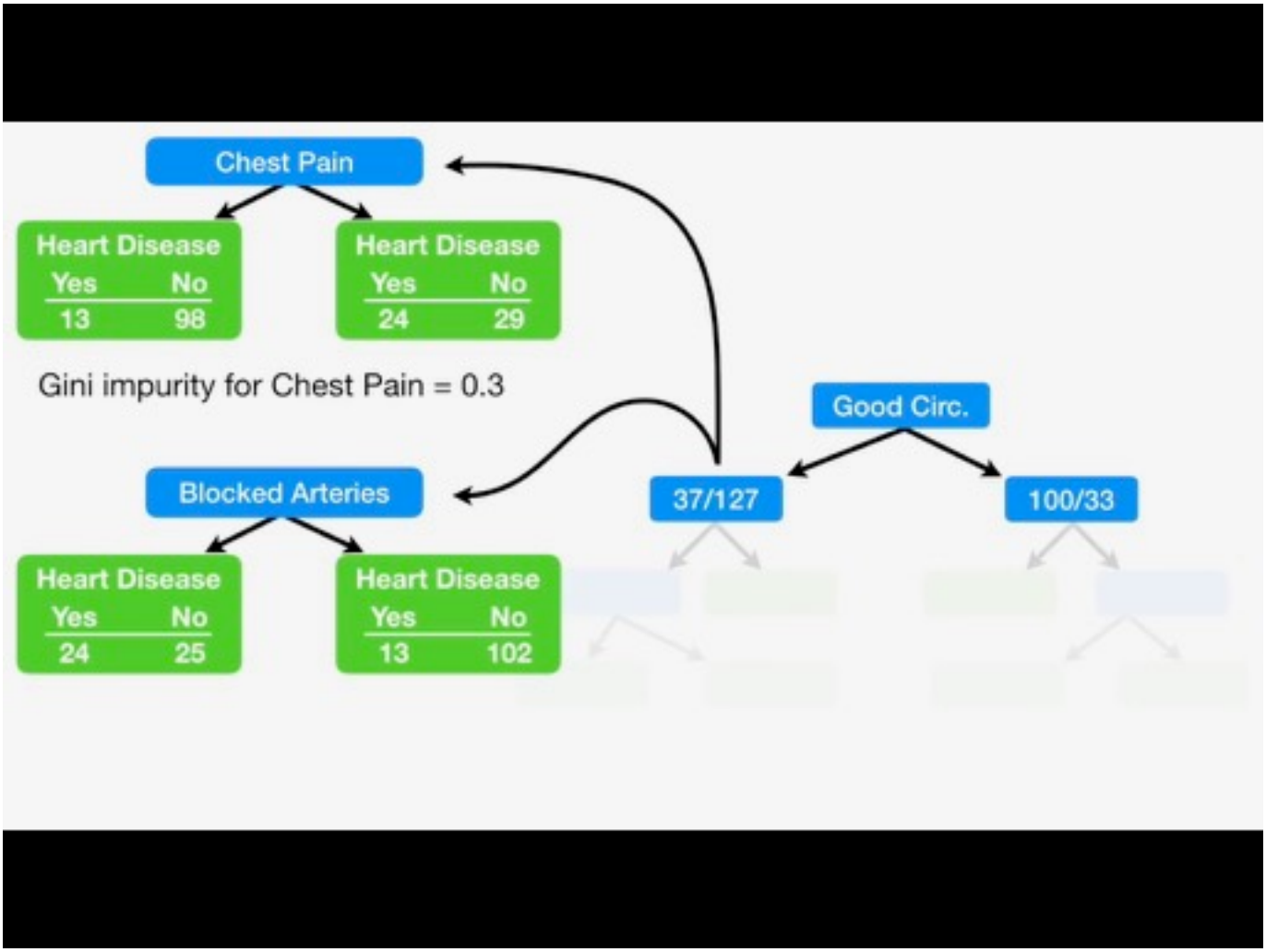

Good Blood Circulation
Heart Disease
Yes No
37 127
Heart Disease
Yes No
100 33
… and the 133 patients with and without heart disease that ended up in this leaf node…
Good Circ.
37/127

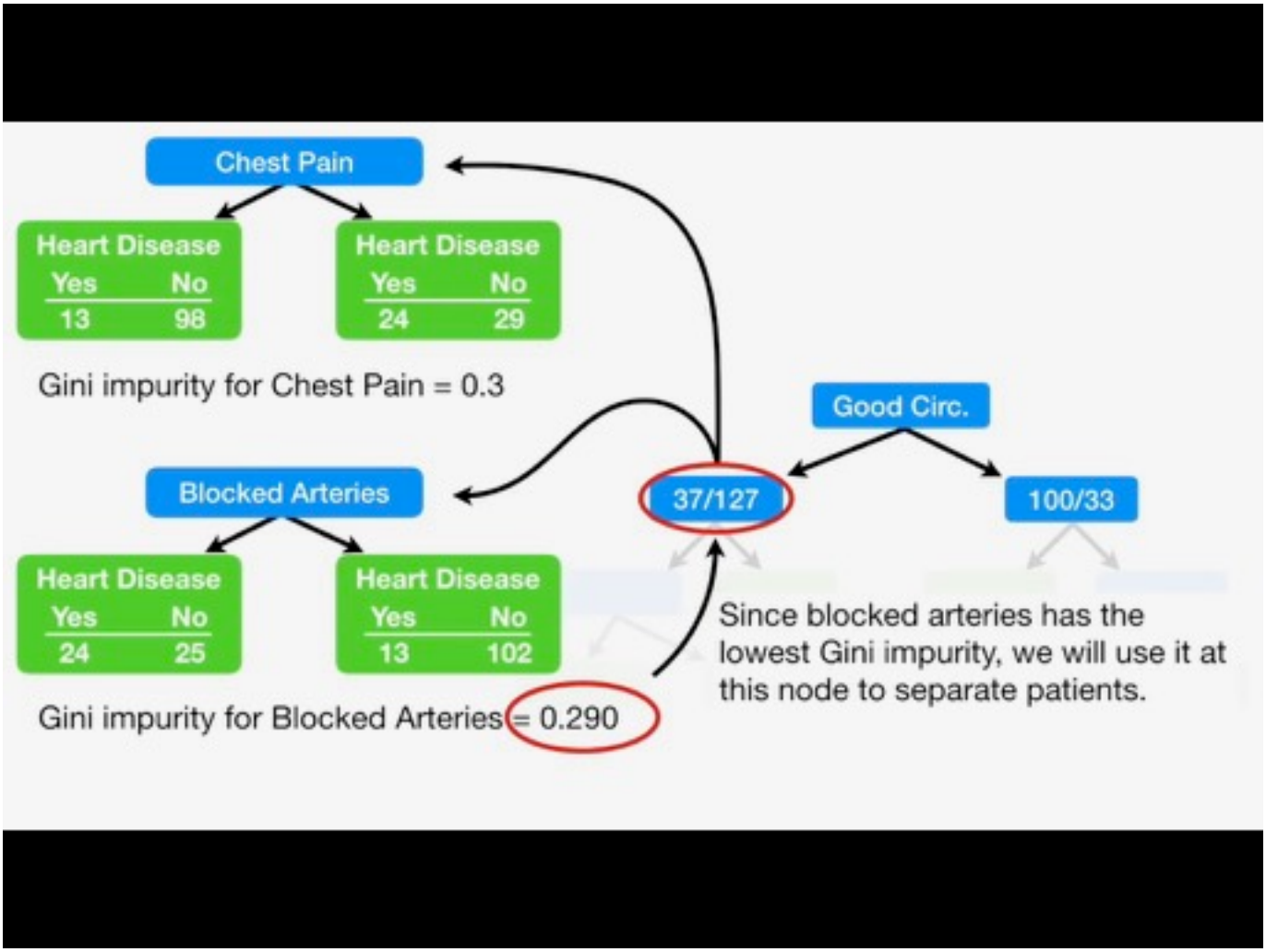

Good Blood Circulation
Heart Disease
Yes
No
37
127
Heart Disease
Yes
No
100
33
...are now in this node in the tree.
Good Circ.
37/127
100/33

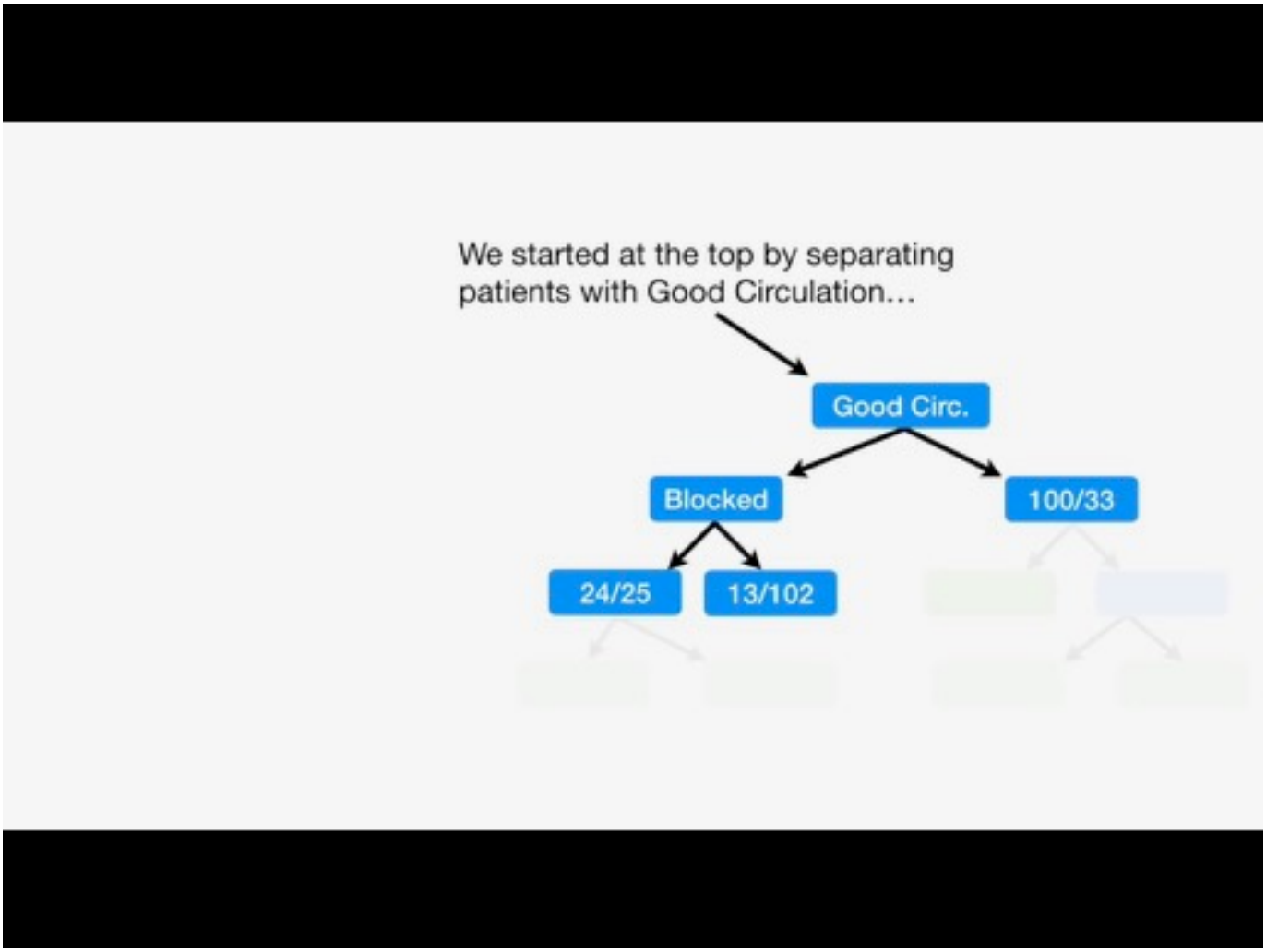
Now we need to figure how well **chest pain** and **blocked arteries** separate these 164 patients (37 with heart disease and 127 without heart disease).



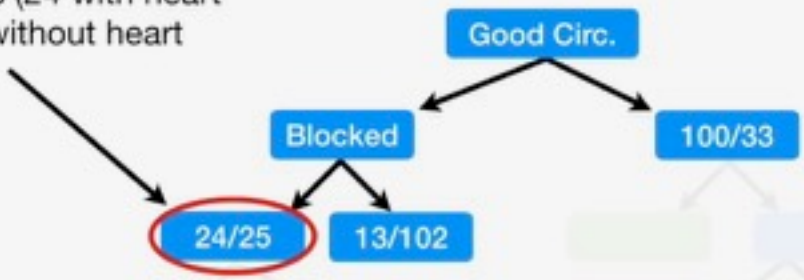

Chest Pain
Heart Disease
Yes No
13 98
Heart Disease
Yes No
24 29
Good Circ.
37/127
100/33

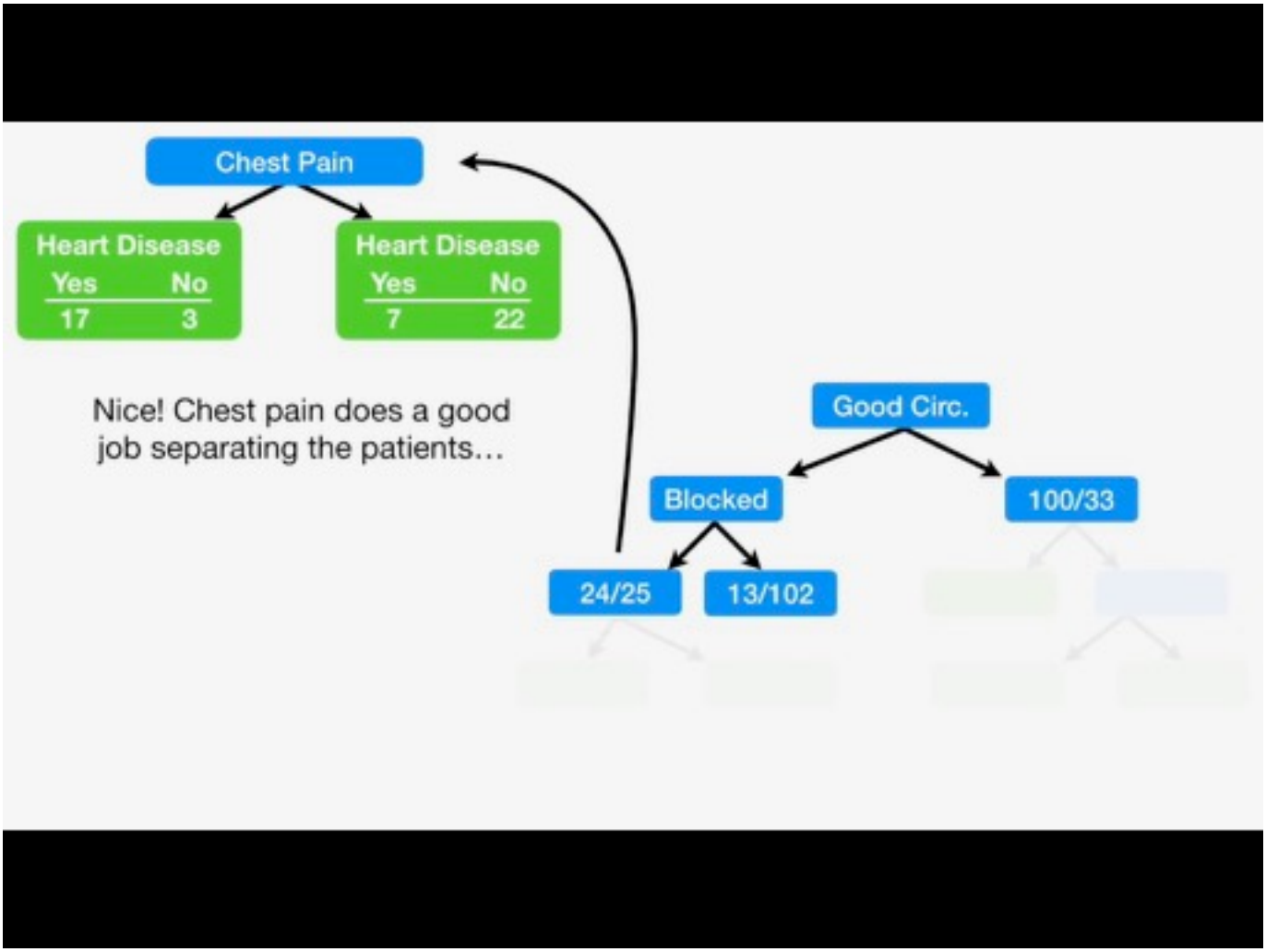

Chest Pain
Heart Disease
Yes No
13 98
Heart Disease
Yes No
24 29
Gini impurity for Chest Pain = 0.3
Blocked Arteries
Heart Disease
Yes No
24 25
Heart Disease
Yes No
13 102
Good Circ.
37/127
100/33

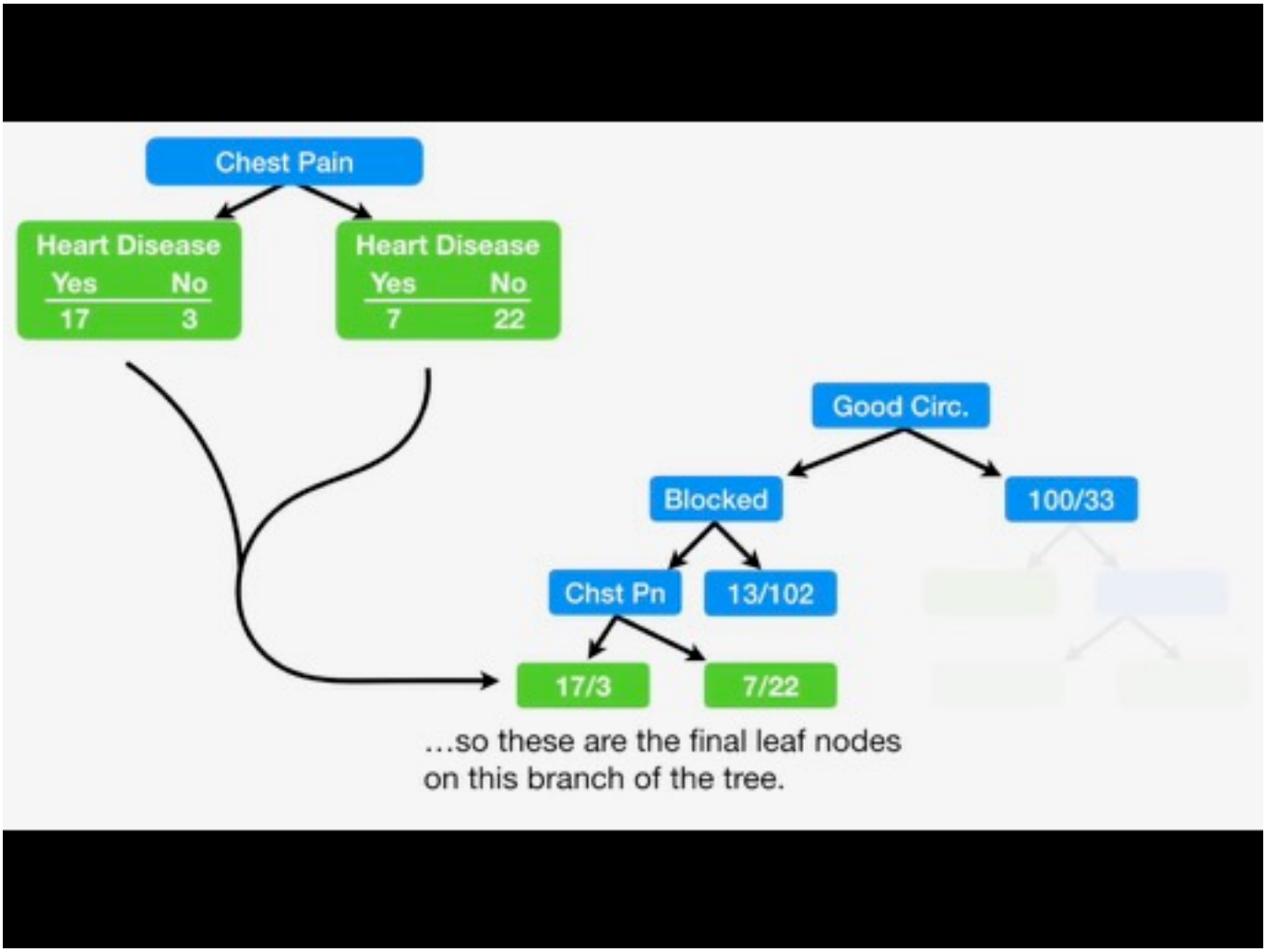

Chest Pain
Heart Disease
Yes No
13 98
Heart Disease
Yes No
24 29
Gini impurity for Chest Pain = 0.3
Blocked Arteries
Heart Disease
Yes No
24 25
Heart Disease
Yes No
13 102
Gini impurity for Blocked Arteries = 0.290
Good Circ.
37/127
100/33
Since blocked arteries has the lowest Gini impurity, we will use it at this node to separate patients.


We started at the top by separating patients with Good Circulation…
Good Circ.
Blocked
100/33
24/25
13/102

All we have left is Chest Pain, so first we'll see how well it separates these 49 patients (24 with heart disease and 25 without heart disease).




Chest Pain
Heart Disease
Yes No
17 3
Heart Disease
Yes No
7 22
Nice! Chest pain does a good job separating the patients…
Good Circ.
Blocked
100/33
24/25
13/102

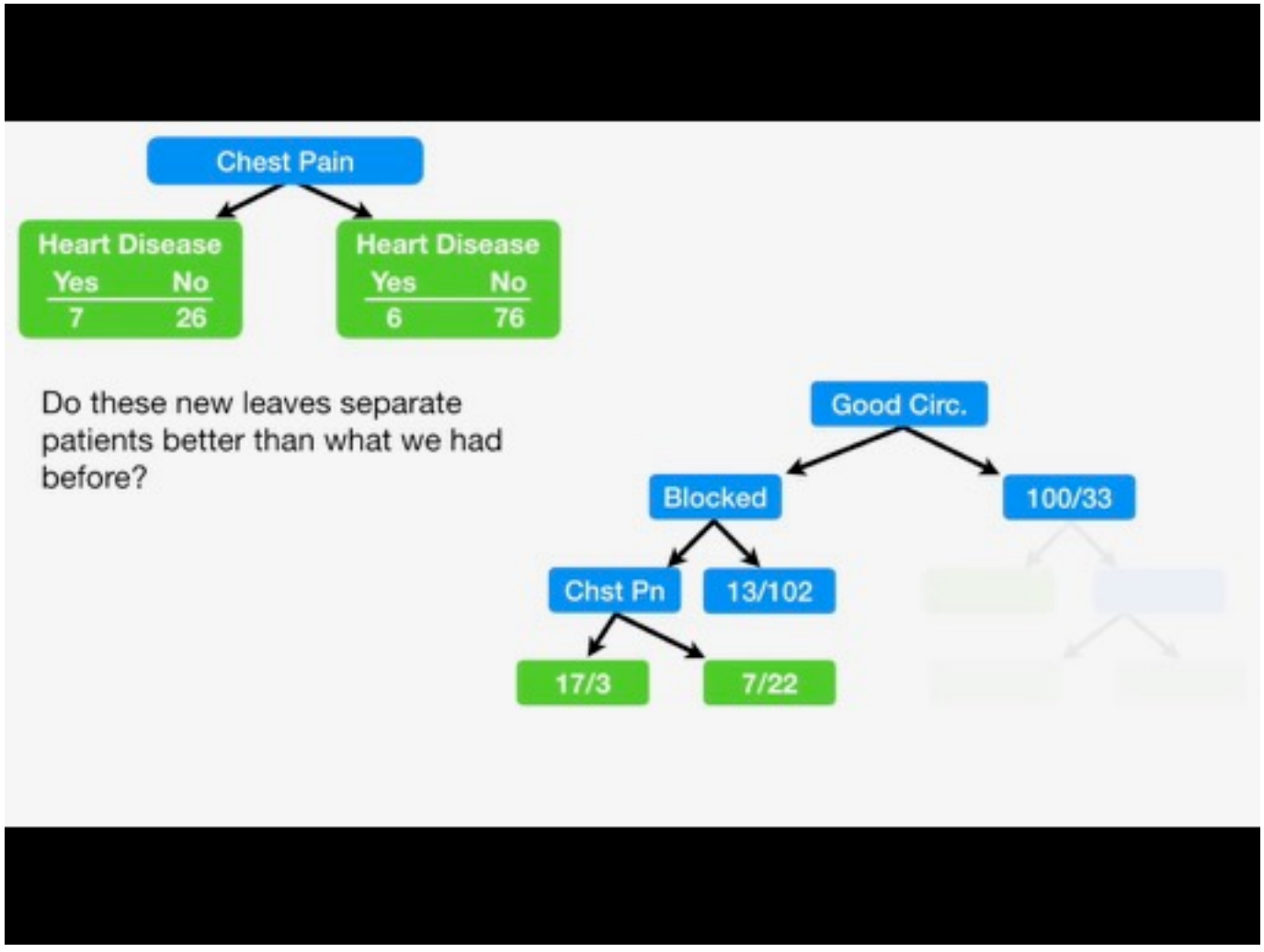

Chest Pain
Heart Disease
Yes No
17 3
Heart Disease
Yes No
7 22
Good Circ.
Blocked
100/33
Chst Pn
13/102
17/3
7/22
…so these are the final leaf nodes
on this branch of the tree.

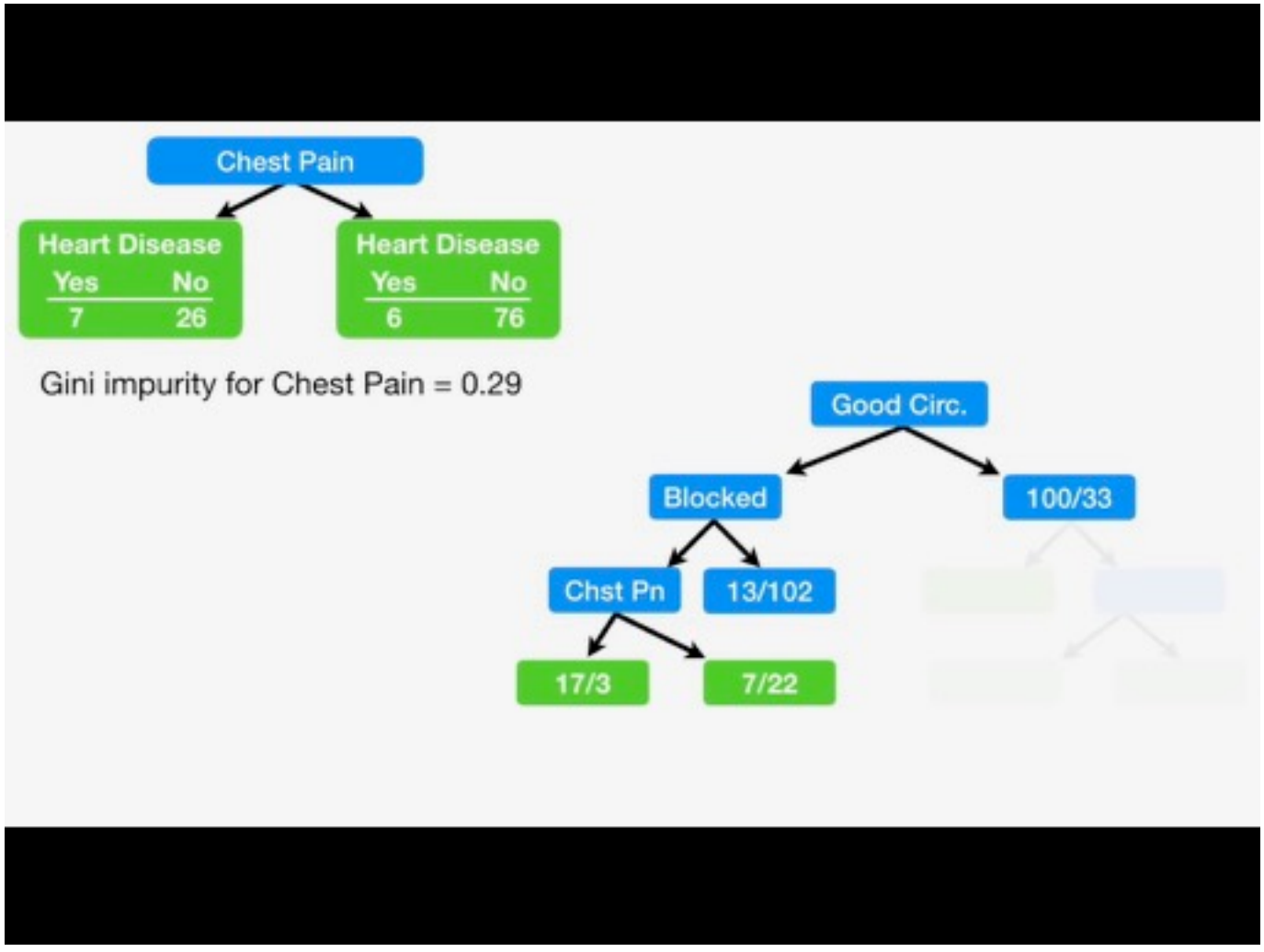
Now let's see what happens when we use chest pain to divide these 115 patients (13 with heart disease and 102 without).

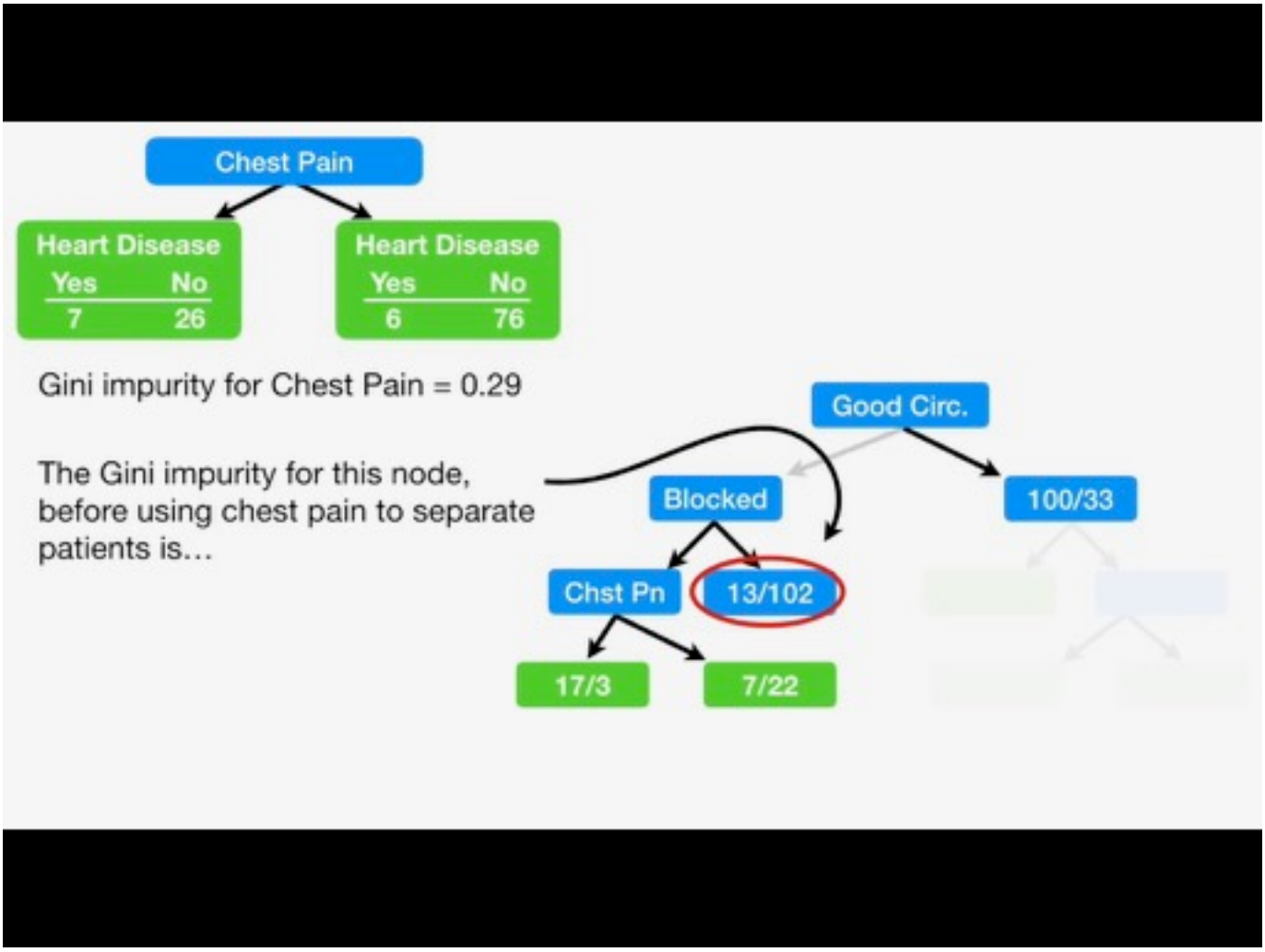
Now let's see what happens when we use chest pain to divide these 115 patients (13 with heart disease and 102 without).

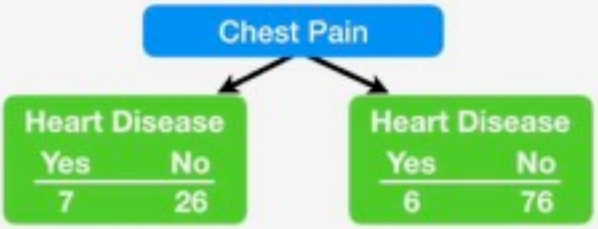
Now let's see what happens when we use chest pain to divide these 115 patients (13 with heart disease and 102 without).

NOTE: The vast majority of the patients in this node (89%) don't have heart disease.


Chest Pain
Heart Disease
Yes No
7 26
Heart Disease
Yes No
6 76
Do these new leaves separate patients better than what we had before?
Good Circ.
Blocked
100/33
Chst Pn
13/102
17/3
7/22


Chest Pain
Heart Disease
Yes No
7 26
Heart Disease
Yes No
6 76
Gini impurity for Chest Pain = 0.29
Good Circ.
Blocked
100/33
Chst Pn
13/102
17/3
7/22


Chest Pain
Heart Disease
Yes No
7 26
Heart Disease
Yes No
6 76
Gini impurity for Chest Pain = 0.29
The Gini impurity for this node, before using chest pain to separate patients is…
Good Circ.
Blocked
100/33
Chst Pn
13/102
17/3
7/22



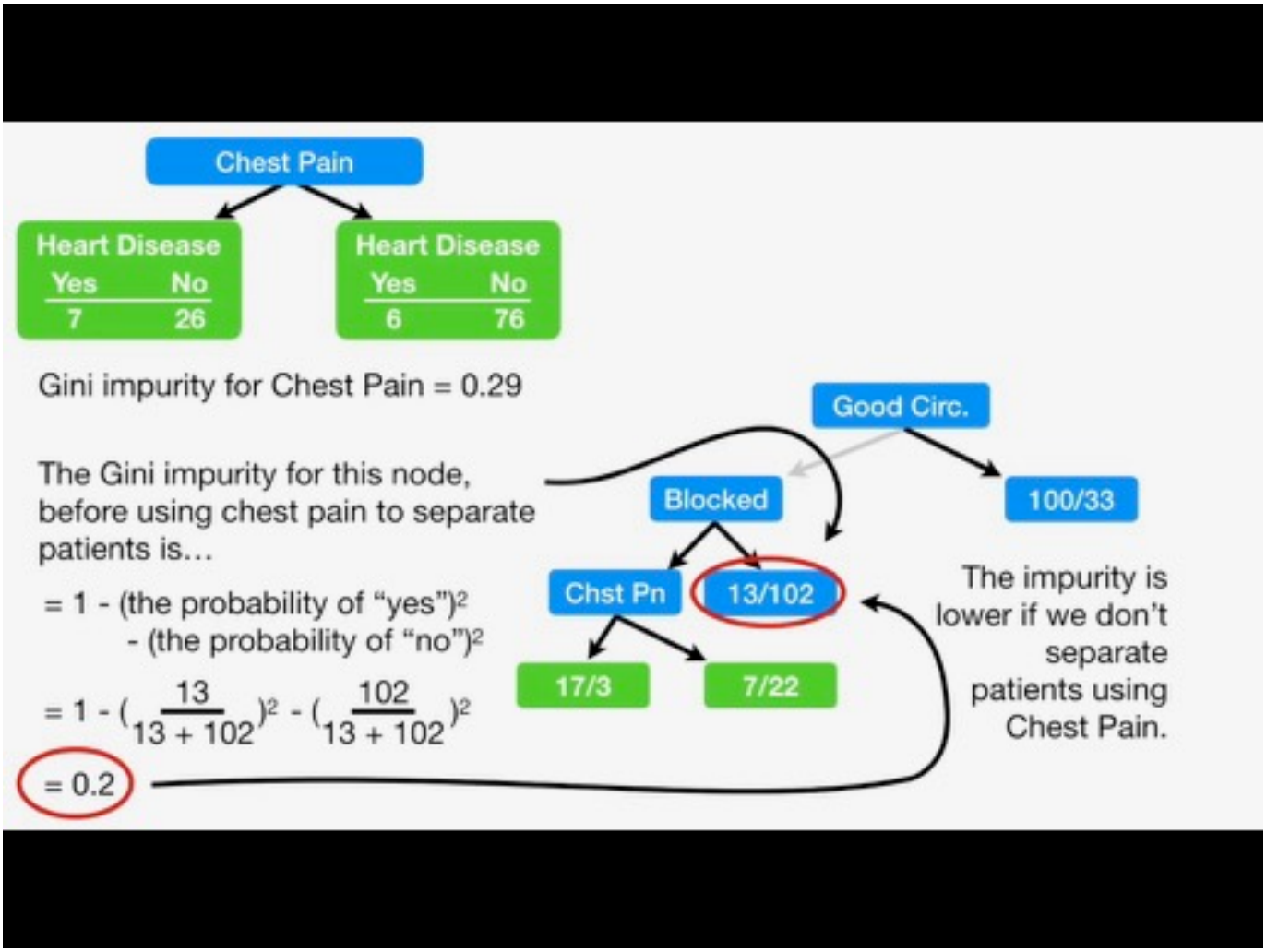
Gini impurity for Chest Pain = 0.29

The Gini impurity for this node, before using chest pain to separate patients is...

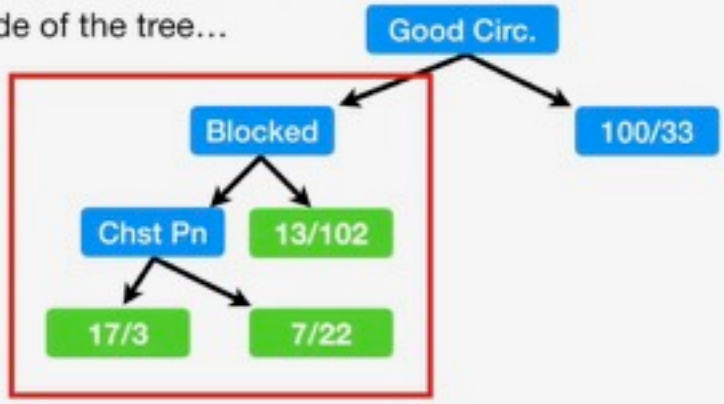
= 1 - (the probability of "yes")^2
- (the probability of "no")^2

$$= 1 - \left(\frac{13}{13 + 102}\right)^2 - \left(\frac{102}{13 + 102}\right)^2$$

= 0.2


Chest Pain
Heart Disease
Yes No
7 26
Heart Disease
Yes No
6 76
Gini impurity for Chest Pain = 0.29
The Gini impurity for this node, before using chest pain to separate patients is…
= 1 - (the probability of "yes")²
- (the probability of "no")²
= 1 - (13/(13 + 102))² - (102/(13 + 102))²
= 0.2
Good Circ.
Blocked
100/33
Chst Pn
13/102
17/3
7/22
The impurity is lower if we don't separate patients using Chest Pain.

OK, at this point we've worked out the entire left side of the tree…



The good news is that we follow the exact same steps as we did on the left side:

1) Calculate all of the Gini impurity scores.

2) If the node itself has the lowest score, than there is no point in separating the patients any more and it becomes a leaf node.

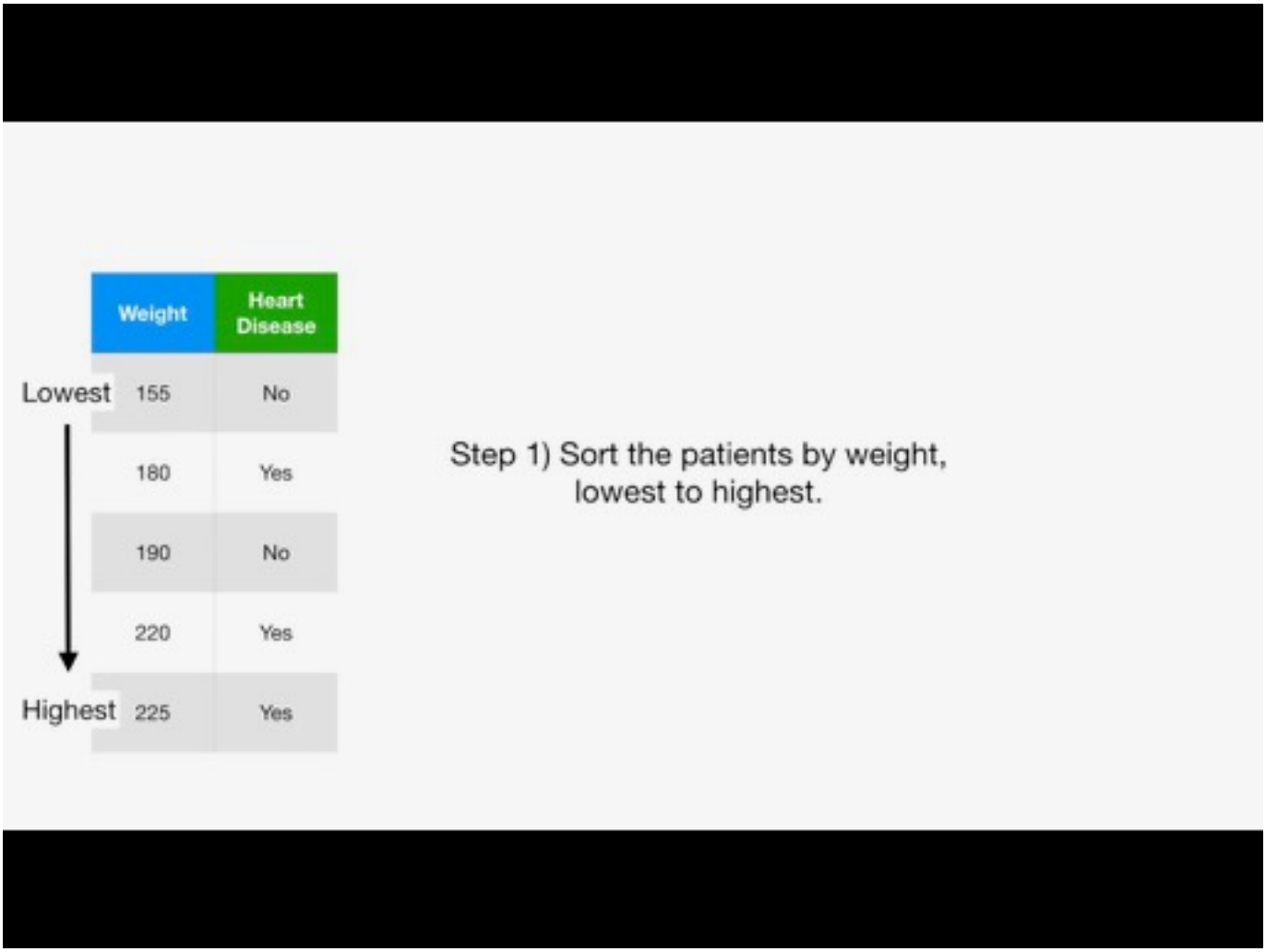
3) If separating the data results in an improvement, than pick the separation with the lowest impurity value.

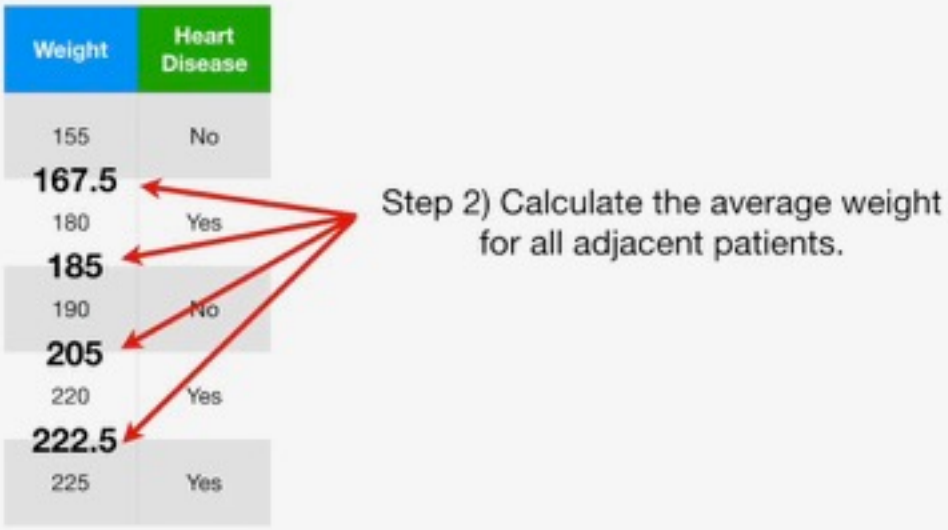
Hooray!!! We made a decision tree!!!

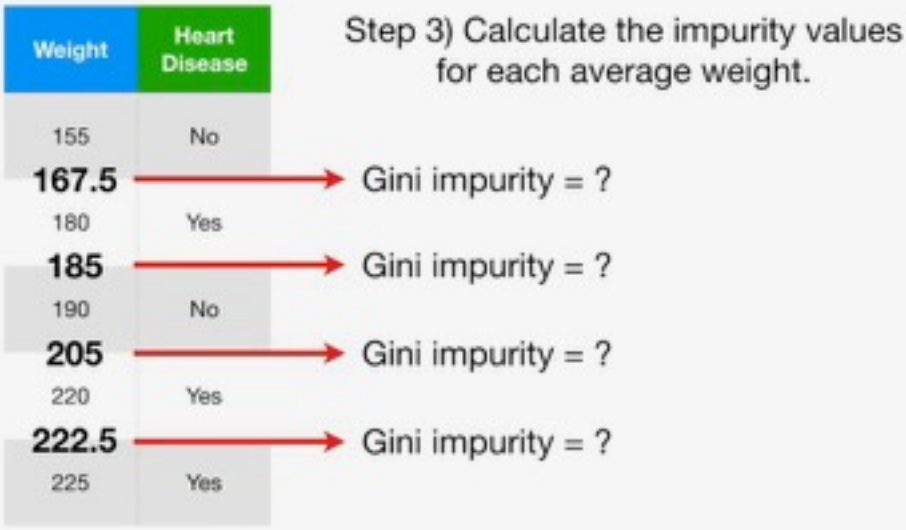
So far we've seen how to build a tree with "yes/no" questions at each step…

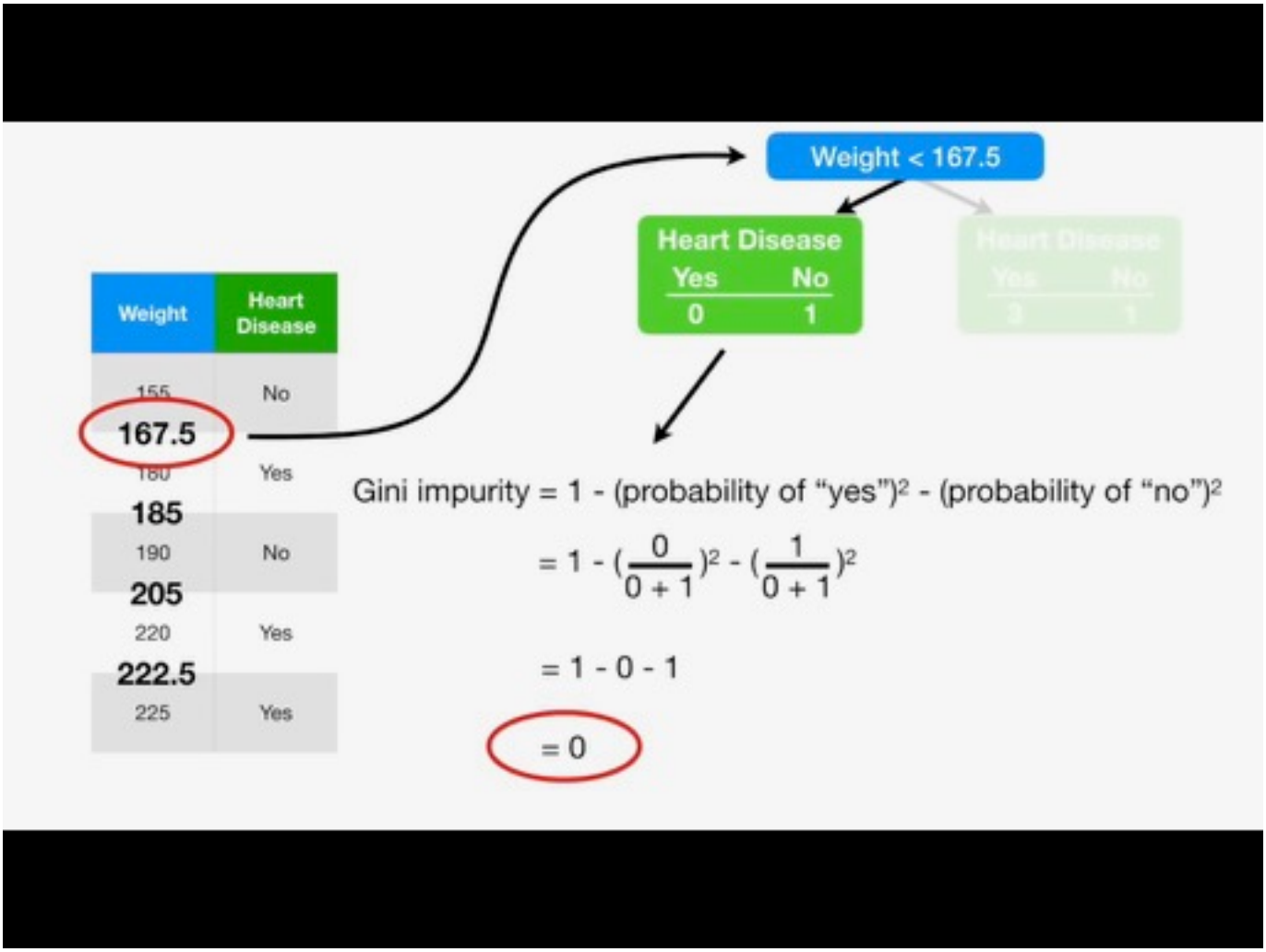
Weight	Heart Disease
220	Yes
180	Yes
225	Yes
190	No
155	No

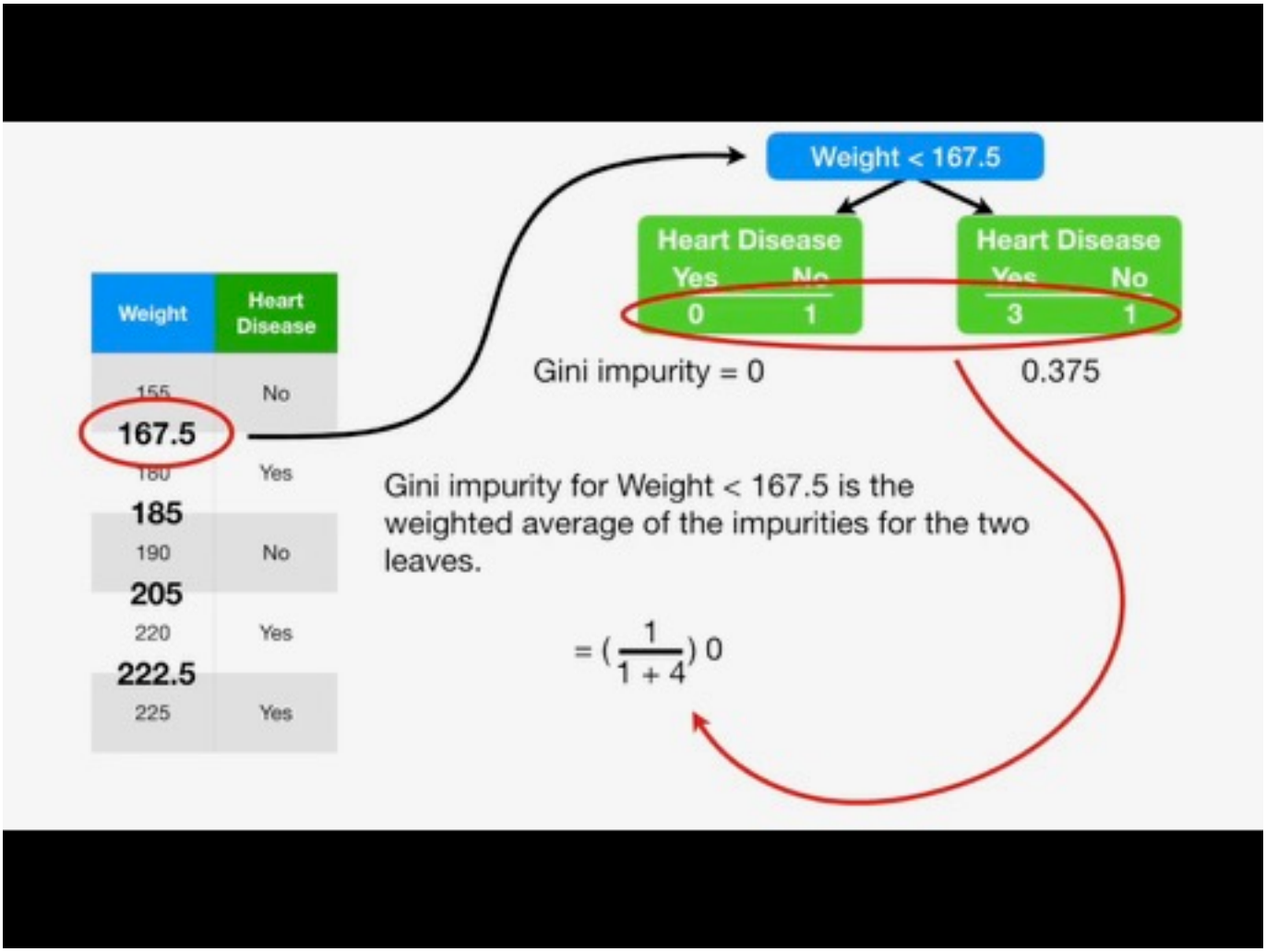
Imagine if this were our data...

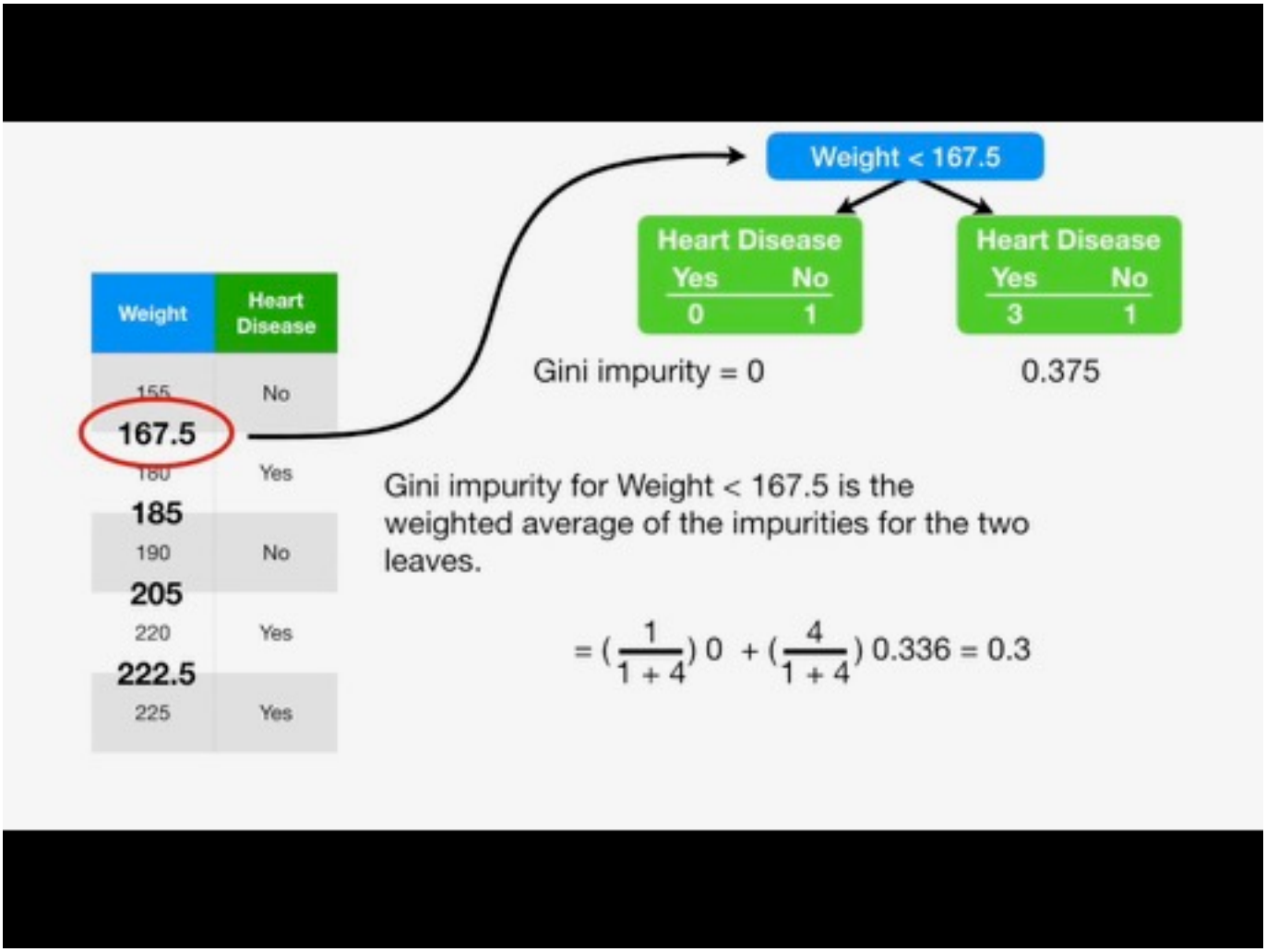

Weight
Heart Disease
Lowest
155
No
180
Yes
190
No
220
Yes
Highest
225
Yes
Step 1) Sort the patients by weight, lowest to highest.


Weight
Heart Disease
155
No
167.5
180
Yes
185
190
No
205
220
Yes
222.5
225
Yes
Step 2) Calculate the average weight for all adjacent patients.

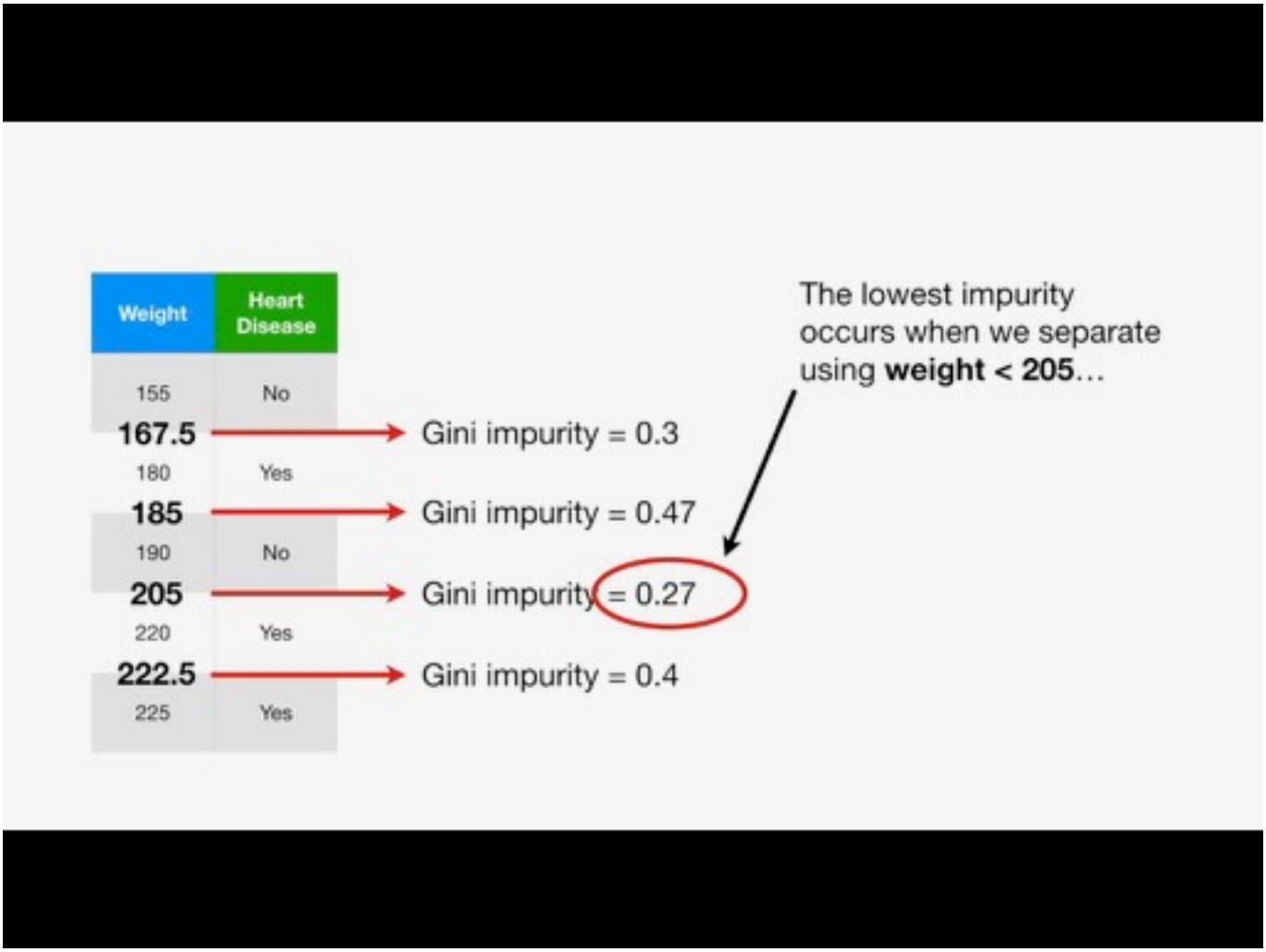

Weight
Heart Disease
155
No
167.5
Gini impurity = ?
180
Yes
185
Gini impurity = ?
190
No
205
Gini impurity = ?
220
Yes
222.5
Gini impurity = ?
225
Yes
Step 3) Calculate the impurity values for each average weight.

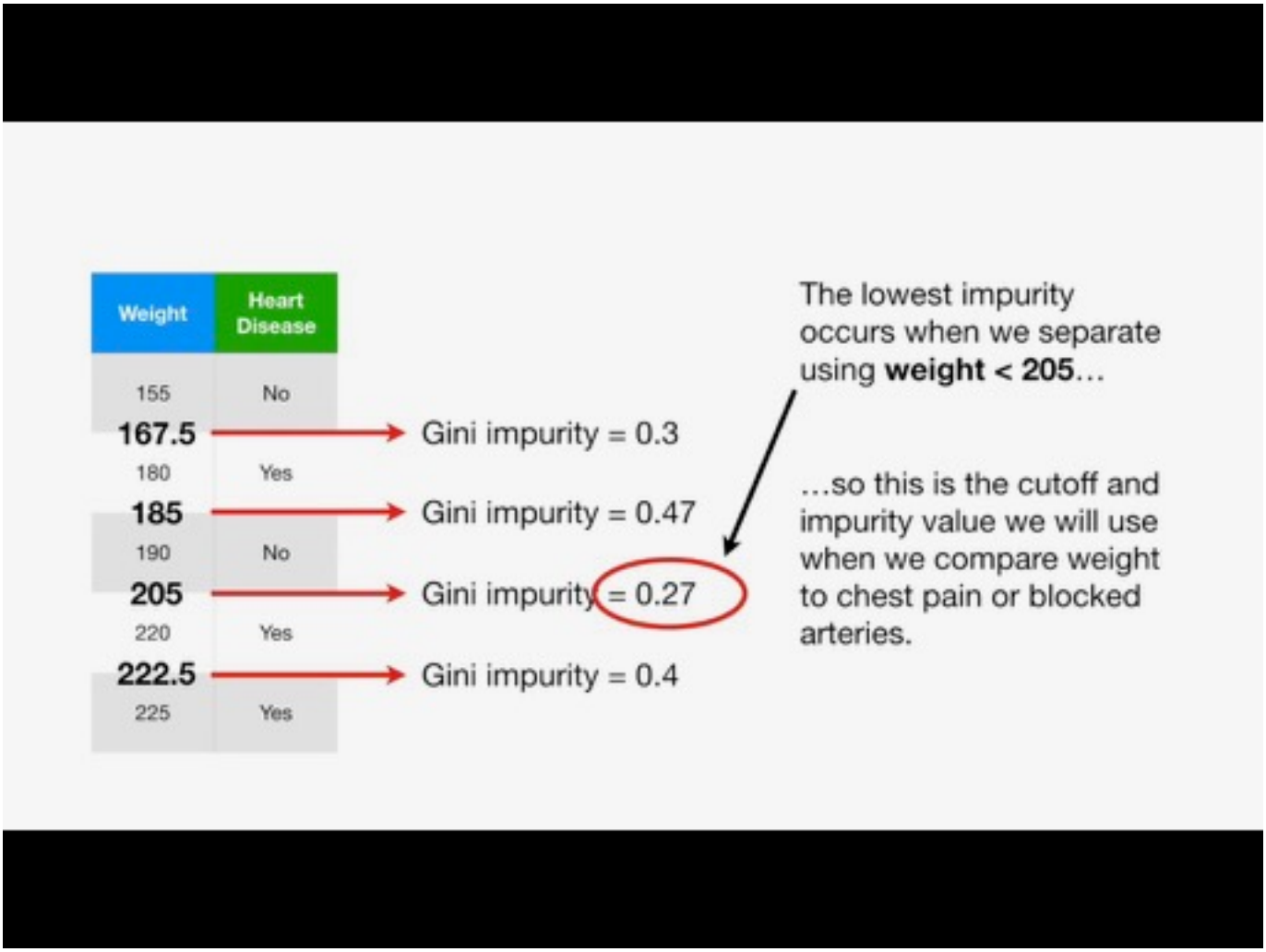

Weight < 167.5
Heart Disease
Yes No
0 1
Weight
Heart Disease
155
No
167.5
160
Yes
185
190
No
205
220
Yes
222.5
225
Yes
Gini impurity = 1 - (probability of "yes")² - (probability of "no")²
= 1 - (0 / 0 + 1)² - (1 / 0 + 1)²
= 1 - 0 - 1
= 0


Weight < 167.5
Heart Disease
Yes
No
0
1
Heart Disease
Yes
No
3
1
Weight
Heart Disease
155
No
167.5
160
Yes
185
190
No
205
220
Yes
222.5
225
Yes
Gini impurity = 0
0.375
Gini impurity for Weight < 167.5 is the weighted average of the impurities for the two leaves.
$= (\frac{1}{1+4}) 0$


Weight < 167.5
Heart Disease
Yes No
0 1
Heart Disease
Yes No
3 1
Gini impurity = 0
0.375
Weight
Heart Disease
155
No
167.5
160
Yes
185
190
No
205
220
Yes
222.5
225
Yes
Gini impurity for Weight < 167.5 is the weighted average of the impurities for the two leaves.
$= (\frac{1}{1+4})\,0 + (\frac{4}{1+4})\,0.336 = 0.3$


Weight
Heart Disease
155
No
167.5
Gini impurity = 0.3
180
Yes
185
Gini impurity = 0.47
190
No
205
Gini impurity = 0.27
220
Yes
222.5
Gini impurity = 0.4
225
Yes


Weight
Heart Disease
155
No
167.5
Gini impurity = 0.3
180
Yes
185
Gini impurity = 0.47
190
No
205
Gini impurity = 0.27
220
Yes
222.5
Gini impurity = 0.4
225
Yes
The lowest impurity occurs when we separate using **weight < 205**...


Weight
Heart Disease
155
No
167.5
Gini impurity = 0.3
180
Yes
185
Gini impurity = 0.47
190
No
205
Gini impurity = 0.27
220
Yes
222.5
Gini impurity = 0.4
225
Yes
The lowest impurity occurs when we separate using weight < 205...
...so this is the cutoff and impurity value we will use when we compare weight to chest pain or blocked arteries.

Now we've seen how to build a tree with…

1) "yes/no" questions at each step…

Now we've seen how to build a tree with…

1) "yes/no" questions at each step…

2) Numeric data, like patient weight…