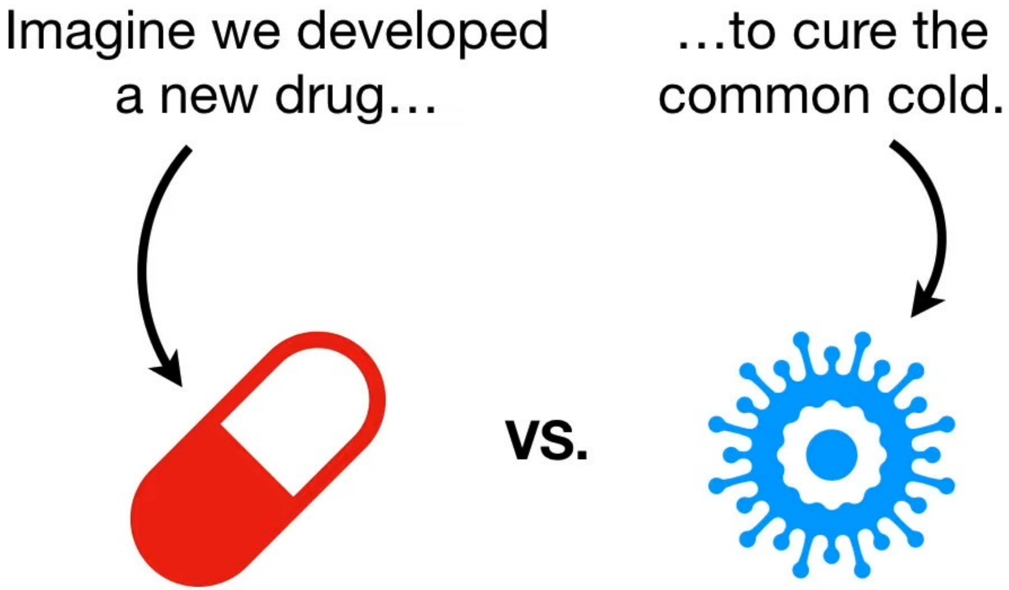
Regression Tree



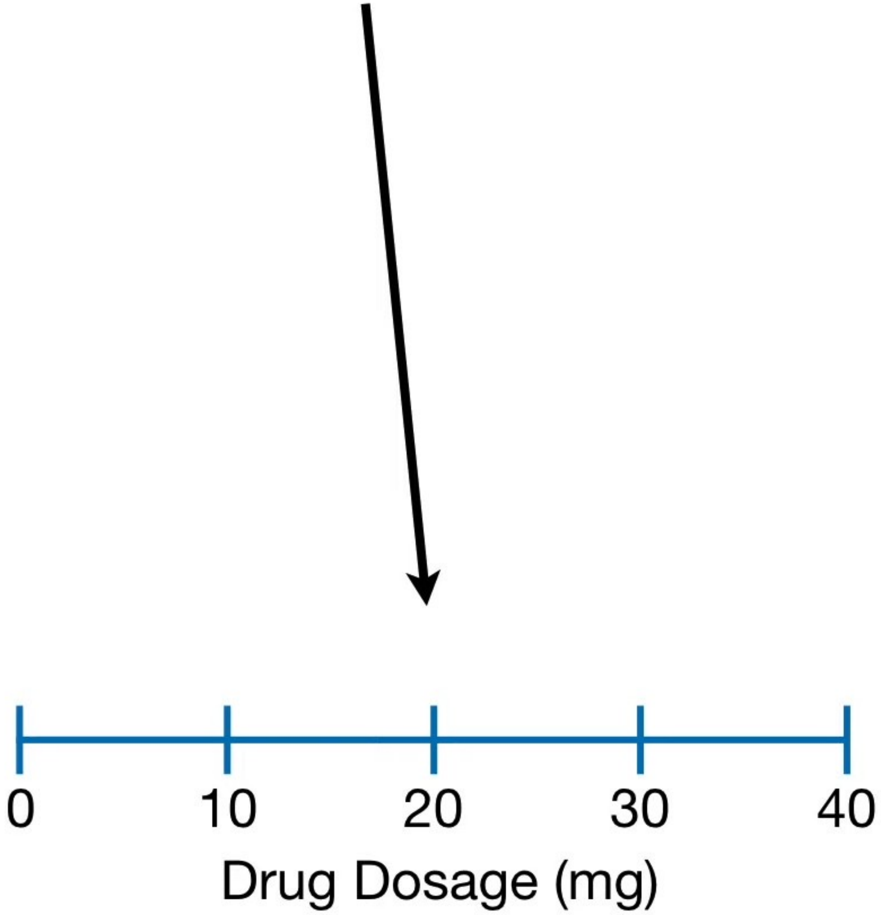
Imagine we developed a new drug…

…to cure the common cold.

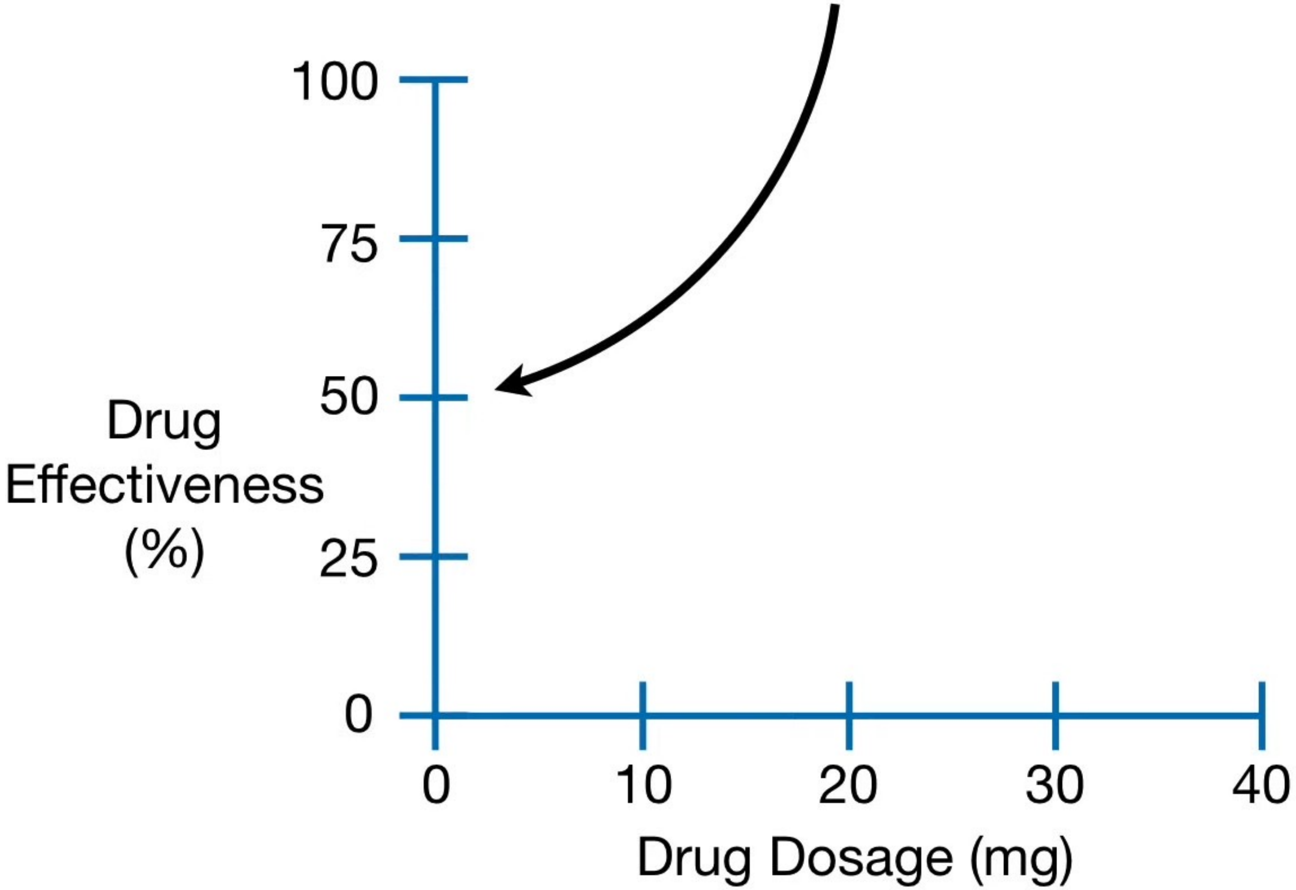
vs.

However, we don’t know the optimal dosage to give to patients.

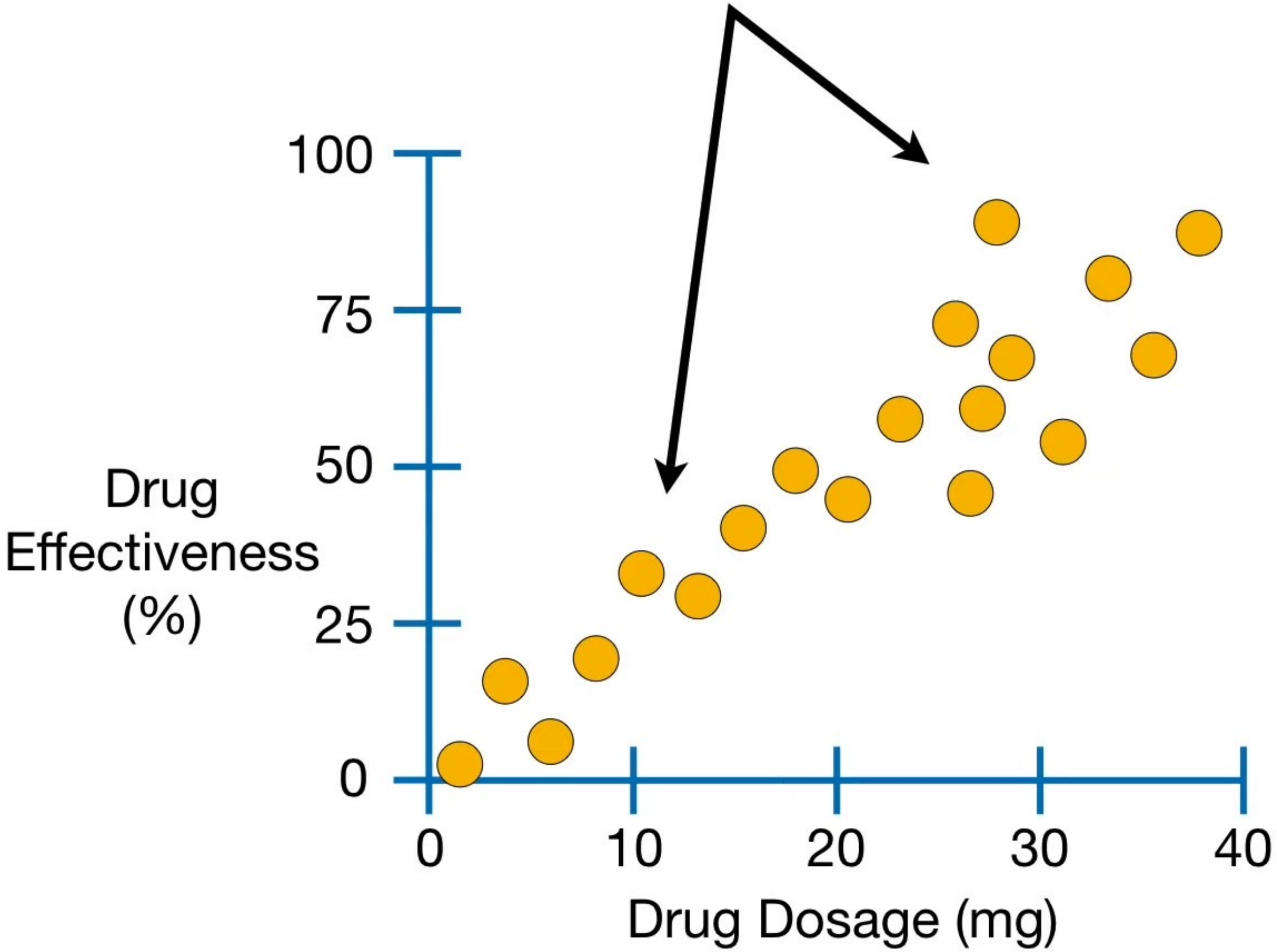
So we do a clinical trial with different dosages…



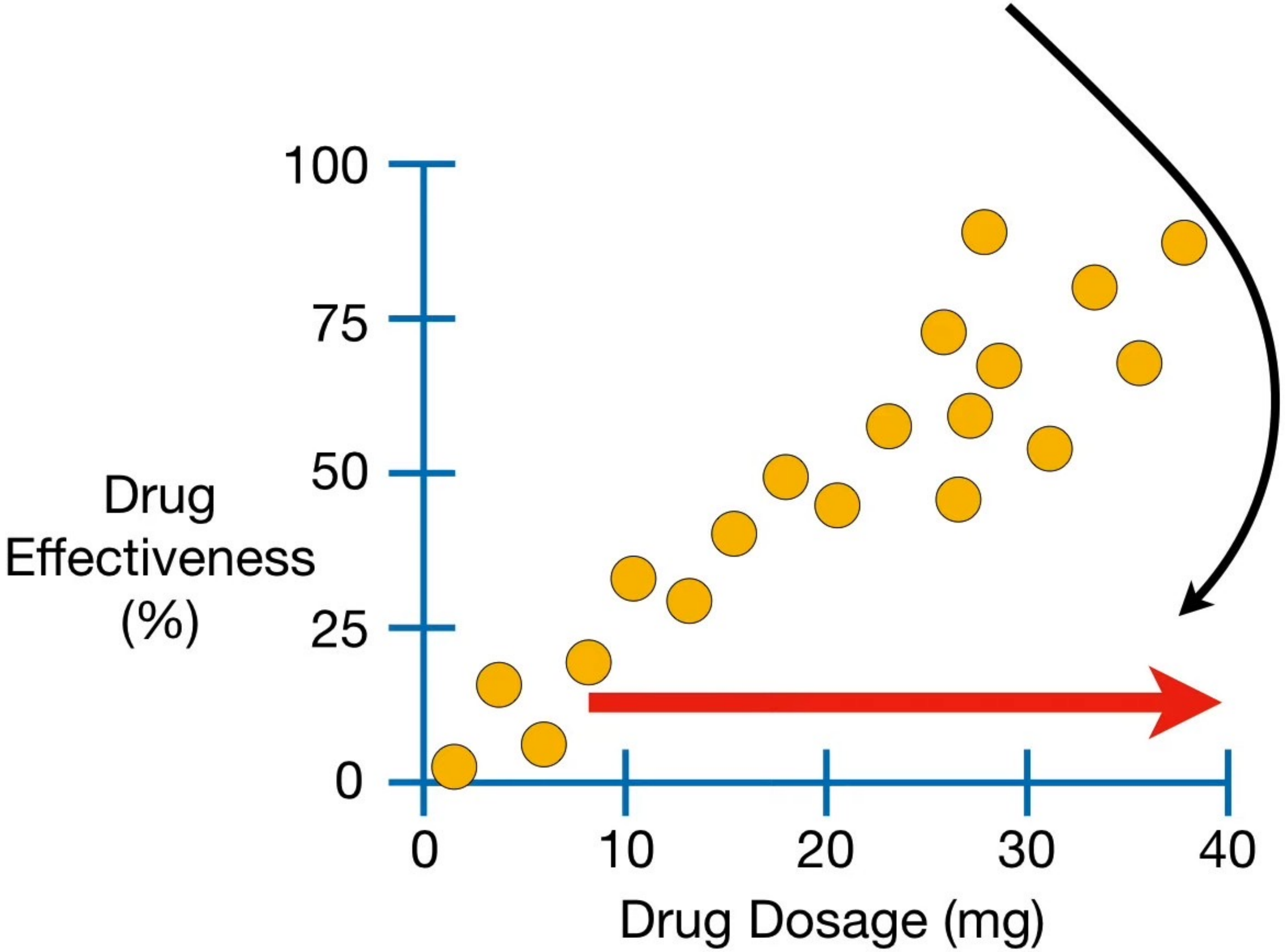
…and measure how effective each dosage is.



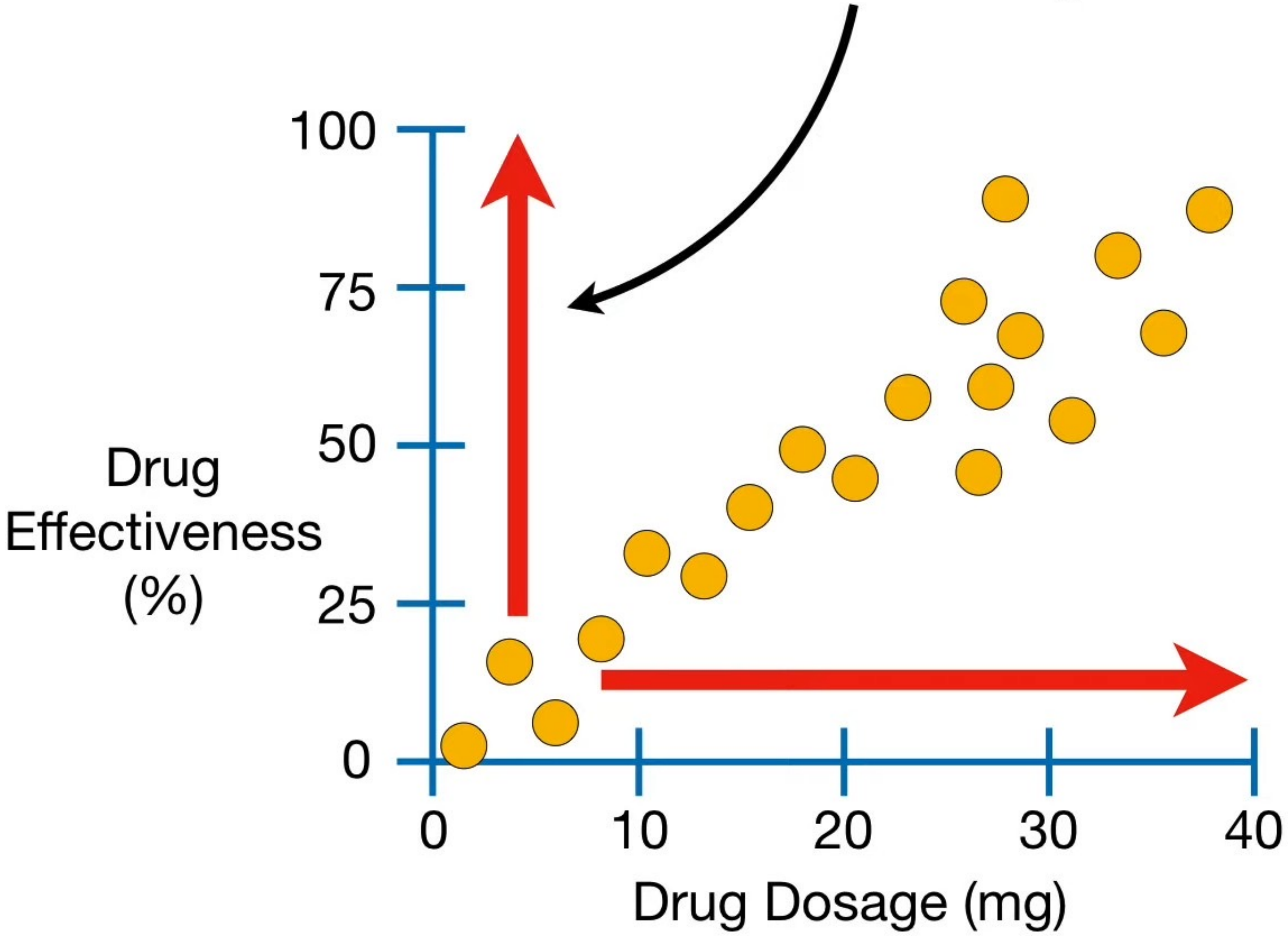
If the data looked like this…



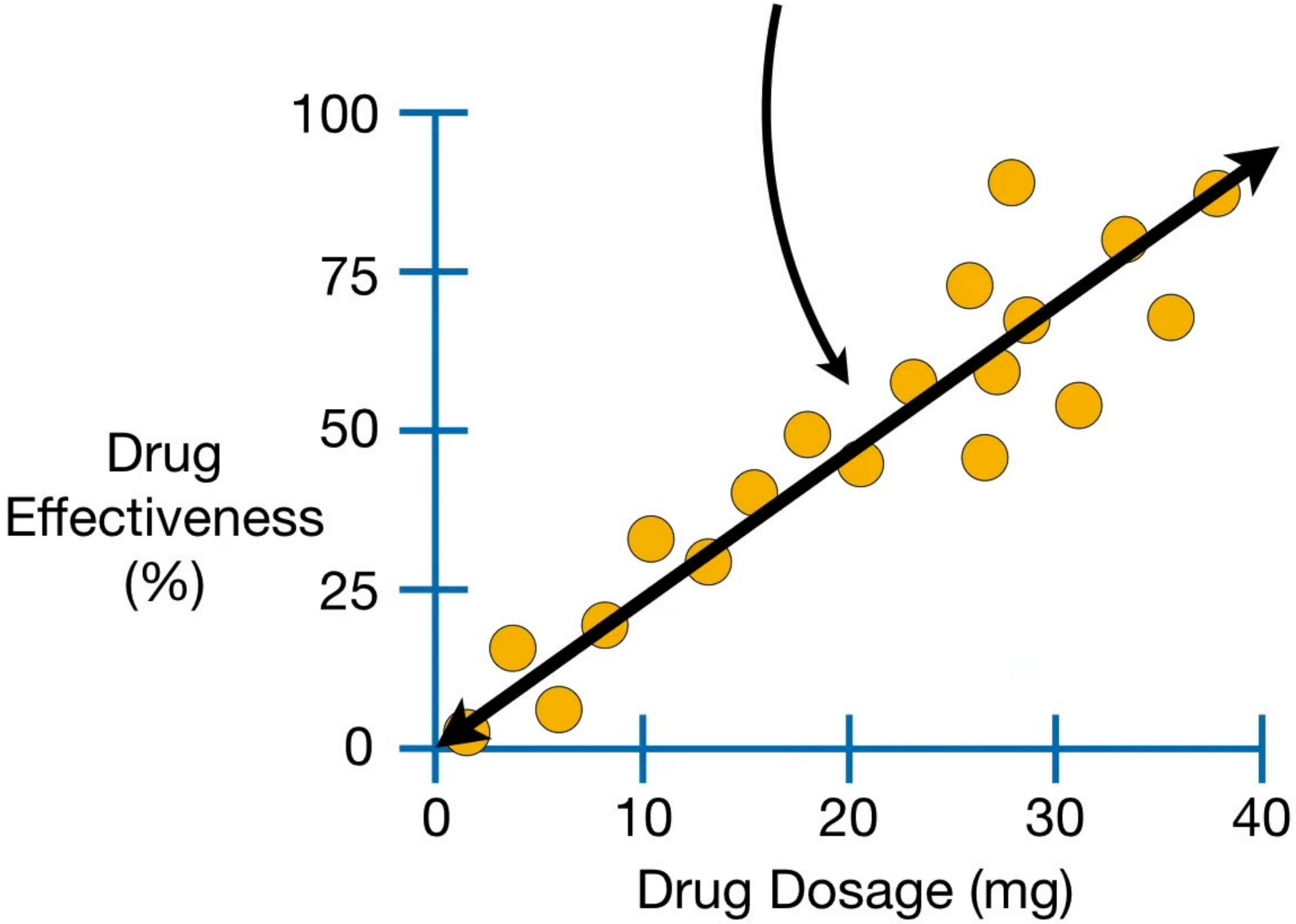
…and, in general, the higher the dose,



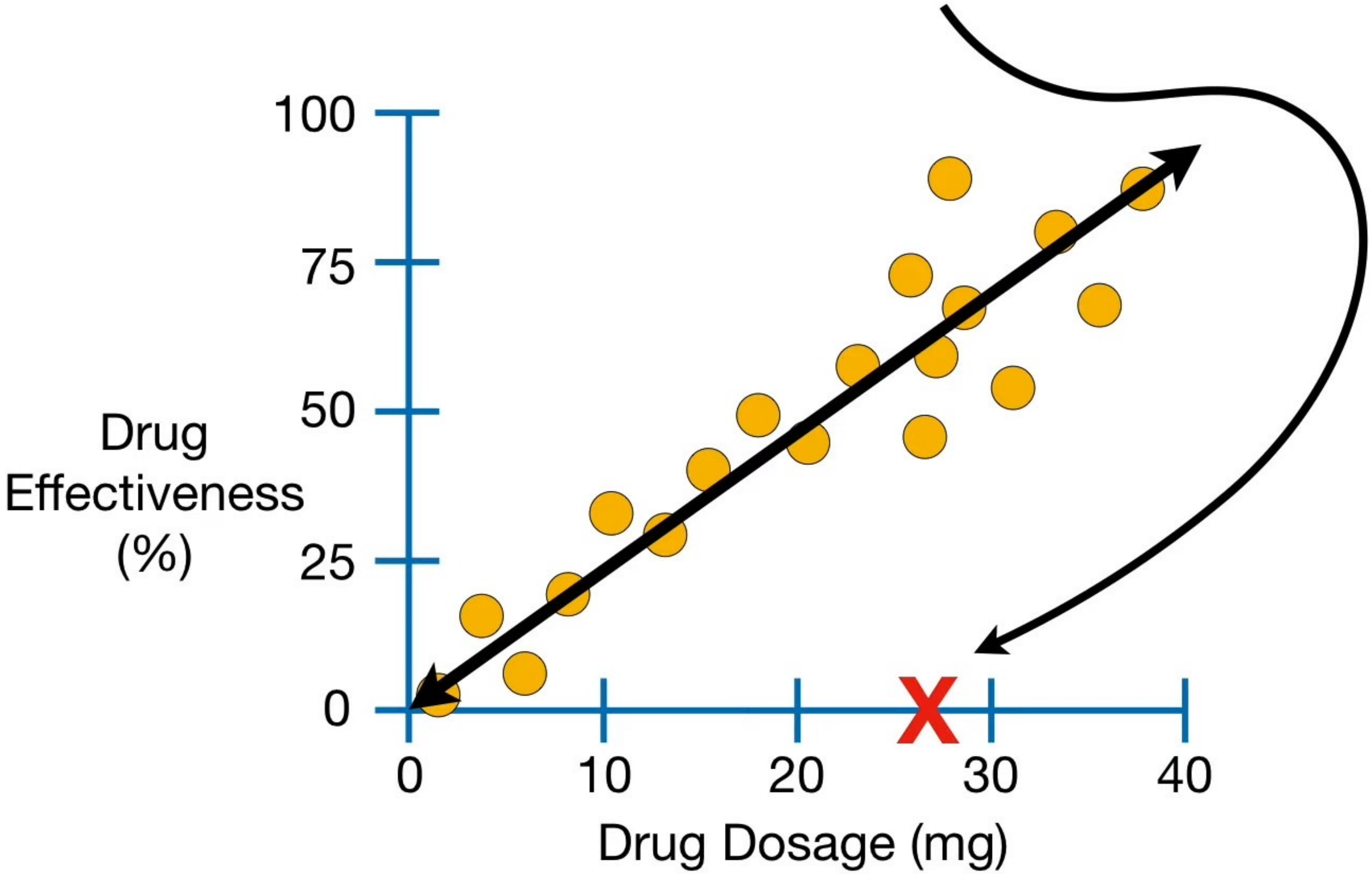
…and, in general, the higher the dose, the more effective the drug…



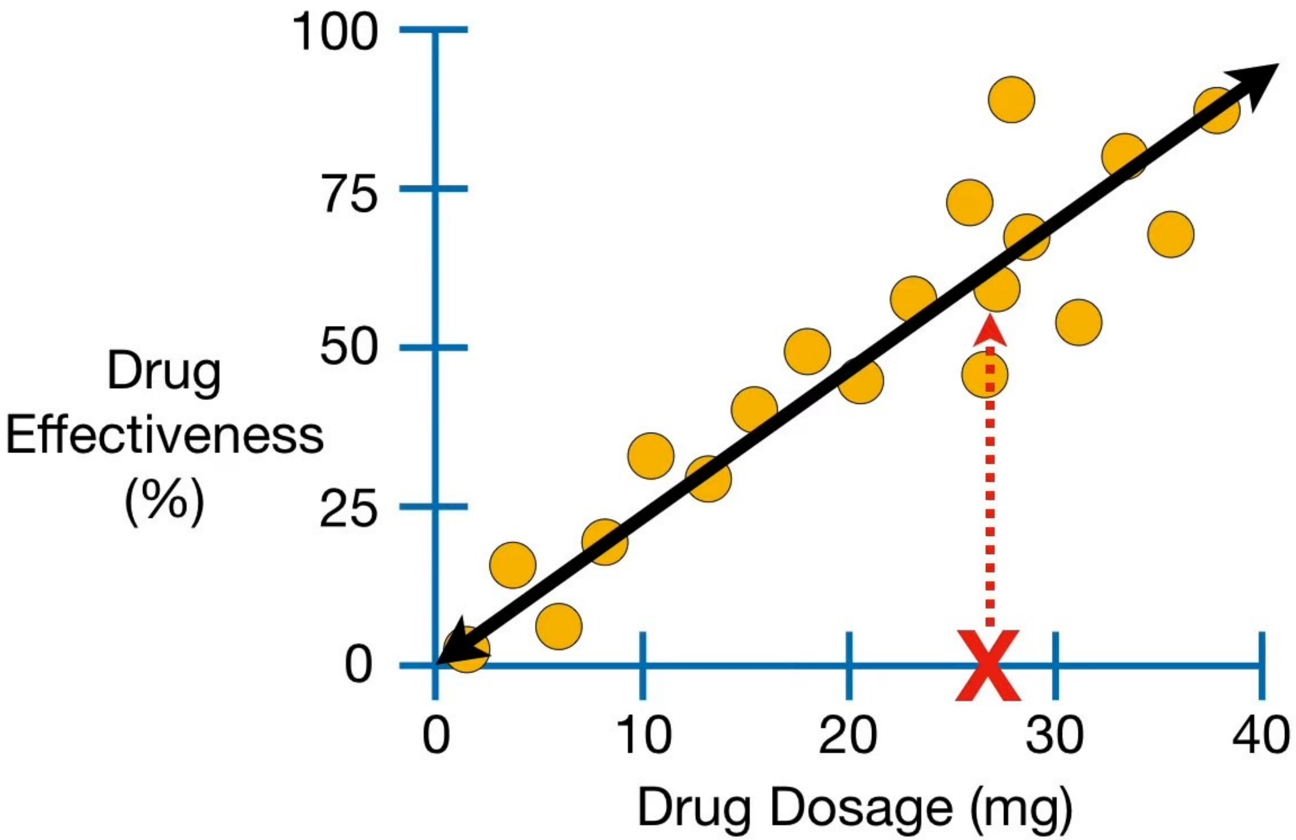
…then we could easily fit a line to the data…



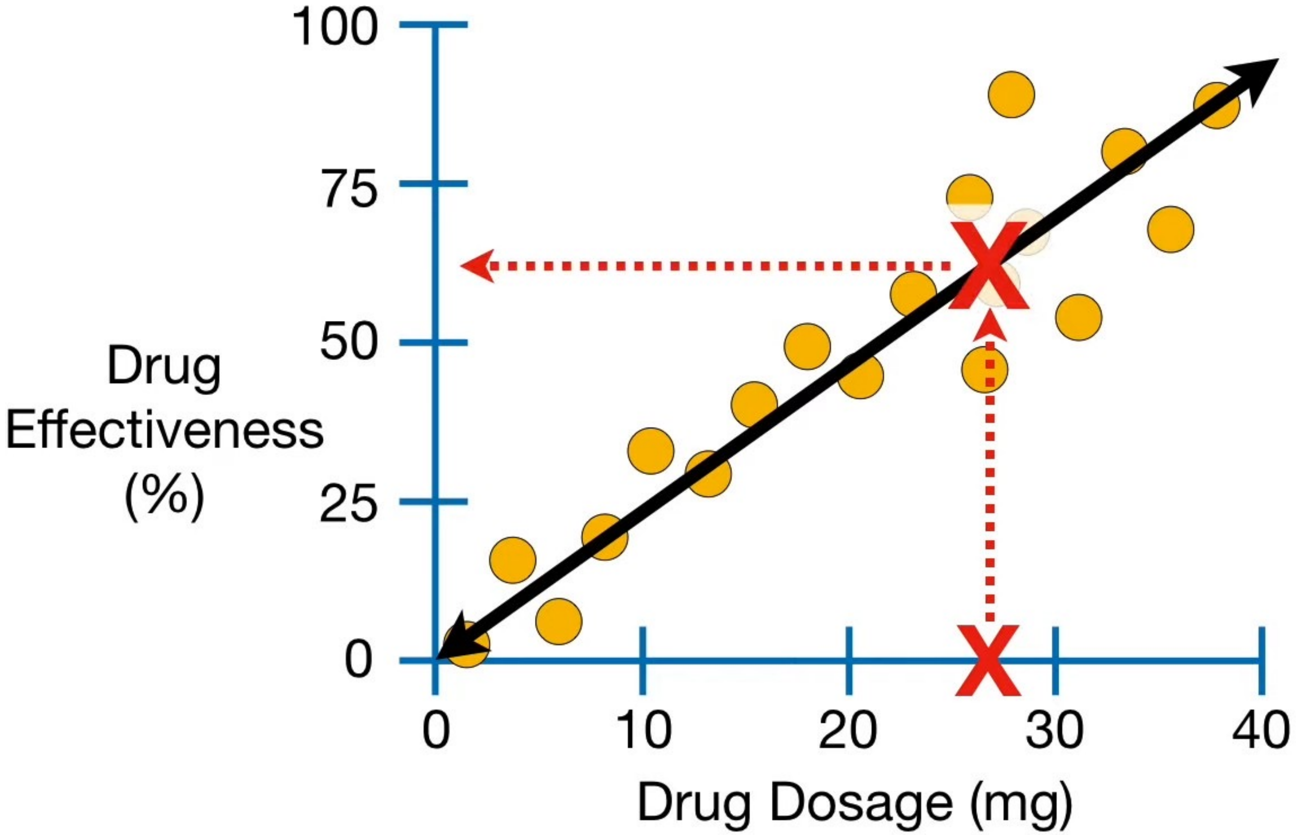
…and if someone told us they were taking a **27 mg Dose**…



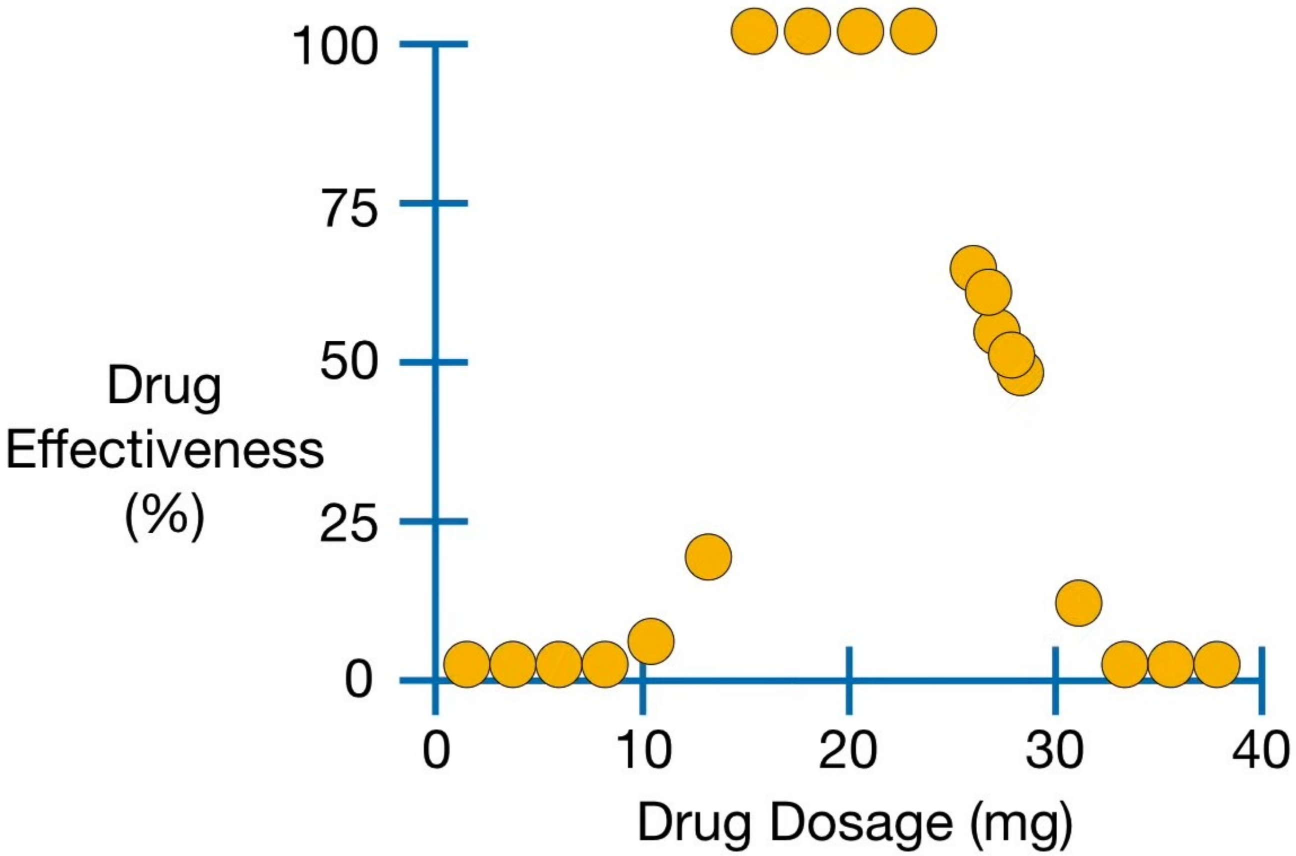
…we could use the line to predict that a **27 mg Dose** should be **62% Effective**.



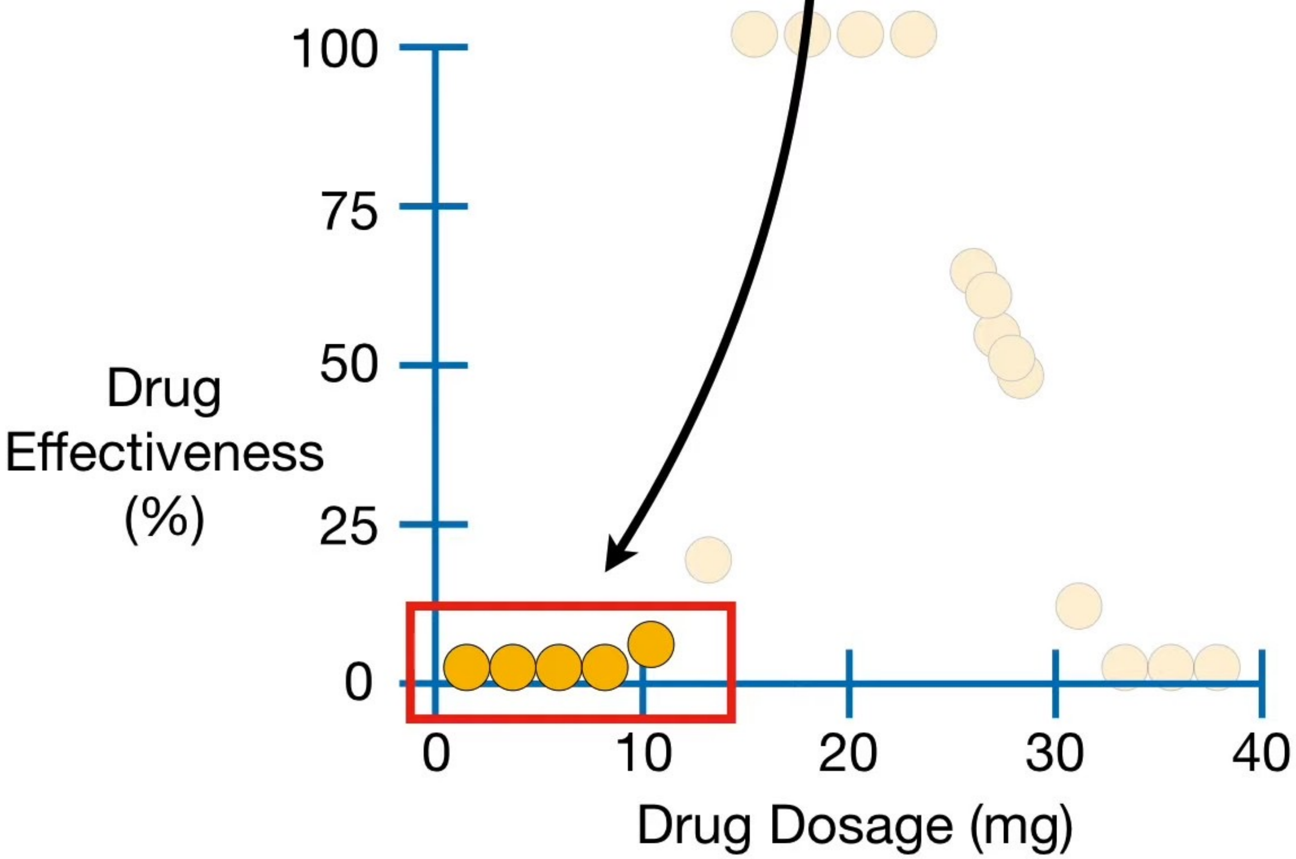
…we could use the line to predict that a **27 mg Dose** should be **62% Effective**.



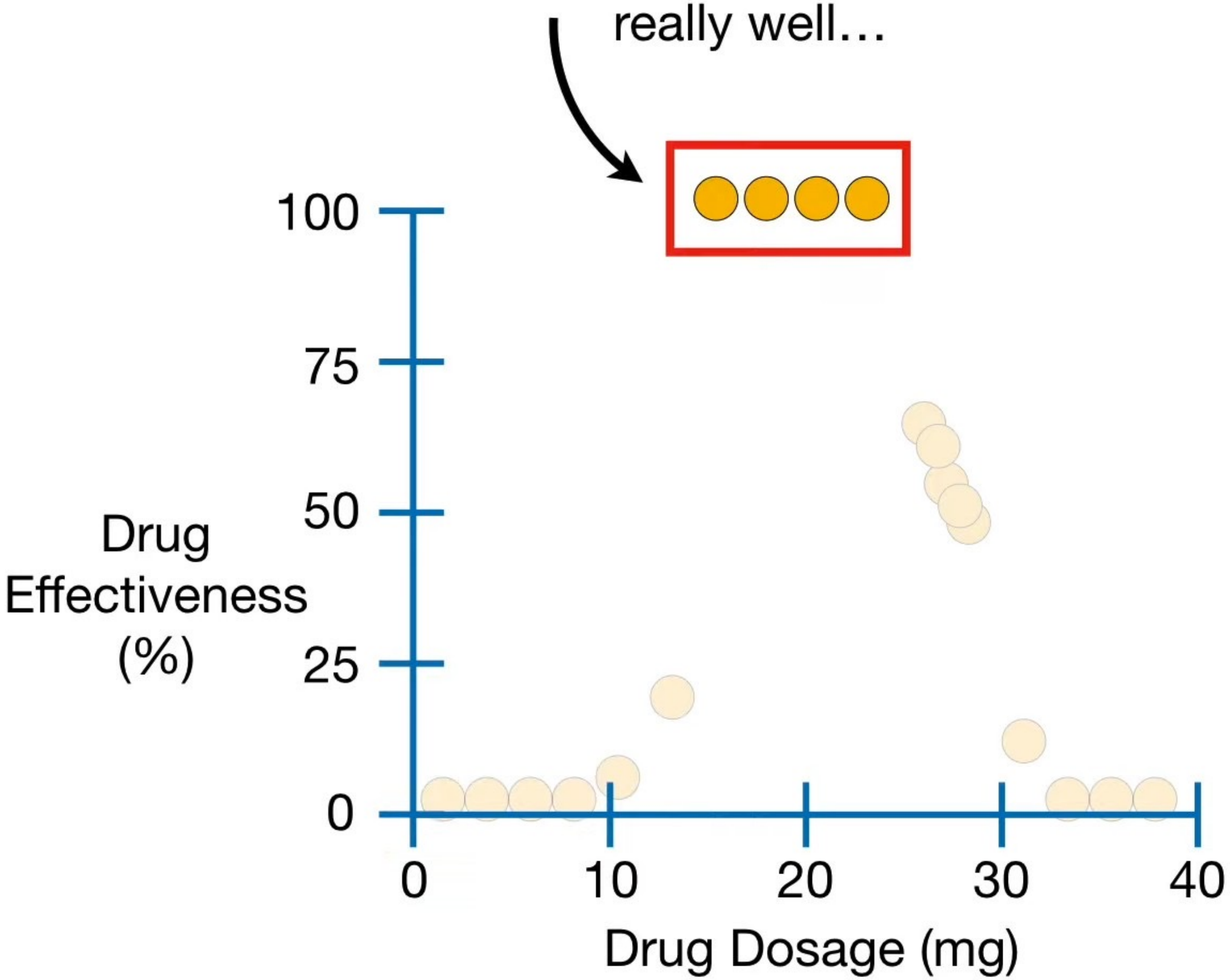
However, what if the data looked like this?



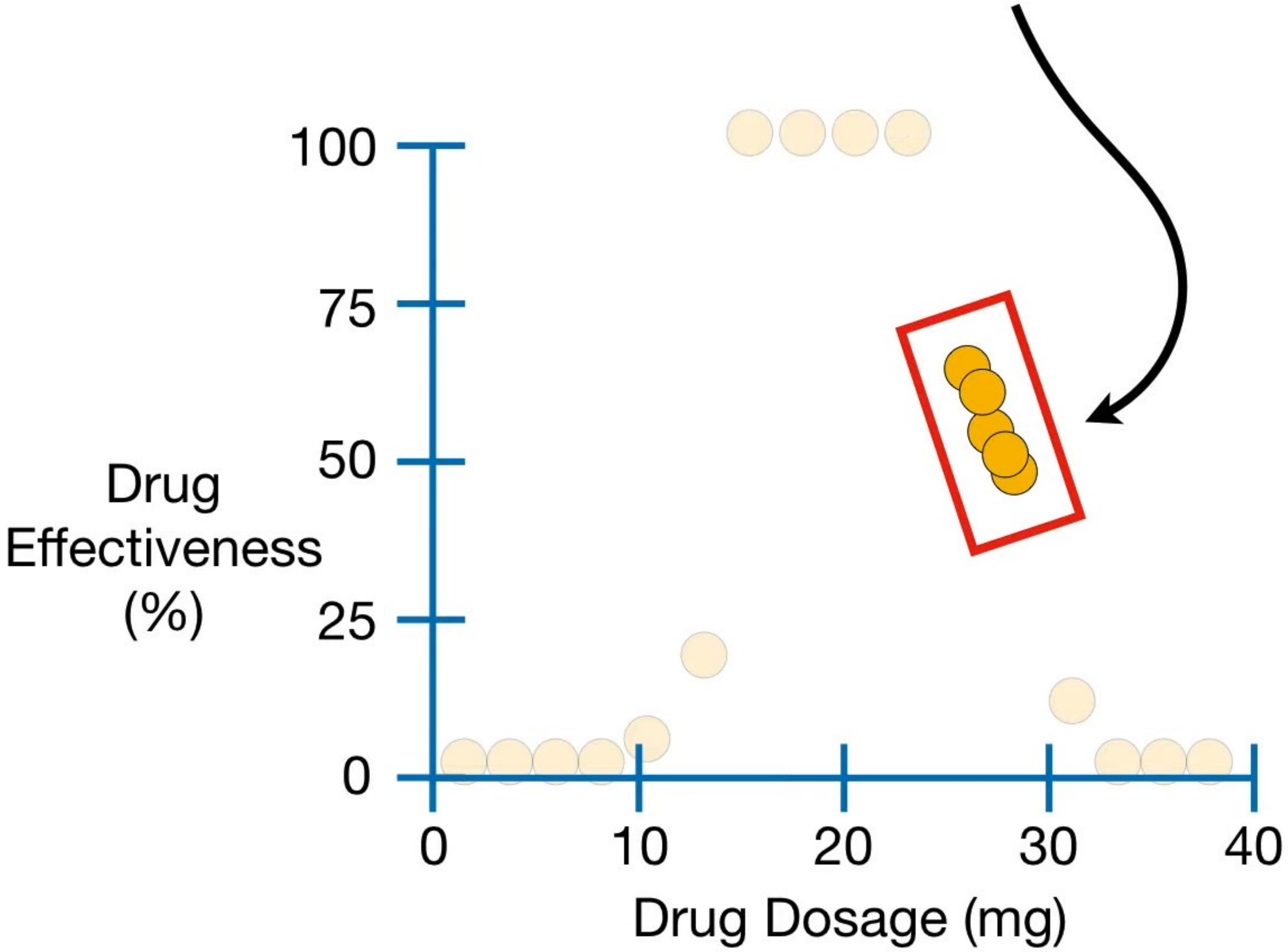
Low dosages are not effective…



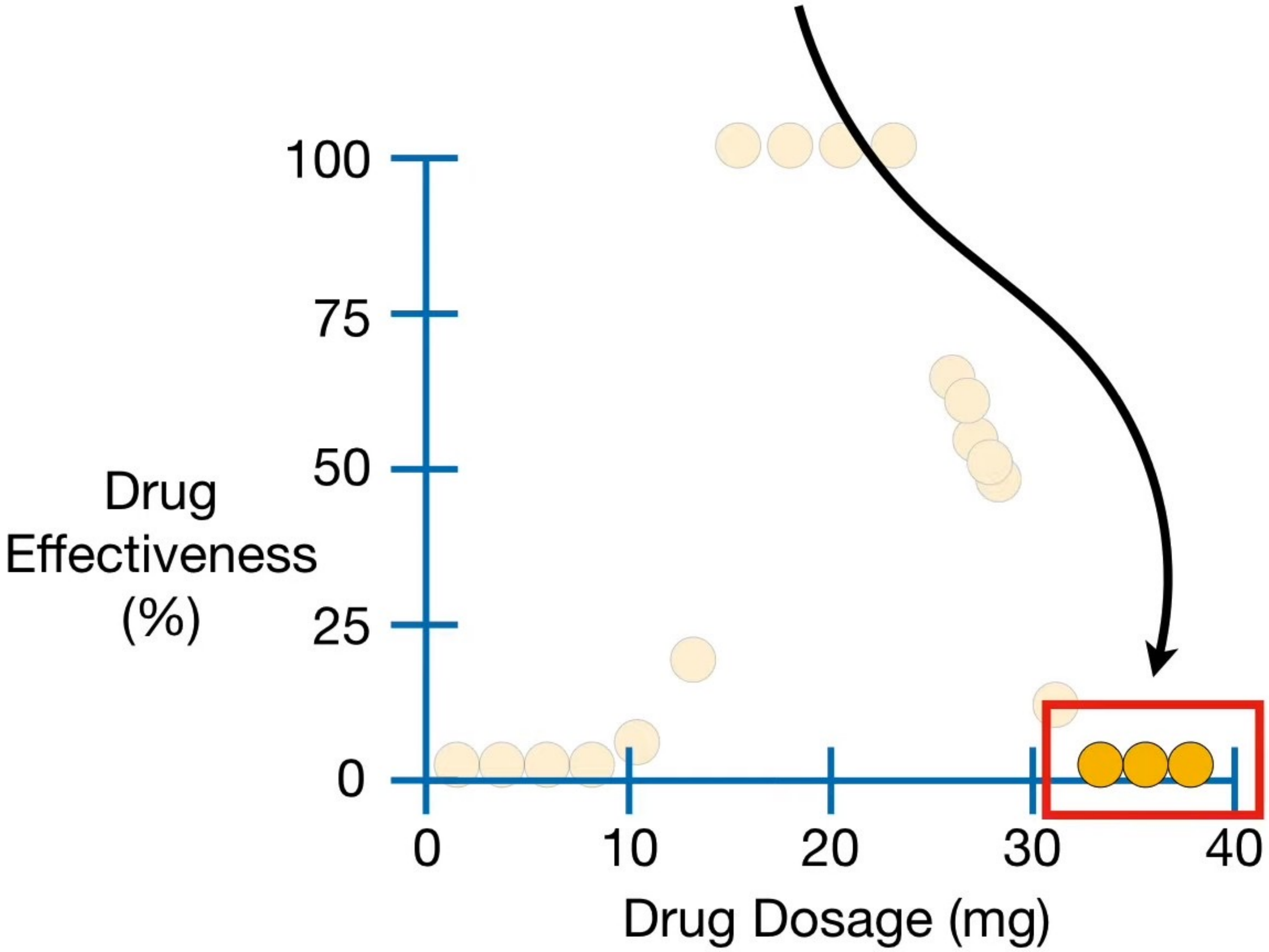
…moderate dosages work really well…



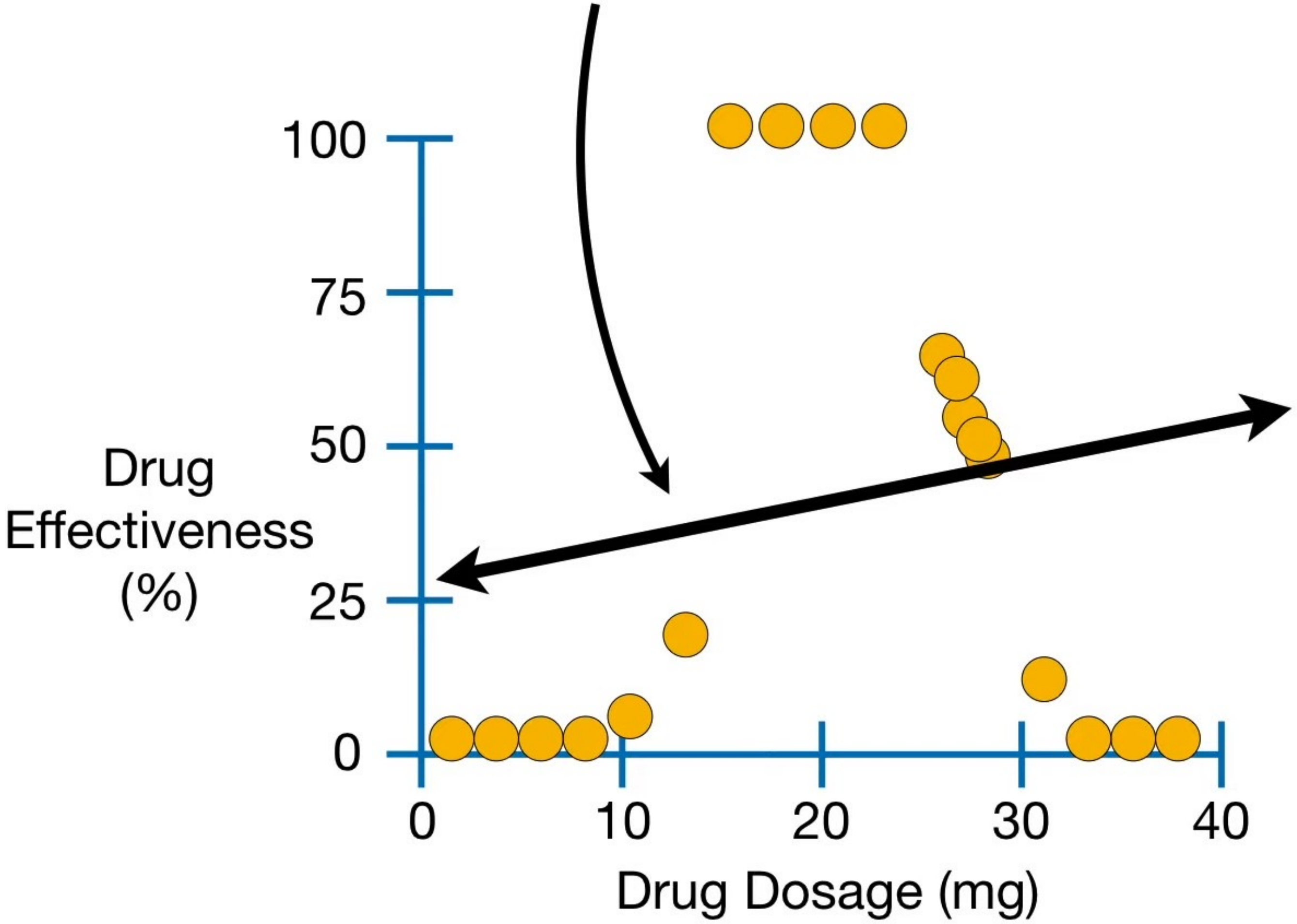
…somewhat higher dosages work at about **50%** effectiveness…



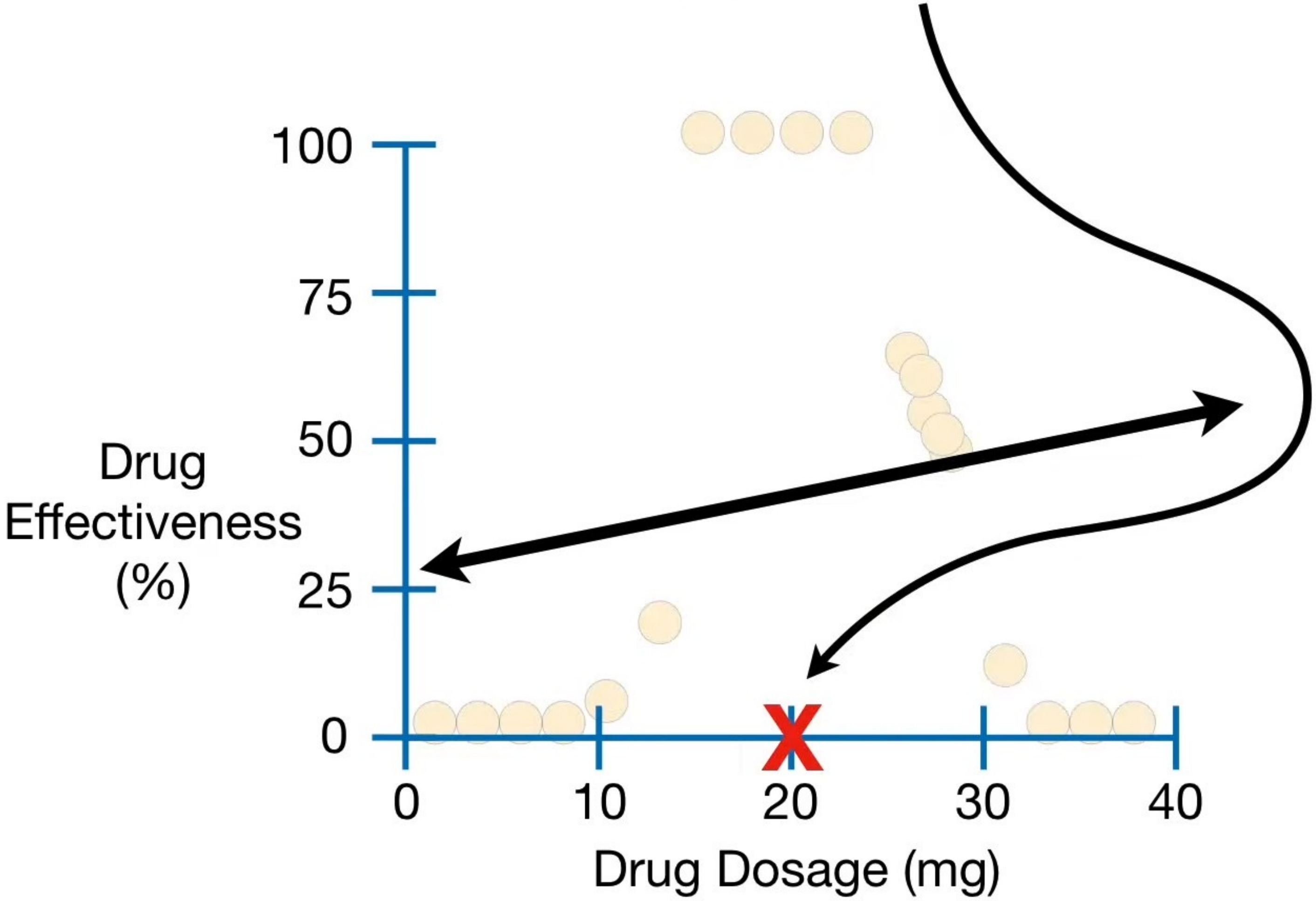
…and high dosages are not effective at all.



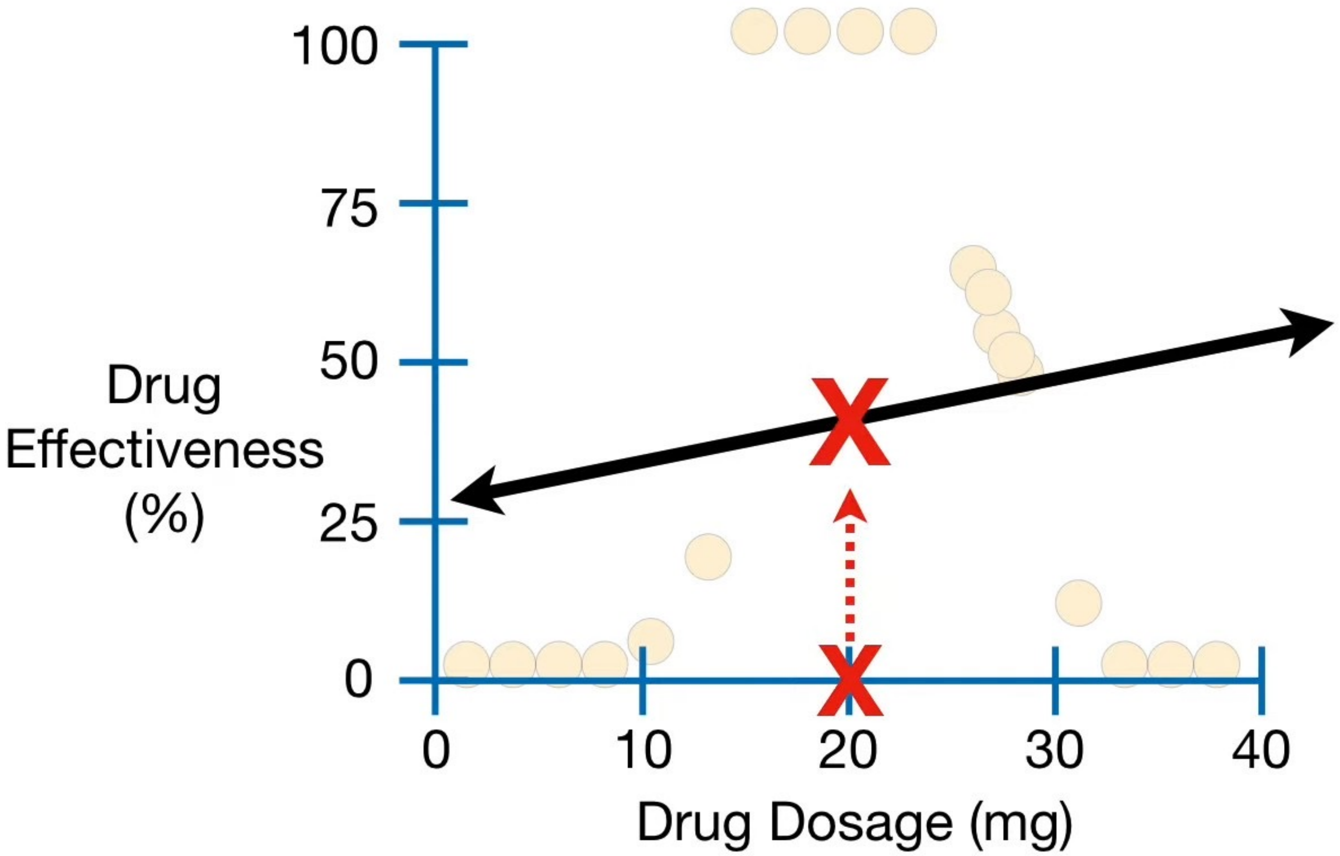
In this case, fitting a straight line to the data will not be very useful.



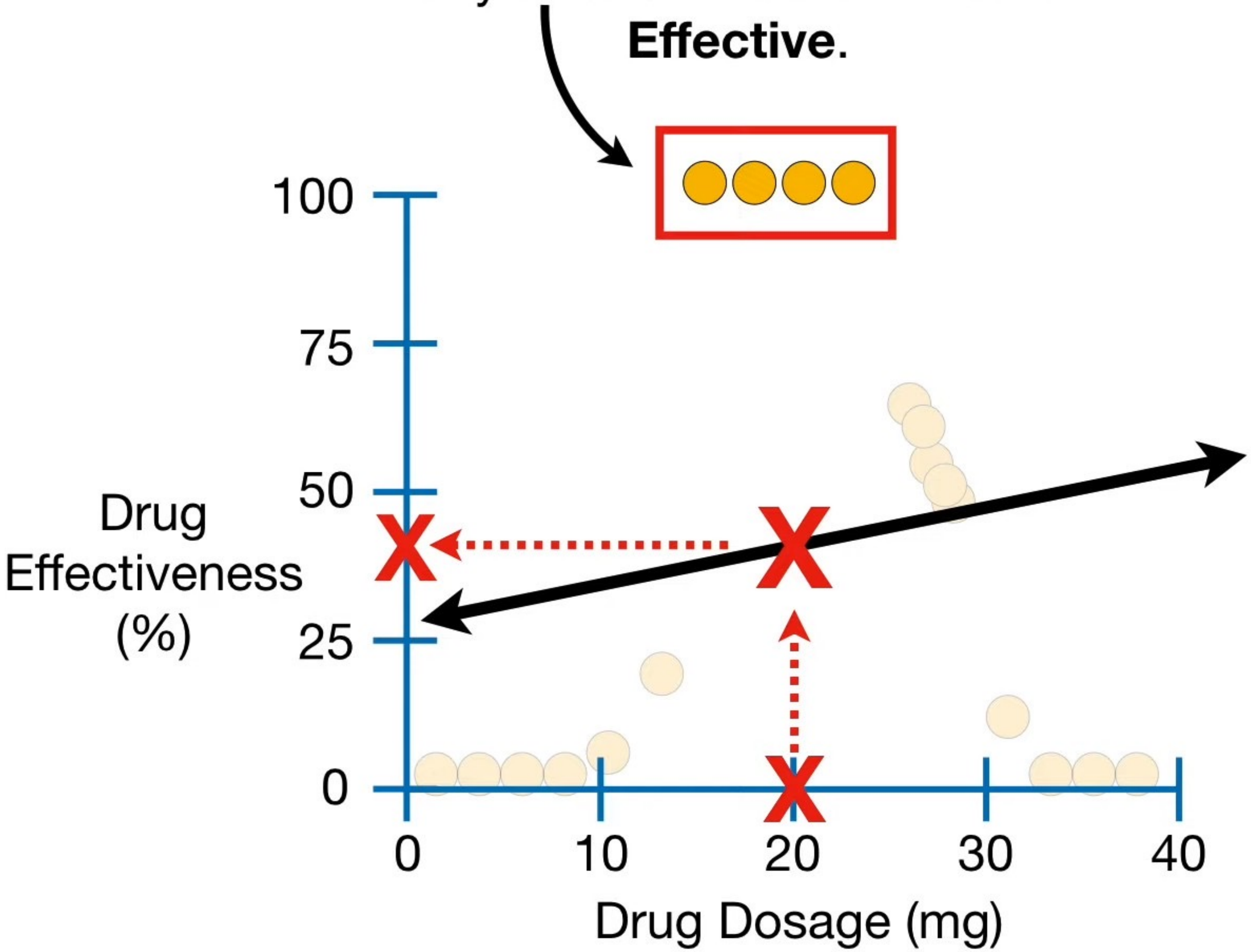
For example, if someone told us they were taking a **20 mg Dose**…



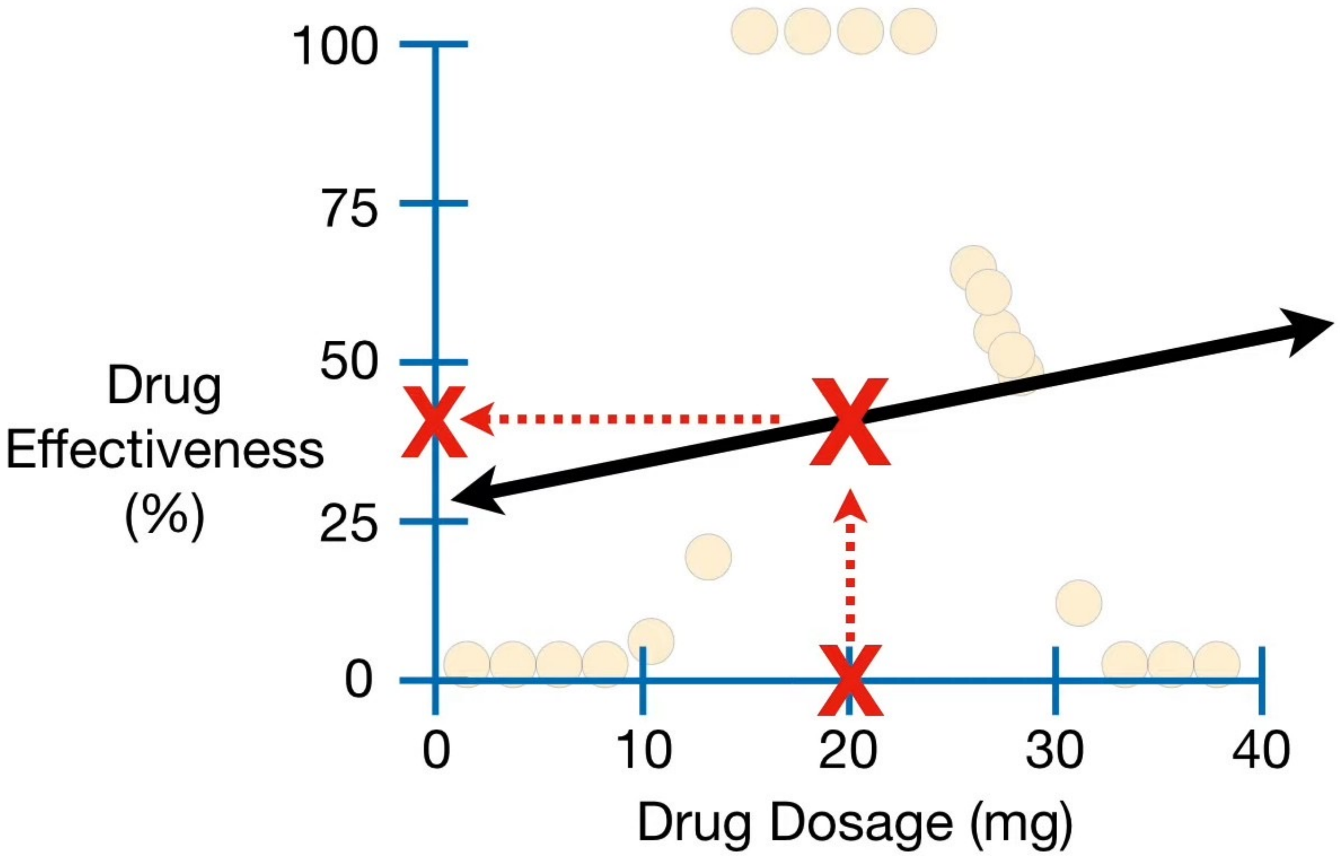
…then we would predict that a **20 mg Dose** should be **45% Effective**…



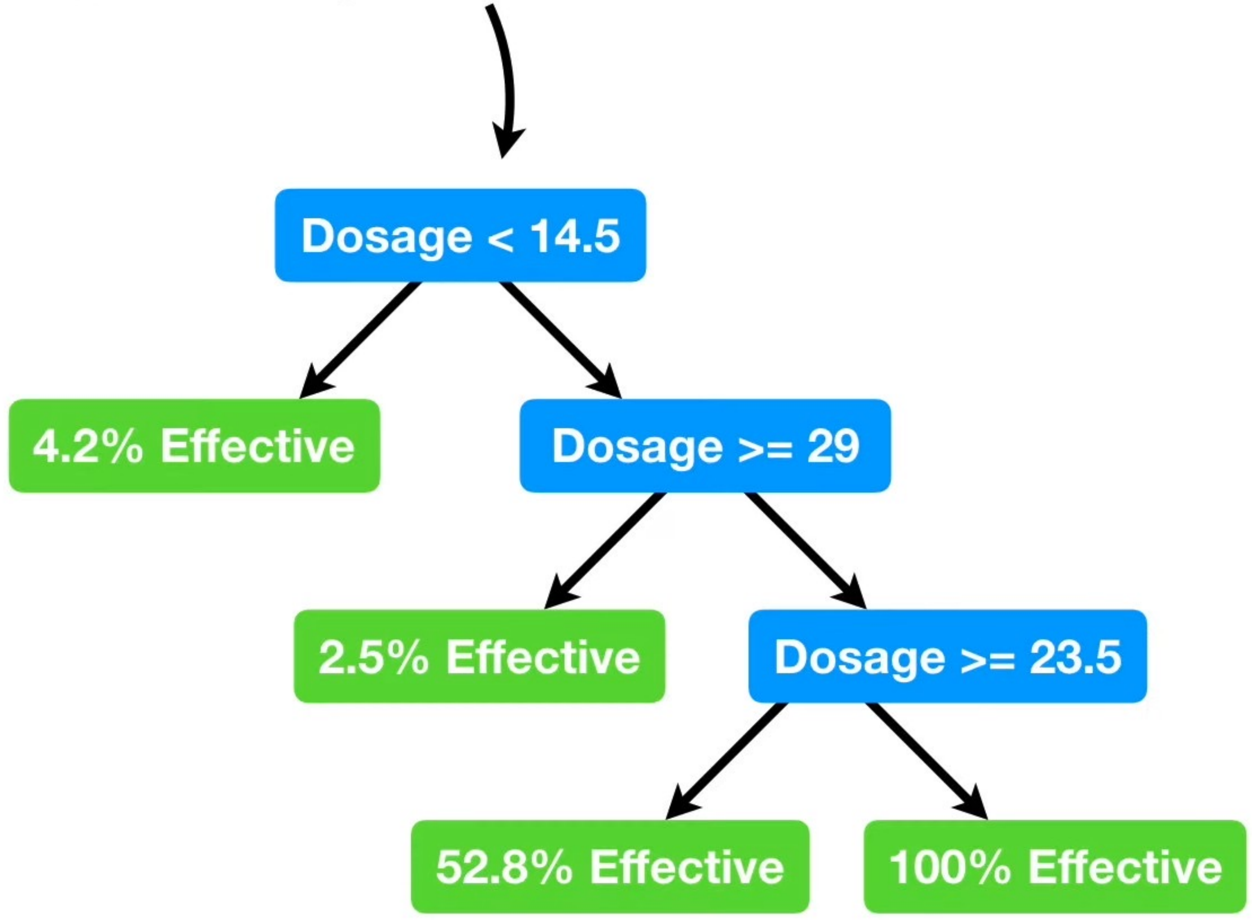
…even though the observed data says that it should be **100% Effective.**



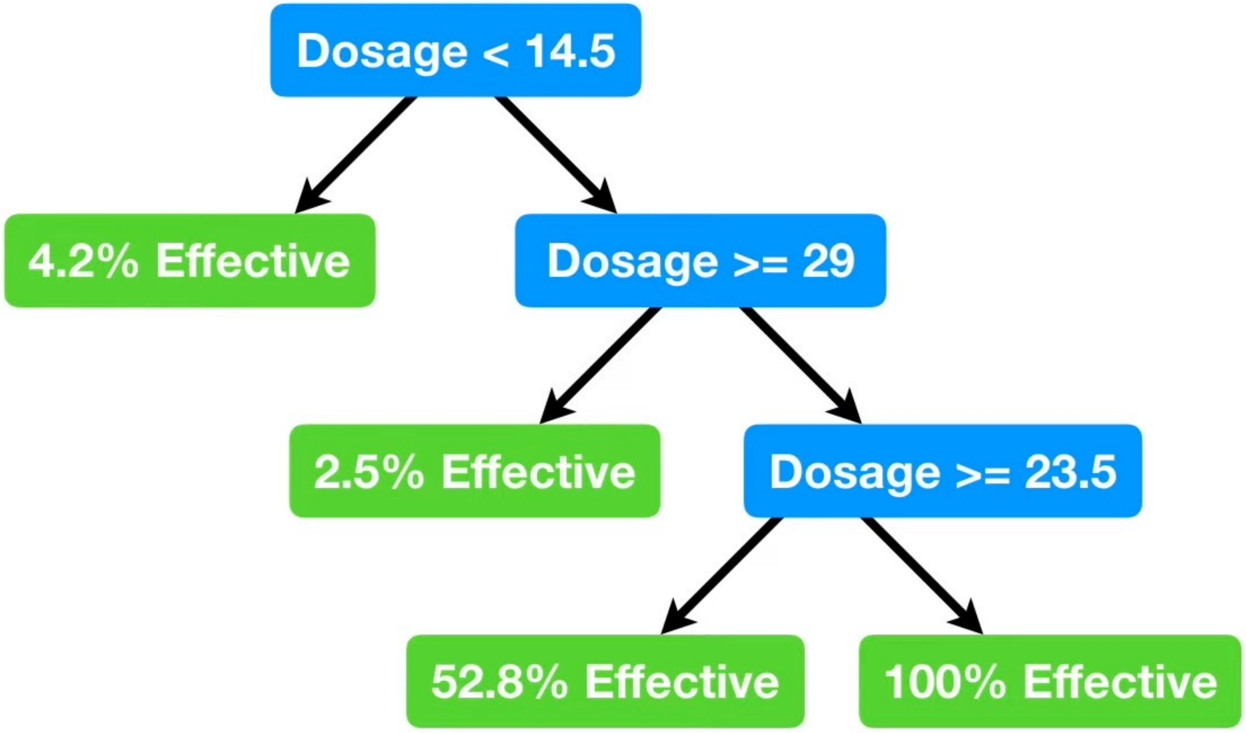
So we need to use something other than a straight line to make predictions.



One option is to use a **Regression Tree**.



Regression Trees are a type of **Decision Tree.**



Dosage < 14.5

4.2% Effective

Dosage >= 29

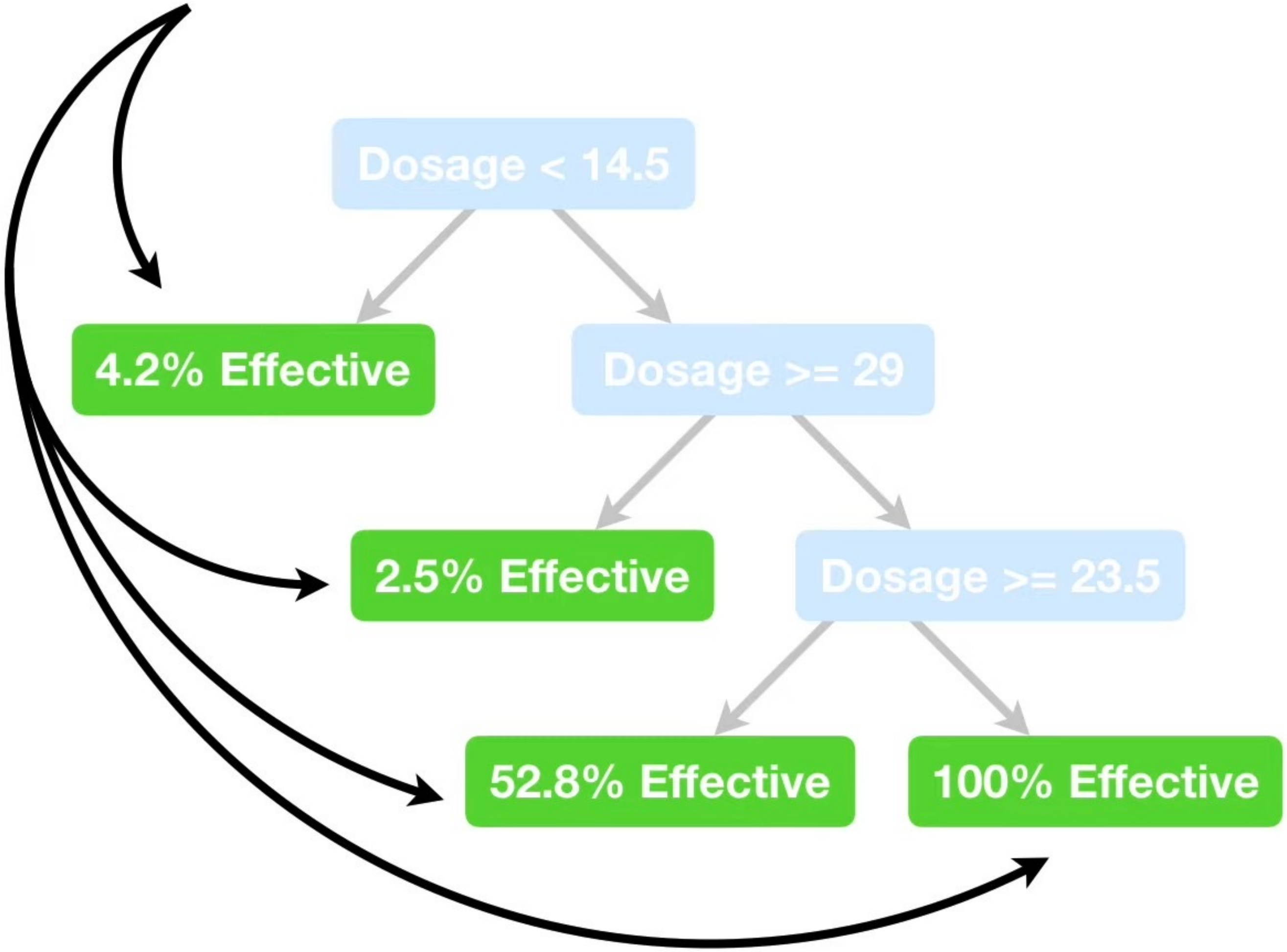
2.5% Effective

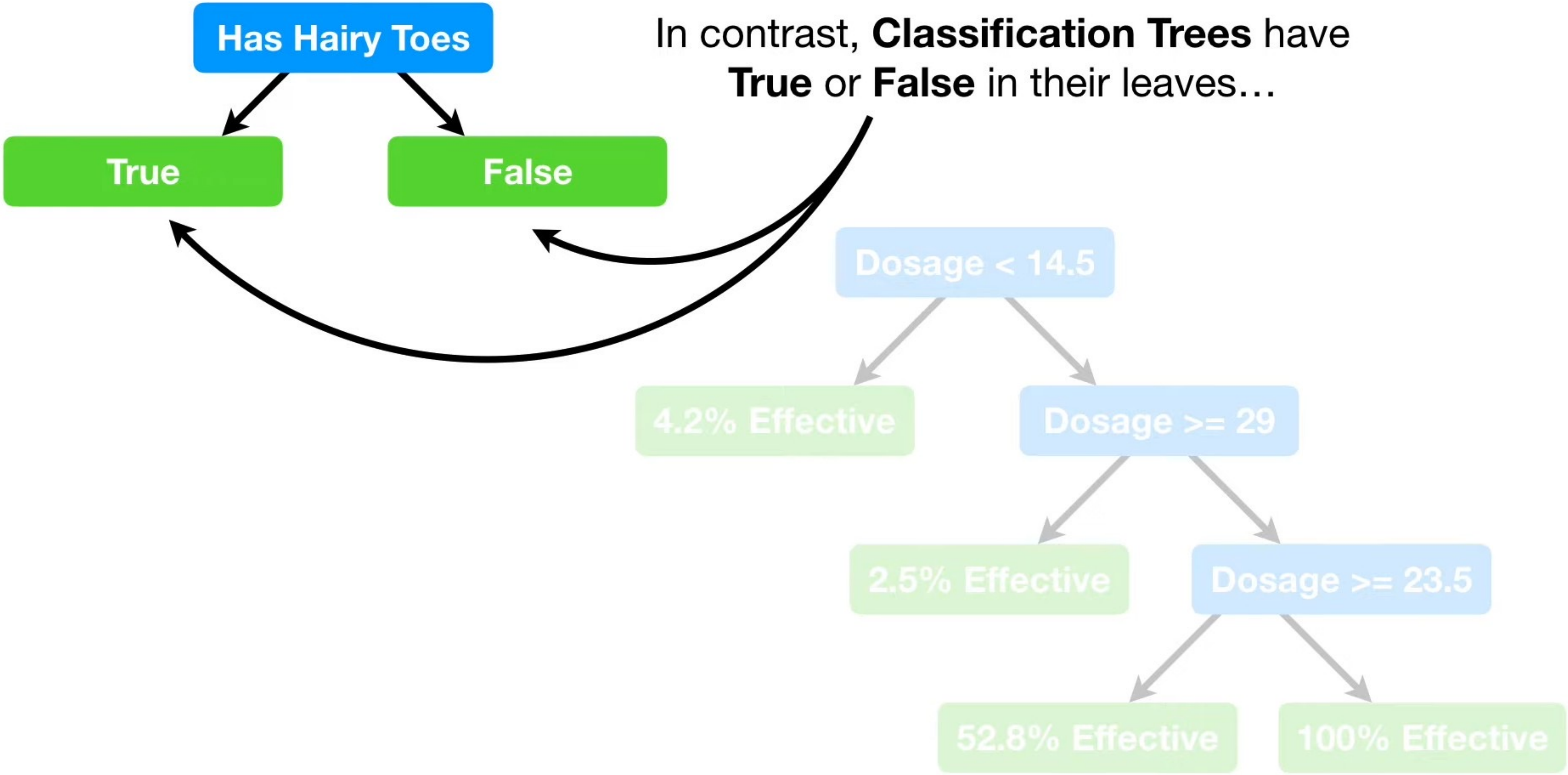
Dosage >= 23.5

52.8% Effective

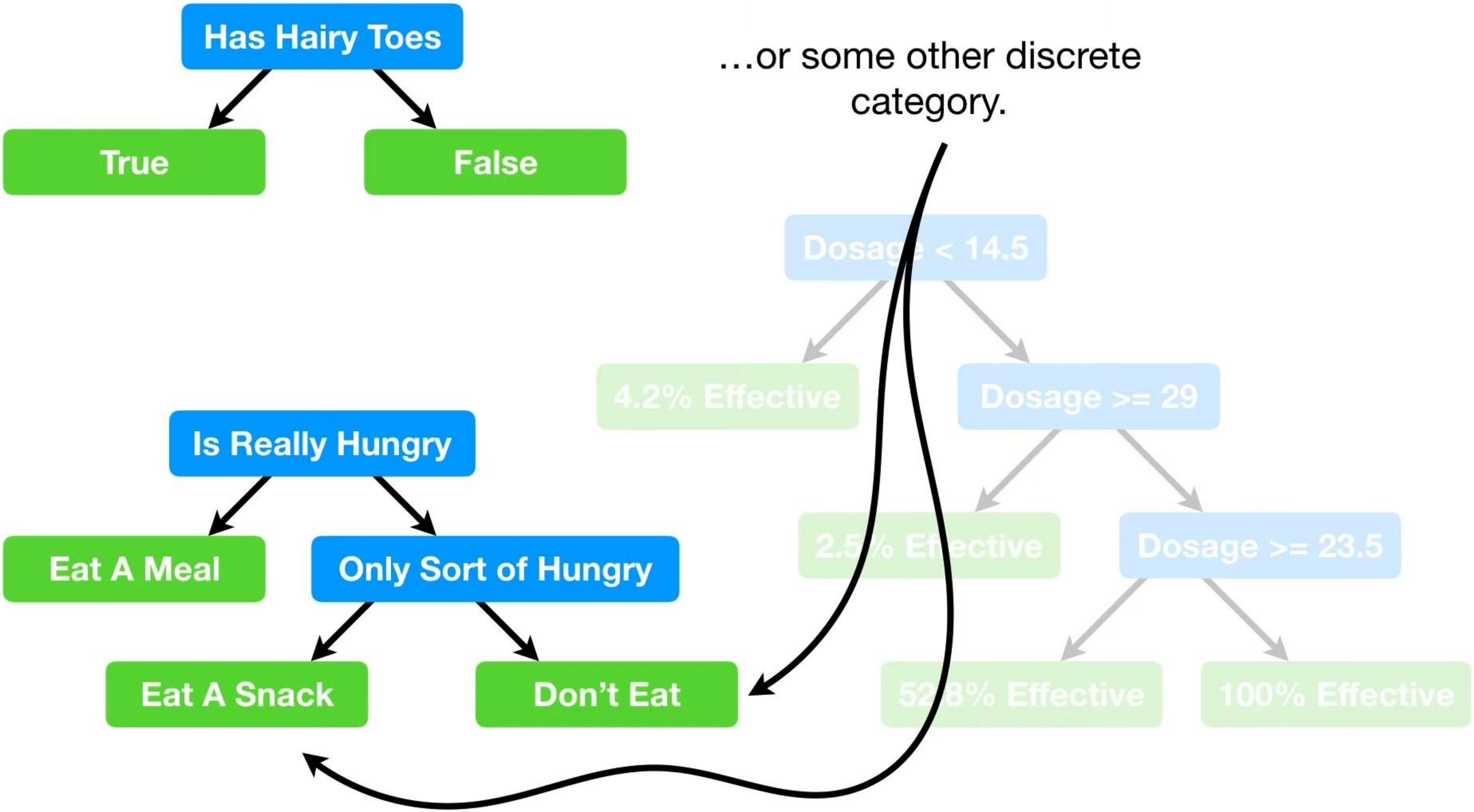
100% Effective

In a **Regression Tree**, each leaf represents a numeric value.





In contrast, **Classification Trees** have **True** or **False** in their leaves…



Has Hairy Toes

True

False

Is Really Hungry

Eat A Meal

Only Sort of Hungry

Eat A Snack

Don't Eat

Dosage < 14.5

4.2% Effective

Dosage >= 29

2.5% Effective

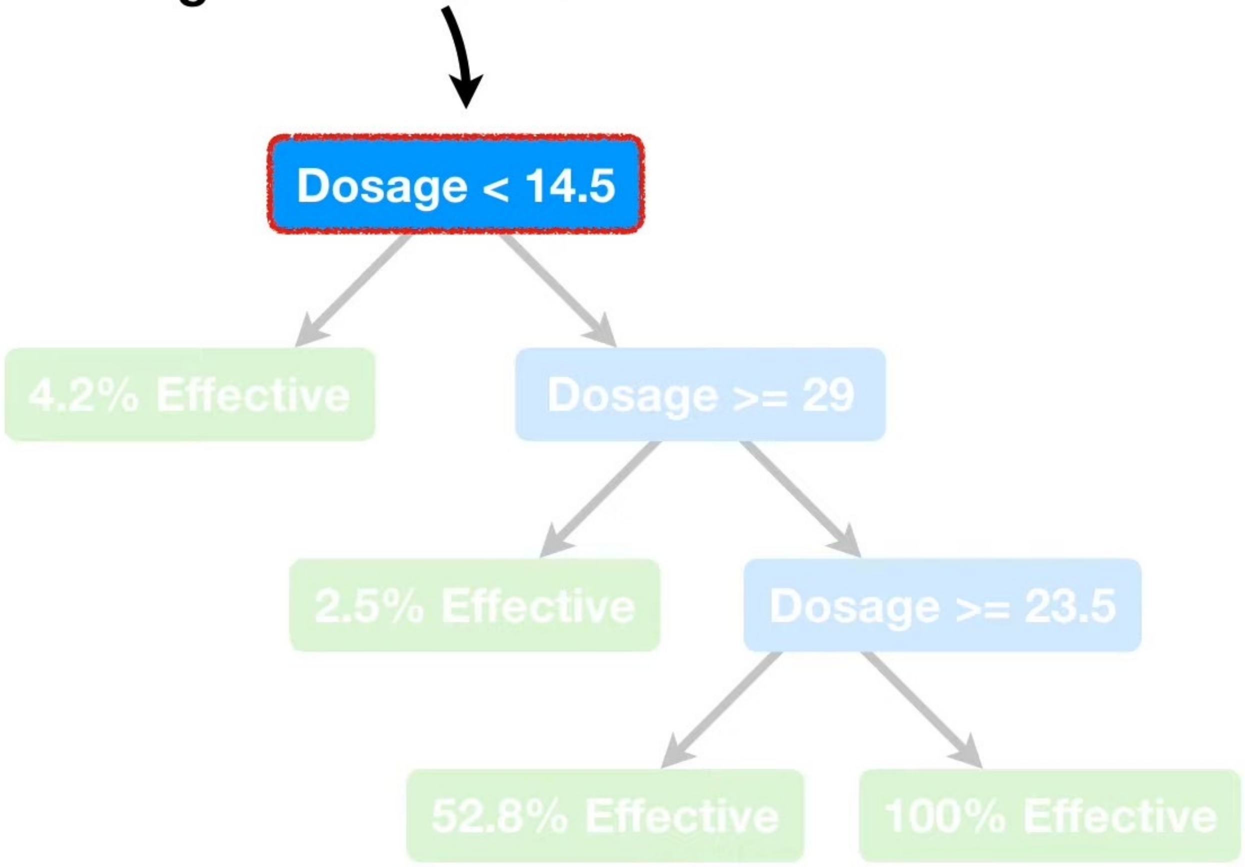
Dosage >= 23.5

52.8% Effective

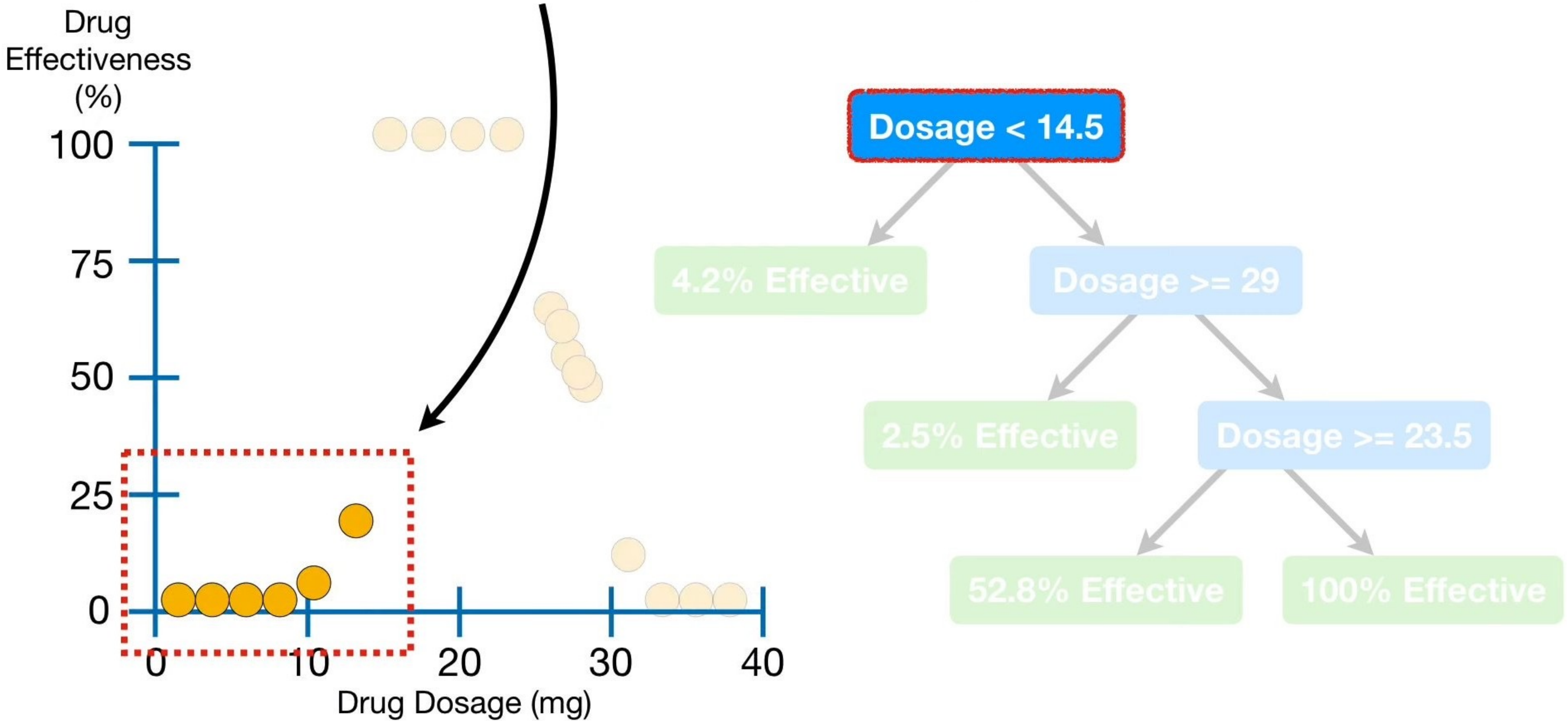
100% Effective

…or some other discrete category.

With *this* **Regression Tree**, we start by asking if the **Dosage** is less than **14.5**.



…if so, then we are talking about these **6** observations in the training data…



Dosage < 14.5

4.2% Effective

Dosage >= 29

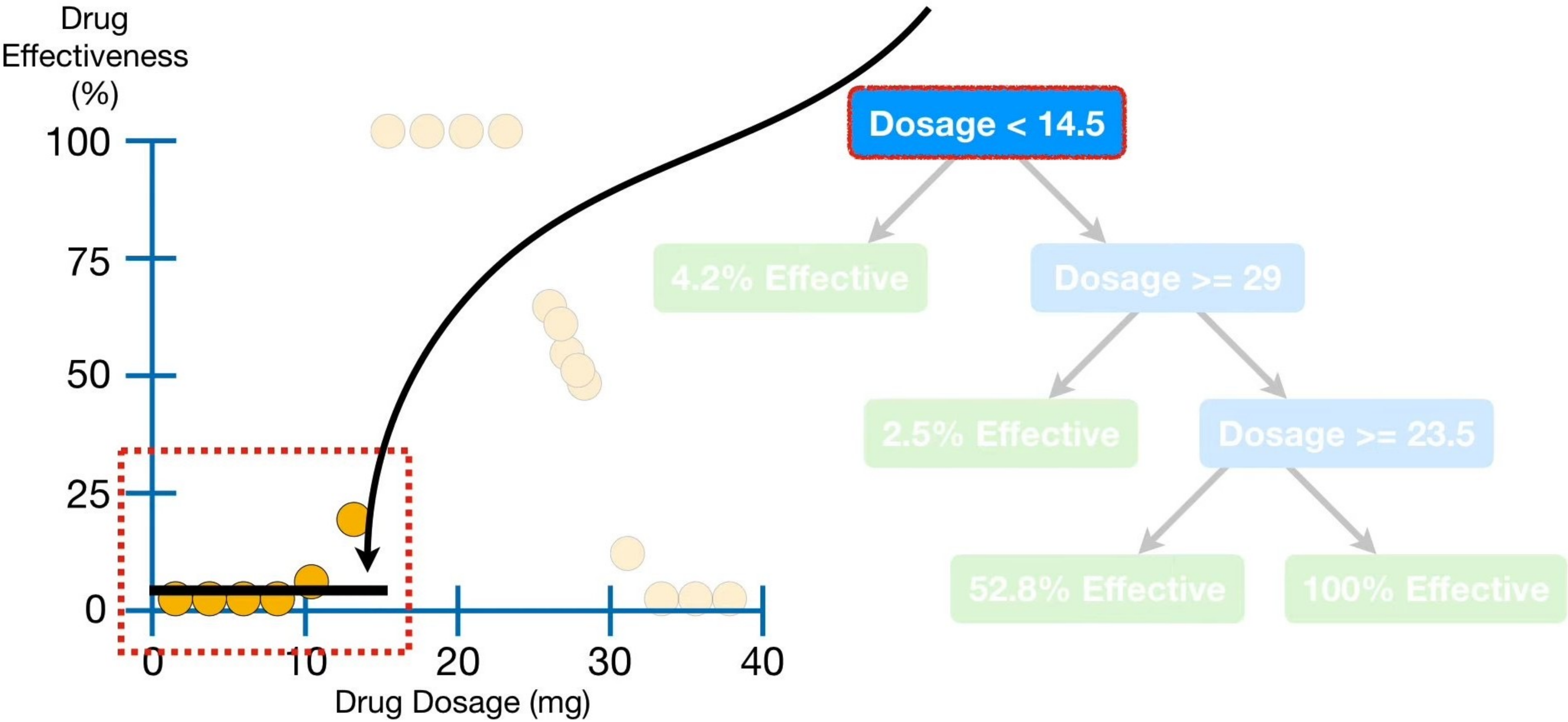
2.5% Effective

Dosage >= 23.5

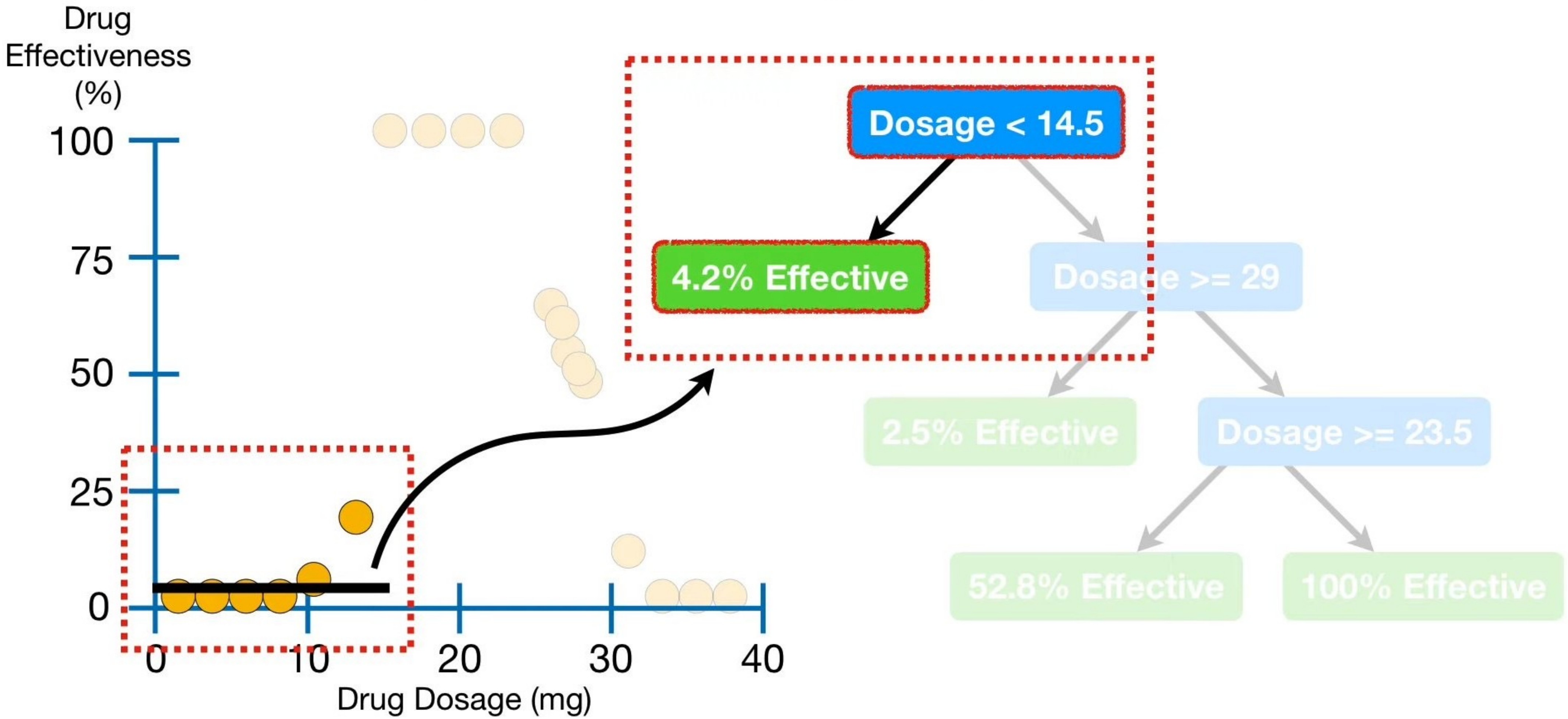
52.8% Effective

100% Effective

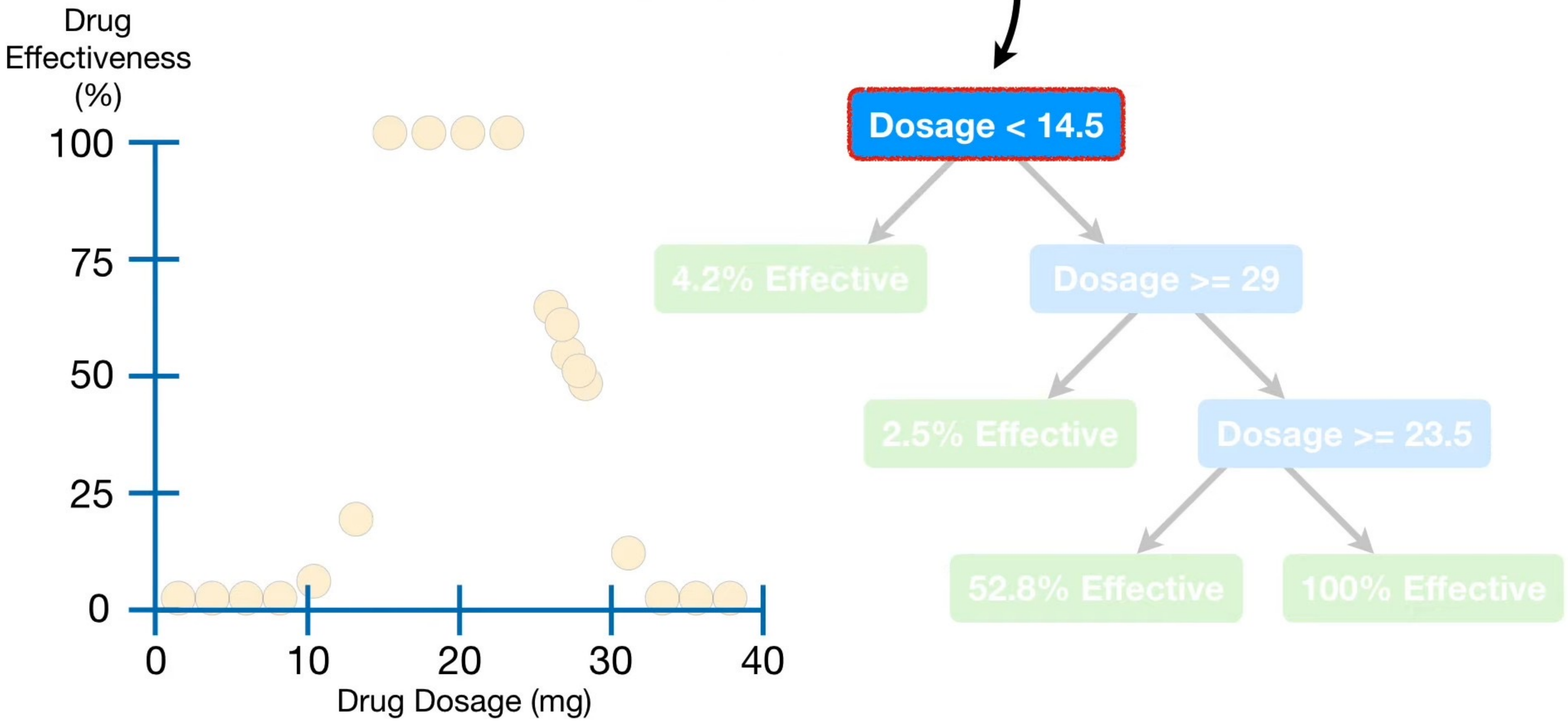
…and the average **Drug Effectiveness** for these **6** observations is **4.2%**…



…so the tree uses the average value, **4.2%**, as its prediction for people with **Dosages < 14.5**.



On the other hand, if the **Dosage >= 14.5**…



Dosage < 14.5

4.2% Effective

Dosage >= 29

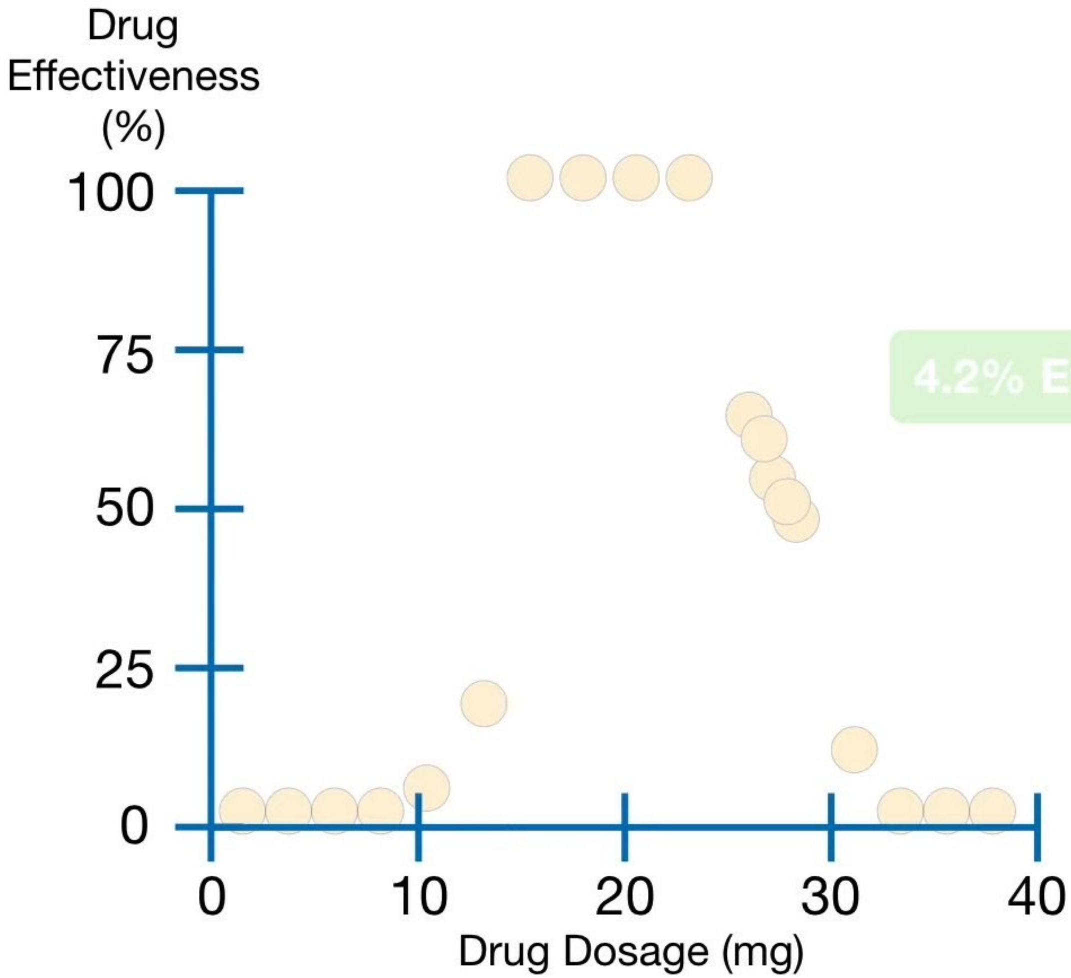
2.5% Effective

Dosage >= 23.5

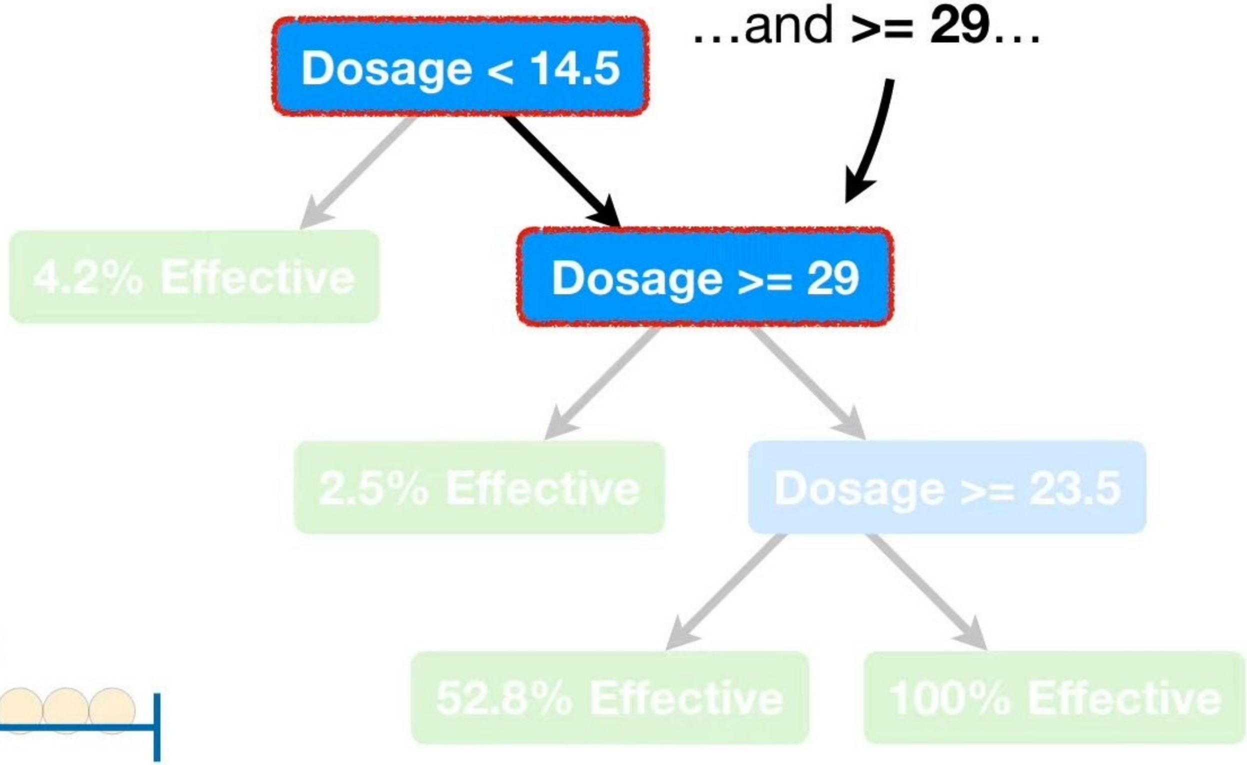
52.8% Effective

100% Effective

On the other hand, if the **Dosage >= 14.5**…



…and **>= 29**…


Dosage < 14.5

4.2% Effective

Dosage >= 29

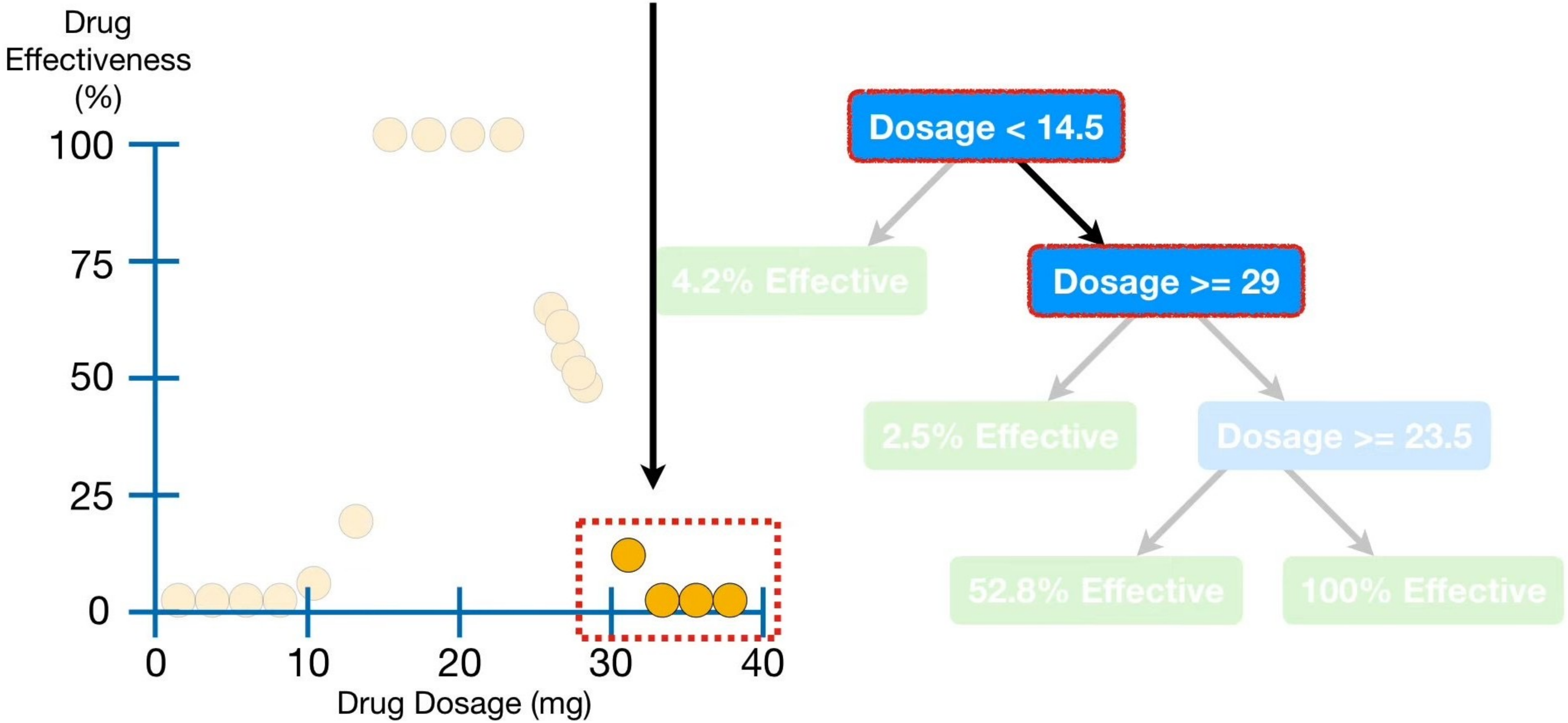
2.5% Effective

Dosage >= 23.5

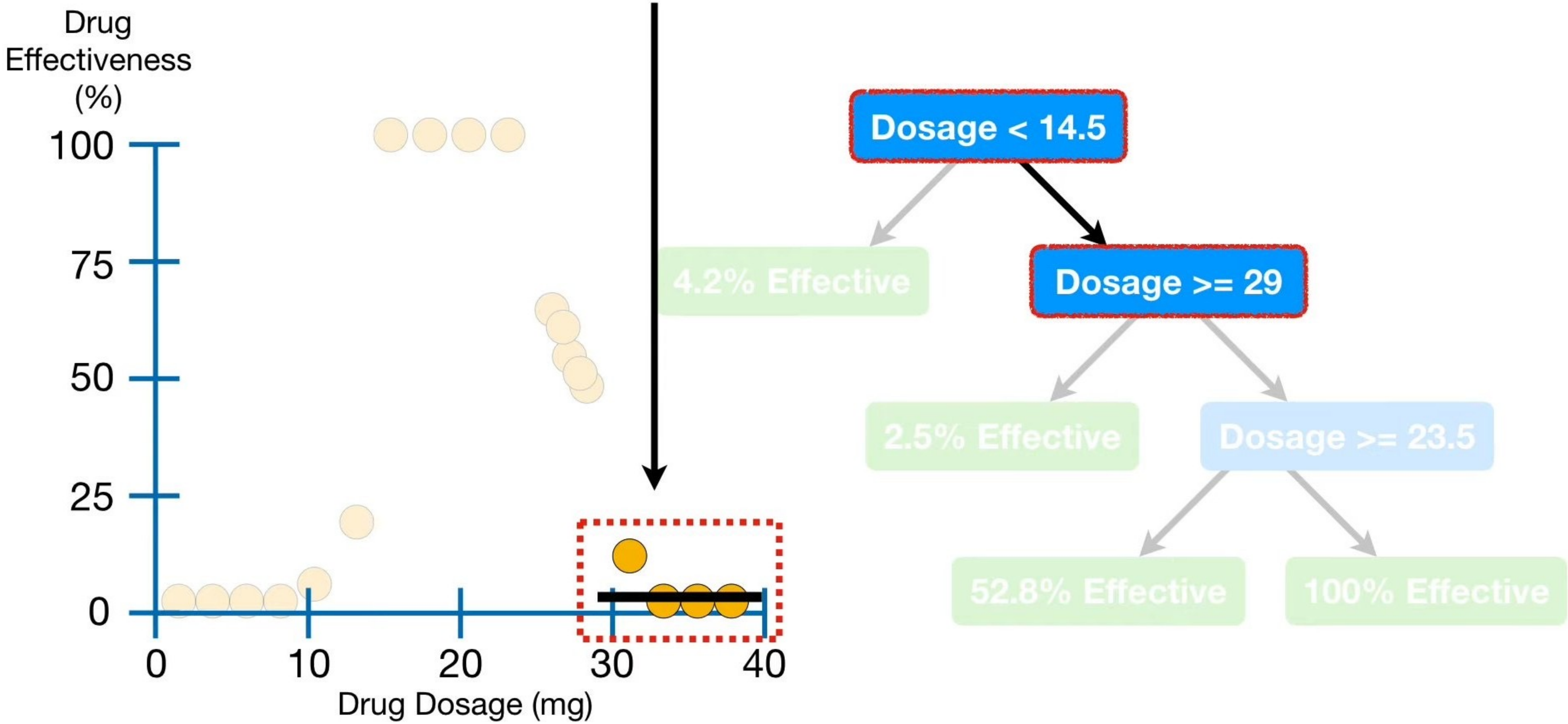
52.8% Effective

100% Effective

…then we are talking about these **4** observations in the training dataset…



…and the average **Drug Effectiveness** for these **4** observations is **2.5%**…



Dosage < 14.5

4.2% Effective

Dosage >= 29

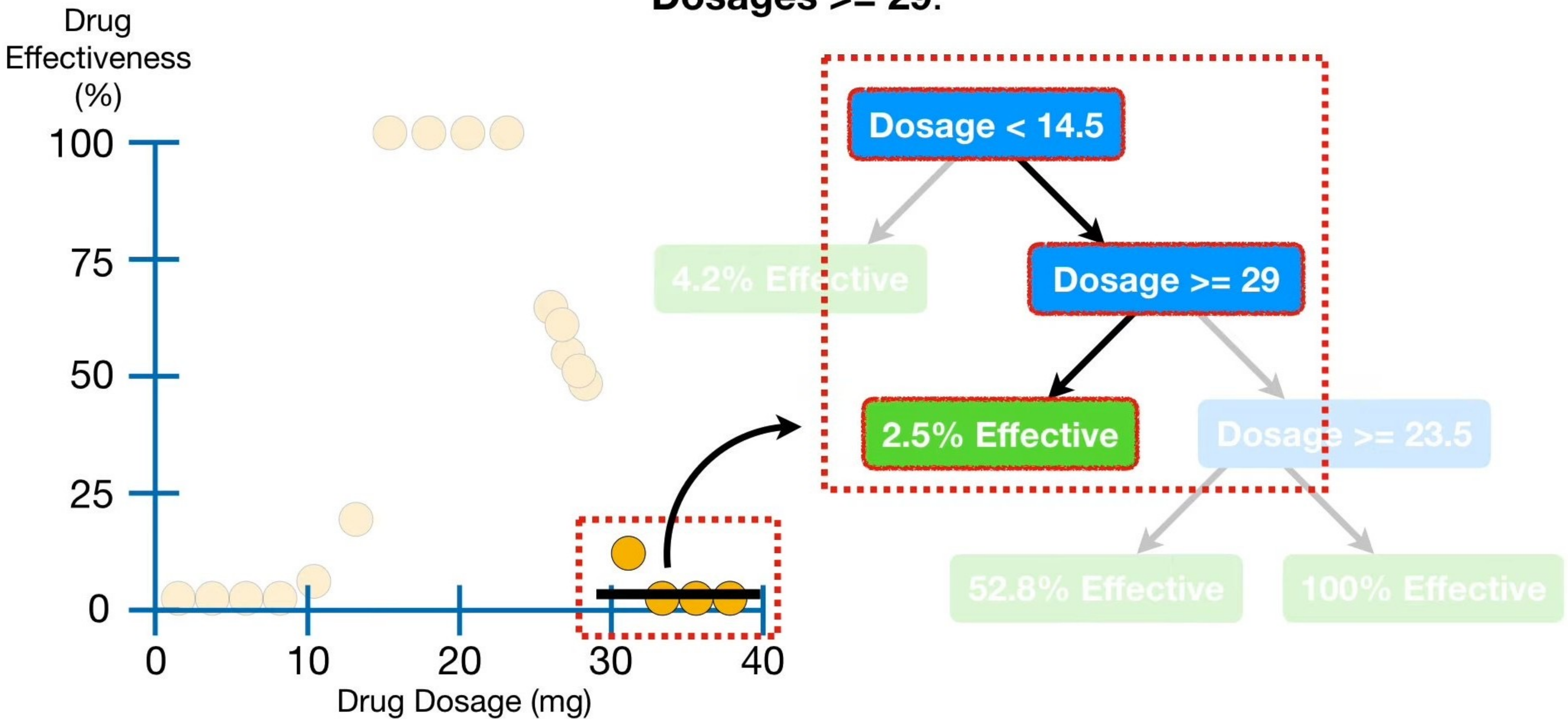
2.5% Effective

Dosage >= 23.5

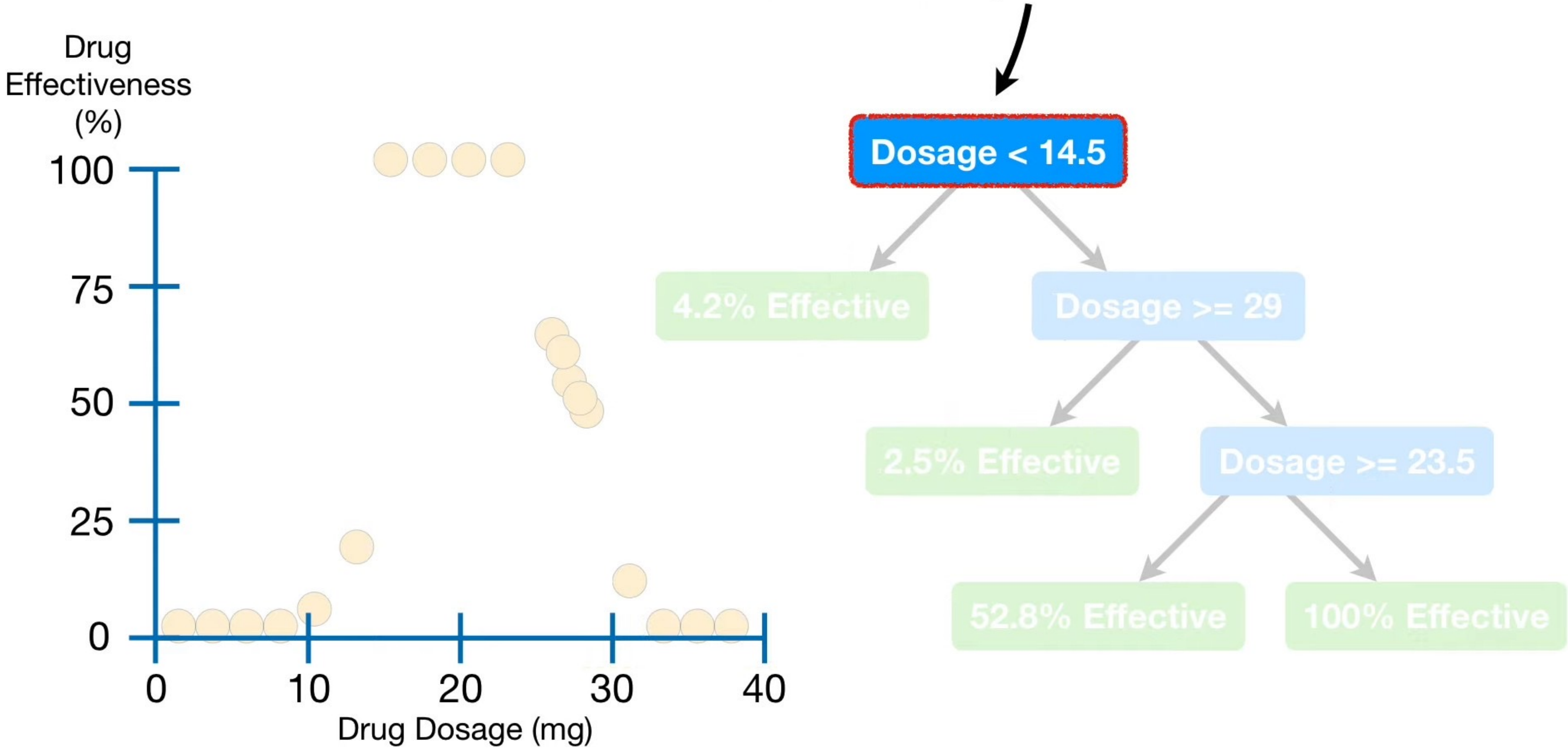
52.8% Effective

100% Effective

…so the tree uses the average value, **2.5%**, as its prediction for people with **Dosages >= 29**.



Now, if the **Dosage >= 14.5**…



Dosage < 14.5

4.2% Effective

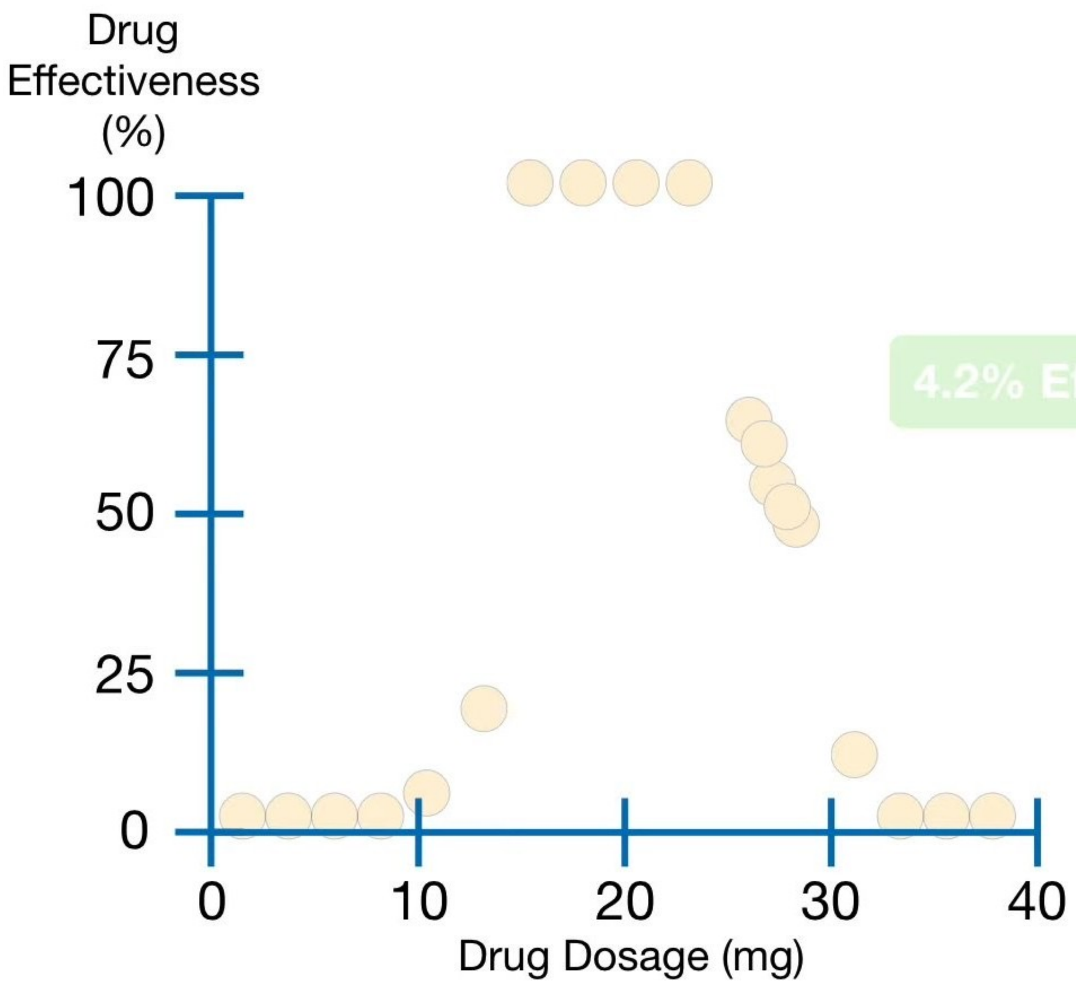
Dosage >= 29

2.5% Effective

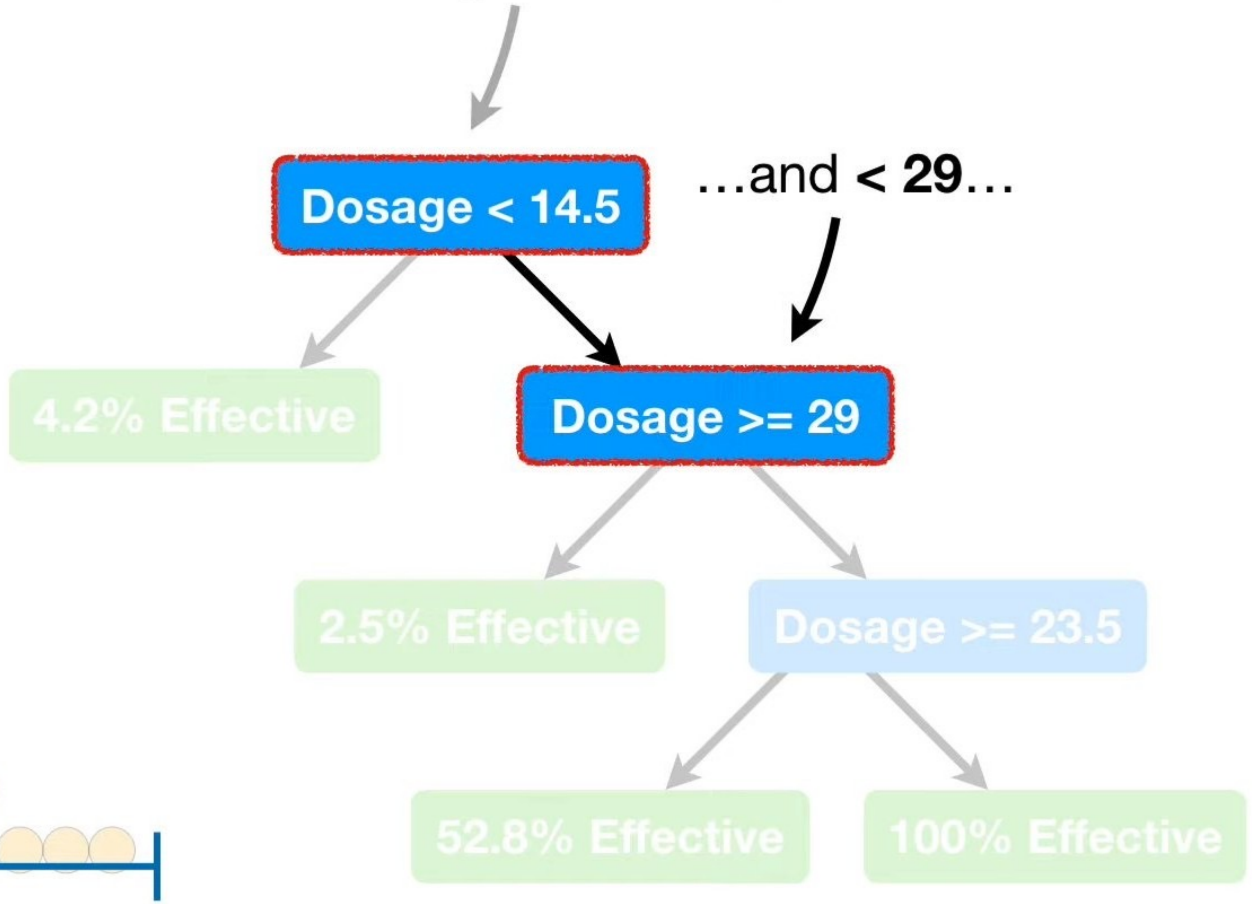
Dosage >= 23.5

52.8% Effective

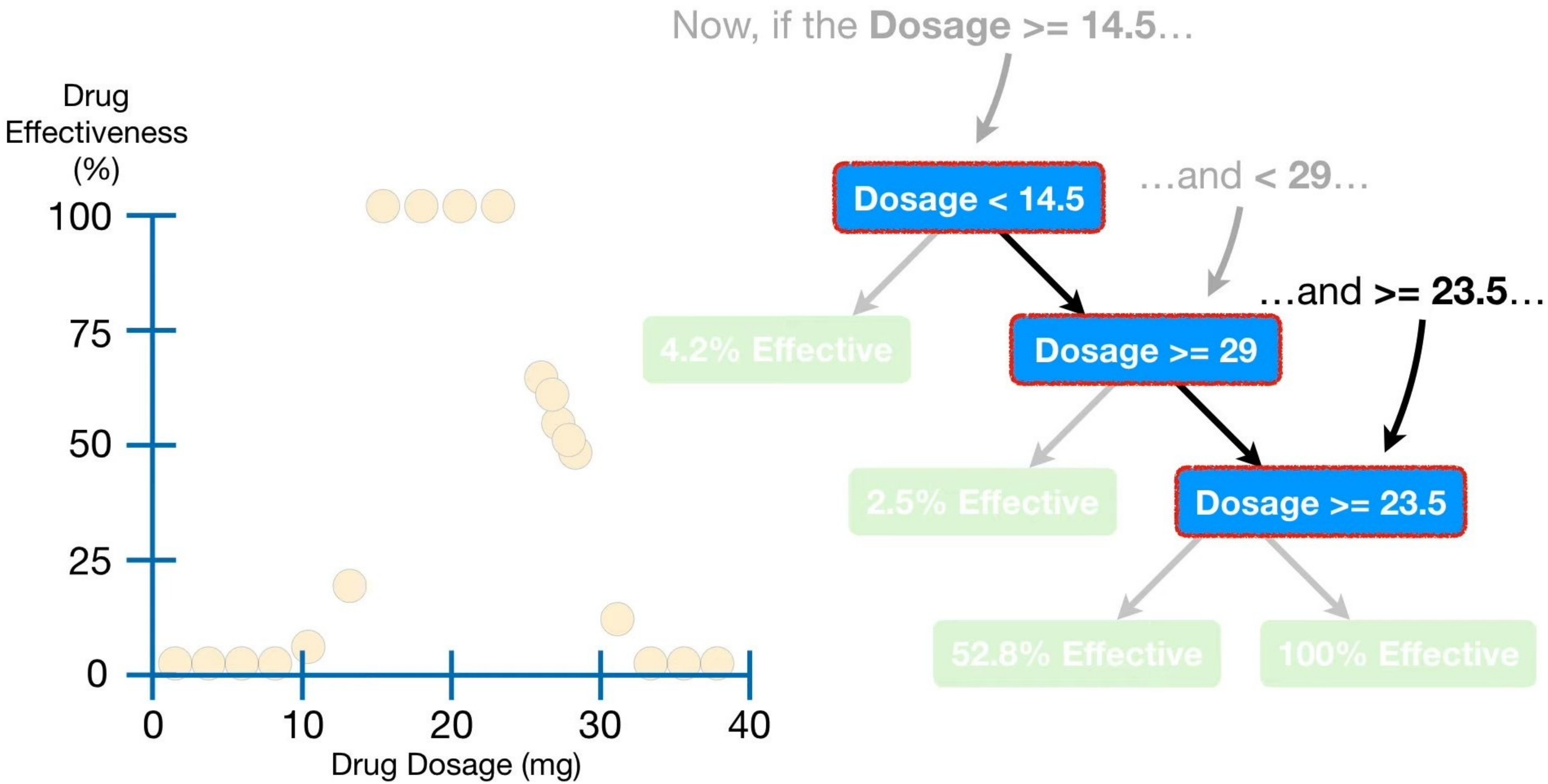
100% Effective



Now, if the **Dosage >= 14.5**…



…and **< 29**…



Now, if the **Dosage >= 14.5**…

…and **< 29**…

…and **>= 23.5**…

Dosage < 14.5

4.2% Effective

Dosage >= 29

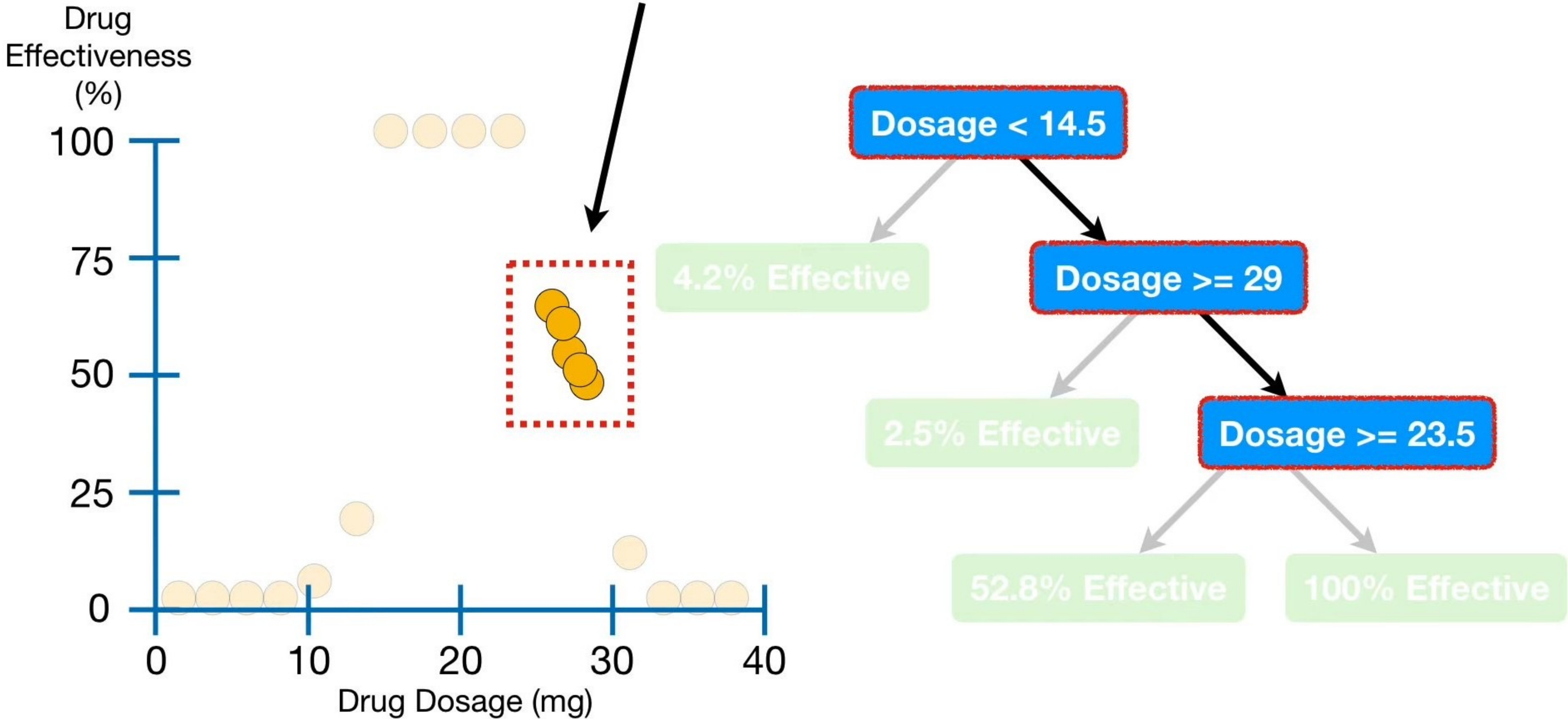
2.5% Effective

Dosage >= 23.5

52.8% Effective

100% Effective

…then we are talking about these **5** observations in the training dataset…



Dosage < 14.5

4.2% Effective

Dosage >= 29

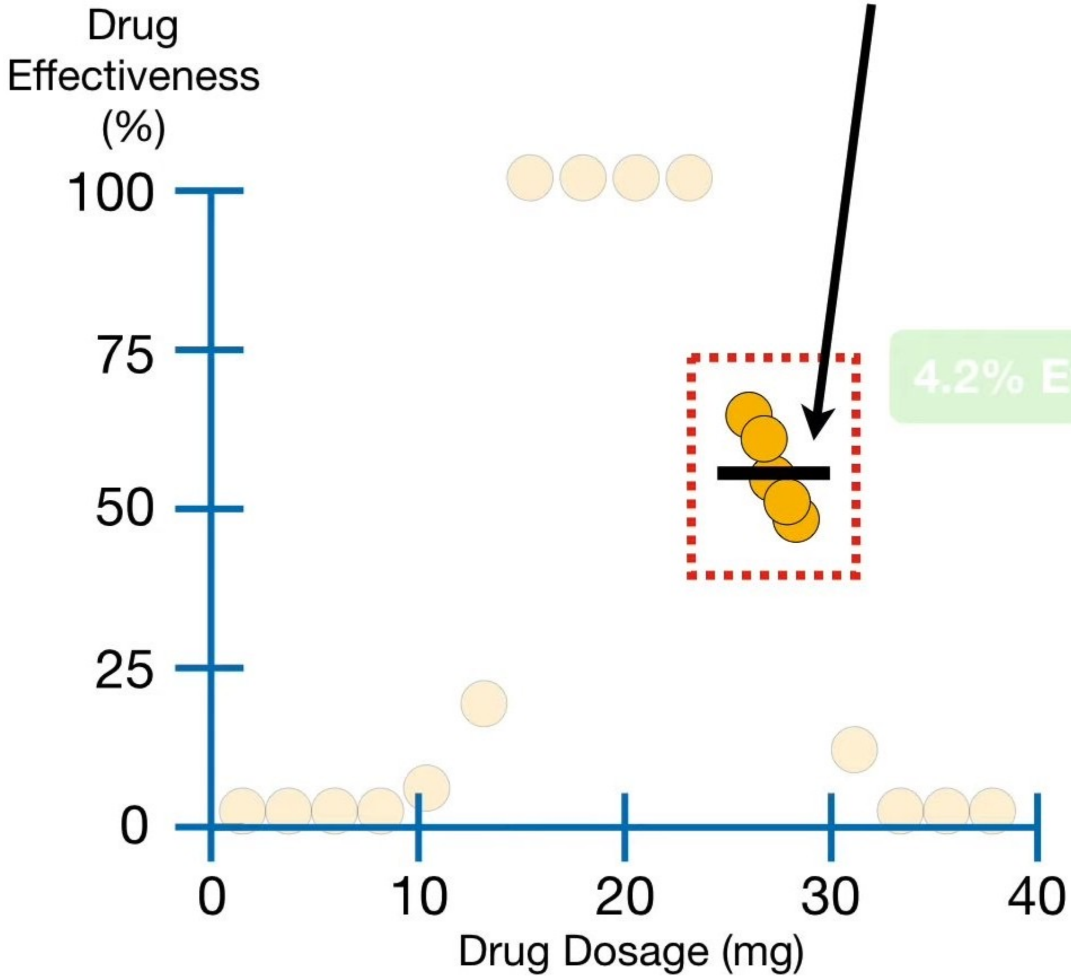
2.5% Effective

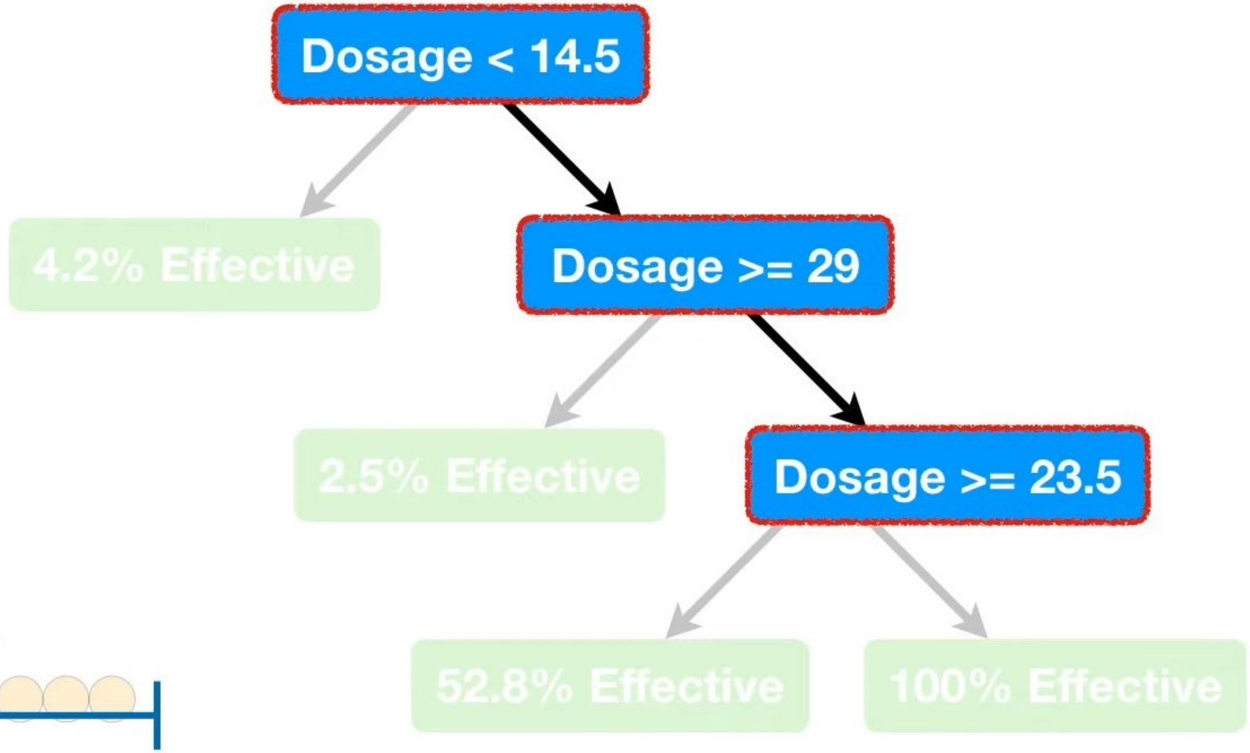
Dosage >= 23.5

52.8% Effective

100% Effective

…and the average **Drug Effectiveness** for these **5** observations is **52.8%**…




Dosage < 14.5

4.2% Effective

Dosage >= 29

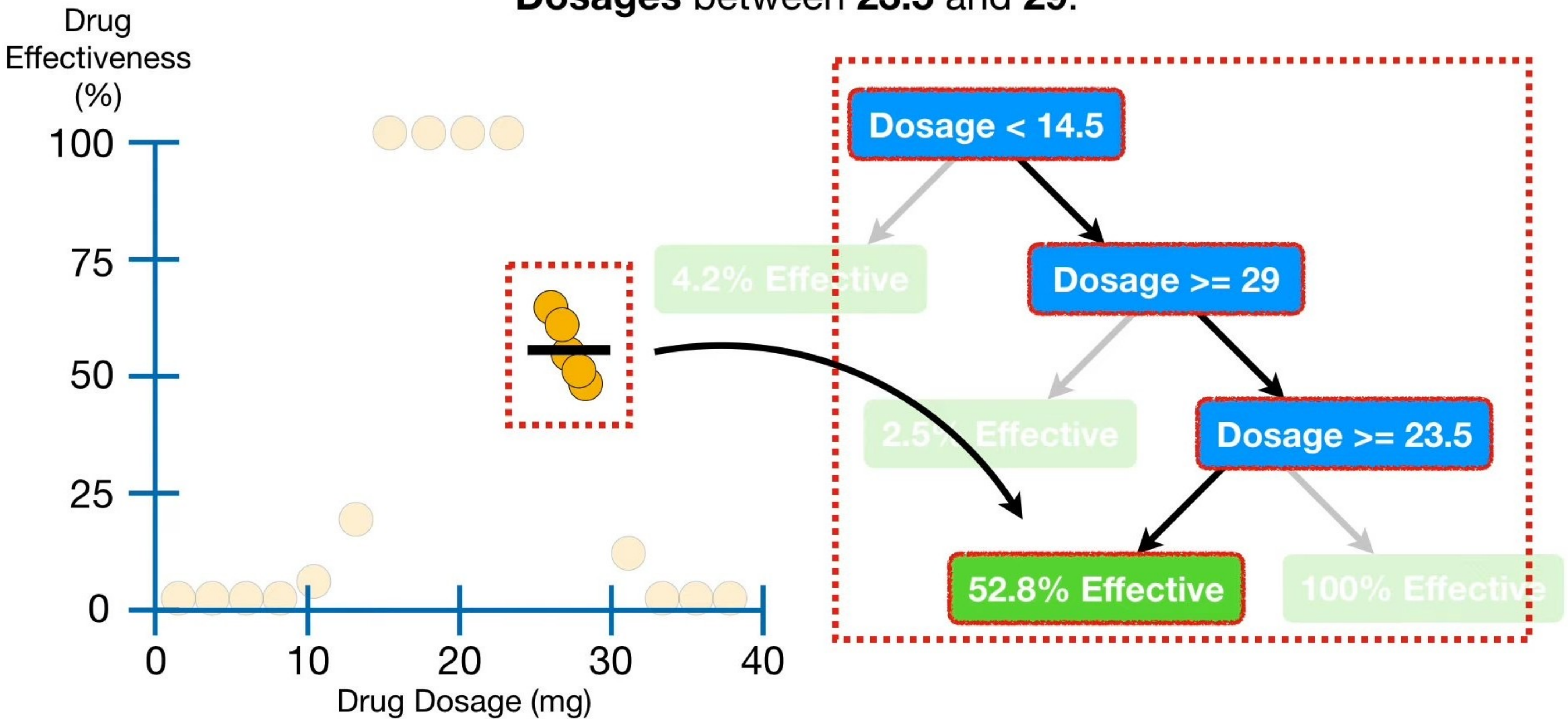
2.5% Effective

Dosage >= 23.5

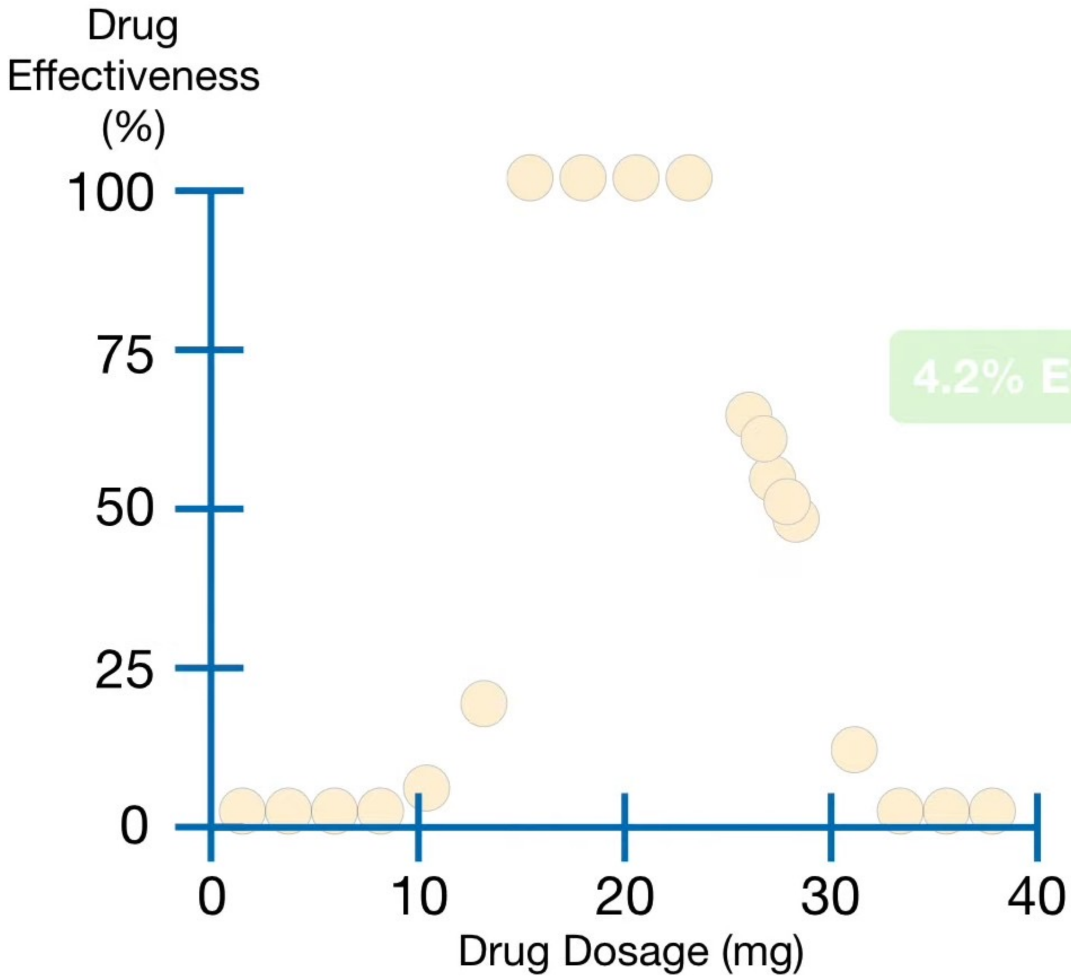
52.8% Effective

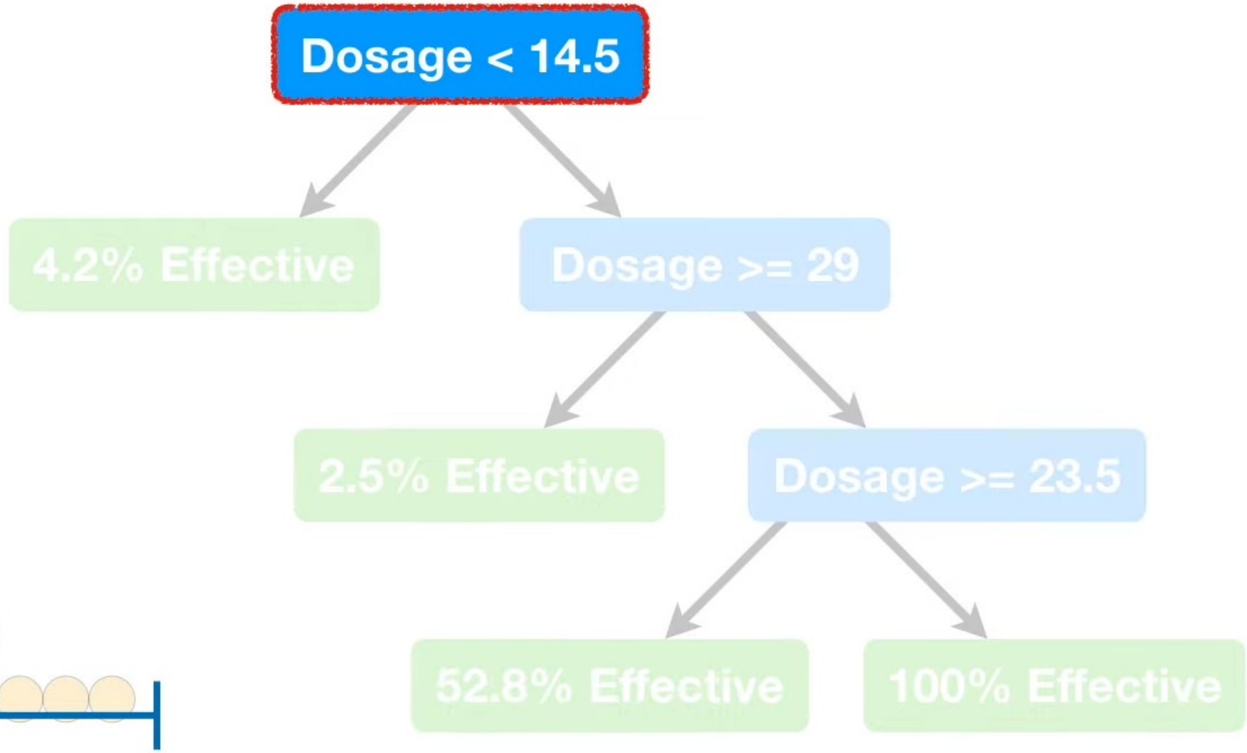
100% Effective

…so the tree uses the average value, **52.8%**, as its prediction for people with **Dosages** between **23.5** and **29**.



Lastly, if the **Dosage >= 14.5**…





Dosage < 14.5

4.2% Effective

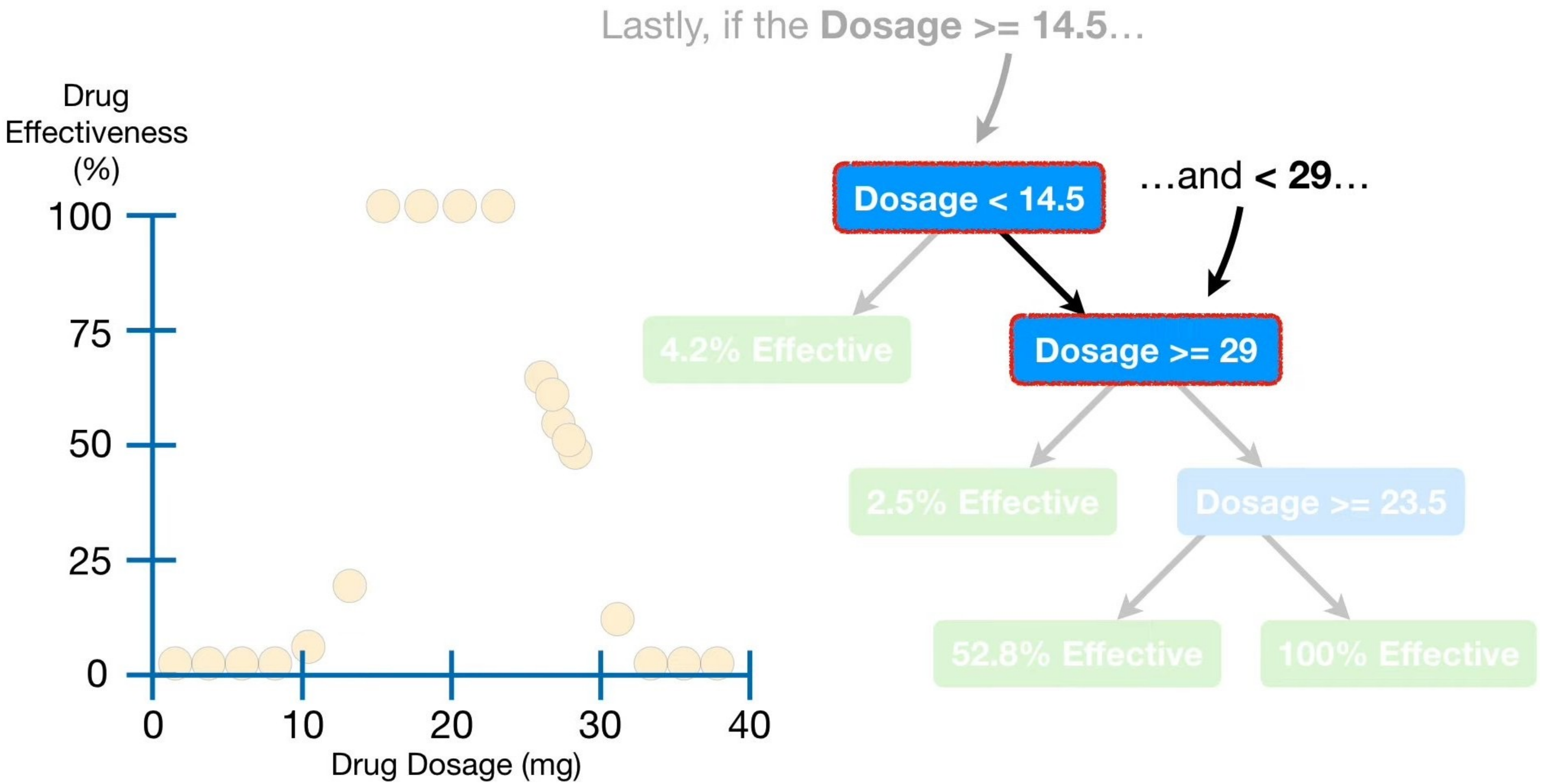
Dosage >= 29

2.5% Effective

Dosage >= 23.5

52.8% Effective

100% Effective



Lastly, if the **Dosage >= 14.5**…

Dosage < 14.5

…and **< 29**…

4.2% Effective

Dosage >= 29

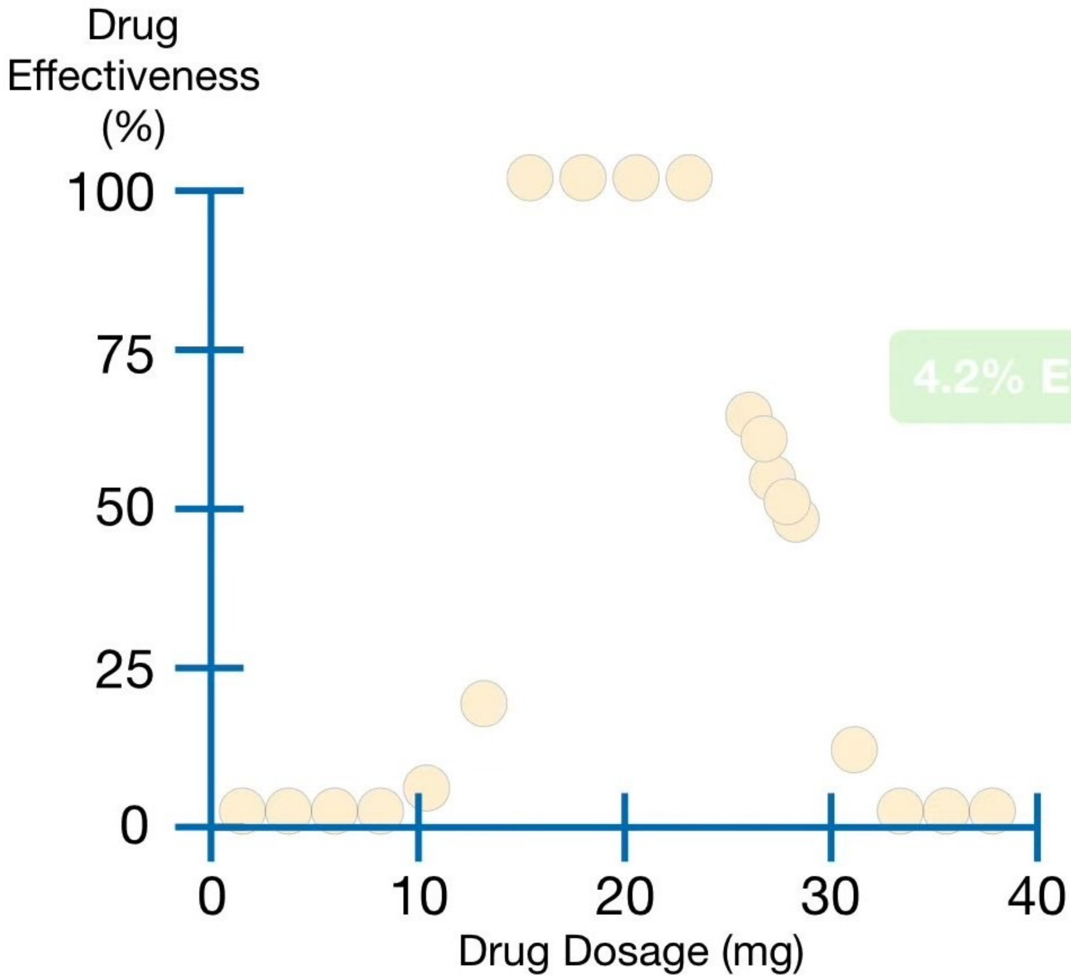
2.5% Effective

Dosage >= 23.5

52.8% Effective

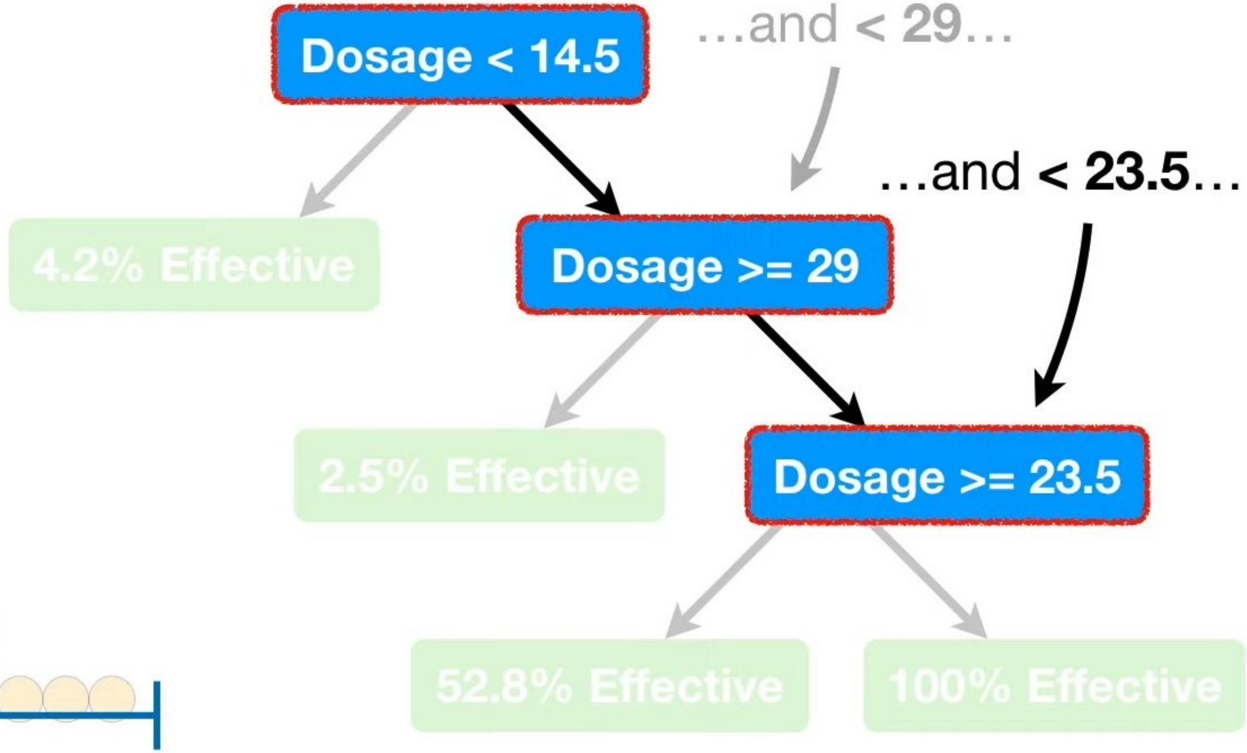
100% Effective

Lastly, if the **Dosage >= 14.5**…



…and **< 29**…

…and **< 23.5**…



Dosage < 14.5

4.2% Effective

Dosage >= 29

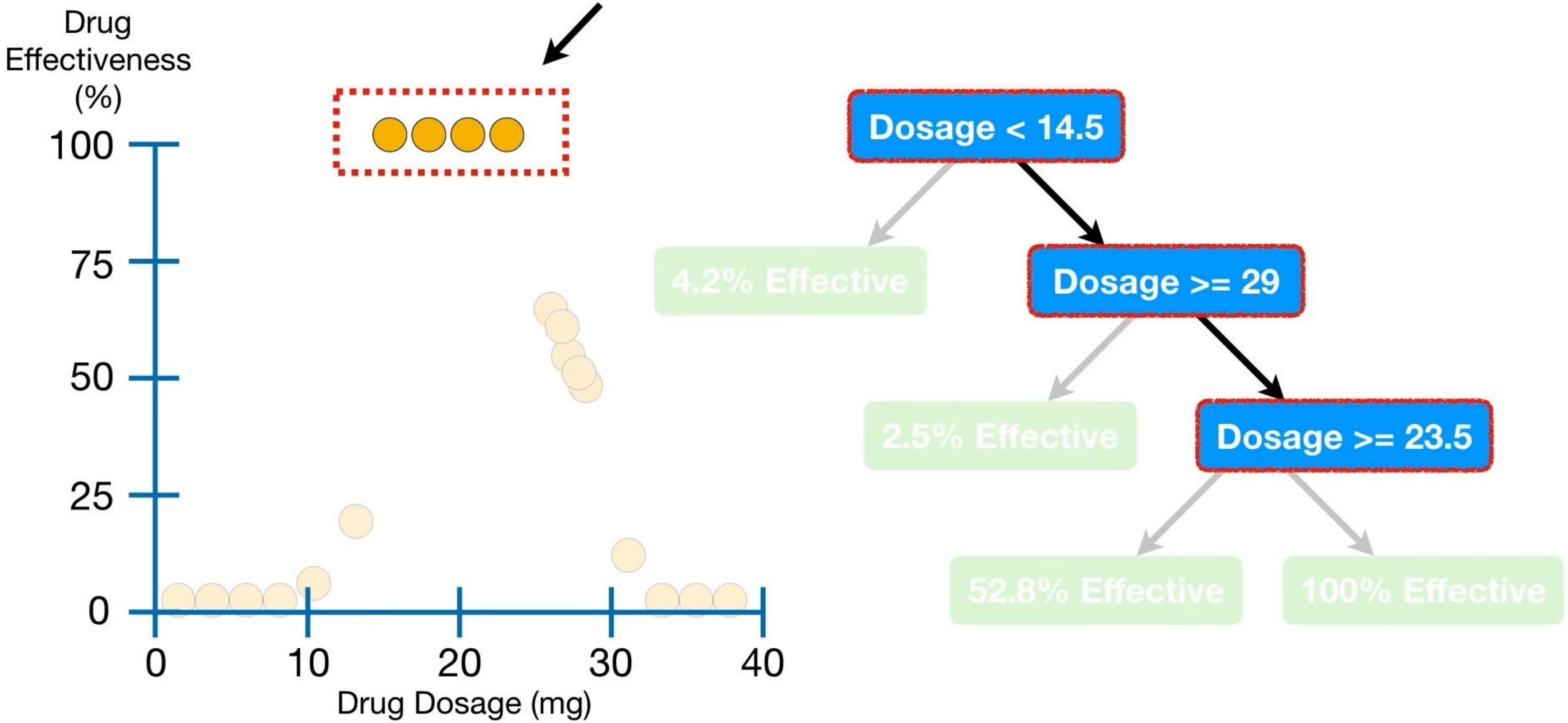
2.5% Effective

Dosage >= 23.5

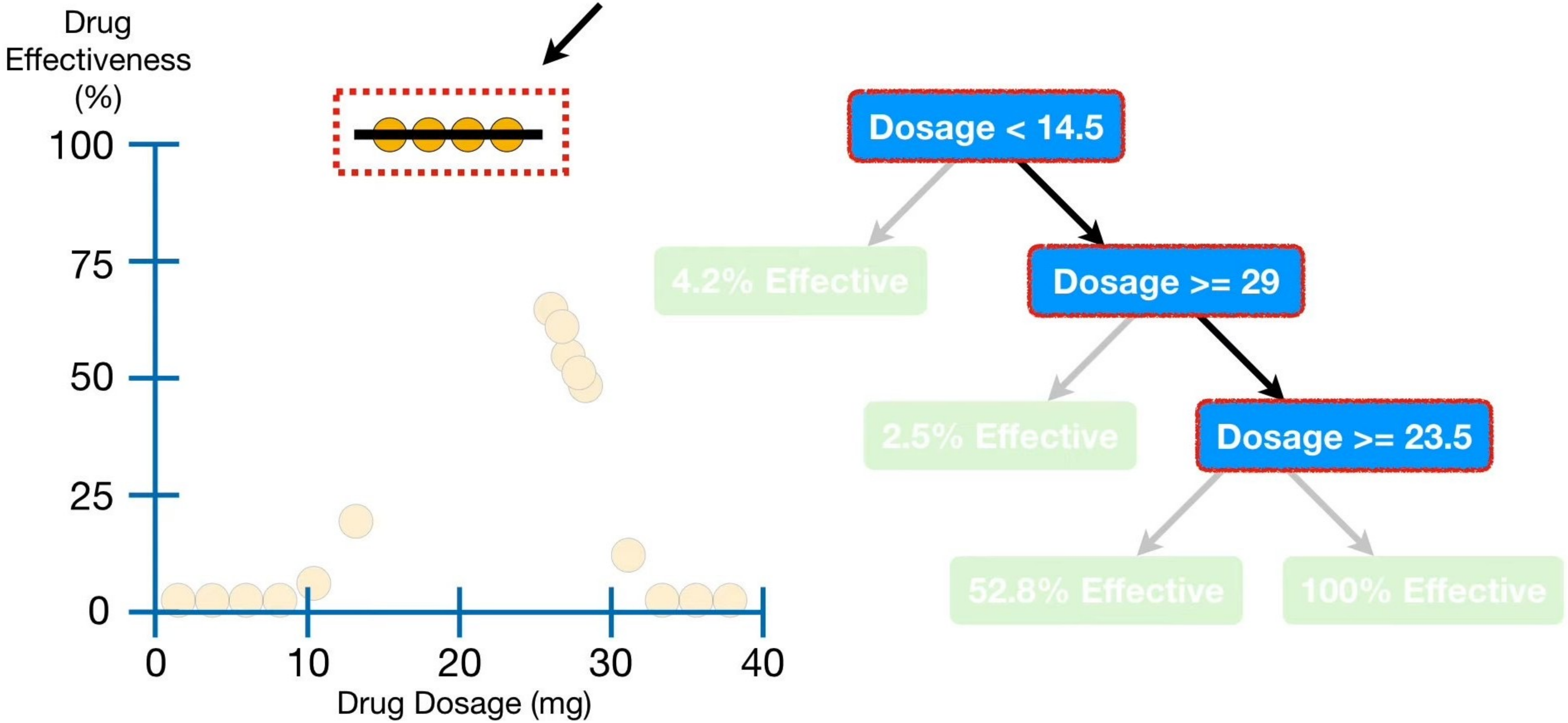
52.8% Effective

100% Effective

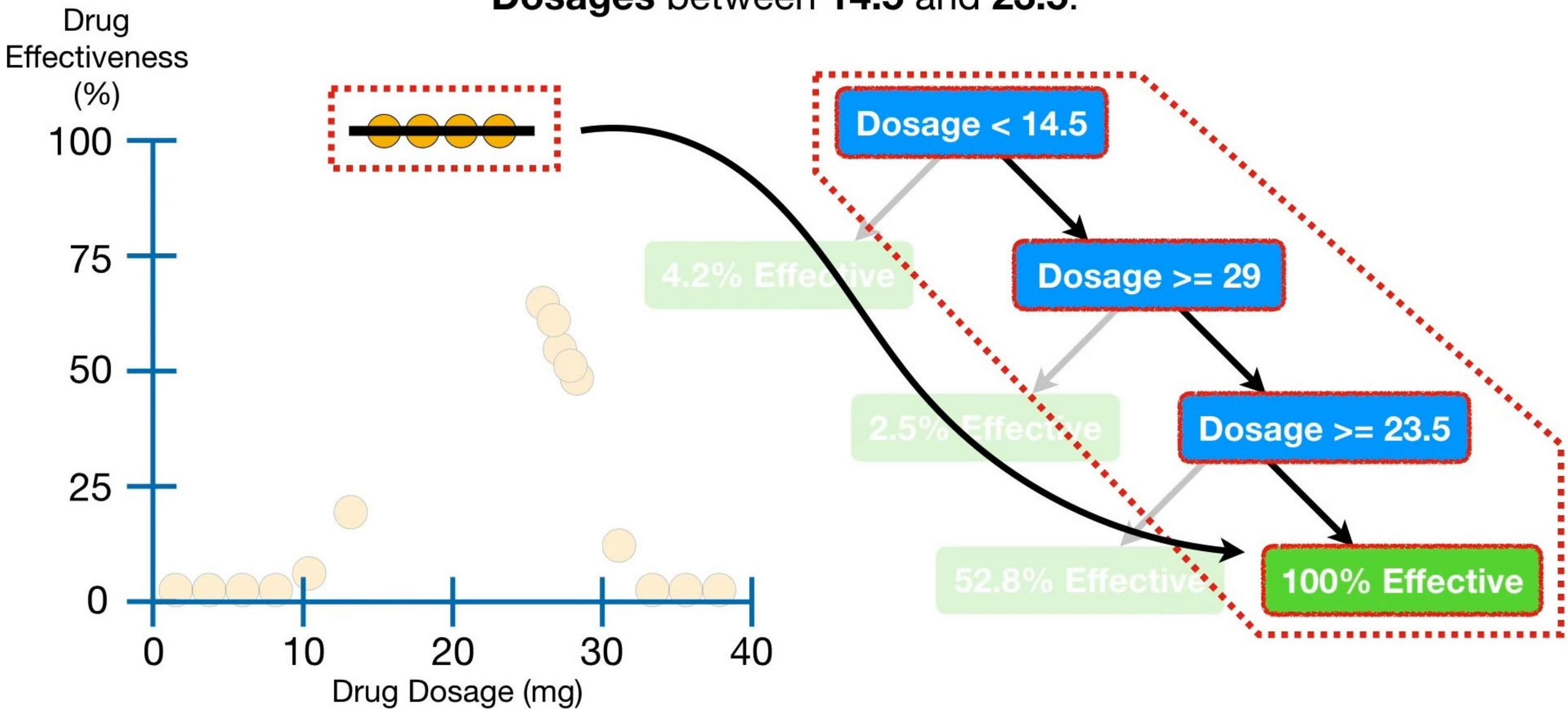
…then we are talking about these **4** observations in the training dataset…



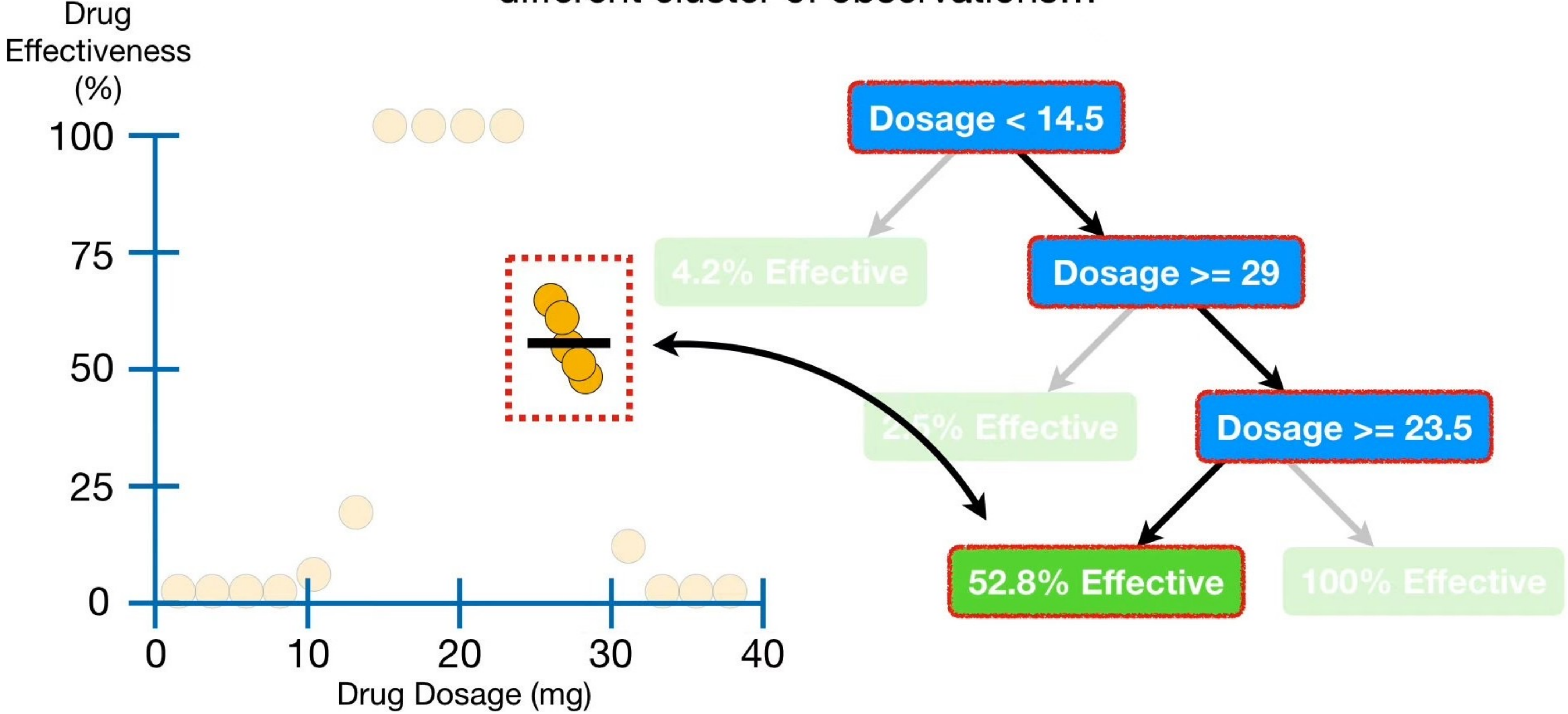
…and the average **Drug Effectiveness** for these **4** observations is **100%**…



…so the tree uses the average value, **100%**, as its prediction for people with **Dosages** between **14.5** and **23.5**.



Since each leaf corresponds to the average **Drug Effectiveness** in a different cluster of observations…



Dosage < 14.5

4.2% Effective

Dosage >= 29

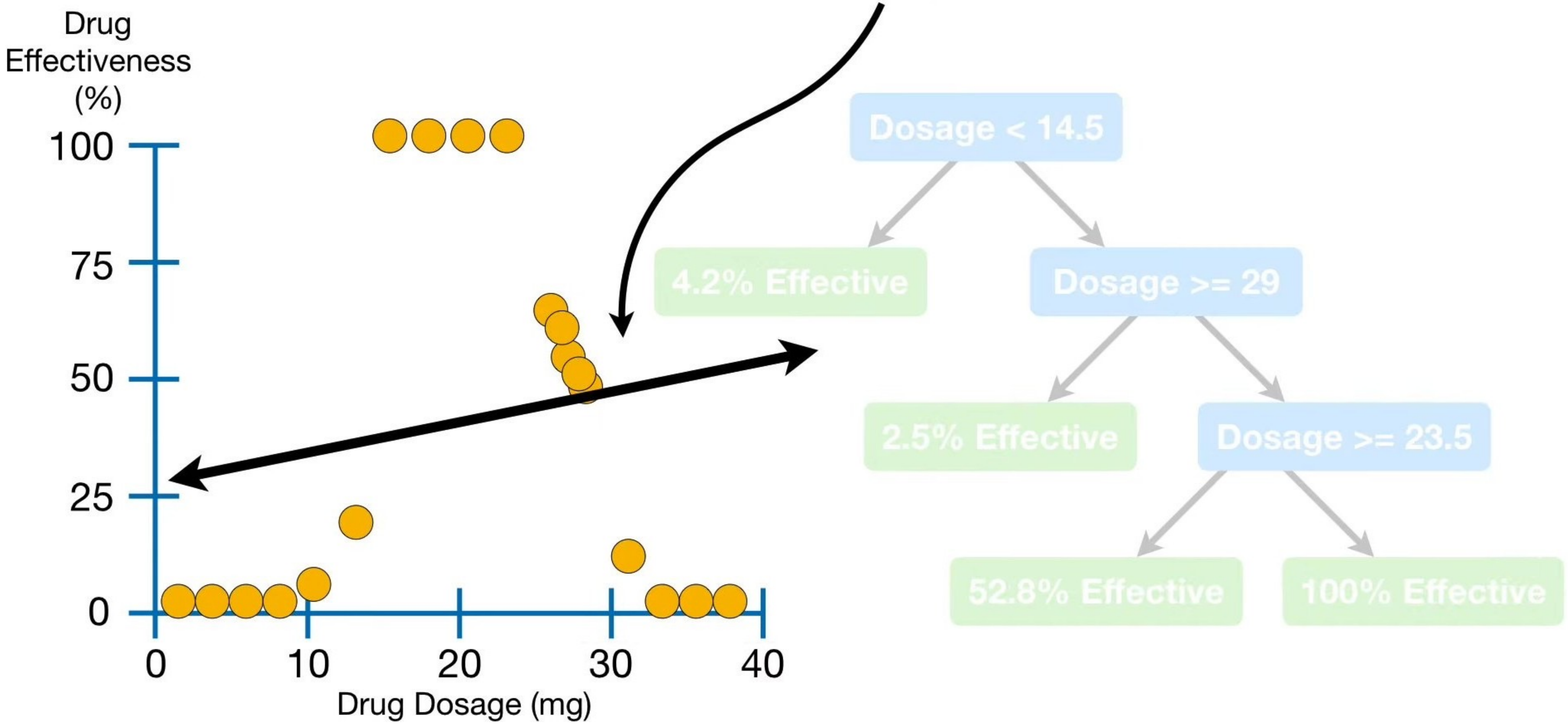
2.5% Effective

Dosage >= 23.5

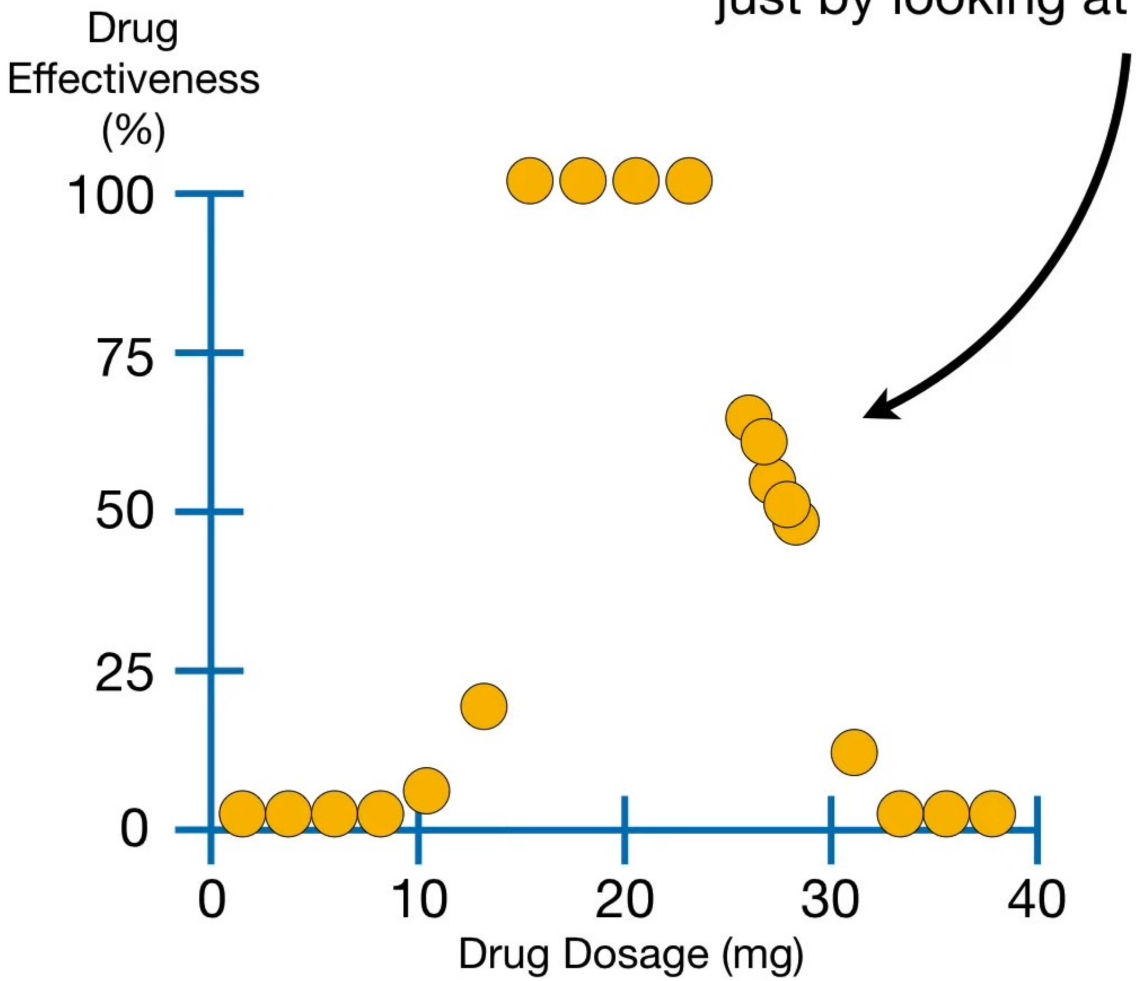
52.8% Effective

100% Effective

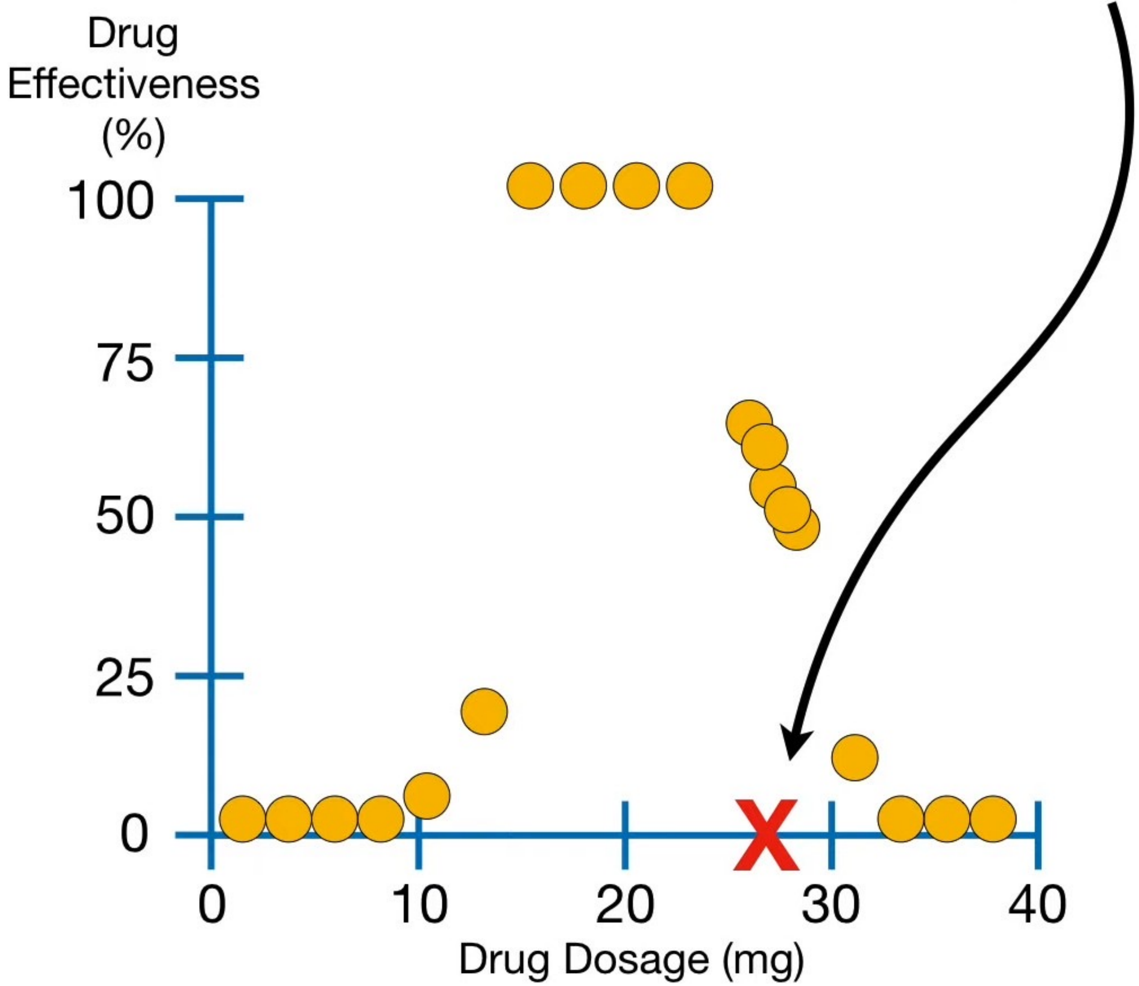
…the tree does a better job reflecting the data than the straight line.



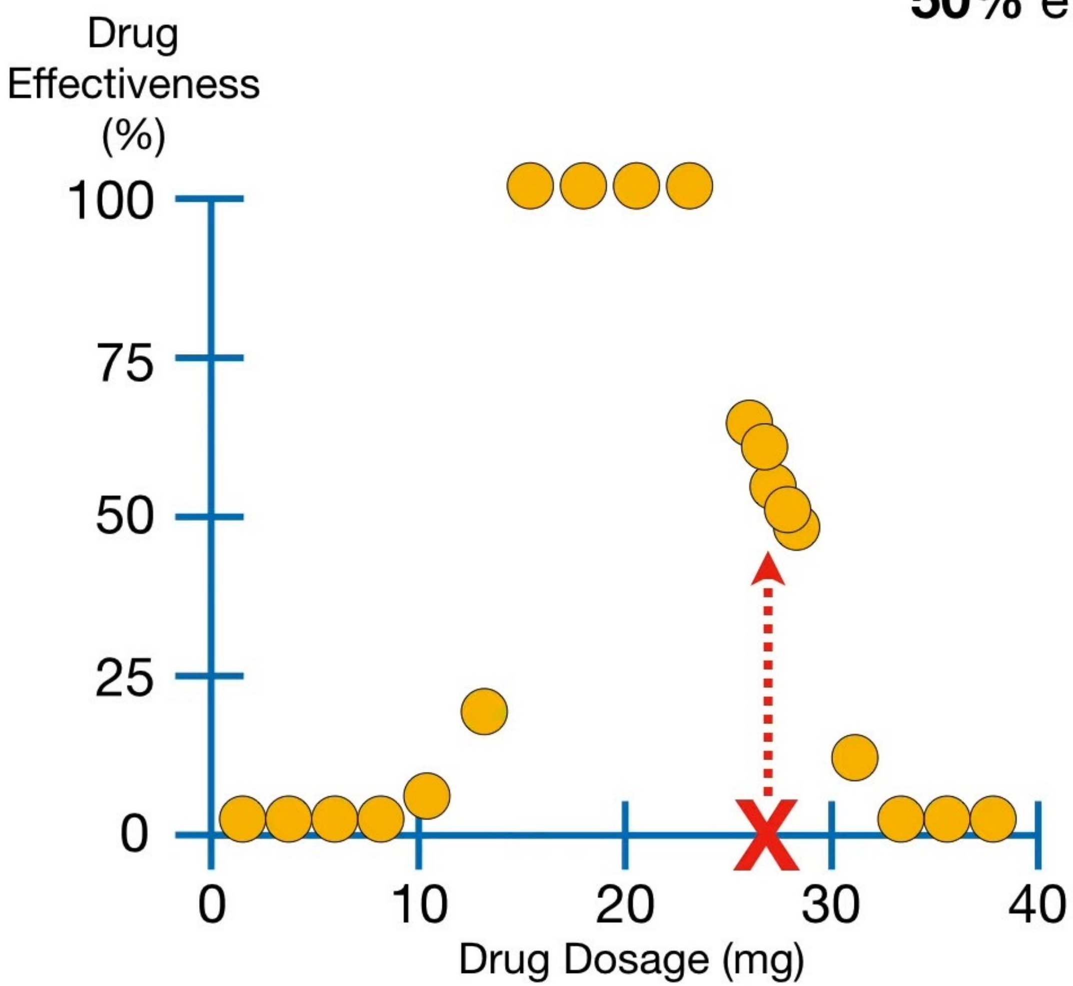
At this point you might be thinking, “The **Regression Tree** is cool, but I can also predict **Drug Effectiveness** just by looking at the graph…”



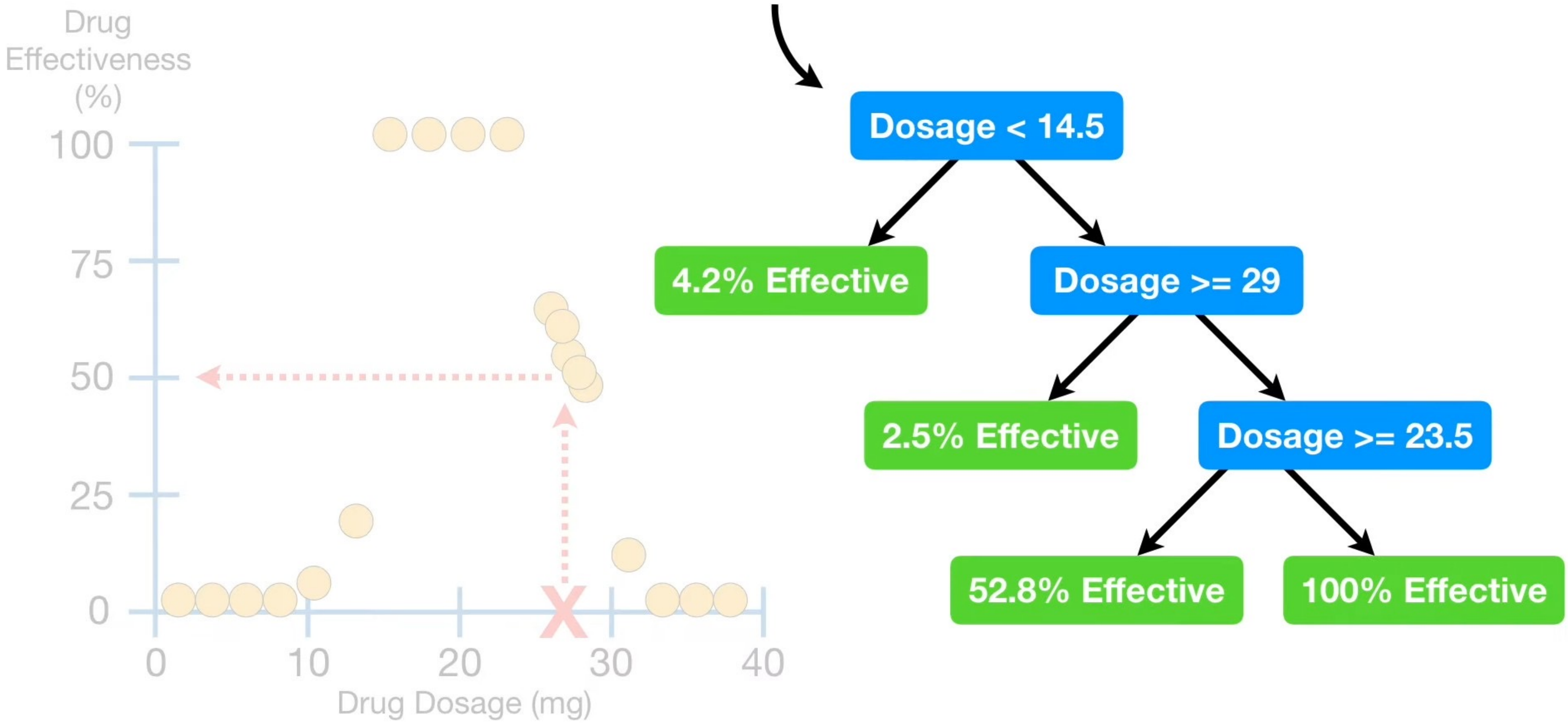
For example, if someone said they were taking a **27 mg** dose…



…then, just by looking at the graph, I can tell that the drug will be about **50%** effective.



So why make a big deal about the **Regression Tree**?



Dosage < 14.5

4.2% Effective

Dosage >= 29

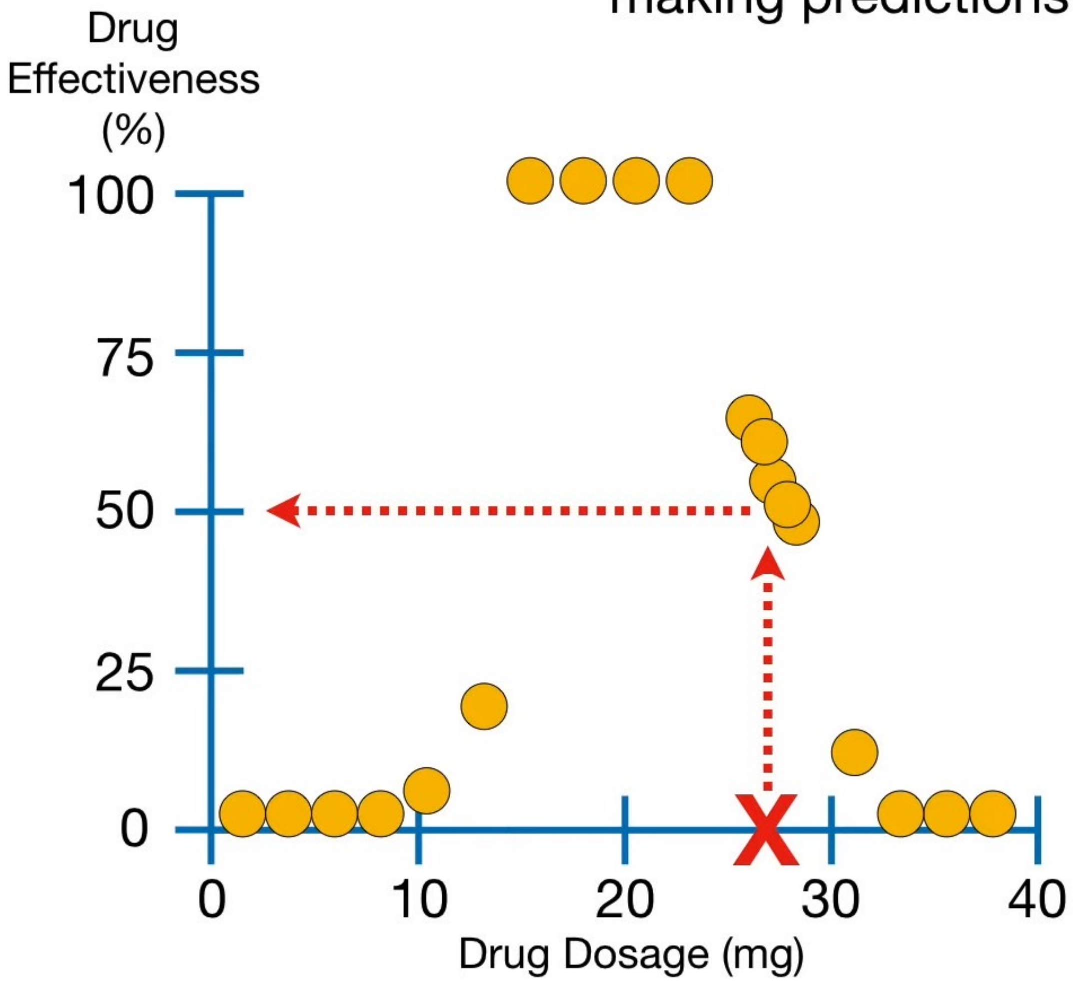
2.5% Effective

Dosage >= 23.5

52.8% Effective

100% Effective

When the data are super simple and we are only using one predictor, **Dosage**, to predict **Drug Effectiveness**, making predictions by eye isn't terrible.

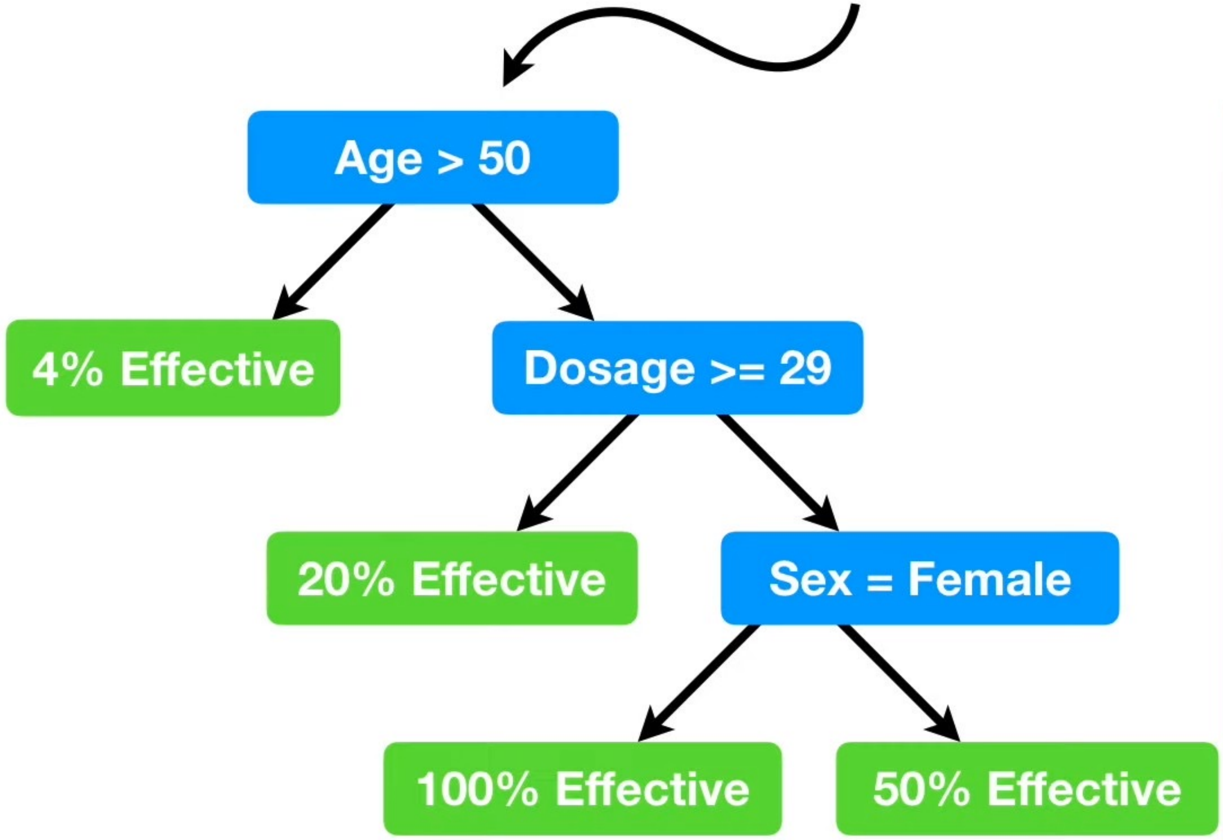


Dosage	Drug Effect.
10	58
20	60
35	57
5	44
etc…	etc…

But when we have **3** or more predictors, like **Dosage**, **Age** and **Sex**, to predict **Drug Effectiveness**, drawing a graph is very difficult, if not impossible.

Dosage	Age	Sex	Etc.	Drug Effect.
10	25	Female	…	98
20	73	Male	…	0
35	54	Female	…	100
5	12	Male	…	44
etc…	etc…	etc…	etc…	etc…

In contrast, a **Regression Tree** easily accommodates the additional predictors.


Age > 50

4% Effective

Dosage >= 29

20% Effective

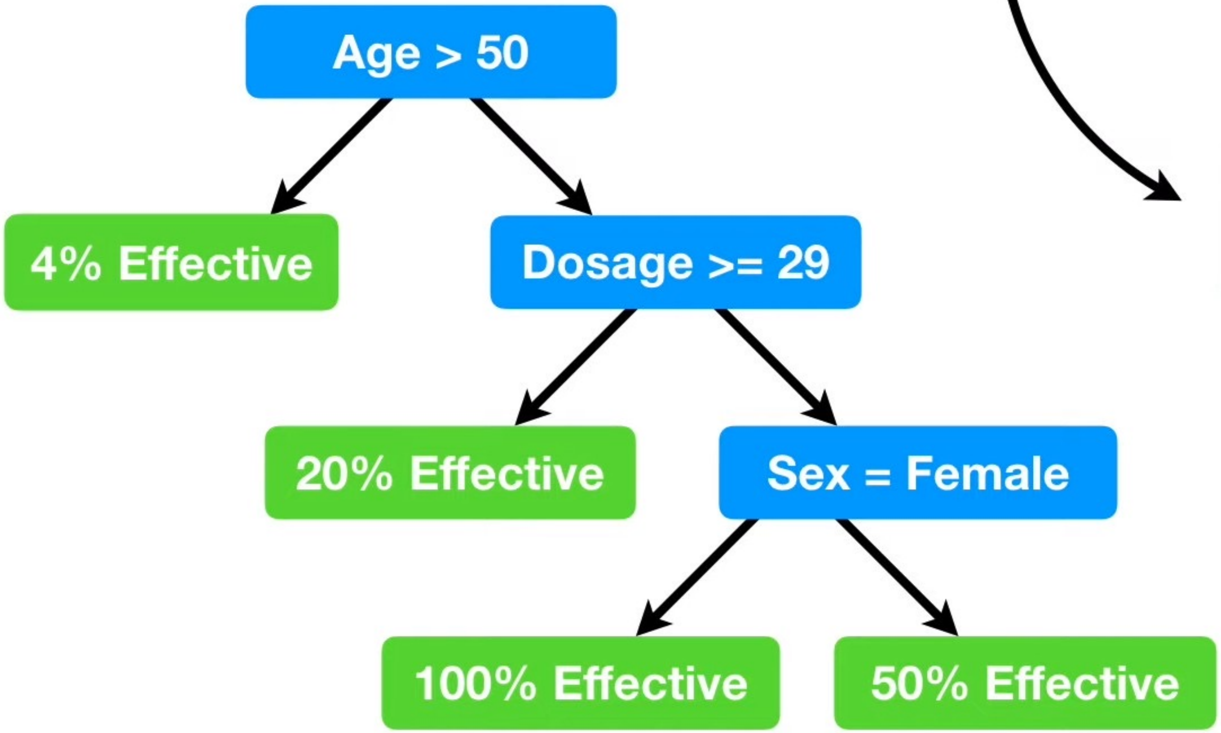
Sex = Female

100% Effective

50% Effective

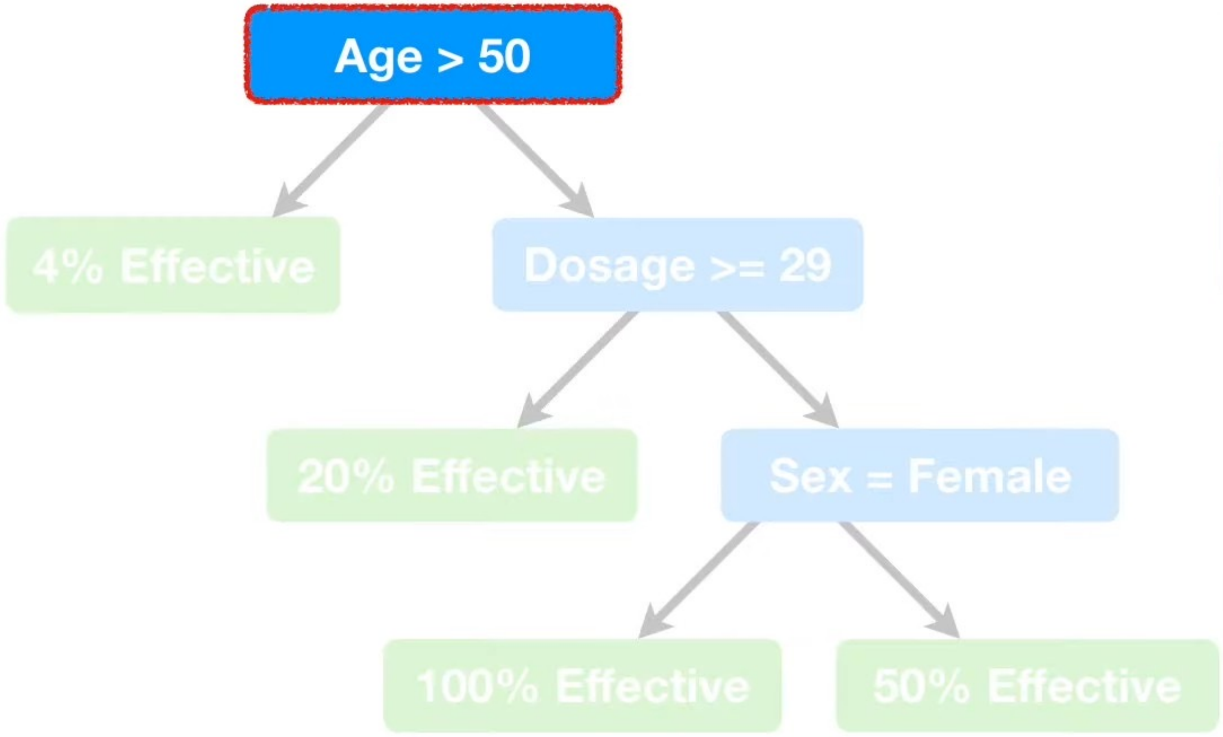
Dosage	Age	Sex	Etc.	Drug Effect.
10	25	Female	…	98
20	73	Male	…	0
35	54	Female	…	100
5	12	Male	…	44
etc…	etc…	etc…	etc…	etc…

For example, if we wanted to predict the **Drug Effectiveness** for this patient…



Dosage	Age	Sex	Etc.	Drug Effect.
10	25	Female	…	98
20	73	Male	…	0
35	54	Female	…	100
5	12	Male	…	44
etc…	etc…	etc…	etc…	etc…

…we start by asking if they are older than **50**…



Age > 50

4% Effective

Dosage >= 29

20% Effective

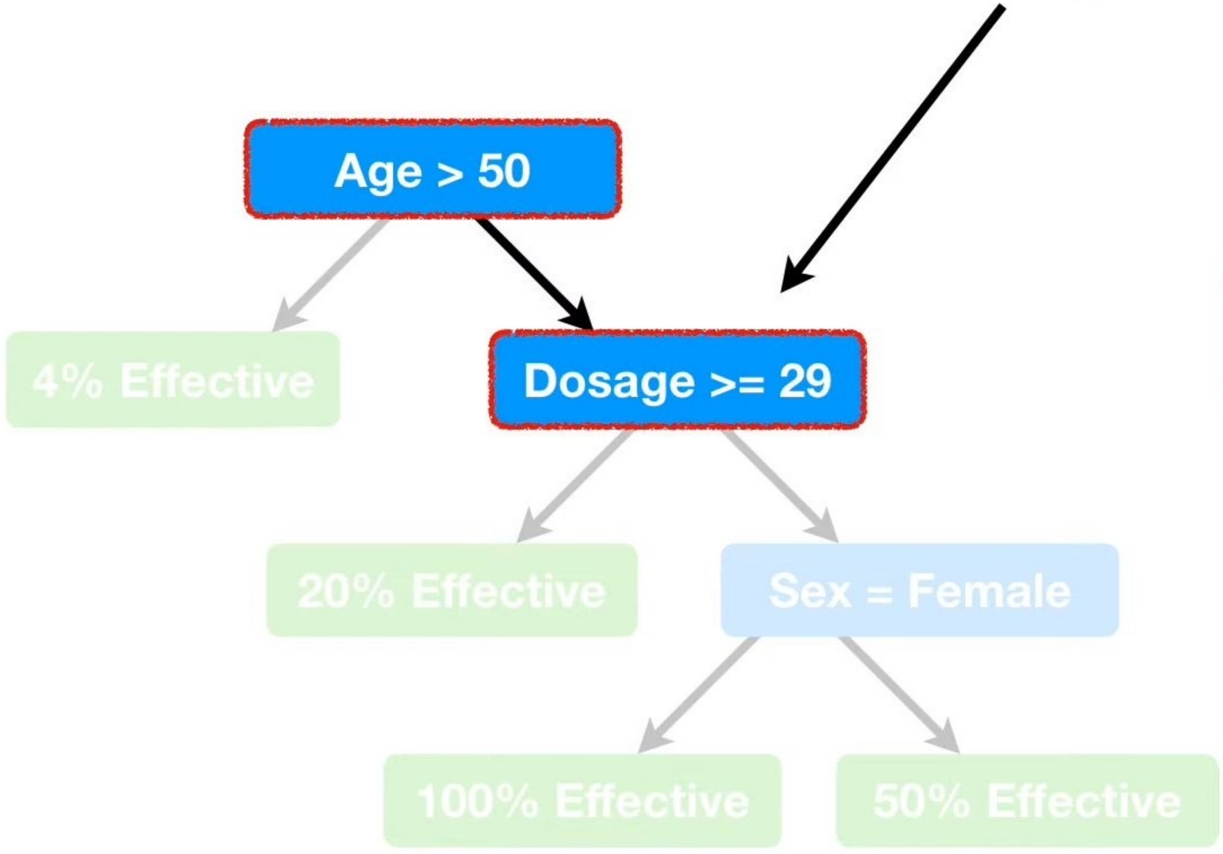
Sex = Female

100% Effective

50% Effective

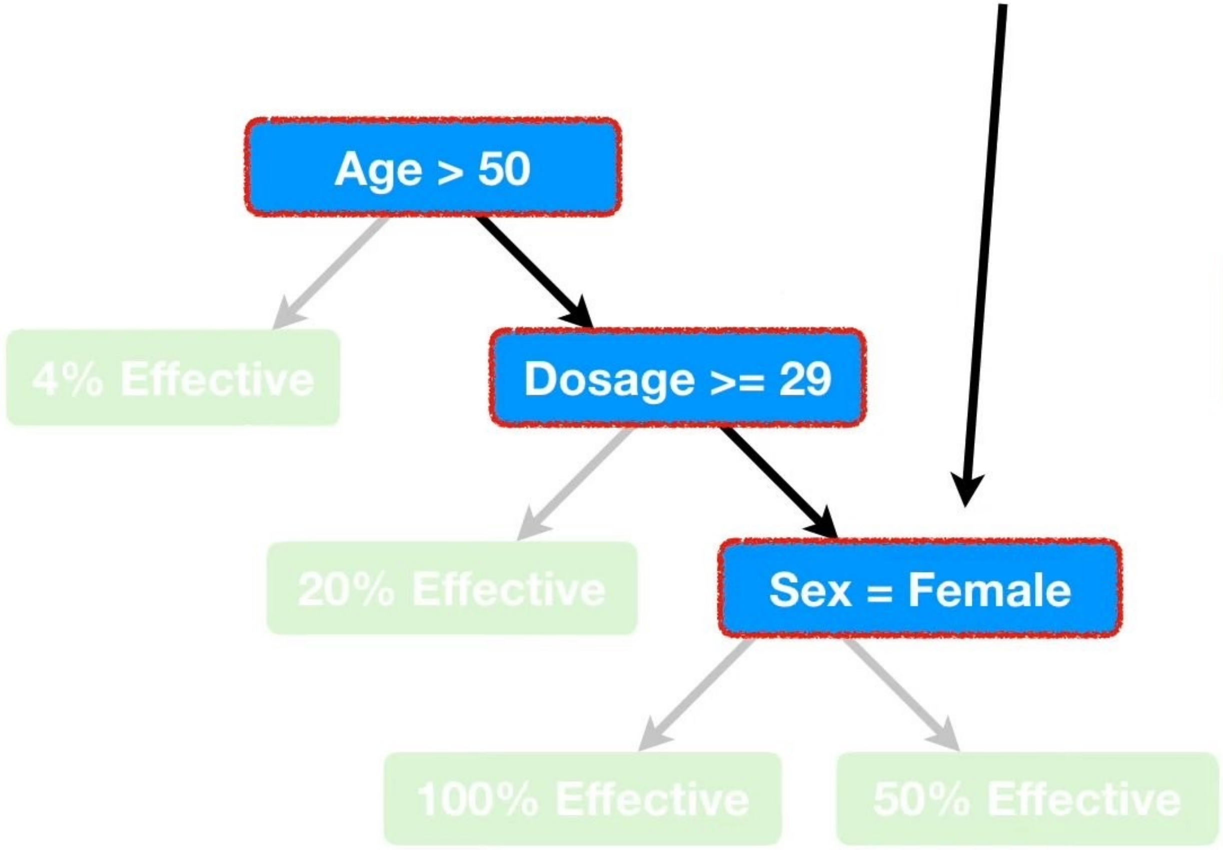
Dosage	Age	Sex	Etc.	Drug Effect.
10	25	Female	…	98
20	73	Male	…	0
35	54	Female	…	100
5	12	Male	…	44
etc…	etc…	etc…	etc…	etc…

…and since they *not* over **50**, we follow the branch on the *right* and ask if their **Dosage >= 29**…



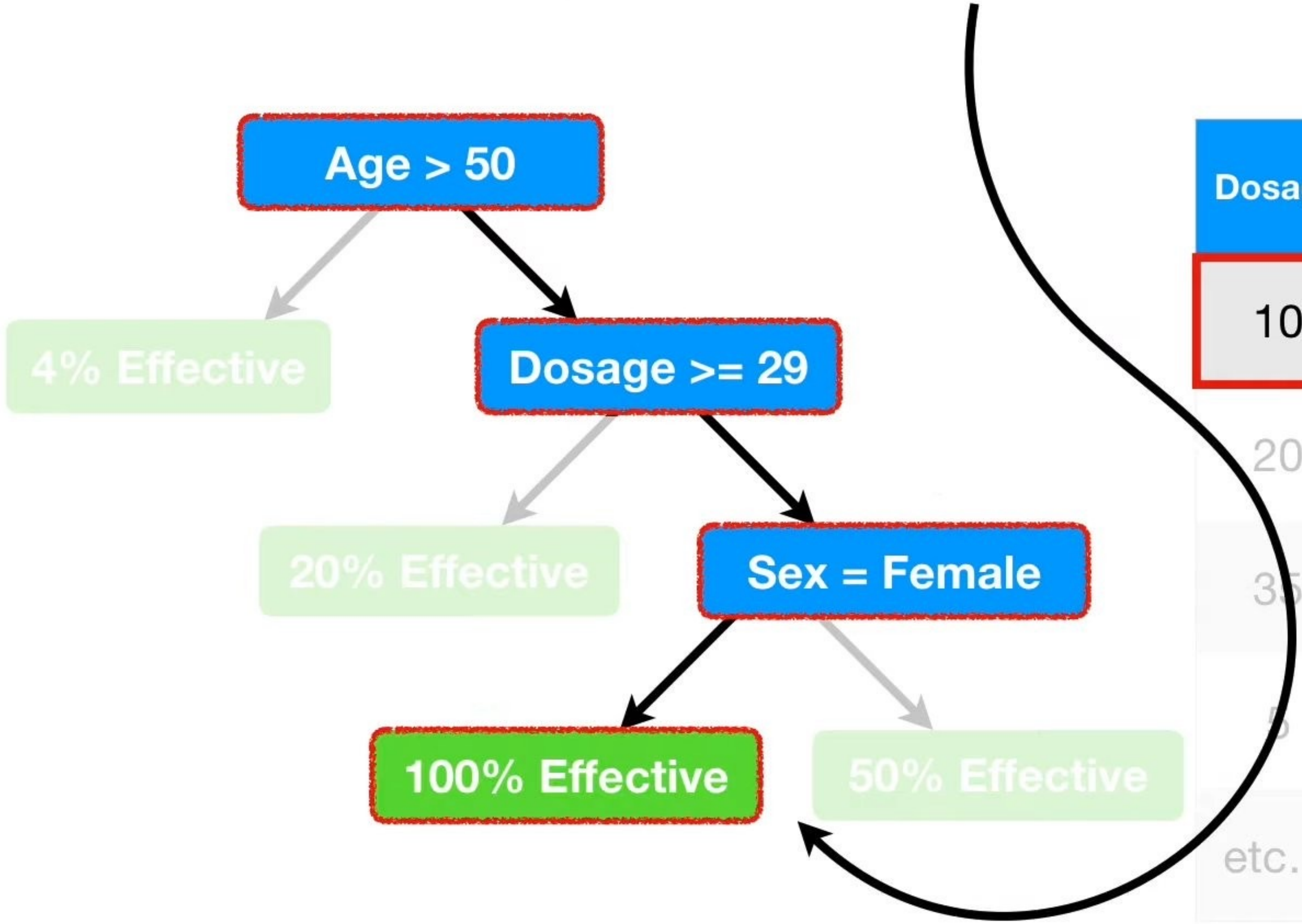
Dosage	Age	Sex	Etc.	Drug Effect.
10	25	Female	…	98
20	73	Male	…	0
35	54	Female	…	100
5	12	Male	…	44
etc…	etc…	etc…	etc…	etc…

…and since their dosage is *not* **>= 29**, we follow the branch on the *right* and ask if they are **Female**…



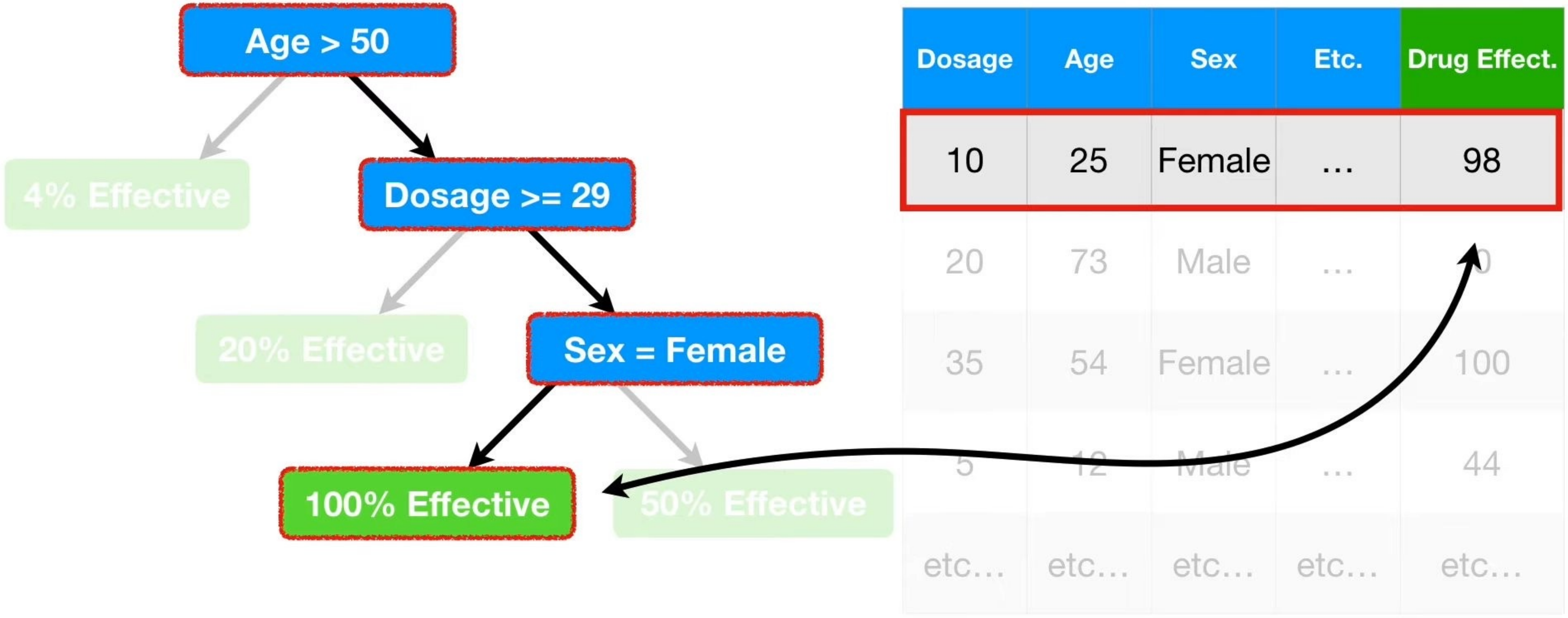
Dosage	Age	Sex	Etc.	Drug Effect.
10	25	Female	…	98
20	73	Male	…	0
35	54	Female	…	100
5	12	Male	…	44
etc…	etc…	etc…	etc…	etc…

…and since they *are* **Female**, we follow the branch on the *left* and predict that the dosage will be **100% Effective**…



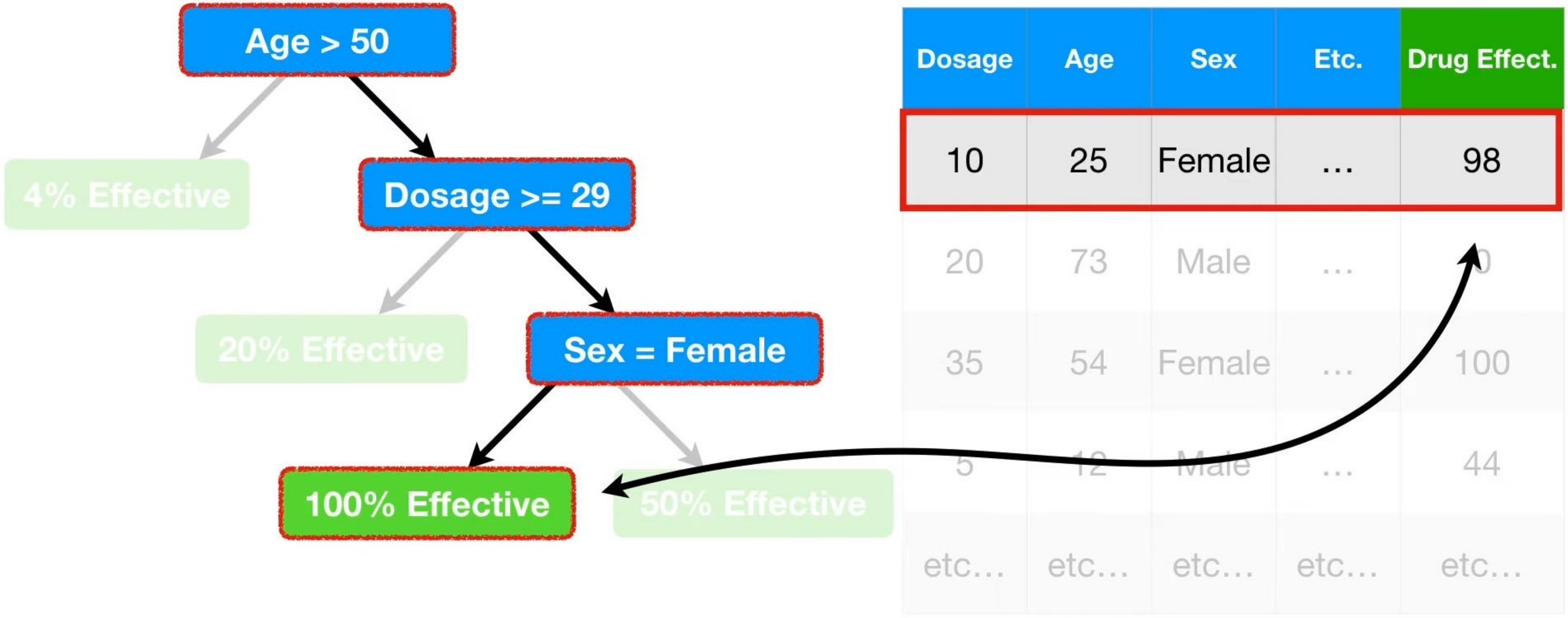
Dosage	Age	Sex	Etc.	Drug Effect.
10	25	Female	…	98
20	73	Male	…	0
35	54	Female	…	100
5	12	Male	…	44
etc…	etc…	etc…	etc…	etc…

…and that’s not too far off from the truth, **98%**.



Dosage	Age	Sex	Etc.	Drug Effect.
10	25	Female	…	98
20	73	Male	…	0
35	54	Female	…	100
5	12	Male	…	44
etc…	etc…	etc…	etc…	etc…

OK, now that we know that **Regression Trees** can easily handle complicated data…

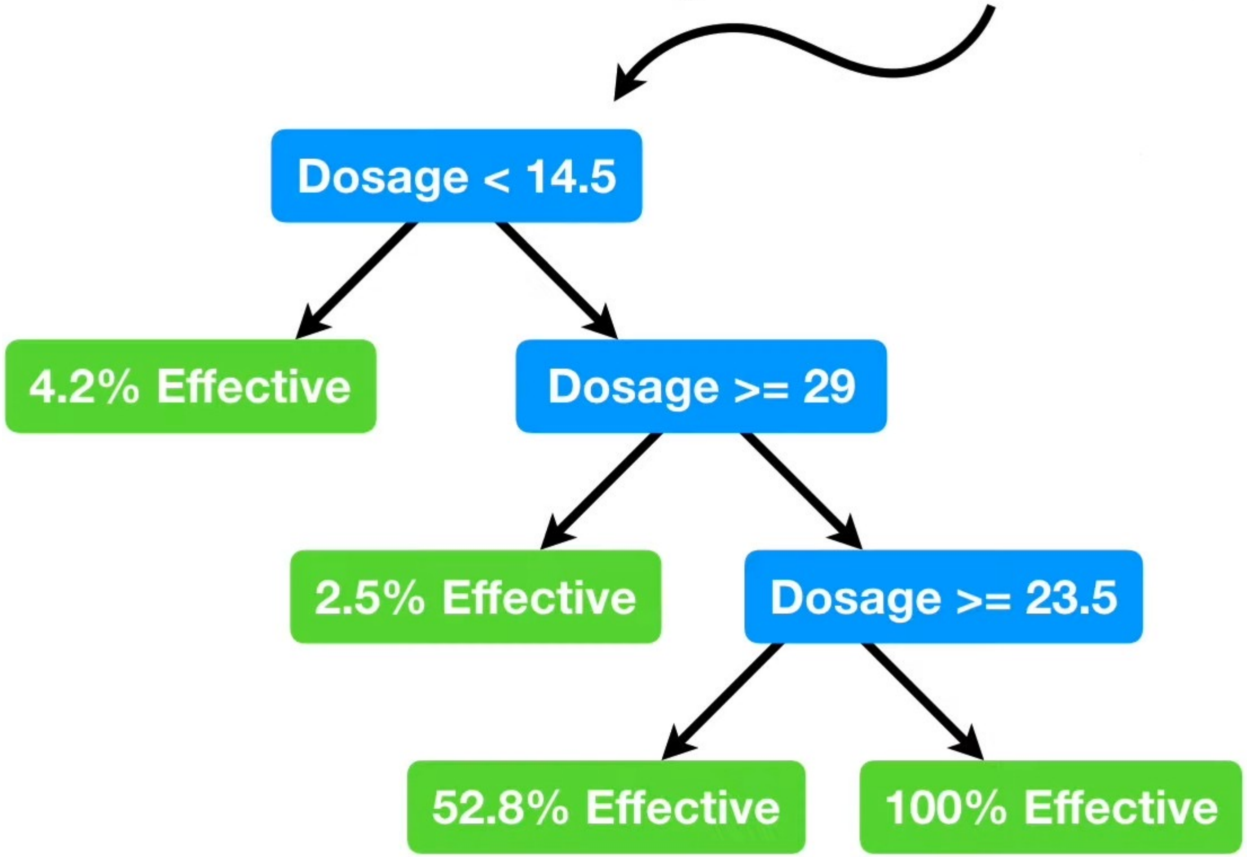


Dosage	Age	Sex	Etc.	Drug Effect.
10	25	Female	…	98
20	73	Male	…	0
35	54	Female	…	100
5	12	Male	…	44
etc…	etc…	etc…	etc…	etc…

…let's go back to the original data, with just one predictor, **Dosage**…

Dosage	Drug Effect.
10	58
20	60
35	57
5	44
etc…	etc…

…and talk about how to build this **Regression Tree** from scratch…



Dosage < 14.5

4.2% Effective

Dosage >= 29

2.5% Effective

Dosage >= 23.5

52.8% Effective

100% Effective

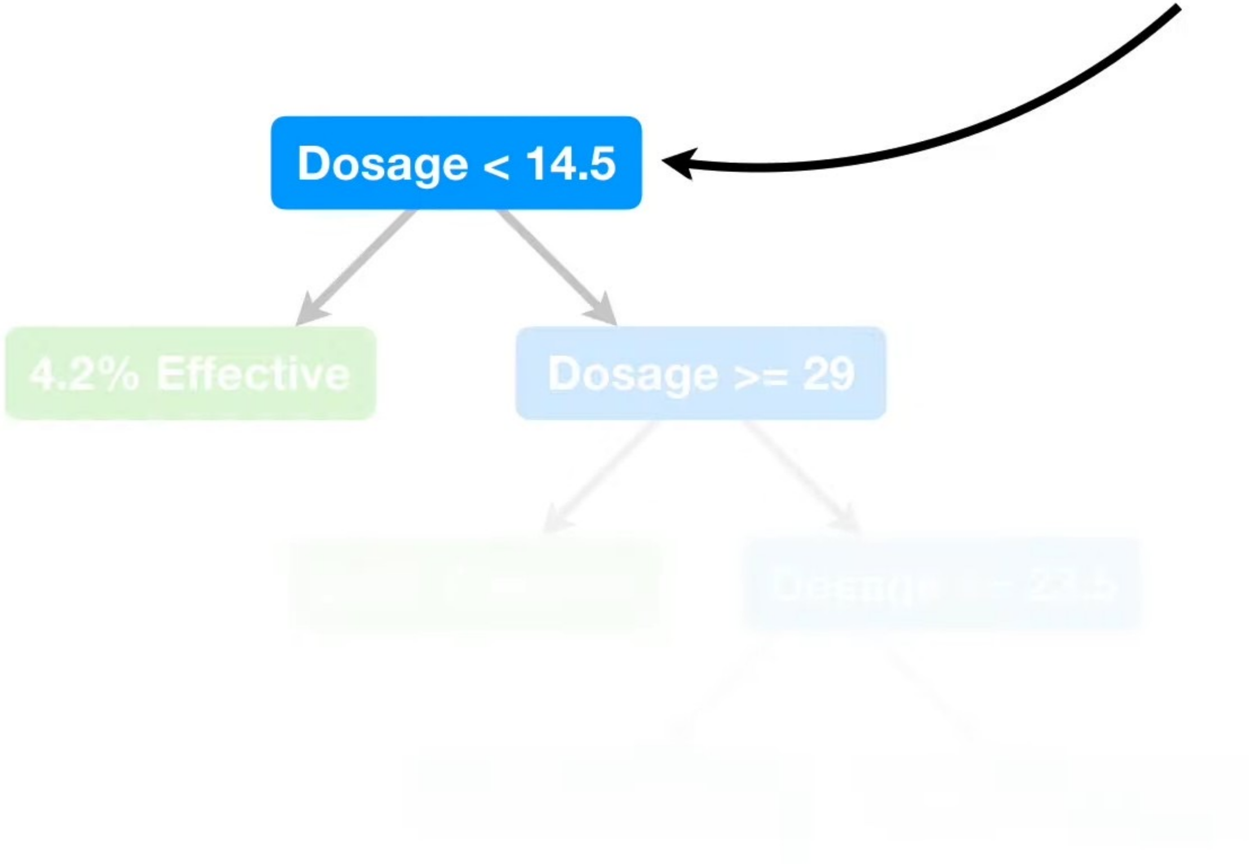
Dosage	Drug Effect.
10	58
20	60
35	57
5	44
etc…	etc…

…and since **Regression Trees** are built from the top down…

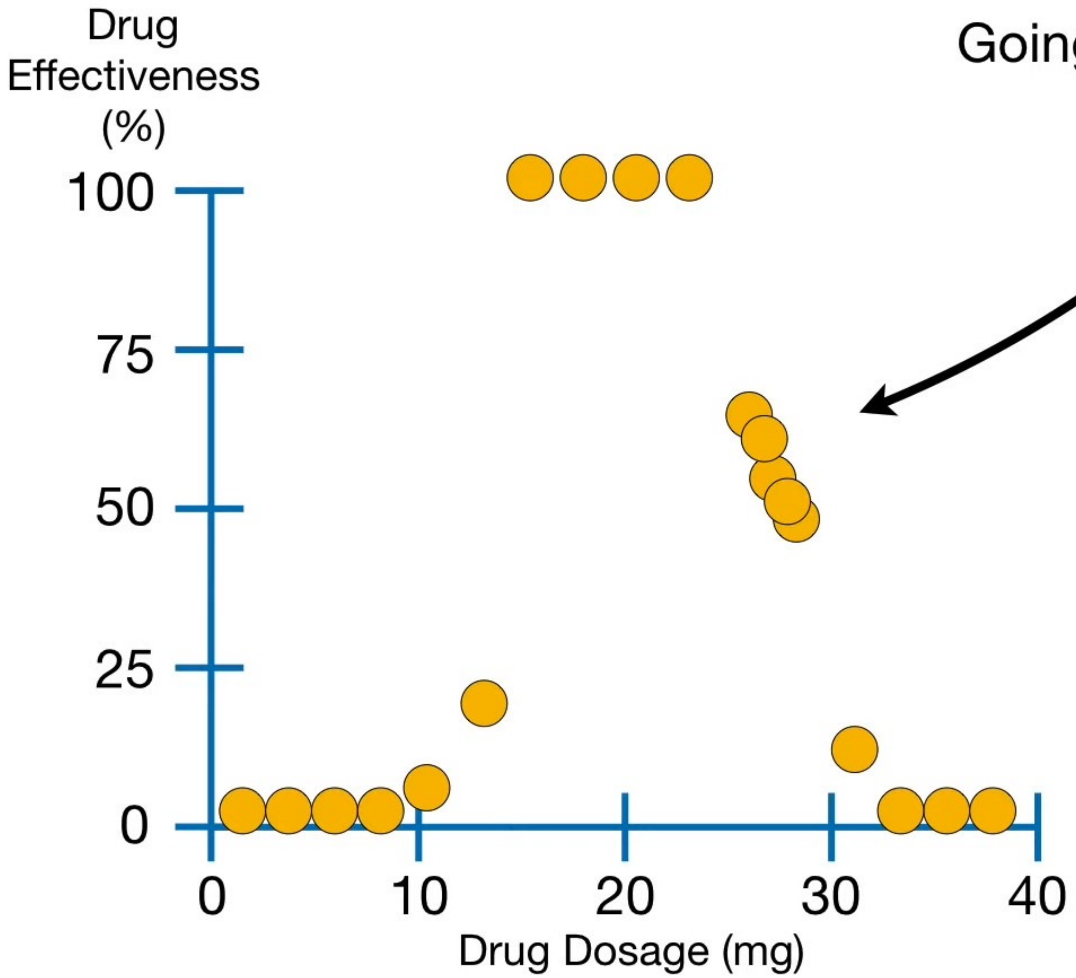
Dosage < 14.5

Dosage	Drug Effect.
10	58
20	60
35	57
5	44
etc…	etc…

…the first thing we do is figure out why we start by asking if **Dosage < 14.5**.

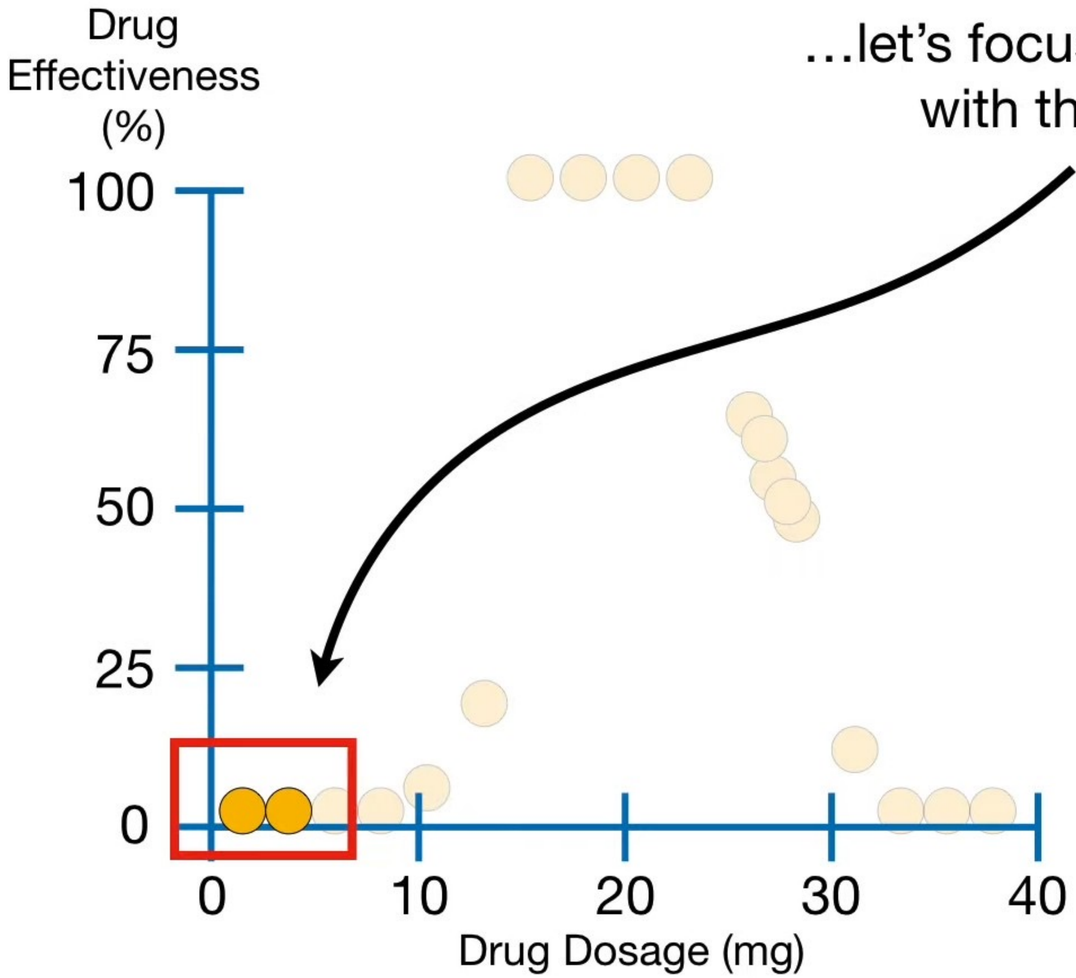


Dosage	Drug Effect.
10	58
20	60
35	57
5	44
etc…	etc…

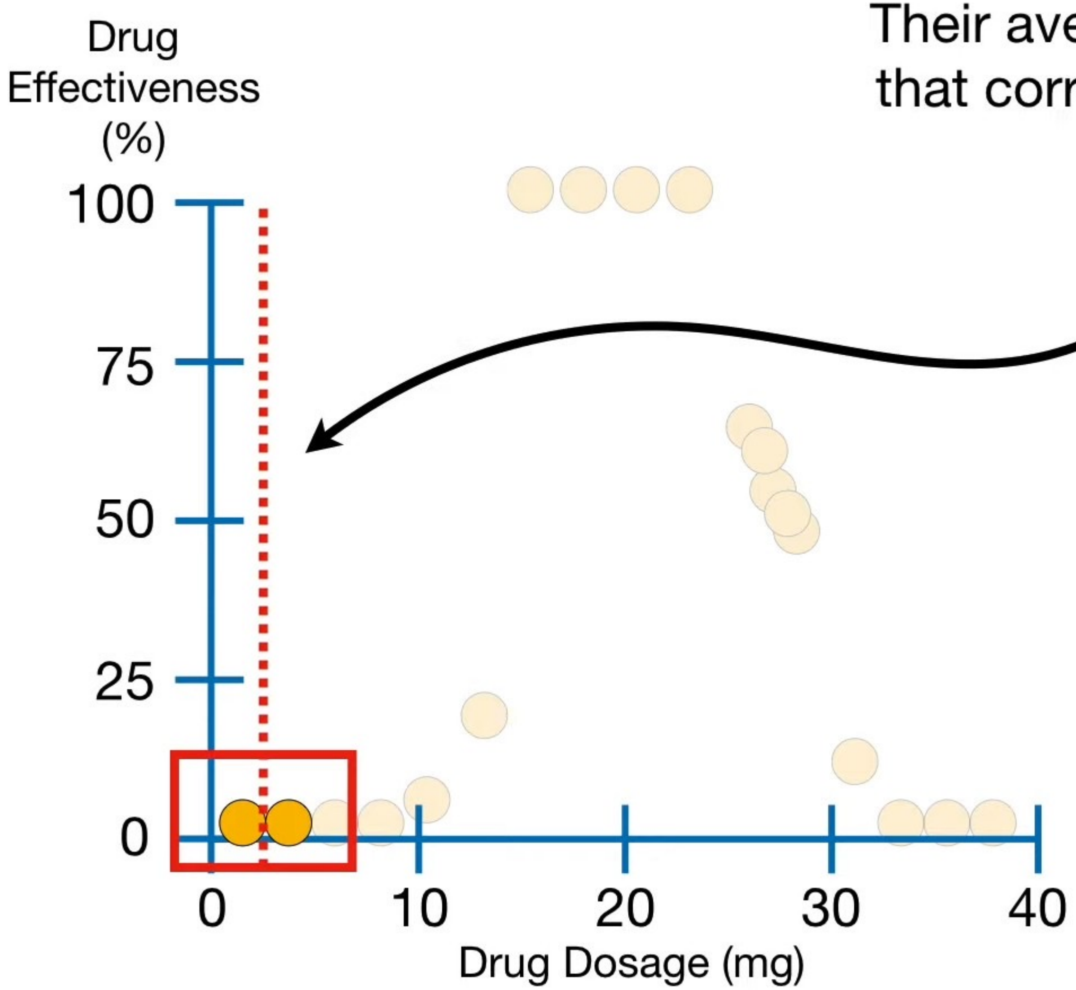


Going back to the graph of the data…

Dosage	Drug Effect.
10	58
20	60
35	57
5	44
etc…	etc…

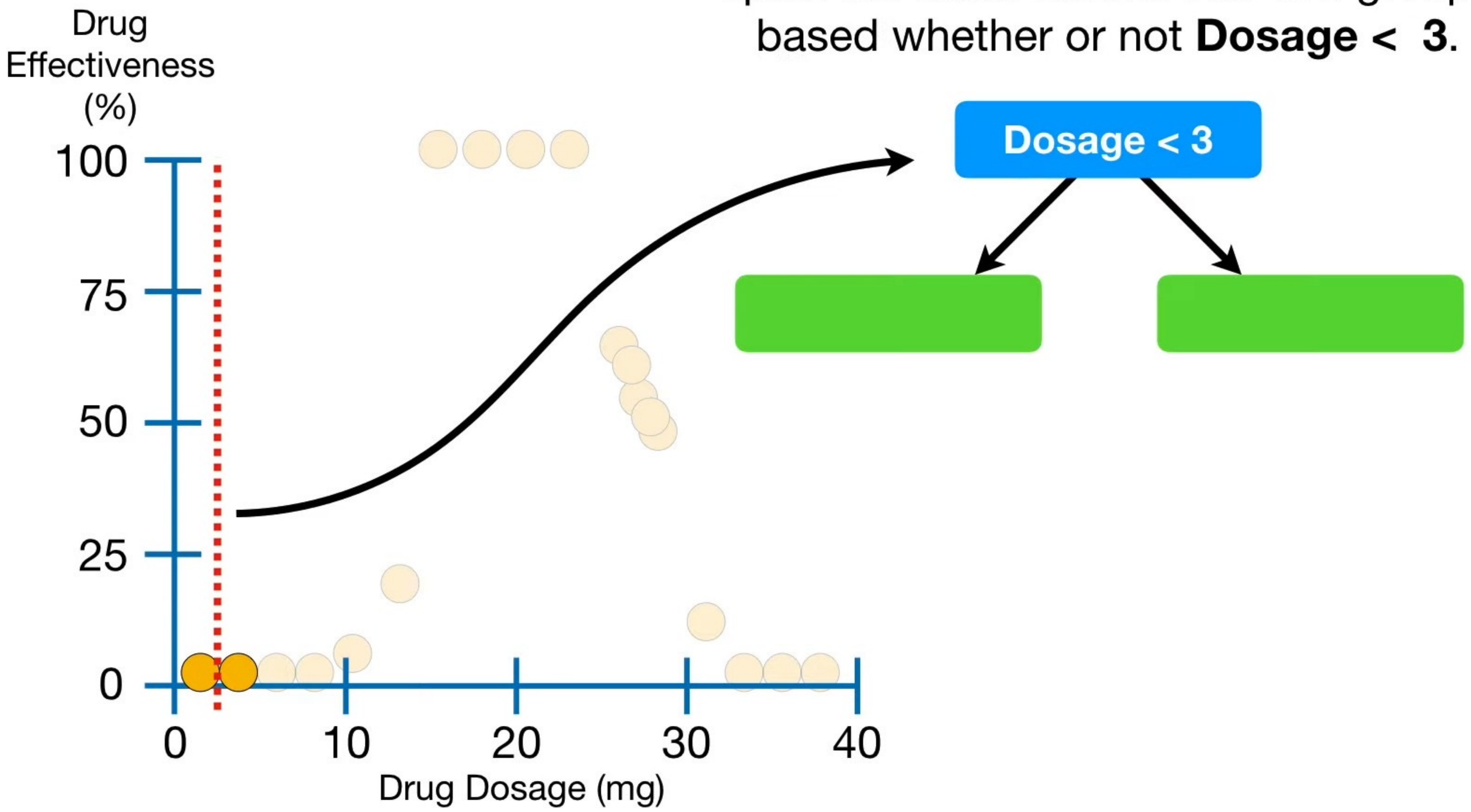


…let's focus on the two observations with the smallest **Dosages**.

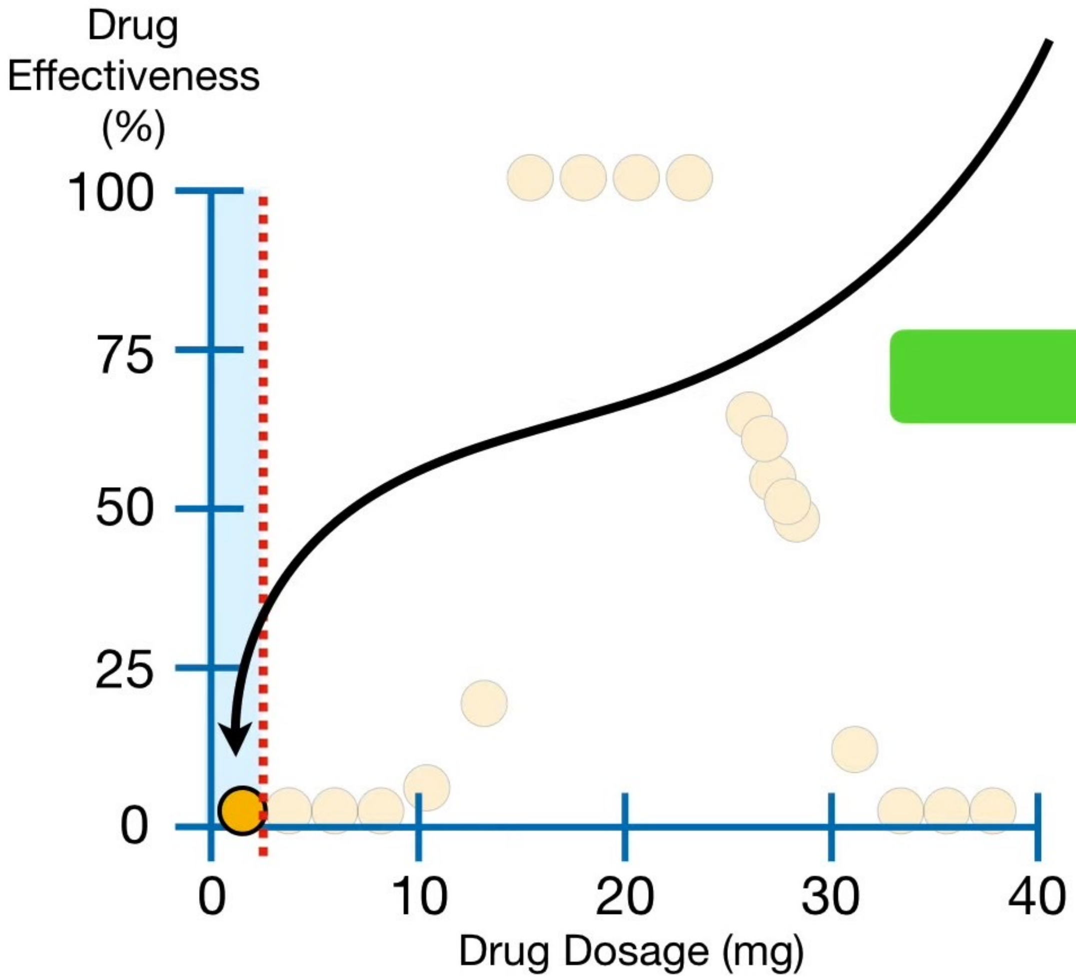


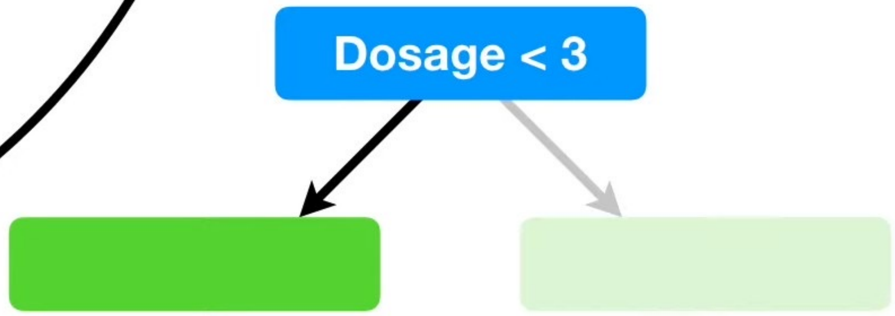
Their average **Dosage** is **3**, and that corresponds to this dotted **red line**.

Now we can build a very simple tree that splits the observations into two groups based whether or not **Dosage < 3**.



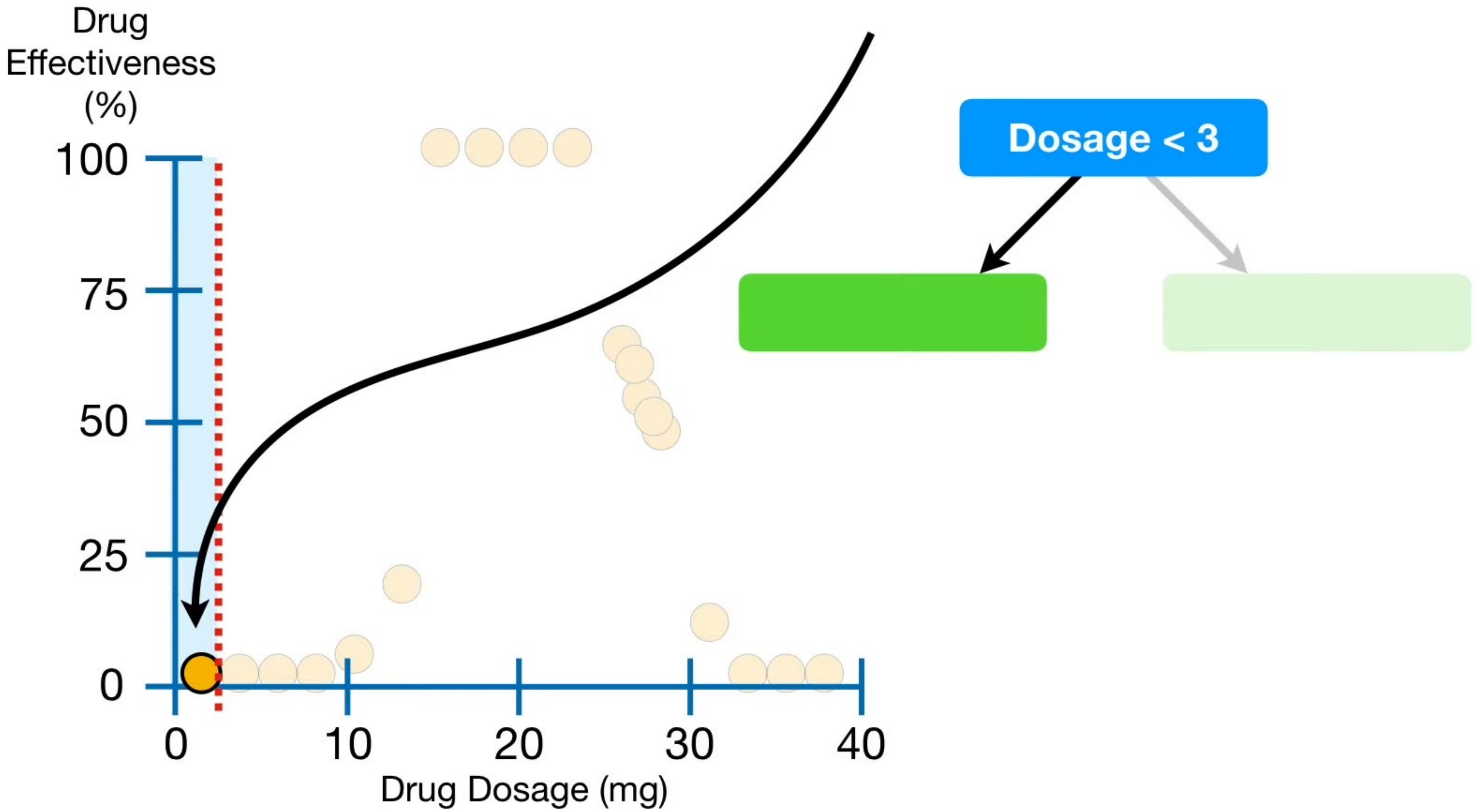
The point on the far left is the only one with **Dosage < 3**…



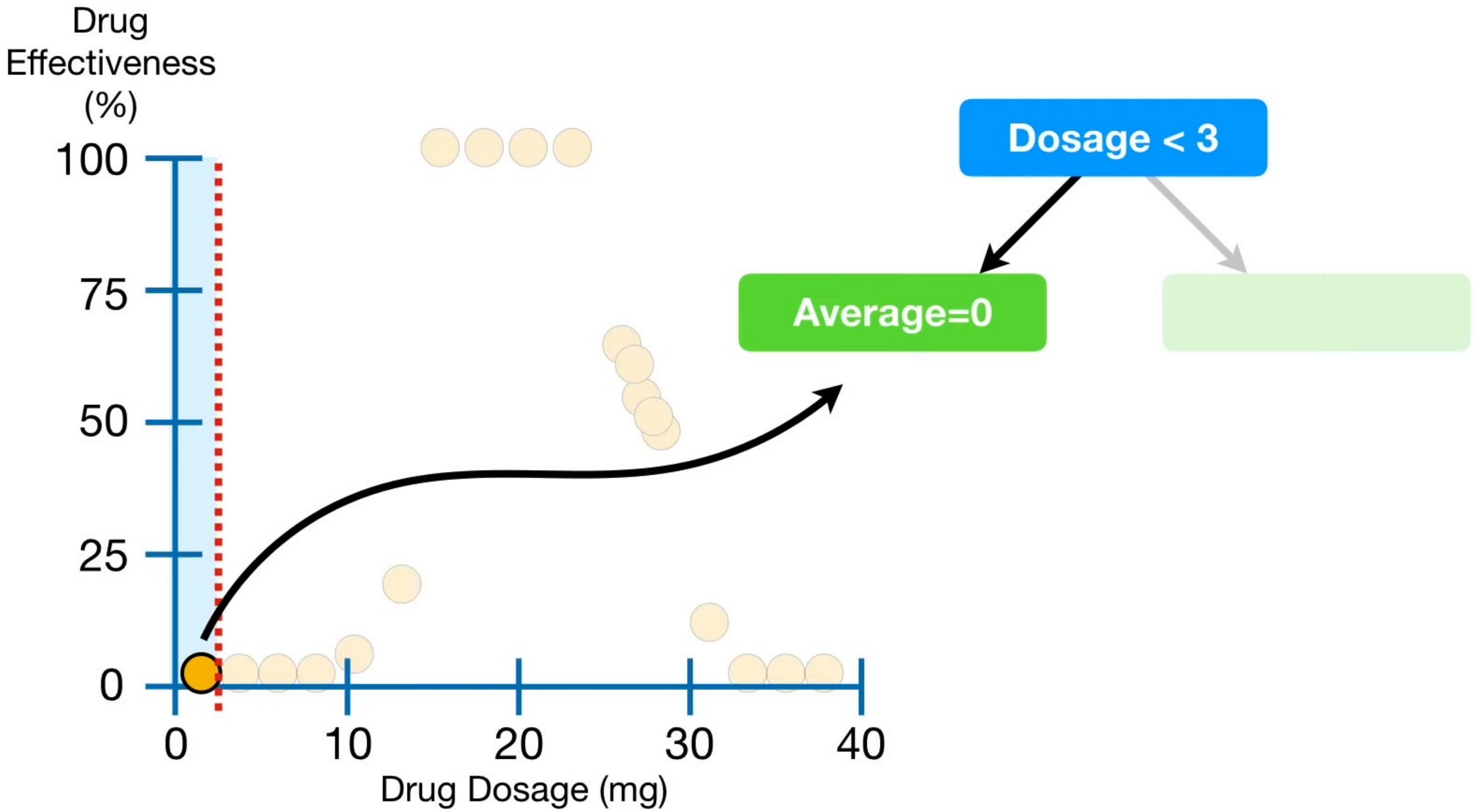


Dosage < 3

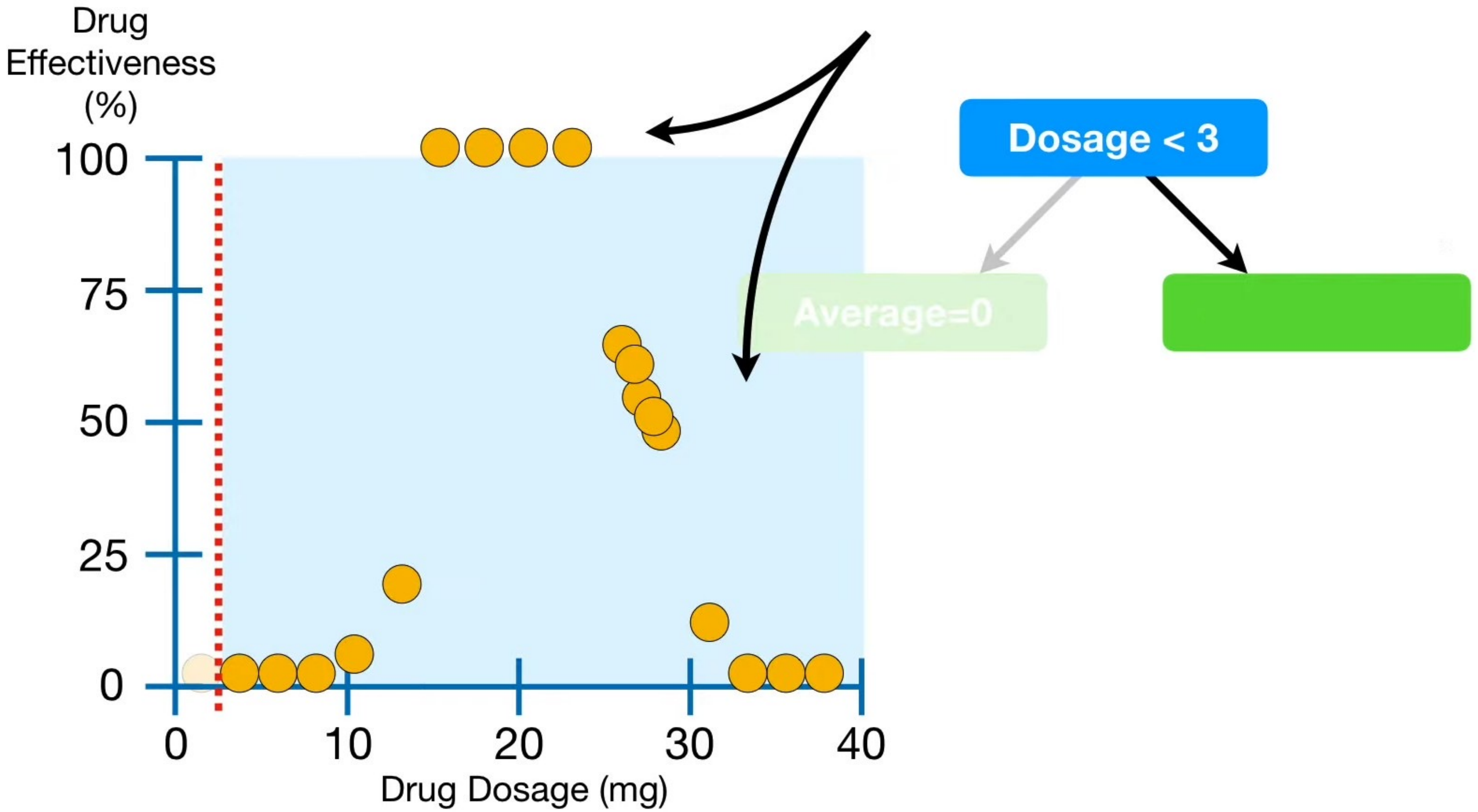
…and the average **Drug Effectiveness** for that one point is **0**…



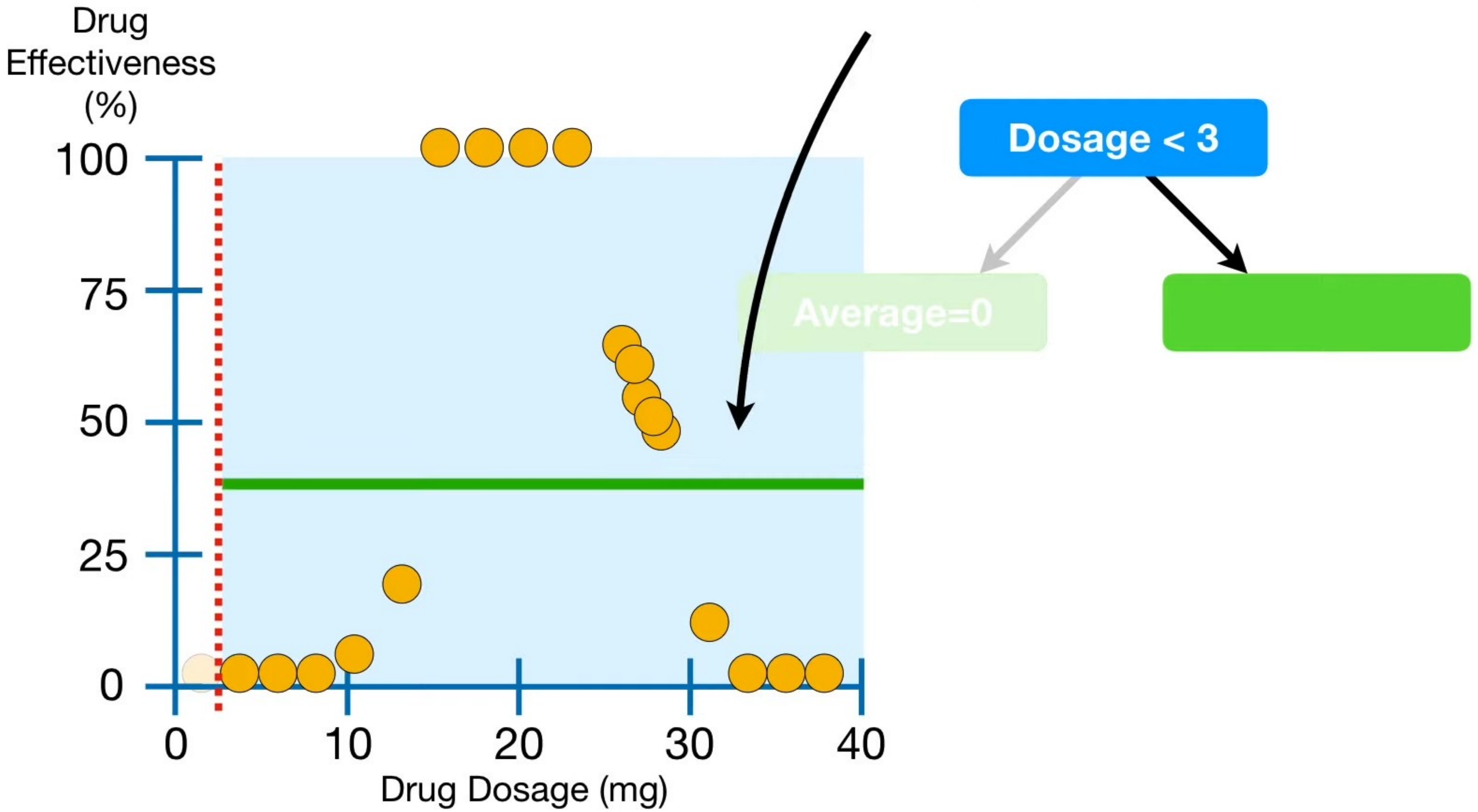
…so we put **0** in the leaf on the left side, for when **Dosage < 3**.



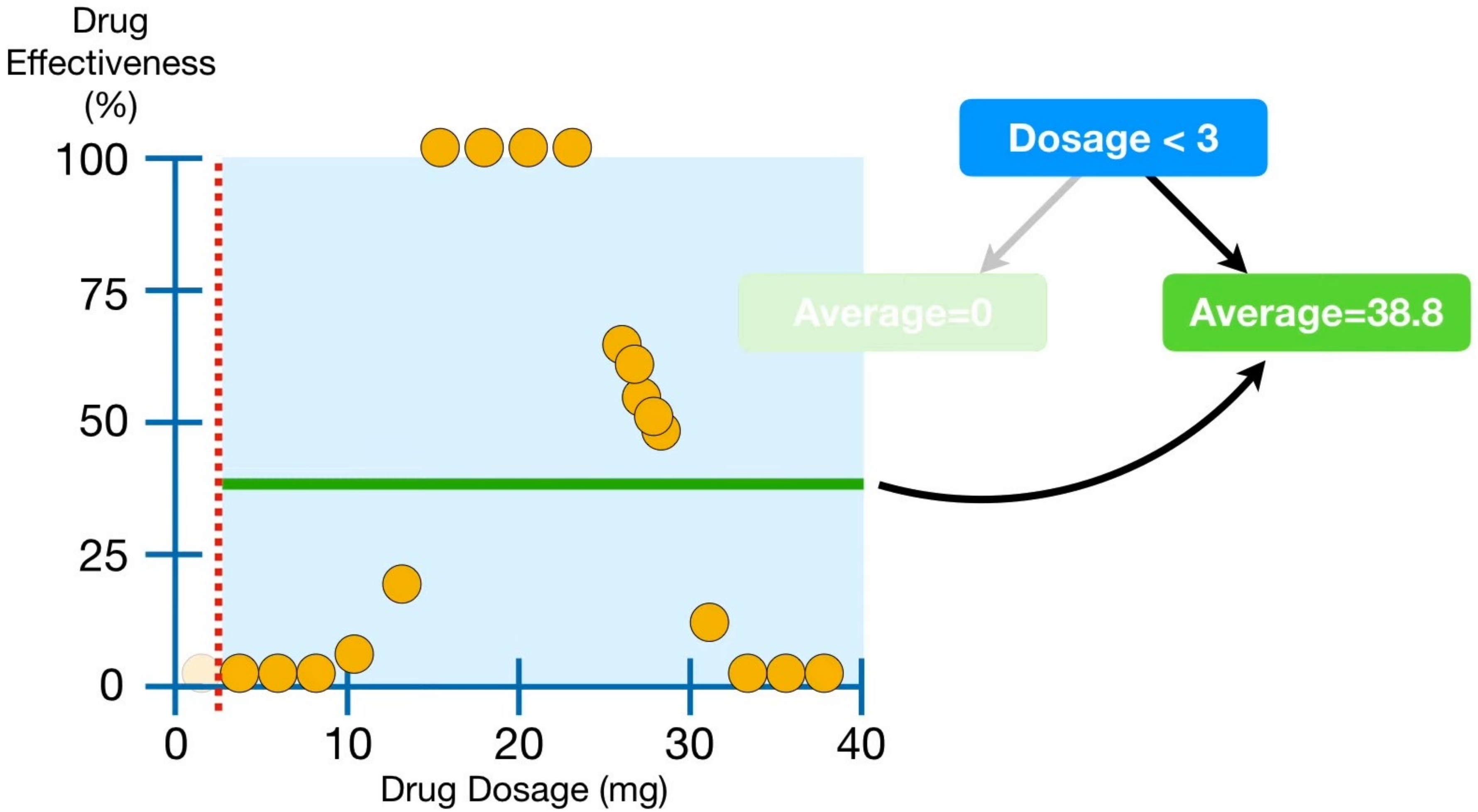
All of the other points have
Dosages >= 3…



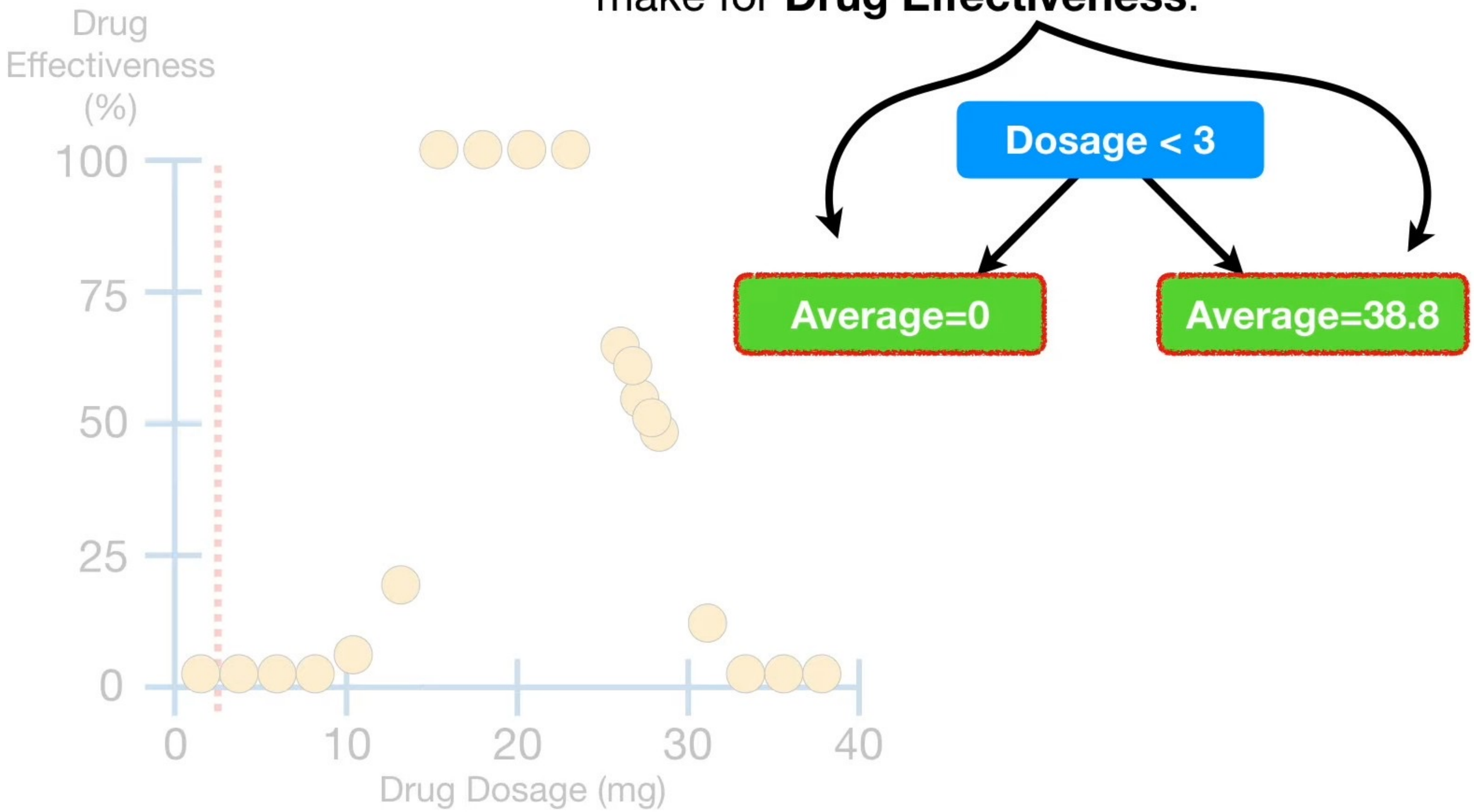
…and the average **Drug Effectiveness** for all of the points with **Dosages >= 3** is **38.8**, (the **green line**)…



…so we put **38.8** in the leaf on the right side, for when the **Dosage >= 3**.



The values in each leaf are the predictions that this simple tree will make for **Drug Effectiveness**.

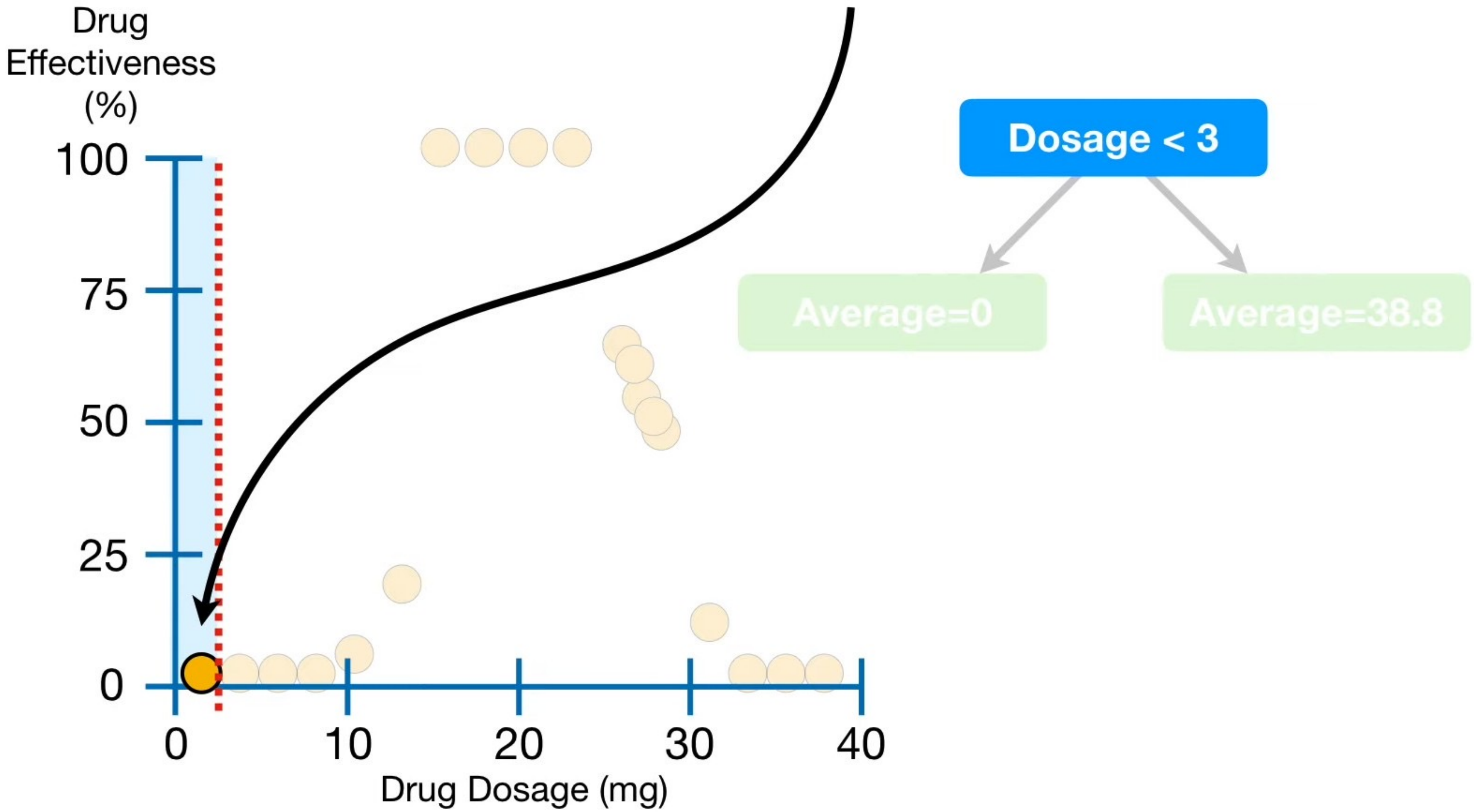


Dosage < 3

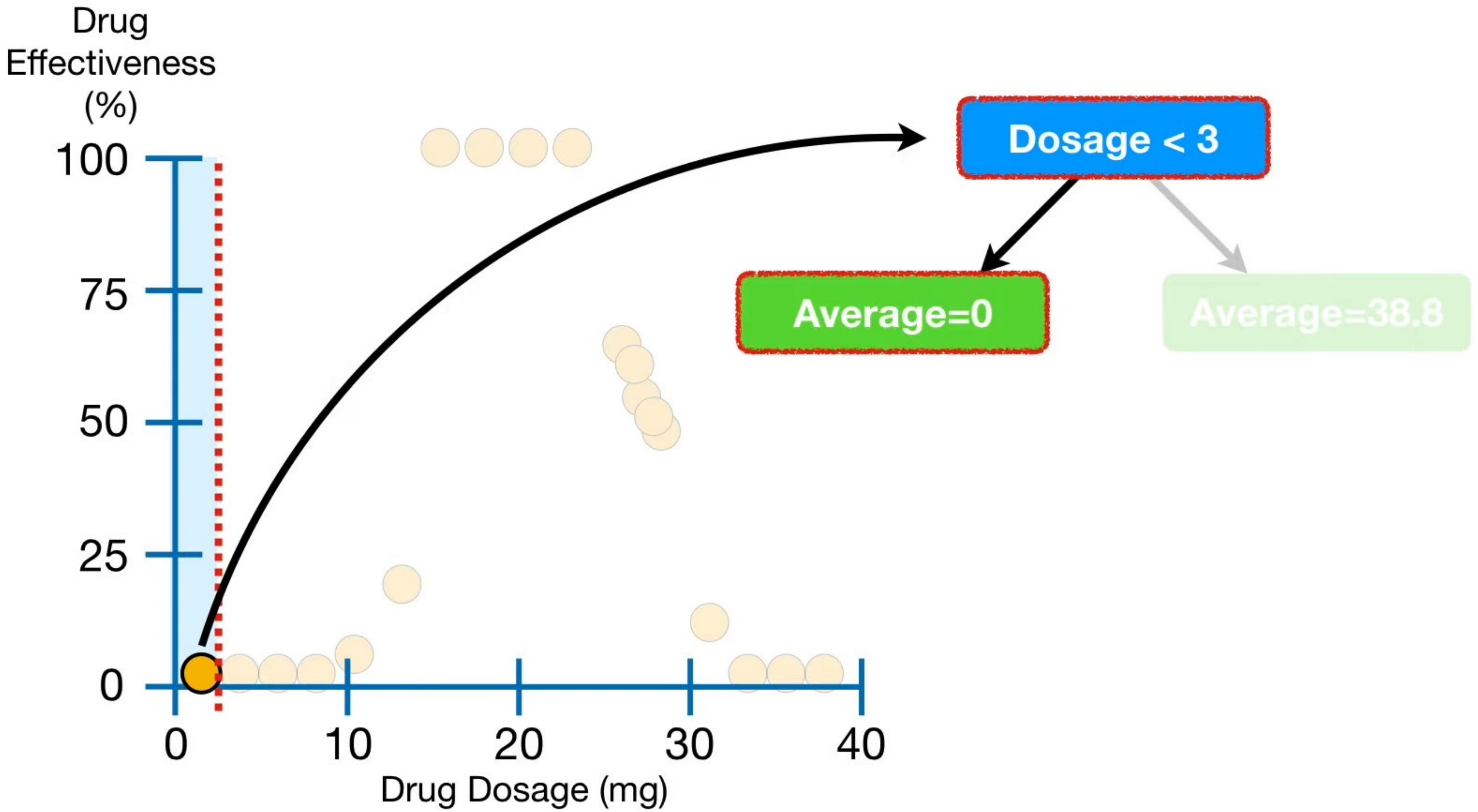
Average=0

Average=38.8

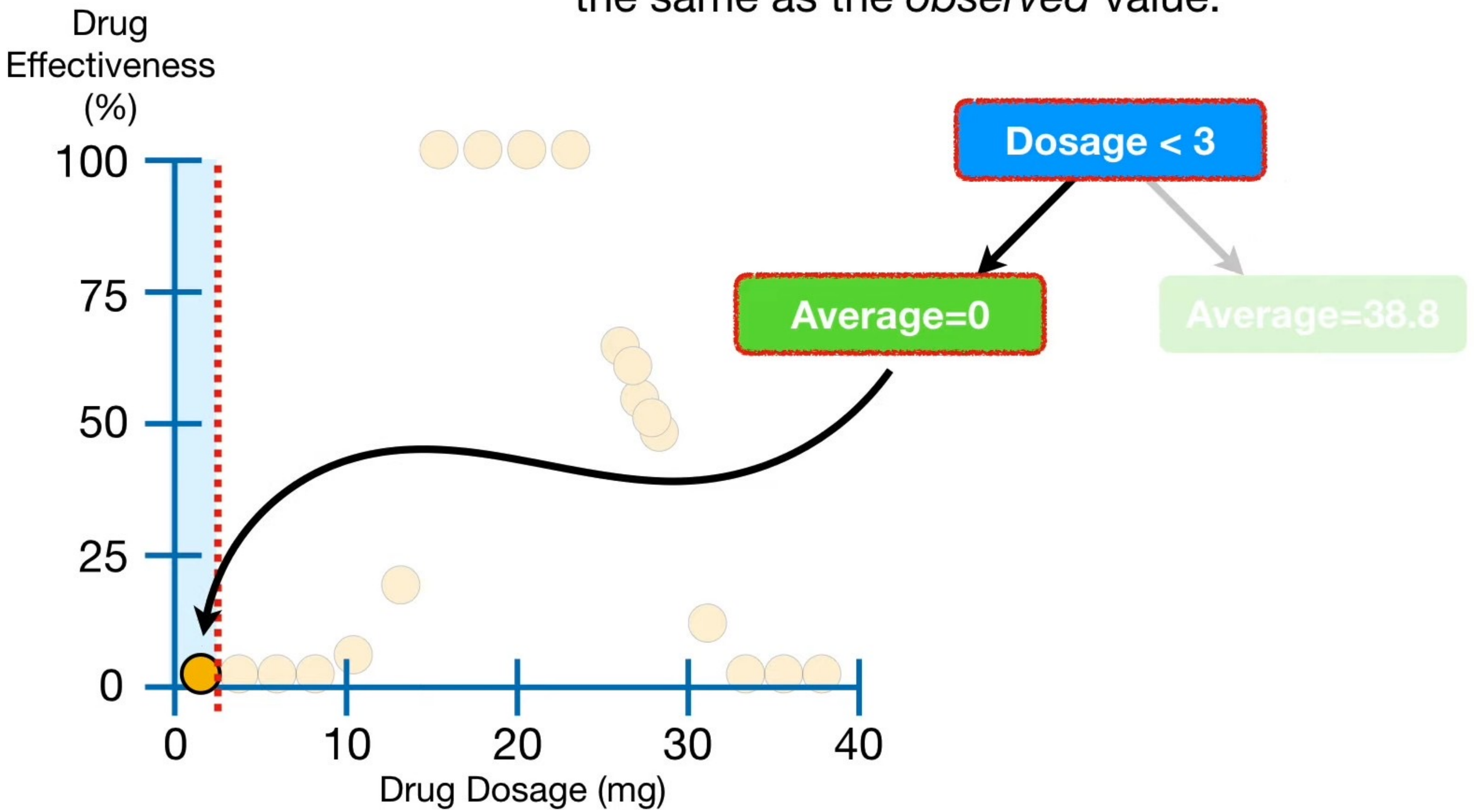
For example, this point, on the far left, has **Dosage < 3**…



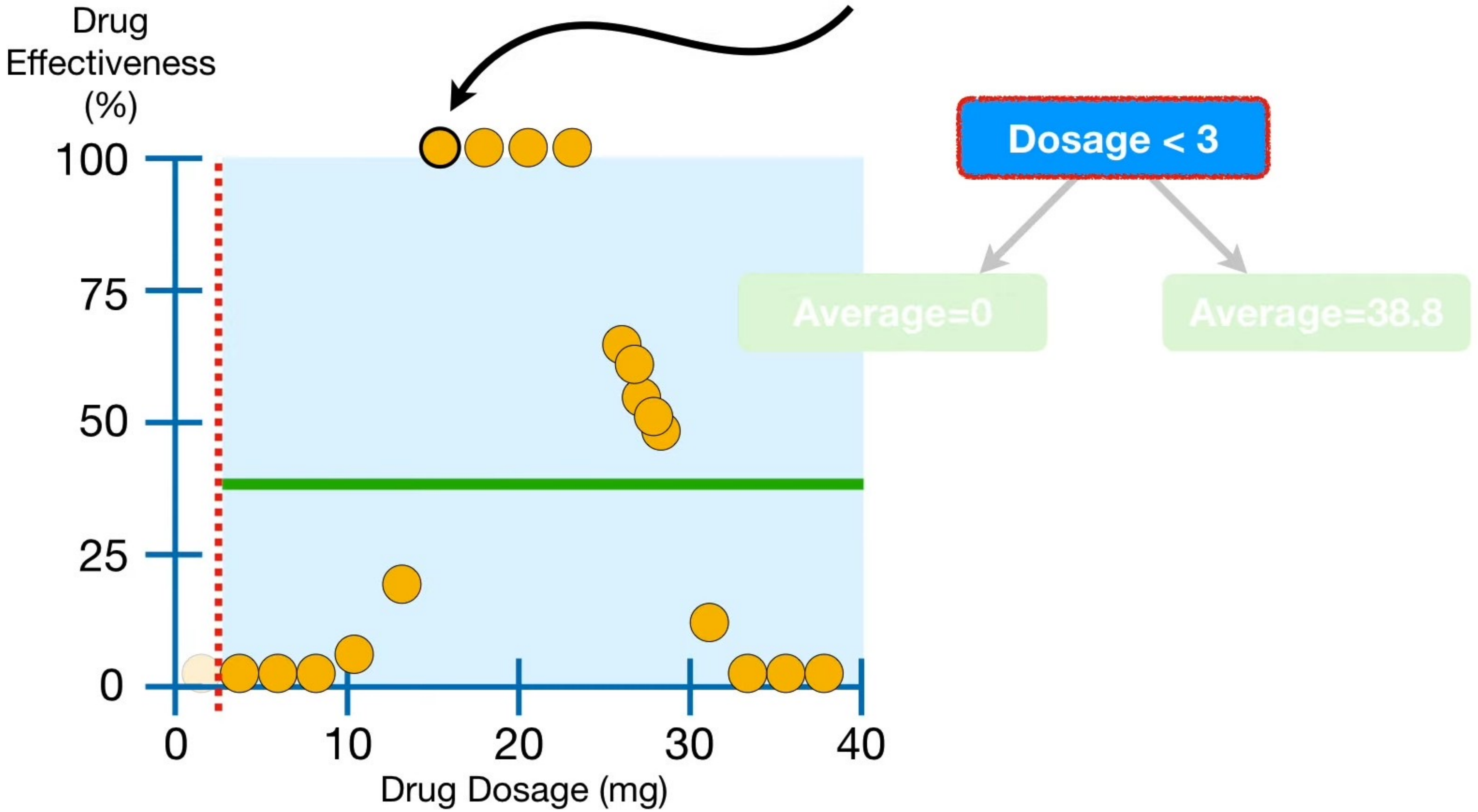
...and the tree predicts that the **Drug Effectiveness** will be **0**.



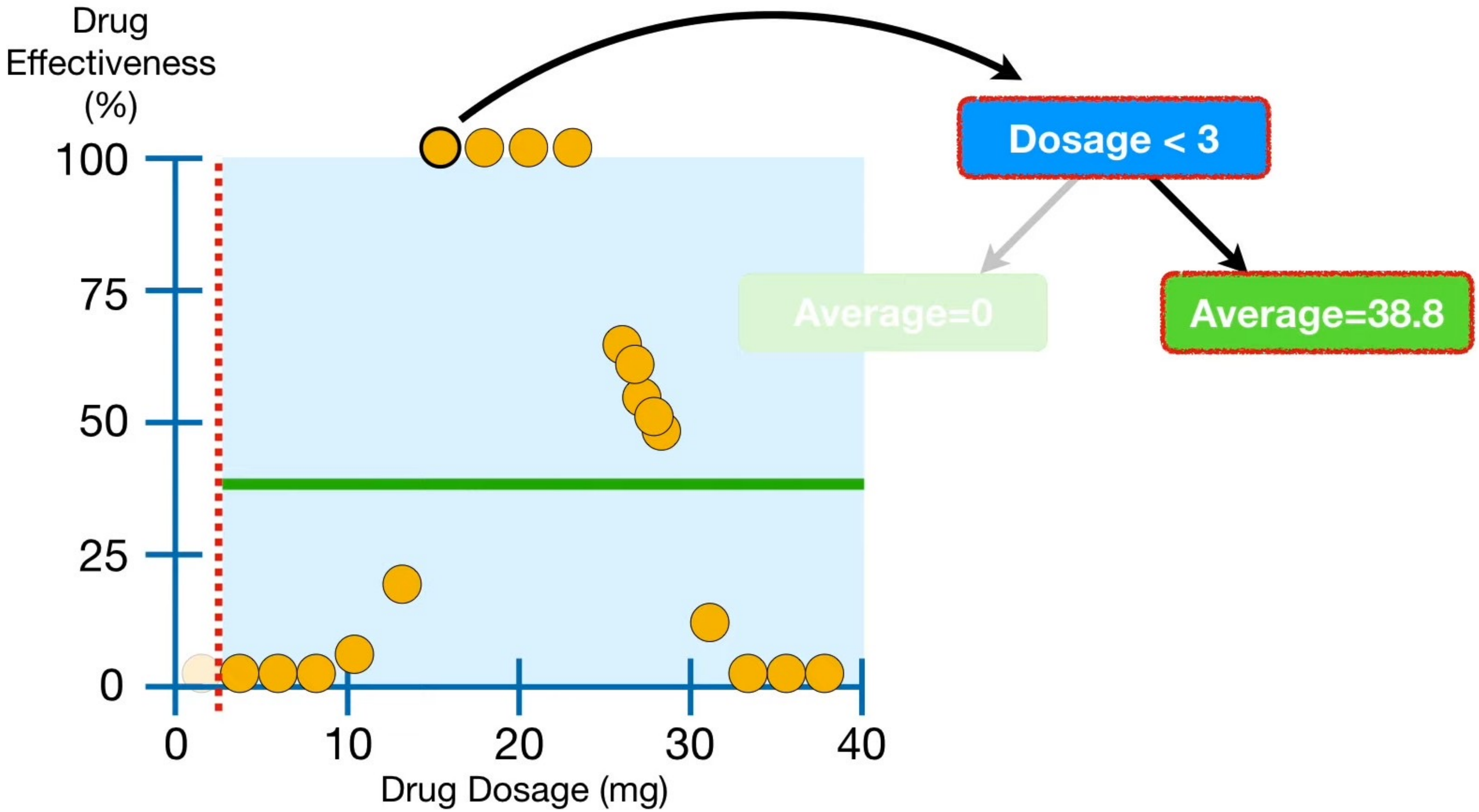
The *prediction* for this point, **Drug Effectiveness = 0**, is pretty good since it is the same as the *observed* value.



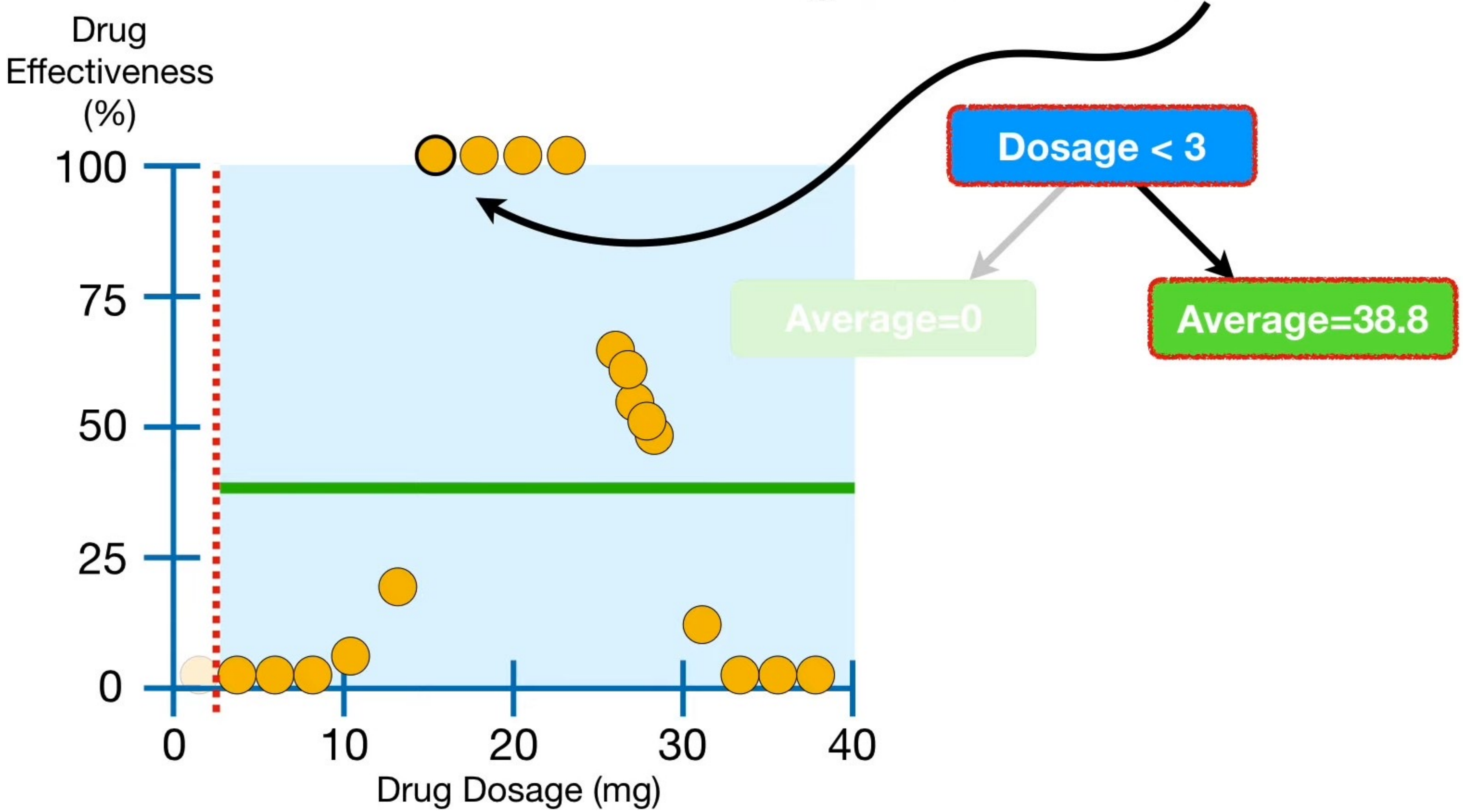
In contrast, for this point, which has **Dosage > 3**…



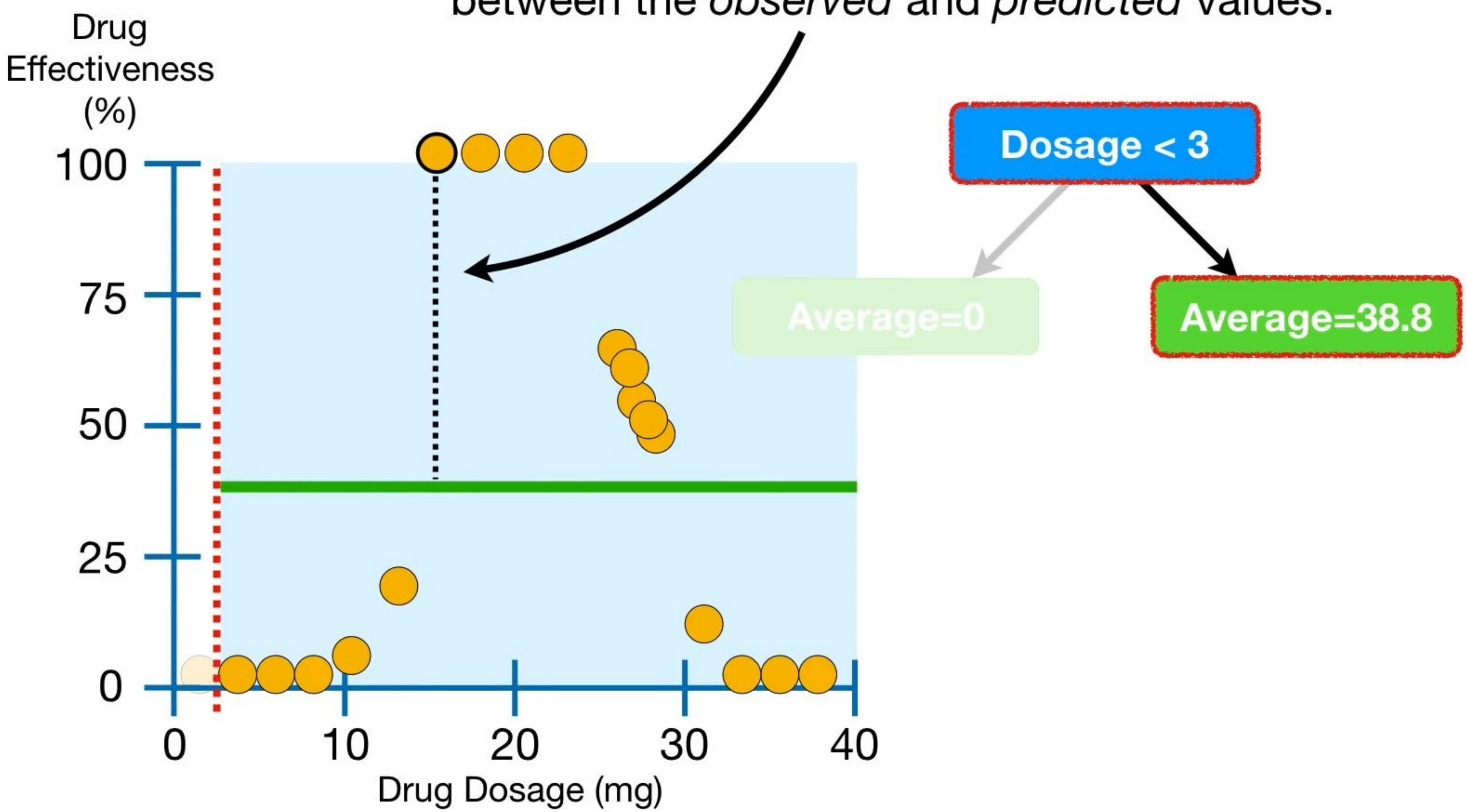
…the tree predicts that the **Drug Effectiveness** will be **38.8**…



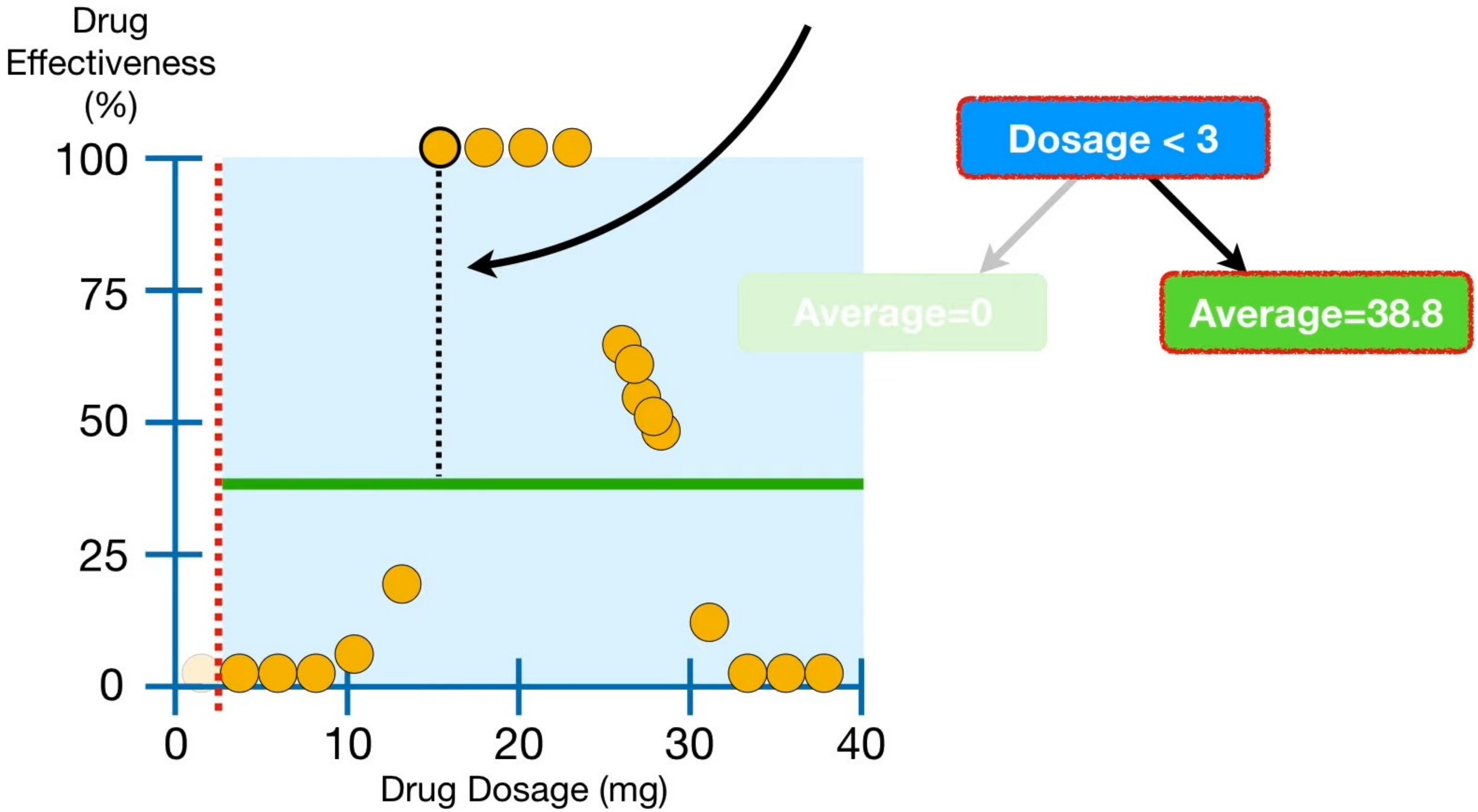
…and that *prediction* is not very good, since the *observed* **Drug Effectiveness** is **100%**.



NOTE: We can visualize how bad the prediction is by drawing a dotted line between the *observed* and *predicted* values.



In other words, this dotted line is a **residual**.

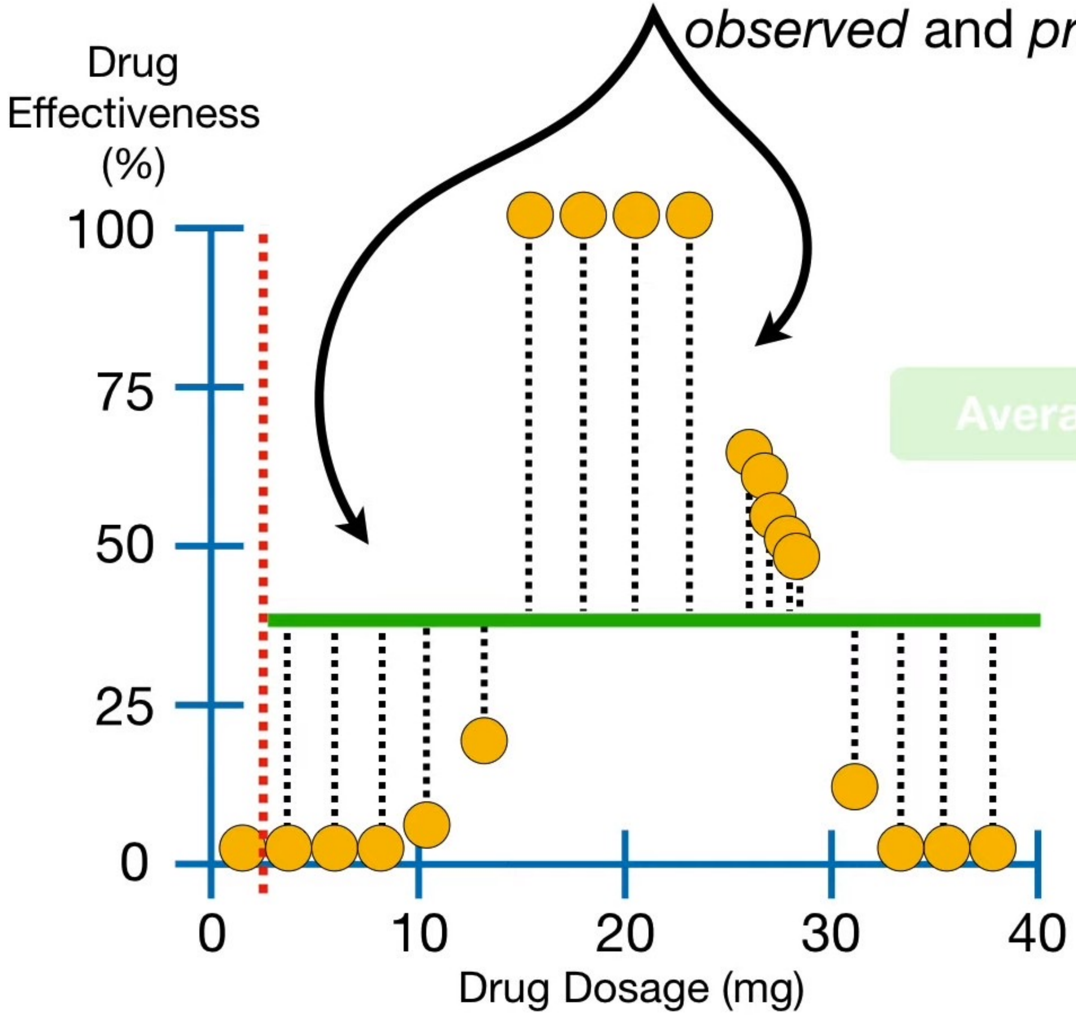


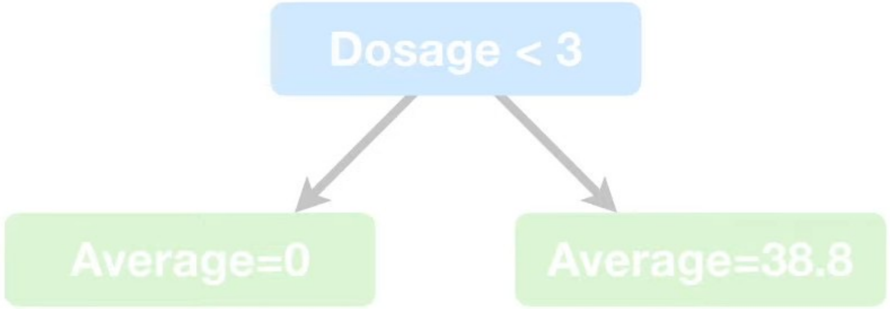
Dosage < 3

Average=0

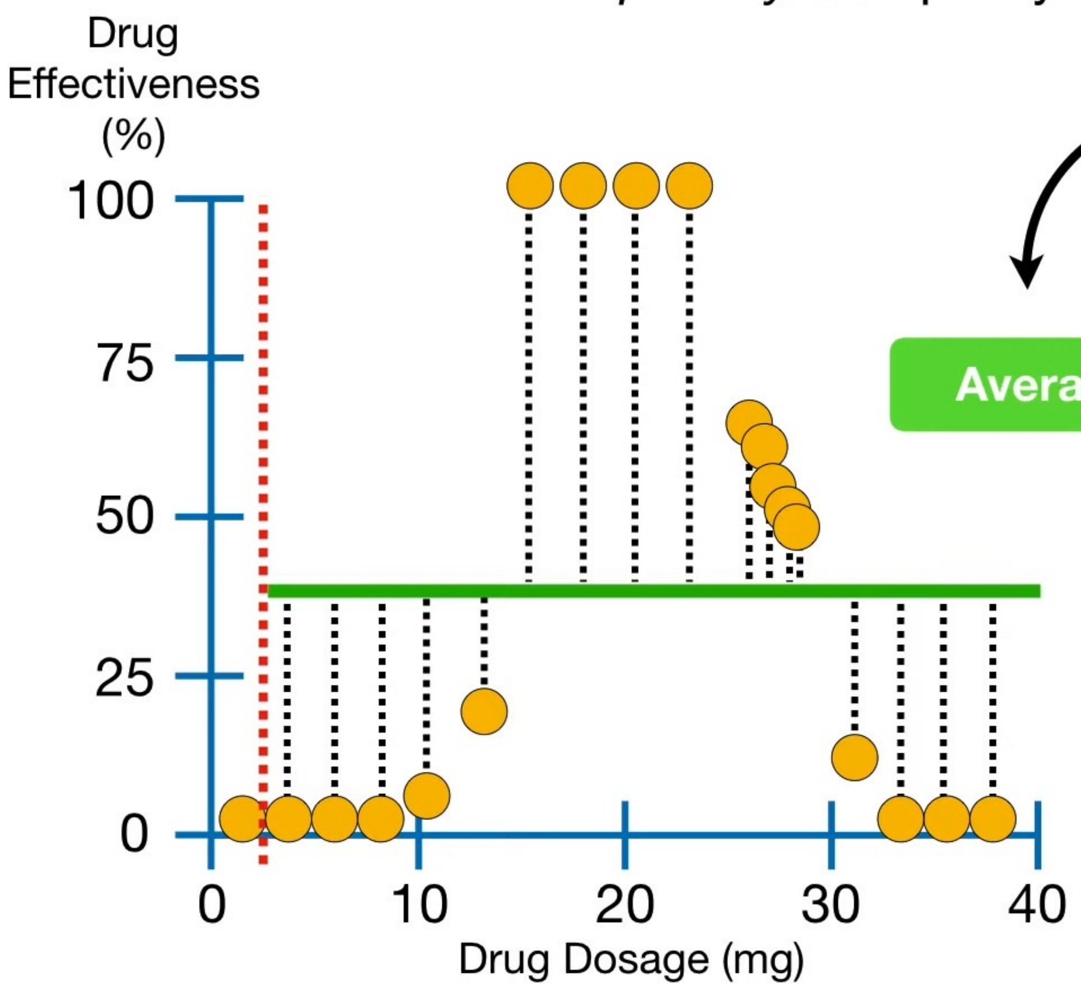
Average=38.8

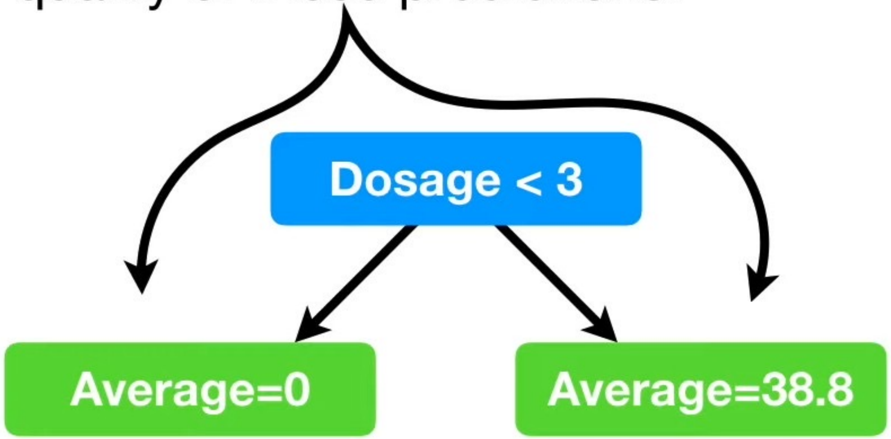
For each point in the data, we can draw its **residual**, the difference between the *observed* and *predicted* values…



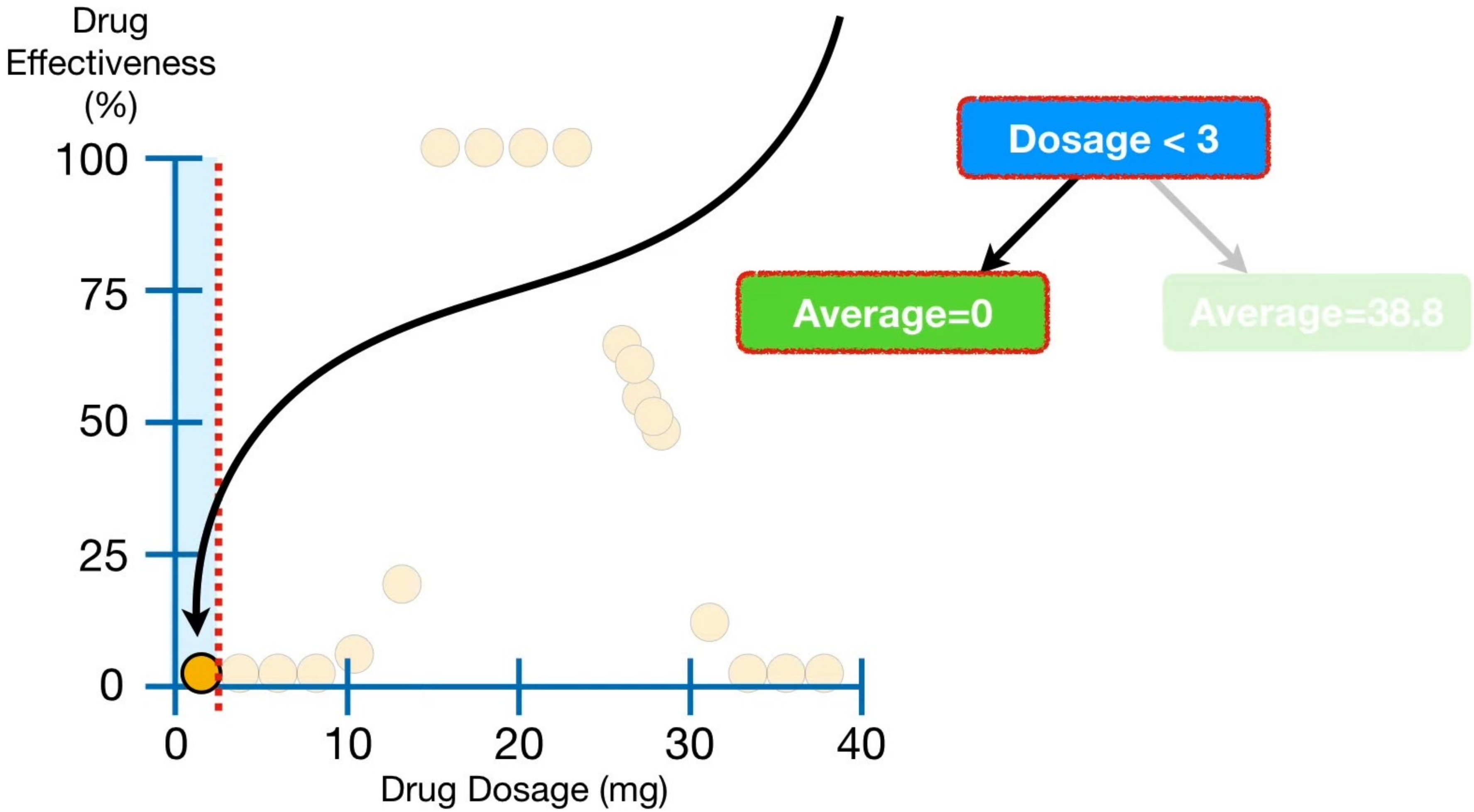


…and we can use the **residuals** to *quantify* the quality of these predictions.

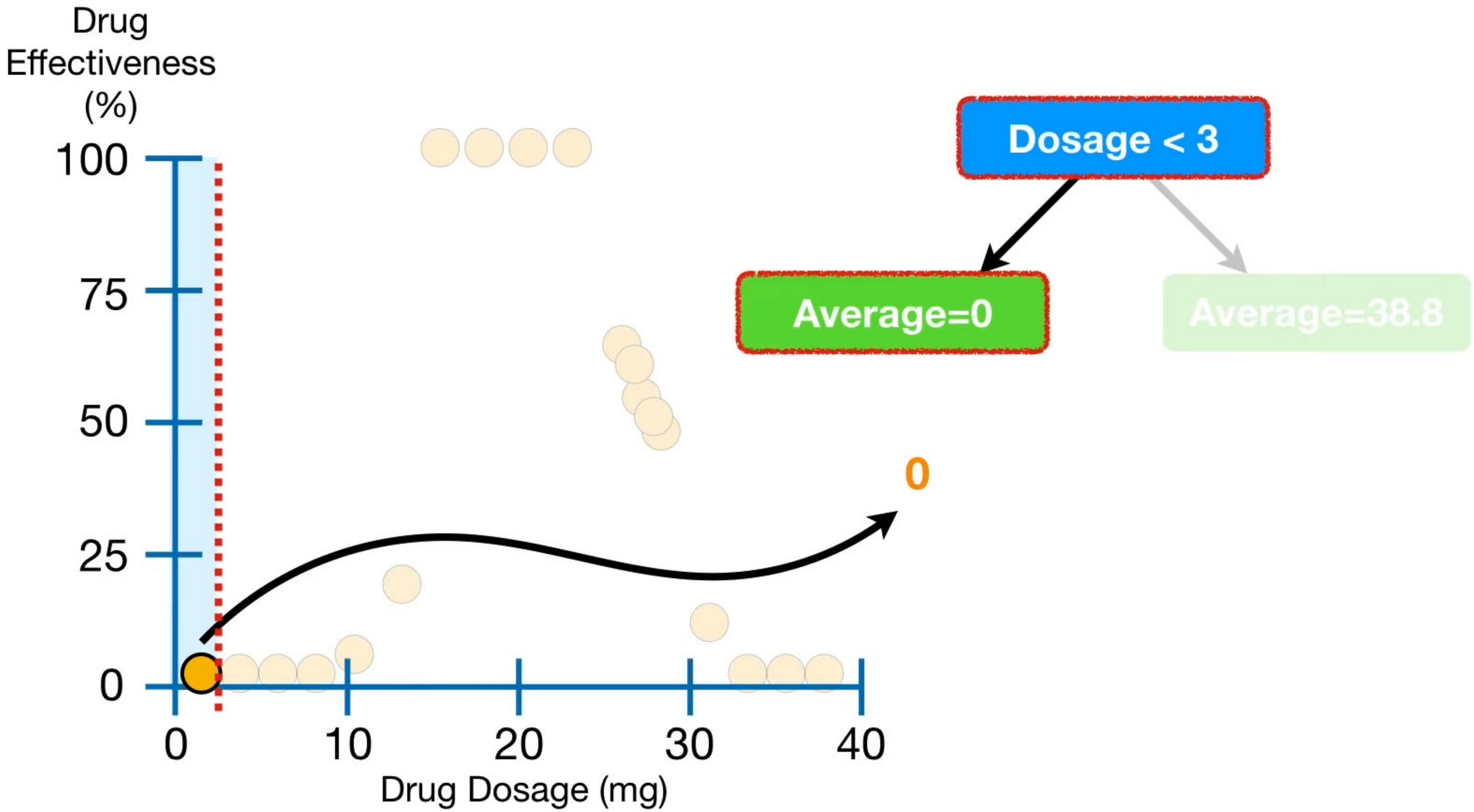




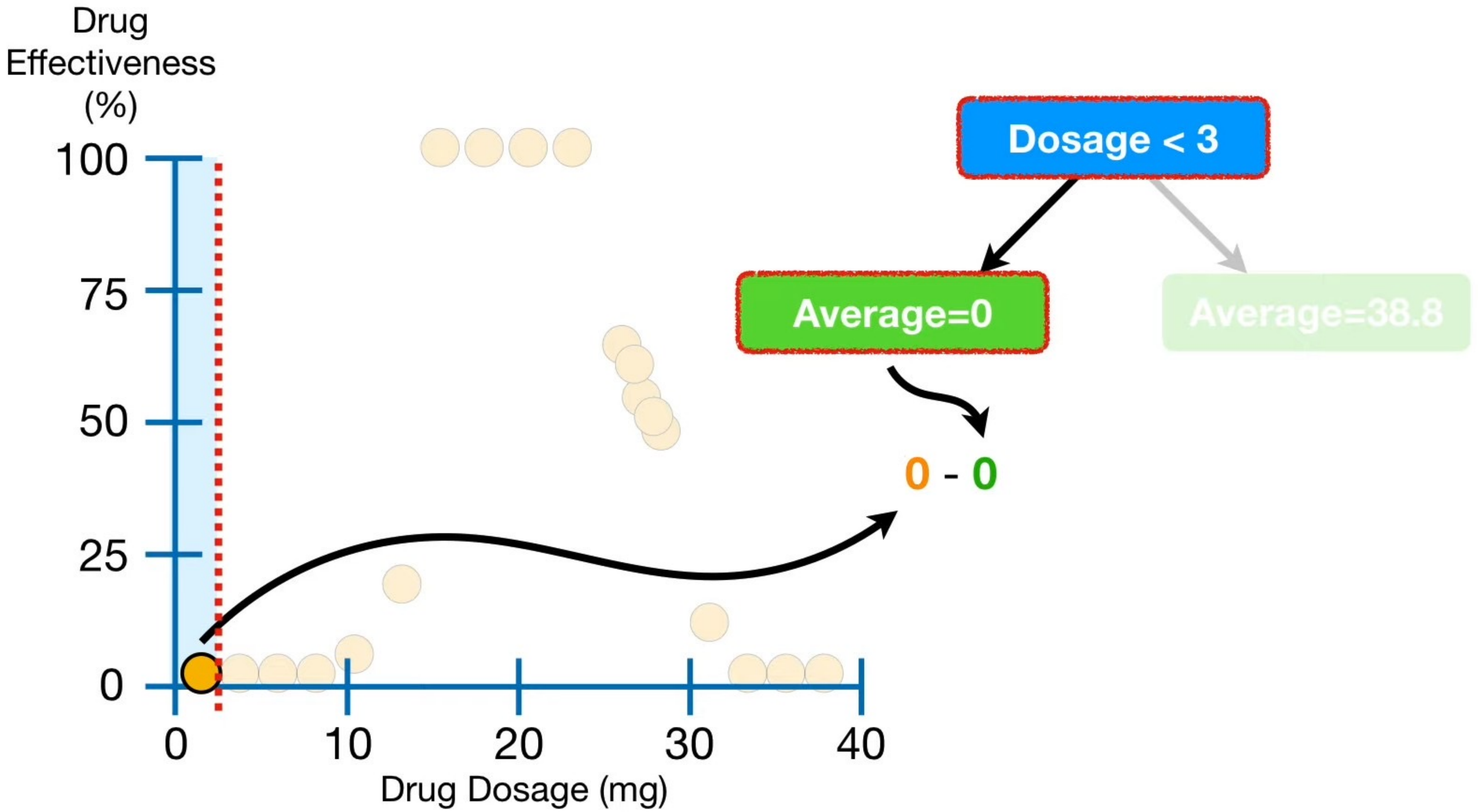
Starting with the only point with **Dosage < 3…**



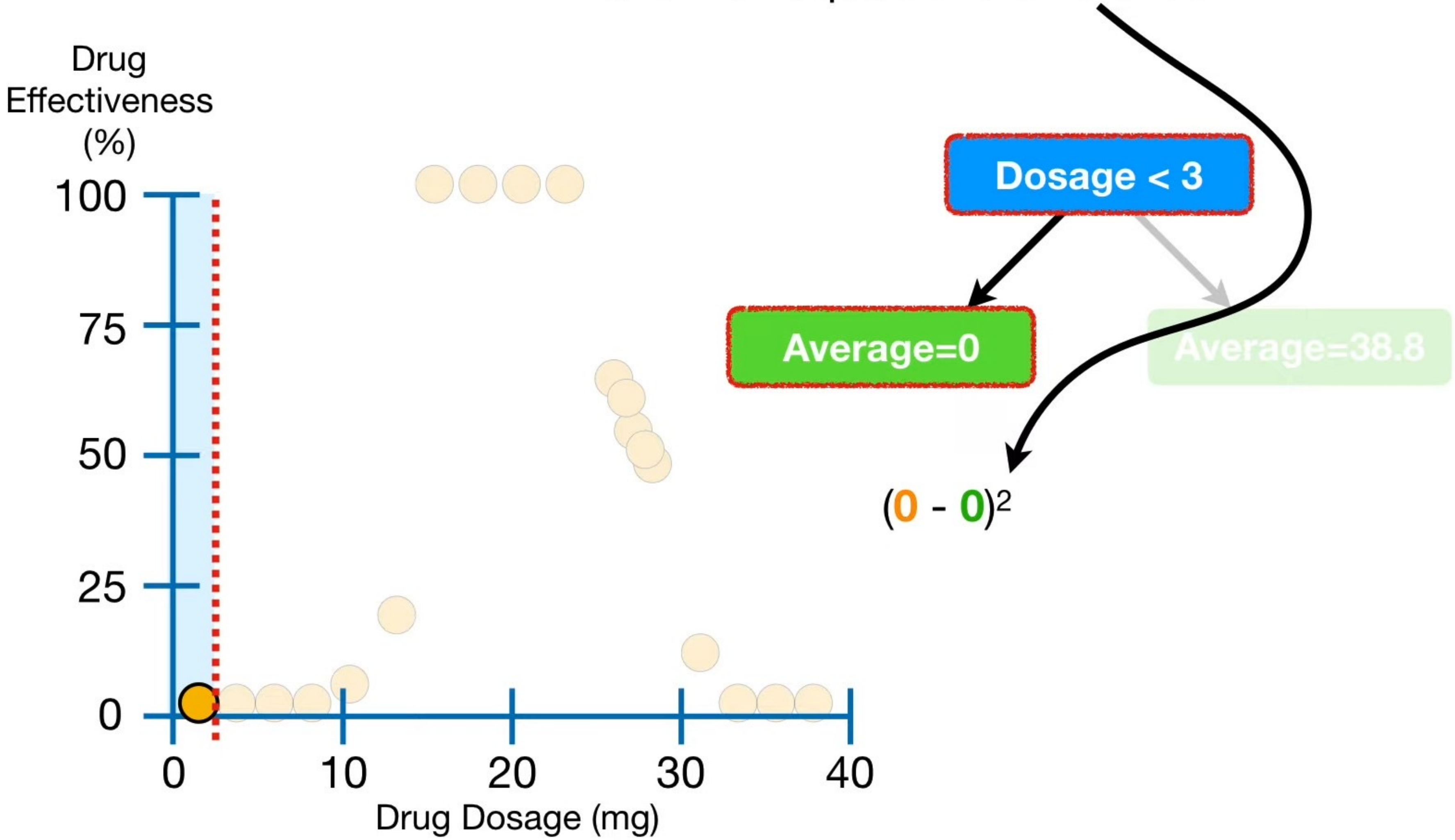
…we calculate the difference between its *observed* **Drug Effectiveness**, **0**,…



…and the *predicted* **Drug Effectiveness**, **0**,…

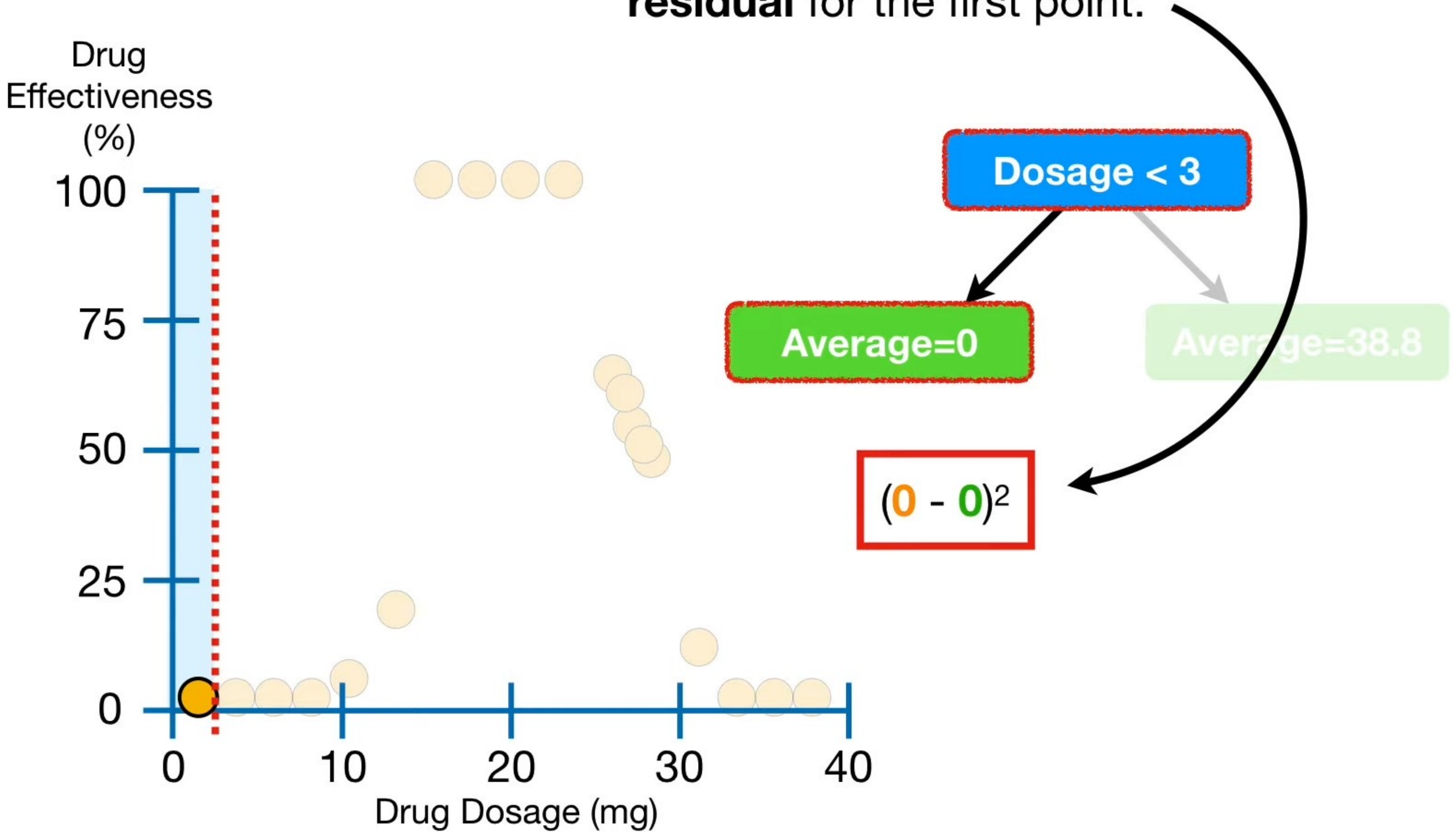


…and then square the difference.

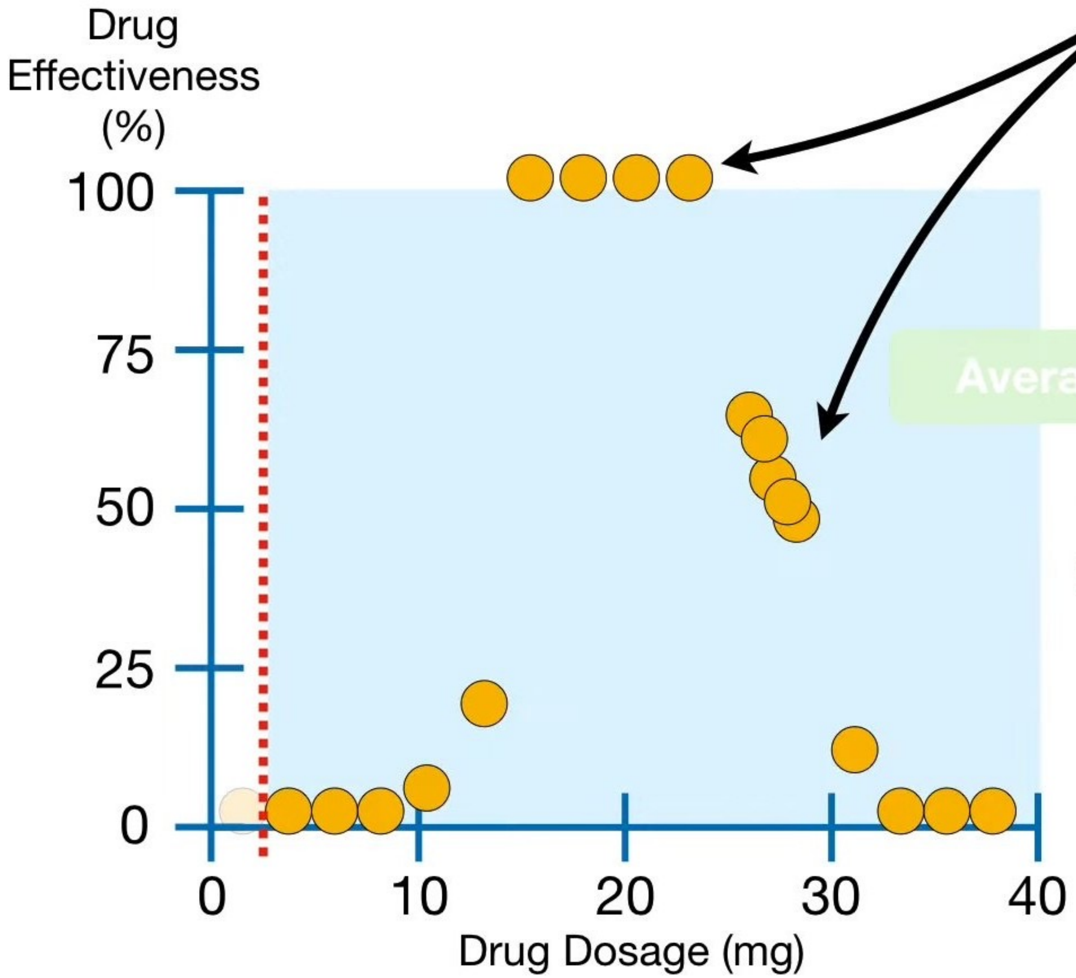


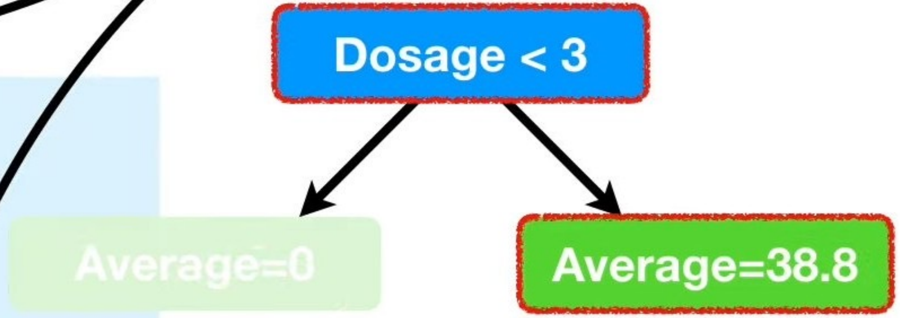
$({\color{orange}0} - {\color{green}0})^2$

In other words, this is the **squared residual** for the first point.



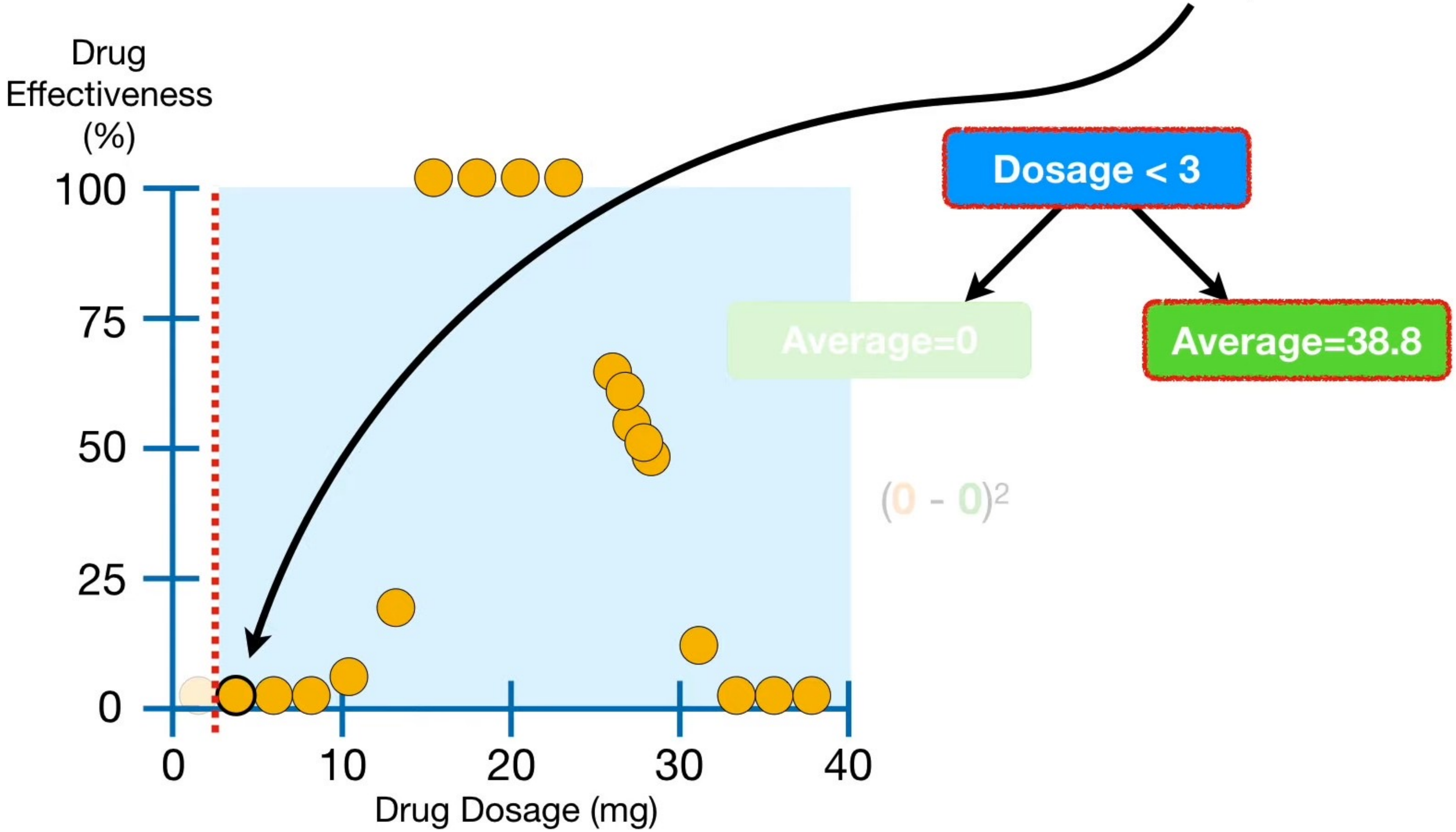
Now we add the squared residuals for the remaining points with **Dosages >= 3**.



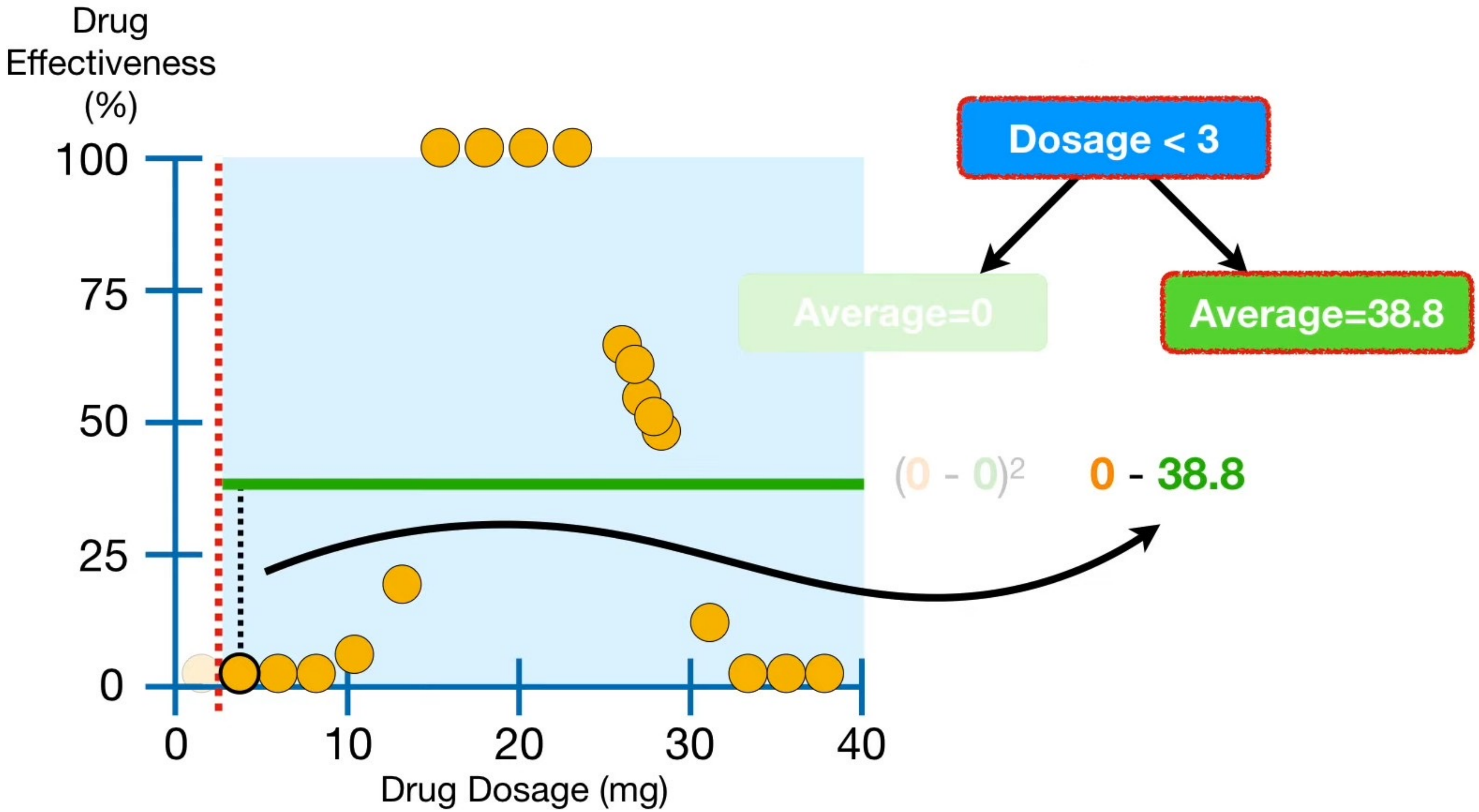


$(0 - 0)^2$

In other words, for this point…

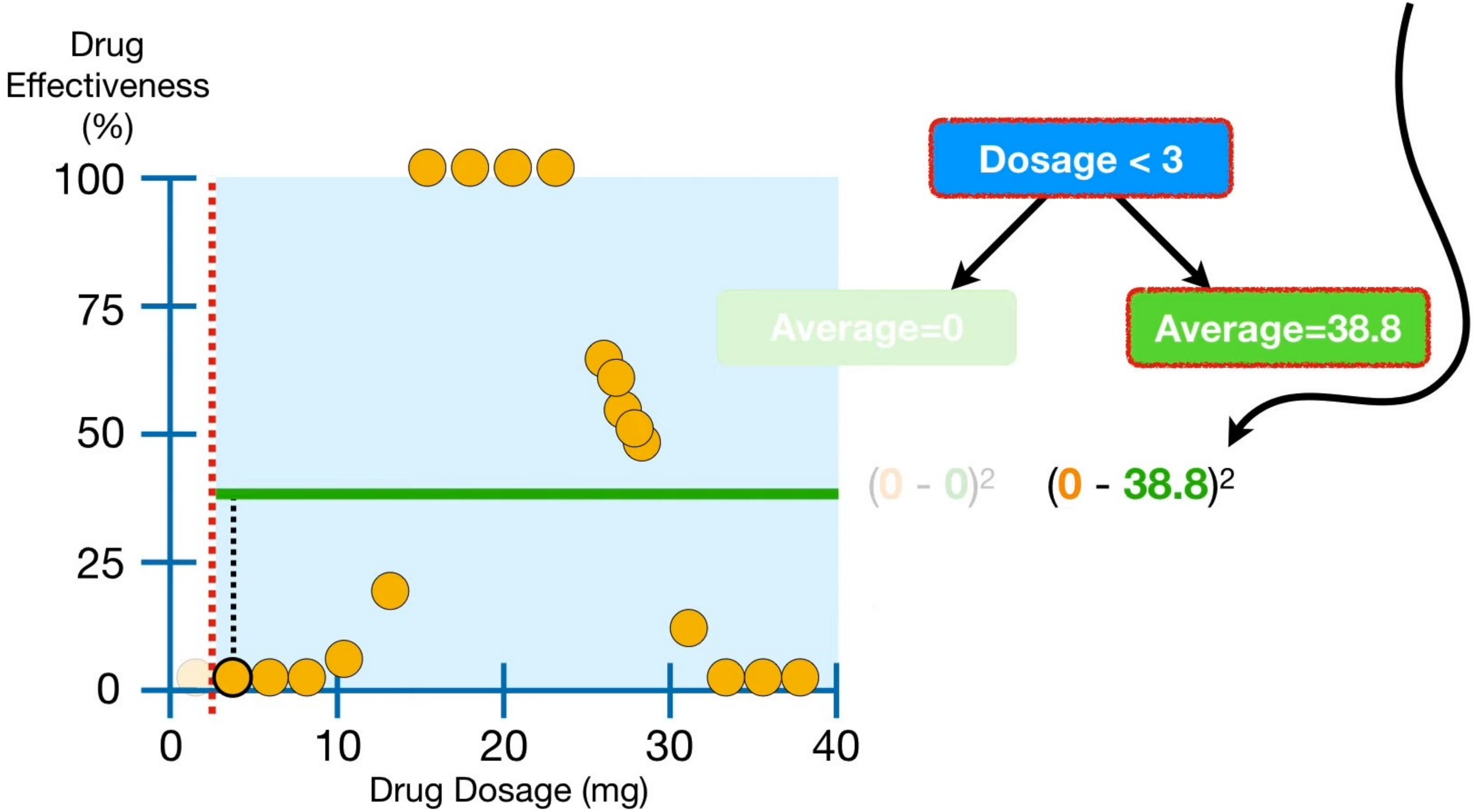


…we calculate the the difference between the *observed* and *predicted* values…



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

…and square it..



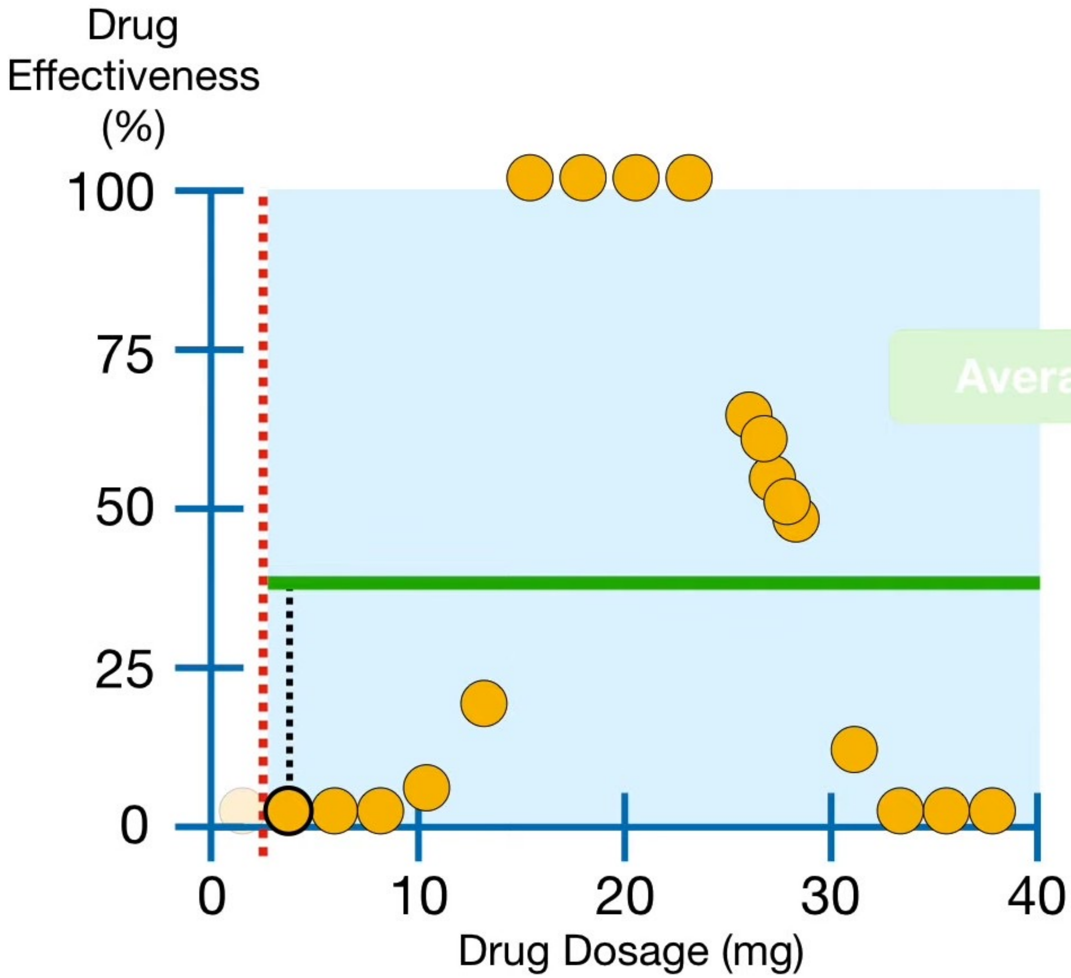
Dosage < 3

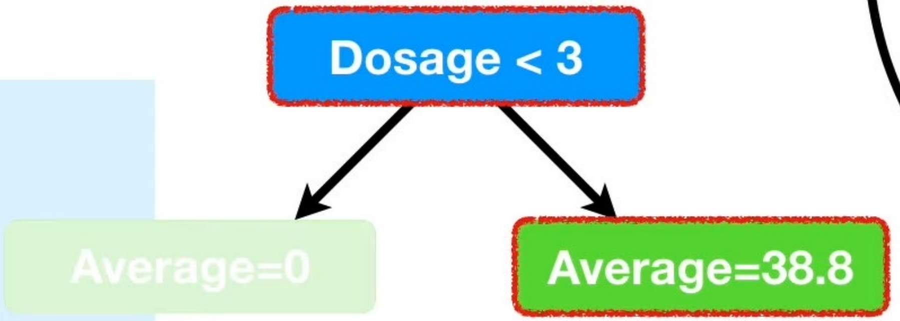
Average=0

Average=38.8

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

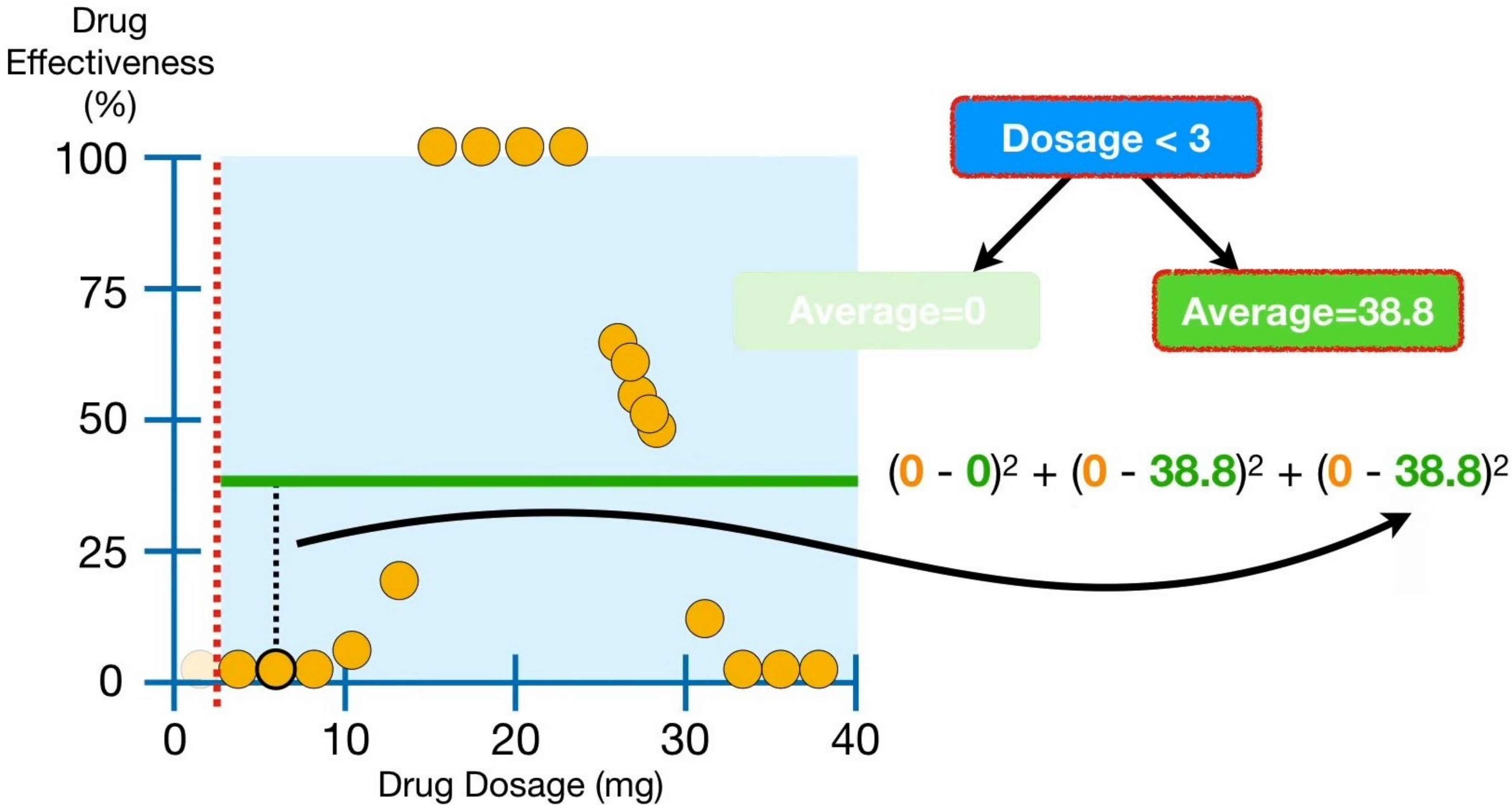
…and then add it to the first term.





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

Then we do the same thing for the next point…



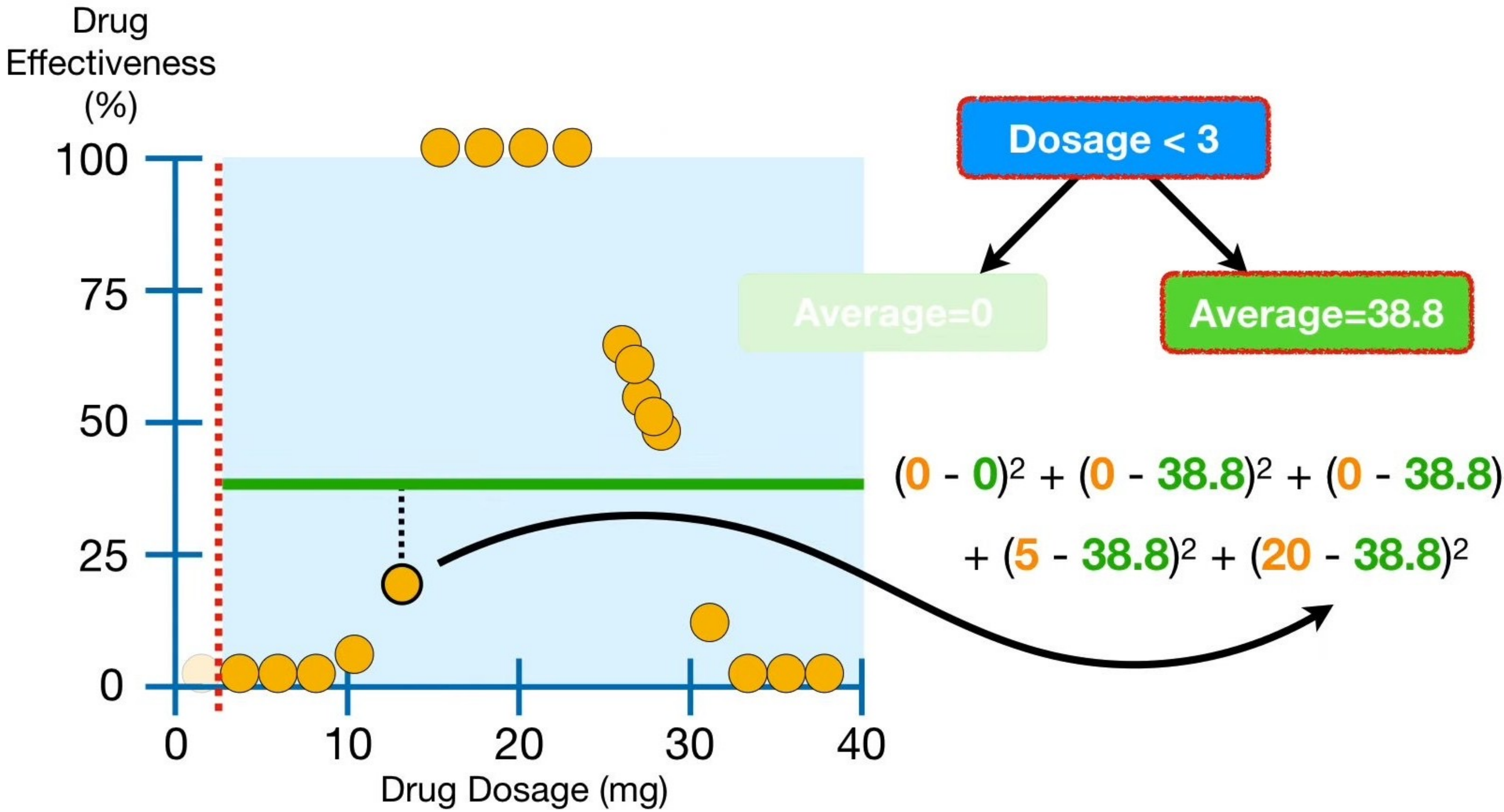
Dosage < 3

Average=0

Average=38.8

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

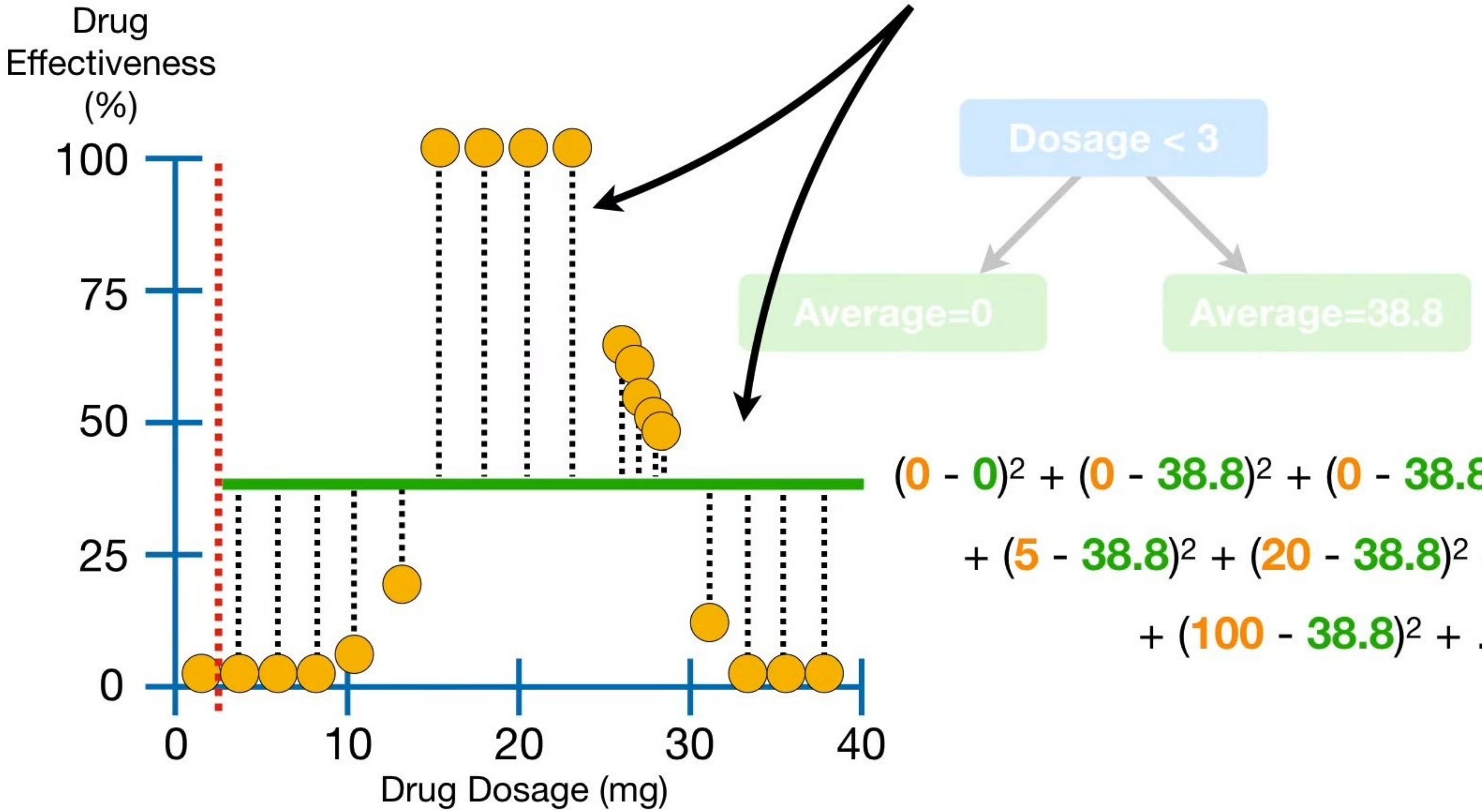
…and the rest of the points…



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

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

…until we have added squared residuals for every point.

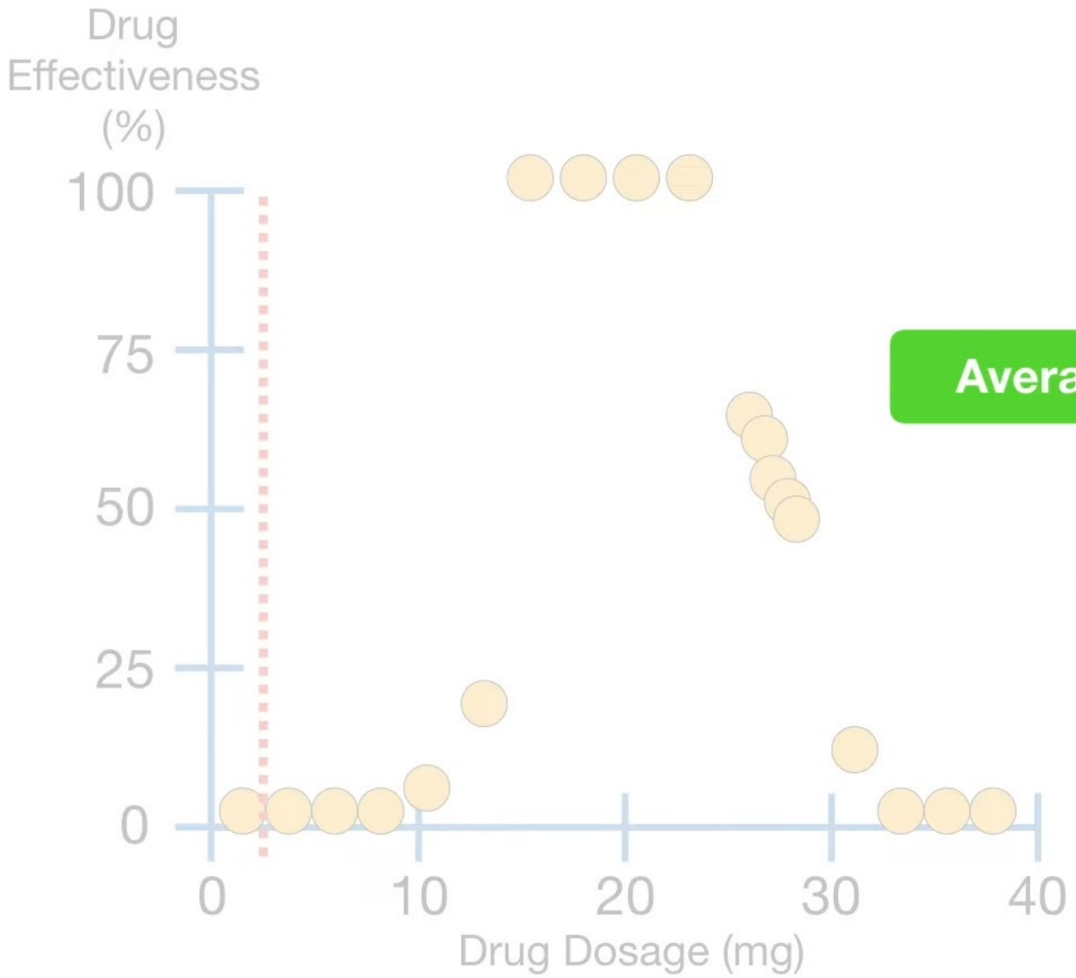


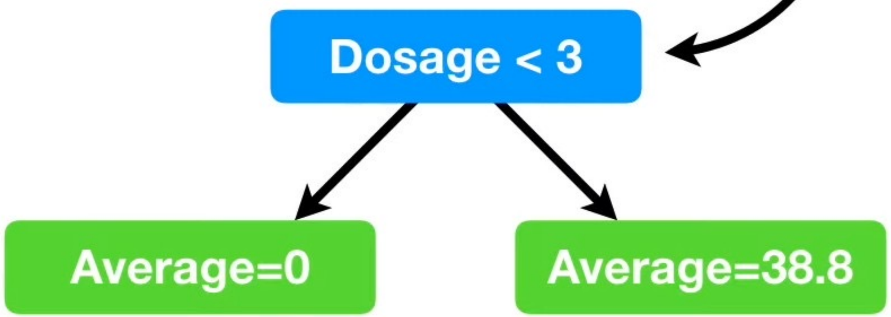
$$(0 - 0)^2 + (0 - 38.8)^2 + (0 - 38.8)^2 + (0 - 38.8)^2$$

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

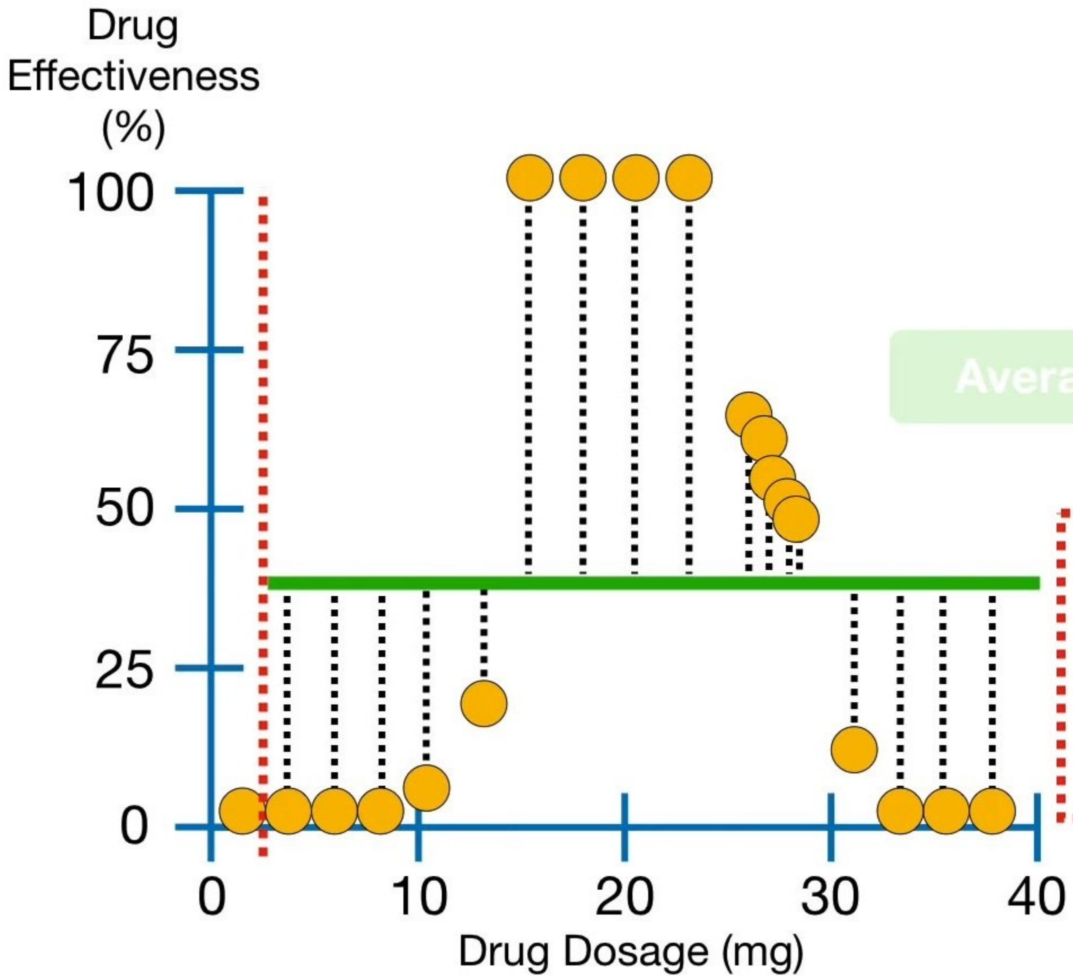
$$+ (100 - 38.8)^2 + \ldots + (0 - 38.8)^2$$

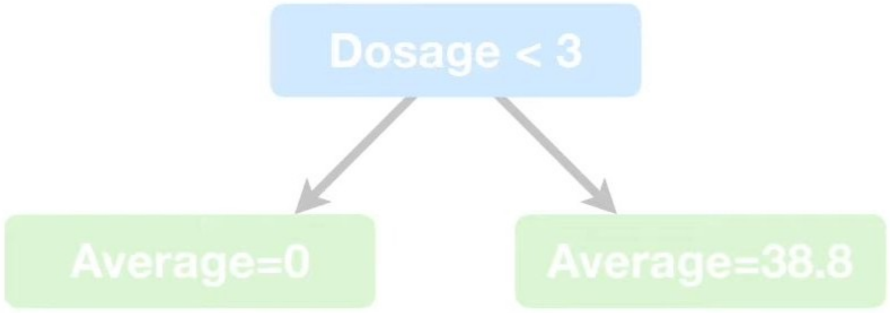
Thus, to evaluate the predictions made when the threshold is **Dosage < 3**…





$$(0 - 0)^2 + (0 - 38.8)^2 + (0 - 38.8)^2 + (0 - 38.8)^2$$
$$+ (5 - 38.8)^2 + (20 - 38.8)^2 + (100 - 38.8)^2$$
$$+ (100 - 38.8)^2 + \ldots + (0 - 38.8)^2$$

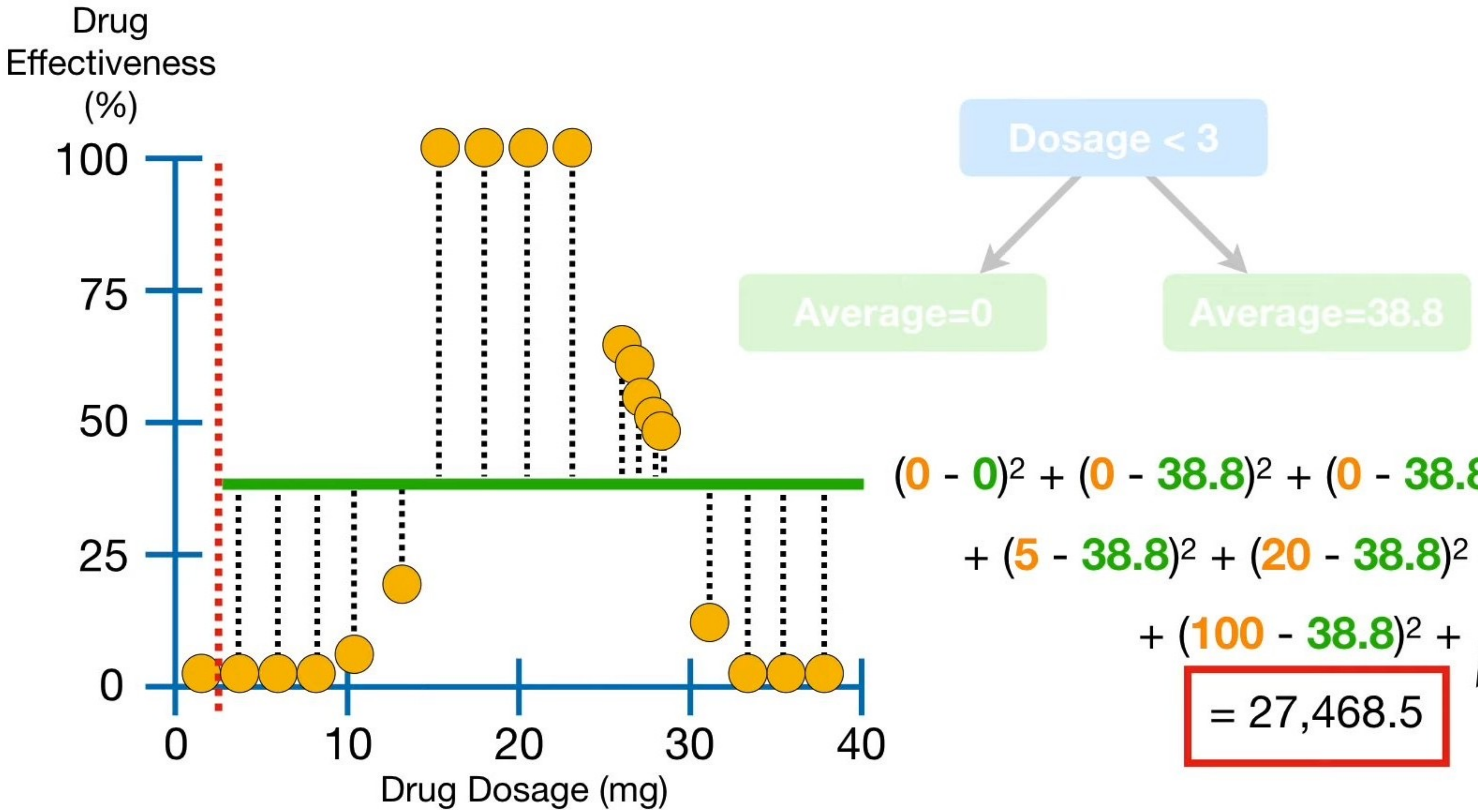




…we add up the squared residuals for every point…

$$\begin{aligned}(0 - 0)^2 + (0 - 38.8)^2 + (0 - 38.8)^2 + (0 - 38.8)^2 \\ + (5 - 38.8)^2 + (20 - 38.8)^2 + (100 - 38.8)^2 \\ + (100 - 38.8)^2 + \ldots + (0 - 38.8)^2\end{aligned}$$

…and get **27,468.5**.



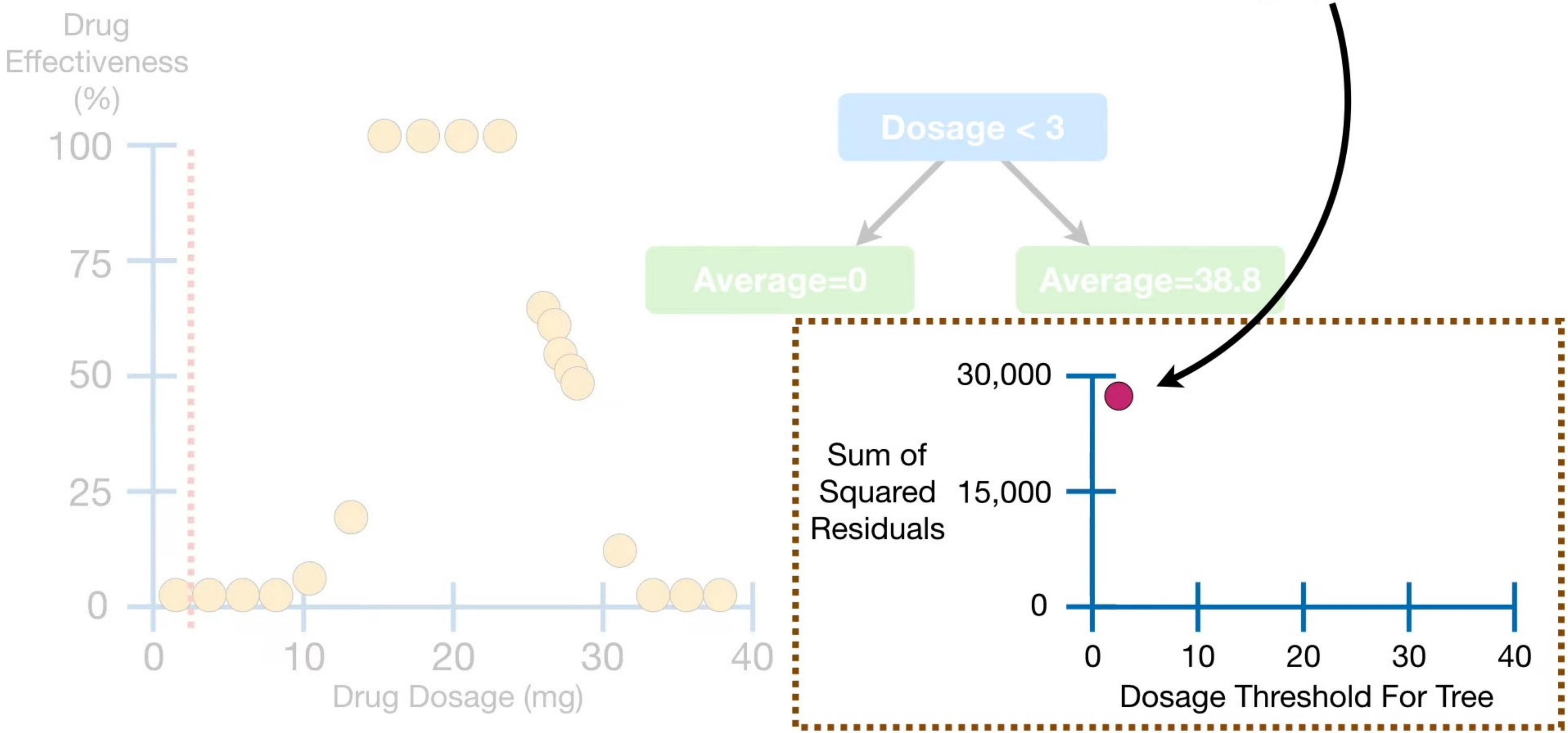
$$(0 - 0)^2 + (0 - 38.8)^2 + (0 - 38.8)^2 + (0 - 38.8)^2$$

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

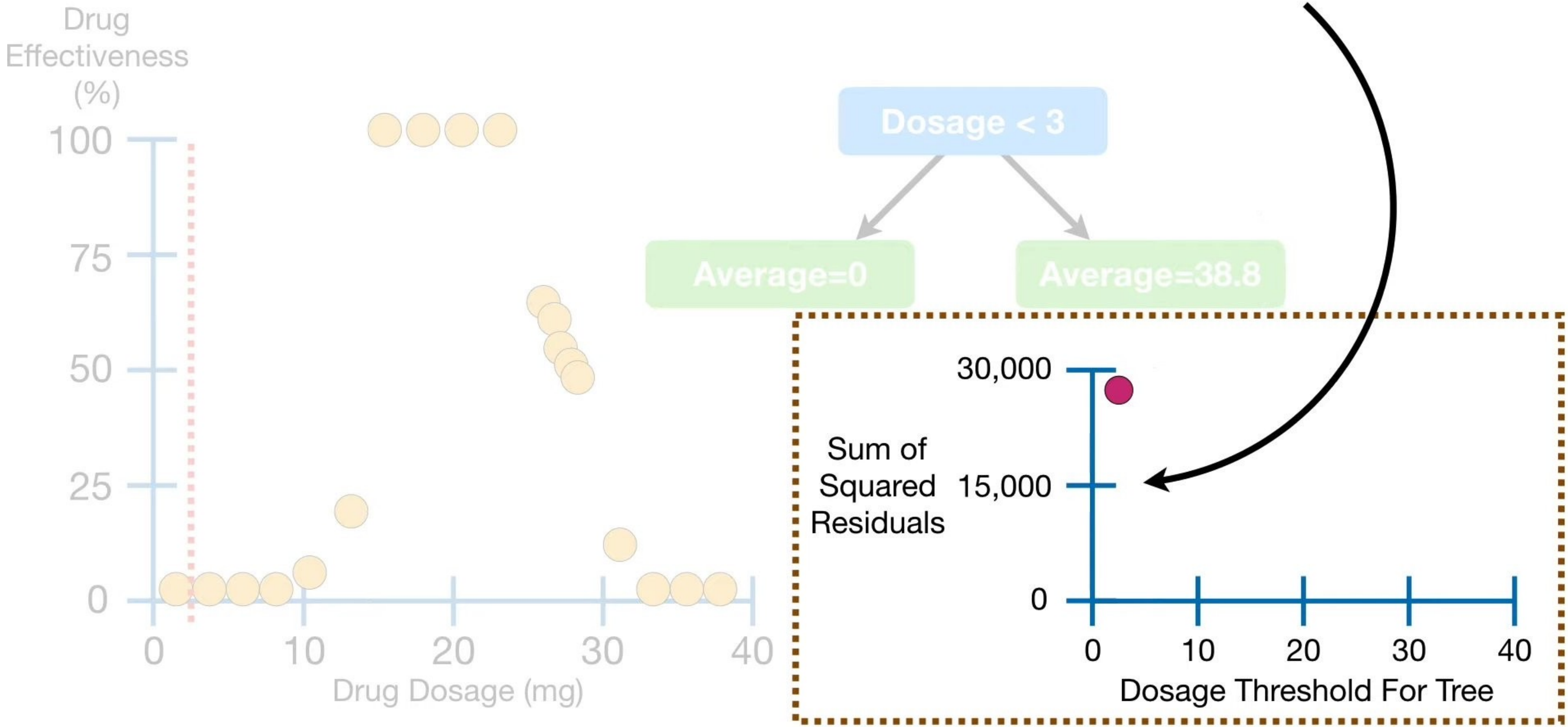
$$+ (100 - 38.8)^2 + \ldots + (0 - 38.8)^2$$

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

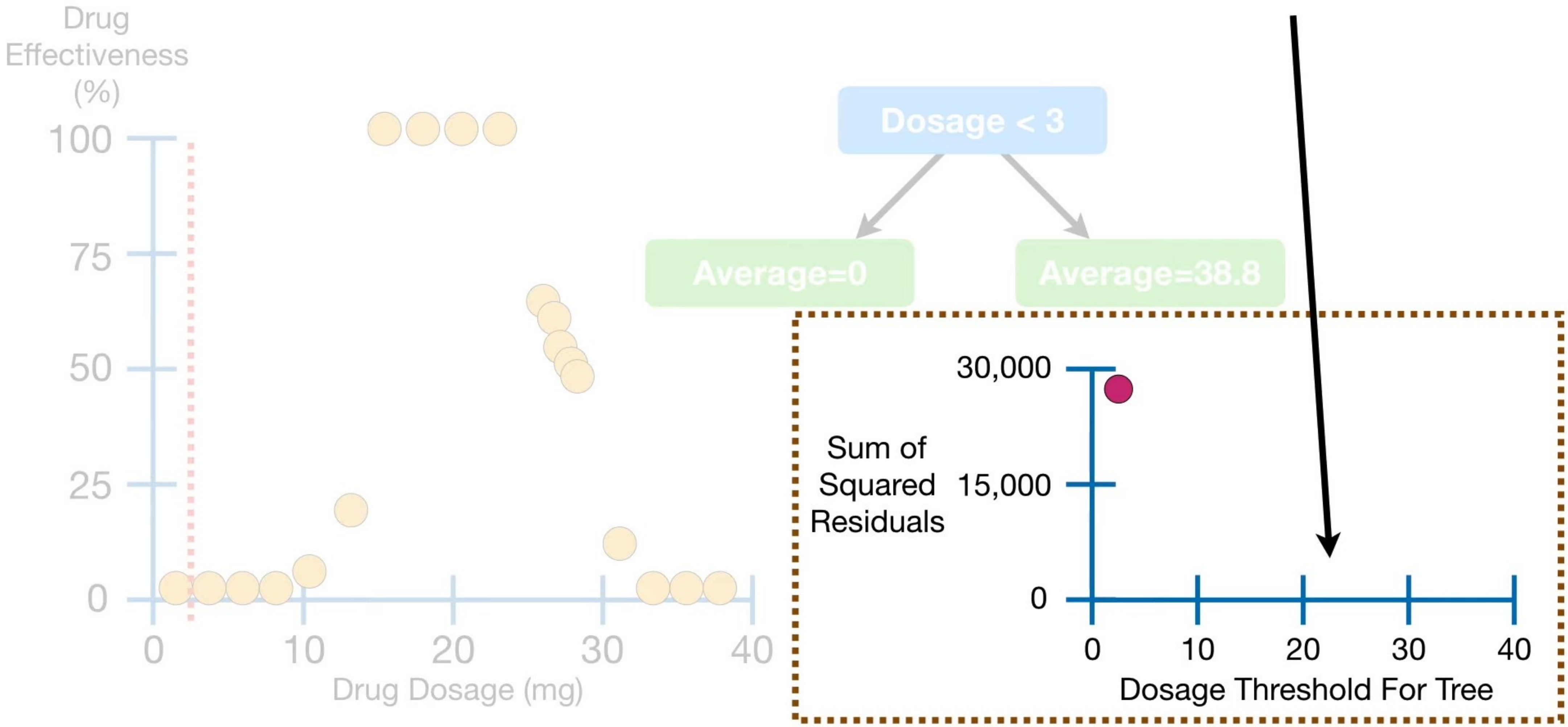
NOTE: We can plot the sum of squared residuals on this graph.



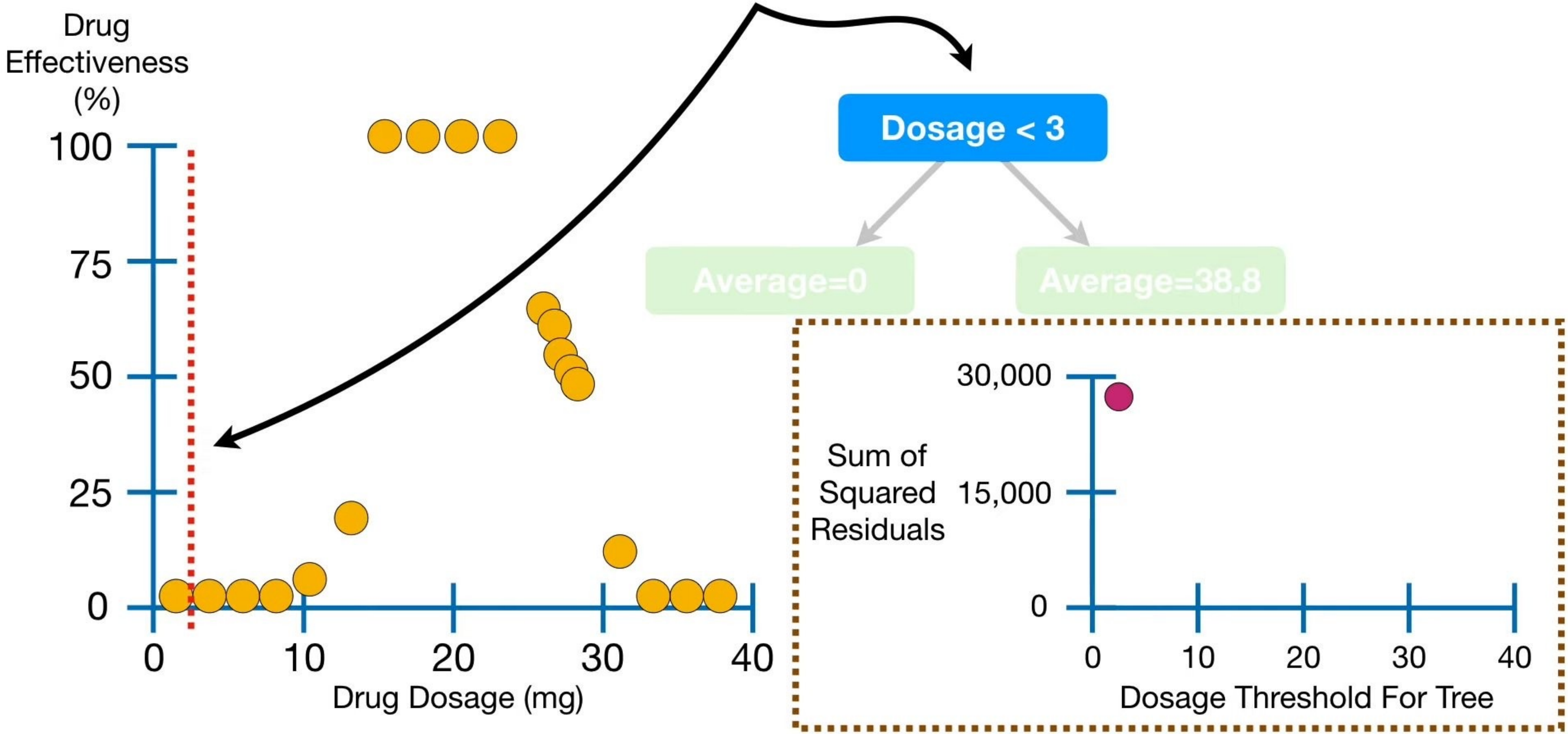
The **y-axis** corresponds to the sum of squared residuals…



…and the **x-axis** corresponds to **Dosage** thresholds.



In this case, the **Dosage** threshold was **3**…

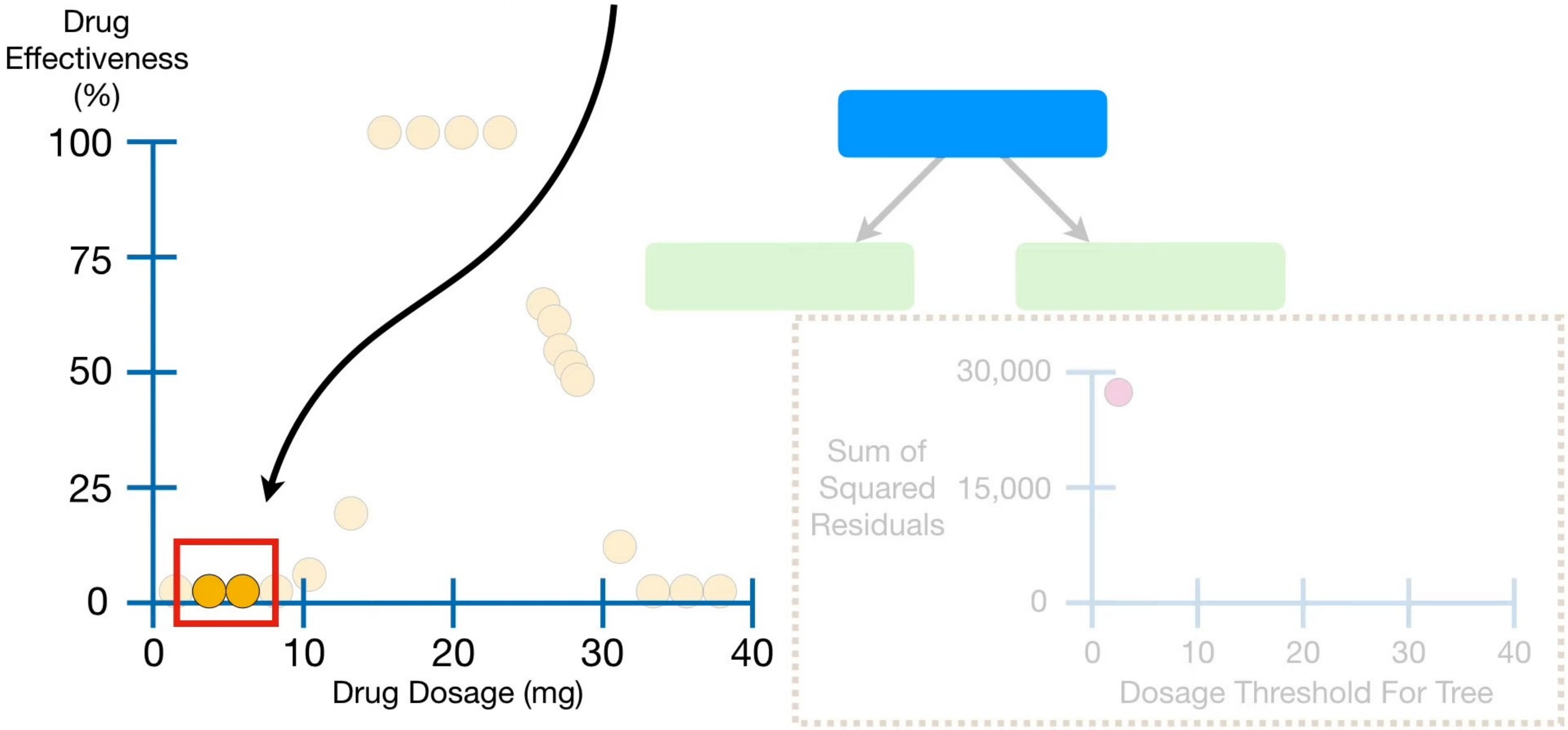


Dosage < 3

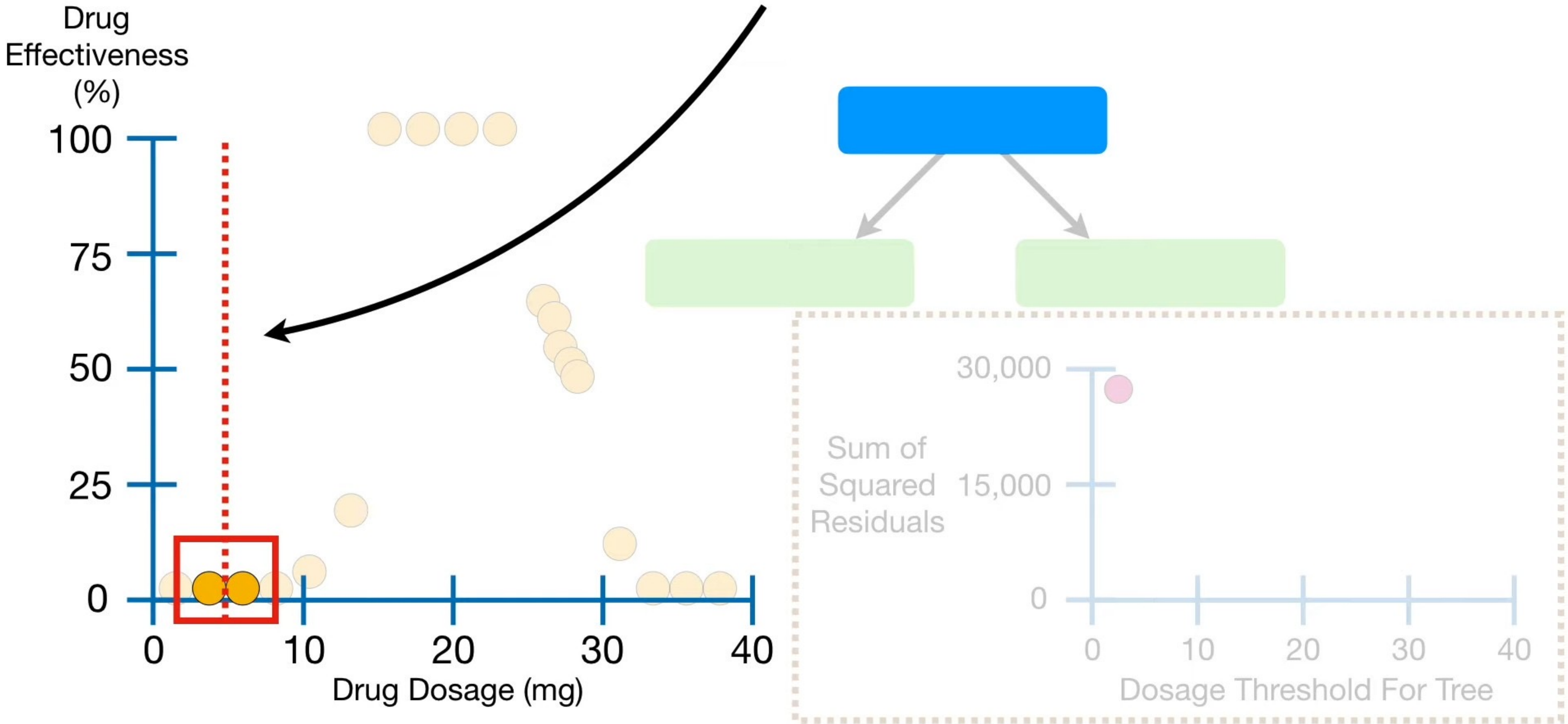
Average=0

Average=38.8

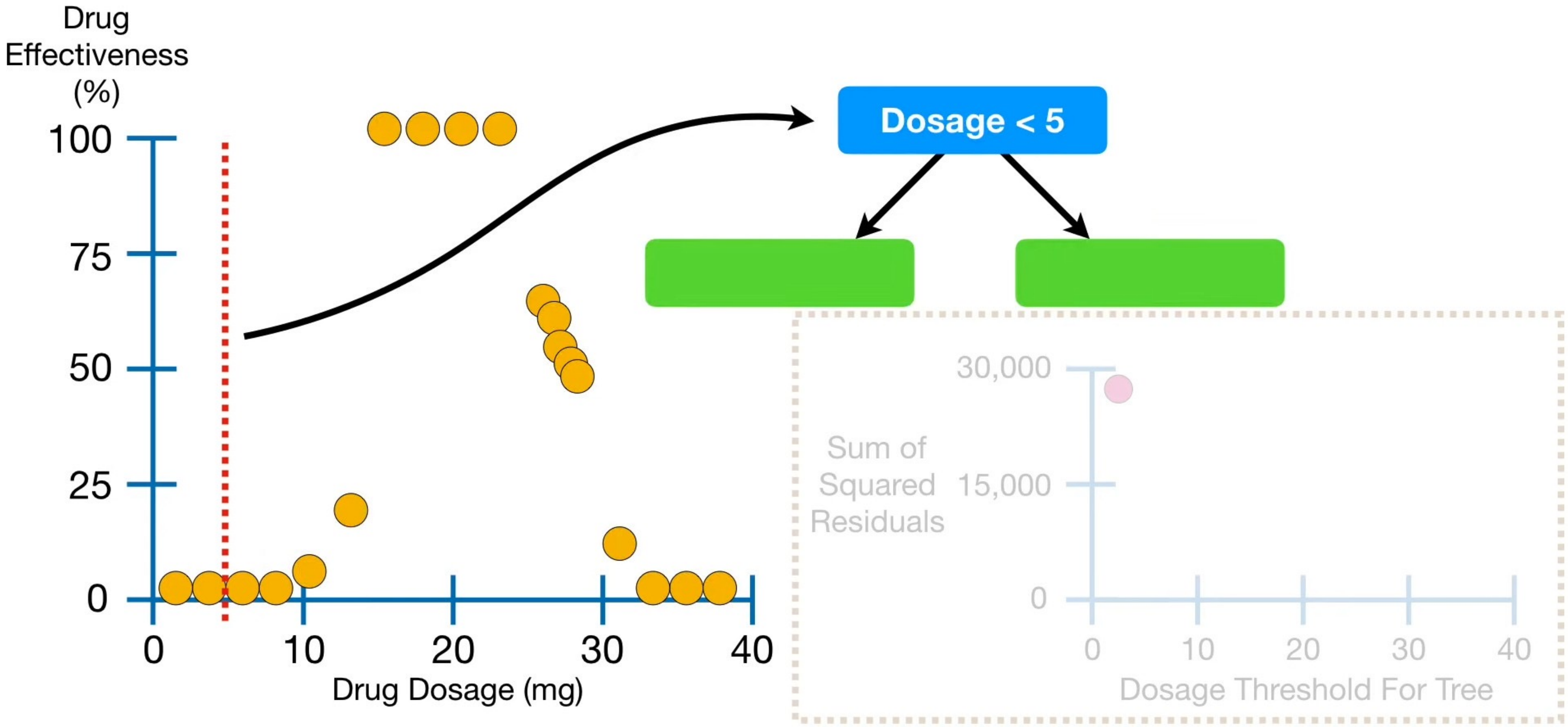
…but if we focus on the next two points in the graph…



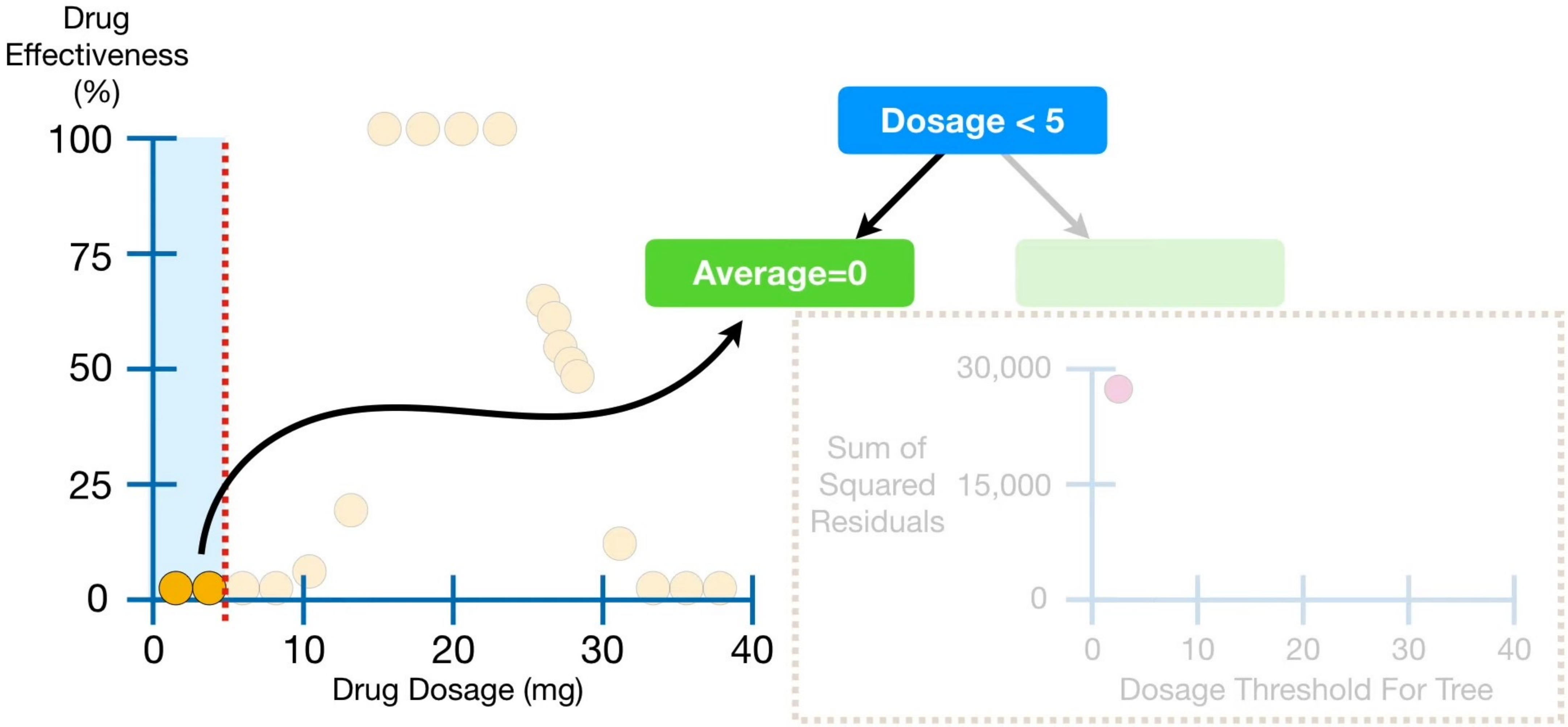
…and calculate their average **Dosage**, which is **5**…



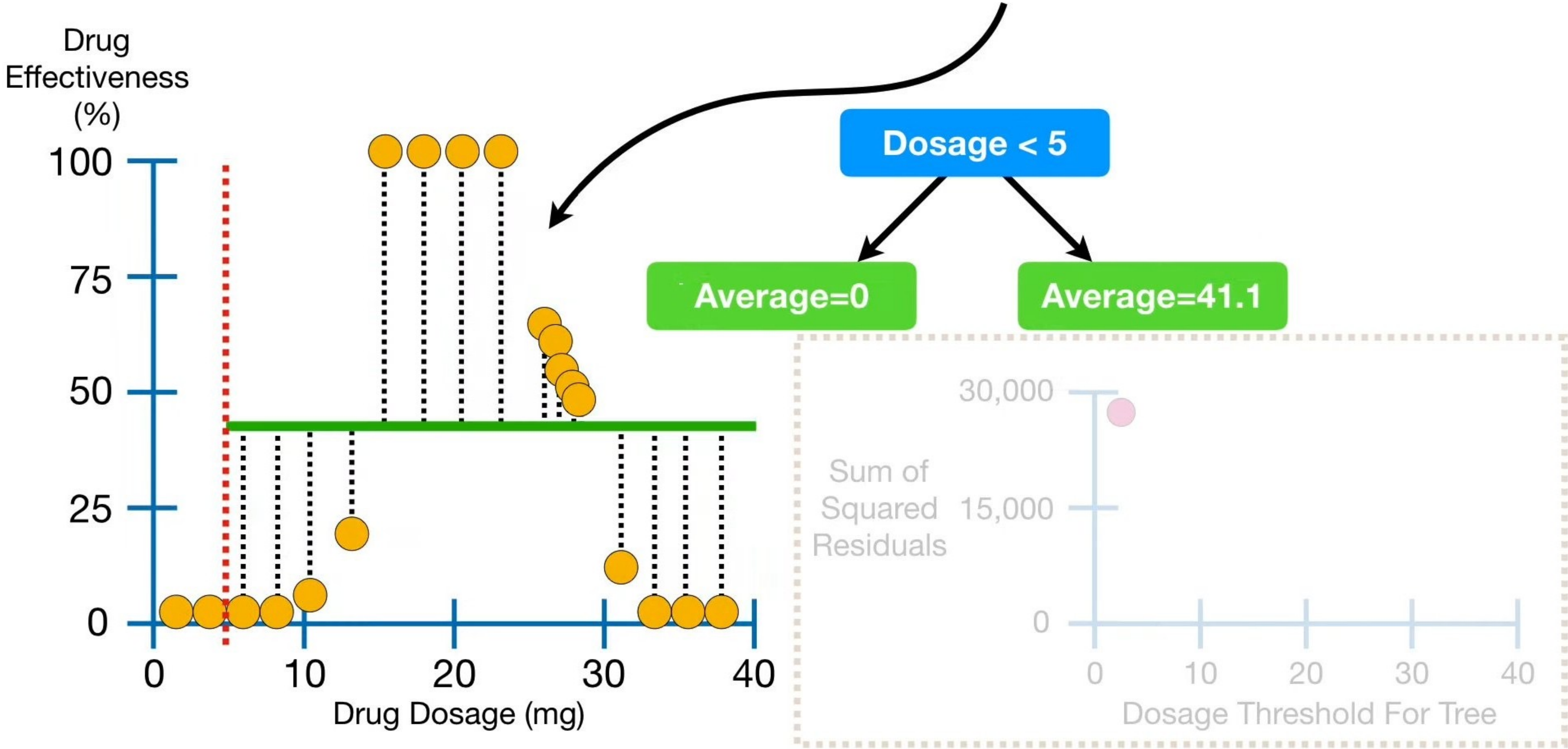
…then we can use **Dosage < 5** as a new threshold.



Using **Dosage < 5** gives us new predictions…



…and new residuals…

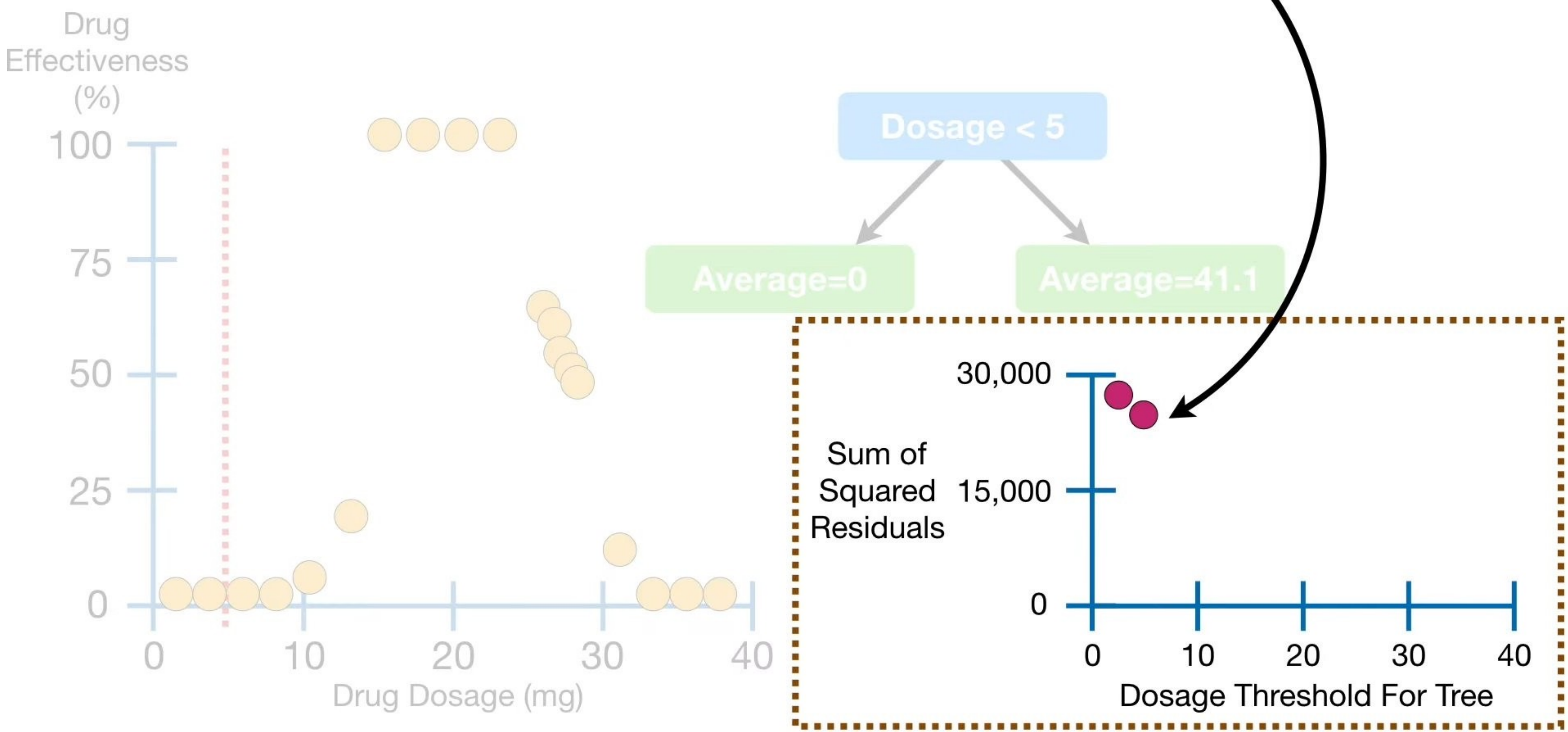


Dosage < 5

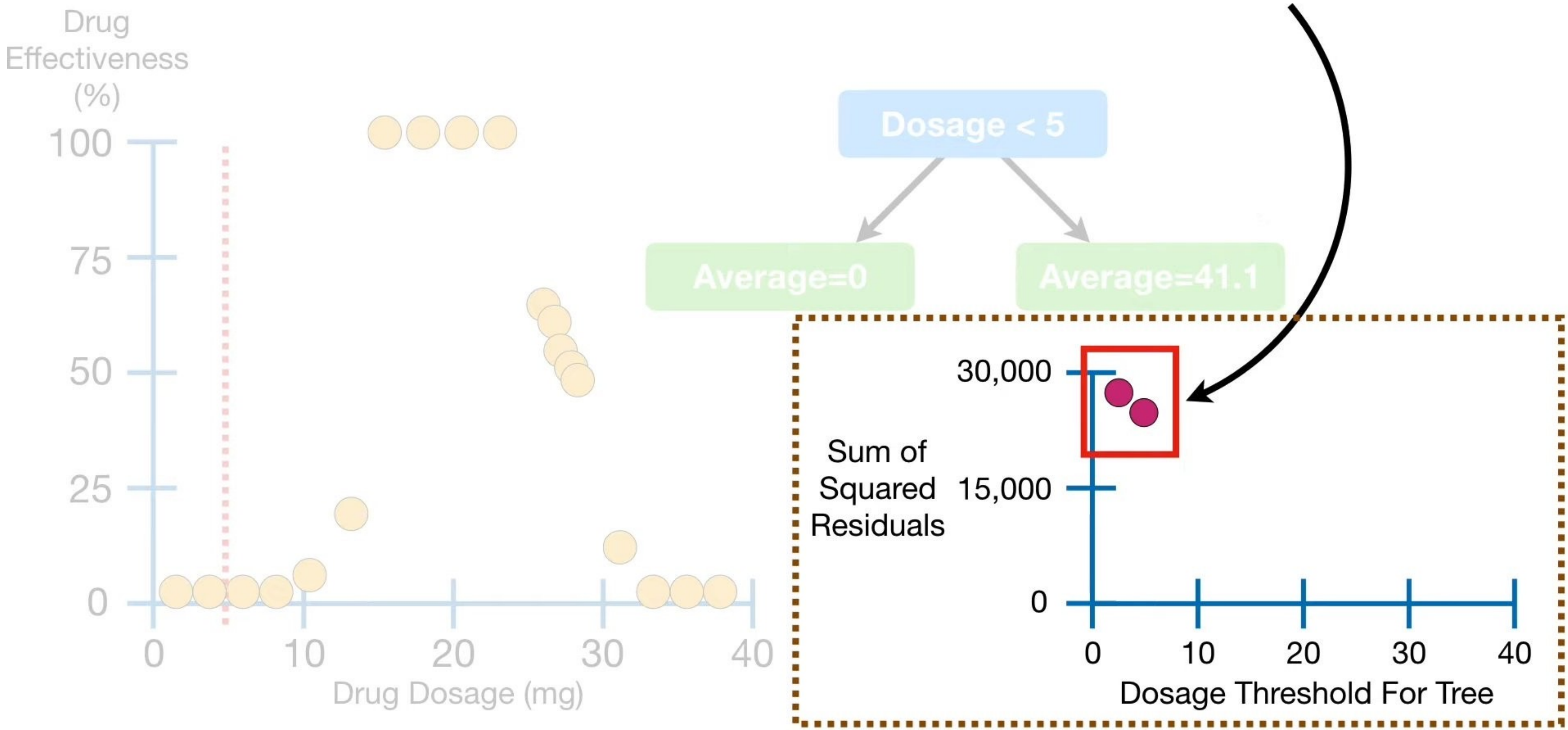
Average=0

Average=41.1

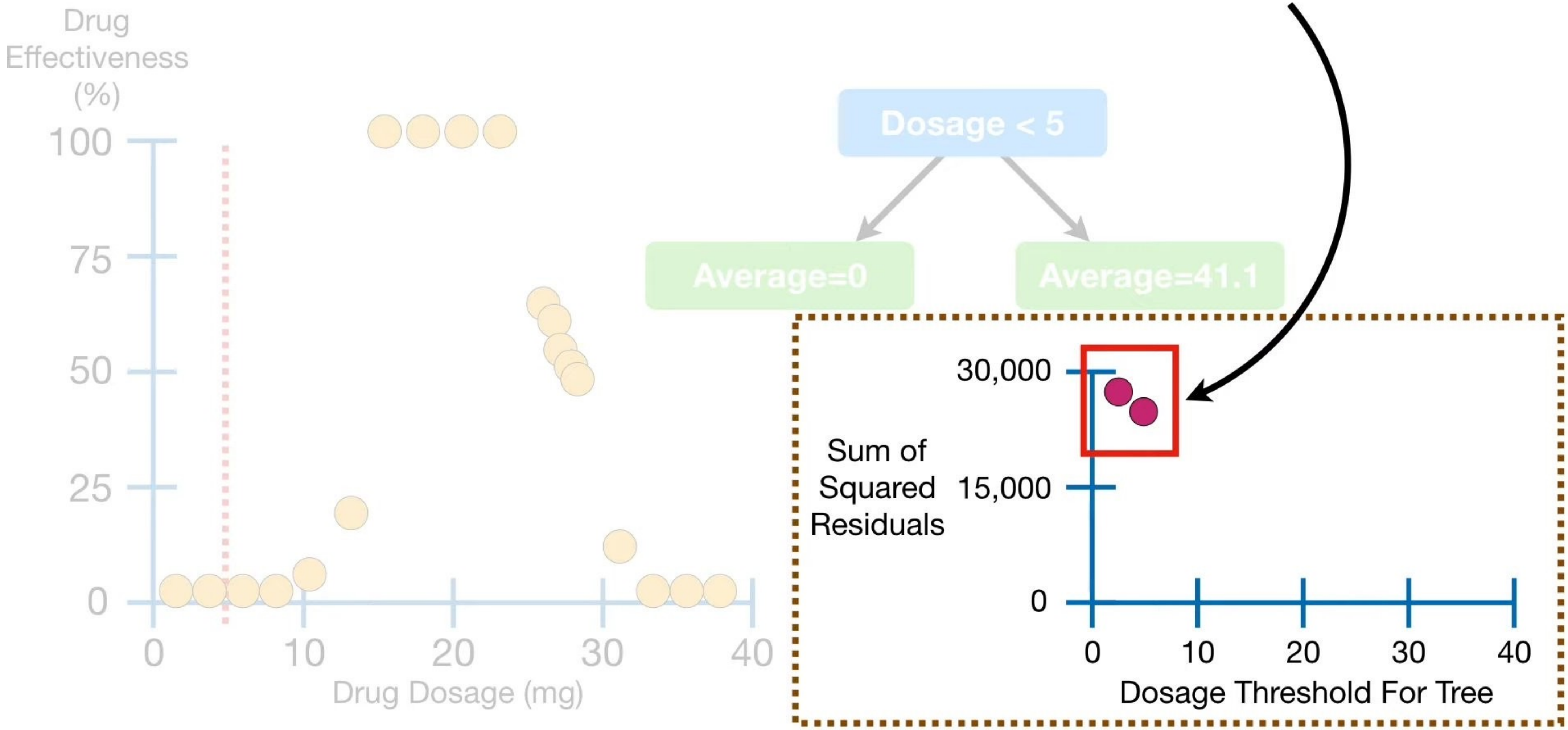
…and that means we can add a new sum of squared residuals to our graph.



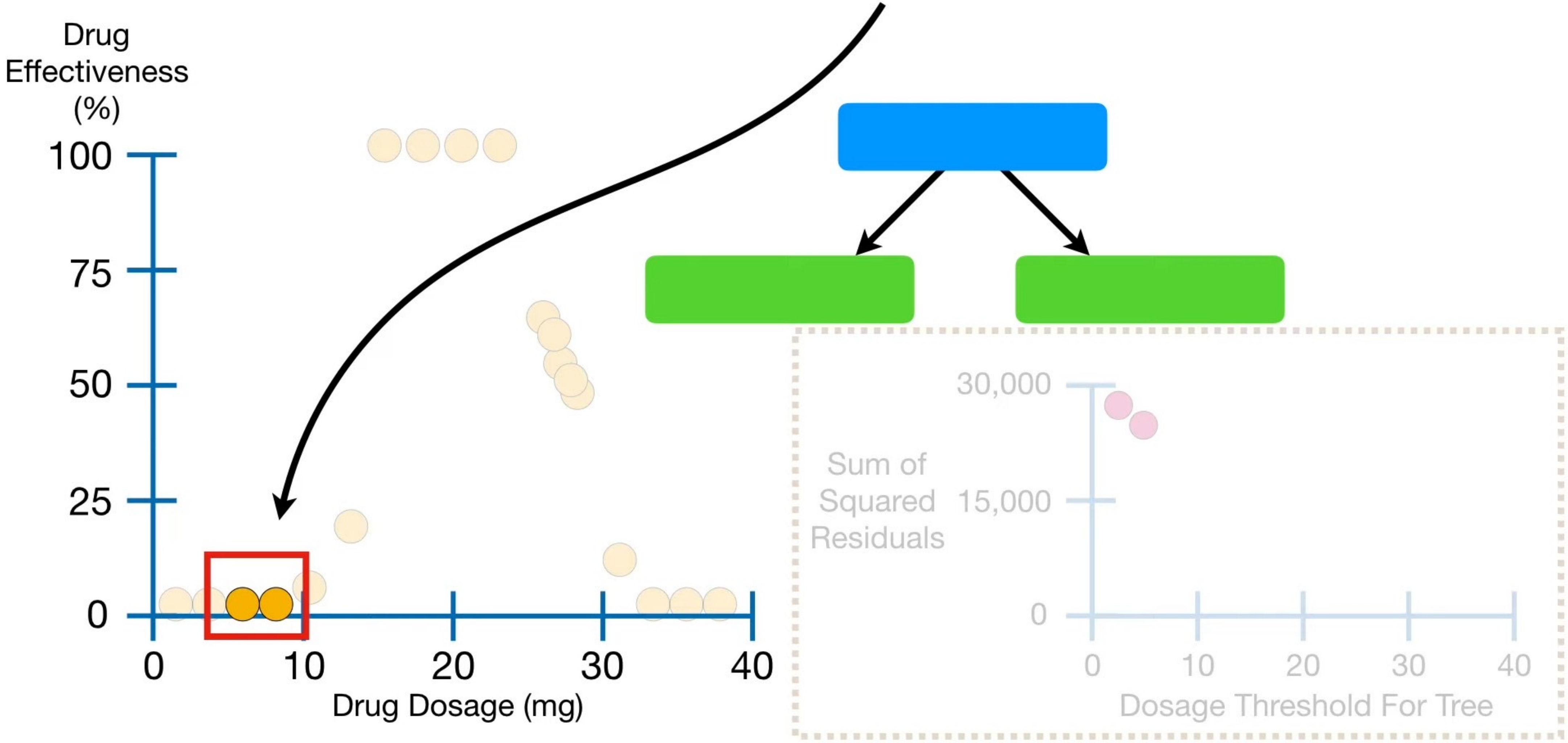
In this case, the new threshold, **Dosage < 5**, results in a smaller sum of squared residuals…



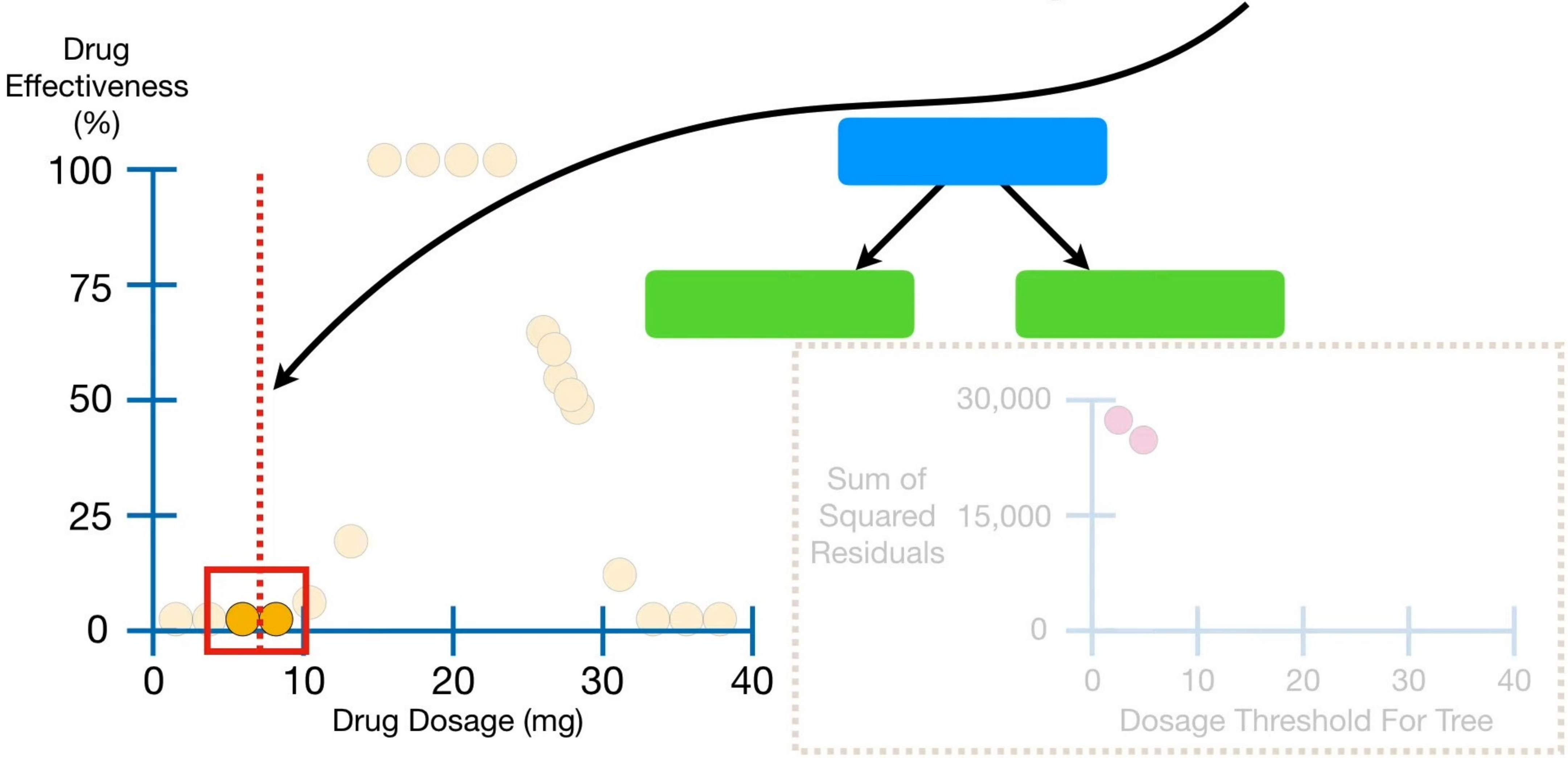
...and that means using **Dosage < 5** as the threshold resulted in better predictions over all.



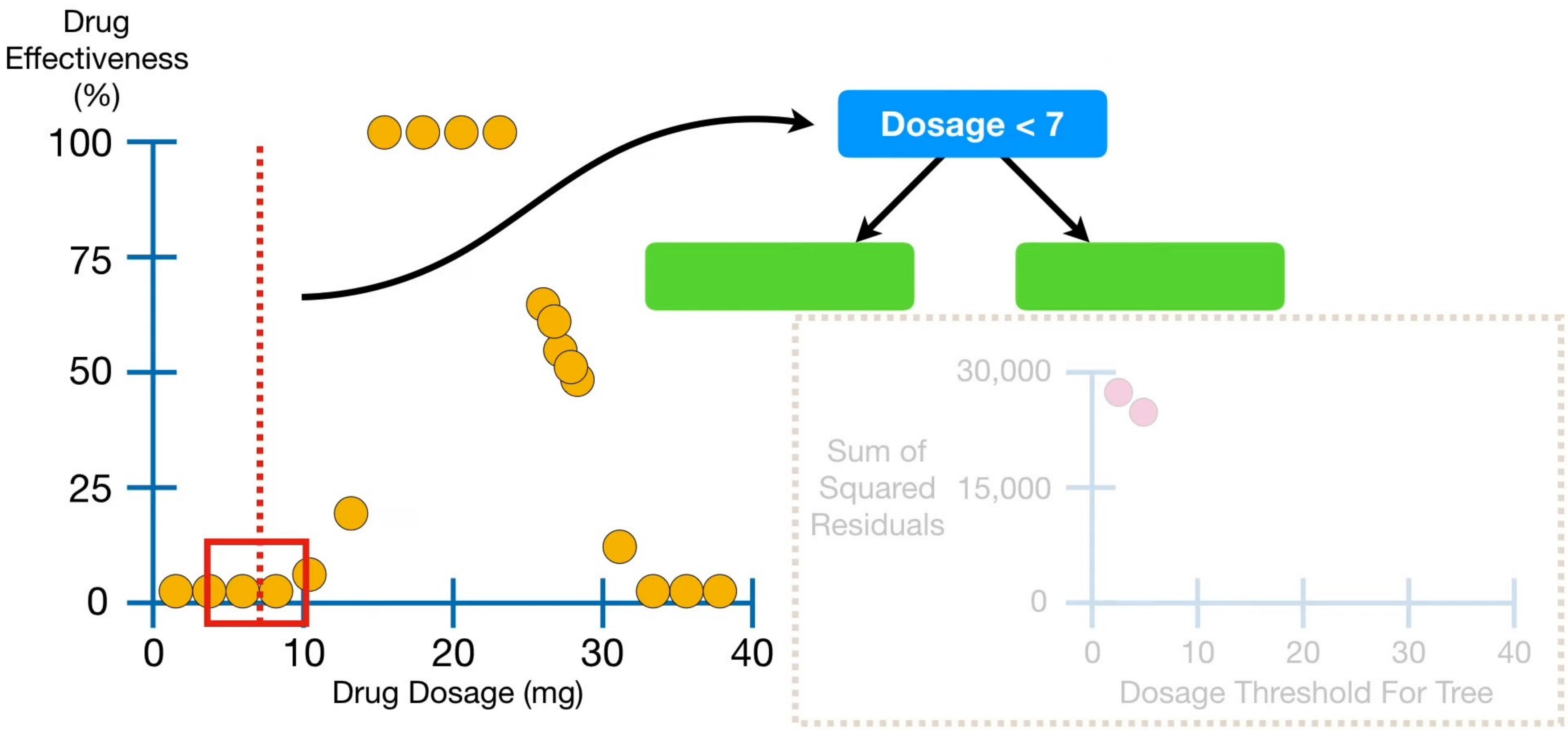
Now let's focus on the next two points…



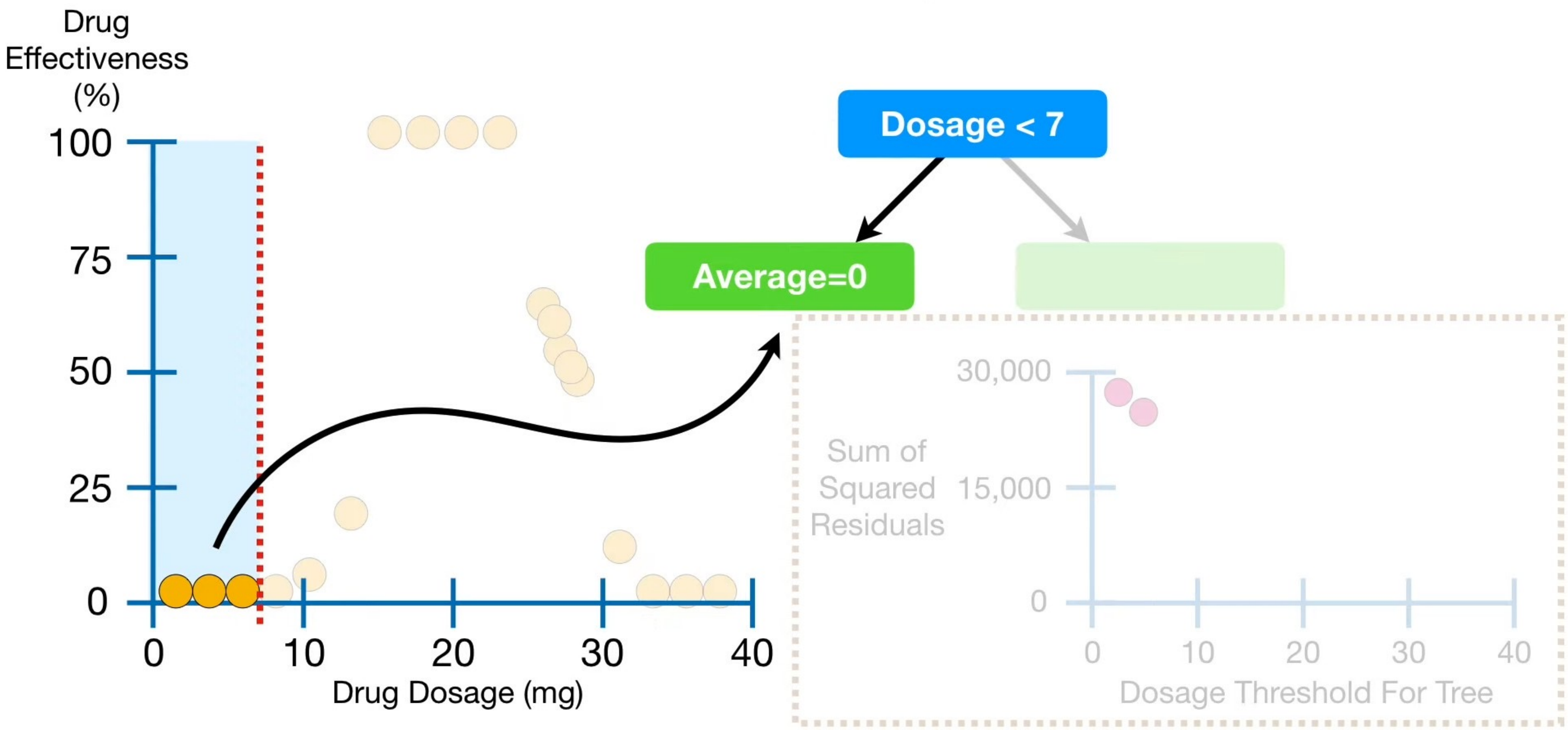
…calculate their average, which is **7**…



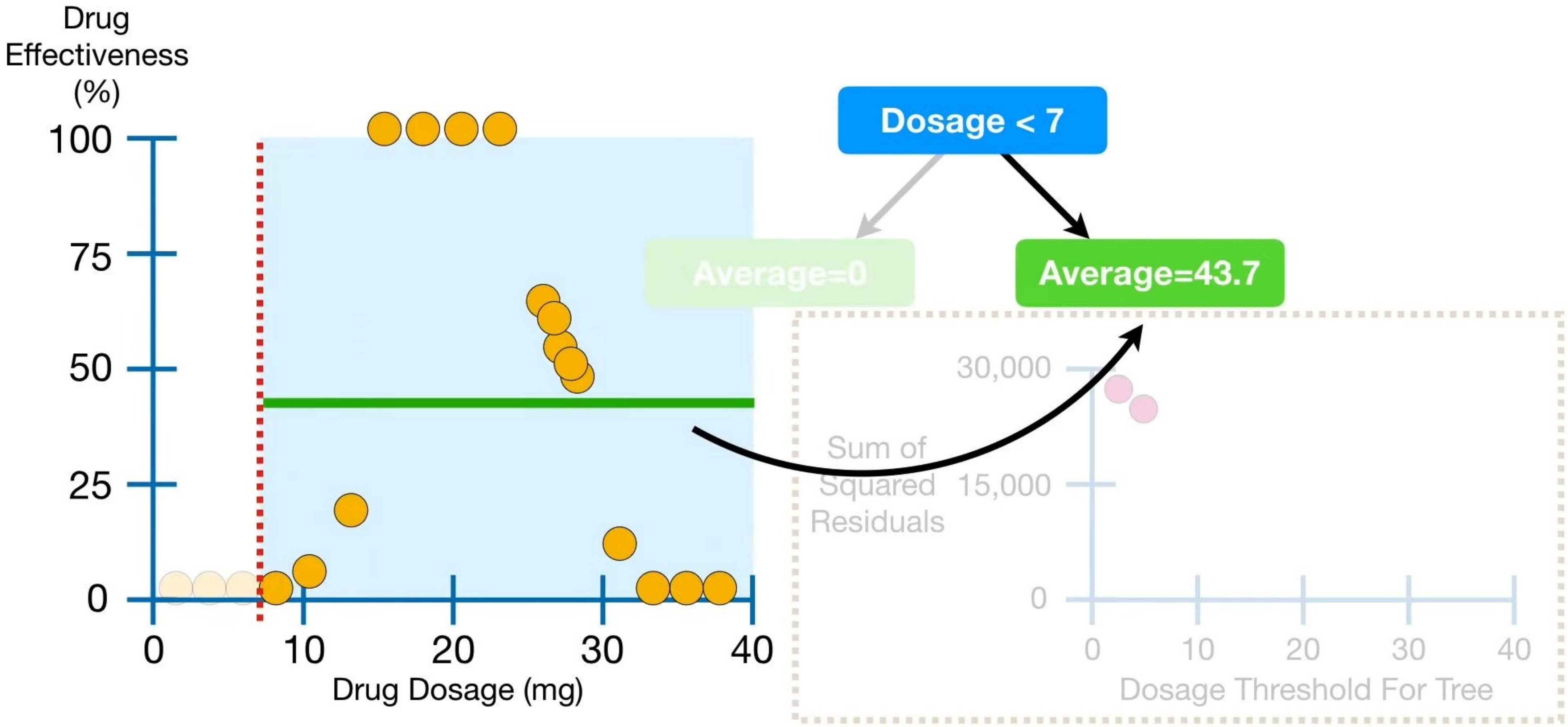
…and use **Dosage < 7** as a new threshold.



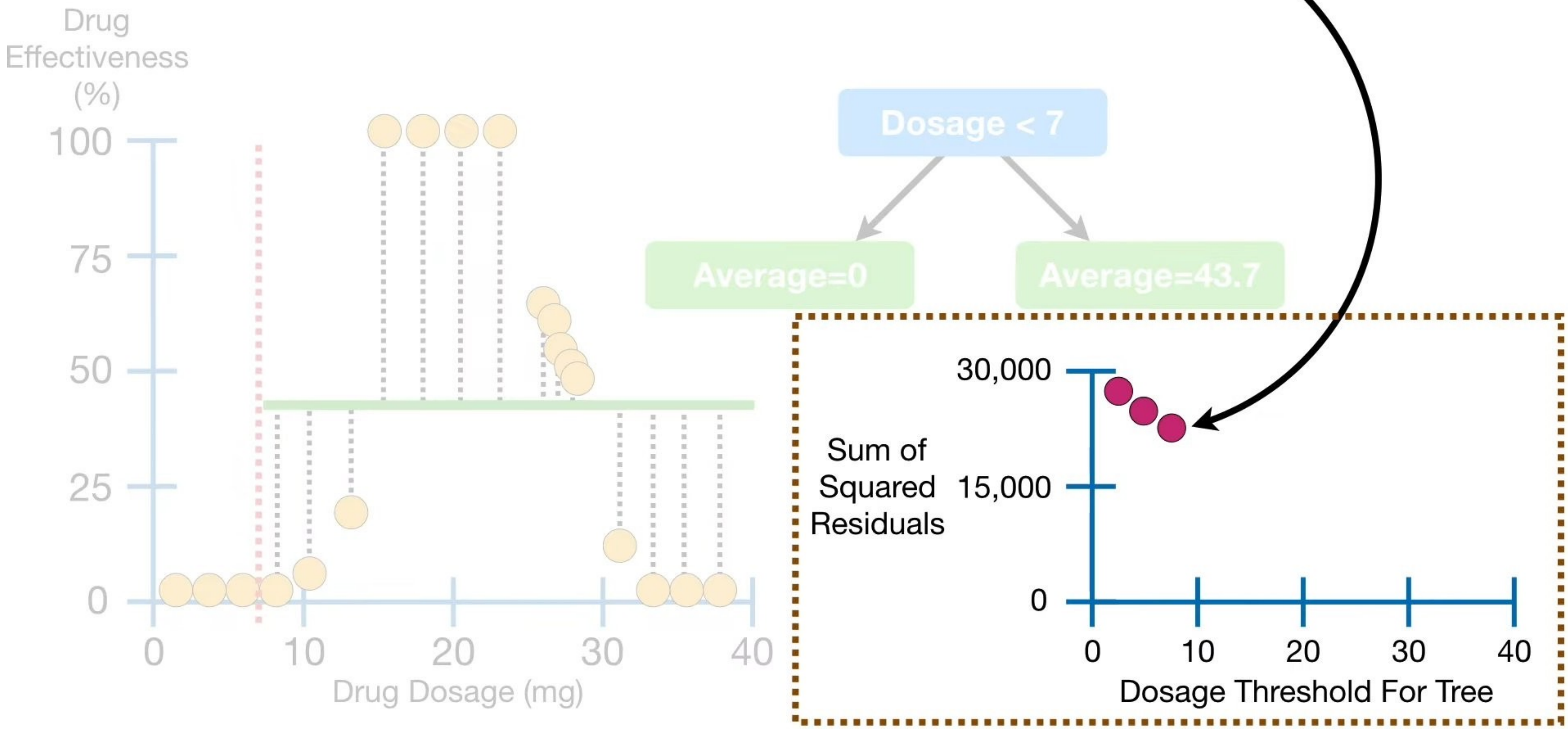
Again, the new threshold gives us new predictions…



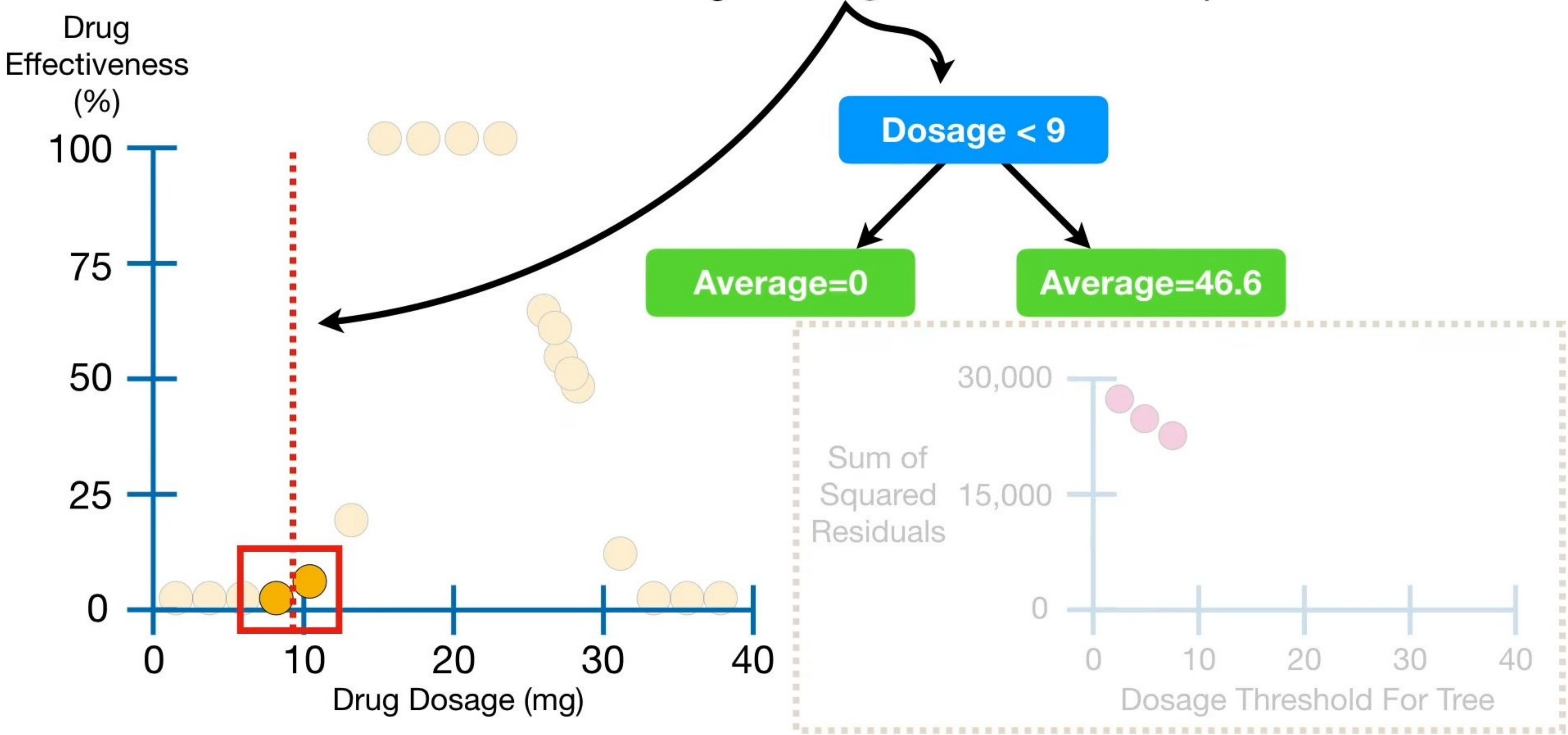
Again, the new threshold gives us new predictions…



…and a new sum of squared residuals.



Now shift the threshold over to the average **Dosage** for the next two points…

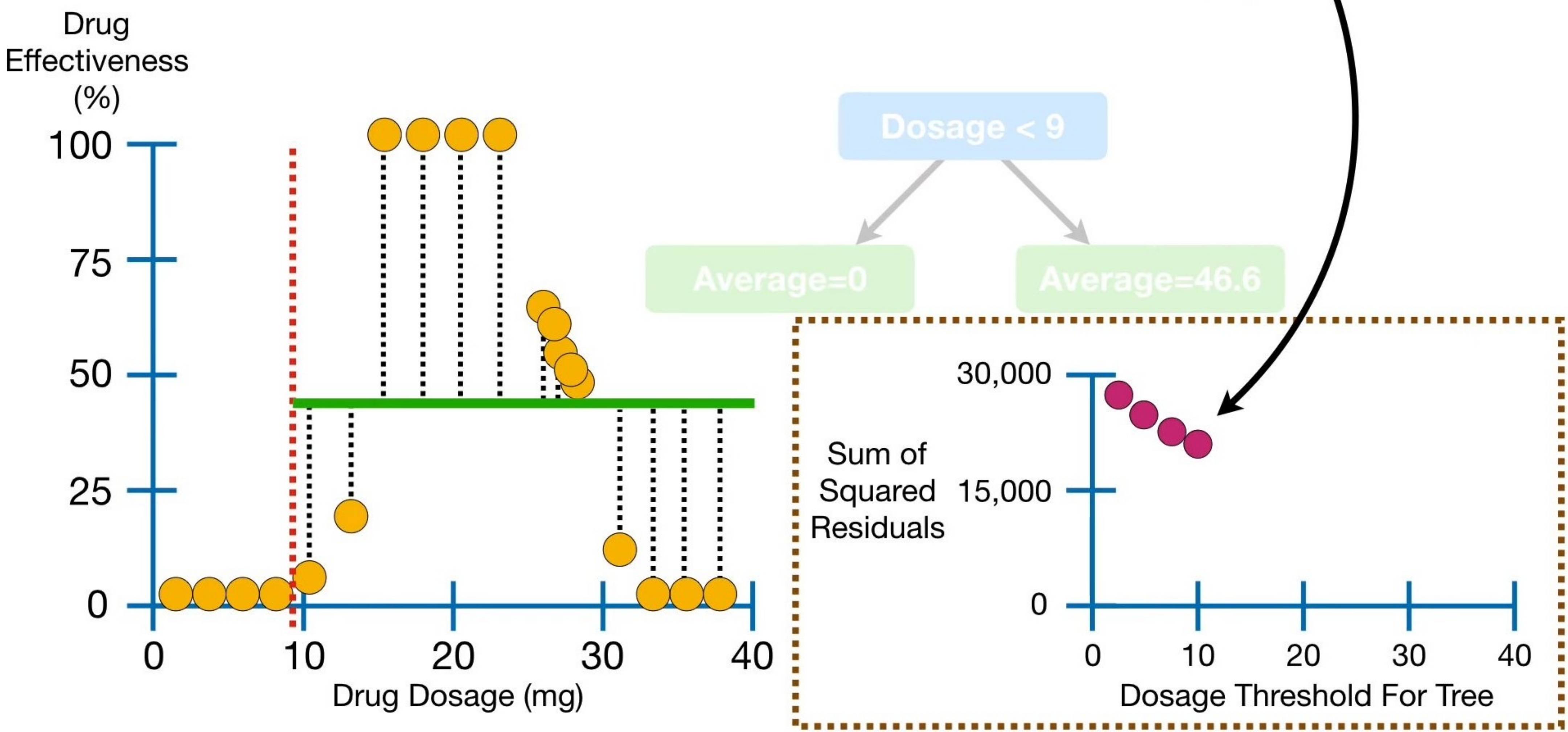


Dosage < 9

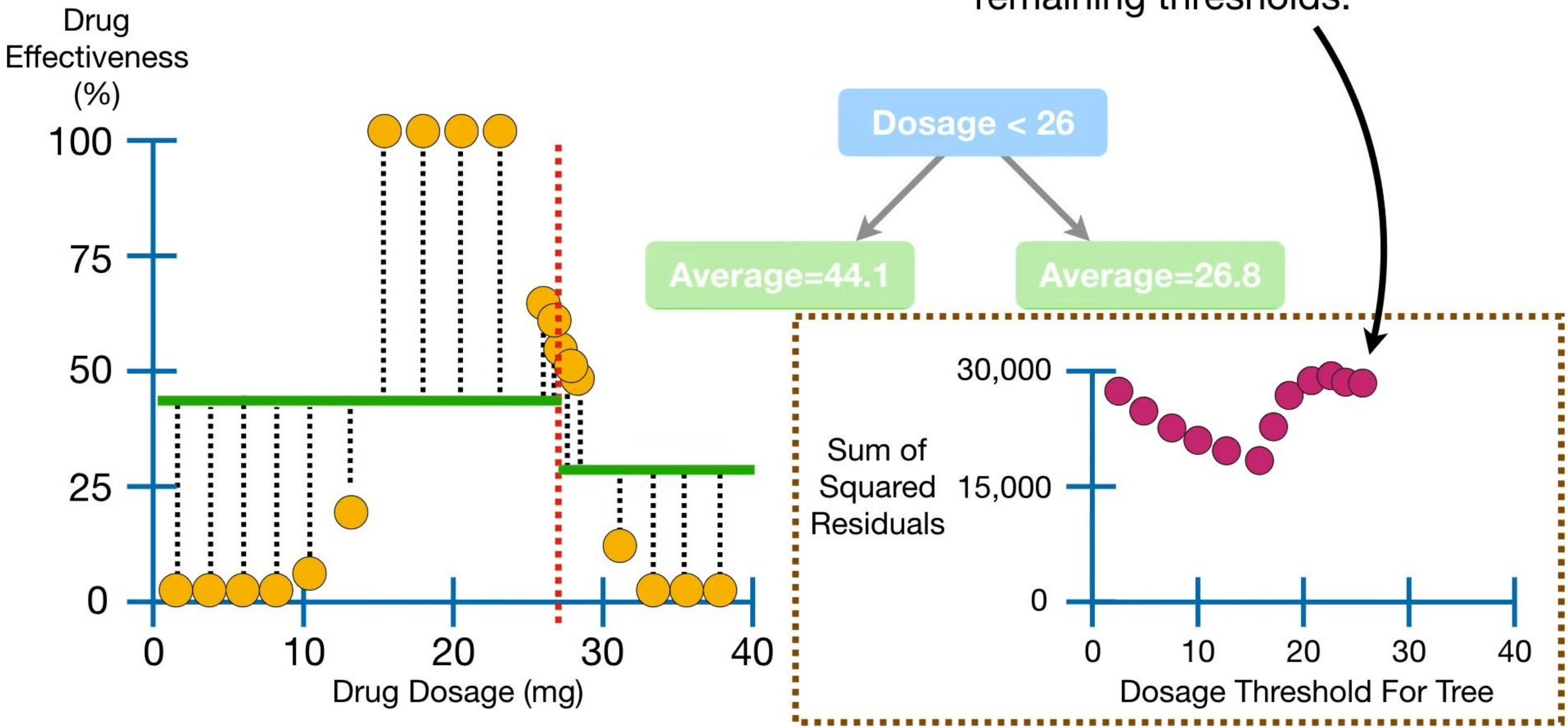
Average=0

Average=46.6

…and add the new sum of squared residuals to the graph.



And we repeat until we have calculated the sum of squared residuals for all of the remaining thresholds.

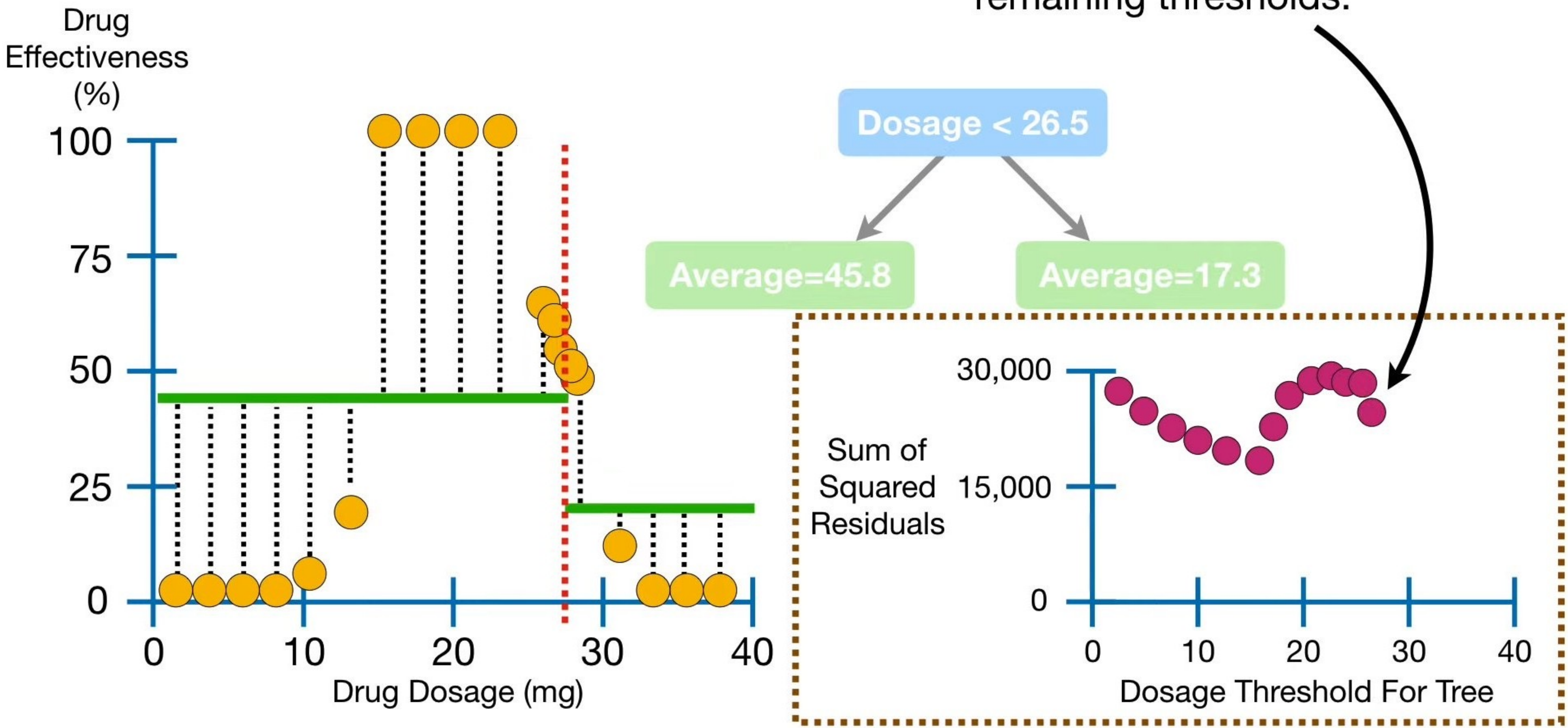


Dosage < 26

Average=44.1

Average=26.8

And we repeat until we have calculated the sum of squared residuals for all of the remaining thresholds.

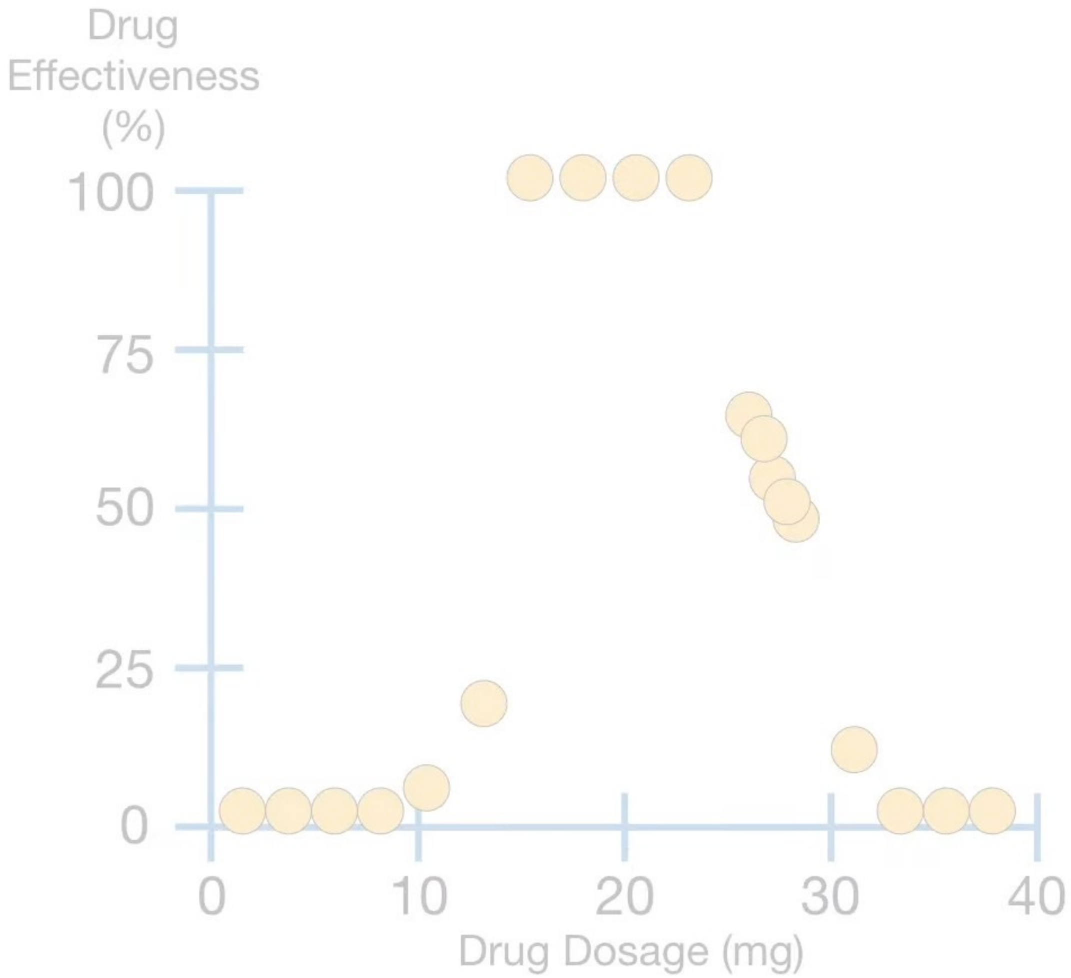


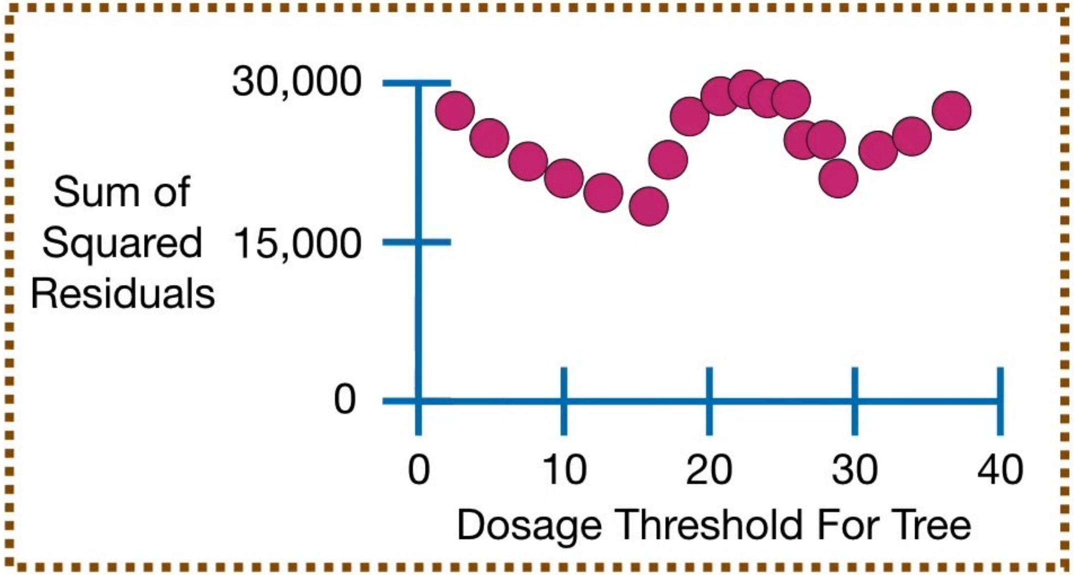
Dosage < 26.5

Average=45.8

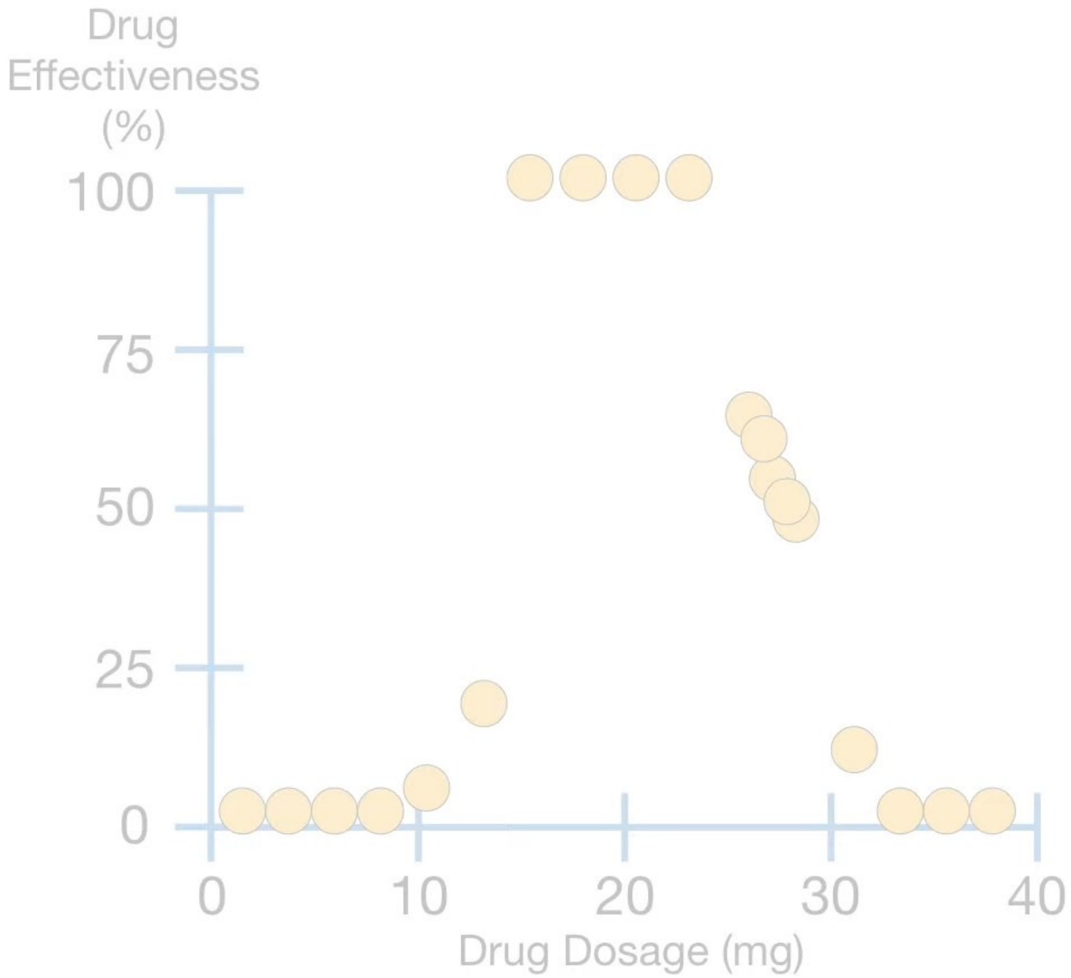
Average=17.3

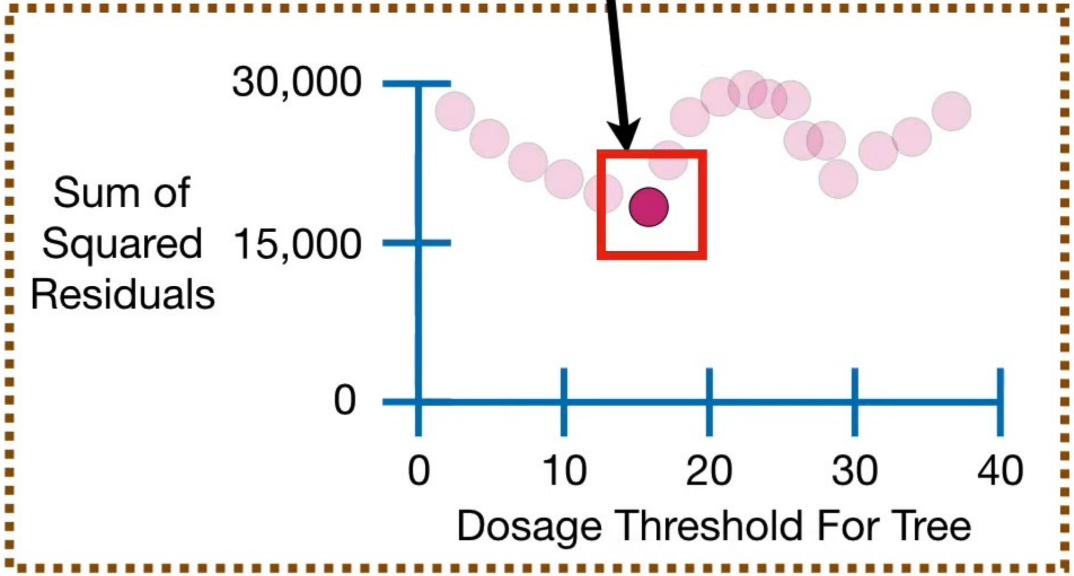
Now we can see the sum of squared residuals for all of the thresholds…



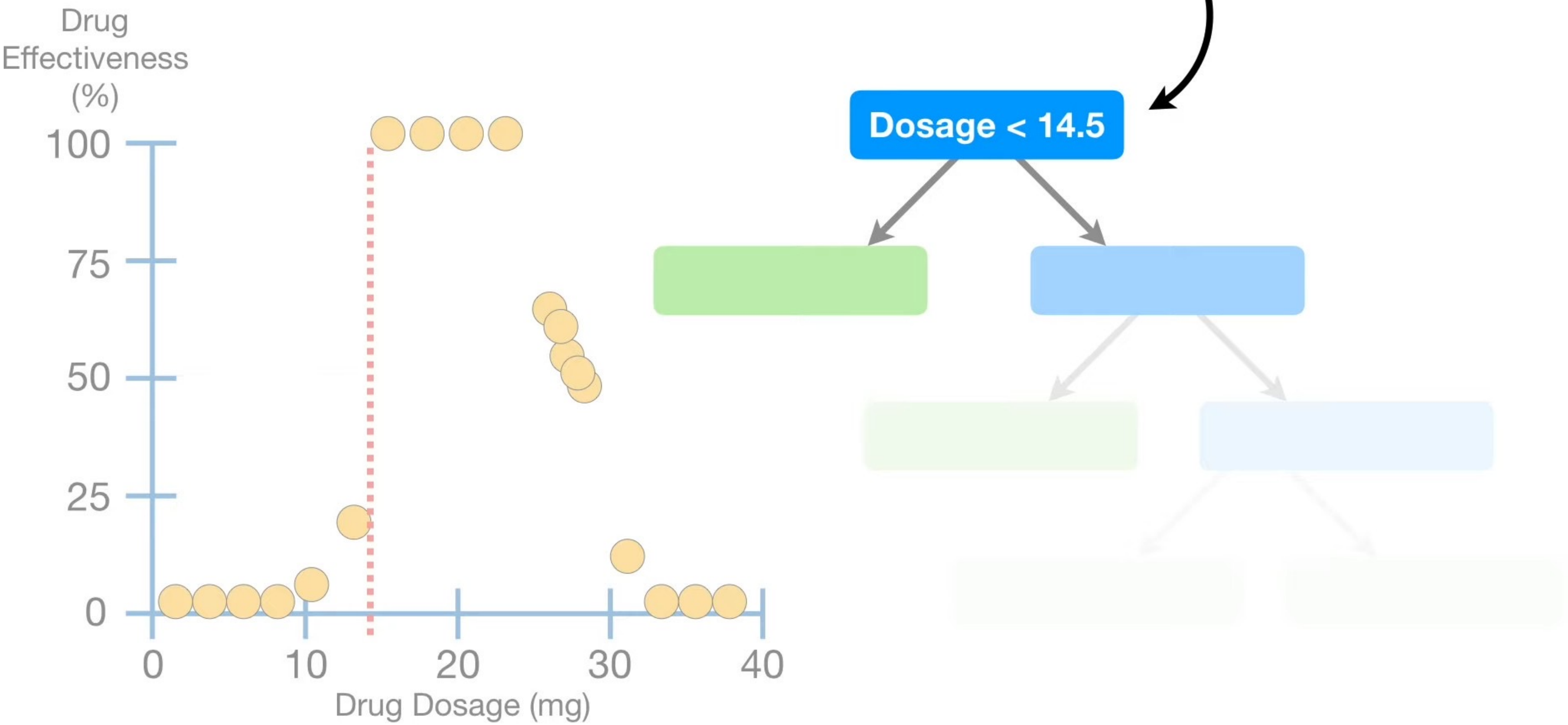


…and **Dosage < 14.5** had the smallest sum of squared residuals…

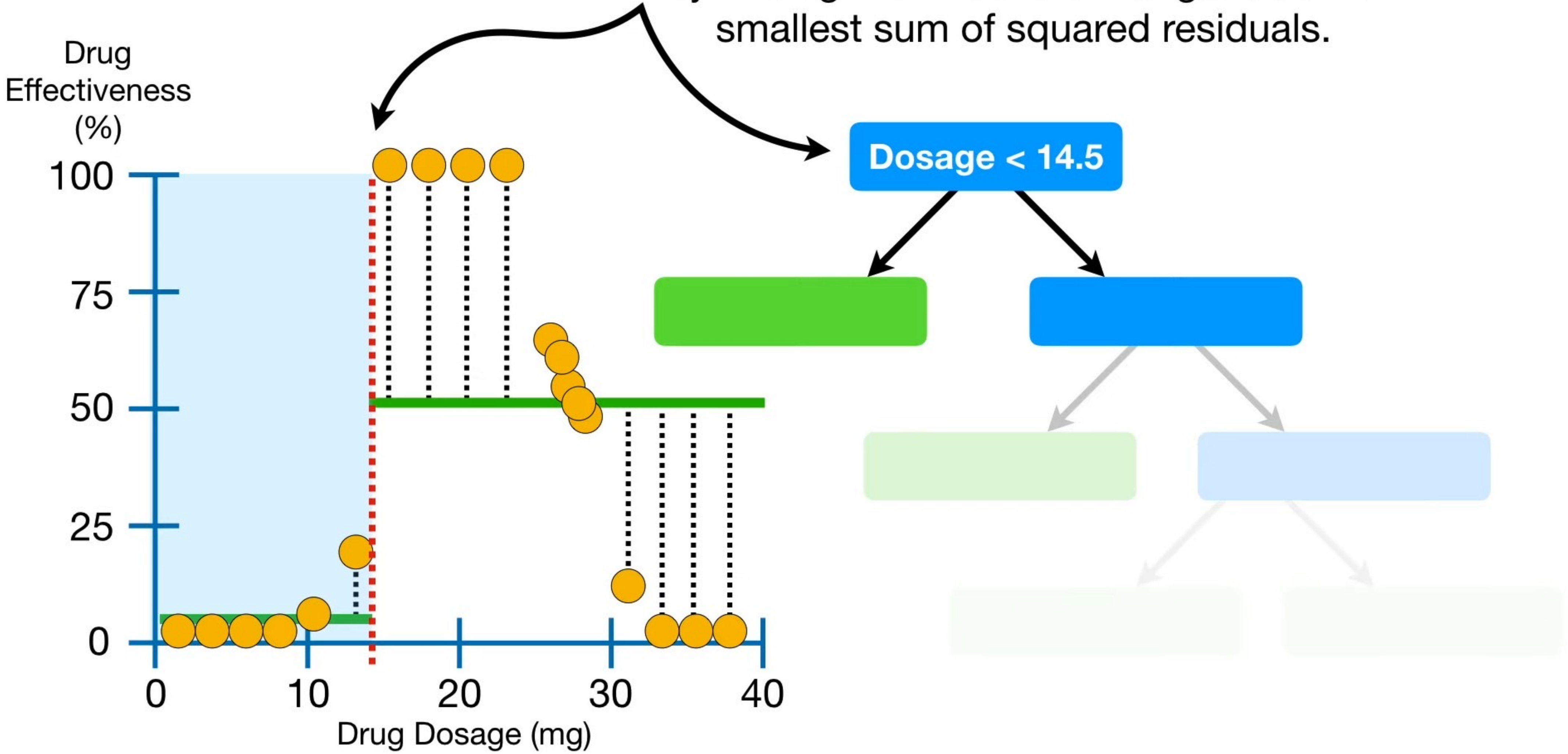




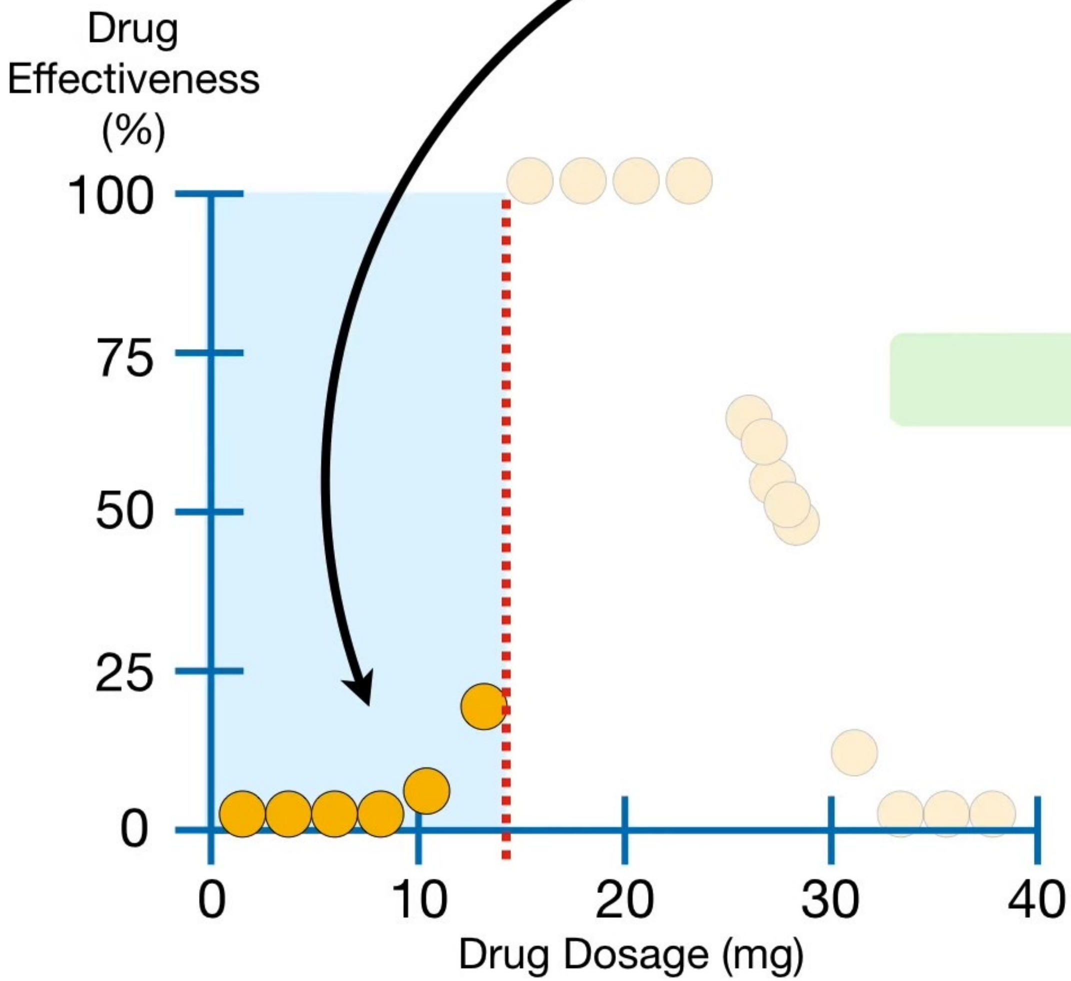
…so **Dosage < 14.5** will be root of the tree.

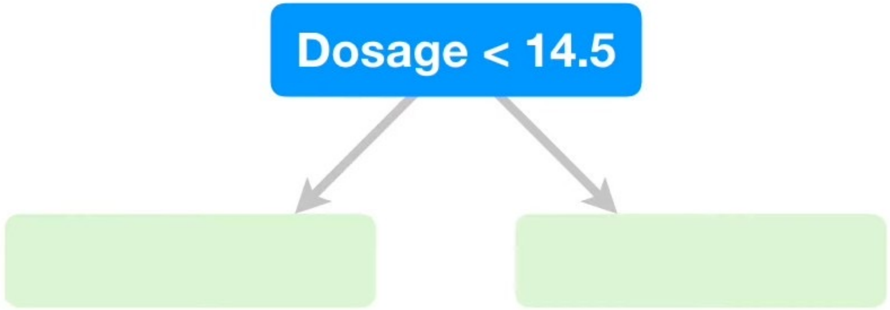


In summary, we split the data into two groups by finding the threshold that gave us the smallest sum of squared residuals.

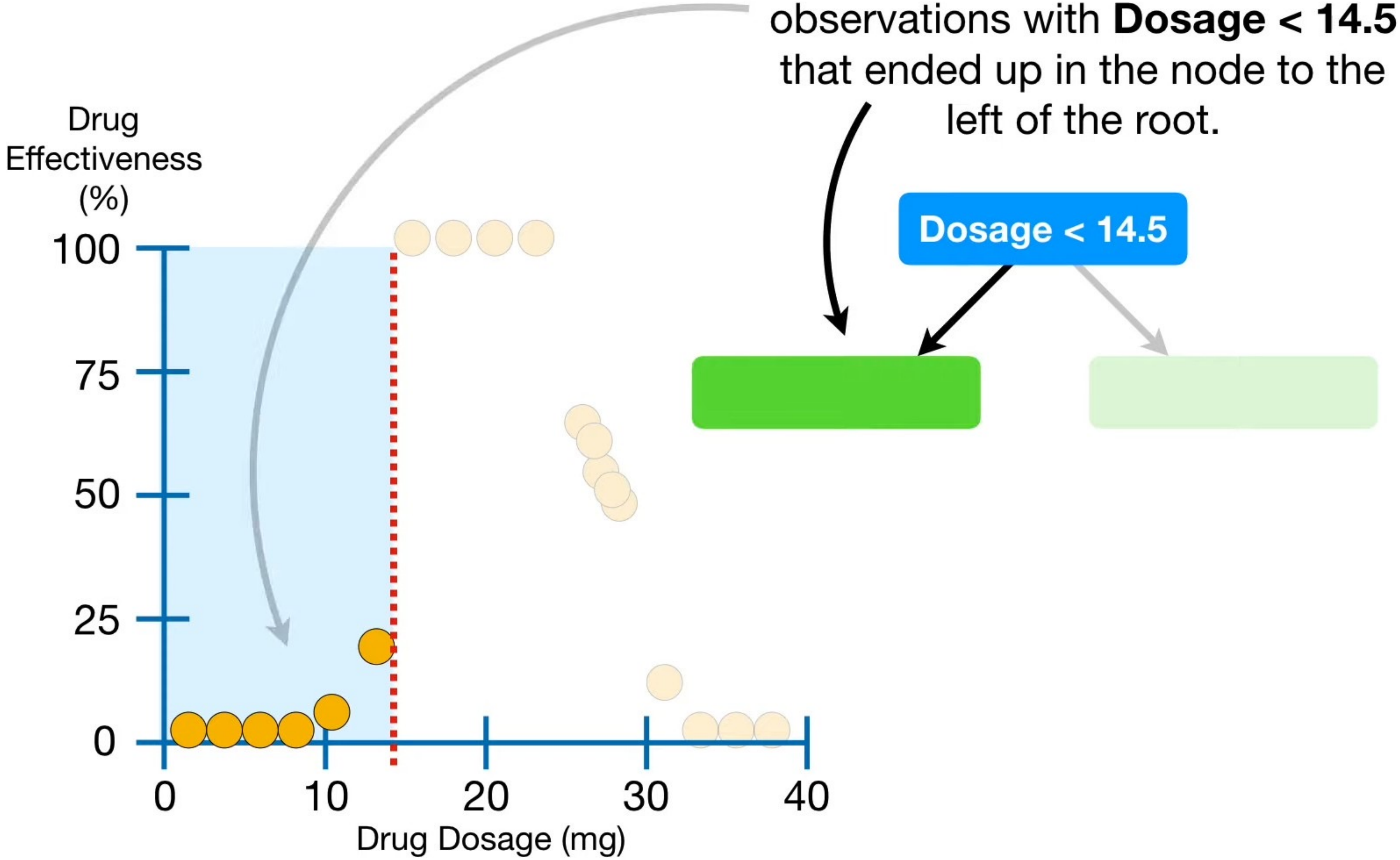


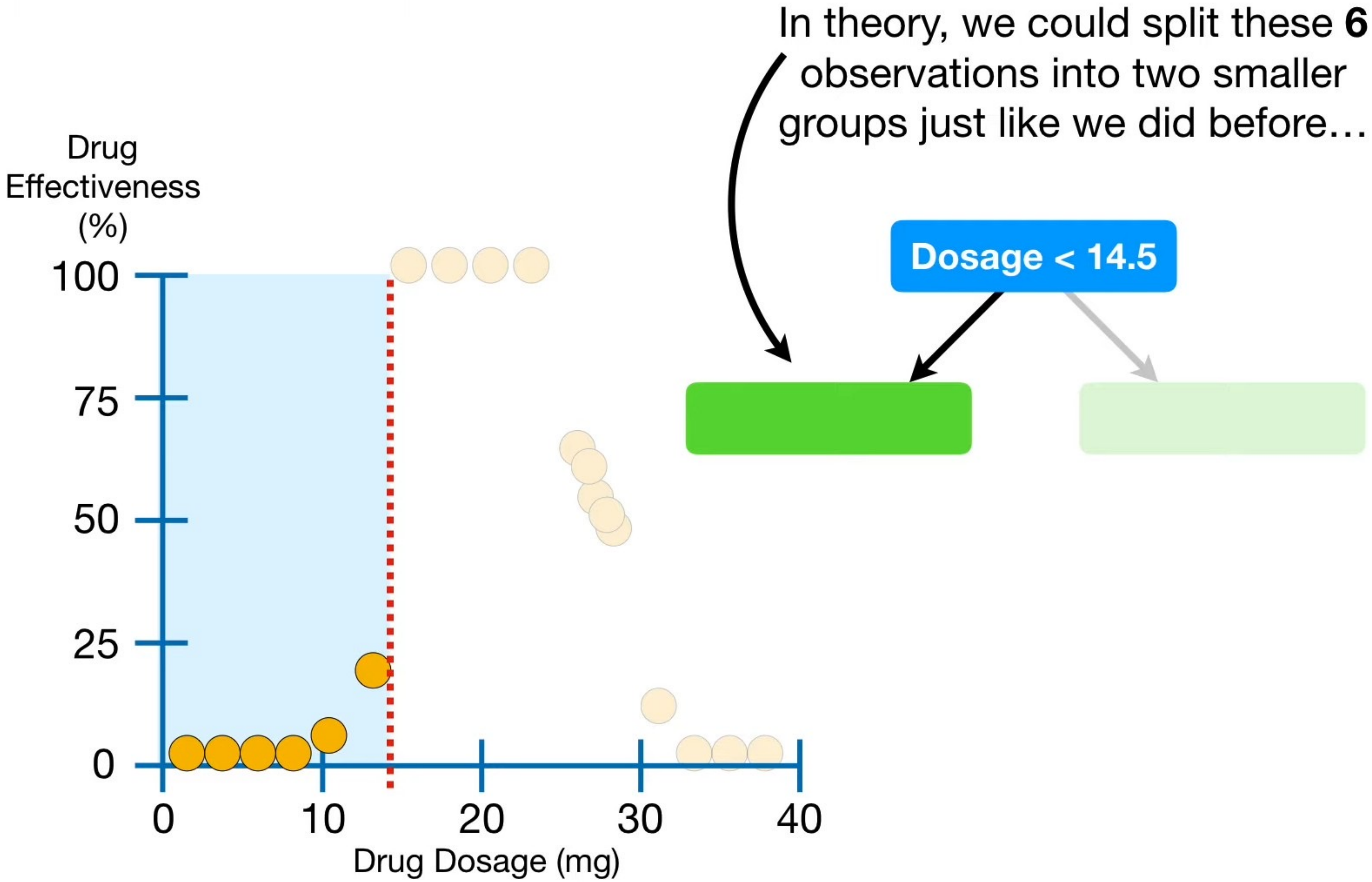
Now let’s focus on the **6** observations with **Dosage < 14.5** that ended up in the node to the left of the root.





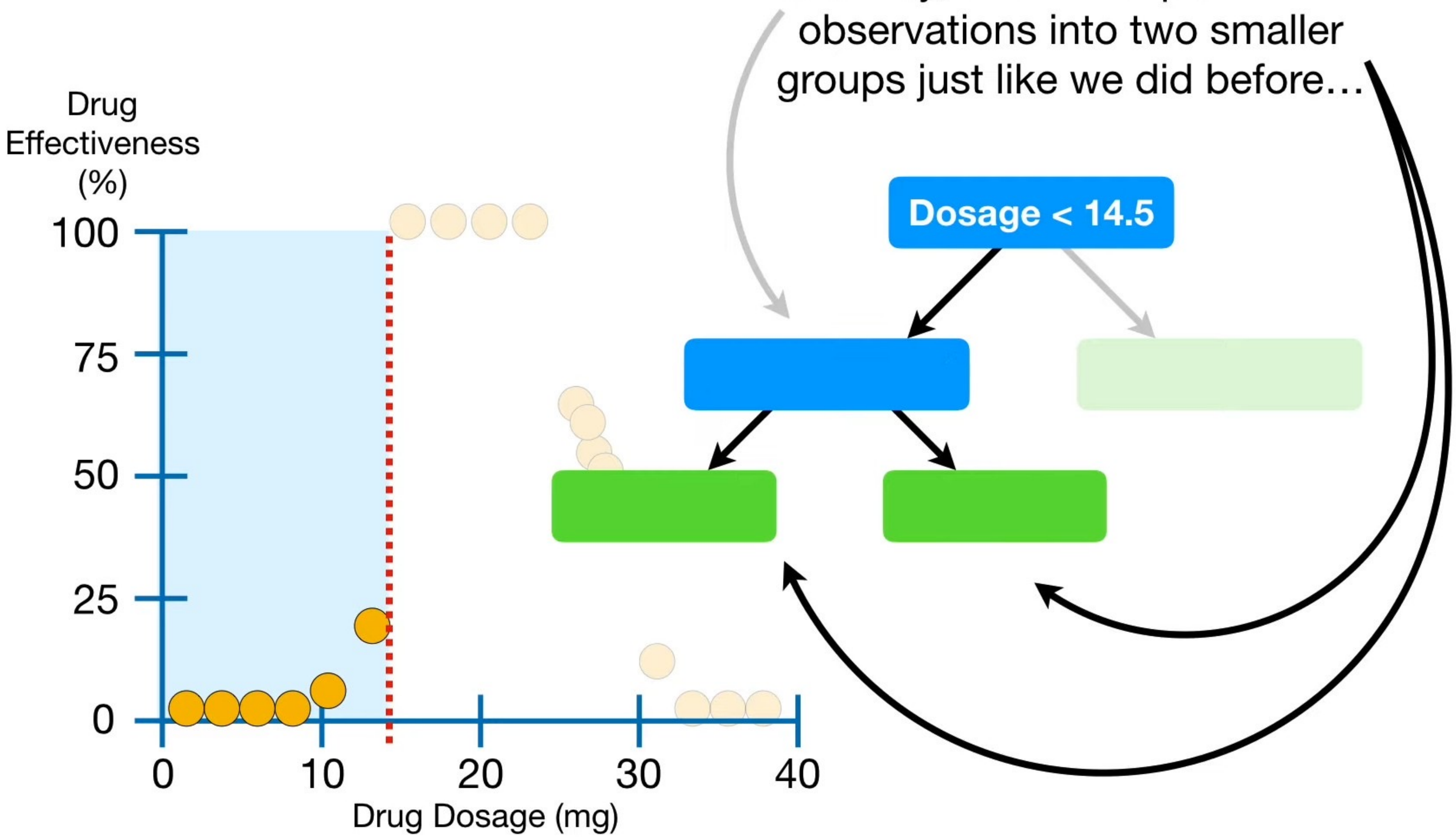
Now let's focus on the **6** observations with **Dosage < 14.5** that ended up in the node to the left of the root.



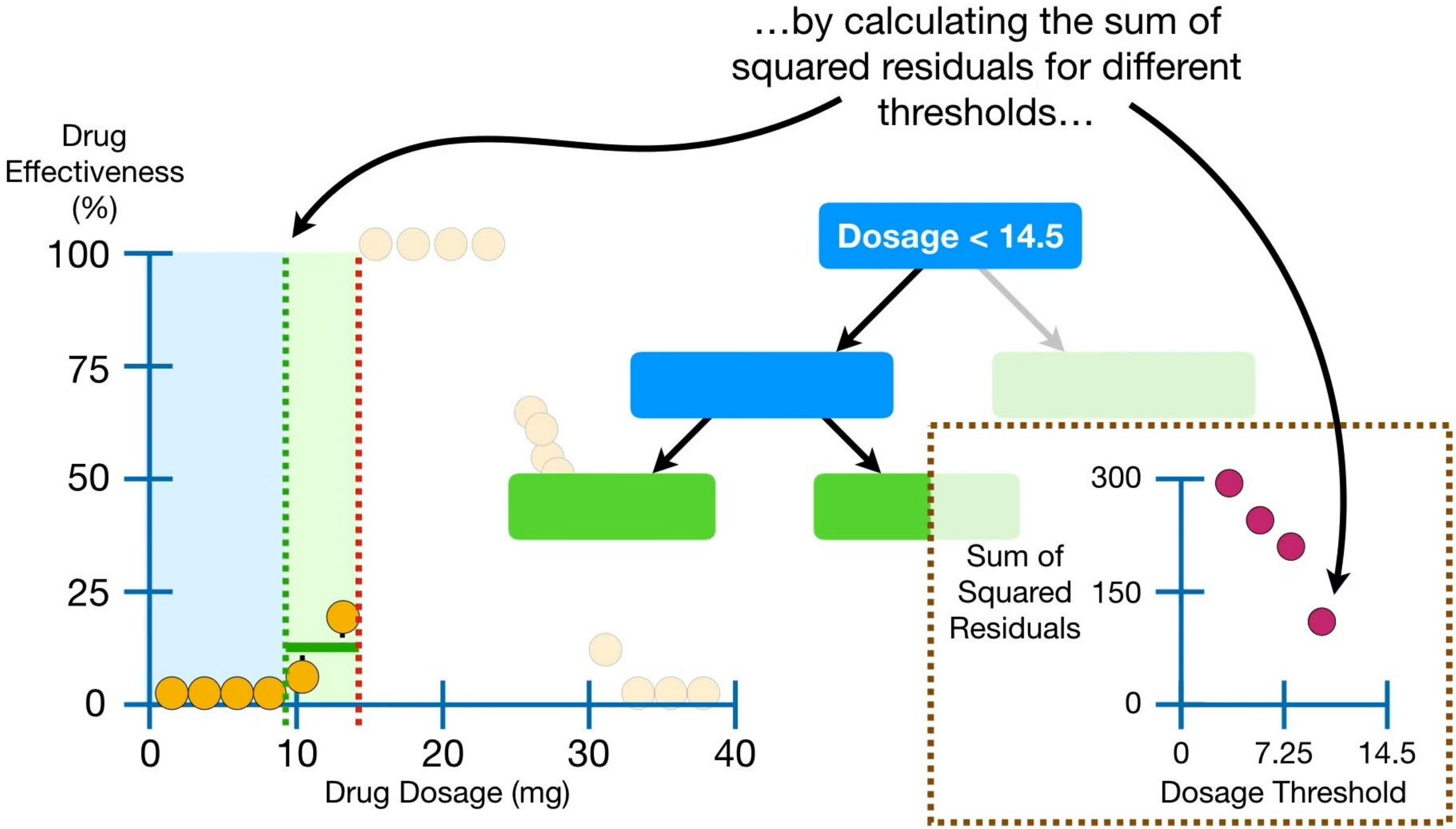


In theory, we could split these **6** observations into two smaller groups just like we did before…

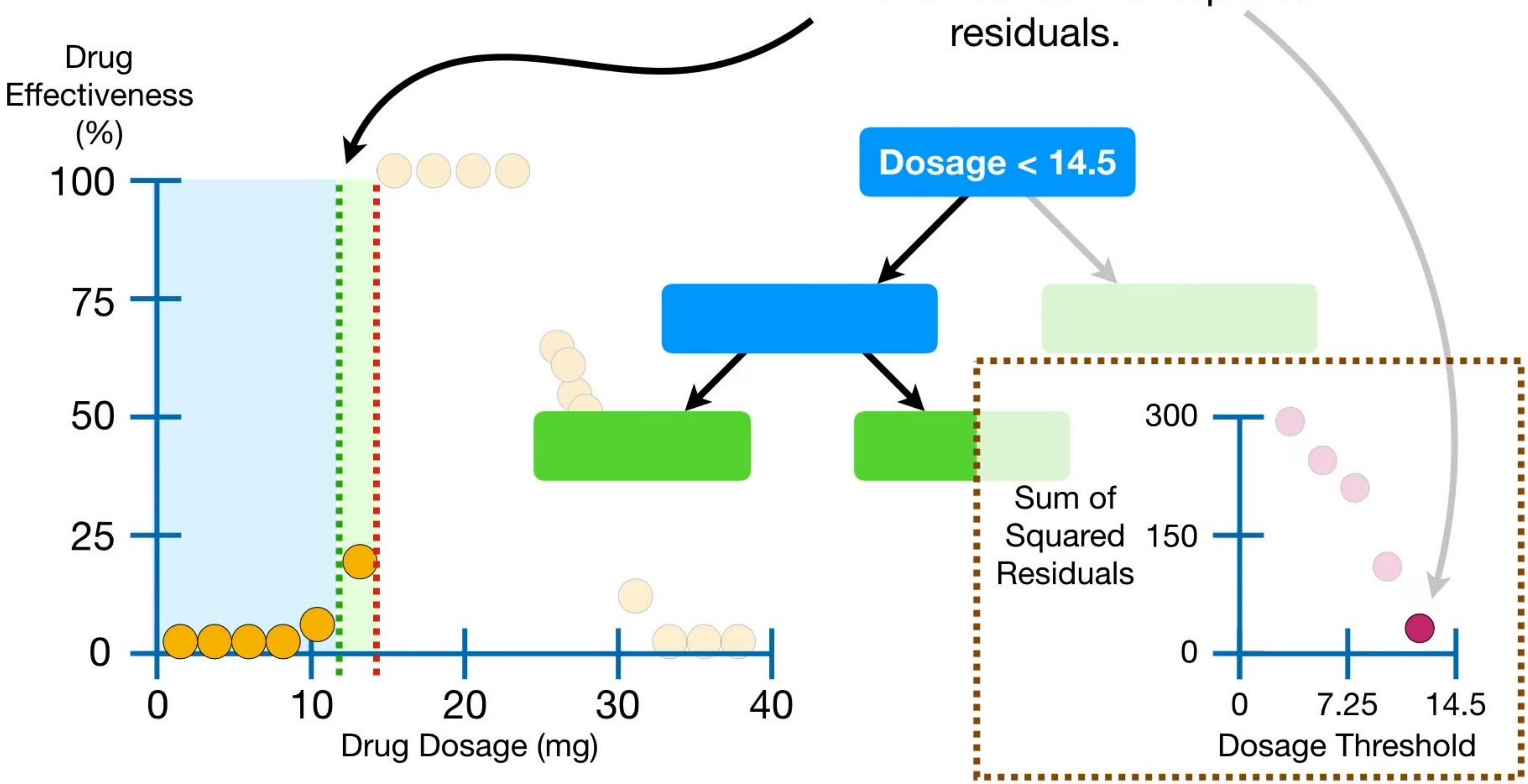
In theory, we could split these **6** observations into two smaller groups just like we did before…



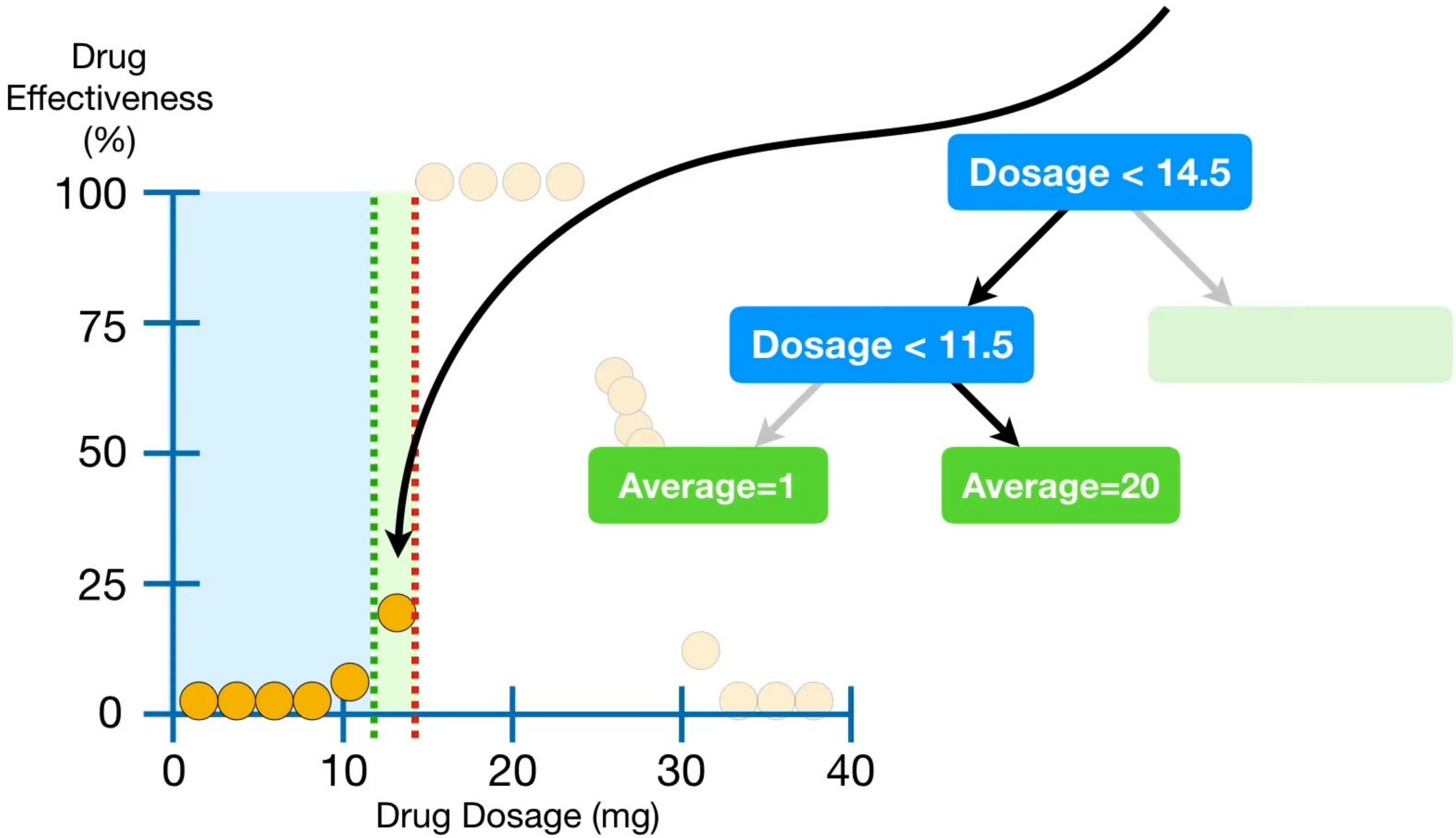
…by calculating the sum of squared residuals for different thresholds…



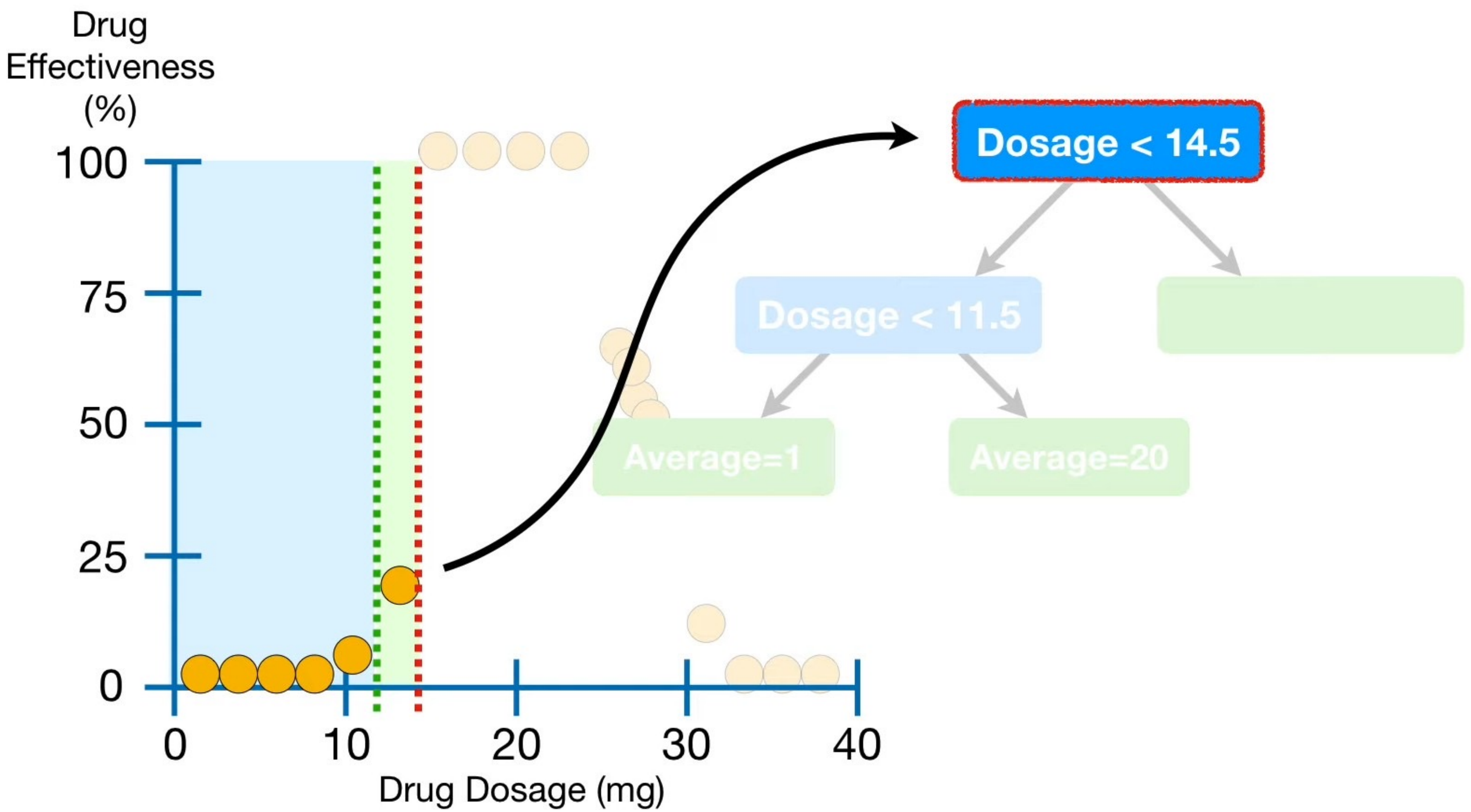
…and choosing the threshold with the lowest sum of squared residuals.



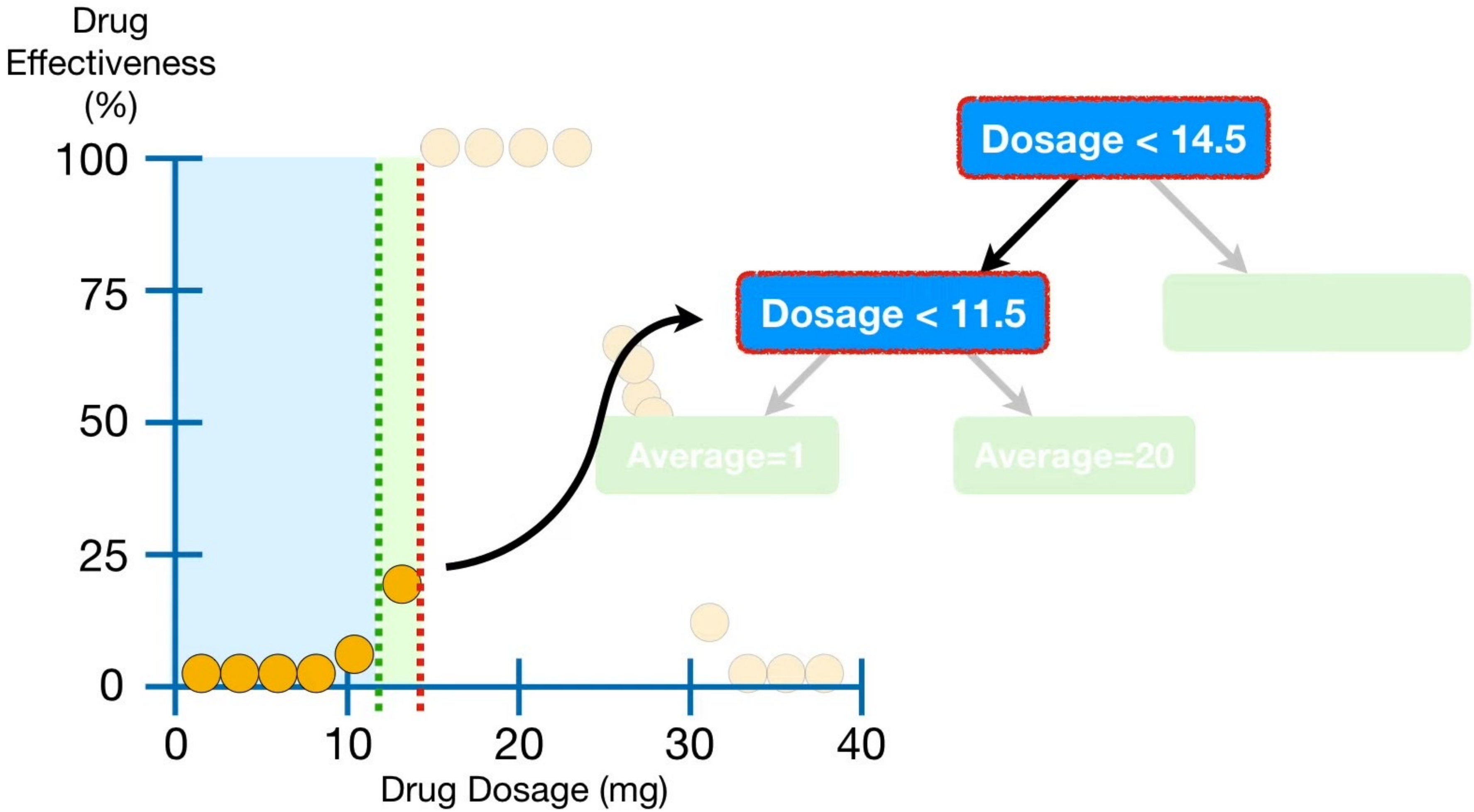
NOTE: This observation…



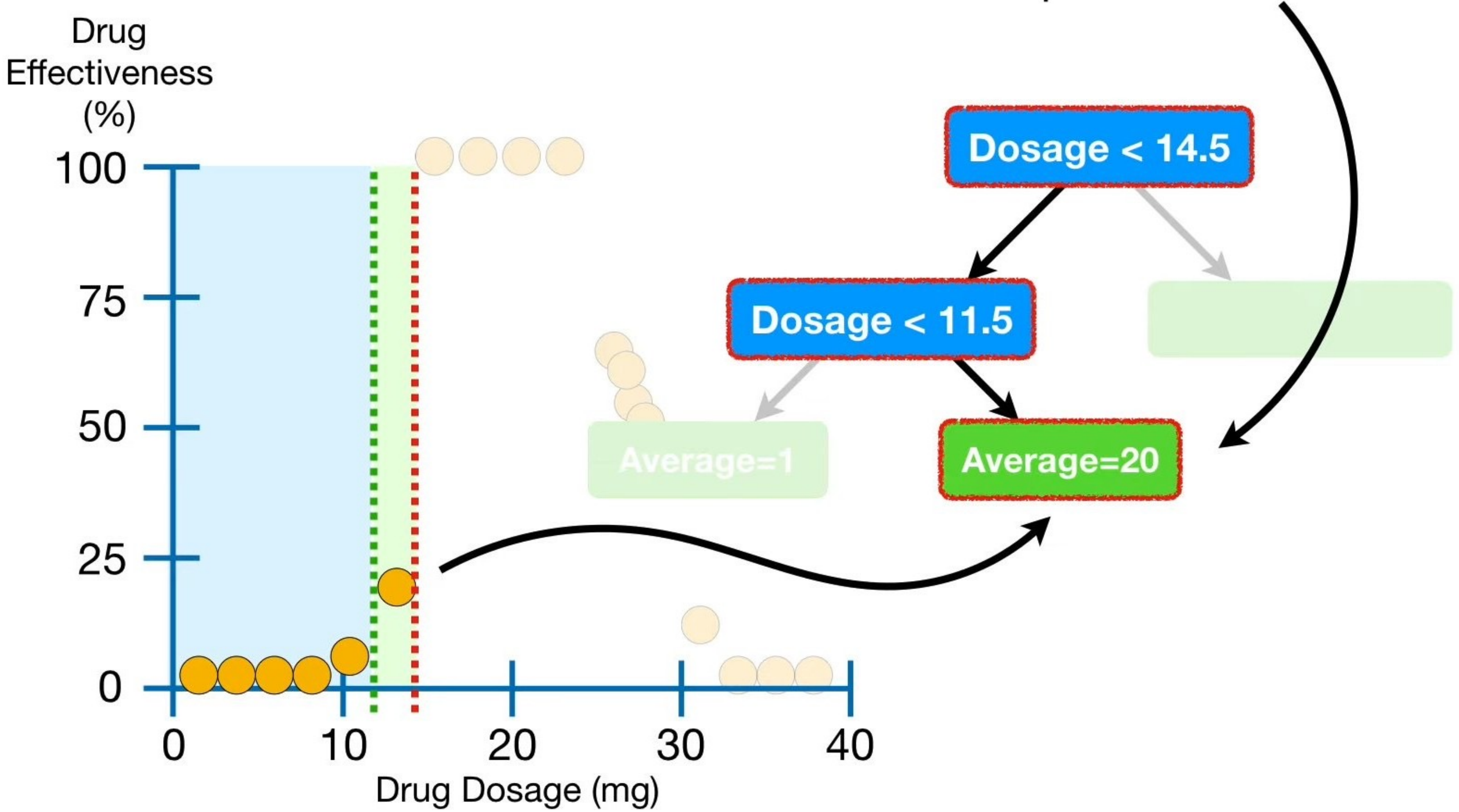
…has **Dosage < 14.5**…



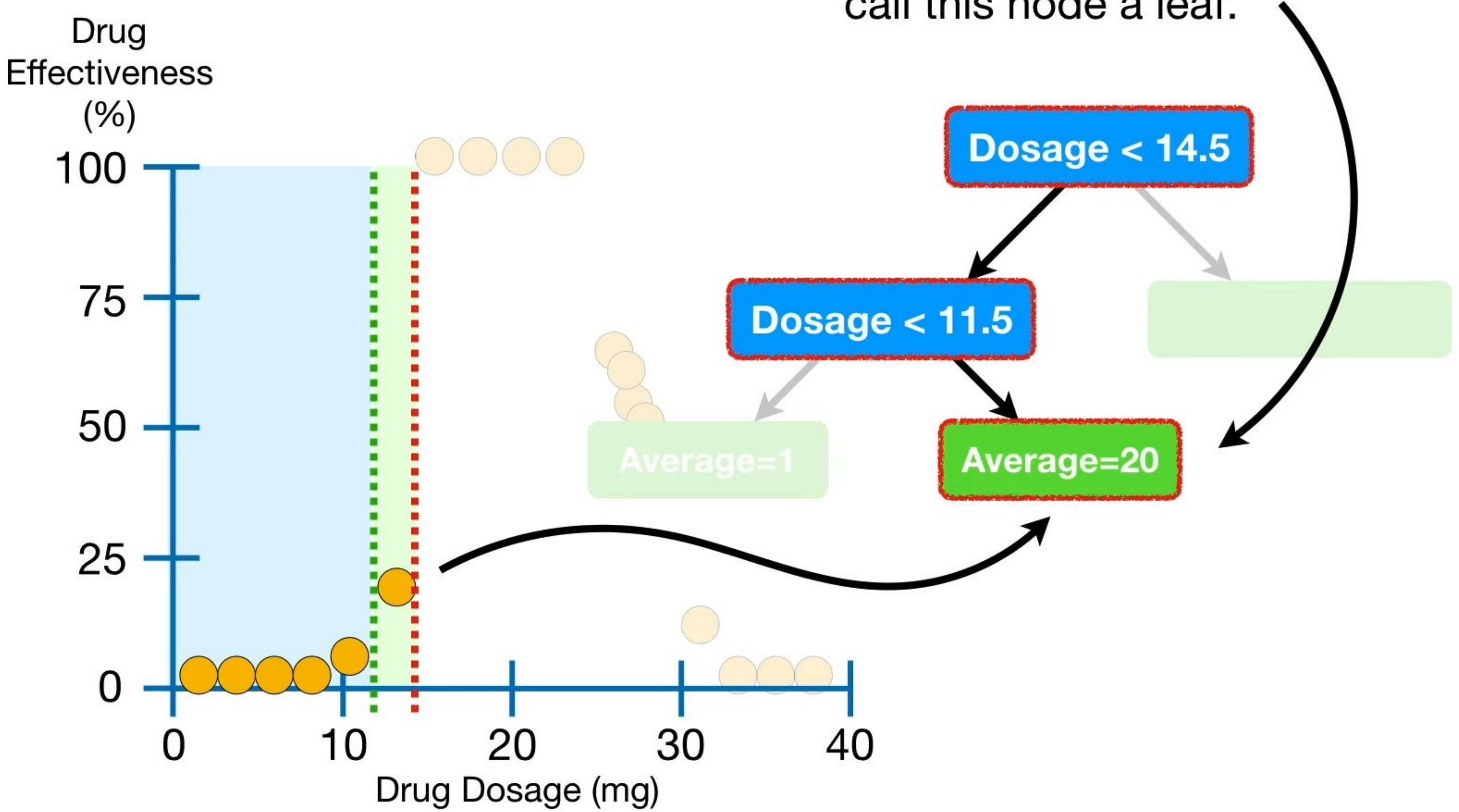
…and does *not* have **Dosage < 11.5…**



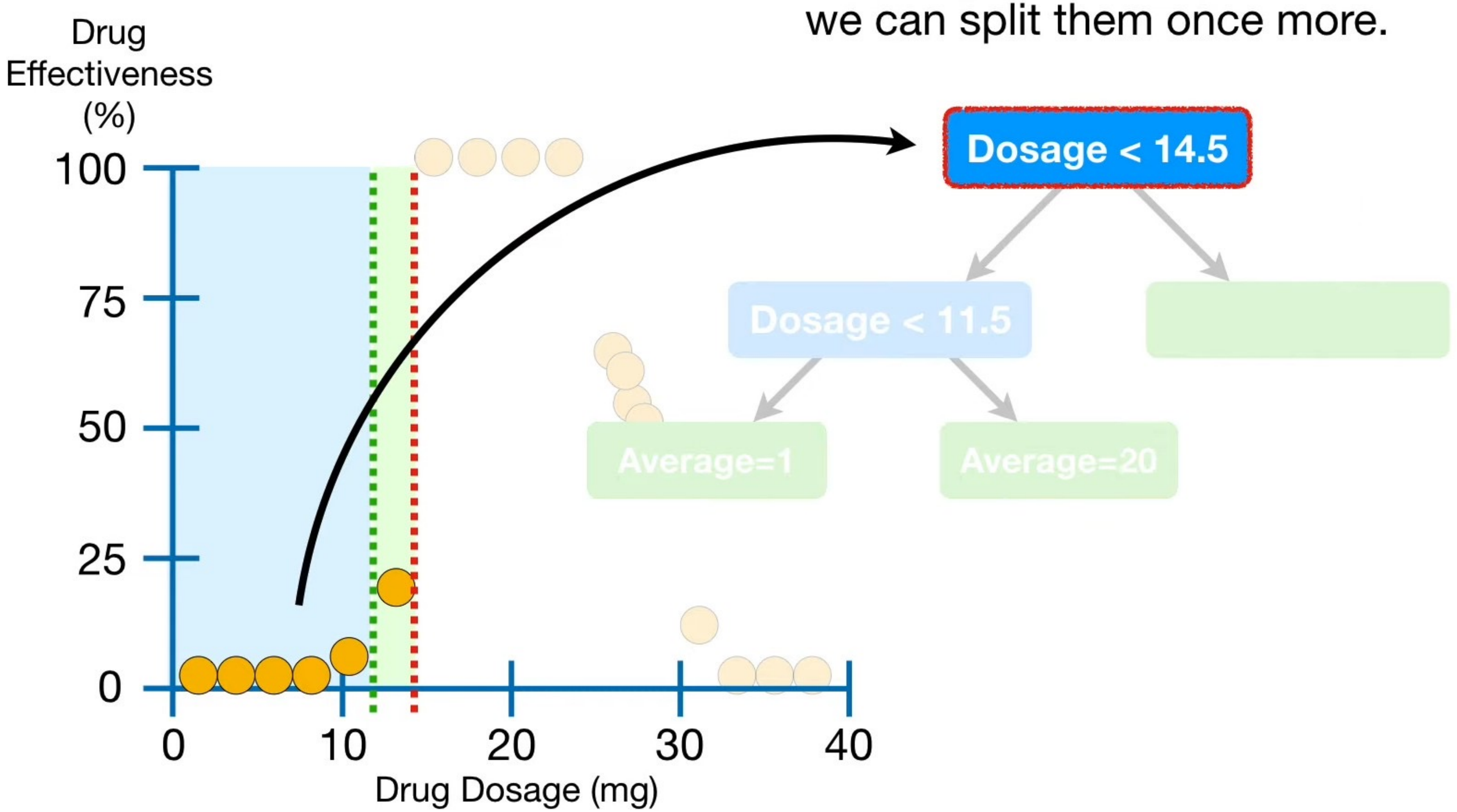
…so it is the only observation to end up in this node…



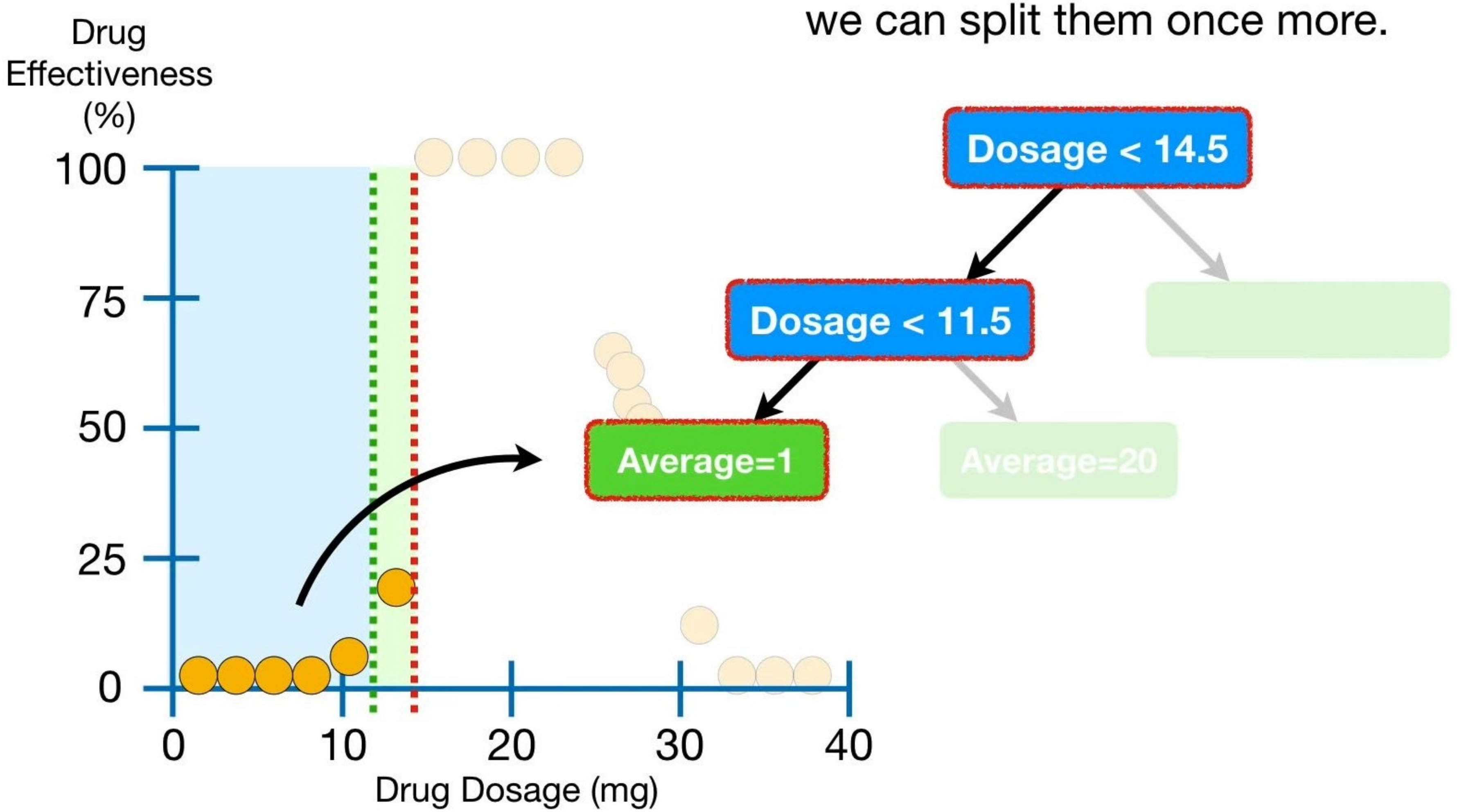
…and since we can't split a single observation into two groups, we will call this node a leaf.



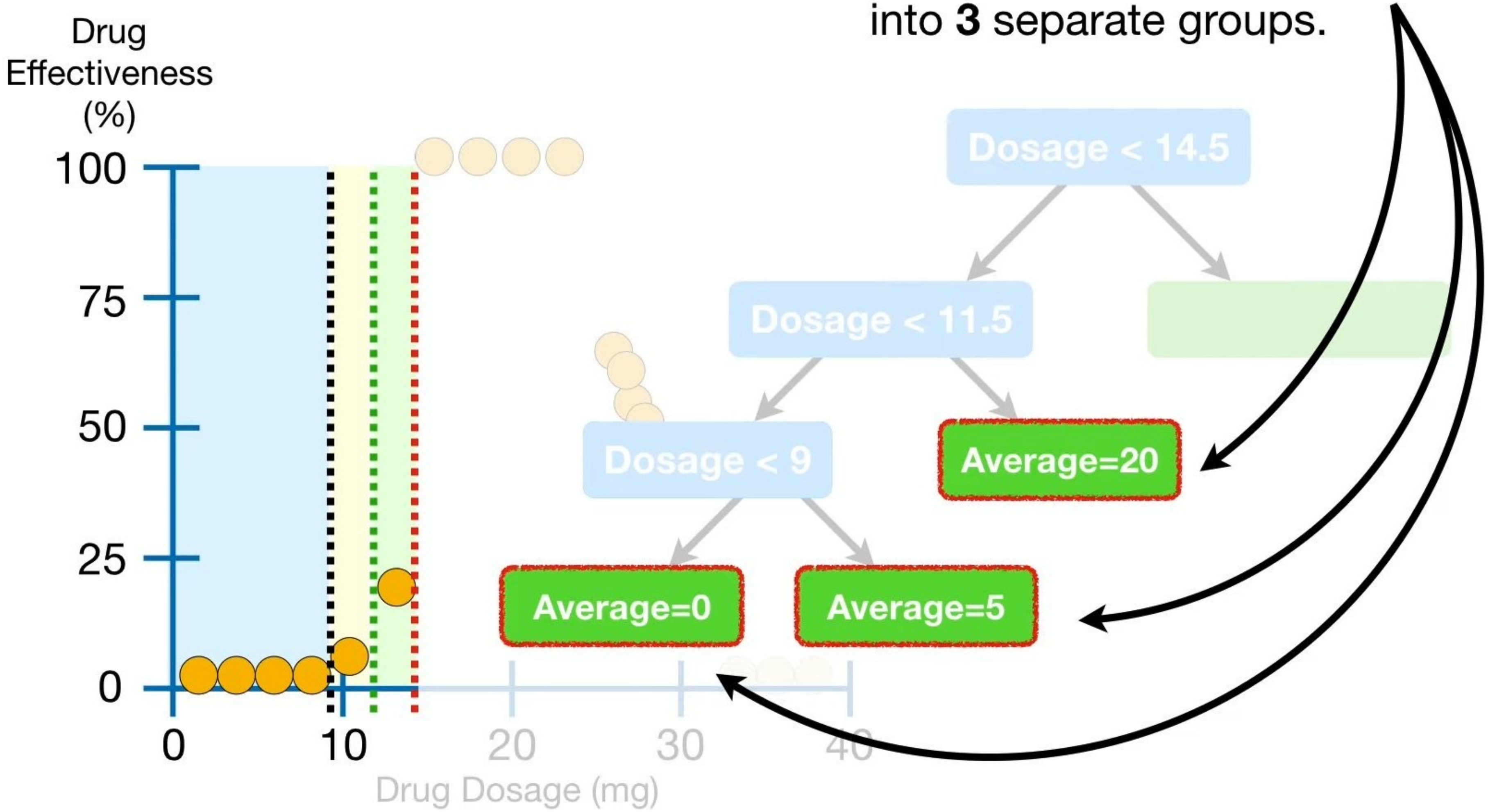
However, since the remaining **5** observations go to the other node, we can split them once more.



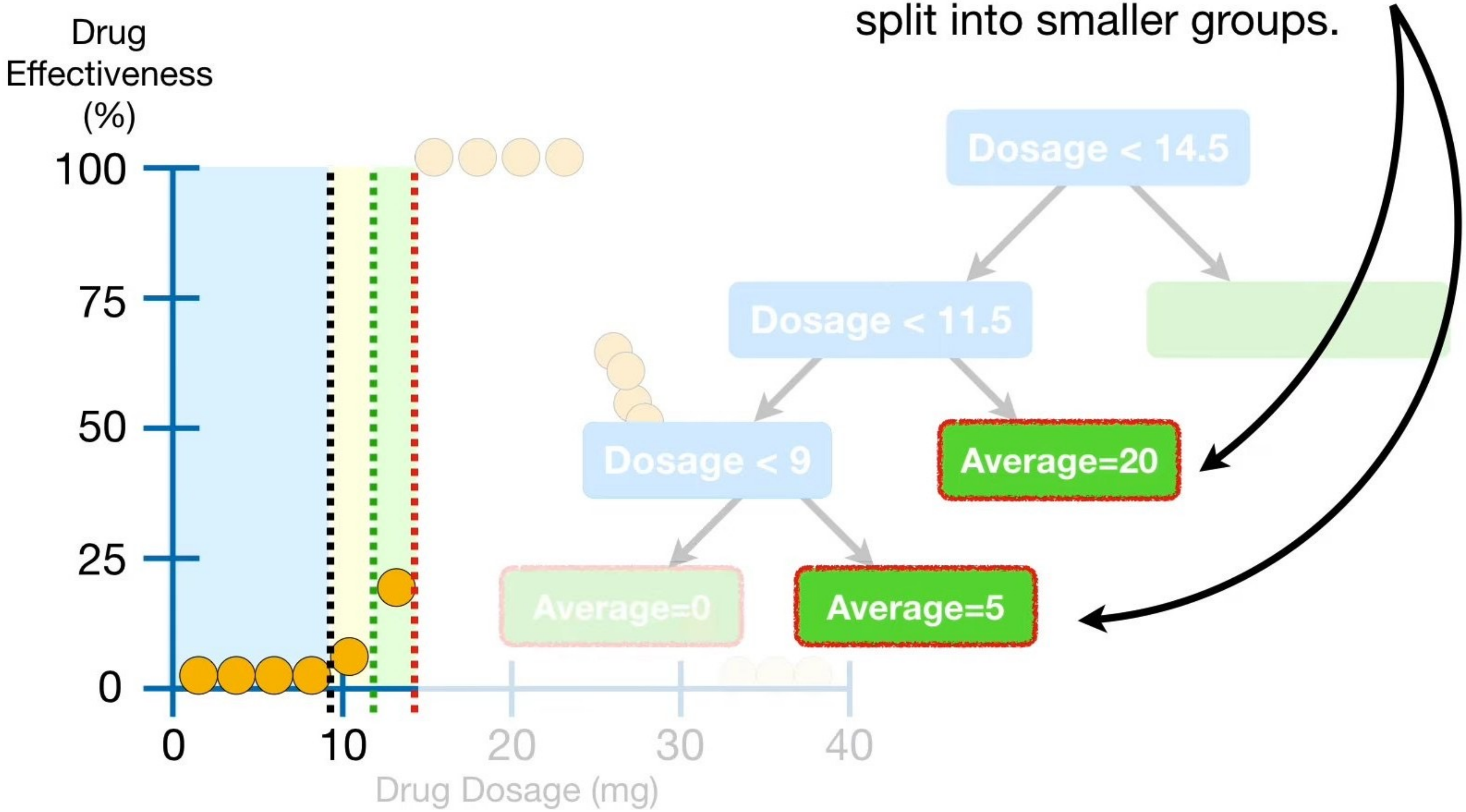
However, since the remaining **5** observations go to the other node, we can split them once more.



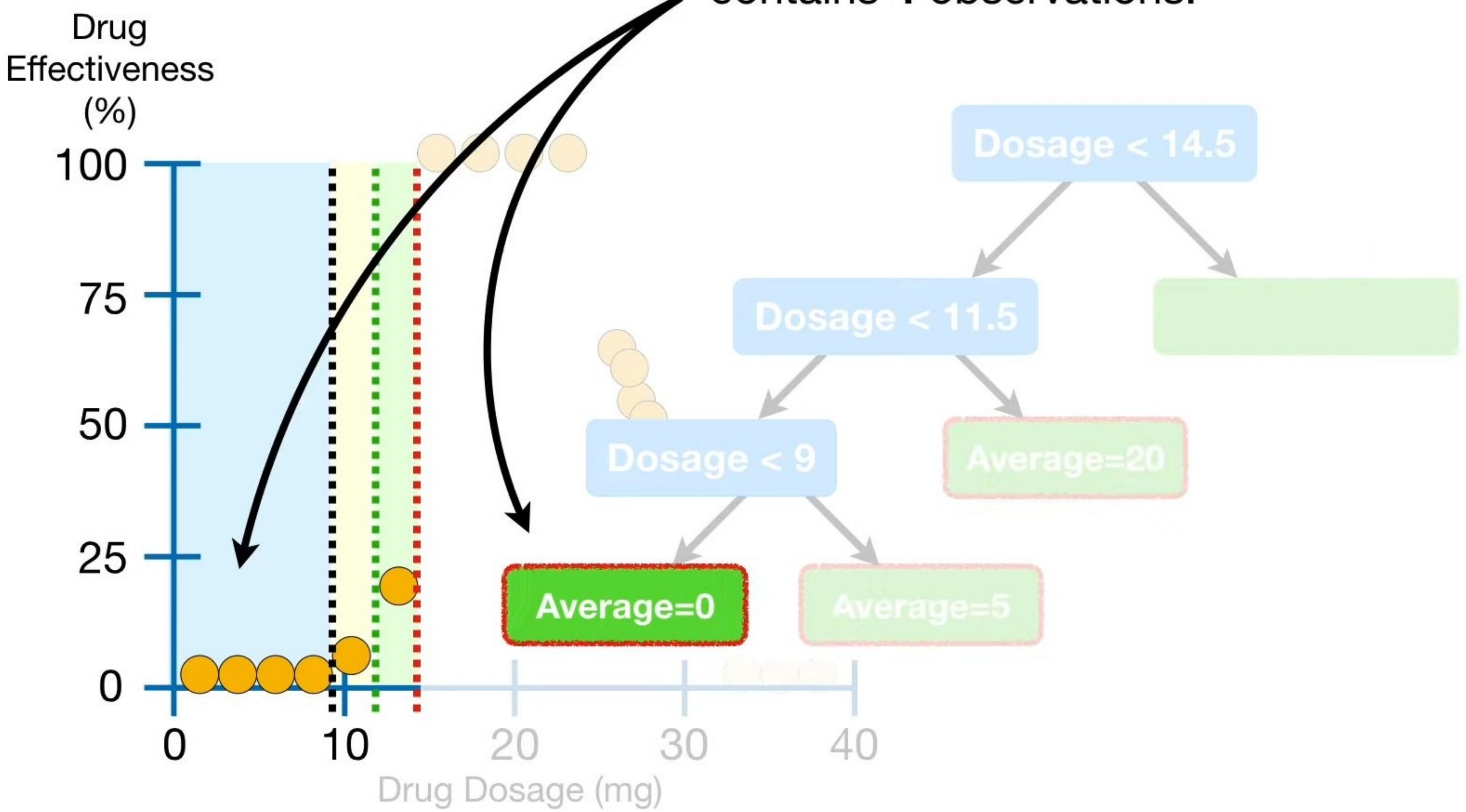
Now we have divided the observations with **Dosage < 14.5** into **3** separate groups.



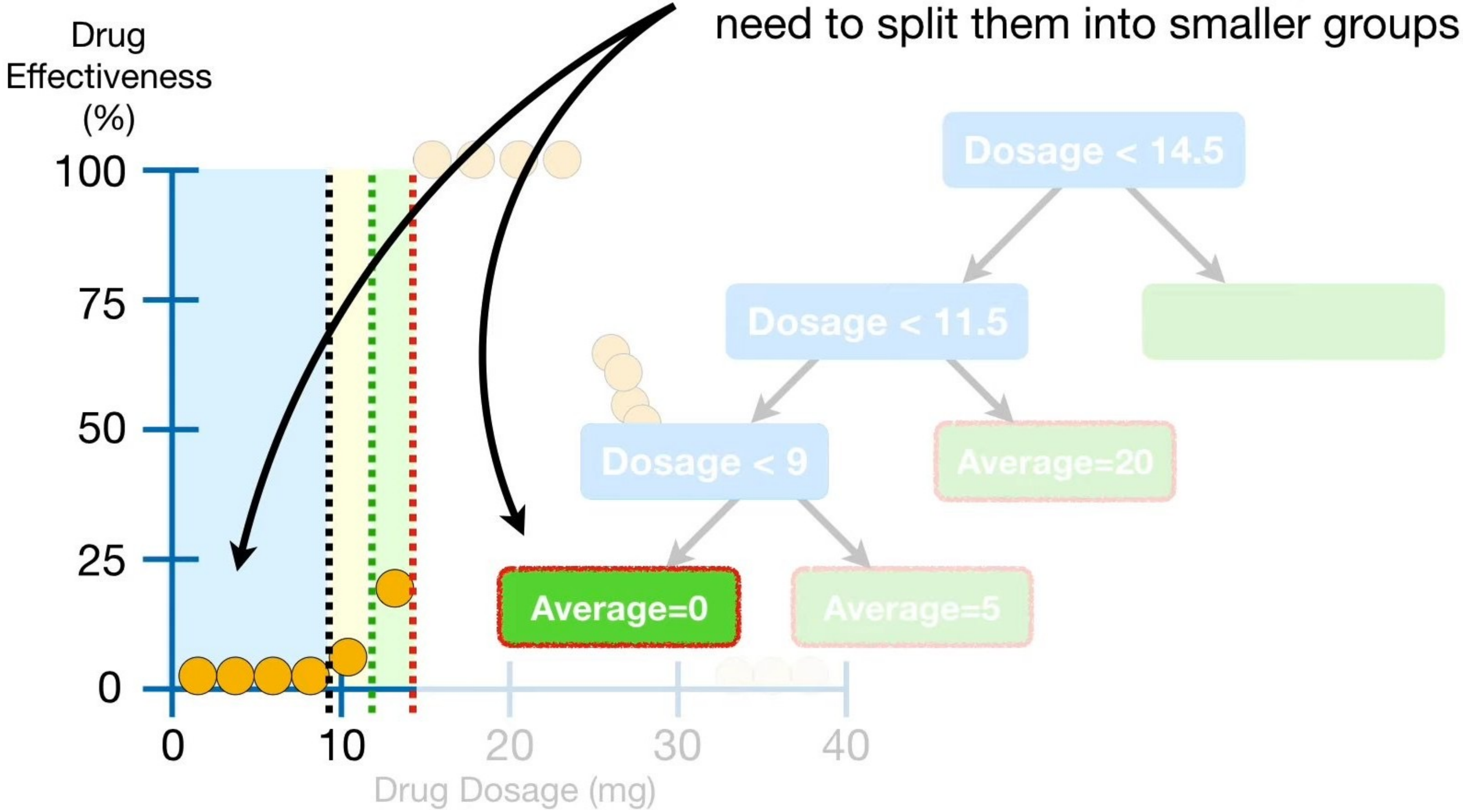
These two leaves only contain one observation each and can not be split into smaller groups.



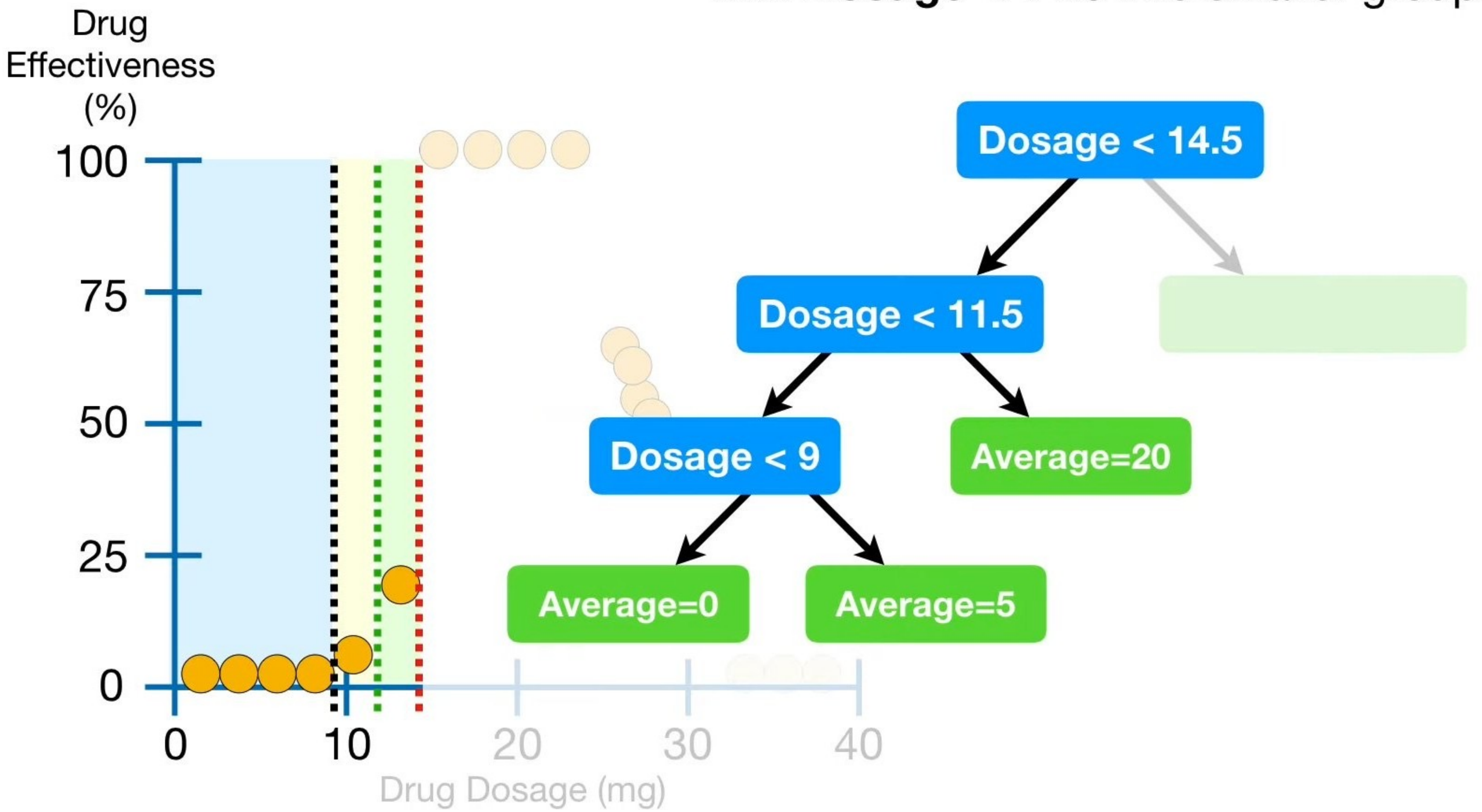
In contrast, this leaf contains **4** observations.



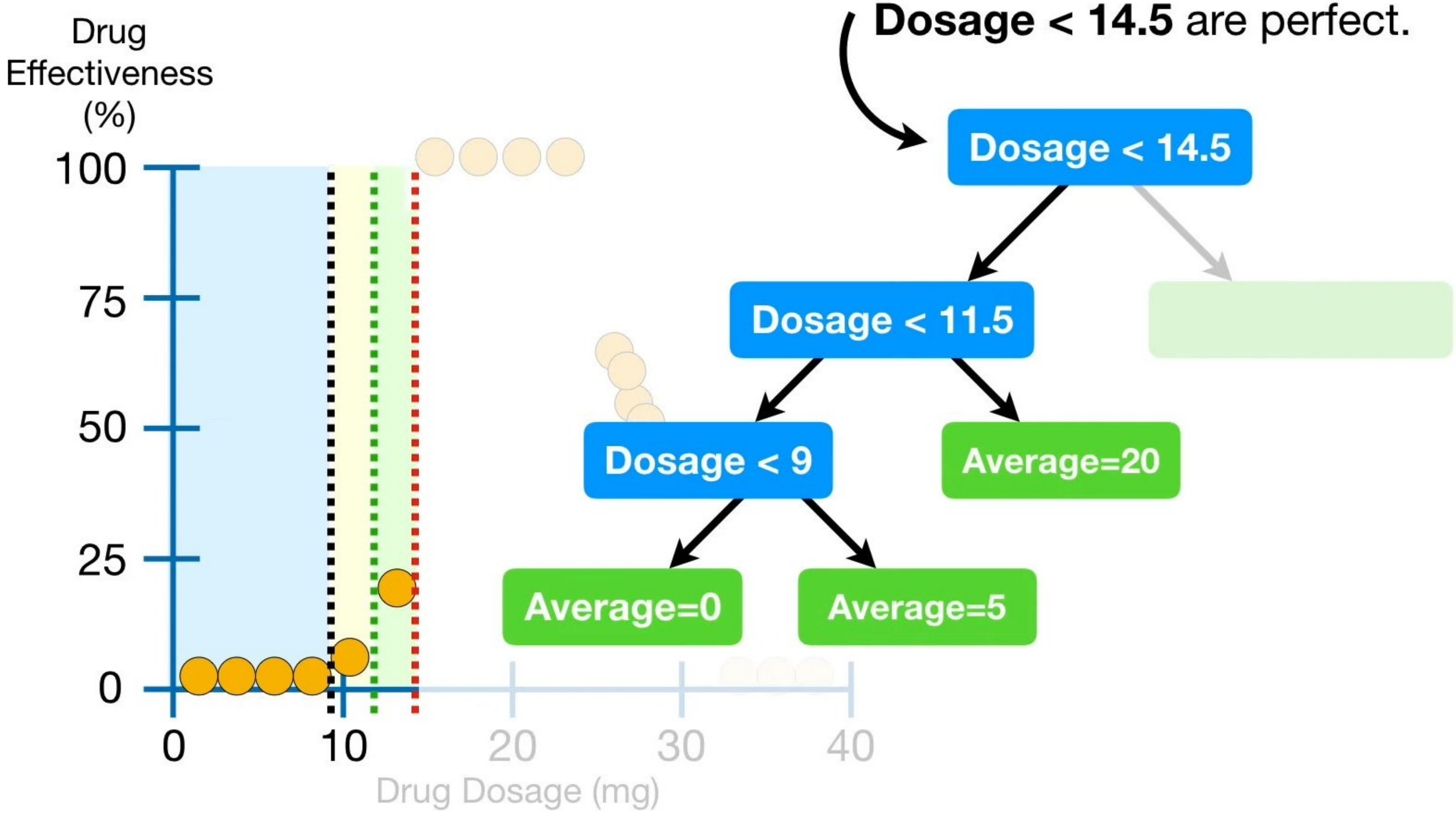
That said, those **4** observations all have the same **Drug Effectiveness**, so we don't need to split them into smaller groups.



So we are done splitting the observations with **Dosage < 14.5** into smaller groups.



NOTE: The *predictions* that this tree makes for all observations with **Dosage < 14.5** are perfect.



Dosage < 14.5

Dosage < 11.5

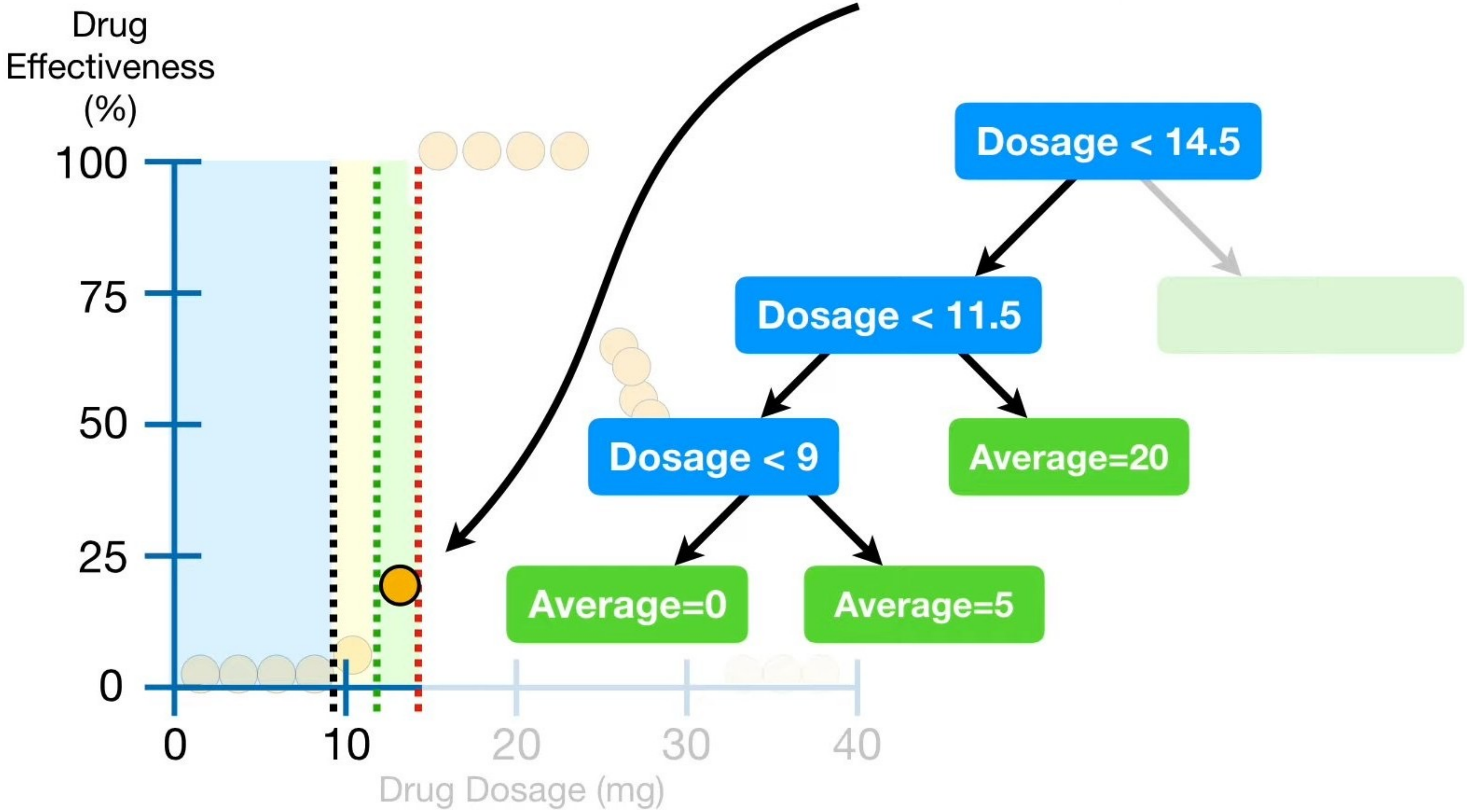
Dosage < 9

Average=20

Average=0

Average=5

In other words, this observation has **20% Drug Effectiveness**…



Dosage < 14.5

Dosage < 11.5

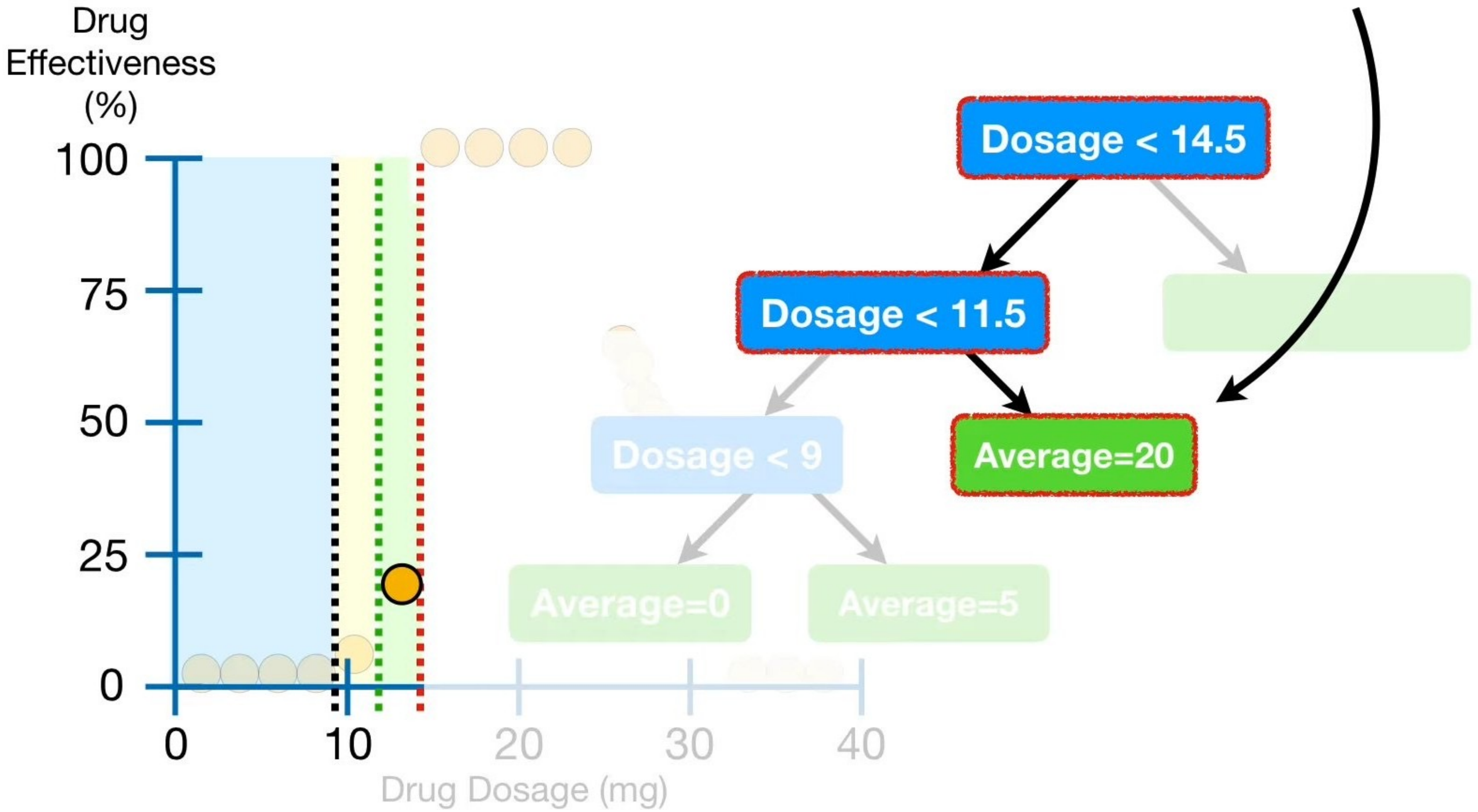
Dosage < 9

Average=20

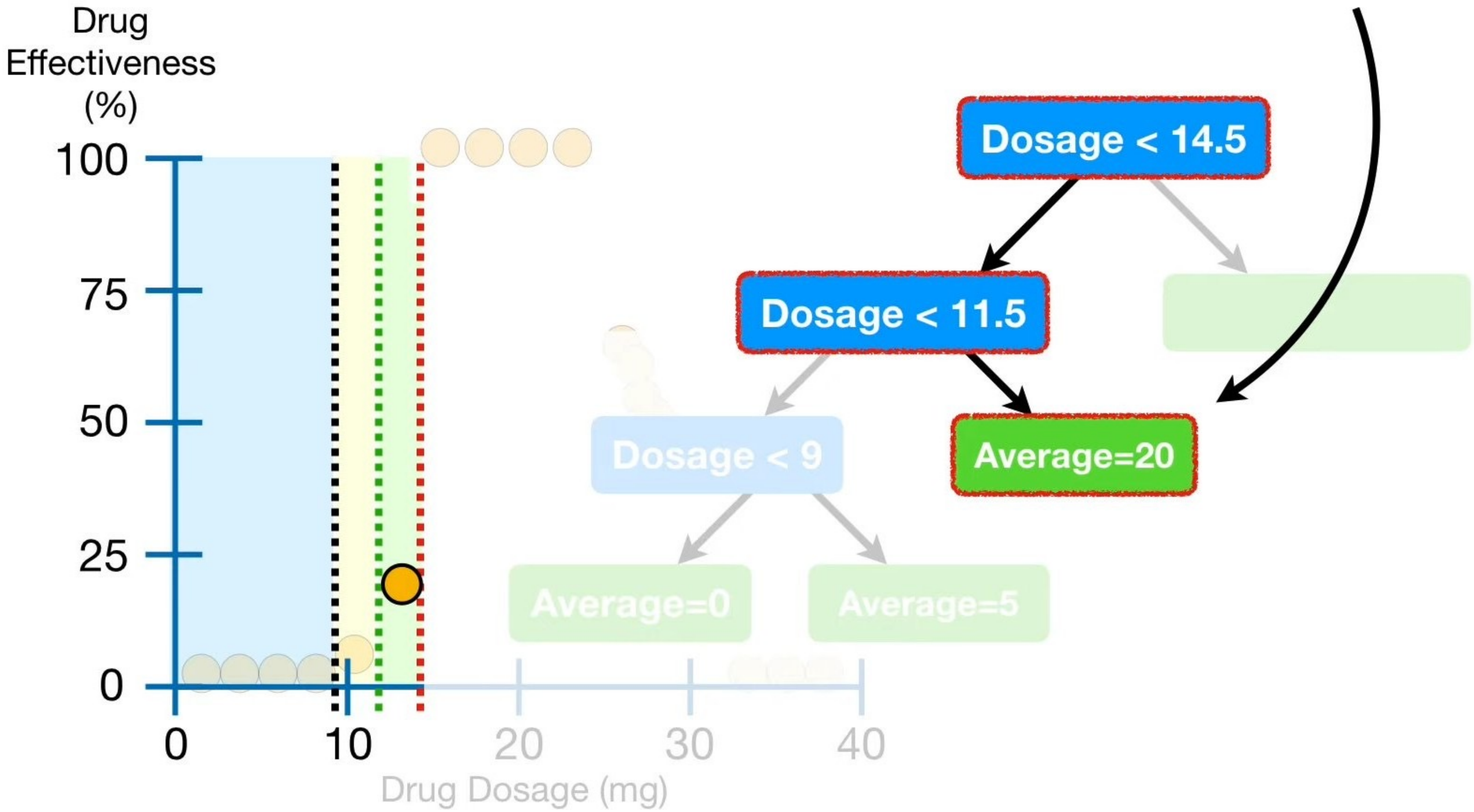
Average=0

Average=5

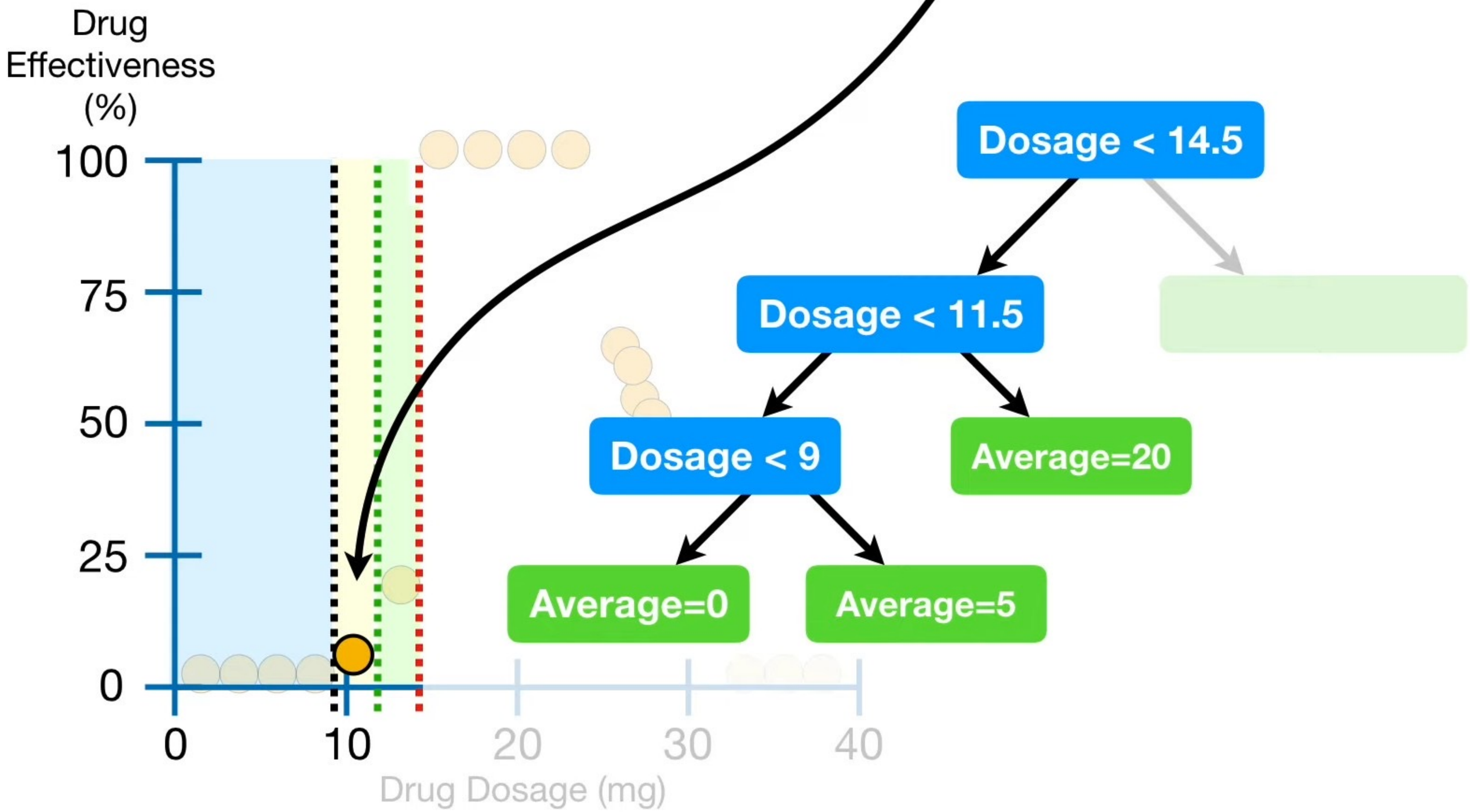
…and the tree predicts **20% Drug Effectiveness**…



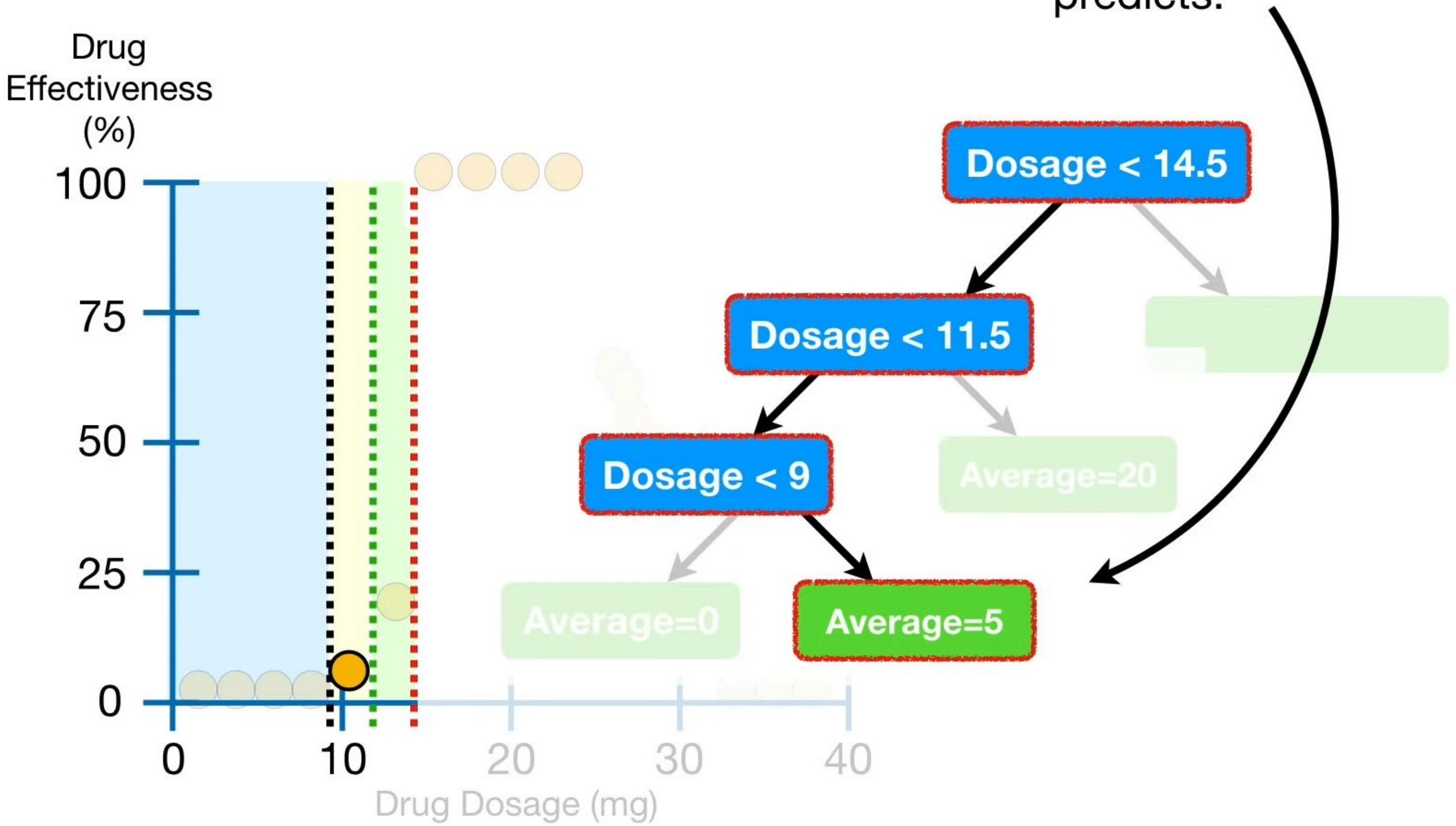
…so the *observed* and *predicted* values are the same.



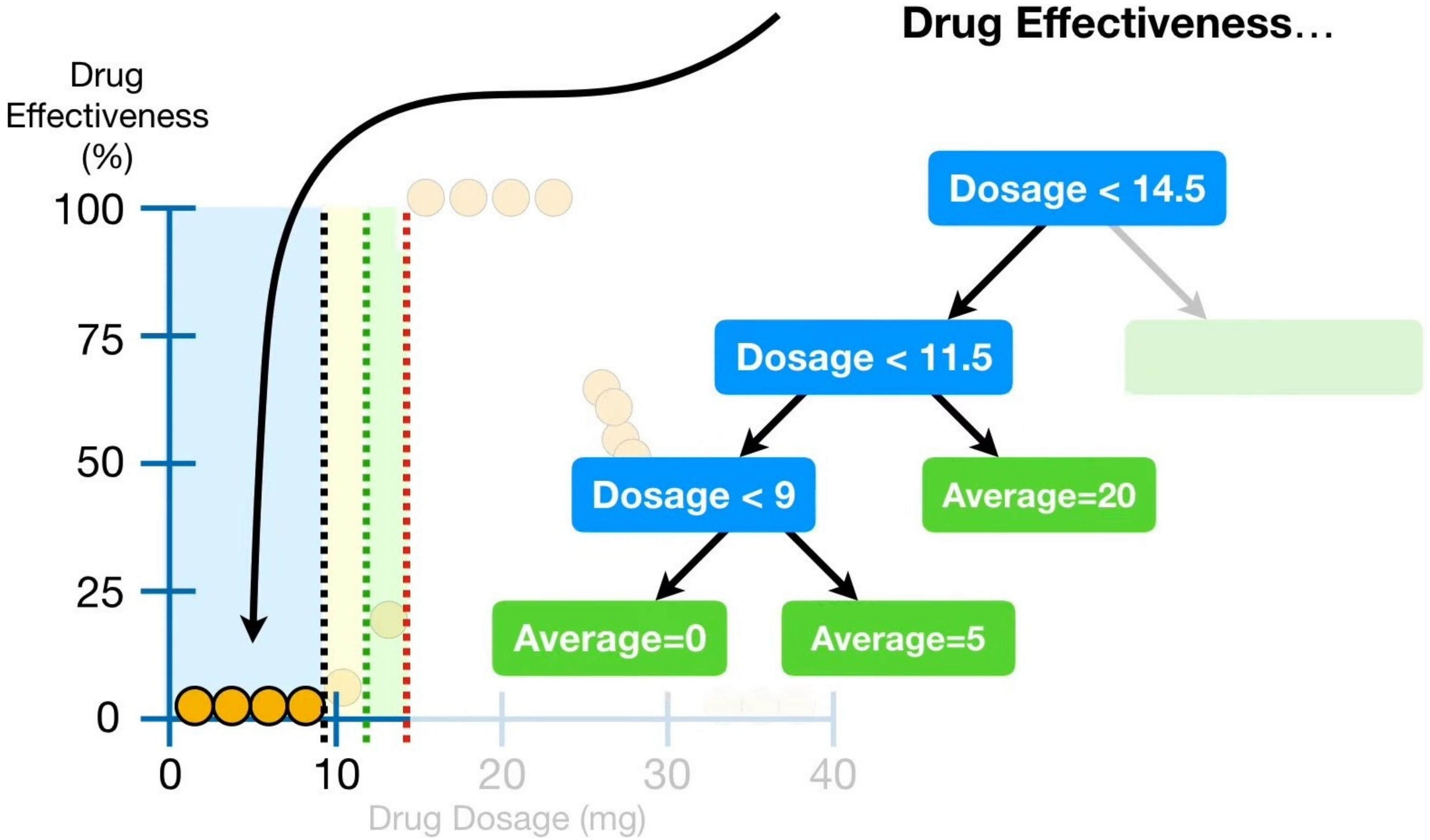
This observation has **5% Drug Effectiveness**…



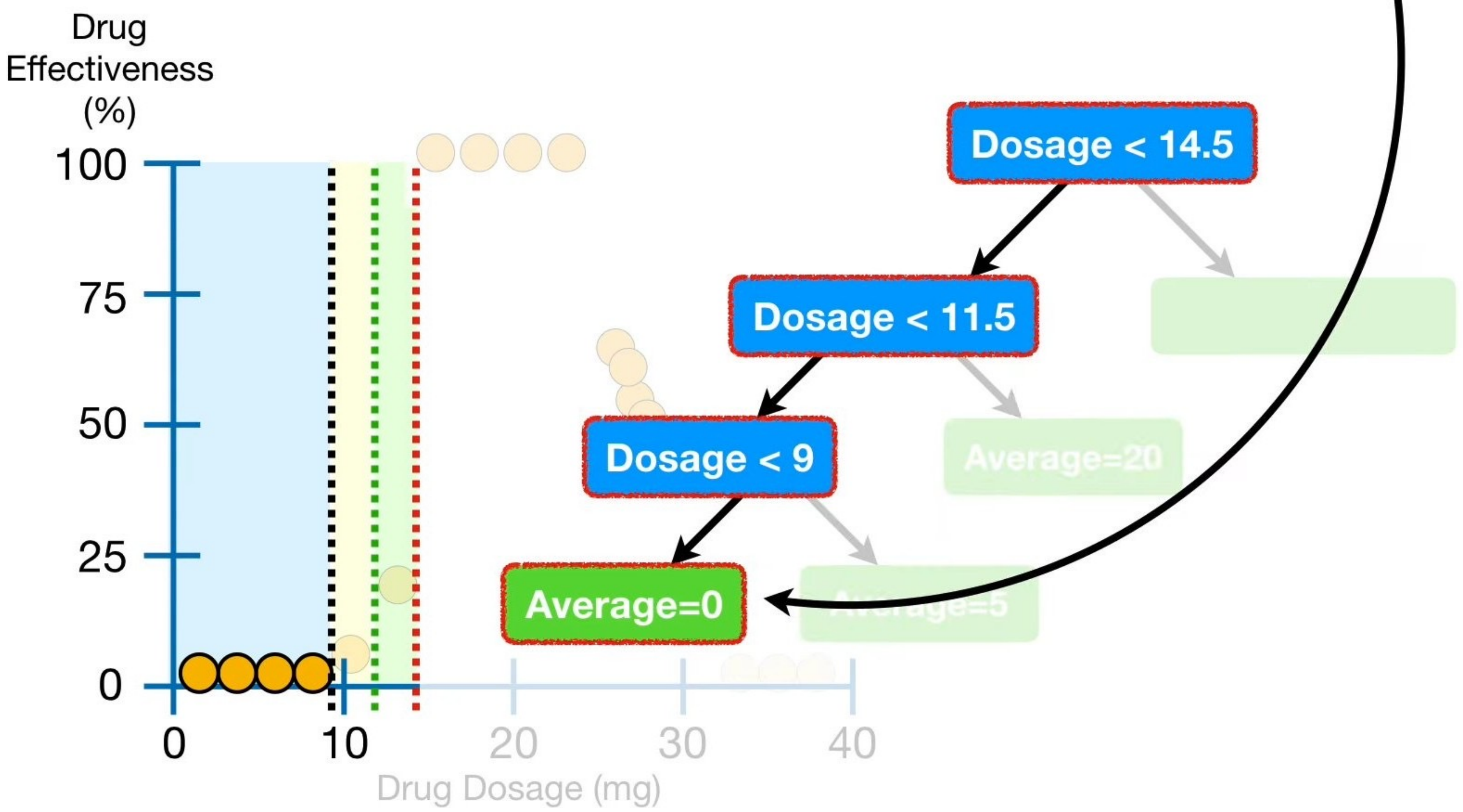
…and that's exactly what the tree predicts.



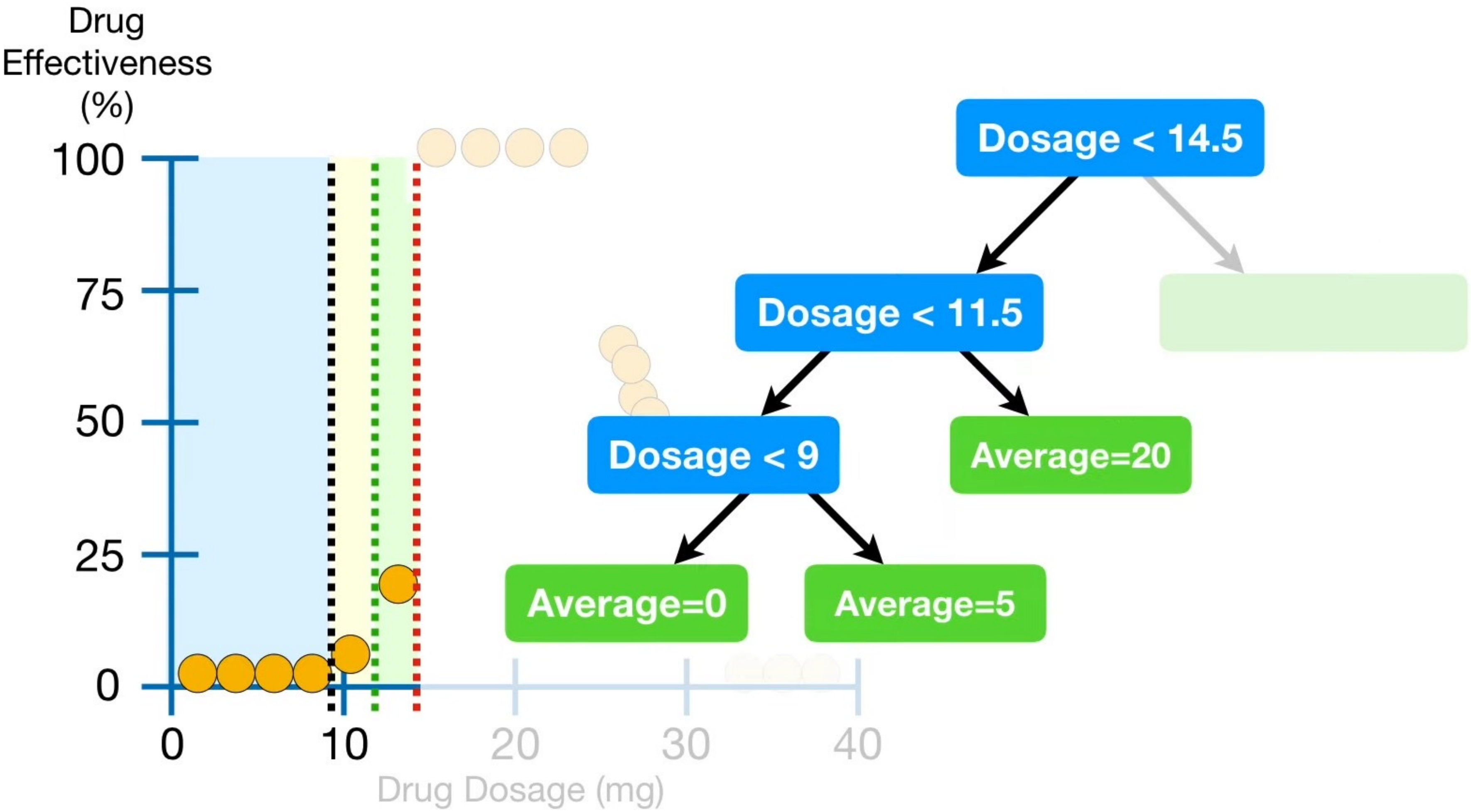
These **4** observations all have **0% Drug Effectiveness**…



…and that’s exactly what the tree predicts.



Is that awesome?



Dosage < 14.5

Dosage < 11.5

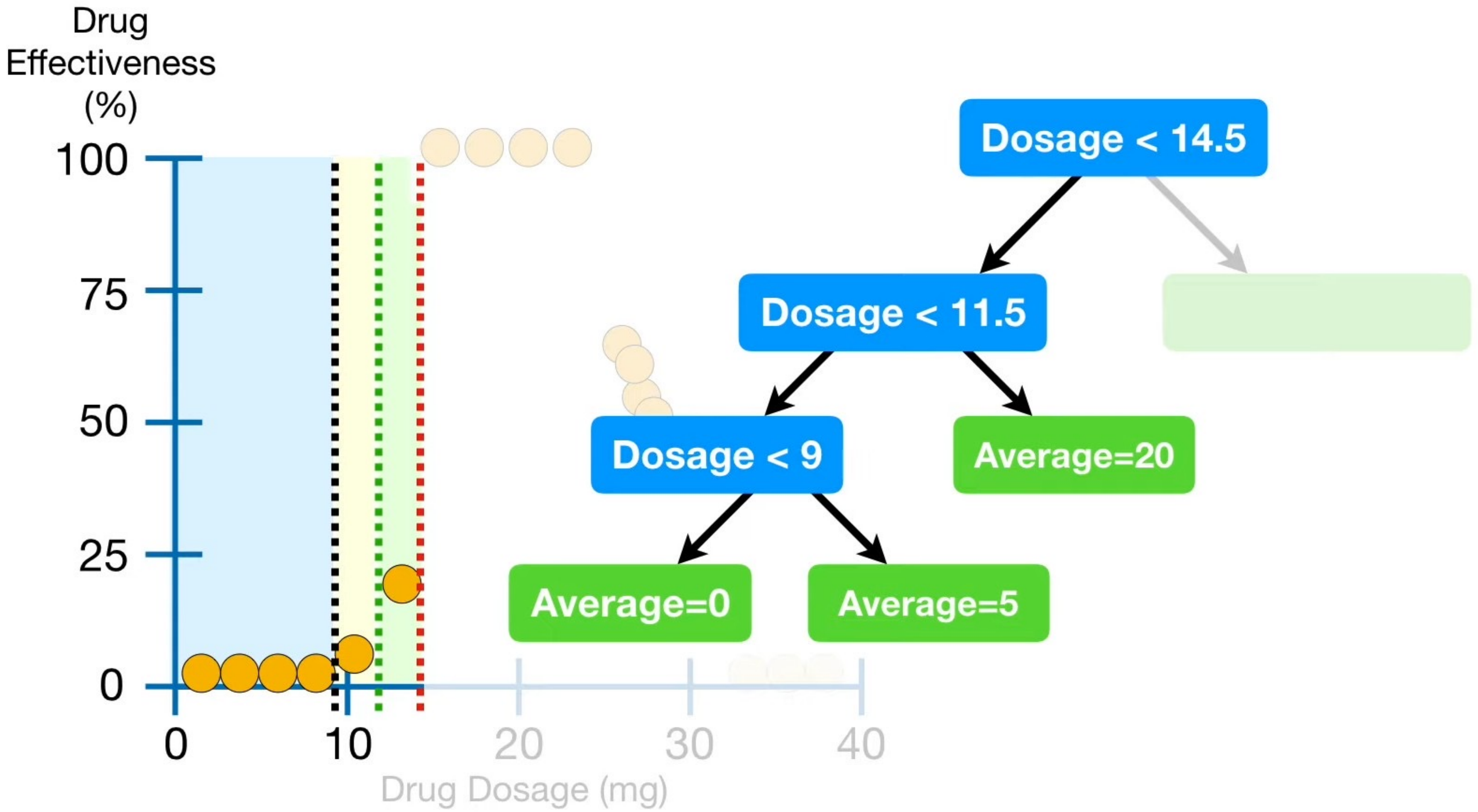
Dosage < 9

Average=20

Average=0

Average=5

No.



Dosage < 14.5

Dosage < 11.5

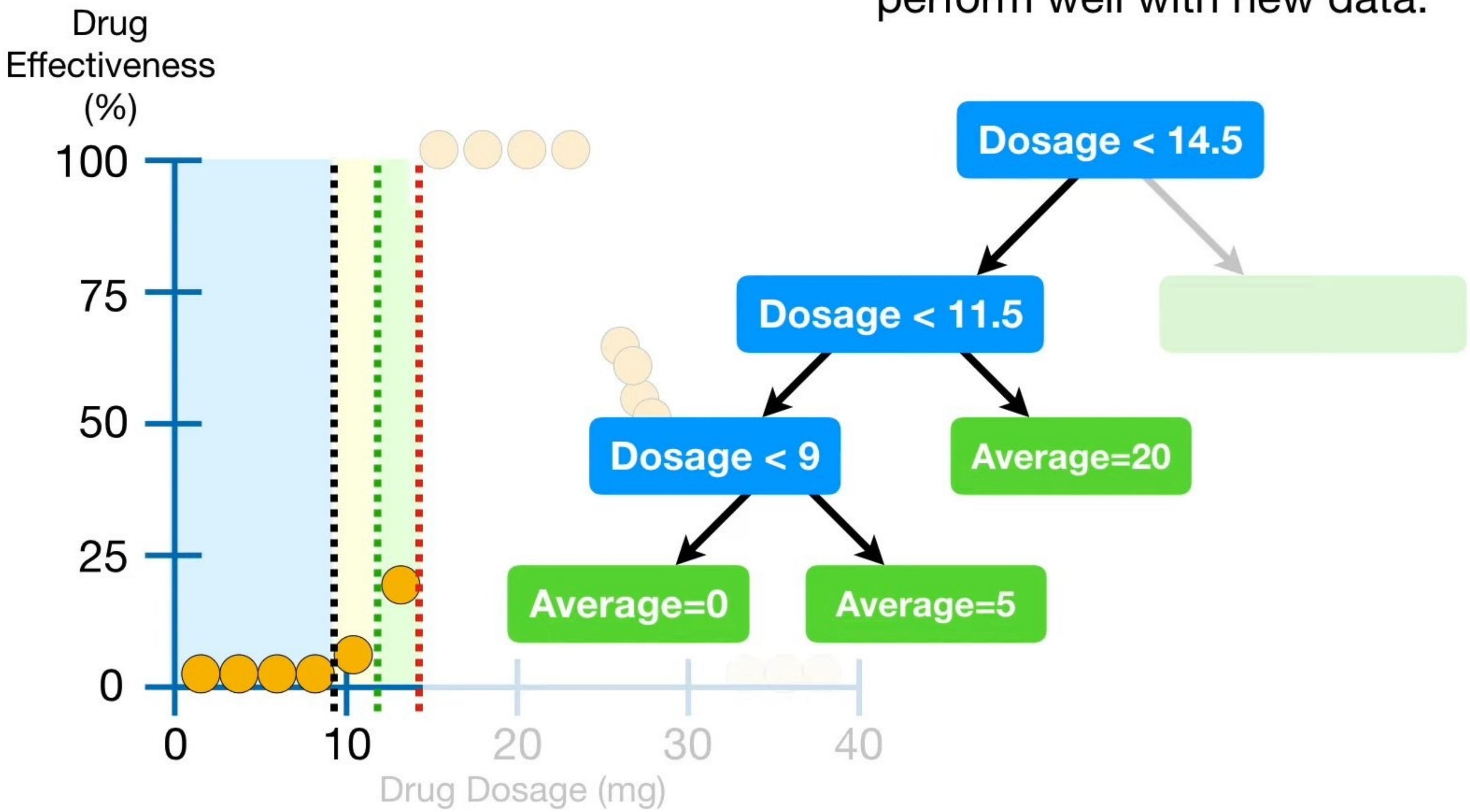
Dosage < 9

Average=20

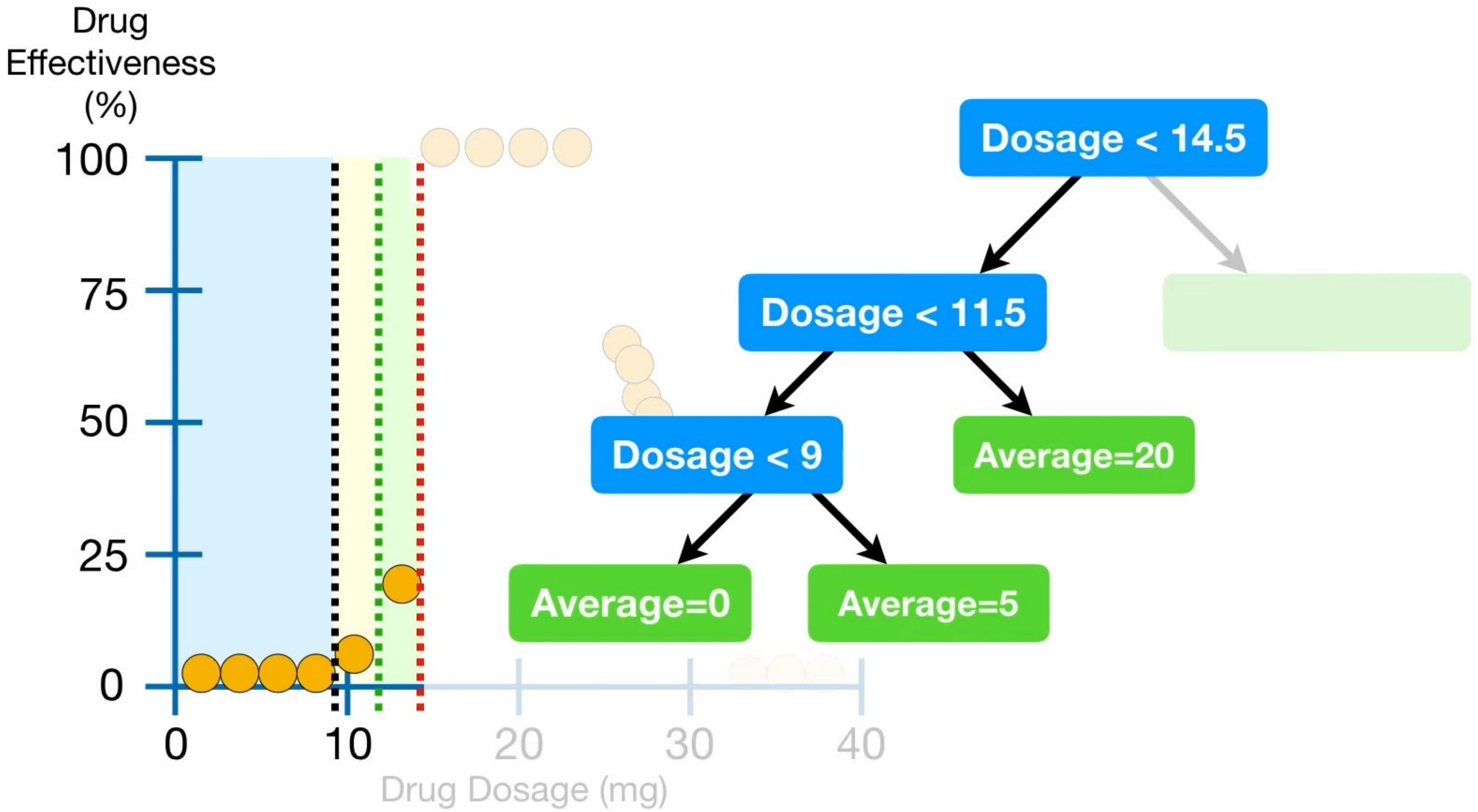
Average=0

Average=5

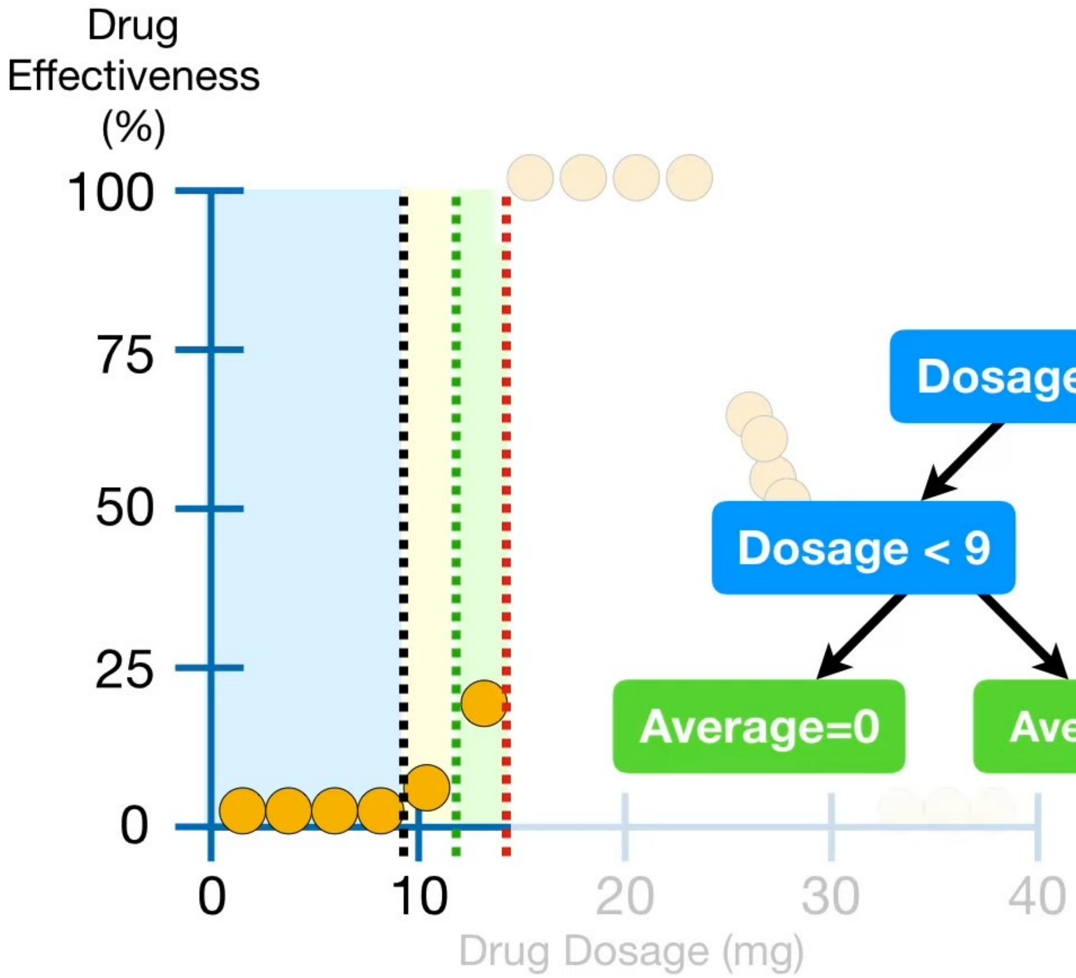
When a model fits the training data perfectly, it probably means it is overfit and will not perform well with new data.

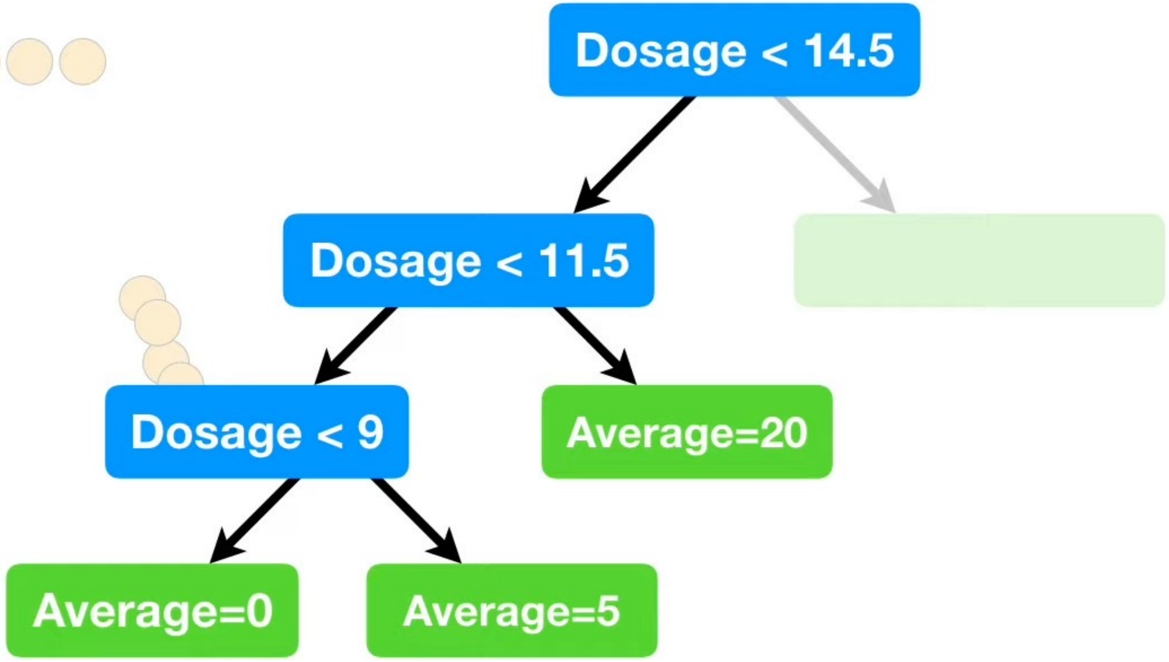


In **Machine Learning Lingo**, the model has no ***Bias***, but potentially large ***Variance***.

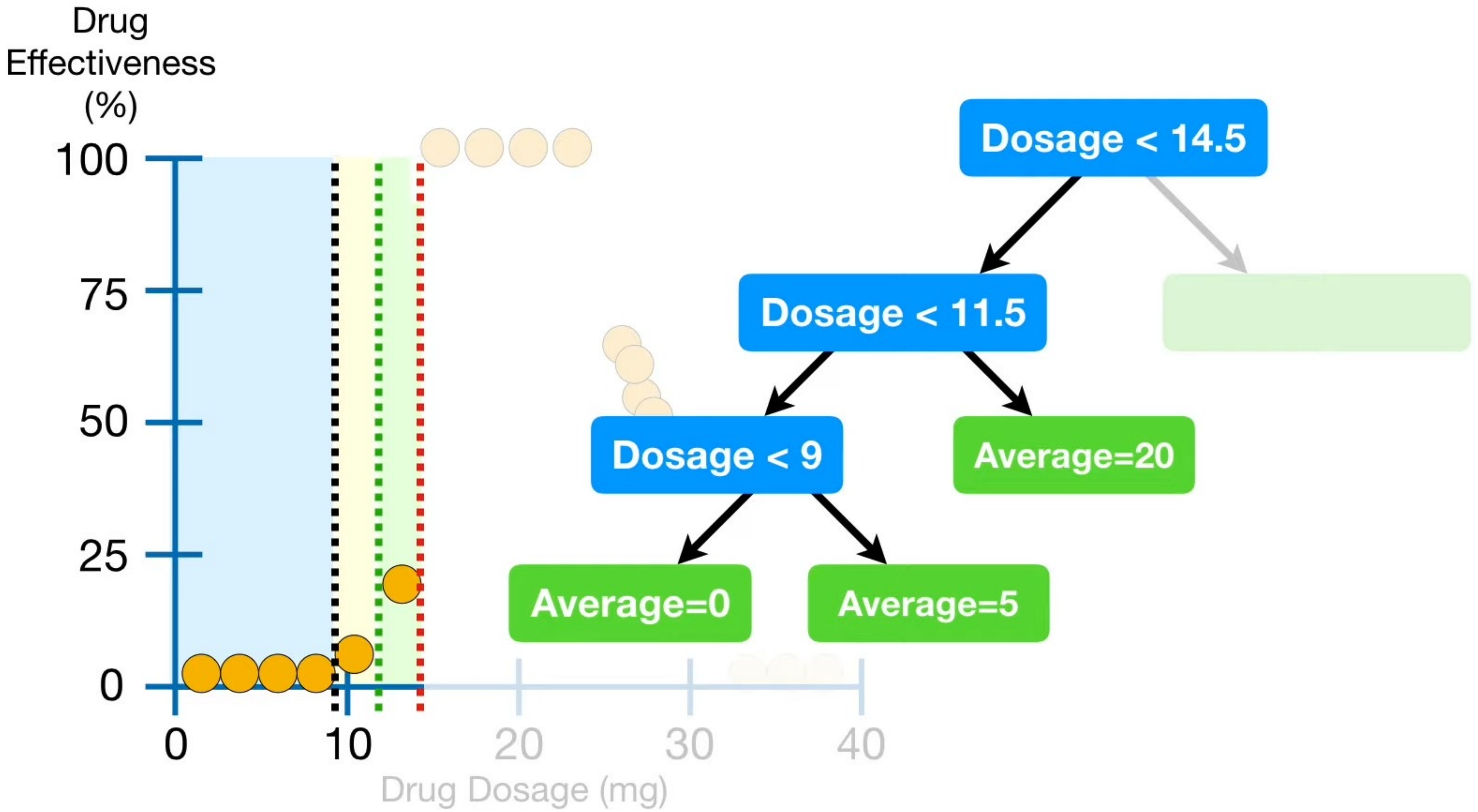


Is there a way to prevent our tree from overfitting the training data?

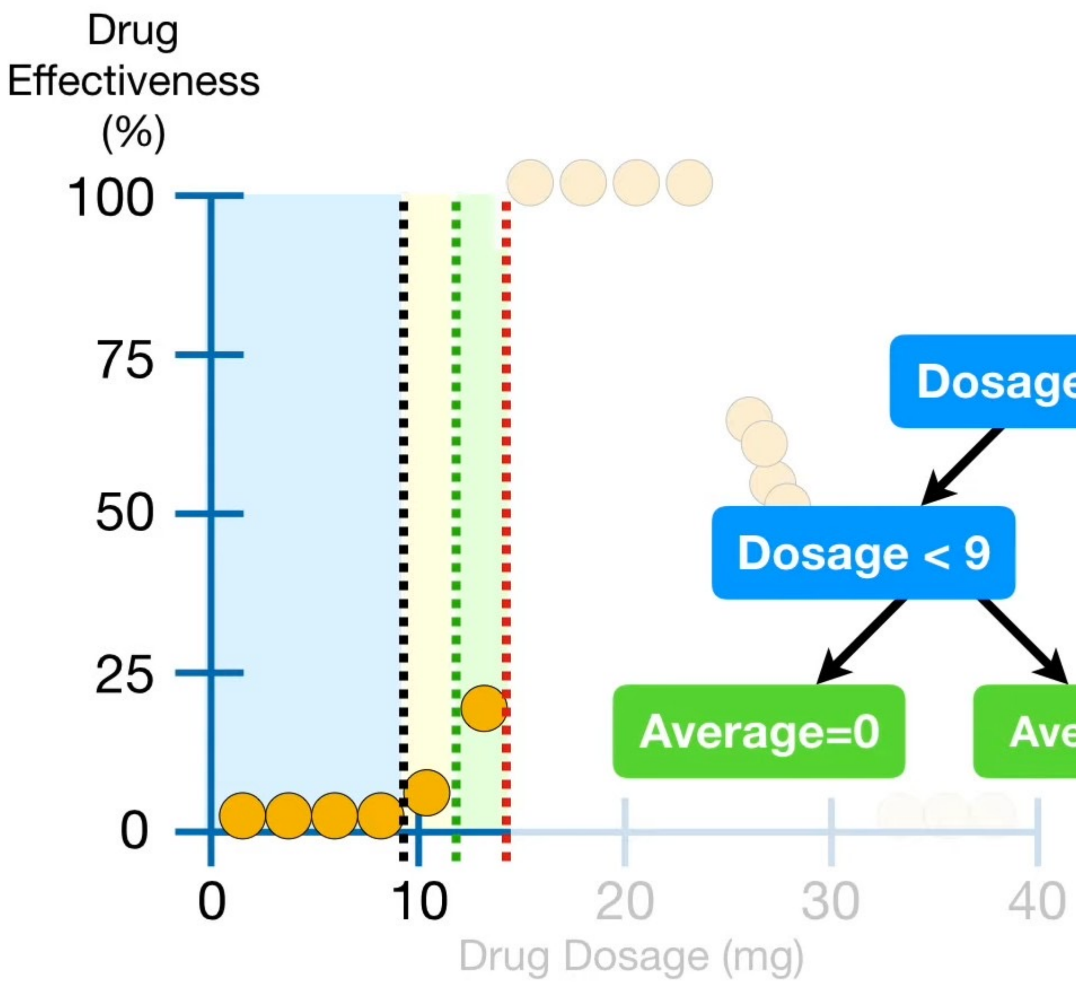


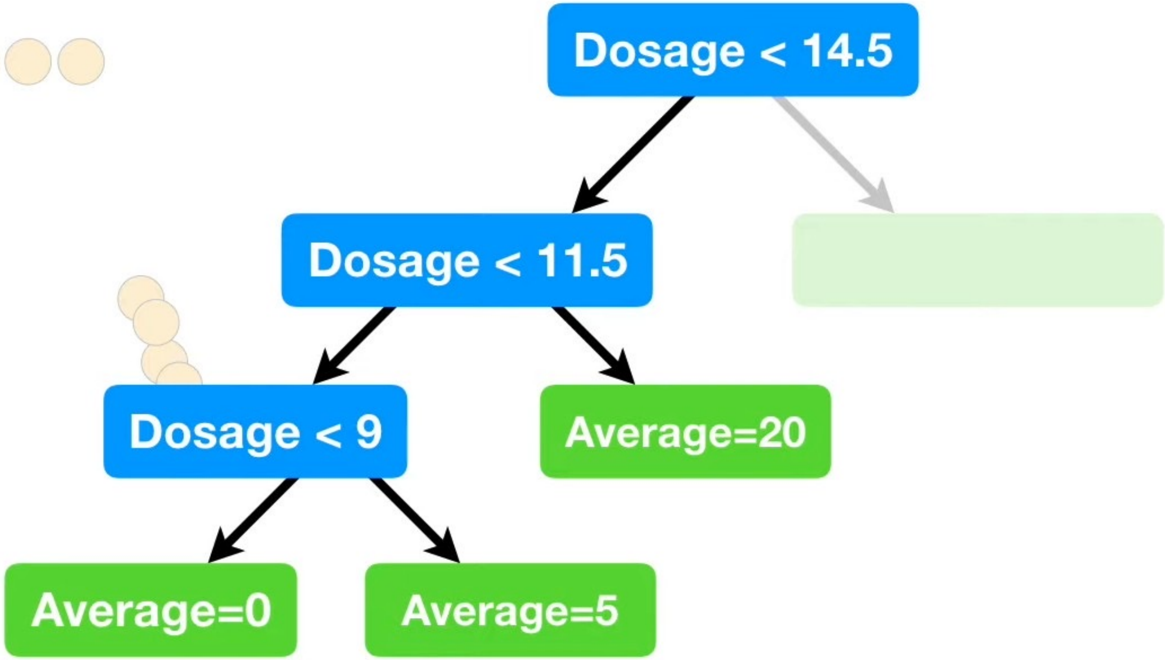


Yes, there are a bunch of techniques.

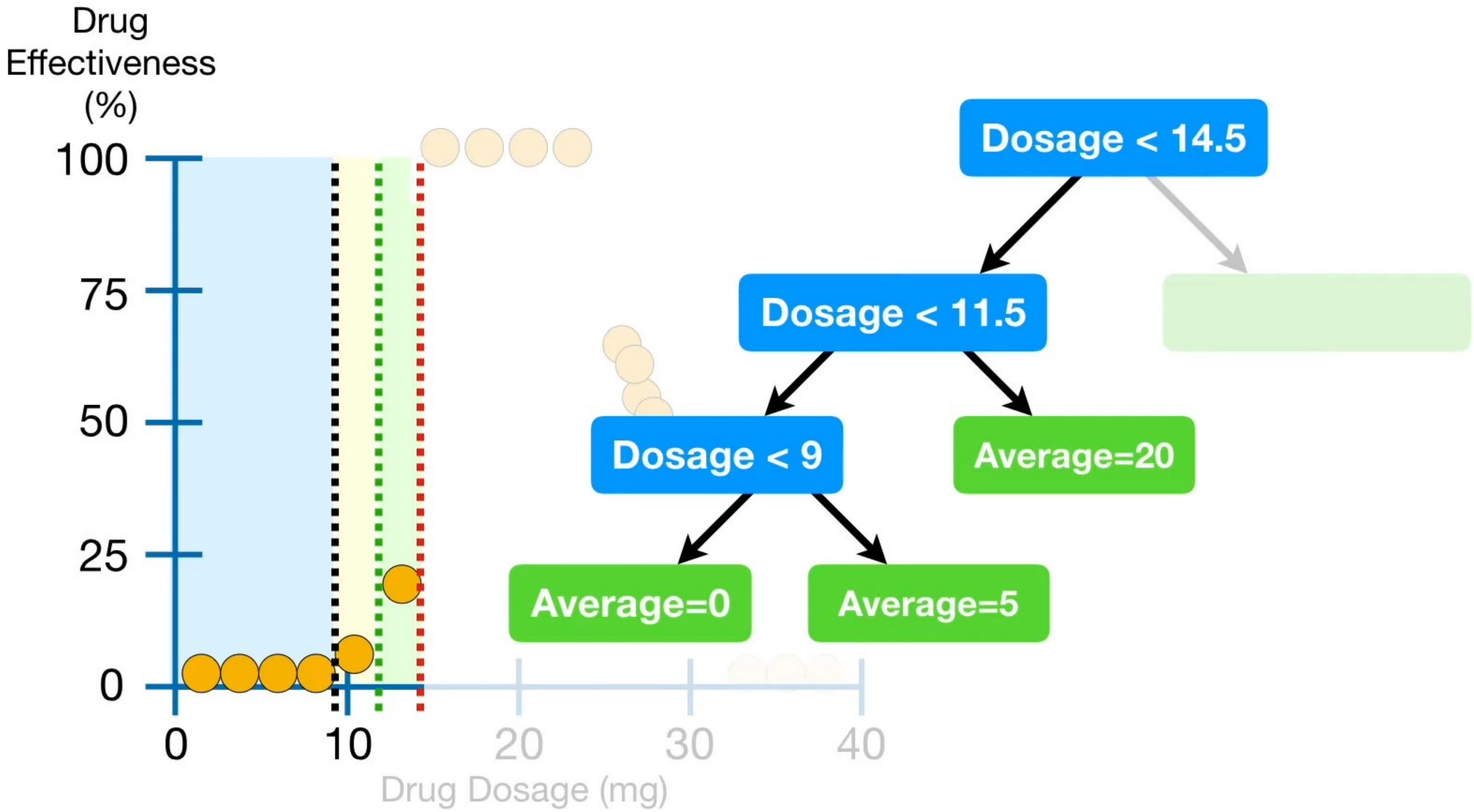


The simplest is to only split observations when there are more than some minimum number.





Typically, the minimum number of observations to allow for a split is **20**.



Dosage < 14.5

Dosage < 11.5

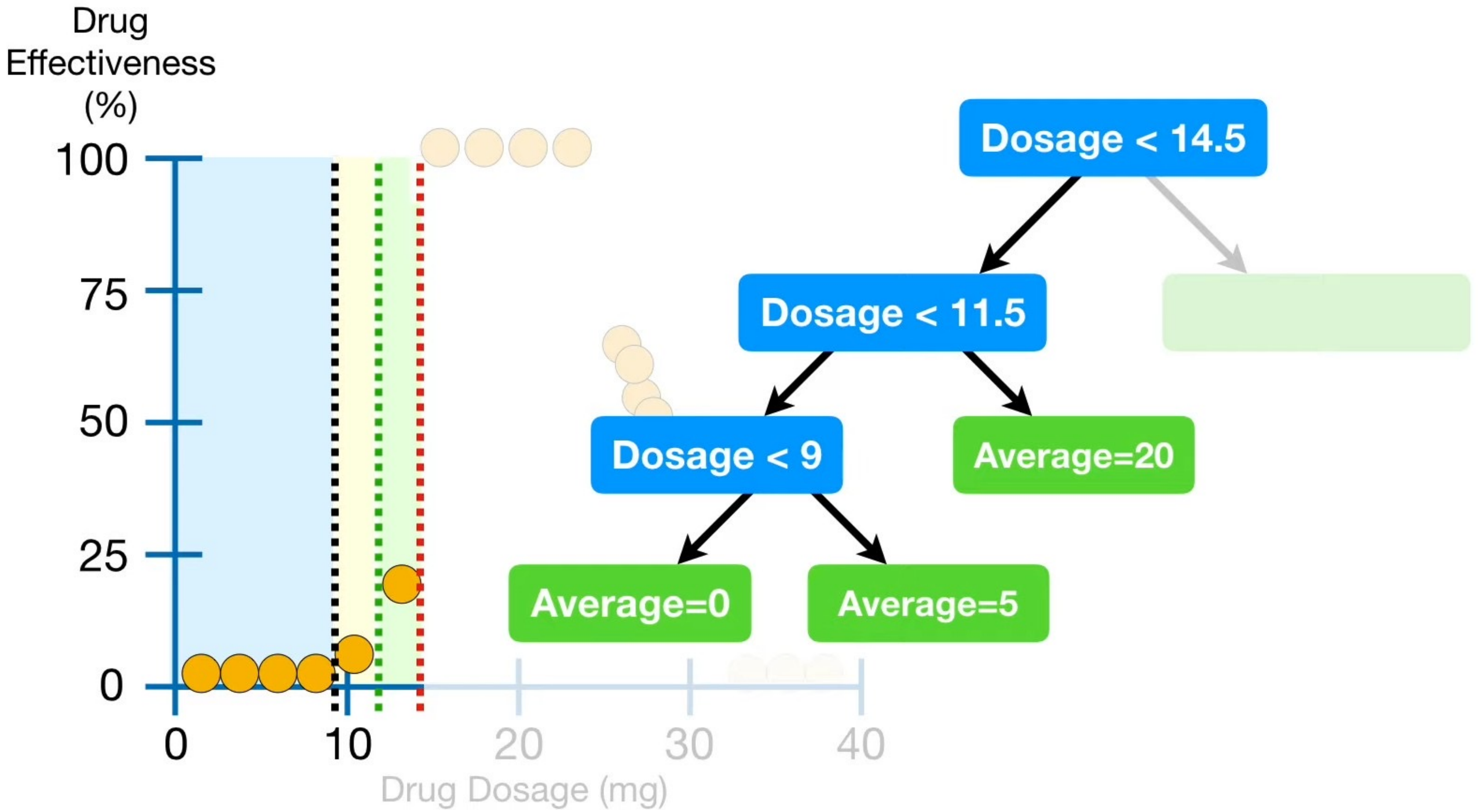
Dosage < 9

Average=20

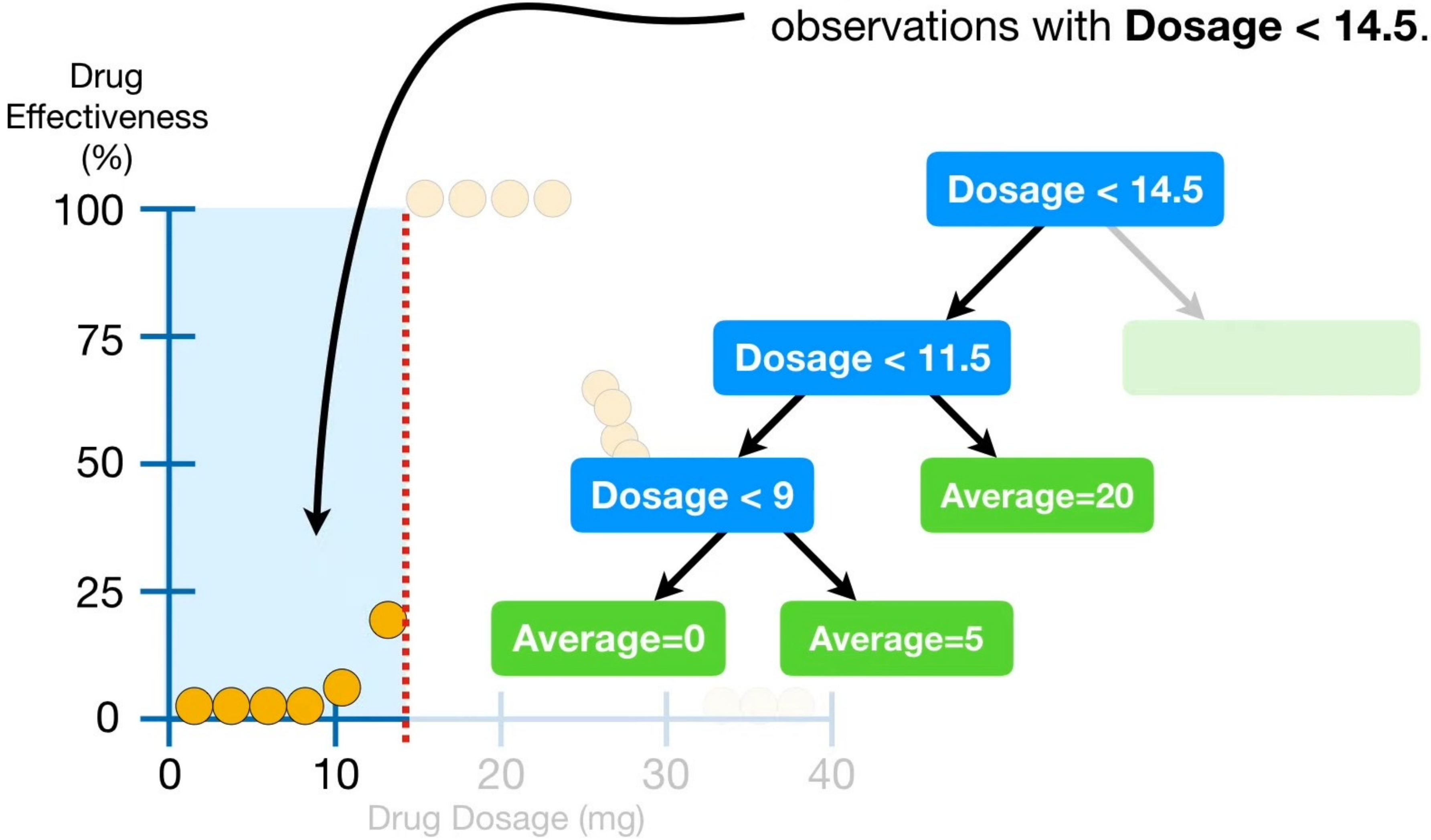
Average=0

Average=5

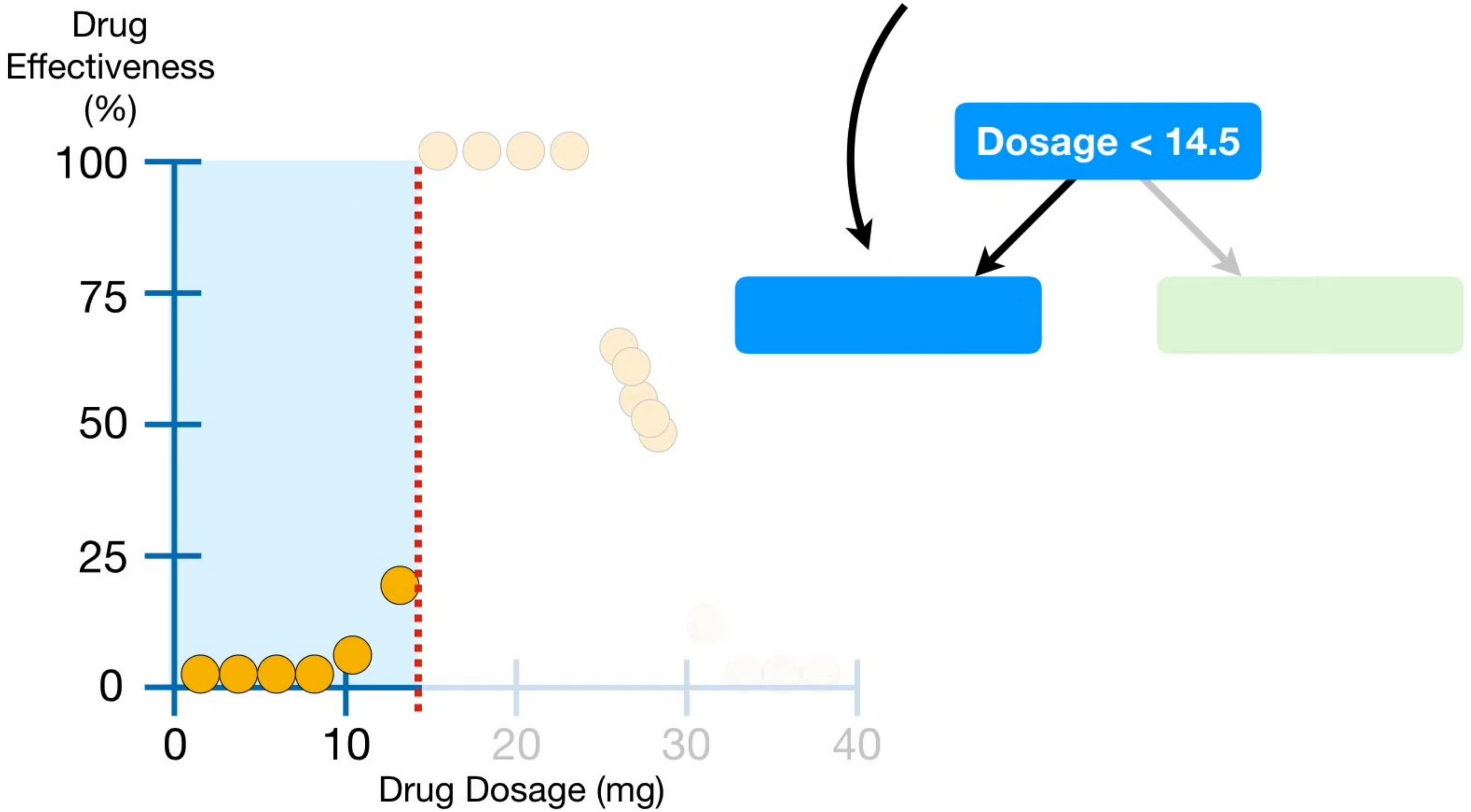
However, since this example doesn't have many observations, I set the minimum to **7**.



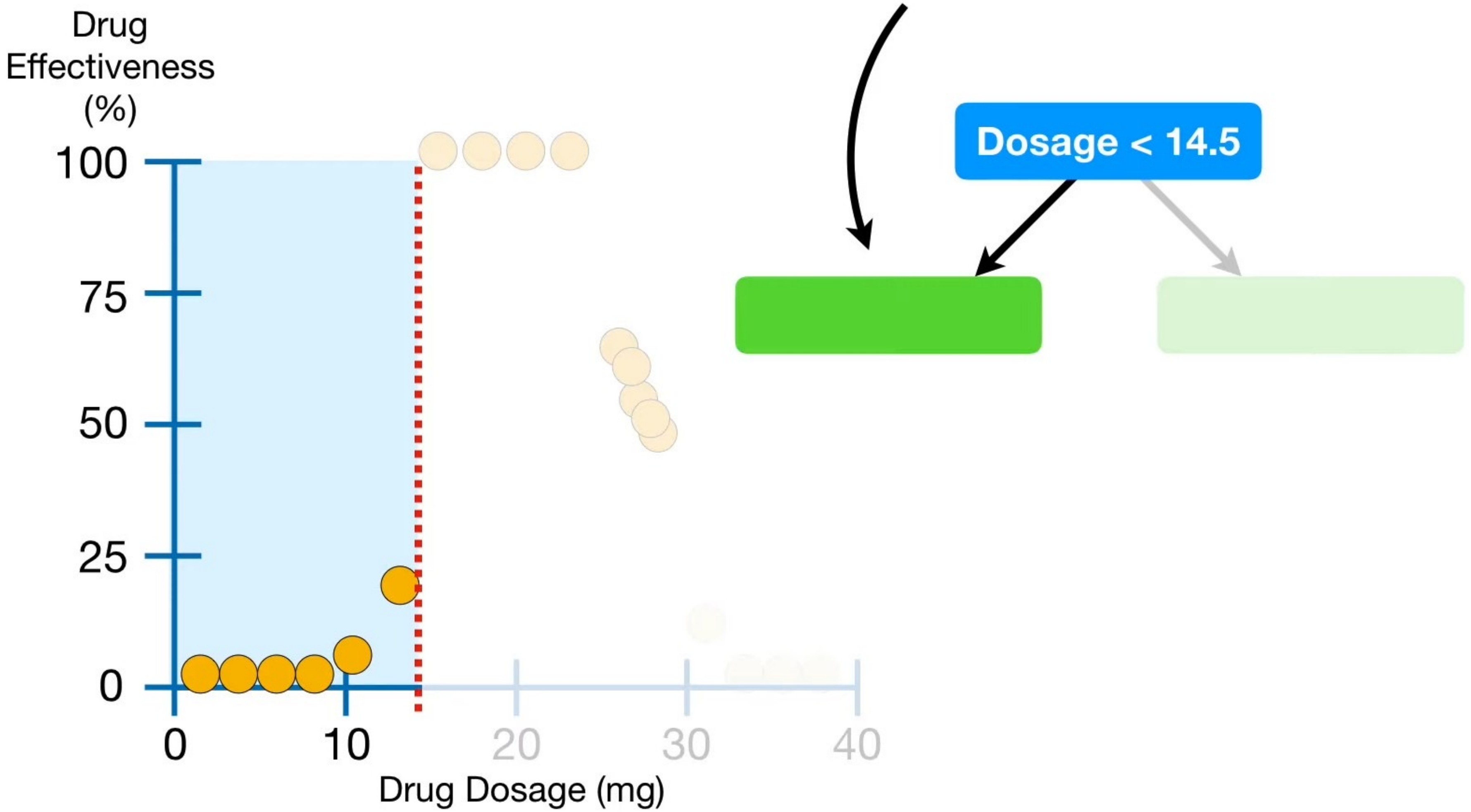
In other words, since there are only **6** observations with **Dosage < 14.5**…



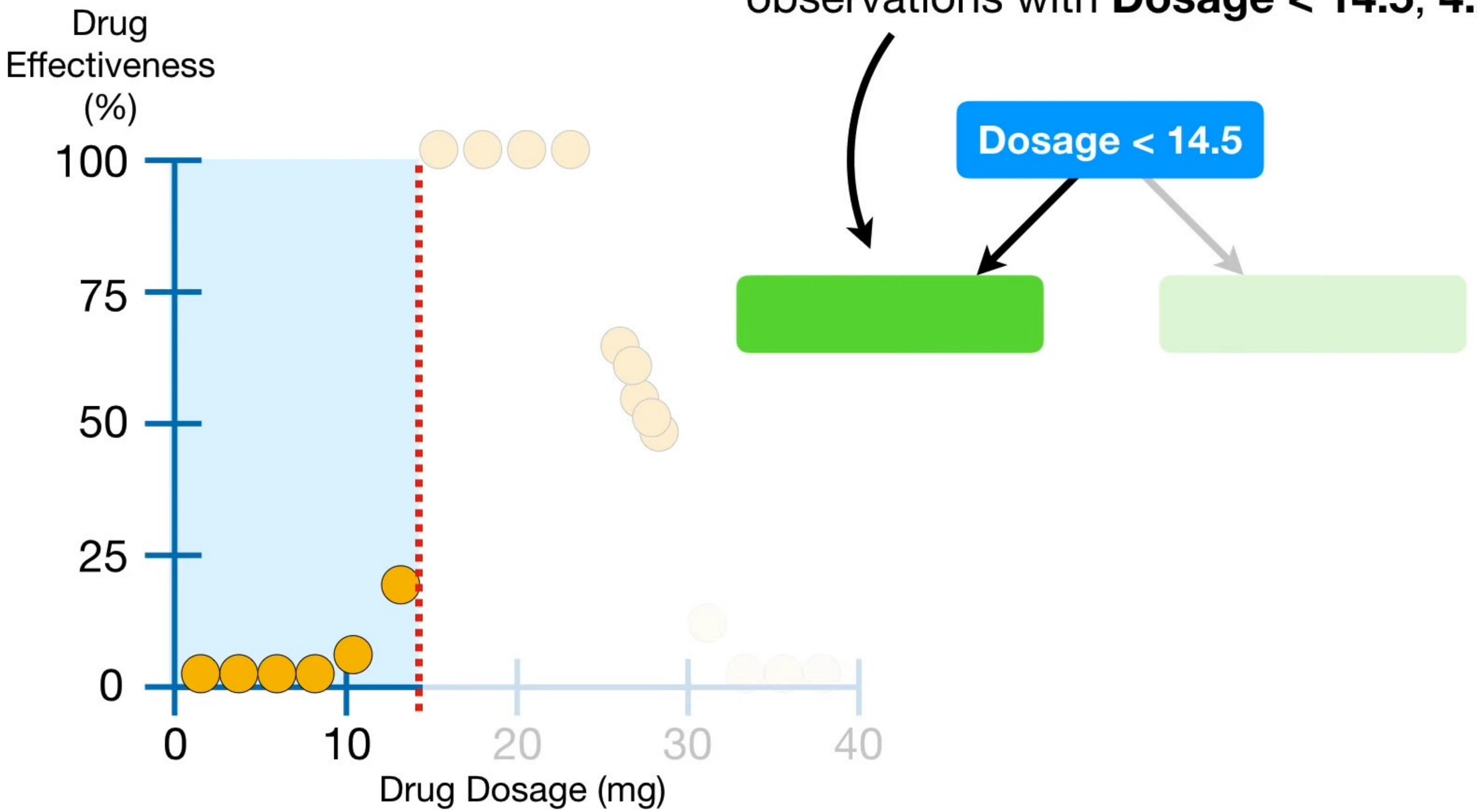
…we will not split the observations in this node.



Instead, this node will become a leaf…

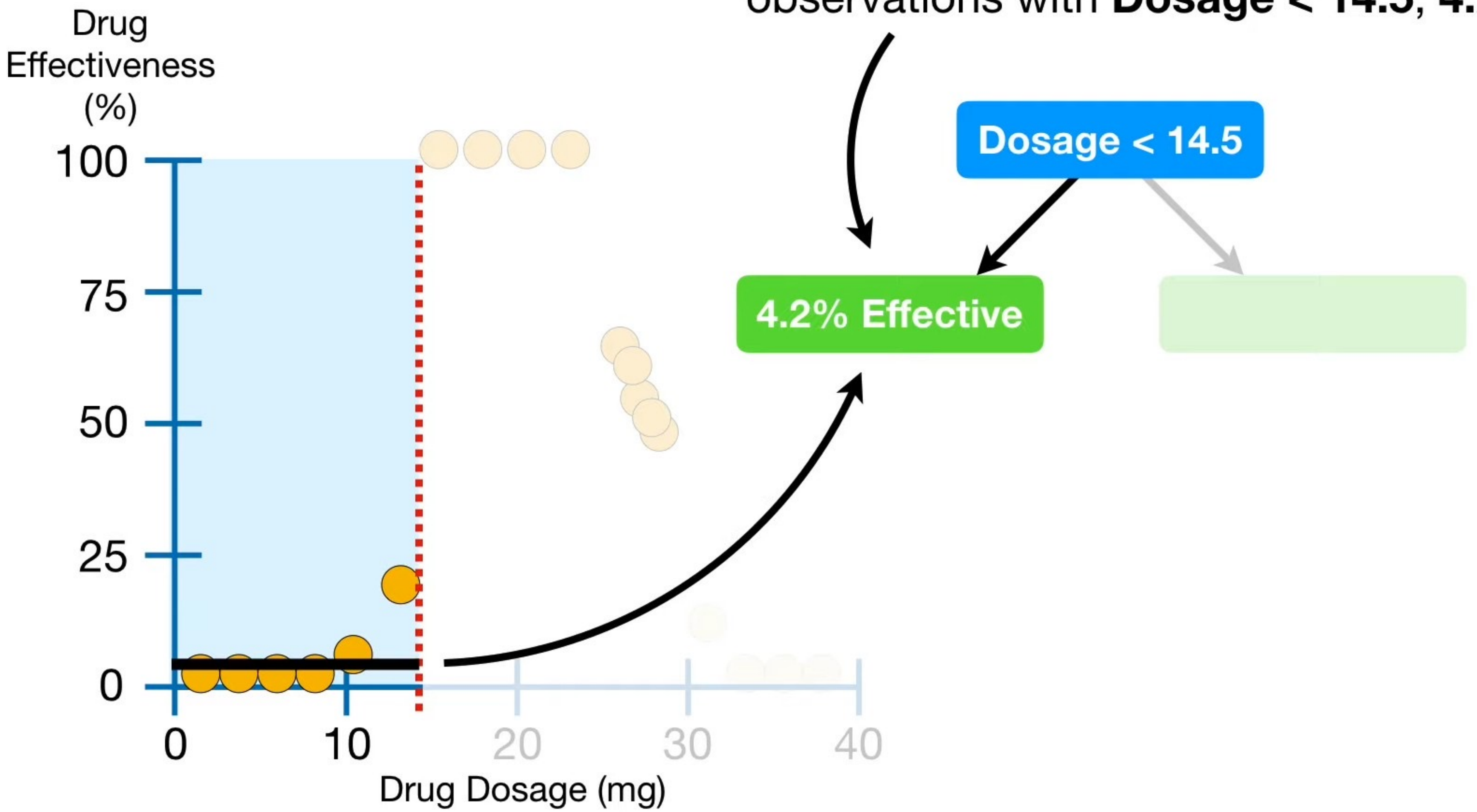


…and the output will be the average **Drug Effectiveness** for the **6** observations with **Dosage < 14.5, 4.2%**.

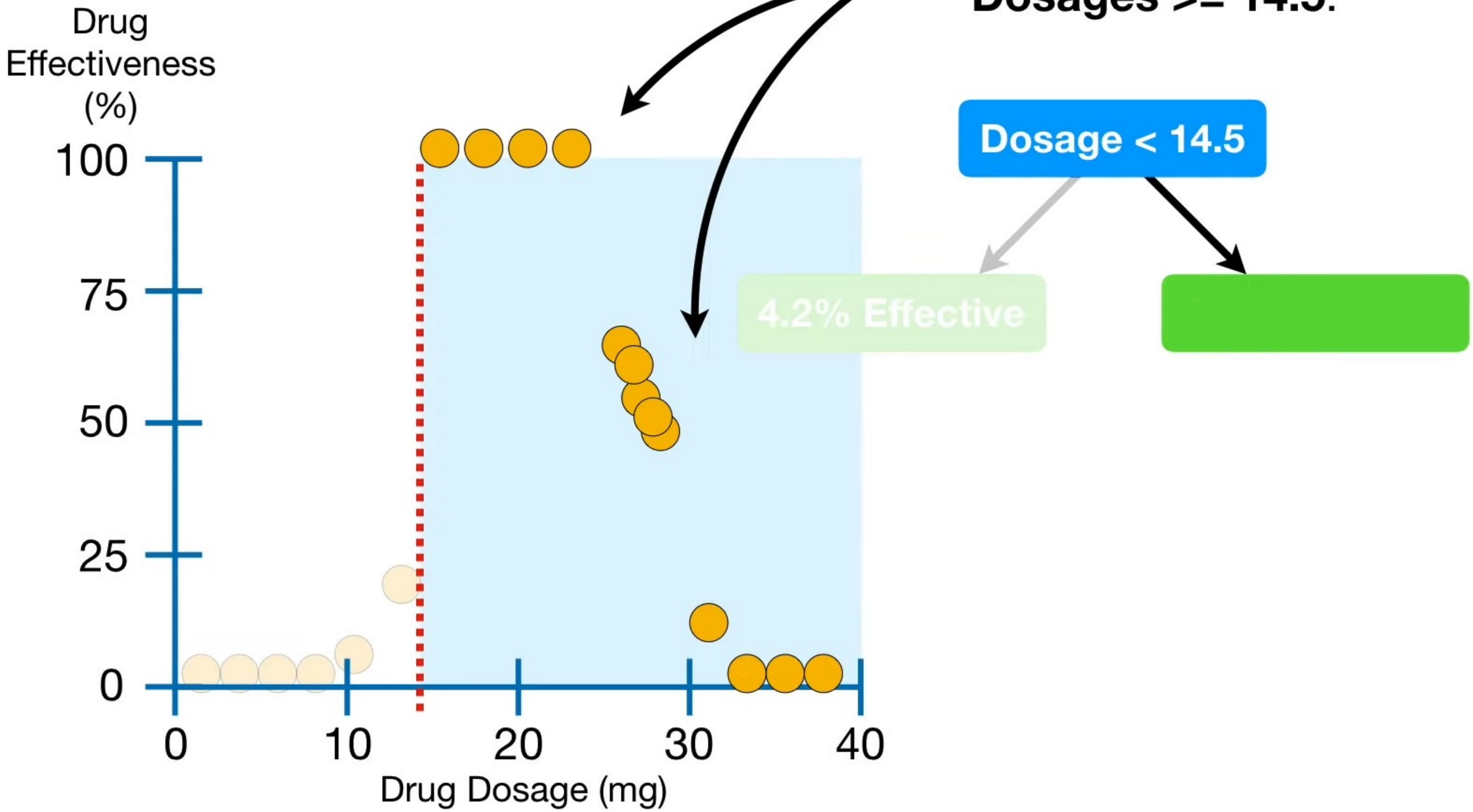


Dosage < 14.5

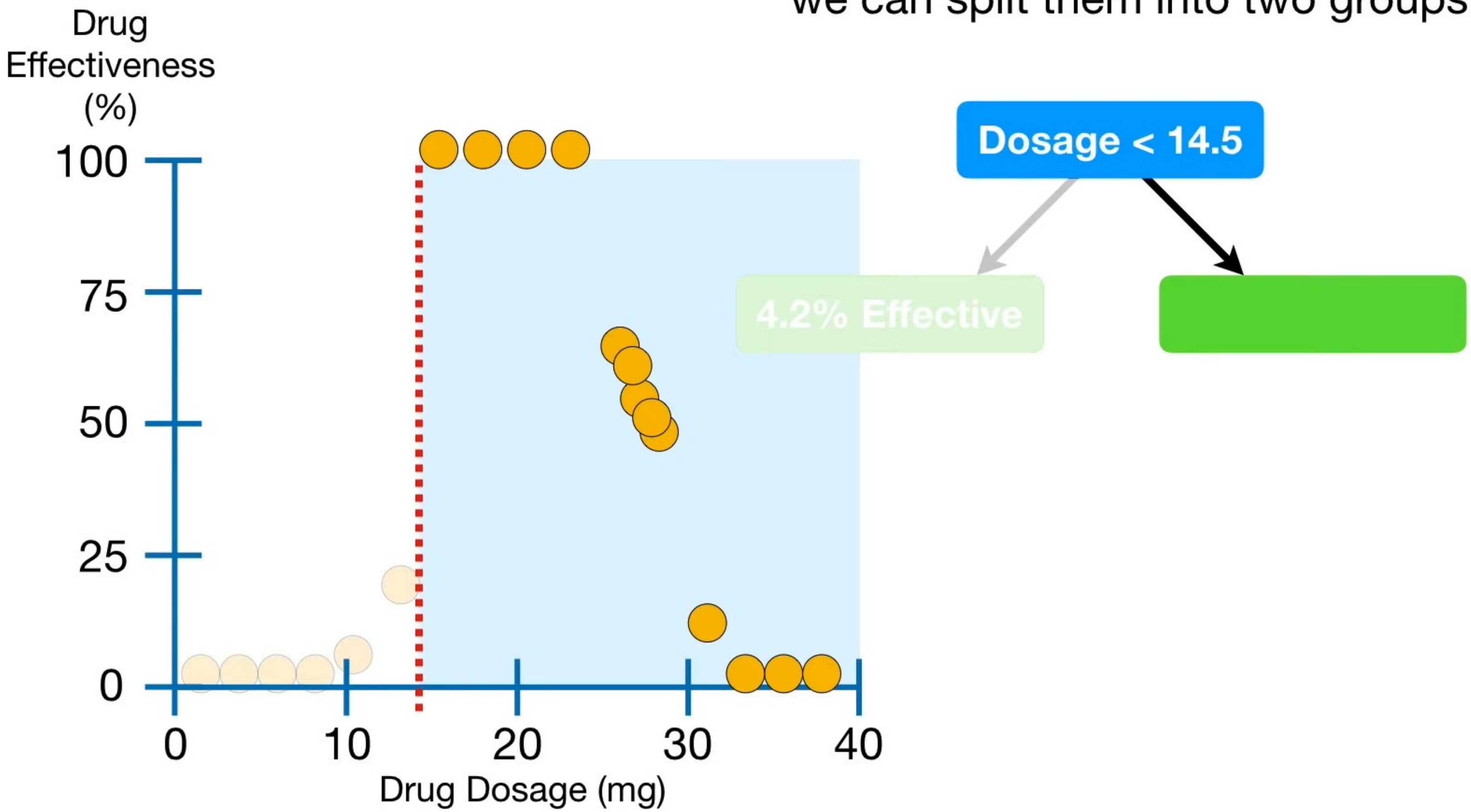
…and the output will be the average **Drug Effectiveness** for the **6** observations with **Dosage < 14.5**, **4.2%**.



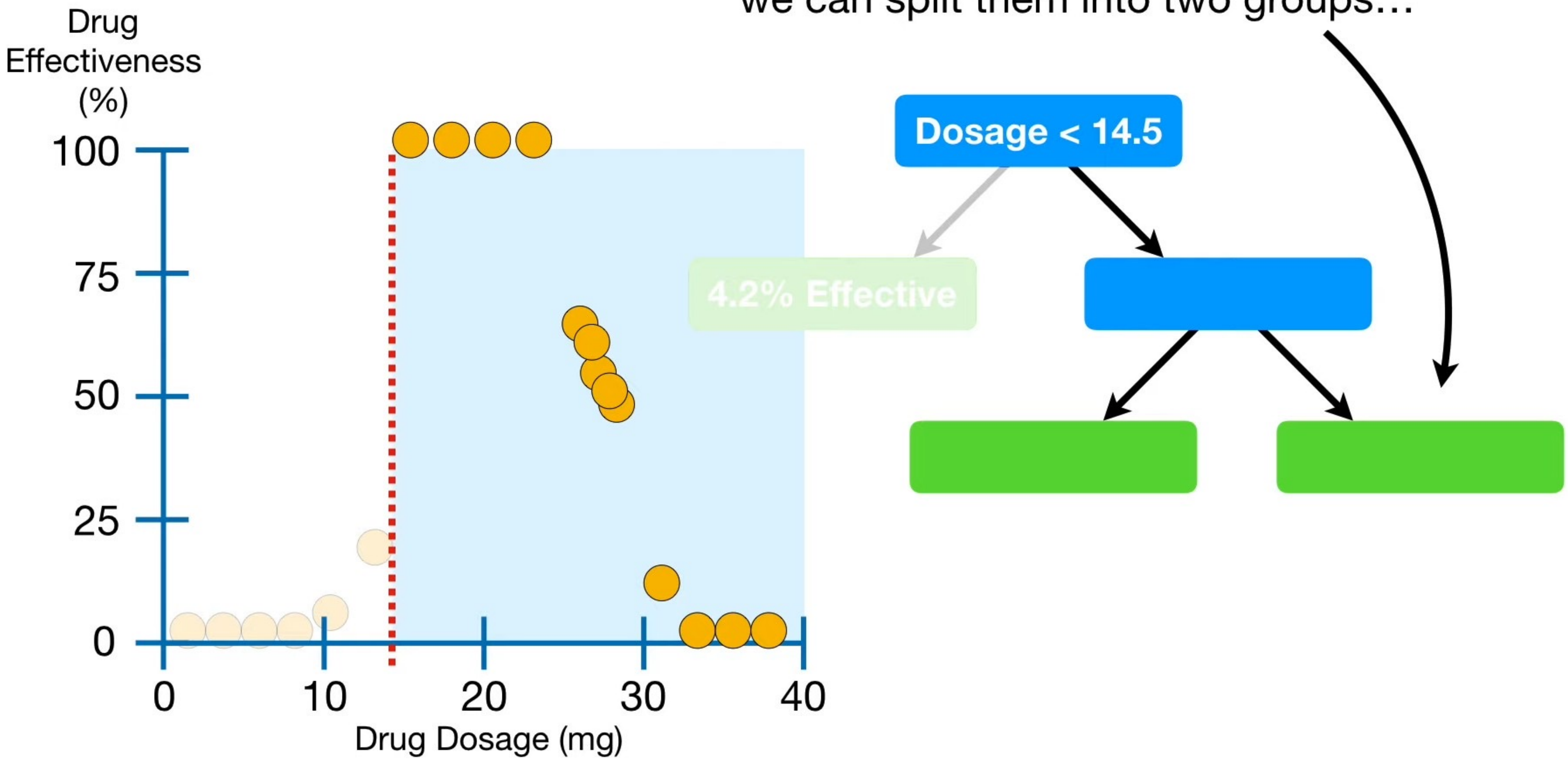
Now we need to figure out what to do with the remaining **13** observations with **Dosages >= 14.5**.



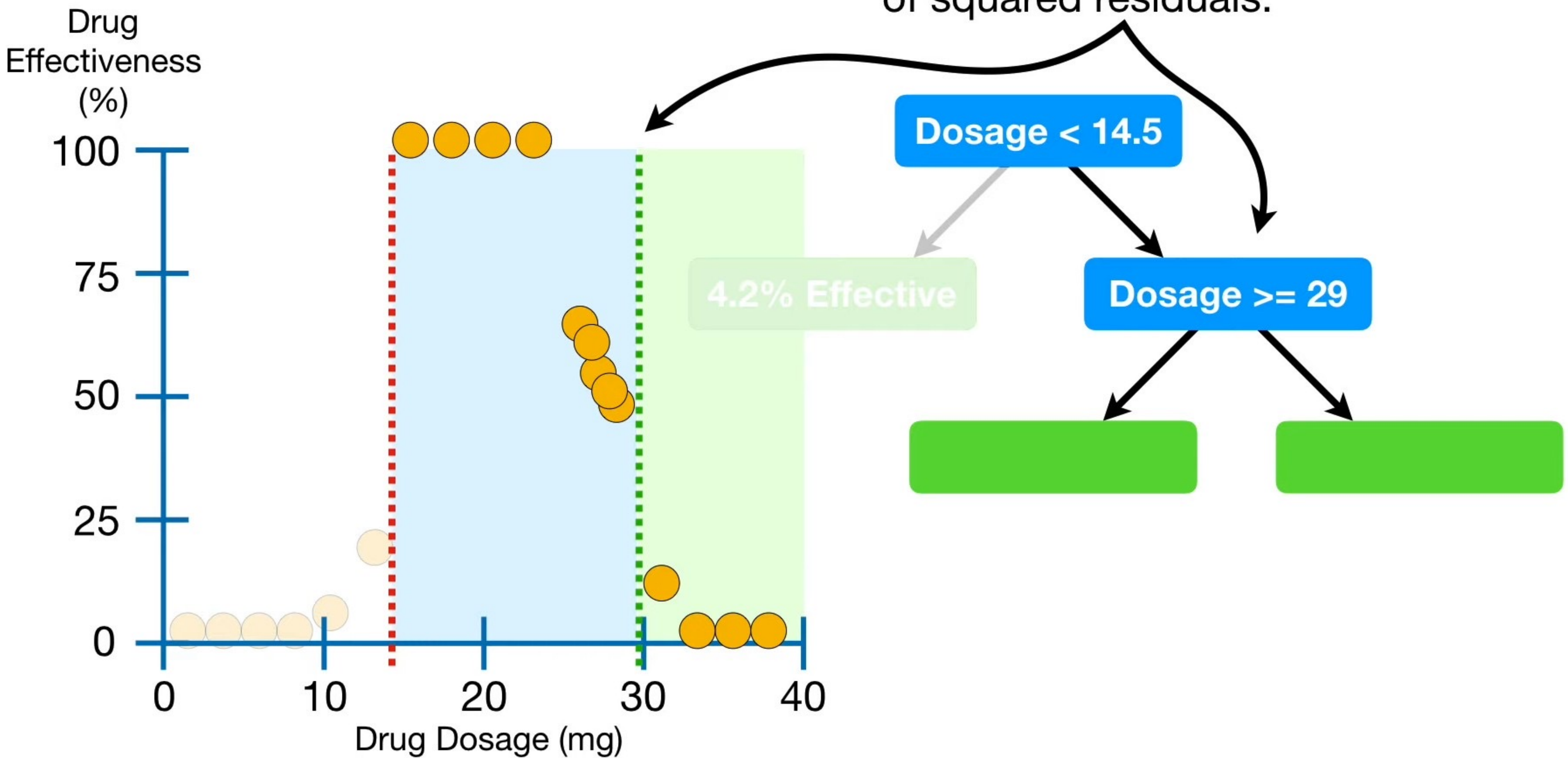
Since we have more than **7** observations on the right side (with **Dosage >= 14.5**), we can split them into two groups…



Since we have more than **7** observations on the right side (with **Dosage >= 14.5**), we can split them into two groups…



…and we do that by finding the threshold that gives us the smallest sum of squared residuals.

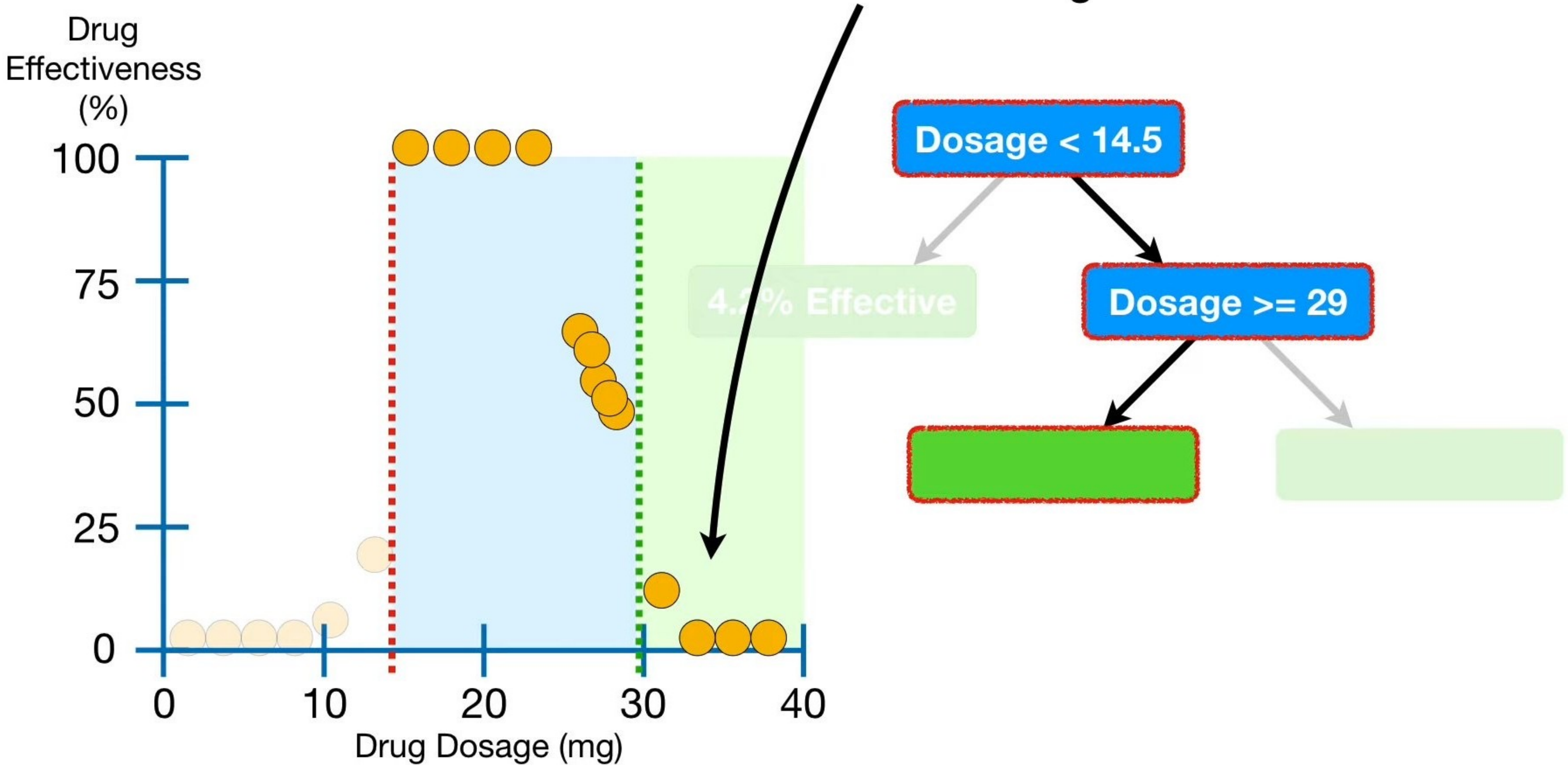


Dosage < 14.5

4.2% Effective

Dosage >= 29

NOTE: There are only **4** observations with **Dosage >= 29**…

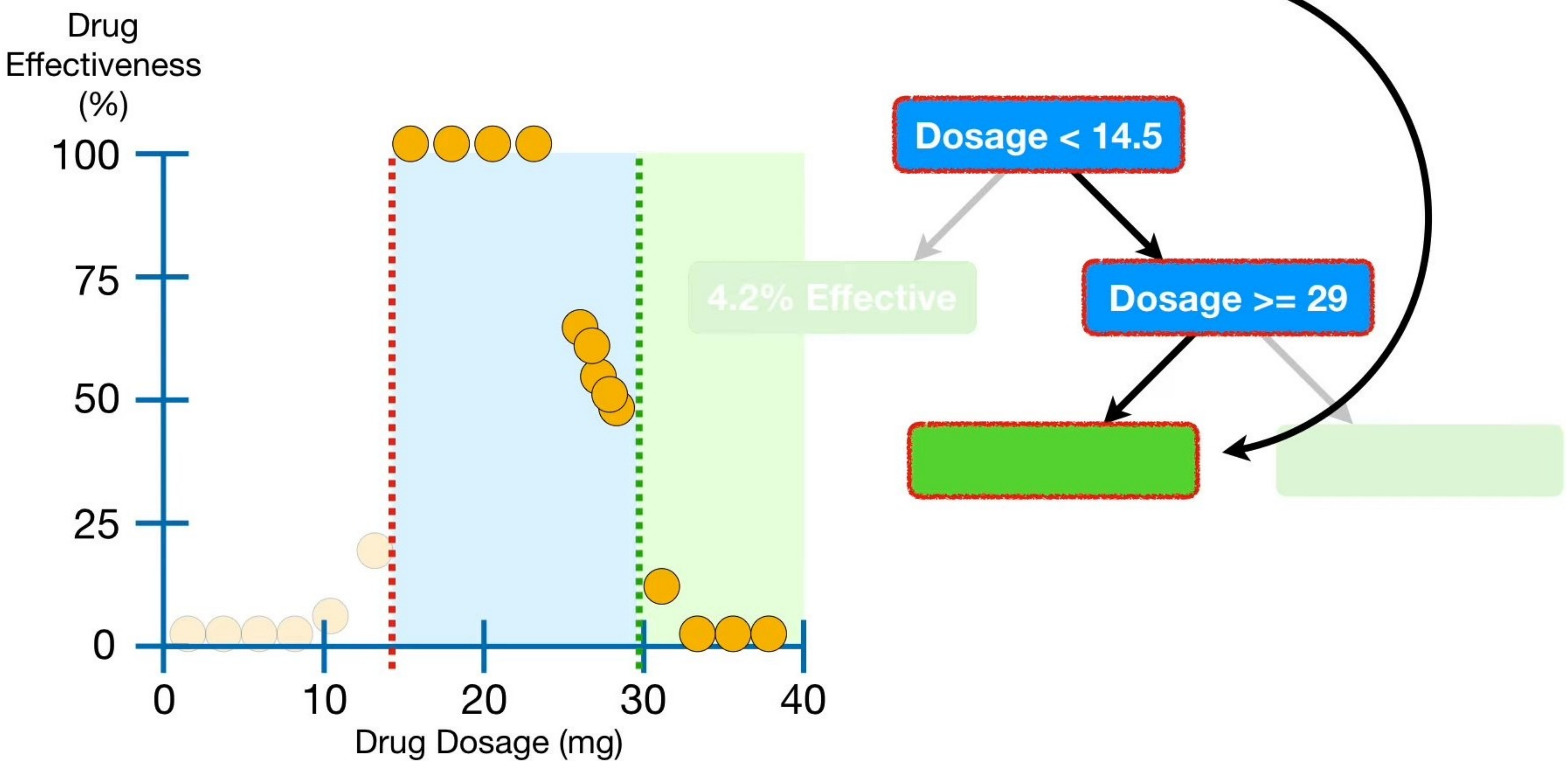


Dosage < 14.5

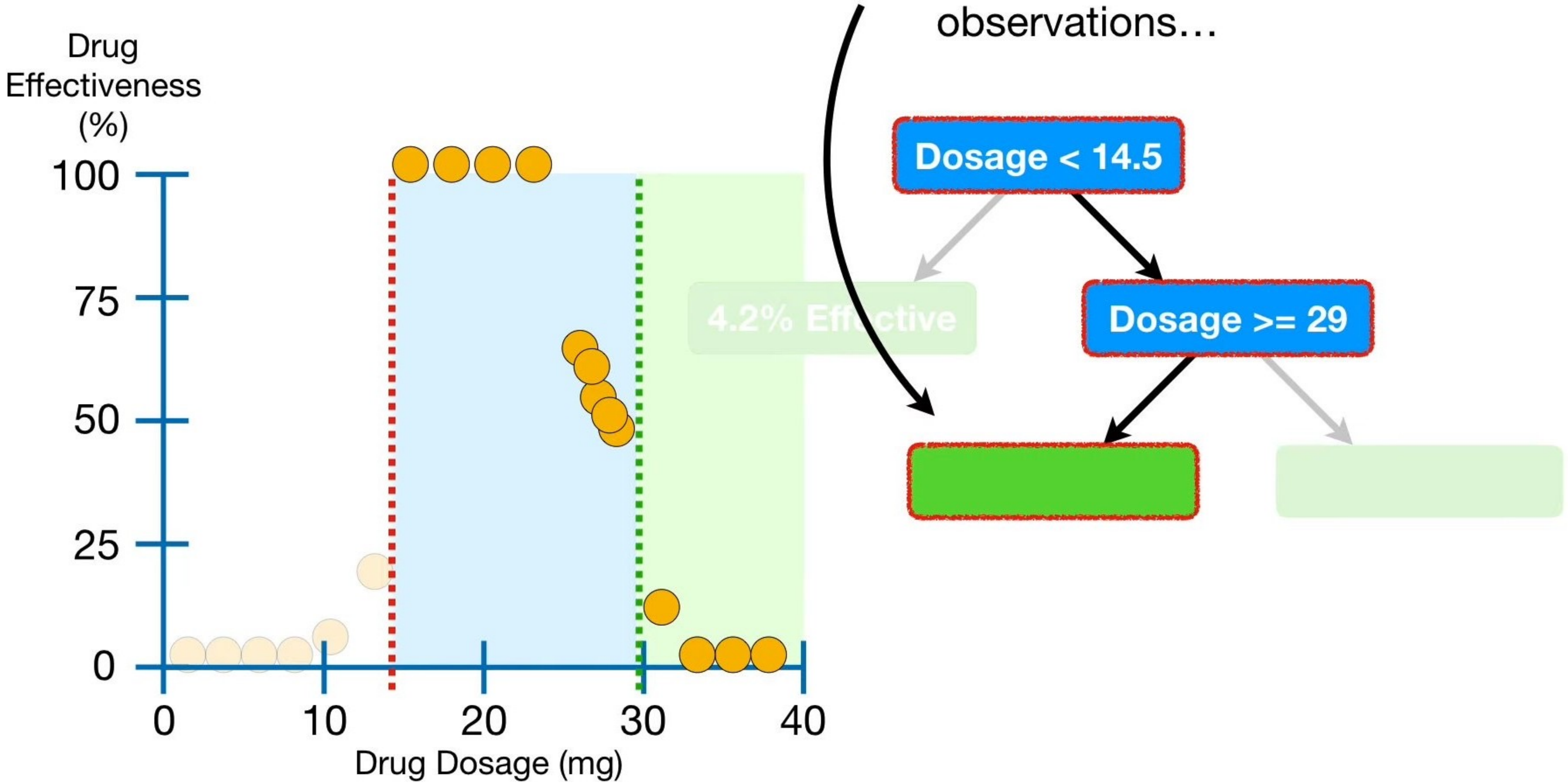
4.2% Effective

Dosage >= 29

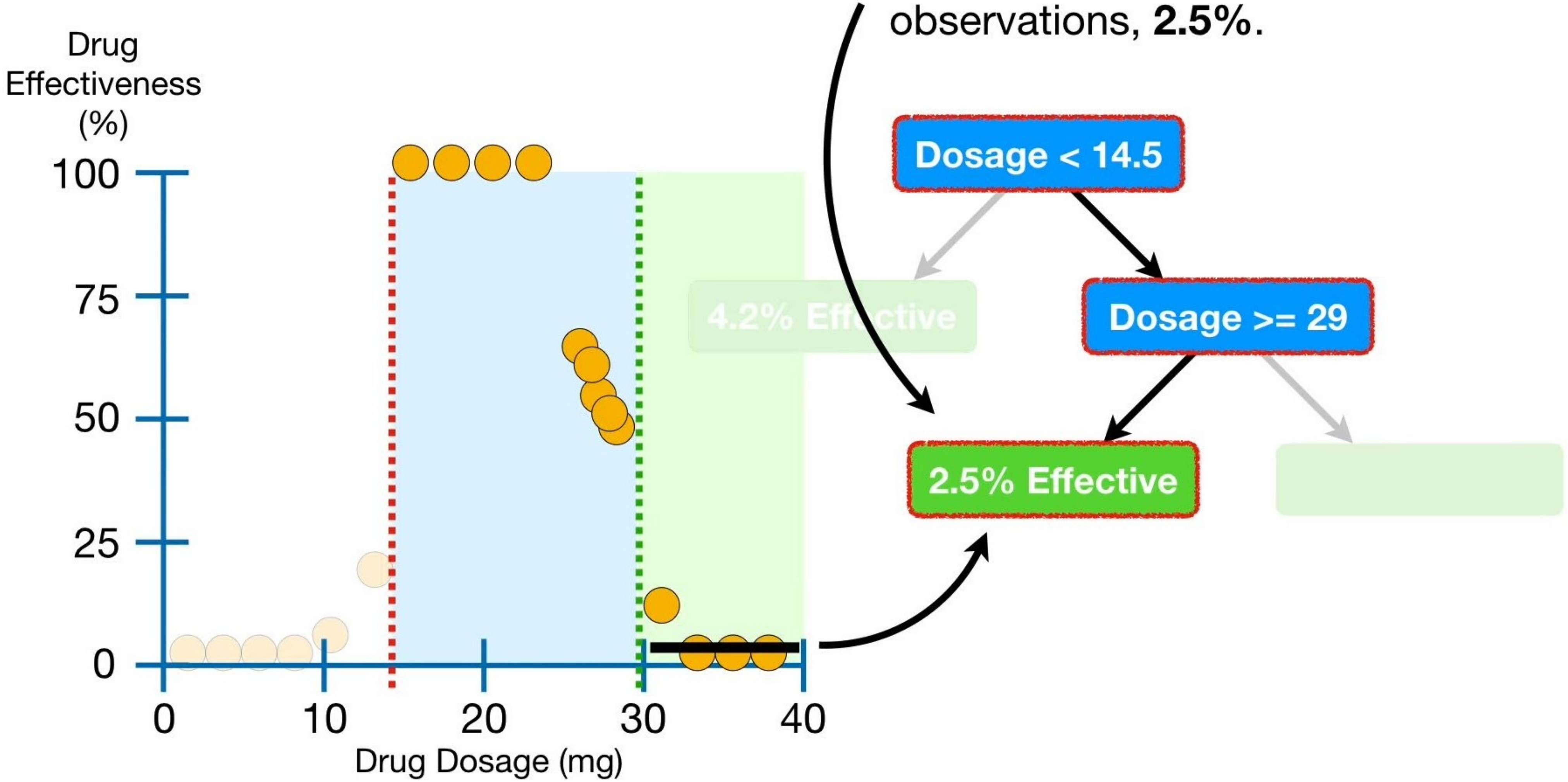
…thus, there are only **4** observations in this node…



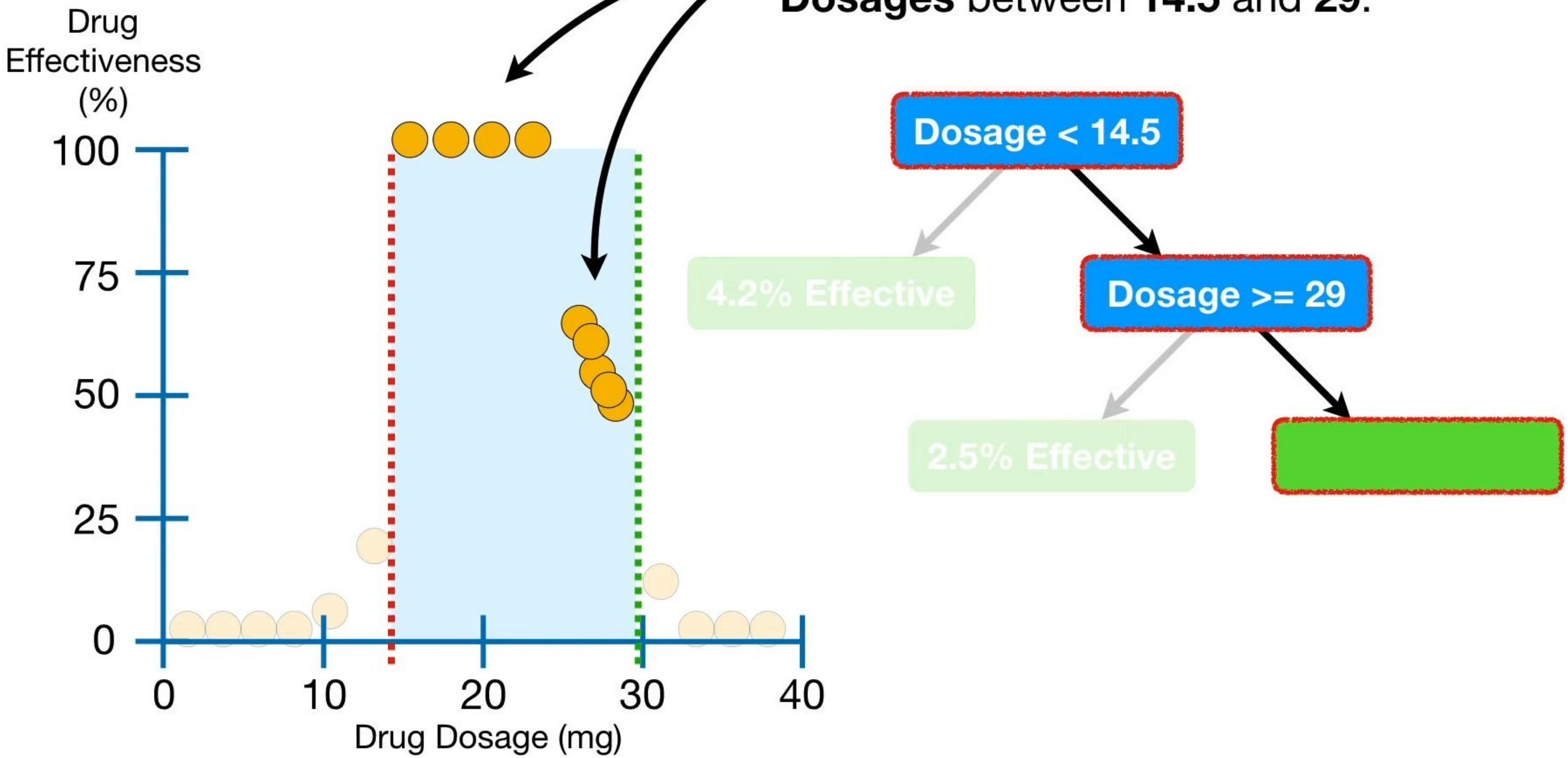
…thus, we will make this a leaf because it contains fewer than **7** observations…



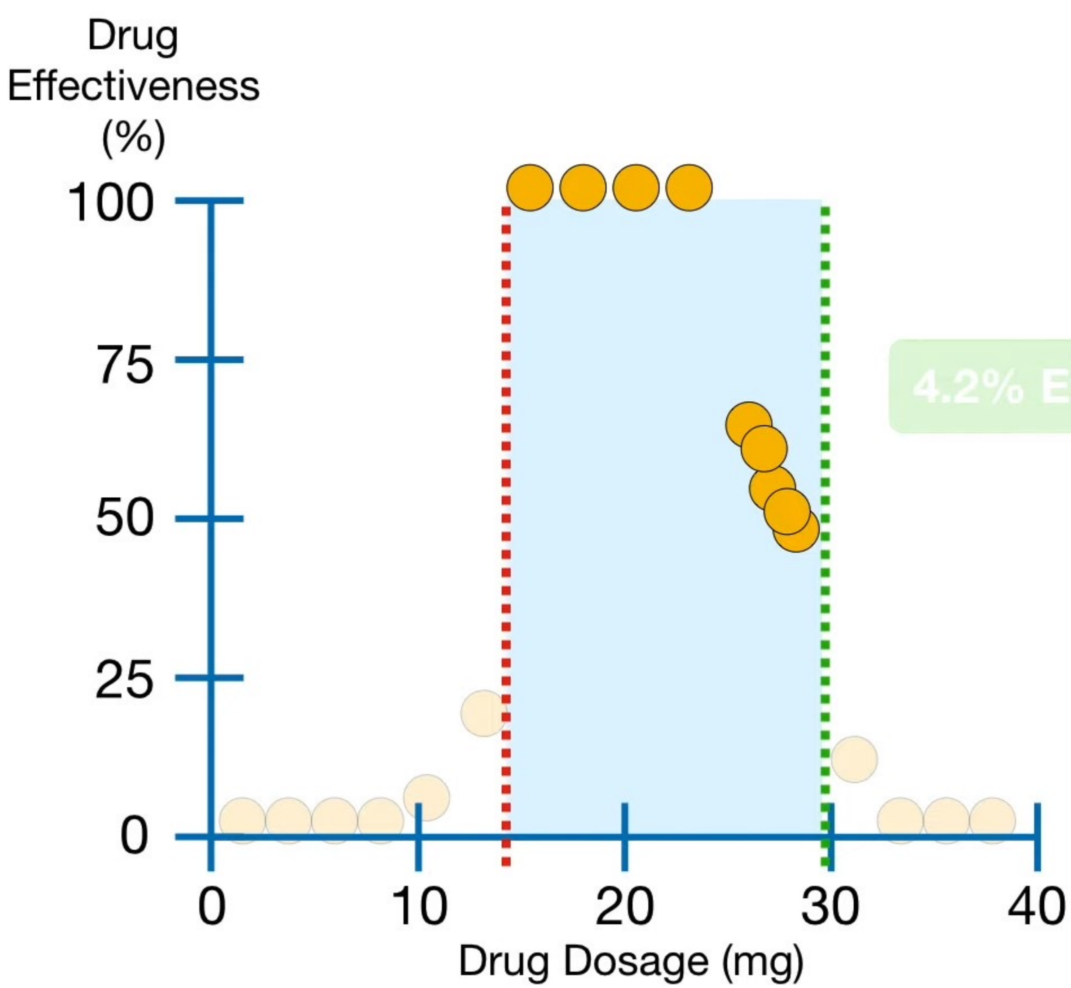
…and the output will be average **Drug Effectiveness** for these **4** observations, **2.5%**.

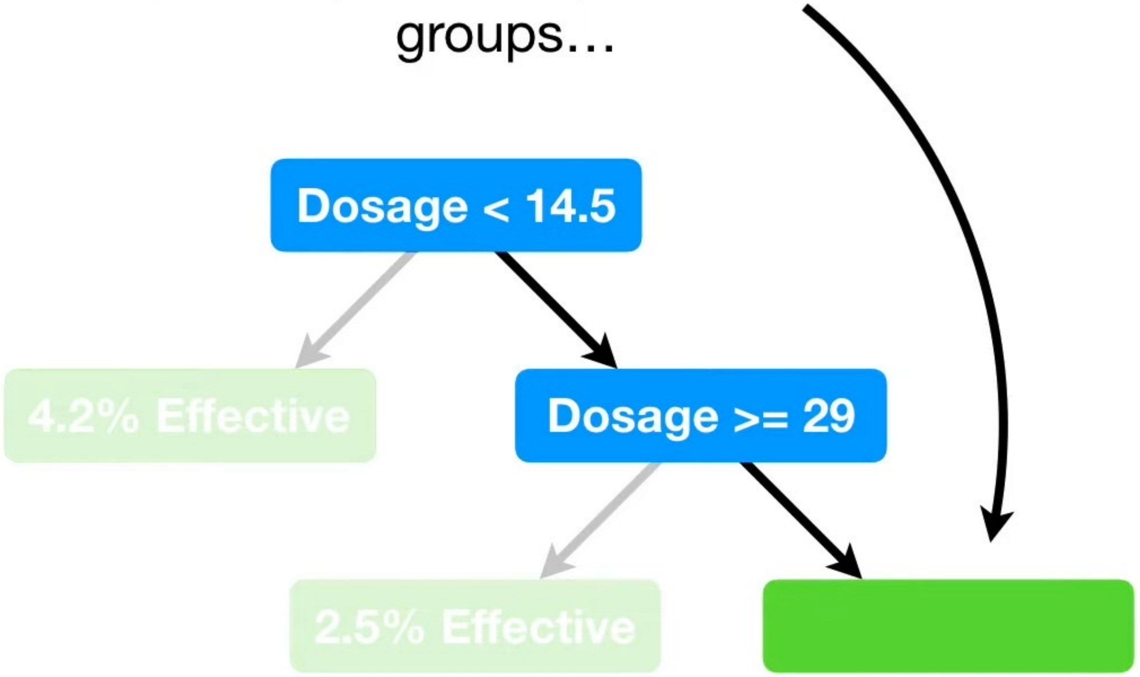


Now we need to figure out what to do with the **9** observations with **Dosages** between **14.5** and **29**.

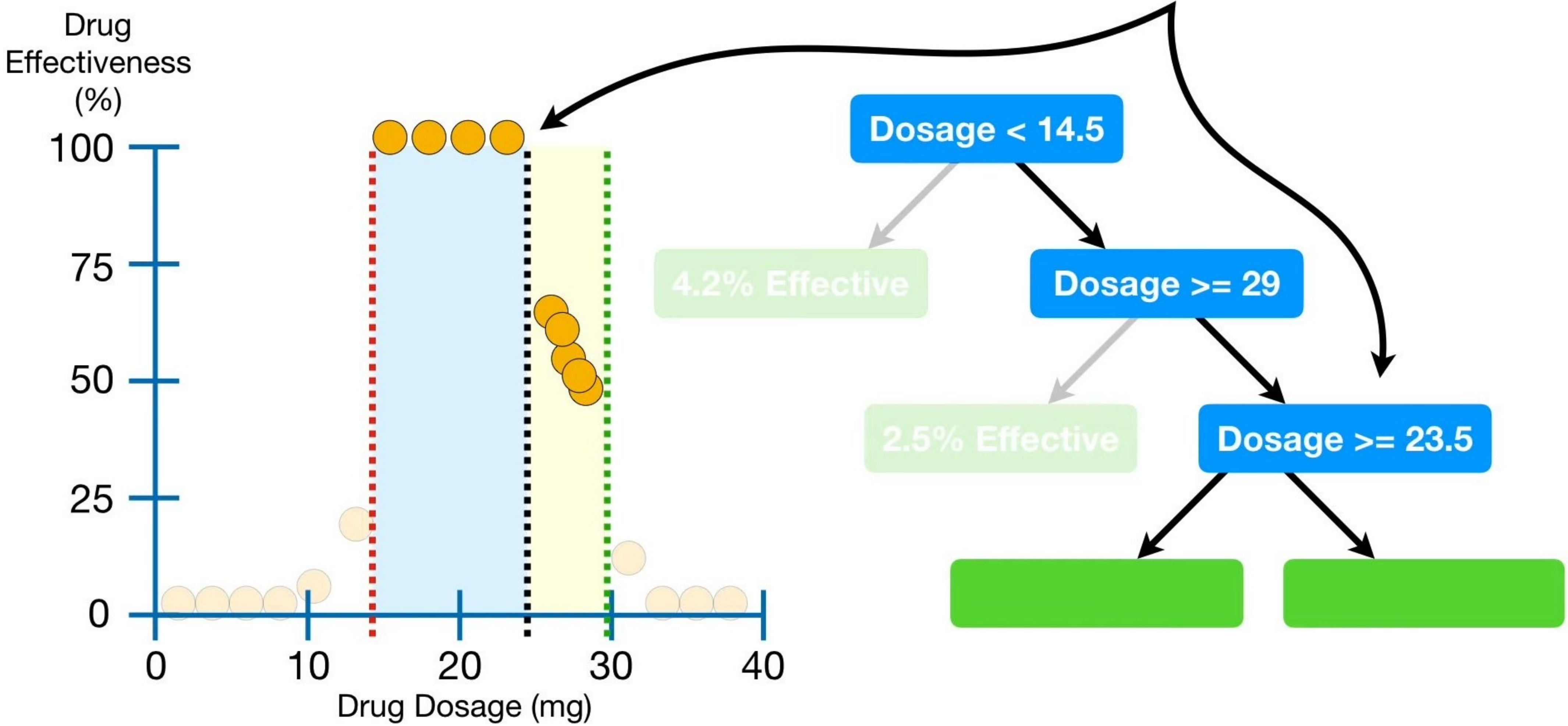


Since we have more than **7** observations, we can split them into two groups…





…by finding the threshold that gives us the minimum sum of squared residuals.



Dosage < 14.5

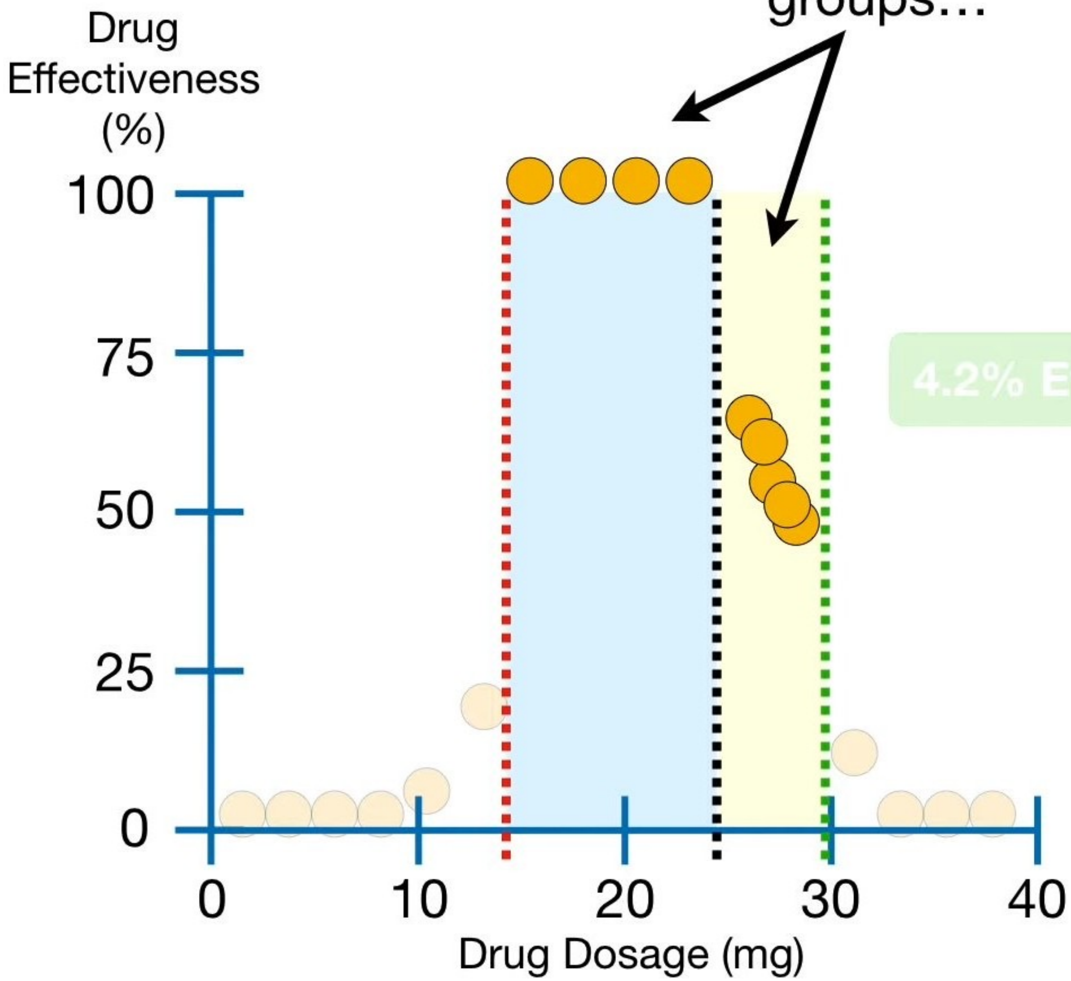
4.2% Effective

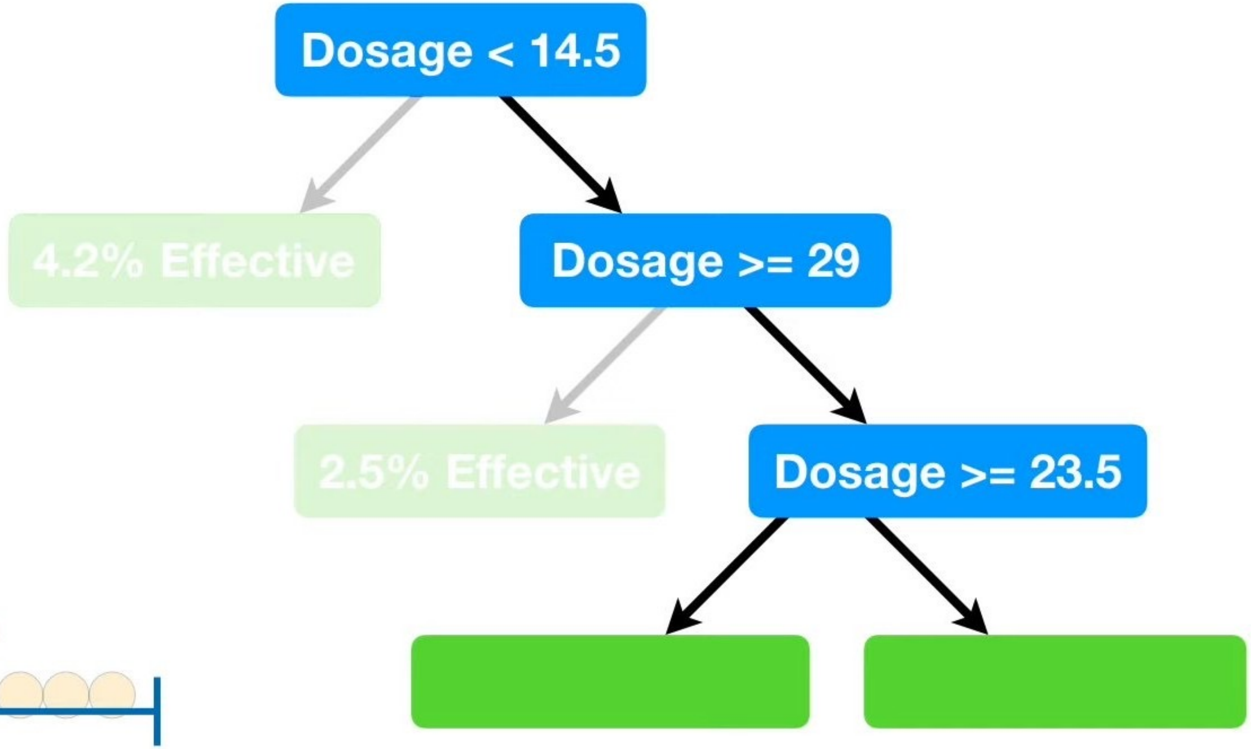
Dosage >= 29

2.5% Effective

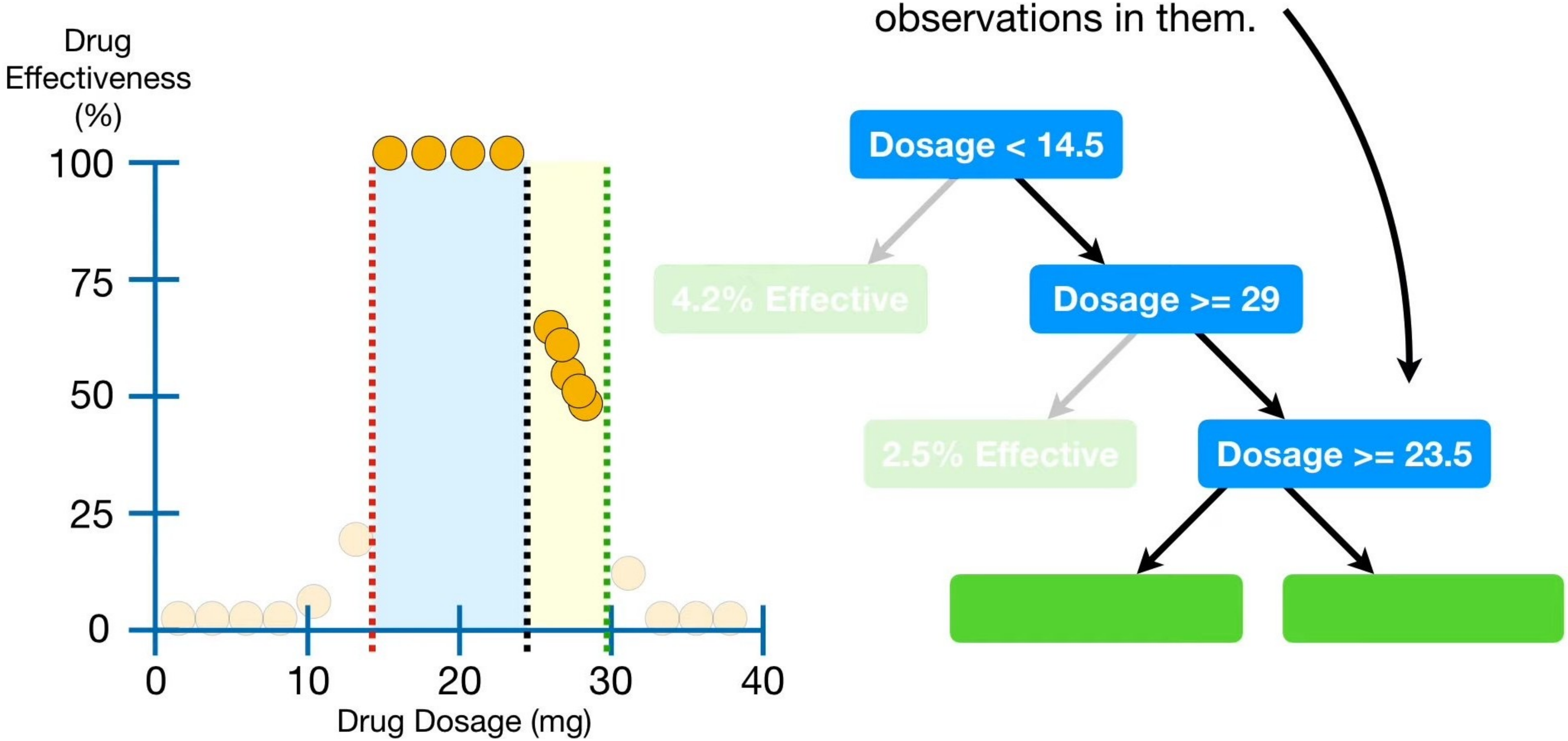
Dosage >= 23.5

NOTE: Since there are fewer than **7** observations in each of these two groups…

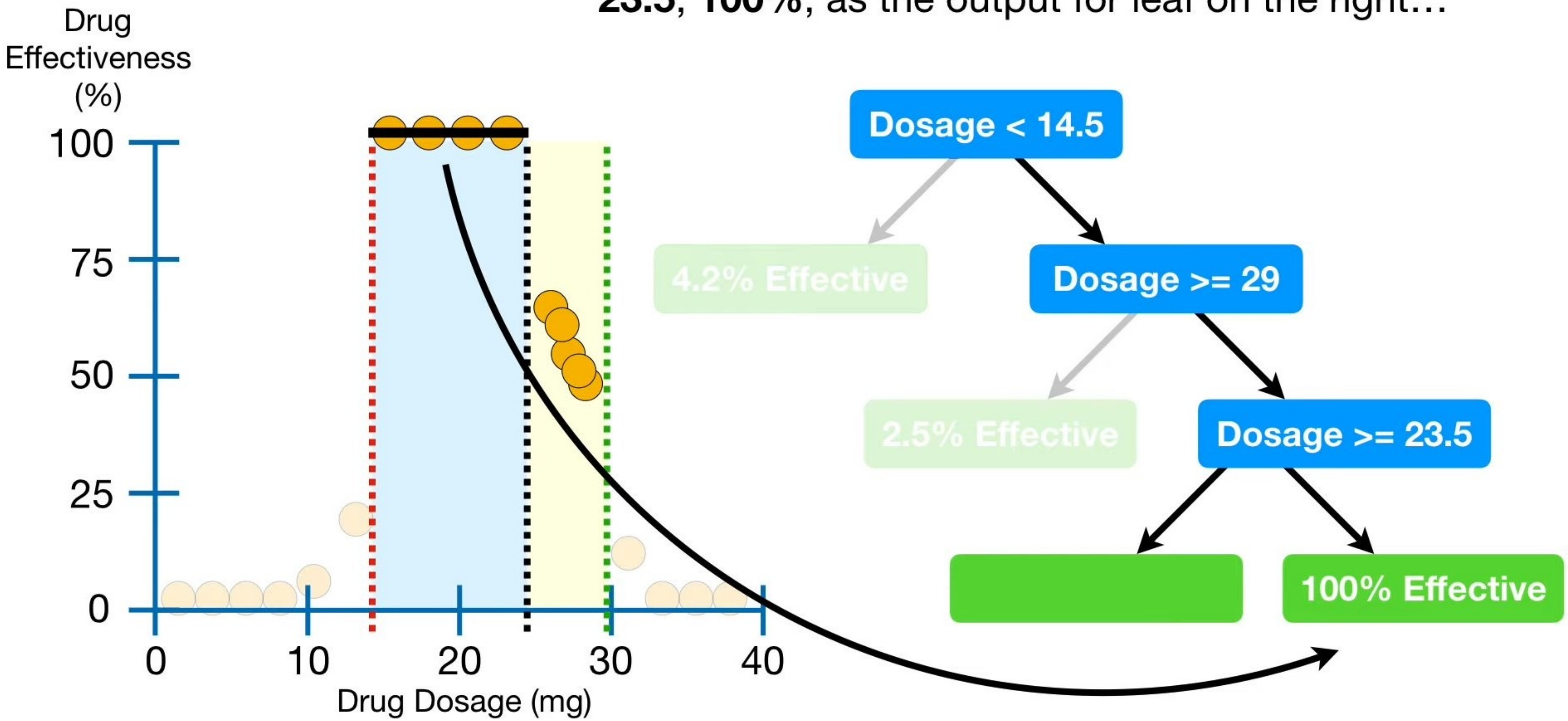




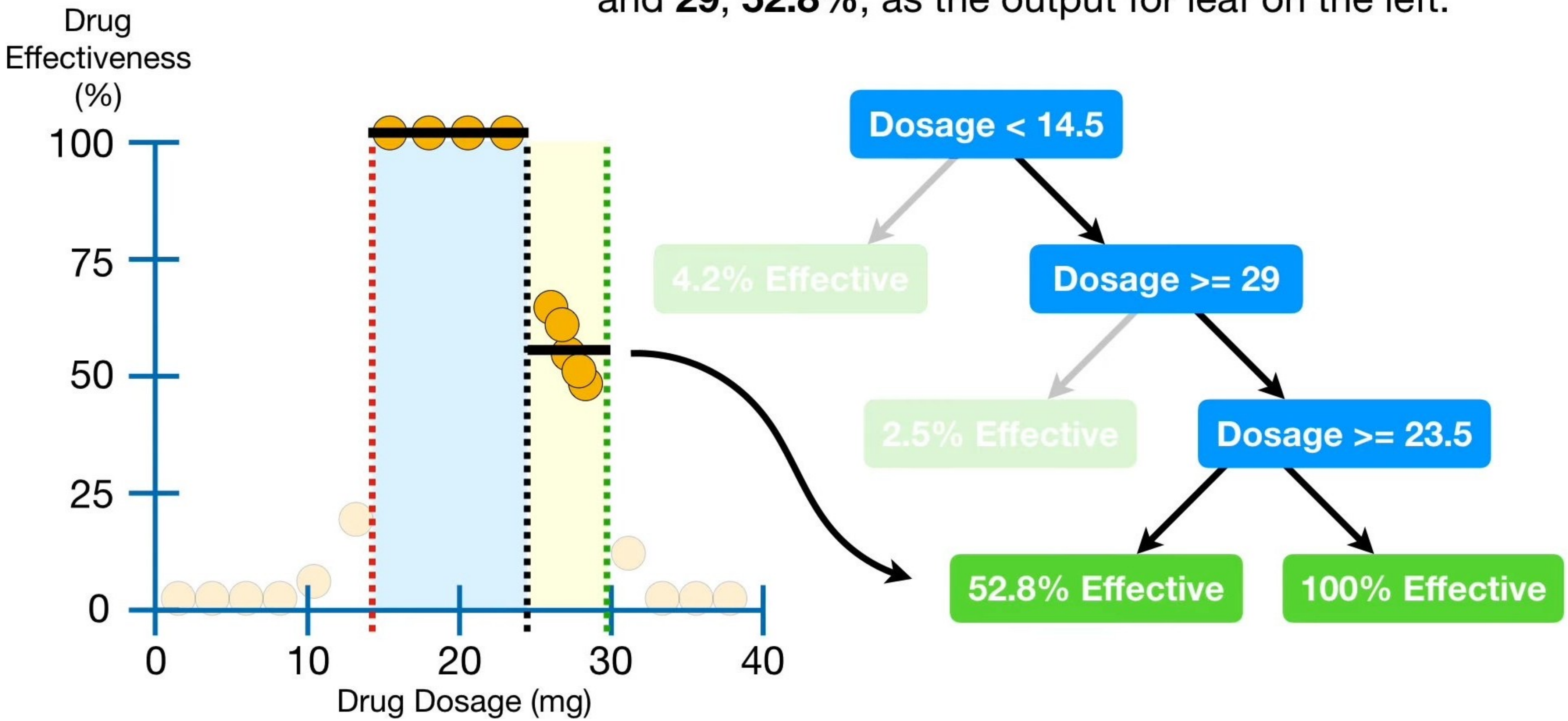
….this is the last split, because none of the leaves have more than **7** observations in them.



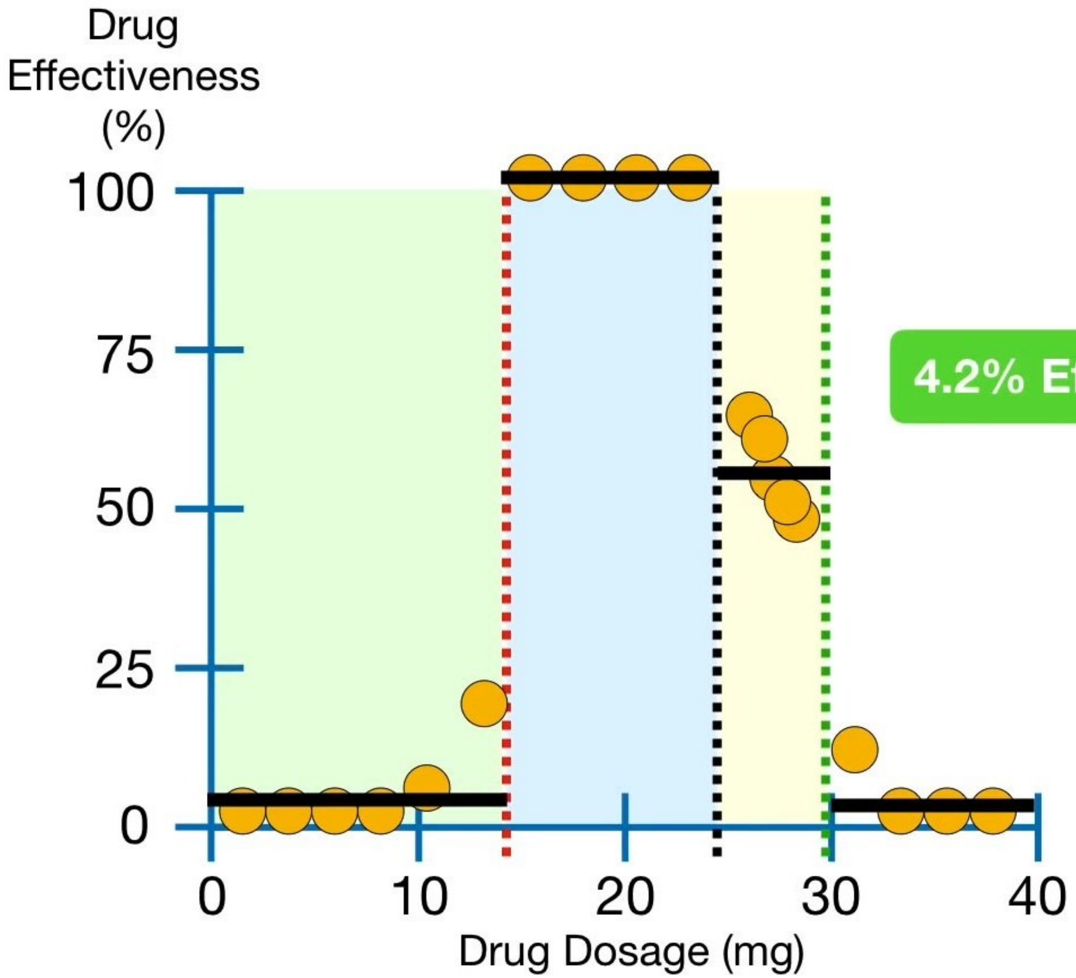
So we use the average **Drug Effectiveness** for observations with **Dosages** between **14.5** and **23.5**, **100%**, as the output for leaf on the right…

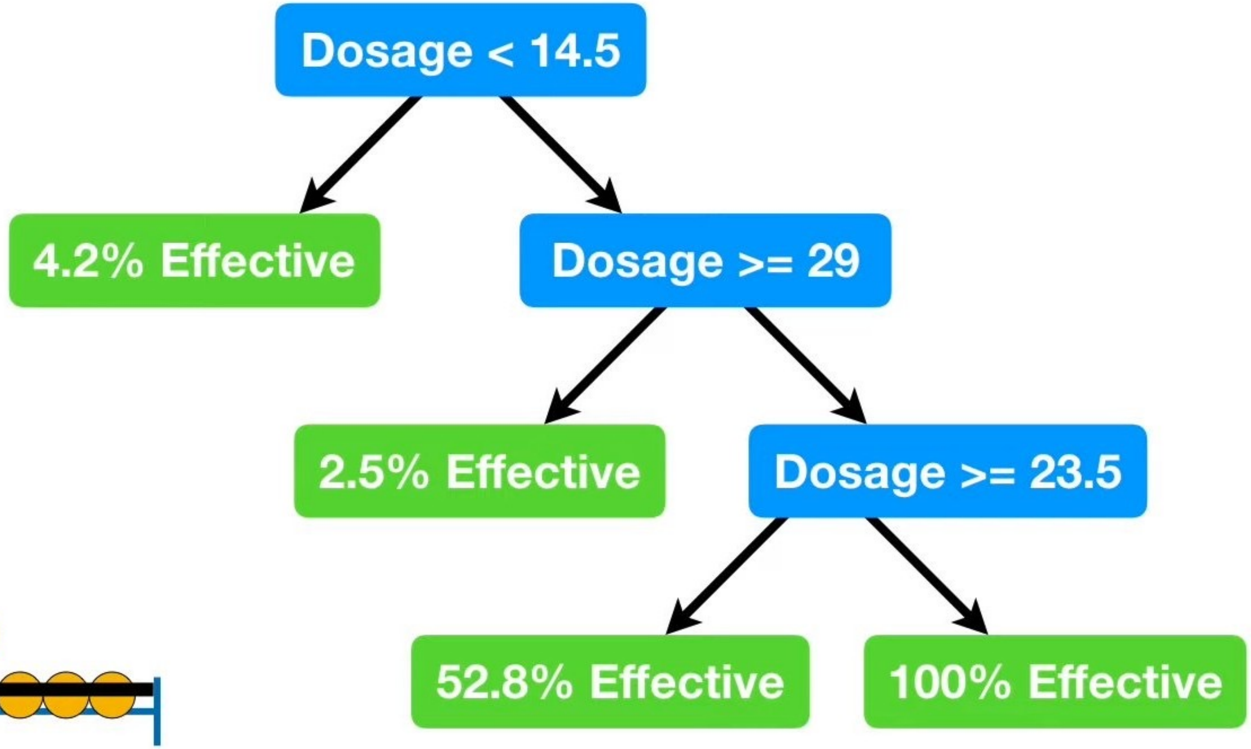


…and we use the average **Drug Effectiveness** for observations with **Dosages** between **23.5** and **29**, **52.8%**, as the output for leaf on the left.



Since no leaf has more than **7** observations in it,…




Dosage < 14.5

4.2% Effective

Dosage >= 29

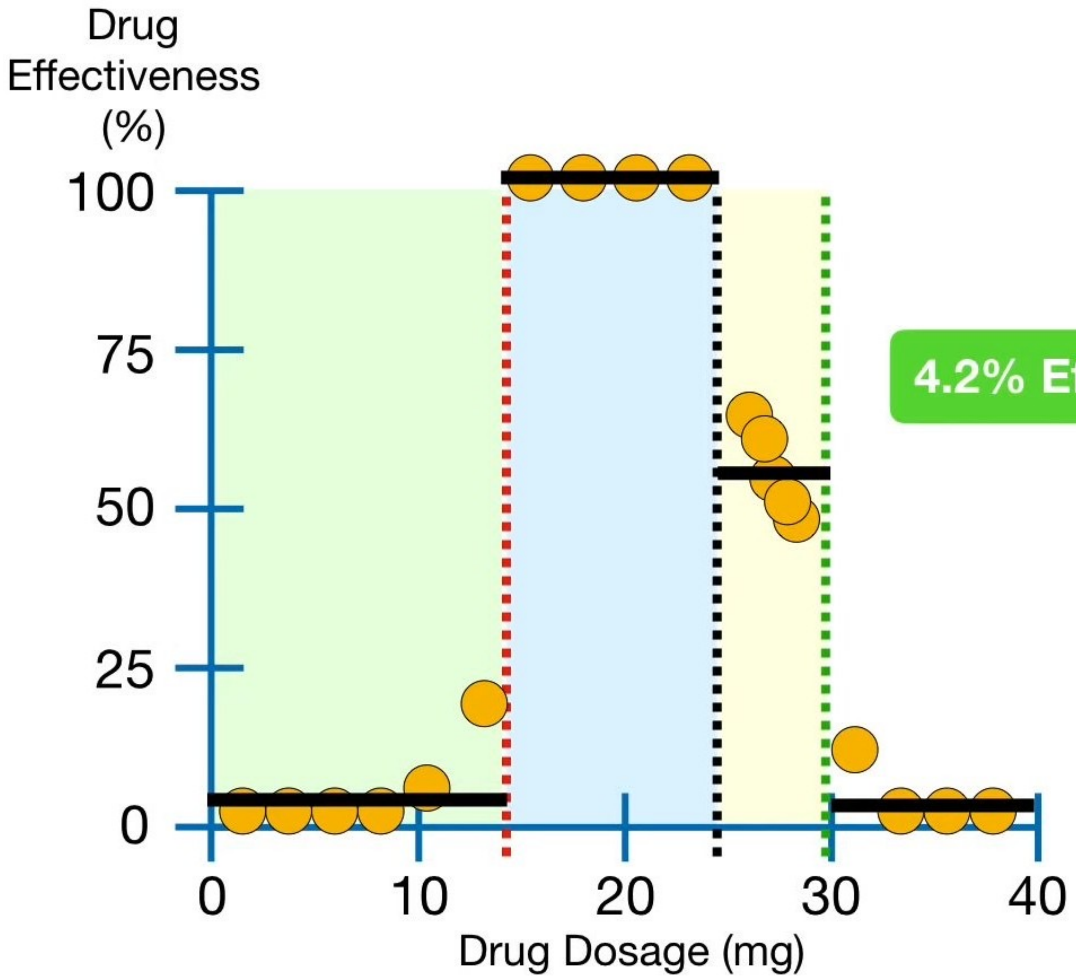
2.5% Effective

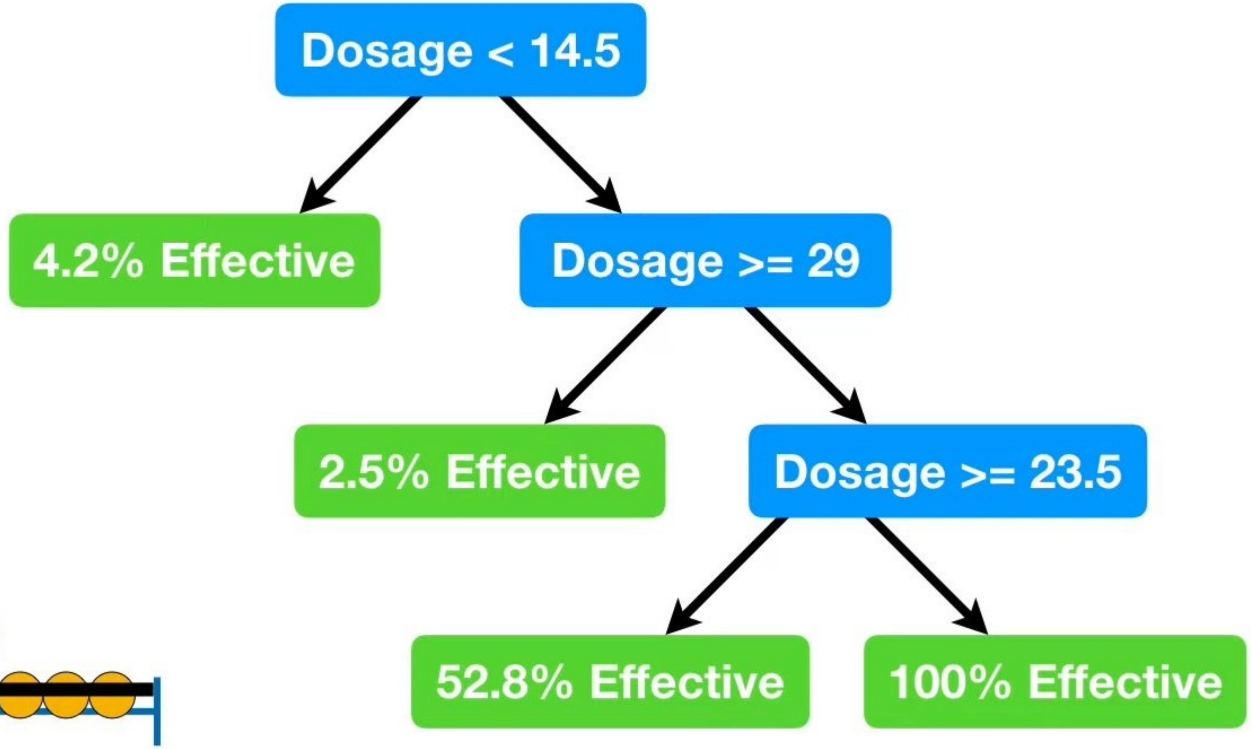
Dosage >= 23.5

52.8% Effective

100% Effective

…we're done building the tree…





Dosage < 14.5

4.2% Effective

Dosage >= 29

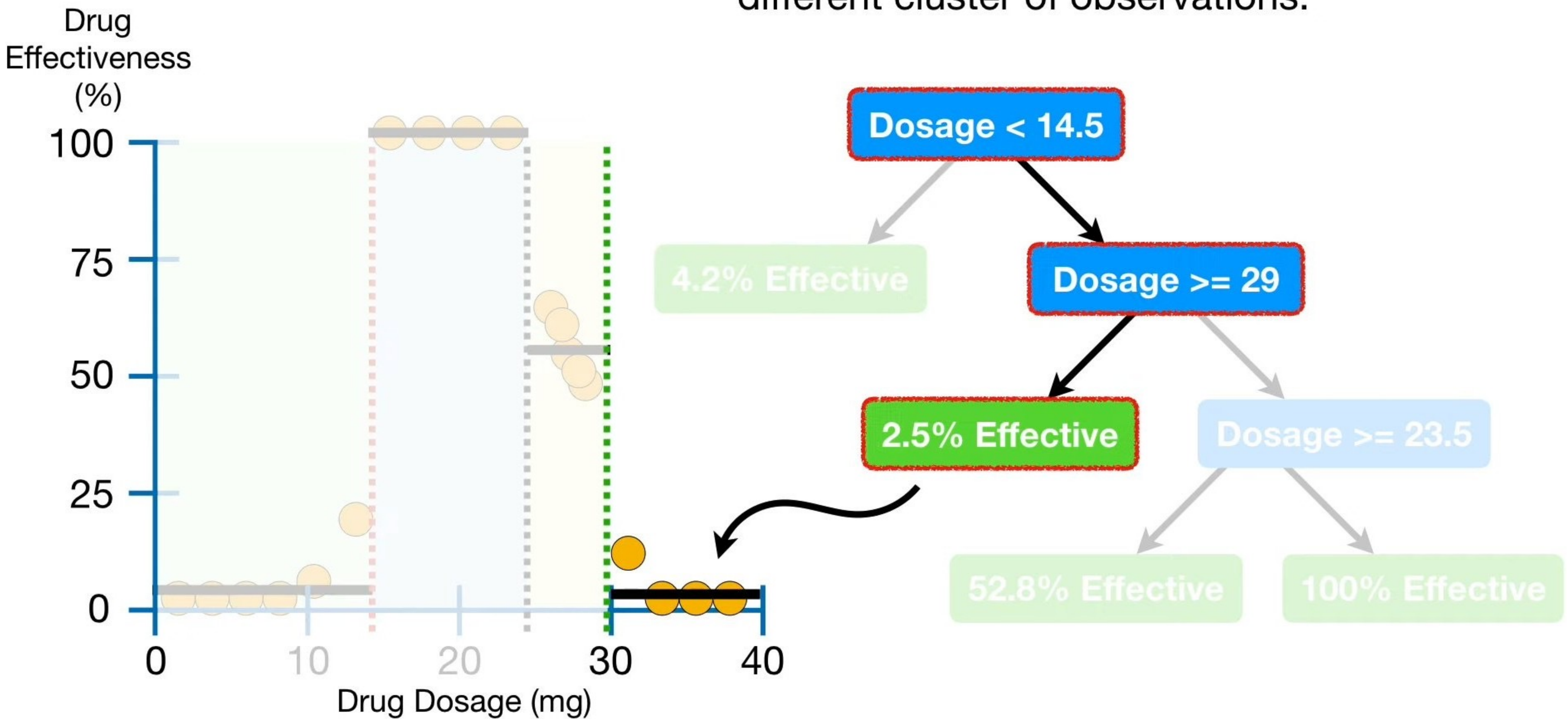
2.5% Effective

Dosage >= 23.5

52.8% Effective

100% Effective

…and each leaf corresponds to the average **Drug Effectiveness** from a different cluster of observations.



Dosage < 14.5

4.2% Effective

Dosage >= 29

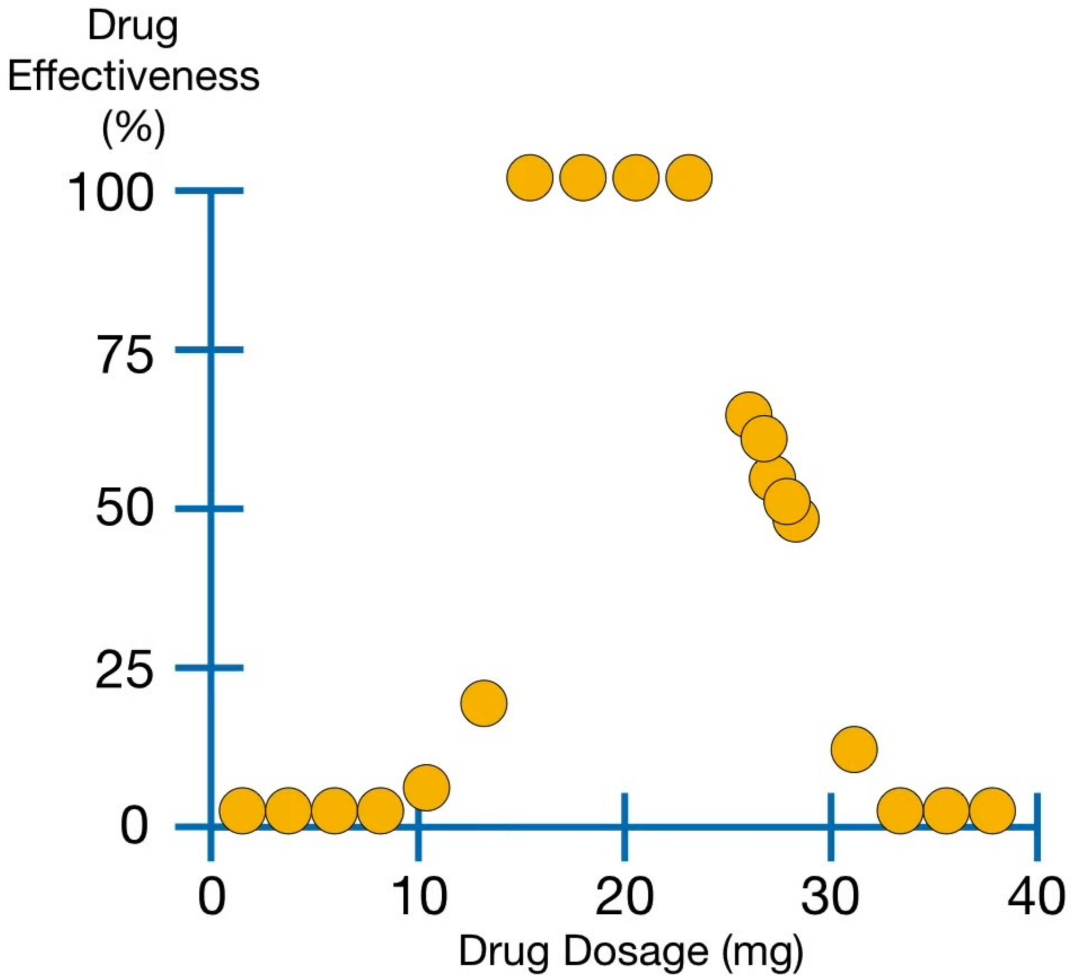
2.5% Effective

Dosage >= 23.5

52.8% Effective

100% Effective

So far we have built a tree using a single predictor, **Dosage**, to predict **Drug Effectiveness**.



Dosage	Drug Effect.
10	58
20	60
35	57
5	44
etc…	etc…

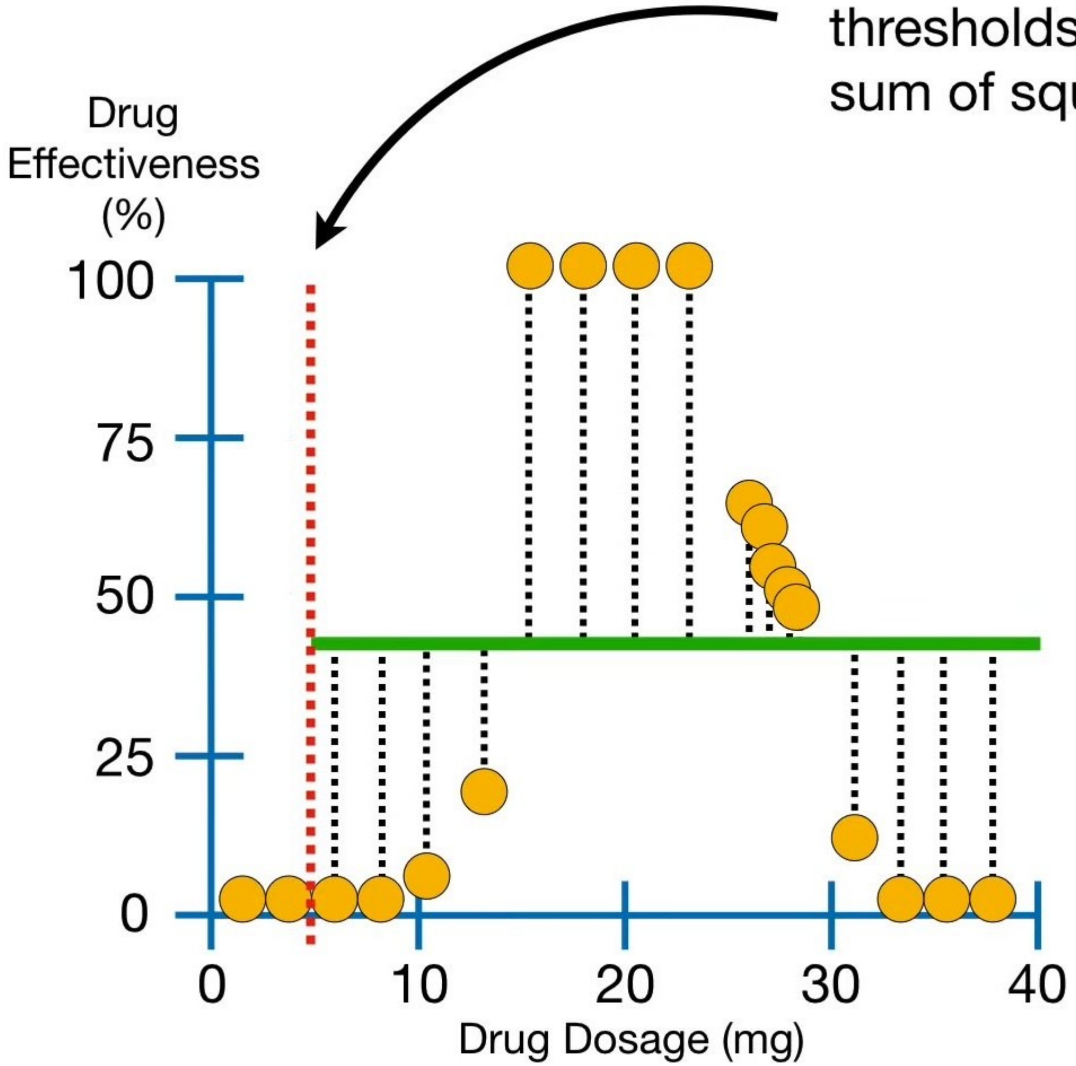
Now let’s talk about how to build a tree to predict **Drug Effectiveness** using a bunch of predictors.

Dosage	Age	Sex	Drug Effect.
10	25	Female	98
20	73	Male	0
35	54	Female	6
5	12	Male	44
etc…	etc…	etc…	etc…

Just like before, we will start by using **Dosage** to predict **Drug Effectiveness**.

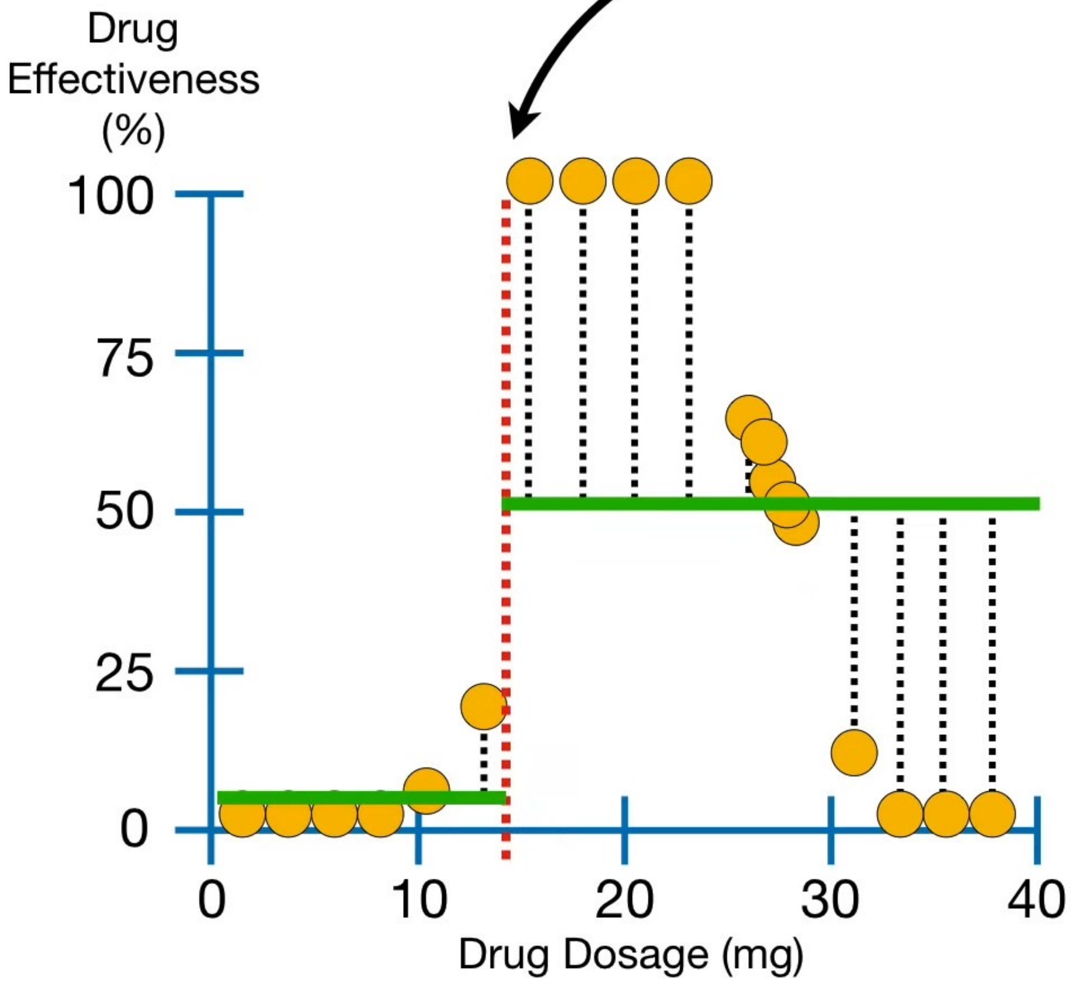
Dosage	Age	Sex	Drug Effect.
10	25	Female	98
20	73	Male	0
35	54	Female	6
5	12	Male	44
etc…	etc…	etc…	etc…

Thus, just like before, we will try different thresholds for **Dosage** and calculate the sum of squared residuals at each step…

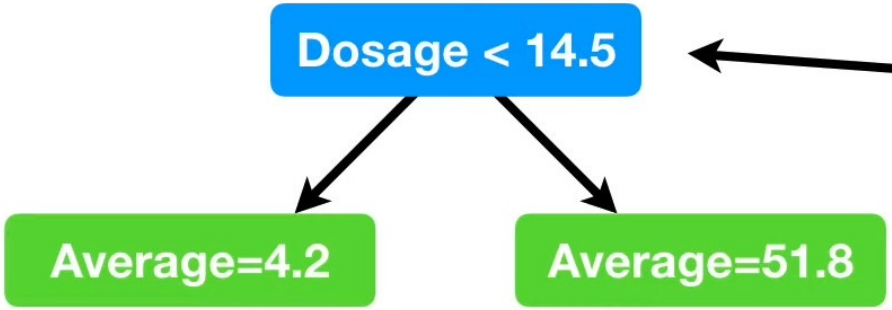


Dosage	Age	Sex	Drug Effect.
10	25	Female	98
20	73	Male	0
35	54	Female	6
5	12	Male	44
etc…	etc…	etc…	etc…

…and pick the threshold that gives us the minimum sum of squared residuals.



Dosage	Age	Sex	Drug Effect.
10	25	Female	98
20	73	Male	0
35	54	Female	6
5	12	Male	44
etc…	etc…	etc…	etc…



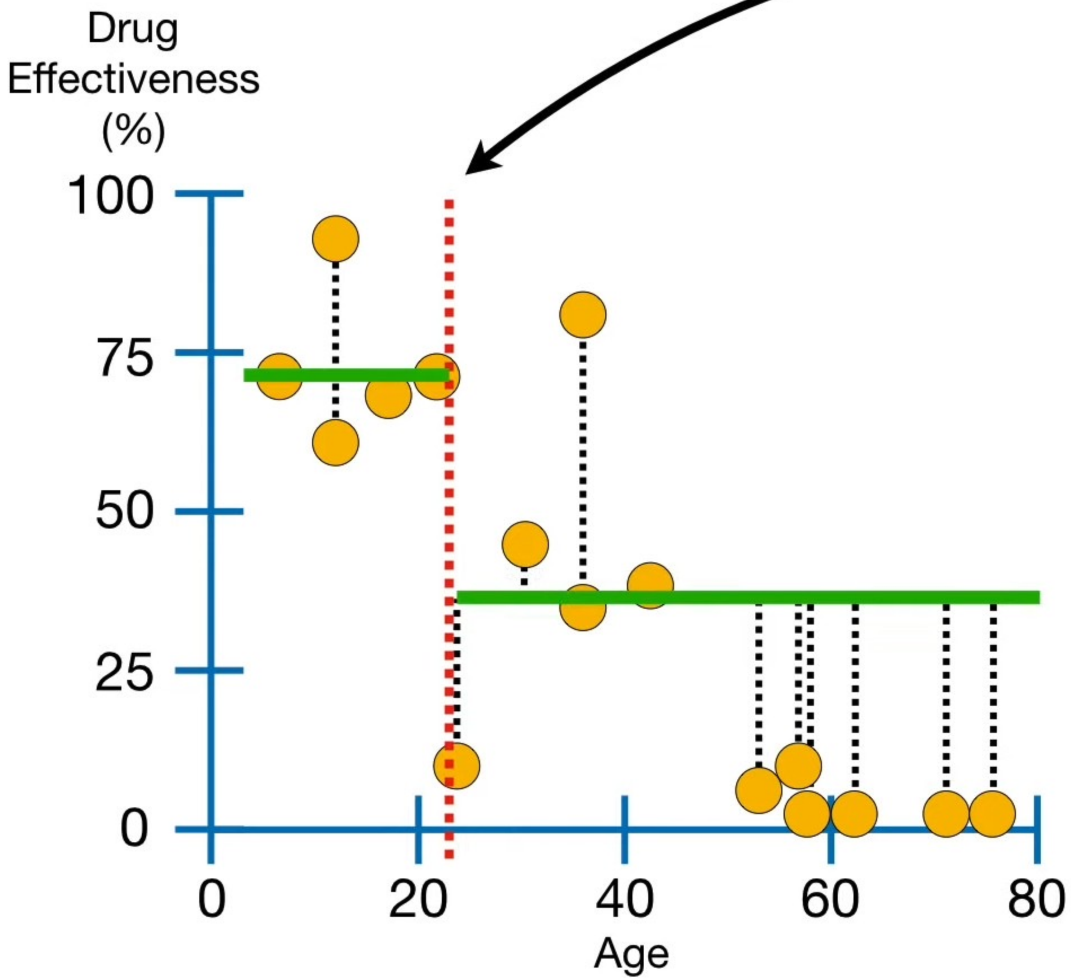
The best threshold becomes a *candidate* for the root.

Dosage	Age	Sex	Drug Effect.
10	25	Female	98
20	73	Male	0
35	54	Female	6
5	12	Male	44
etc…	etc…	etc…	etc…

Now we focus on using **Age** to predict **Drug Effectiveness**.

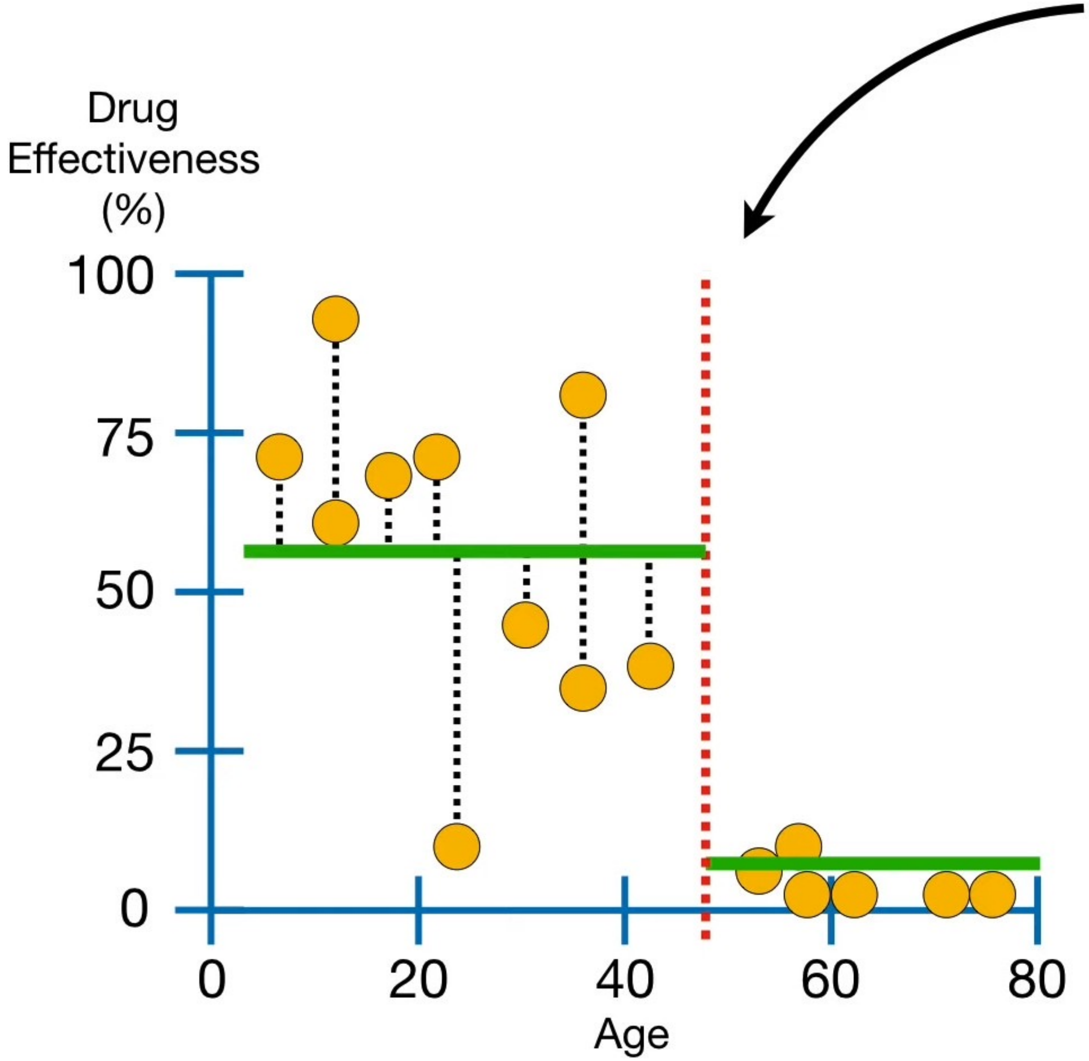
Dosage	Age	Sex	Drug Effect.
10	25	Female	98
20	73	Male	0
35	54	Female	6
5	12	Male	44
etc…	etc…	etc…	etc…

Just like with **Dosage**, we try different thresholds for **Age** and calculate the sum of squared residuals at each step…

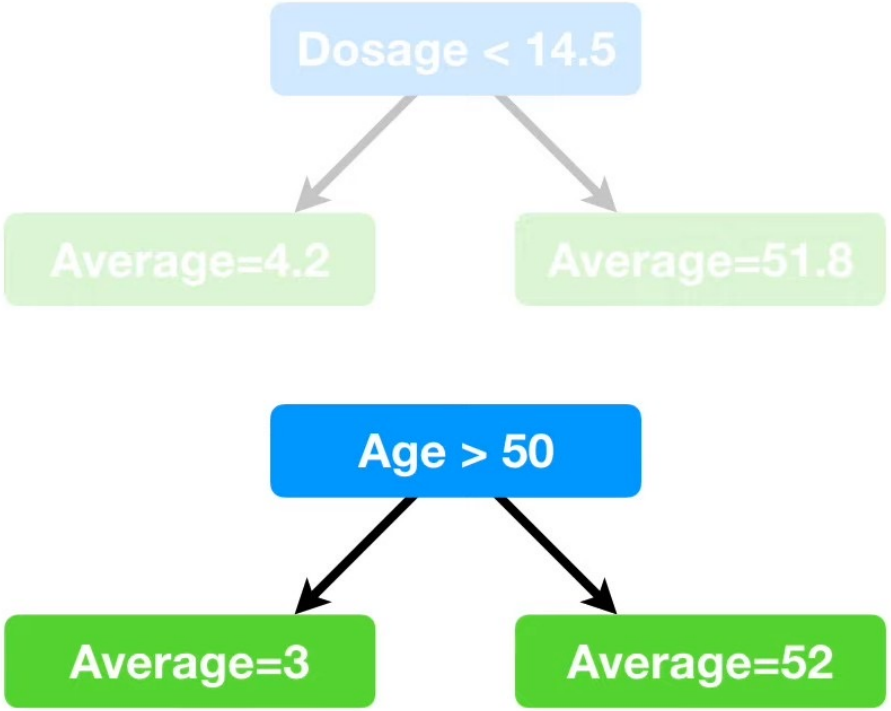


Dosage	Age	Sex	Drug Effect.
10	25	Female	98
20	73	Male	0
35	54	Female	6
5	12	Male	44
etc…	etc…	etc…	etc…

…and pick the one that gives us the minimum sum of squared residuals.



Dosage	Age	Sex	Drug Effect.
10	25	Female	98
20	73	Male	0
35	54	Female	6
5	12	Male	44
etc…	etc…	etc…	etc…

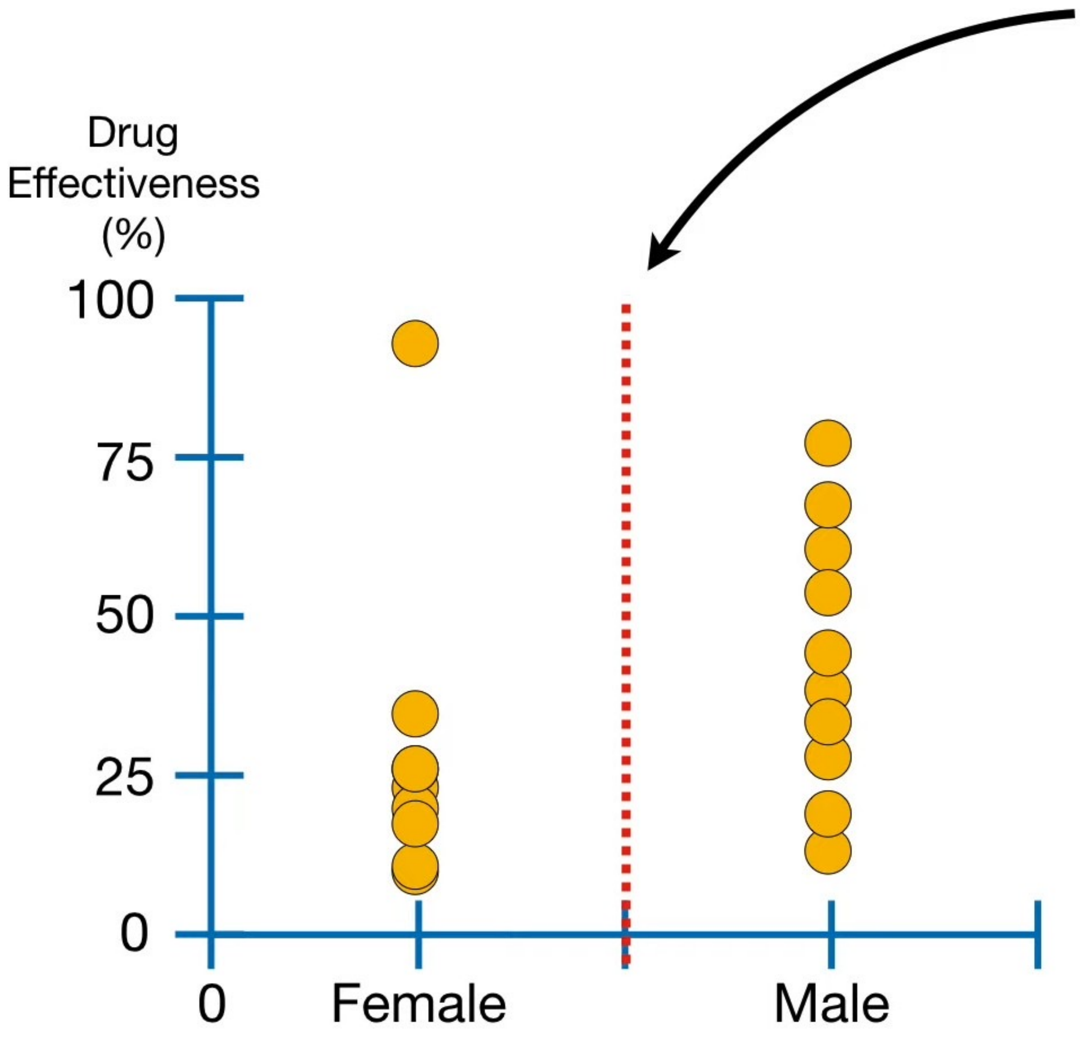


The best threshold becomes another *candidate* for the root.

Dosage	Age	Sex	Drug Effect.
10	25	Female	98
20	73	Male	0
35	54	Female	6
5	12	Male	44
etc…	etc…	etc…	etc…

Now we focus on using **Sex** to predict **Drug Effectiveness**.

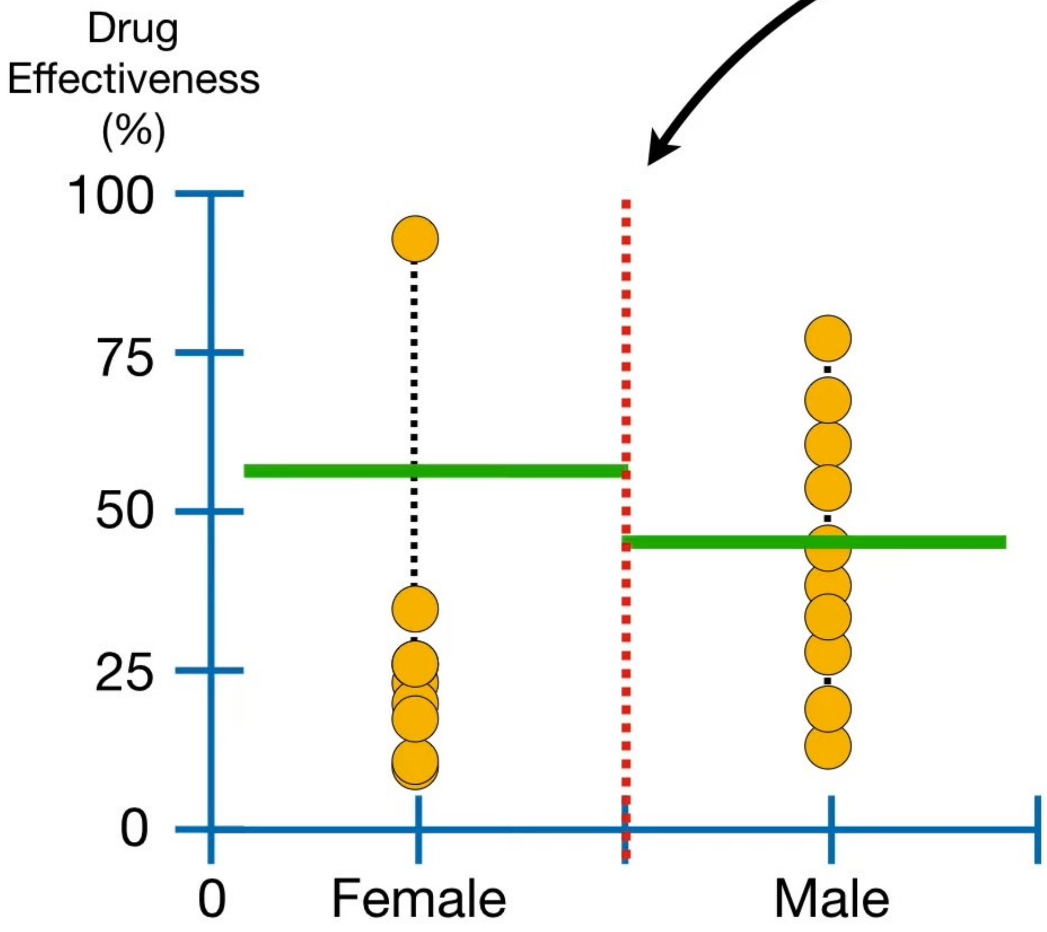
Dosage	Age	Sex	Drug Effect.
10	25	Female	98
20	73	Male	0
35	54	Female	6
5	12	Male	44
etc…	etc…	etc…	etc…



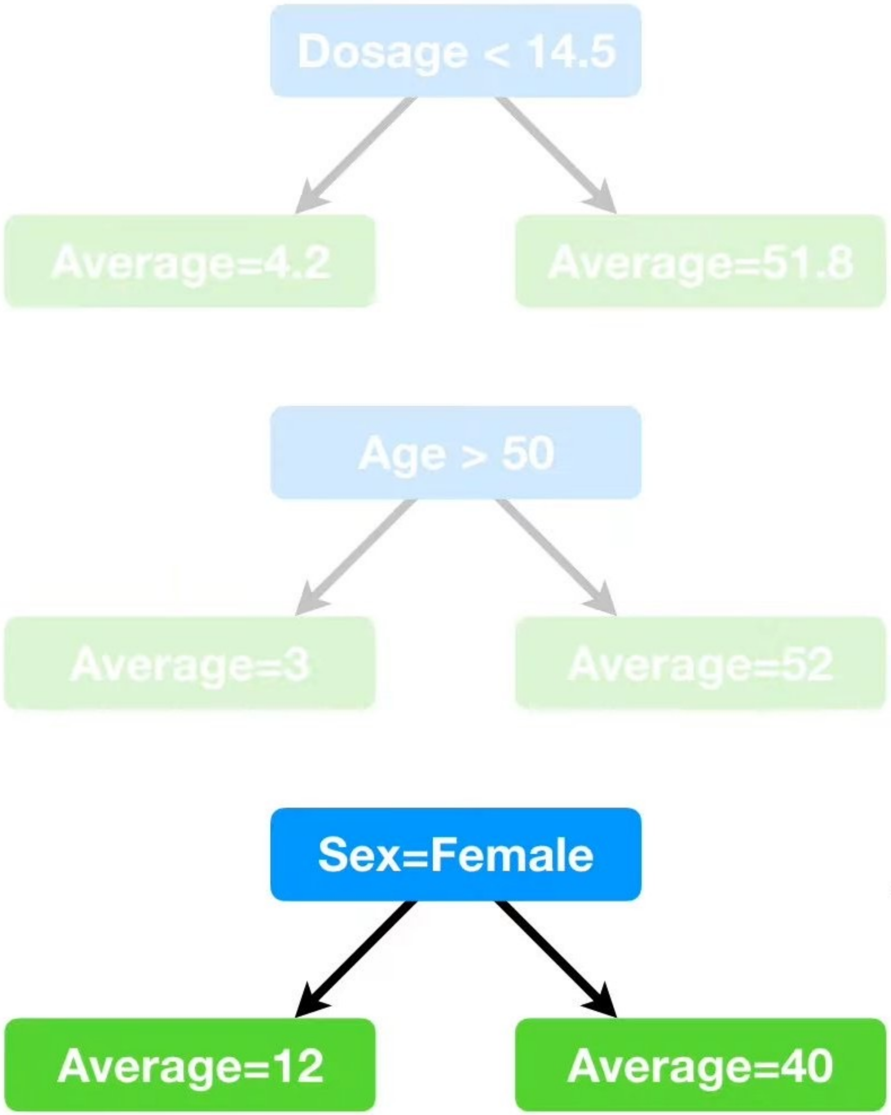
With **Sex**, there is only one threshold to try…

Dosage	Age	Sex	Drug Effect.
10	25	Female	98
20	73	Male	0
35	54	Female	6
5	12	Male	44
etc…	etc…	etc…	etc…

…so we use that threshold to calculate the sum of squared residuals…

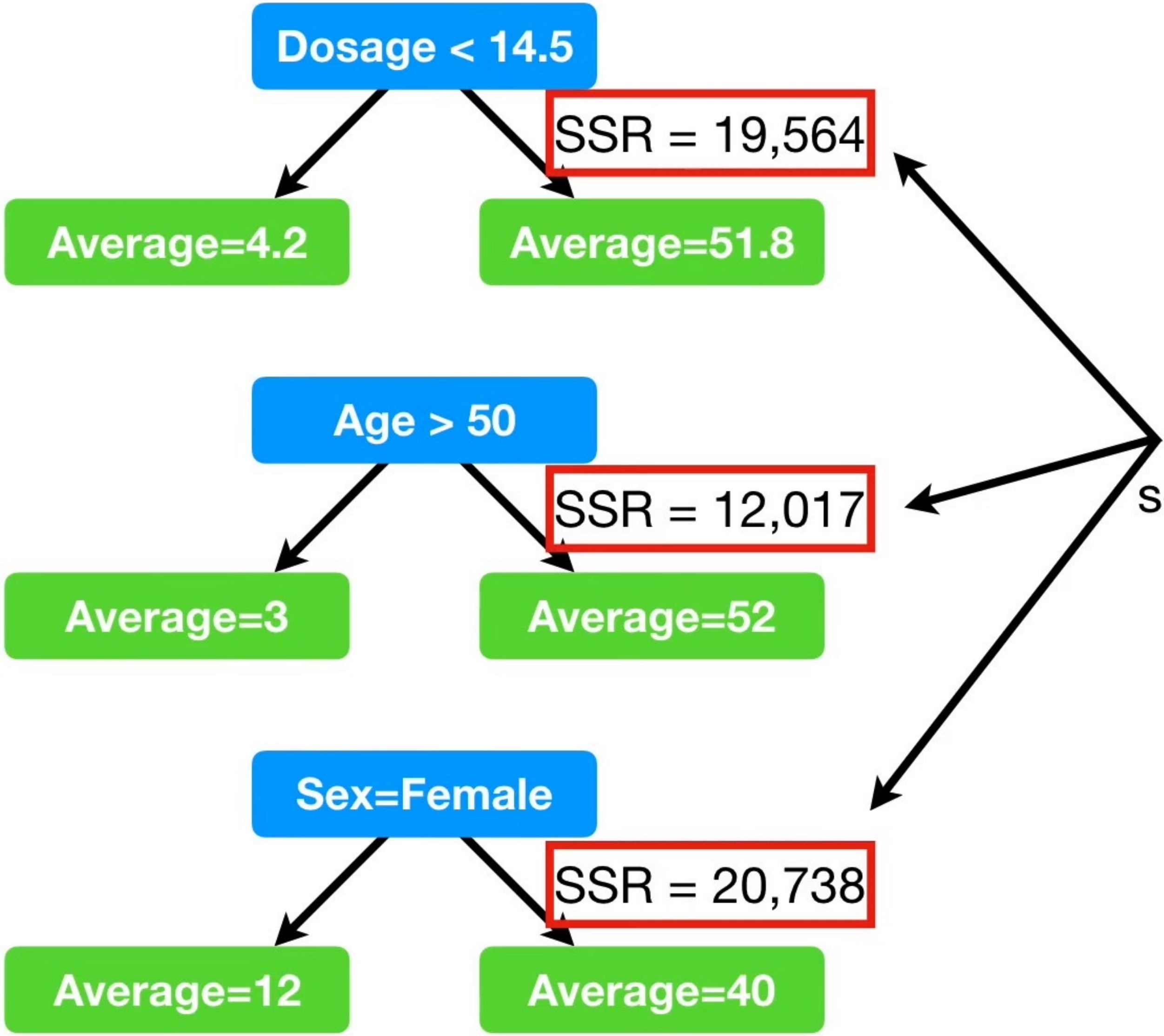


Dosage	Age	Sex	Drug Effect.
10	25	Female	98
20	73	Male	0
35	54	Female	6
5	12	Male	44
etc…	etc…	etc…	etc…



…and that becomes another *candidate* for the root.

Dosage	Age	Sex	Drug Effect.
10	25	Female	98
20	73	Male	0
35	54	Female	6
5	12	Male	44
etc…	etc…	etc…	etc…



Dosage < 14.5

Average=4.2

Average=51.8

SSR = 19,564

Age > 50

Average=3

Average=52

SSR = 12,017

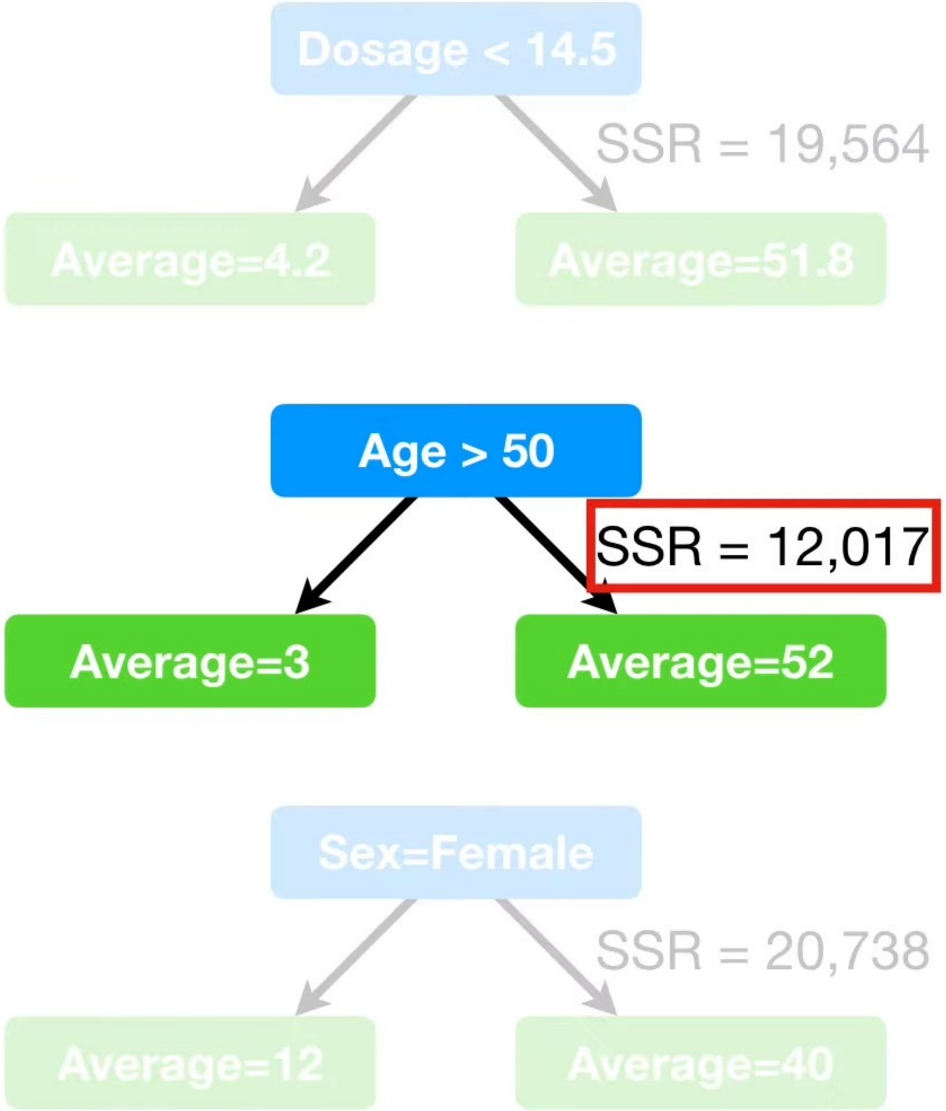
Sex=Female

Average=12

Average=40

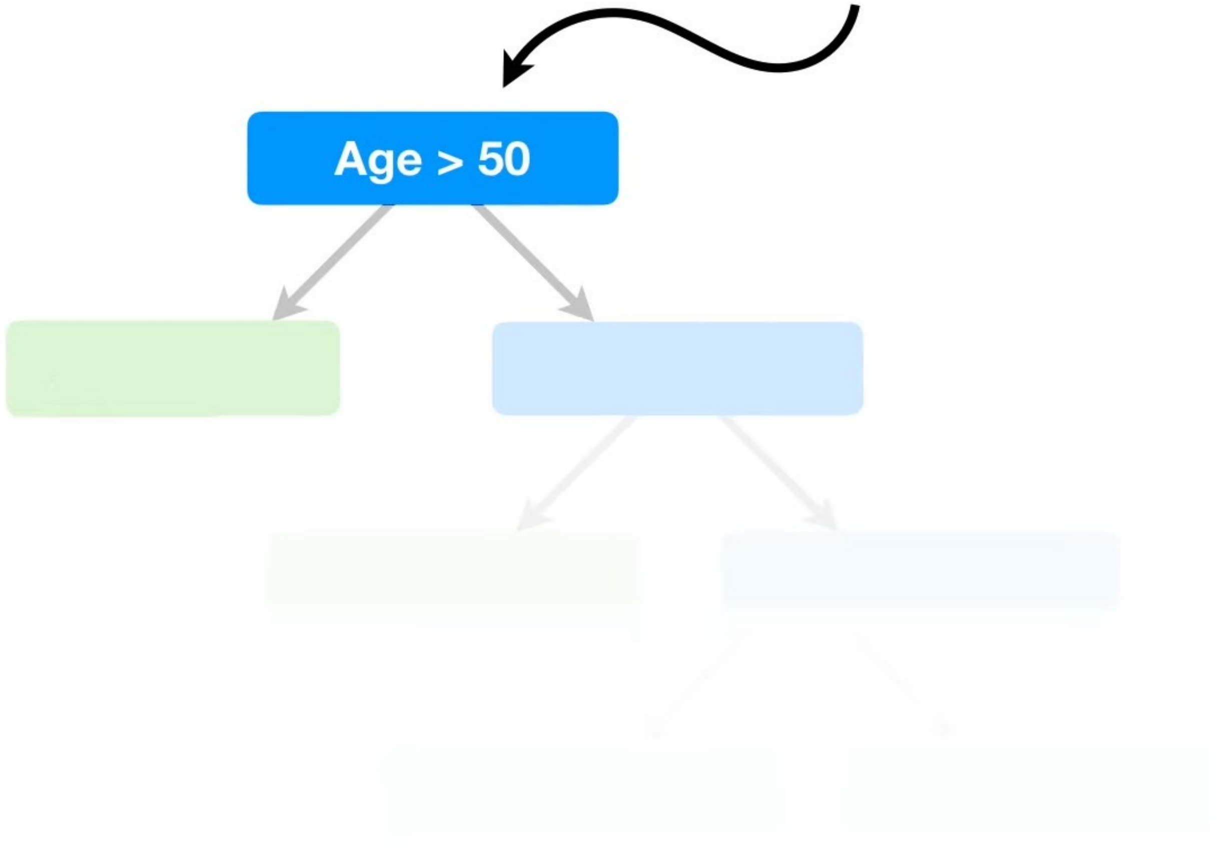
SSR = 20,738

Now we compare the sum of squared residuals (SSRs) for each candidate…



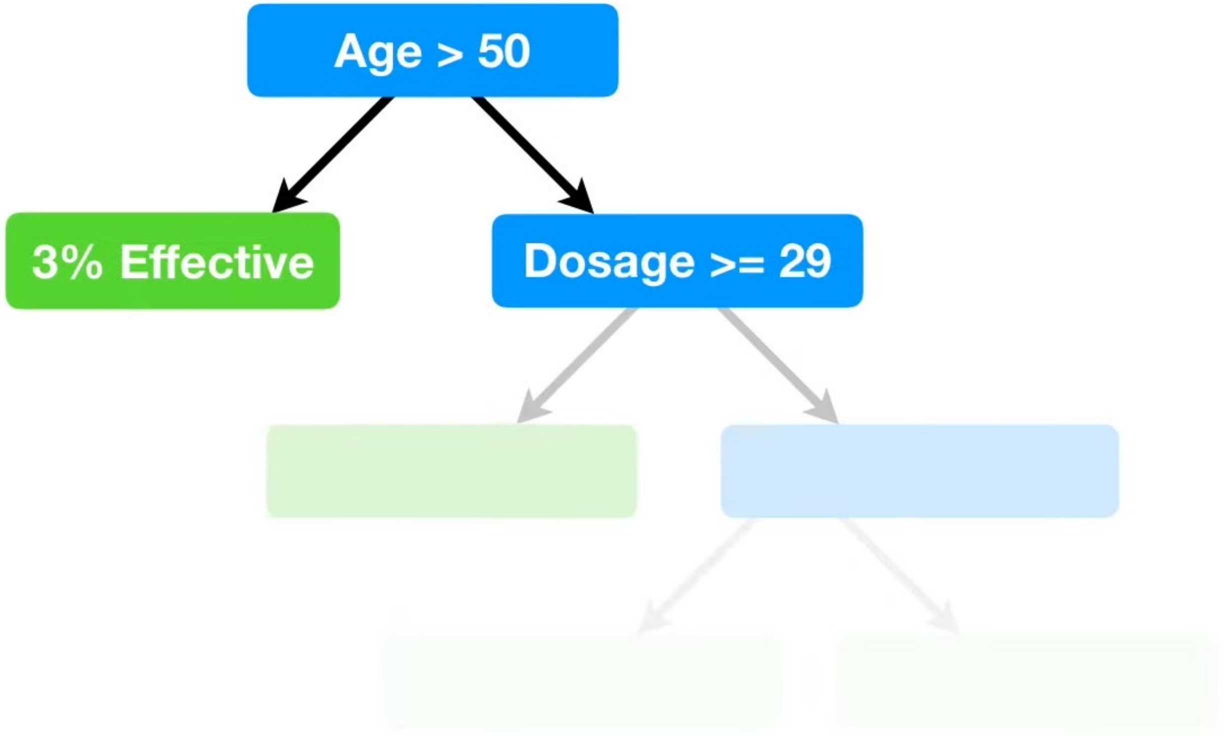
…and pick the candidate with the lowest value.

Since **Age > 50** had the lowest sum of squared residuals, it becomes the root of the tree.



Dosage	Age	Sex	Drug Effect.
10	25	Female	98
20	73	Male	0
35	54	Female	6
5	12	Male	44
etc…	etc…	etc…	etc…

Then we grow the tree just like before, except now we compare the lowest sum of squared residuals from each predictor.



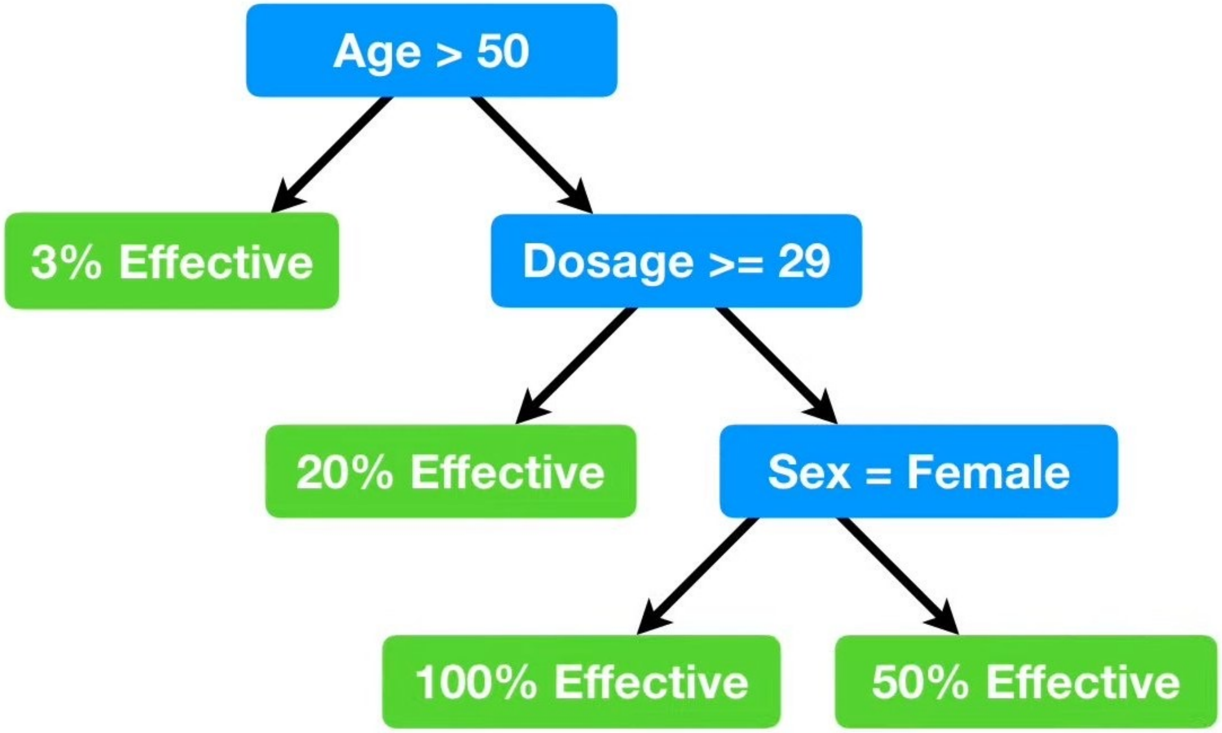
Age > 50

3% Effective

Dosage >= 29

Dosage	Age	Sex	Drug Effect.
10	25	Female	98
20	73	Male	0
35	54	Female	6
5	12	Male	44
etc…	etc…	etc…	etc…

Then we grow the tree just like before, except now we compare the lowest sum of squared residuals from each predictor.



Age > 50

3% Effective

Dosage >= 29

20% Effective

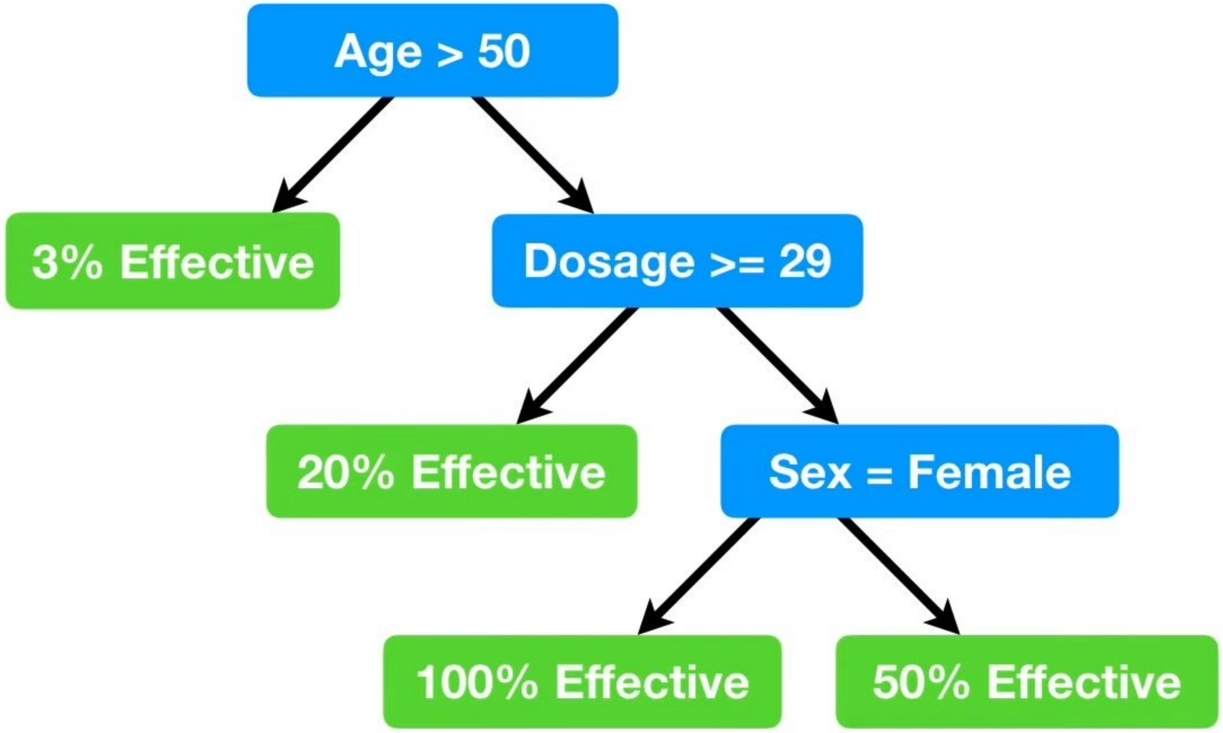
Sex = Female

100% Effective

50% Effective

Dosage	Age	Sex	Drug Effect.
10	25	Female	98
20	73	Male	0
35	54	Female	6
5	12	Male	44
etc…	etc…	etc…	etc…

And just like before, when a leaf has less than a minimum number of observations, which is usually **20**, but we are using **7**, we stop trying to divide them.



Age > 50

3% Effective

Dosage >= 29

20% Effective

Sex = Female

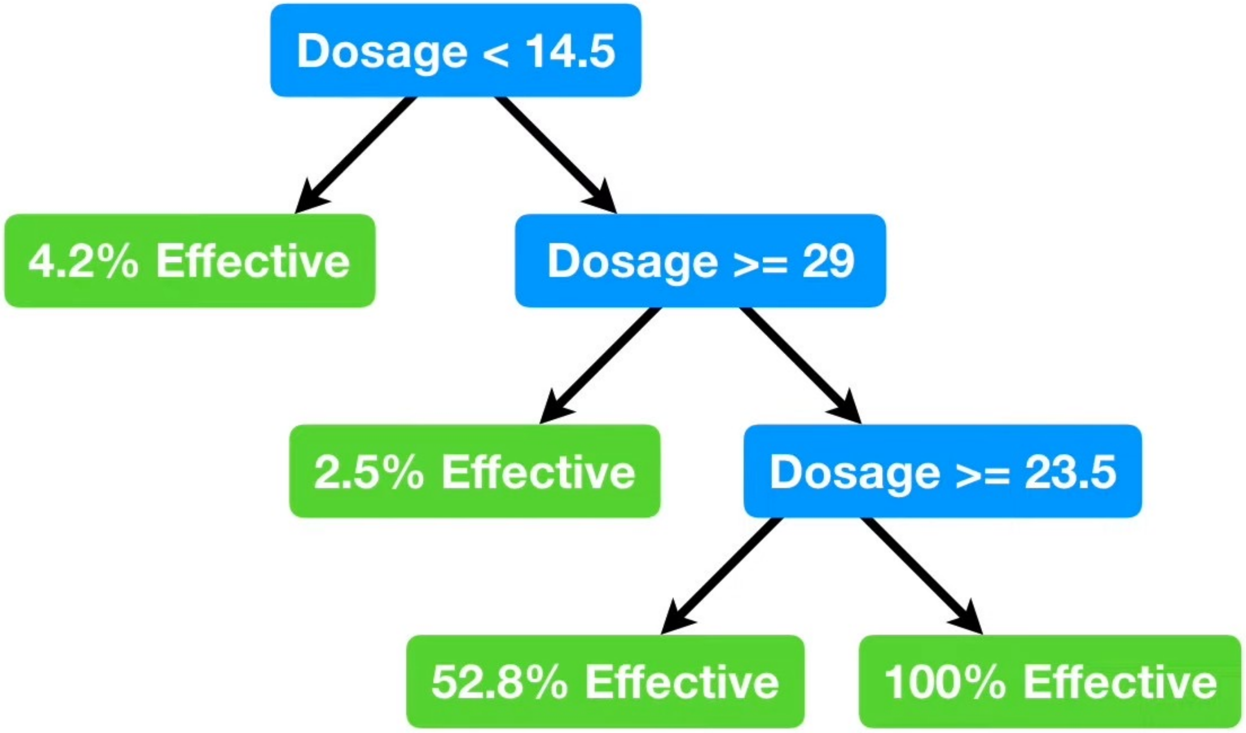
100% Effective

50% Effective

Dosage	Age	Sex	Drug Effect.
10	25	Female	98
20	73	Male	0
35	54	Female	6
5	12	Male	44
etc…	etc…	etc…	etc…

In summary…

Regression Trees are a type of **Decision Tree.**


Dosage < 14.5

4.2% Effective

Dosage >= 29

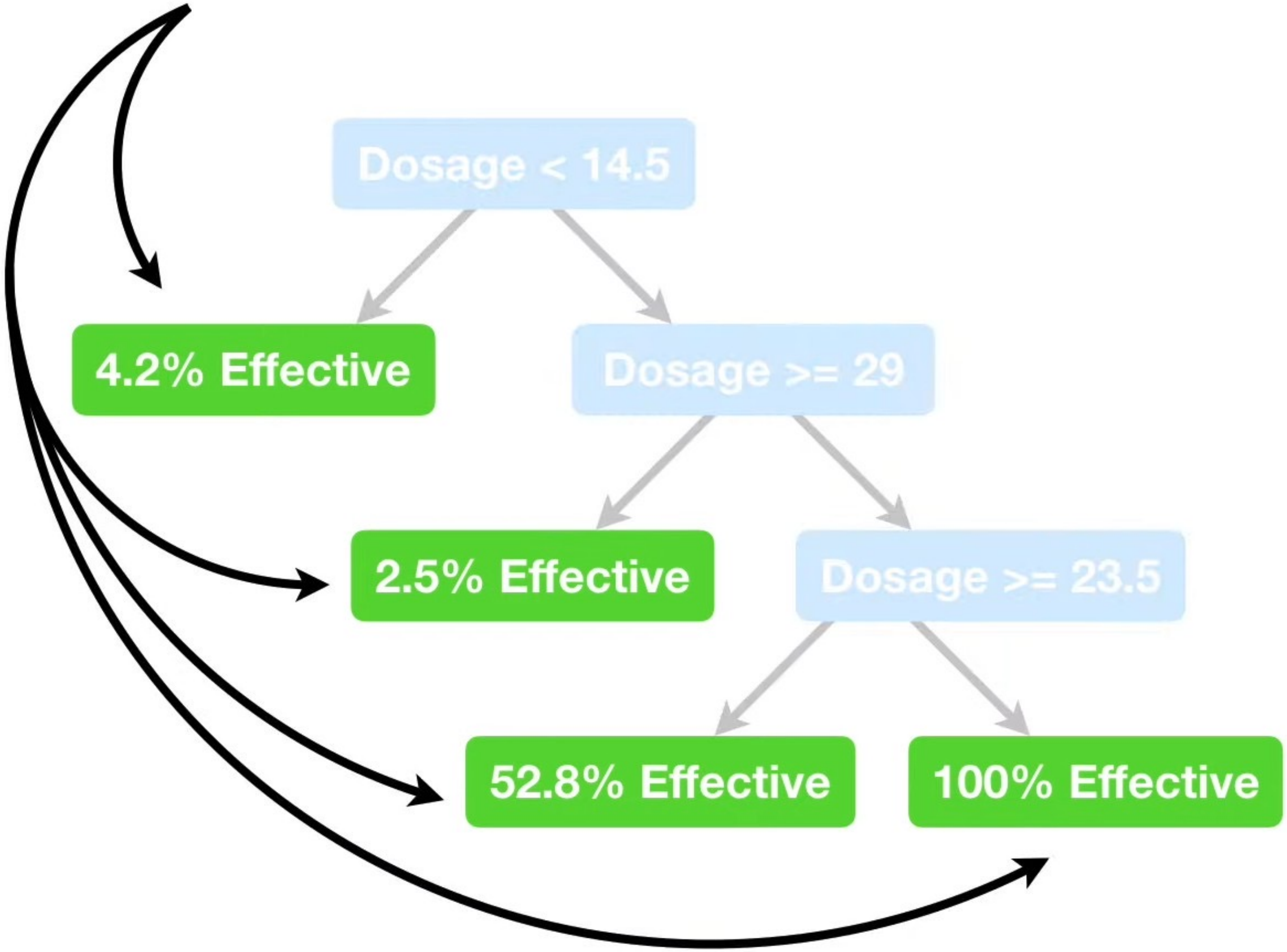
2.5% Effective

Dosage >= 23.5

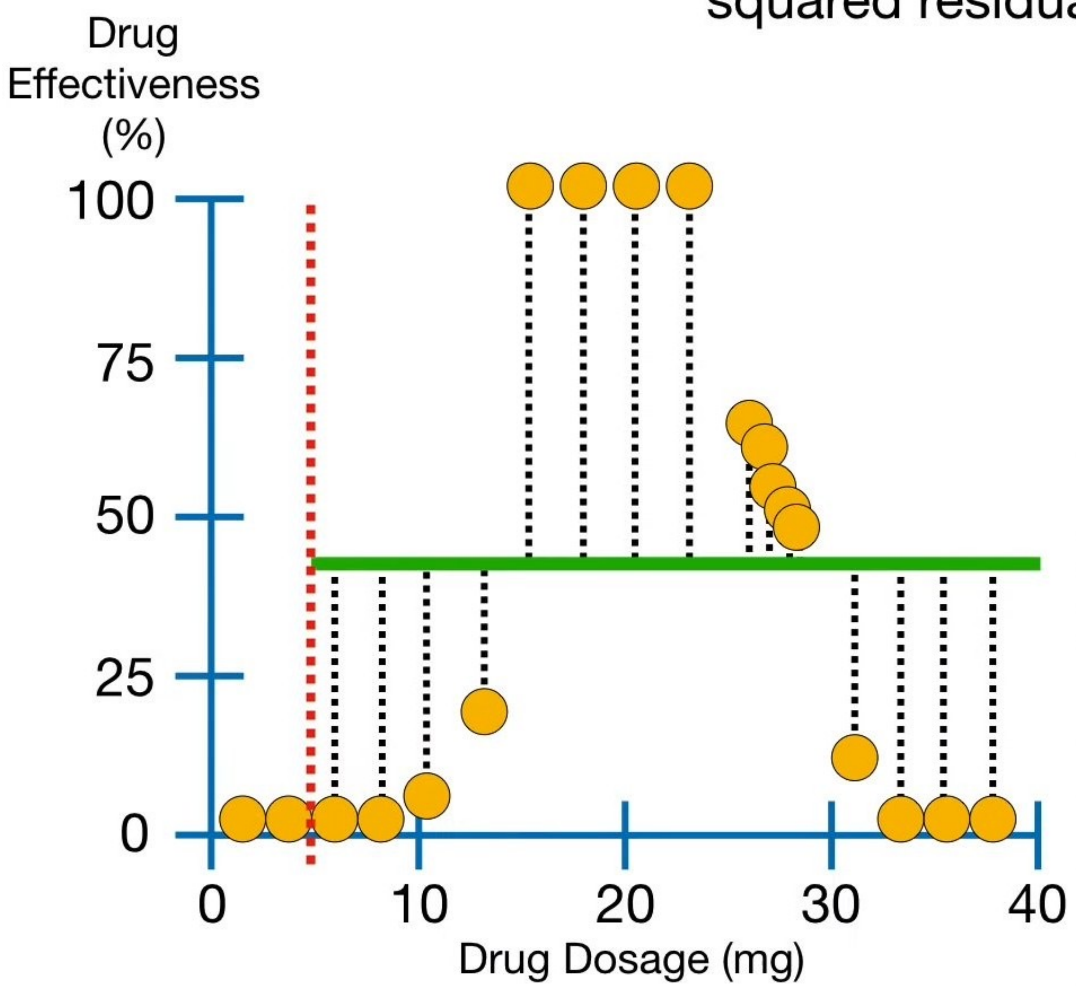
52.8% Effective

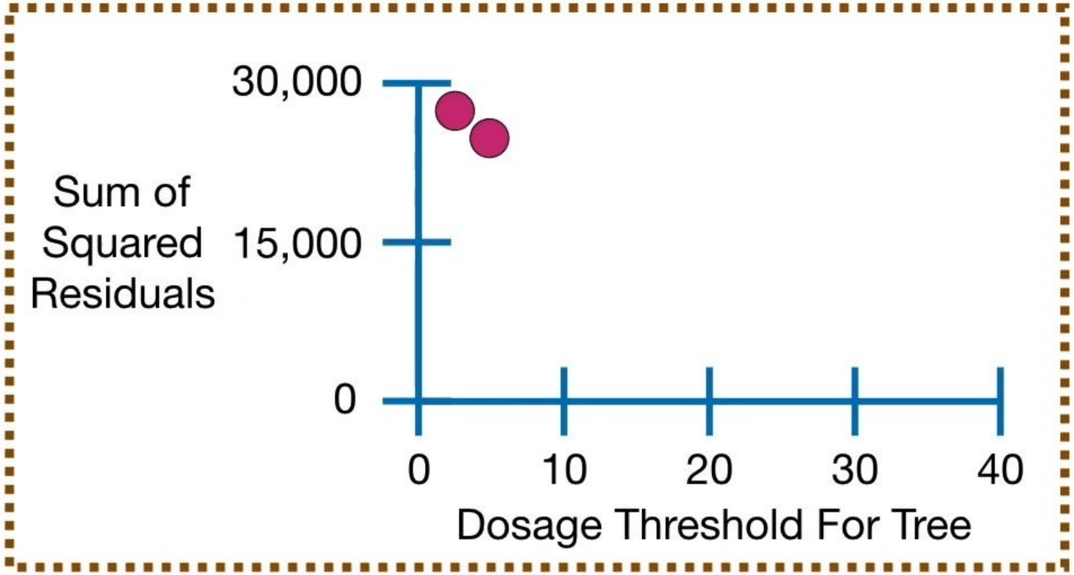
100% Effective

In a **Regression Tree**, each leaf represents a numeric value.

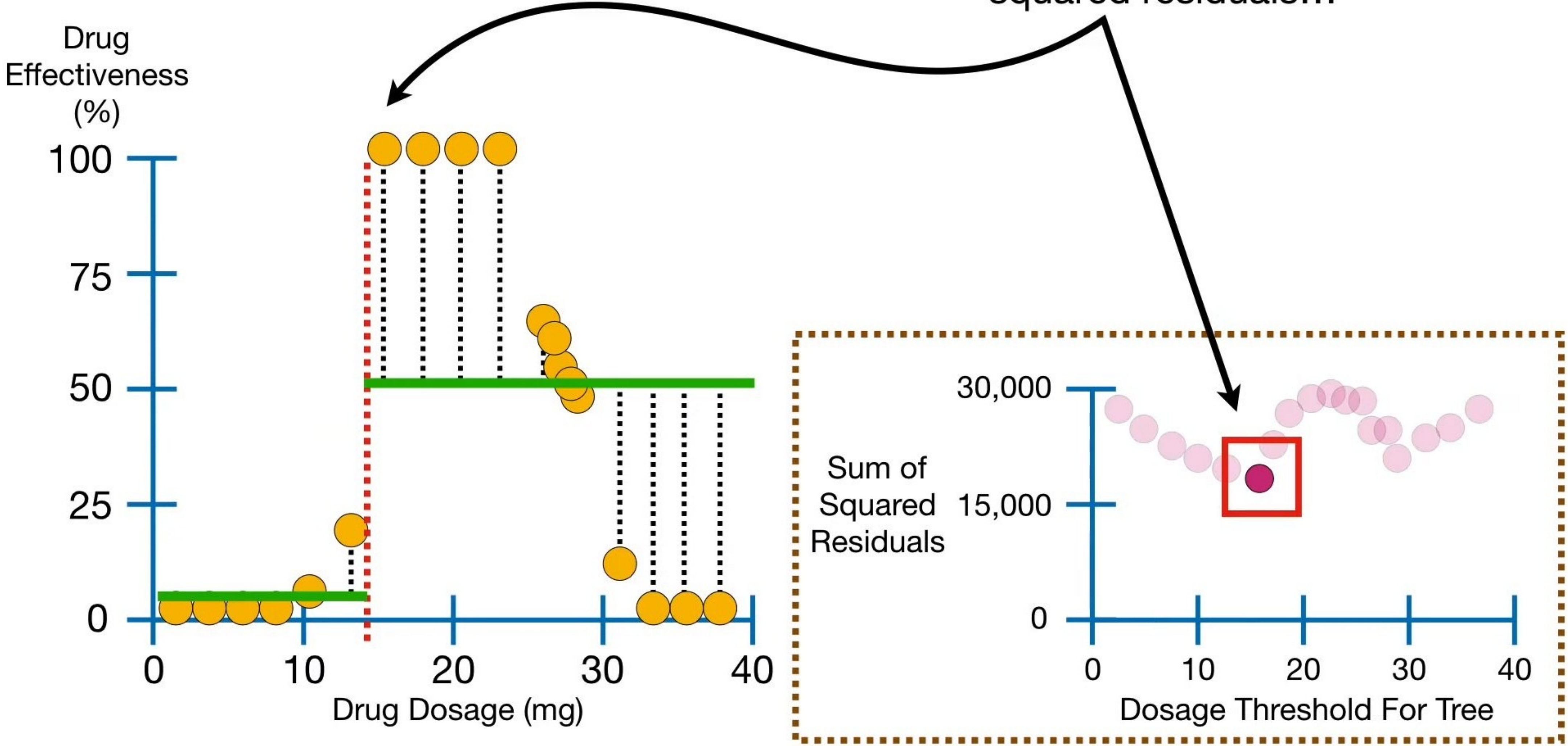


We determine how to divide the observations by trying different thresholds and calculating the sum of squared residuals at each step.

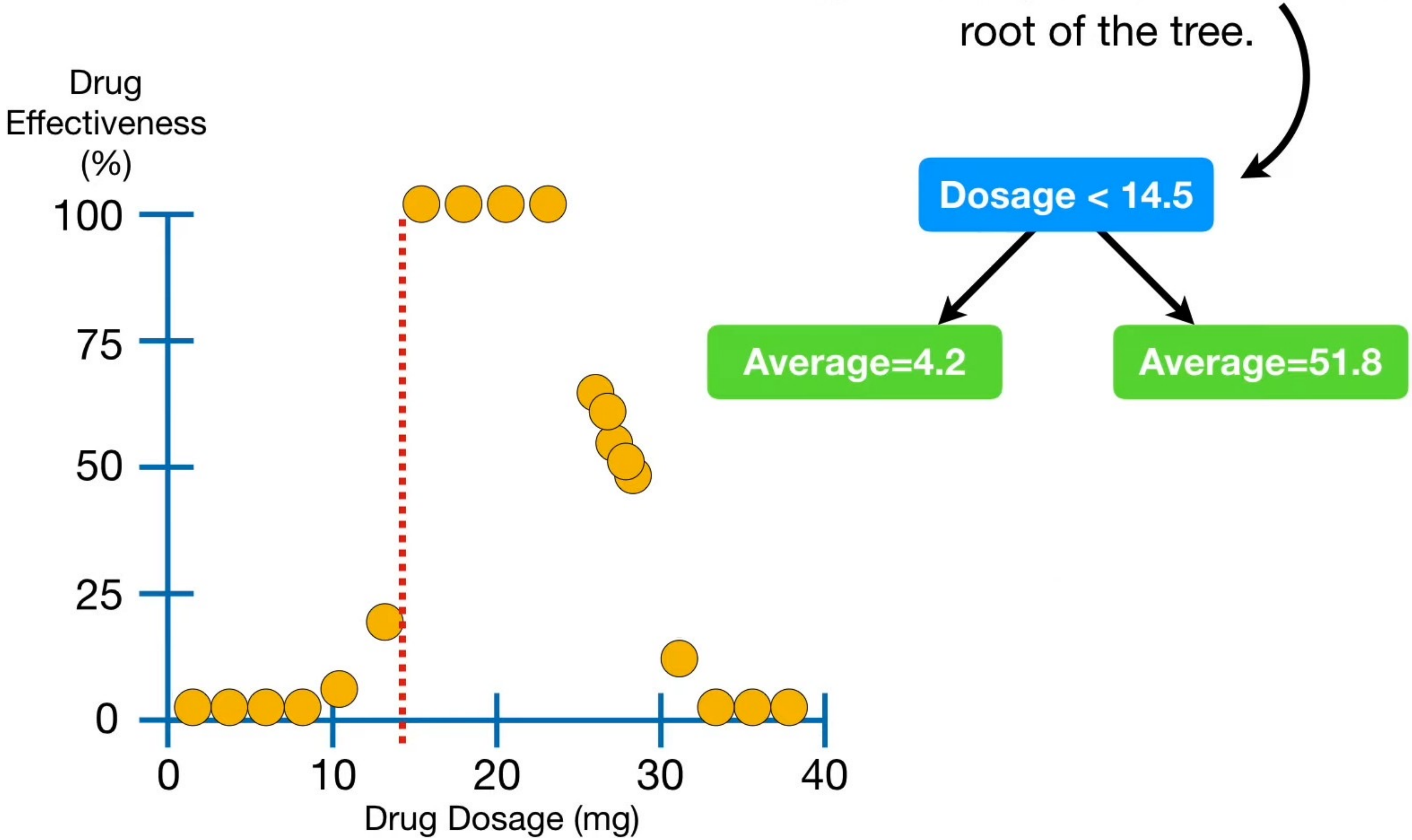




The threshold with the smallest sum of squared residuals…



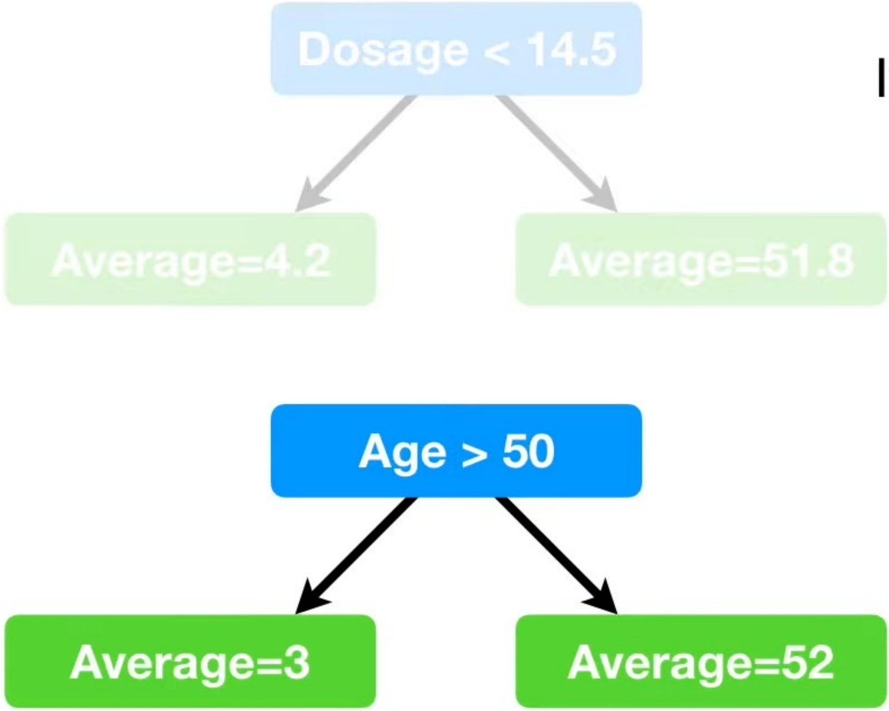
…becomes a candidate for the root of the tree.



Dosage < 14.5

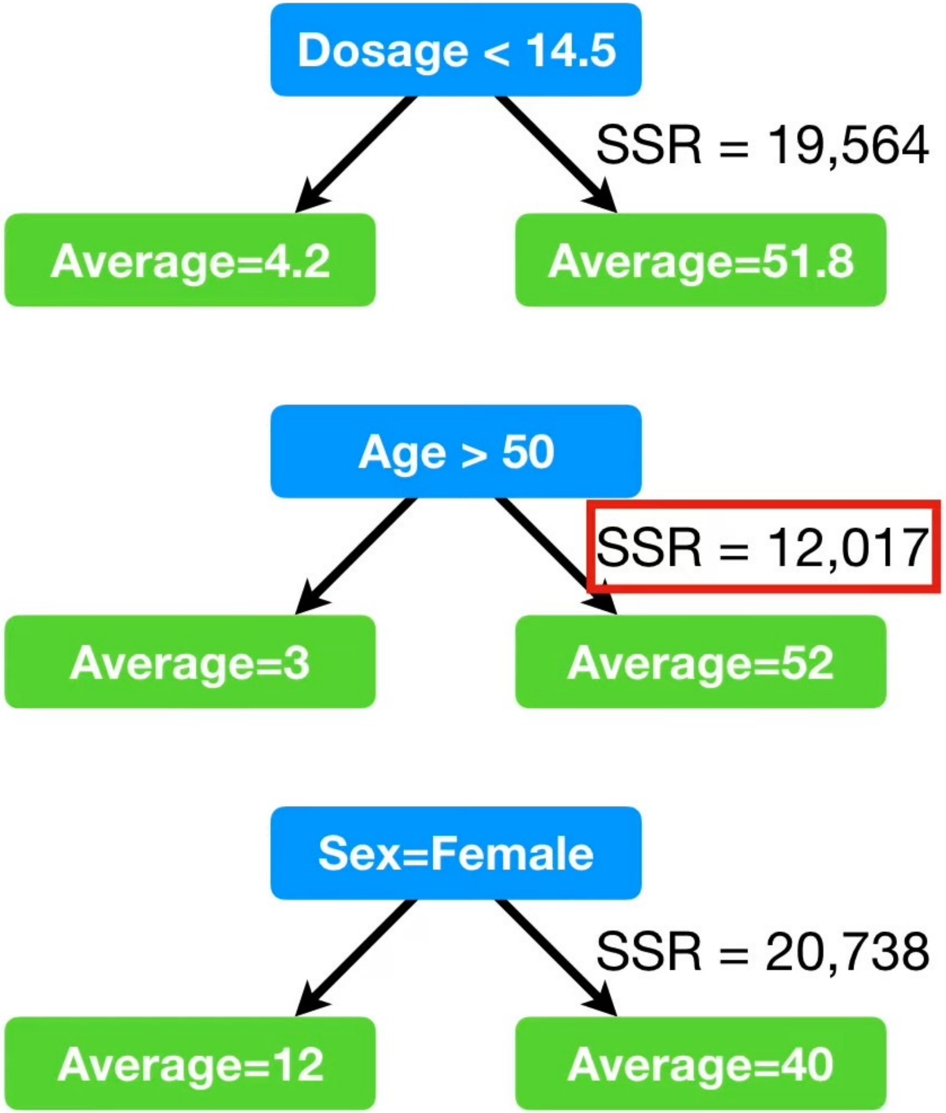
Average=4.2

Average=51.8



If we have more than one predictor, we find the optimal threshold for each one…

Dosage	Age	Sex	Drug Effect.
10	25	Female	98
20	73	Male	0
35	54	Female	6
5	12	Male	44
etc…	etc…	etc…	etc…



Dosage < 14.5

SSR = 19,564

Average=4.2

Average=51.8

Age > 50

SSR = 12,017

Average=3

Average=52

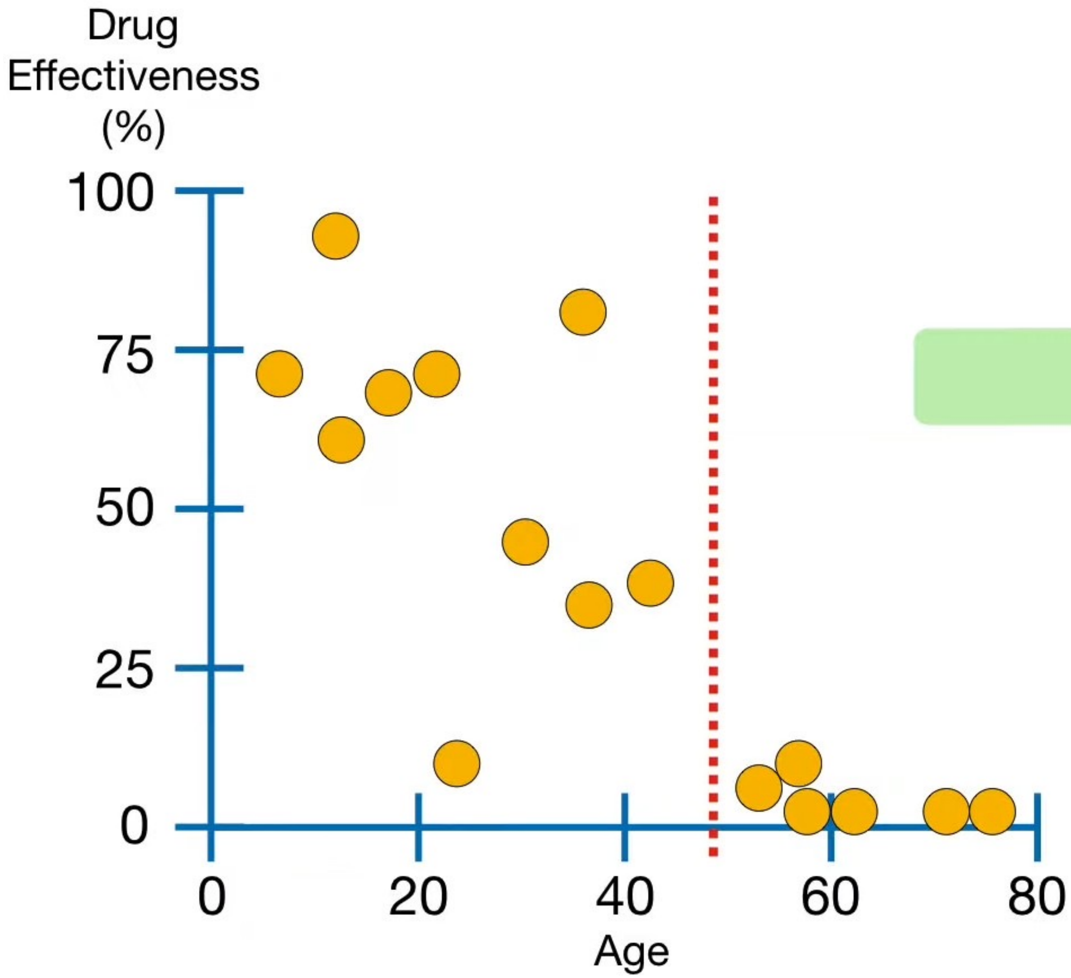
Sex=Female

SSR = 20,738

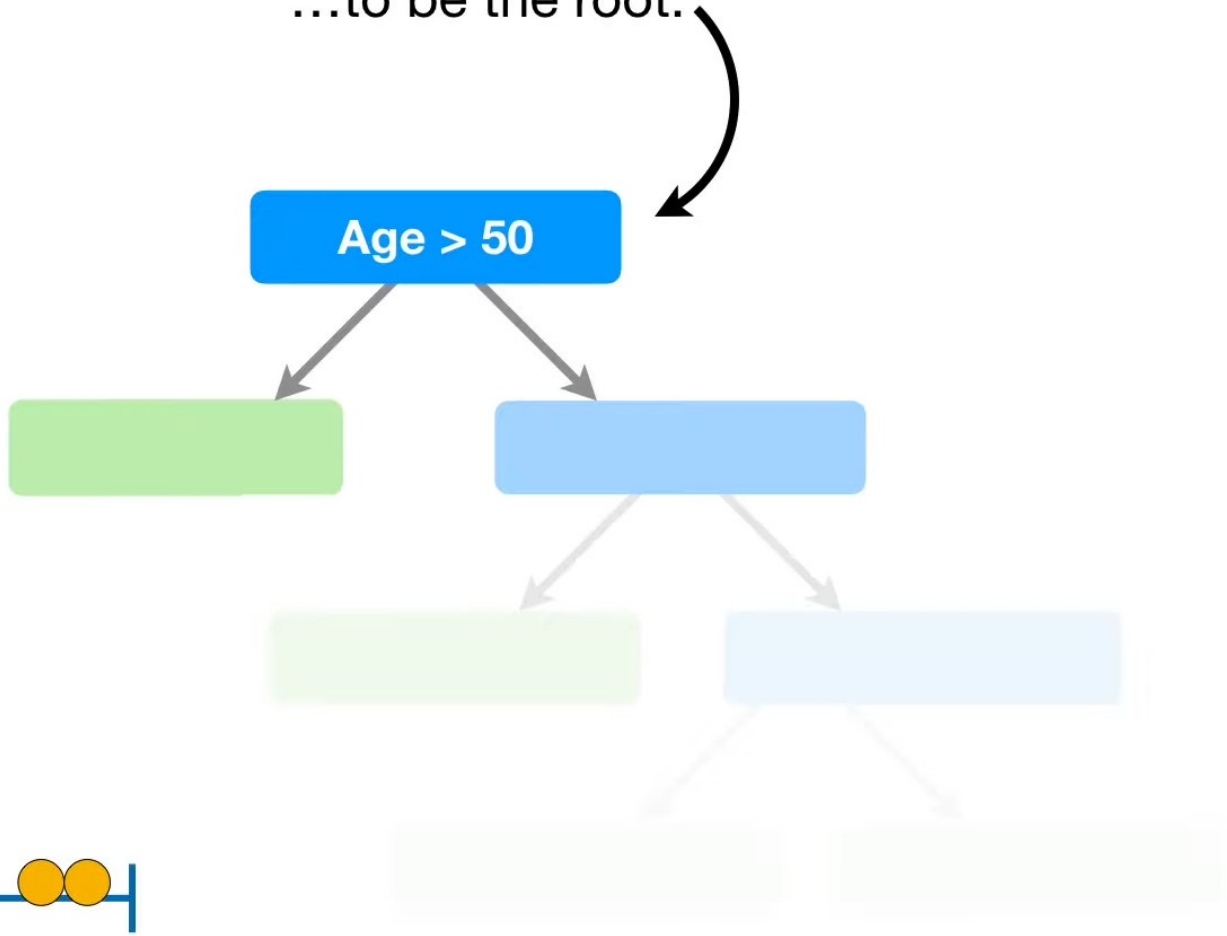
Average=12

Average=40

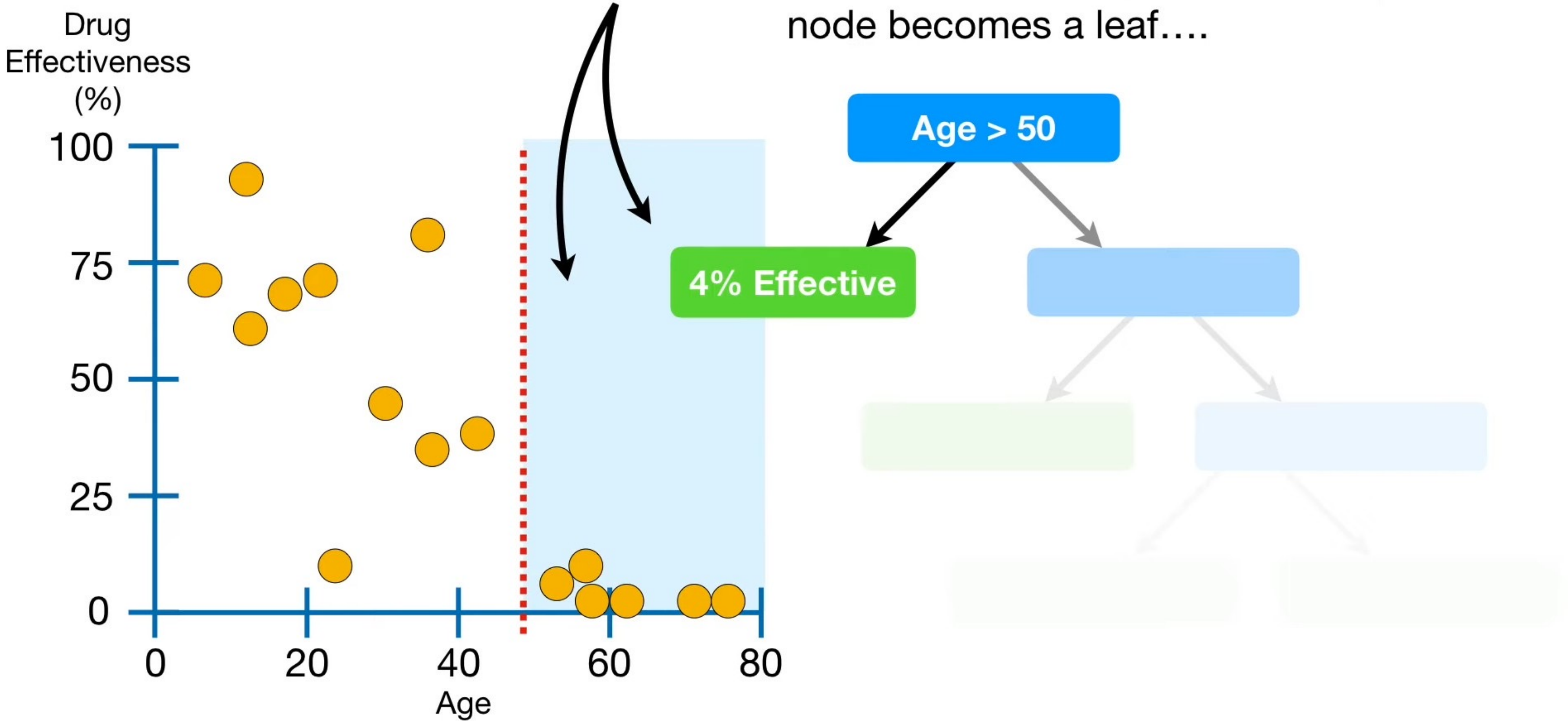
…and we pick the candidate
with the smallest sum of
squared residuals…



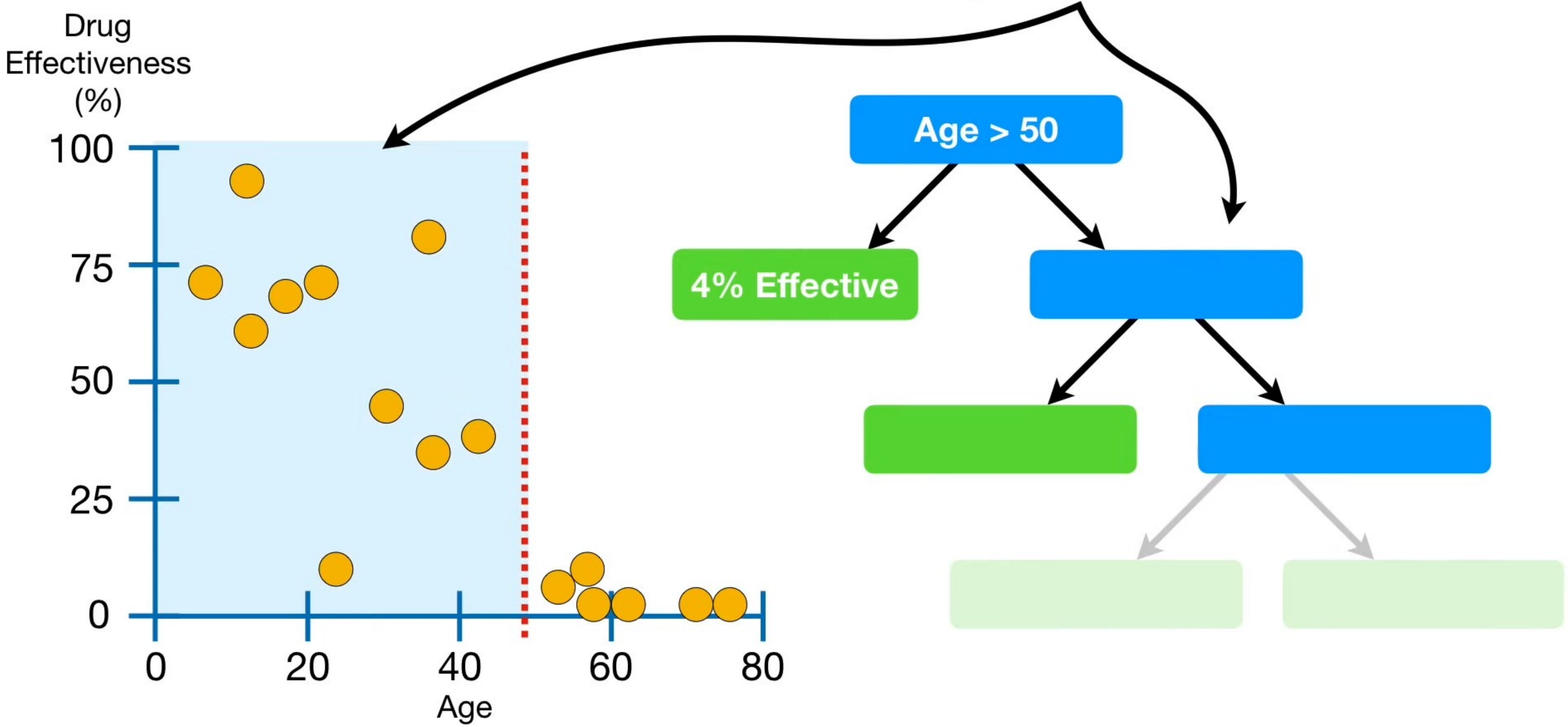
…to be the root.



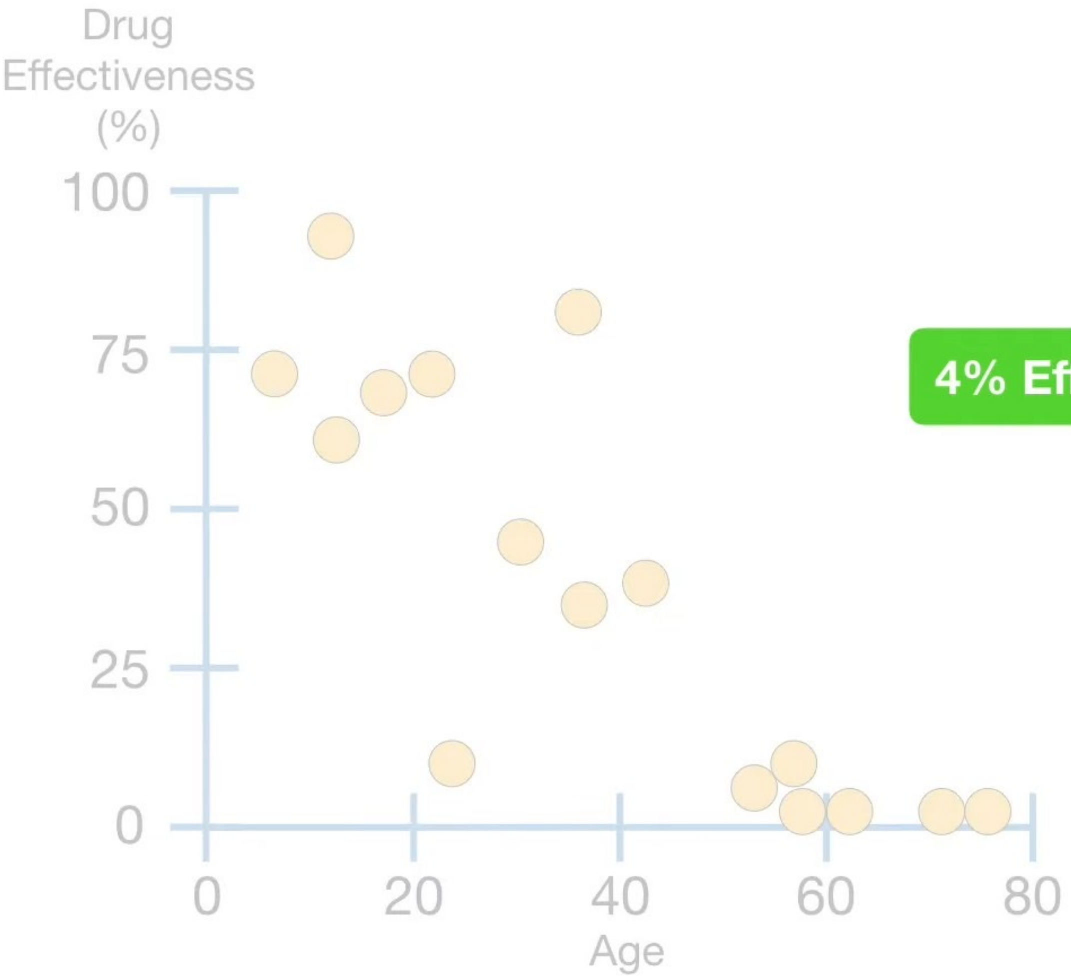
When we have fewer than some minimum number of observations in a node (**7** in this example, but more commonly **20**), then that node becomes a leaf….

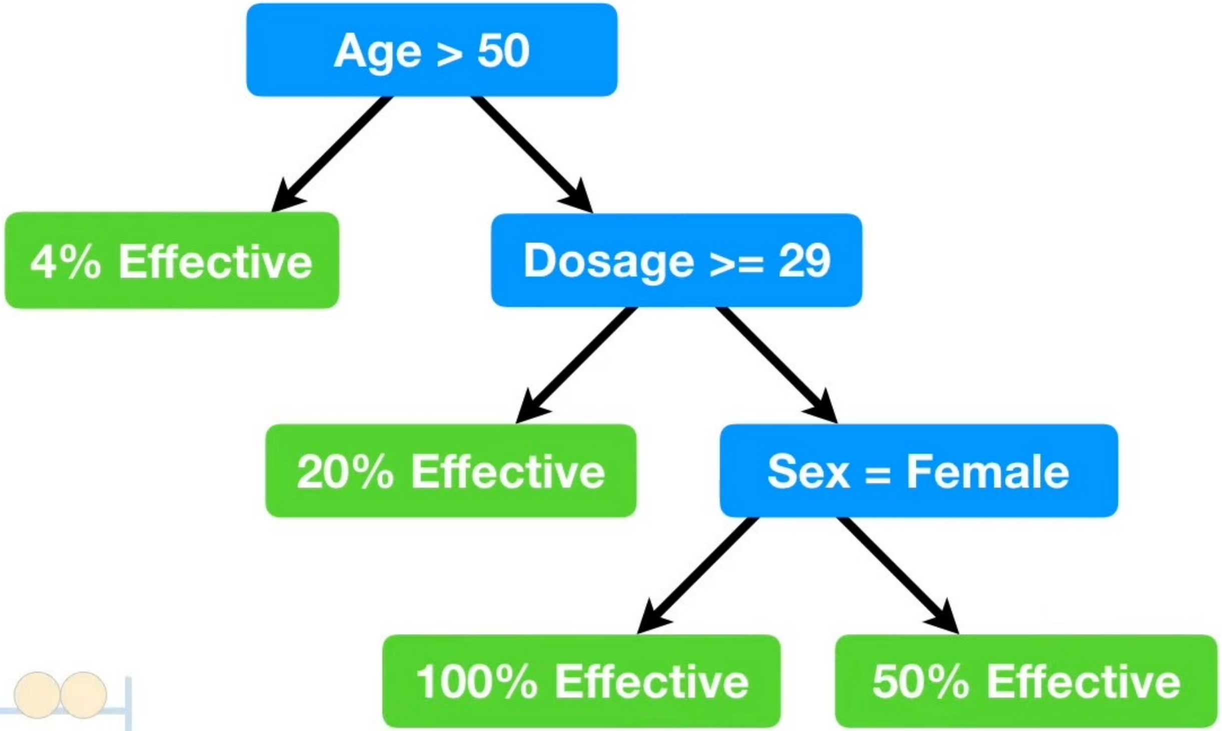


…otherwise we repeat the process to split the remaining observations…



…until we can no longer split the
observations into smaller groups…




Age > 50

4% Effective

Dosage >= 29

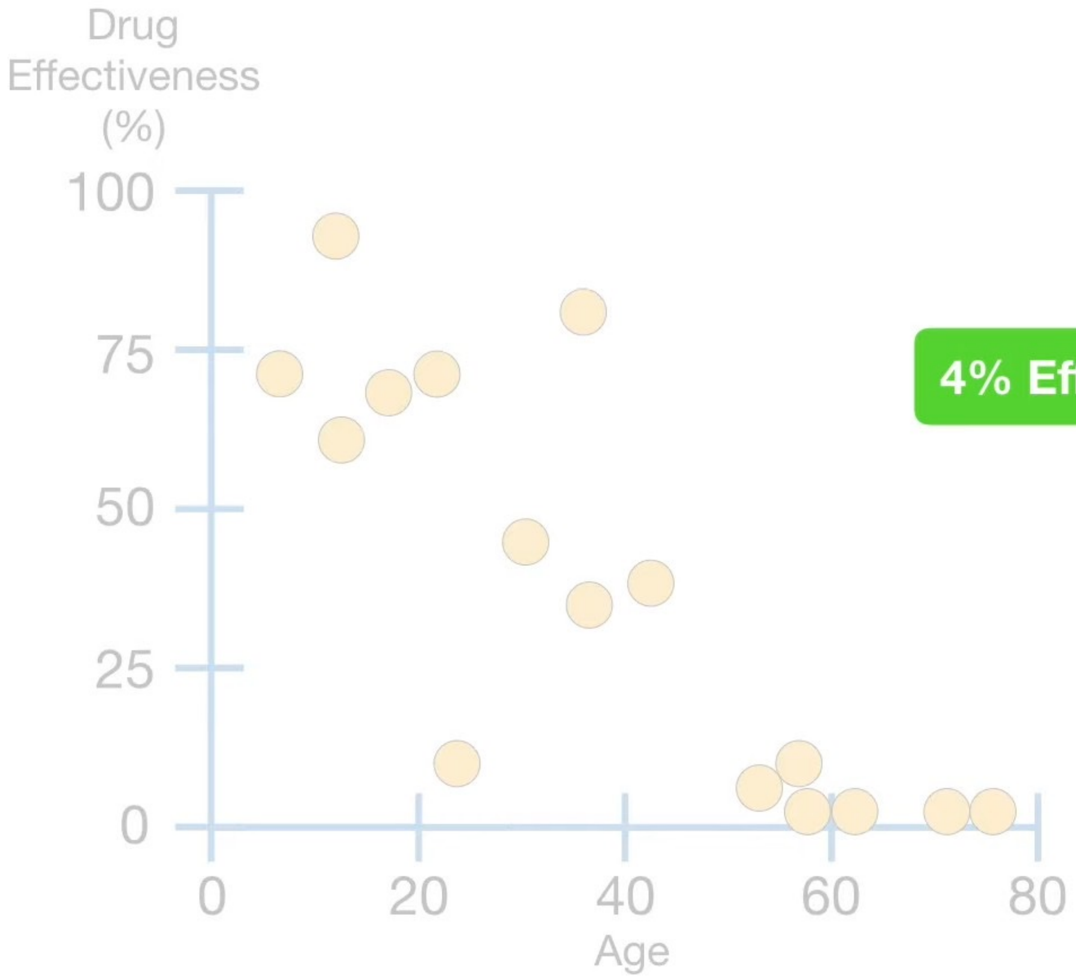
20% Effective

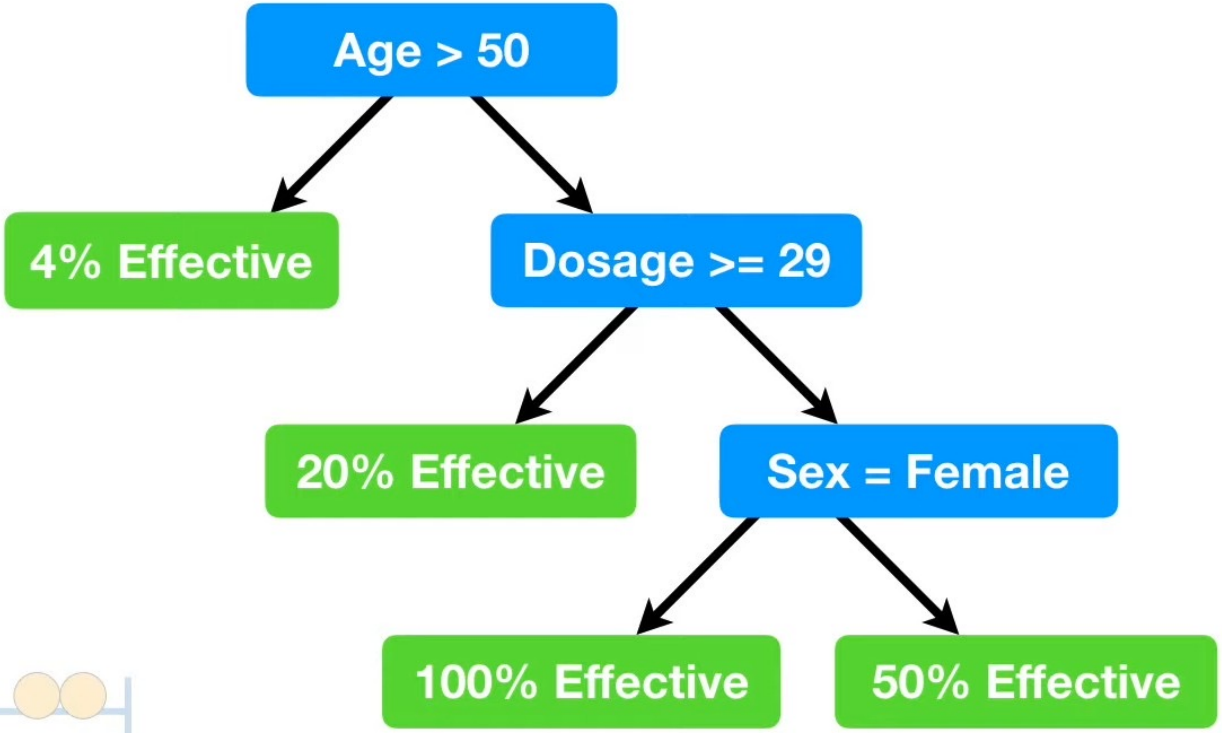
Sex = Female

100% Effective

50% Effective

…and then we are done.




Age > 50

4% Effective

Dosage >= 29

20% Effective

Sex = Female

100% Effective

50% Effective