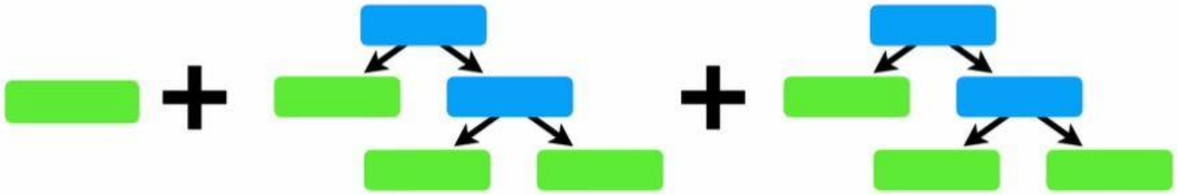
XGBoost
Part 1:
XGBoost Trees for Regression!!!

NOTE: This **StatQuest** assumes that you are already familiar with at least the main ideas of how **Gradient Boost** does **Regression**…

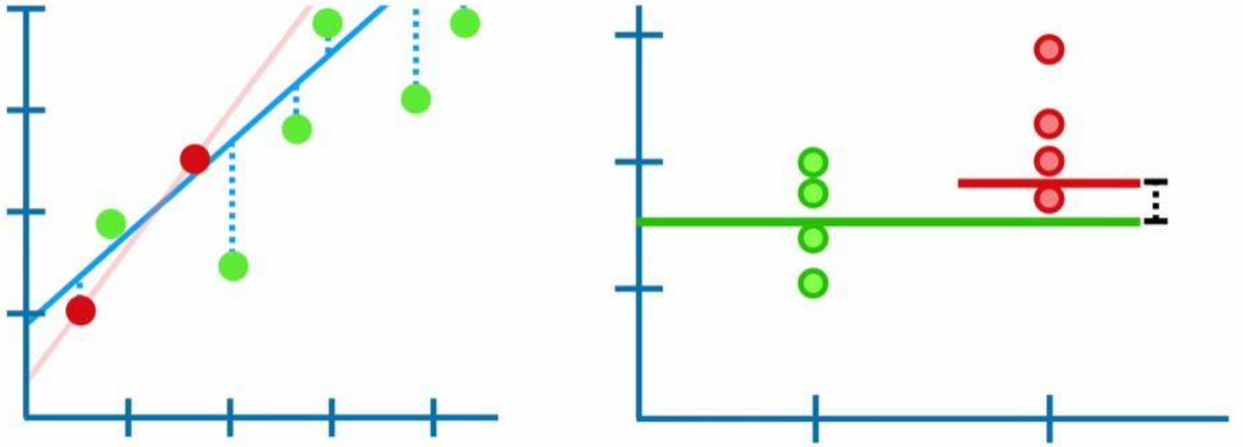
Gradient Boost Part 1…



…Regression Main Ideas!!!

…and you should be familiar with at least the main ideas behind **Regularization**. If not, check out the **'Quests**. The links are in the description below.

Regularization Part 1: Ridge Regression….



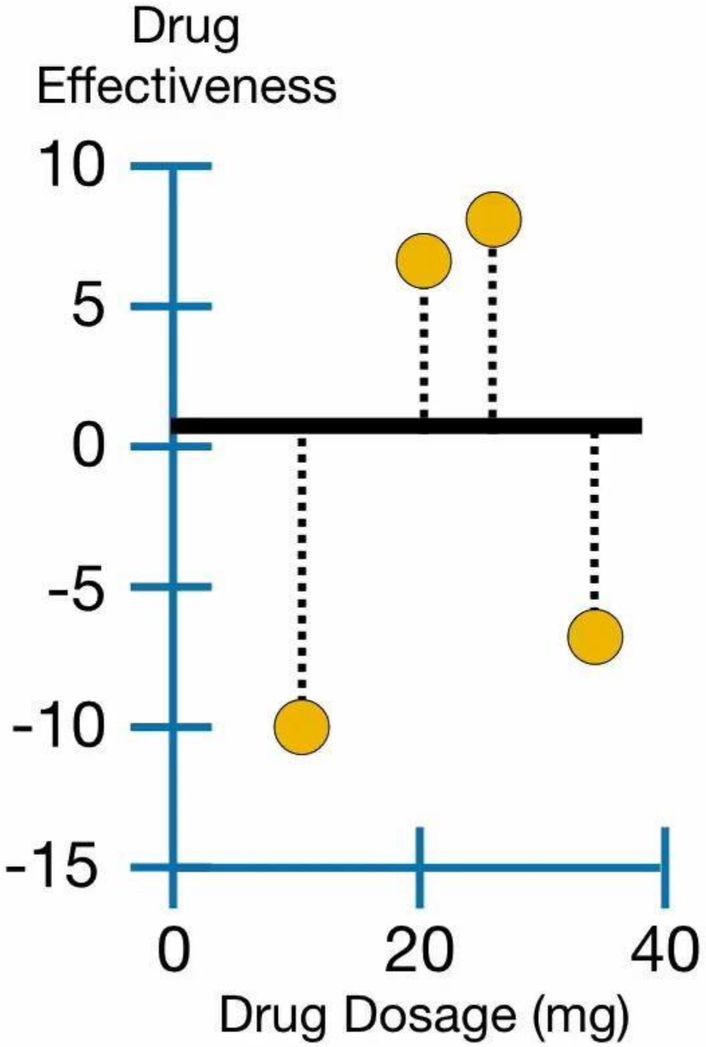
…Clearly Explained!!!

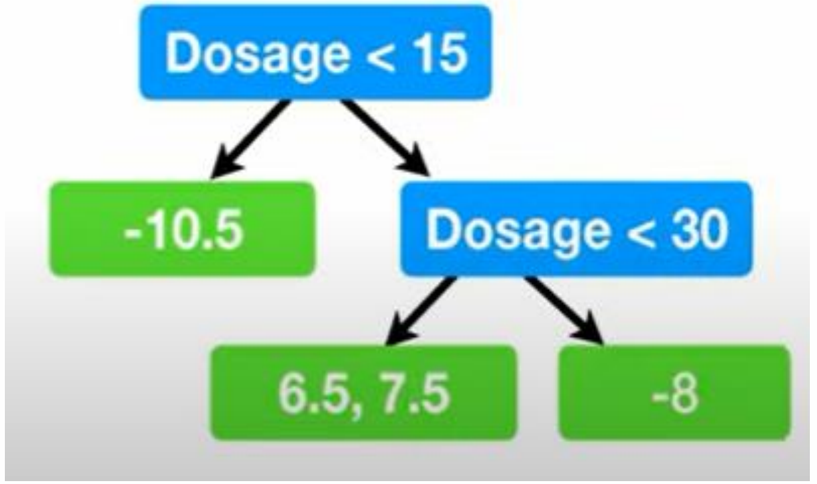
Gradient Boost-(ish)

XGBoost is **EXTREME!!!!!** And that means it's a big **Machine Learning** algorithm with lots of parts.

The good news is that each part is pretty simple and easy to understand, and we'll go through them one step at a time.

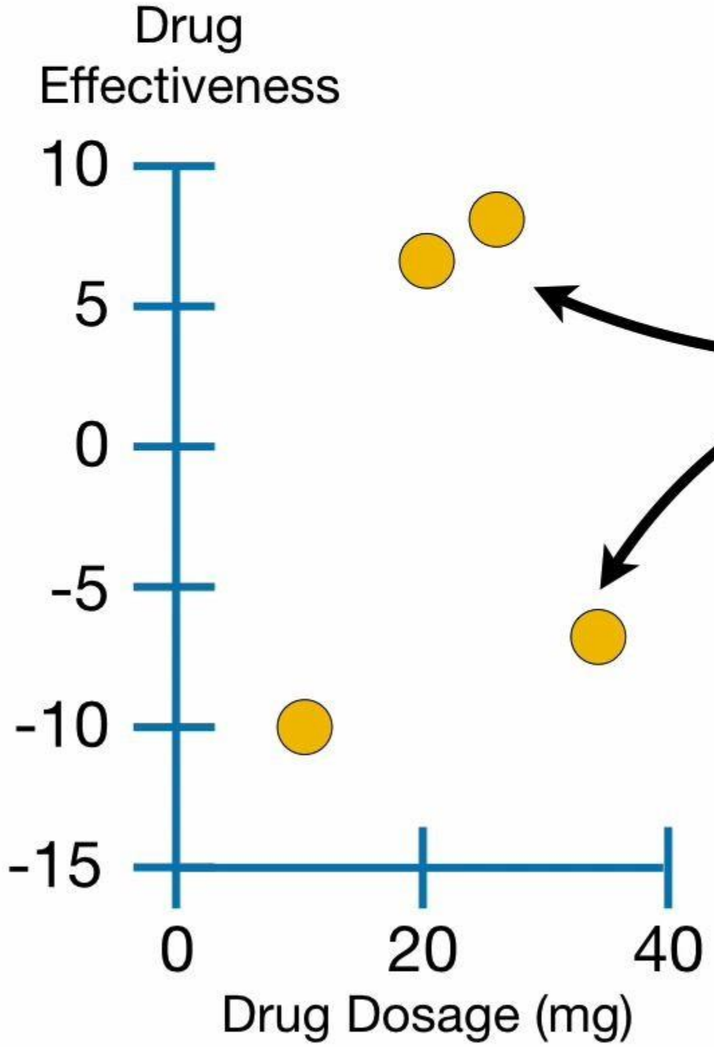
In this **StatQuest**, **Part 1**, we'll build our intuition about how **XGBoost** does **Regression** with its unique trees.



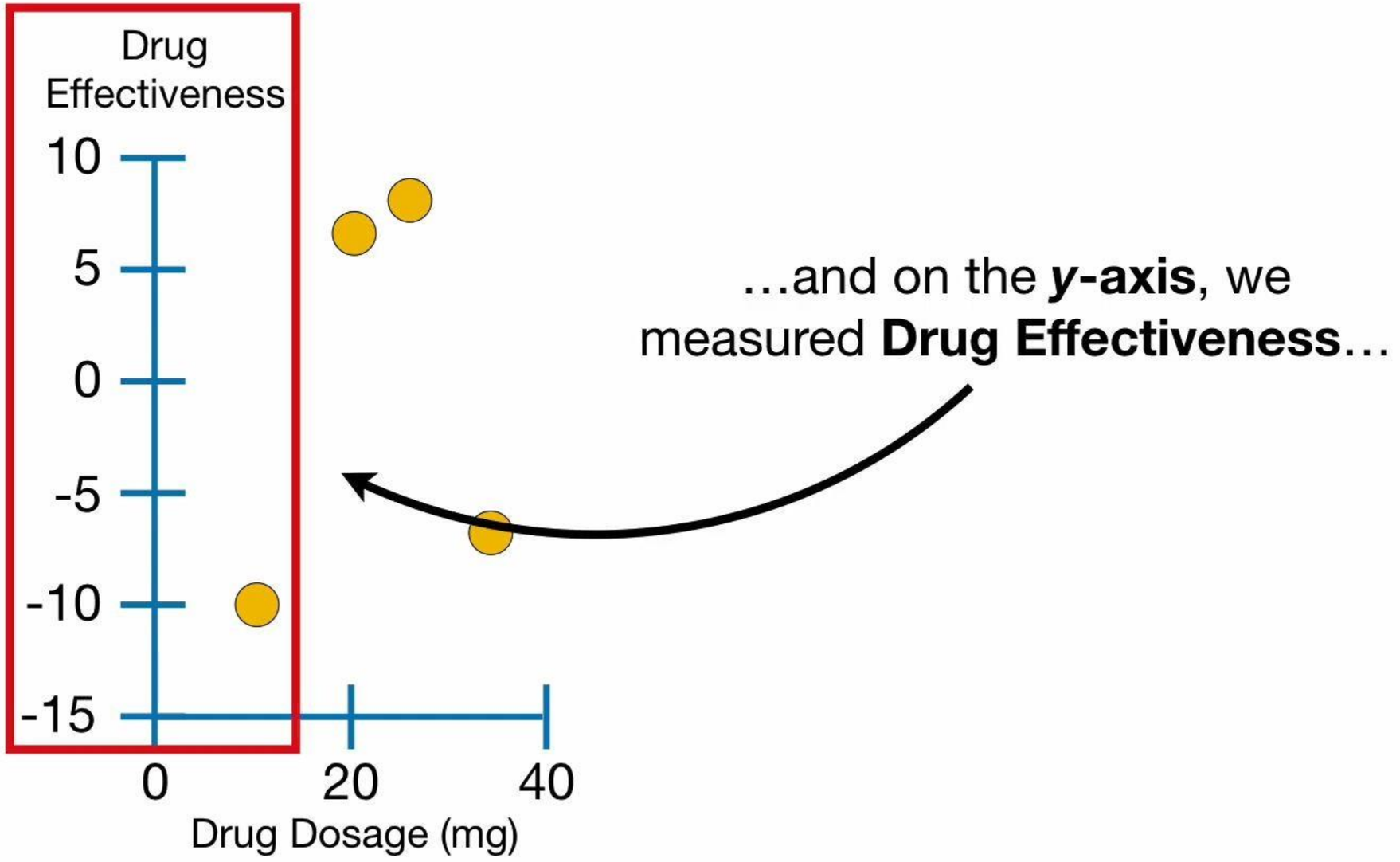


NOTE: XGBoost was designed to be used with large, complicated data sets.

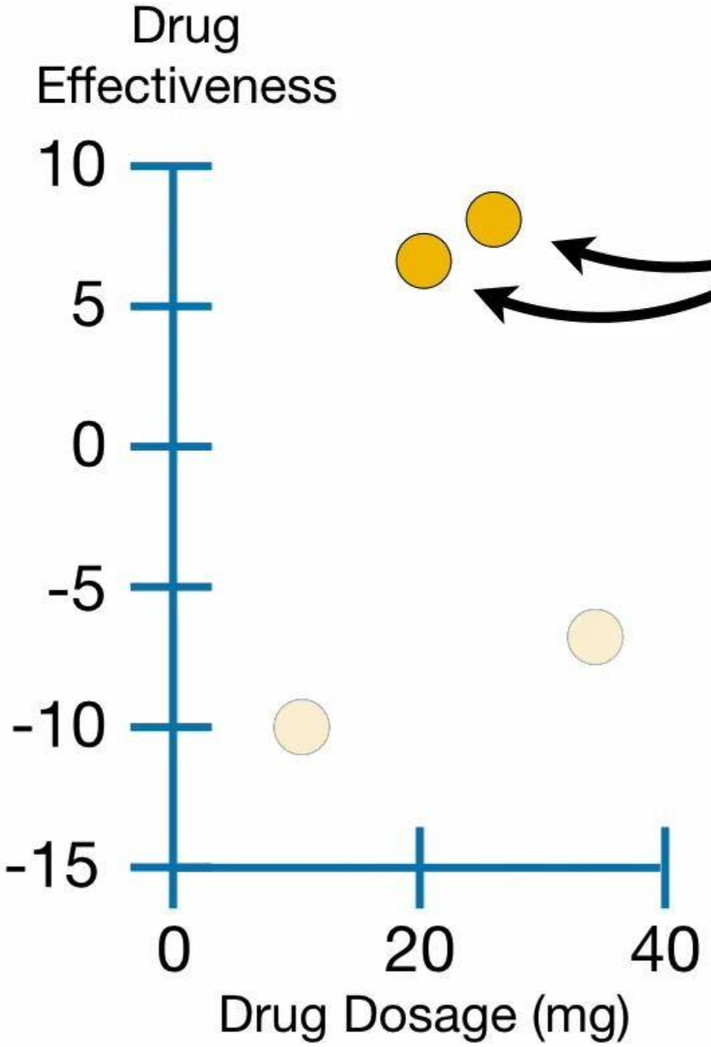
Mass	Age	BP	Color	Movie	Car	Hair	etc.
120	23	102	Brown	T2	Ford	Long	…
150	25	98	Brown	Frozen	Kia	Short	…
165	22	130	Black	Spiderman	Ford	Short	…
123	45	98	Red	T2	Kia	Long	…
156	33`	78	Brown	Frozen 2	Ford	Long	…
…	…	…	…	…	…	…	…



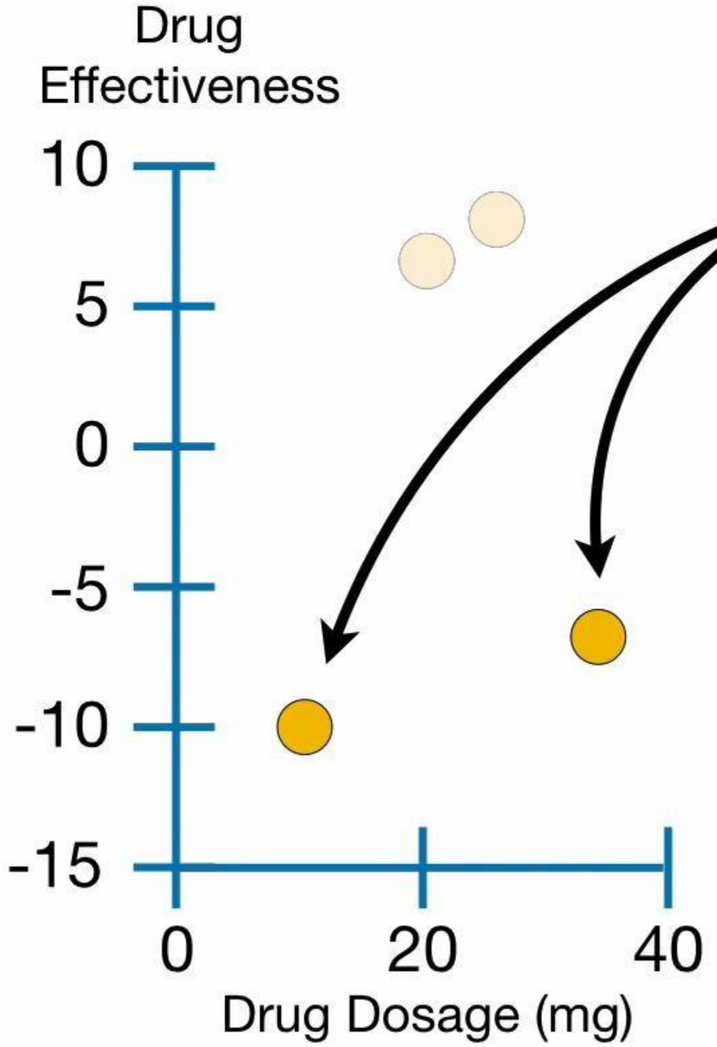
However, to keep the examples from getting out of hand, we will use use this super simple **Training Data.**



…and on the ***y*-axis**, we measured **Drug Effectiveness**…



These two observations have relatively large positive values for **Drug Effectiveness**, and that means that the drug was helpful.

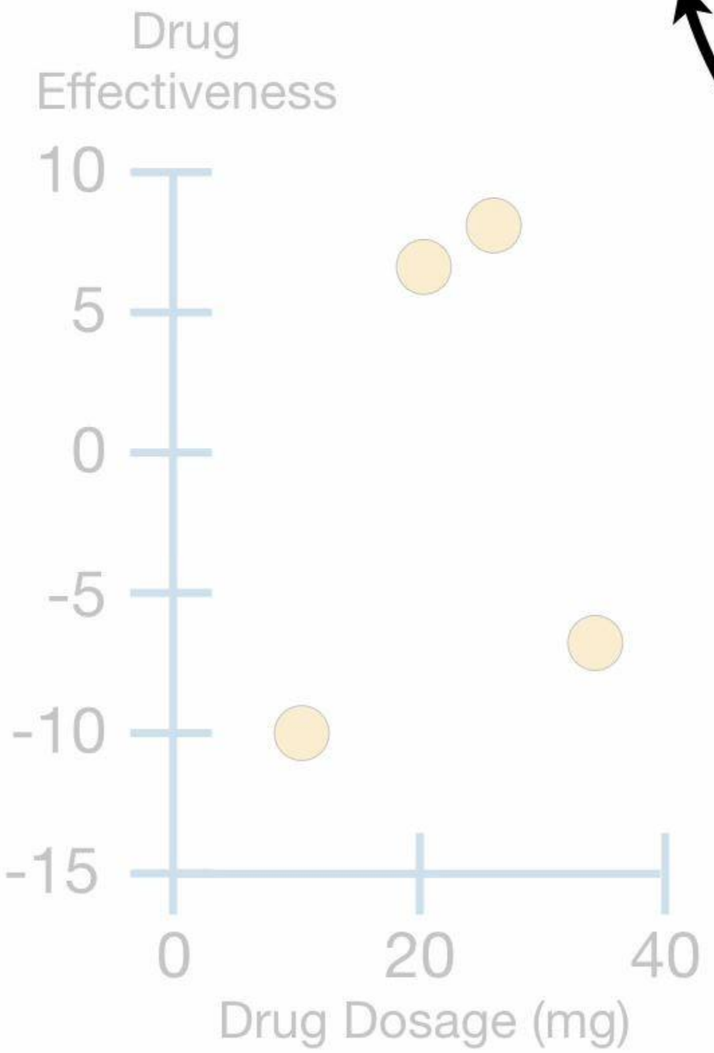


These two observations have relatively large negative values for **Drug Effectiveness**, and that means that the drug did more harm than good.

Predicted Drug Effectiveness

0.5

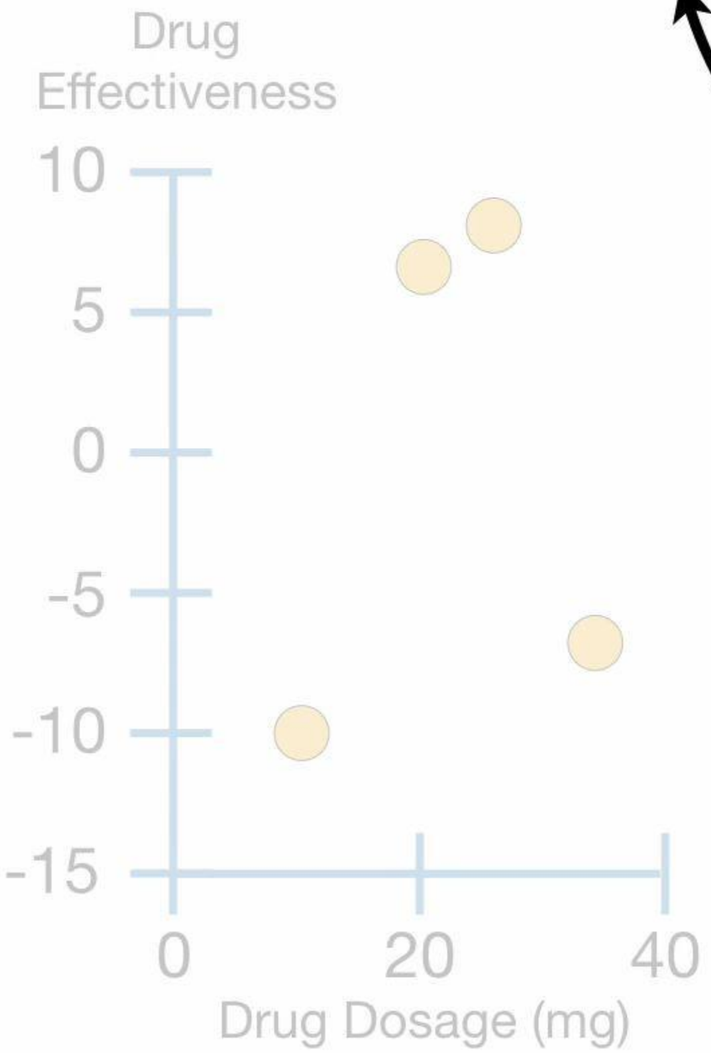
The very first step in fitting **XGBoost** to the **Training Data** is to make an initial prediction.



Predicted Drug Effectiveness

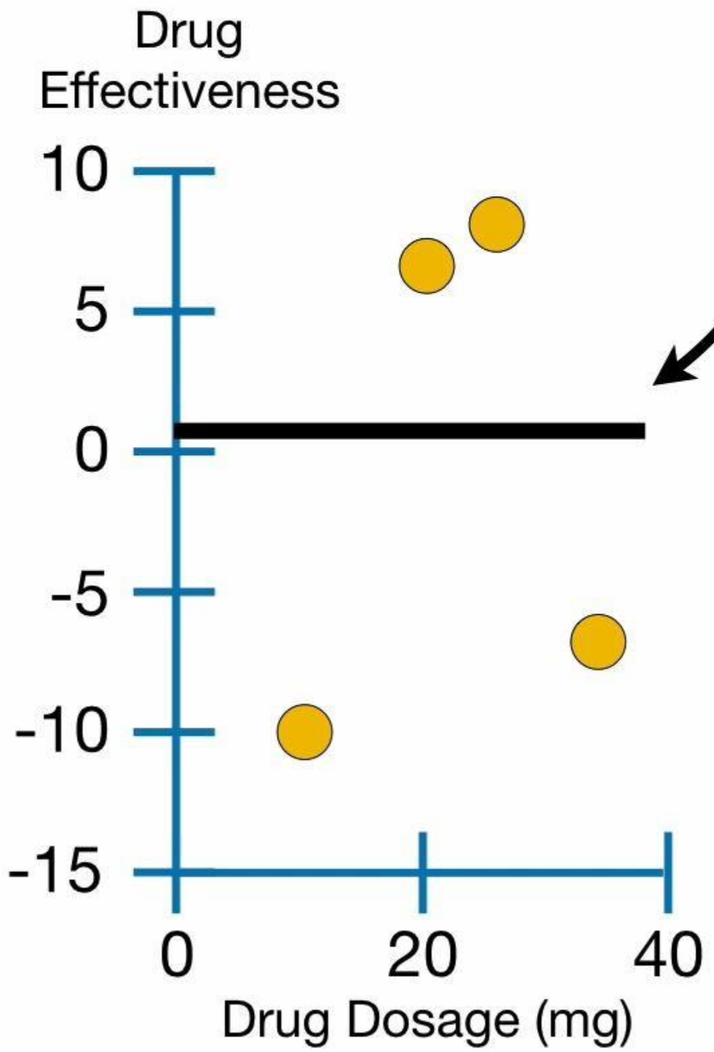
0.5

This prediction can be anything, but by default it is **0.5**, regardless of whether you are using **XGBoost** for **Regression** or **Classification**.



Predicted Drug Effectiveness

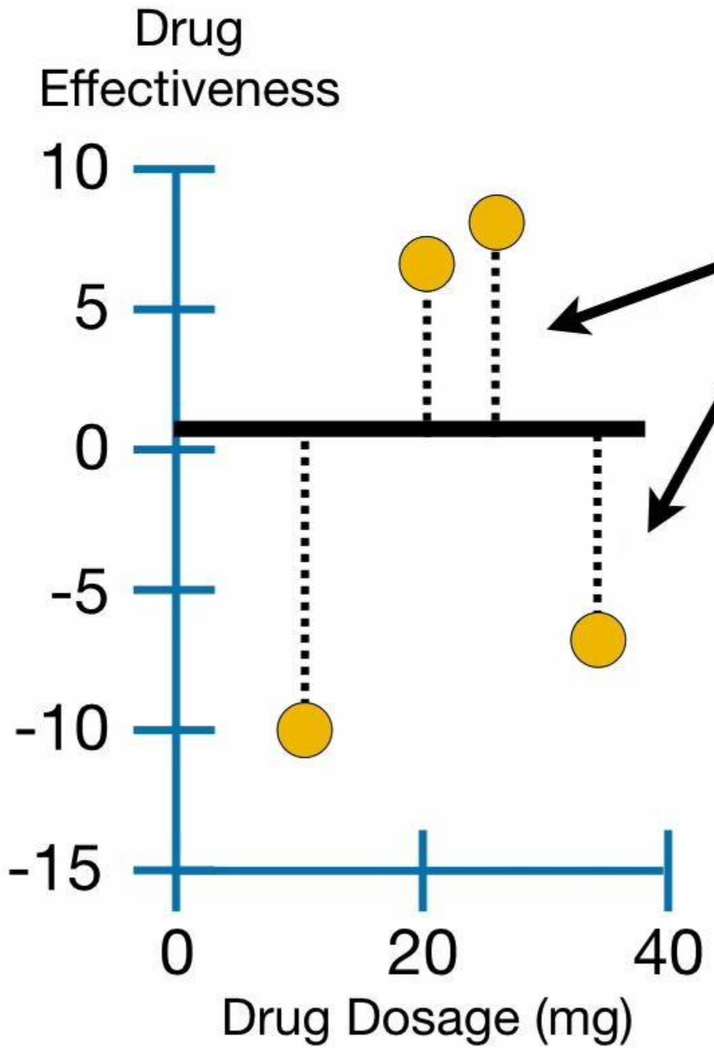
0.5



The prediction, **0.5**, corresponds to this **thick, black, horizontal line**…

Predicted Drug Effectiveness

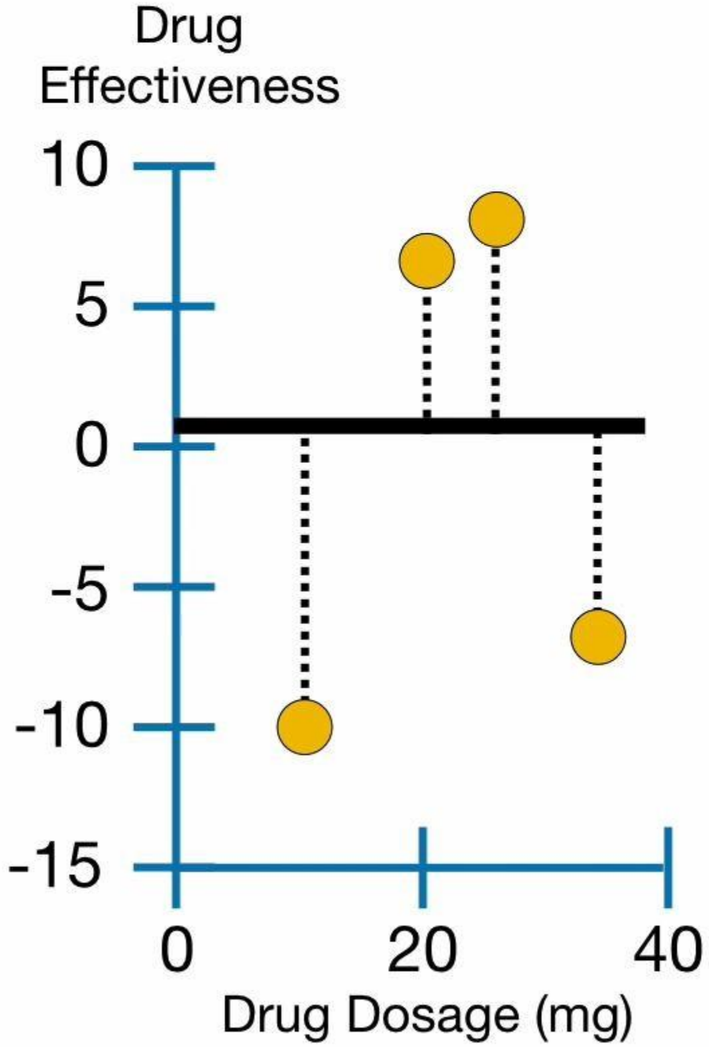
0.5

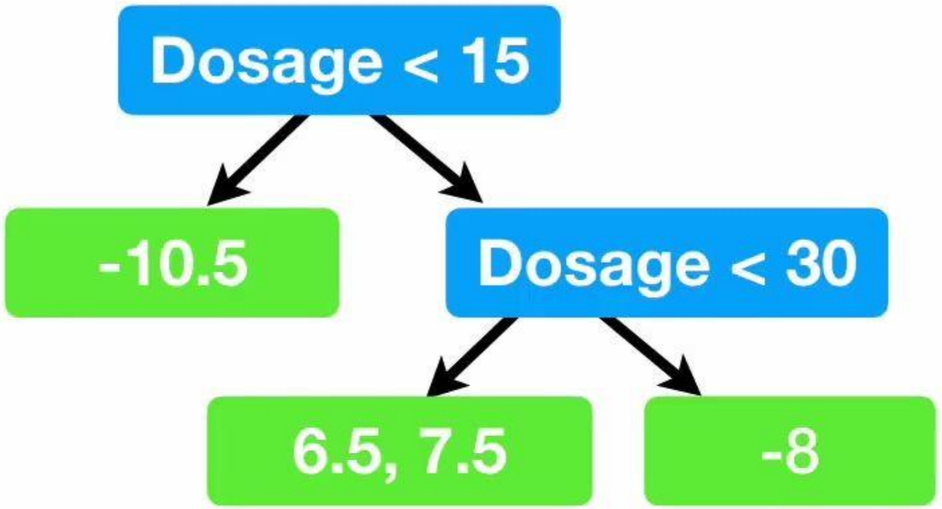


…and the **Residuals**, the differences between the **Observed** and **Predicted** values, show us how good the initial prediction is.

Predicted Drug Effectiveness

0.5

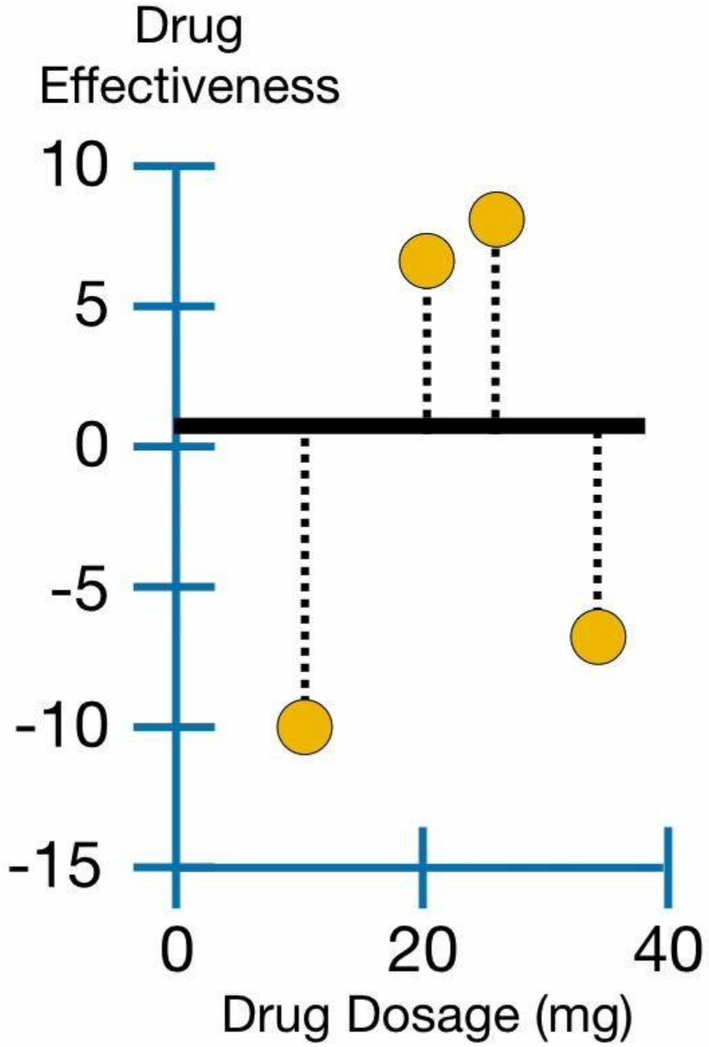


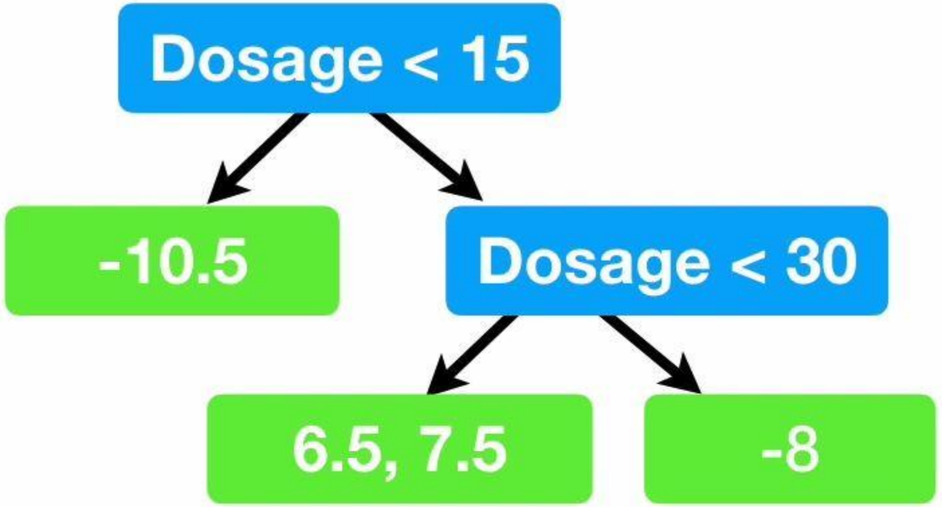


Now, just like unextreme **Gradient Boost**, **XGBoost** fits a **Regression Tree** to the residuals…

Predicted Drug Effectiveness

0.5

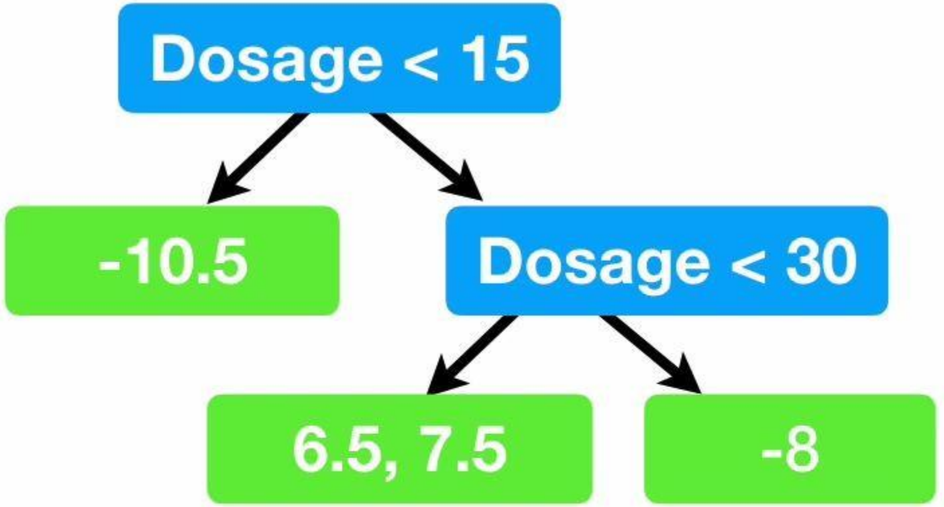


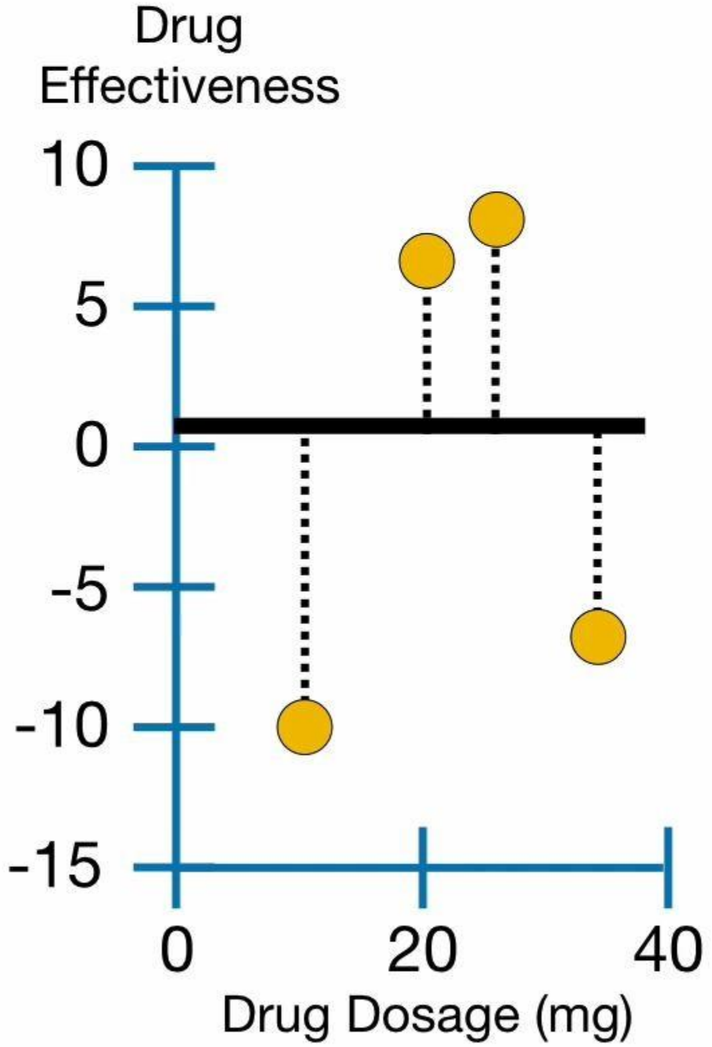


However, unlike unextreme **Gradient Boost**, which typically uses regular, off-the-shelf, **Regression Trees**…

Predicted Drug Effectiveness

0.5

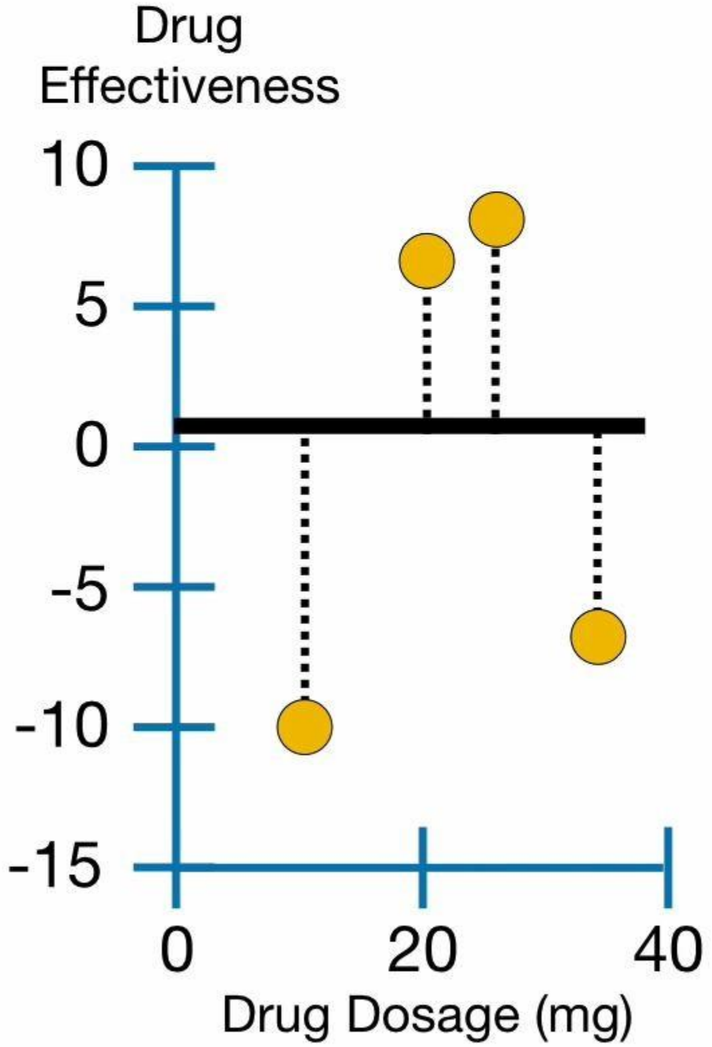


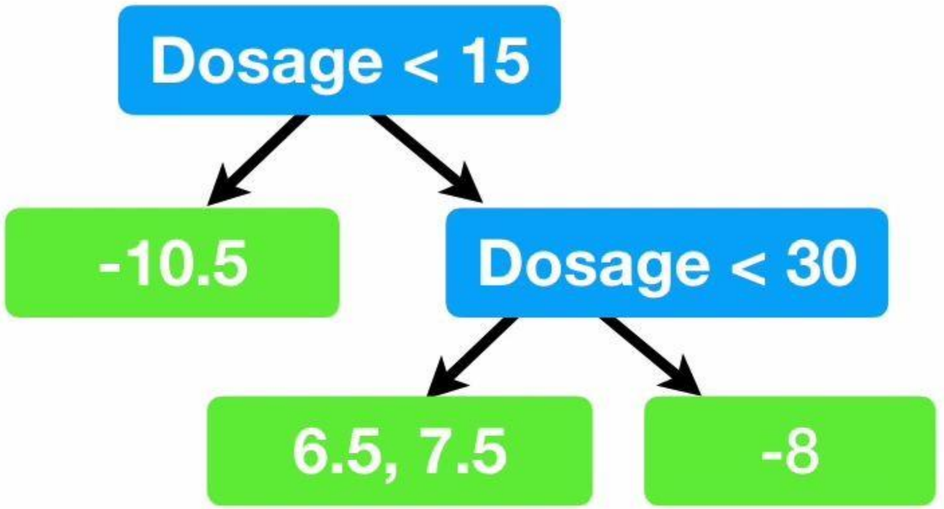


So let's talk about how to build an **XGBoost Tree** for **Regression**.

Predicted Drug Effectiveness

0.5

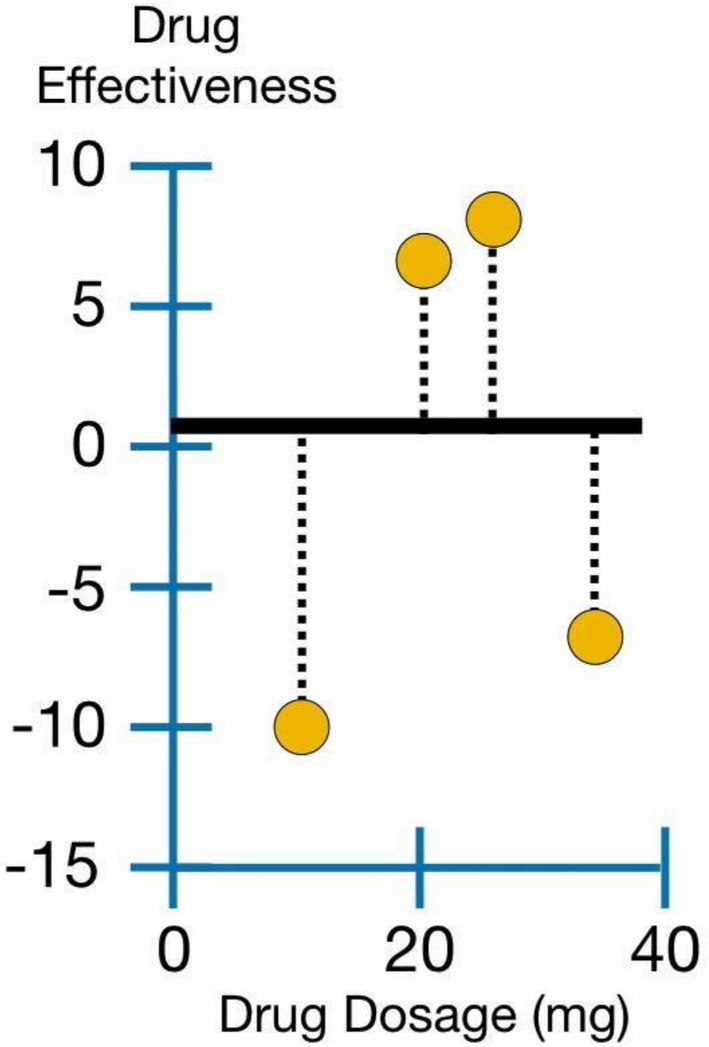




NOTE: There are many ways to build **XGBoost Trees**. This video focuses on the most common way to build them for **Regression**.

Predicted Drug Effectiveness

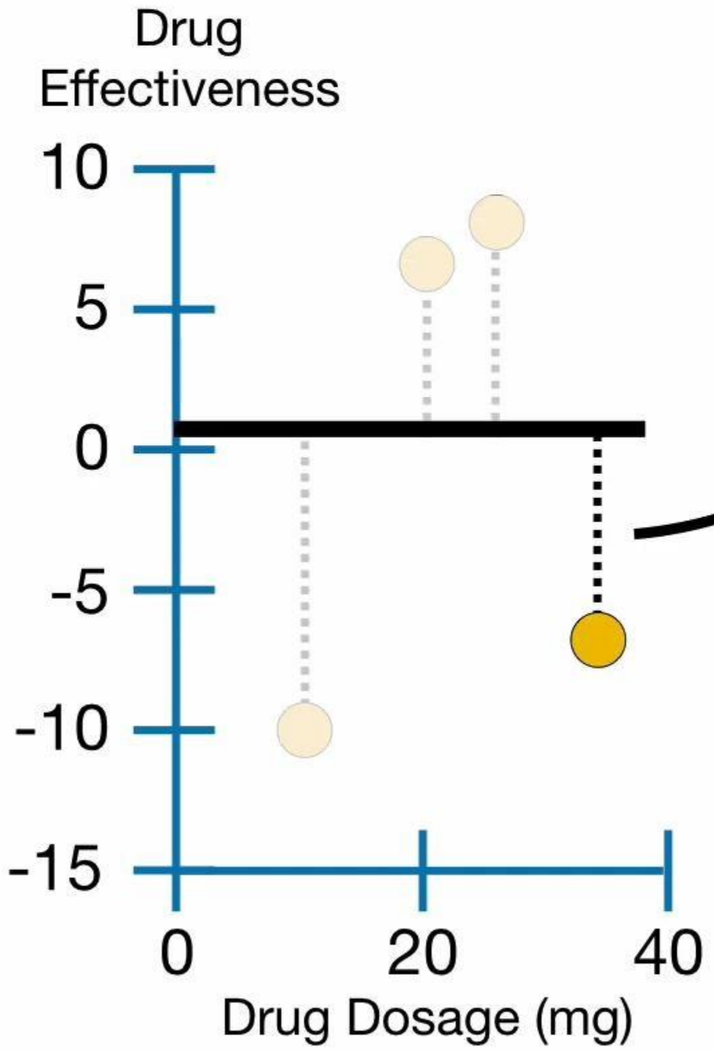
0.5



Each tree starts out as a single leaf…

Predicted Drug Effectiveness

0.5



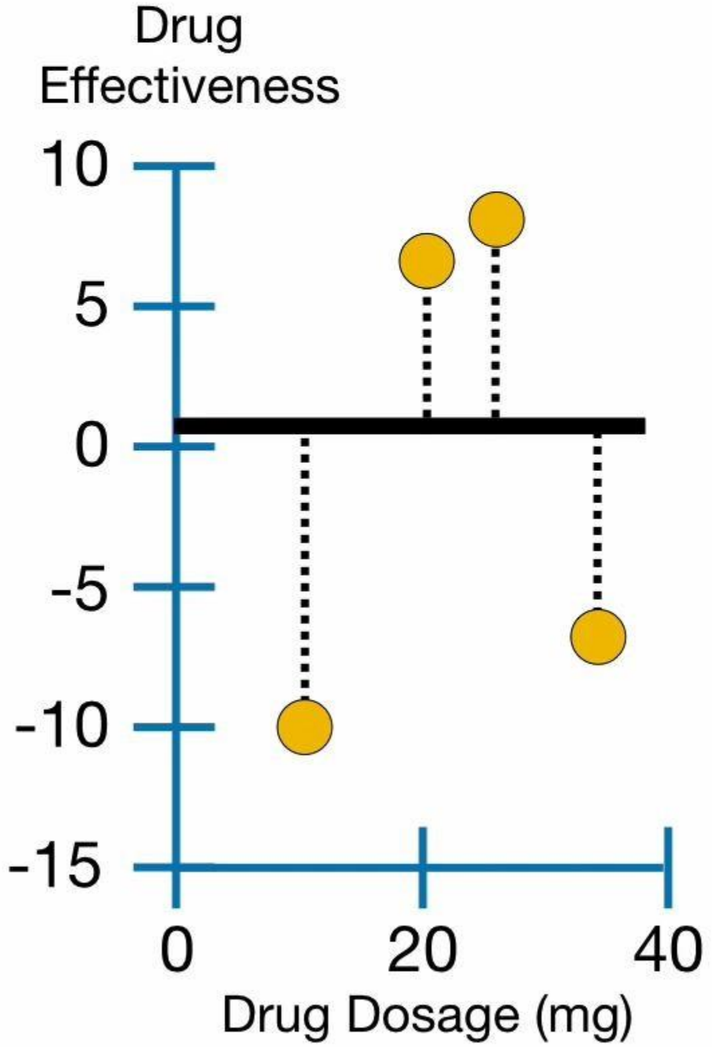
-10.5, 6.5, 7.5, -7.5

…and all of the **Residuals** go to the leaf.

Predicted Drug Effectiveness

0.5

-10.5, 6.5, 7.5, -7.5

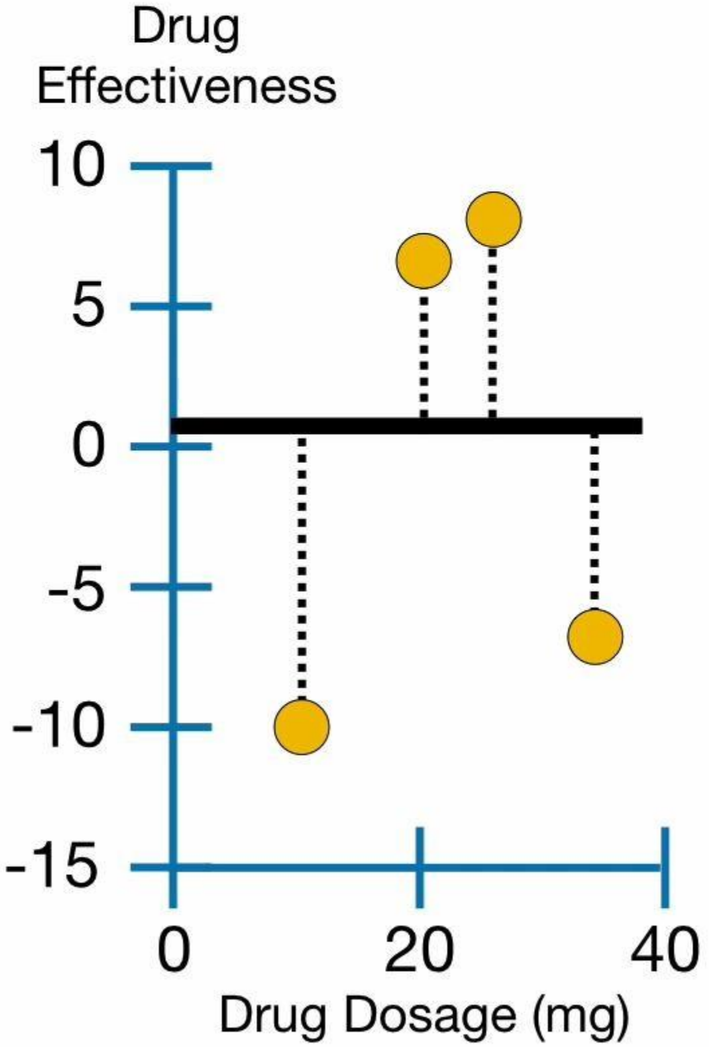


Now we calculate a **Quality Score**, or **Similarity Score**, for the **Residuals**.

Predicted Drug Effectiveness

0.5

-10.5, 6.5, 7.5, -7.5

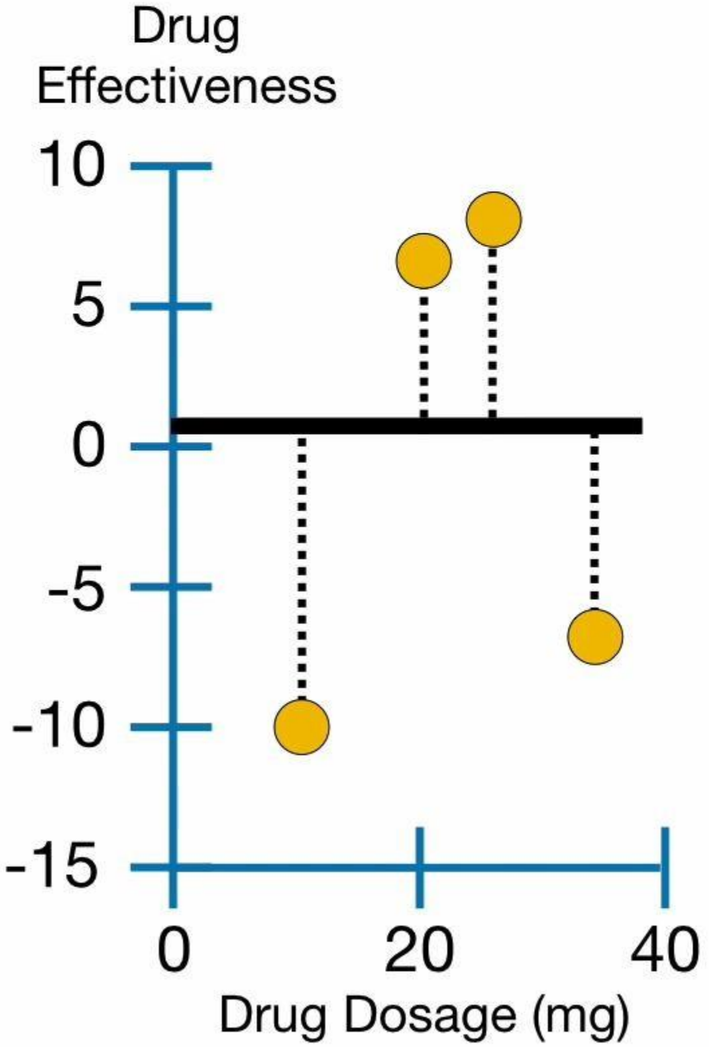


$$\text{Similarity Score} = \frac{\text{Sum of Residuals, Squared}}{\text{Number of Residuals} + \lambda}$$

Predicted Drug Effectiveness

0.5

-10.5, 6.5, 7.5, -7.5



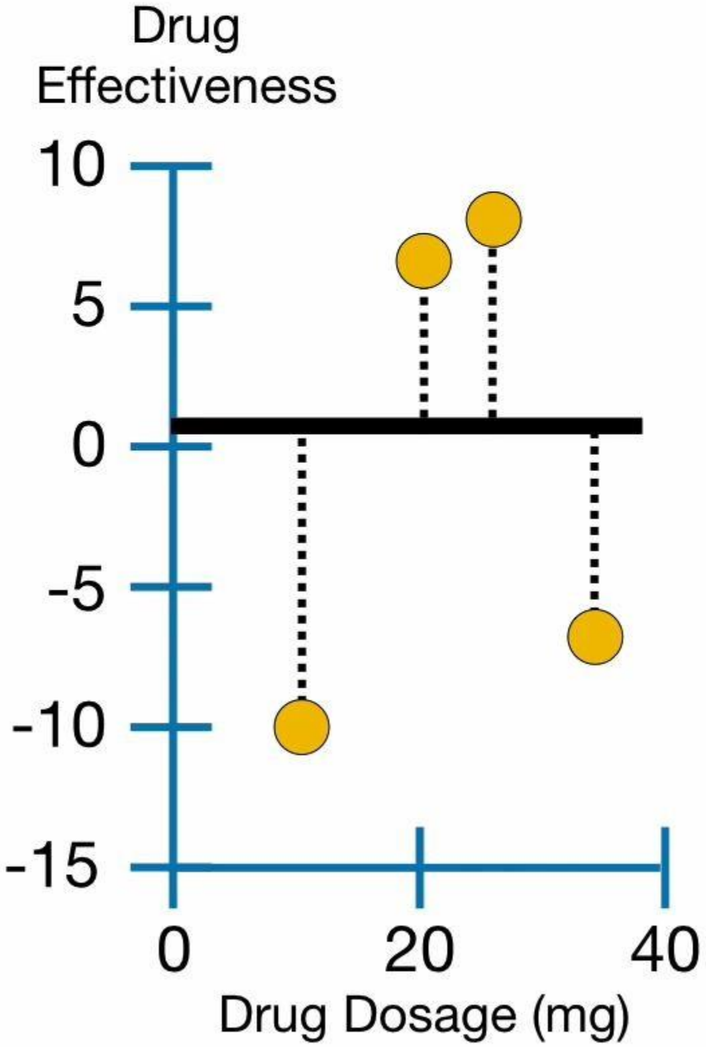
$$\text{Similarity Score} = \frac{\text{Sum of Residuals, Squared}}{\text{Number of Residuals} + \lambda}$$

NOTE: λ (lambda) is a **Regularization** parameter, and we'll talk more about that later.

Predicted Drug Effectiveness

0.5

-10.5, 6.5, 7.5, -7.5



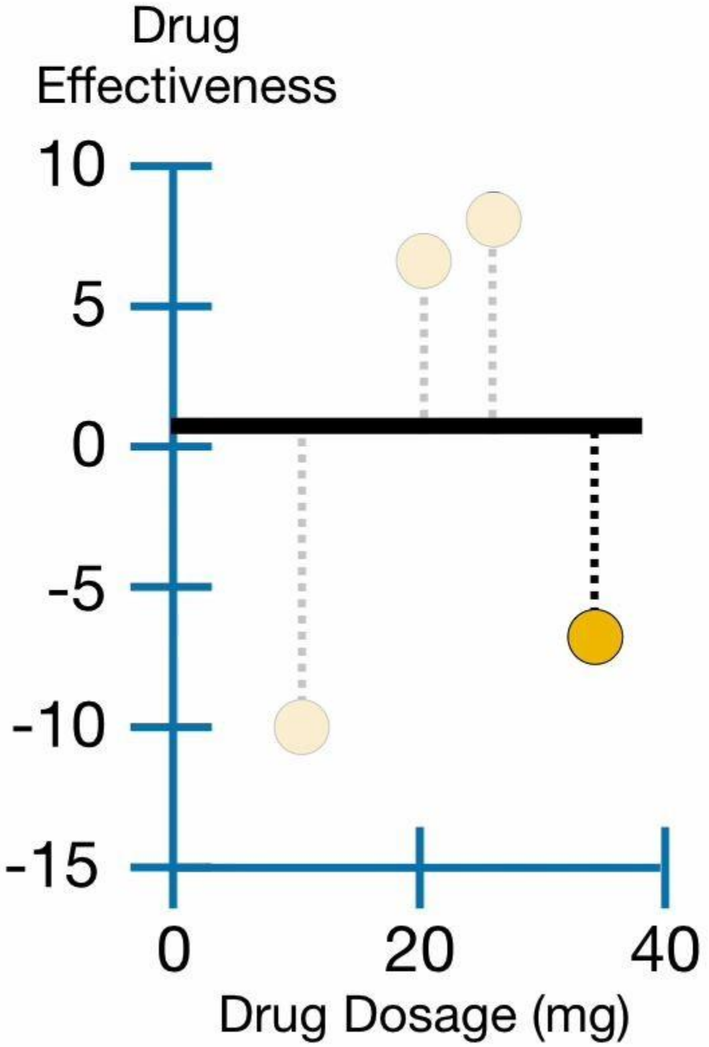
$$\text{Similarity Score} = \frac{\text{Sum of Residuals, Squared}}{\text{Number of Residuals} + 0}$$

Now we plug the **4 Residuals** into the numerator…

Predicted Drug Effectiveness

0.5

-10.5, 6.5, 7.5, -7.5



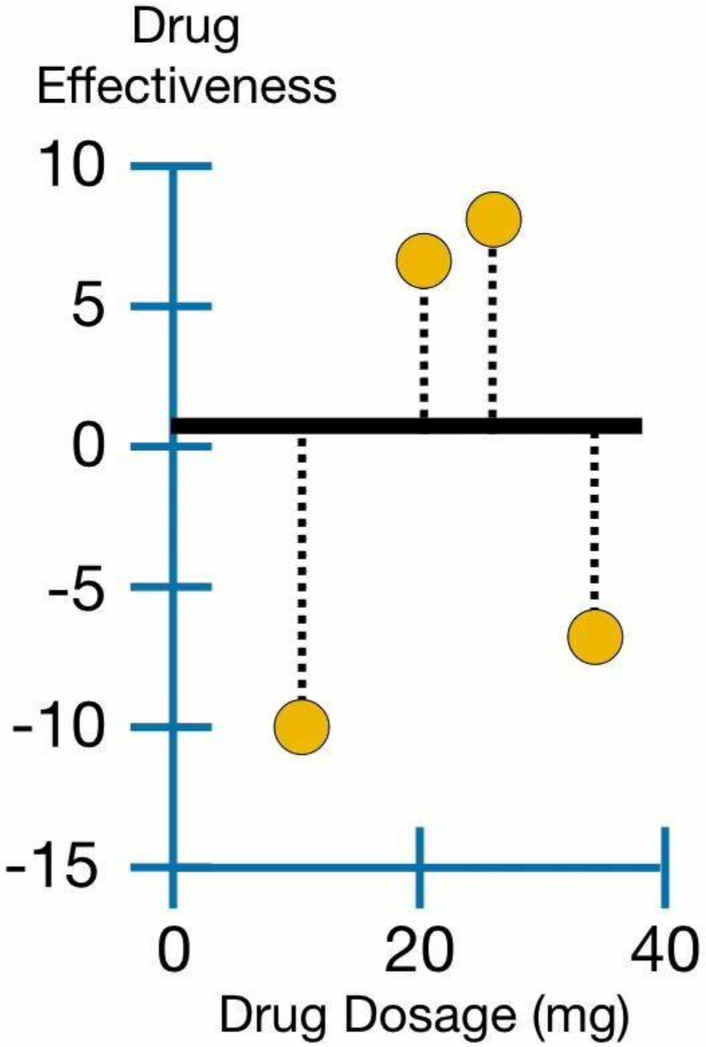
$$\text{Similarity Score} = \frac{(-10.5 + 6.5 + 7.5 + -7.5)^2}{4 + \mathbf{0}}$$

…and since there are **4** **Residuals** in the leaf, we put a **4** in the denominator.

Predicted Drug Effectiveness

0.5

-10.5, 6.5, 7.5, -7.5



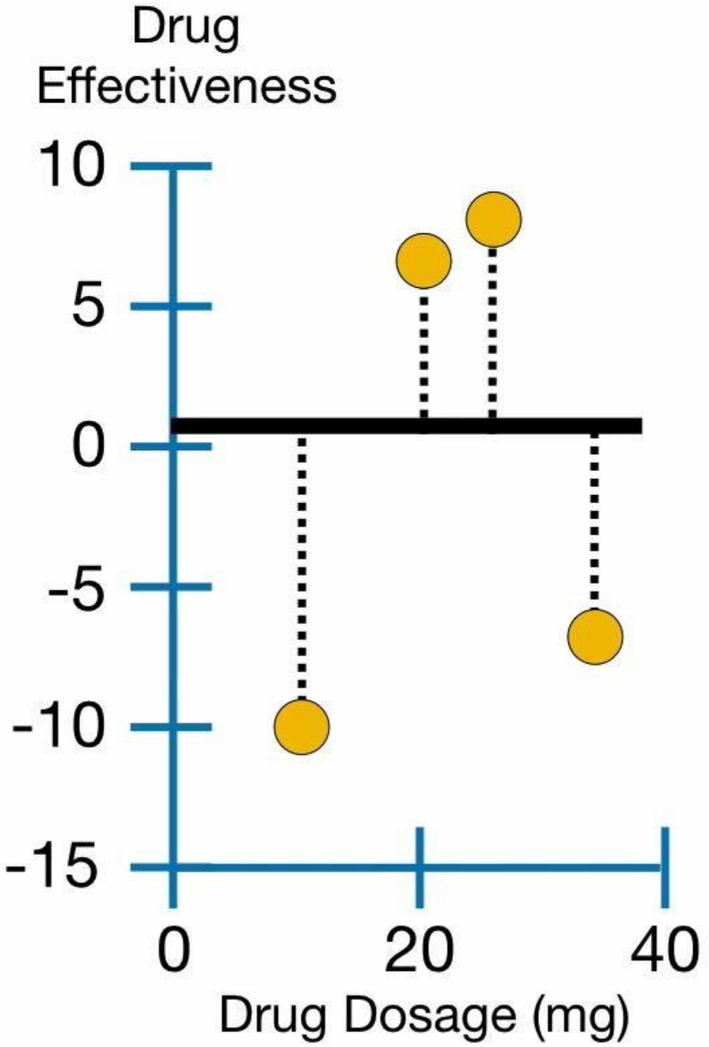
$$\text{Similarity Score} = \frac{(-10.5 + 6.5 + 7.5 + -7.5)^2}{4 + 0}$$

NOTE: Because we do not square the **Residuals** *before* we add them together in the numerator, **7.5** and **-7.5** cancel each other out.

Predicted Drug Effectiveness

0.5

-10.5, 6.5, 7.5, -7.5



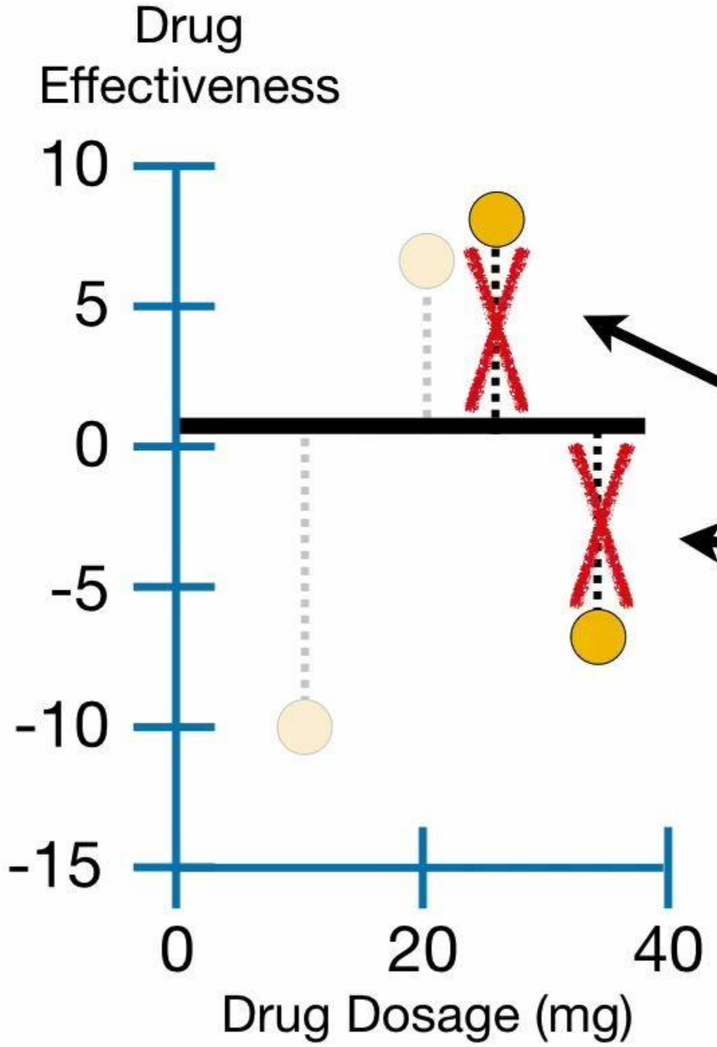
$$\text{Similarity Score} = \frac{(-10.5 + 6.5 + 0)^2}{4 + 0}$$

NOTE: Because we do not square the **Residuals** *before* we add them together in the numerator, **7.5** and **-7.5** cancel each other out.

Predicted Drug Effectiveness

0.5

-10.5, 6.5, 7.5, -7.5



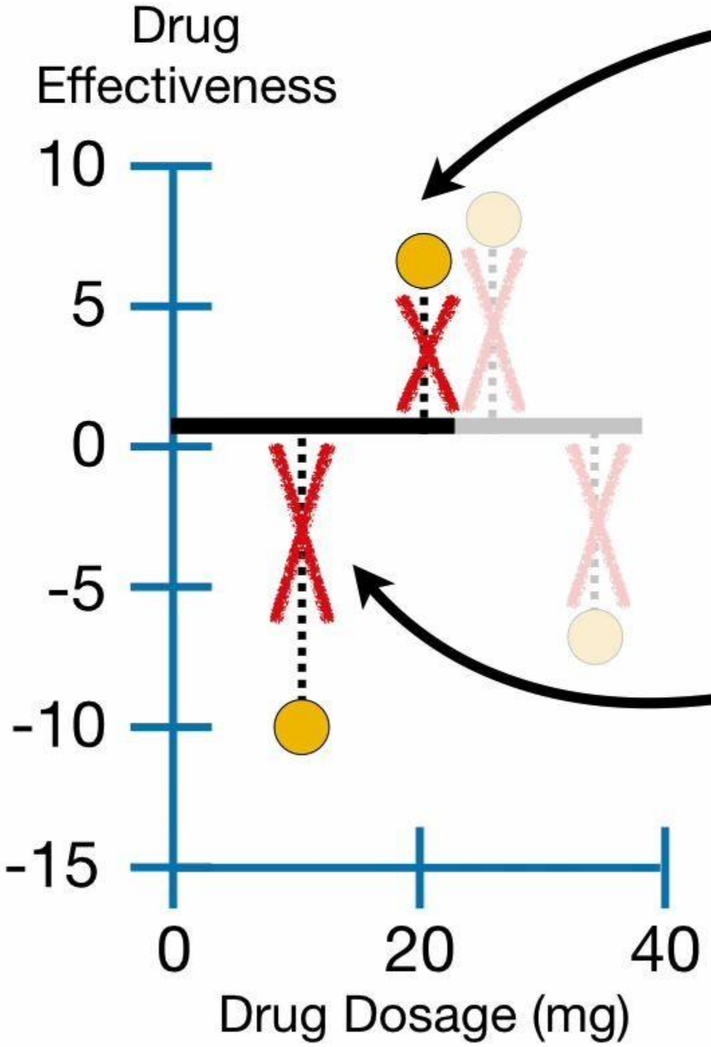
$$\text{Similarity Score} = \frac{(-10.5 + 6.5 + 0)^2}{4 + 0}$$

…they cancel each other out.

Predicted Drug Effectiveness

0.5

-10.5, 6.5, 7.5, -7.5



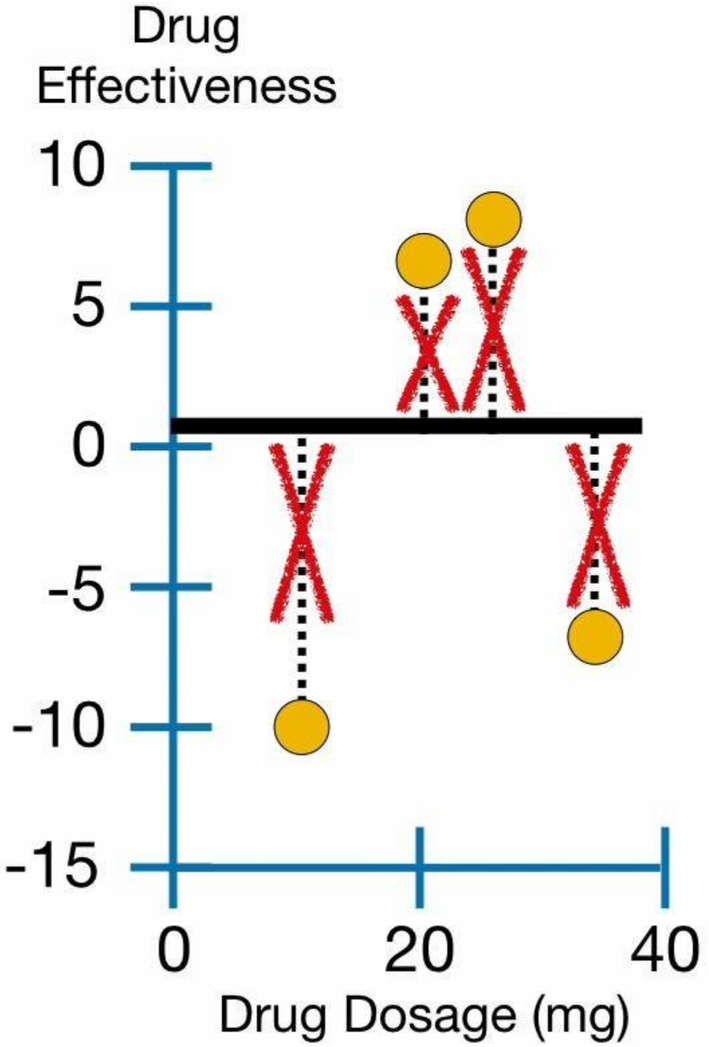
$$\text{Similarity Score} = \frac{(-10.5 + 6.5 - 0)^2}{4 + 0}$$

Likewise, **6.5** cancels out most of **-10.5**.

Predicted Drug Effectiveness

0.5

-10.5, 6.5, 7.5, -7.5

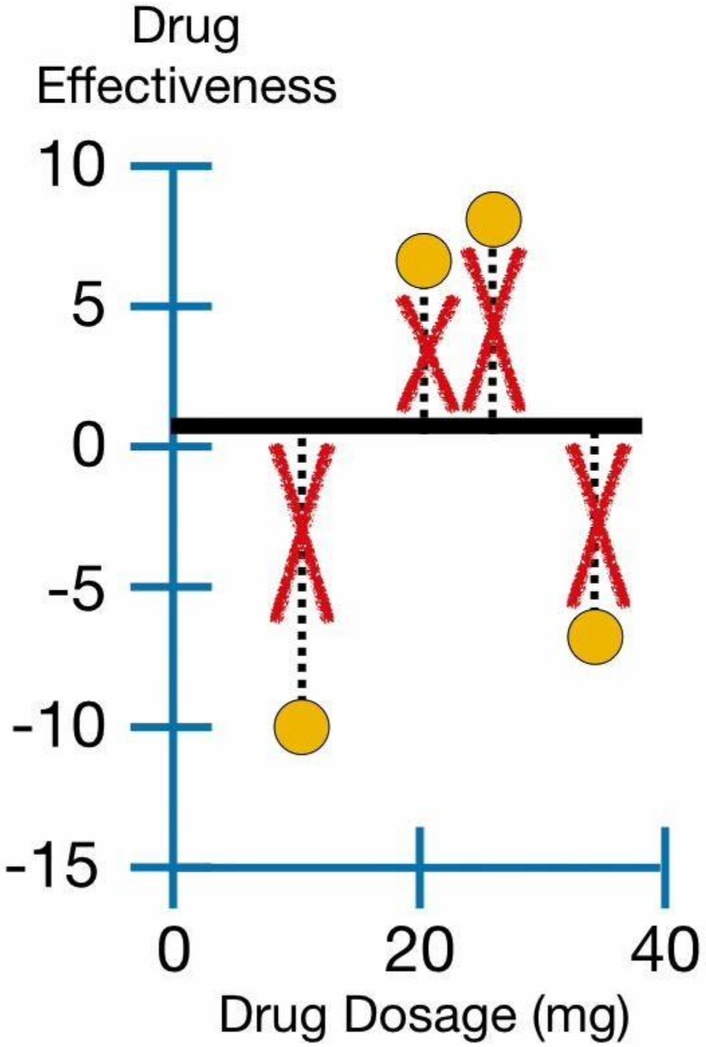


$$\text{Similarity Score} = \frac{(-4)^2}{4 + \mathbf{0}}$$

Thus, the **Similarity Score** for the **Residuals** in the root = **4**.

Predicted Drug Effectiveness

0.5



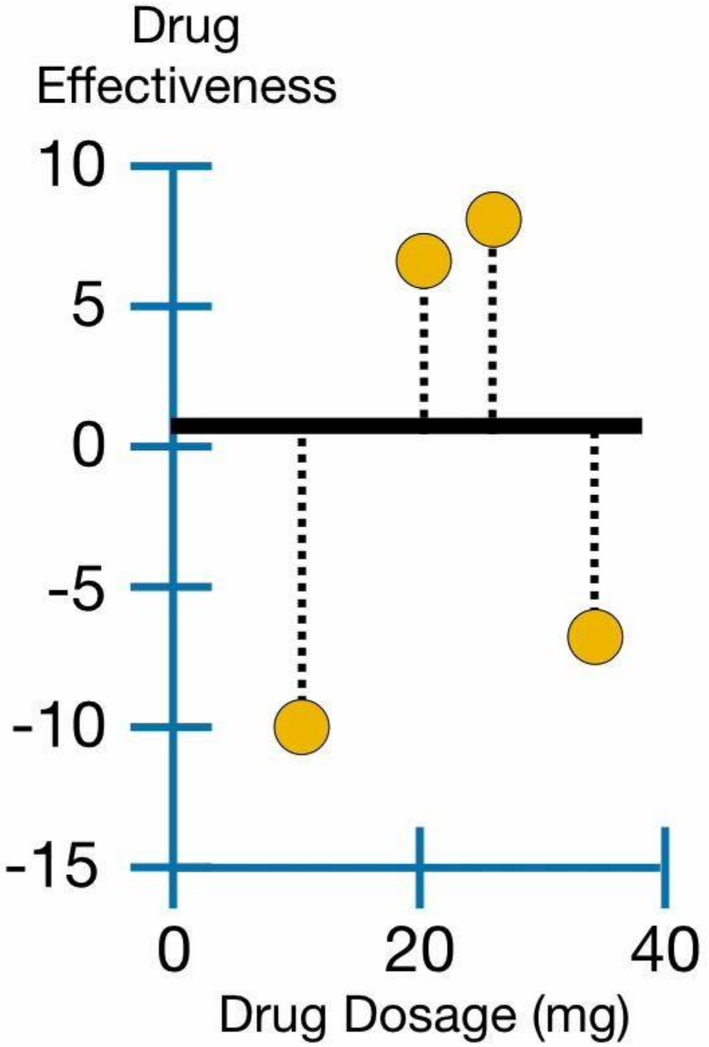
-10.5, 6.5, 7.5, -7.5 Similarity = 4

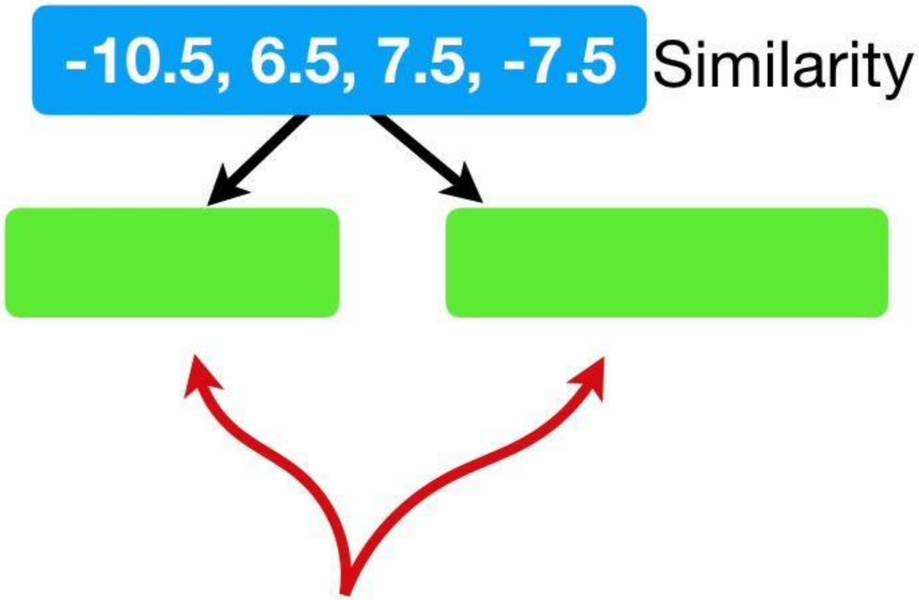
$$\text{Similarity Score} = \frac{(-4)^2}{4 + 0} = 4$$

So let's put **Similarity = 4** up here to can keep track of it.

Predicted Drug Effectiveness

0.5



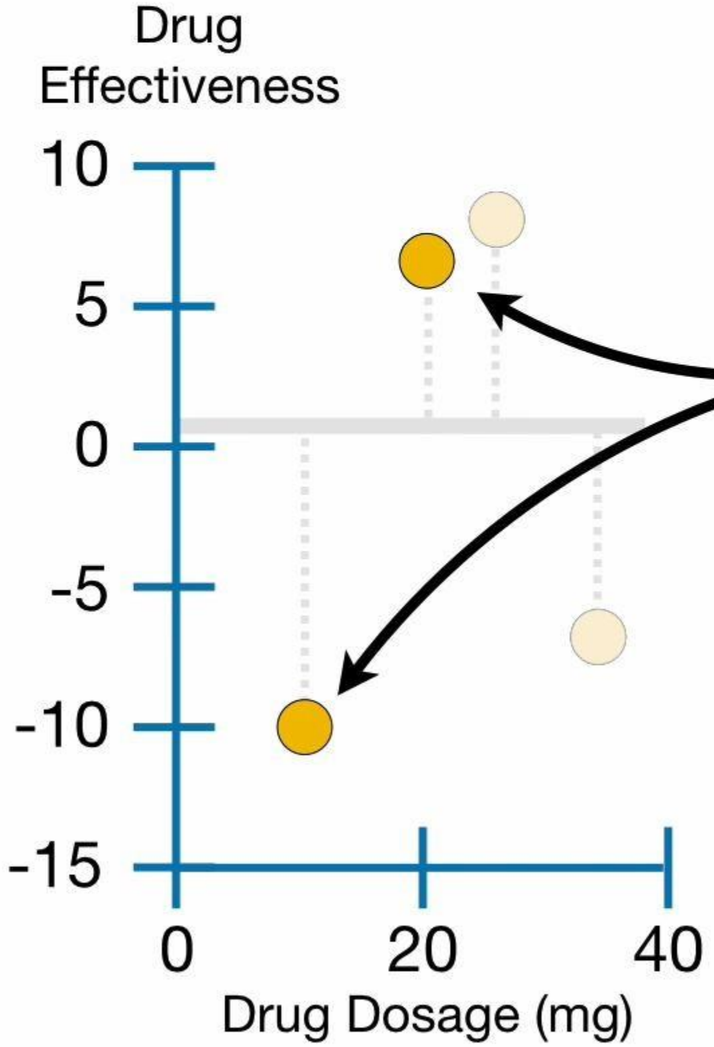


Now the question is whether or not we can do a better job clustering similar **Residuals** if we split them into two groups.

Predicted Drug Effectiveness

0.5

-10.5, 6.5, 7.5, -7.5 Similarity = 4

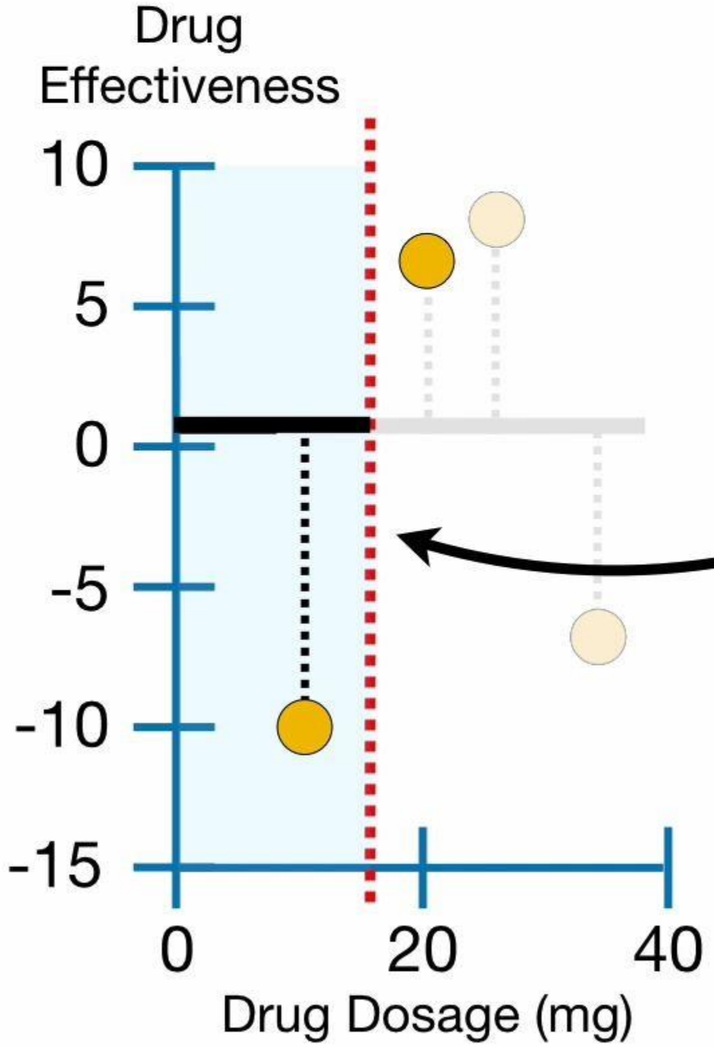


To answer this, we first focus on the two observations with the lowest **Dosages**.

Predicted Drug Effectiveness

0.5

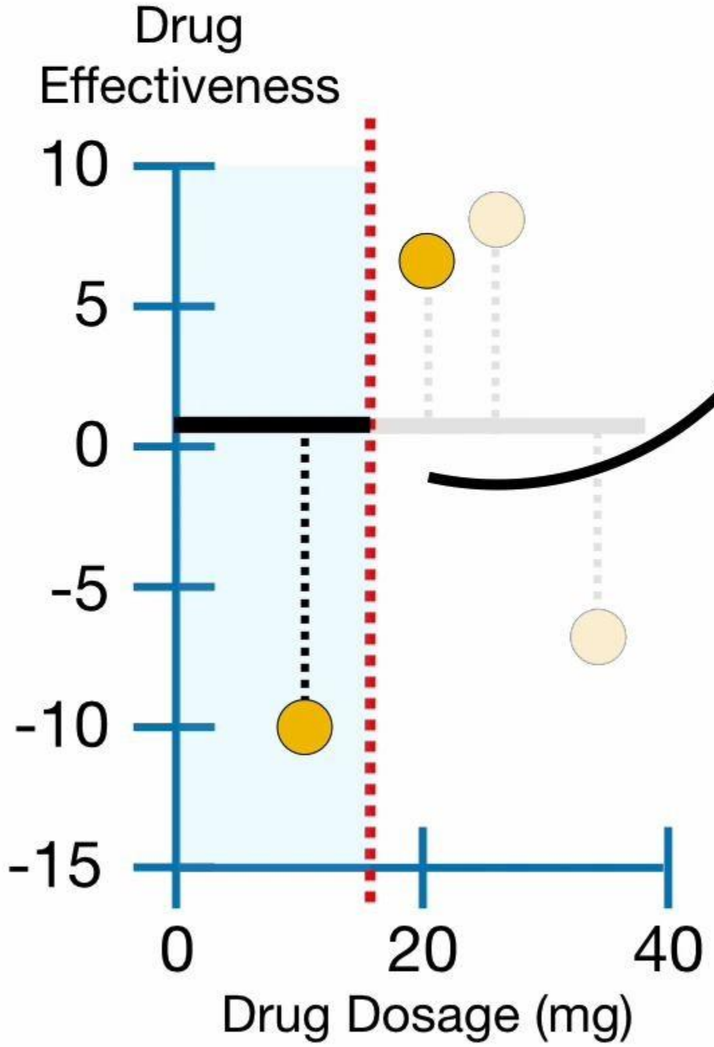
-10.5, 6.5, 7.5, -7.5 Similarity = 4



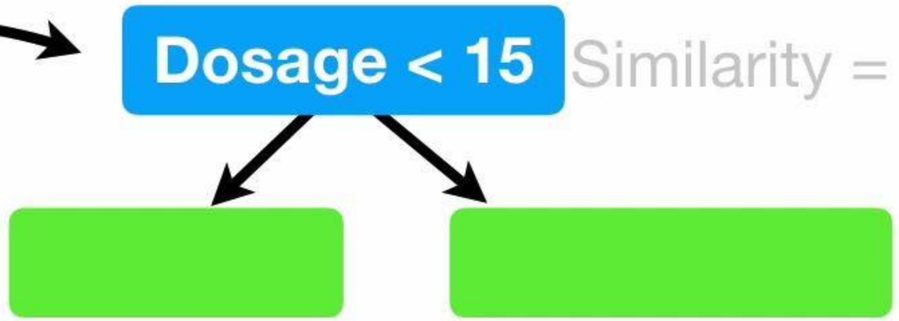
Their average **Dosage** is **15**, and that corresponds to this **dotted red line**.

Predicted Drug Effectiveness

0.5

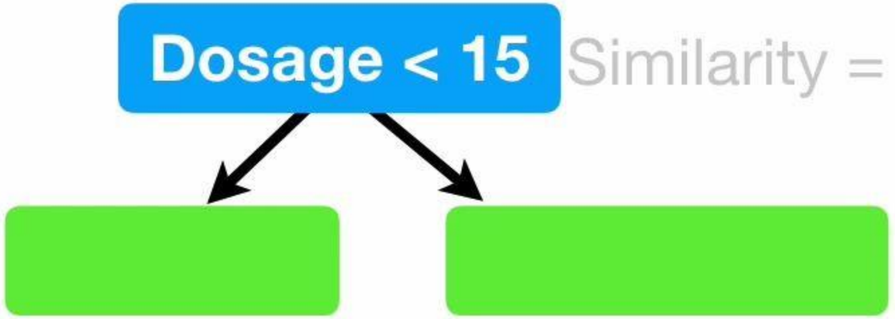


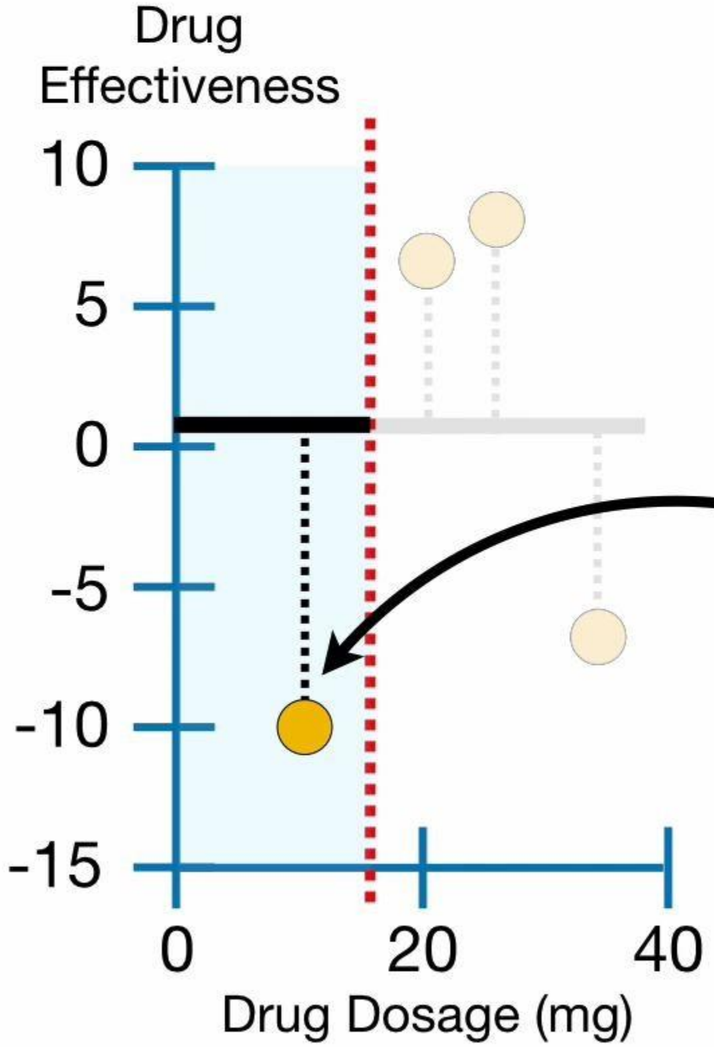
So we split the observations into two groups, based on whether or not the **Dosage < 15**.



Predicted Drug Effectiveness

0.5

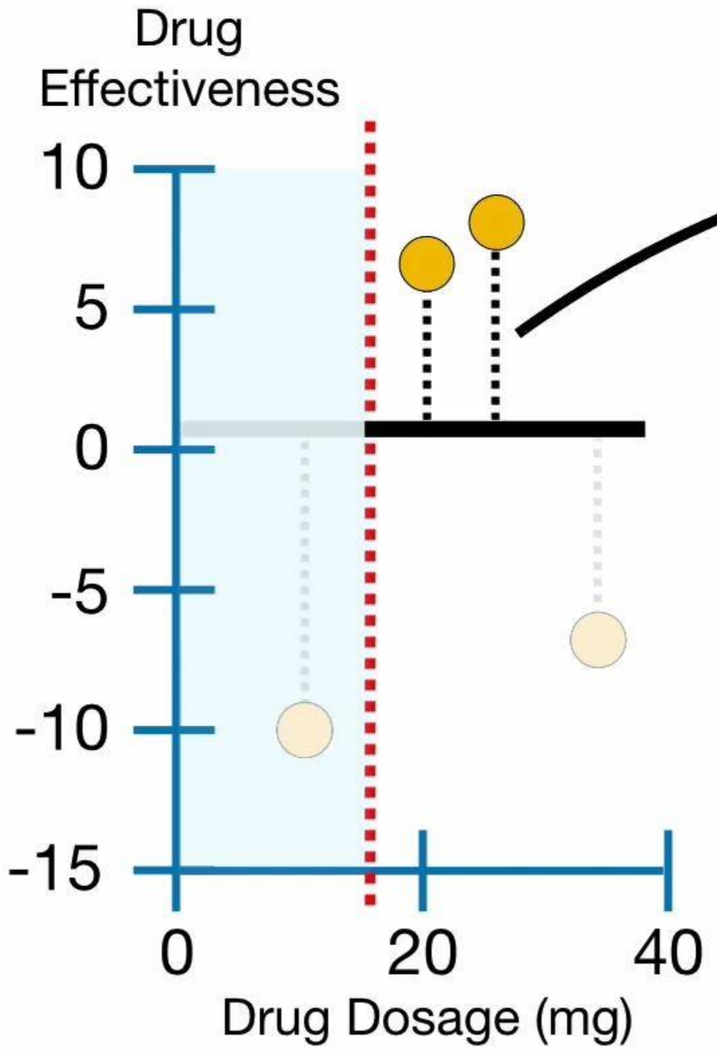


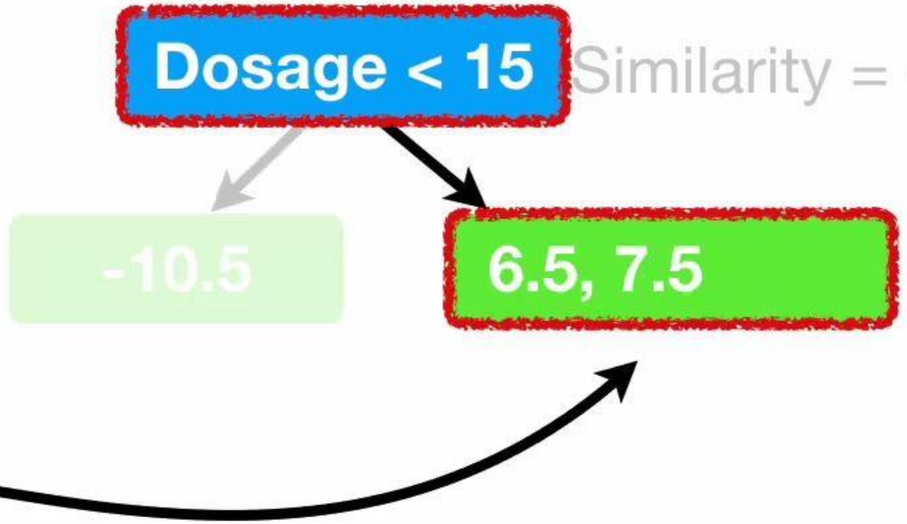


The observation on the far left is the only one with **Dosage < 15**…

Predicted Drug Effectiveness

0.5

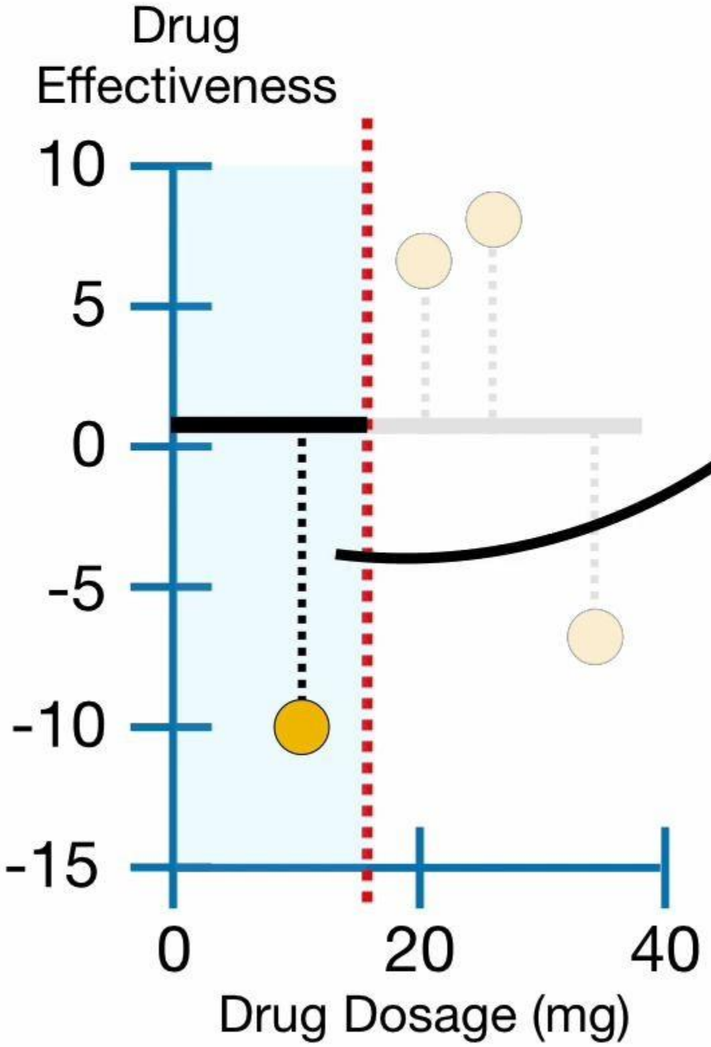




All of the other **Residuals** go to the leaf on the right.

Predicted Drug Effectiveness

0.5



Dosage < 15 Similarity = 4

-10.5

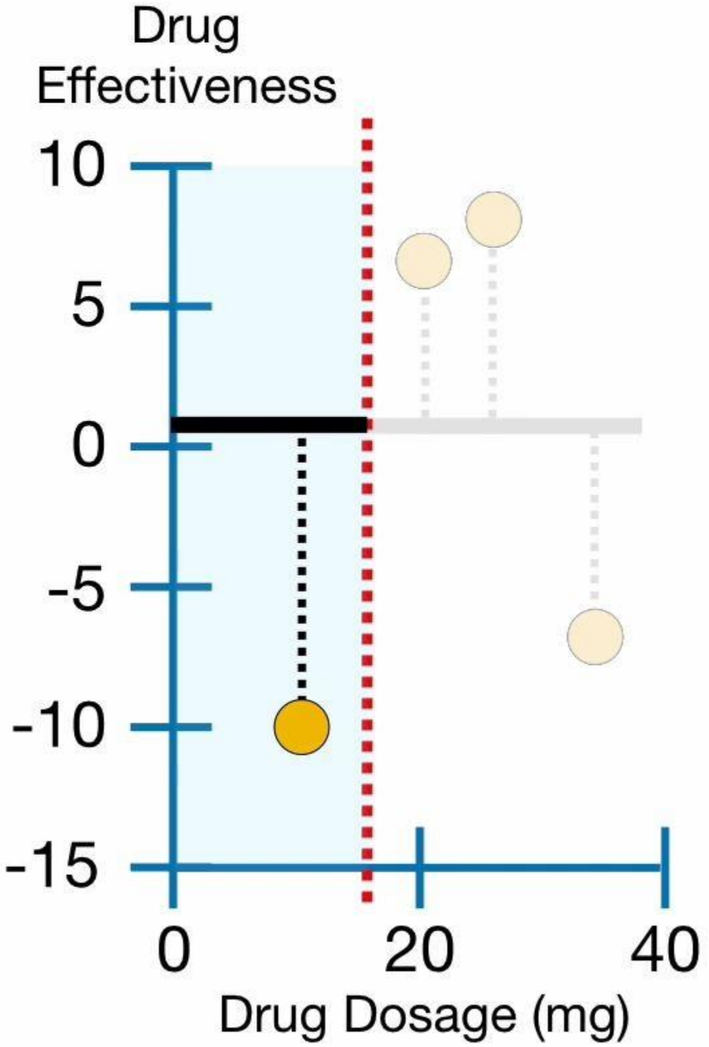
6.5, 7.5, -7.5

$$\text{Similarity Score} = \frac{\text{Sum of Residuals, Squared}}{\text{Number of Residuals} + \lambda}$$

…by plugging the one **Residual** into the numerator…

Predicted Drug Effectiveness

0.5

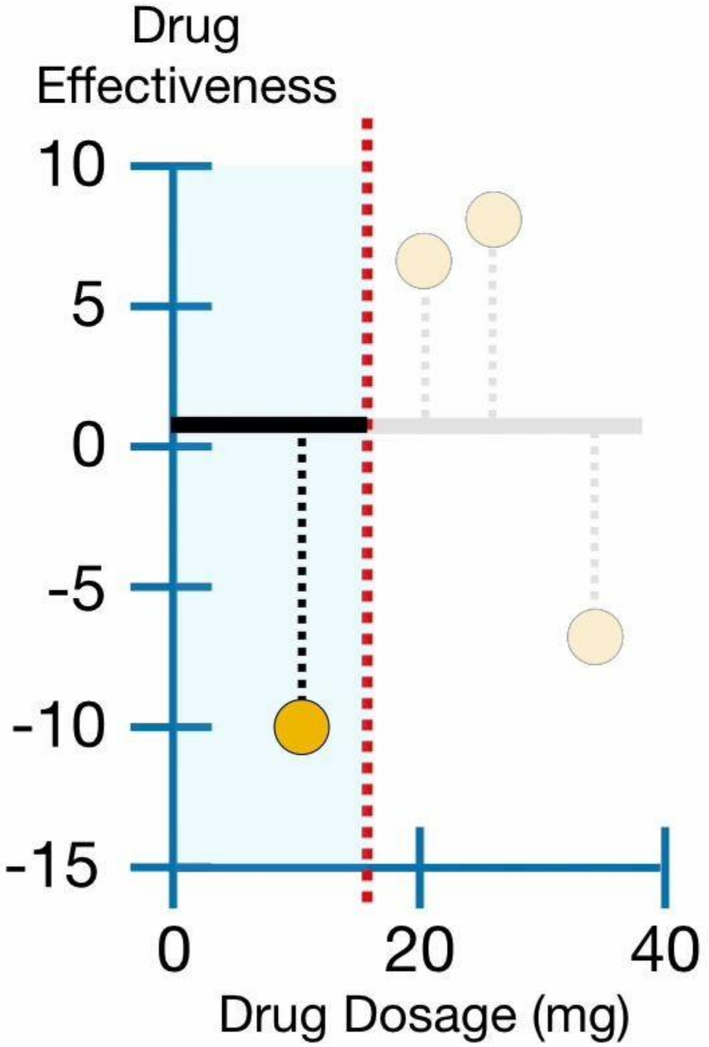


$$\text{Similarity Score} = \frac{-10.5^2}{\text{Number of Residuals} + \lambda}$$

…and since only one **Residual** went to the leaf on the left, the **Number of Residuals = 1**.

Predicted Drug Effectiveness

0.5



Dosage < 15

Similarity = 4

-10.5

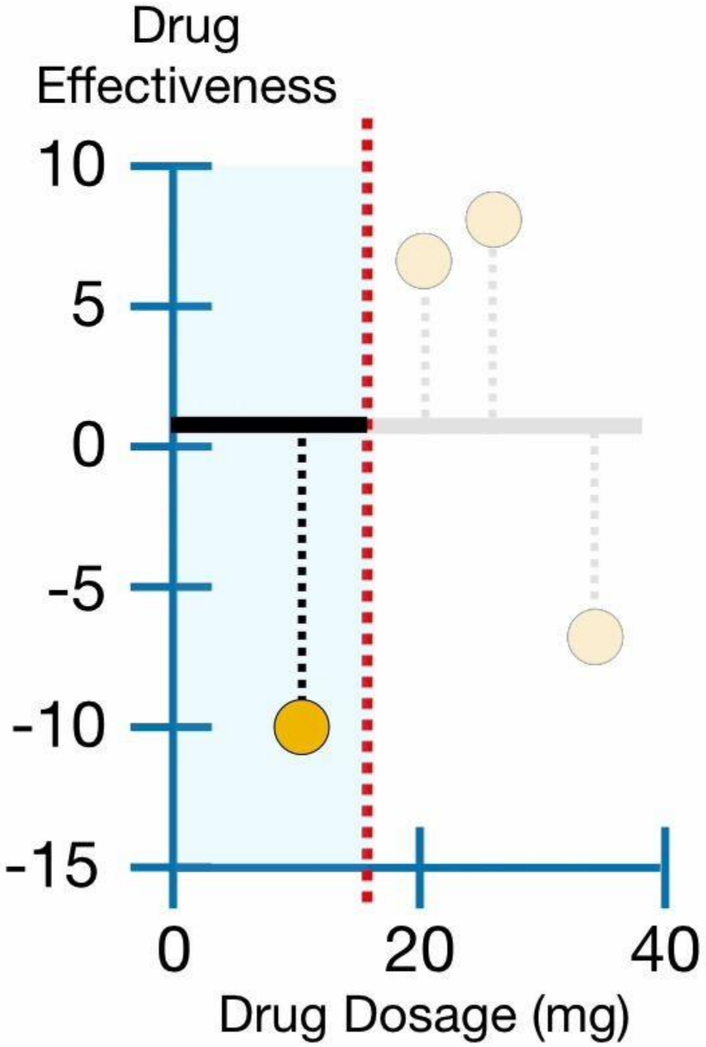
6.5, 7.5, -7.5

$$\text{Similarity Score} = \frac{-10.5^2}{1 + \mathbf{0}}$$

And just like before, we set $\boldsymbol{\lambda = 0}$…

Predicted Drug Effectiveness

0.5

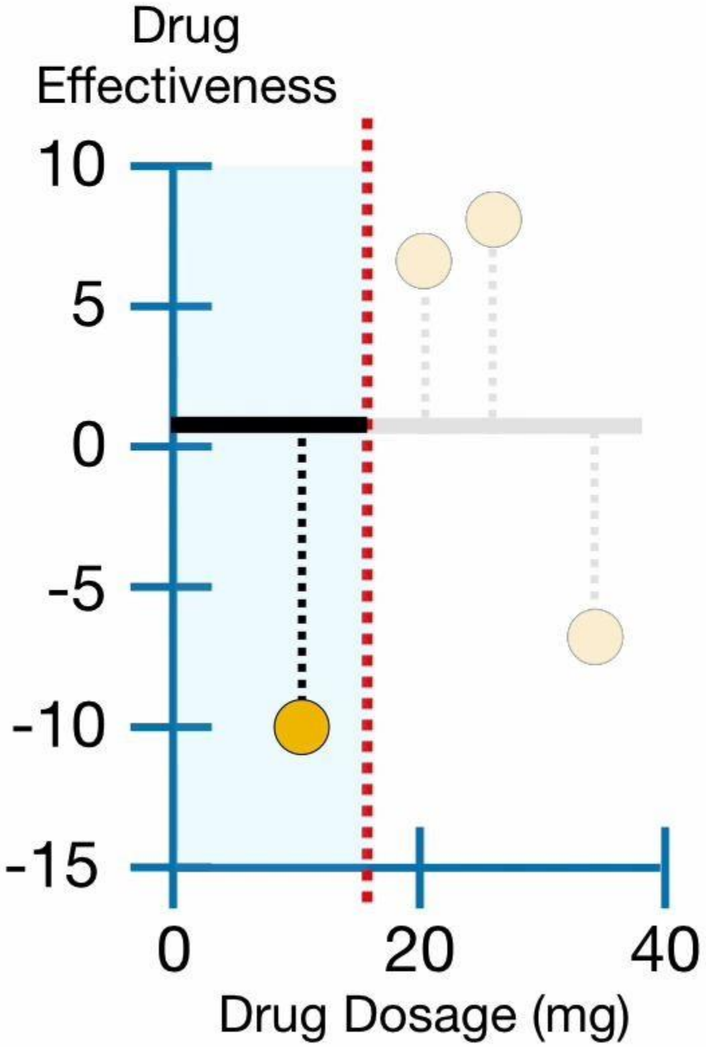


$$\text{Similarity Score} = \frac{-10.5^2}{1 + \mathbf{0}} = 110.25$$

…and the **Similarity Score** for the leaf on the left = **110.25**.

Predicted Drug Effectiveness

0.5



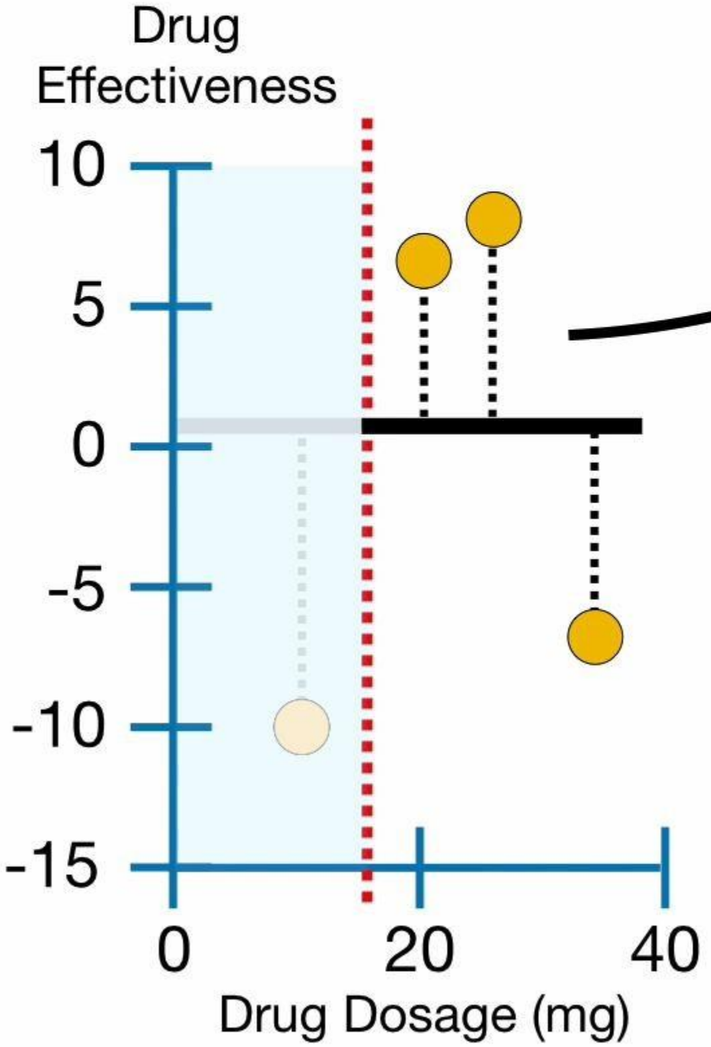
Similarity = 110.25

$$\text{Similarity Score} = \frac{-10.5^2}{1 + 0} = 110.25$$

So let's put **Similarity = 110.25** under the leaf so we can keep track of it…

Predicted Drug Effectiveness

0.5



Dosage < 15

Similarity = 4

-10.5

6.5, 7.5, -7.5

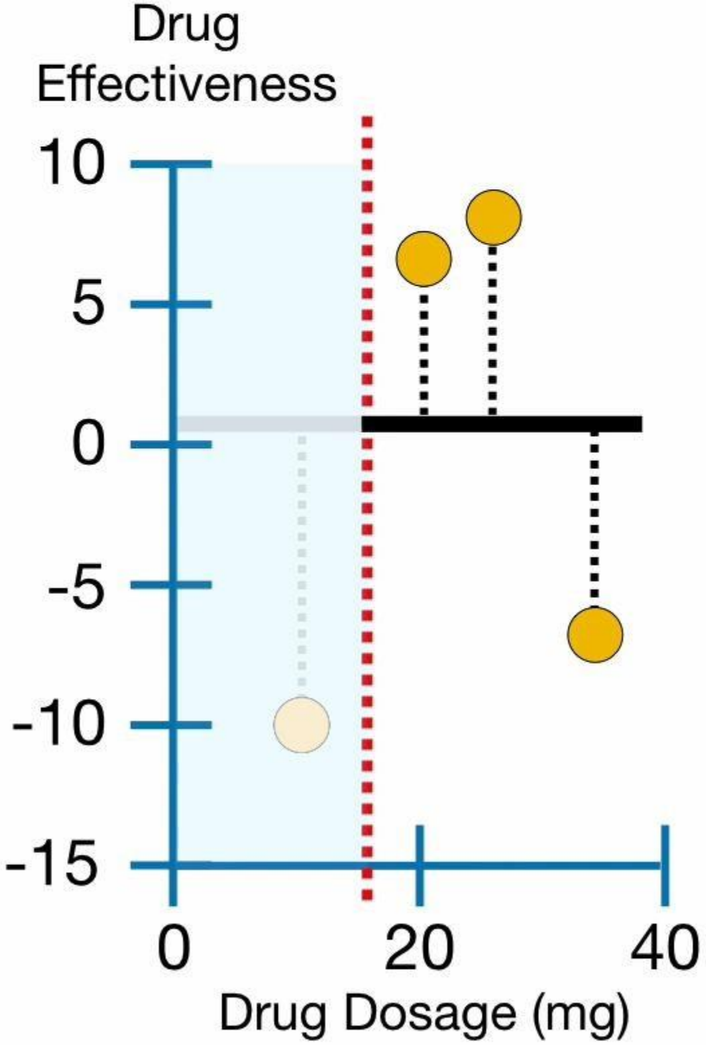
Similarity = 110.25

$$\text{Similarity Score} = \frac{\text{Sum of Residuals, Squared}}{\text{Number of Residuals} + \lambda}$$

We plug the **Sum of Residuals, Squared** into the numerator…

Predicted Drug Effectiveness

0.5

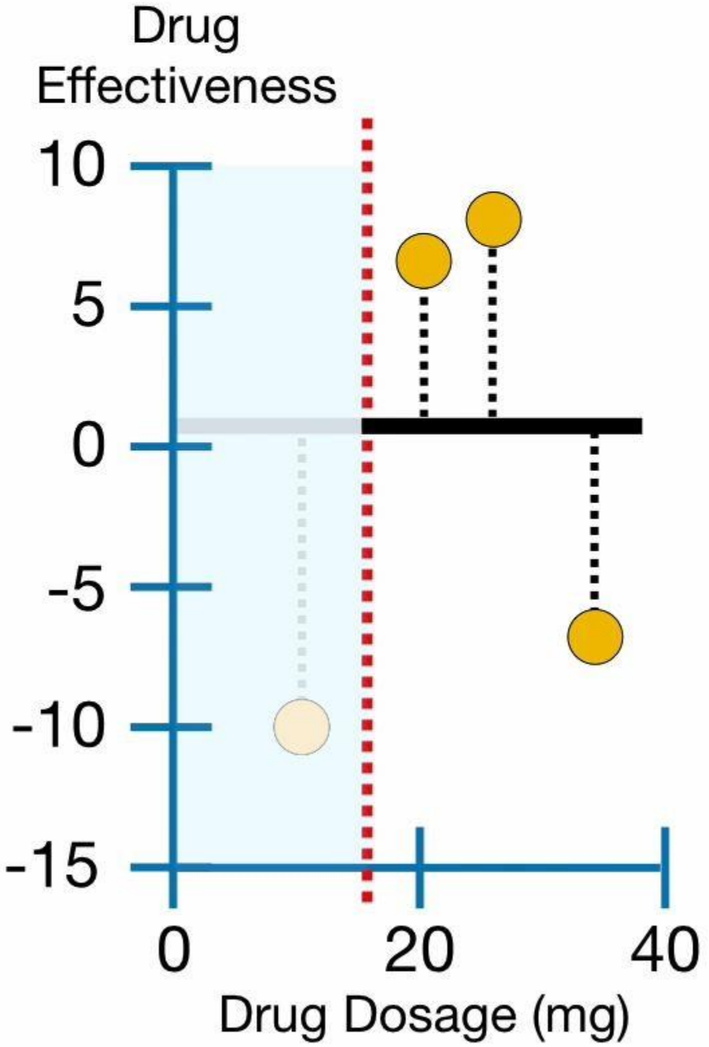


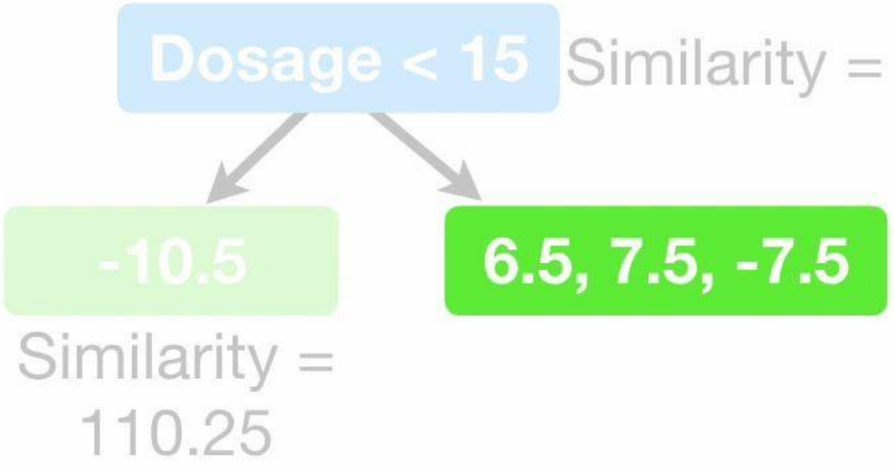
$$\text{Similarity Score} = \frac{(6.5 + 7.5 + -7.5)^2}{\text{Number of Residuals} + \lambda}$$

…and since there are **3 Residuals** in the leaf on the right, we plug **3** into the denominator…

Predicted Drug Effectiveness

0.5



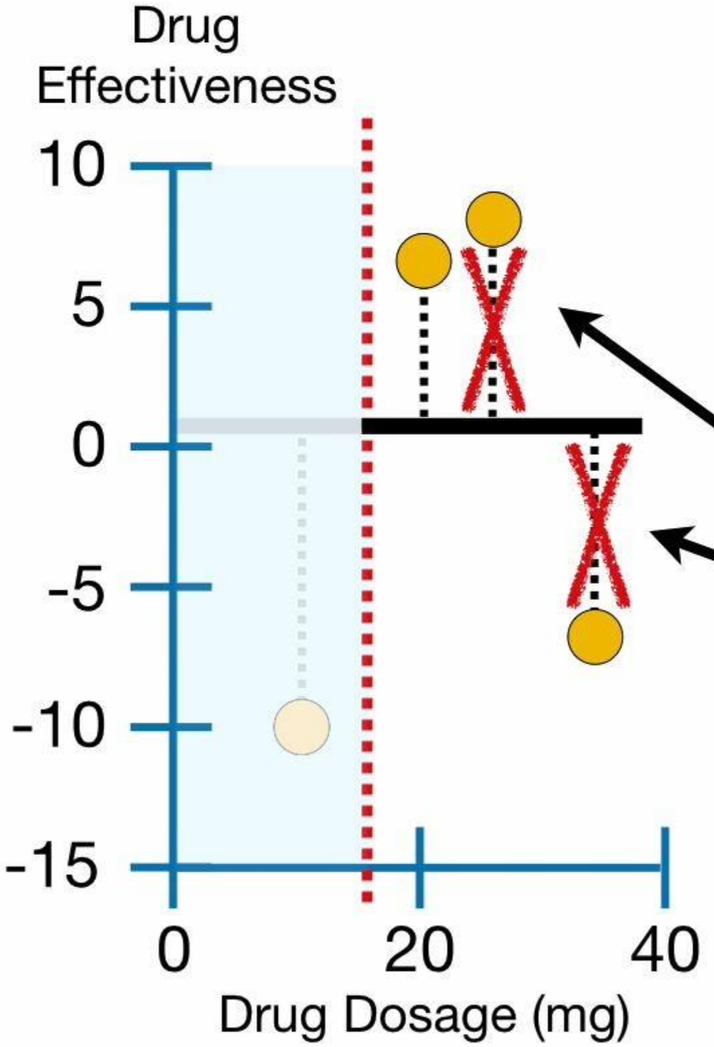


$$\text{Similarity Score} = \frac{(6.5 + 7.5 + -7.5)^2}{3 + \mathbf{0}}$$

…and just like before, let's let $\boldsymbol{\lambda = 0}$.

Predicted Drug Effectiveness

0.5



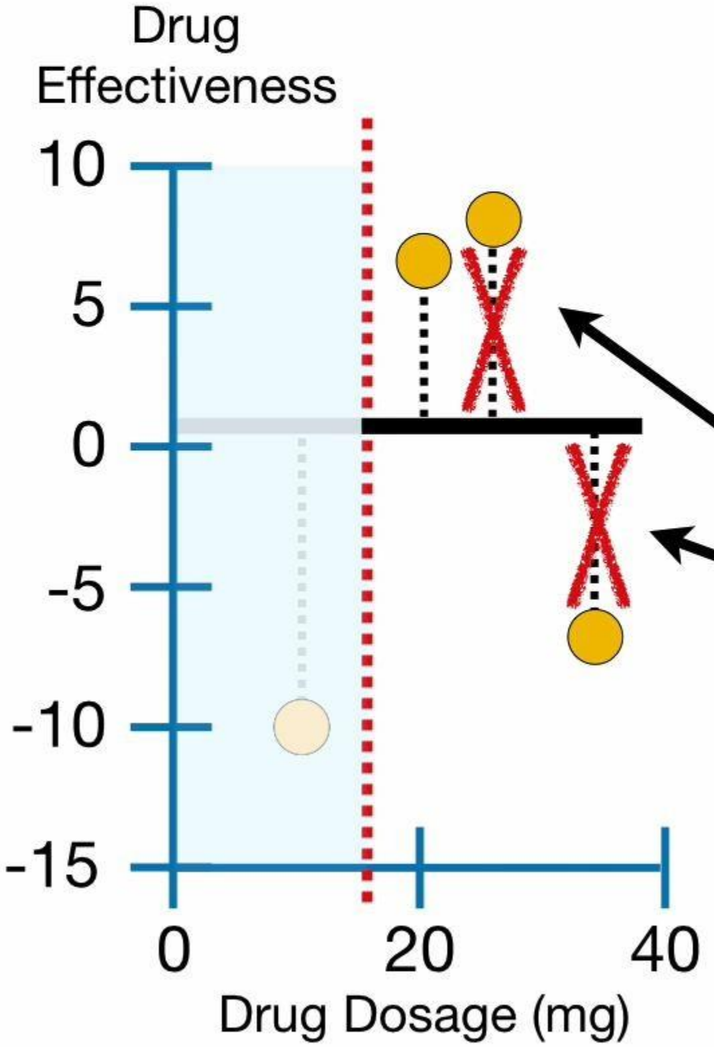
Similarity = 110.25

$$\text{Similarity Score} = \frac{(6.5 + 7.5 + -7.5)^2}{3 + 0}$$

NOTE: Like we saw earlier, because we do not square the **Residuals** *before* we add them together, **7.5** and **-7.5** cancel each other out…

Predicted Drug Effectiveness

0.5



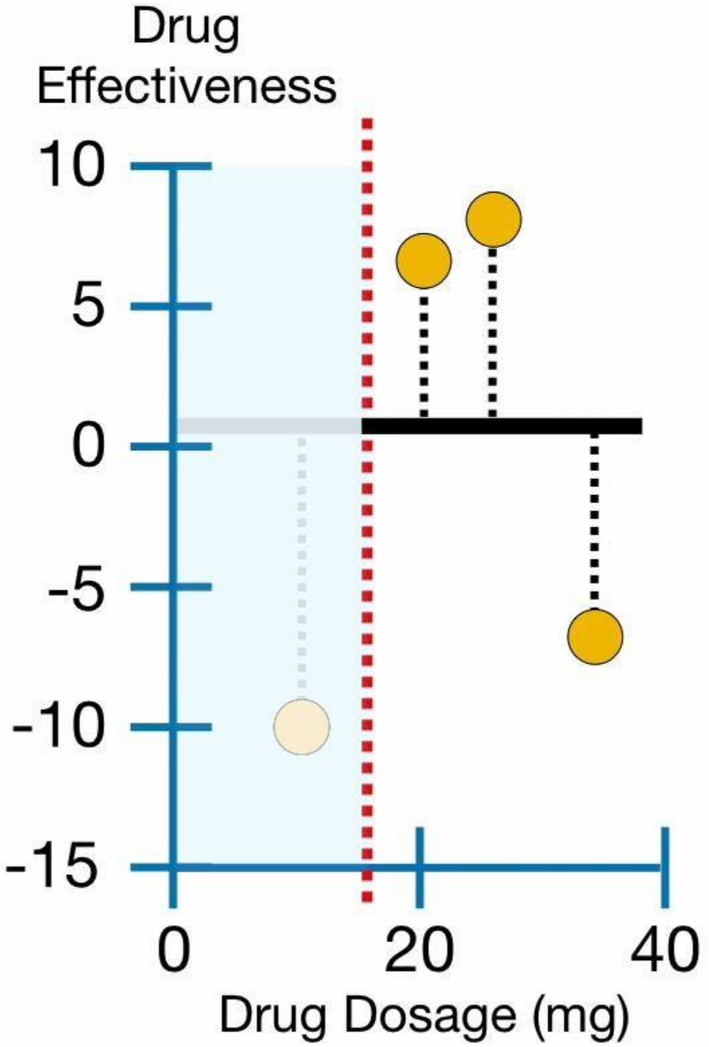
Similarity = 110.25

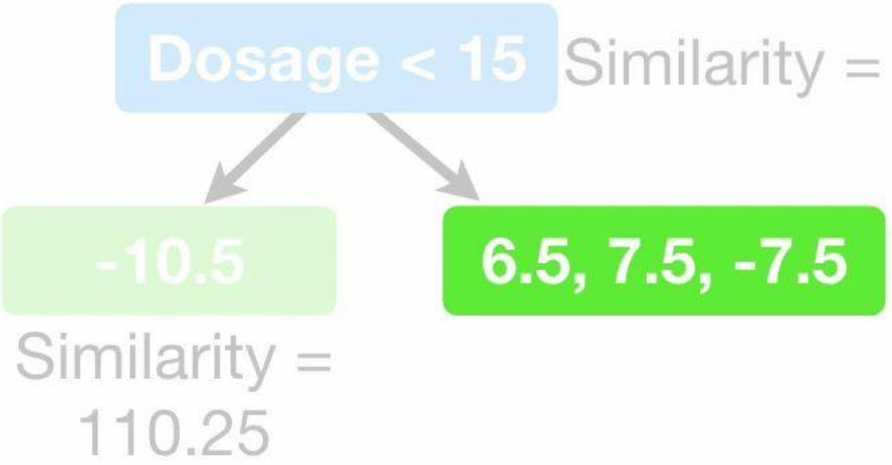
$$\text{Similarity Score} = \frac{(6.5 + 0)^2}{3 + 0}$$

NOTE: Like we saw earlier, because we do not square the **Residuals** *before* we add them together, **7.5** and **-7.5** cancel each other out…

Predicted Drug Effectiveness

0.5



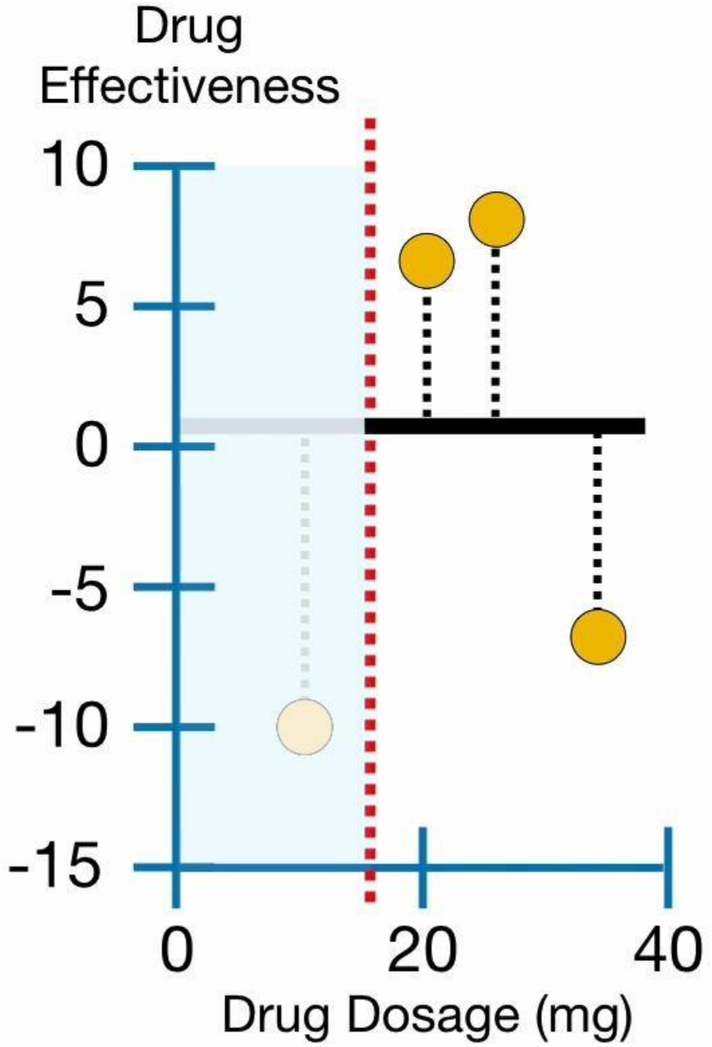


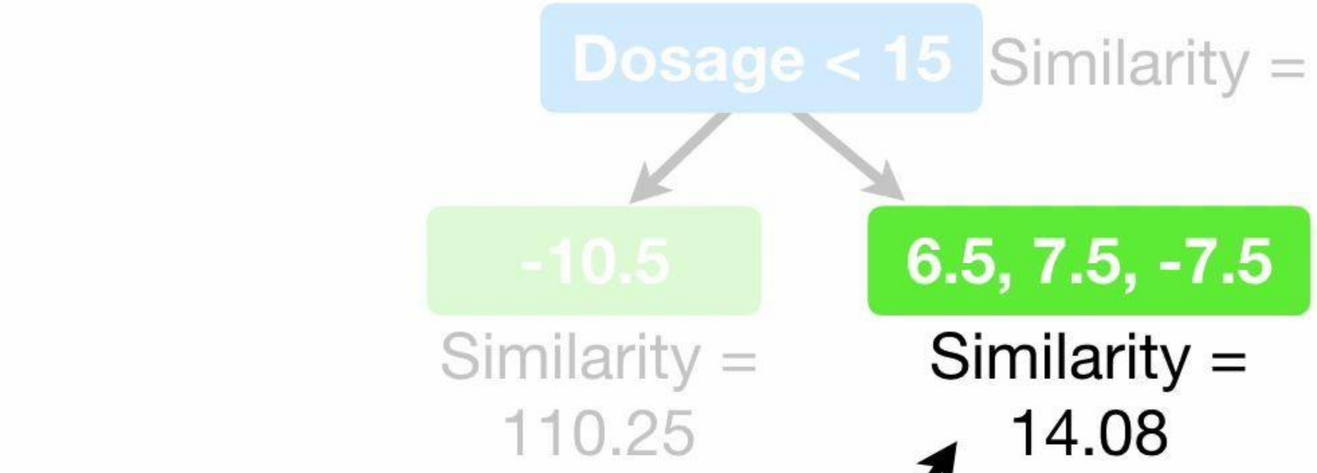
$$\text{Similarity Score} = \frac{6.5^2}{3 + 0}$$

Thus, the **Similarity Score** for the **Residuals** in the leaf on the right = **14.08**.

Predicted Drug Effectiveness

0.5



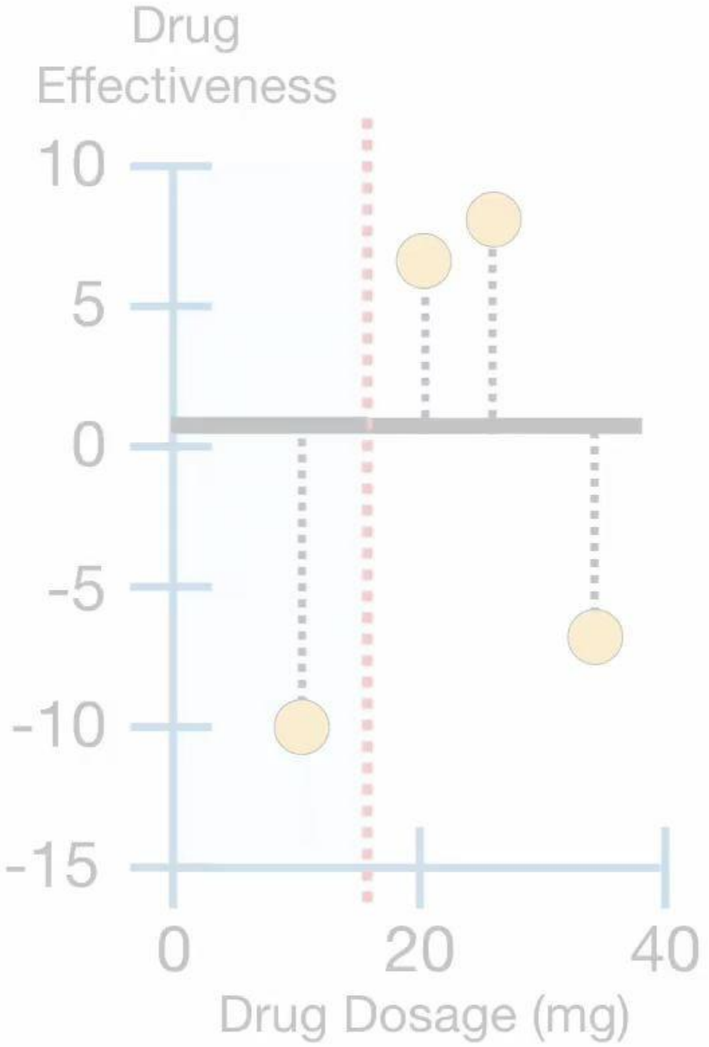


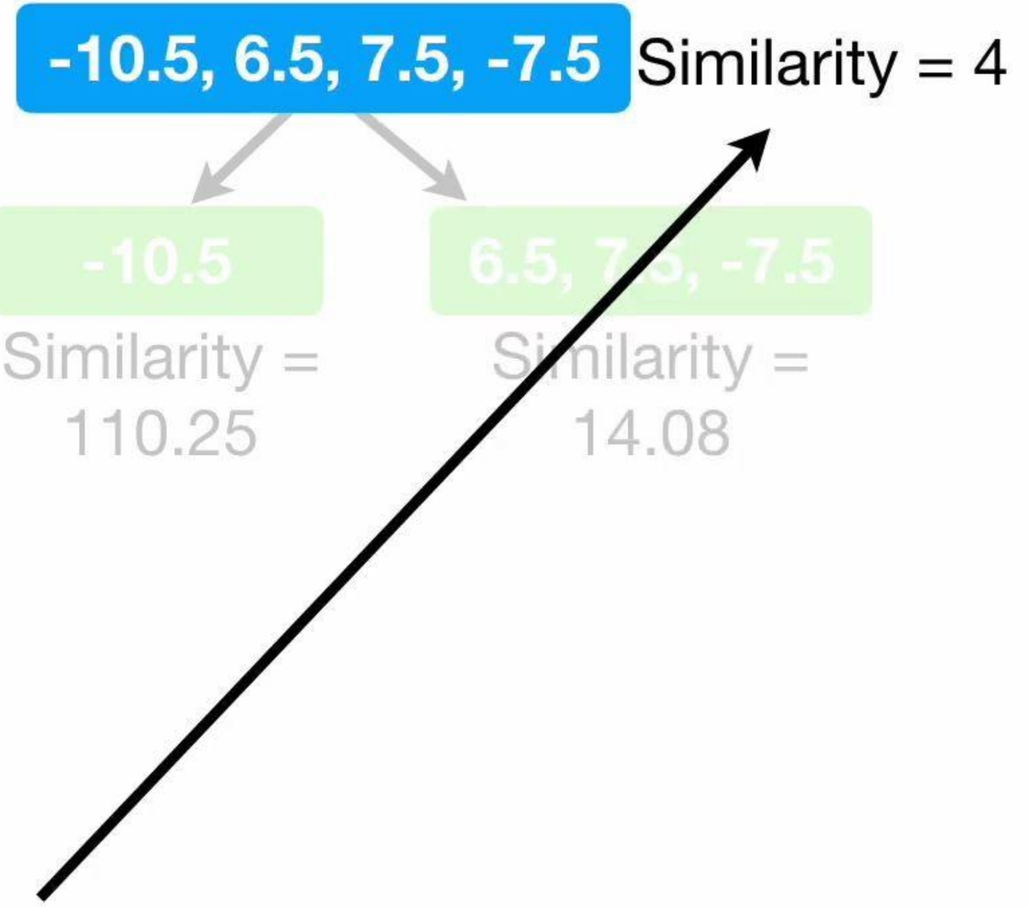
$$\text{Similarity Score} = \frac{6.5^2}{3 + 0} = 14.08$$

So let's put **Similarity = 14.08** under the leaf so we can keep track of it.

Predicted Drug Effectiveness

0.5

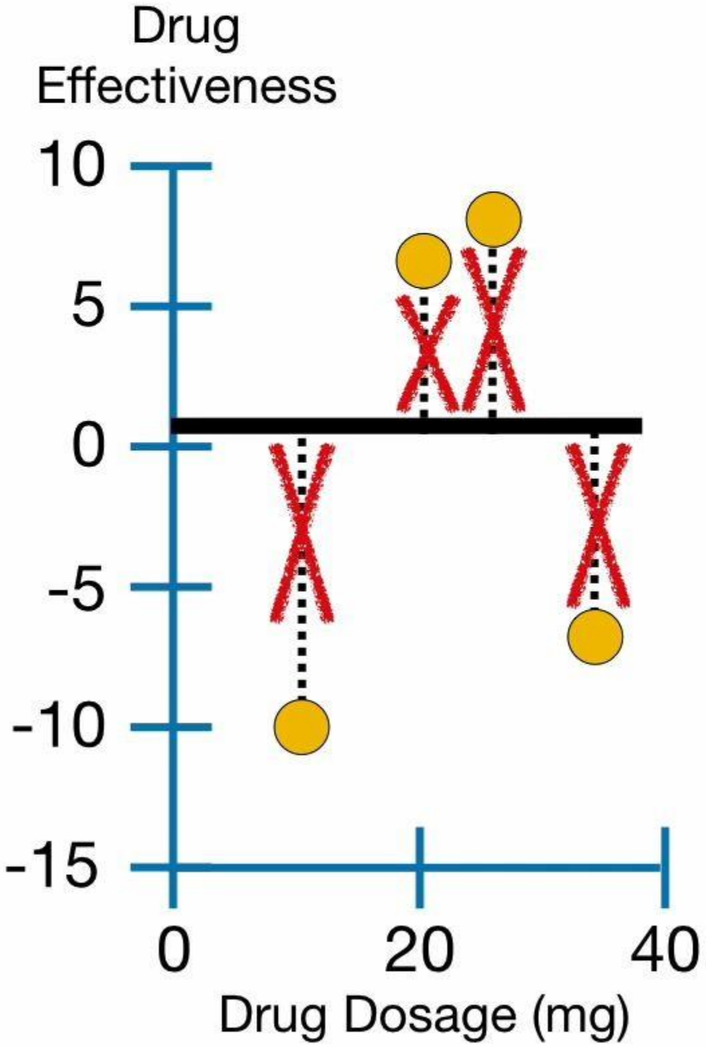


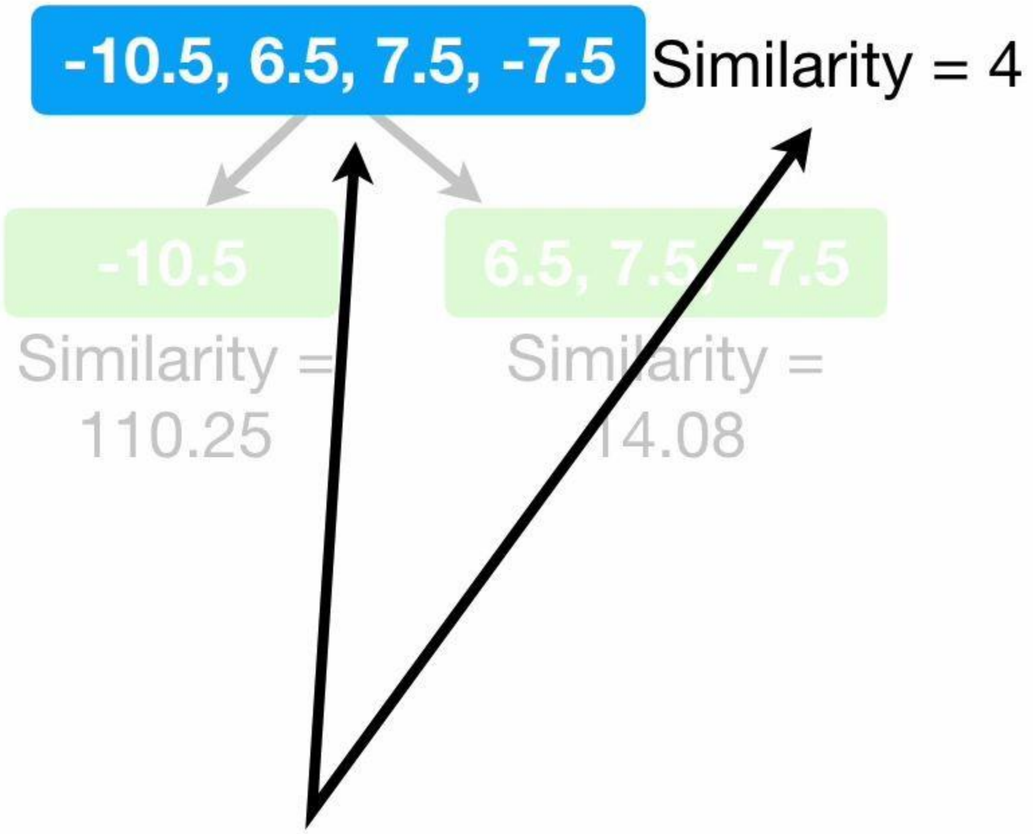


Now that we have calculated the **Similarity Scores** for each node…

Predicted Drug Effectiveness

0.5

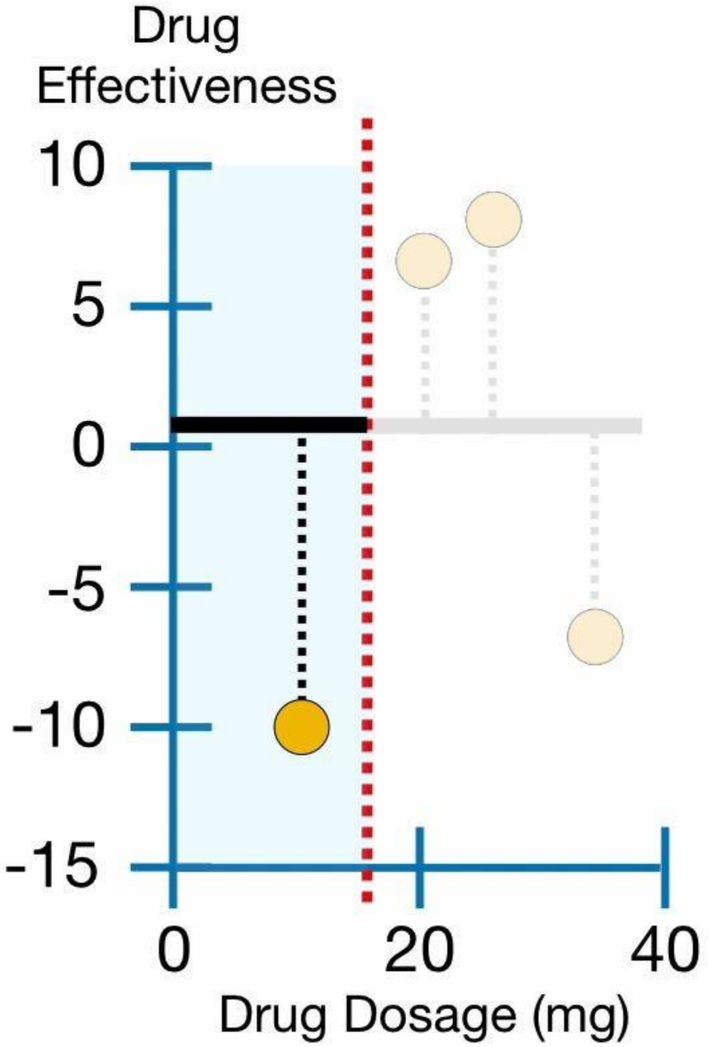




…we see that when the **Residuals** in a node are very different, they cancel each other out and the **Similarity Score** is relatively small.

Predicted Drug Effectiveness

0.5



-10.5, 6.5, 7.5, -7.5 Similarity = 4

-10.5

Similarity = 110.25

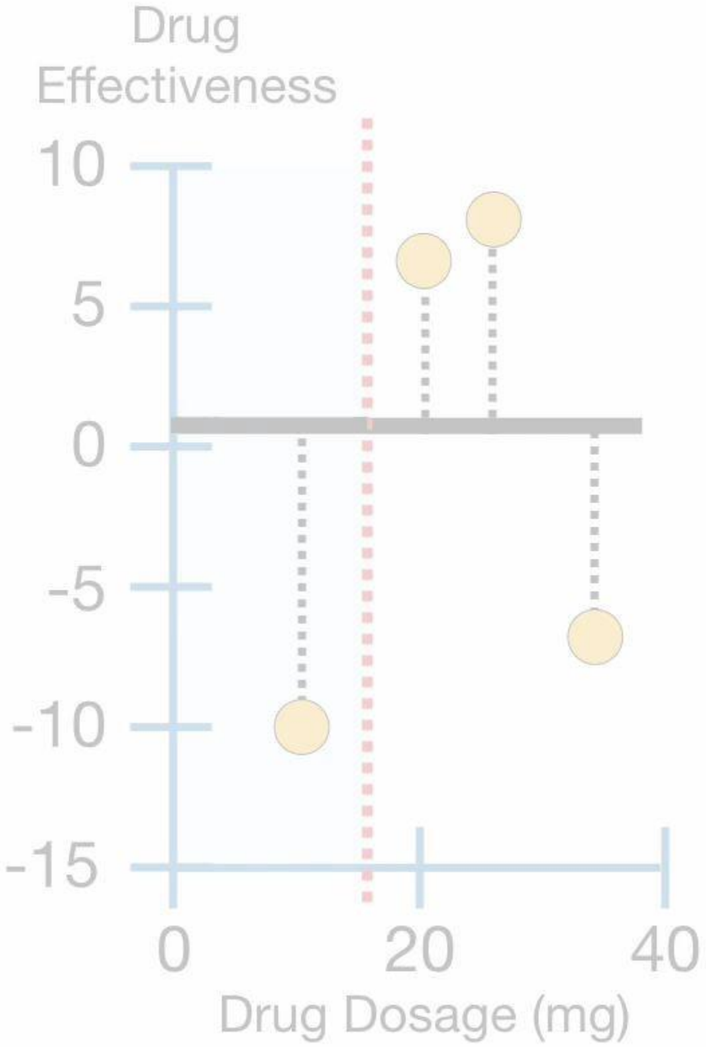
6.5, 7.5, -7.5

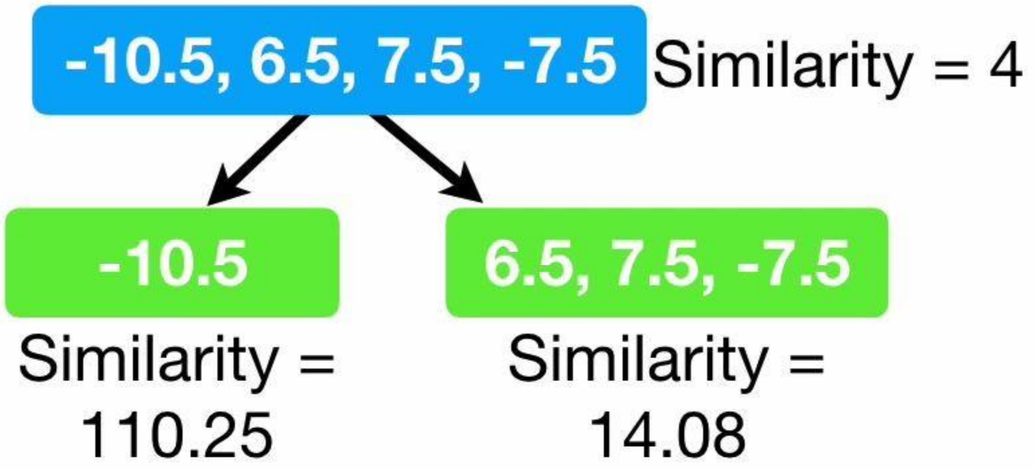
Similarity = 14.08

In contrast, when the **Residuals** are similar, or there is just one of them, they do not cancel out and the **Similarity Score** is relatively large.

Predicted Drug Effectiveness

0.5

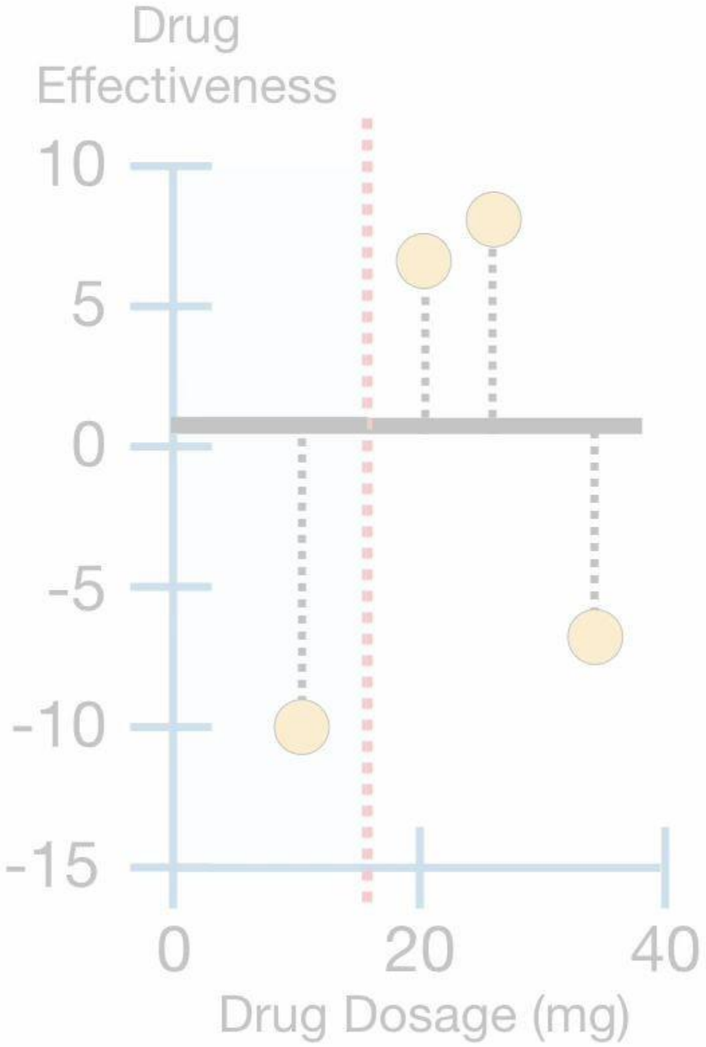


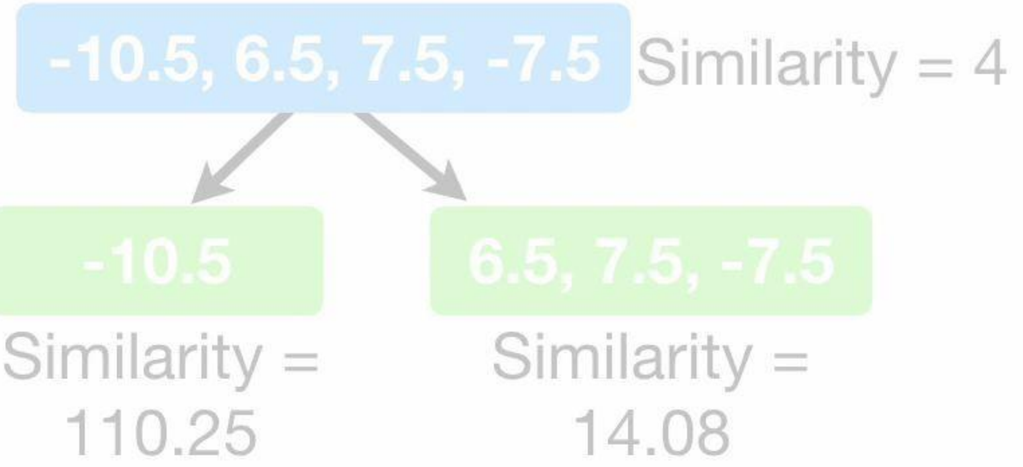


Now we need to quantify how much better the leaves cluster similar **Residuals** than the root.

Predicted Drug Effectiveness

0.5



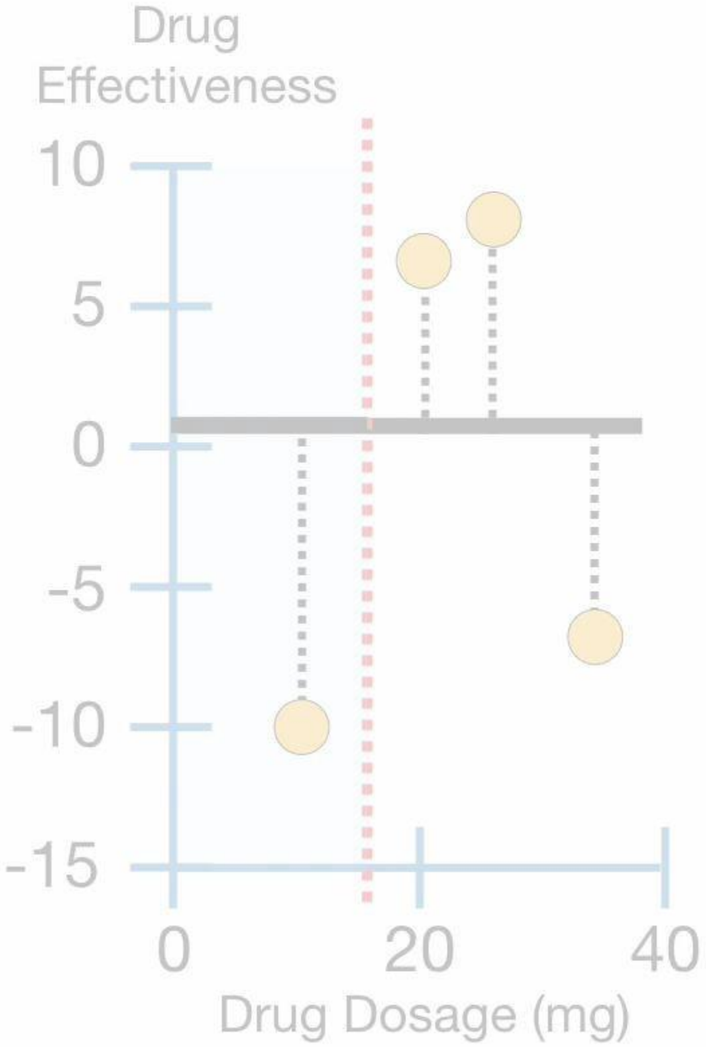


$$\text{Gain} = \text{Left}_{\text{Similarity}} + \text{Right}_{\text{Similarity}} - \text{Root}_{\text{Similarity}}$$

We do this by calculating the **Gain** of splitting the **Residuals** into two groups.

Predicted Drug Effectiveness

0.5



-10.5, 6.5, 7.5, -7.5 Similarity = 4

-10.5

Similarity = 110.25

6.5, 7.5, -7.5

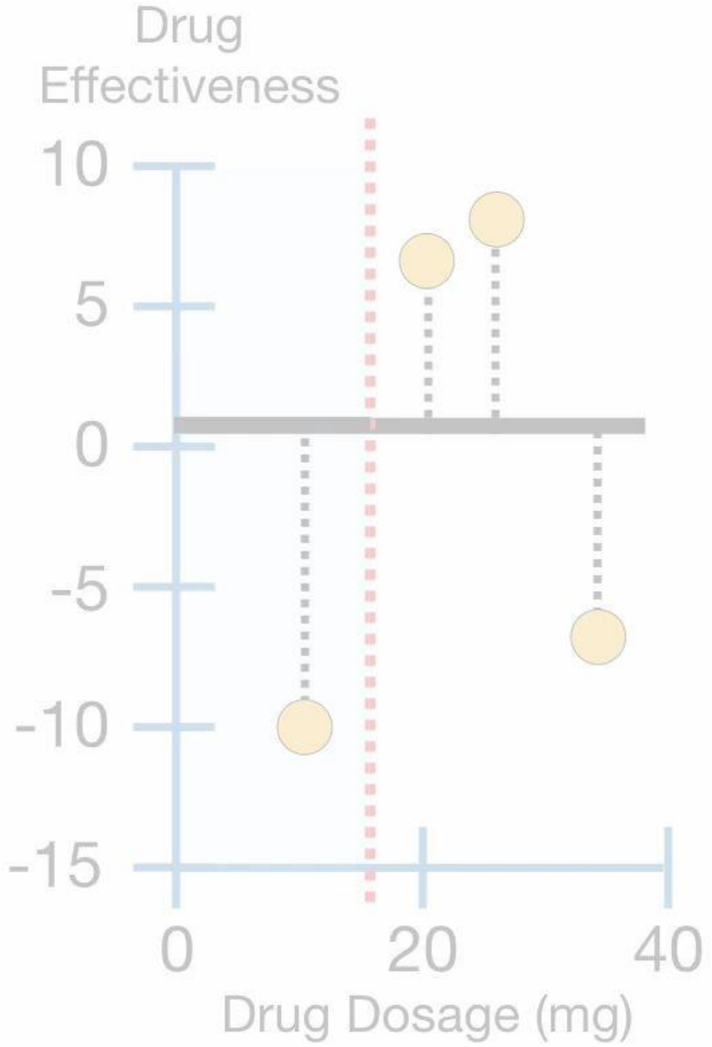
Similarity = 14.08

$$\text{Gain} = \text{Left}_{\text{Similarity}} + \text{Right}_{\text{Similarity}} - \text{Root}_{\text{Similarity}}$$

…plus the **Similarity Score** for the leaf on the right…

Predicted Drug Effectiveness

0.5



-10.5, 6.5, 7.5, -7.5 Similarity = 4

-10.5

Similarity = 110.25

6.5, 7.5, -7.5

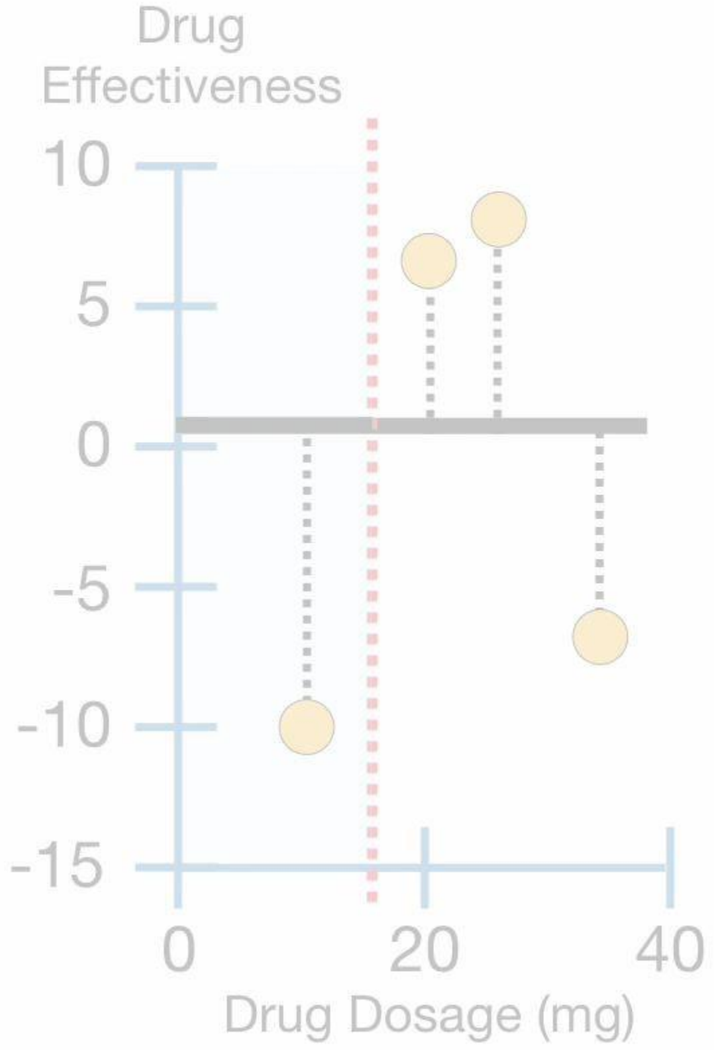
Similarity = 14.08

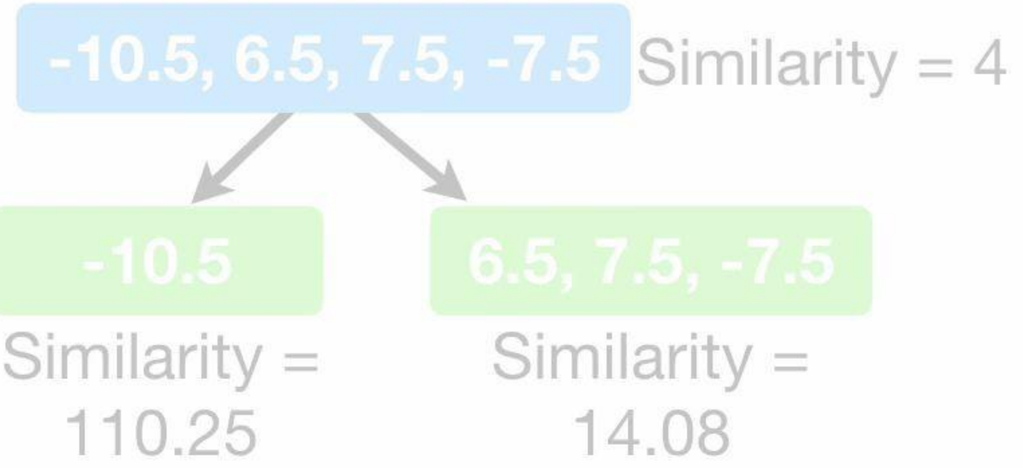
$$Gain = Left_{Similarity} + Right_{Similarity} - Root_{Similarity}$$

Plugging in the numbers…

Predicted Drug Effectiveness

0.5



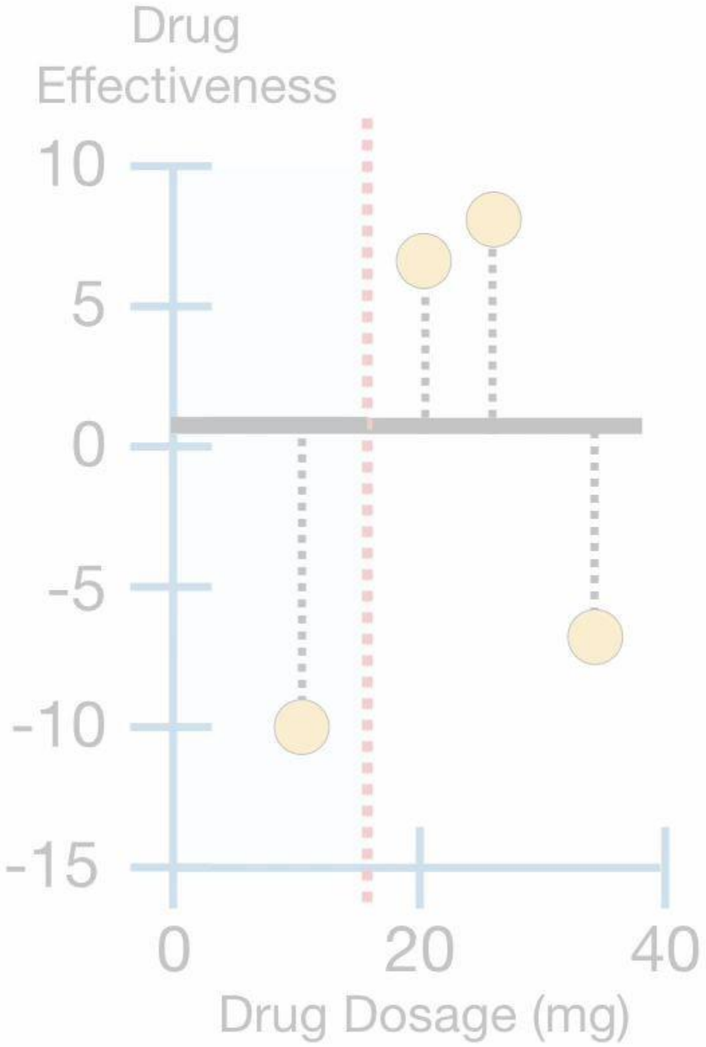


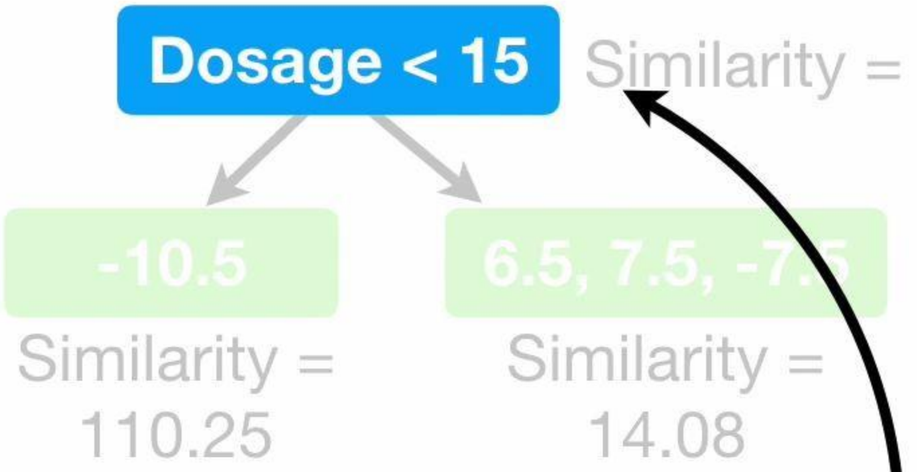
Gain = 110.25 + 14.08 - 4 = 120.33

…gives us **120.33**.

Predicted Drug Effectiveness

0.5



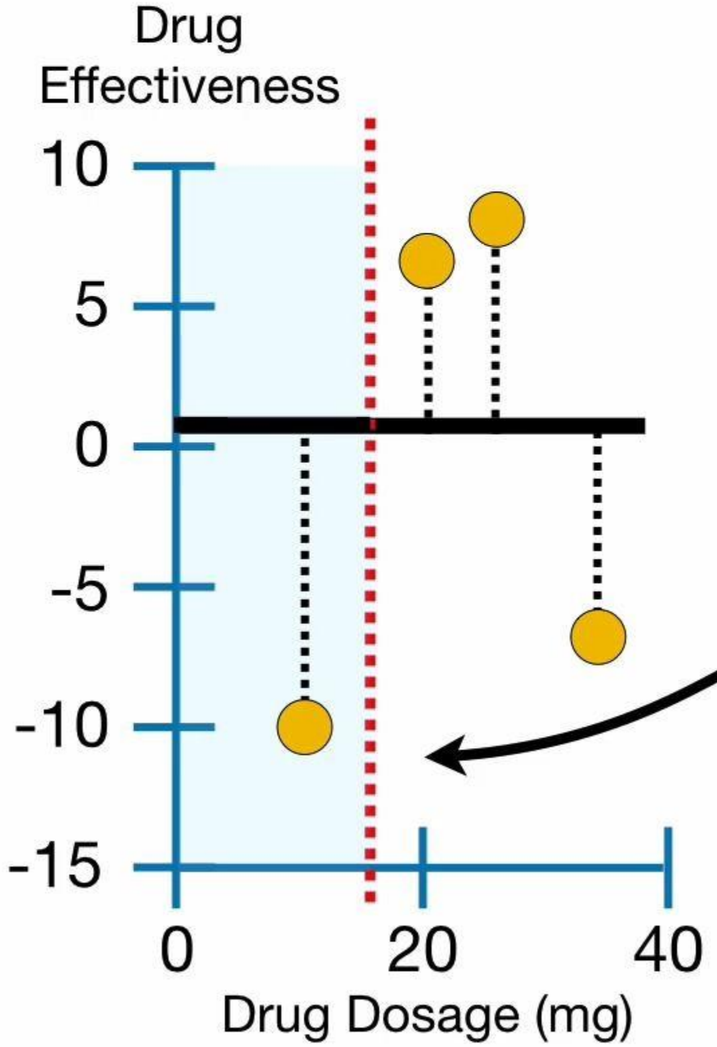


Gain = 110.25 + 14.08 - 4 = 120.33

Now that we have calculated the **Gain** for the threshold **Dosage < 15**, we can compare it to the **Gain** calculated for other thresholds.

Predicted Drug Effectiveness

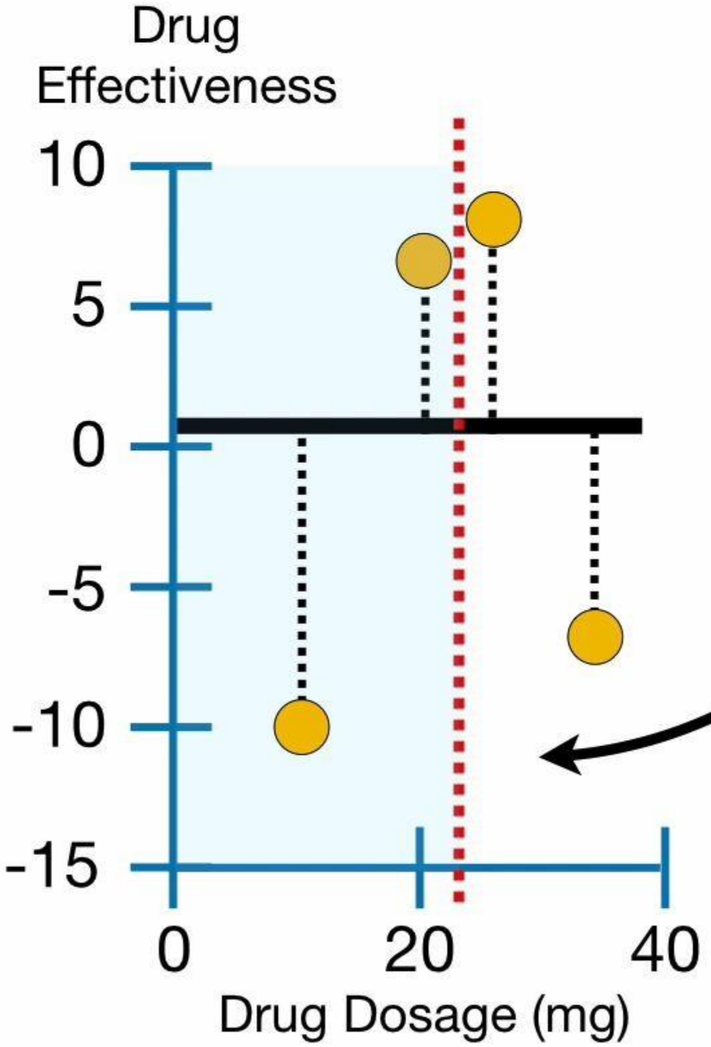
0.5



So we shift the threshold over so that it is the average of the next two observations…

Predicted Drug Effectiveness

0.5



…and build a simple tree that divides the observations using the new threshold, **Dosage < 22.5**.

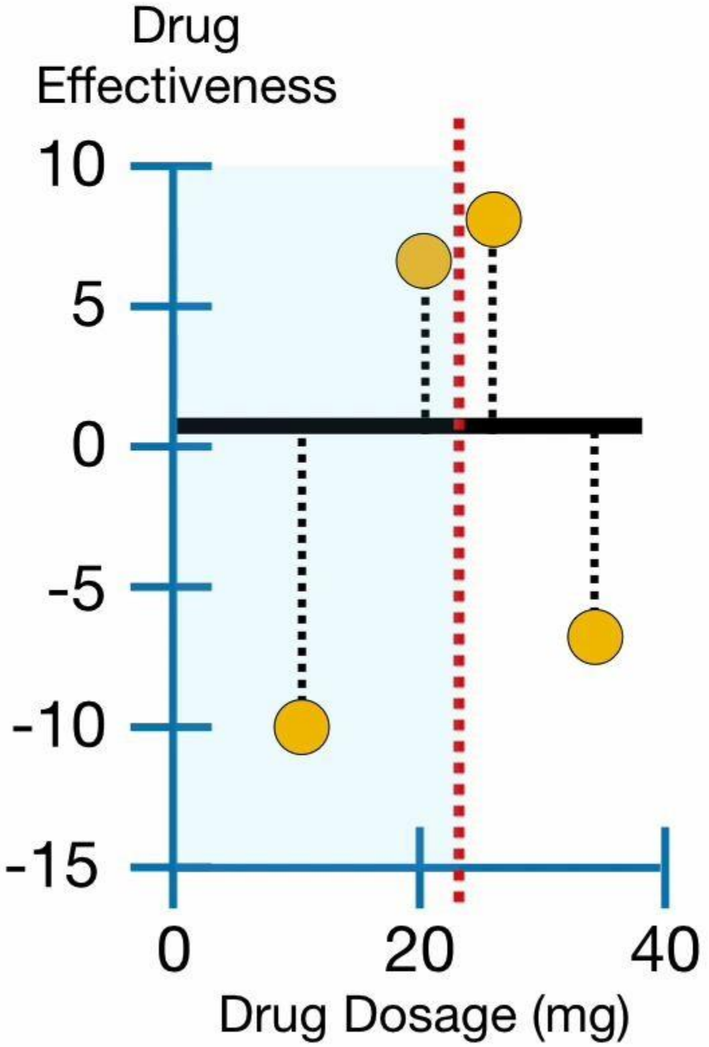
Dosage < 22.5 Similarity = 4

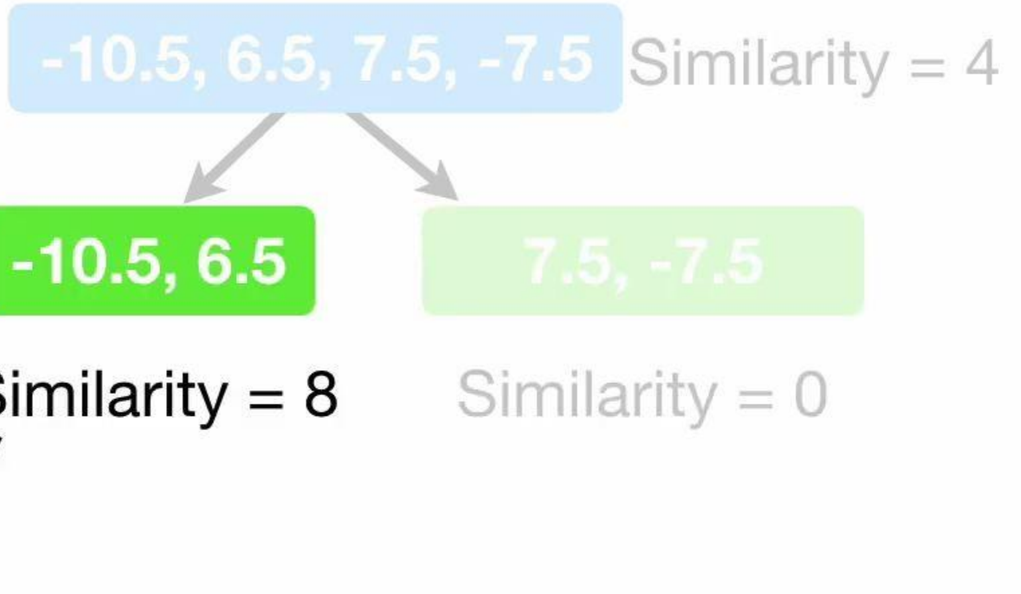
-10.5, 6.5

7.5, -7.5

Predicted Drug Effectiveness

0.5



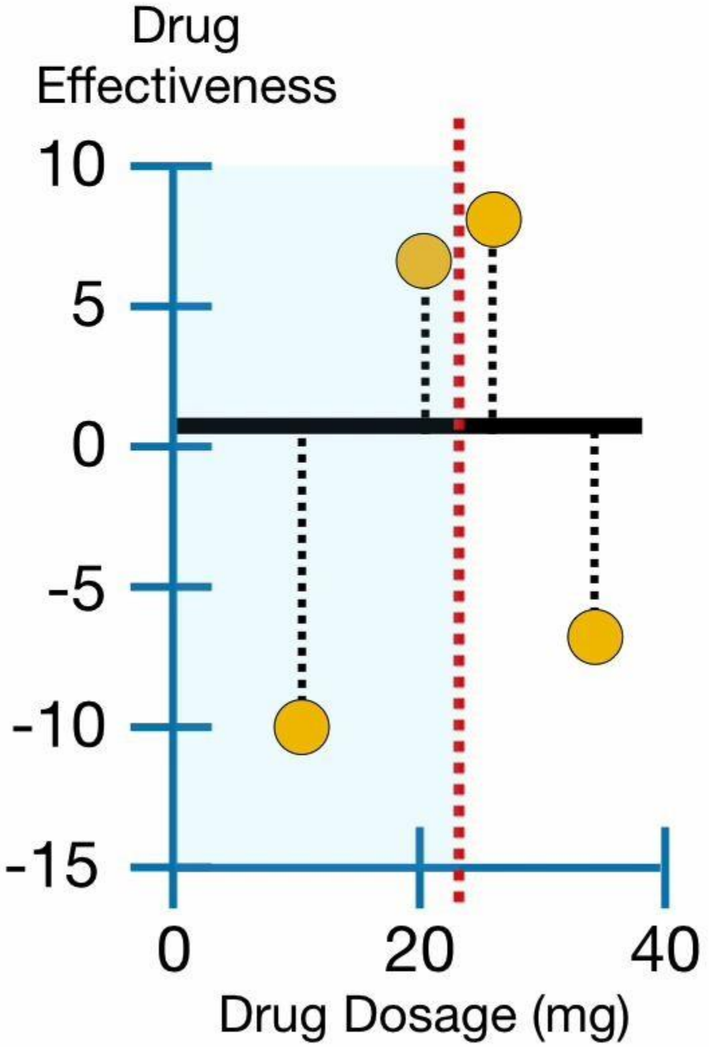


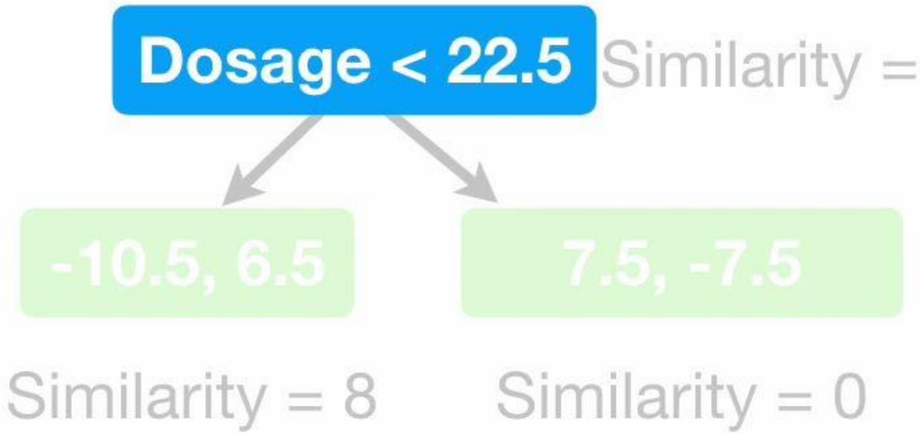
$$\text{Gain} = 8 + \text{Right}_{\text{Similarity}} - \text{Root}_{\text{Similarity}}$$

…and calculate the **Gain**.

Predicted Drug Effectiveness

0.5



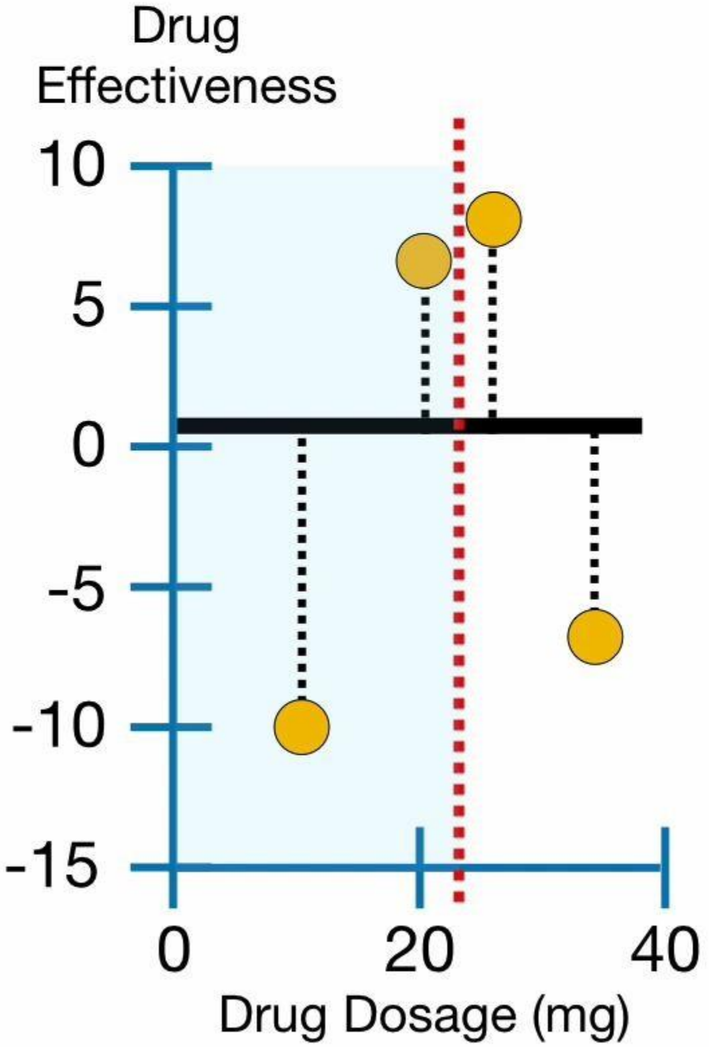


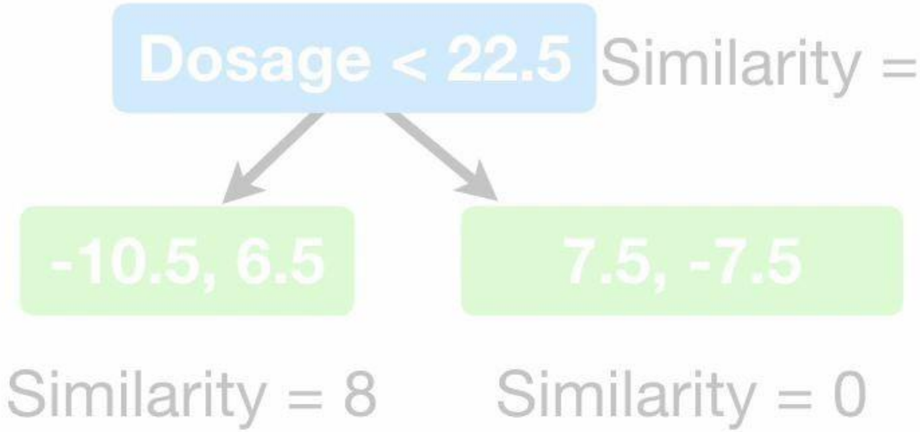
Gain = 8 + 0 - 4 = 4

The **Gain** for **Dosage < 22.5** is **4**.

Predicted Drug Effectiveness

0.5



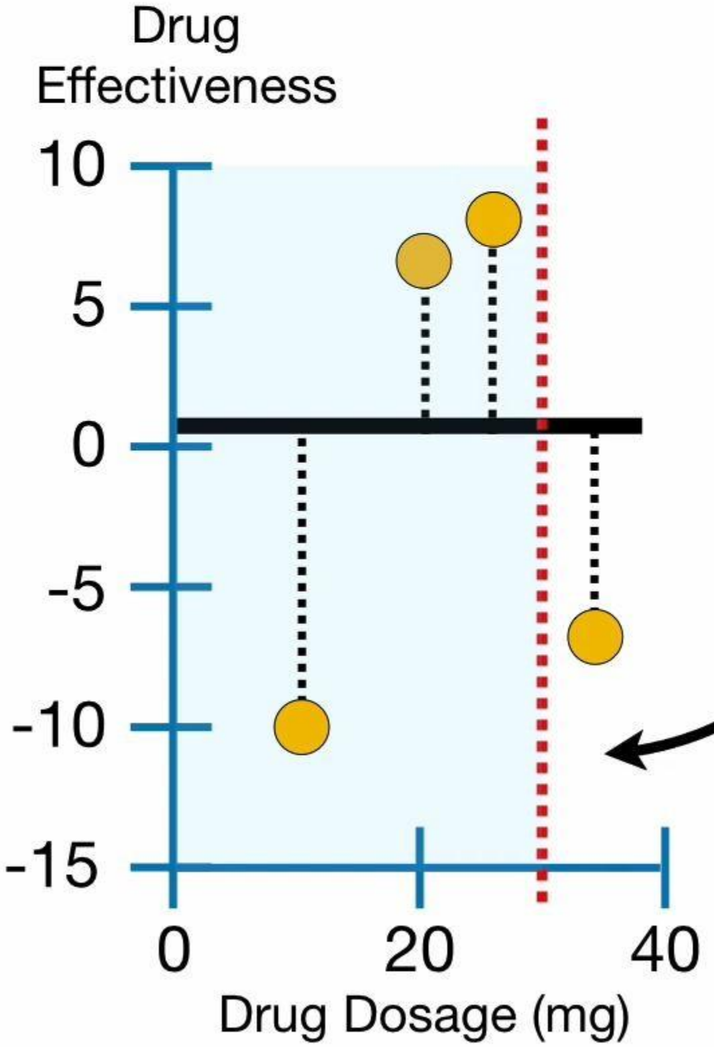


Gain = 8 + 0 - 4 = 4

Since the **Gain** for **Dosage < 22.5** (**Gain = 4**) is less than the **Gain** for **Dosage < 15** (**Gain = 120.33**), **Dosage < 15** is better at splitting the **Residuals** into clusters of similar values.

Predicted Drug Effectiveness

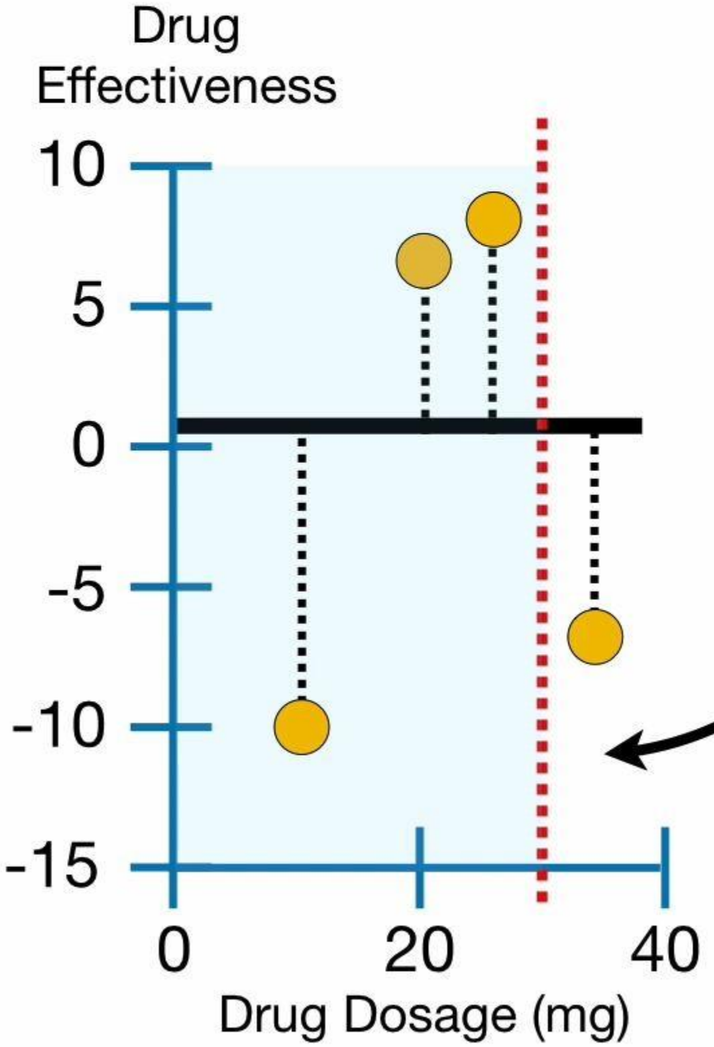
0.5



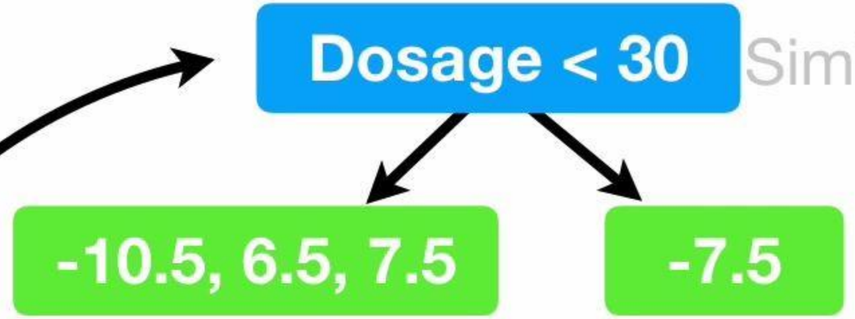
Now we shift the threshold over so that it is the average of the last two observations…

Predicted Drug Effectiveness

0.5

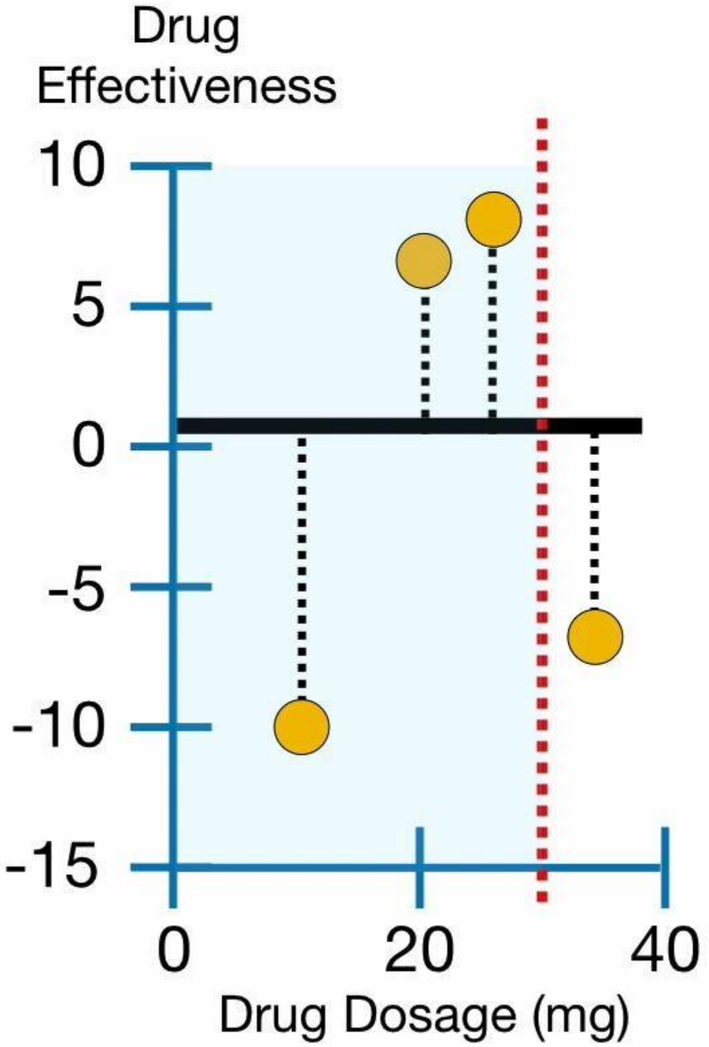


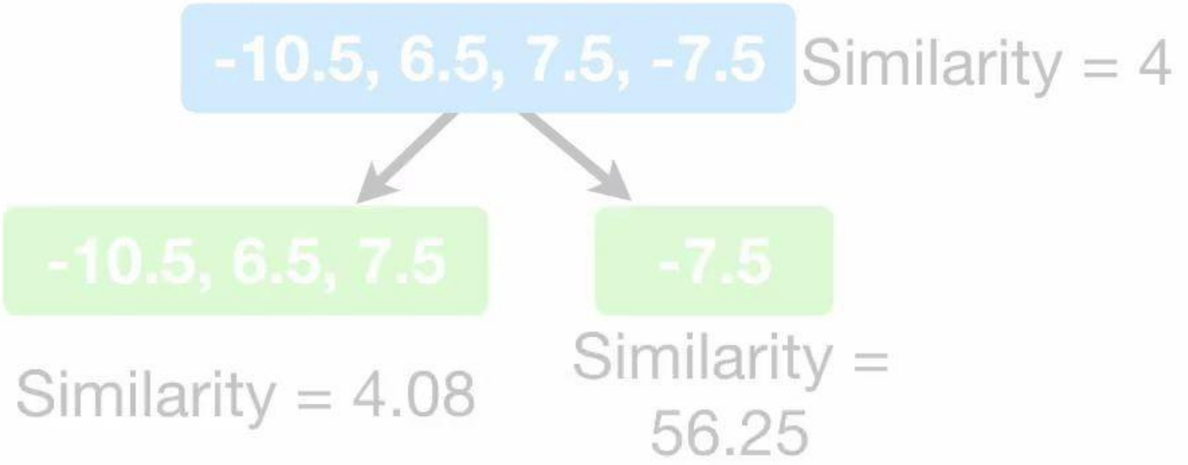
…and build a simple tree that divides that observations using the new threshold, **Dosage < 30**.



Predicted Drug Effectiveness

0.5



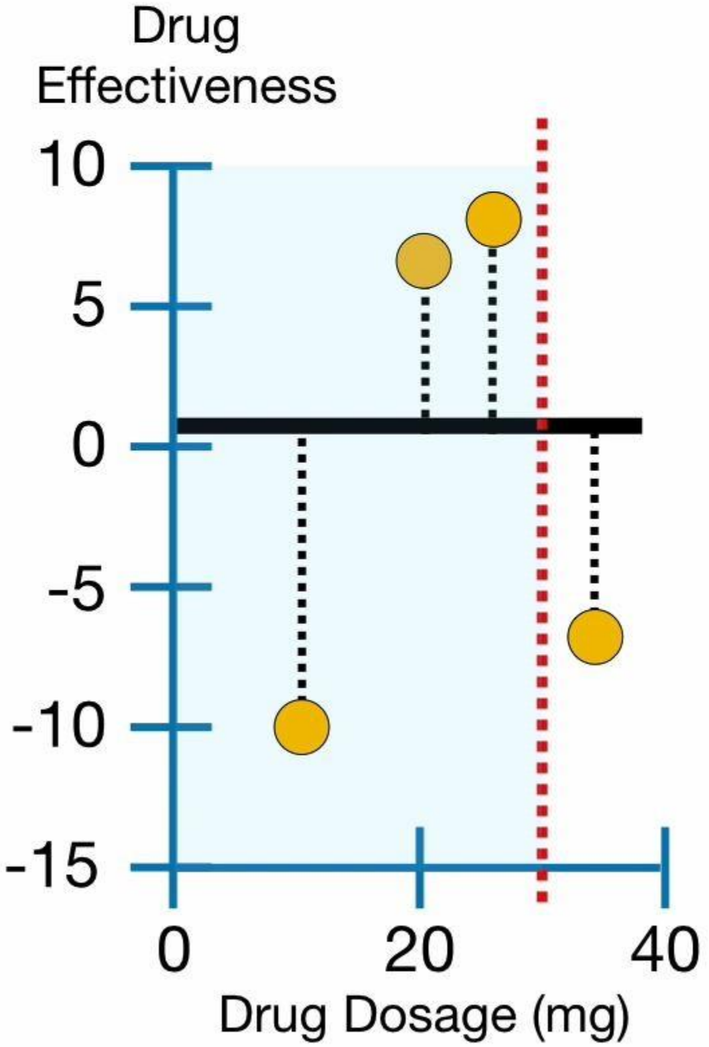


$$Gain = Left_{Similarity} + Right_{Similarity} - Root_{Similarity}$$

…and the **Gain**.

Predicted Drug Effectiveness

0.5



Dosage < 30 Similarity = 4

-10.5, 6.5, 7.5

Similarity = 4.08

-7.5

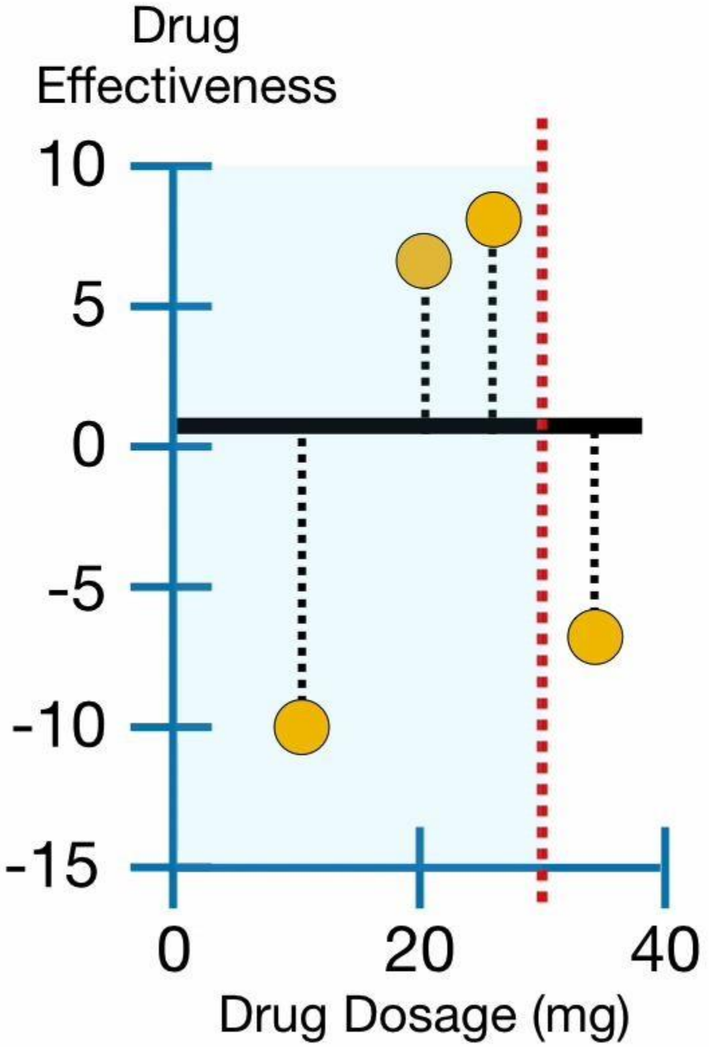
Similarity = 56.25

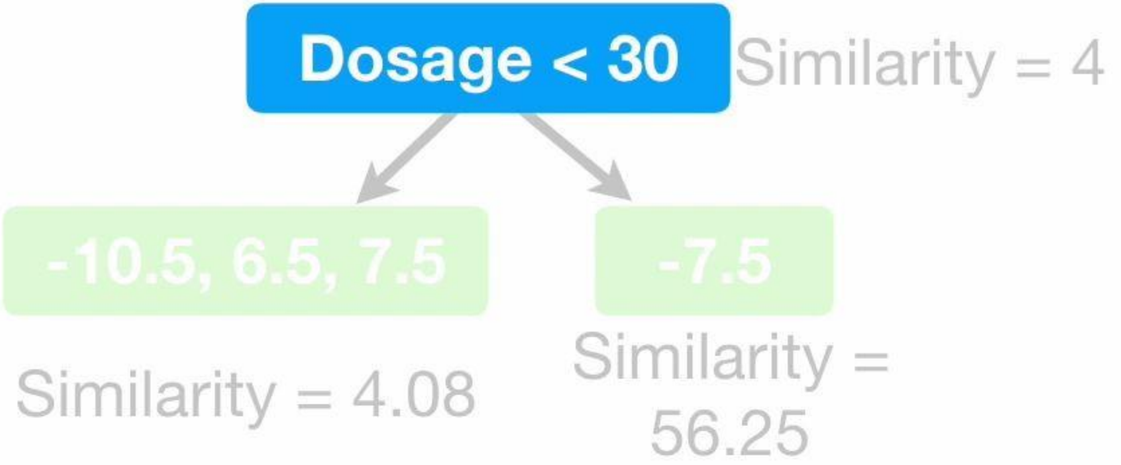
Gain = 4.08 + 56.25 - 4 = 56.33

The **Gain** for **Dosage < 30** = **56.33**

Predicted Drug Effectiveness

0.5



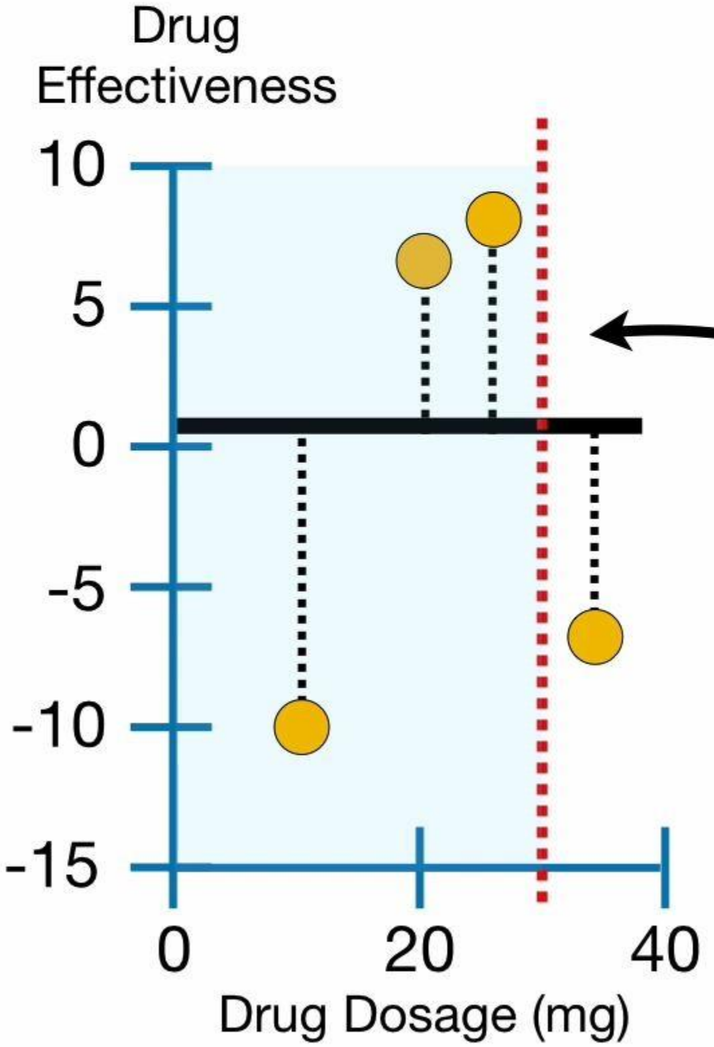


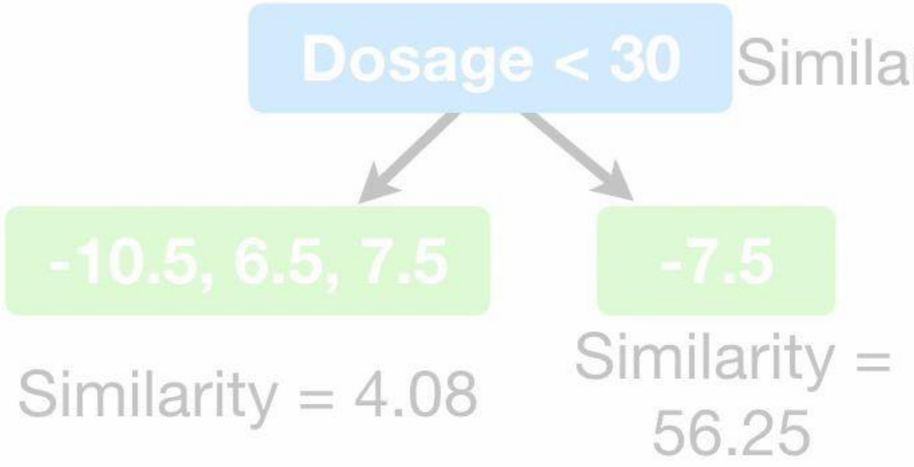
Gain = 4.08 + 56.25 - 4 = 56.33

Again, since the **Gain** for **Dosage < 30** (**Gain = 56.33**) is less than the **Gain** for **Dosage < 15** (**Gain = 120.33**), **Dosage < 15** is better at splitting the observations.

Predicted Drug Effectiveness

0.5

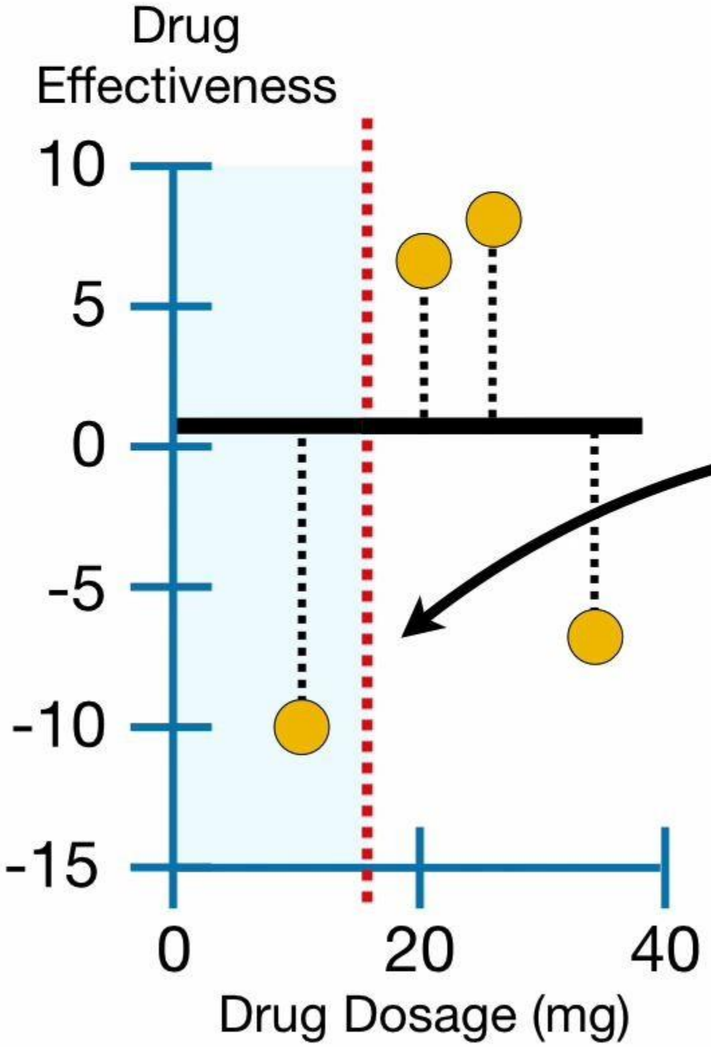


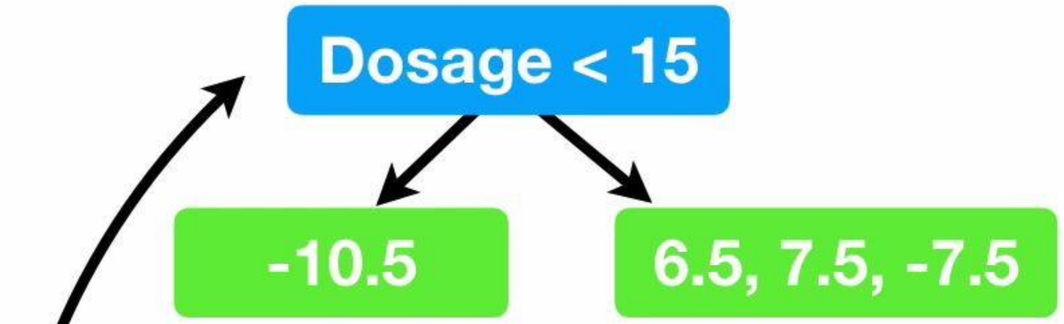


And since we can't shift the threshold over any further to the right, we are done comparing different thresholds…

Predicted Drug Effectiveness

0.5

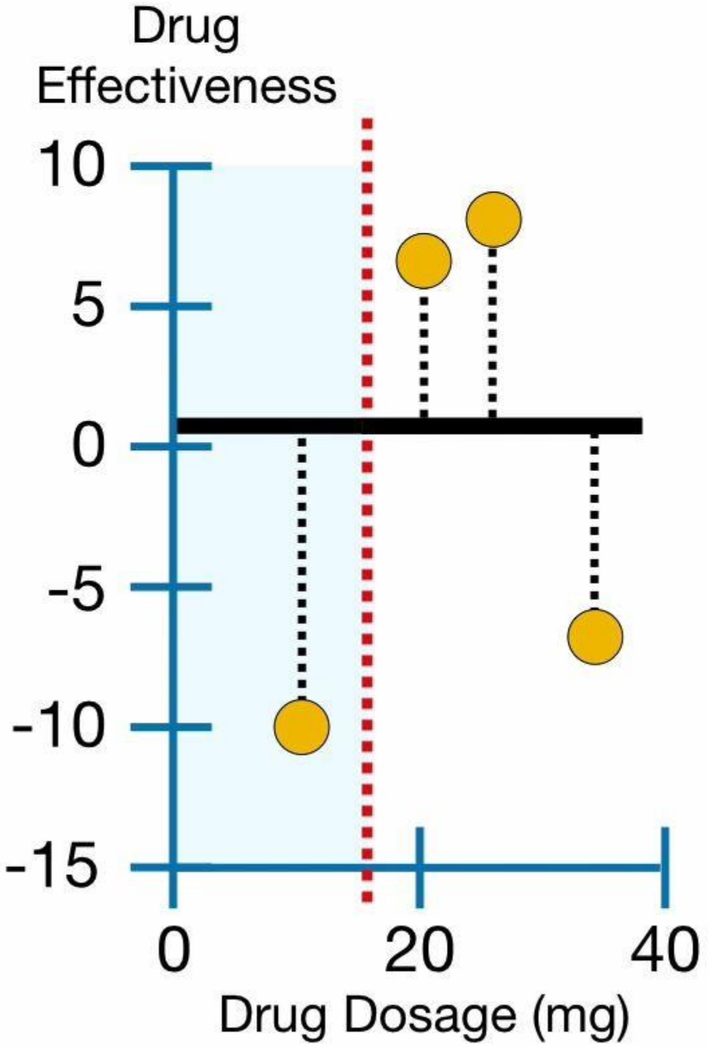




…and we will use the threshold that gave us the largest **Gain**, **Dosage < 15**, for the first branch in the tree.

Predicted Drug Effectiveness

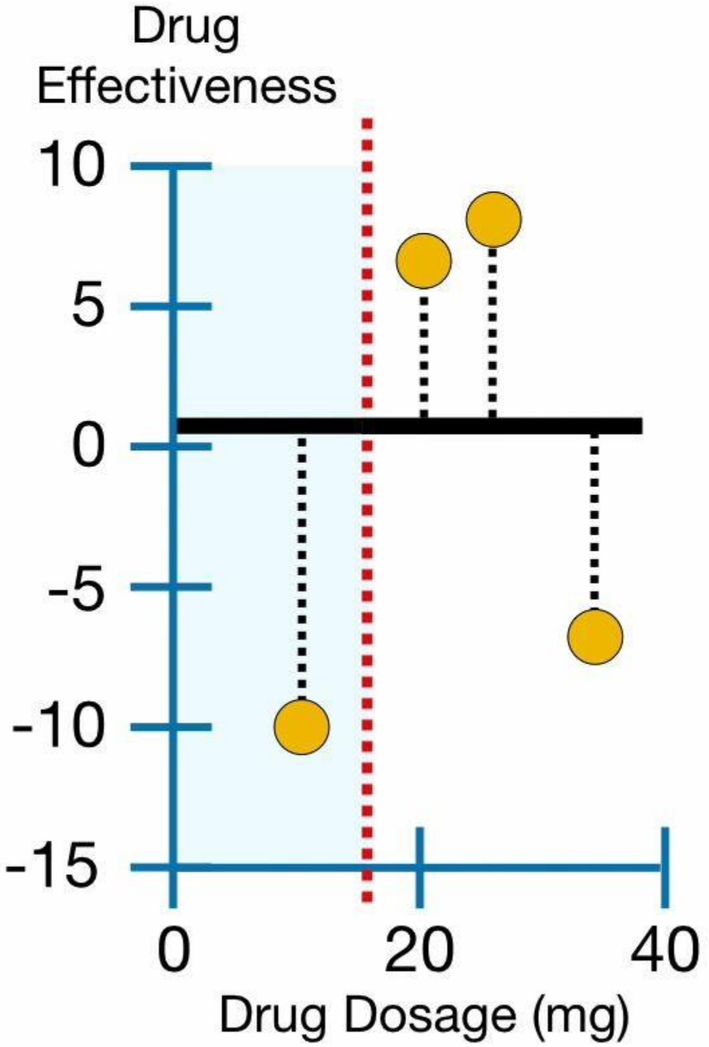
0.5



Now, since there is only one **Residual** in the leaf on the left, we can't split it any further.

Predicted Drug Effectiveness

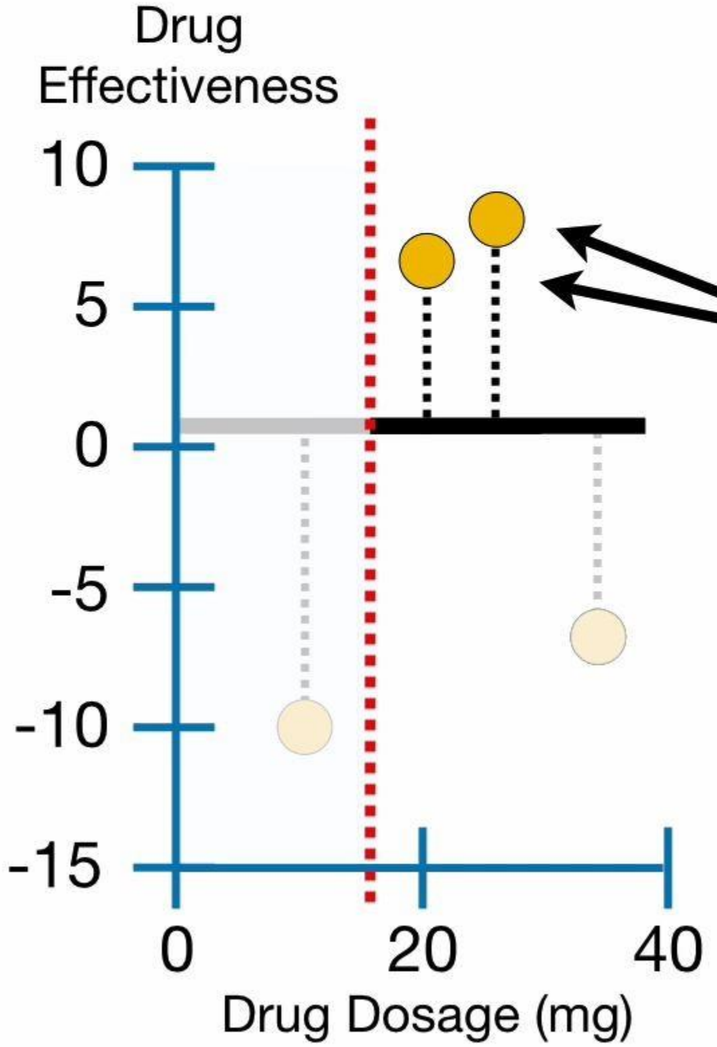
0.5



However, we can split the **3** **Residuals** in the leaf on the right.

Predicted Drug Effectiveness

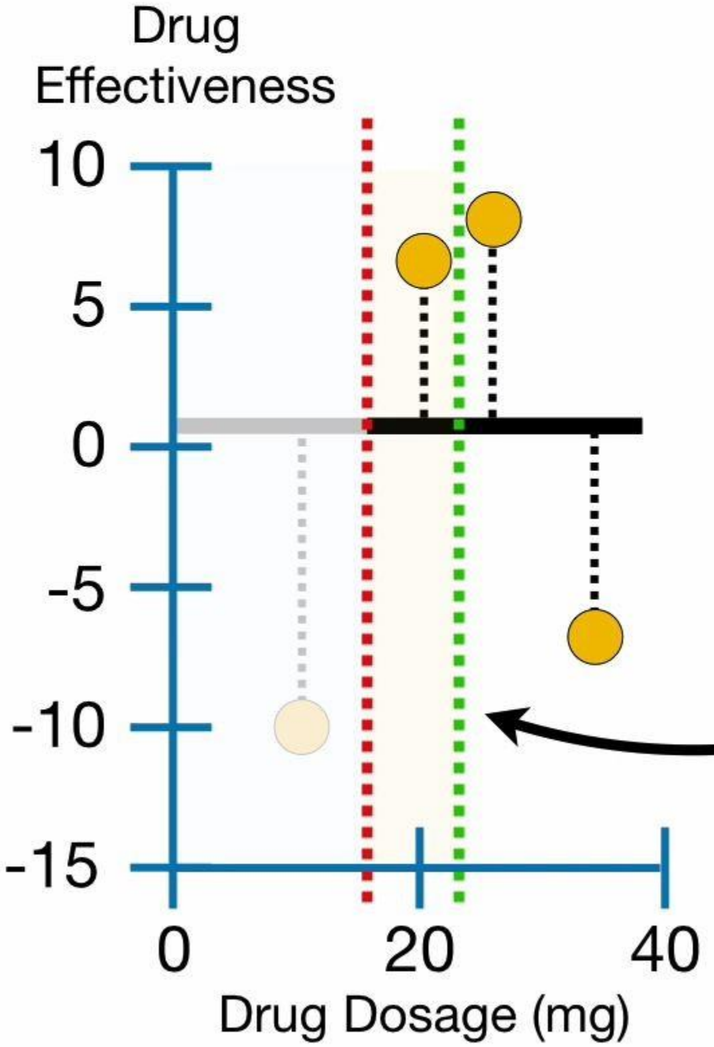
0.5



So we start with these two observations…

Predicted Drug Effectiveness

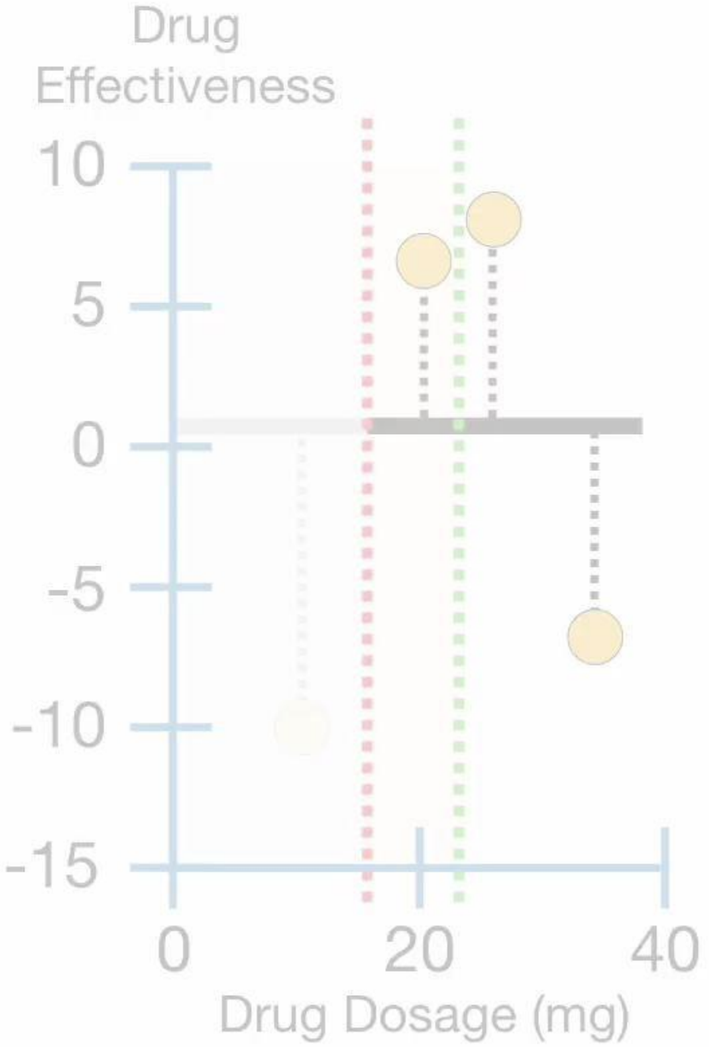
0.5

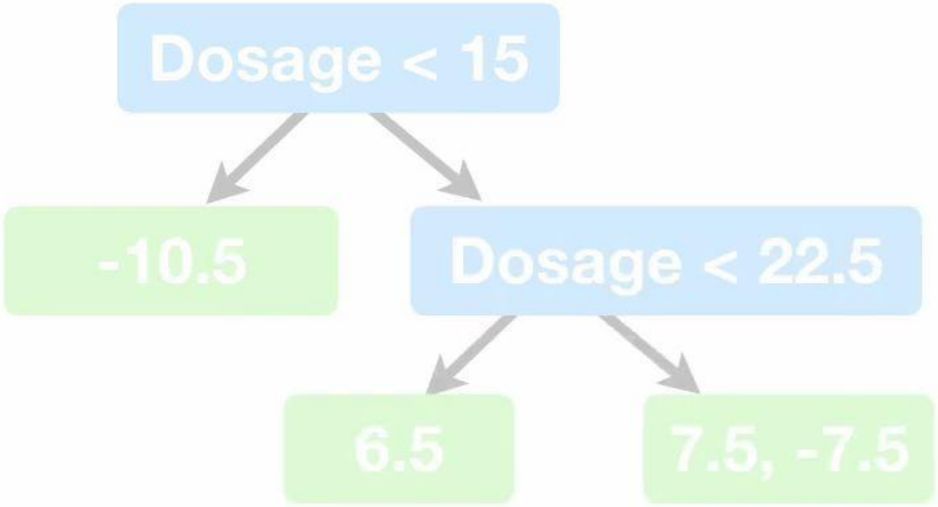


…and their average **Dosage** is **22.5**, which corresponds to this **dotted green line**.

Predicted Drug Effectiveness

0.5



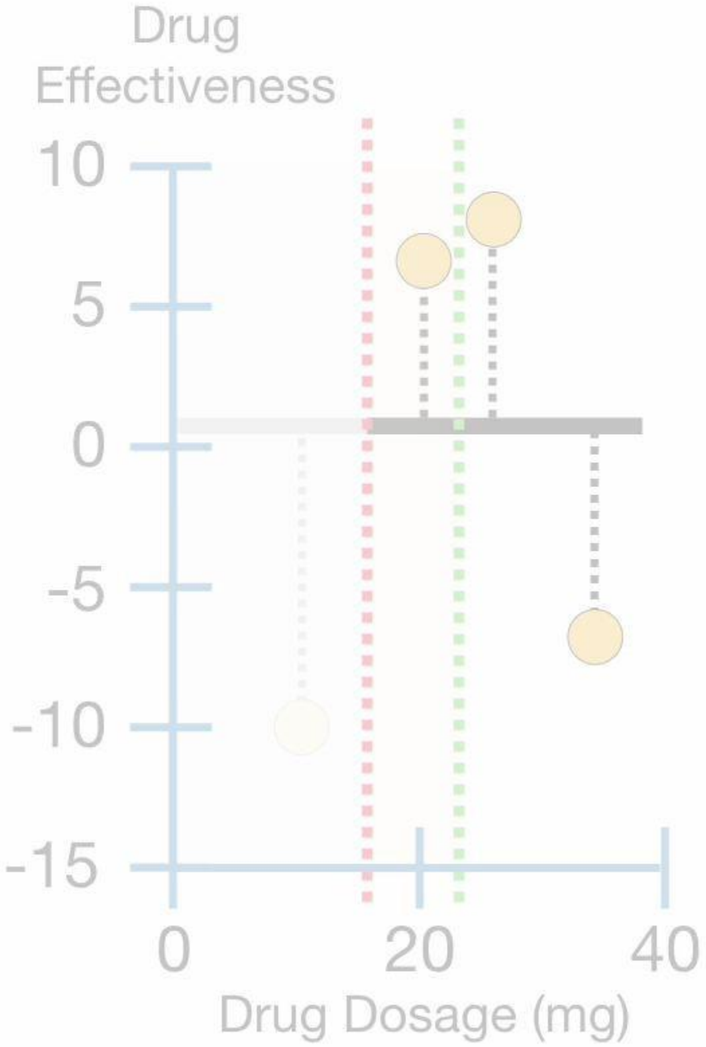


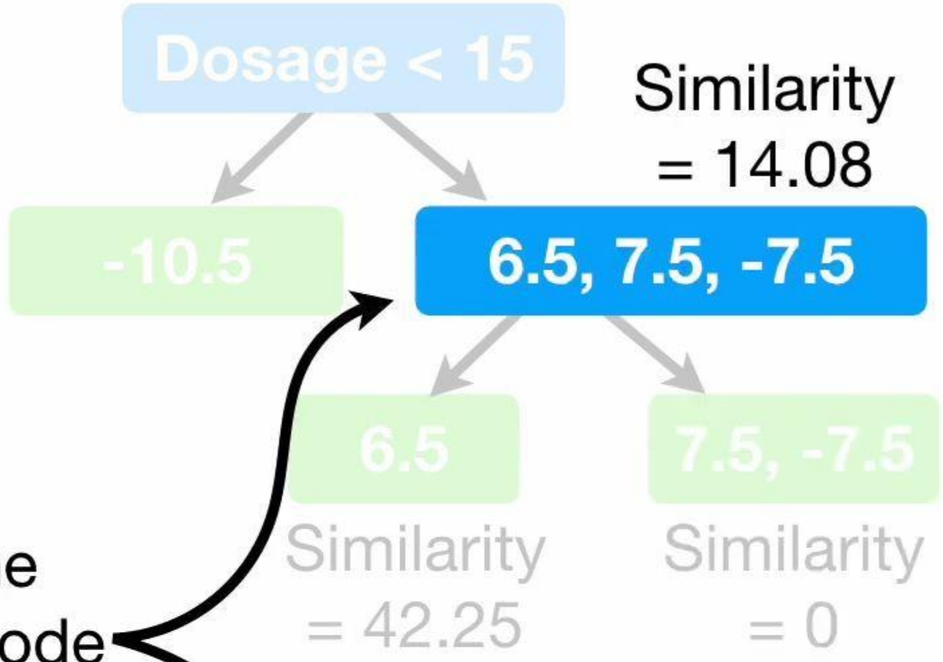
Now, just like before, we calculate the **Similarity Scores** for the leaves.

$$\text{Similarity Score} = \frac{\text{Sum of Residuals, Squared}}{\text{Number of Residuals} + \lambda}$$

Predicted Drug Effectiveness

0.5



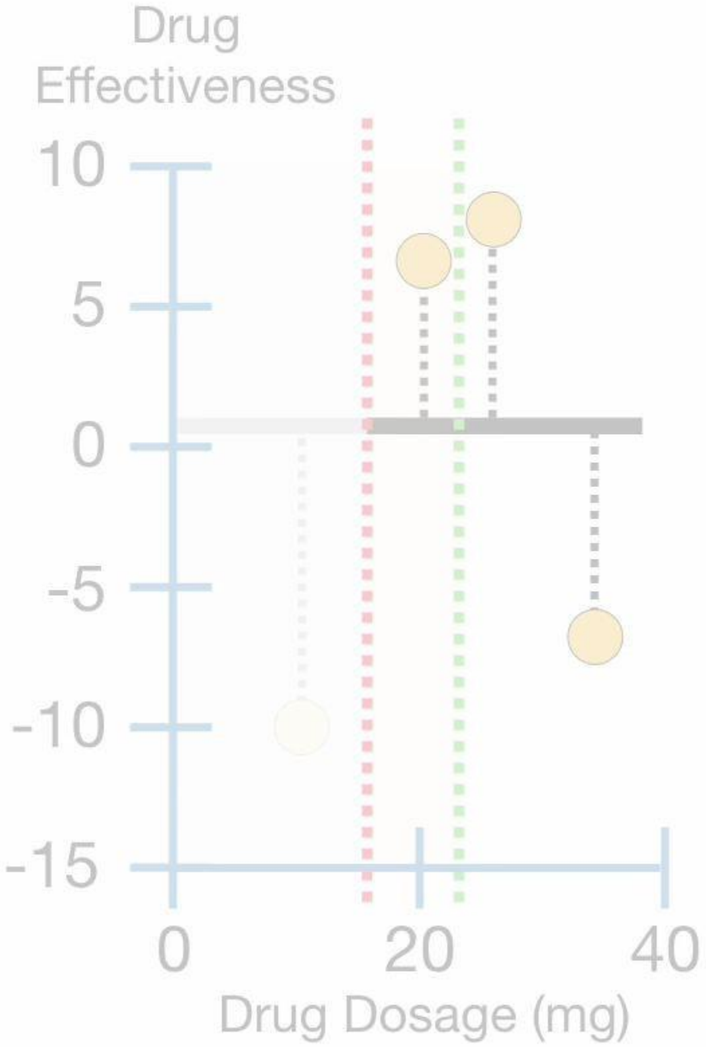


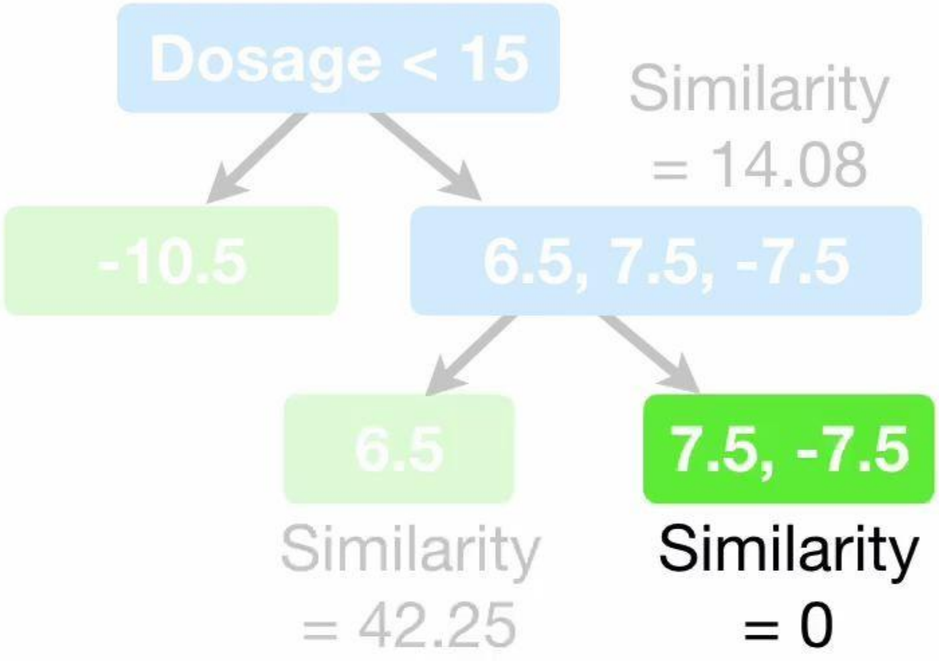
NOTE: We calculated the **Similarity Score** for this node when we figured out how to split the root.

$$\text{Similarity Score} = \frac{(6.5 + 7.5 + -7.5)^2}{3 + \mathbf{0}}$$

Predicted Drug Effectiveness

0.5



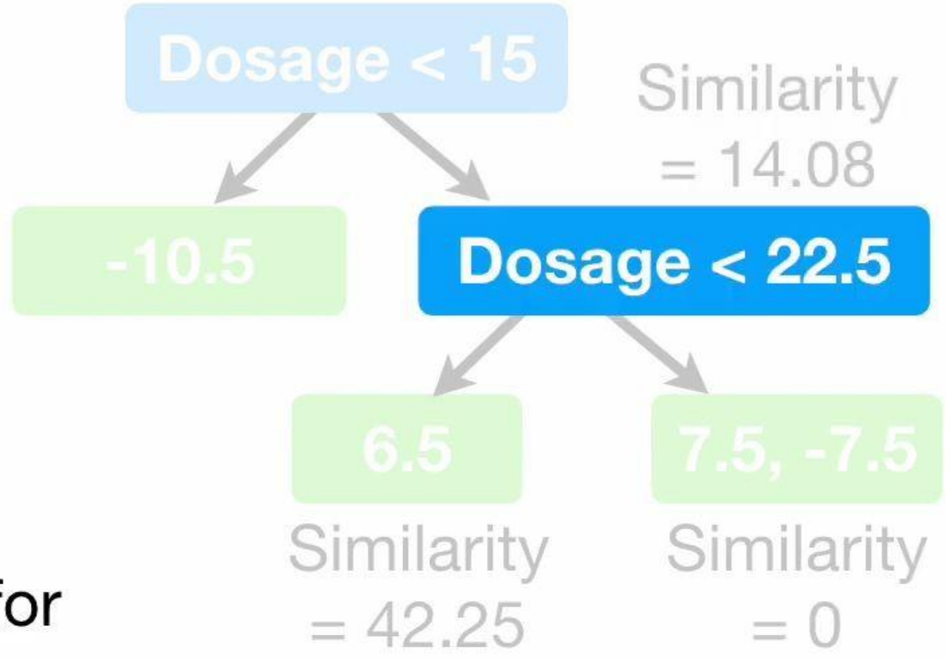


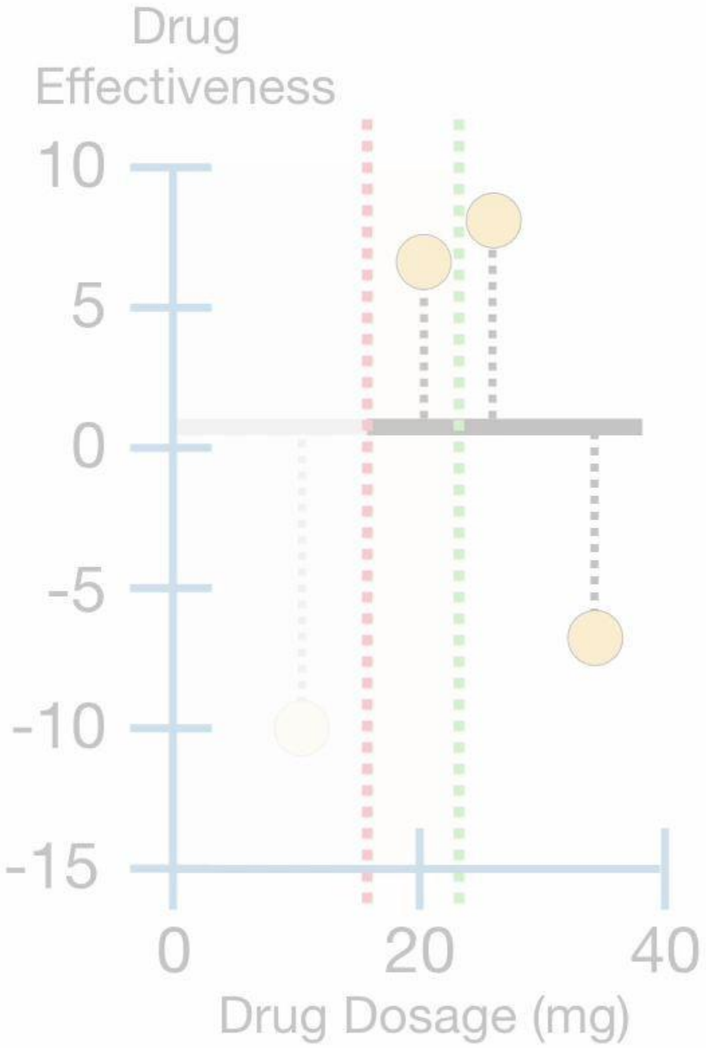
So now we calculate the **Gain**.

$$\text{Gain} = 42.25 + \text{Right}_{\text{Similarity}} - \text{Root}_{\text{Similarity}}$$

Predicted Drug Effectiveness

0.5



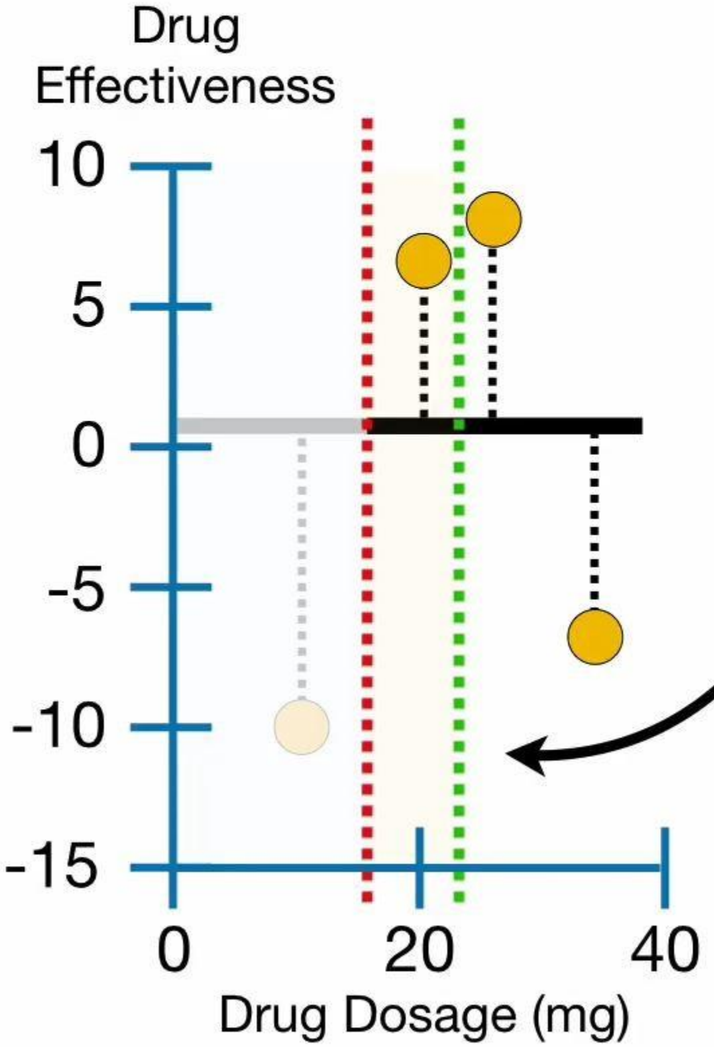


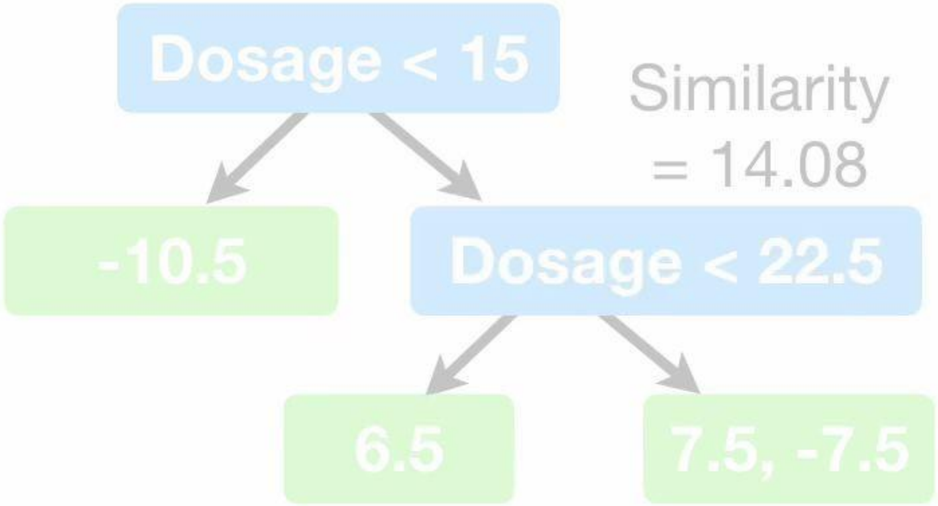
And we get **Gain = 28.17** for when the threshold is **Dosage < 22.5**.

Gain = 42.25 + 0 - 14.08 = 28.17

Predicted Drug Effectiveness

0.5

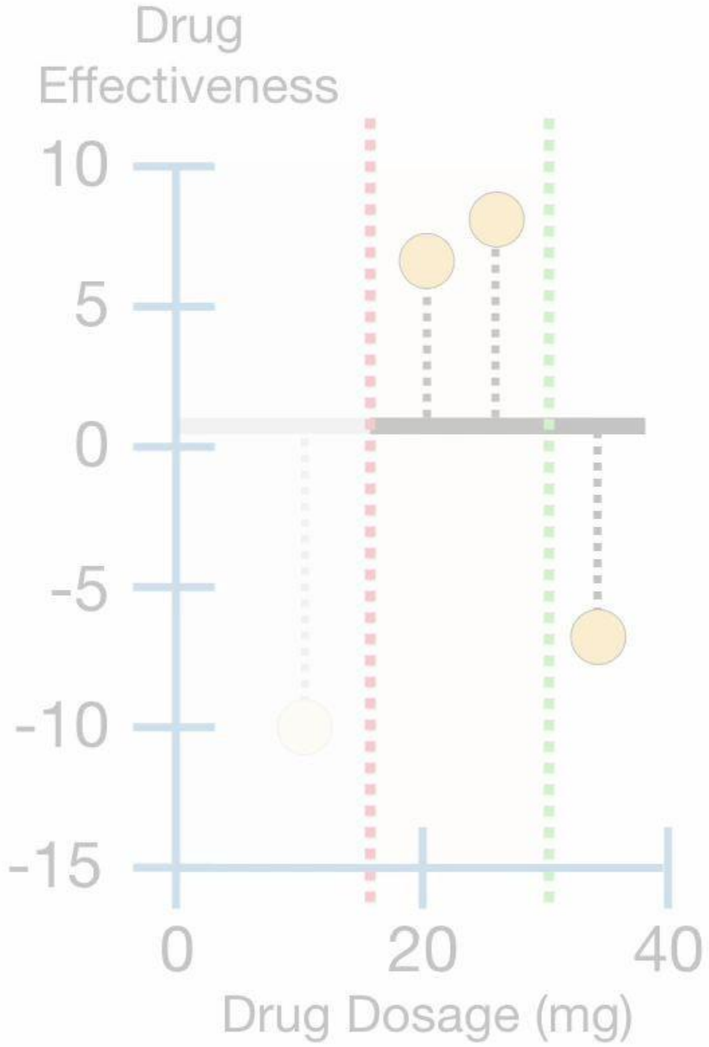


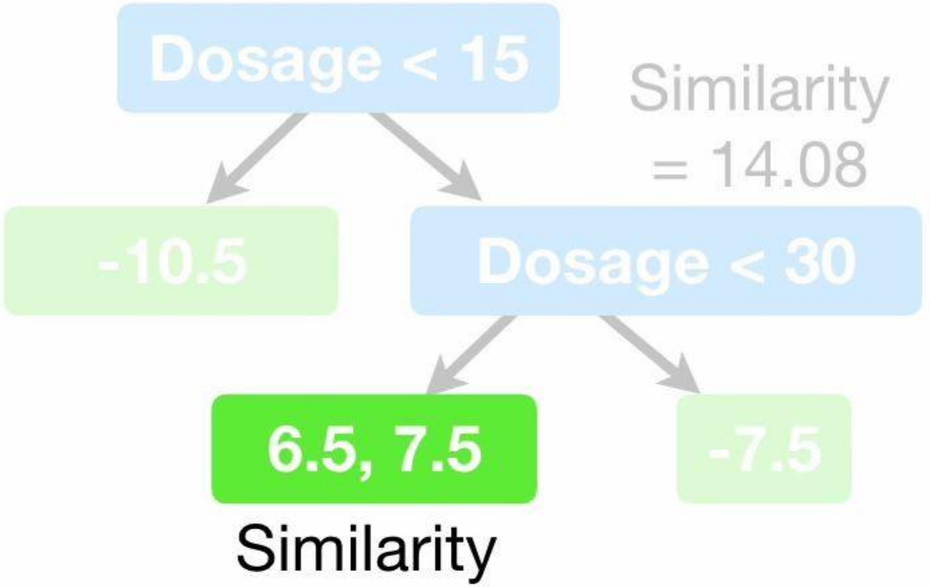


Now we shift the threshold over so that it is the average of the last two observations…

Predicted Drug Effectiveness

0.5



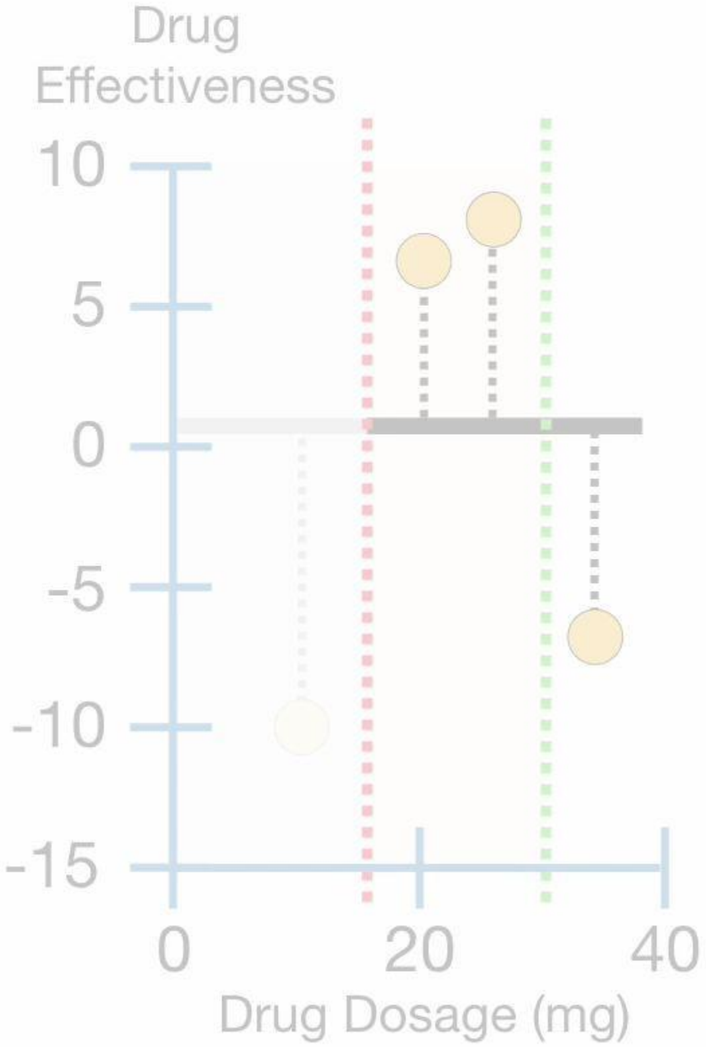


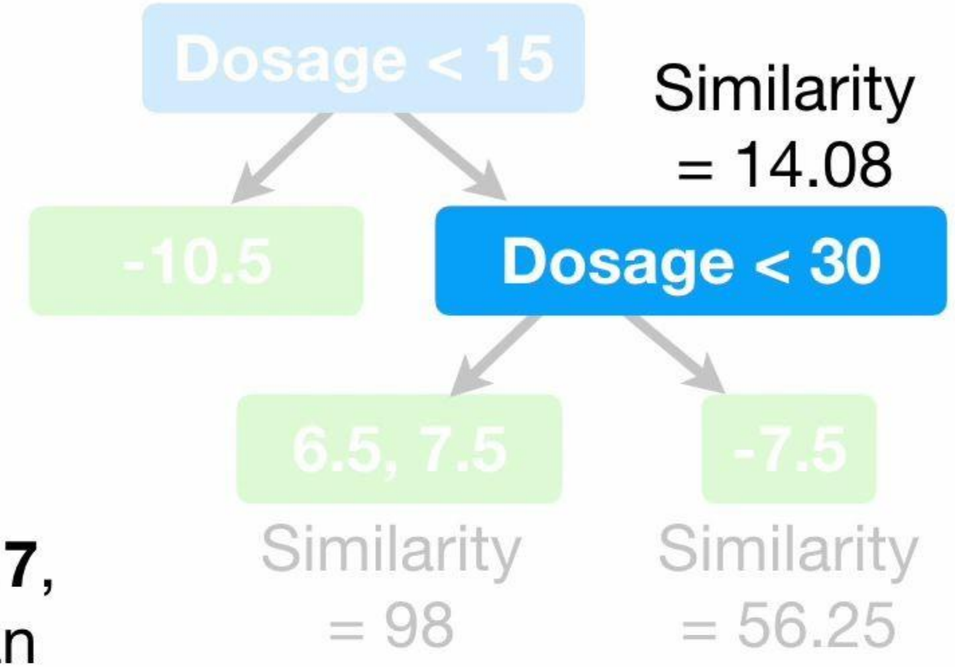
…calculate the **Similarity Scores** for the leaves…

$$\text{Similarity Score} = \frac{(6.5 + 7.5)^2}{2 + 0}$$

Predicted Drug Effectiveness

0.5



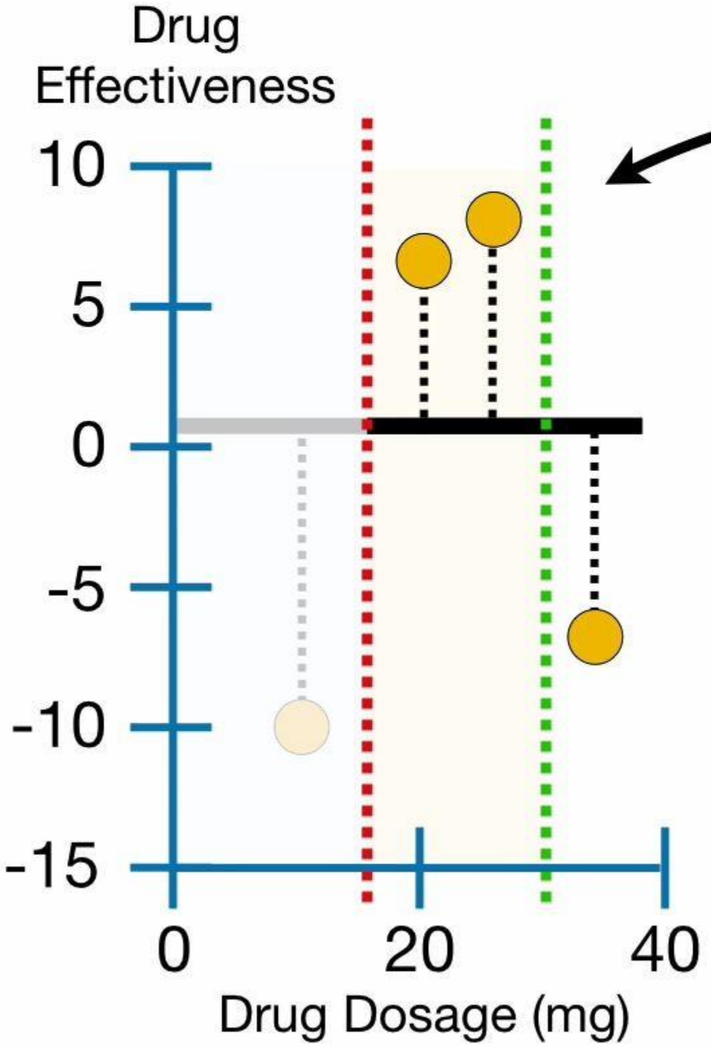


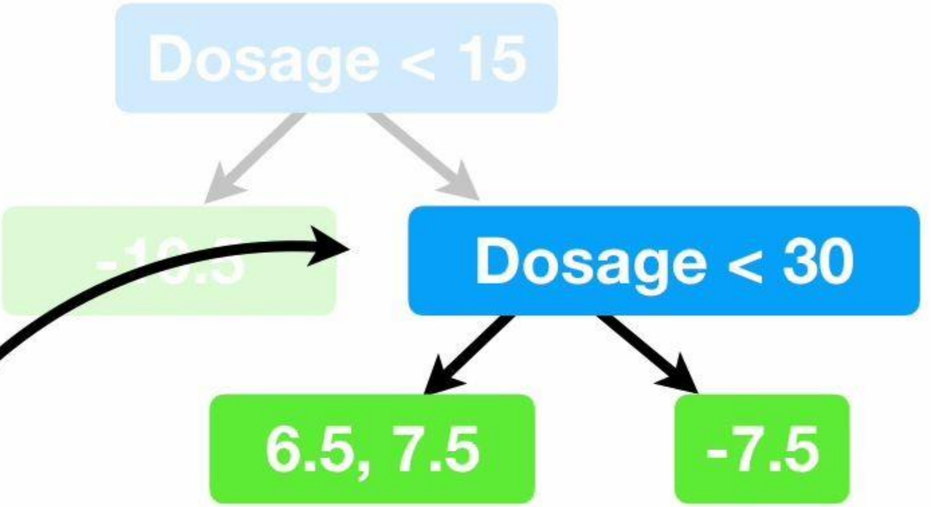
And we get **Gain = 140.17**, which is much larger than **28.17**, when the threshold was **Dosage < 22.5**.

Gain = 98 + 56.25 - 14.08 = 140.17

Predicted Drug Effectiveness

0.5



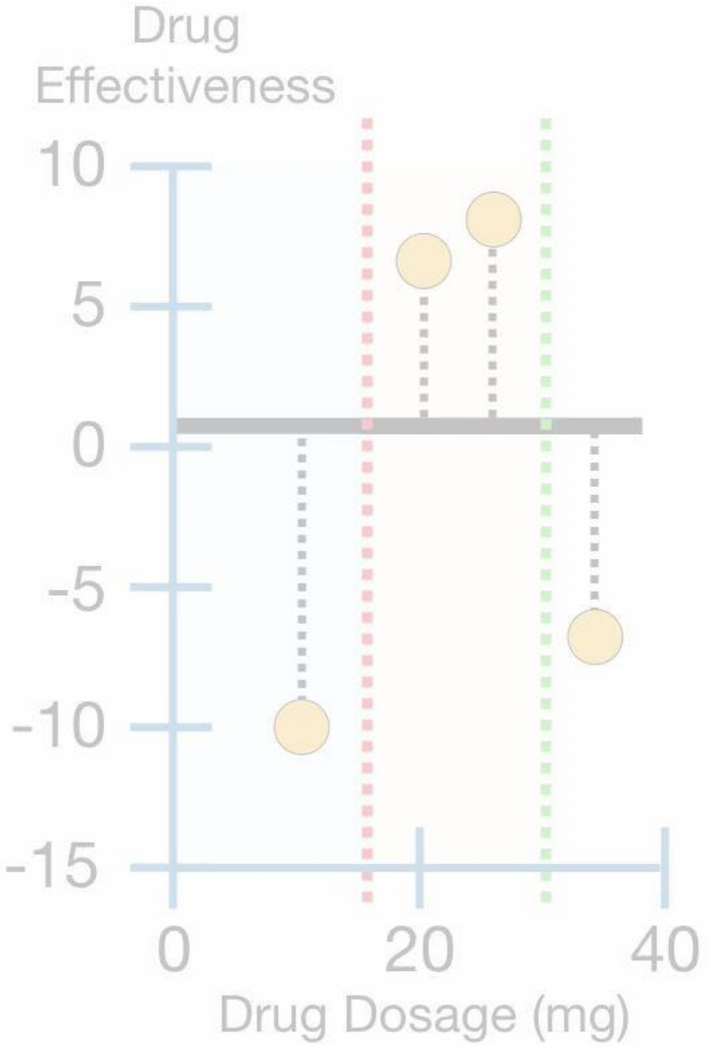


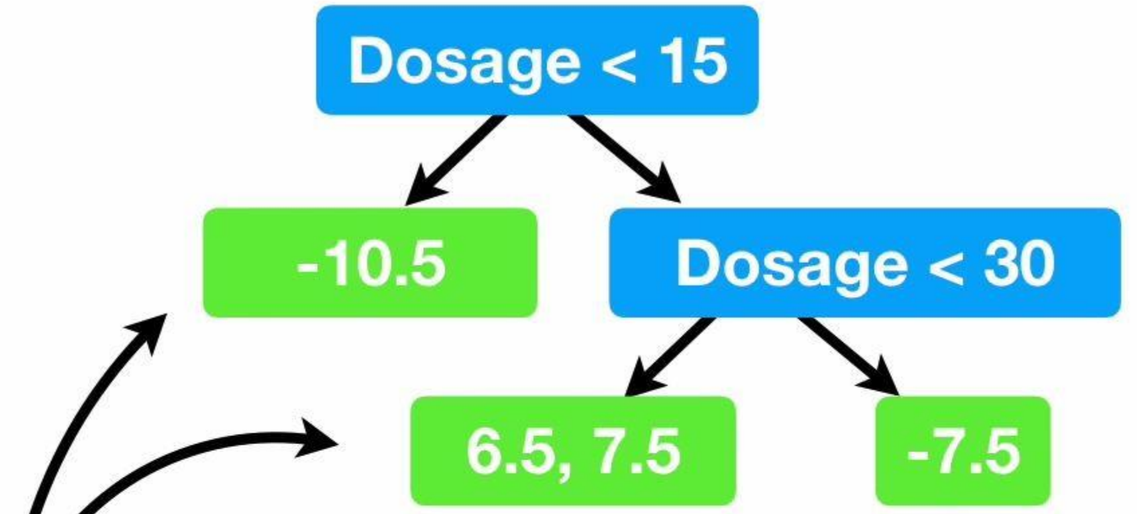
So we will use **Dosage < 30** as the threshold for this branch.

Gain = 98 + 56.25 - 14.08 = 140.17

Predicted Drug Effectiveness

0.5

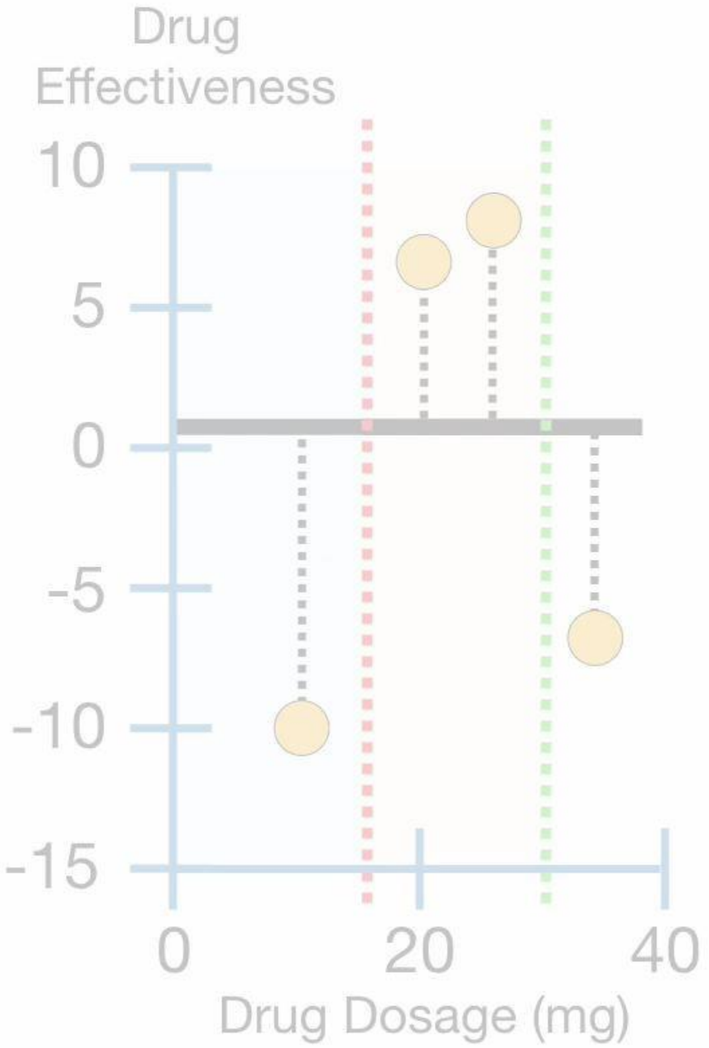


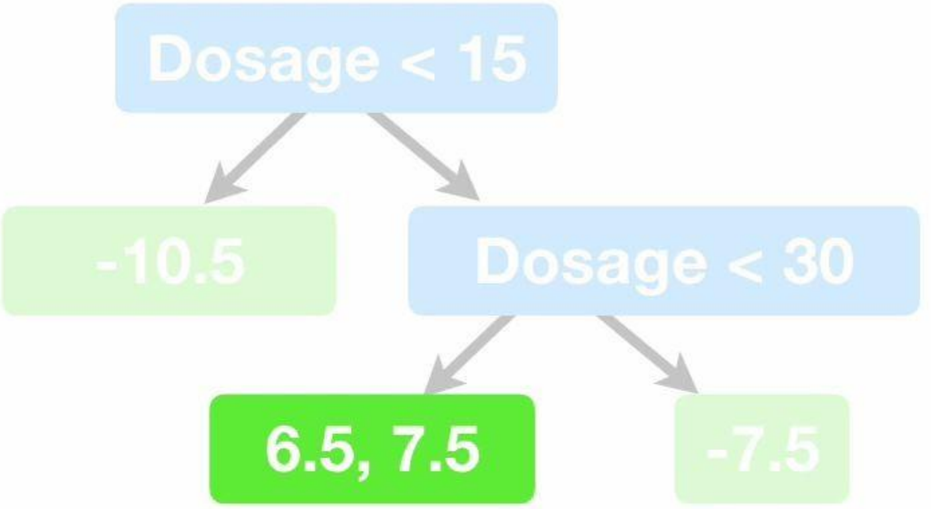


NOTE: To keep this example from getting out of hand, I've limited the tree depth to two levels…

Predicted Drug Effectiveness

0.5

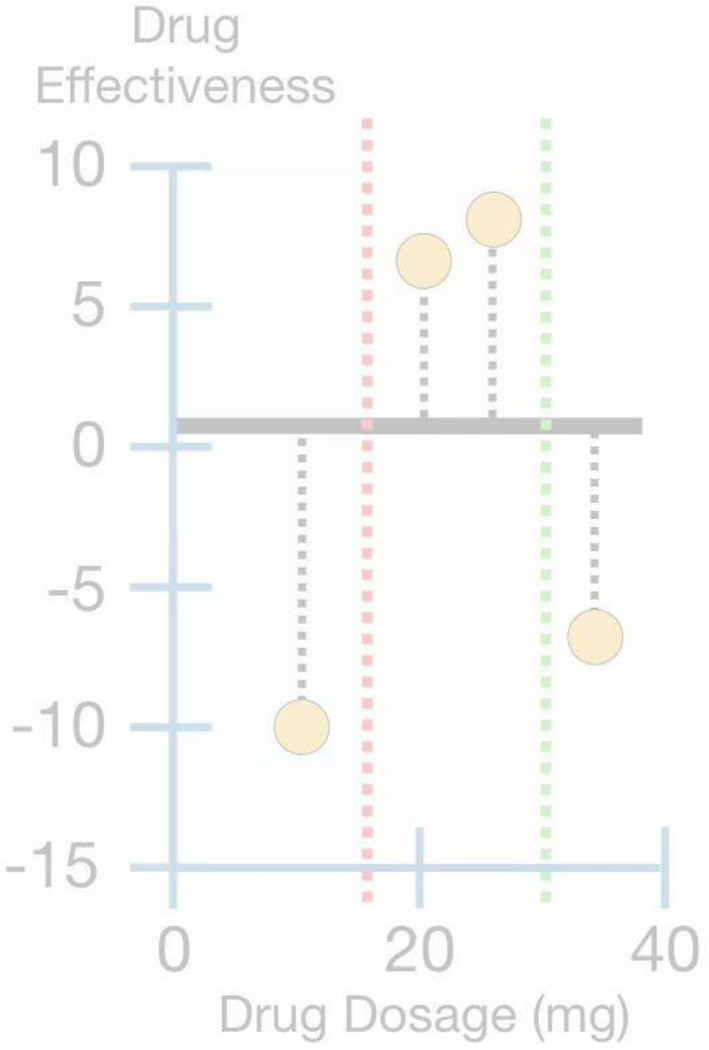


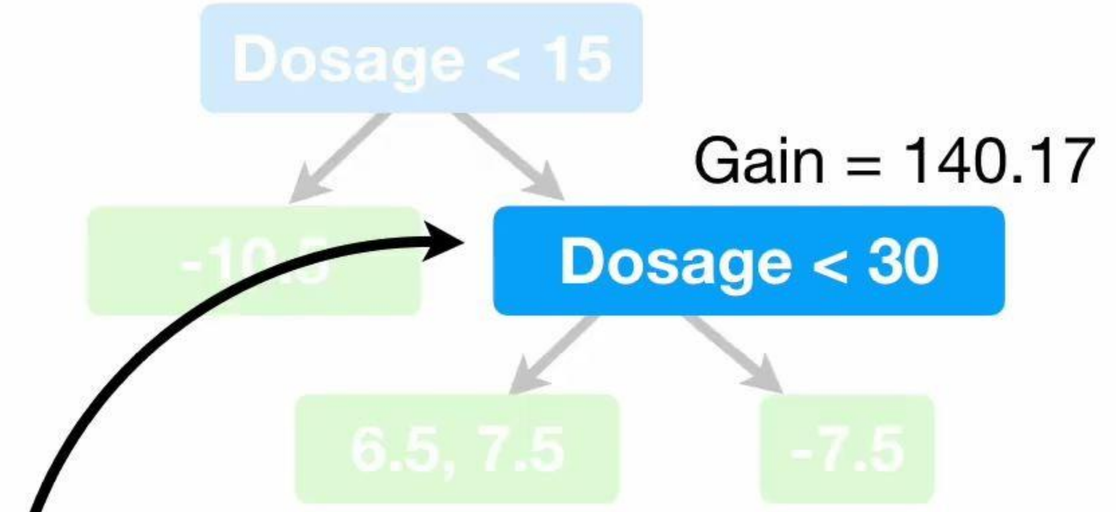


…and this means we will not split this leaf any further, and we are done building this tree.

Predicted Drug Effectiveness

0.5

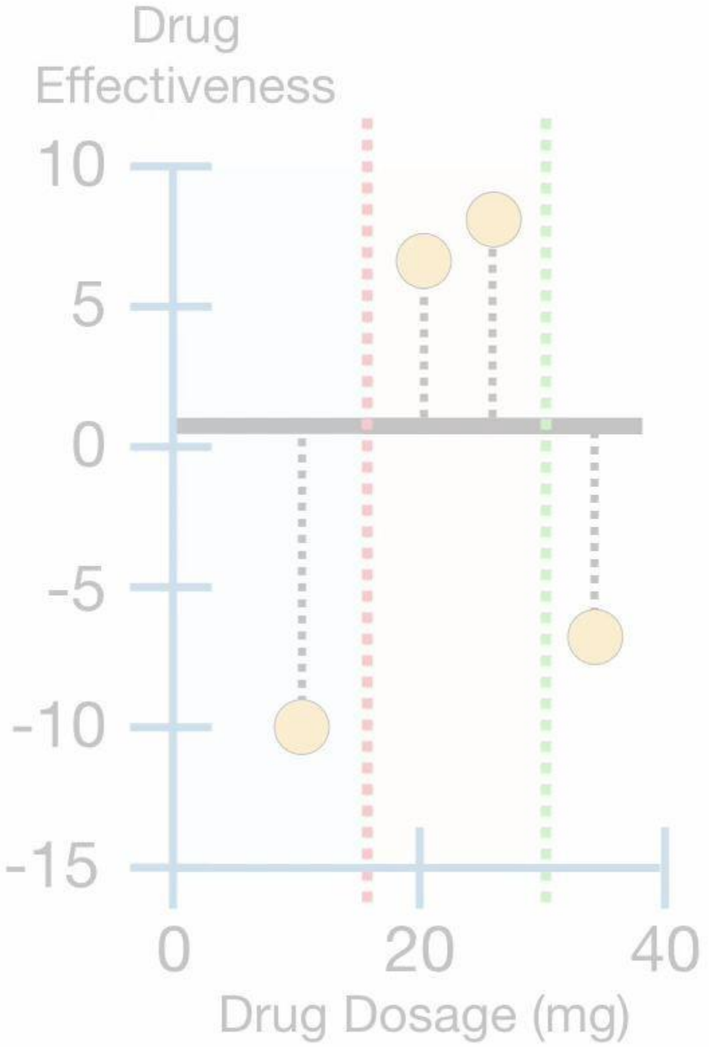




We **Prune** an **XGBoost Tree** based on its **Gain** values.

Predicted Drug Effectiveness

0.5



Gain = 120.33

Dosage < 15

Gain = 140.17

-10.5

Dosage < 30

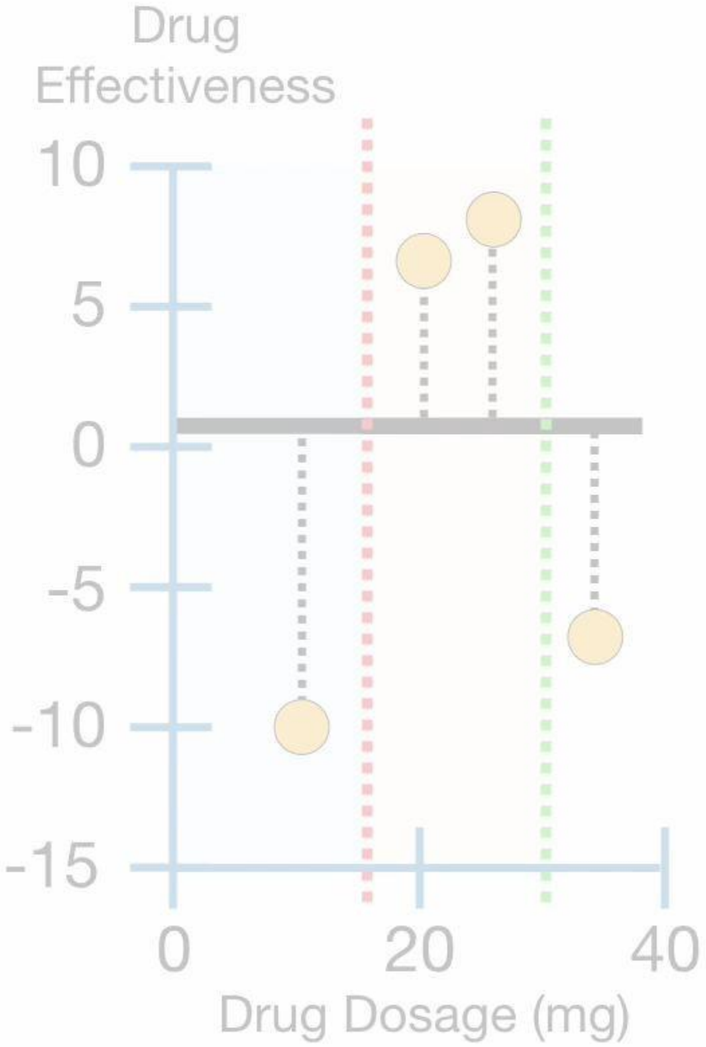
6.5, 7.5

-7.5

We start by picking a number, for example, **130**.

Predicted Drug Effectiveness

0.5



Gain = 120.33

Dosage < 15

Gain = 140.17

-10.5

Dosage < 30

6.5, 7.5

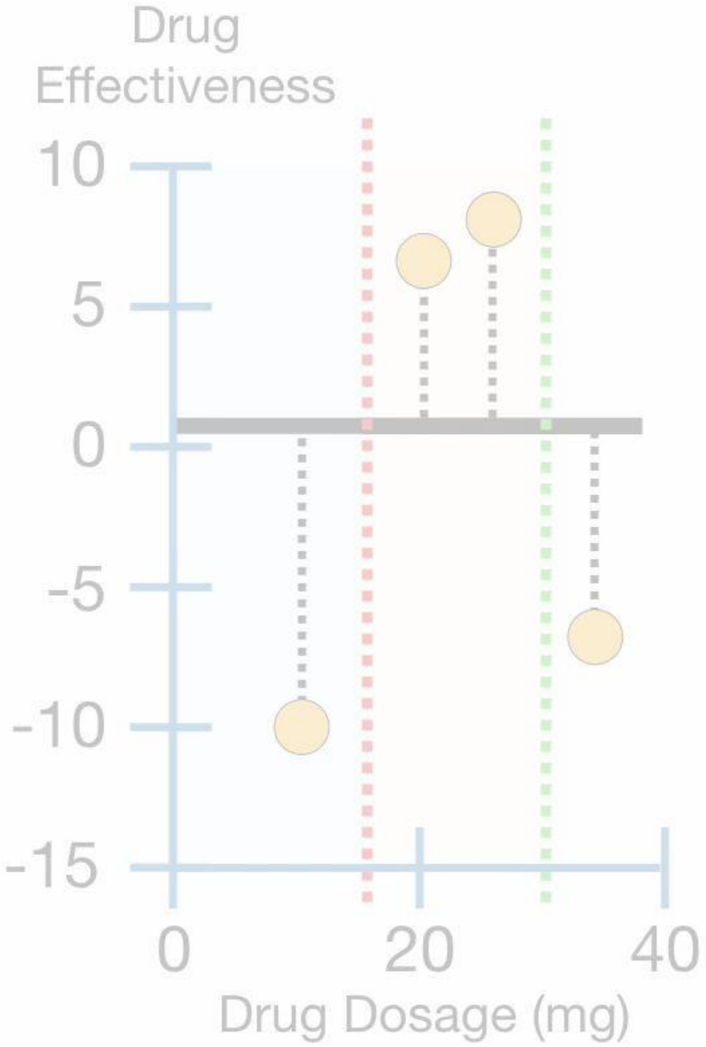
-7.5

We start by picking a number, for example, **130**.

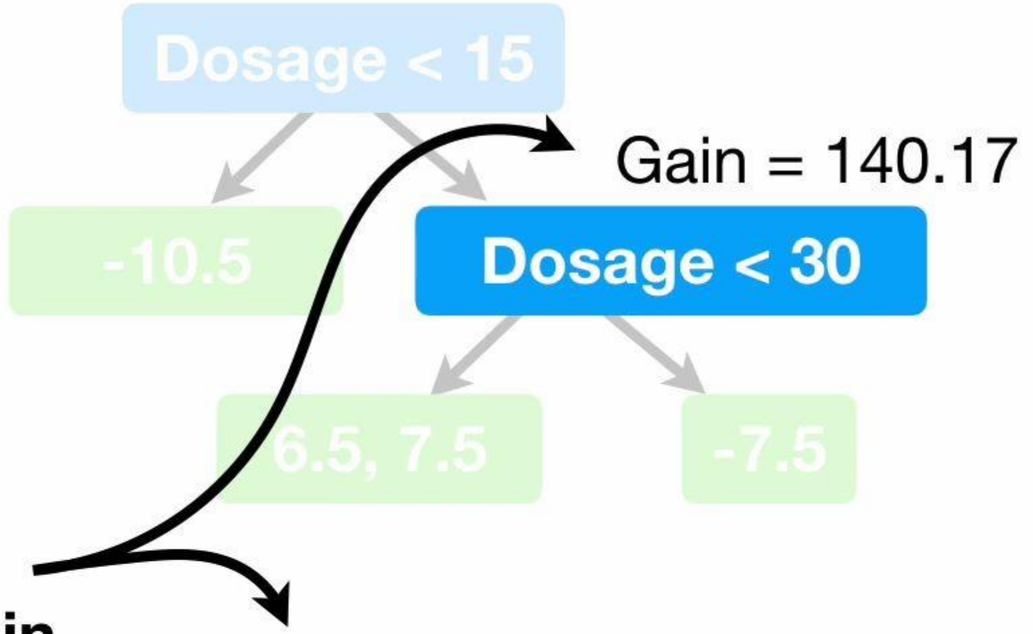
TERMINOLOGY ALERT!!!
XGBoost calls this number ***γ* (gamma)**.

Predicted Drug Effectiveness

0.5

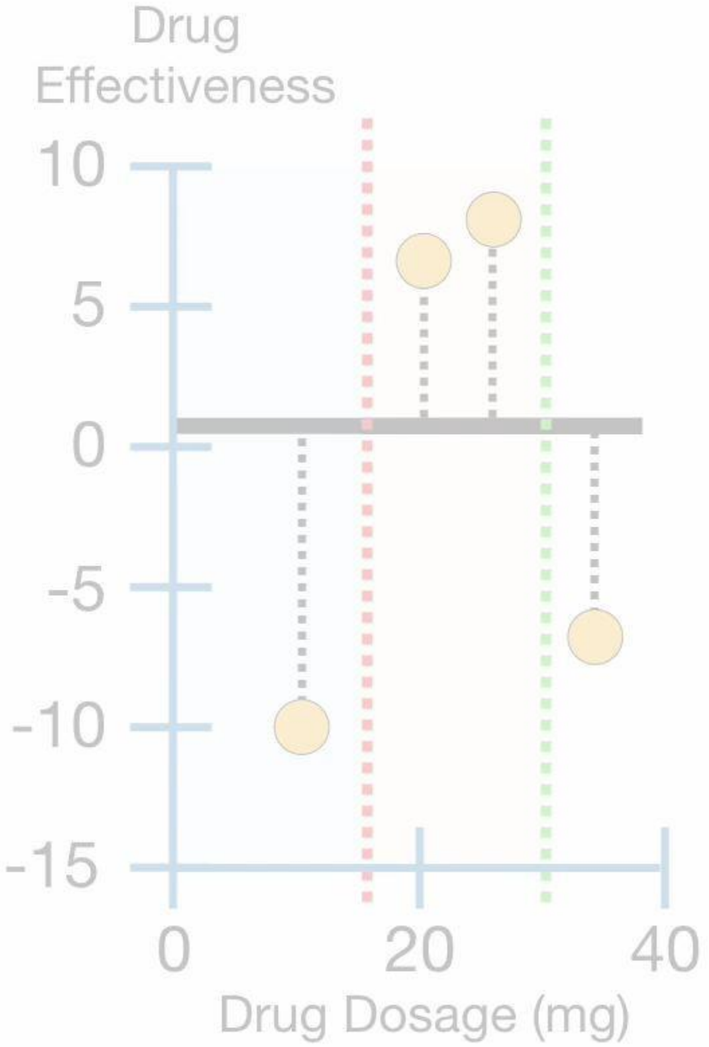


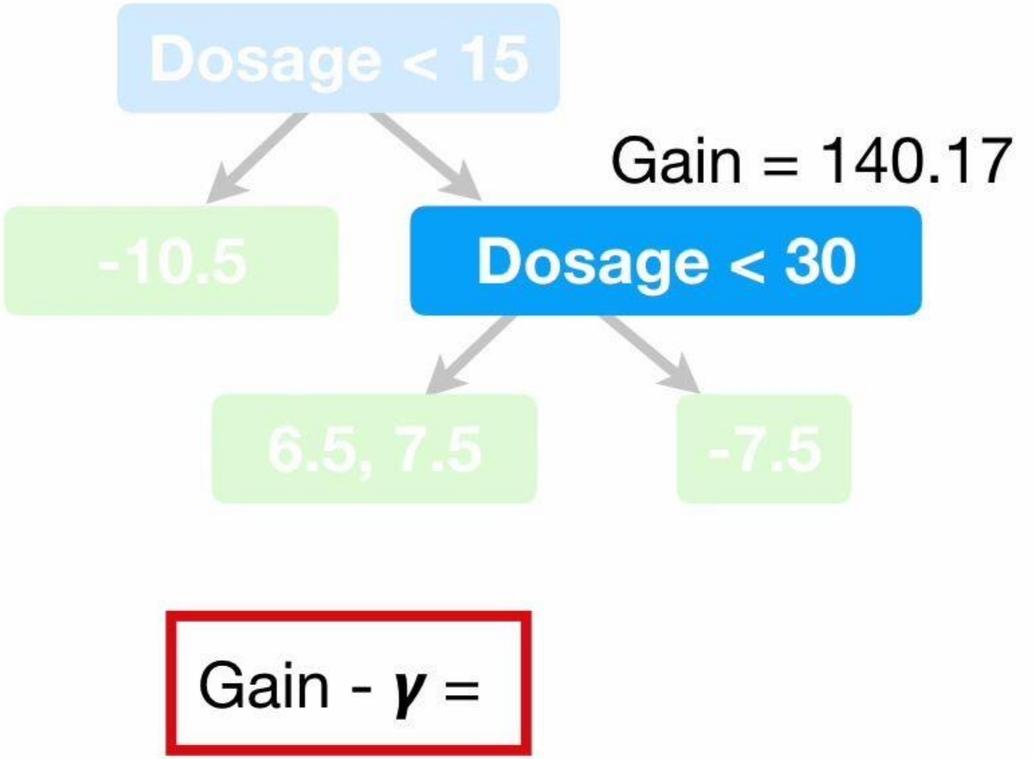
We then calculate the difference between the **Gain** associated with the lowest branch in the tree…



Predicted Drug Effectiveness

0.5



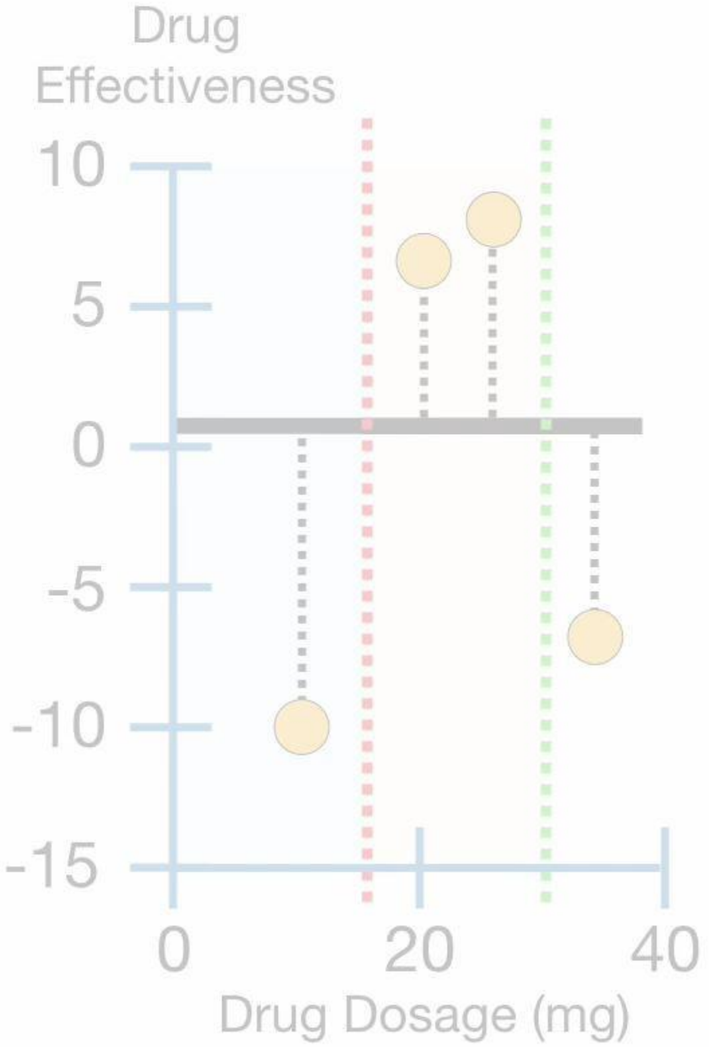


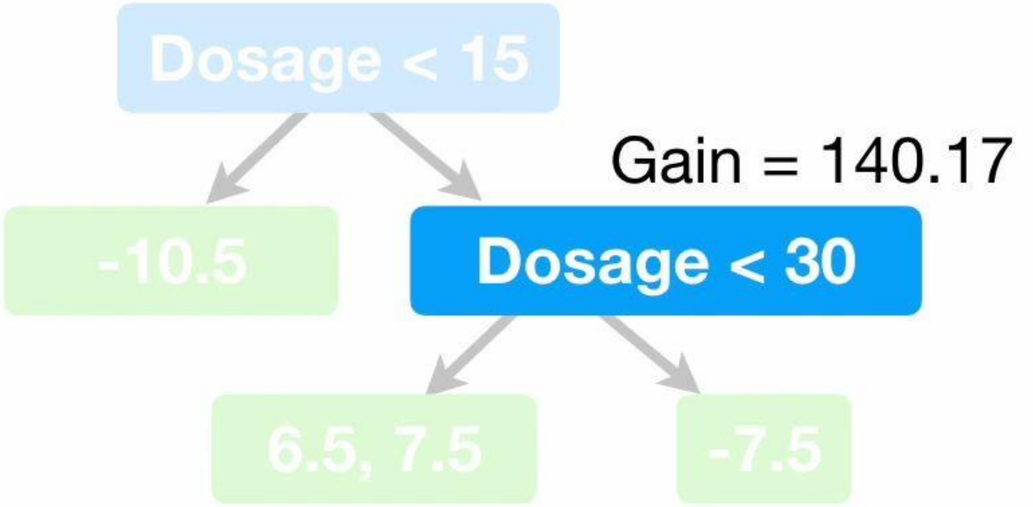
Gain - ***γ*** =

If the difference between the **Gain** and ***γ* (gamma)** is ***negative*** we will remove the branch…

Predicted Drug Effectiveness

0.5



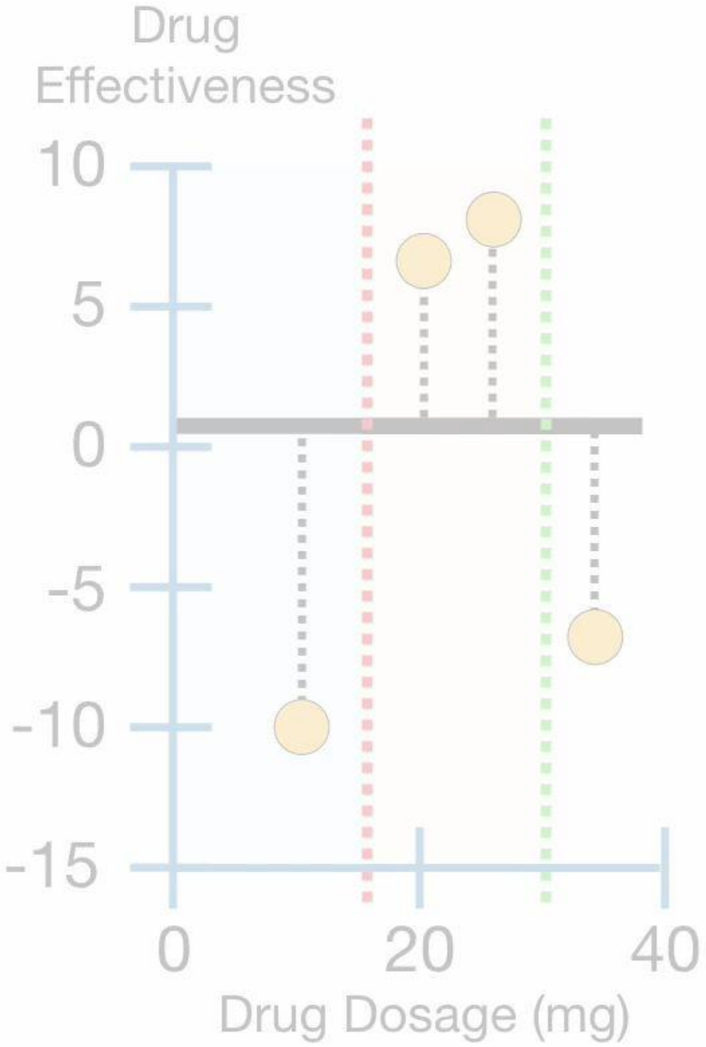


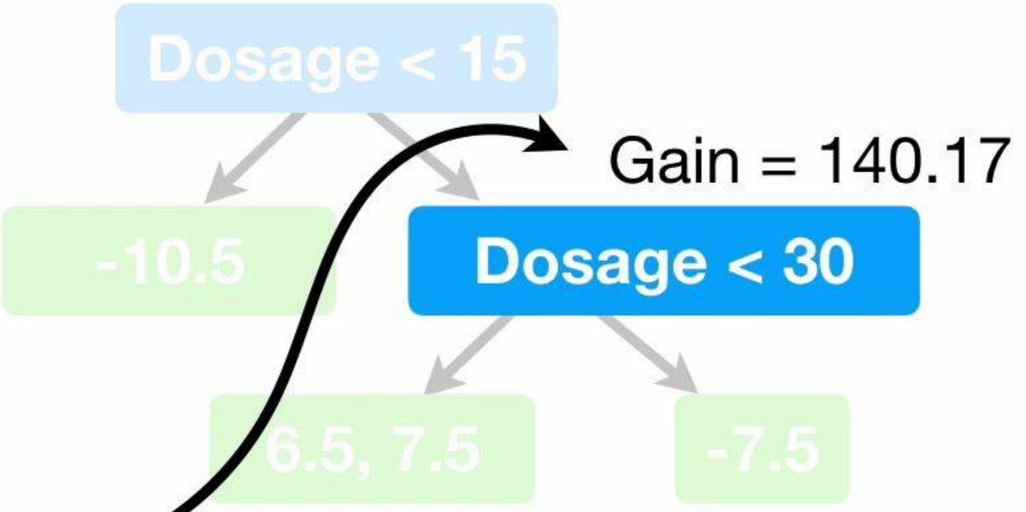
Gain - ***γ*** =

…and if the difference between the **Gain** and ***γ* (gamma)** is ***positive*** we will not remove the branch.

Predicted Drug Effectiveness

0.5





Gain = 140.17

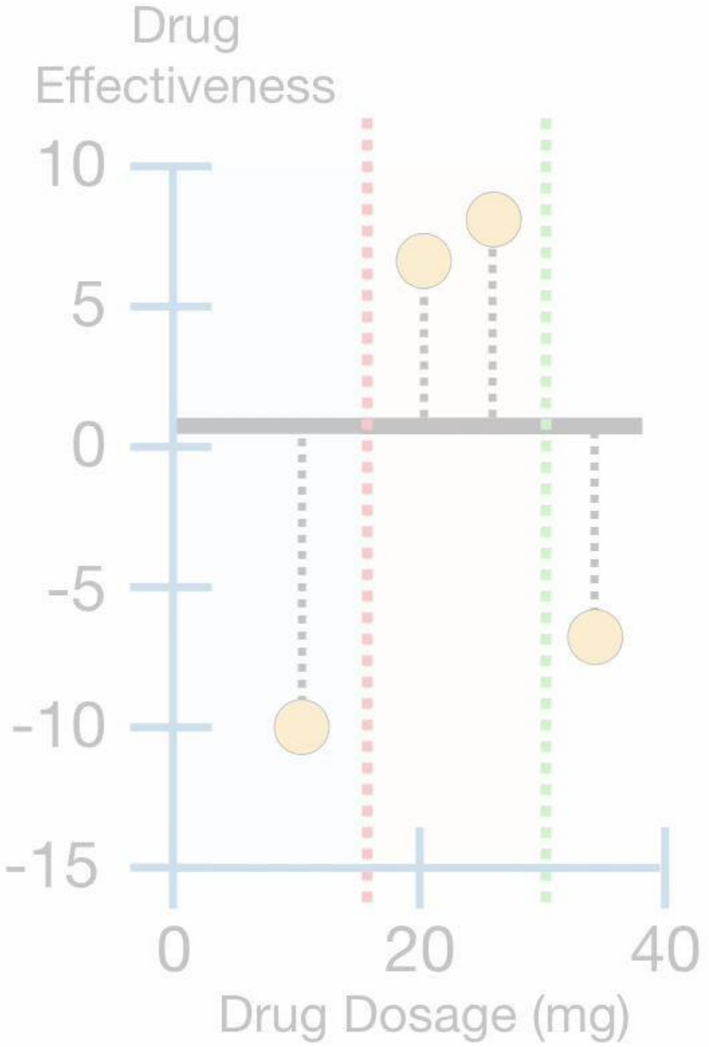
In this case, when we plug in the **Gain**…

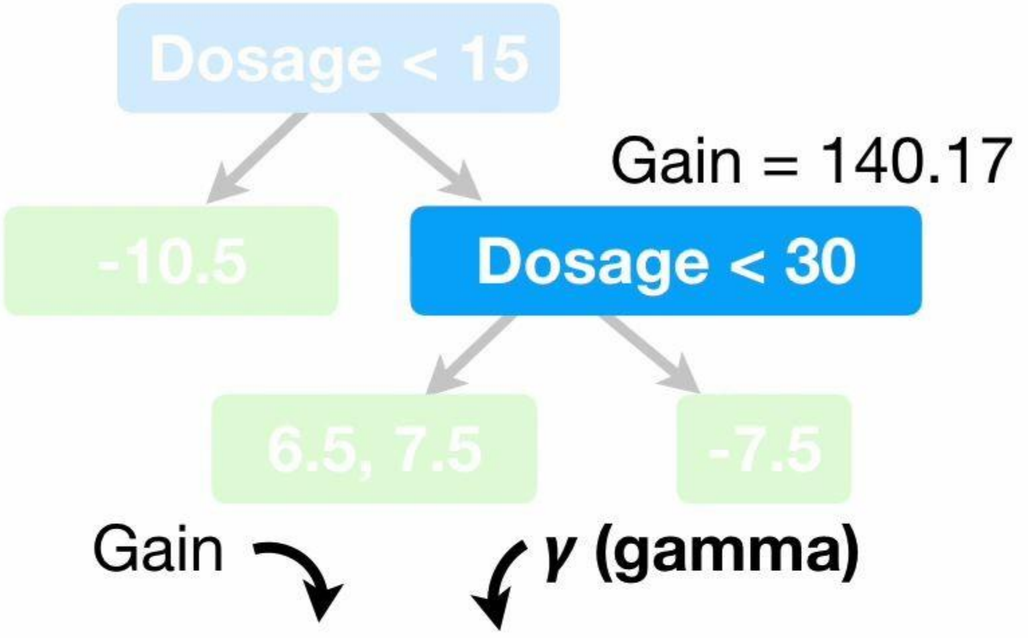
140.17 - 130 =

…and the value for ***γ*** **(gamma)**, **130**…

Predicted Drug Effectiveness

0.5

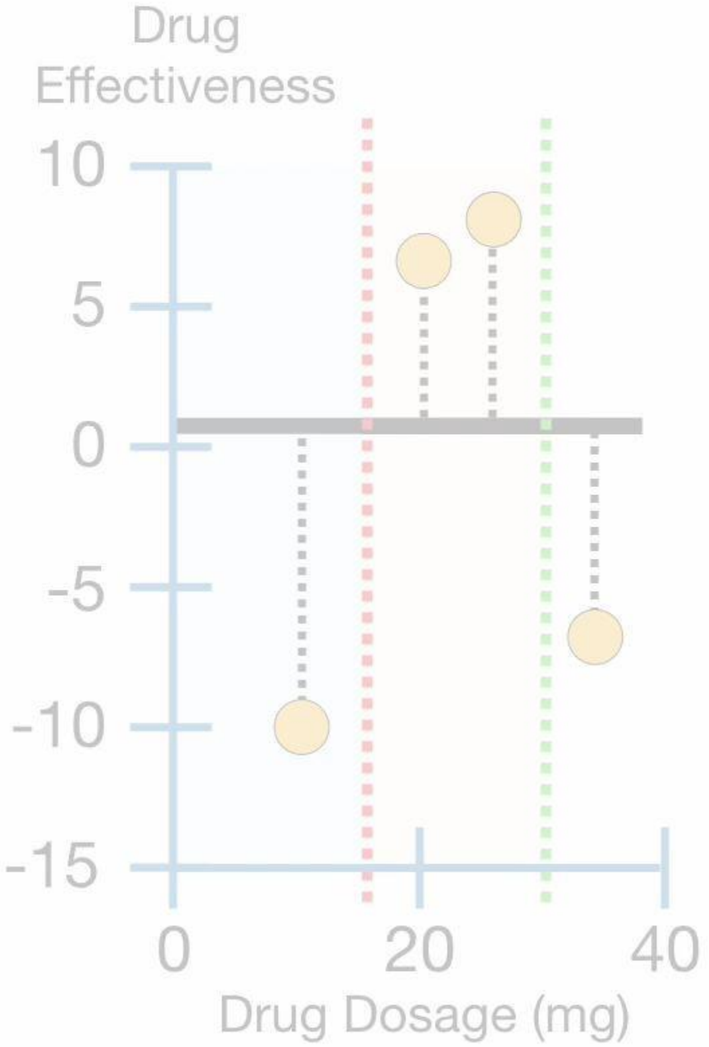




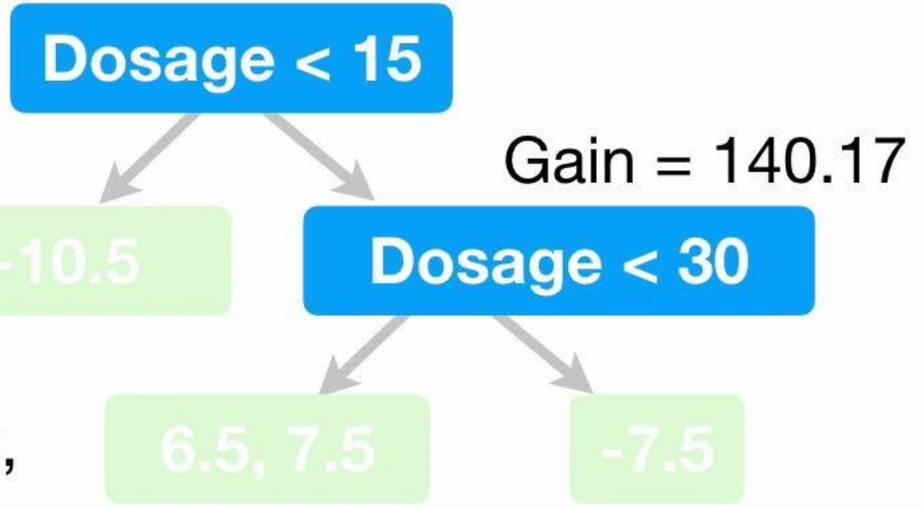
…we get a ***positive*** number, so we will not remove this branch and we are done pruning.

Predicted Drug Effectiveness

0.5



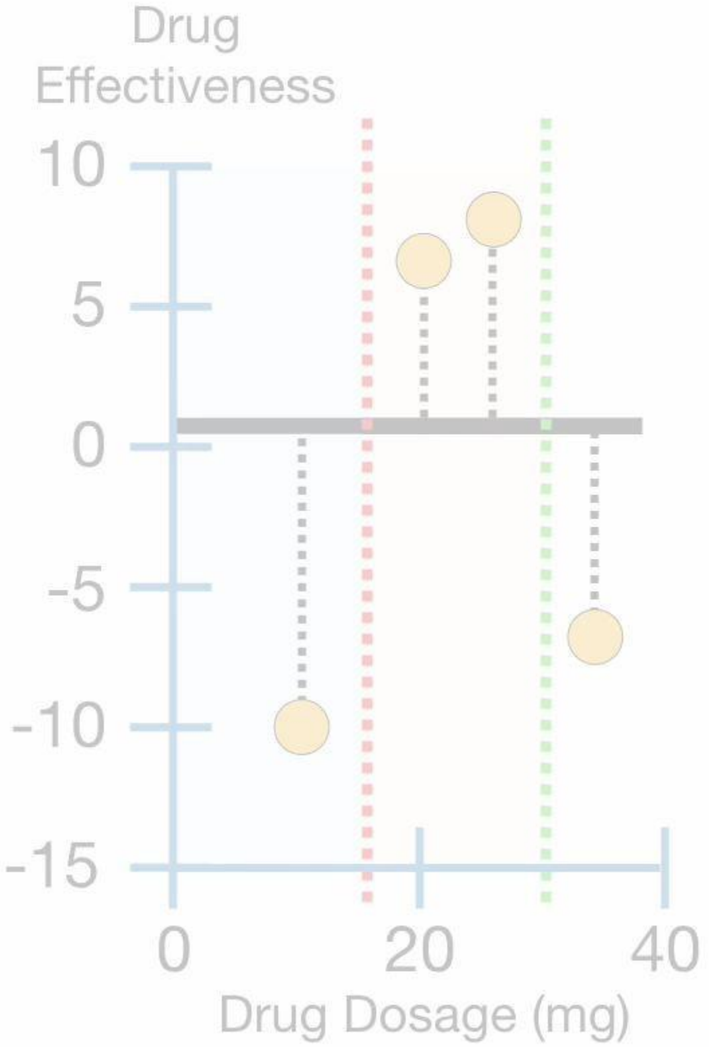
Gain = 120.33



NOTE: The **Gain** for the root, **120.3**, is less than **130**, the value for ***γ* (gamma)**, so the difference will be ***negative***.

Predicted Drug Effectiveness

0.5



Gain = 120.33

Dosage < 15

Gain = 140.17

-10.5

Dosage < 30

6.5, 7.5

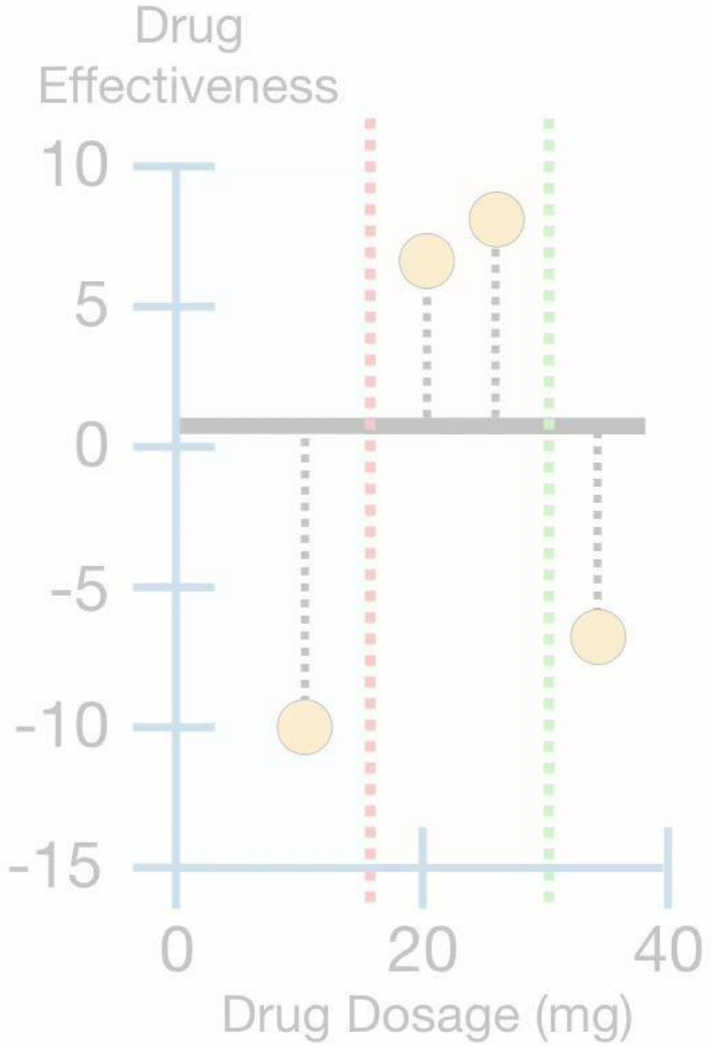
-7.5

NOTE: The **Gain** for the root, **120.3**, is less than **130**, the value for ***γ* (gamma)**, so the difference will be ***negative***.

However, because we did not remove the first branch, we will not remove the root.

Predicted Drug Effectiveness

0.5



Gain = 120.33

Dosage < 15

-10.5

Gain = 140.17

Dosage < 30

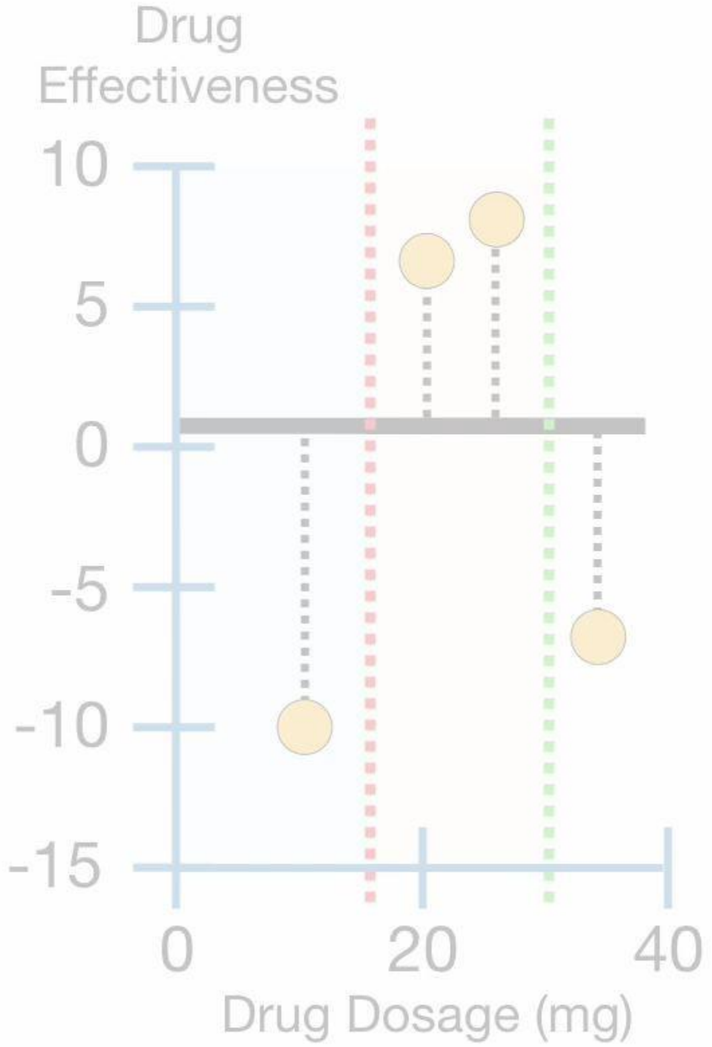
6.5, 7.5

-7.5

In contrast, if we set ***y* = 150**, then we would remove this branch because…

Predicted Drug Effectiveness

0.5



Gain = 120.33

Dosage < 15

-10.5

Gain = 140.17

Dosage < 30

6.5, 7.5

-7.5

In contrast, if we set ***γ*** = **150**, then we would remove this branch because…

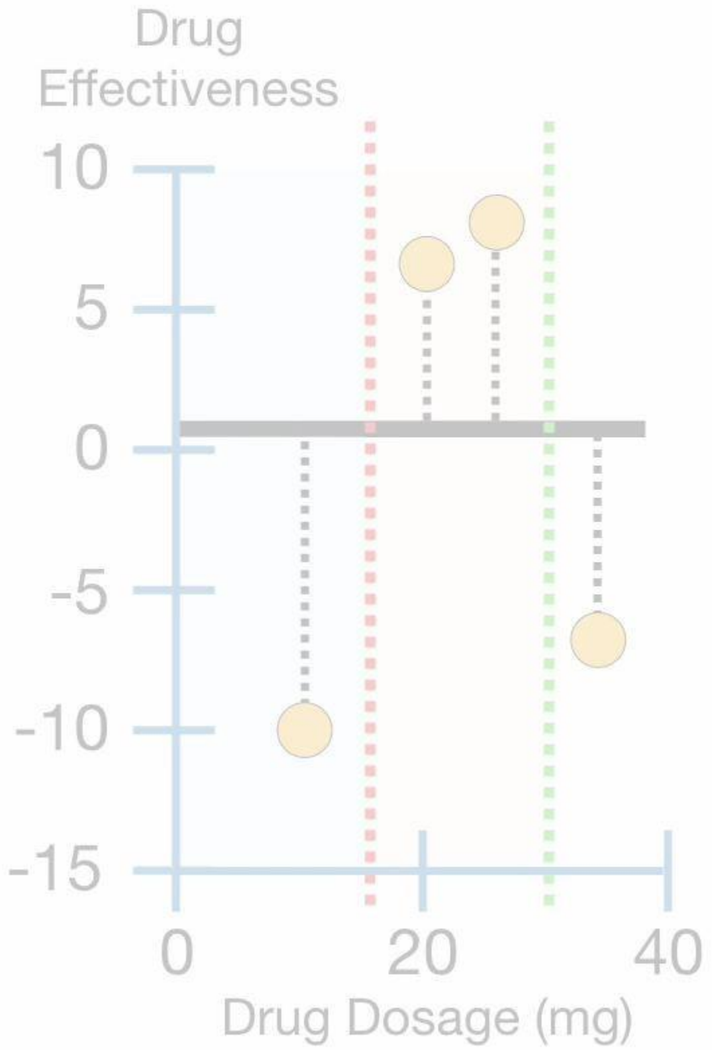
140.17 - 150 = a negative number.

Gain

***γ* (gamma)**

Predicted Drug Effectiveness

0.5



Gain = 120.33

Dosage < 15

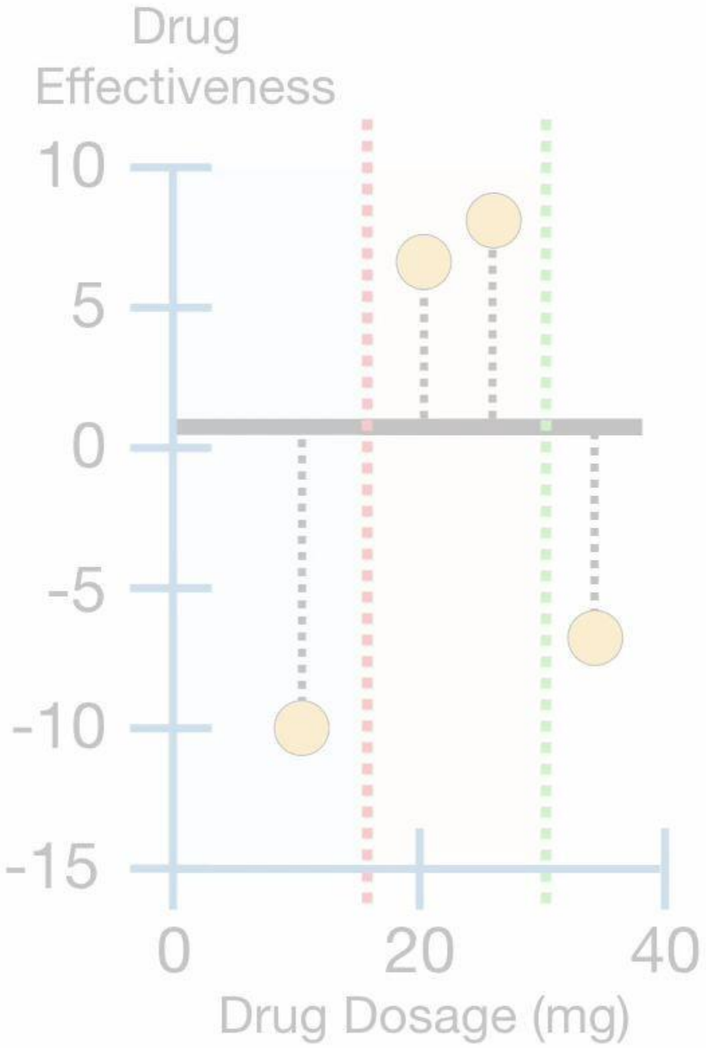
-10.5

6.5, 7.5, -7.5

Now we will subtract ***γ* (gamma)** from the **Gain** for the **Root**.

Predicted Drug Effectiveness

0.5



Gain = 120.33

Dosage < 15

-10.5

6.5, 7.5, -7.5

Since

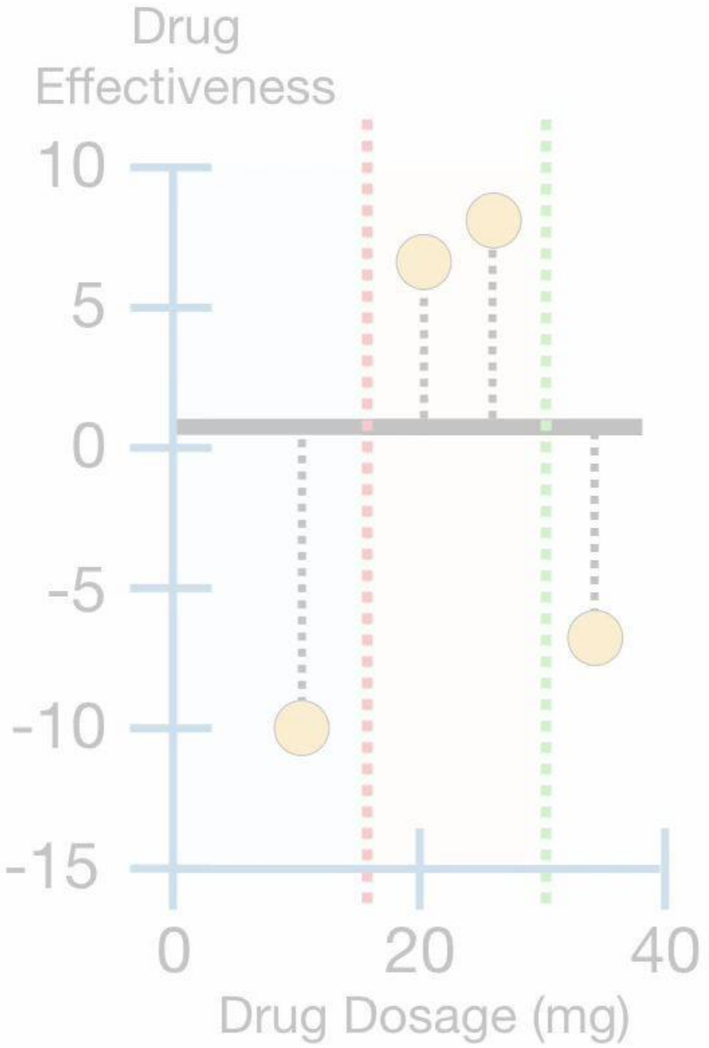
120.33 - 150 = a negative number…

Gain

***γ* (gamma)**

Predicted Drug Effectiveness

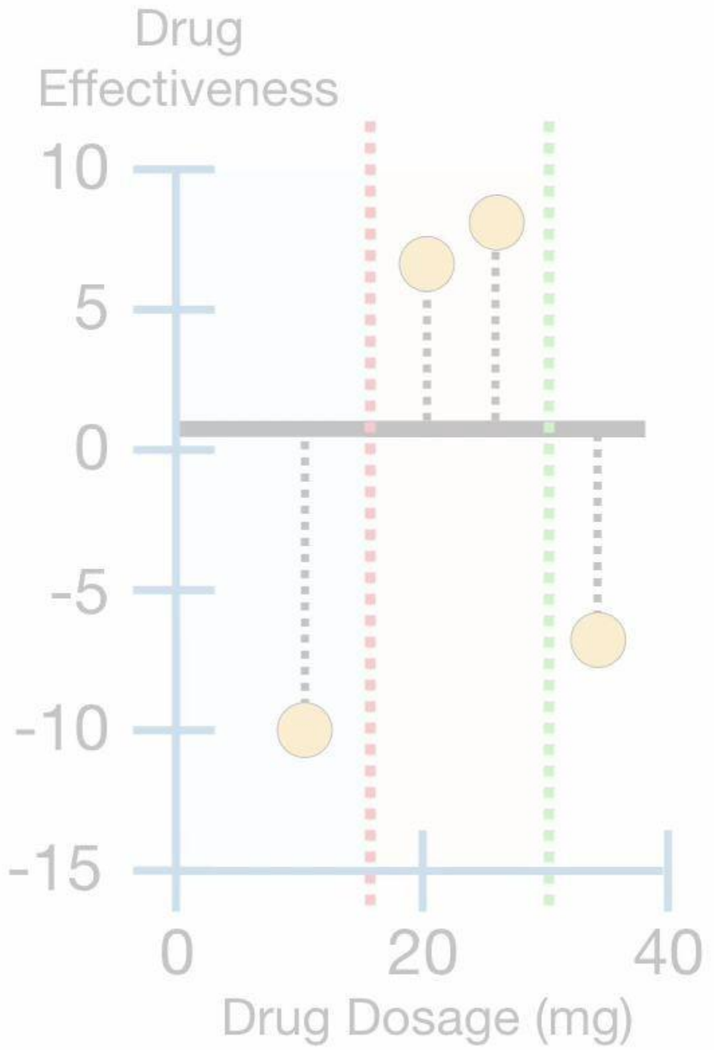
0.5



…and all we would be left with is the original prediction, which is pretty extreme pruning.

Predicted Drug Effectiveness

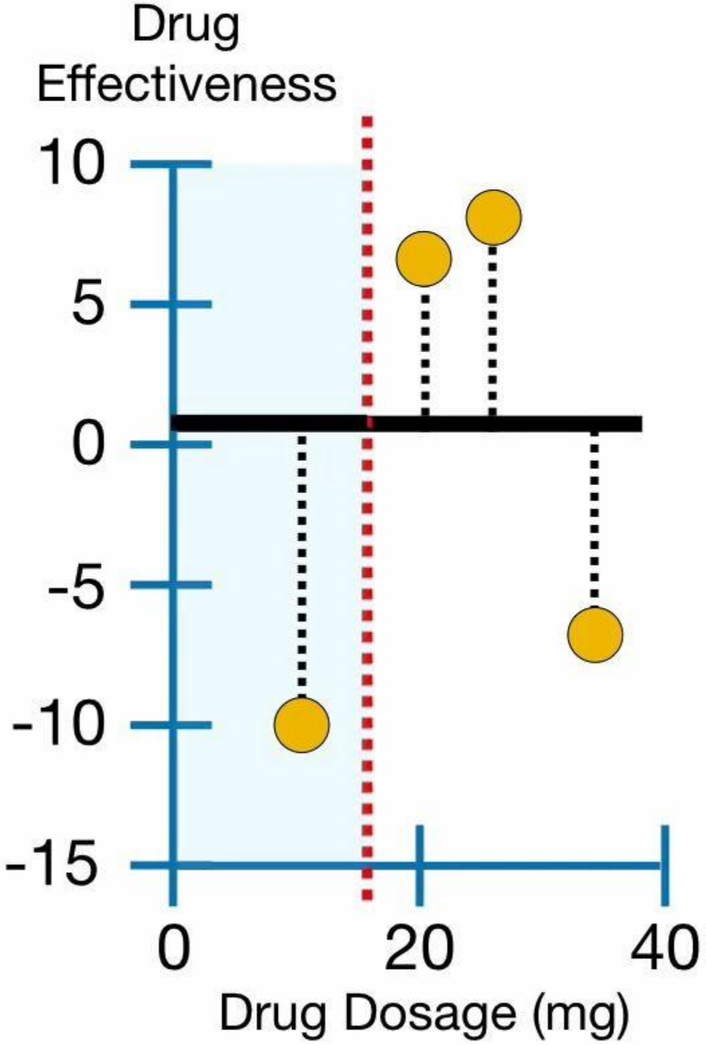
0.5



So, while this wasn't the most nuanced example of how an **XGBoost Tree** is pruned, I hope you get the idea.

Predicted Drug Effectiveness

0.5

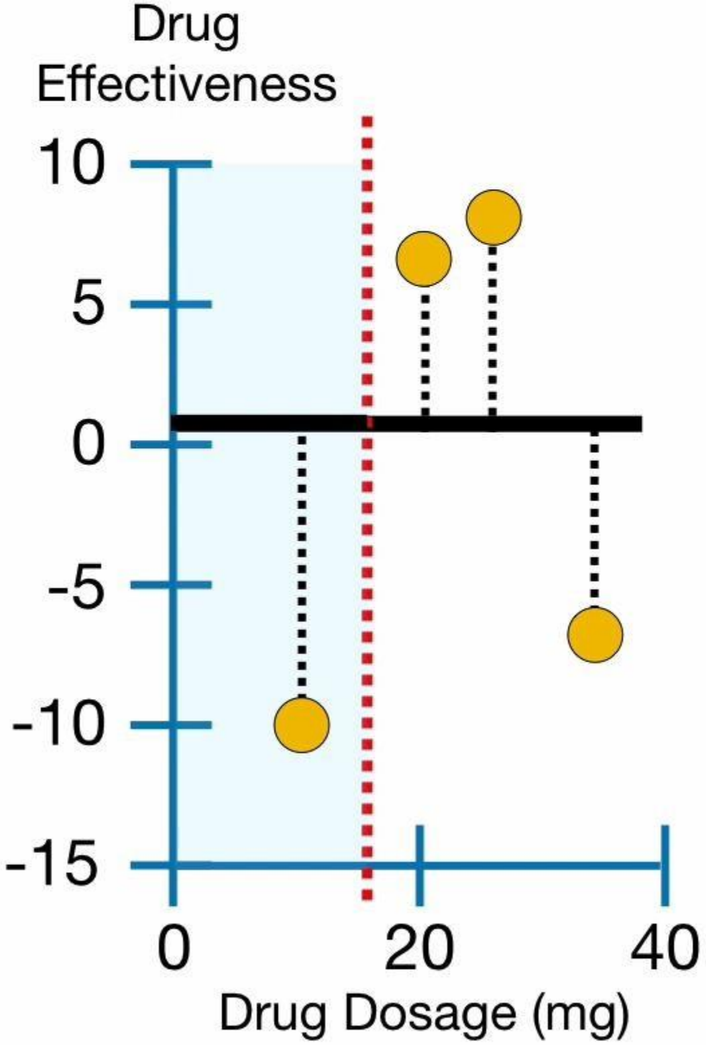


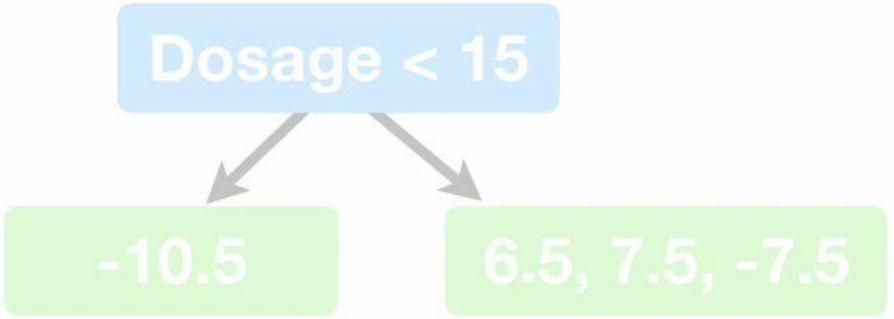
$$\text{Similarity Score} = \frac{\text{Sum of Residuals, Squared}}{\text{Number of Residuals} + \lambda}$$

…only this time, when we calculate **Similarity Scores**, we will set **λ (lambda) = 1**.

Predicted Drug Effectiveness

0.5



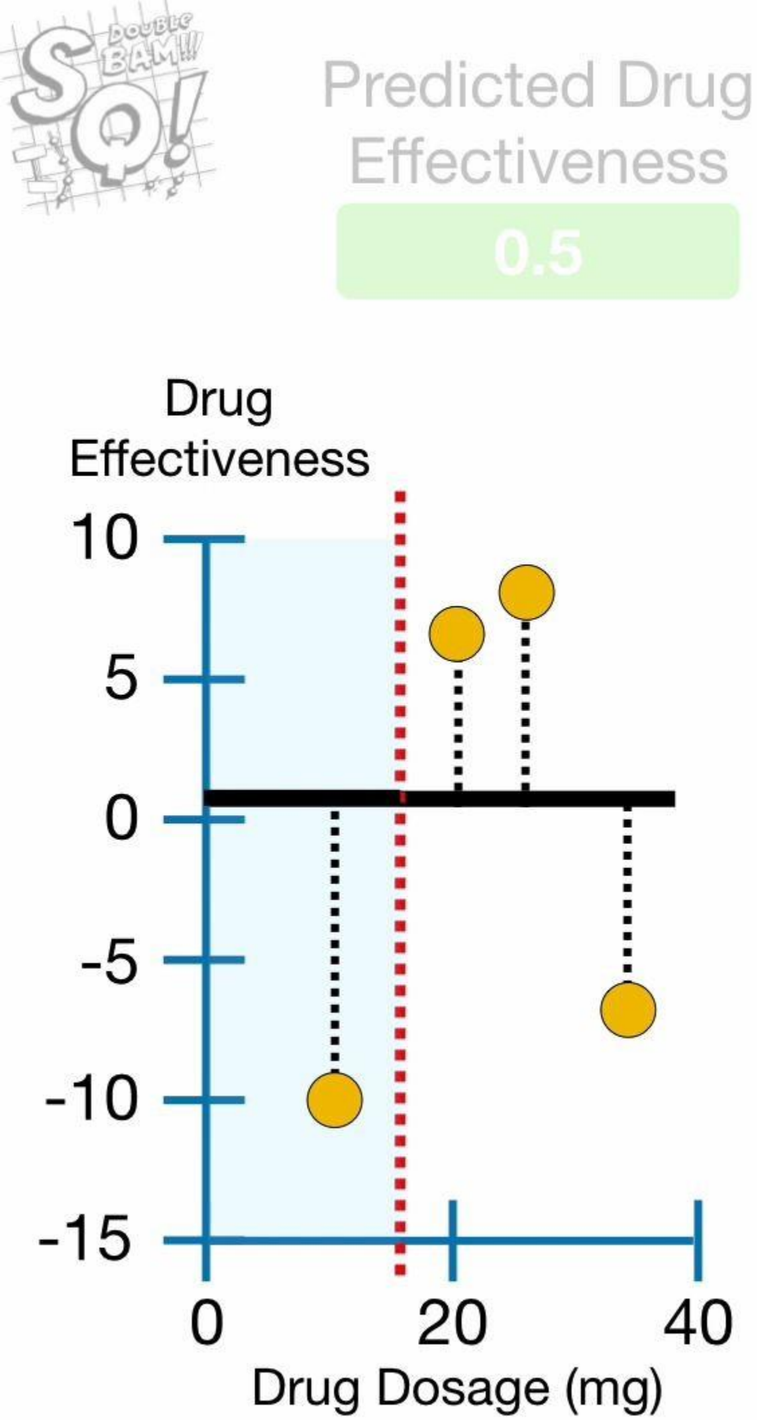


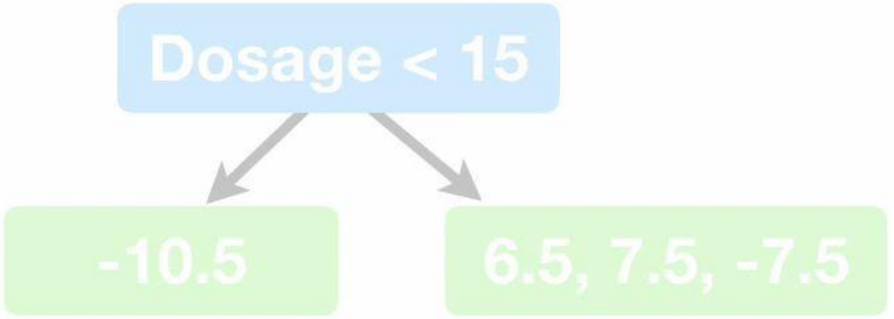
$$\text{Similarity Score} = \frac{\text{Sum of Residuals, Squared}}{\text{Number of Residuals} + \mathbf{1}}$$

…only this time, when we calculate **Similarity Scores**, we will set $\boldsymbol{\lambda}$ **(lambda) = 1**.

Predicted Drug Effectiveness

0.5



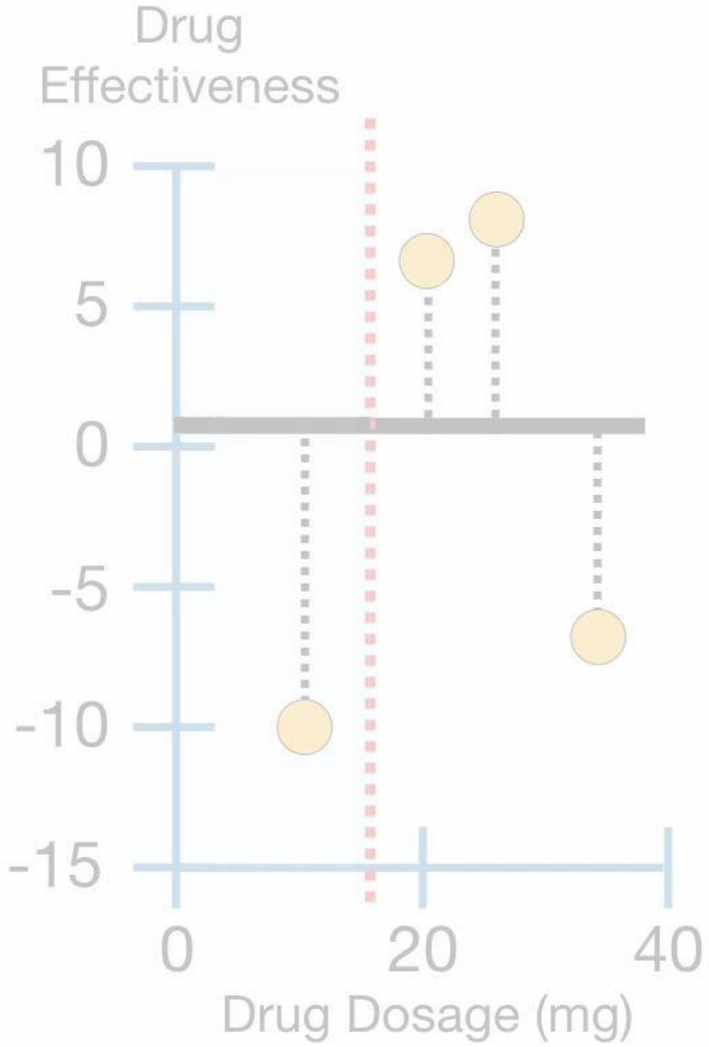


$$\text{Similarity Score} = \frac{\text{Sum of Residuals, Squared}}{\text{Number of Residuals} + \mathbf{1}}$$

Remember **λ (lambda)** is a **Regularization Parameter**, which means that it is intended to reduce the prediction's sensitivity to individual observations.

Predicted Drug Effectiveness

0.5



-10.5, 6.5, 7.5, -7.5

Similarity = 3.2

-10.5

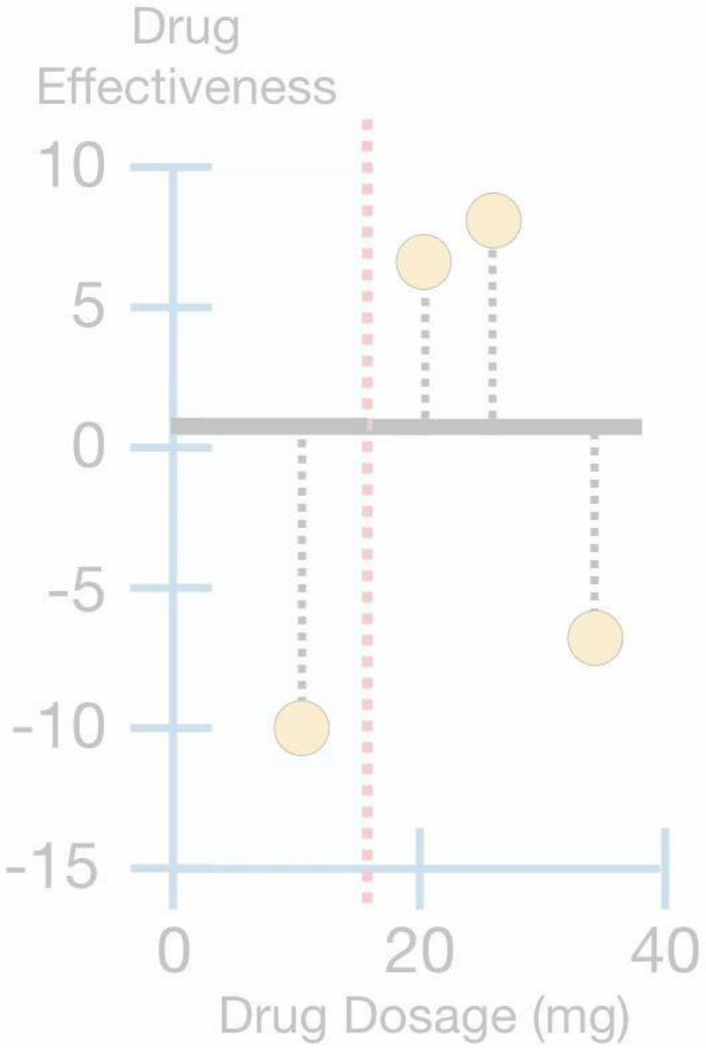
6.5, 7.5, -7.5

$$\text{Similarity Score} = \frac{(-10.5 + 6.5 + 7.5 + -7.5)^2}{4 + \mathbf{1}} = 3.2$$

…**3.2**, which is **8/10s** of what we got when $\boldsymbol{\lambda = 0}$.

Predicted Drug Effectiveness

0.5



Dosage < 15

Similarity = 3.2

-10.5

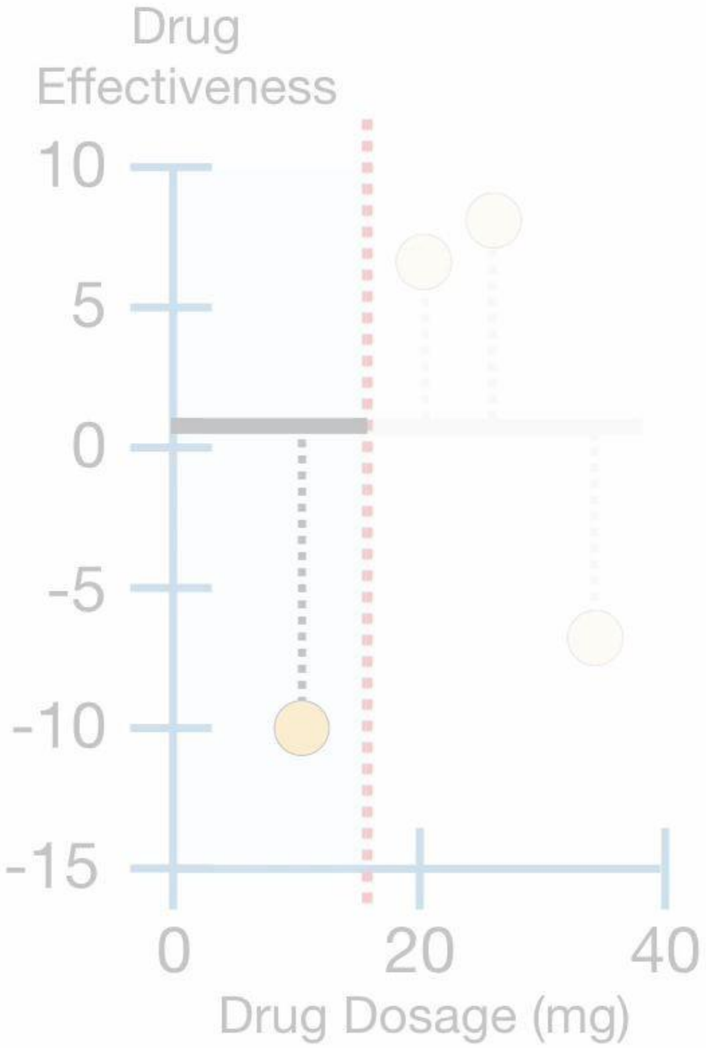
6.5, 7.5, -7.5

$$\text{Similarity Score} = \frac{\text{Sum of Residuals, Squared}}{\text{Number of Residuals} + 1}$$

When we calculate the **Similarity Score** for the leaf on the left…

Predicted Drug Effectiveness

0.5



Dosage < 15

Similarity = 3.2

-10.5

6.5, 7.5, -7.5

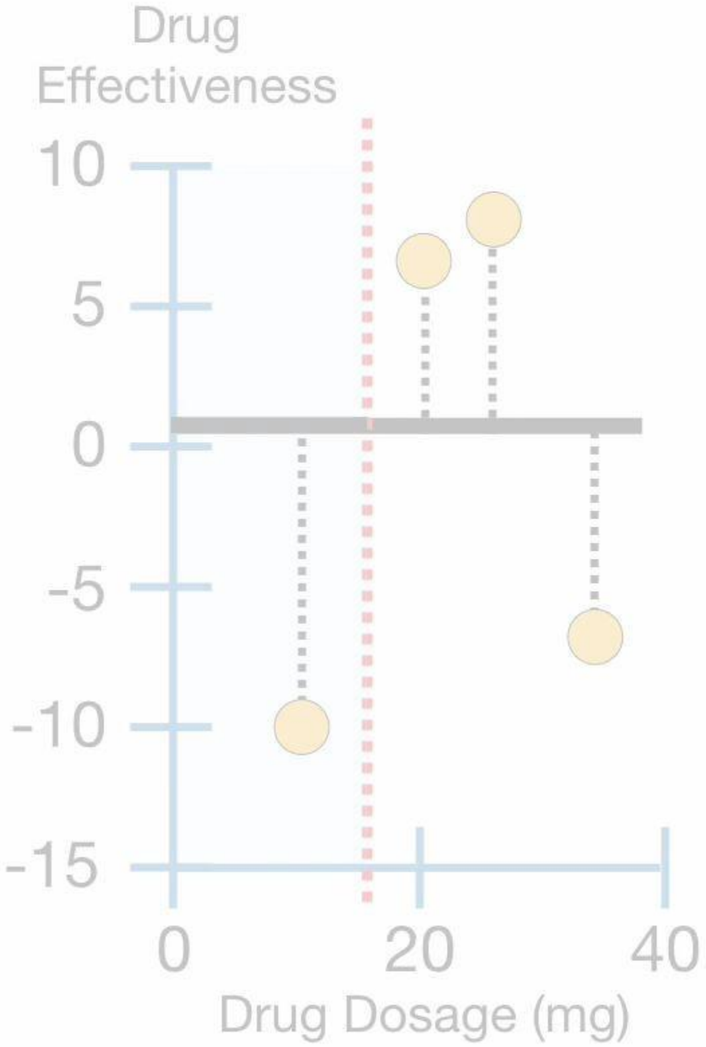
Similarity = 55.12

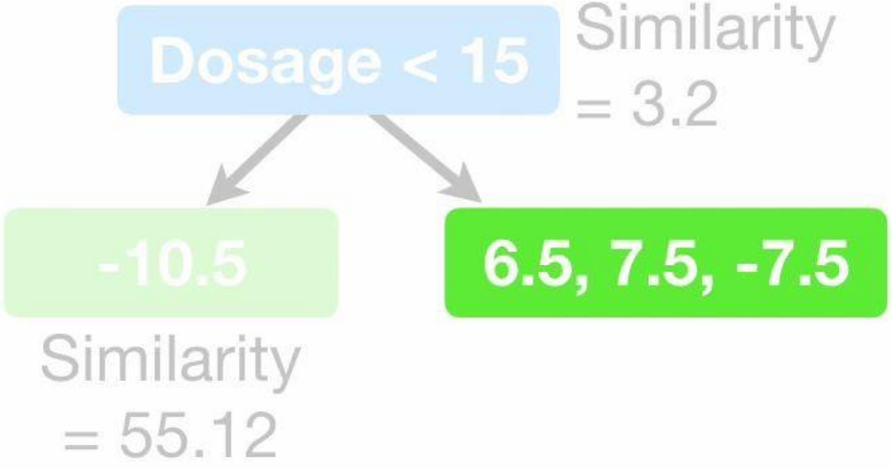
$$\text{Similarity Score} = \frac{-10.5^2}{1 + \mathbf{1}} = 55.12$$

…we get **55.12**, which is half of what we got when $\boldsymbol{\lambda = 0}$.

Predicted Drug Effectiveness

0.5



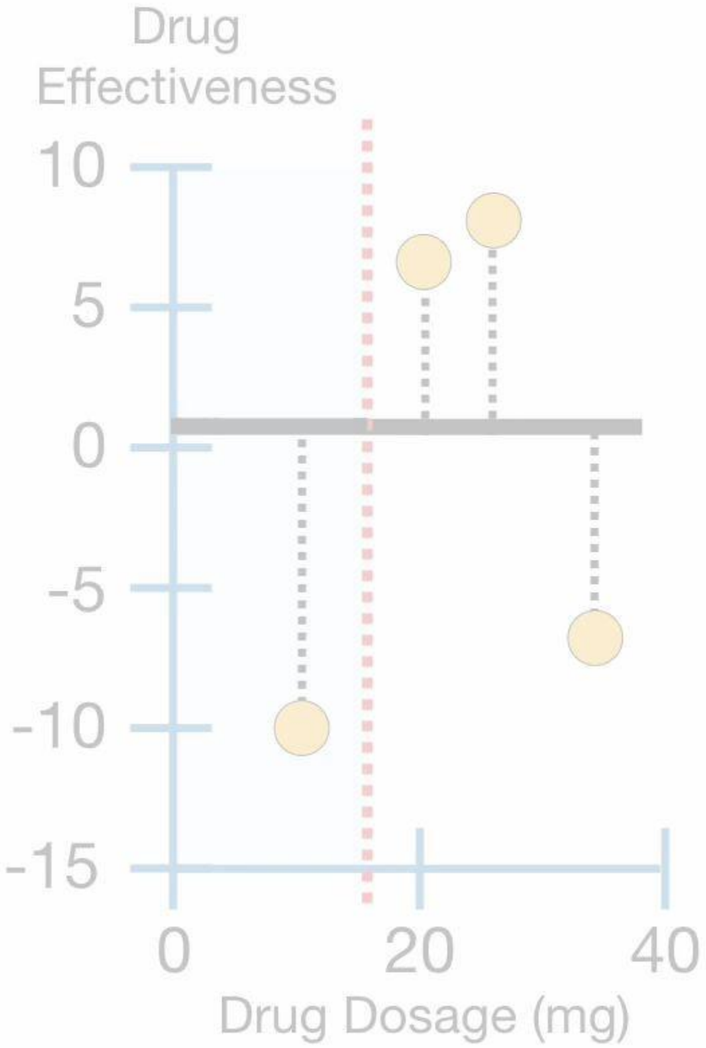


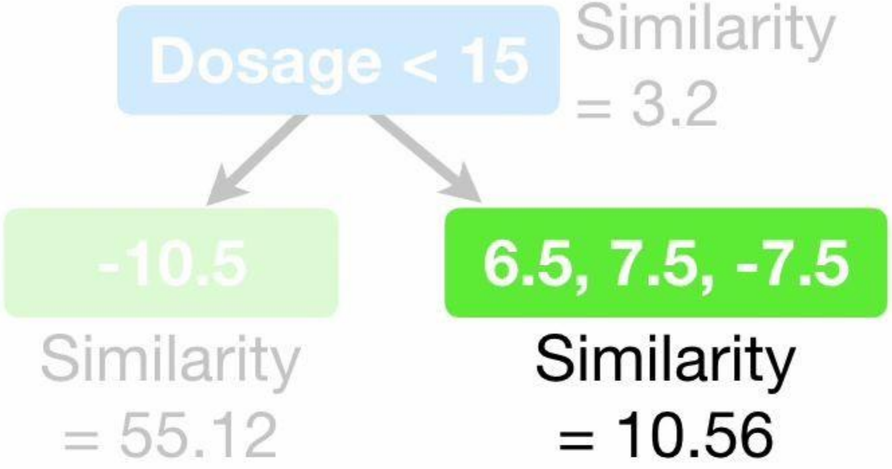
$$\text{Similarity Score} = \frac{(6.5 + 7.5 + -7.5)^2}{3 + 1}$$

And when we calculate the **Similarity Score** for the leaf on the right…

Predicted Drug Effectiveness

0.5



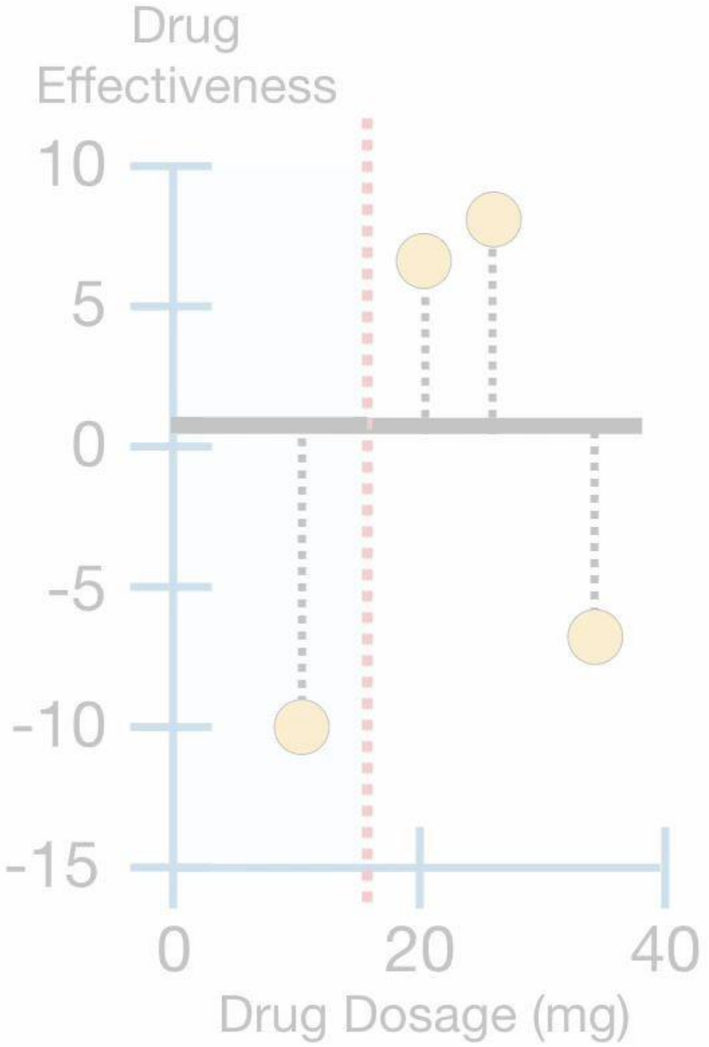


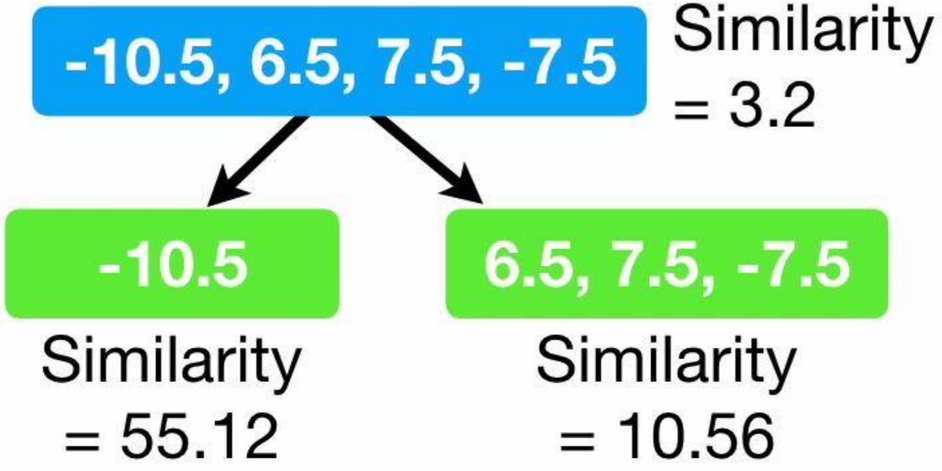
$$\text{Similarity Score} = \frac{(6.5 + 7.5 + -7.5)^2}{3 + \mathbf{1}} = 10.56$$

…we get **10.56**, which is **3/4s** of what we got when $\boldsymbol{\lambda = 0}$.

Predicted Drug Effectiveness

0.5

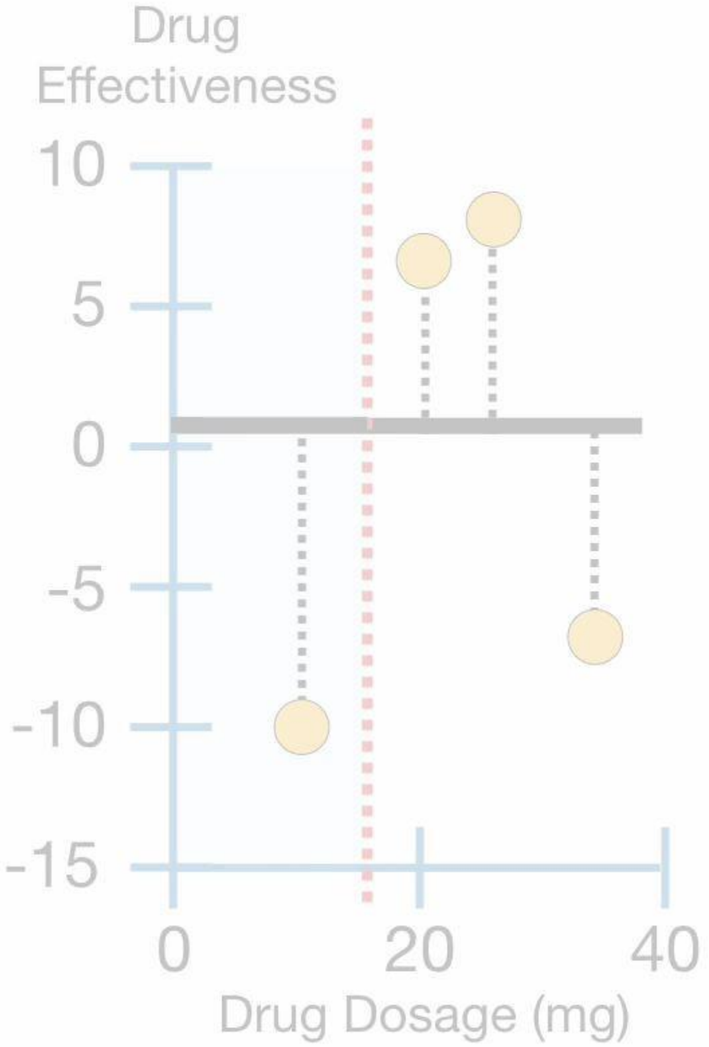


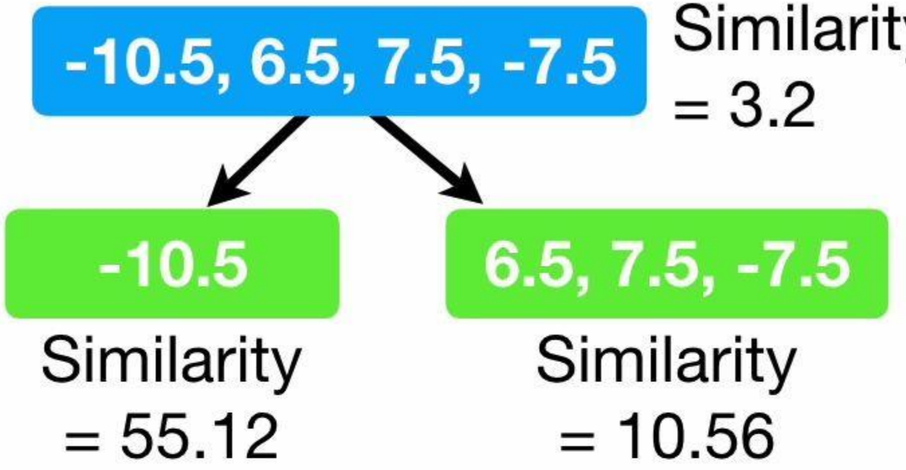


So, one thing we see is that when **λ > 0**, the **Similarity Scores** are smaller…

Predicted Drug Effectiveness

0.5

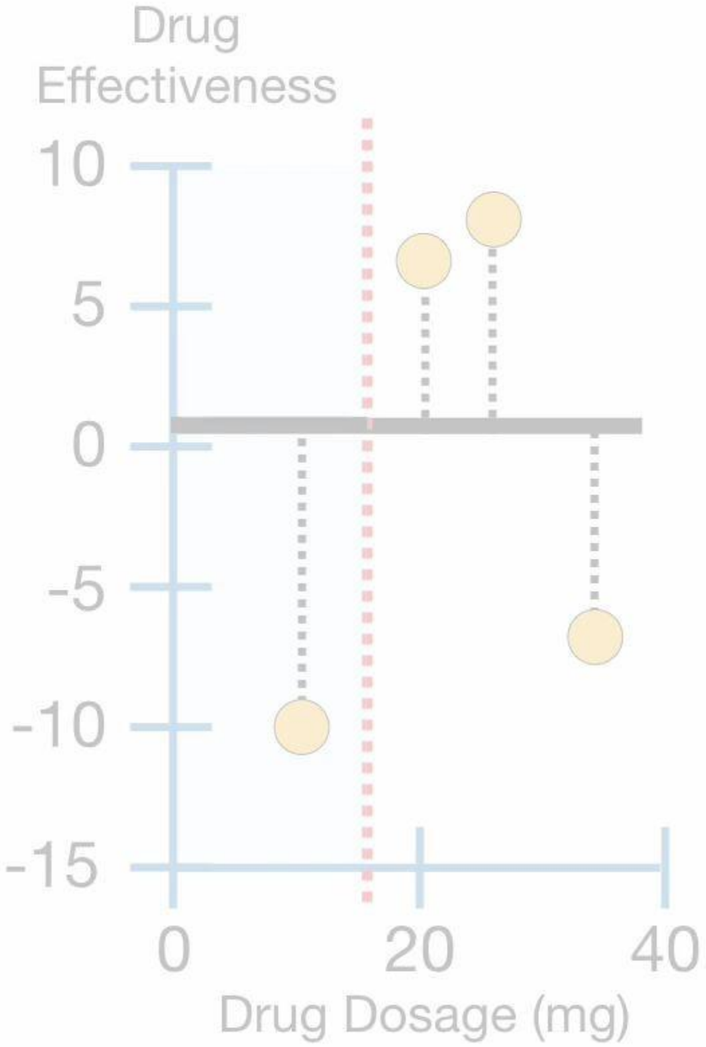




…and the amount of decrease is **inversely proportional** to the number of **Residuals** in the node.

Predicted Drug Effectiveness

0.5



-10.5, 6.5, 7.5, -7.5

Similarity = 3.2

-10.5

Similarity = 55.12

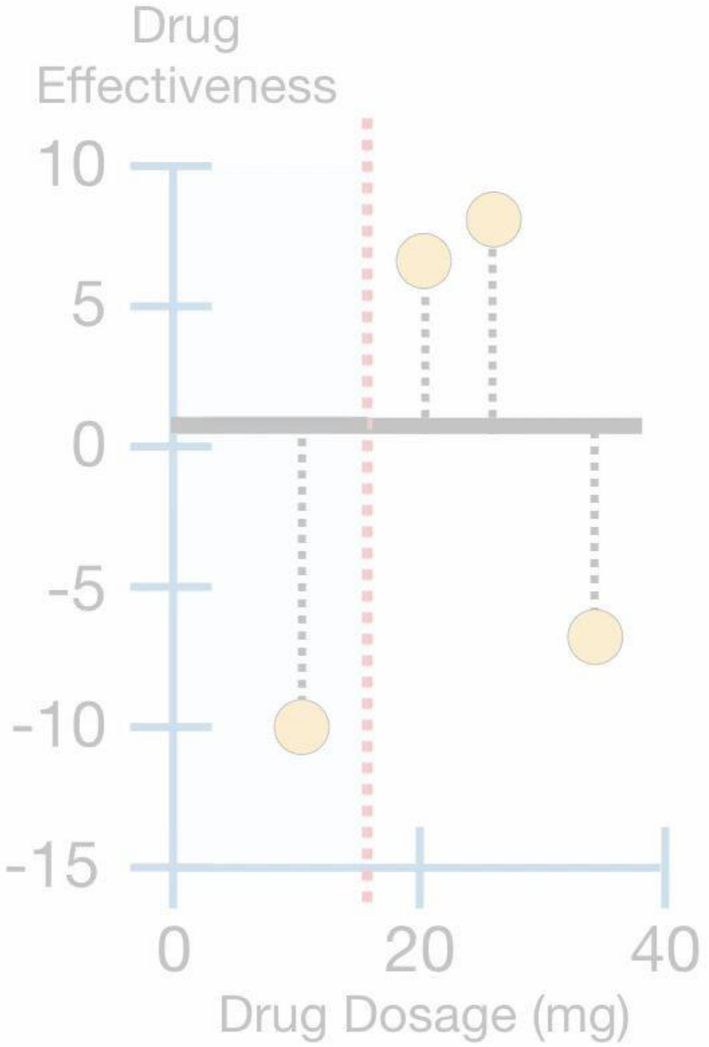
6.5, 7.5, -7.5

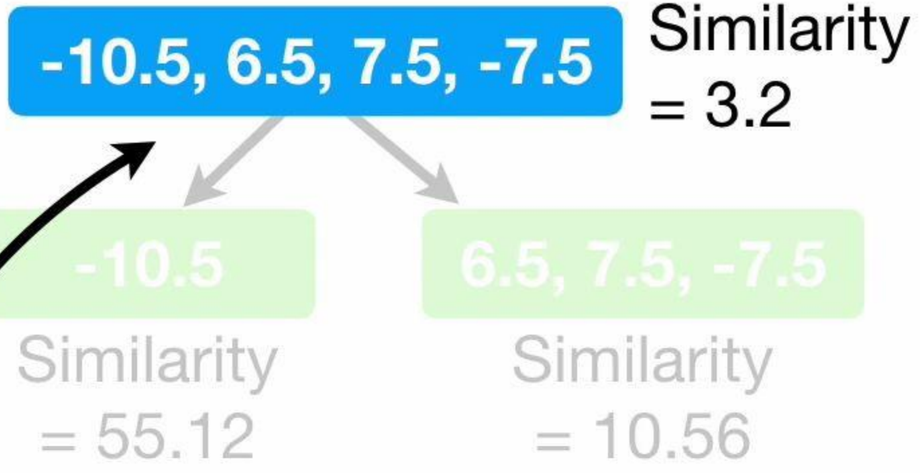
Similarity = 10.56

In other words, the leaf on the left had only **1 Residual**, and it had the largest decrease in **Similarity Score**, **50%**.

Predicted Drug Effectiveness

0.5

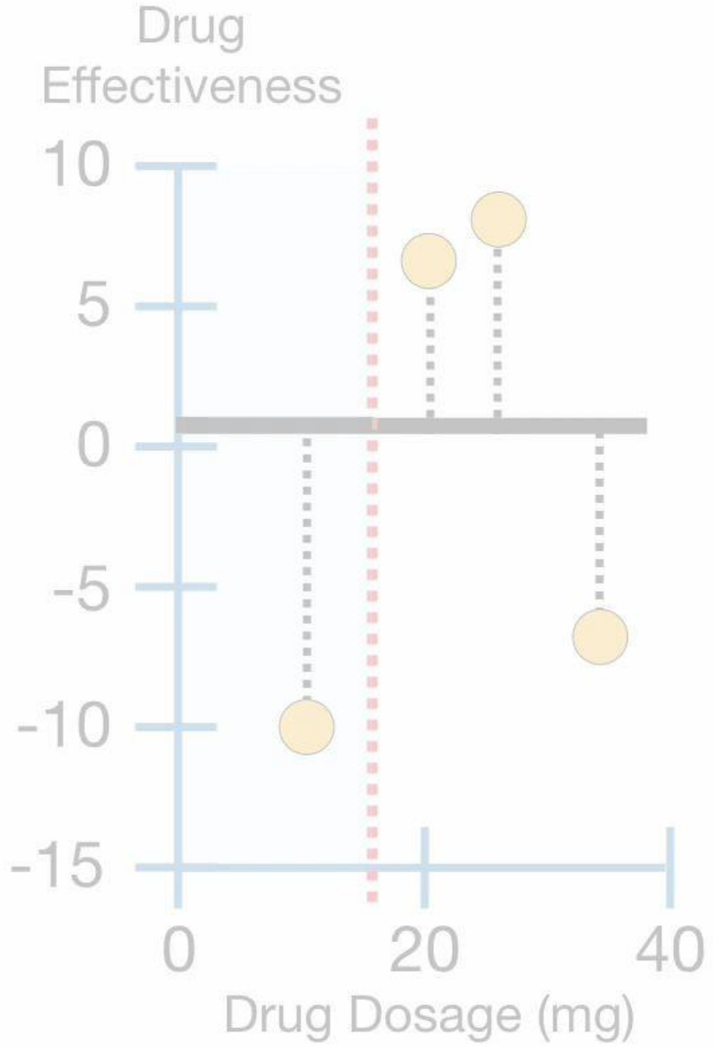


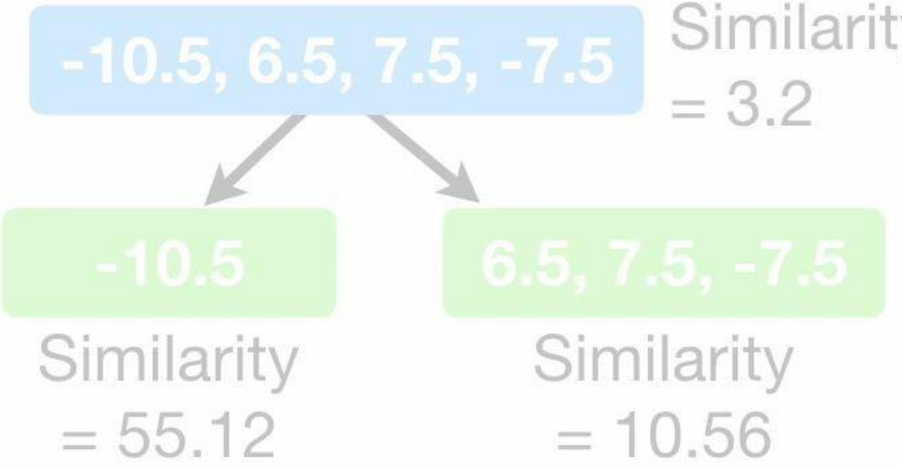


In contrast, the root had all **4 Residuals** and the smallest decrease, **20%**.

Predicted Drug Effectiveness

0.5



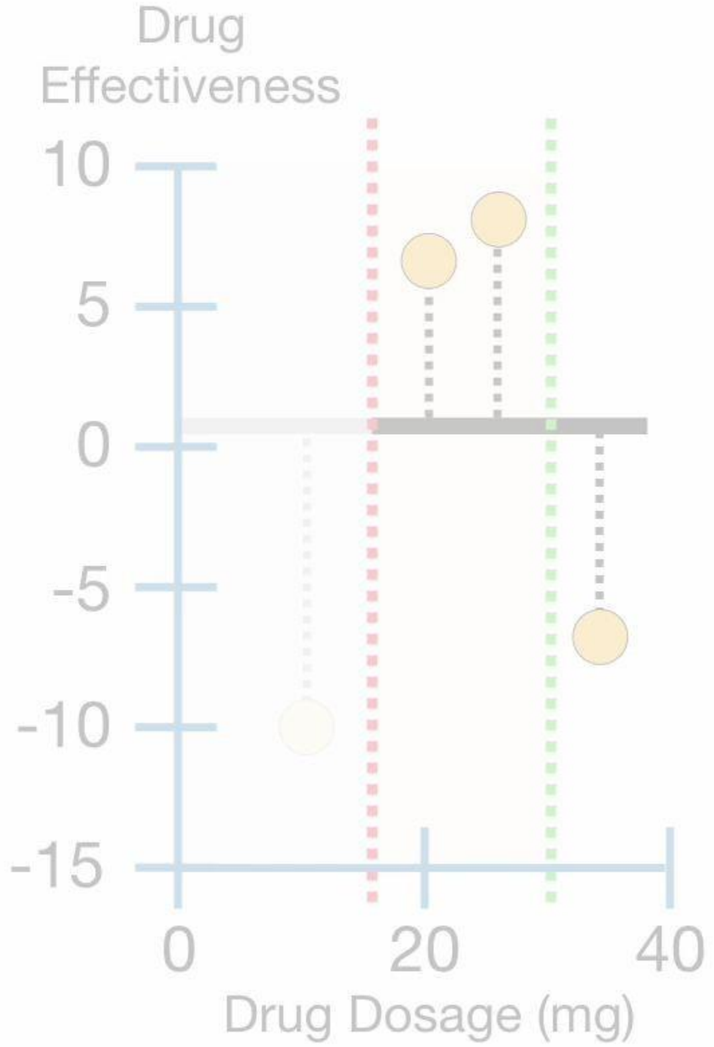


Gain = 55.12 + 10.56 - 3.2 = 62.48

…we get **66**, which is a lot less than **120.33**, the value we got when $\lambda = 0$.

Predicted Drug Effectiveness

0.5



Gain = 62.49 **Dosage < 15**

10.5

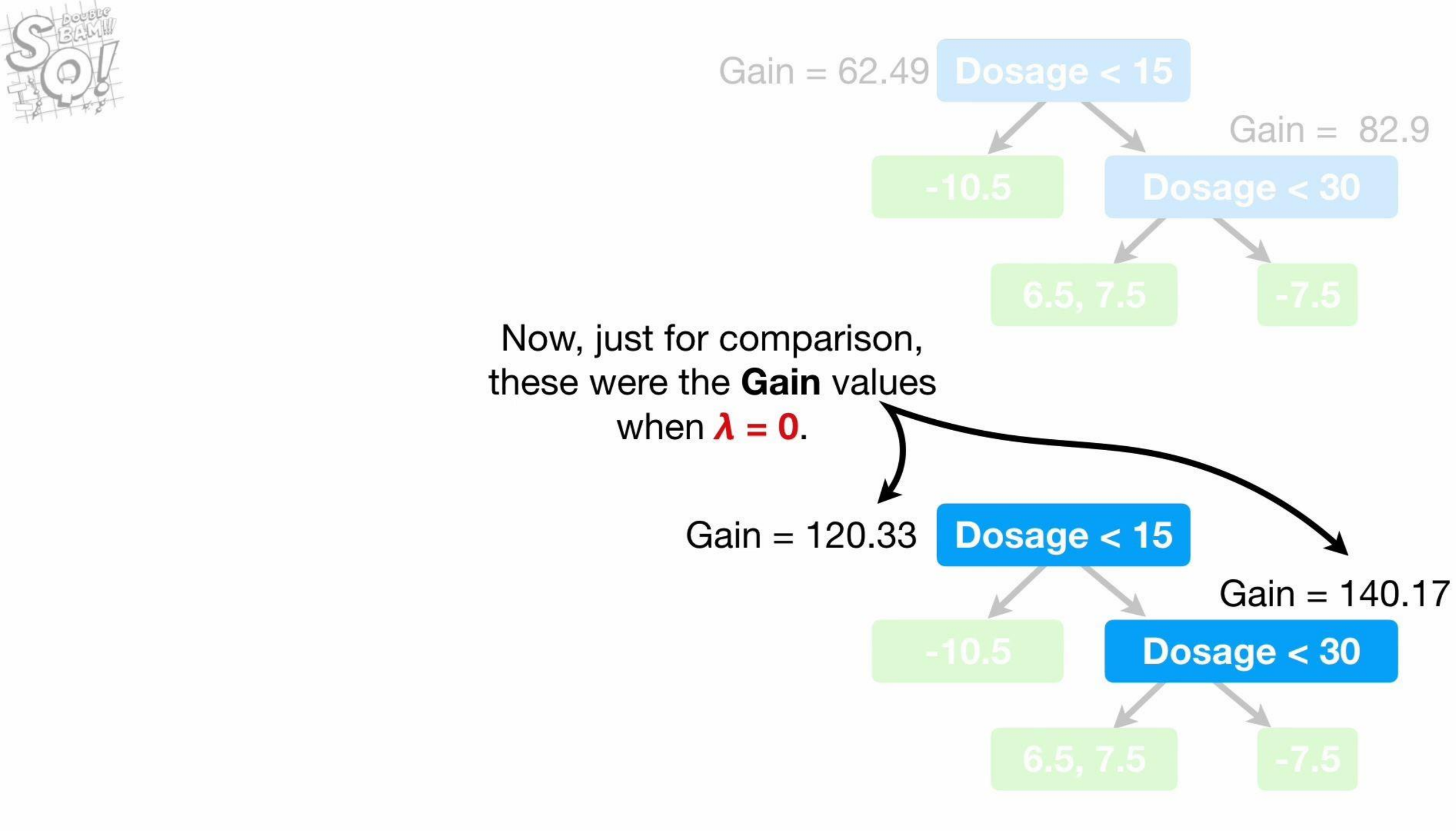
Dosage < 30

Gain = 82.9

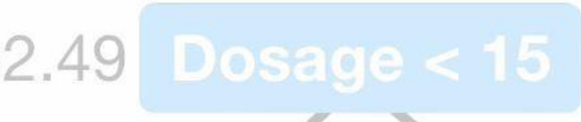
6.5, 7.5

-7.5

Similarly, when $\lambda = 1$, the **Gain** for the next branch is smaller than before.


Gain = 62.49
Dosage < 15
Gain = 82.9
-10.5
Dosage < 30
6.5, 7.5
-7.5
Now, just for comparison, these were the **Gain** values when **λ = 0**.
Gain = 120.33
Dosage < 15
Gain = 140.17
-10.5
Dosage < 30
6.5, 7.5
-7.5

Gain = 62.49

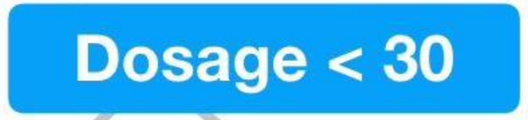


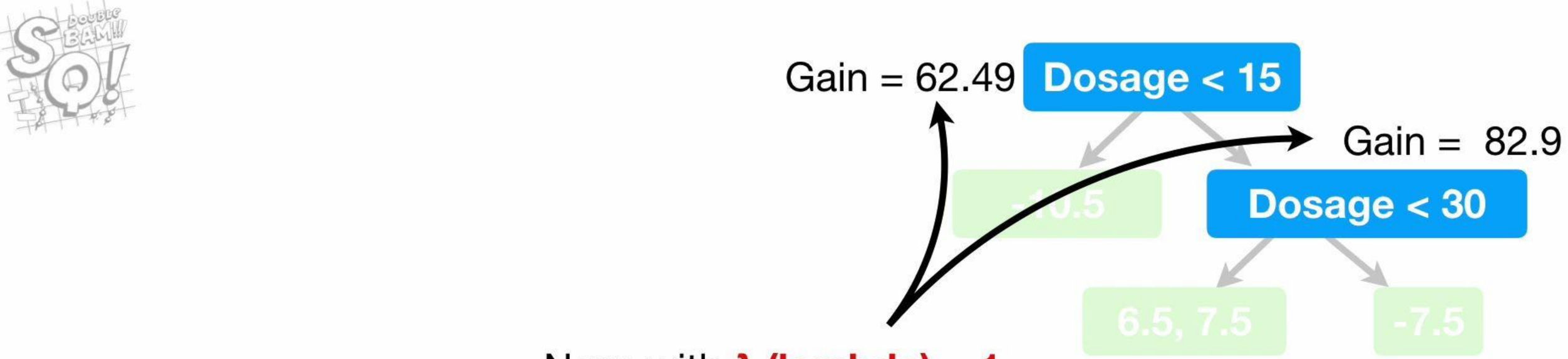
Gain = 82.9

…and because, for the lowest branch in the first tree, **Gain - *y* = a positive number**, we did not prune at all.

Gain = 120.33

Gain = 140.17

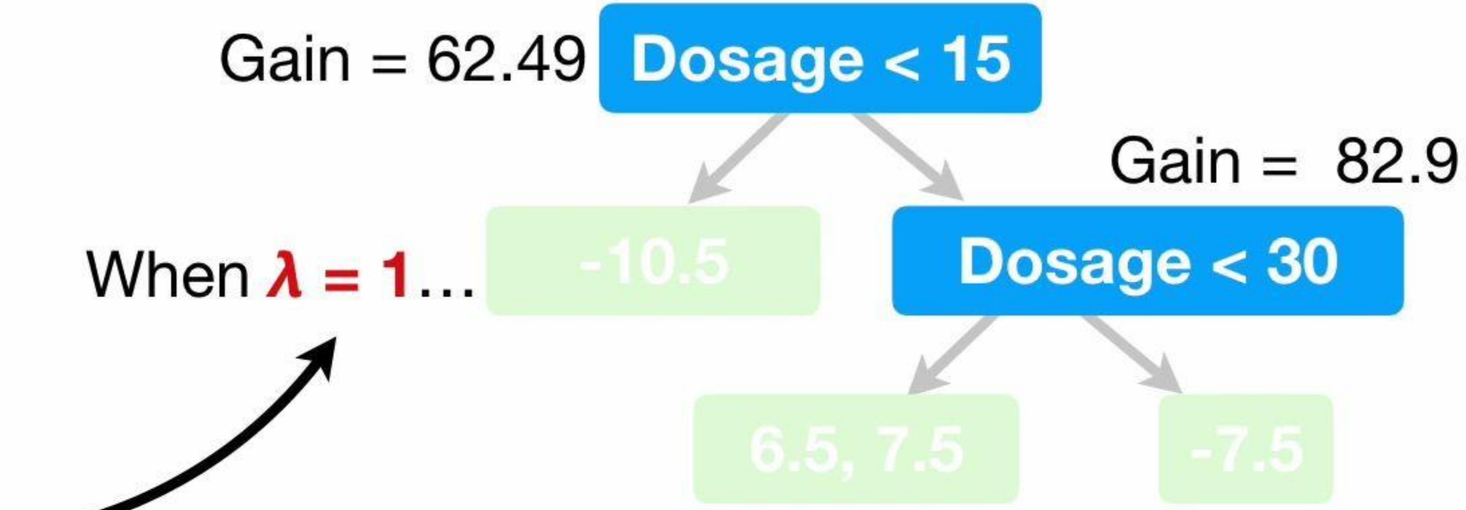




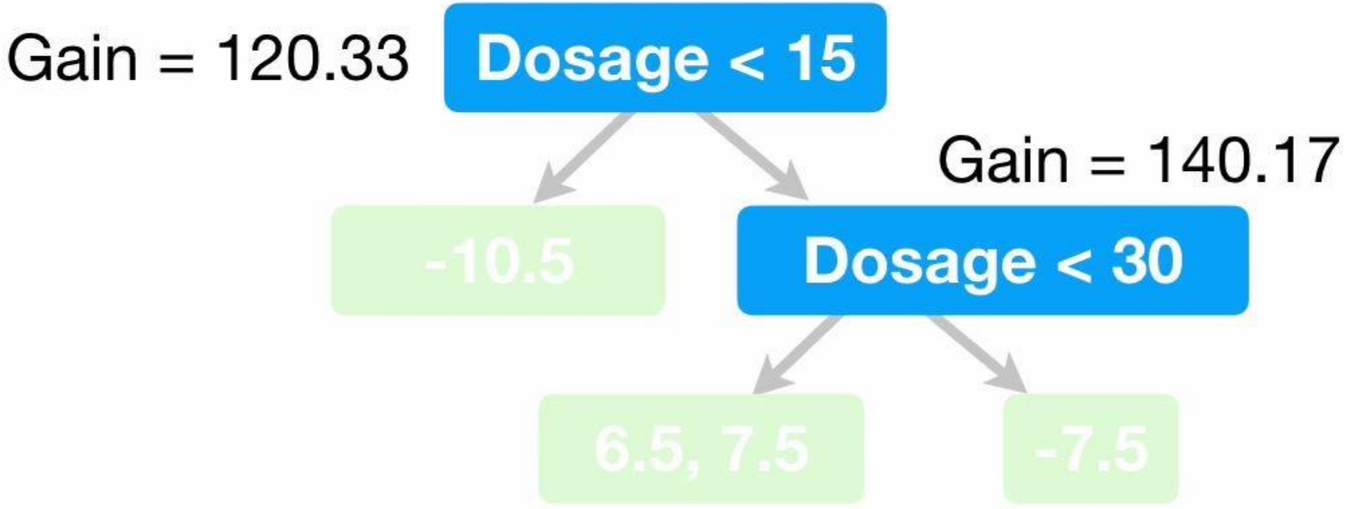
Now, with **λ (lambda) = 1**,
the values for **Gain** are
both **< 130**…

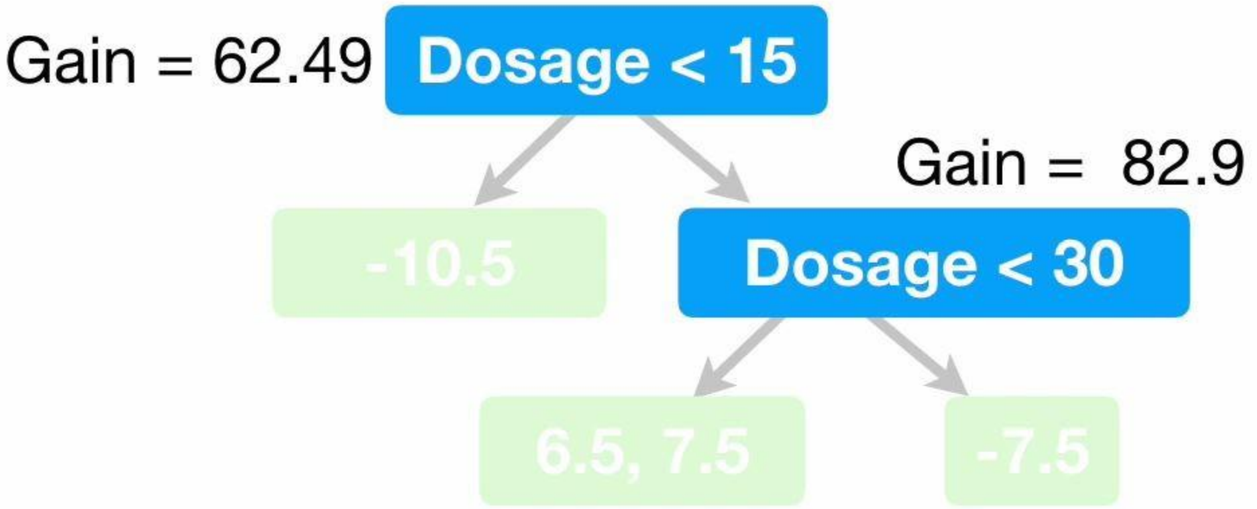
So when $\lambda > 0$, it is easier to prune leaves because the values for **Gain** are smaller.

When $\lambda = 1$…



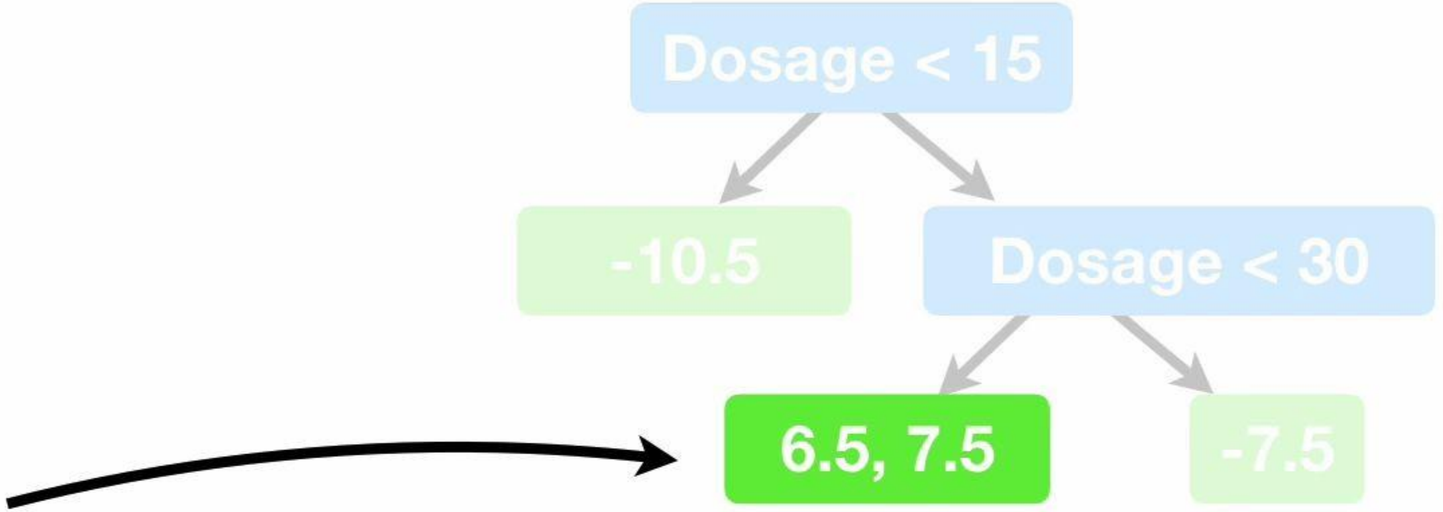
When $\lambda = 0$…

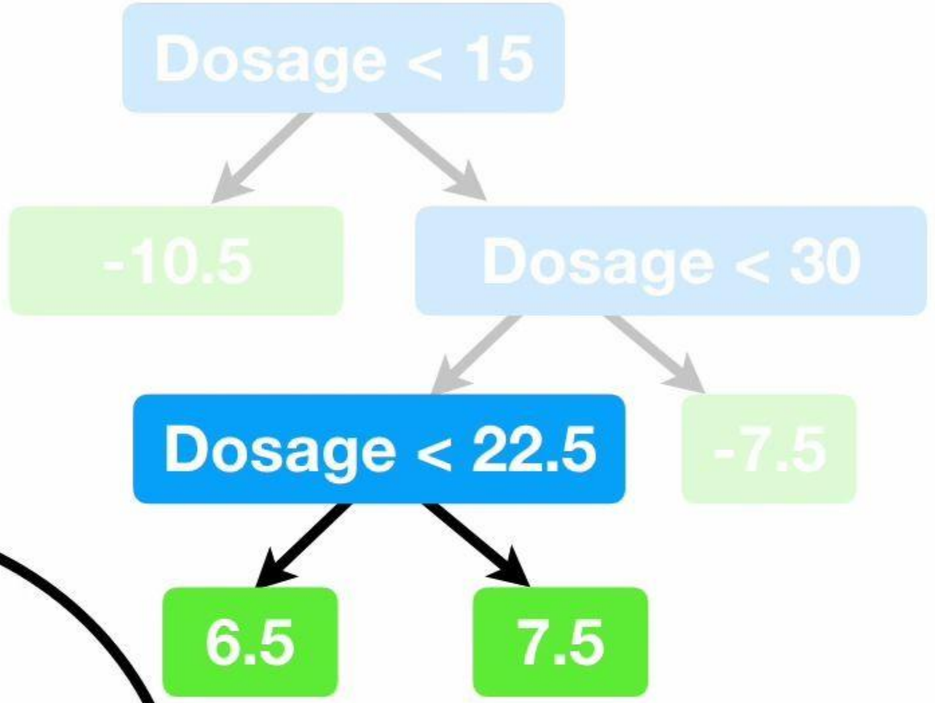




NOTE: Before we move on, I want illustrate one last feature of ***λ* (lambda)**.

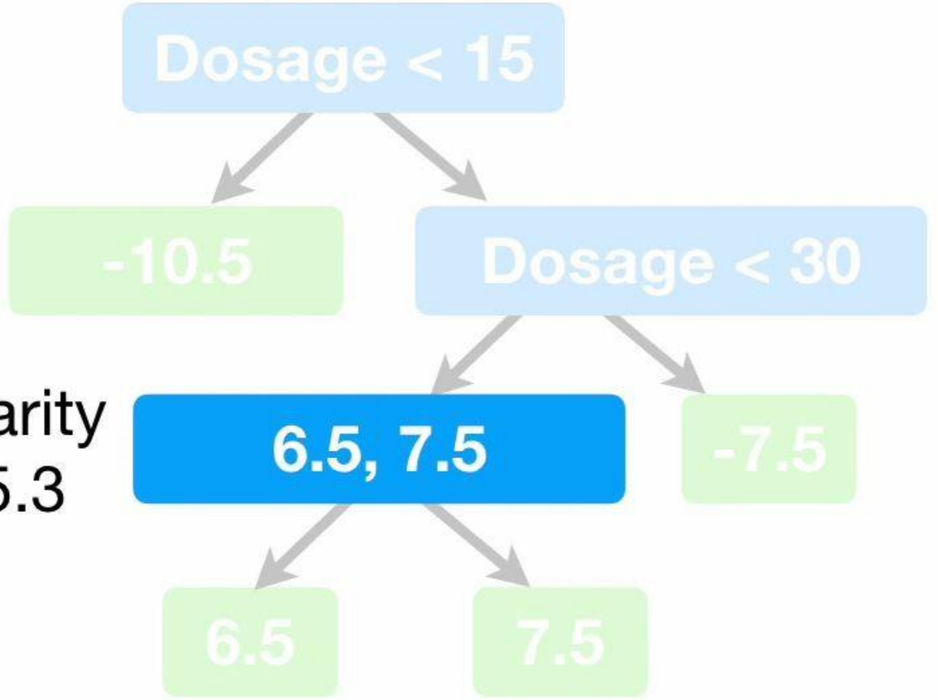
For this example, imagine we split this node into two leaves.





Now let's calculate the **Similarity Scores** with **λ (lambda) = 1**.

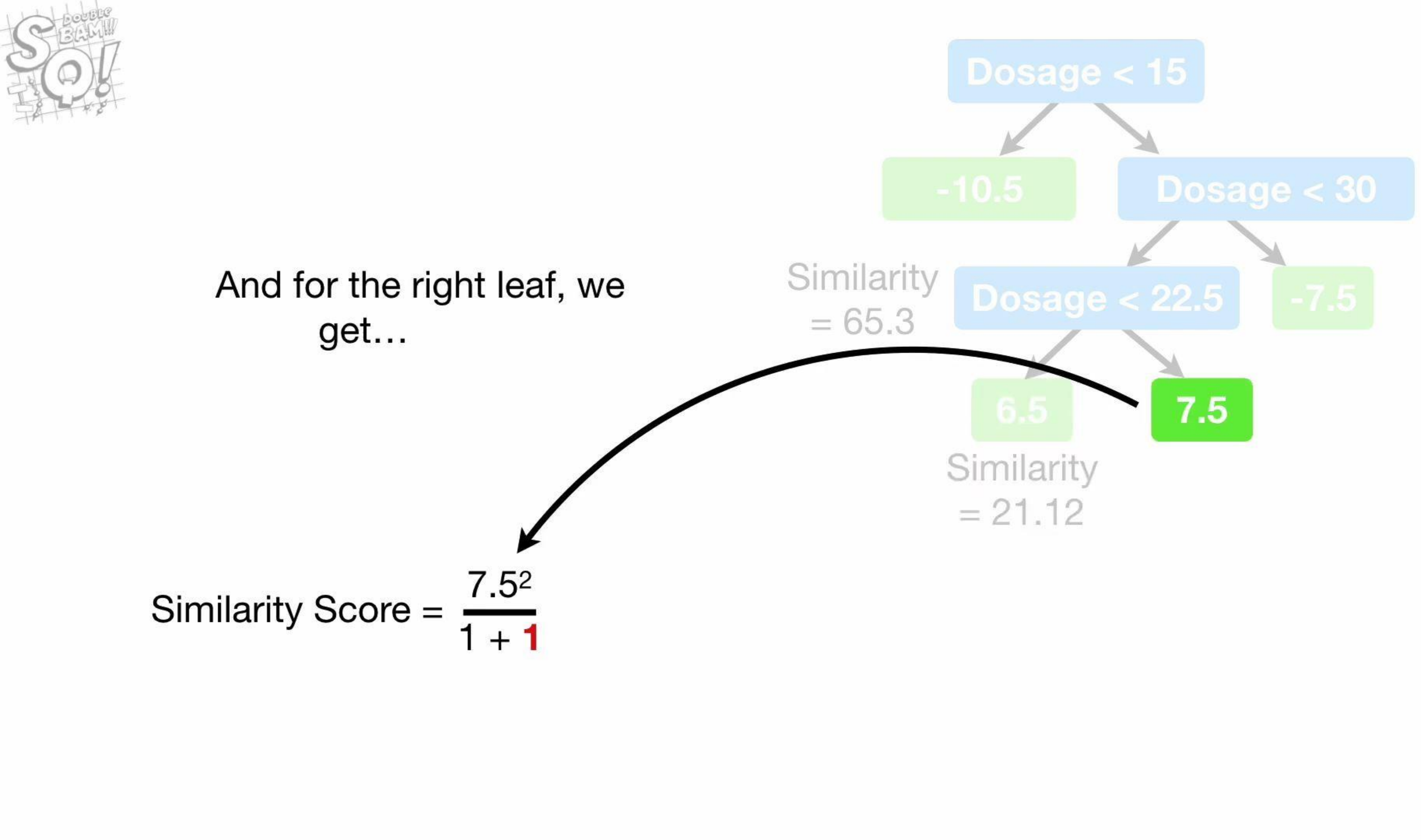
$$\text{Similarity Score} = \frac{\text{Sum of Residuals, Squared}}{\text{Number of Residuals} + \lambda}$$



Similarity = 65.3

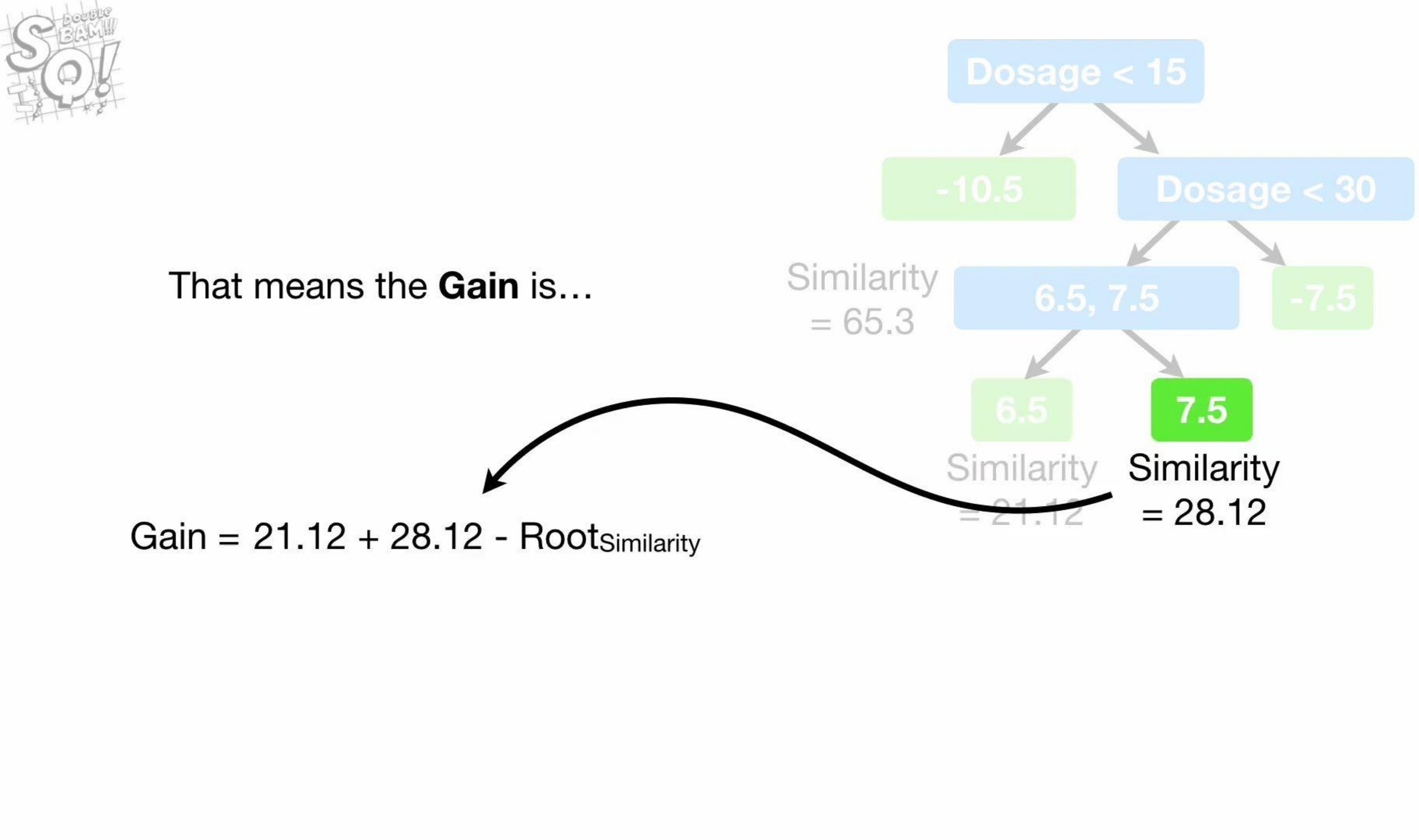
For the branch, we get…**65.3**.

$$\text{Similarity Score} = \frac{(6.5 + 7.5)^2}{2 + 1} = 65.3$$



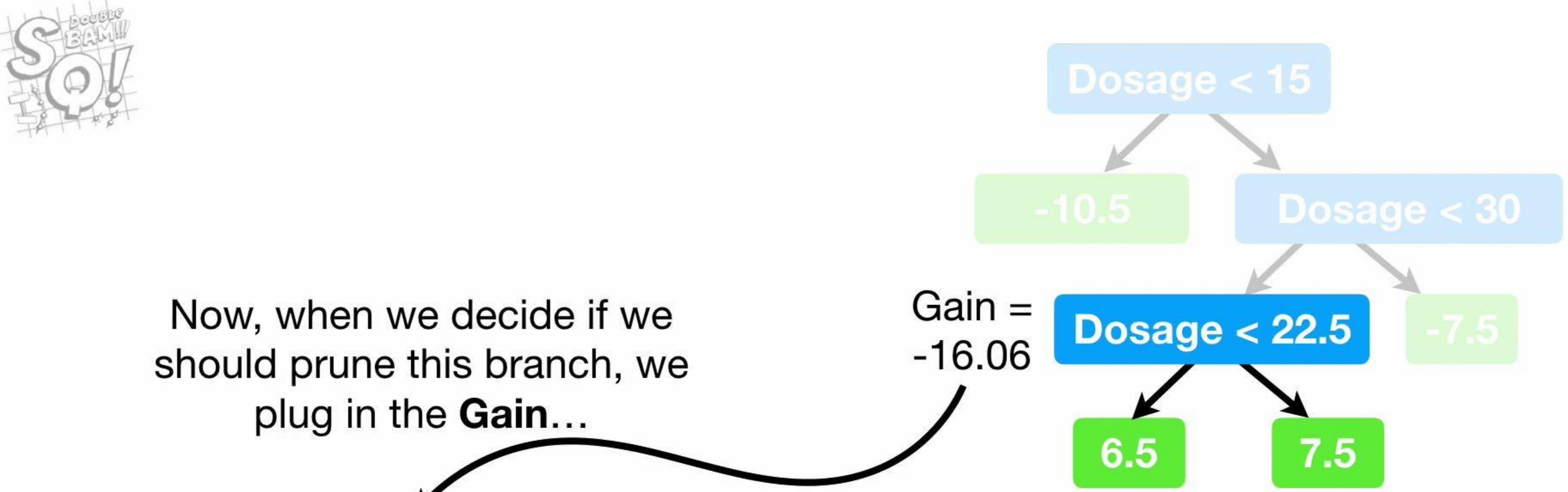
And for the right leaf, we get…

$$\text{Similarity Score} = \frac{7.5^2}{1 + 1}$$



That means the **Gain** is…

$$\text{Gain} = 21.12 + 28.12 - \text{Root}_{\text{Similarity}}$$



Now, when we decide if we should prune this branch, we plug in the **Gain**…

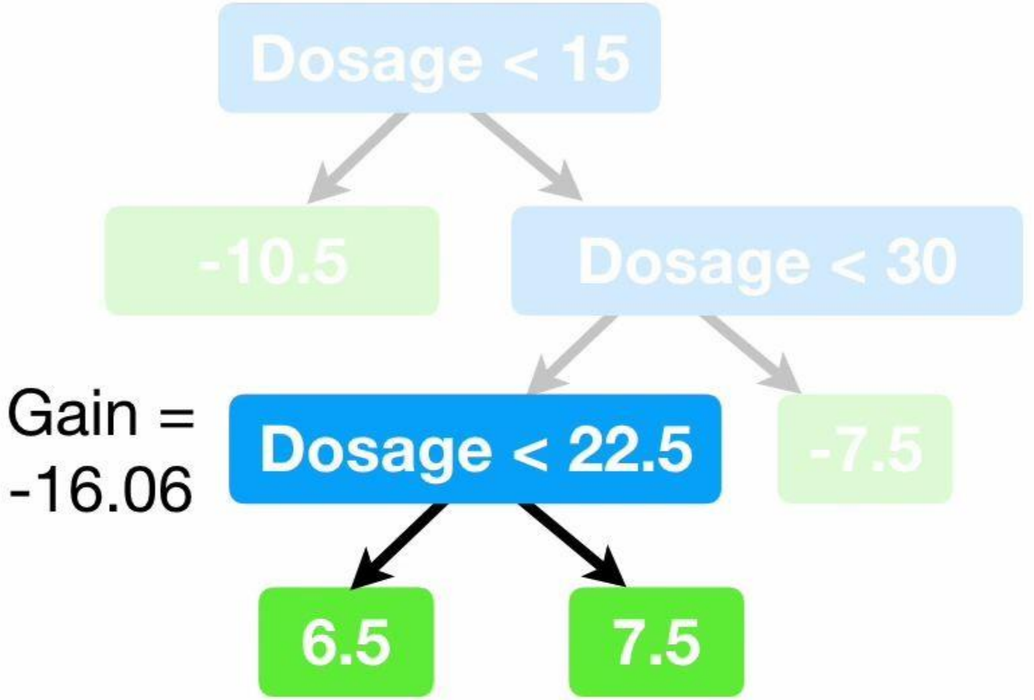
Gain = -16.06

Gain - γ =

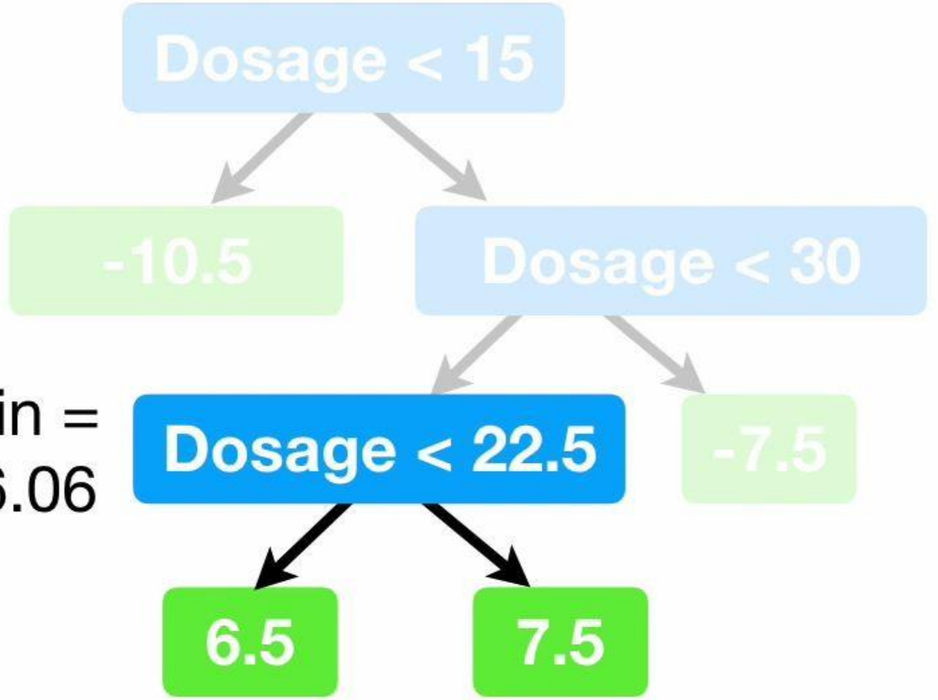
NOTE! If we set $\gamma = 0$…

-16.07 - γ =

Gain

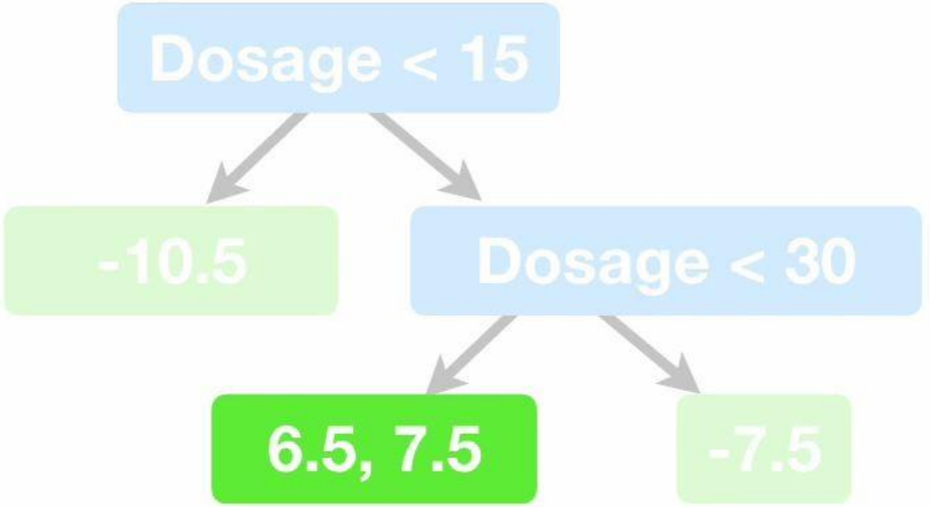


…and we will prune this branch, even though $\gamma = 0$.



-16.07 - 0 = -16.07

Gain — γ **(gamma)**



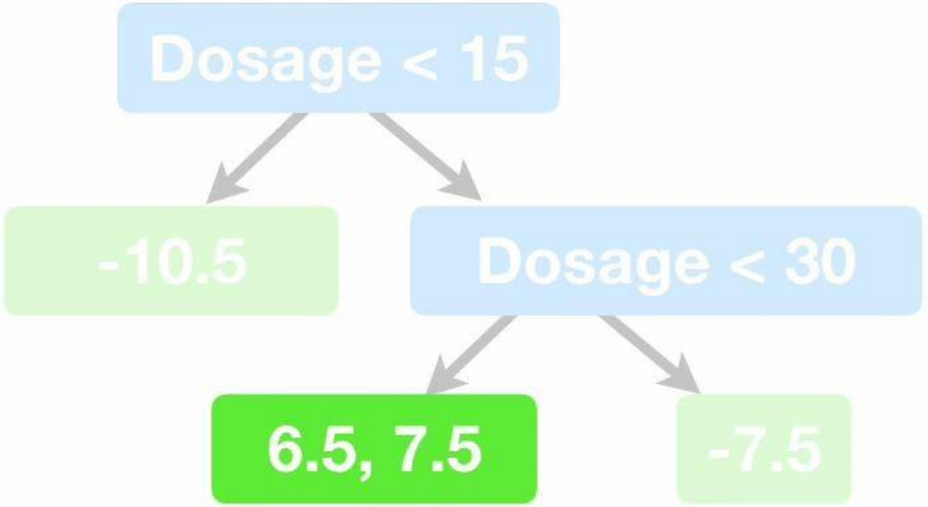
In other words, setting ***γ* = 0** does not turn off pruning.

-16.07 - 0 = -16.07

Gain

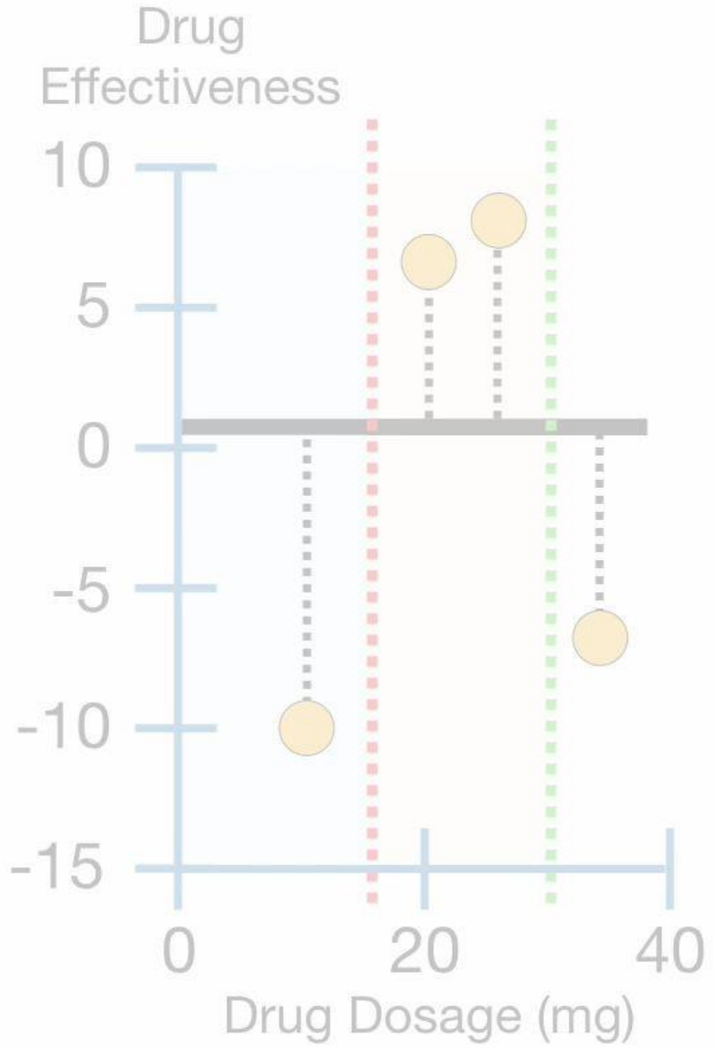
***γ* (gamma)**

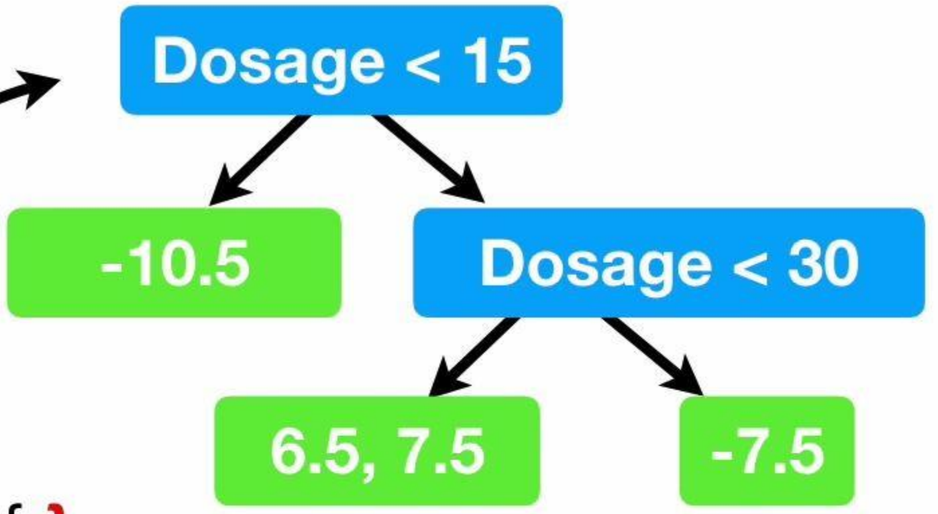
On the other hand, by setting **λ (lambda) = 1**, **λ** did what it was supposed to do; it prevented over fitting the **Training Data**.



Predicted Drug Effectiveness

0.5

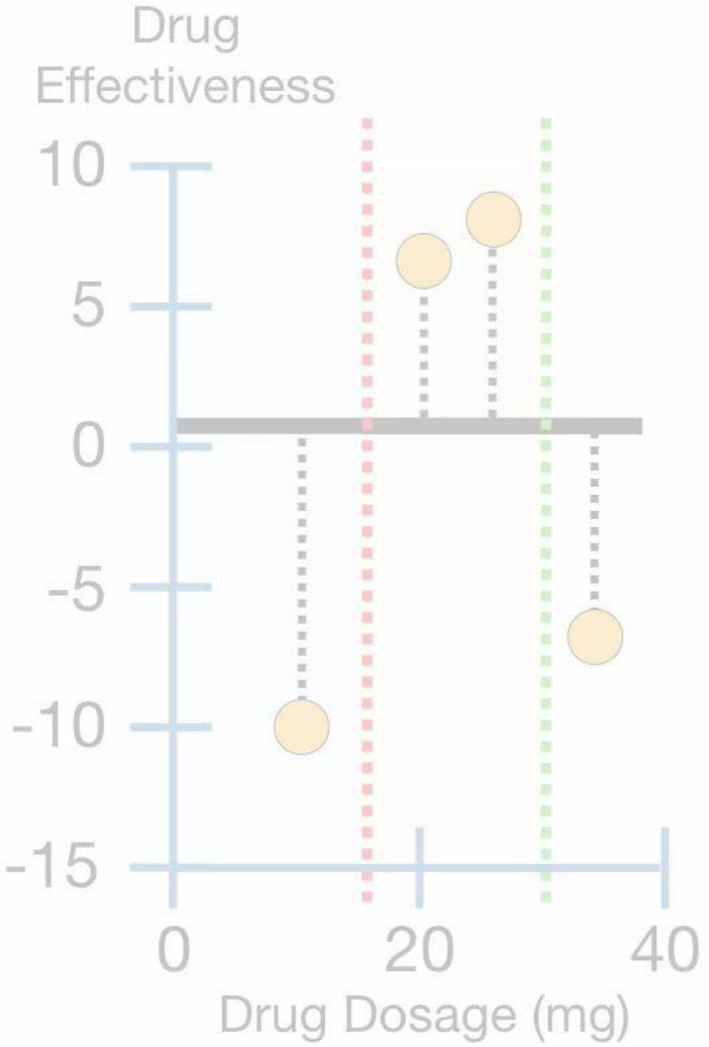


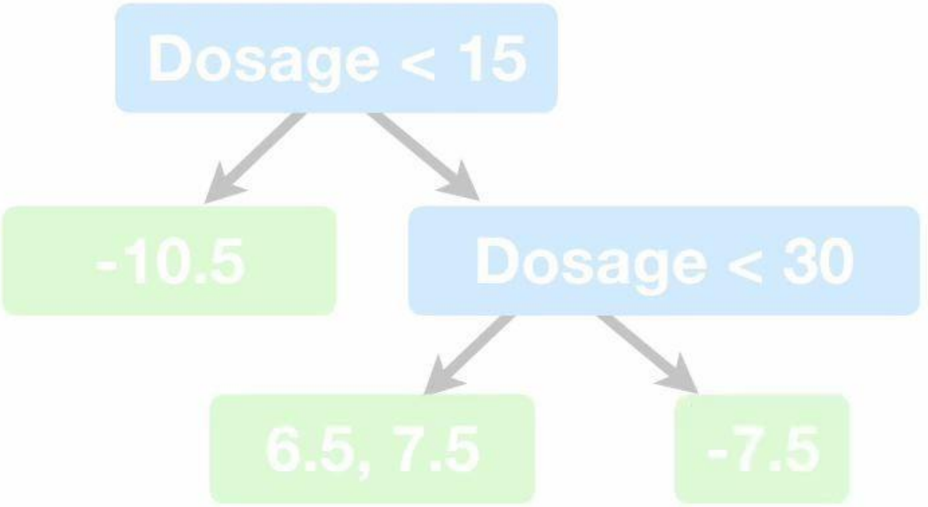


For now, regardless of ***λ*** **(lambda)** and ***γ*** **(gamma)**, let's assume this is the tree we are working with…

Predicted Drug Effectiveness

0.5

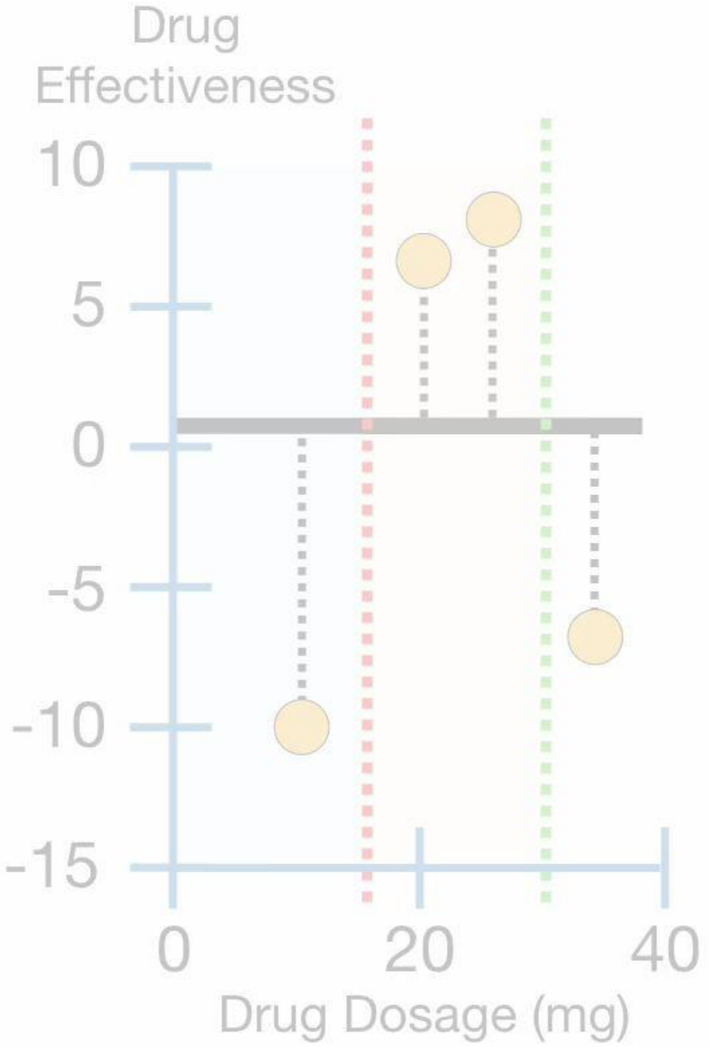


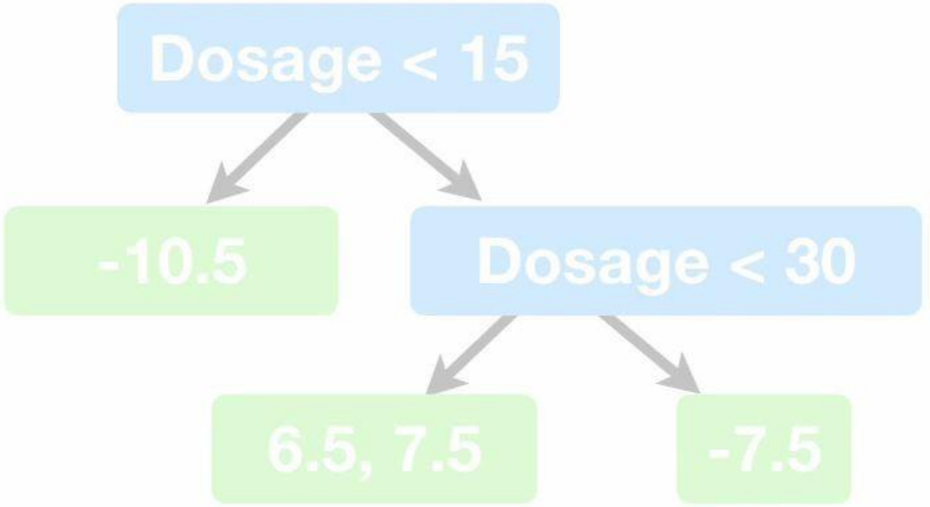


$$\text{Output Value} = \frac{\text{Sum of Residuals}}{\text{Number of Residuals} + \lambda}$$

Predicted Drug Effectiveness

0.5

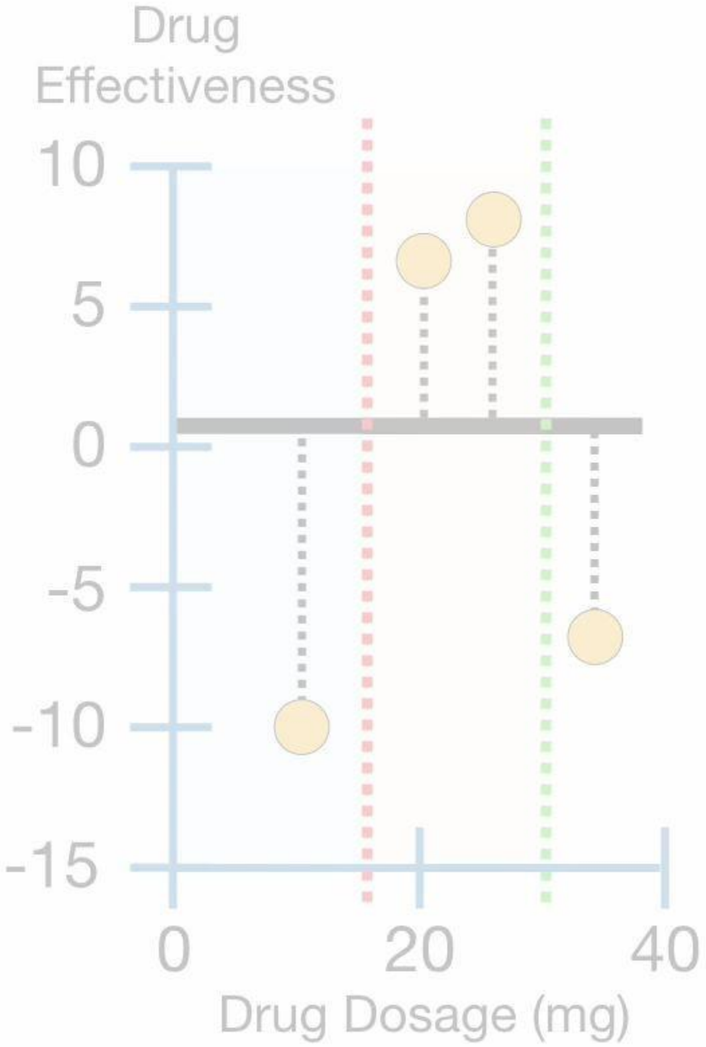


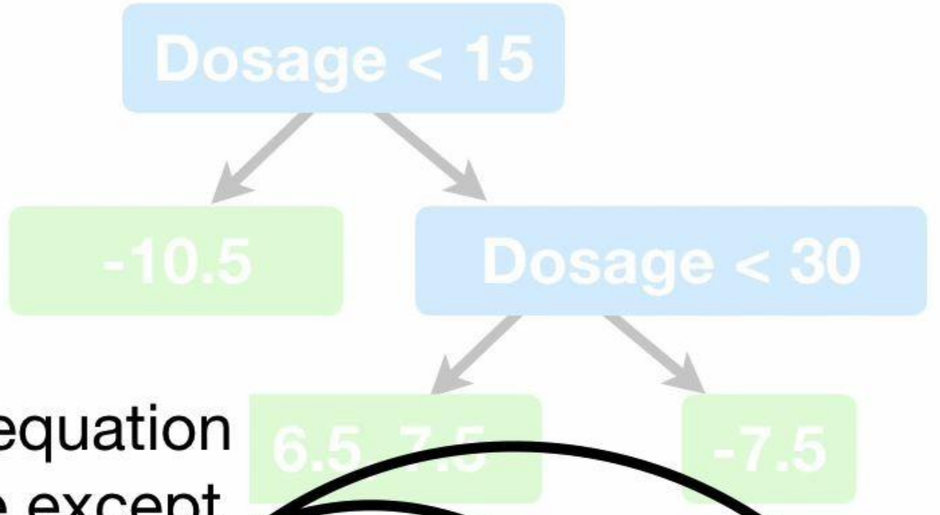


$$\text{Output Value} = \frac{\text{Sum of Residuals}}{\text{Number of Residuals} + \lambda}$$

Predicted Drug Effectiveness

0.5





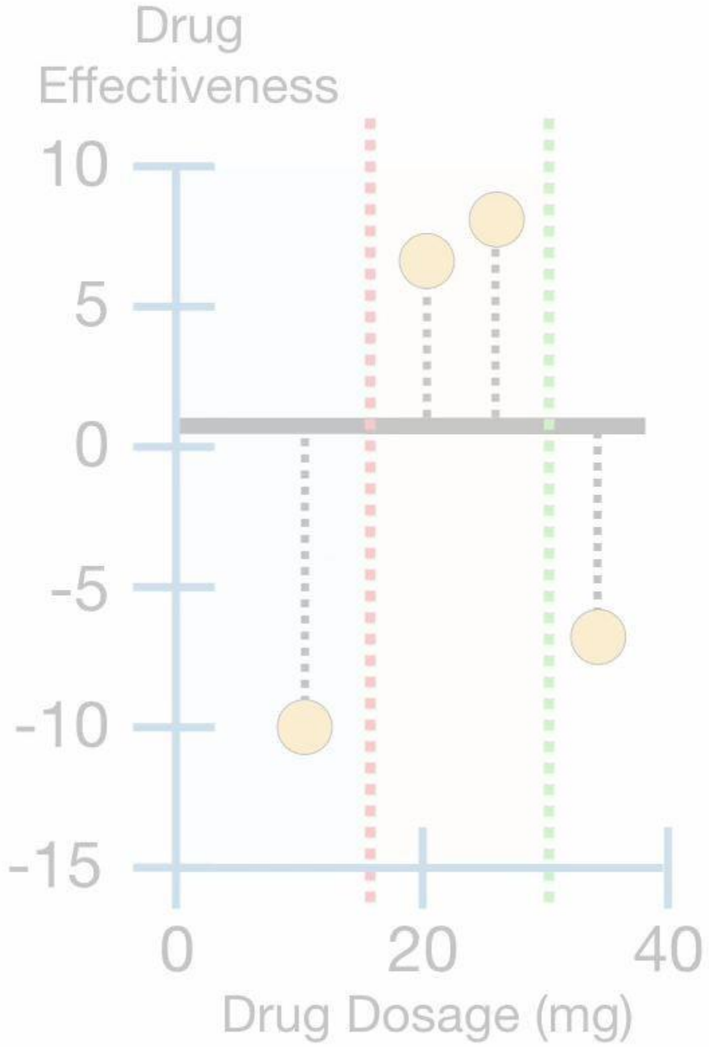
NOTE: The **Output Value** equation is like the **Similarity Score** except we do not square the sum of the residuals.

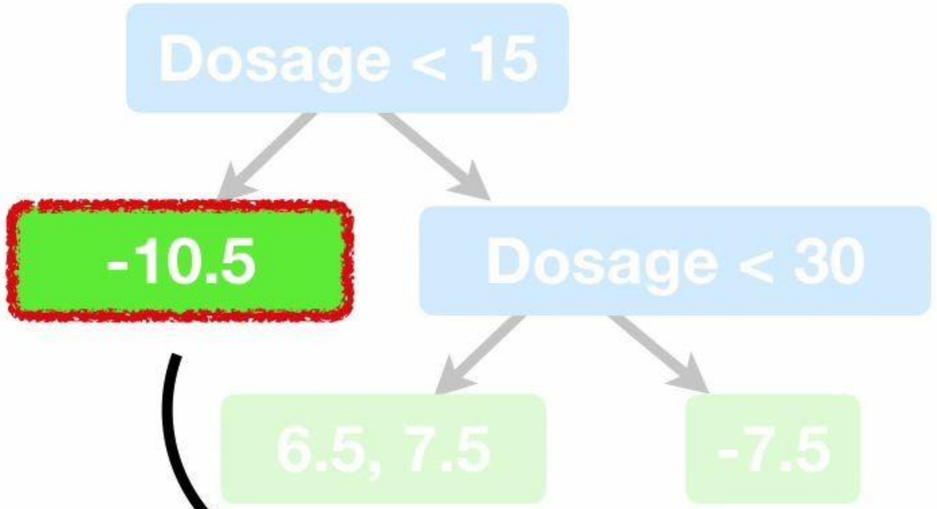
$$\text{Output Value} = \frac{\text{Sum of Residuals}}{\text{Number of Residuals} + \lambda}$$

$$\text{Similarity Score} = \frac{\text{Sum of Residuals, Squared}}{\text{Number of Residuals} + 1}$$

Predicted Drug Effectiveness

0.5



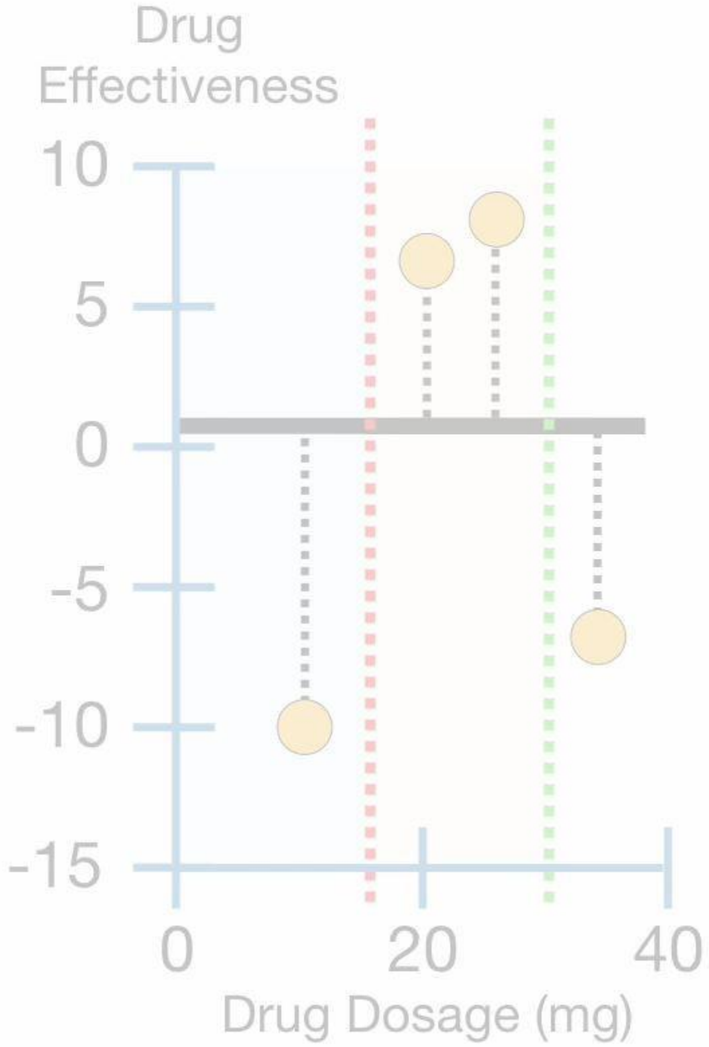


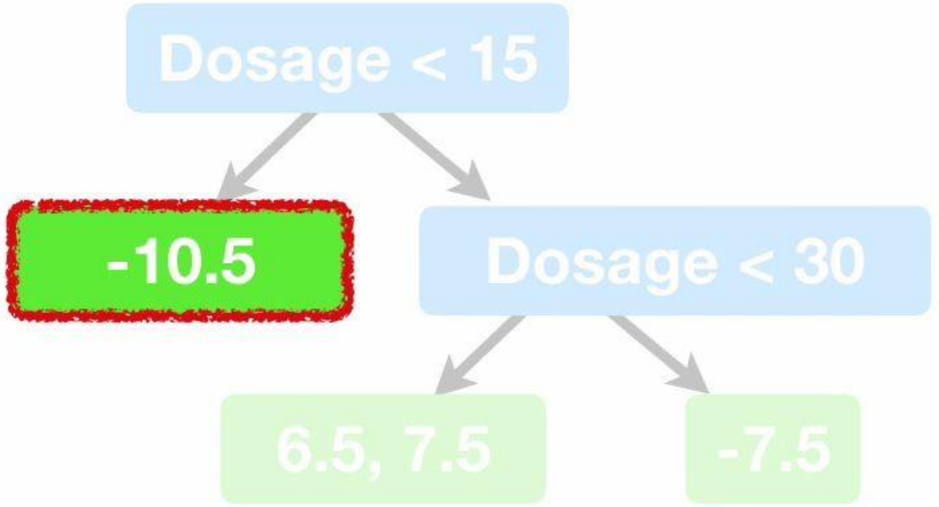
…we plug in the **Residual**, **-10.5**…

$$\text{Output Value} = \frac{\text{Sum of Residuals}}{\text{Number of Residuals} + \lambda}$$

Predicted Drug Effectiveness

0.5



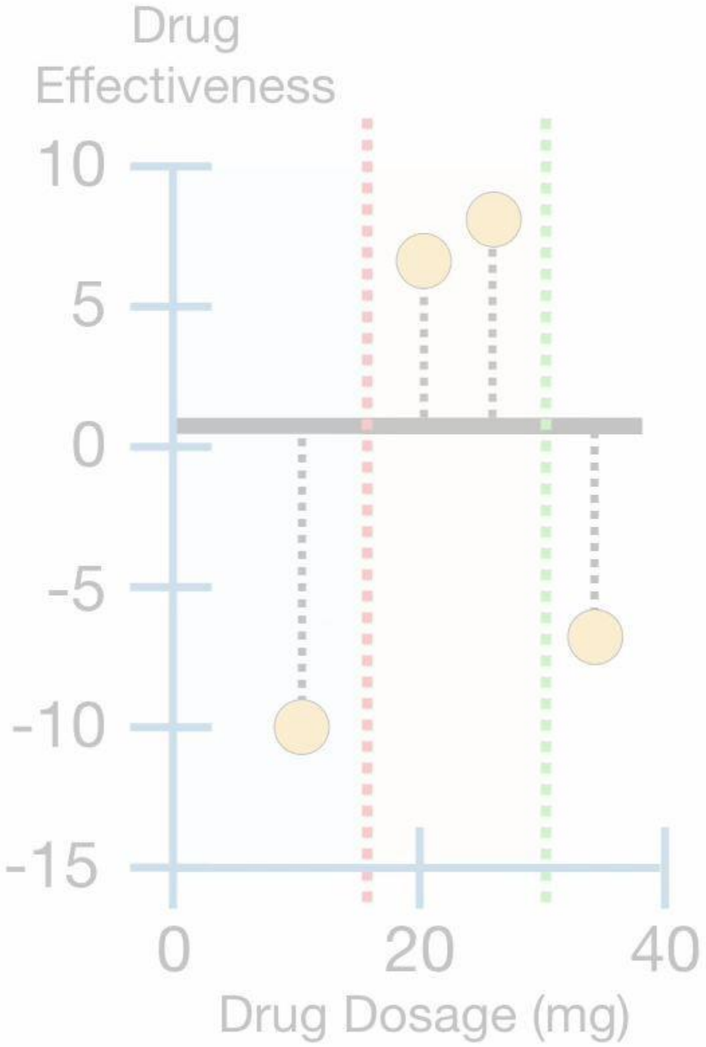


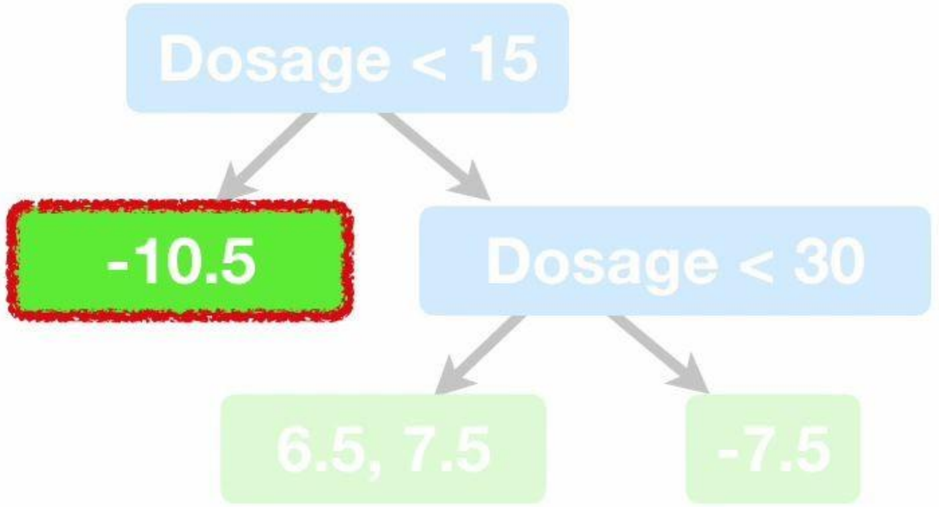
…and the value for the **Regularization Parameter**, λ.

$$\text{Output Value} = \frac{-10.5}{1 + \lambda}$$

Predicted Drug Effectiveness

0.5



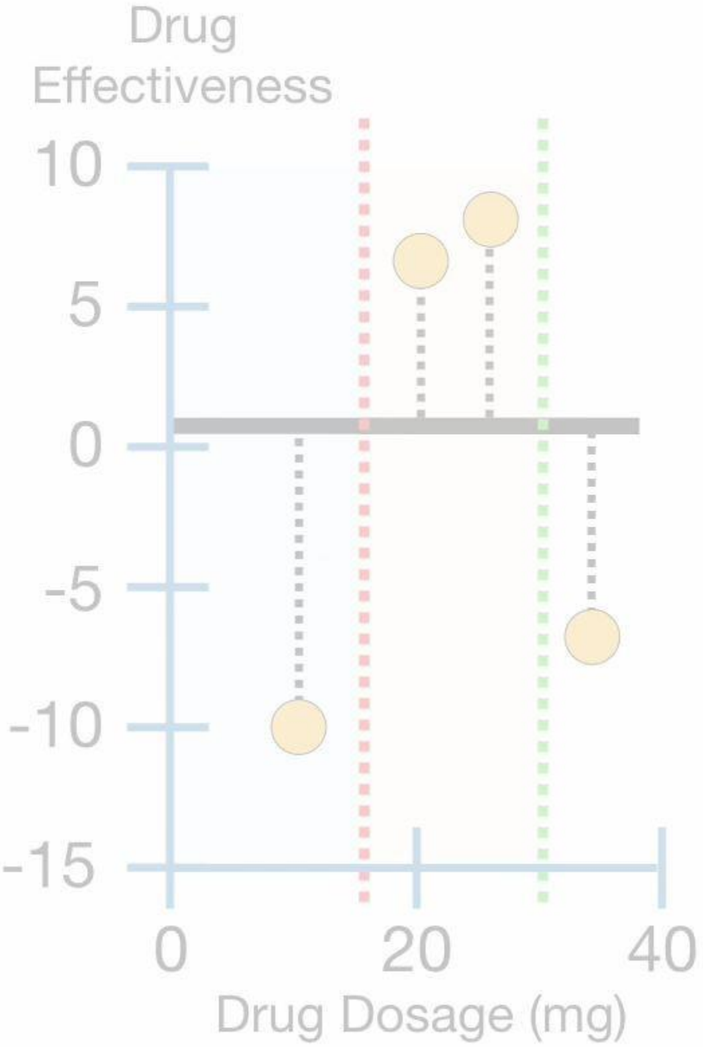


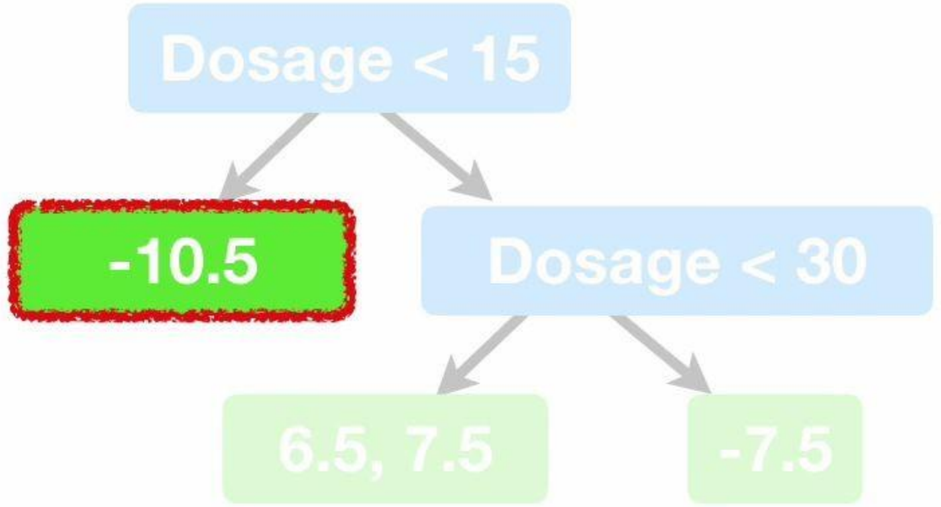
If $\lambda = 0$, then there is no **Regularization** and the **Output Value = -10.5**.

$$\text{Output Value} = \frac{-10.5}{1 + 0}$$

Predicted Drug Effectiveness

0.5



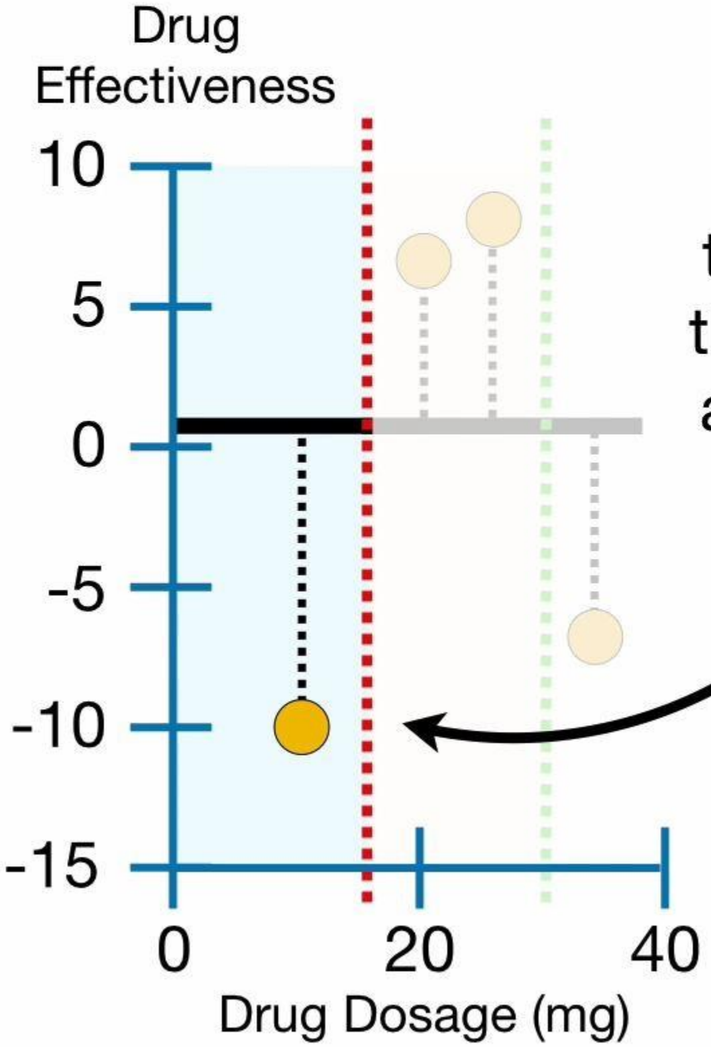


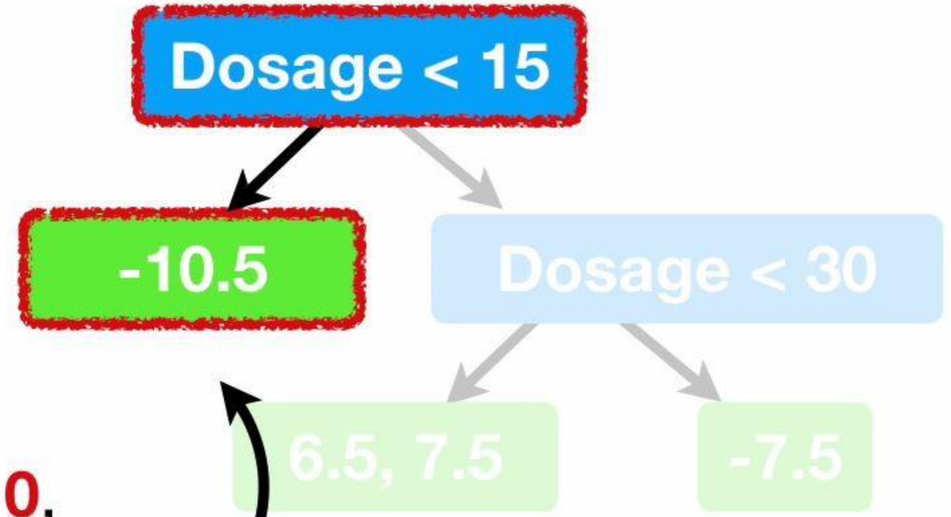
On the other hand, if $\lambda = 1$…

$$\text{Output Value} = \frac{-10.5}{1 + 1}$$

Predicted Drug Effectiveness

0.5



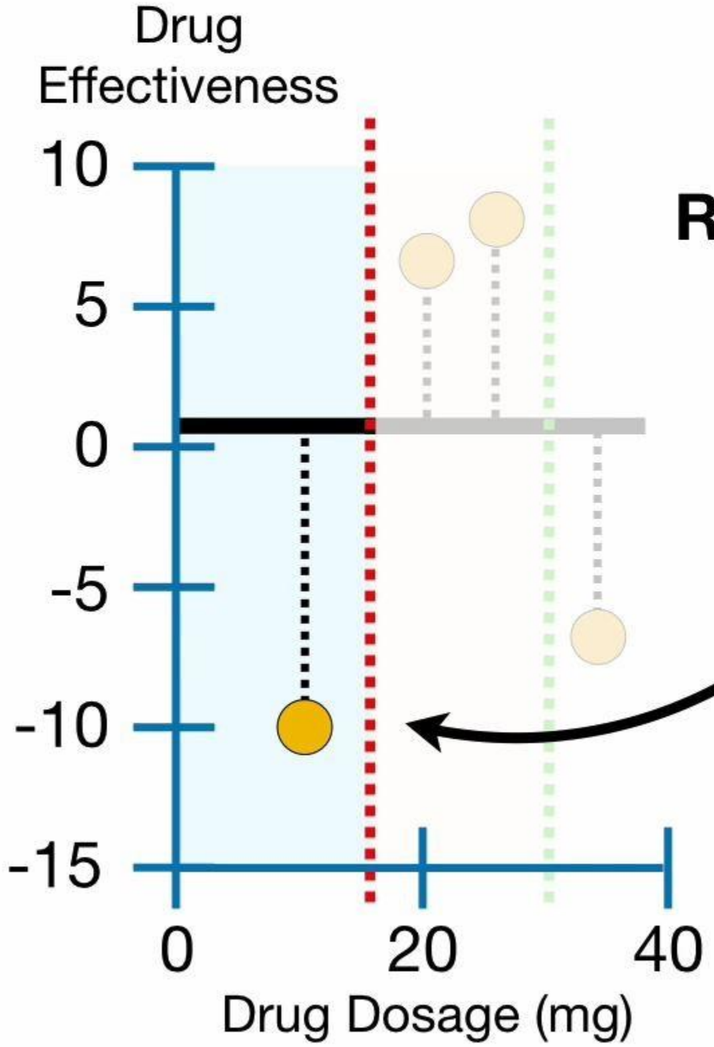


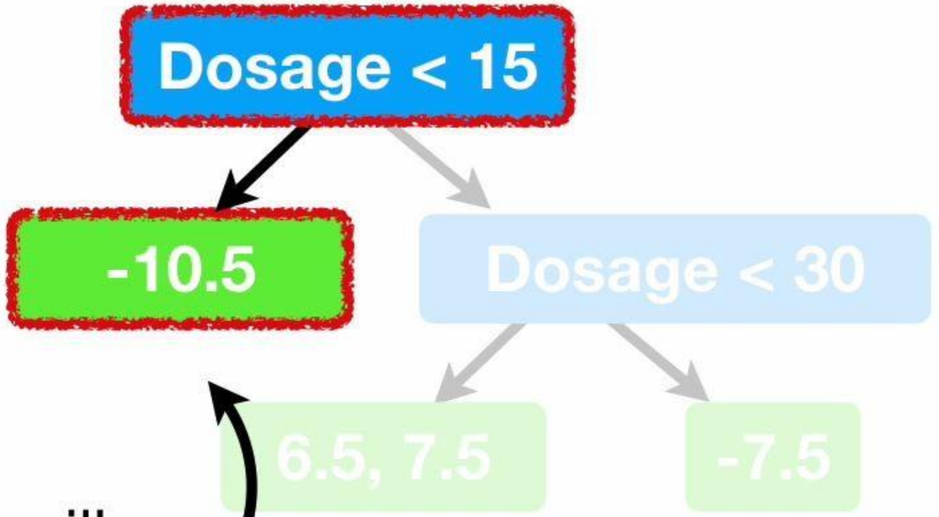
In other words, when $\lambda > 0$, then it will reduce the amount that this individual observation adds to the overall prediction.

$$\text{Output Value} = \frac{-10.5}{1 + 1} = -5.25$$

Predicted Drug Effectiveness

0.5



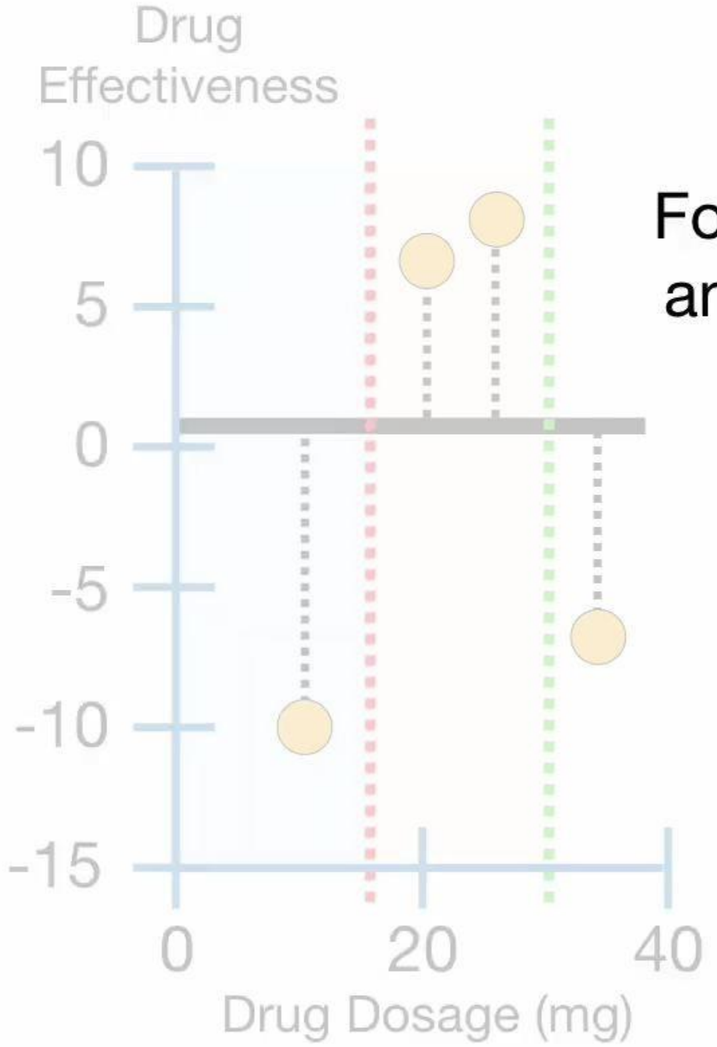


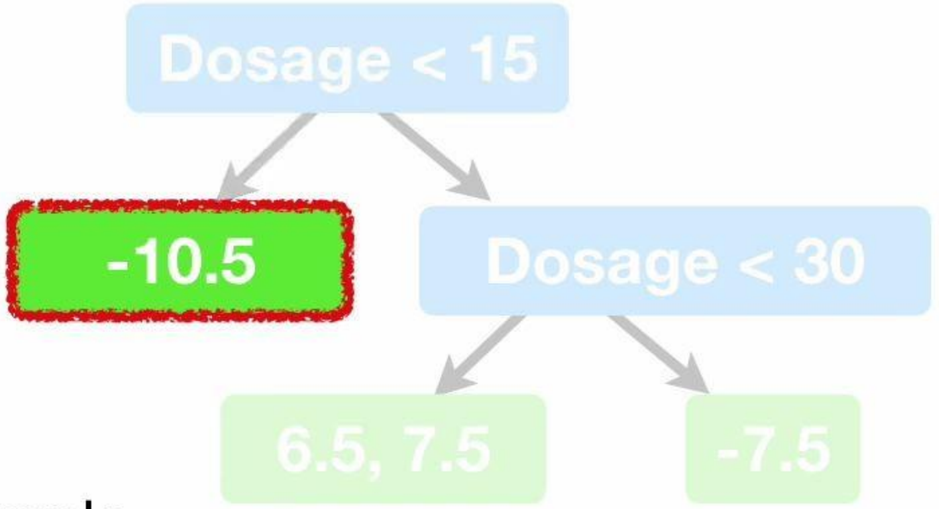
Thus, ***λ* (lambda)**, the **Regularization Parameter**, will reduce the prediction's sensitivity to this individual observation.

$$\text{Output Value} = \frac{-10.5}{1 + \mathbf{1}} = -5.25$$

Predicted Drug Effectiveness

0.5



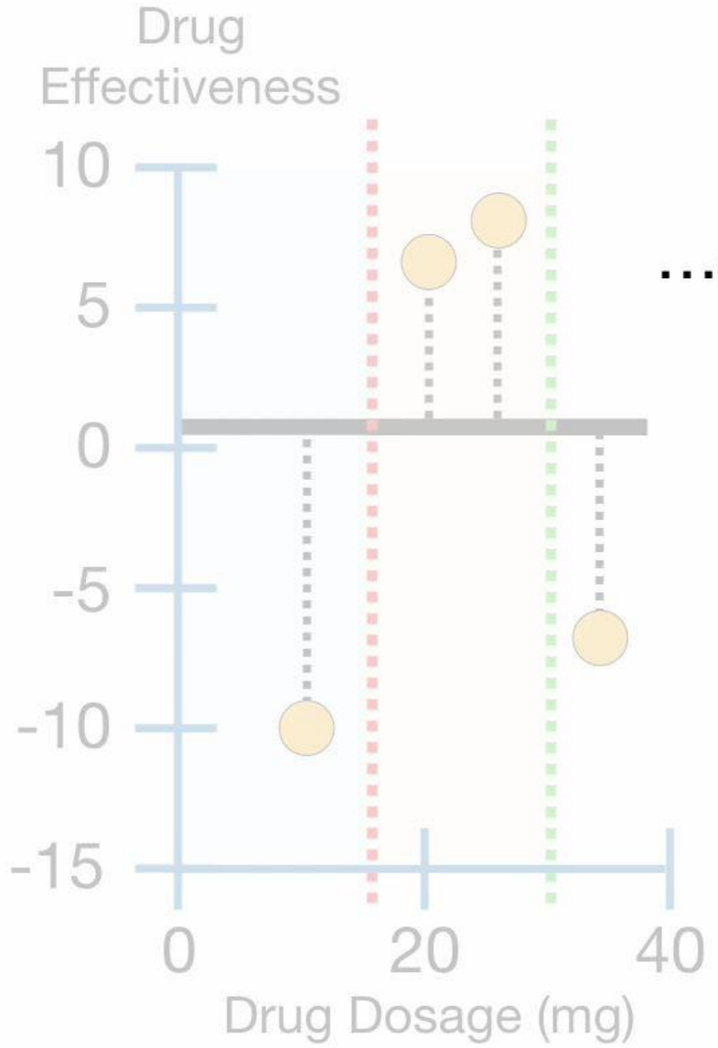


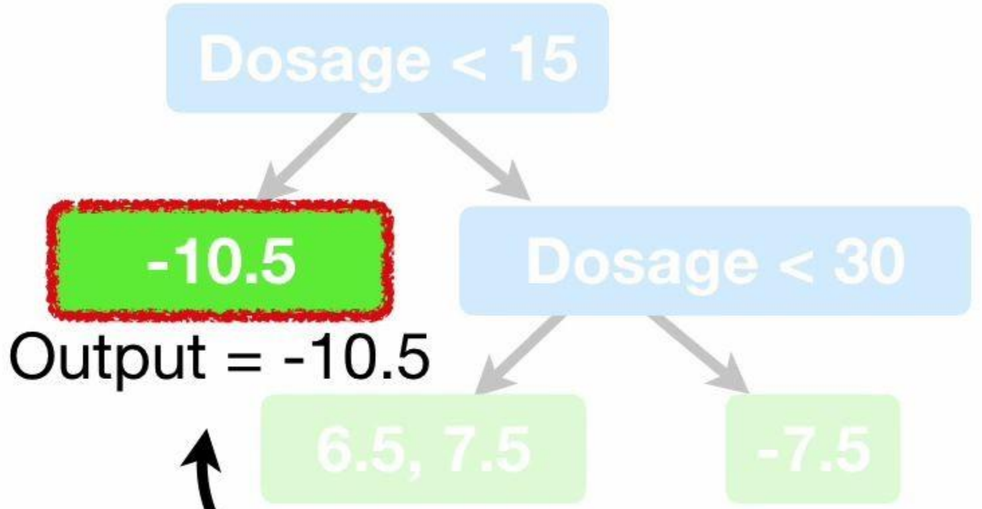
For now, we'll keep things simple and let $\lambda = 0$, because this is the default value…

$$\text{Output Value} = \frac{-10.5}{1 + 0} = -10.5$$

Predicted Drug Effectiveness

0.5





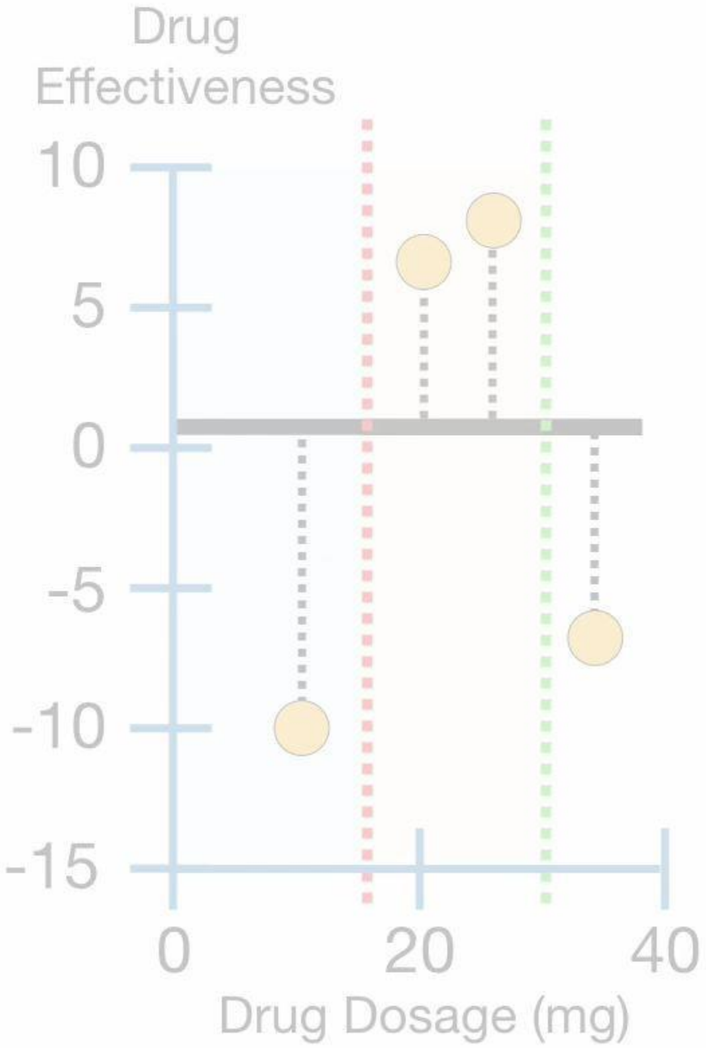
Output = -10.5

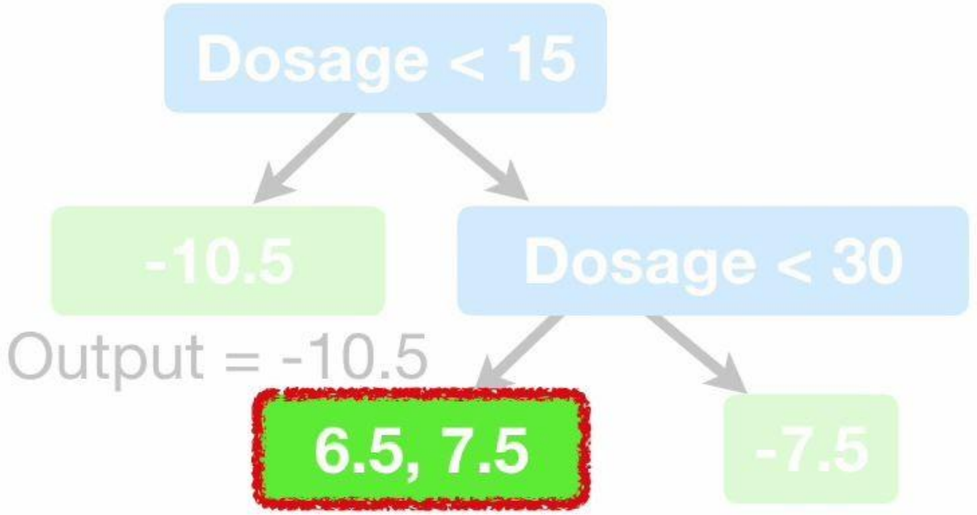
…and put **-10.5** under the leaf so we will remember it.

$$\text{Output Value} = \frac{-10.5}{1 + \mathbf{0}} = -10.5$$

Predicted Drug Effectiveness

0.5



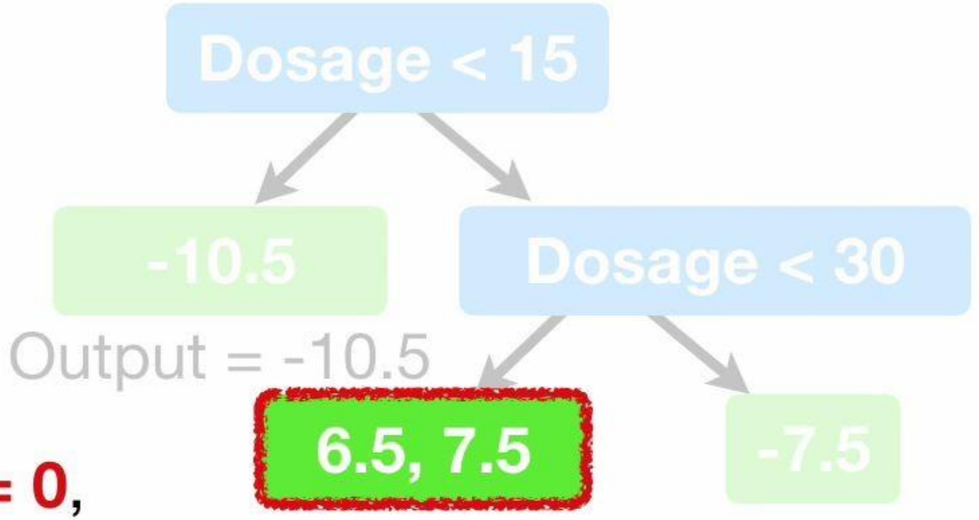


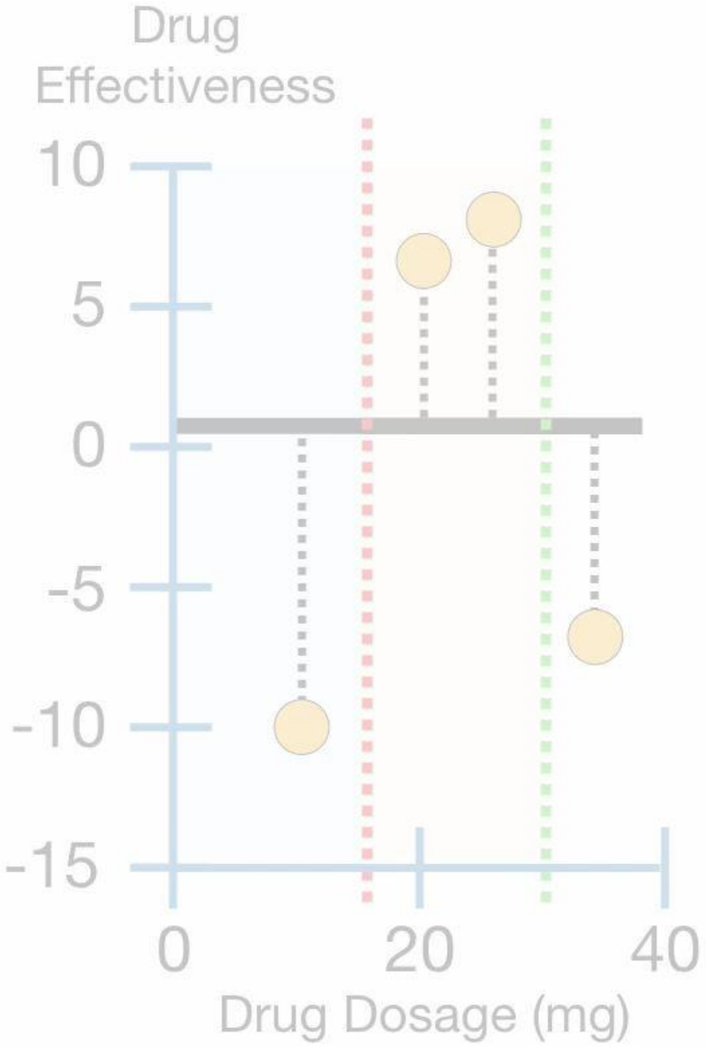
When $\lambda = 0$, the **Output Value** is…

$$\text{Output Value} = \frac{\text{Sum of Residuals}}{\text{Number of Residuals} + 0}$$

Predicted Drug Effectiveness

0.5



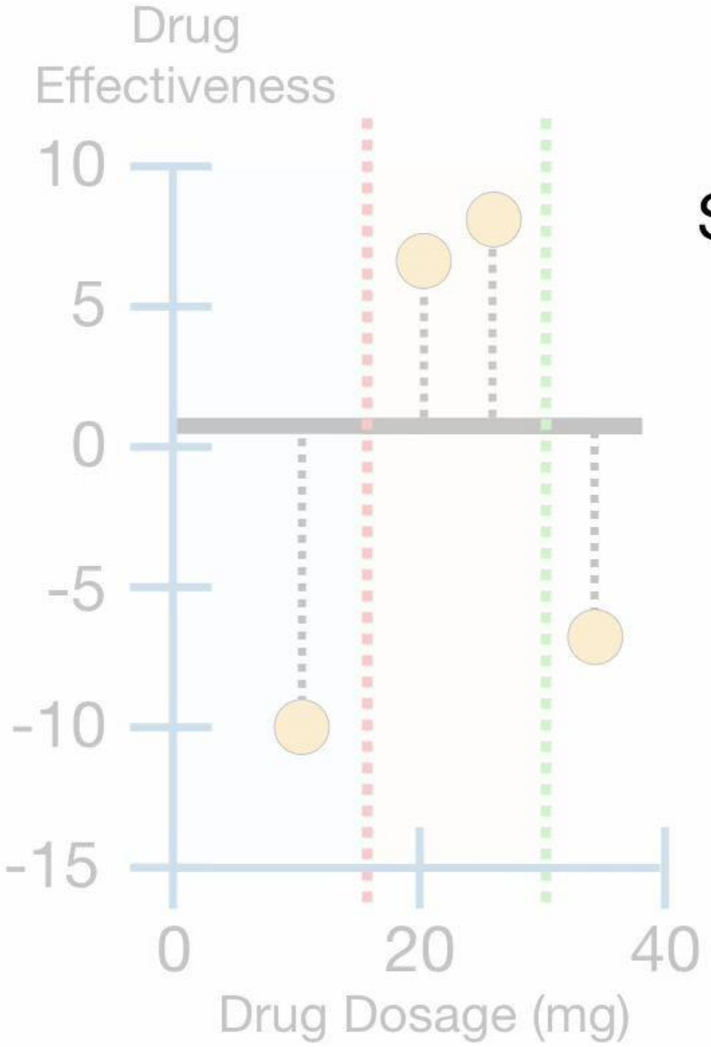


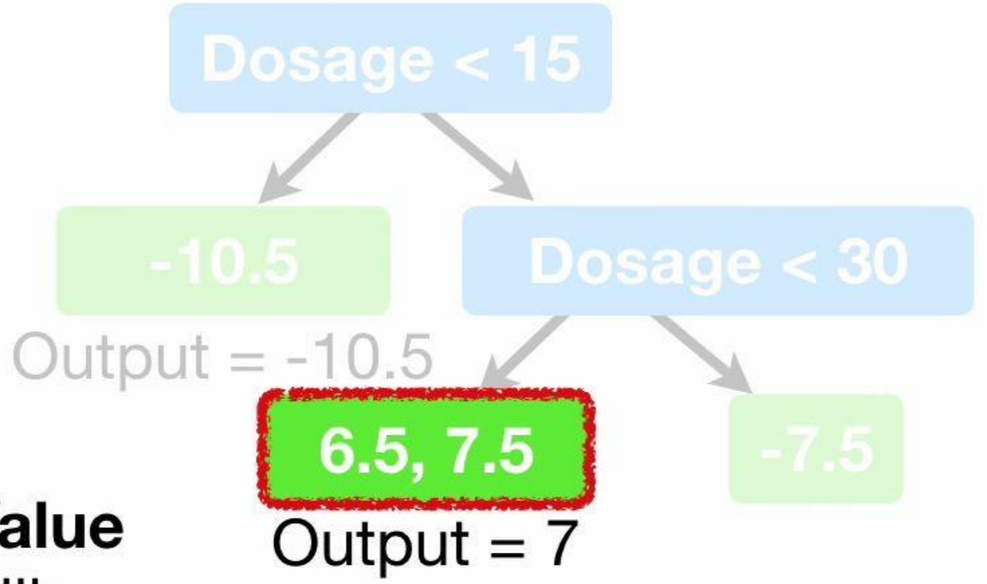
In other words, when $\lambda = 0$, the **Output Value** for a leaf is simply the *average* of the **Residuals** in that leaf.

$$\text{Output Value} = \frac{6.5 + 7.5}{2 + 0} = 7$$

Predicted Drug Effectiveness

0.5



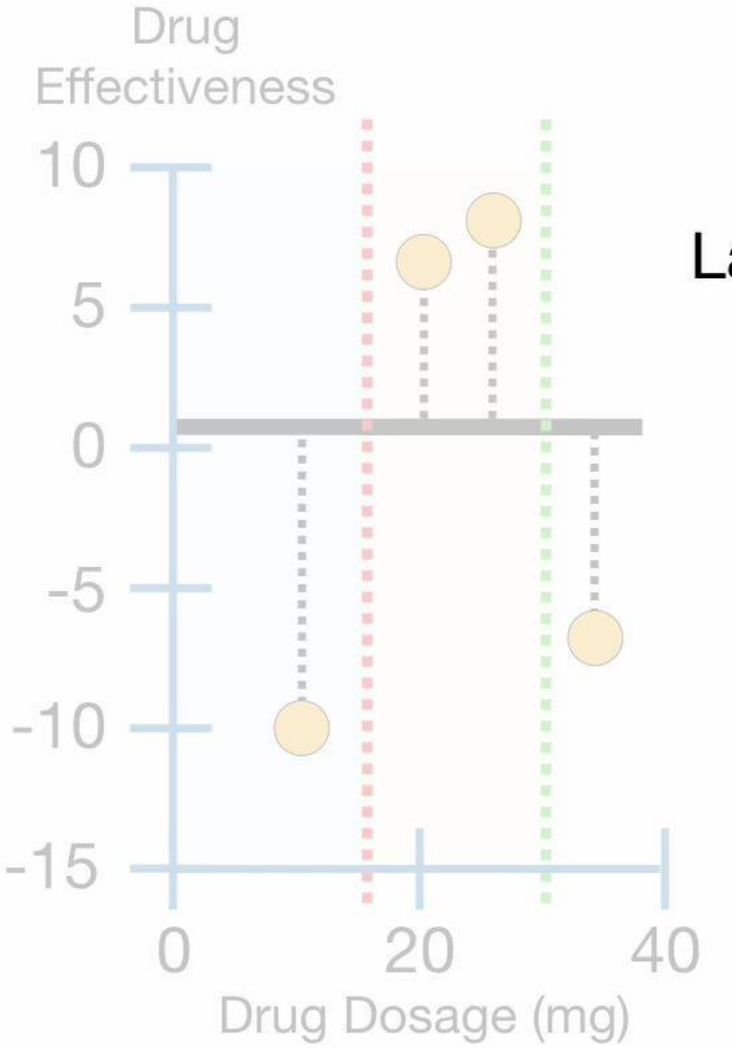


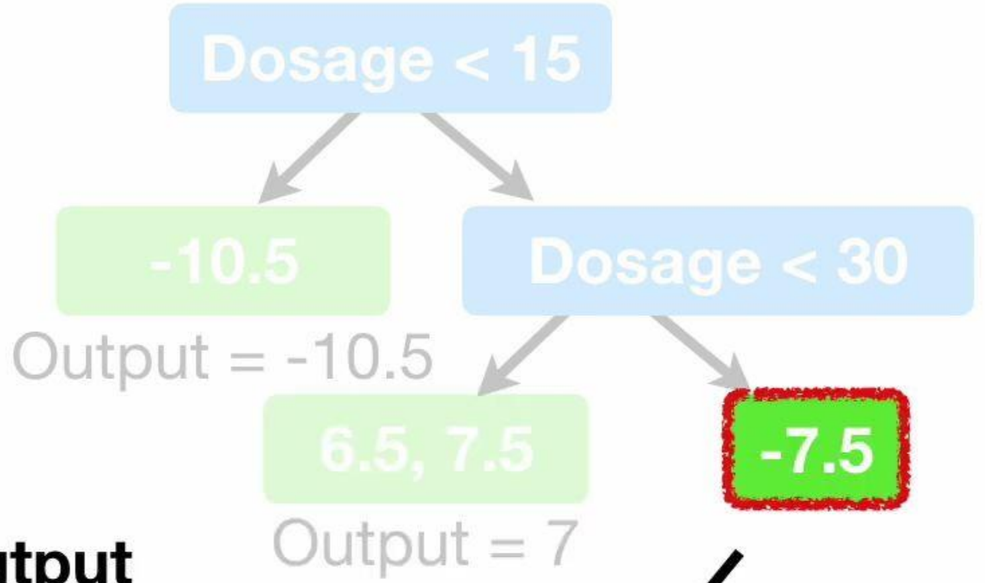
So we'll put the **Output Value** under the leaf so we will remember it.

$$\text{Output Value} = \frac{6.5 + 7.5}{2 + \mathbf{0}} = 7$$

Predicted Drug Effectiveness

0.5



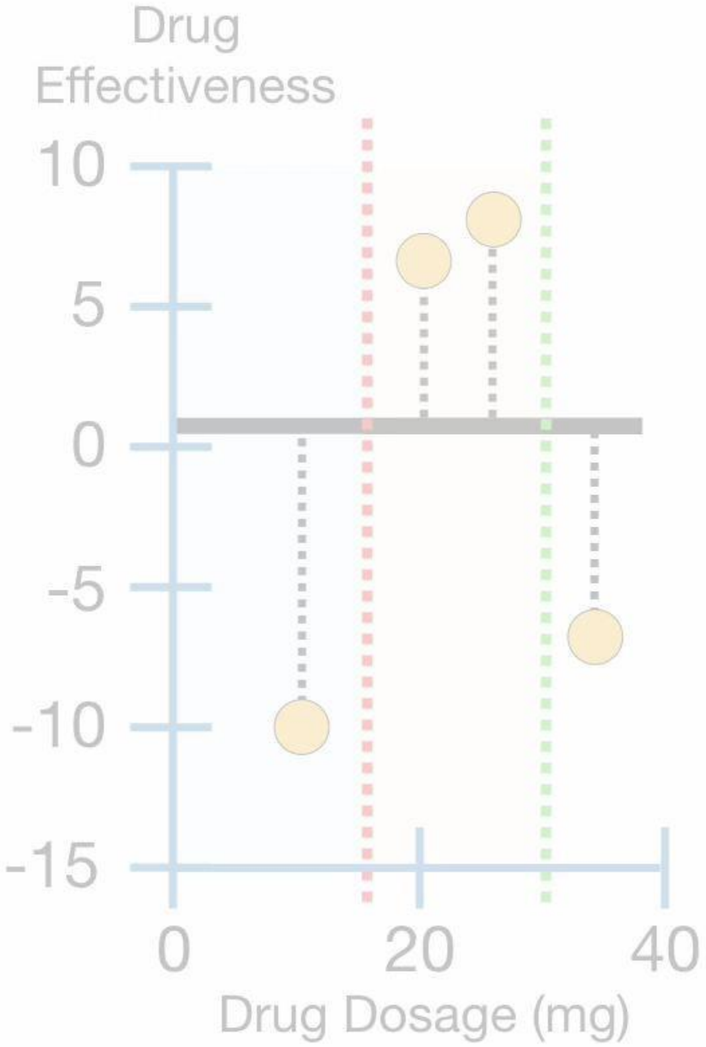


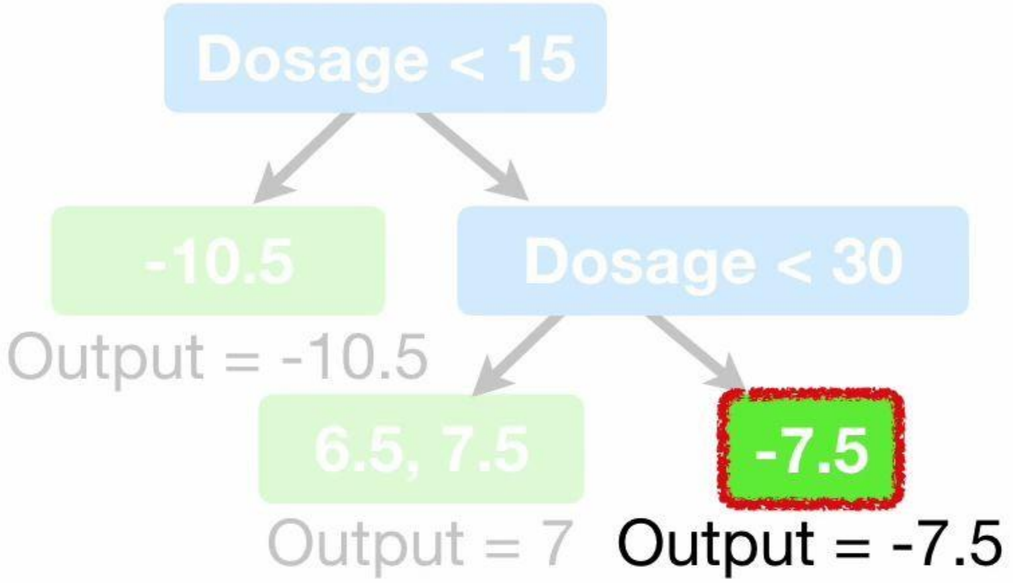
Lastly, when $\lambda = 0$, the **Output Value** for this leaf is…

$$\text{Output Value} = \frac{-7.5}{\text{Number of Residuals} + 0}$$

Predicted Drug Effectiveness

0.5



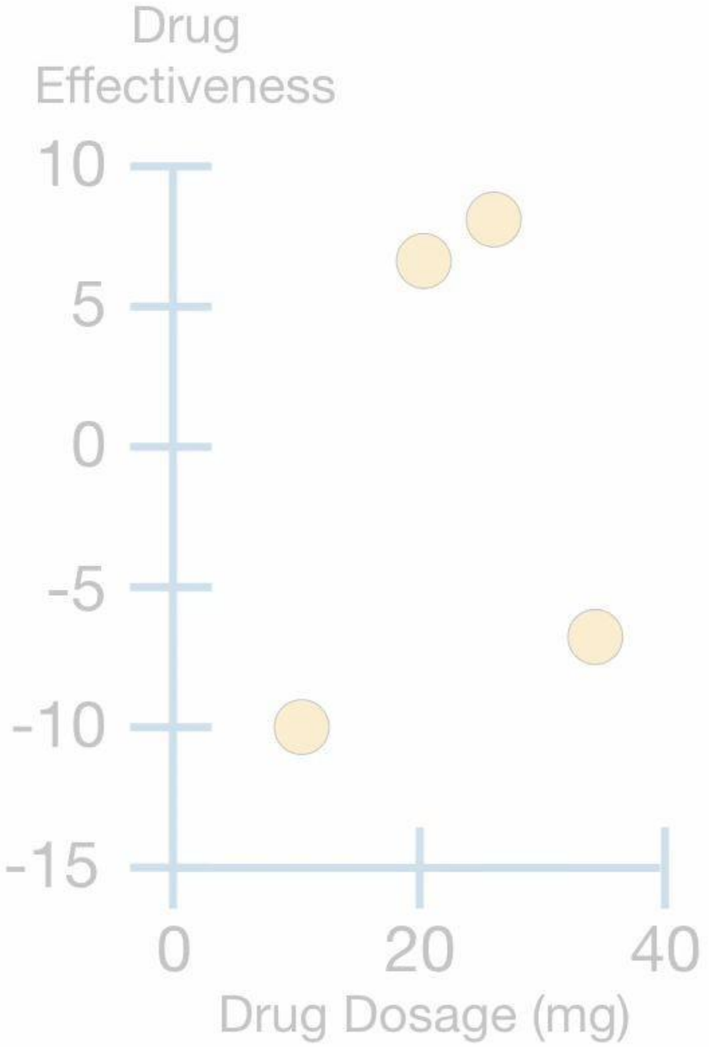


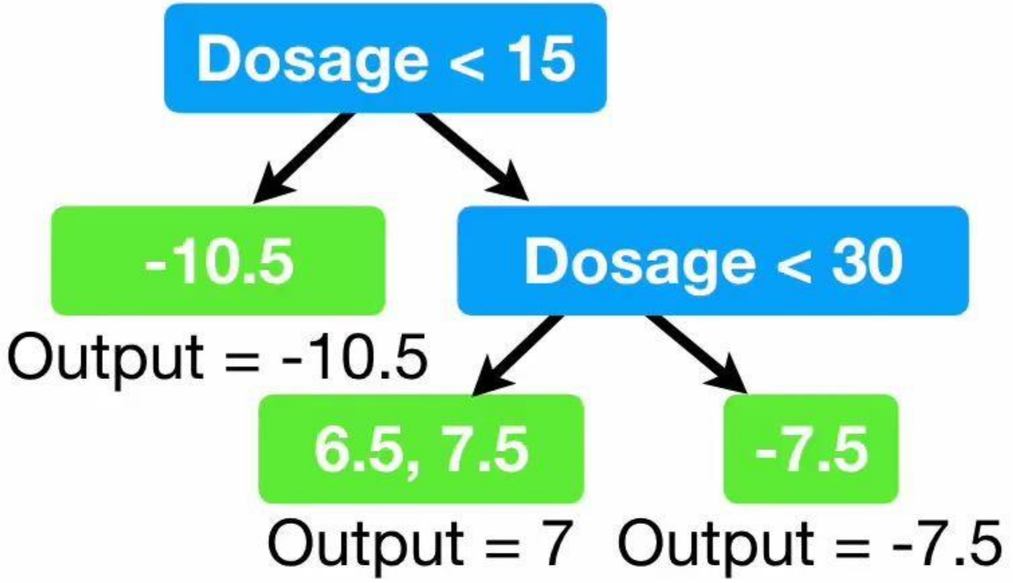
…**-7.5**.

$$\text{Output Value} = \frac{-7.5}{1 + \mathbf{0}} = -7.5$$

Predicted Drug Effectiveness

0.5

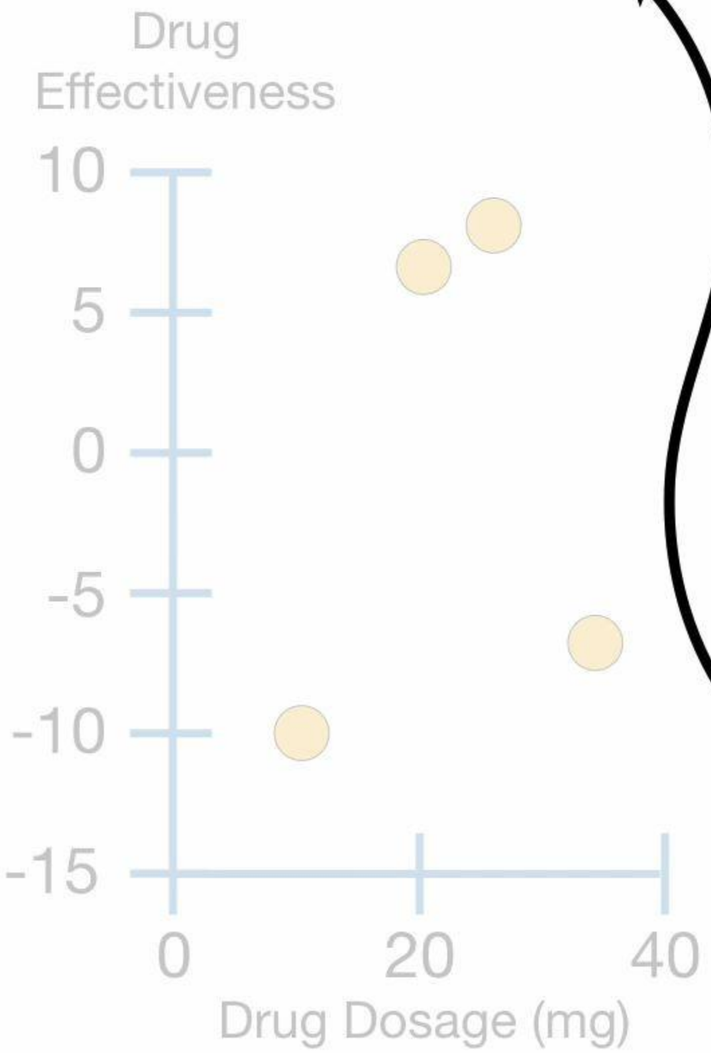


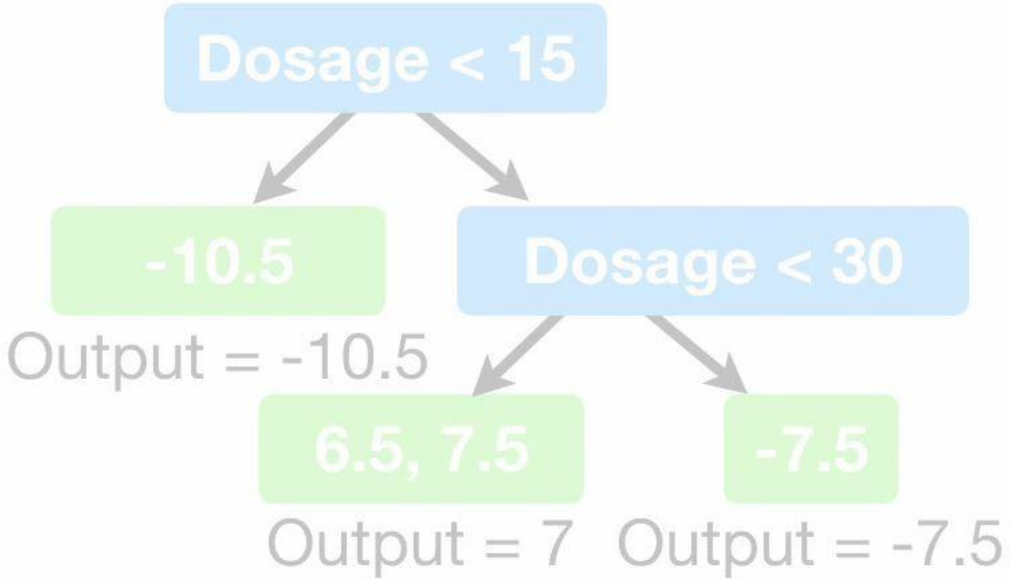


Since we have built our first tree,
we can make new **Predictions**.

Predicted Drug Effectiveness

0.5



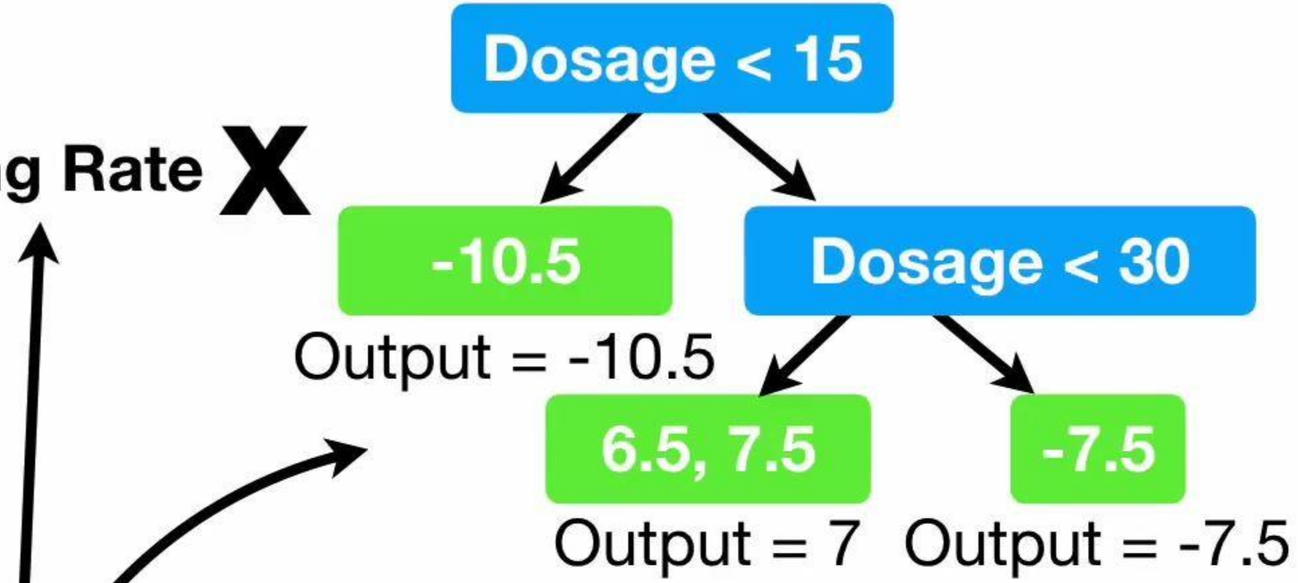


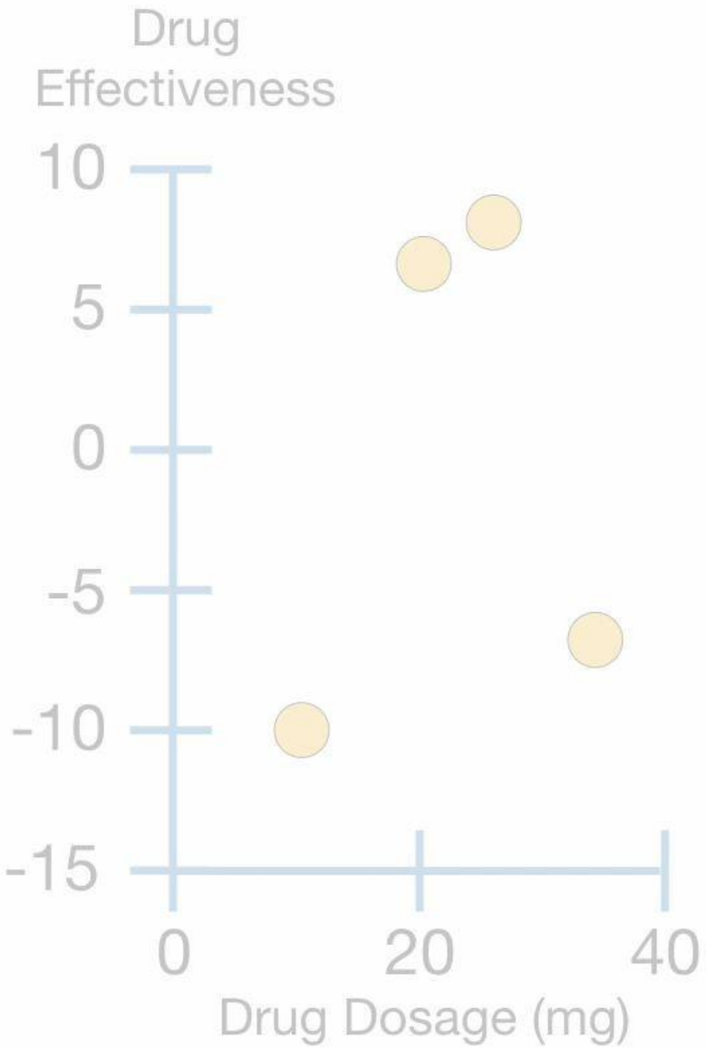
And just like unextreme **Gradient Boost**, **XGBoost** makes new predictions by starting with the initial **Prediction**…

Predicted Drug Effectiveness

0.5

+ Learning Rate X



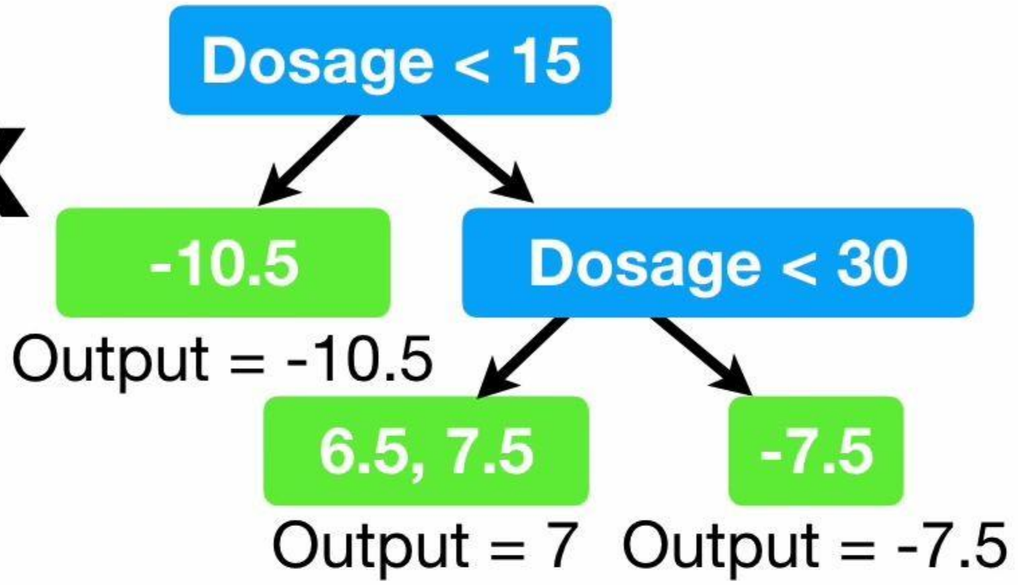


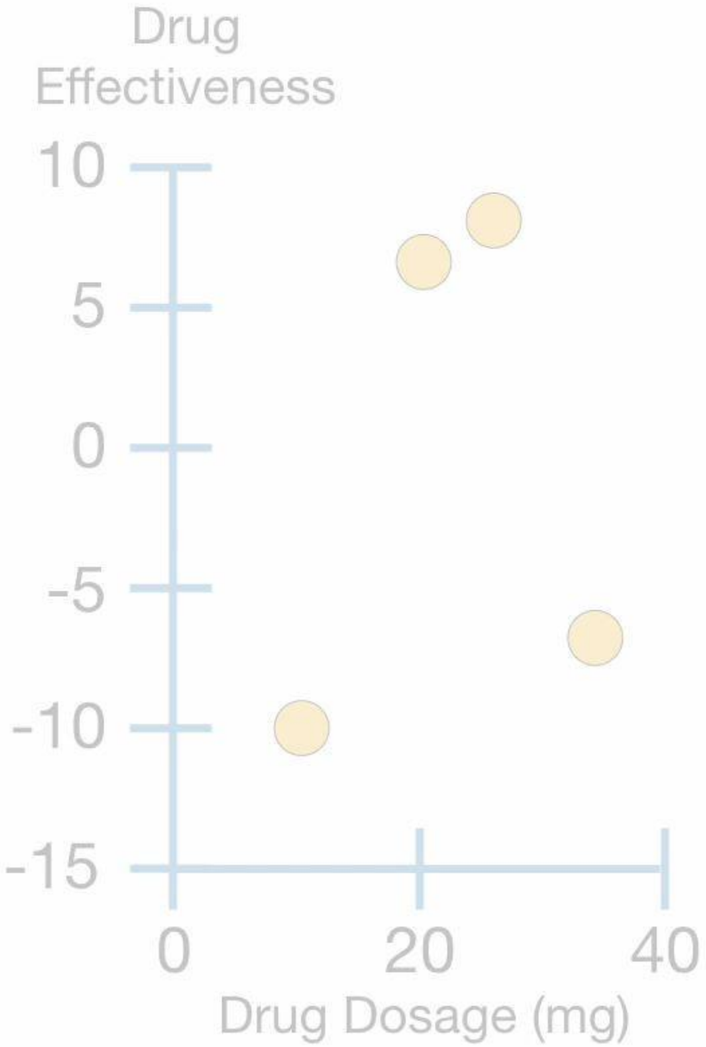
…and adding the output of the **Tree**, scaled by a **Learning Rate**.

Predicted Drug Effectiveness

0.5

+ Learning Rate X





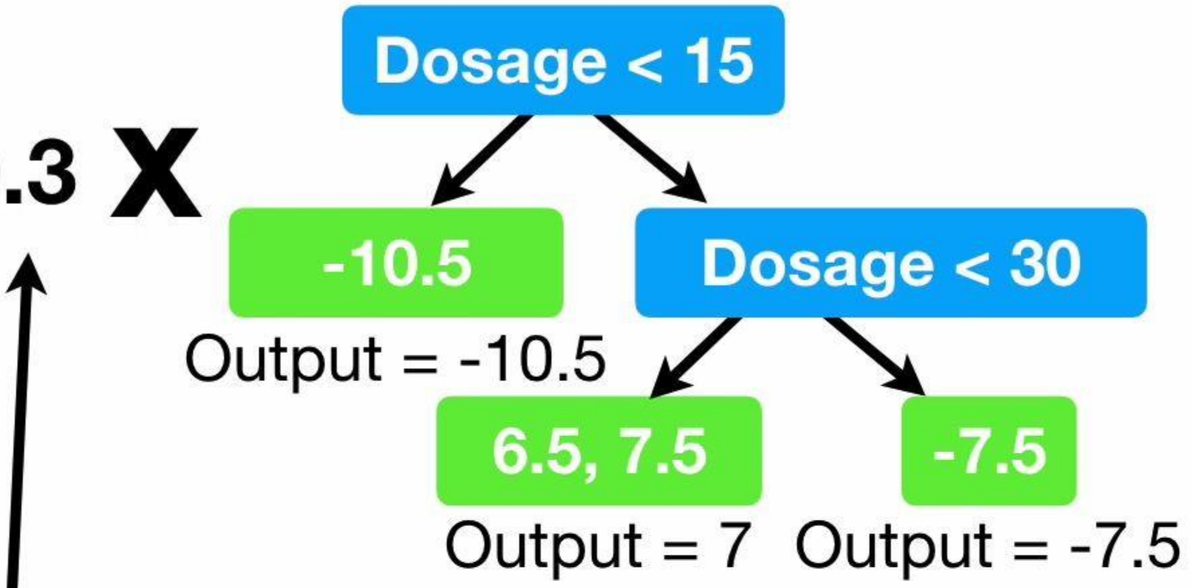
TERMINOLOGY ALERT!!!!

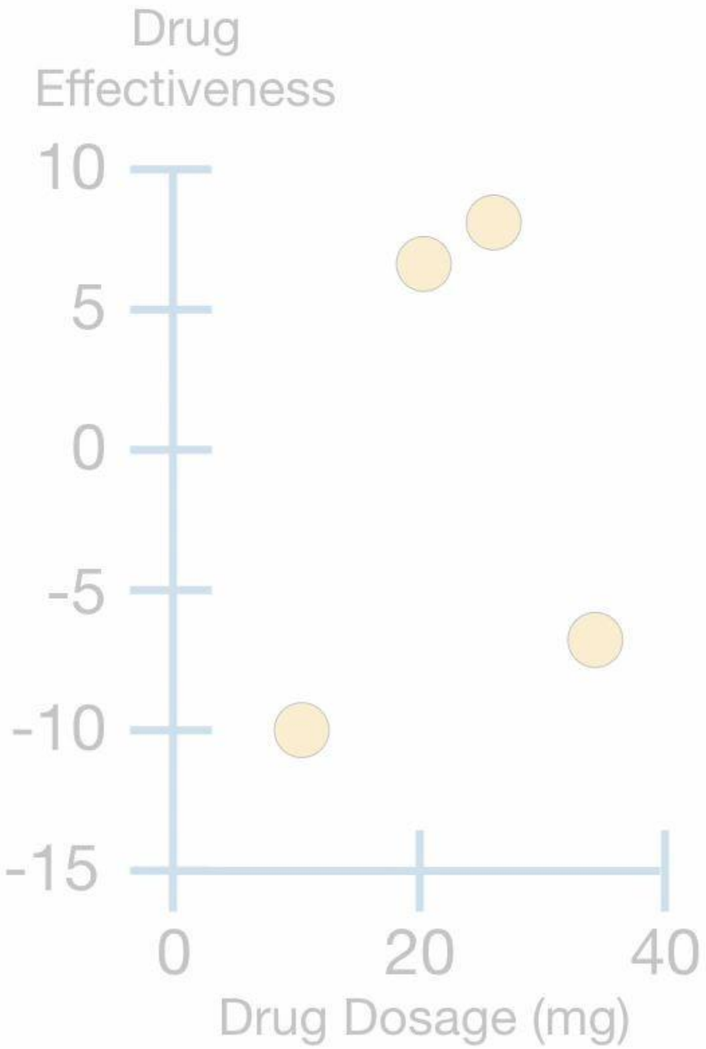
Predicted Drug Effectiveness

0.5

+

0.3 X





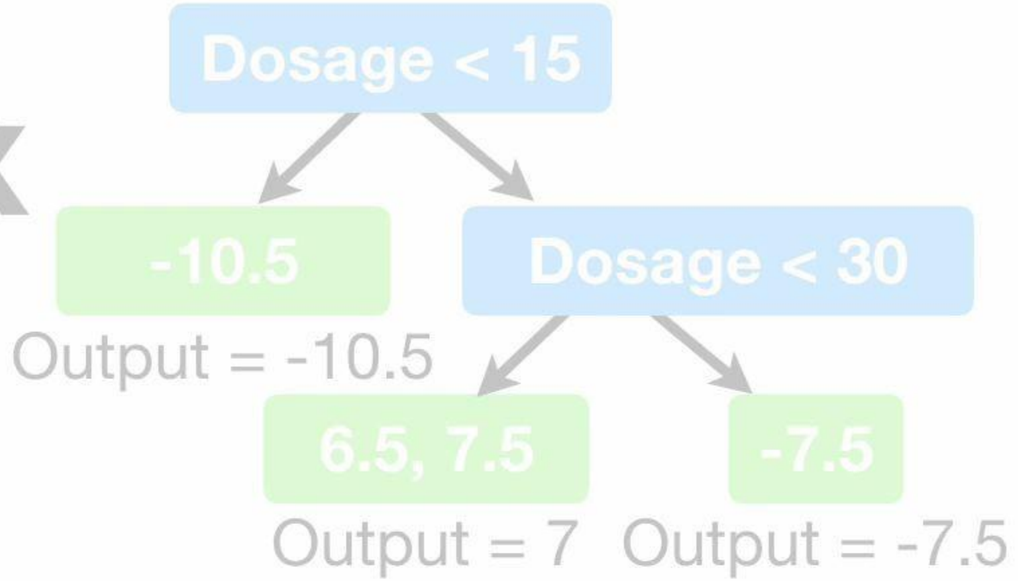
XGBoost calls the **Learning Rate**, ***ε*** **(eta)**, and the default value is **0.3**, so that's what we'll use.

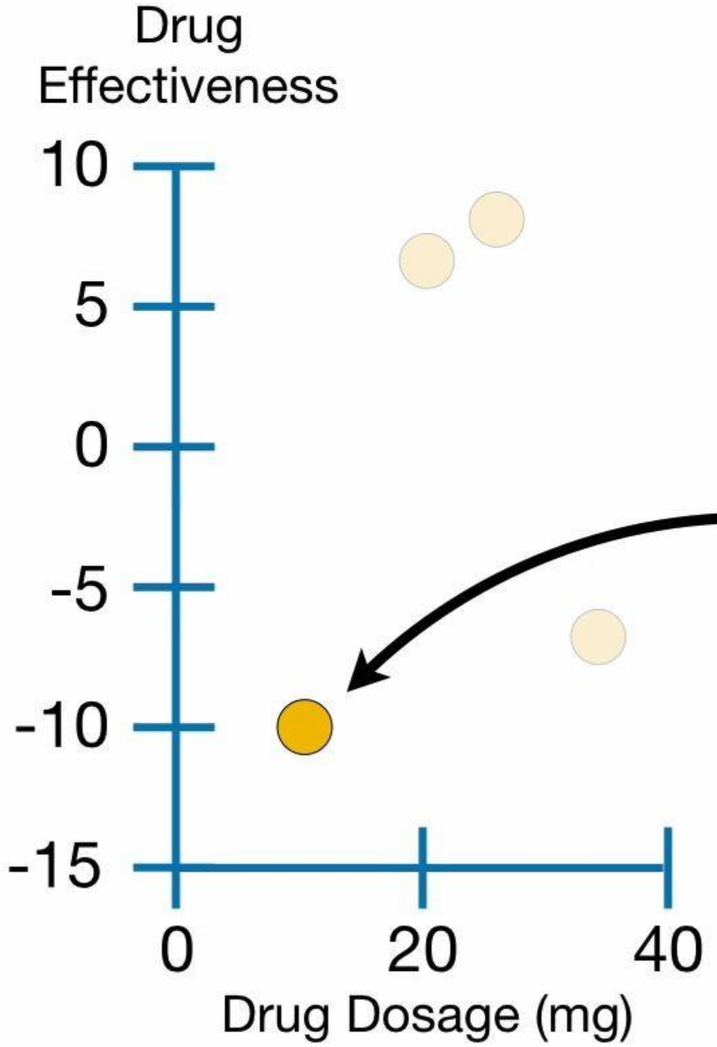
Predicted Drug Effectiveness

0.5

+

0.3 X





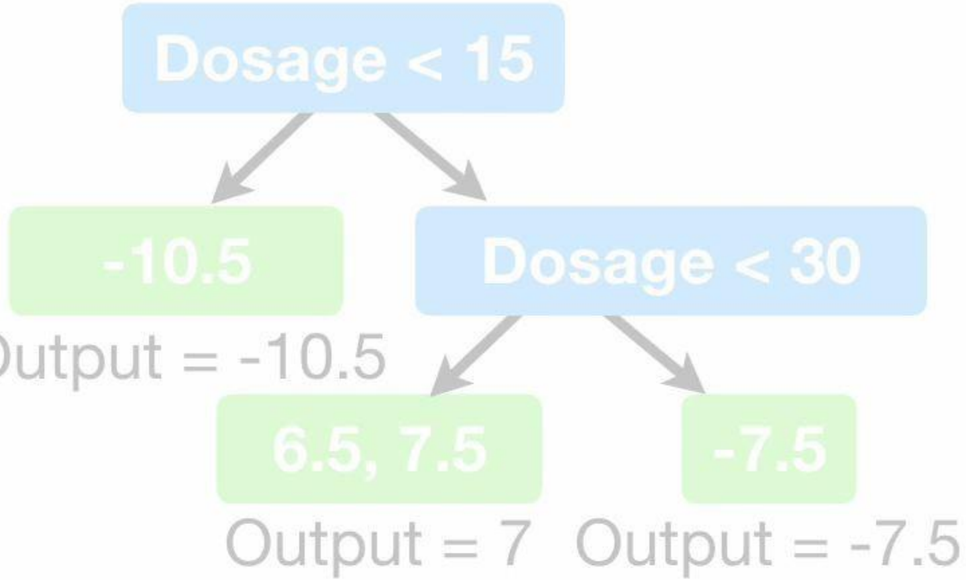
Thus, the new **Predicted** value for this observation, with **Dosage = 10**…

Predicted Drug Effectiveness

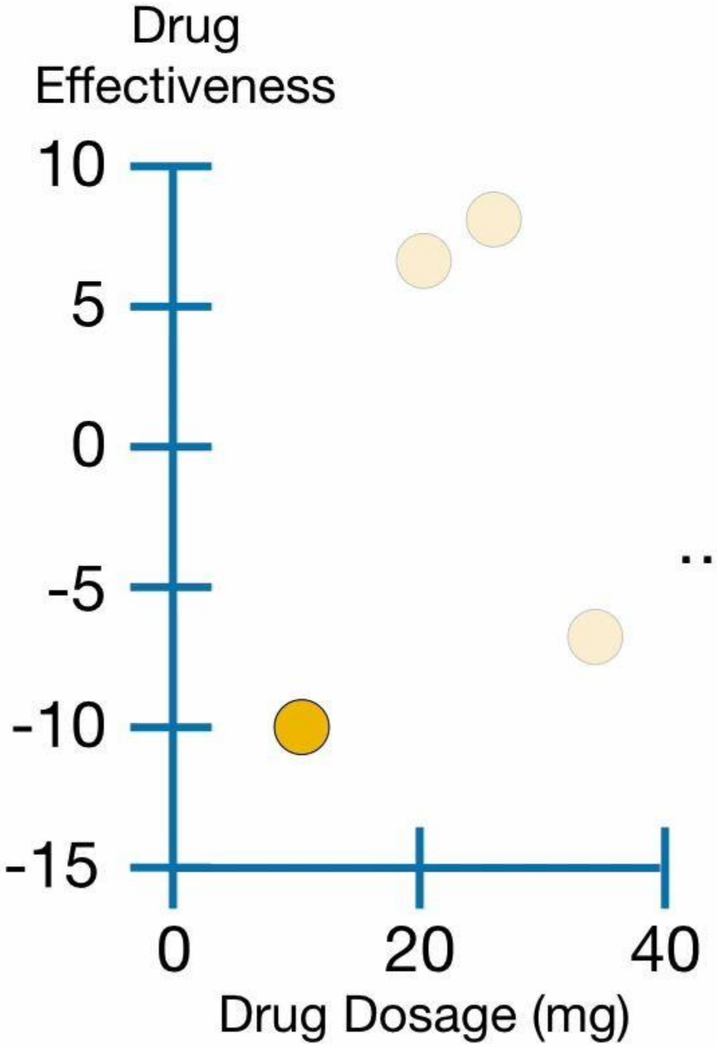
0.5

+

0.3 X



0.5 + (0.3



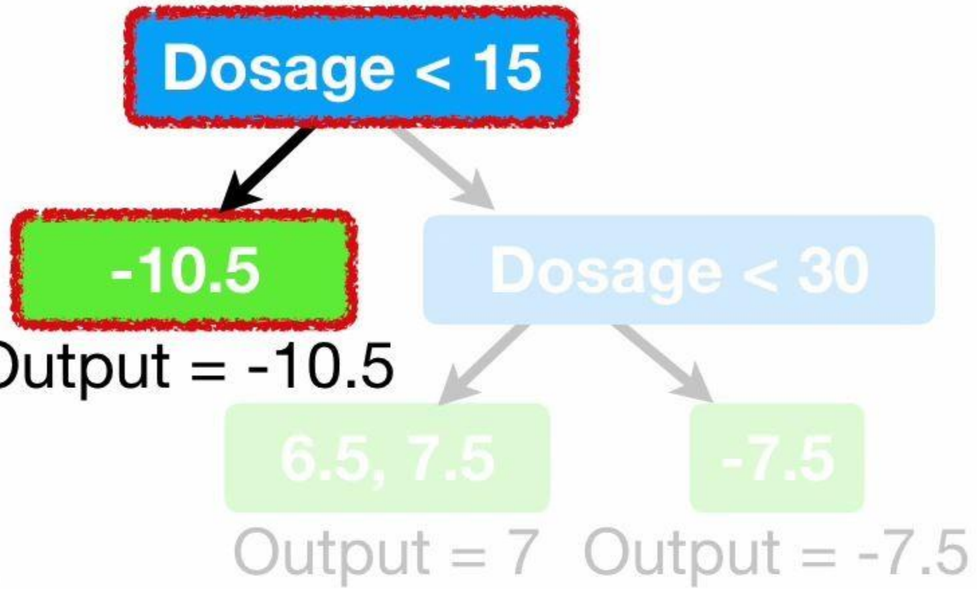
…plus the **Learning Rate, ε (eta), 0.3**…

Predicted Drug Effectiveness

0.5

+

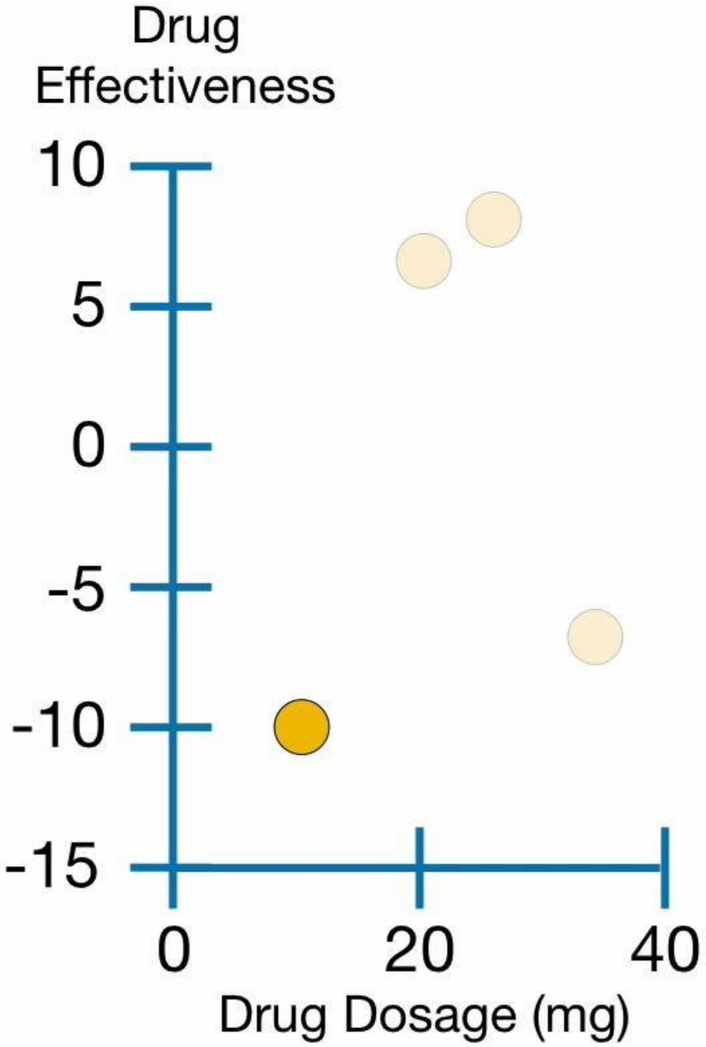
0.3 X



Output = -10.5

Output = 7

Output = -7.5



0.5 + (0.3 x -10.5) = -2.65

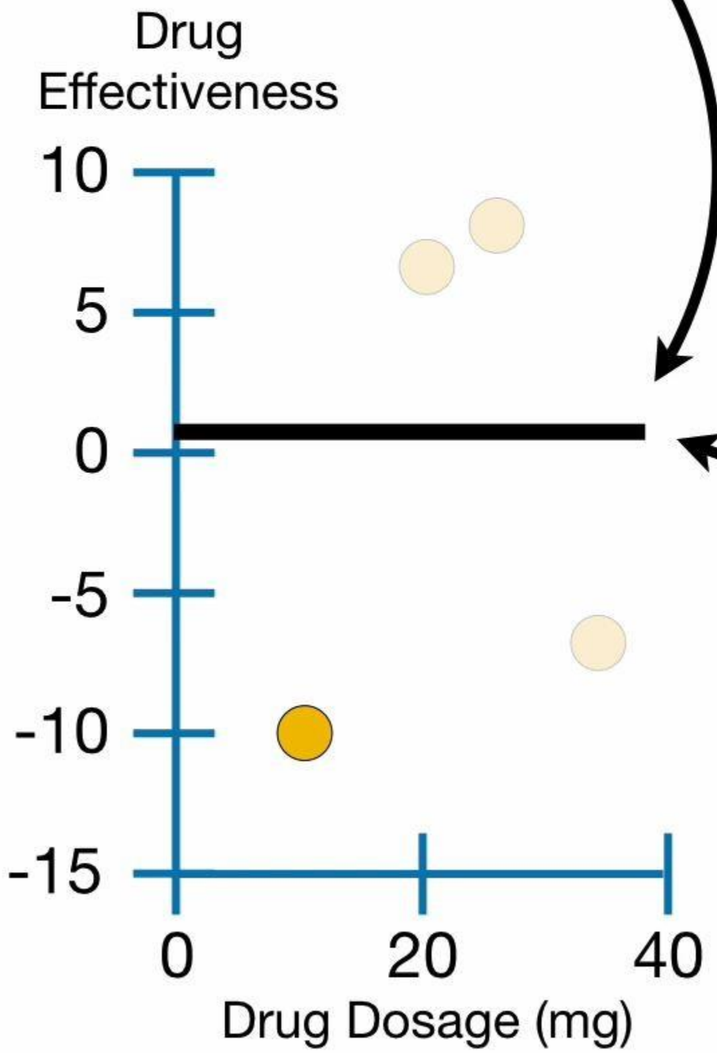
…and that gives us **-2.65**.

Predicted Drug Effectiveness

0.5

+

0.3 X

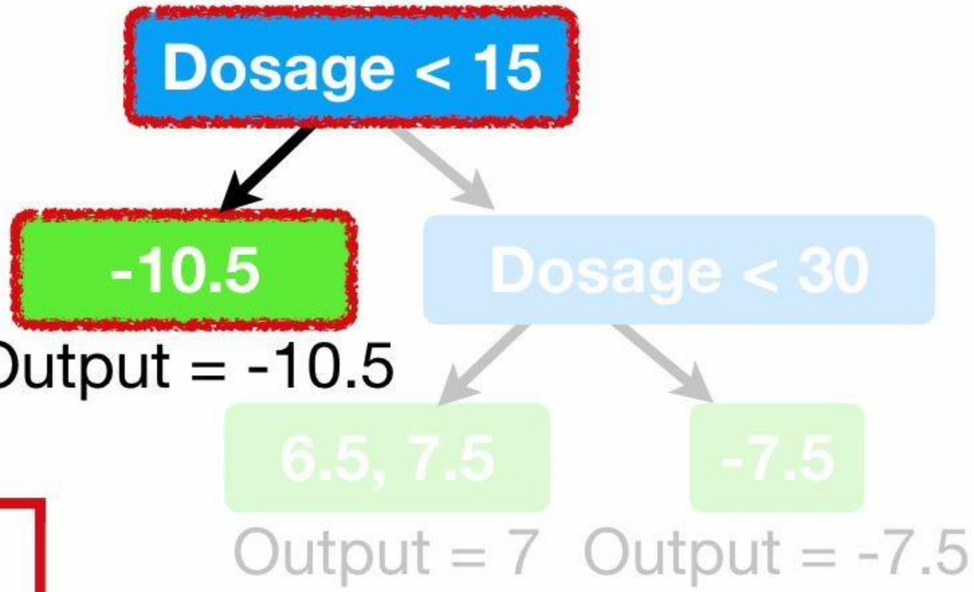


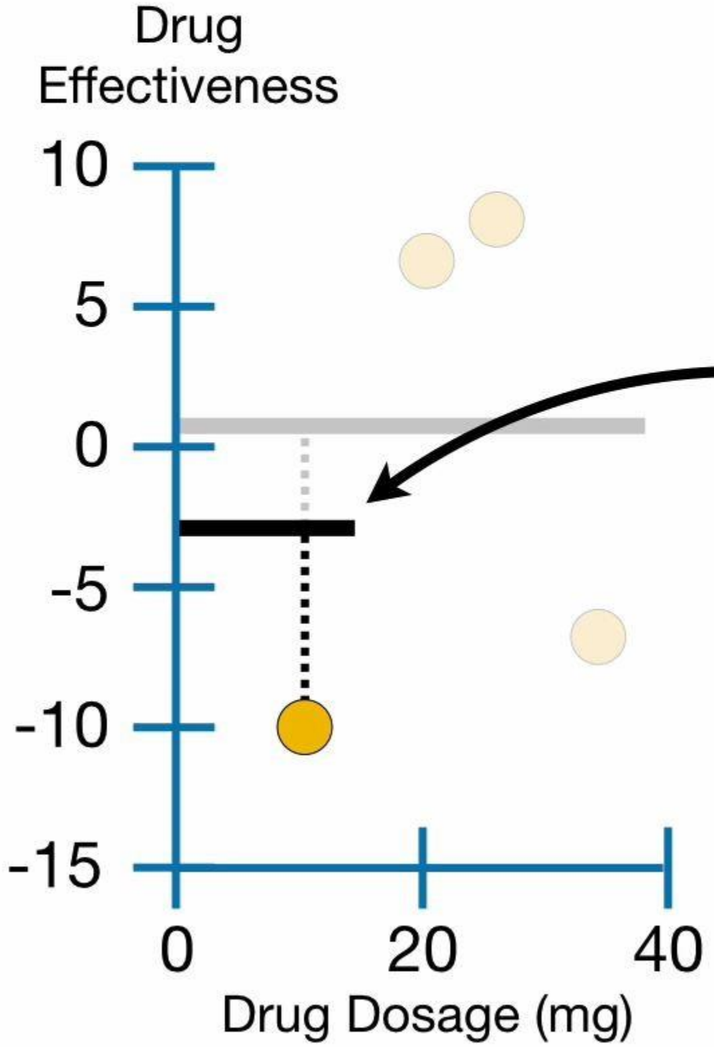
0.5 + (0.3 x -10.5) = -2.65

So, if the original **Prediction** was **0.5**…

Predicted Drug Effectiveness

0.5 + 0.3 X





0.5 + (0.3 x -10.5) = -2.65

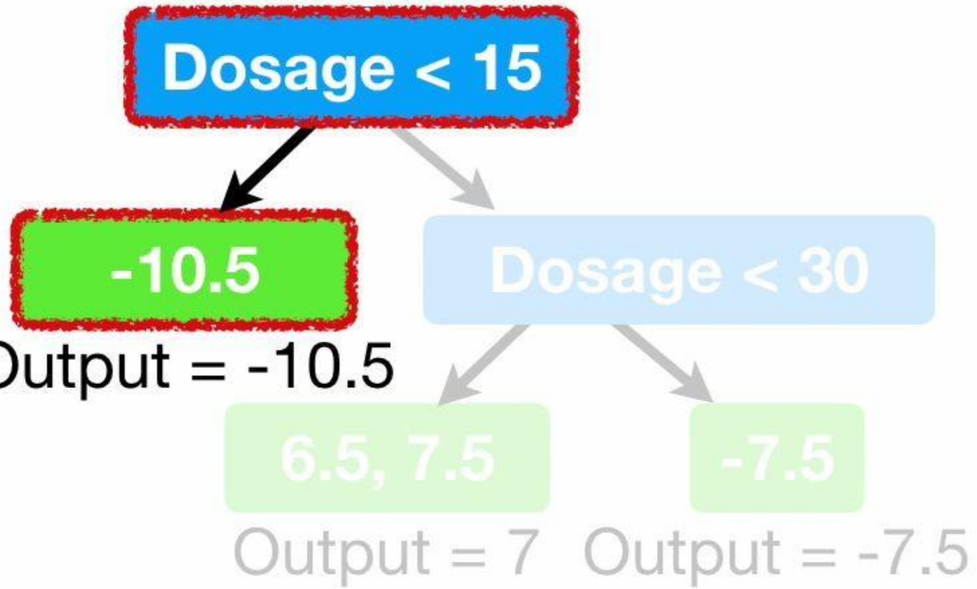
The new prediction is **-2.65**…

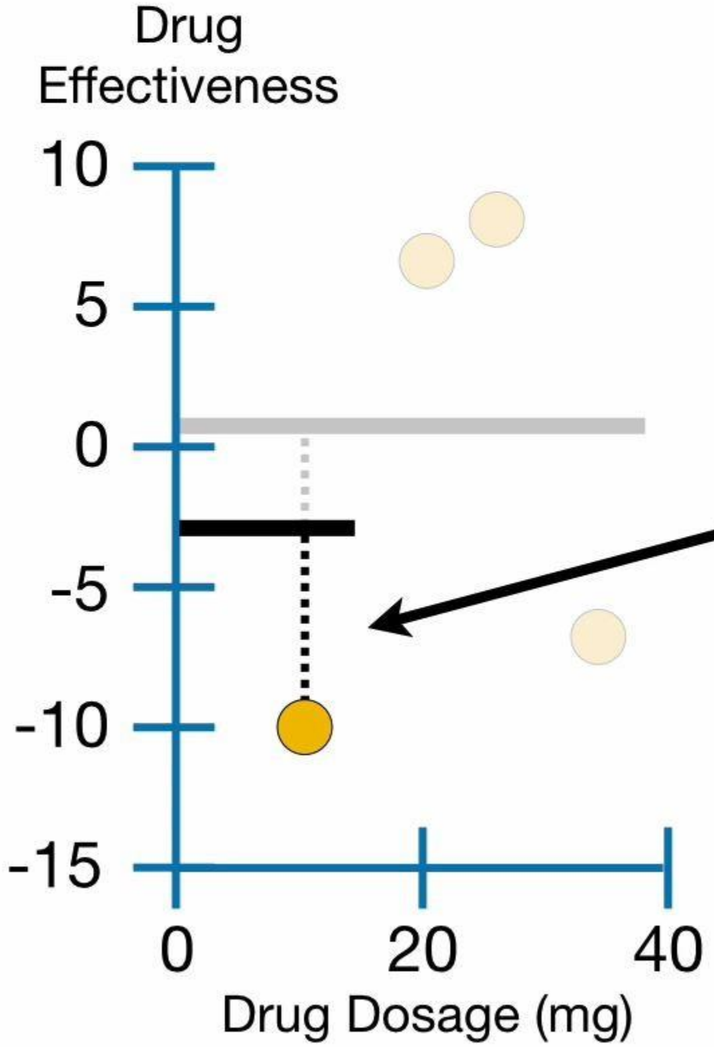
Predicted Drug Effectiveness

0.5

+

0.3 X





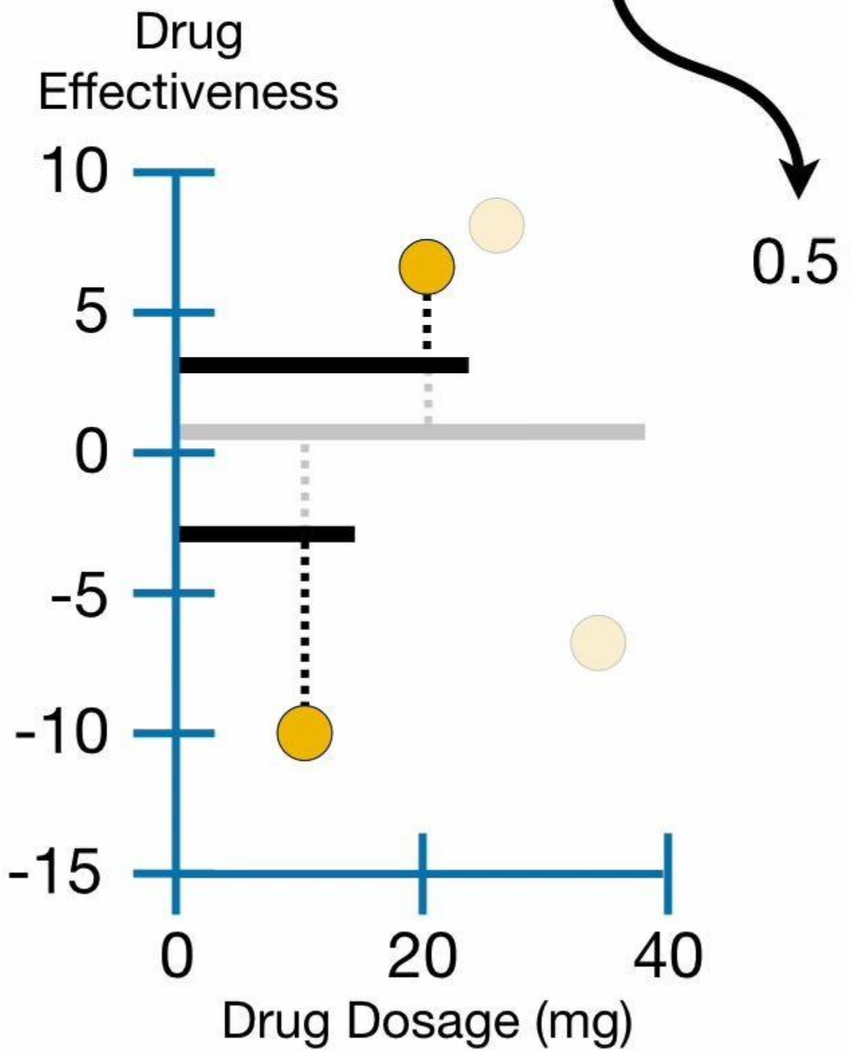
0.5 + (0.3 x -10.5) = -2.65

…and we see that the new **Residual** is smaller than before, so we've taken a small step in the right direction.

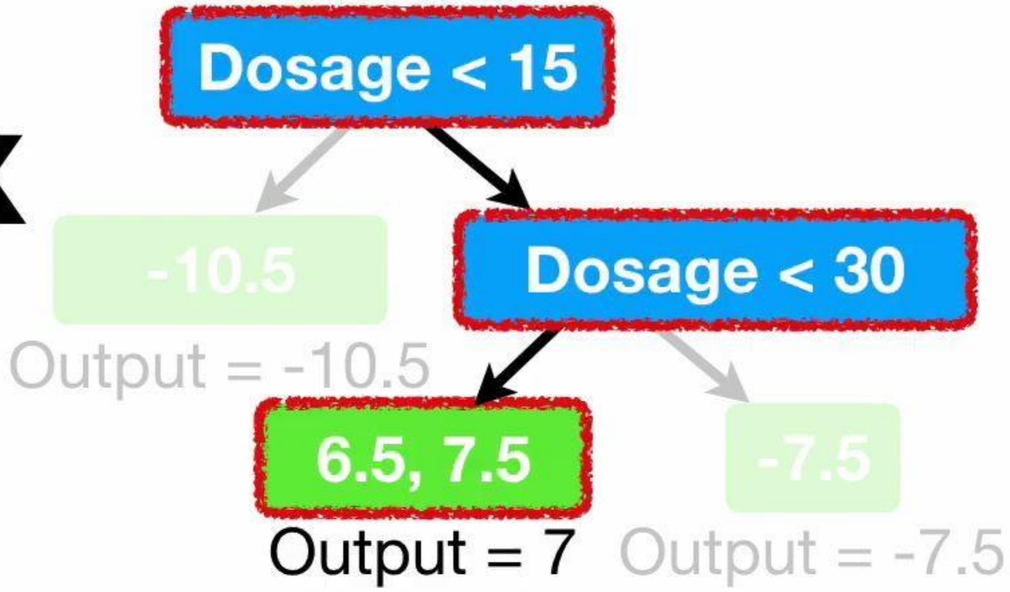
Predicted Drug Effectiveness

0.5

+



0.3 X

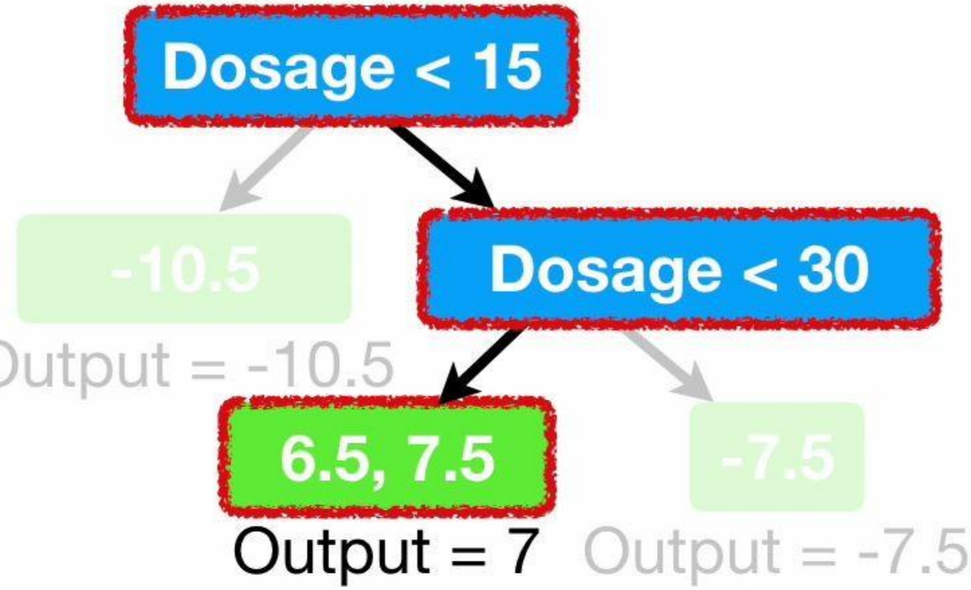


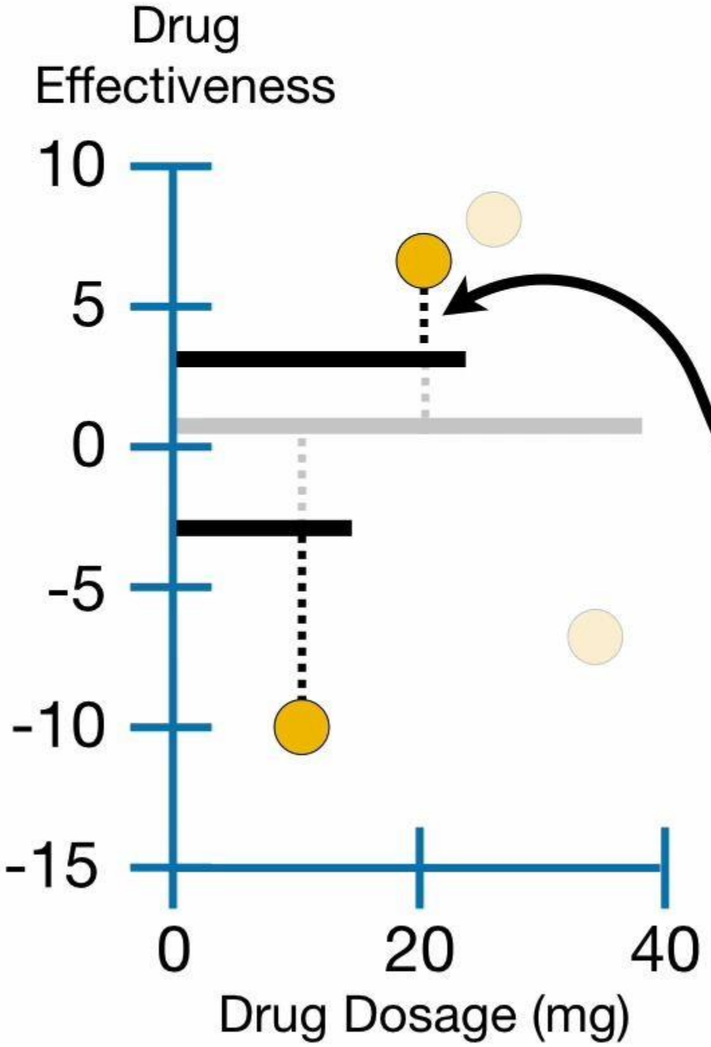
Predicted Drug Effectiveness

0.5

+

0.3 X





$0.5 + (0.3 \times 7) = 2.6$

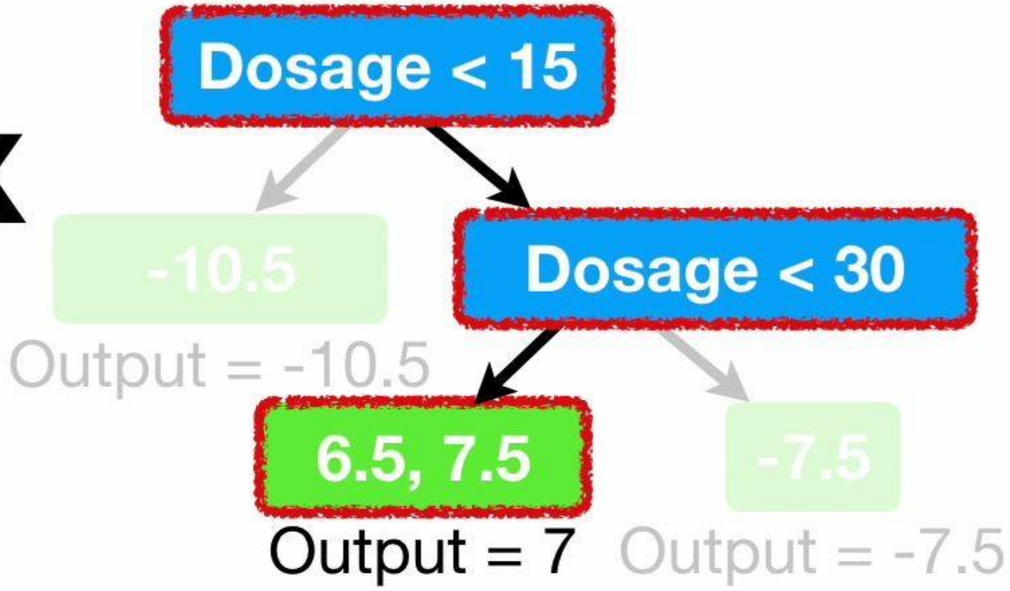
…and the new **Residual** is smaller than before, so we've taken another small step in the right direction.

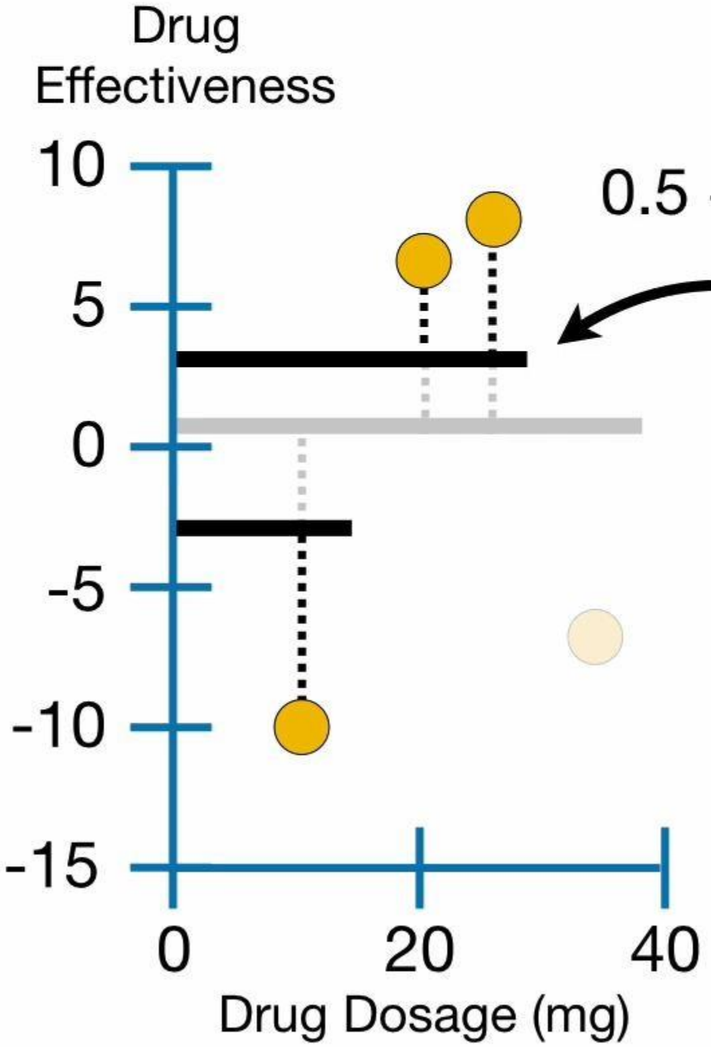
Predicted Drug Effectiveness

0.5

+

0.3 X





0.5 + (0.3 x 7) = 2.6

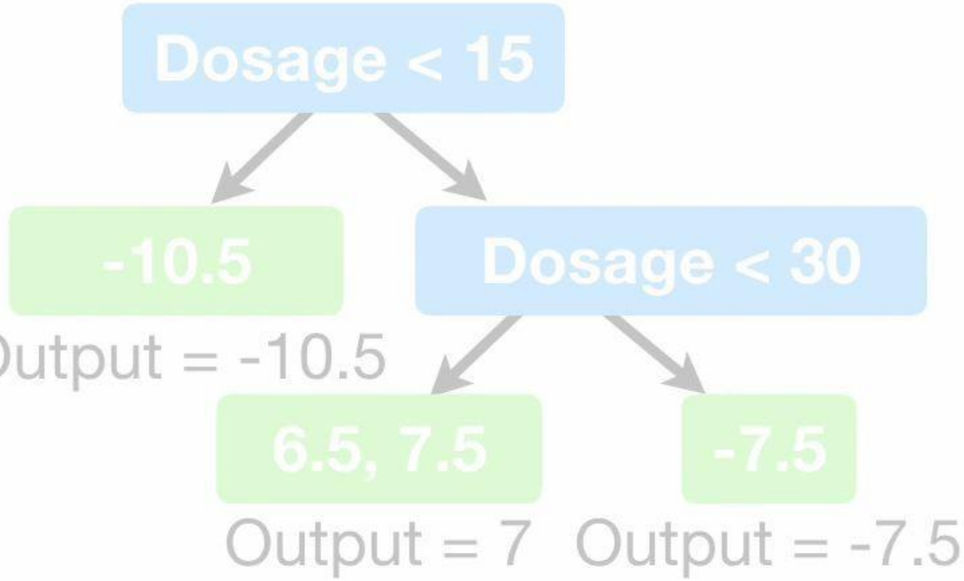
Likewise, the new predictions for the remaining observations have smaller **Residuals** than before, suggesting each small step was in the right direction.

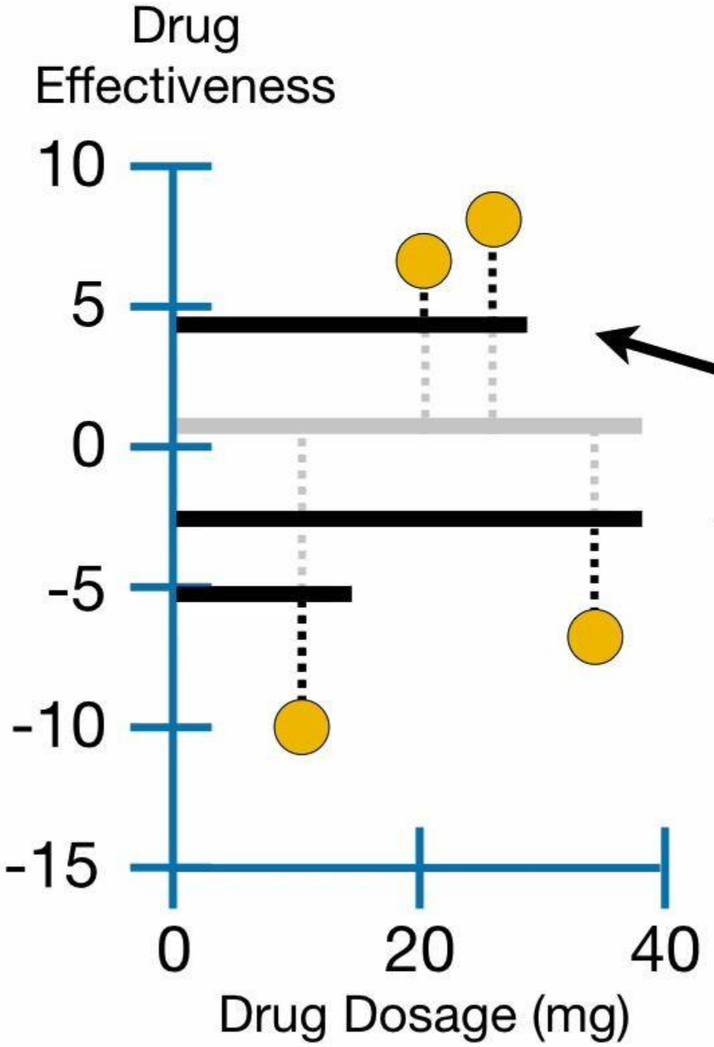
Predicted Drug Effectiveness

0.5

+

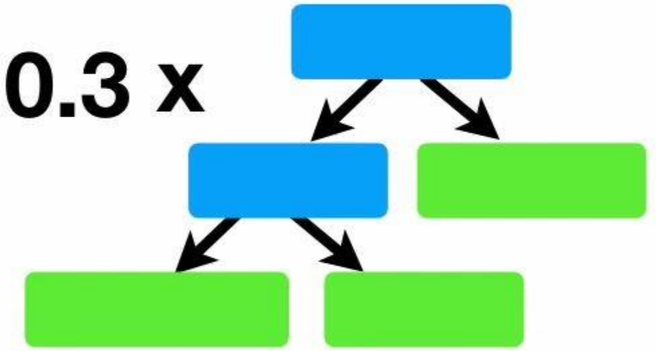
0.3 X





…and make new predictions that give us even smaller residuals…

+ 0.3 x

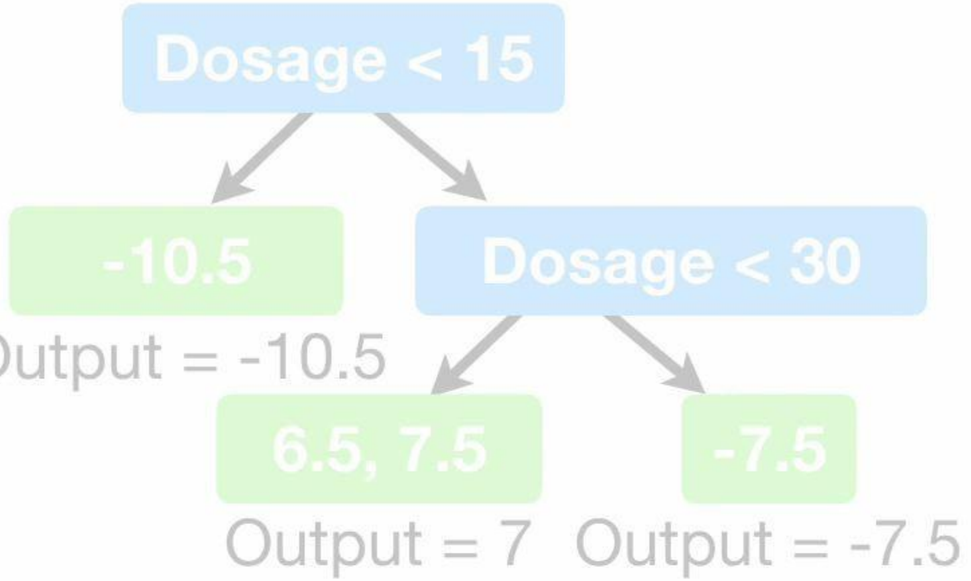


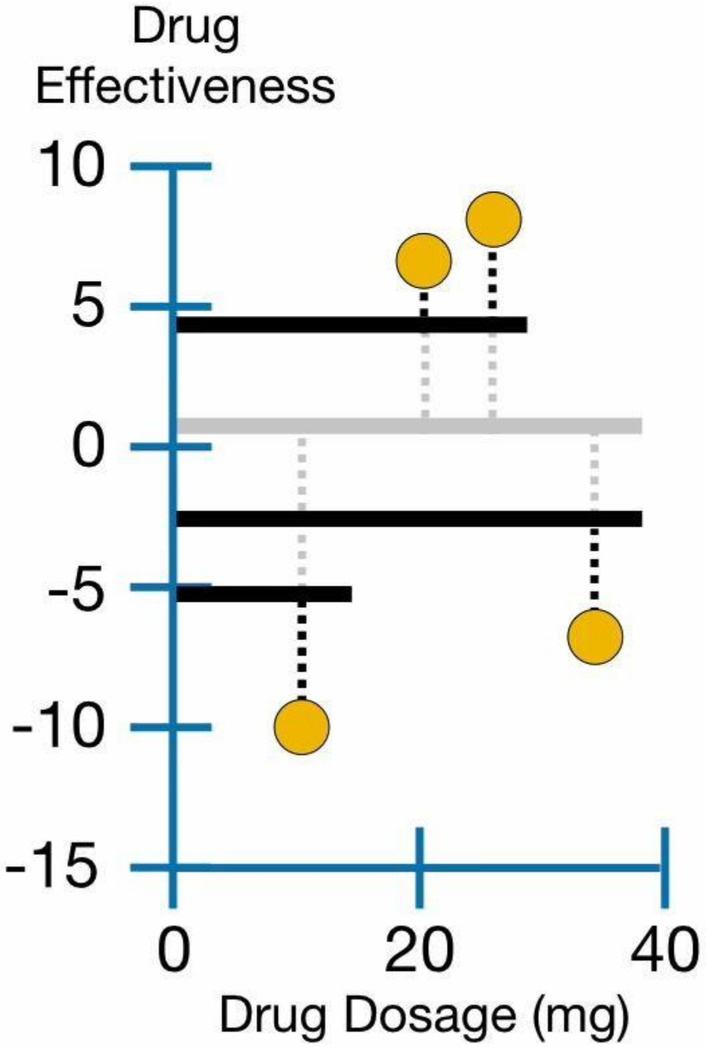
Predicted Drug Effectiveness

0.5

+

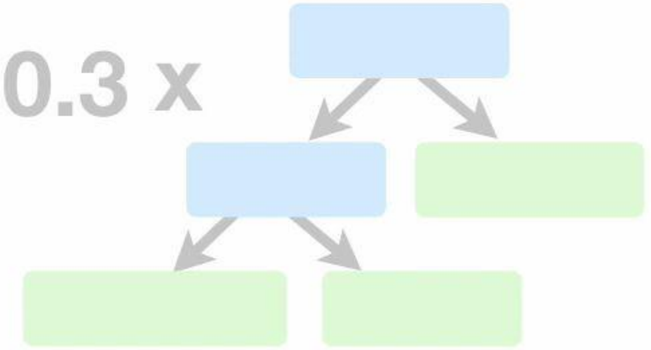
0.3 X



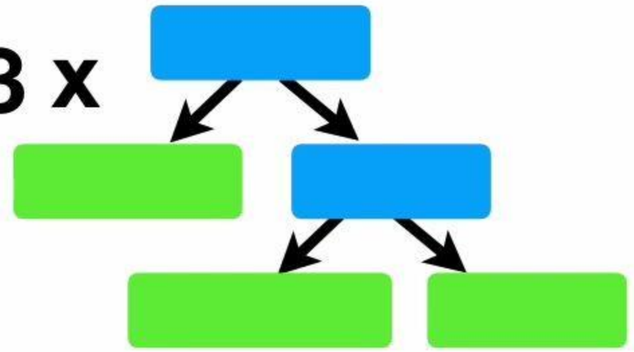


…and then build another tree based on the newest **Residuals**…

+ 0.3 x



+ 0.3 x

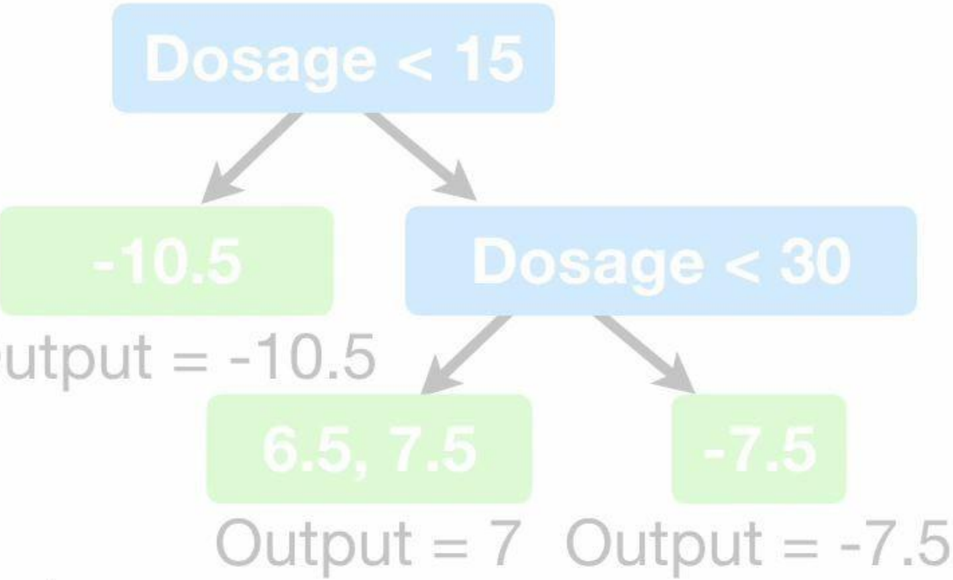


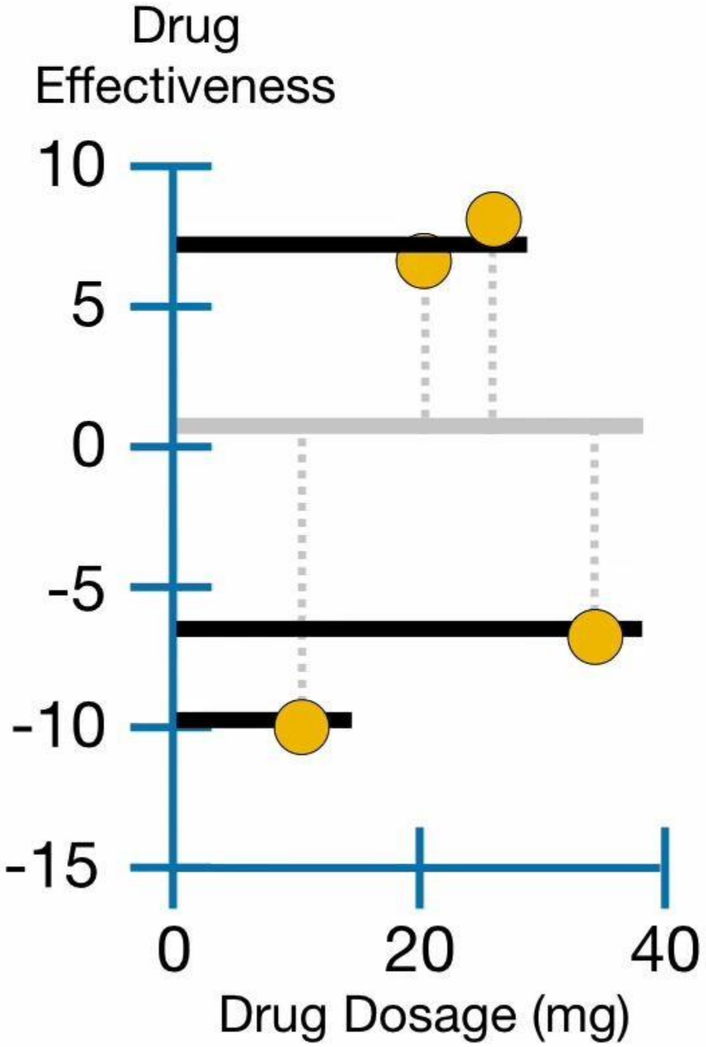
Predicted Drug Effectiveness

0.5

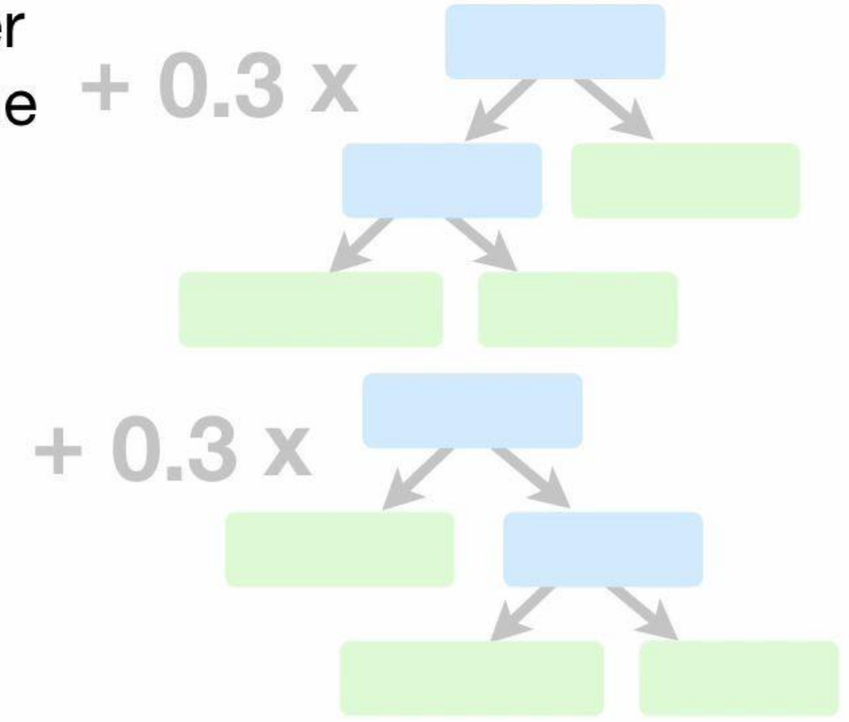
+

0.3 X



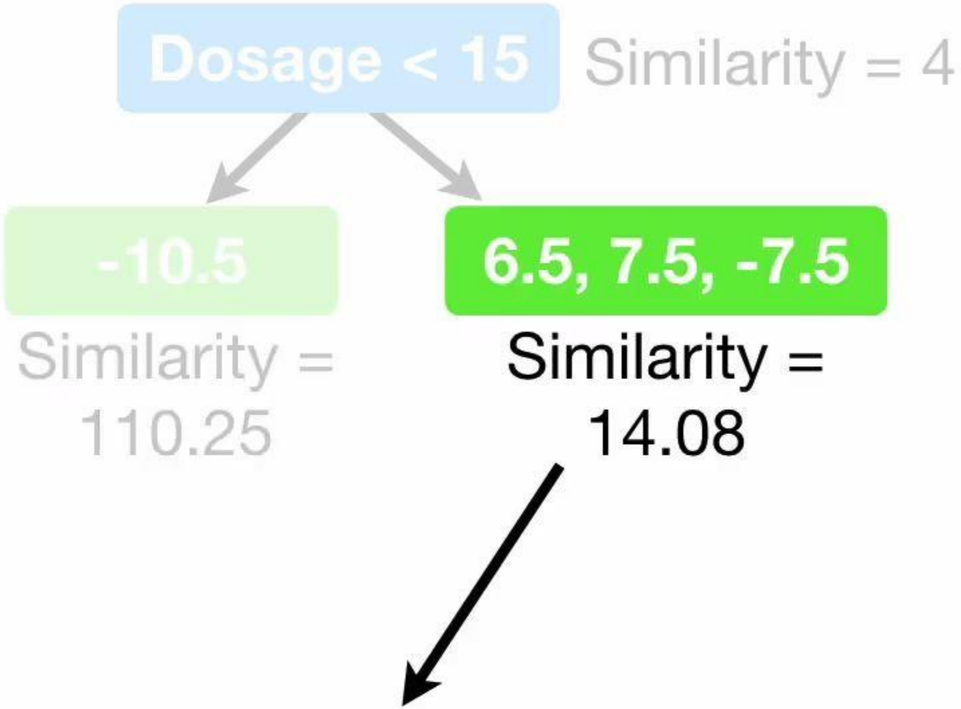


…and we keep building trees until the **Residuals** are super small, or we have reached the maximum number.



In summary, when building **XGBoost Trees** for **Regression**…

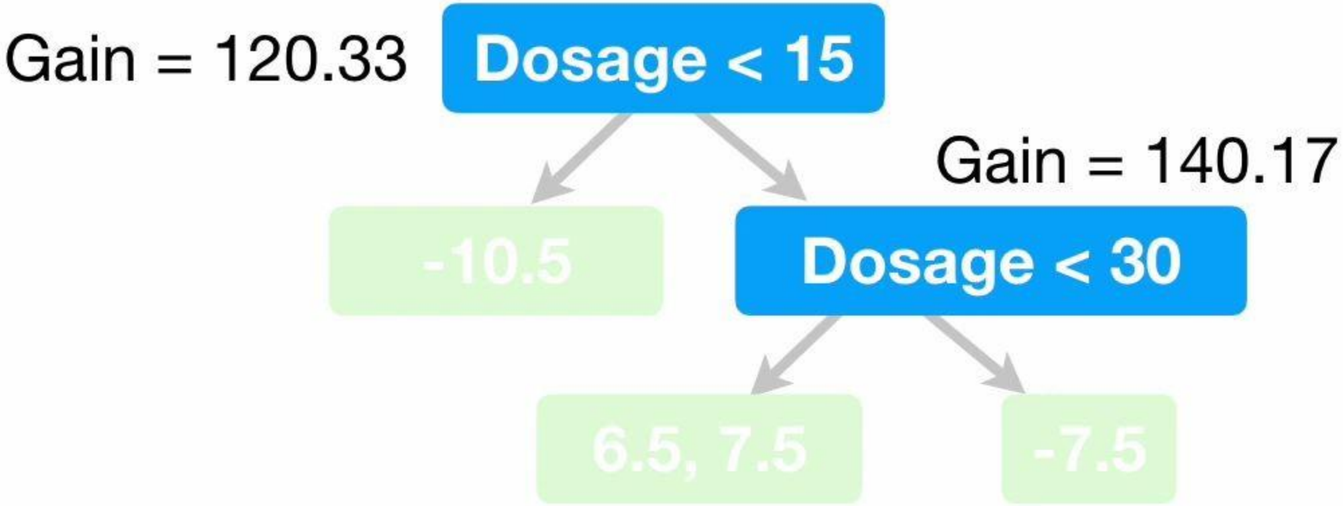
… and **Gain** to determine how to split the data…



$$Gain = Left_{Similarity} + Right_{Similarity} - Root_{Similarity}$$

…and we prune the tree by calculating the differences between **Gain** values and a user defined **Tree Complexity Parameter**, ***γ*** **(gamma)**.

Gain - ***γ*** =

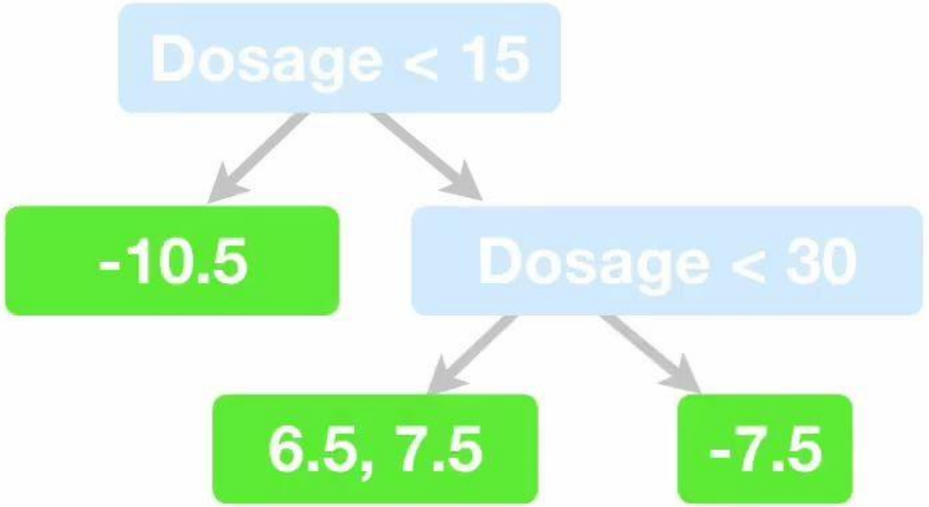


Gain - γ = { If positive, then do not prune.
If negative, then prune.

For example, if we subtract ***γ*** **(gamma)** from this **Gain** and get a negative value, we will prune, otherwise we're done.

$$\text{Gain} - \gamma = \begin{cases} \text{If positive, then do not prune.} \\ \\ \text{If negative, then prune.} \end{cases}$$

Then we calculate the **Output Values** for the remaining leaves…



$$\text{Output Value} = \frac{\text{Sum of Residuals}}{\text{Number of Residuals} + \lambda}$$

…and lastly, **λ (lambda)** is a **Regularization Parameter** and when **λ > 0**, it results in more pruning, by shrinking the **Similarity Scores**, and it results in smaller **Output Values** for the leaves.

…and lastly, **λ (lambda)** is a **Regularization Parameter** and when **λ > 0**, it results in more pruning, by shrinking the **Similarity Scores**, and it results in smaller **Output Values** for the leaves.

$$\text{Similarity Score} = \frac{\text{Sum of Residuals, Squared}}{\text{Number of Residuals} + \lambda}$$

$$\text{Output Value} = \frac{\text{Sum of Residuals}}{\text{Number of Residuals} + \lambda}$$

The End!!!