

Original Dataset

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	???	???	No

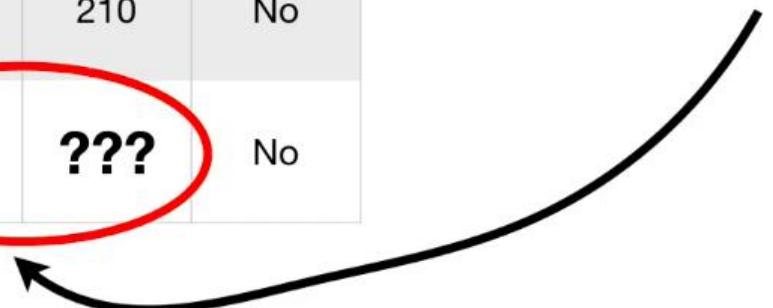


We've got data for 4 patients...

Original Dataset

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	???	???	No

However, for patient #4, we've got some missing data.



Random Forests consider 2 types of missing data...

Original Dataset

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	???	???	No

- 1) Missing data in the original dataset used to create the random forest.
- 2) Missing data in a new sample that you want to categorize.



New Sample

Chest Pain	Good Blood Circ.	Blocked Arteries	Weight	Heart Disease
No	No	No	???	

Random Forests consider 2 types of missing data...

Original Dataset

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	???	???	No

- 1) Missing data in the original dataset used to create the random forest.

We'll start with
this one...

So we want to create a random forest from this data...

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	Yes	???	???	No

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	Yes	???	???	No

← However, we don't know if this patient has blocked arteries or their weight.

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	Yes	???	???	No

The general idea for dealing with missing data in this context is to make an initial guess that could be bad, then gradually refine the guess until it is (hopefully) a good guess.



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	Yes	???	???	No

Because this person ***did not*** have Heart Disease...

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	Yes	???	???	No



...the initial, and possibly bad, guess for the blocked arteries value is just the most common value for “Blocked Arteries” found in the other samples that **do not** have Heart Disease.

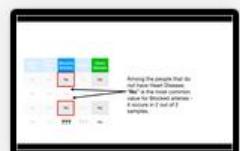
Chest Pain	Good Blood Circ.	Blocked Arteries	Weight	Heart Disease
No	No	No	125	No
Yes	Yes	Yes	180	Yes
Yes	Yes	No	210	No
Yes	Yes	???	???	No

Among the people that do not have Heart Disease, “No” is the most common value for Blocked arteries - it occurs in 2 out of 2 samples.

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	Yes	No	???	No

No

← So “No” is our initial guess.



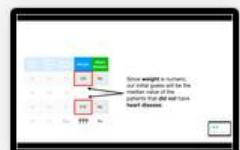
Chest Pain	Good Blood Circ.	Blocked Arteries	Weight	Heart Disease
No	No	No	125	No
Yes	Yes	Yes	180	Yes
Yes	Yes	No	210	No
Yes	Yes	No	???	No

Since **weight** is numeric, our initial guess will be the median value of the patients that **did not** have heart disease.



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	Yes	No	167.5	No

In this case, the median value is **167.5**.



Filled-in Missing Values

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	Yes	No	167.5	No

Here's our new dataset
with the filled in missing
values...



Filled-in Missing Values

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	Yes	No	167.5	No

Now we want to refine these guesses.

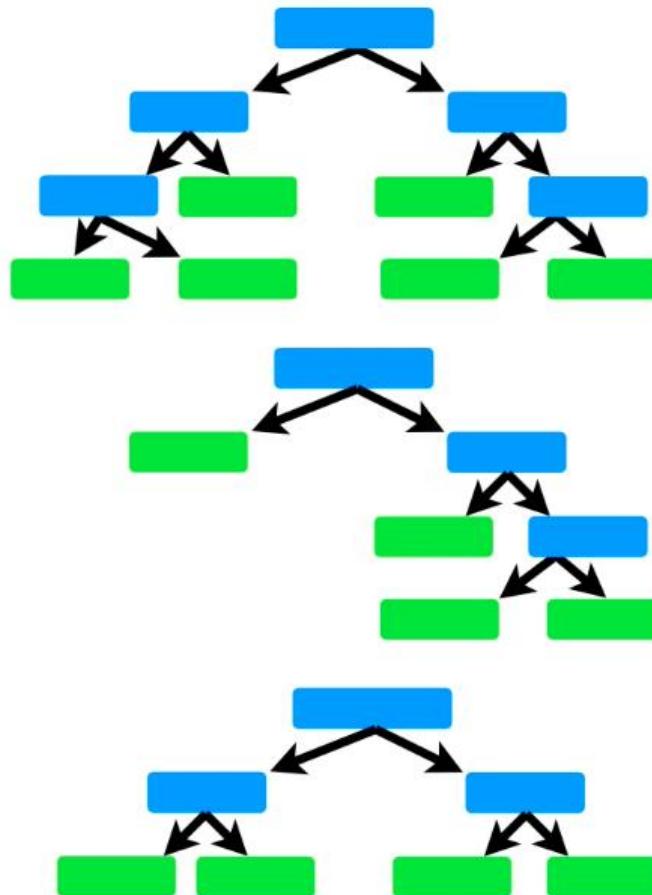
We do this by first determining which samples are similar to the one with missing data.

So let's talk about how to determine similarity...

Step 1: Build a random forest...

Filled-in Missing Values

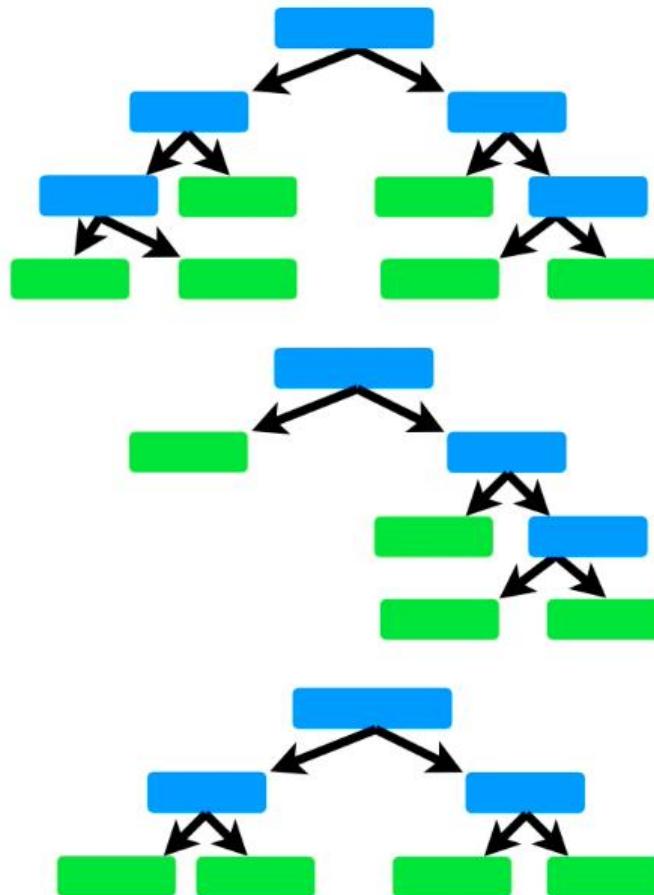
Chest Pain	Good Blood Circ.	Blocked Arteries	Weight	Heart Disease
No	No	No	125	No
Yes	Yes	Yes	180	Yes
Yes	Yes	No	210	No
Yes	Yes	No	167.5	No



Step 2: Run all of the data down all of the trees.

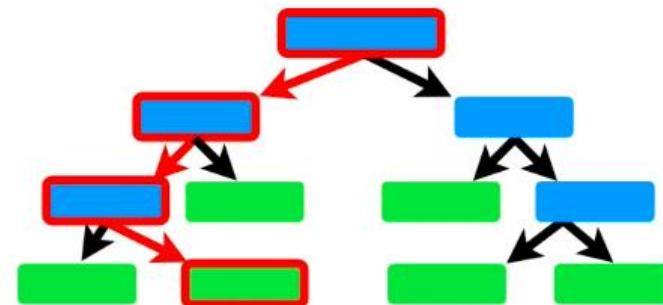
Filled-in Missing Values

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	Yes	No	167.5	No



Filled-in Missing Values

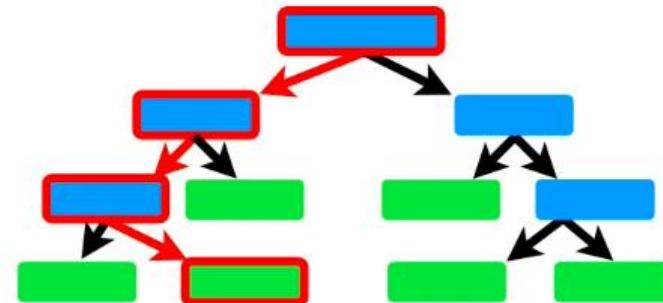
Chest Pain	Good Blood Circ.	Blocked Arteries	Weight	Heart Disease
No	No	No	125	No
Yes	Yes	Yes	180	Yes
Yes	Yes	No	210	No
Yes	Yes	No	167.5	No



Notice that Sample 3 and Sample 4 both ended up at the same leaf node.

Filled-in Missing Values

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	Yes	No	167.5	No



That means they are similar.

Filled-in Missing Values

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	Yes	No	167.5	No

We keep track of similar samples using a “Proximity Matrix”

	1	2	3	4
1				
2				
3				
4				

Filled-in Missing Values

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	Yes	No	167.5	No

The proximity matrix has a row for each sample..

	1	2	3	4
1				
2				
3				
4				

Filled-in Missing Values

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	Yes	No	167.5	No

...and it has a column for each sample.

	1	2	3	4
1				
2				
3				
4				

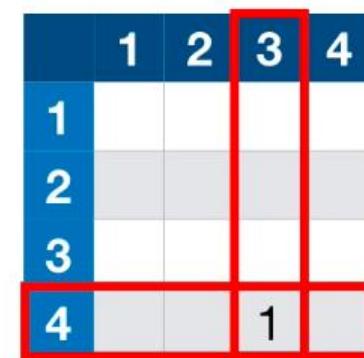


Filled-in Missing Values

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	Yes	No	167.5	No

Because sample 3...

...and sample 4
ended up in the
same leaf
node...



...we put a 1
here.

Filled-in Missing Values

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	Yes	No	167.5	No



	1	2	3	4
1				
2				
3				1
4			1	

We also put a 1 here, since this position also represents samples 3 and 4.

Filled-in Missing Values

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	Yes	No	167.5	No

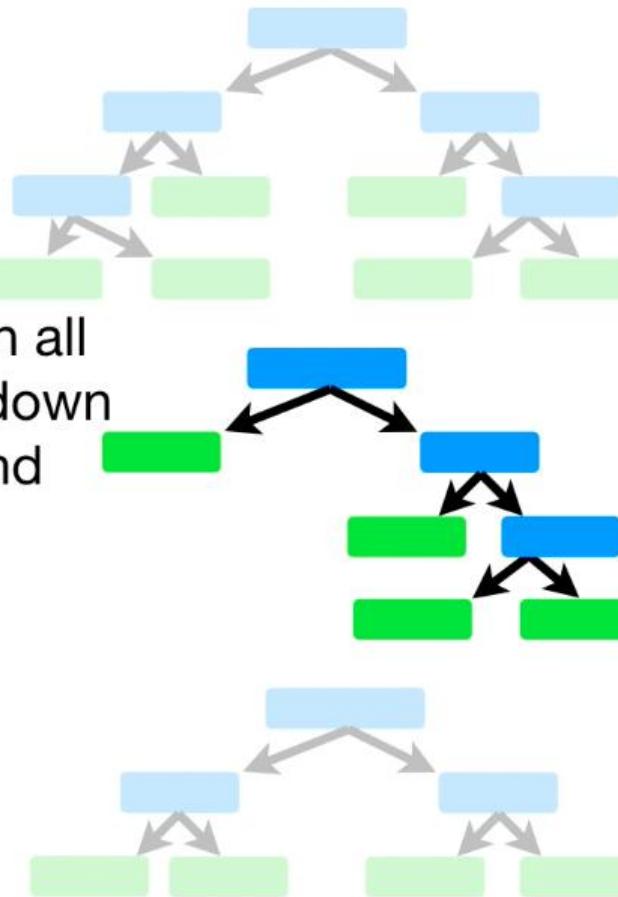
Because no other pair of samples ended in the same leaf node, our proximity matrix looks like this after running the samples down the first tree.

	1	2	3	4
1				
2				
3				1
4			1	

Filled-in Missing Values

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	Yes	No	167.5	No

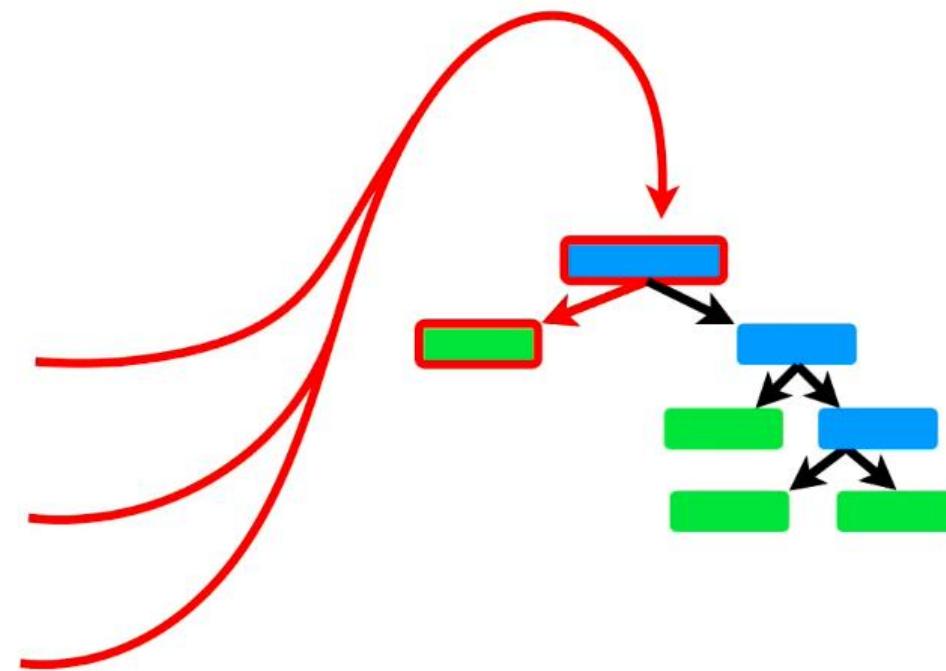
Now we run all
of the data down
the second
tree...



Filled-in Missing Values

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	Yes	No	167.5	No

NOTE: Samples 2, 3 and 4 all ended up in the same leaf node.



Filled-in Missing Values

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	Yes	No	167.5	No

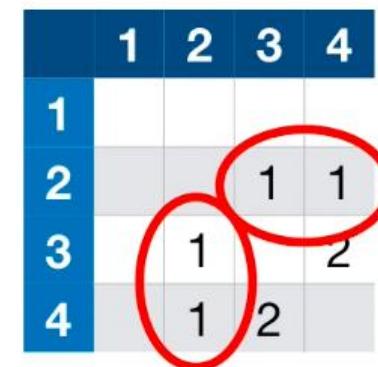
...after the second tree, we add 1 to any pair of samples that ended up in the same leaf node.

	1	2	3	4
1				
2			1	1
3		1		2
4	1	2		

Filled-in Missing Values

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	Yes	No	167.5	No

...after the second tree, we add 1 to any pair of samples that ended up in the same leaf node.

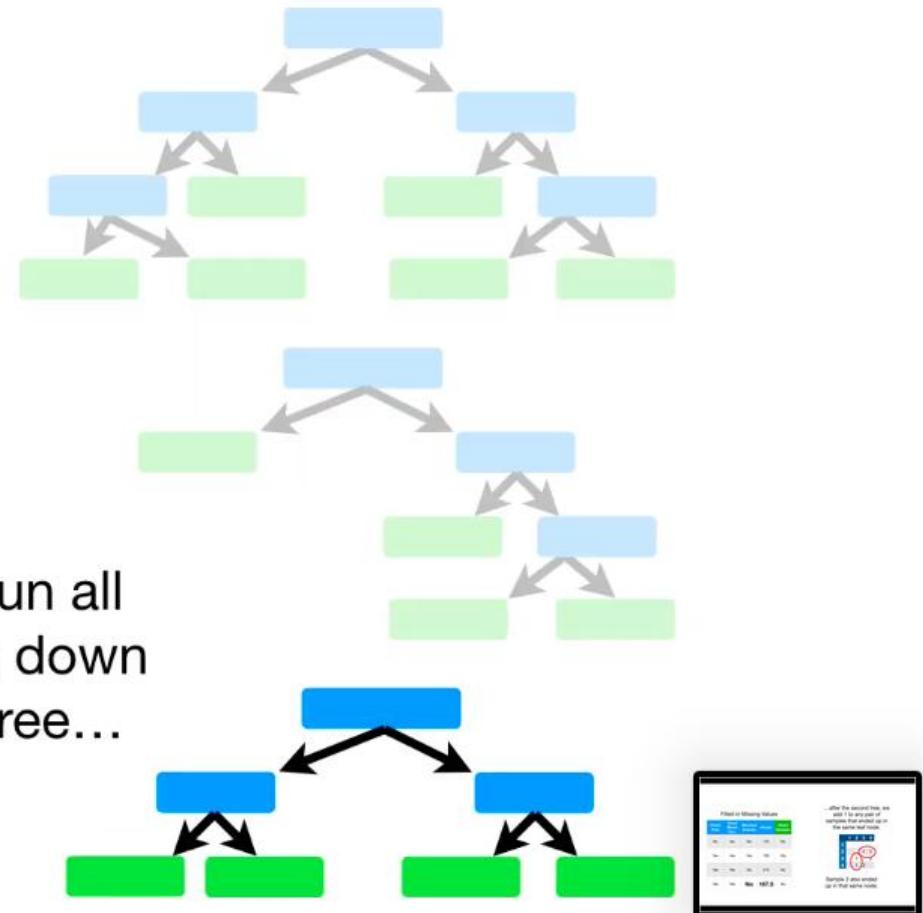


Sample 2 also ended up in that same node.

Filled-in Missing Values

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	Yes	No	180	No

Now we run all
of the data down
the third tree...



Filled-in Missing Values

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	Yes	No	167.5	No

...and here's the updated proximity matrix.

	1	2	3	4
1				
2			1	1
3		1		3
4	1	3		

Only samples 3 and 4 ended up in the same leaf node.

Filled-in Missing Values

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	Yes	No	167.5	No

Ultimately, we run the data down all the trees and the proximity matrix fills in.

	1	2	3	4
1		2	1	1
2	2		1	1
3	1	1		8
4	1	1	8	

Filled-in Missing Values

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	Yes	No	167.5	No

	1	2	3	4
1	0.2	0.1	0.1	
2	0.2	0.1	0.1	
3	0.1	0.1	0.8	
4	0.1	0.1	0.8	

Then we divide each proximity value by the total number of trees. In this example, assume we had 10 trees.

Filled-in Missing Values

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	Yes	???	???	No

Now we use the proximity values for sample 4 to make better guesses about the missing data.

	1	2	3	4
1	0.2	0.1	0.1	
2	0.2		0.1	0.1
3	0.1	0.1		0.8
4	0.1	0.1	0.8	



Filled-in Missing Values

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	Yes	???	???	No

For Blocked Arteries, we calculate the weighted frequency of “Yes” and “No, using proximity values as the weights.

	1	2	3	4
1	0.2	0.1	0.1	
2	0.2		0.1	0.1
3	0.1	0.1		0.8
4	0.1	0.1	0.8	

Filled-in Missing Values

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	Yes	???	???	No

Yes = 1/3

No = 2/3

The frequency
of “No”.

	1	2	3	4
1	0.2	0.1	0.1	
2	0.2		0.1	0.1
3	0.1	0.1		0.8
4	0.1	0.1	0.8	

Filled-in Missing Values

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	Yes	???	???	No

The weighted frequency for “Yes” is...

$$\text{Yes} = \frac{1}{3} \times \text{The weight for “Yes”}$$

Yes = 1/3

No = 2/3

	1	2	3	4
1	0.2	0.1	0.1	
2	0.2		0.1	0.1
3	0.1	0.1		0.8
4	0.1	0.1	0.8	

The weight for “Yes” =

Proximity of “Yes”
All Proximities

Filled-in Missing Values

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	Yes	???	???	No

The weighted frequency for “Yes” is...

$$\text{Yes} = \frac{1}{3} \times \text{The weight for “Yes”}$$

Yes = 1/3

No = 2/3

	1	2	3	4
1		0.2	0.1	0.1
2	0.2		0.1	0.1
3	0.1	0.1		0.8
4	0.1	0.1	0.8	

The weight for “Yes” =

0.1

The proximity value for Sample 2 (the only one with “Yes”)

Filled-in Missing Values

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	Yes	???	???	No

The weighted frequency for “Yes” is...

$$\text{Yes} = \frac{1}{3} \times \text{The weight for “Yes”}$$

Yes = 1/3

No = 2/3

1	1	2	3	4
1		0.2	0.1	0.1
2	0.2		0.1	0.1
3	0.1	0.1		0.8
4	0.1	0.1	0.8	

The weight for “Yes” =

$$\frac{0.1}{0.1 + 0.1 + 0.8}$$

Divided by the sum of the proximities for Sample 4.

Filled-in Missing Values

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	Yes	???	???	No

The weighted frequency for “Yes” is...

$$\text{Yes} = \frac{1}{3} \times 0.1$$

Yes = 1/3

No = 2/3

1	1	2	3	4
1	0.2	0.1	0.1	
2	0.2		0.1	0.1
3	0.1	0.1		0.8
4	0.1	0.1	0.8	

The weight for

$$\text{“Yes”} = \frac{0.1}{0.1 + 0.1 + 0.8} = \frac{0.1}{1} = 0.1$$

Filled-in Missing Values

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	Yes	???	???	No

The weighted frequency for “No” is...

$$\text{Yes} = \frac{1}{3} \times 0.1 = 0.03$$

$$\text{No} = \frac{2}{3} \times \text{The weight for “No”}$$

Yes = 1/3
No = 2/3

	1	2	3	4
1		0.2	0.1	0.1
2	0.2		0.1	0.1
3	0.1	0.1		0.8
4	0.1	0.1	0.8	

The weight for “No” = $\frac{0.1 + 0.8}{0.1 + 0.1 + 0.8} = \frac{0.9}{1} = 0.9$

Filled-in Missing Values

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	Yes	???	???	No

The weighted frequency for “No” is...

$$\text{Yes} = \frac{1}{3} \times 0.1 = 0.03$$

$$\text{No} = \frac{2}{3} \times 0.9 = 0.6$$

$$\text{Yes} = 1/3$$

$$\text{No} = 2/3$$

	1	2	3	4
1	0.2	0.1	0.1	
2	0.2		0.1	0.1
3	0.1	0.1		0.8
4	0.1	0.1	0.8	

Filled-in Missing Values

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	Yes	NO	???	No

The weighted frequency for “No” is...

$$\text{Yes} = \frac{1}{3} \times 0.1 = 0.03$$

$$\text{No} = \frac{2}{3} \times 0.9 = 0.6$$

	1	2	3	4
1	0.2	0.1	0.1	
2	0.2		0.1	0.1
3	0.1	0.1		0.8
4	0.1	0.1	0.8	

$$\text{Yes} = 1/3$$

$$\text{No} = 2/3$$

“No” has a way higher weighted frequency, so we’ll go with it.



Filled-in Missing Values

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	Yes	NO	???	No

	1	2	3	4
1		0.2	0.1	0.1
2	0.2		0.1	0.1
3	0.1	0.1		0.8
4	0.1	0.1	0.8	

For weight, we use the proximities to calculate a weighted average.

Sample 1's
 Weighted average = $125 \times$ weighted average
 weight...

Filled-in Missing Values

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	Yes	NO	???	No

	1	2	3	4
1	0.2	0.1	0.1	
2	0.2		0.1	0.1
3	0.1	0.1		0.8
4	0.1	0.1	0.8	

Sample 1's
 Weighted average = $125 \times$ weighted average
 weight...

$$\frac{0.1}{0.1 + 0.1 + 0.8}$$

Divided by
 the sum of
 the
 proximities.

	1	2	3	4
1	0.2	0.1	0.1	
2	0.2		0.1	0.1
3	0.1	0.1		0.8
4	0.1	0.1	0.8	

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	Yes	NO	???	No

Weighted average = 125×0.1

$$\frac{0.1}{0.1 + 0.1 + 0.8} = \frac{0.1}{1} = 0.1$$

Filled-in Missing Values

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	Yes	NO	???	No

	1	2	3	4
1		0.2	0.1	0.1
2	0.2		0.1	0.1
3	0.1	0.1		0.8
4	0.1	0.1	0.8	

$$\text{Weighted average} = (125 \times 0.1) + (180 \times 0.1)$$

Filled-in Missing Values

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	Yes	NO	???	No

	1	2	3	4
1	0.2	0.1	0.1	
2	0.2		0.1	0.1
3	0.1	0.1		0.8
4	0.1	0.1	0.8	

...the weighted value for 180...

$$\text{Weighted average} = (125 \times 0.1) + (180 \times 0.1) + (210 \times 0.8)$$

$$= 198.5$$

Filled-in Missing Values

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	Yes	NO	???	No

	1	2	3	4
1	0.2	0.1	0.1	
2	0.2	0.1	0.1	
3	0.1	0.1	0.8	
4	0.1	0.1	0.8	

$$\text{Weighted average} = (125 \times 0.1) + (180 \times 0.1) + (210 \times 0.8)$$

$$= 198.5$$

Filled-in Missing Values

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	Yes	NO	198.5	No

	1	2	3	4
1	0.2	0.1	0.1	
2	0.2	0.1	0.1	
3	0.1	0.1	0.8	
4	0.1	0.1	0.8	

The weighted average weight!

Filled-in Missing Values

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	Yes	NO	198.5	No

Now that we've revised our guesses a little bit, we do the whole thing over again...

We build a random forest, run the data through the trees, recalculate the proximities and recalculate the missing values.

We do this 6 or 7 times until the missing values converge (i.e. no longer change each time we recalculate).

Now it's time for an interlude of awesomeness!!!

Let me show you something super cool
we can do with the proximity matrix!!!

	1	2	3	4
1		0.2	0.1	0.1
2	0.2		0.1	0.1
3	0.1	0.1		0.8
4	0.1	0.1	0.8	

This is the proximity matrix
before we divided each value by
10, the number of trees in the
pretend random forest.

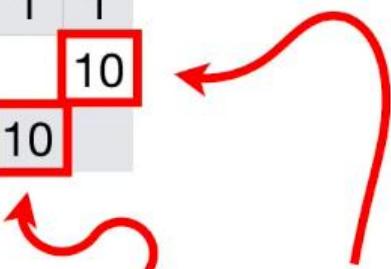
	1	2	3	4
1		2	1	1
2	2		1	1
3	1	1		8
4	1	1	8	

Just for the sake of easy math,
imagine if Samples 3 and 4
ended up in the same leaf node
in all 10 trees.

	1	2	3	4
1		2	1	1
2	2		1	1
3	1	1		10
4	1	1	10	

Just for the sake of easy math,
imagine if Samples 3 and 4
ended up in the same leaf node
in all 10 trees.

	1	2	3	4
1		2	1	1
2	2		1	1
3	1	1		10
4	1	1	10	



Now we have 10 here and here...

After dividing by 10 (the number of trees in the forest), we see that the largest number in the proximity matrix is 1.

	1	2	3	4
1		2	1	1
2	2		1	1
3	1	1		10
4	1	1	10	

	1	2	3	4
1		0.2	0.1	0.1
2	0.2		0.1	0.1
3	0.1	0.1		1
4	0.1	0.1	1	

1 in the proximity matrix means the samples are as close as close can be.

	1	2	3	4
1		2	1	1
2	2		1	1
3	1	1		10
4	1	1	10	

	1	2	3	4
1		0.2	0.1	0.1
2	0.2		0.1	0.1
3	0.1	0.1		1
4	0.1	0.1	1	

That means...

1 - the proximity values
= distance

	1	2	3	4
1		2	1	1
2	2		1	1
3	1	1		10
4	1	1	10	

	1	2	3	4
1		0.2	0.1	0.1
2	0.2		0.1	0.1
3	0.1	0.1		1
4	0.1	0.1	1	

	1	2	3	4
1		0.8	0.9	0.9
2	0.8		0.9	0.9
3	0.9	0.9		0
4	0.9	0.9	0	

That means...

1 - the proximity values
= distance

	1	2	3	4
1		2	1	1
2	2		1	1
3	1	1		10
4	1	1	10	

	1	2	3	4
1		0.2	0.1	0.1
2	0.2		0.1	0.1
3	0.1	0.1		1
4	0.1	0.1	1	

	1	2	3	4
1		0.8	0.9	0.9
2	0.8		0.9	0.9
3	0.9	0.9		0
4	0.9	0.9	0	

Close as can be = no distance between.

That means...

1 - the proximity values
= distance

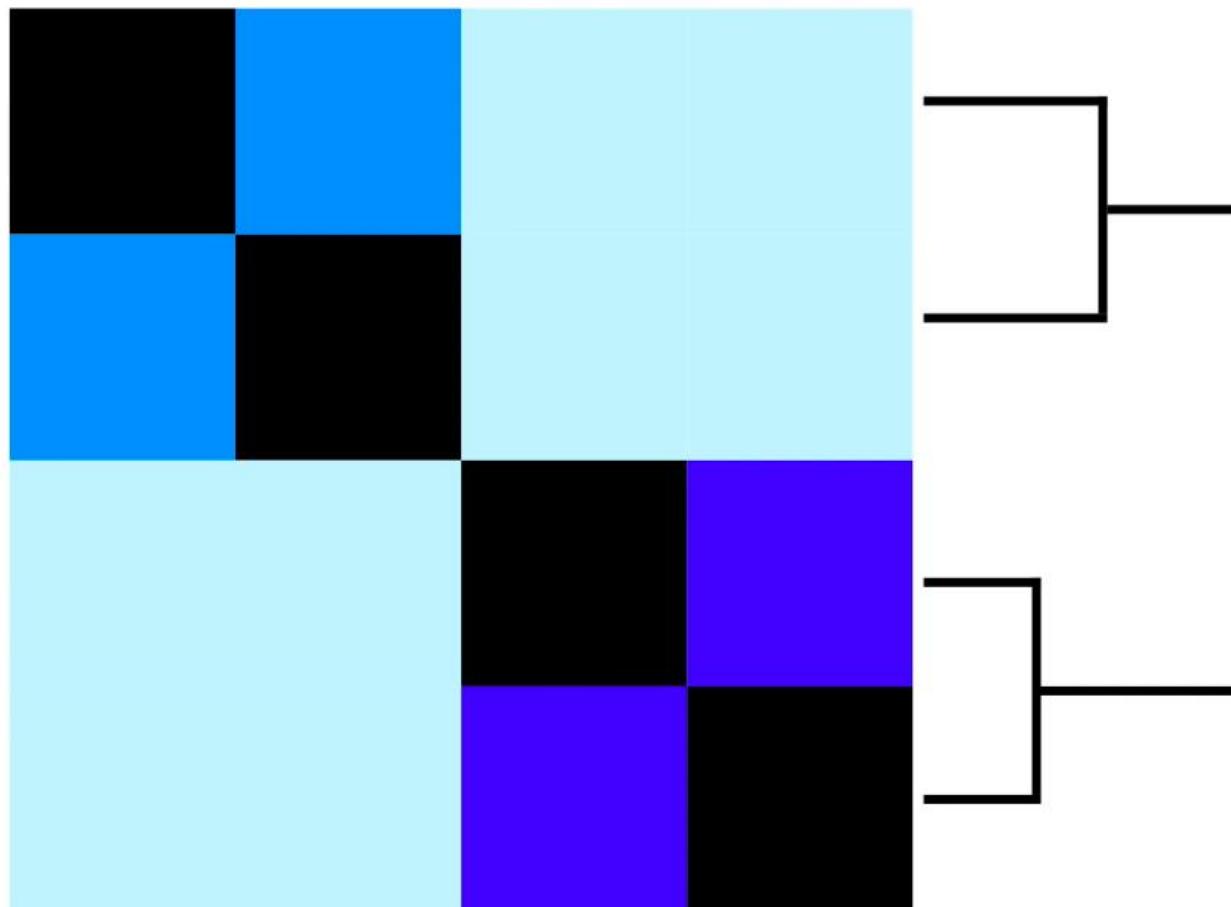
	1	2	3	4
1		2	1	1
2	2		1	1
3	1	1		10
4	1	1	10	

	1	2	3	4
1		0.2	0.1	0.1
2	0.2		0.1	0.1
3	0.1	0.1		1
4	0.1	0.1	1	

	1	2	3	4
1		0.8	0.9	0.9
2	0.8		0.9	0.9
3	0.9	0.9		0
4	0.9	0.9	0	

Not close = far away

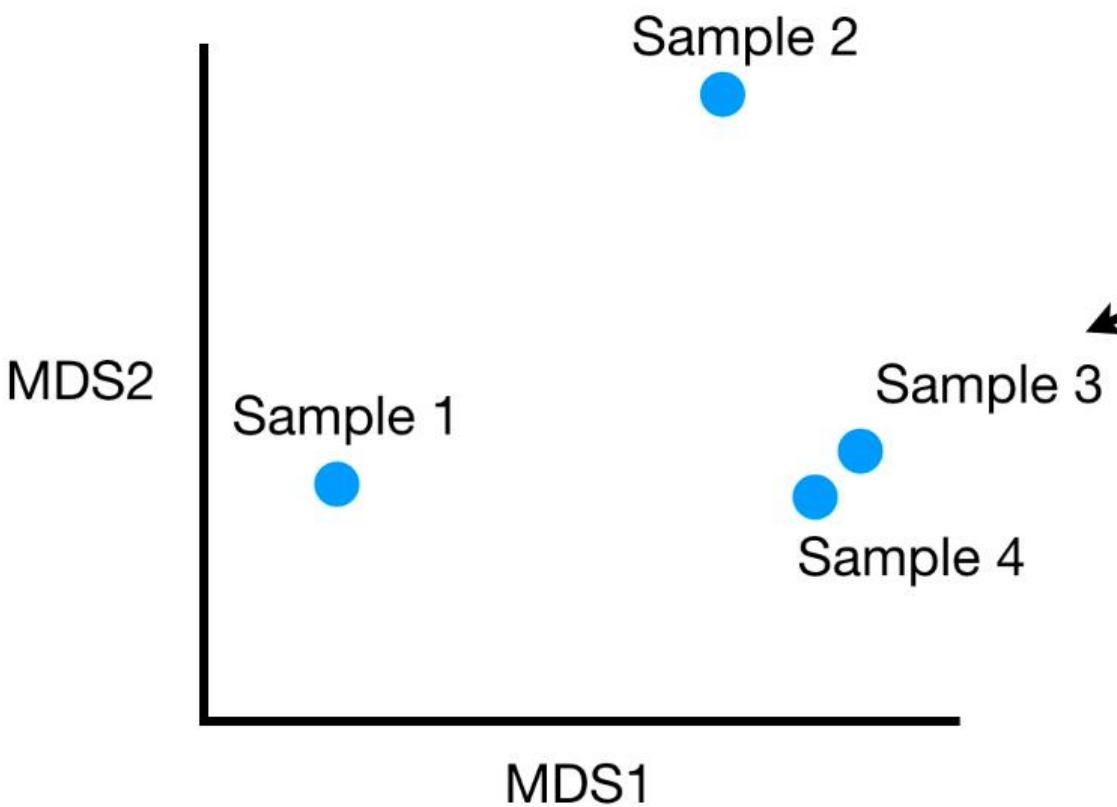
Sample 1 Sample 2 Sample 3 Sample 4



This is a distance matrix...

...and that means we can draw a heatmap with it!!!

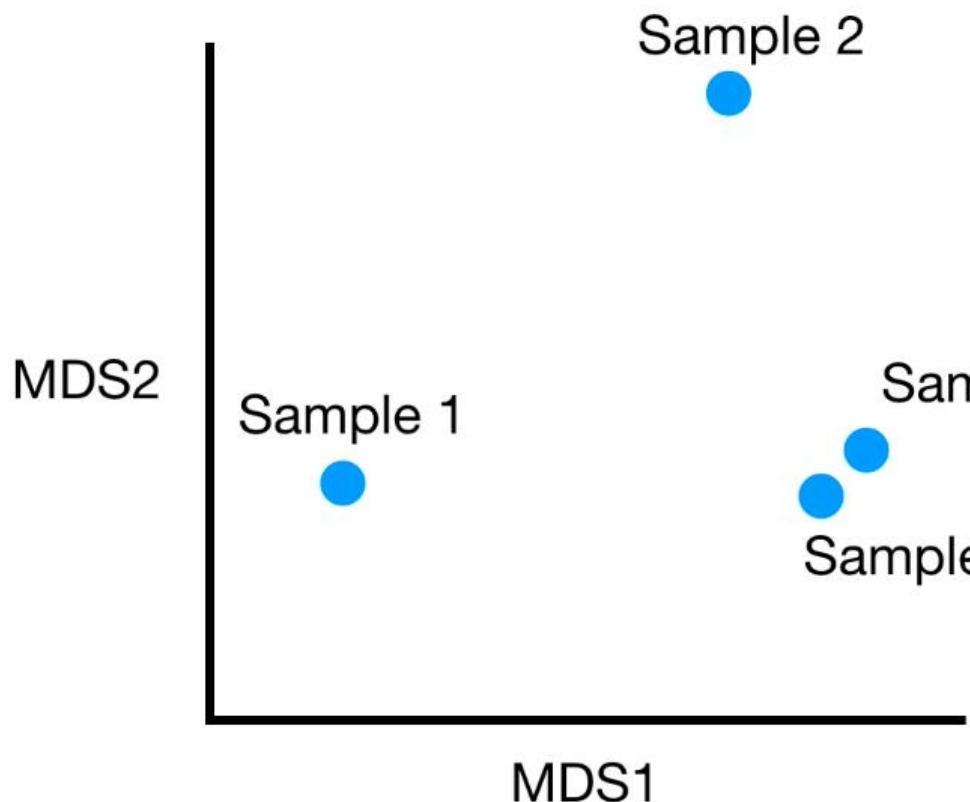
	1	2	3	4
1	0.8	0.9	0.9	
2	0.8		0.9	0.9
3	0.9	0.9		0
4	0.9	0.9	0	



This is a distance matrix...

...and we can also draw an MDS plot with it!

	1	2	3	4
1	0.8	0.9	0.9	
2	0.8		0.9	0.9
3	0.9	0.9		0
4	0.9	0.9	0	



This is super cool because it means that no matter what the data are (ranks, multiple choice, numeric, etc)... if we can use it to make a tree, we can draw a heatmap or an MDS plot to show how the samples are related to each other!!!

	1	2	3	4
1	0.8	0.9	0.9	
2	0.8		0.9	0.9
3	0.9	0.9		0
4	0.9	0.9	0	

Enough fun stuff! Back to missing data!!!

Random Forests consider 2 types of missing data...

- 1) Missing data in the original dataset used to create the random forest.
- 2) Missing data in a new sample that you want to categorize.

At long last, we'll talk about the second method!

New Sample

Chest Pain	Good Blood Circ.	Blocked Arteries	Weight	Heart Disease
No	No	No	???	

Imagine we had already built a Random Forest with existing data and wanted to classify this new patient.

Chest Pain	Good Blood Circ.	Blocked Arteries	Weight	Heart Disease
Yes	No	???	168	

Chest Pain	Good Blood Circ.	Blocked Arteries	Weight	Heart Disease
Yes	No	???	168	

...but we don't know if
they have blocked
arteries...

Chest Pain	Good Blood Circ.	Blocked Arteries	Weight	Heart Disease
Yes	No	???	168	

...so we need to make a guess about Blocked Arteries so we can run the patient down all the trees in the forest.



Chest Pain	Good Blood Circ.	Blocked Arteries	Weight	Heart Disease
Yes	No	???	168	YES

The first thing we do is create two copies of the data, one that has heart disease...

Chest Pain	Good Blood Circ.	Blocked Arteries	Weight	Heart Disease
Yes	No	???	168	NO

Chest Pain	Good Blood Circ.	Blocked Arteries	Weight	Heart Disease
Yes	No	???	168	YES

...and one that doesn't have heart disease.

Chest Pain	Good Blood Circ.	Blocked Arteries	Weight	Heart Disease
Yes	No	???	168	NO



Then we use the iterative method we just talked about to make a good guess about the missing values.

Chest Pain	Good Blood Circ.	Blocked Arteries	Weight	Heart Disease
Yes	No	???	168	YES

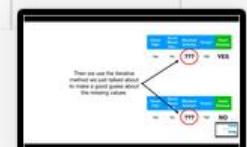
Chest Pain	Good Blood Circ.	Blocked Arteries	Weight	Heart Disease
Yes	No	???	168	NO

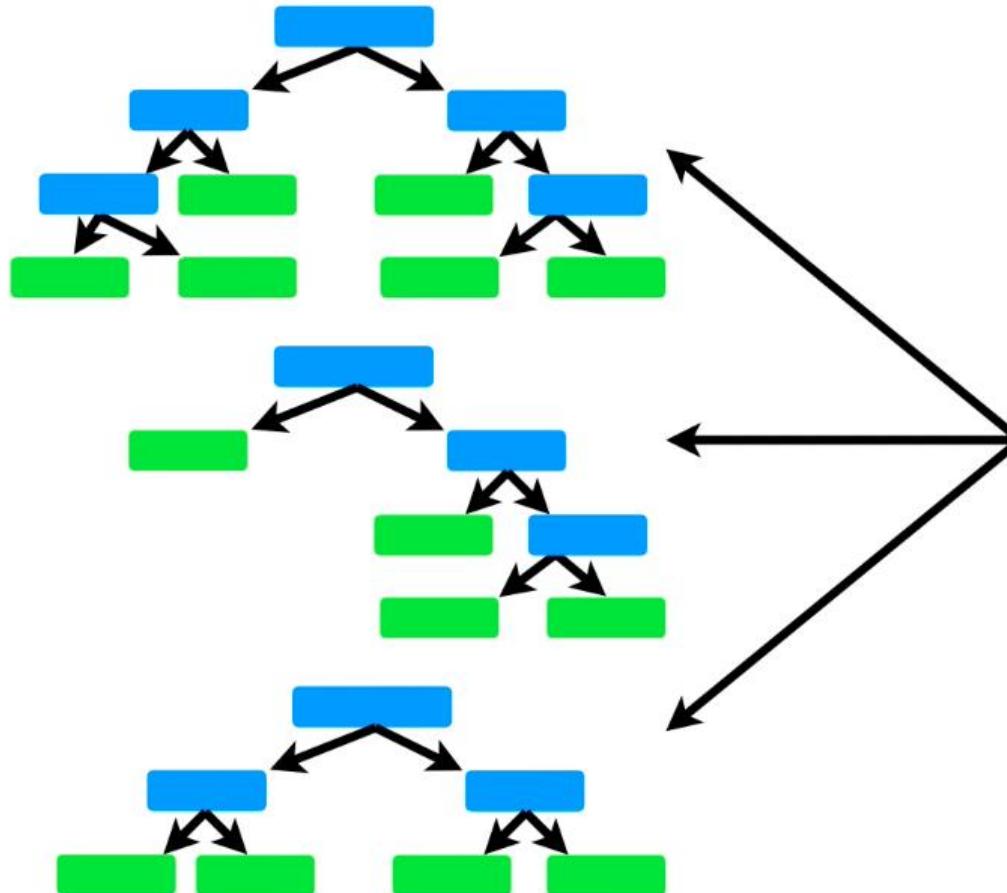


Then we use the iterative method we just talked about to make a good guess about the missing values.

Chest Pain	Good Blood Circ.	Blocked Arteries	Weight	Heart Disease
Yes	No	YES	168	YES

Chest Pain	Good Blood Circ.	Blocked Arteries	Weight	Heart Disease
Yes	No	NO	168	NO

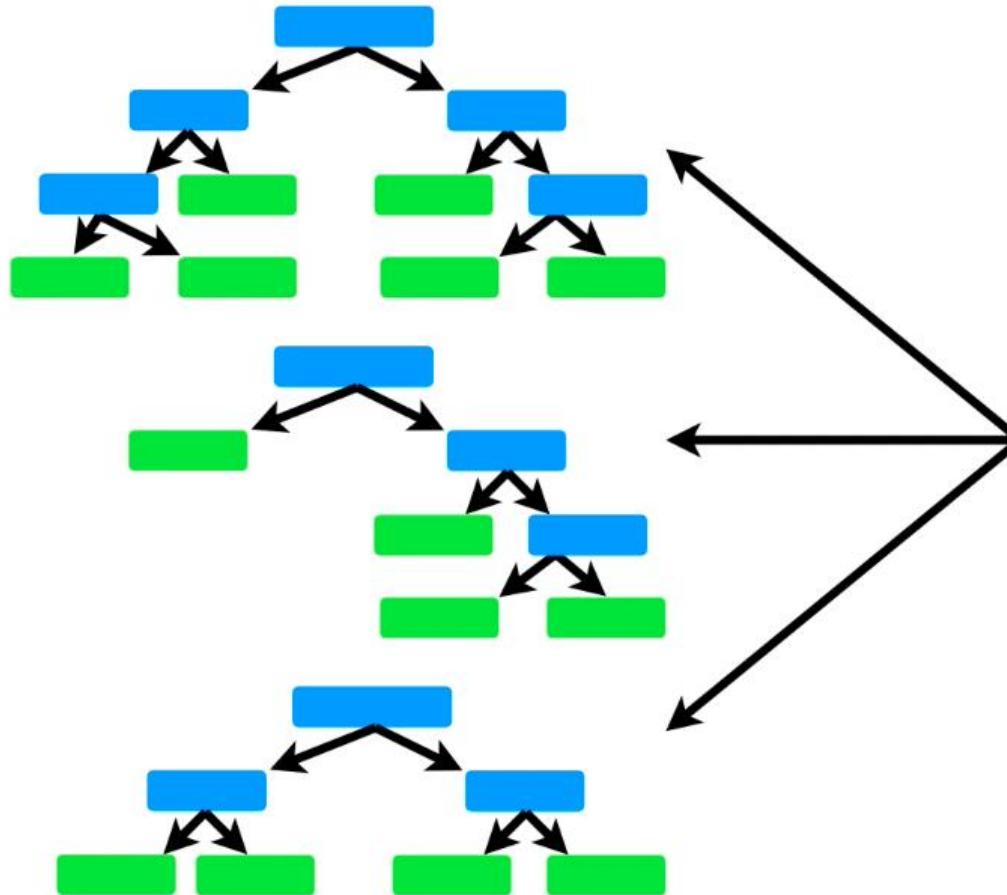




Chest Pain	Good Blood Circ.	Blocked Arteries	Weight	Heart Disease
Yes	No	YES	168	YES

Then we run the two samples down the trees in the forest...

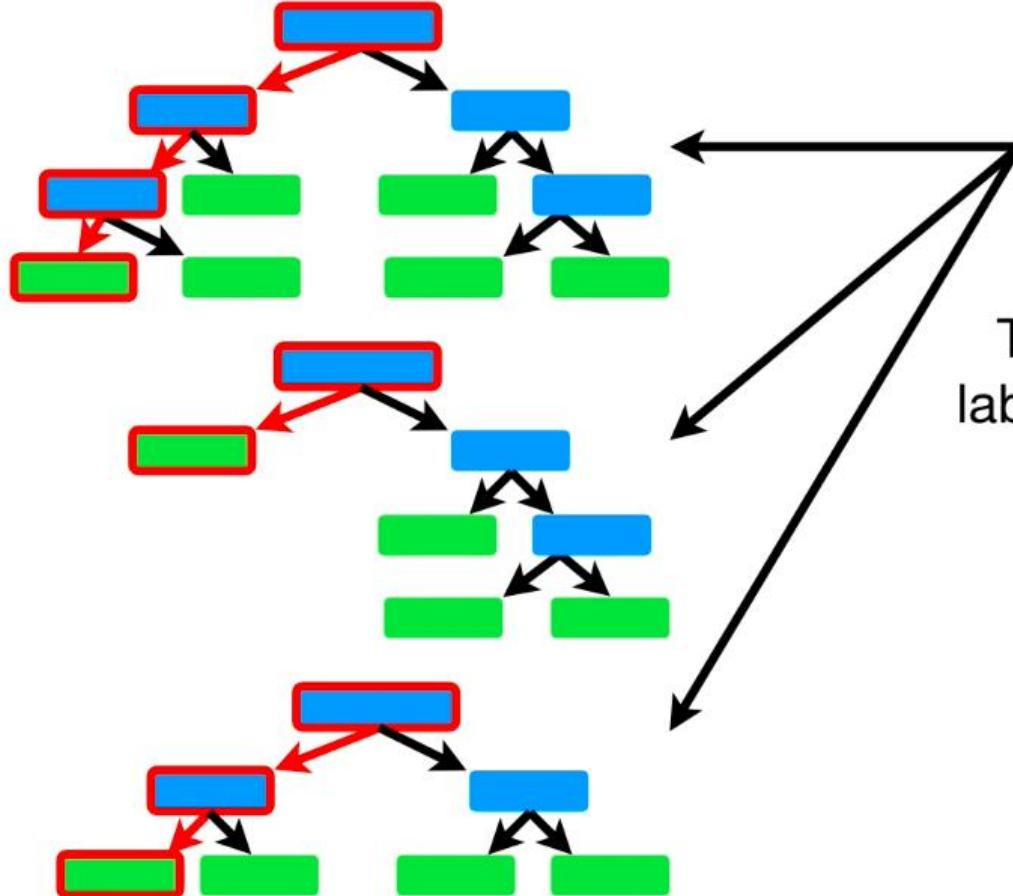
Chest Pain	Good Blood Circ.	Blocked Arteries	Weight	Heart Disease
Yes	No	NO	168	NO



Chest Pain	Good Blood Circ.	Blocked Arteries	Weight	Heart Disease
Yes	No	YES	168	YES

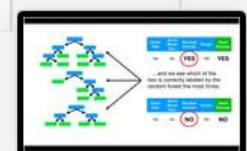
...and we see which of the two is correctly labeled by the random forest the most times.

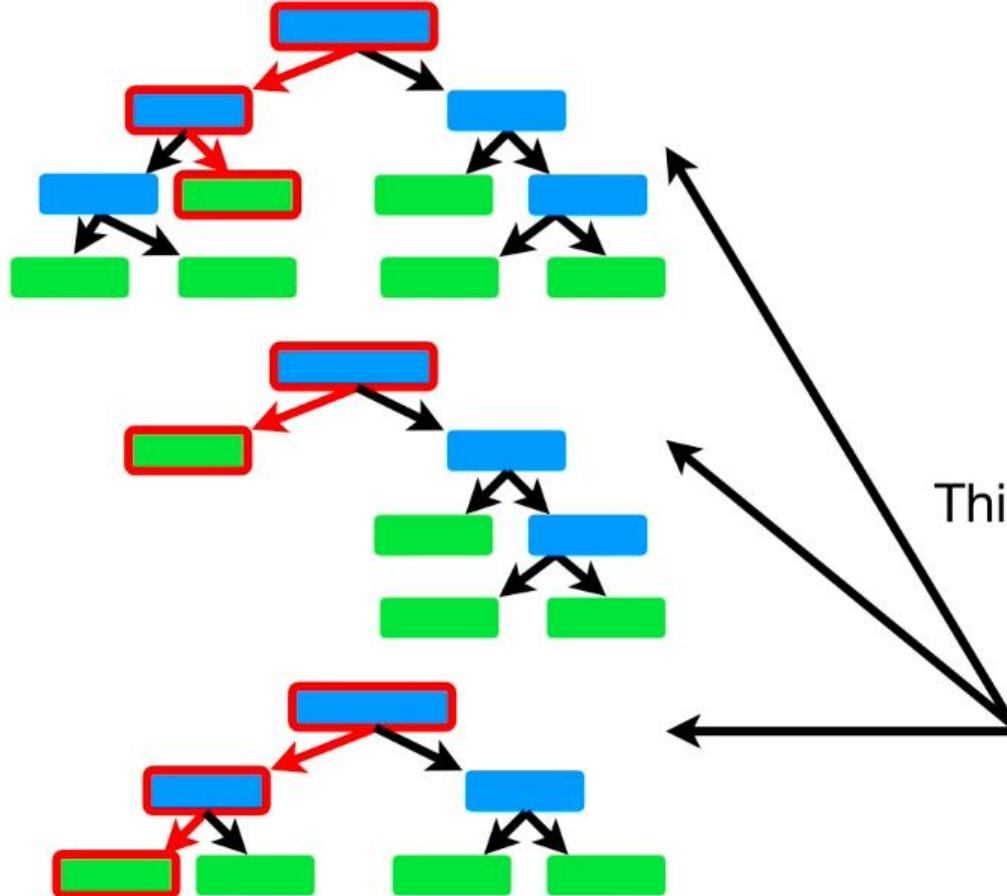
Chest Pain	Good Blood Circ.	Blocked Arteries	Weight	Heart Disease
Yes	No	NO	168	NO



Chest Pain	Good Blood Circ.	Blocked Arteries	Weight	Heart Disease
Yes	No	YES	168	YES

Chest Pain	Good Blood Circ.	Blocked Arteries	Weight	Heart Disease
Yes	No	NO	168	NO

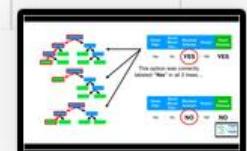


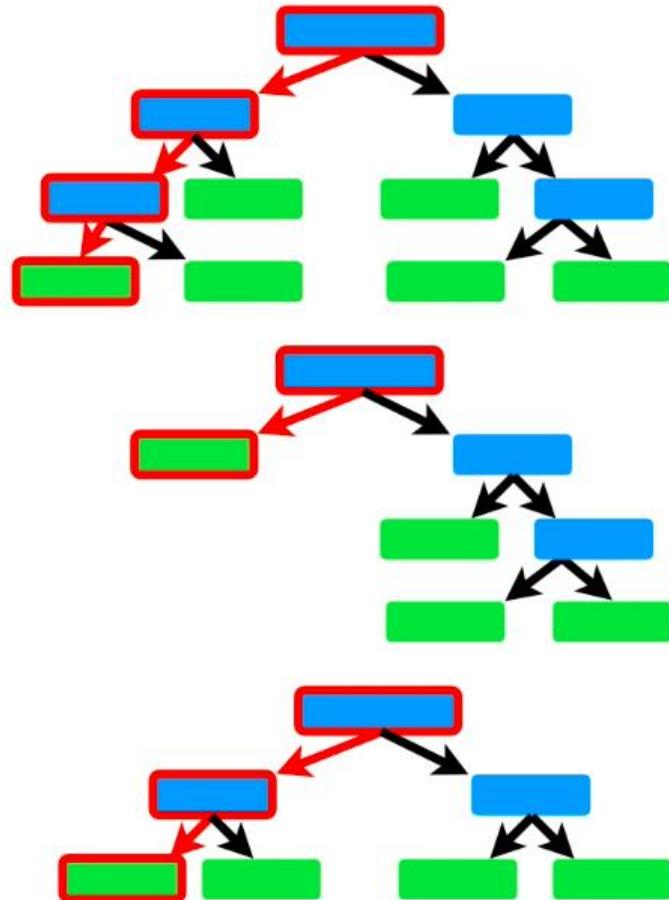


This option was only correctly labeled “**No**” in 1 tree...

Chest Pain	Good Blood Circ.	Blocked Arteries	Weight	Heart Disease
Yes	No	YES	168	YES

Chest Pain	Good Blood Circ.	Blocked Arteries	Weight	Heart Disease
Yes	No	NO	168	NO





Chest Pain	Good Blood Circ.	Blocked Arteries	Weight	Heart Disease
Yes	No	YES	168	YES

This option wins because it was correctly labeled more than the other option.

Chest Pain	Good Blood Circ.	Blocked Arteries	Weight	Heart Disease
Yes	No	NO	168	NO

