Machine Learning

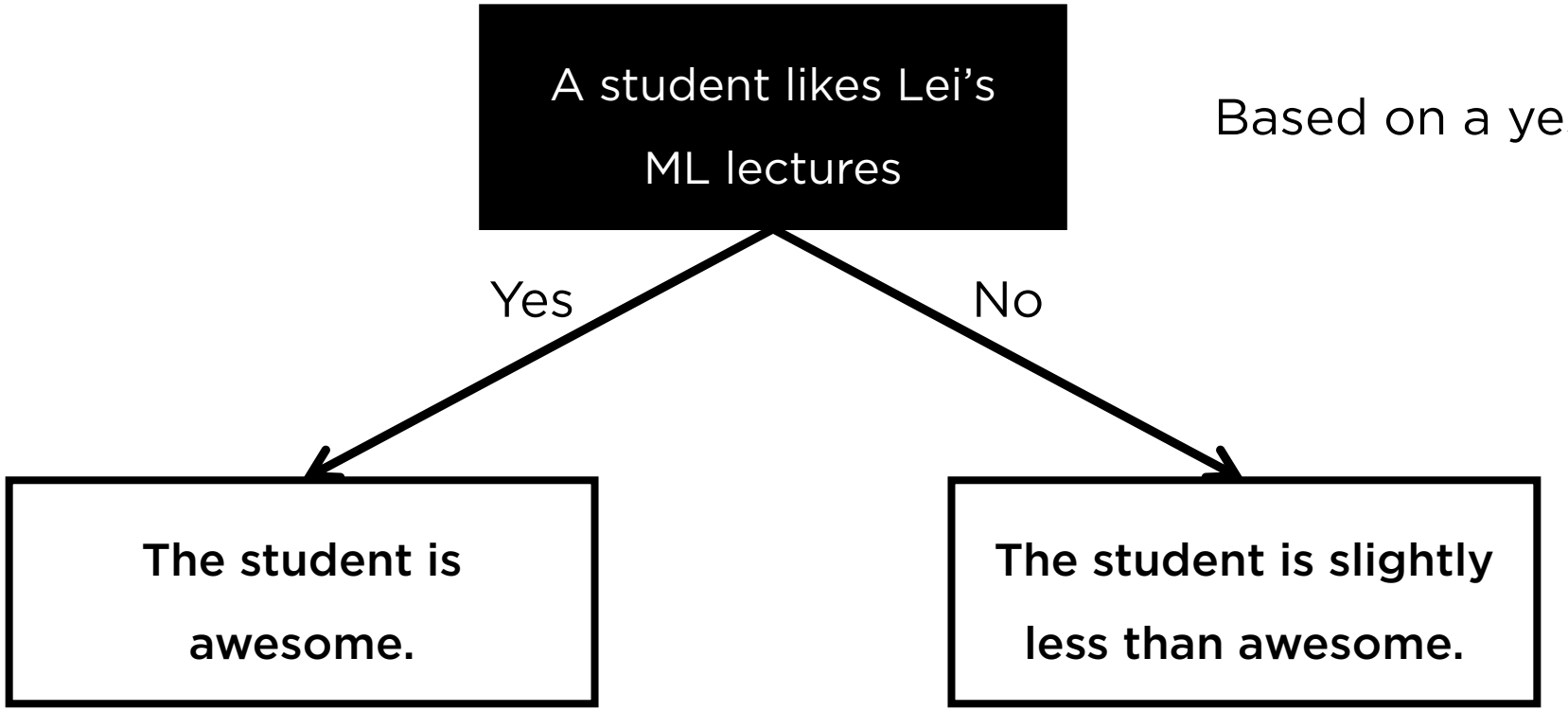
Lecture 6 – Decision Trees and Random Forests

Dr SHI Lei

Decision Trees

Decision Trees

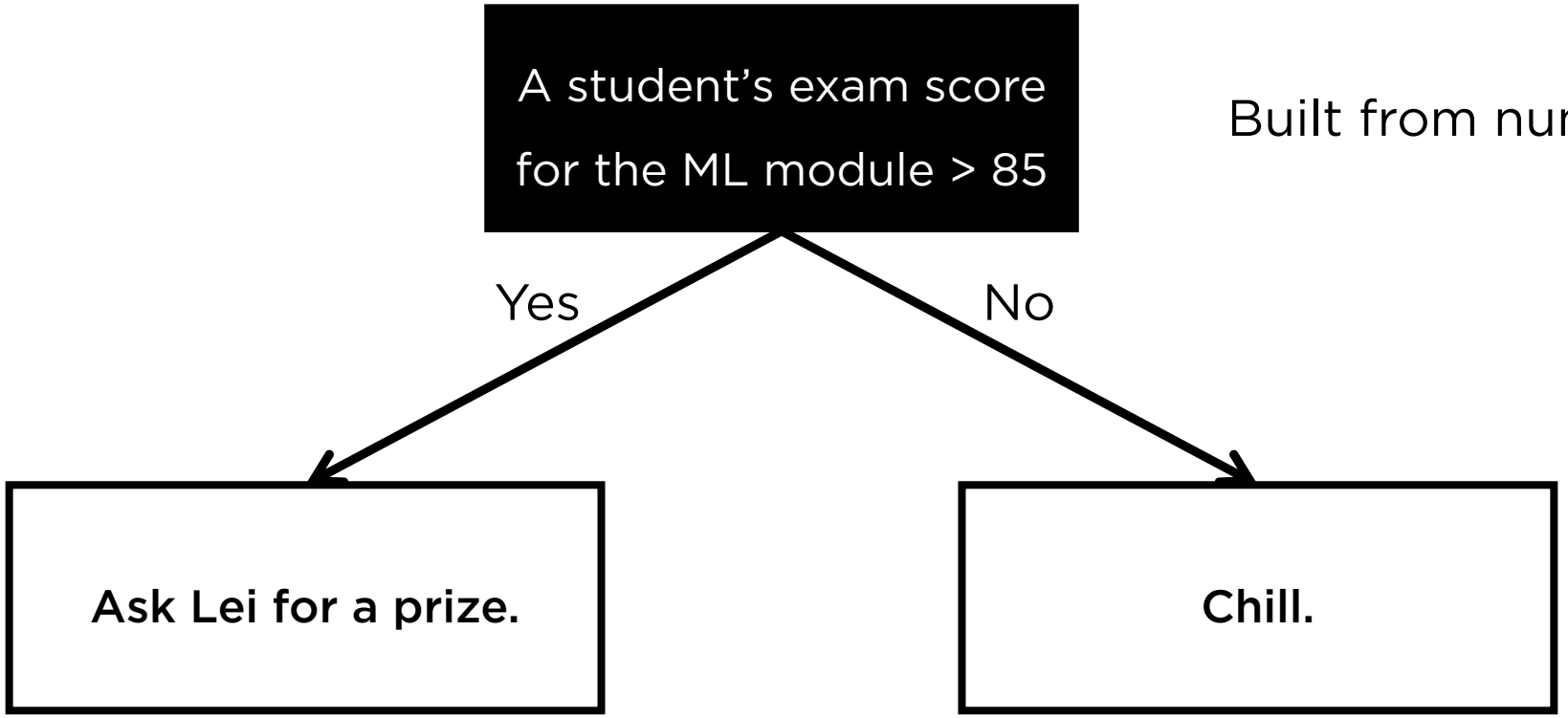
- An example



Based on a yes/no question

Decision Trees

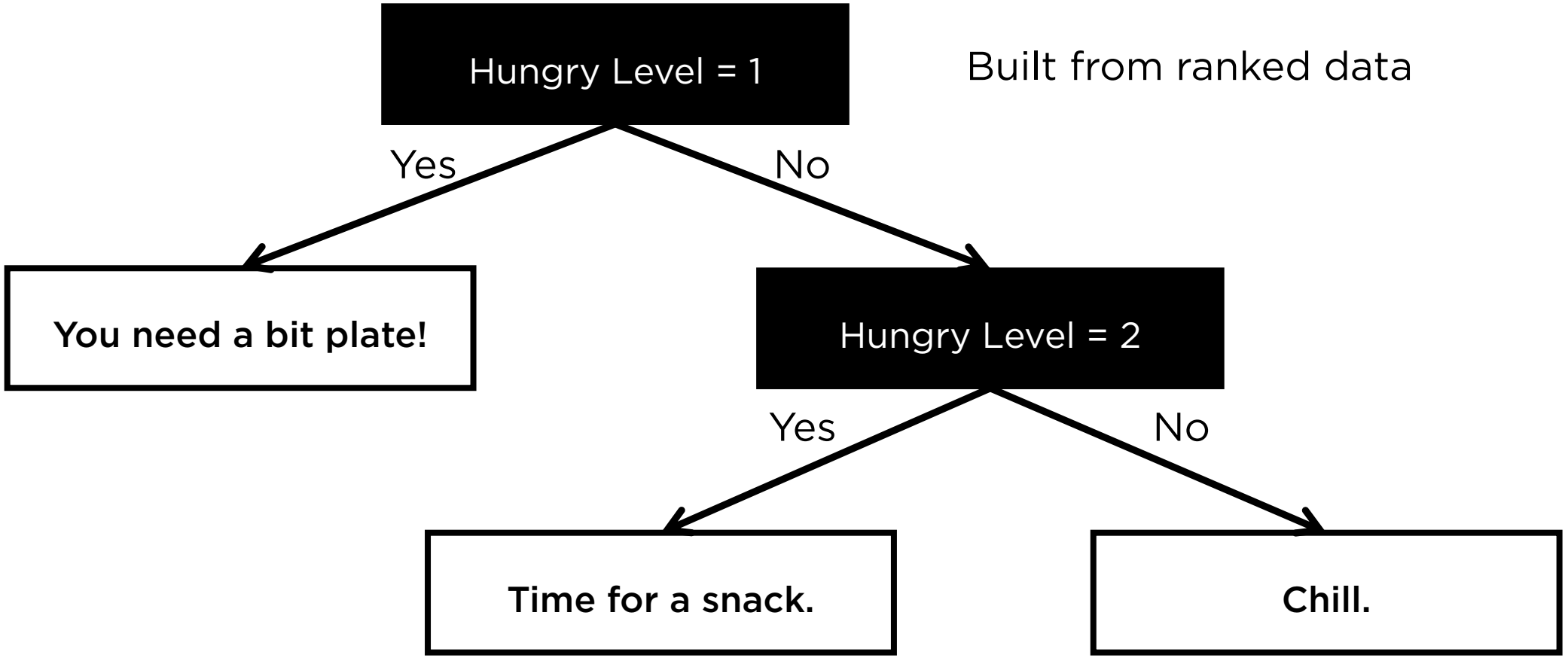
- Another example



Built from numeric data

Decision Trees

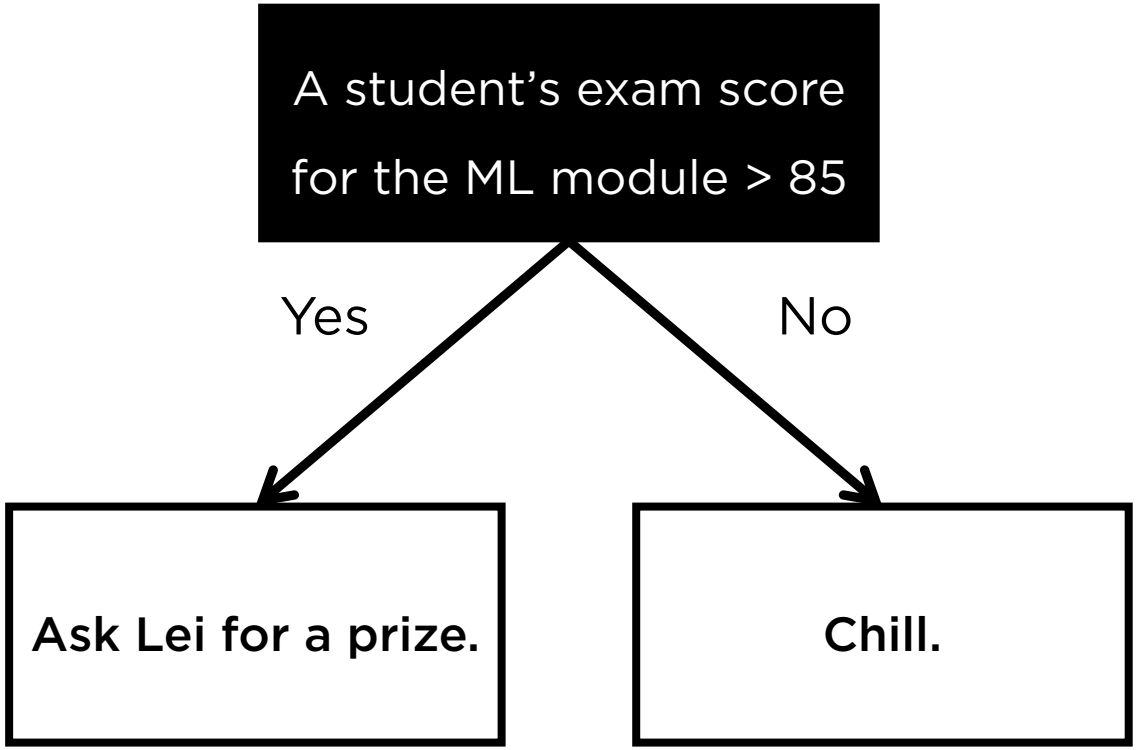
- One more example

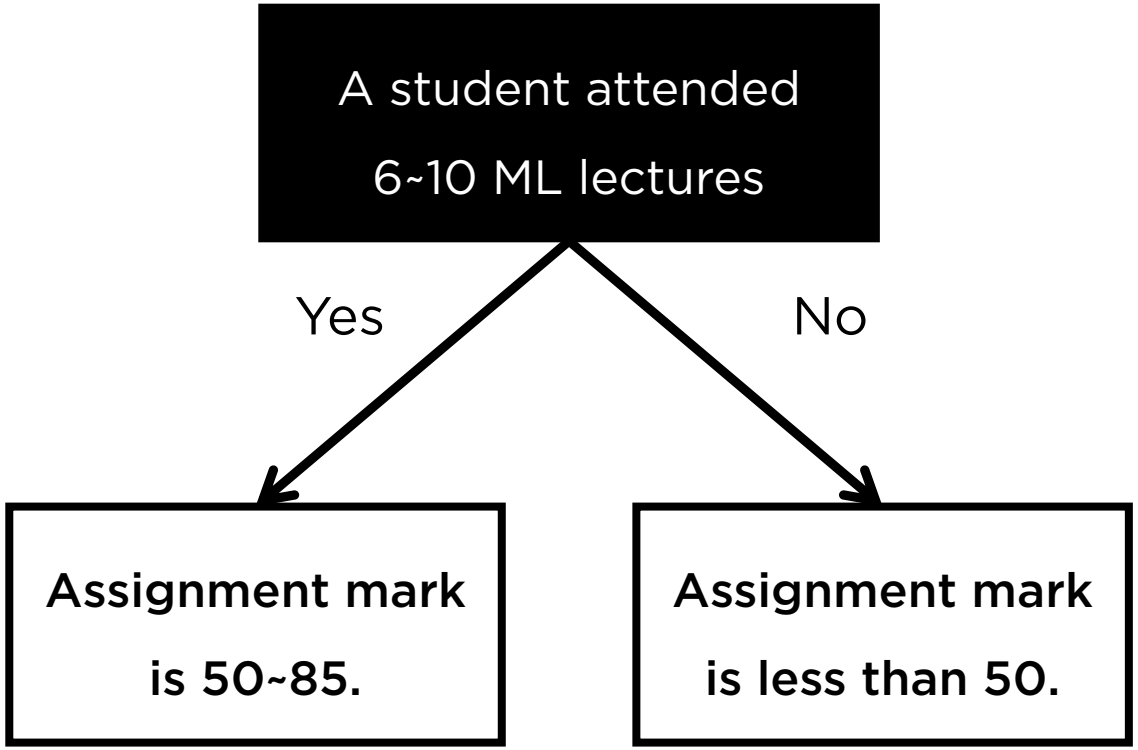


Built from ranked data

Decision Trees

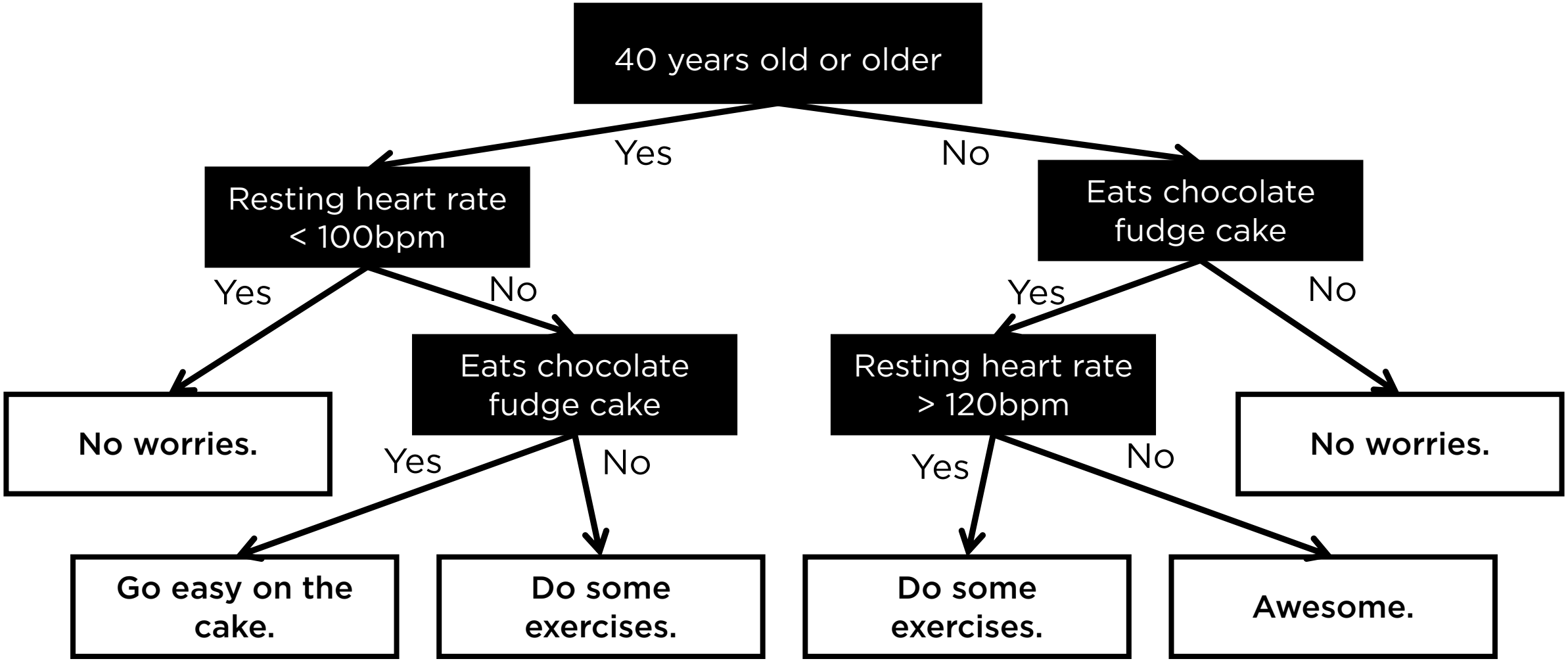
- The classification can be **Categorical** or **Numerical**





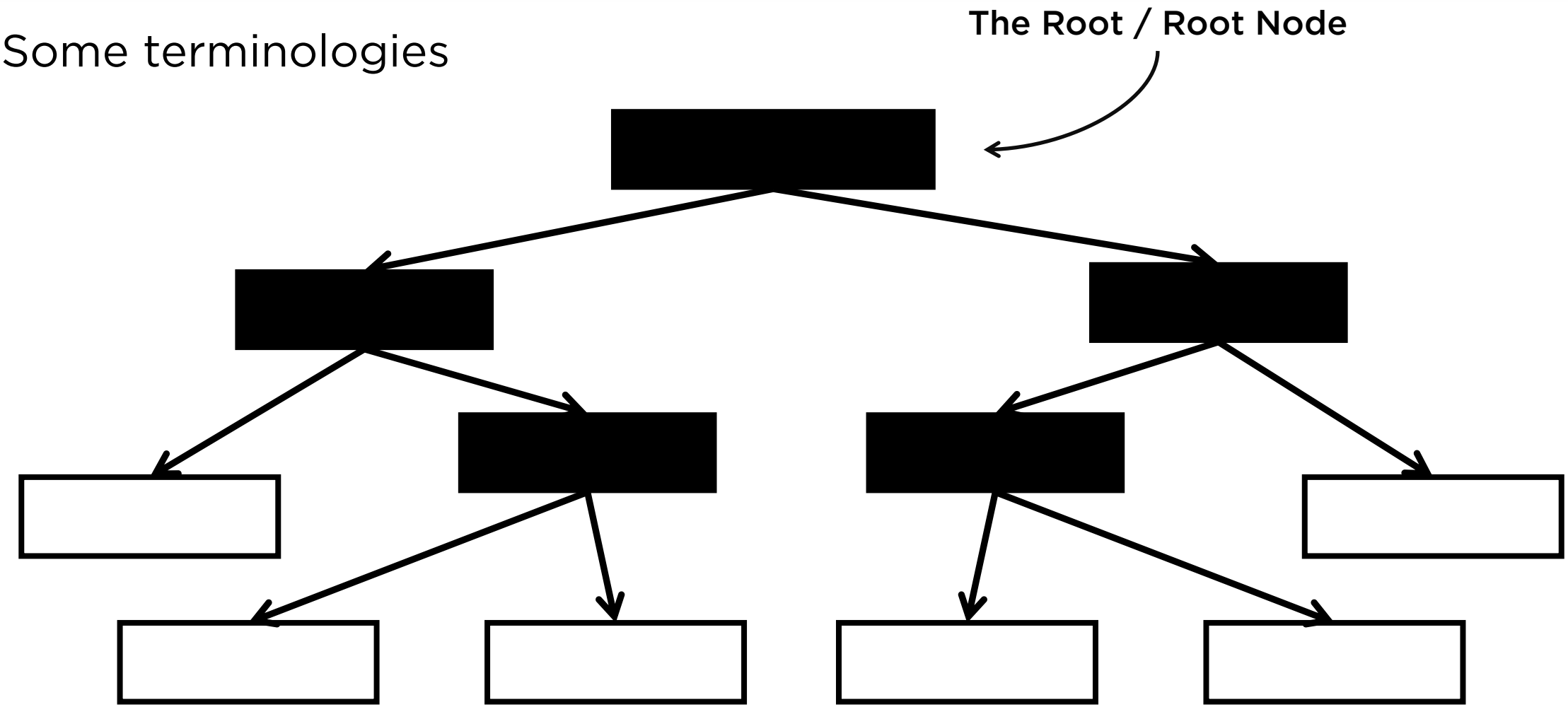
Decision Trees

- A more complicated decision tree



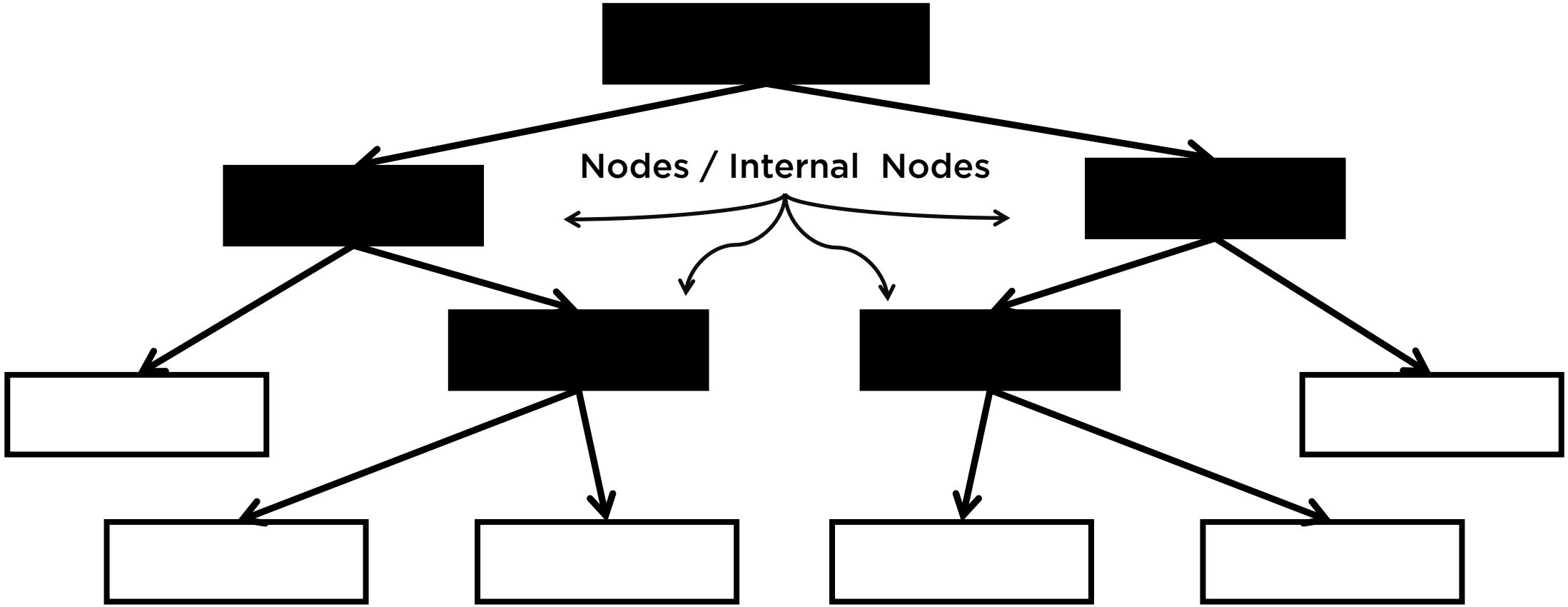
Decision Trees

- Some terminologies



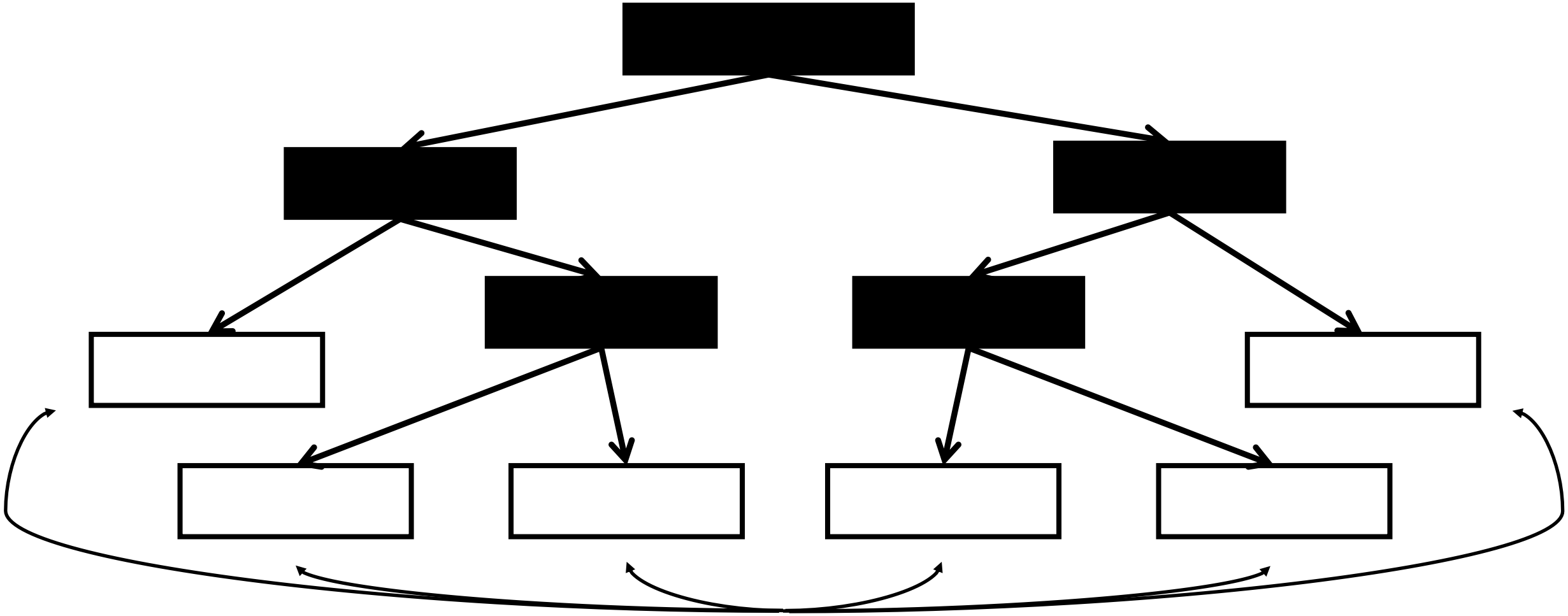
Decision Trees

- Some terminologies



Decision Trees

- Some terminologies

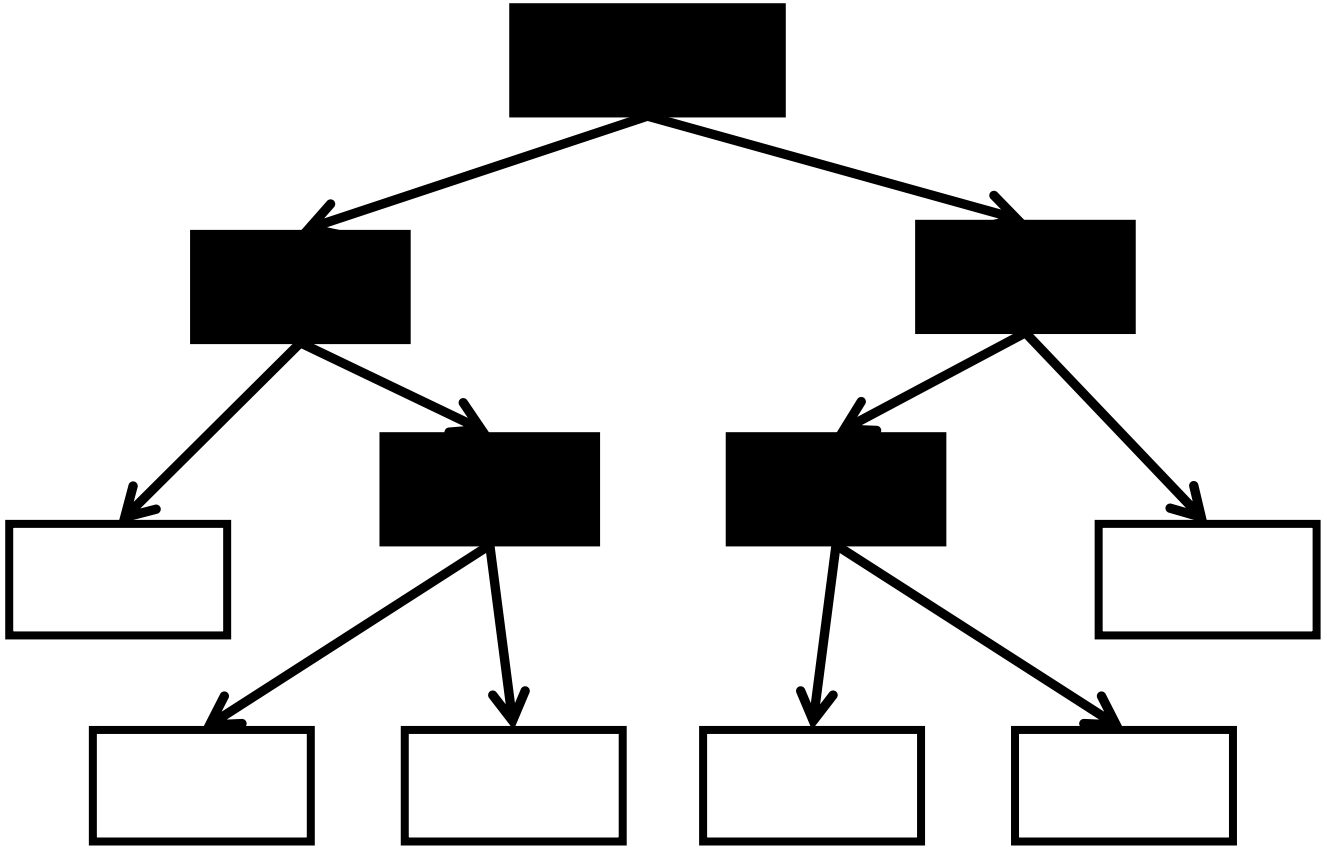


Leaves / Leaf Nodes

Decision Trees

- From a raw table of data to a decision tree

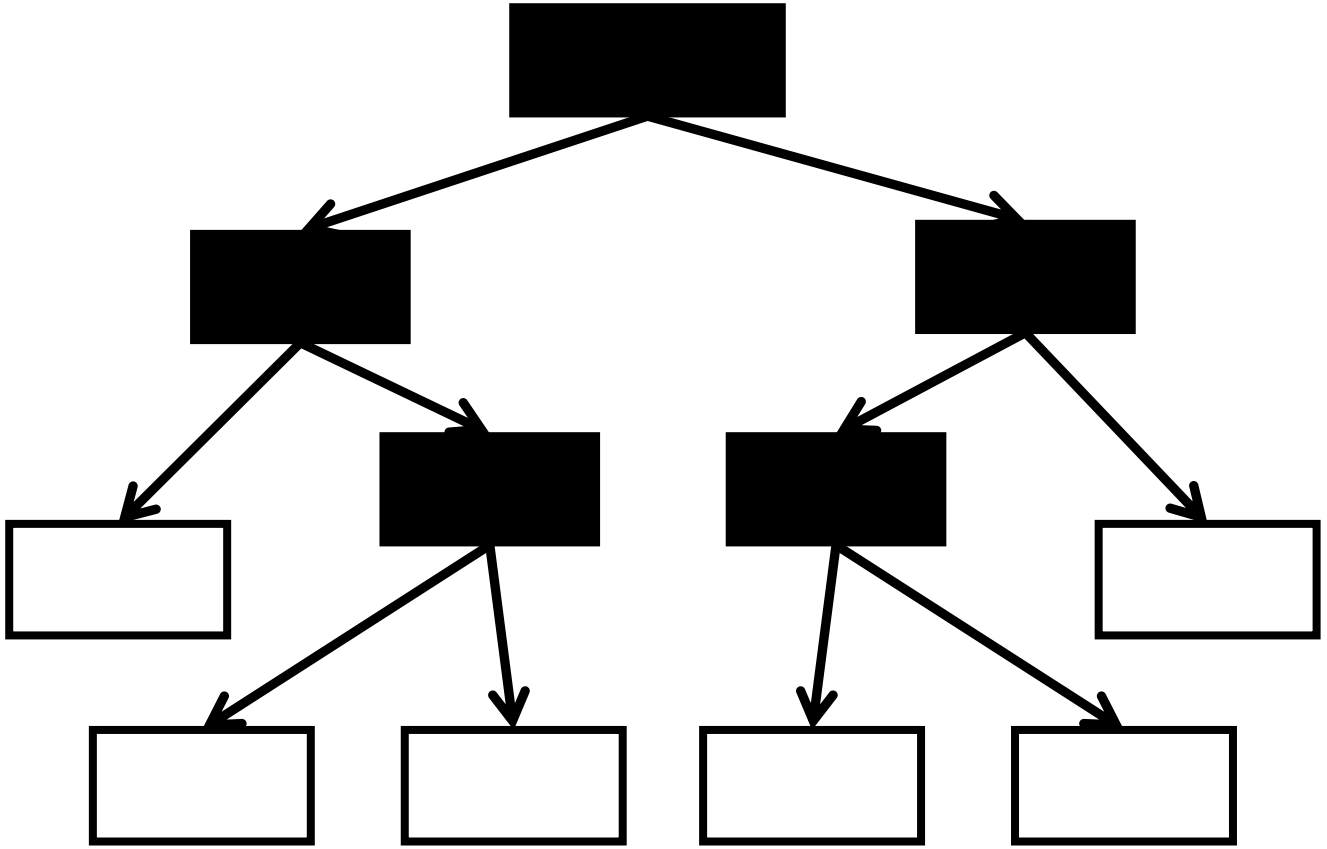
Chest pain	Good blood circulation	Blocked arteries	Heart disease
No	No	No	No
Yes	Yes	Yes	Yes
Yes	Yes	No	No
Yes	No	?	Yes
...	...	...	...



Decision Trees

- From a raw table of data to a decision tree

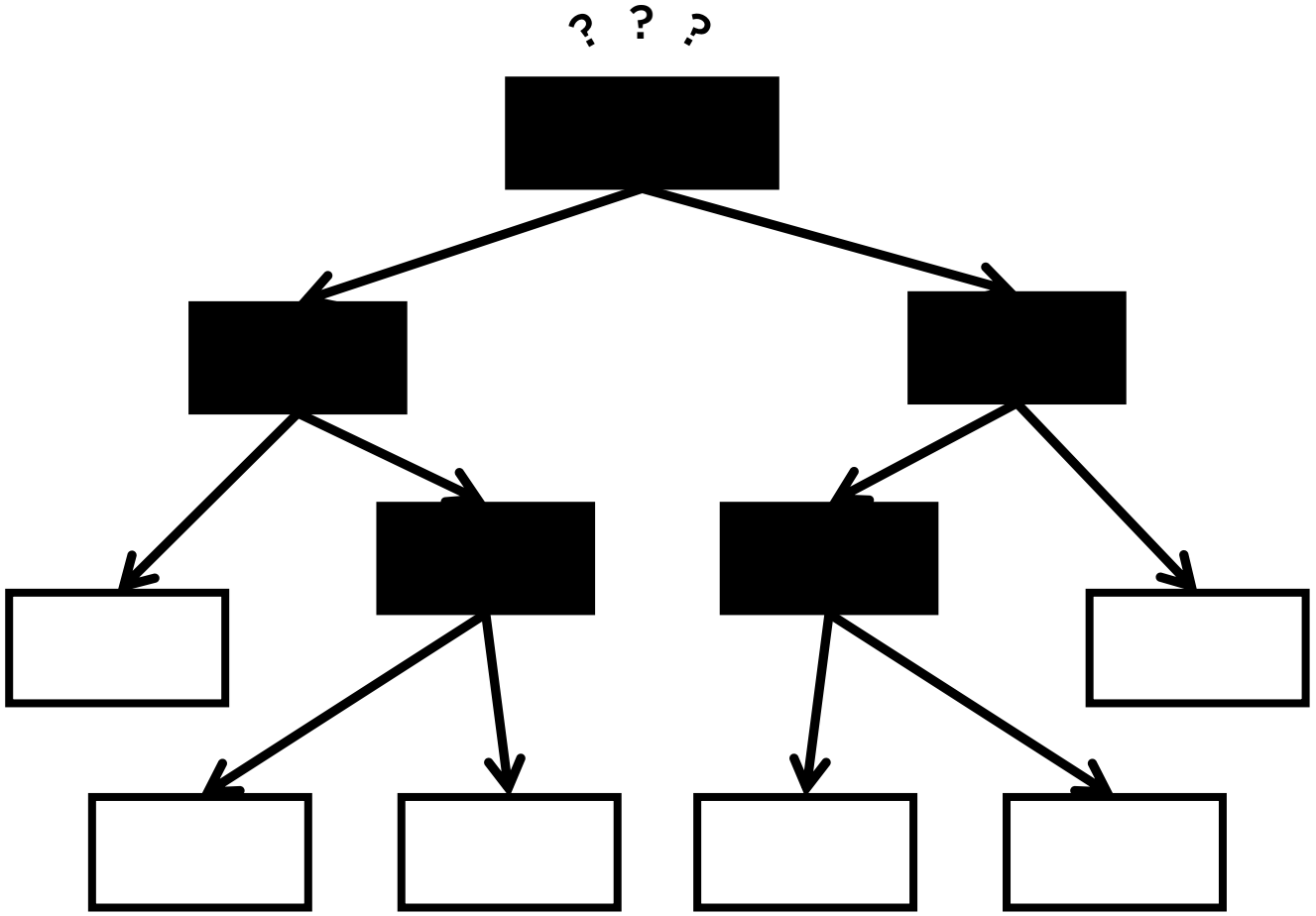
Chest pain	Good blood circulation	Blocked arteries	Heart disease
No	No	No	No
Yes	Yes	Yes	Yes
Yes	Yes	No	No
Yes	No	?	Yes
...	...	...	...



Decision Trees

- From a raw table of data to a decision tree

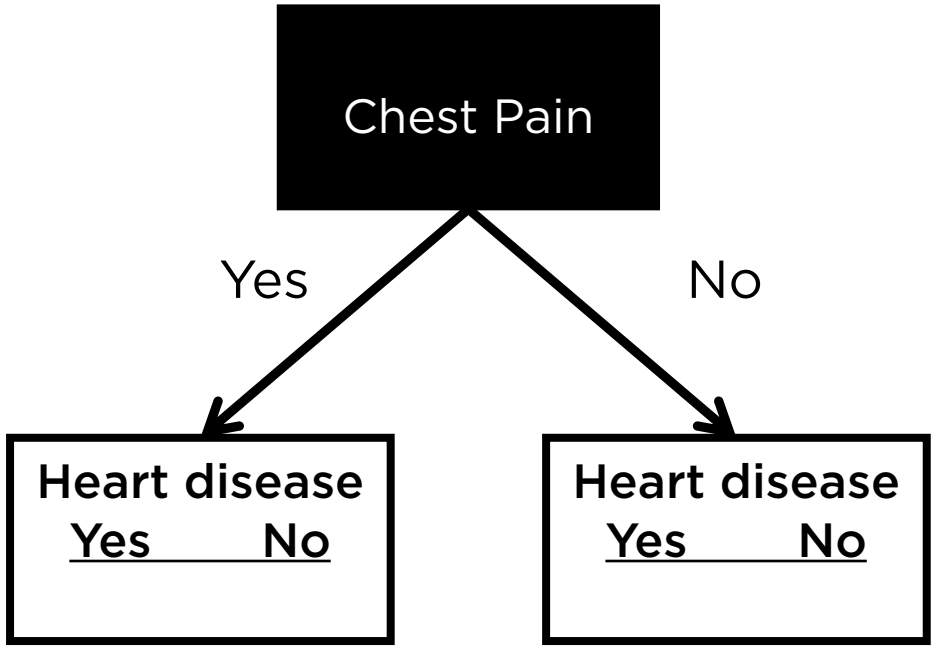
Chest pain	Good blood circulation	Blocked arteries	Heart disease
No	No	No	No
Yes	Yes	Yes	Yes
Yes	Yes	No	No
Yes	No	?	Yes
...	...	...	...



Decision Trees

- From a raw table of data to a decision tree

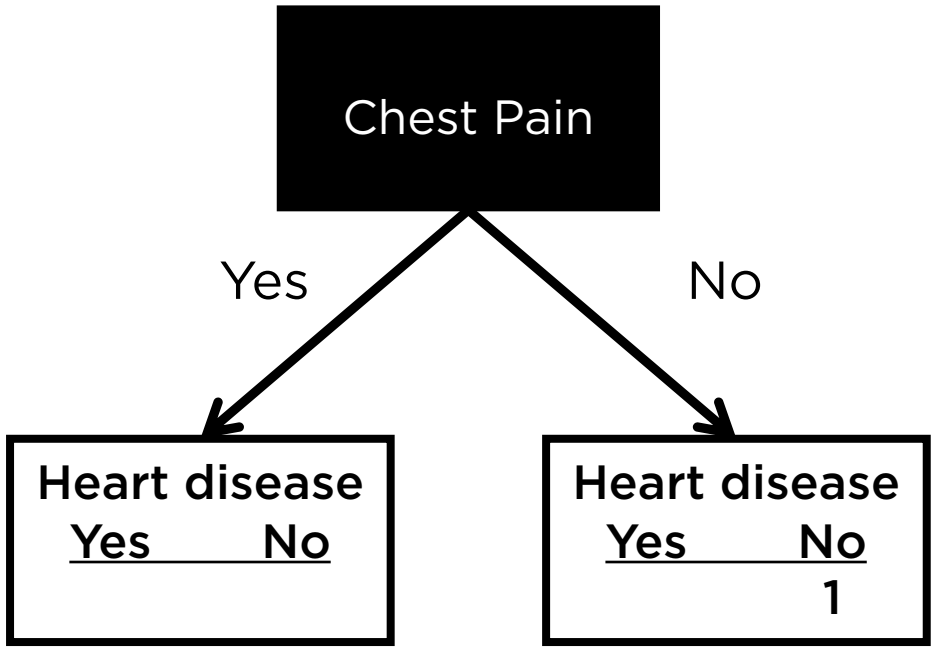
Chest pain	Good blood circulation	Blocked arteries	Heart disease
No	No	No	No
Yes	Yes	Yes	Yes
Yes	Yes	No	No
Yes	No	?	Yes
...	...	...	...



Decision Trees

- From a raw table of data to a decision tree

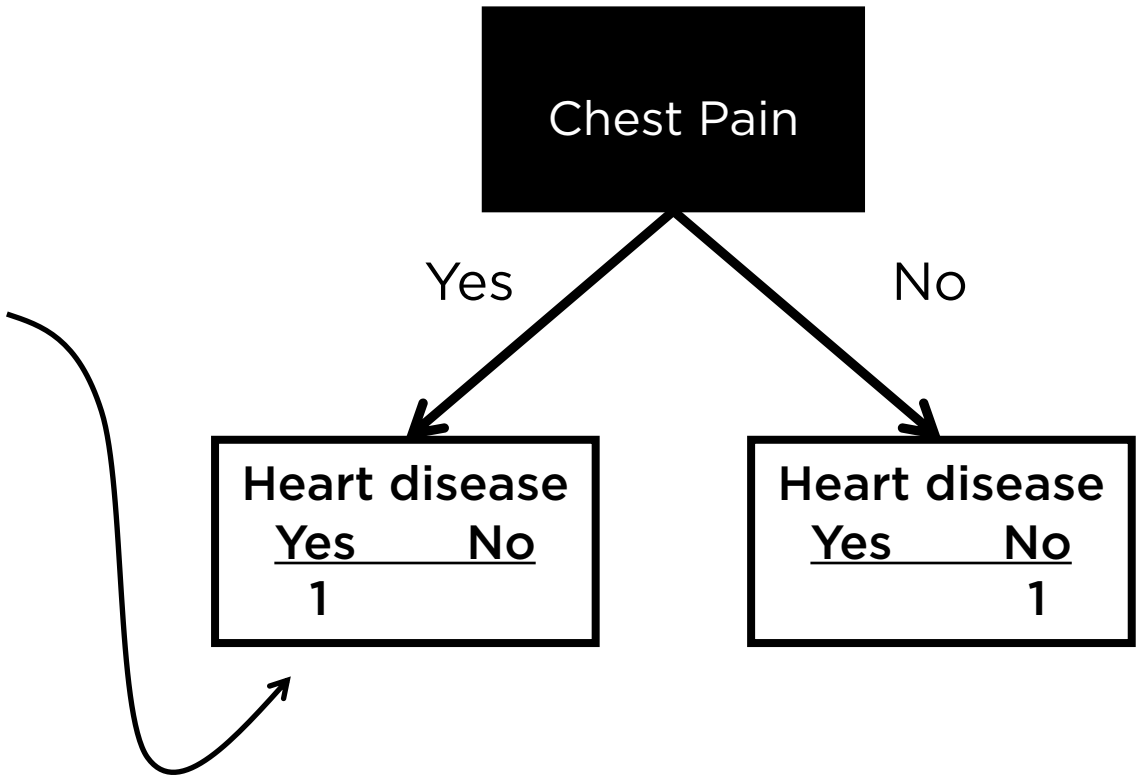
Chest pain	Good blood circulation	Blocked arteries	Heart disease
No	No	No	No
Yes	Yes	Yes	Yes
Yes	Yes	No	No
Yes	No	?	Yes
...	...	...	...



Decision Trees

- From a raw table of data to a decision tree

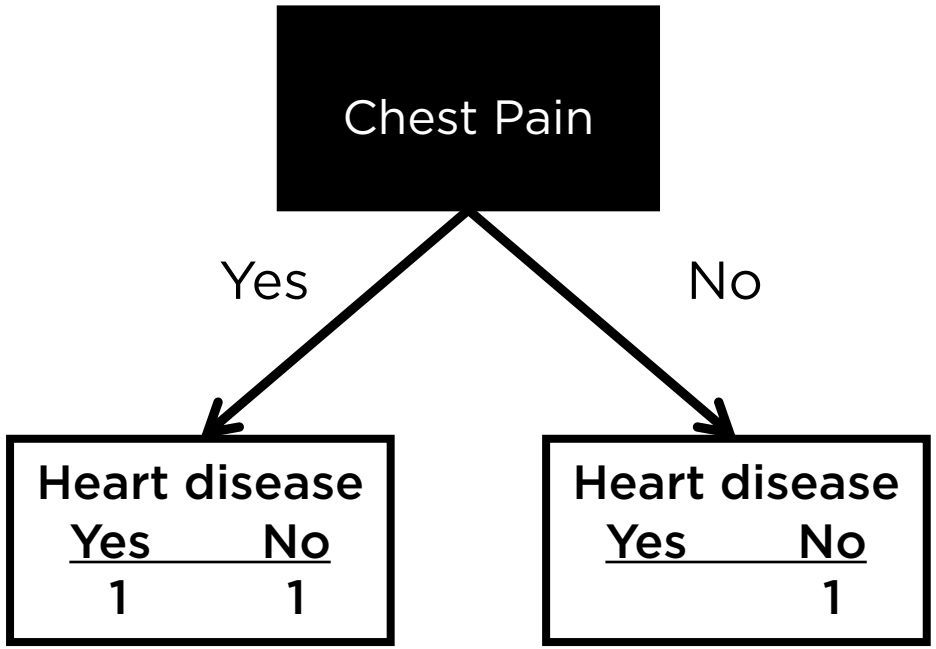
Chest pain	Good blood circulation	Blocked arteries	Heart disease
No	No	No	No
Yes	Yes	Yes	Yes
Yes	Yes	No	No
Yes	No	?	Yes
...	...	...	...



Decision Trees

- From a raw table of data to a decision tree

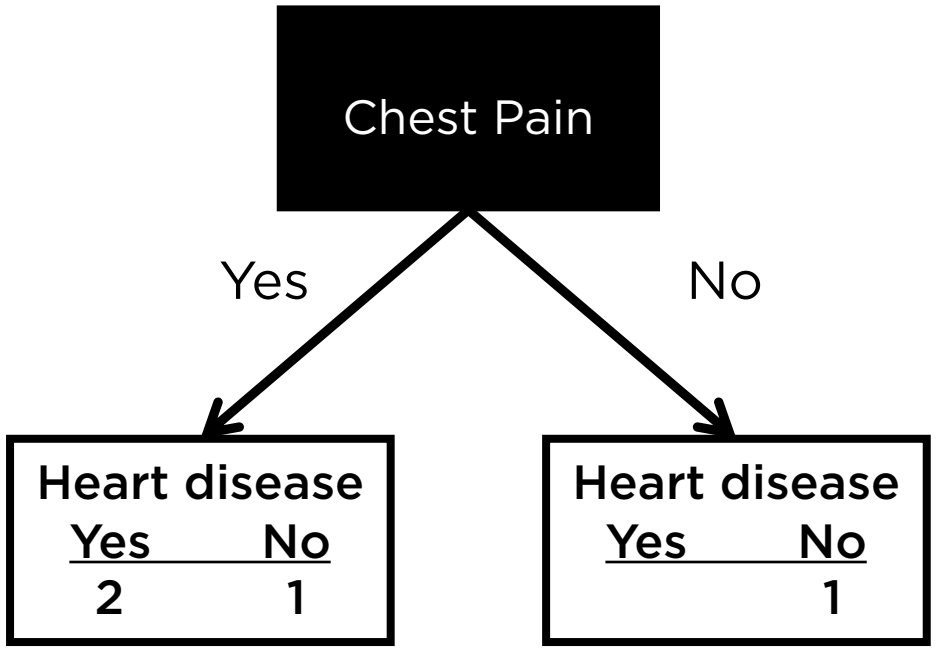
Chest pain	Good blood circulation	Blocked arteries	Heart disease
No	No	No	No
Yes	Yes	Yes	Yes
Yes	Yes	No	No
Yes	No	?	Yes
...	...	...	...



Decision Trees

- From a raw table of data to a decision tree

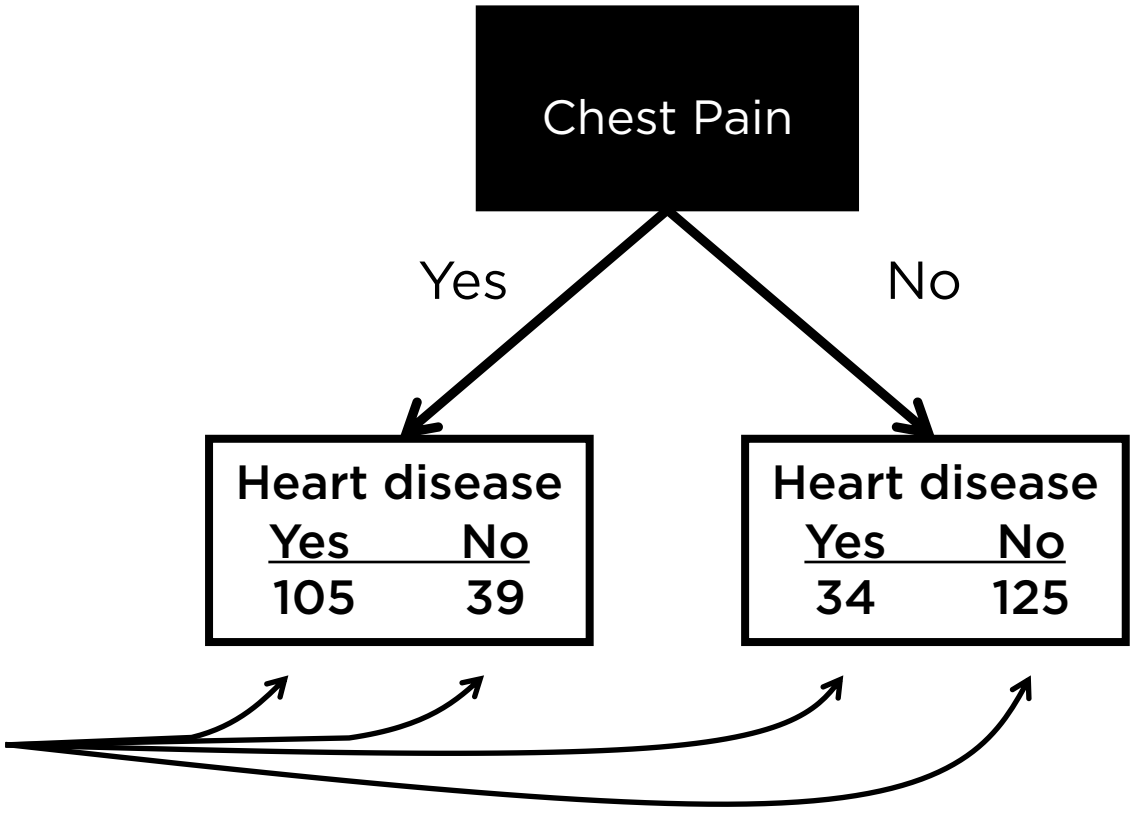
Chest pain	Good blood circulation	Blocked arteries	Heart disease
No	No	No	No
Yes	Yes	Yes	Yes
Yes	Yes	No	No
Yes	No	?	Yes
...	...	...	...



Decision Trees

- From a raw table of data to a decision tree

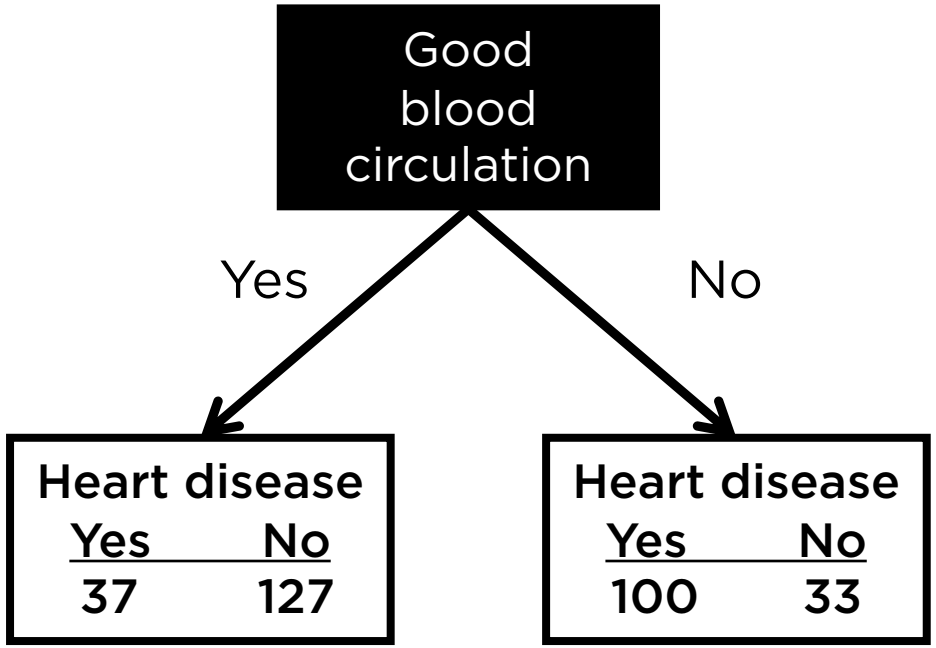
Chest pain	Good blood circulation	Blocked arteries	Heart disease
No	No	No	No
Yes	Yes	Yes	Yes
Yes	Yes	No	No
Yes	No	?	Yes
...	...	...	...



Decision Trees

- From a raw table of data to a decision tree

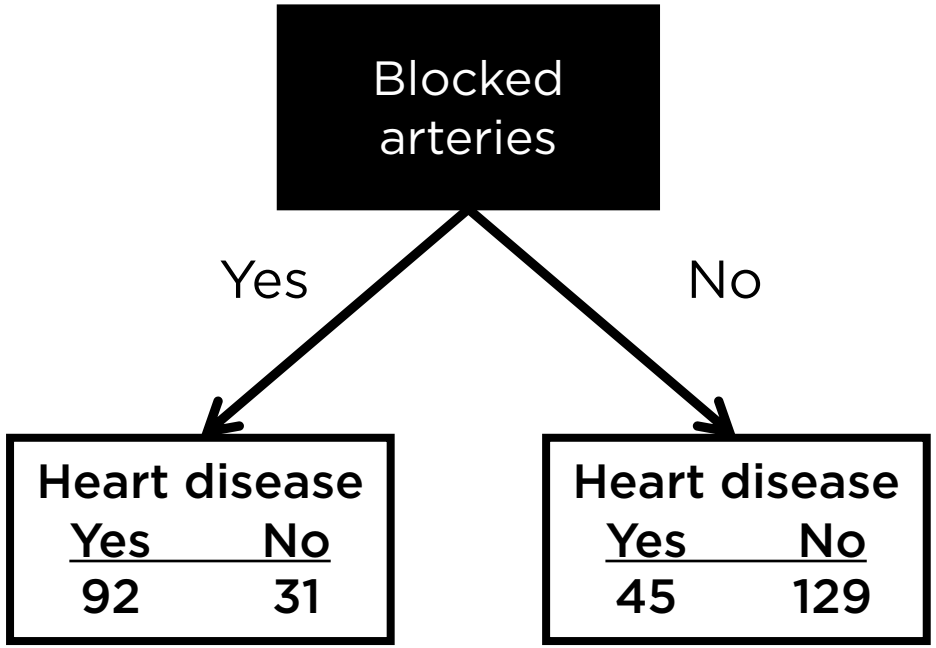
Chest pain	Good blood circulation	Blocked arteries	Heart disease
No	No	No	No
Yes	Yes	Yes	Yes
Yes	Yes	No	No
Yes	No	?	Yes
...	...	...	...



Decision Trees

- From a raw table of data to a decision tree

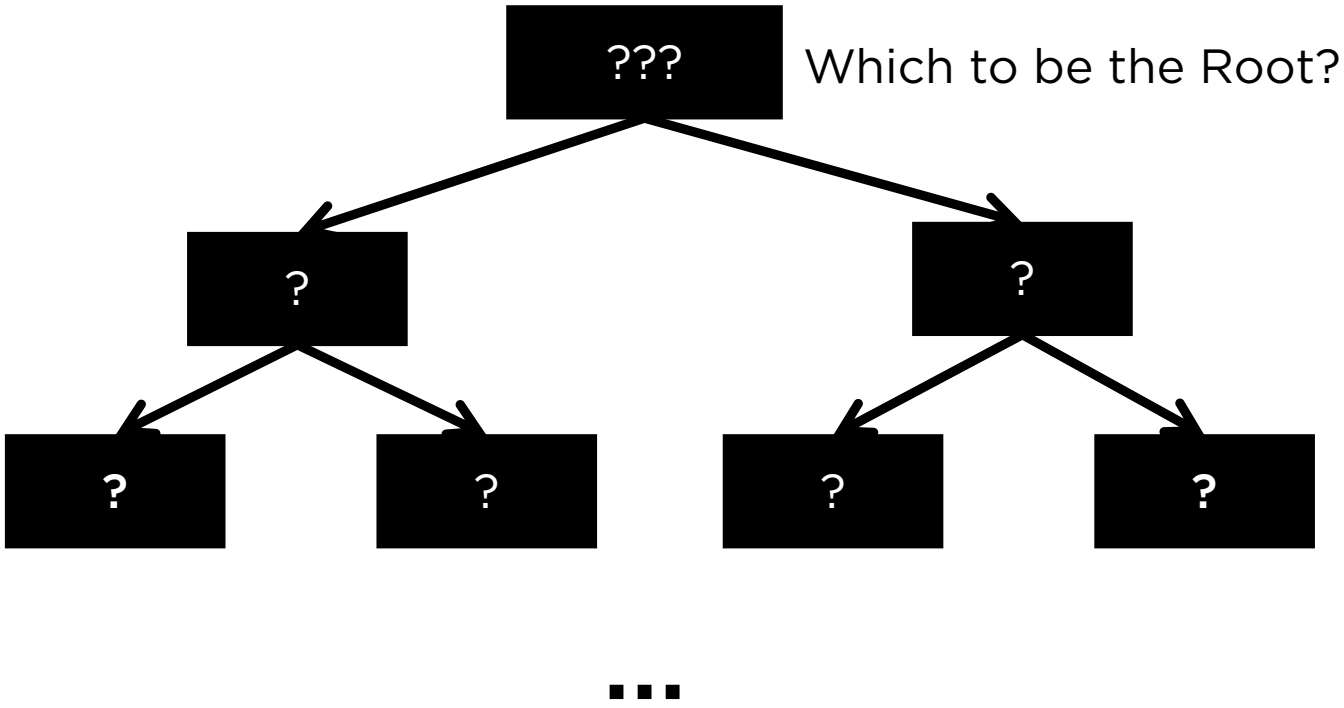
Chest pain	Good blood circulation	Blocked arteries	Heart disease
No	No	No	No
Yes	Yes	Yes	Yes
Yes	Yes	No	No
Yes	No	?	Yes
...	...	...	...



Decision Trees

- From a raw table of data to a decision tree

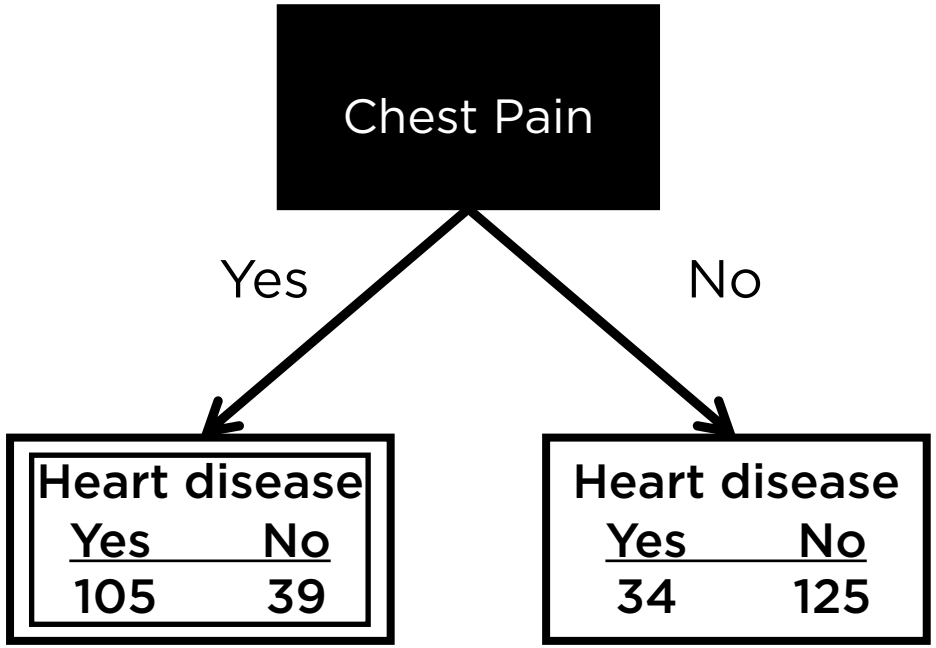
Chest pain	Good blood circulation	Blocked arteries	Heart disease
No	No	No	No
Yes	Yes	Yes	Yes
Yes	Yes	No	No
Yes	No	?	Yes
...	...	...	...



Decision Trees

- From a raw table of data to a decision tree

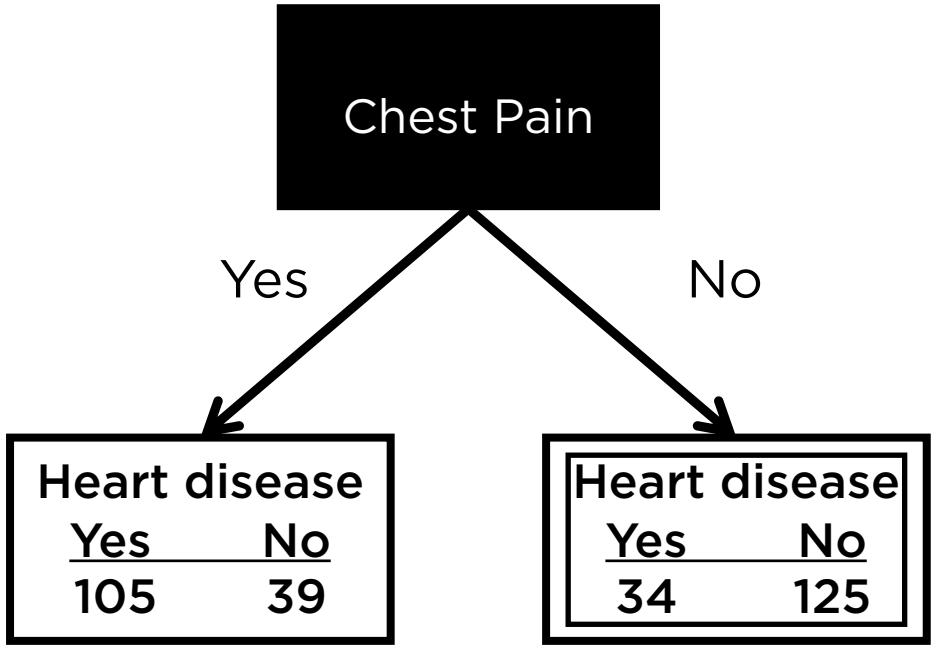
Chest pain	Good blood circulation	Blocked arteries	Heart disease
No	No	No	No
Yes	Yes	Yes	Yes
Yes	Yes	No	No
Yes	No	?	Yes
...	...	...	...



Decision Trees

- From a raw table of data to a decision tree

Chest pain	Good blood circulation	Blocked arteries	Heart disease
No	No	No	No
Yes	Yes	Yes	Yes
Yes	Yes	No	No
Yes	No	?	Yes
...	...	...	...



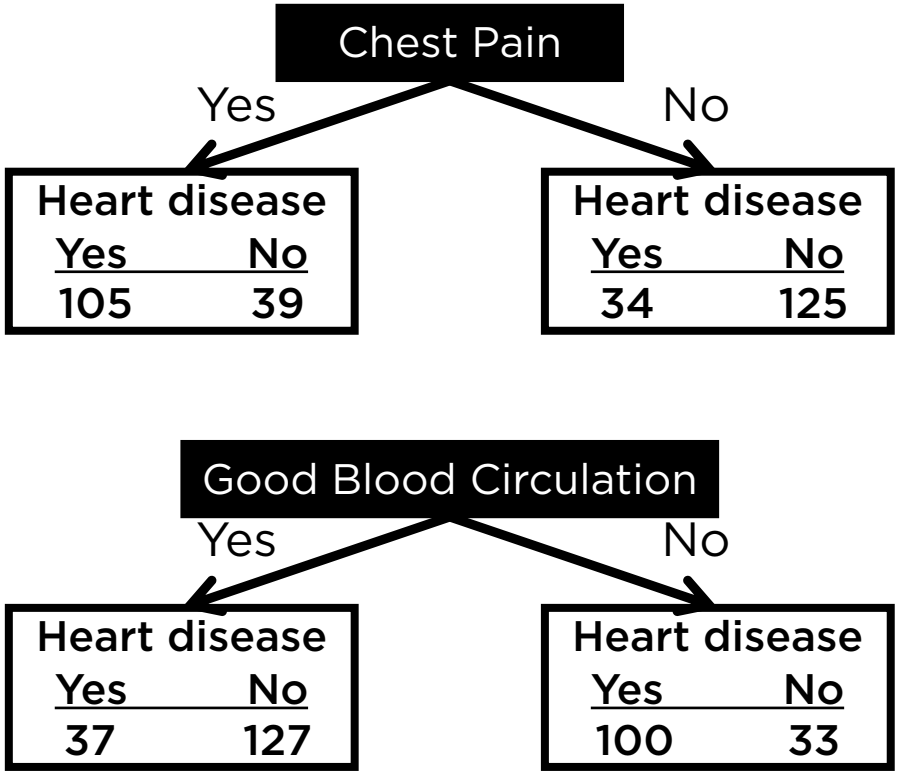
None of the leaf nodes are 100% "Yes Heart Disease" or 100% "No Heart Disease".

→ They are considered **Impure**.

Decision Trees

- From a raw table of data to a decision tree

Chest pain	Good blood circulation	Blocked arteries	Heart disease
No	No	No	No
Yes	Yes	Yes	Yes
Yes	Yes	No	No
Yes	No	?	Yes
...	...	...	...

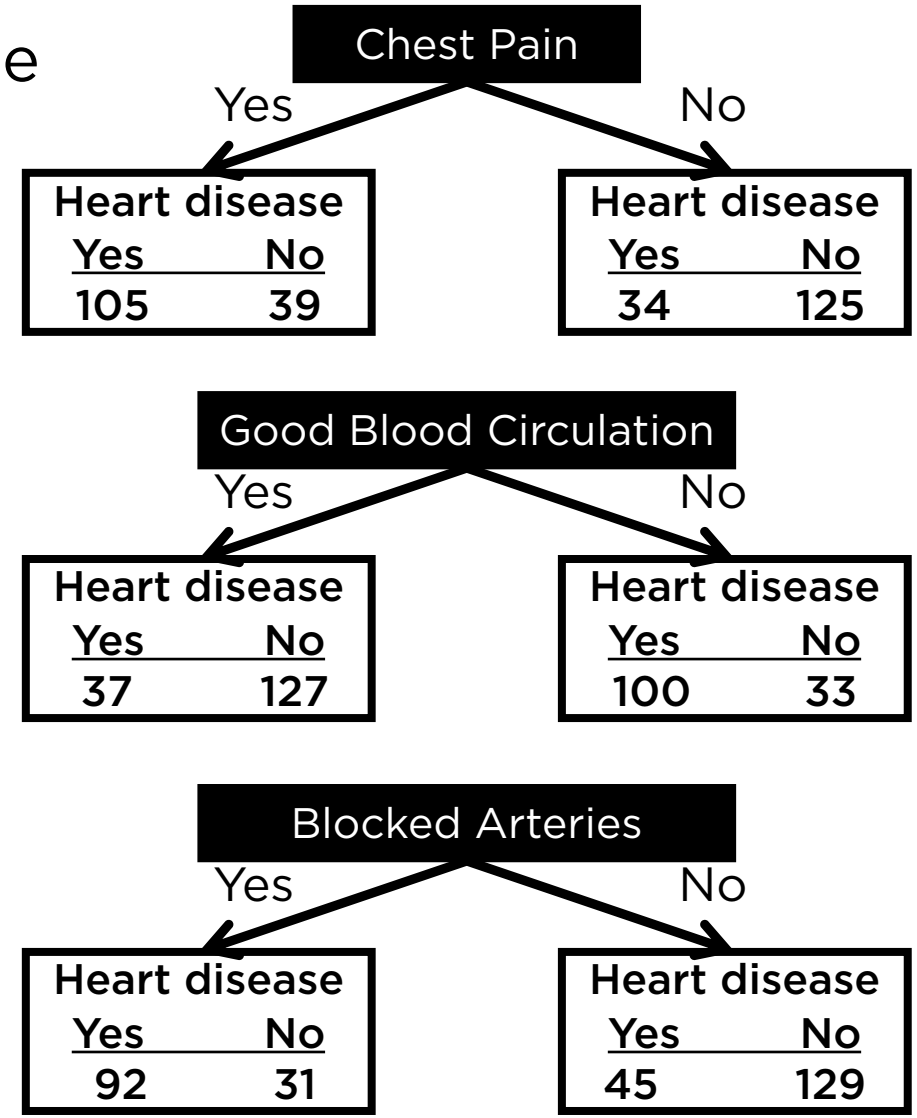


How about **Good Blood Circulation**?

Decision Trees

- From a raw table of data to a decision tree

Chest pain	Good blood circulation	Blocked arteries	Heart disease
No	No	No	No
Yes	Yes	Yes	Yes
Yes	Yes	No	No
Yes	No	?	Yes
...	...	...	...

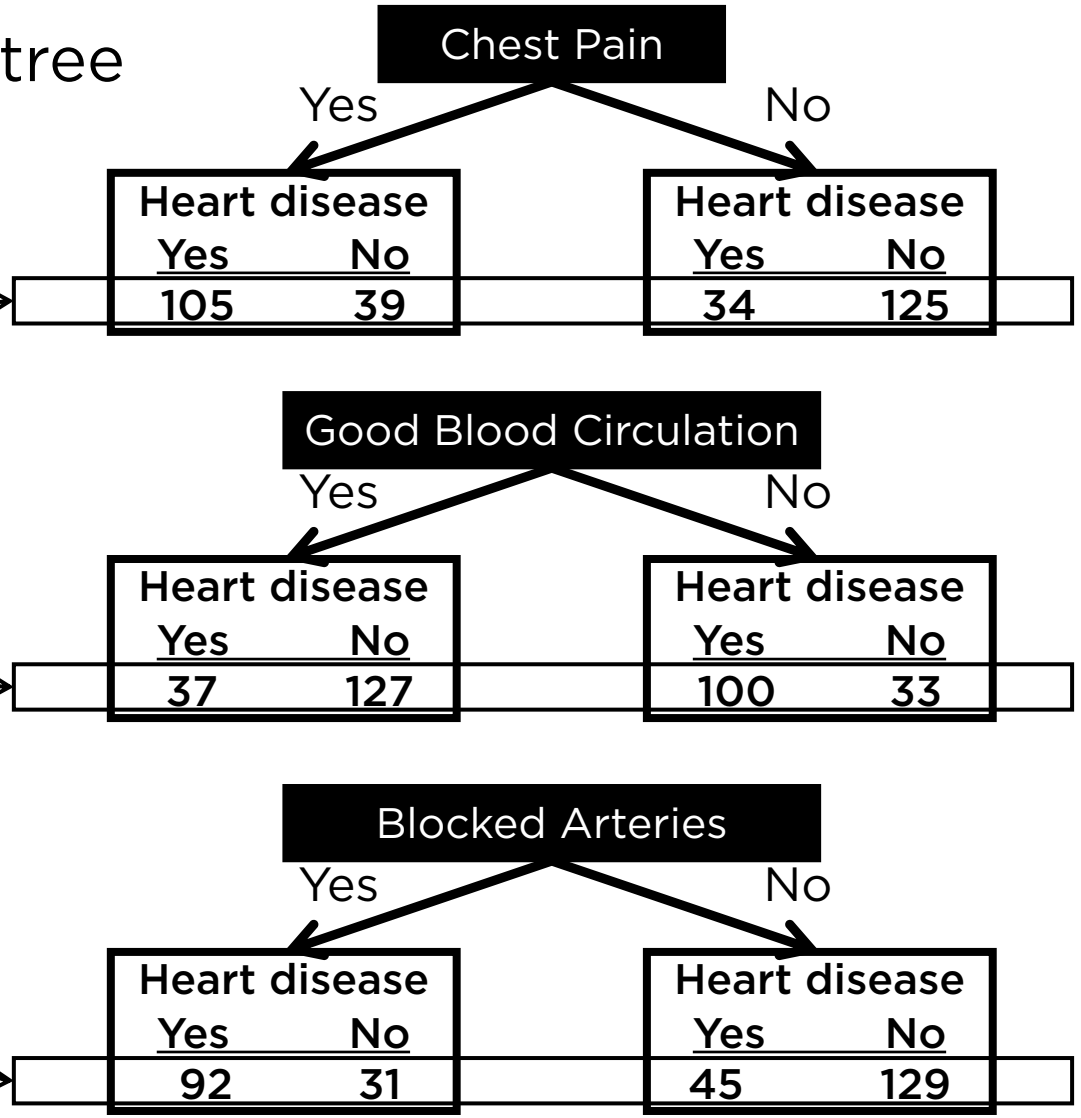


And **Blocked Arteries**?

Decision Trees

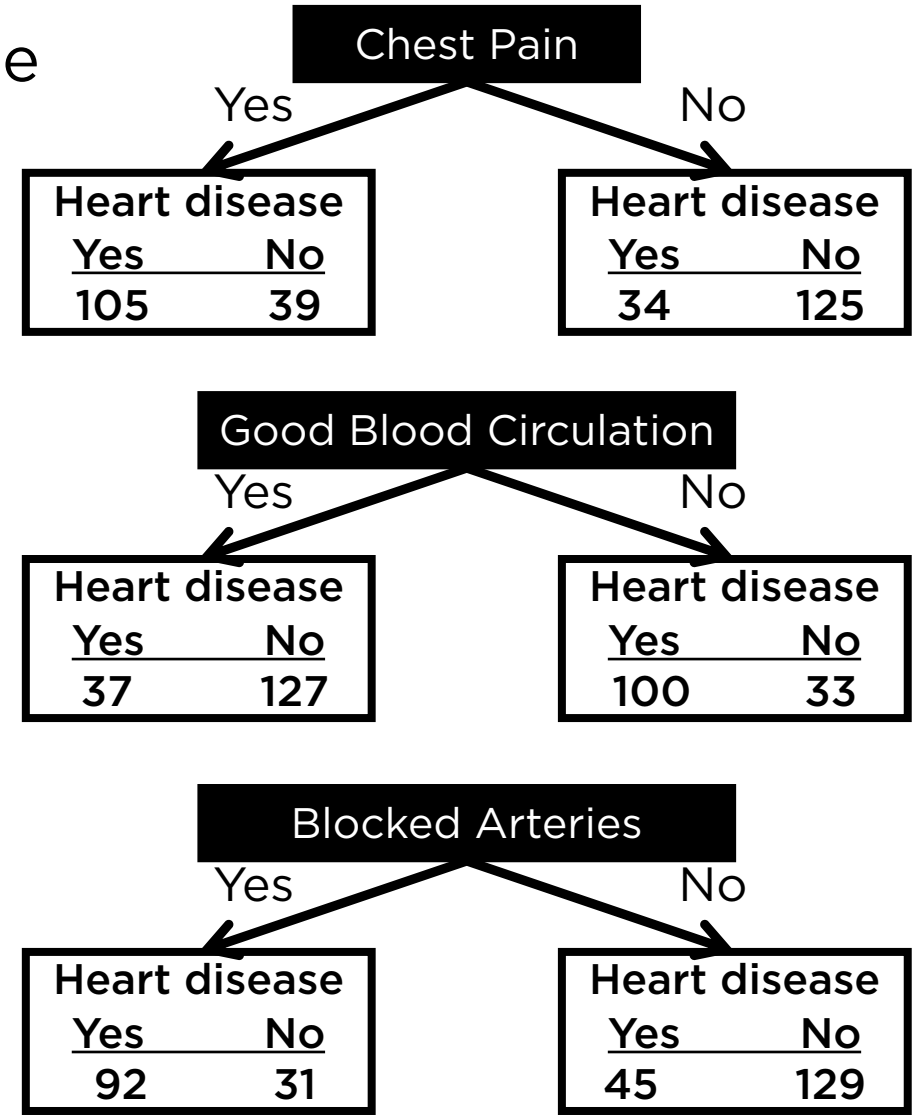
- From a raw table of data to a decision tree

Note: total number is different.



Decision Trees

- From a raw table of data to a decision tree



How to decide which is the best?

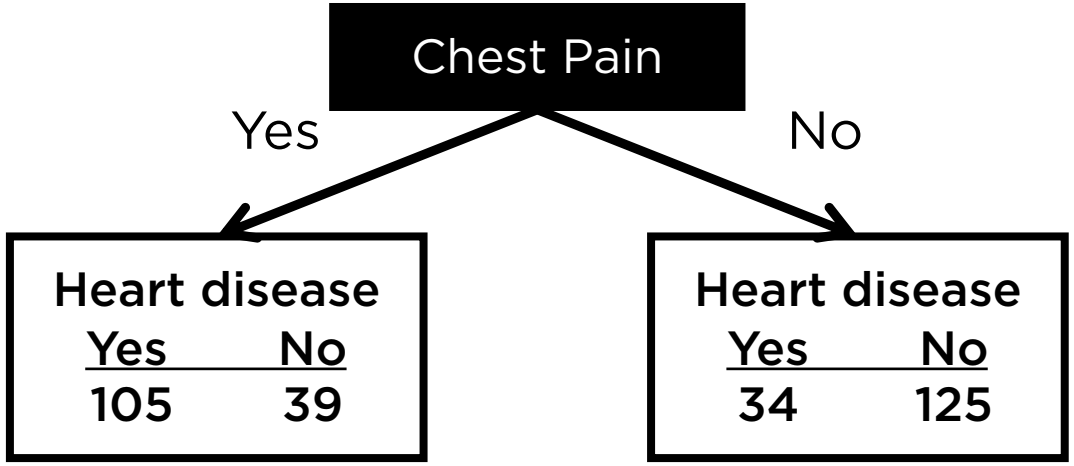
By measuring and comparing **Impurity**.

Gini

Decision Trees

- From a raw table of data to a decision tree

Calculating **Gini Impurity**...



$$Gini\ Impurity = 1 - (the\ probability\ of\ Yes)^2 - (the\ probability\ of\ No)^2$$

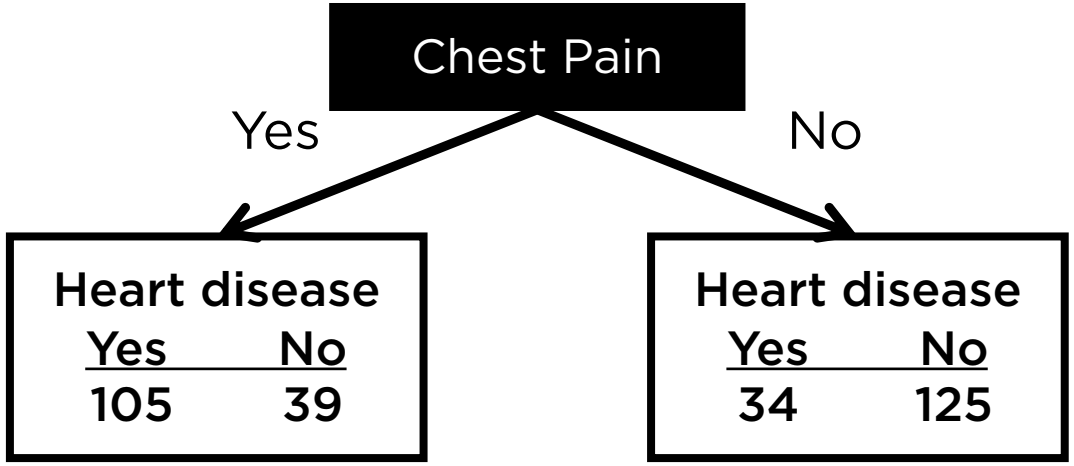
$$= 1 - \left(\frac{105}{105 + 39}\right)^2 - \left(\frac{39}{105 + 39}\right)^2$$

$$= 0.395$$

Decision Trees

- From a raw table of data to a decision tree

Calculating **Gini Impurity**...



$$Gini\ Impurity = 0.395$$

$$Gini\ Impurity = 1 - (the\ probability\ of\ Yes)^2 - (the\ probability\ of\ No)^2$$

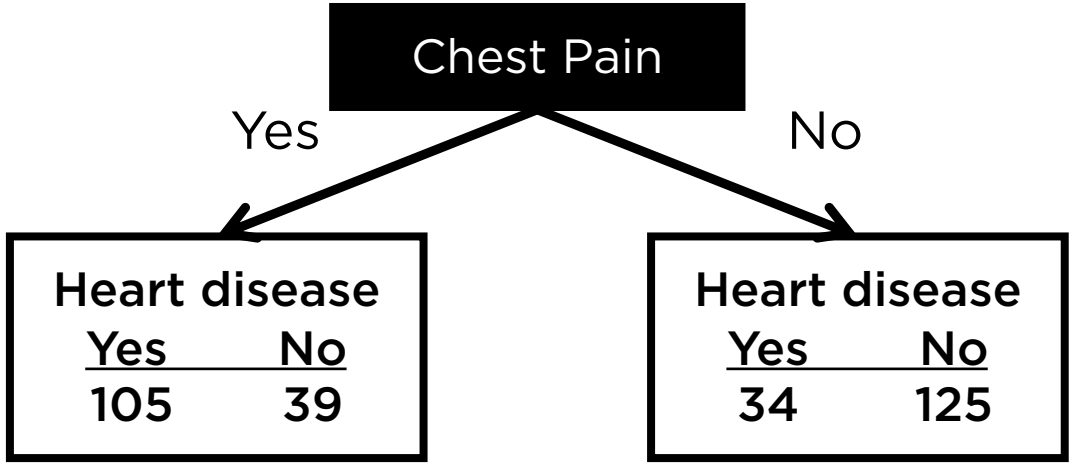
$$= 1 - \left(\frac{34}{34+125}\right)^2 - \left(\frac{125}{34+125}\right)^2$$

$$= 0.336$$

Decision Trees

- From a raw table of data to a decision tree

Calculating **Gini Impurity**...



$Gini\ Impurity = 0.395$ $\quad$ $Gini\ Impurity = 0.336$

$$Gini\ Impurity for\ Chest\ Pain = weighted\ average\ of\ Gini\ impurities\ for\ the\ leaf\ nodes$$

$$= \frac{144}{144+159} \times 0.395 + \frac{159}{144+159} \times 0.336$$

$$= 0.364$$

Decision Trees

- From a raw table of data to a decision tree

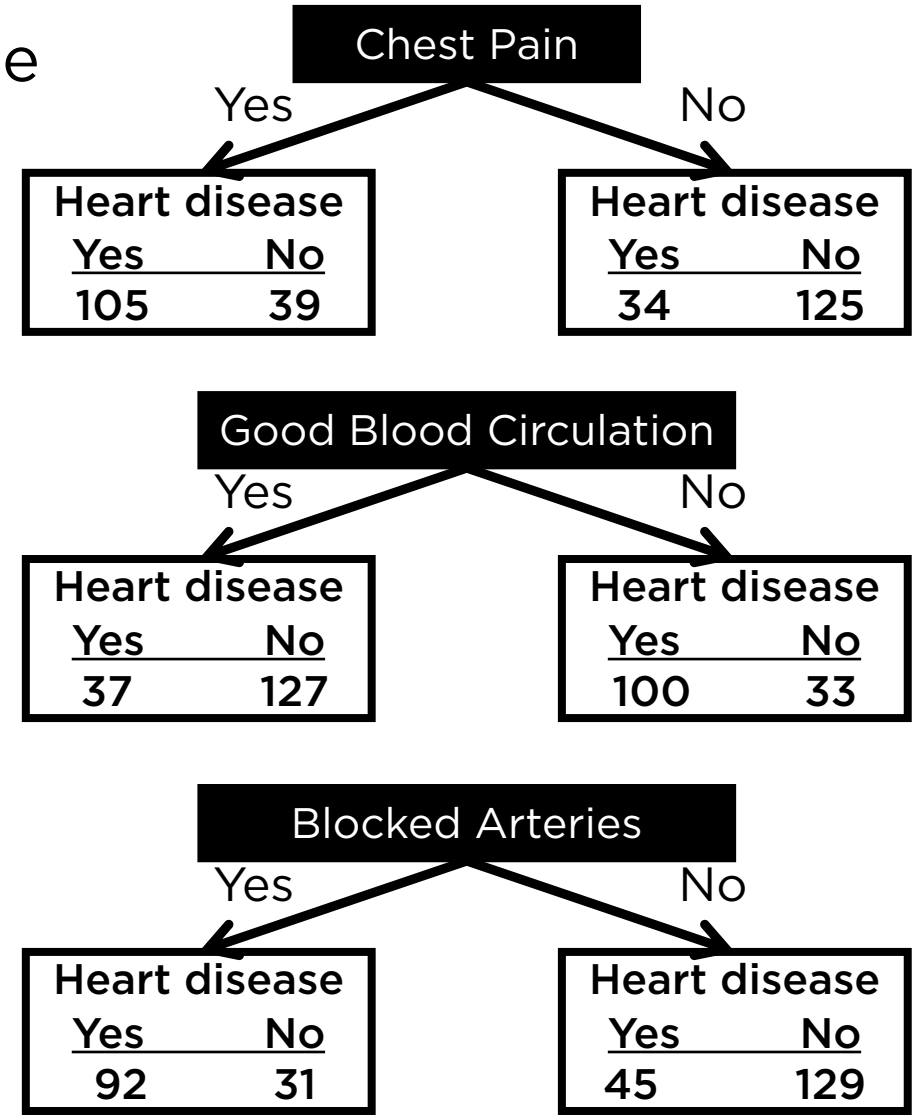
Calculating **Gini Impurity**...

$Gini\ Impurity\ for\ \boldsymbol{Chest\ Pain} = 0.364$

$Gini\ Impurity\ for\ \boldsymbol{Good\ Blood\ Circulation} = 0.360$

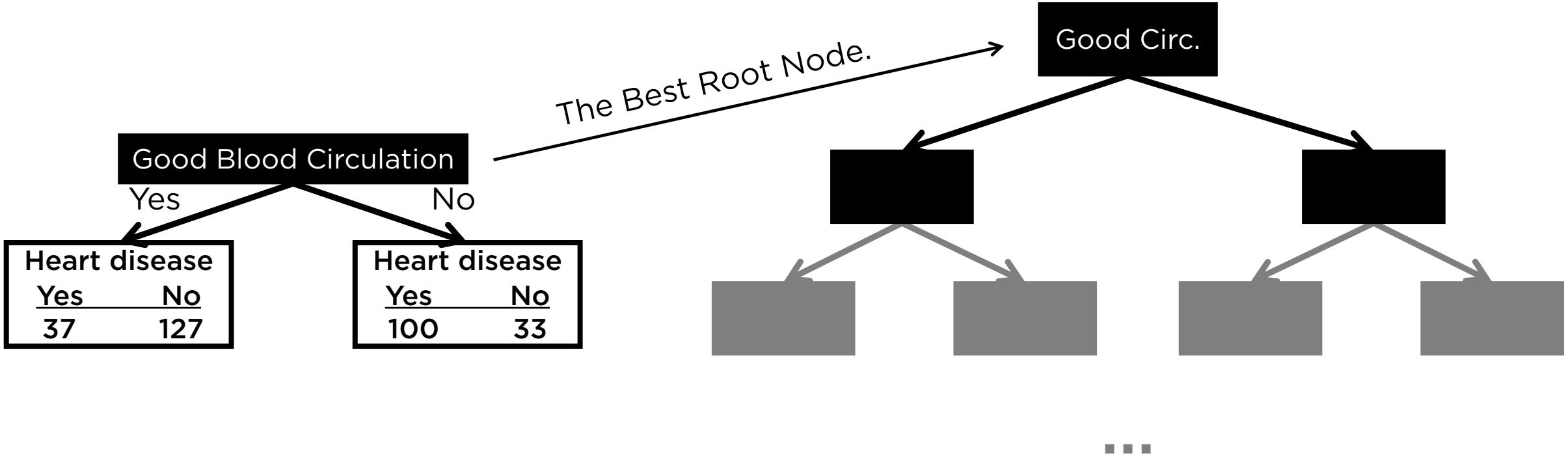
The lowest impurity

$Gini\ Impurity\ for\ \boldsymbol{Blocked\ Arteries} = 0.381$



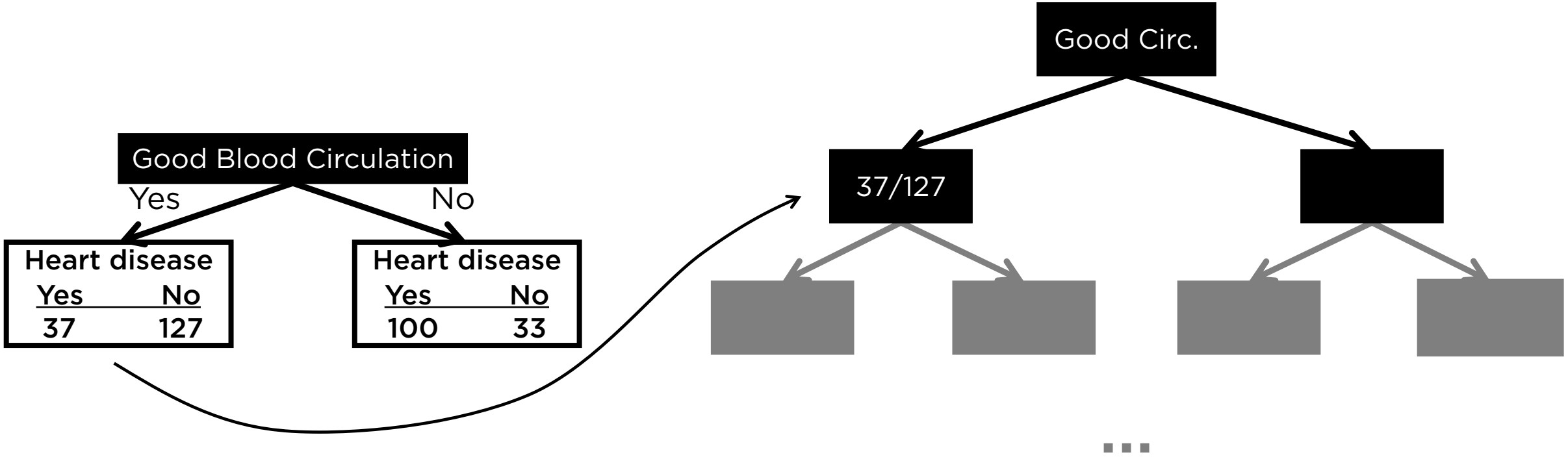
Decision Trees

- From a raw table of data to a decision tree



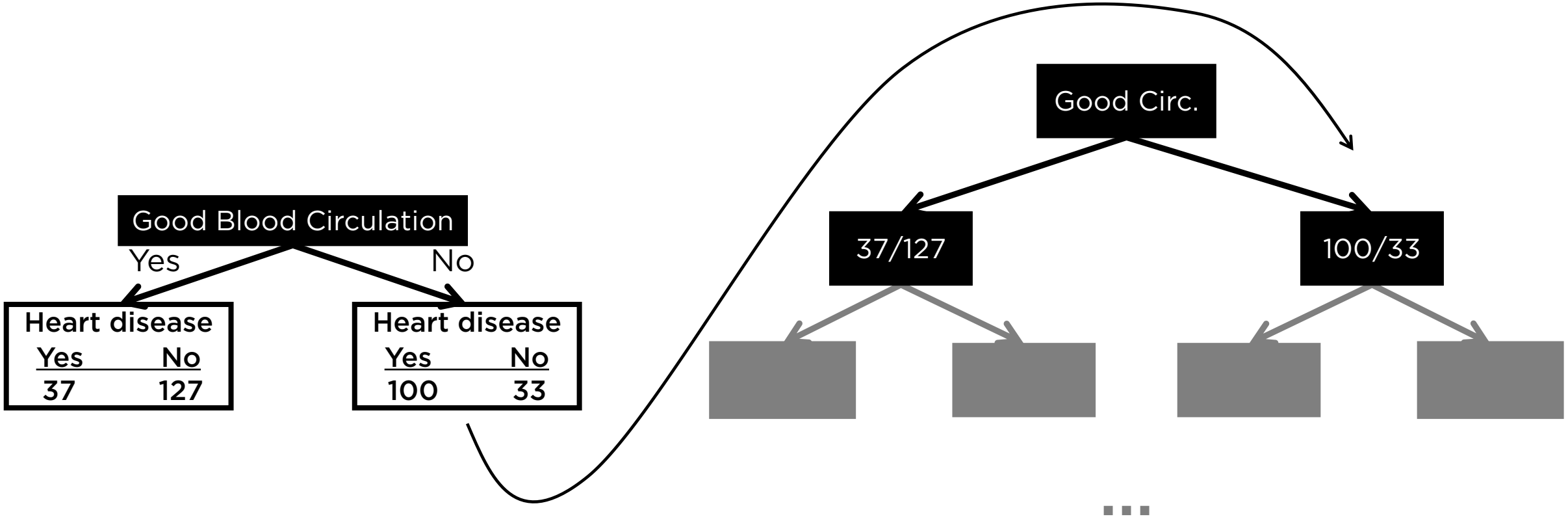
Decision Trees

- From a raw table of data to a decision tree



Decision Trees

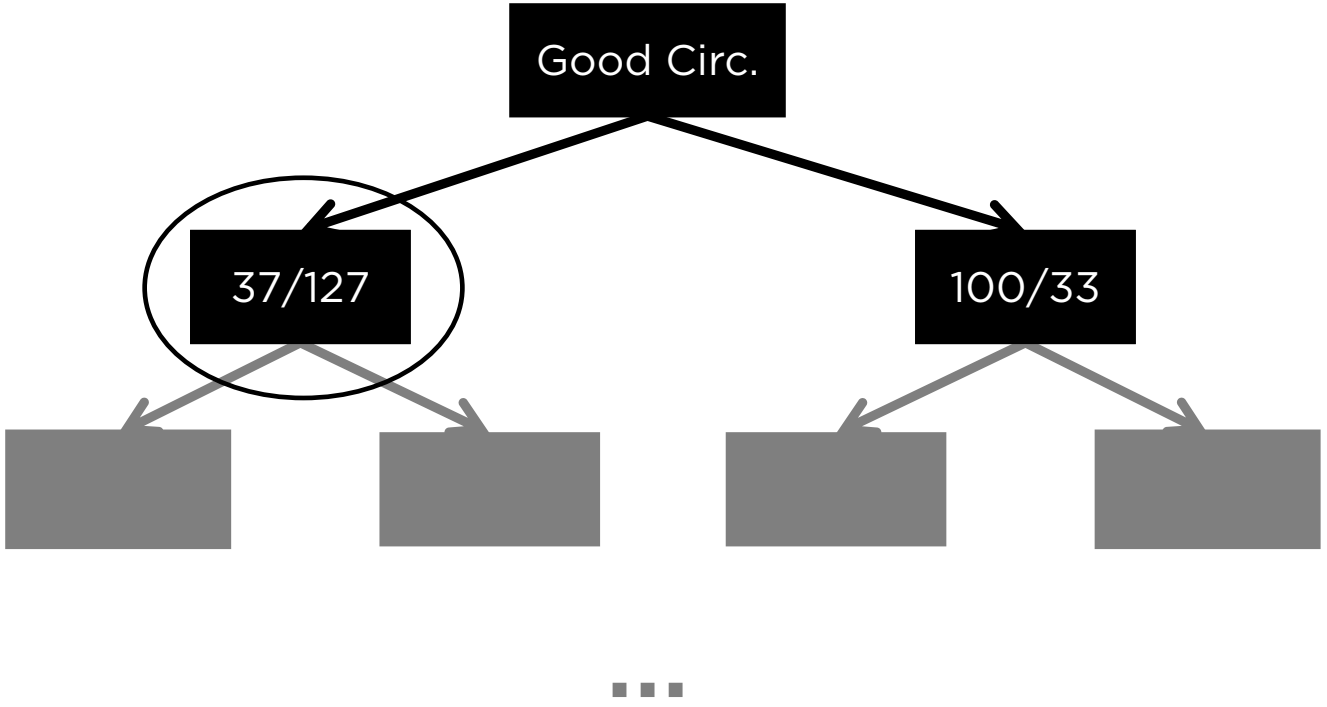
- From a raw table of data to a decision tree



Decision Trees

- From a raw table of data to a decision tree

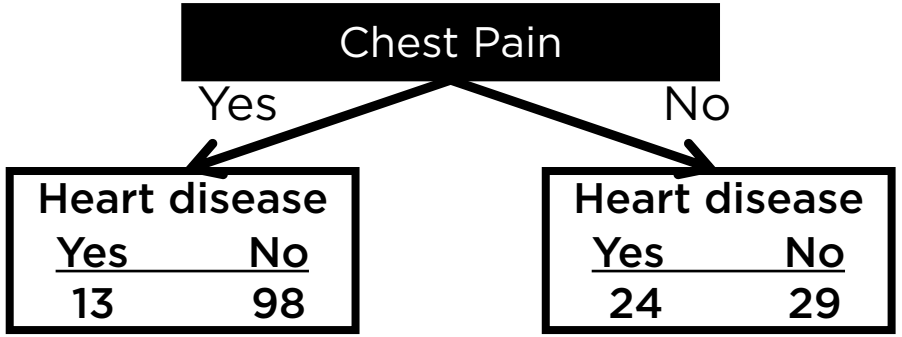
How well **Chest Pain** and **Blocked Arteries** separate these 164 patients?



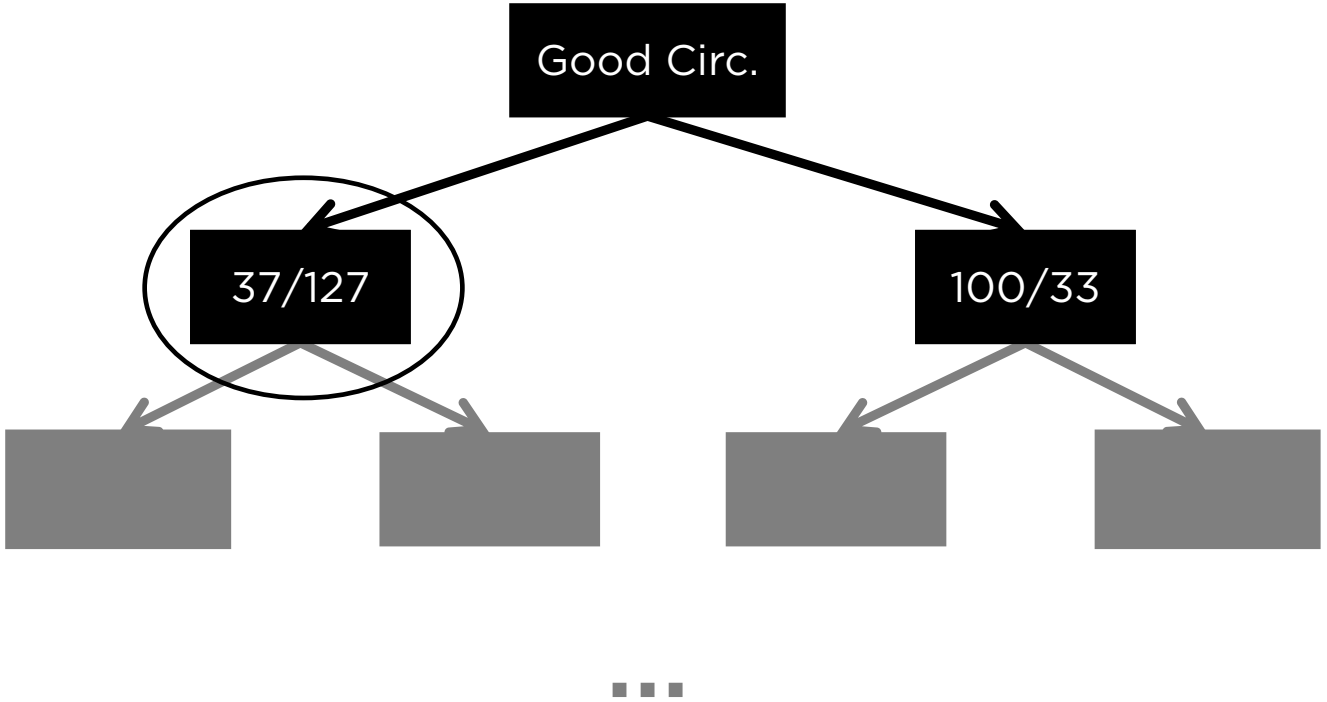
Decision Trees

- From a raw table of data to a decision tree

How well **Chest Pain** and **Blocked Arteries** separate these 164 patients?



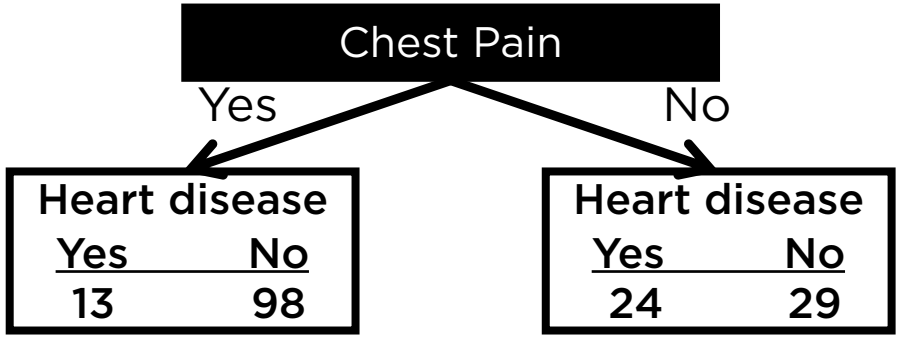
$Gini\ Impurity\ for\ \boldsymbol{Chest\ Pain} = 0.3$



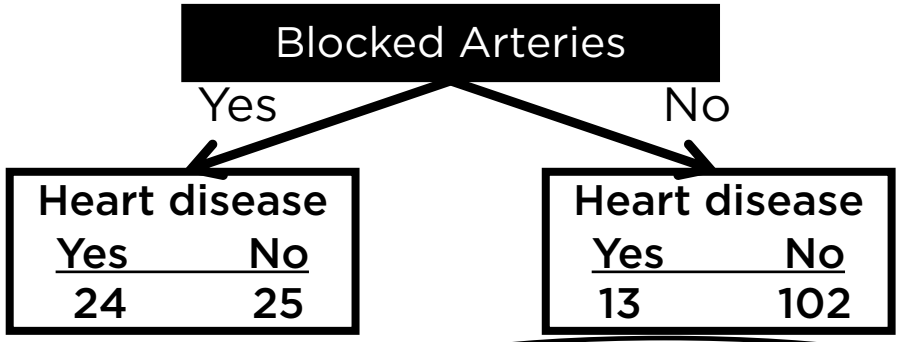
Decision Trees

- From a raw table of data to a decision tree

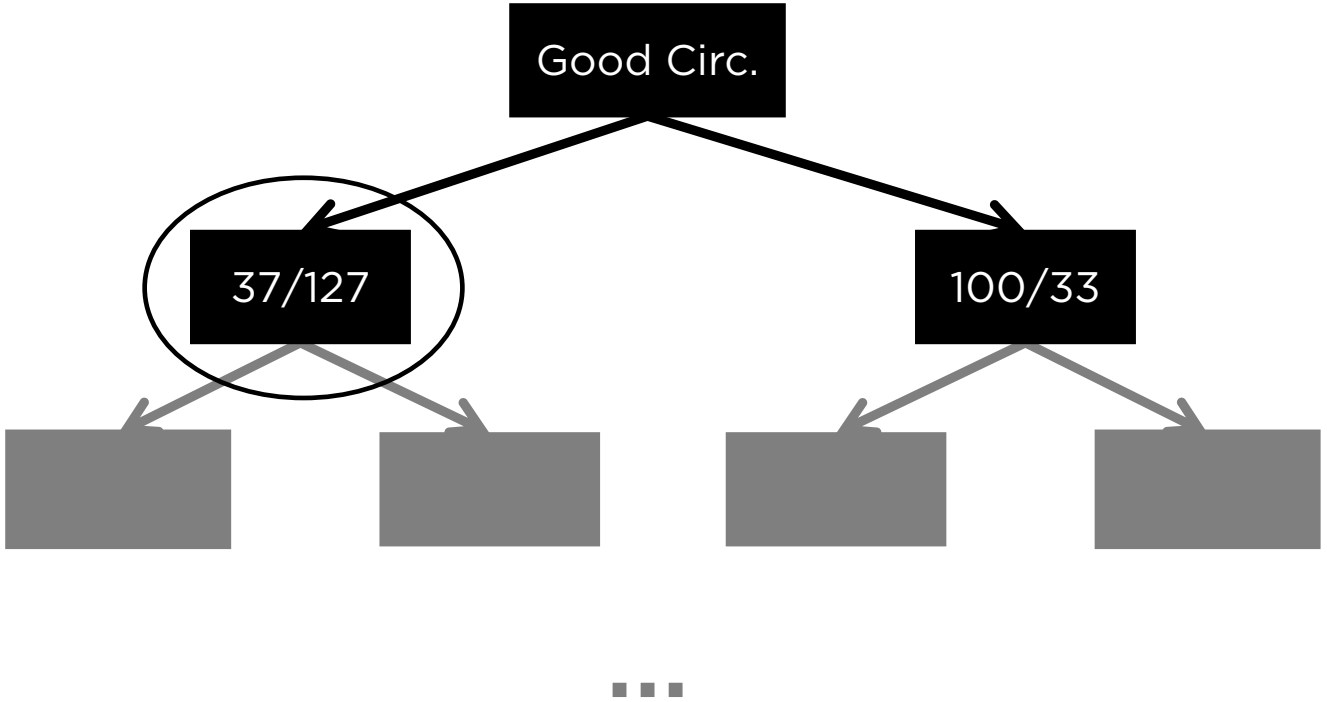
How well **Chest Pain** and **Blocked Arteries** separate these 164 patients?



$Gini\ Impurity\ for\ \boldsymbol{Chest\ Pain} = 0.3$



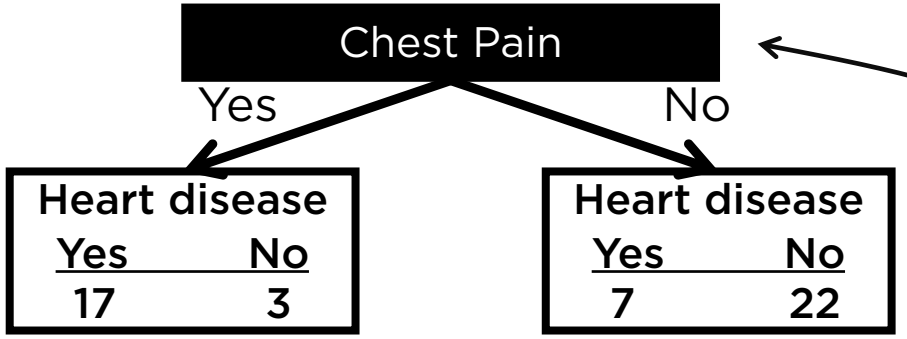
$Gini\ Impurity\ for\ \boldsymbol{Blocked\ Arteries} = 0.290$ The lowest impurity

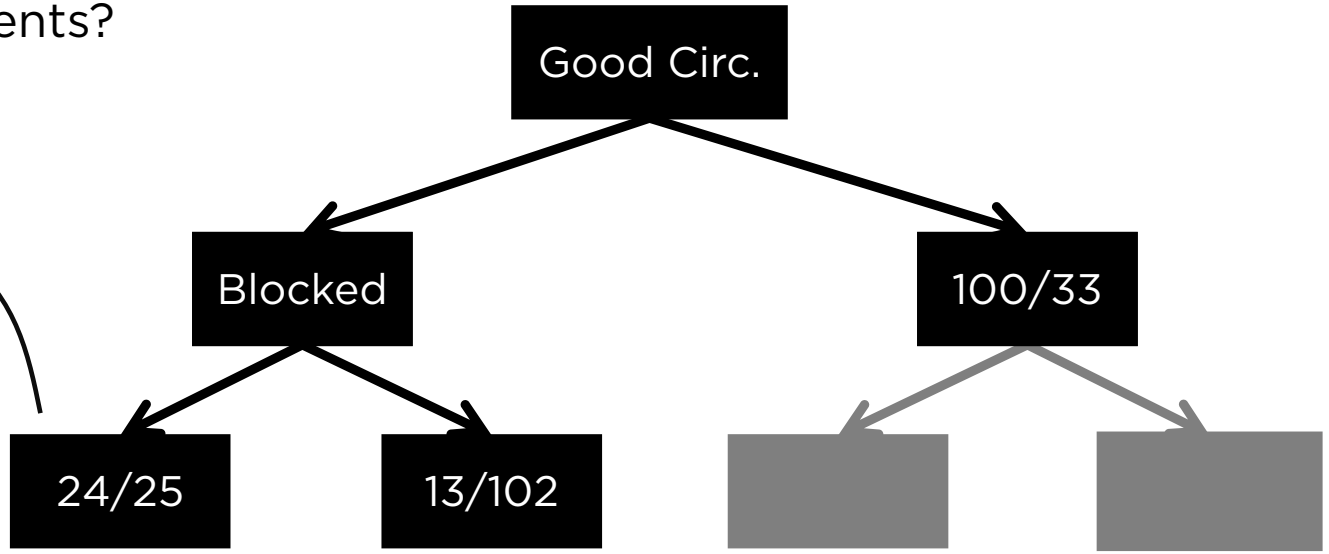


Decision Trees

- From a raw table of data to a decision tree

How well **Chest Pain** separate these 49 patients?



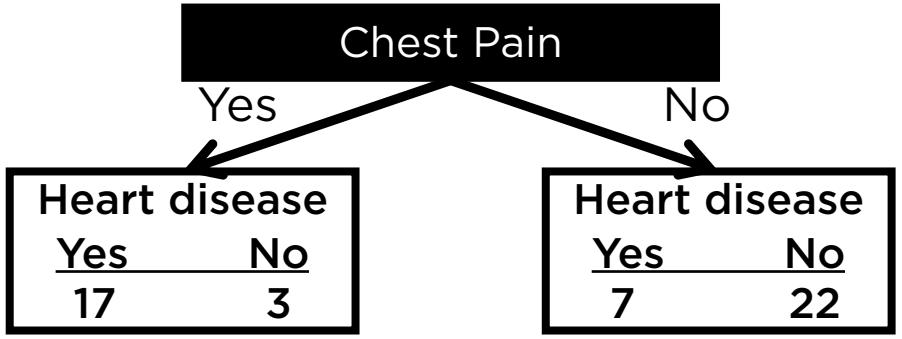


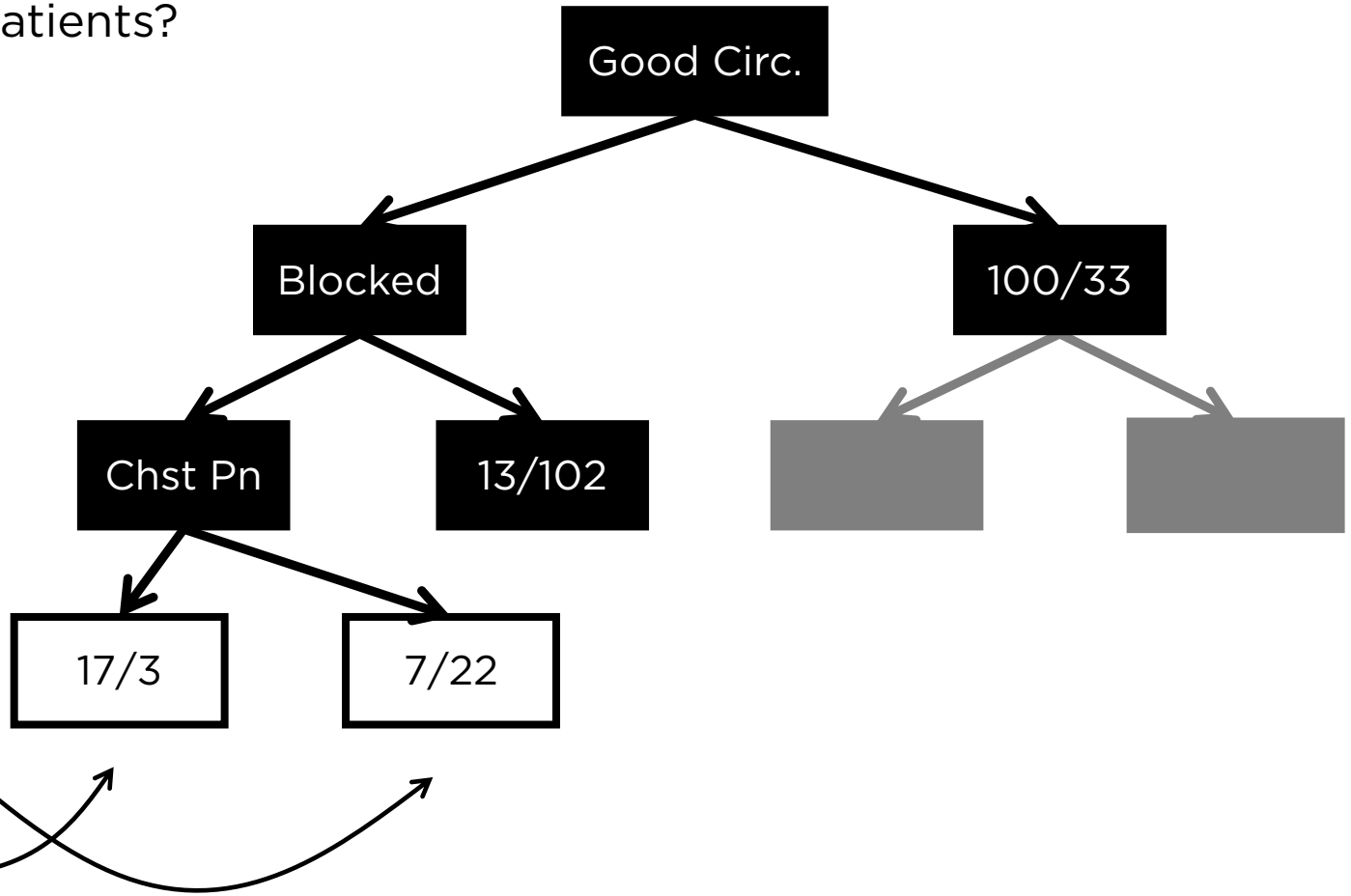
...

Decision Trees

- From a raw table of data to a decision tree

How well **Chest Pain** separate these 49 patients?



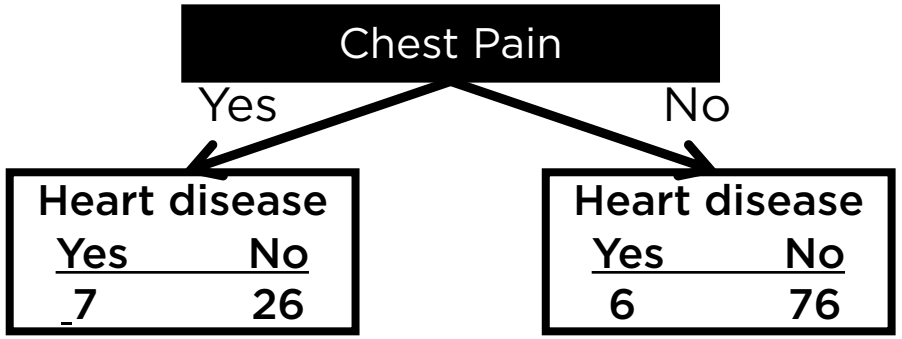


Decision Trees

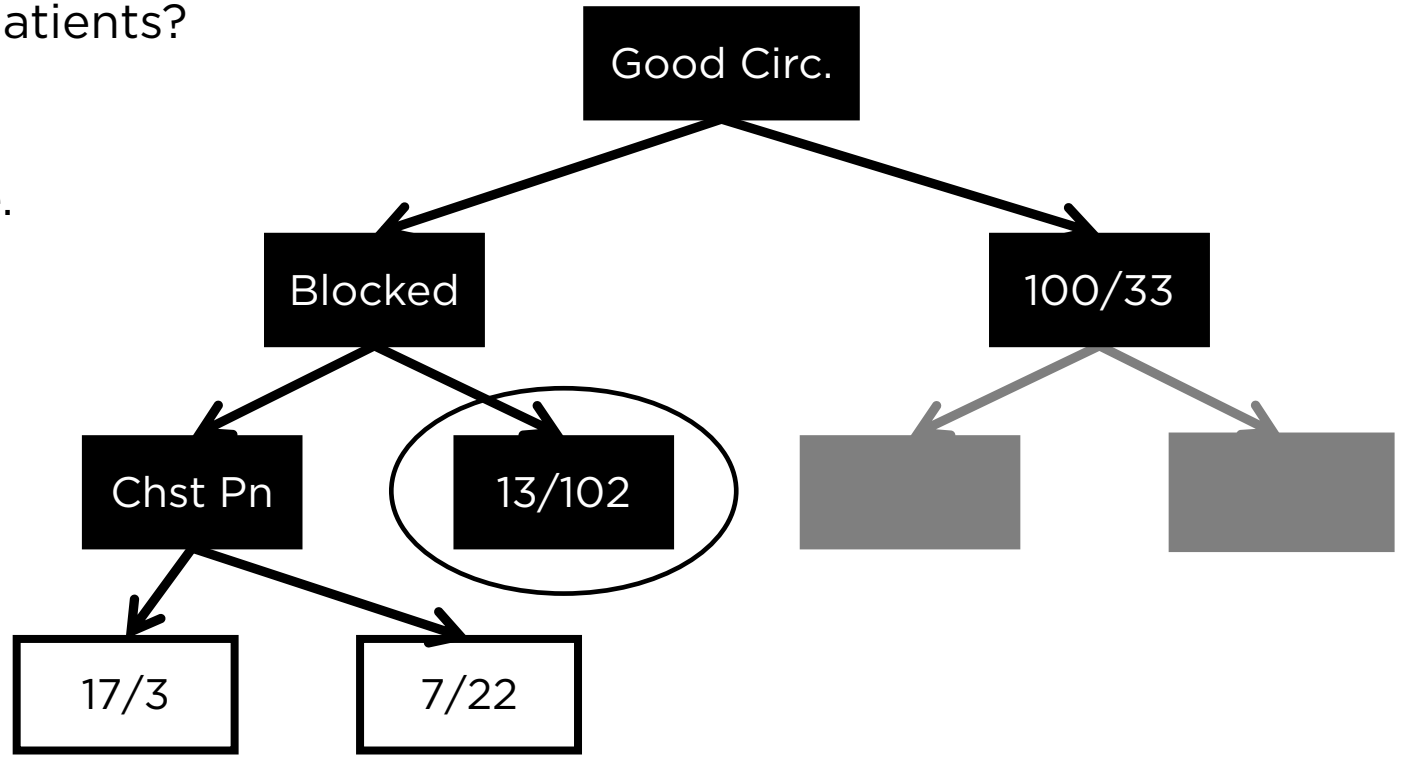
- From a raw table of data to a decision tree

How well **Chest Pain** separate these 115 patients?

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



Do these new leaves separate patients better than before?

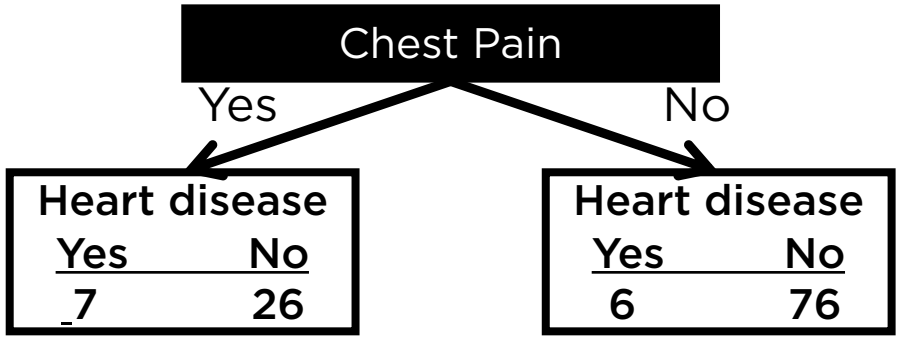


Decision Trees

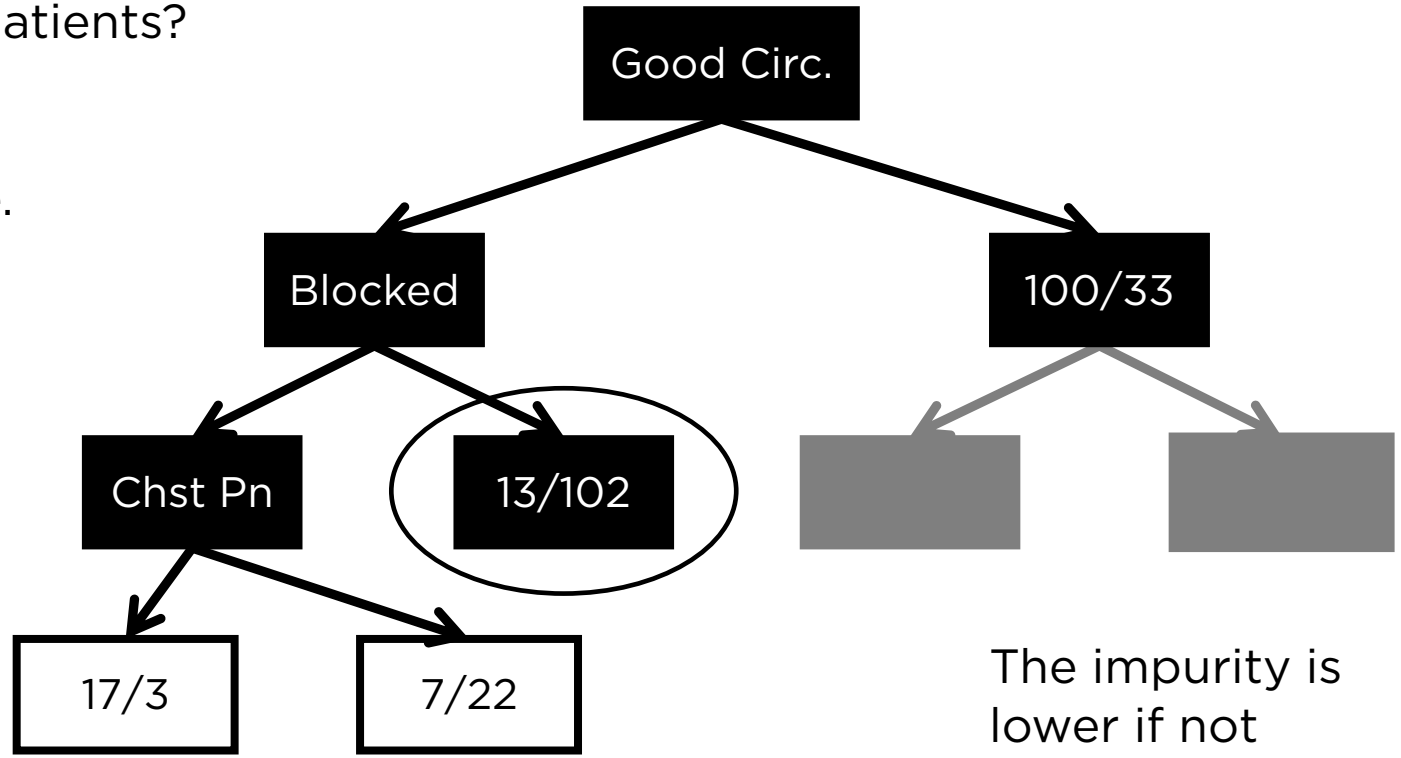
- From a raw table of data to a decision tree

How well **Chest Pain** separate these 115 patients?

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



$$Gini\ Impurity\ for\ \boldsymbol{Chest\ Pain} = 0.29$$



Gini impurity for this node, before using **Chest Pain** to separate patients is...

$$1 - \left(\frac{13}{13+102}\right)^2 - \left(\frac{102}{13+102}\right)^2 = 0.2$$

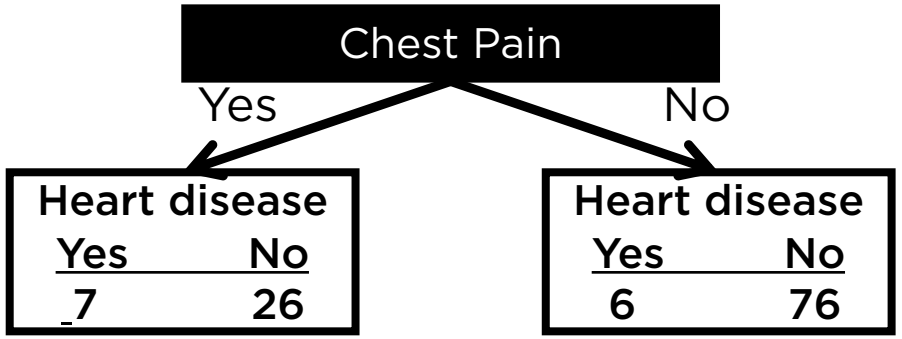
The impurity is lower if not separating patients using Chest Pain.

Decision Trees

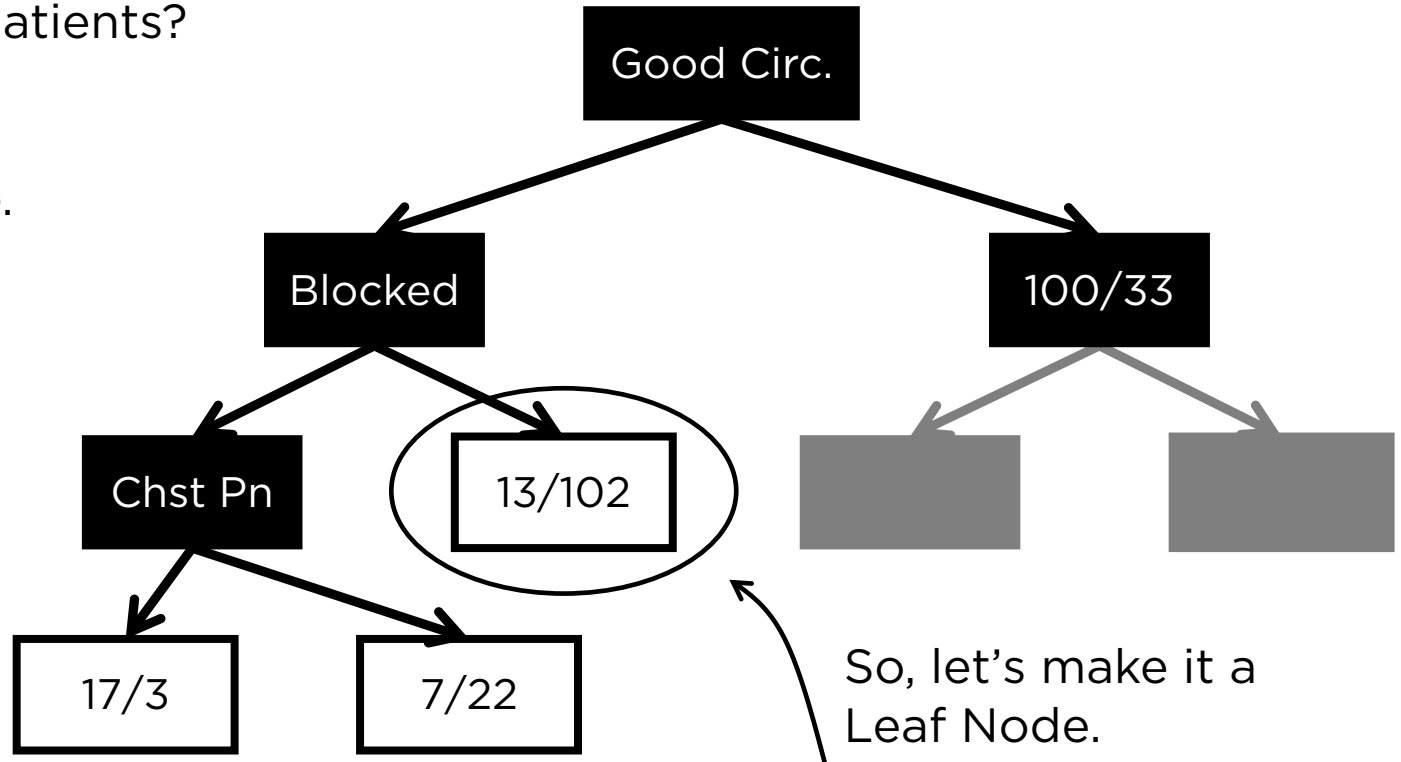
- From a raw table of data to a decision tree

How well **Chest Pain** separate these 115 patients?

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



$$Gini\ Impurity\ for\ \boldsymbol{Chest\ Pain} = 0.19$$



So, let's make it a Leaf Node.

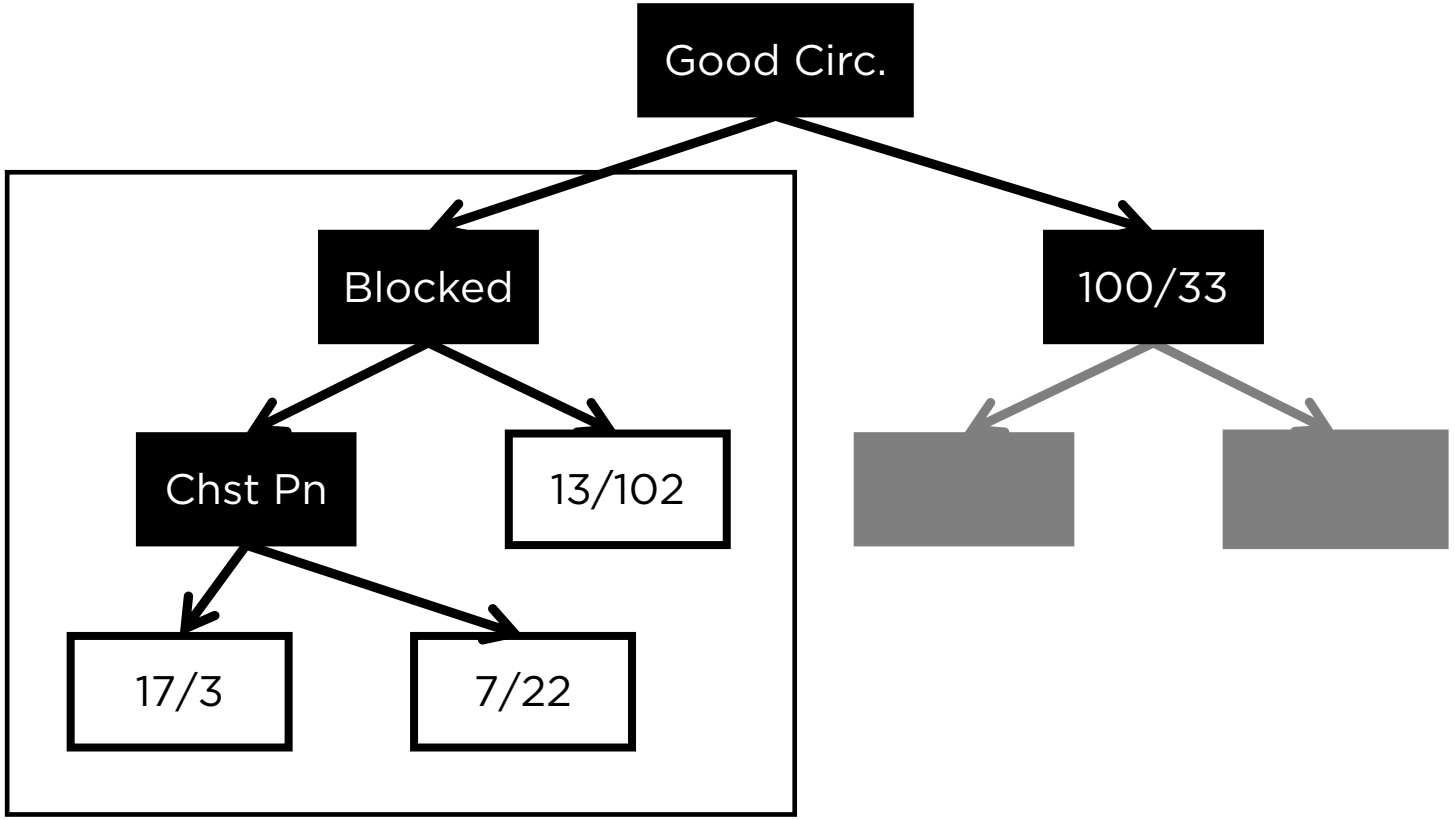
Gini impurity for this node, before using **Chest Pain** to separate patients is...

$$1 - \left(\frac{13}{13+102}\right)^2 - \left(\frac{102}{13+102}\right)^2 = 0.2$$

Decision Trees

- From a raw table of data to a decision tree

Left side of the Tree completed!



Decision Trees

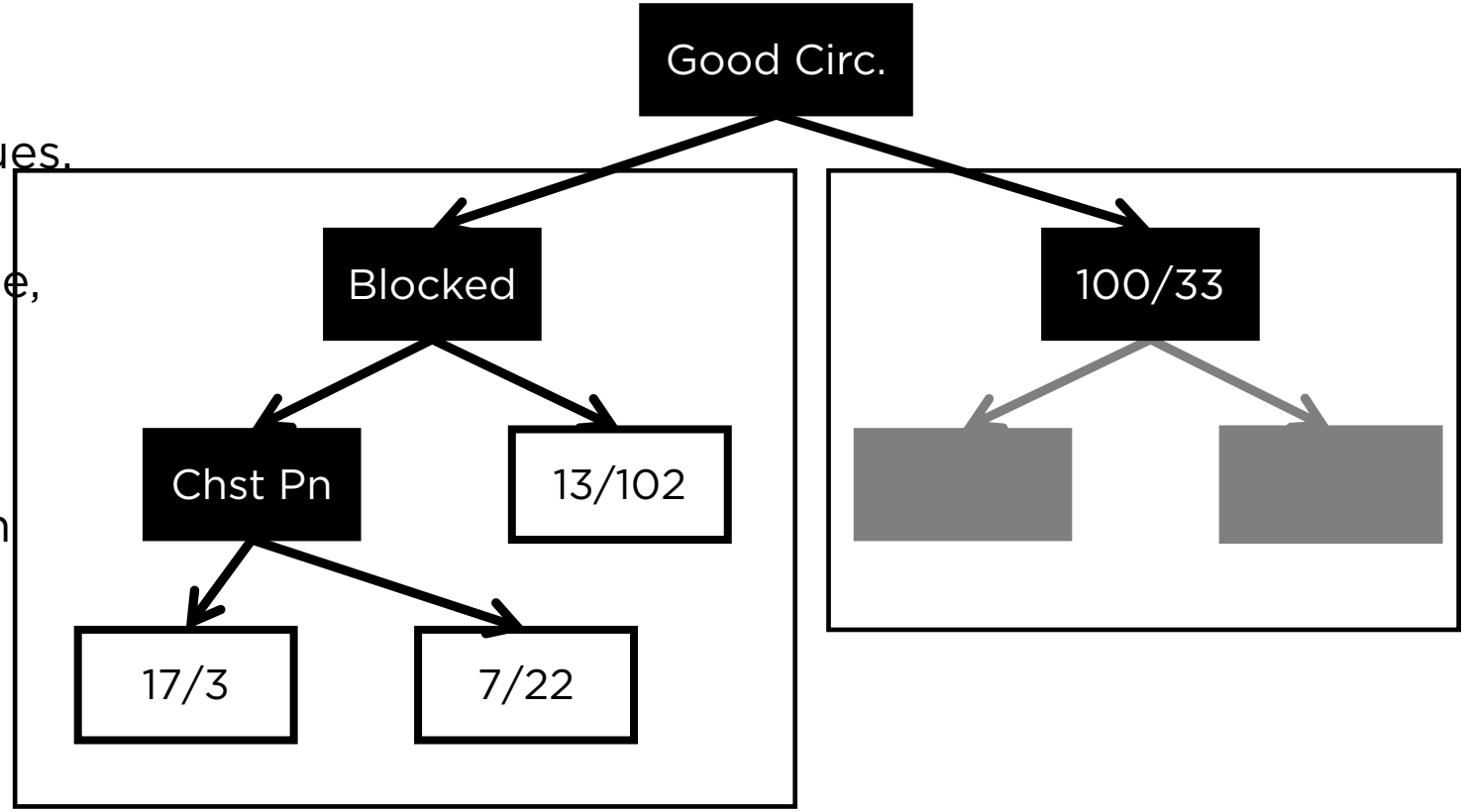
- From a raw table of data to a decision tree

Following the exact same steps:

1. Calculate all of the Gini impurity values.

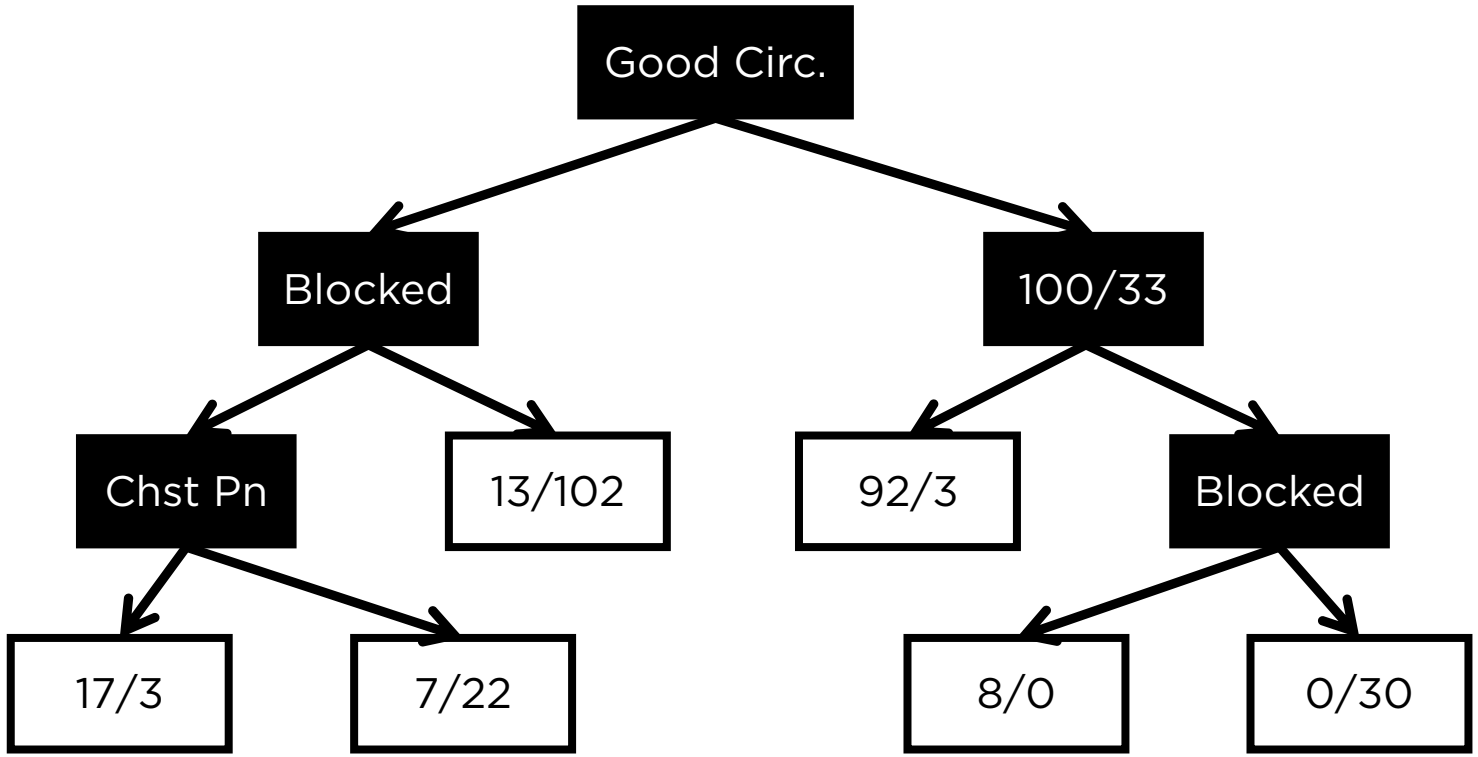
2. If the node itself has the lowest value, leave it as a Leaf node.

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



Decision Trees

- From a raw table of data to a decision tree



Numeric Data?

Decision Trees

- Built tree from **Numeric Data**

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

Decision Trees

- Built tree from **Numeric Data**

lowest

highest

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

Step 1. Sort patients by weight, lowest to highest.

Decision Trees

- Built tree from **Numeric Data**

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.

Decision Trees

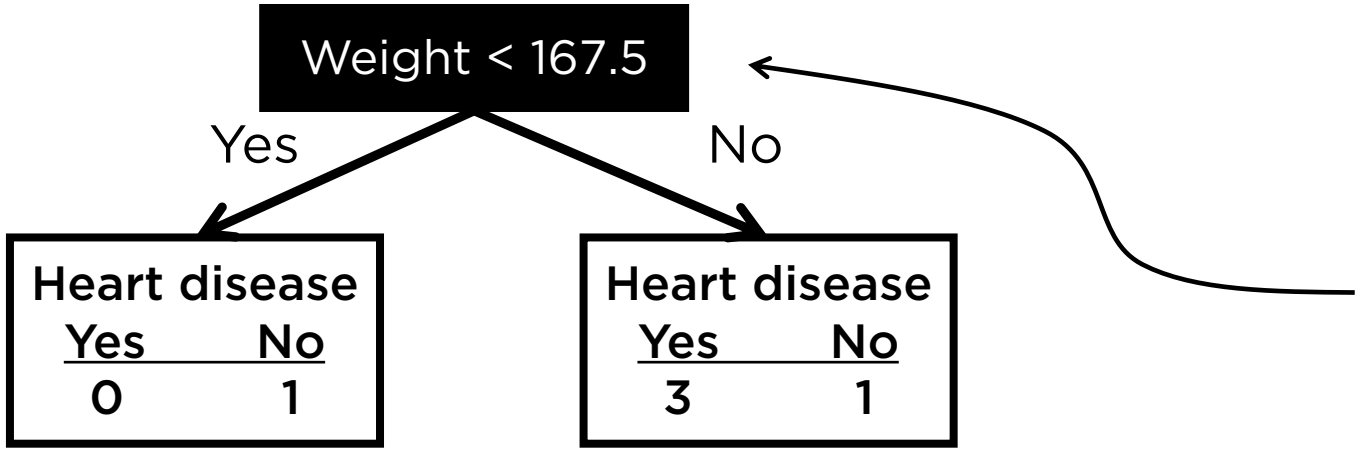
- Built tree from **Numeric Data**

	Weight	Heart disease
	155	No
Gini impurity? ←	**167.5**	
	180	Yes
Gini impurity? ←	**185**	
	190	No
Gini impurity? ←	**205**	
	220	Yes
Gini impurity? ←	**222.5**	
	225	Yes

Step 3. Calculate the impurity values for each average weight.

Decision Trees

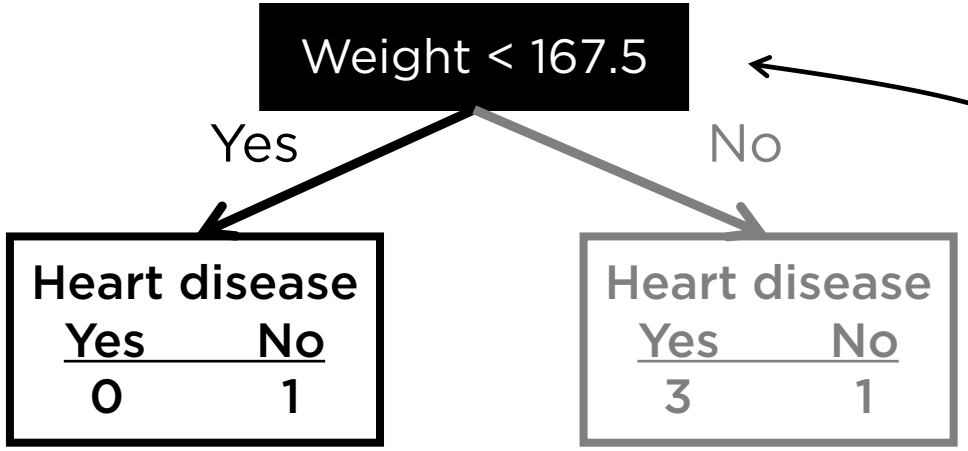
- Built tree from **Numeric Data**



Weight	Heart disease
155	No
167.5	
180	Yes
185	
190	No
205	
220	Yes
222.5	
225	Yes

Decision Trees

- Built tree from **Numeric Data**



Weight	Heart disease
155	No
167.5	
180	Yes
185	
190	No
205	
220	Yes
222.5	
225	Yes

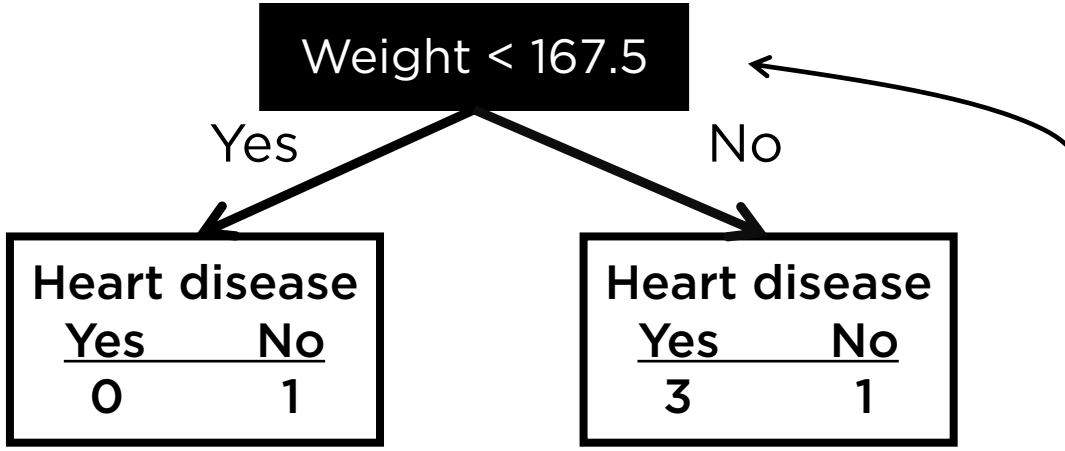
$$Gini\ impurity$$

$$= 1 - (probability\ of\ Yes)^2 - (probability\ of\ No)^2$$

$$= 1 - \left(\frac{0}{0+1}\right)^2 - \left(\frac{1}{0+1}\right)^2 = 0$$

Decision Trees

- Built tree from **Numeric Data**



$Gini\ impurity = 0$ $\quad$ $Gini\ impurity = 0.375$

$$Gini\ impurity\ for\ Weight < 167.5$$

$$= \frac{1}{1+4} \times 0 + \frac{4}{1+4} \times 0.375 = 0.3$$

Weight	Heart disease
155	No
167.5	
180	Yes
185	
190	No
205	
220	Yes
222.5	
225	Yes

Decision Trees

- Built tree from **Numeric Data**

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

Step 3. Calculate the impurity values for each average weight.

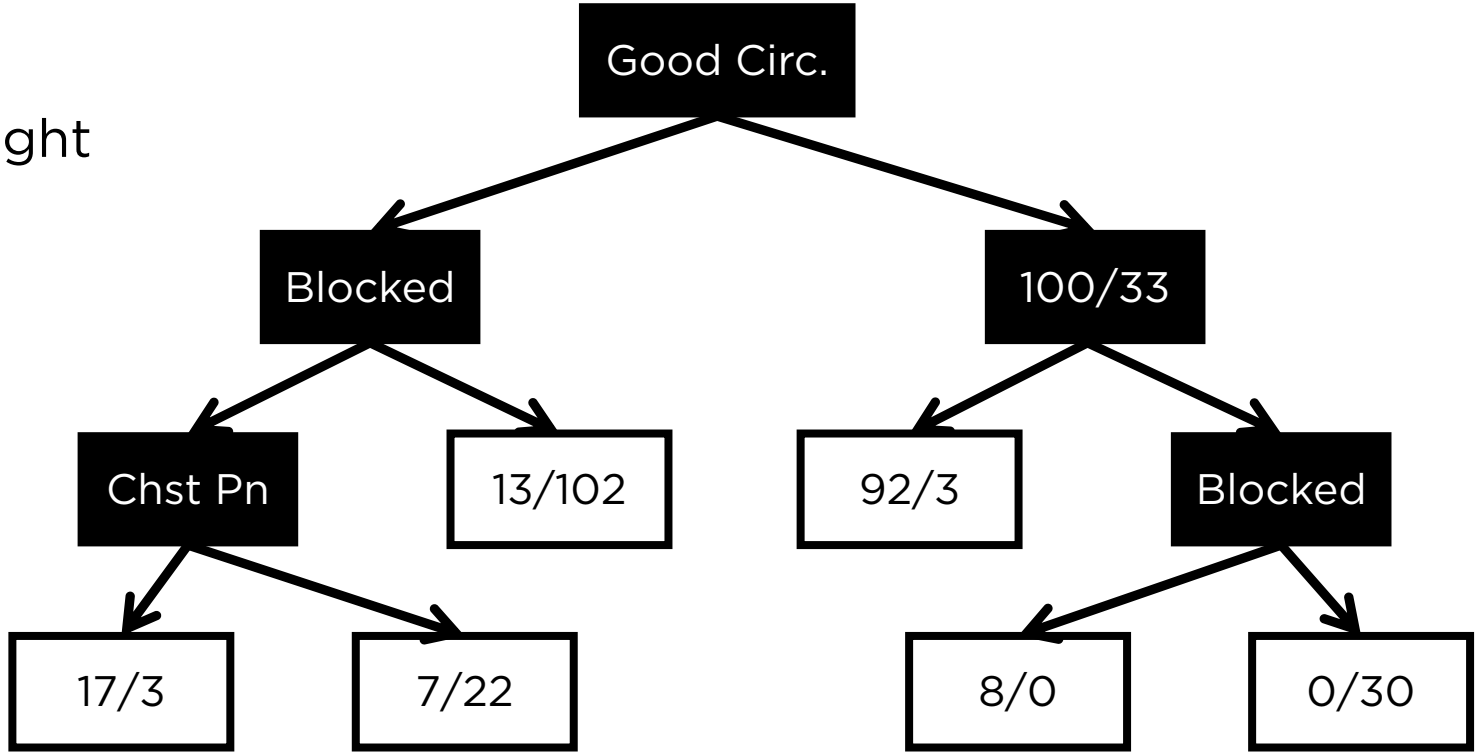
The lowest impurity occurs when using Weight < 205.

This is the cutoff to use when comparing Weight to Chest Pain, or Blocked Arteries.

Decision Trees

- To build a tree

1. yes/no questions at each step
2. Numeric data, like patient Weight



Ranked Data & Multiple Choice Data ?

Decision Trees

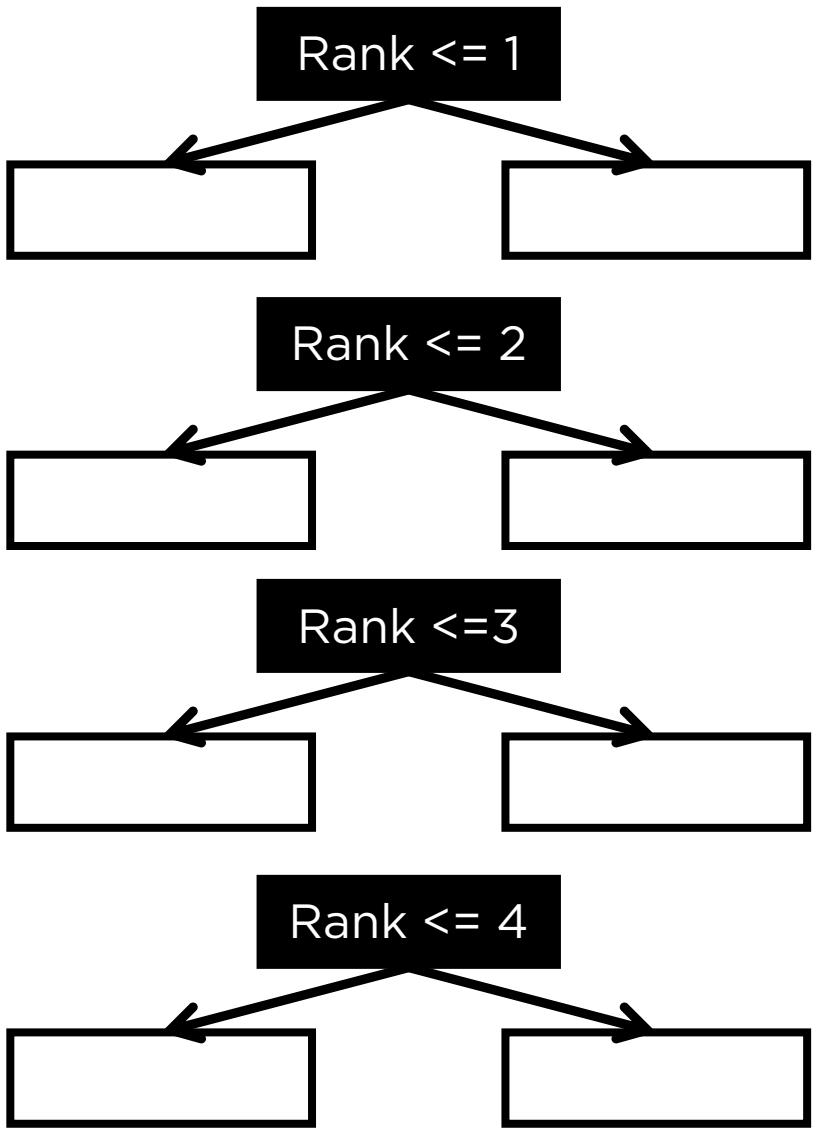
- Built tree from **Ranked Data**

Rank my ML lectures	Pass the exam
2	No
5	Yes
4	Yes
3	No
...	...

Decision Trees

- Built tree from **Ranked Data**

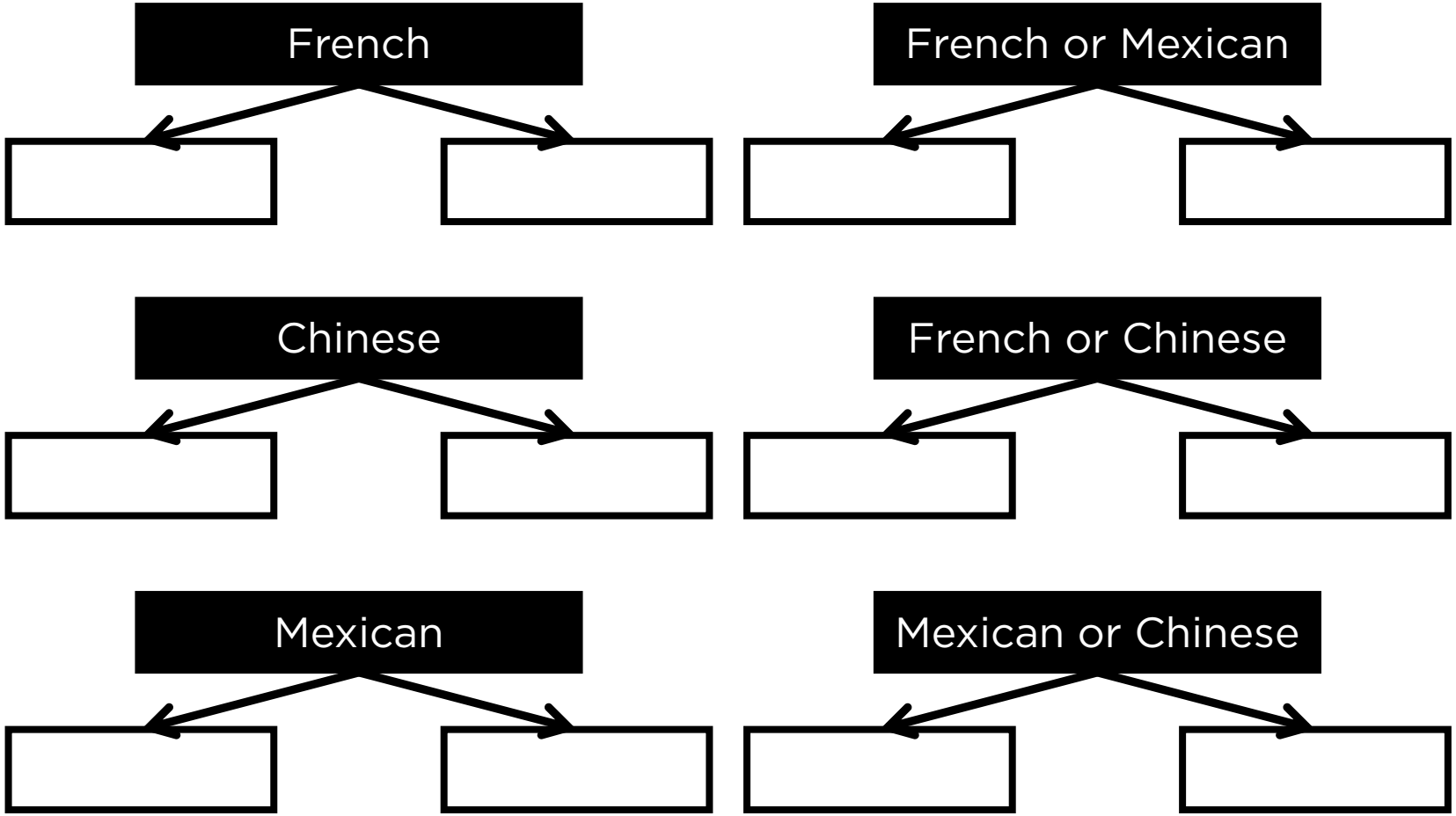
Rank my ML lectures	Pass the exam
2	No
5	Yes
4	Yes
3	No
...	...



Decision Trees

- Built tree from **Multiple Choices Data**

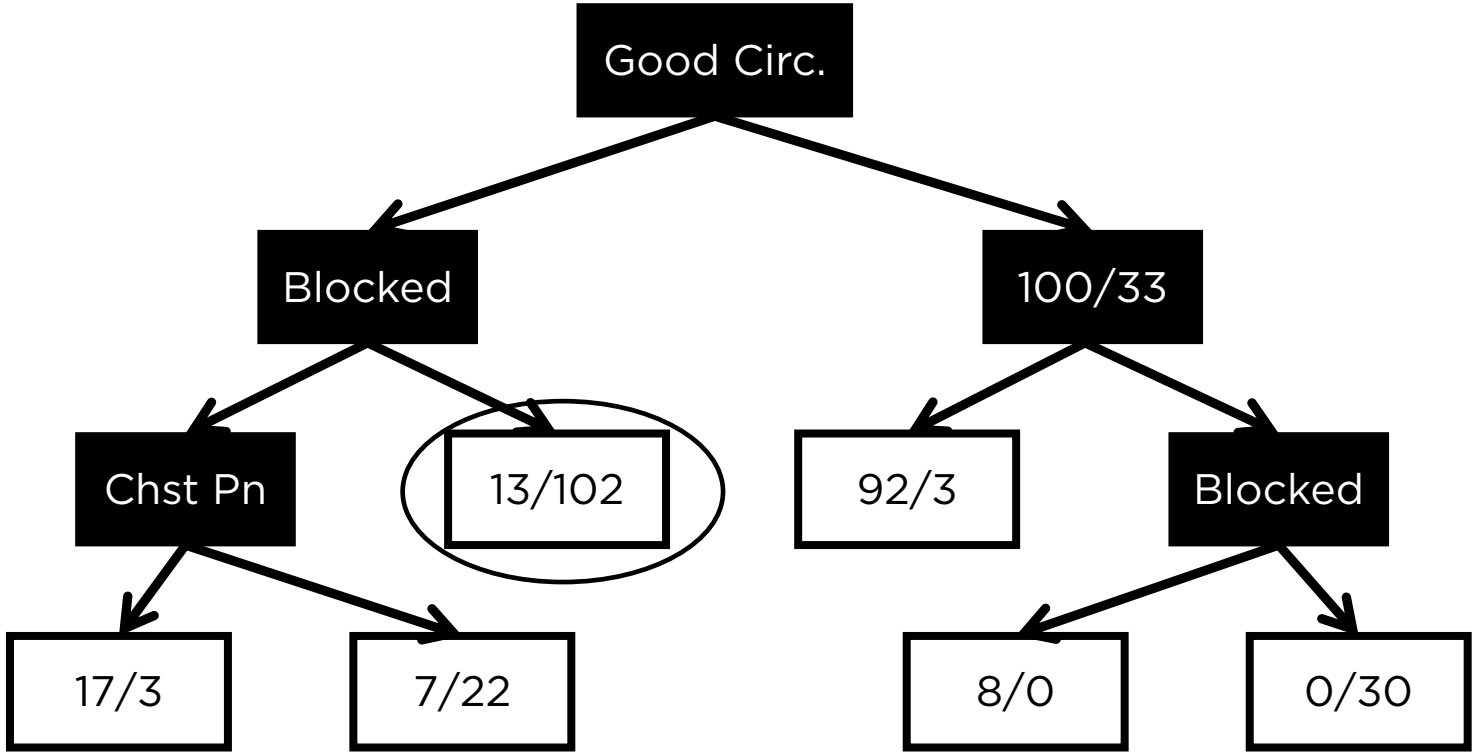
Cuisine Choice	Pass the exam
Mexican	Yes
Chinese	Yes
French	No
Chinese	Yes
...	...



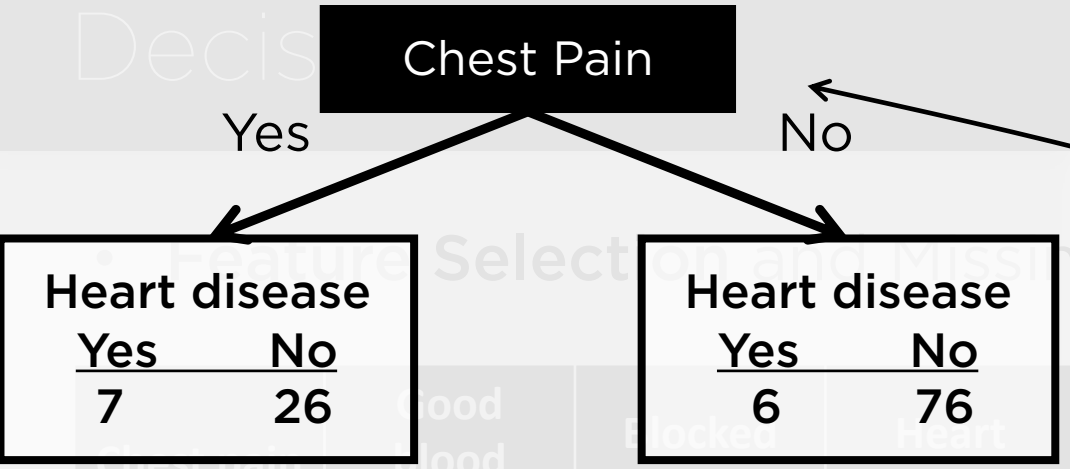
Decision Trees

- **Feature Selection** and Missing Data

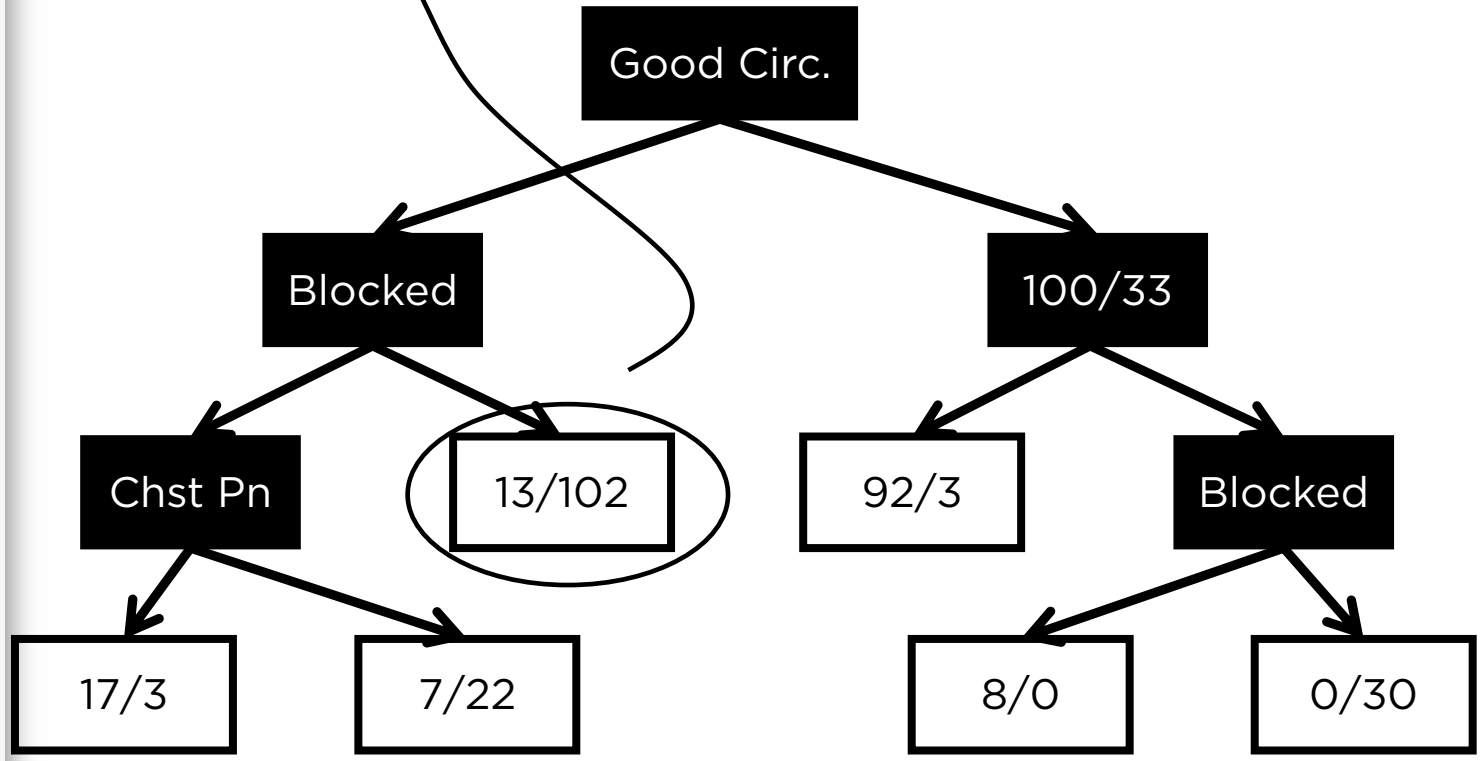
Chest pain	Good blood circulation	Blocked arteries	Heart disease
No	No	No	No
Yes	Yes	Yes	Yes
Yes	Yes	No	No
Yes	No	?	Yes
...	...	...	...



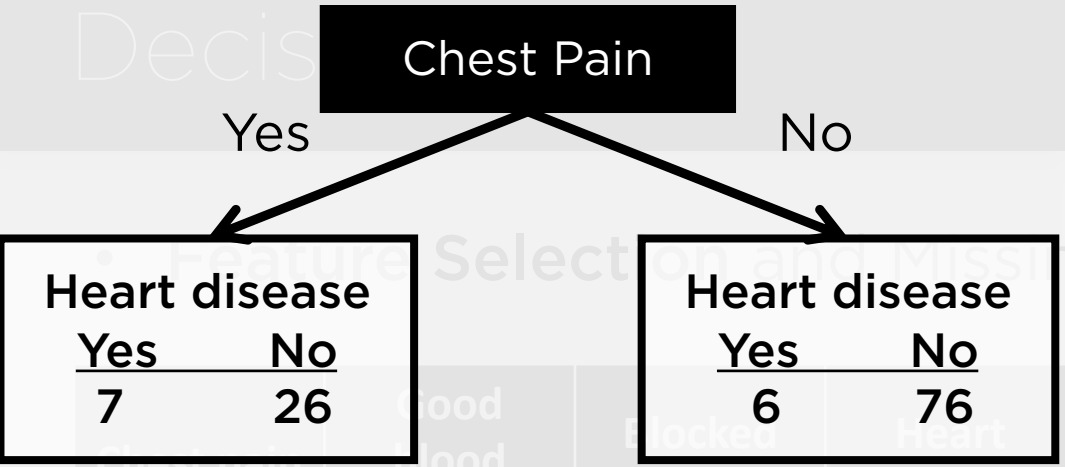
Decis g Data



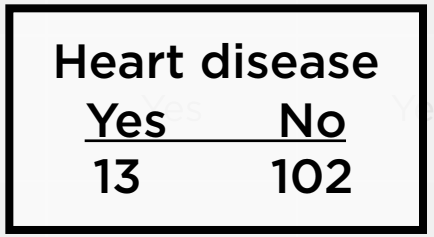
$$Gini\ impurity for\ Chest\ Pain = 0.29$$



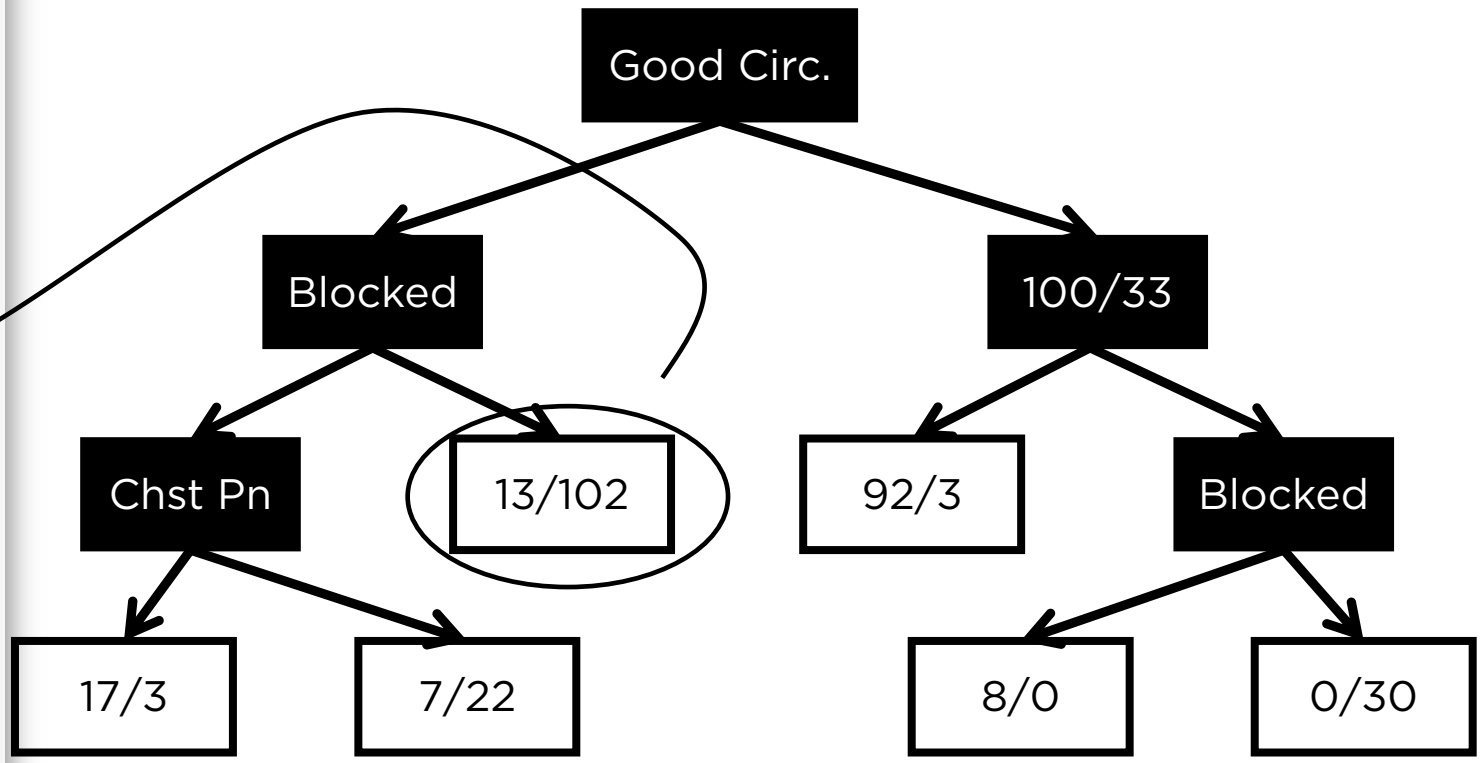
Decis g Data



$$Gini\ impurity for\ Chest\ Pain = 0.29$$



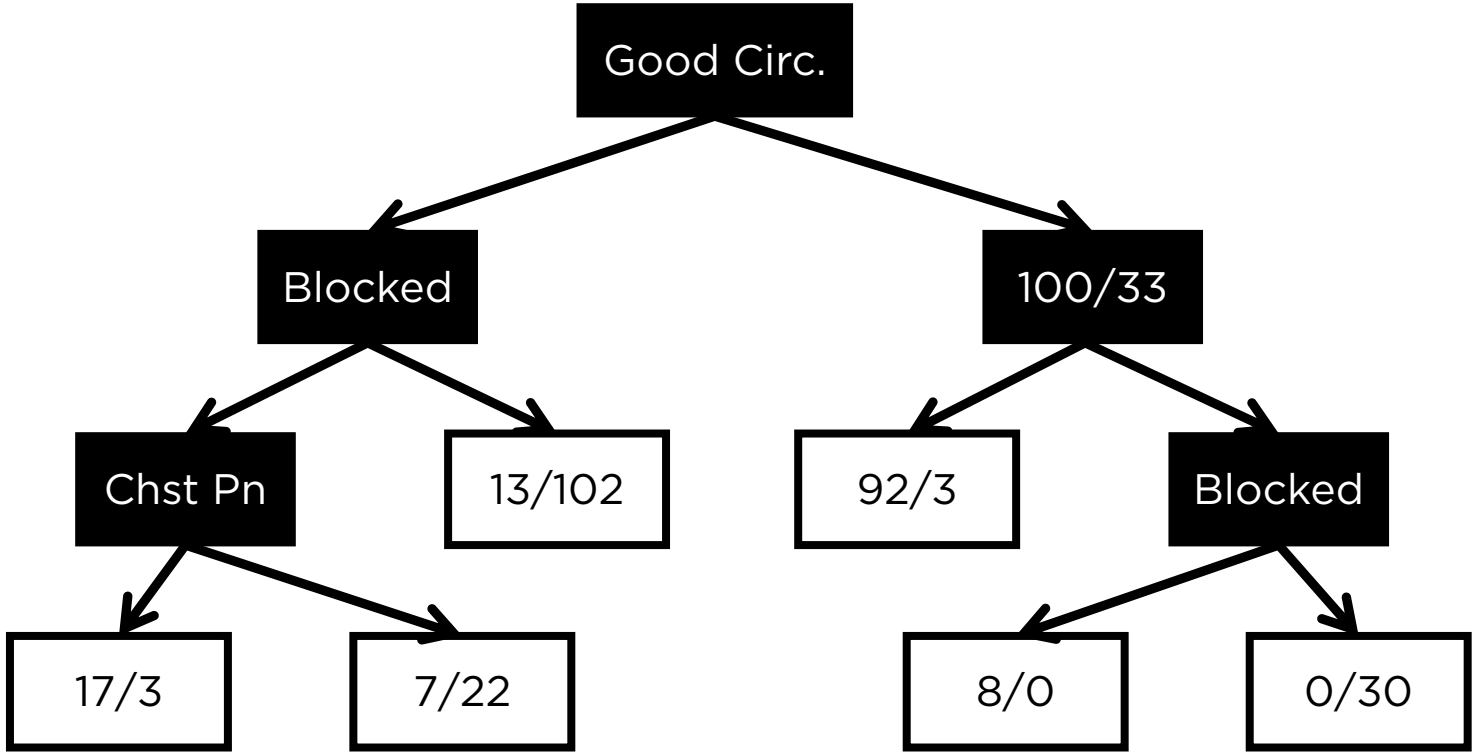
$$Gini\ impurity for\ Chest\ Pain = 0.20$$



Decision Trees

- **Feature Selection** and Missing 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
...	...	...	...



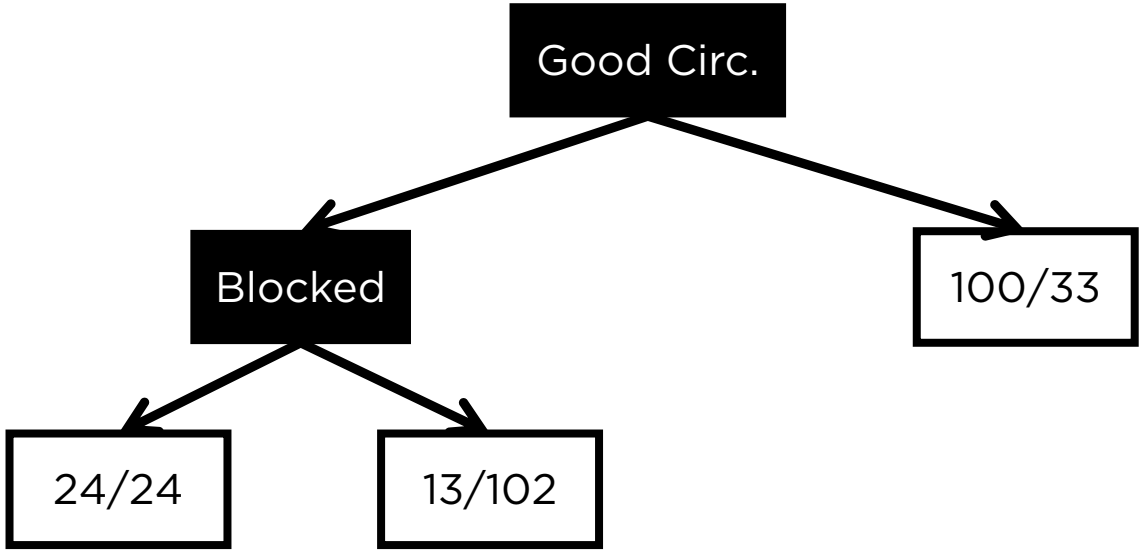
Decision Trees

- **Feature Selection** and Missing Data

If Chest Pain does not reduce impurity score,

It would not be used to separate patients, and it would not be part of our tree.

Even though we have data for Chest Pain, it is no longer part of our tree.



Automatic feature selection

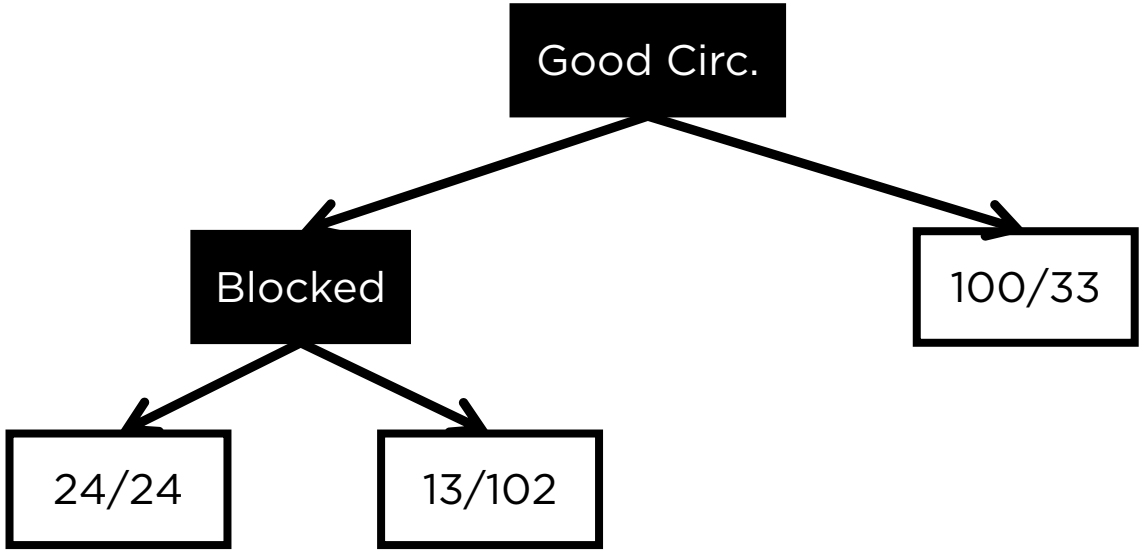
Decision Trees

Over fit: the tree does well the original data – the data used to build the tree – but does not do well with any other data set.

Decision Trees have the downside of often being over fit.

Need each split to make a large reduction in impurity.

- **Feature Selection** and Missing Data



Automatic feature selection

- Not to “over fit”

Decision Trees

- Feature Selection and **Missing Data**

Chest pain	Good blood circulation	Blocked arteries	Heart disease
No	No	No	No
Yes	Yes	Yes	Yes
No	Yes	No	No
Yes	No	? Skipped	Yes
...	...	...	...

Choose the most common value

"No"

Decision Trees

- Feature Selection and **Missing Data**

Chest pain	Good blood circulation	Blocked arteries	Heart disease
No	No	No	No
Yes	Yes	Yes	Yes
No	Yes	No	No
Yes	No	?	Yes
...	...	...	...

Find another column that has the highest correlation with blocked arteries and use that as a guide.

"Yes"

Decision Trees

- Feature Selection and **Missing Data**

Chest pain	Good blood circulation	Weight	Heart disease
No	No	162	No
Yes	Yes	178	Yes
No	Yes	156	No
Yes	No	?	Yes
...	...	...	...

Weight data

- Mean / median

Decision Trees

- Feature Selection and **Missing Data**

Height	Good blood circulation	Weight	Heart disease
170cm	No	162	No
182cm	Yes	178	Yes
168cm	Yes	156	No
176cm	No	?	Yes
...	...	...	...

Find another column that has the highest correlation with weight and use that as a guide.

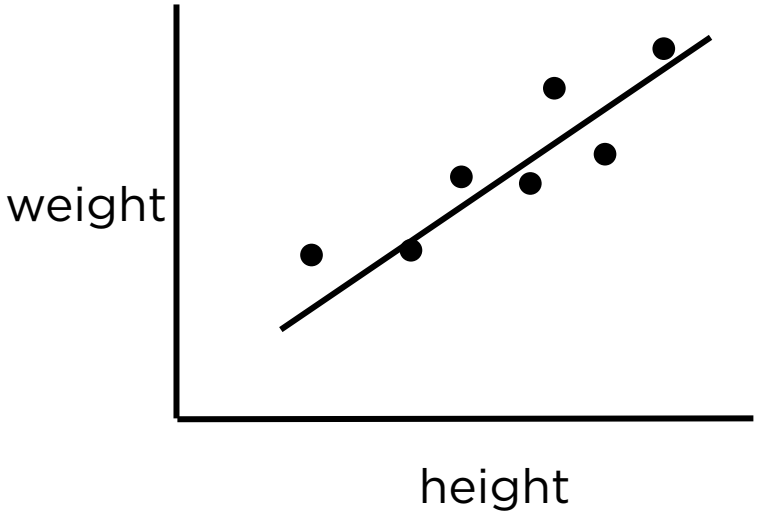
Decision Trees

- Feature Selection and **Missing Data**

Height	Good blood circulation	Weight	Heart disease
170cm	No	162	No
182cm	Yes	198	Yes
168cm	Yes	156	No
176cm	No	?	Yes
...	...	...	...

Find another column that has the highest correlation with weight and use that as a guide.

Do a linear regression on the two columns



Decision Trees

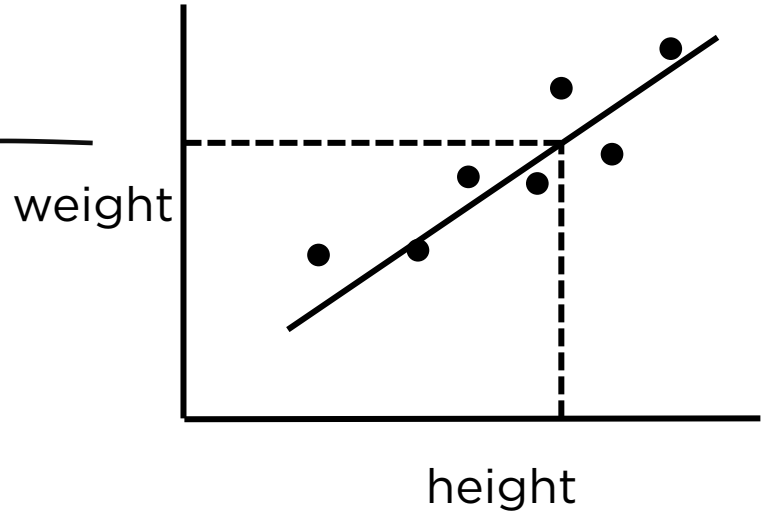
- Feature Selection and **Missing Data**

Height	Good blood circulation	Weight	Heart disease
170cm	No	162	No
182cm	Yes	198	Yes
168cm	Yes	156	No
176cm	No	180	Yes
...	...	...	...

Find another column that has the highest correlation with weight and use that as a guide.

Do a linear regression on the two columns

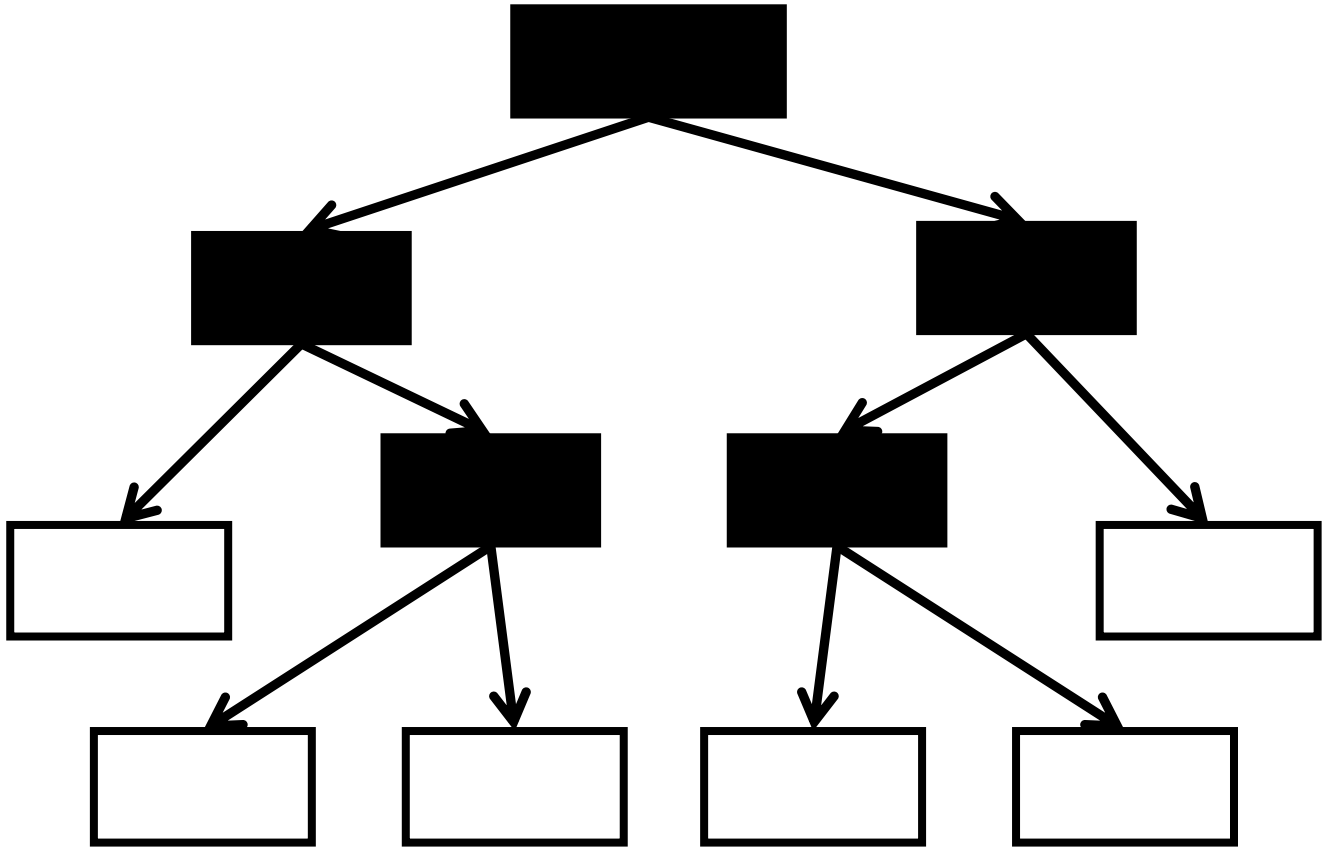
Use least squares line to predict the weight value



Random Forests

Random Forests

Why Random Forests



- Decision Trees are easy to build, use and interpret, but...
- Decision Trees are not flexible when classifying new samples.

  -> accuracy not so good.
- Random Forests combine the simplicity of Decision Trees with flexibility -> **better accuracy**.

How to build a Random Forest?

Random Forests

How to build a Random Forest?

Step 1. Create a “bootstrapped” dataset.

Original Dataset → **Bootstrapped Dataset**

Chest pain	Good blood circulation	Weight	Heart disease
No	No	162	No
Yes	Yes	178	Yes
Yes	Yes	156	No
Yes	No	180	Yes

Chest pain	Good blood circulation	Weight	Heart disease

- Same size as the original dataset.
- Randomly selected samples from the original dataset.
- Samples can be selected more than once.

Random Forests

How to build a Random Forest?

Step 1. Create a “bootstrapped” dataset.

Original Dataset → **Bootstrapped Dataset**

Chest pain	Good blood circulation	Weight	Heart disease
No	No	162	No
Yes	Yes	178	Yes
Yes	Yes	156	No
Yes	No	180	Yes

Chest pain	Good blood circulation	Weight	Heart disease
Yes	Yes	156	No

Random Forests

How to build a Random Forest?

Step 1. Create a “bootstrapped” dataset.

Original Dataset → **Bootstrapped Dataset**

Chest pain	Good blood circulation	Weight	Heart disease
No	No	162	No
Yes	Yes	178	Yes
Yes	Yes	156	No
Yes	No	180	Yes

→

Chest pain	Good blood circulation	Weight	Heart disease
Yes	Yes	156	No
Yes	Yes	178	Yes

Random Forests

How to build a Random Forest?

Step 1. Create a “bootstrapped” dataset.

Original Dataset → **Bootstrapped Dataset**

Chest pain	Good blood circulation	Weight	Heart disease
No	No	162	No
Yes	Yes	178	Yes
Yes	Yes	156	No
Yes	No	180	Yes

Chest pain	Good blood circulation	Weight	Heart disease
Yes	Yes	156	No
Yes	Yes	178	Yes
Yes	No	180	Yes

Random Forests

How to build a Random Forest?

Step 1. Create a “bootstrapped” dataset.

Original Dataset → **Bootstrapped Dataset**

Chest pain	Good blood circulation	Weight	Heart disease
No	No	162	No
Yes	Yes	178	Yes
Yes	Yes	156	No
Yes	No	180	Yes

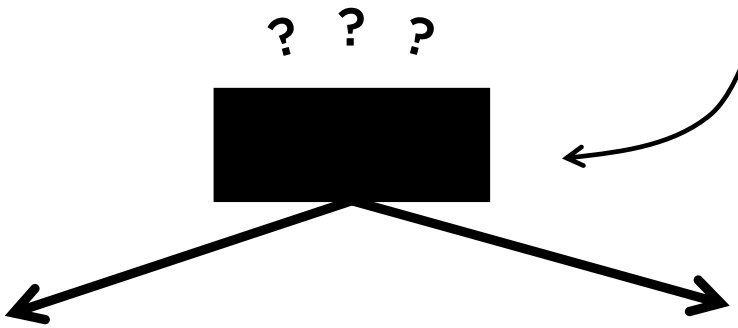
Chest pain	Good blood circulation	Weight	Heart disease
Yes	Yes	156	No
Yes	Yes	178	Yes
Yes	No	180	Yes
Yes	No	180	Yes

Same!

Random Forests

How to build a Random Forest?

Step 2. Build a Decision Tree using the “bootstrapped” dataset, but only use a random subset of variables, e.g. 2



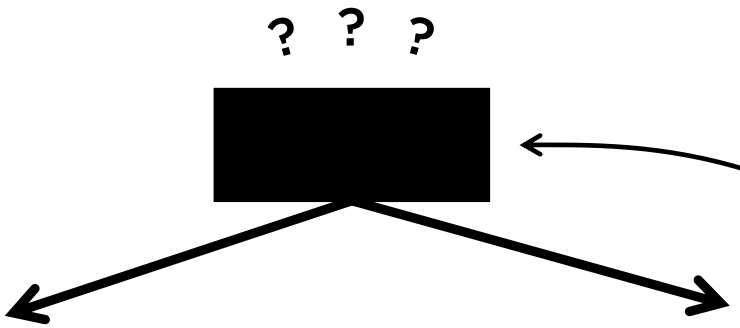
Bootstrapped Dataset

Chest pain	Good blood circulation	Weight	Heart disease
Yes	Yes	156	No
Yes	Yes	178	Yes
Yes	No	180	Yes
Yes	No	180	Yes

Random Forests

How to build a Random Forest?

Step 2. Build a Decision Tree using the “bootstrapped” dataset, but only use a random subset of variables, e.g. 2



Bootstrapped Dataset

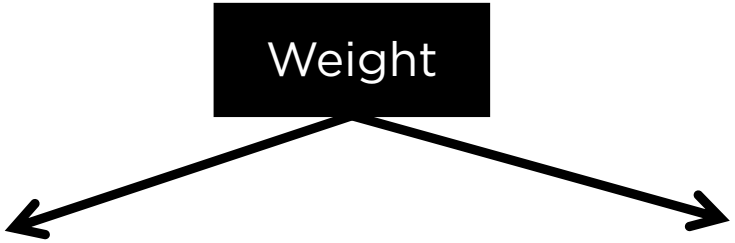
Chest pain	Good blood circulation	Weight	Heart disease
Yes	Yes	156	No
Yes	Yes	178	Yes
Yes	No	180	Yes
Yes	No	180	Yes

Random Forests

How to build a Random Forest?

Step 2. Build a Decision Tree using the "bootstrapped" dataset, but only use a random subset of variables, e.g. 2

Weight did the best job separating the samples.



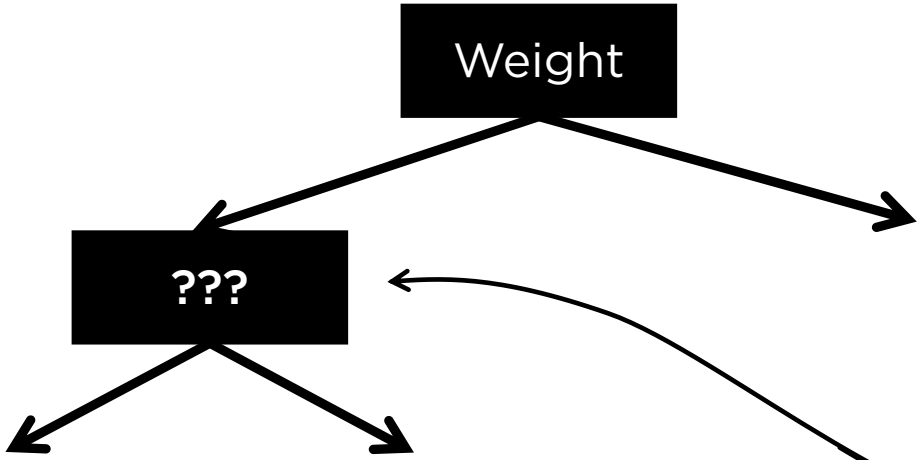
Bootstrapped Dataset

Chest pain	Good blood circulation	Weight	Heart disease
Yes	Yes	156	No
Yes	Yes	178	Yes
Yes	No	180	Yes
Yes	No	180	Yes

Random Forests

How to build a Random Forest?

Step 2. Build a Decision Tree using the “bootstrapped” dataset, but only use a random subset of variables, e.g. 2



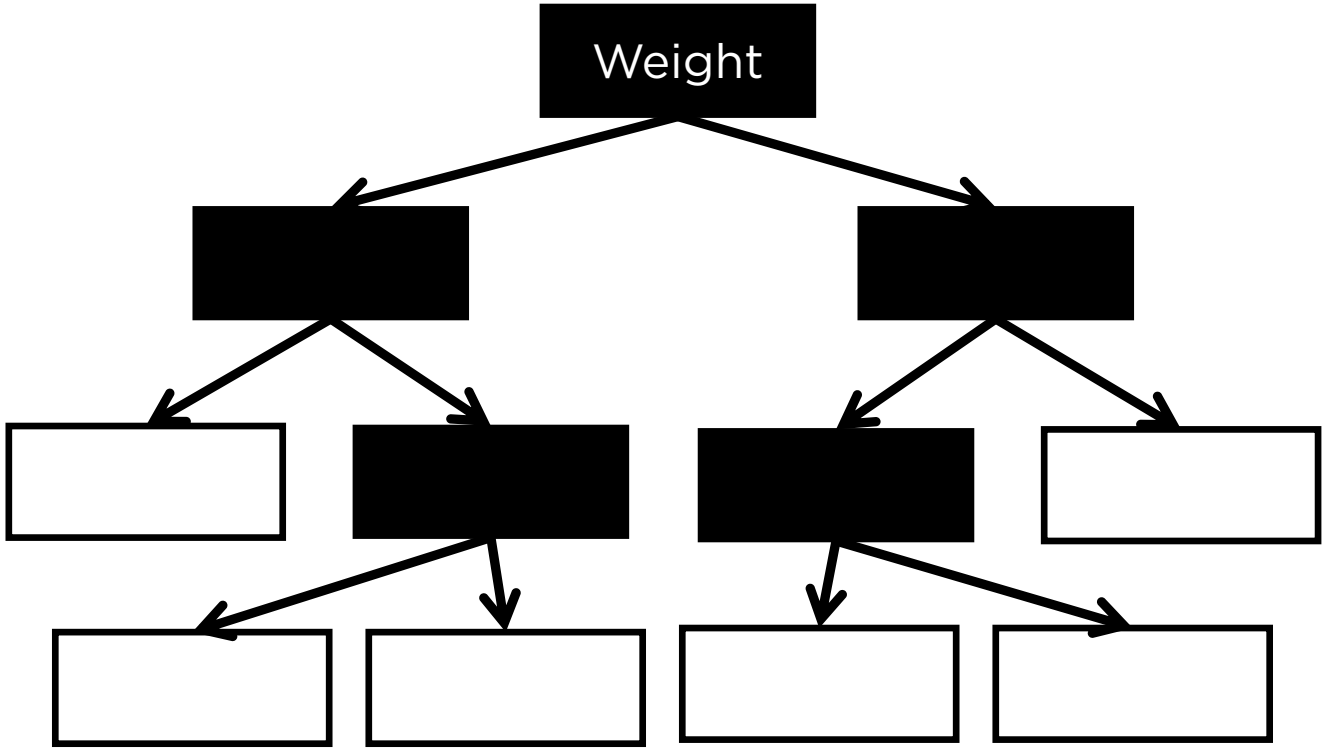
Bootstrapped Dataset

Chest pain	Good blood circulation	Weight	Heart disease
Yes	Yes	156	No
Yes	Yes	178	Yes
Yes	No	180	Yes
Yes	No	180	Yes

Random Forests

How to build a Random Forest?

Step 2. Build a Decision Tree using the “bootstrapped” dataset, but only use a random subset of variables, e.g. 2



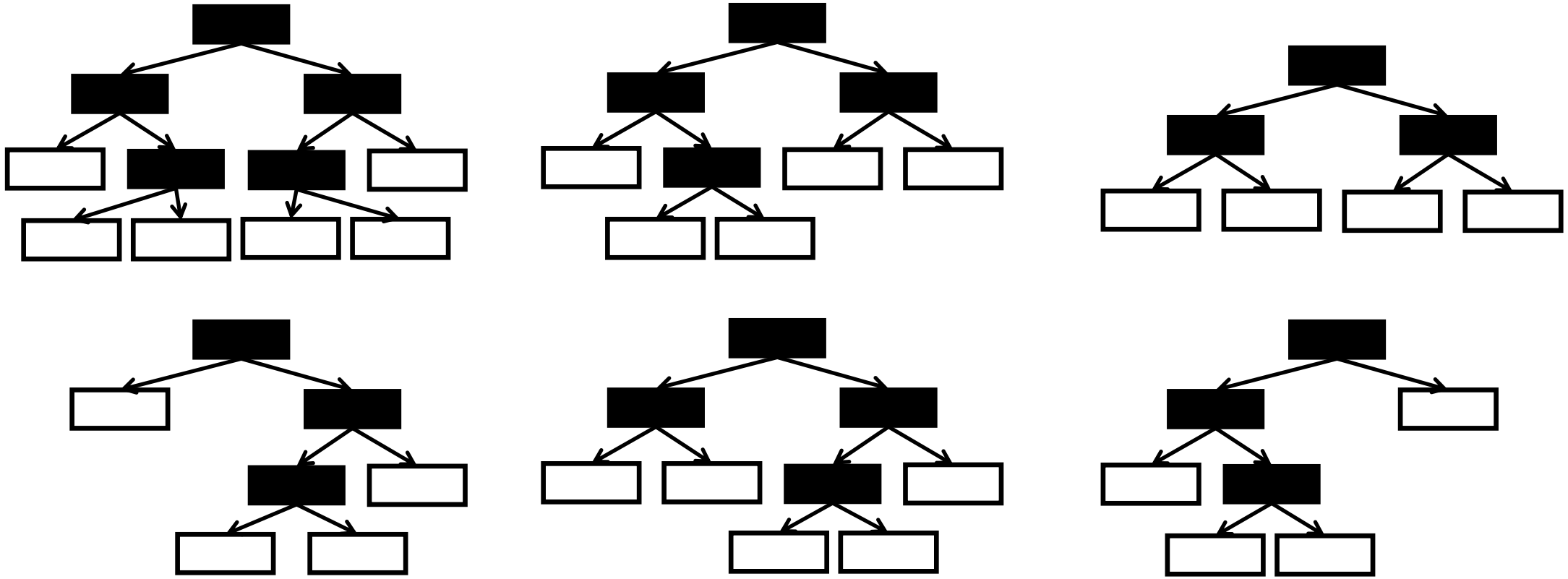
Bootstrapped Dataset

Chest pain	Good blood circulation	Weight	Heart disease
Yes	Yes	156	No
Yes	Yes	178	Yes
Yes	No	180	Yes
Yes	No	180	Yes

Random Forests

How to build a Random Forest?

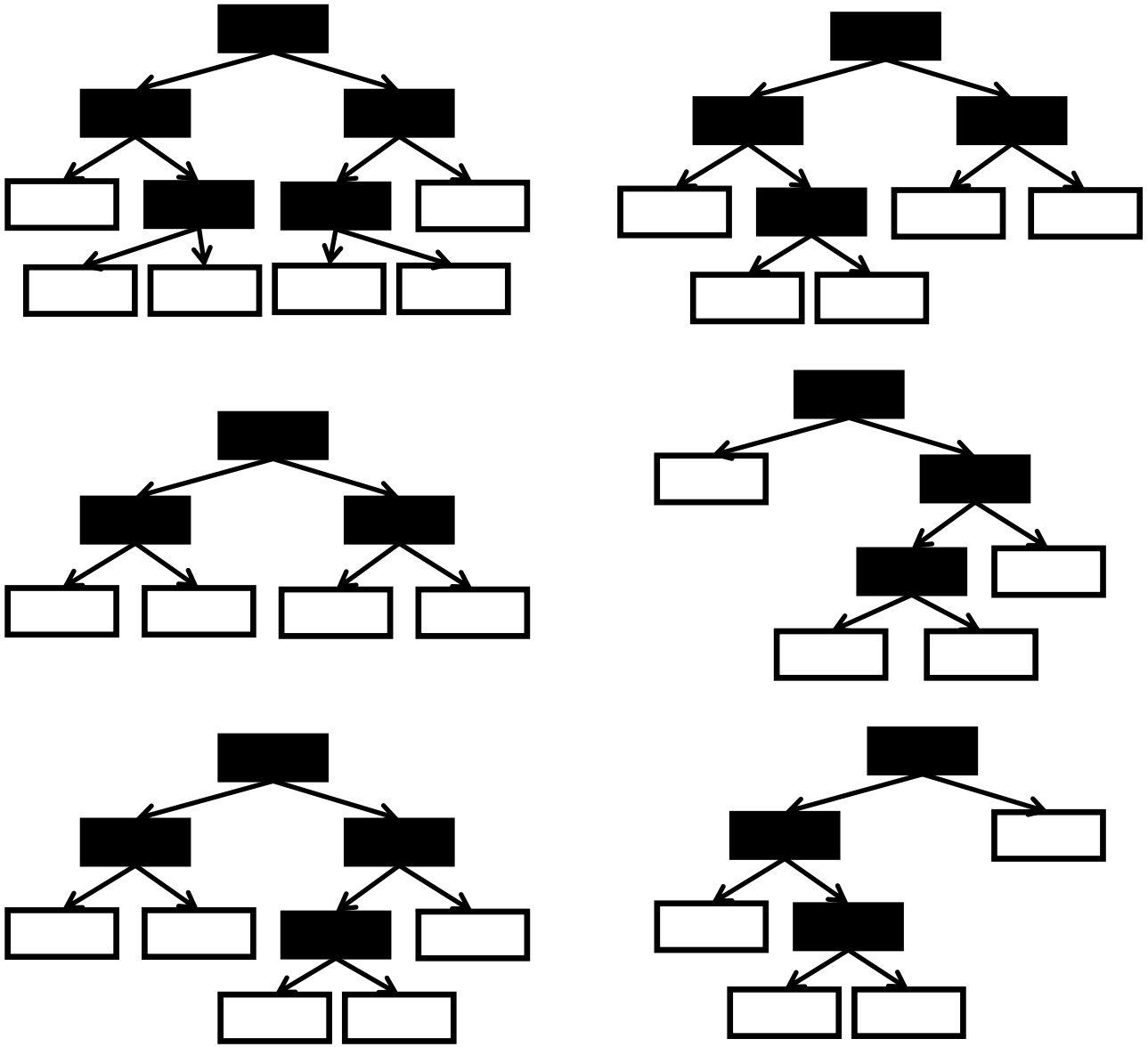
Step 3. Go back to Step 1 and repeat: make a new bootstrap dataset and build a tree considering a subset of variables at each step. (ideally 100's of times)



Random Forests

How to build a Random Forest?

- Using a bootstrapped sample and considering only a subset of the variables at each step results in a wide variety of trees.
- The variety makes Random Forests more effective than individual Decision Trees.



How to use a Random Forest?

Random Forests

How to use a Random Forest?

Get a new patient.

Chest pain	Good blood circulation	Weight	Heart disease
Yes	Yes	156	???

With all the measurements.

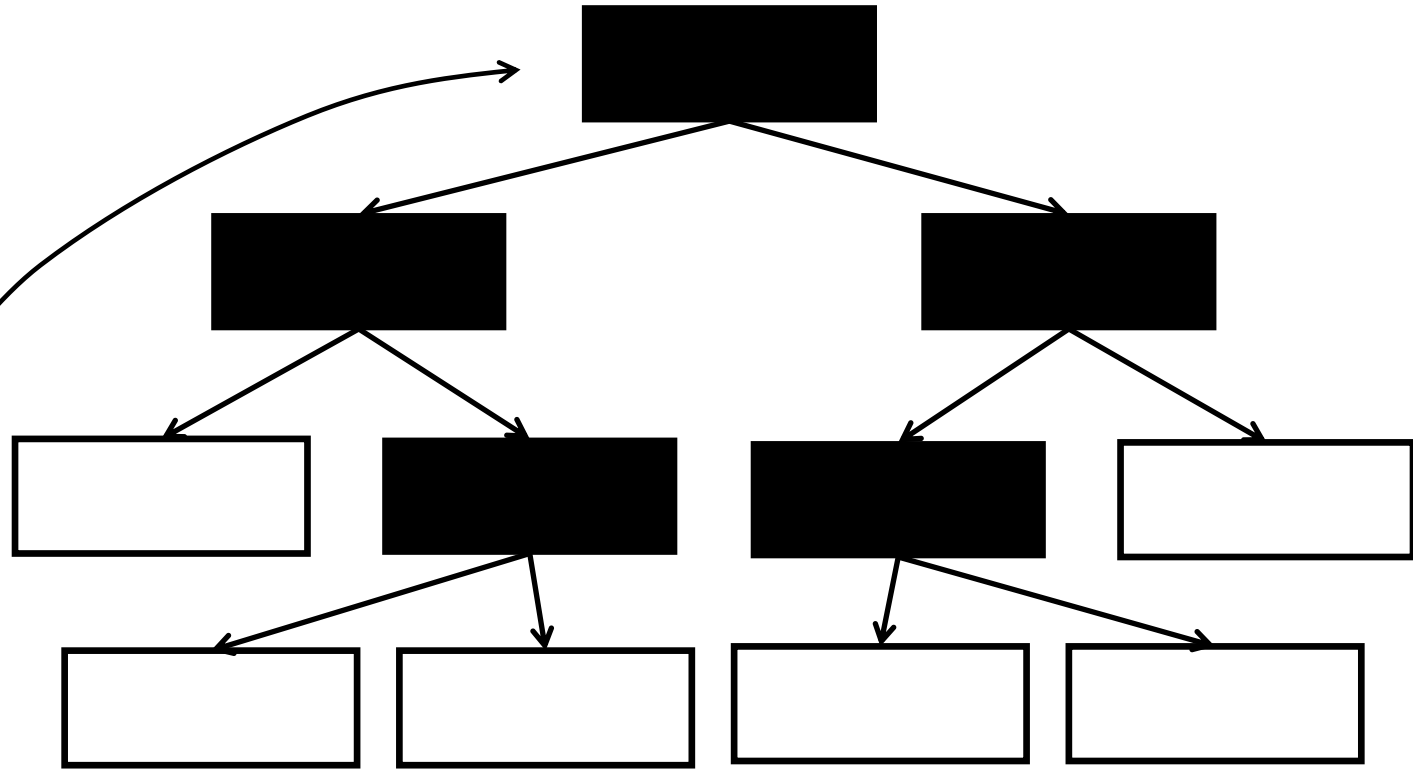
If the patient have heart disease?

Random Forests

How to use a Random Forest?

Get a new patient.

Chest pain	Good blood circulation	Weight	Heart disease
Yes	Yes	156	???



Take the data & run it down the **first tree** we built.

Keep track of the result

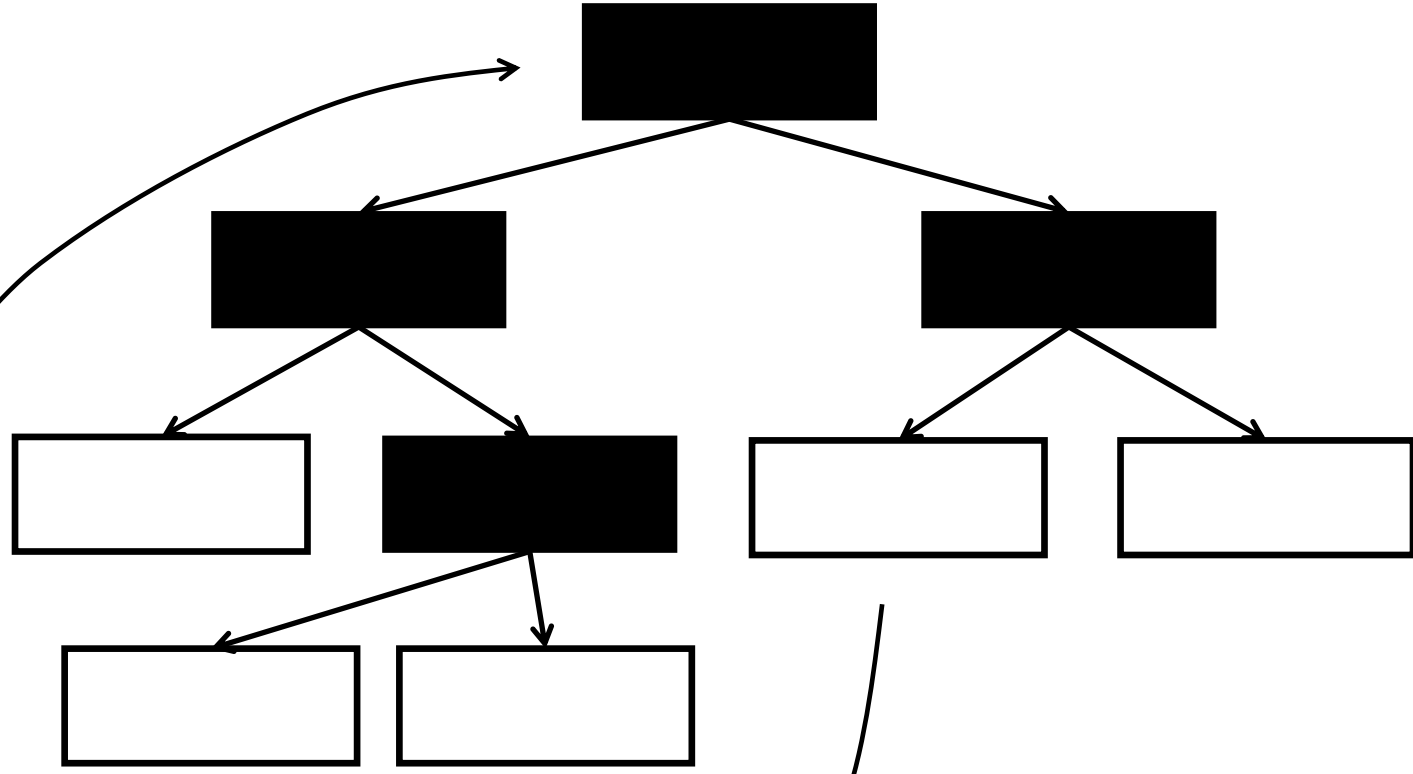
Heart disease	
Yes	No
1	0

Random Forests

How to use a Random Forest?

Get a new patient.

Chest pain	Good blood circulation	Weight	Heart disease
Yes	Yes	156	???



Take the data & run it down the **second tree** we built.

Keep track of the result

Heart disease	
Yes	No
2	0

Random Forests

How to use a Random Forest?

Get a new patient.

Chest pain	Good blood circulation	Weight	Heart disease
Yes	Yes	156	**Yes**

Take the data & run it down the **all the trees** we built.

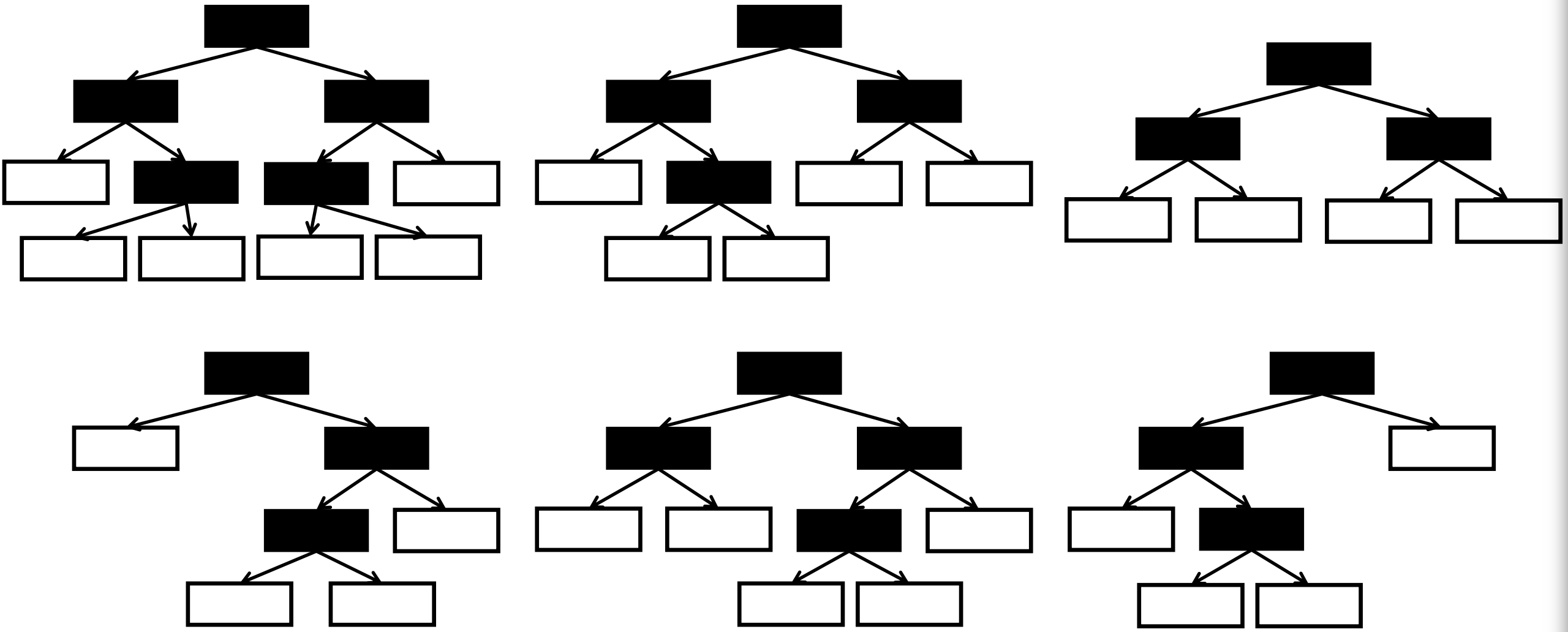
Which option received more votes.

Heart disease	
Yes	No
4	2

Terminology

Bootstrapping the data plus using the **agg**regate to make a decision is called **Bagg**ing.

Random Forest



Is this Random Forest any good?

Random Forests

Typically,

About 1/3 of the original data does NOT end up in the bootstrapped dataset.

Is this Random Forest any good?

Original Dataset

Chest pain	Good blood circulation	Weight	Heart disease
No	No	162	No
Yes	Yes	178	Yes
Yes	Yes	156	No
Yes	No	180	Yes

Not included

Bootstrapped Dataset

Chest pain	Good blood circulation	Weight	Heart disease
Yes	Yes	156	No
Yes	Yes	178	Yes
Yes	No	180	Yes
Yes	No	180	Yes

Duplicate entries

Random Forests

Is this Random Forest any good?

Original Dataset

Chest pain	Good blood circulation	Weight	Heart disease
No	No	162	No
Yes	Yes	178	Yes
Yes	Yes	156	No
Yes	No	180	Yes

Not included

"Out-Of-Bag" Dataset

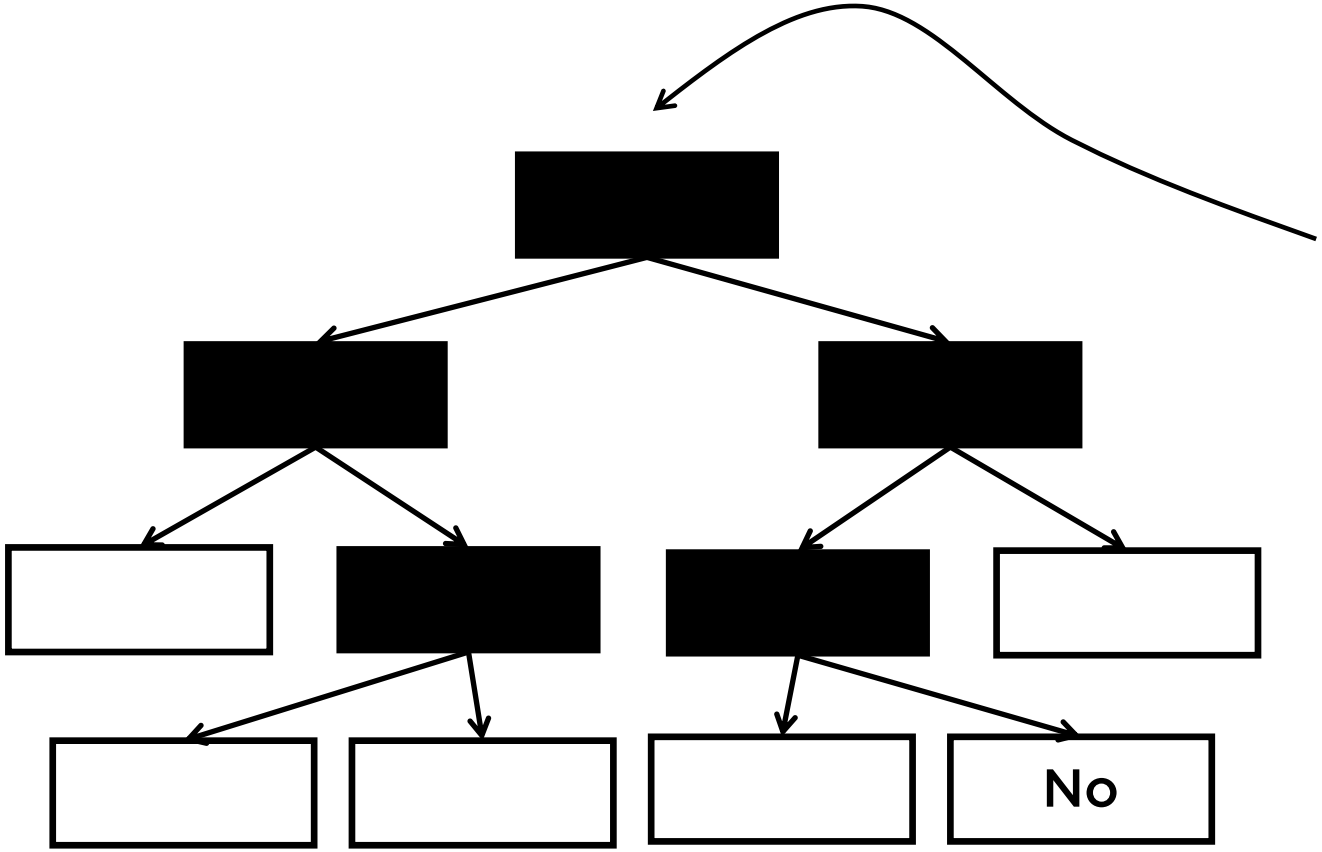
Chest pain	Good blood circulation	Weight	Heart disease
No	No	162	No

Entries Not in the Bootstrapped Dataset

There would be more entry, if the original dataset were larger.

Random Forests

Is this Random Forest any good?



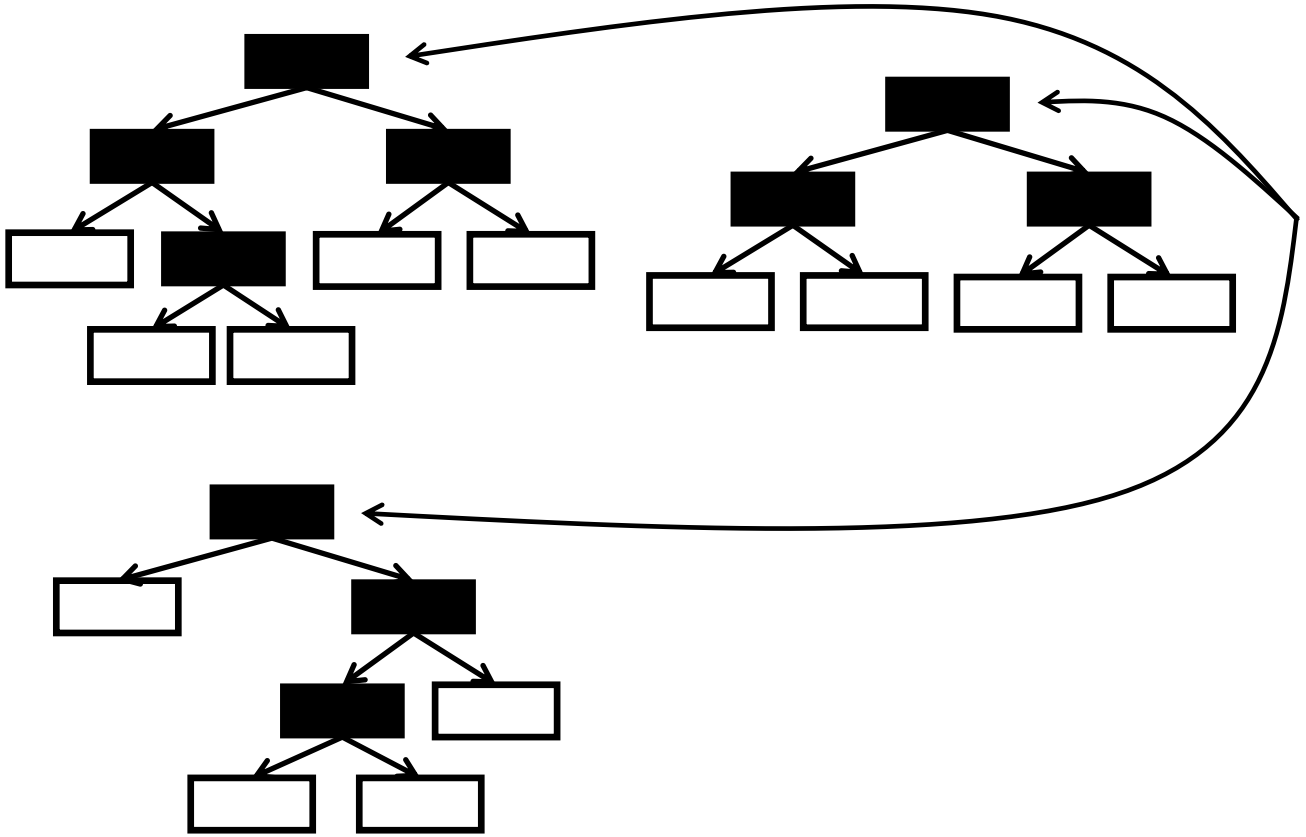
"Out-Of-Bag" Dataset

Chest pain	Good blood circulation	Weight	Heart disease
No	No	162	No

Run the data through and see if it correctly classifies the sample as "No Heart Disease"

Random Forests

Is this Random Forest any good?



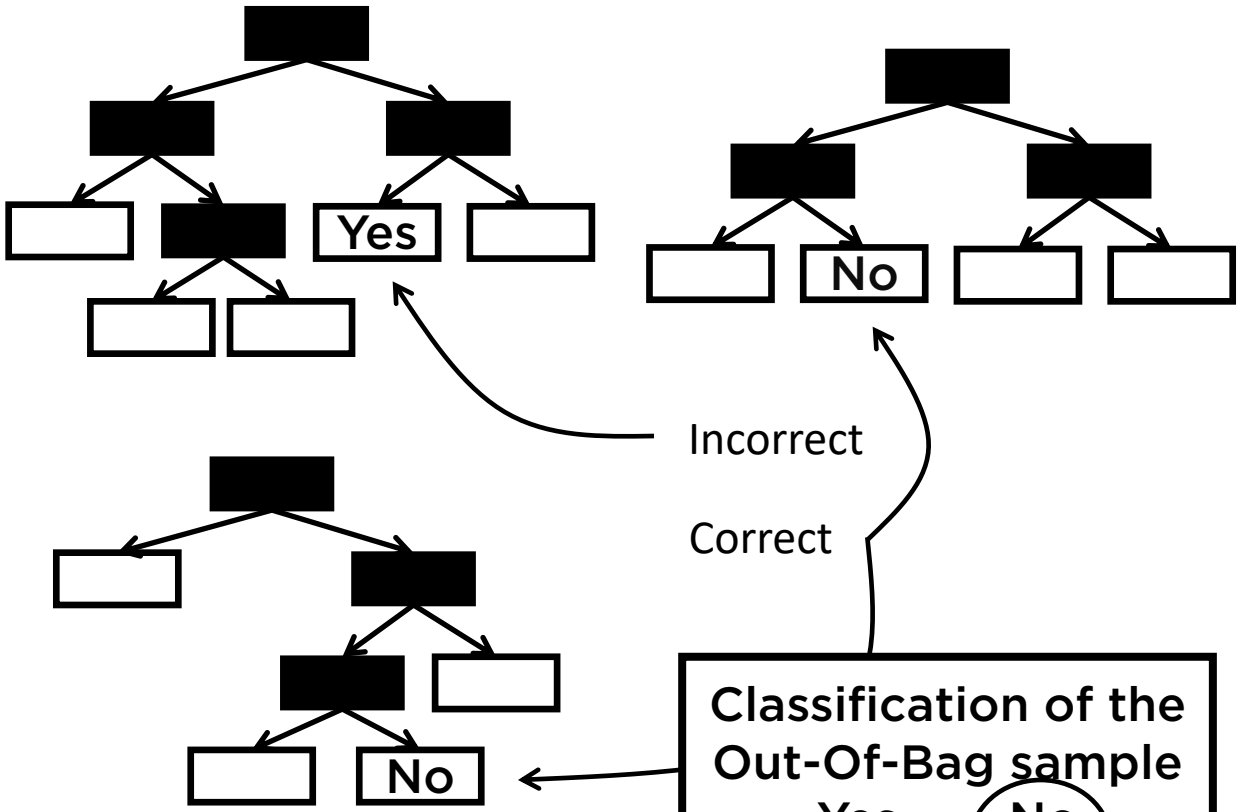
"Out-Of-Bag" Dataset

Chest pain	Good blood circulation	Weight	Heart disease
No	No	162	No

Then, run this Out-Of-Bag sample through all of the other trees that were built without it.

Random Forests

Is this Random Forest any good?



Classification of the Out-Of-Bag sample

Yes	No
1	3

"Out-Of-Bag" Dataset

Chest pain	Good blood circulation	Weight	Heart disease
No	No	162	No

Then, run this Out-Of-Bag sample through all of the other trees that were built without it.

This sample is correctly labeled by the Random Forest.

Random Forests

Is this Random Forest any good?

Classification of the Out-Of-Bag sample

Yes	No
1	3

Classification of the Out-Of-Bag sample

Yes	No
4	0

Correct

"Out-Of-Bag" Dataset

Chest pain	Good blood circulation	Weight	Heart disease
Yes	No	156	Yes

Then, do the same thing for all of the other Out-Of-Bag samples for all of the trees.

Random Forests

Is this Random Forest any good?

Classification of the Out-Of-Bag sample

Yes	No
1	3

Classification of the Out-Of-Bag sample

Yes	No
4	0

Classification of the Out-Of-Bag sample

Yes	No
1	3

Etc. ...

Incorrect

"Out-Of-Bag" Dataset

Chest pain	Good blood circulation	Weight	Heart disease
Yes	Yes	176	Yes

Then, do the same thing for all of the other Out-Of-Bag samples for all of the trees.

Random Forests

Is this Random Forest any good?

Classification of the Out-Of-Bag sample

Yes	No
1	3

Classification of the Out-Of-Bag sample

Yes	No
4	0

Classification of the Out-Of-Bag sample

Yes	No
1	3

Etc. ...

- Measure **accuracy** our Random Forest is by the proportion of Out-Of-Bag samples that were correctly classified by the Random Forest.
- The proportion of Out-Of-Bag samples that were incorrectly classified is the "**Out-Of-Bag Error**"

Random Forests

Now we know how to:

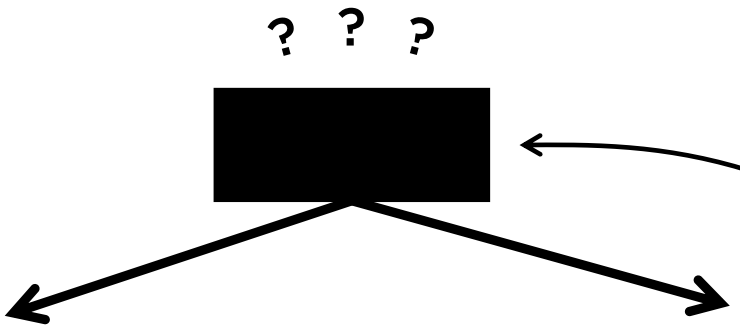
1. Build a Random Forest

2. Use a Random Forest

3. Estimate the Accuracy of a Random Forest

Random Forests

When we built our first tree and we only used two variables to make a decision at each step.



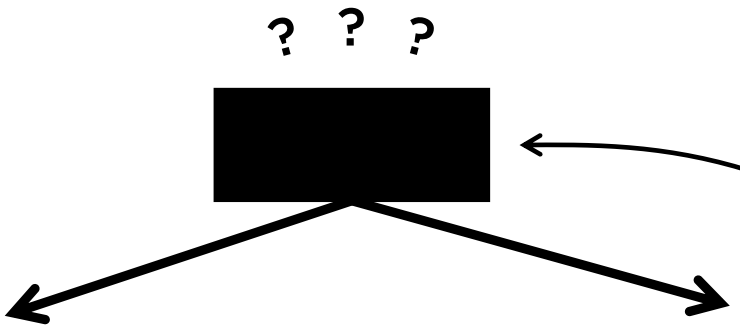
Bootstrapped Dataset

Chest pain	Good blood circulation	Weight	Heart disease
Yes	Yes	156	No
Yes	Yes	178	Yes
Yes	No	180	Yes
Yes	No	180	Yes

Random Forests

Compare Random Forest built using 2 variables per step...

... to Random Forest built using 3 variables per step.



Bootstrapped Dataset

Chest pain	Good blood circulation	Weight	Heart disease
Yes	Yes	156	No
Yes	Yes	178	Yes
Yes	No	180	Yes
Yes	No	180	Yes

Test a bunch of different settings

and choose the most accurate RF.

Random Forests

So...

1. Build a Random Forest

2. Estimate the Accuracy of a Random Forest

Change the number of variables used per step

- Do this for a bunch of times and then choose the one that is most accurate.
- Typically, we start by using the square of the number of variables and then try a few settings above and below that value.

The End