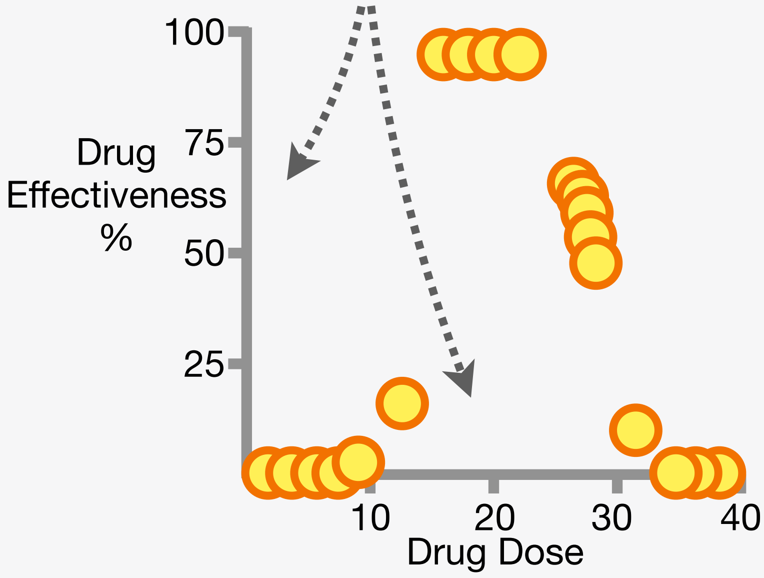
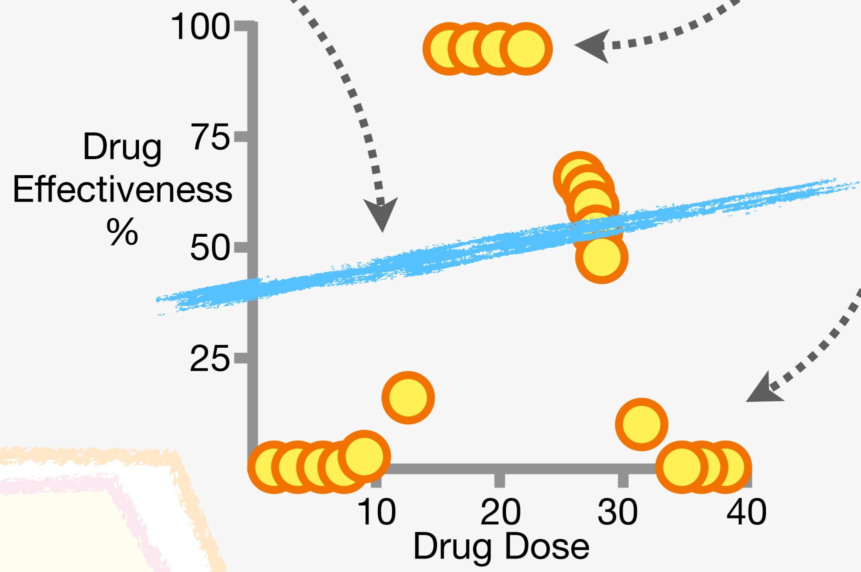
Regression Trees

Regression Trees: Main Ideas Part 1

1 **The Problem:** We have this **Training Dataset** that consists of Drug Effectiveness for different Doses…

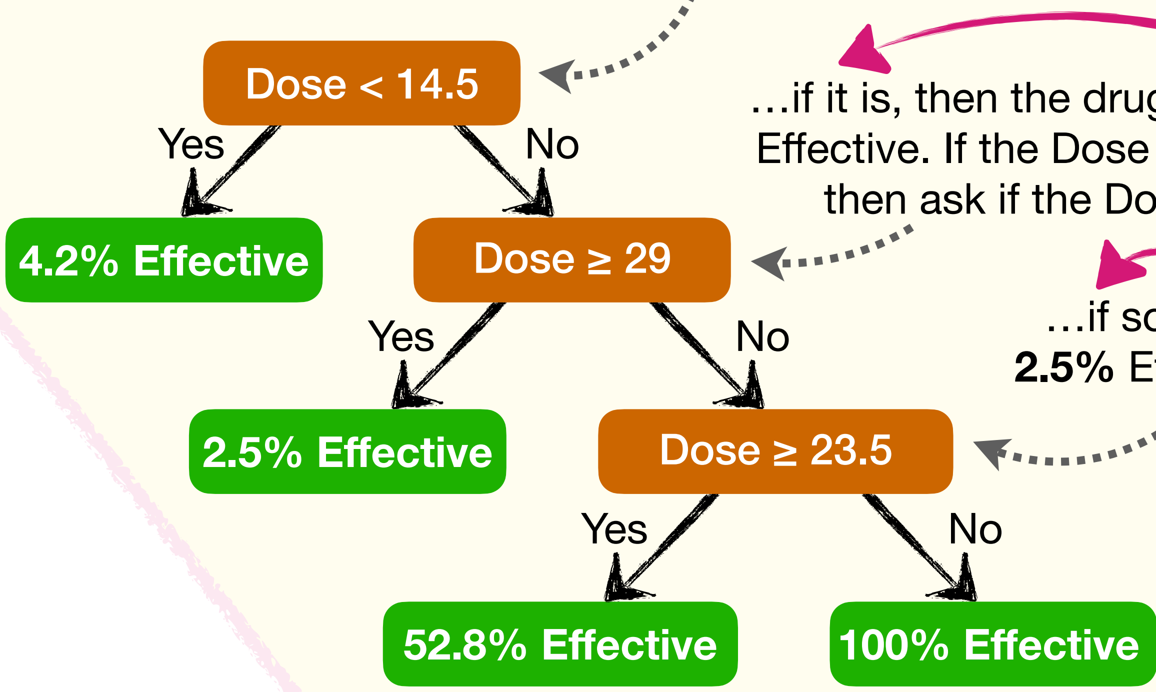


…and fitting a **straight line** to the data would result in terrible predictions because there are clusters of *ineffective* Doses that surround the *effective* Doses.



2 **A Solution:** We can use a **Regression Tree**, which, just like a **Classification Tree**, can handle all types of data and all types of relationships among variables to make decisions, but now the output is a *continuous* value, which, in this case, is Drug Effectiveness.

Just like **Classification Trees**, **Regression Trees** are relatively easy to interpret and use. In this example, if you're given a new Dose and want to know how Effective it will be, you start at the top and ask if the Dose is **< 14.5**…



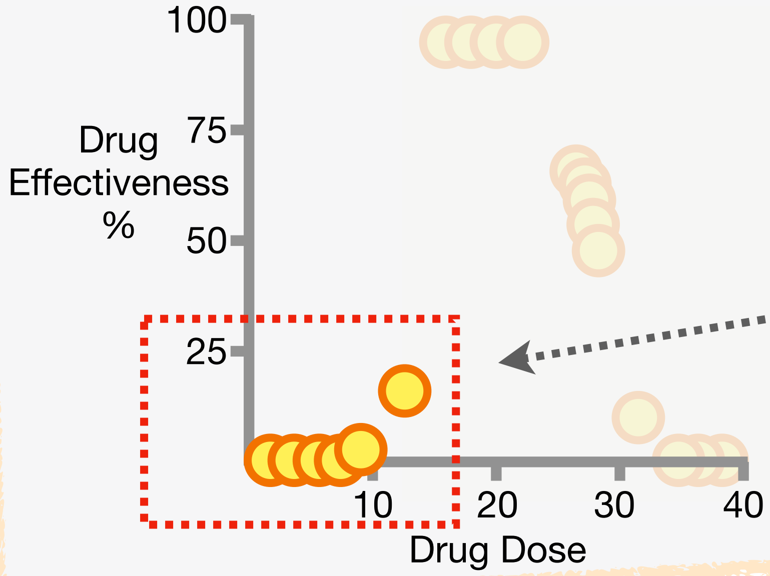
…if it is, then the drug will be **4.2%** Effective. If the Dose is not **< 14.5**, then ask if the Dose ≥ **29**…

…if so, then the drug will be **2.5%** Effective. If not, ask if the Dose ≥ **23.5**…

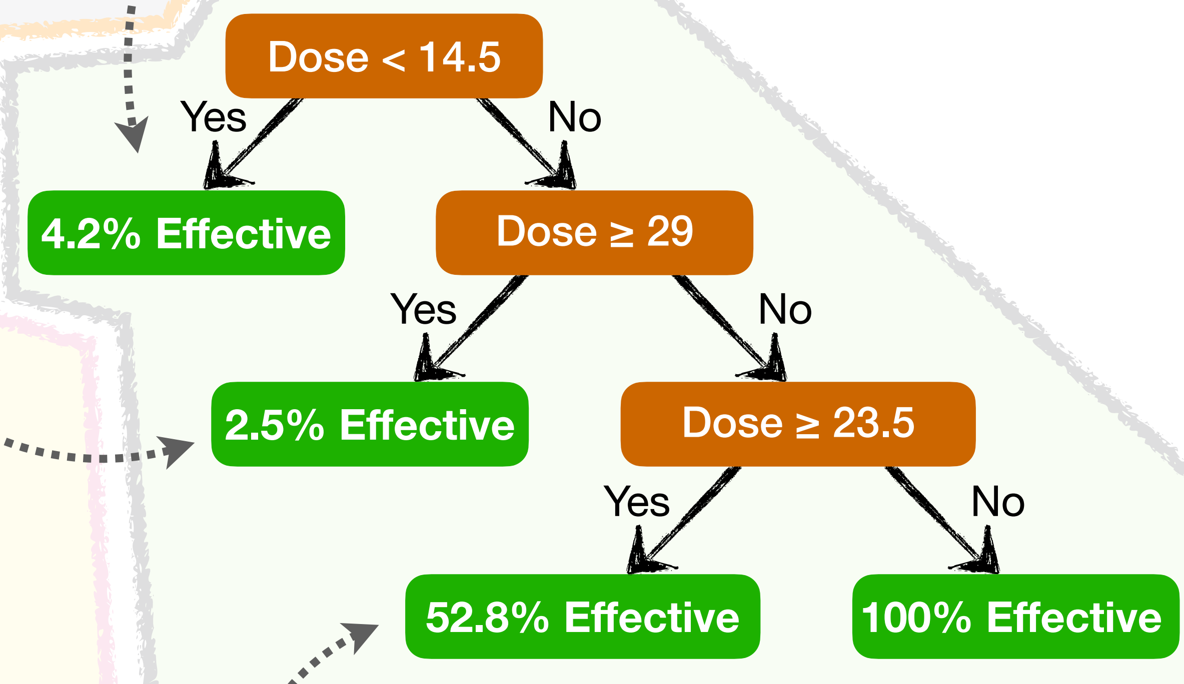
…if so, then the drug will be **52.8%** Effective. If not, the drug will be **100%** Effective…

Regression Trees: Main Ideas Part 2

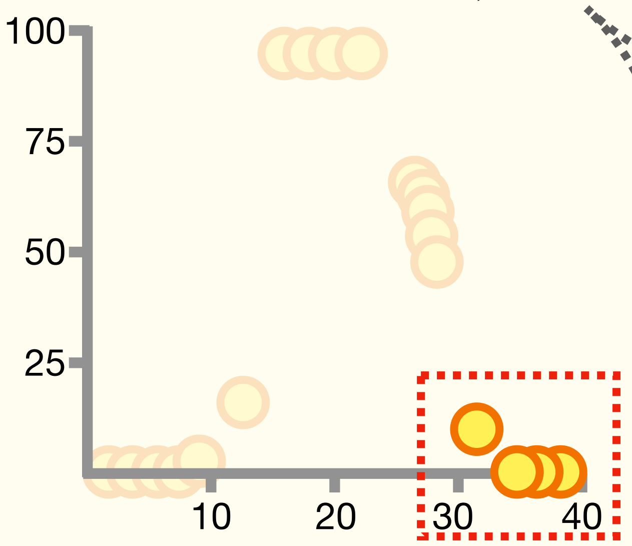
3 In this example, the **Regression Tree** makes good predictions because each **Leaf** corresponds to a different cluster of points in the graph.



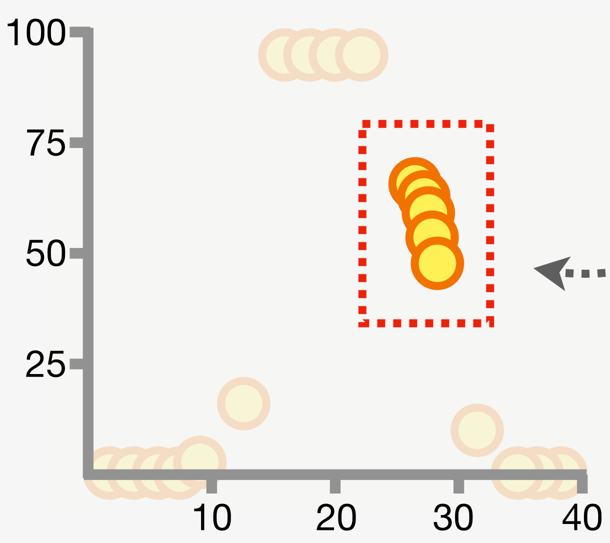
If we have a Dose **< 14.5**, then the output from the **Regression Tree** is the *average* Effectiveness from these **6** measurements, which is **4.2%**.



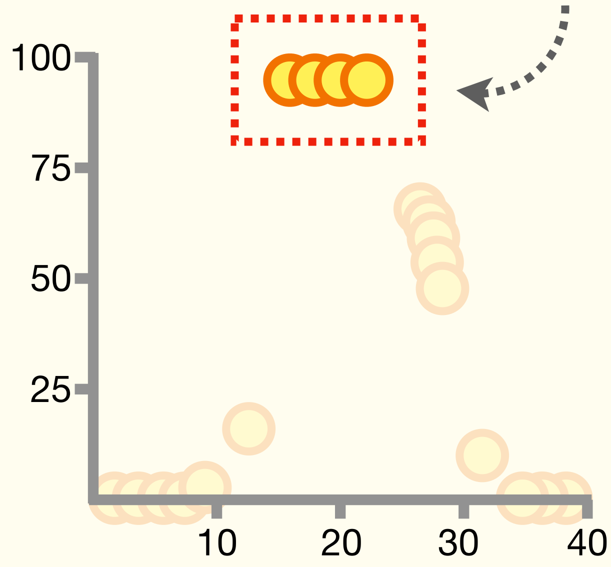
4 If the Dose is **≥ 29**, then the output is the *average* Effectiveness from these **4** measurements, **2.5%**.



5 If the Dose is between **23.5** and **29**, then the output is the *average* Effectiveness from these **5** measurements, **52.8%**.

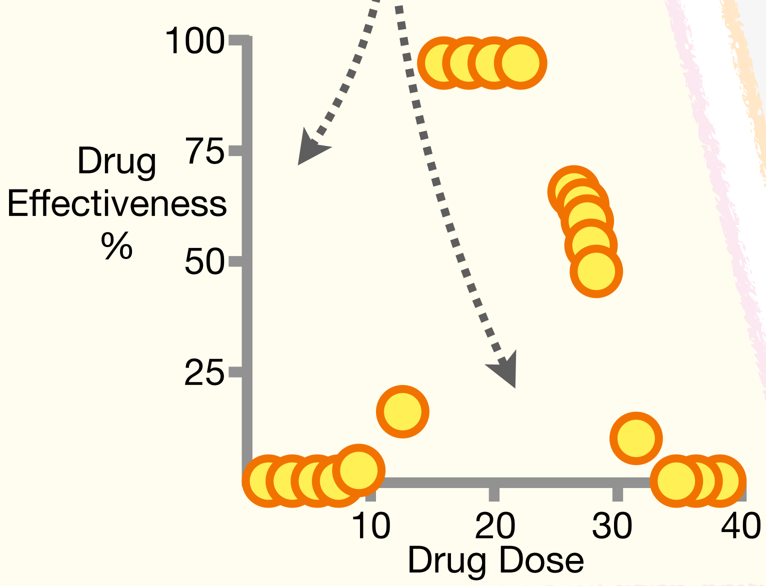


6 If the Dose is between **14.5** and **23.5**, then the output is the *average* Effectiveness from these **4** measurements, **100%**.

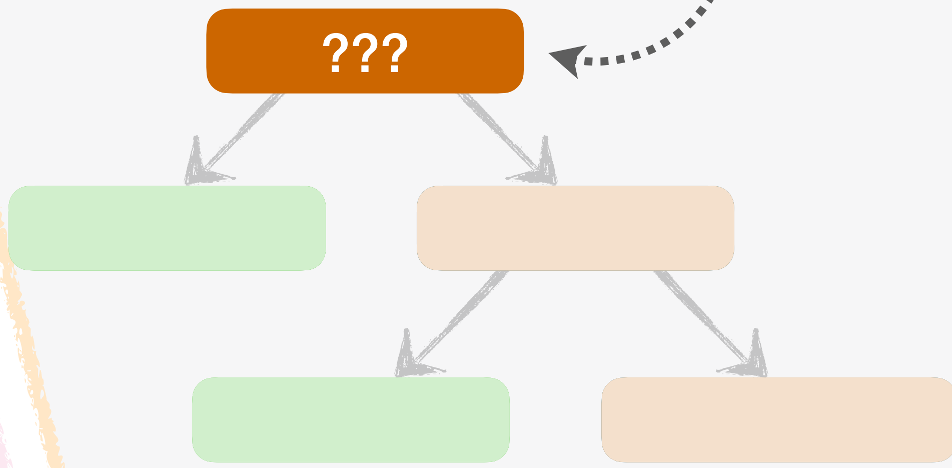


Building a Regression Tree: Step-by-Step

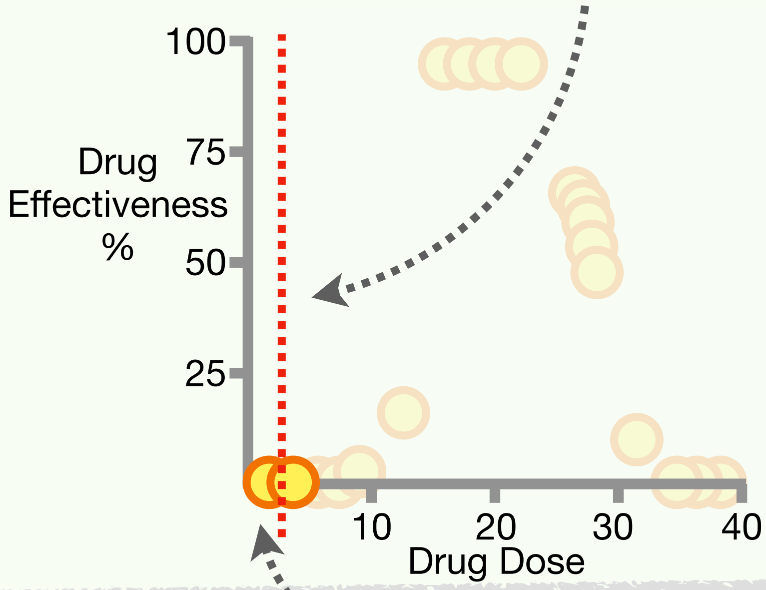
1 Given these **Training Data**, we want to build a **Regression Tree** that uses Drug Dose to predict Drug Effectiveness.



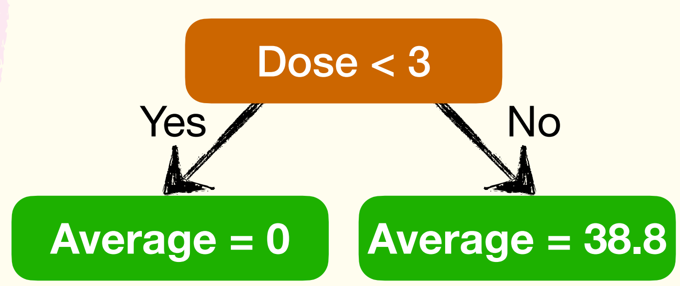
2 Just like for **Classification Trees**, the first thing we do for a **Regression Tree** is decide what goes in the **Root**.



3 To make that decision, we calculate the average of the first **2** Doses, which is **3** and corresponds to this **dotted line**…

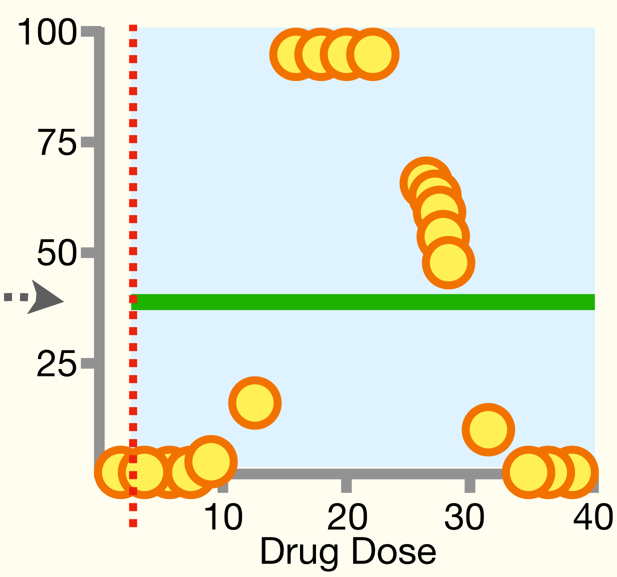


…and then we build a very simple tree that splits the measurements into **2** groups based on whether or not the Dose **<** **3**.



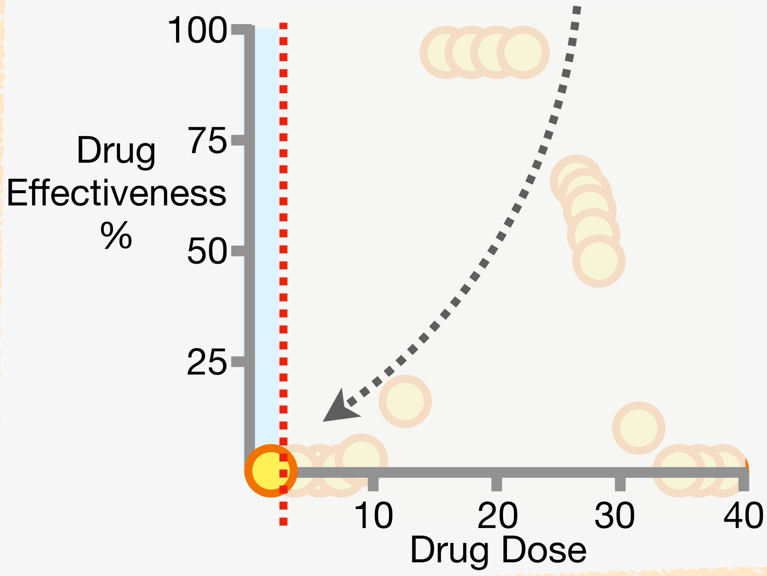
4 Because only one point has a Dose **<** **3**, and its average Effectiveness is **0**, we put **0** in the **Leaf** on the *left*.

All other points have a **Dose ≥ 3**, and their average Effectiveness is **38.8**, so we put **38.3** in the **Leaf** on the *right*.

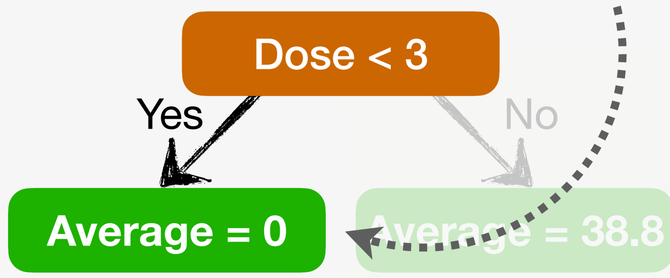


Building a Regression Tree: Step-by-Step

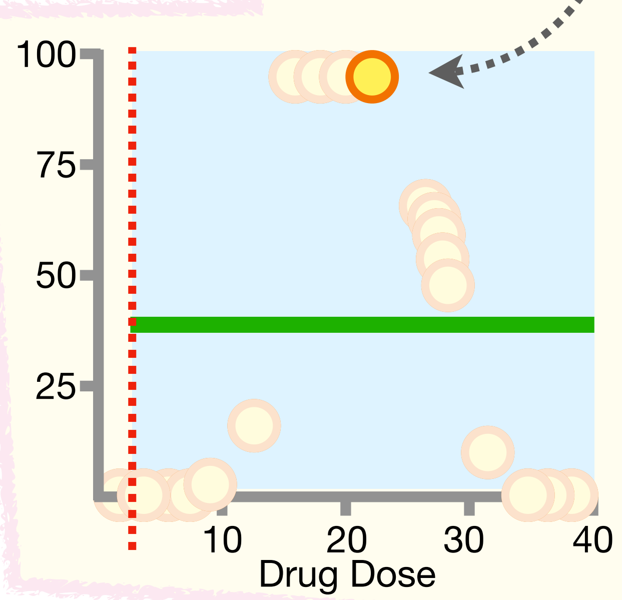
5 For the one point with Dose < **3**, which has Effectiveness = **0**…



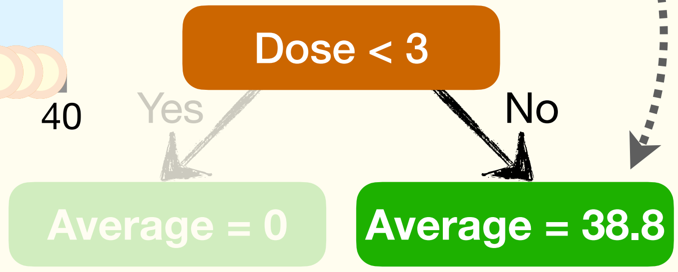
…the **Regression Tree** makes a pretty good prediction, **0**.



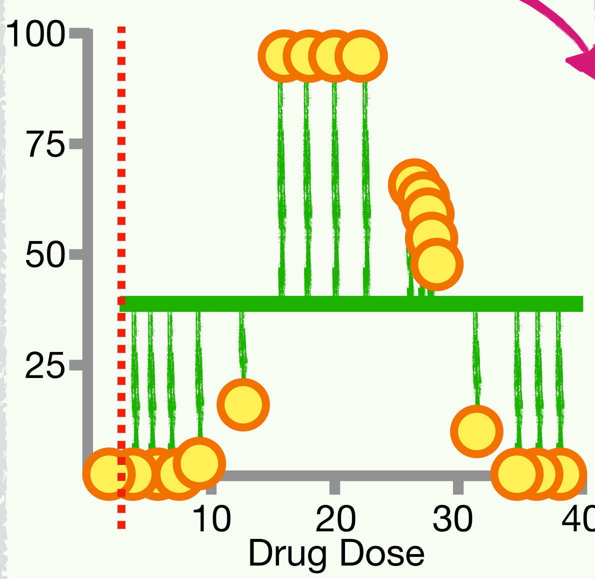
6 In contrast, for this specific point, which has Dose **> 3** and **100%** Effectiveness…



…the tree predicts that the Effectiveness will be **38.8**, which is a pretty bad prediction.



7 We can *visualize* how good or bad the **Regression Tree** is at making predictions by drawing the **Residuals**, the differences between the Observed and Predicted values.



We can also *quantify* how good or bad the predictions are by calculating the **Sum of the Squared Residuals (SSR)**…

$$(0 - 0)^2 + (0 - 38.8)^2 + (0 - 38.8)^2 + (0 - 38.8)^2$$

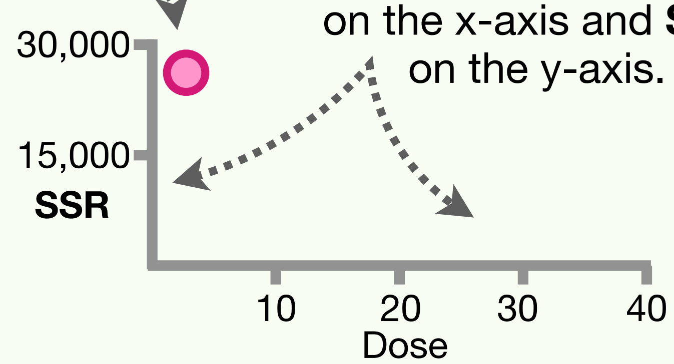
$$+ (5 - 38.8)^2 + (20 - 38.8)^2 + (100 - 38.8)^2$$

$$+ (100 - 38.8)^2 + \ldots + (0 - 38.8)^2$$

$$\boxed{= 27{,}468.5}$$

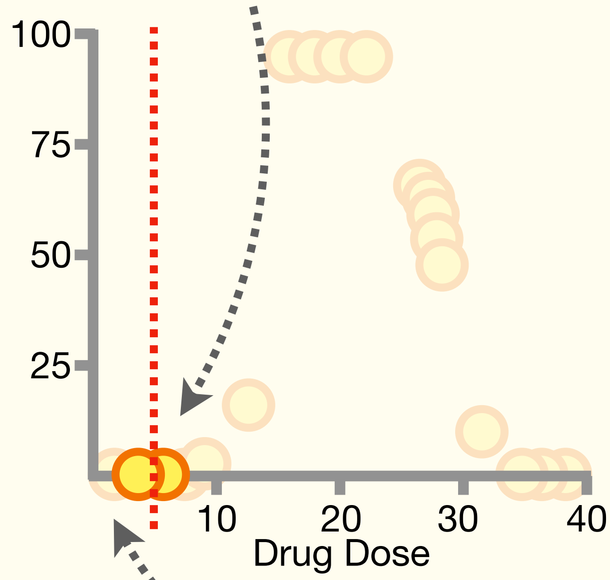
…and when the threshold for the tree is Dose **< 3**, then the **SSR = 27,468.5**.

Lastly, we can compare the **SSR** for different thresholds by plotting them on this graph, which has Dose on the x-axis and **SSR** on the y-axis.

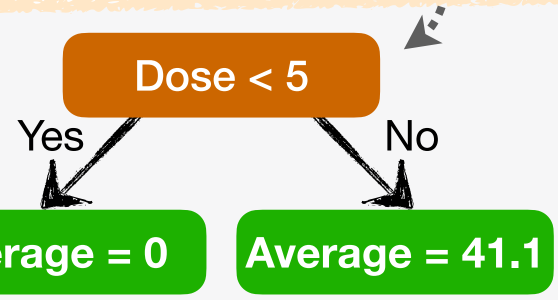


Building a Regression Tree: Step-by-Step

8 Now we shift the Dose threshold to be the average of the second and third measurements in the graph, **5**…

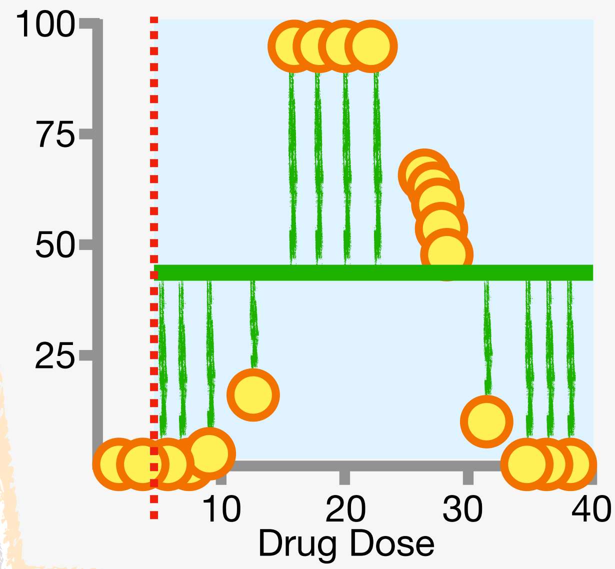


…and we build this super simple tree with Dose **<** **5** at the **Root**.

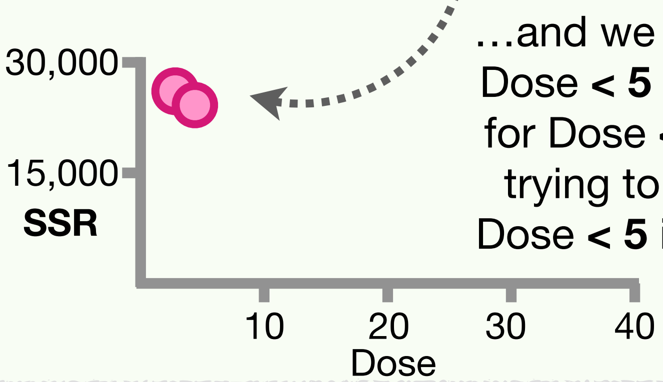


9 Because the average Effectiveness for the **2** points with Dose **<** **5** is **0**, we put **0** in the **Leaf** on the *left*.

All of the other points have Dose **≥** **5**, and their average is **41.1**, so we put **41.1** in the **Leaf** on the *right*.

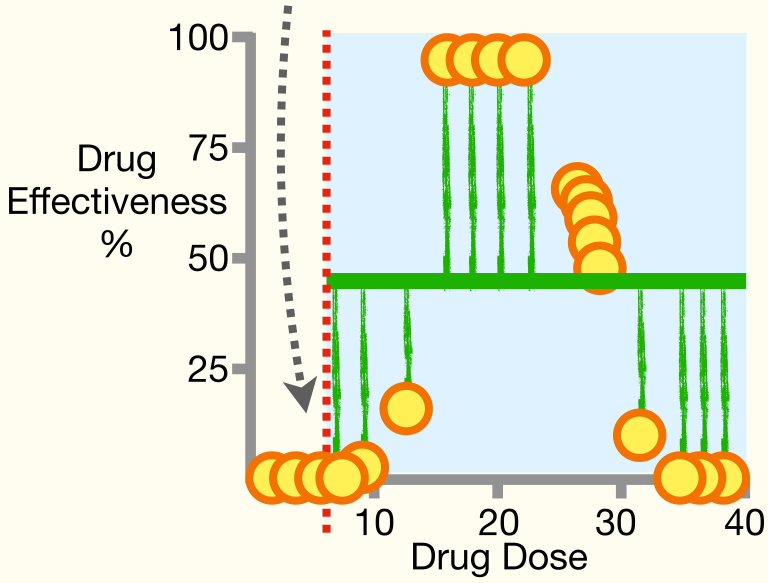


10 Now we calculate and plot the **SSR** for the new threshold, Dose **<** **5**…

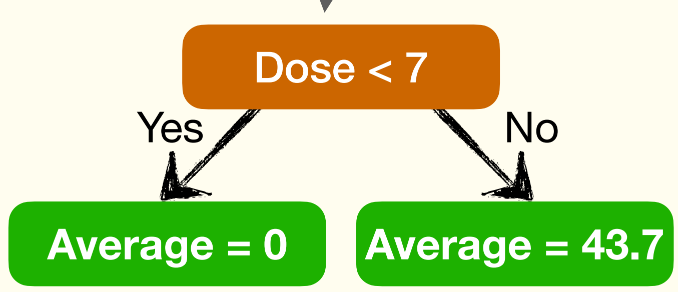


…and we see that the **SSR** for Dose **<** **5** is less than the **SSR** for Dose **<** **3**, and since we're trying to minimize the **SSR**, Dose **<** **5** is a better threshold.

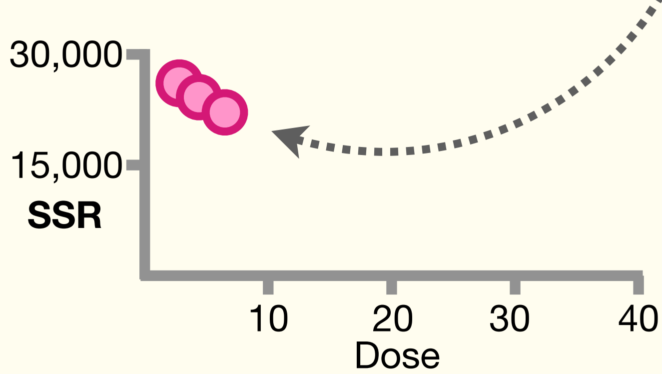
11 Then we shift the threshold to be the average of the third and fourth measurements, **7**…



…and that gives us this tree…

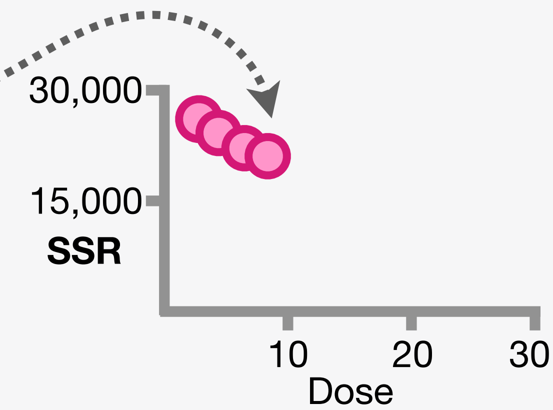


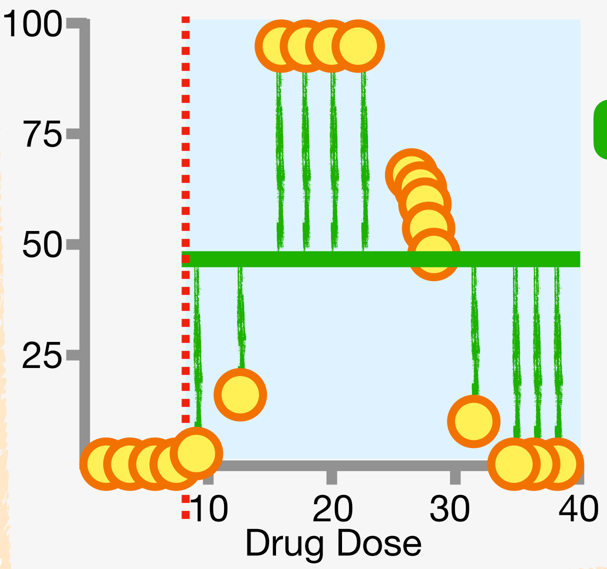
…and this point on the graph.

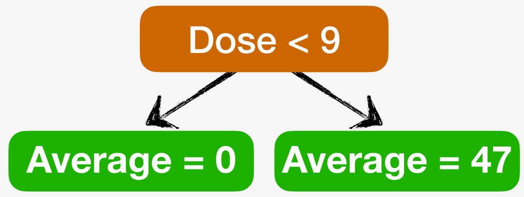


Building a Regression Tree: Step-by-Step

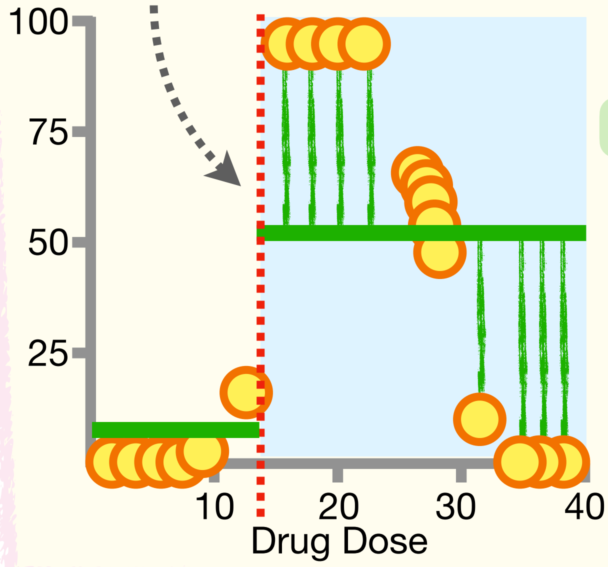
12 Then we just keep shifting the threshold to the average of every pair of consecutive Doses, create the tree, then calculate and plot the **SSR**.

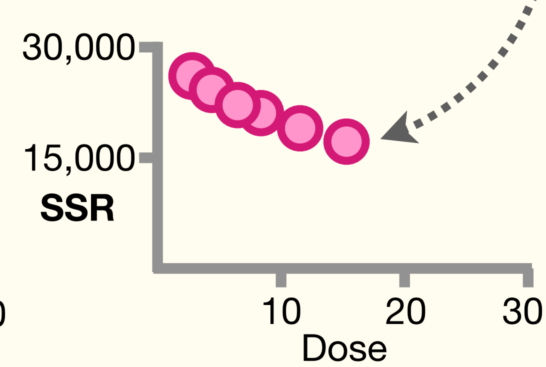




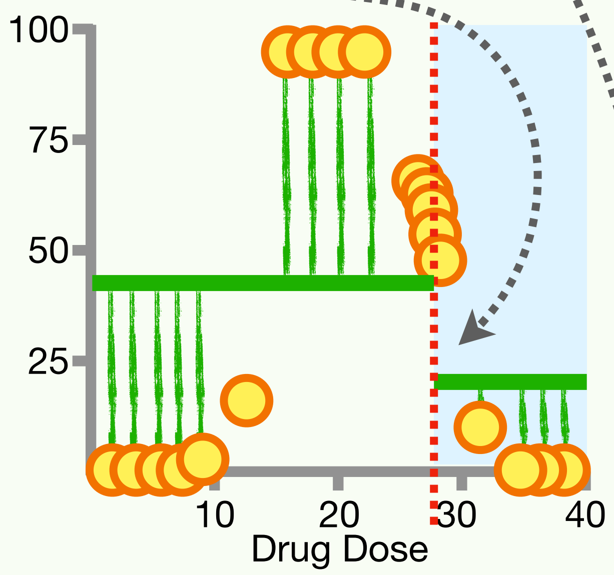


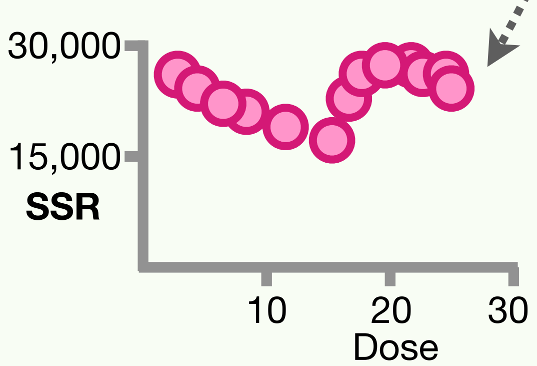
13 After shifting the Dose threshold over **2** more steps, the **SSR** graph looks like this.



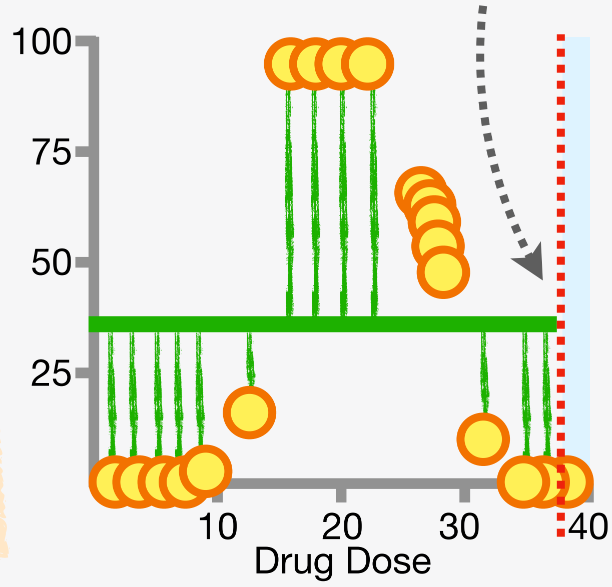


14 Then, after shifting the Dose threshold **7** more times, the **SSR** graph looks like this.

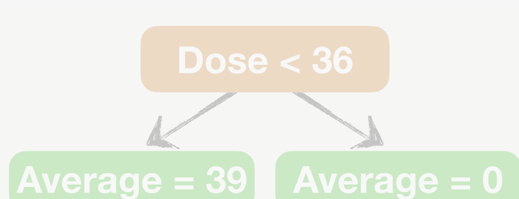


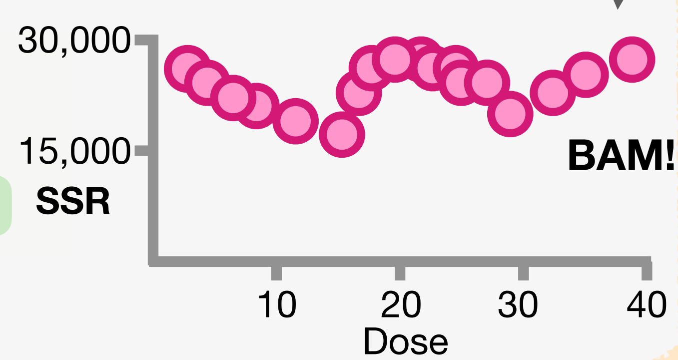


15 And finally, after shifting the Dose threshold all the way to the last pair of Doses…



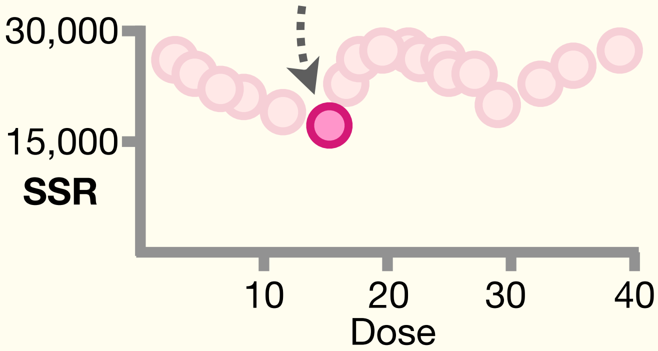
…the **SSR** graph looks like this.



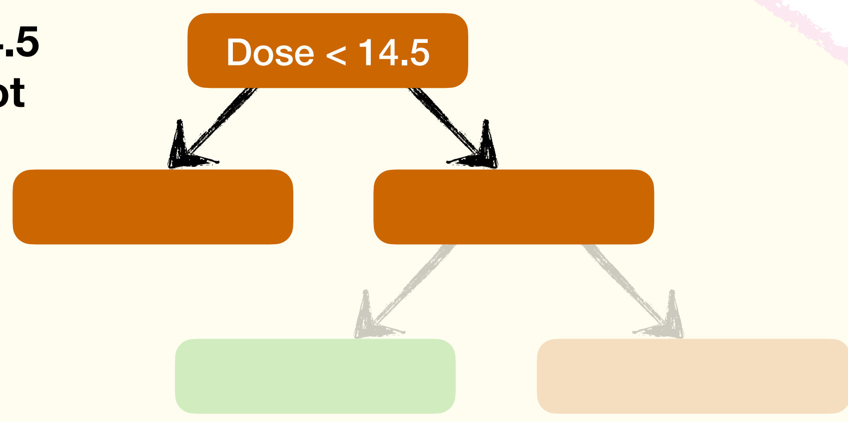


Building a Regression Tree: Step-by-Step

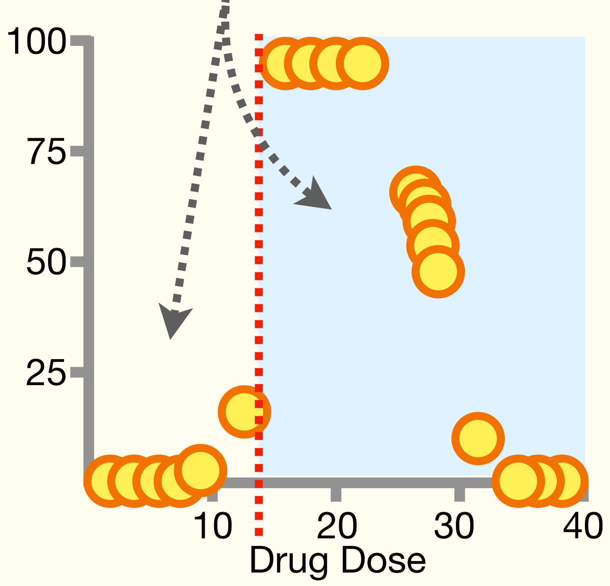
16 Looking at the **SSRs** for each Dose threshold, we see that Dose **< 14.5** had the smallest **SSR**…



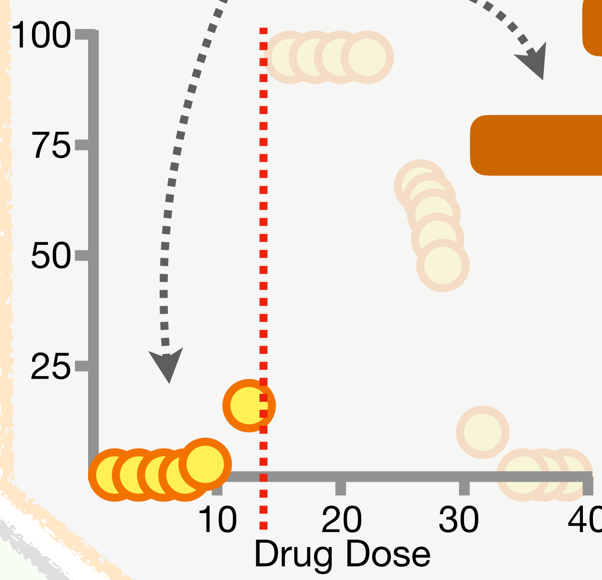
…so Dose **< 14.5** will be the **Root** of the tree…

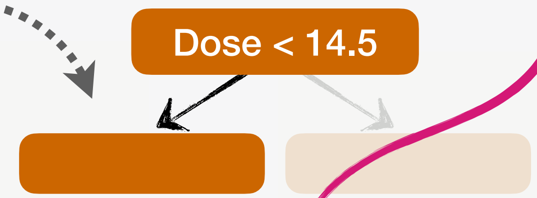


…which corresponds to splitting the measurements into two groups based on whether or not the Dose **< 14.5**.



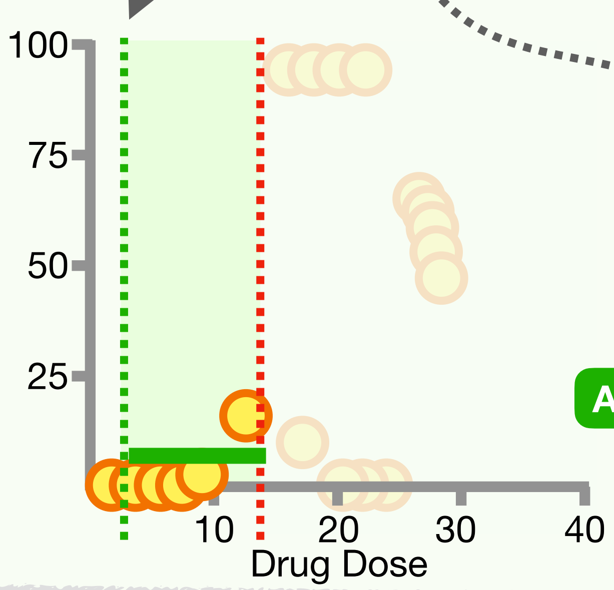
17 Now, because the threshold in the **Root** of the tree is Dose **< 14.5**, these **6** measurements go into the **Node** on the *left*…

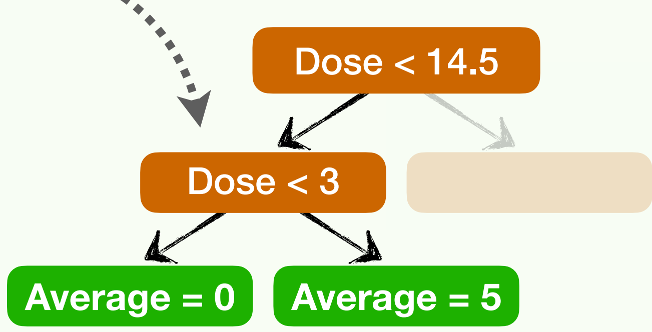




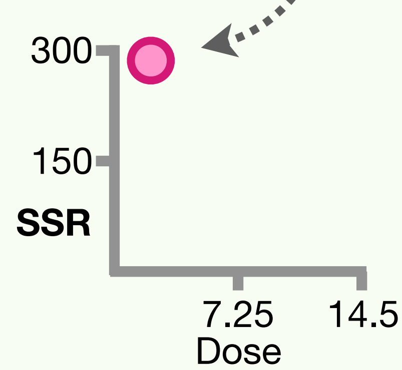
…and, *in theory*, we could subdivide these **6** measurements into smaller groups by repeating what we just did, only this time focusing only on these **6** measurements.

18 In other words, just like before, we can average the first two Doses and use that value, **3**, as a cutoff for splitting the **6** measurements with Dose **< 14.5** into **2** groups…





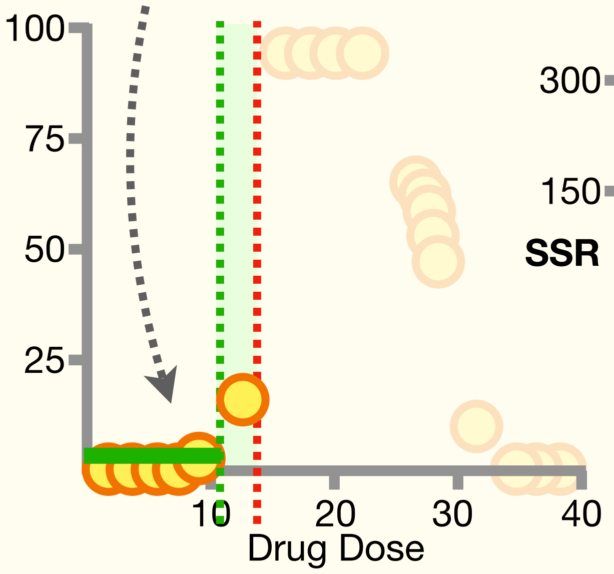
…then we calculate the **SSR** for just those **6** measurements and plot it on a graph.

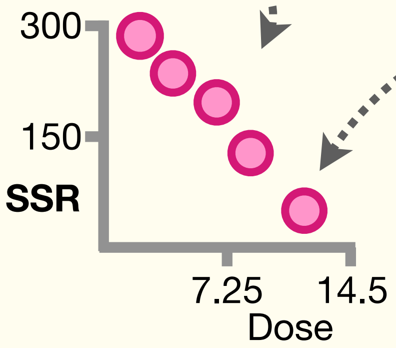


Building a Regression Tree: Step-by-Step

19 And after calculating the **SSR** for each threshold for the **6** measurements with Dose **< 14.5**, we end up with this graph…

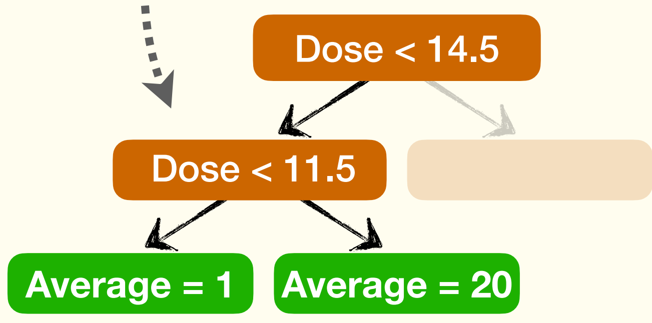
…and then we select the threshold that gives us the lowest **SSR**, Dose **< 11.5**, for the next **Node** in the tree.



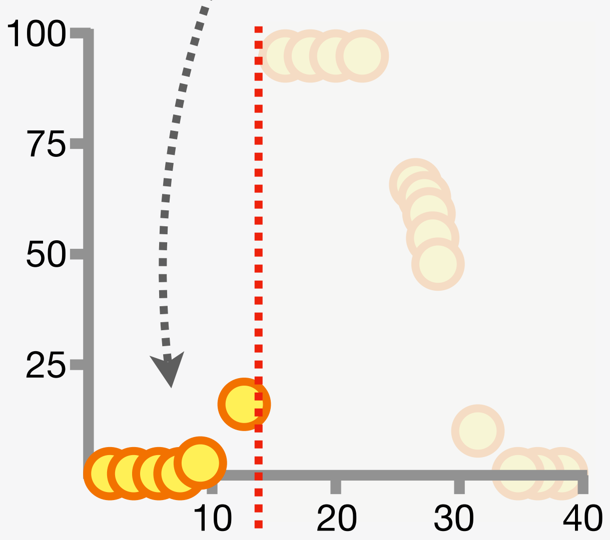


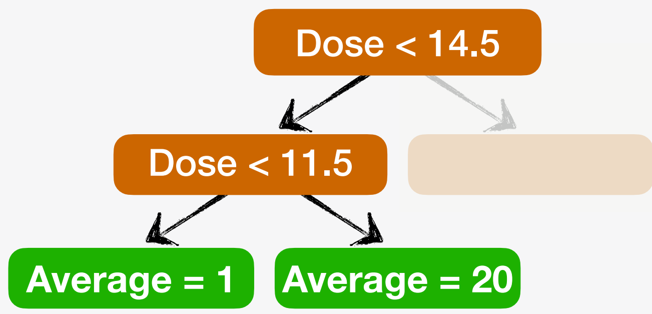
BAM?

No. No bam.

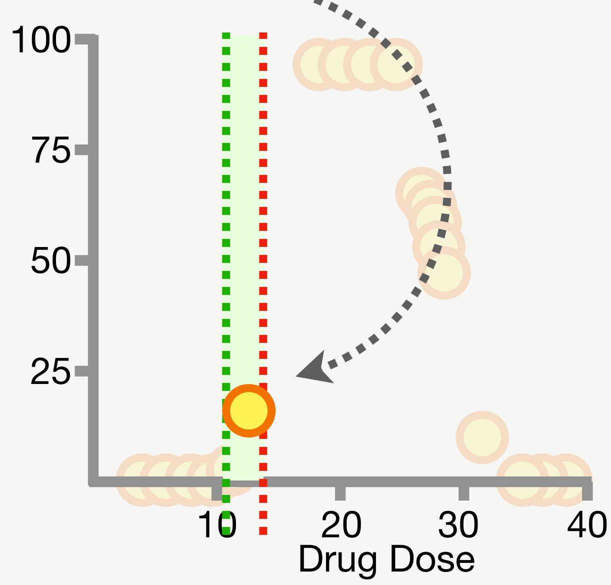


20 Earlier, we said *in theory* we could subdivide the **6** measurements with Dose **< 14.5** into smaller groups…





…but when we do, we end up with a single measurement in the **Leaf** on the *right* because there is the only one measurement with a Dose between **11.5** and **14.5**…

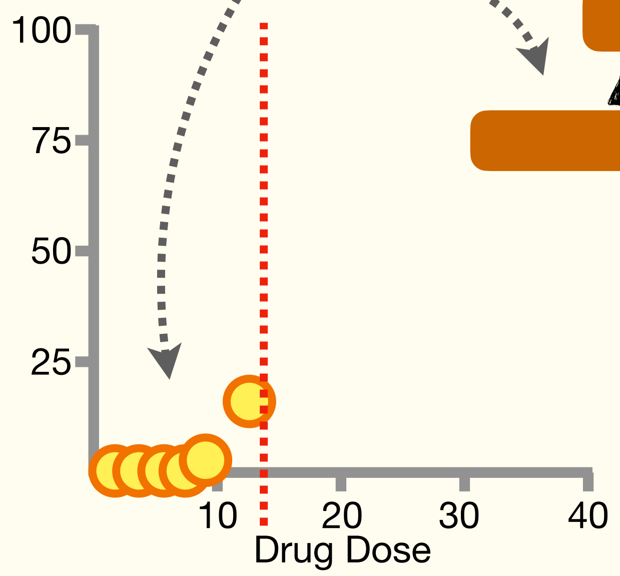


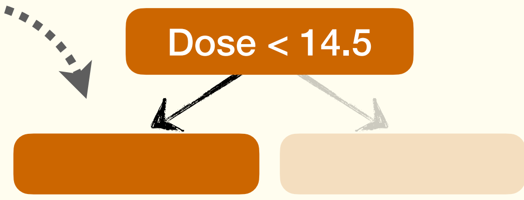
…and making a prediction based on a single measurement suggests that the tree is **Overfit** to the **Training Data** and may not perform well in the future.

The simplest way to prevent this issue is to only split measurements when there are more than some minimum number, which is often **20**. However, since we have so little data in this specific example, we'll set the minimum to **7**.

Building a Regression Tree: Step-by-Step

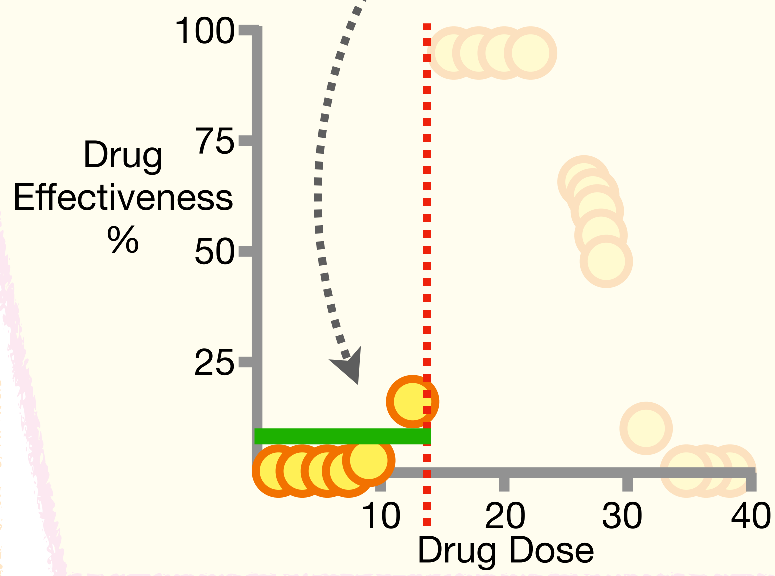
21 Now, because there are only **6** measurements with Dose **< 14.5**, there are only **6** measurements in the **Node** on the *left*…



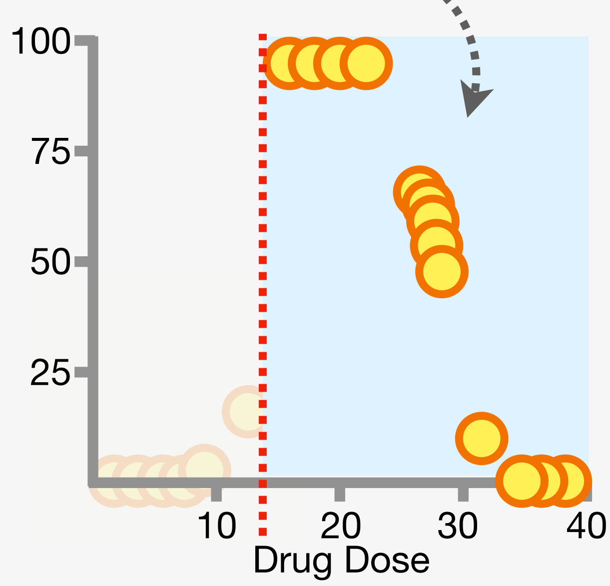


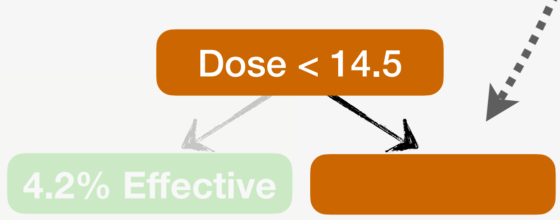
…and because we require a minimum of **7** measurements for further subdivision, the **Node** on the *left* will be a **Leaf**…

…and the output value for the **Leaf** is the average Effectiveness from the **6** measurements, **4.2%.**

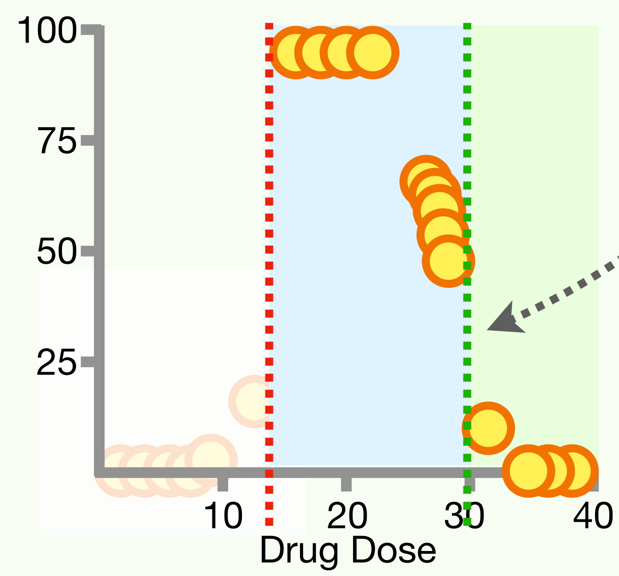


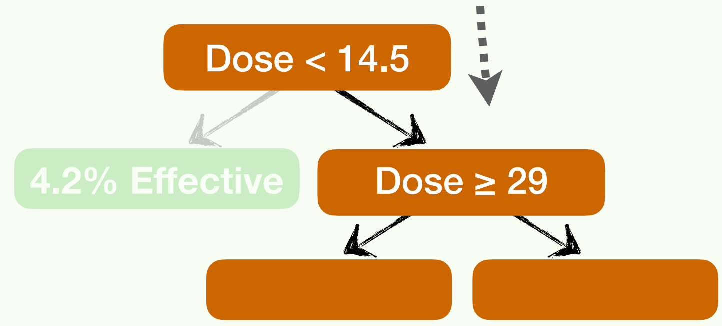
22 Now we need to figure out what to do with the **13** remaining measurements with Doses **≥ 14.5** that go to the **Node** on the *right*.





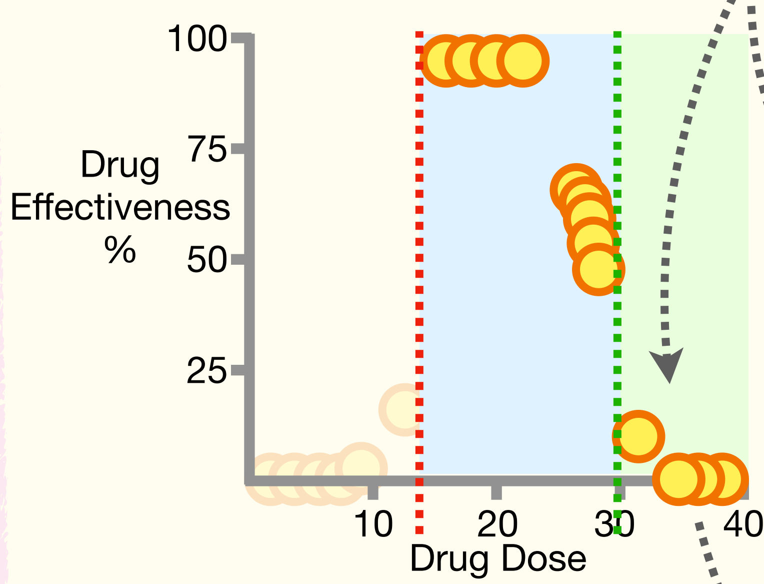
23 Since we have more than **7** measurements in the **Node** on the *right*, we can split them into two groups, and we do this by finding the threshold that gives us the lowest **SSR**.

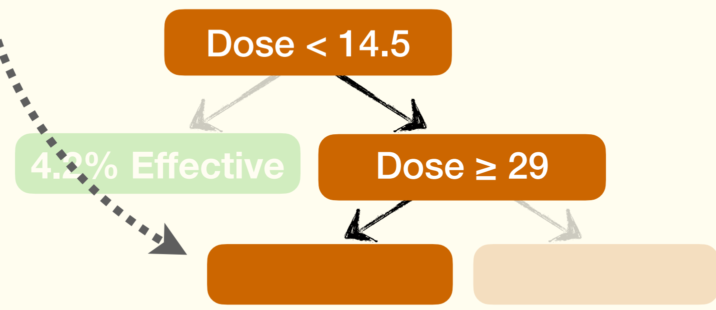




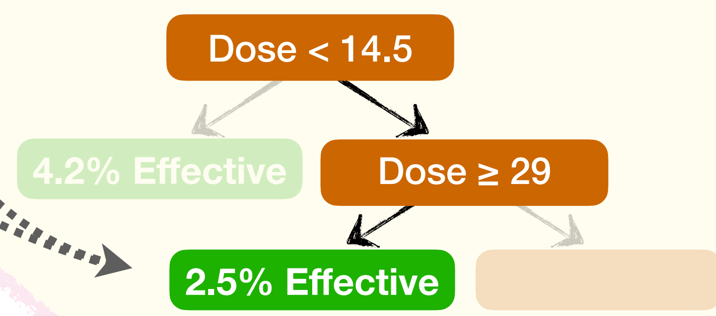
Building a Regression Tree: Step-by-Step

24 And because there are only **4** measurements with Dose ≥ **29**, there are only **4** measurements in the **Node** on the left…

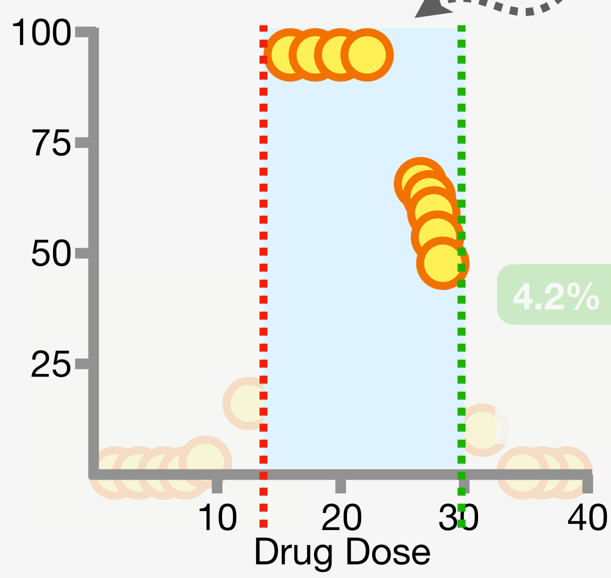


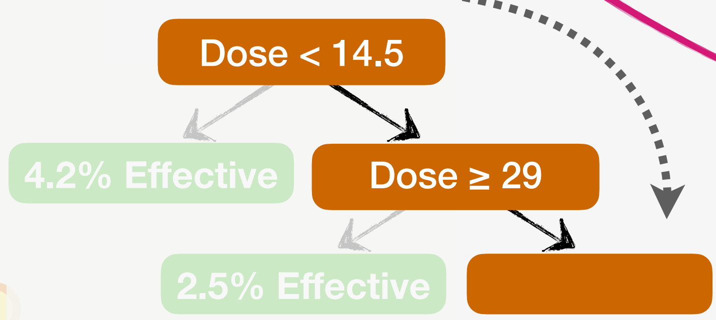


…and since the **Node** has fewer than **7** measurements, we'll make it a **Leaf**, and the output will be the average Effectiveness from those **4** measurements, **2.5%**.

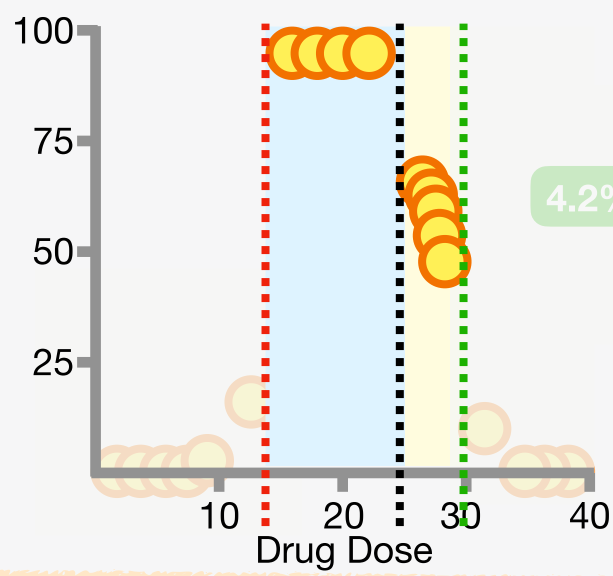


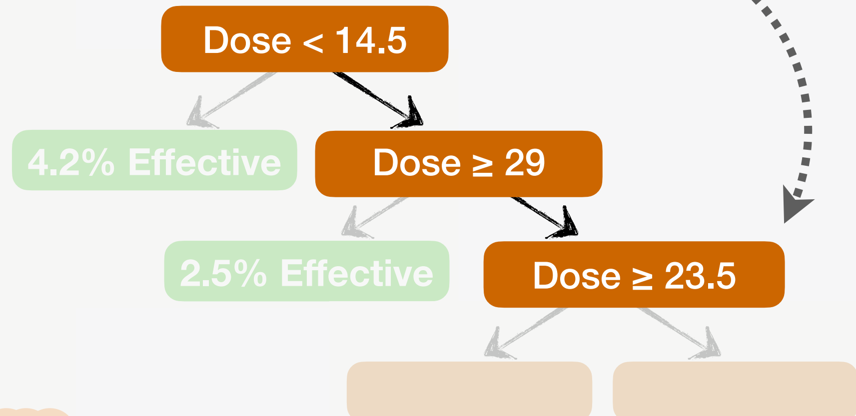
25 Now, because we have more than **7** measurements with Doses between **14.5** and **29**, and thus, more than **7** measurements in the **Node** on the right…





…we can split the measurements into two groups by finding the Dose threshold that results in the lowest **SSR**.

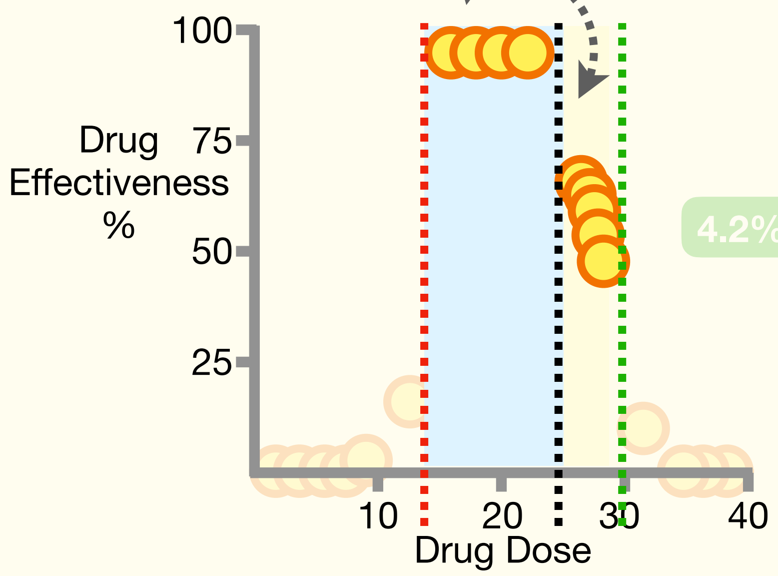




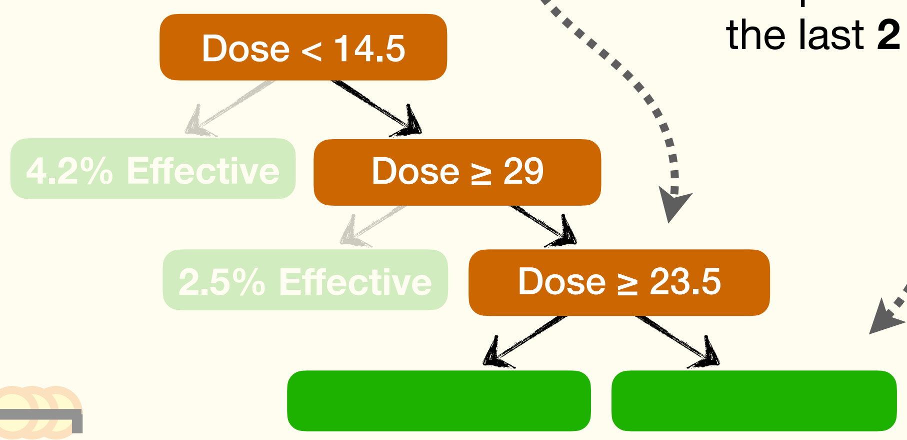
Building a Regression Tree: Step-by-Step

26 And since there are fewer than **7** measurements in each of these two groups…

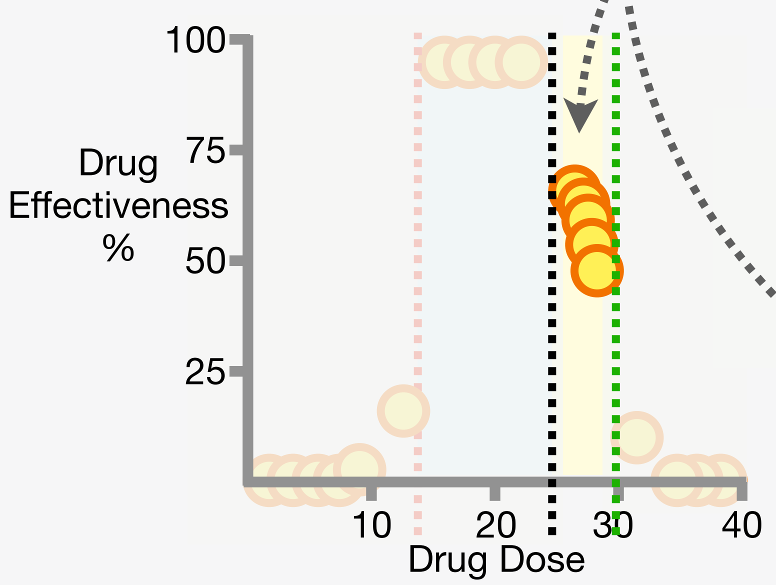
…this will be the last split, because none of the **Leaves** has more than **7** measurements in them.

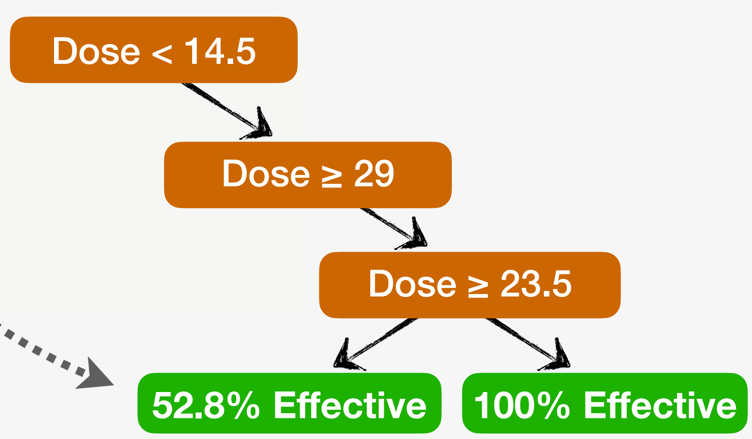


Now, all we have to do is calculate the output values for the last **2 Leaves**.

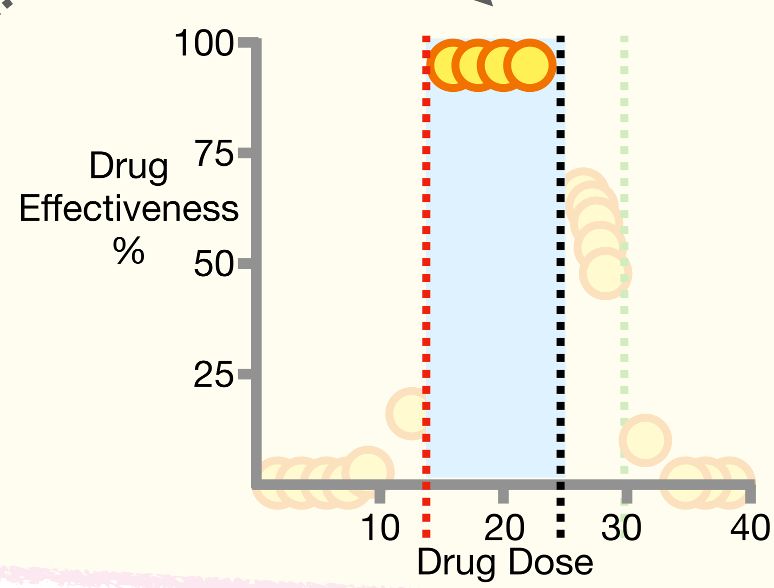


27 So, we use the average Drug Effectiveness for measurements with Doses between **23.5** and **29**, **52.8%**, as the output for **Leaf** on the *left*…





28 …and we use the average Drug Effectiveness for observations with Doses between **14.5** and **23.5**, **100%**, as the output for **Leaf** on the *right*.

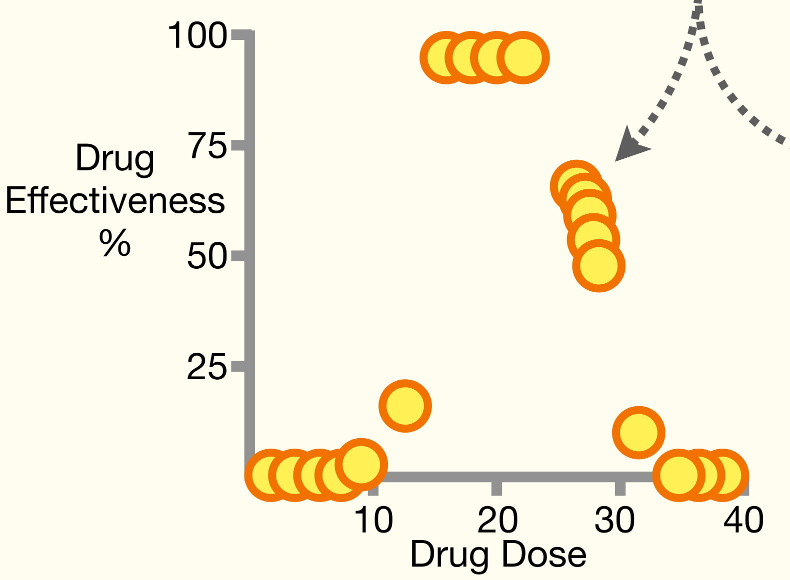


29 Now, at long last, we've finished building the **Regression Tree**.


BUT WAIT!!!
THERE'S MORE!!!

Building a Regression Tree With Multiple Features: Part 1

1 So far, we've built a **Regression Tree** using a single predictor, Dose, to predict Drug Effectiveness.



Dose	Drug Effect
10	98
20	0
35	6
5	44
etc…	etc…

NOTE: Just like for **Classification Trees**, **Regression Trees** can use any type of variable to make a prediction. However, with **Regression Trees**, we always try to predict a continuous value.

2 Now let's talk about how to build a **Regression Tree** to predict Drug Effectiveness using Dose, Age, and Sex.

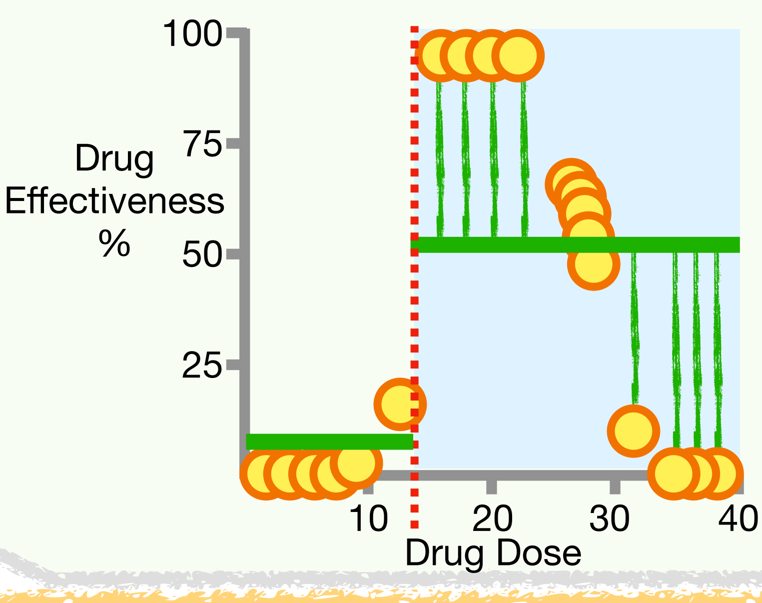
Dose	Age	Sex	Drug Effect
10	25	F	98
20	73	M	0
35	54	F	6
5	12	M	44
etc…	etc…	etc…	etc…

3 First, we completely ignore Age and Sex and only use Dose to predict Drug Effectiveness…

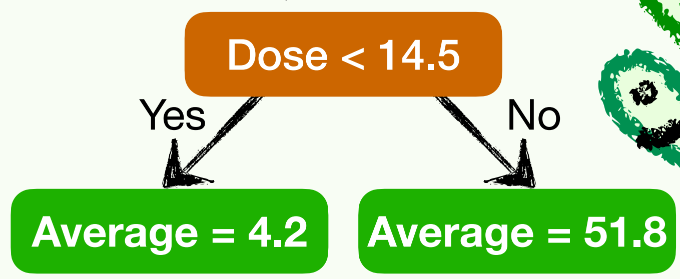
Dose	Age	Sex	Drug Effect
10	25	F	98
20	73	M	0
35	54	F	6
5	12	M	44
etc…	etc…	etc…	etc…

…and then we select the threshold that gives us the smallest **SSR**.

However, instead of that threshold instantly becoming the **Root**, it only becomes a *candidate* for the **Root**.



This might be the **Root**, but we don't know yet.

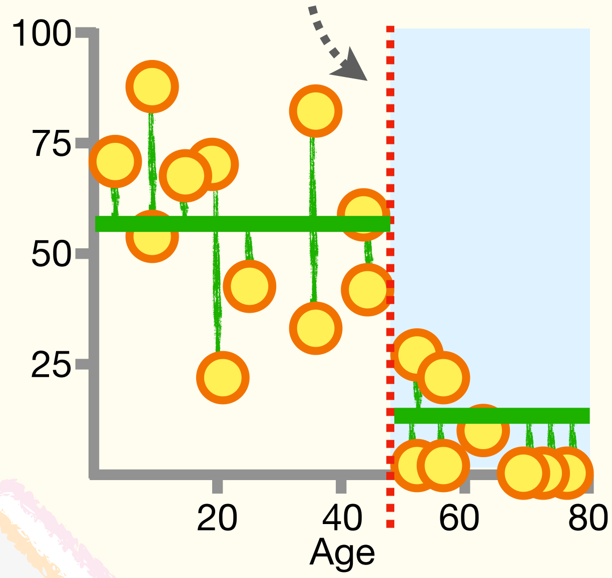


Building a Regression Tree With Multiple Features: Part 2

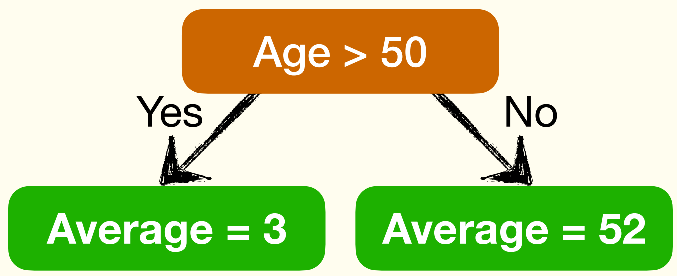
4 Then, we ignore Dose and Sex and only use Age to predict Effectiveness…

Dose	Age	Sex	Drug Effect
10	25	F	98
20	73	M	0
35	54	F	6
5	12	M	44
etc…	etc…	etc…	etc…

…and we select the threshold that gives us the smallest **SSR**…



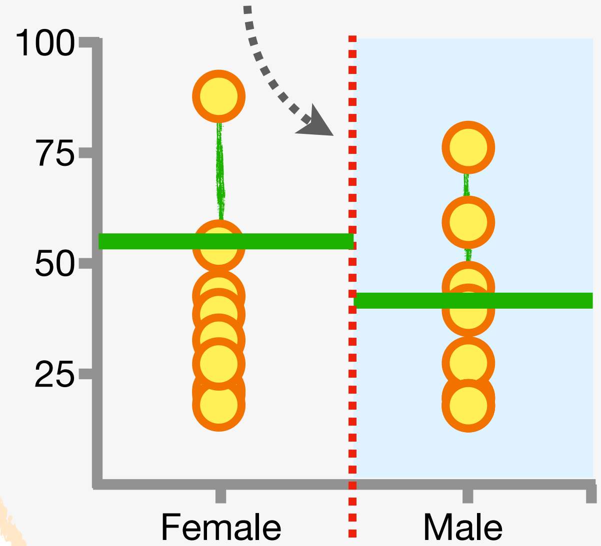
…and that becomes the second *candidate* for the **Root**.



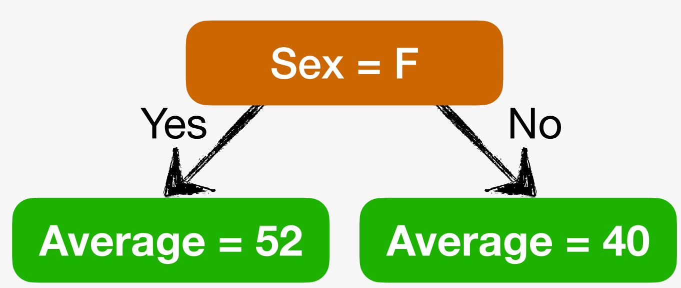
5 Lastly, we ignore Dose and Age and only use Sex to predict Effectiveness…

Dose	Age	Sex	Drug Effect
10	25	F	98
20	73	M	0
35	54	F	6
5	12	M	44
etc…	etc…	etc…	etc…

…and even though Sex only has one threshold for splitting the data, we still calculate the **SSR**, just like before…

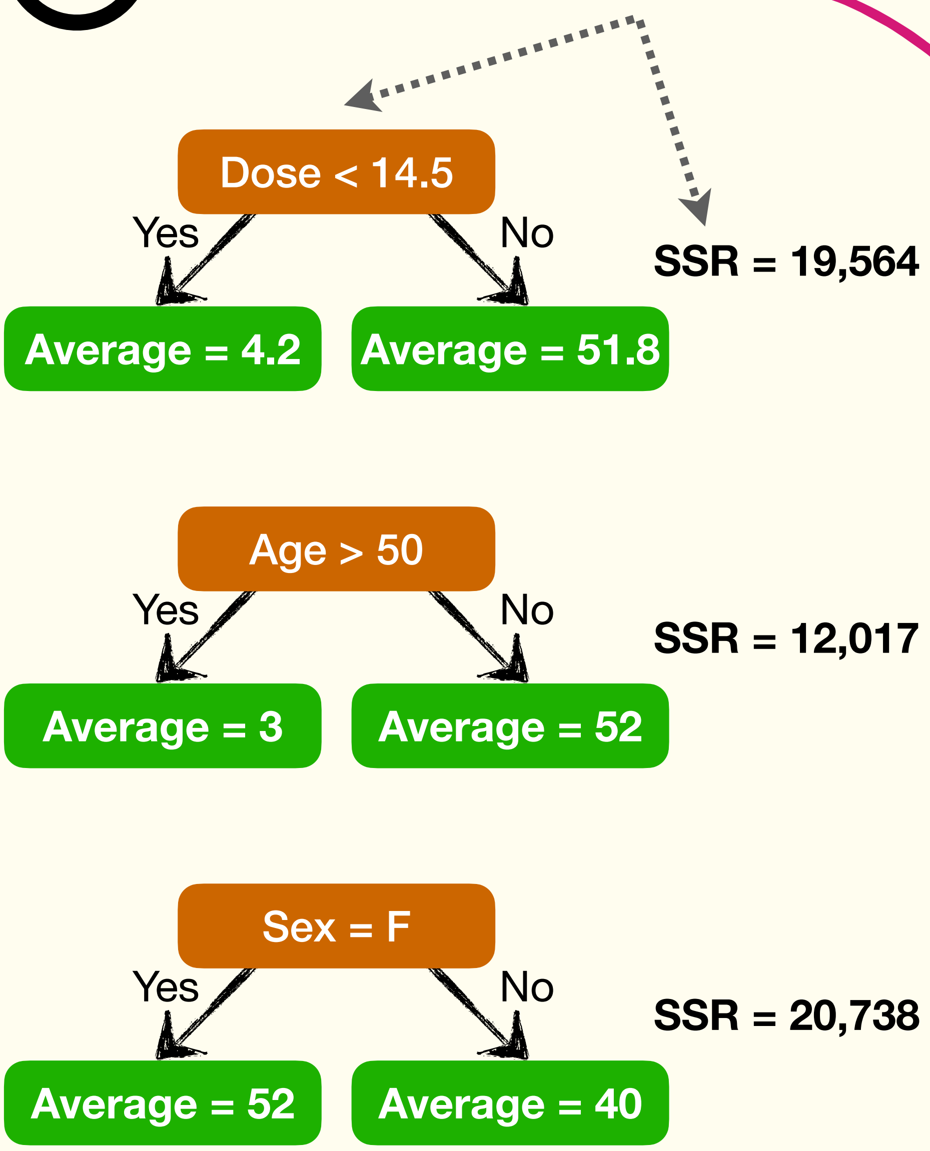


…and that becomes the third *candidate* for the **Root**.



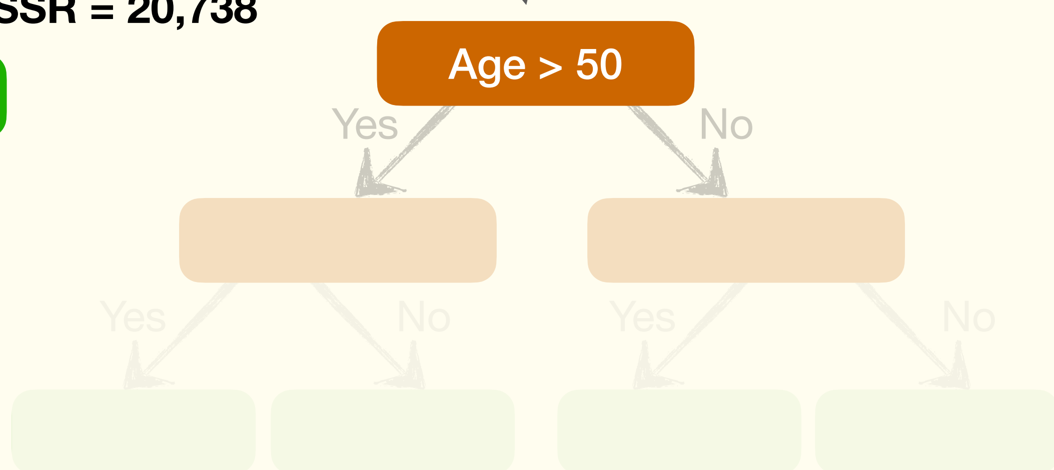
Building a Regression Tree With Multiple Features: Part 3

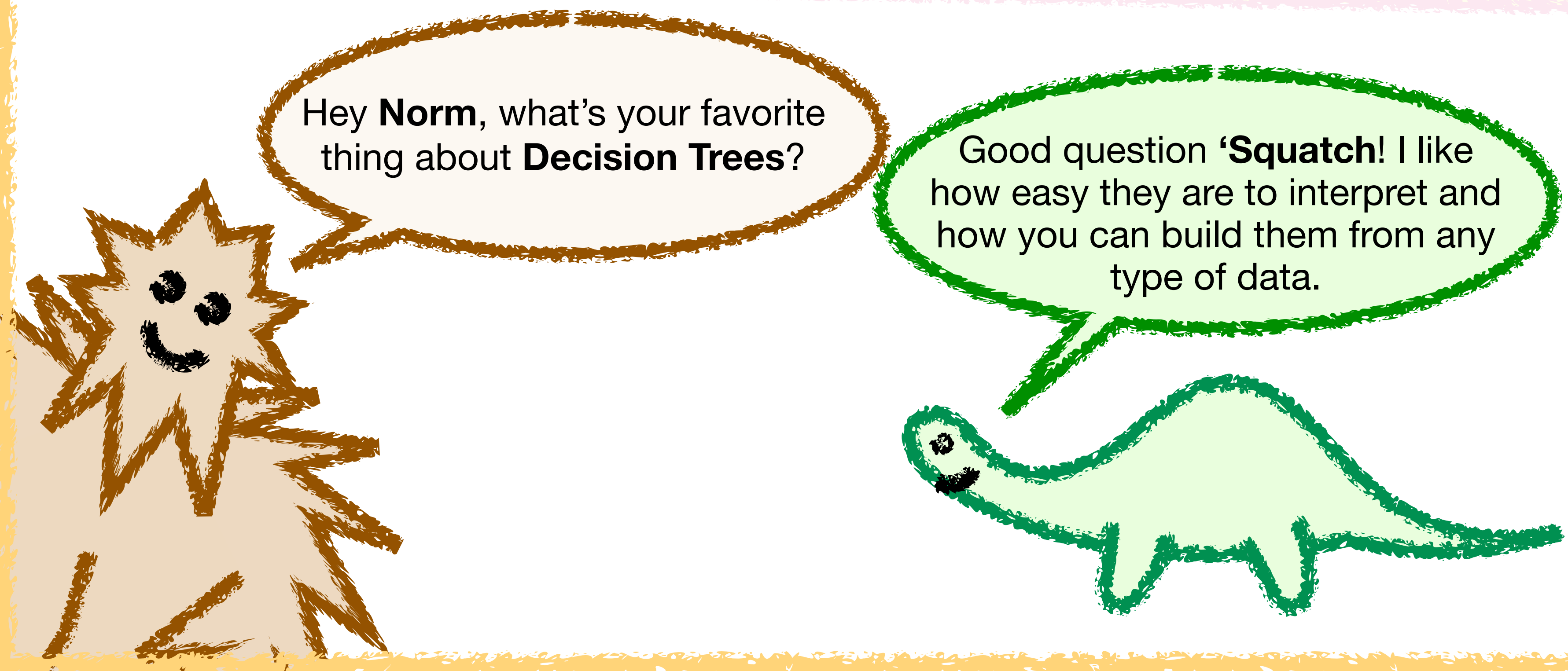
6 Now we compare the **SSRs** for each *candidate* for the **Root**…



…and pick the one with the lowest value…

…and because Age **> 50** had the lowest **SSR**, it becomes the **Root** of the **Regression Tree**.





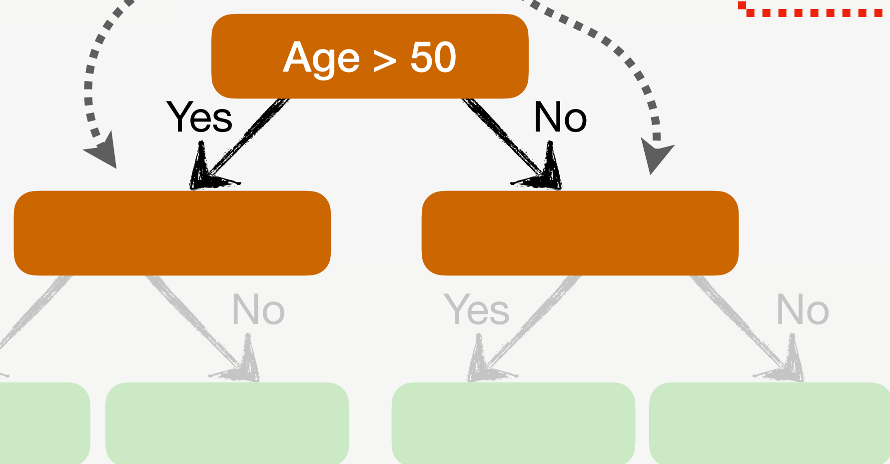
Building a Regression Tree With Multiple Features: Part 4

7 Now that Age **>** **50** is the **Root**, the people in the **Training Data** who are older than **50** go to the **Node** on the *left*…

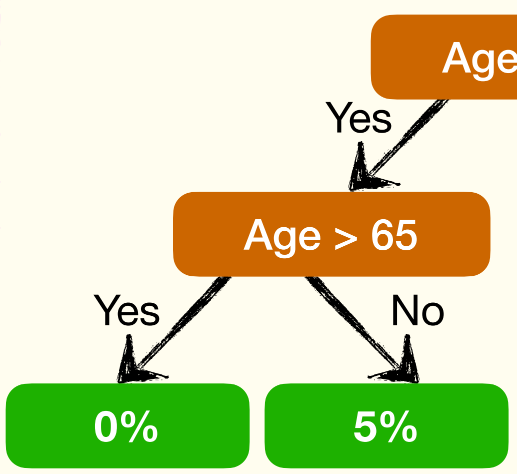
…and the people who are **≤ 50** go to the **Node** on the *right*.

Dose	Age	Sex	Drug Effect
20	73	M	0
35	54	F	6

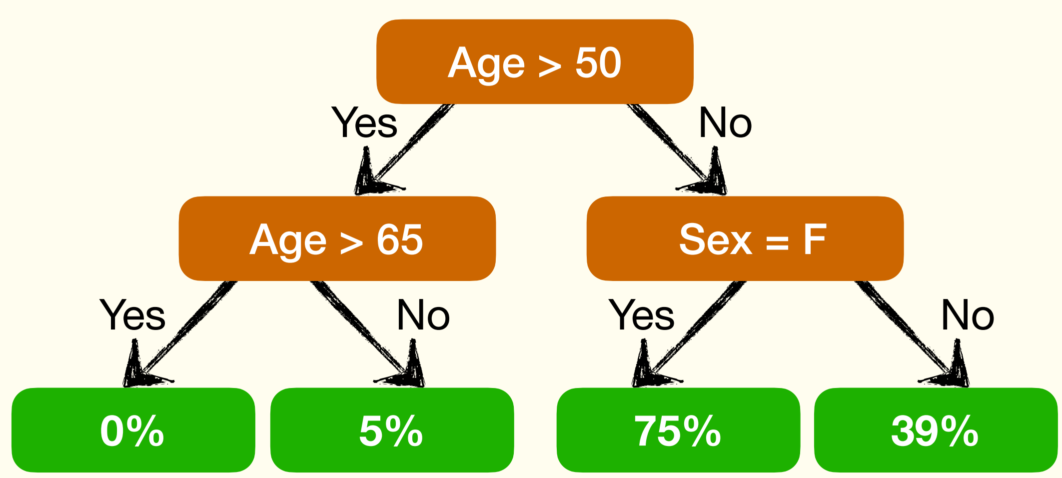
Dose	Age	Sex	Drug Effect
10	25	F	98
5	12	M	44



8 Then we grow the tree just like before, except now for each split we have **3** candidates, Dose, Age, and Sex, and we select whichever gives us the lowest **SSR**…



…until we can no longer subdivide the data any further. At that point, we're done building the **Regression Tree**.



NOTE: The final prediction is Drug Effectiveness.